The Impact of Machine Learning on Credit Card Fraud Detection

Rafael González, Master Student
Polytechnic University of Puerto Rico
regc1024@gmail.com

Roberto Vivas, Undergraduate Student
Polytechnic University of Puerto Rico
robertovivas2@gmail.com

Edna M. Chaar, Doctoral Student
Polytechnic University of Puerto Rico
ednachaar@gmail.com

Dr. Alfredo Cruz, Ph.D.
Polytechnic University of Puerto Rico
alcruz@pupr.edu

Abstract

The global credit card fraud rate increased significantly in 2023. After COVID-19 surfaced in 2019, consumers chose credit cards to prevent touch and infection. This increase makes it more attractive to attackers, where there is also a growth in credit card fraud. In this research, we use two types of machine learning techniques to detect credit card fraud. The study attempted to increase the accuracy of credit card fraud detection, reduce incidents of fraud, and limit false positives by using Random Forest (RF) and Logistic Regression (LR) algorithms, therefore offering high precision in predictions. The two algorithms were compared using the Python programming language and a dataset called "Credit Card Fraud Detection Data" from Kaggel. The results showed that RF showed 99% accuracy and 88% ROC AUC, which we could determine was the most efficient in detecting fraud. These tests and analyses were executed using the Google Colab compiler.

Keywords: Machine Learning, Credit Card Fraud, Financial Security, Supervised Learning, Random Forest, Logistic Regression, SMOTE

Introduction

Ensuring the identification and prevention of fraud in financial transactions is of utmost importance, given that advanced fraudulent activities have exposed the shortcomings of conventional rule-based systems (Kazeem, 2023). Credit cards were declining before COVID-19 arrived in the US. The US had three times more credit cards than debit cards in 2019 (Statista, 2023). Analyzing massive transactional data for patterns and anomalies may help machine learning systems overcome these restrictions. These algorithms detect fraud tendencies and monitor real-time transactions, decreasing financial losses (Kazeem, 2023).

Before the rise of machine learning, credit card fraud detection relied on conventional methods like card security features and heuristic-based pattern analysis (Jendruszak, 2021; Go Digit General Insurance Limited, 2024). However, these approaches often resulted in high false positive rates and missed fraudulent transactions; compounding this challenge was the unique asymmetry in transaction behavior between fraud victims, who have a strong incentive to report unauthorized transactions promptly, and

fraudsters, who aim to extract as much value as possible swiftly (Jendruszak, 2021). As such, machine learning has revolutionized credit card fraud detection by enabling real-time analysis of vast datasets to identify overt and subtle signs of fraudulent activity (Jendruszak, 2021; Go Digit General Insurance Limited, 2024). Unlike rule-based systems, machine learning algorithms continuously learn from transaction data, adapting to new fraud tactics and improving detection accuracy over time (Jendraszak, 2021), moving financial institutions from reactive to proactive fraud detection (Go Digit General Insurance Limited, 2024). This research underscores the transformative impact of machine learning on fraud detection, providing a comprehensive comparison with conventional methods to highlight advancements and the potential for future innovations in the field.

This research examines the efficacy of machine learning methodologies in identifying instances of credit card fraud, explicitly comparing the performance of random forest and linear regression algorithms. The evaluation includes accuracy, precision, recall, F1 score, and ROC AUC metrics. The research includes a thorough investigation of relevant literature, the approach used, experimental procedures, findings, interpretation of conclusions, and suggestions for future research.

Credit Card Fraud: Traditional Methods and ML

Credit card fraud encompasses a range of illicit activities, including the unauthorized acquisition of card information to commit identity theft, Card-not-present (CNP) fraud involving transactions that do not require physical card verification, the misuse of lost or stolen cards, and the creation of counterfeit cards through the replication of magnetic stripe information for unauthorized purchases. Financial fraud is a significant issue that leads to significant losses and undermines public trust in the financial system. Machine Learning models are effective in identifying fraudulent transactions, unlike traditional methods like rule-based technologies and human verification. Machine learning algorithms can adapt to new fraud tendencies, reducing human monitoring and improving fraud prevention and detection efficiency (Kazeem, 2023).

