Detecting Credit Card Fraud Using Machine Learning

Brian Mann

Background

Credit card fraud is one of the most common types of fraud in the United States. In 2021, there were roughly 390,000 reported cases of credit card fraud to the FTC (Egan, 2023). It is estimated that these incidents of fraud cost Americans over $5-6 billion per year (Cruz, 2024). Much of this is not simply caused by lost or stolen cards, but by unauthorized access to personal data and information remotely. This project hopes to aid in the detection and prevention of fraud by using machine learning models such as decision trees and logistic regression. With swift and accurate fraud detection models in place, credit card companies can better safeguard their customers' data, thus saving billions in the process (Dieker, 2024).

The Problem

A hypothetical credit card company has tasked us with analyzing the activity from thousands of customers to create a model that can detect whether a fraudulent transaction has taken place.

Dataset

The data used in this project comes from a simulated list of one million credit card transactions that was attained from Kaggle (Narayanan, 2022). This is a labeled dataset, where it is known which transactions are fraudulent and which are deemed legitimate. There are eight variables included in the dataset:

- **distance_from_home (**float**):** distance from the user's residence to the location of the transaction
- **distance_from_last_transaction (**float**):** distance from the previous transaction's location to the current transaction's location
- **ratio_to_median_purchase_price (**float**):** ratio of the transaction's cost to the median transaction of all the user's purchases
- **repeat_retailer (**bool**):** Has this establishment been used by the user before?

- **used_chip (**bool**):** Did the transaction occur with a chip reader?
- **used_pin_number (**bool**):** Did the transaction occur with PIN entry?
- **online_order (**bool**):** Was the transaction online?
- **fraud (**bool**):** Is this transaction classified as fraud?

Data Prep

Each of the eight variables used in this project is derived from simulated data, so there was no need to clean and prep the dataset. Checks were still performed to detect any missing data, and there was none. Additionally, true or false variables were already converted to numeric form – 0 for False and 1 for True. The only transformations that needed to be made were log-transforms on the continuous variables for clarity, as well as scaling the dataset using a standard scaler during logistic regression.

Methods

In the first part of this project, we will conduct an initial EDA to assess the general trends in the dataset, such as measures of central tendency, correlation, and other distinctive characteristics. Once that is completed, the data will be split into a training and test set in preparation for modeling. Using a python ML libraries such as scikitlearn, modeling will be performed using decision trees and logistic regression. Once these models have been made, they will be tested for accuracy and consistency. A confusion matric and ROC-AUC curve will be used to visualize the results from these tests.

Potential Challenges

One of the challenges with this dataset is that the number of transactions labeled as fraud might be relatively small. If there are too few fraudulent transactions, it might be more difficult to create an accurate decision tree or regression model, as the model may try to report every transaction as non-fraudulent. More research may need to be done to understand exactly how to mitigate the consequences of this type of data. Another potential challenge is making sure that the ratio of fraudulent to non-fraudulent transactions remains consistent between training and testing data. This can be alleviated by performing several folds of cross-validation.

Analysis

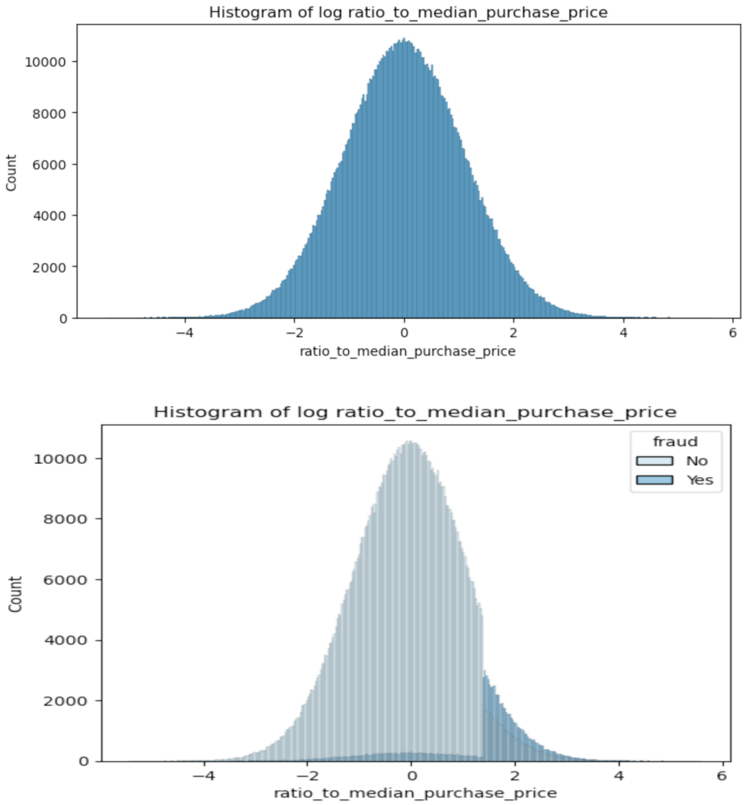


Figure 1: Histogram of Median Purchasing Price (Log-transformed)

