

**BC2407 Analytics II: Advanced Predictive Techniques**

**Project Written Report**

**S06 Team 2**

**Navigating the War for Talent: Predictive HR Analytics for**

**Employee Retention in the Automotive Industry**

|  |  |
| --- | --- |
| **Group Members** | **Matriculation Number** |
| Ang Wei Hao | U2210361E |
| Aung Kaung Kaung | U2221195J |
| Cindrea D’cruz | U2210122K |
| Liu Yu-Hsing | U2221170C |
| Sarkar Mufaddal Adnan | U2210740F |
| Yong Shao En Ernest | U2221153B |

**Contents**

[**Executive Summary** 4](#_Toc162816760)

[**1. Problem Statement** 5](#_Toc162816761)

[**1.1. Introduction** 5](#_Toc162816762)

[**1.2. Practical Motivation** 5](#_Toc162816763)

[**2. Project Objectives** 6](#_Toc162816764)

[**2.1. Business Targets and Outcomes** 6](#_Toc162816765)

[**2.2. Personal and Organisational Factors** 6](#_Toc162816766)

[**3. Methodology** 6](#_Toc162816767)

[**3.1. Data Cleaning** 6](#_Toc162816768)

[**3.2. Exploratory Data Analysis** 7](#_Toc162816769)

[**3.2.1. Univariate Analysis** 7](#_Toc162816770)

[**3.2.2. Bivariate Analysis** 8](#_Toc162816771)

[**3.2.3. Multivariate Analysis** 9](#_Toc162816772)

[**3.3. Feature Transformation** 10](#_Toc162816773)

[**4. Machine Learning and Optimisation** 11](#_Toc162816774)

[**4.1. Logistic Regression** 12](#_Toc162816775)

[**4.1.1. Implementation** 12](#_Toc162816776)

[**4.1.2. Analysis of Logistic Regression** 12](#_Toc162816777)

[**4.1.3. Optimisation** 12](#_Toc162816778)

[**4.1.4 Backward Elimination** 12](#_Toc162816779)

[**4.2. Neural Network** 13](#_Toc162816780)

[**4.2.1. Implementation** 13](#_Toc162816781)

[**4.2.2. Analysis of Neural Network** 13](#_Toc162816782)

[**4.2.3. Optimisation** 13](#_Toc162816783)

[**4.3. CART - Classification Tree** 14](#_Toc162816784)

[**4.3.1. Implementation** 14](#_Toc162816785)

[**4.3.2. Growing the Maximal Tree** 14](#_Toc162816786)

[**4.3.3. Pruning the Tree** 14](#_Toc162816787)

[**4.3.4. Variable Importance** 14](#_Toc162816788)

[**4.4. Random Forest** 15](#_Toc162816789)

[**4.4.1. Implementation** 15](#_Toc162816790)

[**4.4.2. OOB Error Stabilisation** 15](#_Toc162816791)

[**4.4.3. Optimisation of Standard Random Forest** 15](#_Toc162816792)

[**4.4.4. Variable Importance** 16](#_Toc162816793)

[**4.5. XGBoost** 16](#_Toc162816794)

[**4.5.1. Implementation** 16](#_Toc162816795)

[**4.5.2. Optimisation** 16](#_Toc162816796)

[**4.5.3. Variable Importance** 17](#_Toc162816797)

[**4.6. Model Comparison and Evaluation** 17](#_Toc162816798)

[**5. Proposed Business Recommendations** 18](#_Toc162816799)

[**5.1. Business Insights** 18](#_Toc162816800)

[**5.2. Business Solutions** 19](#_Toc162816801)

[**5.2.1. Reducing Attrition Rates** 19](#_Toc162816802)

[**5.2.2. Performance and Productivity - Scorecards** 21](#_Toc162816803)

[**6. Conclusion** 24](#_Toc162816804)

[**6.1. Limitations and Future Work** 24](#_Toc162816805)

[**6.2 Closing** 24](#_Toc162816806)

[**7. References** 25](#_Toc162816807)

[**8. Appendices** 27](#_Toc162816808)

# **Executive Summary**

**Employee attrition** is a growing cause of concern for companies worldwide. An increasing number of employees are voluntarily choosing to leave their jobs in recent years amid “The Great Resignation”. Within the automotive industry, **Tesla Inc.** has been impacted by an employee attrition rate of 70% , which is higher than the industry average of 46%. While most tech companies have adopted machine learning techniques to identify key factors driving attrition and implement targeted retention strategies, Tesla has yet to do the same. As such, there is a pressing need for Tesla to embrace an analytics-driven approach to determine both **personal and organisational factors** driving employees' decisions to leave the company. Therefore, our project aims to align Tesla’s HR practices with broader business goals by leveraging predictive analytics to achieve 3 key business objectives: **reducing attrition rate, achieving cost savings, and increasing employee productivity and performance**.

Our methodology involves conducting a preliminary exploratory data analysis (EDA), followed by the development of 5 predictive models, namely Logistic Regression, Neural Network, Classification and Regression Tree (CART), Random Forest, and XGBoost. In particular, since the cost of failing to predict an employee’s departure when they actually leave (false negatives) is higher than the cost of incorrectly predicting that an employee will leave when they actually remain (false positives), priority was given to **maximising recall** during the optimisation of the model to minimise the risk of Tesla overlooking critical signals of attrition.

By plotting the precision-recall curve of all models, we deduce **Neural Network** to be the most optimal model as it achieves the highest cost savings of $22,316 for every employee that leaves the company, nearly 70% of the maximum cost savings achievable by a perfect classifier. With this, Tesla stands to save **$516 million per year in attrition costs** with the application of our analytics-driven strategy. Moreover, from the variable importance charts of our Random Forest and XGBoost models, we derived the most important variables influencing attrition to be **Overtime, Marital Status, and Stock Option Level**.

The insights from the models were used to inform our business solutions. Due to the demanding nature of Tesla’s operations, it can be difficult to eliminate overtime. As such, we propose a **3-shift system** to alleviate responsibilities and improve employee well-being, as well as **cross-training** employees to enhance operational flexibility. Furthermore, the implementation of a real-time **HR monitoring platform** will enable the proactive identification of employees with high attrition risk. These solutions will be complemented with **scorecards** to assess the employees’ cultural fit score and organisational alignment score. The tabulated scores will then be used to plot the employees’ standing on the **strategic heat map**, which has 4 quadrants (Focus, Maintain, Monitor, and Promote) that each come with tailored strategies to encourage employee retention.

In summary, our project underscores the potential of predictive analytics in transforming HR practices, providing Tesla with a strategic blueprint to effectively reduce attrition risk.

# **1. Problem Statement**

## **1.1. Introduction**

Employee attrition is the loss of employees in a company due to resignations initiated by the employee, resulting in employees leaving faster than new ones being hired (Indeed, n.d.).

In 2022, The Great Resignation witnessed an unprecedented 50.5 million US workers quit their jobs in search of better job opportunities, higher wages and remote work arrangements (Iacurci, 2023), costing US businesses over $1 trillion a year. Exit interviews revealed that 52% of voluntarily exiting employees expressed how steps could have been taken by the organisation to deter them from leaving (McFeely & Wigert, 2019), indicating how attrition is a largely self-inflicted problem.

As such, understanding the reasons that contribute to attrition and taking early intervention could potentially halve the attrition rate. However, identifying these reasons through regular exit interviews and employee surveys are ineffective as attrition is a lagging indicator. When an employee is ready to leave, it is too late to implement changes to persuade them to stay, and considerable loss would have been suffered by the company, making it highly important to track and predict factors leading to attrition in real-time (Luther, 2023).

Therefore, organisations should adopt a preemptive approach to tackle attrition, with Human Resources (HR) departments taking the lead to identify employees with high flight risk and improving company policies based on insights derived from employee data. However, without thorough data analysis, HR managers may rely on intuition to drive decisions, resulting in recommendations that are counterproductive towards reducing attrition due to ineffective policies.

Conversely, the emergence of analytical techniques using machine learning has paved the way for evidence-based HR to allow decisions to be made more objectively, transforming HR into a strategic, data-driven role rather than an administrative one. With analytics, organisations can simultaneously anticipate attrition accurately, while implementing targeted strategies to retain their employees.

## **1.2. Practical Motivation**

Every year, the automotive industry experiences an attrition rate of 46%, with a median employee tenure of only 3.1 years. Tesla, a pioneering electric vehicle (EV) manufacturer, is one such company in the automotive industry grappling with high attrition rates. In 2018, a total of 70% of Tesla’s workforce was replaced, indicating a systemic issue within the organisation (Owusu, 2022).

The cost of replacing an employee ranges from half to two times the employee's annual salary (McFeely & Wigert, 2019). With 140,000 people in Tesla and an average salary of $50,000 (Tesla, 2023), the total cost of attrition is approximately $2.45 to $9.8 billion per year - costs that could have been easily avoided. Beyond financial expenses, attrition also includes softer costs, such as reduced morale and poor reputation among attractive job candidates and existing employees (Luther, 2023). Moreover, high attrition affects customer loyalty as it can lead to inconsistent customer experiences, thereby eroding trust in the company.

To tackle attrition, numerous companies have employed machine learning and analytical techniques to uncover factors contributing to attrition. For instance, Nielsen's predictive model highlighted the importance of the first year in an employee's tenure. By leveraging this insight, they refined their retention strategy, resulting in the successful transition of 40% of vulnerable employees to new roles and increasing employee retention by 48%. Experian’s predictive model revealed that teams with over 10 to 12 members and employees who relocated further from the office had increased flight risk, allowing them to pursue targeted strategies that resulted in cost savings of $8 to $10 million dollars (van Vulpen, n.d.). Through predictive analytics, both companies were able to tailor their solutions to address the specific factors impacting attrition. This led to substantial cost savings by uncovering hidden patterns that traditional qualitative methods would have overlooked.

The cost savings derived through leveraging machine learning is evident. However, not only has Tesla not utilised machine learning in its employee attrition strategy, it has pursued policies that are contrary to employee retention, such as requiring employees to return to office full-time, when many of its Silicon Valley competitors no longer require the same (Jin & Datta, 2022). As societal values evolve across generations, employees may increasingly prioritise aspects such as work-life balance over salary considerations. Moreover, the decision to leave a job is influenced by various factors, making it challenging to discern whether departures stem from organisational factors conflicting with generational norms or personal factors. As such, a more precise methodology is needed to determine the relative proportion of personal factors and organisational factors that contribute to attrition rates.

# **2. Project Objectives**

## **2.1. Business Targets and Outcomes**

3 Key business targets were defined. Fulfilling these targets through qualitative and quantitative solutions indicates a successful analytics-driven business strategy. The business targets are:

1. **Attrition Rate:** Attrition rate is the primary metric indicating the percentage of employees who leave Tesla over a year. Reducing this rate signifies a direct impact of our analytics-driven strategy in improving employee retention.
2. **Cost Savings:** Cost savings is associated with actively retaining employees through employee feedback mechanisms compared to attrition costs arising from not pursuing these strategies. Achieving overall cost savings indicates a successful business strategy, displaying how our analytical approach has helped Tesla to improve employee retention, productivity, and decrease costs.
3. **Productivity and Performance:** Lastly, the implementation of these strategies will positively impact employee productivity and performance. Should our analytics-driven strategy be successful, future employees will be less likely to leave Tesla due to a stronger cultural fit and alignment with its values, thereby fostering greater engagement and involvement in the company. By deducing the most significant factors influencing attrition and the variables associated with these factors, we can refine Tesla’s retention strategies by optimising these factors to improve workplace performance.

