Q1. Examine the data dictionary and download the data

The data dictionary provided outlines categorical variables in the dataset, each with values that represent different levels or states. Education levels range from 1 ('Below College') to 5 ('Doctor'), indicating the highest level of education an employee has achieved. EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, and RelationshipSatisfaction are scored from 1 ('Low') to 4 ('Very High'), reflecting employees' feelings about their work environment, their involvement and satisfaction with their job, and their satisfaction with work relationships, respectively. PerformanceRating ranges from 1 ('Low') to 4 ('Outstanding'), assessing how well employees perform their job duties. Lastly, WorkLifeBalance scores from 1 ('Bad') to 4 ('Best'), indicating the quality of an employee's balance between work and personal life. These variables are crucial for understanding the nuances of employee experiences and can help identify factors leading to employee attrition.

Q2. Perform exploratory data analysis, pre-process the data as required. Provide a bullet-list summary of data pre-processing steps.

- ***Missing Values:*** No missing values were detected in the dataset.
- ***Variance of Features:*** Two features, StandardHours and EmployeeCount, have a variance of 0, indicating they do not vary at all across the dataset and can be considered for removal as they provide no informative value. Other numerical features exhibit varying degrees of variance, which will be important for model training and might require scaling.
- ***High Cardinality Features:*** The highest cardinality is observed in JobRole with 9 unique values, followed by EducationField with 6. These levels of cardinality are manageable for most modeling techniques through encoding. Other categorical features have lower cardinality (3 or fewer unique values), which is generally not considered high and is suitable for straightforward encoding methods. Over18 has only 1 unique value, indicating that it does not vary across the dataset.
- ***Remove Non-Variative and Non-Informative Features:*** StandardHours, EmployeeCount, Over18, and EmployeeNumber will be removed as they provide no informative value for analysis or modeling.
- ***Categorical Variables:*** Identified categorical variables are Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime.
- ***Numerical Variables:*** Identified numerical variables include Age, DailyRate, DistanceFromHome, Education, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MontlyIncome, NumCompaniesWorked, MonthlyRate, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager
- ***Recoding Variables:*** Change Education (representing the level of education), JobLevel (representing the level of the job), and StockOptionLevel (the level of stock options the employee has) to categorical. Leave EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, and RelationshipSatisfaction as numeric, treating them as Likert scale responses.

Feature Name	Data Quality	Index	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
Age		1	Numeric	43	0	36.92	9.13	36	18	60
Attrition	TARGET	2	Categorical	2	0					
BusinessTravel		3	Categorical	3	0					
DailyRate		4	Numeric	886	0	802	403	802	102	1,499
Department		5	Categorical	3	0					
DistanceFromHome		6	Numeric	29	0	9.19	8.10	7	1	29
Education (Categorical Int)		7	Categorical	5	0					
EducationField		8	Categorical	6	0					
EnvironmentSatisfaction		11	Numeric	4	0	2.72	1.09	3	1	4
Gender		12	Categorical	2	0					
HourlyRate		13	Numeric	71	0	65.89	20.32	66	30	100
JobInvolvement		14	Numeric	4	0	2.73	0.71	3	1	4
JobLevel (Categorical Int)		15	Categorical	5	0					
JobRole		16	Categorical	9	0					
JobSatisfaction		17	Numeric	4	0	2.73	1.10	3	1	4
MaritalStatus		18	Categorical	3	0					
MonthlyIncome		19	Numeric	1,349	0	6,503	4,706	4,919	1,009	19,999
MonthlyRate		20	Numeric	1,427	0	14,313	7,115	14,236	2,094	26,999
NumCompaniesWorked		21	Numeric	10	0	2.69	2.50	2	0	9
OverTime		23	Categorical	2	0					
PercentSalaryHike		24	Numeric	15	0	15.21	3.66	14	11	25
PerformanceRating		25	Numeric	2	0	3.15	0.36	3	3	4
RelationshipSatisfaction		26	Numeric	4	0	2.71	1.08	3	1	4
StockOptionLevel (Categorical Int)		28	Categorical	4	0					
TotalWorkingYears		29	Numeric	40	0	11.28	7.78	10	0	40
TrainingTimesLastYear		30	Numeric	7	0	2.80	1.29	3	0	6
WorkLifeBalance		31	Numeric	4	0	2.76	0.71	3	1	4
YearsAtCompany		32	Numeric	37	0	7	6.12	5	0	40
YearsInCurrentRole		33	Numeric	19	0	4.23	3.62	3	0	18
YearsSinceLastPromotion		34	Numeric	16	0	2.19	3.22	1	0	15
YearsWithCurrManager		35	Numeric	18	0	4.12	3.57	3	0	17

Q3. Develop a business case focusing on employee retention, define the metric to be used for model performance evaluation.

