Predicting Credit Card Fraud

Vincent Forca

Nov 7th, 2023

DataCamp Dataset

Python Visualizations

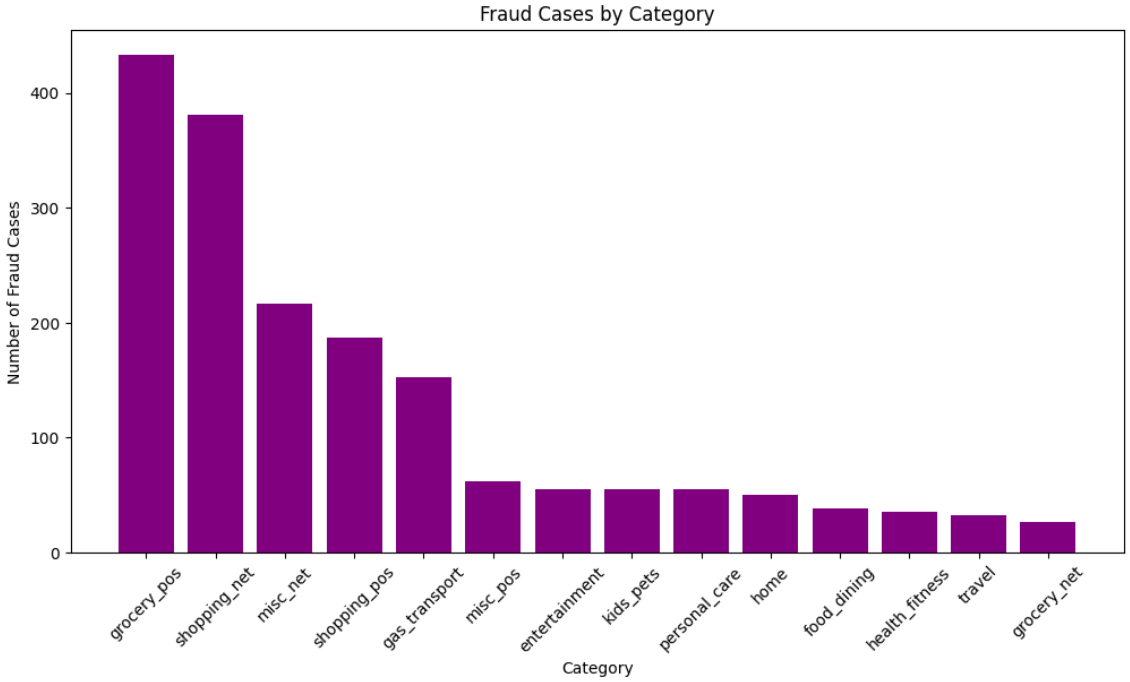
This report aims to investigate, describe, and diagnose fraudulent transactions with intentions of improving fraud detection parameters for the credit card company that provided the data.

