MODELLING THE BEHAVIOUR AND PREDICTING THE SUCCESS OF EQUITYCROWDFUNDING INVESTMENTS

DR MUDIARASAN KUPPUSAMY AND RAASIANNA PARAMALINGAM

SUBMITTED TO

PITCH PLATFORMS SDN BHD

Asia Pacific University of Technology & Innovation (APU)

DECEMBER 2019

ABSTRACT

This study aims to introduce the application of data analytics in an ECF platform inMalaysia. Crowdfunding services are gaining traction in Malaysia, where more platforms are mushrooming as this area grows. Retaining investors and onboarding relevant campaigns would become a challenging phase; thus, taking the lead to understand and identify the potentials of the available data was the purpose of this study. An ECF platform in Malaysia has been engaged. Their anonymous investor's investment history and the successful campaign dataset were provided. Data structuring, pre-processing, data exploration, feature engineering, and dataset modelling were performed. Also, as there was a class imbalance issue with a small number of total observations, the SMOTE technique was explored to balance the class while increasing the observations. Cluster analysis to identify the investortaxonomy was performed using statistical (Hierarchical) and machine learning techniques (k-means).

Additionally, feature selection techniques were explored to identify factors influencing a campaign's success level in the platform using a Tree-Based Algorithm and Boruta technique. Also, dimensional reduction with factor analysis was exploredto identify the model performance with reduced factor. Finally, a prediction model was built to predict success using the Naïve Bayes, Random Forest and Support Vector Machine (SVM) techniques. Dataset Variation, Sampling Techniques, Feature Variation, Optimisation and fine-tuning of parameters were performed to compare and improvethe model performance. A confusion matrix was used to view the prediction versus actual score, where the Accuracy and AUC value was used to evaluate the model performance. Naïve Bayes with Laplace smoothing and SVM with polynomial kernel provided a good fit model utilisingthe features obtained from the SMOTE tree-based and factor analysis principal components.

Table of Contents

List of Figures

List of Tables

SECTION 1
INTRODUCTION

1.1 Introduction

The advent of the Industry 4.0 revolution is shaping the new business frontiers, including data-driven organisation growth. Big data analytics is one of the arms of the Industry 4.0 framework, where organisations use analytics initiatives for value generation. Organisations are doing massive spending to generate great insights of data available at hand, and the financial industry is also among the major industries riding on this bandwagon. The financial services have evolved where new services such as virtual financial services, blockchain technology, cryptocurrency, crowdfunding and many other financial technologies or fintech have emerged (Lee and Shin, 2018).

One of the prominent elements of fintech is Crowdfunding services, where this service has gained traction across the globe from developed to now developing countries. These crowdfundingplatforms arose in response to the 2008 financial crisis when traditional banking had imposedstringent regulations driving start-ups or SMEs with limited access to generating funds for their venture (Hakim Ghazali, 2019). However, this crowdfunding concept originated way back in1885 when Joseph Pulitzer launched a fundraising campaign to complete the construction of the Statue of Liberty via his newspaper. He then collected USD 100,000 from 160,000donors, young adults, politicians, street cleaners, and business people, where 75% of the donations were less than a USD dollar (BBC News, 2013). From newspaper then, to internet now. This internet-enabled form of raising money has a few different models that are donation-based, reward-based, lending-based, peer-to-peer (P2P) or equity-based crowdfunding(ECF) (UNDP (WorldBank, 2013; United Nations Development Programme), 2017).

In Malaysia, ECF is one of the crowdfunding establishments that started in 2015 with only six players than with ten licensed players in 2019 (SC, 2019). This crowdfunding modelis the fastest-growing crowdfunding model (Paschen, 2017). This gained strength when the Department of Statistics Malaysia (DOSM) reported a projected growth of 41 percent in2020 by the Small and Medium Enterprise (SME) compared to 37.1 percent in 2017 (DOSM

2017, SME Corp, 2015). Additionally, SME Corp Malaysia 2016 reported that most business establishments in Malaysia (98.5 percent) are from the SME industry with 66% of employment opportunities. This further drove the Malaysian Security of Commissions (SC) (2019)to establish the governance framework for this fintech industry in 2015 to cater to SME and start-ups' response growth and funding needs.

Since ECF was officiated in Malaysia, 51 successful campaigns were launched with a collection of 48.87 million (Fintech News, 2019; SC, 2019; pitchIn, 2018). Due to the market condition in 2018, there was a drop in the number of campaigns launched and capital raised. Despite that, PitchinIn grew further in 2018 to have a market share of 56% with 100% successful closed campaigns (Fintech News, 2019; pitchIn. 2018), followed by Ataplus (12%), CrowdPlus.asia (12%), Fundedbyme (11%) and Crowdo (9%) as illustrated in Figure 1 (Fintech News, 2019).

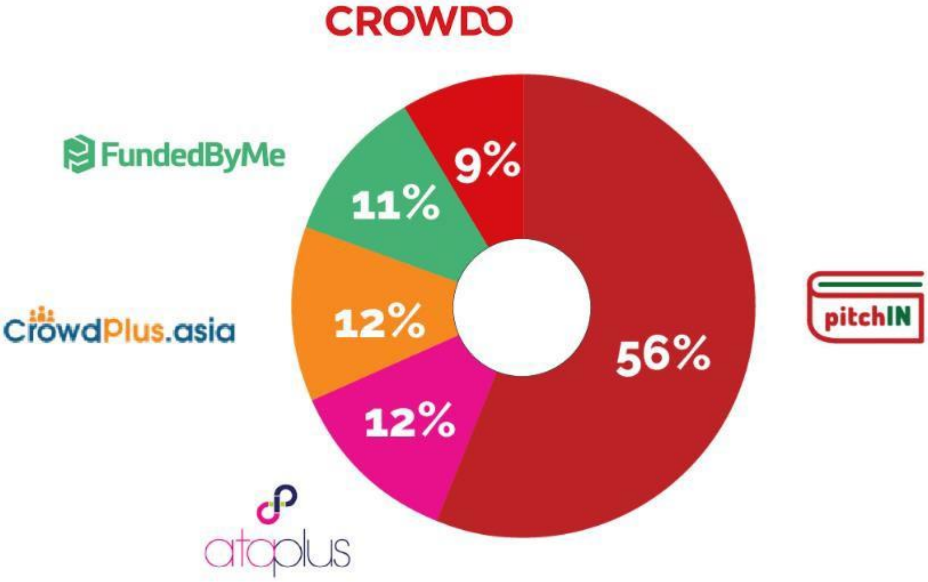


Figure 1: ECF players in Malaysia and their market share as of March 2019 (adopted from (Fintech News, 2019)).

The emergence of the crowdfunding financial model in Malaysia posed the question of what the ECF platform does and needs to remain relevant. Additionally, as investors are the main drivers of this ECF platform, it is crucial to understand the type of investors and their investment preferences based on their profile.

1.2 Research Background

The crowdfunding framework generally involves three parties, the platform, the fundraiser and the investors, also known as the backers (Yu *et al.*, 2018; Asia Institute of Finance Report, 2017). The fundraiser, a team of Entrepreneurs, would raise a fund request on the platform for the new or initial stage service or product they are venturing. The platform works as a governance body to onboard the campaign and performs due diligence before requesting fundraising. Then, investors are the main driver of the platform who decides which campaign to pledge based on their interest and motives.

There are multiple types of crowdfunding platforms, from donation-based, reward base, peer-to-peer (P2P) lending base, and equity crowdfunding (ECF). Each differs in its characteristics: the investor's investment motivation, who would be interested in such investment, the returns model, and the risk level. ECF is different from all other crowdfunding modelswith a small similarity with P2P. P2P provides a confirmed, quick and guaranteed return with interest paid compared to ECF; it provides equity share with a lock period for selling the share with the only dividend paid after a certain year of stabilisation. The risk factor for ECF being higher as the campaign that enters the platform is new start-ups or SMEs seeking initial seed to develop their idea or SMEs looking for early-stage financing. Thus, the success rate of the business is subjective as there could be the probability of the project failing in contra to the P2P platform. Therefore, most businesses are mature with stronger cash flow and a guaranteed return.

Table 1 details the comparison between the crowdfunding models based on their discriminant attributes. In general, donation and Reward crowdfunding attracts philanthropic investors who contribute for a good cause with no expectation of returns. However, a small gift or reward would be given to investors in the reward model. As there is an expectation for returns, both pose a medium to high risk from a risk level. The possibility for lenders to default payment may impact the P2P investors whereby for Equity funding, it is a high risk of losing all money if a project fails after the funding provided (Asia Institute of Finance*, 2017;* Lee and Shin, 2018; Paschen, 2017).

Table 1: Crowdfunding Models and their attributes (adopted from Asia Institute of Finance, 2017; Lee and Shin, 2018; Paschen, 2017)

	Donation	**Reward**	**Equity**	**P2P**
Purpose	Non-profit or philanthropic cause	Small business and creative projects	Generally, start up or SME business	Business or consumer
Return to funders	Nothing in return	A small gift or pre-purchase of product or service	Equity or share (ownership)	Interest at a fix rate and principal at a defined scheme
Examples	GoFundMe, GiveForward	Kickstarter, Crowdfunder, Indiegogo	AngelList, Early Shares, PitchIn	Lending Club Prosper
Risk Level	Not applicable	Not applicable	High as the business are new thus no guarantee of success	Defined to assist investors to understand the requestor
Investor's Intention	Philanthropic	Philanthropic	Social with profit goal	Profit goal

Furthermore, the crowdfunding platform has two different types of return models known as "All-or-Nothing" (AON) and "Keep-it-All" (KIA) model. The AON model only provides the entire collected fund to the entrepreneur once it meets the expected funding goal; else, money would be returned to the investors. In contrast, the KIA model would allow the entrepreneur to keep the collected money despite not achieving the funding goal (Cumming, Leboeuf and Schwienbacher, 2019). The AON model would have a minimum goal and the funding goal, which is the top goal. Thus, the campaign determines success as long as the minimum goal is met. However, there are multiple levels to the success, which could be categorised as the minimum goal, 50% of targeted goal and targeted goal

Most of the models here are AON, thus shifting the risk to the Entrepreneur and encouraging more investors to invest with minimal risk. The risk observed is when the successful project fails to launch; thus, the investors are losing the regular dividend pay-out and may be challenged to sell higher equity to other potential buyers. On top of that, there is another scenario of overfunded project where funds exceeded the funding goal. Then, the platform would participate on a first-come, first-serve (FCFS) mechanism in ensuring the initial investors who had transferred would obtain the investment privilege.

Therefore, having the platform return option and an FCFS mode, investors may need to respond to investment to participate in an investment promptly. As investors are the main driver, their motivation and behaviour towards the campaign may be influenced by many factors among them are the social presence or identity (Gerber and Hui, 2013; Nevin *et al.*, 2017), team members, linguistic (Yu *et al.*, 2018) and perception of information provided by the fundraiser (Tung and Liu, 2019). Furthermore, the sector type of the onboarded campaign mentioned by Vismara (2018) could also classify the investors and gender-driven influence (Mohammadi and Shafi, 2018). The need to understand the investors' investment interest, investment pattern and behaviour would influence the entire growthof the ECF platform (Hornuf and Schwienbacher, 2018; Xue and Sun, 2016).

Subsequently, from a fundraiser perspective onboarding, the right campaign that would entice the majority of the investor would be the main factor in driving the success of a campaign. Therefore, identifying the critical success factor of influencing the success of the campaign and being able to raise the minimum funding is necessary, as highlighted by several pieces of literature (Hornuf and Schwienbacher, 2018; Xue and Sun, 2016; Aprilia and Wibowo, 2018, Yu et al., 2018; Lee and Shin, 2018; Paschen, 2017*)*. Lastly, limited studies on equity crowdfunding platform around the Asia Region and Malaysia was highlighted by literature (Mochkabadi and Volkmann, 2018).

1.3 Problem Statement

The systemic growth of SMEs and start-ups in Malaysia stood at 98.5% (SME Corp Malaysia 2016), thus raising funding needs. The strict regulation by the traditional financing institution has made financing a challenge to the SME and start-up businesses, thus strengthening the ECF platform's growth. However, the ECF platform licencing has increased from six to ten (SC, 2019). Additionally, the economic downturn, especially during 2018, as highlighted by pitchIn

(2018) report, had reduced the number of campaigns raised and campaigns that were raised, struggling to obtain much investors' funding. These had triggered concern over the existing ECF provider on how to stay relevant in this competitive industry. The knock over effect of having more ECF platforms established wouldbe retaining the investors. As all the ECF and P2P platforms are dependent on the investorsas the main driver, investor retention and expansion of new investors are required. However, with the limited understanding of the investor's behaviour, especially on the existing investors,efforts in retaining existing investors may not be possible. Also, an onboarding campaign that isirrelevant to their intention would only drive failure to the campaign.

Similarly, understanding what signals determines the success of a campaign is unknown. These signals are what is observed by investors in driving their investment decision. As neither the platform nor the entrepreneurs are unaware of these signals, it serves as a disadvantage. The platform constantly onboarding campaigns that are irrelevant to the existing investors and Entrepreneurs spending extensive effort to ensure more information are provided. The effort is to waste as it does not fulfil the success signals. This indirectly increases investor churn, which would be a major drawback to the platform.

Finally, as Mochkabadi and Volkmann (2018) mentioned, limited studies around ECF signals and investors in Asia. Wallmeroth (2019) highlighted potential future research to identify different cultural influences impacting crowdfunding platforms and investor behaviour. This further strengthens the need to improve the body of knowledge for the ECF industry in different countries with different cultures and the investor's portfolio and interest from a different region (Xue and Sun, 2016; Wallmeroth, 2019).

1.4 Research Aim and Objective

1.4.1 Aim of Study

This study aims to undertake analytical modelling on the financial behaviour of crowdfunding investors belonging to one of the ECF platforms in Malaysia. This study would provide business insights, especially from an investor perspective, that would assist the ECF platform provider in better understanding their investors. Additionally, to assist the platform in predicting the funding level of successful campaigns based on success signals.

1.4.2 Objective of Study

Following are the objectives of this study:

- To construct the investor's financial behaviour taxonomy
- To analyse the success factors that influence the success level of an equity crowdfunding campaign
- To predict the funding level of a successful campaign based on the campaign's success factors.

1.5 Research Questions

Following are the questions to be answered through this study

- What would the segmentation of investors be like based on their behaviour in the ECF platform in Malaysia?
- How would a factor influence the success level of a campaign within the platform?
- Would these identified potential success factors be able to predict the success levelof those successful projects?

1.6 Scope of Study

The scope of this study would be on the leading ECF Platform in Malaysia, PitchIn. The researcher obtained the dataset from this ECF leading platform. Thus, the scope of data would be within the secondary dataset received from the ECF organisation. This study analyses investors' data against the transactions performed on campaigns in the platform and the campaign's information. The total number of campaigns that would be analysed is 35 successful campaigns and 1645 active investors who had performed transactions within those campaigns. The data to be analysed will be from 2016 to August 2019. The campaigns on the platform are limited to companies within Malaysia; thus, the study would project the ECF performance and growth for Malaysia. The cluster analysis technique is used to identify the different investor types in the platform. Then,the feature selection techniques to identify the most influencing factors in determining the success level of the successful campaign, while dimensional reduction technique factor analysiswas performed to identify the best model. Finally, machine learning techniques Naïve Bayes, Random Forest and SVM was adopted to predict the success level of the successful campaigns.

1.7 Significance of Study

As the projected growth of ECF to be continuously high and the increase of start-ups and SMEs in Malaysia, thus significantly increases the competition between the ECF platforms in Malaysia. Investors, Businesses and their campaigns onboarded are the key factors influencing the growth of the ECF platform. Therefore, ensuring the right and relevant investors are onboarded, the business provides the right information. So, the insights generated from this study would benefit the ECF platforms, investors, businesses (SME or start-ups) and relevant stakeholders such as researchers, students and individuals.

1.7.1 Existing ECF Platform

The existing ECF platform would understand their investor's investment behaviour and the factors that influence their investment decision. Additionally, these insightswould also assist the ECF platform members in decision making of the campaigns to onboard based on the preference of their investors. Also, this outcome could assist the organisation in managing the investors churn also serve as a marketing tool to attract new investors. Besides that,this information also could be utilised as an added service for business to strategise their campaign plan. By doing so, the platform would stay competitive in this growing industry.

1.7.2 Investors

This study would benefit investors, especially new investors seeking to explore the ECF platform for economic diversity. They would understand the factors driving investment behaviour, especially seasoned investors. Knowing seasoned investors behaviour, all investors would be able to follow suit their investment behaviour. Besides, an investor would identify the relevant signals while deciding on their investment.

1.7.3 Business

Entrepreneurs would be able to subscribe to the platform's analytics to understand better the investors profile in ensuring the campaign objective and returns are tailored based on the investors' preference, indirectly attracting more investors and ensuring a successful campaign. Savings on time to ensure only relevant details are planned for their campaign and have targeted marketing on the relevant potential investors.

SECTION 2
LITERATURE REVIEW

2.1 Introduction

The focus of the literature reviews was performed on the ECF domain. However, the practices on other crowdfunding domains were also reviewed in identifying best practices for cross-adoption. The area of studies is categorised into five different categories: Entrepreneur Perspective, Capital Market Perspective, Institutional Perspective, Investor Perspective, and Platform Perspective. Refer to Table 2 on the definition of each categorisation and its sub-themes(Mochkabadi and Volkmann, 2018).

Table 2: Five themes on ECF domain journal contribution

Themes	**Sub-Themes**
Entrepreneur Perspective	• Rationale for ECF • Factors determining the success of a campaign • Gender Issue
Capital Market Perspective	• Functioning and Development of ECF • Potential Role
Institutional Perspective	• Impact of Law • Comparison of legal conditions • Contracting Practices
Investor Perspective	• Investment motivation • Investment evaluation • Investor Type • Investment Dynamics • Return of Investment
Platform Perspective	• Platform Design • Shareholder risk

Thus, the two domains that would be our key focus are the Entrepreneur and Investor perspectives. Factors determining the campaign's success from an entrepreneur's perspective would indicate the signals that investors would look upon while making investment decisions, thus predicting the success level of a campaign. Additionally, from an investor's perspective, investors would be reviewed on their motivation for investment and what drives their investment decision.

2.2 Investor Perspective

The investors' perspectives include the motivation that drives an investor towards ECF, the thought process on decision making, and the type of investors in the ECF domain utilisingpast investment history data. The subsequent section will further elaborate on this investor perspective.

2.2.1 Investor Motivation

Motive is a developed and content-specific physiology disposition, while motivation is the behaviour when a motive is activated (Bretschneider and Leimeister, 2017). The general motivation in an ECF platform is financially driven due to the monetary returns one would obtain by investing in a campaign (Bretschneider and Leimeister, 2017; Goethner, Luettig and Regner, 2018; Moysidou and Spaeth, 2016; Cholakova and Clarysse, 2015). Among the motivations that was studied by the literatures were financial (Bretschneider and Leimeister, 2017; Goethner, Luettig and Regner, 2018; Cholakova and Clarysse, 2015), altruism, recognition, lobbying, image, liking (Bretschneider and Leimeister, 2017), pro-social, community (Bretschneider and Leimeister, 2017; Goethner, Luettig and Regner, 2018; Cholakova and Clarysse, 2015; Mohammadi and Shafi, 2018) and herding (Bretschneider and Leimeister, 2017; Hornuf and Schwienbacher, 2018).

Bretschneider and Leimeister (2017) performed an empirical study on a model developed regarding backer's motivation on incentive-based crowdfunding, e.g. equity, reward and lending platform. They concluded a significant negative correlation between recognition motivation towards investment was due to certain information, e.g. amount funded in this example was not made visible. Thus, recognition would be received despite the amount invested.

Altruism motive does not exist in his scenario. The ECF nature does not uplift poverty or donate money to an SME in developing countries, especially those in need. Other motives such as lobbying, liking, financial, herding, and image are the main reasons for individual participation in an ECF platform.

Europe based crowd-investing platform where investors motives and strategies were investigated (Goethner, Luettig and Regner, 2018). Goethner, Luettig and Regner (2018) investigated the financial, community and social motivations together with the investors experience to determine the clusters of investors. An investor's experience is determined by the number of projects invested and the average amount invested. As the ECF platform is financially driven with equity return, the diversification of the investment portfolio was reviewed to identify the risk-reducing strategy a financially motivated investor would analyse before making decisions. Also, calculating the participation share (average share per amount spent) represents the investor's financial reward, thus an indicator of a financially motivated investor (Goethner, Luettig and Regner, 2018).

As for socially motivated investors, the innovation indicator identifies them and an investor's experience. Innovation could be flagged to a campaign when the following element is noted 1) Patent or Trademarked applied 2) Significant R&D 3) Serves a market where no direct competitor 4) the only service or product provider in the market. Therefore, there is a high risk of such a project failing due to its infancy stage. Thus, this indicates socially driven investors who wish to encourage ideas rather than be financially motivated. The average share per EUR5 would be low for socially driven investors (Goethner, Luettig and Regner, 2018).

A sense of community also would be visible among investors via the average number of backers and financial motivation indicators. When the financial motivation is low, the higher average number of backers indicates community motivated individuals (Goethner, Luettig and Regner, 2018; Mohammadi and Shafi, 2018). The herding behaviour could clearly distinguish this as for herding behaviour. The financial motivation indicators would be higher in comparison to a community-driven individual. Hornuf and Schwienbacher (2018) investigated the investment dynamics where one of the motivations studied was herding behaviour. Here, he observed if an investor had invested or withdrawn its pledge amount from investment due to the influence of a sophisticated or more experienced investor. This was observed by the amount pledged, and the study confirmsthe existence of herding motivation by investors.

2.2.2 Investor Evaluation

The Decision-making process is the thought process an individual would perform before decidingon any situation that requires an action to be taken. The evolution of decision making from rationally driven model to later cognitive model and now the cognitive-affective model. Consumer decision making is influenced by the available information, the cognitive limitation and the finite time for decision making (Moysidou and Spaeth, 2016). Moysidou and Spaeth (2016) confirm that ECF is a more complex crowdfunding structure, thus having a more cognitive approach. Cognitive decision-making is data-driven, rational and practical where one would learn, develop its knowledge, think, and make decisions based onthe analysis done. In comparison, effective decision making is purely irrational, impulsive and intuitive driven were one with the information at hand and the feeling at that moment determining the decision was taken (Moysidou and Spaeth, 2016).

Functional, social, emotional, epistemic and conditional were the five perceived factors determining consumers' decisions. Later studies concluded that consumer decision making is driven by two dimensions of functional and affective, where functional focus on rational and economic evaluation. In contrast, affective is influenced by emotional and social aspects (Moysidou and Spaeth, 2016).

Moysidou and Spaeth (2016) proposed a framework combining the cognitive and affective model where financial, functional, and information value is grouped as cognitive. One would think, understand, and interpret a campaign before deciding. At the same time, the other part is effective, which is influenced by emotional, social, novelty and aesthetic values. His motives were to identify the difference of values and their effects on a different form of crowdfunding, especially equity, loan, and presales. ECF nature is predominantly cognitive, rational and data-driven, where functional and informational are the main values that drive this crowdfunding platform. The latter concluded that a backers' decision depends on the type of crowdfunding performed (Moysidou and Spaeth, 2016).

Several studies were performed to identify investing interests (Hornuf and Neuenkirch, 2017; Zheng, 2016). Some additionally looked into deciding the amount to invest (Zunino, van Praag and Dushnitsky, 2017). The evaluation could be divided into four different perspectives from a fundraiser perspective (Zunino, van Praag and Dushnitsky, 2017; Zheng, 2016), project perspective (Zunino, van Praag and Dushnitsky, 2017; Hornuf and Neuenkirch,2017), platform perspective (Zunino, van Praag and Dushnitsky, 2017) as well as investor perspective (Hornuf and Neuenkirch, 2017).

From a fundraiser perspective, the human capital information, such as their skills and past failure or success, was taken into account in deciding for investment and the amount to invest (Zunino, van Praag and Dushnitsky, 2017). Zunino, van Praag and Dushnitsky (2017) concluded that the stigma attached to failure where the past failure would deteriorate future campaigns does not exist. When there is a potential good signal of their skills, the past failure vanishes. Thus, new campaigns would deem new and willingness to invest would increase. Zheng(2016), on the other hand, viewed value congruence and social interactions ties. The similarity between the fundraiser and funder creates trust, thus enhancing the willingness to invest. However, the social interactions from an information flow between the parties via the platform do not significantly affect the willingness to invest (Zheng, 2016).

Progress in funding campaign, including operational costs, audio or video media and frequent updates (Zunino, van Praag and Dushnitsky, 2017; Hornuf and Neuenkirch, 2017), pre-valuation, funding goal are among the characteristics of a campaign that is observed in the decision-making process (Hornuf and Neuenkirch, 2017). A platform also impacts the decision to invest. Zheng (2016) concluded that the perceived accreditation on the platform might deem the platform reliable and drive more investors to the platform, thus indirectly trusting the fundraiser and campaigns. Investors' experience, income level, number of pledges, and the average amount spent is the investor’s characteristics that determine their investment willingness to invest (Hornuf and Neuenkirch, 2017).

2.2.3 Investor Type

Several studies have explored the difference between investors. Gender was the first differentiationamong the investor type. Mohammadi and Shafi (2018) investigated the gender differences in investment pattern, where he found that female has a risk-averse attitude and they are contributors of herding behaviour especially having a biased opinion that a male investor's decision would be greater compared to another female (Mohammadi and Shafi, 2018). Wallmeroth (2019) also highlighted that males invest more predominantly, and females invest more in less risky projects. However, when a certain amount of investment is made (large-amount-investment, EUR 5,000), the gender is no longer significant, indicating that men and women invest at an equal rate (Wallmeroth, 2019).

Lin, Boh and Goh (2014) identified four main investor types: active, trend follower, generous, and crowd. This has a different motive that drives its characteristic, as detailed in Table 3. The authors then further broke down subtypes for the large composition group Crowd (55%) and identified similar grouping as the main type active, trend followers and crowd except altruistic as the motive of crowd investor is financially driven. Thus, this would not be visible in the subgroup. Despite not having a distinct character as the main type but a mild characteristic within the main type was noted (Lin, Boh and Goh, 2014).

Goethner, Luettig and Regner (2018) identified three types of investors: the Sophisticated Investors who are very active and experienced but comprise a small group of people. Additionally, crowd enthusiasts are motivated by pro-social campaigns, and most individuals are casual investors who are merely motivated on the monetary returns (Goethner, Luettig and Regner, 2018). Large-amount-investment were observed to invest less frequently and in fewer industries indicating a business-angel-like investor type where they would only perform minimal investment based on utilitarianism rather than emotion (Wallmeroth,2019). This is similar to the Sophisticated investors has highlighted by Goethner, Luettig and Regner (2018).

The amount invested and the number of investors at the beginning and end of the campaign was reviewed by Abrams (2017) and concluded that the first-week investment was performed by family, friends and fools where campaigns with huge debt and minimal assets were supported. It also concluded that a sophisticated investor would only invest after the first seven days after digesting the information provided by the fundraiser. The information that drives more sophisticated investors is those with less debt, more assets and more information produced to the US SEC board (Abrams, 2017).

Distance effects on investment were also explored by scholars Guenther, Johan and Schweizer (2018). He studied the influence of geographical distance and its influence on the investor's decision. Here, he concluded that home country investors would be sensitive to distance as they are based in the fundraiser's local area, thus making communication smoother, especially when technology advances are not fully incorporated. On the other hand, overseas investors are not sensitive to distance, thus investing in campaigns not within their geographic distance. This is technology dependent and indirectly creates home bias, especially when distance-sensitive platforms (Guenther, Johan and Schweizer, 2018).

Wallmeroth (2019) assessed the profile of the individuals who invested less frequently. However, these individuals have contributed 51% to the raised capital of the first 59 campaigns; however, only 3 per cent of the total investment was performed. Thus, identify the profile and the reasoning of this individual not investing more. He concluded two profiles of newcomers and sophisticated investors with different characteristics. They needed the platform to invest in marketing effort or retention plan to encourage more new investors and perks to retain the existing members to avoid churn (Wallmeroth, 2019).

Table 3: Investors Type details, motives and characteristics

Studies	**Crowdfunding Type**	**Investor Type**	**Motives**	**Cluster (%)**	**Characteristics**
Mohammadi and Shafi, (2018)	ECF	Gender	Financial, Herding	NA	female are risk-averse, herding behaviour especially mirroring the male investors
Wallmeroth (2019)	CI (loan, investment-based)	Gender	Financial, Herding	NA	male investing more predominantly, female invests on less risky projects, certain amount of investment made gender is no longer significant
		Sophisticated	Financial, Utilitarian	NA	Invest less frequently, less diversified portfolios, large amount invested, fewer comments, less likelyto be a returning investor (fewer active days)
		Newcomer	Financial	NA	large amount invested less frequently
Lin, Boh and Goh (2014)	Reward-based	Active	Social, Recognition	9	Back large projects, project creators themselves, post more comments, invest in diversified portfolios
		Trend Followers	Herding, Financial	24	Risk-averse (a back project with a large number of backers, back projects with small average goal),back less risky but highly popular project
		Altruistic	Prosocial, Community	12	Back projects that have no reward, Less risk-averse (a back project with high average goal), the fewer number of backers

		Crowd	Financial	55	Focused on reward, Risk-averse (back projects with small average goal), smaller number project, limited portfolio diversification, do not create the project, likely to leave comments
Goethner, Luettig and Regner (2018)	ECF	Sophisticated	Social, Recognition, Financial	4	Very active and experienced, small group of people, high average amount invested, actively commenting,
		Crowd	Prosocial, Community	35	Low comments posted, low investment, the high amount invested on innovative projects, low average participation share, the highest number of investors per project
		Casual	Financial	61	The highest number of funded projects, the lowest amount invested per project, lower share on innovation projects, the average number of investors is small; participation share is highest, a small amount in less innovative projects, less risky with high interest
Abrams (2017)	ECF	Sophisticated	Financial	NA	Less debt, more assets, more information disclosed to regulatory board
Guenther, Johan and Schweizer (2018)	ECF	Distance		NA	Overseas investors are not sensitive to distance whereby home investors do

Hornuf and Schmitt (2016)	ECF	Family & Friend	Liking/Personal Connection		Less response to comments, invest in focal start-up rather than any other start-up, large investment in their focal start-up, no more than three other start-ups, local bias
		Angel-Like Investors	Recognition		Invest high amount, the main driver for campaign success, invest during the day time and weekdays, a large amount beyond EUR 5,00 is local bias except if the contribution was made within three days.
		Diversified	Financial		High financial literacy, the higher average amount spent, not bias

The risk-averse perspective was brought by using the number of investors invested in a campaign, thus driving trend followers to pledge when there are many backers (Lin,Boh and Goh, 2014). This was interpreted differently by Goethner, Luettig and Regner (2018),where a pro-social or community-driven crowd enthusiast could be identified with a similar variable. Additionally, Lin, Boh and Goh (2014) investigated the altruistic type of investors by observing the reward return compared to Goethner, Luettig and Regner (2018), who observed the number of investors on a campaign to determine if it was socially driven. Both these studies were performed in two different types of crowdfunding environments. Goethner, Luettig and Regner (2018) performed on an ECF environment and Lin, Boh and Goh(2014) performed on a reward-based crowdfunding platform. ECF platform is financially motivated while Reward-based has an option of not receiving reward thus the difference in motivation and variable for observation as documented in Table 3.

Both the literature used the crowd terminology; however, the interpretation of motive was different. Lin, Boh and Goh (2014) were financially driven, and Goethner, Luettigand Regner (2018) referred to the prosocial driven individual. The composition of members differs between both. Thus, the mapping for the crowd's literature would be as documented in Table 4. Here, Goethner, Luettig and Regner (2018) did not observe the herdingbehaviour compared to trend follower by Lin, Boh and Goh (2014).

Table 4: Comparison of investor type between Reward-based and ECF

Reward Based (Lin, Boh and Goh, 2014)	**ECF (Goethner, Luettig and Regner, 2018)**	**ECF (Wallmeroth, 2019)**	**ECF Hornuf and Schmitt (2016)**
Active	Sophisticated	Sophisticated	Angel-like, Diversified
Altruistic	Crowd	NA	NA
Crowd	Casual	NA	NA
Trend Follower	NA	NA	NA
NA	NA	Newcomer	NA
NA	NA	NA	Family & Friend

Both studies were performed on the ECF platform Goethner, Luettig and Regner (2018) and Wallmeroth (2019), where both identified the similar characteristic for sophisticated users to be investing a large amount. However, Goethner, Luettig and Regner (2018) had a different perspective than Wallmeroth (2019) on the sophisticated individual's motive to have social and recognition-driven this group would provide. Thus, more comments on the project were identified as their characteristic. Wallmeroth (2019) highlighted the less diversified portfolio that sophisticated individuals would adopt with very less frequent on their investments.

As we have observed all the scholars and their grouping of investor types, a summary of our discussion comparing the characteristics and the expected outcome for each type of investor is documented in Table 5. The characteristic could be grouped to the investor profile, investor funding history, campaign, firm and social interaction. Therefore, based on thesummary of previous studies, we would look at the following motives to perform the clusteringactivity:

Altruism, Pro-Social or Community would support a campaign indirectly to support the idea and cause for the community. This group tends to support new innovative campaigns with the high risk involved. This could also be observed in the average goal funded by an investor where investment is performed on a high funding goal campaign. However, it will drive the failure of a campaign byproceeding to do so. It highlights a sense of altruism or community in the cluster.

Financial Motives are individuals who are heavily attracted to the financial returns from the investment. Less innovative or community elements exist in this cluster.Additionally, this group of people are risk-averse where they would diversify their investmentand low amount invested per project but still have the highest number of projects backed.

Table 5: Summary of characteristics by investor type for ECF domain

Area	**Characteristic**	**Sophisticated**	**Altruistic**	**Casual**	**Trend Follower**	**Newcomer**	**Friend & Family**
Investor Profile	**Active**	Yes	Yes				
	Experience	High	Med				
	Composition	Small	Moderate	Majority			
	Campaign Creators	Yes		No			
	Gender Significance	No		Yes	Yes		
	Local Bias	No with condition					Yes
Investor Funding History	**Amount invested**	High	Low			High	
	Average amount invested	High		Low		Low	
	Average amount invested – Innovation		High	Low			
	Average amount invested – Start-up						High
	Number of funded campaigns	Low		High		Low	
	Average Participation Share		Low	High			

Campaign	**Innovative**		Yes				
	Number of Investors		High	Low	High		
	High Interest			Yes			
	Portfolio Diversification	Low					
	Average Goal		High	Small	Small		
	No more than three start-up						Yes
Firm	**Low Debt**	Yes					
	More assets	Yes					
Social Interaction	**Comments**	Active	Low	Low			Low

2.3 ECF Success Factors

Information asymmetries are a major concern in this crowdfunding domain. Short investment duration, limited investment experience, and inadequate face to face sessions with the campaign creator trigger this problem. Thus, many studies identify campaign signals that could assist investors with their investment decision. These signals could be observed from aninvestor perspective, the campaign, investor's past investment history, the firm and platform. These signals were used to identify successful campaign (Vismara, 2016; Stebro *et al.*, 2017; Hervé *et al.,* 2019; Vulkan, Åstebro and Sierra, 2016; Block, Hornuf and Moritz, 2018; Li *et al.*, 2016; Piva and Rossi-Lamastra, 2018), identify the number of investment made in a particular day (Hornuf, Lars; Schwienbacher, 2015; stebro *et al.*, 2017), amount pledged (Stebro*et al.*, 2017; Vulkan, Åstebro and Sierra, 2016) and total number of investors (Piva and Rossi-Lamastra, 2018).

Signals used in the ECF platform could be categorised as human capital (Li *et al.*, 2016); Goethner, Luettig and Regner, 2018; Piva and Rossi-Lamastra, 2018) which comprises of management team size, education of team members, industrial and entrepreneurial experience. The The percentage of equity offered to investor is another type of signal indicating equity retention(Vismara, 2016a). There are social capital signal which indicates the number of contacts in social network together with communication signal on number of contents provided and updates being made to the content (Block, Hornuf and Moritz, 2018; Li *et al.*, 2016; Li, Rakeshand Reddy, 2016; Yu *et al.*, 2018; Vismara, 2016a). Other third-party signals include investors as partners, product certification, social proof, prominent affiliates, intellectual property rights as we as grants (Goethner, Luettig and Regner, 2018). Campaign characteristics is also another type of signal that could explain an ECF using information such as capital gained at first week of investment, largest single investment, # of investors, prior funding amount collected, minimum investment amount, campaign duration, business to customer orientation (Li *et al.*, 2016; Vismara, 2016; Vulkan, Åstebro and Sierra, 2016). Finally, the last type of signal is on post campaign where number of management team members, presence of professional investors, second successful campaign could further indicate the success of ECF campaign (Mochkabadi and Volkmann, 2018).

2.3.1 Human Capital of project investor

Li *et al.* (2016) and Goethner, Luettig and Regner (2018) conclude from their findings that human capital plays an important form of identifying the firm's involvement. Here, information such as the team’s education level (Goethner, Luettig and Regner, 2018) or the total number of full-time workers and business age (Li *et al.,* 2016) could be used as an indicator of human capital of the project investor. Piva and Rossi-Lamastra (2018) specifically highlighted the education level of the entrepreneur and the differentiation between their education and their working experience. Where for education, he analysed the difference between business-related education in comparison to industry-related education. Additionally, he compared the entrepreneurial experience to an industry specific experience. Business-related education and entrepreneurial experience contribute to a campaign's success (Piva and Rossi-Lamastra, 2018). Goethner, Luettig and Regner (2018)’s study highlights that Sophisticated and altruism investors would mostly invest on campaign that has human capital signal.

2.3.2 Funding Stage

Early investment serves as an indicator to other late investors that the campaign has value and could be trustworthy. Thus, when there is huge amount or sophisticated investors invested during the early stage of the campaign, this would attract more investors to invest. Additionally, if there is no lead investors but high amount of investment been collected during the early days, this is a strong indicator of successful project(Vismara, 2016). Vulkan, Åstebro and Sierra (2016) highlighted the importance of having an initial strong start where a high percentage of amount invested in week 1 signals a strong growth for a campaign. The campaign goal, single backer pledge large amount as well as the number of backers in a campaign is a great signal for a successful campaign. Goethner, Luettig and Regner (2018) also highlighted on the prior sophisticated investors does not influence the investment of casual investors. In contrast, altruism investors tend to follow the decision of experienced peer investors (Goethner, Luettig and Regner, 2018). Lin, Boh and Goh (2014) identified that trend followers tend to back project at the later stage in comparison to Sophisticated, Altruism and Casual investors whom would invest in a platform at early stage.

2.3.3 Social Influence

Vulkan, Åstebro and Sierra (2016) highlighted the importance of having an initial strong start where a high percentage of amount invested in week 1 signals a strong growth for a campaign. The campaign goal, single backer pledge large amount as well as the number of backers in a campaign is a great signal for a successful campaign. Li *et al.* (2016) found otherwise where lead investor's investment information had negatively affected the number of backers potentially due to the assumption that this lead investors has a connection with the fundraiser themselves. Lin, Boh and Goh (2014) found that a project with the number of backer pledge would measure the social influence of the future backers to back for a project. Trend followers tend to back project with large number of backers whereas the Altruistic backers tend to have smaller number of backers.

2.3.4 Investor Competence

Goethner, Luettig and Regner (2018) tested on disclosure of financial projection such as planned revenue, expenditures, earning before investment and taxes as information to highlight on this signal. His findings highlights that Sophisticated and casual investors would mostly invest on campaign that provides all this financial details compared to crowd investors (Goethner, Luettig and Regner, 2018). The depth of how investors assess the financial information, or the campaign information provided in making a rational investment decision has limited studies (Nitani and Riding, 2017). Nitani and Riding (2017) embeded this signal to identify how it influences the success rate of a campaign. The findings indicated that prior start up experience, age of firm, EBITDA margin does have an influence of determining the success of a campaign (Nitani and Riding, 2017). Therefore, where the risk level is high, potentially the investment by investors may be minimal thus the funding goal may not be met rather the minimum goal would be met.

2.4 Analytical Method

The previous studies was reviewed on the techniques used to analyse the investor's behaviour as well as examine the different investment decision made by the different investor's portfolio based on the campaign signals.

2.4.1 Cluster Analysis

The exploratory clusters analysis techniques was adopted by both Lin, Boh and Goh (2014) and Goethner, Luettig and Regner (2018) to investigate if the investors could be distinctly differentiated. Two stage cluster analysis procedure was adopted where to identify the appropriate number of clusters, the hierarchical clustering was employed. Then, to optimise the validity of final clustering, the non-hierarchical approach was taken using the k-means technique (Lin, Boh and Goh, 2014; Goethner, Luettig and Regner, 2018). Both these studies performed the similar cluster analysis techniques, however the difference was on the features observed as both analysis was performed on a different crowdfunding platform Lin, Boh and Goh (2014) on Reward-base and Goethner, Luettig and Regner (2018) on ECF. Multicollinearity needs to be treated by selecting variables that are not highly correlated. This could be performed with by performing the factor analysis technique (Hair et.al, 2013). Both these studies did not perform the factor analysis as it had employed features that was confirmed by previous study. Table 6 summarises the techniques used for cluster analysis by the prior similar studies.

Table 6: Cluster Analysis Technique Details with Crowdfunding dataset

Journal	Technique	Features (motives - variable)	Transformation	Distance Matrix	Cluster Agglomeration	Cluster validation
Goethner, Luettig and Regner (2018)	2 stage clustering – hierarchical and non-hierarchical	Financial – Participation Share Social – Innovation indicator Community Benefits– Average number of investors per project, Number of comments Experience – Number of Project and average amount invested	Z-transformational	Euclidean distance measure	Ward's minimum-variance	Choice model using logistic regression
Lin, Boh and Goh (2014)	2 stage clustering – hierarchical and non-hierarchical	Social Benefits – Number of projects backed, number of projects created, number of comments Rewards – percentage of reward offered, Average Goal Reputation – average backers, number of varieties	Standardize	Euclidean distance measure	Ward's minimum-variance	Choice model using logistic regression

2.4.2 Prediction

All, the above literatures around the ECF success were based on a statistical model for analysis. The tecnhiques used was reviewed where, Hornuf and Schwienbacher (2018) adopted fixed-effect negative binominal (FENB) estimator as this model has an advantage of removing any unobserved, time-invariant heterogeneity for the campaign data. Aprilia and Wibowo (2018) utilized 2 different model ordinary least regression (OLS) robust standard errors regression and logistic regression where OLS was used first to indentify the influence of 3 dimensions to the success of a project and to the number of investors whom participated in a project. Lastly, a logistic regression was used to compare between the actual funds raised versus the expected goal, this to validate if the 3 dimensions impact the likelihood of success of the project to achieve 100% funding (Aprilia and Wibowo, 2018). Most of all the studies utilised the Regression technique either Probit Regression, Negative Binominal Regression (Vismara, 2018), OLS Regression (Mohammadi and Shafi, 2018; Vulkan et al., 2016), Linear Regression (Xue, J. and Sun, F.F, 2016; Li *et al.*, 2016), Tobit Regression (Piva and Rossi—Lamastra, 2017) or FENB (Block et al., 2018b).

Therefore, from the reward-based crowdfunding platform there were advances machine learning techniques adopted in predicting the success of a campaign . For the best model of the similar prediction of success objective, deep learning technique of Multi-Layer Perceptron (MLP) with an accuracy of 93% was obtained (Yu et al, 2018). The common machine learning techniques used by most studies was Neural Network, Random Forest (RF) (Yu et al, 2018; Kamath and Kamat, 2016) and Naïve Bayes (Kamath and Kamat, 2016)). Optimisation was performed on neural network by Yu et al (2018) with first-order gradient-based optimization of stochastic was performed. Other optimisation techniques such as AdaBoost was also adopted by some studies which provided high accuracy similar to Random Forest (Yu et al, 2018; Liao et al, 2017).

2.5 Research Design

This research is designed to identify the type of investors available in the platform as illustrated in Figure 2 with the highlighted features. Also, for the prediction of success level, the following attributes as highlighted in Figure 3 would be used as the initial starting point.

How to identify the type of investors?

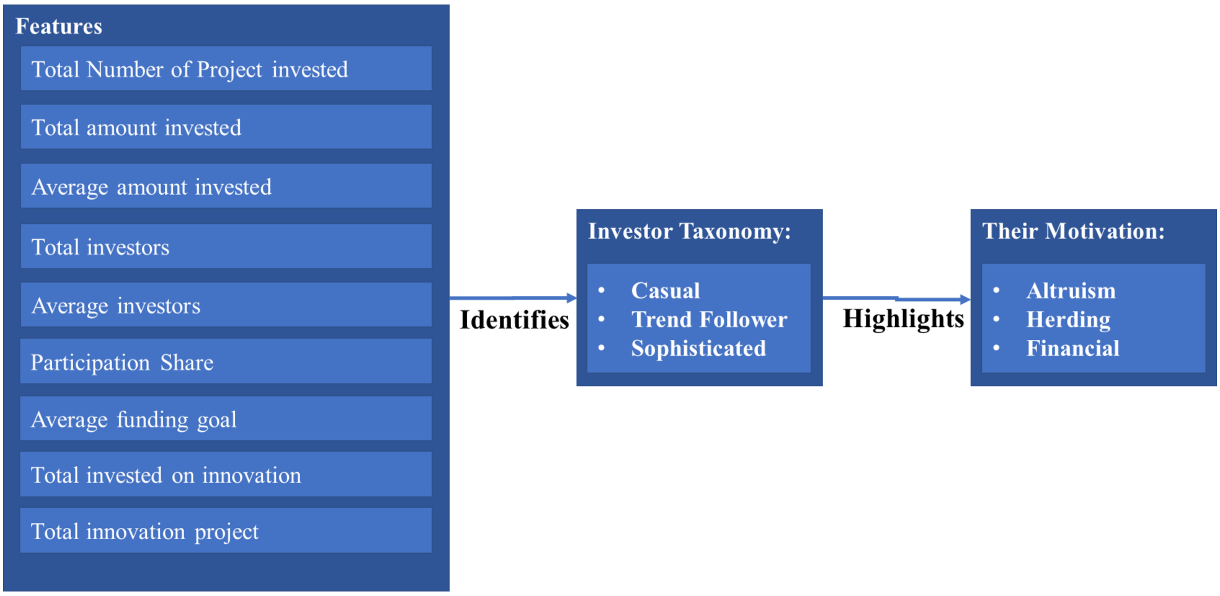


Figure 2: Cluster Analysis Research Framework

All the attributes that could be used to predict the success level of ECF successful campaigns.

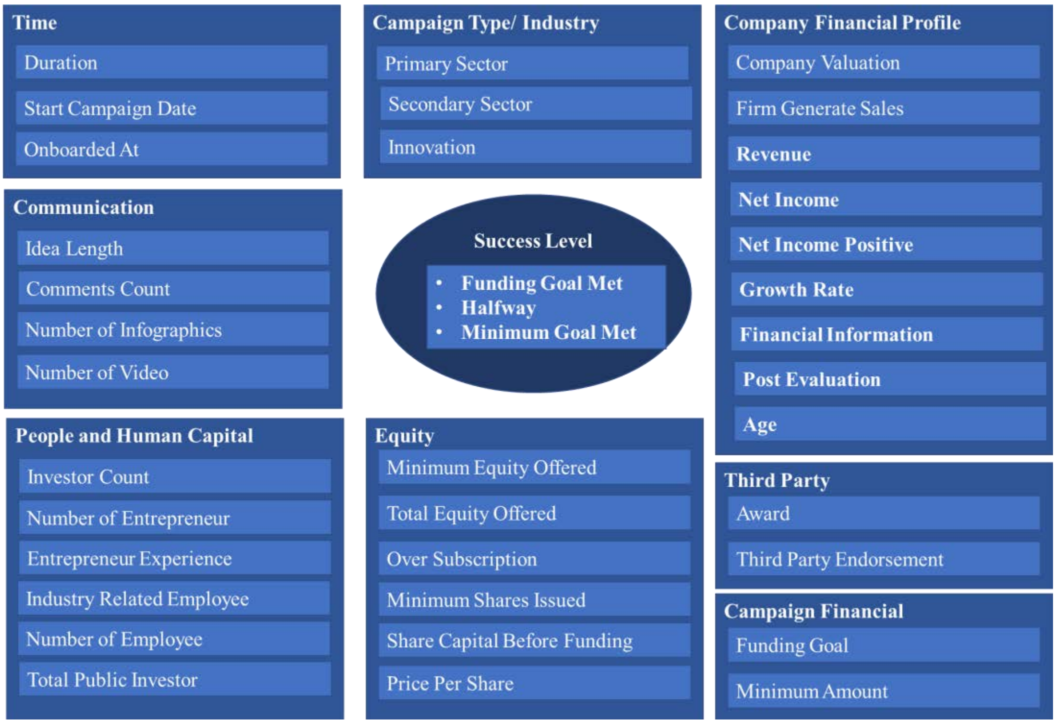


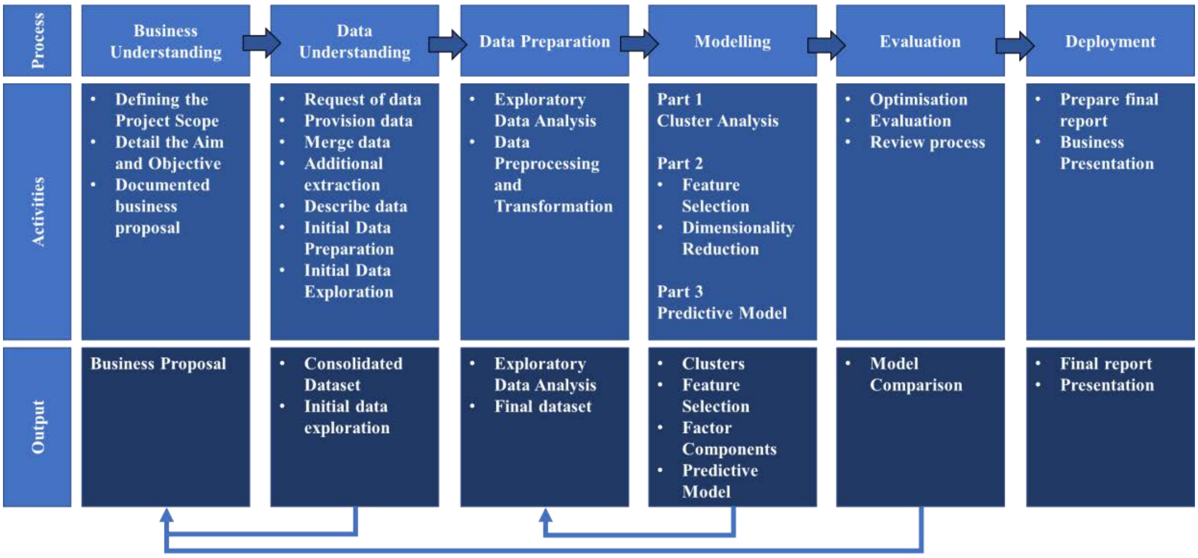
Figure 3: Predictive Model's Attributes for prediction

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter explain the detail data, process, methods, tool and technology that was utilised to undertake this research study. The CRISP-DM model was adopted for this study where end to end process flow and the planned activities as illustrated in Figure 4.



