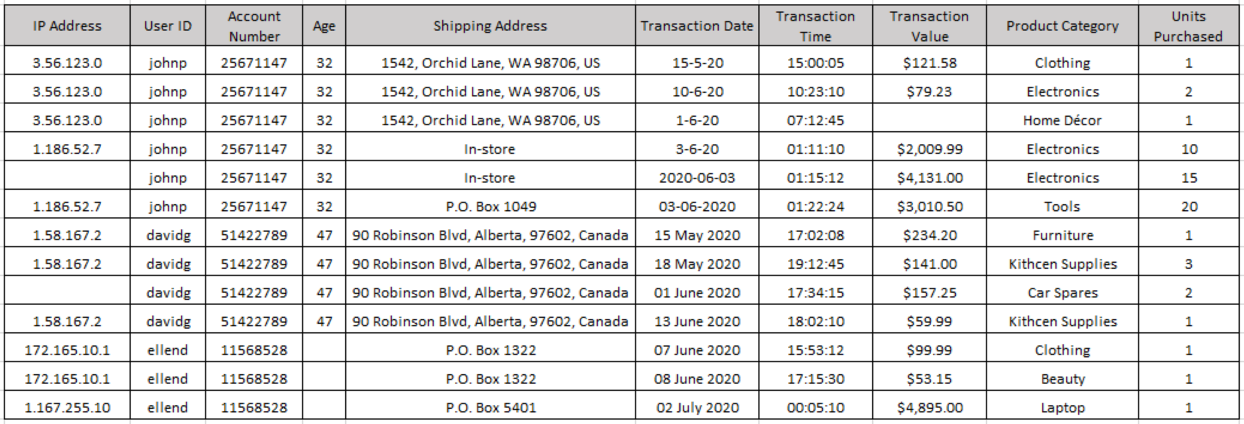
Using Data Analysis for Detecting Credit Card Fraud

Companies today are employing analytical techniques for the early detection of credit card frauds, a key factor in mitigating fraud damage. The most common type of credit card fraud does not involve the physical stealing of the card, but that of credit card credentials, which are then used for online purchases.

Imagine that you have been hired as a Data Analyst to work in the Credit Card Division of a bank. And your first assignment is to join your team in using data analysis for the early detection and mitigation of credit card fraud.

In order to prescribe a way forward, that is, suggest what should be done in order for fraud to get detected early on, you need to understand what a fraudulent transaction looks like. And for that you need to start by looking at historical data.

Here is a sample data set that captures the credit card transaction details for a few users.



IP Address	User ID	Account Number	Age	Shipping Address	Transaction Date	Transaction Time	Transaction Value	Product Category	Units Purchased
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	15-5-20	15:00:05	$121.58	Clothing	1
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	10-6-20	10:23:10	$79.23	Electronics	2
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	1-6-20	07:12:45		Home Décor	1
1.186.52.7	johnp	25671147	32	In-store	3-6-20	01:11:10	$2,009.99	Electronics	10
	johnp	25671147	32	In-store	2020-06-03	01:15:12	$4,131.00	Electronics	15
1.186.52.7	johnp	25671147	32	P.O. Box 1049	03-06-2020	01:22:24	$3,010.50	Tools	20
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 May 2020	17:02:08	$234.20	Furniture	1
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	18 May 2020	19:12:45	$141.00	Kithcen Supplies	3
	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	01 June 2020	17:34:15	$157.25	Car Spares	2
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	13 June 2020	18:02:10	$59.99	Kithcen Supplies	1
172.165.10.1	ellend	11568528		P.O. Box 1322	07 June 2020	15:53:12	$99.99	Clothing	1
172.165.10.1	ellend	11568528		P.O. Box 1322	08 June 2020	17:15:30	$53.15	Beauty	1
1.167.255.10	ellend	11568528		P.O. Box 5401	02 July 2020	00:05:10	$4,895.00	Laptop	1

Descriptive techniques of analysis, that is, techniques that help you gain an understanding of what happened, include the identification of patterns and anomalies in data. Anomalies signify a variation in a pattern that seems uncharacteristic, or, out of the ordinary. Anomalies may occur for perfectly valid and genuine reasons, but they do warrant an evaluation because they can be a sign of fraudulent activity.

