DEBIT CARD FRAUD TRANSACTION PREDICTION

SUMMARY REPORT

The dataset consists of 50000 unique rows and 21 features related to debit card transactions. Machine learning models are trained with the dataset as the input to predict the fraudulent transaction in future and alert the bank/customer.

Data Preprocessing:

1. Dataset has no missing/null values.
2. There are no duplicate records.
3. Card_age in months feature is converted to card_age in years for easy analysis.
4. Outliers are present in the column Transaction_Amount and the outliers are replaced with the maximum value calculated.
5. Month name is extracted from the Timestamp feature for easy analysis of the data.
6. Features Transaction_ID, User_ID, Timestamp and Card_age in months are removed which are no longer required for the analysis.

Univariate and Bivariate Analysis:

1. With the analysis, Binary columns IP_Address_Flag, Previous_Fraudulent_Activity and Is_Weekend has no significant relation with the target variable Fraud_Label.

2. Analysis of Categorical features against the target variable Fraud_Label.
   - The percentage of fraudulent transactions to the total number of transactions are same irrespective of card type, device type, transaction type, Merchant category, location and authentication method because of which it is not possible to conclude the reason/pattern of the fraudulent transactions.
   - As seen from the analysis, it is not possible to conclude any significant relationship between the independent variables and the target variable.

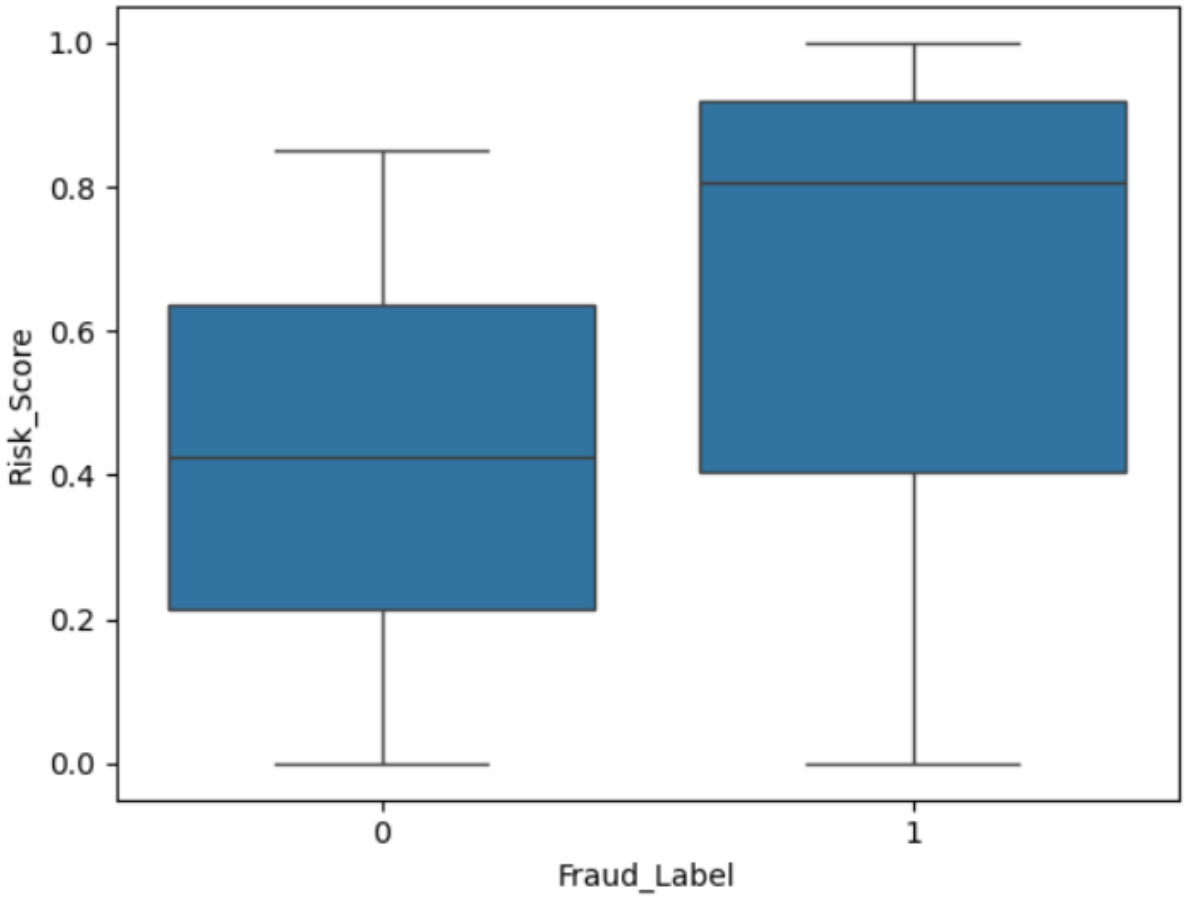
3. Analysis of Risk_score vs Fraud_label

   Two Way ANOVA Test:

	df	sum_sq	mean_sq	F	PR(>F)
IP_Address_Flag	1.0	0.099953	0.099953	0.538470	0.463071
Risk_Score	1.0	1623.187969	1623.187969	8744.462269	0.000000
IP_Address_Flag:Risk_Score	1.0	0.251035	0.251035	1.352378	0.244868
Residual	49996.0	9280.491263	0.185625	NaN	NaN

   Risk_Score pValue is less than 0.05 which means the result does possess a statistically significant effect on fraud_Label.

Graphical Representation:



- Greater the risk score, the higher the possibility of the fraudulent transaction to occur.
- Almost 90% of the fraudulent transactions are with the Risk_Score greater than 0.4.

4. Relation between Failed transaction count and Fraud Label

	Failed_Transaction_Count_7d	Number of Fraudulent transactions
0	4	9954
1	3	1601
2	0	1542
3	1	1531
4	2	1439

We can see that there is an occurrence of fraudulent transaction in the card profile with a greater number of failed transactions. i.e., Count of fraud transactions are more for the higher failed transaction count.

Feature Selection and Model Selection:

I have used SelectKBest algorithm to select the important features for training the model.
The selected features are

	Feature_name	Score
1	Failed_Transaction_Count_7d	0.579029
2	Risk_Score	0.419716
0	Account_Balance	0.001255

It is inferred that features Failed transaction count and Risk score are highly correlated with the target variable Fraud label.
Below are the Models created and the metrics are evaluated for all the models to identify the good performing models.
Grid Search is used in the model creation to identify the best combination of hyper tuning parameters which yields the best model performance.
The best model identified with Grid Search is

RandomForestClassifier(criterion='gini', max_features='sqrt', n_estimators=100)

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	logistic	0.80152	0.722082	0.634218	0.675304	0.896175
1	KNN	0.99704	0.995331	0.995575	0.995453	0.999968
2	Naive Bayes	0.89104	0.854374	0.801868	0.827289	0.948643
3	DecisionTree Classifier	0.99960	0.999508	0.999263	0.999385	0.999513
4	Random Forest Classifier	1.00000	1.000000	1.000000	1.000000	1.000000

From the above table, we can create the model with Random Forest Classifier since all the metrics are equal to 1 which denotes a very good performance of the model.