Capstone Project: Loan Default Prediction

Executive Summary:

The project uses Classification's model to predict customers who are likely to default on their loans and advise banks on important characteristics to consider when approving loans. To this end, the bank will use the guidelines of the Equal Credit Opportunity Act to help establish an empirically derived and statistically sound model for credit scoring. This model will be based on data obtained from applicants who have recently received credit through the existing loan underwriting process.

Problem Summary:

Most retail banks' profits come from interest in the form of home loans, usually borrowed by fixed-income or high-income customers. However, if a large number of defaulters bring a large number of non-performing loans (NPAs) to banks, this result will lead to a large number of bank profits being eaten up. As a result, some banks are beginning to approve loans from customers with more careful scrutiny. However, manual loan approval is inefficient, because it takes a lot of time in this process, and it is prone to some human errors or biases, which lead to wrong judgment or approval. However, with the advent of data science and machine learning, the focus has shifted to building machines that can learn this approval process and make it unbiased and more efficient. At the same time, it also ensures that the machine does not learn biases that have previously sneaked in due to the human approval process.

Solution Design:

First: As part of the solution design, exploratory data analysis (EDA) and visualization of the data were performed through univariate analysis to explore the range of values for each variable, as well as worthy central trends. Use bivariate analysis to find associations between variables. Finally, through multivariate analysis to explore the correlation between the data or clarify the structure of the data

Second: The data is classified and pre-judged through the decision tree. Figure 1 shows the pre-judgment result of the decision tree. From the figure, we observe that Debt-to-income (DEBTINC) is an important factor in judging whether a customer defaults on a loan. Through the model of the decision tree, it can be observed that if the customer does not have DEBTINC, the customer will default on the loan. If the customer's DEBTINC is greater than 43.745, the customer will default on the loan. If the customer's DEBTINC is less than 43.745 but the customer's number of delinquent credit lines is greater than 175.168, then the customer will also have the problem of defaulting on the loan.

