ANALYSIS OF CREDIT CARD

Predicting Whether a Credit Card Transaction is Fraud.

About the Data:

The 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. The goal here is to accurately predict whether a transaction is fraudulent or not.

Data Preprocessing:

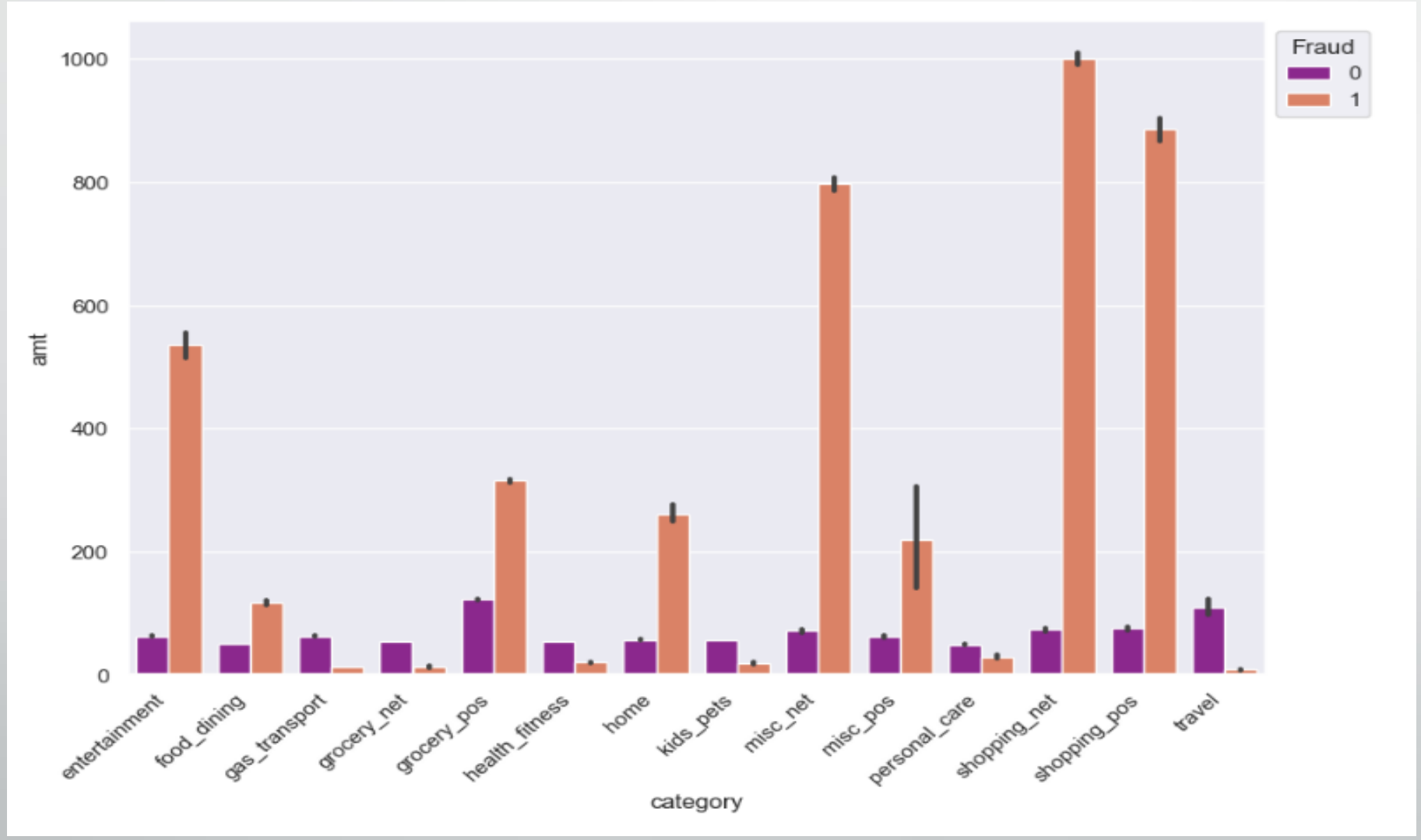
The dataset contains 339607 records and total of 15 columns. After carrying out data preprocessing, there was no missing entries.

Relationship Analysis

The correlation matrix:

	amt	lat	long	city_pop	merch_lat	merch_long
amt	1.000000	0.002817	-0.007017	0.006050	0.002957	-0.006891
lat	0.002817	1.000000	-0.151364	-0.236336	0.993646	-0.151267
long	-0.007017	-0.151364	1.000000	-0.066080	-0.150385	0.998960
city_pop	0.006050	-0.236336	-0.066080	1.000000	-0.235019	-0.066052
merch_lat	0.002957	0.993646	-0.150385	-0.235019	1.000000	-0.150278
merch_long	-0.006891	-0.151267	0.998960	-0.066052	-0.150278	1.000000

The correlation matrix shows that, there is a high positive correlation between longitude location of purchase and longitude of merchant with correlation value, r = 0.9989; and also, with latitude location of merchant with the latitude location of purchase with r=0.99.


Fraud
0
1
amt
1000
800
600
400
200
0
entertainment
food_dining
gas_transport
grocery_net
grocery_pos
health_fitness
home
kids_pets
misc_net
misc_pos
personal_care
shopping_net
shopping_pos
travel
category

The bar chart in the previous slides shows the relationship between the amount of purchase and the transaction categories. Considering the categories and the amount of transactions, we can see that fraudulent transactions occur more in shopping net, shopping pos, misc net, and entertainment. Hence, we can say that shopping net and shopping pos can be more instances of fraud.

Descriptives Statistics

Out of the total records received, from the analysis performed, the proportion of the population that was defrauded is about 0.01% of the population. This implies that, the proportion of the population that was not defrauded is 99.4% of the population.

Building a Machine Learning Model

A KMeans classification model was built to make predictions whether a transaction was fraudulent or not. The model predictions accuracy score was about 87%.