**OLA – Business Case – Ensemble Learning**

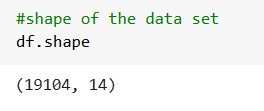
**Problem Statement :**

The major issue faced by OLA is the high churn among drivers. We need to work with the analytics department focused on driver team attrition. From the monthly information of drivers in 2019 and 2020, we need to predict whether a driver will be leaving the company or not based on their various attributes, not limited to, demographics, tenure information, historical data regarding performance of the driver.

**Exploratory Data Analysis**

Q) What is the shape of the data set?

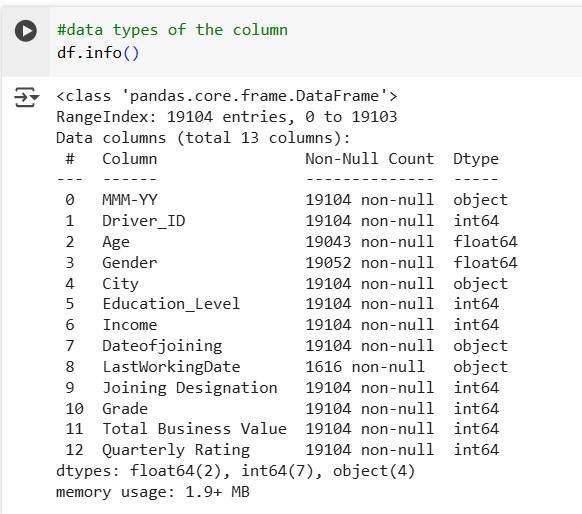
Ans)



Insights : There are 19104 records in the total data set. There are 14 columns for the data set which gives the attributes for the data set.

Q) What are the data types of the columns / features

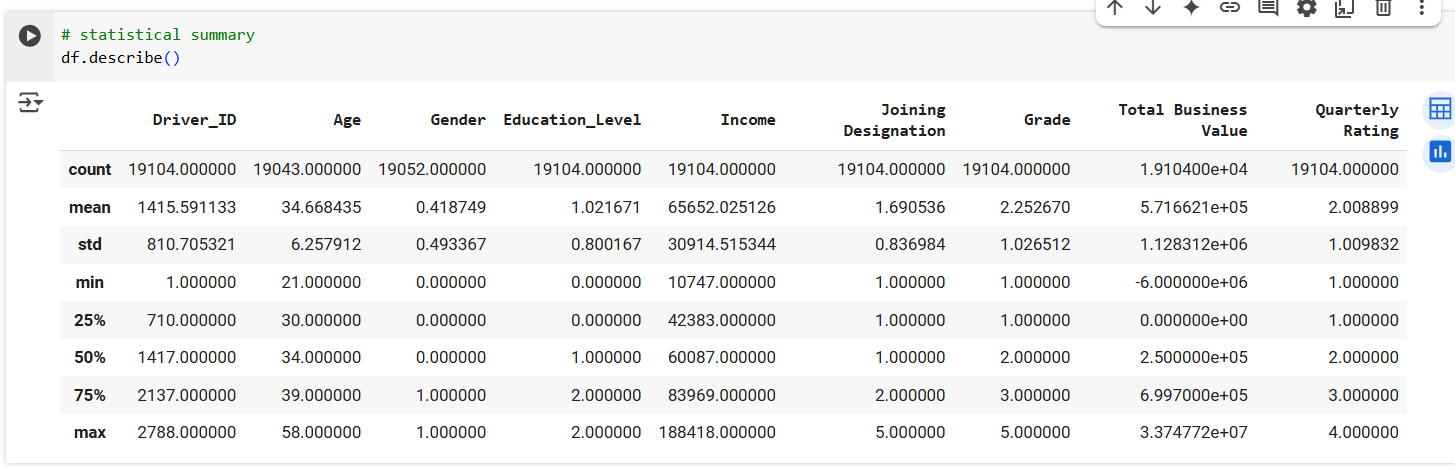
Ans)



Insights: There are total of 13 columns of which 2 of them are float data types, 7 of them are integer data types. 4 of the columns are object data types

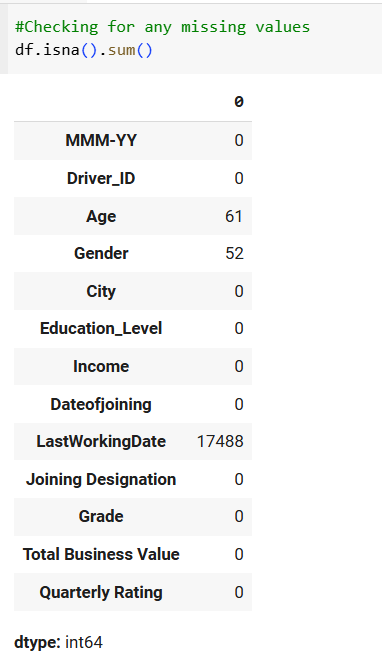
Q) Describe the statistical summary of the data set

Ans)



Q) Missing value detection

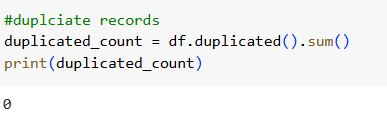
Ans)



Insights : The maximum number of missing data is from feature / column – LastWorkingDate with a count of 17488. The other columns which has missing values are Age (61) and Gender(52). These two columns are having a very low impact when compared to the total records of the data set 19104. But LastWorkingDate with count of 17488 is having a significant impact on the outcome of the analysis.

