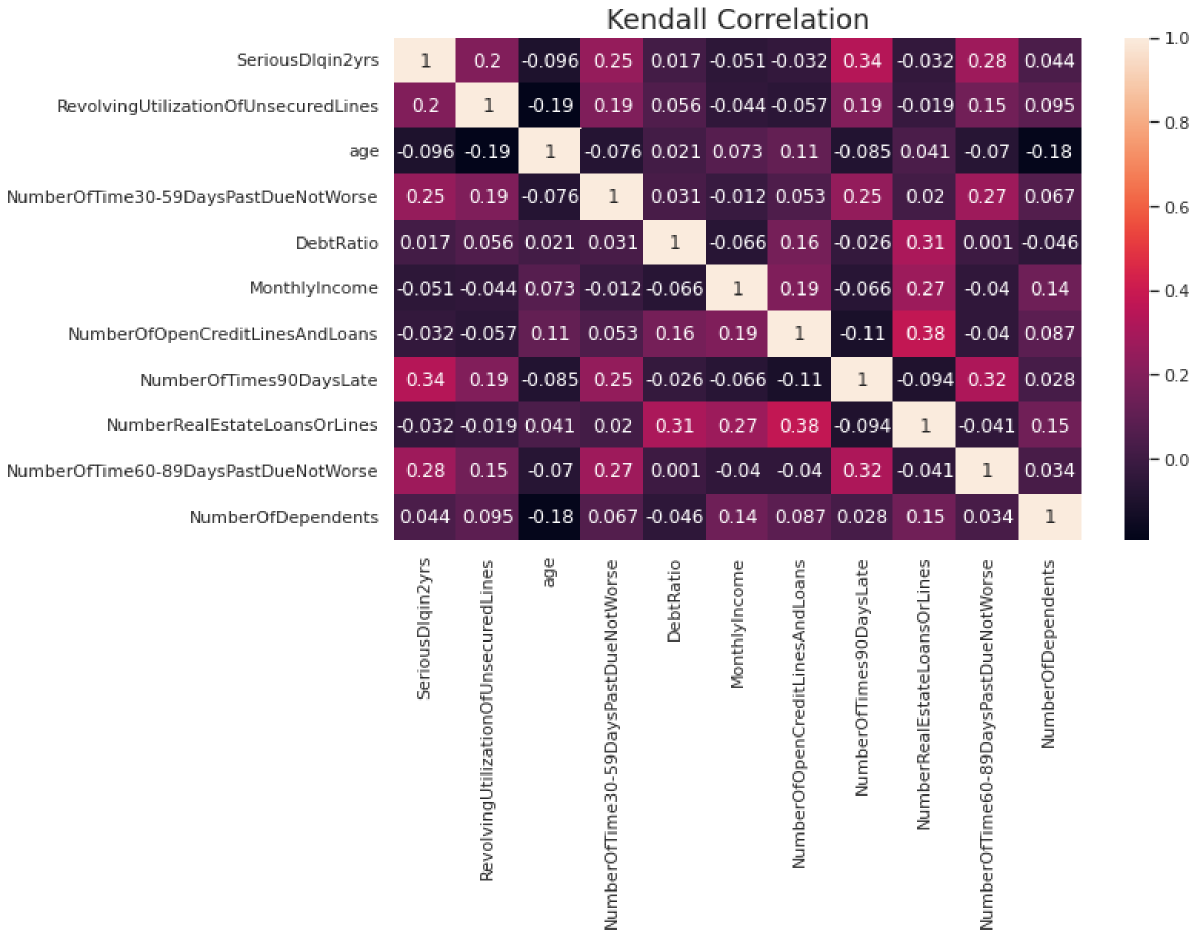
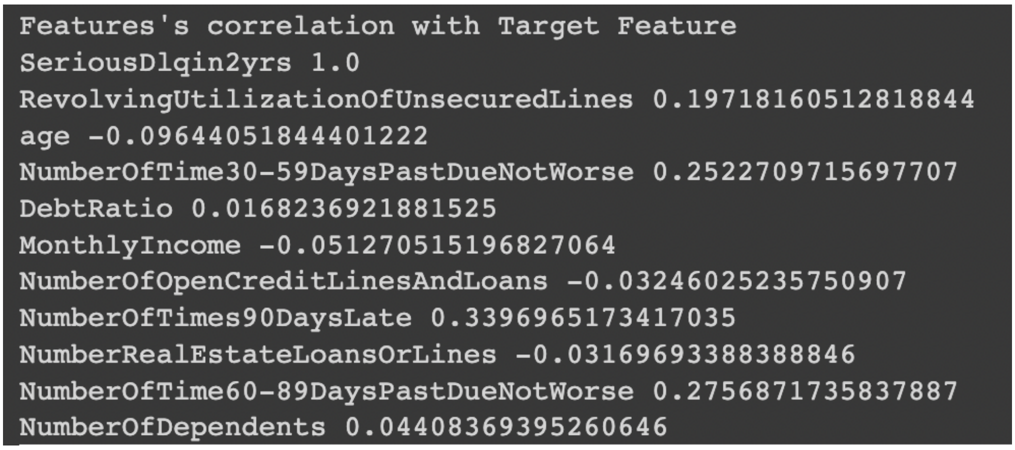
1. What are the factors that have a high correlation with the probability of loan default?



