

Credit Card Fraud Detection

Name:	Anuradha Mahato
Registration No./Roll No.:	19347
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	February 02, 2022
Date of Submission:	April 24,2022

1 Introduction

Problem Overview- The data set contains credit card transactions made by European cardholders. The data set is modified with Principal Component Analysis (PCA) to maintain confidentiality. The features include 'time', 'amount', and the other ones (V1, V2, V3, up to V28) are the principal components obtained using PCA. The number of fraudulent cases is very high compared to the non-fraudulent. 5.

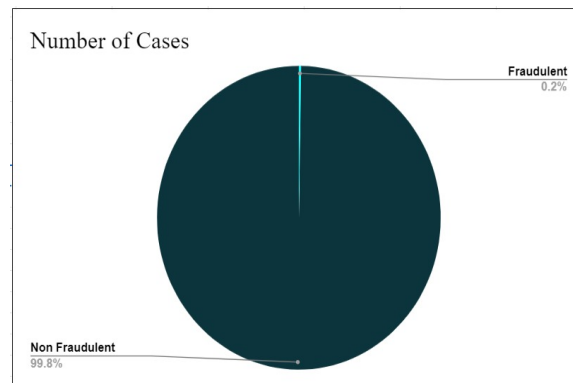


Figure 1: Fraud vs Genuine

Project Plan-

- **EDA on Data:** Data Analysis and visualization. The size of the train data is 57116x30. Here 28 columns are PCA transformed. The class labels is of size(57116,1) and size of the test data is (14280, 30).
No null values are present in the data. The fraud and valid transactions are: Fraud Cases: 142
Valid Transactions: 56974
- **Pre-processing of the data:** Handling missing values, scaling the data using Robust scaler for time and amount as all the other features are PCA transformed.
- **Train Test Split:** Splitting data into train and test data.
- **Sampling the data :** Performing undersampling, over sampling, SMOTE as the data is highly unbalanced. While building models we have performed them - without handling data imbalance, over sample, undersampling.
- **Building and evaluating the models:** Implementing supervised classification algorithm (Random Forest, KNN, SVM, Logistic Regression, Decision Trees, Linear SVC), did parameter tuning for them to improve the F1 score accuracy using various evaluation metrics.

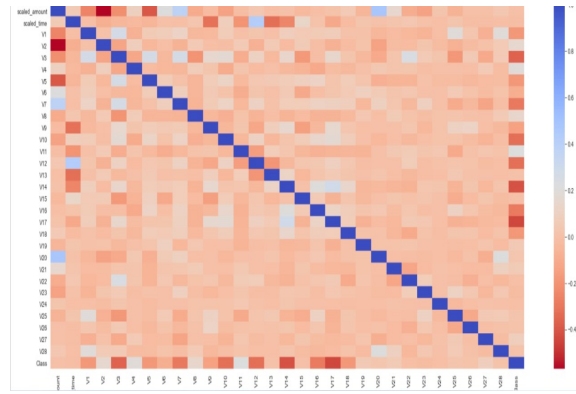


Figure 2: Correlation Matrix of the train data

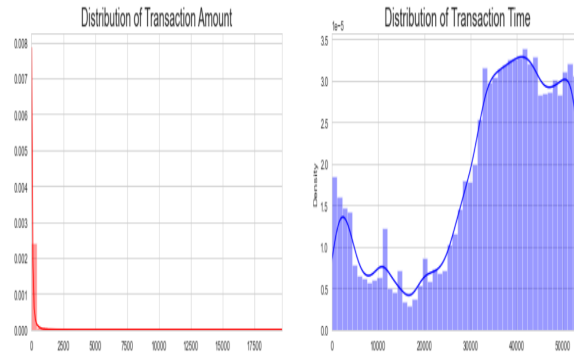


Figure 3: Understanding transaction amount, transaction time

2 Methods

We have done the model training for the data- 'Without sampling', 'With sampling - over and under'. Here we take the best classifiers by considering their evaluation metrics.

