DSO 562 - Fraud Analytics
Group 103

Credit Card Transactions
Fraud Analysis Report

Overview

Executive Summary	3
Data Description	4
Data Cleaning	5
Candidate Variables	6
Variable/Feature Selection	10
Model Algorithms	16
Results	21
Conclusions	29
Appendix	30

Executive Summary

The cost of credit card fraud is a burden on organizations around the world. According to statista.com, $32 billion were lost worldwide by 2021 as a result of credit card fraud.As with this report, our goal was to build a model that would help predict fraud effectively and efficiently.An analysis was conducted of a real-world dataset with a list of government-related credit card transactions for 2006.As the data included a column showing a transaction's fraud status (fraud or not), it presented a supervised problem.In addition, it contained information about each transaction such as the credit card number, merchant name, or state of the merchant.Overall, the dataset contained 96,753 records and 10 data fields.

An exploratory analysis of each of the 10 data fields in the dataset was conducted.To get a sense of the distribution of each data field, we provided a histogram or table.Of the total 96,753 records in the dataset, 1,059 were fraudulent (1.09%).

Before we could start creating variables, we needed to clean the dataset.In particular, we removed one outlier record due to its extremely high value.Additionally, we analyzed only records that had a type of transaction equal to "P" (purchase) in order to work on a more focused dataset.Finally there were missing values in the "Mercnum," "Merch state," and "Merch zip" data fields, so careful data imputation techniques were used to fill those spaces.

Having cleaned the dataset, we created as many candidate variables as possible.A total of ten different types of variables were created: amount variables, frequency variables, days-since variables, velocity change variables,velocity/days since ratio variables, cross entity uniqueness variables, acceleration variables, variability variables, Benford's Law variables, and a day-of-the-week risk table variable.

Following the process of defining variables, we then performed feature selection using filter and wrapper methods to determine the most useful variables for building machine learning models that can predict potentially fraudulent behavior.After choosing the final variables, the data was divided into three sections for modeling and analysis: training, testing, and out-of-time (OOT).Various models were then run using these variables, including logistic regression, boosted trees, random forests, and neural networks.We found the best model to predict fraud to be Neural Network , which resulted in a ~60% FDR at 3% of the OOT data i.e it was able to capture around 60% of the fraud in the top 3% of the population.

Data Description

The dataset contains observations of credit card transactions conducted by a government organization, and holds transaction details such as card number, date, merchant description, merchant zip, amount and a fraud label associated with each transaction, classified as “1” (fraud) and “0” (no fraud).
There are 96,753 records from the year of 2006, and has 10 fields - 1 Numerical, 8 Categorical and 1 Record field.

Table 1: Numerical Records

Name	# of records	% populated	# of unique values	# records with zero	Mean	SD	Min	Max
Amount	96,753	100%	34,909	None	427.89	10,006.14	0.01	3,102,045.53

Table 2: Categorical Records

Name	# of records	% populated	# of unique values	# records with zero	Most common field value
Cardnum	96,753	100%	1,645	None	5142148452
Date	96,753	100%	365	None	2010-02-28
Merchnum	93,378	96.51%	13,091	None	930090121224
Merch Description	96,753	100%	13,126	None	GSA-FSS-ADV
Merch State	95,558	98.76%	227	None	TN
Merch Zip	92,097	95.19%	4,567	None	38118.0
Transtype	96,753	100%	4	None	P
Fraud	96,753	100%	2	95,694	0

Data Cleaning

First step was to identify exclusions and bad records. For the “Amount” fields, there was a single large amount of 3,102,045.53, which was excluded from the dataset.
As for the transaction type, only records labeled “P” were considered, the rest was filtered out.

The second step was to impute missing values for Merchnum, Merch state and Merch Zip.

1. Merchnum

We first grouped by Merch description and replaced with most frequent values, and out of 3198 null values we ended up having 2038. The “Retail Credit/Debit Adjustment” transactions have been replaced by “Unknown”, and after that we were left with 1347 null values. For the last step, we replaced all the 1347 with 481 unique random numbers based on “Merchant Description”.

2. Merch State

As the first step, for some of the known zip codes we imputed the state code values, after which we were left with 1000 null values. Then we grouped by Merchnum, and replaced with the most frequent values. After this step we still had 968 null values. The “Retail Credit/Debit Adjustment” transactions were replaced by “Unknown”, and after that we were left with 315 null values.

3. Merch Zip

First we grouped by Merchnum and replaced with most frequent values, and out of 4300 null values we ended up having 1101 null values. Then we grouped by Merch description and replaced with the most frequent values, after which we were left with 1073 null values. The “Retail Credit/Debit Adjustment” transactions were replaced by “Unknown”. Afterwards, for both Merch State and Merch Zip, all the null values were replaced with “Unknown”.

Candidate Variables

The goal of the candidate variable creation process was to create as many variables as possible based on the original data fields that could be used to inform a machine learning model about patterns that lead to transaction fraud. The creation of these variables can be split into three broad areas: target encoding, field combination/aggregation, and measurement distribution consideration (Benford's Law).

I. Target Encoding

The goal of the target encoding process is to directly encode what we're trying to predict based on a categorical variable, while reducing potentially dimensionality by not using other categorical field encoding like One-Hot Encoding. The two variables that were chosen for target encoding were the day of week for the transaction and the merchant state of the transaction, and these variables were encoded by including the proportion of frauds based on each category.

