Table of Contents

Extract Transform Load (ETL) 2

Relevant Analytical Models 6

Selected Analytical Models 6

Decision Tree Classifier 6

AdaBoost Classifier 7

Critically Evaluation of Analytical Models 7

Decision Tree Model Evaluation 7

AdaBoost Model Evaluation 7

Logistic Regression Evaluation 8

Critical Analysis of chosen model 8

Correlation Matrix 9

Confusion Matrix 10

Confusion Matrix of AdaBoost Model 11

Recommendation 11

References 13

Extract Transform Load (ETL)

ELT is the first process while working on the credit card dataset in the data analytics framework (Wang, et al., 2020). For analysis of the credit card data, we have to need to preprocess the data while using pre-processing libraries in python language. For this purpose, first loaded the dataset in the implementation environment and then performed the extraction & transformation process (Ge, et al., 2018). Provided credit card data contains the 100000 observations and 15 variables in which 14 variables are independent variables and fraud is the dependent or target variable. This project is related to fraud prediction so there is a need to implement the classification model. But before this, required to extract & transform the data that will help in exploratory data analysis (EDA) and analytical model implementation (Sahoo, et al., 2019).

Extract: Data already extracted from a data source or database and provided in CSV format that needs to be preprocessed, transform, and load in the development environment.

Transform: The raw data contains the character, string, object, and int datatypes. Amount variable is also given in object format that needs to be transformed into numeric data. All the variables expect "fraud & age" require transforming into numeric data and for this purpose, I have used the "Label encoding" technique in data preprocessing. The analytical model requires numeric data as an input so transformed the object data types variables into numeric (integer).

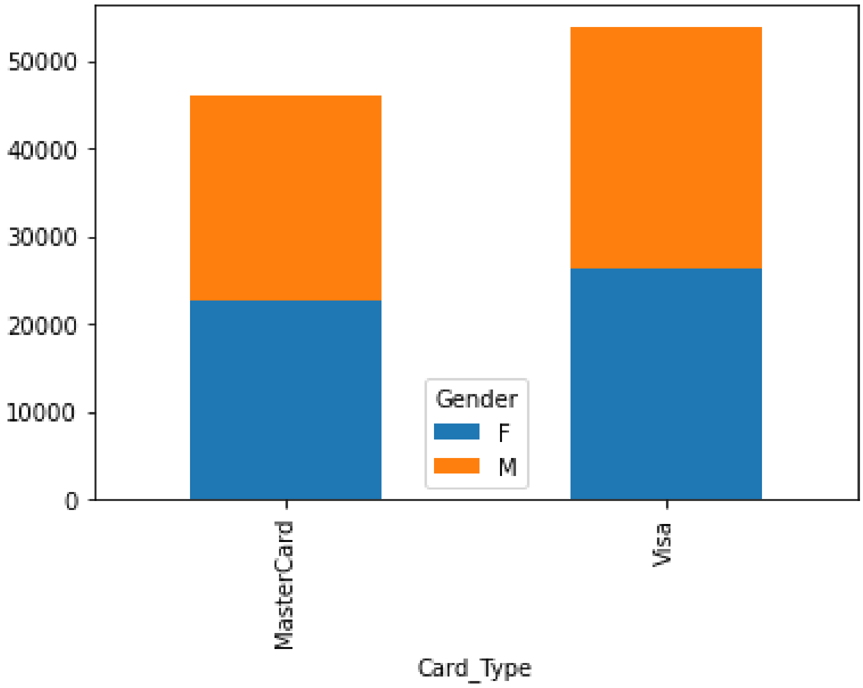
Load: After preprocessing load the data for performing exploratory data analysis and analytical model implementation. Data load into two data frames in which the first data frame contains the original data and in the second data frame data contains the transformed data.

From the provided data I have found some "ambiguities, assumptions, and anomalies" but there are not any missing values in the data. The first three variables are not related to the data analysis, so I have removed these anomalies from the data.

