Predicting Fraud in Credit Card Transactions

Christopher Piacesi

ITCS 5156-051 Project
Spring 2024 Semester

cp1aces1/cpiacesi-5156-project (github.com)

Chung, Jiwon, and Kyung-Ho Lee. “Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression.” *Sensors*, vol. 23, no. 18, Sept. 2023

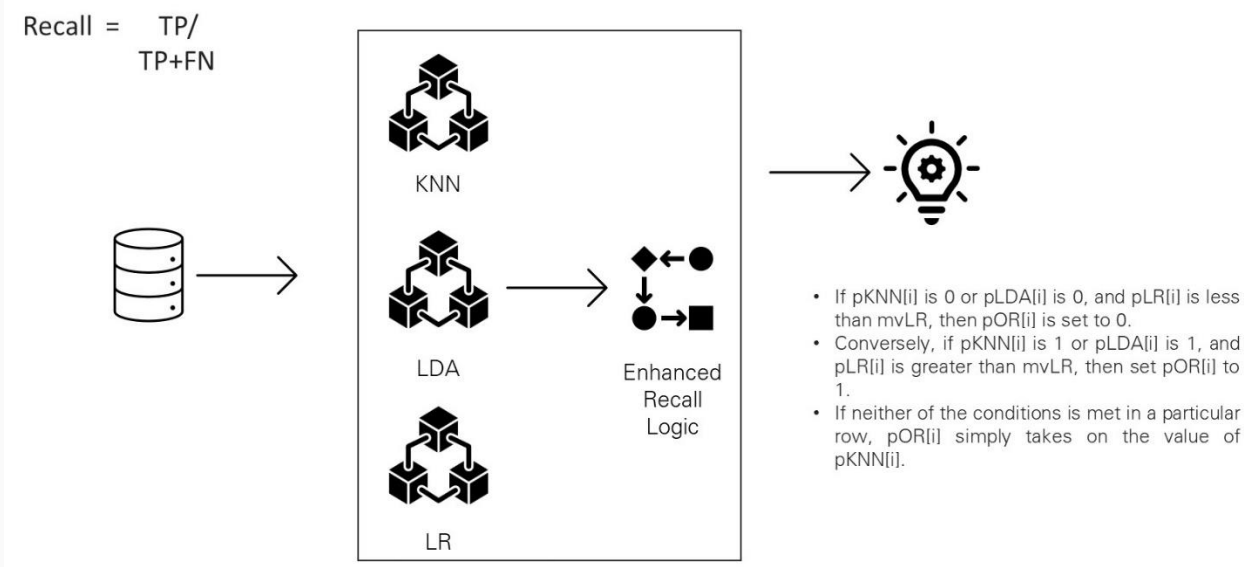
1 Introduction

Credit card transaction fraud is a prevalent problem in society. Payment card fraud losses worldwide exceeded $32 billion in 2021, of which nearly $12 billion was in the US. Losses to fraud worldwide increased by 14% in 2021. Higher fraud losses in the US were attributable to a 25% increase in purchases made by credit cards after a 9% drop in 2020. Also impacting fraud in the US was the continued growth in card-not-present transactions such as those that occur when spending online. Online purchases leave merchants more vulnerable to fraud[4]. Most credit card holders have been subjected to fraudulent charges including 65 percent of credit and credit card holders at some point in their lives. This equates to about 151 million Americans[5].

While increasingly powerful Machine Learning and Artificial Intelligence tools are at our disposal, the numbers indicate the problem is still pervasive. The ability to fully secure, detect, and prevent credit card fraud remains unresolved. As a technologist in the banking industry, the challenges of detecting credit card fraud represented an interesting topic to explore. The objective: apply existing Machine Learning models and tools to the problem and evaluate how well they can identify fraudulent transactions.

The approach taken with this project is to evaluate several Machine Learning models with a primary focus on the supervised learning strategy to enhance recall. The primary paper used as the inspiration for this project suggests, that recall holds significant importance in fraud detection[1]. Providing a highly accurate detection approach that limits misclassifying fraudulent transactions significantly lowers the industry's losses.

