DETECTION OF CREDIT CARD FRAUD USING MACHINE LEARNING MODELS

BY VERONIKA POTIIKO

APRIL, 2024

PROJECT SUBMITTED AS FULFILLMENT OF REQUIREMENTS OF THE ST3189 MACHINE LEARNING COURSEWORK AT THE UNIVERSITY OF LONDON.

Table of contents

1. Substantive issue - Fraud Detection

Credit card fraud has emerged as a widespread business and user challenge, with the rise of online transactions granting high accessibility. Due to the high frequency of transactions occurring daily, fraud detection - the process that detects illegitimate uses of credit cards and prevents scammers from obtaining money by unauthorized means, is significant for safeguarding assets as businesses could be involved in money-laundering, terrorist financing, or fraud.
Financial companies and banks rely on machine learning algorithms to flag, meaning identify as suspicious, transactions and / or users of the service. As a financial crime investigator, a part of my role revolves around identifying if the machine learning algorithms have falsely or correctly flagged individuals by analyzing transactional attributes such as sending and receiving locations, relationships to counterparties, amounts sent and more. The following paragraph explains how machine learning algorithms fail to properly identify suspicious transactions which includes an example of completely hypothetical individuals and any similarity with existing individuals is a coincidence.
A common tactic employed by fraudsters is 'splitting', wherein large sums of money received is divided into smaller chunks and distributed among several recipients in a short timeframe. This method is frequently used to cover the illicit origin of funds. Furthermore, some geographic regions serve as 'red flags' for illicit activity, including the sanctioned territories, such as Iran and North Korea, and high-risk countries, for example, Sudan and Romania. Consider the scenario of Andrei Stefan Creţu, a Romanian national residing in Germany, sending monetary family support to their relatives located in Romania. And Andreas Bauer, a German citizen receiving from the United States and sending to rural areas in Romania, claiming the counterparties are acquaintances that they have met on holiday in Croatia. But is in fact engaged in transactions linked to human trafficking. A financial crime analyst can discern between legitimate and suspicious individuals by examining the ethnicity, receiver's profiles, destination of transactions on city level, and identifying splitting.
Current machine learning algorithms lack the advances to analyze these multifaceted elements contributing to fraudulent transactions. There is, however, a potential to reduce time consumption and enhance detection capabilities for analysts by identifying which factors are most influential for fraud detection, which is the goal of this investigation.

2. Identify Research Questions (RQs)

1. Which attributes are not influential for determining if the transaction is fraudulent?
2. Are there underlying patterns or structures within the data that could contribute to fraud detection?
3. Which supervised learning classification method performs best in detecting credit card fraud, and how do their performance metrics compare?
4. Which regression technique performs the best in predicting fraudulency of a transaction?

3. Explanation of the data set

The data set used for credit card fraud detection was used which illustrates transactions made through credit cards by European cardholders in September of 2013. A total of 284,807 transactions were made over a two-day period of which 492 were fraudulent. The dataset comprises solely numerical input, due to concerns of privacy, there are variables resulting from PCA transformation. Features V1 through V28 (columns) represent principal components derived from PCA, with only 'Time', 'Amount', 'Class; remaining unchanged. 'Time' refers to seconds elapsed between the first and indicated transaction, 'Amount' is the transactional amount in euro currency, and 'Class' is the response variable which indicates fraudulent or non-fraudulent transactions used 1 and 0 respectively. The dataset has the potential to exhibit class imbalance, therefore, the degree of imbalance needs to be assessed.