When making these variables, it is important to mitigate overfitting as much as possible, since we are directly encoding these variables based on the response variable we're trying to predict (fraud label). Therefore, we made sure to only target encode based on the first 10 months of data, using the last 2 months as our out-of-time data set. Lastly, statistical smoothing was implemented to avoid the effect of too many/too few samples in a single category. The distributions based on these encodings are down below

Figure 1: Day of Week based on Fraud Percentage

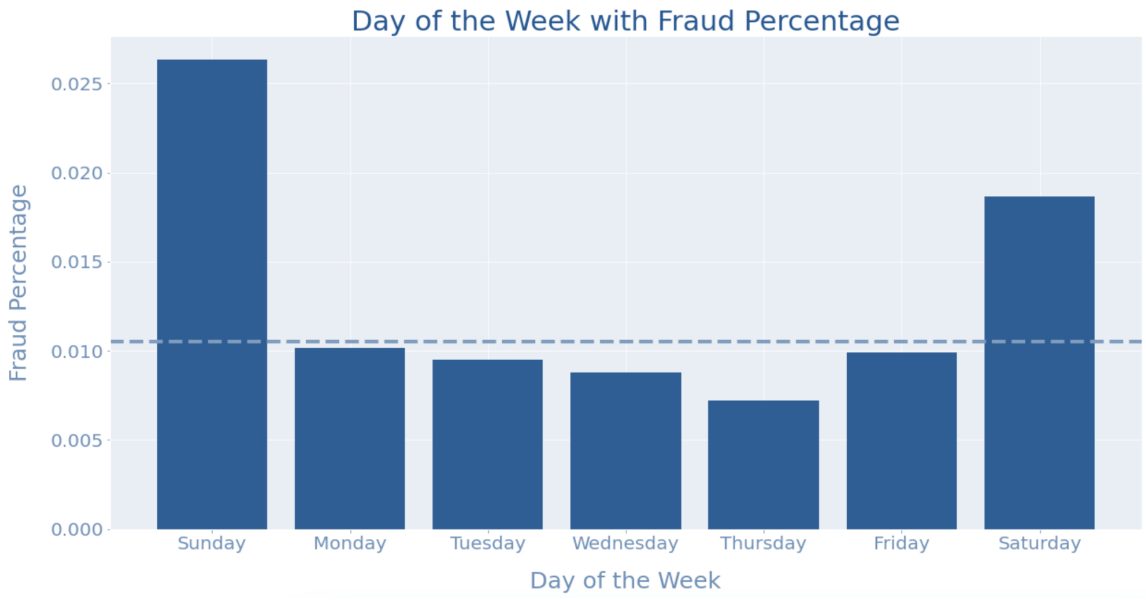
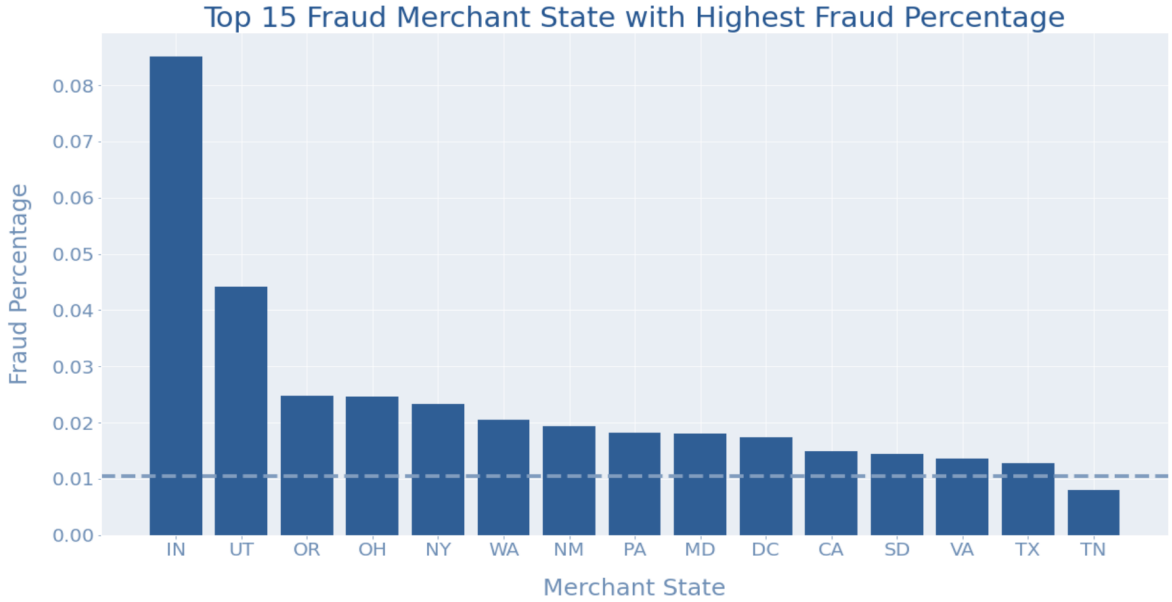


Figure 2: Merchant State based on Fraud Percentage (top 15 states)



II. Field Combination/Aggregations

Before creating variables related to the fields included in our data set relating to the card, merchant, and amount spent, we created 6 core relationship variables that were combinations of some of the original fields to create more descriptive entities that we could create variables based on. These new entities are:

1) merchstate_zip: a combination of the merchant's state and zip code (to investigate transactions at the same merchant in the same zip code)
2) cardnum_merchnum: a combination of the card number and merchant number (to investigate transactions with the same card at the same merchant)
3) cardnum_merchstate: a combination of the card number and merchant state (to investigate transactions with the same card in the same state)
4) cardnum_merchzip: a combination of the card number and merchant zip (to investigate transactions with the same card in the same zip code)
5) merchnum_state: a combination of the merchant number and merchant state (to investigate transactions with a merchant within a single state)
6) merchnum_zip: a combination of merchant number and merchant zip code (to investigate transactions with a merchant within a single zip code)

These 6 new entities were combined with the original fields "Cardnum", "Merchnum", and "Merch zip" to create our final entity list for the field combination/aggregation variables (the amount of candidate variables created for each candidate variable type is included in parentheses).

