Anita Fong
Practical Data Science/Classification

Loan Default Prediction

Executive Summary

Problem Summary

—

Retail bank profits rely heavily on the interests they earn from home loans. The loans are borrowed by regular income/high earning clients. But banks are also fearful for defaulters as bad loans are a huge loss in profits so banks are judicious while approving loans for their customer base. The approval loan process is extremely effort-intensive, multifaceted, and prone to human error/biases.

Our objective is to build a machine that can learn this approval process and make it more efficient and free of biases. The machine must not learn the same biases that humans make. We need to build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan.

Proposed Solution

—

Based on the Home Equity dataset from the consumer's credit department, we will build a prediction model based on approved people with loans that either repaid or defaulted on it. We will look at what key important factors to review when approving someone for a loan.

There were various models that were tested and the best performing one was the tuned random forest model.

Random forest is a type of decision tree model so it is easy to understand and easily interpretable. It reduces the overfitting in decision trees through the method of taking averages of multiple independent decision trees and there's no need for pruning. At the same time, it has all the pros in a decision tree which are less misclassification than logistic regression, easy to visualize, mirrors human decision making more closely, and requires little data preparation. It also outlines the important features we are looking to review when approving a loan along with a high recall and precision score.

Key Takeaways and Next Steps

—

The debt to income ratio seems to be the most important feature but it is also the variable that has the highest missing values. DELINQ, CLAGE, YOJ, VALUE, DEROG, CLNO, LOAN, MORTDUE, NINQ also hold some importance but there are many missing values in those variables too.

We suggest to the business to make all fields mandatory so the bank can get a full view into the person's finances and be able to make an even better, more confident prediction.

From the data, it seems people with low DEBTINC ratio, low DELINQ, high CLAGE would be great candidates for the loan because the model suggests these people would repay back their loans.

Since there were so many missing values, we would like for the bank to make fields mandatory so we can re-analyze and reassess to see if we draw the same conclusions with the tuned random forest model. Would DEBTINC feature still be the most important feature or would the other variables hold more importance than before? Those are

some next steps we would take and further validate or re-explore what features are important for the bank to make the loan decision.

Data Exploration

Data Description

—

We were provided the Home Equity dataset from the existing loan underwriting process on recent applicants who had been approved for the credit.
The dataset contains the loan performance information for recent home equity loans on 5,960 clients. The loan defaulting occurred in 1,189 cases (20 percent). The target (BAD) is a binary variable where 1 = “Client defaulted loan” and 0 = “loan repaid” along with 12 other input variables regarding 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 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 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.

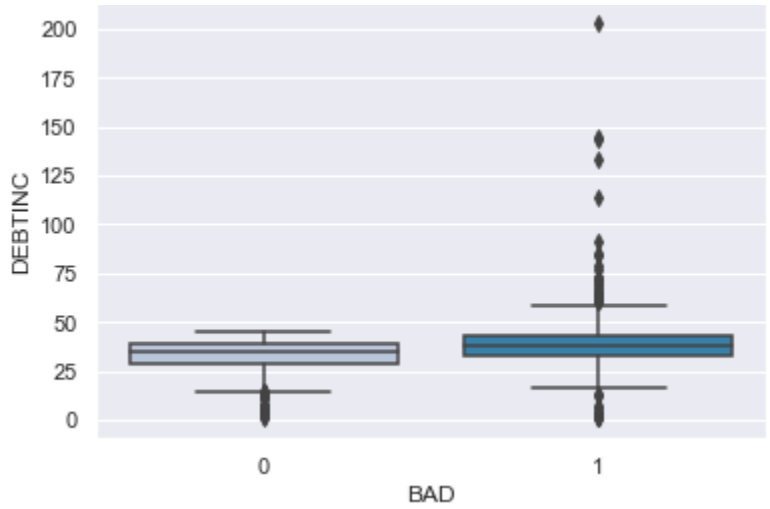
Data Quality

There are lots of missing values in this dataset with 21% of DEBTINC values missing, 11% of DEROG, 9% of DELINQ and more. Almost all variables have some percentage of missing which makes it challenging to analyze because we can't drop the variables since they might have very important information. We can fill the missing values with various methods but there is still difficulty in accurately inferring the missing values and it is still not as accurate as if we were to have real inputs.

