Capstone Project Proposal
for
Credit Card Fraud Detection

Credit Card Fraud Detection

To classify if the given credit card transactions are fraudulent or genuine. The datasets contain purchases made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days; it includes 492 frauds out of 284,807 transactions

1. Objective

   The goal is to classify if the given credit card transaction is fraudulent or genuine in the candidate dataset. The dataset is provided by Kaggle(www.kaggle.com), and it contains credit card transactions of EU users from Sep 2013.

   As part of this, I would like to find

   1. What are main characteristics of a fraudulent transaction?
   2. What are top 5 categories of a fraudulent transaction?
   3. What is the most common modus operand of a fraudulent operation?
   4. What is the min and max time lapsed before a fraudulent sale?
   5. Can we predict a fraud using ML techniques in real time?

2. The Client

   The client for this project is Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. The purpose is to find an ML model which can be used to detect a fraud in real time. It will not only save money(e-commence) but also boost customer confidence in using plastic money.

3. Data source

   The dataset is provided by Worldline and the ULB (Université Libre de Bruxelles) through www.kaggle.com.

   The data set credit: 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

4. Solution Approach

   My solution approach is specifically designed to achieve a controlled and structured approach to minimize data quality issues that may be present or introduced to protect personally identifiable information(PII). The solution is sub-divided into three phases

   a) Dataset preparation: This phase of the project is designed to clean (drop or adding missing data), add or update attributes and transform the dataset for analysis and discovery.

   b) Analysis and Discovery: This phase of the project is designed to validate and explore the dataset for all the problems listed in the “Problem” section of this proposal.

   c) Fraud Detection Model: In this phase of the project I will be exploring various machine learning algorithms to find the best model to classify a fraud.

   Please note that the above steps are NOT sequential in nature. I plan to take an iterative approach to improve the classification accuracy.

5. Project Deliverables

   a) Project deliverables are listed below.

      1. An analysis report (a .pdf document) on:
         a. List the main characteristics of a fraudulent transaction
         b. Identify top 5 categories under which of a fraudulent transaction is happening.
         c. Detailing the most common modus operand of a fraudulent transaction
         d. The min and max time elapsed before a fraudulent transaction is detected.
         e. ML model to detect a credit card fraud
      2. All project artifacts like – Design document, scripts, and code, test case, how to setup and reproduce the results.