

Project Three

Dsc 680



November 10, 2022

Colton Cook

**Table of Contents**

**Business Problem………………………………………………………………………………………………………………….2**

**Background/History………………………………………………………………………………………………………………2**

**Data Explanation……………………………………………………………………………………………………………………2**

**Methods………………………………………………………………………………………………………………………………..3**

**Analysis…………………………………………………………………………………………………………………………………3**

**Modeling Setup……………………………………………………………………………………………………………………..8**

**Conclusion…………………………………………………………………………………………………………………………….9**

**Assumptions………………………………………………………………………………………………………………………….9**

**Limitations…………………………………………………………………………………………………………………………….9**

**Implementation Plan……………………………………………………………………………………………………………10**

**Challenges……………………………………………………………………………………………………………………………10**

**Future Uses/Additional Applications……………………………………………………………………………………10**

**Ethical Assessment………………………………………………………………………………………………………………10**

**Business Problem**

From traditional to emerging sectors, there is not one single business that is fully immune from fraud. Some studies show that frauds of various kinds could cost businesses 1%-1.75% of their annual sales, this translates to around $200 billion a year!

As one of the most common types of fraudulent activities, credit card transaction fraud impacts around 127 million people, or approximately $8 billion in attempted fraudulent charges on Americans’ credit and debit cards. It is therefore imperative for credit card companies to understand the characteristics of a fraudulent transaction and develop predictive models accordingly to flag down potentially risky activities for fraud prevention

**Background/History**

Credit card fraud is hard to deal with as it doesn’t happen enough within the total amount of transactions that occur within the industry to give a great set of predictors. It can be hard to train something with such a unbalanced set. Some methods have been developed to look into individual people and monitor their spending habits to give an idea of when something goes wrong. That doesn’t always work though.

**Data Explanation**

**A picture containing chart

Description automatically generated**

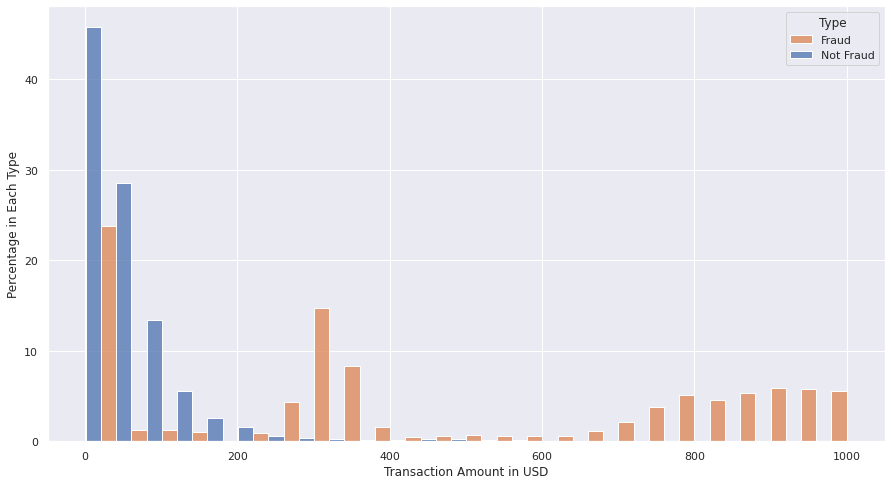
As can be seen from the results above, the training dataset contains 23 columns that detail the time of the credit card transaction, the merchant, the spending category, the transaction amount, and personal information about the credit card holders, including their names, genders, locations and birthdays. It also contains a column called "is\_fraud" which marks fraudulent transactions as 1 and non-fraudulent as 0. There is no missing data in the dataset and we also remove any duplicated observations in the data set to make it ready for further analysis.

**Methods**

Using simple EDA methods to look at various groups of data and compare them with graphs.

