Predictive Data Analytic Modeling

Carrie McDowell

Author Note

Carrie McDowell, Student at the Southern New Hampshire University

This research was to complete an assignment for DAT 640, Professor Genon

Correspondence concerning this article should be addressed to Carrie McDowell, Student, Southern New Hampshire University, New Hampshire, 03106

Contact: [carrie.mcdowell@snhu.edu](mailto:carrie.mcdowell@snhu.edu)

**Organizational Background**

ABC Insurance Company, (AIC) opened their first agency in 1956 in Chicago west side. Beginning only with home owner’s insurance, they branched out to renter’s insurance, automobile, and life insurance. AIC has 150 offices through the United States and employ more than 1,500 employees and provide insurance to more than 1 million individuals. The privately held company holds true to old time traditions of slow but steady growth with little risk. The newly appointed president, John Adams Jr., the great-great grandson of the original owner, has some forward-thinking ideas and wants to bring the company into this century.

Like changes to major medical, gone are the days of vehicle insurance coverage only for major accidents. A new concept of repair insurance is taking form and John wants to get ahead of the competition. John has asked that his marketing team perform analytics on their existing customer data base which is something never done. Their marketing efforts thus far has not changed in fifty years and no one has suggesting that they look at the data that they have collected over several years.

To make sense of all the data, John decided to open a new division for data analytics and hired a team of six to clean, model, and analyze the data. His goal is to one day be able to build a predictor model to predict a customer needs based off matching variables. This will not only allow his team to identify existing customers for the new product but will enable his team to target new customers as well.

**Data Set**

Currently, the data collected include 86 variables of various types with 5821 observations and no missing data. The data is held in a tab delimited data file. The data that will be used include the customers current products, age, marital status, demographics, household income and size, education level, employment type and income. The data captured at the time an insurance policy was written, has not been used for any real purpose. While the company is steady and true, competitors have benefited from leveraging data from their customers for a few years and may be slightly ahead of the game whereas AIC is just beginning to understand its value. The goal of the data miners is to understand and analyze the descriptive data and then create a prediction model using the CRISP-DM methodology. Once complete agents of each office and self-service customers will use the model to get the best mix of value-added services.

**CRISP-DM**

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology has six very important parts. The CRISP-DM process is a well-defined process and will be most helpful to the data team before testing and deployment of the predictor model. The process includes: 1) business understanding, 2) data understanding, 3) data preparation, 4) modeling, 5) evaluation, and 6) deployment.

Before data prep can proceed it is imperative to speak with all stakeholder and identify their objectives and then the understanding of the data itself. Data prep is the most important at this stage and may be revisited several times before an actual prediction model can be completed.

Once complete, the predictor model will add value to customers and the company alike. In addition, company directors will see value in capturing and analyzing customer data for other aspects of the business.

**Predictive Analytics**

Predictive analytics sets out to answer a business question and then predict outcome for the business based on learning patterns gleaned from historical data. Once the predictive question has been answered, we can build a model to predict results based on old customer data and apply it to new customer value-added sales. The predictive question will come from what will be done with the results (Parthasarathy, 2018). In my case, I will use the model to predict how likely new customers will benefit from and purchase automobile repair insurance. My task is to sell the product to my existing customer base, analyze the descriptive data, and then use an algorithm to predict the probability based on attributes of the customers who have adopted the new product and then build the prediction model to predict how likely a new customer will benefit from and purchase the new insurance. It is very important to spend an appropriate amount of time analyzing, cleaning, and formatting the descriptive data.

My question will be a forecasting/regression question, in that I would like to know how many new customers will purchase the new product, and therefore the business question I will answer is: “what common variables do customers share that have purchased this new type of automobile insurance”.

To answer the question, I will offer the new product to a random selection of existing customers. I will then use modeling techniques such as a decision tree to follow the customer’s course of action and then create a probability figure based on the results. Then I will determine which variables are seen in most customers who adopt the new insurance using a linear regression model and frequency chart. A regression algorithm is ideal to predict an outcome with many moving parts (Edwards, 2018).

**Where to Begin**

Before creating the predictive algorithm, I will first need to collect, transform, and clean the data using R. I will not use all data that is collected, so I will need to choose the right variables to answer the question. The time frame has been established, which is as soon as possible. I will lead a team to collect and transform the data, this step will take the longest as the results need to be accurate. Next, I will generate a baseline forecast which I will tweak until I find the right sequence of variables to include in the prediction model (Wilson, 2018). Once complete, I plan to use the model to predict new customer behavior. As time goes on, I will assign a team member to test the model each month for continued accuracy.

