Vikram Achuthan
September, 2022

What is Prosper?

Prosper is a peer-to-peer lending platform where people can invest their money in personal loans requested by other people.

What is the problem you are trying to solve?

It can be difficult for a prospective investor to ascertain if a certain loan will be paid back on time given the data provided to the investor. I aim to build a model that adds an extra degree of confidence for the investor to know that they will be paid back.

What is the data set available from Prosper?

The dataset contains more than 100,000 entries, where each entry represents a loan taken out by a user through Prosper, and 81 columns of parameters describing the loan.

What is your solution?

In this project, I performed a binary classification using a Logistic Regression Model on a set of loan data from Prosper. Binary classification is a supervised ML algorithm that classifies new observations into one of two states (0 and 1) . In this case, the “0” state is a loan that was not appropriately paid back, and the “1” sate is a loan that was paid back in time.

Using the scikit-learn machine learning library, I implemented a Logistic Regression Model to classify loans as one of the two states described above.

How did you choose the features?

The parameters I chose in this model were the column headers that are also provided to a prospective investor on Prosper’s investor platform. These include, but are not limited to:

- FICO range
- Employment status
- Occupation
- Stated Income
- State (VA, MD, NY..)
- Inquiries Last 6 Months
- First Credit Line
- Current/Open Credit Lines
- Total Credit Lines

- Revolving Credit Balance
- Bankcard Utilization
- Has Mortgage
- Debt/Income Ratio

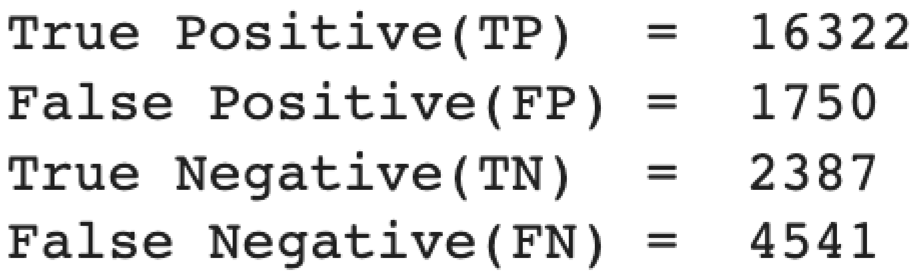
How did you choose the label?

For this model, the only label is the binary paid back/not paid back classification described in section 1.

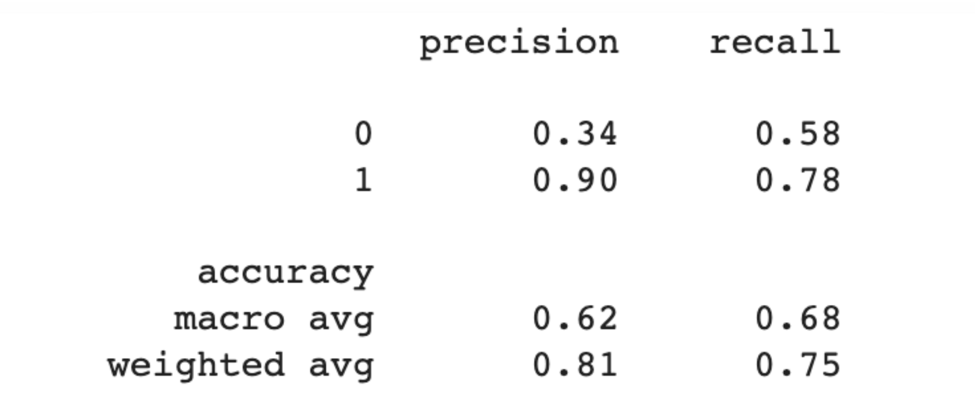
Explain your results

A binary classifier is evaluated based on 4 parameters: True Positive, False Positive, True Negative, and False Negative

How well did the model perform?



```
True Positive(TP)  =  16322
False Positive(FP) =  1750
True Negative(TN)  =  2387
False Negative(FN) =  4541
```



```
              precision    recall

           0       0.34      0.58
           1       0.90      0.78

    accuracy
   macro avg       0.62      0.68
weighted avg       0.81      0.75
```

The table above shows the precision and recall ratios of the model. Precision quantifies the proportion of positive identifications that were accurate TP/(TP + FP) while recall quantifies the number of *actual* positives that were classified correctly TP/(TP + FN). It is clear that the model did a relatively good job of identifying “good” potential loans correctly, and did not do a good job identifying “bad” potential loans correctly.

Next Steps:

In order to make this model more accurate and useful, I plan to:

1. Study the distribution of the feature set and the labels to understand the biases and correct them using appropriate sampling.
2. Generate a variety of models using a combination of features and compare the performance of these models.
3. Implement a good methodology to compare the model performance on various characteristics like accuracy, AUC, F1-score etc.