Predicting Digital Payment Fraud

Vijaykumar Mehtre

Bellevue University

Predictive Analytics 630

Andrew Hua

Introduction

This milestone outlines the initial plan for my individual project, focusing on Detecting Digital Payment Fraud with Machine Learning.

Digital transactions have become a part of daily life like purchasing a product online, sending money to friends, depositing cash in bank account, investment purposes etc., They had a lot of benefits so does paved way for fraudulent activities. People started using digital money transactions medium to launder money and make the money look like it comes from a legal source.

Data Source

The dataset used for this analysis is sourced from https://www.kaggle.com/datasets/ealaxi/paysim1 providing Type of transactions like CASH-IN, CASH-OUT or PAYMENT etc., Amount transacted and account type like CC" (Customer to Customer), "CM" (Customer to Merchant), "MC" (Merchant to Customer), "MM" (Merchant to Merchant).

This dataset was a sample of a much larger dataset (not available on Kaggle) generated from a simulation that closely resembles the normal day-to-day transactions including the occurrence of fraudulent transactions.

Modeling and Methods

As mentioned, the objective is identifying the online fraudulent transaction. So, target variable i.e. fraud_label should have true or false value.

Four machine learning algorithms were used to train and test the dataset after which they were evaluated using different metrics for best performance and deployment. We selected our models and target variable. The algorithms used were Logistic Regression, Random Forest, Decision Tree, and K-Nearest Neighbors. We trained the dataset on 80% while testing on 20%.

Both the Decision Tree and Random Forest models outperform the Logistic Regression and K-Nearest Neighbors model by a wide margin. Since they both have similar recall scores, we performed a cross-validation of the two models so we may declare which is the best performer with more certainty. Out of the 4 Machine Learning Models, Random Forest performs best with prediction accuracy 99.97% and recall accuracy 87% which is important for our problem statement where false negative is our priority.

Result Interpretation

The Decision Tree model with default parameters yields 99.96% accuracy on training data.

Precision Score: This means that 82% of all the things we predicted came true. that is 82% of clients transactions was detected to be a fraudulent transaction.

Recall Score: In all the actual positives, we only predicted 82% of it to be true.

Random Forest Tree model with default parameters yields 99.97% accuracy on training data.

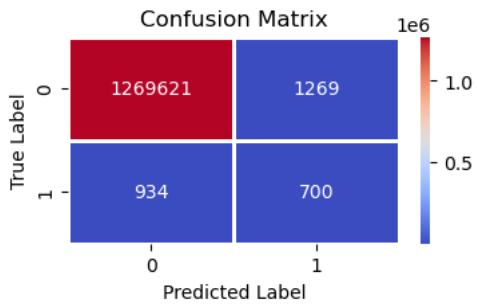
Precision Score: This means that 99% of all the things we predicted came true. that is 99% of clients' transactions was detected to be a fraudulent transaction.

Recall Score: In all the actual positives, we only predicted 81% of it to be true.

Both the Decision Tree and Random Forest models outperform the Logistic Regression and K-Nearest Neighbors model by a wide margin. Since they both have similar recall scores, we should perform a cross-validation of the two models so we may declare which is the best performer with more certainty.

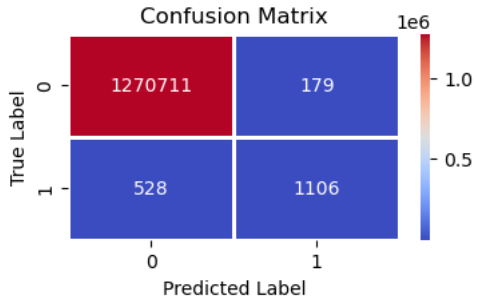
For **LogisticRegression**, Accuracy score is **0.9982687949303903**

	precision	recall	f1-score	support
0	1	1	1	1270890
1	0.36	0.43	0.39	1634
accuracy			1	1272524
macro avg	0.68	0.71	0.69	1272524
weighted avg	1	1	1	1272524



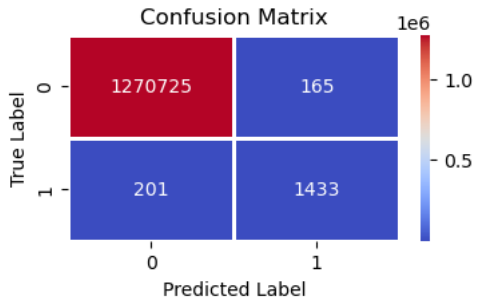
For **KNeighborsClassifier**, Accuracy score is **0.999444411264542**

	precision	recall	f1-score	support
0	1	1	1	1270890
1	0.86	0.68	0.76	1634
accuracy			1	1272524
macro avg	0.93	0.84	0.88	1272524
weighted avg	1	1	1	1272524



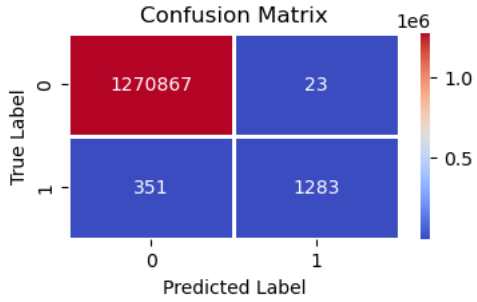
For **DecisionTreeClassifier**, Accuracy score is **0.9997123826348265**

	precision	recall	f1-score	support
0	1	1	1	1270890
1	0.9	0.88	0.89	1634
accuracy			1	1272524
macro avg	0.95	0.94	0.94	1272524
weighted avg	1	1	1	1272524



For **RandomForestClassifier**, Accuracy score is **0.9997060959164621**

	precision	recall	f1-score	support
0	1	1	1	1270890
1	0.98	0.79	0.87	1634
accuracy			1	1272524
macro avg	0.99	0.89	0.94	1272524
weighted avg	1	1	1	1272524



Conclusion

Upon training and evaluating our classification model, we found that the Random Forest model performed the best by a narrow margin.

Therefore, Random Forest performs best with recall cross-validation accuracy of 87% which is important for our problem statement where false negative is our priority.

Random Forest Classifier model should be deployed by bank because for this business problem, recall score is more relevant because it measures how many of the actual fraudulent payments the model identified as fraud.

Transaction History and Frequency - if unaccounted transactions occurs frequently bank should confirm genuinity of the transaction with the customer

Reference

Edgar Lopez-Rojas . Synthetic Financial Datasets For Fraud Detection from Kaggle website:

https://www.kaggle.com/datasets/ealaxi/paysim1