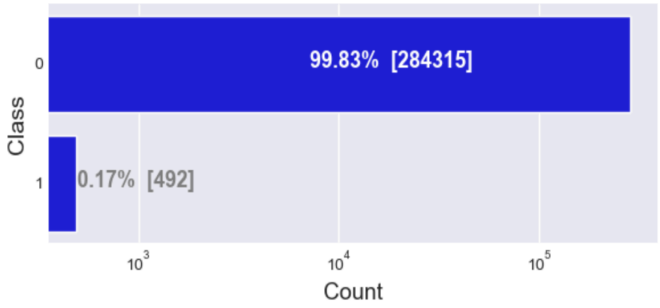
4. Methodology

The process utilized existing libraries that allowed the data set to be processed, transformed, graph relevant charts, and apply different learning techniques. The libraries include NumPy, Pandas, Sklearn, MatPlotLib, and SeaBorn. The data set was imported, inspected, and checked for null values of which none were found.

4.1 Class imbalance analysis

As seen in Figure 1, a bar graph was created to understand the degree of class imbalance present in the data set. The fraudulent class consisted of only 0.17% of the dataset, making it highly imbalanced. This may lead to biased model performance. It is important to choose appropriate learning methodologies and perform mitigation of skewness by transforming the data set.

Figure 1: Bar graph of class distribution



4.2 Predictor variable analysis

Non-PCA features, 'Amount' and 'Time' were graphed against 'Class' to understand visually if there is a significant difference between these features. Figure 2 demonstrates that the majority of legitimate transactions are in quantities of less than 10000 euros, with the presence of potential outliers. Fraudulent transactions consist of smaller amounts never exceeding 5000 euros mark.

Figure 2: Scatter plot of 'Class' against 'Amount'

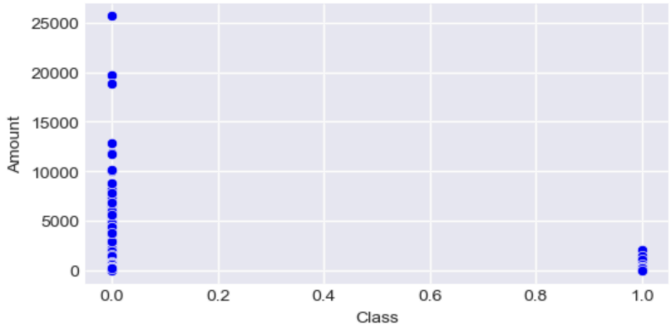
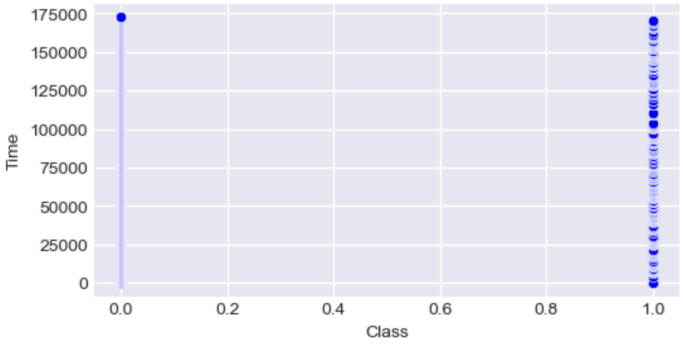


Figure 3 shows that there isn't a particular pattern difference between the classes as the data points are evenly distributed throughout the 2 day period. 'Time' variable was dropped during the analysis as there weren't significant differences that could cause detection between fraudulent and non-fraudulent transactions.

Figure 3 : Scatter plot of 'Class' against 'Time'



Analysis of the PCA features was conducted by creating a grid of Kernel Density Estimation comparing the graph and area under the graph of fraudulent and non-fraudulent transactions. Figures weren't attached to the report as there are 29 PCA variables and would exceed page limit. Please, refer to section '1.3 Analyze Predictor's Variables' of the python notebook for the figures.

Columns of 'Time', 'V13', 'V15', 'V22', 'V24', 'V25', 'V26', and V28' were dropped, as area covered by fraudulent and non-fraudulent transactions were similar, indicating that these variables will not be a good predictor of fraudulency of the transaction.

The research question 'Which attributes are not influential for determining if the transaction is fraudulent?' can be answered as the 'Time' columns in other PCA features were dropped meaning some attributes are indeed not useful for determining if a transaction is fraudulent.

