Credit Card Transaction Data Fraud Detection Report

USC Marshall, DSO 562
Team 112 Members: Freda Lin, Jingchen Liu, Pizheng Zhang, Yafan Zeng
05/05/2022

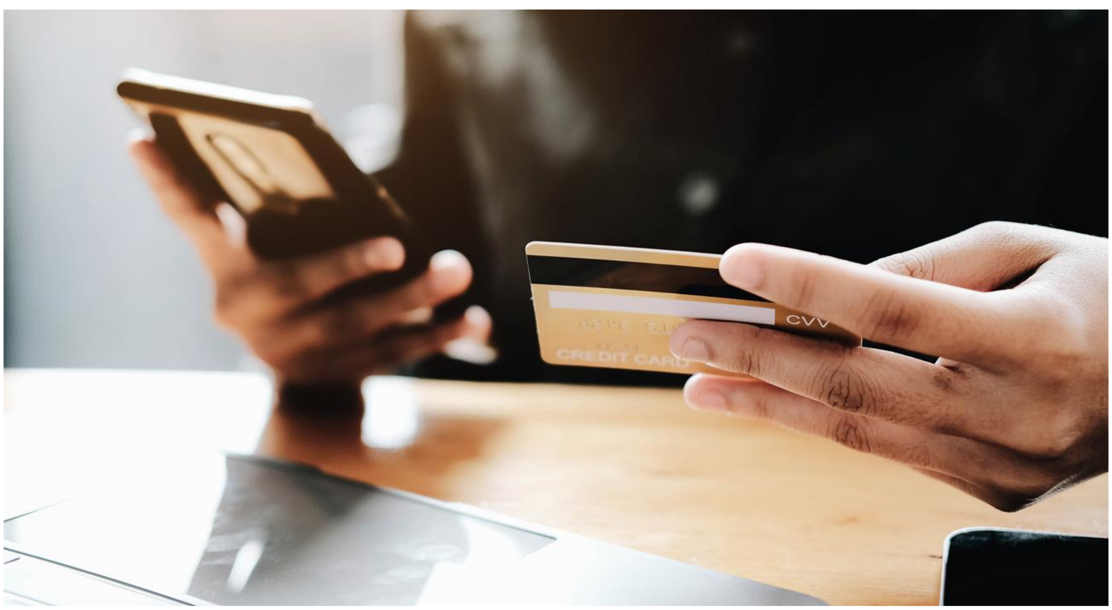


Table of Contents

Executive Summary

The purpose of conducting this project is to identify signals of credit card transaction fraud and apply data analytics and machine learning techniques to detect credit card frauds. According to the Nilson report, the U.S. has the greatest number of cases of credit card fraud worldwide, and the number of credit card fraud occurrences has risen 161 percent since 2015. An American credit card preference study has shown that a third of Americans have been victims of credit card fraud, so it is important to research and detect credit card fraud. Some common methods of credit card fraud include lost credit cards, “card not present” fraud, phishing, card skimming, and online account hacking.

This report examines and analyzes the best model to predict credit card transaction fraud with supervised machine learning algorithms and analytics tools. The final model algorithm can be scaled up to detect fraud in real-world problems. The data source of this project is from a US government agency in 2006 provided by Professor Stephen Coggeshall. This report contains descriptions of data, the data cleaning process, the mechanism to create candidate variables, the feature selection process, model algorithms, results, conclusions, and an appendix. In deploying model algorithms and fitting data, Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree, and Neural Network were used to select the best models for Project 3.

The model selected for the final algorithm is Random Forest. The Random Forest model has 20 variables with 12 trees, 14 layers, a split criterion of Gini impurity, a min_sample_leaf of 70, a min_sample_split of 10, and a max number of features used in each tree of 20. The model has an average fraud detection rate (FDR) of 80.73% for the training set, 76.11% for the testing set, and 58.10% for the Out-of-Time (OOT) set at a 3% cutoff. And we have concluded that the best selected Random Forest model can maximize the overall savings and eliminate 58.10% of the fraud by declining 3% of the transactions.

Description of Data

Data Description
The transaction dataset is a dataset of credit card transactions from a US government agency in 2006 with synthetic fraud labels. There are 10 fields and 96,763 records in the dataset.

The following table has the summary statistics of all numeric fields. There are two numeric fields in this dataset: the date of the transaction and the amount of each transaction. The mean of 'Amount' is significantly larger than the median, indicating that the distribution of 'Amount' is right-skewed, which might be caused by some extremely large outliers.

Table 1. Summary Table for Numeric Fields

Field Name	% Populated	Min	Max	Mean	Stdev	Median
Date	100	2006-01-01	2006-12-31	N/A	N/A	2006-06-27
Amount	100	0.01	3,102,045.53	427.89	10,006.14	137.98

The following table contains the summary statistics of all categorical fields. There are eight categorical variables in this dataset: the unique record number of the transaction, the card number of the transaction, the merchant number, the merchant description, merchant state, merchant zip code, the transaction type, and the fraud label. We noticed that there are missing values in the 'Merchnum', 'Merch state', and 'Merch zip' fields.

Table 2. Summary Table for Categorial Fields

Field Name	% Populated	# Unique Values	Most Common Value
Recnum	100.00	96,753	N/A
Cardnum	100.00	1,645	5142148452
Merchnum	96.51	13,091	930090121224
Merch description	100.00	13,126	GSA-FSS-ADV
Merch state	98.76	227	TX
Merch zip	95.19	4,567	38118.0
Transtype	100.00	4	P
Fraud	100.00	2	0

Detailed Description of Important Variables

Fraud

This field classifies if the transaction is a fraud or not. A value of 1 represents a fraud application, and a value of 0 represents a regular application. The following plot illustrates the distribution of application fraud.

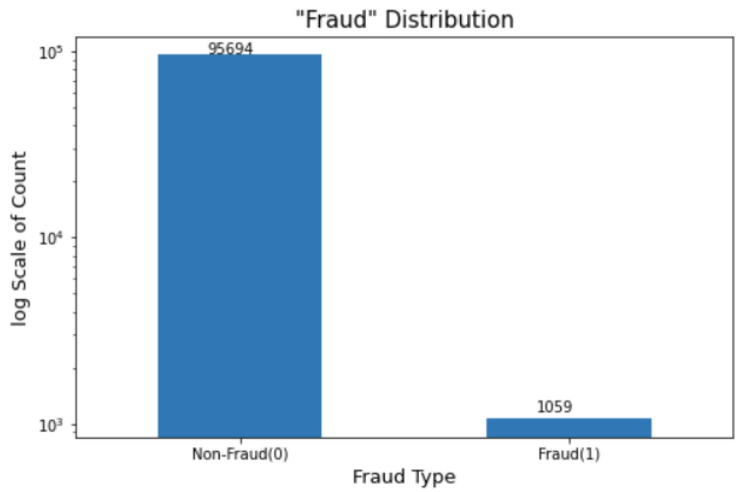


Figure 1. Fraud Label Distribution

The second and third graphs demonstrate the weekly distribution of transactions. The left graph shows the seasonality of transactions. There is an obvious drop in October because the government agency fiscal year starts in October, and there is an intention that employees would spend less at the beginning of the year and spend more toward the end of the year. The right graph separates the transaction distribution based on its fraud label. The green line represents the distribution of the good transactions, and the red line represents the distribution of the bad transactions. It clearly indicates that there is more fluctuation in the fraud transaction distribution.

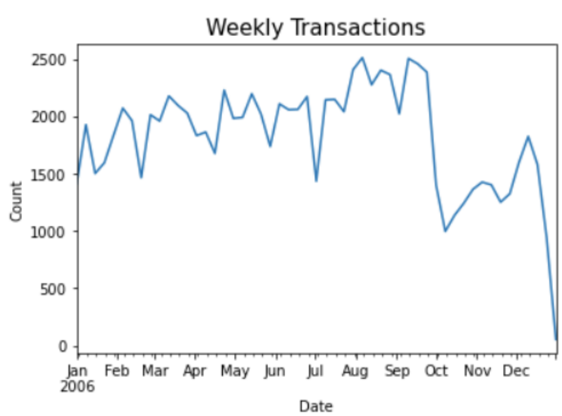


Figure 2a. Weekly Transaction Count

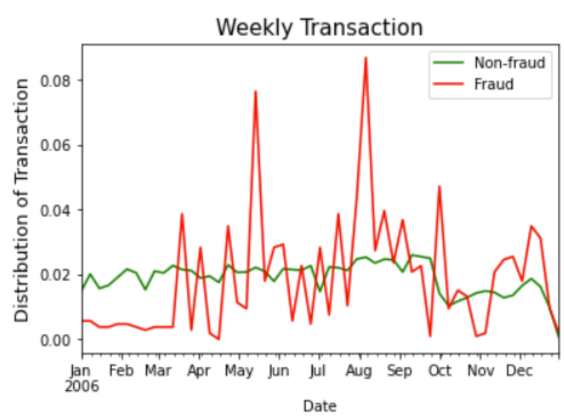


Figure 2b. Weekly Transaction Count by Fraud

Merchnum

This field contains the merchant number of each card transaction. There are 13,092 unique merchant numbers in this data set. The most frequent merchant number is 930090121224, and it appears 9310 times in the dataset, which indicates a higher likelihood of fraud. The following graph shows the distribution of the top 15 most common merchant numbers.

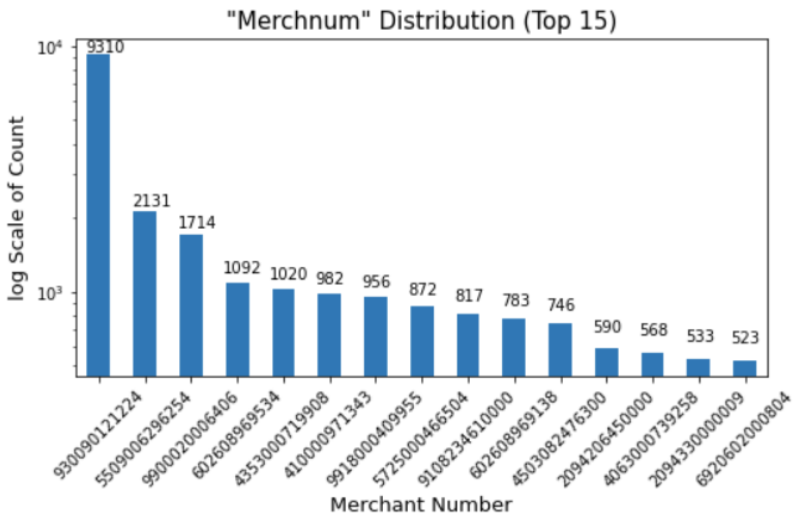


Figure 3. Merchant Number Distribution

Merch zip

This field contains the merchant zip code of each card transaction. There are 4,568 unique merchant numbers in this data set. The most frequent merchant number is 38118.0, and it appears 11,868 times in the dataset. This has a significantly higher frequency than other zip codes, and it might be a signal of fraud. The following graph shows the distribution of the top 15 most common merchant zip codes.

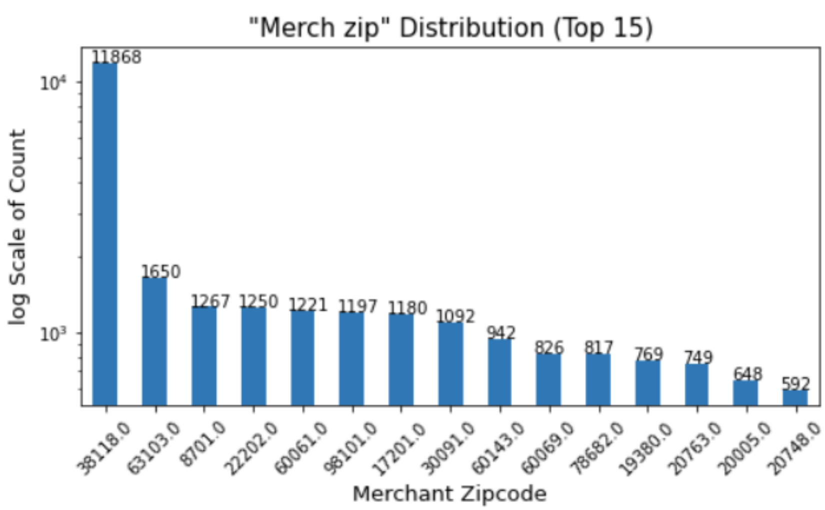


Figure 4. Merchant Zip Distribution

Data Cleaning

<u>Dropping Records</u>

The purpose of data cleaning is to remove duplicate information, fix existing errors, and provide data consistency. After a close look at the dataset, we noticed that there is an outlier in the amount field which has an extremely high value, so we removed this data point to avoid it influencing our model result. Moreover, because the model was focusing on purchase transactions, we dropped all transaction records other than "P" in transaction type. After the initial cleaning, 356 rows were removed, making the dataset a shape from (96753,10) to (96397,10).

<u>Imputation</u>

In the data description section, we also identified three fields with missing values: Merchnum, Merch state, and Merch zip. In order to keep as many records as possible for model training, we used the following imputation steps to fill in the missing values. During the imupation process, we have also found a number of transactions that has a single-digit zip code or a merchant number of 0, which is clearly incorrect. Therefore, we replaced these entries with Nan and re-assigned their values based on the imputation logic explained below.

1. Merch state (1,020 missing)
   To impute the missing values, the following steps were taken:
   - For records with available zip codes in the range 600-799 or 900-999, assign "PR" (Puerto Rico) as the state (980 missing)
   - For remaining records with zip codes, match the zip code with the U.S. zip_state dataset (imported separately) and assign the corresponding state (953 missing)
   - Use the mode of Merchnum to fill in missing states (924 missing)
   - Use the mode of Merch description to fill in missing states (268 missing)
   - Use the mode of Cardnum to fill in missing states (38 missing)
   - Assign "Unknown" to the rest of the missing states (0 missing)

2. Merchnum (3,198 missing)
   To impute the missing values, the following steps were taken:
   - Replace Merchnum with "0" to Nan (3,251 missing)
   - Use the mode of Merch description to fill in missing Merchnum (2,094 missing)
   - Use the mode of Cardnum to fill in missing Merchnum (57 missing)
   - Use the mode of Merch state to fill in missing Merchnum (38 missing)
   - Assign "Unknown" to the rest of the missing Merchnum (0 missing)

3. Merch zip (4,300 missings)
   To impute the missing values, the following steps were taken:
   - Replace single-digit zip codes with Nan (4,412 missing)
   - Use the mode of Merchnum to fill in missing zips (797 missing)
   - Use the mode of Cardnum to fill in missing zips (36 missing)
   - Use the mode of Merch state to fill in missing zips (33 missing)
   - Use the mode of Merch description to fill in missing zips (31 missing)
   - Assign "Unknown" to the rest of the missing zips (0 missing)

The following tables show the first 5 rows of the cleaned dataset and the data type of each column:

Table 3. Cleaned Dataset

	record	date	ssn	firstname	lastname	address	zip5	dob	homephone	fraud_label
0	1	2016-01-01	379070012	XRRAMMTR	SMJETJMJ	6861 EUTST PL	02765	000000-1	1797504115	0
1	2	2016-01-01	387482503	MAMSTUJR	RTTEMRRR	7280 URASA PL	57169	19340615	4164239415	1
2	3	2016-01-01	200332444	SZMMUJEZS	EUSEZRAE	5581 RSREX LN	56721	000000-3	0216537580	0
3	4	2016-01-01	747451317	SJJZSXRSZ	ETJXTXXS	1387 UJZXJ RD	35286	19440430	0132144161	0
4	5	2016-01-01	024065868	SSSXUEJMS	SSUUJXUZ	279 EAASA WY	03173	19980315	6101082272	0

Candidate Variables

<u>Build Candidate Variables</u>

After completing the data cleaning process, we built a series of potential candidate variables. We started the process by creating two numerical variables for the day of the week and the state of each transaction. The purpose of encoding is to convert categorical data into numerical data. In order to prevent increasing the dimension of the variables, we chose the target encoding. We used the first ten months' data to calculate the fraud proportion of each weekday. The numerical values were based on the average fraud label on each category and were smoothed afterward. Figure 5 below indicates the relationship between the smoothed fraud proportion and the day of the week.

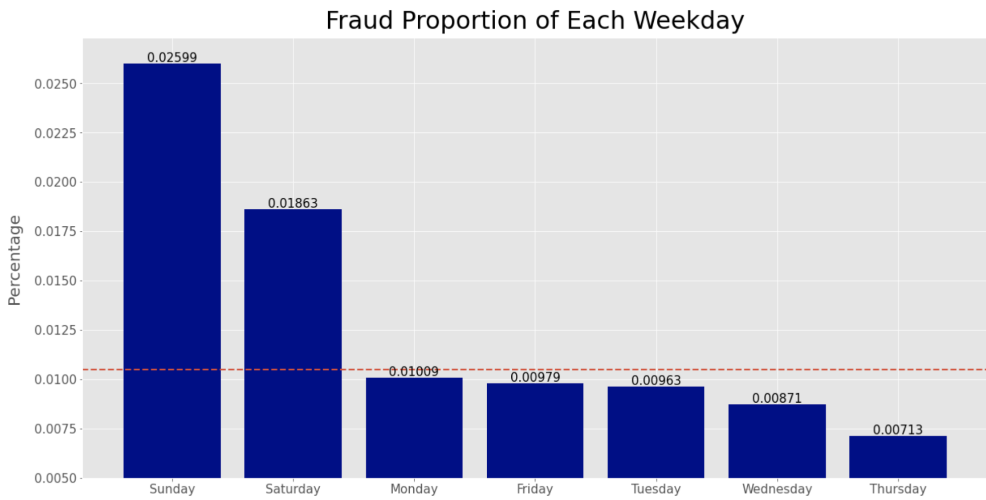


Figure 5. Target Encoding for Day of Week

A similar target encoding was done for each state. The numerical values were created based on the average fraud label of each state and were smoothed afterward. Figure 6 below indicates the relationship between the smoothed fraud proportion and 20 transaction states with the highest fraud proportion values.

