CREDIT CARD FRAUD ANALYSIS AND CLAUSTERING

Objective

Analyze credit card transactions by looking for patterns, and training different clustering models, as well as picking the best clustering model that could aid the early detection of fraudulent transactions.

Description of Dataset

The credit card fraud dataset is designed for fraud detection analysis. The dataset contains 10,000 rows and several features that represent various characteristics of credit card transactions. Each row in the dataset represents a single transaction, and one of the columns indicates whether the transaction is fraudulent or not. The columns that make up the dataset are:

1. **Transaction_Amount:** The amount of money involved in the transaction. This could be a strong indicator of fraud if unusual large amounts are involved compared to the customer's typical spending pattern.
2. **Transaction_Time:** The timestamp of when the transaction occurred, which often is represented in seconds since it is a specific point in time (e.g., Unix epoch time). Fraudulent transactions may often occur at unusual times (e.g., late at night or at irregular intervals).
3. **Transaction_Type:** The type of transaction (e.g., online payment, in-store purchase, ATM withdrawal). Online transactions, in particular, are more prone to fraudulent activity than in-store purchases.
4. **Account_Age:** Age of the account in days or months.
5. **Location_Match:** Binary feature indicating if the transaction location matches the account holder's usual transaction location. Fraudulent transactions often occur in locations far from the cardholder's usual location, or even in different countries. This feature could be one of the signals for fraud detection.
6. **Device_Suspicion_Score:** A numerical score representing how suspicious the device used in the transaction is (higher means more suspicious).
7. **Merchant_Risk_Score:** A score representing how risky the merchant is based on previous transaction history. This shows the category of the merchant where the transaction was made (i.e grocery store, electronics store, online retailer). Some categories might be more prone to fraud, especially online or high-value goods merchants.

8. **Previous_Fraud_Flag:** A feature indicating whether an account has previously been involved in fraudulent activity (i.e the transaction was initially flagged as suspicious by the bank or payment processor.
9. **Fraud_Label:** The target column, where 1 indicates a fraudulent transaction and 0 indicates a legitimate one.

Data Exploration

Data Cleaning

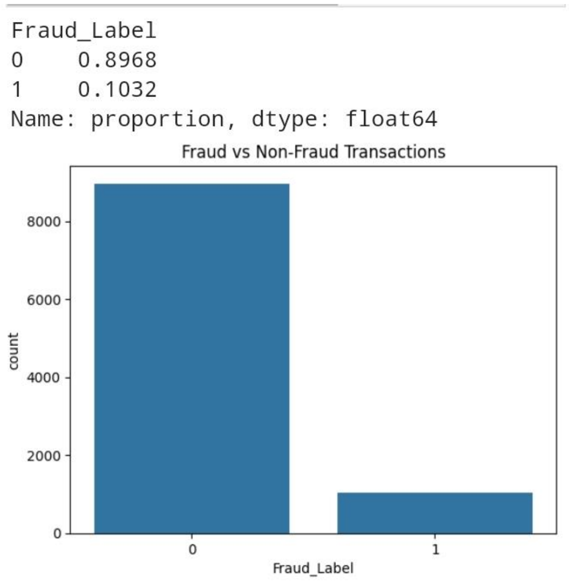
The data cleaning process started with checking for missing values, duplicates, and outliers. The dataset was found to have no missing values, no duplicate values, and no outliers.

Statistical Analysis

The distribution and skewness of the dataset was checked using the histogram plot. The target variable (Fraud_Label column) was found to be highly skewed. The skewness of the target variable was seen to be due to the high level of imbalance between the non fraudulent transactions (0) and the fraudulent transaction (1).

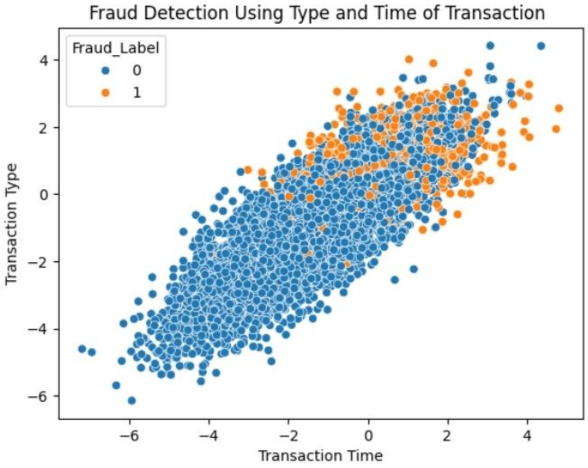
Data Visualization

Patterns from different data visuals was studied, and deep insight gotten, of how well data of some features exposed fraudulent transactions, as well as suspecting and detecting that a credit card transaction is fraudulent. The chart from these Visualizations are shown and explained below:

