Capstone Project: Classification

LOAN DEFAULT PREDICTION

Executive Summary

The proposed solution best suited for this project of predicting loan defaults is the Tuned Random Forest Model which had the best recall score among all the different models that were built. A primary target of the bank is to reduce the number of approved loans that eventually get defaulted because these become considerable losses that greatly affect their profits. The Tuned Random Forest model reliably achieves this reduction of false negatives.

Aside from predicting if a client would likely default or not, the factors/features from the client data that have the most effect on loan default can be identified by generating the features importance of this model. This information is then able to provide the basis for the justification of a specific decision arrived at, either the approval or the rejection of the loan. The output of this model shows that the 3 topmost features that have the most relevance to the likelihood of a loan default are the client's debt-to-income ratio, the age of his oldest credit line, and the number of his delinquent credit lines. The bank must then take note of these when they assess loan applications.

To improve the Tuned Random Forest Model, it can be tuned further to increase its recall score. When this is done, there is a possibility that a different set of variables/factors which are most relevant to loan defaults is generated. In addition, as enough new client data has been gathered, the model has to be trained and tested with the updated dataset so that it performs optimally as new loan applications are received. A challenge of this model is that it is not explicit as to the values of the features that would classify a client to be a credit risk. In this regard, the bank has to set their own cut-off values based on the data of their client base or they ca use the standards available within the banking industry, such as a debt-to-income ratio of no more than 36% to qualify for a loan.

Implementing this predictive model accomplishes the goal of having a loan approval process that is more efficient and with no human bias and errors that are commonly found in the manual loan approval method.

Problem and Solution Summary

Problem

The Bank's Consumer Credit Department wants to streamline its home equity line of credit (HELOC) approval process. Since interests from this type of loans make up a large proportion of the bank's profits, improving its current system is one of the bank's top priorities. Although the bank has an existing underwriting process, it is performed manually and is therefore complex and cumbersome and is accompanied by human biases and errors coming from the personnel tasked to evaluate the loan applications.

The bank consequently wants to employ a predictive machine learning model to facilitate its loan approval process. The bank will adopt the guidelines of the Equal Credit Opportunity Act and use available client data gathered from its current loan underwriting process to base this model on. A chief criteria is that the model must be interpretable enough to provide a justification for a decision made of either an approval or a rejection of a HELOC application. In addition, a foremost consideration to be factored in is the risk of loan defaults. The bank considers unpaid loans as a major loss. With this in mind, the model has to be designed so as to lessen the probability of false negatives (approving loans that turn out to be getting defaulted).

Solution

Different machine learning models were built in this project. Since loan defaults is the bank's paramount concern because losses from these eat up a substantial chunk of their profits, the performance metric that must be of prominent focus is the recall score. This performance metric has to be maximized since the greater the recall

score, the lower the false negatives. The model with the highest recall would assist the bank in decreasing their losses due to approved loans that eventually turn out getting defaulted.

Table of Model Performances

Machine Learning Model	Test_Accuracy	Test_Recall	Test_Precision
Logistic Regression	0.657159	0.627451	0.318182
LR w/ Optimal Threshold	0.746085	0.347339	0.359420
Decision Tree	0.867450	0.610644	0.689873
Tuned Decision Tree	0.836689	0.764706	0.567568
Random Forest	0.909396	0.652661	0.859779
RF w/ 20:80 Class Weights	0.899329	0.613445	0.839080
RF w/ CW of balanced subsample	0.895973	0.610644	0.822642
Tuned Random Forest	0.721477	0.803922	0.401399