Description of dataset provided on last page

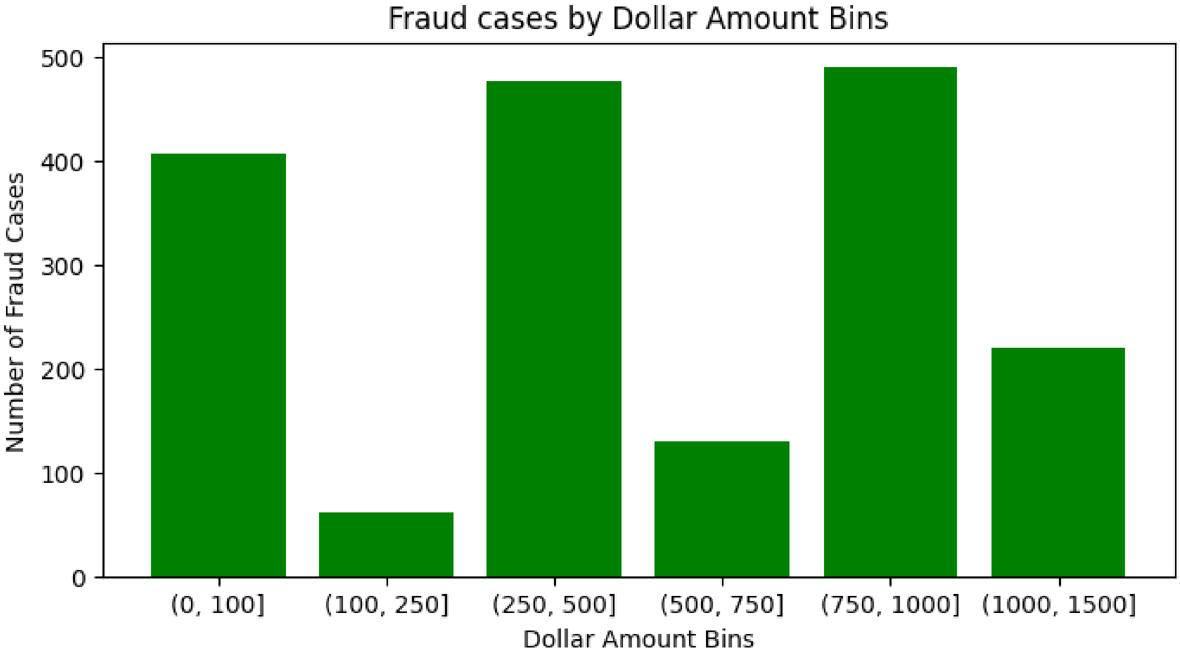
CREDIT CARD FRAUD REPORT:

By Vincent Forca: Original Workspace

Out of 339607 transactions, 1782 (about 0.5%) of them were fraudulent totalling $923,192.65 dollars in fraudulent charges. The Probability of randomly pulling a fraudulent transaction is about 0.00525. Here are the purchase categories of transactions resulting in fraud:



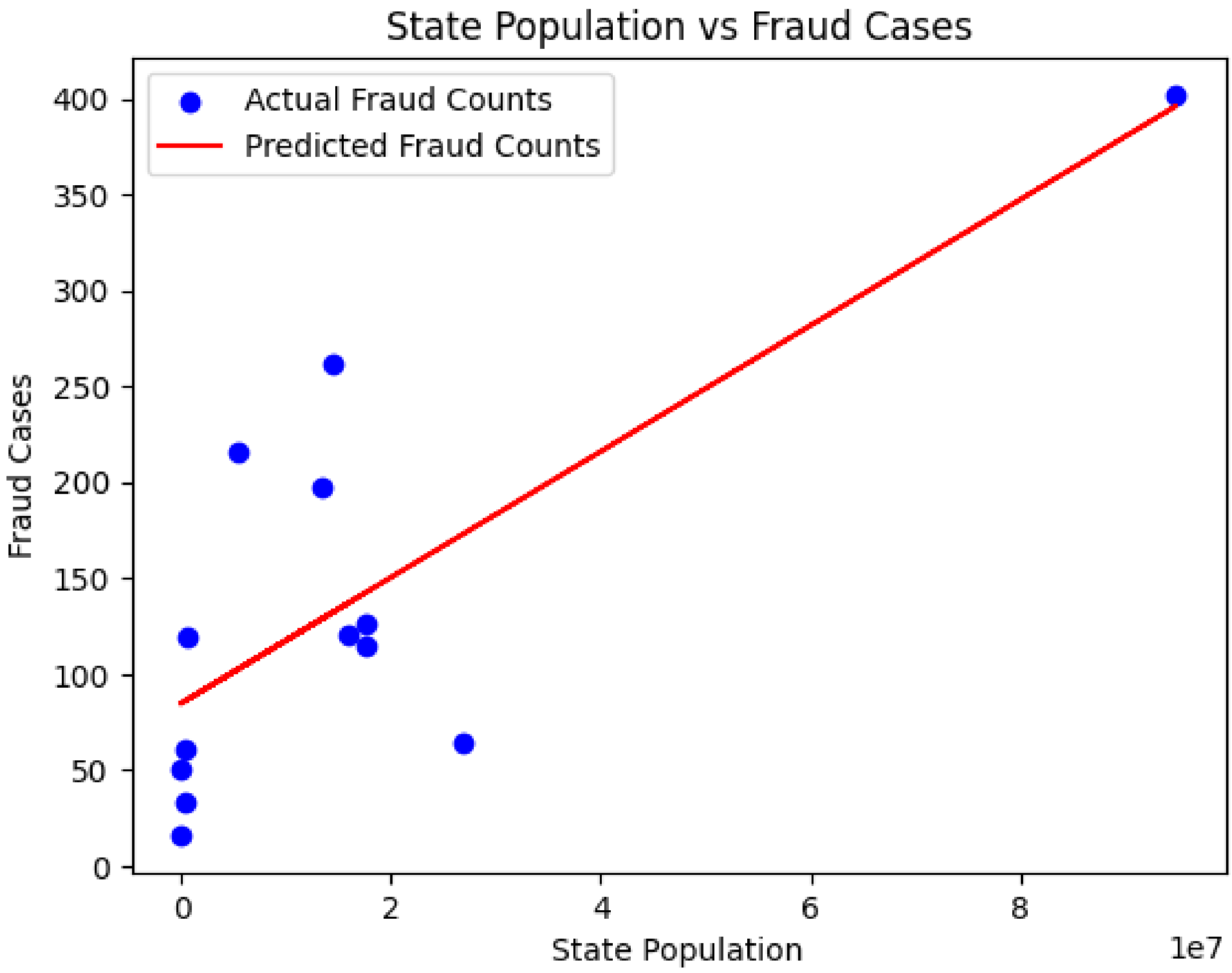
The majority of fraud (76.94%) takes place in the top 5 categories above: grocery stores, net-shopping, in-person shopping, and gas/transportation. It was discovered that grocery stores account for almost 100% of fraudulent charges within $250-$500 dollars (see below), while online shopping and in-person shopping accounts for most high dollar charges ranging from $750 to $1500:



