Need To Approve Loans?

Understand The Metrics Required to Get Your Money Back

Mithil Patel

Bellevue University

DSC550-T302 Data Mining (2231-1)

Prof. Brett Werner

November 19, 2022

Since ancient history, humans have accepted money as a form of payment to exchange goods and services. Money allows an individual have access to countless resources to thrive in life such as food, healthcare, education, etc. Therefore, many people find themselves at the mercy of a lender to supply money as a loan in hopes to repay the amount in the future with loan interest. From a lender's perspective, providing loans is a profitable deal since the borrower has agreed to return the full amount with an interest – profit for the bank— determined during the agreement. However, giving out loans can be fatal for lenders if the borrower is unable to return the agreed amount, essentially forcing the lender to lose the borrowed money. As a matter of fact, one of the major problem people are currently facing is consumer debt. According to financial experts, approximately 80% of Americans have some sort of consumer debt, reportedly around a total of $14 trillion is owed collectively with 13% of Americans not ever being able to fully pay off their debts. This has led to many companies and banks losing money that they were not able to acquire back from giving loans out to consumers, which could result in major entities going out of business nationwide and potentially an economic recession similar to the 2008 Financial Crisis in the United States.

To combat the issue, I ought to create a model that would be able to analyze attributes for the lenders (i.e., banks) giving out loans and help prevent lenders from losing billions of dollars. Before creating the model, I shall search for a good dataset because the dataset is fundamental for a data scientist to tackle issues and instill confidence in the given resolution. For this project, a dataset was extracted from Kaggle, an online platform to share datasets, that include attributes related to credit history, loan amount, applicant income, education, marital status, credit loan approval, etc. The dataset helped determine the most common attributes companies used to determine how they were giving out loans to individuals.

As part of the data preparation process, numerous data transformation techniques in conjunction with data visualization were utilized to clean the dataset. The following transformation was performed to our dataset: converted categorical variable to numerical, removed irrelevant columns such as loan ID, replaced null values with column median, and feature scaling. Additionally, data visualizations were created to explore data to filter attributes pertaining to consumers' ability to repay loans. The correlation heatmap showed that credit history was the primary factor used by companies to allow loans (Fig. 1). Furthermore, marital status is a good indicator for companies to approve loans. There were a lot more given approvals for loans than non-married couples and men seemed to receive the majority of the loans. Count plots shown in Fig. 3 provide further evidence to support our previous claims regarding credit history and marital status variables. An income vs loan amount scatterplot depicts the loan approval rate amongst male participants as a significant rate compared to females (Fig. 2). Only the relevant columns extracted from analyzing charts were included in the final dataset. To maintain our statistical analysis significant and avoid skewing our results, boxplots were created to display extreme data points in our dataset (Fig. 5 & 6). To handle outliers, a z-score was calculated for the columns with outliers, and any data point higher than 3 was removed.

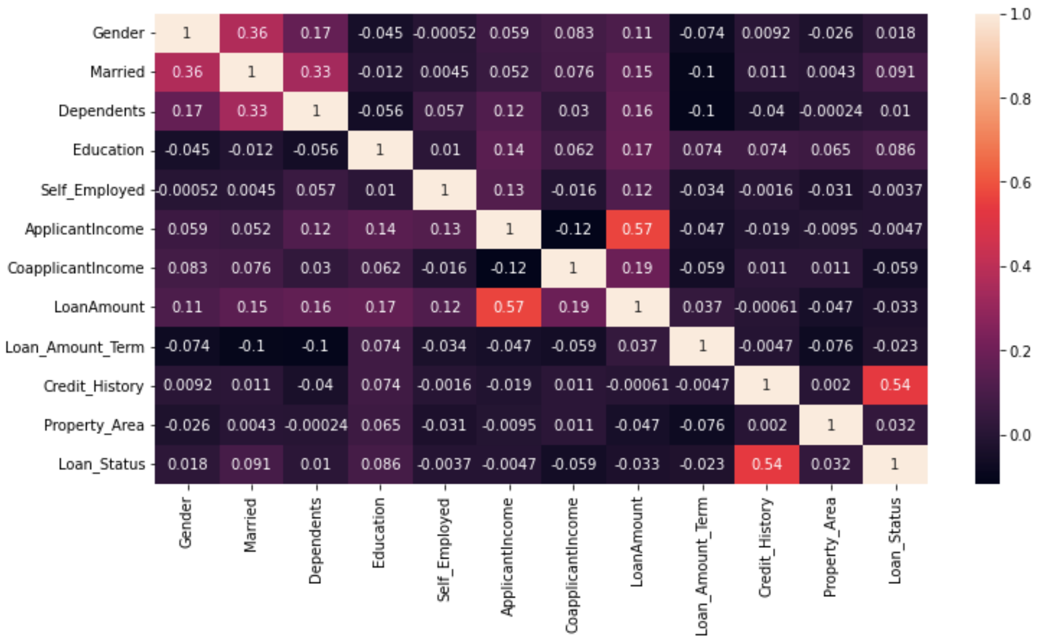

Fig 1. A correlation heatmap of all the variables along the associated correlation coefficient values. The relationship between the two variables strengthens as the square becomes lighter in color.

