**Data Cleaning**

**Remove outliers**



As can be seen from the amount distribution plot, there is an apparent outlier. We check it in the raw data and found it’s not a fraud. This amount is far larger than other normal ones, so we decide to remove it. According to the Data Quality Report (Appendix), all the frauds appear in the transaction type P. Thus, we only include data of type P in our supervised model.

**Fill in missing values**

|  |  |  |  |
| --- | --- | --- | --- |
| **variable** | **# NaN** | **% populated** | **# unique** |
| Recnum | 0 | 100 | 96753 |
| Cardnum | 0 | 100 | 1645 |
| Date | 0 | 100 | 365 |
| Merchnum | 3375 | 96.51174 | 13091 |
| Merch description | 0 | 100 | 13126 |
| Merch state | 1195 | 98.7649 | 227 |
| Merch zip | 4656 | 95.18775 | 4567 |
| Transtype | 0 | 100 | 4 |
| Amount | 0 | 100 | 34909 |
| Fraud | 0 | 100 | 2 |

Notes: # NaN stands for number of missing values. % populated means the percentage that non-missing values take up for each variable. # unique means the number of unique values/categories in this variable.

The above table shows the summary of each variable. Merchnum, Merch state and Merch zip need to be filled in reasonable values.

* Merchnum

Firstly, a fair amount of 0 values are found in Merchum, which makes no sense, so we regard those as missing values. Secondly, after observations we find Merch description corresponds well with Merchnum, so we fill the missing values with the most frequent merchnum in this Merch description. With only 100 more records are filled, we turn to use Merch zip as an indicator and fill in with the most frequent Merch zip.

* Merch zip

We also find that Merch description corresponds very well with Merch zip, so we fill missing zip with the zip number in the same Merch description. It is not uncommon that the same Merchant number and Card number will appear in one zip at most of time. Thus, we fill missing zip with the most frequent zip value in the same Merchnum and Cardnum respectively.

* Merch state

Apparently, a Merch state could be well inferred by using Merch zip, hence, that’s the first method to fill in NA. While most zip and state are missing at the same time, other methods should be used. The same to Merch zip, we use Merch description to help fill missing states in the second step. Lastly, we also find the most frequent state in its Merchnum group and Cardnum group and fill in.

With steps mentioned above to fill in missing values, there are still a small number of records remained missing. But the missing values only represent for less than 2% of all values in each variable, we decide to drop them.

**Candidate Variables Creation**

**Amount expert variables**

Based on our domain knowledge, those amounts larger than normal purchase would be a good signal in detecting frauds. Card number, merchant number and card number at this merchant, in this zip code or in this state, those 5 entity combinations could serve as important groups. Then, for each group, we calculate the average, total amount and how far the actual amount is away from the average and total spent over the past 1, 4, 7, 14 and 30 days.

加图

**Frequency expert variables**

Burst of activity at different merchants or card owners is usually a signal of fraud. Thus, we create frequency expert variables to detect whether the activity is abnormal.

**Days since last purchase variables**

It is not uncommon that if a card is left behind for a long time, the chance to be reused is smaller. We could infer that day intervals since last purchase would be a good variable in detecting the fraud, so for each important group mentioned above, we calculate the time difference between current transaction date and last most recent transaction date.

**Velocity** **deviation expert variables**

These variables aim to compare the number of transactions over the past 1 day and average daily number of transactions over 7, 14 and 30 days in card number and merchant number groups.

