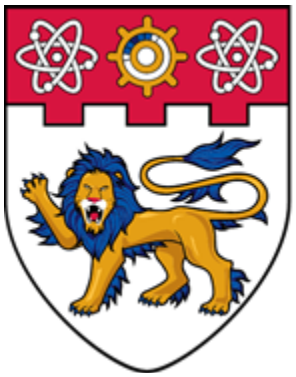


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2407 - ANALYTICS II

Semester 2 2022-2023

SEM 1 Group 7

Improving Employee Retention in Coinbase: An Analytics Framework for Talent Management in the Cryptocurrency Industry

Group Member	Matriculation Number
Jindal Jai	U2123034K
Koh Xin Yi Clarice	U2110183D
Lee Pei Yee	U2122590E
Shaun Lim Shi Lun	U2110811D

Table of Contents

Executive Summary

Coinbase is a cryptocurrency trading and investing platform with over 4,500 employees and 245,000 ecosystem partners in over 100 countries. After the Great Resignation of 2022, The U.S. hit a record 50.5 million people leaving their jobs, with Coinbase as no exception. With an average tenure period of 0.8 years, and one of the fastest turnover rates in the country, this has directly affected Coinbase's profitability and revenue generation. To progress on their "Go Broad, Go Deep" strategy, Coinbase is looking to focus on retaining their existing employees to enhance cost-savings. Therefore, there is an urgent need to predict and identify employees who might be considering leaving, and implementing policies to retain them.

The report aims to identify the main driving factors of employee attrition through analytical models and proposes a solution for Coinbase to predict and address employee attrition using a predictive app and a dashboard. By tapping on data analytics, Coinbase can gain a better understanding of the root causes through a combination of Logistic Regression and Random Forest as our feature selection methods to shortlist 13 common important variables from the initial 26 features. In addition, based on these features, our Random Forest model helps Coinbase to predict and identify employees who are at risk of leaving the company. This also ensures that Coinbase can implement effective retention policies that targets the important factors and the right group of employees, preventing wastage of scarce resources and costs.

The report proposes a two-tier approach for Coinbase to tackle their employee attrition problem. The solution comprises a WebApp and a Dashboard for usage by Senior Management and the Human Resources Team. Through the WebApp, users are able to predict whether an employee is likely to leave, and what are the key factors driving their decision. Personalised Recommendations are then provided, for HR to implement. Lastly, for more complex, non quantifiable factors, we derived insights from a second dataset, to come up with the respective recommendations.

The dashboard on the other hand, gives a broader long-term overview, allowing the users to track statistics and identify key contributory factors in varying job roles. This aids their analysis on the effectiveness of current retention strategies, as well as provide insights on how to further develop new & improved strategies.

The report concludes by highlighting the benefits of adopting a **combination** of the two solutions to predict the likelihood of attrition of Coinbase employees, while providing further recommendations to continually improve Coinbase's retention policies.

1. Introduction

1.1. Overview of Coinbase

Coinbase is a cryptocurrency trading and investing platform that offers users the ability to buy, sell, and exchange over 200 tradable cryptocurrencies. Today, there are approximately 110 million verified users and over $80 billion in assets on the platform, making it the biggest cryptocurrency exchange in the U.S. (Coinbase, n.d)

With more than 4,500 employees and 245,000 ecosystem partners in over 100 countries, Coinbase faces challenges in talent management, especially after the year of Great Resignation in 2022. The U.S. hit its record high of 50.5 million people quitting their jobs according to the U.S. Bureau of Labor Statistics. (Iacurci, 2023) Without exception, Coinbase experienced an average tenure period of employees working in Coinbase going as low as 0.8 years. (Murphy, 2022) Despite being a remote-first company, many past employees still complained on Glassdoor about the poor work life balance when they were working in the company. (Glassdoor, n.d) This raises a possibility that the company's demanding and unhealthy working environment discourages current employees to remain in the job, leading to the high attrition rate. However, more in-depth research is needed to confirm this justification, thus this report seeks to help Coinbase **identify the main driving factors of attrition** among employees through the use of analytical models.

1.1.1. Current Measures

In 2022, Coinbase implemented four "recharge weeks" per year for employees to take a break off work. (ETF Trends, 2022) Previously, there were already two recharge weeks scheduled per year and 52% of employees said recharge days and weeks were the primary tool that helped them rest and recover in 2021. Thus, the company believes that the 'sprint and recharge' approach is the best way for them to manage the pace of the cryptocurrency industry and still ensure employees could all take the time off they need. (Taylor, 2022)

However, some professionals are sceptical towards the policy's effectiveness in the long run as it does not solve the root causes of low retention rate. Just recently in March 2023, Coinbase made a public statement to accelerate operationalizing its "Go Broad, Go Deep" international strategy in globally recognized markets across six continents over the next eight weeks. (Murugesan & Duff Gordon, 2023) The current "recharge weeks" policy may be effective now but as the workload becomes more intense, it is unclear if Coinbase can afford to give more recharge weeks and whether the recharge weeks will be sufficient to keep employees from leaving the company to seek for a better work life balance. As the company is working towards a rapid expansion, it will be increasingly difficult to let employees take recharge weeks so often. (Berlin, 2022) Hence, Coinbase needs to reevaluate their retention policies to ensure the retention issues are not just dealt with temporarily but resolved permanently. This report thus also serves to provide **meaningful data insights** to help Coinbase come up with **more effective long-term and short-term retention policies**, targeting the root causes of attrition identified using our analytical models.

1.2. Defining Problem & Approach

1.2.1. Current Situation

From a survey conducted, Coinbase was listed as one of the companies with the fastest turnover rates. (Murphy, 2022) Furthermore, the ongoing global tech worker shortage makes retaining developers (a role which requires high technological expertise) in Coinbase more challenging. Having a high turnover rate directly affects the profitability and revenue generation of a company. (Turner, n.d.) This is further supported by Coinbase's statistics, where revenue has been constantly dropping, as low as $576 million in the 3rd Quarter of 2022. (Sigalos, 2023) From a non-intangible aspect viewpoint, high turnover results in lowered workplace morale and quality of work drops. This may lead to a trend of continually decreasing productivity, which is suboptimal for the company.

1.2.2. Business Problem

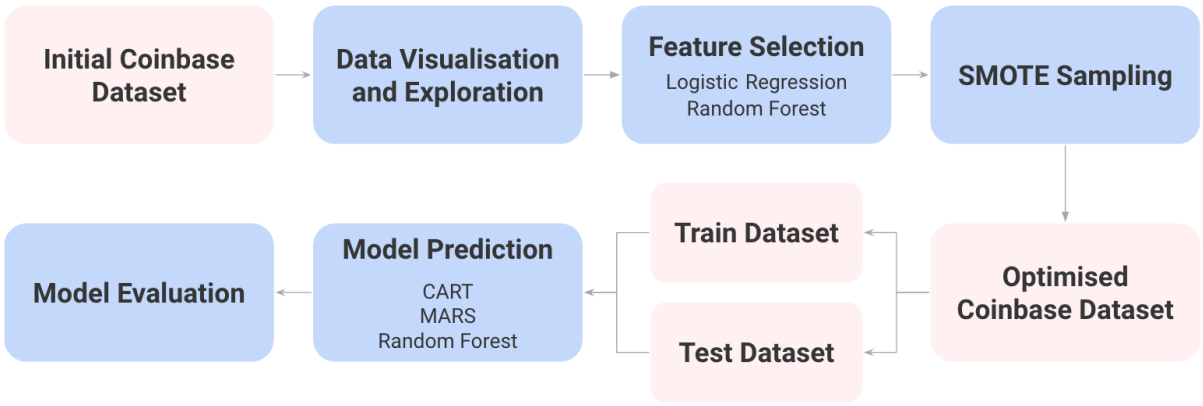
With reduced revenue and increasing costs in recent quarters, profitability has been hard-hit. Now, if Coinbase is to progress on their "Go Broad, Go Deep" international strategy as planned, they would need to set aside a huge sum of money, which will require the company to lower their cost in other areas. In particular, Coinbase can achieve greater cost-savings by focusing on retaining their existing employees, especially developers. The cost-savings can then be reallocated to fund this expansion, which eases their finance strains. Thus, it is crucial to identify employees in Coinbases's development team that are at high risk of leaving, and to then undertake various short gap and long term actions to meet these employees evolving needs, so as to increase retention rates. The secondary benefits of increasing retainment include building a positive work culture, which allows for sustained or even increased productivity.

1.2.3. Opportunity in the Crypto Industry

In the Cryptocurrency Industry, many companies have yet to undertake actions to promote employee retention and thus face the issue of skilled IT worker shortage. (Mearian, 2019) Solving this business problem would hence give Coinbase a competitive edge in terms of being able to attract talented developers. This is supported by Glint's Employee Wellbeing Report where findings show that many employees highly value employers who care about their wellbeing. With our solution, Coinbase will be able to better retain employees so that there will be sufficient manpower resources to ensure continued sustainability and innovation during their expansion.

1.2.4. Our Approach

We will start by identifying significant factors that influence the decision of employees to leave Coinbase. Among these factors, we can also highlight the areas employees feel that Coinbase is underperforming. We can then use supplementary datasets to provide a deeper analysis on these factors, so as to provide more substantial insights on how Coinbase can move forward and meet their needs.



The diagram above illustrates the machine learning workflow that we will be performing to obtain our final machine learning model. Each of the stages in the workflow will be covered in detail in the next few segments.

Upon developing our final model, we will be developing an AI based app, that uses our predictive model, to identify in the short-term, employees who are at high risk of leaving and point out possible key causes to the HR personalised to each individual. This would help HR to narrow down and focus on how to better meet their important needs, and employ various suggested stop-gap solutions that will be more effective. The app will also recommend long-term actions the company can undertake, to prevent employees from leaving. This information will be presented in a dashboard format for ease of visualisation and understanding that Coinbase's HR team can then use to solve this business problem. Over time, HR can leverage on the dashboard to measure the effectiveness of their retention policies.

1.2.5. Key Objectives

We seek to answer these interesting key business questions:

1. What are the primary factors contributing to the high employee turnover rate at Coinbase?

2. What are the factors that are important to developers' job satisfaction and how can Coinbase leverage this information to fuel their growth and better retain more developers?

3. How can Coinbase design customised retention plans that take into account the particular requirements and worries of various employee groups?

4. How can Coinbase track and assess the effectiveness of its retention guidelines over time and make any necessary data-driven adjustments?

2. Data Preparation

2.1. Dataset Description

Our main dataset is "Employee Attrition and Factors" obtained from Kaggle. It is a comprehensive dataset with 35 columns and 1470 rows. In order to make the dataset more relevant to Coinbase, we removed the following columns:

- "DistanceFromHome" because Coinbase is a remote first working culture, thus distance from home is not important in this context
- "EmployeeCount" is all 1
- "Over18" is all "Yes"
- "StandardHours" is all 40 hours
- "EducationField" is redundant as there is another column named "Education"
- "EmployeeNumber" is arbitrary ID to identify employees
- "HourlyRate", "DailyRate", "MonthlyRate" are redundant because there is a "MonthlyIncome" column already

This will be our **Coinbase Dataset** to use for our project. It has the remaining 26 columns, {Age, Attrition, BusinessTravel, Department, Education, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager}.

2.2. Data Cleaning and Preprocessing

Our **Coinbase Dataset** has no missing values. In our data cleaning process, we performed one-hot encoding to convert classes within each categorical variable into numerical numbers. Then, we change the data type of these variables to factors using the factor() function. This ensured that the data was in the correct format for analysis, and that the model was able to use all available data.

2.3. Data Visualisation and Exploration

We created visualisations that would help us better identify trends of Attrition against other factors and these are some key insights. Attrition = 1 refers to employees who have left Coinbase and Attrition = 0 refers to employees who stayed in Coinbase.

From Figure 1, we noted that the younger demographic were more likely to leave the company compared to those of the older age group.

When it came to Business Travelling (Figure 2), we also noted that as employees were sent on business travels more frequently, the likelihood of attrition increased too.
Across different departments, Marketing & Communication (Department 4) has the highest attrition rate (Figure 3) while Security and Privacy (Department 6) has the lowest attrition rate.

We did however also note a relatively flat and similar amount of attrition for different education levels of employees and also their gender. (Figure 4 and 5). However, a consistent number of attrition did not mean that the factor was not important at determining attrition, rather, it just showed that there was consistent attrition regardless of the level of said factor.

Figure 6 identified a decreasing trend on the proportion of employees who decide to leave Coinbase across an increasing level of job involvement required. This suggests that employees who need to be more involved in their jobs, are more likely to stay as compared to those who are not required to be as involved.

From Figure 7, we can observe that the business roles such as Marketing & HR (Job Role 6), Finance (Job Role 3) and Product Specialist (Job Role 1) have the highest attrition rates as compared to the technical roles like Information Security (Job Role 4) and Software Engineers (Job Role 0). However, when looking at absolute numbers of employees who left in each role, Software Engineers (Role 0) is ranked top. On the other hand, management roles across all departments (Job Role 5) have the lowest attrition rate and absolute number of employees who left as compared to all other roles.

There is an overall decreasing trend of attrition over an increasing level of job satisfaction observed in Figure 8. This suggests that employees who stay in the company have higher job satisfaction as compared to those who decide to leave, emphasising that the job satisfaction is very likely to influence the employees' decisions to leave Coinbase.

In Figure 9, a low environmental satisfaction of 1 gives rise to the highest attrition rate while there is no significant difference in the attrition rates from environment satisfaction of 2 to 4. This suggests that employees may not be greatly motivated to stay in Coinbase simply because of a satisfied work environment. This could be due to the remote-first working culture in Coinbase that reduces the importance of an office environment. As employees have the freedom to choose where they want to work, Coinbase just needs to ensure that satisfaction of the office environment is maintained above 2.

From Figure 10, we have observed that for the "Monthly Income" variable, employees who retained in Coinbase received a mean monthly income of approximately $5000. Meanwhile, those who left received a mean monthly income of approximately $2500. Hence, this suggests that employees who receive a generally lower income are more likely to leave.

Figure 11 found that Divorced and Married employees have a similar proportion(approximately 12.5%) of employees who leave. Interestingly, we found that for Single employees, there is a doubled proportion(approximately 25%) of employees who leave.

Analysing Figure 12, we found that approximately 25% of employees who have worked at 5 or more companies before, leave Coinbase. This is a much higher proportion than employees who worked at 4 or less companies. The exception to this, is that employees who have worked at only 1 company before, are also relatively highly likely to leave.

From Figure 13, we found that 30% of employees who worked overtime, left Coinbase. Meanwhile, out of the employees who did not need to work overtime, only 15% left the company. This suggests that employees are more likely to retain in Coinbase, if they are not required to work overtime.

Figure 14,15,16,17 discusses Attrition with relation to Job Level, Relationship Satisfaction, Performance Rating and Stock Options Level respectively. It was inferred from the count of people who left that Stock Options might play a part with Attrition, as majority of those who left had low to no stock options offered to them.

Seen in Figure 18, more than 50% of employees who left Coinbase have less than 10 years of working experience. These employees include fresh graduates who just entered the workforce while more than 50% of employees who remained in Coinbase have at least 10 years of experience.

Figure 19 depicted the relation between Attrition and the Number of Trainings an employee undertook in the past year. We could identify that there was a similar distribution of training in terms of people who left and stayed, which might suggest that it was not a key factor in determining attrition.

In Figure 20, we can observe that employees who have a lower work life balance (WorkLifeBalance at 1 and 2) have a higher tendency to leave than employees who have a better work life balance (WorkLifeBalance at 3 and 4) On the other hand, it is apparent that the attrition rate is higher among employees with an extremely good work life balance (WorkLifeBalance at 4) than employees with WorkLifeBalance at 3, hence other factors apart from work life balance influence these employees to leave.

Figure 21, More than 50 % of employees who left Coinbase worked in the company for less than 5 years but more than 50% of employees who remained in Coinbase worked for more than 5 years there. This implies that Coinbase should focus on retaining employees who have just joined the company for less than 5 years as they are more at risk of leaving.

The boxplots in Figure 22 show that more than 50% of employees who stayed at Coinbase have been in their current role for long years while more than 50% of employees who decided to leave Coinbase just recently transitioned into the current role a few years ago. This potentially suggests that most employees in Coinbase prefer to remain in their current roles and job scope and a change into a new role may deter employees from staying in Coinbase.

Figure 23 depicts the attrition rate compared to the years since an employee's last promotion. This was a surprising statistic, showing us that the majority of employees who left were actually promoted within the last 5 years. This meant Coinbase had trouble retaining talented employees, who were possibly leaving for greener pastures.

Figure 24 depicts that among employees who left Coinbase, 50% have an average of 2 years with their current managers as compared to those who stayed in Coinbase. This suggests that working relationships with managers do have a significant influence on the attrition decision of employees.

3. Identification of Key Features

3.1. Logistic Regression

Logistic Regression is implemented as one of our feature selection methods to remove features that exhibit high multicollinearity with other independent variables based on the Variation Inflation Factor (VIF). Iterative elimination of factors with the highest VIF until we are left with a subset of features with VIF scores below 5 is important as a high multicollinearity refers to a high linear dependence on 2 or more other independent variables, which undermines its statistical significance and results in less reliable statistical inferences.

Upon creating the Logistic Regression on the **Coinbase Dataset**, we discovered that "Department" has a **-1 alias coefficient,** suggesting a perfect correlation with other variables such as "Job Role". We found that each department is mainly made up of a specific role, thus there is no value in keeping "Department" in the dataset. Then, we analysed the VIF scores of the remaining factors and found that "Job Level" has a GVIF of 43.259, which represents a significantly high multicollinearity. Our team decided to remove it as it is closely related to the "Monthly Income" - when an employee's position increases, monthly income will inevitably be higher. Hence, "Job Level" was also removed, leaving 24 variables in our final dataset for the next step of feature selection. With this set of variables in our revised logistic regression, we further shortlisted the 17 most important features based on the p-values and absolute values of coefficient estimates. A p-value lower than 0.05 significance level and larger coefficient estimates suggest that the variable is more important. (Figure 40, 41, 42)

3.2. Random Forest

Random Forest is chosen as our second Feature Selection model as it is able to automatically and quickly filter through a large number of variables and remove unimportant features, without having us to manually check through each variable, which is prone to human error. Furthermore, it is not affected by multicollinearity, thus the trained model can still produce highly accurate results despite the presence of correlated variables.

Variable Importance in Random Forest is obtained by measuring and ranking the extent of mean decrease in Accuracy and Gini Impurity when each variable is permuted within the bootstrap

samples, reducing its influence on the prediction model. Hence, variables with higher mean decrease in either Accuracy or Gini Impurity are more important. By evaluating every factor based on the 2 metrics, 19 out of the 26 variables are selected by the Random Forest Model as important. Furthermore, the Random Forest trained on these 19 features (87.07%) has proven to produce higher predictive accuracy than Random Forest trained on all 26 features (86.39%), thus it is evident that these 19 features are significant (Figure 43 to 48).

3.3. Final Key Predictors

```
> print(top_features_dataset)
 [1] "JobRole"                 "BusinessTravel"          "OverTime"                "WorkLifeBalance"
 [5] "MaritalStatus"           "EnvironmentSatisfaction" "StockOptionLevel"        "Education"
 [9] "YearsSinceLastPromotion" "YearsAtCompany"          "YearsWithCurrManager"    "JobSatisfaction"
[13] "MonthlyIncome"
```

Among these 2 sets of variables, we find that these 13 variables (as shown in the screenshot above) are common and present in both sets. Hence, our **Coinbase Dataset** is now reduced to only these 13 variables, which are identified as the final set of important variables to be used in the prediction of attrition in Coinbase.

3.4. Final Dataset Processing

The **Coinbase Dataset** is severely imbalanced, with 1233 instances of 0 (employees who stayed), and 237 instances of 1 (employees who left) in Attrition (Figure 49). Therefore, there is a need to balance the dataset before training our models to prevent overpredicting the majority class. We used a custom command to imitate SMOTE (Synthetic Minority Oversampling Technique) to balance our dataset. Our SMOTE Balancing creates 862 instances of 0s and 862 instances of 1s, making it a 50/50 balance with 1724 total instances.

