**K. J. Somaiya College of Engineering, Mumbai-77**

# REPORT ON
# CREDIT CARD FRAUD DETECTION:
# A CLASSIFIACTION ANALYSIS

Name: Jeel Shah

Roll No. : 1921012

Paper link: **http://ieeexplore.ieee.org/document/8653770**

## INTRODUCTION

Credit card fraud, act committed by any person who, with intent to defraud, uses a credit card that has been revoked, cancelled, reported lost, or stolen to obtain anything of value. Using the credit card number without possession of the actual card is also a form of credit card fraud. Stealing a person’s identity in order to receive a credit card is another more threatening form of credit card fraud, because it works in conjunction with identity theft. Credit card fraud is a problem that affects the entire consumer credit industry. It is one of the fastest-growing types of fraud and one of the most difficult to prevent.

## WHY DO WE NEED TO FIND FRAUD TRANSACTIONS?

For many companies, fraud detection is a big problem because they find these fraudulent activities after they experience high loss. Fraud activities happen in all industries. We can't say only particular companies/industries suffer from these fraudulent activities or transactions.

But when it comes to financial-related companies, this fraud transaction becomes more of an issue/problem. So these companies want to detect fraud transactions before the fraud activities turn into significant damage to their company.

In the current generation, with high-end technology, still, on every 100 credit card transactions, 13% are falling into the fraudulent activities reported by the creditcards website.

Here the point is not only fraud activities increase, but the way of doing scams also increases badly. Companies suffer from detecting fraud, and due to these fraudulent activities, many companies worldwide have lost billions of dollars yearly.

## FRAUD DETECTION APPROACHES

Companies start to detect these fraud activities automatically by using smart technologies.

First, companies hire few people only for the detection of these kinds of activities or transactions. But here they must and should be experts in this field or domain, and also the team should have knowledge of how frauds occur in particular domains. This requires more resources, such as people's effort and time.

Second, companies changed manual processes to rule-based solutions. But this one also fails most of the time to detect frauds.

Because in the real world, the way of doing frauds is changing drastically day by day. These rule-based systems follow some rules and conditions. If a new fraud process is different from others, then these systems fail. It requires adding that new rule to code and execute.

Now companies are trying to adopt Artificial Intelligence or machine learning algorithms to detect frauds. Machine learning algorithms performed very well for this type of problem.

## CREDIT CARD FRAUD DETECTION

The credit card fraud classification problem is used to find fraud transactions or fraudulent activities before they become a major problem to credit card companies.

It uses the combination of fraud and non-fraud transactions from the historical data with different people's credit card transaction data to estimate fraud or non-fraud on credit card transactions.

## DIFFERENT TYPE OF FRAUD TECHNIQUES

Cheats can be extensively requested into three groupings, i.e., standard card related fakes, broker related fakes and Internet fakes. The different sorts of methodologies for giving Visa fakes are depicted underneath.

A. MERCHANT RELATED FRAUDS Seller related traps are started either by proprietors of the shipper foundation or their representatives. The sorts of fakes started by shippers are portrayed underneath:

a. Vendor Collusion: This sort of pressure happens when dealer proprietors, or their agents intend to submit contortion utilizing the cardholder accounts or by utilizing the individual data. They pass on the data about cardholders to fraudsters.

b. Triangulation: Triangulation is a kind of bending which is done and works from a site.. Right when the fraudsters get these unnoticeable parts, they compose stock from a good old fashioned site utilizing stolen charge card unassuming segments. The fraudsters by then by utilizing the Mastercard data buy the things.

B. WEB RELATED FRAUDS

The web is the base for the fraudsters to make the fakes in the fundamentally and the most clear way. Fraudsters have beginning late wore down a to a great degree transnational level. With the extension of trans-edge, cash related and political spaces, the web has changed into an alternate universe's market, getting customers from most nations around the globe. The underneath depicted are most generally utilized methods in Internet compulsion:

a. Site cloning: Site cloning is the place fraudsters close a whole site or basically the pages from which the client made a buy.

