# CREDIT CARD FRAUD DETECTION UNDER EXTREME CLASS IMBALANCE

**Project Description:** —Credit card based online payments has grown intensely, compelling the financial organizations to implement and continuously improve their fraud detection system. It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. However due to the imbalance data issue, many observations can be predicted as Normal transaction, whereas they are actually Fraud transactions. If we predict that a transaction is fraudulent and turns out not to be, is not a massive problem compared to the opposite situation. Naturally, a credit card owner will not be happy if the credit card is blocked by the bank when no actual fraud had taken place. Machine Learning algorithms are very likely to produce faulty classifiers when they are trained with imbalanced datasets. These algorithms tend to show a bias for the majority class, treating the minority class as a noise in the dataset. With many standard classifier algorithms, such as Logistic Regression, Naive Bayes and Decision Trees, there is a likelihood of the wrong classification of the minority class. This project speaks about the different methods to handle the imbalanced data and the importance.

**Goals and objectives:**

* The goal of this project is to detect anonymous credit card transactions and label it as fraudulent or genuine in transactional data. Fraud detection involves monitoring the behavior of users to estimate, detect, or avoid undesirable behavior.
* The objective is to handle data imbalance problem and create simple and commonly used machine learning models like logistic regression, Random forest and maybe others to compare how they perform regarding the metric chosen (Precision, Recall etc) for the task of predicting fraudulent credit card transactions.

**Project Requirements:**

* [Python](https://www.python.org/)
* NumPy (for documentation:<http://www.numpy.org/>)
* Pandas (for documentation:<http://pandas.pydata.org/>)
* Scikit-Learn (for documentation:<http://scikit-learn.org/stable/>)
* itertools (<https://docs.python.org/3/library/itertools.html>)
* Tensorflow
* Keras

**Problems to be addressed**

* ***Imbalanced data*** can be a serious problem for building predictive models, as it can affect our prediction capabilities and mask the fact that our model is not doing so good

**Potential pitfalls & challenges**

***Challenge 1:*** Choosing the best algorithm which predicts the fraud data accurately. Recall metric (i.e. what percentage of fraud cases were actually detected as fraud) can be used.

**Background Research:** While an incredible amount of transactions are made everyday, it is no doubt that some transactions may be fraudulent. As a result, credit card companies usually deploy sophisticated algorithms to identify suspicious transactions and prevent them from going through. Classification techniques are the promising solutions to detect the fraud and non-fraud transactions. Unfortunately, in a certain condition, classification techniques do not perform well when it comes to huge numbers of differences in minority and majority cases. Hence this project speaks about techniques that were applied in the credit card dataset to overcome the rare events.

* <https://thesai.org/Downloads/Volume9No11/Paper_55-Handling_Class_Imbalance_in_Credit_Card_Fraud.pdf>
* <https://arxiv.org/pdf/1608.06048.pdf>
* <https://www.hindawi.com/journals/complexity/2018/5764370/>
* <https://pdfs.semanticscholar.org/0be1/e1f748845244bf8ff4041bb5e7d35b9057ee.pdf?_ga=2.64650601.731729570.1553044261-1175306196.1553044261>
* <https://www.researchgate.net/publication/326986162_Credit_Card_Fraud_Detection_Using_Machine_Learning_As_Data_Mining_Technique>
* Given the imbalanced dataset, taking these papers as reference, we will first evaluate different resampling methods such as Random oversampling, Random under sampling, SMOTE etc. to overcome the skewed distribution in class. We will then build different models and compare the overall performance.
* We will also be building predictive models

**Algorithms and code sources:**

1. Imblearn provides some great functionality for dealing with imbalanced data.

* **For Random Under sampling:** <https://imbalanced-learn.org/en/stable/generated/imblearn.under_sampling.RandomUnderSampler.html>
* **For Random Oversampling:** <https://imbalanced-learn.readthedocs.io/en/stable/over_sampling.html>
* **For SMOTE:**