**Selecting a Predictive Algorithm**

The type of data, the complexity of the problem, the size of the data model and objectives will dictate the algorithm I use to create the predictive model (Zicari & Boweles, 2016). Data used is typically structured or unstructured.

Algorithms are developed to answer business problems, changing variables and running several scenarios will help to define the best algorithm to use for the problem statement. Additionally, choosing more than one algorithm called ensemble modeling, will increase the chances of success. Ensemble modeling runs two or more related but different models on the data to eliminate biases within the data (Bari, Chaouchi, & Jung). Using Ensemble models in sequence provides the optimum number of comparative variables using the same data set.

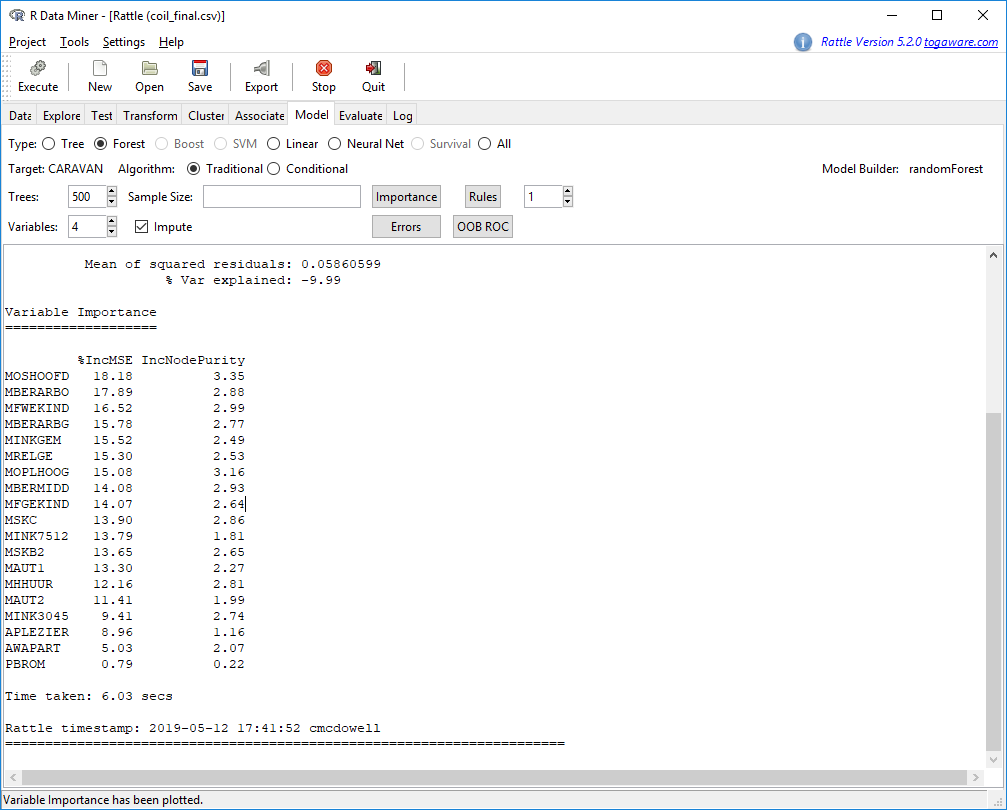
**Specifications**

AIC is pushing a new automobile repair insurance. Using historic and observed data we will answer the question, “what common variables do customers have that have purchased this new insurance”.

The data set itself is very large, so just 25% of existing customers will be contacted. The outcome of the offer will be binary with a scored result of one for yes and zero for no. AIC will select the appropriate variables from the account of those customers scoring one to continue to hone in on the best modeling techniques moving forward.

Data derived from descriptive analysis performed by the AIC marketing team is very structured in nature and can easily be classified into social categories. Using a clustering algorithm multiple times will show association among the customers. Additionally, assigning a weight to the variables identified can then be used in basic tree models. AIC will use a combination of clustering modeling, and forest modeling and then observe the outcome of a linear regression model to create a predictor model for future sales of automobile repair insurance.