While oversampling and undersampling can help to balance the dataset, they also have some limitations. Random oversampling can result in overfitting, as the model is trained on duplicated instances. Random undersampling can result in information loss, as important instances may be removed from the dataset. Moreover, both methods do not take into account the distribution of the features and can generate irrelevant or uninformative instances. (He & Garcia, 2009)

SMOTE is preferred over oversampling and undersampling because it generates synthetic instances based on the distribution of the features, which can lead to better generalisation and prevent overfitting. Moreover, SMOTE can create relevant and informative instances, which can improve the performance of the model. (Chawla et al., 2002)

Finally, we conducted a 70-30 train-test split on this SMOTE balanced dataset, producing a training dataset of 604 instances of 0s and 1s each, and a testing dataset of 258 instances 0s and 1s each.

4. Model Development

4.1. CART

4.1.1. Initial Model

Classification and Regression Trees (CART) is a machine learning algorithm that is known for its high predictive power. It can handle categorical and numerical variables well in our **Coinbase Dataset**. CART uses values of certain predictor variables (also known as the split condition) as a guide to partition the full dataset into smaller subsets. By doing so, a decision tree is created, and forms a visual representation of binary decisions that lead to the predicted outcomes.

In the first phase, known as the growth phase, CART considers and tests all predictor variables, to identify the best binary split. The best split is determined by obtaining a pair of child nodes with the highest purity. After which, multiple next-best splits continually occur, until a stopping criteria is met.

We started off by growing our CART model to the maximum, however due to overfitting, we pruned it by using a complexity penalty (cp), to obtain the optimal tree. (Figure 50) Our final model is a smaller tree, which can be effectively used to categorise if an employee is likely to leave, based on the employee's information in accordance with the splits. (Figure 51)

4.1.2. Performance Evaluation

We produced a confusion matrix for our CART model, for both the train set and the test set.

Trainset	Testset

`> cart.cm_train`
`cart.yhat_train`
`hr.trainset$Attrition 0 1`
`0 603 0`
`1 0 603`
	`> cart.cm`
	`cart.yhat`
	`hr.testset$Attrition 0 1`
	`0 219 40`
	`1 0 259`

An accuracy of 100% on the train set, with no False Positive (FP) and False Negative(FN) rate were achieved. (Figure 52) Meanwhile, the testset accuracy reached 92.2%, FP rate of 15.4%, and a FN rate of 0, reaching an almost 8% drop from the training accuracy. (Figure 53)

4.2. MARS

4.2.1. Initial Model

We explored the idea of using MARS to better predict the attrition rate with our dataset as MARS allows us to account for non-linear relationships between variables. We ran MARS with a degree of 4 on the train set (Figure 54), meaning interactions of up to 4 different variables are considered in the regression model. Following that, we ran evimp to find the variable importance of the factors in our dataset (Figure 55). We then generated a confusion matrix of each prediction, in order to properly compare and evaluate the performance of MARS on our dataset.

4.2.2. Performance Evaluation

The confusion matrix for the MARS model was as follows:

Trainset	Testset
mars_pred_factor_train 0 1 0 498 82 1 105 521	mars_predicted 0 1 0 205 37 1 54 222

The train set accuracy reached 84.4%, 17.4% for FP rate and 13.6% for FN rate (Figure 56) For the test set, an accuracy of 82.4%, a FP rate of 20.8% and a FN rate of 14.3% were produced (Figure 57). There is no overfitting as the decrease in accuracy between test set and train set is small at 2% and FP and FN rates increase slightly by 3.4% and 0.7% respectively too.

4.3. Random Forest

4.3.1. Initial Model

Random Forest (RF) is a machine learning model that leverages on the idea of Bootstrap Aggregation (Bagging). Using replacement from the original data, each decision tree is constructed using a separate random sample and then independently trained. Given that this is a categorization issue, the model will use the consensus of all the trees as the outcome. (Aggregation)

RF can lessen biases and the impact of a potent predictor variable by bagging. Additionally, it is able to reduce single tree bias (as found in CART) by taking the consensus of 500 trees and does not overfit. As a result, it can offer better predictive accuracy on an unknown test set.

One disadvantage of RF is its black box nature, which makes it difficult to interpret how predictors influence predictions. However, RF has an inbuilt permutation feature importance that allows us to gain insight into the features with predictive power. The idea is that important features contain valuable information, and if that information is destroyed by randomly shuffling its values, prediction accuracy will suffer. If the decrease in accuracy is significant, the predictor has a significant impact on predictions.

4.3.2. Performance Evaluation

Finally, generating the confusion matrix for the RF model, we got the following conclusions.

Trainset	Testset
Reference Prediction 0 1 0 603 0 1 0 603	Reference Prediction 0 1 0 247 0 1 12 259

For our train set, we managed to get 100% accuracy, with 0% FP and FN rate, just like the CART model. (Figure 58) Fortunately, our RF model on our testset actually managed to get an accuracy of 97.7%, with an FP rate of 4.6% and a FN rate of 0%. The 2.3% deviation of test set accuracy from the train set is not significant, thus the model is not overfitted (Figure 59).

5. Evaluation of Models

5.1. Evaluation based on Metrics

Model	Accuracy	False Positive Rate	False Negative Rate	Precision	Recall
CART	92.2%	15.4%	0%	86.6%	100%
MARS	82.4%	20.8%	14.3%	80.4%	85.7%
Random Forest	97.7%	4.6%	0%	95.6%	100%

Based on 5 metrics (Accuracy, FP Rate, FN Rate, Precision and Recall), we identified that Random Forest (RF) is the best model for us to adopt when predicting employee attrition in Coinbase. The main factors which put RF above the rest of the models are its high accuracy of 97.7%, which was 5.5% more than CART, paired with the absence of False Negatives. This meant that we would not wrongly predict an employee would stay when he was actually leaving.

As such, we decided to adopt Random Forest as our main model.

5.2. Limitations of Models

There are two main limitations of our model.

Firstly, it was important to note the 4.6% False Positive rate of our model. This false positive rate meant that for every 1000 employees, we might predict 46 of them who might be leaving the company, but in reality, they were happy with their current job and had no intention to leave. Although it was preferable to have it at 0%, False Positives are unavoidable.

Secondly, there was a 2.3% chance that our model would predict an employee's attrition likelihood wrongly, meaning there is a possibility of missing an employee who needs our support.

As our models are running on the default parameters in-built into the functions, we can perform hyperparameters tuning to find the optimal hyperparameters for the respective MARS and Random Forest that will most likely help to lower this false positive rate further. For example, we can find the optimal number of features to select for each split and number of classification trees to create using grid search. Meanwhile, for MARS, we can perform 10-fold cross validation to obtain better predictive results and accuracy.

5.3. Conclusion

Although the limitations of our model might seem worrying, we identified that False Positives were acceptable. This is because it was perfectly fine to support employees which we predicted to

be leaving, but actually had no intentions to leave. This would allow us to have an even higher likelihood of retaining them. Due to the lack of false negatives, along with the still relatively low false positive rate, we are confident our model would still be useful for Coinbase. It is also accepted that the 2.3% of inaccurate prediction would be manageable for our HR team, and therefore none of the identified limitations of our model poses a significant challenge for us.

6. Recommendation/ Solution

6.1. Predictive App

6.1.1. Description

To implement our machine learning models and bring our analytics to fruition, it became imperative to develop a platform that would assist the Human Resources Department at Coinbase in identifying employees who are likely to leave, enabling Coinbase to take proactive measures to retain them, such as providing additional training, better benefits, or adjusting their compensation.

The app's personalised recommendations for each employee can boost retention even further by providing specific actions that can be taken to address the unique concerns of each employee. By addressing the specific factors that may be contributing to an employee's dissatisfaction or disengagement, the company can increase their job satisfaction and decrease their likelihood of leaving.

To build our Webapp, we have used our random forest model that gives us an accuracy rate of 98% . We have a 0% false negative rate which ensures that 100% of the employees who are at a risk of leaving the company will be correctly identified and Coinbase will be able to address their issues before it is too late. This Webapp has been built using the Streamlit module in Python and has been published online. You can use it by scanning the QR Code or copying the URL given below:

https://jaijindal-bc2407-s1team7-attrition-ap-bc2407-s1team7-app-7bzpob.streamlit.app/

Overall, the impact of this app can be significant for Coinbase as the company will be more aware of the profile of employees that are highly likely to leave and the possible recommendations on what Coinbase can do to retain them. This will be especially useful when there is a specific group of well-performing employees that Coinbase wishes to keep. Hence, with the help of the App, Coinbase can achieve better retention of valuable employees, reduction of turnover, and overall improvement in employees' productivity, morale, and ultimately - profitability.

6.1.2. User Interface

This Webapp has 3 main features:

1) Company Employee Attrition Prediction based on Profile

The app provides the user with a dropdown menu to select a Coinbase employee whose attrition probability they want to predict. Once a name is selected, the app automatically assigns and displays a set of feature values for that employee loaded from Coinbase's databases, and then uses a pre-trained random forest classifier to predict their attrition probability. (Figure 60)

2) Highly Personalised Recommendations

Based on the employee's features values, the Webapp then correctly identifies the problems/variables that the employee is likely influenced by that led to the risk of him leaving Coinbase. Then, for every highlighted problem, the App makes highly personalised recommendations to guide Coinbase to change specific work policy to retain that specific employee. (Figure 63)

3) Variable Toggling to explore Options

The toggling function in the app allows Coinbase to explore various extent of changes in values of the factors identified in the recommendation section (in part 2 above) and then rerun the model to see how much each variable has to be adjusted to change the attrition prediction. Coinbase can hence gain a deeper insight on the most optimised changes they could make to keep the employee happy while ensuring minimal losses. (Figure 61, Figure 62)

6.2. Dashboard Visualisation

6.2.1. Purpose of Dashboard

The second part of our solution is a dashboard. This dashboard can be used by Coinbase's senior management, to measure and track employee attrition statistics. By viewing these statistics, one is able to measure the effectiveness of current ongoing retention policies. This dashboard can also be utilised by the Human Resource team. They will be able to pinpoint which job role has the highest proportion of attrition, and identify the key features that contribute to these respective rates. The insights gained, can be used to develop future job-role wide strategies to further lower attrition.

6.2.2. Dashboard Features

From the main dashboard, we can view certain Key Performance Index(KPIs), such as the total number of employees, number of employees that have left, as well as the attrition rate for each job role. (Figure 64)

We can proceed by navigating to one of the four statistical domains: Demographic, Wellbeing, Performance and Others. In each domain, we can view the related statistics, filtered by the job role specified by the user.

1) Demographic(Figure 65) - In this dashboard, we can view Attrition statistics regarding Age, Gender, Marital Status, Education, Experience & Total Working Years.

2) Wellbeing(Figure 66) - In this dashboard, we can view Attrition statistics regarding Job Involvement, Job Satisfaction, Work Life Balance, Environmental Satisfaction & Relationship Satisfaction

3) Performance(Figure 67) - In this dashboard, we can view Attrition statistics regarding Monthly Income, Percent Salary Hike, Performance Rating, Overtime & Business Travel

4) Others(Figure 68) - In this dashboard, we can view Attrition statistics regarding Years in Current Role, Years in Company, Years since last Promotion, Years with Current Manager & No. of training times last year

After viewing the statistics in a particular domain, there is a navigation button on the top right of the screen, to toggle back to the main dashboard. After which, we can select an alternative domain to view.

6.2.3. Dashboard Usage

After specifying which job role to focus on, we can observe the barcharts or box plots of the related attributes. If we identify that the proportion of attrition is much higher in a particular category, for example, viewing the "Overtime" attribute, 56.76% of employees who had to work overtime left Coinbase, whereas only 24.49% of employees who did not have to overtime, left(Figure 67). We can then conclude that working overtime is one of the key contributing attributes to high attrition rates in that job role. This implies that current policies or regulations regarding overtime work are not effective in lowering attrition, and that a new or improved job-role wide strategy can be devised to reduce these rates.

6.3. Limitations and Mitigation

While our model is effective in highlighting potential employees who are leaving and the possible reasons for it, our model poses some limitations in recommending retention policies to the senior management and HR team, especially for subjective and qualitative factors like Job Satisfaction. These causes are hard to address as there are many underlying factors that constitute Job Satisfaction. Hence, in order to implement more targeted policies that aim to improve Job Satisfaction, our group used **a second dataset from "Stack Overflow Developer Survey"** in Kaggle to help us uncover deeper insights into Job Satisfaction. The survey consists of 154 questions from 51400 respondents but we only focus on full-time professional developers working in similar conditions as Coinbase - a less than 5000 employees, publicly-traded corporation. Due to the high number of developers who are the core of Coinbase's daily operations, leaving the company, it is essential to conduct deeper analysis in this aspect. There are 15 "Important Benefits" full-time developers value and a series of data visualisations were plotted in Appendix 9.2. Using

these responses, we seek to evaluate what developers of different job satisfactions look out for, thus allowing us to have a better understanding on the types of retention policies to implement to cater to these needs and wants.

6.3.1. Data Exploration and Visualisation

In Figure 25, we identified that Retirement benefits were significant to a sizable amount of developers. This was also seen in Figure 26 and 27 with regards to Vacation Days and Health Benefits like access to health insurance. However, Figure 28 indicated low interest in a charitable match of their retirement accounts (401K).

Figure 29, 30 and 31 found that an employee's Annual Bonus, Expected Working Hours and Equipment provided to them were essential in providing them with substantial job satisfaction. Long-Term Leave showed an increasing trend towards developers with high job satisfaction in Figure 32. Hence, even though a small proportion of employees selected it, it is an increasingly valued benefit. On the other hand, in Figure 33, having a Private Office was shown to have little to no effect on their Job Satisfaction.

Figure 34 was great in showing us that Remote Options were very significant in a developer's level of job satisfaction. Fortunately. Coinbase is already a remote-first workplace, meaning we have already succeeded in providing what we deem a basic necessity for our developers. Along with that, providing employees with opportunities for Professional Development (Figure 35), also contributed to their job satisfaction, albeit not as much.

Providing free meals was not important to most developers, as seen from Figure 36. What was important to them was instead Stock Options (Figure 37), which tied well to Figure 17 of the previous dataset, where employees who left usually had little to no stock options. Figure 38 and 39 showed us that Education Sponsorship and Child/Elder Care were not significantly important to developers, we suspect it was due to the fact that the developers hired were already highly skilled, and did not see the need for further education outside of their normal job. Majority of developers were also working remotely, meaning it was not likely they required child/elder care as they were able to manage this while working from home.

Job Satisfaction and Job Seeking Status in the second dataset are also calculated to be highly dependent using a Chi-Squared Test (p-value less than 0.05 significance level), which directly reflects the strong relationship between Job Satisfaction and Attrition in our **Coinbase Dataset.**

6.3.2. Conclusion

In conclusion, we shortlisted a list of important benefits based on 2 criteria - if a majority of respondents selected this benefit (more than 50%) or if there is a significant trend across different levels of Job Satisfaction.

Benefits that fulfil both conditions are {Retirement, Vacation/days off, Remote options}, while benefits that only fulfil one of the conditions are {Annual Bonus, Expected work hours, Health benefits, Long-term leave, Equipment}.

Moreover, benefits like Retirement are valued more by employees with lower Job Satisfaction while benefits like Remote options and Vacation/days off are valued more by employees with higher Job Satisfaction. Since Coinbase has a remote-first culture and have already started implementing "recharge weeks", more attention can be placed on rolling out better retirement incentives to increase employees' Job Satisfaction. Although Long Term Leave, Annual Bonus and Equipment are not every employees' top concerns, HR should not compromise on employees' flexibility to take long term leave, annual bonuses and provision of high quality equipment in office so as to maintain the willingness to continue in the company among the highly satisfied employees. Lastly, Expected work hours and Health benefits are widely chosen by all employees regardless of their job satisfaction, thus Coinbase should prioritise offering these benefits to improve Job Satisfaction of the general employee population.

To sum up, benefits such as Expected work hours, Vacation/days off, Health, Retirement and Remote Options are the primary needs of employees Coinbase should aim to meet. Meanwhile, benefits such as Annual Bonus, Long Term Leave and Equipment can be adjusted depending on the current distribution of employees' Job Satisfaction.

6.4. Future Expansion of Solutions

To enhance the usefulness of our solutions, our App can provide real-time updated insights on the likelihood of attrition among the existing employees. Whenever there are changes in the personal and working information of employees such as having a job promotion, the model will run new predictions and update the App so that Coinbase (HR team) will be notified first-hand if there are employees who have become inclined to leave. Employees will also be categorised according to their likelihood of attrition, and the App will highlight to the HR team which group of employees they need to pay more attention to when implementing retention policies.

In addition, the dashboard can be linked to the App to get access to the prediction data so that Coinbase can forecast and visualise the predicted attrition rates. With this forecast, Coinbase can pre-empt possible problems in the future and thus have more time to think of effective solutions for them. The dashboard can also retain past data, and display it in a graph format. With this, management can view how these statistics change over time. The identified trends and patterns can aid them in analysing if their executed strategies are effectively helping to lower attrition rates.

7. Conclusion

Our original goal was to answer 4 key business questions:

1. What are the primary factors contributing to the high employee turnover rate at Coinbase?

2. What are the factors that are important to developers and how can Coinbase leverage this information to fuel their growth and better retain more developers?

3. How can Coinbase design customised retention plans that take into account the particular requirements and worries of various employee groups?

4. How can Coinbase track and assess the effectiveness of its retention guidelines over time and make any necessary data-driven adjustments?

After deep insight, we found that our analytical models helped us to gain better understanding on the root causes of the high attrition rate that Coinbase is experiencing today. We have identified that not just work life balance, there are other factors such as job satisfaction and business travel that are influencing employees to leave Coinbase.

As the models are able to provide us with a complete perspective of the problem, Coinbase can take the right actions that targets the problem directly. For example, the 4 “recharge weeks” implemented by Coinbase last year may not achieve good results as it only focuses on one of the 13 key factors identified in the report earlier. Although Vacation and Day offs is one way to increase employees’ job satisfaction, the de-incentivising influences from 12 other factors may still outweigh the benefits of this policy.

From our data analysis, it is apparent that in order to achieve maximum effectiveness, a combination of retention policies that targets multiple root causes is required. Therefore, our 2 solutions - App and Dashboard - serve to help Coinbase tackle the retention issues more efficiently. Our App takes into account the combined effects of the different key factors to determine if a particular employee is likely to leave Coinbase or not. The App uses Random Forest model as it produces the highest predictive accuracy among all the machine learning models. The robustness of the predictive model ensures that recommendations of retention policies to Coinbase which rely on the prediction will be effective. As the app also allows HR and senior management to see exactly who might be leaving, it enables them to tailor retention plans not just for a general employee group (e.g. Developers), but it is possible for managers to fine tune plans down to the individual in question. Managers can use the predicted factor of an employee leaving to offer him tailored plans such as additional leave, less business travelling, to help improve his quality of life. This helps Coinbase emphasise how important each and every employee is to them. On the other hand, the Dashboard is a platform for Coinbase to get a bigger picture of the attrition problem and track the effectiveness of the retention policies in place through a wide range of visualisations.

Currently, our attrition rate stands at about 16.2%. With the implementation of the Dashboard and App, we aim to bring the attrition down by about 5%. This would result in a saving of over $3 million/year for Coinbase, at an average yearly pay of $80,000/year for each employee. By tracking the retention rate over the course of the year, Coinbase will be able to see if their prediction is accurate, and how their retention guidelines might have helped to reduce the attrition rate. In the long term, we might also notice a change in the key factors that would make employees leave with the occurrences of new economic events. Using new and updated data, we would be able to fine-tune our model to include new factors in conducting its prediction, ensuring we are able to account for every employee in Coinbase.

In summary, our 2 solutions work hand in hand - Dashboard provides broad and comprehensive insights at the company level while App gives detailed and customised predictions and recommendations at an employee level. Like all applications of machine learning, our solutions can potentially help Coinbase save a lot of costs and increase profits by retaining as many good employees as possible but they should be used responsibly for the sole intended purpose of solving attrition issues.

