**ML1819 Research Assignment 1: Report**

1. **Data & Pre-Processing**

We downloaded ‘creditcard.csv’ dataset from www.kaggle.com, with 285k cases and 31 variables. The target is to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

The first impression of the dataset is that there is no missing data and other than Time and Amount, all other columns are already scaled from 0 to 1. Data of Amount column has been scaled using the ‘preprocessing.scale’ function in sklearn and hence the resulted data can be more suitable for algorithms such as Random Forest, Logistic Regression and Naïve Bayes. We introduced artificial noise to existing data and then once again applied all the algorithms mentioned above to evaluate the impact of noise on different ML algorithms.

-----------A graph to show scaled data of all the features ---------

1. **Algorithm & Feature Selection**

The algorithms that we used for carrying out this project are Logistic regression, Random Forest and Naïve Bayes.

Random forest can construct a multitude of decision trees at same time and make decisions by voting. Although sometimes this model maybe a little bias, the variance of its results is very small. The algorithm parameter to be selected is the number of trees. By trying from 1 tree to 100 trees and generating a graph which is shown as below to see how the accuracy value changed with the growing of the number of trees. The plot graph below shows the progression of the value of accuracy (using 10-fold cross-validation) as the number of trees is increased. The value of accuracy stabilized around 0.57 when there are over 60 trees, and the number 80 was chosen to build the model.

-----------A graph to show scaled data of all the features ---------

A Logistic regression model is a representation………………………..

-----------A graph to show scaled data of all the features ---------

A Naïve Bayes model represents………………………..

-----------A graph to show scaled data of all the features ---------

1. **Evaluation**

The dataset was split into train set and test set with a 70:30 split. Since this is to be considered as classification(replace) problem, accuracy/recall/precision is chosen as evaluation metric. The results of these three algorithms before introducing noise are shown below.

-----------A graph/table to show scaled data of all the features ---------

The accuracy values were -----------------------------

Firstly, selecting fewer features which are clearly important in predicting the quality of wine. The result of experiment shows that for random forest, the most important variables are---------. Whereas for NB, important feature ------------------------ much more important variables.

Then by introducing noise in dataset and applying Random Forest, Logistic Regression and Naïve Bayes ML algorithms, the below graphs shows the result. The 1st one shows the progression of the value of accuracy as the number of trees is increased in random forest, while the 2nd one shows how of the value of accuracy changes from ---------- to --------.

-----------A graph/table to show scaled data of all the features ---------

-----------A graph/table to show scaled data of all the features ---------

-----------A graph/table to show scaled data of all the features ---------

After repeating the procedure of training and predicting, the accuracy/recall/precision value with confidence intervals is 0.860.045 for Random Forest and 0.870.023 for NB respectively as shown in the graph below.

-----------A graph/table to show scaled data of all the features ---------

1. **Conclusion**

Before introducing noise, the values of ---------------- accuracy was higher/lower than 0.6. However, after imputing noise, the result for all the algorithms decreased to 0.8. All algorithms’ performances were degraded. Comparing their precision/recall/accuracy values before and after imputing noise, it is easy to tell that performance degrades in the presence of noise in dataset.

As a conclusion,--------------------------------.