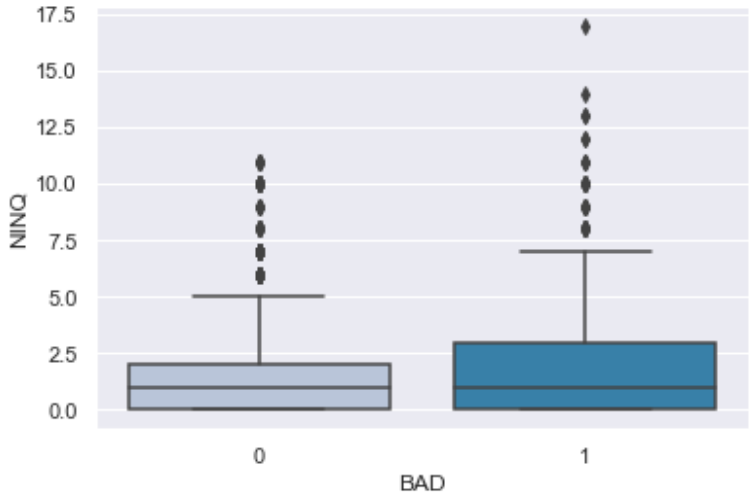
Initial Exploratory Data Analysis

1. There are a good amount of variables that have missing values - DEBTINC has the most missing values so that variable data may not be useful for the analysis
2. Majority of people have zero derogatory reports and delinquent credit lines with a few outliers
3. Majority have 0-1 recent credit inquiries with a few trailing outliers
4. People who repaid their loans seem to be a wider distribution of a large number of existing credit lines. There's more big outliers, some reaching to 70 number of existing credit inquiries for people who repaid their loans vs people who defaulted
5. Correlation heatmap shows that there's not much correlation anywhere other than MORTDUE & VALUE which makes sense. All other variables are very low.
6. People with Sales jobs have the highest percentage of loans being repaid, followed by Self and Manager. People with office jobs seem to have the highest percentage of defaulting on a loan. Job type might have some effect on loans being repaid
7. People with higher CLNO seem to default on their loans as you can see on the graph, there are many high CLNO outliers that defaulted. Same for high DEBTINC and NINQ (graphs below)

