HarvardX PH125.9x Credit Card Fraud Prediction Project

Rogelio Montemayor

June 25, 2021

1 Introduction and overview

1.1 Introduction

The goal of this project is to create a model that identifies fraudulent credit card transactions.

The dataset contains transactions made over two days in September 2013 by European cardholders. The main problem with this type of problem is that the data is highly unbalanced. Only 0.173% of the transactions in this dataset are fraudulent. There are 284,807 transactions of which 492 are fraudulent transactions.

Due to privacy and confidentiality issues, features ***V1*** to ***V28*** are the principal components of the original features after PCA transformation. There are only two features that are not transformed: ***Time*** and ***Amount***. The target variable is ***Class***, and it takes the value of 1 in case of fraud and 0 otherwise.

This is a classification project.

It is recommended to use Area Under the Precision-Recall Curve (AUPRC) to measure accuracy because the classes are so unbalanced. Even though Area Under Receiver Operating Characteristic (AUROC) is more common, it is not recommended for highly unbalanced classification.

The best results were obtained using a Extreme Gradient Boosting algorithm. The AUPRC of the best model was **XXXXX**.

Credit card fraud increases costs for everyone and machine learning techniques can help flag fraudulent transactions and lower costs for banks, merchants and their customers.

1.1.1 Acknowledgements

My version of the dataset was downloaded from Kaggle: https://www.kaggle.com/mlg-ulb/creditcardfraud.

The dataset 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.

I want to thank Max Kuhn (https://topepo.github.io/caret/) and DataCamp (https://www.datacamp.com)

More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project.

Please refer to the following papers:

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

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. ***Learned lessons in credit card fraud detection from a practitioner perspective***, Expert systems with applications, 41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. ***Credit card fraud detection: a realistic modeling and a novel learning strategy***, IEEE transactions on neural networks and learning systems, 29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea ***Adaptive Machine learning for credit card fraud detection*** ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. ***Scarff: a scalable framework for streaming credit card fraud detection with Spark***, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. ***Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization***, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi ***Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection***, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi ***Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection*** Information Sciences, 2019

Yann-Aël Le Borgne, Gianluca Bontempi ***Machine Learning for Credit Card Fraud Detection*** - Practical Handbook

1.2 Overview

These are the steps we will follow to go from raw dataset to model and insights:

- Decompress, read, and build the dataset
- Analysis
  - Initial Exploratory Analysis
  - Visual Analysis
  - Data cleaning and feature engineering

    - Modeling approach
        - Split the dataset
        - Preprocess and Setup
        - Train models
        - Algorithm selection
- Results and final model
- Conclusion and insights

1.3 Read in the compressed file

2 Analysis

2.1 Initial Exploratory Analysis

2.2 Visual Analysis

2.3 Data Cleaning and Feature Engineering

2.4 Modeling Approach

2.4.1 Regularized Logistic Regression with glmnet

2.4.2 Support Vector Machines with svmLinear

2.4.3 Random Forest with ranger

2.4.4 Extreme Gradient Boosting with xgbTree

2.4.5 Algorithm Selection

3 Results and Final Model

4 Conclusion