Process	Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Activities	• Defining the Project Scope • Detail the Aim and Objective • Documented business proposal	• Request of data • Provision data • Merge data • Additional extraction • Describe data • Initial Data Preparation • Initial Data Exploration	• Exploratory Data Analysis • Data Preprocessing and Transformation	Part 1 Cluster Analysis Part 2 • Feature Selection • Dimensionality Reduction Part 3 Predictive Model	• Optimisation • Evaluation • Review process	• Prepare final report • Business Presentation
Output	Business Proposal	• Consolidated Dataset • Initial data exploration	• Exploratory Data Analysis • Final dataset	• Clusters • Feature Selection • Factor Components • Predictive Model	• Model Comparison	• Final report • Presentation

Figure 4: End to End Process, activities and expected outcome as per CRISP-DM methodology

The subsequent subtopic would explain each of the phases in detail.

3.2 Data Understanding

Th is a quantitative study utilising a secondary dataset obtained from the Malaysia ECF platform in Malaysia, PitchIn. PitchIn was founded in 2016 where as of August 2019, PitchIn had successfully funded 35 campaigns indirectly being the market leader for ECF platform in Malaysia. As of August 2019, there are 5665 registered members in the platform where only 1645 members have actively invested.

Additionally, PitchIn has an AON model where a minimum goal and funding goal would be determined. Thus, as long the minimum goal is met it would be determined as a successful campaign. The success was broken down to three level 1) minimum goal met 2) Halfway between minimum goal met to almost funding goal met 3) Funding Goal Met. Therefore, if the minimum goal is not achieved, they would refund the collected money.

This platform also identifies the investors based on the investor type which is determined on the amount planned to invest in the platform. The type of investors are broken into Sophisticated investor, angle-like investors as well as retail investors. The amount allowed for investing has no limits for Sophisticated investors, maximum of RM500,000 per annum for angle-like investors and a cap of RM5 000 for companies or RM50,00 per annum for retail investors. This profile is not derived based on the motives of investment neither their investment history rather a governance indicator of the amount allowed to be spent. Additionally, the platform allows potential investors to view the investments made by invested investors if the investor's profile has been set to public (anonymous is set to false).

The objective of this study is to identify the type of investors in the PitchIn platform based on their investment history, identify the influencing factors that distinguishes among success level and subsequently predict the success level based on the influencing factors. Table 7 highlights the initial dataset provided by the platform for this analysis.

Table 7: ECF Dataset and the attributes

Dataset	**Attribute**
Investors	Investor ID, Date of Birth, Anonymous, Investor Type, Created at, Investment count, gender, draft amount, pledged amount, banked in amount, cancelled amount, waiting amount, private placement amount
Investors Transaction	Investor ID, Pitch ID
Campaign	Pitch ID, Idea, Financial Overview, Investment Terms, Funding Goal, Minimum amount spent, Funding Block, Funding duration, Minimum equity offered, Oversubscription equity offered, Total equity offered, Video, Status, Start Campaign,

	Created At, Oversubscription, Oversubscription amount, Company Valuation, Min shares issued, Max shares issued, Share Capital Before Funding, Price per share, draft amount, pledged amount, banked in amount, cancelled amount, waiting amount, private placement amount, Business ID
Business/Firm	Business ID, Status, Sectors, Created At, Social Links, Comments Count, Valid Pitch Count, Investors Count, Gone Live, Coming Soon

The data was provided in excel format where each campaign has a set of four excel sheet that represents the business information, the campaign, the investors as well as the transaction details of the investors for the campaign. Consolidation all four files into a single excel sheet for each of the areas to obtain a master file for business, campaign, investor and transaction would be performed. Additional information on the campaign would be manually scrapped from the website. Following are the additional variables that was captured:

i. Description

- Infographics – Yes or No
- Video Count

ii. Human Capital

- Number of Entrepreneur
- Entrepreneur Experience - Yes or No
- Total number of industry related employee
- Total number of employees

iii. Risk-return

- Firm already generated sales – Yes or No
- Net Income Positive – Yes or No
- Anticipated Growth Rate – in percentage group
- Firm Age in year
- Financial Information if all this information are available – Yes/No

- Projected Revenue
- Expenditure
- Earning before investment
- Audited Account

iv. Innovative Elements

- Innovation Indicator as suggested by Goethner, Luettig and Regner (2018) if existence any one of them – Yes/No
    - Intellectual property patents
    - Start-up on R&D strategy
    - New market with no direct competitor
    - Only service provider for the service or product

v. Others

- Third Party Endorsement – Yes/No
- Awards – Yes/No

Initial Data Preparation

As this dataset is raw and transaction level, an aggregated dataset would be produced for the cluster analysis known as the investor view. As for the feature engineering and predictive model, a campaign view consolidated dataset would be derived. Additionally, new features would be generated as documented in Table 8 to be incorporated in both views.

Table 8: New features for the aggregated Investor view

File	Attribute	Pre-Processing Method
Investors	**For Cluster Analysis**	
	Average amount invested	Mean of the amount invested for individual project
	Average number of investors per project	Mean of number of investors per project
	Amount invested – innovation	Where Project Innovation = Yes, the total amount spent
	Average amount invested – innovation	Where Project Innovation = Yes, the average amount spent

	Participation share	For each RM5 invested = 1 share, thus to see the total number of shares invested
	Total innovation project	
	Innovation Share	Total innovation over total project invested
	Average Goal	Mean of funding goal backed
	Total Variety	Total number of different industries investor had invested
	Early Investor	invested in first t days of a campaign
	Late Investor	invested last days of a campaign
Campaign	**For Predictive Modelling**	
	Percentage Raised	Total amount raised by the campaign divided by the campaign goal. if overfunding, the variable takes a value that is greater than 100 percent
	Public Profile	number of investors profile public over total investors made to the campaign
	% covered in t	Total collection over goal accumulated within first t day of time

Initial Data Exploration

Then, the two files of investor view and campaign view would be explored to identify the data type if it is a numerical or character data, attribute properties if it is categorical, continuous or date time. Also, a statistical summary of the attributes would be performed to identify the distribution, missing value, inconsistent variables or potential outliers.

3.3 Data Preparation

Once, the initial data understanding performed, the exploratory data analysis would be performed based on the aggregated and consolidated dataset. Here, the dependencies between attributes would be explored to identify any useful insights.

From the initial data exploration, the potential data quality issue would be known thus the necessary data pre-processing activities as well as transformation would be performed. For missing value, if it is a categorical attribute, mod imputation would be performed and if it is the continuous dataset the mean imputation would be performed.

On the outlier treatment, cluster analysis would use scale do normalise the data. Subsequently, for the predictive modelling dataset log transformation would be performed for any skewed attribute. If a quantitative dataset is required for modelling, the necessary data type transformation would be performed from categorical to numerical.

Additionally, as the percentage of classes among the three-success level is not balanced, SMOTE technique would be used to perform class balancing. Here, as the number of observations in the dataset is small, thus we would increase the observations while balancing the classes by producing a dataset called SMOTE.

3.4 Modelling

3.4.1 Cluster Analysis

The Hierarchical Clustering as well as K-Means method would be used to identify the optimal number of clusters. The cluster analysis involves 5 stages, following are the details of the stage, the required input and techniques to be used (Hair et.al, 2013; Datanovia, 2018).

Stage 1: Select Objective and the Clustering Variables

The objective for this research is to identify the taxonomy of the investors, the cluster analysis would be performed. The variables that would be used for this analysis as populated in Table 9.

Table 9: List of Variables for Cluster Analysis

Attribute
Total Number of Project invested
Total Amount invested
Average amount invested
Total Investors
Average number of investors
Total innovation project

Average amount invested – innovation
Participation share
Average Funding Goal

Stage 2: Research Design

Firstly, if there is any outliers to perform standardisation of data. As all the variables are metric variable thus the Euclidean Distance Measure would be used to identify the similarity between variables (Lin, Boh and Goh, 2014; Goethner, Luettig and Regner, 2018). Thus, the largest dissimilarity between observation to variable would be identified. This would be noted when performing the hierarchical clustering, if these variables are isolated entirely, then it would be a confirmed outlier and removed from the observation.

Stage 3: Assumption

Ensuring the sample represents the population as we would be utilising the entire investors thus no sub sampling would be performed. To identify if the dataset is cluster-able, this could be performed by visualising with cluster plot or dendrogram.

Stage 4 (Step 1) : Hierarchical Method to determine the optimal number of clusters

The two-stage cluster technique would be used where the hierarchical method would be used to identify the number of clusters for a non-hierarchical method. Several agglomeration techniques are available to measure the dissimilarity between clusters known as complete linkage, average, single and Ward's minimum variance for clustering. The complete and single linkage performs a pairwise dissimilarities comparison between cluster one and two where complete linkage produces more compact cluster by taking the largest dissimilarity distance and single takes the smallest dissimilarity distance thus a looser cluster. Average linkage takesthe average dissimilarity value between two clusters. Ward's distance reduces the within-cluster variance by merging clusters with minimum between-cluster distance (UC, NA). Agglomeration coefficient that is closer to one would be determine as the best method for clustering thus a cluster-able dataset.

Also, other machine learning methods to determine the optimal number of clusters for hierarchical from the R packages such as Elbow, Silhouette and Gap Statistics would be explored to identify the optimal number of cluster (Datanovia, 2018).

Stage 4 (Step 2): k-means method

Then, the k-means method is performed to fine tune the results by profiling and validating the cluster solution. The k-means optimising algorithm would be used to reassign cluster till minimum level of heterogeneity reached. Similarly, the R packages such as Elbow, Silhouette and Gap Statistics would be explored to identify the optimal number of cluster (Datanovia, 2018).

Stage 5 : Interpretation of Clusters

Then, the clusters would be interpreted to identify the avatars of the investors.

3.4.2 Feature Selection and Dimensionality Reduction

To identify the influencing factors that determine the success level of the campaigns, the feature selection technique would be performed. Here, Tree-based Algorithm and Boruta would be explored to identify the important features that influences the success level. These techniques would be explored on the original and SMOTE dataset. Both these techniques are a wrapper technique where the Tree-based algorithm works from the decision tree while Boruta works from the random forest model. These techniques would run to identify the influence of a feature to the target variable thus decision made to add or remove unimportant features, thus the outcome of important features. From these techniques we could compare the selected features.

Next, to minimise the effect of multicollinearity, the factor analysis method would be embedded to identify the principal components of the factors. Then, to determine the number of factor or principal components to be selected, eigenvalue above one would be observed. The selected principal components would be used to model for model comparison. This technique is known as the dimensional reduction technique, where no attributes are removed from the dataset rather its grouped to a similar factor and that principal component is used to predict a model. Here, rather using the huge number of attributes its reduced to minimal number of attributed without losing any information from all the attributes.

3.4.3 Predictive Model

As the target variable is a multiclass scenario as mentioned below:

- 1 – Minimum goal met
- 2 – Halfway
- 3 – Funding Goal met

Thus, the Naïve Bayes, Random Forest and Support Vector Machine (SVM) machine learning techniques would be explored. The variation of testing, parameter tuning, and optimisation would be as in Table 10.

Table 10: List of experiments for this project

Main Area	**Experiment**
Dataset Variation	All features
	Selected Features with Tree Based Algorithm
	Selected Features with Boruta
	Principal Component Features (Factor Analysis)
	SMOTE
Sampling	Random
	Stratified
Parameter Tuning	Naïve Bayes: Laplace
	Random Forest: grid search
	Random Forest: random search
	SVM: Radial
	SVM: Polynomial

3.5 Model Evaluation

As the dataset is a multiclass scenario, we would visualise with the confusion metric to indicate the difference between the actual and predicted value. For model comparison the test accuracy and Area Under Curve (AUC) value would be utilised to describe the best model. AUC is used as this measurement would calculate the balance between the sensitivity and specificity measure (Lantz, B, 2015).

3.6 Tools and Software

Following are the tools, functions and packages that would be used through this project for each of these phases as tabulated in Table 11.

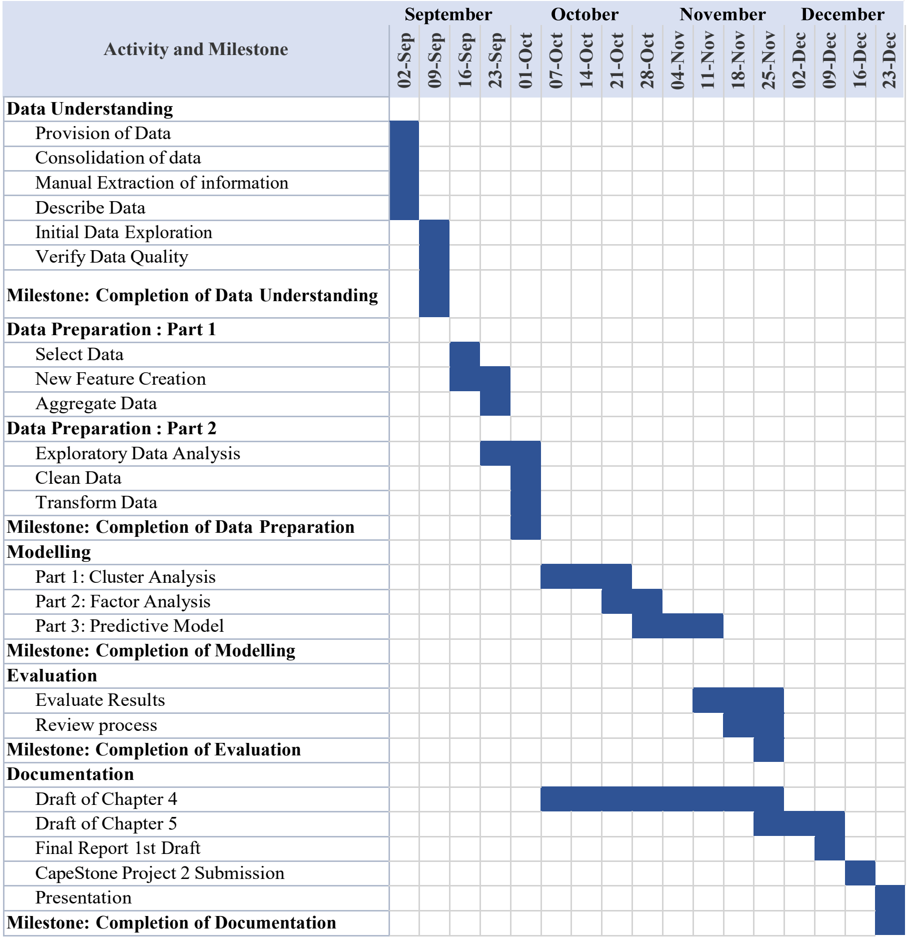
Table 11: Tools, Functions and Packages to be utilised for this project

Phase	**Activity**	**Tool**	**Function/Package**
Data Understanding	Excel Consolidation	Excel	Get Data
	Data Description	R	DataExplorer, Tidyverse, data.table
	Initial Data Exploration	R	Hmisc, psych, ggplot2, Corrplot, Broom, Cowplot, Histogram
Data Preparation	Initial Data Preparation	Tableau Prep	
	New Features	Tableau Prep and R	Dplyr
	Exploratory Data Analysis	Tableau, R	Hmisc, psych, ggplot2,
	Pre-Processing and Transformation	R	Dplyr, mltools
Modelling	Cluster Analysis	R	Factoextra - clustering visualisation NbClust, Clustertend, cluster - clustering Dendextend - colour and compare dendogram
	Feature Selection	R	Mlbench, Caret, Boruta
	Factor Analysis / Principal Component	R	Stats
	Stratified Sampling	R	caTools
	Naïve Bayes	R	Caret, e1071, klaR

	Random Forest	R	RandomForest, MLmetrics, rpart, rpart.plot, party, ROCR
	SVM	R	e1071, MLmetrics
	Class Imbalance	R	DMwR, smotefamily
Evaluation	ROC and AUC	ROC, AUC	pROC

3.7 Research Plan

The research plan is detailed as illustrated in Figure 5 below.



Activity and Milestone	September				October					November				December			
	02-Sep	09-Sep	16-Sep	23-Sep	01-Oct	07-Oct	14-Oct	21-Oct	28-Oct	04-Nov	11-Nov	18-Nov	25-Nov	02-Dec	09-Dec	16-Dec	23-Dec
Data Understanding																	
Provision of Data	■																
Consolidation of data	■																
Manual Extraction of information	■																
Describe Data	■																
Initial Data Exploration		■															
Verify Data Quality		■															
Milestone: Completion of Data Understanding		■															
Data Preparation : Part 1																	
Select Data			■														
New Feature Creation			■	■													
Aggregate Data				■													
Data Preparation : Part 2																	
Exploratory Data Analysis				■	■												
Clean Data					■												
Transform Data					■												
Milestone: Completion of Data Preparation					■												
Modelling																	
Part 1: Cluster Analysis						■	■	■									
Part 2: Factor Analysis								■	■								
Part 3: Predictive Model									■	■	■						
Milestone: Completion of Modelling																	
Evaluation																	
Evaluate Results											■	■	■				
Review process												■	■				
Milestone: Completion of Evaluation													■				
Documentation																	
Draft of Chapter 4						■	■	■	■	■	■	■	■				
Draft of Chapter 5													■	■	■		
Final Report 1st Draft															■		
CapeStone Project 2 Submission																■	
Presentation																	■
Milestone: Completion of Documentation																	■

Figure 5: Research Plan

The project milestone would be used to communicate the research progress and share the outputs at the end of the respective week to the supervisor as mentioned in Table 12.

Table 12: Project Plan: High Level Phases, Milestone and Duration

Phases Milestone	**Expected Output**	**Completion week**
Business Proposal	Business Proposal Completion	Completed
Data Understanding	Consolidated dataset and initial data exploration	09/09
Data Preparation – Part 1	Aggregated dataset, New Features	23/09
Data Preparation – Part 2	EDA and Final transformed dataset	01/10
Modelling – Part 1	Clusters	21/10
Modelling – Part 2	Factors	28/10
Modelling – Part 3	Predictive Modelling	11/11
Evaluation	Evaluation Completion	28/11
Final Project Submission	1st Draft Review Final Report Business Presentation	09/12 20/12 13/01/2020

CHAPTER 4

ANALYSIS AND DISCUSSION

4.1 Introduction

This chapter outlines the procedures involved in the data analytics. The chapter is divided in two sections for Cluster Analysis as well as the Predictive modelling. The steps involved in each of those sections are similar. Firstly, the initial data preparation was performed to merge all the raw file and produce an aggregated (Investor View) and consolidated (Campaign View) dataset for Cluster Analysis and Predictive Modelling respectively. Then, the initial data exploration was performed to identify the data structure, data type, missing value, outlier and multicollinearity. Formerly, the Exploration data analysis would be performed to identify any useful insights from the data.

Upon then, the data pre-processing was performed to treat the data on missing value, outlier and class imbalanced. Then, the modelling would start for cluster analysis while for predictive modelling an extra phase to perform feature selection and dimensional reduction had occur, subsequently the modelling was performed. Here, the output of all the models was reviewed and discuss under Section 4.2.6 for Cluster Analysis and Section 4.3.7 for the feature selection and predictive modelling.

The investor view and campaign view file was prepared from the raw files. Firstly, all the separate files was consolidated. Then, additional information was manually scrapped from the PitchIn website. Appendix 4: Metadata highlights the list of attributes and the description. Tableau Data Prep was used to create the aggregated dataset for the cluster analysis known as the investor view. As for the feature engineering and predictive model, a campaign view dataset was consolidated.

4.2 Cluster Analysis

4.2.1 Initial Data Preparation

Investor view was created to be utilised for the Clustering Analysis. All 3 files was used for the creation of this view. The first step was to clean each of the file and merge them into a single view while creating new features required for the analysis as illustrated in Figure 6 and the detailed out in Table 13.

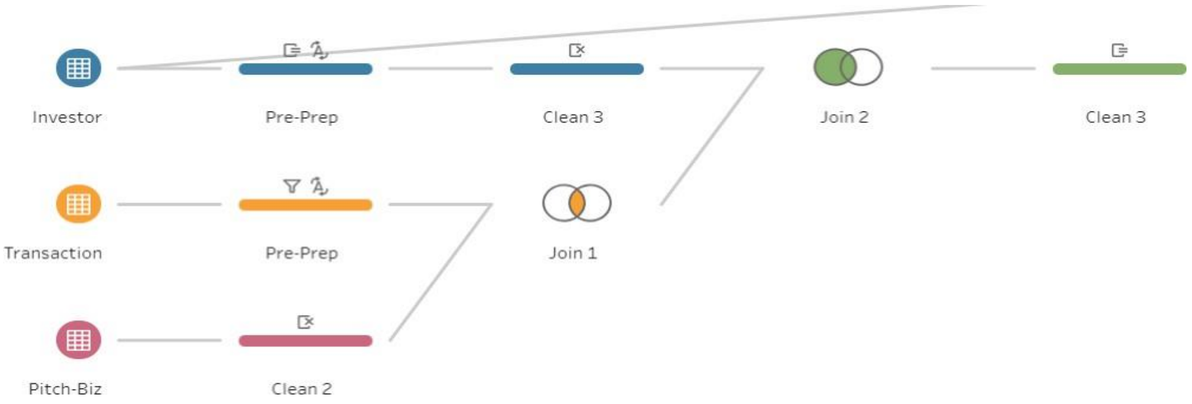


Figure 6: Initial data preparation in Tableau Prep – Step 1

Table 13: Initial Data Preparation – Step 1

File	Attribute	Action	New Attribute	Code
Pitch-Business	Removed 44 fields that is not required for this analysis.			
	Sector	Split with "," to extract the first 2 sectors from the list of sectors involved.	Primary Sector Secondary Sector	
	State	Clean up the state to standardise the naming convention		
	Description	Length of description		LEN(Description)
	Idea	Length of Idea		LEN(Idea)
	Banked In Amount, Funding Goal	Calculate the percentage raised	Percentage Raised	ROUND((([Banked In Amount]/[Funding Goal])*100),0)
	Funding Goal, Min Target	Calculate the halfway value based on the funding goal and minimum target	Halfway Value	(([Funding Goal]-[Min Target])/2)+[Min Target]

	Funding Goal, HalfwayValue, Banked In Amount,	Create the dependent variable Success Level for the predictive analytics	Success Level	IF [Banked In Amount] >= [Funding Goal] THEN 'Funding Goal Met'
				ELSEIF [Banked In Amount] >= [HalfwayValue] AND [Banked In Amount] < [Funding Goal] THEN 'Halfway'
				ELSE 'Minimum Goal Met'

				END
	Start Campaign Date, Last Transaction date	Change to date and time type		
	Start Campaign Date, Last Transaction Date	Calculate the funding duration	Campaign Duration	DATEDIFF('day',[Start Campaign Date],[Last Transaction At],'Monday')
Transaction	Created At	Change to date and time type		
		Filter only Status = Banked in		
Investors	DOB	Calculated the age of the investors	Investor Age	DATEDIFF('year', [DOB],#2019-10-11#,'Monday')
	Created At	Change to date and time type		

Next aggregated features from the aggregated investor view for Cluster Analysis was produced as detailed in Table 14 and illustrated in Figure 7.

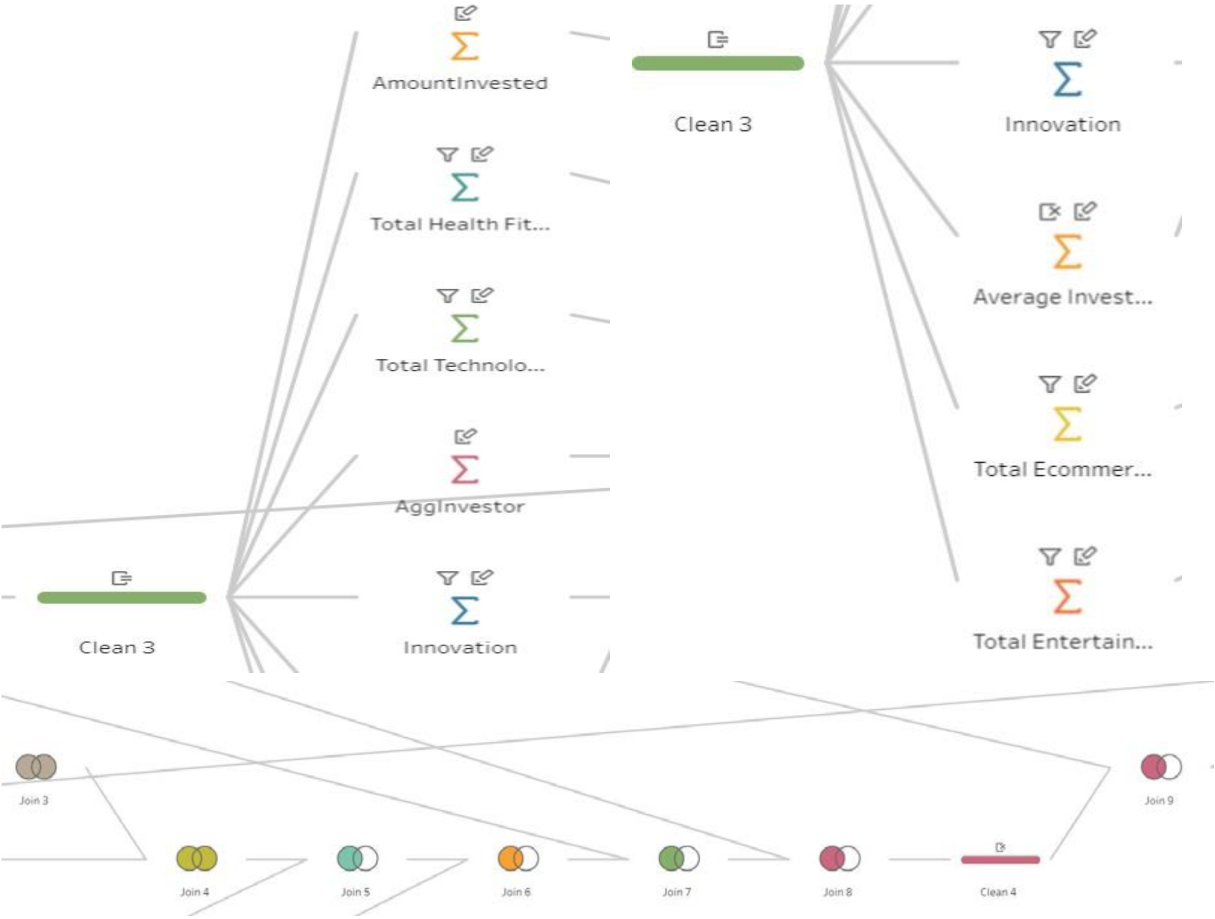


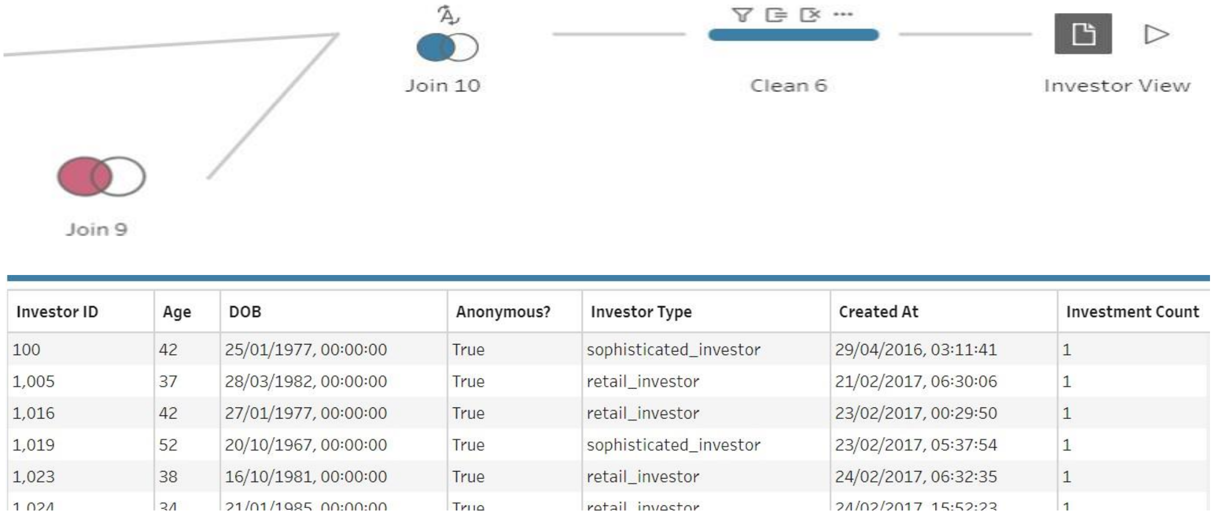
Figure 7: Initial data preparation in Tableau Prep – Step 2

Table 14: Initial Data Preparation – Step 2

Action	**Attribute(s) Used**	**Code**
Created a new attribute Participation Share	Amount Price Per Share	[Amount]/[Price Per Share]
Create Aggregated value for Total Amount Invested	Investor ID Banked In Amount	Group Investor ID, Sum Banked In Amount
Create Aggregated value for Total Project Invested for Health and Fitness	Investor ID Primary Sector Banked In Amount	Group Investor ID, Filter Primary Sector = Health and Fitness Sum Banked In Amount
Create Aggregated value for Total Project Invested for Technology	Investor ID Primary Sector Banked In Amount	Group Investor ID, Filter Primary Sector = Technology Sum Banked In Amount
Create Aggregated value for Total Project, Total Investor, Total Participation Share, Average Amount Invested, Average Funding Goal	Investor ID Project ID Investor Count Participation Share Banked In Amount Funding Goal	Group Investor ID, Count Distinct Project ID – Total Project Sum Investor Count – Total Investors Sum Participation Share – Total Participation Share Average Banked In Amount – Average Amount Invested Average Funding Goal

Create Aggregated value for Total Project Invested for Innovation	Investor ID Innovation Banked In Amount	Group Investor ID, Filter Innovation = Y (Yes) Sum Banked In Amount – Total Invested on Innovation Count Distinct Project ID – Total Innovation Project
Create Aggregated value for Average Investor	Investor ID Investor Count	Group Investor ID, Average Investor Count – Total Investors
Create Aggregated value for Total Project Invested for eCommerce	Investor ID Primary Sector Banked In Amount	Group Investor ID, Filter Primary Sector = eCommerce Sum Banked In Amount
Create Aggregated value for Total Project Invested for Entertainment	Investor ID Primary Sector Banked In Amount	Group Investor ID, Filter Primary Sector = Entertainment Sum Banked In Amount

Then the final Investor view file with 27 fields was produced and an output file was generated as shown in Figure 8.



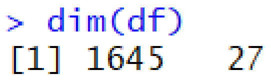
Investor ID	Age	DOB	Anonymous?	Investor Type	Created At	Investment Count
100	42	25/01/1977, 00:00:00	True	sophisticated_investor	29/04/2016, 03:11:41	1
1,005	37	28/03/1982, 00:00:00	True	retail_investor	21/02/2017, 06:30:06	1
1,016	42	27/01/1977, 00:00:00	True	retail_investor	23/02/2017, 00:29:50	1
1,019	52	20/10/1967, 00:00:00	True	sophisticated_investor	23/02/2017, 05:37:54	1
1,023	38	16/10/1981, 00:00:00	True	retail_investor	24/02/2017, 06:32:35	1
1,024	34	21/01/1985, 00:00:00	True	retail_investor	24/02/2017, 15:52:23	1

Figure 8: Investor View Output

4.2.2 Initial Data Exploration

4.2.2.1 Data Description and Data Type

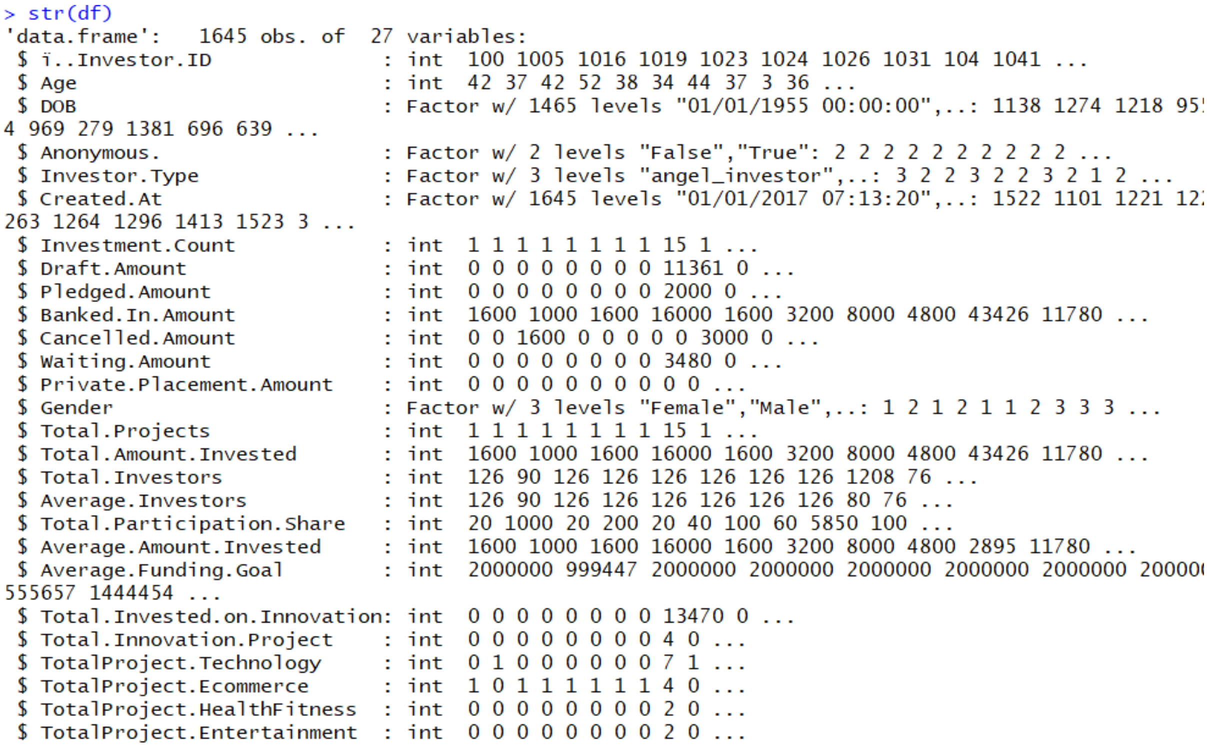
This dataset has a total of 1645 investors whom are active and 27 variables as illustrated in Figure 9.



```
> dim(df)
[1] 1645   27
```

Figure 9: Investor View data dimension

The variables are identified, their data type, length, variable type as well as labels for categorical variable was populated as in Figure 10. The Investor ID is the unique identifier for each investor as this is for cluster analysis, there is no target variable available here.



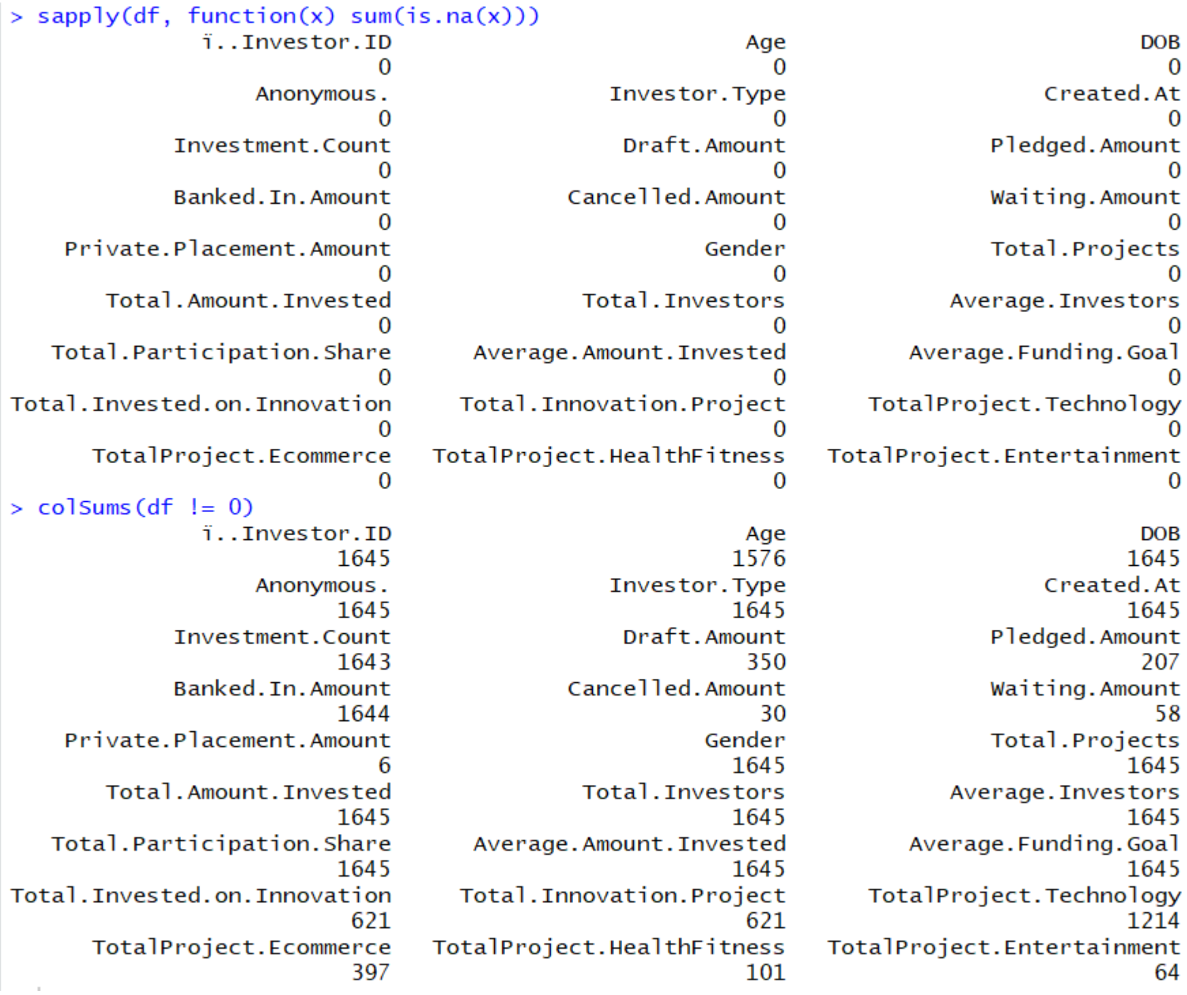
```
> str(df)
'data.frame':   1645 obs. of  27 variables:
 $ ï..Investor.ID             : int  100 1005 1016 1019 1023 1024 1026 1031 104 1041 ...
 $ Age                        : int  42 37 42 52 38 34 44 37 3 36 ...
 $ DOB                        : Factor w/ 1465 levels "01/01/1955 00:00:00",..: 1138 1274 1218 95
4 969 279 1381 696 639 ...
 $ Anonymous.                 : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
 $ Investor.Type              : Factor w/ 3 levels "angel_investor",..: 3 2 2 3 2 2 3 2 1 2 ...
 $ Created.At                 : Factor w/ 1645 levels "01/01/2017 07:13:20",..: 1522 1101 1221 12
263 1264 1296 1413 1523 3 ...
 $ Investment.Count           : int  1 1 1 1 1 1 1 1 15 1 ...
 $ Draft.Amount               : int  0 0 0 0 0 0 0 0 11361 0 ...
 $ Pledged.Amount             : int  0 0 0 0 0 0 0 0 2000 0 ...
 $ Banked.In.Amount           : int  1600 1000 1600 16000 1600 3200 8000 4800 43426 11780 ...
 $ Cancelled.Amount           : int  0 0 1600 0 0 0 0 0 3000 0 ...
 $ Waiting.Amount             : int  0 0 0 0 0 0 0 0 3480 0 ...
 $ Private.Placement.Amount   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Gender                     : Factor w/ 3 levels "Female","Male",..: 1 2 1 2 1 1 2 3 3 3 ...
 $ Total.Projects             : int  1 1 1 1 1 1 1 1 15 1 ...
 $ Total.Amount.Invested      : int  1600 1000 1600 16000 1600 3200 8000 4800 43426 11780 ...
 $ Total.Investors            : int  126 90 126 126 126 126 126 126 1208 76 ...
 $ Average.Investors          : int  126 90 126 126 126 126 126 126 80 76 ...
 $ Total.Participation.Share  : int  20 1000 20 200 20 40 100 60 5850 100 ...
 $ Average.Amount.Invested    : int  1600 1000 1600 16000 1600 3200 8000 4800 2895 11780 ...
 $ Average.Funding.Goal       : int  2000000 999447 2000000 2000000 2000000 2000000 2000000 20000
555657 1444454 ...
 $ Total.Invested.on.Innovation: int  0 0 0 0 0 0 0 0 13470 0 ...
 $ Total.Innovation.Project   : int  0 0 0 0 0 0 0 0 4 0 ...
 $ TotalProject.Technology    : int  0 1 0 0 0 0 0 0 7 1 ...
 $ TotalProject.Ecommerce     : int  1 0 1 1 1 1 1 1 4 0 ...
 $ TotalProject.HealthFitness : int  0 0 0 0 0 0 0 0 2 0 ...
 $ TotalProject.Entertainment : int  0 0 0 0 0 0 0 0 2 0 ...
```

Figure 10: Investor View data structure

All the variables are numeric except there is 3 categorical variables on Anonymous, Investor Type and Gender. There are additional 2 variables that are datetime however being reflected as factor, thus requires a data type clean up.

4.2.2.2 Missing Value Identification

Next to explore if there is any missing value, two approach to first check if there is any NA and subsequently check the total number of 0 in a dataset was performed. The 2nd approach was taken as we expect certain numerical field to not have zero value. Figure 11 indicates the outcome.



```
> sapply(df, function(x) sum(is.na(x)))
            ï..Investor.ID                        Age                        DOB
                         0                          0                          0
                Anonymous.              Investor.Type                 Created.At
                         0                          0                          0
          Investment.Count               Draft.Amount             Pledged.Amount
                         0                          0                          0
          Banked.In.Amount           Cancelled.Amount             Waiting.Amount
                         0                          0                          0
  Private.Placement.Amount                     Gender             Total.Projects
                         0                          0                          0
     Total.Amount.Invested            Total.Investors          Average.Investors
                         0                          0                          0
   Total.Participation.Share    Average.Amount.Invested       Average.Funding.Goal
                         0                          0                          0
Total.Invested.on.Innovation   Total.Innovation.Project   TotalProject.Technology
                         0                          0                          0
      TotalProject.Ecommerce   TotalProject.HealthFitness TotalProject.Entertainment
                         0                          0                          0
> colSums(df != 0)
            ï..Investor.ID                        Age                        DOB
                      1645                       1576                       1645
                Anonymous.              Investor.Type                 Created.At
                      1645                       1645                       1645
          Investment.Count               Draft.Amount             Pledged.Amount
                      1643                        350                        207
          Banked.In.Amount           Cancelled.Amount             Waiting.Amount
                      1644                         30                         58
  Private.Placement.Amount                     Gender             Total.Projects
                         6                       1645                       1645
     Total.Amount.Invested            Total.Investors          Average.Investors
                      1645                       1645                       1645
   Total.Participation.Share    Average.Amount.Invested       Average.Funding.Goal
                      1645                       1645                       1645
Total.Invested.on.Innovation   Total.Innovation.Project   TotalProject.Technology
                       621                        621                       1214
      TotalProject.Ecommerce   TotalProject.HealthFitness TotalProject.Entertainment
                       397                        101                         64
```

Figure 11: Investor View - Missing Value

The output confirmed no NA variable available in the dataset, however noticed there is many variables not having full 1645 non zero value. This is acceptable for the amount, total investor and total project related variables. However, it is not acceptable for Age to have zero value, thus missing value treatment is required for this attribute.

4.2.3 Exploratory Data Analysis

1. Anonymous

Anonymous flag is an indicator on the Profile, if the individual investor has made themselves known as Public Investors or their Profile are unknown to public. This would be visible during investment, where if an investor are a public investor and influential, this may attract more investors to invest on the profile. Herding Behaviour may be presence here as one would invest at campaigns that has the influential individual presence. Figure 12 highlights the distribution of investor based on private and public profile.

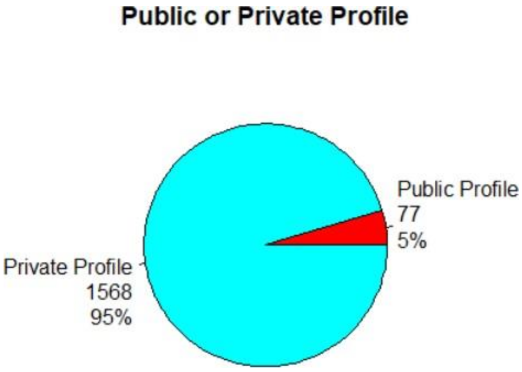


Figure 12: Investor - Public or Private Profile

From the figure above, noted that 95% of the profile are private profile thus investors and their investment are unknown. On the other hand, the Public Profile's investment would be known and if they are deemed influential investors, their investment action would be mirrored thus a herding behaviour.

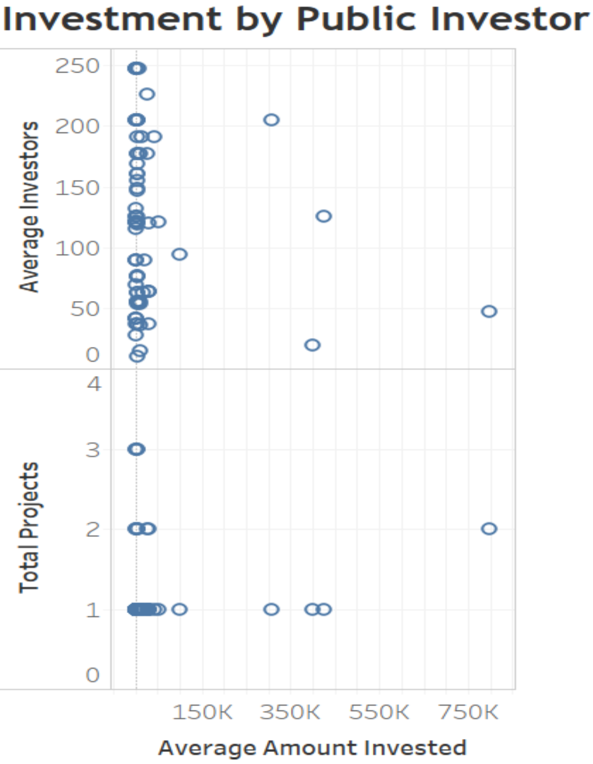


Figure 13: Investor - Public Investor Investment Pattern

Further exploration of this Public profile to identify their spending pattern, based on figure above only 5 public investors have large average amount spent beyond RM150k per campaign and only 1 of them have invested in 2 campaigns. The rest 4 of them are not a frequent investor rather a onetime investor whom have pumped in large amount for a project. Thus, we can conclude them being an influential public profile individual. Additionally, the average investors are high only for 3 campaigns with investors more than 50 investors. The rest of the public investors have only invested in 1 project with an average below RM5k of investment. Only 3 public investors have invested in more than 1 project. Thus, here we could conclude the herding behaviour may not be prominent as most public investor has only invested on 1 project.

2. Investor Type

The investor type here is captured during an investor registering to the platform. This indicated the spending capacity of an individual. There is three investor type sophisticated, angel and retail investor where sophisticated having the most investment capability which is not capped followed by angel investor with a cap of RM500k and retail investor has a limit of RM50k per year.

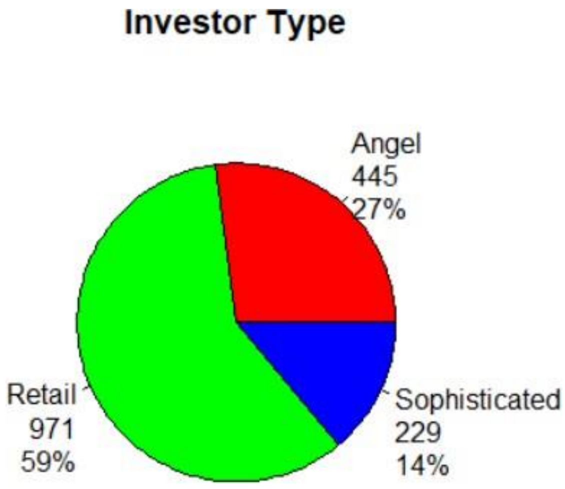


Figure 14: Investor Type

Based on the distribution pie above, 59% of the investors are retail investors followed by 17% of Angel and 14% of sophisticated investors.

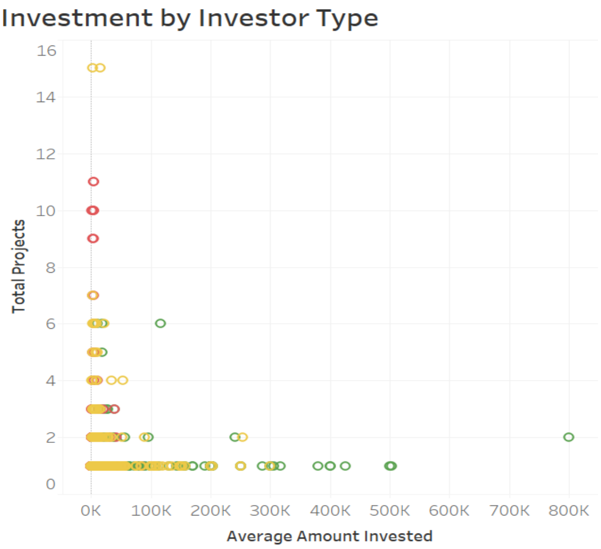


Figure 15: Investment by Investor Type

As illustrated above, majority of the retail investors have invested the most on 3 projects expect a few of them whom have invested between 9 to 12 projects. Most angel investors have invested in one project and several invested between two to 7 projects. On sophisticated investors, at most they have invested in 6 projects however with high average amount.

3. Gender

Investor's gender was explored as illustrated in Figure 16.

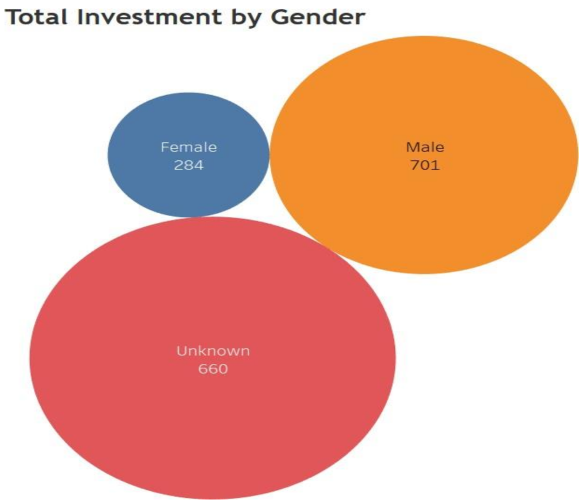


Figure 16: Total Investors and Investment by Gender

43% of investors are male and 17% are female, however 40% of investor's gender are unknown. Observing the total amount invested, the unknown gender has the most amount invested followed by male and female. This variable would not be referenced due to the large unknown gender, by performing a mode imputation the dataset would be biased towards male. Thus, would only be referenced for profiling.

4. Created Year

The onboarding of members was observed, Figure 17. Prior to doing so, the attribute would be converted to datetime type.