Figure 1: Table of Performances of Machine Learning Models

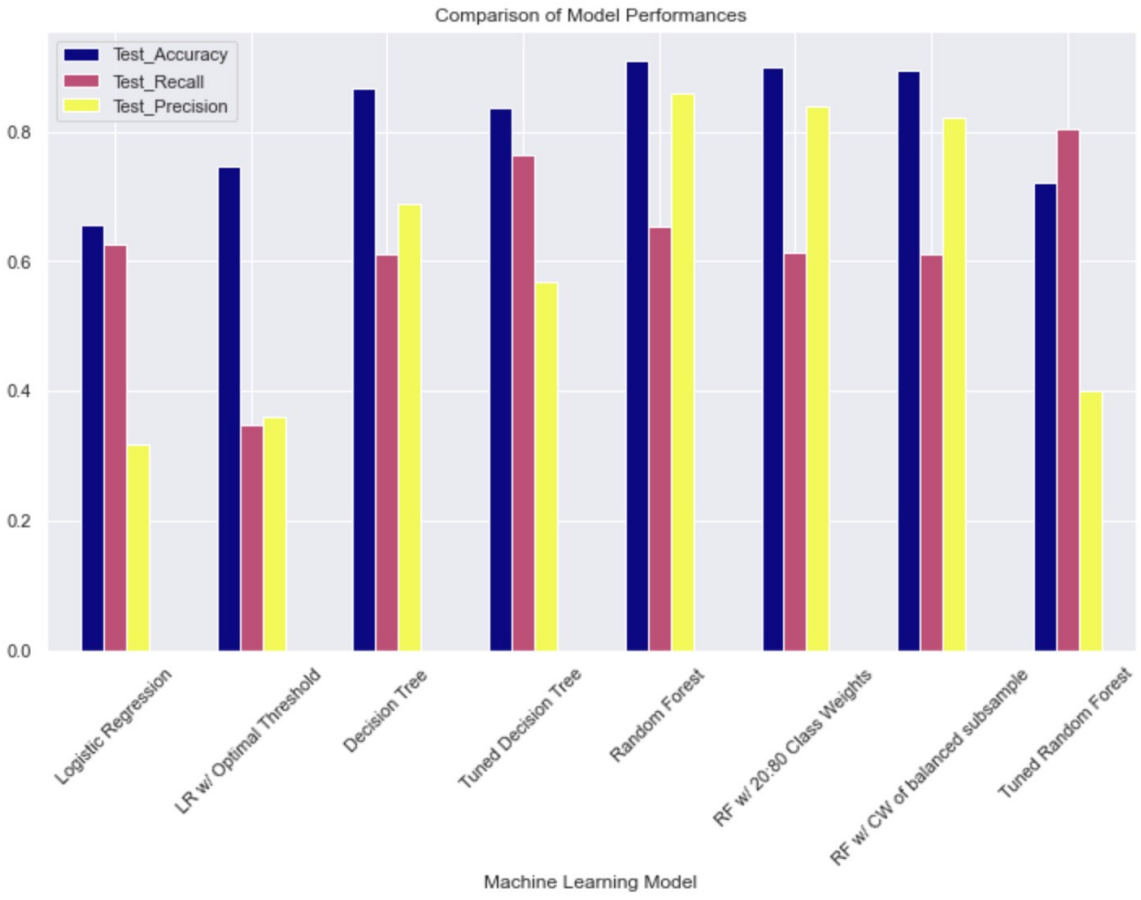

Figure 2: Comparison of Model Performances

The Tuned Random Forest model gave the best performance as it had the highest recall score and is thus proposed to be the machine learning model to be adopted by the bank's consumer credit department for their loan default prediction in their HELOC approval process. It is the leading solution to come up with the best decision for a particular loan approval scenario. Moreover, because of its feature importance capability, the model is interpretable enough to ascertain the important factors that have the most impact on the credit score of a client which determines his eligibility for a HELOC loan. Thus, justification is provided by looking at these important features that have been most relevant to the cause for the rejection or the approval of the loan.