We experimented with **Logistic Regression**, **KNeighboursClassifier**, **Support Vector Classifier**, **Decision Tree Classifier**, **Random Forest Classifier**, **Gaussian Naive Bayes** models by comparing them with various evaluation metrics before doing sampling and after it. Then hyper parameter tuning was done by the classification pipeline discussed in class for the above models.

1. **Random Forest** - Random forests is an ensemble learning method that is used for classification here. It works by constructing multiple decision trees at training time. The final classification is done by storing that output of the random forest which is selected by most trees. [1]

2. **KNN** - It is a type of classification where the function is only approximated locally. It is done by taking a majority vote of its k nearest neighbours. K nearest neighbours

3. **Decision trees** - It predicts the fraud/non-fraud cases by learning simple decision rules that are inferred from the data features.

4. **Logistic Regression** - It is used for predicting the categorical dependent variable using a given set of independent variables

5. **Support Vector Classifier** - It is a non parametric clustering algorithm and does not make any assumption on the number or shape of the clusters in the data. [2, 3]

6. **Gaussian Naive Bayes** - Based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

As the data is highly imbalanced, we first tried scaling data by doing random undersampling, random oversampling and SMOTE (Synthetic Minority Oversampling Technique). oversampling using SMOTE

Stratified sampling

Table 1: Performance Of Different Classifiers Using All Terms

Classifier	Precision	Recall	F-measure
DT	0.79	0.91	0.84
SVM	0.60	0.71	0.63
RF	0.90	0.95	0.92
KNN	0.94	0.91	0.93
LR	0.55	0.94	0.59
SVC	0.69	0.93	0.77

The proposed models are Random Forest, KNN, and Decision Trees

We have tried various parameters for the models using GridSearchCV

GridSearchCV : It helps to loop through predefined hyperparameters and fit our estimator (model) on our training set. So we can select the best parameters from the listed hyperparameters. GridSearchCV

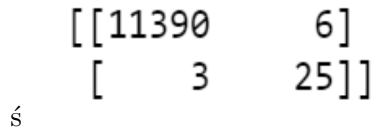
We have performed **select k- best** features for taking the k best features out of the total features we had. Here the chi2 method showed error as the data was having negative values and so we performed the select k best mutual info classif method. ¹ used to implement the classifiers .

Github link-Github link

https://github.com/Anu-14/Credit_Card_Fraud_Detection/blob/main/Anuradha_Mahato_Final_project_code.ipynb

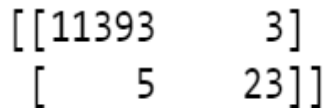
3 Evaluation Criteria

As the data is highly imbalanced, accuracy is not enough for evaluation. Precision and Recall are equally essential as low numbers of false positives and false negatives are essential here. Thus, F1 score will be a good measure here. We computed the classification report for each of the classifier and compared the F1 scores for each of them.



$$\begin{bmatrix} 11390 & 6 \\ 3 & 25 \end{bmatrix}$$

Figure 4: Confusion matrix for Random Forest



$$\begin{bmatrix} 11393 & 3 \\ 5 & 23 \end{bmatrix}$$

Figure 5: Confusion matrix for KNN

¹<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC>

4 Analysis of Results

It was found that under-sampling reduced the F1 score while by Random oversampling on the data set, the F1 score increased. From the above classifiers we proposed in this project, we observed that the F1 scores for random forest and KNN are 0.85 after hyper parameter tuning[4] and feature extraction.

5 Discussions and Conclusion

A significant finding was that Random Forest and KNN were the best possible classifier models for our data set from among the classifiers implemented, which included Logistic Regression, KNeighbors Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and Gaussian Naive Bayes. We can further implement the AdaBoost classifier as we could not perform it due to time constraints and computational complexity.

References

- [1] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. Random forest for credit card fraud detection. In *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, pages 1–6. IEEE, 2018.
- [2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
- [3] I. H. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, third edition, 2011.
- [4] Naoufal Rtayli and Nourddine Enneya. Enhanced credit card fraud detection based on svm-recursive feature elimination and hyper-parameters optimization. *Journal of Information Security and Applications*, 55:102596, 2020.