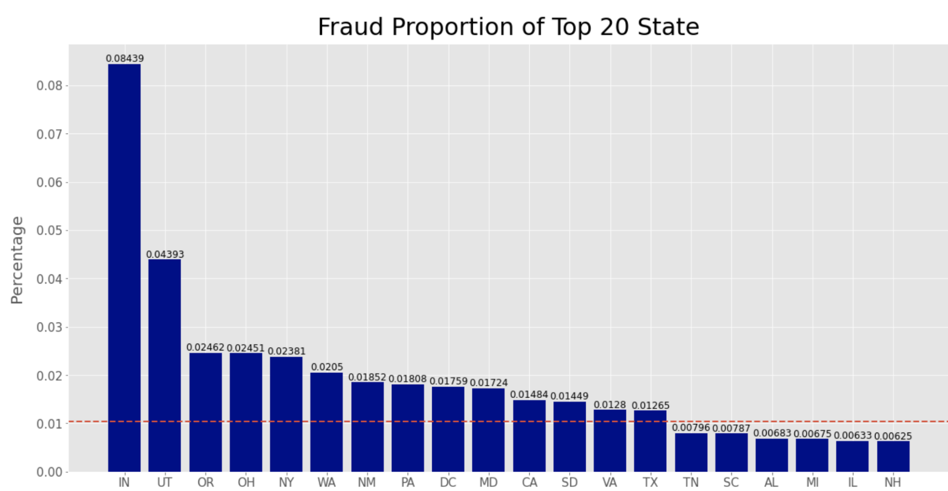


Figure 6. Target Encoding for Transaction State

After target encoding, two Benford's Law variables were created. Benford's Law is the non-intuitive fact that the first digit of many measurements is not uniformly distributed, and the first digit "1" appears about 30% of the time. Thus, our team designed 2 variables on the distribution of the first digit of transaction amount by cardnum and merchnum, to examine if they follow Benford's Law. In order to build Benford's Law variable, all transactions from FedEx were excluded because our team believed that the transaction amount at FedEx was more uniform and might not follow the rule. All the transaction records were then grouped separately by cardnum and merchnum to show the distribution of the first digit of the transaction amounts and measure potential unusualness. Although the original Benford's Law distribution has nine bins corresponding to each digit, our team divided the distribution into two bins: nLow which included records beginning with 1 or 2, and nhigh which included records beginning with 3 to 9 in order to have enough representation in each group. We calculated the distribution ratio with$R = \frac{1.096 * nlow}{nhigh}$, and Unusualness = $max(1, \frac{1}{R})$. In order to reduce the variance of unusualness, we smoothed unusualness by $U^* = 1 + (\frac{U-1}{1+exp^{-t}})$ and $t = \frac{n-15}{3}$ as our final Benford's Law variables. Benford's variables for cardnum and merchnum are shown in the tables below.

Table 4. Benford's Variable for Card Number

Top 40 Cardnum

Cardnum	n	n_low	n_high	R	1/R	U	t	U*
5142253356	66	61	5	13.37	0.07	13.37	17.00	13.37
5142299705	28	25	3	9.13	0.11	9.13	4.33	9.03
5142197563	149	15	134	0.12	8.15	8.15	44.67	8.15
5142194617	38	5	33	0.17	6.02	6.02	7.67	6.02
5142288241	14	1	13	0.08	11.86	11.86	-0.33	5.53

Table 5. Benford's Variable for Merchant Number

Top 40 Merchnum

Merchnum	n	n_low	n_high	R	1/R	U	t	U*
991808369338	181	1	181	0.01	165.15	165.15	55.33	165.15
8078200641472	60	59	1	64.66	0.02	64.66	15.00	64.66
308904389335	53	1	53	0.02	48.36	48.36	12.67	48.36
3523000628102	34	34	1	37.26	0.03	37.26	6.33	37.20
808998385332	37	1	36	0.03	32.85	32.85	7.33	32.83
55158027	28	27	1	29.59	0.03	29.59	4.33	29.22
8916500620062	31	1	31	0.04	28.28	28.28	5.33	28.15

After creating the target encoding and Benford's Law variables, our team created candidate variables based on our understanding of potential fraud signals. Some common signals of fraud we identified were:

- Purchase amount is too big, and it's larger than normal purchase amounts at same or different merchants
- Same card is used at vary state or zip code or used at very different geographical location
- Burst of activity at different merchants but use the same card
- Infrequent recurring charges at the same merchants or have the same amount.

Since the transaction amount and card number could be significant signals of fraud, our team created 14 new entities based on these two fields: *'card_merch', 'card_state', 'card_zip', 'card_transtype', 'Merch_address', 'Merchnum_state'. 'Merchnum_zip', 'Date_state', 'Date_zip', 'Amount_date',' Amount_description', 'Amount_state', 'Amount_zip', 'Amount_transtype'* .

A few groups of candidate variables were then created based on the new entities identified above.

- *Day since:* By calculating the number of days since an application with the same entity has been seen, we were able to tell how long a repeated entry or entity has been recorded in the system since the last time someone tried to enter the entity. If the combination of entries is unique and new, then it would return a value of 365 days for "# days seen".

- *Frequency*: The number of transactions at one entity over the past specific number of days. We chose the record with the same variables over the last 0, 1, 3, 7, 14, and 30 days, and counted the number of their appearances.

- *Amount variables*: The total, maximum, mean, and median of the transaction amount at one entity over the past 0, 1, 3, 7, 14, and 30 days.

- *Amount frequency*: The frequency of actual total, maximum, mean, and median of the transaction amounts at one entity over the past 0, 1, 3, 7, 14, and 30 days.

- *Relative velocity*: The number of transactions by one entity within a short period of time (same day or past day) compared to the number of transactions at that entity over a longer time window of 0, 1, 3, 7, 14, and 30 days.

<ins>List of Candidate Variables</ins>

During the variable creation process, a total of 1,177 candidate variables were created. The following table is the summary of variables created (the full list of all variables created is included in the appendix).

Table 6. Summary of Candidate Variables Created

Description of Variables	# Variables Created
Create New Entities ['card_merch', 'card_state', 'card_zip', 'card_transtype', 'Merch_address', 'Merchnum_state'. 'Merchnum_zip', 'Date_state', 'Date_zip', 'Amount_date',' Amount_description', 'Amount_state', 'Amount_zip', 'Amount_transtype']	14
# days since a transaction with that entity has been seen. Entities are ['Cardnum', 'Merchnum', 'Merch description','Merch state', 'Merch zip','card_merch', 'card_state', 'card_transtype', 'Merch_address', 'Merchnum_state', 'Merchnum_zip',	19

'Amount_transtype', 'Date_state','Date_zip', 'Amount_description', 'card_zip', 'Amount_date','Amount_state', 'Amount_zip']	
# transaction at that entity over the past n days. Entities are [‘card_merch’, ‘card_state’, ‘card_zip’, ‘card_transtype’, ‘Merch_address’, ‘Merchnum_state’. ‘Merchnum_zip’, ‘Date_state’, ‘Date_zip’, ‘Amount_date’,’ Amount_description’, ‘Amount_state’, ‘Amount_zip’, ‘Amount_transtype’], n is {0,1,3,7,14,30}	114
# Amount of mean, max, median, sum number of the entities has been seen over the past n days. Entities are [‘card_merch’, ‘card_state’, ‘card_zip’, ‘card_transtype’, ‘Merch_address’, ‘Merchnum_state’. ‘Merchnum_zip’, ‘Date_state’, ‘Date_zip’, ‘Amount_date’,’ Amount_description’, ‘Amount_state’, ‘Amount_zip’, ‘Amount_transtype’], n is {0,1,3,7,14,30}	456
# Frequency of actual mean, max, median, sum number of the entities has been seen over the past n days. Entities are [‘card_merch’, ‘card_state’, ‘card_zip’, ‘card_transtype’, ‘Merch_address’, ‘Merchnum_state’. ‘Merchnum_zip’, ‘Date_state’, ‘Date_zip’, ‘Amount_date’,’ Amount_description’, ‘Amount_state’, ‘Amount_zip’, ‘Amount_transtype’], n is {0,1,3,7,14,30}	456
# Transaction at that entity seen in a short time window {today:1 or past day:0} compared to # applications at that entity over a longer time window {past 7,14,30}. Entities are [‘card_merch’, ‘card_state’, ‘card_zip’, ‘card_transtype’, ‘Merch_address’, ‘Merchnum_state’. ‘Merchnum_zip’, ‘Date_state’, ‘Date_zip’, ‘Amount_date’,’ Amount_description’, ‘Amount_state’, ‘Amount_zip’, ‘Amount_transtype’	114
Targeting encoding for the day of the week.	1
Targeting encoding for the transaction state	1
Banford’s law variable: unusualness of transaction amount distribution by card number and merchant number	2
Total created variables	**1,177**

Feature Selection Process

Feature selection is extremely important because it helps eliminate redundant variables and deals with overfitting or underfitting in order to improve the model accuracy rate. We divided the progress into two steps in terms of feature selection progress: filter and wrapper.

When applying filters to the feature selection, we sorted all variables by their importance to predict whether the record is fraud or not. We chose the KS and fraud detection scores to represent the variables' importance.

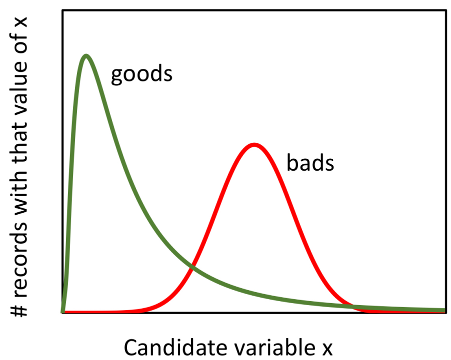


Figure 7a. KS Score

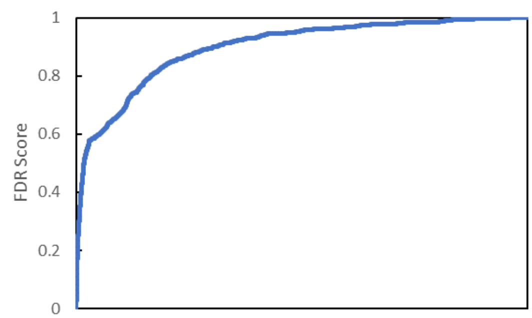


Figure 7b. Fraud Detection Score

We used the univariate Kolmogorov-Smirnov (KS) method among various filter methods because it works well for binary classification problems. For each candidate variable, we plotted the good and bad records separately (as illustrated in figure 7a). We used the formula $Max \int_{x\,min}^{X} [P_{goods}-P_{bads}]\, dx$ to measure the separation between these two curves and get the univariate KS scores and a higher KS score implies a stronger predictor that we want. We also included the fraud detection score in the filter process. We sorted the dataset by KS values for each variable and then calculated the number of fraud transactions detected from the first 3%. We used the mean of the KS score and FDR score as our filter criterion and kept the top 80 variables that have the highest filtering scores.

In the wrapper process, we used forward stepwise regression to find the best subset of 20 variables from the top 80 variables we created in the previous step. The forward selection process looks for the best subset of variables using a boosted tree algorithm starting at the size of 1, and repeats the process by adding one additional predictor for each iteration to find the best subset. The final 20 variables were selected by finding the best 20 variables with the best classification performance. During the forward selection, we used the fraud detection rate at 3% as our reference to measure the goodness of each model, which was calculated by dividing the number of correctly predicted frauds over the total number of actual frauds.

The final 20 variables selected from the wrapper were then used as our best variables to predict whether the record is fraud or not in the following model exploration process.

List of Final Variables

Table 7. List of Final 20 Variables

	Top 20 Variables (by importance)	Description	KS Score	FDR Score	Final Score
1	Card_state_zip_total_3	Total amount of transactions using that card number, merchant state and merchant zip over the past 3 days	0.6778	0.6429	0.6603
2	Card_zip_total_3	Total amount of transactions using that card number and merchant zip over the past 3 days	0.6776	0.6429	0.6602
3	Card_Merchnum_total_3	Total amount of transactions using that card number, merchant number over the past 3 days	0.6751	0.6313	0.6532
4	Card_state_total_3	Total amount of transactions using that card number and merchant state over the past 3 days	0.6742	0.6302	0.6522
5	Card_state_total_7	Total amount of transactions using that card number and merchant state over the past 7 days	0.6697	0.5968	0.6332
6	Card_state_total_1	Total amount of transactions using that card number and merchant state over the past 1 day	0.6591	0.6014	0.6302
7	Card_state_zip_total_1	Total amount of transactions using that card number, merchant state and merchant zip over the past 1 day	0.6607	0.5991	0.6299
8	Card_zip_total_1	Total amount of transactions using that card number and merchant zip over the past 1 day	0.6605	0.5991	0.6298
9	Card_Merchnum_total_1	Total amount of transactions using that card number, merchant number over the past 1 day	0.6582	0.5979	0.6281
10	Card_state_total_14	Total amount of transactions using that card number and merchant state over the past 14 days	0.6691	0.5219	0.5955
11	Card_state_total_0	Total amount of transactions using that card number and merchant state over the past 0 days	0.6106	0.5622	0.5864
12	Card_state_zip_total_0	Total amount of transactions using that card number, merchant state and merchant zip over the past 0 days	0.6100	0.5576	0.5838
13	Card_zip_total_0	Total amount of transactions using that card number and merchant zip over the past 0 days	0.6099	0.5576	0.5838

14	Cardnum_total_7	Total amount of transactions using that card number over the past 7 days	0.6001	0.5184	0.5593
15	Merch description_total_1	Total amount of transactions using that merchant description over the past 1 day	0.6117	0.4724	0.5420
16	Card_state_total_30	Total amount of transactions using that card number and merchant state over the past 30 days	0.6356	0.4447	0.5402
17	Card_state_max_30	Maximum amount of transactions using that card number, merchant state and merchant zip over the past 30 days	0.5982	0.4781	0.5382
18	Merchnum_state_zip_total_1	Total amount of transactions using that merchant number, merchant state and merchant zip over the past 1 day	0.6032	0.4677	0.5355
19	Merch description_total_3	Total amount of transactions using that merchant description over the past 3 days	0.6249	0.4182	0.5215
20	Merchnum_state_total_3	Total amount of transactions using that merchant number and merchant state over the past 3 days	0.6107	0.4205	0.5156

Model Algorithms

A total of 5 model algorithms have been tested after the feature selection process. During the modeling process, the first two weeks' records have been excluded to avoid variables that were not well-formed. The last two months' records have been taken out as the out-of-sample dataset, and the remainder was randomly split into 70% training set and 30% test set for model fitting. Each model algorithm went through multiple iterations with a different number of variables and/or hyperparameters for the best fraud detection performance, and the final best model was selected after comparing the best models of each algorithm.

A list of detailed algorithm explanations is provided below.

Logistic Regression
This model is the baseline for the model selection process. The logistic regression does a logistic transformation to a linear regression surface to solve binary classification problems. It has a regression function of y = $\frac{1}{1+e^{-\Sigma b_i x_i}}$ that returns the probability of y=1. Logistic regression has a maximum output value of 1 and a minimum output value of 0.

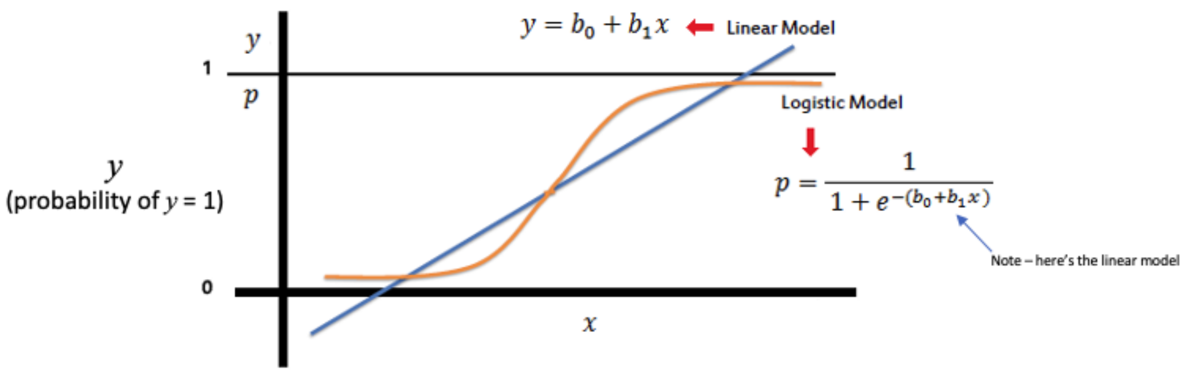


Figure 8. Logistic Regression Algorithm

Decision Tree
Decision tree divides an independent variable space into boxes and assigns a model output to each box by averaging the dependent variable y in that box. After passing the minimum number of points for a leaf, this model algorithm makes a decision to "cut" in this dimension by finding the lowest total impurity in the 2 boxes. This cut will continue until it cannot slide anymore due to restrictions, and different measures of impurity such as Gini or Entropy can be selected during the model fitting process. This process is repeated in all dimensions during fitting. The decision tree works better with model separation lines parallel to axes than curved ones.

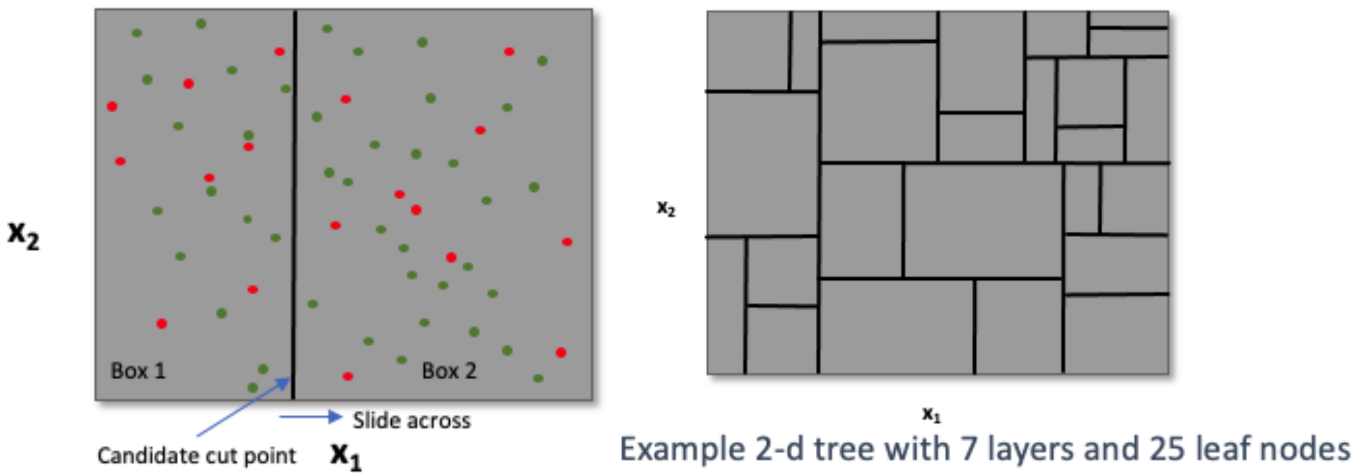


Figure 9. Decision Tree Algorithm

<u>Random Forest</u>

Random Forest is an improved algorithm of Division Tree which builds a collection (ensemble) of many independent strong trees, and the final output is the average (for regression) or voting (for classification) of the outputs of each tree. For each individual tree in the algorithm, it uses a randomly chosen subset of records to ensure some randomness in each tree.

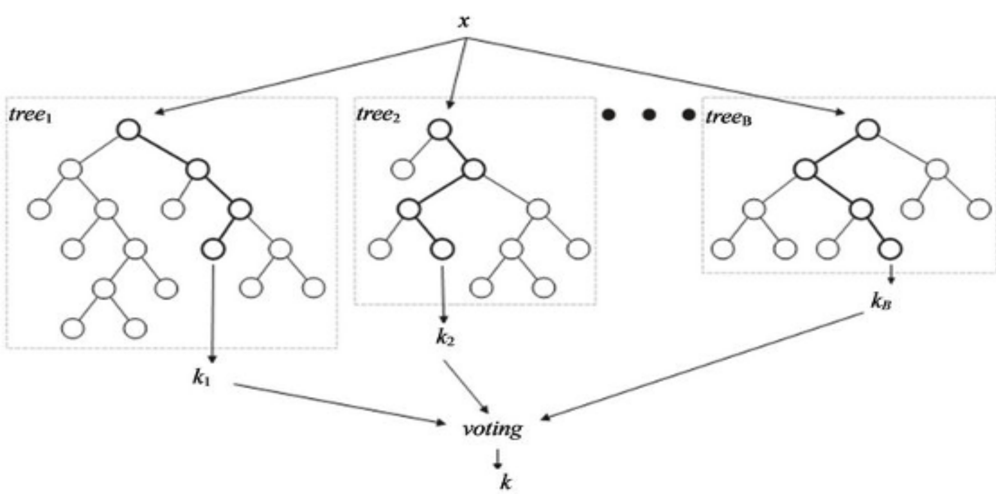


Figure 10. Random Forest Algorithm

Boosted Tree

Boosted Tree is also an evolved algorithm of Division Tree which builds a collection of many weak trees sequentially with each tree adding corrections to the residual error of the current sum. The final boosted tree is a linear combination of all the weak trees.

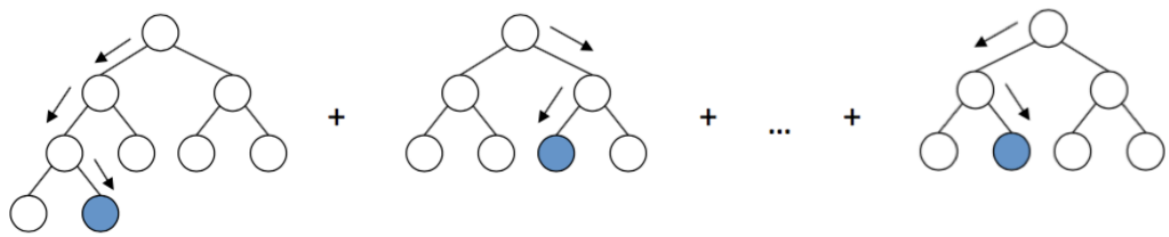


Figure 11. Boosted Tree Algorithm

Neural Network

Neural Network mimics how a neuron in the human brain works by sending variables through the algorithm to get outputs. A neural network model consists of an input layer, some hidden layers, and an output layer. The input layer contains all the independent variables that can get linearly combined for the hidden layer. Nodes in the hidden layer then utilize different activation functions to transform the inputs multiple times based on the number of epochs and generate final outputs.

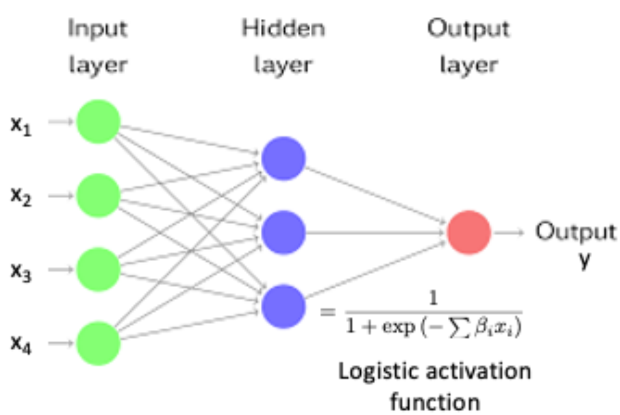


Figure 12. Neural Network Algorithm

Model Exploration

During the model exploration, various supervised machine learning models were explored after the final 20 variables have been selected. All the variables were standardized first to avoid discrepancies in scales. The last two months of data were taken out as the out-of-time (OOT) set, and the rest was kept as the training and testing set. The training and testing set was then split into 30% as the testing set and 70% as the training set for model fitting.

During the fitting process, a series of model algorithms were tested and hypertuned for the best model performance. For each model, the simple and default parameter values were used first. The hyperparameters were then tuned to increase the model complexity until overfitting occurred. During the process of each tuning trial, the model performance, such as the OOT FDR, was derived by running the same model 10 times and taking the average since the dataset and the number of bads was not large enough. In addition, manual cross-validation has been applied for each model by re-splitting the training and testing set to ensure the consistency of model results without compromising the training set size.

Below is the summary table for the model tuning and selection process.

Table 8. Model Exploration and Performance

Model	Parameters								Average FDR at 3%		
	Iteration	# Variables	penalty	C	max_iter	random_state	solver		Train	Test	OOT
Logistics Regression	1 (default)	15	l2	1	100	none	lbfgs		59.51%	60.09%	24.58%
	2	15	l2	1	1000	none	lbfgs		59.80%	59.75%	23.80%
	3	15	l2	1	500	none	lbfgs		59.51%	60.13%	25.25%
	4	15	l2	2	500	none	lbfgs		59.91%	60.59%	25.53%
	5	15	l2	5	500	none	lbfgs		60.45%	60.06%	25.59%
	6	15	l2	8	500	none	lbfgs		60.05%	59.24%	25.92%
	7	15	l2	10	500	none	lbfgs		60.60%	58.64%	24.08%
	9	15	l2	8	800	none	lbfgs		60.23%	58.99%	25.53%
	10	15	none	8	500	none	lbfgs		59.78%	60.13%	25.47%
	11	15	none	8	500	none	sag		60.36%	58.83%	24.69%
	12	15	none	8	500	none	saga		60.07%	59.78%	25.42%
	13	15	none	8	500	1	saga		59.78%	60.11%	23.85%
	14	15	none	8	500	5	saga		60.02%	59.72%	24.08%
	15	15	none	8	500	10	saga		60.15%	59.37%	24.30%
	16	15	none	8	500	10	saga		60.15%	59.37%	24.30%
	17	20	l2	1	500	none	lbfgs		59.64%	58.74%	24.64%
	18	20	l2	5	500	none	lbfgs		59.77%	59.71%	25.14%
	19	20	l2	8	500	none	lbfgs		59.85%	59.12%	24.64%
	20	20	l2	5	1000	none	lbfgs		59.08%	59.50%	23.46%
	21	20	none	8	500	none	lbfgs		59.43%	60.78%	25.14%
	22	20	none	8	1000	none	lbfgs		60.13%	58.88%	24.80%
	Iteration	criterion	max_depth	max_features	min_samples_leaf	min_samples_split	min_weight_fraction_leaf	splitter	Train	Test	OOT
Decision Tree	1 (default)	gini	None	None	1	2	0	best	100.00%	58.48%	28.88%
	2	gini	None	auto	10	4	0.1	best	20.72%	21.13%	8.88%
	3	gini	None	auto	10	4	0,5	best	4.81%	4.62%	8.72%
	4	gini	None	sqrt	10	4	0,1	best	21.06%	21.00%	15.98%
	5	gini	None	sqrt	100	4	0,1	best	21.22%	21.32%	12.85%
	6	gini	None	sqrt	60	100	0.1	best	21.89%	21.79%	16.48%
	7	gini	None	sqrt	60	300	0.1	best	20.22%	21.46%	11.51%
	8	gini	None	sqrt	60	100	0.1	random	3.81%	3.87%	9.55%
	9	gini	None	sqrt	60	20	0.1	ransom	5.31%	4.98%	11.34%
	10	entropy	None	None	50	100	0	best	84.94%	74.81%	47.43%
	11	entropy	None	None	50	100	0.1	best	22.05%	22.31%	16.70%
	12	entropy	None	None	50	150	0	best	83.81%	74.80%	46.70%
	13	entropy	None	auto	50	100	0	best	80.91%	72.47%	40.50%
	14	entropy	None	None	60	100	0	best	83.78%	75.02%	48.44%
	15	entropy	None	None	70	100	0	best	82.13%	75.91%	48.83%
	16	entropy	None	None	75	100	0	best	82.99%	74.84%	47.21%
	17	entropy	None	None	80	100	0	best	81.76%	73.64%	43.97%
	18	entropy	None	sqrt	70	100	0	best	79.11%	72.58%	41.34%
	19	entropy	None	auto	70	100	0	best	78.92%	73.55%	39.22%
	20	entropy	None	auto	70	100	0	random	71.24%	67.34%	38.55%
	21	entropy	None	None	71	100	0	best	82.71%	75.22%	48.88%
	22	entropy	None	None	72	100	0	best	82.54%	74.35%	50.61%
	23	entropy	None	None	73	100	0	best	82.33%	74.51%	50.39%
	24	entropy	None	None	72	110	0	best	82.73%	75.48%	43.35%
	25	entropy	None	None	72	95	0	best	82.50%	74.86%	50.50%
	26	entropy	None	None	72	98	0	best	82.41%	74.35%	45.47%
	27	entropy	5	None	72	20	0	best	74.03%	70.48%	53.63%
	28	entropy	5	None	72	21	0	best	72.59%	71.44%	50.84%
	29	entropy	4	None	72	20	0	best	67.91%	65.07%	43.63%

	Iteration	boosting_type	num_leaves	max_depth	learning_rate	n_estimators	min_child_samples	Train	Test	OOT
Gradient Boosted Tree	1 (default)	gbdt	31	-1	0.1	100	20	99.04%	79.49%	38.99%
	2	gbdt	100	-1	0.1	100	20	100.00%	78.23%	42.51%
	3	gbdt	100	5	0.1	100	20	94.72%	79.51%	48.94%
	4	gbdt	150	5	0.1	100	20	94.70%	78.73%	42.40%
	5	gbdt	150	8	0.1	100	20	99.71%	77.12%	38.10%
	6	gbdt	150	5	0.1	120	20	96.05%	80.04%	44.53%
	7	gbdt	50	5	0.1	50	20	90.58%	78.71%	49.16%
	8	gbdt	50	3	0.1	50	20	80.67%	76.84%	51.79%
	9	gbdt	50	3	0.15	50	20	82.16%	75.56%	52.12%
	10	gbdt	50	3	0.2	50	20	82.36%	75.67%	48.60%
	11	gbdt	50	2	0.15	50	20	75.83%	73.76%	52.07%
	12	gbdt	50	1	0.1	50	20	65.99%	65.66%	43.02%
	13	gbdt	80	2	0.15	30	20	73.46%	70.70%	52.12%
	14	dart	80	2	0.15	30	20	72.44%	69.09%	50.17%
	15	goss	80	2	0.15	30	20	73.86%	70.78%	52.51%
	16	goss	70	1	0.1	30	20	64.56%	65.16%	35.92%
	17	goss	70	2	0.3	30	20	56.74%	52.99%	38.83%
	18	goss	70	2	0.05	30	20	72.21%	71.30%	53.74%
	19	goss	70	2	0.05	50	20	72.91%	72.81%	53.46%
	20	goss	50	2	0.05	50	20	72.77%	71.85%	53.41%
	21	goss	50	2	0.05	50	30	73.14%	70.80%	52.79%
	22	goss	50	2	0.05	50	10	73.01%	72.28%	53.74%
	23	goss	50	2	0.05	50	8	72.93%	72.09%	53.69%
	24	goss	50	2	0.05	30	8	73.03%	70.54%	53.30%
	25	gbdt	50	2	0.05	30	8	71.71%	70.67%	50.50%
	26	goss	50	2	0.03	30	8	71.67%	71.17%	50.89%
	27	goss	60	2	0.03	30	8	71.11%	69.26%	50.84%
	28	goss	55	2	0.03	30	8	71.27%	71.01%	51.17%
	29	goss	55	2	0.02	30	8	70.03%	69.51%	49.44%
	30	goss	75	2	0.02	30	8	70.13%	69.71%	50.95%
	31	goss	75	2	0.02	30	20	70.34%	69.21%	52.57%
	32	goss	75	2	0.02	30	22	69.71%	69.31%	51.68%
	33	goss	80	2	0.02	30	22	70.19%	68.98%	50.78%
	34	goss	80	3	0.02	30	22	73.57%	73.12%	52.51%
	35	goss	80	4	0.02	30	22	76.01%	74.65%	54.25%
	36	goss	80	5	0.02	30	22	80.27%	76.13%	50.73%
	37	goss	80	4	0.02	20	22	76.63%	74.05%	52.96%
	38	goss	80	4	0.02	40	22	77.92%	74.97%	53.13%
	39	goss	80	4	0.02	35	22	77.10%	75.74%	53.63%
	40	goss	50	4	0.02	30	22	77.21%	72.77%	53.46%
	41	goss	90	4	0.02	30	22	76.17%	76.12%	54.41%
	42	goss	100	4	0.02	30	22	75.87%	75.27%	53.85%
	43	goss	90	4	0.02	30	10	76.74%	75.68%	53.24%
	44	goss	90	0	0.02	30	30	90.74%	80.48%	51.56%
	45	goss	200	4	0.02	30	22	76.60%	75.05%	53.35%
	46	goss	200	4	0.001	30	22	71.02%	70.09%	52.79%

	Iteration	n_estomators	criterion	max_depth	min_samples_leaf	min_samples_split	max_feature	Train	Test	OOT
Random Forest	1 (default)	100	gini	None	1	2	auto	100.00%	79.74%	54.02%
	2	100	entropy	None	1	2	auto	100.00%	80.73%	52.01%
	3	200	entropy	None	1	2	auto	100.00%	79.57%	51.12%
	4	100	gini	5	20	5	auto	73.49%	71.48%	53.80%
	5	100	gini	5	30	5	auto	72.53%	71.43%	54.97%
	6	100	gini	5	40	5	auto	72.14%	71.06%	55.53%
	7	100	gini	5	50	5	auto	72.31%	70.60%	55.87%
	8	100	gini	5	80	5	auto	71.25%	69.82%	55.70%
	9	150	gini	5	70	5	auto	71.35%	69.88%	55.92%
	10	200	gini	5	70	5	auto	72.02%	69.59%	55.75%
	11	150	gini	6	70	10	auto	74.23%	72.18%	55.92%
	12	150	gini	7	70	10	auto	75.09%	73.92%	56.09%
	13	150	gini	10	70	10	auto	79.20%	74.21%	57.26%
	14	150	gini	12	70	10	auto	80.98%	75.73%	57.21%
	15	150	entropy	12	70	10	auto	81.88%	77.73%	54.08%
	16	150	gini	14	70	10	auto	81.15%	77.17%	57.54%
	17	150	gini	15	70	15	auto	81.26%	76.69%	57.26%
	18	150	gini	16	70	10	auto	80.74%	77.33%	57.26%
	19	200	gini	15	70	15	auto	81.44%	77.17%	57.49%
	20	200	gini	14	70	10	auto	81.10%	76.57%	57.21%
	21	180	gini	14	70	10	auto	81.10%	77.57%	57.26%
	22	**120**	**gini**	**14**	**70**	**10**	**auto**	**80.77%**	**76.14%**	**57.77%**
	23	130	gini	14	70	10	auto	81.08%	75.19%	57.60%
	24	80	gini	30	100	50	auto	78.26%	75.61%	56.20%
	25	80	gini	14	70	10	auto	81.38%	75.86%	57.54%
	26	120	gini	14	100	10	auto	78.99%	73.68%	57.21%
	27	120	gini	14	70	10	5	80.88%	76.75%	56.93%
	28	120	gini	14	70	10	log2	81.00%	76.84%	57.21%
	29	120	gini	14	70	50	auto	81.04%	76.49%	57.65%

	Iteration	activation	alpha	learning_rate	learning_rate_init	power_t	max_itera	Train	Test	OOT
Neurak Network	1 (default)	relu	0.0001	constant	0.001	0.5	200	79.11%	74.28%	47.37%
	2	relu	0.001	constant	0.001	0.5	200	78.46%	74.45%	52.35%
	3	relu	0.01	constant	0.001	0.5	200	75.13%	71.15%	54.97%
	4	relu	0.1	constant	0.001	0.5	200	68.75%	68.28%	47.99%
	5	relu	0.01	constant	0.01	0.5	200	73.41%	71.16%	51.79%
	6	relu	0.01	constant	0.001	1	200	73.92%	71.37%	53.30%
	7	relu	0.01	constant	0.001	0.4	200	74.44%	72.81%	53.13%
	8	relu	0.01	constant	0.001	0.5	300	73.94%	72.27%	52.68%
	9	relu	0.01	constant	0.001	0.5	150	74.46%	72.41%	54.47%
	10	relu	0.01	constant	0.001	0.5	220	73.60%	71.85%	53.46%
	11	identity	0.01	constant	0.001	0.5	200	60.21%	59.61%	29.27%
	12	logistic	0.01	constant	0.001	0.5	200	67.96%	66.47%	51.12%
	13	tanh	0.01	constant	0.001	0.5	200	72.90%	73.20%	56.15%
	14	tanh	0.01	invscaling	0.001	0.5	200	73.86%	71.40%	55.53%
	15	tanh	0.01	adaptive	0.001	0.5	200	73.69%	70.59%	55.70%
	16	tanh	0.01	constant	0.001	0.4	200	73.25%	71.88%	55.14%
	17	tanh	0.01	constant	0.001	0.5	50	73.05%	72.04%	55.81%
	18	tanh	0.01	constant	0.001	0.5	300	73.30%	72.35%	55.81%
	19	tanh	0.01	constant	0.001	0.5	500	73.49%	72.25%	56.09%
	20	tanh	0.01	constant	0.001	0.5	1000	73.20%	71.33%	56.37%
	21	tanh	0.01	constant	0.001	0.5	1200	73.81%	71.63%	56.26%
	22	tanh	0.01	constant	0.001	1	1000	73.51%	71.13%	55.14%
	23	tanh	0.01	constant	0.001	0.3	1000	73.58%	72.32%	55.31%
	24	tanh	0.01	constant	0.001	0.6	1000	73.40%	71.73%	55.31%

Results

Based on the model performance listed in the last section, the final model for this project was selected which has a high out-of-time fraud detection rate and relatively close fraud detection rates for the training and testing sets to avoid overfitting. Therefore, a random forest algorithm with the following parameters was chosen for the final implementation of this project.

<u>Final Algorithm</u>
Model: **Random Forest**
Number of Variables: **20**
N_estimators: **12**
Criterion: **Gini**
Max_depth: **14**
Min_samples_split: **10**
Min_samples_leaf: **70**
Max_feature: **20**

<u>Final Model Results</u>
Average FDR at 3%: **80.7% (Train), 76.1% (Test), 58.1% (OOT)**
The following tables show the model performance of the final model for the training, testing, and OOT sets across different population bins, respectively.

Table 9. Final Model Performance for Training Set

Training	# Records		# Goods		# Bads		Fraud Rate					
	59010		58377		633		1.07%					
	Bin Statistics					Cumulative Statistics						
Population bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR	KS	FPR
1	590	258	332	43.73	56.27	590	258	332	0.44	52.45	52.01	0.78
2	590	460	130	77.97	22.03	1180	718	462	1.23	72.99	71.76	1.55
3	590	541	49	91.69	8.31	1770	1259	511	2.16	80.73	78.57	2.46
4	590	566	24	95.93	4.07	2360	1825	535	3.13	84.52	81.39	3.41
5	590	573	17	97.12	2.88	2950	2398	552	4.11	87.20	83.10	4.34
6	591	574	17	97.12	2.88	3541	2972	569	5.09	89.89	84.80	5.22
7	590	579	11	98.14	1.86	4131	3551	580	6.08	91.63	85.54	6.12
8	590	582	8	98.64	1.36	4721	4133	588	7.08	92.89	85.81	7.03
9	590	584	6	98.98	1.02	5311	4717	594	8.08	93.84	85.76	7.94
10	590	585	5	99.15	0.85	5901	5302	599	9.08	94.63	85.55	8.85
11	590	583	7	98.81	1.19	6491	5885	606	10.08	95.73	85.65	9.71
12	590	587	3	99.49	0.51	7081	6472	609	11.09	96.21	85.12	10.63
13	590	586	4	99.32	0.68	7671	7058	613	12.09	96.84	84.75	11.51
14	590	582	8	98.64	1.36	8261	7640	621	13.09	98.10	85.02	12.30
15	591	589	2	99.66	0.34	8852	8229	623	14.10	98.42	84.32	13.21
16	590	586	4	99.32	0.68	9442	8815	627	15.10	99.05	83.95	14.06
17	590	587	3	99.49	0.51	10032	9402	630	16.11	99.53	83.42	14.92
18	590	588	2	99.66	0.34	10622	9990	632	17.11	99.84	82.73	15.81
19	590	590	0	100.00	0.00	11212	10580	632	18.12	99.84	81.72	16.74
20	590	589	1	99.83	0.17	11802	11169	633	19.13	100.00	80.87	17.64

Table 10. Final Model Performance for Testing Set

Testing	# Records		# Goods		# Bads		Fraud Rate					
	25290		25043		247		0.98%					
	Bin Statistics						Cumulative Statistics					
Population bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR	KS	FPR
1	253	120	133	47.43	52.57	253	120	133	0.48	53.85	53.37	0.90
2	253	216	37	85.38	14.62	506	336	170	1.34	68.83	67.48	1.98
3	253	235	18	92.89	7.11	759	571	188	2.28	76.11	73.83	3.04
4	253	247	6	97.63	2.37	1012	818	194	3.27	78.54	75.28	4.22
5	252	246	6	97.62	2.38	1264	1064	200	4.25	80.97	76.72	5.32
6	253	251	2	99.21	0.79	1517	1315	202	5.25	81.78	76.53	6.51
7	253	251	2	99.21	0.79	1770	1566	204	6.25	82.59	76.34	7.68
8	253	250	3	98.81	1.19	2023	1816	207	7.25	83.81	76.55	8.77
9	253	252	1	99.60	0.40	2276	2068	208	8.26	84.21	75.95	9.94
10	253	251	2	99.21	0.79	2529	2319	210	9.26	85.02	75.76	11.04
11	253	253	0	100.00	0.00	2782	2572	210	10.27	85.02	74.75	12.25
12	253	252	1	99.60	0.40	3035	2824	211	11.28	85.43	74.15	13.38
13	253	252	1	99.60	0.40	3288	3076	212	12.28	85.83	73.55	14.51
14	253	253	0	100.00	0.00	3541	3329	212	13.29	85.83	72.54	15.70
15	253	252	1	99.60	0.40	3794	3581	213	14.30	86.23	71.94	16.81
16	252	251	1	99.60	0.40	4046	3832	214	15.30	86.64	71.34	17.91
17	253	253	0	100.00	0.00	4299	4085	214	16.31	86.64	70.33	19.09
18	253	253	0	100.00	0.00	4552	4338	214	17.32	86.64	69.32	20.27
19	253	251	2	99.21	0.79	4805	4589	216	18.32	87.45	69.12	21.25
20	253	253	0	100.00	0.00	5058	4842	216	19.33	87.45	68.11	22.42

Table 11. Final Model Performance for Out-of-Time Set

OOT	# Records		# Goods		# Bads		Fraud Rate					
	12097		11918		179		1.48%					
	Bin Statistics						Cumulative Statistics					
Population bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Cumulative Goods	FDR	KS	FPR
1	121	74	47	61.16	38.84	121	74	47	0.62	26.26	25.64	1.57
2	121	87	34	71.90	28.10	242	161	81	1.35	45.25	43.90	1.99
3	121	98	23	80.99	19.01	363	259	104	2.17	58.10	55.93	2.49
4	121	120	1	99.17	0.83	484	379	105	3.18	58.66	55.48	3.61
5	121	120	1	99.17	0.83	605	499	106	4.19	59.22	55.03	4.71
6	121	115	6	95.04	4.96	726	614	112	5.15	62.57	57.42	5.48
7	121	119	2	98.35	1.65	847	733	114	6.15	63.69	57.54	6.43
8	121	118	3	97.52	2.48	968	851	117	7.14	65.36	58.22	7.27
9	121	118	3	97.52	2.48	1089	969	120	8.13	67.04	58.91	8.08
10	121	120	1	99.17	0.83	1210	1089	121	9.14	67.60	58.46	9.00
11	121	119	2	98.35	1.65	1331	1208	123	10.14	68.72	58.58	9.82
12	121	121	0	100.00	0.00	1452	1329	123	11.15	68.72	57.56	10.80
13	121	120	1	99.17	0.83	1573	1449	124	12.16	69.27	57.12	11.69
14	121	121	0	100.00	0.00	1694	1570	124	13.17	69.27	56.10	12.66
15	121	120	1	99.17	0.83	1815	1690	125	14.18	69.83	55.65	13.52
16	121	119	2	98.35	1.65	1936	1809	127	15.18	70.95	55.77	14.24
17	120	116	4	96.67	3.33	2056	1925	131	16.15	73.18	57.03	14.69
18	121	119	2	98.35	1.65	2177	2044	133	17.15	74.30	57.15	15.37
19	121	119	2	98.35	1.65	2298	2163	135	18.15	75.42	57.27	16.02
20	121	119	2	98.35	1.65	2419	2282	137	19.15	76.54	57.39	16.66

Time Dependency of Fraud Scores

With the selected model, our team went back to the original dataset and examined the time dependency of the fraud score. To help visualize the concept, two 10-day fraud score graphs were plotted for a selected card number (5142212038) and merchant number (965610600330) shown below.

For the fraud score over time graph of the chosen card number (Figure 13), there is a spike between July 18th and July 21st because 39 transactions happened within the 2 days, which was about 72% of its total transactions throughout the year. Additionally, by looking at the cumulative transaction count vs. fraud score plot on the bottom, the fraud score didn't increase to the peak value immediately when the first few transactions occurred on July 18th. This is because a fraud score takes time to recognize a flow of unusual transactions. In this case, the first few transactions looked normal to the model and then started to look more suspicious as the number of transactions kept increasing. As a result, the fraud score rose rapidly to its peak at the end.

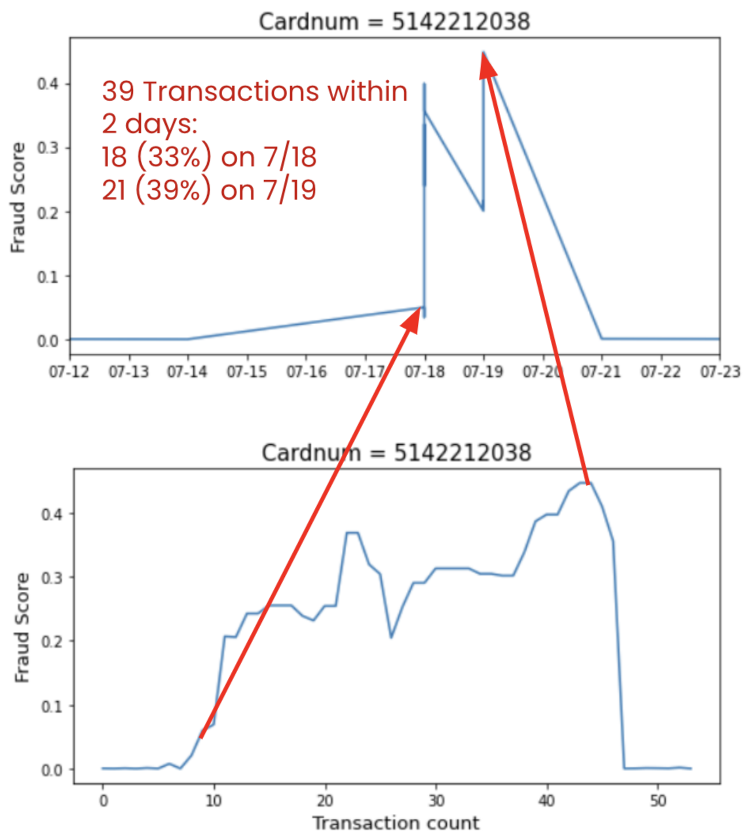


Figure 13. Fraud Score Change for Card Number=5142212038

Similarly, for the fraud score over time graph of the selected merchant number in Figure 14, about 50% of its annual transactions took place within 5 days (July 17th - July 21st), which explained why there were 3 peaks in the fraud score on July 17th, 19th, and 21st. By looking at its fraud score by transaction count, it also matched the previous observation that the fraud score takes time to increase with activity. In addition, it is clearly shown that the 3 jumps on the bottom matched the 3 rapid increases in the transactions on the 3 days identified on the top.

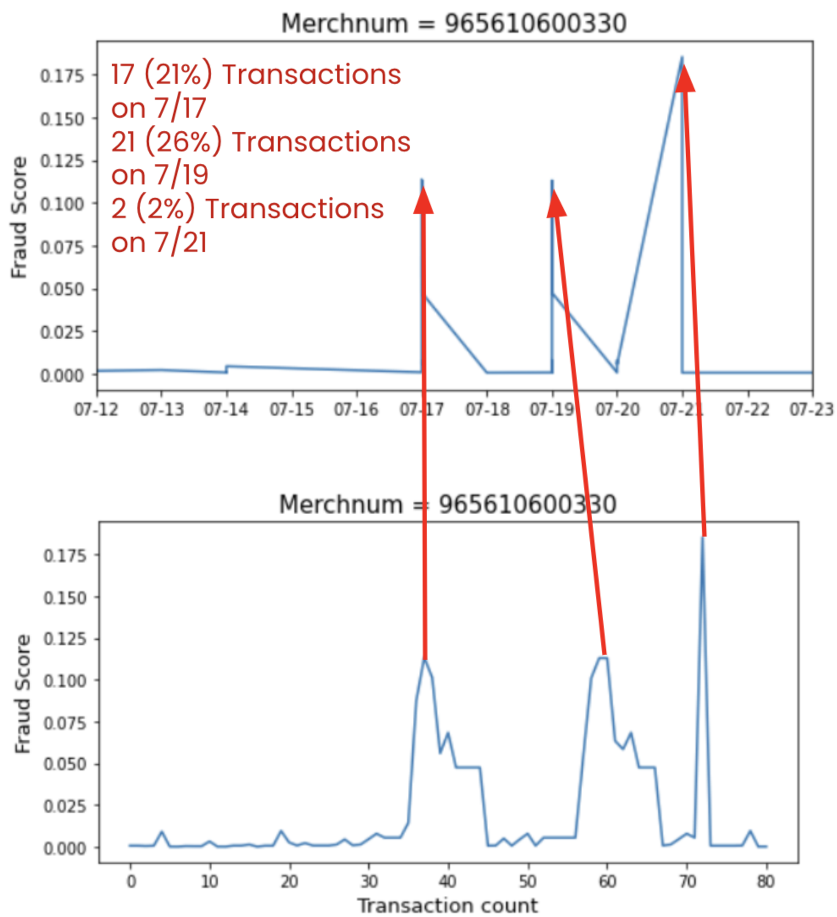


Figure 14. Fraud Score Change for Merchant Number = 965610600330

Model Financials by Cutoff

To analyze the financial performance of our model, a plot of the estimated savings, lost sales, and overall savings across different bins of the out-of-time population is shown in Figure 15 below.

During the analysis, a $2,000 gain was assumed for each fraud that our model could catch. A $50 loss was assumed for each false positive transaction that our model would decline. The overall savings were calculated by subtracting the lost sales from fraud savings at each population bin. With the comparison, we recommend a cutoff point at 3% for this model because it would maximize the overall savings and also push the cutoff as left (small) as possible to prevent too many transactions from being declined.

With this analysis, our team believes that we could have an annualized total overall saving of $1,170,000 by implementing our model and setting the cutoff line at 3%.

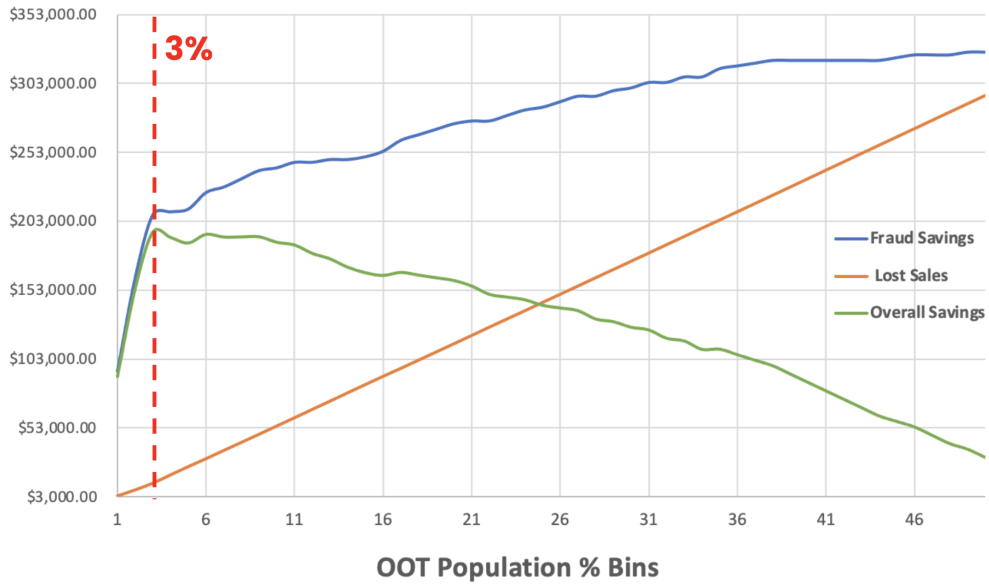


Figure 15. Financial Analysis of Final Model

Conclusions

To examine and analyze the best model to prevent credit card transaction fraud with supervised machine learning algorithms and analytics tools, the team built a strong model to find the best algorithms to use for model predictions. The team identified data type and distribution as the initial step. Next, the team cleaned the data to filter extreme values and fill in missing values in “Merchnum”, “Merch State”, and “Merch Zip”. Furthermore, based on our assumption of fraud signals, the team created as many entities and variables as possible, such as transaction velocity, day since, and relative velocity. The team also included variables from target encoding and Benford’s Law in the variable creation process. The team then used a filter and wrapper to select the best 20 variables for the feature selection process. Using the best 20 variables, 5 different algorithms, including Logistic Regression, Decision Tree, Random Forest, Boosted Tree, and Neural Network, were deployed to find the best model. Finally, the team selected Random Forest as the optimal algorithm. We have concluded that the best selected Random Forest model can eliminate 58.10% of the fraud by declining 3% of the transactions.

If time allows, the team can ask domain experts for advice when creating new variables for future improvement. Since the class distribution is imbalanced in the original dataset, the team may also use techniques such as downsampling or upsampling to balance the data better. Lastly, the team can implement more models for model algorithms such as K-nearest neighbors (KNN) and support vector machines to potentially increase the fraud detection rate.

Appendix

A.Data Quality Report - Application Data

I. Data Description

The card transactions dataset contains transactional information from a U.S. government organization that issues charge cards to its employees for government business purposes. The dataset has ten fields, including fraud labels, and about 100,000 records. The fraud records are synthetic for research and educational purposes. The dataset is provided by Dr. Coggeshall for DSO562.

II. Field Summary

Numerical Field Summary Table

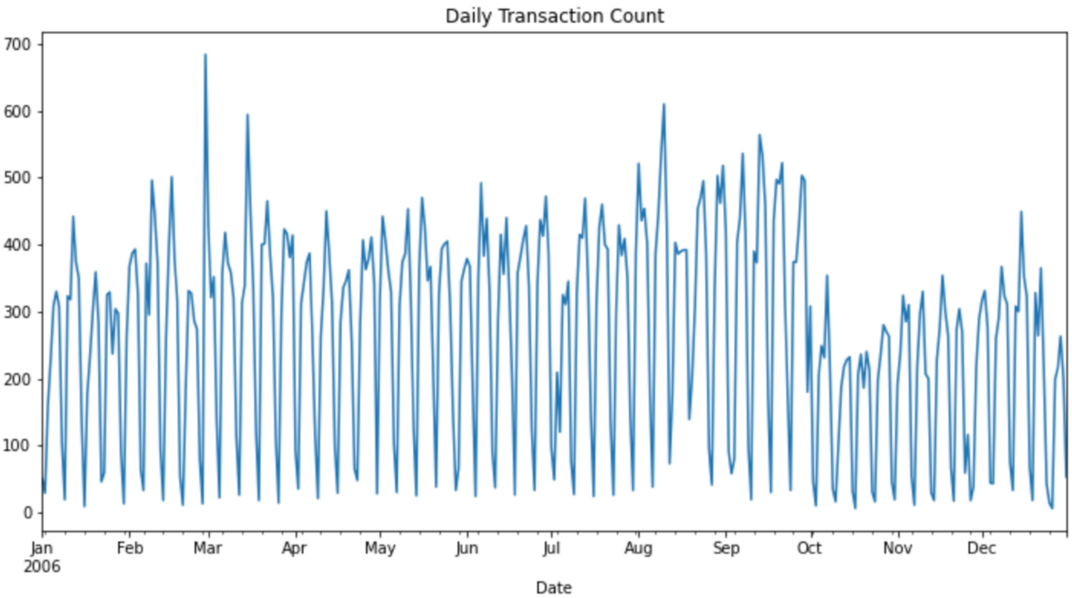
Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
Date	100	2006-01-01	2006-12-31	N/A	N/A	0
Cardnum	100	0.01	3,102,045.53	427.89	10,006.14	0

Categorical Field Summary Table

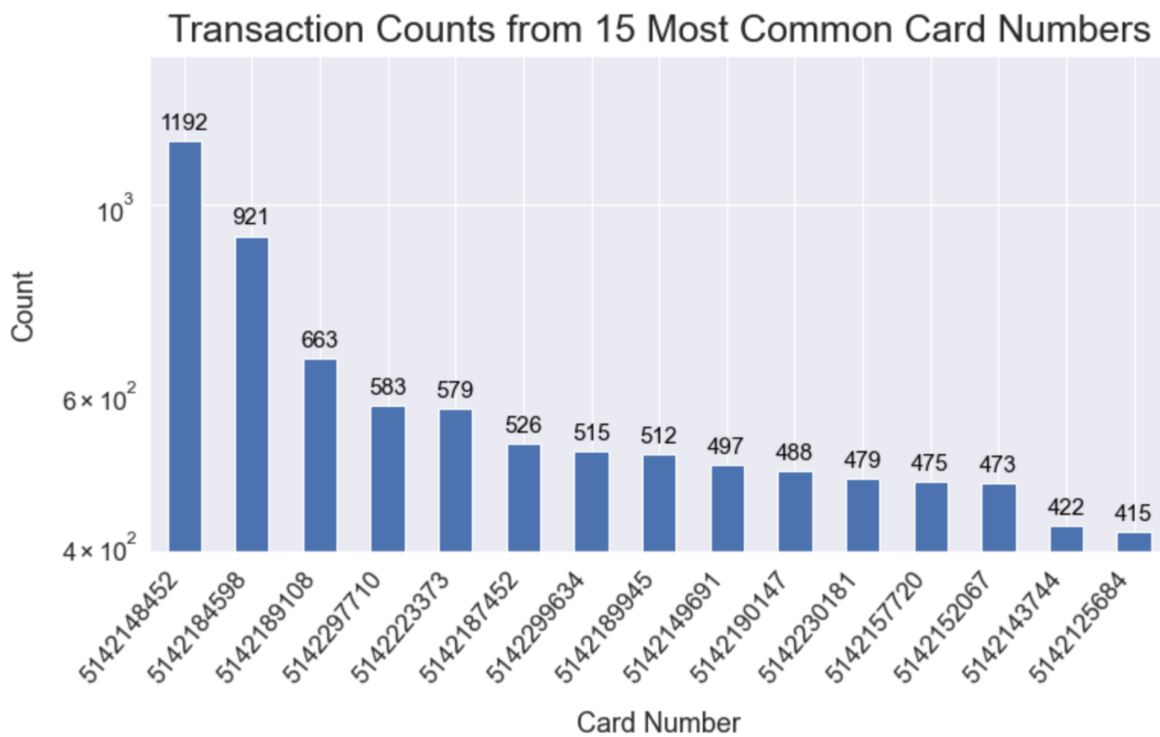
Field Name	% Populated	# Unique Values	Most Common Value
Recnum	100	96,753	N/A
Cardnum	100	1,645	5142148452
Merchnum	96.5	13,092	930090121224
Merch description	100	13,126	GSA-FSS-ADV
Merch state	98.8	228	TN
Merch zip	95.2	4,568	38118
Transtype	100	4	P
fraud	100	2	0

III. Field Description

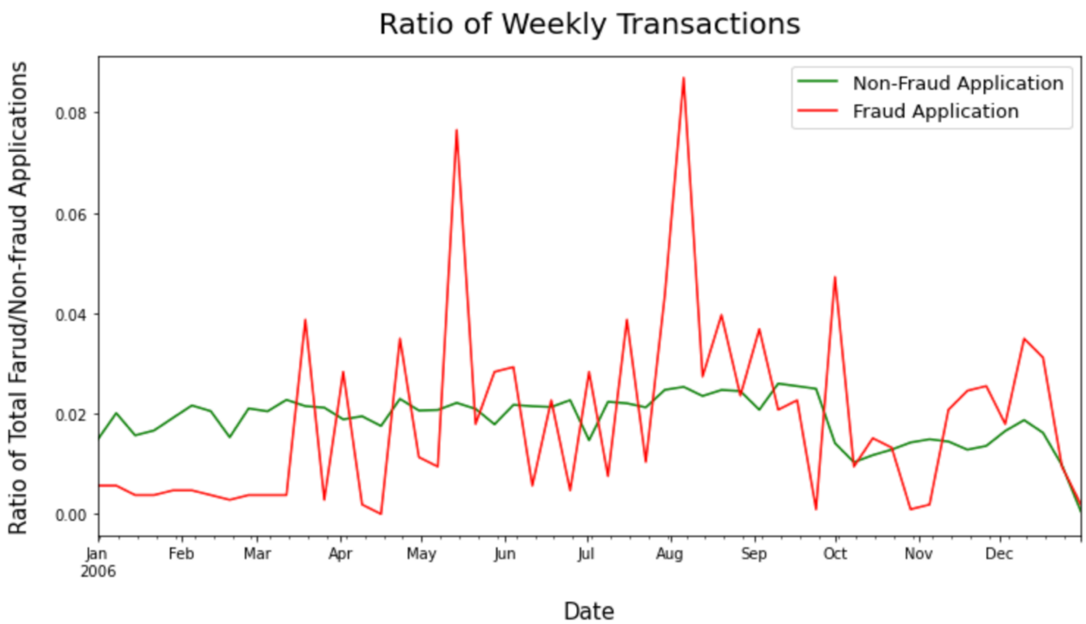
1. **Recnum**: Unique record number of each transaction; the plot below indicates the distribution of total daily transactions



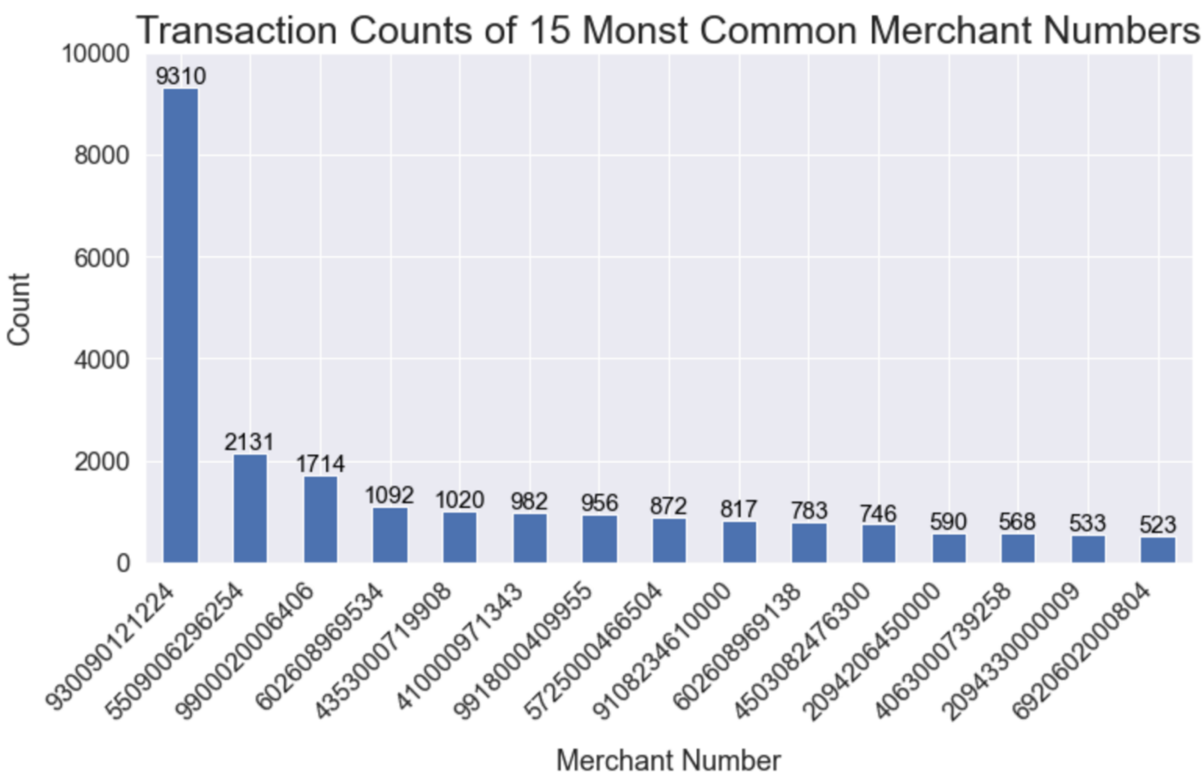
2. **Cardnum**: Card number of each transaction; the plot below covers the total transaction counts from the 15 most common card numbers in dataset



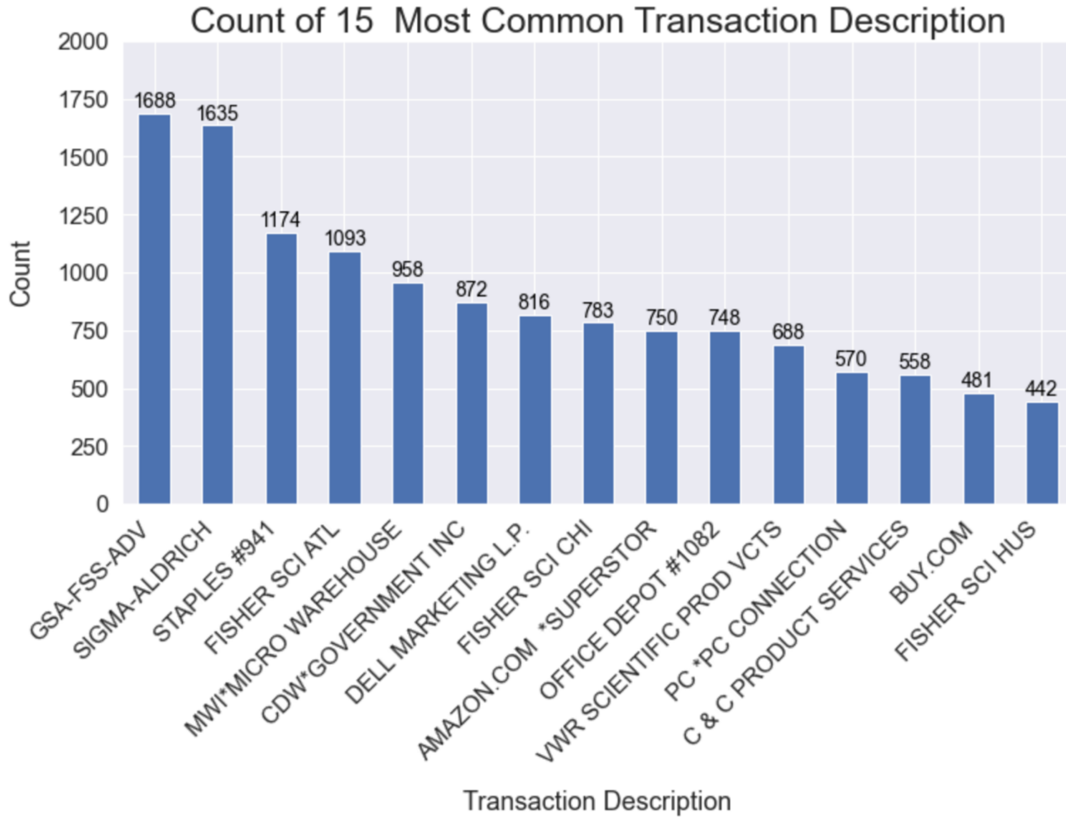
3. **Date**: Transaction date of each record; the plot below reflects the ratio of weekly transactions by their fraud labels



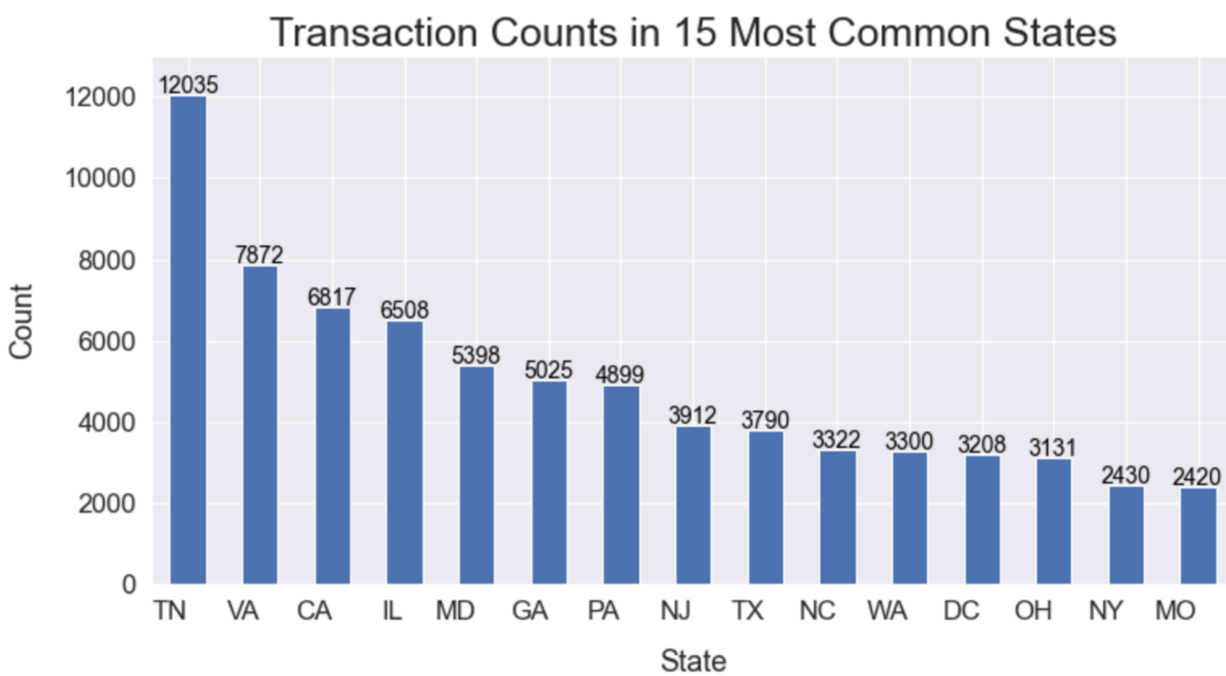
4. **Merchnum**: Merchant number of the merchant in each transaction; the plot below covers the total transaction counts associated with the 15 most common merchant numbers