Grocery stores account for almost 25% of all fraud scattered throughout the western states, yet all counts of grocery fraud were found to be a charge amount between $250 to $400 dollars. This does not seem random -- why would there be no grocery charges ranging from $0 to $249 dollars?

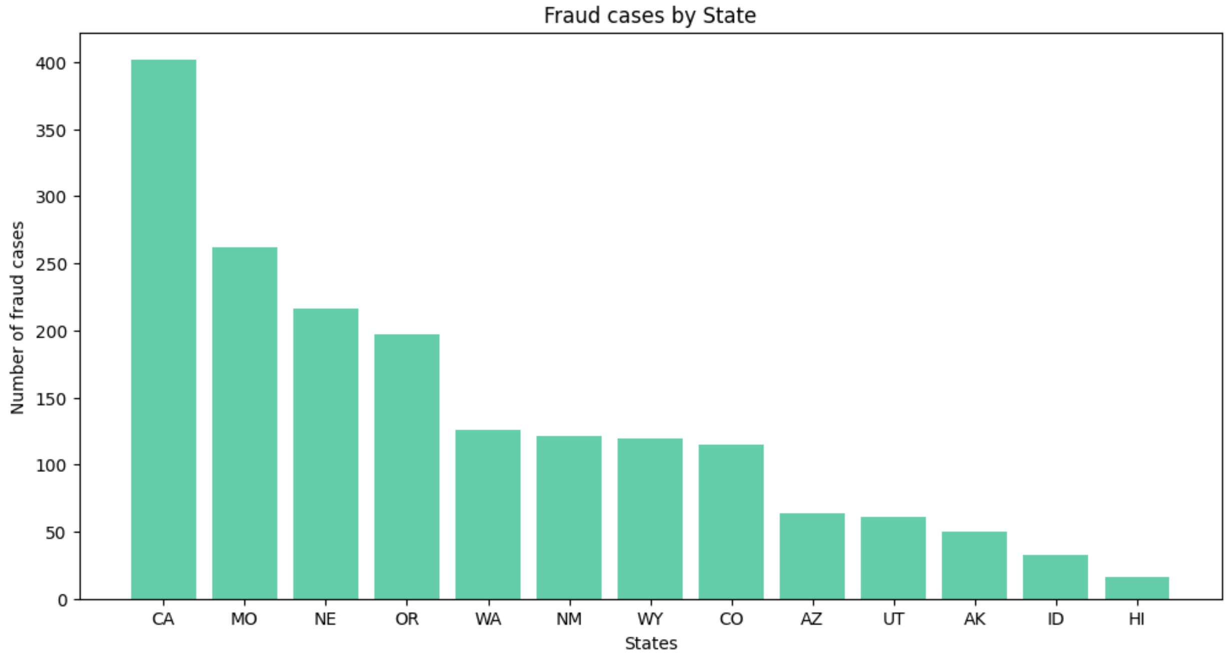
More data is required, but a theory could be that a criminal fraud group is scattered throughout the States and operating/exchanging/distributing the fraud information online as well as instructions on how much they can get away with.

