Credit Card Fraud Detection

Using SQL query

Hadari kimchi

About Dataset

- Digital payments are evolving, but so are cyber criminals.
- According to the Data Breach Index, more than 5 million records are being stolen on a daily basis, a concerning statistic that shows - fraud is still very common both for Card-Present and Card-not Present type of payments.
- In today's digital world where trillions of Card transaction happens per day, detection of fraud is challenging.

There are 1,000,000 rows & 8 columns:

- **Distance_from_home** - the distance from home where the transaction happened.
- **Distance_from_last_transaction** - the distance from last transaction happened.
- **Ratio_to_median_purchase_price** - Ratio of purchased price transaction to median purchase price.
- **repeat_retailer** - Is the transaction happened from same retailer. (Boolean variable)
- **used_chip** - Is the transaction through chip (credit card). (Boolean variable)
- **Used_pin_number** - Is the transaction happened by using PIN number. (Boolean variable)
- **online_order** - Is the transaction an online order. (Boolean variable)
- **fraud** - Is the transaction fraudulent. (Boolean variable)

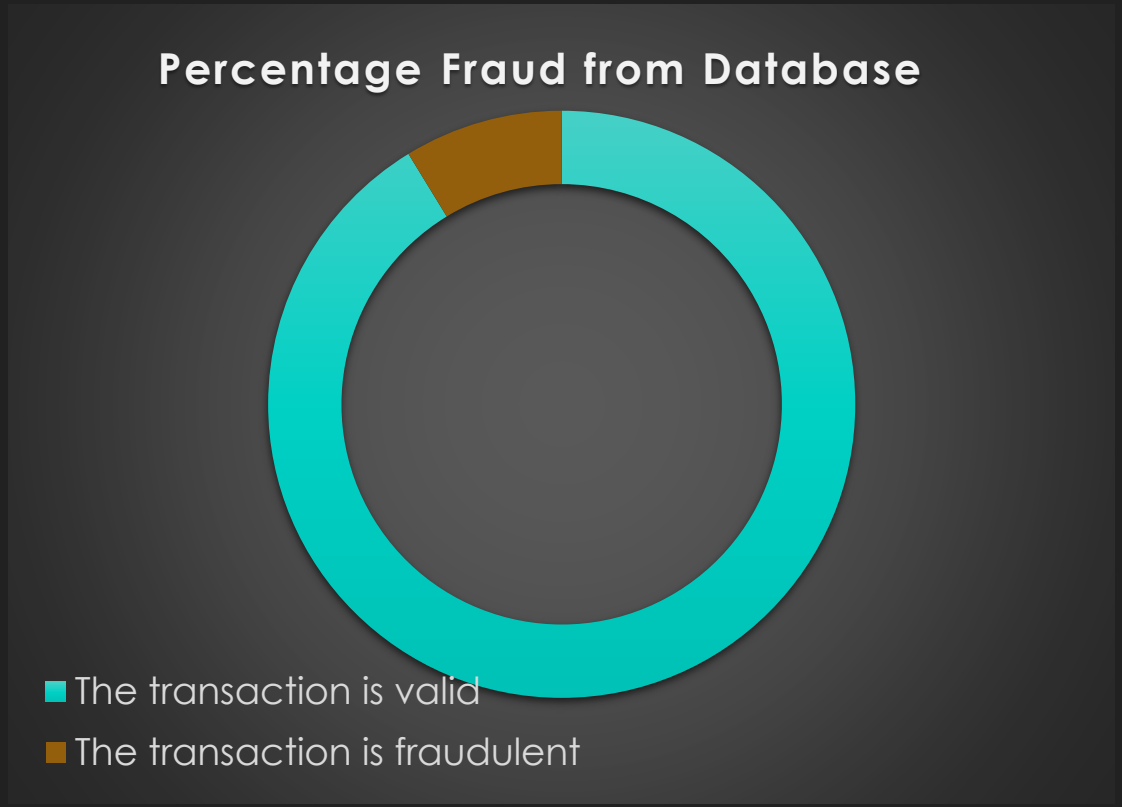
The project's goal

To examine what parameters indicate a significant connection to the commission of a credit card fraud.

Fraud

The first figure I wanted to check is the percentage of fraud out of all the data

The results showed that 91.26% of the purchases are legal and the rest, 8.74% are fraud.



