Credit Card Fraud Detection – Exploratory Data Analysis

Dataset summary

I am using Credit Card Fraud Detection dataset for this assignment. I found a few datasets on Kaggle related to fraud detection. However, I decided to work with the one of the smaller datasets with 3075 rows and 12 columns (11 feature columns and 1 target column)
Dataset URL: https://www.kaggle.com/shubhamjoshi2130of/abstract-data-set-for-credit-card-fraud-detection?select=creditcardcsvpresent.csv

Here is the output showing the column names in the dataset file.

```
In [5]: filepath = "data/creditcardcsvpresent.csv"
        data = pd.read_csv(filepath)
        data.head()
```

Out[5]:

	Merchant_id	Transaction date	Average Amount/transaction/day	Transaction_amount	Is declined	Total Number of declines/day	isForeignTransaction	isHighRiskCountry	Daily_chargeback_av
0	3160040998	NaN	100.0	3000.0	N	5	Y	Y	
1	3160040998	NaN	100.0	4300.0	N	5	Y	Y	
2	3160041896	NaN	185.5	4823.0	Y	5	N	N	
3	3160141996	NaN	185.5	5008.5	Y	8	N	N	
4	3160241992	NaN	500.0	26000.0	N	0	Y	Y	

```
In [6]: data.shape
Out[6]: (3075, 12)

In [7]: data.columns
Out[7]: Index(['Merchant_id', 'Transaction date', 'Average Amount/transaction/day',
               'Transaction_amount', 'Is declined', 'Total Number of declines/day',
               'isForeignTransaction', 'isHighRiskCountry', 'Daily_chargeback_avg_amt',
               '6_month_avg_chbk_amt', '6-month_chbk_freq', 'isFradulent'],
              dtype='object')
```

'IsFradulent' is the target column, the rest are the features of this dataset.

Data exploration

I am using some of the techniques described in the course.

1. Find data types for each column/feature

```
Merchant_id                       int64
Transaction date                float64
Average Amount/transaction/day  float64
Transaction_amount              float64
Is declined                      object
Total Number of declines/day      int64
isForeignTransaction             object
isHighRiskCountry                object
Daily_chargeback_avg_amt          int64
6_month_avg_chbk_amt            float64
6-month_chbk_freq                 int64
isFradulent                      object
dtype: object
```

2. Examine null/missing values in columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3075 entries, 0 to 3074
Data columns (total 12 columns):
Merchant_id                         3075 non-null int64
Transaction date                    0 non-null float64
Average Amount/transaction/day      3075 non-null float64
Transaction_amount                  3075 non-null float64
Is declined                         3075 non-null object
Total Number of declines/day        3075 non-null int64
isForeignTransaction                3075 non-null object
isHighRiskCountry                   3075 non-null object
Daily_chargeback_avg_amt            3075 non-null int64
6_month_avg_chbk_amt                3075 non-null float64
6-month_chbk_freq                   3075 non-null int64
isFradulent                         3075 non-null object
dtypes: float64(4), int64(4), object(4)
memory usage: 288.4+ KB
```

The highlighted column in yellow does not have any data.

3. Find duplicates

4. Find outliers

5. Find the string/non-numeric columns for categorical encoding

```
['Is declined', 'isForeignTransaction', 'isHighRiskCountry', 'isFradulen
t']
```

6. Find minimum, maximum, range, mean and median in numeric columns.

7. Find skew in numeric columns.

8. Find the correlation between each of the measurements.

Data cleaning

Since ‘Transaction date’ column does not have any data, this column can be dropped for this analysis. It is hard to conclude at this time if transaction data could relate to fraudulent transaction.

Feature engineering

1. Perform binary encoding on the string columns. The 4 string columns have only two values, i.e., ‘Y and N’. That can be encoded as 1 and 0 respectively.

2. Perform feature scaling on the following numeric columns.

