CREDIT CARD FRAUD DETECTION.

A study on credit card fraud detection predictive modelling.

By

Deependra Dhakal

Bellevue University

MS Data Science

DSC630 T302-2223 Winter 2021

Predictive Analysis

Date: 03/03/2022

Abstract.

Ever since the use of e-commerce payment systems some people have found ways to access someone's finances illegally. This has been a major issue as the transactions can be easily completed online. The data breach leads to monetary theft and loss of customers along with the company's reputation.

Credit card fraudulent activities cost tens of billions of dollars worldwide. According to the Statistic Brain Research Institute [Ins18], one American over ten has been a victim of credit card fraud with a median amount of $399. According to the European Central Bank (ECB), the total level of card fraud losses amounted to 1.8 billion euros in the Single European Payment Area (SEPA).

The scope of this project to see if the use of predictive analysis can help in identifying the fraudulent activities and help prevent them in the future.

Executive Summary.

Credit card fraudulent activities cost tens of billions of dollars worldwide. It has negative consequences in the financial industry. This has been a major issue. The data breach leads to monetary theft and loss of customers along with the company's reputation.

Data mining has been applied on the dataset to automate and analyse frauds in online transactions. However, it was a challenging process as the profiles of genuine and fraud transactions changed frequently and the data was highly skewed.

The project was to investigate and check the performance of 4 different models (Random Forest Clasifier, AdaBoost Classifier, XGBoost Classifier, and LightGBM Classifier) on the dataset containing 284786 transactions of European cardholders obtained from Kaggle "Credit Card Fraud Detection: Anonymized credit card transactions labelled as fraudulent or genuine".

The performance of the models was evaluated using confusion matrix which is useful for measuring recall, precision, specificity, accuracy, and AUC-ROC curves. The results indicated that XGBoost had the highest accuracy of 98% followed by LightGBM classifier with accuracy of 95%, Random Forest Classifier 85%, and AdaBoost with accuracy 83%.

Contents

List of Figures.

1 INTRO/BACKGROUND.

1.1 Business Understanding & Problem.

According to the Credit Card Fraud Detection: Top ML Solutions in 2021. (2021, October 18), "Unauthorized card operations hit an astonishing amount of 16.7 million victims in 2017. Additionally, as reported by the Federal Trade Commission (FTC), the number of credit card fraud claims in 2017 was 40% higher than the previous year's number. There were around 13,000 reported cases in California and 8,000 in Florida, which are the largest states per capita for such type of crime. The amount of money at stake will exceed approximately $30 billion by 2020."

Online shopping is on the rise with stay home and covid situation. It's no surprise that there is an increase in credit card fraud. Consumers and businesses have adapted to the current scenario of pandemic and hence increasing the online credit card transactions which has opened a bigger playground for fraudulent activities.

According to the Credit card transaction fraud continues to climb to new heights. (2021, April 05). Retrieved from https://www.ncr.com/blogs/payments/credit-card-fraud-detection, "A 2018 study by the Federal Reserve showed the amount of card-present fraud in the U.S. declined from $3.68 billion in 2015 to $2.91 billion in 2016. Unfortunately, during the same period the loss from CNP fraud jumped from $3.4 billion to $4.57 billion. Another study by Javelin Strategy & Research revealed that CNP fraud is 81 percent more likely to occur than card present fraud."

There are various scenarios where fraudster may successfully perform fraudulent payments with a credit card. Consumers and businesses both suffer and pay the price which can be staggering. $34.66 billion is expected to be at loss related to payment cards in 2022 which has prompted the technologies being developed to guard and protect information online such as a data mining process called anomaly detection.

1.2 Scope.

The scope of this project to see if the use of predictive analysis can help in identifying the fraudulent activities and help prevent them in the future. The project will explain how we can analyse the transactions using modern datamining techniques.

2 METHODS

2.1 Data Understanding and Preparation.

For this project, I will be using dataset from Kaggle "Credit Card Fraud Detection: Anonymized credit card transactions labelled as fraudulent or genuine".

The dataset contains credit cards transactions of European cardholders in September 2013. The dataset presents transaction that occurred in two days. The dataset is highly unbalanced with 492 frauds out of 284,807 transaction is those two days accounting for 0.172% of frauds.