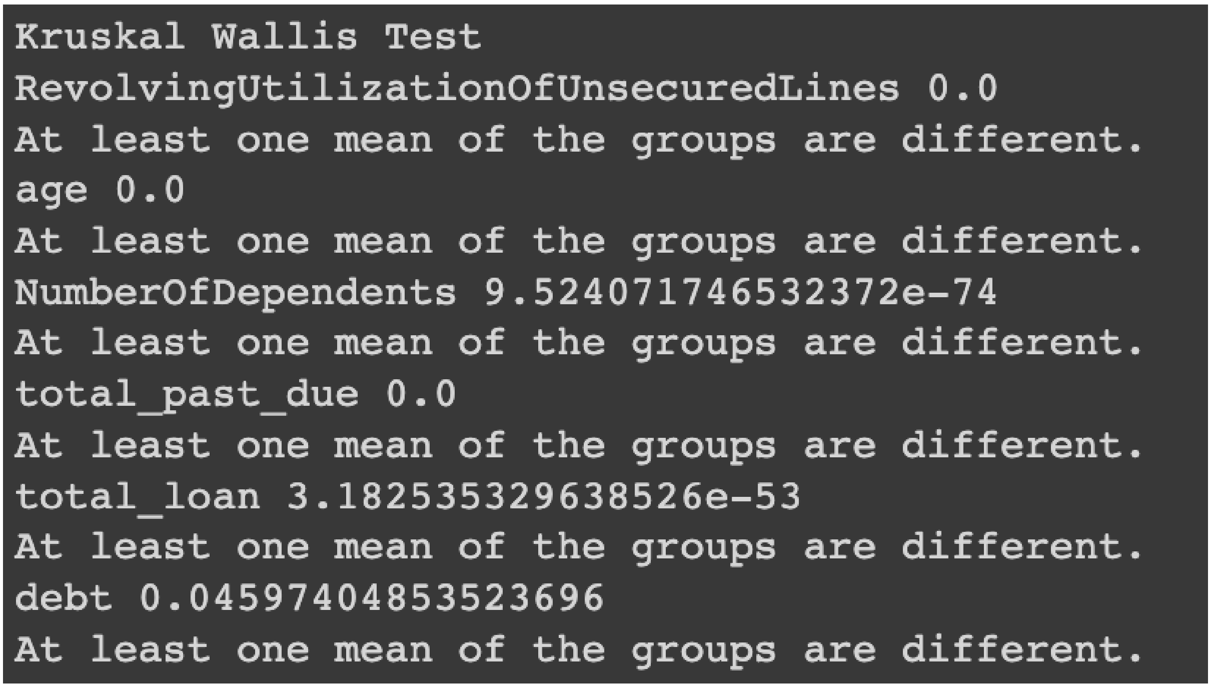
- The highest correlation with the target feature is achieved by “NumberofDays90DaysLate”(0.34) followed by similar features representing the frequency of late due payments.
- I have plotted the Kendall correlation as Pearson and Spearman would not go with the categorical target feature.
- The complete list of independent features and their respective correlation with target variable is mentioned below,



```
Features's correlation with Target Feature
SeriousDlqin2yrs 1.0
RevolvingUtilizationOfUnsecuredLines 0.19718160512818844
age -0.09644051844401222
NumberOfTime30-59DaysPastDueNotWorse 0.2522709715697707
DebtRatio 0.0168236921881525
MonthlyIncome -0.051270515196827064
NumberOfOpenCreditLinesAndLoans -0.03246025235750907
NumberOfTimes90DaysLate 0.3396965173417035
NumberRealEstateLoansOrLines -0.03169693388388846
NumberOfTime60-89DaysPastDueNotWorse 0.2756871735837887
NumberOfDependents 0.04408369395260646
```

2. **Are there interaction effects occurring among the variables?**

- Interaction effects mean the effect of one variable on another variable through a third variable. There are a couple of techniques such as OLS, Two-way ANOVA, and the Kruskal-Wallis test.
- I have implemented the Kruskal Wallis test,
    1. Additive effect of NumberOfOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines
    2. The multiplicative effect of DebtRatio and MonthlyIncome

