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. The other models that were built did not perform as well as this model. It is to be noted that the client dataset that was available in building these models had missing values for which imputation was carried out to fill these in. It would benefit the bank to have a more comprehensive data gathering system to gain more accuracy and a more optimal predictive model for their loan services.

Aside from predicting if a client would likely default or not, the factors/features from the client data that have the most effect on the likelihood of a 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 top 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 take note of the prioritization of the features as they assess loan applications.

To improve the Tuned Random Forest Model, further tuning can be undertaken to increase its recall score. While doing this, accuracy and precision must not be neglected but a balance of the effects of these metrics must be arrived at based on the bank's overall vision for their loan offerings. The bank has to take into account how much false positives (precision) can they forego. When tuning is performed, there is a possibility that a different set of variables/factors most relevant to loan defaults is generated and the bank personnel must remember and heed these as they appraise loan applications. 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. There is a tradeoff between model performance and model interpretability, but since the foremost concern of the bank is loan defaults, the Tuned Random Forest model is the better option than the Decision Tree which may have better interpretability for it is able to provide decision rules but gives a lower recall score. The challenge of this model then 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 can use the standards available within the banking industry, such as, for example, adopting a debt-to-income ratio of no more than 36% to qualify for a loan.

Implementing this predictive model accomplishes the bank's goal of having a better loan approval process that is simplified, more efficient, faster, and with no human biases 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 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

Exploratory data analysis was first undertaken and it showed a dataset having 5,960 observations and 13 columns/features. The loan default rate is at 20%. Clients who repaid their loans have older credit lines than those who defaulted. Clients who defaulted on their loans had more recent credit inquiries, more existing credit lines, and had higher debt-to-income ratio compared to the clients who repaid their loans. Different machine learning models were built in this project. There were outliers present in the client dataset and so they were treated for the Logistic Regression model. The Decision Tree and Random Forest models are not sensitive to outliers. There were also missing values in the client dataset which were imputed with median for the numeric columns and mode for the categorical ones.

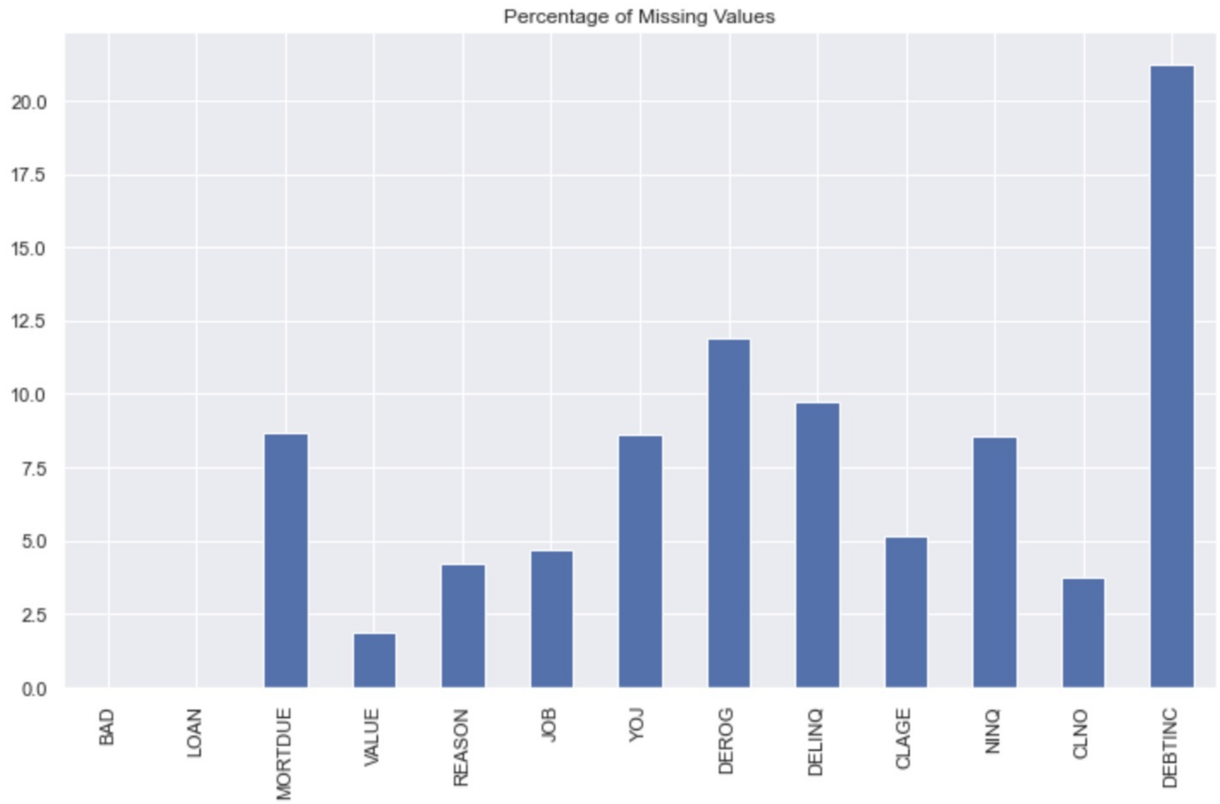


Figure 1: Percentage of Missing Values per Feature

As the above plot shows, DEBTINC has the greatest percentage of missing values and DEBTINC happens to be the most important feature, as will be shown later, 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.