Dataset contains only numerical input variables which are the result of a PCA transformation. Due to the confidentiality issues, the original features and the background information has not been provided. Features V1, V2, V3,……,V28 are the principal components obtained with PCA, 'Time' and 'Amount' as the only features which have not been transformed. 'Time' consists of the seconds between each transaction and the first transaction in the dataset whereas 'Amount' is the transaction amount. Feature 'Class' is the response variable which takes 1 and 0 for fraud and genuine transactions respectively.

2.2 Modelling.

Machine learning as suggested learns and improve from experience without being explicitly programmed. Classifier is an instance of supervised learning algorithm that can learn from training data with correctly identified observations.

The data was investigated, checked for data unbalancing, visualized, and relationship between different features were analysed. After that, four predictive models will be used to perform validation by splitting data into 3 parts, a train set, validation set, and a test set. Following models were selected:

- RandomForrestClassifier.
- AdaBoostClassifier.
- XGBoost.
- LightGBM.

I will only use train and test set for the first two models and will use validation set for the third model. And for the fourth, I will use both train-validation split and cross-validation to evaluate model effectiveness to predict class value.

2.2.1 Random Forest Classifier

Random Forest, a classification and regression algorithm, is a collection of decision trees classifiers. A decision tree is built from a subset of the training set sampled randomly to train each individual tree.

As each tree is trained independently of the others, training is fast in random forest even for larger data sets with numerous features and data instances. Random Forest algorithm has been found to provide a good estimate of generalization error and to be resistant to overfitting.

2.2.2 AdaBoost Classifier.

AdaBoost (Adaptive Boosting) classifier is another ensemble classifier which combines weak classifier algorithm to form stronger classifier. AdaBoost can be applied to learn from classifiers shortcoming and propose more accurate model.

Combining multiple classifiers and assigning right amount of weight in final voting can have good accuracy score overall. It retains the algorithm by choosing the training set based on accuracy of previous training. AdaBoost assigns weight to each training item, misclassified item gets higher weight so that they appear in the training subset of next classifier with higher probability. Weight is assigned to the classifier based on accuracy after each training. More accurate classifier has higher weight as it will have more impact in outcome.

2.2.3 XGBoost Classifier.

XGBoost (Extreme Gradient Boosting) is a gradient-boosted decision tree machine learning library for regression, classification, and ranking problems which provides parallel tree boosting. XGBoost is highly accurate implementation of gradient boosting to push the limitations of computing power for boosted tree algorithms.

2.2.4 LightGBM Classifier.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms and grows tree vertically. LightGBM grows free leaf-wise whilst other grow level-wise. Leaf-wise algorithm when growing the same leaf can reduce more loss than a level-wise algorithm.

2.2.5 Area Under Curve (AUC).

AUC-ROC (Area Under Curve- Receiver Operating Characteristics) is a performance measurement metrics. It is one of the most important evaluation metrics to check the performance of the multi-class classification problem. ROC is the probability curve whereas AUC represents degree or measure of separability. Higher the AUC, better the accuracy of predictions.

3 Exploratory Data Analysis.

Some of the results after exploratory data analysis (EDA) are as follows:

- There is no null values or missing values in the dataset.
- The data is highly unbalanced with respect with the target variable ‘class’. Only 492 out of 284,807 transactions are fraudulent.
- Fraudulent transactions have more even distribution than valid transactions.