Q) Is there any presence of duplicated records in the dataset ?

Ans)

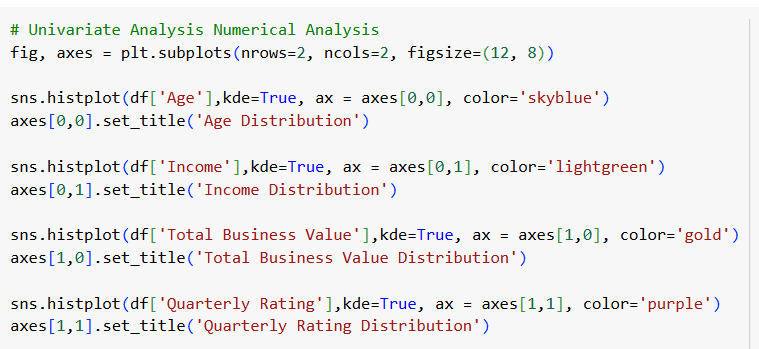


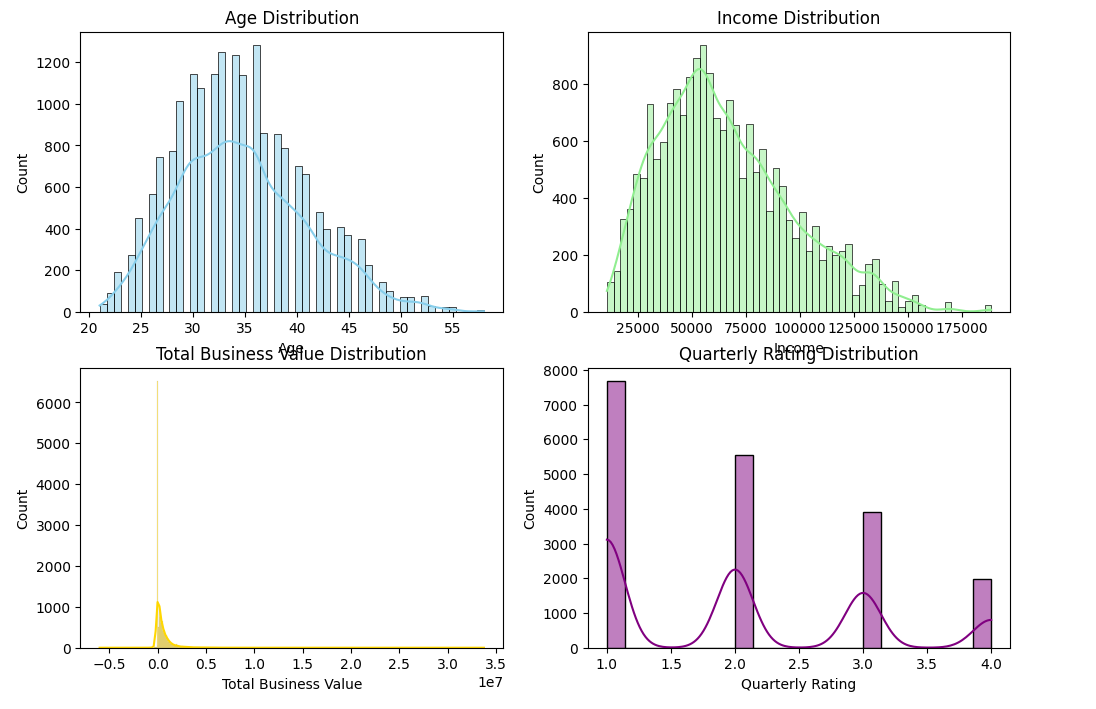
Insights : When we check the dataset, there are no duplicate records present. This implies that when we do the analysis the outcome that we find out will be in tandem with the original with no distortion as there is no duplicate records in the data set.

**Uni-variate Analysis**

Q) Conduct the uni-variate analysis for the numerical columns

Ans)





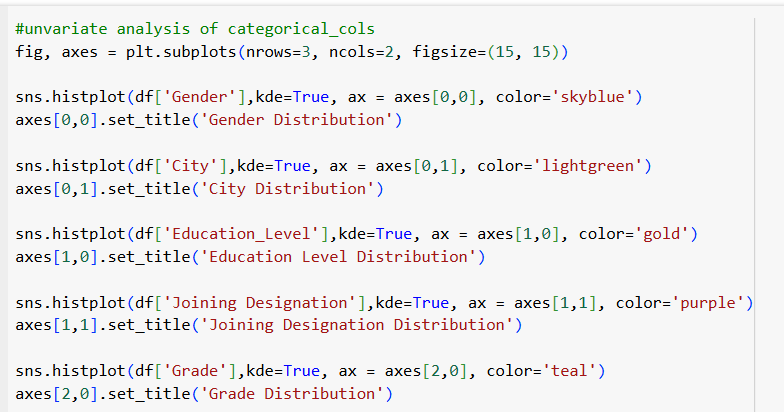
Insights: The graphs for Age and income distribution seems to be a bit right skewed i.e. there are individuals who have more income / higher age than the rest of the population that makes the graph skewed. The normal range for the distribution for Age is in the group of 30-35 years and for income it is around 50,000. The spread of the Age distribution is from 20-55. The spread for Income distribution is from 20,000 to 175000. The Age graph points to the presence of a younger population while the Income distribution helps to infer that the small and mid range of income earners are more when compared to high income group.