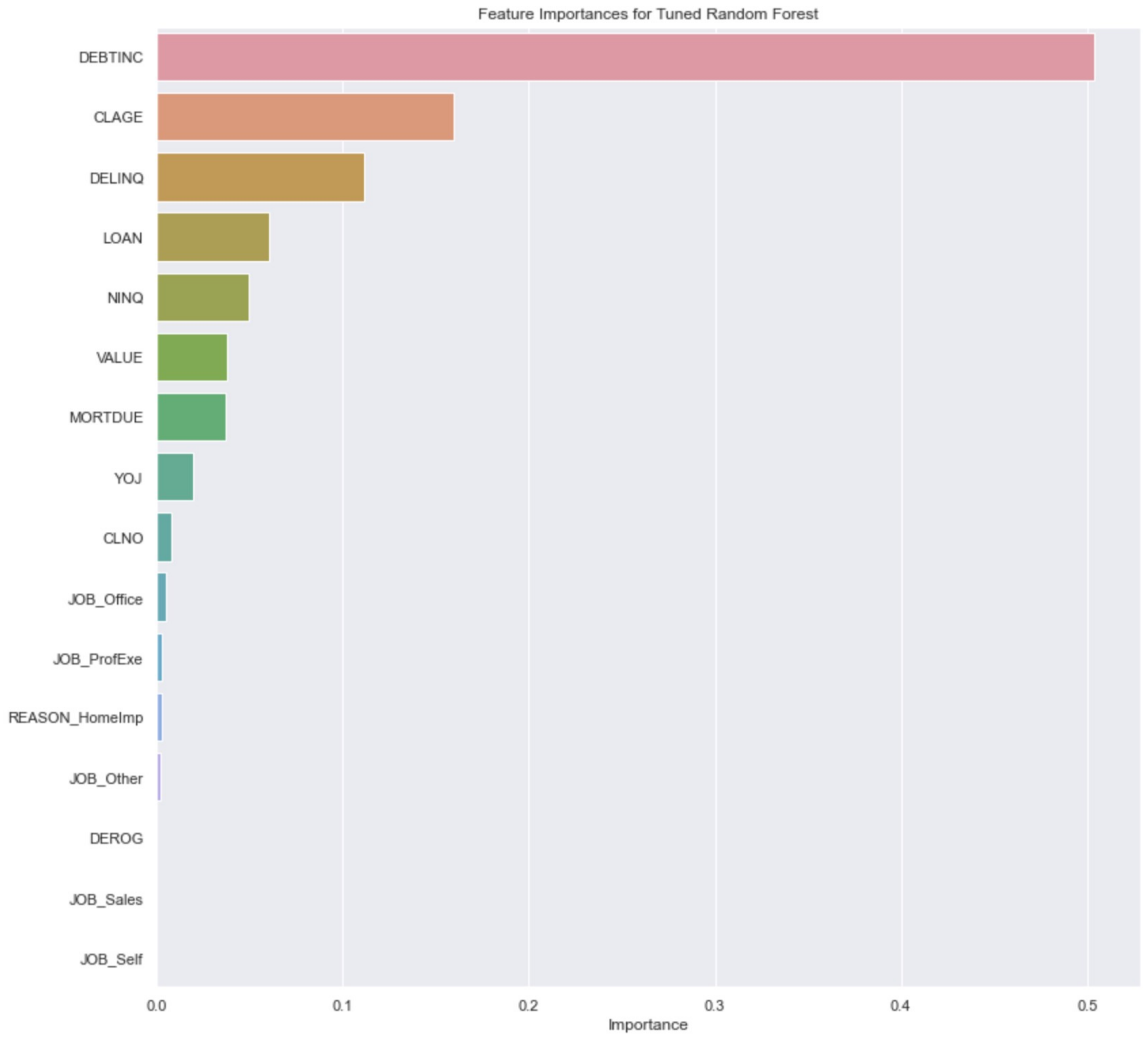

Figure 3: Features Importances of Tuned Random Forest Model (refer to Appendix for Data Description)

From the above plot, it shows that the client's debt-to-income ratio (DEBTINC), the age of his oldest credit line (CLAGE), and the number of his delinquent credit lines (DELINQ) are the topmost 3 features that have the greatest impact to the likelihood of a loan default. DEBTINC is outstandingly the most important feature and, as such, the bank must give due attention to this factor when appraising loan applications.

Given then a set of information about a client and based on this available client data, this model has the capability to predict if the client is likely to default on the loan or not and also give the basis for the prediction in accord with the client's data. Hence, the Tuned Random Forest model would be able to facilitate the bank's

decision-making process in HELOC approvals and at the same time reduce the losses that they would have incurred due to erroneous loan approvals that turn into loan defaults.

The Tuned Random Forest model that was produced in this project can be improved upon by adjusting the values of the hyperparameters and/or adding more hyperparameters and tune the model to further increase its recall score. This tuning also affects the values of the features importances and therefore may result in a different set and prioritization of the variables that most affect the loan default rate which the credit department must take note of when justifying a HELOC decision.

