greatlearning

Fraud Detection using Supervised Learning Classification

Mentor Subramanian Palamarneri

Team Members

Ajay Krishnan, Babloo Kumar S, Panduranga A, Rabi Bastin J, Gopi Selvaraj

Problem Definition

- The objective of this project is to develop a robust and accurate fraud detection model using supervised learning classification techniques. The model should be able to classify transactions or activities into two categories: "fraudulent" and "non-fraudulent," based on historical data and relevant features.
- The dataset provided for this project contains historical transaction data, where each transaction is described by a set of features. These features includes such as transaction amount, transaction date and time, user information, and other relevant details. Additionally, each transaction is labeled as either "fraudulent" or "non-fraudulent."
- References:
- https://www.fraud.com/post/the-history-and-evolution-of-fraud (Background Research)
- https://www.amygb.ai/blog/how-fraud-detection-works-in-banking (Application)
- https://www.sciencedirect.com/science/article/pii/S0957417421017164 (Past Reasearch)
- https://ieeexplore.ieee.org/document/10085493 (Ongoing Research)

About Dataset

greatlearning

S.No	Field name	Description	Data type
1	Accountnumber	Unique identifier for the account associated with the transaction.	Float64
2	Customerid	Unique identifier for the customer associated with the account.	Float64
3	Creditlimit	The credit limit assigned to the account.	Float64
4	Availablemoney	The available balance in the account.	Float64
5	Transactiondatetime	Date and time of the transaction.	Object
6	Transactionamount	The amount of money involved in the transaction.	Float64
7	Merchantname	Name of the merchant where the transaction took place.	Object
8	Acqcountry	Country where the acquiring bank is located.	Object
9	Merchantcountrycode	Country code of the merchant's location.	Object
10	Posentrymode	Point of service (POS) entry mode for the transaction.	Int64
11	Posconditioncode	Condition of the POS at the time of the transaction.	Int64
12	Merchantcategorycode	Code indicating the category of the merchant.	Object
13	Currentexpdate	Expiration date of the card at the time of the transaction.	Object
14	Accountopendate	Date when the account was opened.	Object
15	Dateoflastaddresschange	Date of the last address change on the account.	Object
16	Cardcvv	CVV (card verification value) of the card.	Int64
17	Enteredcvv	CVV entered during the transaction.	Int64
18	Cardlast4digits	Last 4 digits of the card number.	Int64
19	Transactiontype	Type of the transaction (e.G., Purchase, cash advance).	Object
20	Echobuffer	An echo buffer associated with the transaction.	Object
21	Currentbalance	Current balance in the account.	Float64
22	Merchantcity	City where the merchant is located.	Object
23	Merchantstate	State where the merchant is located.	Object
24	Merchantzip	ZIP code of the merchant's location.	Object
25	Cardpresent	Indicator whether the card was present during the transaction.	Bool
26	Posonpremises	Indicator whether the POS was on the merchant's premises.	Object
27	Recurringauthind	Indicator for recurring authorization.	Object
28	Expirationdatekeyinmatch	Indicator whether the expiration date matches during key-in transactions.	Bool
29	Isfraud	Indicator whether the transaction is fraudulent.	Bool

Shape and Spread of Target Variable

- The Dataset has **786363** Observation & **29** Variables, The Complexity in finding the solution to the problem based on the Chosen Sampling Techniques (**Stratified & Smote Sampling**)

- When Calculating the Proportion to Target Variables Major Class of Fraud Transaction is False Minor Class of Fraud Transaction is True with ratio of **98.5 : 1.5**

- **Python Version:**

  '3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)]'

- **Dataset**

  Reference: https://www.kaggle.com/datasets/iabhishekbhardwaj/fraud-detection

greatlearning

Pre Processing Data Analysis

Variables Name	Zero Values	Missing Values	% of Total Values	Total Zero & Missing Values	% Total Zero & Missing Values	Data Type
echoBuffer	0	786363	100	786363	100	float64
merchantCity	0	786363	100	786363	100	float64
merchantState	0	786363	100	786363	100	float64
merchantZip	0	786363	100	786363	100	float64
posOnPremises	0	786363	100	786363	100	float64
recurringAuthInd	0	786363	100	786363	100	float64
acqCountry	0	4562	0.6	4562	0.6	object
posEntryMode	0	4054	0.5	4054	0.5	float64
merchantCountry Code	0	724	0.1	724	0.1	object
transactionType	0	698	0.1	698	0.1	object
posConditionCode	0	409	0.1	409	0.1	float64

Data Preparation

- Only **posEntryMode & posConditionCode** are in the from discrete number however the data types is in float hence changed to Object.

POS Entry Mode	Description
2	PAN auto-entry via magnetic stripe
5	PAN auto-entry via chip
9	PAN entry via electronic commerce, including remote chip
80	Chip card at chip-capable terminal was unable to process transactions using data on the chip–terminal defaulted to magnetic stripe-read PAN
99	PAN auto entry via magnetic stripe-full track data has been read without alteration or truncation

POS Condition Code	Description
1	**Sale:** This code indicates that a sale transaction has been processed. The merchant must collect the customer's signature for this type of transaction.
8	**Void:** This code indicates that a previously processed sale transaction has been voided. The merchant must not collect the customer's signature for this type of transaction.
99	**Refund:** This code indicates that a previously processed sale transaction has been refunded. The merchant must collect the customer's signature for this type of transaction.

Reference:
https://developer.mastercard.com/fld-fraud-submission/documentation/parameters/annexure-1/#table-8-pos-entry-mode1%20%E2%80%93

Dropping Variables

- Dropping columns with 100% null values, which includes

echoBuffer	MerchantZip
merchantCity	PosOnPremises
MerchantState	RecurringAuthInd.

- Excluding the following variables from the dataset as they are irrelevant to the target and may introduce confusion during model building:

merchantName	cardCVV
accountNumber	enteredCVV
customerId	currentExpDate
cardLast4Digits	dateOfLastAddressChange
transactionDateTime	accountOpenDate

Missing Value Imputation

Method 1

The Most Common Method In the Field of Data Science (Imputation Using **Mode: Highest Frequency Observed Values**)