Machine learning, including subdivisions like supervised, unsupervised, and reinforcement learning, excels at complex issues like image recognition and predictive analytics. Decision trees are used in credit card fraud detection to classify transactions as legal or fraudulent. This enables the identification of complex fraud patterns and improves the overall accuracy of the detection process (Kazeem, 2023).

Case Studies, Comparative Analysis, and Challenges

Capital One has implemented machine learning for credit card fraud detection, leveraging datasets including transaction details, cardholder information, and contextual data, thereby enhancing fraud identification accuracy. In contrast, Coinbase, a leading cryptocurrency exchange, uses machine learning for image analysis in online ID authentication, applying face-similarity algorithms for document verification, which boosts security and fraud detection (Lech, 2023).

As seen in Table 1, comparative investigations into machine learning for credit card fraud detection reveal diverse outcomes, titled Comparative Analysis of Previous Machine Learning Techniques. It can be considered that Bhanusri et al. (2020) study is better than Lipare (2023) due to its overall better results regarding Precision, Recall, F1 Score, and Accuracy measurements; however, it is believed that this is not the case. Bhanusri et al. (2020) use the Credit Card Fraud Detection dataset found in Kaggle (ULB, 2018), which has attributes Time, V1, V2, V3, …, V26, V27, V28, Amount, and Class. There is a lack of information about the attributes of V1, V2, V3, …, V26, V27, and V28, only that their values resulted from PCA transformations to protect customers' confidentiality. Due to the descriptions found in ULB (2018), Bhanusri et al. (2020) as well as other related projects found in Kaggle, it seems that values of attributes V1, V2, V3, …, V26, V27, and V28 were computed using the complete original dataset, which would lead to data leakage, also known as Train - Test Contamination or Bias (Géron, 2023; Lukita, 2023; Matharu, 2019; QHarr, 2017). This would explain why model evaluation results tend to be extremely high in the Training Phase, possibly showing Overfitting, and, if not, higher in the Testing Phase when using this dataset.

Table 1 Comparative Analysis of Previous Machine Learning Techniques

Study	Technique Name	Case	Precision	Recall	F1 Score	Accuracy
(Lipare, 2023)	Logistic Regression	Fraud	0.03	0.57	0.07	0.91
		No Fraud	1.00	0.91	0.95	
	Random Forest	Fraud	0.08	0.80	0.14	0.94
		No Fraud	1.00	0.94	0.97	
(Bhanusri et al., 2020)	Logistic Regression	Fraud	0.99	1.00	1.00	0.99
		No Fraud	1.00	0.99	0.99	
	Random Forest	Fraud	1.00	1.00	1.00	1.00
		No Fraud	1.00	1.00	1.00	

Research Methodology

The research analyzes genuine and fraudulent transactions using Logistic Regression (LR) and Random Forest (RF) machine learning models. LR is a simple benchmark, while RF can identify refined fraud patterns using its decision-making skills. Each model's credit card fraud detection effectiveness is assessed using precision, recall, F1 score, ROC AUC, and accuracy. This comparison examines the pros and cons of Logistic Regression (LR) and Random Forest (RF) in fraud detection.

Figure 1 depicts our study ML Model Fraud Detection Workflow for training and testing models. Once datasets are added, 'Initialization' starts. Data visualization, preparation, feature selection, and scaling comprise the 'Data Preprocessing step.' The dataset is 80% 'Training' and 20% 'Testing.' In the 'Training Model' step, Logistic Regression and Random Forest create two prediction models. Accuracy, precision, recall, F1 Score, and ROC-AUC evaluate model effectiveness. Using the testing set improved in 'Generate Final Model', the 'Testing Model' step extrapolates. The 'Results' output tags transactions as 'No Fraud' or 'Fraud.'