I plan to use the forest model in R and will use the attribute CARAVAN as the target against the following input attributes which scored the highest frequencies.



|  |  |  |  |
| --- | --- | --- | --- |
| MOSHOOFD Customer main type | PBROM Contribution moped policies | MFWEKIND Household with children | MFGEKIND Household without children |
| MBERARBG Skilled laborer’s | MINKGEM Average income | MOPLHOOG High level education | APERSAUT Number of car policies |
| PPLEZIER Contribution boat policies | MAUT2 2 cars | MAUT1 1 car | MINK3045 Income 30-45.000 |
| MINK7512 Income 75-122.000 | MBERMIDD Middle management | APLEZIER Number of boat policies | MRELGE Married |
| MSKC Social class C | MSKB2 Social class B2 | MHHUUR Rented house | Caravan |

Aside from obeying the law, purchasing insurance is a gamble. The insurer is betting that you will not need to use it, and the insured is betting they will need it. Often, certain types of customers will fit into this mold.

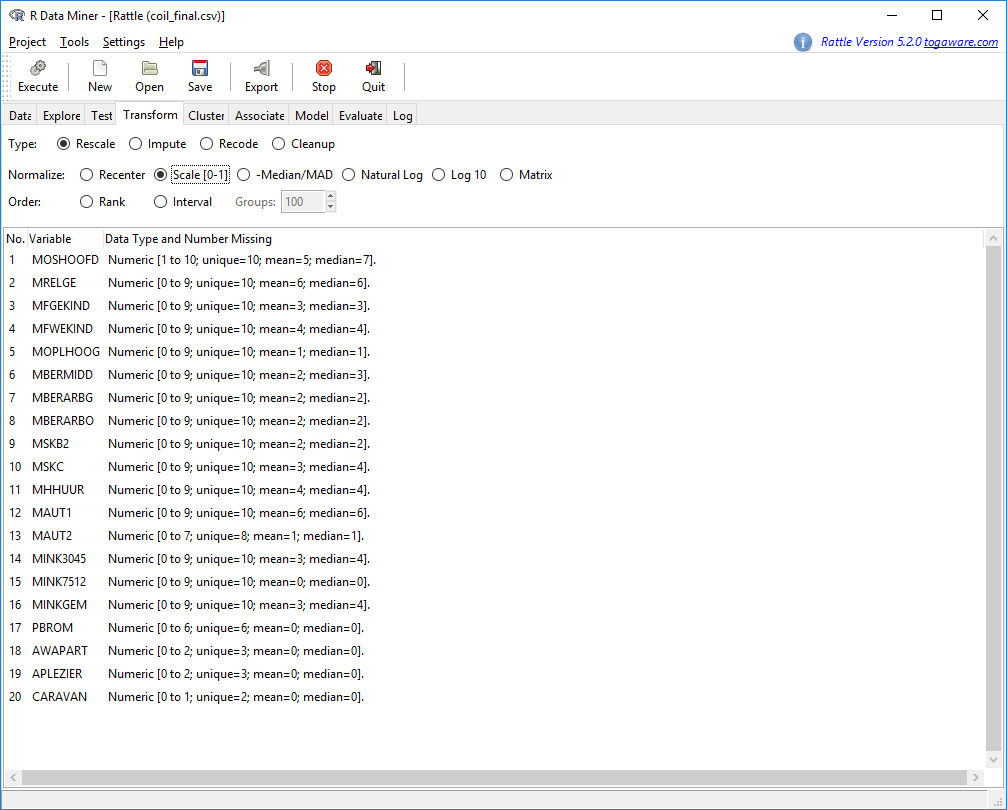
**Relationship**

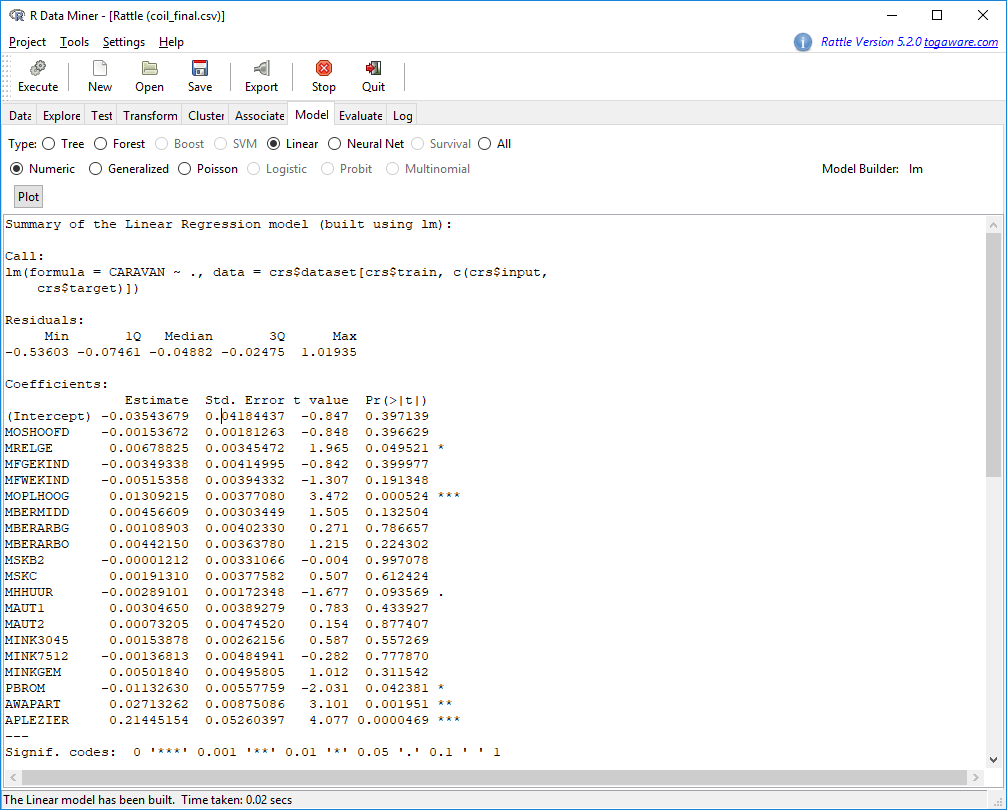