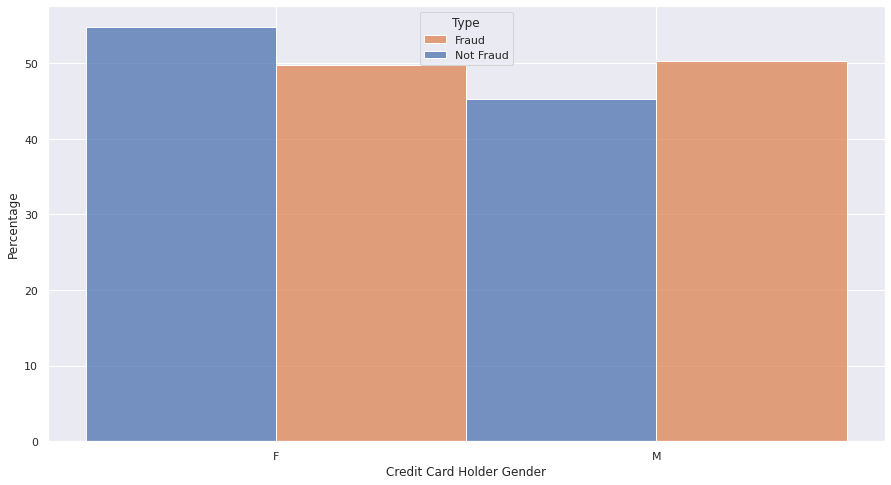
**Analysis**

With the dataset cleaned, we can now start to examine how various features relate to fraud. First we will see how the distribution of transaction amount differs between fraudulent and normal activities. As there are extreme outliers in transaction amount, and the 99 percentile is around $546, we subset the data for any transaction amounts below \$1,000 to make the visualizations more readable.



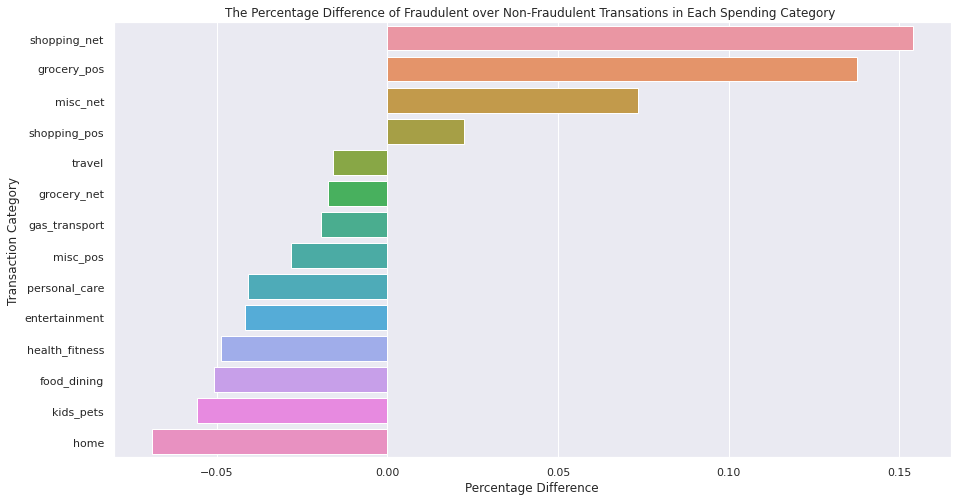
The result is very interesting! While normal transactions tend to be around $200 or less, we see fraudulent transactions peak around \$300 and then at the $800-\$1000 range. There is a very clear pattern here!

Examining whether one gender is more susceptible to fraud than the other.