The methodology inputs the prediction results of K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Linear Regression (LR) models into conditional logic to determine the final prediction. Because Linear Regression does not provide a discrete classification the algorithm uses a binning approach by comparing each prediction with the mean value across all predictions.



2 Related Works

The literature and research on credit card fraud detection are voluminous and this survey just scratched the surface. One goal of this work was to identify multiple approaches to identifying fraud that used the same dataset to get comparable results validation across the proposed solutions. There is no shortage of recommended solutions in the literature, but this survey represents an introduction to evaluating some of them.

In Afriyie et al. 2023, the researchers compared the results of applying Decision Tree, Logistic Classification, and Random Forest models [2]. The team used the Synthetic Minority Oversampling Technique (SMOTE) to balance the Credit Card Transactions Fraud Detection Dataset [7] before validating each approach. Their results led to Random Forest as the best-performing model of this set.

In Xia 2022, the research included an approach applying a Support Vector Machine using the same dataset [3]. With this approach, the researcher did not balance the dataset and achieved modest results. Their conclusions pointed to future research with this model using SMOTE to enhance the results. Xia was correct in their conjecture as improved performance was seen in the experiments completed for this work.

3 Method

This section describes the Machine Learning methods implemented based on guidance across all three papers surveyed. Pre-published code was not sought out for this work. The approach included replicating the preprocessing and model development using described techniques and hyperparameters for independent validation of model performance. Additional experimentation with different hyperparameters provided from a grid search was conducted to attempt better performance. The project GitHub contains all the code independently developed for this work.

3.1 Data

Credit card transaction data is confidential as it includes personal data. While obtaining raw real-world data in the public domain for experimentation is impossible, there are some workable options. The IEEE-CIS Fraud Detection dataset[6] is based on real-world transactions from Vesta Corporation's e-commerce platform. The data is mostly encoded with generic feature names to protect confidential and proprietary information. The positive aspect is working with realistic data but with the downside of not getting the transparency that would be useful in feature engineering. Other options include a variety of synthetic data sets. The Credit Card Transaction Fraud Dataset [7] is one popular option

used across the papers surveyed in this project. This dataset was used across all the models as a baseline to compare the documented results from the prior research with the results of this work. Both datasets include a fraud classification label for each transaction, making them applicable to supervised learning techniques.

3.1.1 IEEE-CIS Fraud Detection [6]

This data set includes 590,540 Rows and 394 columns from Vesta Corporation's e-commerce platform transactions.

Here is a summary of the features included in the dataset:

- TransactionDT: timedelta from a given reference datetime (not an actual timestamp)
- TransactionAMT: transaction payment amount in USD
- ProductCD: product code, the product for each transaction
- card1 - card6: payment card information, such as card type, card category, issuing bank, country, etc.
- addr: address
- dist: distance
- P_ and (R__) email domain: purchaser and recipient email domain
- C1-C14: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
- D1-D15: time delta, such as days between previous transaction, etc.
- M1-M9: match, such as names on card and address, etc.
- Vxxx: Vesta engineered rich features, including ranking, counting, and other entity relations.

3.1.2 Credit Card Transactions Fraud Detection Dataset [7]

This dataset includes a file with 1,296,675 rows and 23 columns of training data along with a test data file that includes 555,719 rows with the same 23 columns.

Features:

trans_date_trans_time	The date and time of the transaction
cc_num	Credit Card Number used
merchant	The merchant name for the transaction
category	The product category of the purchase

amt	The amount of the purchase
first	First name of the card holder
last	Last name of the card holder
Gender	Gender of the card holder
Street	Credit card holder address
city	
state	
zip	
lat	Latitude and Longitude coordinates for the credit card holder's address
long	
city_pop	City population of the credit card holder
job	Credit card holder's occupation
dob	Credit card holder date of birth
trans_num	The transaction number
unix_time	The column name implies a Unix time data element. Unknown what this represents as it does not correspond to either the transaction date or the cardholder's date of birth. The month and day approximately correspond with the transaction date, but the year is off.
merch_lat	Merchant address latitude/longitude coordinates.
merch_long	

