Group: The Lone Wolf
Members: Ben Shaw
Date: 10/28/2022
Competition: AMEX default prediction

Checkpoint 2 - Report

The purpose of this competition is to predict the probability that a customer will default on their credit card payment. Training data and labels are given, which training data is approximately half the size of the test data. In the training data, approximately 74% of the labels are attributed to the “0” class: that is, a non-default. Therefore, the mean accuracy of all models must be above 74%. Notably, the scoring metric for this dataset is not the mean accuracy, but rather is the mean of the normalized Gini Coefficient and the default rate captured at 4%.

For each customer, and for each statement date, 188 additional features are attributed. The training and test data are given in the form of a single .csv file (respectively), so that a single customer’s data is given on multiple rows, each with a different value for the statement date. Thus, in order to obtain the statement data for a single customer, the .csv file needs to be searched for all rows with matching “customer_ID.” There are 924,621 customers in the test dataset.

The first task in preprocessing was to split the data so that each customer ID in the training and test data has its own .csv file. This was a time-consuming process, taking several days to complete. Next, .csv files were created which contained only the last statements from each customer: the early models would be based only on the last statement data. “NaN” values were changed to 0.

It was intended that the first models would be based on the random forest algorithm. For simplicity, these first models would be based only on the last statement for each customer. A pandas dataframe was thus created: each row represented the last statement for each customer in the training set. The test dataframe was constructed likewise. Four features were removed from the data for training and testing: the customer ID, the statement date, and two features whose values were strings.

Four random forest models were trained on the training data. The first model used 100 trees and no maximum depth; the second used 100 trees with a maximum depth of 5; the third used 1000 trees with a maximum depth of 5; the fourth used 1000 trees with no maximum depth. To estimate which model had better performance, the mean validation accuracy of each model was calculated: the validation set accounted for approximately 33% of the training data. The first and fourth models exhibited the two highest mean accuracies and where subsequently used to make predictions on the training data: the mean accuracies were approximately 89.5%.

For each of the two models, the class probabilities for the test data were computed and used to create a submission .csv file. The submissions were evaluated, and the public scores are as follows: 0.76060 (first random forest), 0.76592 (fourth random forest). The private scores are as follows: 0.76841 (first random forest), 0.77409 (fourth random forest). It is notable that these scores are higher than the “benchmark random forest” score of approximately 0.75595.

The plan for the rest of the semester is to run different models on the data. Specifically, models that treat time series data are desired. A basic neural network architecture will be trained on the “last statement” data that the random forest model was trained on. Subsequently, the data will be treated as a time series, where the time step is the statement date.