OLA

DATA ANALYSIS REPORT

ANIL

INTRODUCTION

Ola is India's largest mobility platform and one of the world's largest ride-hailing companies, serving 250+ cities across India, Australia, New Zealand, and the UK. The Ola app offers mobility solutions by connecting customers to drivers and a wide range of vehicles across bikes, auto-rickshaws, metered taxis, and cabs, enabling convenience and transparency for hundreds of millions of consumers and over 1.5 million driver-partners.

Recruiting and retaining drivers is seen by industry watchers as a tough battle for Ola. Churn among drivers is high and it's very easy for drivers to stop working for the service on the fly or jump to Uber depending on the rates. As the companies get bigger, the high churn could become a bigger problem. To find new drivers, Ola is casting a wide net, including people who don't have cars for jobs. But this acquisition is really costly. Losing drivers frequently impacts the morale of the organization and acquiring new drivers is more expensive than retaining existing ones. You are working as a data scientist with the Analytics Department of Ola, focused on driver team attrition. You are provided with the monthly information for a segment of drivers for 2019 and 2020 and tasked to predict whether a driver will be leaving the company or not based on their attributes

Understanding the Dataset

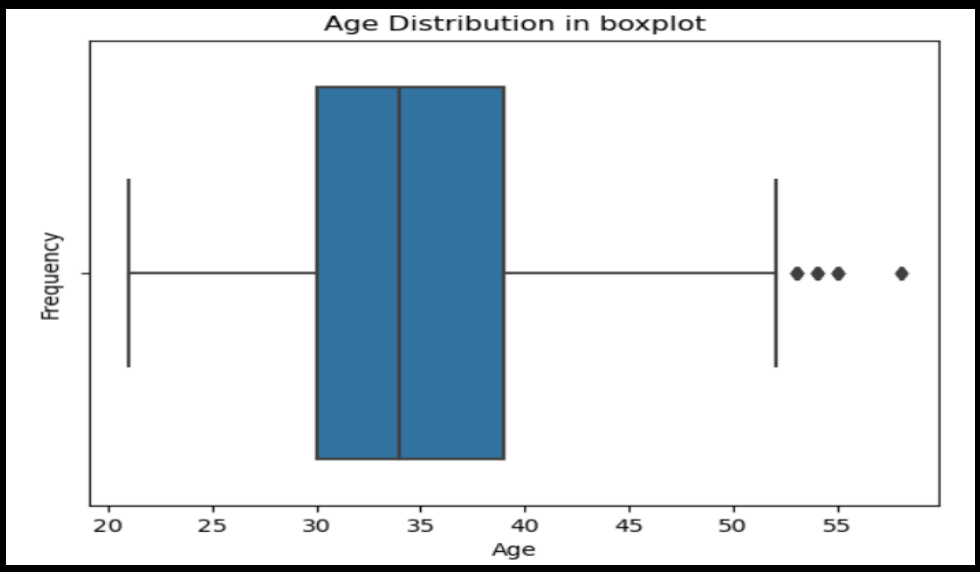
The dataset contains 19104 entries across 13 columns.

- ***Overview of the dataset:***
  - *There are three types column present in the dataset which are "Numerical", " Categorical" and "Datetime".*
  - *Columns like Driver_ID , Age, Gender, Educational_Level, Income, Joining Designation, Grade, Total Business Value and Quarterly Rating represent numerical features relevant to Driver's operations.*
  - *Columns like City represent categorical data such city names.*
  - *DateOfJoining and LastWorkingDate represent time-related data.*
  - *Some columns, like Age, Gender and LastWorkingDate contain missing values, which will need handling in preprocessing.*

- ***Statistical Insights:***
  - *Numerical Data Summary:*
    - *Each column has 19,104 records, except for Age (19,043) and Gender (19,052), indicating some missing values.*
    - *The average values for key variables: Age: 34.67 years, Income: 65,652, Total Business Value: 571,662 and Quarterly Rating: 2.01 (on a 1-4 scale).*
    - *Standard Deviation: The variability in income (30,914) and business value (1.13M) is quite high, suggesting significant disparities among individuals.*
    - *Total Business Value has a **negative** minimum value (-6M), which might indicate an error.*
    - *Income ranges from **10,747 to 188,418**, showing a large disparity in earnings.*
    - *Quarterly Rating follows a 1-4 scale.*
  - ***Outlier Alert**: min_item_price has a minimum of -86, which is clearly an error and requires correction or removal.*

- ***Categorical Data:***
  - *City: 29 different cities are present.*
  - *The most common city (C20) has **1,008 employees**.*

- ***Datetime Data:***
  - ***MMM-YY: 24 different months are represented.***
  - ***Dateofjoining: 869 unique dates, indicating a large spread of joining dates.***
  - ***LastWorkingDate: 493 unique values, but only 1,616 records are non-null, suggesting missing values.***

- ***Missing Values:***
  - ***Missing Data:*** *Age, Gender, and LastWorkingDate contain missing values.*
  - ***Potential Data Issues:*** *Negative Total Business Value values need investigation.*
  - ***Skewed Data:*** *Income and business value distributions have a large range, likely indicating outliers.*
  - 

