Credit Card Fraud Detection Dataset Documentation

Executive Summary/Data Source and access information: The dataset used in the dashboard building has been taken from publicly open source kaggle website https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?utm_source .The dataset contains transactions made by European credit card holders back in September 2013.Due to confidentiality constraints only the input values has been made available for the public users .The column values that are made available are Time, amount, Class, Fraud label, Amount category .

Time: Contains information about elapsed seconds between first and each transaction made.

Amount: Defines the total amount where the transactions have been made and using this information insights can be generated

Class: Categorized as 0 or 1 for 1 as fraud and 0 as normal transactions.

Amount Category: In the dashboard the ranges 0-100 , 101-500 and 500+ have been defined to identify fraudulent transactions and their count. **This data has been used using calculated functions like Dax measures.**

Fraud Label: In the dashboard with the help of Dax a function has been created that distinguishes the column values with 1 as Fraud and 0 as Normal using the Class column from the dataset.

Business Question:

1. **What is the overall fraud rate and how does it seem to be observed ?**
   As per the analysis and the data sets the total transaction is 73.16k and the fraudulent transaction seemed to happen over amount 500+ (exact figure according to report is 529) which makes up to 0.138% of the overall .Since due to less number of rows there seems to be a drop in trend for the fraudulent transaction taking place as zeroth hour(hour when business operations are closed and activity isn't tracked down and only those transactions are visible that start with the first hour of the day).

2. **Do high value transactions seem to be targeted often?**
   Looking into the insights and analysis it seems that scammers/hackers try to target the accounts with higher values to gain maximum profit .This Clearly states the amount band categorized as Amount category like 0-100,101-500 and 500+.

3. **Which graph shows the distribution of fraud Vs non -fraudulent transaction?**
   The stacked column chart visual shows the distribution along with labels classified as fraud or non fraud(normal) based on the amount Where transactions have been made.

4. **What actions can be taken to set alerts or threshold?**

As per my analysis the baseline rate calculation should be based on total transaction/fraud transaction *2 and once this threshold reaches for a particular account more than 2 an alert should be set by flagging and with a check in every 30-60 min during hours of operations and auto set during off hours such as 11 pm-3 am .

5. **Should we keep a check on larger transactions ?**
   In my opinion yes we should as accounts with higher transactional rate have more chances of being at risk.

Business Questions as per my report:

1. **What is the overall fraud rate and how can it be answered?**
   0.138% approx (Total transaction/Fraudulent transaction)

2. **What is the transactional amount that has been observed as fraud ?**
   529 with the category falling under an amount ranging 500+.

3. **What is the total percentage of transactions as per the report?**
   100% with total number as 2.

4. **What is the Dax measure that has been used to create a band or range for the transaction to distinguish as fraud or non fraud?**

```
Amount_Category =
 SWITCH(TRUE(),
'CC_FraudData'[Amount]<=100,"LOW(0-100)",
CC_FraudData[Amount]<=500,"Medium(101-500)",
"High(500+)"
)
```

5. **What Calculative functions have you used to identify which transaction is Fraud?**

```
Total_Fraud_Transaction_Amount =
CALCULATE(SUM(CC_FraudData[Amount]),'CC_FraudData'[Fraud_label]="Fraud")
```

Technical Approach: For the credit card fraud detection dataset there were few technical approaches consisting of Calculated columns over rows using Dax queries for logical reasoning, measures to perform any kind of aggregations such as Sum, Count to build KPI's. The below ones show the columns and measures used in the report.

1. **Amount Category(Calculated Column)**:

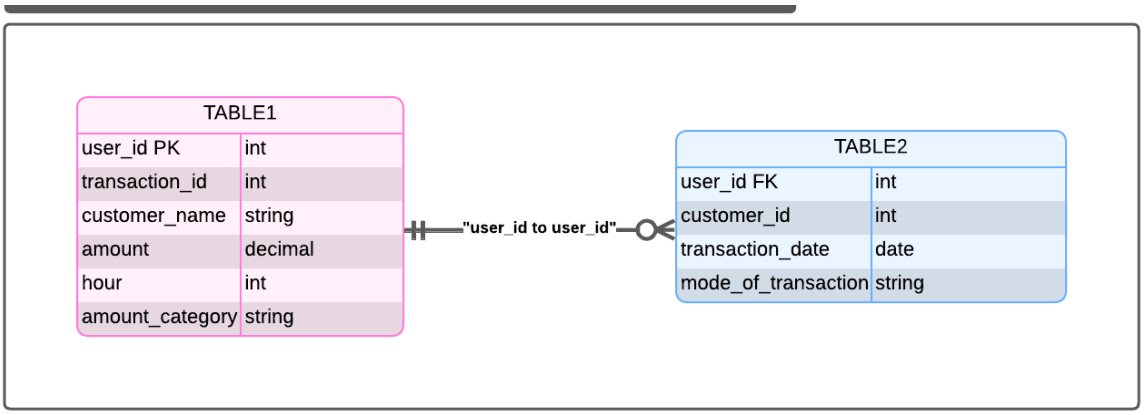
```
Amount_Cataegory =
SWITCH(TRUE(),
'CC_FraudData'[Amount]<=100,"LOW(0-100)",
CC_FraudData[Amount]<=500,"Medium(101-500)",
"High(500+)"
)
```