Note: For the following variables, when "past n days is used," we used {0,1,3,7,14,30,60} days as values for n (0 indicates same day).

"Days Since" Variables (9)

These variables represent the days since a transaction with the same entity value was seen. The notion behind creating these variables was to look into how rarely or commonly a certain type of transaction was made (depending on the entity).

"Frequency" Variables (63)

These variables find how many times a certain entity value was seen in the past n days. These variables allow us to narrow down into a certain time frame and investigate how many transactions were made over a period of time. If "bursts" of many transactions over a period of time indicate fraud, these variables should help a machine learning model capture that behavior.

"Amount Variable" (504)

These variables compute various statistics around the amount spent for transactions over the past n days with the same entity value. These statistics are the average (mean), median, maximum and total amount spent, as well as the actual amount spent divided by these 4 statistics. These variables were created to bring attention to amounts spent that were abnormal in comparison to what is expected, the maximum amount spent that we have seen, or if the amount spent contributes to a significant share of the total amount spent for all transactions with a certain entity value.

"Velocity Change" Variables (72)

These variables find a ratio between the frequency of that entity value ("Frequency" variable) over the past 0 or 1 days and the average amount of daily transactions that happen for transactions with the entity values over the past n days (excluding 0, 1, and 3). This variable was created to understand how often a card was used in comparison to how often we expect it to be used in a single day.

"Velocity Change/Days Since Ratio" Variables (72)
These variables correspond to the ratio between the "Velocity Change" variable and "Days Since" variables. They were created to include information about the abnormality/normality of the amount of transactions in a short period of time with the time that has passed since any transactions were made that had the same entity value.

"Cross Entity Uniqueness" Variables (72)
These variables represent the unique number of entity values for another particular entity for all possible distinct entity pairs. For example, the number of unique card number/merchant number combinations for each merchant zip code. These variables allow us to investigate the diversity in other entity values per entity, and may uncover potential preferences in geographical regions or merchants for fraudulent transactions.

"Acceleration" Variables (72)
These variables are the rate at which the occurrence of an entity value is speeding or accelerating in the number of transactions in 0 or 1 days compared to the past n days (excluding 0, 1, and 3). These variables were created to look into the ramp up of the number of transactions with a particular entity value, with high acceleration indicating a spike in velocity, which could potentially be indicative of fraud.

"Variability" Variables (162)
These variables add information about the average, median, and maximum variability in transaction amounts over the past n days (excluding 6) for an entity value. These variables can give insight into how the number of transactions fluctuates for a certain entity value.

III. Benford's Law Variables
Benford's Law is a proven rule that states that the distribution of the first digit of a set of measurements should not be uniform, and should follow a certain distribution. We wanted to use these candidate variables to investigate whether or not the distribution of transaction amounts spent for a certain card number or merchant number followed this rule. Since a fraudster who is making up transactions is unlikely to make them up such that the first digit distribution follows Benford's Law, this variable should, in theory, be able to identify made-up transaction amounts.

Note: Before starting to create these variables, we removed all FedEx related transactions, since these transactions do not follow Benford's Law but are also known to be non-fraudulent.

According to Benford's law, the proportion of measurements that begin with the digits 3 through 9 should be about 1.096 times the proportion of measurements starting with 1 or 2. Based on this we can find the number

of transactions in the “high” bin, and the number of transactions in the “low” bin (doing this also allows us to deal with not having transactions that span all 9 starting digits for a certain card number/merchant). Using these counts, we define the metric R:

$$R = \frac{1.096 \times n_{\text{low}}}{n_{\text{high}}}$$

We expect R to be close to/exactly 1 if a set of amounts spent follows Benford’s Law, and use this intuition to construct an unusualness score, U:

$$U = \max(R, 1/R).$$

Finally, we incorporate statistical smoothing to deal with high variability that can occur with very different amounts of transactions for all given cards/merchants. This yields our final variable, U*, which is calculated for both a record’s card number and merchant number.

$$U^* = 1 + \left(\frac{U - 1}{1 + \exp^{-t}}\right) \qquad t = (n - n_{mid})/c$$

The values for C and n_mid were 3 and 15, respectively

The total number of candidate variables created were 2 (Target Encoded Variables) + 1,026 (Field Combination/Aggregation Variables) + 2 (Benford’s Law Variables) = 1,030 candidate variables. A full list of these variables is included in the Appendix

Feature Selection Process

After creating the 1030 candidate variables we plan to use for our fraud analysis, we want to find which of these variables are most informative towards fraud and only use those. The feature selection process will allow for a reduction in dimensionality (helping with the curse of dimensionality), while also ensuring the data that is kept is meaningful and contributes significantly towards our final goal of accurately detecting fraud. The process that was followed was to first use a filter and then a wrapper

Feature selection is performed in such a way so that the most statistically important features are kept from the original input. Also, features that are either highly correlated with others or not significant to perform an accurate prediction are ignored. Feature selection is pivotal in modern statistical learning as it often allows reduced computational costs and increased model performance. There are three main methods of feature selection:

1) Filter Methods: Filter methods use statistical measures to evaluate the relationship (correlation) of two distributions and measure the correlation between the distribution of each of the classes of each feature and

the dependent variable. The features that are chosen are the ones with the highest correlation with the dependent variable.

2) Wrapper Methods: Wrapper methods utilize statistical models to evaluate the performance of each feature (or a subset of features) based on a performance metric (accuracy, AUC, f1 score, etc.). A common wrapper method is recursive feature elimination, in which a model recursively uses smaller and smaller sets of features until a desired number of features is reached.

3) Embedded Methods: Embedded methods perform feature elimination as the model is built. A common embedded method for feature selection is regularization, in which a norm is included in the loss function of a statistical model to penalize the number of features used.

Filter Methods

For our analysis, we performed feature selection using the Kolmogorov-Smirnov distance and fraud detection rate.

Kolmogorov–Smirnov (KS) distance:

Kolmogorov-Smirnov (KS) distance measures the maximum distance between two distributions to determine how well the distributions are separated. A higher KS distance value correlates to a better separation between the two distributions and therefore a better feature in the context of feature selection. For this analysis, we used the KS distance metric by calculating the univariate KS value as a filtering method to aid in determining which features provide a better separation between the “Fraud” values of 1 and 0. Meaning, for each numerical candidate variable, we generated the distribution of the two classes (1 and 0) based on the dependent variable ("Fraud” data field). Subsequently, we measured the KS distance between the

distributions of the two classes for each of the numerical candidate variables. More formally,

$$KS = max_x = \sum_{x_{min}}^{x} \left(P_{goods} - P_{bads}\right)$$

We then rank-ordered the KS distance value from high to low for each of the numerical candidate variables and used this ranking to evaluate the importance of each variable.

We filtered out the top 80 variables using this process.

Wrapper Methods

The wrapper was used to go from the 80 to the final 25 variables that were going to be passed into our models. Out of the 3 most common wrapper methods, forwards selection was chosen. Forward selection starts with an empty set of variables, then builds n separate 1-dimensional models (a simple LGBM classifier was used to evaluate performance) and keeps the top variable. At each subsequent step, variables that haven't been chosen are added to a model with the chosen one-at-a-time, keeping the best performing variable every step. This process stops when 25 variables have been chosen,

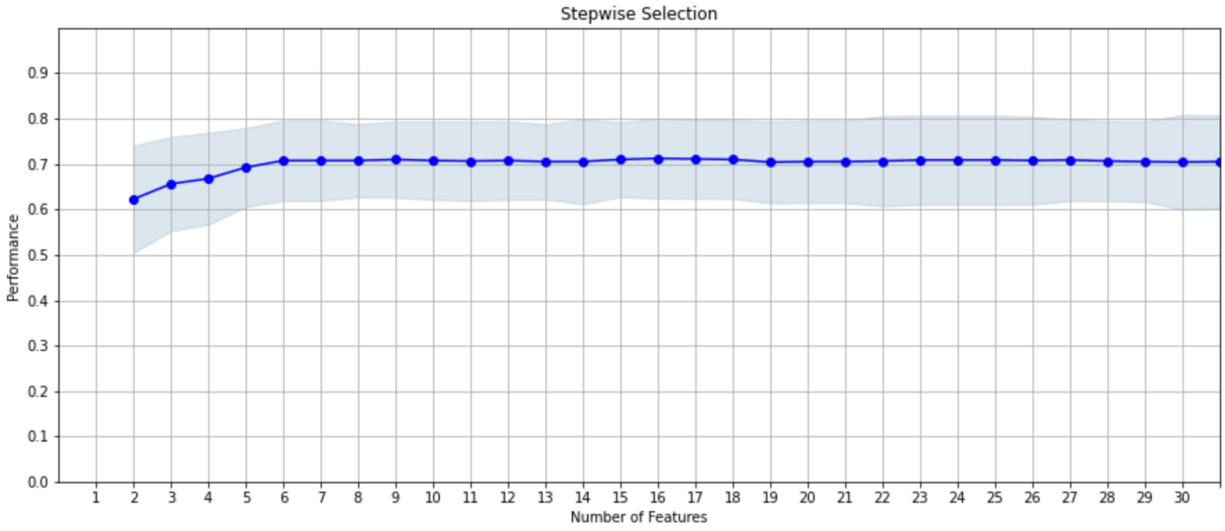
The scoring used for choosing each variable was the Fraud Detection Rate at 3%

Fraud Detection Rate (FDR) at 3%:

In general, the FDR is the percentage of all the frauds that are detected up to a particular cutoff point. In the context of this analysis for feature selection, we used a cutoff threshold of 3% and calculated the multivariate FDR for each addition of candidate variable. The FDR at 3% was determined by first sorting the numerical candidate variables in descending order based on the predicted probability of fraud, and then computing the percentage of frauds in the top 3%

Final list of candidate variables:

We ran the wrapper model multiple times, and checked the resulting model performance with the list of final variables that the wrapper model gave us. Below is a graph indicating model performance as each new variable is added to the model



The table below shows us the top 25 variables sorted by their multivariate importance (No 1 being the most important) and their univariate KS scores .