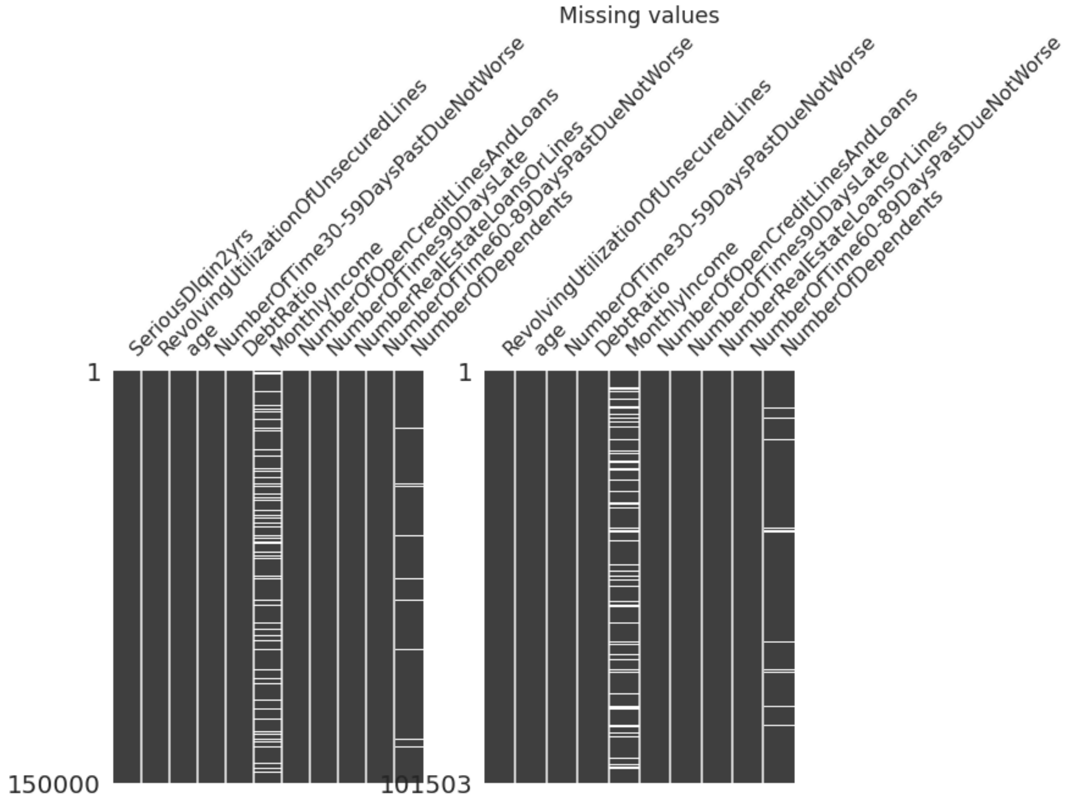


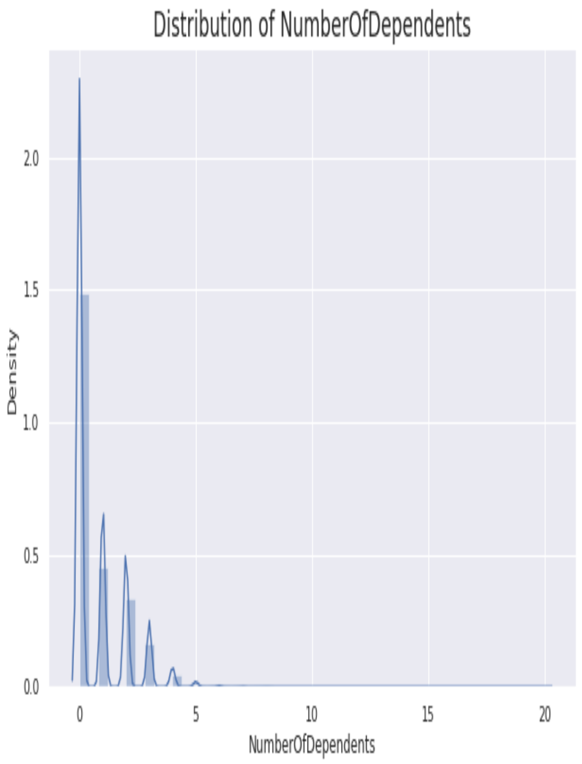
```
Kruskal Wallis Test
RevolvingUtilizationOfUnsecuredLines 0.0
At least one mean of the groups are different.
age 0.0
At least one mean of the groups are different.
NumberOfDependents 9.524071746532372e-74
At least one mean of the groups are different.
total_past_due 0.0
At least one mean of the groups are different.
total_loan 3.182535329638526e-53
At least one mean of the groups are different.
debt 0.04597404853523696
At least one mean of the groups are different.
```

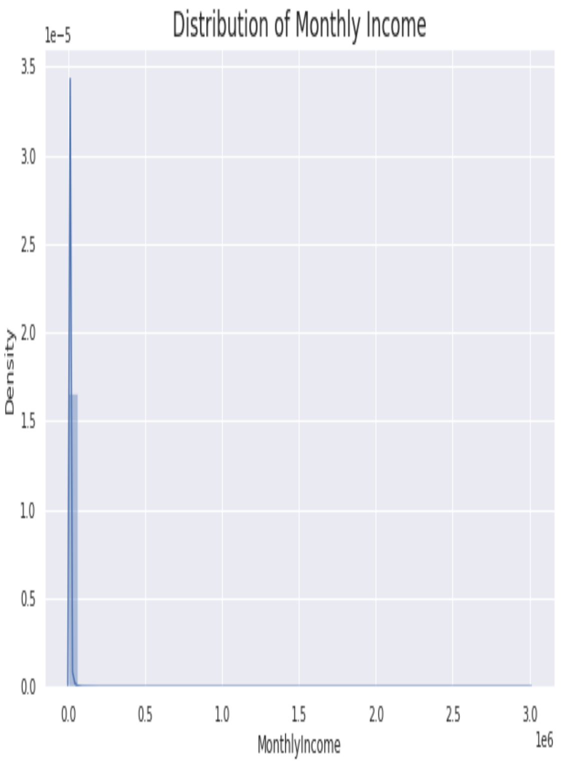
- The p-value of both the new variables is less than 0.05 so we can reject the null hypothesis.

- But, the overall performance of the baseline model decreased to a great extent so I haven’t included the new variables in the final model.
- In contrast to that, To overcome multicollinearity, I performed a weighted summation of three other variables which boosted the performance of the model.