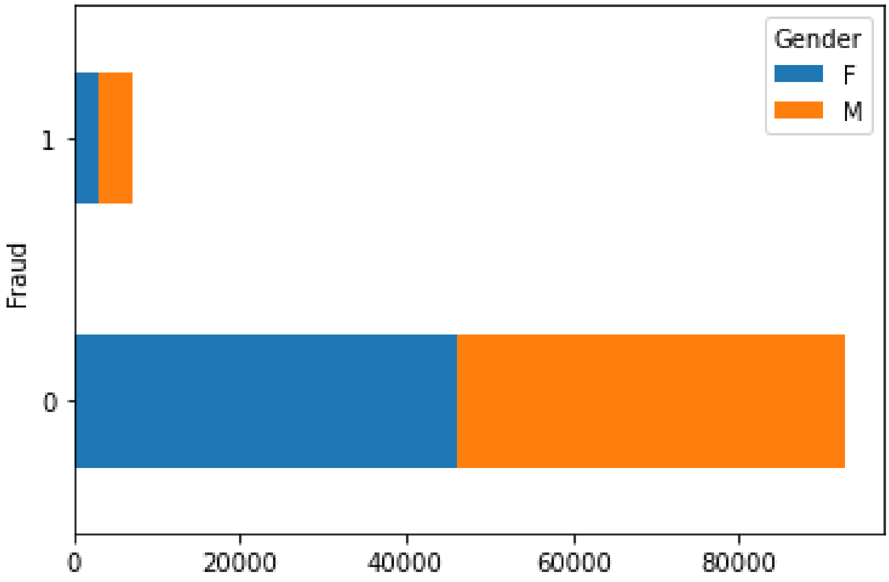
Exploratory Data Analysis (EDA)

Exploratory data analysis is a visualization process that helps to understand & explore the important aspects of the dataset. (Sahoo, et al., 2019) in this project, I have explored credit card fraud data that helped me to find the important information. Original data is used to perform the exploratory data analysis (EDA) because with transformed data we can’t get the actual understanding & exploration from the data.

Card Usage & Fraud by Gender

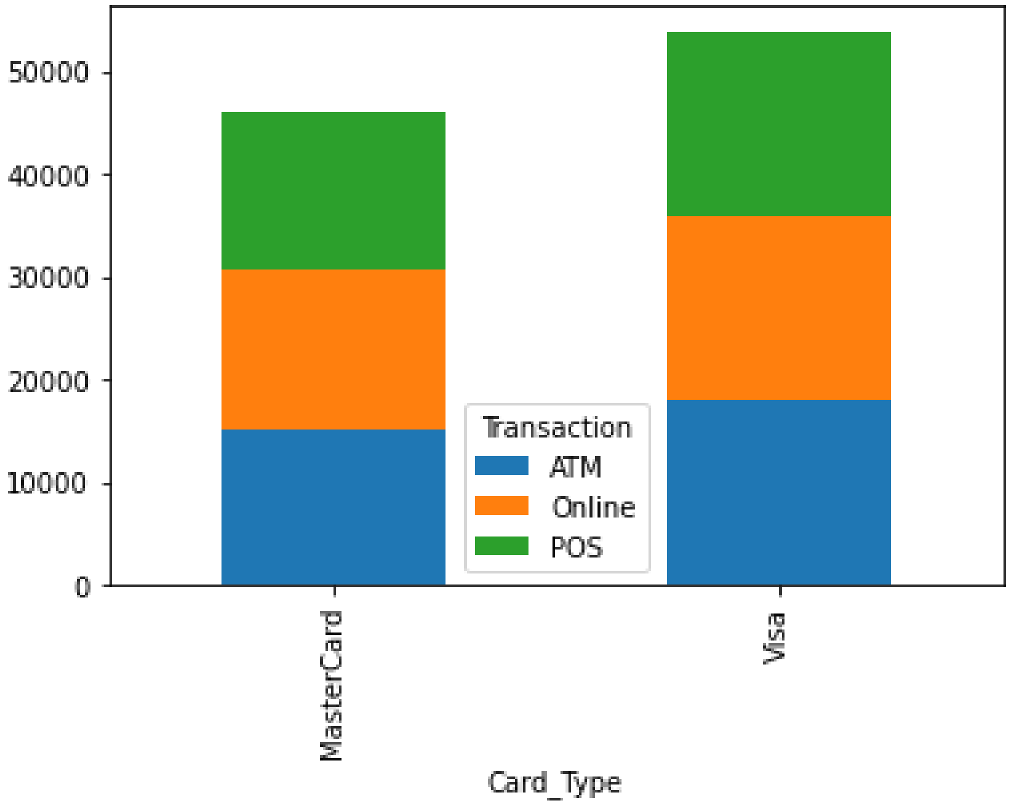


The visa credit card has the highest transactions in the data regarding gender which mean male & females mostly used the Visa credit card. Male customers have the highest Visa credit card accounts as compared to the female.