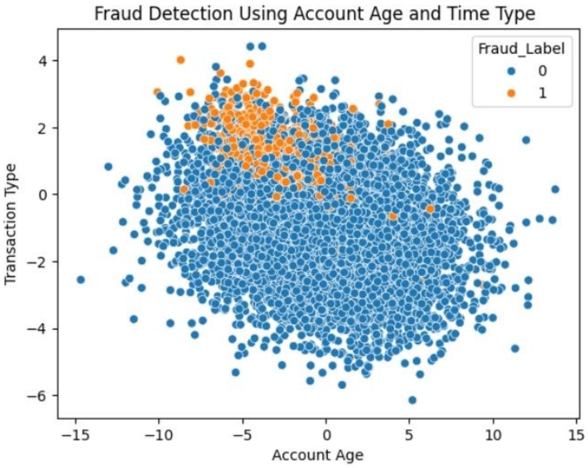


```
Fraud_Label
0    0.8968
1    0.1032
Name: proportion, dtype: float64
```

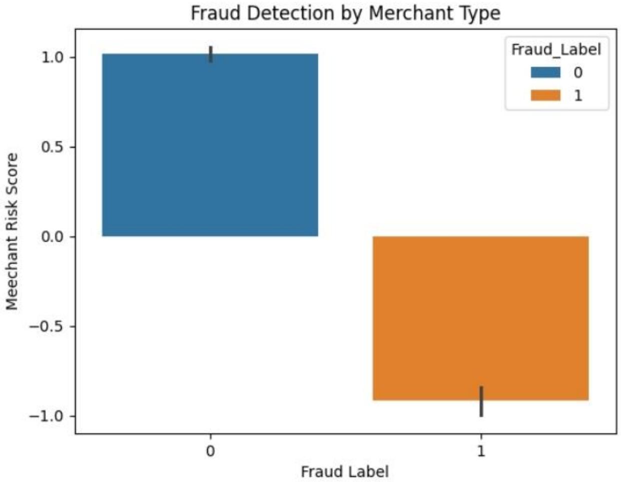
The above chart shows the level of imbalance between the fraudulent and non fraudulent transactions, as well as the scores of the imbalanced classes



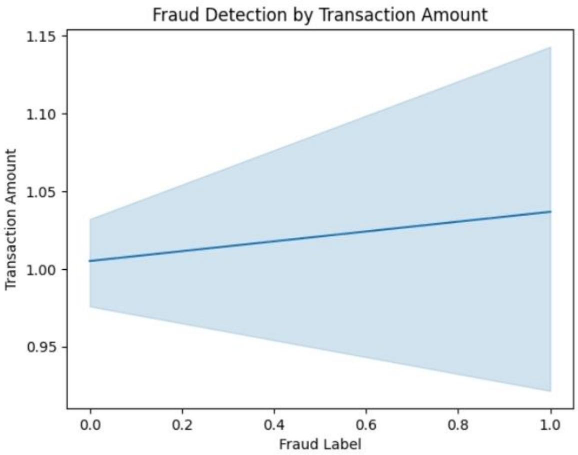
The above chart shows that most fraudulent transactions occurred at odd times, with regards to the type of transaction.



The above chart shows that fraudulent transactions are mainly carried out using new / young accounts, with consideration to the type of transaction.



The fraud detection by merchant chart tells us that merchants with merchant risk score below zero (0) are mainly used to carry out fraudulent transactions.



The fraud detection by transaction amount chart shown above shows that higher transactions are mainly fraudulent transactions, depending on customers typical spending pattern.

Features Selection

Features for training the models were selected using:

- ★ Pearson Correlation with target features.
- ★ Tree-base feature selection method.
- ★ Laplacian scoring Method.

From these methods of features selection, a feature that appears in at least two of the methods of selection, was used for training the model.

Model Creation and Evaluation

4 different clustering models were trained and evaluated, with all models using the same random and cluster parameters for training. They are:

- ★ Kmeans Clustering Model.
- ★ DBSCAN Model.
- ★ Agglomerative Clustering Model.
- ★ Gaussian Mixture clustering Model.

The four Clustering models showed their Clustering capabilities which was evaluated using Silhouette Score and Davies-Bouldin Score.

Recommended Clustering Model

The target feature label is : 0=0.8968 and 1=0.1032, as seen from visuals and analysis.