I chose the following attributes for various reasons. It is a well-known fact that most people who are financially secure but may live paycheck to paycheck will purchase insurance if it will prevent a large expense in the future so skilled laborers and middle management made the list as well as the income variable. For MOSHOOD, the customer main type I will focus on is the driven growers, average family, and conservative families. Additionally, I will target customers with children living in the household as children cost money. There will be less money to pay for car repairs.

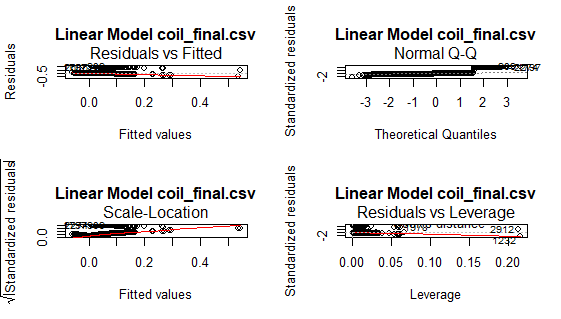
Less money leads me to income, I will target an income between 30 to 75K, and then I will also look at the number of automobile policies they already have.

**Which Algorithms to Use?**

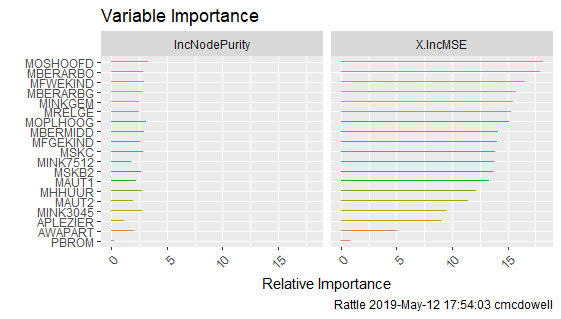
Using R, I first transformed my data by rescaling it so that each variable has an approximate size to another. This will help during K-Means testing. I then derived the data above using the forest model and then retrieving the variables with the largest frequencies.



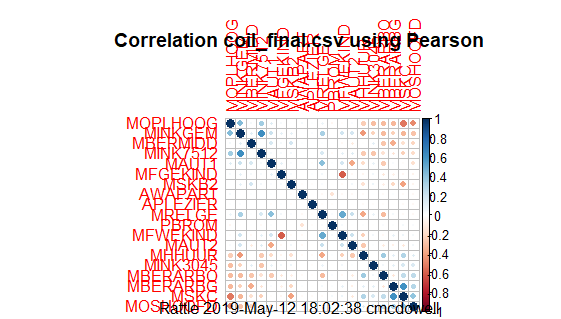
To analyze the data above, I will use a series of algorithms. First to find the best attributes, I will do a linear regression analysis and then a forest analysis. The linear regression describes a line that best identifies a relationship between variables using coefficients (Le, 2018).