## **2.2. Personal and Organisational Factors**

To understand the relationship between factors causing attrition, a definition of Personal and Organisational factors are needed. Personal factors are characteristics of the employee that are largely out of the control of the company, while Organisational factors are those that can be changed by Tesla. The list of factors are broken down as such:

**Personal:** Age, DistanceFromHome, Education, EducationField, EnvironmentSatisfaction, Gender, JobSatisfaction, MaritalStatus, NumCompaniesWorked, PerformanceRating, RelationshipSatisfaction, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole

**Organisational:** BusinessTravel, Department, JobInvolvement, JobLevel, JobRole, MonthlyIncome, OverTime, PercentSalaryHike, StockOptionLevel, TrainingTimesLastYear, WorkLifeBalance, YearsSinceLastPromotion, YearsWithCurrManager

# **3. Methodology**

## **3.1. Data Cleaning**

**Identifying Missing Values**

The dataset is complete and there are no missing values. Checking for missing values increases the reliability of the dataset, thereby improving the precision of subsequent analyses performed on the dataset and the model’s ability to accurately identify factors that contribute towards a higher attrition rate.

**Removing Duplicate Rows**

Duplicate rows in the data may arise from errors during data entry. Checking and removing duplicate rows is essential as it prevents the same rows from being present in the train and test data during model building, which may cause the models to “memorise” the data rather than mapping the patterns within the variables.

**Feature Selection**

As the columns StandardHours, EmployeeCount, and Over18 contained identical values for all records, providing no useful information in relation to attrition, they were dropped. Likewise, EmployeeNumber was dropped as it contained a unique identification number for each employee and is not relevant to attrition. Lastly, as MonthlyIncome is a more accurate representation of an employee’s salary accounting for additional remuneration, DailyRate, HourlyRate, and MonthlyRate were removed to reduce multicollinearity, which occurs when multiple independent variables are highly correlated with one other, affecting the interpretation of parameters for regression models (Bayman & Dexter, 2021). The remaining 27 variables are all associated with attrition rates based on domain knowledge and were selected.

**Identifying Outliers**

To identify outliers in each column, boxplots were used for clear visual representation to allow for easier interpretation and comparisons between each column (Appendix A).

## **3.2. Exploratory Data Analysis**

|  |  |  |
| --- | --- | --- |
| **3.2.1. Univariate Analysis** | | |
| **Distribution of Attrition Cases** | **Distribution of Monthly Income** | **Distribution of Performance Rating** |
|  |  |  |
| The dataset is highly imbalanced, with 16.1% of employees having Attrition: Yes and 83.9% having Attrition: No. This is expected as employee attrition in companies are typically minority cases. However, predicting all “No”s for Attrition would already result in a high baseline accuracy score of 83.9%. As such, accuracy might not be a suitable indicator for model performance. Additionally, steps should be taken to balance the dataset to prevent the models from overpredicting on the majority class. | The histogram for MonthlyIncome displays a large positive skew, with the majority of employees’ monthly income within the $2,500 bracket. However, the median ($4,919) and mean monthly incomes ($6,503) are significantly higher than the mode. The skew is affected by a minority group of employees who earn disproportionately more, indicating a large pay disparity within the company. This suggests that fairer compensation policies can be taken to improve employee satisfaction. | The PerformanceRating for every employee is “Excellent” or greater, indicating that every employee is motivated and performing well. However, this might suggest a lack of differentiation in the performance review process, possibly undermining the effectiveness in identifying and rewarding high-performing employees as well as underperformers. Additionally, as attrition rates persist, this could indicate that the issue lies in poor organisational policies rather than personal factors. |

|  |  |
| --- | --- |
| **3.2.2. Bivariate Analysis** | |
| **Mosaic Plot of Overtime against Attrition** | **Stacked Barplot of Marital Status against Attrition** |
|  |  |
| The mosaic plot compares the relative proportion of employees who work OverTime against Attrition. Given that an employee has to work overtime, the conditional probability of the employee leaving is significantly higher (30%) compared to that for no overtime (10%). Employees who work overtime have significantly higher rates of attrition than the overall average attrition rate of 16.1%, while those who do not work overtime fall below the average rate. As such, Tesla can look into policies that reduce employees’ working hours. However, given the nature of Tesla where extreme working hours are required (Hawkins, 2023), addressing long working hours can be difficult to resolve. | The barplot illustrates that single employees showcase a notably higher attrition rate compared to their married and divorced counterparts. Specifically, 25.53% of single employees contribute to this rate, whereas only 12.48% and 10.09% of married and divorced employees respectively contribute towards a higher attrition rate. This indicates that single employees are twice as likely to leave the company. Consequently, Tesla can look into implementing targeted retention strategies to mitigate the higher attrition rates amongst single employees and increase overall employee retention within companies. |
| **Analysis of Relationship between Monthly Income and Attrition** | |
| **Boxplot of Monthly Income against Attrition** | **Bar Chart of Monthly Income by Department** |
|  |  |
| The median monthly income for Attrition: “Yes” group ($3,202) is significantly lower than Attrition: “No” group ($5,204), suggesting how employees with lower monthly income are more likely to leave the company. Consequently, the conditional probability of an employee leaving given that they have low income is higher than the standalone probability of them leaving (Appendix B). This is consistent with the understanding that poor remuneration is linked to higher rates of attrition. Given that most employees have monthly income between $2,000 to $3,000 from the univariate analysis, a substantial portion of the workforce falls within a salary range that may contribute to their decision to leave the company. Hence, this indicates that monthly income could be important in determining and predicting employee attrition. | The average monthly income is always lower for those in Attrition: “Yes”’ group. Human Resources has the highest income disparity between employees who stay and employees who leave, with the pay gap between Attrition: “Yes” and Attrition: “No” decreasing for Research & Development and Sales. One surprising finding is that despite employees in Human Resources and Sales who stay having a similar average income, there is a distinct difference in average income for employees who leave. This indicates that employees in Sales are more likely to leave for smaller pay disparities than Human Resources, which might be due to factors other than MonthlyIncome related to the job. |

|  |  |
| --- | --- |
| **3.2.3. Multivariate Analysis** | |
| **Proportion of Attrition and Overtime Rates of Employees across Job Roles** | **Average Environment Satisfaction by Job Role**  **factored by Attrition** |
|  |  |
| The graph compares the relative proportion of employees in different roles who leave the company against whether they work overtime or not. Employees who work overtime tend to have higher rates of attrition across all job roles with the exception of Healthcare Representative, which has slightly lower attrition rates when working overtime. Sales Representatives display the highest attrition rates among all the roles, with 66% of Sales Representatives leaving the company. Interestingly, attrition rates in response to working overtime is perceived differently across job roles, with Managers being 4 times more likely to leave when having to work overtime than when they do not have to work overtime. | Employees in the Attrition: “Yes” group have lower EnvironmentSatisfaction levels on average compared to those in Attrition: “No” group across all roles with the exception of “Research Director”.  This suggests that across almost every job role, employees are dissatisfied with their working environment, notably except for research directors, possibly due to having a better workplace culture within the department as compared to other job roles. |

|  |  |  |
| --- | --- | --- |
| **3.3. Feature Transformation** | | |
| **Before Transformation** | **After Transformation** | **3.3.1. Log and Square-root Transformation**  To address skewness in data, log-transformation was applied on highly right-skewed variables MonthlyIncome, DistanceFromHome, and PercentSalaryHike. As TotalWorkingYears, YearsAtCompany, and NumCompaniesWorked have data values between 0 to 1, square-root transformation was applied instead to avoid negative values, which would impact the effectiveness of model training.  The effect of log and square-root transformation decreases the skewness of the variable, as seen in the density plot becoming more symmetric and normal (Appendix C). The Q-Q plot better approximates a straight line, where the points from -1 to 1 on the x-axis lie close to the red line, which shows that the assumption of normality is better fulfilled. While the tail ends of the points deviate slightly from the red line, with over 1400 data points, the Central Limit Theorem will mitigate these deviations. This means that slight non-normality in the residuals will not compromise the model’s validity. |
| **Density Plot** | **Density Plot** |
| **Q-Q Plot** | **Q-Q Plot** |
| **Before Winsorization** | **After Winsorization** | **3.3.2. Winsorization**  For variables that remained highly skewed after log-transformation, winsorization was used instead. Winsorization changes extreme values in the data to less extreme ones, allowing us to retain the data in extreme observations while reducing skewness. As outliers largely lie above the upper quartile of the boxplot, the top 10% values of the variables TrainingTimesLastYear, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrManager were reduced to the 90th percentile.  After Winsorization, the skewness of the density plot is reduced, becoming slightly more normal. A larger peak is seen at the tail end where the values are capped off. The outliers in the boxplot are replaced with a larger upper quartile. Thus, the distribution of the variables becomes closer to normal, improving the performance of models that assume normality of data, such as logistic regression. Moreover, the effect of outliers and extreme values is reduced. This can prevent overfitting and improve the generalisation ability of the models. |
| **Density Plot** | **Density Plot** |
| **Boxplot** | **Boxplot** |

# **4. Machine Learning and Optimisation**

The target variable, Attrition, is a binary categorical variable that represents the probability of an employee departing from Tesla. A positive outcome denotes that an employee has left the company, while a negative outcome denotes that an employee remained in the company.

A train-validation-test split (70%-15%-15%) was used, with each model trained on the train set and optimised on the validation set to prevent overfitting the model on the test set, and finally predicted on the test set which is completely unseen. To optimise the models, GridSearchCV was used, which performs an exhaustive search of specified parameter values to determine the most appropriate combination of values that results in the best performance for our dataset.

To balance the dataset, Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) were used to balance the dataset by filling up the minority class (Attrition: “Yes”) with samples generated with similar features based on the K-Nearest Neighbours algorithm. This helps to reinforce learning patterns when training the models. Finally, since identifying attrition is the goal, it is important to identify all cases where attrition is present even if they are wrongly identified. Consequently, a better performance metric than accuracy is needed, namely the recall and precision metrics.

**Recall** measures the ratio of attritions predicted over the total number of actual attritions:

**Recall =**

Prioritising recall over accuracy increases the number of positive predictions, increasing both True Positive Rate (TPR) (correct predictions of attrition) and False Positive Rate (FPR) (incorrect predictions of attrition). As the number of actual attritions are small, this decreases the overall accuracy, but correctly identifies a larger segment of actual positives. However, relying solely on recall may result in the models making incorrect predictions for attrition excessively, incurring unnecessary costs. As such, the precision metric was also considered.

**Precision** measures the number of actual attritions out of the number of attrition predictions:

**Precision =**

This ensures that the model correctly identifies employees at risk of attrition without unnecessarily penalising those who are not.

Since the cost of failing to predict an employee’s departure when they actually leave (false negatives) is higher than the cost of incorrectly predicting that an employee will leave when they actually remain (false positives), priority was given to **maximising recall** during the optimisation of the modesl. This would minimise the risk of Tesla overlooking critical signals of attrition, thus ensuring that necessary interventions can be taken to retain valuable employees.