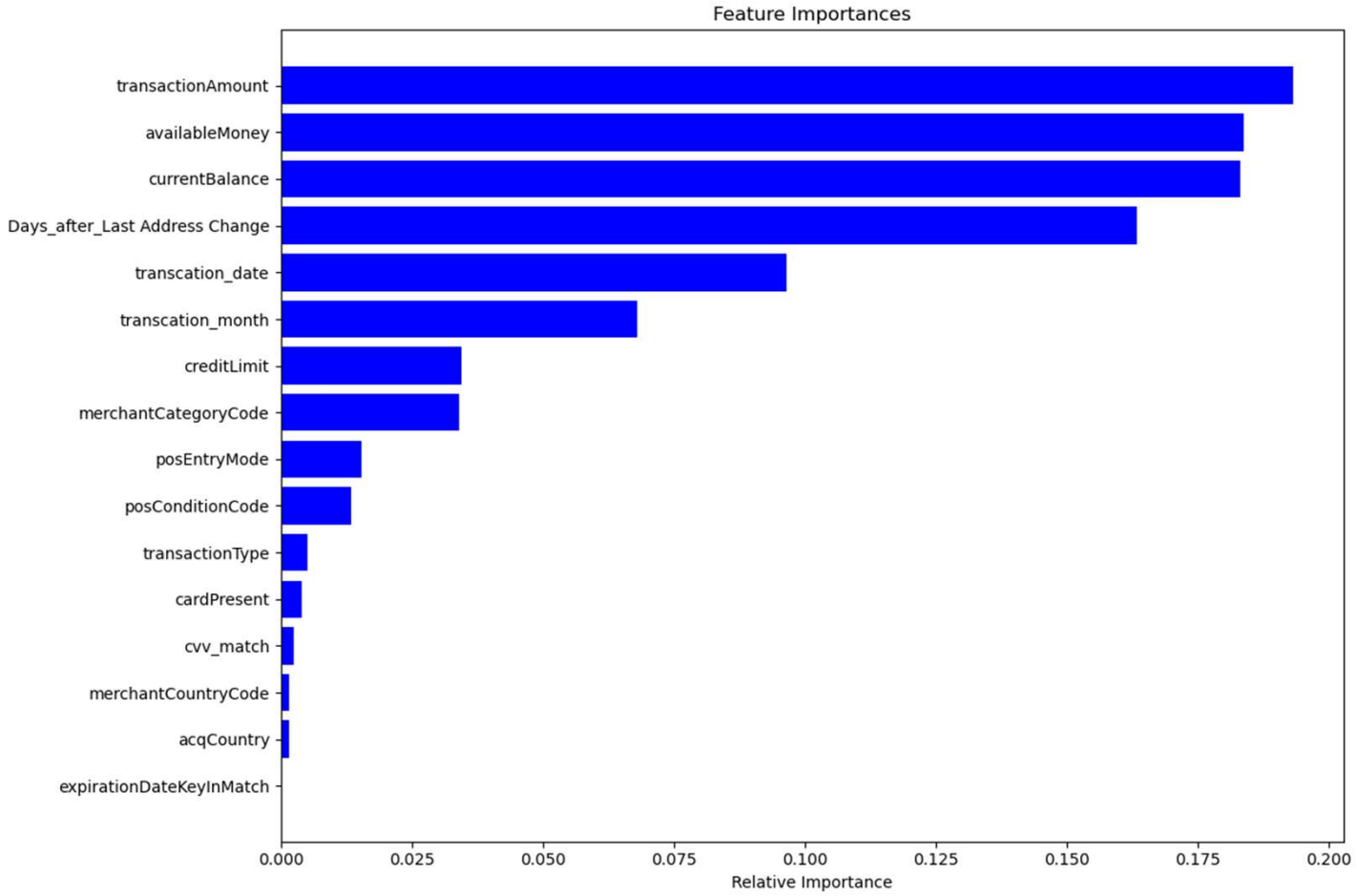
Method 2

Using Prediction Method, Random Forest Classification

Method 3

Removing Missing Values Observation from the Data Set

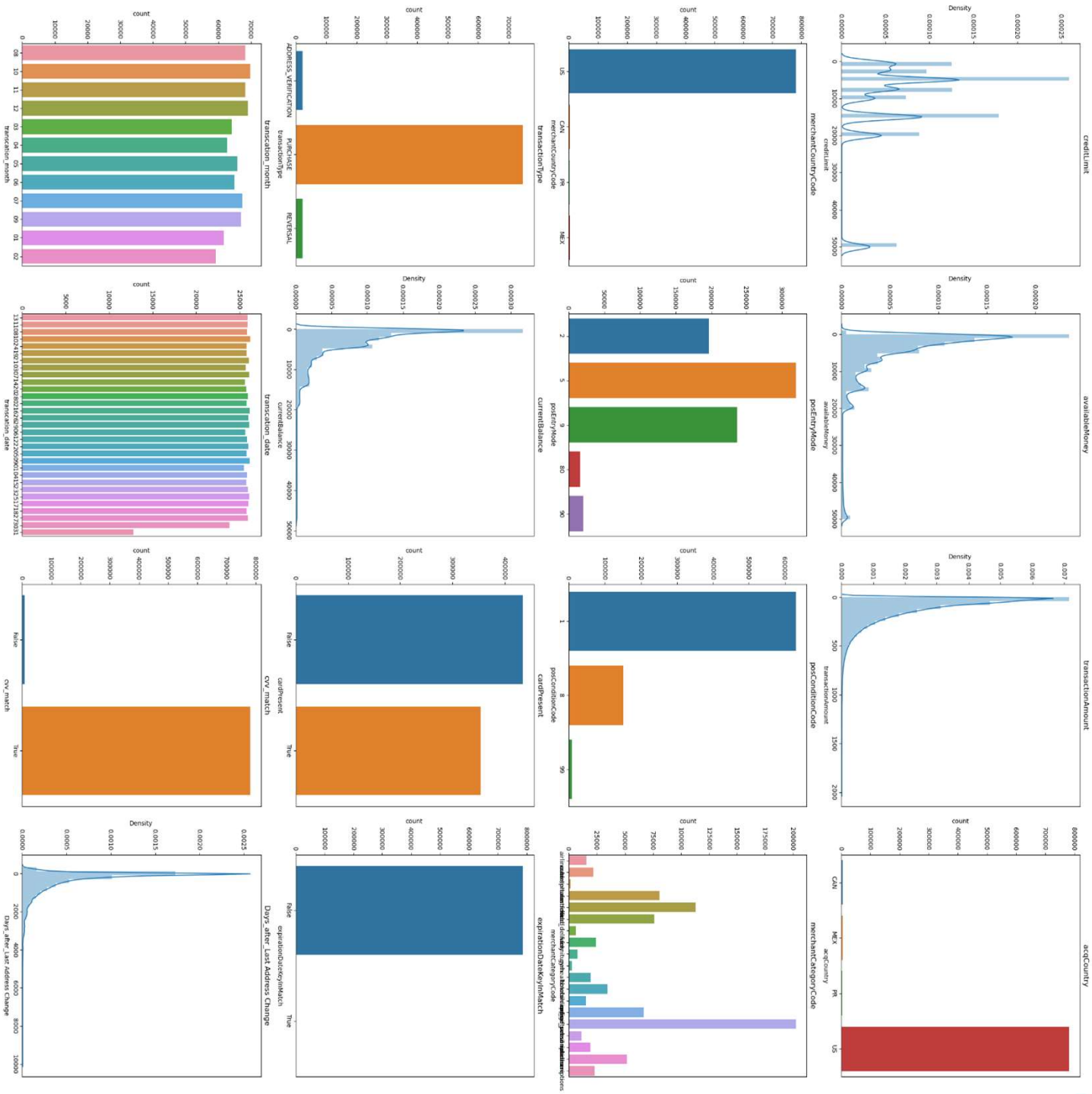
Feature Importance Using Random Forest Classification before Analysing



Exploratory Data Analysis

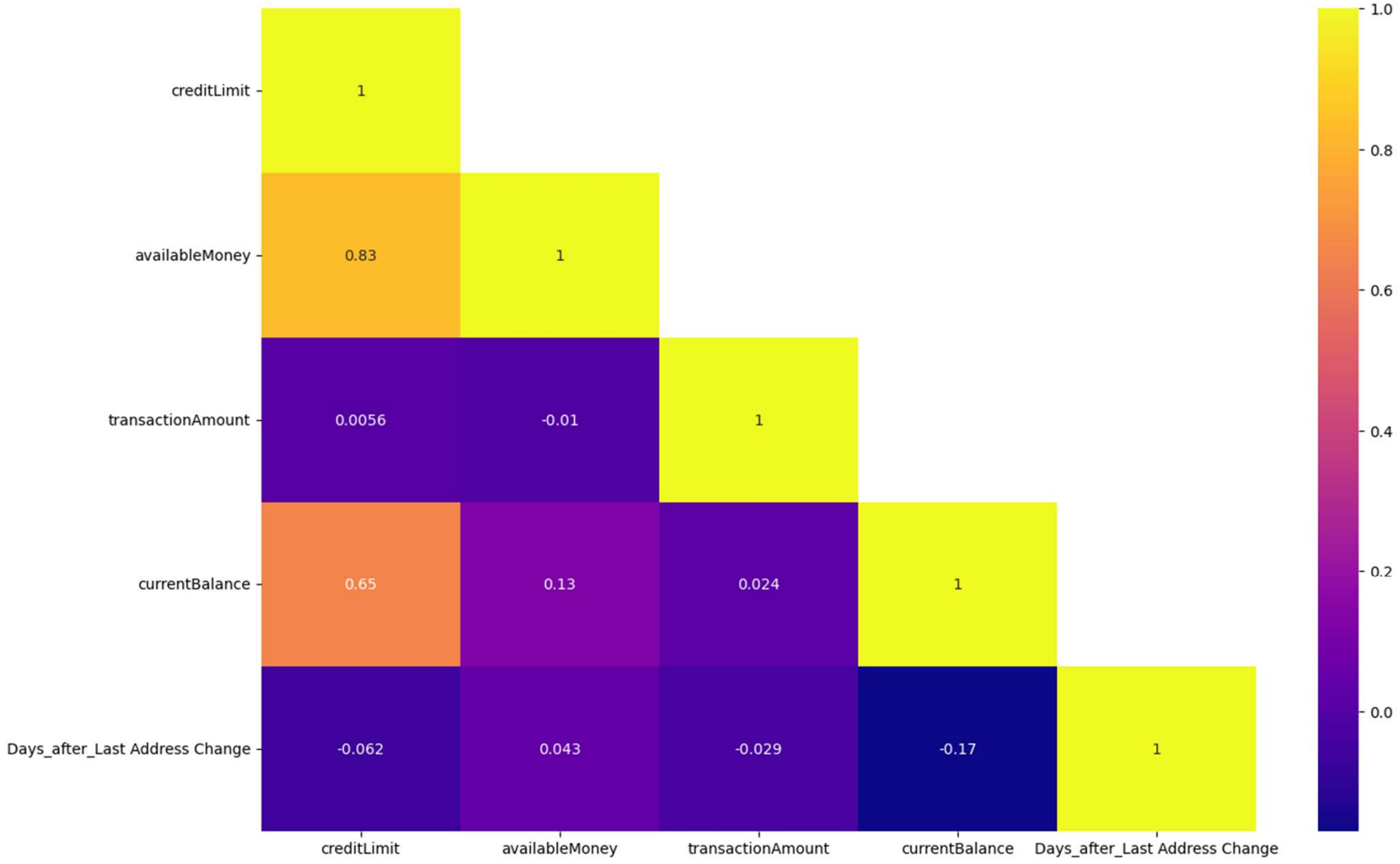
- Univariate Analysis
- Bivariate Analysis
- Multi- Variate Analysis