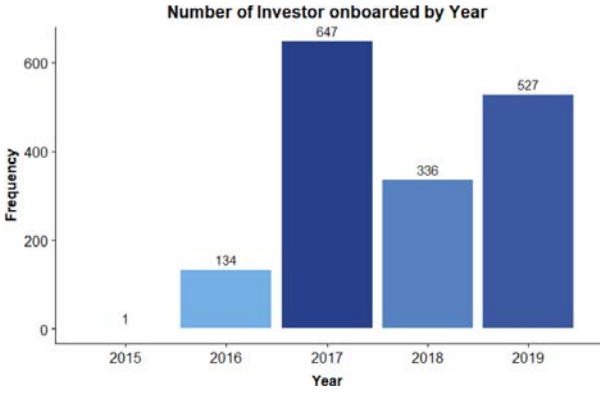
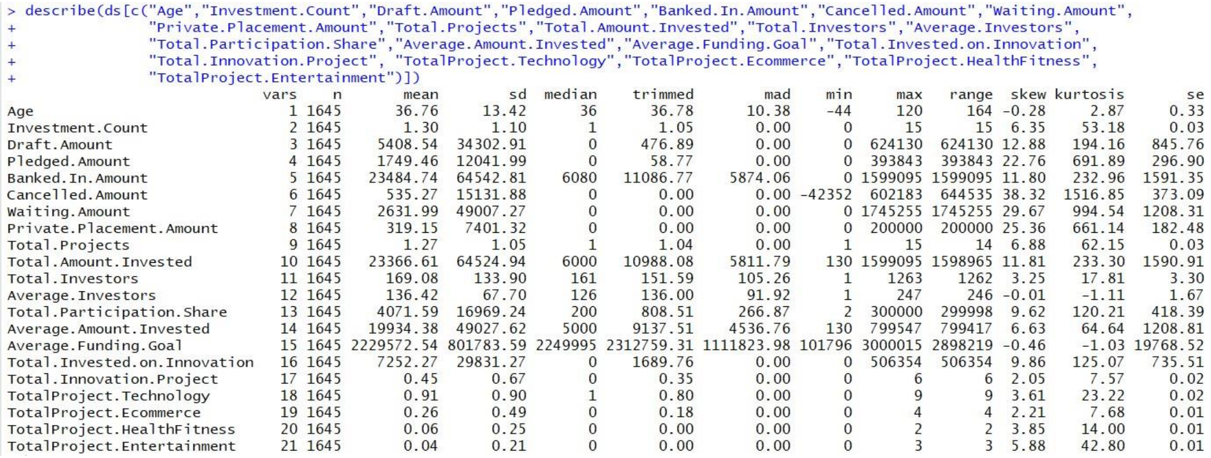


Figure 17: Investor Growth by Year

Majority of investors was onboarded in 2017 of 647 of investors. Observed a drop if investors onboarding in 2018 and subsequently picked up in 2019 where additional 527 investors onboarded in the platform then. The 2018 drop could be triggered due to the slowdown in Malaysia economy due to the political instability during the first half of the year. Subsequently, we would be able to observe the total project onboarded and total investment drop during the same period.

Numerical Variables.

A statistic summary of all numerical variables was obtained as illustrated in Figure 18.



```
> describe(ds[c("Age","Investment.Count","Draft.Amount","Pledged.Amount","Banked.In.Amount","Cancelled.Amount","Waiting.Amount",
+              "Private.Placement.Amount","Total.Projects","Total.Amount.Invested","Total.Investors","Average.Investors",
+              "Total.Participation.Share","Average.Amount.Invested","Average.Funding.Goal","Total.Invested.on.Innovation",
+              "Total.Innovation.Project", "TotalProject.Technology","TotalProject.Ecommerce","TotalProject.HealthFitness",
+              "TotalProject.Entertainment")])
                             vars    n       mean        sd  median    trimmed        mad    min     max   range  skew kurtosis       se
Age                             1 1645      36.76     13.42      36      36.78      10.38    -44     120     164 -0.28     2.87     0.33
Investment.Count                2 1645       1.30      1.10       1       1.05       0.00      0      15      15  6.35    53.18     0.03
Draft.Amount                    3 1645    5408.54  34302.91       0     476.89       0.00      0  624130  624130 12.88   194.16   845.76
Pledged.Amount                  4 1645    1749.46  12041.99       0      58.77       0.00      0  393843  393843 22.76   691.89   296.90
Banked.In.Amount                5 1645   23484.74  64542.81    6080   11086.77    5874.06      0 1599095 1599095 11.80   232.96  1591.35
Cancelled.Amount                6 1645     535.27  15131.88       0       0.00       0.00 -42352  602183  644535 38.32  1516.85   373.09
Waiting.Amount                  7 1645    2631.99  49007.27       0       0.00       0.00      0 1745255 1745255 29.67   994.54  1208.31
Private.Placement.Amount        8 1645     319.15   7401.32       0       0.00       0.00      0  200000  200000 25.36   661.14   182.48
Total.Projects                  9 1645       1.27      1.05       1       1.04       0.00      1      15      14  6.88    62.15     0.03
Total.Amount.Invested          10 1645   23366.61  64524.94    6000   10988.08    5811.79    130 1599095 1598965 11.81   233.30  1590.91
Total.Investors                11 1645     169.08    133.90     161     151.59     105.26      1    1263    1262  3.25    17.81     3.30
Average.Investors              12 1645     136.42     67.70     126     136.00      91.92      1     247     246 -0.01    -1.11     1.67
Total.Participation.Share      13 1645    4071.59  16969.24     200     808.51     266.87      2  300000  299998  9.62   120.21   418.39
Average.Amount.Invested        14 1645   19934.38  49027.62    5000    9137.51    4536.76    130  799547  799417  6.63    64.64  1208.81
Average.Funding.Goal           15 1645 2229572.54 801783.59 2249995 2312759.31 1111823.98 101796 3000015 2898219 -0.46    -1.03 19768.52
Total.Invested.on.Innovation   16 1645    7252.27  29831.27       0    1689.76       0.00      0  506354  506354  9.86   125.07   735.51
Total.Innovation.Project       17 1645       0.45      0.67       0       0.35       0.00      0       6       6  2.05     7.57     0.02
TotalProject.Technology        18 1645       0.91      0.90       1       0.80       0.00      0       9       9  3.61    23.22     0.02
TotalProject.Ecommerce         19 1645       0.26      0.49       0       0.18       0.00      0       4       4  2.21     7.68     0.01
TotalProject.HealthFitness     20 1645       0.06      0.25       0       0.00       0.00      0       2       2  3.85    14.00     0.01
TotalProject.Entertainment     21 1645       0.04      0.21       0       0.00       0.00      0       3       3  5.88    42.80     0.01
```

Figure 18: Investor numerical variables statistical summary

Based on the figure above, observed all the variables are skewed where skewness value is beyond positive 2 and below negative 2. The kurtosis beyond 3 further confirms the presence of outlier in the dataset. Age, Average Investor and Average Funding Goal are the only two variables does not have outlier within the dataset. This confirms that scaling is required for this dataset to be able to perform the cluster analysis.

5. Age

The Age attribute was derived from the DOB attribute. Prior to doing so, the DOB attribute would be converted to datetime type.

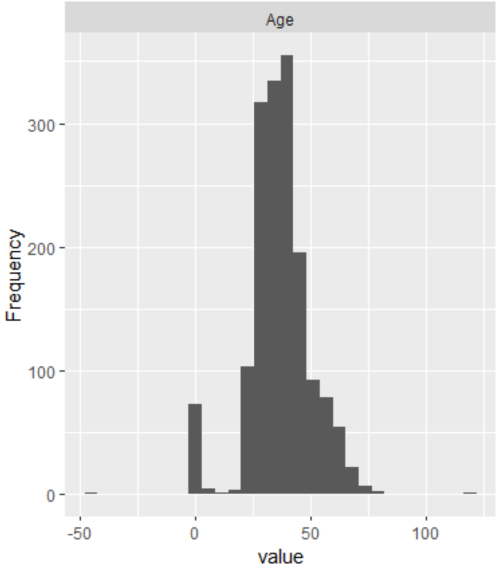


Figure 19: Investor Age Distribution

Based on the Histogram above, noted that around 75 investors are at age -44, 0 and 120 which are incorrect data, thus this value would be converted to missing value and the missing value would be mean imputed. Majority of the investors are at the age of 37.

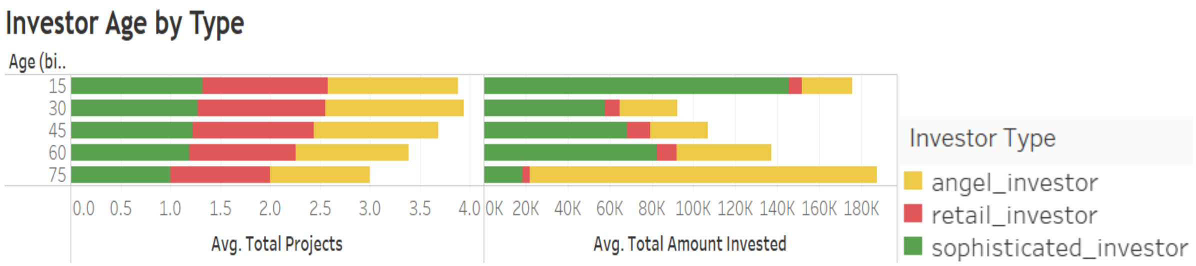


Figure 20: Investor Age by type and average amount invested

Here, we noted that highest average number of projects are invested by the Age group of 30 to 45 years with average of 4 investments. Next followed by the centennials between 3 to 4 projects and the least number of projects invested by age group beyond 75 years of age. However, observing the total amount invested by those age group, age 75 and above had invested the most beyond RM180k, thus explains that minimal project but with maximum amount is spent by that age group. For the 30 to 45 years group, noted that they invested the least at around RM90k. However, surprisingly the younger generation 15 to 35 years, has invested the second highest. As financial capacity may not be possible at this age, further zoom on the dataset identified an investor whom have pumped in close to RM1.5 mil on two different project and is a sophisticated investor. Thus, the influence on that age group. Here, we could conclude that the older generation has the highest amount to invest however does not randomly invest on many projects and on the other hand the 30 to 45 years of investors are eager to explore however limits their investment amount.

6. Campaigns

Total Campaigns, Innovation Campaigns, Sector related campaigns such as Technology, Ecommerce, Health and Fitness and Entertainment were explored. The histogram for these variables was explored as in Figure 21.

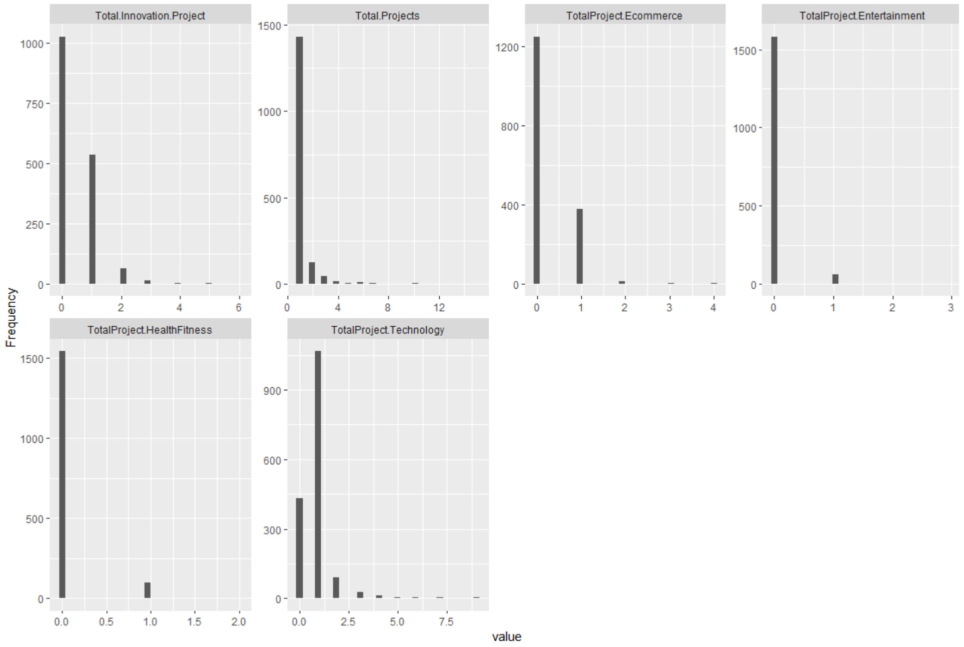


Figure 21: Campaign Distributions

Majority investors have invested in total 1 project however there is few whom had invested close to 10 projects, thus the positive skewness towards right. Zooming further into the number of innovation project invested, majority investors are not keen in innovation project and around 500 plus investors have invested in at least 1 innovation project. The maximum that one have invested on an innovation project is 5 projects. Looking at the sector of investment, majority investment are made in the Technology industry with the highest investors had invested at least once followed by the ecommerce sector. Further exploration to observe the diversification of campaigns invested based on the current investor type was performed, Figure 22.

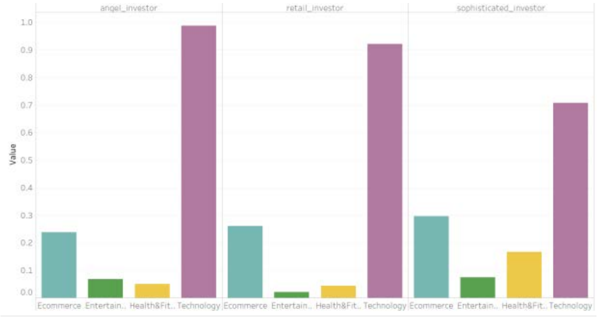


Figure 22: Total Campaign invested by campaign category and investor type

Here, noted that all investor type has invested the most on Technology driven campaigns followed by ecommerce related. Sophisticated Investors has an additional investment made on

Health and Fitness related campaigns indication a slight diversification on investment performed.

7. Draft Amount, Pledged Amount and Total Amount Invested

The amount of investment was observed, where the draft and pledge amount are the pre committed amount by an investor and subsequently provides the real amount which is the amount invested. Here, an observation on the investor behaviour where more amount was pledged or invested but eventually lesser amount was transferred as illustrated in Figure 23.

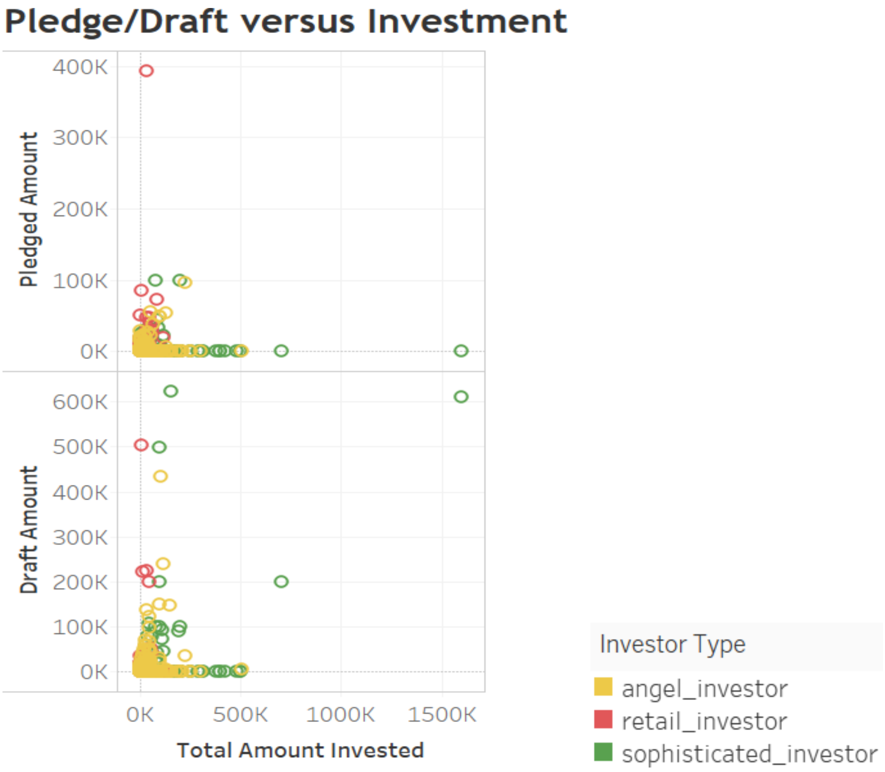


Figure 23: Amount Invested Distribution

Here, the observing the horizontal line along the RM0k is an indication of certain individual whom are not keen to pledge however they would still end up investing. Here we noted that a consistent distribution of individuals along that axis where majority of them are sophisticated investors with very few angel investors there. Then the vertical line along the RM0k noticed very few but there is existence of individual whom does not invest, however they do pledge/draft. Most of the individuals are the retail investors along that line. Rest of the majority investors are closer to the central point, indicating a neutral behaviour where amount pledge are the amount invested. Very few outliers was observed having pledge/draft less amount but ended investing bigger amount and vice versa.

Additionally, the participation share was observed. This attribute is an indicator of financial reward one would obtain based on the amount spend. The more shares they receive the more financially motivated they are, Figure 24.

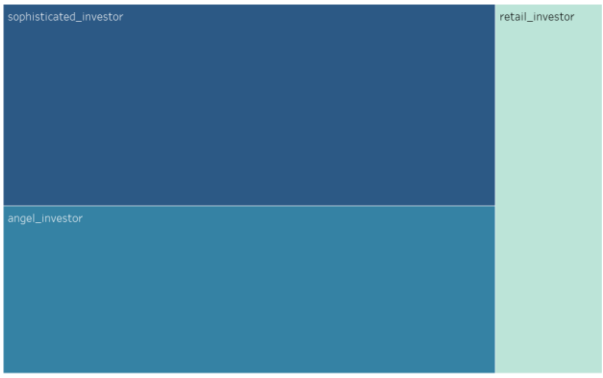


Figure 24: Participation Share Distribution by investor type

Here, as expected Sophisticated and Angel investors would have the highest number of participations share due to the huge amount spent by them. However, this does not express the financial motivation by them, thus further exploration on the total amount spent by the participation share was performed, as in Figure 25.

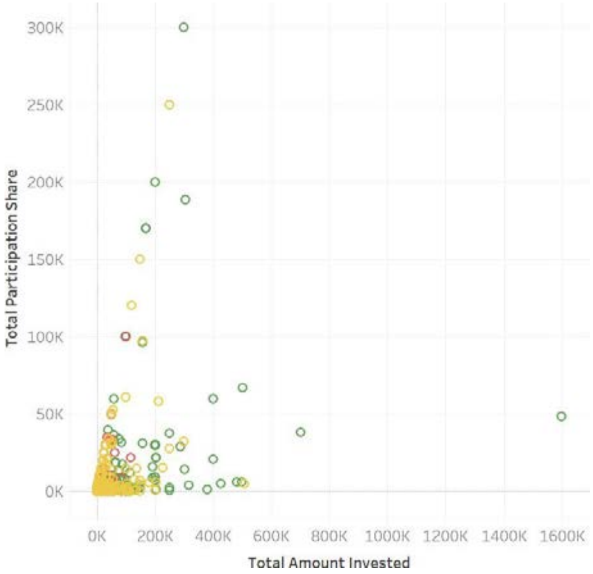


Figure 25:Relationship between Participation Share and Total amount invested

As above, noted that the more amount invested the less participation share is an indication of less share rewarded due to high share price. Thus, here the financial motivation is minimum rather the Entrepreneurship experience on the prospect of a project is looked upon, here sophisticated investors are the one whom fall within this category. On the other trend, noted that as the participation share increases in parallel to the amount invested, here we see a mix group of all investor type and this is the financial motivated individuals where the more they spend, the more shares the receive. Majority others falls closer to the central point where minimum amount spend and minimum return received.

8. Total Investors and Average Investors

The distribution of investors was observed as in Figure 26. Here, we would like to observe what are the average number of investors available when an investor is investing.

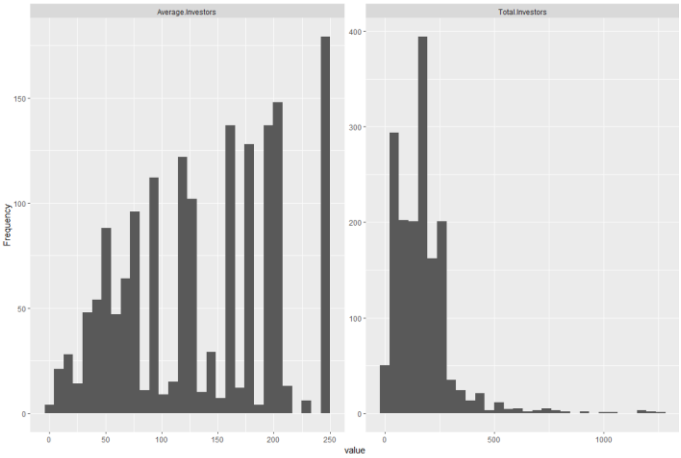


Figure 26: Investors Distribution

From distribution above, noted that total investors skewed towards right where there are campaigns with maximum 1500 investors and majority campaigns has around 240 investors. The average number of number of investors is also 250 investors during most investors invest. Next, to observe the relationship between average amount invested to the average number of investors available. Here if there is more money invested assume more individuals would invest as it's an indication of a successful project, thus the existence of herding behaviour.

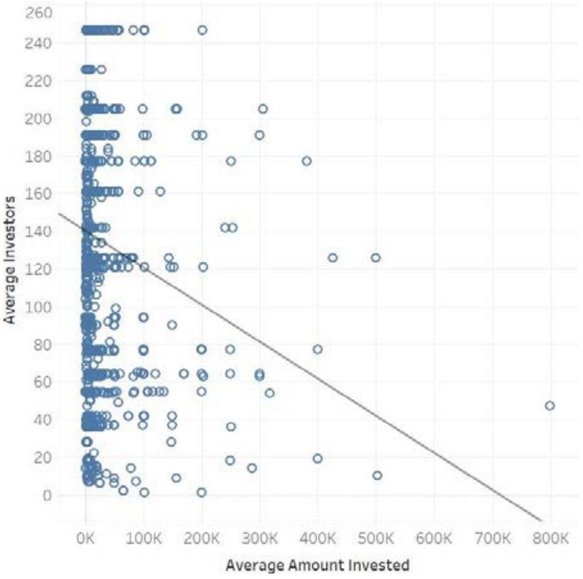


Figure 27: Relationship between Average Amount Invested and Average Investors

As per figure above, noted that a negative correlation between amount invested to the average investors was observed. Here, noted that the more money invested, does not attract more investors. Thus, potentially minimal herding behaviour amongst investors. Moreover, this further substantiated due to the anonymous identity of investors, thus an indication of financially literate community whom does not judge a campaign based on amount invested.

9. Average Funding Goal

Funding Goal is basically the target value of funding that is set by the campaign owner. Here, Figure 28 highlights the distribution of investors across the funding goal.

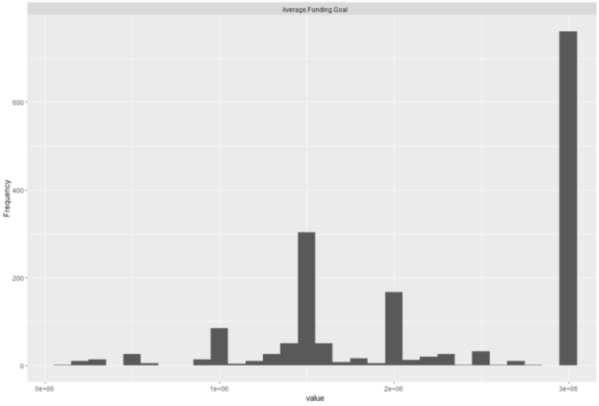


Figure 28: Average Funding Goal Distribution

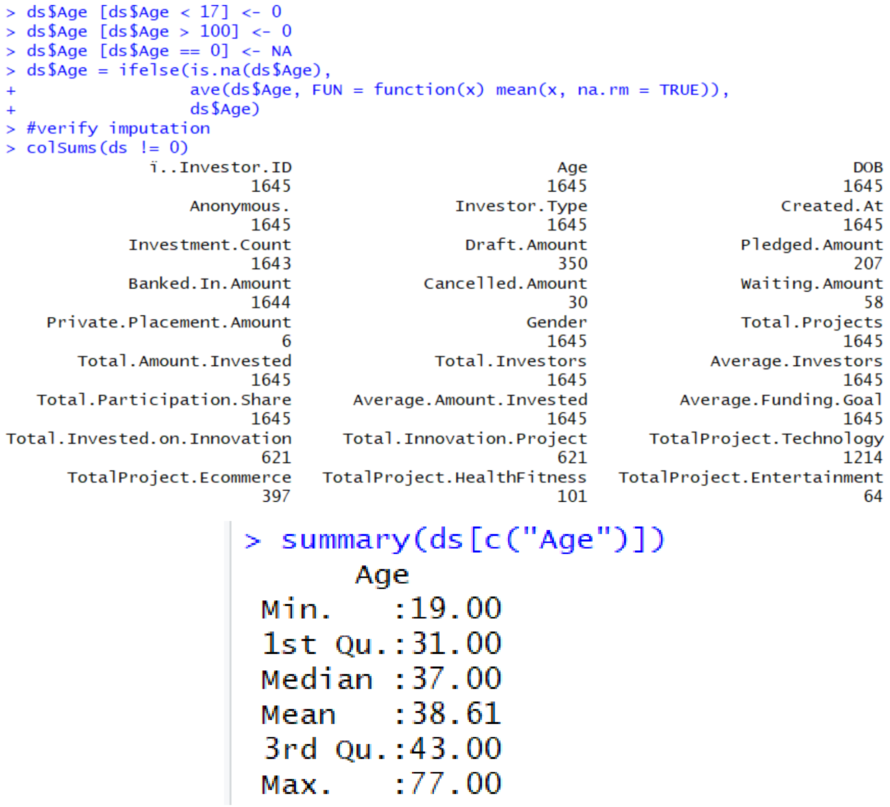
Noted, majority investors have invested on campaigns with high funding goals that is close to RM3 mil followed by average investors investing on campaigns with funding goal of RM2.2 mil. Lower funding goals has attracted lower investors, this probably relates to the maximum oversubscription that is capped for a campaign.

4.2.4 Data Pre-processing and Transformation

During data exploration, certain pre-processing activities was performed especially transforming the data type of DOB and Created At from factor to datetime. Now the other data quality issues would be treated accordingly.

4.2.4.1 Missing Value Treatment

The Age attribute was the only variable identified to have irrelevant value zero, age below 17 and above 100. Thus, this value would be converted to missing value, thus a mean imputation would be performed to fix these variables, the output as illustrated in Figure 29.



```
> ds$Age [ds$Age < 17] <- 0
> ds$Age [ds$Age > 100] <- 0
> ds$Age [ds$Age == 0] <- NA
> ds$Age = ifelse(is.na(ds$Age),
+                 ave(ds$Age, FUN = function(x) mean(x, na.rm = TRUE)),
+                 ds$Age)
> #verify imputation
> colSums(ds != 0)
               ï..Investor.ID                          Age                          DOB
                         1645                         1645                         1645
                   Anonymous.                Investor.Type                   Created.At
                         1645                         1645                         1645
             Investment.Count                 Draft.Amount               Pledged.Amount
                         1643                          350                          207
             Banked.In.Amount             Cancelled.Amount               Waiting.Amount
                         1644                           30                           58
     Private.Placement.Amount                       Gender               Total.Projects
                            6                         1645                         1645
        Total.Amount.Invested              Total.Investors            Average.Investors
                         1645                         1645                         1645
      Total.Participation.Share      Average.Amount.Invested          Average.Funding.Goal
                         1645                         1645                         1645
   Total.Invested.on.Innovation     Total.Innovation.Project       TotalProject.Technology
                          621                          621                         1214
        TotalProject.Ecommerce     TotalProject.HealthFitness   TotalProject.Entertainment
                          397                          101                           64
```

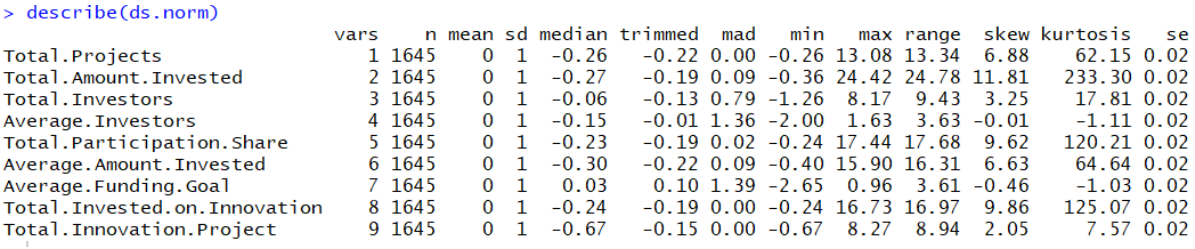
```
> summary(ds[c("Age")])
      Age
 Min.   :19.00
 1st Qu.:31.00
 Median :37.00
 Mean   :38.61
 3rd Qu.:43.00
 Max.   :77.00
```

Figure 29: Investor View - Missing value treatment

Here, noted that now age has 1645 observation with no age zero value. Additionally, the mean age is at 38 after the transformation where the minimum age is 19 and the maximum age of investor is at 77.

4.2.4.2 Outlier Treatment

As we noted most variables have outliers as the kurtosis value is beyond 3, here for cluster analysis scaling the numerical variable would be performed. Figure 30 highlights the scale transformation performed and the statistical summary of the data upon transformation.



```
> describe(ds.norm)
                            vars    n mean sd median trimmed  mad   min   max range  skew kurtosis   se
Total.Projects                 1 1645    0  1  -0.26   -0.22 0.00 -0.26 13.08 13.34  6.88    62.15 0.02
Total.Amount.Invested          2 1645    0  1  -0.27   -0.19 0.09 -0.36 24.42 24.78 11.81   233.30 0.02
Total.Investors                3 1645    0  1  -0.06   -0.13 0.79 -1.26  8.17  9.43  3.25    17.81 0.02
Average.Investors              4 1645    0  1  -0.15   -0.01 1.36 -2.00  1.63  3.63 -0.01    -1.11 0.02
Total.Participation.Share      5 1645    0  1  -0.23   -0.19 0.02 -0.24 17.44 17.68  9.62   120.21 0.02
Average.Amount.Invested        6 1645    0  1  -0.30   -0.22 0.09 -0.40 15.90 16.31  6.63    64.64 0.02
Average.Funding.Goal           7 1645    0  1   0.03    0.10 1.39 -2.65  0.96  3.61 -0.46    -1.03 0.02
Total.Invested.on.Innovation   8 1645    0  1  -0.24   -0.19 0.00 -0.24 16.73 16.97  9.86   125.07 0.02
Total.Innovation.Project       9 1645    0  1  -0.67   -0.15 0.00 -0.67  8.27  8.94  2.05     7.57 0.02
```

Figure 30: Investor View Outlier Treatment

Only 9 variables are selected for this Cluster Analysis as documented in Section 2.5 by other literatures thus the scaling was performed for those variables only.

4.2.5 Construction of Model and Interpretation

Next, the Cluster Analysis would be performed where Agglomerative Hierarchical Clustering would be performed. The HCLUST and Cluster package in R was used. First, a dissimilarity matrix would be performed to obtain the distance between the observations using the Euclidean distance and the complete method of dissimilarity would be selected.

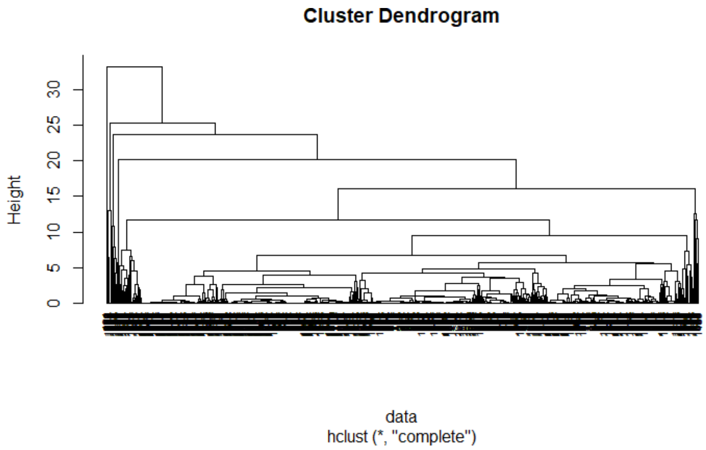
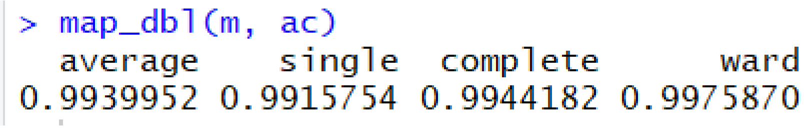


Figure 31: Agglomerative Hierarchical Clustering with HCLUST – Complete

Here, noted a very compact clusters was obtained. In ensuring a good spread of clusters is obtained, an equal distance between clusters should be seen. Despite seeing a good pattern at height 8, however there is still 2 group of outliers that was visible in both end that could not be merged. The final equal distribution of cluster was only at height 27 where 2 clusters was obtained at that point. Next, the Cluster package was used to calculate the agglomerative coefficient to observe between the different agglomerative method to identify which is the best method. The value closer to 1 indicates a strong clustering structure.



```
> map_dbl(m, ac)
  average    single  complete      ward 
0.9939952 0.9915754 0.9944182 0.9975870 
```

Figure 32: Agglomerative Hierarchical Clustering – Agglomerative Coefficient

Based on the coefficient value, the ward distance is the closest to 1 indicating the best method to measure the dissimilarity. So, a dendrogram would be produced to observe the clustering using the HCLUST package.

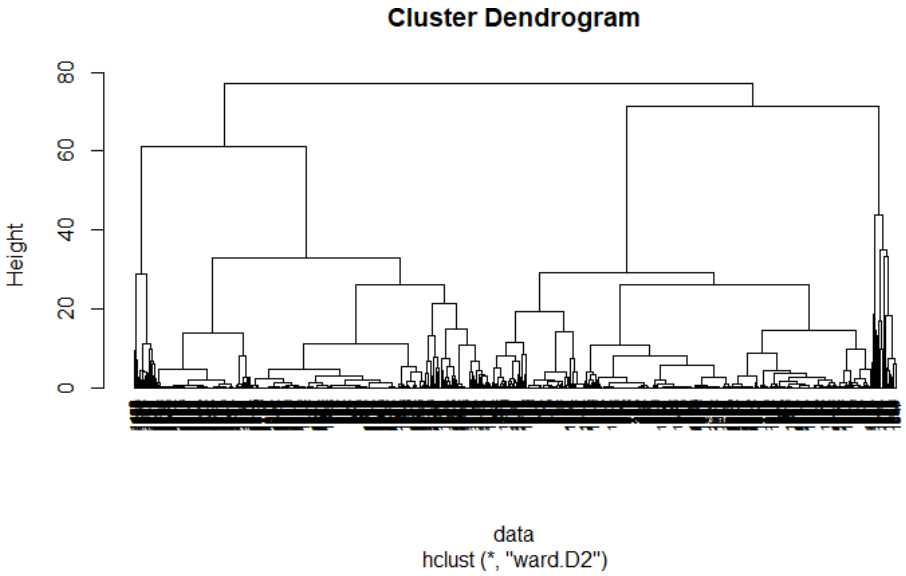
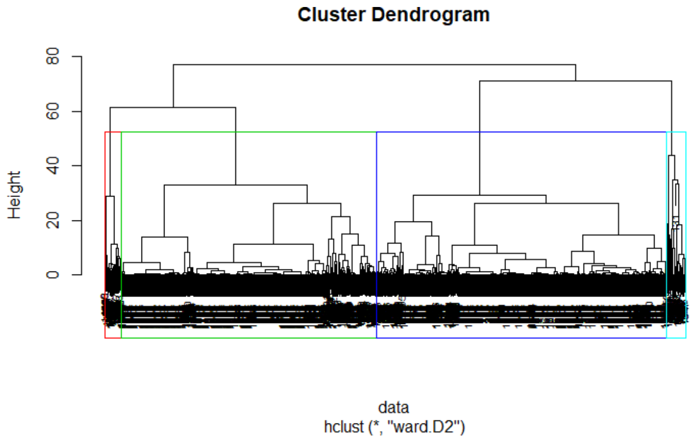
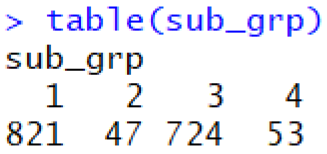


Figure 33: Agglomerative Hierarchical Clustering with HCLUST – Ward

Figure 33 illustrates a good distance between cluster achieved at height 50. This is an indication of a good clustering where 4 clusters was derived at that height. In comparison to the previous dendrogram, it only achieved a clear equal distance with 2 clusters. Thus, concludes Ward dissimilarity method being the good method. However, we still a separate small group that is determined as outliers in this dataset. So, the tree would be cut to k equal to 4 to observe the distribution of observations within those groups, Figure 34 highlights the output.





```
> table(sub_grp)
sub_grp
  1   2   3   4 
821  47 724  53 
```

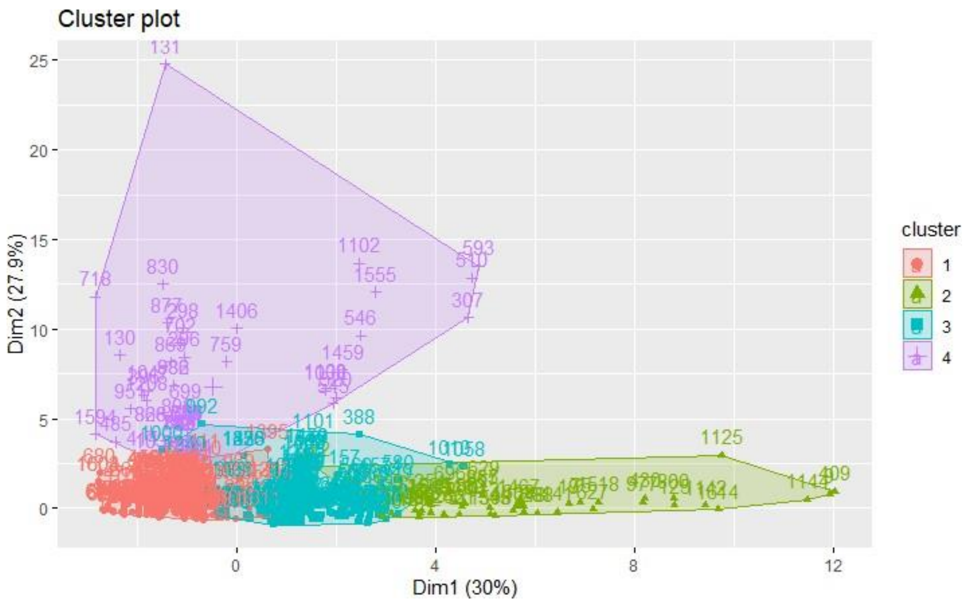


Figure 34: Hierarchical Cluster Output

As the Agglomeration Hierarchical Clustering suggested 4 clusters, next to further explore the optimal number of clusters, the elbow, average silhouette and gap statistic techniques would be used for hierarchical clustering. Following are the output, Figure 35.

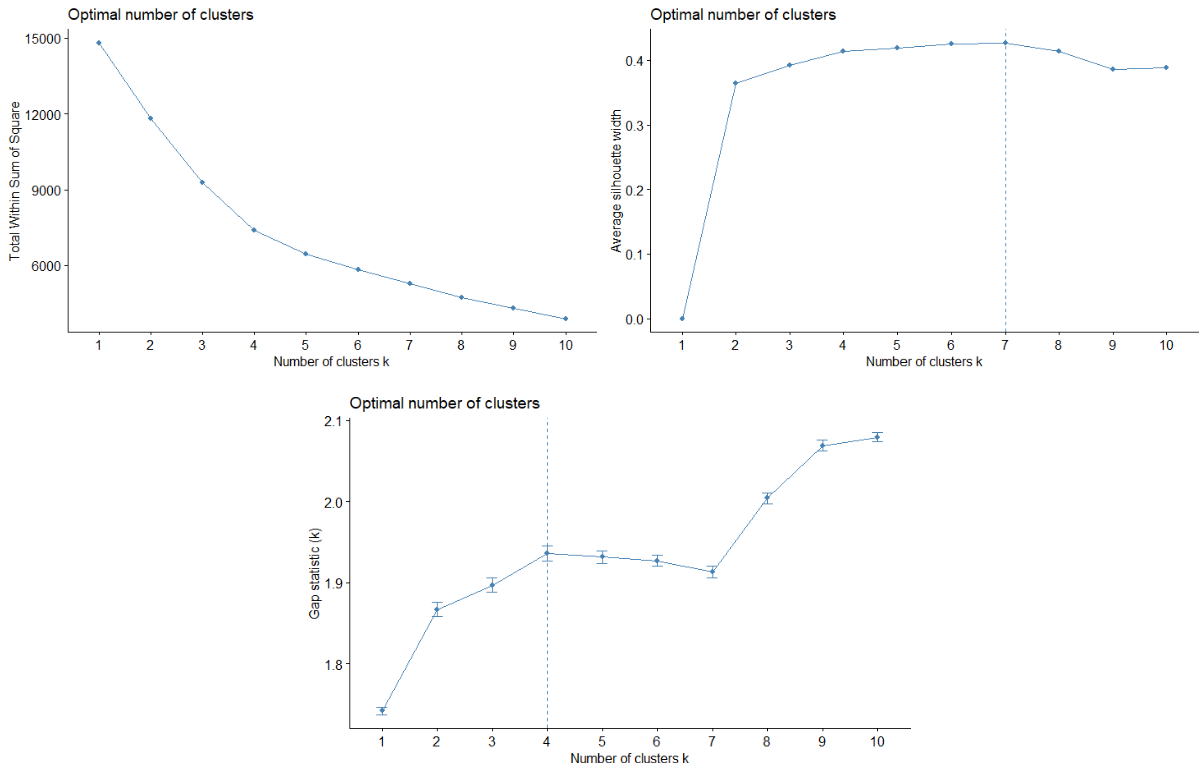
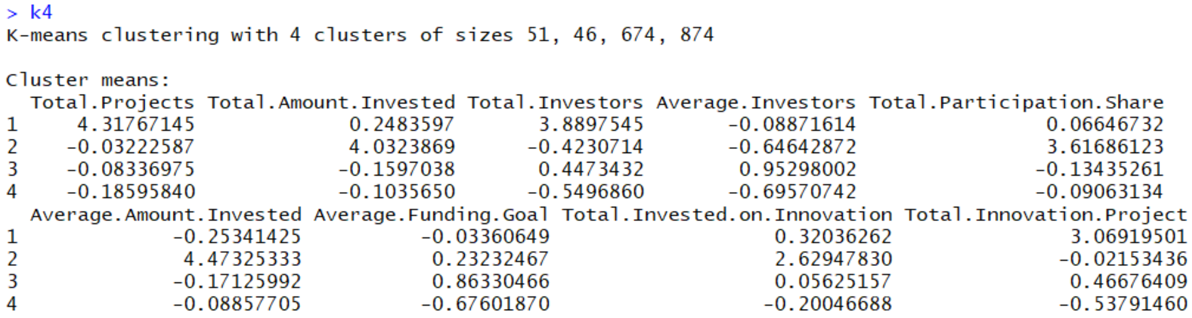


Figure 35: Hierarchical Optimal Clustering - Elbow, Average Silhouette and Gap Statistic

The Elbow method ensure the minimal total within-cluster variation is minimised. Thus, at the point of the elbow bending is considered the optimal number of clusters. Here the elbow method has suggested 4 clusters. Next the average silhouette measure the quality of a cluster, the higher the average silhouette, indication of a good quality clusters. Here, noted that the best number of clusters are 7, however between 4,5 and 6 has almost same average silhouette value as 7. Gap statistics on the other hand measures the total intra-cluster variation for a different set of k, here the optimal number of clusters suggested was 4. Here, we conclude with Hierarchical clustering and utilising the optimal identification techniques, the optimal number of clusters is

4. Next, as we have identified the k, we would use this for the unsupervised machine learning technique for clustering, k-means. This technique used to further explore the clustering to ensure a good quality clusters are obtained, the output of the k = 4 cluster as below Figure 36.



```
> k4
K-means clustering with 4 clusters of sizes 51, 46, 674, 874

Cluster means:
  Total.Projects Total.Amount.Invested Total.Investors Average.Investors Total.Participation.Share
1     4.31767145             0.2483597       3.8897545       -0.08871614                0.06646732
2    -0.03222587             4.0323869      -0.4230714       -0.64642872                3.61686123
3    -0.08336975            -0.1597038       0.4473432        0.95298002               -0.13435261
4    -0.18595840            -0.1035650      -0.5496860       -0.69570742               -0.09063134
  Average.Amount.Invested Average.Funding.Goal Total.Invested.on.Innovation Total.Innovation.Project
1             -0.25341425          -0.03360649                   0.32036262               3.06919501
2              4.47325333           0.23232467                   2.62947830              -0.02153436
3             -0.17125992           0.86330466                   0.05625157               0.46676409
4             -0.08857705          -0.67601870                  -0.20046688              -0.53791460
```

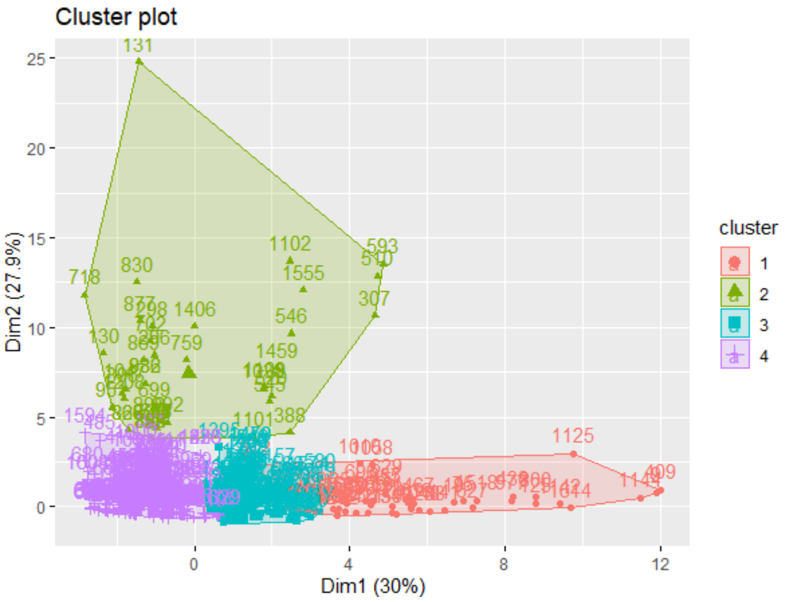


Figure 36: K-Means Clustering (k=4)

Compared to the hierarchical clustering, here we see a small difference in the number of observations in each cluster. Additionally, looking at the clustering plot, noted that for hierarchical clustering there was an obvious overlapping between cluster and here it has been minimised, thus the changes in the observations per cluster. Next, similarly as above the optimal clustering technique would be utilised to confirm if k=4 is the optimal number of clusters, Figure 37.