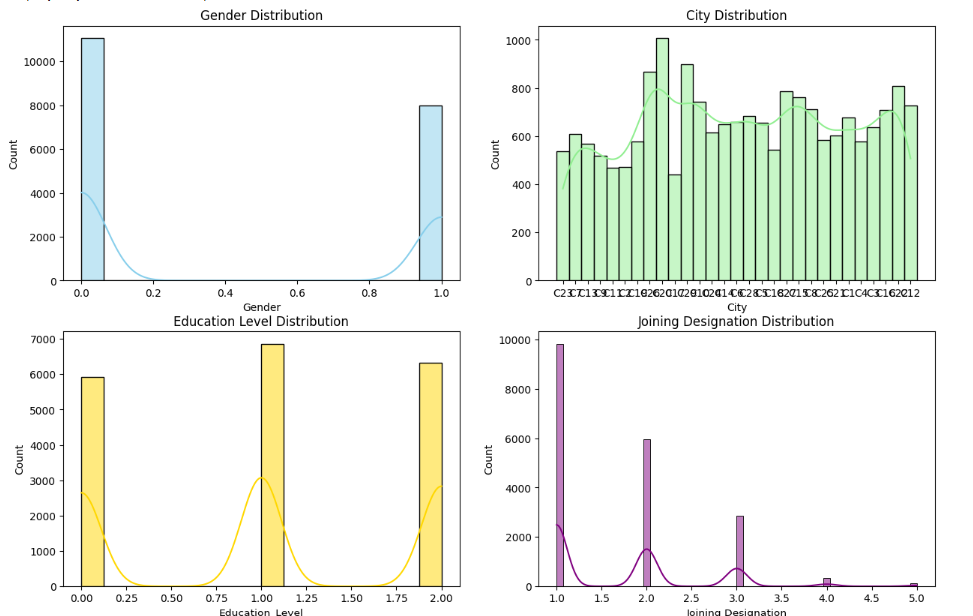
The total business value is clustered near to zero telling us that vast majority of business value is relatively low. i.e. there are very few high value businesses being done while most of them are low value businesses.

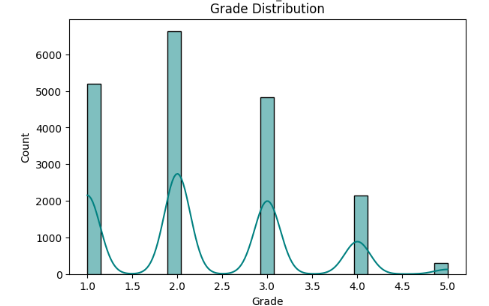
In case of quarterly rating it seems to be a multi modal graph with several peaks, while the concentration is near rating 1 and 2. The higher rating is comparatively less .

Q) Univariate analysis for categorical columns

Ans)







Insights :

Gender distribution – 0 => male and 1 => female. As per the graph its seems that the majority of the drivers are male, accounting for about 10000+. The number of females drivers are not less. They are nearly comparable. They account for about 8000. So we can say that there is no much disparity among the gender representation.

In case of City Distribution – The graph shows the geographical distribution of the data. Certain cities seem to have a larger representation in the dataset than others.C20 city seems to have the highest representation among drivers with about count of 1000. The lowest being C17 with just above 400 count . On average if we check, we can see that the city average to be around 700

With respect to Education level, all the three levels are almost in similar range with the max count reported for group 1 ( for 12+ ) with around cout of 6500. The next highest is for group 2(for graduate) with count of around just above 6000 and last being for group 0(for 10+) with count of just below 1000. This shows that among drivers all sectors of education background can be seen with no bias for people from particular educational background.

On analysing Joining designation distribution we can understand that the maximum drivers has joined for level 1 and the minimum number of drivers have joined for the level 5. The count for level 1 nearing to 10000 and the goes decreasing for the rest of the levels with level 5 a meagre value around 130.

If we check the Grade of the driver when joining, we can see that the maximum count is for the grade 2 followed by 1,3,4 and 5. The grade 2 has a count near to 6500 followed by grade 1 with just above 5000. The lowest count would be for grade 5 near to zero a value around 305 or so.

Recommendations :

Though the count for females are comparable more offers could be given to female drivers like joining bonuses so that we can attract more females drivers. Presence of female drivers will improve the commutation by female passengers as they would be more comfortable with female drivers in most of the cases. This increased count will inturn bring us more revenue.