Univariate Analysis



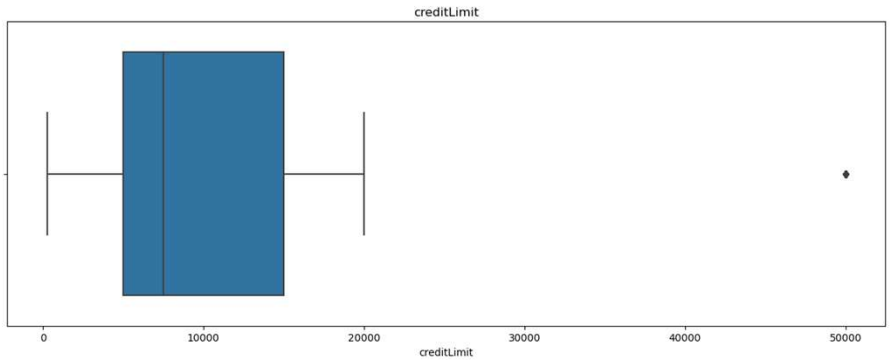
Univariate Numerical Analysis (Correlation Matrix of Heatmap)

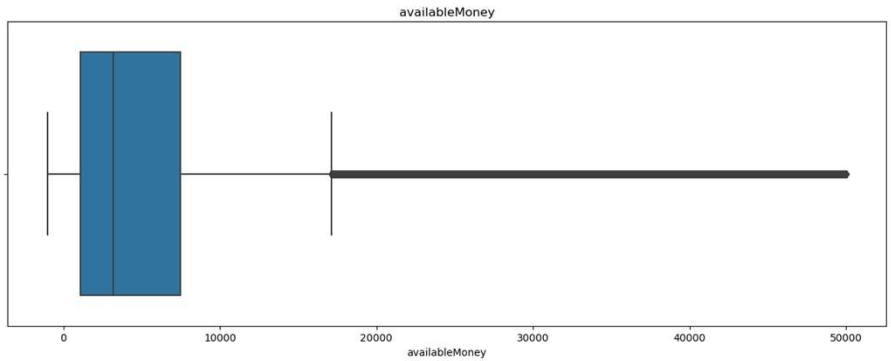
- Multi-Collinearity Between Variables Between Credit Limit to Current Balance & Available Money, However we Can Ignore As bank requires this Variables in Order to Prevent Fraud and to Avoid Risk to the Bank & Customer

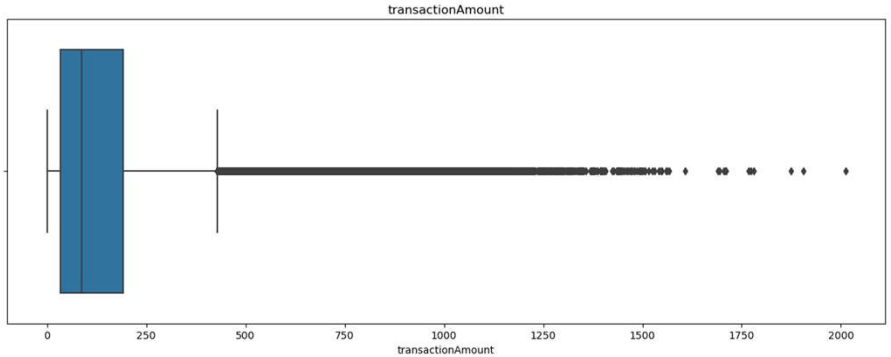


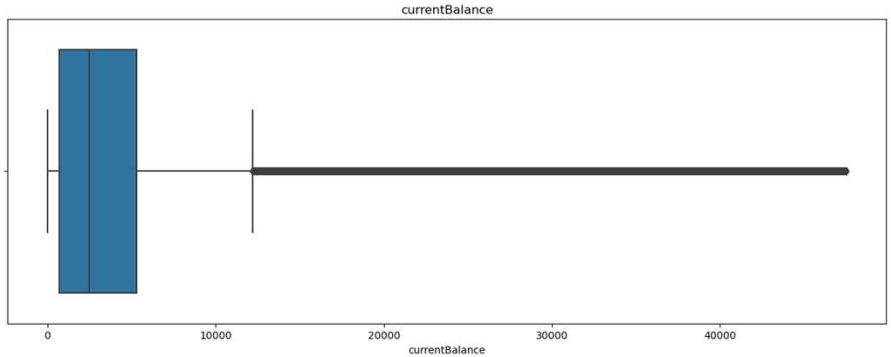
greatlearning

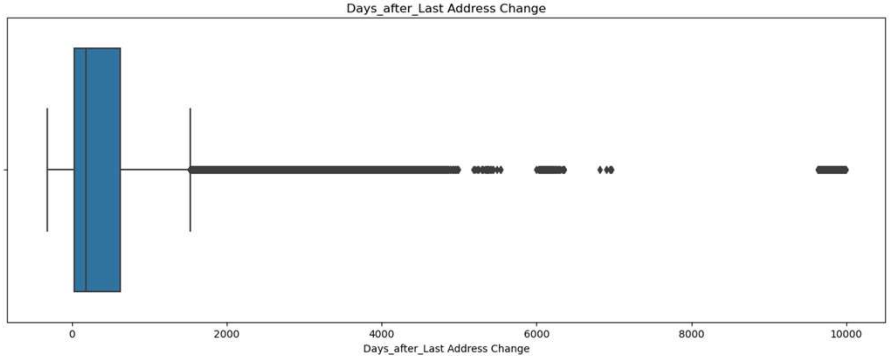
Univariate Numerical Analysis (Outlier Detection)