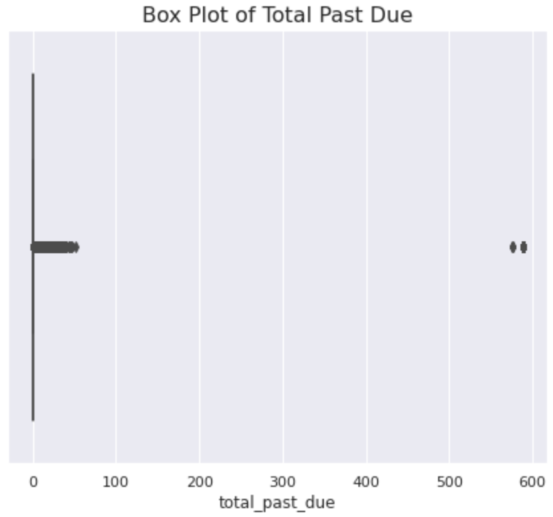
3. Any other preliminary analysis of the given dataset?

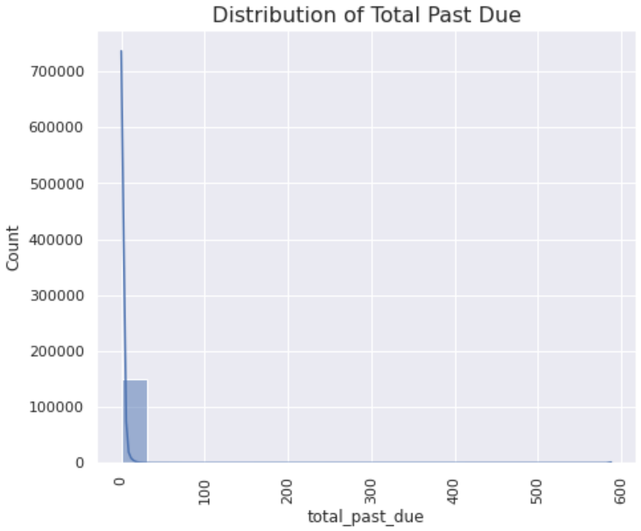




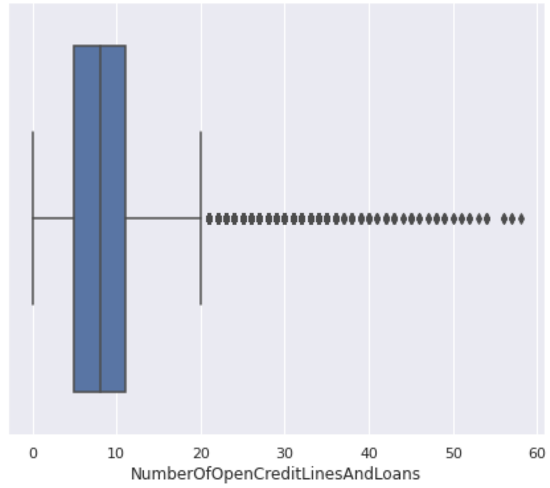


- The first plot depicts the missing values in two features namely the Number of Dependents and Monthly Income.
- The histogram shows that they are not normally distributed so I have replaced the missing values with respective medians.
- The below images show the boxplots, histogram, and scatter plot of the independent variables.

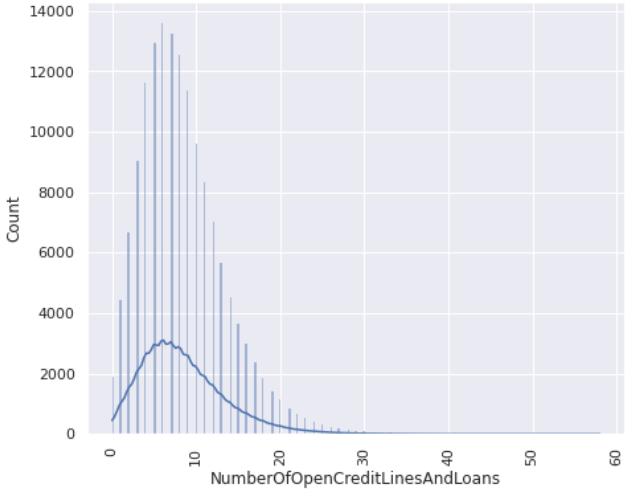




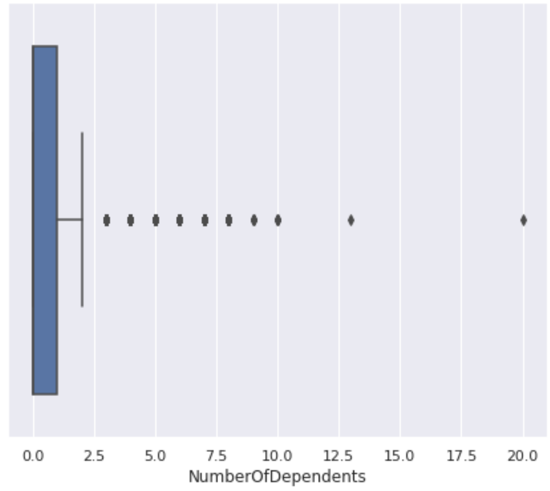
Box Plot of NumberOfOpenCreditLinesAndLoans



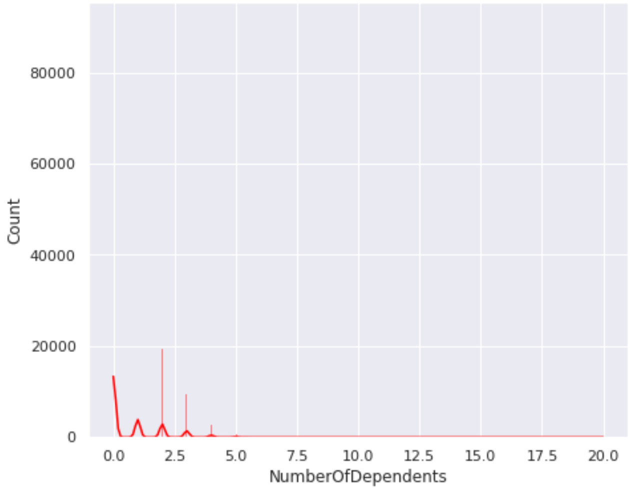
Distribution of NumberOfOpenCreditLinesAndLoans