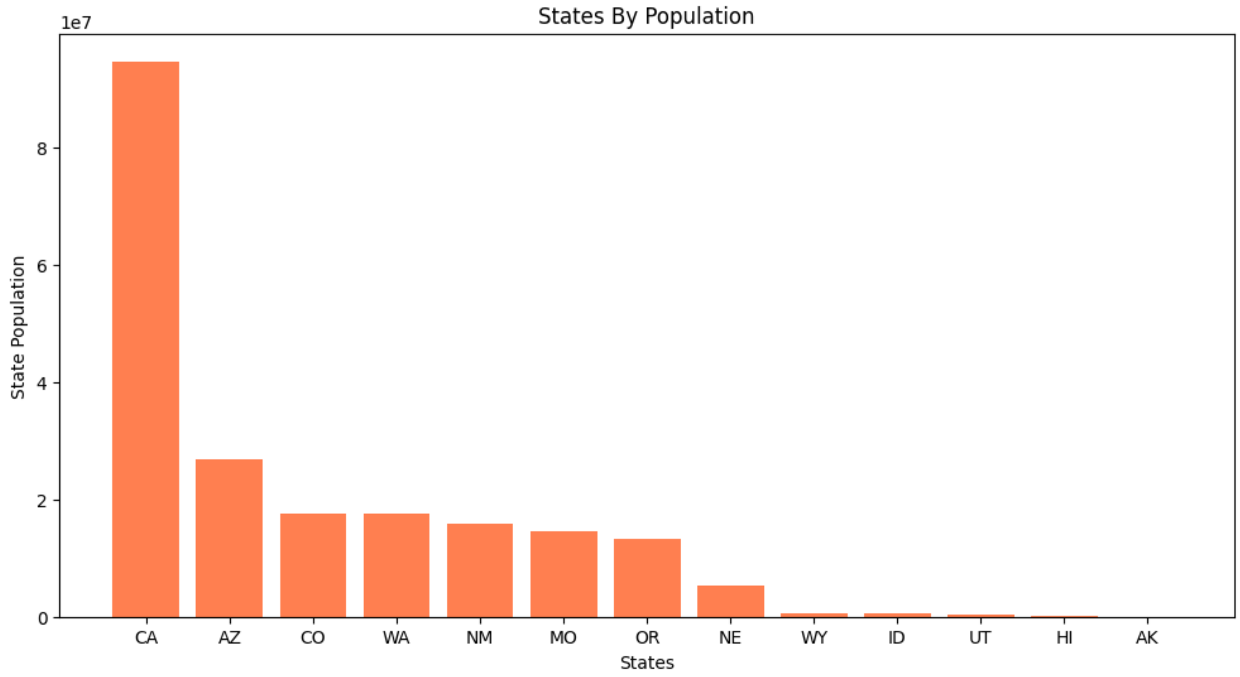
There are 332 unique merchants making up all of the fraudulent charges; certain merchants are associated with multiple fraud cases. We will now look at the populations of states.



The above analysis showed that the state population and fraud rate have a moderate positive correlation. We expect the high population states to yield a higher number of frauds.

