

**Topic: Credit Card Fraud Detection**

**Name & ID:**

1)Tanjim Ahmed - 19301259

2) Adib Muhammad Amit - 21241062

3) Md. Sifat Kamal - 20101231

4) Syeda Raisa Afsar - 19101007

Section: 04

**Submitting to :**

Nazibur Rahman

Sumaiya Akter

**Date of Submission:**

21/08/2022

**Introduction:** Credit card fraud is a cumbersome and troubling issue that is alarmingly increasing due to the large number of money transactions that take place today. Since it is almost effortless to withdraw money through credit cards, fraudulent transactions can easily be done by escaping a few technicalities. Credit card fraud can take place in many forms and manners such as: phishing through email, getting access to credit card numbers through stealing, rummaging through discarded receipts, scam calls, credit card skimmers placed at the ATM by scammers, etc. To mediate this issue, we have taken a predictive approach using a foolproof dataset containing transactions via a credit card and set some parameters to help the simulation of the experiment i.e. valid, mismatched, missing, mean etc. The main objective of this project is to help in aiding companies as well as their customers to recognize fraudulent actions involving their credit cards by using algorithms to detect credit card fraud. The process and the execution of this project can be witnessed as we traverse throughout this report.

**Methodology:** For this project, we chose a dataset resourced from Kaggle and implemented three algorithms to observe the accuracy, precision, recall, and the F1 score of each implementation and compare which of the three algorithms is the best for this scenario. The three algorithms that we deemed fit for this project are: Random Forest, Logistic Regression, and Decision Tree.

**Dataset Description:** The dataset 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 numeric 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-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification. The correlation between the features has been visualized in the diagram below.

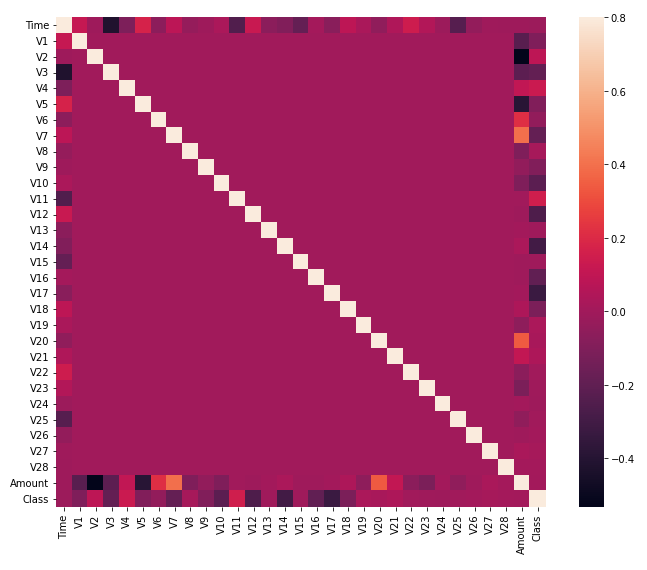


Fig: Correlation Matrix/Heat Map

**Pre Processing:**

Pre-processing in Machine Learning refers to cleaning and organizing the raw data at one’s disposal in order to make it suitable for building and training machine learning models. This can be accomplished by the following steps: 1) acquiring the dataset(first step),2) importing crucial libraries i.e. NumPy, Pandas, Matplotlib and etc., 3) Importing the dataset, 4) Identifying and handling the missing values, 5)Encoding the categorical data, 6) Splitting the dataset(into ratios i.e. training data and testing data such as 70:30 or 80:20) and 7) Feature Scaling(end of preprocessing, method to standardize independent variables of a dataset within a specific range.)

For the first step, we acquired the dataset from Kaggle, a platform consisting of a myriad of datasets for machine learning enthusiasts to utilize and it is also a community for data scientists. Then for the second step, the libraries that we had to import were: NumPy, Pandas, Matplotlib(for generating the accuracy and loss curves and other graphs), and seaborn(to visualize random distribution/statistical data). In the third step, we imported the dataset into google colab, and for the fourth step, we identified and handled the missing values but since it is a pristine dataset, there were no missing values and therefore we did not drop any columns. However, the dataset we acquired is a very unbalanced dataset i.e there are about 284315 legit or normal transactions compared to only 492 fraudulent transactions. We can see that the difference in two classes are very big and hence we had to undersample the number of normal transactions of our dataset in order to balance the dataset and fit our models so that the models could provide more accurate results compared to the results that would be obtained using an unbalanced dataset. During undersampling we have randomly taken 492 legit data and 492 fraud data and merged them together and made another dataset.

For the fifth and the sixth step, we encoded the categorical data and split the dataset into 80:20 ratio. After that we scaled the dataset and started implementing the algorithms to the dataset in order to achieve the result parameters i.e. accuracy, precision etc. The models applied to our dataset have been described beyond this point in this report.

**Models Applied:** The models that have been applied/implemented have been described below:

1. **Random Forest:**

Random forests are a combination machine learning algorithm. Which are combined with a series of tree classifiers, each tree casts a unit vote for the most popular class, then combining these results get the final sort result. Random Forest possesses high classification accuracy, tolerates outliers and noise well and never gets overfitting. Random Forest has been one of the most popular research methods in the data mining area and information to the biological field. Since our dataset contains targeted values as Yes/No i.e 1/0 for fraud transactions and valid or legit transactions respectively, the Random Forest classifier model is one of the best fit for our dataset.

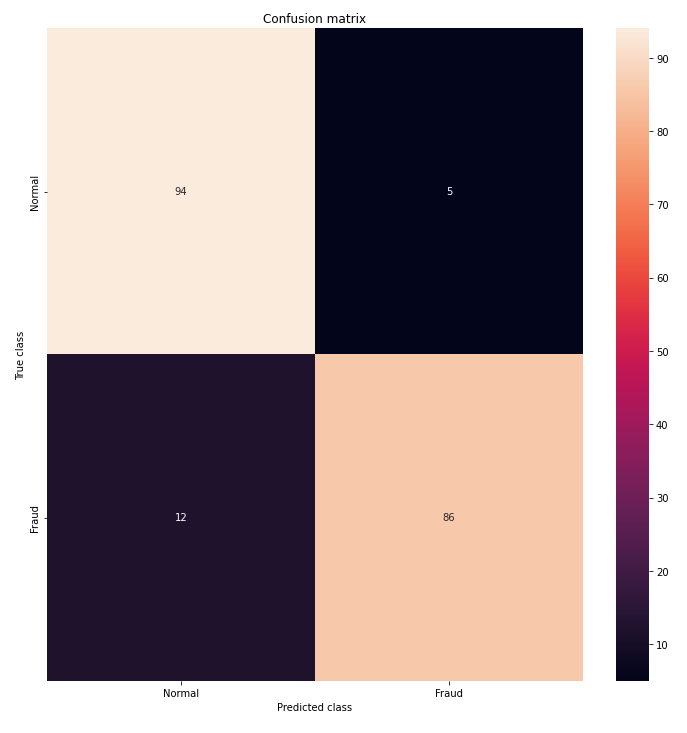


Fig: Random Forest Confusion Matrix

1. **Logistic Regression:**

This algorithm is basically a supervised learning algorithm that is used for the prediction probability of a target variable. In this algorithm, the target variable must be binary. That means here the outcome will be either 1 or 0. If we take a look at this algorithm in the mathematical way here basically the model predicts P(Y=1) as a function of X.

Here in this project, we have to find whether a credit card transaction was fraud or legit. Since here the outcome will be either Yes or No so this is a binary classification problem and for this reason, we can use the logistic regression model.

In this dataset, there is a column named class which is the actual outcome. The outcome is either 0 or 1 where 1 means it's a fraudulent transaction and 0 means it's a legit transaction. So here first of all the dataset is very unbalanced. The legit transaction amount is much higher than the fraud. The amount of legit is 284315 and the fraud amount is only 492 which can be possible that won't provide a good outcome. So for this reason here for balancing the data, we have randomly taken 492 legit data and 492 fraud data and merged them together and made another dataset. Then by splitting the data into X and Y and by training the data, we have got our outcome.