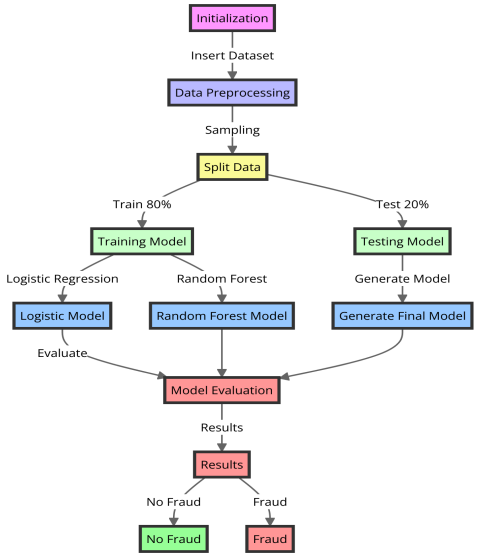

Figure 1 ML Model Fraud Detection Workflow

Gholamy et al. (2018) suggest splitting the data into training and testing sets to prevent overfitting. The training data is used to establish the model's parameters, while the testing data is used to assess the model's correctness. Research has shown that optimal outcomes are achieved by assigning 20-30% of the data for testing and the rest 70-80% for training. This division provides correct estimates that avoid overestimating the model's accuracy and provide the most precise valid estimates, thereby addressing the empirical observation mentioned in the research. Consequently, our research dataset utilizes an 80% (training) and 20% (testing) division.

During the data preprocessing stage, missing entries were addressed through mean and standard deviation-based imputation, deletion of affected rows and columns, and other standard techniques. Key statistical formulas, including the mean (μ) illustrated in Equation (1) and standard deviation (σ) shown in Equation (2), were utilized to summarize and understand the data, with the mean representing centrality and the standard deviation quantifying variability (Khan Academy, 2016). These statistical formulas are fundamental to data analysis.

The mean formula is given by:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

The standard deviation formula is given by:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2}$$

After data cleaning, data visualization included geographical and gender distributions, fraud and legitimate transaction percentages, unique values, major merchant categories, transaction amounts, and states with the most fraudulent transactions. The influence of correlations on feature selection was also examined. After appending the dataset, it was separated into training and testing sets (80% and 20% sample percentages). Encoders, transformers, resamplers, and feature scaling were chosen based on data classification, imbalance, scaling, or combination. A figure showed how cross-validation divided the dataset into ten subgroups for performance assessment using unknown data.

Following that, feature selection was achieved. Recursive Feature Elimination Cross Validation (RFECV) was utilized to obtain the best subset of features for the estimator by repeatedly deleting features and assessing the model's cross-validation score (The Sci-kit YB developers, n.d.). RFECV starts with model fitting for feature selection. After that, it eliminates the least significant attribute(s) until a specified number remains (The Sci-kit YB developers, n.d.). The model was then evaluated. This involved tweaking hyperparameters using a random search to choose random values within predetermined ranges (Brownlee, 2020). Finally, conclusions were reached. The models scored highly in F1 and accuracy.

According to Aslam and Hussain in 2024, logistic regression and random are considered 'state-of-the-art machine learning algorithms', further validating the algorithms used for the dataset present. It is also mentioned that they are advanced algorithms that are proven for their effectiveness. Based on the research (Rawat, 2022), we can determine that random forest is versatile in its application and capable of effectively addressing regression and classification problems. This fits perfectly within this juncture since credit card fraud detection, according to Iconiq Inc., n.d., presents a case of binary classification within supervised learning.