3.2 Preprocessing

3.2.1 IEEE-CIS Fraud Detection Dataset

The IEE-CIS Fraud Detection dataset included 394 columns of data. One column was the Transaction ID and one was the fraud classification label, leaving 392 potential features for model training validation. The dataset included high occurrences of missing values across many of the features. The first step in preprocessing was dimensionality reduction to eliminate the features that contained the top 25th percentile of missing values. This left 199 features for training the model. The remaining missing values were imputed to the mean value for the feature. The categorical data was label encoded and all features were scaled to the same range using the MinMaxScaler. This dataset only included one usable data file requiring an 80%/20% split to create training and testing data. For the SVM model, only 10% of the training data was used as done in [3] to obtain reasonable runtimes.

3.2.2 Credit Card Transactions Fraud Detection Dataset

The Credit Card Transactions Fraud Detection Dataset included 21 Features with no missing values. The data preprocessing only required label encoding the 12 categorical features and scaling all features using the MinMaxScaler. This data set included a second testing file that was similarly preprocessed and used for validating the trained models.

3.2.3 Imbalanced Data

Fraud data is naturally unbalanced with the not fraud classification dominating the dataset. Both datasets, as expected, exhibited this characteristic.

IEE - CIS Fraud Detection Data Set

Credit Card Fraud Transactions Detection Dataset

To prevent this unbalance from biasing the model to the not fraud classification both undersampling of the majority class and Synthetic Minority Oversampling Technique (SMOTE) were applied to the training data in various runs to provide the model with balanced data.

3.3 Models

3.3.1 Enhanced Recall

The Enhanced Recall model uses ensemble learning by applying conditional logic to predictions from KNN, Linear Discriminant Analysis, and Linear Regression to get the final result. Because Linear Regression does not provide a discrete classification the algorithm uses a binning approach by comparing each prediction with the mean value across all predictions. A Linear Regression value below the mean equates to a negative classification and above the mean equates to a positive classification. The conditional logic, in summary, first checks if either KNN or LDA indicates the negative class for fraud. If LR is also negative, the final prediction is set to negative. The logic then checks if KNN or LDA indicates the positive class for fraud. Again if LR agrees it sets the final prediction to fraud. If neither of those conditions exists, the model defaults the final prediction to the outcome predicted by the KNN model.

- If pKNN[i] is 0 or pLDA[i] is 0, and pLR[i] is less than mvLR, then pOR[i] is set to 0.
- Conversely, if pKNN[i] is 1 or pLDA[i] is 1, and pLR[i] is greater than mvLR, then set pOR[i] to 1.

- If neither of the conditions is met in a particular row, pOR[i] simply takes on the value of pKNN[i].