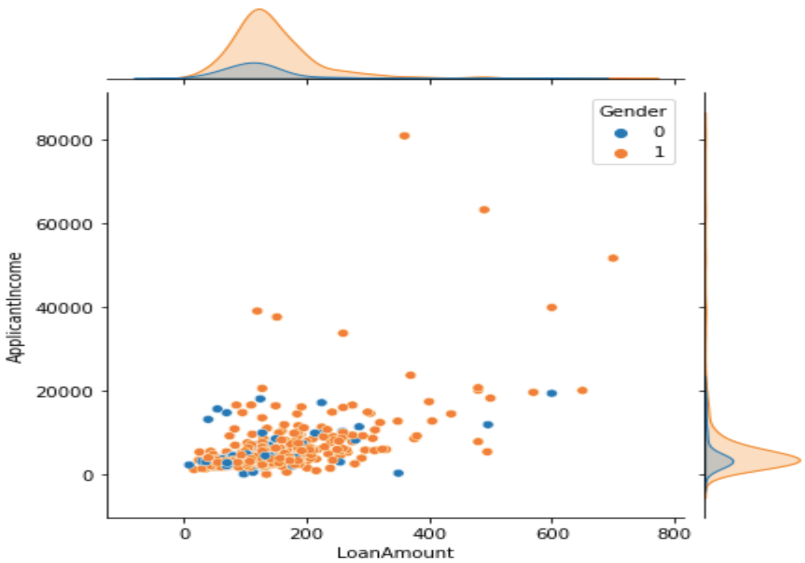

Fig 2. A scatterplot of applicant income vs loan amount along with a density plot ordered by gender to show how the data points are distributed.

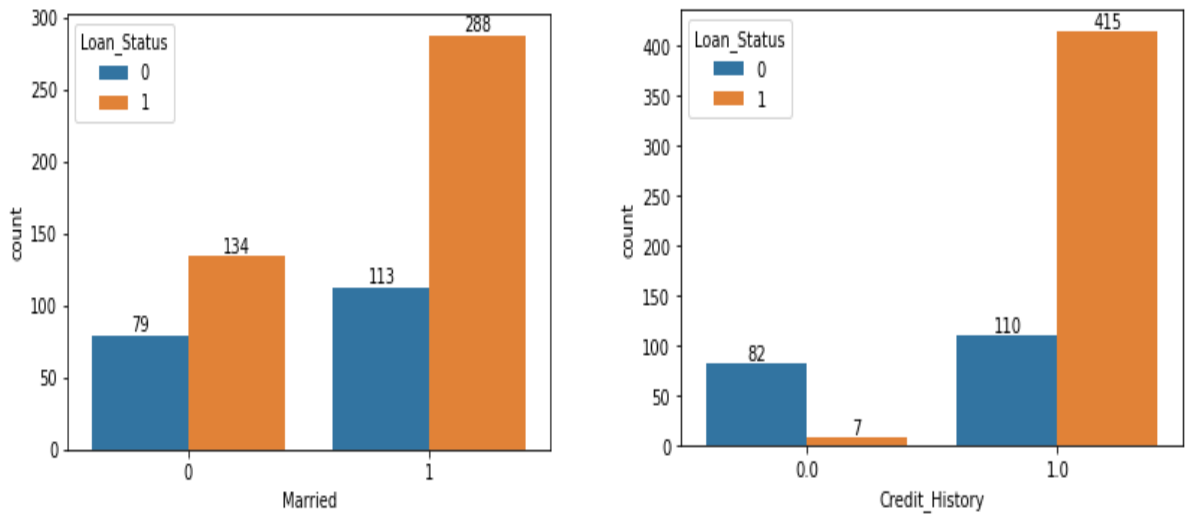

Fig 3. Count plots of marital status (left) and credit history (right) grouped by loan status to show the observational values in each category.

