# Final Project

Law Neely

2024-02-27

## Data Background:

For my Research Report, I decided the dataset I wanted to work on was a credit card fraud transaction dataset. The dataset has over 550,000 rows of data. The variables in the dataset that I will be using are:

merchant: The name of the merchant where the transaction occurred. category: The category of the transaction (e.g., entertainment, kids/pets, health/fitness, personal care). amt: The amount of the transaction. city: The city where the transaction occurred. state: The state where the transaction occurred. job: The occupation or job title of the cardholder. is_fraud: A binary indicator (0 or 1) indicating whether the transaction is fraudulent.

I removed some of the columns such as first(first name), last(last name), etc. from my dataset because they were unneeded for the research I was doing.

I wanted to work on this dataset because during my time in this class I became intrigued with the risk management organization and utilizing data analytics to solve questions in that field. I am looking at it as a future career path and I believe this was a great idea of a real world scenario a Risk Analyst/Modeler may face.

### Research Questions:

I had a plethora of questions I wanted to answer about the dataset such as:

What are the common characteristics of fraudulent transactions compared to legitimate transactions? Is there a correlation between transaction amount and the likelihood of fraud? Do certain merchant categories have higher instances of fraudulent transactions? Are there specific geographical locations (cities or states) where fraud is more prevalent? Is there a pattern in the time of day or day of the week when fraudulent transactions occur? Do fraudulent transactions tend to involve certain types of jobs or industries? Can we predict the likelihood of a transaction being fraudulent based on certain features such as transaction amount, merchant category, location)? What are the most significant fraud categories? What are the most common fraud categories based on state?

The questions I chose to focus on for this report are: Is there a correlation between transaction amount and the likelihood of fraud? How do transactions amounts, categories, and region interact to affect the probability of fraud in credit card transactions?

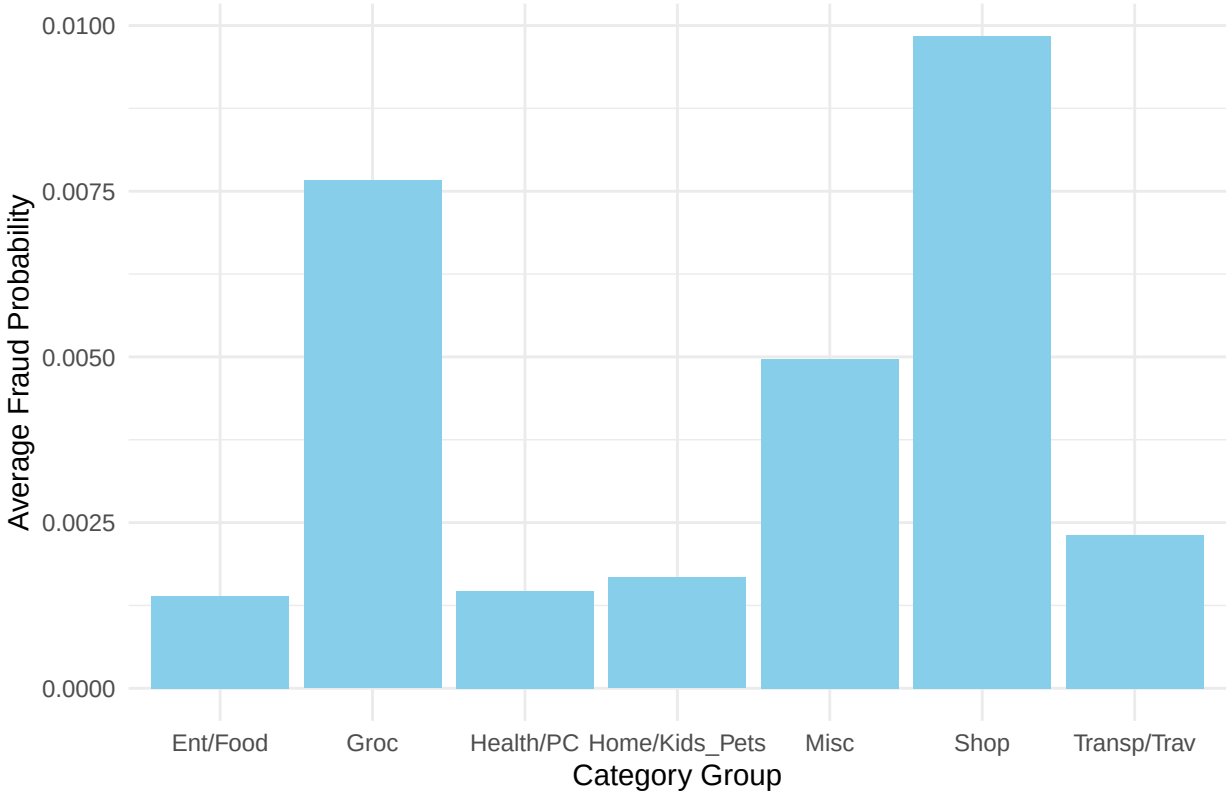
First things I needed to load in the data and I wanted to partition the dataset to test with 10% of the original dataset. So then we are only looking at around 55,000 rows compared to 550,000. I had to make sure that it was a proportionate amount of fraudalent and legitimate transactions so I used the createDatapartition() function from the caret package to achieve a stratified sample.

