**HW2: Fraud Analytics**

**-Rijul Sherathia**

**Data Exploration and Cleaning :**

1. The dataset in CSV format contains several columns with missing values denoted as 'NaN'. We will exclude these columns from our analysis.
2. The data type of the date field in the dataset is currently set as 'object'. We will convert it to the 'datetime' format.
3. To comply with the given requirements, we will retain only the data where the transaction type is 'P’ and remove all other transactions from the dataset.
4. In addition, there is an outlier or spike in the data where the transaction amount exceeds $3000000. We will remove this outlier from the dataset.
5. After cleaning the dataset, we will check the number of rows that contain missing or 'NaN' values.

**Clean and impute ‘merchnum’ field :**

1. Initially, there were 3198 null rows in the 'Merchnum' field. We dropped all the columns with NaN values.
2. Rows with 'Merchnum' as 0 were replaced with NaN values, increasing the count of NaN rows to 3251.
3. We created a dictionary 'merchdes\_merchnum' to store unique non-null 'Merch description' with their corresponding 'Merchnum' values. We skipped rows where 'Merchnum' and 'Merch description' were both null.
4. We replaced all Null or NaN values in the 'Merchnum' column using the 'merchdes\_merchnum' dictionary, reducing the null values from 3198 to 2094.
5. We manually set 'Merchnum' as 'unknown' for all rows with 'Merch description' - 'RETAIL CREDIT ADJUSTMENT' or 'RETAIL DEBIT ADJUSTMENT', further reducing null rows to 1403.
6. We created a new dictionary 'merchnum\_create' to map unique non-null 'Merch description' values to corresponding new unique 'Merchnum' values for rows with missing 'Merchnum' values. We iterated over the dataset to replace all NaN values in the 'Merchnum' column with the corresponding value in the 'merchnum\_create' dictionary.
7. As a result, the number of null/NaN values in the 'Merchnum' column was reduced to 0. The imputation of the 'Merchnum' field is complete.

However, at this point, we observe that except the columns ‘Merch description’, ‘Merch state’ and ‘Merch zip’ all others have 0 null counts and hence we now proceed towards cleaning and imputing these fields.

**Clean and Impute ‘Merch state’ Field :**

1. Zipcodes and states are related, so we'll start by analyzing them.
2. We find 'Merch zip' with non-missing values and 'Merch state' with missing values, and create a dictionary called 'zip\_state' to store unique zipcodes as keys and their corresponding state values.
3. We populate 'zip\_state' with unique values of 'Merch zip' and their corresponding 'Merch state' values where 'Merch zip' has non-missing values. We manually add some known key-value pairs to the dictionary.
4. We create another dictionary called 'merchnum\_state' to store unique 'Merchnum' values and their corresponding 'Merch state' values where 'Merchnum' has non-missing values.
5. We create a third dictionary called 'merchdes\_state' to store unique 'Merch description' values and their corresponding 'Merch state' values where 'Merch description' has non-missing values. We fill the missing 'Merch state' values using these three dictionaries, reducing the number of missing values.
6. We replace missing 'Merch state' values with 'unknown' where the 'Merch description' is 'RETAIL CREDIT ADJUSTMENT' or 'RETAIL DEBIT ADJUSTMENT', further reducing the number of missing values.
7. We store all possible state values in an array and set the value of all foreign states to 'foreign' if they are not in the array.
8. Finally, we replace all remaining null values in the 'Merch state' column with 'unknown', reducing the number of null or missing data in the 'Merch state' column to 0.

**Clean and Impute Merch zip Column:**

To impute missing values in the 'Merch zip' column, the following steps were taken:

1. Created dictionaries 'merchnum\_zip' and 'merchdes\_zip' to store non-null values of 'Merchnum' and 'Merch description' columns as keys and their corresponding 'Merch zip' column values as values.
2. Used the above dictionaries to replace null values in the 'Merch zip' column in accordance with the respective 'Merchnum' and 'Merch description' column values.
3. Replaced missing values in the 'Merch zip' column with 'unknown' where the 'Merch description' column was 'RETAIL CREDIT ADJUSTMENT' or 'RETAIL DEBIT ADJUSTMENT'.
4. Created a dictionary 'state\_zip' to store non-null values of the 'Merch state' column as keys and corresponding non-null value of one of the zip codes belonging to this state.
5. Used the above dictionary to replace null values in the 'Merch zip' column with the zip code from the dictionary according to the 'Merch state' column value.
6. Replaced remaining null values in the 'Merch zip' column with 'unknown'. This resulted in the missing data count for 'Merch zip' being reduced to 0.

**Point 4,5 are newly applied imputation logic.**

**Summary Table:**

|  |  |
| --- | --- |
| **Description** | **# Variables Created** |
| **Date of week target encoded**:  average fraud percentage of that day | 1 |
| **Month target encoded**:  average fraud percentage of that month | 1 |
| **State Risk target encoded**:  Top 15 Fraud Merchant State with highest fraud percentage | 1 |
| **Benford’s Law Card number Variable**:  count the # of first digits in card number | 1 |
| **Benford’s Law Merchant number Variable**:  count the # of first digits in Merchant number | 1 |
| **Benford’s Law Merchant Zip Variable**:  count the # of first digits in Merchant Zip | 1 |
| **Day Since Variable**:  Number of days since application with that entity was seen | 18 |
| **Velocity Variable**:  Number of records with the same entity over the last {0,1,3,7,14,30,60} days | 1134 |
| **Relative Velocity Variable**:  Count of records and total amount with same entities seen in the past {0,1} day divided by the number of applications with those same entities seen in the last {7,14,30,60} days | 288 |
| **Velocity Density Variable**:  Count of records with same entities seen in the past {0,1} day divided by day since those same entities seen in the last {7,14,30,60} days | 144 |
| **Counts by Entity:**  Number of unique entities for a particular field over the last {0,1,3,7,14,30,60} days | 1836 |
| **Cross Entity Uniqueness Variable**:  Number of records with unique combinations of all the entities | 306 |
| **Entity Amount Variability**:  Amount variability with the amount of same entity seen over the last {0,1,3,7,14,30,60} days | 324 |
| **Relative Velocity Variable (squared)**:  Count of records and total amount with same entities seen in the past {0,1} day divided by square of the number of applications with those same entities seen in the last {3,7,14,30} days | 144 |
| **Binning Amount:**  Creates 5 equal sized bins to divide amount column and assigns labels (1,5) to each bin | 1 |

Additional Variables Created:

1. Benford’s Law Merchant Zip Variable.
2. Month target encoded.