Capstone 1 Project : In-depth Analysis
Fraud Detection in Mobile Payment Data
Springboard Data Science Career Track

Eric Cruz

Model Selection

Since this type of fraud detection is a binary classification problem, there are several Machine Learning Models which are usually suggested as good candidates for this type of problem. As we review each candidate, we will describe some model characteristics and parameters, in order to determine whether the results confirm our intuition about its suitability for this particular problem.

Note that our data includes labels for the classification task, so we will start with Supervised Learning models such a regression (Logistic Regression), decision trees (SVM), and clustering (KNN). We will also include a Bayesian Inferential Statistics approach, or Naive Bayes, such as MultiNomialNB() and GaussianNB(). A Neural Network MLP Classifier, Ensemble Learning methods, Random Forest Classifier and Gradient Boosting Classifier will be tested. We will use Cross Validation and Pipelining, and test hybrid combinations to develop the final model.

Unsupervised Learning models will be tested in the ensemble pipeline. PCA is an unsupervised dimensionality reduction technique that will be used in the feature selection analysis and implementation. For the sake of comparison, a KMeans clustering model will be compared to its supervised counterpart, KNN.

Preprocessing and Feature Selection

The data will be centered and normalized using the preprocessing function StandardScaler(). Further, we can convert the “Type” variable, for which we rejected the null hypothesis that distribution of type is independent of fraud, into a categorical feature indicating true/false for each of the 5 types. We will also create new features by combining some which are intuitively related, such as the difference between old and new balance because the fraud problem was described in the dataset description as typically emptying out an account’s entire balance. In addition to checking for correlation between all original, transformed, and created features, we will use GridsearchCV (cross validation) to measure and compare feature importance using various combinations. Those results will help determine which features to include in the final model.

Parameter and Hyperparameter Selection

The selection of parameters, and review of hyperparameters will be described for each model in the presentation of results, as we will iteratively review the tradeoffs between performance metrics and their applicability to our business goals.

Sampling Techniques and Imbalanced Data

The data will be partitioned automatically, according to the Train, Test, Split parameter selection, and we will create a separate holdout set while evaluating the optimal size for splitting. We will review SMOTE undersampling and oversampling results, again with the intention of tuning to the optimal split.

Measuring Model Performance

The results will be evaluated using metrics from the model fitting scores. We will unpack the confusion matrix, comparing the classification report and score function results to clarify the ordering of the matrix labels. We will plot the ROC curve, and also review the Precision, Accuracy, Recall, and F1 scores. Interpreting the metrics will be subject to our business goals, and the relative tradeoffs can be measured by this output.