Faculty Of Engineering

Cairo University

CMPN451 – Data Mining, Big Data and Analytics

Project Proposal

Names:

Ahmed Mohamed Ismail Nabeel	-	1180501
Moaaz Mohamed Elsherbini	-	1180528
Mostafa Ashraf Ahmed Kamal	-	1180406
Nader Youhanna Adib Khalil	-	1180477

I. Idea:

- Our idea consists of detecting fraudulent credit card transactions. We are given some information about the transaction, such as:

  - distance_from_home - the distance from home where the transaction happened.
  - distance_from_last_transaction - the distance from last transaction happened.
  - ratio_to_median_purchase_price - Ratio of purchased price transaction to median purchase price.
  - repeat_retailer - Is the transaction happened from same retailer.
  - used_chip - Is the transaction through chip (credit card).
  - used_pin_number - Is the transaction happened by using PIN number.
  - online_order - Is the transaction an online order.

  And try to predict whether fraud has happened.

- According to the Data Breach Index, more than 5 million records are being stolen on a daily basis, a concerning statistic that shows - fraud is still very common both for Card-Present and Card-not Present type of payments.

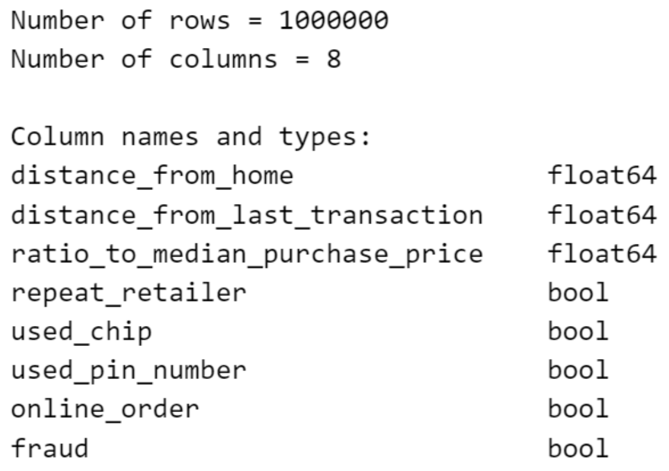
II. Dataset:

We are going to use this dataset: [Credit Card Fraud | Kaggle]

This dataset has 1 Million labeled examples. Each example has 7 features which are a combination of numeric and categorical features.

The label is binary (0: No Fraud, 1: Fraud)

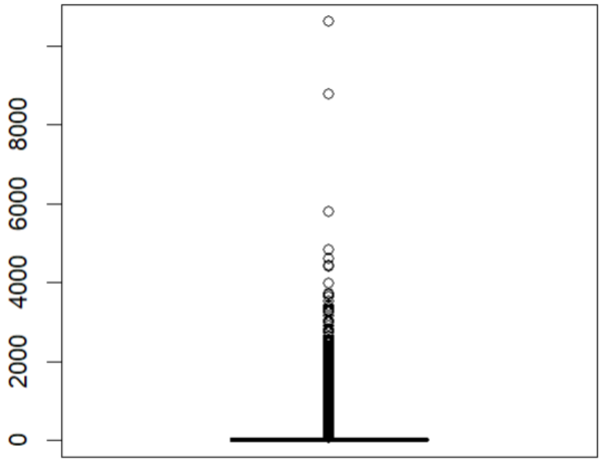
8.74% of the examples are fraudulent transactions, while 91.26% are non-fraudulent.



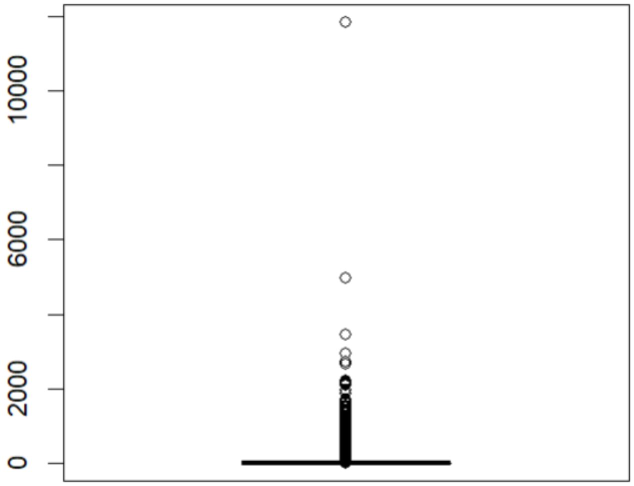
```
Number of rows = 1000000
Number of columns = 8

Column names and types:
distance_from_home                float64
distance_from_last_transaction    float64
ratio_to_median_purchase_price    float64
repeat_retailer                   bool
used_chip                         bool
used_pin_number                   bool
online_order                      bool
fraud                             bool
```