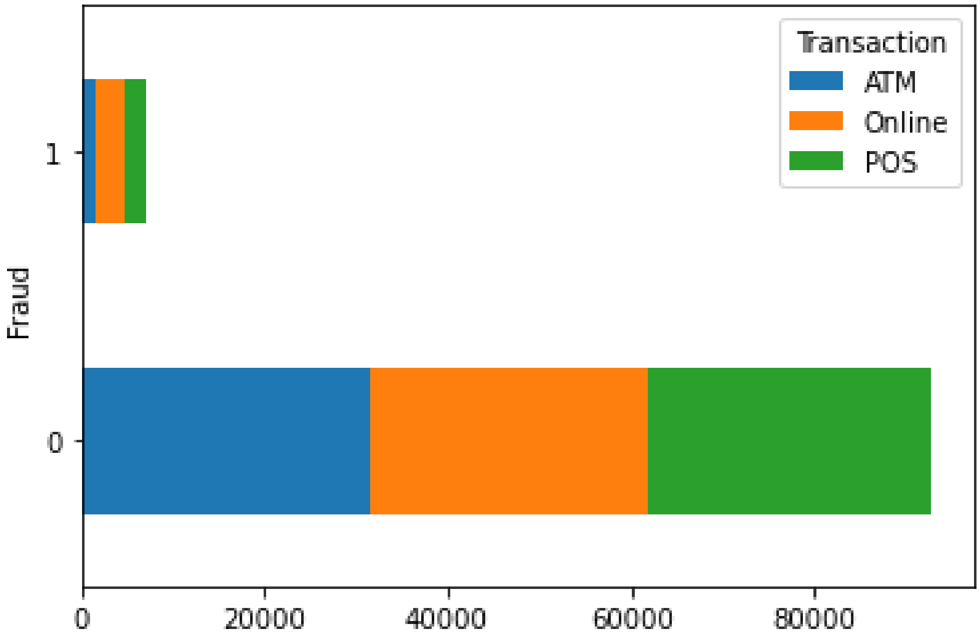


Male & females both have performed fraudulent transactions but there are the highest fraudulent transactions recorded against male customers. It means male customers are creating problems for the banks.

Transaction by Card Type & Fraud by Transaction

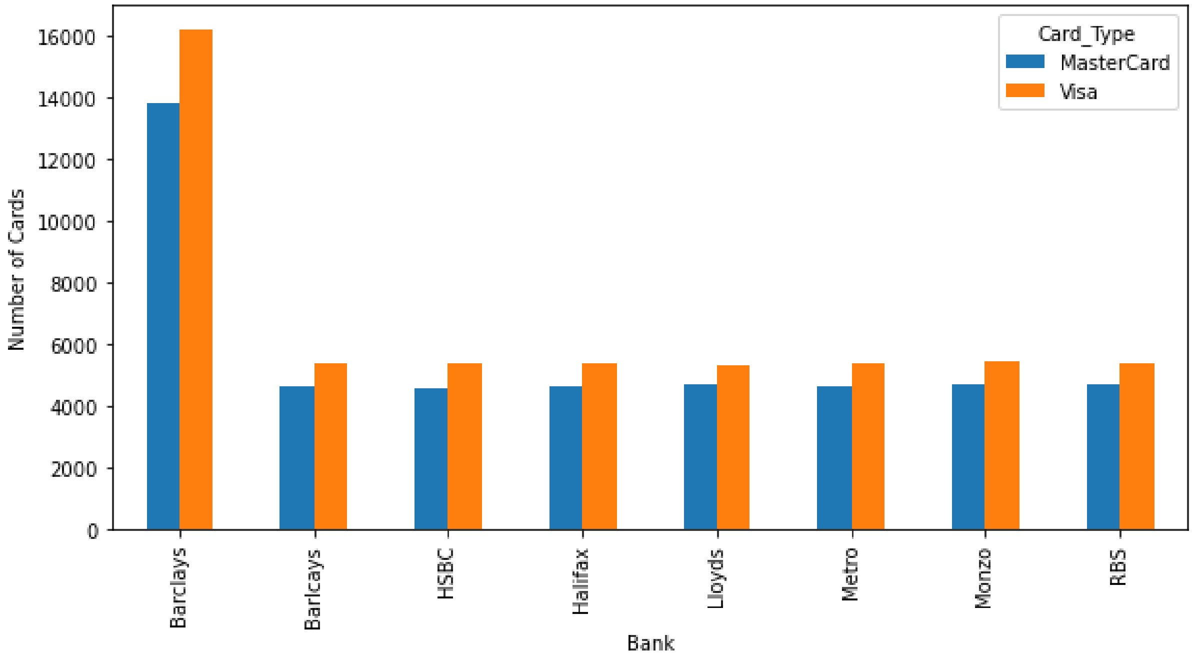


In the above analysis, we can see that most of the customers used Visa credit cards instead of MasterCard. Customers mostly performed transactions online, but ATM & POS represents almost equal transactions history.