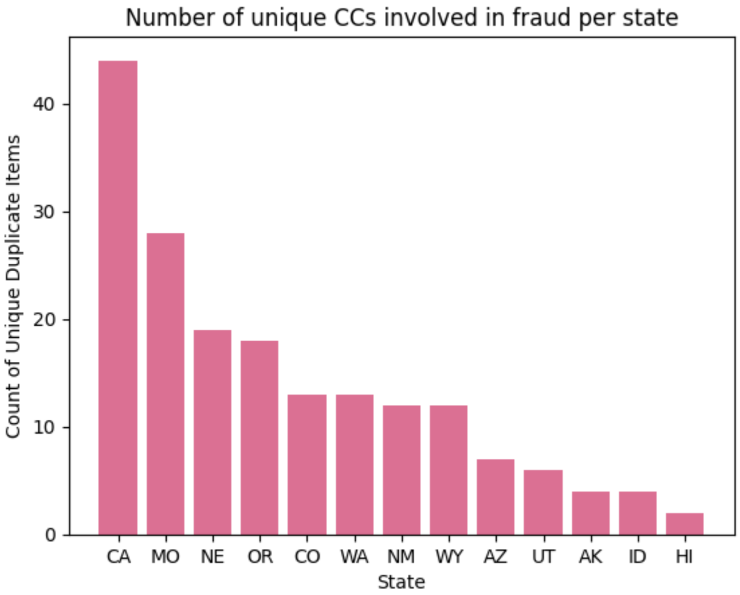
Fraud rates typically go up with an increasing population, but that is not the only factor in high fraud rates for each state. We wanted to find out what causes variation in fraud rates relative to population:

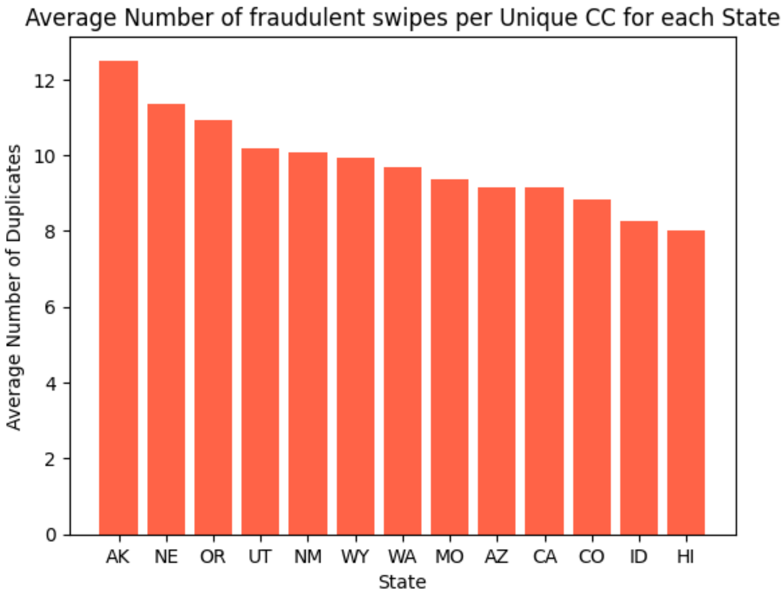




Certain states seem to have high fraud rates relative to their population rank. Diving deeper into these transactions, it was discovered that one credit card is often used for multiple cases of fraud. We found this information by grouping transactions with duplicate DOB, job, and city parameters.

