Employing Machine Learning for Fraudulence Detection across Credit Card Transactions

Nick Austin, Wes Bovee, Brayden Christensen, Cole Edgren, Spencer Jorgensen
CS 201R, Winter 2024
Department of Computer Science
Brigham Young University

Abstract

Recent efforts to combat the ongoing surge in financial fraud have included the harnessment of machine learning techniques to construct anomaly detection models for various forms of payment. In this paper, we analyze the utility of machine learning algorithms in classifying the legitimacy of credit card transactions. Data was collected from Kaggle.com and subsequently preprocessed for training. We selected features from the original dataset and engineered new features to incorporate more relevant variables into the data. Novel machine learning models, including Decision Tree, Logistic Regression, and XGBoost were implemented, trained, and evaluated against an alloted test portion of the final prepared dataset on their ability to correctly mark a credit card transaction as legitimate or fraudulent. Initial results demonstrated excellent performance ($>$ 99% classification accuracy) with tree-based models. The high proficiency of these models in detecting anomalies across hundreds of thousands of transactions emphasizes the potential of machine learning for effective fraud prevention.

1 Introduction

Financial fraud is an evolving identity crime that affects millions worldwide annually. In particular, the advent of convenient and accessible electronic payment systems such as those linked through mobile devices, personal accounts, or software applications carries a heightened risk for financial security breaches. These emerging technologies combined with the substantial amount of transactions performed each day warrants a need for autonomous monitoring that extends beyond the practical scope of human capability. With modern advancements in data science, automatic early threat detection poses as a solution to mitigating the consequences of compromised financial information. Corporations have recently begun to implement early alert systems by training anomaly detection machine learning models on payment transaction data. Depending on the type of transaction, certain indicators may be important predictors in flagging a fraudulent payment, such as the amount of money transacted, the time of the transaction, or whether or not a card chip was used to complete the payment.

As with many machine learning experiments, data quality, characteristics, and amount is paramount to capturing the subtle patterns and relationships necessary to make accurate predictions. Unfortunately, in our search for appropriate data, we were faced with limitations from publicly available datasets due to customer privacy concerns. Despite this bottleneck, we postulated that general trends for financial fraud could still be replicated within available datasets and help formulate valuable and interpretable insights from prediction results. Throughout this paper, we collect and prepare transaction data, implement a variety of machine learning models, and examine classification ability through multiple metric assessments.

2 Methods

2.1 Data Collection

Due to the confidential nature of consumer financial information, directly obtaining nonsynthetic data of adequate quality and scope proved difficult; in fact, we found that acquiring such data with interpretable features would not be possible with current legal restrictions. Eventually, we were able to locate two datasets from Kaggle.com, one which was synthteic, and the other containing genuine anonymized credit card transactions. Both datasets contained labeled instances that were marked as fraudulent by a binary value of 1 and 0 otherwise.

2.2 Dataset 1

The initial dataset, which we will refer to as Dataset 1, contained precisely 1,000,000 unique transactions characterized by several features as outlined in Table 1. Continuous features were standardized using standard score. Upon further inspection of the dataset, an immediate concern arose from the highly imbalanced class distribution: less than 1/10 instances were classified as fraudulent, while the majority of the other instances were legitimate. Prior to addressing the class distribution of the data, we decided to retain a copy of this imbalanced data for experimentation with machine learning models suited for anomaly detection to gather preliminary insights.

In addition to the original imbalanced data for the anomaly detection models, we further prepared balanced data for more general use algorithms. In an effort to combat potential bias that could favor patterns across the legitimate transactions, we oversampled the minority fraudulent class using ADASYN (Adaptive Synthetic Sampling). Our choice of ADASYN instead of conventional SMOTE (Synthetic Minority Oversampling) is based on ADASYN's algorithmic design of generating synthetic minority samples through K-Nearest Neighbors that are closer to the majority samples and ultimately more difficult to learn. By this approach, we hypothesized subtle relationships in the minority class had a better chance of being captured by our models.