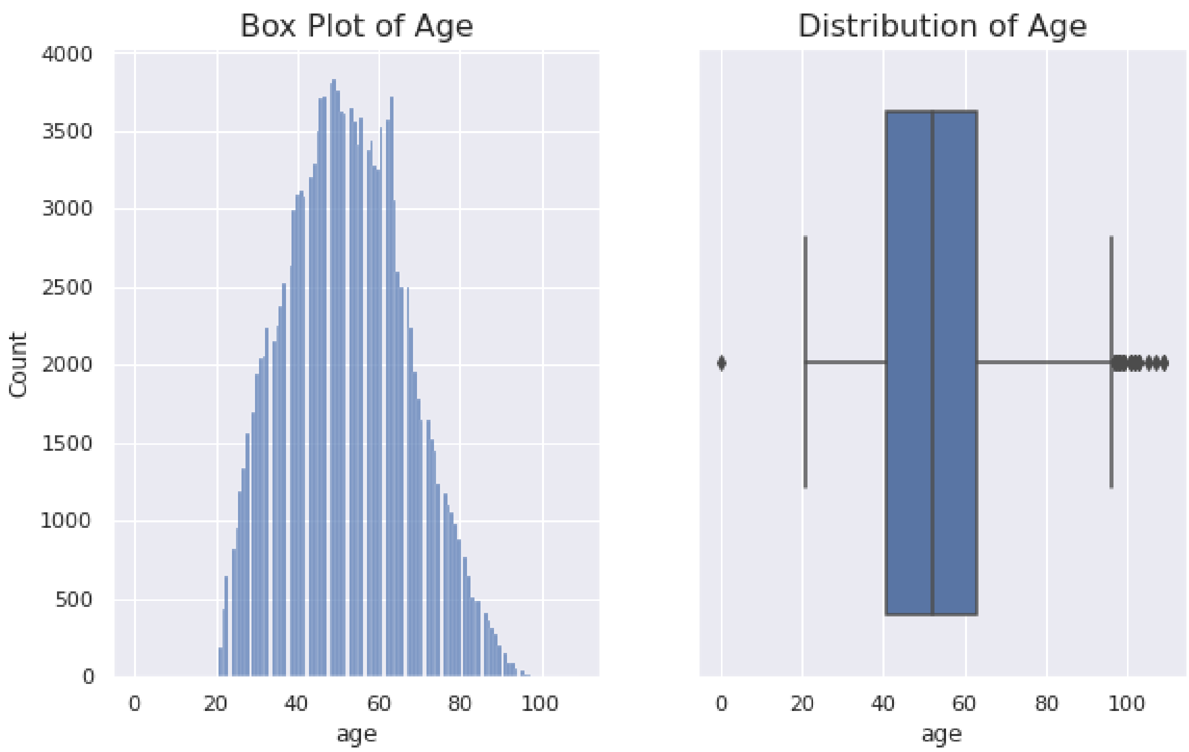


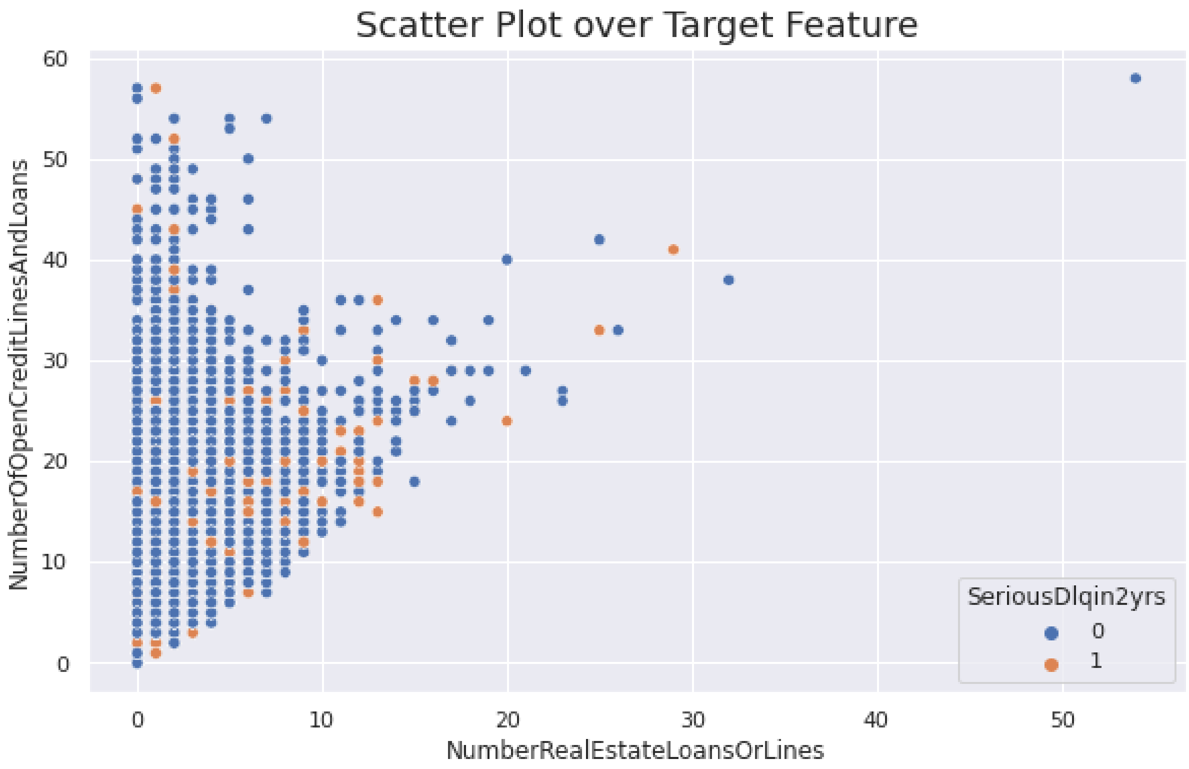
Box Plot of NumberOfDependents



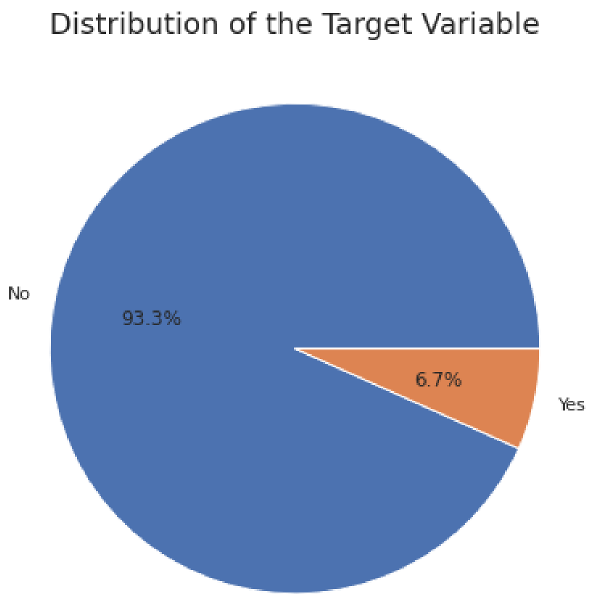
Distribution of NumberOfDependents







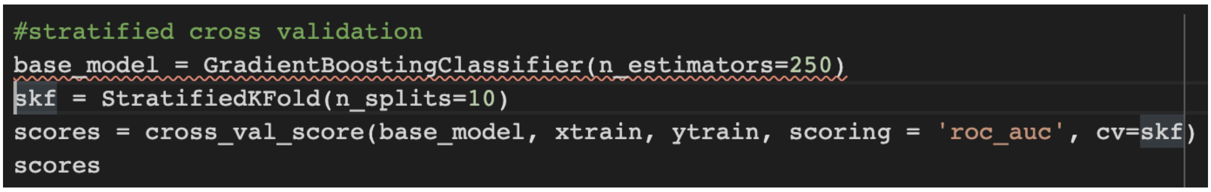
- The target features’ distribution is as follows,



Part 2

1. **Tell us how you validate your model and why you chose such an evaluation technique(s).**

   For model validation, I have used Stratified cross-validation. The major motive for using this technique over k-fold cross-validation is to make sure that each fold has equal distribution of