Business Case Study: Driver Attrition Prediction Using Ensemble Learning

Problem Description

App-based taxi and auto-rickshaw services have revolutionized urban transportation, offering convenience, safety, and efficiency. By leveraging mobile technology, these services allow users to book rides with a few taps on their smartphones, eliminating the need to hail cabs on the street or wait at taxi stands. These platforms provide real-time tracking, fare estimation, and various payment options, enhancing the user experience.

They also offer multiple vehicle choices, from economical auto-rickshaws to premium sedans, catering to different customer preferences and budgets. Additionally, app-based services have improved the livelihood of drivers by providing them with a steady stream of passengers and flexible working hours. Integrating GPS and customer feedback mechanisms ensures better safety and service quality, making app-based taxi and auto-rickshaw services essential to modern urban mobility.

Recruiting and retaining drivers is indeed a significant challenge for these companies. The competition for drivers is fierce, with multiple platforms vying for a limited pool of qualified individuals. Drivers often switch between services based on incentives, bonuses, and overall earning potential, creating ongoing challenges for retention. Additionally, fluctuating fuel prices, maintenance costs, and the commission rates imposed by the platforms can impact drivers' earnings, leading to dissatisfaction and turnover.

Ride-hailing companies must balance offering attractive incentives and sustainable business models to address these issues. Providing better support, transparent policies, and benefits such as insurance and financial services can help improve driver loyalty. Ensuring a positive work environment and promptly addressing drivers' concerns are crucial strategies for retaining a reliable and motivated driver workforce. Losing drivers frequently impacts the organization's morale, and acquiring new drivers is more expensive than retaining existing ones.

This case study is focused on driver attrition prediction using ensemble learning. The monthly information for a segment of drivers for the 2019 and 2020 years are considered. The work predicts whether a driver will leave the company based on the attributes of the datasets. The following features are analyzed: Demographics (city, age, gender, etc.), Tenure information (joining date, Last Date), and Historical data regarding the performance of the driver (Quarterly rating, Monthly business acquired, grade, Income).

The dataset consists of 13 columns naming

1. MMMM-YY: Reporting Date (Monthly) in DD-MM-YYYY format
2. Driver_ID: Unique ID for drivers
3. Age: Age of the driver
4. Gender: Gender of the driver – Male: 0, Female: 1
5. City: City Code of the driver
6. Education_Level: Education level – 0 for 10+ ,1 for 12+ ,2 for graduate
7. Income: Monthly average Income of the driver
8. Date Of Joining: Joining date for the driver
9. LastWorkingDate: Last date of working for the driver
10. Joining Designation: Designation of the driver at the time of joining
11. Grade: Grade of the driver at the time of reporting
12. Total Business Value: The total business value acquired by the driver in a month (negative business indicates cancellation/refund or car EMI adjustments)
13. Quarterly Rating: Quarterly rating of the driver: 1,2,3,4,5 (higher is better)

Business Questions to be answered from Analysis

1. Which factor has the highest impact on driver retention?
2. What is the relationship between quarterly rating and driver retention?
3. How does income change affect driver retention?
4. Does gender play a significant role in driver retention?
5. Which age group has the highest driver retention rate?
6. What is the impact of joining designation on driver retention?
7. Which machine learning model performs best in predicting driver retention?
8. What is the F1 score for the best-performing model?
9. How can companies use ROC curves to improve driver retention?

10. What are some actionable insights that companies can implement to improve driver retention?

Companies in the ride-hailing industry rely heavily on understanding the reasons behind driver churn (drivers leaving the company). They can gain valuable insights by analyzing factors like performance reviews, compensation changes, demographics, and the effectiveness of different machine learning models. These insights can then be used to develop targeted retention strategies that address the specific needs of varying driver groups. Ultimately, the goal is to obtain actionable recommendations and optimize identifying drivers at risk of leaving, leading to a more stable driver workforce and improved business performance.

Methodology

The proposed work analyses the given dataset to detect the most influential features that affect driver attrition prediction and builds a predictive model according to the following phases.

1. Exploratory Data Analysis

To analyze a dataset, start by examining its structure and characteristics, including data types and column names. Next, identify and address any missing values. Perform univariate analysis to understand the distribution and key statistics of individual variables. Then, conduct bivariate analysis to explore relationships between pairs of variables. Assess the correlation among independent variables to see how they interact with each other. Finally, provide a statistical summary of the dataset, highlighting key measures such as mean, median, and standard deviation after any necessary transformations or cleaning.

2. Data Preprocessing

   To prepare the dataset, start by aggregating data to eliminate duplicate driver entries. Use KNN imputation to handle missing values. In feature engineering, create columns to indicate if a driver’s quarterly rating or monthly income has increased (assigning a value of 1 if true else 0) and a target variable to show if a driver has left the company (value of 1 if the last working day is present else 0). Address class imbalance issues, standardize the data, and apply one-hot encoding to categorical variables for better model performance.