5. Unsupervised learning

Methods of machine learning include supervised and unsupervised methods of learning. Unsupervised learning analyzes unlabelled data and clusters them into groups while supervised learning is fed labeled data to predict outcomes and identify patterns according to Delua, 2021.
Clustering is grouping data points based on common features. K-means clustering was implemented, which is the grouping of K clusters, in which K is the number of clusters that the data set is divided into (IBM, 2021). This technique was chosen as it can partition data in a simple manner for visual comprehension.

5.1 Elbow method for k-determination

Since k-means clustering stores centroids which are used to define clusters within a data set, it is important to cluster the data into an optimal number of clusters for accurate performance. The elbow method, which determines the within-cluster sum of squares, was employed due to its ability to quantify the optimality of the chosen k-number. Figure 4 demonstrates the results of the visualizer, in which the 'elbow dent' or the linear decrease of inertia, is clearly visible at 2. Therefore, 2 clusters were implemented. The result can be attributed to the fact that there were 2 classes of transactions within the data set.

Figure 4 : Elbow Method

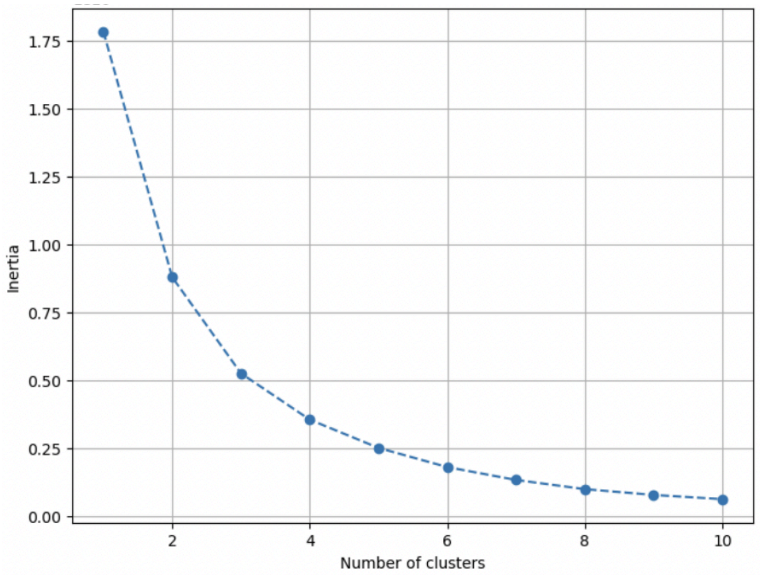
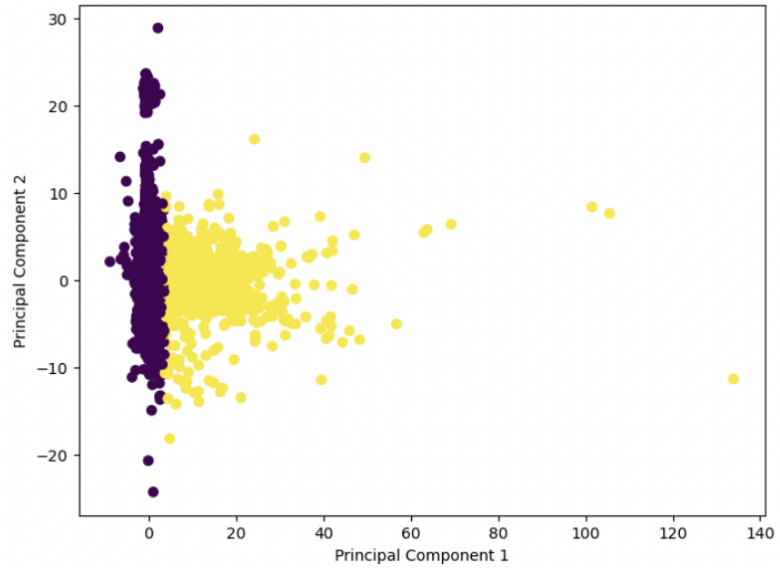