Table 3: Top 25 Variables

No	Variable name	KS Score	No	Variable name	KS Score
1	Cardnum_zip_total_3	0.62	14	Cardnum_zip_total_1	0.71
2	Cardnum_zip_max_30	0.66	15	Cardnum_zip_max_7	0.71
3	Cardnum_total_1	0.67	16	Cardnum_Merchnum_max_7	0.71
4	Merchnum_state_total_1	0.69	17	Cardnum_state_avg_7	0.71
5	Cardnum_zip_max_60	0.71	18	Merch zip_total_0	0.70
6	Merchnum_total_1	0.71	19	Cardnum_state_max_1	0.70
7	Cardnum_state_max_7	0.71	20	Cardnum_Merchnum_max_1	0.70
8	Cardnum_state_total_1	0.71	21	Cardnum_zip_avg_14	0.71
9	Merchnum_zip_total_1	0.72	22	Cardnum_state_total_0	0.71
10	Cardnum_Merchnum_max_60	0.71	23	Cardnum_zip_max_1	0.71
11	Cardnum_total_3	0.71	24	State_zip_total_0	0.71
12	Cardnum_Merchnum_max_3	0.70	25	Cardnum_Merchnum_total_0	0.71
13	Cardnum_state_max_14	0.70			

Final Model Variables and Descriptions

Cardnum_zip_total_3
- the total amount spent by a card in the same zip code over the past 3 days

Cardnum_zip_max_30
- the maximum amount spent for a card in the same zip code over the past 30 days

Cardnum_total_1
- the total amount spent by a card over the past day

Merchnum_state_total_1
- the total amount spent by a merchant in the same state over the past day

Cardnum_zip_max_60
- the maximum amount spent for a card at the same zip over the past 60 days

Merchnum_total_1
- the total amount spent by a merchant over the past day

Cardnum_state_max_7
- the maximum amount spent for a card at the the same state over the past 7 days

Cardnum_state_total_1
- the total amount spent by a card at the the same state over the past day

Merchnum_zip_total_1
- the total amount spent by a merchant in the same zip over the past day

Cardnum_Merchnum_max_60
- the maximum amount spent for a card at the same merchant over the past 60 days

Cardnum_total_3
- the total amount spent by a card over the past 3 days

Cardnum_Merchnum_max_3
- the maximum amount spent for a card at the same merchant over the past 3 days

Cardnum_state_max_14
- the maximum amount spent for a card at the same state over the past 14 days

Cardnum_zip_total_1
- the total amount spent by a card at the same zip over the past day

Cardnum_zip_max_7
- the maximum amount spent for a card at the same zip over the past 7 days

Cardnum_Merchnum_max_7
- the maximum amount spent for a card at the same merchant over the past 7 days

Cardnum_state_avg_7
- the average amount spent by a card in the same state over the past 7 days

Merch zip_total_0

- the total amount spent at the same merchant/zip code combination in the same day

Cardnum_state_max_1

- the maximum amount spent by a card in the same state over the past day

Cardnum_Merchnum_max_1

- the maximum amount spent by a card at the same merchant over the past day

Cardnum_zip_avg_14

- the average amount spent by a card in the same zip code over the past 14 days

Cardnum_state_total_0

- the total amount spent by a card in the same state in the same day

Cardnum_zip_max_1

- the maximum amount spent for a card in the same zip code over the past day

State_zip_total_0

- the total amount spent in the same state/zip code combination in the same day

Cardnum_Merchnum_total_0

- the total amount spent by a card at the same merchant in the same day

Model Algorithms

After the top 25 variable selection, our team applied multiple supervised machine learning algorithms in order to build the fraud detection models. The main goal was to catch the most fraudulent records from the top 3% of the whole dataset ranked by the model output.

High level summary of the process to went through:

1. In order to build the fraud detection models, employed the following algorithms:
    a. Logistic Regression
    b. Random Forest
    c. Boosted Tree
    d. Neural Networks
2. For each one of them, several models have been built, with a number of different values of parameters and the number of variables.
3. We split the data into 3 sets of testing data and training data for every model, in order to perform a 3-fold cross validation.
4. For each fold we trained the model with the training dataset. We also predicted the training, testing and OOT dataset outputs through application of the fitted model.
5. We sorted the output records in descending order. Afterwards, we computed the FDR using the following formula:

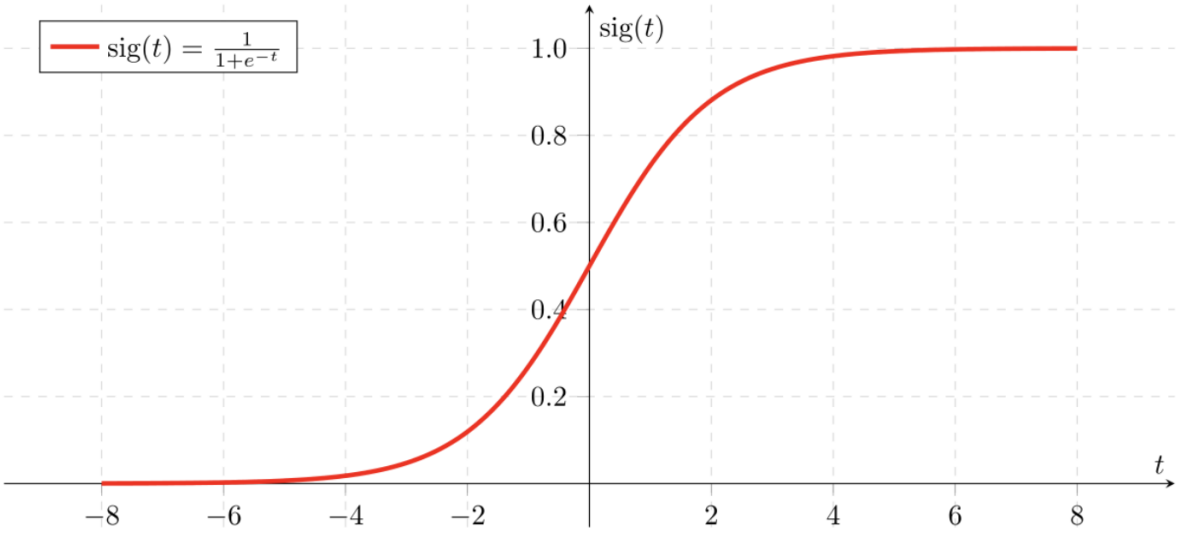
$$\text{FDR at 3\%} = \left(\frac{\text{Number of bad records in the top 3\% of the data}}{\text{Number of bad records in the dataset}} \right)$$