During the initial phase of exploratory data analysis, it was discovered that for each of the continuous variables in the dataset – distance from home, distance form last transaction, and ratio to median purchase – were all log-normally distributed. From Figure 1, for fraudulent transactions, there tends to be a higher rate of high-than-median purchases. There is a similar trend for the distance metrics, but the difference was not as pronounced.

We then compared the boolean variables based on whether they were fraudulent or non-fraudulent (valid) transactions. Whether the transaction was fraudulent or not seemingly had no effect on if a repeated retailer was used. However, for chip-reader usage and pin-number usage, as well as online ordering there was a significantly fraudulent and valid transactions had a noticeable difference. For example, only 0.3% of fraudulent transactions included pin number entry. Likewise, over 94% of fraudulent transactions were online, compared with only 62% of valid transactions (Figure 2).

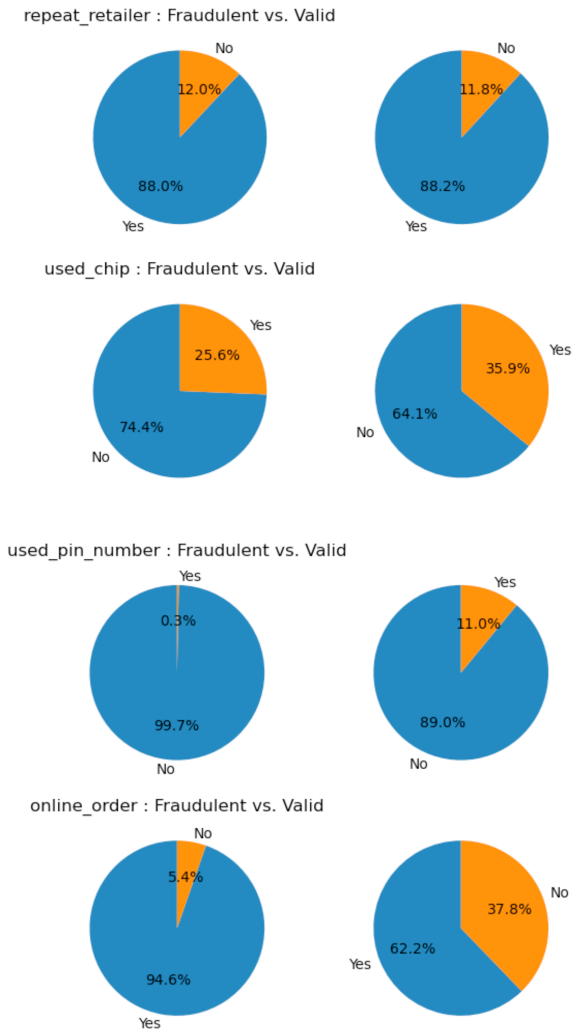


Figure 2: Pie Charts of Fraudulent vs Valid Transactions

Subsequently, a correlation table was generated that identified the Pearson-R value for each of the variables with each other. From this table (Figure 3), we were able to find that distance from home and distance from the last transaction had a slight positive correlation with fraudulent transactions, but ratio to median price had a pretty substantial correlation. In other words, a much higher-than-normal purchase would be the biggest indicator of fraud, according to the dataset. Additionally, there was also a slight positive correlation with online ordering and fraud. On the other hand, pin and chip usage, as well as repeat retailer usage both had very slight negative correlations with fraud.