  - *Here our dataset is normally distributed that's why impute it with mean value.*
  - ```
from sklearn.impute import KNNImputer
imputer=KNNImputer(n_neighbors=5)
gender=np.array(df['Gender']).reshape(-1,1)
imputer.fit(gender)
gender_imputed=imputer.transform(gender)

df['Gender']=gender_imputed.round()
```
  - *From above process we can impute gender missing values.*

## *Target Column*

- *We need to identify whether the Employee Currently working in our company or not. Based on the Last Working date filled rows means Employee Quit the Company and null rows means Employee still working with us.*
- *Above The Observation So We can say that the 17488 Employee Still Working with us and 1616 Employee Left the Company.*
- *Hasleftcompany is our target column.*
```

Creating New Features

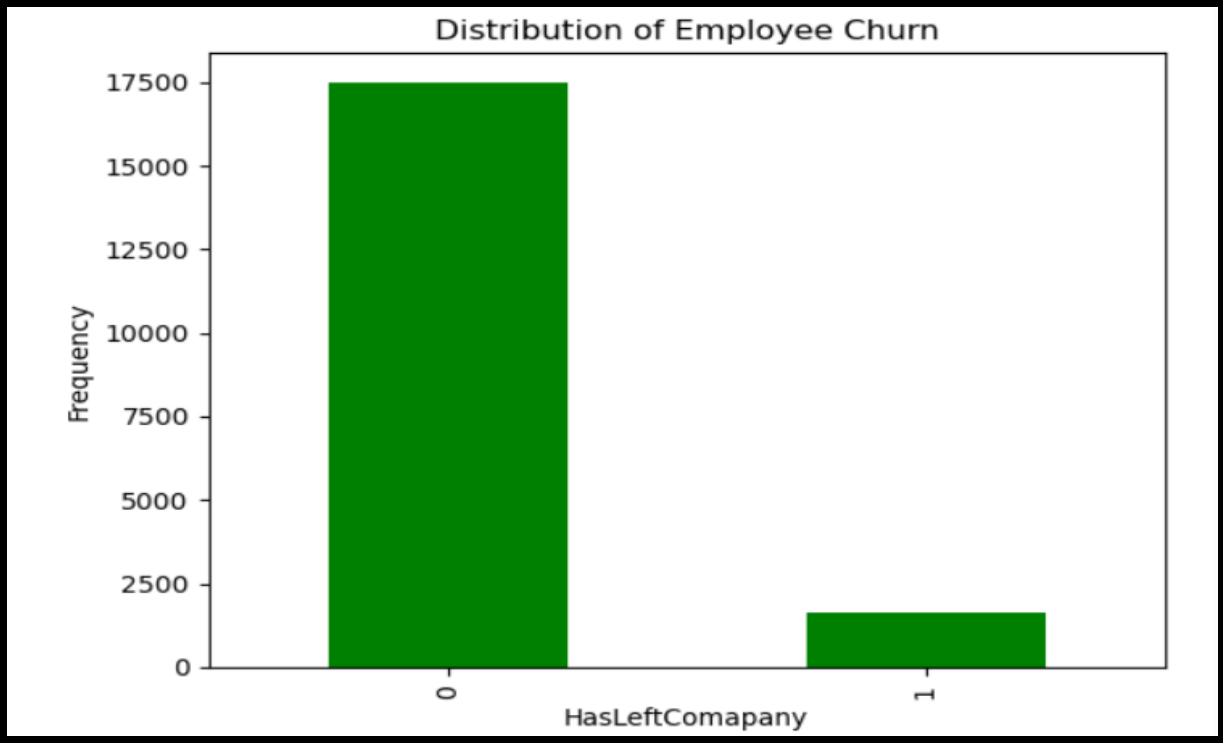
- *Birthyear and AgeAtJoining are the new features are created by the help of existing Dateofjoining column.*
- *Also, Tenure is also created using Dateofjoining and LastWorkingDate. We can also identify how many days each employees working our Organisation.*
- *Using cut to create Salary_Range on the basis of Income and Quarterly_Range on the basis of Quarterly Rating.*
- *Creating Riders_Age_Category on the basis of Age.*