Next I removed the unneccessary columns and created new variables such as transaction_type and category_g to convert those variables to categorical variables. Then I added a new column for the amount range to put the amounts in different groups to make it easier to analyze.

## Data Visualization

To begin my Data Visualizations; first I did a simple bar chart to see the average fraud probability by category group. I was able to identify that the top 3 category groups for fraudalent transactions are Shopping, Groceries, and Miscallenaous

Fig 1: Average Fraud Probability by Category Group

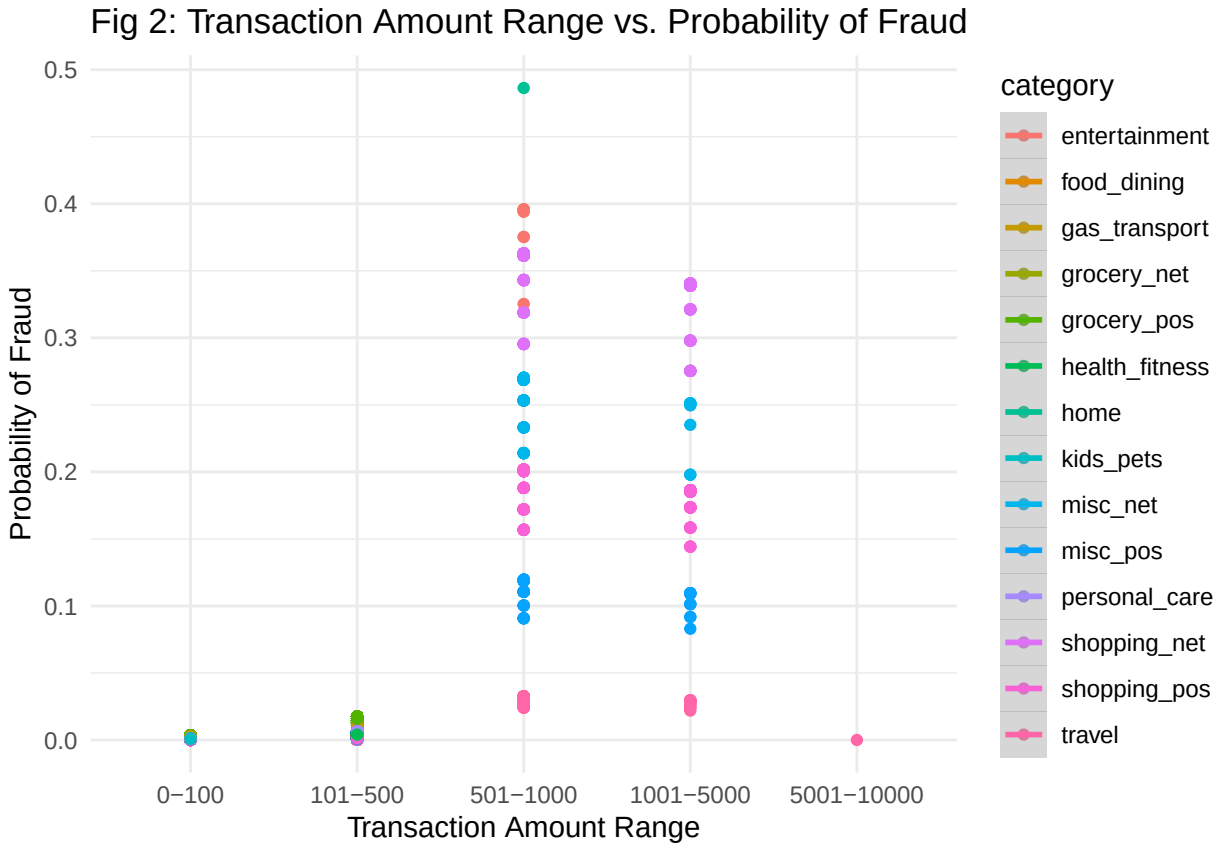



The next question I wanted to identify was if the US regions had any significance to fraudalent transactions. I add a new column called 'region' and split the states up into their appropriate region. I then completed a logistic regression to determine if the regions had any significance but unfortunately the regions did not.

Thanks to the logistic regression I was able to identify that the amount ranges of 101-500,501-1000, and 1001-5000 had a high significance which let me identify that transactions within those ranges have a higher chance of being fraudalent.