Experiments

The experimental technique of this work includes preprocessing, cross-validation, feature selection, and model assessment, as outlined in the Research Methodology section. The dataset information was first extracted using Python 3's Pandas Library, explicitly using its DataFrame data type. The dataset we use for the experiments contains 20,000 items with 23 columns or characteristics, 11 of which are numerical and the remaining categorical, as shown in Table 2. The collection includes data from all 50 states in the United States, collected between January 1, 2019, and December 31, 2020. The dataset comprises 693 distinct merchant firm names, 14 transaction types, two genders (9,029 men representing 45.145% and 10,971 females representing 54.855%), 105 fraud instances, and 19,895 non-fraud cases, accounting for 0.525% and 99.475% of the total, respectively. The transaction categories, listed in descending order of frequency, are: gas_transport, grocery_pos, home, shopping_pos, kids_pets, shopping_net, entertainment, personal_care, food_dining, health_fitness, misc_pos, misc_net, grocery_net, and travel. Categories containing the substrings _pos and _net indicate whether a transaction was conducted in person or online (Lipare, 2023).

Table 2 Attributes of Shbham Lipare's Credit Card Fraud Detection Data Dataset

Attributes of Shubham Lipare's Credit Card Fraud Detection Data Dataset		
Unnamed (Index)	gender (Sex of Card Holder)	job (Job of Card Holder)
trans_date_trans_time (Transaction Timestamp)	street (Transaction Address)	dob (Date of Bird of Card Holder)
cc_num (Credit Card Number)	City (Transaction City)	trans_num (Transaction Number)
merchant (Merchant Name)	state (Transaction State)	unix_time (Unix Time)
category (Transaction Category)	zip (Transaction Zip Code)	merch_lat (Merchant Latitude)
amt (Transaction Amount)	lat (Transaction Latitude)	merch_long (Merchant Longitude)
first (First Name of Card Holder)	long (Transaction Longitude)	is_fraud (Nature of Transaction)
last (Last Name of Card Holder)	city_pop (Population of City)	

As such, the **is_fraud** attribute is the label for studying supervised machine learning models, and the rest of the attributes are the features used to predict it. Moreover, it was also found that none of the column entries were incomplete; henceforth, data cleaning procedures, such as imputing data, were not required. The dataset is appropriate for doing transactional analysis, detecting fraud, and studying customer behavior. The non-null count guarantees feature amt (transaction amount), gender, and last (last name of card holder).

Figure 2, entitled 'Histogram of Some Features Before Feature Scaling', presents the distribution of selected variables from the dataset central to this study, right before being normalized but after the dataset was transformed and feature-selected. The figure comprises four individual histograms, each corresponding to a different variable: 'amt', 'year', 'gender', and 'last.' The 'amt' histogram, which represents the amount of money for each transaction, shows a unimodal distribution, indicating a prevalence of small transactions over larger ones, a common characteristic of financial data. The 'year' histogram is not depicted with actual data in the image provided; typically, it would outline the frequency of transactions across a specified time frame, in this case, as previously stated, from January 1, 2019, to December 31, 2020. The 'gender' histogram illustrates two distinct bars corresponding to the genders of

the male and female clients involved in the transactions. The final histogram, labeled 'last', should depict the distribution of customers' last names, although this is not a conventional type of data for a histogram and may be a mislabeling. Together, these histograms highlight the raw state of the data, underscoring the need for feature scaling. This preprocessing step is crucial in machine learning to ensure that all independent variables or features have a standardized range before the division of data into training and testing sets.