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.

With the problem being about classifying a client whether he is likely to default on his loan or not, the following classification algorithms were considered: Logistic Regression, Decision Tree, and Random Forest. There are other classification models that are available but, at this project’s stage, the above 3 are being considered. In addition, optimization via tuning was performed to come up with better models from these basic forms. Below is the outcome of the performances of these different models that were built.

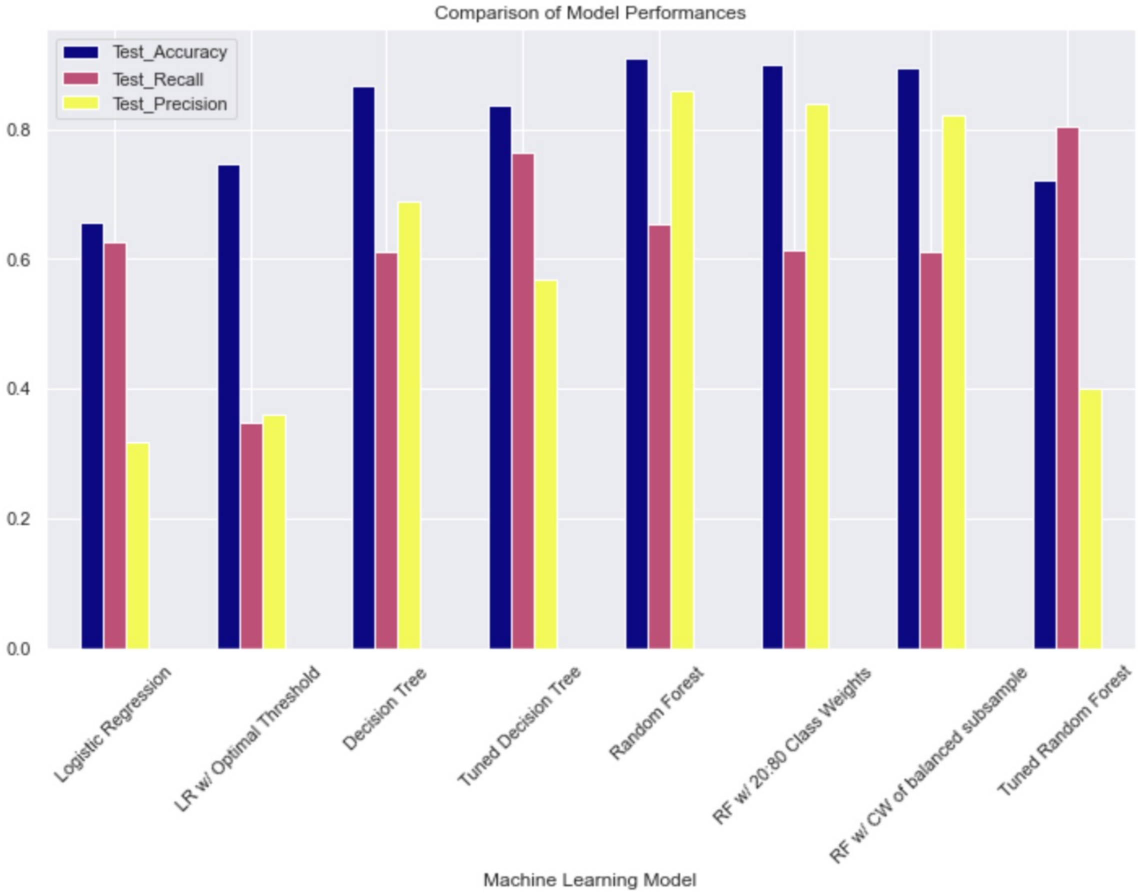
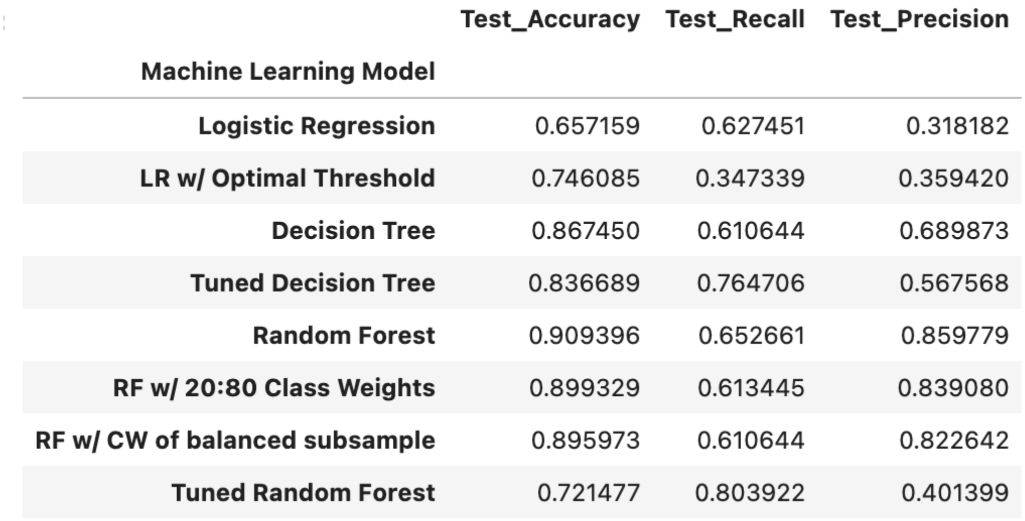


