Strategic Intervention to the Personal Lending Crisis

By Sara Herbstreit

Project Overview

The ability to correctly assess an individual's likelihood of defaulting on a personal loan is of utmost importance to the personal loan lending industry. There is approximately $305 billion dollars in personal loan debt in the United States alone[1]. The current default rate on personal loans is 3.3%, which is more than double that of mortgage and auto loans[2]. This uniquely high default rate accounts for massive financial losses to the personal lending industry each year.

Due to the inherent complexity of an individual's financial background, it is impossible to correctly assess the likelihood of default from single factors such as credit score alone. Assessing the client should be based on a myriad of factors. Historical loan data can be used to evaluate trends in the preceding factors leading to defaulted personal loans. These factors need to be analyzed in-depth to fully understand the relationships between them and the chance of default. By using machine-learning algorithms, many details of thousands of people can be analyzed simultaneously to reveal the most influential factors. This selection of factors can then train a machine learning model capable of predicting the risk of default for any potential client.

An improved approval methodology can reduce defaults, while also increasing the client base by approving worthy candidates that may have been denied by outdated models.

Data Processing and Feature Selection

The data on historical personal loan defaults was collected from Kaggle.com[3]. The target variable in this dataset is loan status, which shows whether the loan was defaulted or paid as expected. The original independent variables were: borrower age, borrower income, borrower home ownership status, borrower employment length, borrower default history, length of borrower credit history, loan grade, loan amount, interest rate on loan, and debt-to-income ratio.

Downsampling was applied due to the data being imbalanced in terms of default and non-default data points. Erroneous values, such as borrower age being greater than 100, and length employed being greater than 70 years, were removed from the dataset. The feature mean was used to fill any missing data points, and the categorical features were converted to numerical using one-hot encoding.

The independent variables were analyzed for relevancy to the target variable using the SelectKBest algorithm from the Sklearn library. The chi^2 statistic was chosen as a means of scoring each feature. A high score indicates the feature is not randomly related to the target variable; therefore the relationship between the feature and target is significant. **Figure 1**

---

[1] Stefan.lembo-Stolba. "Consumer Debt Study." Experian, Experian, 9 Mar. 2020, www.experian.com/blogs/ask-experian/research/consumer-debt-study/.

[2] "Personal Loan Statistics." LendingTree, www.lendingtree.com/personal/personal-loans-statistics/.

[3] Tse, Lao. "Credit Risk Dataset." Kaggle, 2 June 2020, www.kaggle.com/laotse/credit-risk-dataset.

shows the scores of the top 15 features. The elbow method was used to determine the cutoff point of features, which eliminated the lowest three features in figure 1 and all others with a lower score from the data set.

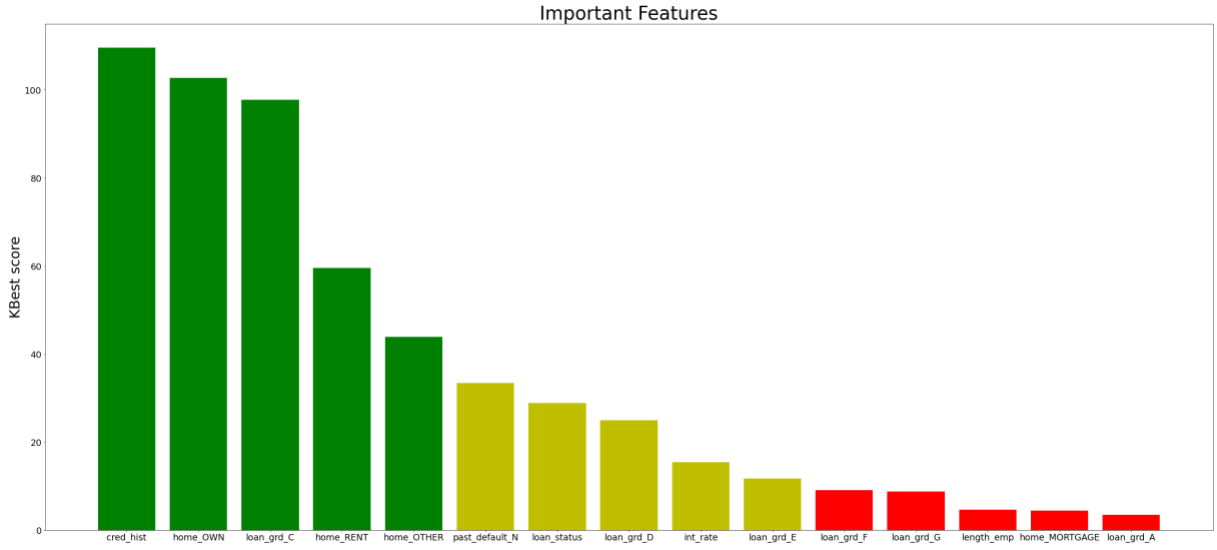

Figure 1

Model Methodology

The data was split into a training and test set, with the training set used to evaluate two different model algorithms: Random Forest Classifier and Support Vector Classifier. The hyperparameters were tuned on each model to produce the best fit on the training data. The performance of each model was evaluated using the Area Under the Curve (AUC) metric. This metric measures the ability of the model to correctly assign the matrix of features to the target variable and graph the prediction as a curve. **Figure 2** shows both model graphs. The x-axis indicates the false-positive to true-negative ratio, while the y-axis indicates the true-positive to false-negative ratio of both models. The net performance is shown on the lower right of the graph.

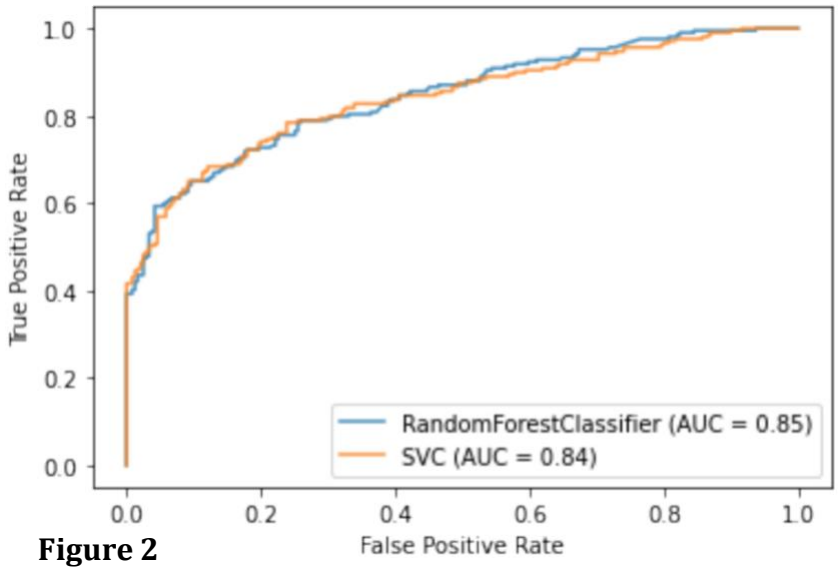

Figure 2

Conclusion

Both model selections performed well, with the RandomForest model outperforming the SVC by a slight margin at 0.85 vs. 0.84. Only minor variations can be seen between the two models predictions. Both models had better success correctly distinguishing data in the negative class (did not default) over the positive (defaulted). With a performance score of 85%, the Random Forest model has the potential to reduce the current default rates on personal loans by up to 2.8%. Future improvements could be made through optimization of the data preprocessing techniques.