| **4.1. Logistic Regression** | | |
| --- | --- | --- |
| **4.1.1. Implementation** Logistic regression is a statistical modelling technique used for classification. Independent variables are fitted in the linear layer of the model to produce the log-odds, which is the log of the probability of Attrition: “Yes” divided by the probability of Attrition: “No”. The sigmoid function is used to map the log odds to a value between 0 and 1, which determines whether a row will be classified as Attrition: “Yes” or “No” based on the classification threshold (default: >0.5 represents Attrition: “Yes”). | | |
|
| Train Set    Accuracy: 0.907677  Precision: 0.893617  Recall: 0.497041 | Validation Set    Accuracy: 0.895455  Precision: 0.722222  Recall: 0.419355 | **4.1.2. Analysis of Logistic Regression** The initial logistic regression model has a high accuracy on the validation set (0.90) due to the large number of true negatives, which are instances correctly predicted as Attrition: “No”. However, this is misleading as the model tends to predict “No Attrition” for most instances in order to increase the overall accuracy score. Logistic regression optimises the log-likelihood objective function, which assumes equal importance for each binary class. Hence, the imbalance in the dataset with significant underrepresentation of Attrition: “No”, results in a bias towards predicting the majority class Attrition: “Yes”. |
| Train Set (ADASYN)    Accuracy: 0.824912  Precision: 0.808390  Recall: 0.846793 | Test Set    Accuracy: 0.778281  Precision: 0.414286  Recall: 0.783784 | **4.1.3. Optimisation** Oversampling was used to correct the issue of imbalance. Consequently, GridSearchCV (Appendix D) with ADASYN achieved the best test set results. While test set accuracy decreased slightly from 0.90 to 0.78, this is justified by the large increase in recall from 0.42 to 0.78. This means that by decreasing the number of true predictions for Attrition: “No” by 12%, the model’s prediction for Attrition: “Yes” increases by 36%. As employee attrition is generally more costly than employee retention, prioritising recall helps to predict a larger segment of employees about to leave, which suits Tesla’s specific business needs to achieve greater cost savings. |
| **4.1.4 Backward Elimination** Logistic Regression allows us to perform backward elimination for feature selection. Coefficients of variables with high p-value (>0.05) are removed from the equation one at a time and the model is fitted again. A p-value of 0.05 and greater indicates a 5% or greater probability that the coefficient has no relationship with Attrition (i.e. coefficient = 0). Variables with low absolute coefficients (<0.05) were also removed. The coefficient represents the practical significance of the variable on the log-odds of Attrition, with higher coefficient magnitudes indicating a greater impact on the final log-odds. Consequently, all other variables had low p-values (<0.05) and high absolute coefficients (>0.1). The linearity assumption of each variable and the log-odds of Attrition was also fulfilled, ensuring the reliability of the interpretation of the model’s predictions. | | |

|  |  |  |
| --- | --- | --- |
| **4.2. Neural Network** | | |
| **4.2.1. Implementation** A Neural Network is defined with one input, one output, and a number of hidden layers. Since Attrition is a categorical variable, the sigmoid function is used in the output layer to compress the output to a value between 0 and 1, and the model is trained to reduce binary cross-entropy (loss function). To allow the model to pick up non-linear patterns, an activation function, rectified linear unit (ReLU), is used to transform the data in each of the training layers. At the start, each neuron is fitted with random weights, performs a weighted sum of the inputs, applies the activation function to the sum, and then passes the result to the neurons in the next layer repeatedly until a final prediction for Attrition is made in the output layer. The weights are then fine-tuned during backpropagation by revising the weights on all the neuron paths to reduce the overall loss, producing a model that fits to the patterns in the train set. | | |
| Train Set    Accuracy: 0.927114  Precision: 0.935185  Recall: 0.597633 | Validation Set    Accuracy: 0.890909  Precision: 0.769231  Recall: 0.322581 | **4.2.2. Analysis of Neural Network** The model achieves a very high recall (0.70) and precision (0.92) on the train set, but poor recall (0.27) and precision (0.57) on the test set. This means that out of all attrition cases, only 27% were successfully identified, which suggests that the model is overfitted to the train data and lacks generalisation.  The model can be optimised by changing the number of hidden layers to map out more complex features, varying the number of nodes to reduce the impact of irrelevant features, changing the activation function to allow the model to identify non-linear patterns in a different way, reducing the batch size to allow the model to fit the train data better at the expense of time, and using early stopping to ensure that the model does not overfit to train data. |
| Train Set (ADASYN)    Accuracy: 0.820799  Precision: 0.796031  Recall: 0.857482 | Test Set    Accuracy: 0.782805  Precision: 0.422535  Recall: 0.810811 | **4.2.3. Optimisation** After optimisation, ADASYN produces the best model with the highest balance of recall (0.64) and precision (0.56) compared to ADASYN’s recall (0.67) and precision (0.38). The model can now predict 64% of actual attritions, and correctly identifies 56% of attrition cases among its attrition predictions. Hence, creating additional samples to simulate the patterns of cases with positive attrition helps to reinforce the model’s learning, improving its reliability.  While the model can be used to make predictions on specific employees with attrition risk, the neural network is not a suitable model to derive business insights due to randomness in the backpropagation process. As such, other models have to be considered to determine feature importance. |

|  |  |  |
| --- | --- | --- |
| **4.3. CART - Classification Tree** | | |
| **4.3.1. Implementation** The CART model determines the best splitting criteria by applying samples from the train set at the root node. The Gini Impurity or Entropy is minimised when choosing the best splitting criteria (Appendix D),resulting in two purest child nodes through a binary split at each parent node. This process of splitting is performed iteratively until a predefined lenient stopping condition is reached, producing the maximal decision tree with the greatest number of nodes. | | |
|  | | **4.3.2. Growing the Maximal Tree** Due to its large size, the maximal tree has accuracy of 1 on the train set, which indicates complete overfitting on the train set. However, the validation set accuracy of only 0.75 suggests poor generalisation on unseen data. As the size of the tree increases, the Cross Validation (CV) error decreases until it reaches a minimum. The node with the lowest CV error can be found. The standard error of that node is added to the CV error to determine the error cap. The tree will be pruned on the first node that falls within the error cap as trees further down that node are statistically equivalent or worse in terms of errors. This process produces the simplest tree that reduces prediction error. |
| Train Set    Accuracy: 0.94655  Precision: 0.925373  Recall: 0.733728 | Test Set    Accuracy: 0.791855  Precision: 0.378378  Recall: 0.378378 | **4.3.3. Pruning the Tree** As such, the optimal trade-off between model complexity and predictive accuracy is one where the tree is pruned at the optimal complexity parameter (CV error = 0.678), resulting in the most significant reduction in impurity.  After pruning, the model’s accuracyon the validation (0.8) and test (0.79) sets improve, which indicates that it no longer overfits the train set and is able to generalise well to new data. However, recall (0.38) is still relatively low. This is because a single decision tree is particularly unstable as small variations in the train set can result in large differences to the tree built, leading to different predictions for the same validation set (Last et al., 2002). |
|  | | **4.3.4. Variable Importance** CART is suitable for businesses as it is highly interpretable and offers HR a tool to visualise variables that are significantly correlated with attrition. The variable importance chart ranks the features used in predicting attrition based on how influential they were in determining the splitting criteria, the number of times they appear in the tree, and their relative order within the tree. This provides business insights as to what variables Tesla should prioritise to achieve the most significant impact. MonthlyIncome has the highest variable importance of 0.24, which accounts for 24% of the variability in predicting Attrition, suggesting that income is the most significant factor associated with Attrition. However, due to the limitations of a single decision tree, the variables presented may not be reliable and more robust techniques are needed to improve model stability. |

|  |  |  |
| --- | --- | --- |
| **4.4. Random Forest** | | |
| **4.4.1. Implementation** As such, Random Forest, which uses multiple decision trees to aggregate the predictions from its trees (via majority voting), is implemented to produce a more stable model. Random forest employs the Bagging technique, which uses repeated sampling with replacement from the dataset to form a bootstrap sample, ensuring that small variations within the train set will not lead to significant differences in the model. Another technique, Random Subset Feature (RSF) selection, is used to prevent any variable from dominating the trees, stabilising model performance. Thus, Random Forest corrects the limitations of CART by producing more reliable results. | | |
|  | | **4.4.2. OOB Error Stabilisation** Due to random sampling, rows that are not chosen in each bootstrap sample for training the model are defined as Out-Of-Bag (OOB). The OOB samples, which are unseen data, act as additional test sets to be predicted on by each tree, where the combined OOB set error is computed. As more trees are used, the OOB error falls and stabilises at the minimum value of 0.054 for ≥1500 trees, indicating the best fit to the train set that generalises well to unseen data. |
| OOB Set (ADASYN)    Accuracy: 0.932432  Precision: 0.952677  Recall: 0.908551 | Train Set (ADASYN)    Accuracy: 0.996475  Precision: 1.000000  Recall: 0.992874 | **4.4.3. Optimisation of Standard Random Forest** After optimisation using the ADASYN train set defined on 1500 trees, Random Forest’s accuracy is 0.996 and recall is 0.993. The validation set achieves 0.88 accuracy and 0.35 recall, which is similar to the test set accuracy of 0.86 and 0.35 recall, validating and justifying the models’ results.  While the train set accuracy is close to 1, which might suggest overfitting, this is not the case for Random Forest due to RSF selection, which allows different variables to be tested despite each decision tree being grown to the maximum.  Finally, while the test set accuracy and recall is low, the OOB sample acts as unseen data on the original sample, and the high accuracy and recall on the OOB set indicates that the model is able to generalise well to unseen data. Thus, Random Forest is able to mitigate the weaknesses of a single CART model, producing more reliable predictions and insights. |
| Validation Set    Accuracy: 0.881818  Precision: 0.647059  Recall: 0.354839 | Test Set    Accuracy: 0.859729  Precision: 0.650000  Recall: 0.351351 |
|  | | **4.4.4. Variable Importance** Similar to CART, Random Forest computes the importance of each variable. The OOB error after random permutation of each variable is tested against the original OOB error. Large differences in OOB error after permutation indicates that the order of that variable is important in predicting attrition, and hence has high variable importance. This ensures reliability in determining the relative importance of each variable.  From the variable importance chart, Overtime is the most important feature with a score of 0.212. This means that Overtime accounts for 21.2% of the variability in predicting Attrition. The second most important is MaritalStatus: “Single”, with a score of 0.085, which is significantly lower than Overtime. Thus, whether an employee works Overtime disproportionately affects their chances of leaving the company. |

|  |  |  |
| --- | --- | --- |
| **4.5. XGBoost** | | |
| **4.5.1. Implementation** XGBoost is another model that aims to correct the high variance of CART by building a series of CART models sequentially to minimise the log-loss of the previous trees. An initial decision tree with a high bias and variance is built, with a second tree defined to fit the residuals of log-loss on the first tree. The two trees are combined to produce a better model that reduces the binary cross entropy error. These steps are performed until a certain stopping criteria is fulfilled, resulting in a model with a low log-loss on the train set. | | |
| Validation Set    Accuracy: 0.859091  Precision: 0.500000  Recall: 0.290323 | Test Set    Accuracy: 0.859729  Precision: 0.636364  Recall: 0.378378 | **4.5.2. Optimisation** XGBoost requires several hyperparameters to be tuned to prevent overfitting (Appendix D), resulting in a dramatic decrease in the binary cross entropy error. Randomised search was used to find the best hyperparameters, with the SMOTE dataset producing the best recall (0.29) and precision (0.5) on the validation set. Consequently, the best XGBoost model produced a recall of 0.38 and a precision of 0.64 on the test set. As the performance metrics on the validation and test sets are similar, this indicates that the model is able to generalise well to unseen data, and the insights gathered from the model are reliable. |

|  |  |
| --- | --- |
|  | **4.5.3. Variable Importance** Since XGBoost is built on CART, the variable importance can also be determined. The most significant variable used to predict attrition is OverTime with a score of 0.171, accounting for 17.1% of the variability in predicting Attrition, with the second most important variable being StockOptionLevel: 1 with a score of 0.043. This corresponds to the most important variable from Random Forest. Likewise, OverTime has a significantly higher variable importance than the second most important variable, which confirms the findings that addressing the reasons for OverTime will produce the greatest impact on attrition rates. |

## **4.6. Model Comparison and Evaluation**

The results of each optimised model are compared using the Precision-Recall curve to identify the optimal Precision-Recall ratio across different classification thresholds (a threshold, above which, results in a positive prediction) to determine the best model. This is achieved by finding the cost savings or loss incurred by each true or false prediction for each Precision-Recall ratio produced by the models that achieves the greatest reduction in attrition costs. The cost savings and losses are estimated as such:

**False Negative (FN)**: Should the model fail to identify a case of employee attrition correctly, no additional losses will be incurred as the employee would already have left. Rather, it signifies a missed opportunity in further cutting down attrition costs. Hence, the **loss** per employee for false negative prediction is derived as:

**FN Additional Loss per Employee = $0**

**True Positive (TP):** According to Gallup, it is estimated that the cost of replacing an employee ranges from half to two times the employee's annual salary (McFeely & Wigert, 2019). Furthermore, Tesla reports that the average annual salary of its employees is $50,000 (Tesla, 2023). Hence, by multiplying the average cost of replacing an employee by the average annual salary of a Tesla employee, the loss per employee attrition is derived as **$62,500**.