In response to the challenge of high employee turnover rates, HR departments and organizational decision-makers, particularly within companies like IBM, are increasingly recognizing the costs associated with this issue. These costs are not just financial, encompassing expenditures on recruitment and training for new hires, but also include the intangible losses of organizational knowledge and cohesion. For a globally recognized company like IBM, which prides itself on innovation and leadership in the technology sector, the impact of high turnover extends beyond immediate costs to affect team dynamics, productivity, and the overall morale and culture of the workforce. This, in turn, can precipitate a cycle of continued attrition. In this analysis, we will leverage advanced analytics and machine learning to develop an employee retention strategy. This strategy involves the deployment of machine learning models—ranging from Neural Networks and K-Nearest Neighbors to Logistic Regression, Random Forest, Gradient Boosted Trees, and Support Vector Machines. The objective is to

analyze and understand the underlying factors driving employee attrition. This knowledge empowers IBM to craft targeted interventions aimed at enhancing facets of employment that are pivotal to employee satisfaction, such as job satisfaction, work-life balance, career development opportunities, and compensation packages. The focus of the evaluation framework for this retention strategy is the Maximum Payoff metric. This metric is designed to quantify the expected financial returns from the implementation of the model-driven retention recommendations, essentially providing a tangible measure of the return on investment (ROI) of these strategies. The Maximum Payoff metric is especially pertinent for a company like IBM, as it directly correlates the efficacy of the retention strategies with the company's bottom line, ensuring that the efforts to curb turnover are not only effective in enhancing employee retention but also in mitigating the financial and operational repercussions of turnover. Through the strategic application of the models to identify and address the predictors of employee attrition, IBM is positioned to significantly enhance its employee retention rates. This helps in reducing the costs associated with high turnover and bolsters IBM's reputation as an employer of choice, committed to the growth, satisfaction, and well-being of its workforce.

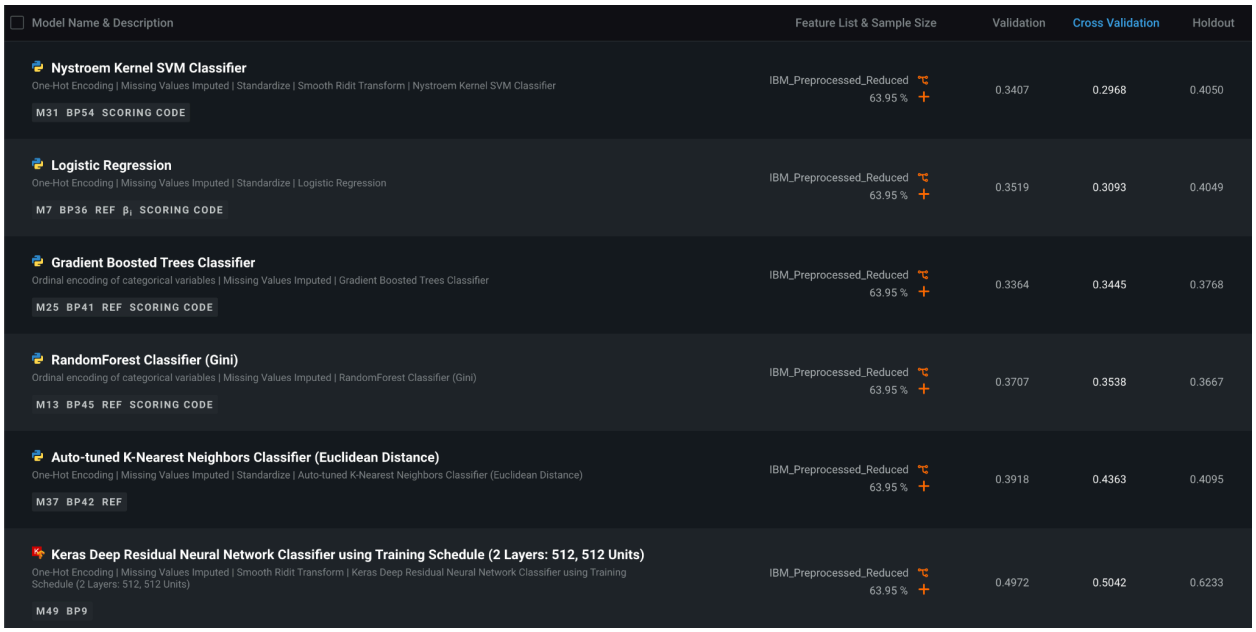
Financial Implications:

- ***Cost of losing and replacing an employee:*** The average cost of replacing an individual employee is conservatively estimated to be 150% of their annual salary, according to Gallup. The estimated average salary for IBM employees is around $118,768 per year, according to online research. As such, the cost of losing an employee would be a conservative estimate of about $178,152.
- ***Cost of implementing a retention strategy per employee:*** It can be estimated that a retention strategy would cost approximately 10% of an employees annual salary, which would equal $11,876.80.
- ***Net savings from successful retention (considering the cost of intervention):*** Subtracting the cost of implementing a retention strategy from the cost of losing an employee would equal $166,275.20 ($178,152 - $11,876.80).