8. References

Berlin, M. P. I. (2022, March 30). Is a 'recharge break' the key to your employees' happiness? Sifted. Retrieved February 23, 2023, from https://sifted.eu/articles/recharge-week-employee-holiday-bitpanda/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Coinbase Reviews. (n.d.). Glassdoor. https://www.glassdoor.sg/Reviews/Coinbase-Reviews-E779622.htm

ETF Trends. (2022, January 13). A Novel Approach to Employee Retention From Coinbase. Nasdaq. Retrieved February 23, 2023, from https://www.nasdaq.com/articles/a-novel-approach-to-employee-retention-from-coinbase

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Iacurci, G. (2023, February 1). 2022 was the 'real year of the Great Resignation,' says economist. CNBC. Retrieved February 23, 2023, from https://www.cnbc.com/2023/02/01/why-2022-was-the-real-year-of-the-great-resignation.html

Mearian, L. (2019, April 8). Blockchain jobs remain unfilled, while skilled workers are being poached. Computerworld. Retrieved February 25, 2023, from https://www.computerworld.com/article/3387441/blockchain-jobs-remain-unfilled-while-skilled-workers-are-being-poached.html

Murphy, A. (2022, June 30). 20 Companies with the Worst Staff Retention Rates. https://www.insightsforprofessionals.com. Retrieved February 23, 2023, from https://www.insightsforprofessionals.com/hr/recruitment-and-onboarding/20-companies-with-the-worst-staff-retention-rates

Murugesan, N., & Duff Gordon. (2023, March 8). *Our 8-week international expansion drive in 6 continents*. Coinbase. Retrieved April 1, 2023, from https://www.coinbase.com/blog/our-8-week-international-expansion-drive-in-6-continents

Sigalos, M. (2023, February 21). Coinbase reports better-than-expected user numbers even as third-quarter revenue plunges. CNBC. Retrieved February 25, 2023, from https://www.cnbc.com/2022/11/03/coinbase-coin-earnings-q3-2022.html

Taylor, A. (2022, January 17). *Coinbase explains why it gave employees a month of 'recharge time' in addition to Flexible Time Off*. Yahoo! Finance. Retrieved April 1, 2023, from https://sg.finance.yahoo.com/news/coinbase-explains-why-gave-employees-140000539.html

Turner, R. (n.d.). Why high turnovers rate are so detrimental to your business. TechDay. Retrieved February 24, 2023, from https://techdayhq.com/community/articles/why-high-turnovers-rate-are-so-detrimental-to-your-business

9. Appendix

9.1. Dataset 1 - Data Visualisation and Exploration

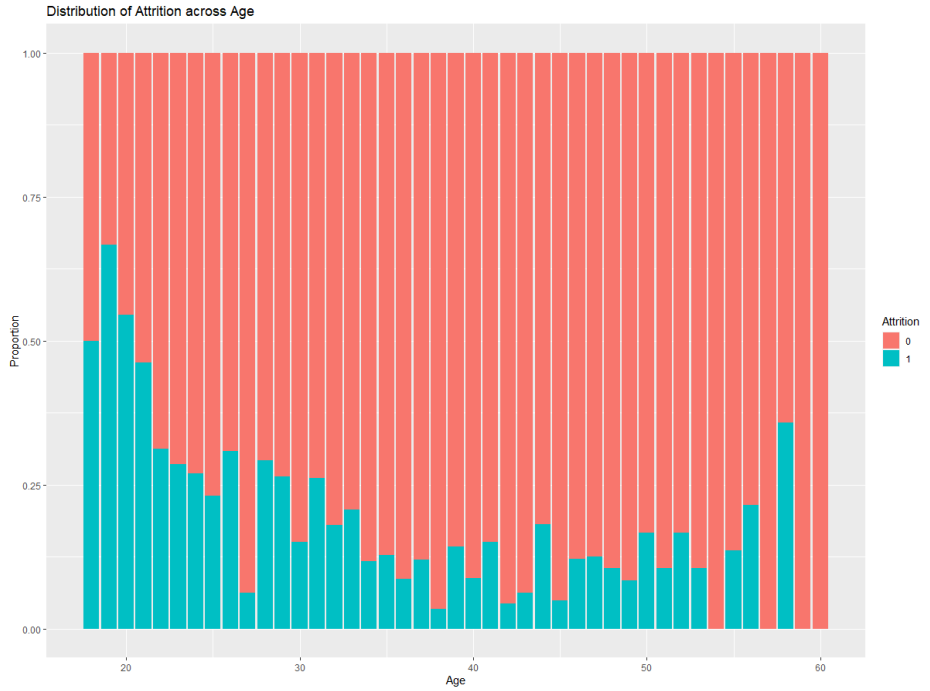


Figure 1 (Attrition vs Age)

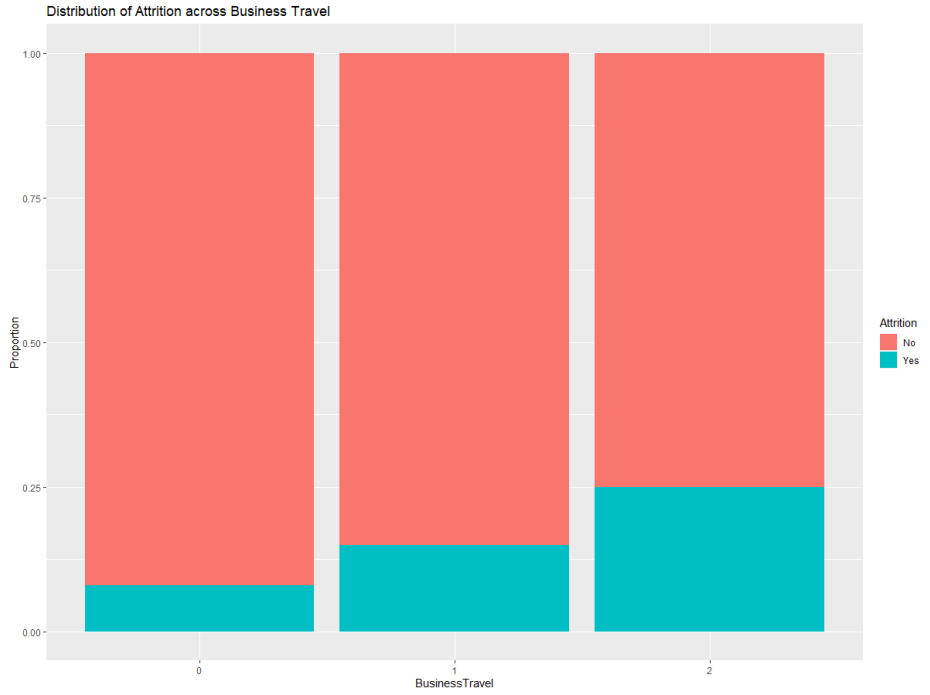


Figure 2 (Attrition vs Business Travel)

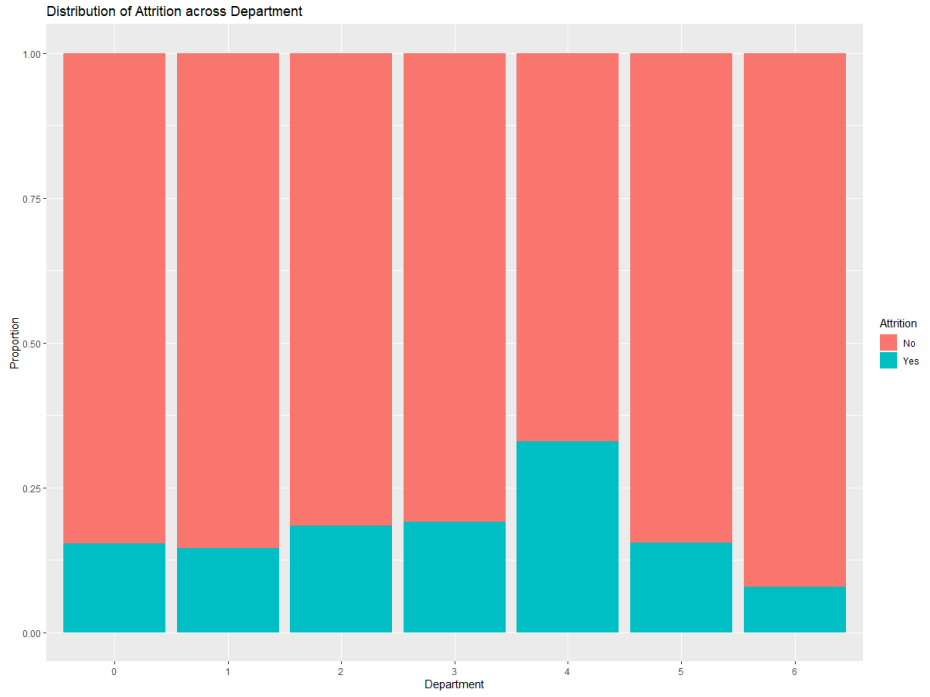


Figure 3 (Attrition vs Department)

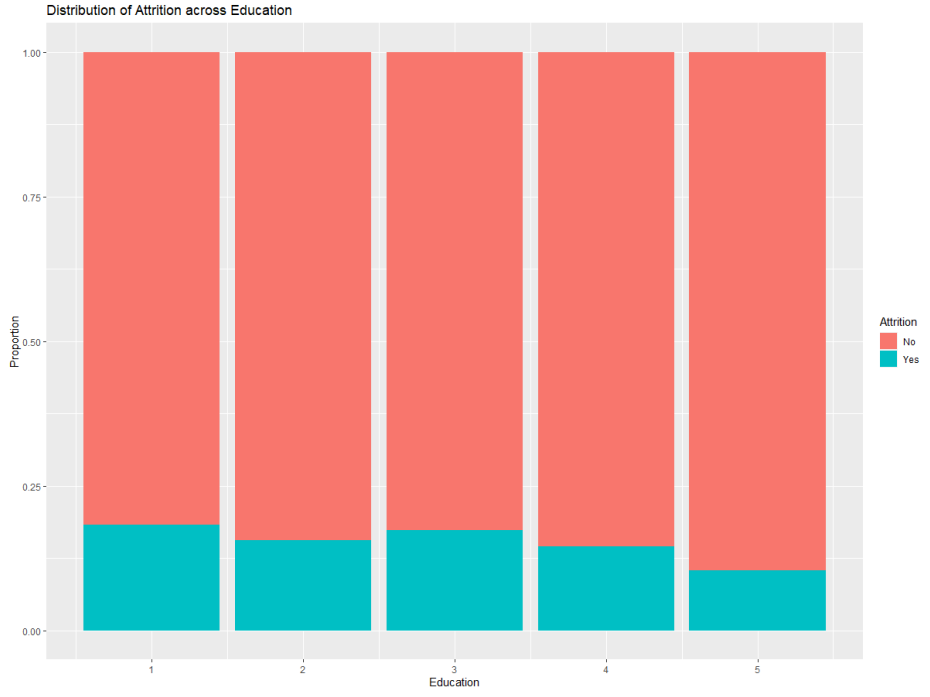


Figure 4 (Attrition vs Education)

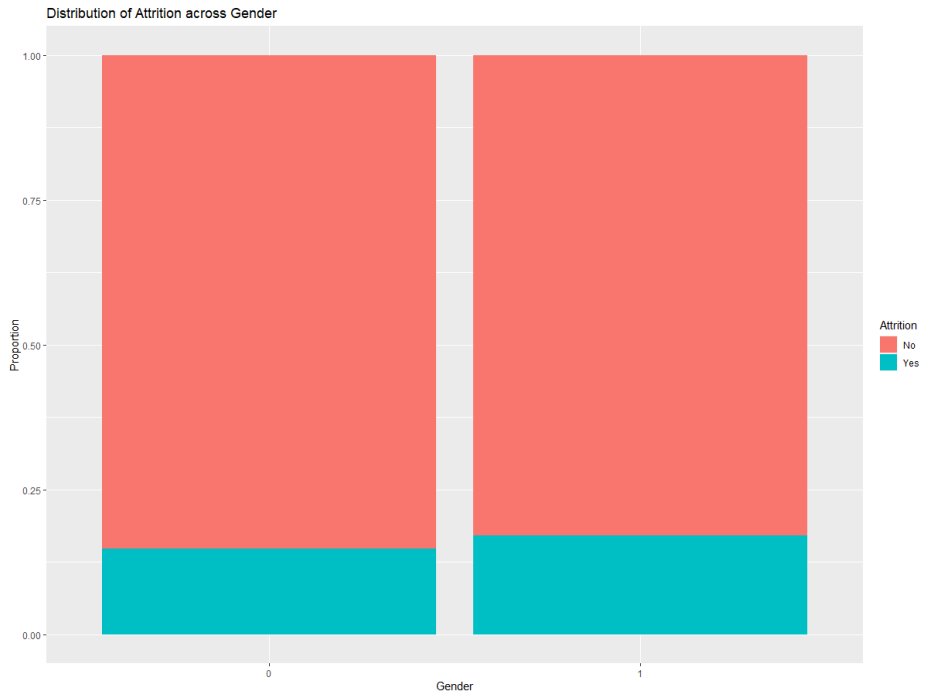


Figure 5 (Attrition vs Gender)

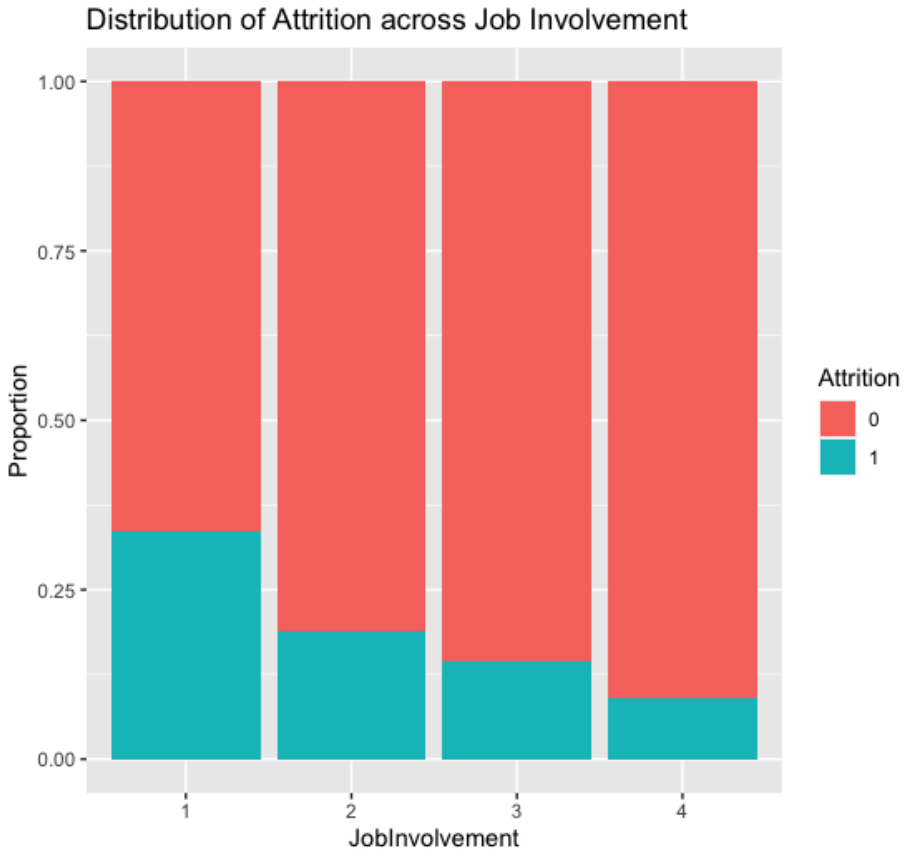


Figure 6 (Attrition vs JobInvolvement)

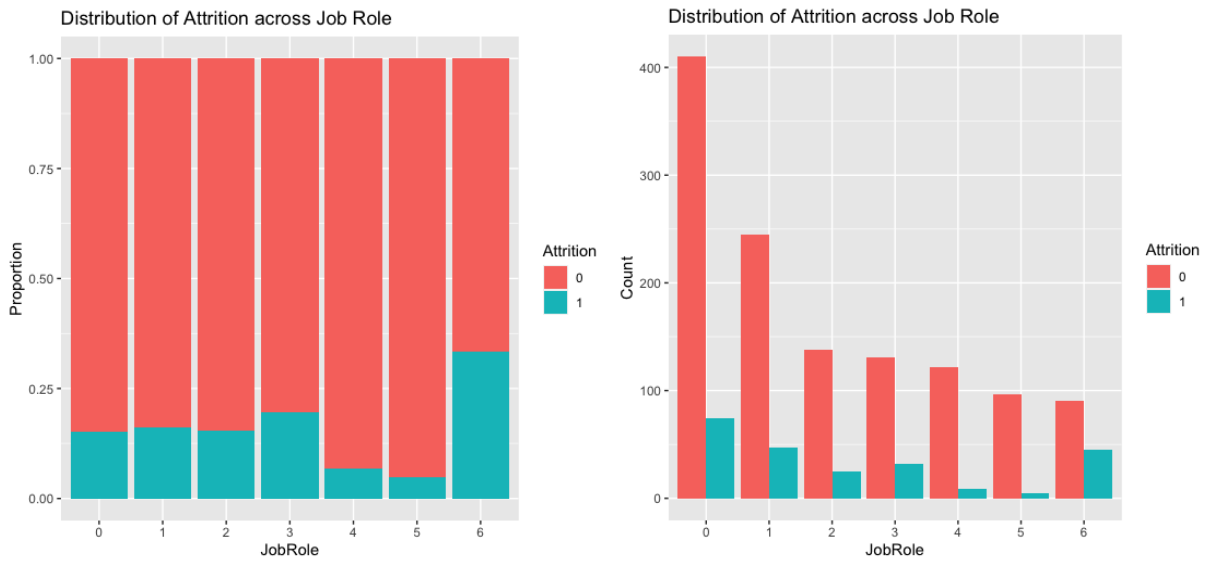


Figure 7 (Attrition vs Job Role)

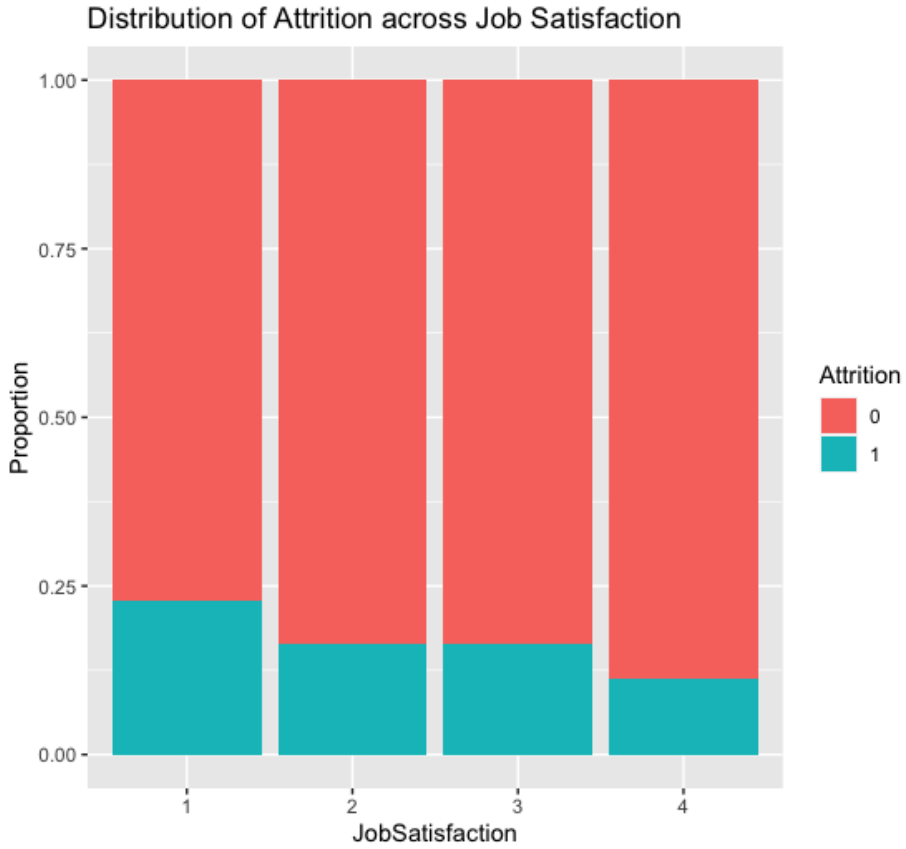


Figure 8 (Attrition vs Job Satisfaction)