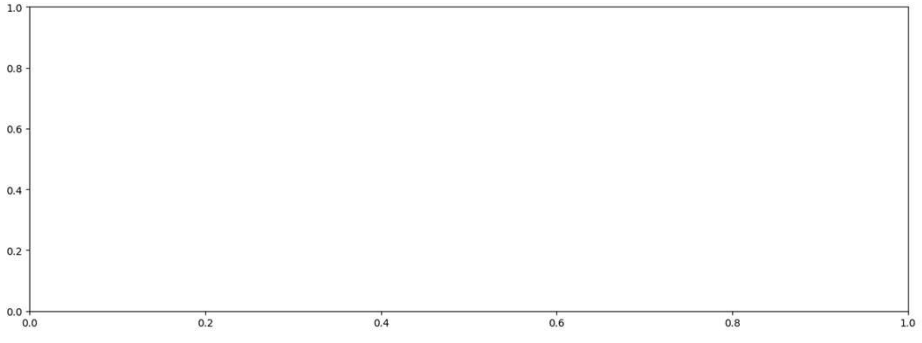




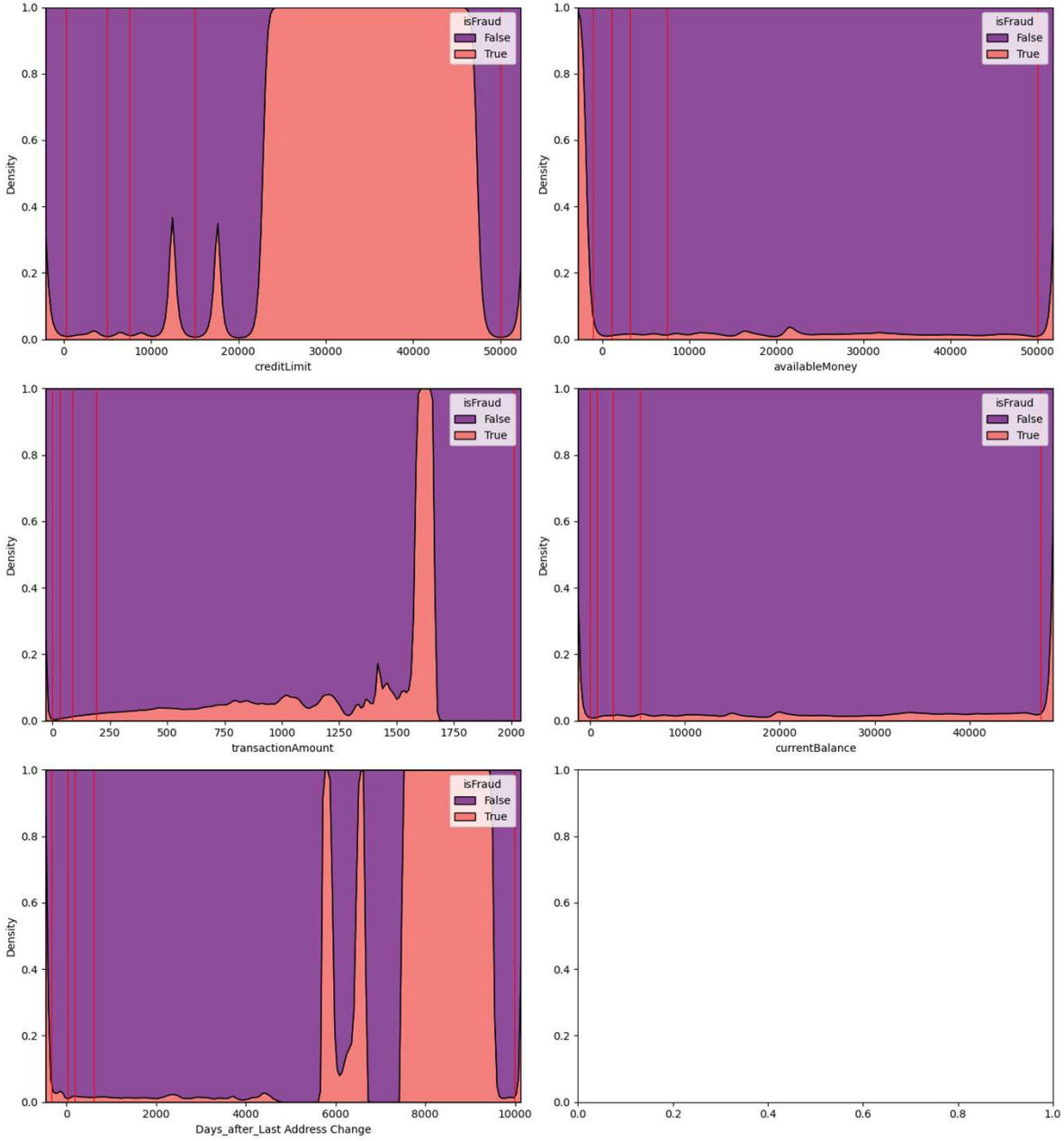








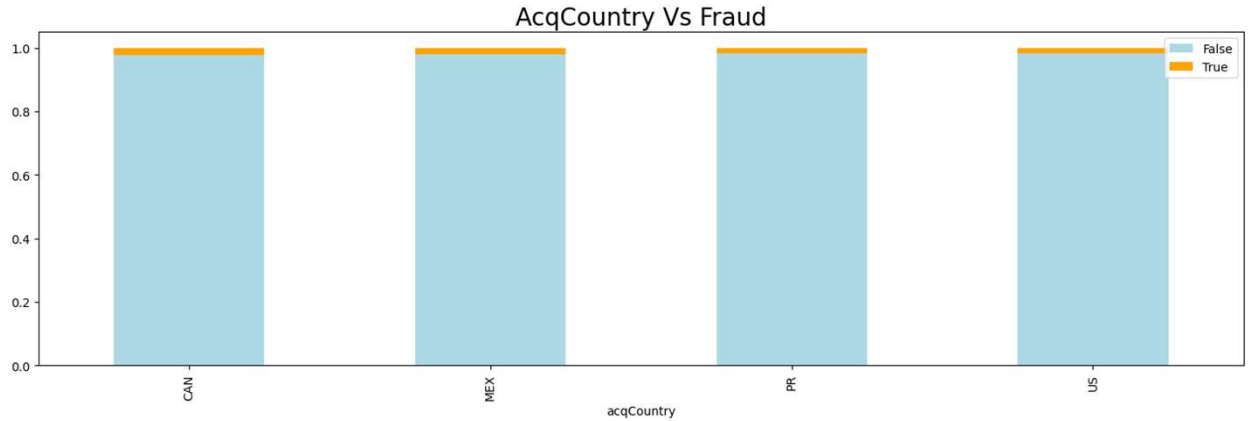
Univariate Numerical Analysis (Distribution of Numbers with Distplot)

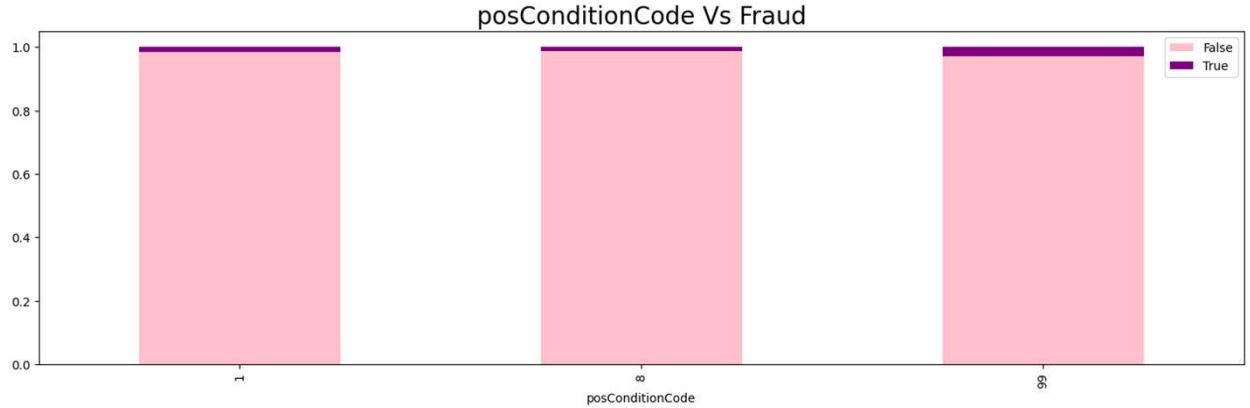


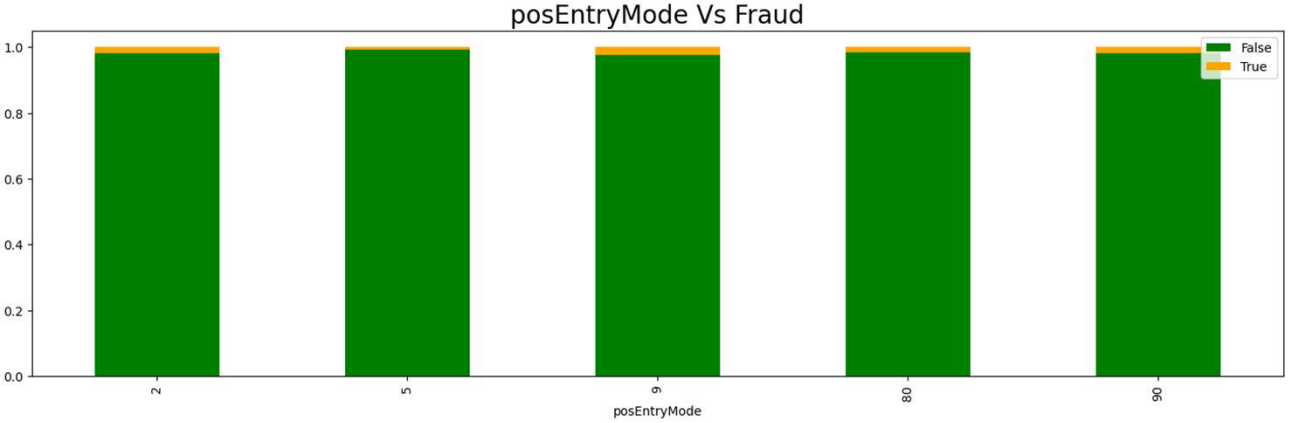
Performance Statistical Test, **Chi-Square test of Independence** for **Categorical** vs **Target Variables** assuming 5% level of significance.