A breakdown the payoff matrix in this scenario would be the following:

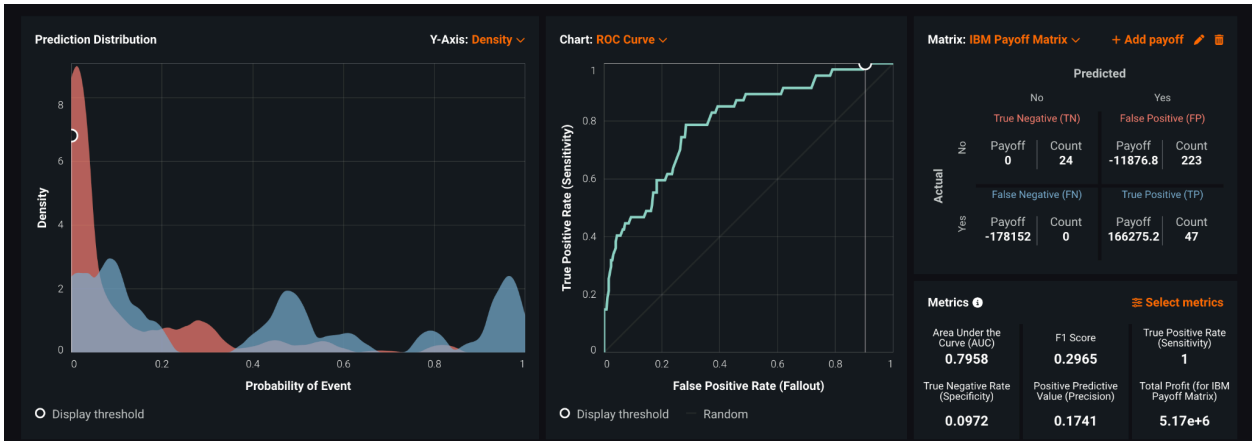
- ***True Positives (TP):*** This represents the net savings from not having to replace an employee due to a successful retention intervention, after accounting for the cost of that intervention.
    - Hypothetical Value: +$166,275.20 per employee
- ***False Positives (FP):*** This is the cost of implementing a retention strategy for an employee who would not have left anyway, representing an unnecessary expense.
    - Hypothetical Value: -$11,876.80 per employee
- ***True Negatives (TN):*** No cost or savings are directly associated with this outcome because it involves employees who were not at risk of leaving and for whom no intervention was made.
    - Hypothetical Value: $0
- ***False Negatives (FN):*** This represents the cost of losing and replacing an employee when a potentially successful retention intervention was not implemented.
    - Hypothetical Value: -$178,152 per employee

Q4. Evaluate all appropriate models discussed in the course. Provide a table with the summary of model performance. Provide screenshots of individual model performance summaries on relevant metrics.

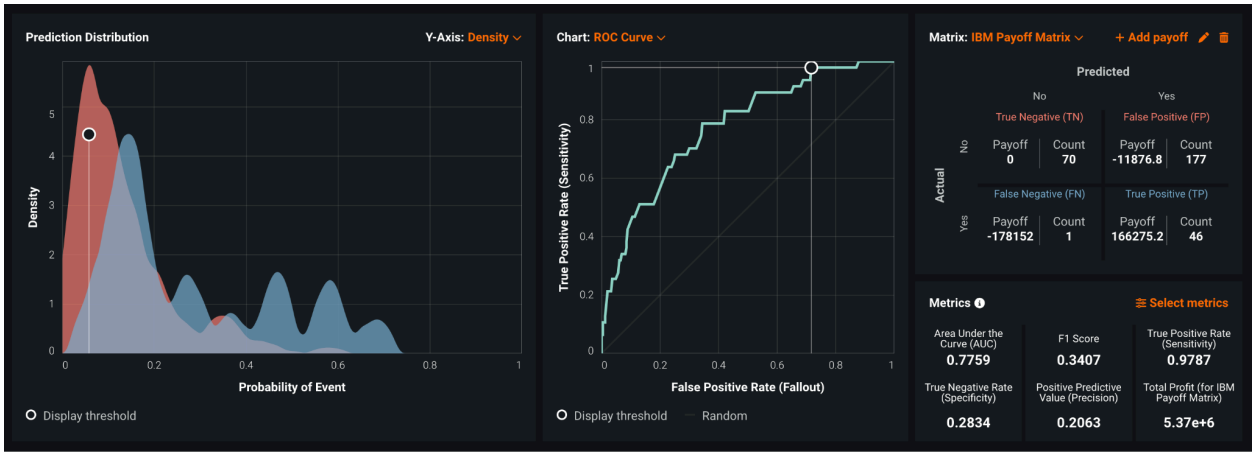


Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M31 BP54 SCORING CODE	IBM_Preprocessed_Reduced 63.95 %	0.3407	0.2968	0.4050
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M7 BP36 REF β_i SCORING CODE	IBM_Preprocessed_Reduced 63.95 %	0.3519	0.3093	0.4049
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M25 BP41 REF SCORING CODE	IBM_Preprocessed_Reduced 63.95 %	0.3364	0.3445	0.3768
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M13 BP45 REF SCORING CODE	IBM_Preprocessed_Reduced 63.95 %	0.3707	0.3538	0.3667
Auto-tuned K-Nearest Neighbors Classifier (Euclidean Distance) One-Hot Encoding | Missing Values Imputed | Standardize | Auto-tuned K-Nearest Neighbors Classifier (Euclidean Distance) M37 BP42 REF	IBM_Preprocessed_Reduced 63.95 %	0.3918	0.4363	0.4095
Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units) One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units) M49 BP9	IBM_Preprocessed_Reduced 63.95 %	0.4972	0.5042	0.6233