<https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html>

* **For SMOTE-ENN:** <https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.combine.SMOTEENN.html>
* **For ADASYN:** <https://imbalanced-learn.org/en/stable/generated/imblearn.over_sampling.ADASYN.html>

1. For most of the classification algorithms

<https://scikit-learn.org/stable/supervised_learning.html>

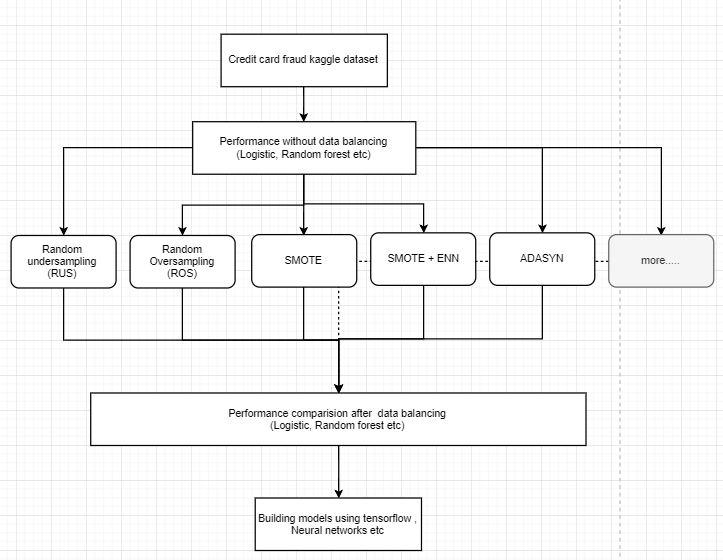
* For Gradient Boosting Classifier [**http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html**](http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html)
* For Random Forest Algorithm

[**https://www.kaggle.com/randyrose2017/using-scikit-learn-and-keras-for-fraud-detection**](https://www.kaggle.com/randyrose2017/using-scikit-learn-and-keras-for-fraud-detection)

1. We use t- SNE for visualizing the data and tensor flow to build the predictive model.

[**https://www.datascience.com/blog/fraud-detection-with-tensorflow**](https://www.datascience.com/blog/fraud-detection-with-tensorflow)

**Block Diagram of our system in more detail (optional)**



**Step 1:** Exploratory data analysis

**Step 2:** Fraud Detection (classification) analysis without data balancing.

**Step 3:** In the third step, we will use techniques such as RUS, ROS, SMOTE, SMOTE-ENN among others to take care of imbalance in the dataset before applying classification algorithms such as Logistic Regression, Random Forest to compare the results with the results of step-2 of the analysis.

**Step 4:** Predictive models to see how accurate they are in detecting whether a transaction is a normal payment or a fraud.

**Data sources:** The required dataset is available on the Kaggle website.

Link: <https://www.kaggle.com/qyzhan/credit-card-fraud-detection/data>.In this data we do not have the original features since we have the PCA transformation of it. The datasets contain 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.

***Variables of the dataset***

* V1, V2, ... V28 are the principal components obtained with PCA
* 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
* 'Amount' is the transaction Amount
* 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise

**References:**

* 1. <https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9>
  2. <https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8>
  3. <https://www.analyzeinsights.com/single-post/2017/10/16/Part-1-Credit-Card-Fraud-Detection-Analysis-on-Imbalanced-Data>
  4. <https://www.kaggle.com/zhouhq/detecting-credit-fraud-through-neural-network>
  5. <https://thesai.org/Downloads/Volume9No11/Paper_55-Handling_Class_Imbalance_in_Credit_Card_Fraud.pdf>
  6. <https://www.kaggle.com/gargmanish/how-to-handle-imbalance-data-study-in-detail>
  7. <https://medium.com/coinmonks/handling-imbalanced-datasets-predicting-credit-card-fraud-544f5e74e0fd>

**Project members:**

* Monisha Vodnala – NEU ID 001825522
* Harshitha Sanikommu- NEU ID 001423529
* Fairy Dmonte - NEU ID 001498529