- *Now we can also know the salary distribution amongs the Employee. We have maximum number of Employee whose salary fall between the 'Low Salary' and 'Medium Salary' range.*
- *We are also be able to identify the Quarterly Rating information. And We can see that our maximum rider falls between the Old Riders and Medium Riders.*
- ***Rating_Increased: -** Using groupby the Driver-ID and Quarterly Rating to find 1 is Yes and 0 is No.*
- ***Salary_Increased: -** Using groupby the Driver-ID and Income to find 1 is Yes and 0 is No.*

Class Imbalance Treatment

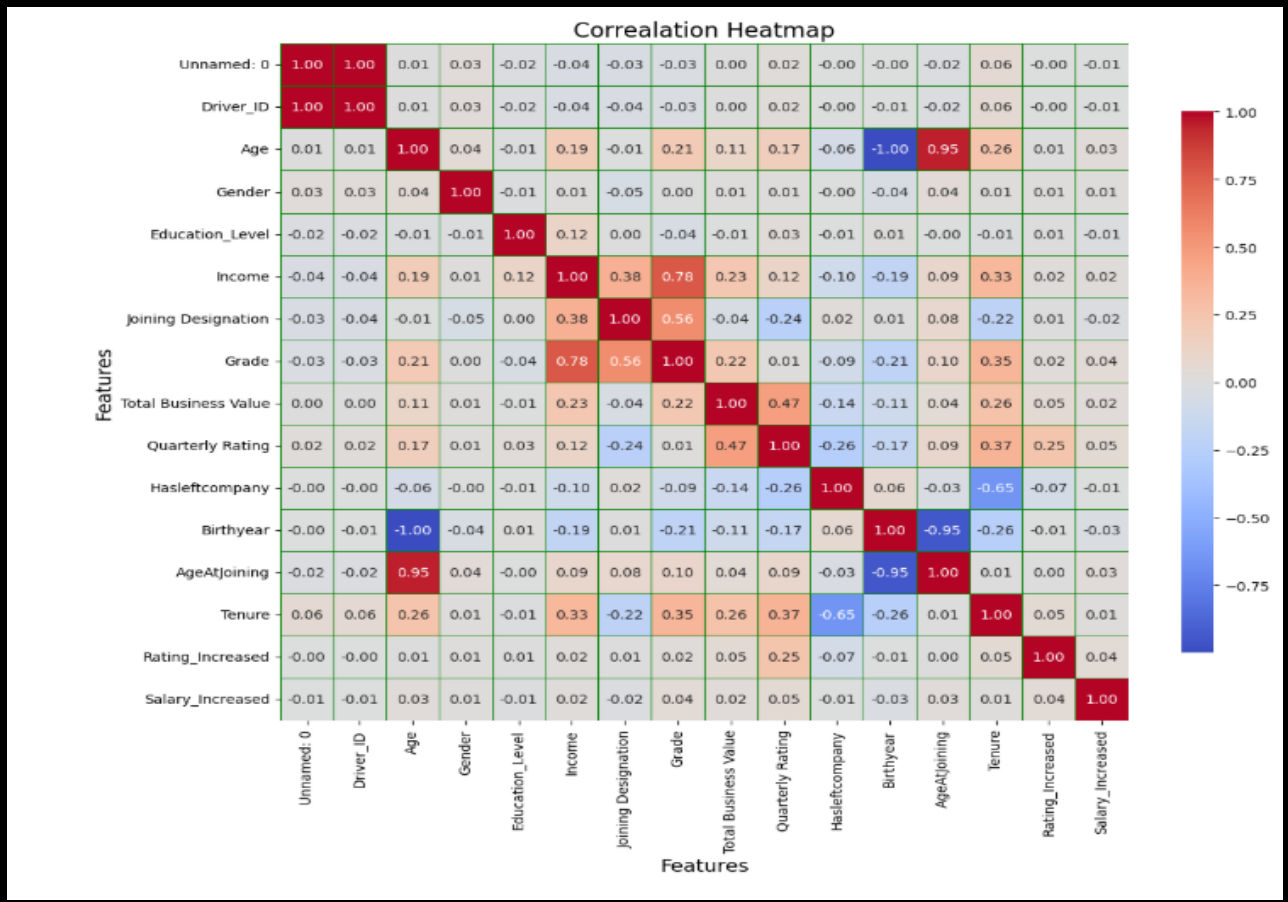
- *Here our target Variable is df["HasLeftCompany"] According to that we try to find how other features are impacted our target variable and how many Employees are still working and how many left the company.*
- *Address imbalance using techniques like oversampling, under sampling or synthetic data generation if necessary.*

Data Visualization

- ***Distribution of Employee Churn***

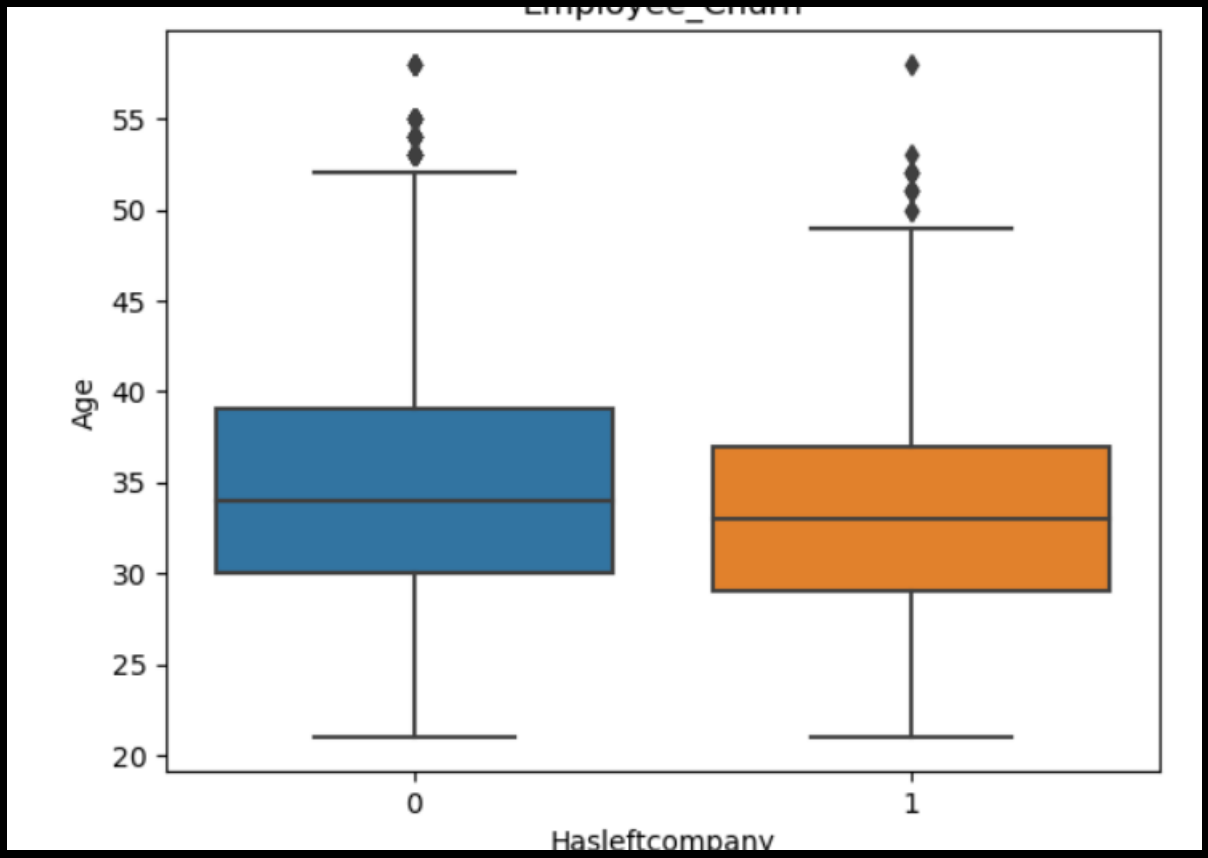


- ***Correlation Heatmap***



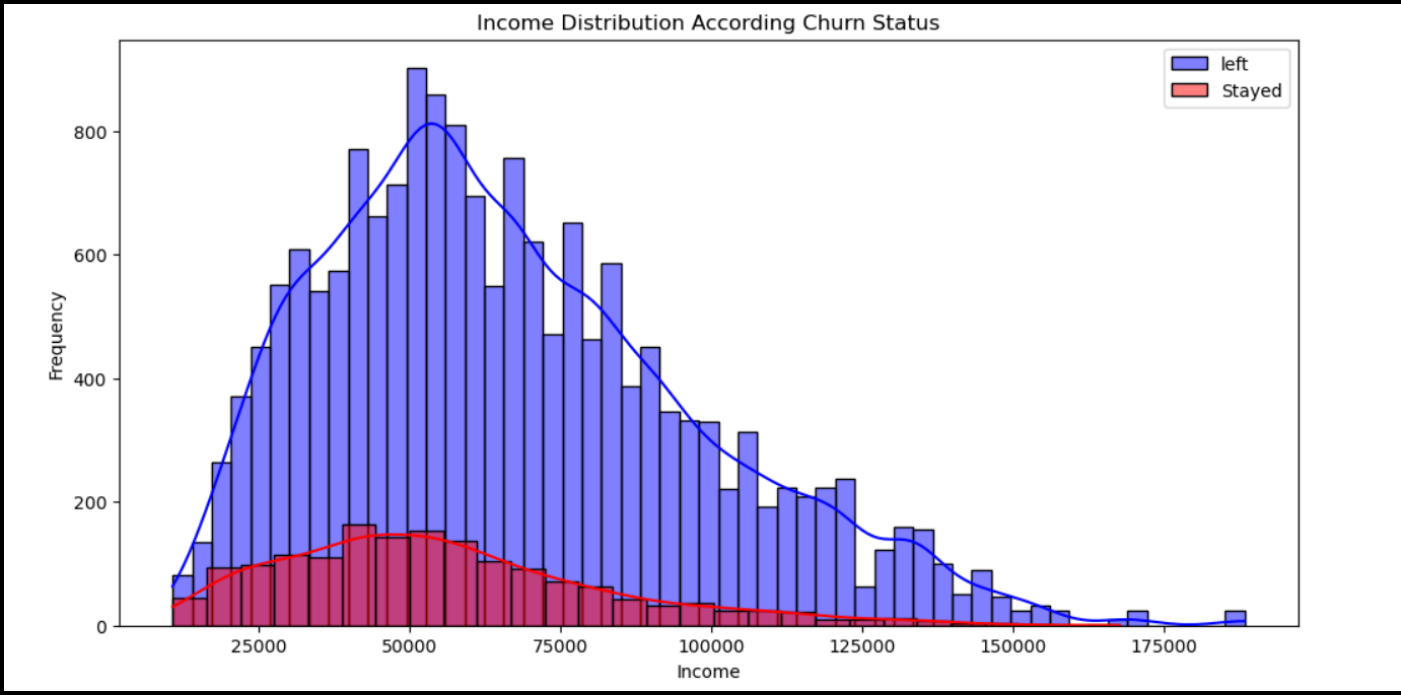
- *Here we can see that correlation between all features in the heatmap (a). The Darker color represents the feature is highly corelated, lighter color indicates the less corelated to each other and white color is indicate the no corelation or the almost negligible corelation between the features.*
(b). The highly corelated represents with the positive (1), highly negative corelated represents with the negative (-1) and the negligible corelation represents with the (0).

- ***Employee Churn***



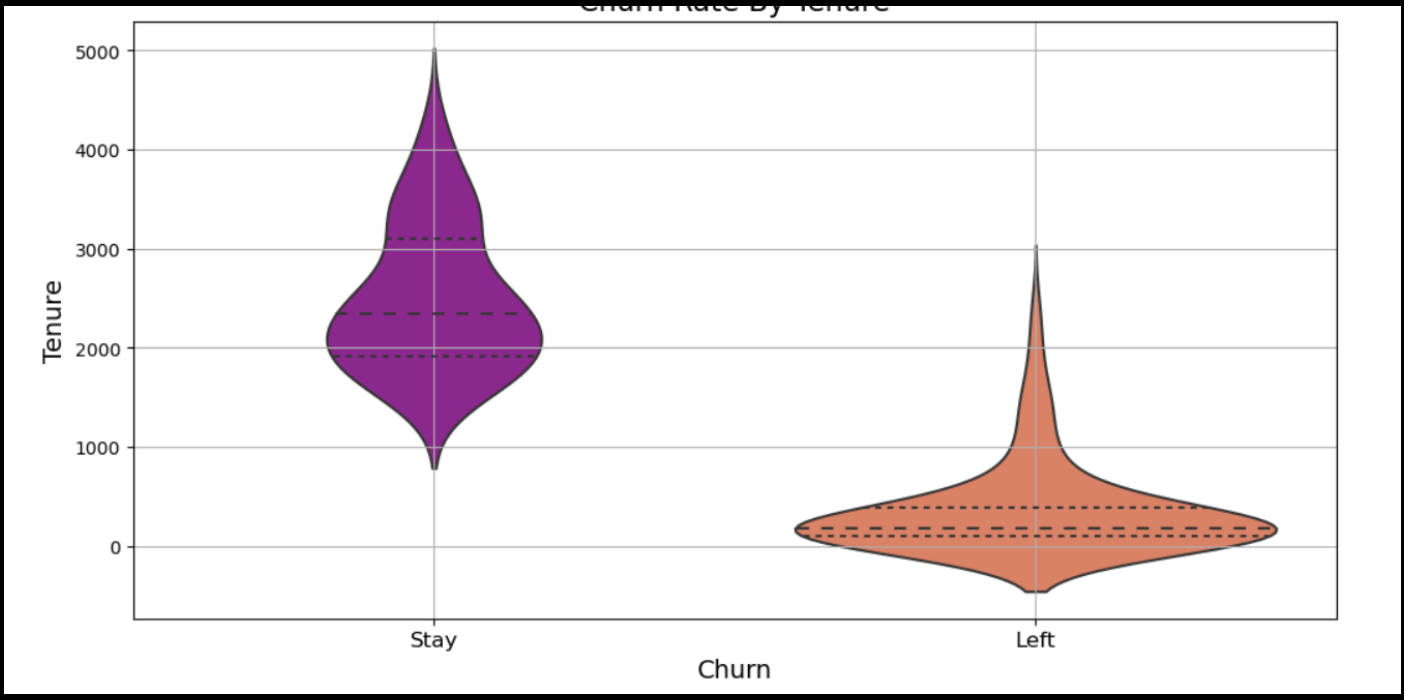
- *This boxplot compares the age distribution of employees who stayed (0) and those who left (1).*
- *The median age for both groups is similar, around 32-35 years.*
- *The interquartile range (IQR) for both groups is similar, with slightly more older employees among those who stayed.*
- *There are several outliers in both groups, particularly for employees older than 50.*
- *Overall, age does not appear to be a strong distinguishing factor for churn.*

- ***Income Distribution According Churn Status***



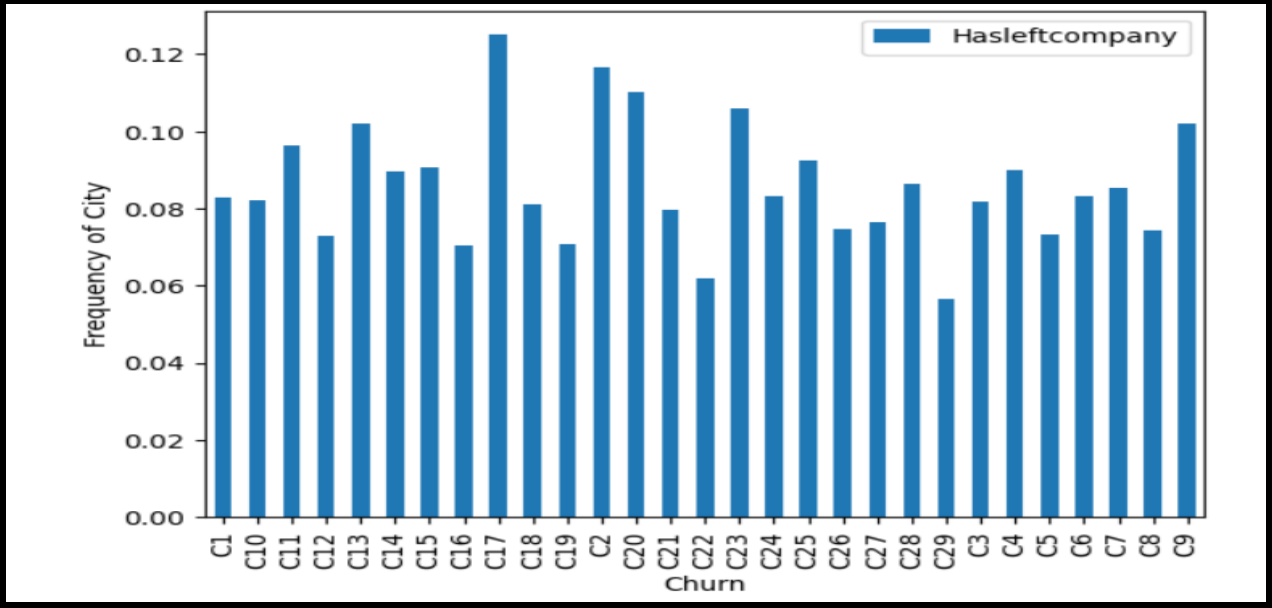
  - *This histogram compares the income distribution of employees who left (blue) vs. those who stayed (red).*
  - *The majority of employees who left had lower incomes (below $50,000), while higher-income employees were more likely to stay.*
  - *The income distribution for those who stayed is more concentrated at lower income levels, but they exist across the entire income spectrum.*
  - *Employees with higher incomes (above $100,000) seem to have a lower churn rate, as indicated by fewer red bars in that range.*
  - *This suggests that lower-income employees are more likely to leave.*