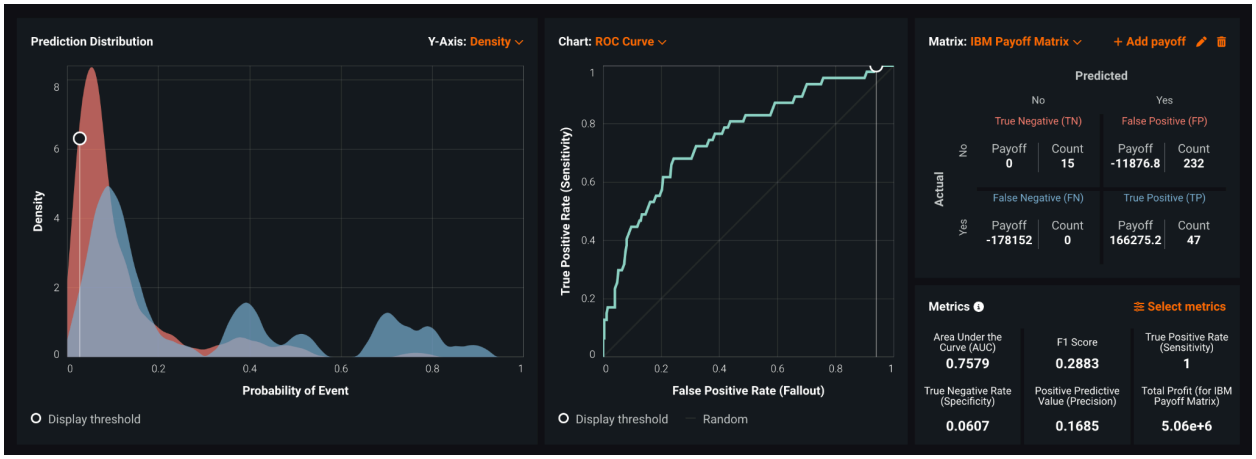
Logistic Regression: We use Logistic Regression for its simplicity and interpretability, making it a great baseline model for binary classification problems like predicting employee attrition based on linear relationships between features.



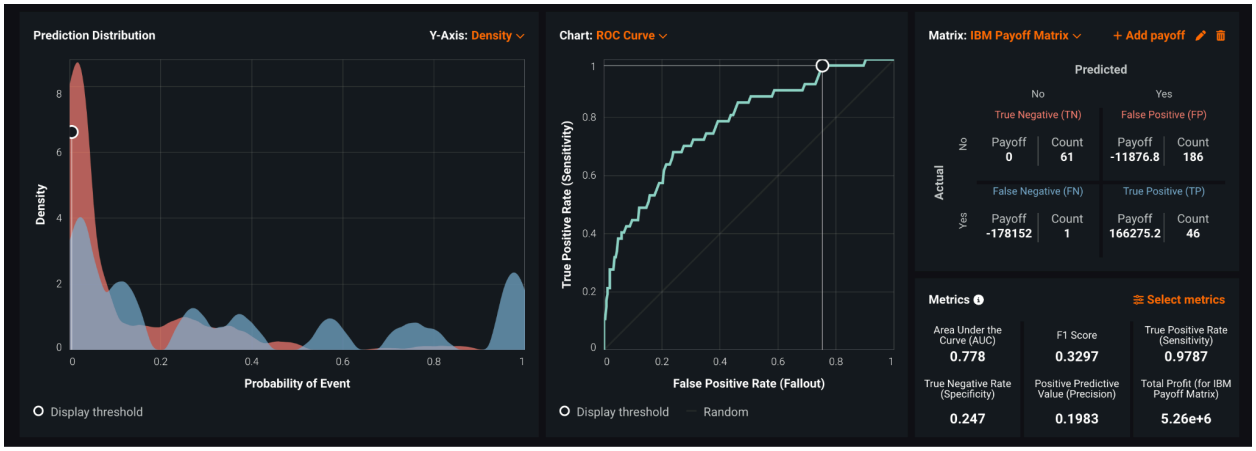
RandomForest Classifier (Gini): The RandomForest Classifier is utilized for its ability to handle non-linear relationships and interactions between variables without requiring feature scaling, leveraging multiple decision trees to improve prediction accuracy and reduce overfitting.