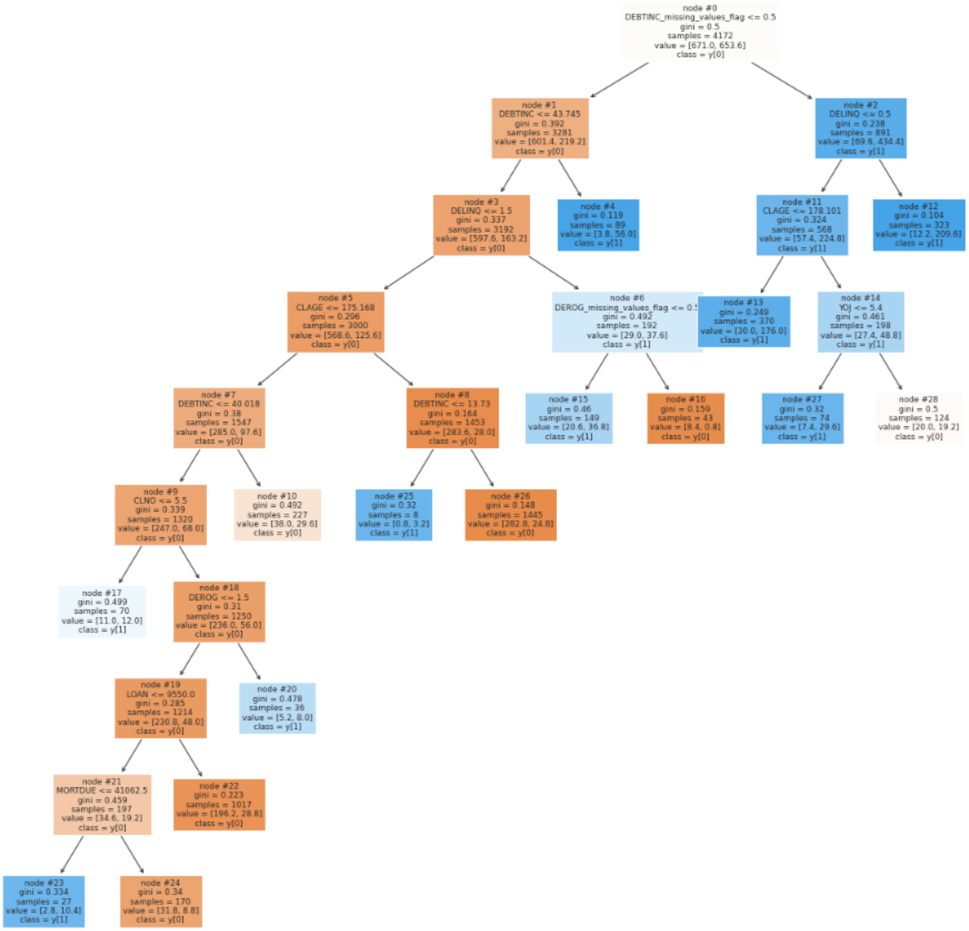


Figure One：Decision Tree

Create a random forest model, use this model to extract samples from the training data, and make decision tree predictions for each sample to find out how important each sample is. Through the analysis results of random forest in Figure 2, it can be seen that DEBTINC missing value is the most important, followed by DEBTINC and DELINQ.

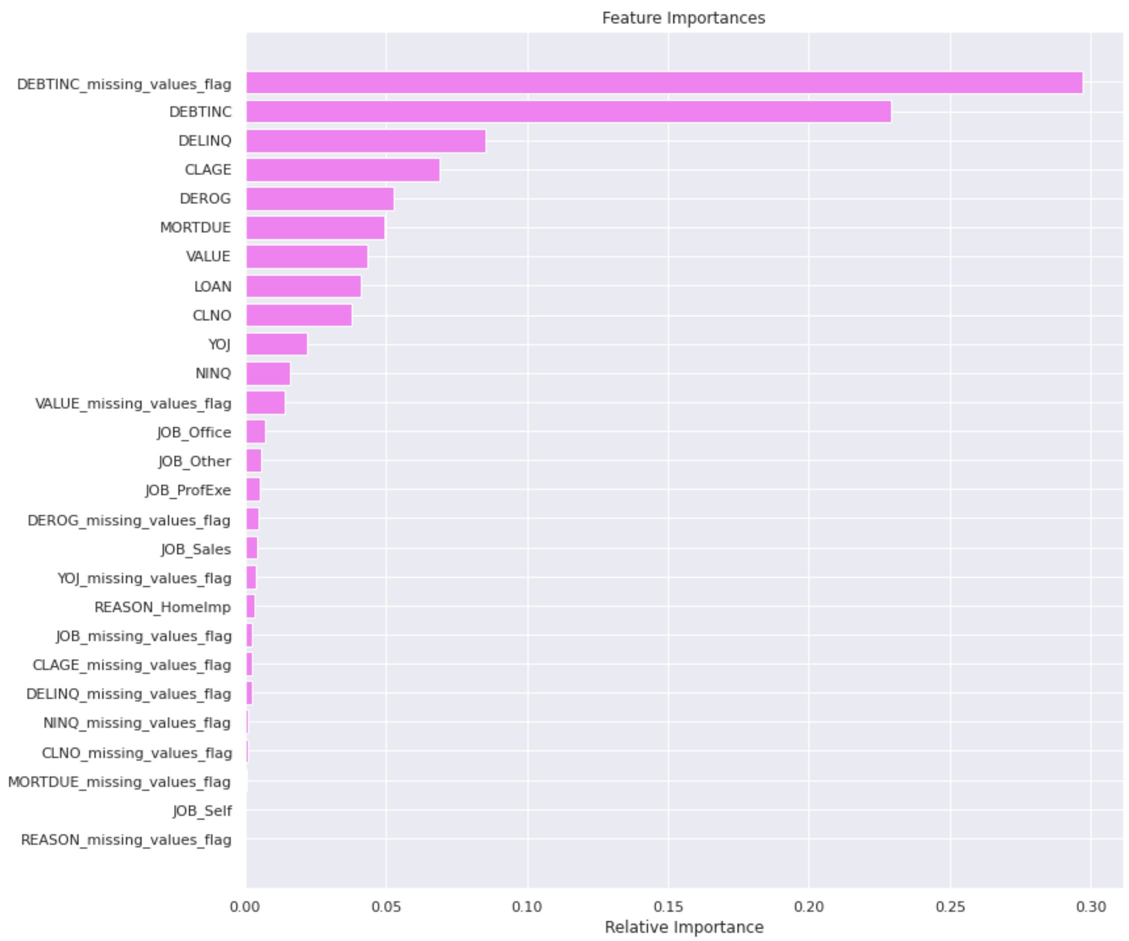
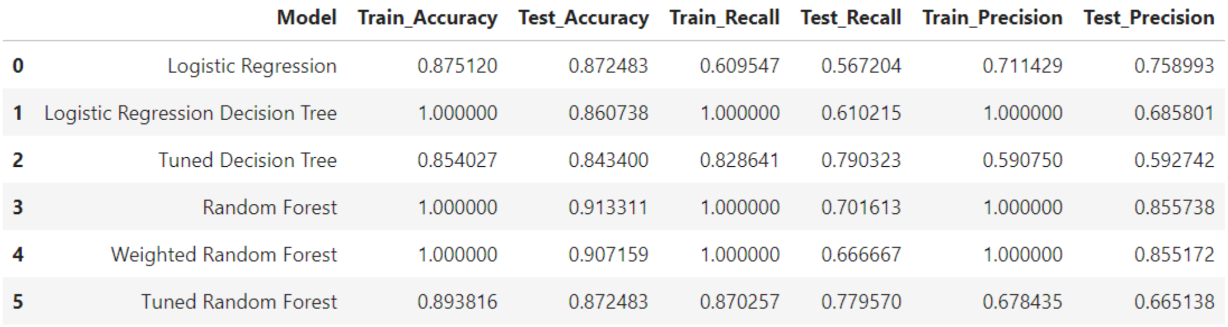