- P >= 0.05: Hypothesis Accepted
- P < 0.05: Hypothesis Rejected

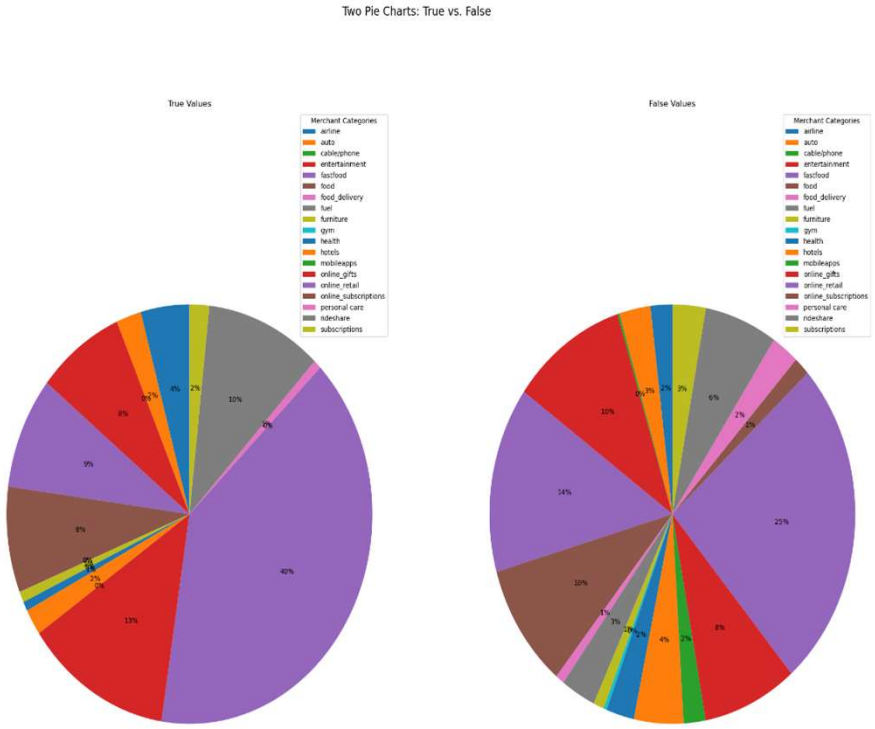
Column	p-value chi2_contingency	Result
acqCountry	0.0067	There is significant association between the two variables and we reject the null hypothesis
merchantCountryCode	0.0049	There is significant association between the two variables and we reject the null hypothesis
posEntryMode	0.0000	There is significant association between the two variables and we reject the null hypothesis
posConditionCode	0.0000	There is significant association between the two variables and we reject the null hypothesis
merchantCategoryCode	0.0000	There is significant association between the two variables and we reject the null hypothesis
transactionType	0.0000	There is significant association between the two variables and we reject the null hypothesis
cardPresent	0.0000	There is significant association between the two variables and we reject the null hypothesis
expirationDateKeyInMatch	0.4605	There is no significant association between the two variables and we fail to reject the null hypothesis
isFraud	0.0000	There is significant association between the two variables and we reject the null hypothesis
transcation_month	0.0000	There is significant association between the two variables and we reject the null hypothesis
transcation_date	0.6952	There is no significant association between the two variables and we fail to reject the null hypothesis
cvv_match	0.0000	There is significant association between the two variables and we reject the null hypothesis


AcqCountry Vs Fraud
False
True
1.0
0.8
0.6
0.4
0.2
0.0
CAN
MEX
PR
US
acqCountry


posConditionCode Vs Fraud
False
True
1.0
0.8
0.6
0.4
0.2
0.0
1
8
99
posConditionCode

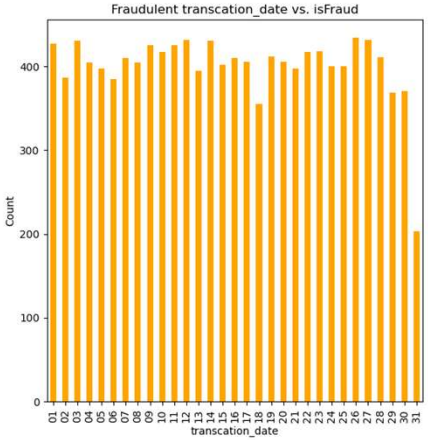

posEntryMode Vs Fraud
False
True
1.0
0.8
0.6
0.4
0.2
0.0
2
5
9
80
90
posEntryMode

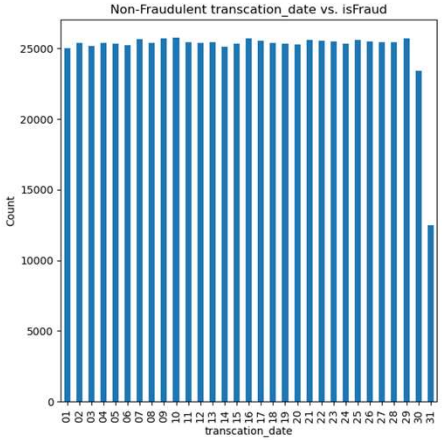
Merchant Category Code vs is Fraud (Target Variable)

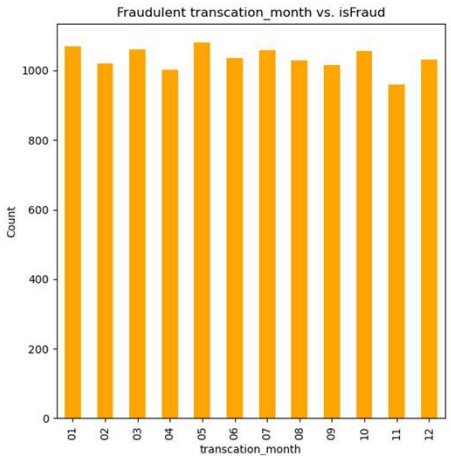


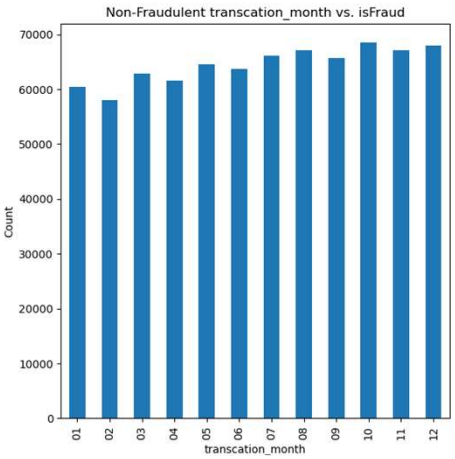
	Category	Merchant Fraud Percentage (True)	Merchant Fraud Percentage (False)
0	Airline	4%	2%
1	Auto	2%	3%
2	Cable/Phone	0%	0%
3	Entertainment	8%	10%
4	Fastfood	9%	14%
5	Food	8%	10%
6	Food Delivery	0%	1%
7	Fuel	0%	3%
8	Furniture	1%	1%
9	Gym	0%	0%
10	Health	1%	2%
11	Hotels	2%	4%
12	Mobileapps	0%	2%
13	Online_Gifts	13%	8%
14	Online_Retail	40%	25%
15	Online_Subscriptions	0%	1%
16	Personal Care	1%	2%
17	Rideshare	10%	6%
18	Subscriptions	2%	3%

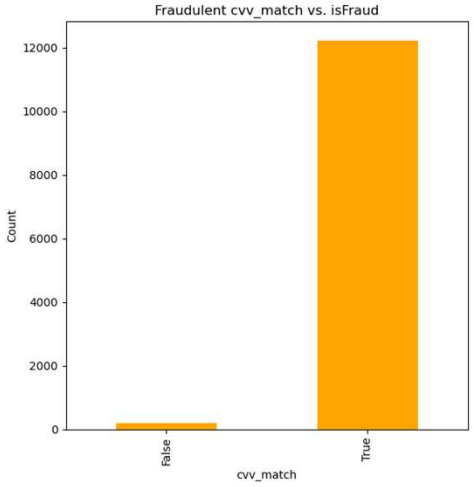
Transaction Month & Year , Card Expiry Month & Year CVV Match Vs Is Fraud (Target Variable)

