**Credit Card Fraud Prediction**

**DSC630 Course Project: Milestone 2 - Data Selection and Project Proposal**

**Ganeshkumar Muthusamy, Vasanthakumar Kalaikkovan & Venkata Kanaparthi**

**Introduction**

**Background**

A credit card can help us to build credit, make convenient payments, and meet everyday expenses in our life. Unlike a debit card, which takes money from your checking account, a credit card uses the issuer’s money and then bills you later.

Most of the credit card provider has 24-hour cardholder information and assistance by phone to all cardholders calling from anywhere in the world, provided by Global Customer Care Services.

There are many kinds of credit cards available like rewards cards, secured cards, charge cards having common functionalities along with varied add-ons on them. When used correctly, credit cards can be a wonderful addition to your wallet.

**Problem Statement**

Credit card fraud is a major problem in financial services and costs billions of dollars every year. Credit card fraud continues to increase due to the rise and acceleration of Phone Order / Mail Order / E-Commerce. There has been tremendous use of credit cards for online shopping which led to a high amount of fraud related to credit cards. Financial institutions like Visa, MasterCard, Amex, and all debit networks have mandated that banks and merchants introduce EMV card technology to counter the fraud. In 2018, a total of $24.26 Billion was lost due to payment card fraud across the globe, and the USA is the most fraud-prone country. Credit card fraud was ranked the number one type of identity theft fraud. Credit card fraud increased by 18.4% in 2018 and is still climbing. There can be two kinds of card fraud, card-present fraud, and card-not-present fraud.

**Scope**

One of the main issues with credit card fraud detection is the unavailability of real datasets. Since the credit card data will have sensitive details about the customer, there are not many datasets available due to privacy issues. Another challenge with the credit card dataset is that it is highly imbalanced where there are more legitimate transactions and few fraudulent transactions. Millions of transactions are processed every day and the size of the dataset will be huge. The scope of this project would be to create a prediction model to detect credit card fraud based on the features in this dataset.

**Document Overview**

The overview for this project document contains the following aspects:

1. **Introduction**
   1. Background – the overview of the problem and the industrial background.
   2. Problem statement – the problem for which we are going to provide a solution.
   3. Scope – about how this project will be helpful in the real-time application.
   4. Document Overview – contains the overview of this proposal.
2. **Preliminary Requirement**
   1. Technical approach – elaborating the analytical approach to develop the project.
   2. Data source – mentioning the details of columns and the source for the dataset.
   3. Analysis – briefing the analysis going to perform in this project.
   4. Requirement Development – details about the required IDE and libraries used.
   5. Model deployment – brief discussion about the deployment of the model.
   6. Testing and evaluation – details of testing and evaluating the model.
3. **Execution and management of the project**
   1. Project Plan – milestone and timeline for the project.
   2. Project risk – contains any missing data for the analytics.

**Preliminary Requirement**

**Technical Approach**

The Cross-Industry Standard Process for Data Mining (CRISP-DM). This process model has 6 phases that naturally describe the data science life cycle for this project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

During every phase of this project lifecycle, we might discover new aspects/finding which we will incorporate them in ways to improve the efficiency of our model.

**Data Sources**

The dataset that is selected has transactions from European cardholders made in 2013. It has 285,000 transactions out of which 492 are fraudulent. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Due to privacy concerns, some principal components are PCA transformed. Time and Amount values are not transformed

1. Time - Number of seconds elapsed between this transaction and the first transaction.
2. V1- V28 – These are the result of a PCA Dimensionality reduction to protect user identities and sensitive features.
3. Amount – Transaction amount
4. Class – This is a response variable and has the values of 1 for fraudulent transactions, and 0 for non-fraudulent transactions.

Kaggle link for a dataset: <https://www.kaggle.com/mlg-ulb/creditcardfraud>

**Analysis**

We are planning to perform feature reduction and dimensionality reduction to select the most relevant variables for our model.  Many predictive algorithms assume the model variables follow a normal distribution. There are inherent advantages to using normally distributed variables, so our approach will focus on columns that closely follow this distribution. We will determine which variables are normally distributed by conducting the following analyses:

* Summary statistics on all variables: concentrating on mean and SD values.
* Identify the best model.

**Requirement Development**

For this project, we are planning to use python. To complete this project in python the following are required for our development:

IDE: Jupyter Notebook

Libraries:

1. Numpy – extensively used for data analysis to handle multidimensional arrays.
2. Pandas – high-performance data structures and analysis tools for the labeled data.
3. Matplotlib –powerful visualizations which create several stories with the data visualized.
4. SciPy – used for high-level technical computations.
5. Seaborn – interface for drawing attractive and informative statistical graphics.

**Model Deployment**

We will use feature selection techniques to finalize our feature list for models. Using the results from this step, we will build a couple of classification models and evaluate their performance. Below are some example models.

* **Random forest:** It builds multiple decision trees and merges them to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.
* **Logistic regression:** It is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.
* **Decision Tree:** It belongs to the family of supervised learning algorithms. It can be used for solving regression and classification problems. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from training data. In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. Based on the comparison, we follow the branch corresponding to that value and jump to the next node.

**Testing and Evaluation**

We will be splitting the data into 70% training and 30% test dataset. Using the test dataset, we will test the model. confusion matrix, AUC, and F1 scores will be used for the evaluation. cross-validation will also be used to decide the best model.

**Expected Results**

Using this model to determine the transactions are fraudulent or not.

**Execution and Management of Project**

**Project Plan**

Week 1: Team Information and Communication Plan

Week 2: Data Selection and Project Proposal

Week 3: Data set cleanup

Week 4: Graph Analysis & EDA

Week 5: Peer review & discussion

Week 6: Review different classification models

Week 7: Train and test the models

Week 8: Improve model efficiency

Week 9: Merge and finalize

Week 10: Review and adjust accordingly

Week 11: Work on documentation

Week 12: Final project paper and presentation

**Project Risk**

The biggest challenge is we don’t have the visibility of the type of data in columns V1-28. Which would be required to improve the efficiency of the model.