AMEX Default Prediction

Project Goal

- Predict Default Probabilities

Data Info

- Training/valid: 458000+ rows. Test: 924000+ rows.
- 190 columns, including statement dates, customer ID, two categorical. Each customer has 1-13 statements (rows).

Data Preprocessing

- Group the data by customer ID, split into separate .csv files (time consuming)
- “NaN” to 0.0, removed the two categorical columns (string values).
- Statement dates and customer ID’s were not used in training.

Evaluation metric

- The evaluation metric was an AMEX custom metric The competition ended, and the code to calculate the metric used in the leader board scoring was no longer available (higher is better).
- Trained using accuracy for the substitute evaluation metric.

Early Models: last statement data (after dummy submission)

- Random forest classifier: 88% accuracy and score of 0.768 on private leader board using 100 trees. With 1000 trees: 89.5% accuracy, 0.774 private score. Benchmark Random forest: ~0.75 private score.
- SVM: 0.71 private score, 87.7% accuracy.
- Simple Neural Network: 3 FC layers, 2 relu and softmax as final. Adam optimizer, NLLLoss, many epochs. Accuracy: 89.7%. private score: 0.698.

Later Models: incorporating all statements.

- “Success” was found combining 13 classifiers into one model, depending on number of statements per customer. Statements were appended so that each customer had one row of data, varying number of columns.
- 13 Random forests: linearly increased the number of trees with the number of statements. 1-11 statements: ~80% accuracy. 12: ~84%; 13: ~90%. Private score: 0.759.
- 13 Neural networks: similar accuracy. Private score: 0.72.

Reflection

- Despite most models obtaining near 89%-90% accuracy, private scores varied (determined by class probabilities—perhaps why overconfident NN’s failed).
- NN’s and RF methods are preferred so that training can be done in a parallel manner.

Future work

- Incorporate categorical data?
- Build a more robust NN, perhaps use regularization and other methods to decrease class probability confidences.