This variable importance chart, identifies variables seen with the most frequency in the linear model above.



Again using the same data, we see revelance within the variavbles in this Correlation chart.

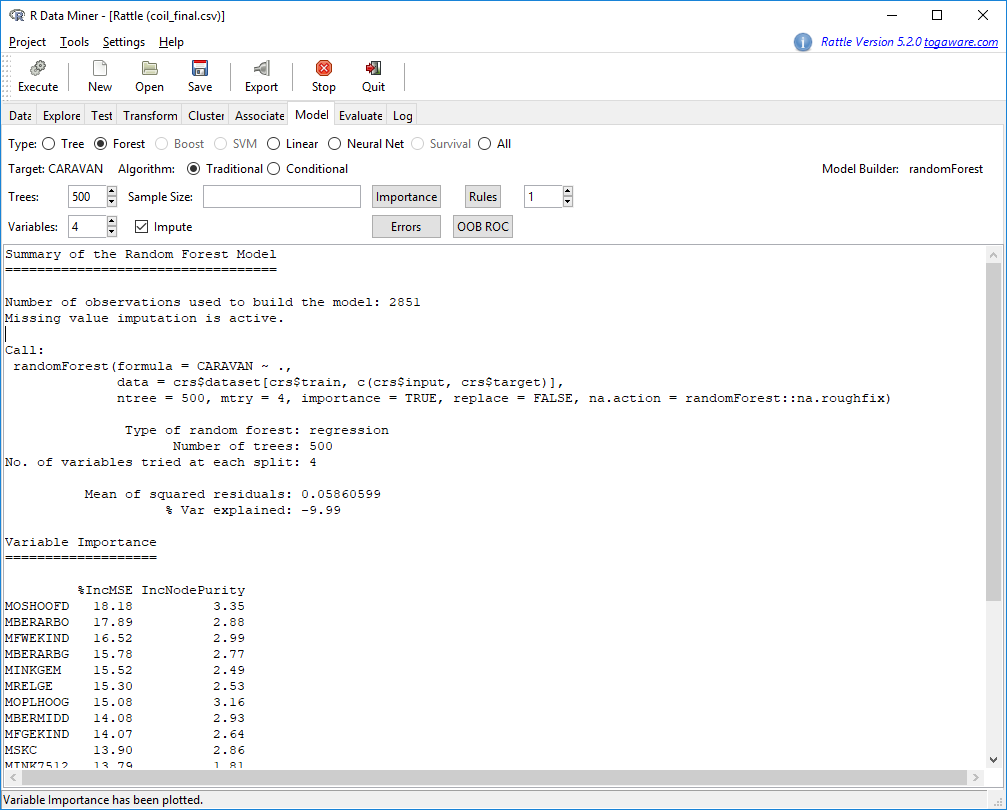


Creating a decision tree using Caravan as the target is a great way to identify the percentage of customers who will purchase car repair insurance.