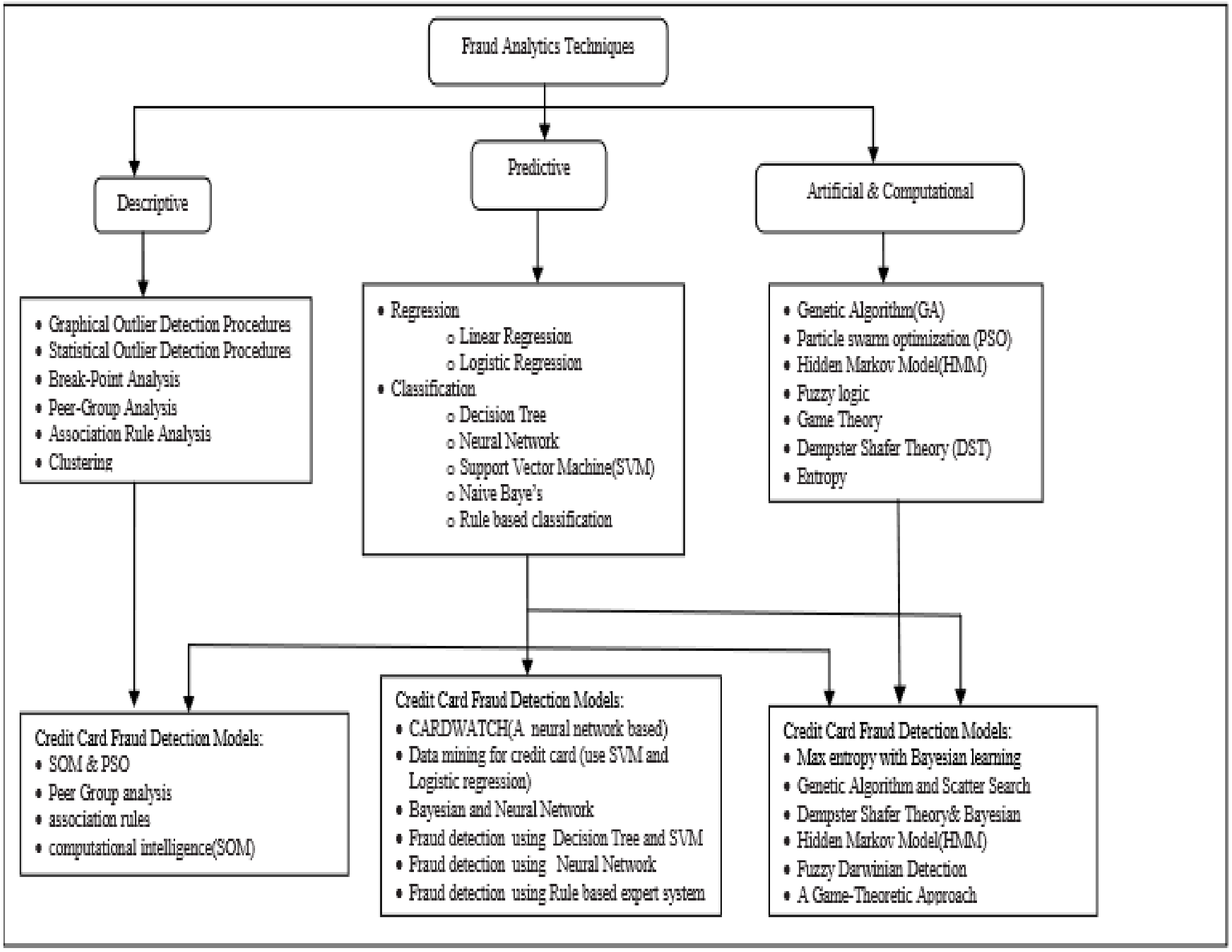
b. False merchant areas: Some objectives routinely offer a broken down association for the clients. That site asks for the client to fill his total motivations behind eagerness, for example, name and pass on to get to the site page where the client gets his required things motivations behind interest.

c. Mastercard generators: These are the PC programs that make liberal Mastercard numbers and expiry dates. These generators

## OTHER FRAUD TECHNIQUES

- Lost/Stolen Cards
- Account Takeover
- Cardholder-Not-Present (CNP)
- Fake and Counterfeit Card
- Erasing the alluring strip
- Making a fraud card
- Skimming
- Phishing