Feature	Data Type
distance_from_home	continuous
distance_from_last_transaction	continuous
ratio_to_median_purchase_price	continuous
repeat_retailer	boolean
used_chip	boolean
used_pin_number	boolean
online_order	boolean

Table 1: Dataset 1 Features

2.3 Dataset 2

In an effort to diversify our findings, we also trained models on a dataset that contains anonymized credit card transactions of European cardholders from 2013, which we will refer to as Dataset 2. It is the only publicly available dataset of our knowledge that contains actual, non synthetic transaction instances. As mentioned previously, customer privacy regulations necessitate a need for anonymity. Following this, the data was normalized and transformed by performing Principal Component Analysis to the original raw data, creating principal components V1-V28, which are the accompanied by only the following features: `Time`: time since transaction in seconds, `Amount`: amount of transaction in EUR, and the class label. Despite the clear drawback of being unable to interpret literal meanings behind various features, we understood that important relationships within the data are still retained. This indicated a promising opportunity to train on a more realistic representation of fraudulent behaviour.

As we prepared this dataset, we were met with similar difficulties as with our initial dataset, albeit much more significant: roughly 500 out of the 284,807 transactions were marked fraudulent. Employing the ADASYN technique once again, we corrected the class distribution ratio to 1:1.

2.4 Model Selection

We implemented models from the scikit-learn library to train on both datasets. Initially, an Isolation Forest model was applied to address imbalanced datasets, followed by Decision Tree classifiers, Multi-layer Perceptron, Logistic Regression, Random Forests, and XGBoost classifiers for balanced data. Our rationale for selecting these models was based on the interpretability of Decision Trees and Logistic Regression, the capability of MLP to discover more nuanced variations within the data, and with the proven performance of XGBoost and Random Forest as a dependable ensemble methods. All data was organized into training and testing splits of 80% and 20%, respectively.

3 Initial Results

3.1 Dataset 1

As previously mentioned, in light of the disproportionately large number of legitimate transactions compared to fraudulent ones in credit card fraud detection, we opted to address this challenge by employing an Isolation Forest algorithm. Isolation Forest operates by assigning anomaly scores to every instance; these scores fall within the range of [-1, 1], where values closer to -1 signify anomalies and those closer to 1 represent normal instances. Initial experimentation with our model yielded an area under the receiver operating characteristic curve (AUC) of 0.75.

Seeking to enhance our model's performance, we decided to test the model's performance on the balanced version Dataset 1. We were surprised to find that the resulting AUC was 0.60. Upon further reflection, we found that this outcome aligns with expectations, as the augmentation of data via ADASYN eliminated outliers, thereby reducing the model's ability to discern anomalous instances.

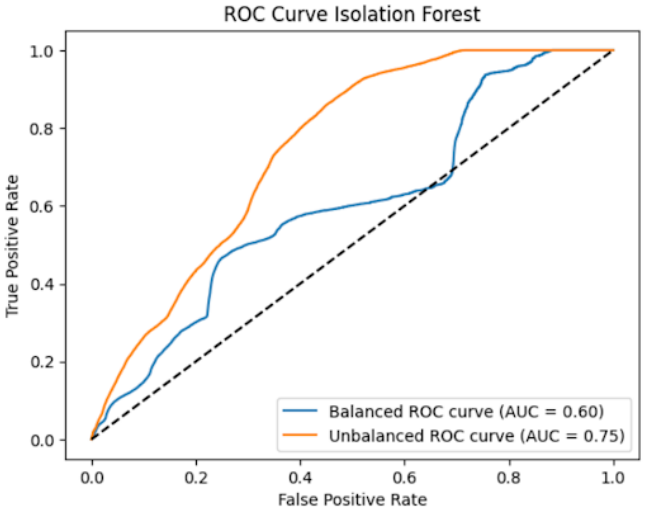

Figure 1: ROC curves for Isolation Forest

Following these findings, we proceeded to train various models on the balanced version of Dataset 1. Table 2 demonstrates various performance metrics for the chosen models.