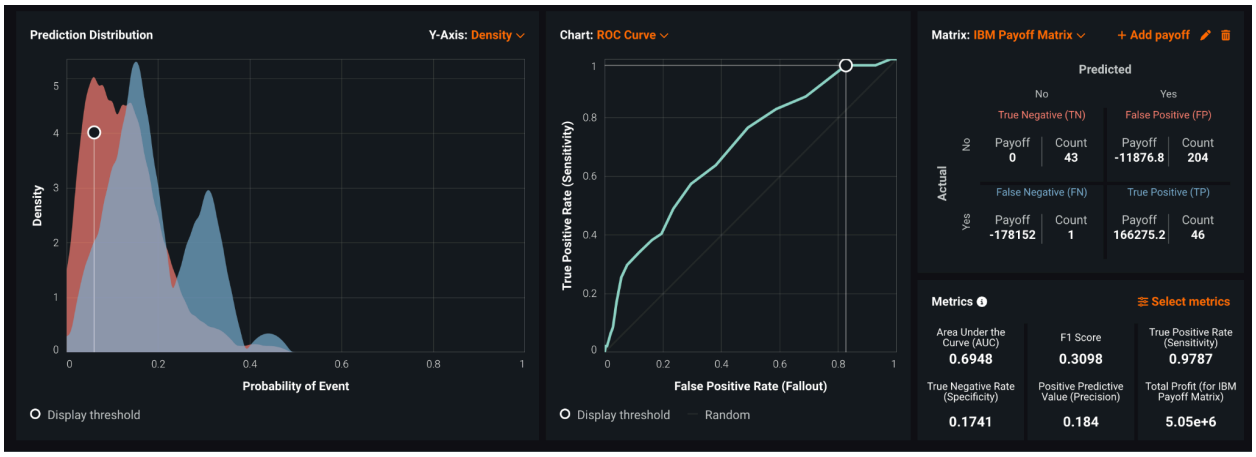
Gradient Boosted Trees Classifier: This model is chosen for its high performance and efficiency in handling complex datasets with non-linear relationships, by sequentially correcting errors of previous trees, leading to improved predictive accuracy.



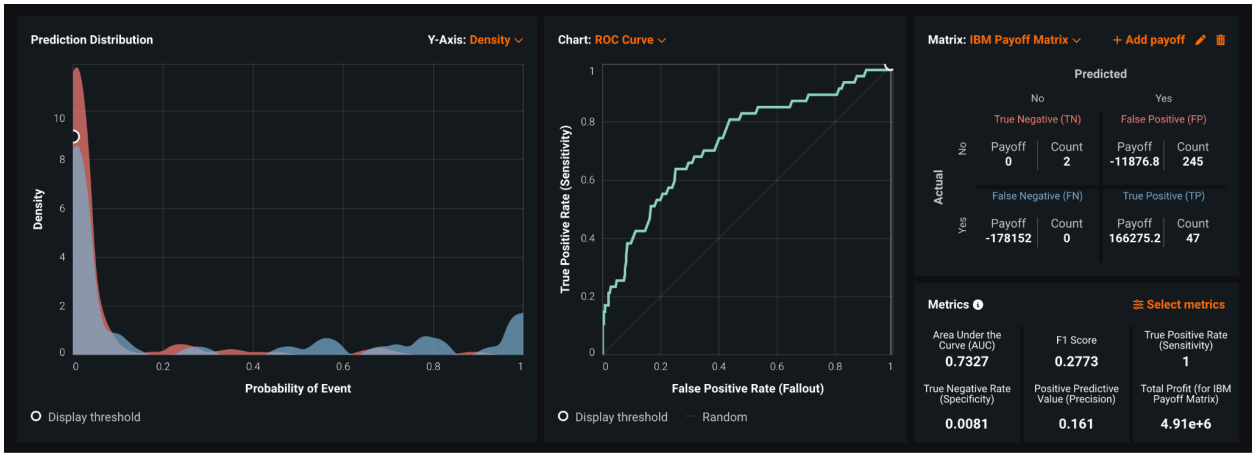
Nystroem Kernel SVM Classifier: The Nystroem Kernel SVM is used to efficiently approximate the kernel trick, allowing us to capture complex, non-linear relationships in the data, while significantly reducing computational cost compared to traditional SVMs.



Auto-tuned K-Nearest Neighbors Classifier (Euclidean Distance): This variant of KNN is employed for its simplicity and effectiveness in capturing the local structure of the data, with auto-tuning ensuring the optimal number of neighbors and distance metric for best performance.



Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units): This deep learning model is chosen for its ability to learn complex patterns and interactions through a deep architecture, with residual connections enhancing training efficiency and mitigating the vanishing gradient problem, making it suitable for high-dimensional data.



	Logistic Regression	Random Forest Classifier (Gini)	Gradient Boosted Trees Classifier	Nystroem Kernel SVM Classifier	Auto-tuned K-Nearest Neighbors Classifier	Keras Deep Self Normalizing Residual Neural Network Classifier
Recall	1	0.9787	1	0.9787	0.9787	1
Precision	0.1741	0.2063	0.1685	0.1983	0.184	0.161

F1	0.2965	0.3407	0.2883	0.3297	0.3098	0.2773
Specificity	0.0972	0.2834	0.0607	0.247	0.1741	0.0081
ROC AUC	0.7958	0.7759	0.7579	0.778	0.6948	0.7327
Maximum Payoff	$5,170,000	$5,370,000	$5,060,000	$5,260,000	$5,050,000	$4,910,000
LogLoss	0.3519	0.3707	0.3364	0.3407	0.3918	0.4972
Cross Validation	0.3093	0.3538	0.3445	0.2968	0.4363	0.5042
Holdout	0.4049	0.3667	0.3768	0.4050	0.4095	0.6233
Threshold	0.0004	0.0578	0.0274	0.0045	0.06	0.0001