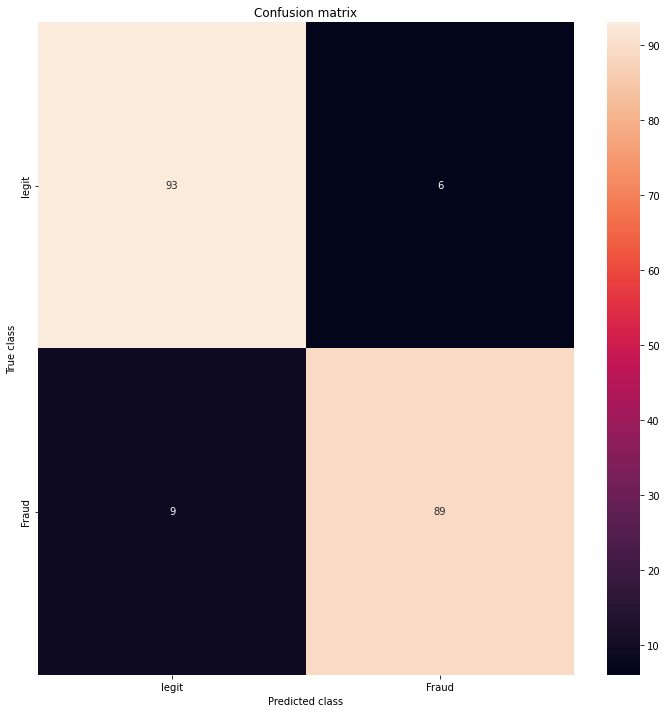


Fig: Logistic Regression Confusion Matrix

1. **Decision Tree:**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees. This algorithm is suitable for our dataset because the outcome of our dataset is also based on Yes/No for fraudulent transactions.

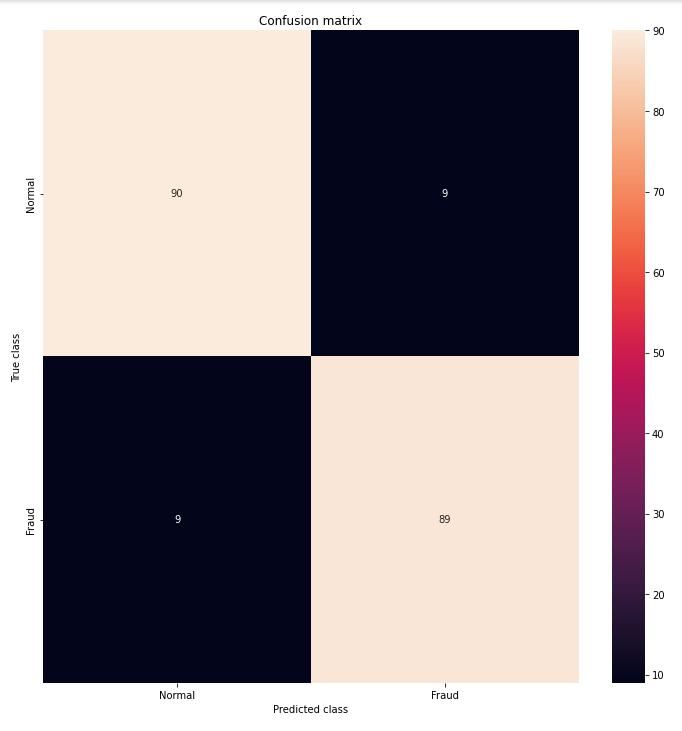


Fig: Decision tree Confusion Matrix

**Results:** As we have previously discussed, we have implemented three models on this dataset and have derived results i.e. accuracy, precision etc for each of the models to compare and contrast amongst the three models. The three models and their results have been given below:

1)Random Forest:

1. Accuracy: 91.37%
2. Precision: 94.50%
3. Recall: 87.75%
4. F-1 Score: 91%
5. Matthews correlation coefficient: 82.94%

2)Logistic Regression

1. Accuracy: 94.92%
2. Precision: 97.83%
3. Recall: 91.84%
4. F-1 Score: 94.74%
5. Matthews correlation coefficient: 90.01%

3)Decision Tree:

a) Accuracy: 90.86%

b) Precision: 90.82%

c) Recall: 90.82%

d) F-1 Score: 90.82%

e) Matthews correlation coefficient: 81.73%

Amongst the three models used, the one with the highest accuracy is Logistic Regression which makes it the best model amongst the three.

References:

1. *Credit Card Fraud Detection*. (2018, March 23). Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
2. *Project 10. Credit Card Fraud Detection using Machine Learning in Python | Machine Learning Projects*. (2021, April 9). [Video]. YouTube. <https://www.youtube.com/watch?v=NCgjcHLFNDg&t=1464s>
3. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
4. Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon
5. Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE
6. Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)
7. Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier
8. Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing
9. Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019
10. Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019
11. Yann-Aël Le Borgne, Gianluca Bontempi Reproducible machine Learning for Credit Card Fraud Detection - Practical Handbook
12. Bertrand Lebichot, Gianmarco Paldino, Wissam Siblini, Liyun He, Frederic Oblé, Gianluca Bontempi Incremental learning strategies for credit cards fraud detection, IInternational Journal of Data Science and Analytics