Credit Risk Analysis Report – May 19, 2024

Ryan MacFarlane

I am writing this report to highlight the performance of the logistic regression model I performed on the data found within the ‘lending\_data.csv’ file I was provided. I was given this financial data in hopes of producing a model that can predict whether a mortgage will be healthy or could instead be considered high risk. I separated out the target variable (loan\_status), and then performed a logistic regression. And I will now present the results of this model and provide a recommendation for it.

A logistic regression is a model that uses aspects of the data to help predict a binary choice, for example – a 0 or 1. In this instance it is being used to determine if mortgages in the input data set are healthy or could be considered at-risk of failure or default. The target variable in the input data set has the following breakdown :

0 – healthy mortgages – 96.8%

1 – at-risk mortgages – 3.2%

What follows is a brief description of each of the metrics used to grade the model’s performance. There are four categories of possible prediction outcomes for a model in cases like this: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

* Accuracy is the metric that gives a good view of overall model performance, by measuring the overall correct predictions (TP + TN) against the overall number of predictions made (TP + TN + FP + FN). It can, however, be a little misleading depending on how your input data is skewed. For example, if positive results are heavily outweighed by negative results.
* Precision focuses on positive predictions made, both correctly and incorrectly, and can sometimes give a more accurate reflection of a model’s performance than accuracy alone. The equation to determine precision is TP / (TP + FP)
* Recall is another metric which focuses only on positive outcomes – it measures the number of correct true predictions against all actual true outcomes - TP / (TP + FN)
* The F1 score is the average of precision and the recall. In my opinion, it is the best and most balanced metric for gauging model performance across different types of input sets.

The overall scores of the logistic regression I performed are as follows :

Precision – 0.99

Recall- 0.99

F1 Score – 0.99

The model was near-perfect at predicting healthy mortgages, and the performance regarding at-risk mortgages was also strong, although not quite as fantastic as the healthy mortgages. Of all the logistic regressions I have performed, this is the best performance I have seen, and I am giving it my highest recommendation. We should put it into production immediately.

Thank you very much for your time!

Ryan MacFarlane