Practicum Project Proposal

DSA-5900: Professional Practice (Summer 2022)
Credit Hours: four

Credit Card Fraud Detection

By

David Nnamdi

Faculty Sponsor: Dr. Brian Fiedler

This proposal is presented in partial fulfilment of the requirements of the DSA 5900 course

Data Science and Analytics Institute
Gallogly College of Engineering
University of Oklahoma
Norman, USA

1.0 Introduction

Credit card companies require a means for definitively identifying fraudulent credit card transactions. This is done for a variety of reasons including ensuring the customers are not charged and to make loss insurance claims.

This project aims to develop a classification model to predict whether a logged credit card transaction is fraudulent or not. The dataset was collected and analyzed during a research collaboration between Worldline and Université Libre de Bruxelles (ULB) and was made available on the Kaggle dataset repository in 2018.

2.0 Project Objectives

The objectives of this project are given below:

- Explore 30 input variables (including 28 anonymized input variables resulting from PCA) and identify the key variables that predict fraudulent transactions
- Explore different sampling techniques for dealing with inherent data imbalance
- Create a well-tuned Machine Learning model(s) that classifies a credit card transaction as fraudulent or not using:
    - Traditional ML / Deep Learning methods
    - Auto ML methods (e.g., PyCaret which uses Python package for automated preprocessing and modeling)
- Compare Traditional ML vs Auto ML methods performance

3.0 Plan

Data Description

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. Due to confidentiality issues, original features and more background information could not be made publicly available.

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**'. '**Time**' contains the seconds elapsed between each transaction and the first transaction in the dataset. '**Amount**' is the transaction Amount. Feature '**Class**' is the response variable, and it takes value 1 in case of fraud and 0 otherwise

Modeling Methods

General statistical and machine learning methods will be applied at various stages of the project. In the initial stages, exploratory data analysis will play a leading role in identifying key predictors of fraudulent transactions. This will be done using various visualization techniques of gathered data, statistical correlations will also be explored to determine real key factors necessary for content creation.

4 different types of sampling techniques will be explored to deal with data imbalance. They are:

- Under sampling (the majority class, i.e., non-fraudulent transactions)
- Oversampling (the minority class, i.e., fraudulent transactions)
- ROSE sampling
- SMOTE sampling

Dimension reduction methods such as LDA or clustering techniques such as t-SNE to be evaluated

In the modeling stages, core ML classification models will be assessed. Methods such as Logistic regression, Bayesian modeling (e.g., Naïve Bayes), SVM and other more advanced tree and ensemble methods may be evaluated. Deep learning methods using Artificial neural networks will also be explored. For each promising model built there will be hyperparameter tuning to improve model performance. Methods such as traditional **Grid Search** and more advanced methods such as **Optuna** will be utilized for hyperparameter tuning.

Subsequently, the Auto ML tool **PyCaret** will be used to build and tune classification models and then a comparison of performance will be done to determine the feasibility of fast prototyping while still maintaining good accuracy will be explored.

Final model selection in both cases will be based on performance and loss/accuracy metrics

4.0 Deliverables

There are two major deliverables from this project:

- A Machine Learning model that predicts the fraudulent credit card transactions
- An analysis on the core signals that are of high importance in classification
- A comparison of traditional ML to Auto ML methods for model building

5.0 Schedule

Milestones for this project are listed below:

- Milestone 1: Setting up environment, extracting dataset, exploratory data analysis (EDA) and identifying candidate features (Due 06/10/2022)
- Milestone 2: Creating training/validation datasets with the identified features after dealing with imbalance (Due 06/17/2022)
- Milestone 3: Training and Tuning the models (Due 07/01/2022)
- Milestone 4: Assessing the model quality and analyzing top features (Due 07/8/2022)
- Milestone 5: Creating a comprehensive report on all findings (Due 07/15/2022)