CLASSIFICATION OF FRAUD ANALYTICS TECHNIQUES


Fraud Analytics Techniques
Descriptive
Predictive
Artificial & Computational
• Graphical Outlier Detection Procedures
• Statistical Outlier Detection Procedures
• Break-Point Analysis
• Peer-Group Analysis
• Association Rule Analysis
• Clustering
• Regression
o Linear Regression
o Logistic Regression
• Classification
o Decision Tree
o Neural Network
o Support Vector Machine(SVM)
o Naive Baye's
o Rule based classification
• Genetic Algorithm(GA)
• Particle swarm optimization (PSO)
• Hidden Markov Model(HMM)
• Fuzzy logic
• Game Theory
• Dempster Shafer Theory (DST)
• Entropy
Credit Card Fraud Detection Models:
• SOM & PSO
• Peer Group analysis
• association rules
• computational intelligence(SOM)
Credit Card Fraud Detection Models:
• CARDWATCH(A neural network based)
• Data mining for credit card (use SVM and Logistic regression)
• Bayesian and Neural Network
• Fraud detection using Decision Tree and SVM
• Fraud detection using Neural Network
• Fraud detection using Rule based expert system
Credit Card Fraud Detection Models:
• Max entropy with Bayesian learning
• Genetic Algorithm and Scatter Search
• Dempster Shafer Theory& Bayesian
• Hidden Markov Model(HMM)
• Fuzzy Darwinian Detection
• A Game-Theoretic Approach

DECISION TREE CLASSIFIER

The decision tree is the simplest and most popular classification algorithm. For building the model the decision tree algorithm considers all the provided features of the data and comes up with the important features.

Because of this advantage, the decision tree algorithms also used in identifying the importance of the feature metrics. Which used in handpicking the features.

Once the important features identified then the model trains with the training data to come up with a set of rules. These rules used in predicting future cases or for the test dataset.

It is a type of supervised learning algorithm. The decision tree uses ID3 technique for building decision tree by considering entropy of dataset. The entropy is used to measure the amount of uncertainty in set of data. The splitting criteria in design of decision tree are decided by calculating entropy of each attribute. The entropy of the different state can be calculated by equation as

$$H(p1, p2\ldots\ldots p_s) = \sum_{i=1}^{s}(pilog\left(\frac{1}{pi}\right))$$

Where P1,P2,…Ps are the probabilities of the attributes of dataset. The entropy of each attribute in dataset is calculated and gain is found by subtracting entropy of entire dataset with entropy of splitting attribute.

LOGISTIC REGRESSION

Logistic regression is a type of probabilistic statistical classification model and uses logistic curve for fraud detection. The formula for univariate logistic curve is

$$p = \frac{e^{(c_0+c_1x_1)}}{1+e^{(c_0+c_1x_1)}}$$

The logistic curve gives a value between 0 and 1, so it can be interpreted as the probability of class membership. To perform the regression, the logarithmic function can be applied to logistic function as given shown below.

$$loge\left(\frac{p}{1-p}\right)$$

Here P is the probability of tuple being in class and 1-P is the probability of tuple not in class. However the model chooses values of coefficient C0 and C1 that maximizes the probability of incoming transaction.

KNN

This algorithm uses a labelled training dataset to learn to predict unlabelled data based on certain attributes. The main idea of the KNN algorithm is to predict the class(fraud or non-fraud) of a certain observation(in our case a claim) based on the K nearest neighbours, where K is a certain number that can be set. Nearness is based on a certain distance measure, which evaluates the distance between the attributes of two observation. When the K nearest neighbours are determined, the mode of the labels of those neighbours is predicted as the class for the observation of interest. This process is repeated for each observation

For the K-NN this means intuitively that the chance of a non-fraudulent near-neighbour is higher than that of a fraudulent claim, simply because the data is imbalanced. This could bias the algorithm towards predicting more non-fraud than it

## DATASET INFORMATION

### Context

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.

### Content

The datasets contains 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.

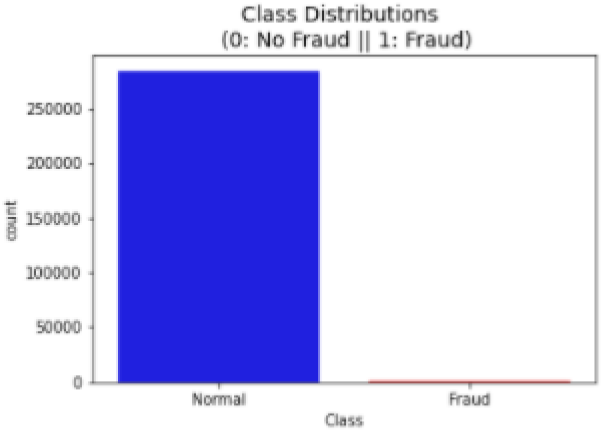
It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Dataset has 284807 rows and 31 features.