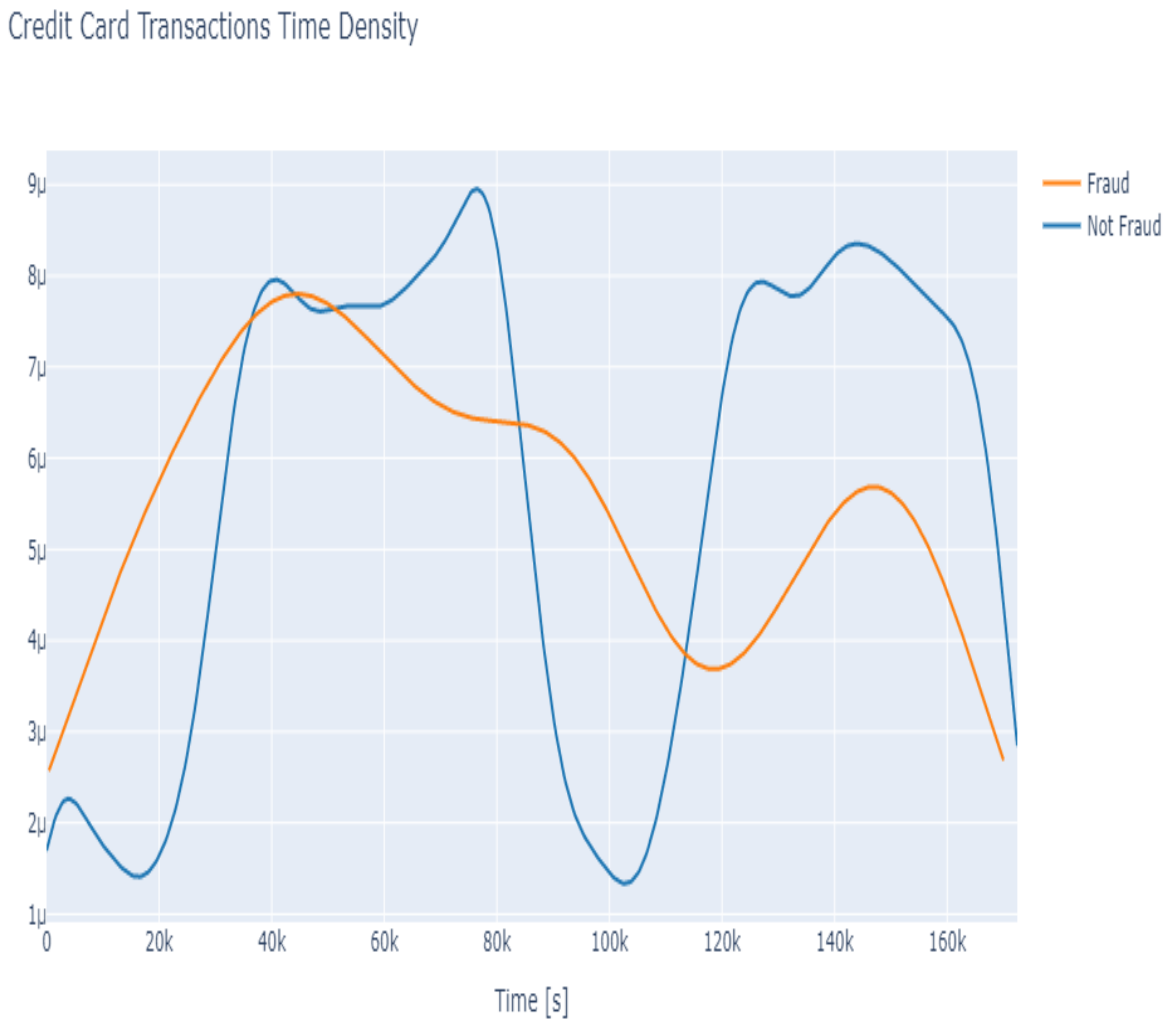

Fig 3.1: Credit Card Transaction Time Density.

- There is no notable correlation between the features.