Results from all other chosen models (Decision Tree, MLP, Logistic Regression, XGBoost) demonstrated excellent classification performance across many metrics, including accuracy, precision, F1 score, recall, and AUC. Skeptical of these ideal results, we sought to infer different feature interactions and importances through the usage of the Logistic Regression model's coefficients. Using the coefficients, we could

Model	Accuracy	Precision	F1	Recall
LR*	0.93	0.92	0.93	0.95
Decision Tree	0.99	0.99	0.99	0.99
XGBoost	0.99	0.99	0.99	0.99

Table 2: Dataset 1 Evaluation Metrics (*LR denotes Logistic Regression)

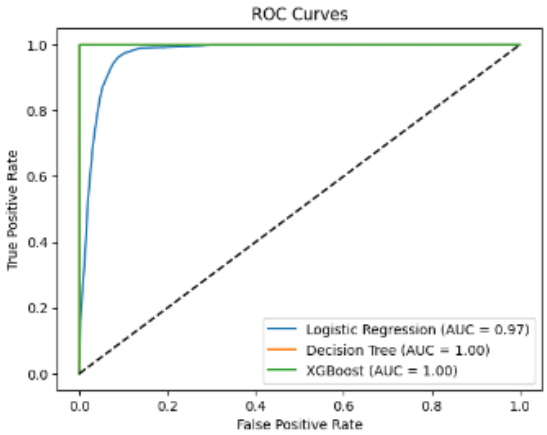

Figure 2: ROC curves for other models

interpret how each feature impacts the log-odds of a classification being legitimate or fraudulent. Regression coefficients for each feature are outlined in Table 3.

Feature	Coefficient
distance_from_home	1.89
distance_from_last_transaction	1.78
ratio_to_median_purchase_price	2.99
repeat_retailer	-0.97
used_chip	-0.73
used_pin_number	-4.38
online_order	3.49

Table 3: Features and Regression Coefficients

Intuitively inferring these coefficients, we were able to make the following observations:

- Holding all else constant, for every unit increase in `ratio_to_median_purchase_price` we expect an increase of **2.99** in likelihood that the purchase was **fraudulent**.
- Holding all else constant, when the purchase was made with a `repeat_retailer` we expect an increase of **0.97** in likelihood that the purchase was **legitimate**.
- Holding all else constant, when the purchase was made with a `used_pin_number` we expect an increase of **4.38** in likelihood that the purchase was **legitimate**.
- Holding all else constant, when the purchase was made with a `online_order` we expect an increase of **3.34** in likelihood that the purchase was **fraudulent**.

3.2 Dataset 2

Accounting for the results using the balanced version of Dataset 1, we proceeded straightway using the balanced version of Dataset 2 on Decision Tree, Logistic Regression, MLP, and XGBoost models. Performance metrics remained substantially high across all models despite using entirely new data (see Table 4 and Figure 3).

Model	Accuracy	Precision	F1	Recall
LR	0.96	0.98	0.96	0.95
MLP	0.99	0.99	0.99	1.0
Decision Tree	0.99	0.99	0.99	0.99
XGBoost	0.99	0.99	0.99	1.0

Table 4: Dataset 2 Evaluation Metrics

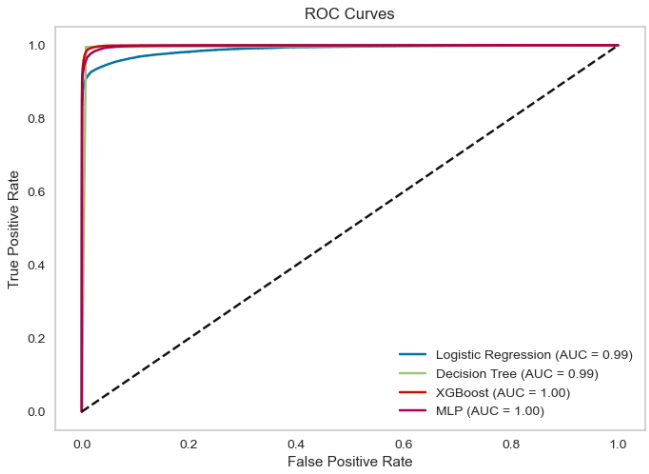

Figure 3: ROC curves for other models

Evaluating feature importance among the principal components in these results, we decided to use the XGBoost 'Gain' metric which is given by the following formula:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

Essentially, each term within the brackets represent the score of splitting on a node, or how well the node achieves a reduction in loss, minus the regularization parameter. We found that one particular feature significantly maximized 'Gain' over all other features, which was `V14` (see Figure 4).