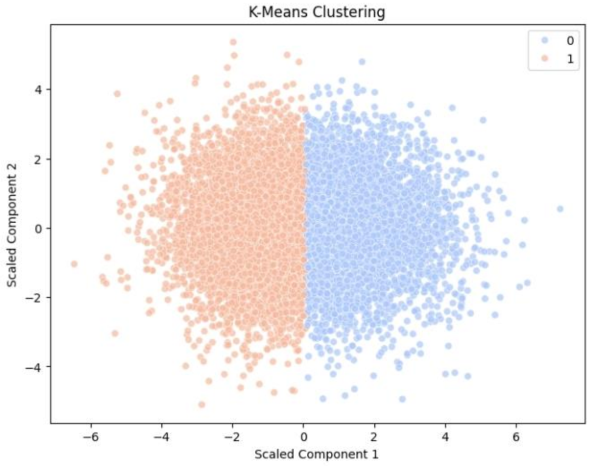
MODELS	Silhouette Score	Davies-Bouldin Score.	Label	
			0.	1
Kmeans	0.3030	1.2402	0.523	0.477
DBSCAN	0.3007	7.3085	0.8927	0.1073
Agglomerative	0.2852	1.2886	0.5133	0.4867
Gaussian Mixture	0.1340	1.7614	0.9003	0.0997

Table showing the different clustering models and their evaluation

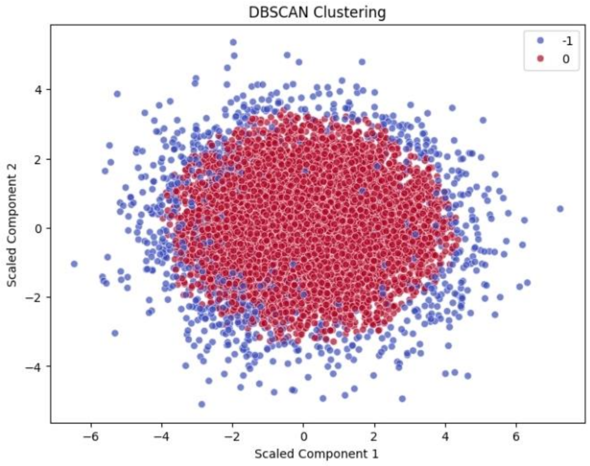
Given the results, none of the models stands out as perfect. However, DBSCAN and Gaussian Mixture Model (GMM) have label distributions closest to the target (0.8968, 0.1032), which suggests they capture the class balance more accurately. Despite this, DBSCAN's very high Davies-Bouldin (DB) score (7.3085) and GMM's low Silhouette score (0.1340) indicate poor cluster separation.

Key Findings and Insights

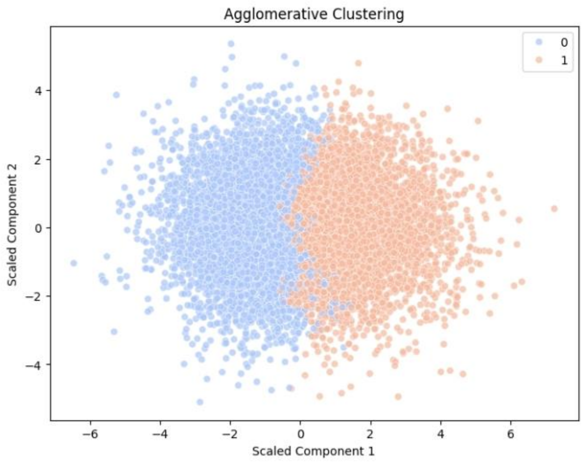
The Kmeans Clustering Model showed balanced clustering abilities but not close to the target distribution. The Silhouette and DB scores were decent but the model did not show exceptional performance.



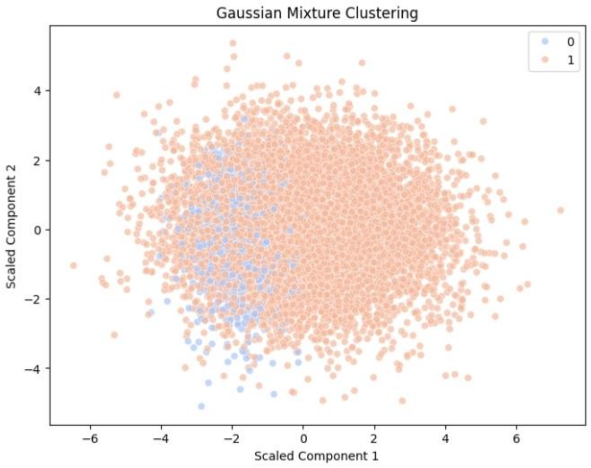
The DBSCAN Clustering Model have the closest label values with the target distribution, but a high DB score shows poor cluster compactness, indicating noisy or overlapping clusters.



The Agglomerative Clustering Model have similar performance to KMeans model, with balanced but incorrect label distribution and slightly worse clustering quality.



The Gaussian Mixture Model (GMM) also have a close match to the target distribution, but the very poor Silhouette score suggests overlapping clusters, with moderate DB score indicating poor separation.



Suggestions for next steps in analyzing this data

Further evaluation is needed to find a balance between cluster quality and matching label distribution before a recommended model can be chosen. Experimenting with different parameters for the current models or trying other clustering techniques could improve the results..