Cap Stone Proposal

Project: Credit Card Fraud Detection

Domain Background

North America was the continent most affected by data breaches in 2014, accounting for 1,164 or 76 percent of breaches in the world. The United States accounted for 1,107 of those breaches -- 72 percent of breaches in the world. Next in line were the United Kingdom (8 percent), Canada (4 percent), Australia (2 percent), Israel (1 percent) and China (1 percent).[1]

Fifty-four percent of data breaches in 2014 related to identity theft, 17 percent aimed at financial access and 11 percent sought account access. The remainder were considered nuisance attacks or attempts to get at intellectual property or classified information.[1]

Among the most highly-publicized breaches in recent years:

- EBay: 145 million records accessed.[1]
- Home Depot: 109 million records accessed.[1]
- JP Morgan Chase: 83 million records accessed.[1]
- Michael's Stores: 3 million records accessed.[1]
- Staples: 1.16 million records accessed.[1]
- Domino's Pizza: 650,000 records accessed.[1]
- Sony Pictures Entertainment: 47,000 records accessed.[1]
- Target: 40 million credit card numbers and 70 million addresses accessed.[2]
- Nieman Marcus: 350,000 cardholders impacted.[2]

Card fraud

Most card fraud occurs in the United States. In fact, a 2015 research note from Barclays stated that the U.S. is responsible for 47 percent of the world's card fraud despite only accounting for 24 percent of total worldwide card volume[2]

The high level of debit and credit card fraud in the United States also impacts other countries. Among U.K.-issued cards in 2015, 35 percent of fraud-related losses occurred in the United States, compared to 10 percent in France and Australia, 9 percent in Canada and 6 percent in Germany[3]

Cross-border fraud occurs when criminals use a consumer's credit or debit card data in one country to make fraudulent transactions in another country. In 2014, 47 percent of fraudulent cross-border transactions on U.K. credit cards took place in the United States[4]

U.S. credit card fraud is on the rise, too. About 31.8 million U.S. consumers had their credit cards breached in 2014, more than three times the number affected in 2013[5]

That fraud isn't cheap. Nearly 90 percent of card breach victims in 2014 received replacement credit cards, costing issuers as much as $12.75 per card [5]

Problem Statement

Anonymized credit card transactions to be labeled as fraudulent or genuine

Datasets and Inputs

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. 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-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

The dataset [7] has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on http://mlg.ulb.ac.be/BruFence and http://mlg.ulb.ac.be/ARTML

Solution Statement

The data is already processed, the features are not directly available due to security constraints, but PCA components are available instead. Trying with different supervised algorithms like Random Forest Classifier and Decision Tree Classifier and varying different parameters would help us in arriving at solution.

References:-

[0] https://www.kaggle.com/dalpozz/creditcardfraud

[1] Gemalto's 2014 Breach Level Index

[2] Barclays' Security in Payments: A Look into Fraud, Fraud Prevention, & the Future, May 22, 2015

[3] Financial Fraud Action UK's Fraud The Facts 2015

[4] FICO press release, June 25, 2015

[5] Javelin Strategy & Research 2015 Data Breach Fraud Impact Report

[6] http://www.creditcards.com/credit-card-news/credit-card-security-id-theft-fraud-statistics-1276.php

[7] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015