Explaining the Metrics

- ***Recall:*** The proportion of actual positive cases correctly identified by the model. High recall means few false negatives.
- ***Precision:*** The proportion of positive identifications that were actually correct. High precision means few false positives.
- ***F1:*** The harmonic mean of precision and recall. It balances both to provide a single measure of a test's accuracy.
- ***Specificity:*** The proportion of actual negatives correctly identified. High specificity means few false positives.
- ***ROC AUC:*** The Area Under the Receiver Operating Characteristic curve. It measures the model's ability to distinguish between classes.
- ***Maximum Payoff:*** An estimate of the maximum financial benefit that could be achieved with the model's predictions.
- ***LogLoss:*** A measure of the accuracy of a classifier. It penalizes false classifications; lower values indicate better models.
- ***Cross Validation:*** A technique to assess how the results of a statistical analysis will generalize to an independent dataset. It helps in mitigating overfitting.
- ***Holdout:*** A portion of the dataset not used during model training, used later to test the model's performance to ensure it generalizes well.
- ***Threshold:*** The decision point between different classifications. Adjusting it affects the trade-off between recall and precision.

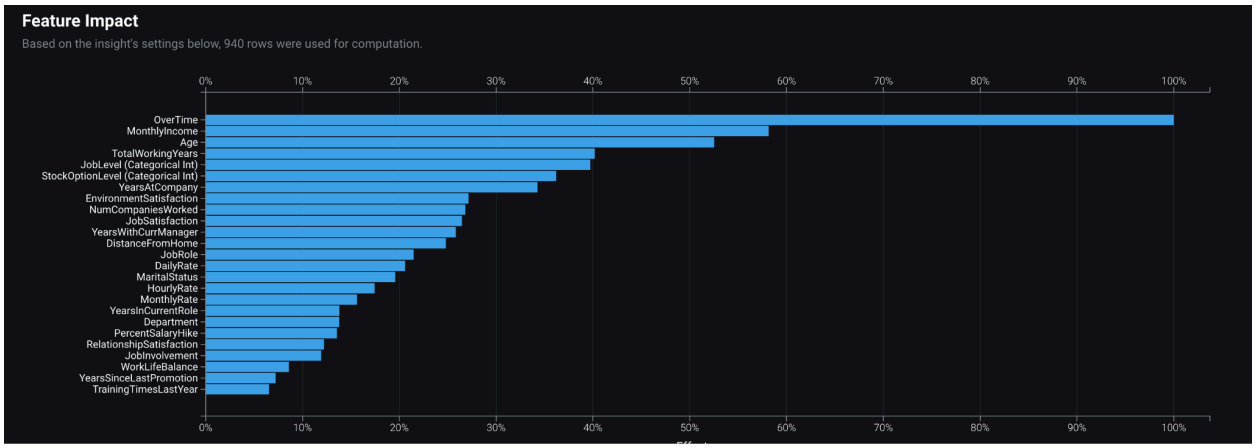
Best Metric

The "Maximum Payoff" metric is crucial here because it quantifies the financial benefits of using each model to predict and address employee attrition, effectively measuring the ROI of implementing the model's recommendations.

Best Model

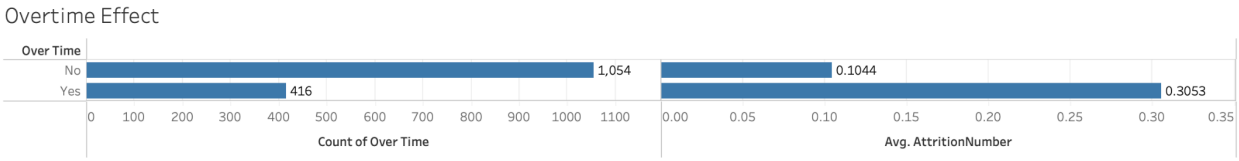
- The Random Forest Classifier (Gini) has the highest maximum payoff of $5,370,000. However, it's important to compare its performance on other metrics to ensure it's not only financially beneficial but also robust and reliable. High recall indicates the model's ability to catch as many true positives (actual attritions) as possible. The Gradient Boosted Trees Classifier, Nystroem Kernel SVM Classifier, and the Keras Deep Self-Normalizing Residual Neural Network Classifier all share the highest recall of 1, indicating perfect sensitivity. Precision shows how many of the positively predicted cases were actually positive. The Random Forest Classifier leads in precision with 0.2063, suggesting it has a higher rate of correct positive predictions relative to the total positive predictions it makes. The F1 Score shows the balance between Precision and Recall. The Random Forest Classifier also shows a strong F1 score of 0.3407, indicating a good balance between precision and recall. Finally the ROC AUC reflects the model's ability to distinguish between the classes across all thresholds. Logistic Regression has the highest ROC AUC of 0.7958, indicating superior performance in discriminating between the classes.
- Considering all these metrics, the Random Forest Classifier (Gini) offers the highest maximum payoff and maintains competitive performance across the key predictive metrics, particularly in Precision and F1 Score, which are crucial for balancing the accuracy of positive predictions and the model's ability to correctly identify as many actual positives as possible. While other models might excel in specific metrics (like Recall or ROC AUC), Random Forest's balance across financial return and predictive performance metrics makes it the best choice in this context. Its strong performance in F1 Score and the highest Precision among models, coupled with the highest Maximum Payoff, suggests it's the most effective model for maximizing financial benefits while maintaining robust prediction capabilities.