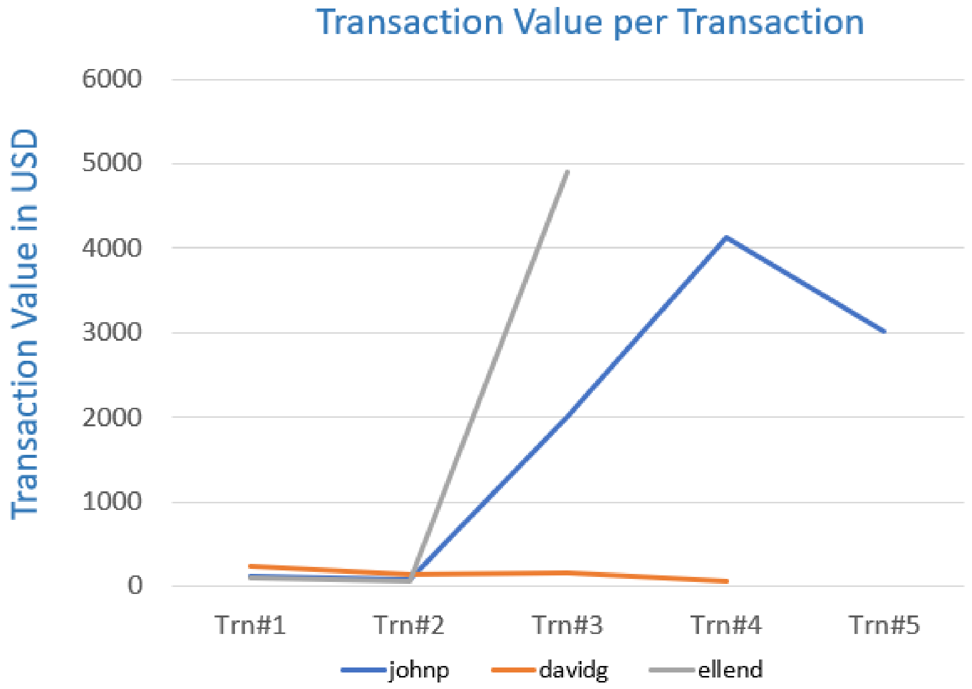
Past studies have suggested that some of the common events that you may need to watch out for include:

- A change in frequency of orders placed, for example, a customer who typically places a couple of orders a month, suddenly makes numerous transactions within a short span of time, sometimes within minutes of the previous order.
- Orders that are significantly higher than a user's average transaction.
- Bulk orders of the same item with slight variations such as color or size—especially if this is atypical of the user's transaction history.
- A sudden change in delivery preference, for example, a change from home or office delivery address to in-store, warehouse, or PO Box delivery.
- A mismatched IP Address, or an IP Address that is not from the general location or area of the billing address.

Before you can analyze the data for patterns and anomalies, you need to:

- **Identify and gather all data points that can be of relevance to your use case.** For example, the card holder's details, transaction details, delivery details, location, and network are some of the data points that could be explored.
- **Clean the data.** You need to identify and fix issues in the data that can lead to false or incomplete findings, such as missing data values and incorrect data. You may also need to standardize data formats in some cases, for example, the date fields.

Finally, when you arrive at the findings, you will create appropriate visualizations that communicate your findings to your audience. The graph below samples one such visualization that you would use to capture a trend hidden in the sample data set shared earlier on in the case study.



In the next section you will be asked to answer the following 5 (five) questions based on this case study:

1. List at least 5 (five) data points that are required for the analysis and detection of a credit card fraud.

The Data analysis process is comprised of Discovery, Transformation, Validation, and Publication. During the analysis phase of the Credit Card, the Fault Detection model data point plays an important role. Consider the 5 data points for credit card fault detection as given below,

1> IP Address: IP Address is one of the crucial points in the Credit Card Fraud Detection Predictive Model. A sudden change in the IP Address can indicate unauthorized use of a credit card. During the Data Mining process, Data Analysts could use the clustering technique to group similar IP Addresses together and can easily identify the address that is different from than real Owner's IP Address. This could help to trace the exact location of the unauthorized entity.

2> Shipping Address: The Second data point important for the analysis is Shipping Address. With the help of sequence analysis, the data analysts will track down a series of related events, identifying the address change. Once the address change is detected, the Data analyst will analyze the transaction amount related to that address. This point will help to identify the unauthorized access of the card. 3> Transaction Value: It will be useful in identifying the sudden data spike from the set of values. Data Analysts could isolate sudden large transaction spikes and compare them with other transactions to detect fraudulent activity.

4> Transaction Date: The Transaction Date log can be correlated with the customer complaint log. If the customer has placed the complaint on a specific date and the date with unauthorized transaction matches that could be a big lead while performing the analysis.

5> Units Purchased: One of the important data points is Units Purchased, With Affinity Grouping, one could identify the correlations between the Units purchased by a specific customer from his Account Number. The Central Tendency function mode will be useful to identify the maximum units purchased from given units purchased list of a specific customer.