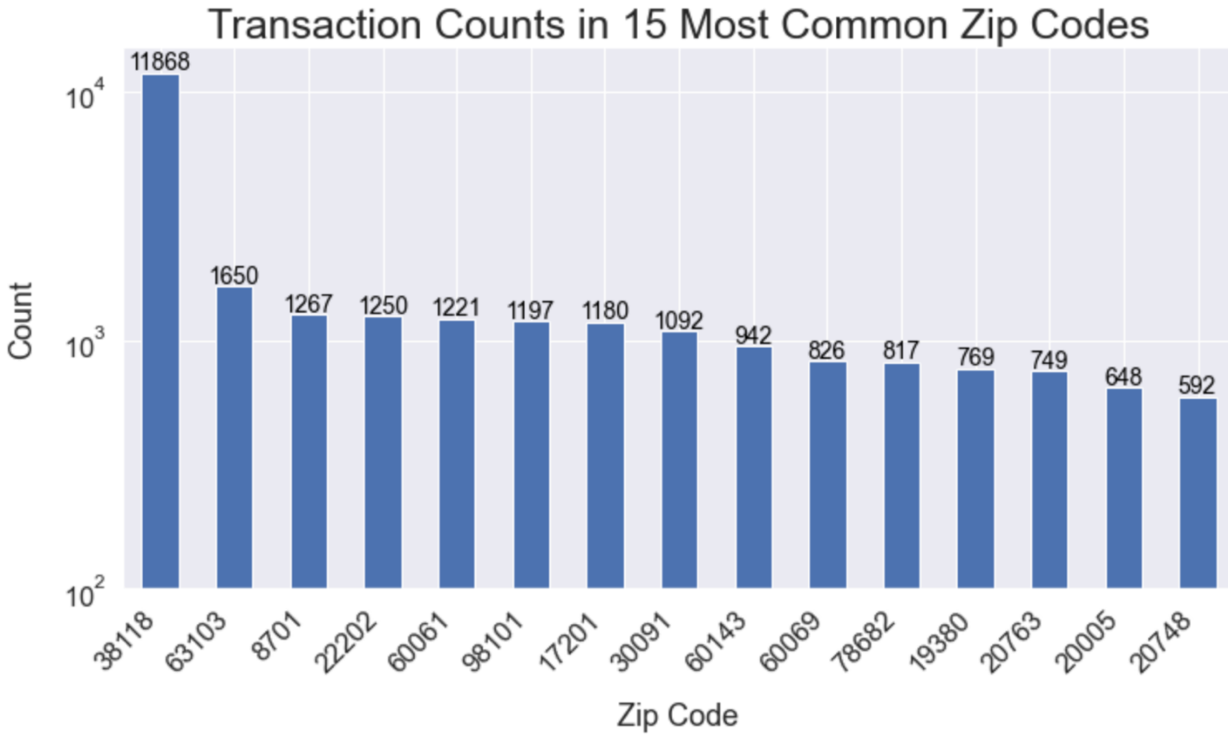
5. **Merch description**: Description of each transaction; the plot below covers the total transaction counts for the top 15 most common transaction descriptions



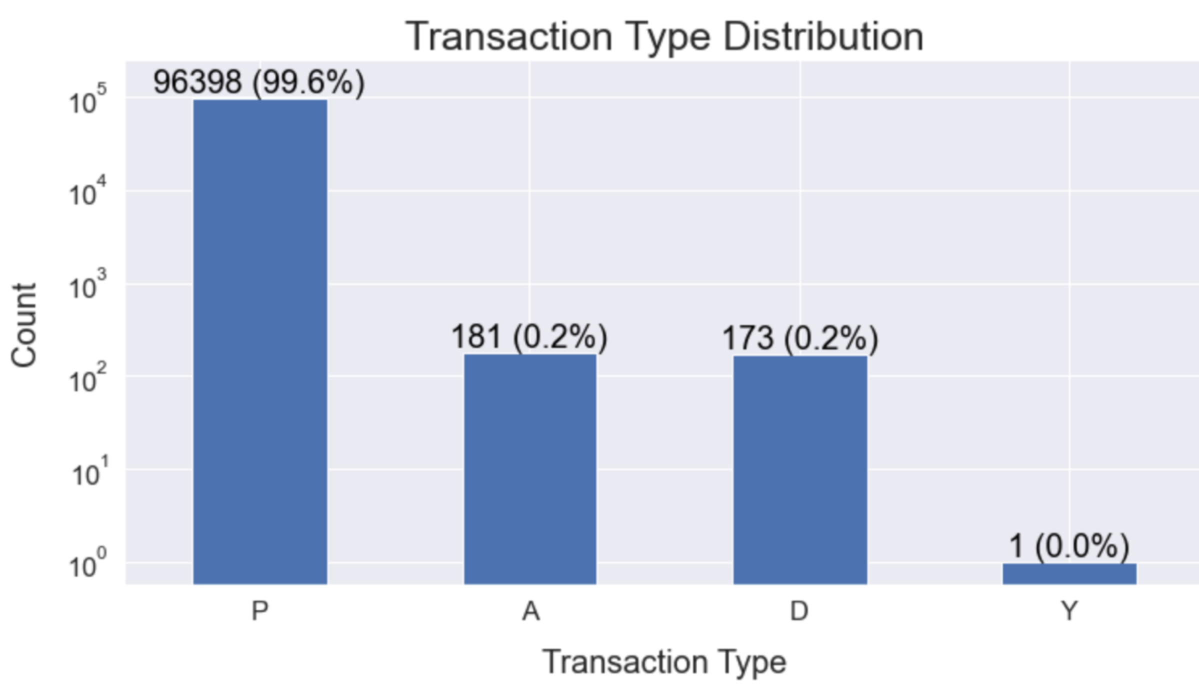
6. **Merch state**: State where each transaction occurred; the plot below covers the total number of transactions in the 15 most common states



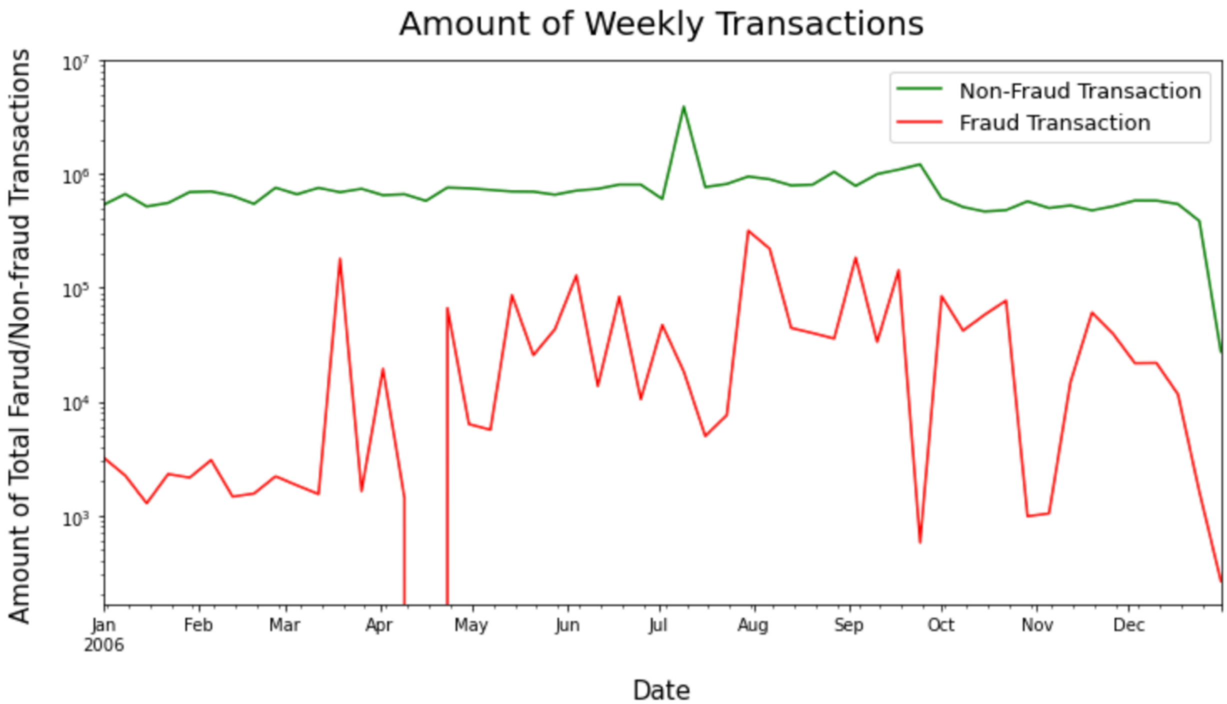
7. **Merch zip**: Zipcode of each transaction; the plot below covers the total transaction counts associated with the 15 most common zip codes



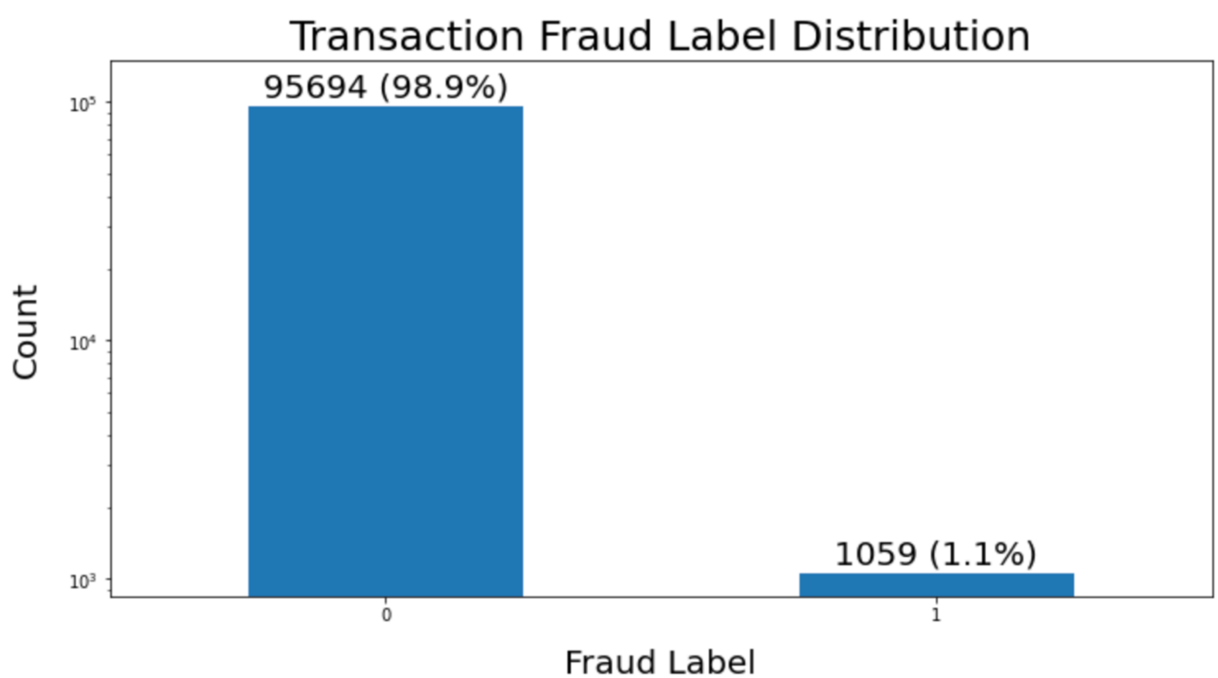
8. **Transtype**: Type of each transaction; the plot illustrates the distribution of all the transaction types in dataset



9. **Amount**: The dollar amount of each transaction; the plot presents the weekly total transaction amount by fraud type



10. **Fraud**: Fraud label of each transaction record with 1 = fraud transaction and 0 = regular transaction; the plot below presents the distribution of transaction fraud from the entire dataset



B. Candidate Variables

```
['Recnum',
 'Cardnum',
 'Date',
 'Merchnum',
 'Merch description',
 'Merch state',
 'Merch zip',
 'Transtype',
 'Amount',
 'Fraud',
 'card_merch',
 'card_state',
 'card_zip',
 'card_transtype',
 'Merch_address',
 'Merchnum_state',
 'Merchnum_zip',
 'Date_state',
 'Date_zip',
 'Amount_date',
 'Amount_description',
 'Amount_state',
 'Amount_zip',
 'Amount_transtype',
 'Cardnum_day_since',
 'Cardnum_count_0',
 'Cardnum_avg_0',
 'Cardnum_max_0',
 'Cardnum_med_0',
 'Cardnum_total_0',
 'Cardnum_actual/avg_0',
 'Cardnum_actual/max_0',
 'Cardnum_actual/med_0',
 'Cardnum_actual/toal_0',
 'Cardnum_count_1',
 'Cardnum_avg_1',
 'Cardnum_max_1',
 'Cardnum_med_1',
 'Cardnum_total_1',
 'Cardnum_actual/avg_1',
 'Cardnum_actual/max_1',
 'Cardnum_actual/med_1',
 'Cardnum_actual/toal_1',
 'Cardnum_count_3',
 'Cardnum_avg_3',
 'Cardnum_max_3',
 'Cardnum_med_3',
 'Cardnum_total_3',
 'Cardnum_actual/avg_3',
 'Cardnum_actual/max_3',
 'Cardnum_actual/med_3',
 'Cardnum_actual/toal_3',
 'Cardnum_count_7',
 'Cardnum_avg_7',
 'Cardnum_max_7',
 'Cardnum_med_7',
 'Cardnum_total_7',
 'Cardnum_actual/avg_7',
 'Cardnum_actual/max_7',
 'Cardnum_actual/med_7',
 'Cardnum_actual/toal_7',
 'Cardnum_count_14',
 'Cardnum_avg_14',
 'Cardnum_max_14',
 'Cardnum_med_14',
 'Cardnum_total_14',
 'Cardnum_actual/avg_14',
 'Cardnum_actual/max_14',
 'Cardnum_actual/med_14',
 'Cardnum_actual/toal_14',
 'Cardnum_count_30',
 'Cardnum_avg_30',
 'Cardnum_max_30',
 'Cardnum_med_30',
 'Cardnum_total_30',
 'Cardnum_actual/avg_30',
 'Cardnum_actual/max_30',
 'Cardnum_actual/med_30',
 'Cardnum_actual/toal_30',
 'Merchnum_day_since',
 'Merchnum_count_0',
 'Merchnum_avg_0',
 'Merchnum_max_0',
 'Merchnum_med_0',
 'Merchnum_total_0',
 'Merchnum_actual/avg_0',
 'Merchnum_actual/max_0',
 'Merchnum_actual/med_0',
```

```
 'Merch zip_avg_3',
 'Merch zip_max_3',
 'Merch zip_med_3',
 'Merch zip_total_3',
 'Merch zip_actual/avg_3',
 'Merch zip_actual/max_3',
 'Merch zip_actual/med_3',
 'Merch zip_actual/toal_3',
 'Merch zip_count_7',
 'Merch zip_avg_7',
 'Merch zip_max_7',
 'Merch zip_med_7',
 'Merch zip_total_7',
 'Merch zip_actual/avg_7',
 'Merch zip_actual/max_7',
 'Merch zip_actual/med_7',
 'Merch zip_actual/toal_7',
 'Merch zip_count_14',
 'Merch zip_avg_14',
 'Merch zip_max_14',
 'Merch zip_med_14',
 'Merch zip_total_14',
 'Merch zip_actual/avg_14',
 'Merch zip_actual/max_14',
 'Merch zip_actual/med_14',
 'Merch zip_actual/toal_14',
 'Merch zip_count_30',
 'Merch zip_avg_30',
 'Merch zip_max_30',
 'Merch zip_med_30',
 'Merch zip_total_30',
 'Merch zip_actual/avg_30',
 'Merch zip_actual/max_30',
 'Merch zip_actual/med_30',
 'Merch zip_actual/toal_30',
 'card_merch_day_since',
 'card_merch_count_0',
 'card_merch_avg_0',
 'card_merch_max_0',
 'card_merch_med_0',
 'card_merch_total_0',
 'card_merch_actual/avg_0',
 'card_merch_actual/max_0',
 'card_merch_actual/med_0',
 'card_merch_actual/toal_0',
 'card_merch_count_1',
 'card_merch_avg_1',
 'card_merch_max_1',
 'card_merch_med_1',
 'card_merch_total_1',
 'card_merch_actual/avg_1',
 'card_merch_actual/max_1',
 'card_merch_actual/med_1',
 'card_merch_actual/toal_1',
 'card_merch_count_3',
 'card_merch_avg_3',
 'card_merch_max_3',
 'card_merch_med_3',
 'card_merch_total_3',
 'card_merch_actual/avg_3',
 'card_merch_actual/max_3',
 'card_merch_actual/med_3',
 'card_merch_actual/toal_3',
 'card_merch_count_7',
 'card_merch_avg_7',
 'card_merch_max_7',
 'card_merch_med_7',
 'card_merch_total_7',
 'card_merch_actual/avg_7',
 'card_merch_actual/max_7',
 'card_merch_actual/med_7',
 'card_merch_actual/toal_7',
 'card_merch_count_14',
 'card_merch_avg_14',
 'card_merch_max_14',
 'card_merch_med_14',
 'card_merch_total_14',
 'card_merch_actual/avg_14',
 'card_merch_actual/max_14',
 'card_merch_actual/med_14',
 'card_merch_actual/toal_14',
 'card_merch_count_30',
 'card_merch_avg_30',
 'card_merch_max_30',
 'card_merch_med_30',
 'card_merch_total_30',
 'card_merch_actual/avg_30',
 'card_merch_actual/max_30',
```