both the classes as the target is skewed by a large margin. The ROC AUC was used as a scoring technique.

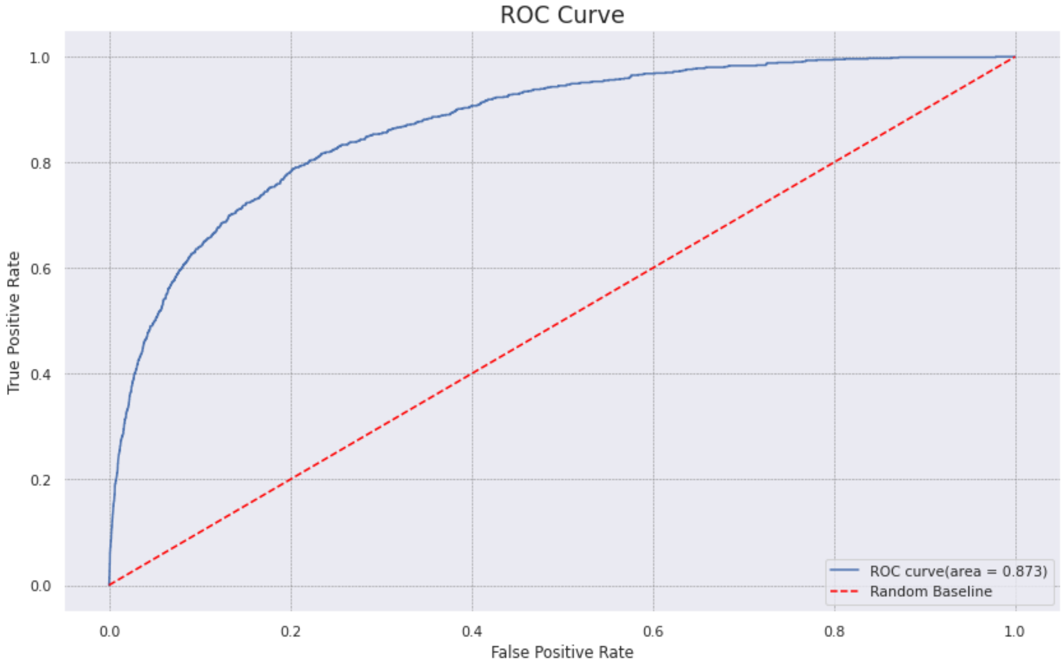


```
#stratified cross validation
base_model = GradientBoostingClassifier(n_estimators=250)
skf = StratifiedKFold(n_splits=10)
scores = cross_val_score(base_model, xtrain, ytrain, scoring = 'roc_auc', cv=skf)
scores
```

The process was followed for different values of K folds such as 5, 10, and 20. Each time the results were pretty similar where there were no traits of bias and variance. The results for each fold were close to each other.

