SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Permanently Affiliated to the University of Mumbai)

A Project Report on

"CREDIT CARD FRAUD DETECTION"

by

Name	SAP ID
Anuj Gupta	60002180012
Kashish Shah	60002180050

PLOT NO. U-15. JUHU SCHEME, BHAKTIVEDANTA SWAMI MARG, VILE PARLE (WEST), MUMBAI 400 056

Tel: +91 022 4233 5000/42335001 Email: info@djsce.ac.in Website: www.djsce.ac.in

PROBLEM DEFINITION

Nowadays Internet or online transactions are growing as new technology is coming day by day. In these transactions Credit cards hold the maximum share.Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Every year due to fraud Billions of amounts are lost. To analyze the fraud there is a lack of research. There is a rapid increase in the credit card transaction which has led to substantial growth in fraudulent cases. Many data mining and statistical methods are used to detect fraud. Many fraud detection techniques are implemented using artificial intelligence, pattern matching. Detection of fraud using efficient and secure methods are very important. Many machine learning algorithms are implemented to detect real world credit card fraud.

ALGORITHM USED

1)Logistic Regression: Logistic regression works with sigmoid function because the sigmoid function can be used to classify the output of a dependent feature and it uses the probability for classification of the dependent feature. This algorithm works well with less amount of data set because of the use of sigmoid function if the value of the sigmoid function is greater than 0.5 the output will 1 if the output the sigmoid function is less than 0.5 then the output is considered as the 0. But this sigmoid function is not suitable for deep learning because if deep learning when we backtracking from the output to input we have to update the weights to minimize the error in weight update. We have to do differentiation of sigmoid activation function in the middle layer neuron then results in the value of 0.25 this will affect the accuracy of the module in deep learning.

2) Decision Tree: Decision tree can be used for the classification and regression problems working for both is the same but some formulas will change. Classification problems use the entropy and information gain for the building of the decision tree model. Entropy tells about how the data is random and information gain tells about how much information we can get from this

feature. Regression problem uses the gini and gini index for the building of the decision tree model. In classification problems the root node is selected by using information gain that the root node t id selected by using is having the high information again and low entropy. In Regression problems the root node is selected by using gini , the feature which is having the least gini is selected as the root here Depth of the tree can be determined by using hyper parameter optimization, this can be achieved by Using grid search cv algorithm.

3) Random Forest: The random forest randomly selects the features that are independent variables and also randomly selects the rows by row sampling and the number of decision trees can be determined by using hyper parameter optimization. For classification problem statements the output is the maximum occurrence outputs from each decision tree model inside the random forest. This is one the widely used machine learning algorithm in real word scenarios and in deployed models. And in most of the Kaggle computation challenges this algorithm is used to solve the problem statement.

TECHNOLOGY USED

- Python version 3.9
- Jupyter notebook

CONCLUSION

After conducting various methods to predict fraudulent credit card transaction it can be concluded that

- More frauds were committed if-
    - Amount was more
    - Time was less
- The problem to be solved was based on classification, hence the use of logistic regression, random forest and decision tree methods

- Logistic regression yielded an accuracy of 99.92% which decreased(99.89%) when we increased the training dataset for more range of values
- Decision Tree's accuracy was 99.905% which remained the same for additional training dataset
- Random Forest classification gave us the best results with an accuracy of 99.95% which improved further to 99.6% after additional training dataset
- To summarize the results all three algorithms were successful with a accuracy greater than 99%