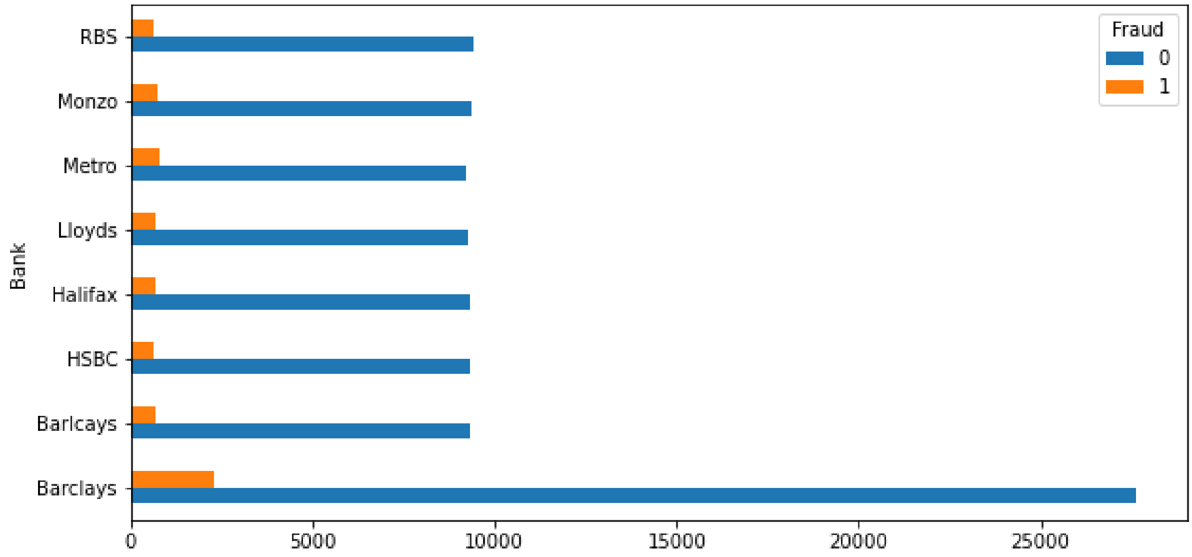


Most of the fraudulent transactions have been performed through online transactions because 1 represents the fraud and we can see stack graph against online has maximum records.

Cards & Fraud information by Banks

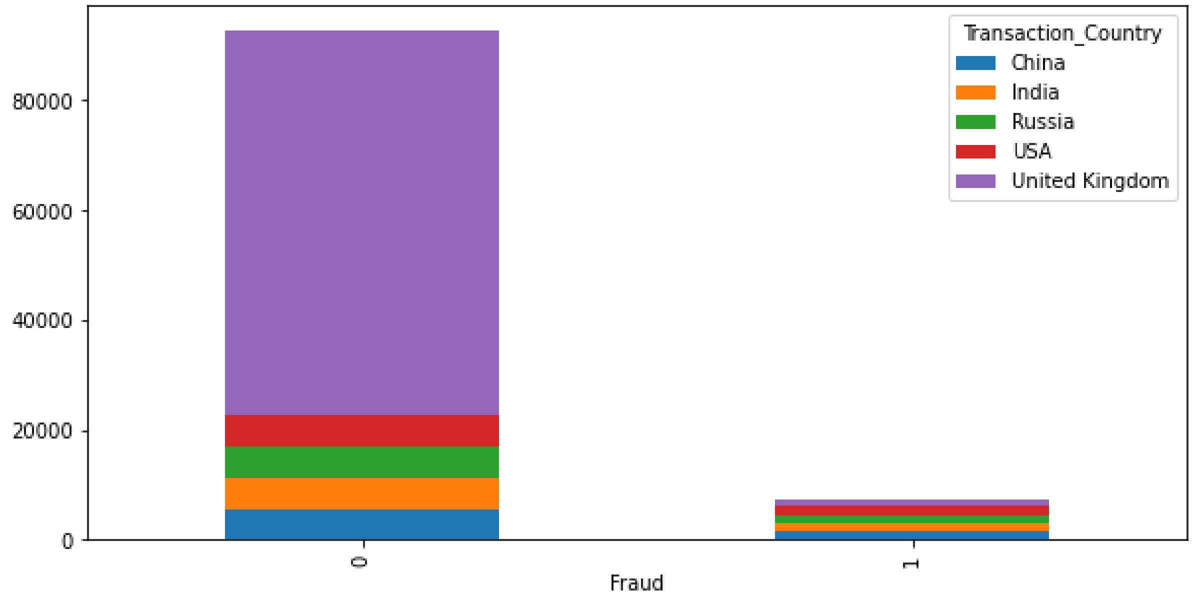


The above plot represents the distribution of credit cards regarding banks. Barclays bank has issued most credit cards (MasterCard & Visa) to their customers. All other banks have shown an equal distribution of both types of credit cards.



