Project Milestone 2

Harlan Wittlieff

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

June 26, 2022

Topic

A small finance company, Loans Today, is launching an initiative to streamline our loan approval process through leveraging automation. The project is called “Approval AI”

Background/History

Currently, Loans Today (a fictional company offering loans to individuals) employs a small team of Loan Officers. This was an adequate staffing level when the operation started. However, the manual loan approval process is creating a bottleneck in the loan operations. Due to today’s tight labor market and increasing compensation package demands, our finance company does not have the budget to hire more Loan Officers. Additionally, it would take a great deal of time to train new hires and get them up-to-speed on the loan approval process. Therefore, the company is turning to AI for a solution to this problem.

Business Problem

The goal of this data science project is to create an algorithm that can predict loan status as approved, yes or no. This will streamline the loan application review process. However, our company does not intend to replace the Loan Officers completely, but rather we will use this as a first stage of screening loan applications. Essentially, the Approval AI could weed out the obviously unqualified loan applications to be flagged for denial, leaving only the borderline and fully qualified loan applications for Loan Officer review. This project will directly impact the efficiency of the loan approval process.

Data Explanation

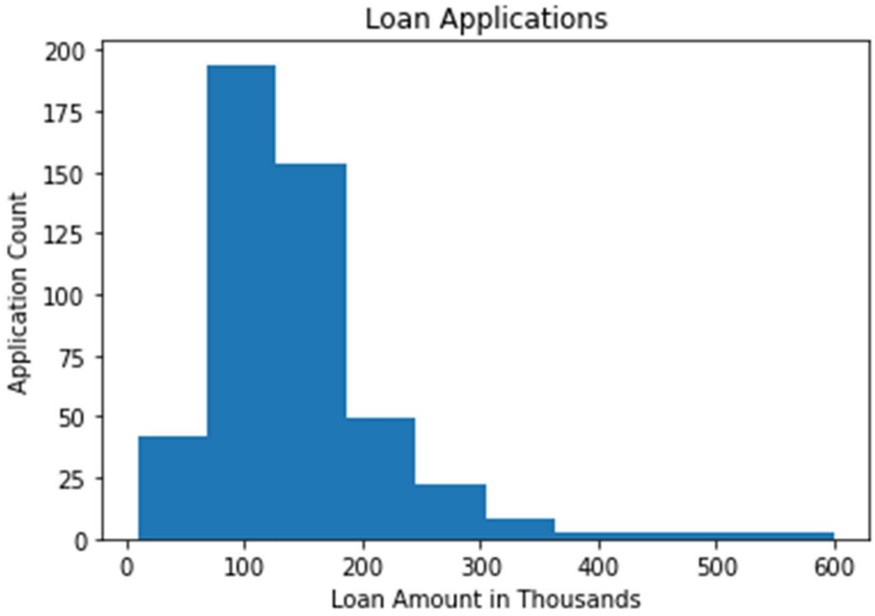
Dataset

The Loan Approval Data Set on Kaggle will serve as the primary dataset for this data science project. The dataset features the key variables included on the Loans Today loan application. These variables include the following:

- Loan ID

- Gender
- Married
- Dependents
- Education
- Self Employed
- Applicant Income
- Co-applicant Income
- Loan Amount
- Loan Amount Term
- Credit History
- Property Area
- Loan Status

Overall, the dataset contains 614 applications for loans for loans from $9,000 to $600,000. The requested loan amounts follow the distribution below.