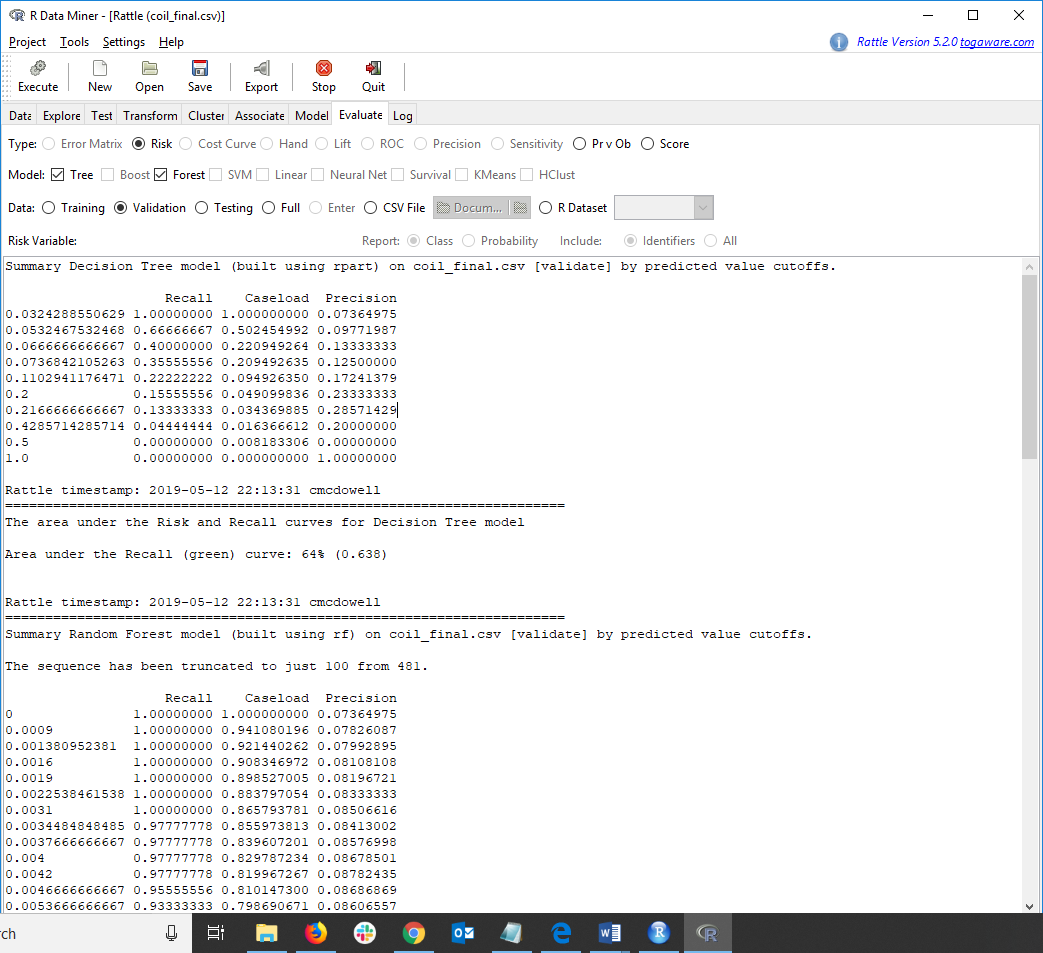
In this example, there is a probability that 83 % of the customers who are 40 years old and older and have one or more boat insurance policies, will purchase car repair insurance. 17% of the customers observed did not carry a third-party insurance and therefore the probability that they will purchase automotive repair insurance is null. These results are important because it gives stakeholders the opportunity to discuss the possibility of extending repair insurance to other vehicles types and to target certain customer types.

The random forest approach estimates an entire statistical model by taking multiple samples of the data, constructing several models and then generates a prediction based on an averaged output value. In addition, I will use the random forest model to evaluate the model.

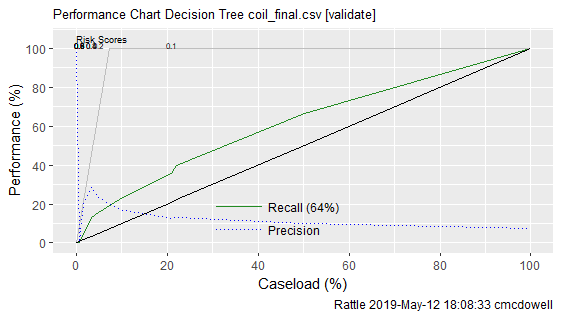


**Evaluating the Model**

Using the forest model , here is the summary:

