Thomas Gartley

Josh Ehlke

Leonardo Rodriguez

Chandler Foster

Predicting Payment Card Fraud with Machine Learning

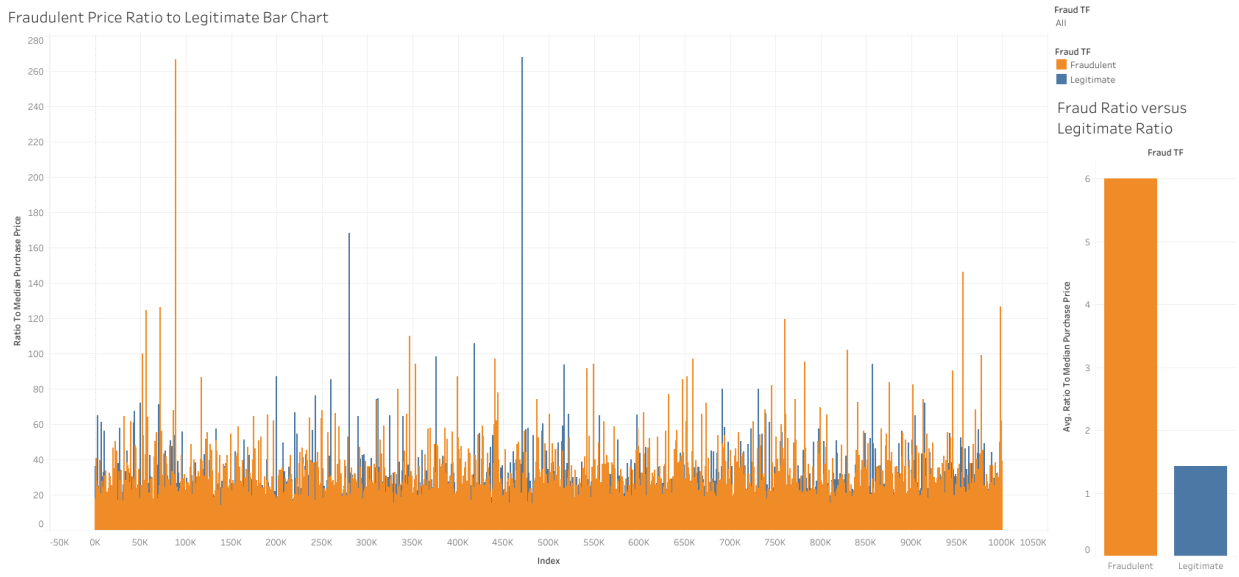
The purpose of our project was to determine if fraud could be accurately predicted based on factors contained in a dataset. This topic interested us because we have all been affected by card fraud at some point or another. Globally, payment card fraud results in $30 billion in losses annually. We wanted to create a system that could catch fraud before it happened.

Our data set consisted of a sample size of 1 million transactions. For each sample transaction, our data provided the distance between that transaction and the card holder's address, the distance between that transaction and the previous transaction, the ratio of the purchase price of the transaction to the card holder's median transaction price, whether or not the transaction took place at a repeat retailer, whether or not the transaction used the card's chip, whether or not the cardholder entered their PIN, whether or not the transaction occurred online, and whether or not the transaction was ultimately fraudulent. This data set did not provide the actual amount of money each transaction was for, nor did it provide the name of the retailer involved in each transaction.

Our data required minimal cleaning. The only preparation step that was required was to add an index column to our data so that each transaction would have a unique identifying key.

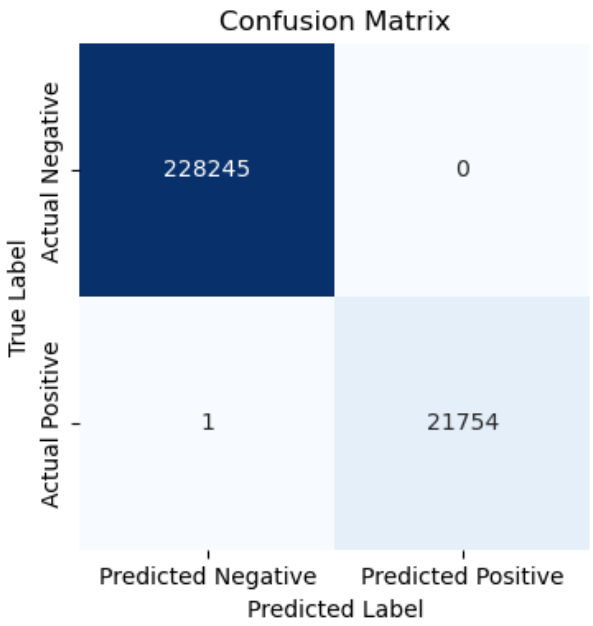
We chose to look into how useful it was to compare a given transaction's purchase price to that cardholder's median purchase price. In order to visualize this, we created a bar graph that showed this ratio for every transaction and color-coded the bars by fraud or legitimate