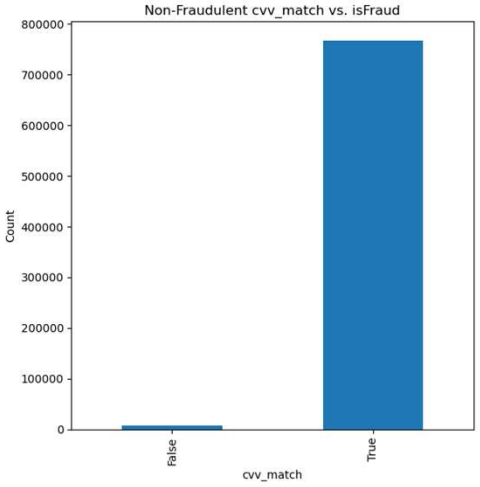


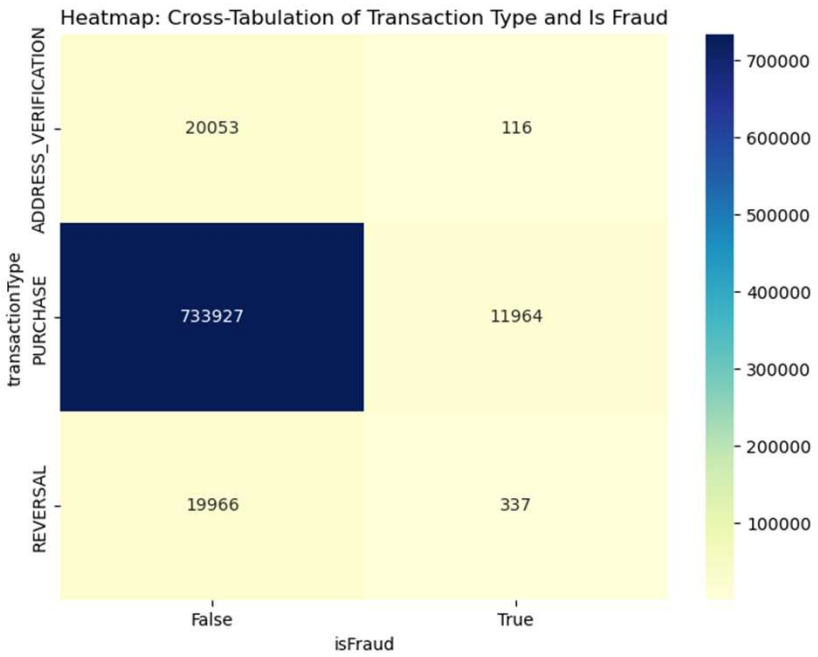






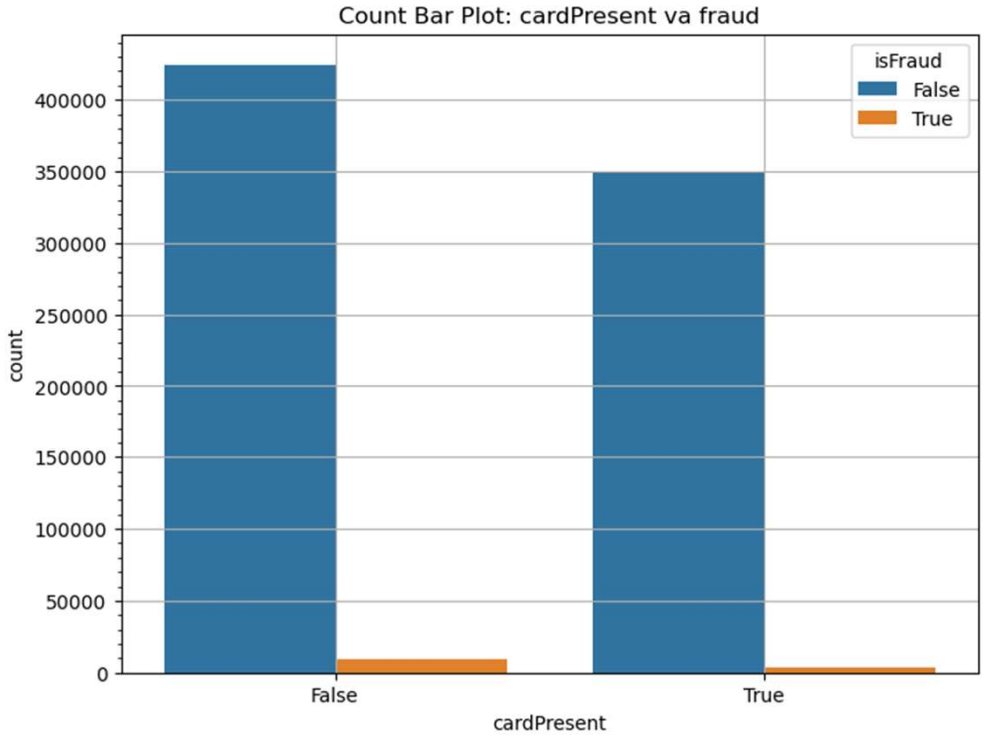






Transaction Type Vs Is Fraud (Target Variable)

Card Present Vs Is Fraud (Target Variable)

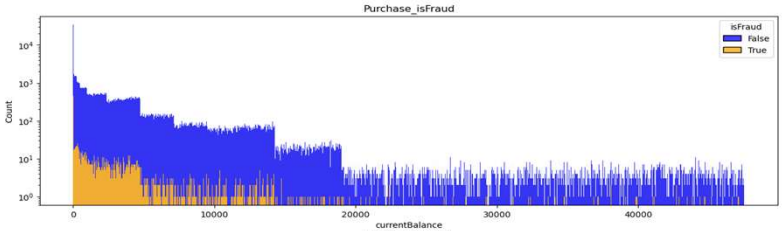


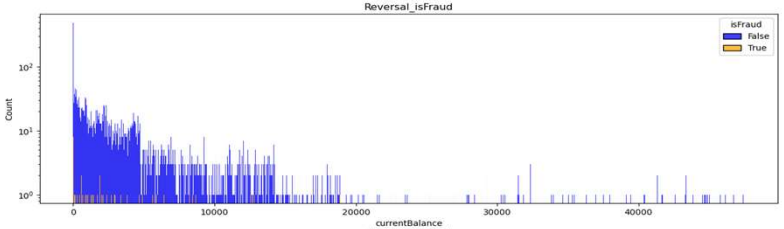
Multivariate Statistical Test & Method Followed

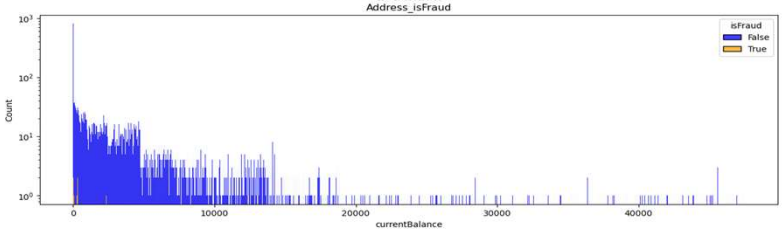
Anova & Chi2_Contigency Test : We Assumed The Data Distribution is normal and Performed these parametric Statistical Test

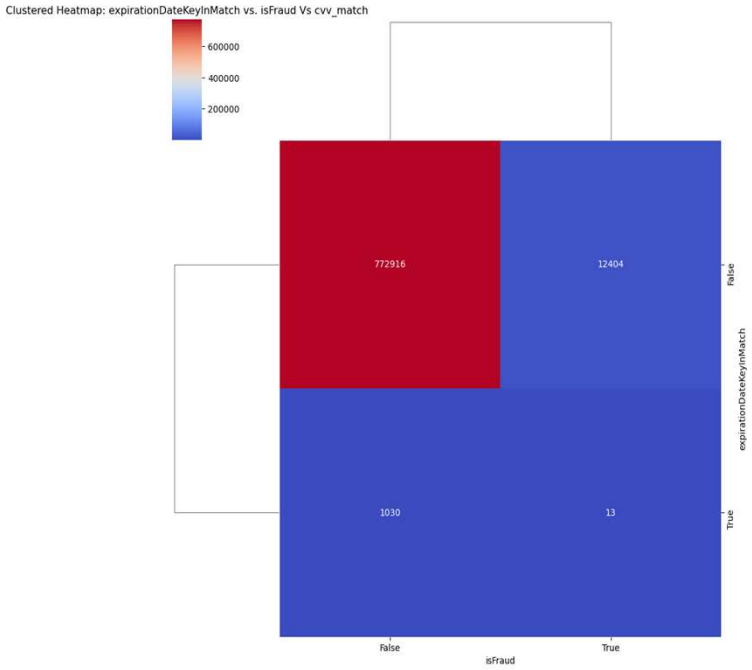
- P >= 0.05: Hypothesis Accepted
- P < 0.05: Hypothesis Rejected

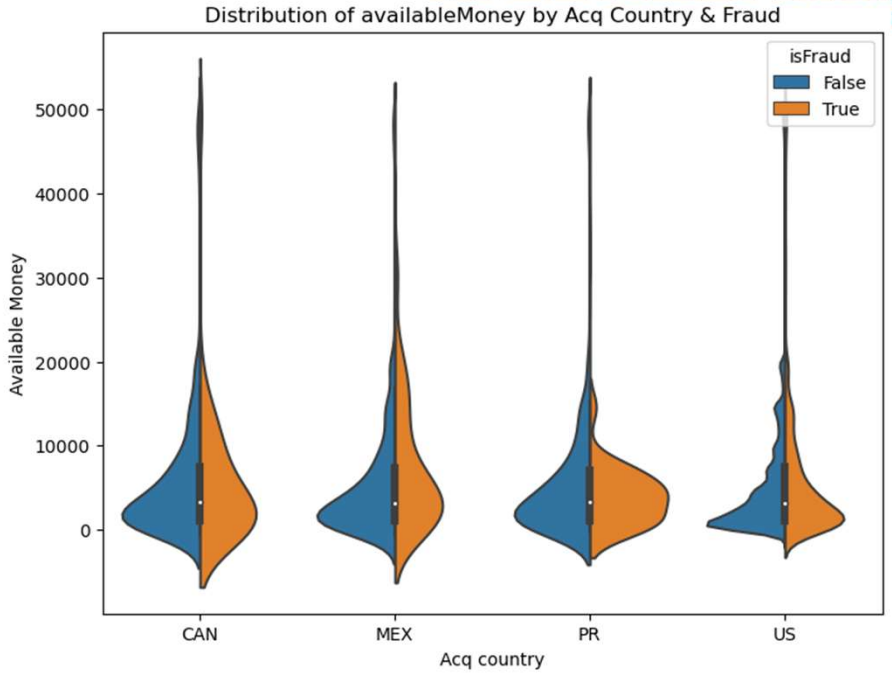
Name_of_Stat	Type of Test	P-Value	Result
Transaction Type Vs Current Balance Vs Fraud	Anova One Way Test	0.0000	There is significant association among the variables and we reject the null hypothesis
Expiration Date Key In Match Vs cvv_match Vs Fraud	Chi2_ Contingency Test	0.8218	There is no significant association among the variables variables and we fail to reject the null hypothesis
Transaction Amount Vs Acq Country Vs Fraud	Anova One Way Test	0.0000	There is significant association among the variables and we reject the null hypothesis
Available Money Vs Acq Country Vs Fraud	Anova One Way Test	0.1077	There is no significant association among the variables and we fail to reject the null hypothesis


Purchase_isFraud
isFraud
False
True
Count
currentBalance


Reversal_isFraud
isFraud
False
True
Count
currentBalance


Address_isFraud
isFraud
False
True
Count
currentBalance


Clustered Heatmap: expirationDateKeyInMatch vs. isFraud Vs cvv_match
600000
400000
200000
772916
12404
1030
13
False
True
expirationDateKeyInMatch
isFraud


Distribution of availableMoney by Acq Country & Fraud
isFraud
False
True
Available Money
Acq country
CAN
MEX
PR
US


Mean transaction amount by Acq country and Fraud
isFraud
False
True
Mean Transaction Amount
Acq country
CAN
MEX
PR
US

Feature Engineering	
1. Transaction Date Time	Transaction Month & Date
2. Card Cvv & Entered CVV	CVV Match (**True** 779348 & **False** 7015**)**
3. Transaction Date Time & Date of Last Address Change	Difference Days

Scaling The Data

- We Have Scaled the numerical variables so the model understands the pattern more efficiently and also to reduce the skewness and brings the numerical value under one scale.

- Alternatively, in future phases we are going scale the data after spilt test and train and check for the performance of the model.

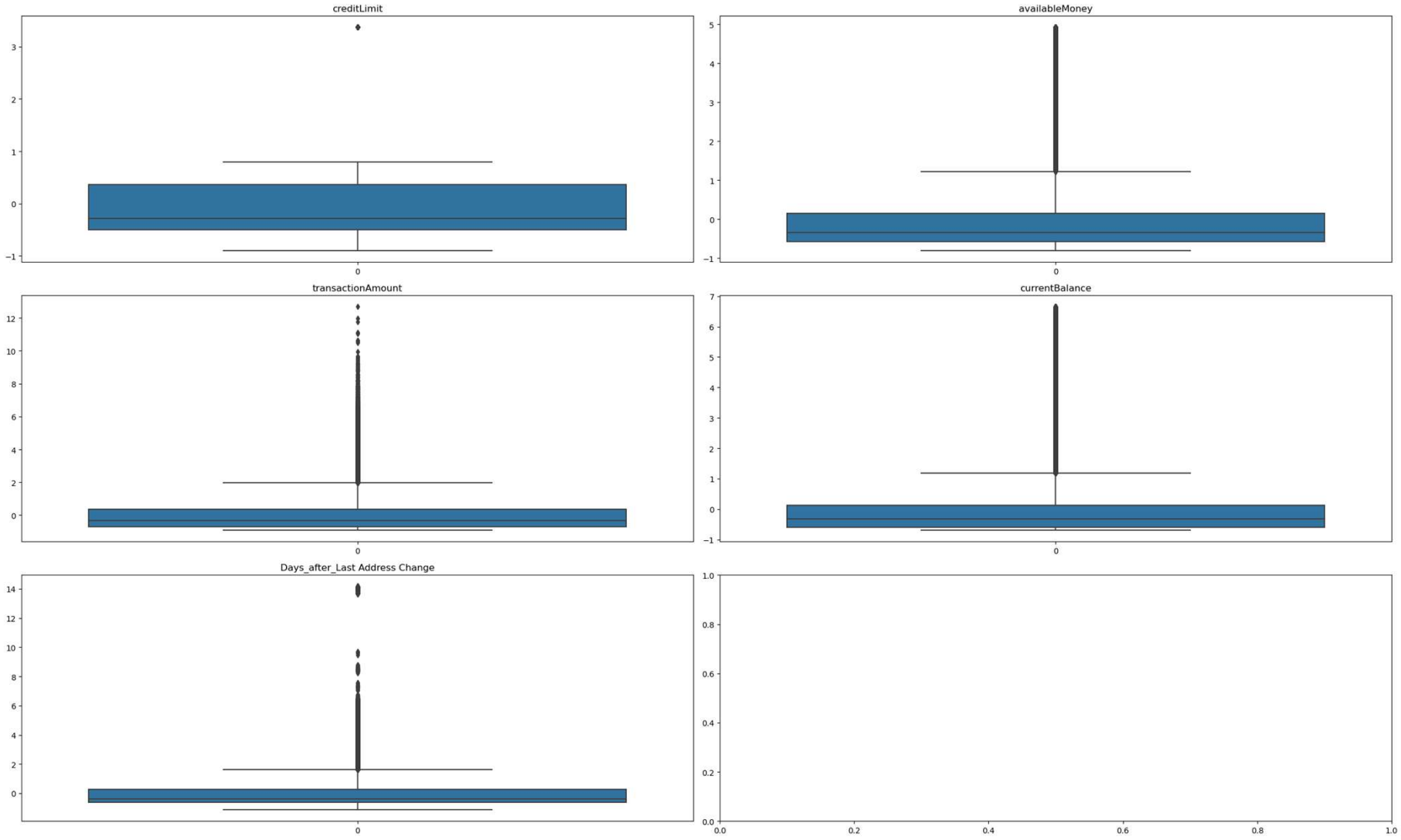
Feature selection

- We will Perform this Method During Model Building as Part of Combination Top Performing Feature & Least Performing Feature

- We will use Feature Selection Forward, Feature Selection Backward & Recursive Feature Selection as part of Model Building

greatlearning

After Scaling the Numerical Variables.



Encoding the Variables

Label Encoding

acqCountry	merchantCategoryCode	merchantCountryCode	transactionType	cardPresent	expirationDateKeyInMatch	isFraud	cvv_match
3	14	3	1	0	0	0	1
3	14	3	1	0	0	0	1
3	14	3	1	0	0	0	1
3	14	3	1	0	0	0	1
3	14	3	1	0	0	0	1

Model Building & Evaluation

- The Modeling is the core of any machine learning project. This step is responsible for the results that should satisfy or help satisfied the project goals.

- Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data.

- Our problem statement come under the Classification thus we have decided to use various models namely

1. Logistic Regression Model
2. Decision Tree Model
3. Random Forest Model
4. KNN Model
5. Naïve Bayes Model
6. Ada Boost Technique
7. XG Boost Technique
8. Gradient Boosting Technique

- **Evaluation**: Recall, Precision, Accuracy & F1 Score

Stratified Sampling & SMOTE Techniques

greatlearning

- Since our Data Set is Very Large classification data is an imbalanced data, it is desirable to sample the dataset into Sampling Datasets in a way that preserves the same proportions of examples in each class as observed in the Original Dataset. **This is called a Stratified Sampling**.
- We can achieve this by setting the **"Stratify"** argument to the **Merchant Category Code** component of the original dataset.
- Once the Sample is taken from various random state of 3 this Sample will be used for the various model with various combination, various ensemble and Stacking Techniques
- Smote Has Been Done on Train Data Only and Test is Passed for Evaluation Metrics

Population Target Variable Proportion		Sample Target Variable Proportion	
False	98.42096	False	98.1009
True	1.579042	True	1.899098

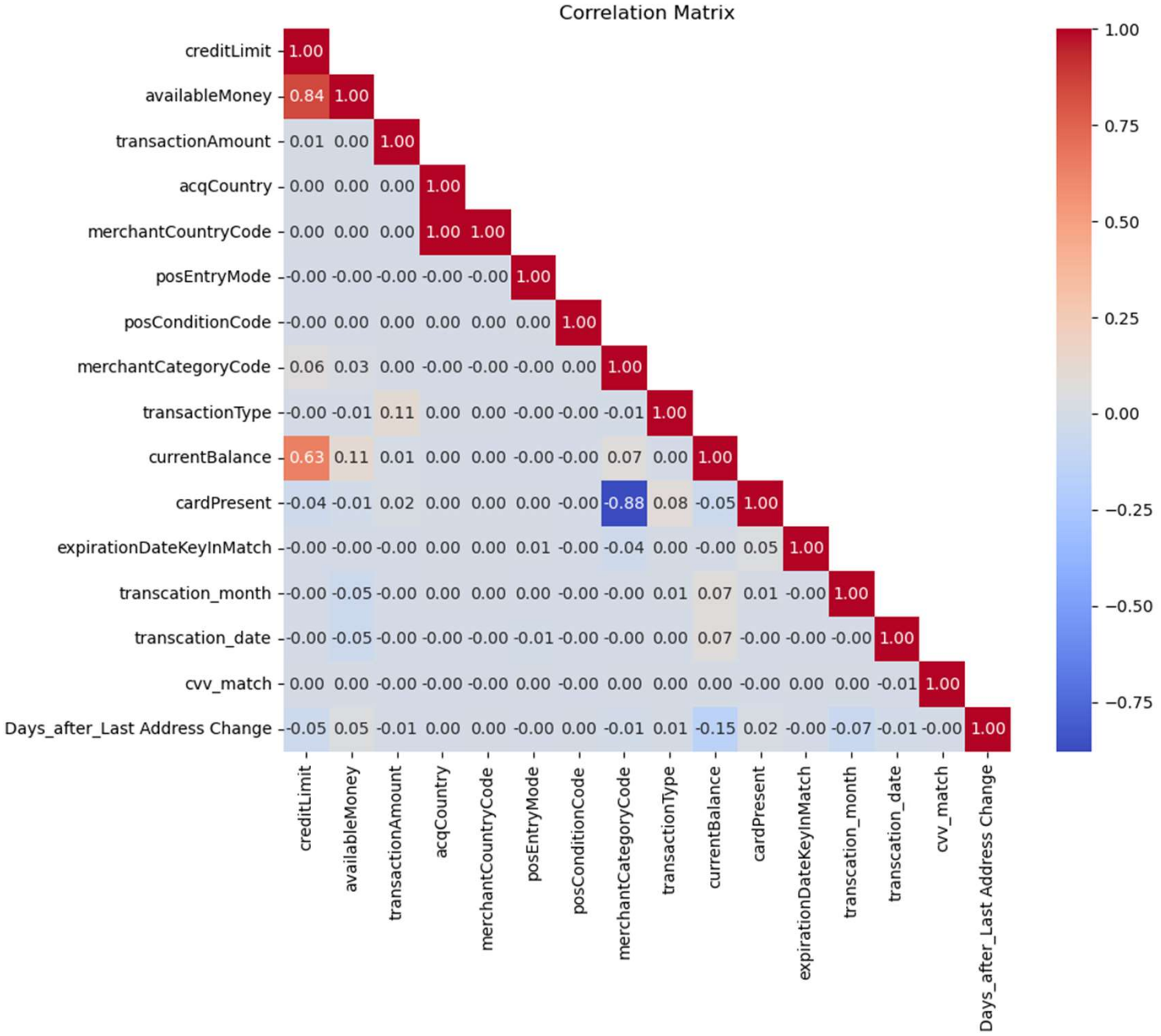
Value Counts of Train and Test

Target Variable Value Counts Train on SMOTE: 153048

Target Variable Value Counts Test: 19504

Assumption Check on Logistic Regression

Assumption	Result	Inference
Binary Outcome Variable	The outcome variable is binary.	The Target Variable IsFraud is True or False. Thus, the Results are in binary
Sufficiently Large Sample Size	Sample size is sufficiently large.	It calculates the sample size needed to achieve a certain level of statistical power and compares it to the actual sample size we have. **sample size** = int(100 * (1 - model. Score(X, y)) / (model. Score(X, y) * 0.20**2)) The Required Sample size is 49 in our Case
Absence of Multicollinearity	There is Presence Multicollinearity in some variables	- **Card Present & Merchant Category Code** has Negative Correlation - Available Money & Creditlimit, Current Balance & Credit Limit , AcqCountry Code and Merchant Country Code Exhibits a High Positive Correlation


Correlation Matrix
creditLimit 1.00
availableMoney 0.84 1.00
transactionAmount 0.01 0.00 1.00
acqCountry 0.00 0.00 0.00 1.00
merchantCountryCode 0.00 0.00 0.00 1.00 1.00
posEntryMode -0.00 -0.00 -0.00 -0.00 -0.00 1.00
posConditionCode -0.00 0.00 0.00 0.00 0.00 0.00 1.00
merchantCategoryCode 0.06 0.03 0.00 -0.00 -0.00 -0.00 0.00 1.00
transactionType -0.00 -0.01 0.11 0.00 0.00 -0.00 -0.00 -0.01 1.00
currentBalance 0.63 0.11 0.01 0.00 0.00 -0.00 -0.00 0.07 0.00 1.00
cardPresent -0.04 -0.01 0.02 0.00 0.00 0.00 -0.00 -0.88 0.08 -0.05 1.00
expirationDateKeyInMatch -0.00 -0.00 -0.00 0.00 0.00 0.01 -0.00 -0.04 0.00 -0.00 0.05 1.00
transcation_month -0.00 -0.05 -0.00 0.00 0.00 0.00 -0.00 -0.00 0.01 0.07 0.01 -0.00 1.00
transcation_date -0.00 -0.05 -0.00 -0.00 -0.00 -0.01 -0.00 -0.00 0.00 0.07 -0.00 -0.00 -0.00 1.00
cvv_match 0.00 0.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 0.00 0.00 -0.00 0.00 0.00 -0.01 1.00
Days_after_Last Address Change -0.05 0.05 -0.01 0.00 0.00 -0.00 0.00 -0.01 0.01 -0.15 0.02 -0.00 -0.07 -0.01 -0.00 1.00
creditLimit
availableMoney
transactionAmount
acqCountry
merchantCountryCode
posEntryMode
posConditionCode
merchantCategoryCode
transactionType
currentBalance
cardPresent
expirationDateKeyInMatch
transcation_month
transcation_date
cvv_match
Days_after_Last Address Change
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Performance Score Table for each Model

greatlearning

Model	Test_size	Accuracy	Recall	Precision	F1 Score
Logistic Regression RS 3 Train	0.2	0.662557	0.635095	0.672004	0.653028
Logistic Regression RS 3 Test	0.2	0.682886	0.591667	0.034075	0.064438

Conclusion (Interim):

Overall, the performance scores for the Logistic Regression RS 3 model suggest that the model is able to identify a significant portion of fraudulent transactions, but it also produces a significant number of false positives. The model could be improved by further tuning the hyperparameters or by using a different machine learning algorithm.

- The model performs better on the test set than on the train set, which suggests that it is not overfitting the train data.
- Overall, the performance scores suggest that the model could be used to identify fraudulent
- transactions, but it should be used with caution due to the high number of false positives.

Future action points:

- We are going to implement different missing values method and test our model and compare the performance for each model
- We are going to transform the data after split and compare the performance of the model
- We are going to implement various algorithm and ensemble techniques & stacking methods to train the model better. (Such as with & without SMOTE & DT, RFC, KNN, naïve bays)
- We will add hypermeter and fine tune the model & compare the performance of each model
- Once the optimum model is identified we are going to test it with unseen data and compare the performance

greatlearning

Thanks For Your Time

Please Share Your Feedback….!!

&

Any Question…?