'Merch description_actual/avg_14',
'Merch description_actual/max_14',
'Merch description_actual/med_14',
'Merch description_actual/toal_14',
'Merch description_count_30',
'Merch description_avg_30',
'Merch description_max_30',
'Merch description_med_30',
'Merch description_total_30',
'Merch description_actual/avg_30',
'Merch description_actual/max_30',
'Merch description_actual/med_30',
'Merch description_actual/toal_30',
'Merch state_day_since',
'Merch state_count_0',
'Merch state_avg_0',
'Merch state_max_0',
'Merch state_med_0',
'Merch state_total_0',
'Merch state_actual/avg_0',
'Merch state_actual/max_0',
'Merch state_actual/med_0',
'Merch state_actual/toal_0',
'Merch state_count_1',
'Merch state_avg_1',
'Merch state_max_1',
'Merch state_med_1',
'Merch state_total_1',
'Merch state_actual/avg_1',
'Merch state_actual/max_1',
'Merch state_actual/med_1',
'Merch state_actual/toal_1',
'Merch state_count_3',
'Merch state_avg_3',
'Merch state_max_3',
'Merch state_med_3',
'Merch state_total_3',
'Merch state_actual/avg_3',
'Merch state_actual/max_3',
'Merch state_actual/med_3',
'Merch state_actual/toal_3',
'Merch state_count_7',
'Merch state_avg_7',
'Merch state_max_7',

'Merch state_med_7',
'Merch state_total_7',
'Merch state_actual/avg_7',
'Merch state_actual/max_7',
'Merch state_actual/med_7',
'Merch state_actual/toal_7',
'Merch state_count_14',
'Merch state_avg_14',
'Merch state_max_14',
'Merch state_med_14',
'Merch state_total_14',
'Merch state_actual/avg_14',
'Merch state_actual/max_14',
'Merch state_actual/med_14',
'Merch state_actual/toal_14',
'Merch state_count_30',
'Merch state_avg_30',
'Merch state_max_30',
'Merch state_med_30',
'Merch state_total_30',
'Merch state_actual/avg_30',
'Merch state_actual/max_30',
'Merch state_actual/med_30',
'Merch state_actual/toal_30',
'Merch zip_day_since',
'Merch zip_count_0',
'Merch zip_avg_0',
'Merch zip_max_0',
'Merch zip_med_0',
'Merch zip_total_0',
'Merch zip_actual/avg_0',
'Merch zip_actual/max_0',
'Merch zip_actual/med_0',
'Merch zip_actual/toal_0',
'Merch zip_count_1',
'Merch zip_avg_1',
'Merch zip_max_1',
'Merch zip_med_1',
'Merch zip_total_1',
'Merch zip_actual/avg_1',
'Merch zip_actual/max_1',
'Merch zip_actual/med_1',
'Merch zip_actual/toal_1',
'Merch zip_count_3',

'Merchnum_actual/toal_0',
'Merchnum_count_1',
'Merchnum_avg_1',
'Merchnum_max_1',
'Merchnum_med_1',
'Merchnum_total_1',
'Merchnum_actual/avg_1',
'Merchnum_actual/max_1',
'Merchnum_actual/med_1',
'Merchnum_actual/toal_1',
'Merchnum_count_3',
'Merchnum_avg_3',
'Merchnum_max_3',
'Merchnum_med_3',
'Merchnum_total_3',
'Merchnum_actual/avg_3',
'Merchnum_actual/max_3',
'Merchnum_actual/med_3',
'Merchnum_actual/toal_3',
'Merchnum_count_7',
'Merchnum_avg_7',
'Merchnum_max_7',
'Merchnum_med_7',
'Merchnum_total_7',
'Merchnum_actual/avg_7',
'Merchnum_actual/max_7',
'Merchnum_actual/med_7',
'Merchnum_actual/toal_7',
'Merchnum_count_14',
'Merchnum_avg_14',
'Merchnum_max_14',
'Merchnum_med_14',
'Merchnum_total_14',
'Merchnum_actual/avg_14',
'Merchnum_actual/max_14',
'Merchnum_actual/med_14',
'Merchnum_actual/toal_14',
'Merchnum_count_30',
'Merchnum_avg_30',
'Merchnum_max_30',
'Merchnum_med_30',
'Merchnum_total_30',
'Merchnum_actual/avg_30',
'Merchnum_actual/max_30',

'Merchnum_actual/med_30',
'Merchnum_actual/toal_30',
'Merch description_day_since',
'Merch description_count_0',
'Merch description_avg_0',
'Merch description_max_0',
'Merch description_med_0',
'Merch description_total_0',
'Merch description_actual/avg_0',
'Merch description_actual/max_0',
'Merch description_actual/med_0',
'Merch description_actual/toal_0',
'Merch description_count_1',
'Merch description_avg_1',
'Merch description_max_1',
'Merch description_med_1',
'Merch description_total_1',
'Merch description_actual/avg_1',
'Merch description_actual/max_1',
'Merch description_actual/med_1',
'Merch description_actual/toal_1',
'Merch description_count_3',
'Merch description_avg_3',
'Merch description_max_3',
'Merch description_med_3',
'Merch description_total_3',
'Merch description_actual/avg_3',
'Merch description_actual/max_3',
'Merch description_actual/med_3',
'Merch description_actual/toal_3',
'Merch description_count_7',
'Merch description_avg_7',
'Merch description_max_7',
'Merch description_med_7',
'Merch description_total_7',
'Merch description_actual/avg_7',
'Merch description_actual/max_7',
'Merch description_actual/med_7',
'Merch description_actual/toal_7',
'Merch description_count_14',
'Merch description_avg_14',
'Merch description_max_14',
'Merch description_med_14',
'Merch description_total_14',

'card_merch_actual/med_30',
'card_merch_actual/toal_30',
'card_state_day_since',
'card_state_count_0',
'card_state_avg_0',
'card_state_max_0',
'card_state_med_0',
'card_state_total_0',
'card_state_actual/avg_0',
'card_state_actual/max_0',
'card_state_actual/med_0',
'card_state_actual/toal_0',
'card_state_count_1',
'card_state_avg_1',
'card_state_max_1',
'card_state_med_1',
'card_state_total_1',
'card_state_actual/avg_1',
'card_state_actual/max_1',
'card_state_actual/med_1',
'card_state_actual/toal_1',
'card_state_count_3',
'card_state_avg_3',
'card_state_max_3',
'card_state_med_3',
'card_state_total_3',
'card_state_actual/avg_3',
'card_state_actual/max_3',
'card_state_actual/med_3',
'card_state_actual/toal_3',
'card_state_count_7',
'card_state_avg_7',
'card_state_max_7',
'card_state_med_7',
'card_state_total_7',
'card_state_actual/avg_7',
'card_state_actual/max_7',
'card_state_actual/med_7',
'card_state_actual/toal_7',
'card_state_count_14',
'card_state_avg_14',
'card_state_max_14',
'card_state_med_14',
'card_state_total_14',
'card_state_actual/avg_14',
'card_state_actual/max_14',
'card_state_actual/med_14',
'card_state_actual/toal_14',
'card_state_count_30',
'card_state_avg_30',
'card_state_max_30',
'card_state_med_30',
'card_state_total_30',
'card_state_actual/avg_30',
'card_state_actual/max_30',
'card_state_actual/med_30',
'card_state_actual/toal_30',
'card_transtype_day_since',
'card_transtype_count_0',
'card_transtype_avg_0',
'card_transtype_max_0',
'card_transtype_med_0',
'card_transtype_total_0',
'card_transtype_actual/avg_0',
'card_transtype_actual/max_0',
'card_transtype_actual/med_0',
'card_transtype_actual/toal_0',
'card_transtype_count_1',
'card_transtype_avg_1',
'card_transtype_max_1',
'card_transtype_med_1',
'card_transtype_total_1',
'card_transtype_actual/avg_1',
'card_transtype_actual/max_1',
'card_transtype_actual/med_1',
'card_transtype_actual/toal_1',
'card_transtype_count_3',
'card_transtype_avg_3',
'card_transtype_max_3',
'card_transtype_med_3',
'card_transtype_total_3',
'card_transtype_actual/avg_3',
'card_transtype_actual/max_3',
'card_transtype_actual/med_3',
'card_transtype_actual/toal_3',
'card_transtype_count_7',
'card_transtype_avg_7',
'card_transtype_max_7',
'card_transtype_med_7',
'card_transtype_total_7',
'card_transtype_actual/avg_7',
'card_transtype_actual/max_7',
'card_transtype_actual/med_7',
'card_transtype_actual/toal_7',
'card_transtype_count_14',
'card_transtype_avg_14',
'card_transtype_max_14',
'card_transtype_med_14',
'card_transtype_total_14',
'card_transtype_actual/avg_14',
'card_transtype_actual/max_14',
'card_transtype_actual/med_14',
'card_transtype_actual/toal_14',
'card_transtype_count_30',
'card_transtype_avg_30',
'card_transtype_max_30',
'card_transtype_med_30',
'card_transtype_total_30',
'card_transtype_actual/avg_30',
'card_transtype_actual/max_30',
'card_transtype_actual/med_30',
'card_transtype_actual/toal_30',
'Merch_address_day_since',
'Merch_address_count_0',
'Merch_address_avg_0',
'Merch_address_max_0',
'Merch_address_med_0',
'Merch_address_total_0',
'Merch_address_actual/avg_0',
'Merch_address_actual/max_0',
'Merch_address_actual/med_0',
'Merch_address_actual/toal_0',
'Merch_address_count_1',
'Merch_address_avg_1',
'Merch_address_max_1',
'Merch_address_med_1',
'Merch_address_total_1',
'Merch_address_actual/avg_1',
'Merch_address_actual/max_1',
'Merch_address_actual/med_1',
'Merch_address_actual/toal_1',
'Merch_address_count_3',
'Merch_address_avg_3',
'Merch_address_max_3',
'Merch_address_med_3',
'Merch_address_total_3',
'Merch_address_actual/avg_3',
'Merch_address_actual/max_3',
'Merch_address_actual/med_3',
'Merch_address_actual/toal_3',
'Merch_address_count_7',
'Merch_address_avg_7',
'Merch_address_max_7',
'Merch_address_med_7',
'Merch_address_total_7',
'Merch_address_actual/avg_7',
'Merch_address_actual/max_7',
'Merch_address_actual/med_7',
'Merch_address_actual/toal_7',
'Merch_address_count_14',
'Merch_address_avg_14',
'Merch_address_max_14',
'Merch_address_med_14',
'Merch_address_total_14',
'Merch_address_actual/avg_14',
'Merch_address_actual/max_14',
'Merch_address_actual/med_14',
'Merch_address_actual/toal_14',
'Merch_address_count_30',
'Merch_address_avg_30',
'Merch_address_max_30',
'Merch_address_med_30',
'Merch_address_total_30',
'Merch_address_actual/avg_30',
'Merch_address_actual/max_30',
'Merch_address_actual/med_30',
'Merch_address_actual/toal_30',
'Merchnum_state_day_since',
'Merchnum_state_count_0',
'Merchnum_state_avg_0',
'Merchnum_state_max_0',
'Merchnum_state_med_0',
'Merchnum_state_total_0',
'Merchnum_state_actual/avg_0',
'Merchnum_state_actual/max_0',
'Merchnum_state_actual/med_0',

'Merchnum_state_actual/toal_0',
'Merchnum_state_count_1',
'Merchnum_state_avg_1',
'Merchnum_state_max_1',
'Merchnum_state_med_1',
'Merchnum_state_total_1',
'Merchnum_state_actual/avg_1',
'Merchnum_state_actual/max_1',
'Merchnum_state_actual/med_1',
'Merchnum_state_actual/toal_1',
'Merchnum_state_count_3',
'Merchnum_state_avg_3',
'Merchnum_state_max_3',
'Merchnum_state_med_3',
'Merchnum_state_total_3',
'Merchnum_state_actual/avg_3',
'Merchnum_state_actual/max_3',
'Merchnum_state_actual/med_3',
'Merchnum_state_actual/toal_3',
'Merchnum_state_count_7',
'Merchnum_state_avg_7',
'Merchnum_state_max_7',
'Merchnum_state_med_7',
'Merchnum_state_total_7',
'Merchnum_state_actual/avg_7',
'Merchnum_state_actual/max_7',
'Merchnum_state_actual/med_7',
'Merchnum_state_actual/toal_7',
'Merchnum_state_count_14',
'Merchnum_state_avg_14',
'Merchnum_state_max_14',
'Merchnum_state_med_14',
'Merchnum_state_total_14',
'Merchnum_state_actual/avg_14',
'Merchnum_state_actual/max_14',
'Merchnum_state_actual/med_14',
'Merchnum_state_actual/toal_14',
'Merchnum_state_count_30',
'Merchnum_state_avg_30',
'Merchnum_state_max_30',
'Merchnum_state_med_30',
'Merchnum_state_total_30',
'Merchnum_state_actual/avg_30',
'Merchnum_state_actual/max_30',

'Merchnum_state_actual/med_30',
'Merchnum_state_actual/toal_30',
'Merchnum_zip_day_since',
'Merchnum_zip_count_0',
'Merchnum_zip_avg_0',
'Merchnum_zip_max_0',
'Merchnum_zip_med_0',
'Merchnum_zip_total_0',
'Merchnum_zip_actual/avg_0',
'Merchnum_zip_actual/max_0',
'Merchnum_zip_actual/med_0',
'Merchnum_zip_actual/toal_0',
'Merchnum_zip_count_1',
'Merchnum_zip_avg_1',
'Merchnum_zip_max_1',
'Merchnum_zip_med_1',
'Merchnum_zip_total_1',
'Merchnum_zip_actual/avg_1',
'Merchnum_zip_actual/max_1',
'Merchnum_zip_actual/med_1',
'Merchnum_zip_actual/toal_1',
'Merchnum_zip_count_3',
'Merchnum_zip_avg_3',
'Merchnum_zip_max_3',
'Merchnum_zip_med_3',
'Merchnum_zip_total_3',
'Merchnum_zip_actual/avg_3',
'Merchnum_zip_actual/max_3',
'Merchnum_zip_actual/med_3',
'Merchnum_zip_actual/toal_3',
'Merchnum_zip_count_7',
'Merchnum_zip_avg_7',
'Merchnum_zip_max_7',
'Merchnum_zip_med_7',
'Merchnum_zip_total_7',
'Merchnum_zip_actual/avg_7',
'Merchnum_zip_actual/max_7',
'Merchnum_zip_actual/med_7',
'Merchnum_zip_actual/toal_7',
'Merchnum_zip_count_14',
'Merchnum_zip_avg_14',
'Merchnum_zip_max_14',
'Merchnum_zip_med_14',
'Merchnum_zip_total_14',

'Merchnum_zip_actual/avg_14',
'Merchnum_zip_actual/max_14',
'Merchnum_zip_actual/med_14',
'Merchnum_zip_actual/toal_14',
'Merchnum_zip_count_30',
'Merchnum_zip_avg_30',
'Merchnum_zip_max_30',
'Merchnum_zip_med_30',
'Merchnum_zip_total_30',
'Merchnum_zip_actual/avg_30',
'Merchnum_zip_actual/max_30',
'Merchnum_zip_actual/med_30',
'Merchnum_zip_actual/toal_30',
'Amount_transtype_day_since',
'Amount_transtype_count_0',
'Amount_transtype_avg_0',
'Amount_transtype_max_0',
'Amount_transtype_med_0',
'Amount_transtype_total_0',
'Amount_transtype_actual/avg_0',
'Amount_transtype_actual/max_0',
'Amount_transtype_actual/med_0',
'Amount_transtype_actual/toal_0',
'Amount_transtype_count_1',
'Amount_transtype_avg_1',
'Amount_transtype_max_1',
'Amount_transtype_med_1',
'Amount_transtype_total_1',
'Amount_transtype_actual/avg_1',
'Amount_transtype_actual/max_1',
'Amount_transtype_actual/med_1',
'Amount_transtype_actual/toal_1',
'Amount_transtype_count_3',
'Amount_transtype_avg_3',
'Amount_transtype_max_3',
'Amount_transtype_med_3',
'Amount_transtype_total_3',
'Amount_transtype_actual/avg_3',
'Amount_transtype_actual/max_3',
'Amount_transtype_actual/med_3',
'Amount_transtype_actual/toal_3',
'Amount_transtype_count_7',
'Amount_transtype_avg_7',
'Amount_transtype_max_7',

