Financial Transaction Fraud Detection

Abhinav Rawat
IIIT, Delhi
abhinav19132@iiitd.ac.in

Aruj Deshwal
IIIT, Delhi
aruj19024@iiitd.ac.in

Sudeep Reddy
IIIT, Delhi
sudeep19313@iiitd.ac.in

1.Abstract

Fraud Detection is a vital topic that applies to many in-industries including banking, insurance, law enforcement and government agencies. Fraud instances have seen a rise in the past few years so this topic is as critical as ever. Thus we need to be able to distinguish between authentic and fraudulent financial transactions. As the world moves towards digitization more transactions become cashless. The use of credit cards and online payment methods have increased. Increase in fraud rates in these kinds of transactions causes huge losses for financial institutions and users. Thus we will do a comprehensive review of the various methods to detect fraud.
Github Repository: link

2. Introduction: Problem statement

The objective of our project is to identify fraudulent transactions from a skewed dataset. We aim to find optimal algorithms to recognize such instances of fraud to better combat this problem. We would be trying different algorithms such as SVMs, Random Forests, Neural Networks, etc. and find the optimal hyperparameters such as number of epochs and learning rate. The dataset contains skewed data which we are planning to counter using techniques such as under sampling and oversampling.

3. Literature Survey

Credit card Fraud Detection Using Machine Learning and Data Science

Maniraj et al. [1] predict fraudulent activity in credit card transactions using Local Outlier Factor and Isolation Forest Algorithm. The local outlier factor finds the anomaly score of each sample and measures the local deviation from its neighbors. The isolation forest algorithm arbitrarily selects a feature and then randomly selects a split value between maximum and minimum values. Recursive partitioning eventually creates a tree. The paper uses a skewed dataset from kaggle, the values of the data have been put through PCA to protect sensitive information. They reach an accuracy of 99.6% and a precision of 33%. These results are attributed to a large imbalance in genuine and fraudulent transactions.

A Predicting Model For Accounting Fraud Based On Ensemble Learning

Y. Sun et al [2] predict fraud using XGBoosting which is a gradient boosting algorithm for decision trees. It combines many low-accuracy trees into a high one. The model is made up of decision trees with different weights. Each tree makes predictions by selecting split points for the feature data. The greater the change in prediction effect caused by a new split of the feature, the more important that feature is. It compares XGBoosting to Logistic Regression and AdaBoosting. The results show that XGBoosting beats both the models and performs better with more data.

4. Dataset

We used a synthetic Credit Card Transaction Dataset from kaggle [link]. Most credit card transaction data contains privileged information and having PCA done on the columns and feature analysis is not possible. The data contains 24,000,000 transactions for 2,000 synthetic consumers in the US. The data also covers gender, debt, income and Fico Score data. Analysis on the data shows that it is a reasonable match for real data in terms of fraud rates, purchase amounts etc. Out of the transactions only 30,000 are fraudulent in nature. Thus it is highly skewed in nature and the authentic transactions are downsampled to 270,000 to help increase precision and f1 scores.

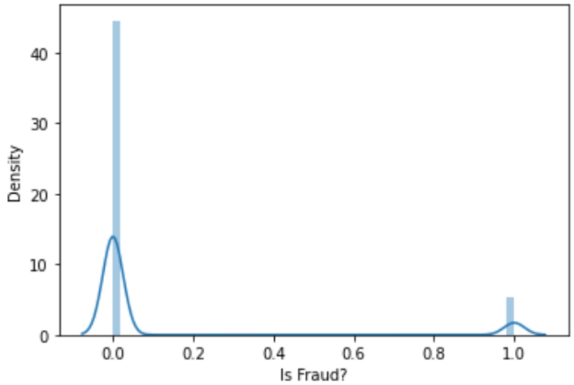


Fig 1 - Distribution of labels in dataset

The dataset consisted of 3 csv files containing transaction, user and card data. These were merged using customer id and card index values as keys. The dataset is composed of attributes such as user, card, amount, transaction error, card type, age, gender, yearly income, fico score etc. Attributes such as year, month, state, zip code, card cvv, number of cards issued, expiry date, card number, latitude and longitude of users, name etc. were dropped as they have low correlation with the nature of transaction. All the categorical variables were encoded to suit the model. All string objects were mapped to integer values. After merging the dataset was shuffled. We plotted the distributions and box plots of the features.