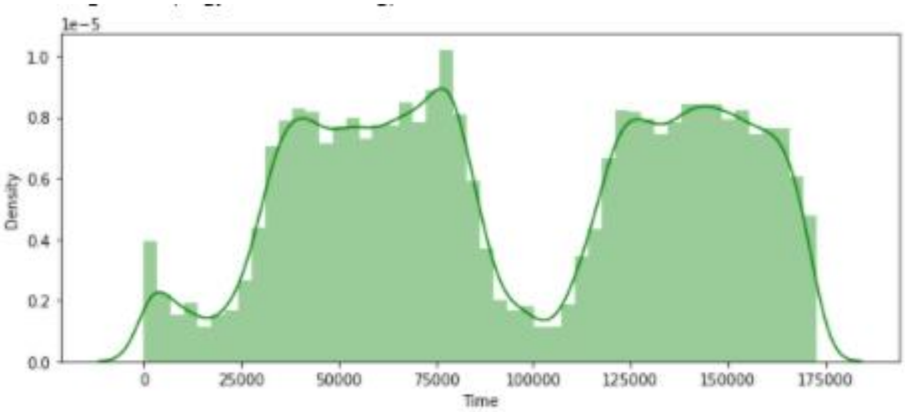
Link to the dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud

## IMPLEMENTATION

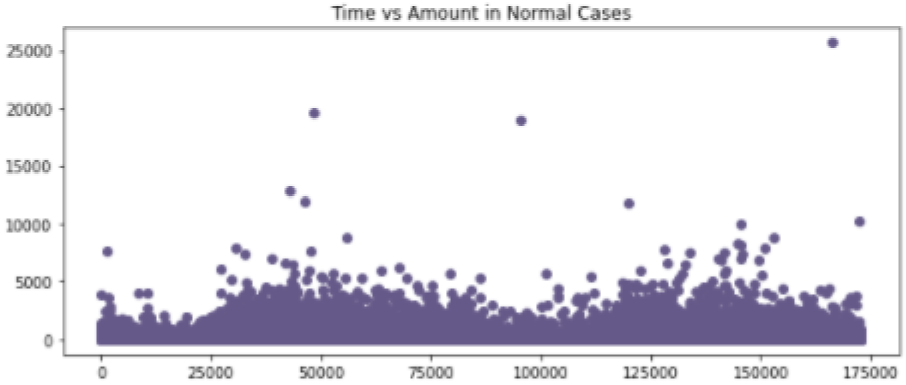
### Visualization

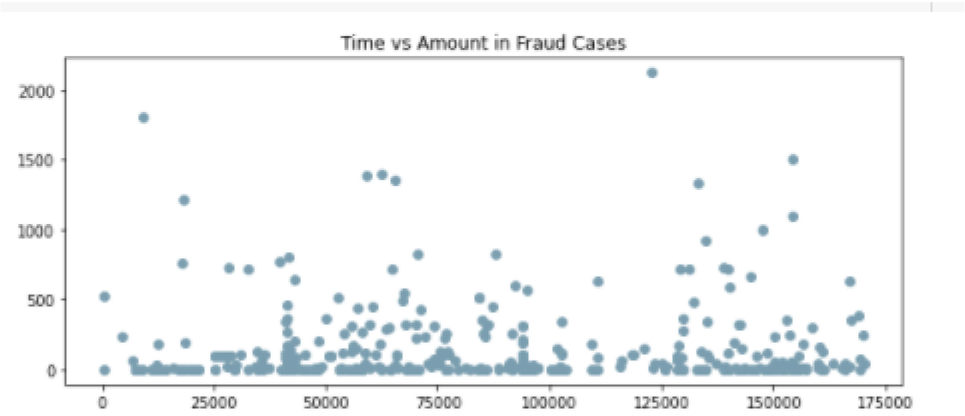


This visualization shows that it is a highly imbalanced dataset

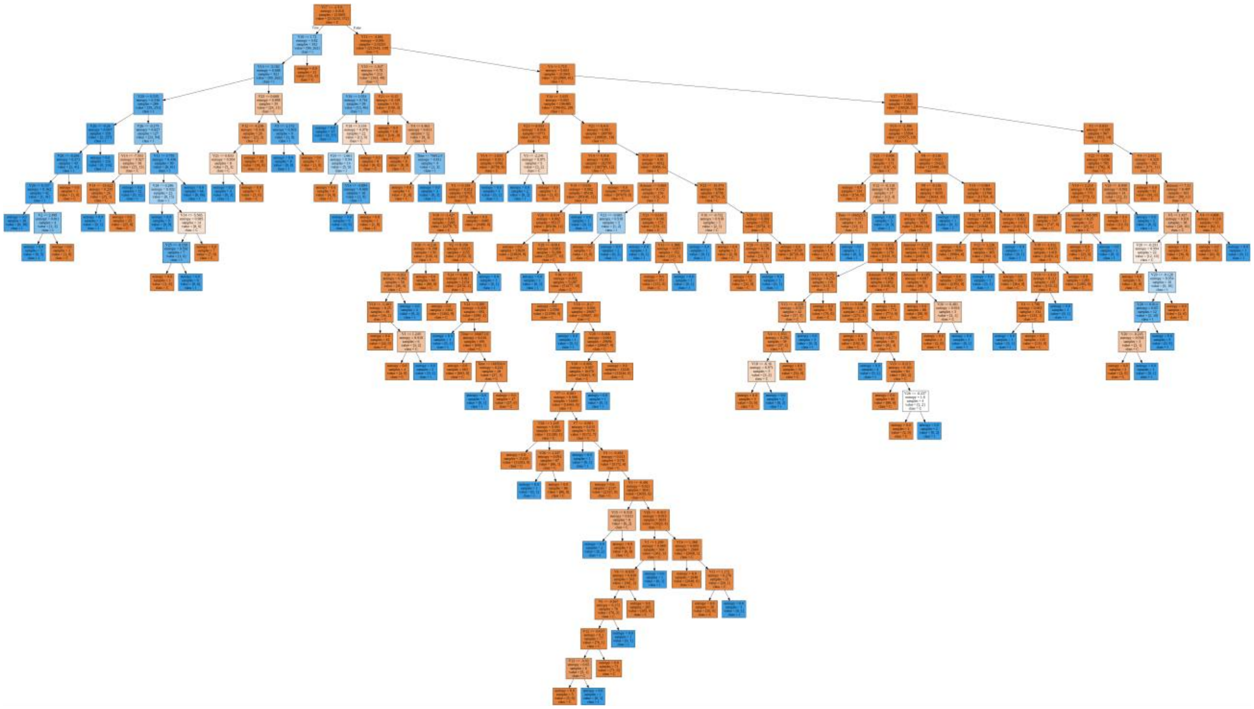


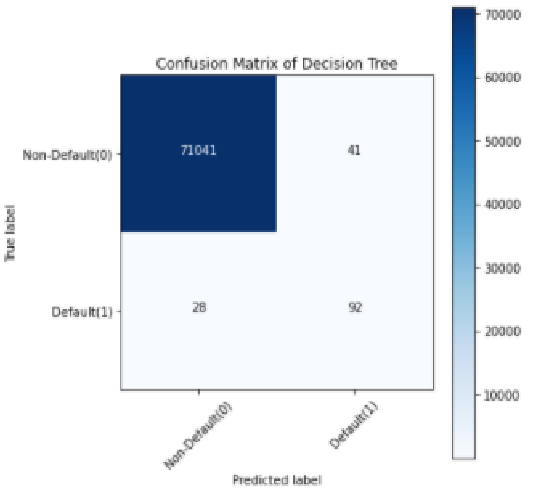
Distribution of Time variable

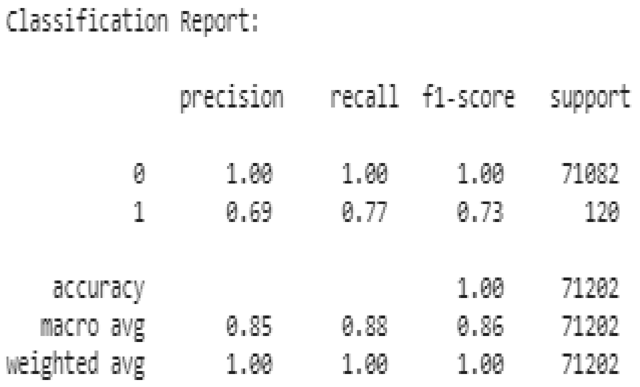




Decision tree Classifier

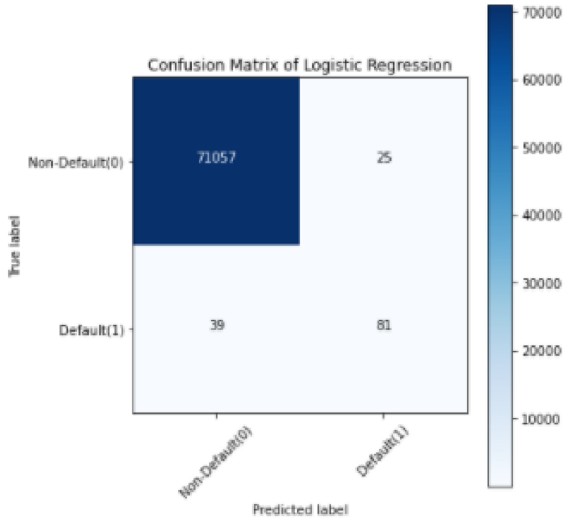


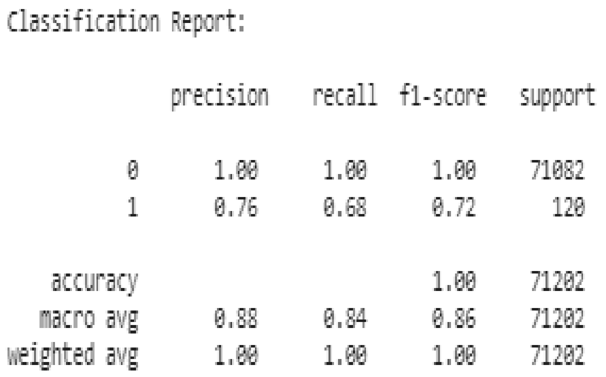




```
Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     71082
           1       0.69      0.77      0.73       120

