# CIE 417 Course Project Final Report

Rawan Eldalil
s-rawaneldalil@zewailcity.edu.eg

January 14, 2021

## Problem Definition and Motivation

Credit-card fraud occurs when some individuals attempt to obtain financial gains from other people unlawfully [1]. These frauds happen by means of:

- Owners losing their credit cards or get them stolen by someone.
- Owners accidentally exposing their credit card number to strangers.
- Criminals getting a counterfeit credit cards.

As these kind of activities causes financial losses and get people involved in financial activities they didn't commit, the detection and prevention of these criminal activities is a necessity to protect people's interests. Credit-card fraud detection is a set precautionary measures/activities applied to identify the presence of fraudulent activities, then take necessary actions to prevent them. Fraud detection can be achieved using a Machine Learning (ML) model which learns from a collected data to estimate the probability of whether a transaction is fraudulent or not. In the course project, one's aim is to train different classifiers for the pre-mentioned purpose and check which classifier has the most reasonable performance.

## Literature Review

The dataset used for training the model is 'Credit Card Fraud Detection' Kaggle dataset; it can be found in [2]. The dataset suffers from high imbalance, as the number of the fraudulent transactions are only 492 out of 284,807 total transactions. As one searched for other projects tackling the same problem, one found most-voted Kaggle notebook that is tackling the same problem using the same dataset as one tackles in this project [3]. That project uses different classifiers to see how they can accurately detect whether a transaction is fraudulent or not. For that purpose, the random under-sampling technique is used to address the problem of high imbalance. Furthermore, the classifiers that were used in that project were logistic regression, decision tree, SVC and KNN models. The logistic regression classifier turns out to have the highest score among other classifiers; its F1 score was 82%. Yet, it doesn't account for the runtime each model take to make predictions; the significance of knowing the runtime of each model lies in the fact that the detection problem is real-time problem (i.e. if there is a fraudulent transaction, the time taken to detect it as fraudulent should be minimal). Accordingly, the runtime is accounted for beside trying ensemble models vs the logistic regression model to see if they perform better than the logistic regression model.

## Approach and Methodology

### Data Exploration and Pre-processing

All data exploration steps were made by Pandas Profiling; it is a Python module which helps to visualize and understand the distribution of each attribute/feature in the dataset by generating a web-formatted report with all the dataset information displayed [4]. The results of exploratory analysis guided us to existence of 1081 duplicate rows, which were dropped, and low correlation between the 'Class' and some of the dataset features, which were identified and dropped for enhanced performance and better results.

### Applying the Random Under-sampling Technique

As previously mentioned, the dataset is highly imbalanced; one way to solve this problem is to use the random under-sampling technique. The way this technique works is to select a random sample of the majority class that is equal in number to the minority class, then, another dataset is created by concatenating the minority class records with the taken random sample. Now, a new dataset, that will be used in the model train-test process, is obtained.

### Model Selection

Four different sklearn classifiers were chosen for learning, three of which are ensemble models and the forth is logistic regression model. The ensemble models trained were Random Forest classifier, AdaBoost classifier and Gradient-Boosting classifier. The reason behind choosing ensemble models that they often have the capability to make better predictions over single predictive models. As one stated previously based on the work in [3] that the logistic regression model's performance was the best among other classifiers, this is the motivation behind choosing it to compare its performance to the ensemble models.

### Model Evaluation

The model performance is evaluated based on different metrics in addition to the average runtime taken from a classifier to make predictions; the metrics used are the accuracy, recall, precision and the F1 score, yet the principal metric is F1 score. One also evaluates the performance of the model based on the F1 score metric of the predictions made on data that are not included in the under-sampled dataset (i.e. unseen data)

## Results

The result of each classifier came as follows:

- The Random Forest Classifier's F1 score is 0.9231, and its average runtime is 0.0153 seconds.
- The Ada-Boost Classifier's F1 score is 0.9231, and its average runtime is 0.0206 seconds.
- The Gradient-Boosting Classifier's F1 score is 0.9125, and its average runtime is 0.0039 seconds.
- The Logistic Regression Classifier's F1 score is 0.9337, and its average runtime is 0.0023 seconds.

The logistic regression model again have the best performance among the other models. Regarding, the model's F1 score on unseen data is 0.9869.

## Discussion

### Model Performance

Looking at the confusion matrix of the logistic regression model, one shall find 17 are predicted to be non-fraudulent transaction, while actually being fraudulent; this is still to some extent large number in regards to our problem, and accordingly the model needs some improvements. Furthermore, one took 2% of the records in the original 'ready dataset' other than the ones in the under-sampled dataset, which are all non-fraudulent transactions, to further test the model; the result was out of 5656 transactions, there are 139 transactions predicted to be fraudulent, while they are not.This is considered to be also large number of falsely fraudulent-reported transactions.

### Ethical and Professional Considerations

The model can make predictions unfairly in demographic groups other than those out of which the transactions data were collected (i.e. the European credit-card holders). In regard to model's bias, ever since the original features of the dataset are not provided, one cannot really tell if there are unwarranted associations the model can make. Also, the dataset used in this project was collected in 2013 and may not be a

good choice for training a model can be used/work in 2021, for instance. So, when developing a model that treats a problem like fraud detection, something like *Concept Drift* should be taken into account. In real world, the data collected about credit-card holders can change in accordance with changes taking place nationally or internationally; these changes can be economical, social...etc, and it could cause the holders' behavior to change, resulting in the so-called concept drift. Accordingly, the model may not accommodate these changes and consequently result in unsatisfying performance because the model learning process didn't accommodate these kind of changes. So, solving this problem may require one to use algorithms/-metrics for drift detection [5]. Alternatively, a periodic update with new historical data to the model can be conducted [6].

## Conclusion

The course project addresses the problem of the credit-card fraud by developing a ML model that helps detecting frauds by making predictions whether a transaction is fraudulent or not. Different models were trained for that purpose; they are Random Forest classifier, AdaBoost classifier, Gradient-Boosting classifier and the Logistic Regression classifier. The logistic regression model has the most reasonable performance among other models, yet it still needs improvements. The model can make predictions unfairly in demographic groups other than those out of which the transactions data were collected. Also, it is more convenient to use a more newly-collected dataset for more adequately-working model.

## References

[1] says: IHUOMABASIL. Credit Card Fraud Detection Solutions To Secure Your Business [Internet]. 2020 [cited 2021Jan14]. Available from: https://spd.group/machine-learning/credit-card-fraud-detection/

[2] ULB MLG-. Credit Card Fraud Detection [Internet]. Kaggle. 2018 [cited 2021Jan14]. Available from:https://www.kaggle.com/mlg-ulb/creditcardfraud

[3] Janiobachmann. Credit Fraud || Dealing with Imbalanced Datasets [Internet]. Kaggle. Kaggle; 2019 [cited 2021Jan14]. Available from: https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets

[4] Lafuente AS. Exploratory Data Analysis with Pandas Profiling [Internet]. Medium. Towards Data Science; 2020 [cited 2021Jan14]. Available from: https://towardsdatascience.com/exploratory-data-analysis-with-pandas-profiling-de3aae2ddff3

[5] Data Drift and Machine Learning Model Sustainability [Internet]. Analytics Insight. 2020 [cited 2021Jan14]. Available from: https://www.analyticsinsight.net/data-drift-and-machine-learning-model-sustainability/

[6] Brownlee J. A Gentle Introduction to Concept Drift in Machine Learning [Internet]. Machine Learning Mastery. 2020 [cited 2021Jan14]. Available from: https://machinelearningmastery.com/gentle-introduction-concept-drift-machine-learning/