6. At 3% of FDR for training, testing and OOT datasets, we averaged the results derived from 3 folds respectively.
7. Last but not least, for the final fraud detection model, we chose the highest average FDR model.

Explanation of Algorithms

1. Logistic Regression

Logistic Regression is the baseline and most commonly used model to solve binary classification problems. It observes the relationship between dependent binary target variable and independent variables. It returns an “S” shaped squiggle following a sigmoid function. The model can take any real-valued number and return an outcome that is in between 0 and 1.



We have chosen six parameters: penalty, C(regularization), solver, max_iter, multi_class and warm_start, and the results can be observed below.

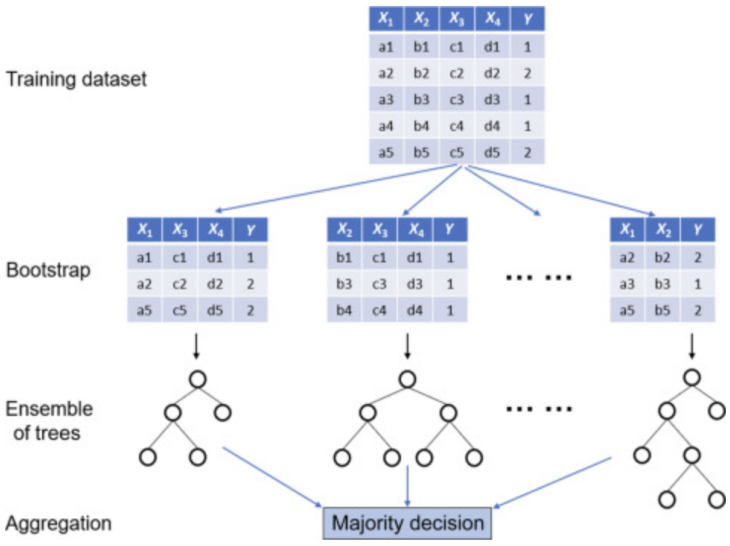
The following combinations were used to build various logistic regression models:
- Number of variables: 10, 20, 30
- Solver: liblinear, lbfgs
- Penalty: l1, l2
- C(regularization): 0.1, 1

Table 4: Average FDR at 3% using Logistic Regression

Logistic Regression	# of variables	solver	penalty	C	Train	Test	OOT
1	10	liblinear	l1	0.1	52.265	52.151	50.978
2	20	liblinear	l1	1	53.197	52.961	51.313
3	30	lbfgs	l2	1	52.174	54.532	50.964
4	10	lbfgs	l2	1	53.197	52.928	51.285
5	20	liblinear	l1	0.1	53.107	52.336	51.021
6	30	lbfgs	l2	0.1	52.426	52.529	50.936

2. Random Forest

Random Forest is a group learning method applied for classification and regression problems. It applies bootstrapped samples from a dataset together with a subset of variables to come up with decision trees. It controls the overfitting problem through random selection of a subset of all attributes every time a decision tree is being built.



Schematic structure of the RF algorithm

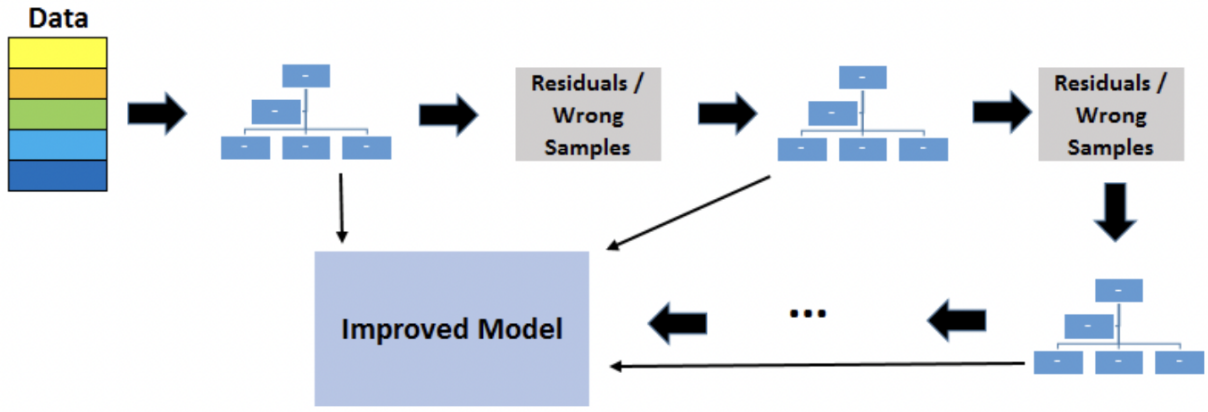
The following combinations were used to build various random forest models:
- Number of variables: 30
- Number of trees: 100, 150, 200, 250
- Max_features: 7
- Max_depth: 60, 70, 80