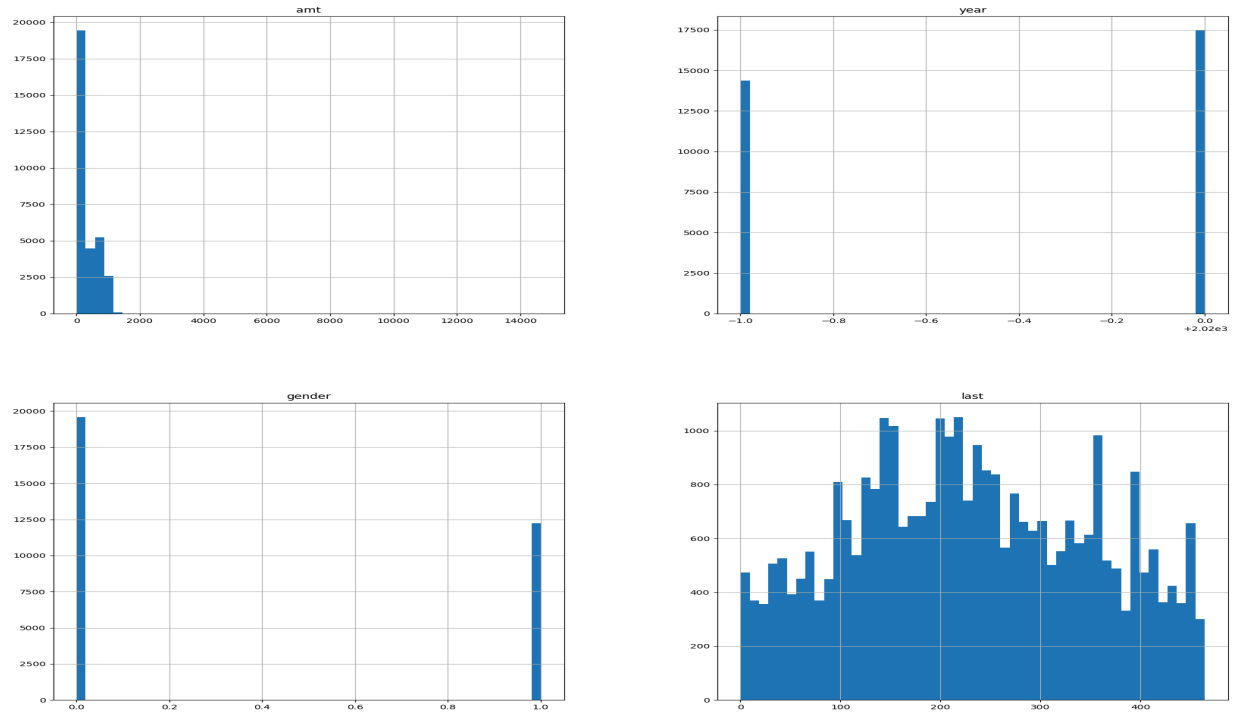

Figure 2 Histogram of Some Features Before Feature Scaling

Logistic Regression and Random Forest machine learning models were examined using K-fold cross-validation (K = 10) to choose which model to utilize with RFECV for feature selection. The Random Forest model was used for this stage. The characteristics found on the dataset columns are the following: cc_num, merchant, category, amt, first, last, street, city, state, zip, lat, long, city_pop, job, trans_num, unix_time, merch_lat, merch_long, month, hour, day, age, lat_dist, and long_dist.

Both machine learning models were fine-tuned using Scikit-Learn's Python 3 Randomized Search function and assessed for precision, recall, F1, ROC AUC, and accuracy. Both models used Logistic Regression: Solver (newton-cholesky), Penalty (l2), and C (Inverse Regularization Strength of 1.0) with a Random State of 42 and Random Forest: N Estimators (200), Max Features (sqrt), Max Depth (45), and Min Samples Split (5).

Experiment Results

Table 3 presents a comparison of the performance metrics of Logistic Regression and Random Forest algorithms in categorizing situations as "Fraud" or "No Fraud." Logistic Regression has stronger recall for fraud situations but poorer accuracy, resulting in more non-fraud transactions being misclassified as fraud. Logistic Regression has a much greater recall rate, suggesting its superior ability to detect all instances of fraud despite its tendency to incorrectly classify more non-fraud cases as fraud. The F1 score, including both false positives and false negatives, indicates poor performance in fraud detection. The ROC-AUC values are identical for both models, suggesting they have comparable performance in differentiating between fraud and non-fraud situations. Random Forest has superior performance in

overall accuracy compared to Logistic Regression, indicating it may be the better well-rounded model for this dataset.