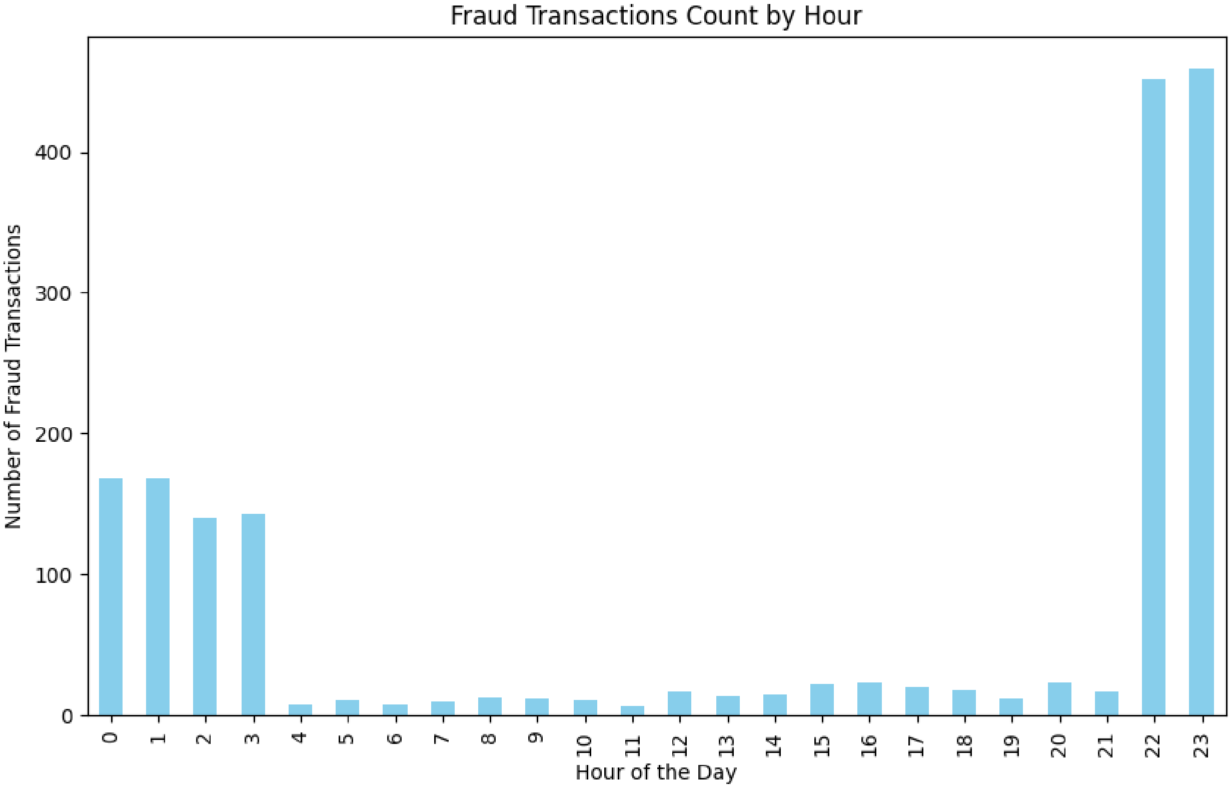
There were found to be 182 unique Credit Cards making up the 1782 fraudulent transactions. The average amount of fraudulent transactions for each stolen credit card is about 9.8, averaging roughly $518 dollars per swipe.