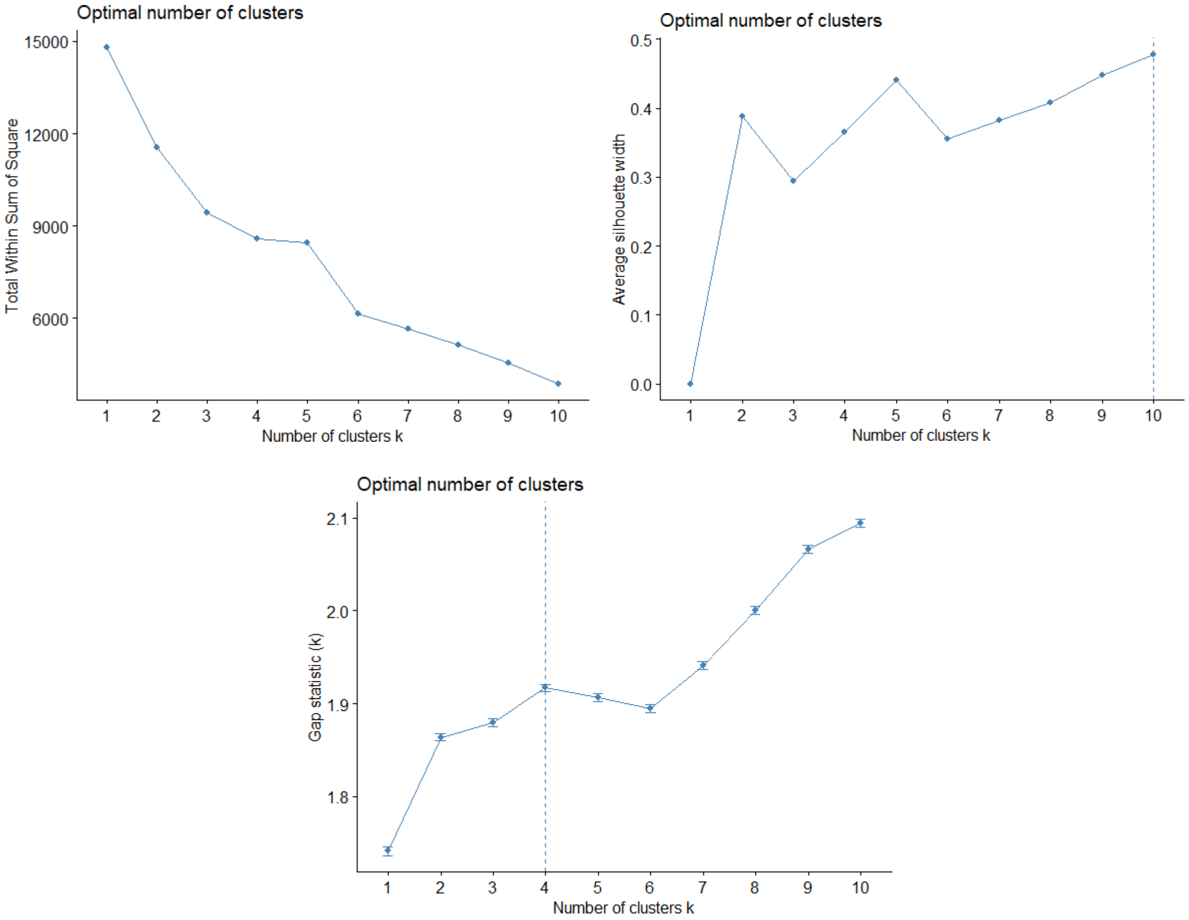
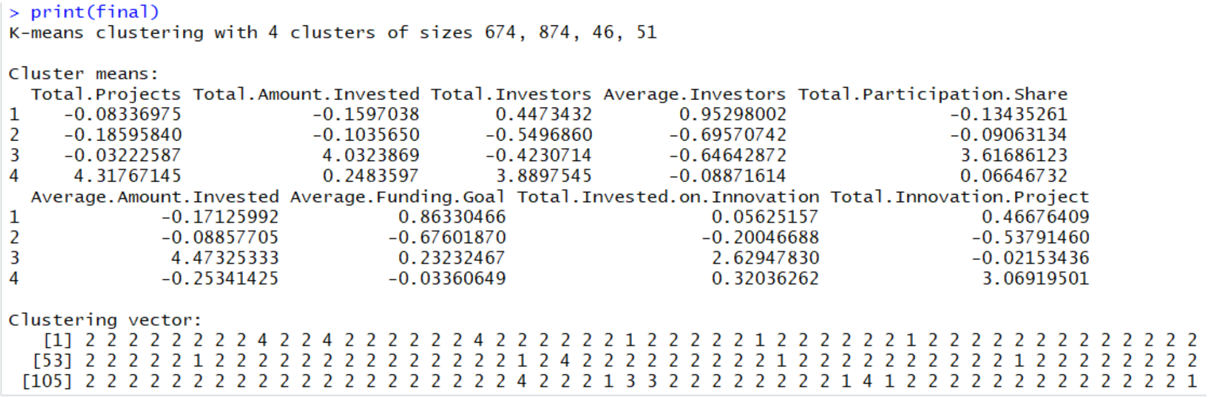


Figure 37: K-means Optimal Clustering - Elbow, Average Silhouette and Gap Statistic

Based on the optimal clustering technique, elbow and gap statistics has produced similar output of having 4 clusters but Average Silhouette suggested 10 as it has the highest average value followed by 5 clusters. Thus, here we conclude 4 clusters as the optimal number of clusters for this dataset, the final output as in Figure 38.



```
> print(final)
K-means clustering with 4 clusters of sizes 674, 874, 46, 51

Cluster means:
  Total.Projects Total.Amount.Invested Total.Investors Average.Investors Total.Participation.Share
1    -0.08336975            -0.1597038       0.4473432        0.95298002               -0.13435261
2    -0.18595840            -0.1035650      -0.5496860       -0.69570742               -0.09063134
3    -0.03222587             4.0323869      -0.4230714       -0.64642872                3.61686123
4     4.31767145             0.2483597       3.8897545       -0.08871614                0.06646732
  Average.Amount.Invested Average.Funding.Goal Total.Invested.on.Innovation Total.Innovation.Project
1             -0.17125992           0.86330466                   0.05625157               0.46676409
2             -0.08857705          -0.67601870                  -0.20046688              -0.53791460
3              4.47325333           0.23232467                   2.62947830              -0.02153436
4             -0.25341425          -0.03360649                   0.32036262               3.06919501

Clustering vector:
   [1] 2 2 2 2 2 2 2 2 4 2 2 4 2 2 2 2 2 2 4 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2
  [53] 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 4 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2
 [105] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 1 3 3 2 2 2 2 2 2 2 2 2 1 4 1 2 2 2 2 2 2 2 2 2 2 2 2 1
```

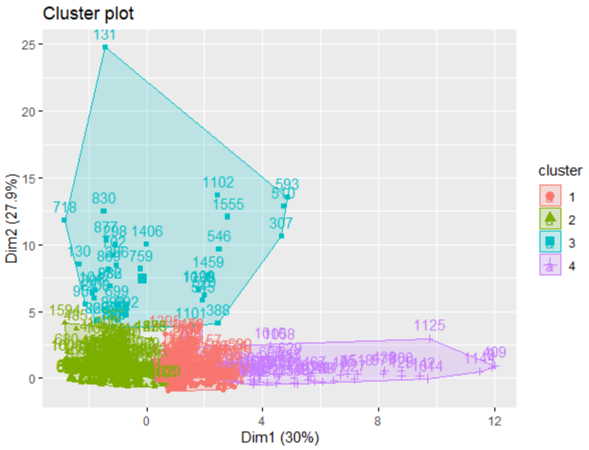


Figure 38: Investor Cluster with k=4

Cluster Interpretation

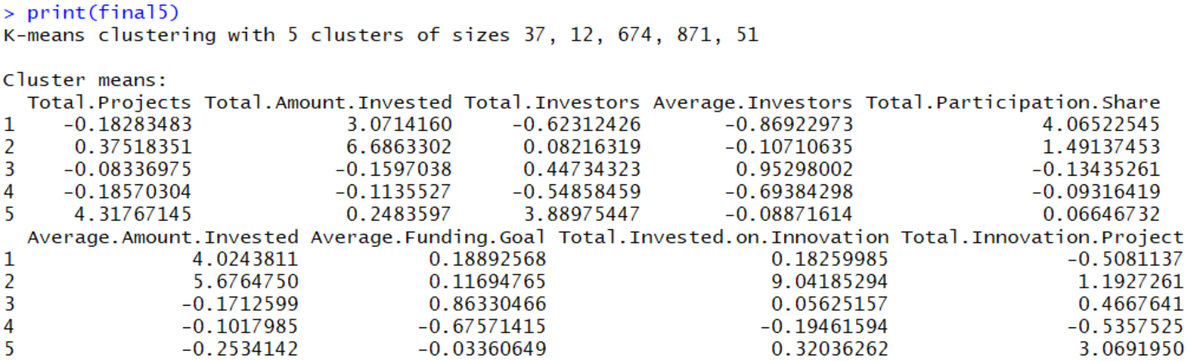
Based on the 4 clusters, we conclude that each cluster has a distribution of 674, 874, 46 and 51 respectively. Cluster 4 having only 51 observations with highest number of campaigns invested on, however low amount invested. Cluster 3 on the other hand has the highest amount invested, however low number of investments done with only 46 investors in that cluster.

Investors is an indicator used to highlight if there exist altruism, herding behaviour as well as financially driven individual. This could be observed together with the Average funding goal, where if see high funding goal with high number of investors, we see an altruism in there. This is due to high funding goal would determine high chances of failure, however despite that we see more investors supporting that initiative, thus an altruism nature. Additionally, if we see low funding goal but having high number of followers, this could be due to herding financially driven as they follow the majority. Here, noted that Cluster 1 and 4 has high number of total investors, however Cluster 4 having a low average funding goal and Cluster 1 having the highest number of investors. This, we could potentially see herding nature in cluster 4 and an altruism nature in Cluster 1. However, as the number of clusters is small, we may be generalising a mix group within that cluster, thus having more break down could further see a clearer grouping.

Next, the total participation share was used to identify the financial motive of an individual, here we expect to see the more amount of money spend would have high number of participations share as they target low share price account thus having more share in return. The one whom spend more with higher participation share are financially motivated and one with high amount spent with less participation share are altruism. Also, one whom has low average amount spend but with high participation share would indicate a financially driven individual. Here, as we know cluster 3 has the highest amount spent, the participation share indicates high as well. The other clusters does not provide enough details, thus an indication of insufficient breakdown of cluster.

Total invested on innovation and amount invested on innovation reflect on altruism driven. This is where we see despite least investment was done, more investment was done for innovation related, Cluster 1. Additionally, on the amount invested Cluster 3 has the highest amount where most of them are innovation driven project. However, we could not conclude this due to that cluster having the largest number of investments thus the effect. In contra, Cluster 1 has the least amount spent but quiet high amount was spent on innovation. This is an indication that Cluster 1 is altruism driven group of people.

This 4 cluster tends to generalise these three attributes where Cluster 2 could not be clearly profile and Cluster 3 and Cluster 4 has a mix nature. Thus, further breakdown of cluster could assist to identify the investor better especially from financially driven, herding and altruism perspective. Thus, further breakdown is required focusing on the attribute total investors, average investor, total participation share and average funding goal. Here, a cluster with k=5 was produced as illustrated in Figure 39.



```
> print(final5)
K-means clustering with 5 clusters of sizes 37, 12, 674, 871, 51

Cluster means:
  Total.Projects Total.Amount.Invested Total.Investors Average.Investors Total.Participation.Share
1    -0.18283483             3.0714160     -0.62312426       -0.86922973                4.06522545
2     0.37518351             6.6863302      0.08216319       -0.10710635                1.49137453
3    -0.08336975            -0.1597038      0.44734323        0.95298002               -0.13435261
4    -0.18570304            -0.1135527     -0.54858459       -0.69384298               -0.09316419
5     4.31767145             0.2483597      3.88975447       -0.08871614                0.06646732
  Average.Amount.Invested Average.Funding.Goal Total.Invested.on.Innovation Total.Innovation.Project
1               4.0243811           0.18892568                   0.18259985               -0.5081137
2               5.6764750           0.11694765                   9.04185294                1.1927261
3              -0.1712599           0.86330466                   0.05625157                0.4667641
4              -0.1017985          -0.67571415                  -0.19461594               -0.5357525
5              -0.2534142          -0.03360649                   0.32036262                3.0691950
```

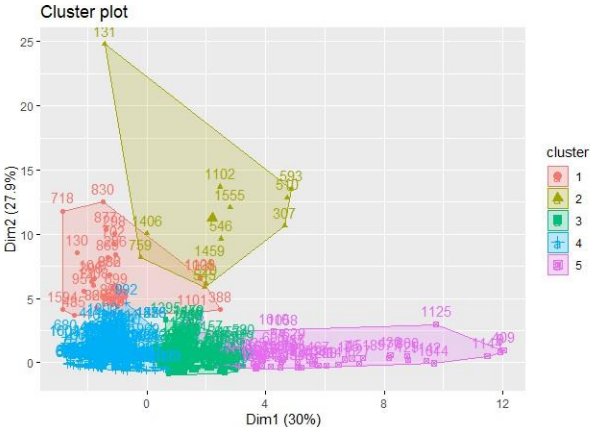
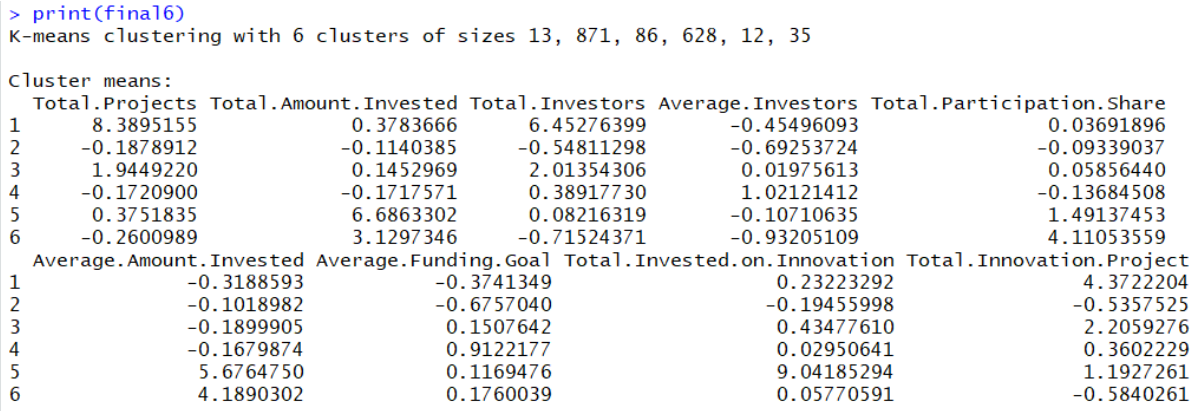
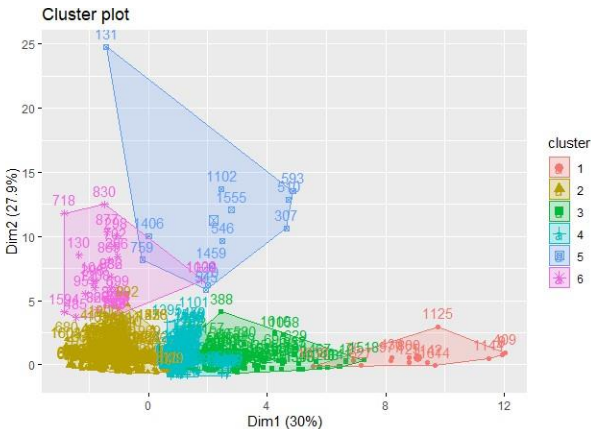


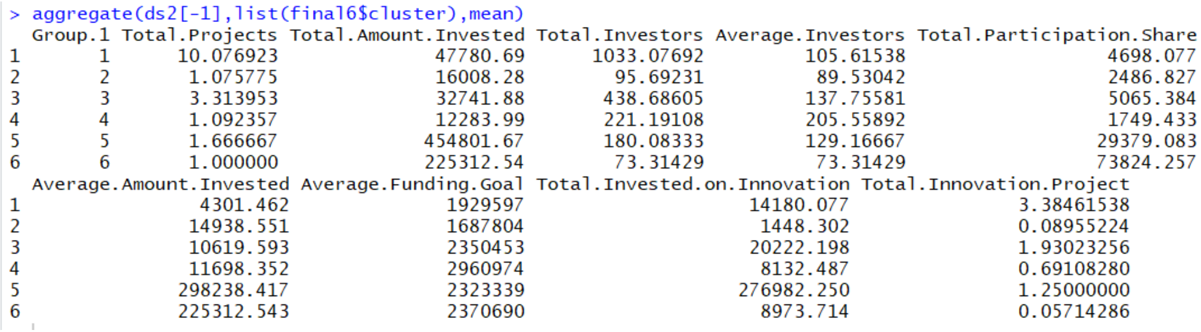
Figure 39: Investor Cluster with k=5

From above, noted that the previously Cluster 3 was further broken down into 2 clusters, Cluster 1 and 2. Noted previously Cluster 2 had only 3 observations redelegated and previously Cluster 4 remained the same. Based on the 5 clusters, the cluster distribution are 37, 12, 674, 871 and 51 respectively. For Cluster 3, noted previously the total number of projects was low and now it was divided between low and high category. However, the total amount invested remained high where those with very little project has the highest amount spent. As highlighted earlier, the Participation share of cluster 3 could not be further distinguish if this highly investing investors have financial motivation on their investment. Here, this k=5 cluster have further broken down that group to highlight those with Altruism nature where despite spending the most, having low participation share, Cluster 2. Cluster 1 on the other hand has high amount invested with even higher participation share, indicating a financially driven investor. However, the previous Cluster 2 and Cluster 4 did not go thru any changes in this k=5 cluster. Thus, the k=6 was further performed to explore the grouping, refer Figure 40.



```
> print(final6)
K-means clustering with 6 clusters of sizes 13, 871, 86, 628, 12, 35

Cluster means:
  Total.Projects Total.Amount.Invested Total.Investors Average.Investors Total.Participation.Share
1      8.3895155             0.3783666      6.45276399       -0.45496093                0.03691896
2     -0.1878912            -0.1140385     -0.54811298       -0.69253724               -0.09339037
3      1.9449220             0.1452969      2.01354306        0.01975613                0.05856440
4     -0.1720900            -0.1717571      0.38917730        1.02121412               -0.13684508
5      0.3751835             6.6863302      0.08216319       -0.10710635                1.49137453
6     -0.2600989             3.1297346     -0.71524371       -0.93205109                4.11053559
  Average.Amount.Invested Average.Funding.Goal Total.Invested.on.Innovation Total.Innovation.Project
1              -0.3188593           -0.3741349                   0.23223292                4.3722204
2              -0.1018982           -0.6757040                  -0.19455998               -0.5357525
3              -0.1899905            0.1507642                   0.43477610                2.2059276
4              -0.1679874            0.9122177                   0.02950641                0.3602229
5               5.6764750            0.1169476                   9.04185294                1.1927261
6               4.1890302            0.1760039                   0.05770591               -0.5840261
```





```
> aggregate(ds2[-1],list(final6$cluster),mean)
  Group.1 Total.Projects Total.Amount.Invested Total.Investors Average.Investors Total.Participation.Share
1       1      10.076923              47780.69     1033.07692         105.61538                  4698.077
2       2       1.075775              16008.28       95.69231          89.53042                  2486.827
3       3       3.313953              32741.88      438.68605         137.75581                  5065.384
4       4       1.092357              12283.99      221.19108         205.55892                  1749.433
5       5       1.666667             454801.67      180.08333         129.16667                 29379.083
6       6       1.000000             225312.54       73.31429          73.31429                 73824.257
  Average.Amount.Invested Average.Funding.Goal Total.Invested.on.Innovation Total.Innovation.Project
1                4301.462              1929597                    14180.077               3.38461538
2               14938.551              1687804                     1448.302               0.08955224
3               10619.593              2350453                    20222.198               1.93023256
4               11698.352              2960974                     8132.487               0.69108280
5              298238.417              2323339                   276982.250               1.25000000
6              225312.543              2370690                     8973.714               0.05714286
```

Figure 40: Investor Cluster with k = 6

Here, finally we observed the initial Cluster 4 has been broken-down further to cluster 1 and 3. Initially Cluster 1 has regrouped 56 observation into the current Cluster 3. The previously Cluster 4 value had changed where now Cluster 1 and Cluster 3 has clearly distinguish between the total number of campaigns, total and average investors and average funding goals. Cluster 1 has the highest campaigns thus the high total number of investors but a low average investor. Cluster 3 on the other hand, has moderate total project with less amount invested where the average funding goal is also moderate, thus a potential mild herding behaviour. As we noticed the Cluster 2 was never further broken-down till now, thus the k would be set to 6 and further exploration of the cluster was performed.

4.2.6 Discussion on findings

As per our first objective is to identify the investor's financial behaviour taxonomy, the cluster analysis above have managed to distinguish six different clusters to indicate the investor's financial behaviour. Next, we would analyse each variable and the respective clusters to identify their nature. The total projects by total amount spent was observed as illustrated in Figure 41.

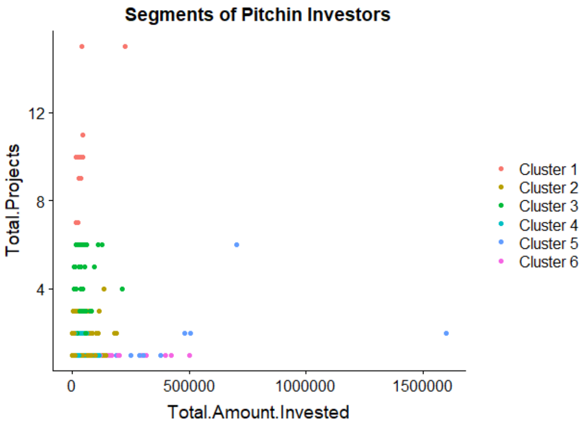


Figure 41: Investor Cluster - Total Project by Total Amount Spent

Based on plot above, the number of projects invested by Cluster 1 are between 7 to 15 projects with a mean of 10 projects. However, the amount spent is less which is within RM 48k. Cluster 3 on the other hand has moderate number of projects invested between 1 to 7 projects and amount invested is also moderate around RM33k. Cluster 5 invested the most around RM455k followed by Cluster 6 invested around RM225k, however Cluster 6 has only invested once whereby Cluster 5 has investors whom have invested more than 1 campaign. Cluster 2 and 4 have low number of total projects as well as low amount invested between RM12k to RM16k. A new comer whom are financially motivated would generally have less investment performed but with large amount invested (Wallmeroth, 2019), this behaviour is seen in mainly in Cluster 6 and slight on Cluster 5. If the average amount spent is also less, thus confirms to be a new comer. Next, an observation on the average amount invested by the clusters would be performed. Here we would be able to analyse the spending capabilities of the investors within a cluster, outcome as illustrated in Figure 42.

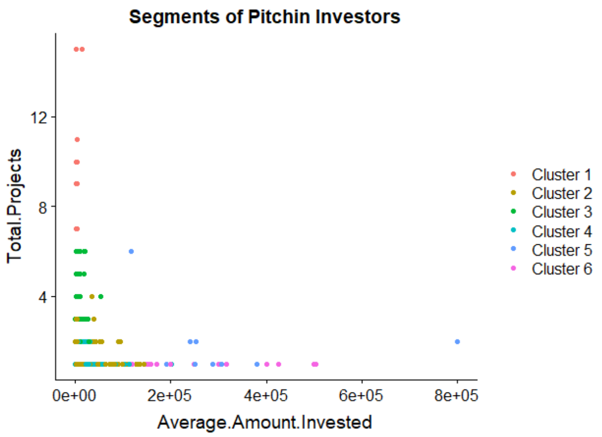


Figure 42: Investor Cluster - Total Project by Average Amount Spent

As described earlier, Cluster 5 having the most amount spent thus the average amount spent is also the highest at RM228k followed by Cluster 6 with no difference in the average amount spent at RM225k. The difference between both this cluster is the number of campaigns invested where Cluster 6 invest less compared to Cluster 5. Additionally, we noticed a liking behaviour pattern where investors in Cluster 6 had only invested big amount on just 1 investment with no difference between the total amount invested to the average amount spent. This indicates they are a family or friend or partner of the campaign thus the intention to pump in large investment on the campaign(Hornuf and Schmitt, 2016). On the other hand, Cluster 2 and Cluster 4 had the lowest amount spent, thus a lower average amount spent was foreseen. The difference was noted in Cluster 1 and 3 where they had a moderate level of amount spent, however looking at the average spent in each campaign, it is the low at RM4k and RM11k. Cluster 1 having the lowest average amount invested despite having a higher total amount spent, this could be due to the highest number of project invested by them. This explains that this Cluster 1 and Cluster 3 spends on many projects but at an average they invest less amount on each project which are financially motivated individuals whom are casual investors (Goethner, Luettig and Regner, 2018). Next, to further distinguish the financial and altruism motive, the total participation share was observed, Figure 43 highlights the output. If more money spent and obtain more participation share, that is an indication of financially motivated investors and vice versa for altruism investors (Goethner, Luettig and Regner, 2018).

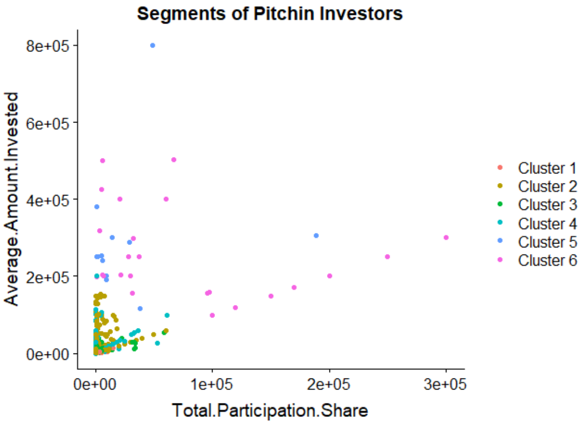


Figure 43: Investor Cluster - Total Project by Total Participation Share

However, here we noted that Cluster 5 whom has spent the most, but the participation share is moderate to low. This indicates Cluster 5 individuals are not financially driven and more community driven. Cluster 6 on the other hand with high investment done, also has a higher participation share thus a confirmed financially motivated individual. Thus, the exponential increase noted for Cluster 6 in the graph. The participation share could not clearly distinguish the other clusters due to the low amount invested. As the total participation share highlights on the financial motivation, the next graph would explore the relationship between average investors to the total participation share as illustrated in Figure 44. For those whom are financially motivated, we would expect to see low average number of investors and for those community motivated, highest number of investors (Goethner, Luettig and Regner, 2018).

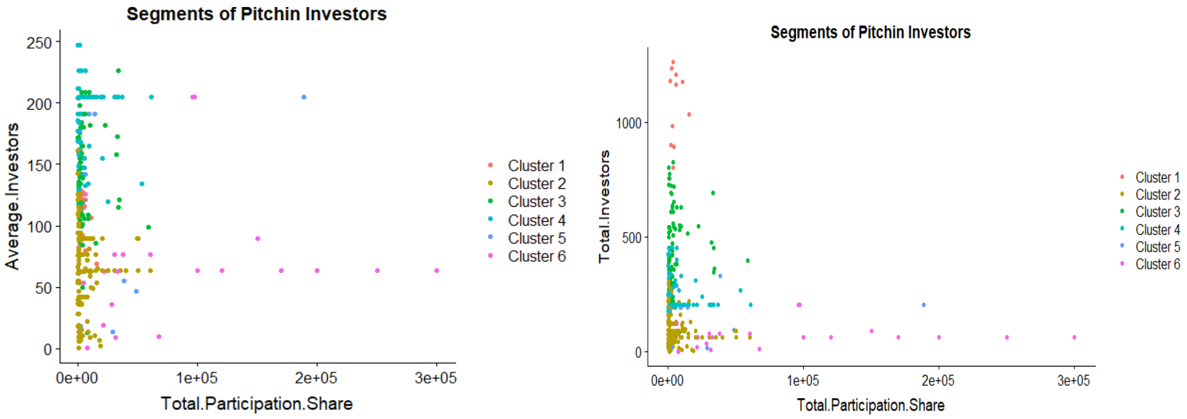


Figure 44: Investor Cluster – Average Investor and Total Investors by Total Participation Share

Cluster 6 was known for being highly financially motivated, figure above noted that the average number of investors are at 73 being the lowest. This confirms that this group are financially driven and no community element. Cluster 5 as noted earlier to be having altruism nature, here noted that the total investors are moderate to low thus could still potentially be a community driven behaviour. Cluster 4 on the other hand, having lowest number of projects invested as well as amount invested but has the highest average investor indicating a potential herding behaviour where they invest on popular projects with less amount. A very prominent herding behaviour was not obvious in this dataset as there was no cluster with high financial motivation and high average investors. But Cluster 3 has moderate to low financial motivation and moderate average investors, further analysis on this cluster is required to confirm the motivation and behaviour of that cluster. Next, to confirm the altruism nature of the cluster, the Innovation indicator was observed in Figure 45. Here we expect to see low investment, but high investment done on innovation thus a community driven motive. On the other hand, high funded project with high participation share and small amount on innovation, confirms a community driven individuals (Goethner, Luettig and Regner, 2018).

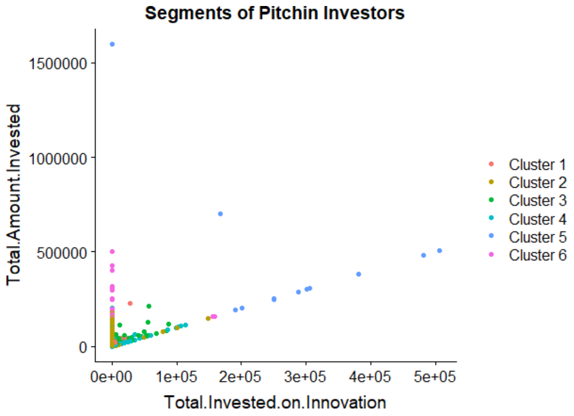


Figure 45: Investor Cluster – Total Amount Invested by Total Amount Invested on Innovation

Based on the plot above, we could confirm the altruism behaviour of investors in Cluster 5 where the more amount invested, it is invested on innovative campaigns. Additionally, Cluster 6 confirms the non-altruism whom are financially motivated investors where the more money spent, it is not spent on innovation campaign. Cluster 1 and Cluster 3 was flagged to be financially motivated, the innovation investment observed and noted that Cluster 3 despite financially motivated it has a mild community conscious where more two third of invested money was done on innovation driven project. Instead, Cluster 1 only have around 30% of what they have invested, invested on innovation. Average Funding Goal was meant to use to identify further the difference between altruism and financial driven investors. Here for a altruism cluster, we expect to see high average funding goal whereby for a financial motivated cluster, we would see low average funding goal as minimal risk is expected to be taken by those investors (Lin, Boh and Goh, 2014). Noted from the cluster mean summary, there is no clear distinguish between the clusters for Average Funding Goal thus an indication that the variable is not a strong variable to distinguish the attributes. However, an observation on those financially motivated clusters (1,3,6) and altruism clusters (3,5) to see if we could observe a clear difference on the average funding goals, refer Figure 46.

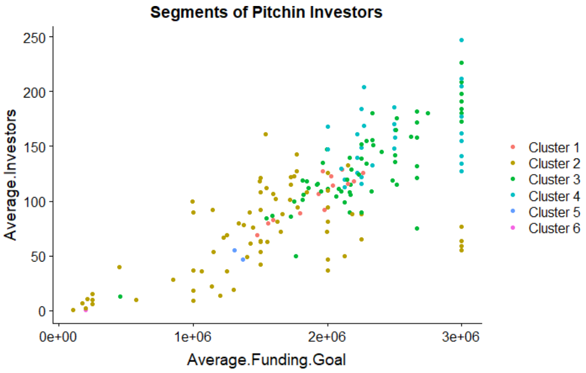


Figure 46: Investor Cluster – Average Investor by Average Funding Goal

Noted that the Cluster 1 having low average funding goal thus confirms further the financially motivated cluster. For Cluster 6, it was a moderate outcome thus this attribute could not be used to confirm the financial motivation of this cluster. Cluster 3 on the other hand having a mix of behaviour between altruism and financially motivated, this average funding goal was at moderate to high thus confirming further the mix nature. Cluster 2 and 5 could not be further distinguished by this variable. The summary of the discussion above is tabulated in Table 1 below.

Table 15: Summary of Investor Clusters

Criteria	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Total Observations	13	871	86	628	12	35
Total Project	7 to 15	1 to 4	1 to 7	1 to 2	1 to 6	1
Average Total Project	10	1	3	1	1	1
Total Amount Invested	RM48k	RM16k	RM33k	RM12k	RM455k	RM225k
Average Amount Invested	RM4k	RM15k	RM11k	RM12k	RM298k	RM225k
Participation Share #	5k	3k	5k	2k	29k	74k
Participation Share	Moderate	Low	Moderate	Low	Lower	Very High
Average Funding Goal	Low	Low	Moderate	High	Moderate	Moderate
Total Investors	High	Low	Moderate	Moderate	Moderate	Low
Average Investor	Moderate	Low	Moderate	High	Moderate	Low
Total Invested on Innovation	RM14k	RM1k	RM20k	RM8k	RM277k	RM9k

In conclusion, these variables was able to distinguish most of all clusters expect Cluster 2. Cluster 2 having the most observations but none of the variables could clearly distinguish this cluster's motivation or behaviour expect we see almost all amount invested was invested on Innovation Project, thus we see an altruism driven individual with very low investment done and amount spent. Table 16 summarises each of the cluster with their behaviour or motivation and a profile name would be given for each cluster.

Table 16: PitchIn's Investor Taxonomy

Cluster	Cluster Profile	Motivation	Attributes Details
1	Active Casual	Financial	• **Highest** number of invested projects • **Small** quantum of investment per project • **Less** focused on innovation driven project • Self-reliant investors • Seeks **moderate** level of participation share
2	Common	Innovation	• **Majority** investment was made on innovation project
3	Altruistic Casual	Financial, Altruism	• **Moderate** number of invested projects • **Moderate** quantum of investment per project • **Slightly** focused on innovation driven project • Self-reliant investors • Seeks **moderate** level of participation share and funding goals
4	Trend Followers	Herding	• **Small** number of invested projects • **Small** quantum of investment • **Dependent** investors
5	Altruistic Sophisticated	Altruism	• Participation share is **low** despite **high** quantum of investment indication of **high share price** • Community driven projects with **high** number of investors • **Highly** focused on innovation project
6	Sophisticated	Financial, Liking	• Invested **once** with **large** quantum of investment • Participation share is **high** equally **high** on quantum of investment – **low share price** • Self-reliant investors • **Less** focused on innovation driven project • **Moderate** funding goal

4.3 Predictive Modelling

4.3.1 Initial Data Preparation

Pitch view was created for the Predictive Modelling. All 4 files was used for the creation of this view and the clean-up as detailed in Figure 47.

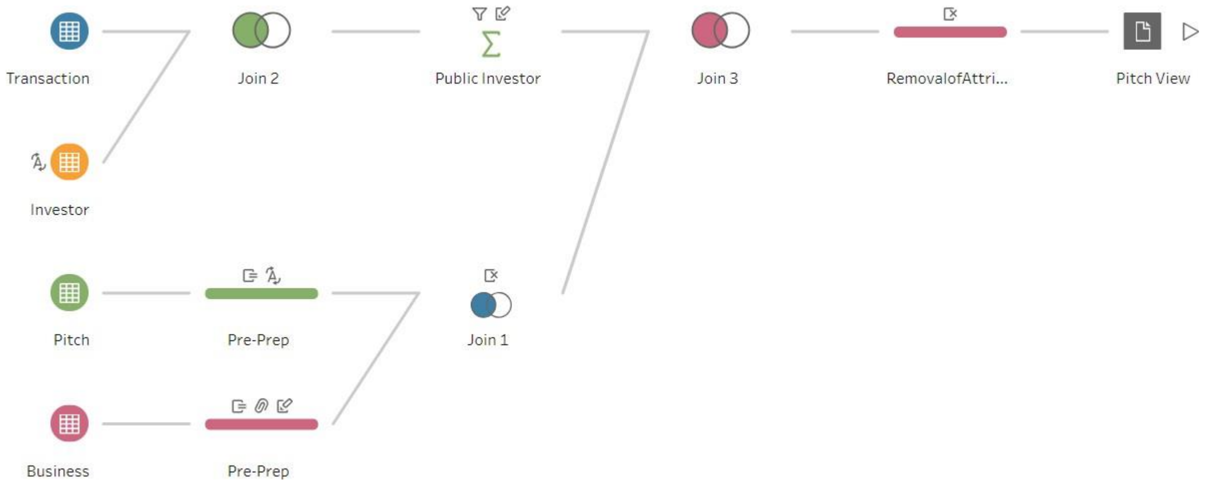


Figure 47: Initial data preparation in Tableau Prep

The details activity that was performed documented in Table 17.

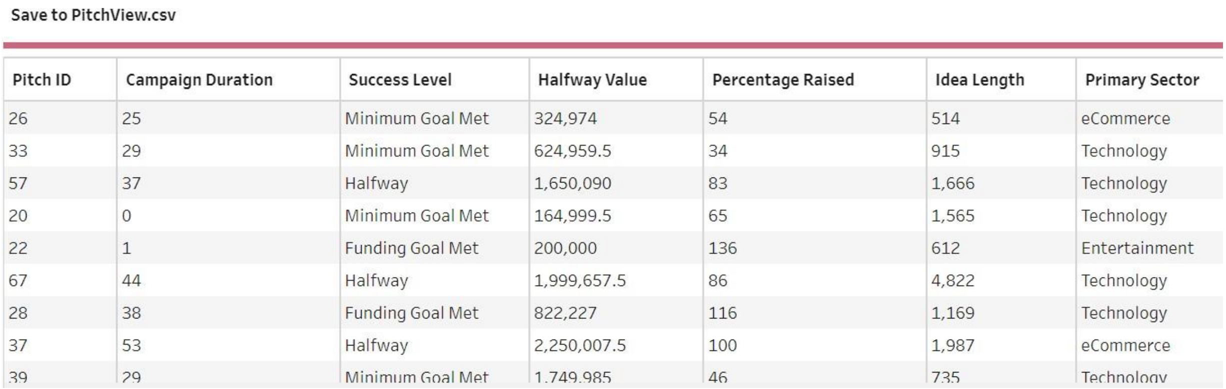
Table 17: Pitch View Detail Initial Preparation

File	**Attribute**	**Action**	**New Attribute**	**Code**
Business	Sector	Split with "," to extract the first 2 sectors from the list of sectors involved.	Primary Sector Secondary Sector	
	State	Clean up the state to standardise the naming convention		
	Description	Length of description		LEN(Description)
Pitch	Idea	Length of Idea		LEN(Idea)
	Banked In Amount, Funding Goal	Calculate the percentage raised	Percentage Raised	ROUND((([Banked In Amount]/[Funding Goal])*100),0)
	Funding Goal, Min Target	Calculate the halfway value based on the funding goal and minimum target	Halfway Value	(([Funding Goal]-[Min Target])/2)+[Min Target]

	Funding Goal, Halfway Value, Banked In Amount,	Create the dependent variable Success Level for the predictive analytics	Success Level	IF [Banked In Amount] >= [Funding Goal] THEN 'Funding Goal Met'
				ELSEIF [Banked In Amount] >= [Halfway Value] AND [Banked In Amount] < [Funding Goal] THEN 'Halfway'
				ELSE 'Minimum Goal Met'
				END

	Start Campaign Date, Last Transaction date	Change to date and time type		
	Start Campaign Date, Last Transaction Date	Calculate the funding duration	Campaign Duration	DATEDIFF('day',[Start Campaign Date],[Last Transaction At],'Monday')
Transaction - Investor	Merge Both files			
	Pitch ID Anonymous? Investor ID	Create Aggregated value for Total Public Investors	Total Public Investor	Group Pitch ID, Filter Anonymous? = False Count Distinct Investor ID

The final output was obtained by removing 3 redundant fields with total 55 fields, Figure 48 illustrates the output.



Save to PitchView.csv

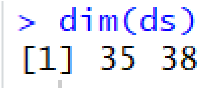
Pitch ID	Campaign Duration	Success Level	Halfway Value	Percentage Raised	Idea Length	Primary Sector
26	25	Minimum Goal Met	324,974	54	514	eCommerce
33	29	Minimum Goal Met	624,959.5	34	915	Technology
57	37	Halfway	1,650,090	83	1,666	Technology
20	0	Minimum Goal Met	164,999.5	65	1,565	Technology
22	1	Funding Goal Met	200,000	136	612	Entertainment
67	44	Halfway	1,999,657.5	86	4,822	Technology
28	38	Funding Goal Met	822,227	116	1,169	Technology
37	53	Halfway	2,250,007.5	100	1,987	eCommerce
39	29	Minimum Goal Met	1,749,985	46	735	Technology

Figure 48: Pitch View Output

4.3.2 Initial Data Exploration

4.3.2.1 Data Description and Data Type

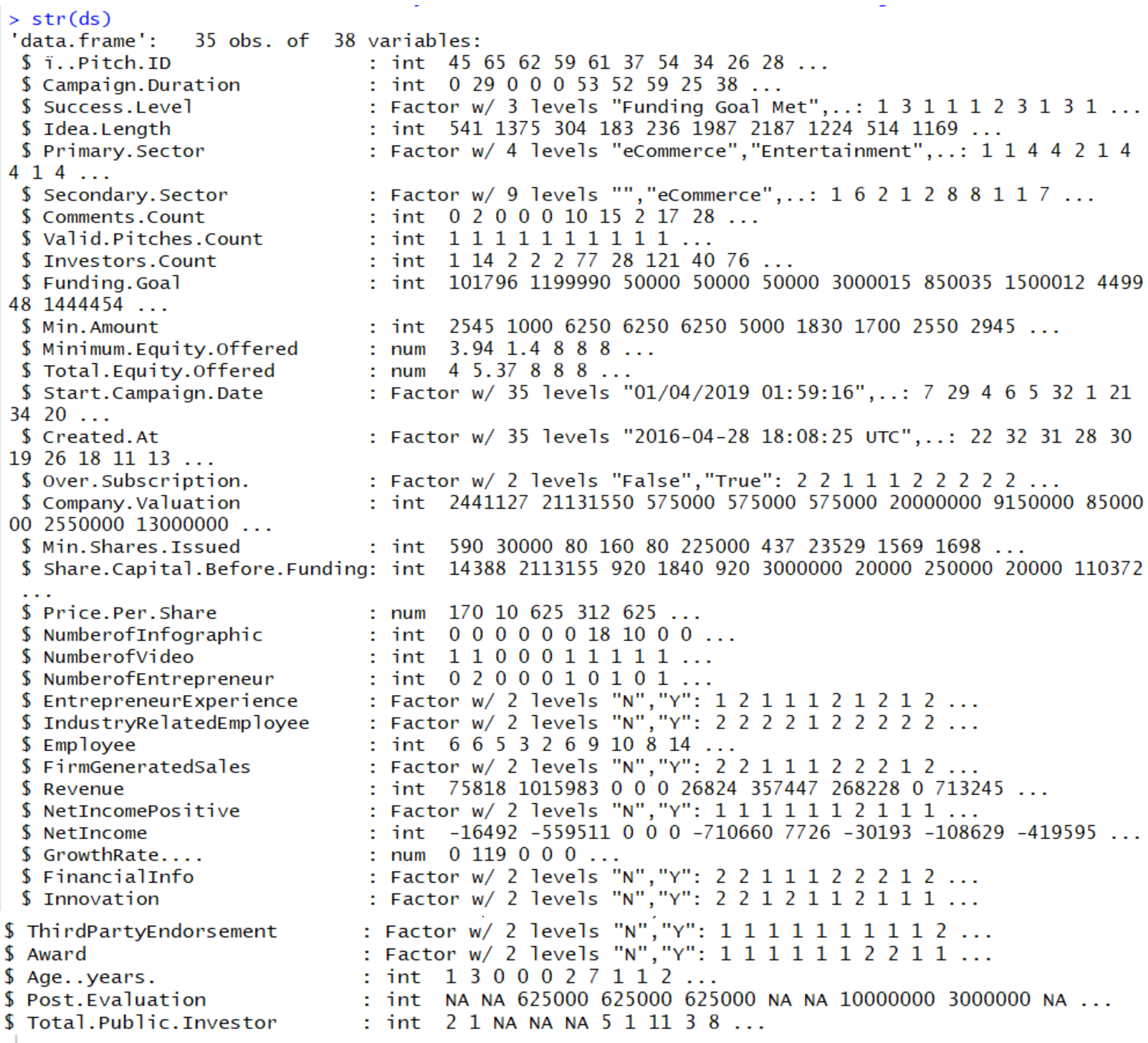
This pitch view dataset has a total of 35 campaigns which are successful, and 55 variables. However, not all 55 variables are relevant thus removal of redundant variables was performed and the data description after the removal as illustrated in Figure 49.



```
> dim(ds)
[1] 35 38
```

Figure 49: Pitch View data dimension

Total 36 variables would be used for this predictive modelling. he variables are identified, their data type, length, variable type as well as labels for categorical variable was populated as in Figure 50. The Pitch ID is the unique identifier for each campaign and the target variable is Success Level which is a multiclass with three labels.



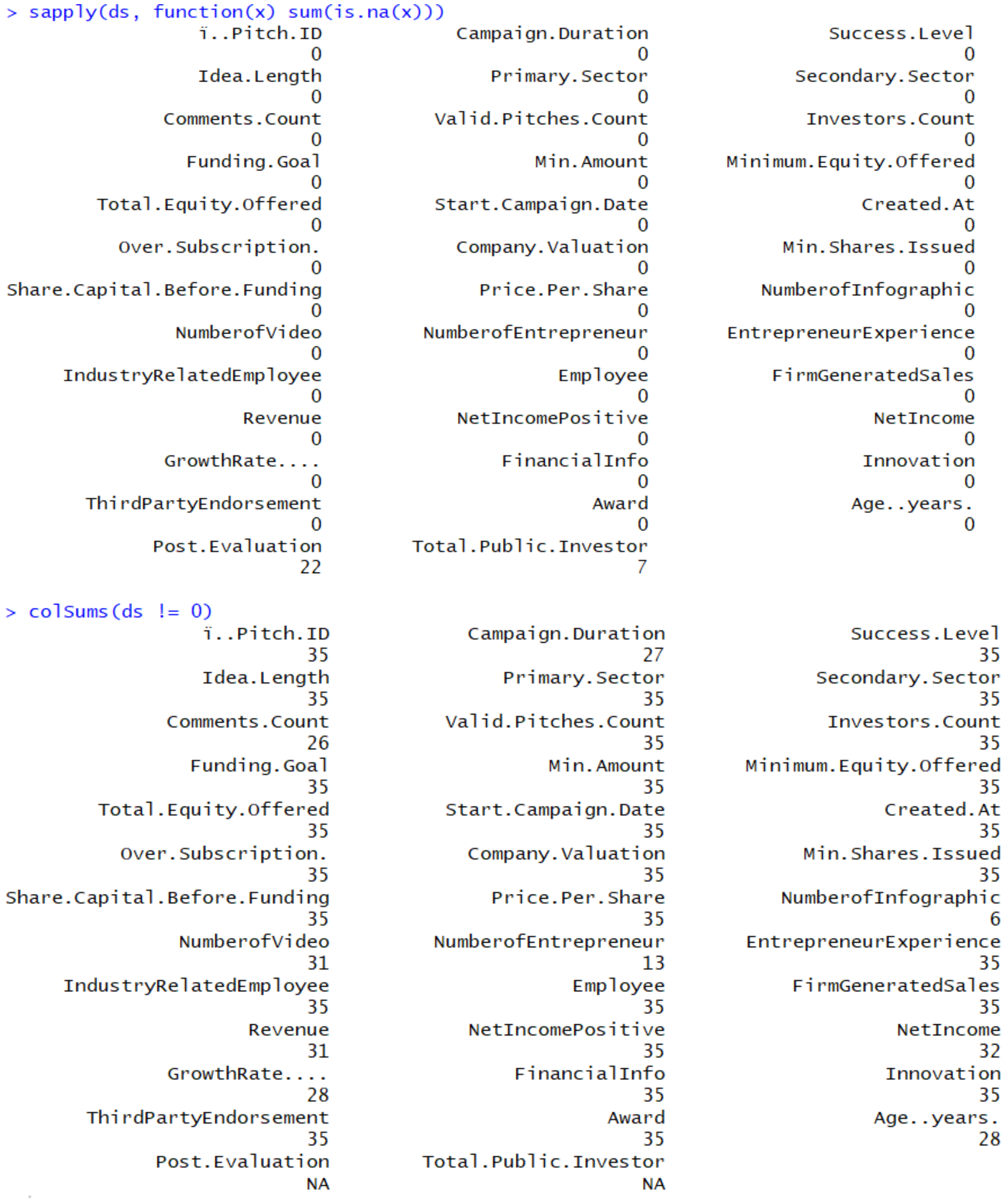
```
> str(ds)
'data.frame':	35 obs. of  38 variables:
 $ ï..Pitch.ID                : int  45 65 62 59 61 37 54 34 26 28 ...
 $ Campaign.Duration          : int  0 29 0 0 0 53 52 59 25 38 ...
 $ Success.Level              : Factor w/ 3 levels "Funding Goal Met",..: 1 3 1 1 1 2 3 1 3 1 ...
 $ Idea.Length                : int  541 1375 304 183 236 1987 2187 1224 514 1169 ...
 $ Primary.Sector             : Factor w/ 4 levels "eCommerce","Entertainment",..: 1 1 4 4 2 1 4 
4 1 4 ...
 $ Secondary.Sector           : Factor w/ 9 levels "","eCommerce",..: 1 6 2 1 2 8 8 1 1 7 ...
 $ Comments.Count             : int  0 2 0 0 0 10 15 2 17 28 ...
 $ Valid.Pitches.Count        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Investors.Count            : int  1 14 2 2 2 77 28 121 40 76 ...
 $ Funding.Goal               : int  101796 1199990 50000 50000 50000 3000015 850035 1500012 4499
48 1444454 ...
 $ Min.Amount                 : int  2545 1000 6250 6250 6250 5000 1830 1700 2550 2945 ...
 $ Minimum.Equity.Offered     : num  3.94 1.4 8 8 8 ...
 $ Total.Equity.Offered       : num  4 5.37 8 8 8 ...
 $ Start.Campaign.Date        : Factor w/ 35 levels "01/04/2019 01:59:16",..: 7 29 4 6 5 32 1 21 
34 20 ...
 $ Created.At                 : Factor w/ 35 levels "2016-04-28 18:08:25 UTC",..: 22 32 31 28 30 
19 26 18 11 13 ...
 $ Over.Subscription.         : Factor w/ 2 levels "False","True": 2 2 1 1 1 2 2 2 2 2 ...
 $ Company.Valuation          : int  2441127 21131550 575000 575000 575000 20000000 9150000 85000
00 2550000 13000000 ...
 $ Min.Shares.Issued          : int  590 30000 80 160 80 225000 437 23529 1569 1698 ...
 $ Share.Capital.Before.Funding: int  14388 2113155 920 1840 920 3000000 20000 250000 20000 110372
 ...
 $ Price.Per.Share            : num  170 10 625 312 625 ...
 $ NumberofInfographic        : int  0 0 0 0 0 0 18 10 0 0 ...
 $ NumberofVideo              : int  1 1 0 0 0 1 1 1 1 1 ...
 $ NumberofEntrepreneur       : int  0 2 0 0 0 1 0 1 0 1 ...
 $ EntrepreneurExperience     : Factor w/ 2 levels "N","Y": 1 2 1 1 1 2 1 2 1 2 ...
 $ IndustryRelatedEmployee    : Factor w/ 2 levels "N","Y": 2 2 2 2 1 2 2 2 2 2 ...
 $ Employee                   : int  6 6 5 3 2 6 9 10 8 14 ...
 $ FirmGeneratedSales         : Factor w/ 2 levels "N","Y": 2 2 1 1 1 2 2 2 1 2 ...
 $ Revenue                    : int  75818 1015983 0 0 0 26824 357447 268228 0 713245 ...
 $ NetIncomePositive          : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 2 1 1 1 ...
 $ NetIncome                  : int  -16492 -559511 0 0 0 -710660 7726 -30193 -108629 -419595 ...
 $ GrowthRate....             : num  0 119 0 0 0 ...
 $ FinancialInfo              : Factor w/ 2 levels "N","Y": 2 2 1 1 1 2 2 2 1 2 ...
 $ Innovation                 : Factor w/ 2 levels "N","Y": 2 2 1 2 1 1 2 1 1 1 ...
$ ThirdPartyEndorsement       : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 2 ...
$ Award                       : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 2 2 1 1 ...
$ Age..years.                 : int  1 3 0 0 0 2 7 1 1 2 ...
$ Post.Evaluation             : int  NA NA 625000 625000 625000 NA NA 10000000 3000000 NA ...
$ Total.Public.Investor       : int  2 1 NA NA NA 5 1 11 3 8 ...
```

Figure 50: Pitch View data structure

There are a mix of numerical and categorical variables. Also, 2 variables (Start Campaign Date and Created At) that are datetime however being reflected as factor, thus requires a data type clean up if it is being utilised.

4.3.2.2 Missing Value Identification

Next the check if there is any missing value, two approach to first check if there is any NA and subsequently check the total number of 0 in a dataset was performed. The 2nd approach was taken as we expect certain numerical field to not have zero value. Figure 51 indicates if there is any NA or missing value in the dataset.



```
> sapply(ds, function(x) sum(is.na(x)))
               ï..Pitch.ID          Campaign.Duration              Success.Level
                         0                          0                          0
               Idea.Length             Primary.Sector           Secondary.Sector
                         0                          0                          0
            Comments.Count         Valid.Pitches.Count            Investors.Count
                         0                          0                          0
              Funding.Goal                 Min.Amount     Minimum.Equity.Offered
                         0                          0                          0
      Total.Equity.Offered        Start.Campaign.Date                 Created.At
                         0                          0                          0
         Over.Subscription.          Company.Valuation          Min.Shares.Issued
                         0                          0                          0
Share.Capital.Before.Funding           Price.Per.Share        NumberofInfographic
                         0                          0                          0
             NumberofVideo       NumberofEntrepreneur     EntrepreneurExperience
                         0                          0                          0
     IndustryRelatedEmployee                  Employee          FirmGeneratedSales
                         0                          0                          0
                   Revenue           NetIncomePositive                  NetIncome
                         0                          0                          0
             GrowthRate....              FinancialInfo                 Innovation
                         0                          0                          0
     ThirdPartyEndorsement                      Award                Age..years.
                         0                          0                          0
           Post.Evaluation       Total.Public.Investor
                        22                          7

> colSums(ds != 0)
               ï..Pitch.ID          Campaign.Duration              Success.Level
                        35                         27                         35
               Idea.Length             Primary.Sector           Secondary.Sector
                        35                         35                         35
            Comments.Count         Valid.Pitches.Count            Investors.Count
                        26                         35                         35
              Funding.Goal                 Min.Amount     Minimum.Equity.Offered
                        35                         35                         35
      Total.Equity.Offered        Start.Campaign.Date                 Created.At
                        35                         35                         35
         Over.Subscription.          Company.Valuation          Min.Shares.Issued
                        35                         35                         35
Share.Capital.Before.Funding           Price.Per.Share        NumberofInfographic
                        35                         35                          6
             NumberofVideo       NumberofEntrepreneur     EntrepreneurExperience
                        31                         13                         35
     IndustryRelatedEmployee                  Employee          FirmGeneratedSales
                        35                         35                         35
                   Revenue           NetIncomePositive                  NetIncome
                        31                         35                         32
             GrowthRate....              FinancialInfo                 Innovation
                        28                         35                         35
     ThirdPartyEndorsement                      Award                Age..years.
                        35                         35                         28
           Post.Evaluation       Total.Public.Investor
                        NA                         NA
```

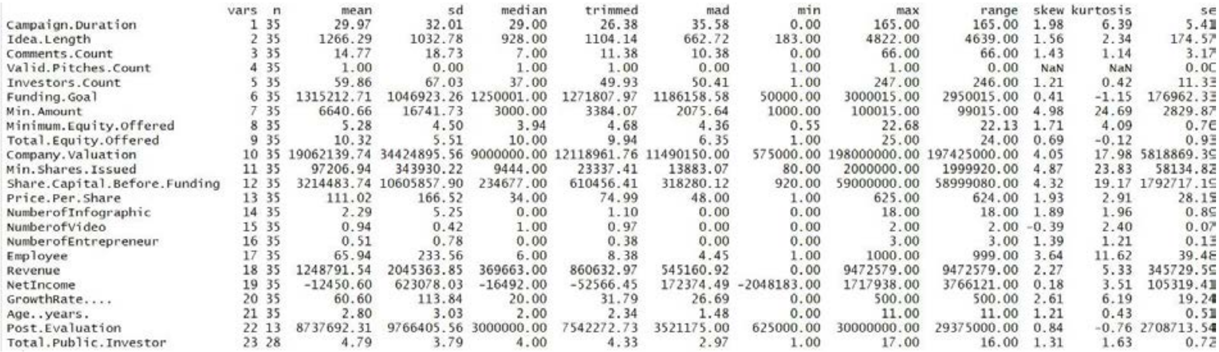
Figure 51: Pitch View - Missing Value

The output confirmed Post Evaluation and Total Public Investor having NA value, thus this requires missing value treatment. On the non-zero value check, noted variables Campaign Duration, Comments Count, Number of Infographic, Number of Video, Number of Entrepreneur, Revenue, Net Income, Growth Rate and Age are having zero in their observations, confirms this is acceptable. Thus, here concludes that only variable Post

Evaluation and Total Public Investor requires missing value treatment where as both are numerical variables, a mean imputation would be performed.