2. **What is AUC? Why do you think AUC was used as the evaluation metric for this challenge? What other metrics do you think would also be suitable for this competition?**

   The main motive of the AUC - ROC curve is to monitor the performance of the model at different probability thresholds. ROC is a probability curve where the True Positive Rate and False Positive Rate are plotted against each other. AUC (Area Under the curve) represents how well the classes can be separated.



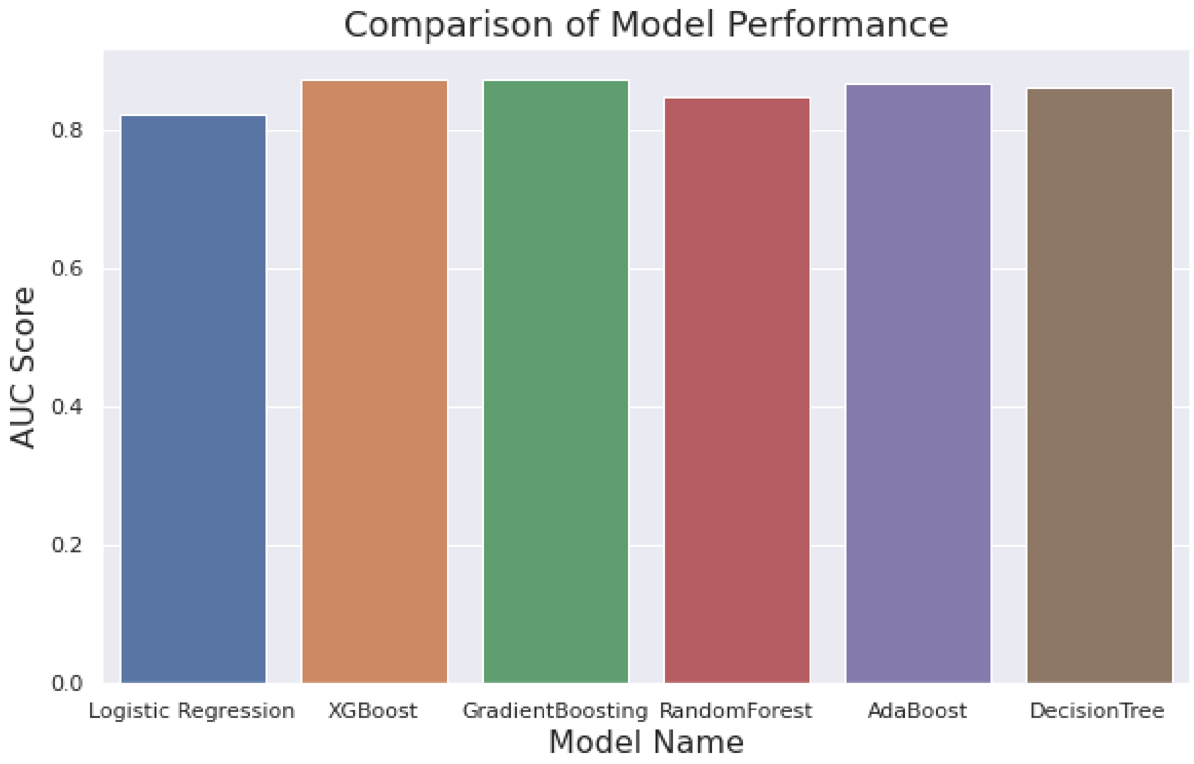
   AUC is used as an evaluation metric as it is not affected by the imbalance of the data like accuracy. Also, it can be considered a good metric when the probabilities of the classes are considered targets. Another metric that can be considered is Brier Score. For classification problems, we can use precision, recall, and F1 score.