Most of the fraudulent transactions are recorded in Barclays bank because 1 represents the fraud and we can see a yellow bar chart against 1 has maximum transactions for this bank.

Fraud by Transaction Country



Most of the fraudulent transactions are recorded in 'USA' because 1 represents the fraud and we can see a red bar chart against 1 has maximum transactions.

Relevant Analytical Models

Based on the data analysis, I have recognized the relevant model which is the classification model. Classification is relevant because, from the exploratory data analysis, I have found there is a need to detect credit card fraud according to the given information. The target variable has binary class information so we can implement the classification model that can detect whether a "card transaction is potentially fraudulent or not". For this purpose, implemented the two classifiers model from the given models and it is a more relevant model according to the nature of the data (Jijo & Abdulazeez, 2021). There are other models available that are relevant models, but this assessment needs to implement only two selected analytical models that’s why used only the decision tree & AdaBoost classifier models that is most suitable & relevant to the analyzed data.

Selected Analytical Models

Decision Tree Classifier

The decision tree has been selected as the first classifier for the implementation of the classification model to detect credit card fraud. This is a supervised learning classification method that is mostly used to implement the classification & regression model (Mienye, et al., 2019). The decision tree algorithm works based on the likelihood

& consequences of an event occurrence and assigns a specific class to each event according to the conditions. In machine learning, it used "Gini index, entropy, and Log loss" functions to set the criteria of conditions in the tree (Jijo & Abdulazeez, 2021). With the help of this classifier, we can train the model while using training input & target variables.

AdaBoost Classifier

AdaBoost is first time introduced by "Yoav Freund and Robert Schapire in 1995" and at that time this is used for statistical classification purposes. In the earlier time, it is categorized under the meta-algorithms but now it is mostly used for ensemble learning under the machine learning algorithms (Wang & Sun, 2021). The AdaBoost algorithm has the potential to increase or boost the efficiency of classification problems such as binary classification. This classifier works based on the iteration process and iteration of the learning process continues until the model found a strong prediction on the testing data (Sailusha, et al., 2020). I have used this classifier as the second approach to implementing the analytical model that can predict whether a transaction is fraudulent.

Critically Evaluation of Analytical Models

Decision Tree Model Evaluation

While developing the analytical model, first divided the processed data into training & testing data while using the train test split package. 70% of data has been used for training and 30% data used for testing data. Decision tree classifier is implemented with the training data while calling the Decision Tree Classifier under the sklearn package (Yang, et al., 2021). Further, train the model while using training data and predict target or fraud values for the testing data. After implementing the model evaluated the model performance and found 96% accuracy in the fraud detection.

The implemented model used all the features but when I used the selected features based on the correlation analysis results model decreased the performance (Bashir, et al., 2020). With selected features, I have found 91% to 92% accuracy which is good but with all features, there is 96% accuracy on the testing data while predicting the fraud value which is good as compared to model with selected features.

AdaBoost Model Evaluation

To develop the AdaBoost analytical model, first split the processed data into training

& testing data while using the train test split package. 70% of data has been used for training and 30% data used for testing data. AdaBoost classifier is implemented with the training data while calling the AdaBoost Classifier under the sklearn package (Prettenhofer, et al., 2018). Further, train the model while using training data and predict target or fraud values for the testing data. After implementing the model evaluated the model performance and found 97.3% accuracy in the fraud detection.

The implemented model used all the features but when I used the selected features based on the correlation analysis results model decreased the performance. With selected features, I have found 92% accuracy which is good but with all features, there is 97.3% accuracy on the testing data while predicting the fraud value which is good as compared to model with selected features & decision tree model.