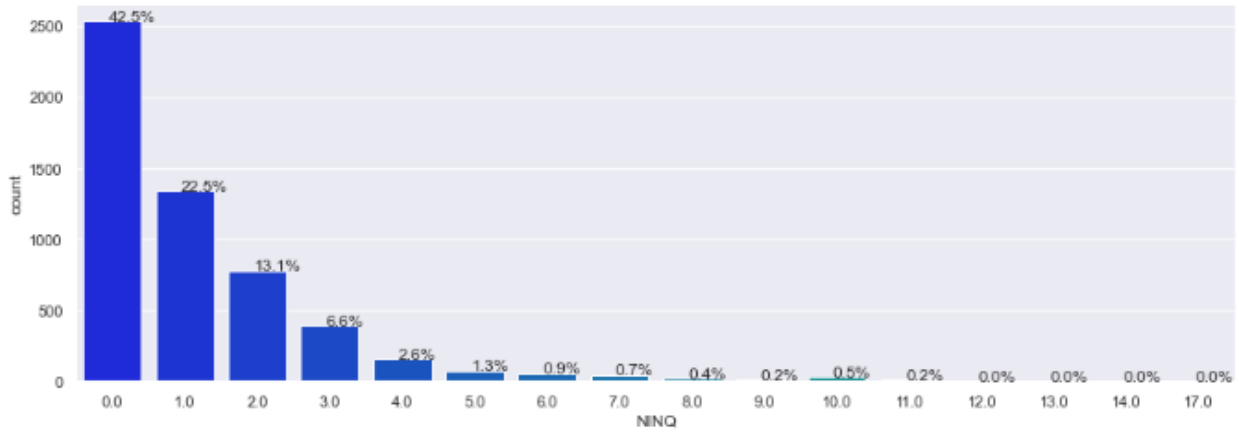
Exploratory Data Analysis



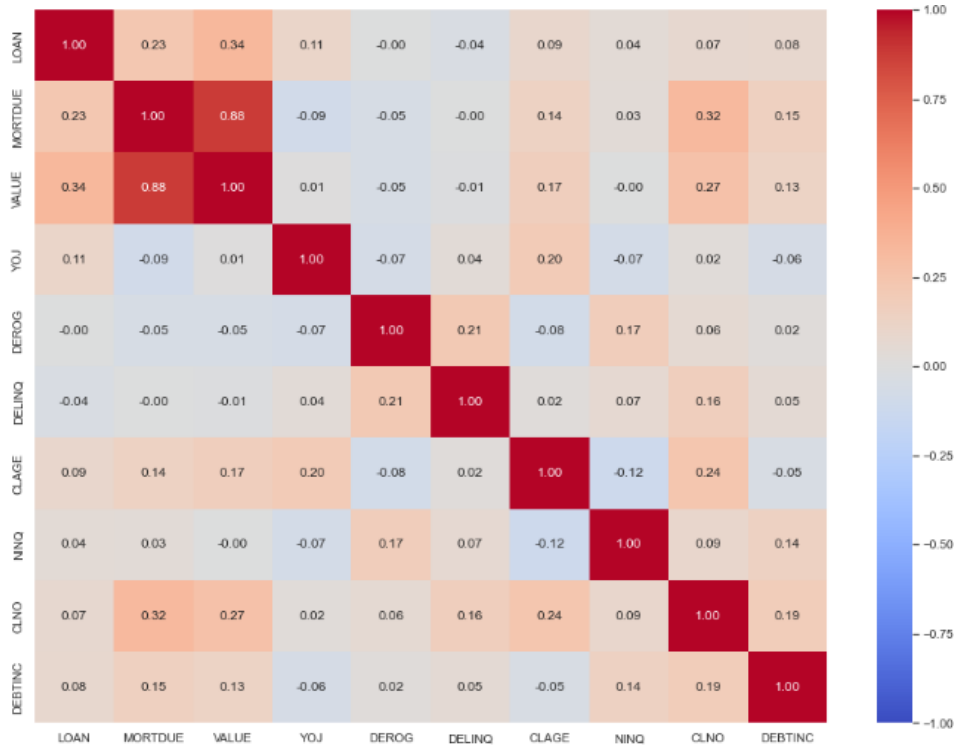
Many high DEBTINC are on “1” which represents people that defaulted on their loan.



This also showcases high NINQ has a higher distribution and many outliers for people who defaulted on their loans



Seems majority (>60%) of people are 0-1 NINQ



Correlation map shows that there is almost no correlation between the variables except VALUE and MORTDUE which makes sense if those two variables correlate. This shows that multicollinearity is not an issue.

Approach

—

We explored and adjusted between the multiple models (logistic regression, optimal threshold logistic regression, decision tree, tuned decision tree, random forest, etc.) to conclude which does the best at predicting with the dataset and which would be also most interpretable for the stakeholders as one of the objectives is to simplify and make this process more efficient. There are pros and cons to the various models for this problem so they must all be tested in order to find out the best model and solution design.

Measure of Success

Since the dataset is uneven, we can't rely on the accuracy score when assessing the models. Both recall and precision is important in this case because if the model predicts the person is eligible for the loan when they are not, that would cost the bank. If the model predicts the person isn't eligible when they are, that would also cost the bank that potential customer they could have had. Looking at the F1-score would be a good evaluation metric but I believe recall is the most important metric to look at while reviewing the models because we don't want to approve someone for a loan and have that loan default. With the dataset being unbalanced, some weights would help with lessening the misclassifications.