During exit interviews, 52% of employees who voluntarily left their company shared that proactive measures could have been taken to prevent their departure (McFeely & Wigert, 2019). Should Tesla successfully implement strategies to mitigate attrition, it could anticipate a 52% reduction in attrition among employees who had contemplated leaving. Hence, by multiplying the expected 52% reduction in attrition by the true positive loss per employee, the **cost savings** per employee for true positive prediction is derived as:

**TP Cost Savings per Employee = 0.52 \* 62,500 = $32,500**

**True Negative (TN):** Should the model predict an employee to stay when the employee actually stays, no additional retention measures need to be taken. Hence, the **cost savings** per employee for true negative prediction is derived as:

**TN Cost Savings per Employee = $0**

**False Positive (FP):** The cost associated with false positives involves the unnecessary expenditures on implementing retention strategies. It is estimated that the cost of enacting such retention strategies, including high compensation and benefits packages and customised career development plans, is between $1,000 to $10,000 per employee (Acrisure, n.d.). Hence, assuming that the cost follows a uniform distribution, by taking the mean of this range, we estimate the **loss** per employee for a false positive prediction as:

**FP Additional Loss per Employee = (1,000 + 10,000) / 2 = $5,500**

A formula can be derived for cost savings in terms of precision, recall, and total number of actual attritions (Appendix E). It can be applied for every Precision-Recall ratio produced by the various models to find the model that produces the highest cost savings, which represents the best model.

|  |  |
| --- | --- |
|  | **Business Value**  Thus, while Logistic Regression has the largest area, which generally indicates better classification performance, **Neural Network** remains the best model as it achieves the highest cost savings of $22,316 per actual attrition on average, with a precision of 0.51 and a recall of 0.82 at the classification threshold of 0.395. This is nearly 70% of the maximum cost savings achievable by a perfect classifier.  As such, using the Neural Network at classification threshold = 0.395, we can estimate the **expected cost savings** should Tesla utilise our model to predict attrition on our test set to be **$815,500 per employee**. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Value** | **Cost Savings Per Employee** | **Count** | **Total Cost Savings** |
| **True Negative** | $0 | 155 | $0 |
| **False Negative** | $0 | 7 | $0 |
| **False Positive** | -$5,500 | 29 | -$159,500 |
| **True Positive** | $32,500 | 30 | $975,000 |
| **Total** | **-** | **221** | **$815,500** |

With 140,000 employees, Tesla can **save $516 million per year** in attrition costs. Hence, the direct impact of our analytics-driven strategy can be seen through cost savings, which fulfils our first key business target.

# **5. Proposed Business Recommendations**

## **5.1. Business Insights**

Variable importance determines if a variable is influential in deciding the predictors of the model. As such, Tesla should identify the presence of these variables in particular when studying factors that relate to attrition rates. From the variable importance charts derived from Random Forest and XGBoost, the most important variables associated with a change in Attrition are OverTime, MaritalStatus, and StockOptionLevel. This closely corroborates with the findings from our Exploratory Data Analysis (EDA), where OverTime: “Yes”, MaritalStatus: “Single” and StockOptionLevel: 1 were factors related to higher attrition rates. All other factors had significantly less influence on the models’ decision-making, with variable importance scores below 0.050 or inconsistency in both variable importance charts. Hence, these factors were excluded from our analysis.

**OverTime:** OverTime is the most influential predictor of Attrition, with a variable importance score of 0.212 in Random Forest and 0.171 in XGBoost.Based on the EDA, employees who work overtime have significantly higher attrition rates on average than the overall attrition rate. A study found that the average productivity of a person is approximately three hours a day (Vouchercloud, n.d.), suggesting that overtime hours decrease employee productivity and efficiency. As such, frequent overtime is linked to burnout and higher rates of attrition, and Tesla should prioritise strategies that prevent employees from having to work overtime excessively.

**MaritalStatus:** The second most influential predictor is MaritalStatus with a variable importance score of 0.080 in Random Forest and 0.038 in XGBoost. 25.5% of singles leave the company on average, which is much higher than divorced (10.1%) and married (12.5%) employees. As divorced and married employees are more likely to have families to support than singles, they have more responsibilities to adhere to and value job stability, resulting in lower attrition rates. Consequently, singles are more easily attracted to opportunities provided by other companies and are more likely to leave. However, as MaritalStatus is a Personal factor, policy changes will not impact existing employees. As such, Tesla can provide greater support to married individuals and increase staff benefits such as maternity and paternity leaves to encourage married individuals to join the company as part of its hiring strategy.

**StockOptionLevel:** StockOptionLevel is the third most influential factor with a variable importance score of 0.050 in Random Forest and 0.048 in XGBoost. It refers to the number of company shares that an employee receives for compensation, with higher levels indicating more shares in possession. Referring to the EDA, employees who own more shares are less likely to leave the company. With higher StockOptionLevel, employees perceive a greater stake in the company as their performance within the company correlates to larger dividends from company profits when it is performing well. On the other hand, employees with low StockOptionLevel are not tied to the company and are more likely to leave. As such, compensating employees with company stock could significantly boost retention rates by improving employee loyalty.

## **5.2. Business Solutions**

### **5.2.1. Reducing Attrition Rates**

As the variable importance of OverTime is disproportionately higher than all other variables, Tesla should prioritise strategies to reduce overtime rates as tackling a single variable will lead to the highest decrease in the attrition rate, maximising the overall cost-benefit. We propose a 2-pronged approach to tackle overtime: reducing work hours and helping employees cope with overtime hours in the event that working overtime is unavoidable.

**Phase 1: Reducing Work Hours**

**3-Shift System:** Excessive overtime leads to prolonged physical strain, which poses significant risks to employee well-being and contributes to dissatisfaction (Moran, 2017). At Tesla, factory workers adhere to a compressed workweek schedule, labouring for four 12-hour shifts each week. During critical deadlines, employees often extend their shifts to 13-14 hours to fulfil job expectations. This issue is particularly prevalent among managerial roles, as supervisors must stay back to oversee their teams after regular shifts.

To prevent employees from having to work excessively long, Tesla should optimise its shift system. Transitioning from the current 2-shift system to a 3-shift model, each lasting 8 hours over 5 days, would effectively reduce the likelihood and necessity of same-day overtime as the workload is spread out over more working days. For managers, they could pass their supervision responsibilities to the subsequent shift's manager during shift handover, alleviating the burden of extended hours. In addition, this adjustment would ensure that even if overtime is occasionally necessary, the total working hours per shift remain reasonable, typically under 10 hours. By implementing a 3-shift system, the occurrence of excessively long shifts can be reduced, even during peak production periods. Additionally, staggered shifts and part-time schedules can be implemented to accommodate fluctuating work demands without relying heavily on overtime (Lehndorff, n.d.).

**Cross-Train Employees:** Given the fixed workforce size within each department and the prevalent need for overtime to meet production demands, it is crucial to equip workers with skills across multiple tasks or roles to enhance operational flexibility and responsiveness to workload fluctuations. An area that stands out is the sales department, where our analysis revealed that 66% of attrition is due to excessive overtime demands. In contrast to production line engineers, the skills possessed by sales representatives are less technical and more transferable. This characteristic makes cross-training easier and more feasible. Therefore, Tesla could consider initiating a cross-training program between the sales and customer service departments. Employees in both departments already possess a deep understanding of Tesla's products and services and demonstrate effective communication skills essential for interacting with customers, understanding their needs, and providing solutions. Through cross-training, this approach disperses responsibility more evenly across the departments during peak sales periods, reducing reliance on specific individuals and mitigating the need for excessive overtime. Moreover, implementing a comprehensive training and development program can empower employees to seamlessly transition between tasks or roles. This not only prevents burnout, but also fosters a more dynamic workforce (Truein, 2023).

**Phase 2: HR Monitoring Platform**

Reducing working hours might not be feasible due to Tesla’s ambitious goal of becoming a leading player in the automotive industry through creating affordable electric vehicles. With its emphasis on employee presence and a policy requiring employees to report to office “at least 40 hours a week” (Bursztynsky, 2022), it is inevitable that employees have to work long hours. However, measures can be taken to help employees cope with overtime hours by finding out the factors that make employees stay in spite of having to work overtime.

|  |  |
| --- | --- |
|  | Consequently, a secondary Random Forest model was built on a subset of the data that contains only employees with OverTime: “Yes” to identify significant variables that are associated with employee retention.  MonthlyIncome was the most important variable in predicting attrition among employees who work overtime, with a variable importance score of 0.215. This can be explained by how employees who have to work overtime would expect fair compensation for their extra hours. |
|  | This observation is corroborated by the boxplots shown, which display how employees who leave generally have lower median income than those who chose to stay despite working overtime or reporting low satisfaction levels. Receiving higher pay also means higher opportunity cost for the employee should they choose to leave, as they would have to find a job with a similar level of pay. Hence, increasing monthly income can encourage employees who work overtime to stay, thereby lowering the attrition rate.  Notably, Age and DistanceFromHome were found to be crucial deciding factors for attrition among employees who work overtime, with variable importance scores of 0.1 and 0.085 respectively. This can be explained by higher transportation expenses and longer travelling times incurred on top of working overtime, which leads to shorter rest times. Older employees also require more rest. All these will further drive down motivation to stay in the company, increasing attrition rates among overtime employees. |

However, as Age and DistanceFromHome are Personal factors, they can only be targeted through changes to Tesla’s hiring strategy by ensuring that future employees are a better fit and will be less likely to leave the company.

**Business Value**

From this insight, an interactive dashboard can be prepared to monitor Tesla’s attrition rates, which serves as a monitoring platform for HR (Appendix F). HR can compare attrition rates for different filters applied, with each categorical variable utilising a checkbox and each numeric variable utilising a slider. By adjusting each filter and comparing the attrition rate output by the dashboard, HR can identify the optimal compensation that reduces attrition rates. Finally, the Neural Network is integrated into the dashboard to make predictions on employees in real time, allowing HR to preemptively identify employees at risk of attrition. Thus, our analytics-driven strategy directly reduces the attrition rate, fulfilling our second key business target.

### **5.2.2. Performance and Productivity - Scorecards**

Alongside our employee retention strategy, attrition rates can be reduced through improvements to Tesla’s retention process, ensuring that future employees are a good fit and display characteristics associated with lower levels of attrition. Consequently, the variable importance scores for Organisational and Personal variables (Appendix G) can be used to curate scorecards for 2 use cases. **Personal variables** will be used to create a **cultural fit scorecard**, which will be used to tabulate the employee’s **cultural fit score**. **Organisational variables** will be used to create an **organisational alignment scorecard**, which will be used to tabulate the employee’s **organisational alignment score**.

**Sample Cultural Fit Scorecard**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Weight** | **Scoring Criteria** | **Weighted Score** |
| **Total Working Years** | **19%** | Total working years between 20-40 receives 5 points, 11-20 receives 3 points, 0-10 receives 1 point |  |
| **Age** | **16%** | Age between 41-60 receives 5 points, 21-40 receives 3 points, 18-20 receives 1 point |  |
| **Years At Company** | **12%** | 21-40 years at company receives 5 points, 0-20 years at company receives 1 point |  |
| **No. of Companies Worked At** | **11%** | Working at 5-9 companies receives 5 points, 0-4 companies receives 1 point |  |
| **Distance From Home** | **10%** | Distance of 0-14 from home receives 5 points, 14-29 receives 1 point |  |
| **Environment Satisfaction** | **9%** | High and Very High environment satisfaction receives 5 points, Medium receives 3 points, Low receives 1 point |  |
| **Marital Status** | **8%** | Divorced employees receive 5 points, Married receive 3 points, Single receive 1 point |  |
| **Gender** | **5%** | Female employees receive 5 points, Male receive 1 point |  |
| **Job Satisfaction** | **4%** | Very high job satisfaction receives 5 points, Medium and High receive 3 points, Low receives 1 point |  |
| **Education Level** | **3%** | Doctorate degree receives 5 points, Master degree receives 4 points, College degree receives 3 points, Bachelor degree receives 2 points, Below college education receives 1 point |  |
| **Relationship Satisfaction** | **2%** | Medium to Very High relationship satisfaction receive 5 points, Low receives 1 point |  |
| **Education Field** | **1%** | Medial and other education fields receive 5 points, Life sciences receives 4 points, Marketing receives 3 points, Technical receives 2 points, Human resources receives 1 point |  |

To illustrate, the above sample **cultural fit scorecard** assesses an employee’s **cultural fit score** using 12 variables. For each variable, the employee's data is used to assign points to that relevant variable based on the correlation direction of the variable to attrition. The points received for each variable will then be weighted based on the significance of that variable. For instance, under Education Level, if an employee has a bachelor degree, the employee will receive 4 points with a weight of 3%. The weighted score can then be derived using the following formula:

**Weighted Score = (Assigned Points for the Variable / Maximum Points for the Variable) x Weight**

After calculating the weighted score for each variable, the total score will be calculated. A lower score implies a higher probability of attrition than a higher score.