Data Preparation

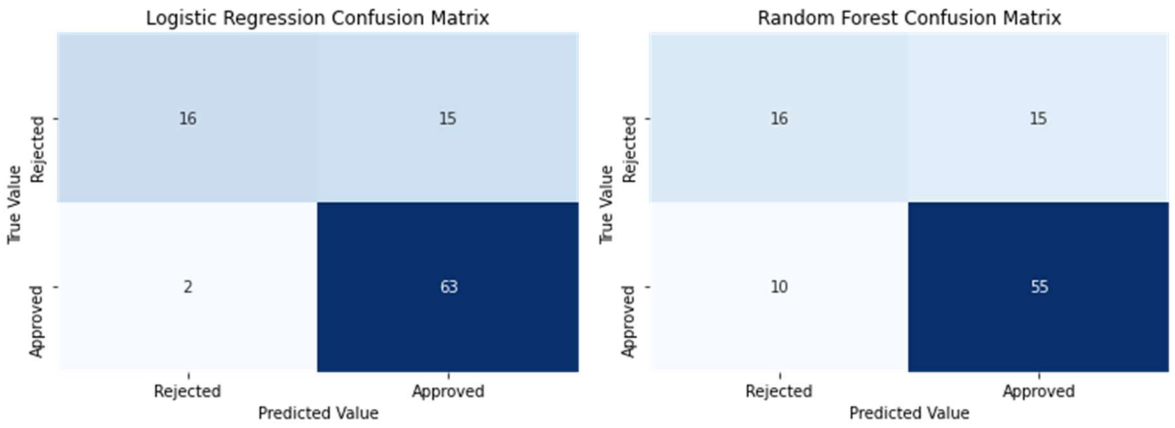
To begin preparing and exploring my data, I checked for null values and removed any records containing nulls. I made this decision since any loan applications with missing information would be resent to the applicants for completion. The most substantial transformation the dataset required was the creation of dummy variables due to many of the fields containing categorical data such as gender, married, if a graduate, self-employed, and property area. The last transformation was to split the dataset into training and testing datasets.

Methods

The target variable for this project is “Loan Status,” as the model will predict whether a loan is approved. Since the classes were not extremely unbalanced, I did not need to leverage a method such as SMOTE. As this is a classification scenario, I built a logistic regression model. A random forest model was built to compare with the results of the logistic regression model.

Analysis

To evaluate the results of both models, I computed the accuracy scores. The logistic regression model yielded an accuracy of eighty-two percent. This accuracy score was higher than that of the random forest model, which came in at seventy-four percent. To further analyze the results, I created a confusion matrix for each model pictured on the following page. The logistic regression model successfully identified sixty-three out of sixty-five loans for approval. However, it incorrectly approved fifteen out of thirty-one loans that should have been rejected. The random forest model has the same number of false positives as the logistic regression model. However, the random forest model falsely rejected an additional eight applications that should have been approved when compared to the logistic regression model.



Conclusion

Based on the model evaluations, the logistic regression model will be implemented as an initial screen for loan applications. Since this model had a very low chance of falsely rejecting loans that should be approved, it will be helpful as a timesaver for our Loan Officers.

Assumptions & Limitations

The dataset used to build the model only contains information about which loans were approved. This gives the assumption that approved loans are loans that are profitable. If a loan was approved and ultimately was not successfully repaid in the future, the model may pick up undesired traits. Additionally, the recommendation for implementation assumes that the lost value from any false rejects is made up for with the savings in the reduction of work for the loan officers.

Challenges

After reviewing the results of the model, I feel a few of the features in the dataset were too vague. For example, Credit History is a binary field that contains a 1 for a favorable credit history and a 0 for an unfavorable history. Utilizing the applicant's credit score may provide richer information for the model to learn from.

Recommendations & Additional Applications

Tracking the loan repayment status and utilizing this metric for the predicted variable will provide for a more profitable model. Additionally, expanding to determine optimal terms of the loan such as repayment periods and interest rate could potentially boost profitability. Expanding the model for lower loan amounts will allow for increased time savings for loan officers and increase profitability.

Implementation Plan

This model will be implemented as an initial filter to weed out loans that have a low chance of approval. The applications not filtered out by the model will be sent to the loan offer for additional review.

Ethical Assessment

Seeing as a loan can significantly impact the livelihood of the applicant, care must be taken to ensure there are not any unfair biases in the model. This due diligence is required not only when building the model, but also following deployment to ensure the results are equitable.

10 Questions from the Audience

1. Is the model's accuracy score high enough to be reliable?
2. What is the timeline for implementation?
3. Should Loan Officers spot check the loan applications that the model rejects?
4. Should we be concerned that the model could potentially adversely impact certain demographic groups?
5. As a Loan Officer, should I be concerned about my job security?
6. How will we know if the model is working as intended?
7. How can we make the model better?
8. How do we encourage buy-in across the company prior to launch?
9. Are our competitors leveraging similar solutions?
10. Do you anticipate implementing predictive analytics or similar solutions in other areas?

References

Ranjith, K. (2020). *Loan Approval Data Set*. Www.kaggle.com. https://www.kaggle.com/datasets/granjithkumar/loan-approval-data-set