4.3.2.3 Outlier Identification

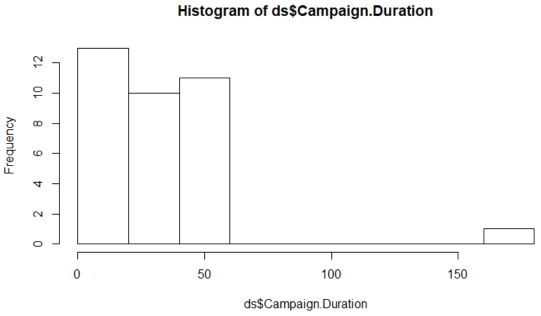
The statistical summary was explored for the numerical variable, the outcome as below Figure 52.

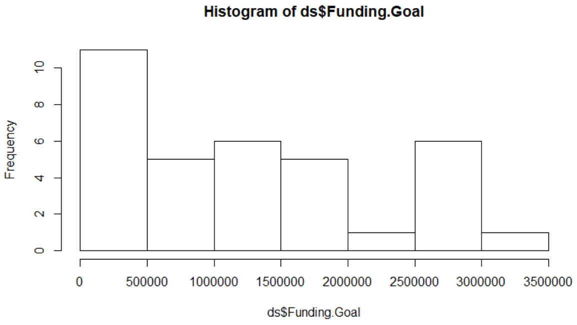


```
                            vars  n        mean          sd     median     trimmed         mad         min          max        range  skew kurtosis          se
Campaign.Duration              1 35       29.97       32.01      29.00       26.38       35.58        0.00       165.00       165.00  1.98     6.39        5.41
Idea.Length                    2 35     1266.29     1032.78     928.00     1104.14      662.72      183.00      4822.00      4639.00  1.56     2.34      174.57
Comments.Count                 3 35       14.77       18.73       7.00       11.38       10.38        0.00        66.00        66.00  1.43     1.14        3.17
Valid.Pitches.Count            4 35        1.00        0.00       1.00        1.00        0.00        1.00         1.00         0.00   NaN      NaN        0.00
Investors.Count                5 35       59.86       67.03      37.00       49.93       50.41        1.00       247.00       246.00  1.21     0.42       11.33
Funding.Goal                   6 35  1315212.71  1046923.26 1250001.00  1271807.97  1186158.58    50000.00   3000015.00   2950015.00  0.41    -1.15   176962.33
Min.Amount                     7 35     6640.66    16741.73    3000.00     3384.07     2075.64     1000.00    100015.00     99015.00  4.98    24.69     2829.87
Minimum.Equity.Offered         8 35        5.28        4.50       3.94        4.68        4.36        0.55        22.68        22.13  1.71     4.09        0.76
Total.Equity.Offered           9 35       10.32        5.51      10.00        9.94        6.35        1.00        25.00        24.00  0.69    -0.12        0.93
Company.Valuation             10 35 19062139.74 34424895.56 9000000.00 12118961.76 11490150.00   575000.00 198000000.00 197425000.00  4.05    17.98  5818869.39
Min.Shares.Issued             11 35    97206.94   343930.22    9444.00    23337.41    13883.07       80.00   2000000.00   1999920.00  4.87    23.83    58134.82
Share.Capital.Before.Funding  12 35  3214483.74 10605857.90  234677.00   610456.41   318280.12      920.00  59000000.00  58999080.00  4.32    19.17  1792717.19
Price.Per.Share               13 35      111.02      166.52      34.00       74.99       48.00        1.00       625.00       624.00  1.93     2.91       28.15
NumberofInfographic           14 35        2.29        5.25       0.00        1.10        0.00        0.00        18.00        18.00  1.89     1.96        0.89
NumberofVideo                 15 35        0.94        0.42       1.00        0.97        0.00        0.00         2.00         2.00 -0.39     2.40        0.07
NumberofEntrepreneur          16 35        0.51        0.78       0.00        0.38        0.00        0.00         3.00         3.00  1.39     1.21        0.13
Employee                      17 35       65.94      233.56       6.00        8.38        4.45        1.00      1000.00       999.00  3.64    11.62       39.46
Revenue                       18 35  1248791.54  2045363.85  369663.00   860632.97   545160.92        0.00   9472579.00   9472579.00  2.27     5.33   345729.59
NetIncome                     19 35   -12450.60   623078.03  -16492.00   -52566.45   172374.49 -2048183.00   1717938.00   3766121.00  0.18     3.51   105319.41
GrowthRate....                20 35       60.60      113.84      20.00       31.79       26.69        0.00       500.00       500.00  2.61     6.19       19.24
Age..years.                   21 35        2.80        3.03       2.00        2.34        1.48        0.00        11.00        11.00  1.21     0.43        0.51
Post.Evaluation               22 13  8737692.31  9766405.56 3000000.00  7542272.73  3521175.00   625000.00  30000000.00  29375000.00  0.84    -0.76  2708713.54
Total.Public.Investor         23 28        4.79        3.79       4.00        4.33        2.97        1.00        17.00        16.00  1.31     1.63        0.72
```

Figure 52: Statistical Summary - numerical variables

Based on the statistical summary above, noted that most variables have skewness value within 2 and beyond -2 except the minimum goal, Company Valuation, Minimum Shares Issued, Share Capital Before Funding, Employee, Revenue and Growth Rate are all slightly skewed towards right indicating a small extreme maximum value. Thus, these variables needs to go through an outlier treatment, as the values are all positively skewed the log transformation would be performed.





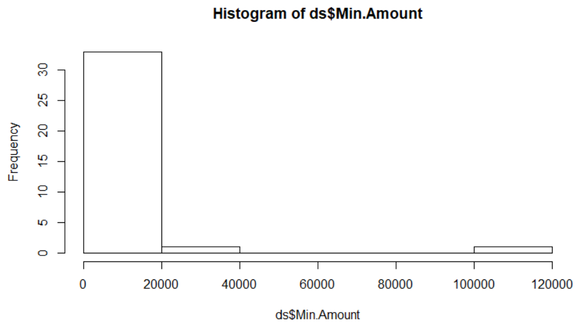


Figure 53: Campaign Duration, Funding Goal and Minimum Goal

For Funding Goal, the minimum a campaign has requested is RM50k and a maximum of RM3 mil has been requested in this platform. On minimum goal amount, there are campaign with the least amount at RM1000 and maximum minimum goal at RM 100k. However, majority campaign has a mean of RM6640 as their minimum goal. In term of campaign duration, there are campaigns that has only campaigned for a day and achieved its funding goal and a maximum of 165 days by certain campaign to achieve its goal.

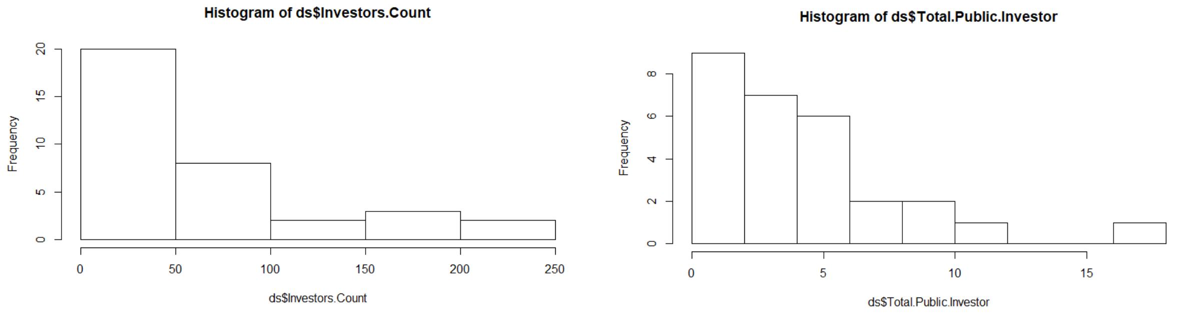


Figure 54: Investor Count and Total Public Investor

As observed above, we noted there are campaign with only 1 investor and maximum of 247 investors in certain campaign. However, mean number of investors in a campaign is around 60 investors. On the number of Public investors, most campaign has around 5 public investors where some campaign had only 1 or there were 1 campaign with 17 public investors.

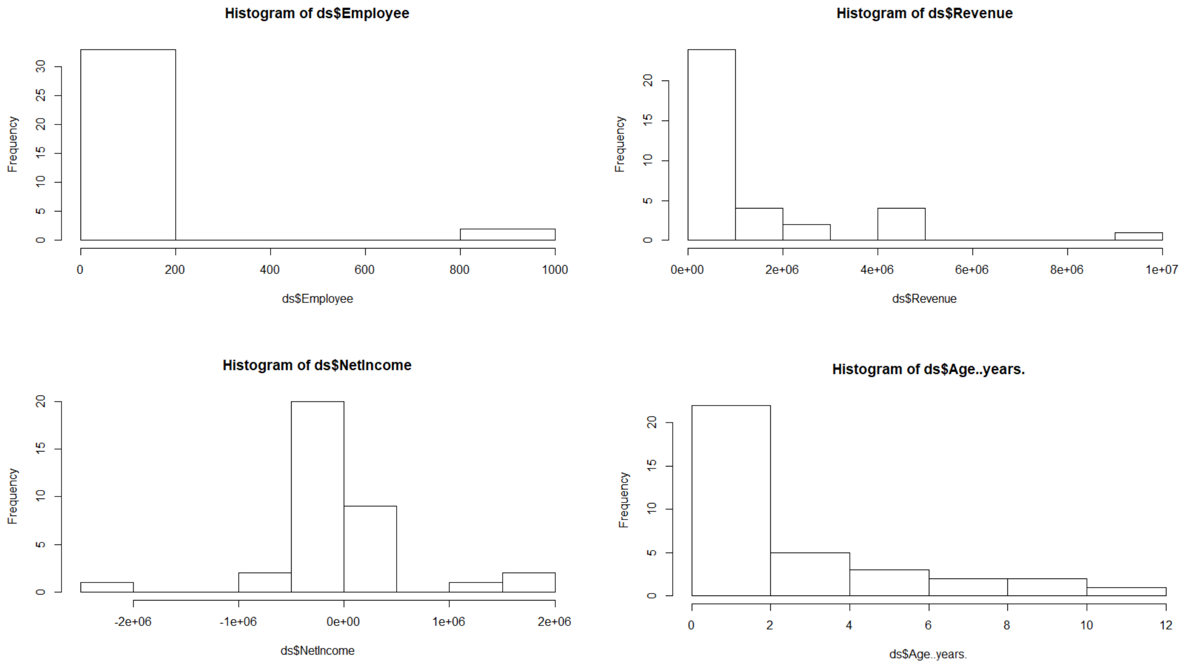


Figure 55: Company Profile – Employee Count, Revenue, Income and Age

From the histogram plot above, noted that around 33 business have within 200 employees with the average having only 66 employees. There were organisation with only 1 employee and 3 organisations having between 800 to 1000 employees. In term of Revenue, there were organisation that are new with no income generated yet and there were organisation whom have generated close to RM10 mil. On the other hand, looking at the net income, there is around 24 organisation whom are having negative net income where average net income at loss of RM 12k. There is 1 organisation that is having loss of close to RM2 mil and there is 3 organisation that has a net income of close to RM200k. Majority organisation has a negative net income of within RM50k. Also, the organisation's number of years established was observed, around 23 organisations have established for the last 2 years where there were certain organisation that are less than a year old and an organisation that is 11 years old. Further exploration of the dataset was performed to observe the dependencies between variables.

4.3.2.4 Multicollinearity Identification

The variables was viewed to identify if there is any strongly correlated variables to be discarded as part of the modelling, Figure 56.

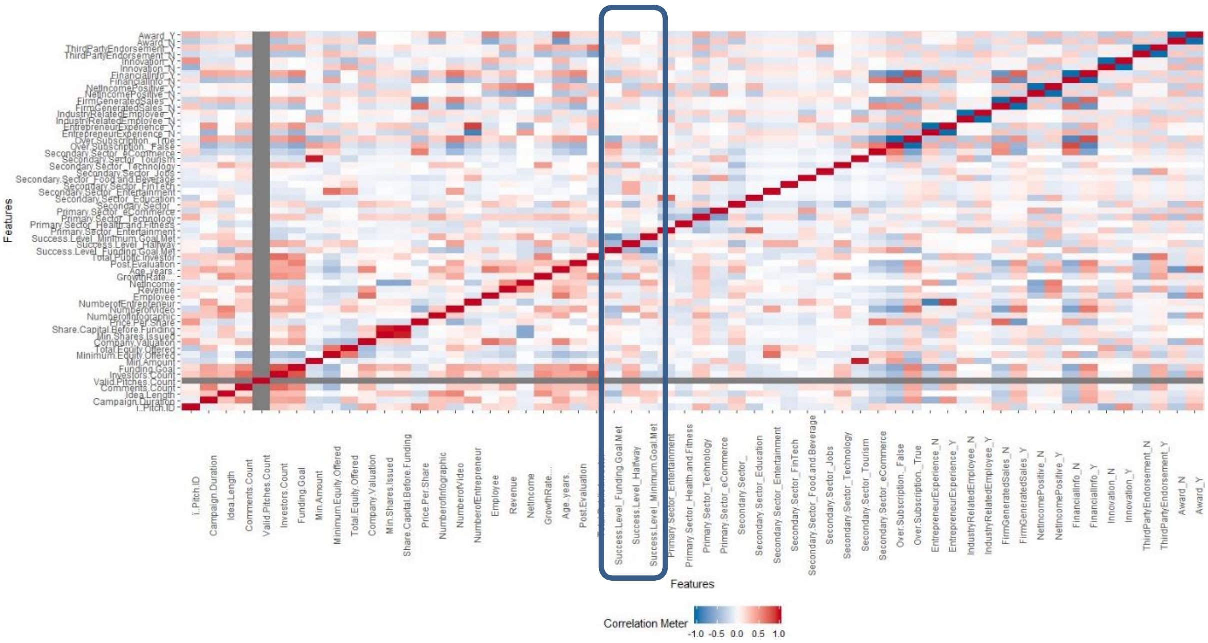


Figure 56: Multicollinearity Plot

Based on the plot above noted, there is no very strong relationship between variables to the dependent variable success levels. However, some moderate positively correlated independent variable Funding Goal, Investor Count and Comments Count towards success level halfway but negatively correlated to minimum goal met and funding goal met. Idea Length are positively correlated to halfway and minimum goal met. For minimum goal met, there are attribute campaign duration, number of infographics, net income not positive and age noted to have a moderate correlation with this target. Funding Goal Met on the other hand has these attributes that are somewhat correlated compared to the other 2 levels, minimum amount, price per share, employee, secondary sector jobs, tourism or ecommerce, false oversubscription and no financial information provided.

The strong correlation between independent variables were also observed to identify any multicollinearity issue, following variables to be highly correlated among themselves:

- Secondary Sector Tourism and Minimum Amount
- Share Capital Before Funding and Min Share Issued
- Number of Entrepreneur and Entrepreneur experience = Yes

This is an interesting correlation between secondary tourism and minimum amount, thus we would leave this attribute to observe the decision made. Share Capital Before Funding and

Minimum Shares issued, however these attributes are not strongly correlated to the dependent variables, thus would be ignored. Similar, the number of entrepreneur and experience has minimum correlation with the dependent variable thus further action required.

4.3.3 Exploratory Data Analysis

1. Target Variable – Success Level

The success level is derived from the total amount collected per campaign. Each campaign has a minimum goal met and maximum funding goal met. Additionally, halfway is campaign whom have collected between halfway to almost reaching the funding goal. Figure 57 highlights the distribution of the campaign based on the success level.

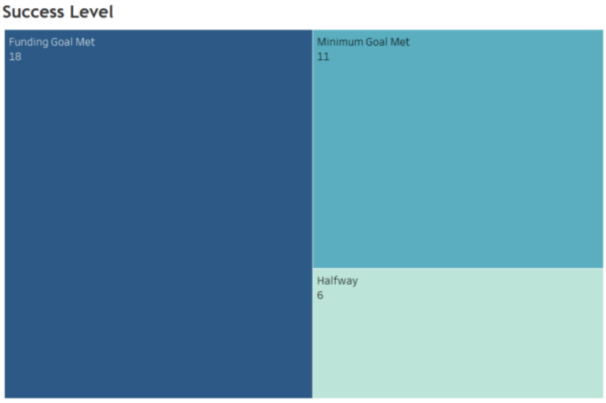


Figure 57: Pitch Success Level

18 campaigns have met its funding goal followed by 11 campaigns whom have met the minimum goal and 6 campaigns came halfway. Here, noted that there is class imbalance issue where the funding goal met has the highest distribution thus halfway may not achieve high accuracy due to the dominant of Funding goal met. Thus, the SMOTE technique to create a balanced dataset would be performed.

2. Minimum goal and Funding Goal by Success Level

Then, further exploration of the success level to the funding goal and the minimum goal is to identify which of those patterns where maximum goal have ridiculously set thus unable to meet the funding goal, Figure 58.

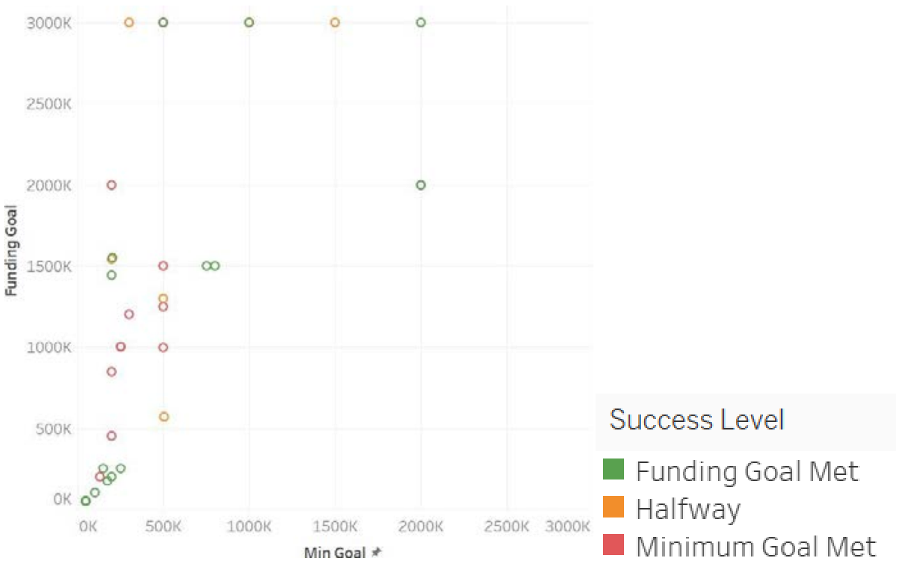


Figure 58: Minimum Goal and Funding Goal by Success Level

Here, noted those with very high funding goal has surprisingly met its funding goal except 2 which at least met halfway. Majority of those whom just met the minimum goal had a huge difference between the minimum to the maximum (2 to 4 times more than the minimum goal). Due such huge variance in the minimum and maximum goal, this provides us an insight of a young start-ups with uncertainty of their idea. However, noted campaigns having 6 times funding goal from the minimum goal, yet successfully meeting the funding goal.

3. Campaigns by Primary Sector

The primary sector was reviewed to identify the most popular sector and their success level, Figure 59 illustrates the outcome.

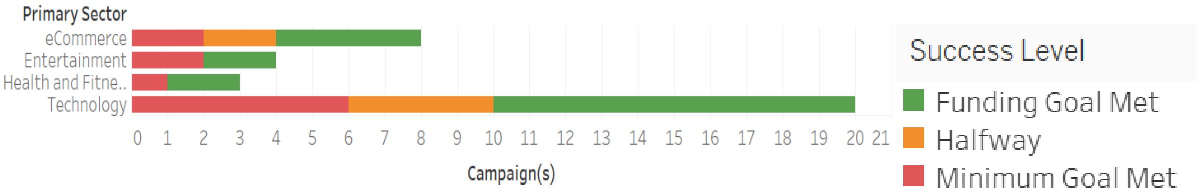


Figure 59: Sectors by success level

Here, noted that there are only 4 main sectors namely eCommerce, Entertainment, Health and Fitness and Technology. Technology sector has the most campaign (20 campaigns) where the same sector has most funding goal met (10 campaigns). Similarly, technology sector too has the most with just minimum funding met (6 campaigns). Other sectors have a balanced between funding goal met and minimum goal. However, here noted that further improvement on the data collection to be performed in order to provide a precise trending.

4. Total Campaign, Success Level and Innovation indicator by duration

The total campaign aggregated by the months was observed. Here, the seasonality element was observed where we noted there were 3 main peak months in April, June and July. In April 63% of the campaigns were successful and 3 of the campaigns only met the minimum fund. In June, despite having 4 campaigns, we noted that only 1 had met the funding goal and 2 have met halfway. This similar effect was observed in July as well.

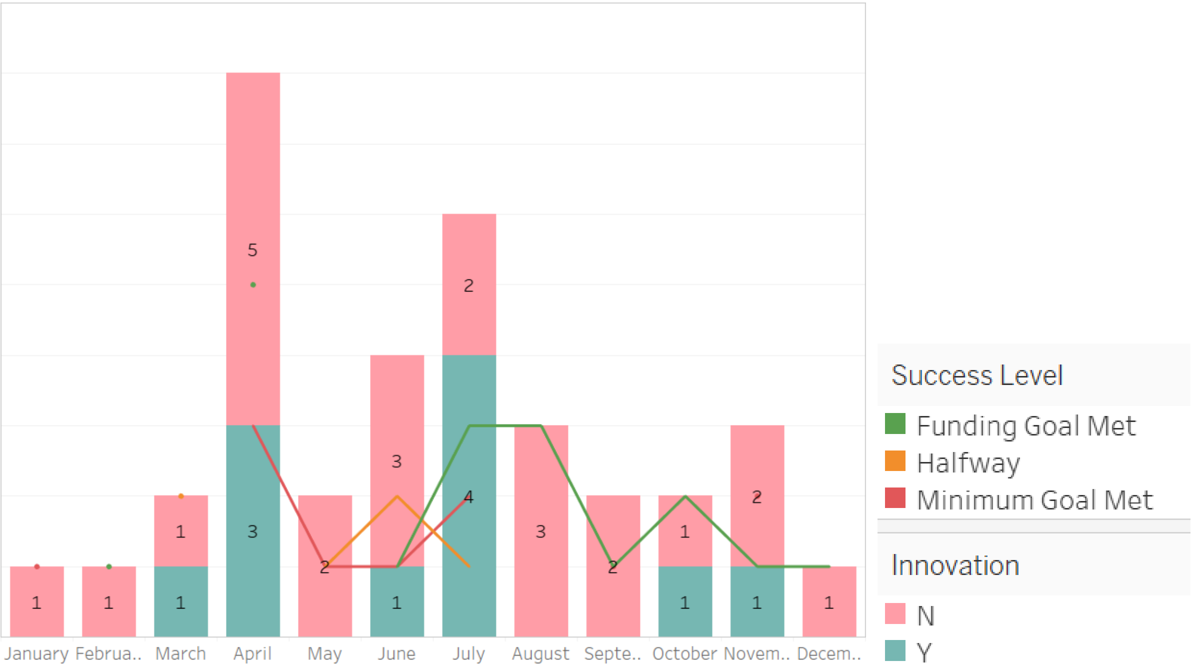


Figure 60: Total Campaign Success Level and Innovation indicator by duration

December through March is deemed to be the low period where minimal campaign was onboarded, however in term of investment by investors remained unchanged. The only duration where we noted high number of halfway and Minimum Goal met campaign was in the month of April, but this could be due to high number of campaigns available in the platform, thus a high competition. Additionally, as we were aware of Innovation being an influencing factor for Investor’s spending, noted that the month of July had more Innovation project thus the highest funding goal met during that period.

5. Funding Goal by Total Received Amount

The funding goal versus amount received was visualised, here we noted that the platform gained traction from February 2017 onwards, where only RM2 mil was requested in total, however the platform managed to collect beyond the funding goal at RM2.5 mil. Then, in June 2019, the new record of RM 7 mil in total of the funding goal set and this in return had collected close to RM6 mil. Here, noted that the platform is gaining traction and more investors are heading towards the platform. However, noted in June 2018 there was a dip in the funding request and similarly the subsequent month we see a drop on investor's investment. This occurred during the political changes that occurred in the country during mid-2018, thus a pause by the investors on their investment. Indirectly an indicator of financial cautiousness practiced by the investors.

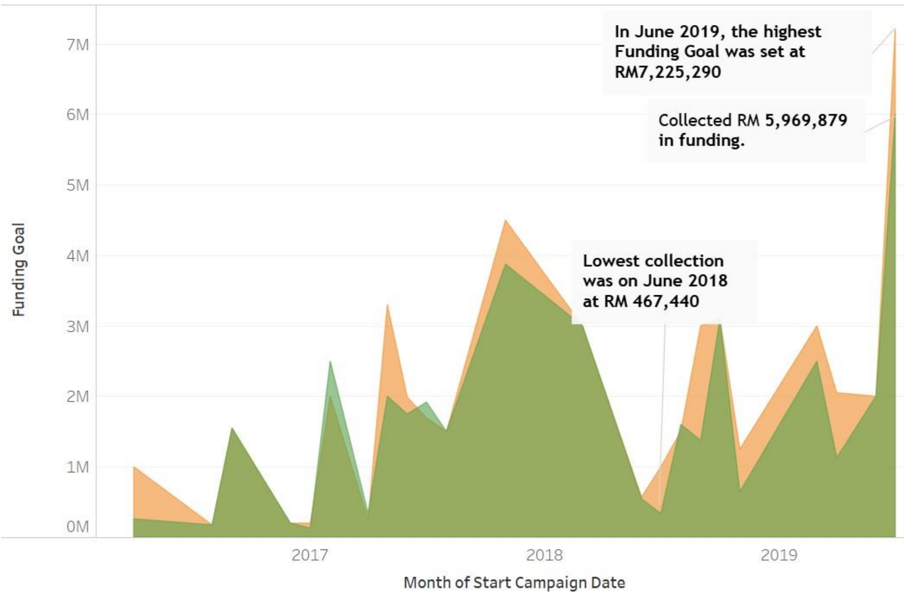


Figure 61: Funding goal versus Total Received

6. Campaign Duration versus Success Rate

As previous studies highlighted on the campaign duration influences the success of a campaign, an observation between both attributes was performed as in Figure 62.

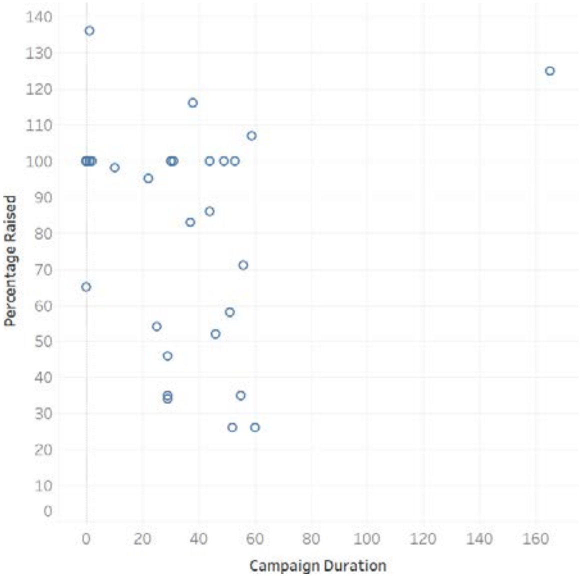


Figure 62: Campaign Duration versus Percentage Raised

Here noted that duration does not have an influence on the percentage raised. A campaign that has close to 160 days also have collected around 125% from its funding goal and a campaign that has between 0 to 60 days too did not managed to attract more. This relates back to a standard fix duration for campaign not practiced in the platform, thus having a standard probably 30 days for a campaign to run, this may assist in certain campaign however not necessarily have more campaign meeting its funding goal.

7. Revenue and Net Income by Success Level

The Revenue and Net Income was observed to identify how it effects the Success Level, Figure 63 illustrates the relationship.

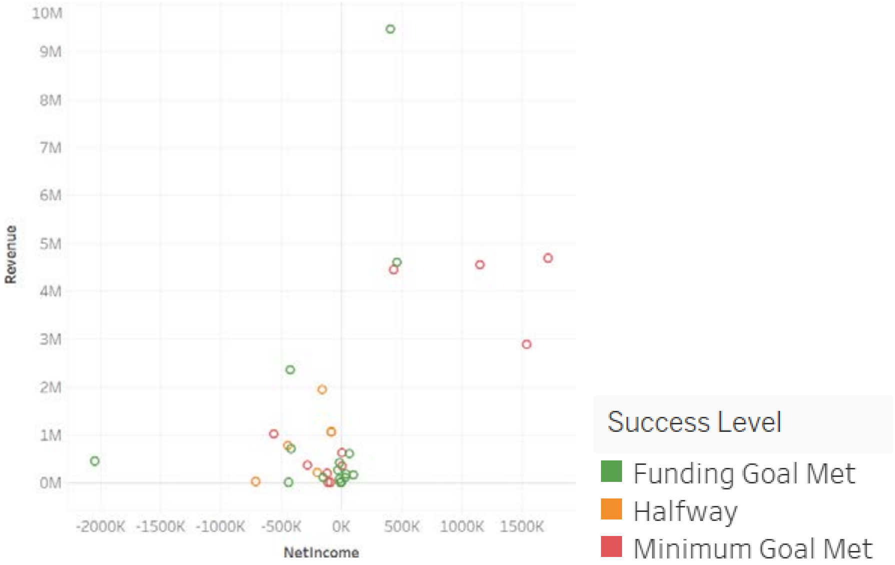


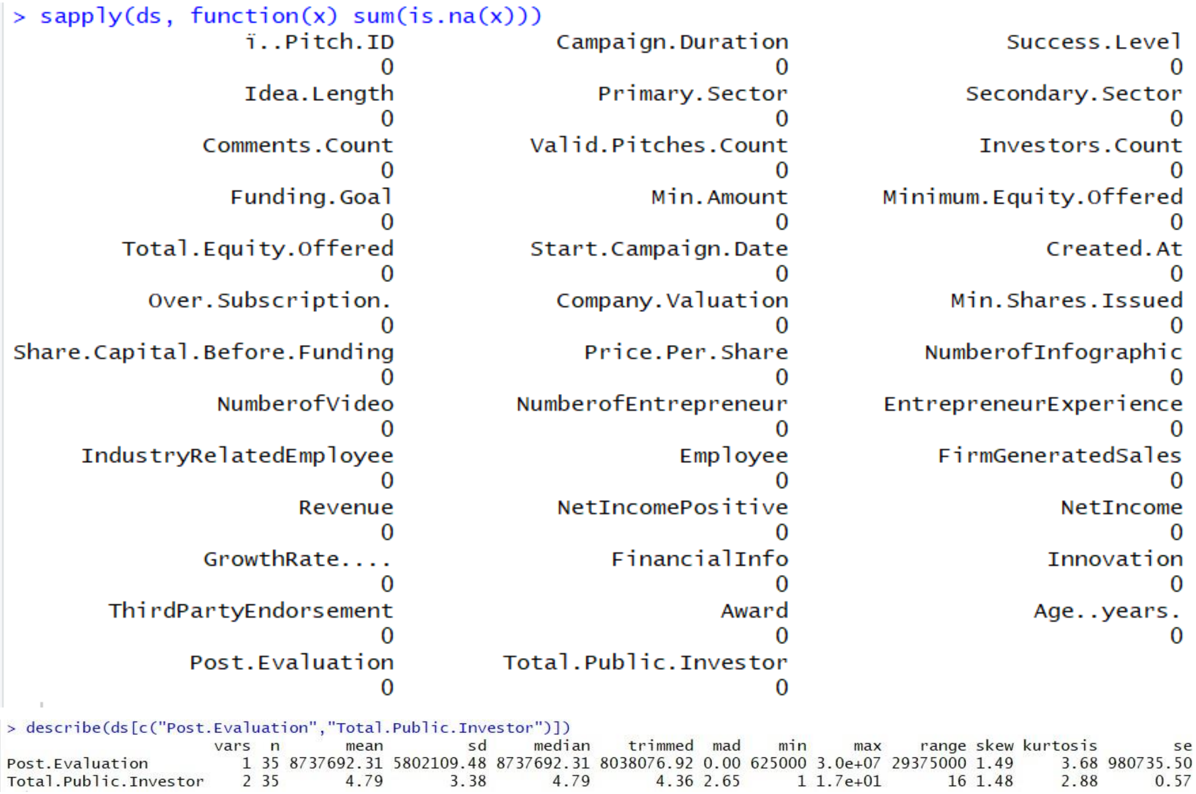
Figure 63: Revenue and Net Income by Success Level

Based on the scatter plot above, observed that despite having a positive higher net income with a reasonable revenue, organisations still could not achieve the funding goal rather the minimum goal is what they could obtain. Additionally, zooming into those campaigns whom has funding goal met, noted that they have a negative net income with minimal revenue achieved to date. Here, we could conclude further that net income and revenue are not heavily looked upon by investors prior to investing. Next, the missing value treatment would be performed together with outlier treatment prior to proceeding with feature selection and predictive modelling.

4.3.4 Data Pre-processing and Transformation

4.3.4.1 Missing Value Treatment

The Post Evaluation as well as Total Public Investor was the two variables identified to have missing value. As both are numerical attributes the mean imputation method would be performed for those attributes, the outcome as illustrated in Figure 64 below.



```
> sapply(ds, function(x) sum(is.na(x)))
                 ï..Pitch.ID          Campaign.Duration              Success.Level
                           0                          0                          0
                 Idea.Length             Primary.Sector           Secondary.Sector
                           0                          0                          0
              Comments.Count        Valid.Pitches.Count            Investors.Count
                           0                          0                          0
                Funding.Goal                 Min.Amount     Minimum.Equity.Offered
                           0                          0                          0
        Total.Equity.Offered       Start.Campaign.Date                 Created.At
                           0                          0                          0
           Over.Subscription.          Company.Valuation           Min.Shares.Issued
                           0                          0                          0
 Share.Capital.Before.Funding            Price.Per.Share        NumberofInfographic
                           0                          0                          0
               NumberofVideo         NumberofEntrepreneur     EntrepreneurExperience
                           0                          0                          0
       IndustryRelatedEmployee                   Employee         FirmGeneratedSales
                           0                          0                          0
                     Revenue          NetIncomePositive                  NetIncome
                           0                          0                          0
               GrowthRate....              FinancialInfo                 Innovation
                           0                          0                          0
        ThirdPartyEndorsement                      Award                Age..years.
                           0                          0                          0
             Post.Evaluation      Total.Public.Investor
                           0                          0
```

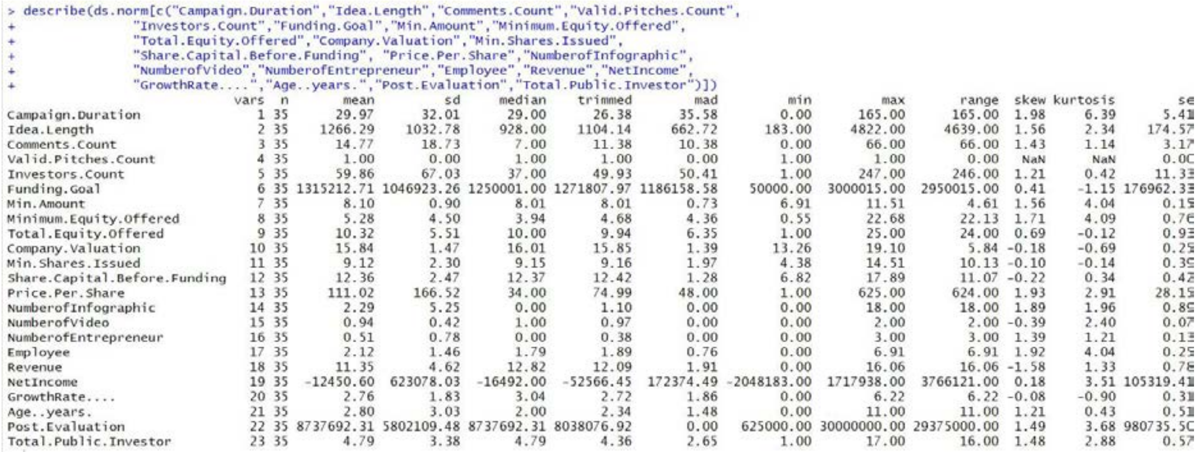
```
> describe(ds[c("Post.Evaluation","Total.Public.Investor")])
                      vars  n       mean         sd     median    trimmed  mad    min     max    range skew kurtosis        se
Post.Evaluation          1 35 8737692.31 5802109.48 8737692.31 8038076.92 0.00 625000 3.0e+07 29375000 1.49     3.68 980735.50
Total.Public.Investor    2 35       4.79       3.38       4.79       4.36 2.65      1 1.7e+01       16 1.48     2.88      0.57
```

Figure 64: Pitch View Missing Value Treatment

Noted the skewness value for the transformed data are within the normal distribution, thus no outlier treatment required.

4.3.4.2 Outlier Treatment

The following attributes was noted of having outliers where the kurtosis value was beyond 3, thus the data would be scaled to obtain a transformation data for model comparison. The variables that requires transformation are minimum goal, Company Valuation, Minimum Shares Issued, Share Capital Before Funding, Employee, Revenue and Growth Rate, the output as Figure 65.



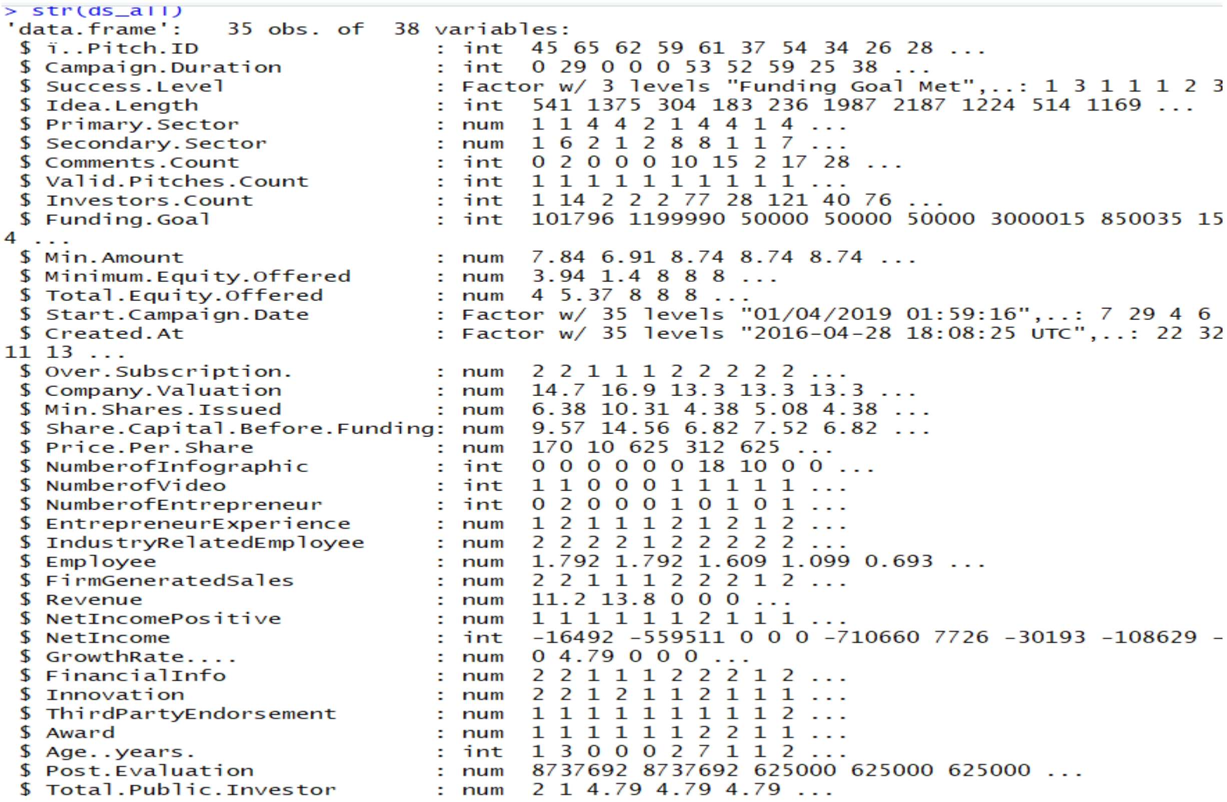
```
> describe(ds.norm[c("Campaign.Duration","Idea.Length","Comments.Count","Valid.Pitches.Count",
+             "Investors.Count","Funding.Goal","Min.Amount","Minimum.Equity.Offered",
+             "Total.Equity.Offered","Company.Valuation","Min.Shares.Issued",
+             "Share.Capital.Before.Funding", "Price.Per.Share","NumberofInfographic",
+             "NumberofVideo","NumberofEntrepreneur","Employee","Revenue","NetIncome",
+             "GrowthRate....","Age..years.","Post.Evaluation","Total.Public.Investor")])
                             vars  n       mean         sd     median    trimmed        mad         min         max       range  skew kurtosis        se
Campaign.Duration               1 35      29.97      32.01      29.00      26.38      35.58        0.00      165.00      165.00  1.98     6.39      5.41
Idea.Length                     2 35    1266.29    1032.78     928.00    1104.14     662.72      183.00     4822.00     4639.00  1.56     2.34    174.57
Comments.Count                  3 35      14.77      18.73       7.00      11.38      10.38        0.00       66.00       66.00  1.43     1.14      3.17
Valid.Pitches.Count             4 35       1.00       0.00       1.00       1.00       0.00        1.00        1.00        0.00   NaN      NaN      0.00
Investors.Count                 5 35      59.86      67.03      37.00      49.93      50.41        1.00      247.00      246.00  1.21     0.42     11.33
Funding.Goal                    6 35 1315212.71 1046923.26 1250001.00 1271807.97 1186158.58    50000.00  3000015.00  2950015.00  0.41    -1.15 176962.33
Min.Amount                      7 35       8.10       0.90       8.01       8.01       0.73        6.91       11.51        4.61  1.56     4.04      0.15
Minimum.Equity.Offered          8 35       5.28       4.50       3.94       4.68       4.36        0.55       22.68       22.13  1.71     4.09      0.76
Total.Equity.Offered            9 35      10.32       5.51      10.00       9.94       6.35        1.00       25.00       24.00  0.69    -0.12      0.93
Company.Valuation              10 35      15.84       1.47      16.01      15.85       1.39       13.26       19.10        5.84 -0.18    -0.69      0.25
Min.Shares.Issued              11 35       9.12       2.30       9.15       9.16       1.97        4.38       14.51       10.13 -0.10    -0.14      0.39
Share.Capital.Before.Funding   12 35      12.36       2.47      12.37      12.42       1.28        6.82       17.89       11.07 -0.22     0.34      0.42
Price.Per.Share                13 35     111.02     166.52      34.00      74.99      48.00        1.00      625.00      624.00  1.93     2.91     28.15
NumberofInfographic            14 35       2.29       5.25       0.00       1.10       0.00        0.00       18.00       18.00  1.89     1.96      0.89
NumberofVideo                  15 35       0.94       0.42       1.00       0.97       0.00        0.00        2.00        2.00 -0.39     2.40      0.07
NumberofEntrepreneur           16 35       0.51       0.78       0.00       0.38       0.00        0.00        3.00        3.00  1.39     1.21      0.13
Employee                       17 35       2.12       1.46       1.79       1.89       0.76        0.00        6.91        6.91  1.92     4.04      0.25
Revenue                        18 35      11.35       4.62      12.82      12.09       1.91        0.00       16.06       16.06 -1.58     1.33      0.78
NetIncome                      19 35  -12450.60  623078.03  -16492.00  -52566.45  172374.49 -2048183.00  1717938.00  3766121.00  0.18     3.51 105319.41
GrowthRate....                 20 35       2.76       1.83       3.04       2.72       1.86        0.00        6.22        6.22 -0.08    -0.90      0.31
Age..years.                    21 35       2.80       3.03       2.00       2.34       1.48        0.00       11.00       11.00  1.21     0.43      0.51
Post.Evaluation                22 35 8737692.31 5802109.48 8737692.31 8038076.92       0.00   625000.00 30000000.00 29375000.00  1.49     3.68 980735.50
Total.Public.Investor          23 35       4.79       3.38       4.79       4.36       2.65        1.00       17.00       16.00  1.48     2.88      0.57
```

Figure 65: Summary Statistics after log transformation

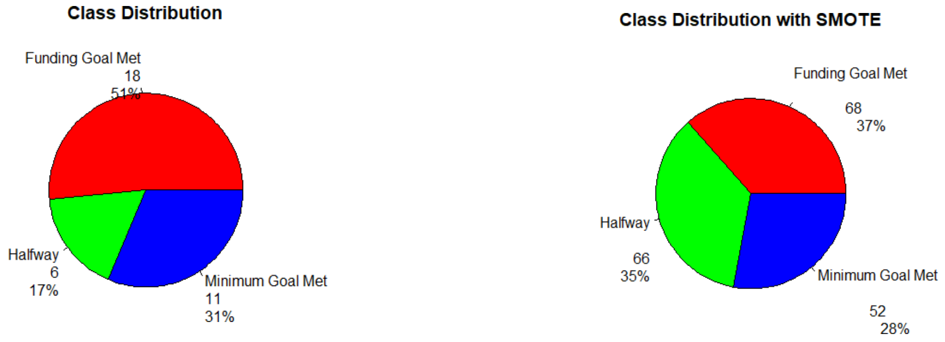
Based on the output above, all attributes had been successfully transformed where the skewness value is beyond -2 and within 2. Next, as the dataset has been cleaned transformed, the feature selection phase would occur to identify the best features to be utilised for the predictive modelling.

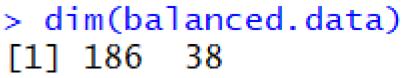
4.3.4.3 SMOTE

As during exploration, we noted the need for SMOTE technique to treat the class imbalanced, this dataset would undergo a SMOTE transformed. Figure 66 highlights the pre and post SMOTE transformation dataset, however prior to using SMOTE package, all factor data would be to be transformed to numeric except the target variable Success Level.



```
> str(ds_all)
'data.frame':	35 obs. of  38 variables:
 $ ï..Pitch.ID                  : int  45 65 62 59 61 37 54 34 26 28 ...
 $ Campaign.Duration            : int  0 29 0 0 0 53 52 59 25 38 ...
 $ Success.Level                : Factor w/ 3 levels "Funding Goal Met",..: 1 3 1 1 1 2 3
 $ Idea.Length                  : int  541 1375 304 183 236 1987 2187 1224 514 1169 ...
 $ Primary.Sector               : num  1 1 4 4 2 1 4 4 1 4 ...
 $ Secondary.Sector             : num  1 6 2 1 2 8 8 1 1 7 ...
 $ Comments.Count               : int  0 2 0 0 0 10 15 2 17 28 ...
 $ Valid.Pitches.Count          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Investors.Count              : int  1 14 2 2 2 77 28 121 40 76 ...
 $ Funding.Goal                 : int  101796 1199990 50000 50000 50000 3000015 850035 15
4 ...
 $ Min.Amount                   : num  7.84 6.91 8.74 8.74 8.74 ...
 $ Minimum.Equity.Offered       : num  3.94 1.4 8 8 8 ...
 $ Total.Equity.Offered         : num  4 5.37 8 8 8 ...
 $ Start.Campaign.Date          : Factor w/ 35 levels "01/04/2019 01:59:16",..: 7 29 4 6
 $ Created.At                   : Factor w/ 35 levels "2016-04-28 18:08:25 UTC",..: 22 32
11 13 ...
 $ Over.Subscription.           : num  2 2 1 1 1 2 2 2 2 2 ...
 $ Company.Valuation            : num  14.7 16.9 13.3 13.3 13.3 ...
 $ Min.Shares.Issued            : num  6.38 10.31 4.38 5.08 4.38 ...
 $ Share.Capital.Before.Funding: num  9.57 14.56 6.82 7.52 6.82 ...
 $ Price.Per.Share              : num  170 10 625 312 625 ...
 $ NumberofInfographic          : int  0 0 0 0 0 0 18 10 0 0 ...
 $ NumberofVideo                : int  1 1 0 0 0 1 1 1 1 1 ...
 $ NumberofEntrepreneur         : int  0 2 0 0 0 1 0 1 0 1 ...
 $ EntrepreneurExperience       : num  1 2 1 1 1 2 1 2 1 2 ...
 $ IndustryRelatedEmployee      : num  2 2 2 2 1 2 2 2 2 2 ...
 $ Employee                     : num  1.792 1.792 1.609 1.099 0.693 ...
 $ FirmGeneratedSales           : num  2 2 1 1 1 2 2 2 1 2 ...
 $ Revenue                      : num  11.2 13.8 0 0 0 ...
 $ NetIncomePositive            : num  1 1 1 1 1 1 2 1 1 1 ...
 $ NetIncome                    : int  -16492 -559511 0 0 0 -710660 7726 -30193 -108629 -
 $ GrowthRate....               : num  0 4.79 0 0 0 ...
 $ FinancialInfo                : num  2 2 1 1 1 2 2 2 1 2 ...
 $ Innovation                   : num  2 2 1 2 1 1 2 1 1 1 ...
 $ ThirdPartyEndorsement        : num  1 1 1 1 1 1 1 1 1 2 ...
 $ Award                        : num  1 1 1 1 1 1 2 2 1 1 ...
 $ Age..years.                  : int  1 3 0 0 0 2 7 1 1 2 ...
 $ Post.Evaluation              : num  8737692 8737692 625000 625000 625000 ...
 $ Total.Public.Investor        : num  2 1 4.79 4.79 4.79 ...
```





```
> dim(balanced.data)
[1] 186  38
```

Figure 66: Before and After SMOTE Transformation

After the SMOTE transformation, the Funding Goal Met was brought to 37% followed by Halfway at 66% and Minimum Goal Met at 52% with total observations now at 186. There is a balance between the success level thus, we would proceed to use this balanced dataset as part of the predictive modelling.