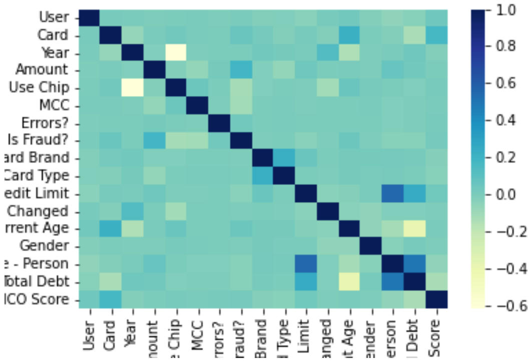


Fig 2- Correlation heatmap of features

The data was split 8:2 for test and train. Copies of the dataset were made and min-max scaling, standard scaling and robust scaling pre-processing techniques were used to determine optimal pre-processing methods.

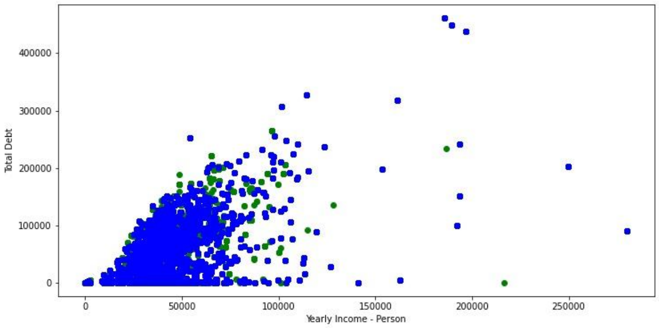


Fig 3- Scatter plot of Debt/Yearly-Income (Green=Not fraud, Blue = fraud)

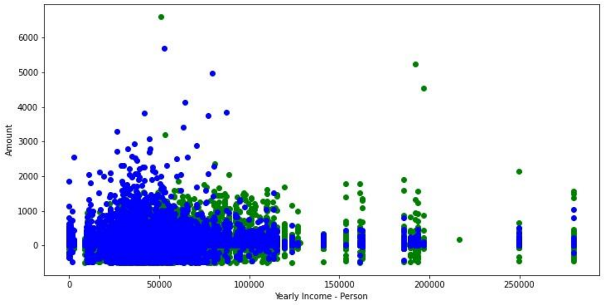


Fig 4- Scatter plot of Amount/Income(Green=Not fraud, Blue = fraud)

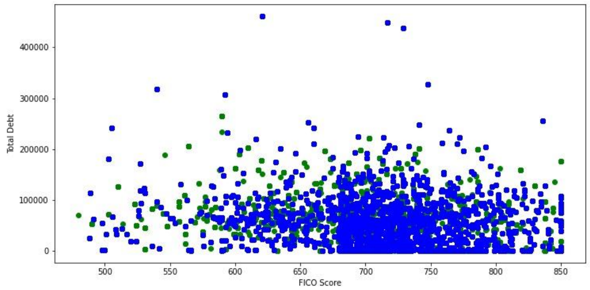


Fig 5- Scatter plot of Debt/Fico score (Green=Not fraud, Blue = fraud)

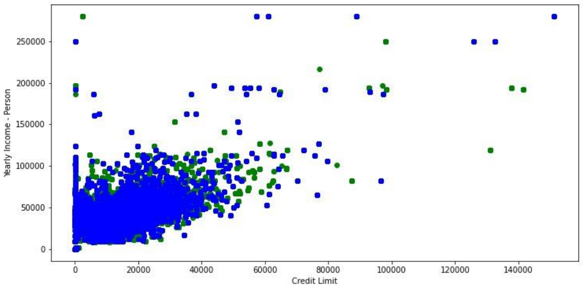


Fig 6- Scatter plot of Yearly-Income/Credit Limit (Green=Not fraud, Blue = fraud)

5. Methodology

Logistic Regression – Logistic regression was used over different sets of pre processed data and metrics such as confusion matrix, f1_score, precision and recall were recorded. Precision-recall vs thresholds were plotted to find the optimal value at which both precision and recall were high. Implemented Grid search to find the best permutation of parameters which would give the maximum precision-recall and accuracy. Parameters over which grid search was implemented include different solvers, l2 penalty and different values of regularisation strengths

Naive Bayes – The naive bayes classifier was used against different sets of pre processed data. We then used different metrics to determine how different preprocessing steps performed.

Decision Trees – We used a Decision Tree classifier on the same dataset with 3 different types of preprocessing. Then compared it with different metrics such as accuracy, precision, recall and f1 scores.

Random Forest Classifier – Random Forest builds multiple decision trees and merges for a more accurate and stable prediction. This allows it to correct the overfitting problem of decision trees. Number of trees taken in the forest is 100.

