**Project Milestone 1**

Adil Khan

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

Oct, 01, 2022

## *Topic*

***Detect Credit Card Fraud***

## Abstract

A machine learning model that can accurately detect whether a credit card transaction seems to be fraudulent. I'll start with exploratory data analysis and utilise machine learning techniques to comprehend and process this data. I will develop, train, and test a number of models to identify which transactions are fraudulent by examining the data set and recognising trends. My final goal is to test and choose the model that is the most accurate for this collection of data. The Machine Learning Group [http://mlg.ulb.ac.be/] and Worldline are the sources of the information used. There are both fraudulent and legitimate transactions included in this dataset of card transactions. There are 31 features and about 285,000 rows of data. For security reasons, 28 of these features are unavailable to the general public; nevertheless, the three I'm interested in are Time, Amount, and Class (fraudulent or non-fraudulent). We have a project opportunity for a classification using this data format. A process called classification uses data that has been divided up into a number of classifications. Such a project aims to determine what class or category a new data item will belong to. In light of the rising automation and electronicization of our environment, it is crucial to examine this data. We must continue to be vigilant in our search for strategies to weed out fraudulent information since credit card theft is on the rise. The created model(s) may prove to be of great use in today's increasingly computerised environment, where millions of transactions—many of them fraudulent—take place every day.

**Which Data?**

A Card Transactions dataset that includes both fraudulent and legitimate transactions makes up the data being used. There are 31 features and about 285,000 rows of data. For security reasons, 28 of these features are unavailable to the general public; nevertheless, the three I'm interested in are Time, Amount, and Class (fraudulent or non-fraudulent). On [https://www.kaggle.com/mlg-ulb/creditcardfraud/home], you may find the data set itself.

**Research Questions? Benefits? Why analyze these data?**

We are able to focus a little bit on our strategy because our data only covers two days' worth of transactions. When we look at our Time feature, we can anticipate that the distribution will be largely normal, which allows us to make a number of assumptions. In order to determine whether the normal Time distribution really does exist, I need to check the credit card transaction amounts as well as the number of fraudulent and legitimate transactions included in the original data set. We will be able to formulate pertinent study questions, such as When do we notice the most fraudulent activity?, by obtaining descriptive data on these factors. How much does a fraudulent transaction typically cost?

**Ethical Consideration**

Given how automated and technological our society is becoming, it is crucial to study this data. We must keep looking diligently for ways to weed out the lies since credit card theft is on the rise.

**What Method?**

I'll start by performing exploratory data analysis and then apply machine learning algorithms to comprehend and process this data. I'll start by using logistic regression to determine how inaccurate a forecast is and random forests to determine whether a transaction is indeed fraudulent. After that, I'll evaluate the models' precision, recall, and accuracy. To more clearly see how each model affects the data, it would also be worthwhile to look at the confusion matrix for this particular data set.

**Potential Issues?**

I think I have all I need to do this analysis using the knowledge I've gained from earlier classes. Having said that, I have had unanticipated project setbacks in the past owing to disparities in data features, previously untried methodologies, etc., that I have not yet been given a responsibility for. Working on a new project on my own will probably include some learning curves, but at this moment I do not anticipate any significant difficulties.

**Concluding Remarks**

In order to construct the most accurate model to determine whether a transaction constitutes fraudulent conduct, I want to evaluate a data set comprising both fraudulent and legitimate credit card transactions. In today's increasingly digitised world, where millions of transactions—many of them fraudulent—take place every day, this kind of approach might be incredibly helpful. I will design and train many models to predict which transactions are fraudulent by examining the data

set and recognising trends. My final goal is to test and choose the best accurate model for this collection of data.

**References:**

The domain from which the data is derived is from the organization Worldline and the Machine Learning Group [http://mlg.ulb.ac.be/].


1. Albon, C. (2018). Machine learning with Python cookbook: practical solutions from preprocessing to deep learning. Sebastopol, CA: OReilly Media.
    a. Alternatively, this textbook presents algorithmic method for data analysis and deep learning using the Python language.
2. Machine Learning Group. (2018, March 23). Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/mlg-ulb/creditcardfraud/home
    a. This is my data source and general description of our data set, housed within Kaggle.
3. Machine Learning Group. (n.d.). DEFEATFRAUD: Assessment and validation of deep feature engineering and learning solutions for fraud detection. Retrieved from https://mlg.ulb.ac.be/wordpress/portfolio_page/defeatfraud-assessment-and-validation-of-deep-feature-engineering-and-learning-solutions-for-fraud-detection/
    a. This project description outlines the overall goals of the Machine Learning Group as they attempt to develop new and improve upon existing mechanisms for detecting credit card fraud transactions using machine learning algorithms.
4. Couronne, R., Probst, P., & Boulesteix, A.-L. (2018, July 17). Random forest versus logistic regression: a large-scale benchmark experiment. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5
    a. This research article discusses the differences, benefits and disadvantages to using random forest versus logistic regression algorithms for regression and classification.
5. DataFlair Team. (2019, October 11). 11 Top Machine Learning Algorithms used by Data Scientists. Retrieved from https://data-flair.training/blogs/machine-learning-algorithms/
    a. This article lays out some of the most commonly used machine learning algorithms for data analysis and model development. The article begins by discussing supervised learning algorithms, and moves into unsupervised algorithms, to cover a breadth of available options.
6. DataFlair Team. (2020, February 19). Project in R - Uber Data Analysis Project. Retrieved from https://data-flair.training/blogs/r-data-science-project-uber-data-analysis/
    a. This is an unrelated project which analyzes Uber pickup data for New York City. The assessment of the data and general work through are helpful as they provide a sort of vague wireframe for how to address a data set with the intent to create predictive models.
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: with applications in R. New York: Springer.
    a. This textbook helps to review many statistical methods for data analysis using R, including logistic regression and random forests.

8. Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Hoboken, NJ: Wiley.
    a. A textbook which offers a guide to data visualization, from the exploratory step through the project presentation step.
9. Nadim, A. H., Sayem, I. M., Mutsuddy, A., & Chowdhury, M. S. (2020, February 13). Retrieved from https://ieeexplore.ieee.org/document/8995753
    a. A secondary article discussed the use of machine learning algorithms to address credit card fraud, investigating the use of various regression algorithms in the process.
10. Puh, M., & Brkić, L. (2019, July 11). Detecting Credit Card Fraud Using Selected Machine Learning Algorithms. Retrieved from https://ieeexplore.ieee.org/document/8757212
    a. This article discusses the growth in interest for applying machine learning techniques to the mission of detection fraudulent credit card transactions and points out the challenges that can arise with these attempts.