Table 3 Performance Analysis for Two Different Algorithms

Technique Name	Case	Precision	Recall	F1 Score	ROC - AUC	Accuracy
Logistic Regression	Fraud	0.05	0.71	0.10	0.86	0.92
	No Fraud	1.00	0.92	0.96		
Random Forest	Fraud	0.13	0.17	0.15	0.88	0.99
	No Fraud	0.99	0.99	0.99		

Figure 3 displays the logistic regression model's performance using a confusion matrix with '0' (Not Fraud) and '1' (Fraud) classes. There were 3,677 real negatives, 239 false positives, and seven false negatives. Seventeen forecasts were correct, resulting in 17 accurate predictions. The ROC curve in Figure 4 shows a progressive increase in correctly detected positives as the number of false positives grows. The model's AUC is 0.86. The findings illuminate model performance.

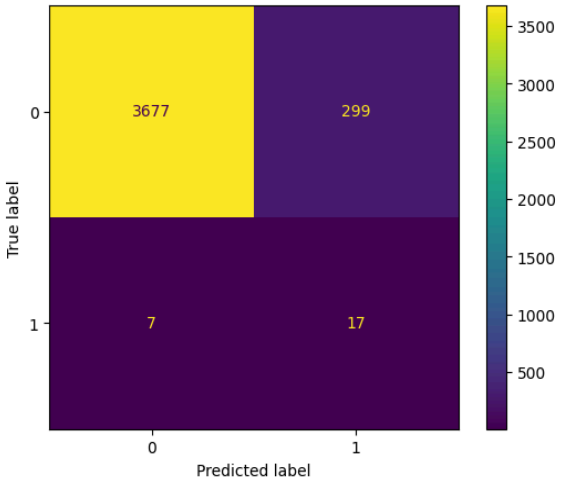

Figure 3 Confusion Matrix of Logistic Regression Results

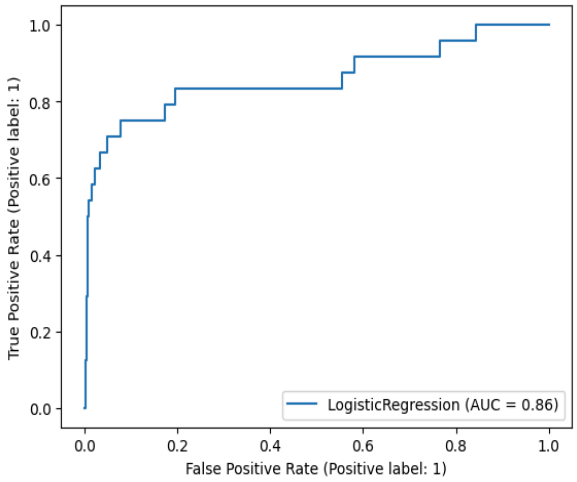

Figure 4 ROC Curve of Logistic Regression Results

Figures 5 and 6 show the Random Forest classifier's effectiveness. The confusion matrix shows model accuracy and precision. It shows that 3949 negative class occurrences were correctly predicted while 27 positive ones were not. Four positive class occurrences were correctly predicted, but 20 negative ones were not. The climbing ROC curve shows the model's diagnostic performance, with an AUC of 0.88, indicating high discrimination. These data would demonstrate the model's prognosis and positive/negative discrimination abilities in a study inquiry report. The model has numerous true negatives and an AUC around 1, indicating good performance, particularly in recognizing the negative class. It shows a good sensitivity-specificity trade-off.