2. **Fraud Label(Calculated Column)**`Fraud_label = IF('CC_FraudData'[Class]=1, "Fraud", "Normal")`
3. **Fraud Transactions(Measure):**`Fraud_label = IF('CC_FraudData'[Class]=1, "Fraud", "Normal")`
4. **Fraud%(Calculated Column):**`Fraud%2 = DIVIDE([Fraud_Transactions],[Total_Transactions],0)`
5. **Hour(Calculated Column):** `Fraud%2 = DIVIDE([Fraud_Transactions],[Total_Transactions],0)`
6. **Total Fraud%: (Parameter Field/Measure)**`Total Fraud% = CC_FraudData[Fraud%]*'Parameter'[Parameter Value]`
7. **Total Fraud Transaction Amount:(Measure)**`Total_Fraud_Transaction_Amount = CALCULATE(SUM(CC_FraudData[Amount]),'CC_FraudData'[Fraud_label]="Fraud")`
8. **Total Normal Transaction :(Measure)**`Total_Normal_Transaction_Amount = CALCULATE(SUM(CC_FraudData[Amount]),'CC_FraudData'[Fraud_label]="Normal")`
9. **Total Transaction amount(Measure):** `Total_Transaction_Amount = SUM(CC_FraudData[Amount])`
10. **Total Transaction amount to Count rows(Measure)**`Total_Transactions = COUNTROWS('CC_FraudData')`
11. **mAmount(Field Parameter):** `mAmount = SUM(CC_FraudData[Amount])`

2. Data Model Design: Since the data used in the report is a single table dataset thats why there wasn't requirement to build data model .If multiple tables would have been used then model would have been created based on following points:

1. Connect the tables with the primary key of the parent table and use the left join foreign key of the child table.
2. As the general rule while building model we use one-many relationship as one field of parent table is always connected to the multiple table of child table

E.g.



3. **Key Decision Made:** Depending on the alert threshold following decisions can be made:
   1. Block high amount transaction without authentication
   2. One time passcode for every transaction
   3. Manual review transaction
   4. Increased monitoring during high risk periods.

4. **Insights:** As per the report the insights that seemed to be observed possibly.

1. Accounts with high amount and transactions are at high risk

2. There could be possible fraudulent activity being observed during peak off hours like 11 pm-3 am as its less possible to track down .

3. Zero hour activity being tracked down i.e only those hours being tracked down when business hours of operations are active and not the peak offhours.

5. **Challenges:** The following were the challenges that were faced during building of dashboard:

1. The dataset was not user friendly as many columns were transformed to numeric values to secure confidential data which was challenging in building insights .

Solution: Used the available existing column value like time, amount to generate insights

2. The original data contained around 284, 807 rows and due to the limitation of software like MS-Excel it only allowed around 1 mb data.

Solution: A minimum of 1,000 rows of data were needed for analysis and dashboard creation in accordance with the assignment requirements. A subset of 1,100 rows was picked from the original dataset in order to satisfy and slightly surpass this criteria while guaranteeing effective processing and usability within the chosen software environment.

This sample size is manageable for analysis, visualisation, and performance optimisation within the reporting tool, and it adequately meets the minimal data volume requirements.

3. Lastly, due to hidden column values or data transformed to numeric, ensuring security compliance on public sites applying advanced features like RLS, Field parameters, what if parameters was a bit challenging.

Solution: To overcome that, I tried to build an advanced solution using columns like amount, hour, class , category and existing Dax measures to meet the requirements.

Data Transformation: Before analyzing the data there were few data transformations steps taken in power query such as:

1. Time column data type was changed from int to time form format.
2. Amount row was sorted in ascending order
3. Removed null/ blank values from the time column.

Note: Portfolio Link https://rsingh96.replit.app