The points were derived based on trends gathered from the EDA. For instance, since the EDA revealed that employees with below college education have higher rates of attrition compared to other education levels, they will receive lower points on the card. Furthermore, each weight was derived from the variable importance chart, offering a direct application of the model’s predictive accuracy on the significance of each variable and resulting in a more reliable indicator of attrition than attributing equal weight to each variable.

|  |  |
| --- | --- |
| **Strategic Heat Map**  Using the tabulated scores from the scorecards, a **strategic heat map** can be plotted to provide a visual representation of where an employee stands (Mondore et al., 2011). The horizontal axis measures the **cultural fit score** from the cultural fit scorecard, while the vertical axis measures the **organisational alignment score** from the organisational alignment scorecard, with both axes centred at the median score of 50%. In turn, the employees’ position on the heat map will guide and inform the development of actionable goals and plans for employee retention. |  |

**Focus (Lower Left Quadrant):** Firstly, employees who fall within this quadrant score below the median on both organisational alignment and cultural fit. These employees are “red flags” who are at the highest risk of leaving as they lack the commitment towards Tesla’s organisational goals nor compatibility with the company’s values and interests, which are crucial for long-term engagement and success. This is where Tesla should focus most of its time, attention, and resources in order to achieve the highest business impact.

Consequently, the heat map facilitates the quick identification of such employees at-a-glance, enabling the development of a **contingency plan** to retain them.

* As a proactive measure, Tesla can conduct **stay interviews**, during which managers will engage in discussions with employees about their reasons for wanting to leave and work together to find solutions to address their issues.
* By doing so, Tesla can **customise retention plans** based on the needs and concerns of these employees, such as providing opportunities for career development, increased recognition and rewards, or modifications to their compensation and benefits packages.

This will enable Tesla to intervene early and take preventive measures to reduce attrition. When employees feel that their needs and concerns are valued, it increases their willingness to stay in the company.

**Maintain (Lower Right Quadrant):** Employees who fall within this quadrant score above the median on cultural fit, but below the median on organisational alignment. While these employees demonstrate a commitment towards Tesla’s values and beliefs, further measures can be taken to ensure their continued involvement in the company’s goals and objectives. To achieve this, the following strategies can be put in place to **increase employee engagement**.

* Tesla can establish **cross-functional teams** to bring together employees from different departments to collaborate on company-wide strategic initiatives.
* **Pulse surveys and focus group discussions** may also be conducted to provide employees with a platform to give feedback and have a greater say in future plans.
* Furthermore, the **performance management system** can be revised to better align key performance indicators (KPIs) with company goals.

These strategies will motivate employees to remain in Tesla and work towards fulfilling its objectives and targets.

**Monitor (Upper Left Quadrant):** Employees who fall within this quadrant score above the median on organisational alignment, but below the median on cultural fit. While most of these individuals would have been identified and turned away during the recruitment process, a few of them might slip through undetected due to shifts in personal priorities and lifestyle choices while working at Tesla.

* To address this issue, Tesla should **consistently highlight its fundamental values, beliefs, and desired behaviours** during onboarding, training programs, and internal management communications.
* **Diversity, Equity, and Inclusion (DE&I) initiatives** can also be established to promote respect and understanding between employees from different backgrounds and create a more welcoming and supportive workplace environment.
* Additionally, Tesla can organise **team-building events and corporate social responsibility (CSR) activities** to foster a closer bond between employees.

These strategies will increase employees’ sense of belonging, making them more likely to stay in the company.

**Promote (Upper Right Quadrant):** Lastly, employees who fall within this quadrant score above the median on both organisational alignment and cultural fit. These valuable employees are important drivers of business outcomes as they display a strong loyalty towards Tesla’s organisational goals, as well as a good match with the company’s values and interests.

* To further strengthen their dedication and foster long-term retention, such employees should be rewarded and **considered for promotion into higher-level roles**. This will demonstrate Tesla’s recognition and appreciation of their efforts and accomplishments, incentivising them to sustain their high performance and remain in the company.

**Business Value**

The heat map provides HR with an objective metric to monitor employees’ well-being within Tesla, allowing the company to pay particular focus on employees who need the most support and the development of targeted approaches aimed at prioritising the unique needs of every employee. With the implementation of the above strategies, Tesla employees will feel more valued and motivated to work towards the goals and objectives of the company, thereby increasing productivity.

Thus, through analytics, a strategy was developed to assess and quantify employee productivity and performance, which fulfils our last key business target.

# **6. Conclusion**

## **6.1. Limitations and Future Work**

**Assumption of Voluntary Attrition:** A limitation in our study is the assumption that all cases of employee attrition at Tesla resulted from **voluntary attrition**. However, **involuntary attrition** such as firing or layoffs may also occur in reality, particularly in the automotive industry. In particular, major automotive companies such as GM, Ford, and Stellantis have laid off a large number of employees as the industry transitions towards electric vehicles (Naughton, 2023).

Therefore, the causes of employee attrition may be misidentified should attrition rates in the dataset be primarily composed of involuntary attrition, which is driven by other factors such as company performance, restructuring, and workforce reductions. As such, studying data that distinguishes between voluntary and involuntary attrition will help to provide more insights and real-world applicability to our findings, especially in a turbulent industry like tech where mass retrenchments often take place.

**Industry-Specific Data:** The dataset used to train our model is exclusive to IBM, a tech company. This could restrict the applicability of our results to Tesla which belongs to the automotive industry. Due to the inherent differences between these sectors, the workplace environment and HR practices at Tesla may vary compared to IBM. Therefore, the predictions of the model may not be directly transferable to Tesla without taking into consideration factors specific to the automotive industry.

Thus, instead of utilising secondary employee data, Tesla should conduct an internal assessment using approaches such as 360-degree feedback, pulse surveys, and focus group discussions to facilitate the collection of data samples that are relevant to the company’s unique culture, operations, and workforce dynamics. By applying the methodological framework of our study to data specific to Tesla, more accurate and contextualised insights can be derived.

**Timeline of Data:** Another constraint of our study is that the employee data represents data from a **single moment in time**, which does not consider how the variables evolve and impact one another over extended periods. Variables such as job satisfaction, monthly income, and performance ratings are not static, but often fluctuate due to changing workplace conditions, policy adjustments, and job roles. This hinders the exploration of temporal dynamics, lagging effects, and attrition patterns over time.

To address this limitation, future studies should adopt **time series analysis** with longitudinal data gathered regularly. Data panels can be created to track the same group of employees over different time periods to examine how variables influence one another over time.

## **6.2 Closing**

In conclusion, our study findings set the foundation for Tesla to implement preventive and prescriptive measures by identifying the relationships between key personal and organisational factors that influence employee attrition. Through this paper, we demonstrate how the insights from our predictive models can be used to develop comprehensive HR strategies to decrease employee attrition. By doing so, we offer companies such as Tesla clear guidance on how to best allocate their time and resources to reduce flight risk.

# **7. References**

Acrisure. (n.d.). Three Cost-Effective Strategies for Employee Retention.<https://www.myhcg.com/blog/three-cost-effective-strategies-for-employee-retention/>

Bayman, E. O., & Dexter, F. (2021). Multicollinearity in Logistic Regression Models. Anesthesia and analgesia, 133(2), 362–365.<https://doi.org/10.1213/ANE.0000000000005593>

Bursztynsky, J. (2022). Elon Musk tells Tesla workers to return to the office full time or resign. CNBC.<https://www.cnbc.com/2022/06/01/elon-musk-reportedly-tells-tesla-workers-to-be-in-office-full-time-or-resign.html>

Hawkins, A.J. (2023). Tesla’s ‘ultra hardcore’ work culture — as told by its employees. The Verge.<https://www.theverge.com/2023/8/16/23833447/tesla-elon-musk-ultra-hardcore-employees-land-of-the-giants>

Iacurci, G. (2023). 2022 was the ‘real year of the Great Resignation,’ says economist. CNBC.<https://www.cnbc.com/2023/02/01/why-2022-was-the-real-year-of-the-great-resignation.html>

Indeed. (n.d.). Why Employee Attrition Matters.<https://www.indeed.com/hire/c/info/employee-attrition>

Jin, H. & Datta, T. (2022). Elon Musk tells Tesla staff: return to office or leave. Reuters.<https://www.reuters.com/technology/musk-memo-tesla-staff-return-office-or-leave-company-2022-06-01/>

Last, M., Maimon, O., & Minkov, E. (2002). Improving Stability of Decision Trees. \*International Journal of Pattern Recognition and Artificial Intelligence\*, 16(2), 145-159. <https://doi.org/10.1142/S0218001402001599>

Lehndorff, S. (n.d.). Working Time and Operating Hours in the European Automotive Industry. Institute for Work and Technology.<https://www.iat.eu/aktuell/veroeff/am/lehndorff00de.pdf>

Luther, D. (2023). 12 Employee Turnover and Retention KPIs to Measure in 2024. Oracle NetSuite.<https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover-kpis-metrics.shtml>

McFeely, S. & Wigert, B. (2019). This Fixable Problem Costs U.S. Businesses $1 Trillion. Gallup.<https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx>

Mondore, S., Douthitt, S., & Carson, M. (2011). Maximizing the impact and effectiveness of HR analytics to drive business outcomes. People and Strategy, 34(2), 20-27.

Moran, J. (2017). Time for Tesla to Listen. Medium.<https://medium.com/@moran2017j/time-for-tesla-to-listen-ab5c6259fc88>

Naughton, N. (2023). The tech layoff contagion is hitting automakers — but it's not like what's happening at Google, Meta, and Microsoft. Business Insider. <https://www.businessinsider.com/layoffs-gm-auto-indsutry-and-tech-are-not-the-same-2023-3>

Owusu, T. (2022). Tesla Worries Heat Up As Major Exec Goes MIA. TheStreet.<https://www.thestreet.com/investing/tesla-worries-heat-up-as-major-exec-goes-mia>

Tesla. (2023). Annual Report pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934.<https://www.sec.gov/Archives/edgar/data/1318605/000162828024002390/tsla-20231231.htm>

TrueIn. (2023).Effective Strategies for Managing Overtime in the Workplace.<https://truein.com/how-to-manage-overtime/>