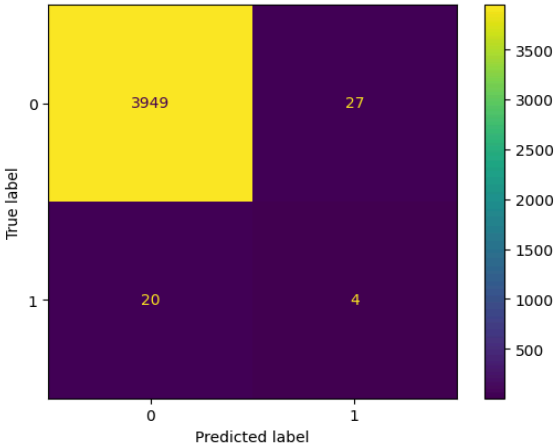

Figure 5 Confusion Matrix of Random Forest Results

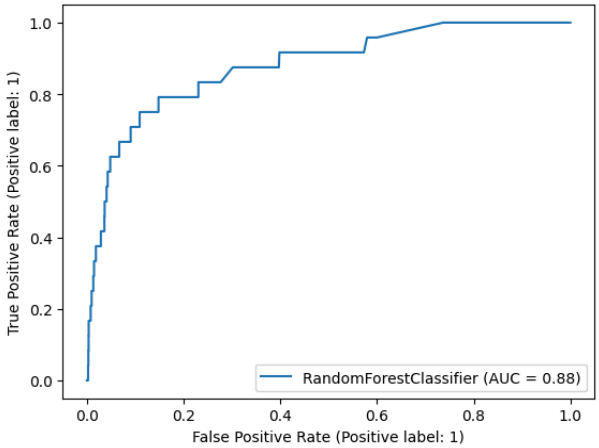

Figure 6 ROC Curve of Random Forest Results

Conclusion

This study uses machine learning model techniques to detect credit card fraud. The procedure known as Preprocessing was implemented to adequately visualize and analyze the data being used throughout this experiment while not negatively affecting the final performances of the models themselves. Processes such as Splitting, Encoding, Balancing, Feature Scaling, Feature Selection, and Hyperparameter Tuning positively impacted the final results of the models. That said, when using SMOTE to balance the Training Set, the newly generated synthetic data caused both the Logistic Regression and Random Forest models to overfit highly, affecting their final performances when faced with new data, i.e., the Testing Set. In essence, it was found that Random Forest, with its accuracy of 99% and ROC - AUC of 88%, was better than the Logistic Regression's 92% and 86% alternatives. In addition, the accuracy of both models performed better than those found in the Lipare (2023) study, which concluded with 91% for its linear regressor and 94% for its random forest. This was due to specific differences in the procedures used throughout this experiment, such as adding the Hyperparameter Tuning process and using RFCEV for Feature Selection.

Fraud detection studies may also benefit from other machine-learning techniques. Exploring alternative fraud detection algorithms may identify strengths and limitations, leading to more robust and complete solutions. Comparing and combining fraud detection approaches may progress research and provide more effective and sophisticated solutions to fight fraud. These recommended future study fields aim to improve fraud detection and prevention using powerful machine-learning algorithms.

Future Work

Future work should focus on conducting additional research to improve the performance of these two models and others for detecting credit card fraud. This can be achieved by exploring the utilization of more hyper-parameters, alternative encoding techniques, and various balancing techniques such as SVM Smote, K-Means SMOTE, and Borderline SMOTE, among others. In addition, research can also be done for Online Learning systems, such as Online Random Forest and Online Deep Neural Networks, to incrementally and constantly train machine learning models with real-time data due to the infeasibility to constantly train over entire datasets. It should be noted that Online Learning systems are sensitive to the quality of the sequential (new) data (ActiveLoop, 2024; Géron, 2019).

Acknowledgment

This material is based upon work supported by, or in part by, the National Science Foundation (NSF-SFS) under contract/award 2140638

References

1000logos.net. (2022, May 29). *Coinbase Logo and Symbol, Meaning, History, PNG, Brand*. 1000 Logos. https://1000logos.net/coinbase-logo/

ActiveLoop. (2024). *What Is Online Random Forest*. https://www.activeloop.ai/resources/glossary/online-random-forest/

Aslam, A., & Hussain, A. (2024). A Performance Analysis of Machine Learning Techniques for Credit Card Fraud Detection. *Journal on Artificial Intelligence*, *6*(1). https://doi.org/10.32604/jai.2024.047226

Bhanusri, A., Sree Valli, K. R., Jyothi, P., Sai, G. V., & Sai Subash, R. R. (2020). Credit Card Fraud Detection Using Machine Learning Algorithms. *Journal of Research in Humanities and Social Science*, *8*(2), 04-11.