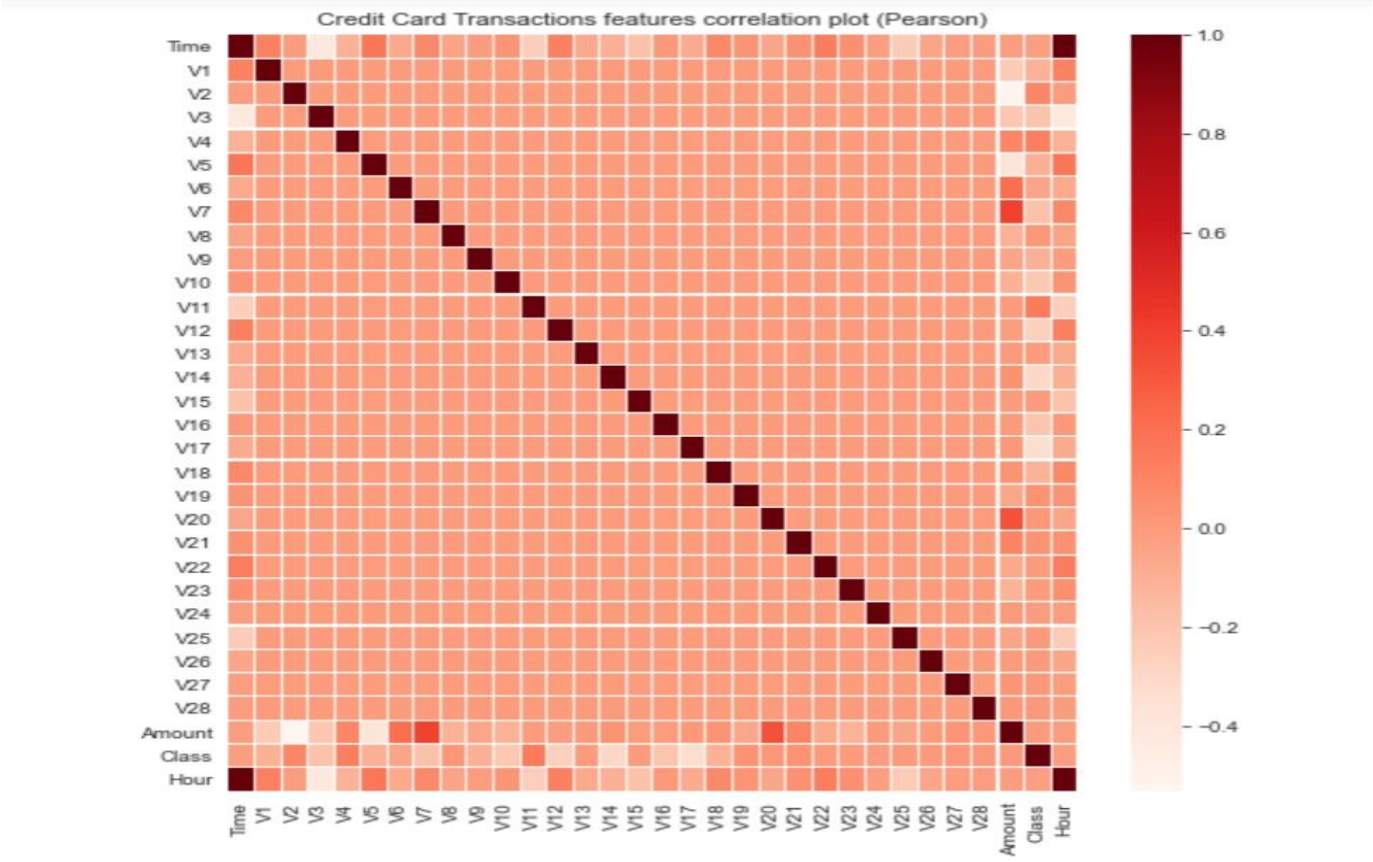

Fig 3.2: Pearson Correlation.

4 Predictive Modelling.

4.1 RandomForest Classifier.

I ran the model using the training set for training and validation set for validation. GINI was used as validation Criteria. Trained RandomForestClassifier using train_df and fit function and predicted larger values for valid_df using predict function. Next step was to visualize the feature importance.