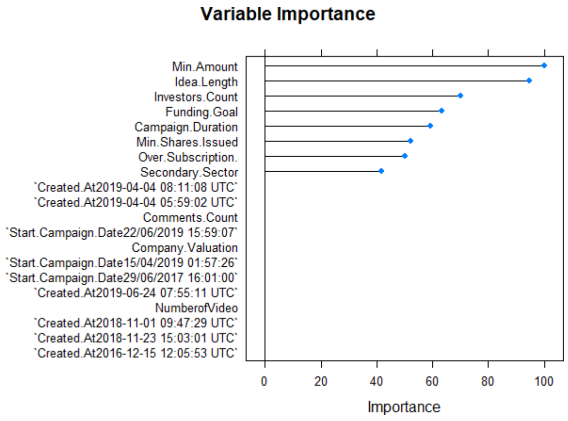
4.3.5 Feature Selection and Dimensionality Reduction

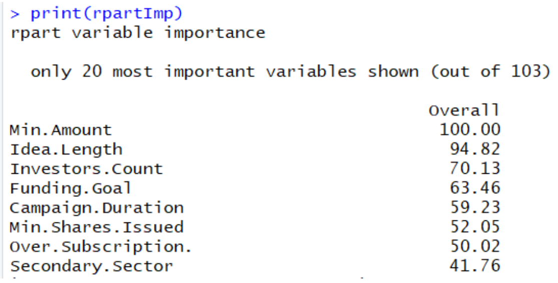
In this section, we would explore two different feature engineering techniques one is the feature selection using Tree Based Algorithm and Boruta to identify the best features. Additionally, the dimensionality reduction technique Factor Analysis would be explored to identify the factorable attributes to be utilised for the modelling.

4.3.5.1 Tree Based Algorithm

The tree-based algorithm uses the Decision Tree technique to identify the variable of importance, following Figure 67.

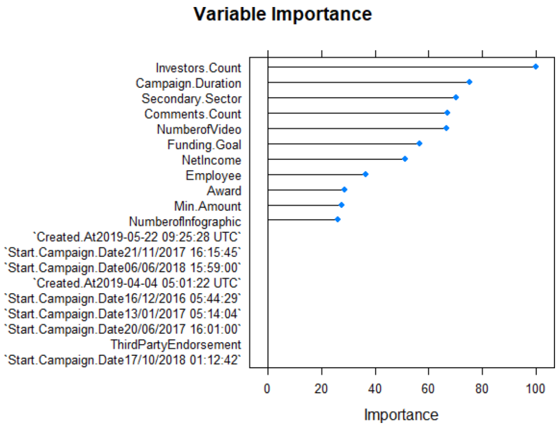
Original

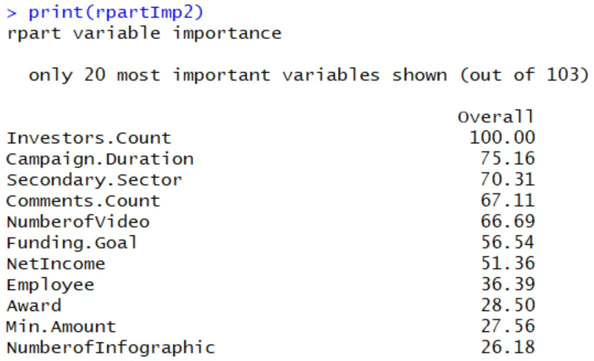




```
> print(rpartImp)
rpart variable importance

  only 20 most important variables shown (out of 103)

                                               Overall
Min.Amount                                      100.00
Idea.Length                                      94.82
Investors.Count                                  70.13
Funding.Goal                                     63.46
Campaign.Duration                                59.23
Min.Shares.Issued                                52.05
Over.Subscription.                               50.02
Secondary.Sector                                 41.76
```

SMOTE





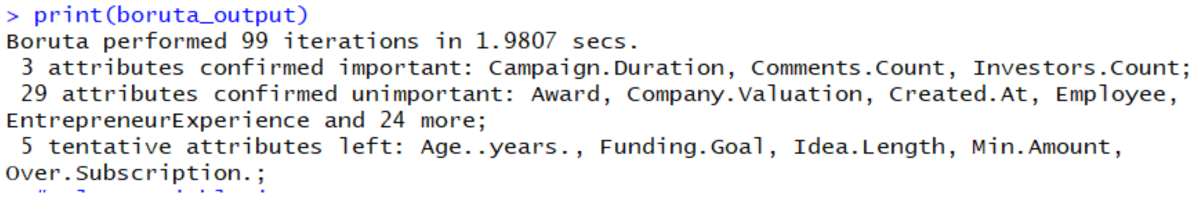
```
> print(rpartImp2)
rpart variable importance

  only 20 most important variables shown (out of 103)

                                               Overall
Investors.Count                                 100.00
Campaign.Duration                                75.16
Secondary.Sector                                 70.31
Comments.Count                                   67.11
NumberofVideo                                    66.69
Funding.Goal                                     56.54
NetIncome                                        51.36
Employee                                         36.39
Award                                            28.50
Min.Amount                                       27.56
NumberofInfographic                              26.18
```

Figure 67: Feature Selection: Tree Based Algorithm

Based on this technique, Minimum Amount Goal, Idea Length, Investor Count, Funding Goal, Campaign Duration, Minimum Shares Issues, Over Subscription and Secondary Sector are all the variables that are importance and has the capability to distinguish the success level from the original dataset. Where, with the SMOTE dataset, additional attributes such as comments count, number of videos, net income, employee, award and number of infographics was added as important variable. Idea Length, Over Subscription and Minimum Shares Issues were identified as not important with SMOTE dataset. This would be used for our predictive modelling experiment.

4.3.5.2 Boruta

The Boruta similarly is built around the random forest classification algorithm to identify the variable of importance, following Figure 68.



```
> print(boruta_output)
Boruta performed 99 iterations in 1.9807 secs.
 3 attributes confirmed important: Campaign.Duration, Comments.Count, Investors.Count;
 29 attributes confirmed unimportant: Award, Company.Valuation, Created.At, Employee,
EntrepreneurExperience and 24 more;
 5 tentative attributes left: Age..years., Funding.Goal, Idea.Length, Min.Amount,
Over.Subscription.;
```

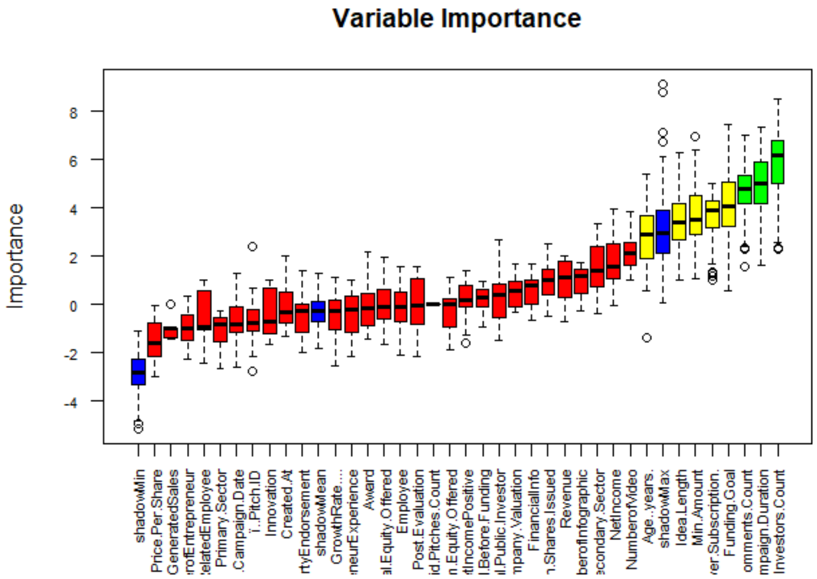
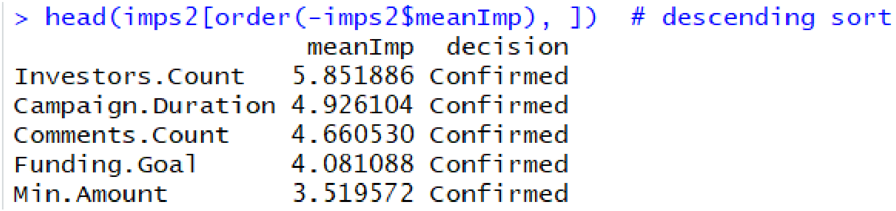


Figure 68: Feature Selection: Boruta

Based on this technique, three variables Campaign Duration, Comments Count and Investor Count has been identified as important variable followed by 5 variables that is tentative Age, Funding Goal, Idea Length, Minimum Amount and Over Subscription. 29 other variables had been determined unimportant. Compared to the previous technique, Age and Comments count are new additions but secondary sector and minimum share ideas are deemed unimportant by this technique. For the tentative variables, additional steps to re-run those variables to re-classify them was performed, the final feature selection output was a below.



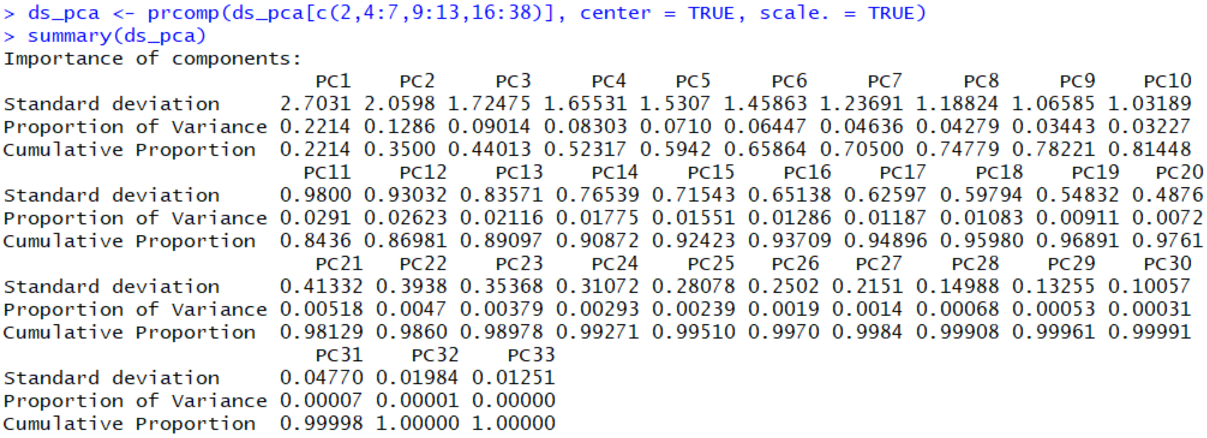
```
> head(imps2[order(-imps2$meanImp), ])  # descending sort
                   meanImp  decision
Investors.Count   5.851886 Confirmed
Campaign.Duration 4.926104 Confirmed
Comments.Count    4.660530 Confirmed
Funding.Goal      4.081088 Confirmed
Min.Amount        3.519572 Confirmed
```

Figure 69: : Feature Selection: Boruta Final

The final output has only 5 variables selected as important excluding the previously tentative variables Age, Idea Length and Over subscription. This technique was performed over SMOTE dataset where all variables except Valid Pitch Count was identified as important. Thus, no further exploration with SMOTE dataset for this technique was performed. This would be used for our predictive modelling experiment.

4.3.5.3 Factor Analysis

As the previous two techniques identifies the most important variable, other unimportant variables would be discarded during modelling. However, this technique does not eliminate the features rather the attributes are grouped into similar principal components. These factors would then be used to predict the success level. Firstly, the original cleaned data without transformation would be used by centring and scaling the data for this activity. Following Figure 70 the Factor Analysis Output.



```
> ds_pca <- prcomp(ds_pca[c(2,4:7,9:13,16:38)], center = TRUE, scale. = TRUE)
> summary(ds_pca)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.7031 2.0598 1.72475 1.65531 1.5307 1.45863 1.23691 1.18824 1.06585 1.03189
Proportion of Variance 0.2214 0.1286 0.09014 0.08303 0.0710 0.06447 0.04636 0.04279 0.03443 0.03227
Cumulative Proportion  0.2214 0.3500 0.44013 0.52317 0.5942 0.65864 0.70500 0.74779 0.78221 0.81448
                         PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19   PC20
Standard deviation     0.9800 0.93032 0.83571 0.76539 0.71543 0.65138 0.62597 0.59794 0.54832 0.4876
Proportion of Variance 0.0291 0.02623 0.02116 0.01775 0.01551 0.01286 0.01187 0.01083 0.00911 0.0072
Cumulative Proportion  0.8436 0.86981 0.89097 0.90872 0.92423 0.93709 0.94896 0.95980 0.96891 0.9761
                          PC21   PC22    PC23    PC24    PC25   PC26   PC27    PC28    PC29    PC30
Standard deviation     0.41332 0.3938 0.35368 0.31072 0.28078 0.2502 0.2151 0.14988 0.13255 0.10057
Proportion of Variance 0.00518 0.0047 0.00379 0.00293 0.00239 0.0019 0.0014 0.00068 0.00053 0.00031
Cumulative Proportion  0.98129 0.9860 0.98978 0.99271 0.99510 0.9970 0.9984 0.99908 0.99961 0.99991
                          PC31    PC32    PC33
Standard deviation     0.04770 0.01984 0.01251
Proportion of Variance 0.00007 0.00001 0.00000
Cumulative Proportion  0.99998 1.00000 1.00000
```

Figure 70: Dimension Reduction: Factor Analysis

Based on the output, 33 principal components (PC1-33) was derived from this dataset where the proportion variance explains the individual variance and the cumulative proportion is the accumulated variance. PC1 explains around 22% of the total variance meaning one fourth of the information in this dataset could be explained by PC1 followed by PC2 explains around 13% of the information of this dataset. As eigenvalue would be used to draw the boundary, eigen value below 1 would be discarded and the those above would be the total principal component required for this dataset, Figure 71 is the scree plot and the cumulative plot.

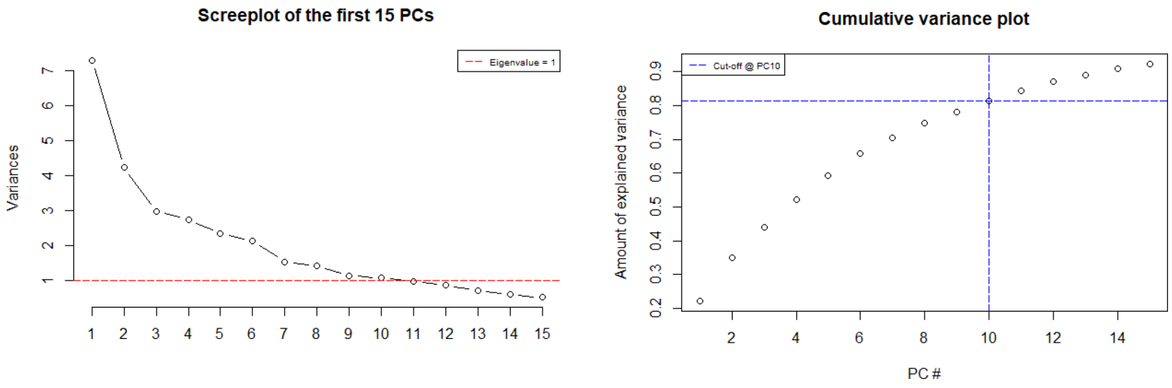


Figure 71: Scree plot and Cumulative Variance Plot

Here, noted that the first 10 principal component has eigenvalue beyond 1 which explains 81% of variance, thus this reduces the dimension from 33 to 10 while losing only 19% of information. Next, to observe individual principal components and the most loading variable for that value, the biplot would be populated where scale would be set to zero to represent the loading, Figure 72.

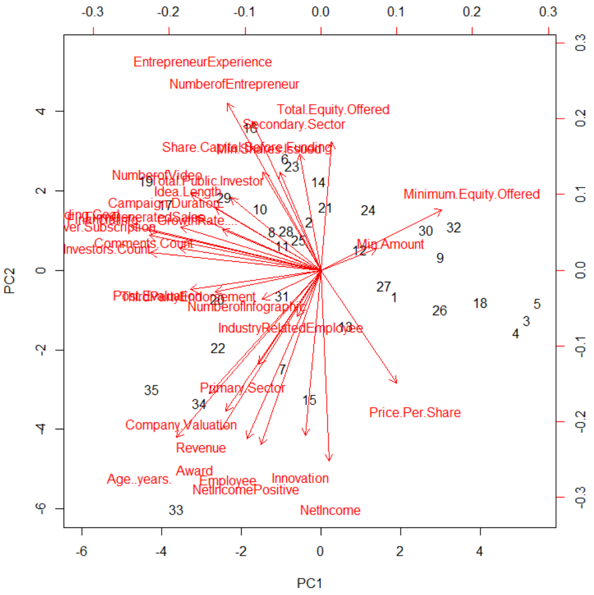


Figure 72: PC1 and PC2

Based on the variable with the heavy loading in PC1 is Minimum Equity Offered and Price Per Share indicating an equity related grouping in PC1. PC2 has Entrepreneur Experience and Number of Entrepreneur as the highest loading for PC2 indicating this principal component to be an Entrepreneurial element. All the ten principal components as elaborated in Figure 73.

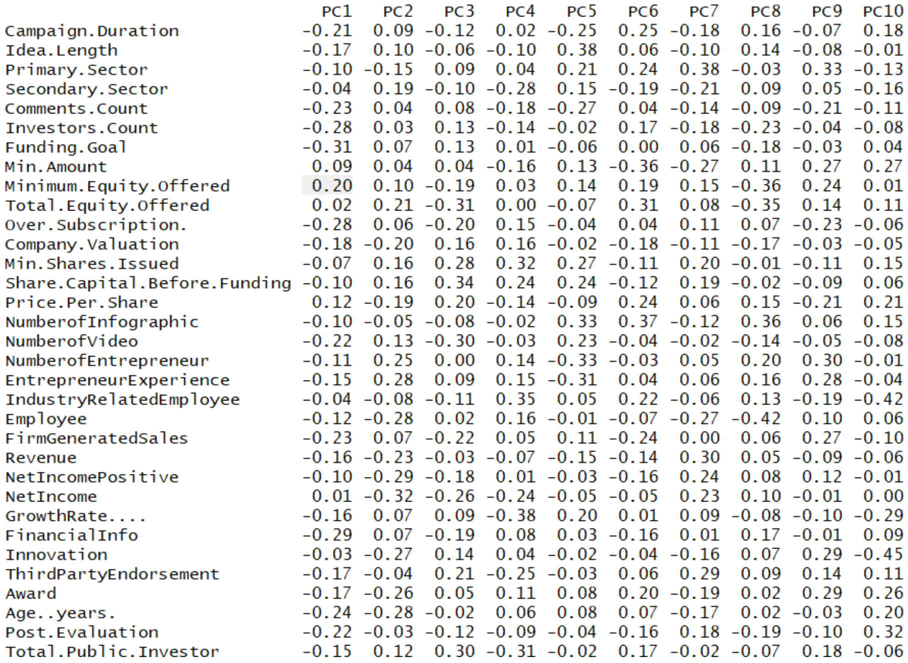

```
                               PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
Campaign.Duration            -0.21  0.09 -0.12  0.02 -0.25  0.25 -0.18  0.16 -0.07  0.18
Idea.Length                  -0.17  0.10 -0.06 -0.10  0.38  0.06 -0.10  0.14 -0.08 -0.01
Primary.Sector               -0.10 -0.15  0.09  0.04  0.21  0.24  0.38 -0.03  0.33 -0.13
Secondary.Sector             -0.04  0.19 -0.10 -0.28  0.15 -0.19 -0.21  0.09  0.05 -0.16
Comments.Count               -0.23  0.04  0.08 -0.18 -0.27  0.04 -0.14 -0.09 -0.21 -0.11
Investors.Count              -0.28  0.03  0.13 -0.14 -0.02  0.17 -0.18 -0.23 -0.04 -0.08
Funding.Goal                 -0.31  0.07  0.13  0.01 -0.06  0.00  0.06 -0.18 -0.03  0.04
Min.Amount                    0.09  0.04  0.04 -0.16  0.13 -0.36 -0.27  0.11  0.27  0.27
Minimum.Equity.Offered        0.20  0.10 -0.19  0.03  0.14  0.19  0.15 -0.36  0.24  0.01
Total.Equity.Offered          0.02  0.21 -0.31  0.00 -0.07  0.31  0.08 -0.35  0.14  0.11
Over.Subscription.           -0.28  0.06 -0.20  0.15 -0.04  0.04  0.11  0.07 -0.23 -0.06
Company.Valuation            -0.18 -0.20  0.16  0.16 -0.02 -0.18 -0.11 -0.17 -0.03 -0.05
Min.Shares.Issued            -0.07  0.16  0.28  0.32  0.27 -0.11  0.20 -0.01 -0.11  0.15
Share.Capital.Before.Funding -0.10  0.16  0.34  0.24  0.24 -0.12  0.19 -0.02 -0.09  0.06
Price.Per.Share               0.12 -0.19  0.20 -0.14 -0.09  0.24  0.06  0.15 -0.21  0.21
NumberofInfographic          -0.10 -0.05 -0.08 -0.02  0.33  0.37 -0.12  0.36  0.06  0.15
NumberofVideo                -0.22  0.13 -0.30 -0.03  0.23 -0.04 -0.02 -0.14 -0.05 -0.08
NumberofEntrepreneur         -0.11  0.25  0.00  0.14 -0.33 -0.03  0.05  0.20  0.30 -0.01
EntrepreneurExperience       -0.15  0.28  0.09  0.15 -0.31  0.04  0.06  0.16  0.28 -0.04
IndustryRelatedEmployee      -0.04 -0.08 -0.11  0.35  0.05  0.22 -0.06  0.13 -0.19 -0.42
Employee                     -0.12 -0.28  0.02  0.16 -0.01 -0.07 -0.27 -0.42  0.10  0.06
FirmGeneratedSales           -0.23  0.07 -0.22  0.05  0.11 -0.24  0.00  0.06  0.27 -0.10
Revenue                      -0.16 -0.23 -0.03 -0.07 -0.15 -0.14  0.30  0.05 -0.09 -0.06
NetIncomePositive            -0.10 -0.29 -0.18  0.01 -0.03 -0.16  0.24  0.08  0.12 -0.01
NetIncome                     0.01 -0.32 -0.26 -0.24 -0.05 -0.05  0.23  0.10 -0.01  0.00
GrowthRate....               -0.16  0.07  0.09 -0.38  0.20  0.01  0.09 -0.08 -0.10 -0.29
FinancialInfo                -0.29  0.07 -0.19  0.08  0.03 -0.16  0.01  0.17 -0.01  0.09
Innovation                   -0.03 -0.27  0.14  0.04 -0.02 -0.04 -0.16  0.07  0.29 -0.45
ThirdPartyEndorsement        -0.17 -0.04  0.21 -0.25 -0.03  0.06  0.29  0.09  0.14  0.11
Award                        -0.17 -0.26  0.05  0.11  0.08  0.20 -0.19  0.02  0.29  0.26
Age..years.                  -0.24 -0.28 -0.02  0.06  0.08  0.07 -0.17  0.02 -0.03  0.20
Post.Evaluation              -0.22 -0.03 -0.12 -0.09 -0.04 -0.16  0.18 -0.19 -0.10  0.32
Total.Public.Investor        -0.15  0.12  0.30 -0.31 -0.02  0.17 -0.02 -0.07  0.18 -0.06
```

Figure 73: Loading for 10 Principal Components

Finally, a new dataset with the component value was generated to be used for one of the predictive modelling testing. This was performed for both the original and SMOTE dataset.

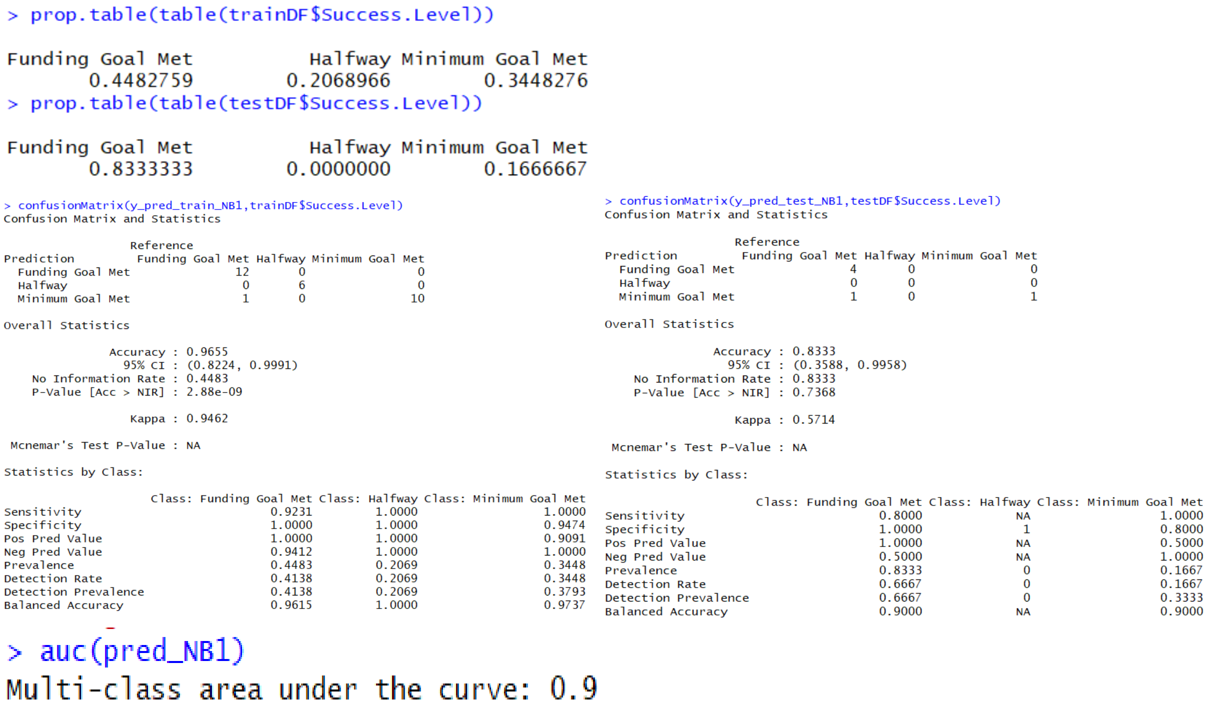
4.3.6 Construction of Model and Interpretation

Now the construction of the predictive models would be performed, the experiments as highlighted in Table 10. Our first experiment would be on all features using random sampling for dataset splitting.

4.3.6.1 Naïve Bayes

Experiment 1: Random sampling with all features and no normalisation performed

The original cleaned dataset with no normalisation performed was used. All features would be used to run this model using random sampling technique for the train test splitting.



```
> prop.table(table(trainDF$Success.Level))

Funding Goal Met            Halfway Minimum Goal Met 
       0.4482759          0.2068966        0.3448276 
> prop.table(table(testDF$Success.Level))

Funding Goal Met            Halfway Minimum Goal Met 
       0.8333333          0.0000000        0.1666667 

> confusionMatrix(y_pred_train_NB1,trainDF$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                12       0                0
  Halfway                          0       6                0
  Minimum Goal Met                 1       0               10

Overall Statistics
                                          
               Accuracy : 0.9655          
                 95% CI : (0.8224, 0.9991)
    No Information Rate : 0.4483          
    P-Value [Acc > NIR] : 2.88e-09        
                                          
                  Kappa : 0.9462          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.9231         1.0000                  1.0000
Specificity                           1.0000         1.0000                  0.9474
Pos Pred Value                        1.0000         1.0000                  0.9091
Neg Pred Value                        0.9412         1.0000                  1.0000
Prevalence                            0.4483         0.2069                  0.3448
Detection Rate                        0.4138         0.2069                  0.3448
Detection Prevalence                  0.4138         0.2069                  0.3793
Balanced Accuracy                     0.9615         1.0000                  0.9737

> auc(pred_NB1)
Multi-class area under the curve: 0.9
```

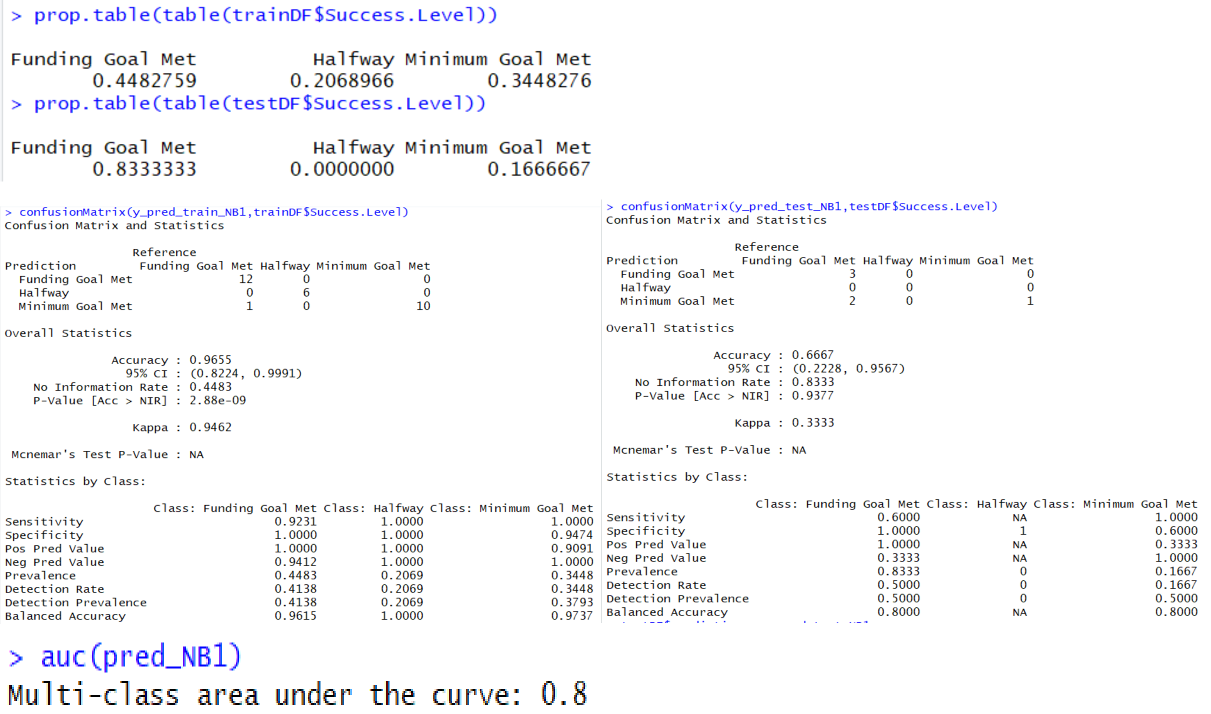
```
> confusionMatrix(y_pred_test_NB1,testDF$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                 4       0                0
  Halfway                          0       0                0
  Minimum Goal Met                 1       0                1

Overall Statistics
                                          
               Accuracy : 0.8333          
                 95% CI : (0.3588, 0.9958)
    No Information Rate : 0.8333          
    P-Value [Acc > NIR] : 0.7368          
                                          
                  Kappa : 0.5714          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.8000             NA                  1.0000
Specificity                           1.0000              1                  0.8000
Pos Pred Value                        1.0000             NA                  0.5000
Neg Pred Value                        0.5000             NA                  1.0000
Prevalence                            0.8333              0                  0.1667
Detection Rate                        0.6667              0                  0.1667
Detection Prevalence                  0.6667              0                  0.3333
Balanced Accuracy                     0.9000             NA                  0.9000
```

Figure 74: Naive Bayes - Experiment 1 Output

Here the distribution of success level had class imbalance issue, where the percentage of halfway was only 21%, Minimum goal met was 34% and Funding Goal Met being the highest at 45%. For test sampling, due to the 70:30 sampling no halfway observation was available for testing. This eventually had impacted the output, where 97% was achieved in training dataset with just 1 observation with funding goal met was wrongly classified as minimum goal met. Then, in testing 83% was achieved with similar 1 observation being misclassified. The AUC value was 90%. Noted that despite being a dominant class, funding goal met still had small number of observations unable to correctly be predicted., thus we would sample with the normalised dataset next.

Experiment 2: Random sampling with all features and normalisation performed

The original cleaned dataset was log transformed on skewed attributes was used. All features would be included in this model where train test splitting would use random samplingtechnique.



```
> prop.table(table(trainDF$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
       0.4482759        0.2068966        0.3448276 
> prop.table(table(testDF$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
       0.8333333        0.0000000        0.1666667 
```

```
> confusionMatrix(y_pred_train_NB1,trainDF$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               12       0                0
  Halfway                         0       6                0
  Minimum Goal Met                1       0               10

Overall Statistics
                                          
               Accuracy : 0.9655          
                 95% CI : (0.8224, 0.9991)
    No Information Rate : 0.4483          
    P-Value [Acc > NIR] : 2.88e-09        
                                          
                  Kappa : 0.9462          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.9231         1.0000                  1.0000
Specificity                           1.0000         1.0000                  0.9474
Pos Pred Value                        1.0000         1.0000                  0.9091
Neg Pred Value                        0.9412         1.0000                  1.0000
Prevalence                            0.4483         0.2069                  0.3448
Detection Rate                        0.4138         0.2069                  0.3448
Detection Prevalence                  0.4138         0.2069                  0.3793
Balanced Accuracy                     0.9615         1.0000                  0.9737
```

```
> confusionMatrix(y_pred_test_NB1,testDF$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                3       0                0
  Halfway                         0       0                0
  Minimum Goal Met                2       0                1

Overall Statistics
                                          
               Accuracy : 0.6667          
                 95% CI : (0.2228, 0.9567)
    No Information Rate : 0.8333          
    P-Value [Acc > NIR] : 0.9377          
                                          
                  Kappa : 0.3333          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6000             NA                  1.0000
Specificity                           1.0000              1                  0.6000
Pos Pred Value                        1.0000             NA                  0.3333
Neg Pred Value                        0.3333             NA                  1.0000
Prevalence                            0.8333              0                  0.1667
Detection Rate                        0.5000              0                  0.1667
Detection Prevalence                  0.5000              0                  0.5000
Balanced Accuracy                     0.8000             NA                  0.8000
```

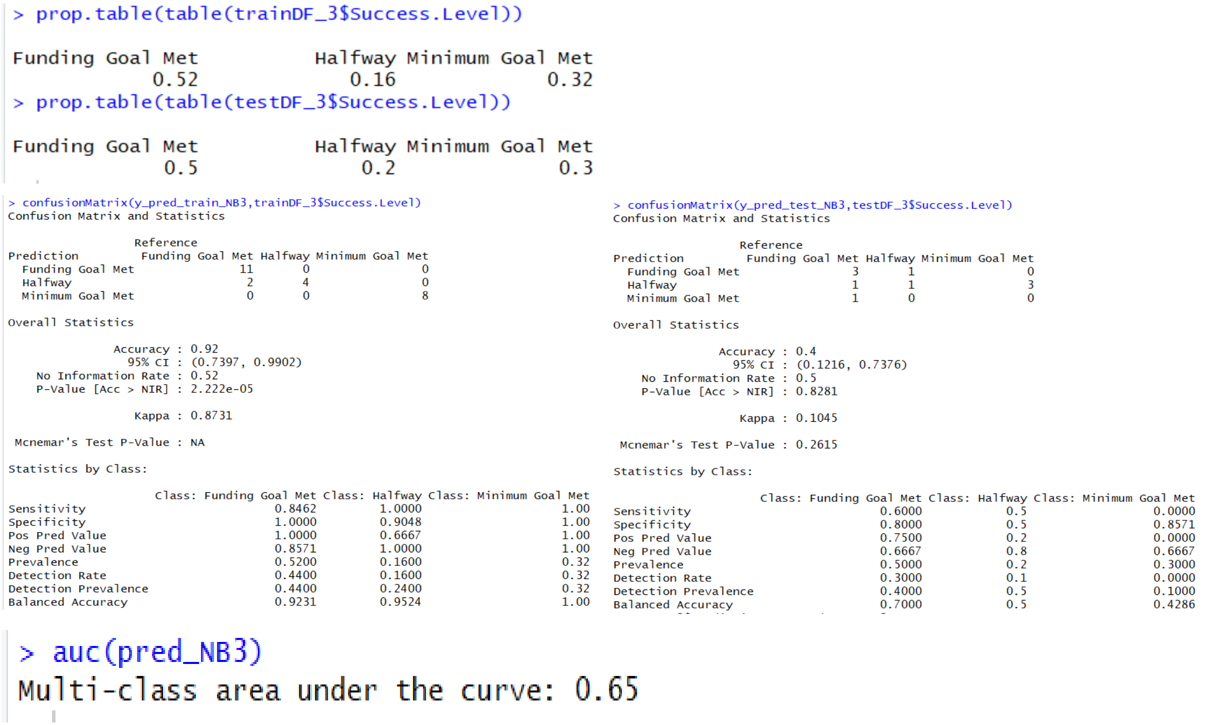
```
> auc(pred_NB1)
Multi-class area under the curve: 0.8
```

Figure 75: Naive Bayes - Experiment 2 Output

Here the distribution of success level and the splitting was similar with the previous experiment. The training accuracy was also similar at 97% with 1 funding goal met was wrongly classified as minimum goal met. However, the testing accuracy had further dropped to 67% confirming the normalisation of the dataset had caused overfitting of the model with 2 funding goal met was wrongly classified. The AUC value was 80%. Noticed no further improvement on the funding goal met rather it had more misclassification. Thus, next to experiment using stratified sampling to see if there is better distribution of data and an improved accuracy.

Experiment 3: Stratified sampling with all features and no normalisation performed

The original cleaned dataset with no normalisation performed was used. All features would be used to run this model using stratified sampling technique for the train test splitting.



```
> prop.table(table(trainDF_3$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
            0.52             0.16             0.32 
> prop.table(table(testDF_3$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
             0.5              0.2              0.3 
```

```
> confusionMatrix(y_pred_train_NB3,trainDF_3$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               11       0                0
  Halfway                         2       4                0
  Minimum Goal Met                0       0                8

Overall Statistics
                                          
               Accuracy : 0.92            
                 95% CI : (0.7397, 0.9902)
    No Information Rate : 0.52            
    P-Value [Acc > NIR] : 2.222e-05       
                                          
                  Kappa : 0.8731          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.8462         1.0000                    1.00
Specificity                           1.0000         0.9048                    1.00
Pos Pred Value                        1.0000         0.6667                    1.00
Neg Pred Value                        0.8571         1.0000                    1.00
Prevalence                            0.5200         0.1600                    0.32
Detection Rate                        0.4400         0.1600                    0.32
Detection Prevalence                  0.4400         0.2400                    0.32
Balanced Accuracy                     0.9231         0.9524                    1.00
```

```
> confusionMatrix(y_pred_test_NB3,testDF_3$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                3       1                0
  Halfway                         1       1                3
  Minimum Goal Met                1       0                0

Overall Statistics
                                          
               Accuracy : 0.4             
                 95% CI : (0.1216, 0.7376)
    No Information Rate : 0.5             
    P-Value [Acc > NIR] : 0.8281          
                                          
                  Kappa : 0.1045          
                                          
 Mcnemar's Test P-Value : 0.2615          

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6000            0.5                  0.0000
Specificity                           0.8000            0.5                  0.8571
Pos Pred Value                        0.7500            0.2                  0.0000
Neg Pred Value                        0.6667            0.8                  0.6667
Prevalence                            0.5000            0.2                  0.3000
Detection Rate                        0.3000            0.1                  0.0000
Detection Prevalence                  0.4000            0.5                  0.1000
Balanced Accuracy                     0.7000            0.5                  0.4286
```

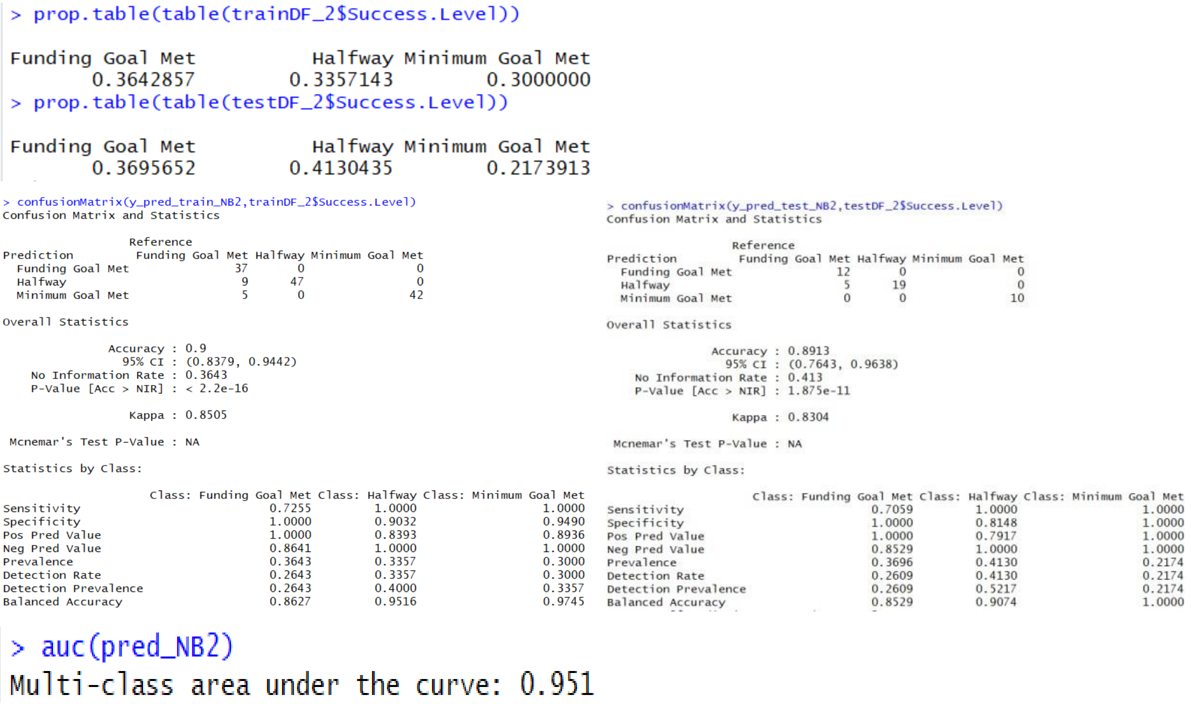
```
> auc(pred_NB3)
Multi-class area under the curve: 0.65
```

Figure 76: Naive Bayes - Experiment 3 Output

Here, the stratified sampling had improved the training and test division where now there is sampling for halfway in the train dataset. However, here the misclassification had increase where 2 observations from funding goal met in both datasets was noted thus a drop in the accuracy rate. Additionally, the halfway in test had 1 misclassification confirming this model to be overfitting thus the very low accuracy 40% compared to experiment 1. Here, we conclude stratified sampling is not suited for this dataset. Next, as this dataset has a class imbalance issue, the SMOTE dataset would be used to perform the modelling with both the normalised and not normalised data.

Experiment 4: Random sampling using SMOTE dataset with all features and normalised

The SMOTE dataset with all features and normalised would be used to run this model using random sampling for the train test splitting.



```
> prop.table(table(trainDF_2$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.3642857         0.3357143        0.3000000
> prop.table(table(testDF_2$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.3695652         0.4130435        0.2173913
```

```
> confusionMatrix(y_pred_train_NB2,trainDF_2$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                37       0                0
  Halfway                          9      47                0
  Minimum Goal Met                 5       0               42

Overall Statistics

               Accuracy : 0.9
                 95% CI : (0.8379, 0.9442)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8505

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7255         1.0000                  1.0000
Specificity                           1.0000         0.9032                  0.9490
Pos Pred Value                        1.0000         0.8393                  0.8936
Neg Pred Value                        0.8641         1.0000                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2643         0.3357                  0.3000
Detection Prevalence                  0.2643         0.4000                  0.3357
Balanced Accuracy                     0.8627         0.9516                  0.9745
```

```
> confusionMatrix(y_pred_test_NB2,testDF_2$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                12       0                0
  Halfway                          5      19                0
  Minimum Goal Met                 0       0               10

Overall Statistics

               Accuracy : 0.8913
                 95% CI : (0.7643, 0.9638)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 1.875e-11

                  Kappa : 0.8304

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7059         1.0000                  1.0000
Specificity                           1.0000         0.8148                  1.0000
Pos Pred Value                        1.0000         0.7917                  1.0000
Neg Pred Value                        0.8529         1.0000                  1.0000
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2609         0.4130                  0.2174
Detection Prevalence                  0.2609         0.5217                  0.2174
Balanced Accuracy                     0.8529         0.9074                  1.0000
```

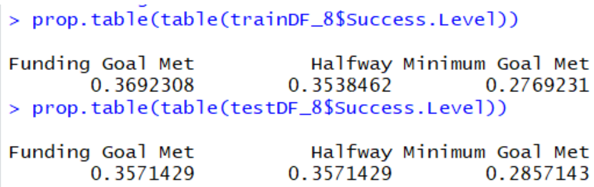
```
> auc(pred_NB2)
Multi-class area under the curve: 0.951
```

Figure 77: Naive Bayes - Experiment 4 Output

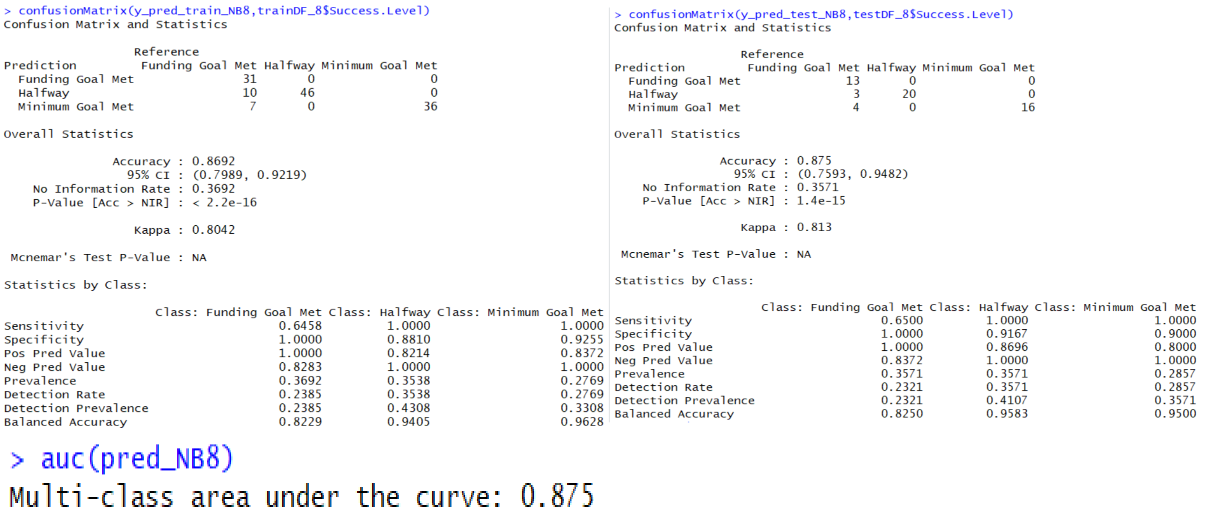
SMOTE dataset has more observation compared to the original dataset. Here, noted the model accuracy had improved further with no overfitting concern like experiment 2 with the normalised dataset. Training had an accuracy of 90% and Test with an accuracy of 89% with AUC of 95% as only 5 funding goal met observation could not be correctly classified. Like previous experiments, halfway and minimum goal met are being able to fully classify whereby funding goal met still has small number of observations unable to be predicted correctly. Next, the SMOTE dataset would be used with the stratified sampling to see if we still encounter overfitting mode.

Experiment 5: Random sampling using SMOTE dataset with

The SMOTE dataset with all features and normalised would be used to run this model using stratified sampling for the train test splitting.



```
> prop.table(table(trainDF_8$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.3692308         0.3538462        0.2769231
> prop.table(table(testDF_8$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.3571429         0.3571429        0.2857143
```



```
> confusionMatrix(y_pred_train_NB8,trainDF_8$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                31       0                0
  Halfway                         10      46                0
  Minimum Goal Met                 7       0               36

Overall Statistics

               Accuracy : 0.8692
                 95% CI : (0.7989, 0.9219)
    No Information Rate : 0.3692
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8042

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6458         1.0000                  1.0000
Specificity                           1.0000         0.8810                  0.9255
Pos Pred Value                        1.0000         0.8214                  0.8372
Neg Pred Value                        0.8283         1.0000                  1.0000
Prevalence                            0.3692         0.3538                  0.2769
Detection Rate                        0.2385         0.3538                  0.2769
Detection Prevalence                  0.2385         0.4308                  0.3308
Balanced Accuracy                     0.8229         0.9405                  0.9628
```

```
> confusionMatrix(y_pred_test_NB8,testDF_8$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                13       0                0
  Halfway                          3      20                0
  Minimum Goal Met                 4       0               16

Overall Statistics

               Accuracy : 0.875
                 95% CI : (0.7593, 0.9482)
    No Information Rate : 0.3571
    P-Value [Acc > NIR] : 1.4e-15

                  Kappa : 0.813

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6500         1.0000                  1.0000
Specificity                           1.0000         0.9167                  0.9000
Pos Pred Value                        1.0000         0.8696                  0.8000
Neg Pred Value                        0.8372         1.0000                  1.0000
Prevalence                            0.3571         0.3571                  0.2857
Detection Rate                        0.2321         0.3571                  0.2857
Detection Prevalence                  0.2321         0.4107                  0.3571
Balanced Accuracy                     0.8250         0.9583                  0.9500
```

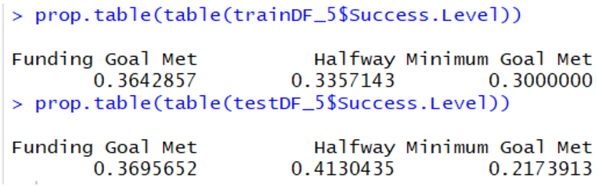
```
> auc(pred_NB8)
Multi-class area under the curve: 0.875
```

Figure 78: Naive Bayes - Experiment 5 Output

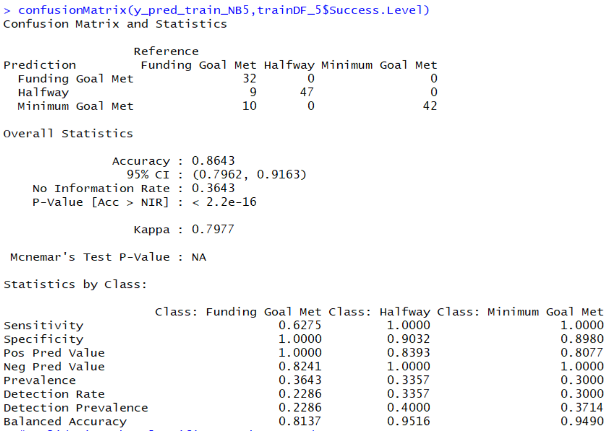
Compared to the results in Experiment 3 which was an overfitting model, here we obtained a better model with an accuracy of 87% for train and 88% for test. A better trained model with higher test outcome. However, compared to the random sampling a balanced split obtained between success level however there is more misclassified funding goal met here where it is misclassified to the other success level. Next, to improve the model better the Laplace smoothing would be performed to observe if an improved model is obtained.