Support Vector Machine - SVM was used after doing PCA of the dataset. The number of components required were selected by taking the components comprising 95% variance of the data, which came out to be 10 components from 16 . The regularization parameter was set to 0.1,1,10 and gamma was set to 0.1, 0.01. The kernel was set to "rbf".

Neural Network - A Neural Network was used with hidden layer sizes 9, 8, 6 ,3 which proved to give best results with learning rate 1e-4.

6. Results and analysis

1) Logistic Regression -

Training Accuracy-0.894
Testing Accuracy-0.891
Logistic Regression performs best when the preprocessing is robust scaling
The optimal threshold for the model was approximately 0.15 where the precision and recall value is approximately 0.3. Standard model gave best results as compared to results obtained from models part of grid search.

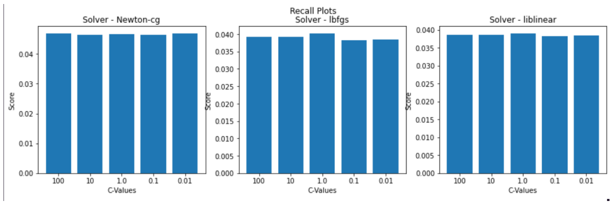


Fig 7- Recall values for 3 distinct solvers and 5 distinct regularisation strengths

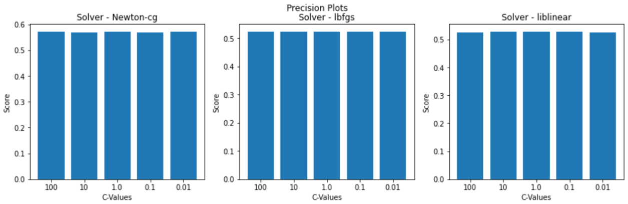


Fig 8- Precision values for 3 distinct solvers and 5 distinct regularisation strengths

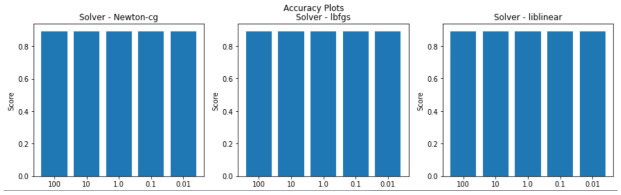


Fig 9- Accuracy values for 3 distinct solvers and 5 distinct regularisation strengths

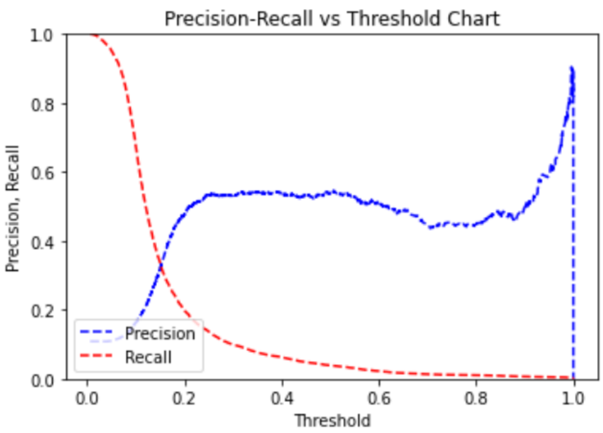


Fig 10- Precision-Recall vs Threshold for Logistic Regression

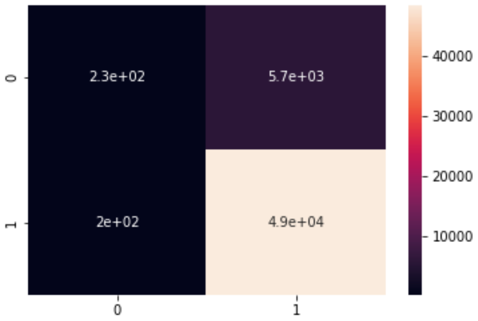


Fig 11- Logistic Regression Confusion Matrix

2) Naive Bayes -
Training Accuracy - 0.881
Testing Accuracy - 0.883
Naive bayes performs best when the preprocessing is robust scaling
The optimal threshold for the model was approximately 0.82 where the precision and recall value is approximately 0.2