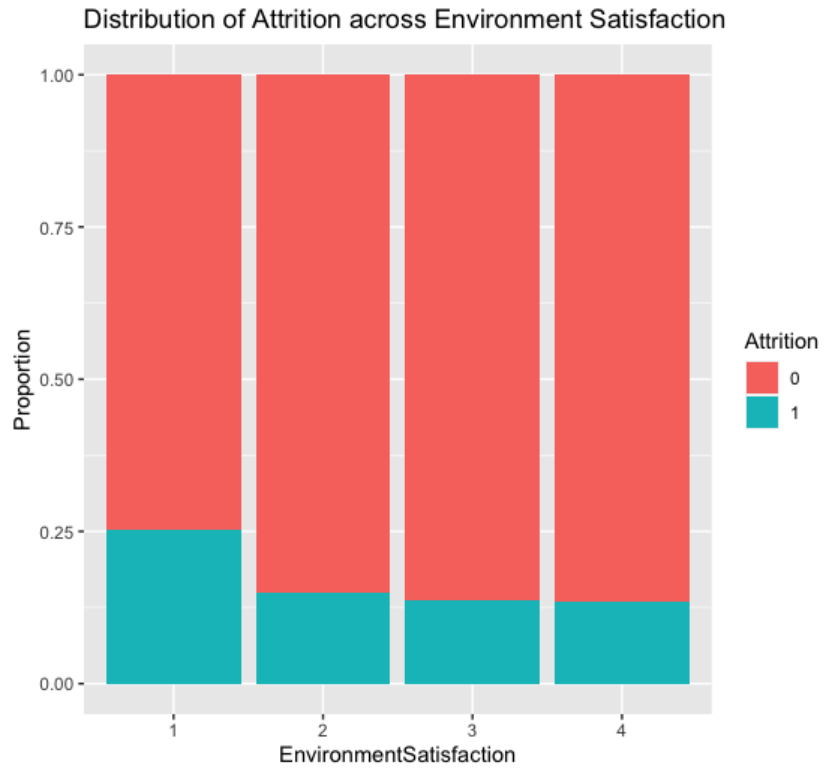


Figure 9 (Attrition vs Environment Satisfaction)

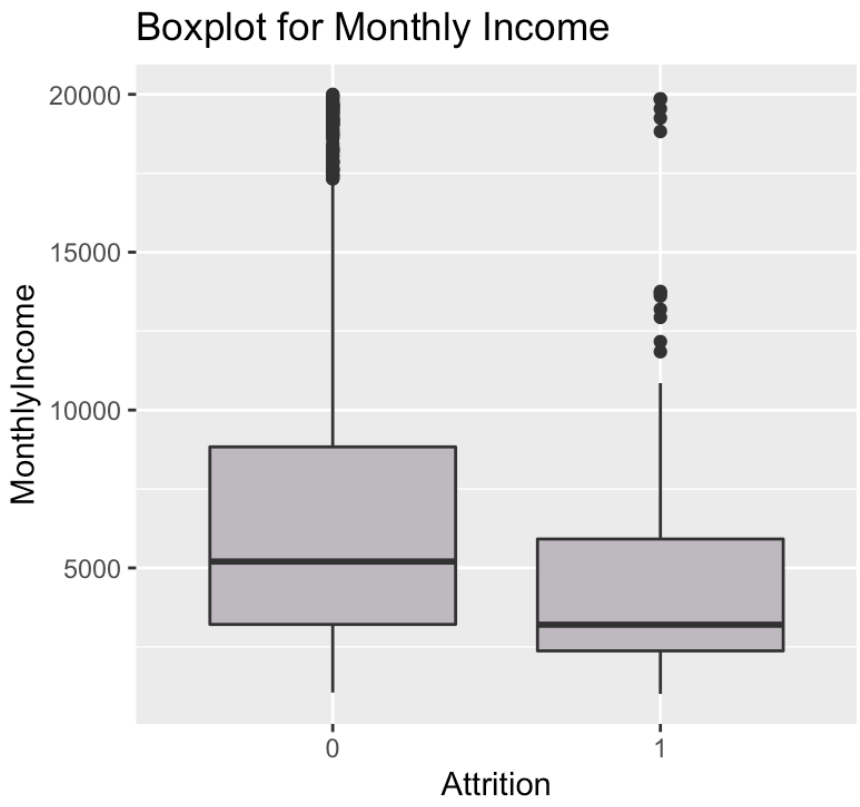


Figure 10 (Attrition vs Monthly Income)

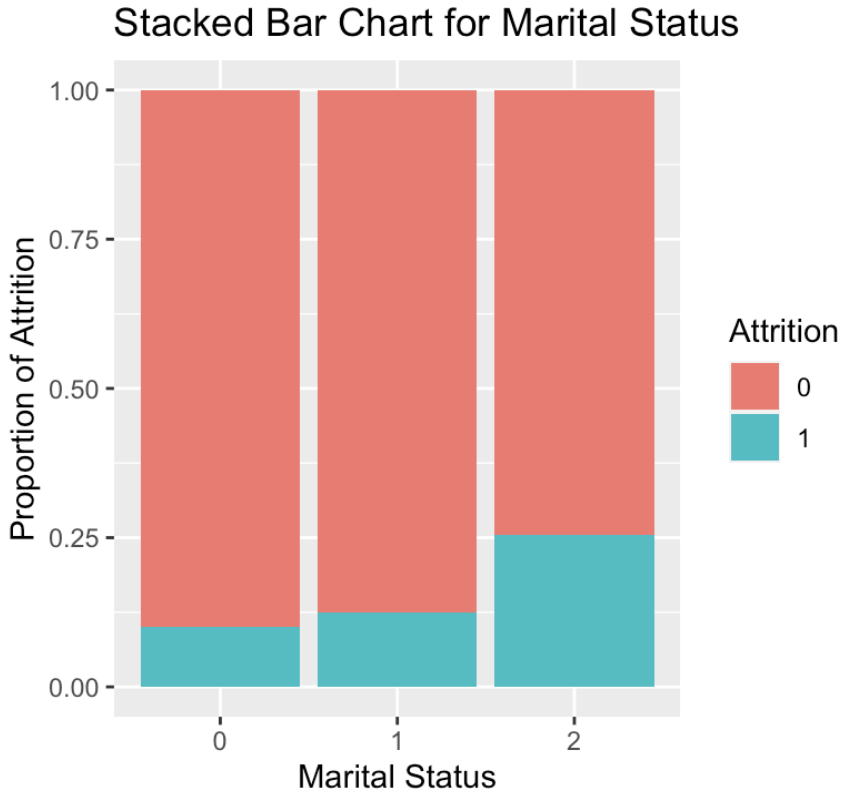


Figure 11 (Attrition vs Marital Status)

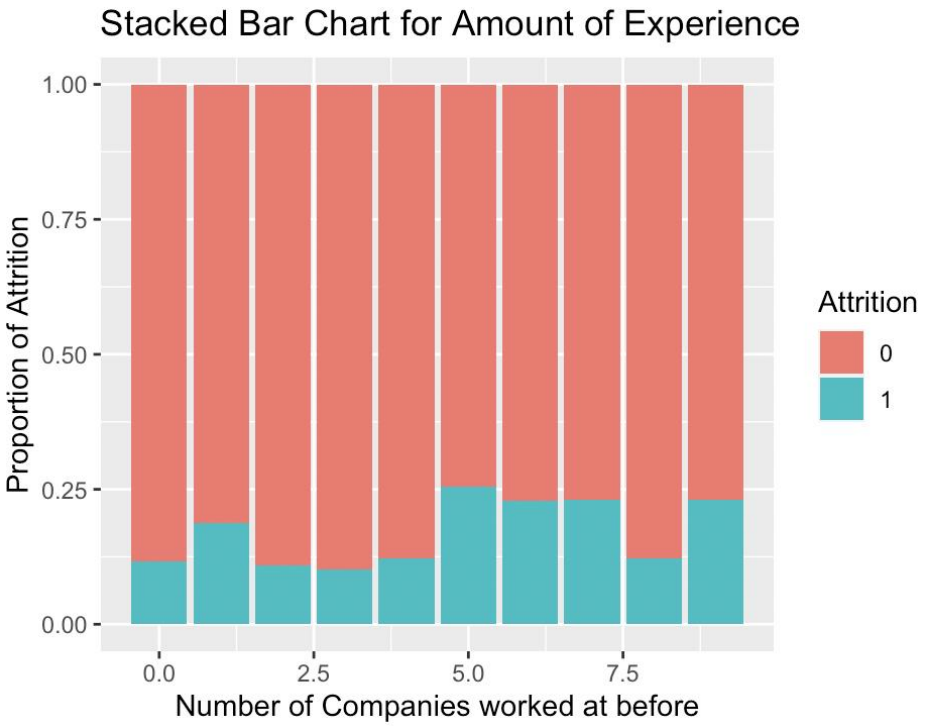


Figure 12 (Attrition vs Amount of Experience)

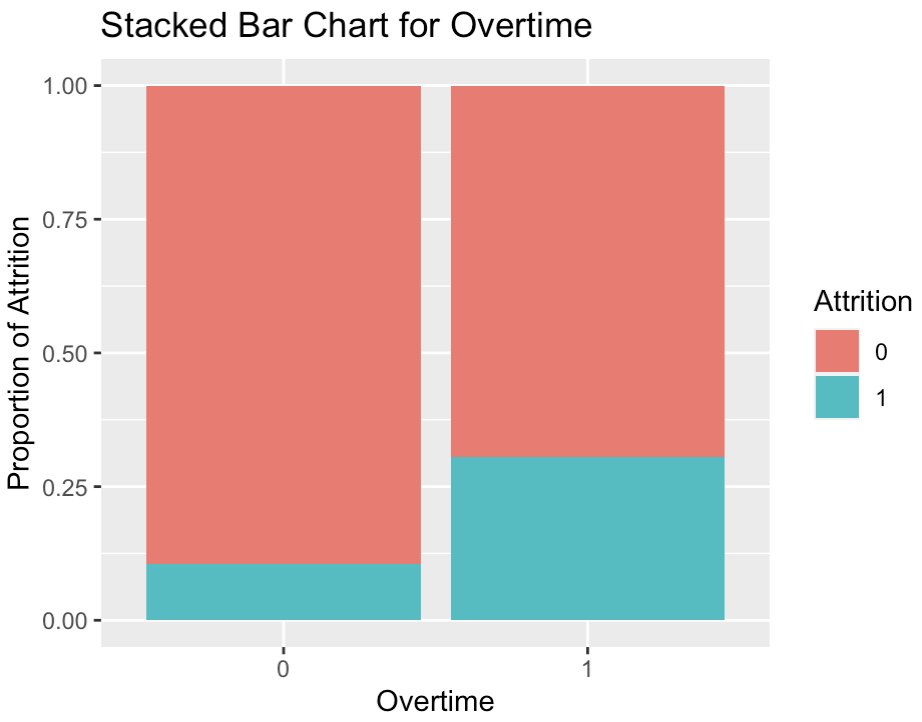


Figure 13 (Attrition vs Overtime)

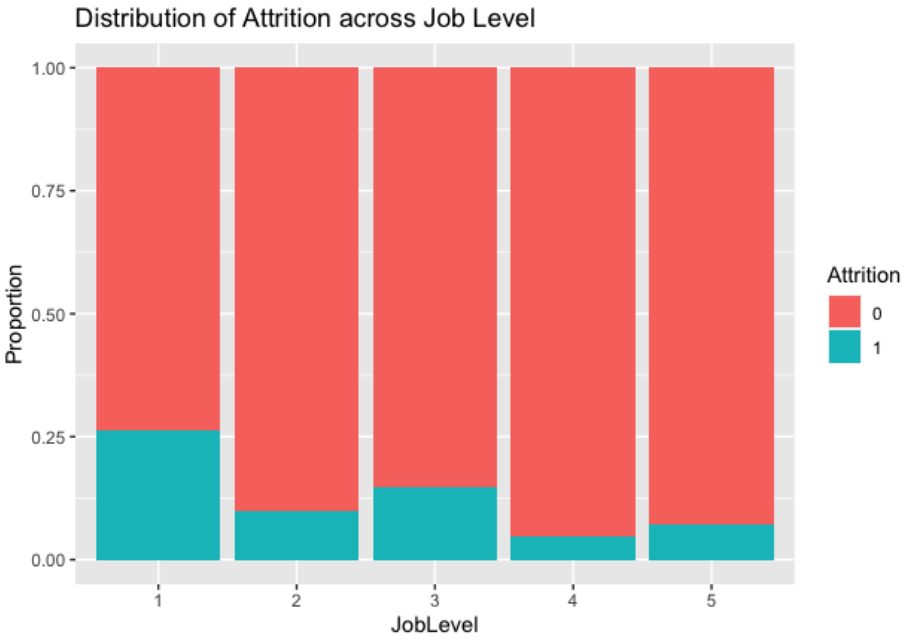


Figure 14 (Attrition vs Job Level)

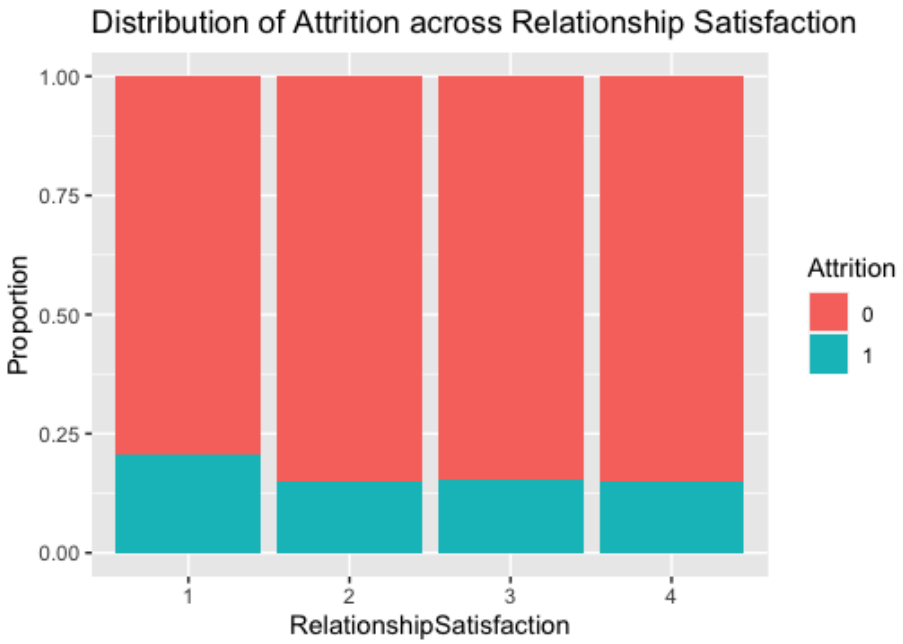


Figure 15 (Attrition vs Relationship Satisfaction)

Figure 16 (Attrition vs Performance Rating)

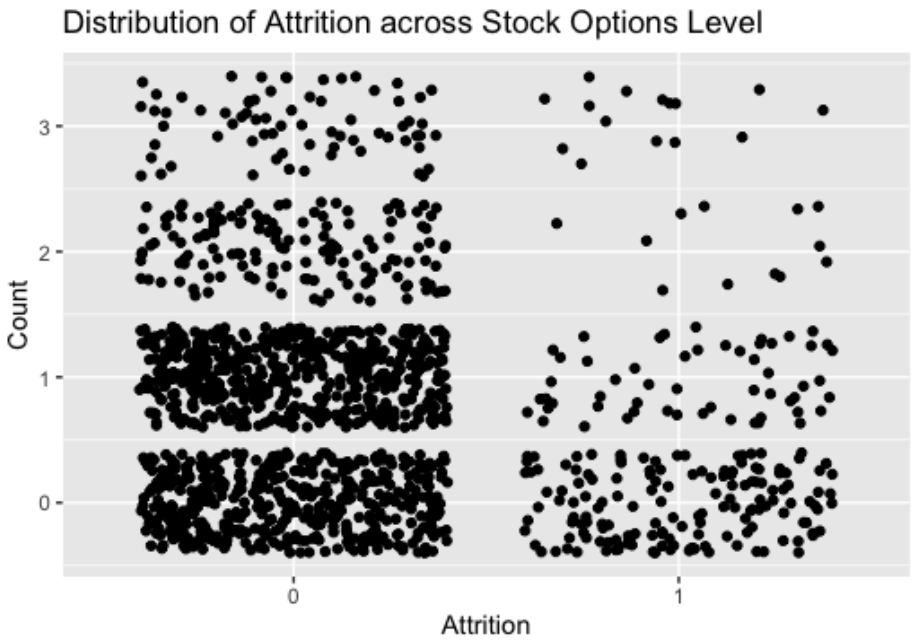


Figure 17 (Attrition vs Stock Options)

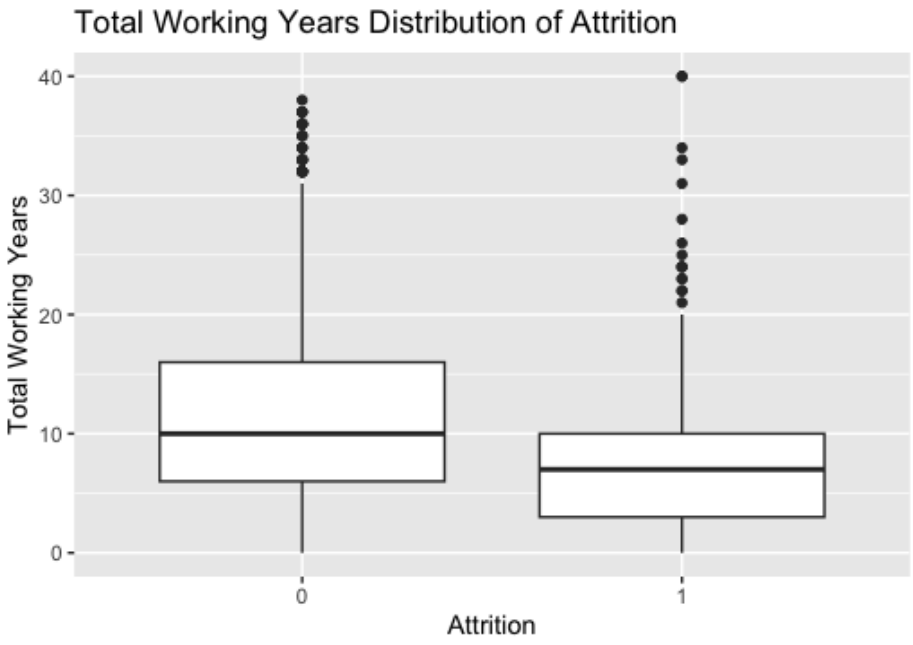


Figure 18 (Attrition vs Total Working Years)

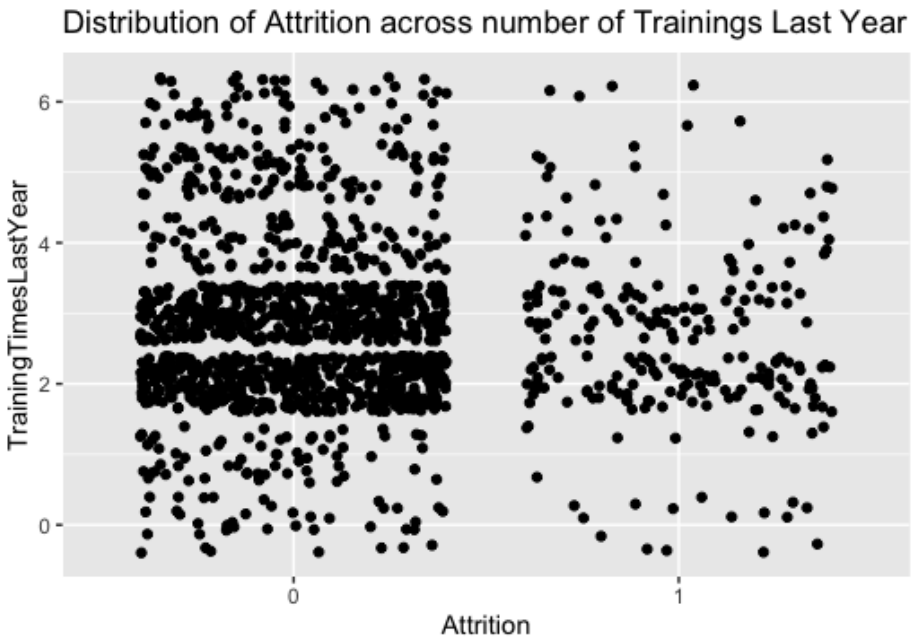


Figure 19 (Attrition vs Number of Trainings Last Year)

Figure 20 (Attrition vs Work Life Balance)