Experiment 6: Random sampling using SMOTE dataset with Laplace

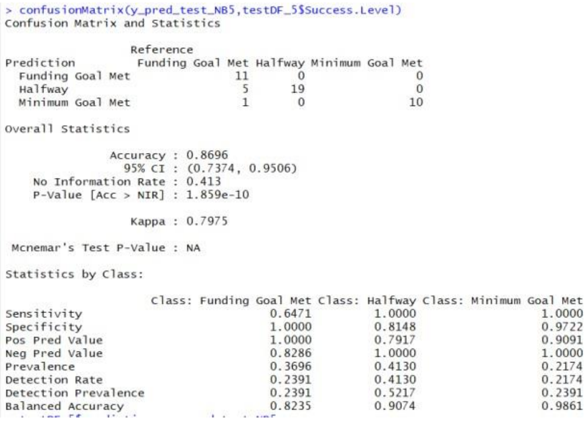
The SMOTE dataset with all features would go through the Laplace smoothing to identify if the model could be further optimised.



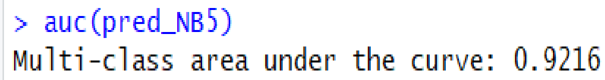
```
> prop.table(table(trainDF_5$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met 
       0.3642857        0.3357143        0.3000000 
> prop.table(table(testDF_5$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met 
       0.3695652        0.4130435        0.2173913 
```



```
> confusionMatrix(y_pred_train_NB5,trainDF_5$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                32       0                0
  Halfway                          9      47                0
  Minimum Goal Met                10       0               42

Overall Statistics

               Accuracy : 0.8643
                 95% CI : (0.7962, 0.9163)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7977

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6275         1.0000                  1.0000
Specificity                           1.0000         0.9032                  0.8980
Pos Pred Value                        1.0000         0.8393                  0.8077
Neg Pred Value                        0.8241         1.0000                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2286         0.3357                  0.3000
Detection Prevalence                  0.2286         0.4000                  0.3714
Balanced Accuracy                     0.8137         0.9516                  0.9490
```



```
> confusionMatrix(y_pred_test_NB5,testDF_5$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                11       0                0
  Halfway                          5      19                0
  Minimum Goal Met                 1       0               10

Overall Statistics

               Accuracy : 0.8696
                 95% CI : (0.7374, 0.9506)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 1.859e-10

                  Kappa : 0.7975

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6471         1.0000                  1.0000
Specificity                           1.0000         0.8148                  0.9722
Pos Pred Value                        1.0000         0.7917                  0.9091
Neg Pred Value                        0.8286         1.0000                  1.0000
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2391         0.4130                  0.2174
Detection Prevalence                  0.2391         0.5217                  0.2391
Balanced Accuracy                     0.8235         0.9074                  0.9861
```



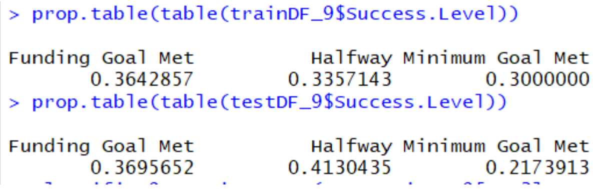
```
> auc(pred_NB5)
Multi-class area under the curve: 0.9216
```

Figure 79: Naive Bayes - Experiment 6 Output

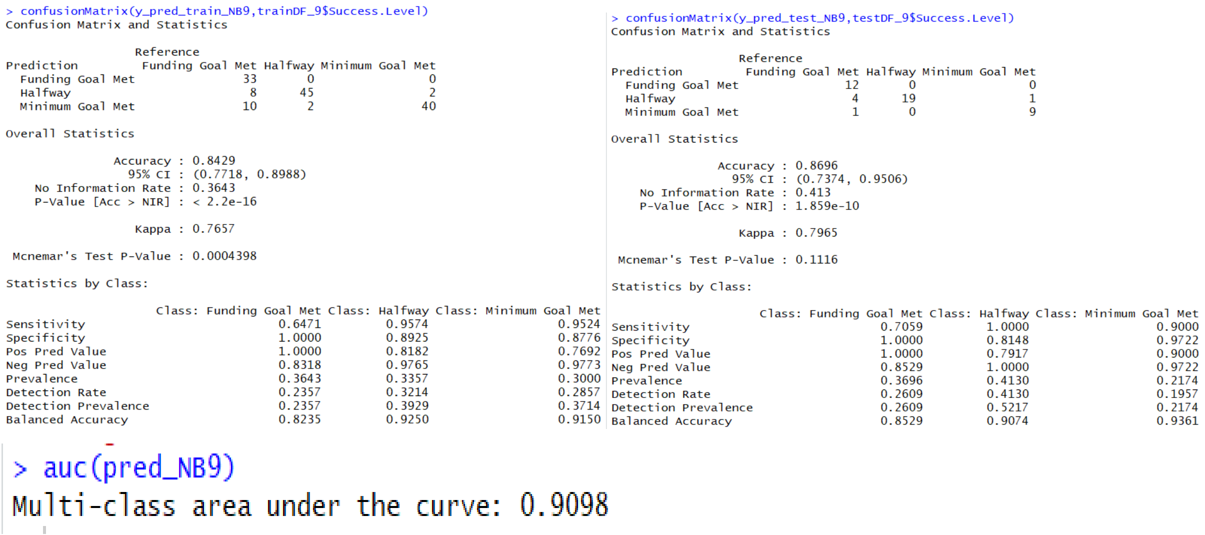
As Laplace smoothing assist to smooth categorical data, here noted the model accuracy did not improve further rather lower by 0.04%. Additionally, training to dropped further with more funding goal met being misclassified. Thus, Laplace smoothing does not smoothen the dataset further. This was also performed on the original dataset, similarly no further improvement on the model was observed. Variation on the pre-processing of the dataset, sampling method and Naïve Bayes optimisation was performed, yet the misclassification of funding goal met remained same through. Thus, there may be features within the dataset that may be causing the confusion in classifying them correctly. Thus, next the features that was identified in the feature selection technique would be used to model the dataset to see the changes on the prediction of the classes.

Experiment 7: Random sampling using SMOTE dataset with Tree Based Algorithm features

The SMOTE dataset with features identified from the tree-based algorithm would be used to run this model using random sampling.



```
> prop.table(table(trainDF_9$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met
       0.3642857        0.3357143        0.3000000
> prop.table(table(testDF_9$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met
       0.3695652        0.4130435        0.2173913
```



```
> confusionMatrix(y_pred_train_NB9,trainDF_9$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                33       0                0
  Halfway                          8      45                2
  Minimum Goal Met                10       2               40

Overall Statistics

               Accuracy : 0.8429
                 95% CI : (0.7718, 0.8988)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7657

 Mcnemar's Test P-Value : 0.0004398

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6471         0.9574                  0.9524
Specificity                           1.0000         0.8925                  0.8776
Pos Pred Value                        1.0000         0.8182                  0.7692
Neg Pred Value                        0.8318         0.9765                  0.9773
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2357         0.3214                  0.2857
Detection Prevalence                  0.2357         0.3929                  0.3714
Balanced Accuracy                     0.8235         0.9250                  0.9150
```

```
> confusionMatrix(y_pred_test_NB9,testDF_9$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                12       0                0
  Halfway                          4      19                1
  Minimum Goal Met                 1       0                9

Overall Statistics

               Accuracy : 0.8696
                 95% CI : (0.7374, 0.9506)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 1.859e-10

                  Kappa : 0.7965

 Mcnemar's Test P-Value : 0.1116

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7059         1.0000                  0.9000
Specificity                           1.0000         0.8148                  0.9722
Pos Pred Value                        1.0000         0.7917                  0.9000
Neg Pred Value                        0.8529         1.0000                  0.9722
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2609         0.4130                  0.1957
Detection Prevalence                  0.2609         0.5217                  0.2174
Balanced Accuracy                     0.8529         0.9074                  0.9361
```

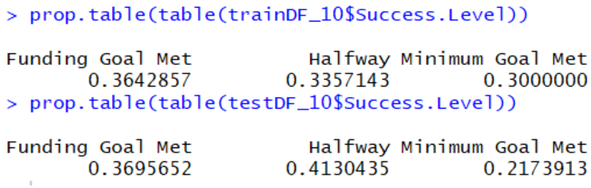
```
> auc(pred_NB9)
Multi-class area under the curve: 0.9098
```

Figure 80: Naive Bayes - Experiment 7 Output

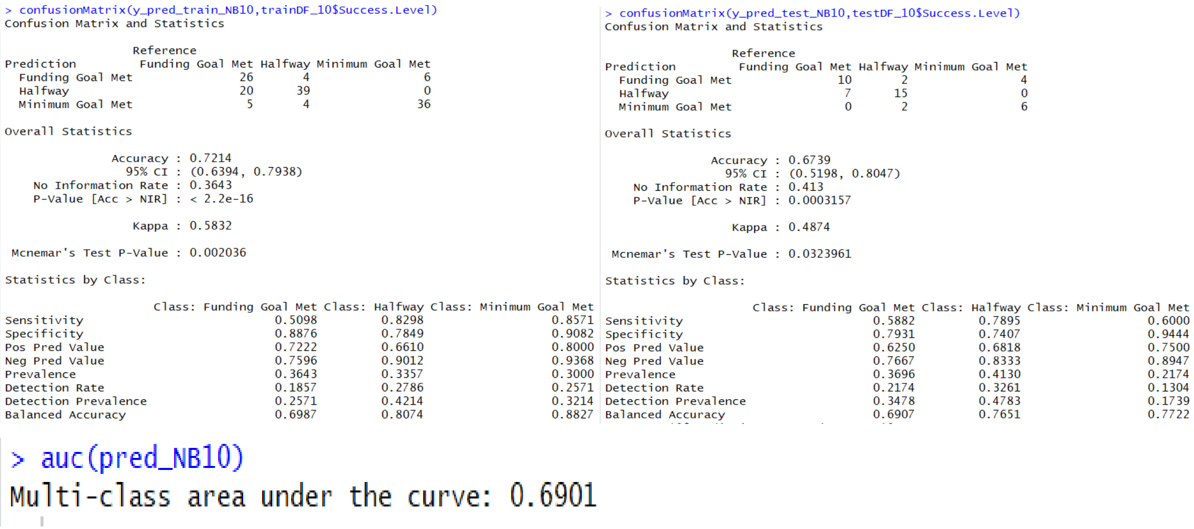
Based on this minimised feature, the accuracy rate had further drop where training at 84% and train at 87% with AUC of 91%. The funding goal met success level had not further improved rather the misclassification rate had increase along with the other success level which had no previous misclassification and now it is being misclassified (minimum goal met). Thus, this features alone would not be able to further distinguish the success level. Next, the features selected by the Boruta technique would be modelled to observe any improvement on the prediction.

Experiment 8: Random sampling using SMOTE dataset with Boruta features

The SMOTE dataset with features identified from the Boruta would be used to run this model using random sampling.



```
> prop.table(table(trainDF_10$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met 
       0.3642857        0.3357143        0.3000000 
> prop.table(table(testDF_10$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met 
       0.3695652        0.4130435        0.2173913 
```



```
> confusionMatrix(y_pred_train_NB10,trainDF_10$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                26       4                6
  Halfway                         20      39                0
  Minimum Goal Met                 5       4               36

Overall Statistics

               Accuracy : 0.7214
                 95% CI : (0.6394, 0.7938)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5832

 Mcnemar's Test P-Value : 0.002036

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.5098         0.8298                  0.8571
Specificity                           0.8876         0.7849                  0.9082
Pos Pred Value                        0.7222         0.6610                  0.8000
Neg Pred Value                        0.7596         0.9012                  0.9368
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.1857         0.2786                  0.2571
Detection Prevalence                  0.2571         0.4214                  0.3214
Balanced Accuracy                     0.6987         0.8074                  0.8827
```

```
> confusionMatrix(y_pred_test_NB10,testDF_10$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                10       2                4
  Halfway                          7      15                0
  Minimum Goal Met                 0       2                6

Overall Statistics

               Accuracy : 0.6739
                 95% CI : (0.5198, 0.8047)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 0.0003157

                  Kappa : 0.4874

 Mcnemar's Test P-Value : 0.0323961

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.5882         0.7895                  0.6000
Specificity                           0.7931         0.7407                  0.9444
Pos Pred Value                        0.6250         0.6818                  0.7500
Neg Pred Value                        0.7667         0.8333                  0.8947
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2174         0.3261                  0.1304
Detection Prevalence                  0.3478         0.4783                  0.1739
Balanced Accuracy                     0.6907         0.7651                  0.7722
```

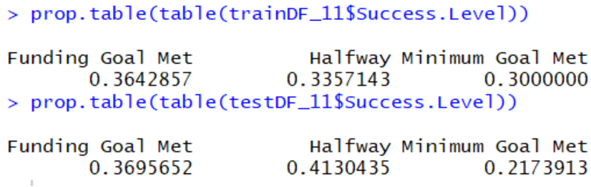
```
> auc(pred_NB10)
Multi-class area under the curve: 0.6901
```

Figure 81: Naive Bayes - Experiment 8 Output

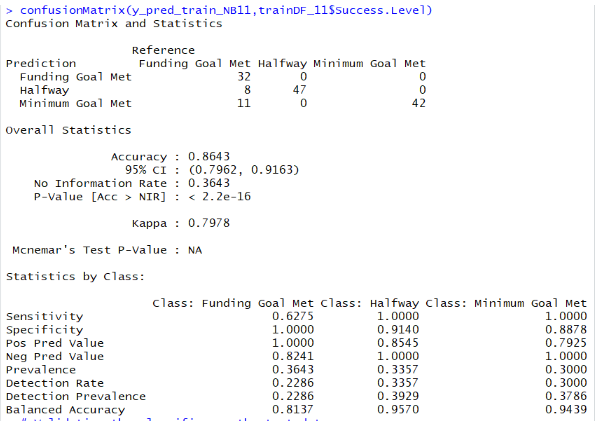
Here, noted that my removing more attributes further does not assist in improving the dataset. Thus, this Boruta technique would not be further explored moreover where it had identified all the attributes from the SMOTE dataset as important.

Experiment 9: Random sampling using SMOTE dataset with SMOTE Tree Based features

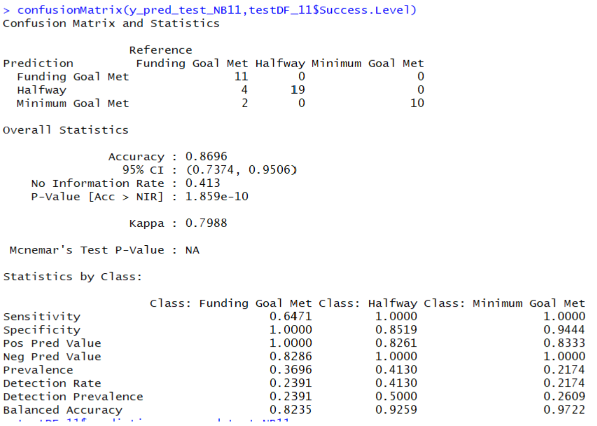
The SMOTE dataset with features identified from the tree-based algorithm using the SMOTE dataset would be used to run this model with random sampling.



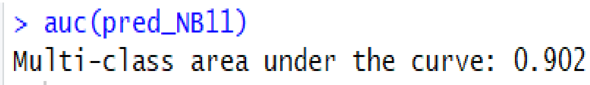
```
> prop.table(table(trainDF_11$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met
       0.3642857        0.3357143        0.3000000
> prop.table(table(testDF_11$Success.Level))

Funding Goal Met          Halfway Minimum Goal Met
       0.3695652        0.4130435        0.2173913
```



```
> confusionMatrix(y_pred_train_NB11,trainDF_11$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                32       0                0
  Halfway                          8      47                0
  Minimum Goal Met                11       0               42

Overall Statistics

               Accuracy : 0.8643
                 95% CI : (0.7962, 0.9163)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7978

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6275         1.0000                  1.0000
Specificity                           1.0000         0.9140                  0.8878
Pos Pred Value                        1.0000         0.8545                  0.7925
Neg Pred Value                        0.8241         1.0000                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2286         0.3357                  0.3000
Detection Prevalence                  0.2286         0.3929                  0.3786
Balanced Accuracy                     0.8137         0.9570                  0.9439
```



```
> confusionMatrix(y_pred_test_NB11,testDF_11$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                11       0                0
  Halfway                          4      19                0
  Minimum Goal Met                 2       0               10

Overall Statistics

               Accuracy : 0.8696
                 95% CI : (0.7374, 0.9506)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 1.859e-10

                  Kappa : 0.7988

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6471         1.0000                  1.0000
Specificity                           1.0000         0.8519                  0.9444
Pos Pred Value                        1.0000         0.8261                  0.8333
Neg Pred Value                        0.8286         1.0000                  1.0000
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2391         0.4130                  0.2174
Detection Prevalence                  0.2391         0.5000                  0.2609
Balanced Accuracy                     0.8235         0.9259                  0.9722
```



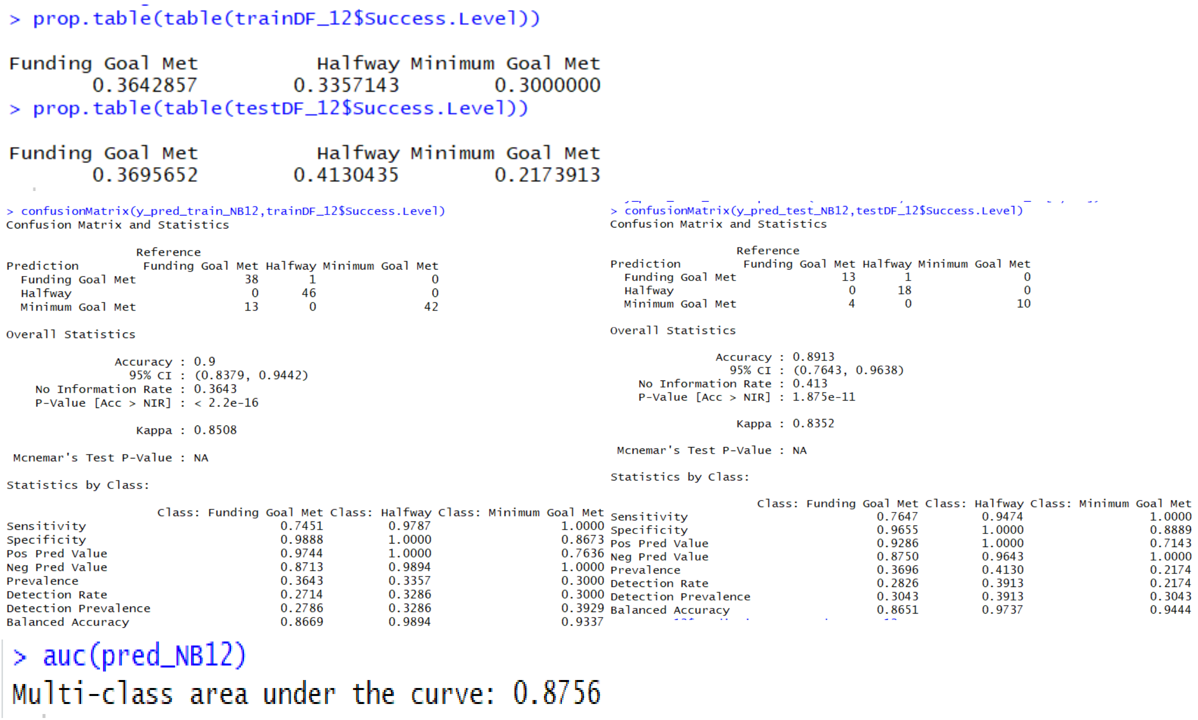
```
> auc(pred_NB11)
Multi-class area under the curve: 0.902
```

Figure 82: Naive Bayes - Experiment 9 Output

The accuracy rate compared to all the feature selection experiment, this had a better learning outcome with training at 86%, however test accuracy had no changes at 87% and AUC with 0.08% drop. Observing the prediction of each success level, this method did not have any misclassification of halfway and minimum funding goal compared to the other feature selection techniques however, comparing to the experiment 4 where all features was used, this model is lower in accuracy. Then an experiment to combine both Original Tree Based and SMOTE Tree based features to identify if the accuracy could be improved further, here noted that the accuracy dropped very low to 77% and AUC down to 73% indicating an overfitted model. To further improve the classification rate, the dimensionally reduced dataset from the factor analysis was used to observe the model outcome.

Experiment 10: Random sampling using SMOTE dataset with Factor Analysis

The SMOTE dataset with the 10 principal components identified from the factor analysis activity used to run this model with random sampling.



```
> prop.table(table(trainDF_12$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
       0.3642857         0.3357143        0.3000000 
> prop.table(table(testDF_12$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met 
       0.3695652         0.4130435        0.2173913 
```

```
> confusionMatrix(y_pred_train_NB12,trainDF_12$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               38       1                0
  Halfway                         0      46                0
  Minimum Goal Met               13       0               42

Overall Statistics
                                          
               Accuracy : 0.9             
                 95% CI : (0.8379, 0.9442)
    No Information Rate : 0.3643          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.8508          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7451         0.9787                  1.0000
Specificity                           0.9888         1.0000                  0.8673
Pos Pred Value                        0.9744         1.0000                  0.7636
Neg Pred Value                        0.8713         0.9894                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2714         0.3286                  0.3000
Detection Prevalence                  0.2786         0.3286                  0.3929
Balanced Accuracy                     0.8669         0.9894                  0.9337
```

```
> confusionMatrix(y_pred_test_NB12,testDF_12$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               13       1                0
  Halfway                         0      18                0
  Minimum Goal Met                4       0               10

Overall Statistics
                                          
               Accuracy : 0.8913          
                 95% CI : (0.7643, 0.9638)
    No Information Rate : 0.413           
    P-Value [Acc > NIR] : 1.875e-11       
                                          
                  Kappa : 0.8352          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7647         0.9474                  1.0000
Specificity                           0.9655         1.0000                  0.8889
Pos Pred Value                        0.9286         1.0000                  0.7143
Neg Pred Value                        0.8750         0.9643                  1.0000
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2826         0.3913                  0.2174
Detection Prevalence                  0.3043         0.3913                  0.3043
Balanced Accuracy                     0.8651         0.9737                  0.9444
```

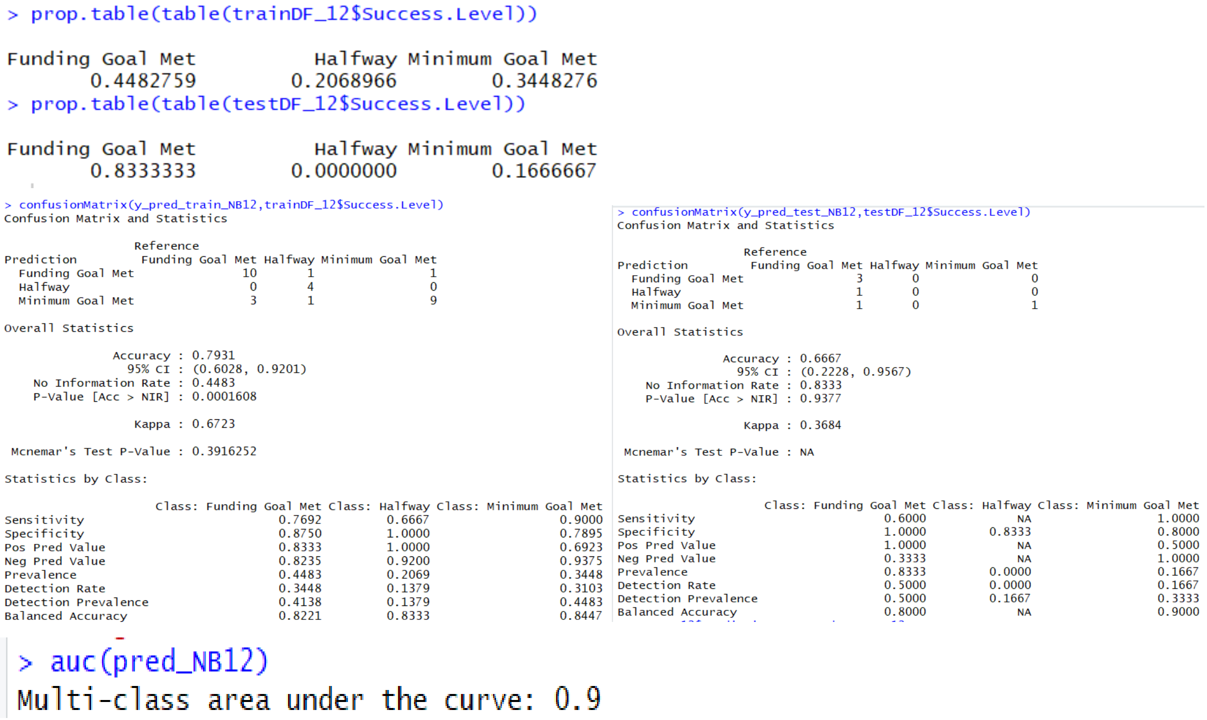
```
> auc(pred_NB12)
Multi-class area under the curve: 0.8756
```

Figure 83: Naive Bayes - Experiment 10 Output

The model utilising the 10 principal component has similar accuracy with Experiment 4 with all 33 variables experiment, with around 0.08% lower in the AUC value due to the increase in misclassification of halfway despite having slight better classification of funding goal. Thus, here this model is considered better as with just 10 principal component we are able to predict the same as experiment 4 by not losing any variables for prediction.

Experiment 11: Random sampling using Original dataset with Factor Analysis

The original dataset with the 10 principal components identified from the factor analysis activity used to run this model with random sampling.



```
> prop.table(table(trainDF_12$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.4482759          0.2068966         0.3448276
> prop.table(table(testDF_12$Success.Level))

Funding Goal Met           Halfway Minimum Goal Met
       0.8333333          0.0000000         0.1666667
```

```
> confusionMatrix(y_pred_train_NB12,trainDF_12$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               10       1                1
  Halfway                         0       4                0
  Minimum Goal Met                3       1                9

Overall Statistics

               Accuracy : 0.7931
                 95% CI : (0.6028, 0.9201)
    No Information Rate : 0.4483
    P-Value [Acc > NIR] : 0.0001608

                  Kappa : 0.6723

 Mcnemar's Test P-Value : 0.3916252

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7692         0.6667                  0.9000
Specificity                           0.8750         1.0000                  0.7895
Pos Pred Value                        0.8333         1.0000                  0.6923
Neg Pred Value                        0.8235         0.9200                  0.9375
Prevalence                            0.4483         0.2069                  0.3448
Detection Rate                        0.3448         0.1379                  0.3103
Detection Prevalence                  0.4138         0.1379                  0.4483
Balanced Accuracy                     0.8221         0.8333                  0.8447
```

```
> confusionMatrix(y_pred_test_NB12,testDF_12$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                3       0                0
  Halfway                         1       0                0
  Minimum Goal Met                1       0                1

Overall Statistics

               Accuracy : 0.6667
                 95% CI : (0.2228, 0.9567)
    No Information Rate : 0.8333
    P-Value [Acc > NIR] : 0.9377

                  Kappa : 0.3684

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6000             NA                  1.0000
Specificity                           1.0000         0.8333                  0.8000
Pos Pred Value                        1.0000             NA                  0.5000
Neg Pred Value                        0.3333             NA                  1.0000
Prevalence                            0.8333         0.0000                  0.1667
Detection Rate                        0.5000         0.0000                  0.1667
Detection Prevalence                  0.5000         0.1667                  0.3333
Balanced Accuracy                     0.8000             NA                  0.9000
```

```
> auc(pred_NB12)
Multi-class area under the curve: 0.9
```

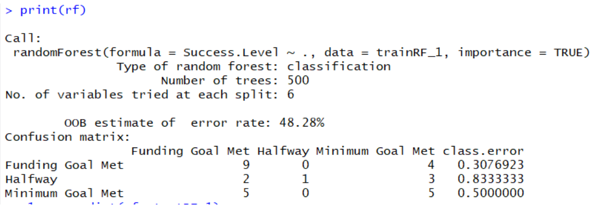
Figure 84: Naive Bayes - Experiment 11 Output

Like previous experiment, here noted a lower accuracy rate but still having a similar AUC value of 90% as experiment 1. Noted a poorer learning occurred here compared to Experiment 1, thus no benefit obtained by using the 10 principal components. As we have explored all angle with Naïve Bayes, next is to improve the model accuracy further using Random Forest model.

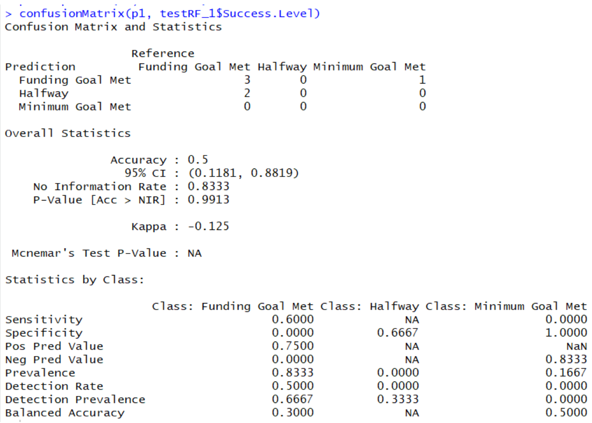
4.3.6.2 Random Forest

Experiment 1: Original dataset

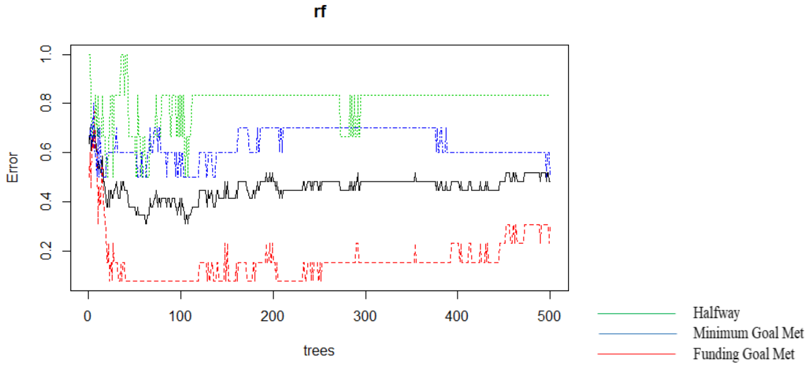
The original dataset would be used to run the Random Forest model with random search, output as illustrated below.

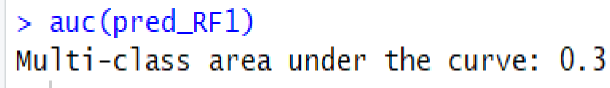


```
> print(rf)

Call:
 randomForest(formula = Success.Level ~ ., data = trainRF_1, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 48.28%
Confusion matrix:
                 Funding Goal Met Halfway Minimum Goal Met class.error
Funding Goal Met                9       0                4   0.3076923
Halfway                         2       1                3   0.8333333
Minimum Goal Met                5       0                5   0.5000000
```



```
> confusionMatrix(p1, testRF_1$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                3       0                1
  Halfway                         2       0                0
  Minimum Goal Met                0       0                0

Overall Statistics

               Accuracy : 0.5
                 95% CI : (0.1181, 0.8819)
    No Information Rate : 0.8333
    P-Value [Acc > NIR] : 0.9913

                  Kappa : -0.125

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.6000             NA                  0.0000
Specificity                           0.0000         0.6667                  1.0000
Pos Pred Value                        0.7500             NA                     NaN
Neg Pred Value                        0.0000             NA                  0.8333
Prevalence                            0.8333         0.0000                  0.1667
Detection Rate                        0.5000         0.0000                  0.0000
Detection Prevalence                  0.6667         0.3333                  0.0000
Balanced Accuracy                     0.3000             NA                  0.5000
```





```
> auc(pred_RF1)
Multi-class area under the curve: 0.3
```

Figure 85: Random Forest - Experiment 1 Output

By using 500 trees with 6 variables used at each split, the training set has an accuracy of 52% and the test data with slight lower accuracy of 50%. Here, noted that funding goal and the minimum goal met has quiet high misclassification observations for both training and test. Halfway has the highest error rate at around 83% due to both the other levels has been misclassified as halfway. Figure 86 highlights the important variables used in this model.

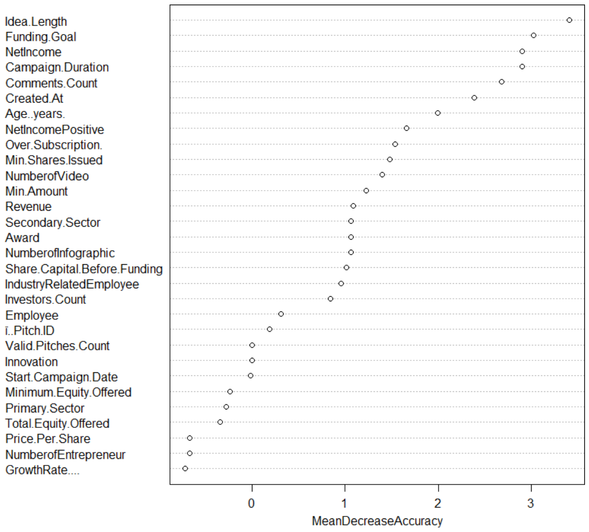
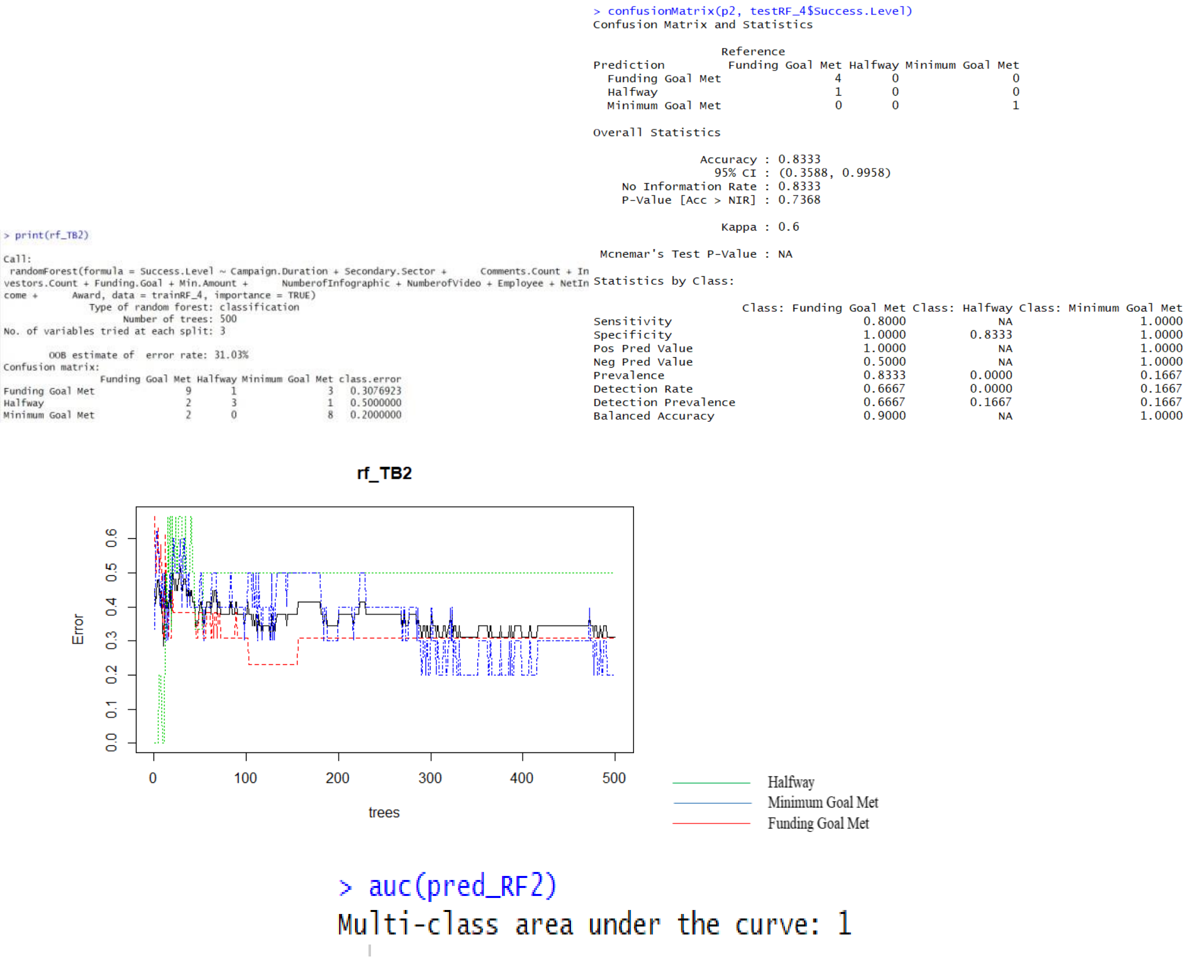


Figure 86: Variable of Importance - RF Exp 1

Idea Length, Funding Goal, Net Income, Campaign Duration and Comments count being the top 5 important variables used in this model. Next, we would experiment the original dataset with the important features selected from the Tree Based Algorithm.

Experiment 2: Original dataset with Tree Based Algorithm Features

The original dataset would be used to run the Random Forest model with random search. However, the features that was determined important from the SMOTE tree-based algorithm would be used for modelling, output as illustrated below.



```
> print(rf_TB2)

Call:
 randomForest(formula = Success.Level ~ Campaign.Duration + Secondary.Sector +      Comments.Count + In
vestors.Count + Funding.Goal + Min.Amount +      NumberofInfographic + NumberofVideo + Employee + NetIn
come +      Award, data = trainRF_4, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 31.03%
Confusion matrix:
                 Funding Goal Met Halfway Minimum Goal Met class.error
Funding Goal Met                9       1                3   0.3076923
Halfway                         2       3                1   0.5000000
Minimum Goal Met                2       0                8   0.2000000
```

```
> confusionMatrix(p2, testRF_4$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                4       0                0
  Halfway                         1       0                0
  Minimum Goal Met                0       0                1

Overall Statistics

               Accuracy : 0.8333
                 95% CI : (0.3588, 0.9958)
    No Information Rate : 0.8333
    P-Value [Acc > NIR] : 0.7368

                  Kappa : 0.6

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.8000             NA                  1.0000
Specificity                           1.0000         0.8333                  1.0000
Pos Pred Value                        1.0000             NA                  1.0000
Neg Pred Value                        0.5000             NA                  1.0000
Prevalence                            0.8333         0.0000                  0.1667
Detection Rate                        0.6667         0.0000                  0.1667
Detection Prevalence                  0.6667         0.1667                  0.1667
Balanced Accuracy                     0.9000             NA                  1.0000
```

```
> auc(pred_RF2)
Multi-class area under the curve: 1
```

Figure 87: Random Forest - Experiment 2 Output

By using 500 trees with 3 variables used at each split, the accuracy of the train data has increase to 69% and the test data with accuracy of 83%. Noted that the model is not learning well as the funding goal and the minimum goal met has quiet high number of misclassification observations, however an improved test results was obtained with lower misclassification error. Halfway error rate has reduced from 83% to 50% indicating the feature selection could distinguish the success level better compared to using all variables. Figure 88 highlights the rearrangement of the importance of the important variables used in this model.

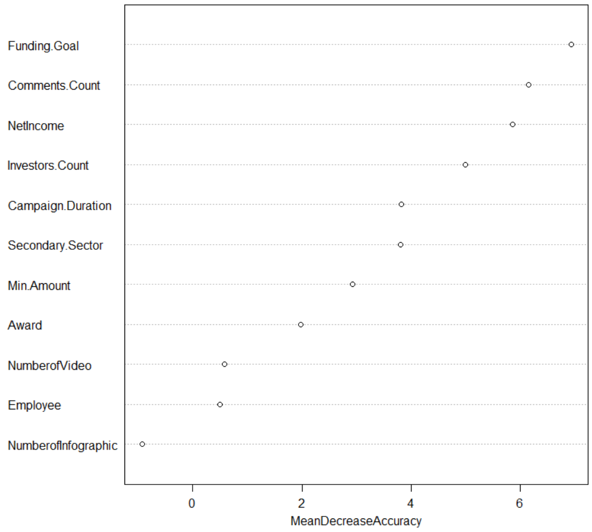
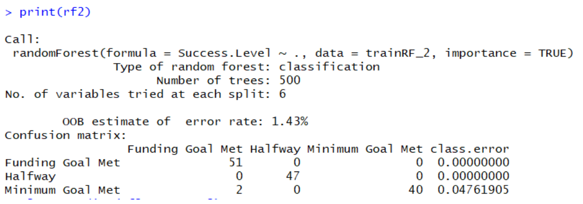


Figure 88: Variable of Importance - RF Exp 2

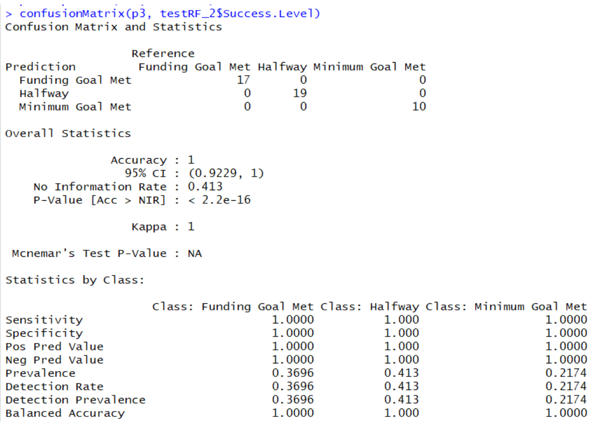
Noted the importance order compared to the SMOTE Tree Based feature selection and this Random Forest outcome has changed. After the random forest modelling, the importance of variables has changed where now the top 5 important variables are Funding Goal, Comments Count, Net Income, Investor Count and Campaign durations. Next, the Random Forest with the SMOTE model would be explored to identify the model outcome.

Experiment 3: SMOTE dataset

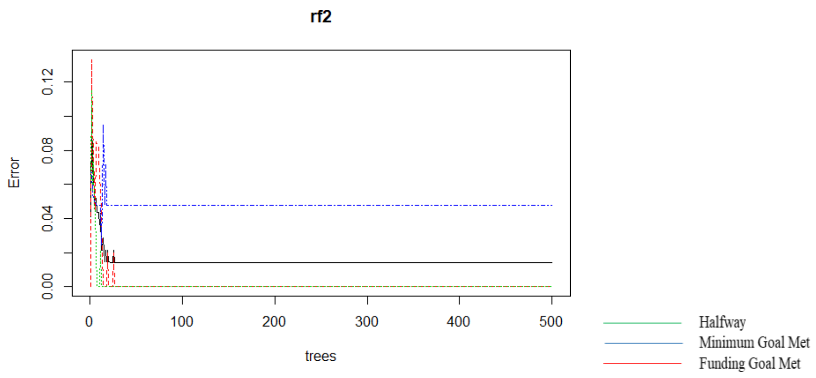
The SMOTE dataset would be used to run the Random Forest model with random search, output as illustrated below.



```
> print(rf2)

Call:
 randomForest(formula = Success.Level ~ ., data = trainRF_2, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 1.43%
Confusion matrix:
                 Funding Goal Met Halfway Minimum Goal Met class.error
Funding Goal Met               51       0                0  0.00000000
Halfway                         0      47                0  0.00000000
Minimum Goal Met                2       0               40  0.04761905
```



```
> confusionMatrix(p3, testRF_2$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                17       0                0
  Halfway                          0      19                0
  Minimum Goal Met                 0       0               10

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9229, 1)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           1.0000          1.000                  1.0000
Specificity                           1.0000          1.000                  1.0000
Pos Pred Value                        1.0000          1.000                  1.0000
Neg Pred Value                        1.0000          1.000                  1.0000
Prevalence                            0.3696          0.413                  0.2174
Detection Rate                        0.3696          0.413                  0.2174
Detection Prevalence                  0.3696          0.413                  0.2174
Balanced Accuracy                     1.0000          1.000                  1.0000
```



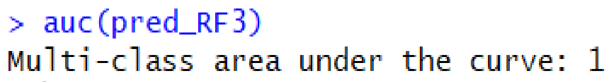
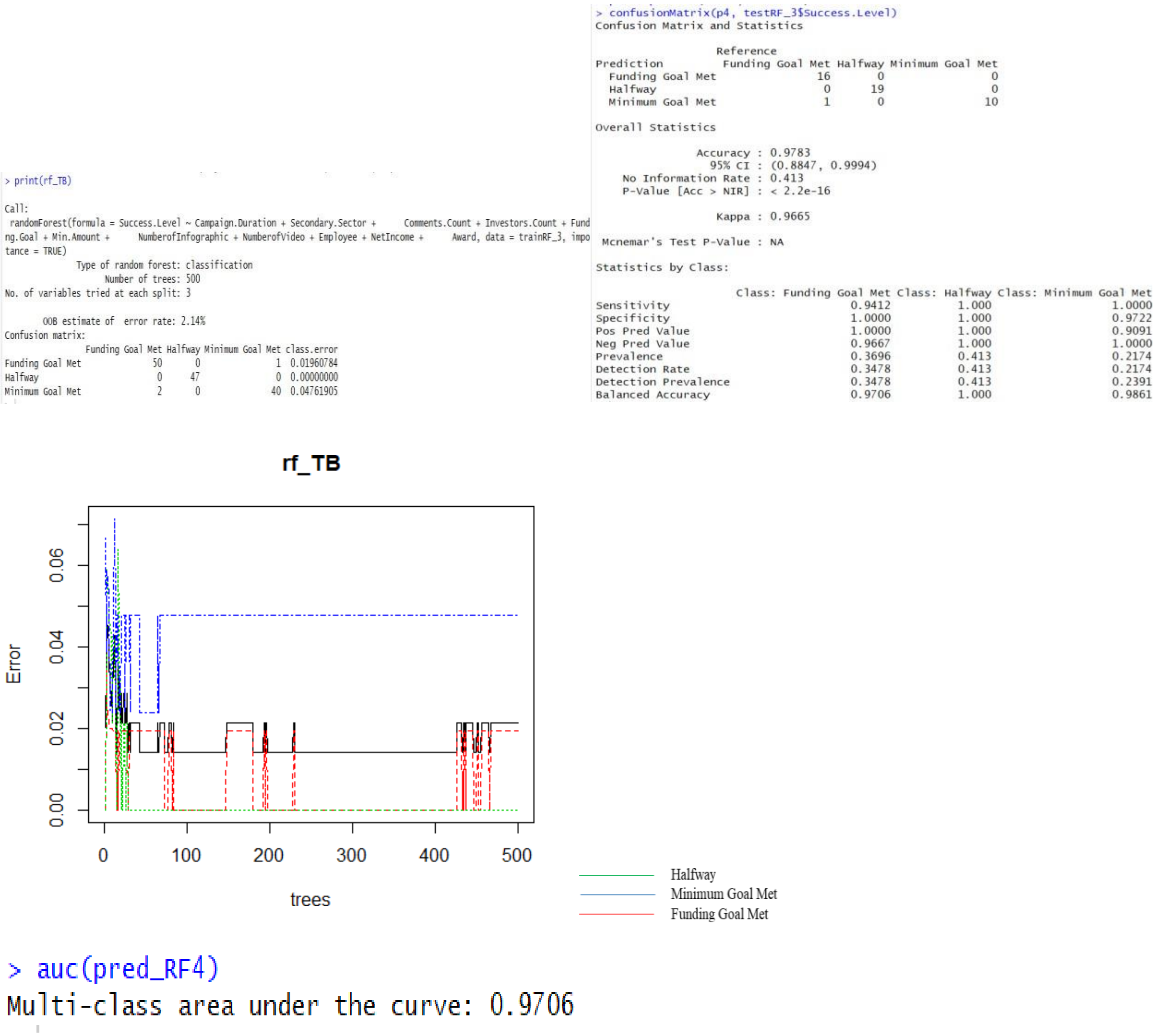

```
> auc(pred_RF3)
Multi-class area under the curve: 1
```

Figure 89: Random Forest - Experiment 3 Output

By using 500 trees with 6 variables used at each split, the accuracy of the train data has increase to 99% and the test data with accuracy of 100%. Here, this dataset with the Random Forest technique has overfitted the dataset. Thus, next we model the same with the selected features from the SMOTE Tree Based Algorithm.

Experiment 4: SMOTE dataset with Tree-Based Algorithm features

The SMOTE dataset would be used to run the Random Forest model with random search. However, the features that was determined important from the SMOTE tree-based algorithm would be used for modelling, output as illustrated below.



```
> print(rf_TB)

Call:
 randomForest(formula = Success.Level ~ Campaign.Duration + Secondary.Sector +      Comments.Count + Investors.Count + Fund
ng.Goal + Min.Amount +      NumberofInfographic + NumberofVideo + Employee + NetIncome +      Award, data = trainRF_3, impo
tance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 2.14%
Confusion matrix:
                 Funding Goal Met Halfway Minimum Goal Met class.error
Funding Goal Met               50       0                1  0.01960784
Halfway                         0      47                0  0.00000000
Minimum Goal Met                2       0               40  0.04761905
```

```
> confusionMatrix(p4, testRF_3$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction         Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               16       0                0
  Halfway                         0      19                0
  Minimum Goal Met                1       0               10

Overall Statistics
                                          
               Accuracy : 0.9783          
                 95% CI : (0.8847, 0.9994)
    No Information Rate : 0.413           
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.9665          
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.9412          1.000                  1.0000
Specificity                           1.0000          1.000                  0.9722
Pos Pred Value                        1.0000          1.000                  0.9091
Neg Pred Value                        0.9667          1.000                  1.0000
Prevalence                            0.3696          0.413                  0.2174
Detection Rate                        0.3478          0.413                  0.2174
Detection Prevalence                  0.3478          0.413                  0.2391
Balanced Accuracy                     0.9706          1.000                  0.9861
```

```
> auc(pred_RF4)
Multi-class area under the curve: 0.9706
```

Figure 90: Random Forest - Experiment 4 Output

Similar as previous experiment, 500 trees with 3 variables used as each split had brought the train and test accuracy to 98% with AUC of 97%. Noted there is a misclassification for the minimum goal met in the training dataset and for testing only funding goal met has misclassified scenario. The importance of the variables was observed as below.