Table 5: Average FDR at 3% using Random Forest

Random Forest	# of variables	# of trees	Max features	Max depth	Train	Test	OOT
1	30	100	7	60	56.316	55.268	53.813
2	30	100	7	70	56.316	55.268	53.813
3	30	100	7	80	56.316	55.295	53.982
4	30	150	7	60	56.292	55.376	53.813
5	30	150	7	70	56.316	55.268	53.856
6	30	150	7	80	56.316	55.295	53.898
7	30	200	7	60	56.316	55.295	53.856

3. Boosted Trees

Boosted trees are sequences of decision trees that are implemented to improve the prediction result. The primary method is iteratively training a series of weak learners in order to result in a strong learner. The ultimate goal is to minimize the residual error to increase the accuracy and improve the prediction. We have used 100 and 800 for the number of estimators, as it is common to add more trees until no further improvements are identified. Adding more trees helps to slow down the overfitting. The learning rates are 0.1 and 0.01, as taking small incremental steps prevents overfitting and increases the chance for better accuracy on the testing data. We multiplied the prediction by learning rate to slow down the fitting process.



Gradient Boosted Decision Trees

The following combinations were used to build various boosted tree models:
- Number of variables: 10, 20, 30
- Max_depth: 3, 7, 8
- Learning_rate: 0.1, 0.01
- Num_leaves: 7, 120, 250

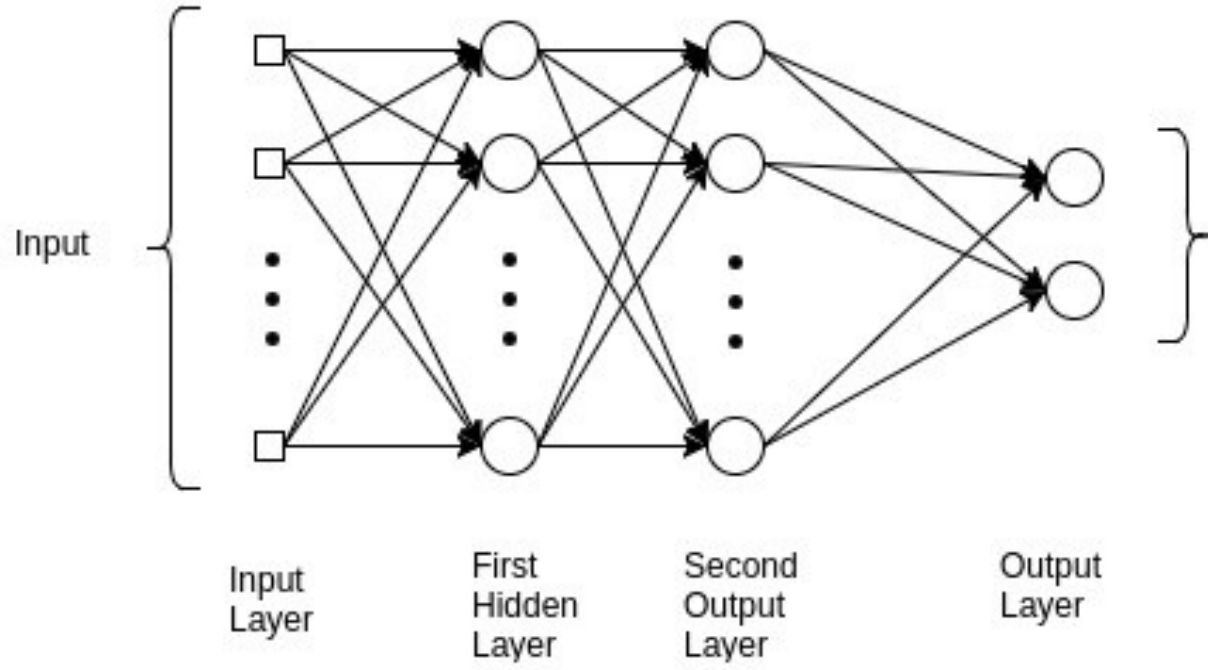
● N_estimators: 100, 800

Table 6: Average FDR at 3% using Boosted Trees

Boosted Trees	# of variables	max_depth	Learning_rate	num_leaves	Train	Test	OOT
1	10	3	0.1	7	55.997	55.985	53.772
2	20	3	0.1	7	55.895	55.827	53.716
3	20	7	0.1	120	56.076	55.853	53.674
4	30	7	0.1	120	55.924	56.145	53.744
5	30	8	0.01	250	55.948	56.073	53.660

4. Neural Networks

In the scope of fraud detection, we also referred to more advanced machine learning algorithms like neural networks. It is a supervised learning algorithm that takes function R(m) -> R(o), where m is the number of input dimensions and o is the number for output dimensions. There can be one or more non-linear (hidden) layers in this model.