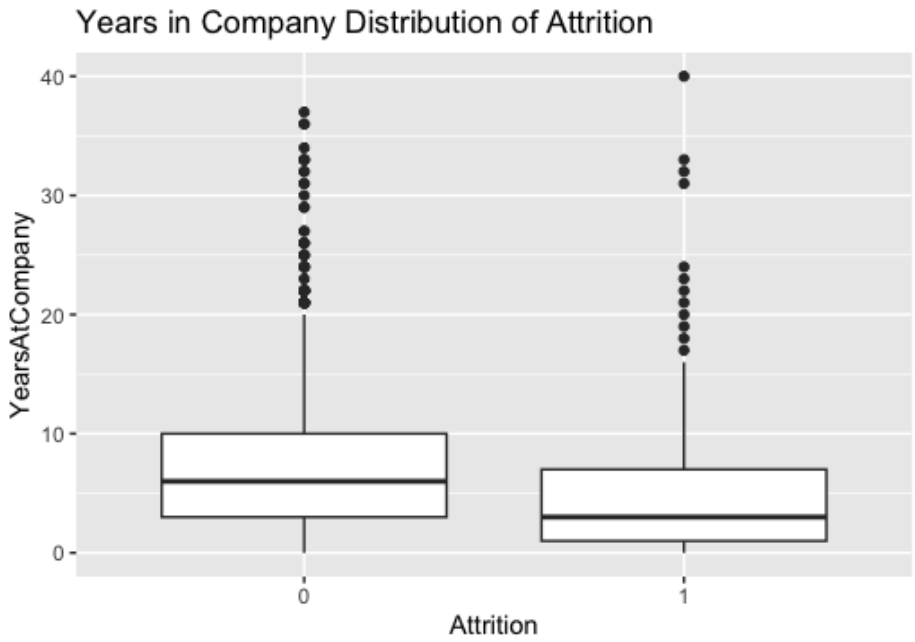


Figure 21 (Attrition vs Years in Company)

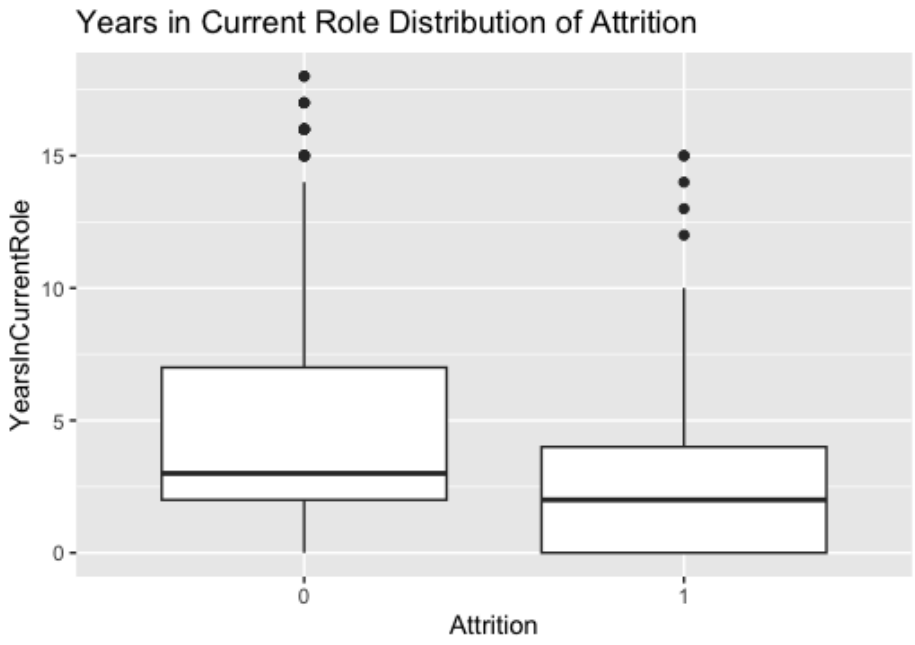


Figure 22 (Attrition vs Years in Current Role)

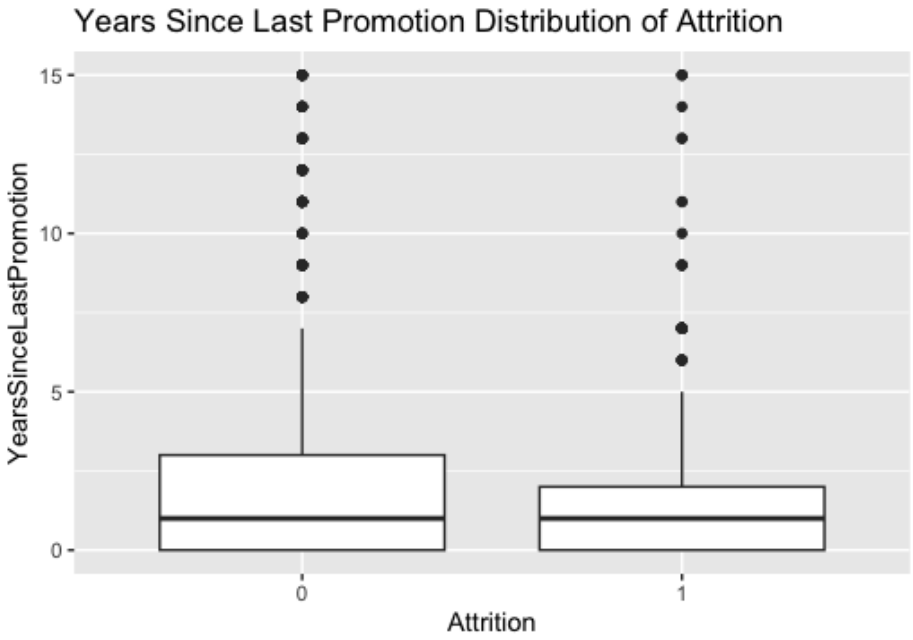


Figure 23 (Attrition vs Years since Last Promotion)

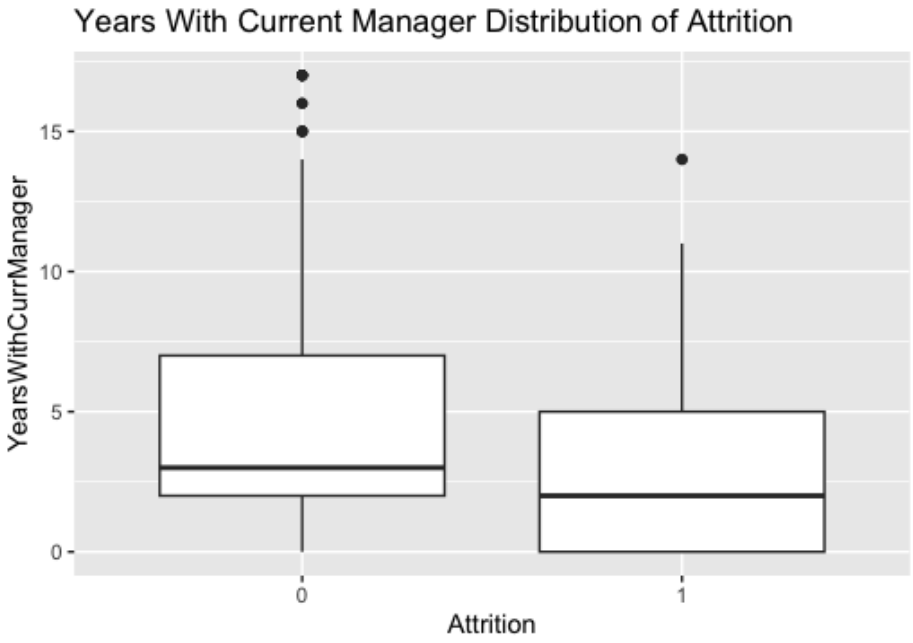


Figure 24 (Attrition vs Years with Current Manager)

9.2. Dataset 2 - Data Visualisation and Exploration

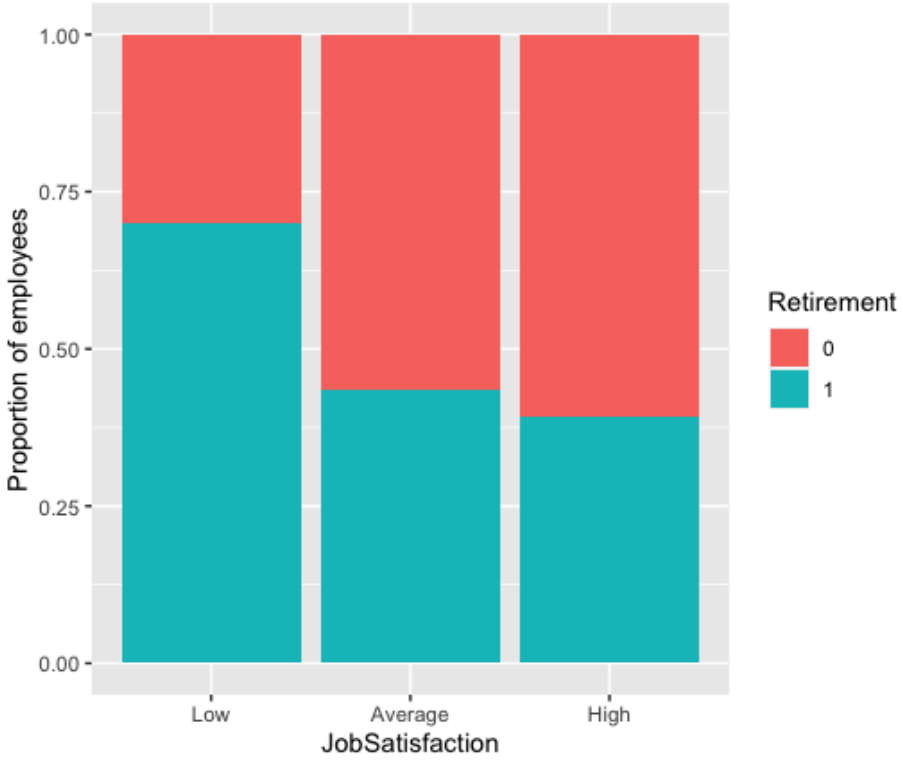


Figure 25 (Job Satisfaction vs Retirement)

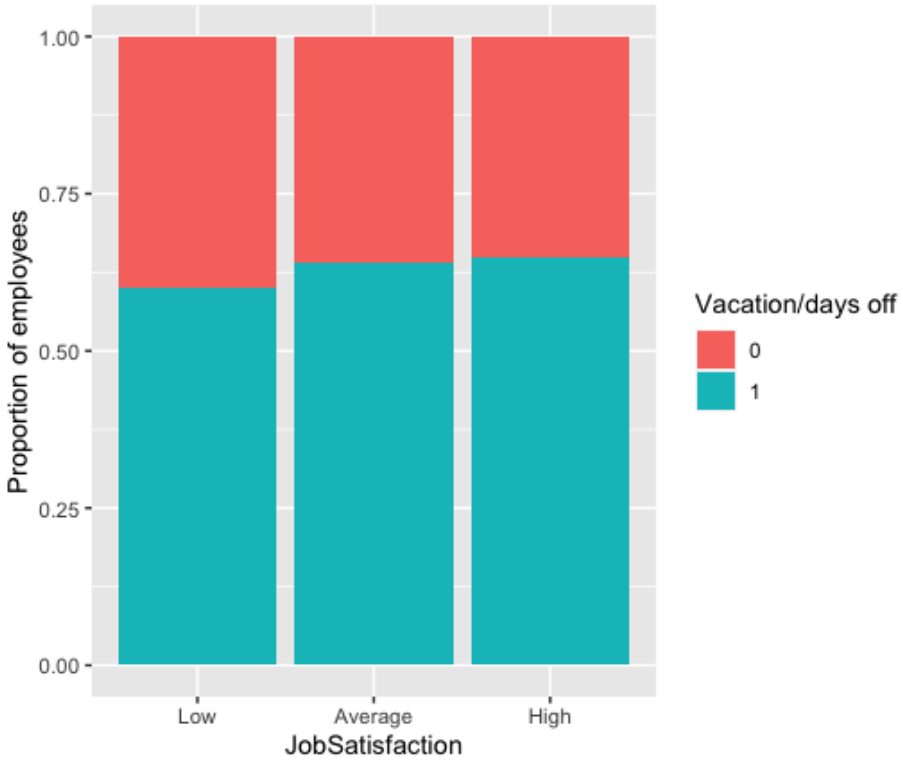


Figure 26 (Job Satisfaction vs Vacations)

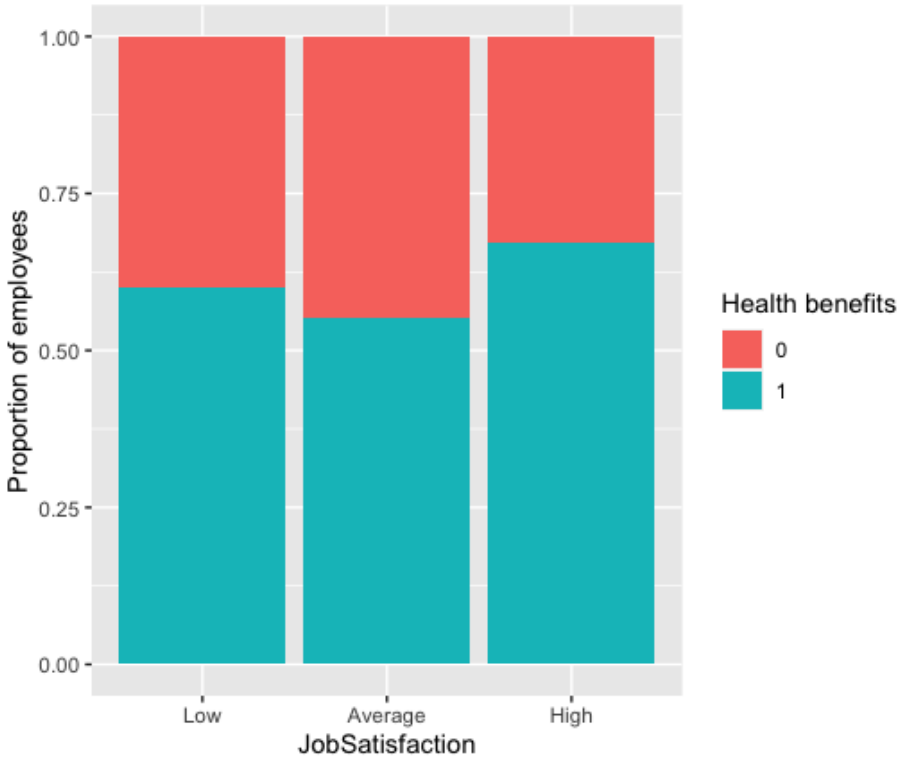


Figure 27 (Job Satisfaction vs Health Benefits)

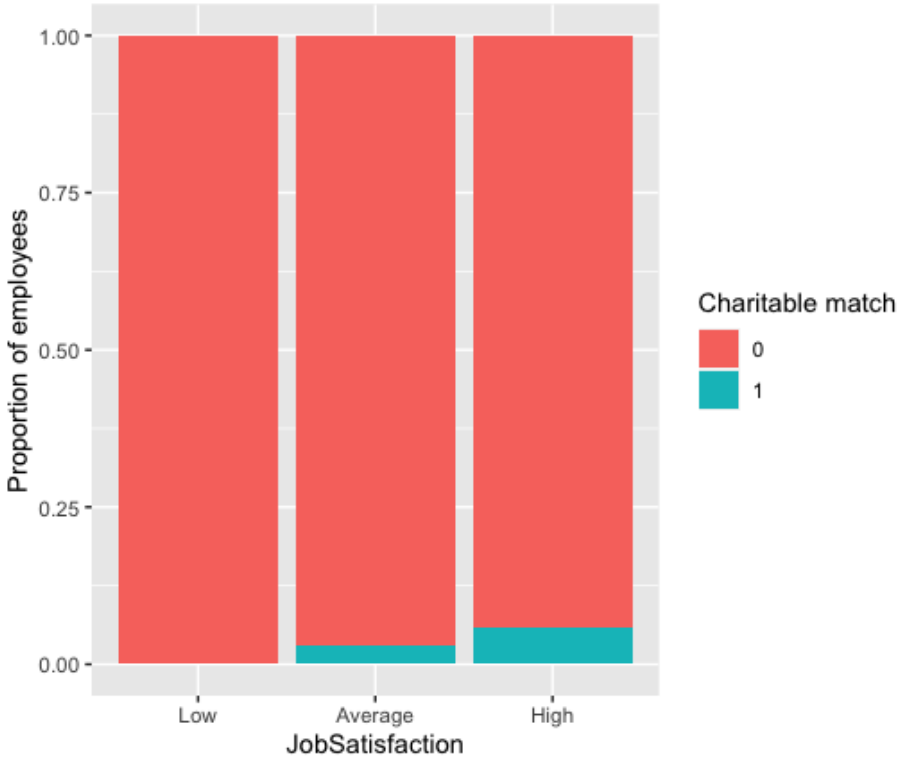


Figure 28 (Job Satisfaction vs Charitable Match)

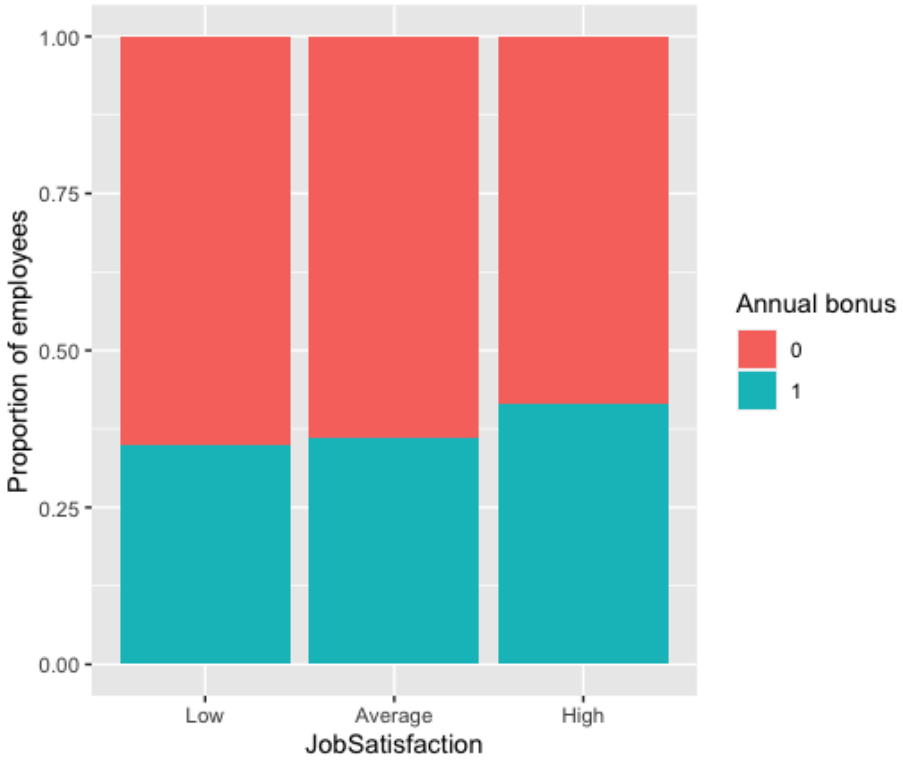


Figure 29 (Job Satisfaction vs Annual Bonus)