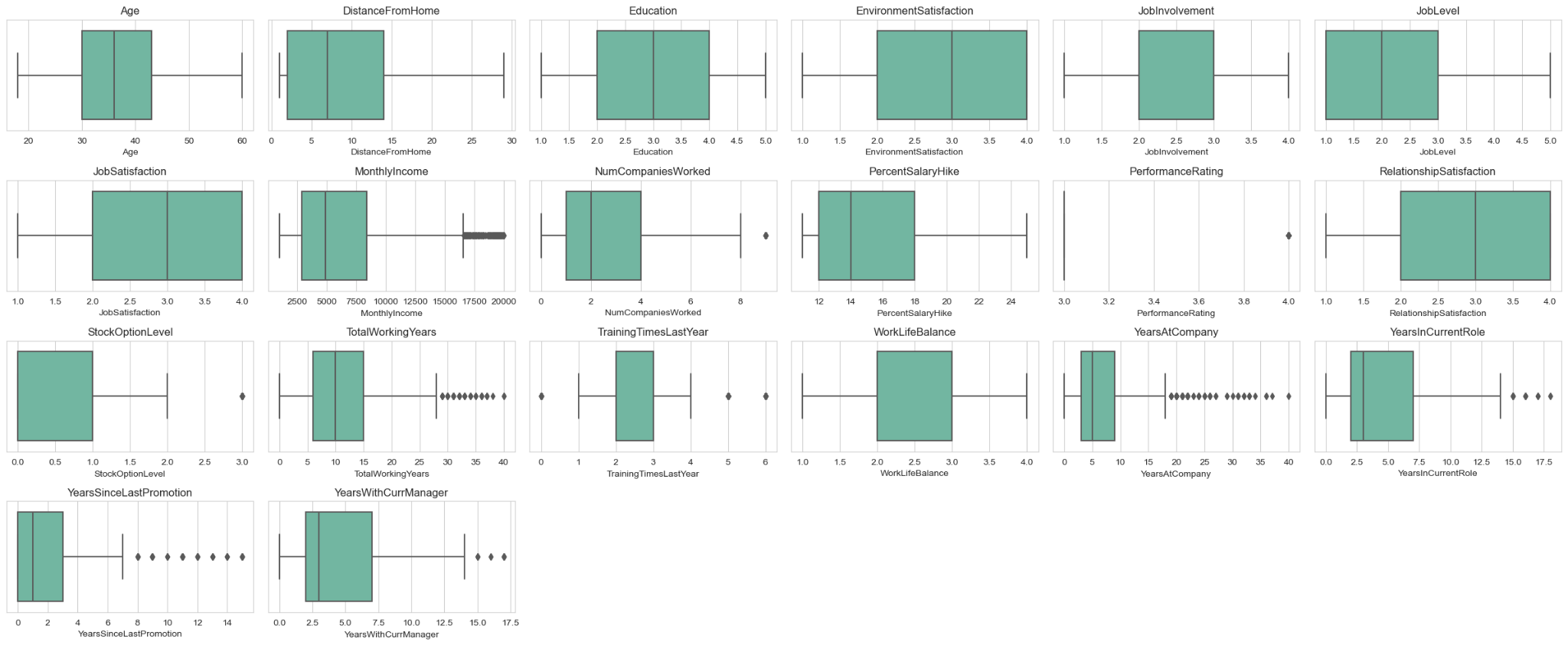
van Vulpen, E. (n.d.). 15 HR Analytics Case Studies with Business Impact. AIHR.<https://www.aihr.com/blog/hr-analytics-case-studies/>

Vouchercloud (n.d.). How Many Productive Hours in a Work Day? Just 2 Hours, 23 Minutes… <https://www.vouchercloud.com/resources/office-worker-productivity>

# **8. Appendices**

**Appendix A**

**Variables with Significant Outliers**

****

Based on the dataset, 9 columns have outliers:

|  |  |
| --- | --- |
| MonthlyIncome  NumCompaniesWorked  StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear | YearsAtCompany  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager |

**Appendix B**

Let x be a threshold for a feature variable X. E.g. X is monthly income and x = $4000.

If

Then

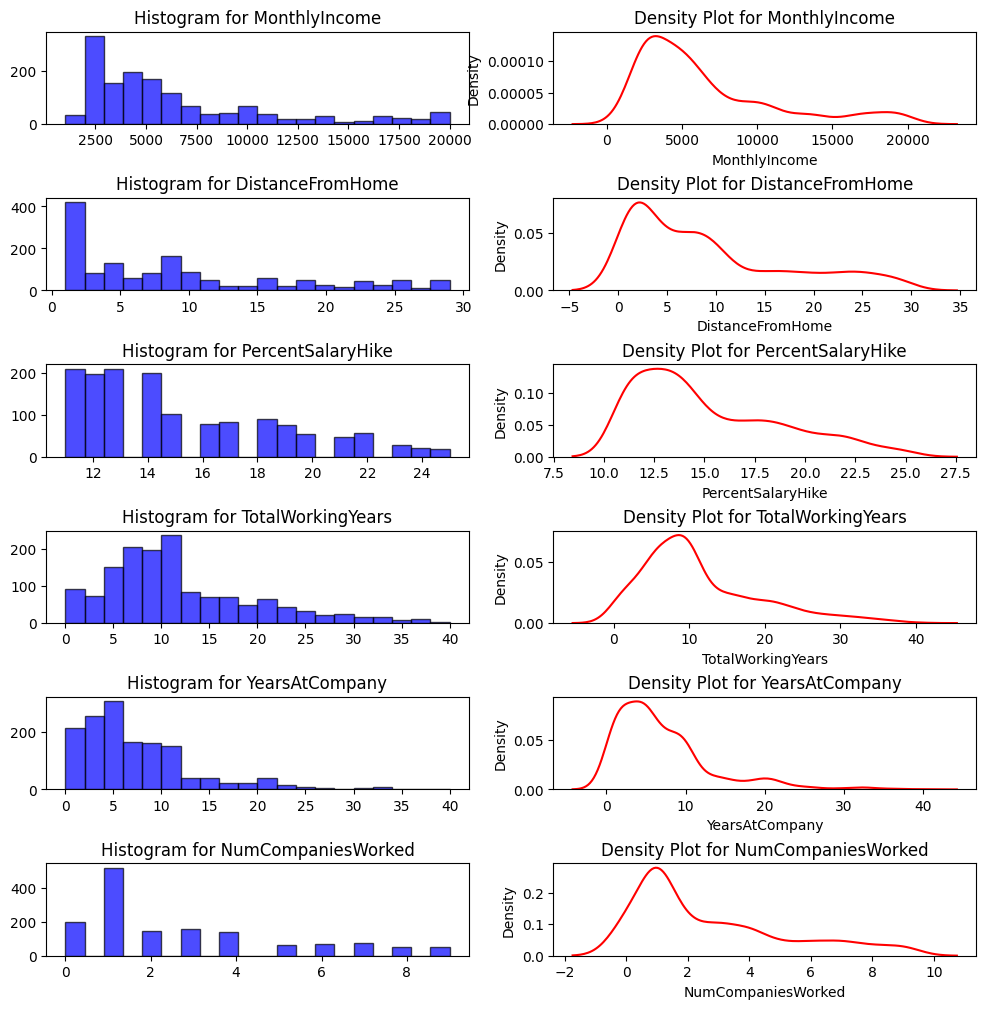
Hence, the given condition increases the probability that

In this case, we found that the conditional probability that an employee’s monthly income is less than $4000 given that the employee leaves is higher than that of when the employee stays. We can thus infer that the conditional probability that an employee leaves given that his monthly income is less than $4000 will be higher than the standalone probability of an employee leaving. As such, monthly income could potentially be an important factor in predicting employee attrition.