Q5. Identify 5 most informative factors in the best performing model. Visualize and summarize the effects in 1-2 sentences for each of the top 5 factors.

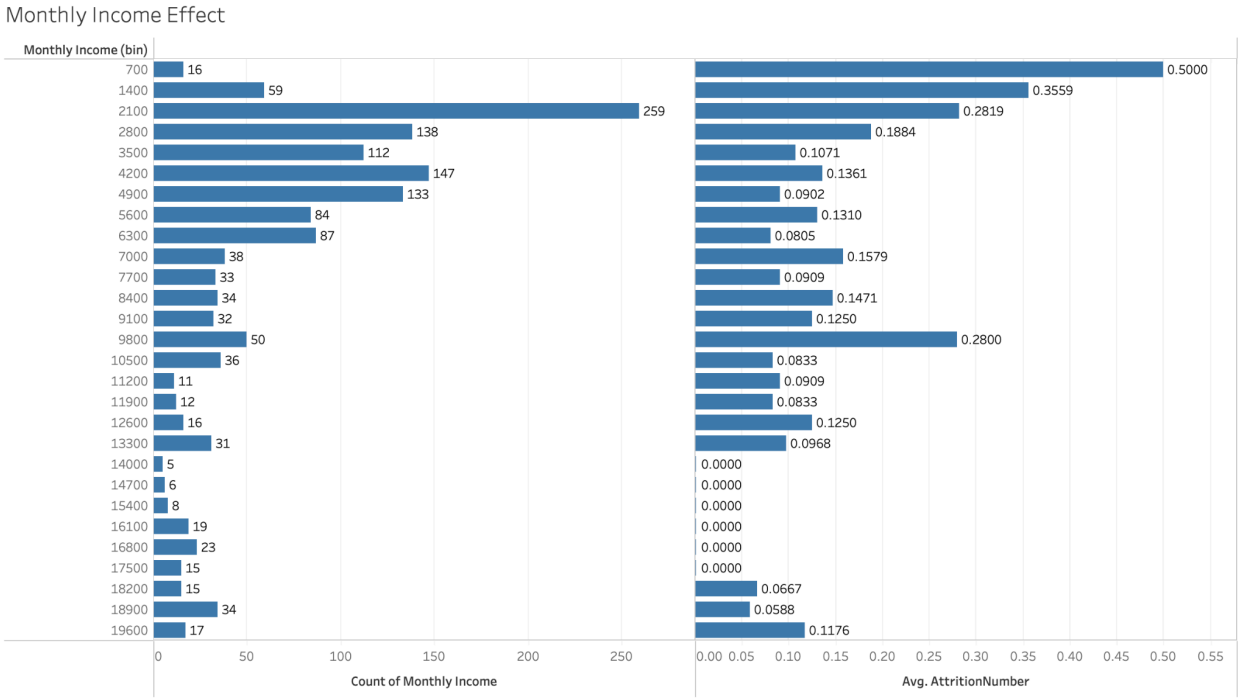


For the Random Forest Classifier (Gini), which emerged as the best-performing model in terms of maximum payoff and balanced prediction metrics, the five most informative factors impacting employee attrition and their effects are as follows:

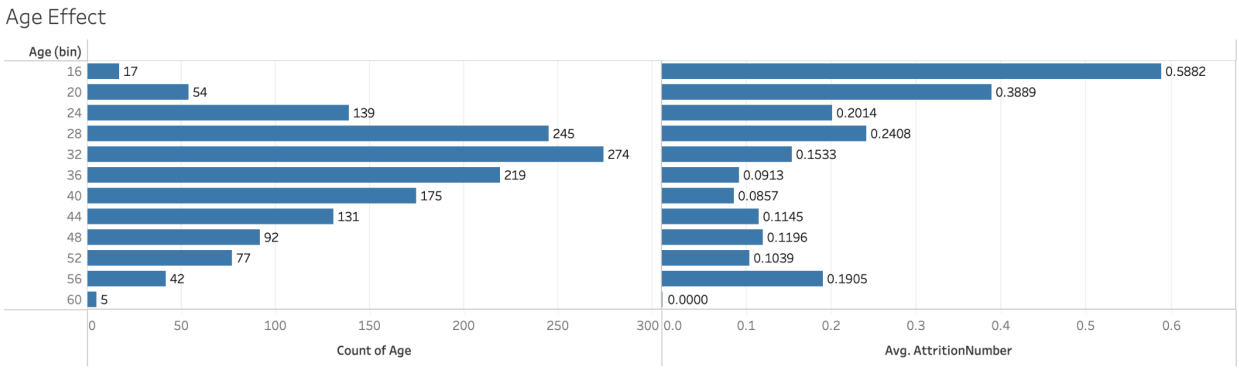
Overtime: Employees who do work overtime appear to have a noticeably higher average attrition rate than those who do not work overtime. This suggests that working overtime is a significant factor in employees' decisions to leave the company, supporting the conclusion that managing overtime could be crucial in reducing overall attrition rates.