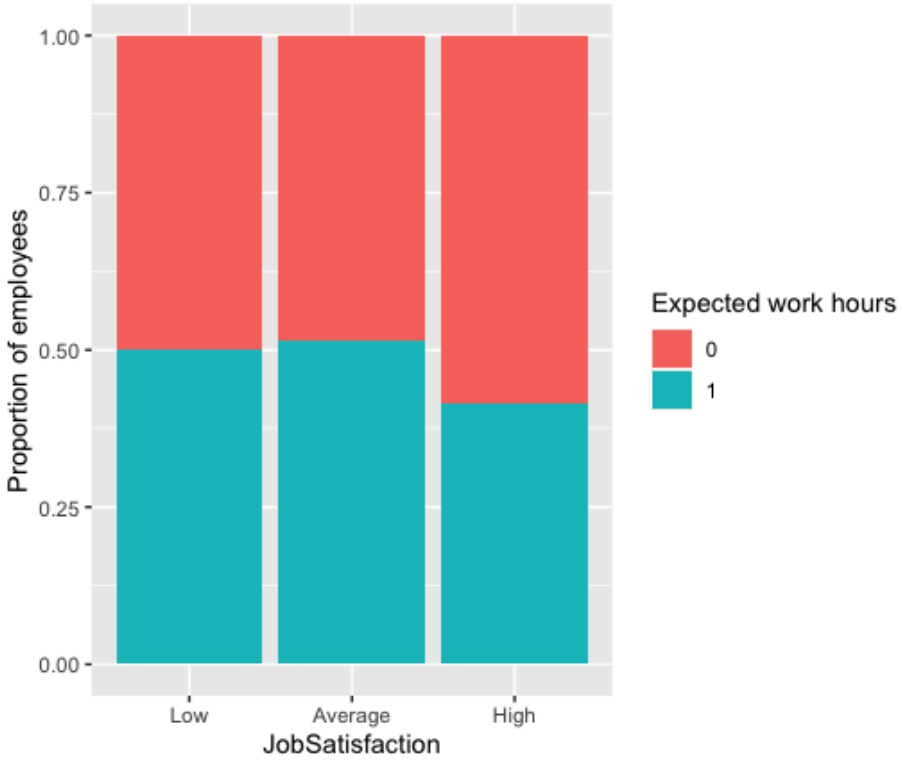


Figure 30 (Job Satisfaction vs Expected Work Hours)

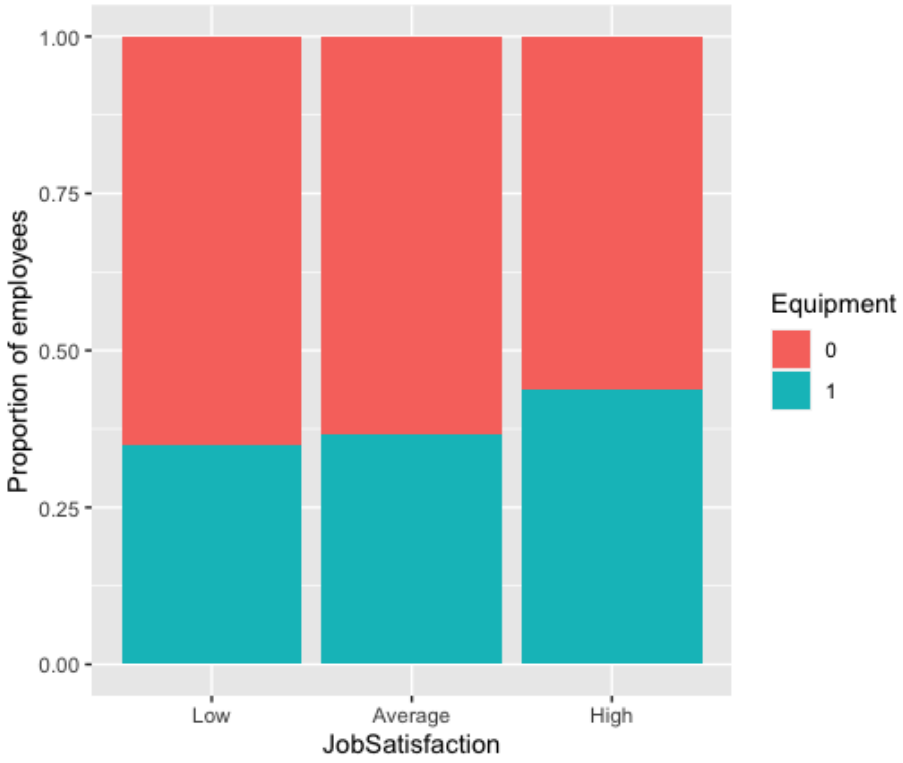


Figure 31 (Job Satisfaction vs Equipment)

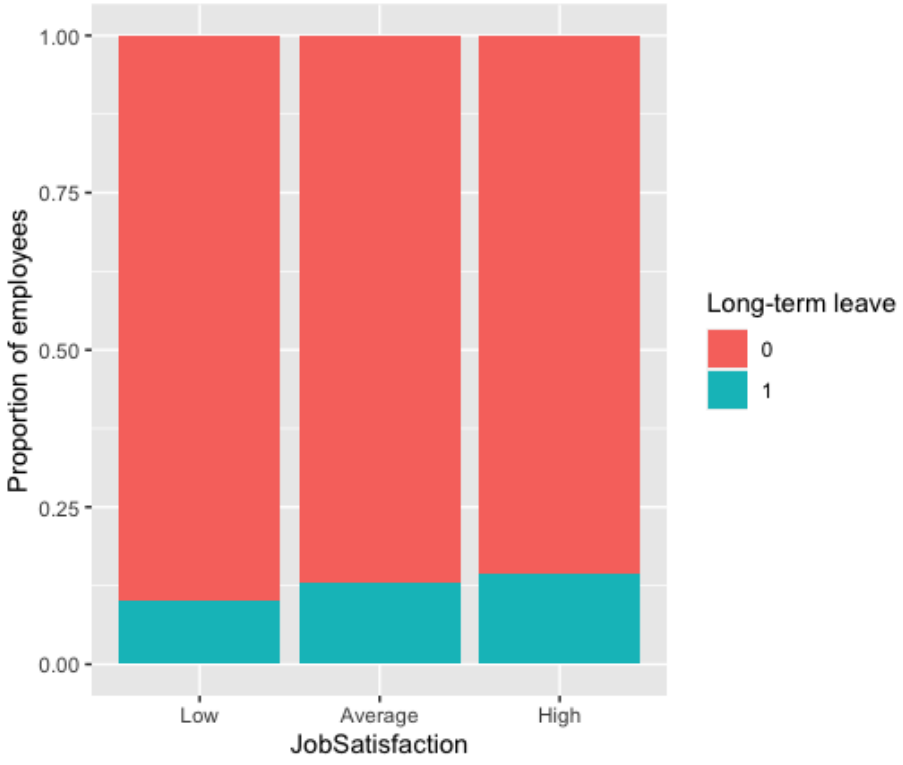


Figure 32 (Job Satisfaction vs Long-Term Leave)

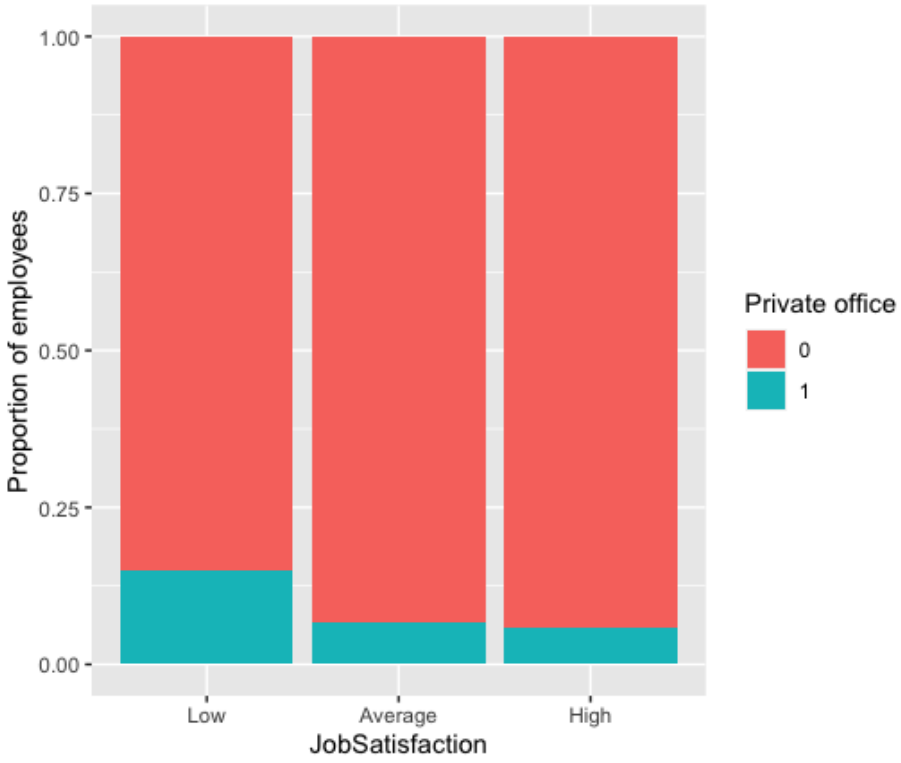


Figure 33 (Job Satisfaction vs Private Office)

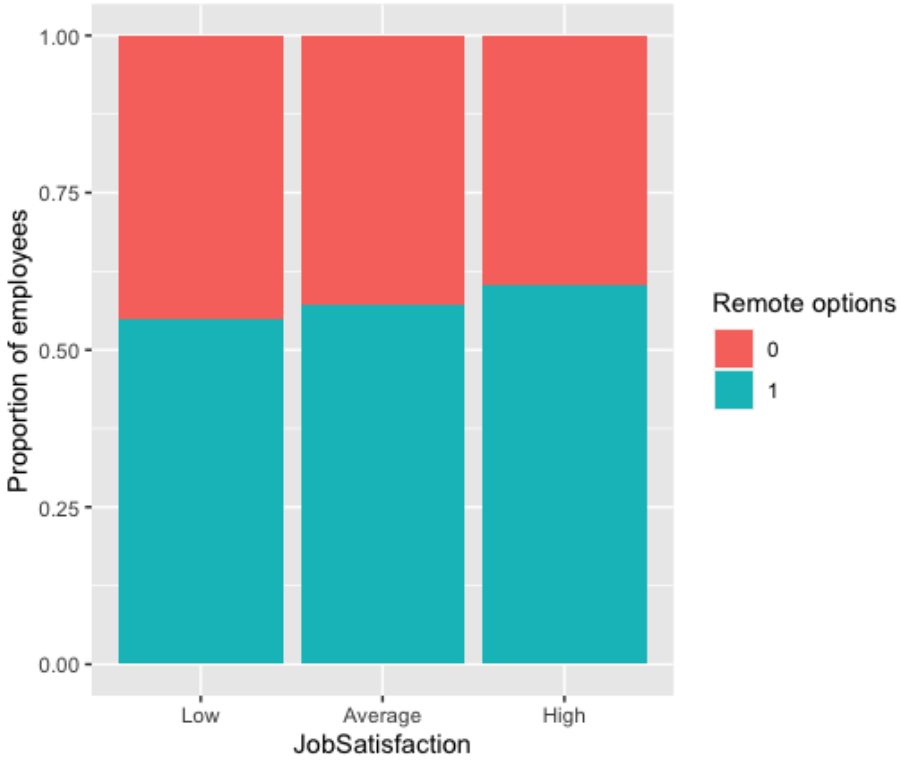


Figure 34 (Job Satisfaction vs Remote Options)

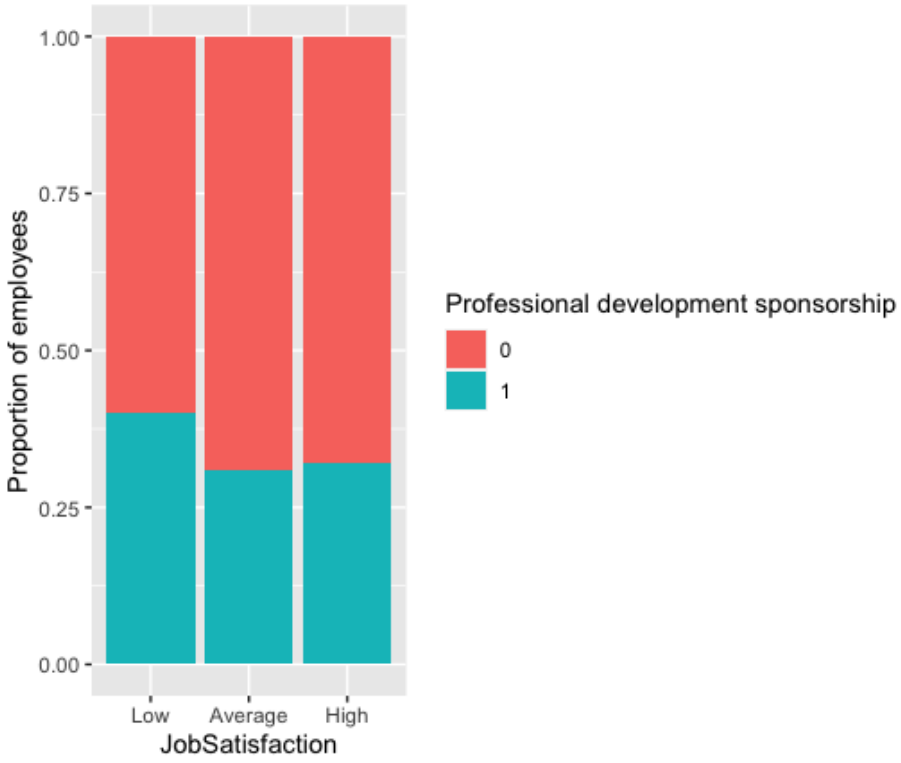


Figure 35 (Job Satisfaction vs Professional Development Sponsorship)

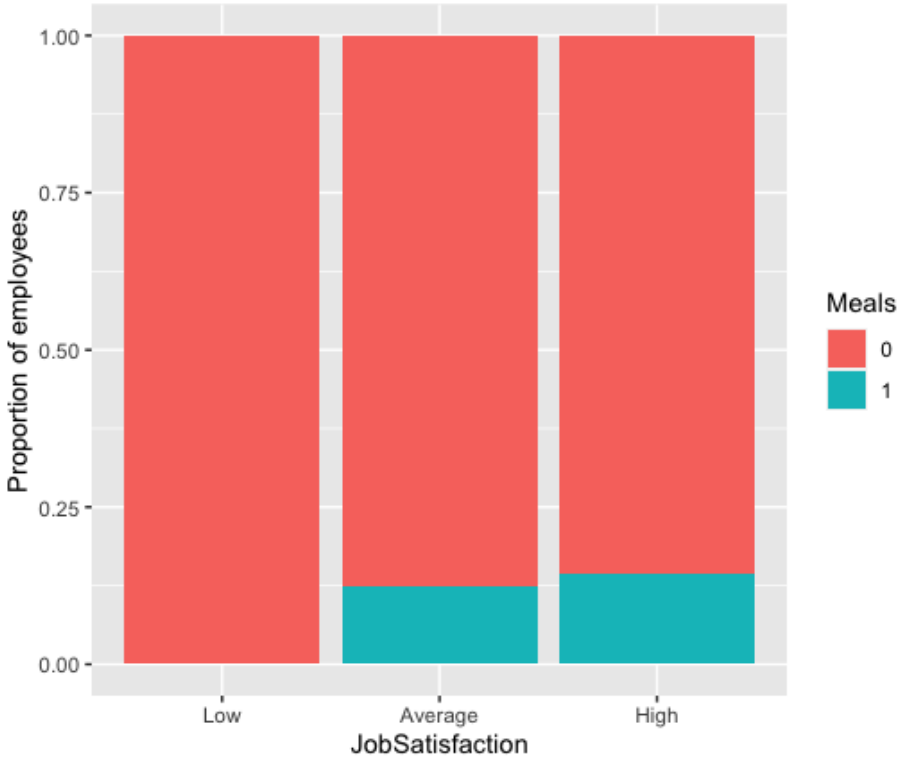


Figure 36 (Job Satisfaction vs Meals)

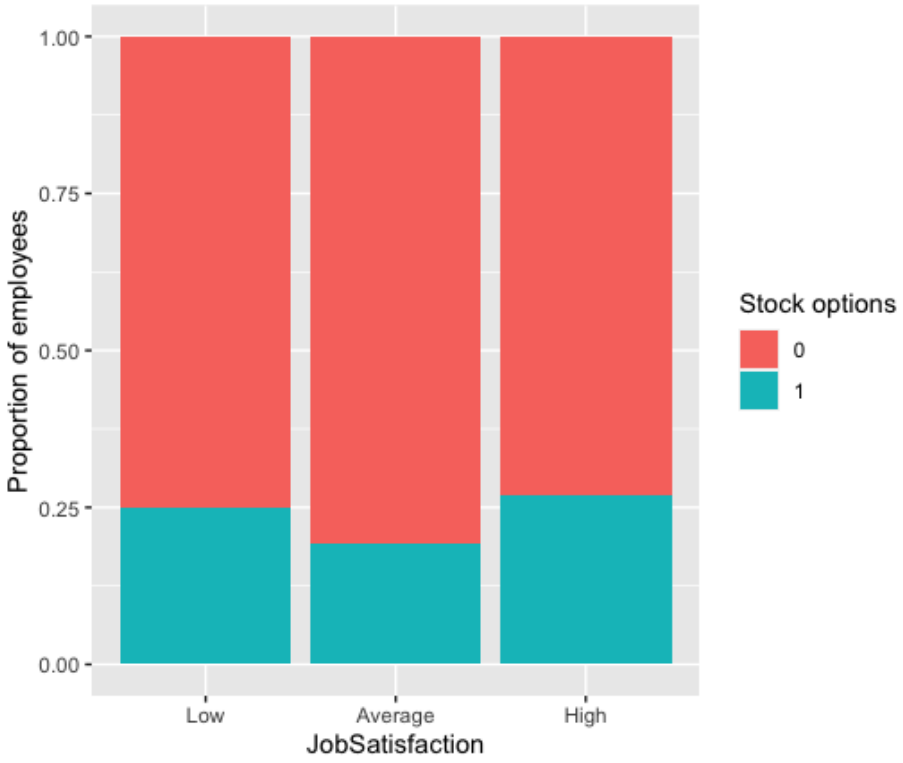


Figure 37 (Job Satisfaction vs Stock Options)

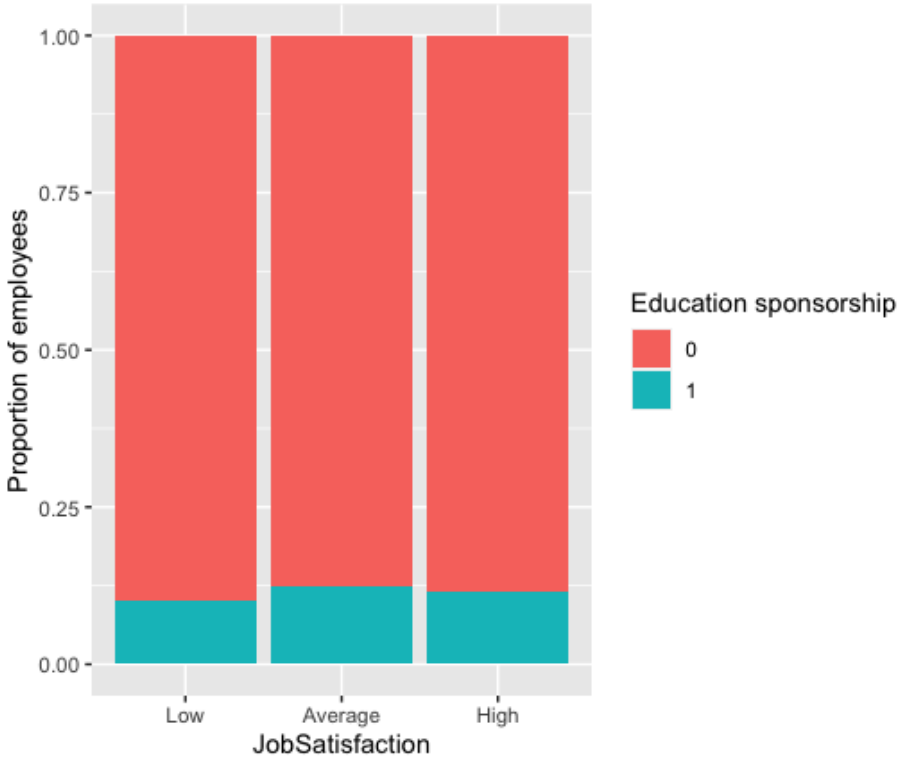


Figure 38 (Job Satisfaction vs Education Sponsorship)