The relationship of each figure individually (in percent) to the question of whether a fraud was committed or not?

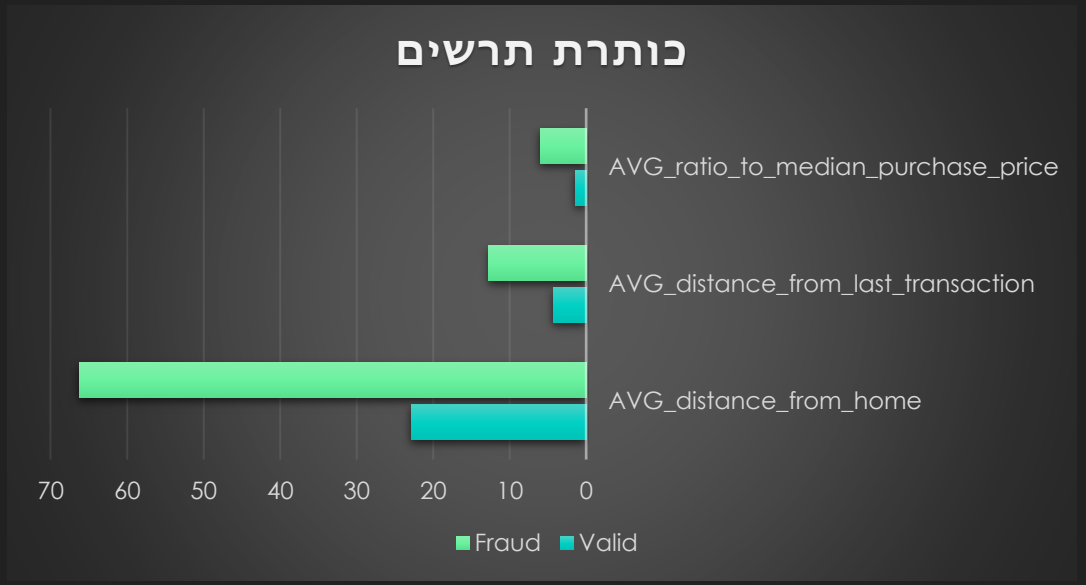
fraud	online_order_pct	used_pin_number_pct	used_chip_pct	repeat_retailer_pct
Valid	62.22	10.99	35.94	88.17
Fraud	94.63	0.13	25.64	88.01

- According to the results, it seems that the figure that has the most impact is:
- "online_order" (94% Vs 62%)
- In addition, "repeat_retailer" has almost no effect.

Examining the relationship between continuous variables and fraud

fraud	AVG_ratio_to_median_purchase_price	AVG_distance_from_last_transaction	AVG_distance_from_home
Valid	1.42	4.3	22.83
Fraud	6.01	12.71	66.26

- For each column with a continuous continuous variable I created an average according to the division of the fraud variable.
- For the purpose of examining the relationship between these columns and fraud, I have chosen the middle of the difference.
- For example - in a query with the variable: "distance_from_last_transaction" the condition will be that the variable is equal to or above 8.



Create different combinations in order to examine the effect on the percentage of fraud.

The following combinations are based on previous results.