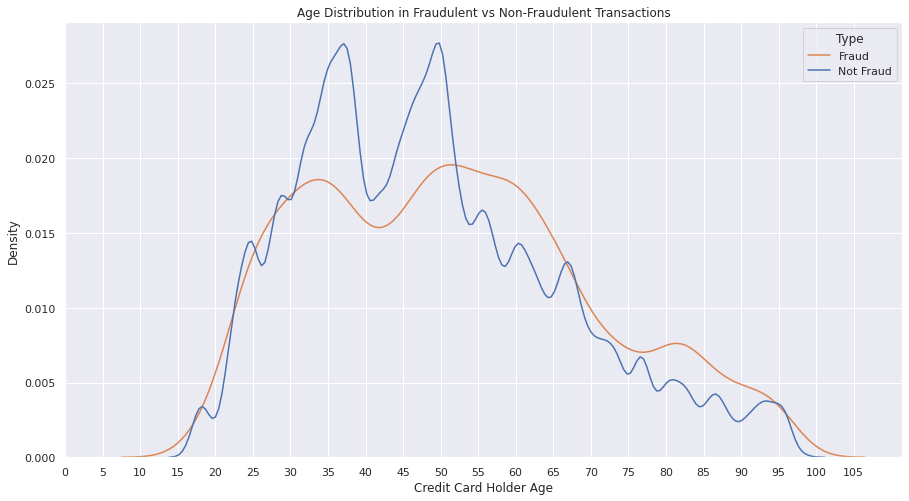
In this case, we do not see a clear difference between both genders. Data seem to suggest that females and males are almost equally susceptible (50%) to transaction fraud. Gender is not very indicative of a fraudulent transaction.

Examining in which spending categories fraud happens most predominantly. To do this, we first calculate the distribution in normal transactions and then the distribution in fraudulent activities. The difference between the 2 distributions will demonstrate which category is most susceptible to fraud. For example, if 'grocery\_pos' accounts for 50% of the total in normal transactions and 50% in fraudulent transactions, this doesn't mean that it is a major category for fraud, it simply means it is just a popular spending category in general. However, if the percentage is 10% in normal but 30% in fraudulent, then we know that there is a pattern.



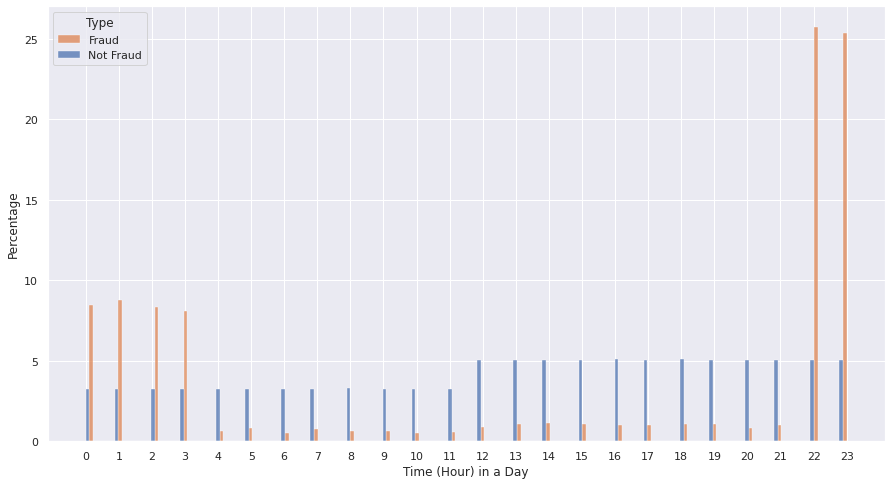
Some spending categories indeed see more fraud than others! Fraud tends to happen more often in 'Shopping\_net', 'Grocery\_pos', and 'misc\_net' while 'home' and 'kids\_pets' among others tend to see more normal transactions than fraudulent ones.

Are older people more prone to credit card fraud? Or is it the other way around? Given the birthday info, we can calculate the age of each card owner (in 2022) and see whether a trend exists.



