Credit card fraud detection system

Problem statement

With the increasing digitalization and online transactions, It becomes ever so important for the credit card companies to be able to recognize “genuine” and “fraudulent” transactions in order to provide their customers with a more secure and a seamless experience.

As a big data engineer, you should architect and design a solution using all the technologies learnt during this program to meet the following requirements:

1.Detect fraudulent transactions at the shortest possible time (Since the transactions are happening in real time, timing constraint plays a very important role). Whenever a card member swipes his/her card for payment, the transaction should be classified as **fraudulent / authentic** based on a set of predefined rules.

2. To resolve the customer complaints and queries, the support team should be made available with the latest customer details(by constantly keeping them updated.)

Rules to categorize a transaction as fraudulent / genuine :

1. **Upper Control Limit (UCL) :**

   Based on the previous transaction patterns of the card user, the maximum limit on the amount per transaction is used as a parameter to authenticate the transaction.

   UCL = (Moving average) + 3 * (Standard deviation) {These are calculated for the last 10 transactions marked as genuine}

2. **Credit Score:**

   The score field in the member_score table is used as a parameter to authenticate the transaction.

   If score < 200, member is a defaulter and the transaction is rejected.

3. **Zip Code distance:**

   The distance with respect to time between the current and previous transaction locations is used as a parameter to determine the authenticity of the transaction.

   If the distance between the transaction locations with respect to time is greater than a particular threshold, then the transaction is marked as fraud.

Understanding the types of data that will be dealt with.

We will be considering the following tables :

card_member -> This table consists of cardholder data and the data is added into this this table by a third party service. Following are the fields in the card_member table.

- card_id (Card number)
- member_id (Member id of the cardholder)
- member_joining_dt (Joining data and time of a new member)
- card_purchase_dt (Purchase date of the card)
- country (Name of the country where the card was purchased)
- city (Name of the city where the card was purchased)

card_transactions -> This table consists of all the transactions carried out by swiping at the POS system. Following are the fields in the card_transactions table.

- card_id (Card number)
- member_id (Member id of the cardholder)
- amount (Amount swiped with the particular card_id)
- postcode (marks the location of the event based on the zipcode at which the card was swiped)
- pos_id (POS terminal ID using which the card was swiped)
- transaction_dt (date and time of transaction)
- Status (Genuine / Fraud value based on whether the transaction was approved or declined)

member_score -> This table consists of all the transactions carried out by swiping at the POS system. Following are the fields in the card_transactions table.

- member_id (Member id of the cardholder)
- score (The score assigned to a member defining his/her credit history)

Apart from the above types of data, There will be real-time data that will be streaming in, generated by the POS(Point of Sale) systems. This data will be in JSON format as shown below :

{

“card_id” : 487654323445689,

“member_id” : 000987654123456,

“amount” : 245600,

“pos_id” : 78765324158934,

“postcode” : 33946,

“transaction_dt” : 09-01-2021 18:00:00

}