Figure Two：Feature Important

Third: By getting Accuracy, Recall, and Precision for six different model performances. After comparison, we can observe that although the Test Recall of Tuned Decision Tree reaches 79%, which is better than other models, but in Test Precision, Tuned Random Forest reaches 66%, which is far better than Tuned Decision Tree's 59%. So we can assume that Tuned Random Forest performs better than other models.



	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	Logistic Regression	0.875120	0.872483	0.609547	0.567204	0.711429	0.758993
1	Logistic Regression Decision Tree	1.000000	0.860738	1.000000	0.610215	1.000000	0.685801
2	Tuned Decision Tree	0.854027	0.843400	0.828641	0.790323	0.590750	0.592742
3	Random Forest	1.000000	0.913311	1.000000	0.701613	1.000000	0.855738
4	Weighted Random Forest	1.000000	0.907159	1.000000	0.666667	1.000000	0.855172
5	Tuned Random Forest	0.893816	0.872483	0.870257	0.779570	0.678435	0.665138

Figure Three：Comparing Model Performances

Recommendation for Policy:

Recommendation: The models suggest that DEBTINC and DELINQ are the most significant features that customers will default on loan. We can also analyze the rate of default on loans in different job categories. So I would suggest that banks can review the DEBTINC and DELINQ of customers of different job categories to reduce unnecessary default on loan customers