Number of unique CCs involved in fraud per state

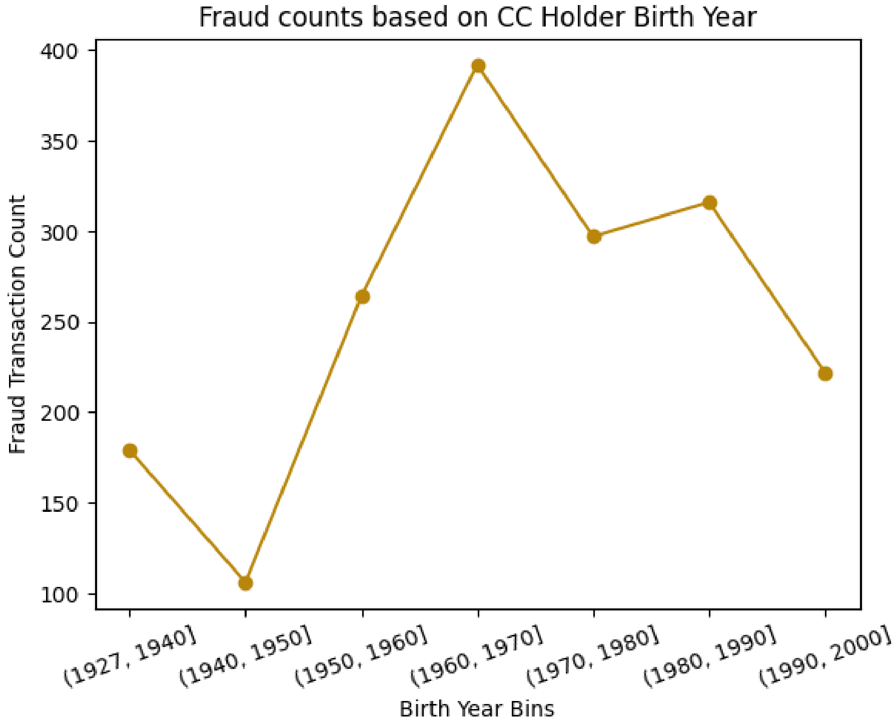

Average Number of fraudulent swipes per Unique CC for each State

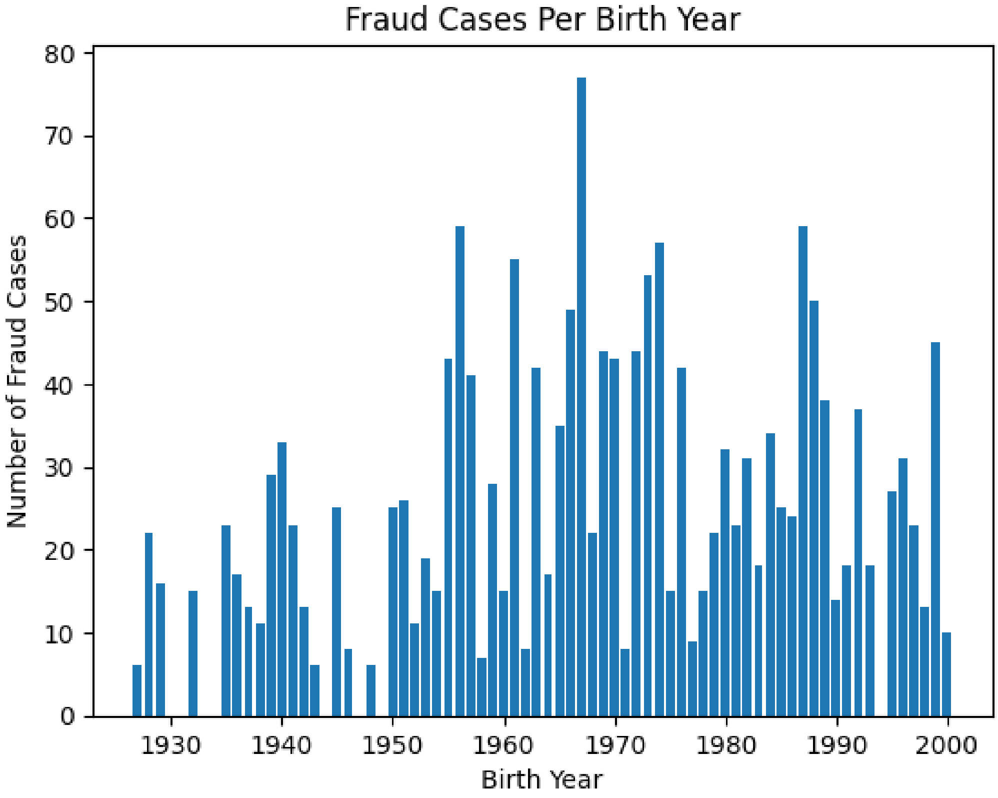
This information helps explain why some states have a higher number of fraud transactions with a smaller population. The states in question (such as NE) had a high average number of swipes per stolen card and/or a high number of unique cards.

It was also found that the time of day could be a strong indicator of future fraudulent transactions. The overwhelming majority of fraud cases (86%) take place from between 10pm - 4am.


Fraud Transactions Count by Hour

We then looked at how the age of the CC holder relates to the amount of fraud. It was determined that a higher age does not directly correlate with a higher number of fraudulent transactions. Although the largest group of fraudulent purchases were made from CC holders with a birthday ranging from 1960-1970, specific birth years with high counts of fraudulent transactions could be based on outside influences such as the timing of data leaks (when and how criminals collect stolen CC information), or multiple repeated transactions on stolen credit cards.





RECOMMENDATIONS:

In order to combat the criminal group distributing the stolen CC information and reduce the overall risk for the company, it is important to catch fraudulent transactions early to prevent continuous fraudulent use of compromised cards and reduce the number of average fraudulent transactions per stolen CC. We can set parameters to flag suspect transactions.