4 Data Improvements and Final Results

The exceptional results we obtained through nearly all methods in both datasets was unprecedented and instigated a pivot from seeking to improve performance metrics. We instead directed our efforts towards optimization through feature engineering and feature space adjustments. We hypothesized

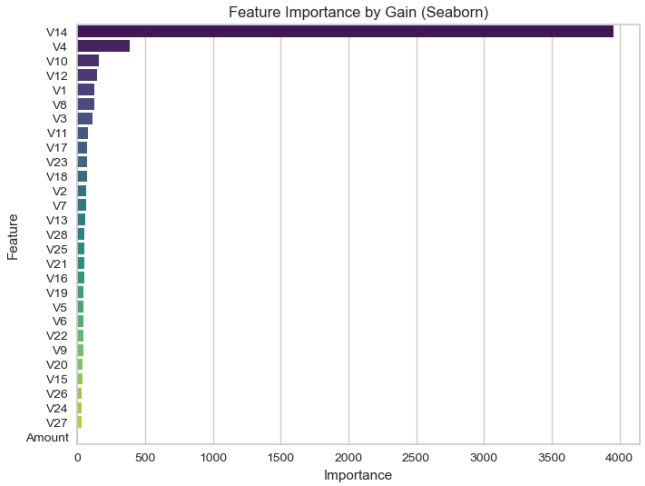

Figure 4: Feature importance by Gain in Dataset 2

that with a simplification of our given feature space, computational efficiency would be increased, and model generalizability would theoretically improve.

4.1 Dataset 1 Feature Engineering

Abstracting the interpretable meanings of the original features in Dataset 1, 3 new features were created, as shown in Table 5.

Feature	Data Type
price_online	continuous
chip_and_pin	boolean
total_distance	continuous

Table 5: New features for Dataset 1

`price_online` is derived through multiplying features `ratio_to_median_purchase_price` and `online_order`. The intuition behind this combination places emphasis on abnormally high or low purchases when made online. `chip_and_pin` is calculated by the negative product of the two boolean features `used_chip` and `used_pin_number`. Essentially, if a chip and pin are used within a transaction, this feature provides more weight to the probability a transaction is legitimate. `total_distance` is measured by taking the sums of the individual logarithmic values of `distance_from_home` and `distance_from_last_transaction`. We speculated that if a distance is within a threshold of a distance away from home or another origin, that feature weight should remain relatively consistent. For instance, if a purchase is made 100 miles away, another purchase made 500 miles away shouldn't increase the probability the transaction is fraud by a direct factor of 5.

The base features from which these new features were calculated were dropped from the dataset. Performance metrics were then evaluated on previous models (see Table 6). We noted that, despite a retention in performance across metrics of accuracy, precision. F1 score, and recall, scores were marginally deducted by roughly 3%. Further, we compared a confusion matrix of our predictions from our transformed data to our previous data (see Figures 5, 6).

Model	Accuracy	Precision	F1	Recall
LR	0.91	0.91	0.91	0.91
Decision Tree	0.94	0.95	0.94	0.95
XGBoost	0.96	0.94	0.96	0.98

Table 6: Transformed Dataset 1 Evaluation Metrics

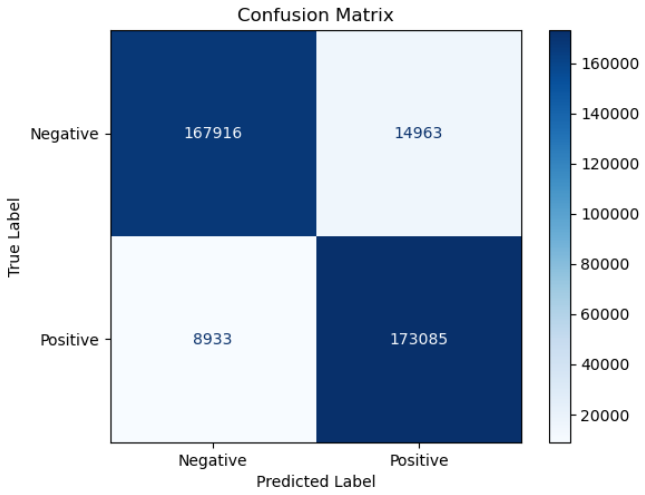

Figure 5: Confusion Matrix for original Dataset 1

As seen by comparing the two confusion matrices, the number of false negatives nearly doubled after we made these modifications to Dataset 1. We believe this is likely attributed to one or more of our engineered features losing original information that better discretized between certain transactions that were fraudulent.