I created a scatterplot for the Transaction Amount Range vs Probability of Fraud to visualize the ranges with the highest probability for fraud while also adding color to symbolize the transaction category.

```
## `geom_smooth()` using formula = 'y ~ x'
```



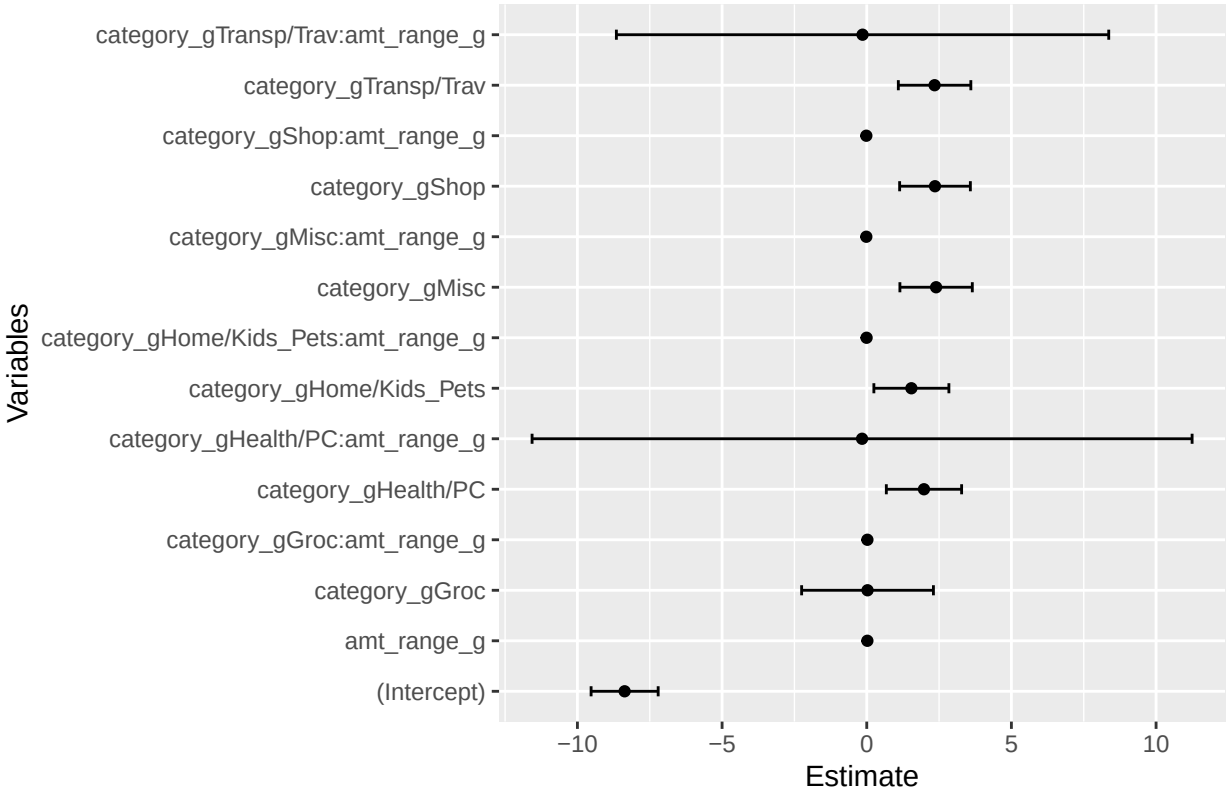

With the visualization above though it can be a little hard to tell which amount range is more significant for which transactional cateogry. To figure that out I decided to do another logistic regression to show the interaction between category and transaction amount range.

To do that I had to first convert the amount range column to numeric. So I added a new column called amt_range_g to do that.

Then I completed another logistic regression instead using category_group * amt_range_g to see the interactions of the two variables.

From that logistic regression I created a plot to show the coefficients from the logistic regression. The coefficients do show that the interaction between category and transaction amount range can actually affect fraud probability especially with Shopping and Miscellaneous categories showing higher probabilities of fraud.

Fig 3: Coeffs from Logistic Regression Model




# Summary:

## Methodology:

I used logistic regressions to analyze the relationship between transaction variables and fraud probability. I think logistic regression was the optimal choice since i had binary outcomes.

My analysis identified significant transaction amount ranges that have higher probabilities of fraud, such as 501-1000 and 1001-5000. The logistic regression model also showed interactions between transaction categories and amount ranges, revealing that Shopping and Miscellaneous categories have higher probabilities of fraud and the probability increases when amount range interacts with them.

## Data Visualization Recap:

I think that my bar chart showing the average fraud probability by category group effectively highlights which categories have higher instances of fraud.

The scatterplot for Transaction Amount Range vs. Probability of Fraud provides a clear visualization of how transaction amount affects fraud probability.

# Conclusion:

In conclusion, this analysis provided insights into the characteristics of fraudulent transactions in credit card data. By understanding these patterns, organizations can better detect and prevent fraud, leading to improved risk management strategies.

## Weaknesses and Implications:

Logistic regression assumes linearity between variables and due to that may not capture the complex interactions. Also, the dataset size could impact my model's performance so further validation with more data and different models could improve some results. From a moral and societal perspective, I think the big takeaway is that understanding fraud patterns can help protect consumers and improve financial security. From working in financial services I know how crucial that is and how models are constantly being testing and improved so this is a never ending battle to attempt to stop fraud but I believe with improved models we can continue to try and lower the probability and percentage of fraudulent transactions taking place.