PHASE 4 REPORT :

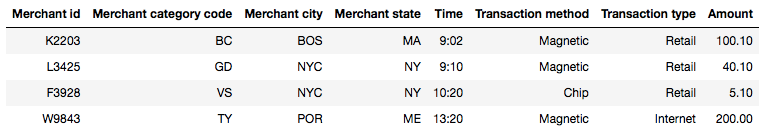
FEATURE ENGINEERING:



Feature engineering is the key to success in modeling.

Credit card fraud happens basically in two types: application fraud and transaction fraud. Application fraud is similar to identity fraud in that one person uses another person’s data to obtain a new card. Transaction fraud happens when a card is stolen or a lost card is obtained to conduct fraudulent transactions. Also, there has been a significant rise in counterfeit cards.

Feature Technicalities:

* **PCA Transformation:**The description of the data says that all the features went through a PCA transformation (Dimensionality Reduction technique) (Except for time and amount).
* **Scaling:** Keep in mind that in order to implement a PCA transformation features need to be previously scaled. (In this case, all the V features have been scaled or at least that is what we are assuming the people that develop the dataset did.)
* A fraudster will try to abuse the card as much as possible in a **short period** before the card is detected and suspended. So we should see **abnormal**transactions in a short period. With this goal, if we aggregate transactions over some time, we shall be able to discover abrupt changes.
* Let me present an example to demonstrate how features can be created. The table shows some transactions of a cardholder. The transactional data include the merchant id, the category of the merchant, the location of the merchant, a timestamp, the transaction method and type, and the transaction amount.
* 

MODEL TRAINING :

These applications are used as machine learning for credit card fraud detection:

Decision Tree and Random Forest

Artificial Neural Network

Naive Bayes

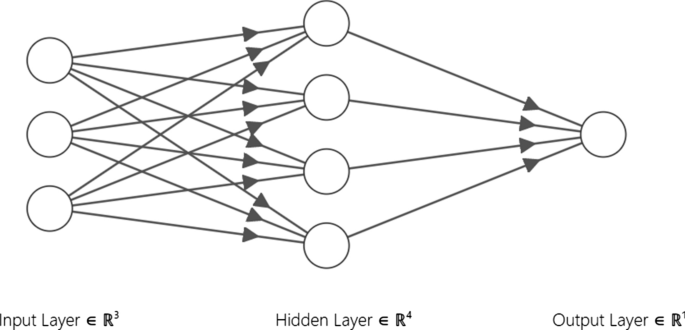
### Decision trees and random forest

Decision Tree (DT) is a supervised ML based approach that is utilized to solve regression and classification tasks. A DT contains the following types of nodes: root node, decision node and leaf node. The root node is the starting point of the algorithm. The decision node is a point whereby a choice is made in order to split the tree. A leaf node represents a final decision [[7](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8#ref-CR7)]. The RF method conducts its predictions by using an ensemble of DTs [[8](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8#ref-CR8)]. In the RF, a decision is reached by majority vote.

Given a number of trees *k*, a RF is defined as, RF = {�(�,��)}, where {��} represents independent identically distributed trees that cast a vote on input vector *X*. The label with the most votes is the prediction.

### Artificial Neural Network

Artificial Neural Network (ANN) is a supervised ML method that is inspired from the inner workings of the human brain. The simplest ANN have the following basic structure: an input layer, one hidden layer and an output layer. The input layer size is based on the number of features in a given dataset. The hidden layer size can be varied based on the complexity of a task and the output layer size depends on the type of problems to be solved. The most basic component of an ANN is a node or neuron. In this research, we consider feed forward ANNs. Therefore, the information flows in one direction (from its input to its output) through a neuron [[12](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8#ref-CR12)]. Figure [1](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8#Fig1) depicts a graphical representation of a simple ANN with 3 nodes in the input layer, a hidden layer with 4 nodes and an output layer with 1 node.

[](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8/figures/1)

### Naive Bayes

The Naive Bayes (NB) is a supervised ML technique that is based on Bayes’ theorem. The NB method assumes the independence of each pair of attributes when provided with the dependant variable (the class). In this research, the Gaussian NB (GNB) classifier was used.

EVALUATION :

To evaluate this machine learning models we considered

two different method namely;

(1) Classification accuracy, which is the ratio of number

of correct prediction to the number of input sample,

as seen in equation 6. But this is very effective only if

there are equal number of samples in each class.

Accuracy = no. of correct prediction/total no. of prediction

(2) Confusion Matrix:

This gives a matrix as output and describe the complete performance of the model. Four

essential measurements are utilized in evaluating the

analyses, to be specific True Positive Ratio (TPR),

True Negative Ratio (TNR), False Positive Ratio

(FPR) and False Negative Ratio (FNR) rates metric

individually.

In which true positive, true negative, false positive and

false negative are the quantity characterized by true positive,

false positive, true negative, and false negative experiments