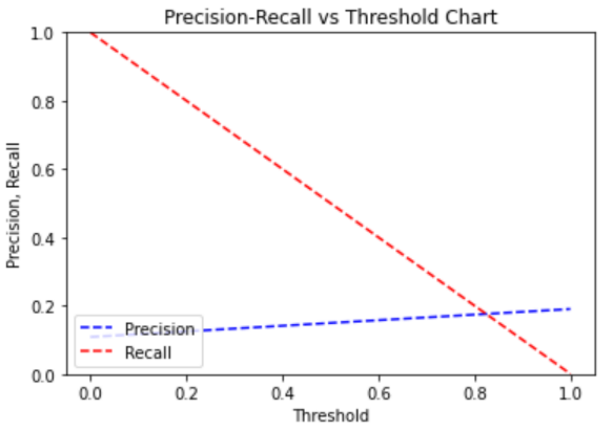


Fig 12- Precision-Recall vs Threshold for Naive Bayes

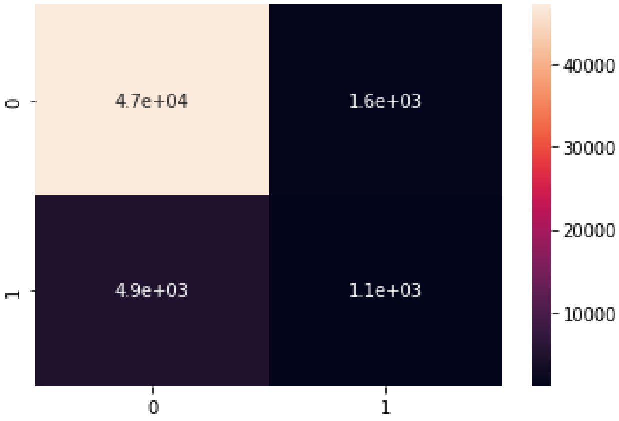


Fig 13- Naive Bayes Confusion Matrix

3) Decision trees -
Training Accuracy - 0.999
Testing Accuracy - 0.941
Decision Tree performs best when the preprocessing is robust scaling. It has a max depth of 38.

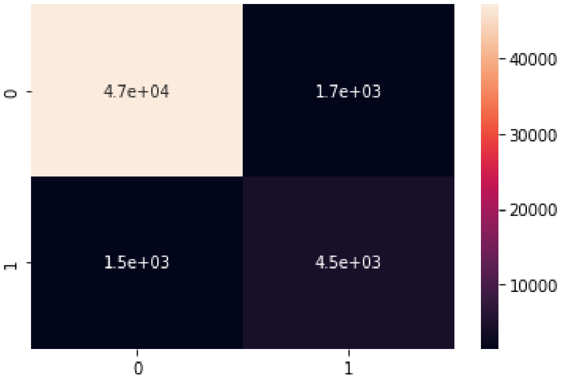


Fig 14- Decision Tree Confusion Matrix

4) Random Forest-
Training Accuracy-0.999

Testing Accuracy-0.961
Random Forest performs best with raw data.

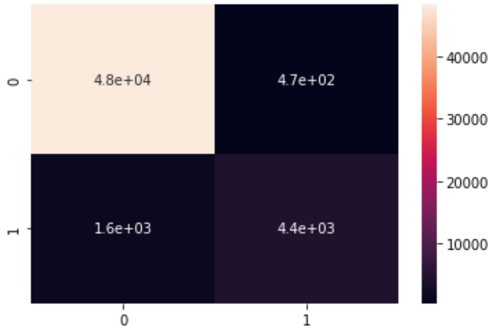


Fig 15- Random Forest Confusion Matrix

5) Support Vector Machine -
Training Accuracy - 0.9018
Testing Accuracy - 0.9022
Precision Score For training - 0.676
Precision Score For testing - 0.660
Recall for testing- 0.1914
F1 for testing -0.296
SVM performed best with regularization set to 1 and gamma set to 0.1.