'Amount_transtype_med_7',
'Amount_transtype_total_7',
'Amount_transtype_actual/avg_7',
'Amount_transtype_actual/max_7',
'Amount_transtype_actual/med_7',
'Amount_transtype_actual/toal_7',
'Amount_transtype_count_14',
'Amount_transtype_avg_14',
'Amount_transtype_max_14',
'Amount_transtype_med_14',
'Amount_transtype_total_14',
'Amount_transtype_actual/avg_14',
'Amount_transtype_actual/max_14',
'Amount_transtype_actual/med_14',
'Amount_transtype_actual/toal_14',
'Amount_transtype_count_30',
'Amount_transtype_avg_30',
'Amount_transtype_max_30',
'Amount_transtype_med_30',
'Amount_transtype_total_30',
'Amount_transtype_actual/avg_30',
'Amount_transtype_actual/max_30',
'Amount_transtype_actual/med_30',
'Amount_transtype_actual/toal_30',
'Date_state_day_since',
'Date_state_count_0',
'Date_state_avg_0',
'Date_state_max_0',
'Date_state_med_0',
'Date_state_total_0',
'Date_state_actual/avg_0',
'Date_state_actual/max_0',
'Date_state_actual/med_0',
'Date_state_actual/toal_0',
'Date_state_count_1',
'Date_state_avg_1',
'Date_state_max_1',
'Date_state_med_1',
'Date_state_total_1',
'Date_state_actual/avg_1',
'Date_state_actual/max_1',
'Date_state_actual/med_1',
'Date_state_actual/toal_1',
'Date_state_count_3',

```
'Date_state_avg_3',
'Date_state_max_3',
'Date_state_med_3',
'Date_state_total_3',
'Date_state_actual/avg_3',
'Date_state_actual/max_3',
'Date_state_actual/med_3',
'Date_state_actual/toal_3',
'Date_state_count_7',
'Date_state_avg_7',
'Date_state_max_7',
'Date_state_med_7',
'Date_state_total_7',
'Date_state_actual/avg_7',
'Date_state_actual/max_7',
'Date_state_actual/med_7',
'Date_state_actual/toal_7',
'Date_state_count_14',
'Date_state_avg_14',
'Date_state_max_14',
'Date_state_med_14',
'Date_state_total_14',
'Date_state_actual/avg_14',
'Date_state_actual/max_14',
'Date_state_actual/med_14',
'Date_state_actual/toal_14',
'Date_state_count_30',
'Date_state_avg_30',
'Date_state_max_30',
'Date_state_med_30',
'Date_state_total_30',
'Date_state_actual/avg_30',
'Date_state_actual/max_30',
'Date_state_actual/med_30',
'Date_state_actual/toal_30',
'Date_zip_day_since',
'Date_zip_count_0',
'Date_zip_avg_0',
'Date_zip_max_0',
'Date_zip_med_0',
'Date_zip_total_0',
'Date_zip_actual/avg_0',
'Date_zip_actual/max_0',
'Date_zip_actual/med_0',
'Date_zip_actual/toal_0',
'Date_zip_count_1',
'Date_zip_avg_1',
'Date_zip_max_1',
'Date_zip_med_1',
'Date_zip_total_1',
'Date_zip_actual/avg_1',
'Date_zip_actual/max_1',
'Date_zip_actual/med_1',
'Date_zip_actual/toal_1',
'Date_zip_count_3',
'Date_zip_avg_3',
'Date_zip_max_3',
'Date_zip_med_3',
'Date_zip_total_3',
'Date_zip_actual/avg_3',
'Date_zip_actual/max_3',
'Date_zip_actual/med_3',
'Date_zip_actual/toal_3',
'Date_zip_count_7',
'Date_zip_avg_7',
'Date_zip_max_7',
'Date_zip_med_7',
'Date_zip_total_7',
'Date_zip_actual/avg_7',
'Date_zip_actual/max_7',
'Date_zip_actual/med_7',
'Date_zip_actual/toal_7',
'Date_zip_count_14',
'Date_zip_avg_14',
'Date_zip_max_14',
'Date_zip_med_14',
'Date_zip_total_14',
'Date_zip_actual/avg_14',
'Date_zip_actual/max_14',
'Date_zip_actual/med_14',
'Date_zip_actual/toal_14',
'Date_zip_count_30',
'Date_zip_avg_30',
'Date_zip_max_30',
'Date_zip_med_30',
'Date_zip_total_30',
'Date_zip_actual/avg_30',
'Date_zip_actual/max_30',
'Date_zip_actual/med_30',
'Date_zip_actual/toal_30',
'Amount_description_day_since',
'Amount_description_count_0',
'Amount_description_avg_0',
'Amount_description_max_0',
'Amount_description_med_0',
'Amount_description_total_0',
'Amount_description_actual/avg_0',
'Amount_description_actual/max_0',
'Amount_description_actual/med_0',
'Amount_description_actual/toal_0',
'Amount_description_count_1',
'Amount_description_avg_1',
'Amount_description_max_1',
'Amount_description_med_1',
'Amount_description_total_1',
'Amount_description_actual/avg_1',
'Amount_description_actual/max_1',
'Amount_description_actual/med_1',
'Amount_description_actual/toal_1',
'Amount_description_count_3',
'Amount_description_avg_3',
'Amount_description_max_3',
'Amount_description_med_3',
'Amount_description_total_3',
'Amount_description_actual/avg_3',
'Amount_description_actual/max_3',
'Amount_description_actual/med_3',
'Amount_description_actual/toal_3',
'Amount_description_count_7',
'Amount_description_avg_7',
'Amount_description_max_7',
'Amount_description_med_7',
'Amount_description_total_7',
'Amount_description_actual/avg_7',
'Amount_description_actual/max_7',
'Amount_description_actual/med_7',
'Amount_description_actual/toal_7',
'Amount_description_count_14',
'Amount_description_avg_14',
'Amount_description_max_14',
'Amount_description_med_14',
'Amount_description_total_14',
'Amount_description_actual/avg_14',
'Amount_description_actual/max_14',
'Amount_description_actual/med_14',
'Amount_description_actual/toal_14',
'Amount_description_count_30',
'Amount_description_avg_30',
'Amount_description_max_30',
'Amount_description_med_30',
'Amount_description_total_30',
'Amount_description_actual/avg_30',
'Amount_description_actual/max_30',
'Amount_description_actual/med_30',
'Amount_description_actual/toal_30',
'card_zip_day_since',
'card_zip_count_0',
'card_zip_avg_0',
'card_zip_max_0',
'card_zip_med_0',
'card_zip_total_0',
'card_zip_actual/avg_0',
'card_zip_actual/max_0',
'card_zip_actual/med_0',
'card_zip_actual/toal_0',
'card_zip_count_1',
'card_zip_avg_1',
'card_zip_max_1',
'card_zip_med_1',
'card_zip_total_1',
'card_zip_actual/avg_1',
'card_zip_actual/max_1',
'card_zip_actual/med_1',
'card_zip_actual/toal_1',
'card_zip_count_3',
'card_zip_avg_3',
'card_zip_max_3',
'card_zip_med_3',
'card_zip_total_3',
'card_zip_actual/avg_3',
'card_zip_actual/max_3',
'card_zip_actual/med_3',
'card_zip_actual/toal_3',
'card_zip_count_7',
'card_zip_avg_7',
'card_zip_max_7',
```

```
'card_zip_med_7',
'card_zip_total_7',
'card_zip_actual/avg_7',
'card_zip_actual/max_7',
'card_zip_actual/med_7',
'card_zip_actual/toal_7',
'card_zip_count_14',
'card_zip_avg_14',
'card_zip_max_14',
'card_zip_med_14',
'card_zip_total_14',
'card_zip_actual/avg_14',
'card_zip_actual/max_14',
'card_zip_actual/med_14',
'card_zip_actual/toal_14',
'card_zip_count_30',
'card_zip_avg_30',
'card_zip_max_30',
'card_zip_med_30',
'card_zip_total_30',
'card_zip_actual/avg_30',
'card_zip_actual/max_30',
'card_zip_actual/med_30',
'card_zip_actual/toal_30',
'Amount_date_day_since',
'Amount_date_count_0',
'Amount_date_avg_0',
'Amount_date_max_0',
'Amount_date_med_0',
'Amount_date_total_0',
'Amount_date_actual/avg_0',
'Amount_date_actual/max_0',
'Amount_date_actual/med_0',
'Amount_date_actual/toal_0',
'Amount_date_count_1',
'Amount_date_avg_1',
'Amount_date_max_1',
'Amount_date_med_1',
'Amount_date_total_1',
'Amount_date_actual/avg_1',
'Amount_date_actual/max_1',
'Amount_date_actual/med_1',
'Amount_date_actual/toal_1',
'Amount_date_count_3',
'Amount_date_avg_3',
'Amount_date_max_3',
'Amount_date_med_3',
'Amount_date_total_3',
'Amount_date_actual/avg_3',
'Amount_date_actual/max_3',
'Amount_date_actual/med_3',
'Amount_date_actual/toal_3',
'Amount_date_count_7',
'Amount_date_avg_7',
'Amount_date_max_7',
'Amount_date_med_7',
'Amount_date_total_7',
'Amount_date_actual/avg_7',
'Amount_date_actual/max_7',
'Amount_date_actual/med_7',
'Amount_date_actual/toal_7',
'Amount_date_count_14',
'Amount_date_avg_14',
'Amount_date_max_14',
'Amount_date_med_14',
'Amount_date_total_14',
'Amount_date_actual/avg_14',
'Amount_date_actual/max_14',
'Amount_date_actual/med_14',
'Amount_date_actual/toal_14',
'Amount_date_count_30',
'Amount_date_avg_30',
'Amount_date_max_30',
'Amount_date_med_30',
'Amount_date_total_30',
'Amount_date_actual/avg_30',
'Amount_date_actual/max_30',
'Amount_date_actual/med_30',
'Amount_date_actual/toal_30',
'Amount_state_day_since',
'Amount_state_count_0',
'Amount_state_avg_0',
'Amount_state_max_0',
'Amount_state_med_0',
'Amount_state_total_0',
'Amount_state_actual/avg_0',
'Amount_state_actual/max_0',
'Amount_state_actual/med_0',
'Amount_state_actual/toal_0',
'Amount_state_count_1',
'Amount_state_avg_1',
'Amount_state_max_1',
'Amount_state_med_1',
'Amount_state_total_1',
'Amount_state_actual/avg_1',
'Amount_state_actual/max_1',
'Amount_state_actual/med_1',
'Amount_state_actual/toal_1',
'Amount_state_count_3',
'Amount_state_avg_3',
'Amount_state_max_3',
'Amount_state_med_3',
'Amount_state_total_3',
'Amount_state_actual/avg_3',
'Amount_state_actual/max_3',
'Amount_state_actual/med_3',
'Amount_state_actual/toal_3',
'Amount_state_count_7',
'Amount_state_avg_7',
'Amount_state_max_7',
'Amount_state_med_7',
'Amount_state_total_7',
'Amount_state_actual/avg_7',
'Amount_state_actual/max_7',
'Amount_state_actual/med_7',
'Amount_state_actual/toal_7',
'Amount_state_count_14',
'Amount_state_avg_14',
'Amount_state_max_14',
'Amount_state_med_14',
'Amount_state_total_14',
'Amount_state_actual/avg_14',
'Amount_state_actual/max_14',
'Amount_state_actual/med_14',
'Amount_state_actual/toal_14',
'Amount_state_count_30',
'Amount_state_avg_30',
'Amount_state_max_30',
'Amount_state_med_30',
'Amount_state_total_30',
'Amount_state_actual/avg_30',
'Amount_state_actual/max_30',
'Amount_state_actual/med_30',
'Amount_state_actual/toal_30',
'Amount_zip_day_since',
'Amount_zip_count_0',
'Amount_zip_avg_0',
'Amount_zip_max_0',
'Amount_zip_med_0',
'Amount_zip_total_0',
'Amount_zip_actual/avg_0',
'Amount_zip_actual/max_0',
'Amount_zip_actual/med_0',
'Amount_zip_actual/toal_0',
'Amount_zip_count_1',
'Amount_zip_avg_1',
'Amount_zip_max_1',
'Amount_zip_med_1',
'Amount_zip_total_1',
'Amount_zip_actual/avg_1',
'Amount_zip_actual/max_1',
'Amount_zip_actual/med_1',
'Amount_zip_actual/toal_1',
'Amount_zip_count_3',
'Amount_zip_avg_3',
'Amount_zip_max_3',
'Amount_zip_med_3',
'Amount_zip_total_3',
'Amount_zip_actual/avg_3',
'Amount_zip_actual/max_3',
'Amount_zip_actual/med_3',
'Amount_zip_actual/toal_3',
'Amount_zip_count_7',
'Amount_zip_avg_7',
'Amount_zip_max_7',
'Amount_zip_med_7',
'Amount_zip_total_7',
'Amount_zip_actual/avg_7',
'Amount_zip_actual/max_7',
'Amount_zip_actual/med_7',
'Amount_zip_actual/toal_7',
'Amount_zip_count_14',
'Amount_zip_avg_14',
'Amount_zip_max_14',
'Amount_zip_med_14',
'Amount_zip_total_14',
```

```
'Amount_zip_actual/avg_14',
'Amount_zip_actual/max_14',
'Amount_zip_actual/med_14',
'Amount_zip_actual/toal_14',
'Amount_zip_count_30',
'Amount_zip_avg_30',
'Amount_zip_max_30',
'Amount_zip_med_30',
'Amount_zip_total_30',
'Amount_zip_actual/avg_30',
'Amount_zip_actual/max_30',
'Amount_zip_actual/med_30',
'Amount_zip_actual/toal_30',
'Cardnum_count_0_by_7',
'Cardnum_count_0_by_14',
'Cardnum_count_0_by_30',
'Cardnum_count_1_by_7',
'Cardnum_count_1_by_14',
'Cardnum_count_1_by_30',
'Merchnum_count_0_by_7',
'Merchnum_count_0_by_14',
'Merchnum_count_0_by_30',
'Merchnum_count_1_by_7',
'Merchnum_count_1_by_14',
'Merchnum_count_1_by_30',
'Merch description_count_0_by_7',
'Merch description_count_0_by_14',
'Merch description_count_0_by_30',
'Merch description_count_1_by_7',
'Merch description_count_1_by_14',
'Merch description_count_1_by_30',
'Merch state_count_0_by_7',
'Merch state_count_0_by_14',
'Merch state_count_0_by_30',
'Merch state_count_1_by_7',
'Merch state_count_1_by_14',
'Merch state_count_1_by_30',
'Merch zip_count_0_by_7',
'Merch zip_count_0_by_14',
'Merch zip_count_0_by_30',
'Merch zip_count_1_by_7',
'Merch zip_count_1_by_14',
'Merch zip_count_1_by_30',
'card_merch_count_0_by_7',
'card_merch_count_0_by_14',
'card_merch_count_0_by_30',
'card_merch_count_1_by_7',
'card_merch_count_1_by_14',
'card_merch_count_1_by_30',
'card_state_count_0_by_7',
'card_state_count_0_by_14',
'card_state_count_0_by_30',
'card_state_count_1_by_7',
'card_state_count_1_by_14',
'card_state_count_1_by_30',
'card_transtype_count_0_by_7',
'card_transtype_count_0_by_14',
'card_transtype_count_0_by_30',
'card_transtype_count_1_by_7',
'card_transtype_count_1_by_14',
'card_transtype_count_1_by_30',
'Merch_address_count_0_by_7',
'Merch_address_count_0_by_14',
'Merch_address_count_0_by_30',
'Merch_address_count_1_by_7',
'Merch_address_count_1_by_14',
'Merch_address_count_1_by_30',
'Merchnum_state_count_0_by_7',
'Merchnum_state_count_0_by_14',
'Merchnum_state_count_0_by_30',
'Merchnum_state_count_1_by_7',
'Merchnum_state_count_1_by_14',
'Merchnum_state_count_1_by_30',
'Merchnum_zip_count_0_by_7',
'Merchnum_zip_count_0_by_14',
'Merchnum_zip_count_0_by_30',
'Merchnum_zip_count_1_by_7',
'Merchnum_zip_count_1_by_14',
'Merchnum_zip_count_1_by_30',
'Amount_transtype_count_0_by_7',
'Amount_transtype_count_0_by_14',
'Amount_transtype_count_0_by_30',
'Amount_transtype_count_1_by_7',
'Amount_transtype_count_1_by_14',
'Amount_transtype_count_1_by_30',
'Date_state_count_0_by_7',
'Date_state_count_0_by_14',
'Date_state_count_0_by_30',
'Date_state_count_1_by_7',
'Date_state_count_1_by_14',
'Date_state_count_1_by_30',
'Date_zip_count_0_by_7',
'Date_zip_count_0_by_14',
'Date_zip_count_0_by_30',
'Date_zip_count_1_by_7',
'Date_zip_count_1_by_14',
'Date_zip_count_1_by_30',
'Amount_description_count_0_by_7',
'Amount_description_count_0_by_14',
'Amount_description_count_0_by_30',
'Amount_description_count_1_by_7',
'Amount_description_count_1_by_14',
'Amount_description_count_1_by_30',
'card_zip_count_0_by_7',
'card_zip_count_0_by_14',
'card_zip_count_0_by_30',
'card_zip_count_1_by_7',
'card_zip_count_1_by_14',
'card_zip_count_1_by_30',
'Amount_date_count_0_by_7',
'Amount_date_count_0_by_14',
'Amount_date_count_0_by_30',
'Amount_date_count_1_by_7',
'Amount_date_count_1_by_14',
'Amount_date_count_1_by_30',
'Amount_state_count_0_by_7',
'Amount_state_count_0_by_14',
'Amount_state_count_0_by_30',
'Amount_state_count_1_by_7',
'Amount_state_count_1_by_14',
'Amount_state_count_1_by_30',
'Amount_zip_count_0_by_7',
'Amount_zip_count_0_by_14',
'Amount_zip_count_0_by_30',
'Amount_zip_count_1_by_7',
'Amount_zip_count_1_by_14',
'Amount_zip_count_1_by_30',
'Cardnum_U*',
'Merch_U*',
'Dow_Risk',
'State_risk" ]
```