Figure 5 : Scatter Plot of KMeans Clustering



5.2 KMeans clustering

Data set was standardized and underwent dimensional reduction using principal component analysis, prior to application of the K Means model.
Figure 5, which demonstrated results of K Means clustering, has x-axis as Principal component 1 (PC1), or direction of original feature space in which data varies the most, and y-axis as Principal component 2 (PC2), perpendicular diffraction to Principal component 1.
Cluster split can be observed at PC1 = 0, suggesting that division in the data along the principal component. Purple cluster has a presence before and after PC2 = 0. The yellow cluster contains a larger and more distributed proportion of data with potential outliers. It is highly likely that PC2 does not contribute as significantly to the difference between clusters identified by Kmeans as the clusters have a similar presence on it, being centered around PC2 = 0 and both containing points beyond it. These patterns suggest that both clusters extend across PC2 = 0, there is a difference in distribution of density, with the purple cluster being slightly highly concentrated which provides an underlying structure of the data and potential factors that contribute to credit card fraud.

The research question ‘Are there underlying patterns or structures within the data that could contribute to fraud detection?’ can be answered as formation of clusters supports that there are underlying patterns that could contribute to fraudulent detection.

6. Supervised learning - Classification

Supervised learning, which uses previously provided labeled data, is split into 2 main aspects - classification and regression. Classification is a method which identifies classes for units of data sets based on previously provided examples. This method does not assume that the data is quantifiable but instead can include categorical data (Chouinard, 2023). Methods of classification that were implemented are logistic regression, decision trees, random forests, and neural networks.
Logistic regression accomplishes binary classification, or splitting into just two possible outcomes, by calculating probability (Kanade, 2022). Logistic regression was chosen due to its concise interpretability and high suitability for the chosen credit card fraud data set, as the transactions are either fraudulent or not.
Decision trees are recursively partitioned the feature space by basing it off the most discriminative features. This creates a tree-like structure, in which internal nodes represent a decision based on the feature and leads to the leaf nodes that are the predicted class labels (IBM, 2023). This method was chosen due to its simple interpretability decision rules.
Random forests utilize multiple decision trees to provide robust predictions. Each tree is trained on a random subset of data which is the reason this method was chosen, as this effectively mitigates overfitting and generalization (Yeon and Wilbern).
Neural networks work through connected nodes which are organized into layers. Each node performs small computation and passes the output to the next layer. During iterative training processes, neural networks adjust the weights between connections and nodes to learn patterns in data (Hardesty, 2017). Neural network was chosen because of its excellent predictive performance but may lack on the interpretability aspect.

6.1 Data pre-processing

Data set was split into dependent (y) and independent sets (x). With ‘x’ set including all information but dropping the ‘Class’ feature, and ‘y’ containing the information of the ‘Class’ feature.
It was further split into a training and testing set, with the test set consisting of 20%. This split ensures the balance between train and test sets, prevents overfitting, uncovering matters which are not actually present or are irrelevant, but still ensures that the model is able to generalize unseen data. The training set was further used so that models could recognize patterns and map function of independent variables. The test was necessary to estimate the performance of the models post testing.
As observed in the data pre-processing section, the set is highly imbalanced. PowerTransformer was applied on the ‘x’ training set to perform a mathematical transformation and mitigate the set's skewness.
The training and testing sets were standardized by applying the StandardScaler feature which scales data to mean of 0 and deviation of 1. The function subtracts the mean and divides by the standard deviation. This helped the models converge faster and perform optimally as the features are on the same scale. Standardization was performed separately for training and testing sets, to prevent data leakages, or sharing data among training and testing sets which resulted in false model performance evaluation.
Two functions were defined for model evaluation, so that it could return model results after its application and be able to input the values into model comparison histogram. To evaluate model performance, accuracy, precision, recall, and f-1 score were utilized. For visual assessment, the AUC, or area under the curve, and ROC curve were calculated and graphed.

