Credit Card Fraud

Significant financial losses at the institutional, commercial, and individual level, can be attributed to credit card fraud and associated crimes such as identity theft. Fraudulent transactions and fraudulent active accounts accounted for 42% of complaints received by the U.S. Federal Trade Commission in 2016. Consumers reported losses in excess of 744 million $ US, with a median payout of us$ 450. Statistically however, credit card fraud accounts for a very minor percentage of transactions – an estimated 0.1%.

Project Societal Relevance

This project, a RoboGarden Machine Learning Bootcamp final project, was made possible through the public release of a Credit Card Dataset that captured two days’ worth of transactions by European cardholders. The relevancies of this project are multifold: it’s a practicum for a student new to the discipline of data science, it marks that students’ completion of their first steps into a the realm of data science, and most importantly, it demonstrates the utility of modern data science techniques to the problem of credit card fraud, thereby advocating for the discipline of Data Science and its growing role in commerce and credit spheres.

The Project

The approach used in this Credit Card Fraud project can be broken into three stages:

* Stage 1: Introduction to the data, in which the dataset is initially accessed and its parameters and facets are examined. It is here that potential data interrelationships, correlations, or unique problems with the dataset may be identified.
* Stage 2: Manipulation of the data. With actions determined by the results of Stage 1, here the data is reshaped and reformatted to correct for perceived deficiencies in the dataset and to prepare the data for final analysis in the next stage.
* Stage 3: in which the data is subject to analysis though regression, correlation or machine learning and analysis techniques as deemed appropriate by the investigator.

Stage One

Upon loading of the dataset and conversion to a dataframe, a first look at the data showed us a dataset consisting of 284807 rows (each row representing one unique transaction) of 31 columns, or features each. A look at the column header showed us that in addition to a timestamp for each row, the rows consisted of a set of blind columns labelled v1 through v28, a column listing the amount of that specific transaction, as well as a class column that differentiated valid and fraudulent transactions (false and true, respectively).

The columns were integers (time) or floats (v1-v28, amount) with the class column being a Boolean value.

A quick count of each of the two class values showed that the dataset consists of 284315 false values (valid transactions) and 492 true (fraudulent transactions): the data is overwhelmingly False, with True values accounting for 0.173% of the total dataset. This is a severely imbalanced dataset.

Dataset imbalance or the unequal distribution of classes in a dataset is a significant feature of our data that will need to be addressed in Stage 2.

Digging further into our data, a quick correlation of the first 30 features (columns 1-30) with our class feature produced few strong correlative instances. Neither time nor transaction amount showed significant correlation with class, while three of the unnamed ‘v’ features had weak (0.1 to 0.2) correlations and 11 ‘v’ features had correlations between -0.1 to -0.35. The rest of the correlations were very poor. The features showing the most significant correlations were used in subsequent investigations.

When plotting transaction Amount vs. Time, a prominent diurnal cyclicity was observed in the data, with the transaction rate making a clear dip, or decline, during what were presumably the late night-early morning hours, before increasing to a higher daily rate. There was no clear trend for the distribution of fraud instances that was observed.

Examining the transaction amount gave us a mean of 88.34, with a standard deviation of 250, a min. of 0.00, a 50% of 22.00, and a max. transaction of 25691.16. This maximum was a significant outlier, but the decision was made to retain it in the dataset.

From the earlier correlation work, the better correlated-with-class features were used in a pairsplot and multiple iterations of 2- and 3D point plots that demonstrated that while the true and false classes frequently had unique plotting distributions or clusters, those distributions tended to overlap and for plotting purposes, were not spatially unique.

Stage Two

As was noted before, the Credit Card Fraud dataset is severely imbalanced. Stage two of our project was where that imbalance, in preparation for analysis, was corrected. Before that, however, it was necessary to scale the time and amount columns in the dataset, as unscaled values in these columns would have a negative impact on subsequent results.

Taking advantage of the scaling, a final method of data visualization was applied to the data, t-Distributed Neighbor Embedding ( t-SNE ). T-SNE is an unsupervised, non-linear method for visualizing higher-dimensional data. The t-SNE plot supported the notion that the True and False classed instances in the dataset had many unique attributes, but the populations did have a degree of overlap. This method of visualization was included as an object of interest by the investigator.

As mentioned, our dataset was imbalanced, with cases of credit card fraud being extremely underrepresented in the dataset. There are two broad ways of responding to an imbalanced dataset: undersampling the dominant class, or oversampling the under-represented class. While both approaches have their advantages and disadvantages, the investigator made the decision to combine them in a two step process, first undersampling the valid transaction class(the ‘false’) then creating additional instances of the true class to arrive at a balanced dataset .

The investigator tried the dominant (valid transaction) dataset to various degrees, finally arriving on an amount of 20,000 samples. Coupled with the original complete number of fraud instances – 492 - arriving at a sample population where the True values (cases of fraud) were 4.7% of the total.

Oversampling was accomplished by first splitting the above dataset into test-train subpopulations at a 75%/25% ratio. The a process of Synthetic Minority Over-sampling Technique (SMOTE) was used to generate unique variations of the training set instances of Fraud sufficient to balance the True and False instances in the training dataset and achieve a 50/50 ratio.

Stage Three

In Stage Three, the processes used to fit multiple models to the training population were Logistic Regression, Linear Discriminant Analysis, K Nearest Neighbors, Decision Tree Classifier, and Random Forest. Each model was then applied to the test data set and the metrics for model accuracy were recorded for comparison.

With an imbalanced dataset, it’s conventionally accepted that using the standard confusion matrix is not the most effective way to gauge accuracy. We chose the AUC-ROC approach to display the performance of our models. Through repeated runs of varying sample sizes, Logistic Regression consistently generated a superior AUC score, closely followed by Linear Discriminant Analysis and Random Forest. At this point it was possible to incorporate confusion matrices for additional context of the results: by examining the confusion matrices we could see that although Logistic Regression produced the fewest misclassifications of false negatives i.e. mislabeling an actual Fraud event a valid transaction, it also produced the largest number of false positives – mislabeling a valid transaction as a fraud.. By these standards, Random Forest produced an outcome where the ratio of false positives to false negatives was the lowest.