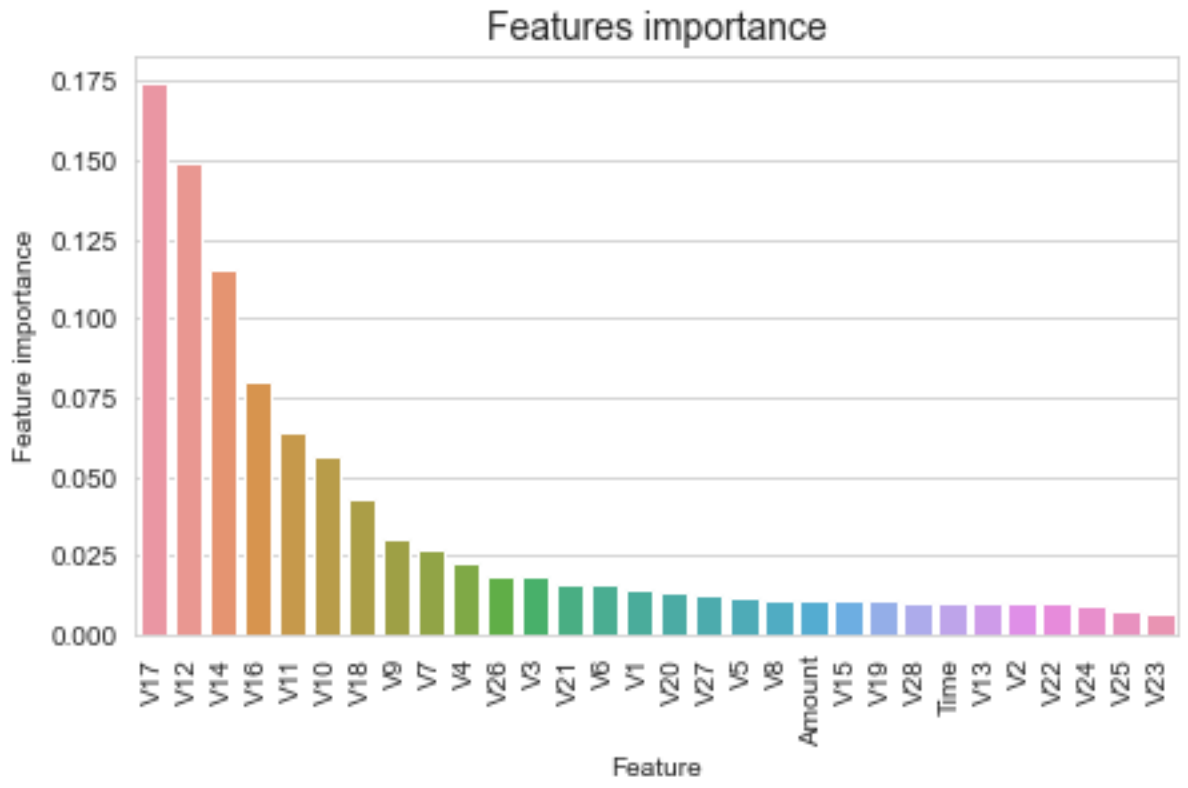

Fig4.1: RandomForest Feature Importance.

As per the figure, the most important features are V17, V12, V14, V16, V11, V10.

Confusion matrix is used to describe the performance of a classification model. It helps in visualizing the performance of the algorithm. It is a summary of predicted results.

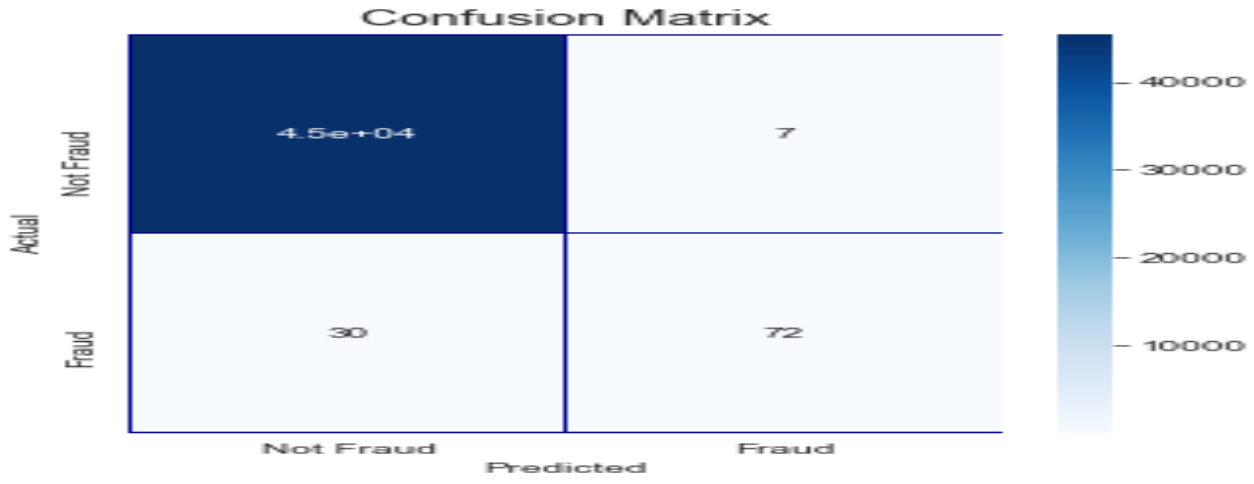

Fig 4.2: RandomForest Confusion Matrix

The ROC-AUC score obtained is 0.85.

4.2 AdaBoost Classifier.

I ran the model using training set and then used valid_df for validation. Trained the AdaBoost and predicted target variables. Then, Visualized the feature importance.