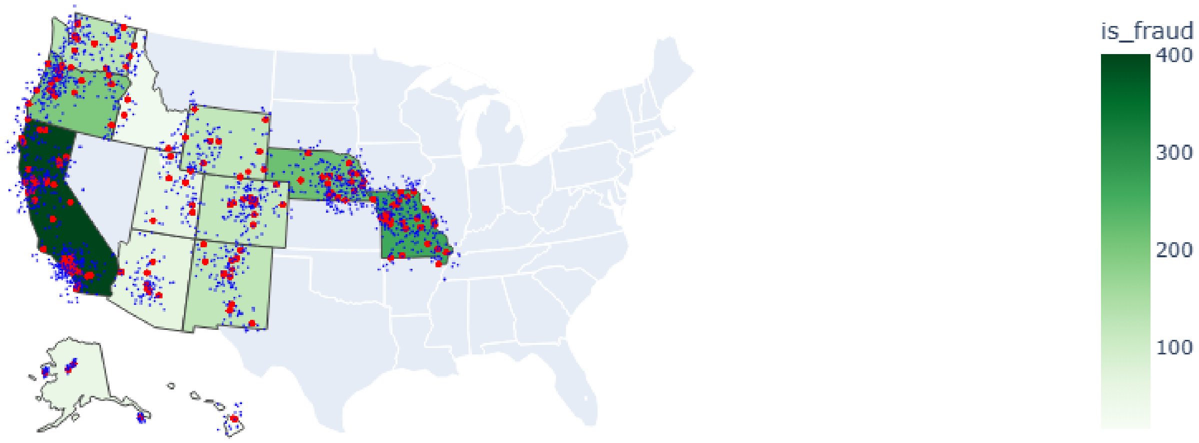
Possible red flags indicating fraud, based on these findings, would be CC purchases in a grocery store (grocery_pos) from $250-$400 dollars, charges for in-store shopping (shopping_pos) from $750+ dollars, or charges for online shopping (misc_net/shopping_net) of amounts exceeding $750+ dollars. These categories make up most of the fraud.

If repeat purchases on a credit card match the descriptions above, with purchase/merchant coordinates that do not align with the CC holder's city or residential area (see heat map of USA below), the transaction should be flagged and investigated.

Transactions taking place late at night that adheres to any of these parameters should especially be flagged. Detection parameters should be tighter at the end of the day – 51% of total cases took place from 10pm - 12am.

To err on the side of caution, we can add stricter parameters around cards where the CC holder's birth year is 1950-1990, which accounts for 71.2% of all known fraud cases. Stricter parameters may include a lower required number of "red flag" transactions before getting flagged by our system. Also, certain merchants had up to 18 fraudulent transactions and can be categorized as "high-risk". Transactions showing warning signs can be flagged when associated with a high-risk merchant.

Heatmap of Fraud Counts by State with Hoverable Cities



Red dots are main cities, blue dots are merchants where fraud took place.

Model:

Based on a random forest test, the model containing the original data is close to perfect at identifying non-fraud transactions, but less effective at identifying fraud. This model only correctly identified 55% of all fraud cases.

A logistic regression model correctly identified 78% of fraud, but was much less effective at identifying non fraud transactions, resulting in 12 thousand more false positives. This is better for discovering fraud, and errs on the side of caution, but more false-positives are annoying to clients and overall bad for business.

We want to adjust the parameters to be more effective at identifying fraud, and more effective at identifying non-fraud transactions.

We can change the model parameters to predictors parallel with our recommendations above.

To be Continued...

Credit Card Fraud

This dataset consists of credit card transactions in the western United States. It includes information about each transaction including customer details, the merchant and category of purchase, and whether or not the transaction was a fraud.

Data Dictionary

transdatetrans_time	Transaction DateTime
merchant	Merchant Name
category	Category of Merchant
amt	Amount of Transaction
city	City of Credit Card Holder
state	State of Credit Card Holder
lat	Latitude Location of Purchase
long	Longitude Location of Purchase
city_pop	Credit Card Holder's City Population
job	Job of Credit Card Holder
dob	Date of Birth of Credit Card Holder
trans_num	Transaction Number
merch_lat	Latitude Location of Merchant
merch_long	Longitude Location of Merchant
is_fraud	Whether Transaction is Fraud (1) or Not (0)

Source of dataset. The data was partially cleaned and adapted by DataCamp.