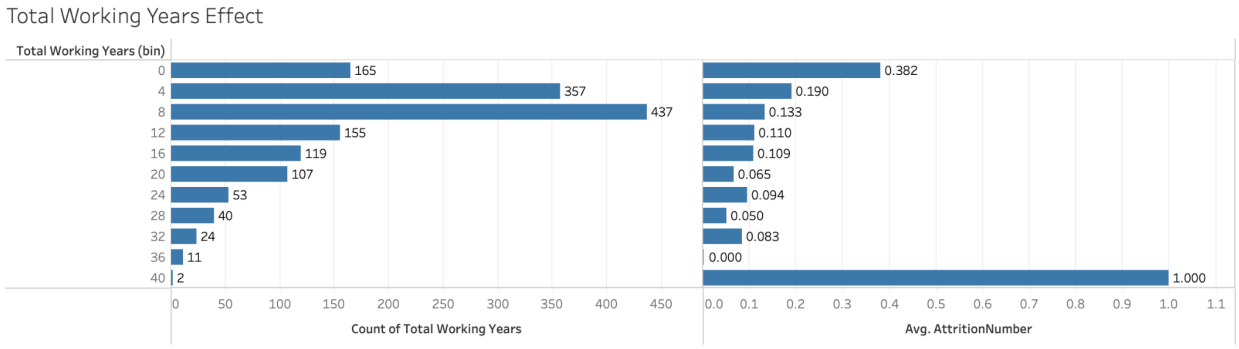
Monthly Income: It appears that lower income brackets have a higher average attrition rate, suggesting that as monthly income increases, the likelihood of employees leaving the company decreases. This trend indicates that higher monthly income may be a significant factor in retaining employees.



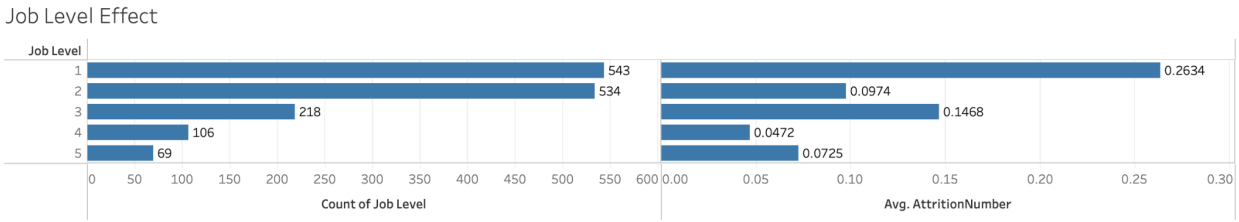
Age: The chart demonstrates the relationship between employee age and attrition rates, with younger employees, particularly those in their 20s, showing the highest average attrition rates. As age increases, the average attrition rate tends to decrease, indicating that older employees are less likely to leave the organization, potentially due to factors such as greater job satisfaction, higher job security, or less desire for job-hopping.



Total Working Years: The graph illustrates a trend where employees with fewer total working years exhibit higher average attrition rates. As the total number of working years increases, the average attrition rate decreases, suggesting that employees with more experience or tenure at the company are less likely to leave, possibly due to greater investments in their careers or more substantial ties to the organization.



Job Level: The chart shows that lower job levels have higher average attrition rates, with the highest attrition observed at the entry-level (Job Level 1), and the attrition rate decreases as job level increases, suggesting that employees at higher job levels, who likely have more responsibilities and higher compensation, are less prone to leave the company. This could reflect increased job satisfaction, investment in the company, or the perceived value of the benefits associated with higher positions.



<u>Q6. Provide actionable recommendations for each of the observed effects. If an observed effect is not actionable, please state so.</u>

- ***Overtime:*** IBM should implement policies to manage overtime more effectively, such as hiring additional staff to alleviate workload or offering time-in-lieu for overtime hours. The company should also encourage a culture that values work-life balance by setting clear expectations around reasonable working hours.
- ***Monthly Income:*** The company should review compensation packages to ensure they are competitive with the market rate, especially for lower-income brackets. It could also consider introducing performance bonuses, pay raises, or other financial incentives to increase employee retention.
- ***Age:*** For younger employees, IBM should develop targeted programs such as mentorship, career development paths, and job rotation opportunities that cater to their career growth aspirations. For older employees, ensure that recognition programs and retirement benefits are competitive to maintain their loyalty.
- ***Total Working Years:*** The company should design retention strategies like loyalty programs, tenure awards, or progressive benefits that increase with years of service to encourage employees to stay longer with the company.
- ***Job Level:*** IBM Should establish clear career progression plans and provide training and development opportunities to help employees move up the ladder. Additionally, it could foster a positive work environment with recognition programs that highlight the contributions of employees at all levels, with a focus on entry-level positions to reduce their higher attrition rates.