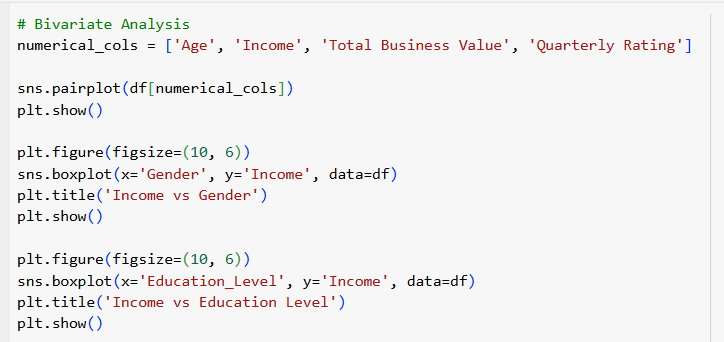
The City representation is almost similar across the various cities. We should look into the cities for the average number of trips conducted per day per driver so that we can increase the induction for more drivers in those cities and offer better incentives.

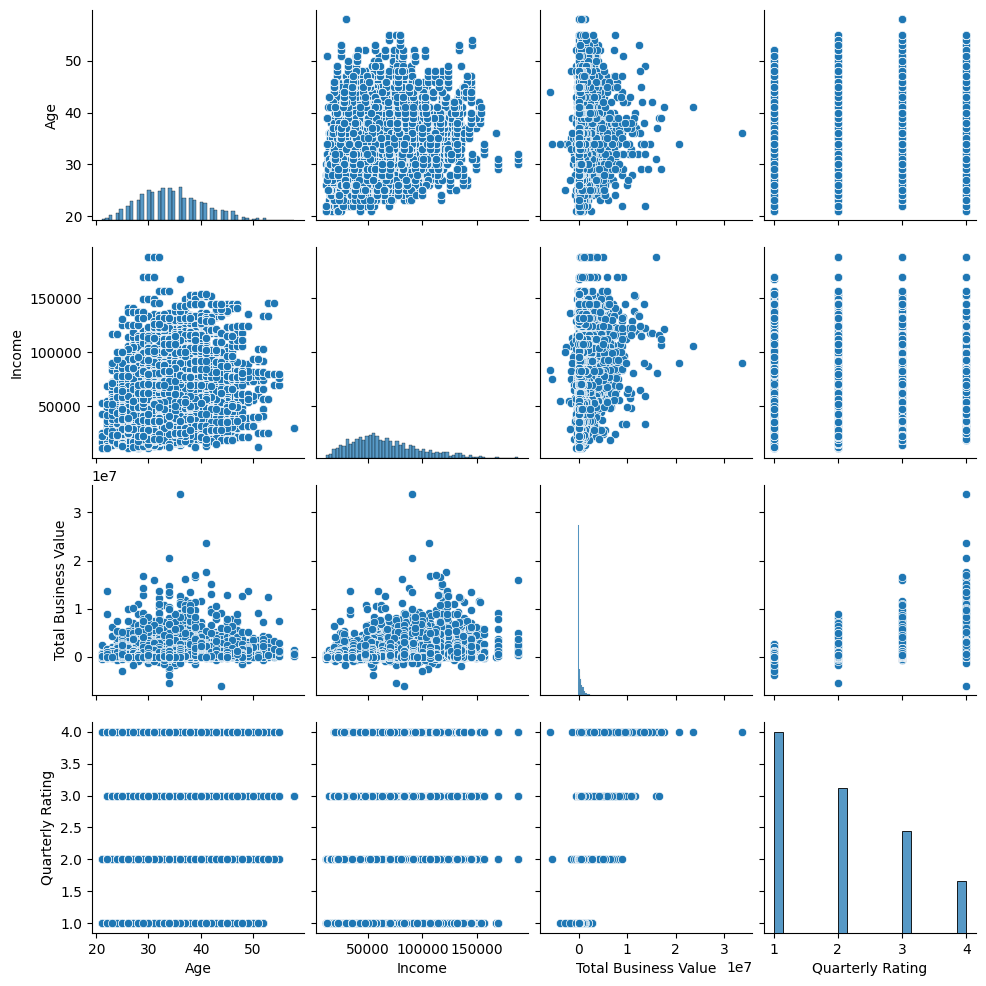
Educational background analysis says that there is no preference to be given for any of the groups since almost all the groups have similar count

Analysis of the Grade and joining designation of drivers tell us that higher grade drivers are very less when compared to the entry level grade. Higher grade employees should be given more offers and incentives to join us so that people would prefer them on a long term basis esp. For the scheduled rides and out station rides which would give us more revenue.

Q) Bi Variate analysis

Ans)





Insights :

Age vs. Income:

There is no clear linear relationship between age and income. Individuals across all age groups seem to have similar income ranges, though younger individuals (below 30) tend to cluster in lower-income brackets.

Age vs. Total Business Value: No strong correlation is evident. Businesses with high total business value are distributed across all age groups. Younger individuals (below 30) appear to have fewer high-value businesses.

Age vs. Quarterly Rating:

Quarterly ratings do not show any significant dependence on age. Ratings are uniformly distributed across all age groups.

Income vs. Total Business Value:

A weak positive relationship exists between income and total business value. Higher-income individuals tend to have businesses with higher total business value, but there are exceptions.