Figure 2: Comparison of Model Performances

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 3: Table of Performances of Machine Learning Models

The Tuned Random Forest model gave 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. The scores for accuracy and precision do warrant further tuning of the model. The hyperparameters can be adjusted and/or other hyperparameters can be added to come up with the best balance between the effects of the respective values of recall, precision, and accuracy that would address the entirety of the bank's needs.

The Tuned Random Forest model is the leading solution among the models to furnish the best decision for a particular loan approval scenario. Moreover, because of its feature importance capability, the model is interpretable enough to establish the important factors that have the most impact on the credit score of a client which determines his eligibility for the HELOC loan. Thus, justification is provided by looking at these important features that have been most relevant in determining his credit worthiness and thus the cause for the rejection or the approval of his loan application.

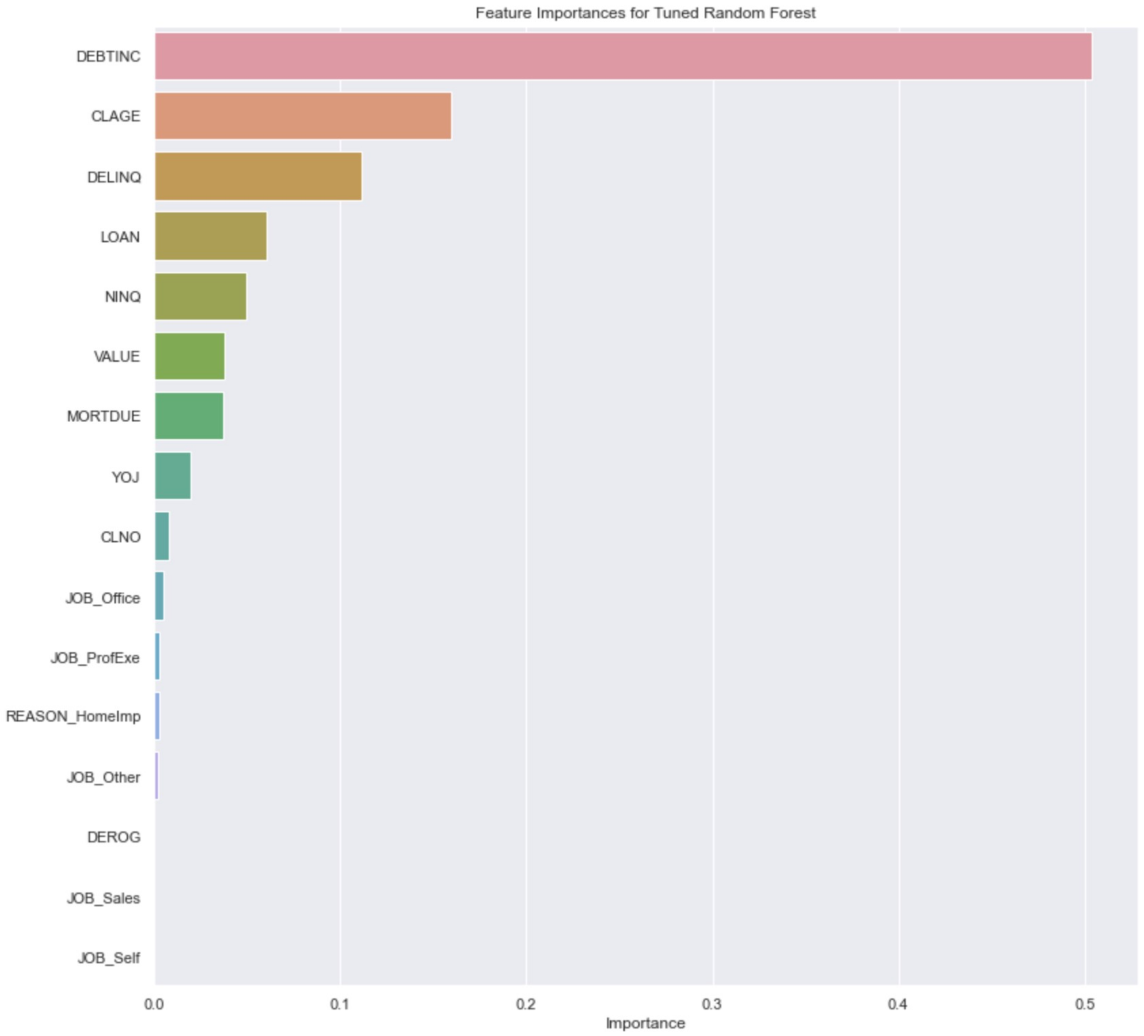


Figure 4: Features Importances of Tuned Random Forest Model (refer to Appendix for Data Description)

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 top 3 features that have the greatest impact on the likelihood of a loan default by a client. DEBTINC is outstandingly the most important feature and, as such, the bank must give prime attention to this factor when appraising loan applications. After DEBTINC, the other factors are accordingly prioritized commensurate with the value of their importance as revealed by the plot.

This thus gives the bank a guide on what important factors to look at as they carry out the decision-making aspect of their loan approval process.

There is a tradeoff between model performance and model interpretability, but since the foremost concern of the bank is loan defaults, the Tuned Random Forest model is the better option than the Decision Tree which may have better interpretability for it is able to provide decision rules but gives a lower recall score. The feature importance capability of this model is not able to provide explicit values as to the cut-off of a feature that would classify a client to likely default or not. For example, it could not supply the debt-to-income ratio value above which would classify a client to be a likely defaulter, and below which the client is likely to be credit-worthy. The bank may set these values based on their client base and/or using the banking industry standards.

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 to give the basis for the prediction in accord with the respective 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 the bank would have incurred due to erroneous loan approvals that turn into loan defaults.

When the Tuned Random Forest model that was produced in this project is improved upon by adjusting the values of the hyperparameters and/or adding more hyperparameters and tuning the model further, this tuning 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 and the bank must be mindful of this while a HELOC decision making is being undertaken.

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.