6.2 Model comparison

Receiver Operating Characteristic (ROC) curve was implemented for comparative analysis, which plots y-axis with the true positive rate, meaning correctly predicting positive class, against the false positive rate, incorrectly predicting positive class, at the x-axis at different threshold settings, as seen in Figure 6. Additionally, a histogram chart has been graphed, as seen in Figure 7, comparing the metric scores of different classification techniques, with the y-axis indicating the score ranging from 0 to 1, and x-axis indicating the metrics with specification.

Figure 6, demonstrated that all models, with the exception of decision trees, demonstrated high true positive and false positive rates, indicating strong discriminatory powers. Figure 7 supported that neural network had the strongest discriminatory power at 97.62%. This suggests that decision trees have a significantly low effectivity in fraud detection.

Accuracy metrics has had a consistent performance across all classification techniques, with the lowest score of 99.91% in decision trees and highest 99.96% in random forest. This means all techniques are highly proficient in correctly classifying instances.

Logistic regression's precision, recall, and f1 scores demonstrate defects in identifying fraudulent transactions. Precision at 81.48% suggests a relatively high proportion of true positives among predicted positives, the recall, however, is much lower at 67.35%, suggesting a high number of false negatives.

The decision tree has exhibited a lower precision and a higher recall than logistic regression, at 74.23% and 73.47% respectively. Indicating that decision tree

Figure 6: ROC curves of classification methods

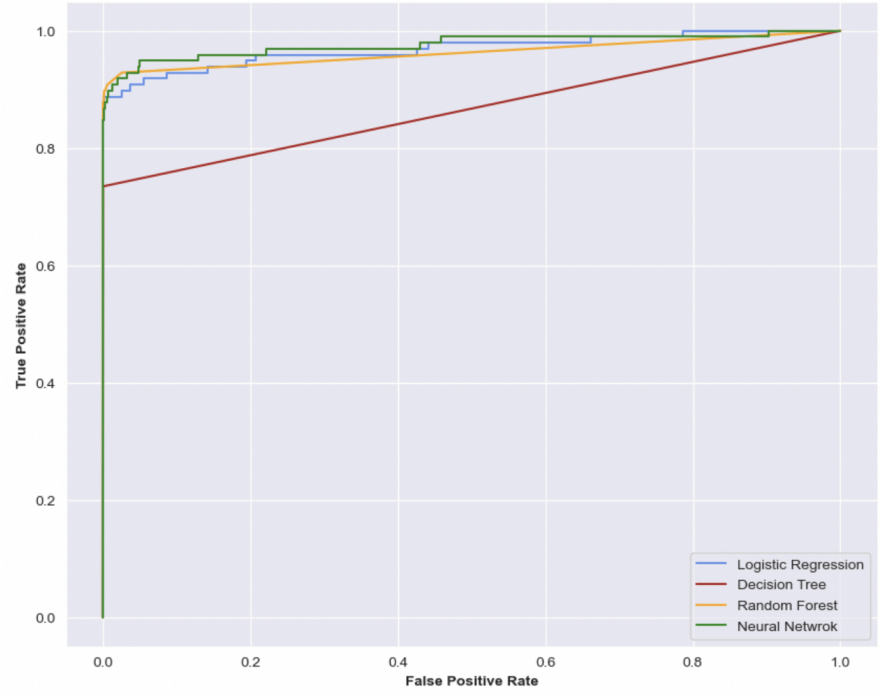
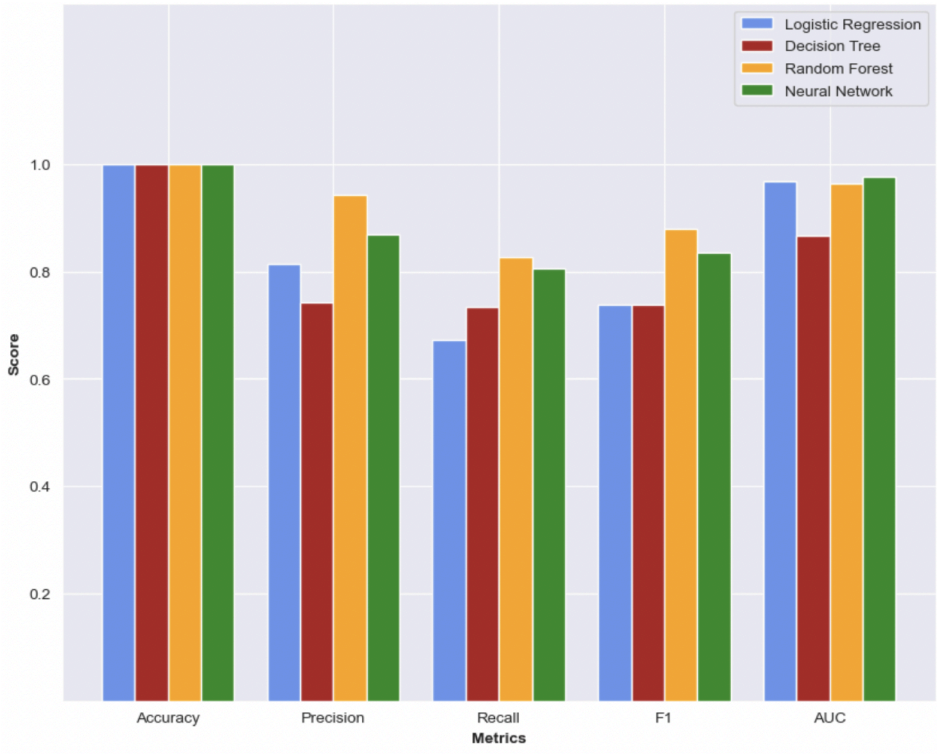