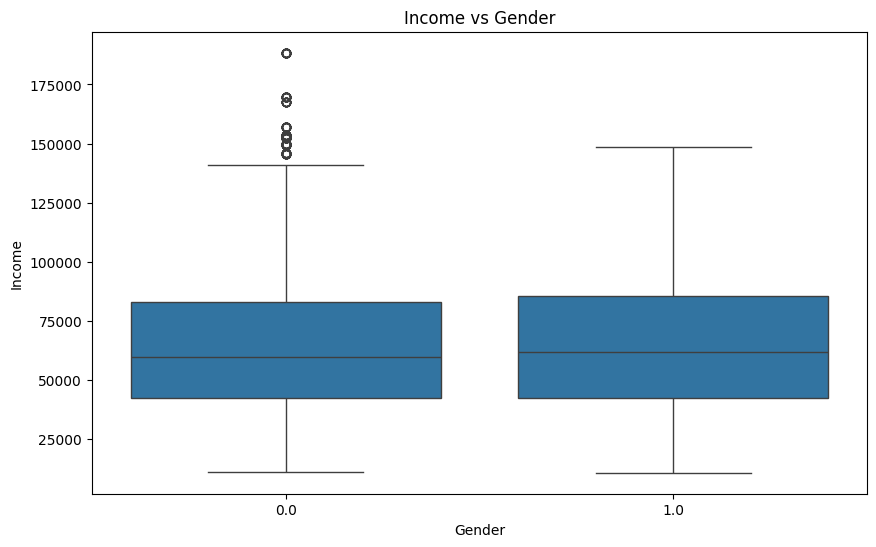
Income vs. Quarterly Rating:

No clear pattern is visible between income and quarterly rating. Individuals with varying incomes receive similar ratings.

Total Business Value vs. Quarterly Rating:

No significant relationship is observed between total business value and quarterly rating. Businesses with both low and high values receive similar ratings.

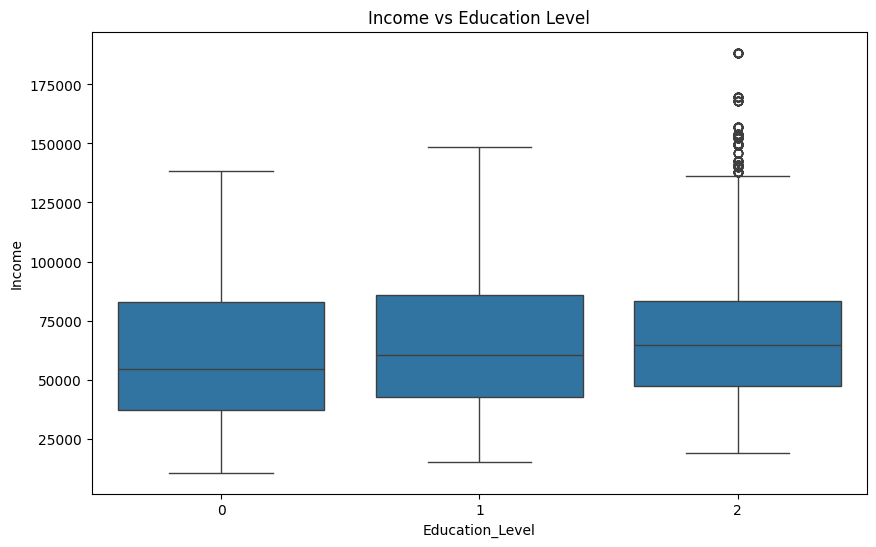
Most variables do not exhibit strong linear relationships, suggesting that other factors might influence outcomes like business value or quarterly ratings. Outliers are evident in income and total business value distributions, which may require further investigation for their impact on overall trends.



Insights : Both gender categories have a similar median income, approximately Rs 55,000 to Rs60,000. This indicates no significant difference in central tendency for income between the two genders. IQR for both genders is similar, with most incomes falling between Rs 40,000 and Rs75,000. Male Gender has several outliers above Rs 125,000, with some incomes exceeding Rs 175,000. Female Gender does not show any visible outliers.

The similarity in median and IQR suggests that income distribution is relatively equal across the two gender categories. However, the presence of outliers in male gender indicates that some individuals in this group earn significantly more than the rest.

Recommendations : The presence of outliers should be checked for any other reasons too. So that those factors if possible should be integrated or consolidated for female drivers too to improve their income so that more female drivers stay with the company.

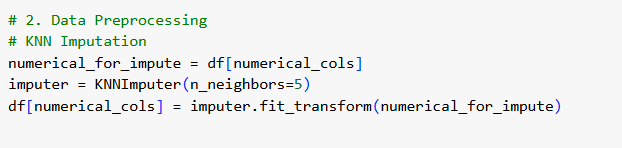


*Insights :*  Individuals with the highest education level (Level 2) exhibit greater variability in income, with some earning exceptionally high amounts. However, the median income does not differ significantly across education levels. The presence of outliers in Level 2 suggests that there are high income section among the level 2 education amongst drivers . The similarity in median incomes across levels implies that factors other than education—such as experience, industry, or geographic location—may play a more significant role in determining income.

*Recommendations :* Since the median lies in the similar region with IQR also in same range, we need check for the outliers present in the level 2. We must find the reasons for the more income being generated for the level2 so that similar methods or culture could be implemented in the other two levels too to improve the income of drivers, thus more retention.