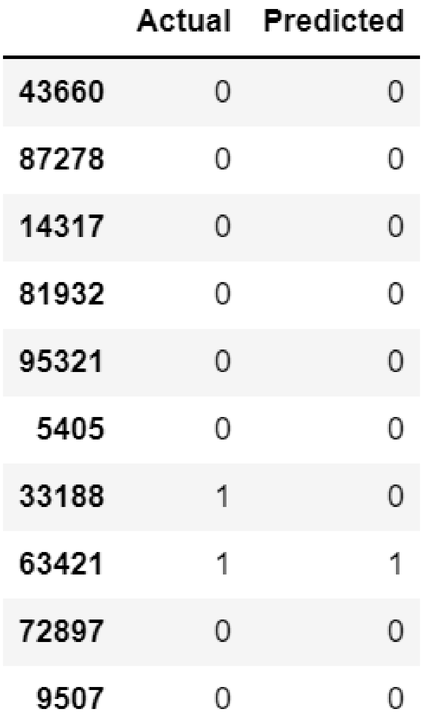
Logistic Regression Evaluation

To develop the Logistic Regression analytical model, first split the processed data into training & testing data while using the train test split package. 70% of data has been used for training and 30% data used for testing data. Further, train the model while using training data and predict target or fraud values for the testing data. After implementing the model evaluated the model performance and found 95% accuracy in the fraud detection.

The implemented model used all the features but when I used the selected features based on the correlation analysis results model decreased the performance. With selected features, I have found 92% accuracy which is good but with all features, there is 95% accuracy on the testing data while predicting the fraud value which is good as compared to model with selected features & decision tree model.

Critical Analysis of chosen model

To evaluate the models, there is a need to select the accuracy metrics and I have selected the confusion matrix and accuracy. Accuracy can calculate while manipulating the confusion matrix, but I have used the accuracy_score function to calculate the models' accuracy and the confusion_matrix function to calculate the confusion matrix terms (Savchuk & Doroshenko, 2021). Sklearn library provides a metrics package that can import to call the discussed accuracy metrics functions and I have called these functions while using this package (Pawan_Dubey & arvindpdmn, 2018). Both accuracy metrics require actual values and prediction values of the target variable. While comparing the actual & predicted values, these functions interpret the model's accuracy & performance.



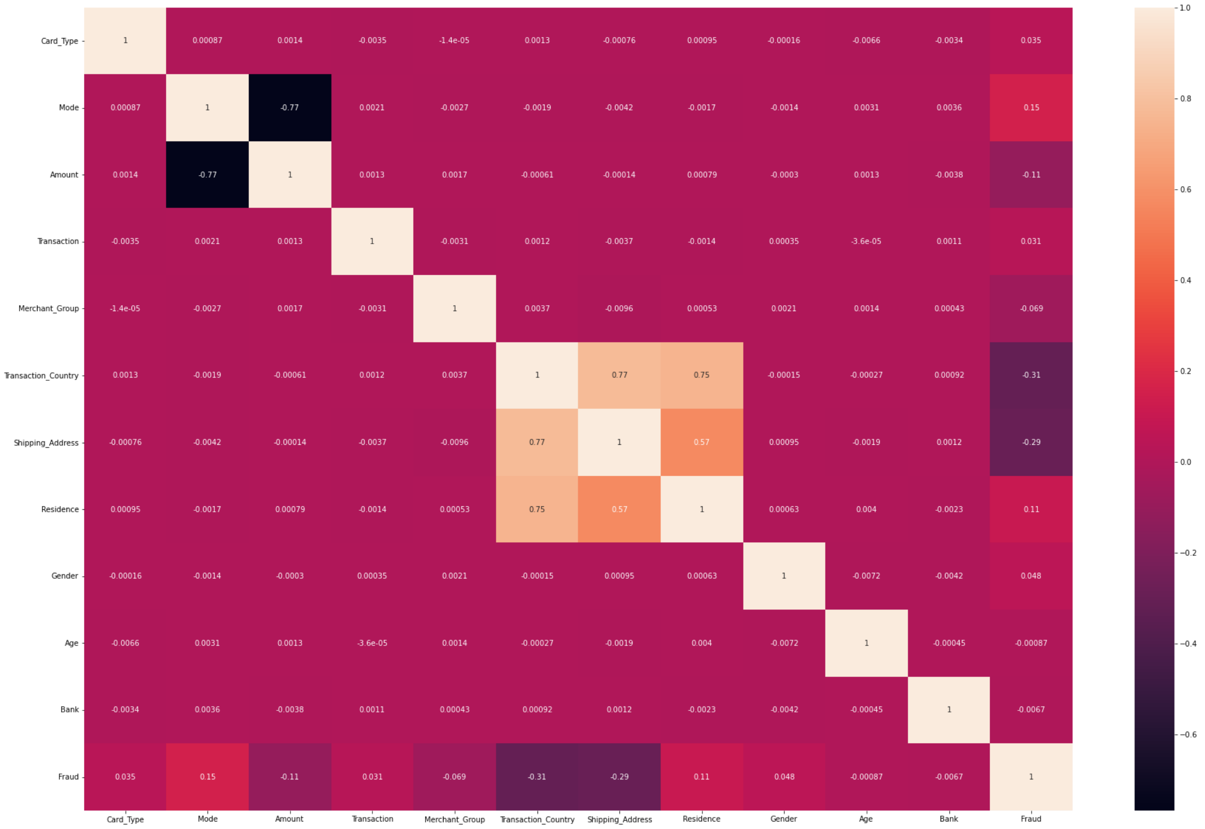
	Actual	Predicted
43660	0	0
87278	0	0
14317	0	0
81932	0	0
95321	0	0
5405	0	0
33188	1	0
63421	1	1
72897	0	0
9507	0	0