Figure 6 : Histogram of metrics scores of classification methods



had a better performance at identifying fraudulent transactions. The f1 score was similar at 73.85%. Despite the minor improvements of the recall score, it still has a low discriminatory power, due to its low AUC score of 86.71%.
There have been found significant improvements across all metrics for the random forest technique. In addition to the highest accuracy of all, its precision at 94.19% indicated a high ability at correctly identifying fraudulent transactions and still keeping a low amount of false positives. The recall score, 82.65%, indicated that the amount of false negatives has majorly reduced compared to previous techniques. F1 score, at 88.04% indicated a balanced performance in precision and recall. Exceptional discriminatory power demonstrated at 96.29% of AUC score, making random forest the technique with the best performance at identifying fraudulent transactions.
Neural network presented a strong performance but not as much as random forest. With accuracy of 99.95%, the neural network was highly accurate at classifying instances. A precision of 96.81% and a recall of 80.61%, the neural network had a balance between identifying true positives and minimizing false negatives, allowing F1 score to reach 83.60%. Neural network had a substantially strong discriminatory power, as seen in the highest AUC score of 97.62%, furthermore proving its high effectiveness in credit card fraud detection. The research question 'Which supervised learning classification method performs best in detecting credit card fraud, and how do their performance metrics compare?' can be answered, as evaluation has demonstrated that random forest performed the best in credit card fraud detection.

6.3 Comparing results to existing study

In the study of Gradient Boosting Techniques for Credit Card Fraud Detection by Kasarami et al. LGBM was the best-performing model with 97% accuracy which outperformed Logistic Regression, Neural Networks, Auto Encoders, K-Means Clustering, and Cat Boost. However in my study random forest was identified as the best performing model. Both suggested potential improvements such as using hybrid models for better performance of credit card fraud detection.