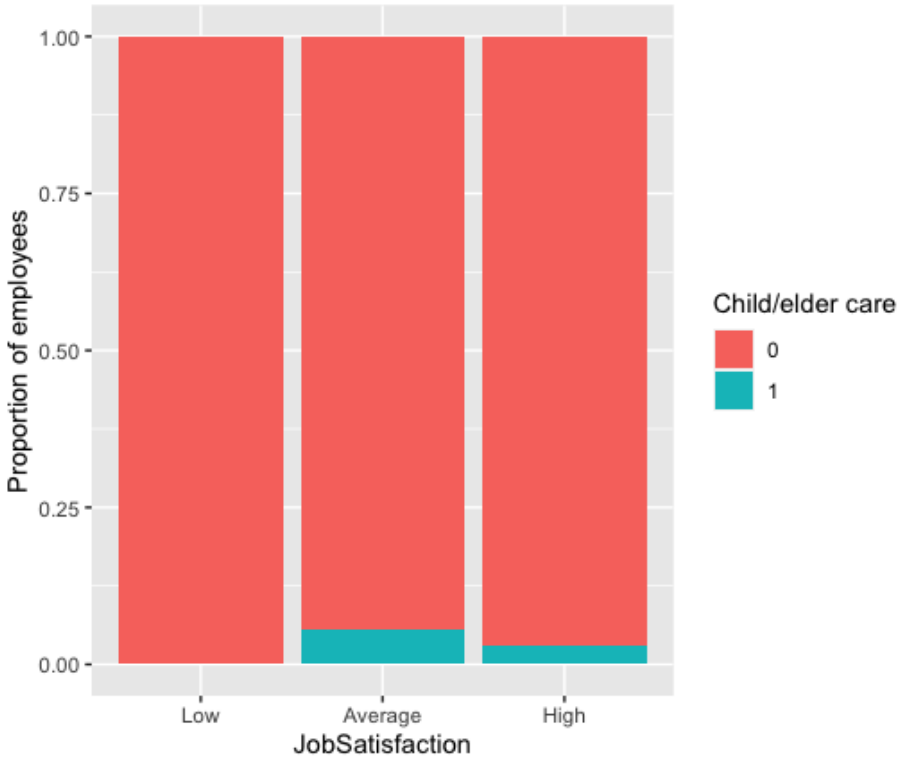


Figure 39 (Job Satisfaction vs Child/Elder Care)

9.3. Logistic Regression Model

```
> vif(lgreg)
                              GVIF Df GVIF^(1/(2*Df))
Age                       1.891649  1        1.375372
BusinessTravel            1.153563  2        1.036359
Education                 1.278016  4        1.031139
EnvironmentSatisfaction   1.187175  3        1.029009
Gender                    1.053471  1        1.026387
JobInvolvement            1.150635  3        1.023661
JobRole                   2.735234  6        1.087467
JobSatisfaction           1.201473  3        1.031064
MaritalStatus             2.116308  2        1.206132
MonthlyIncome             3.426186  1        1.850996
NumCompaniesWorked        1.362511  1        1.167267
OverTime                  1.235829  1        1.111678
PercentSalaryHike         2.550183  1        1.596929
PerformanceRating         2.564143  1        1.601294
RelationshipSatisfaction  1.213494  3        1.032776
StockOptionLevel          1.995478  1        1.412614
TotalWorkingYears         4.678410  1        2.162963
TrainingTimesLastYear     1.042295  1        1.020928
WorkLifeBalance           1.157024  3        1.024606
YearsAtCompany            5.942812  1        2.437788
YearsInCurrentRole        2.597446  1        1.611659
YearsSinceLastPromotion   2.399230  1        1.548945
YearsWithCurrManager      2.862988  1        1.692037
```

Figure 40

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.933e+00  1.025e+00   2.862 0.004208 **
Age                      -3.065e-02  1.356e-02  -2.260 0.023805 *
BusinessTravel1           9.054e-01  3.782e-01   2.394 0.016668 *
BusinessTravel2           1.719e+00  4.067e-01   4.227 2.37e-05 ***
Education2                1.042e-03  3.221e-01   0.003 0.997419
Education3                2.524e-01  2.837e-01   0.890 0.373698
Education4                5.984e-02  3.066e-01   0.195 0.845256
Education5               -2.563e-01  6.319e-01  -0.406 0.685024
EnvironmentSatisfaction2 -1.040e+00  2.726e-01  -3.816 0.000136 ***
EnvironmentSatisfaction3 -1.051e+00  2.442e-01  -4.306 1.67e-05 ***
EnvironmentSatisfaction4 -1.309e+00  2.496e-01  -5.246 1.56e-07 ***
Gender1                   4.469e-01  1.831e-01   2.441 0.014660 *
JobInvolvement2          -1.169e+00  3.517e-01  -3.323 0.000892 ***
JobInvolvement3          -1.472e+00  3.305e-01  -4.455 8.41e-06 ***
JobInvolvement4          -2.055e+00  4.635e-01  -4.433 9.31e-06 ***
JobRole1                 -6.957e-01  2.607e-01  -2.669 0.007606 **
JobRole2                  2.042e-01  3.089e-01   0.661 0.508577
JobRole3                  4.376e-01  3.007e-01   1.455 0.145589
JobRole4                 -8.824e-01  4.245e-01  -2.079 0.037661 *
JobRole5                  2.405e-01  6.781e-01   0.355 0.722825
JobRole6                  7.324e-01  2.797e-01   2.618 0.008833 **
JobSatisfaction2         -6.575e-01  2.664e-01  -2.468 0.013580 *
JobSatisfaction3         -7.068e-01  2.375e-01  -2.975 0.002925 **
JobSatisfaction4         -1.420e+00  2.595e-01  -5.473 4.42e-08 ***
MaritalStatus1            3.295e-01  2.627e-01   1.254 0.209718
MaritalStatus2            1.193e+00  3.321e-01   3.591 0.000329 ***
MonthlyIncome            -1.112e-04  4.436e-05  -2.506 0.012207 *
NumCompaniesWorked        1.733e-01  3.774e-02   4.593 4.37e-06 ***
OverTime1                 1.996e+00  1.928e-01  10.355  < 2e-16 ***
PercentSalaryHike        -1.655e-02  3.801e-02  -0.435 0.663281
PerformanceRating4        3.296e-02  3.883e-01   0.085 0.932360
RelationshipSatisfaction2 -7.896e-01  2.817e-01  -2.803 0.005061 **
RelationshipSatisfaction3 -7.652e-01  2.474e-01  -3.093 0.001979 **
RelationshipSatisfaction4 -9.202e-01  2.546e-01  -3.614 0.000302 ***
StockOptionLevel         -1.326e-01  1.499e-01  -0.885 0.376383
TotalWorkingYears        -5.837e-02  2.869e-02  -2.035 0.041855 *
TrainingTimesLastYear    -1.761e-01  7.143e-02  -2.465 0.013685 *
WorkLifeBalance2         -9.095e-01  3.576e-01  -2.544 0.010973 *
WorkLifeBalance3         -1.313e+00  3.343e-01  -3.926 8.62e-05 ***
WorkLifeBalance4         -9.424e-01  4.077e-01  -2.312 0.020795 *
YearsAtCompany            8.478e-02  3.844e-02   2.206 0.027407 *
YearsInCurrentRole       -1.231e-01  4.518e-02  -2.725 0.006430 **
YearsSinceLastPromotion   1.958e-01  4.216e-02   4.645 3.40e-06 ***
YearsWithCurrManager     -1.609e-01  4.733e-02  -3.399 0.000677 ***
---
```

Figure 41

```
> selected_features1
 [1] "JobInolvement"           "JobRole"                 "BusinessTravel"          "OverTime"
 [5] "WorkLifeBalance"         "RelationshipSatisfaction" "MaritalStatus"          "EnvironmentSatisfaction"
 [9] "StockOptionLevel"        "Gender"                  "Education"               "TrainingTimeLastYear"
[13] "YearsSinceLastPromotion" "YearsAtCompany"          "YearsWithCurrManager"    "JobSatisfaction"
[17] "MonthlyIncome"
```

Figure 42

9.4. Random Forest Model

9.4.1. Original Coinbase Dataset with 26 Variables

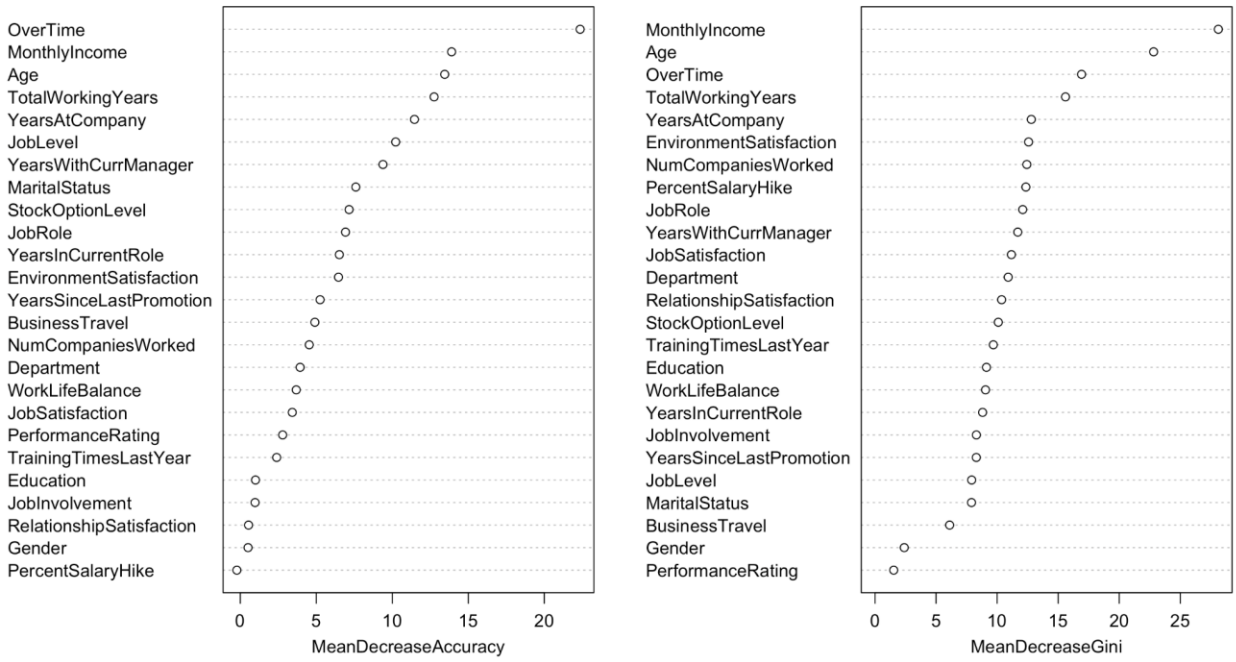


Figure 43

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 366  55
         1   5  15

               Accuracy : 0.8639
                 95% CI : (0.8284, 0.8945)
    No Information Rate : 0.8413
    P-Value [Acc > NIR] : 0.1062

                  Kappa : 0.2827

 Mcnemar's Test P-Value : 2.518e-10

            Sensitivity : 0.21429
            Specificity : 0.98652
         Pos Pred Value : 0.75000
         Neg Pred Value : 0.86936
             Prevalence : 0.15873
         Detection Rate : 0.03401
   Detection Prevalence : 0.04535
      Balanced Accuracy : 0.60040

       'Positive' Class : 1
```

Figure 44 Testing Accuracy (Imbalanced Data)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 250   0
         1   9 259

               Accuracy : 0.9826
                 95% CI : (0.9673, 0.992)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9653

 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 1.0000
            Specificity : 0.9653
         Pos Pred Value : 0.9664
         Neg Pred Value : 1.0000
             Prevalence : 0.5000
         Detection Rate : 0.5000
   Detection Prevalence : 0.5174
      Balanced Accuracy : 0.9826

       'Positive' Class : 1
```

Figure 45 Testing Accuracy (SMOTE Balanced Data)

9.4.2. Minimal Coinbase Dataset with Selected Features

```
> print(selected_features2)
 [1] "OverTime"                "Age"                     "TotalWorkingYears"       "MonthlyIncome"
 [5] "YearsAtCompany"          "JobLevel"                "YearsWithCurrManager"    "YearsInCurrentRole"
 [9] "JobRole"                 "MaritalStatus"           "YearsSinceLastPromotion" "StockOptionLevel"
[13] "EnvironmentSatisfaction" "NumCompaniesWorked"      "BusinessTravel"          "JobSatisfaction"
[17] "Department"              "WorkLifeBalance"         "Education"
```

Figure 46

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 367  53
         1   4  17

               Accuracy : 0.8707
                 95% CI : (0.8358, 0.9006)
    No Information Rate : 0.8413
    P-Value [Acc > NIR] : 0.04884

                  Kappa : 0.3241

 Mcnemar's Test P-Value : 2.047e-10

            Sensitivity : 0.24286
            Specificity : 0.98922
         Pos Pred Value : 0.80952
         Neg Pred Value : 0.87381
             Prevalence : 0.15873
         Detection Rate : 0.03855
   Detection Prevalence : 0.04762
      Balanced Accuracy : 0.61604

       'Positive' Class : 1
```

Figure 47 Testing Accuracy (Imbalanced Data)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 247   0
         1  12 259

               Accuracy : 0.9768
                 95% CI : (0.9599, 0.988)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9537

 Mcnemar's Test P-Value : 0.001496

            Sensitivity : 1.0000
            Specificity : 0.9537
         Pos Pred Value : 0.9557
         Neg Pred Value : 1.0000
             Prevalence : 0.5000
         Detection Rate : 0.5000
   Detection Prevalence : 0.5232
      Balanced Accuracy : 0.9768

       'Positive' Class : 1
```

Figure 48 Testing Accuracy (SMOTE Balanced Data)

9.5. Dataset 1 - Final Dataset Processing

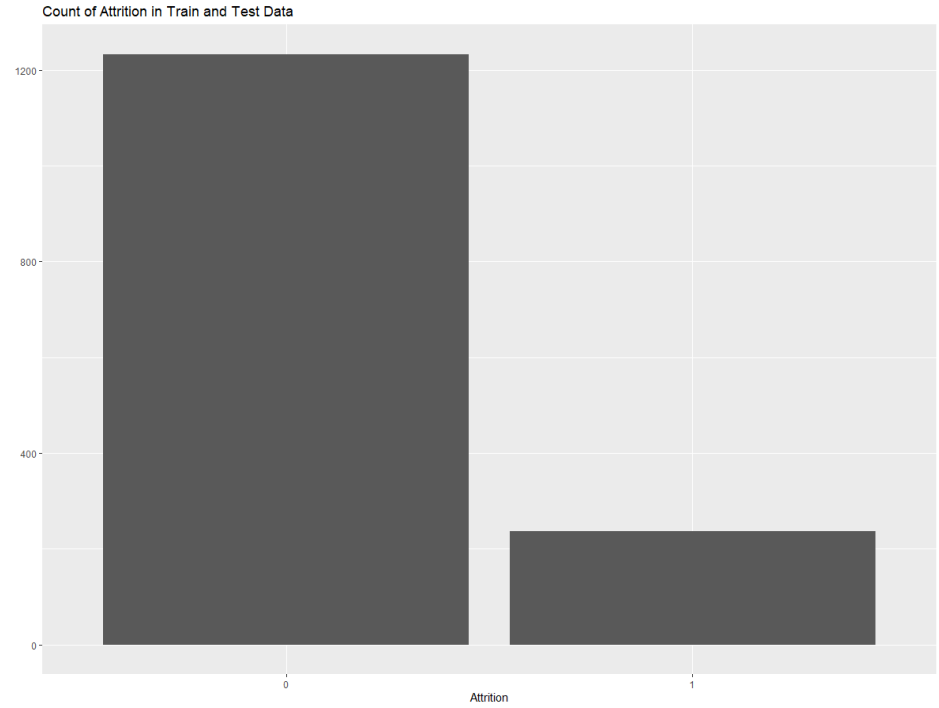


Figure 49 (Number of Instances of 0 and 1 in Attrition)

9.6. CART Model

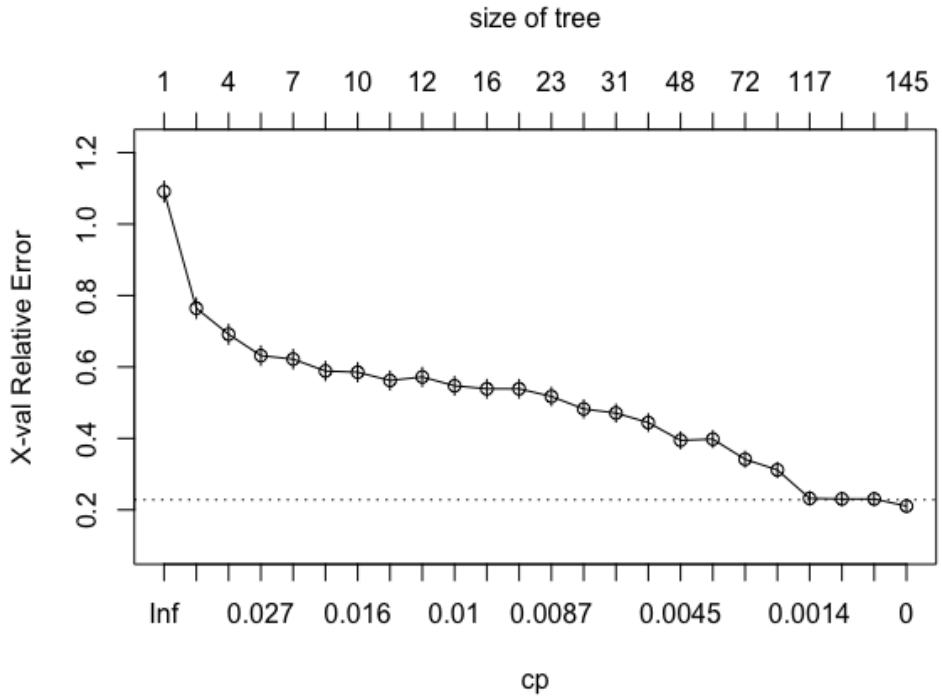


Figure 50

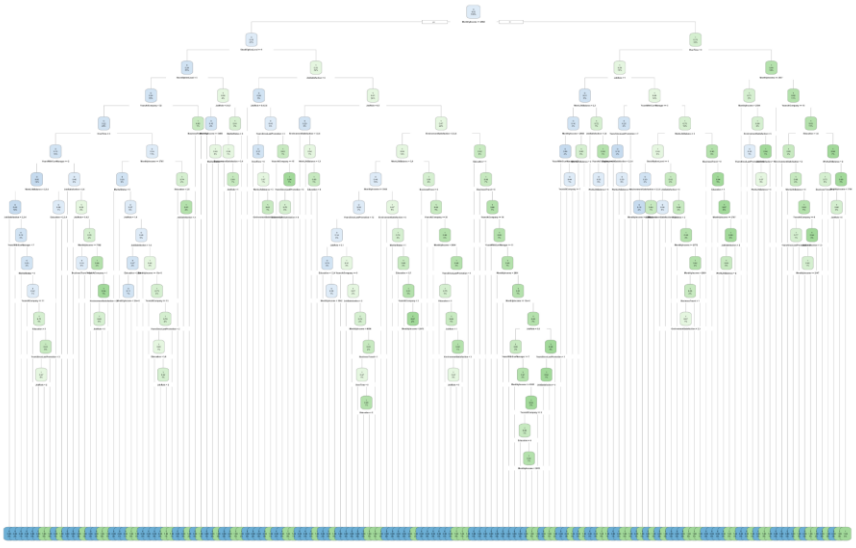


Figure 51

models	accuracy	fp_rate	fn_rate	precision	recall
CART	1	0	0	1	1

Figure 52 (Train)

models	accuracy	fp_rate	fn_rate	precision	recall
CART	0.9227799	0.1544402	0	0.8662207	1

Figure 53 (Test)

9.7. MARS

```
Call: earth(formula=Attrition~., data=hr.trainset, degree=4)