Furthermore, as new data is obtained from new loan applicants who have been given credit, the model has to be trained, tested, and tuned using this new conglomeration of client data so as to assure that the credit department is always using the optimal version of the predictive model.

Recommendations for Implementation

The Tuned Random Forest model is an effective, efficient, and interpretable machine learning model that has the ability to determine the credit-worthiness of a HELOC applicant by predicting if he is likely to default or not on a loan and provide the basis for the prediction, which answers the question as to what client features/data are most important to consider during the HELOC approval process.

The Tuned Random Forest that was arrived at in this project can still be tuned further before a final model is adopted. The solution can still be designed optimally with the goal of achieving a higher recall score that significantly decreases even more the false negatives. This current Tuned Random Forest model that was produced in this project can be improved upon by adjusting the values of the hyperparameters and/or adding more hyperparameters and tune this model to further increase its recall score. And as the model is tuned and improved, the feature importances accordingly change, and therefore a different set of important factors that affect loan default rates can be generated as output from the final model that would be the basis of the justification for a loan approval or rejection. All these aim to attain the bank's target to have a more efficient and effective HELOC approval system in place.

As enough new data is gathered from applicants who are being given credit, the Tuned Random Forest Model must be trained, tested, and tuned using the updated dataset. The performance metrics must be assessed accordingly and aiming for a high recall score must be the ongoing goal. The various variables that affect loan default changes in importance. What was relevant during a previous time period may no longer be valid for the new time period and this is dependent on the updated dataset.

With the current Tuned Random Forest model that has been built in this project, the top 3 features that have the most effect on loan default rates are the following:

- DEBTINC (debt-to-income ratio): it has the most impact on the likelihood of loan defaults out of all the features. Debt-to-income ratio is the percentage of a client's gross monthly income that is used to pay his monthly debt and determines his borrowing risk. During the evaluation of a client's application, the bank has to give most attention to this variable, bearing in mind that the higher the debt-to-income ratio of a client, the more the likelihood of him defaulting on his loan.

- CLAGE is the age of the client's oldest credit line in months and the presence of older credit lines means that it takes the client longer to pay his debts and this can imply a difficulty on his part in paying his loans.

- DELINQ is the number of delinquent credit lines that a client has and the more of this that he has, the harder it is for him to make a payment on his loan which increases the danger of his failure to pay.

Having information such as these, the bank can pinpoint what variables about the client most likely would lead to his loan application being approved or rejected. Deducing the importance of each of the features using this

machine learning model helps the bank determine the credit-worthiness of a client. Their decision would then be supported by logical and sound reasoning.

With the emergence of data science, it is only prudent for banks to be proactive and employ machine learning tools and incorporate these in their underwriting process. The complexity due to the multi-dimensional client data can be handled with ease by this model which makes this system more efficient plus human biases and errors are removed from the equation.

It must be noted that since the precision score decreases as the recall score is increased, the number of false positives increases. The losses that are due to false positives (rejecting a loan application when the client turns out to not be a defaulter) is far outweighed by the losses from false negatives. And implementing the Tuned Random Forest model will aid the bank in reducing these losses from false negatives. The model does not completely eradicate the problem of false negatives but it will certainly help in reducing its occurrence. The bank can also decide on how best to balance the respective effects of precision and recall, such that precision is not totally neglected in the process.

Also, one thing that the bank must try to attain is to have a more comprehensive data gathering process so that the model can perform more accurately. The current dataset had missing values which were addressed thru imputation.