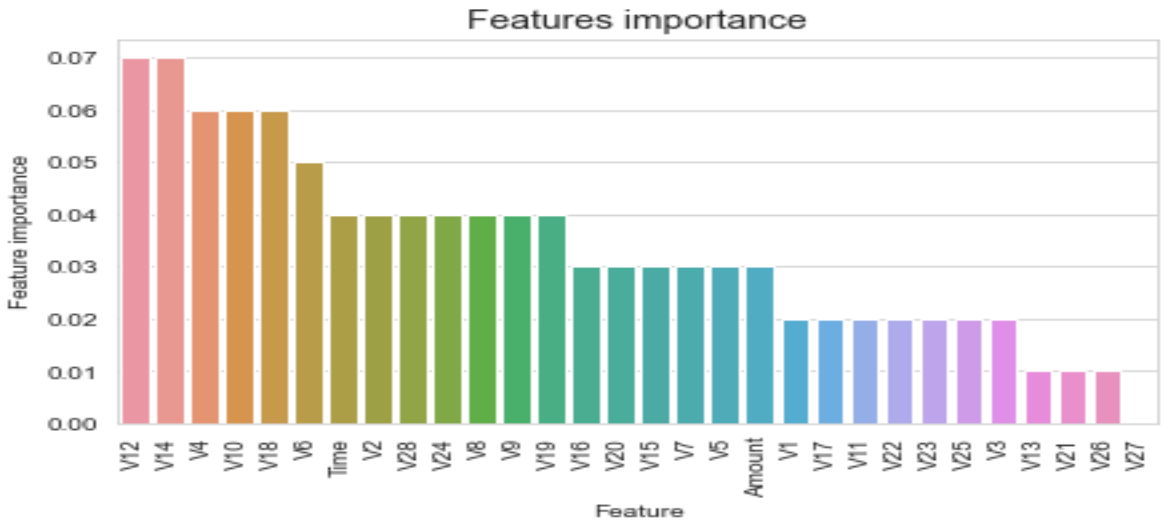

Fig 4.3: AdaBoost Feature Importance

The most important features are V12, V14, V4, V10, V18, V6.

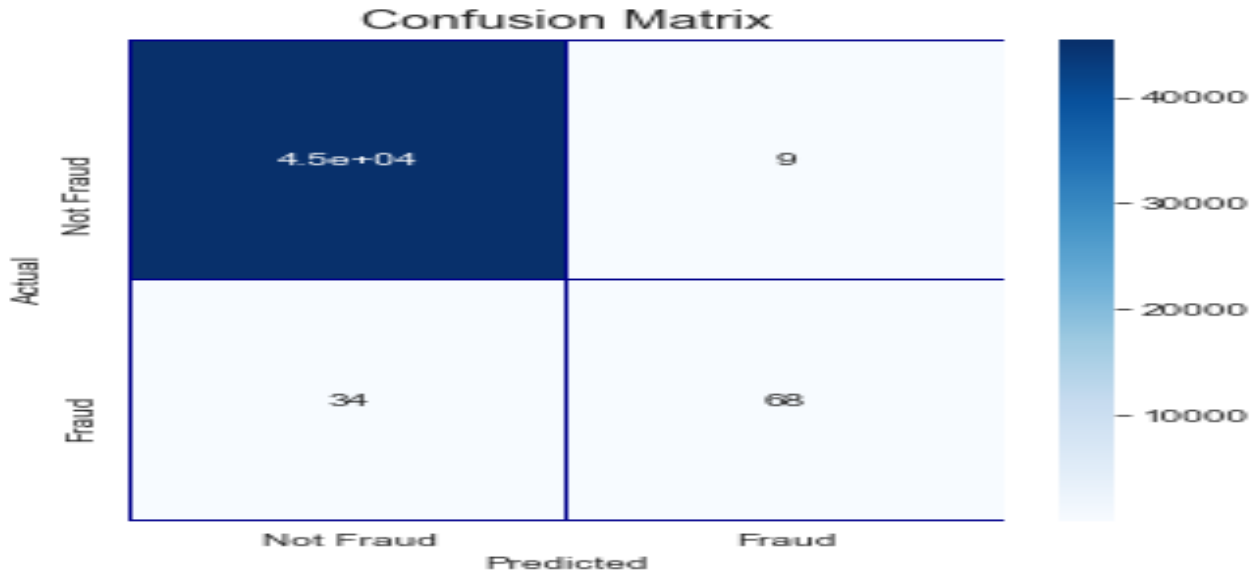

Fig 4.4: AdaBoost Confusion Matrix.

The AUC-ROC obtained was 0.83

4.3 XGBoost Classifier.

I ran the model using the training set and used validation set for validation. The model needed to be prepared before training. Initialized the DMatrix objects for training and validation starting from the dataset. After the model was prepared, trained it using dtrain and params obtained while creating model. Then visualized the feature importance.