When adopted, the Tuned Random Forest model is beneficial to the bank as it will facilitate the loan approval process. The decision-making process becomes more simplified, more efficient, faster, and free from human biases and errors. This streamlined system predicts clients who are likely to default on their loans and gives recommendations to the bank on the important features that need to be considered while approving a loan.

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. Accuracy and precision must not also be neglected but must be given due regard based on the overall goals of the bank. 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 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 may possibly be generated as output from the final model that would be the basis for 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 accordingly be trained, tested, and tuned using the updated dataset. The performance metrics must likewise be assessed and aiming for a high recall score must be the ongoing goal whenever the objective is to lower the chances of loan defaults. Because of this updated dataset which the model will use to learn from, the various variables that affect loan default changes in their respective importance value. 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 from the exploratory data analysis, it showed that the clients who repaid their loans have older credit lines than those who defaulted.

- 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 affect his credit-worthiness and 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 client's probability of defaulting on his loan. Their decision would then be supported by logical and sound reasoning. As loan applications are received, the bank has a guide as to what features they have to be mindful of when they approve loan applications.

It must be noted that since the precision score decreases as the recall score is increased, the number of false positives increases. The bank does consider that losses due to false positives (rejecting a loan application when the client turns out to not be a defaulter) are 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, recall, and accuracy such that precision and accuracy are not totally neglected in the process. The current Tuned Random Forest model can be tuned more to achieve this end.

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

There is a tradeoff between model performance and model interpretability, but since the foremost concern of the bank is loan defaults, the Tuned Random Forest model is the better option than the Decision Tree which may have better interpretability for it is able to provide decision rules but gives a lower recall score.

This Tuned Random Forest model 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.

Overall, the Tuned Random Forest model is the best choice for the bank to adopt for its target of implementing a machine learning model for its loan default prediction process to revamp its current manual method. 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. When adopted, the Tuned Random Forest model would streamline their loan approval process. The decision-making process becomes more simplified, more efficient, faster, and free from human biases and errors. The bank definitely would benefit from a method that facilitates predicting clients who are likely to default on their loans and giving recommendations on the important features that need to be considered while making a decision during the loan approval process.

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.)