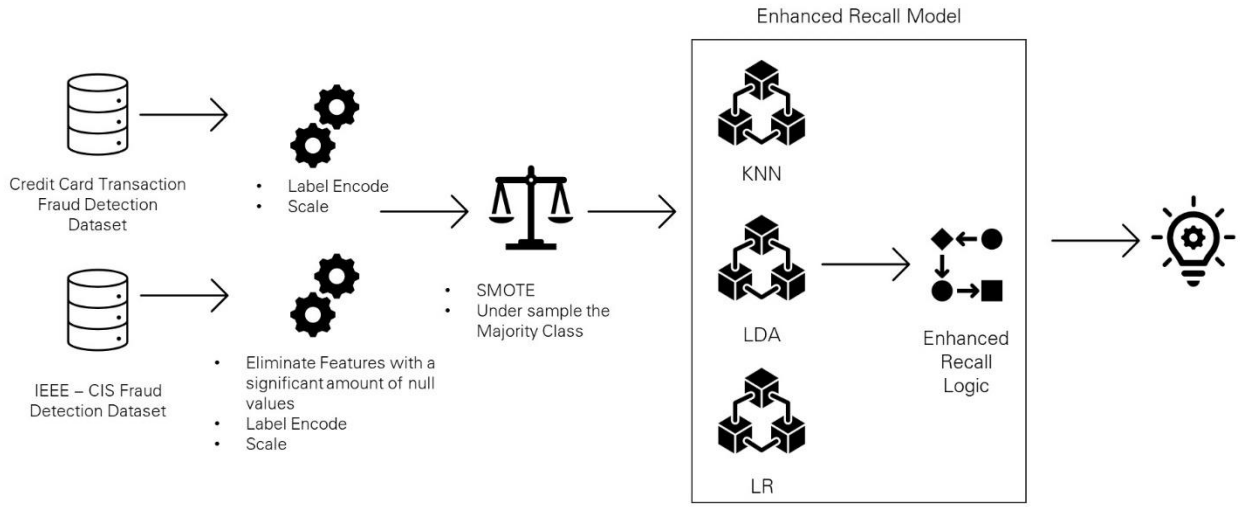


Figure - Full Enhanced Recall Model Machine Learning pipeline.

3.3.2 Additional Models

As part of this survey of Machine Learning models for detecting credit card data fraud additional work included implementing solutions described in [2] and [3].

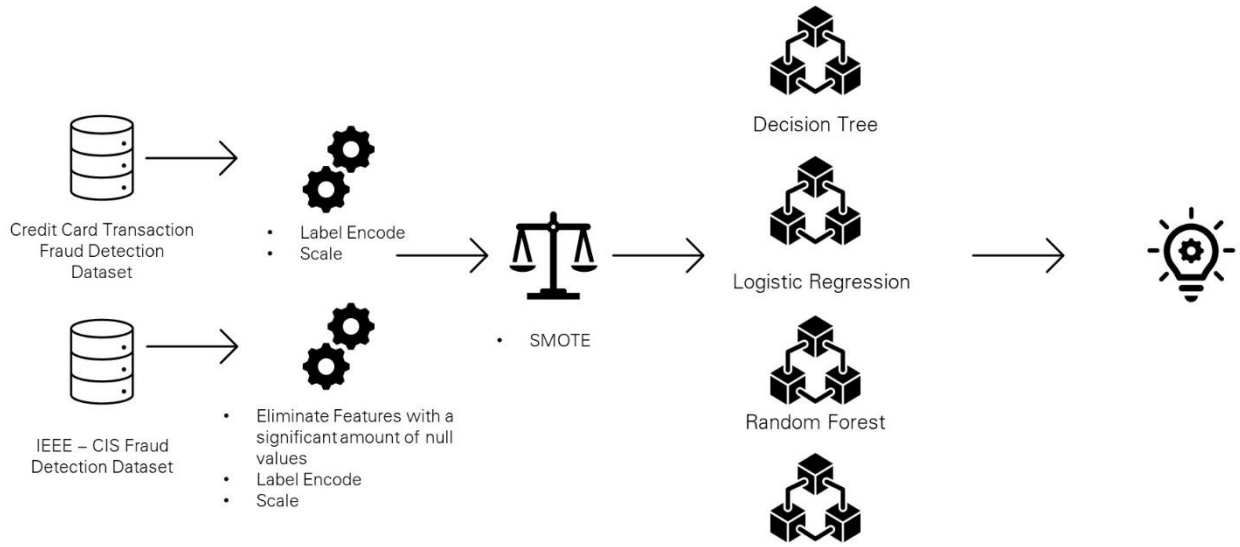


Figure - Secondary Models Machine Learning pipeline.

3.3.3 Grid Search

To experiment with different hyperparameters for each underlying model of the Enhanced Recall ensemble learning approach a Stratified Five Fold Cross-Validation was completed to determine the best hyperparameters for each dataset. This was less successful for the KNN model as excessive run times lasting greater than 24 hours prevented thorough experimentation. Therefore, Grid Search for KNN was completed on a subset of the training data to get results. The experiment did provide timely results for the best hyperparameters on the LDA and LR models. These were applied in re-training the models.

3.3.4 Principal Component Analysis

The IEE-CIS Fraud Detection Dataset has 394 columns of data providing 199 features after manual dimensionality reduction to remove columns with a high rate of missing values. Experiments with applying Principal Component Analysis were completed with n components at 50, 100, and 150 to determine if further dimensionality reduction would improve performance. The resulting transformed data was used to train the Enhanced Recall model with no difference in performance.