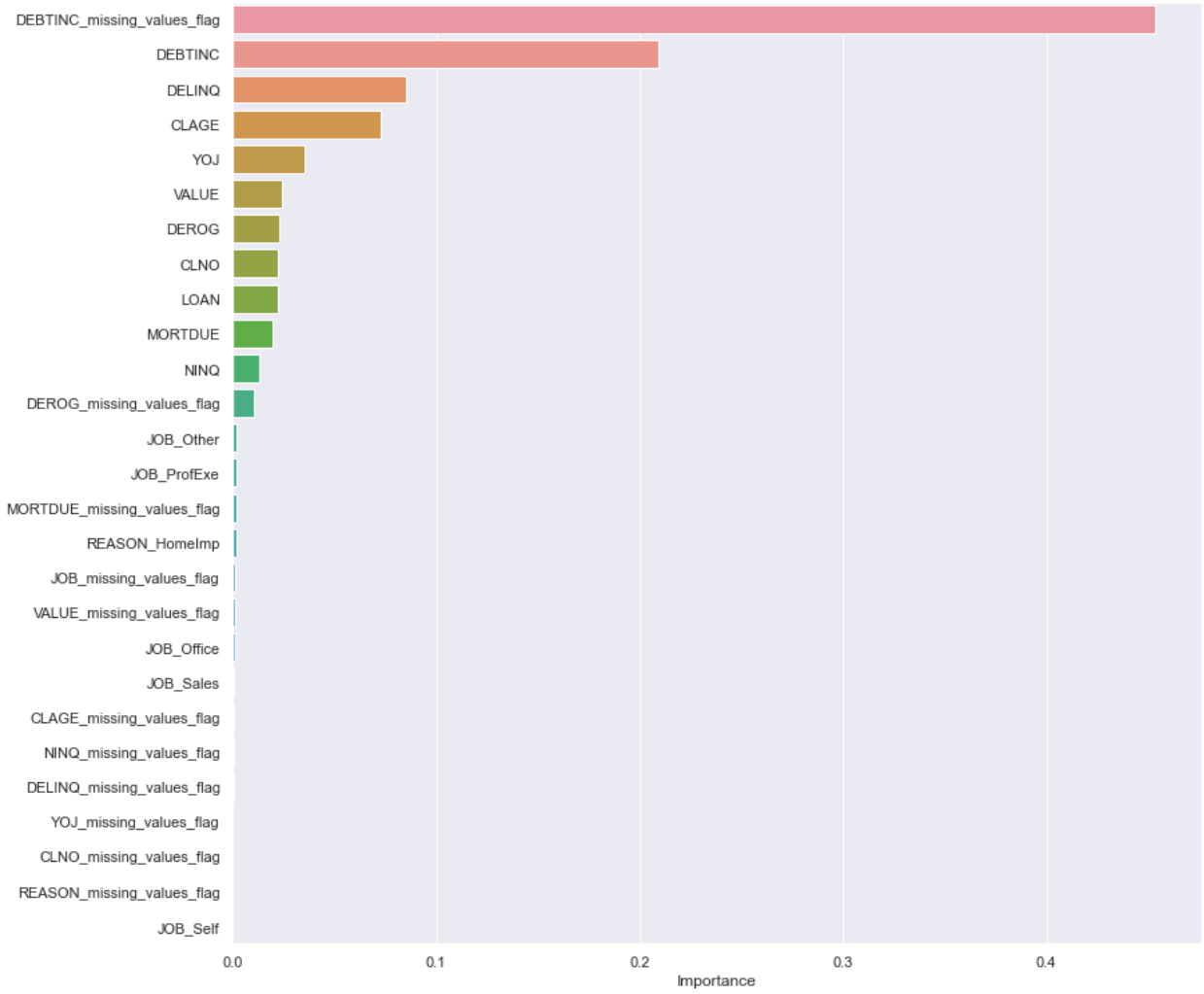
- Model that does well in both training and testing datasets where both aligns closely together
- High recall and precision rate (low rate of false positives and false negatives) with recall slightly more important and needs to be as close to 1 as possible
- Model identifies important features for the bank to consider while approving a loan

Comparing Model Performance

Logistic regression performed poorly, even with the optimal threshold. The recall was around 7%. Decision Tree was overfitting the training dataset and the testing dataset didn't follow closely. The same happened with Random Forest and Weighted Random Forest where it was overfitting and recall was low. Tuned decision tree performed slightly better with testing and training closely aligned, good recall at around 80% but the precision was a little low at around 58%.

The tuned random forest performed the best out of these models with the training and testing data performing closely together. It has one of the highest recall at around 80% and precision at around 70%. It also outlined the key important features to be DEBTINC, DELINQ, CLAGE, YOJ, VALUE, and a few more. As you can see, the DEBTINC is listed twice since we implemented a fix for the missing values.

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	Log Regression	0.811361	0.800336	0.082007	0.077957	0.644231	0.674419
1	Decision Tree	1.000000	0.868568	1.000000	0.629032	1.000000	0.706949
2	Tuned Decision Tree	0.839645	0.841163	0.807834	0.782258	0.563140	0.589069
3	Random Forest	1.000000	0.908277	1.000000	0.674731	1.000000	0.853741
4	Weighted Random Forest	1.000000	0.907159	1.000000	0.666667	1.000000	0.855172
5	Tuned Random Forest	0.890221	0.878076	0.807834	0.755376	0.686785	0.688725



Final Recommendations

We believe the tuned random forest model should be adopted because it performed very well and passed all the measures of success, outlined the key important features and is easily interpretable for stakeholders to understand. To further validate these findings, we would like to recommend implementing mandatory fields for clients to fill out so there are no missing values and we can re-evaluate the important features in the model. We expect this model to make the loan process approval much faster and efficient than the current intensive, prone to human error process. The model will

continue to learn and be able to make even more accurate predictions with all the data available. Key risks and challenges would be that the model does continue to learn so the key important features may change and this might take initial investment into the new process. Asking the clients to fill out these mandatory fields may decrease the number of applicants so we have to find a good balance and potentially make some fields that are on the lower importance scale to be optional. It may have initial errors as it learns since the recall and precision isn't at 100%.