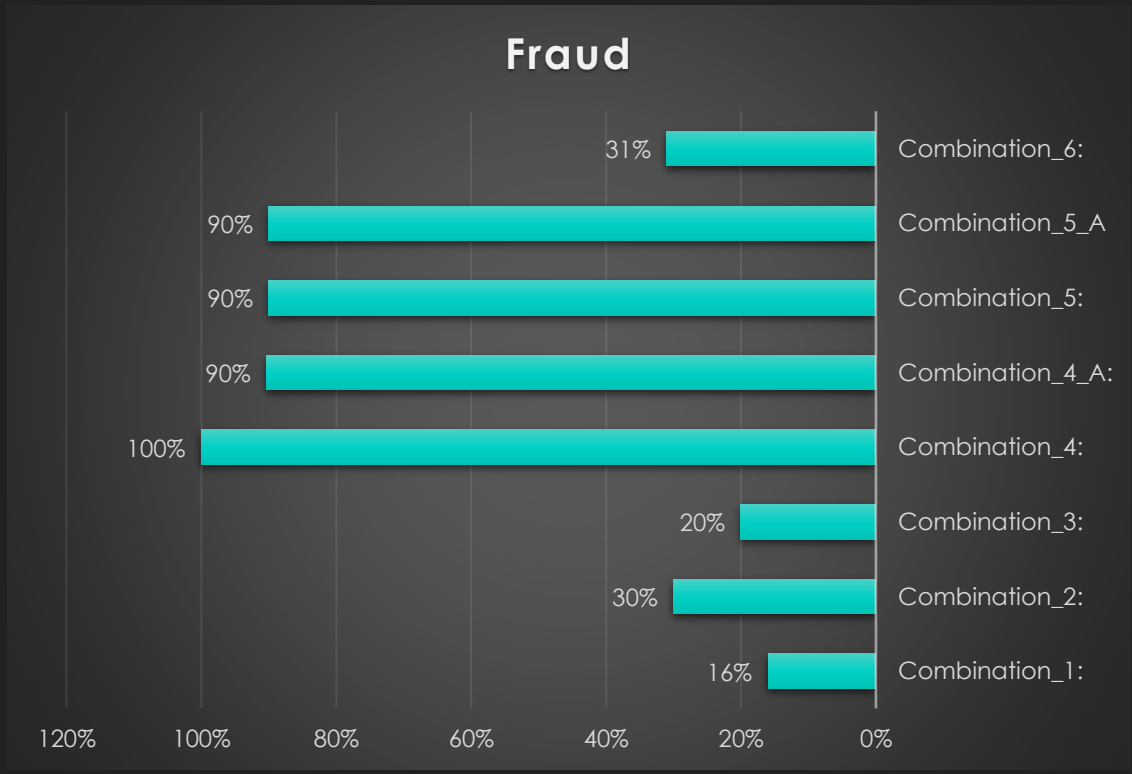
- Combination_1: online_order = 1 & used_pin_number=0 & used_chip = 0
- Combination_2: online_order = 1 & used_pin_number=0 &distance_from_home >=42
- Combination_3: online_order = 1 & used_pin_number=0 &distance_from_last_transaction >=8
- Combination_4: online_order = 1 & used_pin_number=0 &ratio_to_median_purchase_price >=4
- Combination_5: online_order = 1 & ratio_to_median_purchase_price >=4 & distance_from_last_transaction >=8
- Combination_6: online_order = 1 & ratio_to_median_purchase_price >=4 & distance_from_home >=42

During the process, I created two more options:

- Combination_4_A: online_order = 1 & ratio_to_median_purchase_price >=4
- Combination_5_A: online_order = 1 & distance_from_last_transaction >=8

Results

- You can see that there are some combinations that give low results (1-3, 6) and there are combinations where the results are over 90%!
- In addition - I tested combinations of only 2 conditions, after seeing the results to check what the level of impact of the third figure:
- The differences between 4 and 4_A were about 10%. (But still, a high figure)
- The differences between 5 and 5_A were not significant (difference of less than 1%)



Additional data

Before we get to the conclusions and recommendations, I will point out some data that might have shed further light on the main question - is this a fraudulent deal?

- Customer Address and shipping address : To check if there is a match in the shipping address to the customer address?
- Type of purchase: There may be some pattern of purchases (fashion / electronic device / games, etc.) that are mostly scams?
- City - There may be cities where there is a higher likelihood of fraud.

Conclusions and Recommendations

- In presenting recommendations based on the data, the significance of the result should also be taken into account.What is meant by?When we block a customer from a legitimate purchase, we will create a customer who is dissatisfied / disappointed or even angry.On the other hand - it is important for us to reduce the following scams as much as possible.
- Therefore, the final considerations are of the company itself - what is the limit it puts in order to maintain a good customer experience on the one hand and prevent losses on the other.
- When I look at the data - my recommendation is to take the shortest combinations that will give me over 90% which is a scam deal.Therefore, these are the following combinations:
    - Combination_4_A: online_order = 1 & ratio_to_median_purchase_price >=4
    - Combination_5_A: online_order = 1 & distance_from_last_transaction >=8

Extras

- In reality there is no 100%, but companies that do not want to harm their customer experience and still want to reduce scams as much as possible, can use this option:

  Combination_4:
  online_order = 1 & used_pin_number= 0 &ratio_to_median_purchase_price >=4

- In the above analysis I focused on online purchases, but I wanted to see if there is a great combination of data that would give me an indication about scams in offline purchases.
- Unfortunately I did not find a combination that would be a high enough indication of fraud.