Peter Kennedy
12/15/20

Determinants of Loan Approvals

One of the ways that financial institutions earn income is by loaning money to individual people and companies. In particular for home loans, banks need to look at many variables about the applicant in order to determine whether or not to approve the loan. The goal in this project is to explore different variables that financial institutions consider when going through the loan acceptance process and to predict the outcome of loan approvals. These findings can give insights on how banks can use technology to automate the loan approval process, ultimately saving time and money.

The dataset is based off of Dream Housing Finance Company, who deals with housing loans throughout the United States in urban, semi-urban, and rural parts of the country. The dataset contains 614 rows and 13 columns, 11 of which are considered in loan approvals The columns are both numerical and categorical, examples of variables include gender, education, loan amount, credit history, and more. The data used can be found at https://www.kaggle.com/burak3ergun/loan-data-set and https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/.

About 50% of the columns contain missing values. To resolve this issue, I filled them in with the most frequent value in their respective column when running my PCA and logistic regression. I got rid of the loan ID column, which did not provide any insight into the loan acceptance process. I also changed the values in the married and self-employed columns from "Yes" and "No", to more specific values: "married" and "not married", and "self-employed" and "not self-employed", respectively. I did this in order to differentiate the two columns when applying one-hot encoding to the logistic regression analysis.

I start exploring the data with a stacked bar plot of loan approval based on education. **Figure 1** splits the count - college and non-college graduates - into two groups, approved and not approved. College graduates applying for a home loan were approved 70% of the time, while non-college graduates applying for a home loan were approved 61% of the time. The figure also indicates that from the count that was approved for home loans, 81% of them were college graduates.

To further analyze the different variables in our dataset, I used Principal Component Analysis (PCA) to look into the explained variance by all 11 variables. **Figure 2** shows the scaled and unscaled cumulative explained variance by the columns in the dataset. Out of the 11 columns used in the PCA, 7 of them are categorical. To run this analysis, I transformed those 7 columns using one-hot encoding. The unscaled line shows that 3 variables capture 100% of the variance in the dataset. This can be explained by the numerical variables: applicant income, co-applicant income, loan amount, and loan-amount term. These columns have values in the hundreds and in the thousands, compared to the categorical columns being either 1's or 0's because of the one-hot encoding.

Next, I used the 11 explanatory variables to predict loan approvals, stratifying for the loan status column, which I changed to Boolean values. I took my original data and preformed a 75% train, 25% test split, and used a scikit-learn pipeline with the following chain: (**1)** SimpleInputer (using most frequent strategy), **(2)** OneHotTransformer (a class I made that transforms categorical data using OneHotEncoding), **(3)** StandardScaler **(4)** LogisticRegression. The model has an accuracy of 80%, a recall of 98% and precision of 78%. Overall, the classification algorithm performed well in predicting loan approvals, given that the original data had a 69% loan approval rate.

Figure 3 shows the top seven weighted coefficients used in the logistic regression model to predict loan acceptance. The most heavily weighted categorical coefficient in the trial run is credit history, while the largest numerical weighted coefficient is loan amount. Credit history has a positive relationship with loan approval, while loan amount has a negative relationship to loan approval.

The analysis above looks at variables that are involved in the loan approval process. I first looked at the impact that education has on loan approvals, and then looked at the explained variance within each variable. Lastly, I ran logistic regression and found that the heaviest weighted categorical variable in loan approval process was credit history and the heaviest weighted numerical variable was loan amount. Moving forward, I would advise Dream Housing Finance Company and others to prioritize credit history when considering loan approvals for homes.

Peter Kennedy
12/15/20

Figures

Figure 1: Loan Approval Based on Education

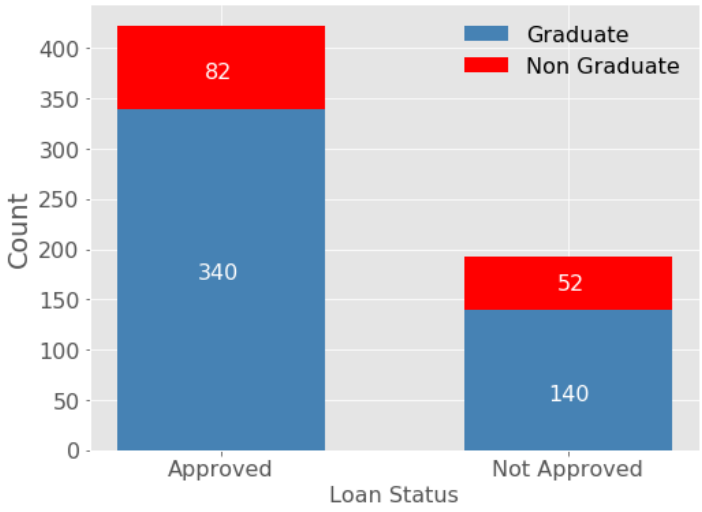

Figure 2: Principal Components of Loans

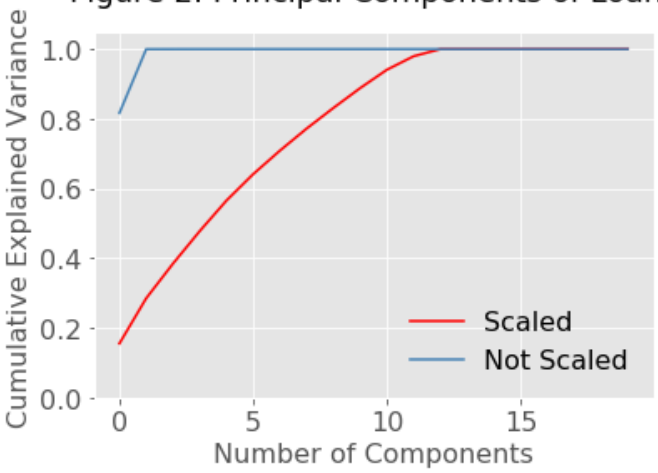

Figure 3: Top Seven Weighted Logistic Regression Coefficients