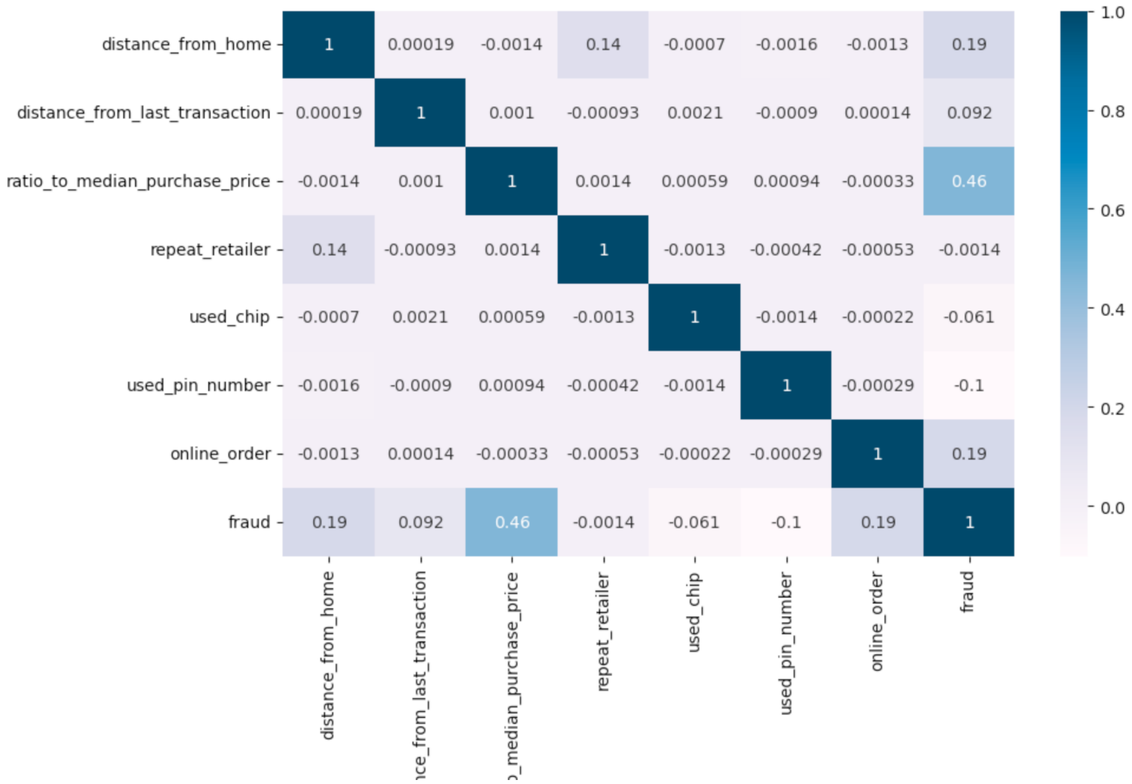


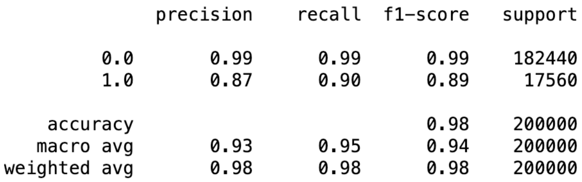
Figure 3: Correlation Plot of Credit Card Data

Modeling

In this project, we completed two separate models: a decision tree model and a logistic regression model. In each case, we first split the data into a training set and a test set in the ratio of 80/20. Our decision tree classifier was generated with a max depth of three and minimum sample split of 100. This was done to limit the effect of overfitting and prevent the decision tree from being overconfident on new data. After fitting the decision tree to our data, we were able to determine three simple indicators of fraud:

1. The ratio to median price is less than 4, distance from home is greater than 100km and the order is online.
2. The ratio to the median price is greater than 4, the order is not online, and the distance from home is greater than 100km.
3. The ratio to the median price is greater than 4, the order is online, and a pin number was not used.

We then tested the accuracy of our model. Out of the test sample of 200,000 transactions, over 196,000 were classified correctly, with an overall weighted accuracy of over 98% (Figure 4). Precision, accuracy, and f1-scores for fraudulent transactions were all around 90%.



```
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99    182440
         1.0       0.87      0.90      0.89     17560

    accuracy                           0.98    200000
   macro avg       0.93      0.95      0.94    200000
weighted avg       0.98      0.98      0.98    200000
```

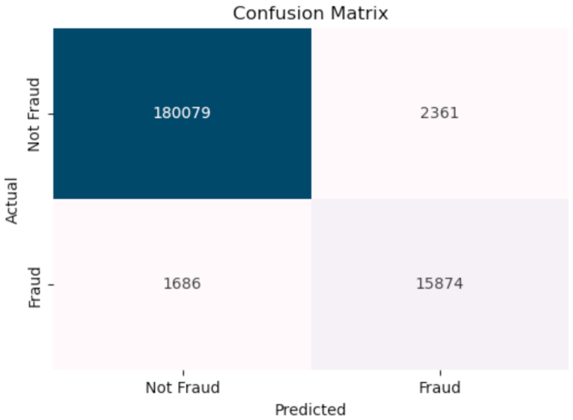
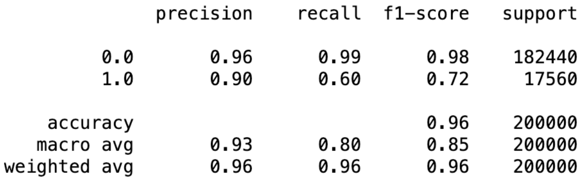