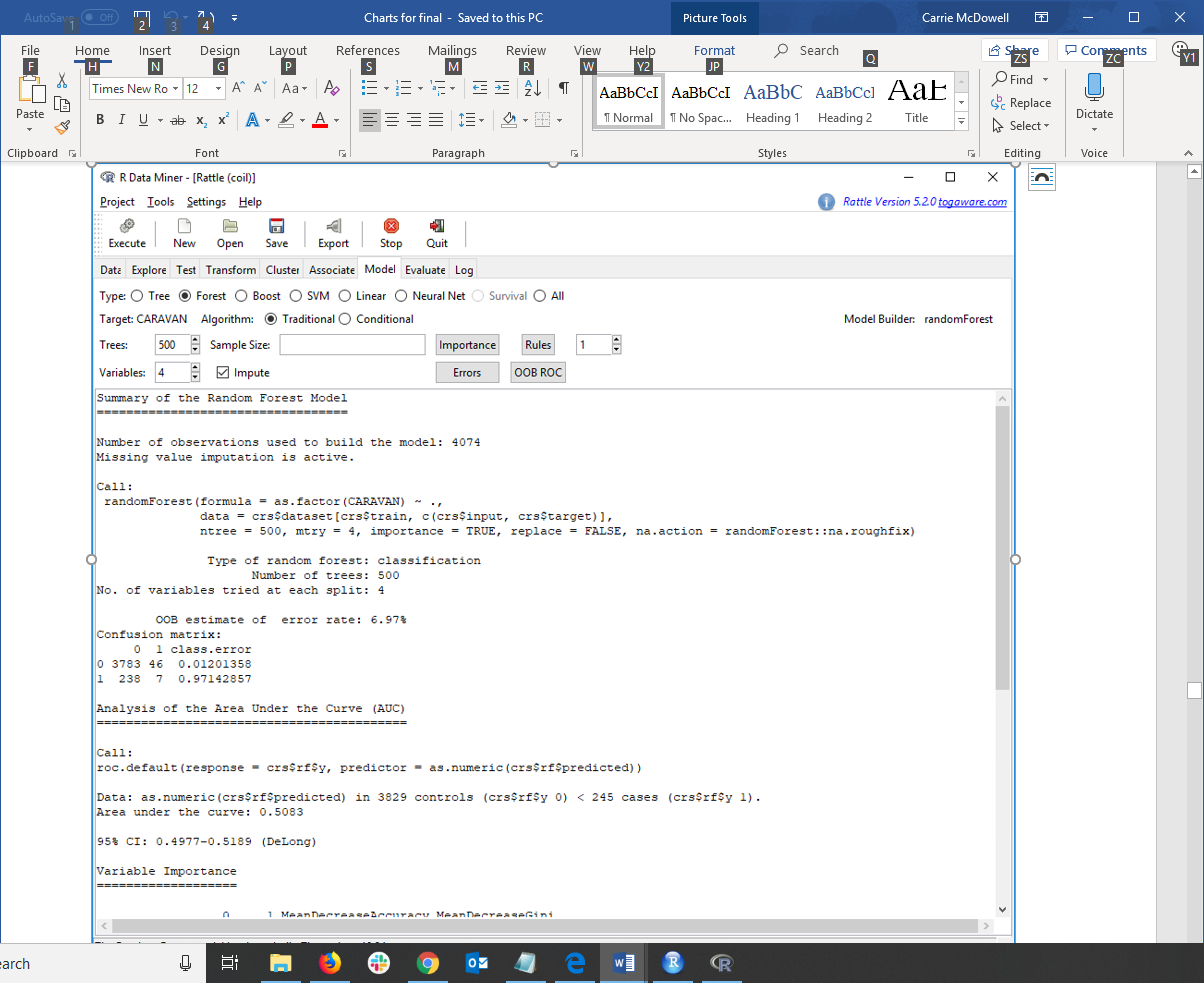


And then the Risk Chart still using the decision tree model. The area under the Risk and Recall curve for the Random Forest model is 64% or .638. Using the data above, I see the analysis of the area under the curve. The area under the under the ROC curve measures the quality of the model. Most models produce an area under the curve between 0.5 and 1. As seen below, the area under the curve is .638, this means that a randomly selected case from data group with a target of 1 has a score larger than that for a randomly chosen case with the target of 0 in just over 63% of the time.



Using this confusion matrix below, I was able to identify the number of correct and

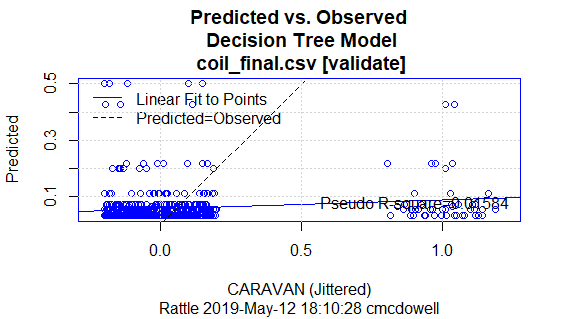
incorrect predictions made using the model compared to the outcomes.



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Confusion Matrix | | **Target** | |  | | | Positive | Negative | | **Model** | Positive | 3783 | 46 | *Positive Predictive Value* | .988 | | Negative | 238 | 7 | *Negative Predictive Value* | .0286 | |  | | *Sensitivity* | *Specificity* | **Accuracy** = .930 | | | .941 | .132 | | |
|  | |

The confusion matrix above proves an accuracy in the model of .93 and there were .988 predicted correctly as does the sensitivity number of .941.

Using the same data, I plotted the predicted vs observed chart. This will also be useful as we monitor the model once put into production as it will display the predicted values against the observed values over time.



**Scoring**

Scoring your model falls within the 5th and is instrumental in the 6th step of the CRISP-DM methodology. Scoring the model will show how well each observation in the dataset predicts a future outcome. For instance, in my dataset, the variables I used to identify purchasers of a certain insurance policy is scored for accuracy to its probability of a positive outcome.

Each line of the scored CSV file lists all variable ID’s as well as a 1 or 0 depending on the probability score. By assigning a score, this allowed me to evaluate the variables used in my dataset and modify it for better results.

The goal of this project was to understand and analyze the data that has been captured over the years on existing customers and then create a prediction model using the CRISP-DM methodology. Once complete agents of each office will use the model to target customers who will most likely benefit from value added services.

**Variable Importance**

Before scoring the variables, I identified a few that did not prove successful in the long run. Scoring allowed me to visualize the variables that returned a positive response and those that did not. I removed some from my model and added others. This only proved the value in data analytics as before scoring, just picking the variables I thought would provide the best result did not pan out.

**New Data Scoring**

In the future, the model will be available to company agents as well as a self-serve model published to the website. An agent or customer will complete a questionnaire, the model will score the answers and then provide a list of optimal product offerings from which the customer will choose from. This model will work with others to return savings for bundling as well. An additional piece, a “scoring engine” will be added to evaluate new data retrieved from the models. The data retrieved will be scored often to ensure that the model is still accurate.

To operate the model and retrieve raw data, customers will simply click on radio buttons or select answers from a drop down. In the background the program will assign a 1 for yes, and 2 for no to fit the answer into a category. The result will provide a list of best choices to the agent or customer and the data will be used in the back ground to maintain the model.

**Deployment**

The model will be deployed in the Predictive Model Markup Language (PMML) and then passed onto application developers who will publish the program to the web. They may create API’s using DeployR or Pattern to facilitate cloud-based communications for all types of devises. Pattern will integrate Hadoop which is another open sourced software meant to handle large files of data. We will use Java in the cloud to handle the scoring engine which will run on a regular basis to test the data.

Once implemented, a team of administrators will monitor the model and results from continual scoring for accuracy and relevance. They will make recommendations back to the data scientist who will pass along enhancements for the model to the application developer for proper maintenance.

**Analytics**

This model proved that the use of data analytic is beneficial to AIM. After implementing this model, the data analytics team continue to capture data from agents and online self-service portals. This data will be analyzed and modeled to be used to either create new customer engagement models or used to tweak the existing model. In time, several models will be utilized to better the health of the company and to create customized products.

References

Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to Choose an Algorithm for a Predictive Analysis Model. Retrieved April 12, 2019, from <https://www.dummies.com/programming/big-data/data-science/how-to-choose-an-algorithm-for-a-predictive-analysis-model/>

Edwards, J. (2018, May 16). What is predictive analytics? Transforming data into future insights. Retrieved April 7, 2019, from <https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html>

Le, J. (2018, January 20). A Tour of The Top 10 Algorithms for Machine Learning Newbies. Retrieved April 26, 2019, from https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

Parthasarathy, S. (2018, November 01). Are You Asking the Right Predictive Questions? Retrieved July 7, 2019, from <https://www.logianalytics.com/predictive-analytics/are-you-asking-the-right-predictive-questions/>

Vorhies, W. 2016. CRISP-DM – a Standard Methodology to Ensure a Good Outcome. Retrieved from <https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>

Williams, G. (2010, August 22). Scoring. Retrieved May 2, 2019, from <https://www.togaware.com/datamining/survivor/Scoring.html>

Williams, G. J. (2013). Data mining with Rattle and R: The art of excavating data for knowledge discovery. New York: Springer.

Wilson, E. (2018, October 10). The 4 Phases Of Predictive Analytics From Start To Finish. Retrieved April 7, 2019, from <http://demand-planning.com/2018/10/10/the-4-phases-of-predictive-analytics/>

Zicari, R., & Boweles, M. (2016, February 2). How to choose a predictive algorithm. Retrieved April 11, 2019, from http://www.odbms.org/2016/02/how-to-choose-a-predictive-algorithm/