    accuracy                           1.00     71202
   macro avg       0.85      0.88      0.86     71202
weighted avg       1.00      1.00      1.00     71202
```

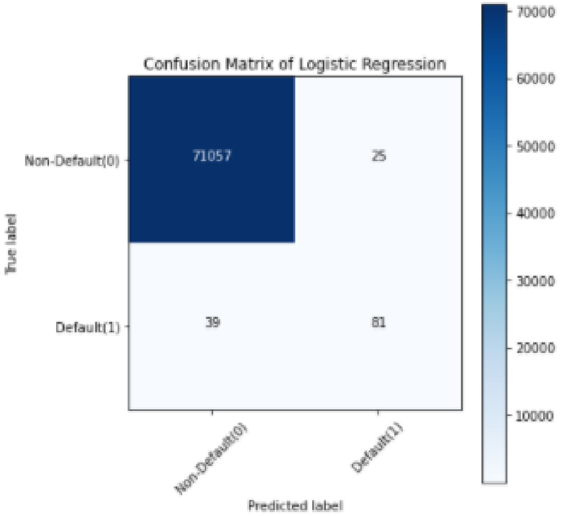
Logistic Regression:

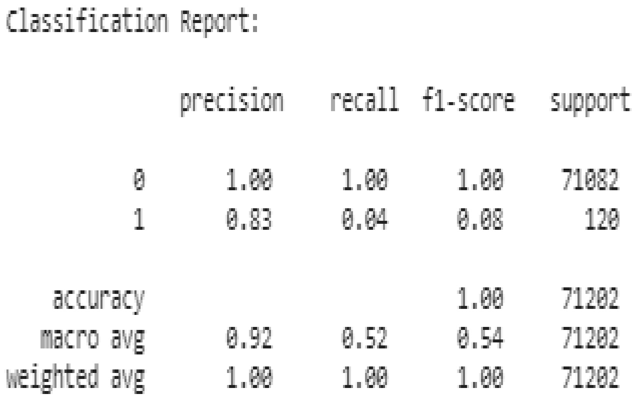




```
Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     71082
           1       0.76      0.68      0.72       120

    accuracy                           1.00     71202
   macro avg       0.88      0.84      0.86     71202
weighted avg       1.00      1.00      1.00     71202
```

KNN:





```
Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     71082
           1       0.83      0.04      0.08       120

    accuracy                           1.00     71202
   macro avg       0.92      0.52      0.54     71202
weighted avg       1.00      1.00      1.00     71202
```

Comparative Results:

| Algorithm | Decision Tree Classifier | Logistic Regression | KNN |
|---|---|---|---|
| Accuracy Score | 99.90% | 99.91% | 99.837% |

## CONCLUSION:

Thus I have implemented Decision Tree Classifier and Logistic Regression out the techniques suggested by the user and these two gave a decent score of accuracy. Since the future scope is not mentioned in the paper, I have Implemented KNN model which gave better accuracy than Decision Tree classifier and Logistic regression. Hence implementation using KNN model can be a part of future scope