Credit Card Fraud Analysis

Syed Shabbir

Bellevue University

Week 12 – Assignment 12.2

Term Project Paper

DSC530 – Exploratory Data Analysis

Credit Card Fraud Analysis

Statistical/Hypothetical Question

According to creditcards.com, there was over £300m in fraudulent credit card transactions in the UK in the first half of 2016, with banks preventing over £470m of fraud in the same period. The data shows that credit card fraud is rising, so there is an urgent need to continue to develop new, and improve current, fraud detection methods.

Credit card fraud is a major concern in the financial industry nowadays. It is estimated that £20M a day were lost due to fraudulent transactions in 2016 alone, totaling almost £770M annually (Financial Fraud Action UK https://www.financialfraudaction.org.uk/fraudfacts16/assets/fraud_the_facts.pdf).

Analyzing fraudulent transactions manually is unfeasible due to huge amounts of data and its complexity. However, given sufficiently informative features, one could expect it is possible to do using Machine Learning. This hypothesis will be explored in the project.

Outcome of your EDA

Using this dataset, I will use machine learning to develop a model that attempts to predict whether or not a transaction is fraudulent. To preserve anonymity, these data have been transformed using principal components analysis.

To begin this analysis, I will first train a random forest model to establish a benchmark, before looping back to Exploratory Data Analysis, looking at the most important predictive variables and testing other models.

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where I have found 492 frauds out of 284,807 transactions. From the outset it seems the dataset is highly unbalanced, instances of fraud account for 0.172% of all transactions. Some variables include:

- transaction amount
- point of sale
- currency
- country of the transaction
- merchant type
- Merchant Category Code
- Merchant city
- Time
- Transaction method

Were there any variables you felt could have helped in the analysis?

- The dataset is extremely unbalanced. Even a “null” classifier which always predicts class=0 would obtain over 99% accuracy on this task. This demonstrates that a simple measure of mean accuracy should not be used due to insensitivity to false negatives.

- Need to figure out the appropriate measures to use on this task.
- How can this dataset be transformed into(Oversampling, Under sampling, SMOTE)

Were there any variables you felt could have helped in the analysis?

Dataset contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response.

What challenges did you face, what did you not fully understand?

Fraud detection is a complex issue that requires a substantial amount of planning before throwing machine learning algorithms at it. Nonetheless, it is also an application of data science and machine learning for the good, which makes sure that the customer's money is safe and not easily tampered with.

Future work will include a comprehensive tuning of the Random Forest algorithm. Having a data set with non-anonymized features would make this particularly interesting as outputting the feature importance would enable one to see what specific factors are most important for detecting fraudulent transactions.