7. Supervised Learning - Regression

Regression is a method which makes qualitative predictions of the dependent, or target, variable based on the independent, or predictor, variable (Kurama, 2018). It aims to find the relationship between the two variables.
Linear regression assumes a linear relationship between input and output variable. The model is represented by a linear equation, where the coefficients show the weights assigned to each independent variable.
By adjusting coefficients, linear regression minimizes the differences between observed and predicted values. Its fundamentality was the reason for choosing this method. Ridge regression is a regularization technique used to address the issue that arises in linear regression - multicollinearity. It ends a 'penalty term' which prevents overfitting by shrinking the coefficient estimates towards zero. Ridge regression is useful for datasets with high dimensionality, which is the reason it was chosen.
Gradient boosting regression is an ensemble learning technique that combines predictions of multiple weak learners. Such as the decision trees, that creates a storing predictive model. It sequentially adds new models to correct errors made by the previous models, which reduces the overall prediction error (Vadapalli, 2020). The results are highly accurate and capture complex relationships in the data, which is why it was employed.

7.1 Model Comparison

The scores of R-squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error were used to analyze effectiveness of the model. R2 measures how a regression model fits the observed data using values closer to 1 which means a greater fit. MAE is the average absolute difference between predicted and actual values, which shows the prediction accuracy. RMSE is the square root of the average squared differences between predicted and actual values.
This demonstrates which model has larger errors and assess predictive performance

Scores were graphed on a bar graph as seen in Figures 7, 8 and 9. Linear and ridge regression have shown similar performance across all scores. R2 values swerve both approximately 0.22, meaning that they explain only 22% of the variance in the data. MAE was approximately 0.165, meaning on average, the predicted values deviate from the actual values by around 0.165 units. And RMSE was 0.406 which indicates the magnitude of the errors in the predicted values, which is quite high.

Gradient boosting, however, strongly outperforms both linear and ridge regression. R2 was approximately 0.95, meaning it explains about 95%n of the variance in the data, suggesting a strong fit. MAE of 0.02 shows that the average absolute difference between the predicted and actual values is very low, which indicates high accuracy. RMSE of 0.14 is significantly lower than other regressions, supporting its better fit to the data set and prediction accuracy.

The research question 'Which regression technique performs the best in predicting fraudulency of a transaction?' can be answered has shown the highest R2 score, indicating its optimal correlation with determining fraudulency of a transaction.