Brownlee, J. (2020, September 13). *Hyperparameter Optimization With Random Search and Grid Search*. Machine Learning Mastery. https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media, Inc.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*. Departmental Technical Reports (CS). 1209. https://scholarworks.utep.edu/cs_techrep/1209

Go Digit General Insurance Limited. (2024, March 5). *Credit Card Fraud Detection: Best Ways & Protection*. Digit Insurance. https://www.godigit.com/finance/identity-theft-and-fraud/detect-credit-card-fraud

Hari, S. (2021, May 19). *Cross Validation and Types*. Nerd for Tech. https://medium.com/nerd-for-tech/cross-validation-and-types-a7498a68f413

Iconiq Inc. (2022, August 17). *Big Data and Data Science Projects - Learn by building apps*. ProjectPro. Retrieved January 24, 2024, from https://www.projectpro.io/project-use-case/credit%20card%20fraud%20detection%20classification%20problem#:~:text=The%20goal%20is%20to%20detect

Jendruszak, B. (2021, December 17). *Credit Card Fraud Detection: The Complete Guide*. SEON. https://seon.io/resources/credit-card-fraud-detection/

Kazeem, O. (2023). *Fraud Detection Using Machine Learning*. https://doi.org/10.13140/RG.2.2.12616.29441

Khan Academy. (2016). *Calculating Standard Deviation Step by Step*. Khan Academy. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/variance-standard-deviation-population/a/calculating-standard-deviation-step-by-step

Lech, E. (2023, November 3). *Machine Learning for Fraud Detection in Fintech*. Pragmatic Coders. https://www.pragmaticcoders.com/blog/machine-learning-for-fraud-detection-in-fintech#:~:text=Capital%20One%20employs%20machine%20learning

Lipare, S. (2023, May 5). *Credit Card Fraud Detection Data*. Kaggle. https://www.kaggle.com/datasets/shubhamlipare/credit-card-fraud-detection-data/data

Lipare, S. (2023, May 5). *Credit Card Fraud Detection Supervised Machine Learning*. Kaggle. https://www.kaggle.com/code/shubhamlipare/credit-card-fraud-detection-supervised-ml/notebook

Logos-World.net. (2024, February 12). *Capital One Logo, Symbol, Meaning, History, PNG, Brand*. Logos-World. https://logos-world.net/capital-one-logo/

Lukita, A. (2023, May 19). *The Dreaded Antagonist: Data Leakage in Machine Learning*. Medium. https://towardsdatascience.com/the-dreaded-antagonist-data-leakage-in-machine-learning-5f08679852cc

Matharu, G. S. (2019, December 15). *Data Leakage in Machine Learning*. Medium.

https://medium.com/@gurupratap.matharu/data-leakage-in-machine-learning-390d560f0969

Nondeterministic Memes for NP Complete Teens. (2018, November 5). *Nondeterministic Memes for NP Complete Teens | Facebook*. Facebook. https://www.facebook.com/npcompleteteens/posts/creds-to-debojeet-chatterjee/488897178271528/

Rawat, T. (2022). Machine Learning For Credit Card Fraud Detection System. *International Journal of Research in Engineering and Science*, *10*(5), 08-14. https://doi.org/10.13140/RG.2.2.17480.60163

Statista. (2023, January 31). *Total Number of Credit Cards and Debit Cards in Circulation in the United States From 2012 to 2020 [Graph]*. In Statista. Retrieved January 24, 2024, from https://ezproxy.pupr.edu:2087/statistics/245385/number-of-credit-cards-by-credit-card-type-in-the-united-states/

The Scikit Yb Developers. (2022, August 21). *Recursive Feature Elimination — Yellowbrick v1.1 Documentation*. Scikit Yb. Retrieved January 25, 2024, from https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html#:~:text=Recursive%20feature%20elimination%20(RFE)%20is