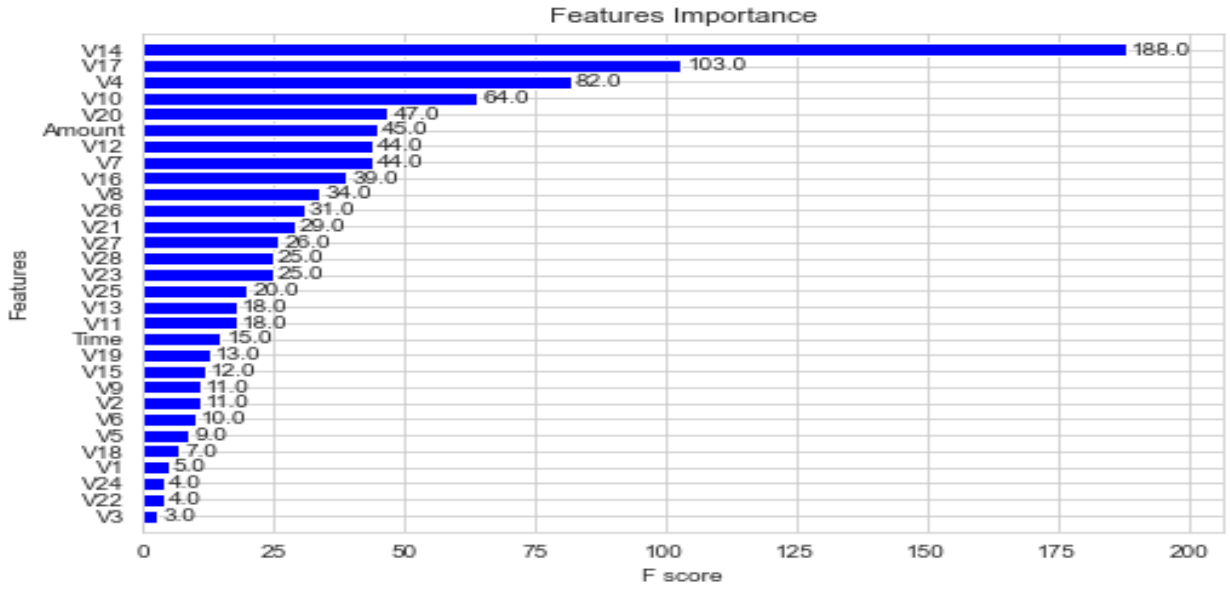

Fig 4.5: XGBoost Feature Importance.

The most important features are V14, V17, V4, V10, V20, Amount. ROC-AUC scare was calculated to be 0.977 using target values and predicted target values obtained in steps ahead.

4.4 LightGBM Classifier.

Parameters were set for the first LGB model. LightGBM was trained using the dtrain datra and params generated while creating model and then feature importance was visualized.