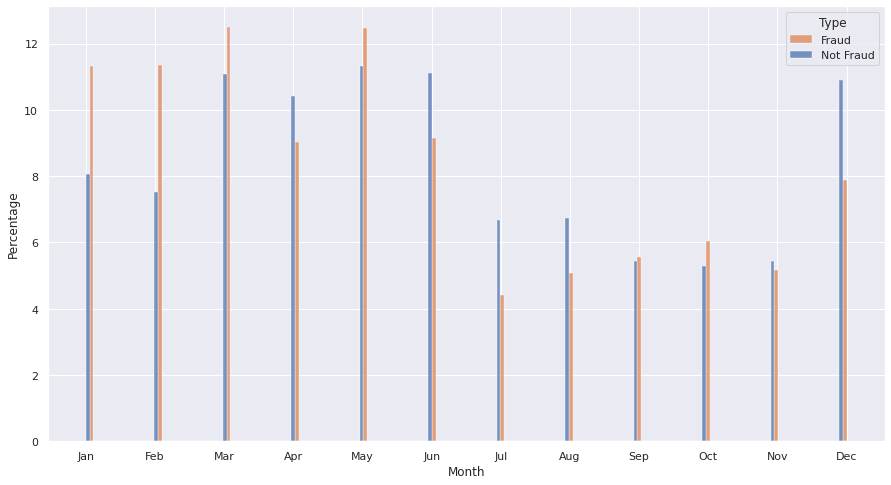
The age distribution is visibly different between 2 transaction types. In normal transactions, there are 2 peaks at the age of 37-38 and 49-50, while in fraudulent transactions, the age distribution is a little smoother and the second peak does include a wider age group from 50-65. This does suggest that older people are potentially more prone to fraud.

How do fraudulent transactions distribute on the temporal spectrum? Is there an hourly, monthly, or seasonal trend? We can use the transaction time column to answer this question.



A very sharp contrast! While normal transactions distribute more or less equally throughout the day, fraudulent payments happen disproportionately around midnight when most people are asleep!

Monthly Trend



Very interesting results! While normal payments peak around December (Christmas), and then late spring to early summer, fraudulent transactions are more concentrated in Jan-May. There is a clear seasonal trend.

**Modeling Setup**

Based on our EDA above, we have found out that the features including transaction amount, credit card holder age, spending category, transaction time and locations all have varying degrees of correlations with credit card fraud. This helps us choose which features we want to include in our data models. The plan is to train the models on the training data set which we have analyzed above and then use the testing dataset to evaluate the model performance.

As data models need numeric input, we need to convert some of our categorical observations into numeric ones. For transaction locations and merchant locations, we already have the longitudinal and latitudinal data. But for shopping categories, we need convert them into dummy variables using pandas.get\_dummies.

We will first try to use Logistic Regression combined with confusion matrix to evaluate the model. As is very common with fraud data, there is always the issue of class imbalance where actual fraud cases are way fewer than normal cases and constitute only a very small part of the dataset. To counter this imbalance, it's important to use the SMOTE (Synthetic Minority Oversampling Technique) method to resample the training dataset so that the model can be trained on more balanced data for better results.

Table

Description automatically generated

**Conclusion**

The model did ok. It looks good with a .9961 accuracy, but its not that great with the confusion matrix. The best way to handle this would be to increase the number of fraudulent actions to give the algorithm a better chance of picking up the right ones to classify as fraud or not.

**Assumptions**

With all data, I am assuming it is accurate and correct. There isn’t a bias in the data and its properly entered. It needs to be a source of truth to really give it preference to correct answers.

**Limitations**

Class imbalance is the biggest issue with this data. Also some of the features probably aren’t allowed to be used in a production environment.

**Implementation Plan**

This model would need more work and better data to really be used in a meaningful way. If I were to deploy it, it would be in the cloud with batch processing. It wouldn’t need to run every day and could be done monthly probably to save on cost. Monitoring it would be important for drift and adding new data to the model would be important as well.

**Challenges**

There are no real challenges with this model setup or running it.

**Future Uses/Additional Applications**

Could look at credit card defaults with similar data I imagine. Would have some of the same issues as before though with gender, age etc.

**Ethical Assessment**

There are some issues. Some features are not ok. Gender, age, location might even be not ok for financial institutions to be making fraud claims. Those would need to be cleaned up and probably more of an unsupervised method would be needed to truly get this model out of development and into production.