1. Card holder / Customer Id
2. Transaction date
3. Transaction time
4. Transaction value
5. Shipping address
6. IP address
7. Device model
8. Location

2. Identify 3 (three) errors/issues that could impact the accuracy of your findings, based on a data table provided.

1> **According to the data provided in the table it needs Transformation**, During the transformation phase, the given data should be checked if it's properly structured or not. One of the important tasks during the transformation is database normalization. It is primarily used to remove the redundancy from the database table. The above database table consists of data redundancy. The normalization can be accomplished by splitting the shipping address column and assigning the account number as the primary key. The shipping address can be placed in the other table in order to satisfy the Second and Third Normal forms. The address should be further divided into the address name and zip code in separate columns.

2> **Missing Value**: Data Cleaning Process would identify the missing values in the given process. The missing values can be replaced by null, a presumed value, or an average of the values.

3> **Identifying the Outliners**: The Transaction Value Analysis would give insights on the outliners having more than three standard deviations. By identifying the outliners one could decide whether to include them in the decision process or not. In order to avoid skew data and maintain data accuracy, data analysts could identify the outlines and eliminate them during the data analysis process.

- Missing transaction value
- Missing IP Address
- Date format inconsistency

3. Identify 2 (two) anomalies, or unexpected behaviors, that would lead you to believe the transaction may be suspect, based on a data table provided. (2 marks)

Consider the following two anomalies,

1> User ID(ellend): When you consider the user with ellend you get to understand that by using clustering and affinity grouping you can see the IP Addresses for the first two transactions are the same and with the owner's valid address. When you move down to the transaction on 02 July 2020, it seems that there is a sudden IP Address change. Moreover, there is a significant Shipping Address change. Data Analysts can further search for the customer database to detect if the account owner has made changes to their residential address or not. If that's, not the case then there is an anomaly in this transaction. One more data point as Transaction Value shows a sudden increase in the payment amount.

2> User ID(johnp): One of the interesting data points with johnp is the transaction date. Data Analyst will closely take a look at transaction date 3-6-20 and establish the correlation between the sudden IP address change and Increased purchases. As two similar product categories have been ordered from the store and one is ordered online at a different shipping address than the original one. This in turn would ensure the suspicious behaviour and present the anomalous data.

4. Briefly explain your key take-away from the provided data visualization chart.

The Provided data visualization chart decribes the Transaction Values and different transactions perfomed by a specific user. According to the chart, johnp is represented by the blue line, where as davidg is represented by the orange color. The ellend represents sudden spike in the transaction value followed by the ending curve. This represents the anamoly in the user transaction. According to the credit card fraud detection predective model, if there is sudden incrase in the

transaction value which sets it apart from all other previous values. It can be considered as a outliner. The more supicious acitivity can be availd by cheking as their is no subsequencet transaction occured after the last transaction on ellend user's account.
Transactions on the ellend user account starts with the Tranq and Tran2 within the normal range and sudden spike can be observed at the Transaction 3 which leads to the suspicious behaviour.
The Line chart for the johnp represents gradual increase in the transaction amount, while doing the highest transaction around $4000 and then doing few more transaction below the same amount. The treanaction1 and 2 are within the normal range where as spending has been increased from the transaction 3. Moreover, Transaction4 would constitue the highest expenditure. There is a significant reduction in the amount at transaction5.

The Line chart for the david represnets the gradual decrease in the spending, clealry indicating that the card has been utilized by the owner. As all the transaction are in the fixed range evenly distibuted over the time. For david there is minute transaction chnage from the transaction 1 to the transaction 4.

5. Identify the type of analysis that you are performing when you are analyzing historical credit card data to understand what a fraudulent transaction looks like. [Hint: The four types of Analytics include: Descriptive, Diagnostic, Predictive, Prescriptive]

Descriptive Analytics: The Descriptive Analytics would be primarily used to gather data from the past events at any time. It will be processing all time historic credit card transaction related data. In credit card fraud detection analysis the descriptive analysis would play an important role. As the transaction history form the previous purchases would act as a backbone for data analysis. The descriptive analytics will give information about the past IP Addresses, Transaction Value, Transaction Data and Shipping addresses. Such information could be crucial while detecting suspicious activities.

Diagnostic Analytics can be Performed on the data collected from the Descriptive analysis in order to implement the deep research over set of values. Diagnostic Analytics will give closer insight on credit history data and help data analyst to detect data anomalies sooner. Diagnostic analysis will help to establish the relationships established on data points provided by the Descriptive Analytics such as transaction date, Ip addresses etc.

Predictive analytics will assist in performing the sequence analysis and recognizing the trend in the transactions. Data Analyst could predict the next transaction initiation location by tracking the IP Address. By corelating Transaction date, Ip Address and the Transaction amount data analyst could trace pattern and identify the next suspicious activity location.