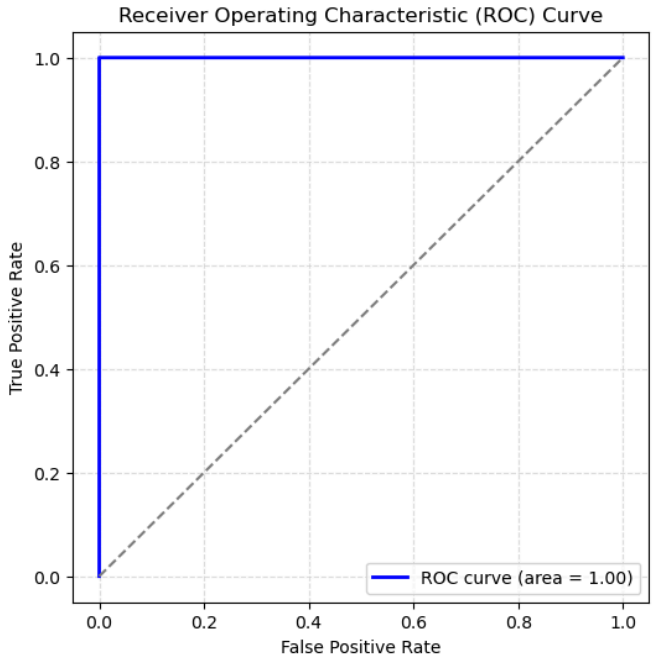
purchases. This graph did not provide entirely clear results, so we directly compared the average ratio for fraudulent purchases to legitimate purchases. It was apparent when comparing the average ratio to median purchase price of fraudulent versus legitimate charges that fraudulent charges were usually for amounts far higher than the cardholder's regular purchase price. We also looked at how the ratio of fraudulent purchases that took place at repeat retailers compared to that ratio of legitimate purchases and determined that this was not a good predictor of fraudulent activity. We also examined whether fraud was more or less common in online purchases. The visualization we created comparing fraudulent versus legitimate online purchases showed that nearly all fraudulent purchases took place online, whereas only a little over half of legitimate purchases took place online. It was clear from the visualizations we created that fraud was much more rampant in online purchases. We also created similar visualizations which showed that fewer fraudulent purchases took place when cardholders used their chip and even fewer took place when cardholders used their PIN.



We started the machine learning process by exploring the data and understanding what each column represented. We then used a scikit-learn pipeline to segment the data into a training and testing set and then scale the values so there was no contamination. After defining a function to help with understanding each model's performance, we tested 5 different models and compared them. Logistic Regression stood out as the worst, with a 60% accuracy on fraudulent charges. The XGBoost, LightGBM, and K Nearest Neighbors models were better, showing a 99% recall and accuracy. Finally, the Random Forest Classifier model was almost perfect, making only a single mistake when tested. After confirming that the model was actually

well-trained by running a 5-fold cross-validation, we decided to move forward with the model





because of its high performance. After retraining the model on the entire dataset, we were ready to use our model to make live insights on the website.

From this data, it can be concluded that use of a PIN is the surest way to prevent payment card fraud. While consumers may feel more comfortable shopping at the same retailers repeatedly, this is not actually a good indicator as to whether or not that merchant will generate a fraudulent purchase. Additionally, weaker correlations suggest that in-store, rather than online, purchases are more secure. Financial institutions should remain vigilant to protect their customers from fraud. By looking with a critical eye at online purchases and purchases where a cardholder's PIN is not submitted, financial institutions may be able to protect their customer and limit losses to fraud. In the future, it would be valuable to examine if specific retailers or retail categories are more likely than others to generate fraudulent transactions. It would also be valuable to see if automatically blocking transactions that are likely to be fraud based on the model we created reduces the actual occurrences of fraudulent transaction activity.

Works Cited

Chatgpt, chatgpt.com/. Accessed 9 Apr. 2025.

Deepseek, chat.deepseek.com/. Accessed 9 Apr. 2025.

R, Dhanush Narayanan. “Credit Card Fraud .” *Kaggle*, Creative Commons, www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud. Accessed 1 Apr. 2025.