- ***Churn Rate by Tenure***

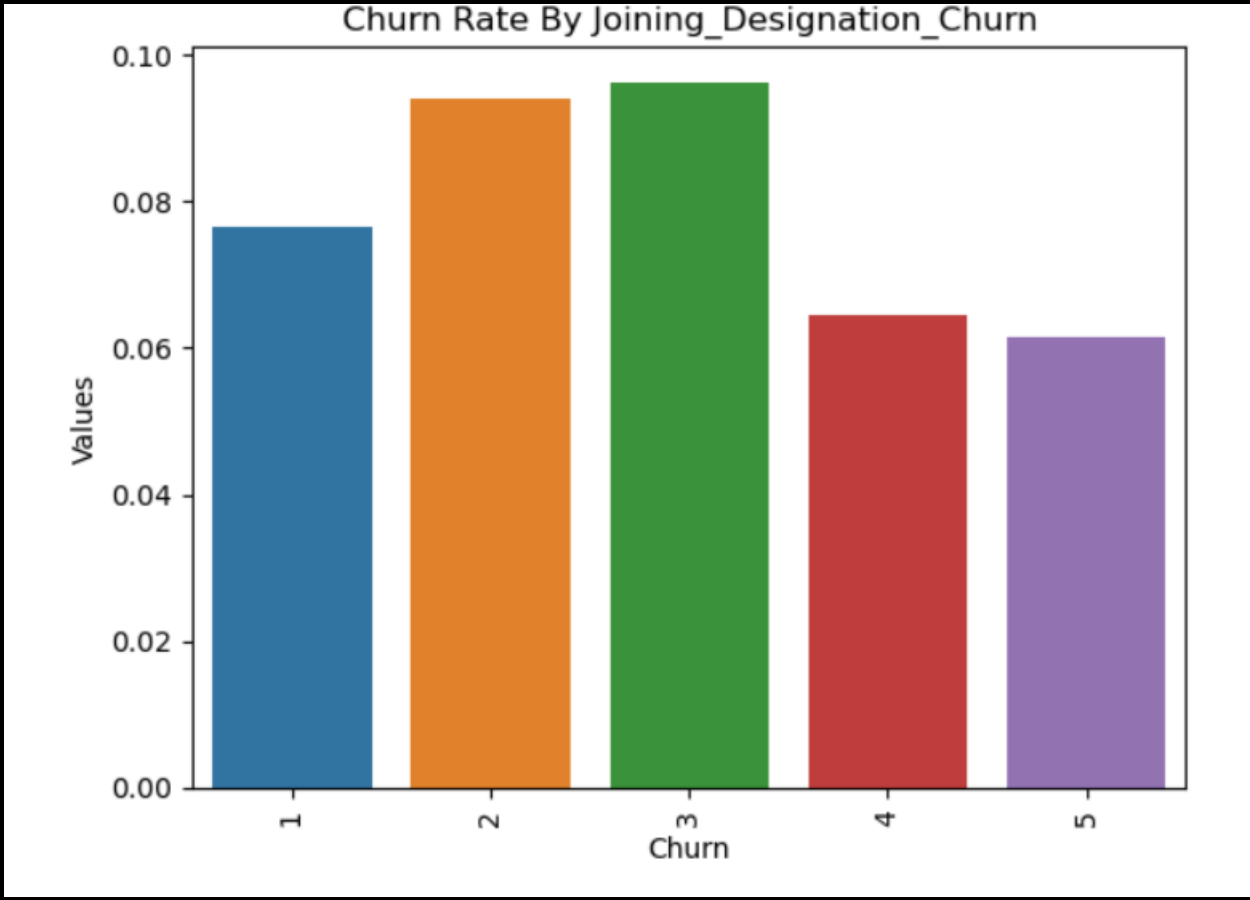


- *This violin plot compares employee tenure between those who stayed (purple) and those who left (orange).*
- *Employees who stayed have a wider distribution of tenure, with a peak around 2,000-3,000 days (~5-8 years).*
- *Employees who left have a much shorter tenure, with most leaving within the first 500-1000 days (~1-3 years).*
- *The shape of the distributions suggests that long-tenured employees are less likely to leave, while early-tenure employees are at higher risk of churn.*

- ***Churn Vs Frequency of City***

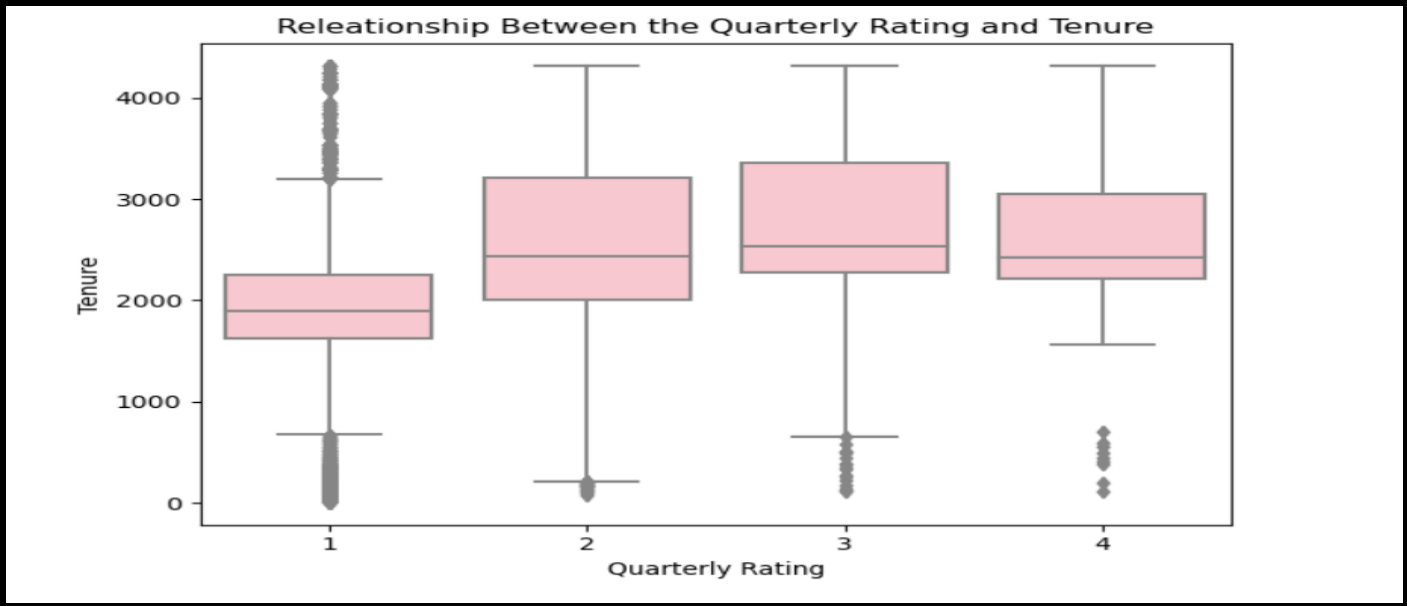


- ***Churn Rate by Joining Designation Churn***



  - *As per above two graphs show how churn rate affected by City and Joining Designation.*
  - *C17 city has highest churn rate and C29 is lowest.*
  - *Joining Designation 3 is highest and 5 is lowest.*

- ***Relationship Between the Quarterly Rating and Tenure***



- *Yes, Drivers with higher Quarterly Rating more likely to stay longer.*

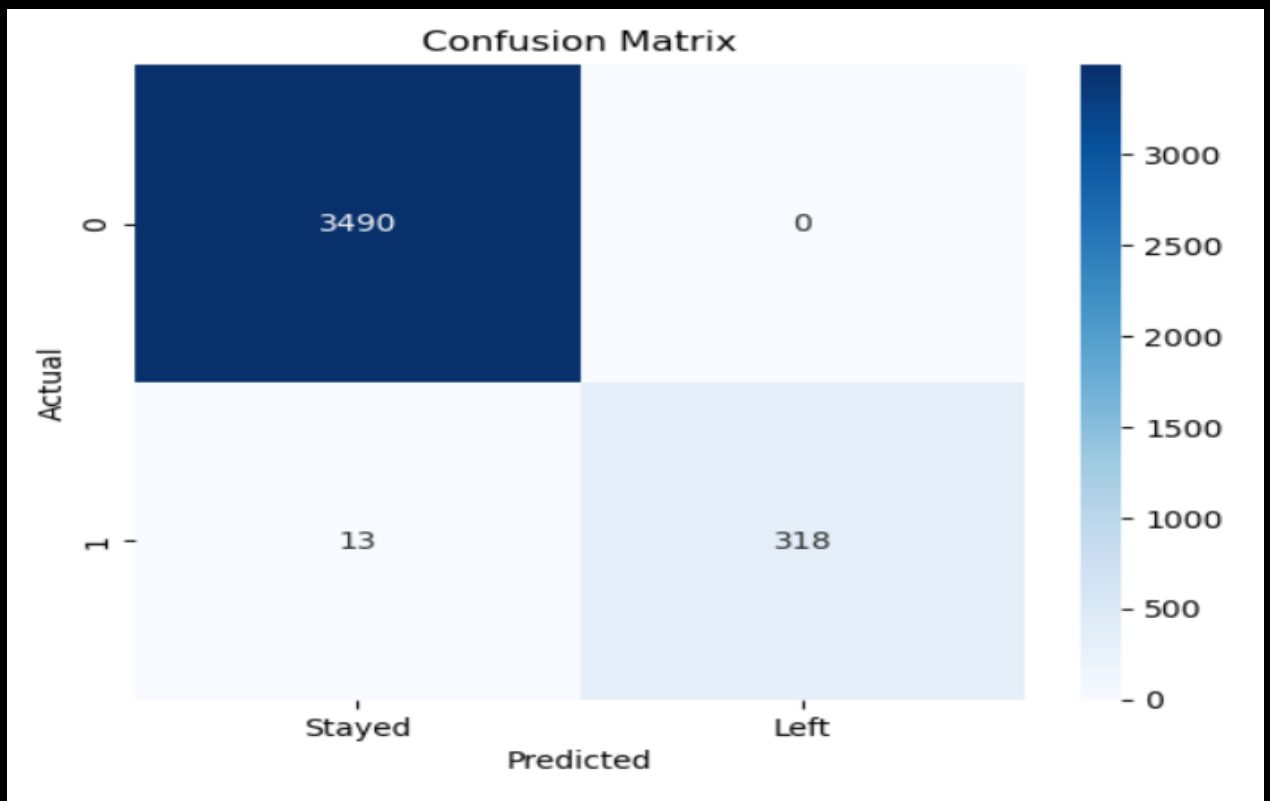
Standardize numerical features to ensure they are on the same scale. Standardize is done using StandardScaler.

Perform one-hot encoding for categorical variables like City, Education_Level, and Joining Designation.