4 Results

Included in this section are the results of the model experimentation for each of the two datasets. Before providing the results from each dataset the following tables include the hyperparameters used for each model.

KNN Hyperparameters

	Algorithm	Leaf Size	Metric	Metric Params	N Jobs	N Neighbors	P	Weights
Documented	Auto	30	Minkowski	None	-1	5	2	Uniform
Grid Search**	Auto	20	Manhatten	N/A		3	1	Distance
Grid Search##								

** Credit Card Transactions Fraud Detection Dataset

IEE-CIS Fraud Detection Dataset

These parameters were found when using a subset of the SMOTE training data as run times limited the experimentation that could be done with KNN hyperparameters.

LDA Hyperparameters

	Covariance Estimator	N Components	Priors	Shrinkage	Solver	Store Covariance	Tol
Documented	None	None	None	None	Svd	False	0.0001
Grid Search**	None	None	None	Auto	Lsqr	True	0.0001
Grid Search##	None	None	None	None	Svd	True	0.0001

** Credit Card Transactions Fraud Detection Dataset

IEE-CIS Fraud Detection Dataset

Linear Regression Hyperparameters

	Copy X	Fit Intercept	Positive
Documented	True	True	False
Grid Search**	True	True	False
Grid Search##	True	True	False

** Credit Card Transactions Fraud Detection Dataset
IEE-CIS Fraud Detection Dataset

Support Vector Machine Hyperparameters

	C	Kernel	Gamma	Class Weight
Documented	10.0	Rbf	0.01	balanced

The following two sections include the test data evaluation scores for each model using the two datasets.

4.1 Credit Card Transactions Fraud Detection Dataset

Strategy for Enhanced Recall Test Data Metrics

	Enhanced Recall			
	Accuracy	Recall	F1	AUC
Unbalanced	0.99	0.49	0.22	0.74
SMOTE	0.92	0.75	0.07	0.84
Under Sample Majority	0.82	0.22	0.01	0.52
SMOTE after Grid Search	0.93	0.75	0.08	0.84
Chung et al.[1]	0.9664	0.9362		

	KNN				LDA			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
Unbalanced	1.0	0.0	0.01	0.5	0.99	0.49	0.22	0.74
SMOTE	0.90	0.02	0	0.46	0.95	0.75	0.11	0.85
Under Sample Majority	0.63	0.22	0	0.43	1.0	0.02	0.03	0.51
SMOTE after Grid Search	0.93	0.02	0	0.47	0.95	0.75	0.11	0.85

Additional Models

	Decision Tree				Random Forest			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
SMOTE	0.98	0.55	0.17	0.77	0.98	0.5	0.16	0.74
Under Sample Majority	0.8	0.02	0.00	0.42	0.87	0.08	0	0.47
Afriyie et al.[2]	0.92	0.93	0.09	94.5	0.96	0.97	0.17	0.989

	Logistic Regression				SVM			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
Unbalanced					0.72	0.52	0.01	0.62
SMOTE	0.96	0.73	0.12	0.85	0.96	0.73	0.13	0.85
Under Sample Majority	0.99	0.16	0.18	0.58				
Afriyie et al.[2]	0.92	0.76	0.08	87.9				
Xia[3]							0.446	0.9

4.2 IEE – CIS Fraud Detection Dataset

Strategy for Enhanced Recall Test Data Metrics

	Enhanced Recall			
	Accuracy	Recall	F1	AUC
Unbalanced	0.97	0.45	0.51	0.72
SMOTE	0.76	0.74	0.18	0.75
SMOTE after Grid Search				

	KNN				LDA			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
Unbalanced	0.98	0.42	0.56	0.71	0.96	0.25	0.34	0.62
SMOTE	0.91	0.78	0.39	0.85	0.76	0.74	0.18	0.75
SMOTE after Grid Search								

	Linear Regression with Binning Applied			
	Accuracy	Recall	F1	AUC
Unbalanced	0.66	0.8	0.15	0.73
SMOTE	0.76	0.74	0.18	0.75
SMOTE after Grid Search				