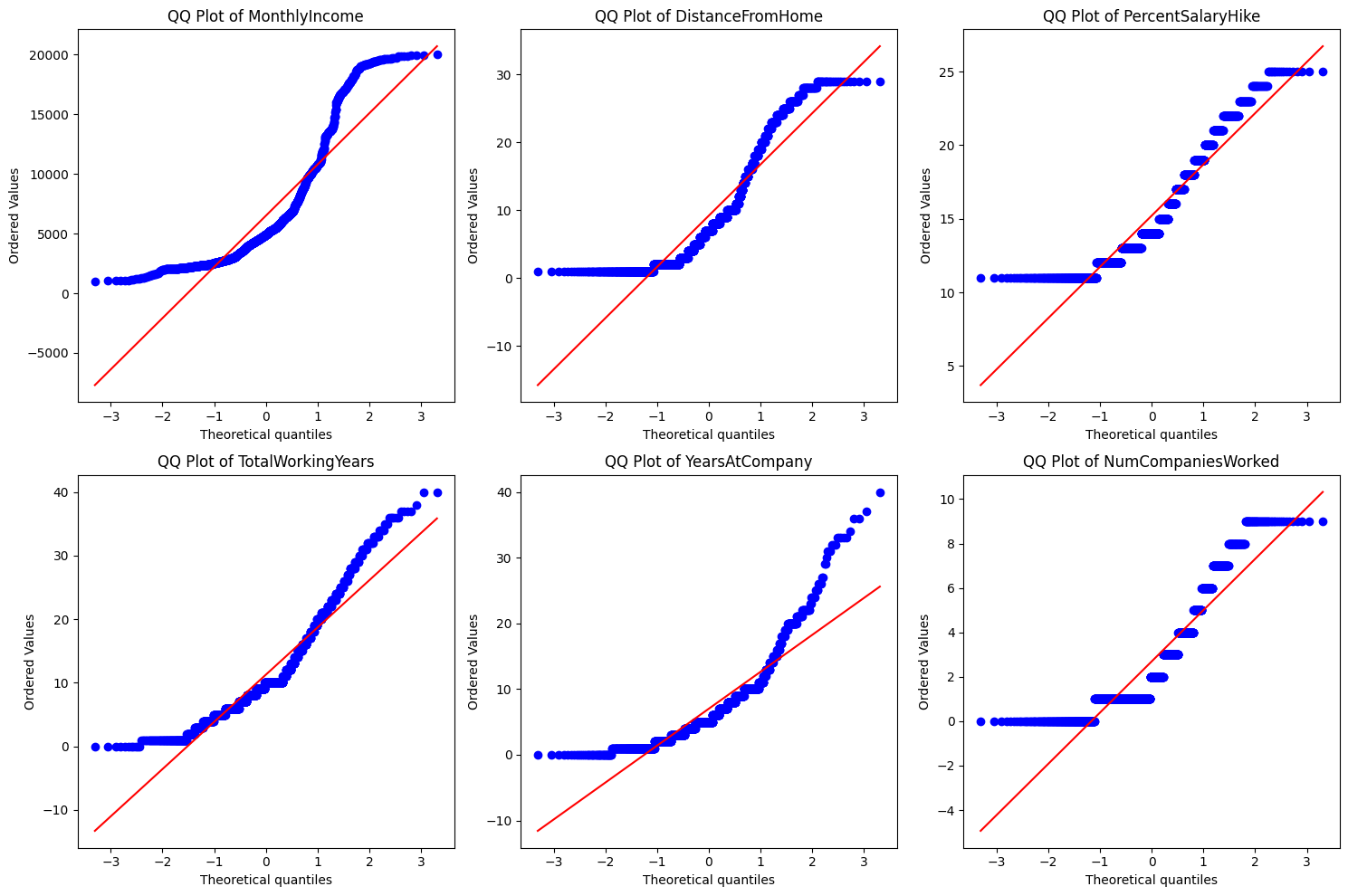
**Appendix C**

**Impact of Log and Square-root Transformation**

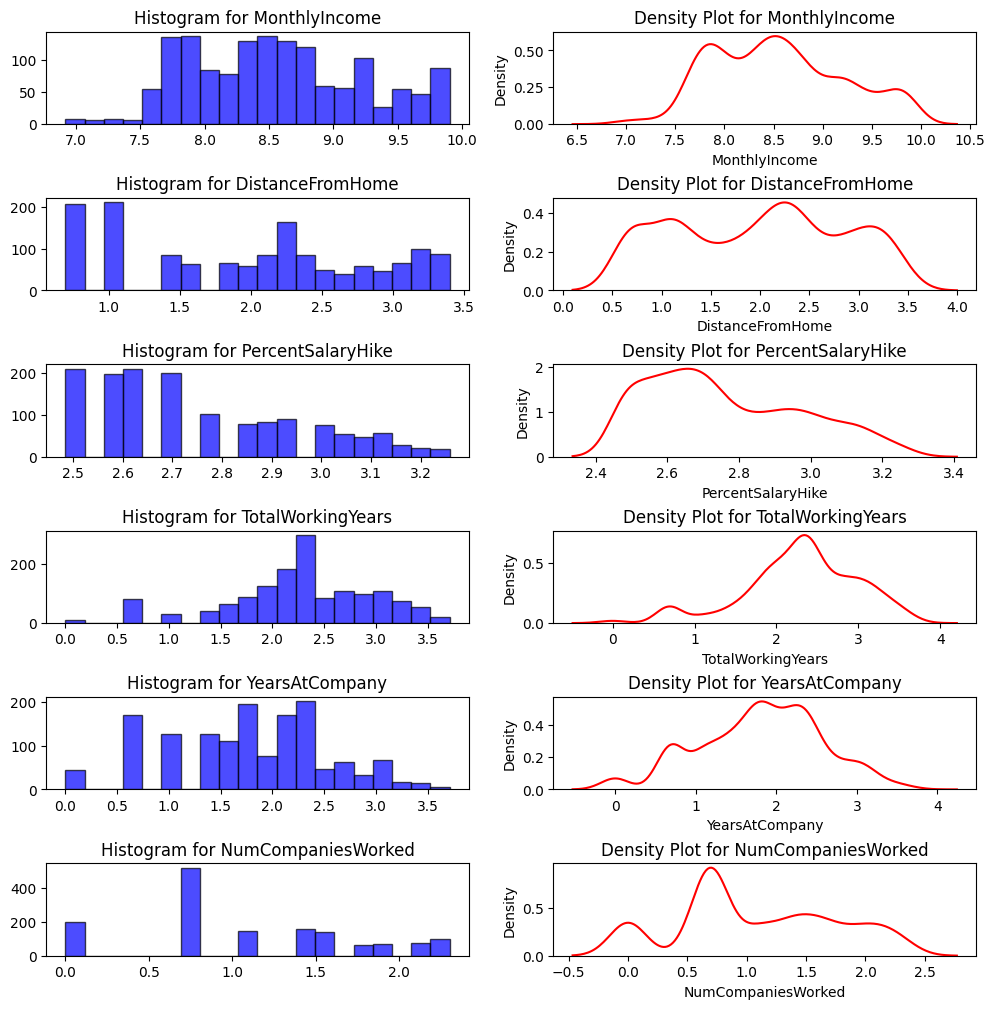
Histogram and Density Plots Before Feature Transformation

****

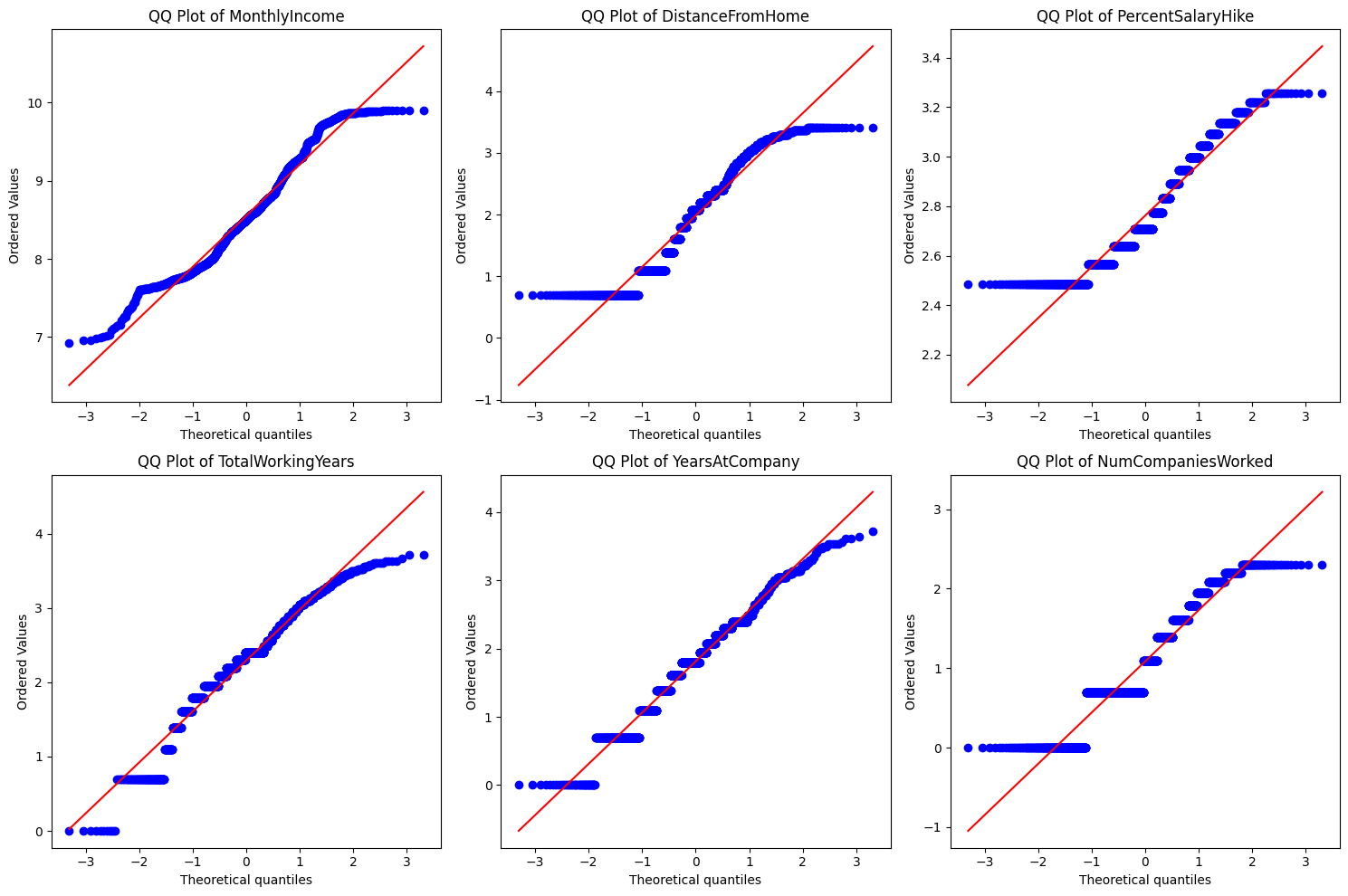
Q-Q Plots Before Feature Transformation

****

Histogram and Density Plots After Feature Transformation

****

Q-Q Plots After Feature Transformation

****

**Appendix D**

**Hyperparameter Optimisation**

**GridSearchCV** was used to perform an exhaustive search of specified parameter values to determine the most appropriate combination of values that results in the best performance for our dataset.

Logistic Regression

3 key parameters were calibrated:

1. The **“Penalty”** term was added to the model’s loss function to penalise large coefficient values and prevent the model from overfitting the train set data. It was set to L1 or L2 regularisation. **L1 regularisation** (lasso regularisation) adds the absolute values of coefficients as the penalty term, effectively performing feature selection by shrinking some coefficients to 0. Conversely, **L2 regularisation** (ridge regularisation) adds the squared magnitudes of the coefficients as the penalty term, allowing for smaller, non-zero coefficients for the features.
2. The **regularisation strength “C”** was configured to find the optimal trade-off between overfitting and underfitting the train set data. It was set to intervals of [0.1, 1, 10, 100]. Smaller values of “C” result in stronger regularisation, while larger values of “C” result in weaker regularisation.
3. The **“Scoring”** term was set to “recall” in order to determine the optimal combination of parameters that maximises the number of correctly predicted attrition cases **(true positives)** out of the actual attrition cases (total positive cases).

“Recall” was selected over “precision” as the consequences of incorrectly predicting employee retention when there is actual attrition (false negatives) are considered to be less severe than incorrectly predicting attrition when there is actual retention (false positives). Such misclassifications could potentially result in implementing strategies that are counterproductive towards reducing attrition.

CART - Classification Tree

The criteria for evaluating the purity of each split was optimised:

1. **Gini impurity** measures the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labelled according to the distribution of labels in the node. Gini impurity is minimised when all instances in the node belong to the same class, and it is maximised when the classes are evenly distributed.
2. **Entropy** measures the level of impurity or disorder in a set of examples. In the context of CART, it represents the amount of information required to classify the samples at a given node. Entropy is minimised when all instances in the node belong to the same class, and it is maximised when the classes are evenly distributed.

Random Forest

2 key parameters were calibrated:

1. **min\_samples\_leaf** is the minimum number of samples required to be at a leaf node of the decision tree. Setting min\_samples\_leaf to a higher value can help prevent overfitting by enforcing a minimum number of samples required for each leaf node. This controls the complexity of the tree by limiting the number of splits. A higher min\_samples\_leaf value results in simpler trees with fewer nodes, which can improve generalisation performance on unseen data but may lead to underfitting if set too high.
2. **max\_features** is the maximum number of features to consider when looking for the best split at each node. This parameter controls the randomness of feature selection in decision tree splits. It can be specified as an integer (representing the number of features to consider) or as a fraction of the total number of features (eg. “sqrt” for the square root of the total number of features). By limiting the number of features considered at each split, max\_features helps to reduce the model's tendency to overfit by introducing randomness and diversity in the tree building process. Using a smaller value for max\_features can lead to simpler trees and reduce the risk of overfitting, especially in high-dimensional feature spaces. However, setting it too low may result in underfitting if important features are excluded from consideration.

XGBoost

XGBoost requires the optimisation of many hyperparameters to determine the best stopping criteria that would minimise overfitting. Consequently, RandomSearchCV was used instead. Unlike GridSearchCV, which evaluates all possible combinations of hyperparameters within a specified grid, RandomSearchCV samples hyperparameter values randomly from specified distributions. A total of 7 parameters were optimised:

1. **n\_estimators** specify the number of decision trees to be built in the ensemble. Increasing n\_estimators generally improves model performance, but also increases computational time and risk of overfitting.
2. **learning\_rate** controls the contribution of each tree to the final prediction. A lower learning rate requires more trees to achieve the same level of performance but improves the generalisation ability of the model.
3. **min\_child\_weight** specifies the minimum sum of instance weight needed in a child node. It helps prevent overfitting by controlling the minimum number of instances required to split a node further, thus reducing the complexity of the trees.
4. **gamma** specifies the minimum reduction in the loss function required to make a further partition on a leaf node of the tree. Higher values of gamma lead to more conservative tree splits, which can help prevent overfitting.
5. **subsample** specifies the fraction of samples used to train each tree. Setting **subsample** to less than 1.0 introduces stochasticity into the training process, which helps prevent overfitting by adding randomness to the model.
6. **colsample\_bytree** specifies the fraction of features to be randomly sampled for each tree. Similar to subsample, it introduces randomness into the feature selection process and helps prevent overfitting by reducing the correlation between trees.
7. **max\_depth** specifies the maximum depth of each tree. Deeper trees can capture more complex relationships in the data but are more prone to overfitting. Setting a suitable max\_depth helps control the complexity of each tree.

**Appendix E**

**Formula Relating Precision, Recall, Cost Savings, and Losses for Tesla**

Let Recall = R, Precision = Pr, total number of actual positives = P, number of true positives = x, number of false positives = y.