**Variable summary**

The table below takes 1 day as an example and describes how each variable is calculated. The same calculation is applied to 7, 14 and 30 days except for 5 actual amounts. In all, we have 166 candidate variables, including 130 amount expert variables, 25 frequency expert variables, 5 days since last purchase variables, 6 velocity deviation expert variables.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Category** |
| Amount\_cn\_actual | The actual total amount spent on each day in this card number | Amount Expert Variables |
| Amount\_mn\_actual | The actual total amount spent on each day in this merchant number |
| Amount\_cn\_mn\_actual | The actual total amount spent on each day in this card number at this merchant number |
| Amount\_cn\_zip\_actual | The actual total amount spent on each day in this card number at this zip |
| Amount\_cn\_st\_actual | The actual total amount spent on each day in this card number at this state |
| Amount\_cn\_1d\_avg | The average of total amount spent over the past 1 day in this card number group |
| Amount\_cn\_1d\_sum | The sum of total amount spent over the past 1 day in this card number group |
| Amount\_cn\_1d\_act-avg | The actual number minus average number of total amount spent over the past 1 day in this card number group |
| Amount\_cn\_1d\_act/avg | The actual number divided by average number of total amount spent over the past 1 day in this card number group |
| Amount\_cn\_1d\_act/sum | The actual number divided by total number of total amount spent over the past 1 day in this card number group |
| Amount\_mn\_1d\_avg | The average of total amount spent over the past 1 day in this merchant number group |
| Amount\_mn\_1d\_sum | The sum of total amount spent over the past 1 day in this merchant number group |
| Amount\_mn\_1d\_act-avg | The actual number minus average number of total amount spent over the past 1 day in this merchant number group |
| Amount\_mn\_1d\_act/avg | The actual number divided by average number of total amount spent over the past 1 day in this merchant number group |
| Amount\_mn\_1d\_act/sum | The actual number divided by total number of total amount spent over the past 1 day in this merchant number group |
| Amount\_cn\_mn\_1d\_avg | The average of total amount spent over the past 1 day in this card number at different merchant numbers number group |
| Amount\_cn\_mn\_1d\_sum | The sum of total amount spent over the past 1 day in this card number at different merchant numbers number group |
| Amount\_cn\_mn\_1d\_act-avg | The actual number minus average number of total amount spent over the past 1 day in this card number at different merchant numbers number group |
| Amount\_cn\_mn\_1d\_act/avg | The actual number divided by average number of total amount spent over the past 1 day in this card number at different merchant numbers number group |
| Amount\_cn\_mn\_1d\_act/sum | The actual number divided by total number of total amount spent over the past 1 day in this card number at different merchant numbers number group |
| Amount\_cn\_zip\_1d\_avg | The average of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_zip\_1d\_sum | The sum of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_zip\_1d\_act-avg | The actual number minus average number of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_zip\_1d\_act/avg | The actual number divided by average number of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_zip\_1d\_act/sum | The actual number divided by total number of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_st\_1d\_avg | The average of total amount spent over the past 1 day in this card number at different states number group |
| Amount\_cn\_st\_1d\_sum | The sum of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_st\_1d\_act-avg | The actual number minus average number of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_st\_1d\_act/avg | The actual number divided by average number of total amount spent over the past 1 day in this card number at different zip number group |
| Amount\_cn\_st\_1d\_act/sum | The actual number divided by total number of total amount spent over the past 1 day in this card number at different zip number group |
| Freq\_cn\_1d | Total number of transactions over the past 1 day in this card number | Frequency expert variables |
| Freq\_mn\_1d | Total number of transactions over the past 1 day in this merchant number |
| Freq\_cn\_mn\_1d | Total number of transactions over the past 1 day in this card number at different merchant numbers |
| Freq\_cn\_zip\_1d | Total number of transactions over the past 1 day in this card number at different zips |
| Freq\_cn\_st\_1d | Total number of transactions over the past 1 day in this card number at different states |
| last\_cn | Days since last transaction in this card number | Days since last purchase variables |
| last\_mn | Days since last transaction in this merchant number |
| last\_cn\_mn | Days since last transaction in this card number at this merchant number |
| last\_cn\_zip | Days since last transaction in this card number at this merchant zip |
| last\_cn\_st | Days since last transaction in this card number at this merchant state |
| velo\_cn\_avg7d | the number of transactions over the past 1 day divided by average daily number of transactions over 7 days at this card number | Velocity deviation expert variables |
| velo\_mn\_avg7d | the number of transactions over the past 1 day divided by average daily number of transactions over 7 days at this merchant number |