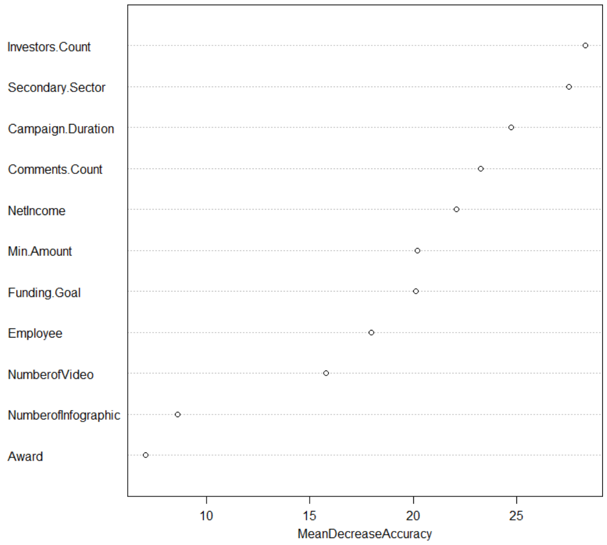


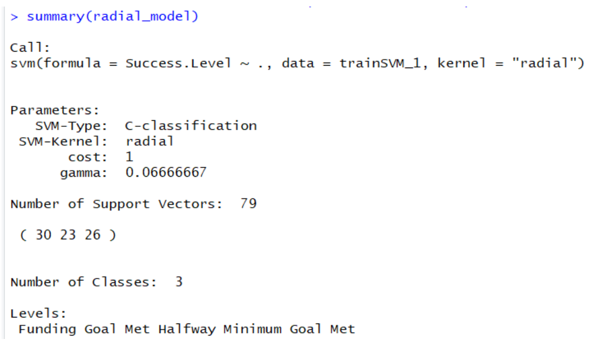
Figure 91: Variables of Importance Random Forest Experiment 4 Output

Noted there is difference in importance of the variables where here the top 5 important variables are Investor Count, Secondary Sector, Campaign Duration, Comments Count and Net Income where we noted a very high mean decrease accuracy was observed the first 2 variables could reduce close to 50% of the accuracy in the dataset. As Random Forest Model tends to create an overfitting mode, Support Vector Machine (SVM) was explored as it supports a high dimension dataset with low number of observations.

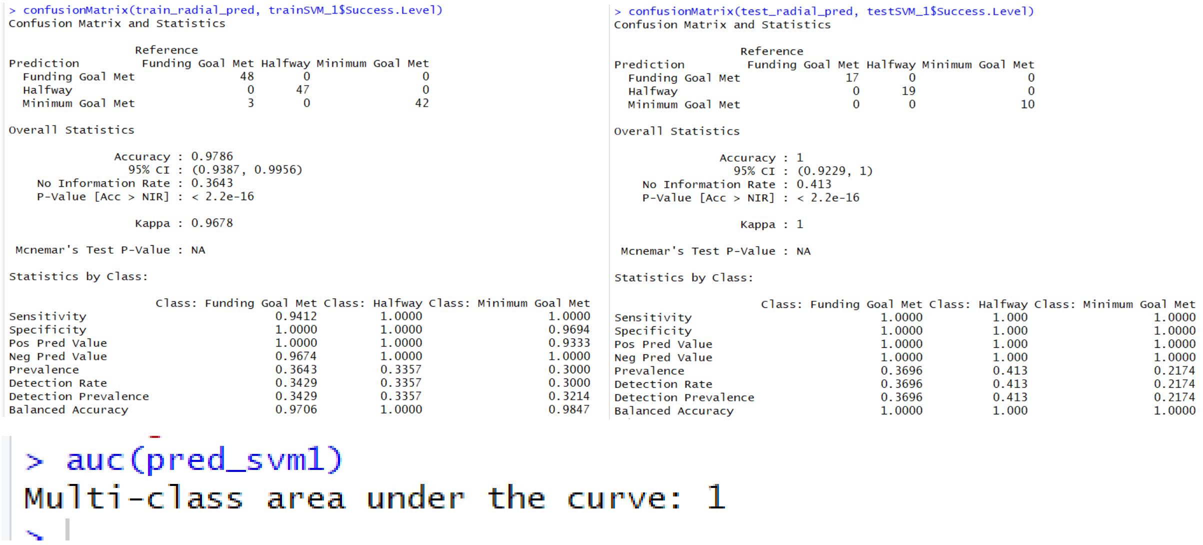
4.3.6.3 Support Vector Machine

Experiment 1: SMOTE dataset with SMOTE Tree Based Features on Kernel = Radial

The SMOTE dataset with the SMOTE Tree Based Features was used where the kernel was set to Radial, the output as highlighted below.



```
> summary(radial_model)

Call:
svm(formula = Success.Level ~ ., data = trainSVM_1, kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.06666667

Number of Support Vectors:  79

 ( 30 23 26 )


Number of Classes:  3

Levels:
 Funding Goal Met Halfway Minimum Goal Met
```



```
> confusionMatrix(train_radial_pred, trainSVM_1$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                48       0                0
  Halfway                          0      47                0
  Minimum Goal Met                 3       0               42

Overall Statistics

               Accuracy : 0.9786
                 95% CI : (0.9387, 0.9956)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9678

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.9412         1.0000                  1.0000
Specificity                           1.0000         1.0000                  0.9694
Pos Pred Value                        1.0000         1.0000                  0.9333
Neg Pred Value                        0.9674         1.0000                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.3429         0.3357                  0.3000
Detection Prevalence                  0.3429         0.3357                  0.3214
Balanced Accuracy                     0.9706         1.0000                  0.9847
```

```
> confusionMatrix(test_radial_pred, testSVM_1$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                17       0                0
  Halfway                          0      19                0
  Minimum Goal Met                 0       0               10

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9229, 1)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           1.0000          1.000                  1.0000
Specificity                           1.0000          1.000                  1.0000
Pos Pred Value                        1.0000          1.000                  1.0000
Neg Pred Value                        1.0000          1.000                  1.0000
Prevalence                            0.3696          0.413                  0.2174
Detection Rate                        0.3696          0.413                  0.2174
Detection Prevalence                  0.3696          0.413                  0.2174
Balanced Accuracy                     1.0000          1.000                  1.0000
```

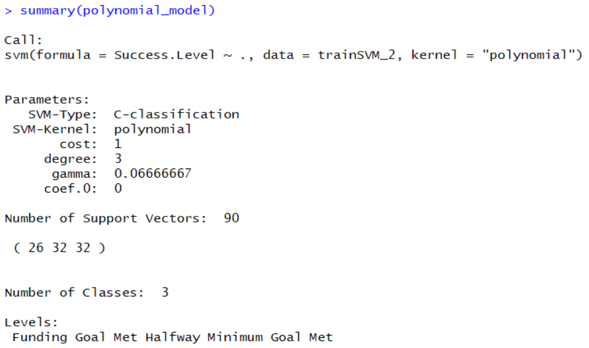
```
> auc(pred_svm1)
Multi-class area under the curve: 1
>
```

Figure 92: SVM - Experiment 1 Output

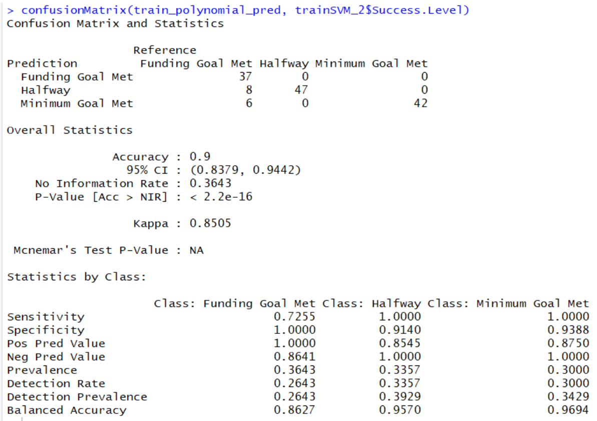
As the RBF/Radial kernel would assist in dividing a non-linear decision boundary, noted an overfitted model had been produced. 79 support vectors was formed to with a training accuracy of 98% and testing accuracy of 100%. This, we have been able to fit all the test samples however making this model rather a very rigid model. Next, we would experiment with the polynomial kernel which works on similar nonlinear dataset.

Experiment 2: SMOTE dataset with SMOTE Tree Based Features on Kernel = Polynomial

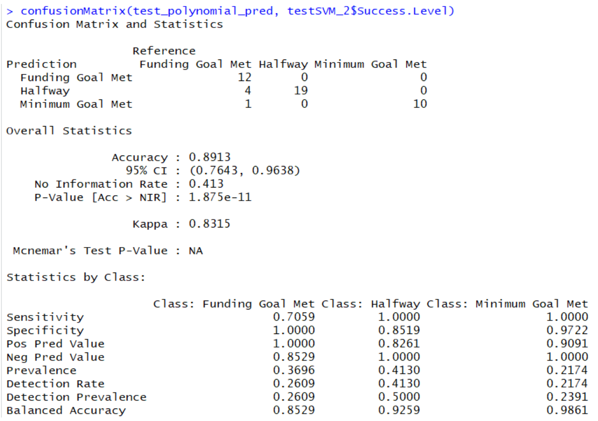
The SMOTE dataset with the SMOTE Tree Based Features was used where the kernel was set to polynomial, the output as highlighted below.



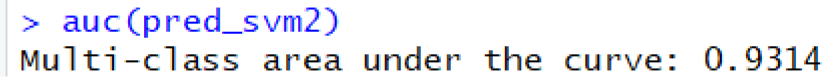
```
> summary(polynomial_model)

Call:
svm(formula = Success.Level ~ ., data = trainSVM_2, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.06666667
     coef.0:  0

Number of Support Vectors:  90

 ( 26 32 32 )


Number of Classes:  3

Levels:
 Funding Goal Met Halfway Minimum Goal Met
```



```
> confusionMatrix(train_polynomial_pred, trainSVM_2$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               37       0                0
  Halfway                         8      47                0
  Minimum Goal Met                6       0               42

Overall Statistics

               Accuracy : 0.9
                 95% CI : (0.8379, 0.9442)
    No Information Rate : 0.3643
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8505

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7255         1.0000                  1.0000
Specificity                           1.0000         0.9140                  0.9388
Pos Pred Value                        1.0000         0.8545                  0.8750
Neg Pred Value                        0.8641         1.0000                  1.0000
Prevalence                            0.3643         0.3357                  0.3000
Detection Rate                        0.2643         0.3357                  0.3000
Detection Prevalence                  0.2643         0.3929                  0.3429
Balanced Accuracy                     0.8627         0.9570                  0.9694
```



```
> confusionMatrix(test_polynomial_pred, testSVM_2$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met               12       0                0
  Halfway                         4      19                0
  Minimum Goal Met                1       0               10

Overall Statistics

               Accuracy : 0.8913
                 95% CI : (0.7643, 0.9638)
    No Information Rate : 0.413
    P-Value [Acc > NIR] : 1.875e-11

                  Kappa : 0.8315

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           0.7059         1.0000                  1.0000
Specificity                           1.0000         0.8519                  0.9722
Pos Pred Value                        1.0000         0.8261                  0.9091
Neg Pred Value                        0.8529         1.0000                  1.0000
Prevalence                            0.3696         0.4130                  0.2174
Detection Rate                        0.2609         0.4130                  0.2174
Detection Prevalence                  0.2609         0.5000                  0.2391
Balanced Accuracy                     0.8529         0.9259                  0.9861
```



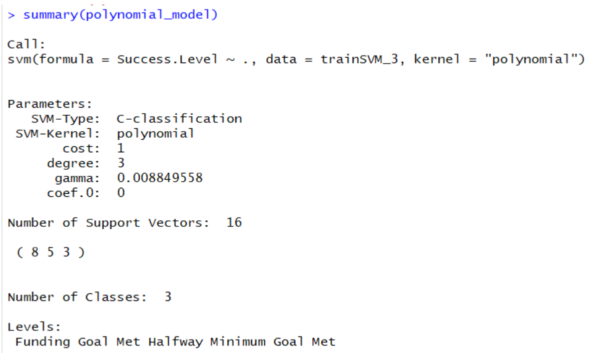
```
> auc(pred_svm2)
Multi-class area under the curve: 0.9314
```

Figure 93: SVM - Experiment 2 Output

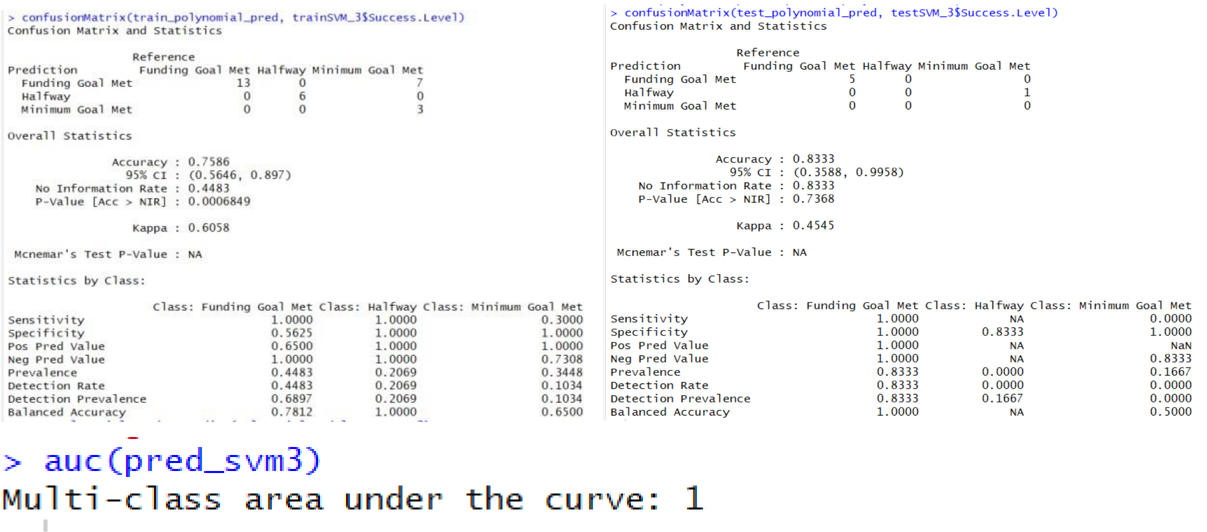
The output of the polynomial model rather provided a better outcome of a good fit model. Here, the training and test accuracy was at 90% and 89% with an AUC of 93% indicating a good fit model with a capability to separate the scenarios accordingly. Here, noted that 90 vectors was used to form this separation. Like all the previous models, funding goal met has a challenge with these features to distinguish the variable thus that having a higher misclassification and these features could clearly distinguish minimum goal met and halfway success level. Then, this polynomial kernel model was used against the normalised original dataset to observe the predicting capability for that dataset.

Experiment 3: Original dataset with all features on Kernel = Polynomial

The original dataset with the SMOTE Tree Based Features was used where the kernel was set to polynomial, the output as highlighted below.



```
> summary(polynomial_model)

Call:
svm(formula = Success.Level ~ ., data = trainSVM_3, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
      gamma:  0.008849558
     coef.0:  0

Number of Support Vectors:  16

 ( 8 5 3 )


Number of Classes:  3

Levels:
 Funding Goal Met Halfway Minimum Goal Met
```



```
> confusionMatrix(train_polynomial_pred, trainSVM_3$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                13       0                7
  Halfway                          0       6                0
  Minimum Goal Met                 0       0                3

Overall Statistics

               Accuracy : 0.7586
                 95% CI : (0.5646, 0.897)
    No Information Rate : 0.4483
    P-Value [Acc > NIR] : 0.0006849

                  Kappa : 0.6058

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           1.0000         1.0000                  0.3000
Specificity                           0.5625         1.0000                  1.0000
Pos Pred Value                        0.6500         1.0000                  1.0000
Neg Pred Value                        1.0000         1.0000                  0.7308
Prevalence                            0.4483         0.2069                  0.3448
Detection Rate                        0.4483         0.2069                  0.1034
Detection Prevalence                  0.6897         0.2069                  0.1034
Balanced Accuracy                     0.7812         1.0000                  0.6500
```

```
> confusionMatrix(test_polynomial_pred, testSVM_3$Success.Level)
Confusion Matrix and Statistics

                  Reference
Prediction          Funding Goal Met Halfway Minimum Goal Met
  Funding Goal Met                 5       0                0
  Halfway                          0       0                1
  Minimum Goal Met                 0       0                0

Overall Statistics

               Accuracy : 0.8333
                 95% CI : (0.3588, 0.9958)
    No Information Rate : 0.8333
    P-Value [Acc > NIR] : 0.7368

                  Kappa : 0.4545

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Funding Goal Met Class: Halfway Class: Minimum Goal Met
Sensitivity                           1.0000             NA                  0.0000
Specificity                           1.0000         0.8333                  1.0000
Pos Pred Value                        1.0000             NA                     NaN
Neg Pred Value                        1.0000             NA                  0.8333
Prevalence                            0.8333         0.0000                  0.1667
Detection Rate                        0.8333         0.0000                  0.0000
Detection Prevalence                  0.8333         0.1667                  0.0000
Balanced Accuracy                     1.0000             NA                  0.5000
```

```
> auc(pred_svm3)
Multi-class area under the curve: 1
```

Figure 94: SVM - Experiment 3 Output

Here, as its 35 observations only 16 support vectors was required to perform the separations among the attribute. Training had produced 76% of accuracy where noted that the minimum goal met could not be clearly separated by this model. Thus, the effect of the misclassification of 1 of the test observations for minimum goal met making the accuracy of test at 83%. Despite that, all funding goal met could be clearly distinguish expect some being wrongly classified as funding goal met. Then, the same dataset with the SMOTE Tree based features was used where noted a much lower training accuracy was obtained indication of poor learning despite the test accuracy is high at 83%. Thus, concluding the use of selected features in a small dataset tends to limit the learning where when all features used a better result was obtained. However, still there is a tendency of overfitting the model and this is due to the small observations available. However, with the SMOTE dataset together with the selected features a better classification of the model was able to achieve without overfitting the model.

4.3.7 Discussion on findings

The outcome of the feature engineering and predictive modelling has been obtained to address the following objectives:

- Objective 2: To examine the success factors of an equity crowdfunding campaign
- Objective 3: To predict the funding level of successful campaign based on the success campaign factors

Success Factors

Two different feature selection method was used, namely Tree Based Algorithm and Boruta which was performed on both the original and SMOTE dataset. This was to identify if there was different variables suggested by the different techniques and datasets. Table 18 highlights on the list of important variables determined by each of the techniques and dataset.

Table 18: Feature Selection Output

Dataset	**Original**		**SMOTE**
Variable	**Tree Based Algorithm**	**Boruta**	**Tree Based Algorithm**
Minimum Amount Goal	1	5	10
Idea Length	2	-	-
Investor Count	3	1	1
Funding Goal	4	4	6
Campaign Duration	5	2	2
Minimum Shares Issues	6	-	-
Over Subscription	7	-	-
Secondary Sector	8	-	3
Comments Count	-	3	4
Number of Video	-	-	5
Net Income	-	-	7
Employee	-	-	8
Award	-	-	9
Number of Infographic	-	-	11

The Boruta for the SMOTE dataset returned almost all variables as important, thus being excluded from this summary. Minimum amount goal, Investor Count, Funding Goal and Campaign Duration has been picked by all the techniques and dataset as being an important variable that could distinguish the success levels. The top five attributes picked from the original dataset with Tree Based Algorithm is Minimum Goal, Idea Length, Investor Count, Funding Goal and Campaign Duration. However, Boruta technique with the same dataset discarded idea length but included comments count.

On the other hand, SMOTE dataset resulted on additional attributes as their top five important variables that are secondary sector and number of videos whereby minimum goal being only at the tenth rank of important variables. Number of Infographic, Employee, Award and Net

Income are additional variables deemed as the factors of distinguishing the success level with the SMOTE dataset. Here, we could conclude with more dataset there are more learning with variation thus the additional attributes identified to influence the success level and vice versa. Idea length, minimum shares issued, and oversubscription identified as important variable with the original dataset using the Tree Based Technique but with the SMOTE dataset as well as Boruta, this has been flagged as not important.

As we have identified the list of variables of importance, this is basically the success factors that influences the success level. Thus, a mapping of the variable to the success signals as highlighted in Section 2.3 was performed. Additionally, as the success level is divided into three different levels, we further utilised the correlation matrix in Section 4.3.2.4 to extract which of these attributes are strongly correlated to the relevant variables. Table 19 highlights the grouping of the variables to the success signals as well as its correlation to the success levels.

Table 19: ECF Success Factors/Signals and its correlation to success level

Success Signals	**Variable**	**Correlation with Success Level**		
		Funding Goal Met	**Halfway**	**Minimum Goal Met**
Campaign Financials	Minimum Amount Goal			Negative
	Funding Goal		Positive	
Time	Campaign Duration	Negative		Positive
Campaign Type/Industry	Secondary Sector	Positive	Negative	
Equity Retention	Minimum Shares Issues			
	Over Subscription - True	Negative		Positive
Communication	Comments Count		Positive	
	Idea Length	Negative	Positive	Positive
	Number of Video	Negative	Positive	
	Number of Infographic	Negative		Positive
Company financial profile	Net Income	Negative	Negative	Positive
Human Capital	Employee	Positive		Negative
	Investor Count		Positive	Negative
Third Party	Award - Yes	Positive	Negative	

The variables of importance by their signals Campaign Financials, Campaign Industry/Type, Equity Retention, Communication, Company Financial Profile, Human Capital and Third Party to be the influencing factors that determine the success level of the campaigns. The signal from a campaign financials are the minimum amount goal and funding goal and a time element of the campaign duration as well as the campaign industry does influence the success level of a campaign.

Then, the correlation between the attributes to the success level was observed. From the correlation table, noted that minimum amount has a negative correlation with minimum goal met, the higher the minimum amount set the chances to meet the other success level is higher. For Halfway, Investor Count and Funding Goal has a positive correlation thus the more investors invest, most projects would at least have meet halfway but not necessarily meeting the funding goal. Also, the higher the funding goal the chances to meet at least halfway of the funding not necessarily full funding goal met. Secondary sector negatively correlates to halfway success level indicating that sector has no influence on the halfway but positively influences the funding goal met. Campaign duration noted a positive correlation with minimum goal met where the more campaign days involve for the minimum funding the more collections could be made by those campaigns while this has negative effect on funding goal me due to certain campaign achieving its funding goal at day 1 itself.

Additionally, equity retention such as the minimum shares issued and if over subscription is allowed for a campaign also determines the success level. The minimum shares issues does not have any strong relationship with the success level however for campaigns that allows oversubscription, noted that it would be able to meet the minimum funding goal. Communication signals such as idea length, comments count, number of video and number of infographics does segregate the success level thus it being signals of success pertaining to this dataset. Comments count could distinguish halfway success level where it has a positive correlation where more comments would at least meet halfway but not necessarily get the full fund. Number of Video and Infographic surprisingly has a negative correlation with funding goal met thus highlight no influence of communication signal in determining the success level. Idea length on the other hand has a negative correlation to funding goal met indicating that the more information captured the least of funding goal would be met.

Investor’s competence to observe the net income to determine the success level, where it has negative correlation with funding goal met and halfway indicating that higher net income does not determine higher success level but at least meeting the minimum goal. From a human capital perspective, the number of Employee influences for achieving a higher success level where it positively correlated with funding goal met and negatively correlates to minimum goal met. Lastly, receiving awards which is third party signals too influences a higher success level and distinguishes that to the halfway meet up point.

Prediction of the Success Level

Naïve Bayes and Random Forest machine learning techniques was used to build the prediction model. The dataset variation with the original dataset of 35 observations and the SMOTE dataset with 186 observations was used for this experiment. In addition, variation of experimentation based on sampling techniques, different features and optimisation with parameter tuning was performed to obtain the best model.

Stratified and Random Sampling with Naïve Bayes was performed where the outcome of the random sampling produced a better accuracy and AUC value compared to the stratified sampling. The stratified sampling has cause overfitting where the training accuracy was at 92% and test accuracy dropped to 42% with the original dataset (Exp 3). But overfitting did not occur when stratified sampling was done on SMOTE dataset (Exp 5) where, an AUC of 88% was achieved. However, the SMOTE dataset’s baseline model with all variables performed better with an AUC of 95%. Thus, confirming further that stratified sampling does not improve the model accuracy moreover with small observation scenario, this may just cause overfitting.

Next, the Laplace smoothing technique was performed to improve the Naïve Bayes model. Here, as we have many variables and few has zero value, tendency for discarding that pattern becomes high due to the zero value. Here, by incorporating Laplace the zero value to be given one, thus a more flexible model would be achieved and an expectation to have a drop in the model accuracy. Noted, the original dataset training accuracy dropped to 72% but test accuracy remained at 83%, here concluding that all patterns are learned thus a better test accuracy and AUC remained at 90%. Similar outcome was observed in the SMOTE dataset as well but with a higher test accuracy of 87% and AUC at 92%. This indicating a good fit model is achieved where more leaning could be performed with an increased test accuracy. The SMOTE dataset was then used to perform the feature modelling as the small number of observations on the original dataset tends to make the model overfit.

The selected features from the Tree Based Algorithm (original and SMOTE) and Boruta was used to perform the modelling with the SMOTE dataset incorporating Laplace smoothing. Here, noted that the SMOTE Tree Based list of features had produced the highest training and test accuracy of 86% and 87% respectively (Exp 9a) where Original Tree Based list of features has almost similar accuracy and AUC that is 1% higher (Exp 7). However, as the SMOTE Tree Based features incorporates all except 3 features (Idea Length, Minimum Shares Issue and Over Subscription) from the original tree-based features, further experiment (Exp 9b) was performed combining all the features from the SMOTE and Original tree based to only produce a model that is overfitted where a low test accuracy of 77% was obtained (Exp 9b) . Thus, here we could conclude that the SMOTE Tree based features produces a good fit model with a better accuracy and AUC value.

Then, as the feature selection technique drops the features thus dimensional reduction technique does not drop the features rather group the similar attributes to similar principal components with the factor analysis techniques. Here, 10 principal components was identified to be the optimal number of factors with eigenvalue beyond 1. Thus, both the original and SMOTE dataset was modelled using the factor analysis outcome where AUC of 90% and 88% was obtained for both datasets respectively. This is an acceptable outcome as AUC is 3% lower than the model with SMOTE Tree based features and by retaining all the variables and grouping them by relevant principle, we could obtain almost a similar accuracy rate as the feature selection model. However, when this is used with the smaller dataset (original) noted a lower accuracy achieved but it is expected as a flexible model is produced.

Then, Random Forest was used to model this dataset by experimenting with the original dataset, SMOTE dataset and with the variation of features. However, due to the small number of observations with the original dataset the model tends to underfit thus produced a very low accuracy rate at 50% and AUC of 30% (Exp 1a). Upon performing grid search to optimise the model, slight improvement on the but it overfitted the model. Next, Random Forest was performed on the SMOTE data as it has a bigger number of observations, we expect to get a better accuracy rate. Experimenting with all attributes overfitted the model by producing 100% of accuracy and AUC similarly when performed with the SMOTE Tree Based features. Slight drop in accuracy was observed when using the Original Tree Based features however this may also deem to be an overfitted model.

The SVM model was built with both the dataset, with radial and polynomial kernel as well as all features, from SMOTE Tree Based and the Factor Analysis. Here, noted that the SMOTE tree-based features with the SMOTE dataset had produced a good fit model with the capability to segregate the success level and avoided overfitting using the polynomial kernel. Similar outcome was achieved with the factor analysis features. The radial kernel was causing an overfitted model with 100% of test accuracy and AUC. The original dataset also produced an overfitted model when all features was used where more error rate was observed especially with minimum funding goal met compared to other models where this was significantly classifiable. Also, when the SMOTE tree-based features with the original dataset was modelled, a lower learning rate was observed.

The SMOTE Tree Based features or the factor analysis principal components being a good set of features with capability to distinguish the success level. However, the SMOTE Tree Based features only works with large observation dataset where with small dataset it tends to reduce the learning rate thus a lower accuracy achieved. However, the factor analysis features in all scenarios had produce a reasonable outcome when tested with both the SMOTE as well as the original dataset. Additionally, in term of the best model to predict this dataset the Naïve Bayes model with Laplace smoothing and SVM with polynomial kernel has produced the best accuracy rate with reasonably high AUC value. Thus, concludes the SMOTE Tree Based features or the Principal Components from the Factor Analysis to be producing the best good fit prediction model taking into consideration of the low volume of observations. Table 20 summarises the output of the experimentation performed.

Table 20: Predictive Modelling Experiment and Outputs

Experiment	Sampling	Normalised	SMOTE	# of Obs	Features	Parameter Tuning	Evaluation		
							Train	Test	
							Accuracy	Accuracy	AUC
Naïve Bayes									
1	Random	N	N	35	All	N	0.9655	0.8333	0.9000
2	Random	Y	N	35	All	N	0.9655	0.6667	0.8000
3	Stratified	N	N	35	All	N	0.9200	0.4000	0.6500
4	Random	Y	Y	186	All	N	0.9000	0.8913	0.9510
5	Stratified	Y	Y	186	All	N	0.8692	0.8750	0.8750
6a	Random	Y	Y	186	All	Laplace	0.8643	0.8696	0.9216
6b	Random	N	N	35	All	Laplace	0.7241	0.8333	0.9000
7	Random	Y	Y	186	Original Tree Based	Laplace	0.8429	0.8696	0.9098
8	Random	Y	Y	186	Original Boruta	Laplace	0.7214	0.6739	0.6901
9a	Random	Y	Y	186	SMOTE Tree Based	Laplace	0.8643	0.8696	0.9020
9b	Random	Y	Y	186	Original & SMOTE Tree Based	Laplace	0.8642	0.7679	0.7327
10	Random	Y	Y	186	Factor Analysis	Laplace	0.9000	0.8913	0.8756
11	Random	N	N	35	Factor Analysis	Laplace	0.7931	0.6667	0.9000
Random Forest									
1a	Random	N	N	35	All	Random Search	0.5172	0.5000	0.3000

1b	Random	Y	N	35	All	Random Search	0.3793	0.5000	0.3000
1c	Random	N	N	35	All	Grid Search	0.4333	0.6667	1.0000
2	Random	N	N	35	Original Tree Based	Random Search	0.6897	0.8333	1.0000
3a	Random	Y	Y	186	All	Random Search	0.9857	1.0000	1.0000
3b	Random	Y	Y	186	SMOTE Tree Based	Random Search	0.9857	1.0000	1.0000
4	Random	Y	Y	186	Original Tree Based	Random Search	0.9786	0.9783	0.9706
Support Vector Machine (SVM)									
1	Random	Y	Y	186	SMOTE Tree Based	Radial	0.9786	1.0000	1.0000
2	Random	Y	Y	186	SMOTE Tree Based	Polynomial	0.9000	0.8913	0.9314
3	Random	Y	N	35	All	Polynomial	0.7586	0.8333	1.0000
3b	Random	Y	N	35	SMOTE Tree Based	Polynomial	0.5517	0.8333	0.5000
4	Random	Y	Y	186	Factor Analysis	Polynomial	0.9077	0.8571	0.9007
4b	Random	Y	N	35	Factor Analysis	Polynomial	0.7241	0.8333	0.5000

CHAPTER 5

CONCLUSION AND RECOMMENDATION

Rapid enhancement in technology and the industrial revolution is driving organisations to adapt to this revolution. Similarly, financial industries have many fintech technologies blooming and disrupting the traditional mean of financial services. Thus, many financial providers besides commercial bank such as financing crowdfunding platforms are exploring the digital transformation of their services. In Malaysia, as we observed, the growth of SME being the primary source of business start-up, thus driving the growth of ECF platforms in Malaysia to finance these SME's. As more ECF platforms evolve in Malaysia, staying relevant and retaining their investors has become a challenge. Ensuring relevant campaigns that fit the requirement of the investors within the platform has become crucial while educating the SME's on factors that drive the success of a campaign. Additionally, based on the works of literature, minimal literature explained the investor's behaviours, especially in the ECF platform in Asia.

Therefore, this study was conducted on a leading ECF platform in Malaysia to identify the investor taxonomy, the factors that drive the success of the campaigns as well as predictive capability of the success level of the successful campaign. The hierarchical and k-means clustering techniques was used to identify the segmentation of investors. Then, the feature selection techniques using a Tree-based Algorithm and Boruta to determine the features that could distinguish between each success level was conducted. Boruta is a random forest technique while Tree-based algorithm is a base of decision tree. Subsequently, predicting the success level with three different machine learning techniques, namely Naïve Bayes, Random Forest and Support Vector Machine, was performed. The predictive modelling had several variations of experiments from a dataset variation to sampling variation and optimisation of the model with parameter tuning was performed.

5.1 Objective 1: Investors Taxonomy

The investor clustering focused on few variables as suggested by previous literature where six clusters were able to distinguish the investors, namely the Active Casual, Altruistic Common, Altruistic Casual, Trend Follower, Altruistic Sophisticated and Sophisticated investors. Each of these clusters has different motivations that have driven their profiling as such. Three identified groups have a strong altruistic element where they encourage innovation-driven project that has a patent, new start-up, new market penetration as well as a sole service provider in the market. Here, we noted these elements drives majority investors. In contra, we also have majority investors who may deem to be trend followers where they are influenced by the majority crowd while investing an only a small amount. Then, another sophisticated group that is not altruism driven, been identified to be motivated by financials as well as a liking nature where they have only invested once, but a considerable large amount invested. Here, we noted that a more family and friend who is encouraging their circle of friend's campaign. The active casual investors are those who are solely financially motivated, where their intention is on minimal risk by investing less in many projects. Here, we noted this platform is having a high number of altruism driven individuals who loves to encourage new innovative ventures.

5.2 Objective 2: Factors influencing the success level

Next, the feature selection technique has concluded fourteen features to be able to distinguish between the success level. For a campaign to achieve the funding goal met, sector, the employee and reward received has influenced a campaign to obtain the full funding. For those who have collected halfway, the comments count, video, funding goal and investor count positively influence the investment to halfway. Campaign duration, oversubscription allowed, number of infographics and net income influence a campaign to achieve the minimum goal.

5.3 Objective 3: Prediction of Success Level

From the Predictive Modelling, we have observed an overfitting scenario, especially with the Random Forest modelling. But Naïve Bayes and SVM provided good fit models. Naïve Bayes using features of SMOTE Tree Based and Factor Analysis with AUC of 90% and 88%, respectively. SVM on the other hand has a prediction capability of 93% with SMOTE Tree based features. Here concludes that the success level of the successful campaign could be predicted with these two models.

5.4 Implications

Similar to the Goethner, Luettig and Regner (2018) which performed the k-means technique identified three relevant clusters but this study, four groups were suggested as the optimal cluster however upon the review of the dissimilarity of the bunch, six clusters model was selected as it could distinguish the individuals better. Just like Wallmeroth (2019), the Sophisticated investors in this platform also has user who invest less but in large amount with less likelihood of returning. Goethner, Luettig and Regner (2018) highlighted a Sophisticated group consist of small number of individuals that are socially motivated. This was a similar outcome in our study, where the Sophisticated Altruistic avatar have twelve individuals with altruistic nature. The Casual and Crowd group that was highlighted by Goethner, Luettig and Regner (2018) was also apparent in this study known as the Active Casual and Altruistic Casual. However, the Cluster 2 Common individuals are the biggest group of almost 800 investors where their character was not prominent, except their investment on innovation was higher compared to their total projects invested.

The trend followers are the second largest group of individuals where the significant behaviour is most of their investment was performed on campaigns with a high average number of investors. Additionally, they only invested less amount and a small number of investment, thus the risk-averse nature with herding behaviour similar to what seen by Lin, Boh and Goh (2014). The difference is his study is a reward-based crowdfunding platform, and this is an ECF platform. However, that is an indicator of a herding nature. In addition to Goethner, Luettig and Regner (2018) study who identified three different clusters, here we managed to identify six clusters with five who are prominent with a different motivation. Like Hornuf and Schmitt (2016), we identified a Sophisticated avatar with liking or personnel connection as their motivation.

On the ECF success factor, campaign financials, duration, industry, equity retention, communications, company financial profile, human capital, and third-party endorsement provides signals that determine the success level in this ECF platform. Just like Li et al. (2016), employee count does influence the success level in this platform. However, as highlighted by Piva and Rossi-Lamastra (2018) on the entrepreneurial experience in driving a successful campaign was not seen in our platform as that feature did not deem citical. This could be due to the difference between these two studies where Piva and Rossi-Lamastra (2018) focused on success and failure, whereby this study looks at success level. Additionally, Lin, Boh and Goh (2014) highlighted that the number of investors does provide signals for the success of a campaign, here we noted that more investors would drive more people to invest beyond the minimum level but not necessarily reaching the funding goal. As for how Goethner, Luettig and

Regner (2018) confirmed that the more information provided of the company financials would attract more investors, in this study net income, was the only variable that appeared necessary moreover it influences to meet the minimum goal rather achieving the funding goal.

As similar studies Lin, Boh and Goh (2014) and Goethner, Luettig and Regner (2018)performed a choice model using logistic regression. Our study has explored the machine learning technique, Naïve Bayes and Random Forest was the models used for a small observation scenario by Kamath and Kamat (2016). In our study, Naïve Bayes and SVM provided a good fit model in contra the Random Forest model tend to create a overfit model.

In terms of the business implication, this study had assisted the platform in identifying the data points required to perform segmentation as well as build a predicting model. The data structure and the features needed for the modelling was identified. Then, the investor taxonomy could assist the platform with specific decision making, such as:

- Most investors are altruism driven, when there is a new start-up or innovation-driven project, the selected avatars could be contacted for targeted promotion
- To identify schemes to encourage more investors to be active in the platform especially those with high potential of investing (Sophisticated Investors)
- Encourage more prominent investors to have a public profile thus increase more investment performed by trend followers
- Additionally, if any financial guidance to be provided especially for new investors or those risk-averse, the active casual and trend followers could be approached to provide investment guidance
- If advisory services were to be given to Entrepreneur, to encourage more innovation-driven opportunities taking into consideration of the preferred sector and the companies strength which is looked upon by investors

The success signals from the feature selection could assist the ECF platform to capture them as their operational checklist, where these attributes should be review when a campaign is being onboarded, especially on setting the minimum goal and funding goal. The predictive model could be embedded into their selection process, where this could assist them to prescribe if a campaign has a potential to meet which success level thus improving relevant resource allocation within the organisation.

5.5 Limitation and future directions

As we have concluded this study, the limitation of this study must be taken into consideration for future improvement or by the ECF platform. Firstly, on the campaign dataset where the failure of campaigns was not recorded by the platform, thus capturing the failure campaigns could improve the learning pattern where a clear distinction between the success and failure could be performed. Additionally, a small number of successful campaigns was also a limitation. Both this data limitation had indirectly limited the identification of the important variables where in this study it has identified the critical variables that could distinguish between the success level rather a success-failure scenario. Thus, if a success-failure scenario,the feature selection and predictive modelling require rebuilding.

Additionally, for the cluster analysis, there was one cluster despite performing a different variation of re-clustering, yet that cluster was not distinguishable. Thus, new features need to be added to achieve the clustering that could distinguish that group further. Also, as the predictive modelling techniques are advancing where now deep learning techniques are being explored with great optimisation options, however, due to the data limitation, this could not be performed in this study.

Lastly, as this study was to observe the investor behaviour of the ECF platform in Malaysia, this study could not be generalised as it only investigates one ECF platform in Malaysia. Thus, extending this study to all the other ECF platform could provide a better outcome of the ECF investor's investment behaviour as well as produce a more robust predictive model to predict the success of a campaign.

Reference

Abrams, E. (2017). Securities Crowdfunding: More than Family, Friends, and Fools?. *SSRN Electronic Journal.* pp. 1–45. doi: 10.2139/ssrn.2902217.

Adnan, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J. and Anwar, S (2018). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research.* Elsevier, 94 (October 2017). pp. 290–301. doi: 10.1016/j.jbusres.2018.03.003.

Aluri, A., Price, B. S. and McIntyre, N. H. (2019). Using Machine Learning To Cocreate Value Through Dynamic Customer Engagement In A Brand Loyalty Program. *Journal of Hospitality and Tourism Research*. 43(1). pp. 78–100. doi: 10.1177/1096348017753521.

Aprilia, Lady and Wibowo, S. S. (2018). The Impact of Social Capital on Crowdfunding Performance. *The South East Asian Journal of Management*. 11(1). pp. 44–57. doi: 10.21002/seam.v11i1.7737.

Asia Institute of Finance (2017). *Crowdfunding Malaysia's Sharing Economy.* [Online]. Available at: https://www.aif.org.my/clients/aif_d01/assets/multimediaMS/publication/AIF_ResearchReport_Crowdfunding.pdf.

Astebro, T.B., Fernández Sierra, M., Lovo, S. and Vulkan, N., (2017). Herding in equity crowdfunding. *Available at SSRN 3084140*. pp. 0–66. doi: 10.2139/ssrn.3084140.

BBC News (2013). *The Statue of Liberty and America's crowdfunding pioneer.* [Online]. Available from: https://www.bbc.com/news/magazine-21932675

Block, J., Hornuf, L. and Moritz, A., (2018). Which updates during an equity crowdfunding campaign increase crowd participation?. *Small Business Economics.* 50(1), pp.3-27. doi: 10.1007/s11187-017-9876-4.

Bretschneider, U., Leimeister, J.M. and Mathiassen, L., (2015). IT-enabled product innovation: Customer motivation for participating in virtual idea communities. *Int. J. Product Development*, *20*(2), pp.126-141.

Chen, X., Chen, X., De Vos, J., Lai, X. and Witlox, F. (2018). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*. Elsevier. 14(August 2018). pp. 1–10. doi: 10.1016/j.tbs.2018.09.002.

Cholakova, M. and Clarysse, B., (2015). Does the Possibility to Make Equity Investments in Crowdfunding Projects Crowd Out Reward–Based Investments?. *Entrepreneurship Theory and Practice*, 39(1), pp.145-172. doi: 10.1111/etap.12139.

Cumming, D.J., Leboeuf, G. and Schwienbacher, A., (2015). Crowdfunding models: Keep-it-all vs. all-or-nothing. *Financial Management*. pp. 1–41. doi: 10.1111/fima.12262.

Datanovia (2018) [Online]. Available from: https://www.datanovia.com/en/lessons/assessing-clustering-tendency/ [Accessed: 28 August 2019]

DOSM (2017) *Small and Medium Enterprises (SMEs) Performance 2017* [Online]. Available from: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=159&bul_id=cEI0bklpZHJaTlhRNDB3d2ozbnFIUT09&menu_id=TE5CRUZCblh4ZTZMODZIbmk2aWRRQT09 [Accessed: 03 March 2019]

Fintech News (2019). *How is Malaysia's Equity Crowdfunding Scene doing in 2019?*. [Online]. Available from: https://fintechnews.my/20194/crowdfunding-malaysia/equity-crowdfunding-report-2019/. [Accessed: 03 June 2019]

Gamal, D., Alfonse, M., M El-Horbaty, E.S. and M Salem, A.B. (2018). Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains. *Machine Learning and Knowledge Extraction.* 1(1). pp. 224–234. doi: 10.3390/make1010014.

Gerber, E. M. and Hui, J. (2013). Crowdfunding: Motivations and Deterrents for Participation. *ACM Transactions on Computing-Human Interaction*. 20(6). pp. 34:1–32. doi: 10.1145/2530540.

Goethner, M., Luettig, S. and Regner, T., (2018). Crowdinvesting in entrepreneurial projects: Disentangling patterns of investor behavior (No. 2018-018). *Jena Economic Research Papers.*

Guenther, C., Johan, S. and Schweizer, D., (2018). Is the crowd sensitive to distance?—How investment decisions differ by investor type. *Small Business Economics*, *50*(2), pp.289-305. doi: 10.1007/s11187-016-9834-6.

Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E., (2013). *Multivariate data analysis: Pearson new international edition.* Pearson Higher Ed.

Hakim Ghazali, N. (2019). Awareness and Perception Analysis of Small Medium Enterprise and Start-up Towards FinTech Instruments: Crowdfunding and Peer-to-Peer Lending in Malaysia. *International Journal of Finance and Banking Research*. 4(1). p. 13. doi: 10.11648/j.ijfbr.20180401.12.

Hervé, F., Manthé, E., Sannajust, A. and Schwienbacher, A., (2019). Determinants of individual investment decisions in investment-based crowdfunding. *Journal of Business Finance & Accounting*, *46*(5-6), pp.762-783. doi: 10.1111/jbfa.12372.

Hornuf, Lars; Schwienbacher, A. (2015) funding dynamics in crowdinvesting.

Hornuf and Schwienbacher (2017). *Crowdfunding How does it work ? When is it feasible ?'.*

pp. 1–6. Available at: http://www.undp.org/content/sdfinance/en/home/solutions/template-fiche12.html.

Hornuf, L. and Neuenkirch, M., (2017). Pricing shares in equity crowdfunding. *Small Business Economics*, *48*(4), pp.795-811. doi: 10.1007/s11187-016-9807-9.

Hornuf, L. and Schmitt, M., (2016). Does a Local Bias Exist in Equity Crowdfunding?. *Max Planck Institute for Innovation & Competition Research Paper*, (16-07). doi: 10.2139/ssrn.2801170.

Hornuf, L. and Schwienbacher, A. (2018). Market mechanisms and funding dynamics in equity crowdfunding. *Journal of Corporate Finance*. The Authors. 50. pp. 556–574. doi: 10.1016/j.jcorpfin.2017.08.009.

Kachamas, P., Akkaradamrongrat, S., Sinthupinyo, S. and Chandrachai, A., (2019). Application of Artificial Intelligent in the Prediction of Consumer Behavior from Facebook Posts Analysis. *International Journal of Machine Learning and Computing*. 9(1). pp. 91–97. doi: 10.18178/ijmlc.2019.9.1.770.

Kang, M., Gao, Y., Wang, T. and Zheng, H., (2016). Understanding the determinants of funders' investment intentions on crowdfunding platforms: A trust-based perspective. *Industrial Management & Data Systems*, *116*(8), pp.1800-1819.

Lantz, B. (2015). Machine learning with R. Packt Publishing Ltd.

Lee, I. and Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons*. Kelley School of Business, Indiana University. 61(1). pp. 35–46. doi: 10.1016/j.bushor.2017.09.003.

Li, X., Tang, Y., Yang, N., Ren, R., Zheng, H. and Zhou, H., (2016). The value of information disclosure and lead investor in equity-based crowdfunding: An exploratory empirical study. *Nankai Business Review International*, *7*(3), pp.301-321.

Li, Y., Rakesh, V. and Reddy, C.K., (2016). Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 247-256). ACM. doi: 10.1145/2835776.2835791.

Liao, Y., Tran, T., Lee, D. and Lee, K., (2017). June. Understanding Temporal Backing Patterns in Online Crowdfunding Communities. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 369-378). ACM. doi: 10.1145/3091478.3091480.

Lin, Y., Boh, W.F. and Goh, K.H., (2014). How different are crowdfunders. Examining archetypes of crowdfunders and their choice of projects. Retrieved November, 19, p.2014. Available at: http://ssrn.com/abstract=2397571.

Malaysia. pitchIn (2018). *pitchIn Equity Crowdfunding Report 2018*. Malaysia: pitchIn Malaysia.

Malaysia. Security Commissions (SC). Malaysia (2019) *Proposed Regulatory Framework for Property Crowdfunding*. Malaysia: SC Malaysia.

Mochkabadi, K. and Volkmann, C.K., (2018). Equity crowdfunding: a systematic review of the literature. *Small Business Economics*, pp.1-44. doi: 10.1007/s11187-018-0081-x.

Mohammadi, A. and Shafi, K. (2018). Gender differences in the contribution patterns of equity-crowdfunding investors. *Small Business Economics*. Small Business Economics. 50(2). pp. 275–287. doi: 10.1007/s11187-016-9825-7.

Moysidou, K. and Spaeth, S., (2016), September. Cognition, emotion and perceived values in crowdfunding decision making. In *Open and User Innovation Conference, Boston, USA*. 14th International Open and User Innovation, (July), pp. 1–11.

Nevin, S., Gleasure, R., O'Reilly, P., Feller, J., Li, S. and Cristoforo, J., (2017), August. Social identity and social media activities in equity crowdfunding. In *Proceedings of the 13th International Symposium on Open Collaboration* (p. 11). ACM. doi: 10.1145/3125433.3125461.

Nitani, M. and Riding, A., (2017), April. On Crowdfunding success: firm and owner attributes and social networking. In *2017 Emerging Trends in Entrepreneurial Finance Conference*. pp. 1–40.

Paschen, J. (2017). Choose wisely: Crowdfunding through the stages of the startup life cycle. *Business Horizons*. Kelley School of Business, Indiana University. 60(2), pp. 179–188. doi: 10.1016/j.bushor.2016.11.003.

Piva, E. and Rossi-Lamastra, C., (2018). Human capital signals and entrepreneurs' success in equity crowdfunding. *Small Business Economics*, *51*(3), pp.667-686. doi: 10.1007/s11187-017-9950-y.

Security Commissions Malaysia (2018). *ECF/P2P Analytics.* Available from: https://www.sc.com.my/analytics/ecfp2p. [Accessed: 07 March 2019]

SME Corp Malaysia (2016). *SME Statistics – Contribution of SMEs in 2016.* [Online]. Available from: http://www.SMEcorp.gov.my/index.php/en/policies/2015-12-21-09-09-49/SME-statistics. [Accessed: 03 March 2019]

SME Corp Malaysia (2015). *SMEs in RMKe-11 – Implication on SMEs* [Online]. Available from: http://www.SMEcorp.gov.my/index.php/en/policies/2015-12-21-09-26-24/rmke-11. [Accessed: 03 March 2019]

Tung, F.W. and Liu, X.Y., (2018), August. Understanding Backers' Motivations and Perceptions of Information on Product-Based Crowdfunding Platforms. In *2018 6th International Symposium on Computational and Business Intelligence (ISCBI)* (pp. 84-88). IEEE. pp. 84–88. doi: 10.1109/ISCBI.2018.00026.

UC (NA). *R Programming Guide: Hierarchical Cluster Analysis.* Available at: https://uc-r.github.io/hc_clustering

UNDP (United Nations Development Programme) (2017). *Crowdfunding How does it work ? When is it feasible ?'.* pp. 1–6. Available at: http://www.undp.org/content/sdfinance/en/home/solutions/template-fiche12.html.

Vismara, S., (2016). Information cascades among investors in equity crowdfunding. *Entrepreneurship Theory and Practice*. 42(3), pp. 467–497. doi: 10.1111/etap.12261.

Vismara, S. (2018). Sustainability in equity crowdfunding. *Technological Forecasting and Social Change*. (January). doi: 10.1016/j.techfore.2018.07.014.

Vulkan, N., Åstebro, T. and Sierra, M.F., (2016). Equity crowdfunding: A new phenomena. *Journal of Business Venturing Insights*, *5*, pp.37-49. doi: 10.1016/j.jbvi.2016.02.001.

Wallmeroth, J., (2019). Investor behavior in equity crowdfunding. *Venture Capital*, *21*(2-3), pp.273-300. doi: 10.1080/13691066.2018.1457475.

WorldBank (2013). Crowdfunding's potential for the developing world.

Xue, J. and Sun, F.F., (2016). August. Influencing factors of equity crowdfunding financing performance—An empirical study. In *2016 International Conference on Management Science and Engineering (ICMSE).* Rome, Italy: IEEE. (pp. 1353-1362). doi: 10.1016/j.tbs.2018.09.002.

Yu, P.F., Huang, F.M., Yang, C., Liu, Y.H., Li, Z.Y. and Tsai, C.H., (2018). Prediction of Crowdfunding Project Success with Deep Learning. *In 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE). China*. IEEE. (pp. 1-8).

Zunino, D., van Praag, M. and Dushnitsky, G., (2017). Badge of Honor or Scarlet Letter? Unpacking Investors' Judgment of Entrepreneurs' Past Failure. SSRN Electronic Journal. doi: 10.2139/ssrn.3041273.