3. **Short explanation of what you tried. What worked and what did not work**

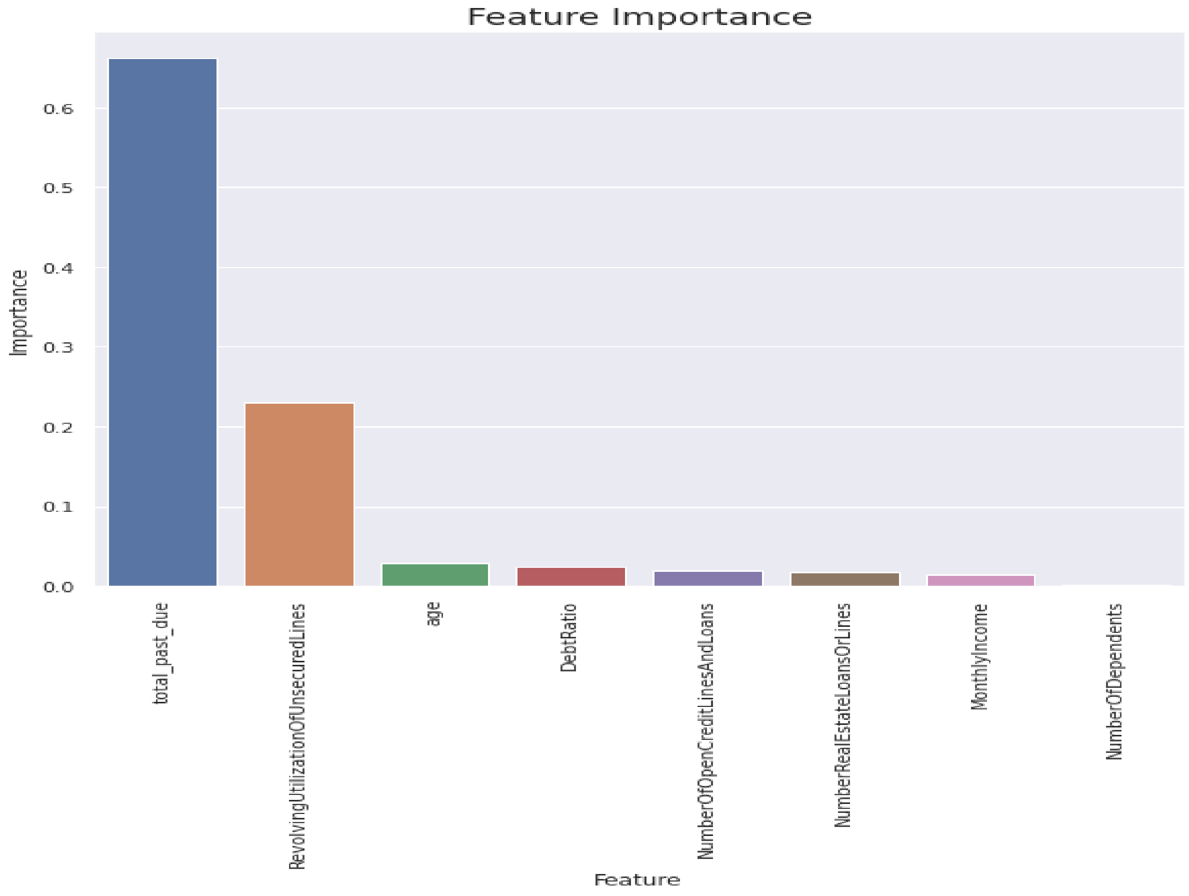
- Treating missing values as mentioned above in preliminary analysis.

- Kendall Correlation to find the features that are highly correlated with the target feature.
- Pearson correlation and Variance Inflation factor to check multicollinearity. Three variables showed signs of multicollinearity which were then combined(weighted summation) together as removing them dropped the performance of the baseline model.
- Plotting histograms, scatterplots, and boxplots helped to find inter-relations between independent variables.
- Treating outliers (all techniques didn't work),
    1. Trimming outliers
    2. Limiting them to the 10th and 90th percentile
    3. Limiting them to the min and max of the respective feature.
- Feature engineering techniques that didn't work:
    1. Combining different features such as multiplying monthly income with debt ratio to get the debt values. Several other combinations didn't work as well.
    2. Step-wise forward feature selection
    3. Principal component analysis
- Stratified Cross-validation to understand the stability and traits of bias and variance.
- Balancing the target feature(training dataset) was essential to increase bias. Implemented oversampling technique. SMOTE and Undersampling didn't work.
- Implemented 5-6 classification models with hyperparameter tuning.
- Stacking best classifiers and Keras ANN to predict probabilities. (Didn't work)

4. What insight(s) do you have from your model(s)?



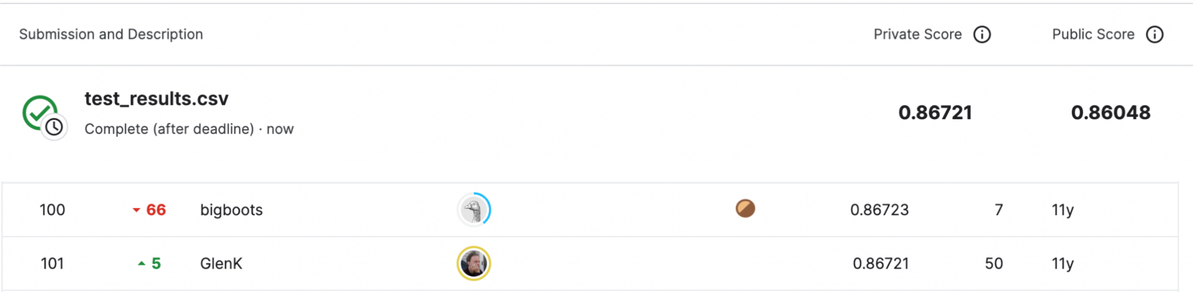
- Gradient boosting, Adaboosting, and XGboost were the top three best-performing models. I have computed the feature importance using the gradient boosting model as it was the best model,



- From the above bar plot, we can conclude that total due, a new variable formed to remove multicollinearity, is the highest contributing feature followed by the Revolving utilization feature.
- The importance of "NumberOfDependents" is negligible among all the independent features. I have also plotted the ROC curve plot which is mentioned in the above sections.

5. Can you get into the top 100 of the private leaderboard or even higher?

My current rank is 101.



Submission and Description	Private Score ⓘ	Public Score ⓘ
test_results.csv Complete (after deadline) · now	**0.86721**	**0.86048**

100	▾ 66	bigboots	0.86723	7	11y
101	▴ 5	GlenK	0.86721	50	11y