3. Model building

For model building, implement one ensemble method using a bagging algorithm, such as Random Forest, to reduce variance and improve accuracy by combining multiple models. Additionally, use a boosting algorithm, like Gradient Boosting, to sequentially build models that correct errors from previous ones, enhancing overall performance and robustness. This combination leverages the strengths of both bagging and boosting to create a more effective predictive model.

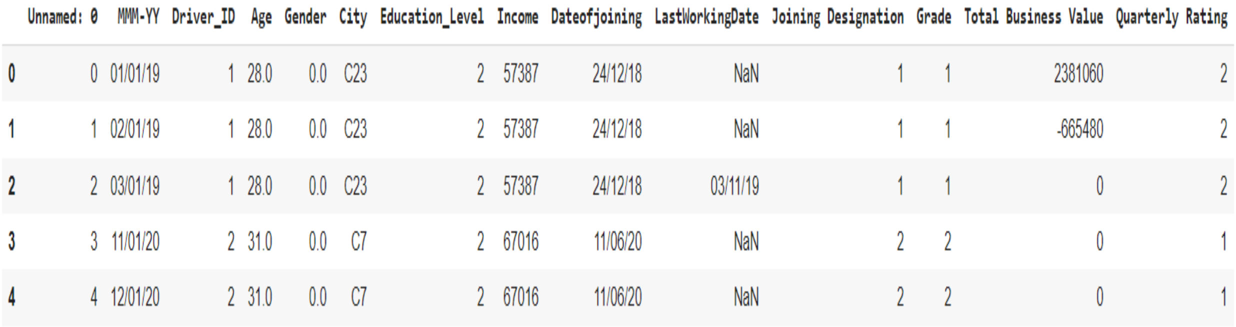
4. Results Evaluation

For results evaluation, use the ROC AUC curve to measure the model's ability to distinguish between classes, providing a single metric for performance. Additionally, generate a classification report, including the confusion matrix, to detail the model's precision, recall, F1-score, and accuracy, offering a comprehensive view of its effectiveness in predicting each class.

Analysis

1. Exploratory Data Analysis

The business case study analyses the given dataset of the driver's retention-related information for the 2019 and 2020 years for driver attrition prediction using ensemble learning. The work predicts whether a driver will leave the company based on the attributes of the datasets. Figure 1.1 displays the five samples of the dataset.



	Unnamed: 0	MMM-YY	Driver_ID	Age	Gender	City	Education_Level	Income	Dateofjoining	LastWorkingDate	Joining Designation	Grade	Total Business Value	Quarterly Rating
0	0	01/01/19	1	28.0	0.0	C23	2	57387	24/12/18	NaN	1	1	2381060	2
1	1	02/01/19	1	28.0	0.0	C23	2	57387	24/12/18	NaN	1	1	-665480	2
2	2	03/01/19	1	28.0	0.0	C23	2	57387	24/12/18	03/11/19	1	1	0	2
3	3	11/01/20	2	31.0	0.0	C7	2	67016	11/06/20	NaN	2	2	0	1
4	4	12/01/20	2	31.0	0.0	C7	2	67016	11/06/20	NaN	2	2	0	1

Figure 1.1: Sample Data of the given dataset

The dataset consists of 19104 rows and 14 columns. The columns of the dataset are: - Unnamed: 0 , MMM-YY, Driver_ID, Age, Gender, City, Education_Level, Income, Dateofjoining, LastWorkingDate, Joining Designation, Grade, Total Business Value, Quarterly Rating. Out of these columns, four columns named MMM-YY, City, Dateofjoining, and LastWorkingDate are categorial features, the remaining nine columns are numerical features, and the first column is the individual row number, not used for the analysis. The unique number of values for the numerical and the categorical features is given in Tables 1.1 and 1.2 respectively.

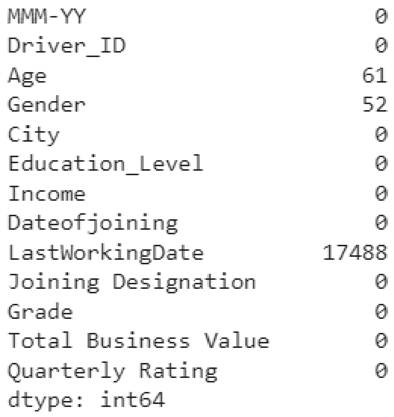
Table 1.1: Numerical Attributes and its number of unique values

Number	Numerical Attributes	Number of Unique Values
1	Driver_ID	2381
2	Age	36
3	Gender	2
4	Education_Level	3
5	Income	2383
6	Joining Designation	5
7	Grade	5
8	Total Business Value	10181
9	Quarterly Rating	4

The data samples were collected monthly for each driver until they left the company and the “LastWorkingDate” column filled only on the month the driver had left the company. The missing value analysis shown in Figure 1.2 also supports these factors and it summarizes around 92 % of “LastWorkingDate” are null values. Therefore, the dataset has to be grouped by driver_ID before starting the modeling.

Table 1.2: Categorical Attributes and its number of unique values

Number	Numerical Attributes	Number of Unique Values
1	MMM-YY (Reporting Date)	24
2	City	29
3	Dateofjoining	869
4	LastWorkingDate	493



```
MMM-YY                      0
Driver_ID                   0
Age                        61
Gender                     52
City                        0
Education_Level             0
Income                      0
Dateofjoining               0
LastWorkingDate         17488
Joining Designation         0
Grade                       0
Total Business Value        0
Quarterly Rating            0
dtype: int64
```

Figure 1.2: The missing value details of the dataset

The correlation heatmap in Figure 1.3 visualizes the Spearman correlation coefficients between numerical features. It reveals positive correlations exist between "Income" and "Grade," as well as between "Total Business Value" and "Quarterly Rating." These correlations hint at underlying connections between these variables, where higher income might correspond to better grades and higher business value might be associated with stronger quarterly ratings. There are no features with a strong negative correlation. Overall, the heatmap provides valuable insights into potential relationships within the data.

Univariate Analysis

The boxplots, visualized in Figure 1.4, reveal some interesting insights about the numerical data. Outliers are present in columns such as Age, Income, Joining Designation, and Total Business
Value. These outliers represent extreme values that might warrant further investigation. Additionally, most of the data appears skewed, meaning it's not evenly distributed. This skewness suggests that the median, represented by the line inside the box, is a better measure of central tendency to the mean. By comparing the medians across different columns, as shown in the boxplots, we can identify significant differences. The size of the boxes (IQR) reflects the data's variability, with larger boxes indicating a wider spread of values. Overall, the boxplots offer a valuable snapshot of the data's distribution, highlighting potential issues like outliers and skewness, and guiding further analysis.

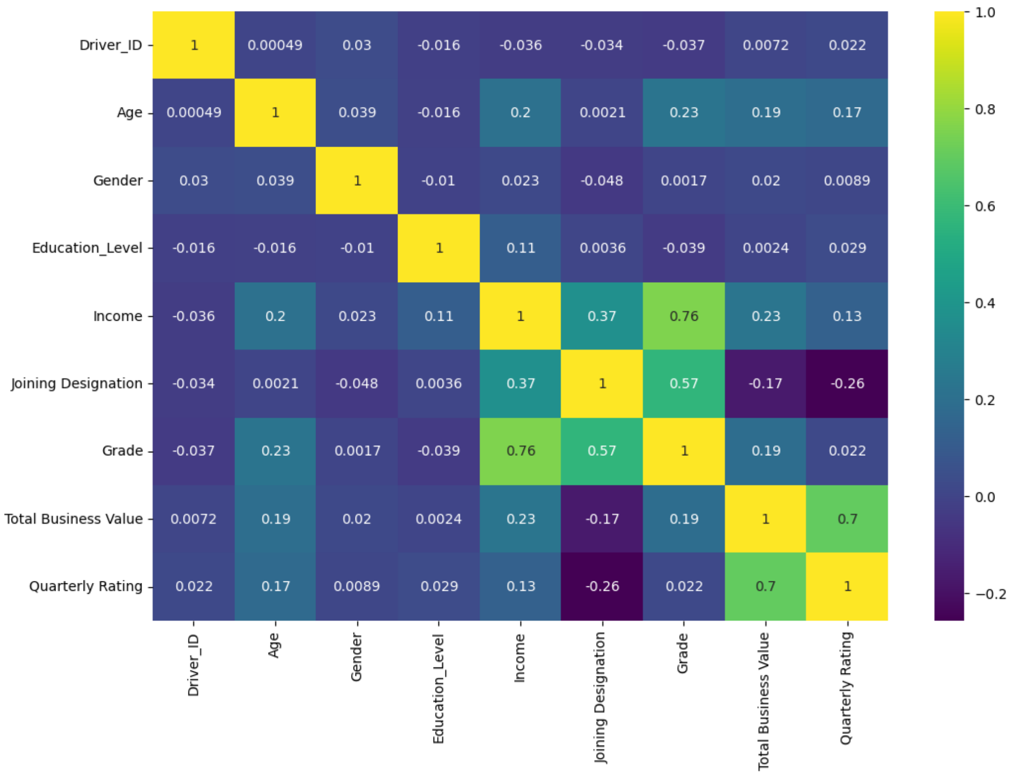


Figure 1.3: Correlation Heatmap of Numerical Attributes

Taking a closer look at numerical features with outliner, through the histograms in Figure 1.5, we see interesting patterns in their distributions. Income exhibits a right skew, meaning a larger portion of employees fall within the lower income range (between 20k and 80k) compared to those with higher incomes. This is visually confirmed by the histogram's shape, with a longer tail extending towards the higher income side. However, there are a few outliers exceeding 100k. Age distribution appears more balanced, resembling a bell curve in the histogram.

Most employees fall between 25 and 45 years old, with a few outliers younger than 20 or older than 60. Total Business Value also displays a right skew, with the majority concentrated between 0 and 2 million as evident from the histogram's peak. Similar to income, there are outliers with Total Business Values exceeding 5 million. These insights highlight potential variations within the workforce and can be valuable for further analysis. Figure 1.4 and Figure 1.5 show that the joining designation has only a few outliers at 4 and 5. About 98% of drivers have joining designation as 1, 2, or 3.

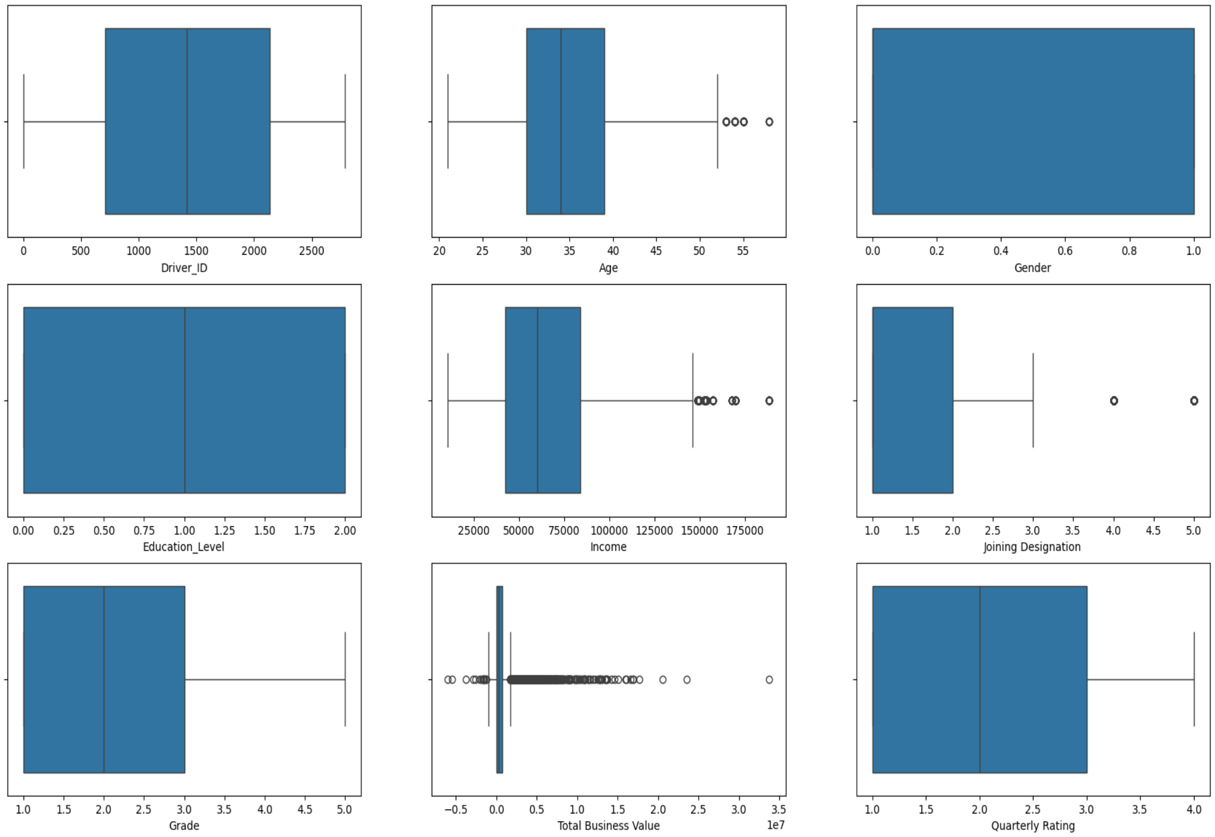


Figure 1.4: Boxplot for Numerical Attributes

Data Preprocessing:

After analyzing the missing values and outliers, we applied the following transformations to the raw data to prepare it for modeling.

1. **Date Wrangling:** Columns representing dates ('MMM-YY', 'Dateofjoining', and 'LastWorkingDate') were converted into a consistent datetime format for further analysis.
2. **City Column Encoding:** The City column was encoded using labelEncoder function and converted the value in the column to numerical values.

3. **Extracting Driver-Specific Information with Aggregation:** The given dataset was grouped based on the 'Driver_ID' column and then used aggregation functions to extract specific information for each driver. The breakdown of the aggregation applied to each column:

   - **Age, Gender, Education_Level, Income, Dateofjoining, LastWorkingDate, Grade, Quarterly Rating, and City_enc:** These columns likely represent attributes that might change over time for each driver. The 'last' aggregation ensures we retain the most recent value for these attributes.

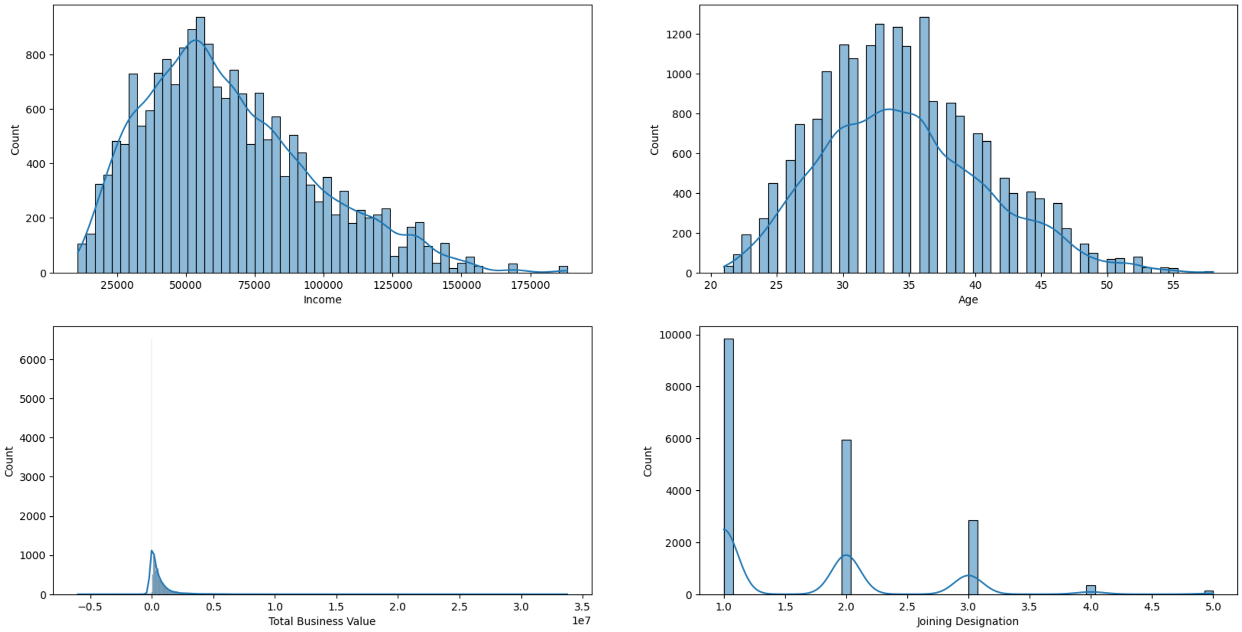


Figure 1.5: Histogram plot for Income, Age, and Total Business Value

- **Joining Designation:** This column might represent the designation assigned during a driver's initial joining. The 'first' aggregation keeps the value assigned at the first record for each driver.
- **Total Business Value:** This column likely represents a numerical value associated with each driver's business contribution. The 'mean' aggregation calculates the average value across all records for each driver.

4. **Total Working Day Calculation:** A new column titled "total_working_days" was created by subtracting the 'Dateofjoining' from 'LastWorkingDate'. The total number of days worked for each driver was computed missing values were filled with 0.

5. **Target Variable Creation:** A binary target variable named 'Target' was created to indicate driver retention. It assigns a 0 to drivers who have left the company and 1 for retention.

6. **Performance Change:** The difference between the first and last quarterly ratings was calculated for each driver. Another binary variable was created to capture whether the rating improved (1) or not (0).

7. **Income Change:** A similar approach was used for income. The difference between the first and last income values was determined, and a binary variable was generated to indicate an increase in income (1) or not (0).

8. **Data Cleaning:** Unnecessary columns ('MMM-YY', 'Dateofjoining', 'LastWorkingDate', and 'Total Business Value') were removed. Additionally, for drivers with multiple entries, only the last row was retained to ensure all relevant information was consolidated into a single record.

9. **Working Day Column Encoding:** The "total_working_days" column was encoded using labelEncoder function and converted the value in the column to numerical values.

10. **Dropping date columns:** The date columns "Dateofjoining" and "LastWorkingDate" are dropped in the dataframe.

11. **Label Encoding of "total_working_days"**: Applied label encoding to the "total_working_days" column and converted the time delta values representing the total working days into numerical labels. The count plot of Gender, Education Level, Joining Designation, Grade and Quarterly Rating consisting of Target is plotted in Figure 1.6. The proportion of retained drivers (Target = 1) is slightly higher for males than females.

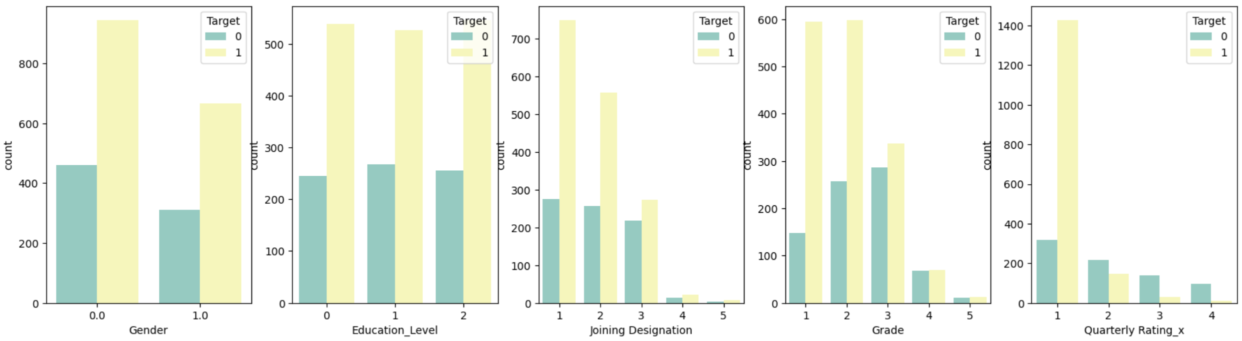


Figure 1.6: The count plot of Gender, Education Level, Joining Designation, Grade and Quarterly Rating consisting of Target

12. **Standardization**: Used the StandardScaler from sci-kit-learn to standardize the numerical features in the DataFrame. StandardScaler scales the data to unit variance and 0 mean. This helps improve the performance of many machine learning algorithms, particularly those sensitive to the scale of the input features.

13. **KNN Imputation:** Used the KNNImputer from sci-kit-learn to impute missing values in the DataFrame. KNNImputer uses the k-nearest neighbors algorithm to estimate the missing values based on the non-missing values in the data frame. The number of neighbors used for imputation was set to 9.

14. **Target Variable:** Extracted the "Target" column from the imputed DataFrame and stored it in a separate variable. The lambda function was applied to the "Target" column to convert values greater than 0 to 1 and 0 otherwise. This ensures that the target variable is binary, with 1 representing retained drivers and 0 representing drivers who left.

15. **Class Distribution:** Counted the number of occurrences of each class in the target variable. There is a wide difference in the distribution of the target classes 0 and 1 with 0 having count 769 and 1 having count 1612.
16. **Class Imbalance Treatment:** The dataset is imbalanced, with a significantly lesser number of drivers who left compared to those who remained. To address this issue, the code applies Synthetic Minority Over-sampling Technique (SMOTE) to the training data. The data is initially split into training and testing sets using `train_test_split`. This ensures that the model is trained on a separate set of data from the one used for evaluation. The `SMOTE` object is used to oversample the minority class (drivers who remained) in the training data. This technique creates synthetic samples of the minority class, effectively balancing the class distribution. The resampled dataset is stored in `X_sm` and `y_sm`. The class distribution is rechecked to confirm that the classes are now balanced. The balanced training data can be used to train a machine-learning model for predicting driver retention.

Model building

1 Ensemble - Bagging Algorithm

Two models considered were the decision tree model and the random forest classifier. A decision tree model with a maximum depth of 4 and a random state of 7 was trained and evaluated using 5-fold cross-validation. The best parameters and their corresponding mean score and rank were printed. The K-Fold Accuracy of the model had 100 mean and 0 standard deviations for both the train and validation data. A random forest model with 100 estimators and a maximum depth of 4 was trained and evaluated using 5-fold cross-validation. The average training and validation accuracy were calculated. The K-Fold Accuracy of the model had 100 mean and 0 standard deviations for both the train and validation data.

The feature importances were calculated and plotted in Figure 1.7 to identify the most important features for predicting driver retention. The feature importance plot shows the relative importance of each feature in the Random Forest model for predicting driver retention. The “total_working_days” feature has the highest importance with a 0.661 value, indicating that the duration of employment is the most relevant feature to predict driver retention. The next feature was “Quarterly_rating” with 0.152, while all other features had lower values.

Figure 1.7: Feature Importance for Driver Retention Prediction

Figure 1.8 shows the confusion matrix for a Random Forest model trained on the balanced training data. The confusion matrix visualizes the performance of the model in predicting driver retention. The diagonal elements of the matrix show the number of correctly classified instances for each class. The off-diagonal elements show the number of misclassified instances. The model accurately classified 393 instances as retained drivers and 191 instances as drivers who left. The F1 Score for the Bagging Algorithm is 1.

Figure 1.9 shows the Receiver Operating Characteristic (ROC) curve for a Random Forest model trained on the balanced training data. The ROC curve visualizes the performance of the model in predicting driver retention at different classification thresholds. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). The model achieves a True Positive Rate of approximately 1 at a False Positive Rate of 0. This indicates that the model correctly identifies 100 % of the drivers who left the company.

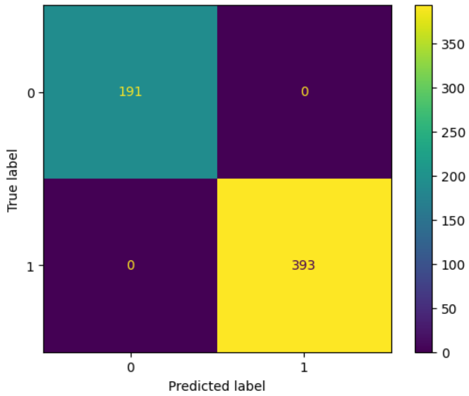


Figure 1.8: Confusion Matrix for Random Forest Model

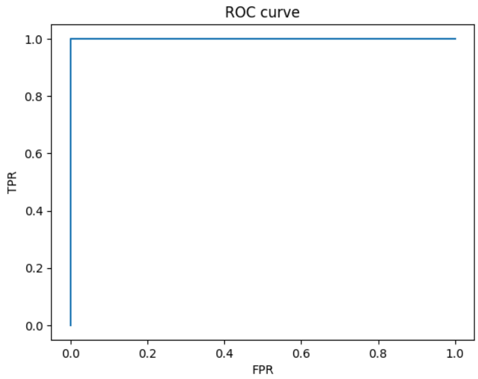


Figure 1.9: ROC Curve for Random Forest Model

1 Ensemble - Boosting Algorithm

A gradient-boosting model with 150 estimators, a learning rate of 0.3, and a maximum depth of 4 was trained. The training and test accuracy were calculated as 1. The loss curve was plotted in Figure 1.10 visualized the model's learning process. The loss function measures how well the model fits the training data. The x-axis represents the iteration number, while the y-axis represents the loss value. The plot shows that the loss function decreases as the number of iterations increases. This indicates that the model is learning and improving its fit to the training data. The plot can be used to assess the convergence of the training process.

Figure 1.11 shows the confusion matrix for a gradient-boosting model trained on the balanced training data. The confusion matrix visualizes the performance of the model in predicting driver retention. The diagonal elements of the matrix show the number of correctly classified instances for each class. The model accurately classified 393 instances as retained drivers and 191 instances as drivers who left. The F1 Score for the Bagging Algorithm is 1.

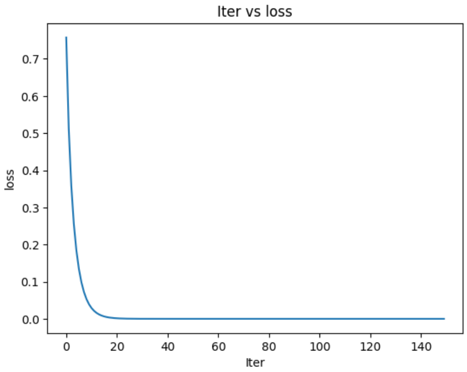


Figure 1.10: Loss Function Values for Gradient Boosting Classifier

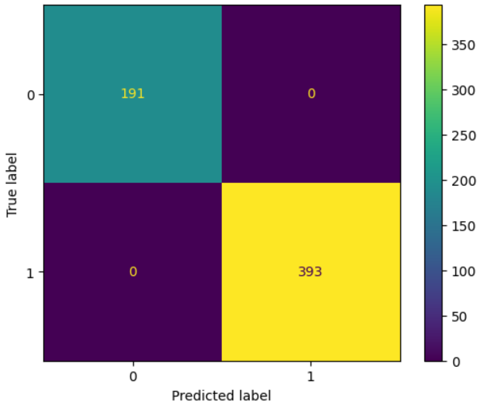


Figure 1.11: Confusion Matrix for Gradient Boosting Model

Figure 1.12 shows the Receiver Operating Characteristic (ROC) curve for a gradient-boosting model trained on the balanced training data. The ROC curve visualizes the performance of the model in predicting driver retention at different classification thresholds. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). The model achieves a True Positive Rate of approximately 1 at a False Positive Rate of 0. This indicates that the model correctly identifies 100 % of the drivers who left the company.