Figure 7: Histogram of R2 scores of Regression

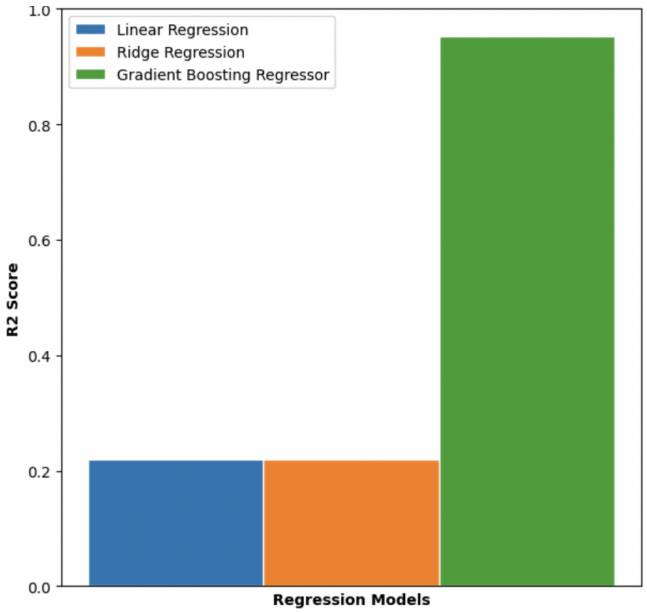


Figure 8: Histogram of MAE of Regression

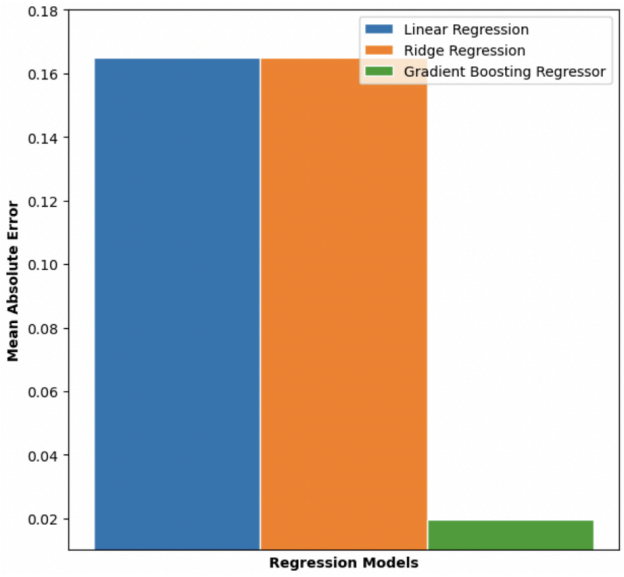
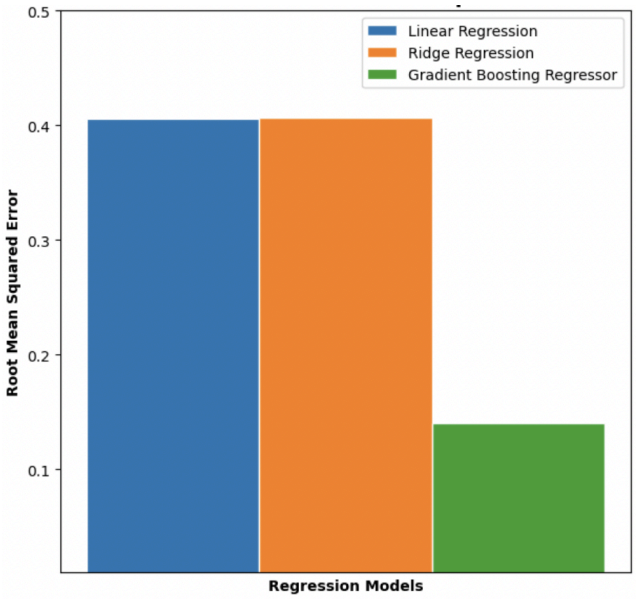


Figure 9: Histogram of RMSE of Regression



8. Conclusion

The exploration of credit card fraud detection has provided insights into addressing the issue of credit card security in the financial industry. Through overcoming the issue of class imbalance, data pre-processing techniques, employing unsupervised and supervised learning techniques, it has highlighted the underlying patterns within the data set, random forest and gradient boosting regression has showcased incredibly effective at determining fraudulency of a transaction.

Reference list

Chouinard, J.-C. (2023). *Classification In Machine Learning*. [online] JC Chouinard. Available at: https://www.jcchouinard.com/classification-in-machine-learning/.

Delua, J. (2021). *Supervised vs. Unsupervised Learning: What's the Difference?* [online] IBM Blog. Available at: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/.

Hardesty, L. (2017). *Explained: Neural networks*. [online] MIT News. Available at: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

IBM (2023a). *What is a Decision Tree | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/decision-trees.

IBM (2023b). *What is Unsupervised Learning? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/unsupervised-learning.

Kanade, V. (2022). *Logistic Regression: Equation, Assumptions, Types, and Best Practices*. [online] Spiceworks. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Practices%20for%202022-.

Kurama, V. (2019). *Regression in Machine Learning: What it is and Examples of Different Models*. [online] Built In. Available at: https://builtin.com/data-science/regression-machine-learning.

Vadapalli, P. (2020). *6 Types of Regression Models in Machine Learning You Should Know About*. [online] upGrad blog. Available at: https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/.

www.ibm.com. (2021). *k-means clustering*. [online] Available at: https://www.ibm.com/docs/en/db2woc?topic=procedures-k-means-clustering [Accessed 25 Mar. 2024].

Yeon, J. and Wilber, J. (n.d.). *Random Forest*. [online] mlu-explain.github.io. Available at: https://mlu-explain.github.io/random-forest/.