Q) Data preprocessing

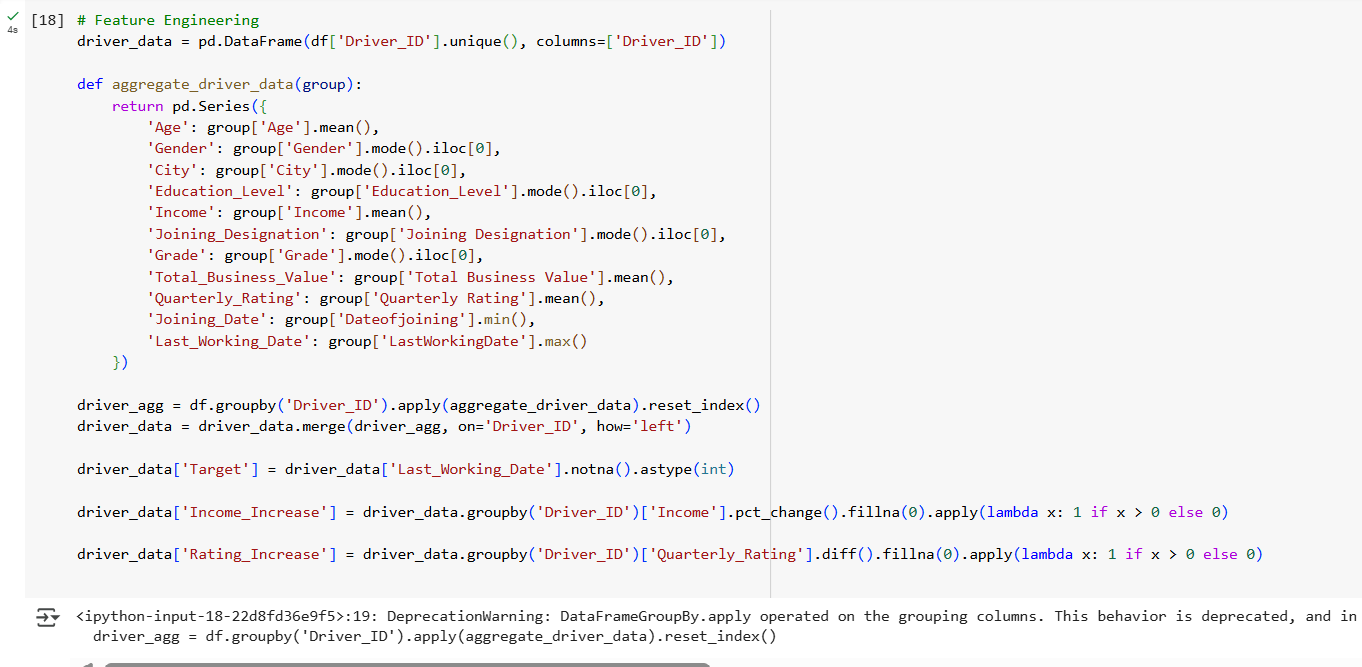
KNN imputation



Insights : when we do the KNN imputation we are replacing the missing values with the estimated values based on the non-missing values of their nearest neighbours in the dataset.

Q) Feature Engineering

Ans)



Insights : we are creating a new data frame by grouping the data of each driver. Each row in the new data frame represents a unique driver\_ID. In the aggregate\_driver\_data function we can see that for certain columns we are directly taking the mean of the various rows for the features like Age, Income, Total business value, quarterly rating. While for features like Gender, City, Education level, Joining designation, grade we are taking the mode value which is the most repeated value among them. The minimum value for joining date is taken as we need the very first value and the max. Value for last working date is taken to get the longest value of dat.

The 'Target' variable likely indicates whether a driver has a recorded last working date (1) or not (0), potentially representing churn or active status

‘Income \_Increase’ column -- This calculates the percentage change in 'Income' for each driver across their records in the original df.

'Rating\_Increase' indicating whether a driver experienced a quarterly rating increase (based on the potentially multiple records in the original df).

The resulting driver\_data DataFrame is a feature-engineered dataset where each row represents a driver, and the columns contain aggregated statistics and derived features that could be useful for further analysis

Q) Class imbalance Treatment

Ans)



One-hot encoding creates new binary (0 or 1) columns for each unique category within the specified categorical columns

Class imbalance – This occurs when one class (e.g., drivers who left) has significantly fewer instances than the other class (e.g., active drivers). This can negatively impact the performance of ML models, as they might be biased towards the majority class.

Synthetic Minority Over-sampling Technique (SMOTE) - SMOTE is a popular oversampling technique used to address class imbalance. It works by creating synthetic instances of the minority class by interpolating between existing minority class instances.

After applying SMOTE, The number of instances in the minority class in y\_resampled will be increased, making the class distribution more balanced compared to the original y.

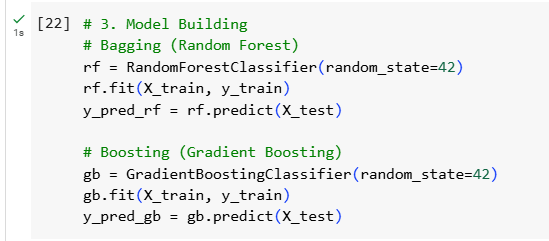
X\_resampled will contain the original data points along with the newly generated synthetic data points corresponding to the oversampled minority class.

Once Standardisation is completed The numerical features in X\_scaled will have a mean of approximately 0 and a standard deviation of approximately 1.

Next we do a Train-Test Split: here we divide the data into training and testing sets to allow for model training and evaluation.