```
Average Amount/transaction/day
Transaction_amount
Daily_chargeback_avg_amt
6_month_avg_chbk_amt
```

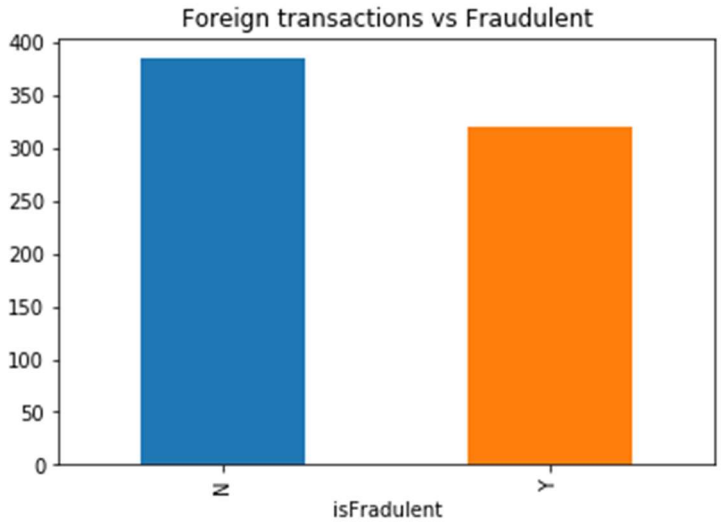
Key Findings and Insights

Based on the exploratory analysis, the dataset is imbalanced. There are large number of non-fraudulent transactions as compared to fraudulent transactions. This could potentially create a bias in training a machine learning model. It would be helpful if we get more data from the source.

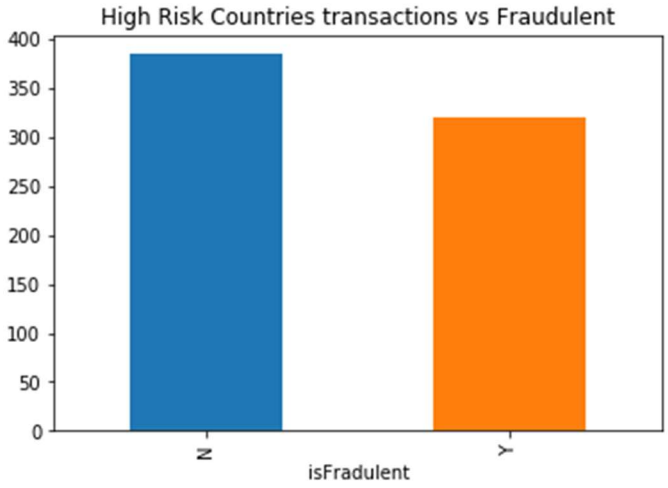
Formulating at least 3 hypothesis

Following are 3 hypothesis:

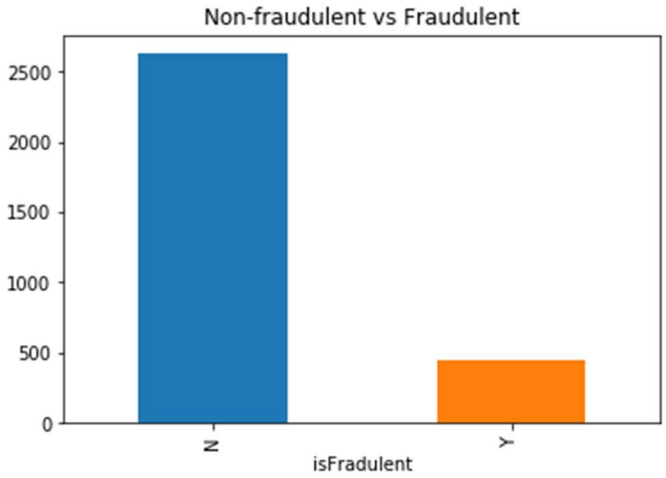
- There is a greater chance (almost 40%) for a transaction to be a fraudulent if it happens in a foreign country

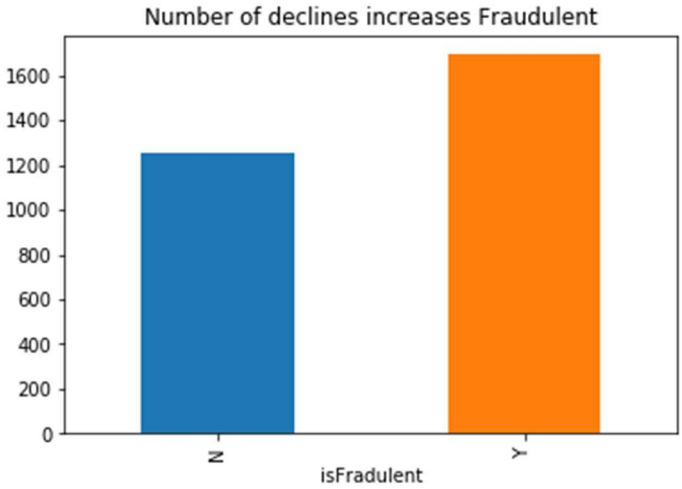


- There is a greater chance for a transaction to be a fraudulent if it happens in a high risk foreign country



- There is a correlation between fraudulent transactions and number of declines per day.





Significance test for one of the hypotheses

It seems that the transactions that are happening in foreign countries, there is a strong possibility that those transactions could be fraudulent. The null and alternate hypothesis will be:

H0 = If the transaction is foreign, there is 40% chance of fraud
H1 = If the transaction is foreign, there is less than 40% chance of fraud

There are 706 foreign transactions in the dataset. Out of those 706 transactions, 321 transactions are fraudulent. That is close to 45% fraud in all foreign transactions as compared to 5% fraud in all non-foreign transactions.

Next steps in analyzing this data

It would be a good idea to fully analyze this data using the available tools we have namely:

- Use histogram to find outliers
- Use Matplotlib for histogram and box plots
- Seaborn for pair plots visualizations and find more correlations
- Perform log transformation for skewed numeric features
- Find non-linear correlations and use of polynomials

Quality of dataset

I would consider this dataset a raw dataset that has enough information to draw hypothesis. However, due to imbalanced nature of the dataset for fraud vs. non-fraud transactions and missing information on transaction date, it would be helpful to get more balanced data as well as acquire missing data that could affect the result of hypothesis. With more analysis performed after getting additional data, it would help deciding on choosing the correct model.