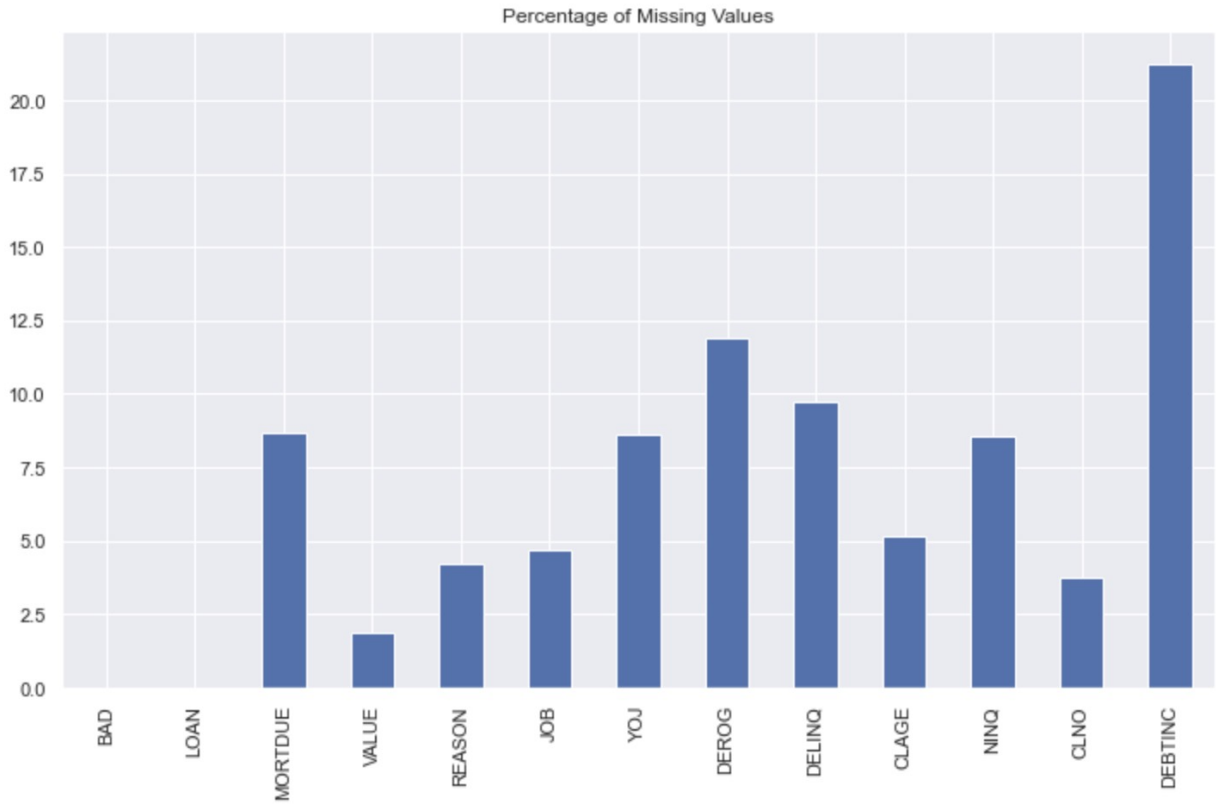

Figure 4: Percentage of Missing Values per Feature

As the above plot shows, DEBTINC has the greatest percentage of missing values and DEBTINC happens to be the most important feature that affects the likelihood of loan default the most. If a more complete dataset is available, the model would be able to produce an output that more resembles actual conditions and therefore be more accurate.

A challenge of this Tuned Random Forest model is that it is not explicit as to the values of the features that would classify a client to be a credit risk. For example, the model does not specify the cut-off value of the debt-to-income ratio that would classify a client to be a probable defaulter or not. In this regard, the bank has to set their own cut-off values based on the data of their client base or use the standards available within the banking industry, such as debt-to-income ratios of no more than 36% to qualify for a loan.

Reference

www.investopedia.com

Appendix

Data Description

The Home Equity dataset (HMEQ) contains baseline and loan performance information for 5,960 recent home equity loans. The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. This adverse outcome occurred in 1,189 cases (20 percent). 12 input variables were registered for each applicant.

- **BAD:** 1 = Client defaulted on loan, 0 = loan repaid
- **LOAN:** Amount of loan approved.
- **MORTDUE:** Amount due on the existing mortgage.
- **VALUE:** Current value of the property.
- **REASON:** Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- **JOB:** The type of job that the loan applicant has, such as manager, self, etc.
- **YOJ:** Years at present job.
- **DEROG:** Number of major derogatory reports (which indicates a serious delinquency or late payments).
- **DELINQ:** Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due).
- **CLAGE:** Age of the oldest credit line in months.
- **NINQ:** Number of recent credit inquiries.
- **CLNO:** Number of existing credit lines.
- **DEBTINC:** Debt-to-income ratio (All of your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.)