Q) Model Building

Ans)

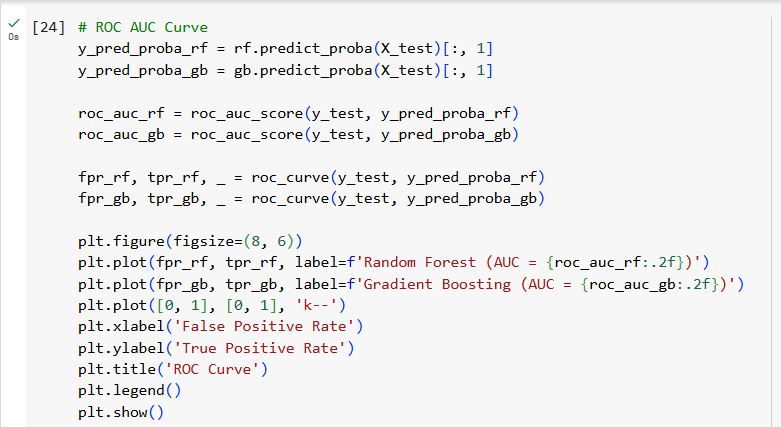


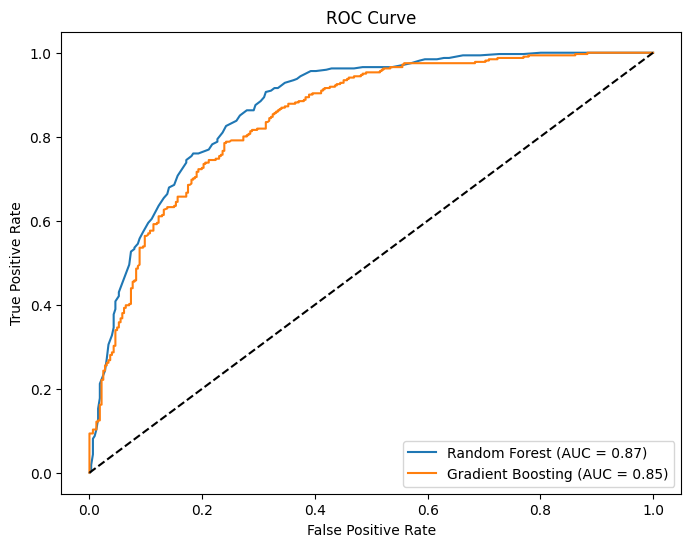
Insights : In model building, we are creating model that will help us giving answers for the new data that would come to us. The model is created based on the training data which will help predict the outcome when a new data is fed. Here the bagging ensemble method is represented by the random Forest classifier. The Boosting ensemble method is represented by the Gradient boosting classifier.

Bagging is basically used for reducing the variance and over fitting while boosting is used to reduce bias and improve accuracy. While former is trained in parallel and independently, the latter is trained sequentially, each trying to correct the previous.

Q) ROC AUC curve

Ans)





Insights : The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The area under the curve summarizes the overall performance of the classifier.

Here the orange curve represents the Gradient boosting ( Boosting ensemble) and the blue curve represents the Random Forest ( Bagging ensemble )

Generally, the higher the AUC, the better the performance of the classifier.

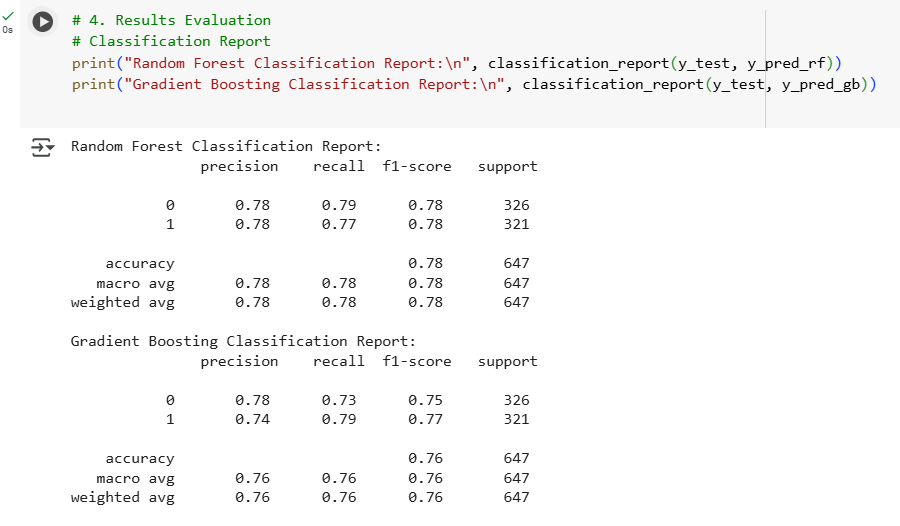
Random Forest (AUC = 0.87): The AUC value of 0.87 indicates that this model has a good ability to distinguish between the positive and negative classes. There's a significant area under the curve, suggesting better than random performance.

Gradient Boosting (AUC = 0.85): The AUC value of 0.85 also indicates good performance, slightly lower than the Random Forest model.

This means that overall, the Random Forest model has a slightly better ability to discriminate between the positive and negative classes on this particular dataset and task.