4.2 Dataset 2 Feature Reduction

After analyzing feature importances through 'Gain' as explored earlier, we decided to drop all features except the top 5 ranked in importance by 'Gain', which were features `V14, V4, V10, V12, V1`. After running previous models on the new feature space, we received the following results contained in Table 7.

Model	Accuracy	Precision	F1	Recall
LR	0.95	0.95	0.95	0.93
MLP	0.98	0.98	0.98	0.97
Decision Tree	0.99	0.99	0.99	0.99
XGBoost	0.98	0.99	0.99	0.99

Table 7: Transformed Dataset 2 Evaluation Metrics

As evidenced by the our results, reducing the dimension of the original feature space from 28 principal components

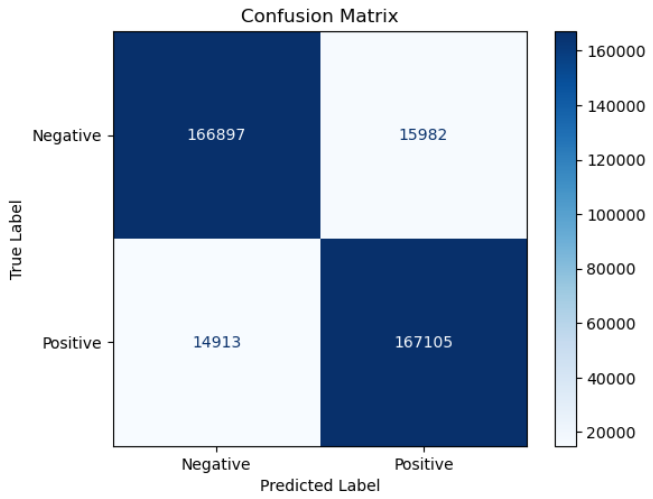

Figure 6: Confusion Matrix for transformed Dataset 1

to 5 preserved nearly the totality of our models' predictive ability. Though the exact meanings behind these important 5 features is only known to the owners of the original dataset, we can clearly see that they effectively capture critical aspects of the patterns between fraudulent and legitimate credit card activity.

5 Conclusion and Discussion

Out of all of our attempted methods, the models trained on the synthetically oversampled version of Dataset 2 with original unaltered features performed best according to metrics of accuracy, precision, F1 score, recall, and AUC. It should be noted that these results are rather trivial when compared to metrics for the rest of our trials; none of our model outcomes resulted in a score of less than 90% for any of the aforementioned metrics. As we sought to progressively optimize training, creating more relevant features and discarding those less relevant, the overall retention of performance is still quite substantial. For instance, the 5 selected principal components suffered metric score decreases of roughly .1%.

Despite our optimal findings, we remain highly skeptical of how generalizable our models may be to new, unseen data. Indeed, we recognize that there is a likelihood our seemingly ideal results are the product of overfitting to the newly distributed synthetic class. Our machine learning algorithms may have learned the limited variations contained within our balanced datasets, and, if tested on genuine credit card transactions, may fail to successfully differentiate between the overwhelming majority of legitimate transactions and the few and far in between fraudulent instances. Returning to our initial limitations as we approached this novel challenge, we conclude that large collections of rich, informative data from authentic credit card transactions would provide a much better starting point to tackling the global issue of financial fraud.

6 Future Considerations

Though it remains difficult to ascertain the full practicality of our machine learning models for building fraud detection systems, the results contained in this paper demonstrate highly promising initial outcomes that indicate the potential of machine learning to effectively classify large volumes of financial data.

As corporations are faced with an exponentially increasing number of transactions performed each day, we speculate that at such a scale and pace, sequential models may be of more utility in assessing dynamic data states, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory Networks (LSTMs). This is especially important under the consideration of the evolving electronic payment sphere, where the way in which fraudulent behavior is manifested is much more unique.

Additionally, the feasibility of extensive data preprocessing must be considered for high volumes of transactional data. For instance, oversampling the minority fraudulent instances may be computationally expensive and impractical for many organizations. Instead, incorporating algorithms designed to predict with minimal data preparation, such as the Isolation Forest we experimented with earlier, may be a better option. Moreover, creating pipelines where data is channeled and evaluated against diverse model types may provide a better overall summary of data characteristics. For instance, comparing results from an Isolation Forest in conjunction with an RNN may be effective.

From our experimentation and considerations, we believe that, with the correct resources and approach, the power to create powerful, reliable fraud detection systems through machine learning is well within reach.