Figure 4: Results of Decision Tree Model Modeling

A second model was then created, this time via logistic regression. To mitigate the overweighting of our continuous data, we first scaled the data with a standard scaler, then the same general principles from our decision tree model were applied. This time, we were only able to achieve an overall accuracy of 96% (Figure 5). However, in this instance, recall was significantly worse – only 60%, thus leading to a lower f1-score (72%).



```
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98    182440
         1.0       0.90      0.60      0.72     17560

    accuracy                           0.96    200000
   macro avg       0.93      0.80      0.85    200000
weighted avg       0.96      0.96      0.96    200000
```

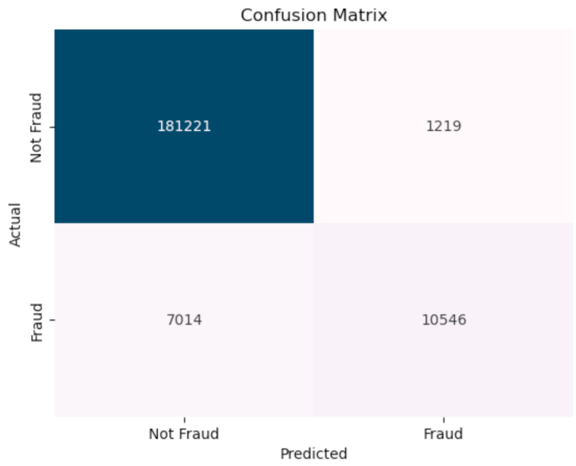


Figure 5: Results of Logistic Regression Modeling

Conclusion

From our decision tree modeling, we conclude that fraudulent transactions can be accurately predicted using a few key indicators. If the transaction has a ratio to the median price of over 4, the distance from home is over 100km, or the order is online, there is a higher likelihood that the transaction is fraudulent, although none of these factors by themselves is sufficient. Moreover, the performance of our logistic regression model was not quite as accurate as the decision tree model, so we would be more likely to opt for the decision tree.

Assumptions and Limitations/Challenges

As there weren't many features to model with, we were limited in how accurate our predictions could be. Despite this, both models we created seemed to be fairly accurate. As the data is simulated, it is likely that there may have been some purposeful engineering on the part of the data creator that led to certain correlations in the data. We are limited to what we have available, but with real world data, the results may not be quite as simple to determine.

Implementation Plan

We believe that our decision tree model is ready to be deployed. To deploy our model, we will run incoming data through our decision tree pipeline. If a transaction is unable to pass through our list of criteria, then it will be flagged as fraudulent, and an alert will be sent to the customer.

Recommendations and Future Applications

We recommend that in the future, we should attempt to integrate the same process of data preparation, analysis and modeling on real-world data. This should give us a better glimpse on how to protect consumers and corporations from credit card fraud. We believe these same processes can be applied to great success in the future.

Ethical Considerations

Typically, when considering the ethical implications to assessing credit card fraud, protecting customer privacy is paramount. As the dataset we are using is completely simulated, this will not be a factor in this project. No customer names, companies, locations, or other personal information has been included. Another factor to consider is that after modeling has been performed, if a malicious actor were to get a hold of such research, they may be able to use this information to circumvent fraud detection software. This is a very real issue, and it is precisely for this reason that most companies performing this type of research spend a lot of time and effort safeguarding their methodologies. It is unlikely that the results of this research would offer a criminal any advantages in avoiding fraud detection, as actual credit card companies use significantly more complex detection systems.

References

Cruz, B. (2024). *Credit Card Fraud Report.* Security.org. https://www.security.org/digital-safety/credit-card-fraud-report/

Dieker, N. (2024). *Know Your Rights When Facing Credit Card Fraud*. Yahoo Finance. https://finance.yahoo.com/news/know-rights-facing-credit-card-190100366.html

Egan, J. (2023). *Credit Card Fraud Statistics.* Bankrate. https://www.bankrate.com/credit-cards/news/credit-card-fraud-statistics/

Narayanan, D. (2022). *Credit Card Fraud*. Kaggle. https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud