Capstone Project Loan Default Prediction

Executive summary

Loan default prediction is one of the most significant and crucial issues that banks and other financial organizations deal with since it has a significant impact on profit. Even though there are numerous established techniques for extracting data from loan applications, most of them appear to be underperforming given that there have been rises in the number of reported loan defaults. In this study, we chose the Gradient Boosting algorithm, among a number of machine learning models to forecast loan default.Gradient Boost model provides the best prediction scores and generates low error on data. Even if the model is the best predictor for Loan defaulting customers, a lot of other factors may increase our bias in future. There might be a lot of secondary variables which are not in our dataset and their effect on customers defaulting might be significant. The forecast is based on loan data from customers, which were previously approved for loan and later on defaulted with their loan payments. Data includes information from the loan application . Additionally, we provide crucial evaluation measures including
Accuracy, Recall, and Precision,ROC. On the confusion matrix we can compare the number the results of models.

Problem Summary

Finding a customer who is eligible to pay off the loan in future is a crucial process in the entire work process of any financial institute. The wrong and non-analytical approach to this process caused the Great Recession in 2008 and ended up with bankruptcy of the top financial instruments, millions of people and slowed down the world economy for a long time. Different Classification Solutions Loan Default Prediction problem has been evaluated for years and researchers come up with different approaches to maximize profit and minimize risk. The main models to predict the customers who are not eligible for future payments are Logistic Regression, Decision tree , Ensemble learning methods, Neural network models and etc. There are 2 main goals from solving this problem .

1. Minimizing the loss

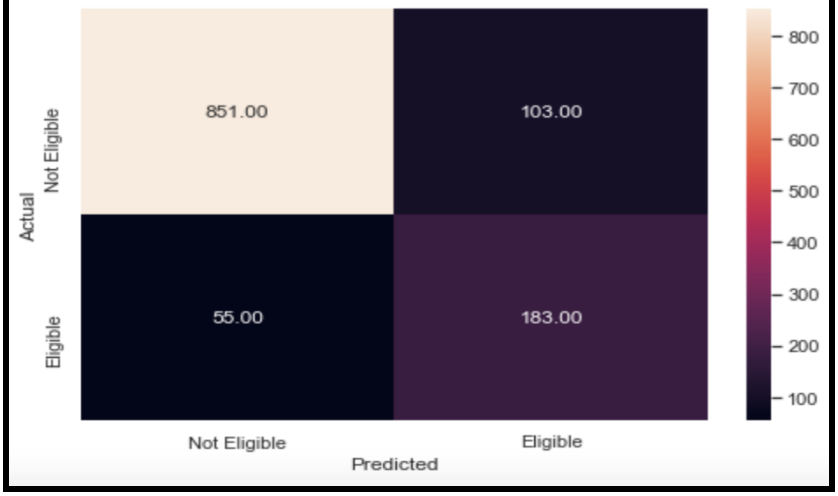
We can minimize loss by predicting the customers who are not eligible to pay off debts in future and can default on payments.

2. Maximize the profit
   Finding customers who have good capacity to pay and customer satisfaction. Rejecting or not recognising customers who won't default could lead to losing a lot of customers.

Solution Design

As we mentioned above, there are two points of view to improve filtering customers. The priority of our models was increasing RECALL score or minimizing False Positive, which means decreasing the number of approved customers by the model, who are not able to pay off loans in future.
Number of classification models were tested to find the method with best scores. Some models such as Logistic regression, Decision tree were able to detect the most amount of False Positives, but it led to Rejection of a lot of potential customers with good paying capacity . Even if our priority is to minimize loss, rejecting the majority of good customers could cause a reverse effect in the business of financial institutes.
Below matrix represents results of Decision tree.As we can see model could significantly decrease amount of False positives , from other side its hurting other customers who should be actually approved. We are losing a large amount of customers.
Because of that reason we decided to trade off between RECALL and PRECISION scores and could significantly improve score.

	precision	recall	f1-score	support
0	0.94	0.89	0.92	954
1	0.64	0.77	0.70	238
accuracy			0.87	1192
macro avg	0.79	0.83	0.81	1192
weighted avg	0.88	0.87	0.87	1192



DECISION TREE SCORE

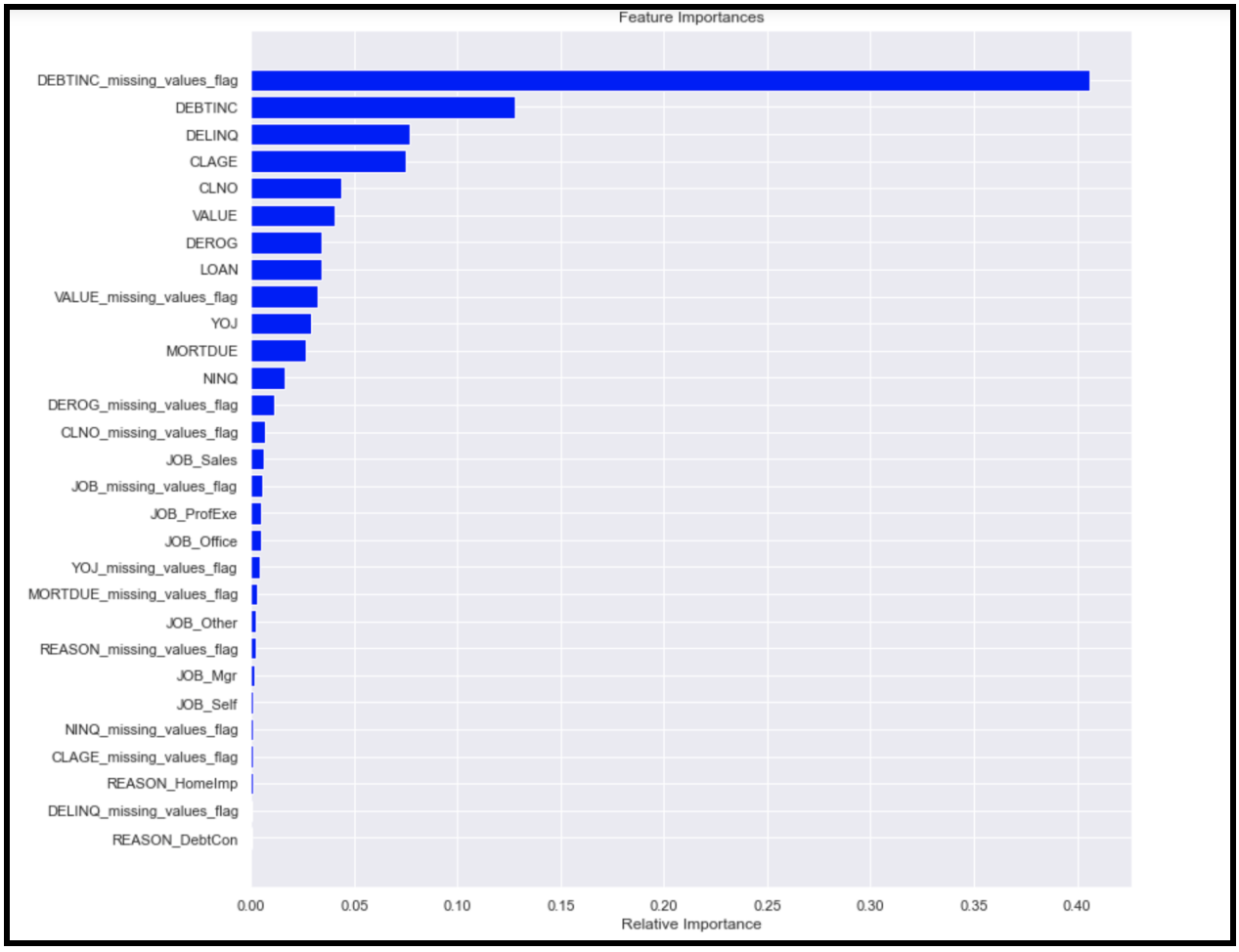
	precision	recall	f1-score	support
0	0.93	0.97	0.95	954
1	0.85	0.71	0.77	238
accuracy			0.92	1192
macro avg	0.89	0.84	0.86	1192
weighted avg	0.91	0.92	0.91	1192

GRADIENT BOOSTING SCORE

Gradient Boosting model Could generate quite an efficient model. As you can compare models, Gradient Boosting model could make balance between RECALL and PRECISION and improve overall ACCURACY score. My personal recommendation for loan default prediction problems is try to keep balance between scores, while giving some priority for RECALL score.

ANALYSIS AND KEY INSIGHTS

It's important that the model works well both on train and test data, so with gradient boosting has been provided an optimization process to increase the predictability of the model. Initially, the model was overfitting our data, but by using the RandomizedSearchCV optimizer , we could find the right parameters under which the model should be analyzed. After optimizer data started providing more clear scores and stopped overfitting. One of the important results of this study is to determine the important features that help the classifier to correctly predict loan default. This helps in business intelligence and decision making. Our model could generate the main factors on how to detect Defaulting Customers. On the chart below we can see the main feature importance of the factors. As we can see , if customers don't have or provide their Debt to Income share information, this is the main red flag to approve the loan. The missing value playing such an important effect in determining defaulting customers might be a specific case for our dataset, as it has over 0.55 importance to the entire outcome. Next important factors are DEBTINC and DELINQ.



Feature importance

LIMITATIONS AND RECOMMENDATIONS FOR FUTURE ANALYSIS

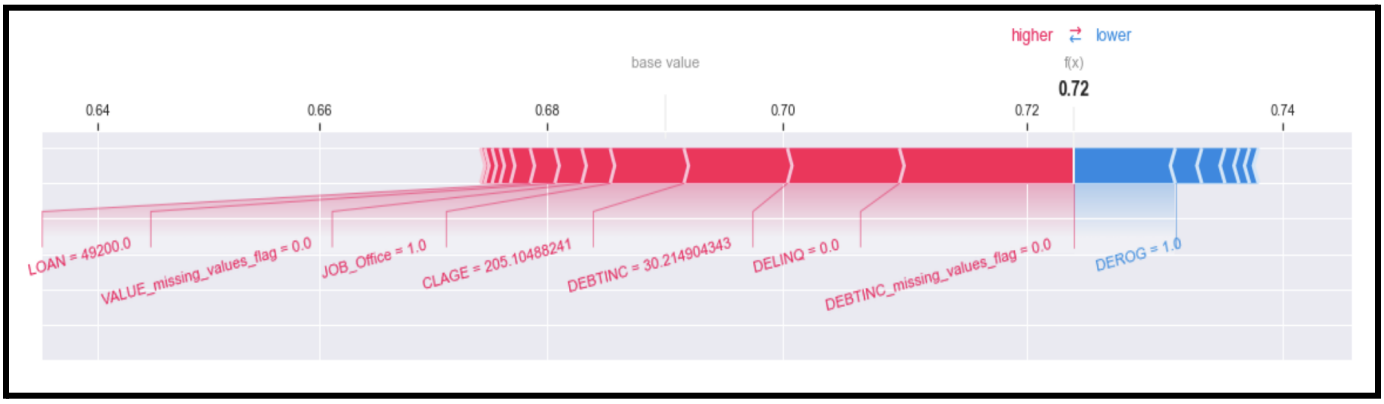
I think the data is missing variables which can really be the determinant of Loan Default problems. The variables as Credit score, interest rate, amount of downpayment and other variables could help improve efficiency of the future models. At the same time, maybe there are tools to determine the effect of secondary effects like unemployment rate, economic stability or inflation rate which are leading customers to default on any payments. At the same time , we are not sure if customers defaulted one payment or they defaulted on the entire loan.

RECOMMENDATION TO IMPLEMENTATION

Overall model is working the most efficiently on the dataset, so it's important to pay attention to the main factors and their effect on the overall result.

The most important factor to pay attention to is to check DEBT income share and if we are missing that information, the application for future loan should be rejected.

On the plot below we can see scenario more clear



Customers who had missed payments in the past have a high probability of Defaulting future payments .

Features with read are increasing predictability of models, so all the red features starting DEBTINC_missing_values_flag have significant effect in determining defaulting customers.

In order to improve credit payability of the future customers , I'd recommend that financial institutes get more data about customers and some meaningful information about them.

Our model isn't perfect and it has an error rate as well, if financial experts and advisors could go through model recommendations and give their approach about the model, a better implementation of the model might be reached.

Bibliography

- https://www.businessinsider.com/personal-finance/what-caused-the-great-recession

- Rising Odegua , Predicting Bank Loan Default with Extreme Gradient Boosting https://arxiv.org/pdf/2002.02011.pdf