In the above figure, we can see almost all the actual & predicted values are the same except for 33188 observations. There are 30000 observations of the testing data and if we analyze all the observations for actual and predicted values there will be 1022 misclassified for the decision tree and 817 for the AdaBoost classifier.

Correlation Matrix

To find the most important features or parameters from the processed & cleaned data, I have used correlation analysis approach. Correlation analysis shows the importance of each feature regarding to the target variable or output feature. Correlation analysis represents the relationship or correlation between two variables, and we can identify the strong or weak relationship between features as per the score of correlation analysis. Correlation or relationship score between two features vary from 1 to -1 and if there is score close to 1 or greater than 0 than we can say the two features have strong & positive relationship between each other (Weichert, et al., 2019). Same as if there is score close to -1 or less than 0 than we can say the two features have weak & negative relationship between each other. Based on the analysis & correlation matrix, I have found 'Card_Type', 'Mode', 'Transaction','Residence', and 'Gender' are the more important features. Only these five features have positive relationship with 'Fraud' or target variable. The most correlated features are 'Shipping Address' with

'Transaction Country' and they have shown strong relationship. Same as 'Residence' with 'Transaction Country', and 'Residence with 'Shipping Address' have the strong & positive correlation. 'Amount' and 'Mode' are the only two features that have shown very low impact on the analytical model because there is negative & weak relationship between these two variables.



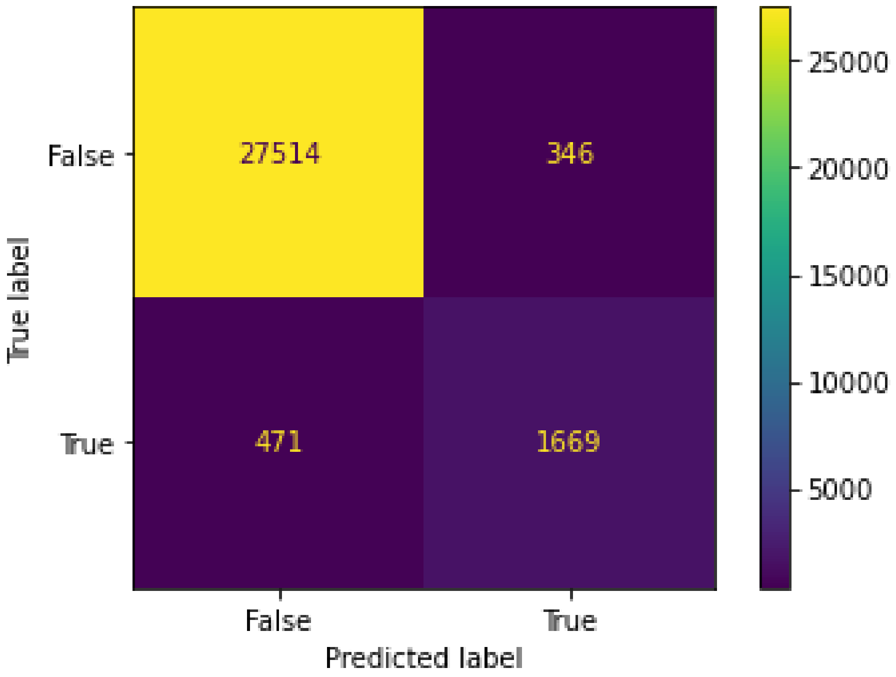
While using the features that extracted from above analysis, I have implemented two approaches with two analytical models. In the first approach used the selected variables and in the second approach used all the variables expect irrelevant variables that have eliminate in the data cleaning process. Second approach has shown highest accuracy on the test data.

Confusion Matrix

The confusion matrix represents the actual and prediction values of the target variable of the testing. It shows how many observations are correctly classified by the implemented model for the testing data and we can identify the model accuracy while interpreting the confusion matrix result. For both models, I have plotted the confusion matrix while using the matplotlib library in python. Confusion matrix values can interpret with the help of the metric library under the sklearn package.

