Probability of Charge-off Neural Net Model

Nate Youngblood

April 2021

1 Introduction

The purpose of this document is to describe a model is designed to predict charge-offs in a retail lending portfolio and was calibrated using data from Lending Club from 2007-2015.

2 Data

Various transformations were applied in order to make input data more logical and useful for modelling.

The following variables were found not to contain useful data for modelling and were therefore dropped:

- URL
- Borrower ID
- Observation ID

Job title was found to contain over 90,000 unique values, and because this many categories of a categorical variable would cause model fitting to exceed capacity, the variable was dropped.

21 variables were found to contain missing values for over 50% of observations, and were therefore dropped in order to avoid potentially biased missingness.

Date variables were converted into a number of seconds since the origin date.

Finally, numeric variables were normalized in order to put them all on the same scale and enhance model accuracy.

3 Model methodology

Data were randomly split into test and train sets, with 20% of data in the test set. The training set was further partitioned to create a separate validation set

for model training. I trained the model on batches of 512 observations over 10 epochs. The model was estimated using Python and Tensorflow in a Jupyter Notebook.

I chose to use a neural net model with two dense layers of 62 units each and a rectified linear unit activation function, a 10% dropout layer, and a sigmoid activation function output layer.

The advantages of this model type are that it can model complex, non-linear relationships without the feature engineering or one-hot encoding required by ensemble or logistic methods. This is especially important in a data set like this one, with many numeric and categorical variables. Furthermore, there are also a large number of observations and variables and further increasing this number by adding indicator variables for each of the many categories would drastically increase the overall size of the data set and processing time for any modeling task.

While this model type can predict relationships with a high degree of accuracy, the disadvantage is that results are more difficult to interpret. While it is possible to estimate feature importance by iteratively removing each variable from the input data and re-estimating model parameters, this method is time intensive and thus outside the scope of this analysis.

Future versions could include calculation of variable importance scores, which would indicate which loan characteristics are associated with higher charge-off rates.

4 Model output

When used to predict test probabilities, the model showed accuracy of 99.919% and a loss rate of 0.334% as calculated by binary cross-entropy.

While this indicates high predictive accuracy and discriminatory power overall, the figure below shows that loss rate is not evenly distributed across time.

Future versions could include month as a categorical variable rather than numeric as there are relatively few months in the analysis period.

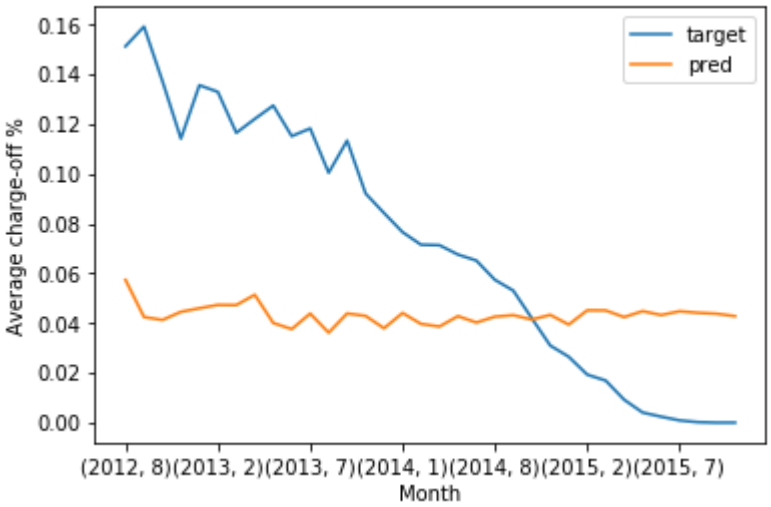

Figure 1: Predicted vs. Actual Charge-off Rates