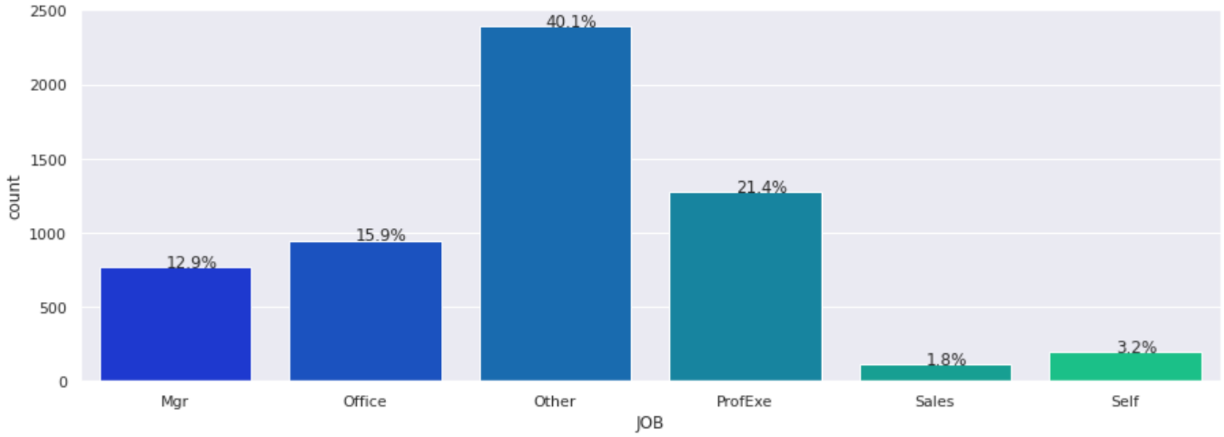


Figure Four：Job

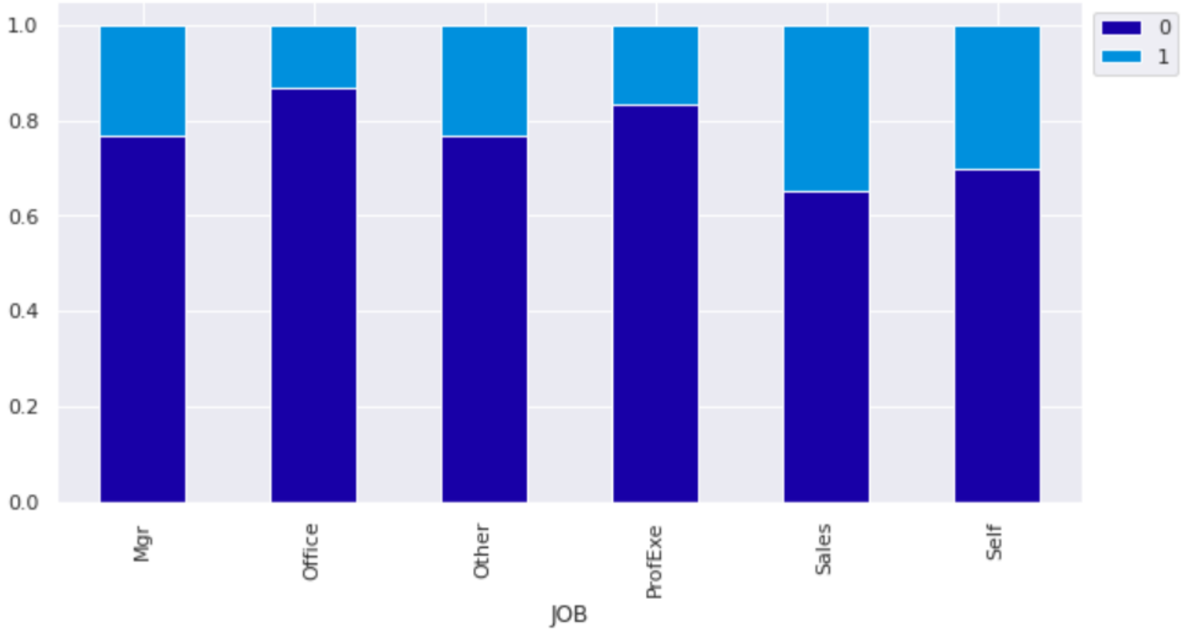


Figure Five：Default on Loan Job Type

Benefit：Through the visualization, it can be found that the work type of most clients is Other (40.1%, followed by ProfExe (21.4%). However, among these clients, the work type is Mgr (12.9%), Sales (1.8%) and Self (3.2%). The proportion of loan delinquency is quite small, but the situation of delinquent loans is the most serious. The proportion of customers of Sales and Self is very small and can be neglected for the time being, but the proportion of customers of Mgr who have delinquent loans is the same as the work type of Other. The proportion of customers who have delinquent loans is very close. This means that among the 12.9% of customers, a relatively large number of customers have delinquent loan problems. So I would think that banks can review Debt-to-income ratio (DEBTINC) and the Number of delinquent credit lines (DELINQ) to reduce delinquency issues.

Risk：Although the largest number of customers are in the job category Other, it is not known which jobs are included. A more in-depth review of loan eligibility for customers in the Other job category will cause a lot of inconvenience to many people who are just starting out, which will reduce some customers.