Additional Models

	Decision Tree				Random Forest			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
SMOTE	0.95	0.5	0.4	0.73	0.98	0.48	0.6	0.74

	Logistic Regression				SVM			
	Accuracy	Recall	F1	AUC	Accuracy	Recall	F1	AUC
SMOTE	0.75	0.73	0.17	0.74	0.78	0.74	0.2	0.76

5 Conclusions

As expected the models all achieved better test evaluation scores when trained using the dataset after applying SMOTE to balance the data. Across all the models Linear Discriminant Analysis was either the best or matched the best Accuracy, Recall, and AUC scores while evaluating the models using the Credit Card Transactions Fraud Detection Dataset. The F1 score was slightly lower than other models, but on the metrics that matter most, it won out. When evaluating the models using the IEEE-CIS Fraud Detection dataset the KNN model achieved the best test data scores across all of the metrics.

This work attempted to match the implementations described in the referenced papers as closely as possible to the described methodologies. Except for the Logistic Regression model, this work did not quite achieve the same test data evaluation scores as published in the referenced papers. This is likely due to a few factors. Without reference code, this work implemented the best interpretation of the proposed solutions but likely differed in some aspects. The specific Machine Learning libraries may have differed along with some hyperparameter and data engineering details. Additional attempts to get better test data scores by completing hyperparameter Grid Searches or Principal Component Analysis did not materially improve model performance.

The planned next step for research on this topic includes the implementation of Deep Learning strategies. Esenogho et al., propose an ensemble technique that uses the LSTM neural network as the base learner in the adaptive boosting (AdaBoost) algorithm. This method is significant for two reasons: the LSTM is a robust algorithm for modeling sequential data. Secondly, the AdaBoost technique builds strong classifiers that are less

likely to overfit [8]. Alarfaj et al. achieved better results using a 20-layer CNN when compared to traditional Machine Learning models [9].

Ultimately, the best solution to this problem may be better security protocols to prevent fraud in the first place. Until that goal is achieved the ability to accurately and quickly detect potentially fraudulent transactions will continue to be studied. While the solutions proposed in the research are interesting, the state-of-the-art is probably only known by those who develop proprietary solutions in use today.

6 References

1. Chung, Jiwon, and Kyung-Ho Lee. "Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression." Sensors, vol. 23, no. 18, Sept. 2023, p. 7788. https://doi.org/10.3390/s23187788.
2. Afriyie, Jonathan Kwaku, et al. "A Supervised Machine Learning Algorithm for Detecting and Predicting Fraud in Credit Card Transactions." Decision Analytics Journal, vol. 6, Mar. 2023, p. 100163. https://doi.org/10.1016/j.dajour.2023.100163.
3. Xia, Jianglin. "Credit Card Fraud Detection Based on Support Vector Machine." Highlights in Science Engineering and Technology, vol. 23, Dec. 2022, pp. 93–97. https://doi.org/10.54097/hset.v23i.3202.
4. Nilson Report [www.nilsonreport.com]. (2022b, December 22). Payment card fraud losses reach $32.34 billion [Press release]. Retrieved April 13, 2024, from https://www.globenewswire.com/news-release/2022/12/22/2578877/0/en/Payment-Card-Fraud-Losses-Reach-32-34-Billion.html
5. Vigderman, A. (2023, May 22). 2023 Credit Card Fraud Report. Security.org. https://www.security.org/digital-safety/credit-card-fraud-report/
6. IEEE-CIS Fraud Detection | Kaggle. (n.d.). https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203
7. Credit Card Transactions Fraud Detection Dataset. (2020, August 5). Kaggle. https://www.kaggle.com/datasets/kartik2112/fraud-detection
8. Esenogho, E., Mienye, I. D., Swart, T. G., Aruleba, K., & Obaido, G. (2022). A Neural Network Ensemble with Feature Engineering for Improved Credit Card Fraud Detection. IEEE Access, 10, 16400–16407. https://doi.org/10.1109/access.2022.3148298
9. Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed M. (2022). Credit card fraud detection using State-of-the-Art machine learning and deep learning algorithms. IEEE Access, 10, 39700–39715. https://doi.org/10.1109/access.2022.3166891