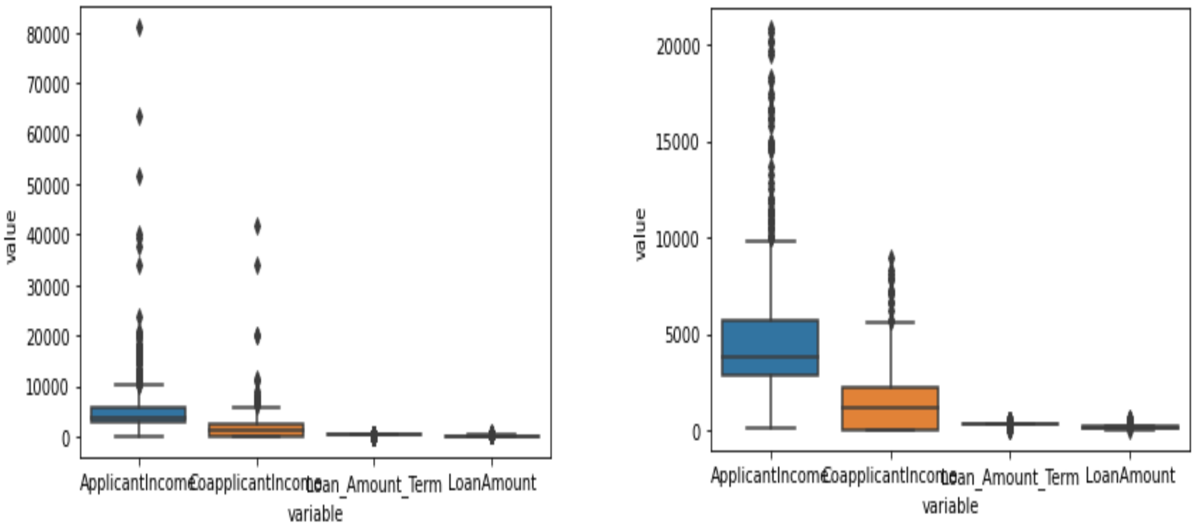

Fig 4. Boxplots of four continuous variables to check for outliers. The left boxplot shows the plot with outliers while the right boxplot shows the plot without outliers.

Once the data preparation step was complete, GridSearchCV, a built-in python library, was utilized along with a logistic regression classifier pipeline to test multiple models to find the best suitable model for the given dataset. GridSearchCV is used to determine the optimal parameters for an estimator from a given input parameter grid. The parameters can be selected based on the highest test score (accuracy) output. However, the technique failed to provide a model with a higher accuracy score than the model I had initially created with logistic regression classifier (84% accuracy) to gauge the baseline model. Additionally, the precision, recall, and F1 values are 0.825, 0.988, and 0.8995, respectively; thus, further providing support that the chosen model (logistic regression classifier) is ideal for our given dataset.

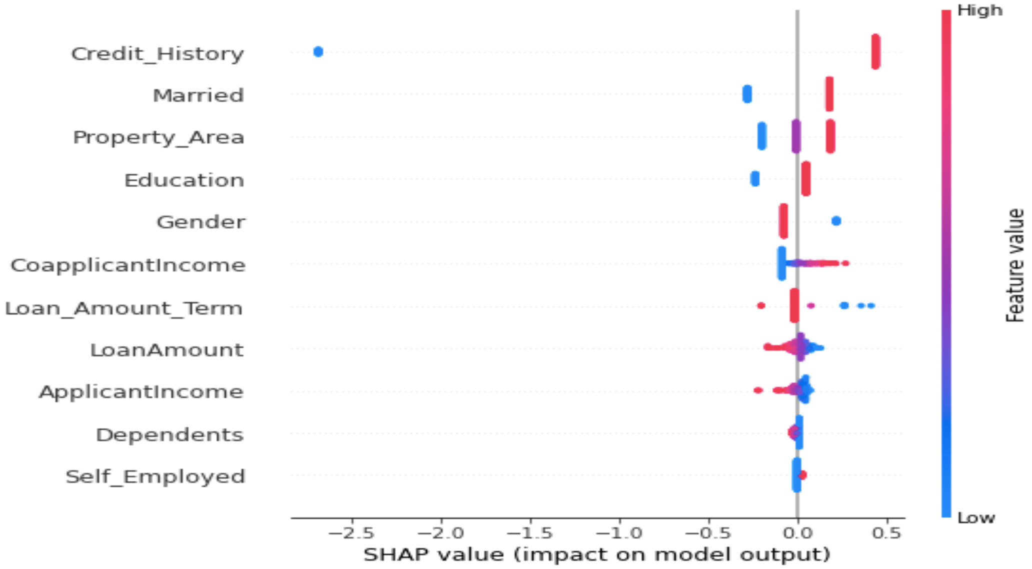

Fig 5. A SHAP value summary plot showing feature importance. The importance of the attributes is ranked in descending order.

Using the logistic regression classifier model, a SHAP values plot was generated to rank the feature in order of importance (Fig. 5). The SHAP plot suggests that there is a high loan approval rate for applicants who satisfy the credit history condition while there is a significantly low chance of approval if the condition is not met. Additionally, there is a high approval chance if an applicant is married, has high property area, is educated, and is a female compared to a single applicant, has a low to no property area, is uneducated, or is a male. Surprisingly, a large number of applicants had loans approved with low income while a low number of applicants with a high income had their loans approved. The insight from our analysis is beneficial to lenders as it encourages them to explore other factors when approving a loan that could increase the probability of a borrower being able to repay the loan, thereby proving more profitable for the lenders. However, the model is in the early stage of development; therefore, it is not yet ready for deployment. There are some areas yet to be explored that could improve the performance of our model. For future development, I highly recommend including additional attributes such as the amount of money a borrower was able to pay. Moreover, I suggest testing a neural network algorithm, specifically an artificial neural network (ANN) as a multi-layer network is more suitable to recognize patterns. Perhaps a neutral network could outperform the logistic regression classifier model that could assist lenders to make calculated decisions. After all, data-driven decisions are impartial compared to gut instincts.

Reference

- “American Debt Statistics.” *Shift Credit Card Processing,* Mar. 2021, shiftprocessing.com/american-debt/.