Using the formula for Recall, we derive:

Similarly, using the formula for Precision, we derive:

Further simplifying the above, we derive:

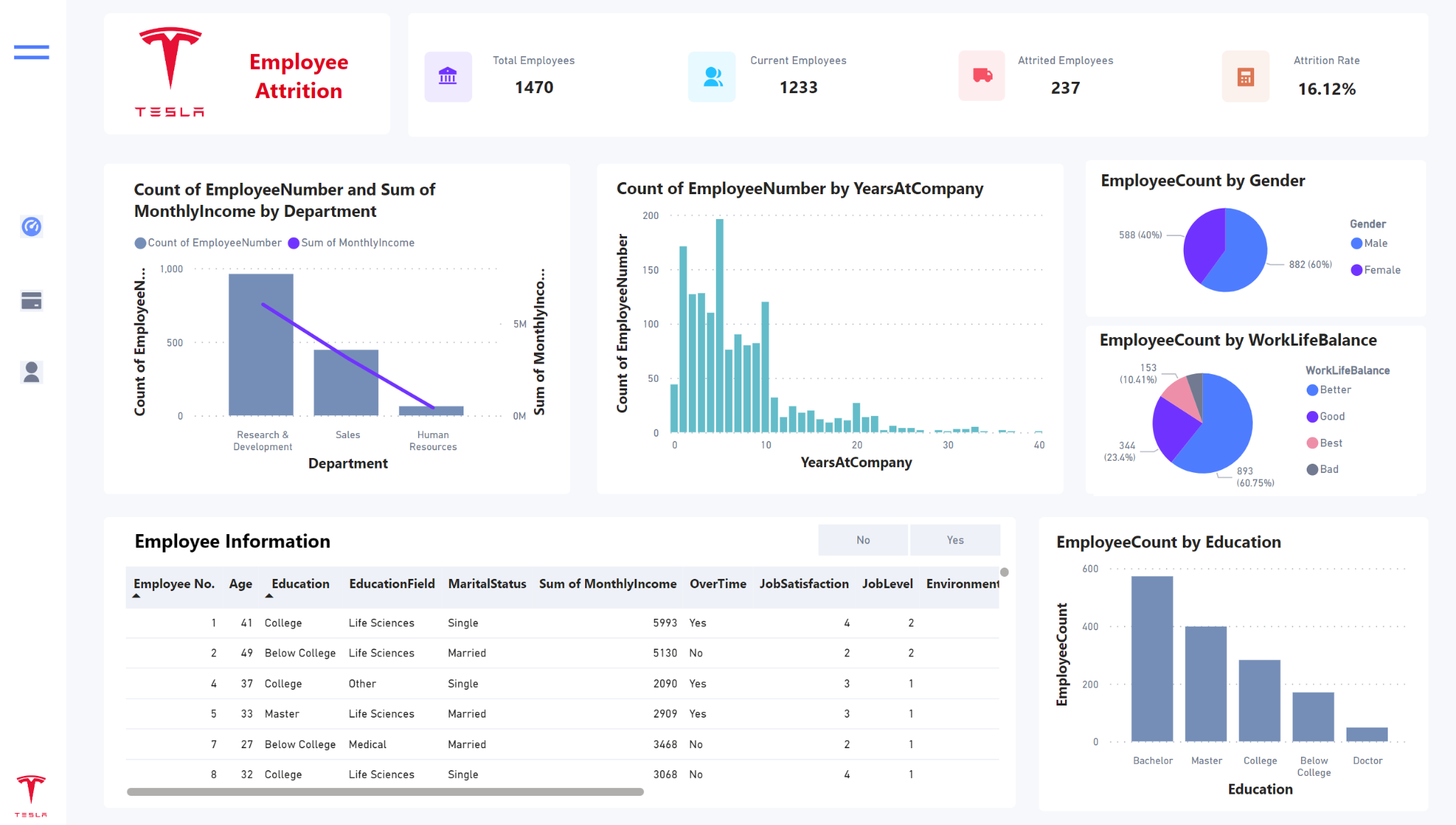
For a fixed P>0, maximum cost savings of $32500P is achieved with a perfect classifier that has R = Pr = 1 (i.e. true positives = total number of actual positives and false positives = 0)

Proof:

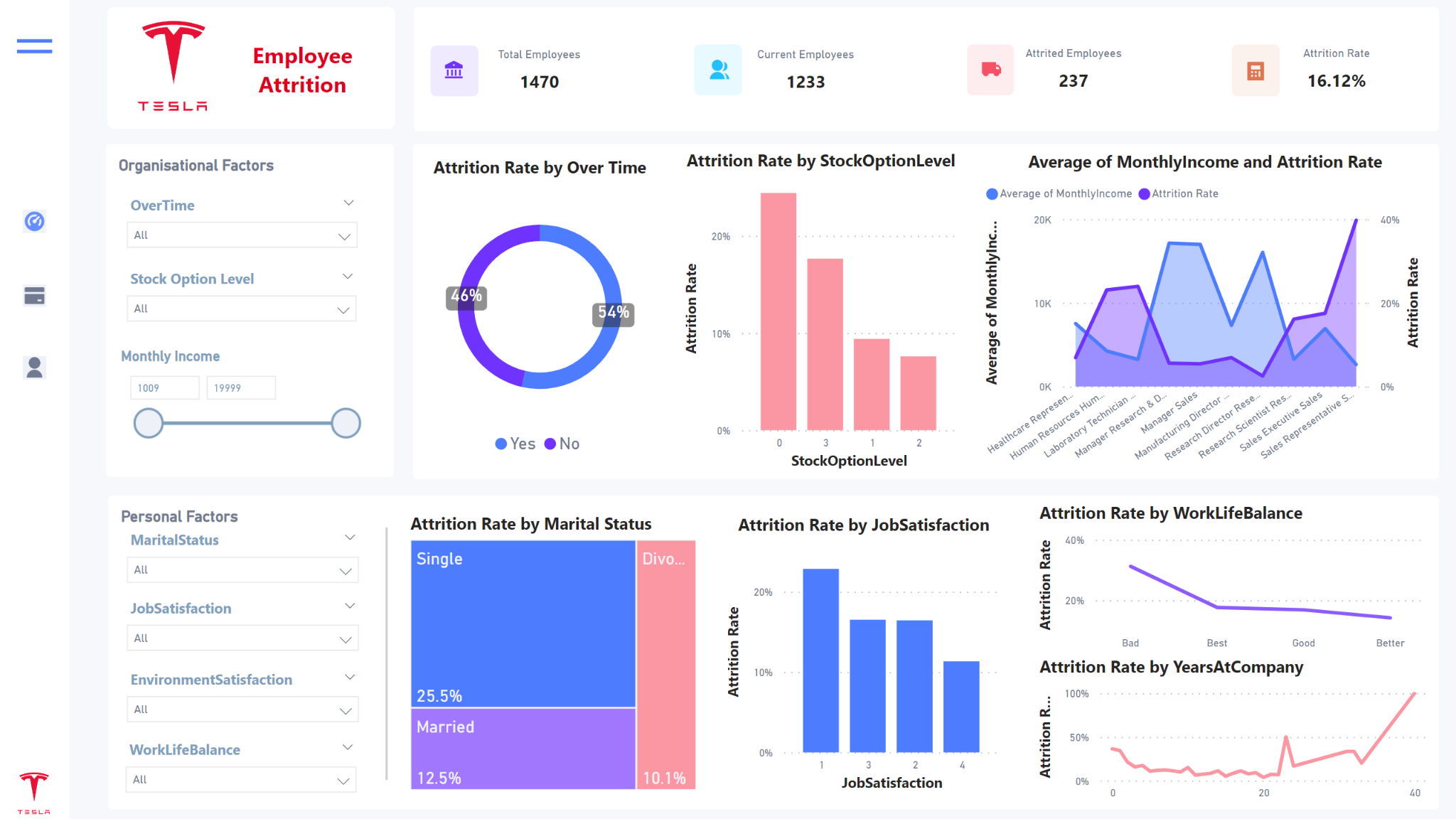
∴S = $32500P is global maximum, achieved when R = Pr = 1

**Appendix F**

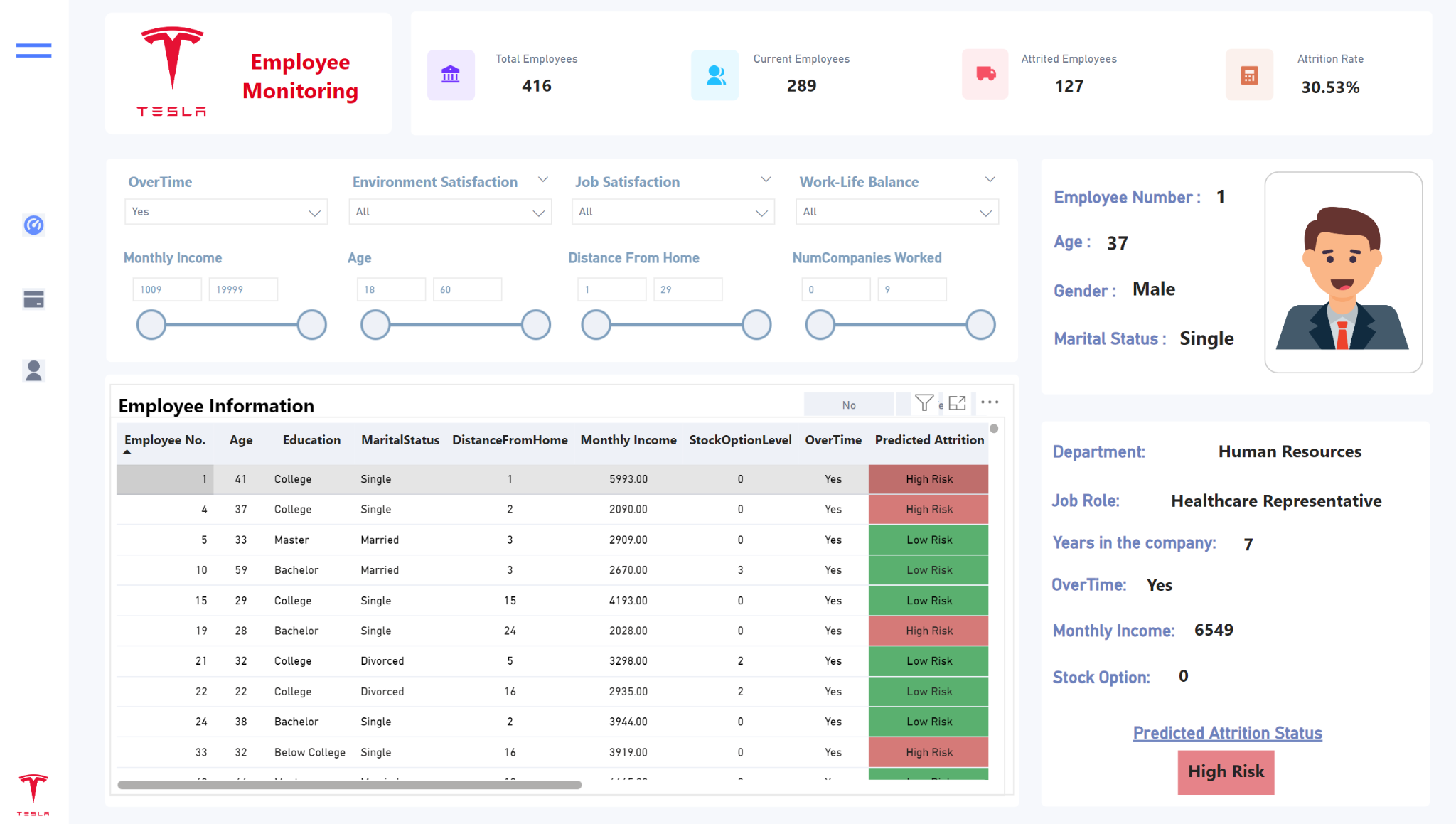
Dashboard View for Univariate Variable EDA

****

Dashboard View for Business Insights (Relationship between variables and attrition rate)

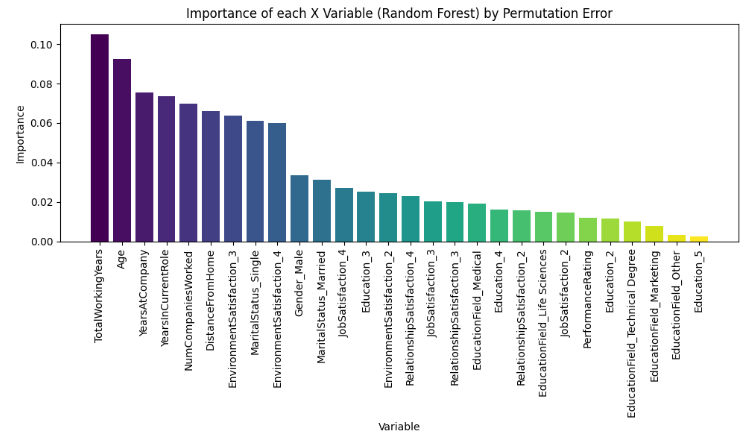
****

Dashboard View to filter Predicted High-Risk Employees (Preemptive)

****

**Appendix G**

Variable Importance Chart (Personal Factors)



Variable Importance Chart (Organisational Factors)

