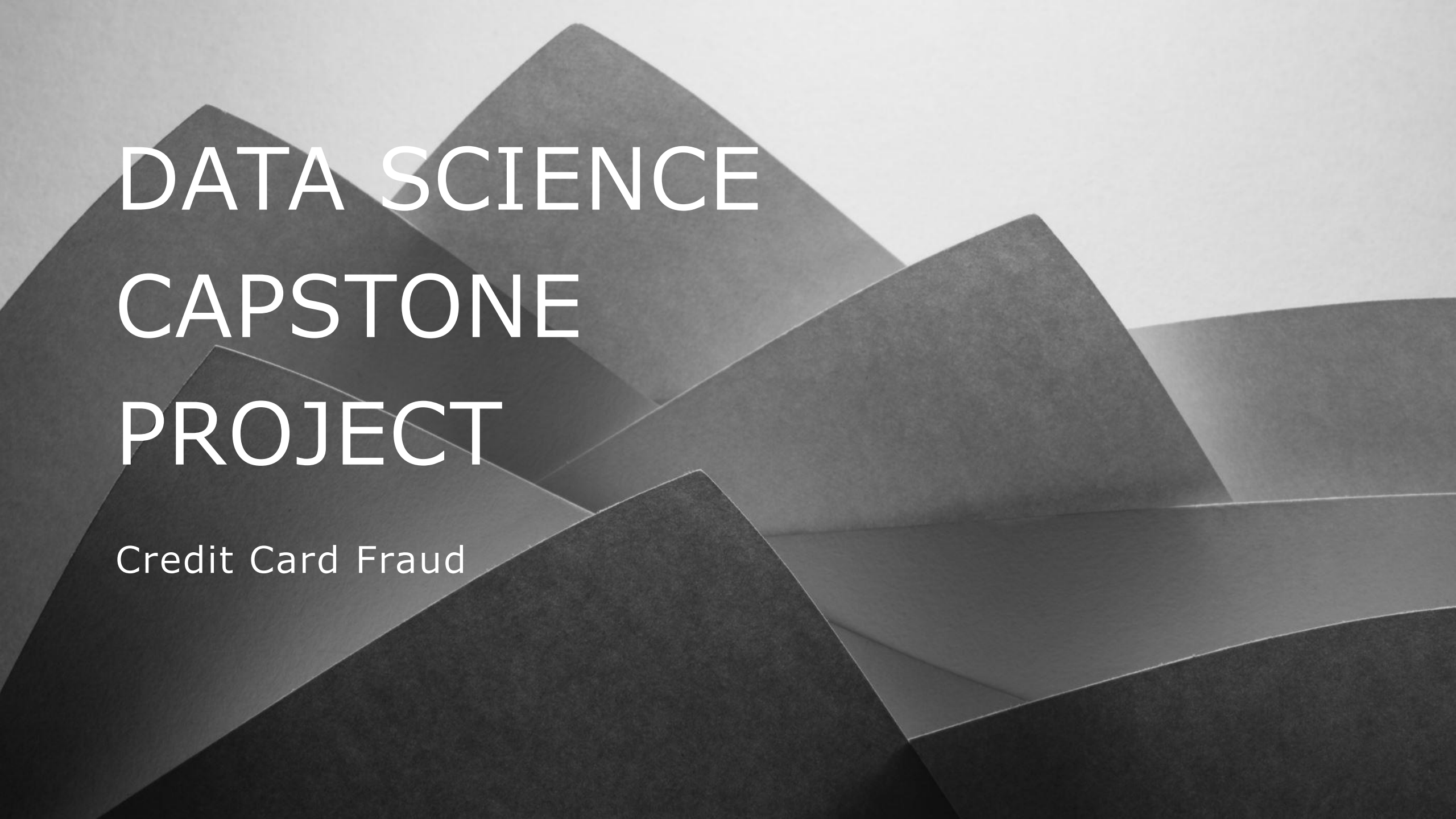


DATA SCIENCE CAPSTONE PROJECT

Credit Card Fraud

Problem Statement

More credit card payments are transacted in our daily life due to various reasons such as cashback benefit, going cashless, online purchase and installments. However, credit card transactions may be susceptible to fraud.

The objective of this project is to build a machine learning model for fraud prediction in credit card transactions.

Dataset Details

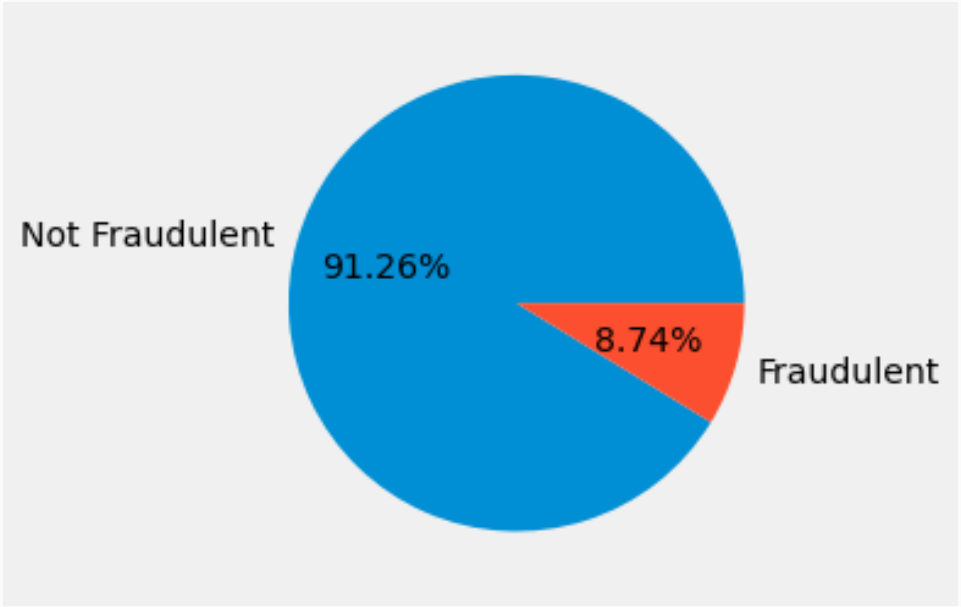
There are 1mio rows of data with 7 different categories for fraud detection.

Columns	Description	Type of Data
distance_from_home	the distance from home where the transaction happened.	Numerical
distance_from_last_transaction	the distance from last transaction happened.	Numerical
ratio_to_median_purchase_price	Ratio of purchased price transaction to median purchase price.	Numerical
repeat_retailer	Is the transaction happened from same retailer.	Categorical
used_chip	Is the transaction through chip (credit card).	Categorical
used_pin_number	Is the transaction happened by using PIN number.	Categorical
online_order	Is the transaction an online order.	Categorical

Datasets source: Credit Card Fraud | Kaggle

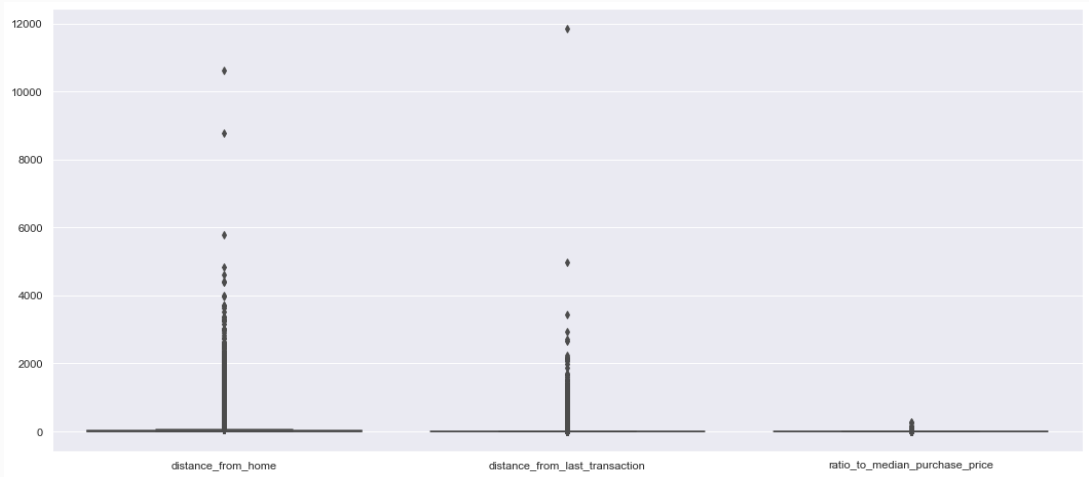
Data Exploration

From the dataset, total of 8.74% are suspected fraudulent cases.

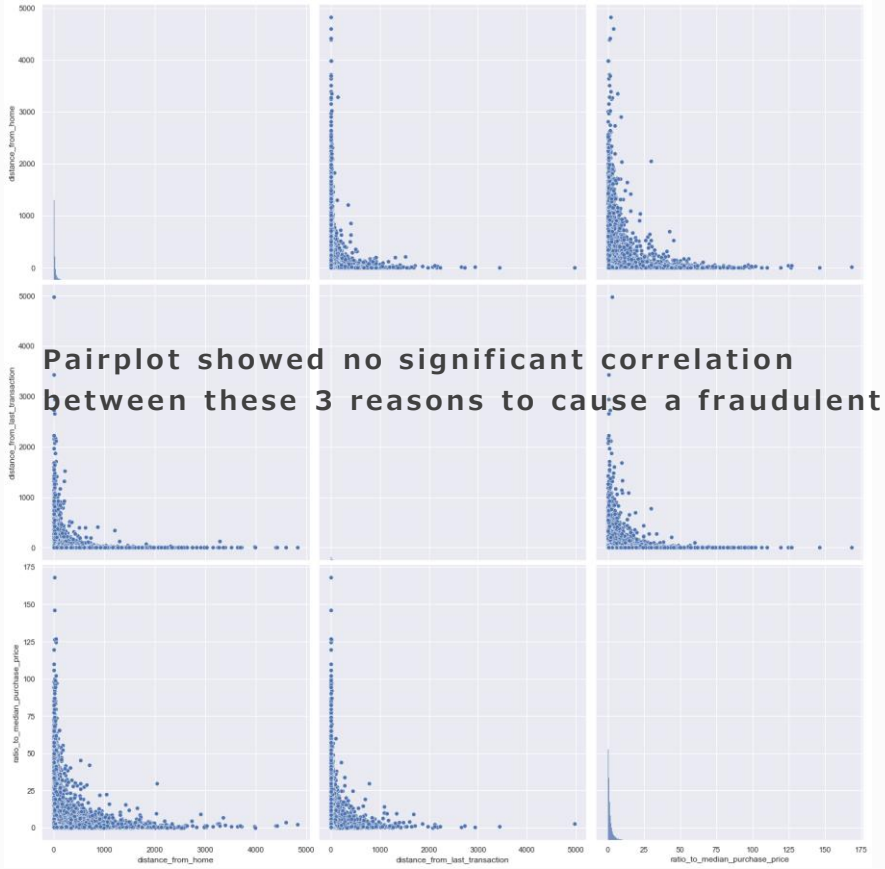


Data Exploration on Numerical Data Type

Investigation of outlier using boxplot, dropping outlier, and find correlation between numerical data type.



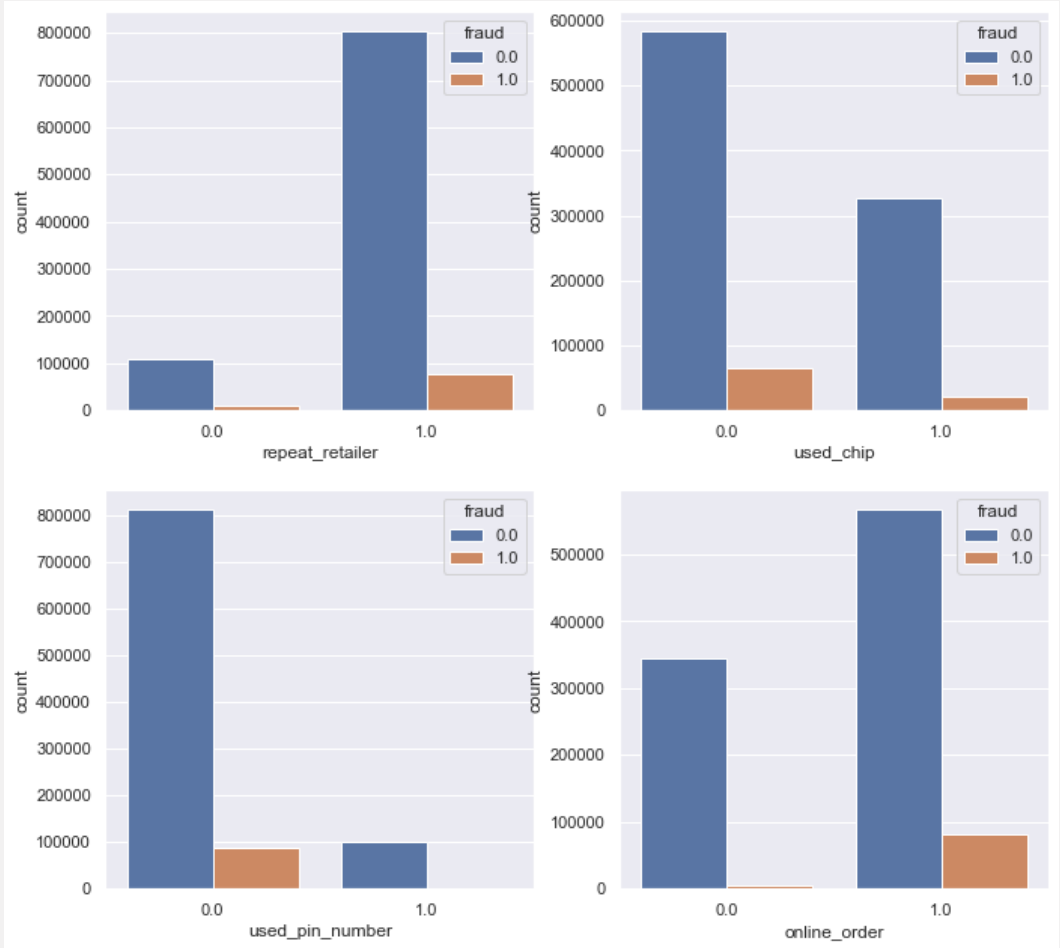
		Count	
Columns	Criteria	Fraudulent	Not Fraudulent
distance_from_home	>5000	1	2
distance_from_last_transaction	>5000	0	1
ratio_to_median_purchase_price	>200	1	1

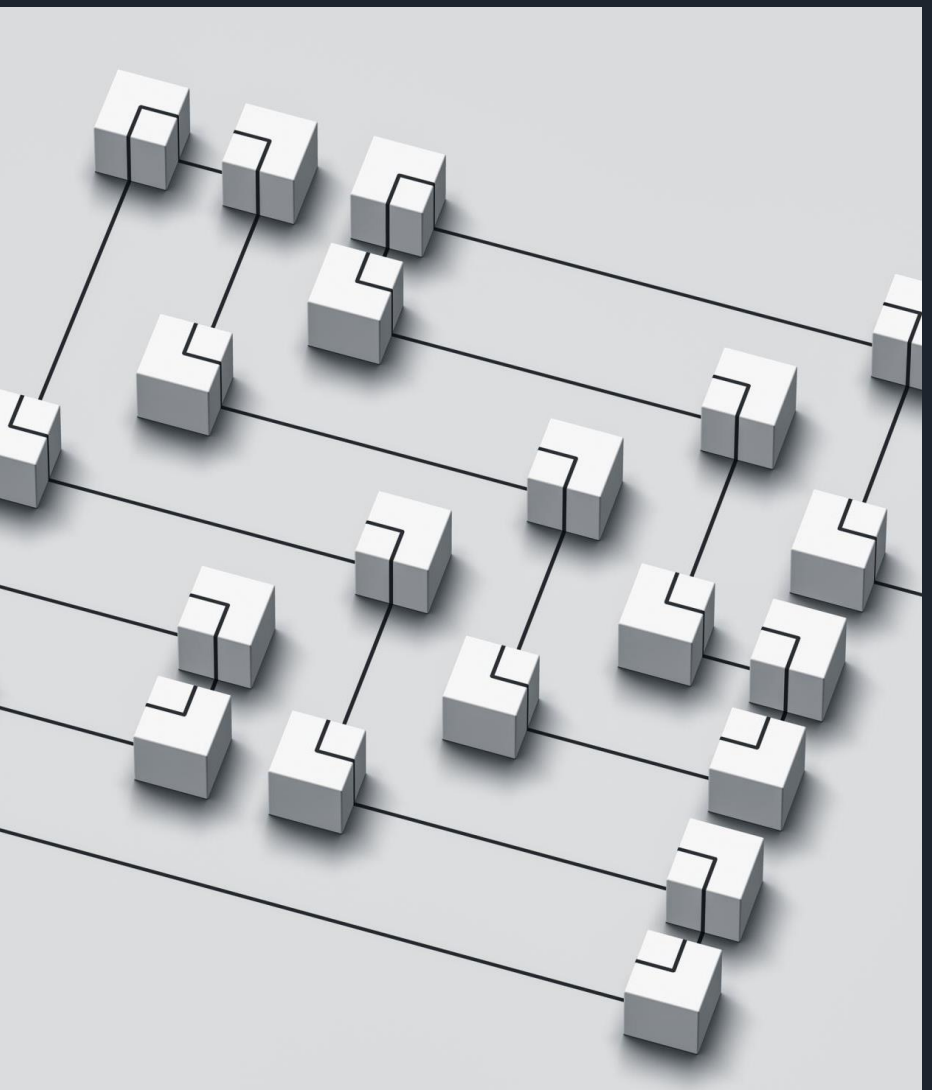


Pairplot showed no significant correlation between these 3 reasons to cause a fraudulent.

Data Exploration on Categorical Data Type

From the countplot, fraud occurred on repeat retailer, transaction without using chip, transaction without using pin number and online order.





Selection of Machine Learning Model

Since numerical data type does not have significant correlation to the fraudulent, the machine learning model I adopt is Logistic Regression due to it is more accurate to predict categorical data type (4 columns of data are categorical. Logistic Regression model is compared with Decision Tree.

Machine Learning Model Results Comparison

Comments:

Both machine learning model gave similar result (AUC close to 1.0). Decision Tree is a better classifier for fraud prediction.

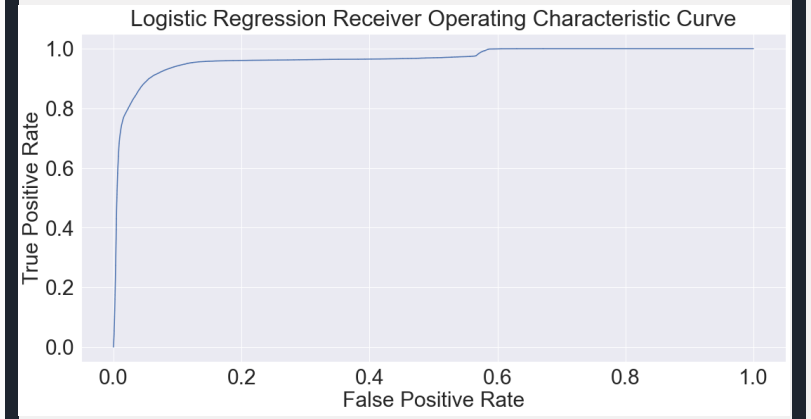
Logistic Regression Model Result

```
LogisticRegression()

Accuracy:  0.9586031953439845
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98    273877
         1.0       0.89      0.60      0.71     26122

    accuracy                           0.96    299999
   macro avg       0.93      0.79      0.85    299999
weighted avg       0.96      0.96      0.95    299999
```



AUC: 0.965

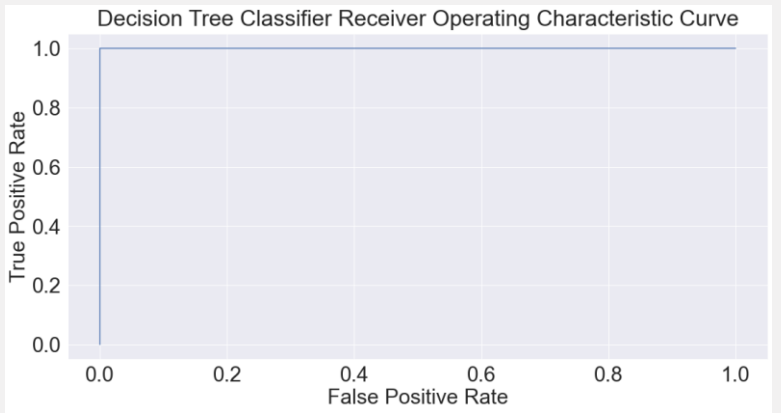
Decision Tree Model Result

```
DecisionTreeClassifier()

Accuracy:  0.9999899999666666
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00    273877
         1.0       1.00      1.00      1.00     26122

    accuracy                           1.00    299999
   macro avg       1.00      1.00      1.00    299999
weighted avg       1.00      1.00      1.00    299999
```



AUC: 0.999