Q) Classification report

Ans)



Insights : 1 represents the driver who has last working date in the data set => who left the company and 0 represents those who do not => Active driver

*Random Forest Classification report :*

Precision is 0.78 for both 0 and 1 i.e. 78% of the prediction were actually true for both the classes.

Recall is the is the ratio of correctly predicted positive observations to the total actual positives.

79% is for class 0 and 77% for class 1. i.e. these are percentages of correctly identified instances among the total dataset of true scenario for both classes.

Accuracy = 0.78. The model correctly classified 78% of the instances in the data set.

*Gradient boosting classification report :*

Precision is 0.78 for class 0 and its 0.74 for class 1 i.e. 78% of the prediction were actually true for class 0 and 74% of the predictions were actually true for class 1.

Recall is 73% is for class 0 and 79% for class 1. i.e. these are percentages of correctly identified instances among the total dataset of true scenario for both classes.

Accuracy = 0.76. The model correctly classified 76% of the instances in the data set.

*Comparison of the Two Models:*

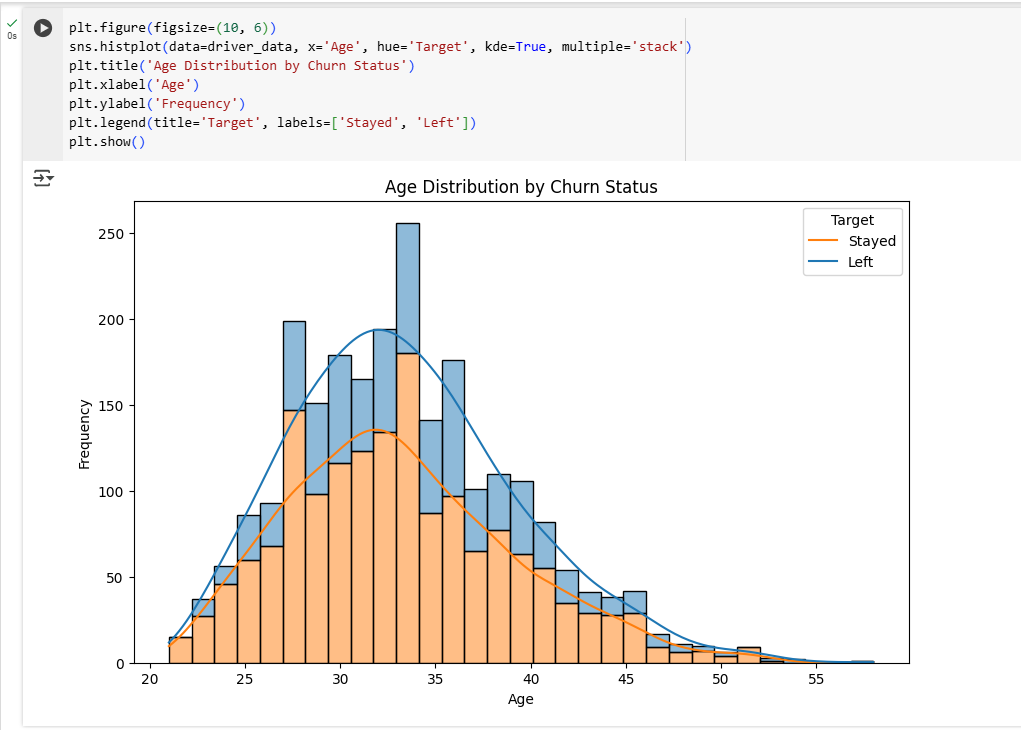
Based on these classification reports, Random Forest Shows slightly better overall performance than Gradient Boosting on this data set, with a higher accuracy (0.78 vs. 0.76) and higher F1-scores for both classes. The precision and recall are also more balanced between the two classes for Random Forest.

Gradient Boosting has a slightly higher recall for class 1 (0.79 vs. 0.77 for Random Forest), meaning it's slightly better at correctly identifying actual instances of class 1. However, it has a lower recall for class 0 (0.73 vs. 0.79 for Random Forest), indicating it misses more actual class 0 instances. Its precision for class 1 is also lower (0.74 vs. 0.78 for Random Forest), meaning when it predicts class 1, it's slightly more likely to be wrong.

It should be noted that the choice of the "better" model might also depend on the specific business problem and the relative importance of precision and recall for each class.

Q) Age v/s Churn rate

Ans)



Insights : this shows that more younger people has left the job. i.e. among young people, the churn rate is high when compared to the senior drivers ( in age)

Q) Actionable insights and Recommendations

Ans)

Insights

* Income and rating increases are significant predictors of driver retention.
* Certain cities or joining designations show higher churn rates.
* Younger drivers tend to have higher churn.
* Rest of the insights are given along with the questions / visual analysis through out the document.

Recommendations:

* Implement targeted retention programs for high-risk driver segments like the young people where the churn rate is high.
* Provide professional development opportunities to improve quarterly ratings as among these the rating 1 is having the highest count. Better rating among the drivers will make them more incentivized as well as them give them more opportunities for rides.
* Revise income and incentive structures to improve driver satisfaction.
* Analyze city-specific churn patterns and retention strategies can be tailor made accordingly so that the its not a one size fit all approach. If not for all the cities individually, based on the churn rate and the rides we can club the areas and create the solutions.