Confusion Matrix of AdaBoost Model

Further, I estimated the AdaBoost analytical model performance while manipulating the confusion matrix. In the below plot, we can see the model performance and there are 27514 observations that are correctly classified by the AdaBoost model regarding the '0' (False) value of the fraud variable. Same as model correctly classified 1669 observations regarding '1' (True) of the fraud variable for the testing data (Weichert, et al., 2019). Almost 817 observations are not correctly classified or predicted by the model which shows the misclassification of the AdaBoost analytical model. Based on the result, we can identify AdaBoost model has classified more observations correctly for both cases as compared to the decision tree analytical model. It means the second model with overall features has a good performance on the prediction of fraud for the testing data.



Recommendation

According to the fraudulent transaction prediction for the credit card dataset, I would like to recommend the second model (AdaBoost analytical model) that can be used to predict the fraud and eliminate the rate of fraud in the market. With the help of this model, the banking sector can predict fraud transactions and can reduce the fraud rate in the future because the AdaBoost model has shown better accuracy & prediction performance as compared to the decision tree analytical model. While performing predictions against fraud transactions, banks can take useful & important decisions for improving their business experience and can grab more customers who are sincere with the company and can lead the business in the future.

References

Ayesha I. T. Tulloch, N. A. S. A.-G. E. B. N. B. C. R. D. G. E. D. O. F. H. G., 2019. A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nature Ecology & Evolution,* 2(1), pp. 1209-1217.

Bashir, D., Montañez, G. D., Sehra, S. & Lauw, P. S. S. &. J., 2020. An Information-Theoretic Perspective on Overfitting and Underfitting. *AI 2020: Advances in Artificial Intelligence,* 5(1), pp. 347-358.

CFI, 2022. *Decision Tree.* [Online] Available at: https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/ [Accessed 8 8 2022].

Ge, M., Bangui, H. & Buhnova, B., 2018. Big Data for Internet of Things: A Survey. *Future Generation Computer Systems,* 87(2), pp. 601-615.

Jijo, B. T. & Abdulazeez, A. M., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of applied science and technology trends,* 2(1), pp. 20-28.

Jijo, B. T. & Abdulazeez, A. M., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends,* 2(1), p. 20 – 28.

Mienye, I. D., Sun, Y. & Wang, Z., 2019. Prediction performance of improved decision tree-based algorithms: a review. *Procedia Manufacturing,* 35(2), pp. 698-703.

Pawan_Dubey & arvindpdmn, 2018. *Confusion Matrix,* s.l.: Devopedia.

Pedamkar, P., 2021. *AdaBoost Algorithm.* [Online] Available at: https://www.educba.com/adaboost-algorithm/ [Accessed 8 8 2022].

Prettenhofer, Peter, Louppe & Gilles, 2018. Gradient Boosted Regression Trees in Scikit-Learn. *UNPUBLISHED CONFERENCE/ABSTRACT (SCIENTIFIC CONGRESSES AND SYMPOSIUMS).*

Sahoo, K., Samal, A. K., Pramanik, J. & Pani, S. K., 2019. Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* 8(12), pp. 2278-3075.

Sailusha, R., Gnaneswar, V., Ramesh, R. & Rao, G. R., 2020. *Credit Card Fraud Detection Using Machine Learning.* Madurai, India, IEEE.

Savchuk, D. & Doroshenko, A., 2021. *Investigation of machine learning classification methods effectiveness.* LVIV, Ukraine, IEEE.

Sekine, S. & Nagata, Y., 2018. Application of AdaBoost to Taguchi's MT Method. *Total Quality Science,* 4(2).

Upadhyay, I., 2021. *Decision Tree in Machine Learning: Types, Advantages, Disadvantages in 5 Points.* [Online] Available at: https://www.jigsawacademy.com/blogs/data-science/decision-tree-in-machine-learning/
[Accessed 8 8 2022].

Wang, J., Yang, Y., Wang, T. & Sherratt, R. S., 2020. BIG DATA SERVICE ARCHITECTURE: A SURVEY. *Journal of Internet Technology,* 21(2), pp. 393-405.

Wang, W. & Sun, D., 2021. The improved AdaBoost algorithms for imbalanced data classification. *Information Sciences,* 563(1), pp. 358-374.

Weichert, D. et al., 2019. A review of machine learning for the optimization of production processes. *The International Journal of Advanced Manufacturing Technology,* 104(2), p. 1889–1902.

Yang, F., Wang, X. & Li, H. M. &. J., 2021. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. *Health Big Data and Artificial Intelligence,* 2(1).