**Credit Card Fraud Detection**

**Abstract**

Now-a-days, Internet has become an important part of human’s life, a person can shop, invest, and perform all the banking task online. Due to the rise and rapid growth of E-Commerce, use of credit cards for online purchases has dramatically increased and it caused an explosion in the credit card fraud. In an era of digitalization, credit card fraud detection is of great importance to financial institutions. In this presentation, we will analyze credit card fraud detection using several Machine Learning classification Algorithms. After analyzing, our aim is to compare all the different methods to handle imbalanced data. The obtained results from databases of credit card transactions show the power of these techniques in the fight against banking fraud comparing them to others in the same field. It was determined that the Support Vector Machine algorithm had the highest performance rate for detecting credit card fraud under realistic conditions.

**Introduction**

Credit card fraud is a form of identity theft that involves an unauthorized taking of another’s credit card information for the purpose of charging purchases to the account or removing funds from it.

The financial services industry and the industries that involve financial transactions are suffering from fraud-related losses and damages. 2016 was a banner year for financial scammers. In the US alone, the number of customers who experienced fraud hit a record 15.4 million people, which is 16 percent higher than 2015. Fraudsters stole about $6 billion from banks last year. A shift to the digital space opens new channels for financial services distribution. It also created a rich environment for fraudsters.

Although less than 0.1% of all credit card transactions are fraudulent, analysts predict that Credit card fraud losses incurred by banks and credit-card companies can surpass $12 billion in the United States in 2020. Evidently, there is a dire need for robust detection of card-present and card-not-present fraudulent transactions to minimize monetary losses.

Technically speaking, two approaches are used to combat these crimes, the fraud prevention systems (FPS) and fraud detection systems (FDS). The FPS are the pro-active hardware and software mechanisms to avoid the fraud occurrence, on the other hand, the FDS are used when the fraudsters have overtaken the FPS. The main objective of the FDS is identify the fraud as soon as possible in order to take the necessary actions to revert it.

The fraud detection problem will come under supervised classification task aiming to determine whether a new transaction is legitimate or fraudulent one. Many techniques have been applied to credit card fraud detection, artificial neural network, support vector machine, decision tree, random forest, naïve bayes, logistic regression, and K nearest neighbors. A comparative analysis of logistic regression and naïve bayes is carried out.

This project evaluates Seven advanced data mining approaches, support vector machines and random forests, KNN, together with logistic regression, as part of an attempt to better detect credit card fraud while neural network, naïve bayes and logistic regression is applied on credit card fraud detection problem.

**Dataset**

The dataset 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-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. There are no “NA” values in the dataset.

**Machine learning Algorithms:**

**Supervised:**

1) Logistic Regression.

2) K-Nearest Neighbors

3) Naïve Bayes

4) Decision Tree

5) Random Forest

6) Support Vector Machines

7) Artificial Neural Network

**Un-Supervised:**

1)K-means Clustering

**Random Under-Sampling:**

"Random Under Sampling" which basically consists of removing data in order to have a more balanced dataset and thus avoiding our models to overfitting. Randomly remove samples from the majority class, with or without replacement. This is one of the earliest techniques used to alleviate imbalance in the dataset, however, it may increase the variance of the classifier and may potentially discard useful or important samples.

Once we determine how many instances are considered fraud transactions (Fraud = "1"), we should bring the non-fraud transactions to the same amount as fraud transactions (assuming we want a 50/50 ratio), this will be equivalent to 492 cases of fraud and 492 cases of non-fraud transactions.

**Over-Sampling:**

To oversample means to artificially create observations in our data set belonging to the class that is under represented in our data. One common technique is SMOTE — Synthetic Minority Over-sampling Technique. At a high level, SMOTE creates synthetic observations of the minority class (in this case, fraudulent transactions).