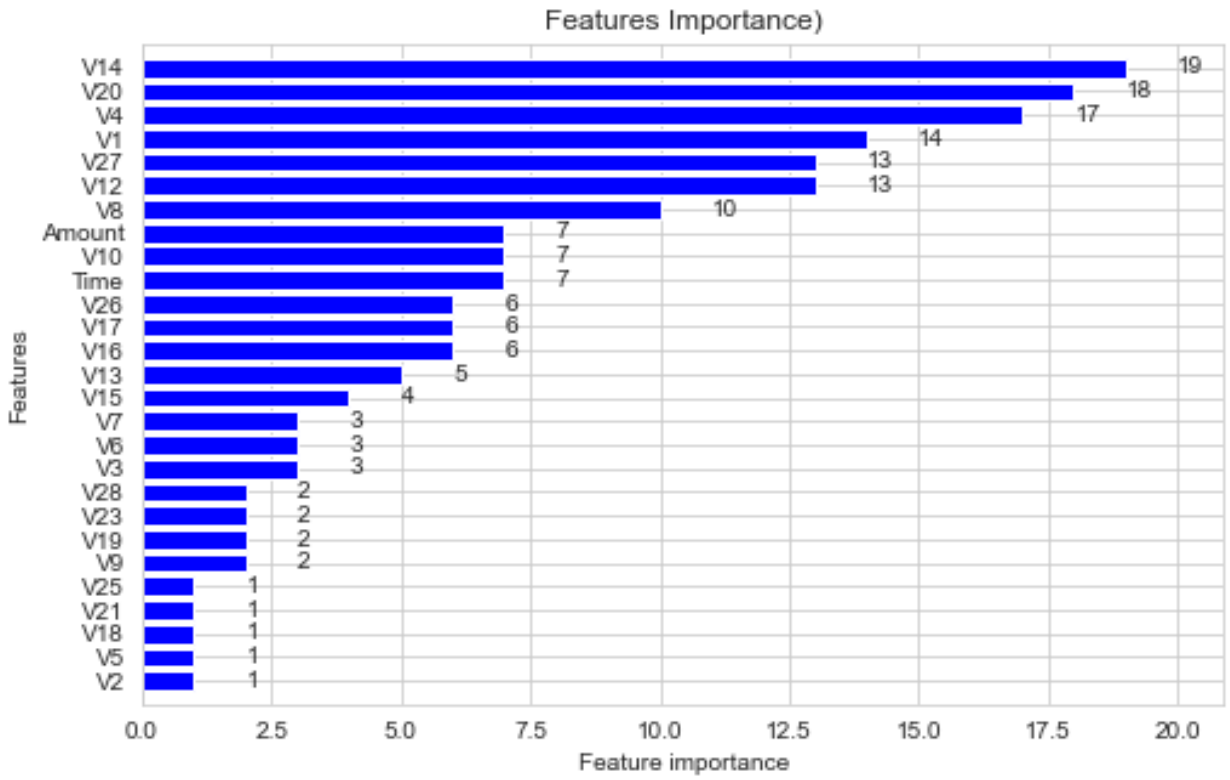

Fig 4.6: LightGBM Feature Importance.

V14, V20, V4, V1, V27, V12 and V8 are the most important features. ROC-AUC was calculated to be 0.95 using target values and predicted target values obtained.

5 DISCUSSION/CONCLUSION

After analyzing the data, checking for data imbalance, visualizing, and understanding the data, I worked with 4 predictive models. The data was split into 3 sets: train set, validation set, and test set. The first two models (RandomForestClassifier and AdaBoostClassifier) used only train and test set. I used the validation set in XGBoost model to validate the training model and then used the model with best training step to predict the target value from the test set. Both train-validation split, and cross-validation was used to evaluate the effectiveness of the LightGBM model.

From the results, it can be concluded that RandomForest Classifier has an accuracy of 85%, AdaBoost 83%, XGBoost 97%, and LightGBM 93%. It concludes that XGBoost Classifier has the higher accuracy in detecting credit card fraud using machine learning.

ACKNOWLEDGEMENTS.

First and foremost, I would like to thank my wife for supporting and motivating me throughout. Another name that I cannot miss is of my fellow Student Wyatt Rasmussen without whom I would have lost track of what I was doing. His reviews helped me be on track throughout the project. I would also like to express my sincere gratitude towards the Bellevue University, my professors and fellow students for valuable guidance and helpful comments.

REFERENCES

- Credit Card Fraud Detection: Top ML Solutions in 2021. (2021, October 18). Retrieved from https://spd.group/machine-learning/credit-card-fraud-detection/
- Credit card fraud scenarios. (n.d.). Retrieved from https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_2_Background/CreditCardFraud.html
- Credit card transaction fraud continues to climb to new heights. (2021, April 05). Retrieved from https://www.ncr.com/blogs/payments/credit-card-fraud-detection
- Using Your Data to Stop Credit Card Fraud: Capital One and Other Best Practices - The Databricks Blog. (2021, July 13). Retrieved from https://databricks.com/blog/2021/07/13/using-your-data-to-stop-credit-card-fraud-capital-one-and-other-best-practices.html
- Adekanye, T. (2021, September 15). Predicting Credit Card Fraud. Retrieved from https://medium.com/low-code-for-advanced-data-science/predicting-credit-card-fraud-b9c67b8b0997
- Brownlee, J. (2021, February 16). A gentle introduction to XGBoost for applied machine learning. Machine Learning Mastery. Retrieved March 3, 2022, from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
- Dwivedi, R. (n.d.). What is LIGHTGBM algorithm, how to use it? Analytics Steps. Retrieved March 3, 2022, from https://www.analyticssteps.com/blogs/what-light-gbm-algorithm-how-use-it
- Kharwal, A. (2021, June 25). Lightgbm in machine learning. Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript. Retrieved March 3, 2022, from https://thecleverprogrammer.com/2021/01/15/lightgbm-in-machine-learning/
- Random Forest: Introduction to random forest algorithm. Analytics Vidhya. (2021, June 24). Retrieved March 3, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
- Sklearn.ensemble.adaboostclassifier. scikit. (n.d.). Retrieved March 3, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html