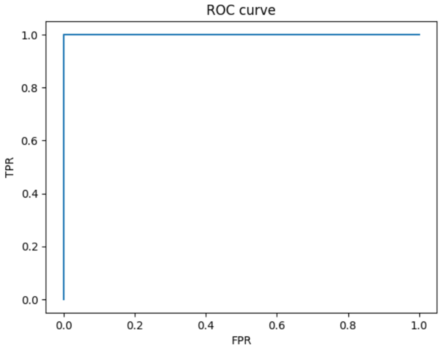


Figure 1.12: ROC Curve for Gradient Boosting Model

XGB Classifier Model

An XGBClassifier model with 100 estimators was trained using a randomized search with cross-validation to find the optimal hyperparameters for maximizing accuracy. The training and test accuracy were calculated. Using this model also, the train and test accuracy was 1.

Actionable Insights

1. High Importance of Total Working Days:
   - The “total working days” feature has the highest importance with a 0.661 value while calculating the relative importance of each feature using the Random Forest model, indicating that the duration of employment is the most relevant feature to predict driver retention.
   - Companies should focus on creating a positive work environment and providing opportunities for growth and development to increase employee retention.

2. Importance of Quarterly Rating and Income:
   - The next feature was Quarterly_rating” with a 0.152 feature importance value.
   - Quarterly rating and income changes are also important factors influencing driver retention.
   - Companies should implement performance management systems that provide regular feedback and opportunities for improvement.
   - Competitive compensation and benefits packages can help attract and retain top talent.

3. Age and Joining Designation:
   - Age and joining designation also play a role in driver retention.
   - Companies should consider implementing targeted retention strategies for different age groups and job roles.

4. Effectiveness of Bagging and Boosting Algorithms:
   - Both bagging and boosting algorithms achieved high accuracy in predicting driver retention.
   - These models can be further optimized and used to identify drivers at risk of leaving the company.

- F1 Score in both bagging and Boosting Algorithms is 1

5. Utilizing ROC Curves:
   - ROC curves visually represent the model's performance at different classification thresholds.
   - Companies can use ROC curves to select the appropriate threshold for their needs, balancing the trade-off between true and false positives.

Recommendations:

1. Focus on Employee Engagement:
   - Implement initiatives to improve employee engagement and satisfaction, such as regular feedback sessions, recognition programs, and professional development opportunities.

2. Competitive Compensation and Benefits:
   - Offer competitive salaries, benefits, and incentives to attract and retain top talent.
   - Regularly review and adjust compensation packages to stay competitive in the market.
   - The company should take action to motivate the drivers to work more days such as providing additional incentives etc.

3. Targeted Retention Strategies:
   - Develop targeted retention strategies for different age groups and job roles based on their unique needs and concerns.
   - Provide personalized support and resources to employees at higher risk of leaving the company.

4. Performance Management and Development:
   - Implement a robust performance management system that provides regular feedback and opportunities for improvement.
   - Invest in employee development programs to help drivers acquire new skills, advance their careers, and improve their quarterly ratings.

5. Data-Driven Insights:
    - Utilize machine learning models and data analysis to identify drivers at risk of leaving the company.
    - Use these insights to address potential issues and implement targeted retention strategies proactively.

Business Questions Answered from Analysis

1. Which factor has the highest impact on driver retention?
   Answer: Total working days.

2. What is the relationship between quarterly rating and driver retention?
   Answer: Drivers with higher quarterly ratings are more inclined to stay with the company.

3. How does income change affect driver retention?
   Answer: Drivers who experience a higher income are more likely to remain with the company.

4. Does gender play a significant role in driver retention?
   Answer: No, gender does not have a significant impact on driver retention.

5. Which age group has the highest driver retention rate?
   Answer: The analysis does not provide information about age groups and retention rates.

6. What is the impact of joining designation on driver retention?
   Answer: Drivers with higher joining designations are more likely to stay with the company.

7. Which machine learning model performs best in predicting driver retention?
   Answer: Both bagging and boosting algorithms achieve high accuracy in predicting driver retention.

8. What is the F1 score for the best-performing model?

   Answer: The F1 score for both bagging and boosting algorithms is 1.

9. How can companies use ROC curves to improve driver retention?

   Answer: ROC curves can help companies select the appropriate threshold for identifying drivers at risk of leaving the company.

10. What are some actionable insights that companies can implement to improve driver retention?

    Answer: Companies can focus on improving employee engagement, offering competitive compensation and benefits, developing targeted retention strategies, implementing performance management systems, and utilizing data-driven insights to identify drivers at risk of leaving the company.