                                                                                                  coefficients
(Intercept)                                                                                       -0.005250993
OverTime1                                                                                          0.164777407
h(18-YearsAtCompany)                                                                               0.017715367
h(3629-MonthlyIncome)                                                                              0.000100565
h(MonthlyIncome-19144)                                                                             0.000766251
JobRole1 * h(18-YearsAtCompany)                                                                   -0.011711398
JobRole2 * h(1-StockOptionLevel)                                                                   0.147379486
JobRole3 * h(MonthlyIncome-3629)                                                                   0.000043595
JobRole4 * h(18-YearsAtCompany)                                                                   -0.020658036
BusinessTravel2 * h(18-YearsAtCompany)                                                             0.011773623
OverTime1 * h(5220-MonthlyIncome)                                                                  0.000063051
EnvironmentSatisfaction4 * h(18-YearsAtCompany)                                                   -0.012083386
h(YearsAtCompany-18) * JobSatisfaction4                                                            0.050110402
h(1-StockOptionLevel) * h(YearsSinceLastPromotion-2)                                               0.038518974
h(1-StockOptionLevel) * h(18-YearsAtCompany)                                                       0.021285161
h(StockOptionLevel-1) * h(18-YearsAtCompany)                                                       0.020691419
h(StockOptionLevel-1) * h(7314-MonthlyIncome)                                                     -0.000051797
h(18-YearsAtCompany) * h(YearsWithCurrManager-2)                                                   0.002962367
h(18-YearsAtCompany) * h(2-YearsWithCurrManager)                                                   0.012422092
h(9-YearsWithCurrManager) * h(MonthlyIncome-3629)                                                 -0.000004757
WorkLifeBalance3 * h(18-YearsAtCompany) * JobSatisfaction4                                        -0.019380535
WorkLifeBalance2 * h(StockOptionLevel-1) * h(18-YearsAtCompany)                                   -0.019879942
WorkLifeBalance3 * h(18-YearsAtCompany) * h(2-YearsWithCurrManager)                               -0.004090886
EnvironmentSatisfaction2 * h(18-YearsAtCompany) * h(2-YearsWithCurrManager)                       -0.011663247
EnvironmentSatisfaction3 * h(18-YearsAtCompany) * h(2-YearsWithCurrManager)                       -0.009765608
h(1-StockOptionLevel) * h(2-YearsSinceLastPromotion) * h(6-YearsAtCompany)                        -0.019143406
h(1-StockOptionLevel) * h(YearsSinceLastPromotion-2) * h(8-YearsAtCompany)                        -0.052255065
h(YearsSinceLastPromotion-3) * h(18-YearsAtCompany) * h(YearsWithCurrManager-2)                    0.001116005
WorkLifeBalance3 * h(YearsSinceLastPromotion-4) * h(18-YearsAtCompany) * h(YearsWithCurrManager-2) -0.002290218
WorkLifeBalance3 * h(4-YearsSinceLastPromotion) * h(18-YearsAtCompany) * h(YearsWithCurrManager-2) -0.001591977

Selected 30 of 43 terms, and 17 of 29 predictors
Termination condition: Reached nk 59
Importance: StockOptionLevel, YearsAtCompany, OverTime1, WorkLifeBalance3, JobSatisfaction4, YearsWithCurrManager,
Number of terms at each degree of interaction: 1 4 15 8 2
GCV 0.1396223    RSS 148.4854    GRSq 0.4424367    RSq 0.507511
```

Figure 54

```
> evimp(mars_model_train)
                          nsubsets   gcv    rss
StockOptionLevel                29 100.0  100.0
YearsAtCompany                  29 100.0  100.0
OverTime1                       28  89.9   90.8
WorkLifeBalance3                27  83.1   84.6
JobSatisfaction4                27  83.1   84.6
YearsWithCurrManager            26  77.2   79.2
MonthlyIncome                   25  72.1   74.5
YearsSinceLastPromotion         24  66.3   69.4
BusinessTravel2                 23  62.4   65.7
JobRole4                        21  54.5   58.5
JobRole1                        19  48.3   52.8
EnvironmentSatisfaction2        16  41.8   46.3
EnvironmentSatisfaction3        15  40.2   44.5
EnvironmentSatisfaction4        14  37.5   42.0
JobRole3                        12  32.9   37.4
WorkLifeBalance2                11  30.2   34.8
JobRole2                         4  14.0   18.1
```

Figure 55

models	accuracy	fp_rate	fn_rate	precision	recall
MARS	0.844942	0.1741294	0.1359867	0.8322684	0.8640133

Figure 56 (Train)

models	accuracy	fp_rate	fn_rate	precision	recall
MARS	0.8243243	0.2084942	0.1428571	0.8043478	0.8571429

Figure 57 (Test)

9.8. Random Forest Final Model

	models	accuracy	fp_rate	fn_rate	precision	recall
Random Forest		1	0	0	1	1

Figure 58 (Train)

	models	accuracy	fp_rate	fn_rate	precision	recall
Random Forest		0.976834	0.04633205	0	0.9557196	1

Figure 59 (Test)

9.9. WebApp

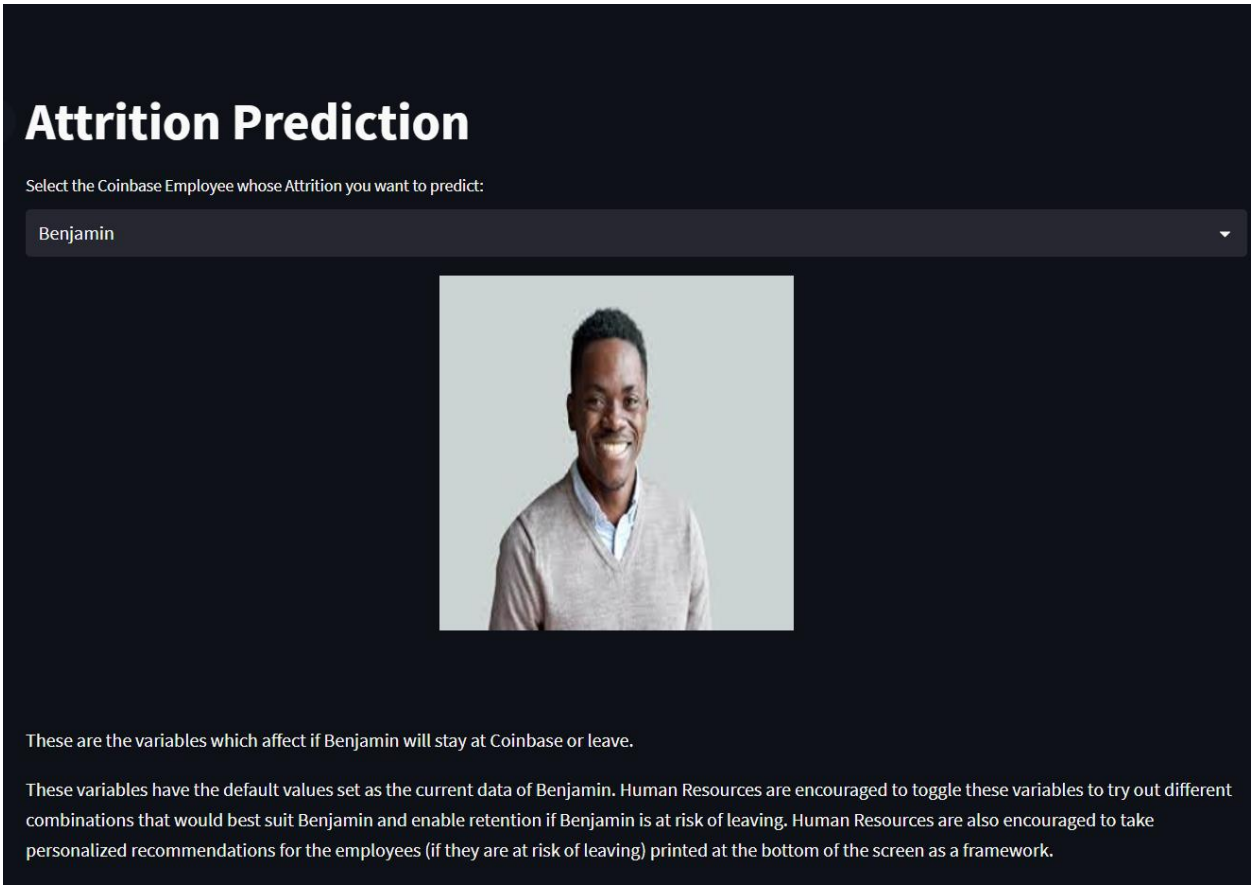


Figure 60

Figure 61

Figure 62

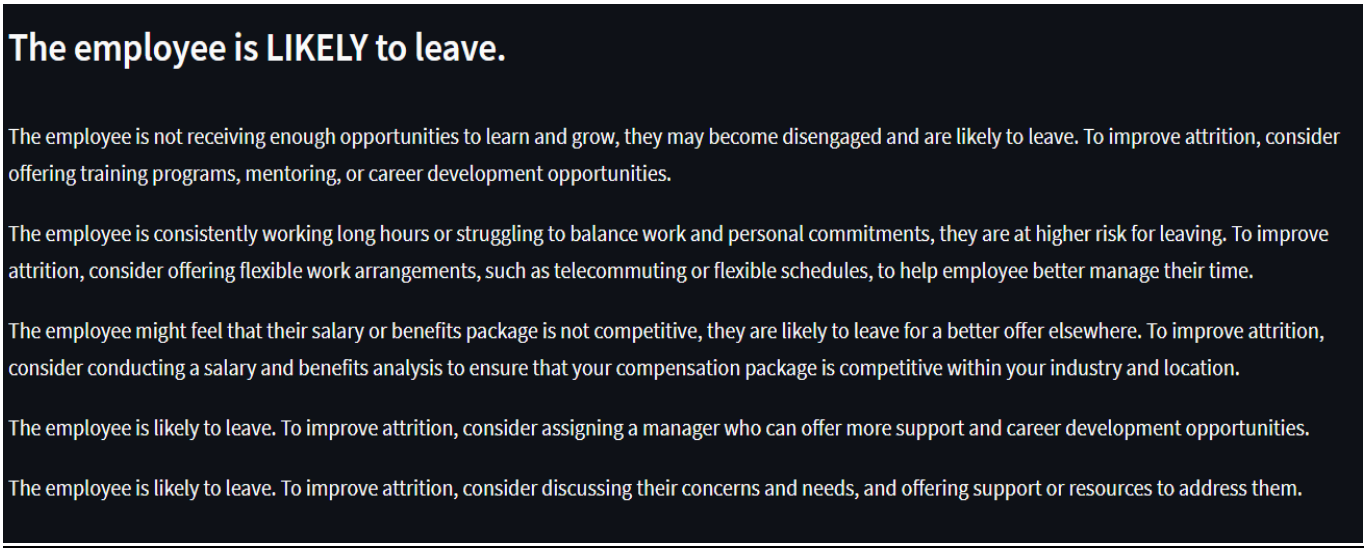


Figure 63

9.10. Dashboard

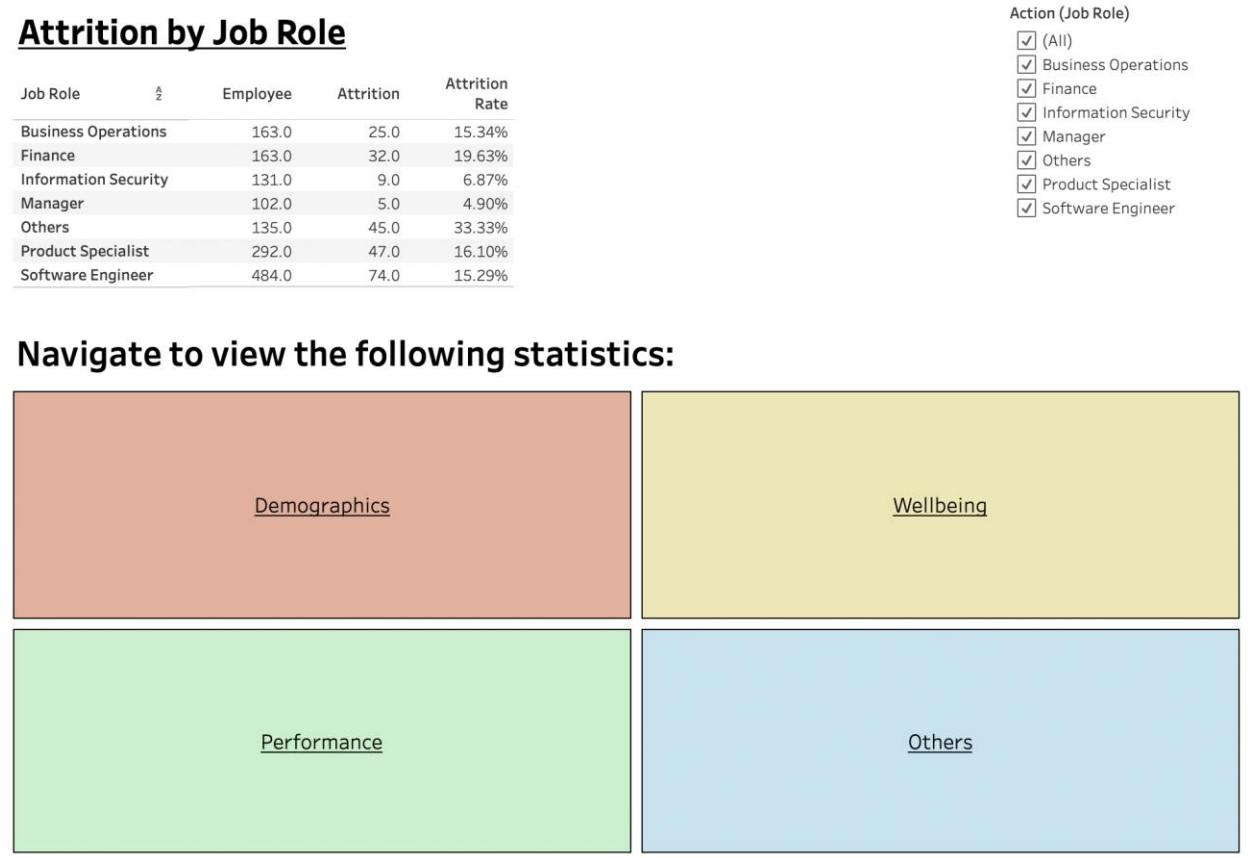


Figure 64

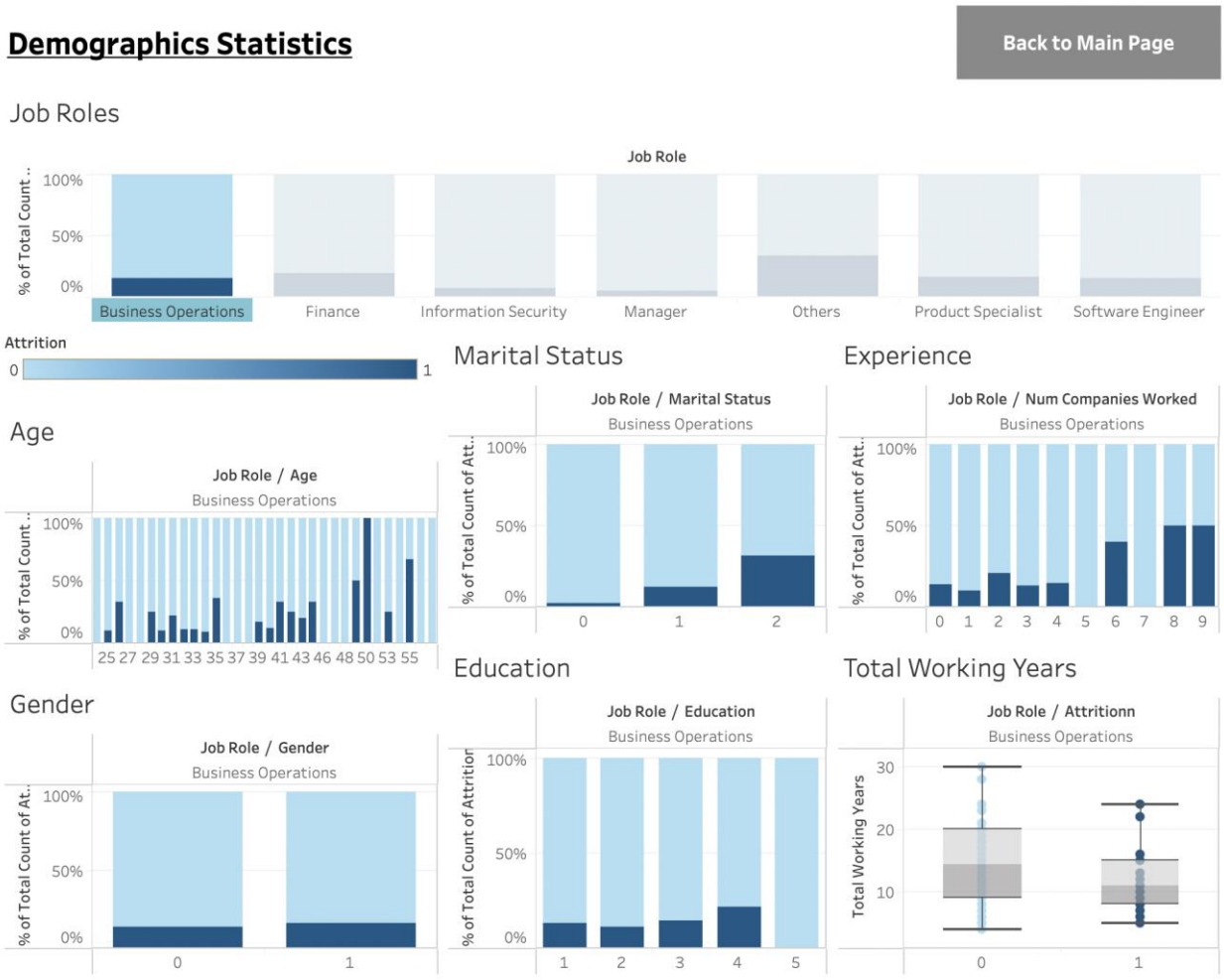

Demographics Statistics
Back to Main Page
Job Roles
Job Role
% of Total Count ..
100%
50%
0%
Business Operations
Finance
Information Security
Manager
Others
Product Specialist
Software Engineer
Attrition
0
1
Marital Status
Job Role / Marital Status
Business Operations
% of Total Count of Att..
Experience
Job Role / Num Companies Worked
Business Operations
Age
Job Role / Age
Business Operations
25 27 29 31 33 35 37 39 41 43 46 48 50 53 55
Education
Job Role / Education
Business Operations
% of Total Count of Attrition
Total Working Years
Job Role / Attritionn
Business Operations
Gender
Job Role / Gender
Business Operations
% of Total Count of At..

Figure 65


Wellbeing Statistics
Back to Main Page
Job Roles
Job Role
% of Total Co..
100%
50%
0%
Business..
Finance
Informat..
Manager
Others
Product ..
Softwar..
Attrition
0
1
Work Life Balance
Job Role / Work Life Balance
Finance
% of Total Count of At..
100%
50%
0%
1
2
3
4
Job Involvement
Job Role / Job Involvement
Finance
% of Total Count of A..
100%
50%
0%
1
2
3
4
Environment Satisfaction
Job Role / Environment Satisfaction
Finance
% of Total Count of..
100%
50%
0%
1
2
3
4
Job Satisfaction
Job Role / Job Satisfaction
Finance
% of Total Count..
100%
50%
0%
1
2
3
4
Relationship Satisfaction
Job Role / Relationship Satisfaction
Finance
% of Total Count of..
100%
50%
0%
1
2
3
4

Figure 66


Performance Statistics
Back to Main Page
Job Roles
Job Role
% of Total Count of At..
100%
50%
0%
Bu s..
Fi..
In f..
M..
O..
Pr o..
So f..
Attrition
0
1
Performance Rating
Job Role / Performance Rating
Others
% of Total Count of A..
100%
50%
0%
3
4
Monthly Income
Job Role / Attritionn
Others
Monthly Income
10K
5K
0.00
1.00
Overtime
Job Role / Over Time
Others
% of Total Count of A..
100%
50%
0%
0
1
Percent Salary Hike
Job Role / Attritionn
Others
Percent Salary H..
25
20
15
10
0
1
Business Travel
Job Role / Business Travel
Others
% of Total Count of A..
100%
50%
0%
0
1
2

Figure 67


Additional Statistics
Back to Main Page
Job Roles
Job Role
% of Total Count of Attr..
100%
50%
0%
Bu s..
Fi..
In f..
M..
O..
Pr o..
So f..
Attrition
0
1
Years Since Last Promotion
Job Role / Attritionn
Information Security
Years Since Last Prom..
15
10
5
0
0
1
Years with Current Manager
Job Role / Attritionn
Information Security
Years With Curr M..
15
10
5
0
0
1
Years in Current Role
Job Role / Attritionn
Information Security
Years In Current..
15
10
5
0
0
1
No. of Training Times Last Year
Job Role / Training Times Last Year
Information Security
% of Total Count of At..
100%
50%
0%
0
1
2
3
4
5
6
Years at Company
Job Role / Attritionn
Information Security
Years At Compa..
40
20
0
0
1

Figure 68