	distance_from_home	distance_from_last_transaction	ratio_to_median_purchase_price
Minimum	0.005	0.000	0.0044
1st Quantile	3.878	0.297	0.4757
Median	9.968	0.999	0.9977
Mean	26.629	5.037	1.8242
3rd Quantile	25.744	3.356	2.0964
Maximum	10632.724	11851.105	267.8029

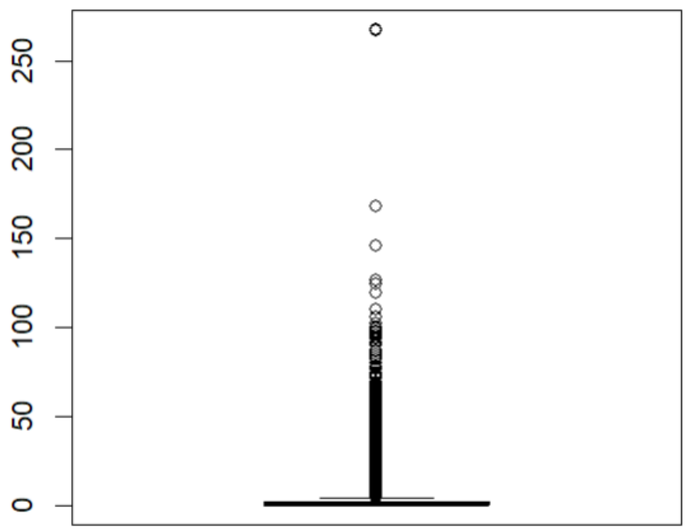
distance_from_home:



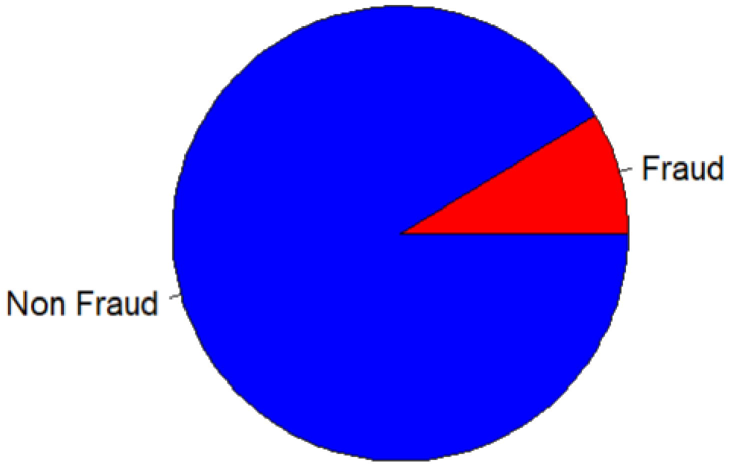
distance_from_last_transaction:



ratio to median purchase price



Fraud vs. Non-Fraud:



We are going to have more insightful visualizations at the end of the project

III. Planned Approach:

We are going to implement supervised as well as unsupervised learning techniques.

Supervised Learning: (To predict the value of the column fraud)

We are going to experiment with different supervised learning techniques and choose what works best, and potentially combine some techniques.

These techniques include:

- Building a classifier such as:
  - Minimum Distance Classifier
  - KNN classifier
  - Naïve Bayes classifier using:
    - Gaussian conditional estimates
    - Parzen density estimates
- Using logistic regression
- Building a Neural Network
- Combining these techniques using algorithms such as AdaBoost

We may also use feature extraction techniques such as:

- Principal Component Analysis (PCA)

Unsupervised Learning: (To generate insights)

We are going to use unsupervised learning techniques such as:

- Clustering
- Association Rules

In order to extract rules such as:

When feature X increases, the probability of a fraudulent transaction increases

When feature Y has a value greater than Z, the transaction is always a fraud

Etc…

We are also going to explore the data well before trying these techniques using visualizations like plots and charts.

MapReduce:

We are going to use one of the big data processing frameworks like Hadoop or Spark integrated with Python.

We are going to use MapReduce in the training phase in some of the following algorithms (algorithm in leture):

- KNN
- Naïve Bayes

- Apriori algorithm
- K-means