Graph for choosing no of components for PCA

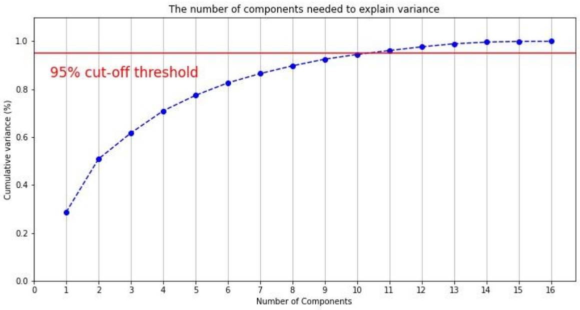


Fig 16- Support Vector Machine Threshold Plot

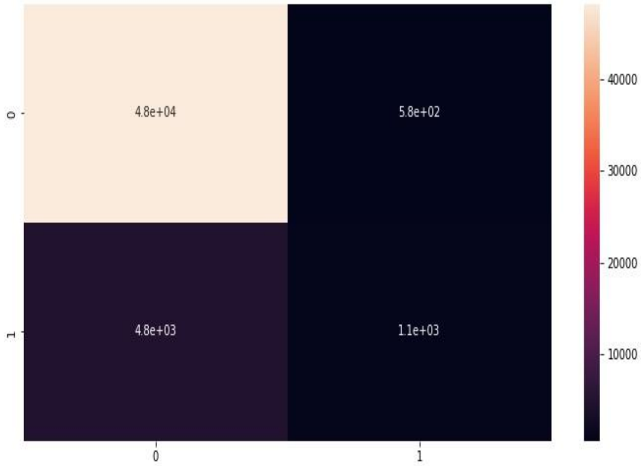
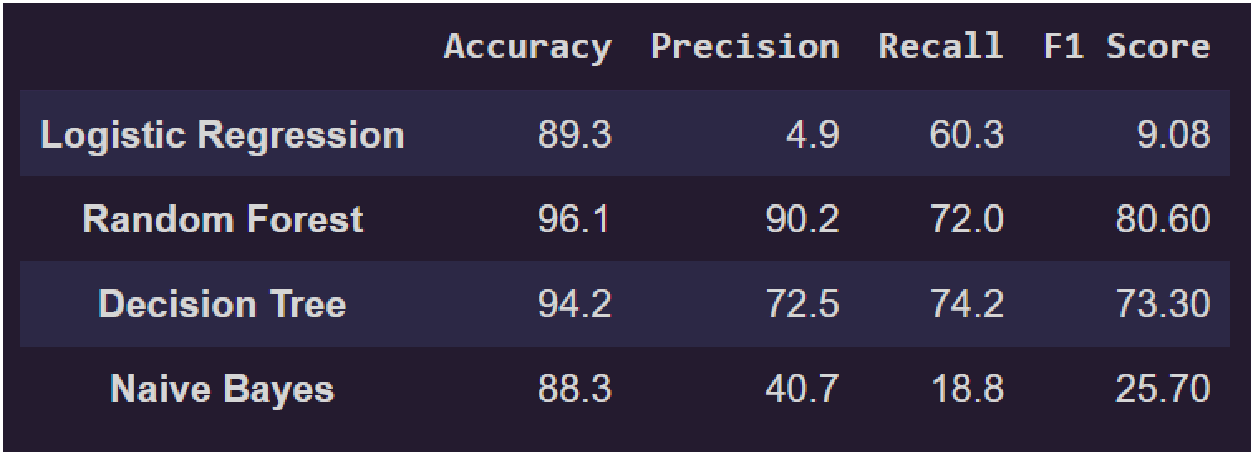


Fig 17- Support Vector Machine Confusion Matrix

6) Neural Network-
Training Accuracy-0.99
Testing Accuracy-0.89
Neural Network performs best with raw data.



	Accuracy	Precision	Recall	F1 Score
Logistic Regression	89.3	4.9	60.3	9.08
Random Forest	96.1	90.2	72.0	80.60
Decision Tree	94.2	72.5	74.2	73.30
Naive Bayes	88.3	40.7	18.8	25.70

Fig 18- Comparison of different model

7. Conclusion

In the majority of the models the best results were seen with robust scaling of the data.

With all the tested models the best results were seen with Random Forest Classifiers with an accuracy of 96.1% and with high precision, recall and f1 score of 90.2, 72.0 and 80.6 respectively. It was seen that oversampling fraud data and undersampling non-fraudulent data allowed for the models to train better and have better f1 scores and were robust enough to detect outliers.

In the majority of the models the best results were seen with robust scaling of the data.

Member contribution

Aruj Deshwal-Random Forest, Pre-Processing, Literature Review, finding Dataset, Hyperparameter Tuning, Neural Network

Abhinav Rawat- Pre-Processing, Logistic Regression, Data Cleaning, Feature analysis ,Finding Dataset, Hyperparameter Tuning

Sudeep Reddy- Pre-processing , Literature Review, Decision Tree, Naive Bayes, Support Vector Machine.

8.References

[1] SP Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed, Credit card Fraud Detection Using Machine Learning and Data Science , *IJITEE* 2019

[2] Y. Sun, Z. Ma, X. Zeng and Y. Guo, "A Predicting Model For Accounting Fraud Based On Ensemble Learning," 2021 IEEE 19th International Conference on Industrial Informatics

[3] Credit Card Transaction Dataset, *Kaggle*