Anita Fong

Practical Data Science/Classification | Milestone 1

January 26th, 2023

Loan Default Prediction

Problem Definition

—

The Context

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.

The Objective

The focus is on building machines 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. 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.

Key Questions

- Which clients are likely to default their loans?
- Are there any trends/patterns in the data on the people who defaulted vs people who paid off their loan?
- Which are key features that can predict whether or not someone will default on their loan?

Problem Formulation

We are trying to utilize data science and predictive modeling techniques to simplify the decision-making process for home equity lines of credit to be accepted. The model must be interpretable enough to provide justification for any adverse behavior (rejections). We need to formulate recommendations to the bank on the important features to consider while approving a loan.

Data Exploration

—

Data Description

A bank's consumer credit department wants to simplify the process for home equity lines of credit to be accepted. They obtained 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.

Observations

- 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
- Majority of people have zero derogatory reports and delinquent credit lines with a few outliers
- Majority have 0-1 recent credit inquiries with a few trailing outliers
- 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
- Correlation heatmap shows that there’s not much correlation anywhere other than MORTDUE & VALUE which makes sense. All other variables are very low.
- 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
- Number of recent inquiries (NINQ) has a bigger distribution for people who repaid the loan. There are also many high NINQ outliers for those that repaid the loan so there might be an association between people with high number of recent credit inquiries and loan repayment

- About 66% of the loans are debt consolidation while the rest are for home improvements; whatever the reason, both reasons have very similar percentage in loan repaid vs loan defaulted

Proposed Approach

—

Potential Techniques

I would like to explore the logistical regression model, decision tree model, and random forest model to see which model would do best in both training and testing datasets.

Potential Solution Design

Explore and adjust between the multiple models (regression/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

- Model that does best in both training and testing datasets
- High precision and recall rate (low rate of false positives and false negatives)
- Model reduces the effort and time required for the loan approval process and minimizes human error and bias
- Model identifies important features for the bank to consider while approving a loan