After the performation of one-hot Encoding on the some features it converted into binary form and it's useful when we train our model.

- ***Predictive Analysis***

  - ***Model_Selection:*** *train_test_split*
  - ***Preprocessing:*** *StandardScaler, LabelEncoder*
  - ***Ensemble:*** *RandomForestClassifier*
  - ***Metrices:*** *classification_report, accuracy_score, confusion_matrix*
  - *Accuracy of 99.6%.*
  - *Classification Report of predictive analysis in which precision, recall, f1-score and support all are nearly 1.*
  - *Feature Importance using RandomForestClassifier where Tenure has highest.*
  - *At result, its best datasets as per the predictive analysis.*



Insights

- ***Income Level: -*** *Those Driver who falls within the low and medium salary range they are left the company within the year. So, the salary plays vital role in the gig economy jobs.*

- ***Quarterly Ratings: -*** *Lower quarterly range also influence the driver attrition if the Quarterly ratings are low, it's indicating the poor performance and dissatisfaction.*

- ***Tenure: -*** *Drivers who have been with company for short period have the faster attrition is often observed that the within the few weeks.*

- ***Education Level: -*** *Education is also affect the driver attrition because more educated drivers are less satisfied their job and seeking better opportunities.*

- ***Age and Gender: -*** *Drivers who falls within the 'Young riders' range switch their job frequently and also male attrition is higher than the female attrition.*

- ***City: -*** *City also affect the attrition because each city has different traffic conditions and competition.*

- ***Salary and Attrition: -*** *Drivers who falls within the lower and medium Salary range they leave the company because they are unsatisfied with the salary for more salary and higher growth in their career exploring the market and goes to other company like uber.*

Recommendations

- ***Competitive Compensation & Incentives:*** *- Increase base salaries or provide incremental salary raises at key tenure milestones. Offer performance-based incentives and bonuses to retain high-performing drivers. Introduce loyalty bonuses for long-term employees to encourage retention.*

- ***Enhanced Early-Tenure Engagement: -*** *Focus on onboarding and training programs to support new drivers in their first 1-3 years. Assign mentors or senior drivers to help new employees adjust and develop career paths. Provide clear career progression plans to give employees a long-term vision within the company.*

- ***Work-Life Balance & Flexibility****: - Offer flexible work schedules or improved shift options to accommodate personal needs. Implement wellness programs, including health benefits, insurance, and mental health support. Recognize and reward consistent performers through non-monetary incentives like extra leave days or awards.*

- ***Exit Interviews & Continuous Feedback:*** *- Conduct structured exit interviews to understand the main reasons for attrition. Implement regular surveys and one-on-one meetings to identify pain points before drivers consider leaving. Use data-driven insights to personalize engagement strategies for different employee segments.*

- ***Improved Scheduling & Route Optimization****: - Implement AI-driven route optimization tools to minimize long hours and fatigue. Allow for more predictable work schedules to help drivers maintain work-life balance. Reduce unnecessary idle time by streamlining dispatch and communication systems.*

- ***Location-Specific Incentives & Support****: - Analyse attrition trends by city to identify high-churn locations and target retention efforts accordingly. In cities with high living costs, provide cost-of-living adjustments or location-based bonuses to ensure competitive compensation. Improve transport and parking facilities in urban areas where congestion and lack of truck-friendly infrastructure may increase stress.*

Power BI link

https://app.powerbi.com/links/IfRjsPKjgS?ctid=ed77d40f-8c11-413d-96e2-467edfd73d60&pbi_source=linkShare


THANK YOU