Tony DePalatis (Kaggle user: adepalatis)

Shamma Kabir (Kaggle user: ShammaKabir)

**Lab 8 Writeup**

For our first submission, we simply trained a logistic regression model with a 60% cut of the training data using the other 40% for cross-validation. We removed the feature MGR\_ID from the data before training. Naïvely, we then made predictions using this model, for which our score was quite low.

**Public score: 0.55115**

After realizing this mistake, we changed our code so that we would retrain the model using the entire training set after doing cross validation. In addition, we used a more thorough cross-validation method where we held 20% of the training data as the CV set and measured the training error, repeating this process 10 times and averaging the error at the end. Additionally, we looked at several discussion posts and gathered that the features ROLE\_ROLLUP\_1 and ROLE\_ROLLUP\_2 created too much variance, so we removed those from the feature set and added MGR\_ID back to it. In addition, we used one hot encoding to encode the data before training and testing.

**Public score: 0.87922**

We trained an XGB model for our second step to improving our score. However the XGB model did not help very significantly. At first the model resulted in a public score of 0.83479. This was lower than our score using a logistic regression; therefore, in order to improve the score we trained our model by increasing the depth and number of estimators. This resulted in an improved score of 0.87922.

**Public score: 0.87922**