The following combinations were used to build various neural networks models:
- Number of variables: 20, 25
- Solver: adam, lbfgs
- Learning_rate: constant, adaptive, invscaling
- Layers: 2, 3
- Hidden_layer_sizes (20, 20), (20, 14), (12, 12, 12), (100)

Table 7: Average FDR at 3% using MLP Classifier

	activation	solver	hidden_layer_s	alpha	learning_rate	max_iter	no. of variables	train	test	oot
Neural Networks	relu	adam	(20, 20)	0.004	constant	140	25	0.84	0.78	0.61
	relu	lbfgs	(20, 14)	0.0001	constant	140	25	0.83	0.77	0.58
	logistic	lbfgs	(20, 14)	0.0001	adaptive	140	25	0.80	0.76	0.57
	logistic	lbfgs	(12, 12, 12)	0.001	invscaling	140	25	0.77	0.73	0.45
	logistic	adam	(16, 16)	0.0001	invscaling	140	25	0.73	0.70	0.40
	identity	lbfgs	(20, 14)	0.001	invscaling	120	25	0.69	0.69	0.36
	relu	sgd	(16, 16)	0.001	adaptive	120	20	0.67	0.66	0.33

Results

Final Model- Neural Networks

In our initial training of the models using logistic regression, boosted trees, random forest, and neural network, our neural network model performed best. As a result of its high fraud detection rates for testing and out-of-time validation datasets, the neural network model consistently outperformed the other models.Comparatively to the other models, not only was the FDR higher, but the variance for the average of the 10 results was also very low.

These are the detailed results tables of the final neural network model for training, testing, and out-of-time datasets:

Table 8: Training Results:

	Bin Statistics					Cumulative Statistics						
Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods_cum	% Bads (FDR)	KS	FPR

1	674	157	517	23.29	76.71	674	157	517	0.23	69.86	69.63	0.30
2	675	585	90	86.67	13.33	1349	742	607	1.11	82.03	80.91	1.22
3	674	643	31	95.40	4.60	2023	1385	638	2.08	86.22	84.14	2.17
4	675	661	14	97.93	2.07	2698	2046	652	3.07	88.11	85.04	3.14
5	674	660	14	97.92	2.08	3372	2706	666	4.06	90	85.94	4.06
6	674	665	9	98.66	1.34	4046	3371	675	5.05	91.22	86.16	4.99
7	675	664	11	98.37	1.63	4721	4035	686	6.05	92.70	86.65	5.88
8	674	667	7	98.96	1.04	5395	4702	693	7.05	93.65	86.60	6.78
9	675	670	5	99.26	0.74	6070	5372	698	8.05	94.32	86.27	7.70
10	674	673	1	99.85	0.15	6744	6045	699	9.06	94.46	85.40	8.65
11	674	670	4	99.41	0.59	7418	6715	703	10.07Hi	95	84.93	9.55
12	675	671	4	99.41	0.60	8093	7386	707	11.07	95.54	84.47	10.45

13	674	672	2	99.70	0.29	8767	8058	709	12.08	95.81	83.73	11.36
14	675	672	3	99.55	0.44	9442	8730	712	13.09	96.21	83.13	12.26
15	674	673	1	99.85	0.15	10116	9403	713	14.09	96.35	82.25	13.19
16	674	671	3	99.55	0.44	10790	10074	716	15.10	96.76	81.65	14.07
17	675	674	1	99.85	0.15	11465	10748	717	16.11	96.89	80.78	14.99
18	674	674	0	100	0	12139	11422	717	17.12	96.89	79.77	15.93
19	675	674	1	99.85	0.15	12814	12096	718	18.13	97.03	78.89	16.85
20	674	674	0	100	0	13488	12770	718	19.14	97.03	77.88	17.78

Table 9: Testing Results:

	Bin Statistics					**Cumulative Statistics**						
Bin%	**# Records**	**# Goods**	**# Bads**	**% Goods**	**% Bads**	**Total # Records**	**Cumulative Goods**	**Cumulative Bads**	**% Goods_cum**	**% Bads (FDR)**	**KS**	**FPR**
1	169	79	90	46.74	53.25	169	79	90	0.47	64.28	63.81	0.88
2	168	149	19	88.69	11.31	337	228	109	1.364	77.86	76.49	2.09
3	169	164	5	97.04	2.96	506	392	114	2.34	81.44	79.08	3.44

4	168	163	5	97.02	2.98	674	555	119	3.32	85	81.68	4.67
5	169	165	4	97.63	2.37	843	720	123	4.31	87.86	83.55	5.85
6	169	166	3	98.22	1.77	1012	886	126	5.30	90	84.70	7.03
7	168	167	1	99.40	0.59	1180	1053	127	6.30	90.71	84.41	8.29
8	169	168	1	99.41	0.59	1349	1221	128	7.30	91.43	84.12	9.53
9	168	164	4	97.62	2.38	1517	1385	132	8.28	94.28	86.00	10.49
10	169	169	0	100	0	1686	1554	132	9.29	94.28	84.99	11.77
11	169	169	0	100	0	1855	1723	132	10.30	94.28	83.98	13.05
12	168	167	1	99.40	0.59	2023	1890	133	11.30	95	83.70	14.21
13	169	169	0	100	0	2192	2059	133	12.31	95	82.68	15.48
14	168	168	0	100	0	2360	2227	133	13.32	95	81.68	16.74
15	169	169	0	100	0	2529	2396	133	14.33	95	80.67	18.01
16	169	169	0	100	0	2698	2565	133	15.34	95	79.66	19.28
17	168	168	0	100	0	2866	2733	133	16.34	95	78.65	20.55
18	169	169	0	100	0	3035	2902	133	17.36	95	77.64	21.82
19	168	168	0	100	0	3203	3070	133	18.36	95	76.64	23.08
20	169	169	0	100	0	3372	3239	133	19.37	95	75.63	24.35

Table 10: OOT Tables:

	Bin Statistics					**Cumulative Statistics**						
Bin%	**# Records**	**# Goods**	**# Bads**	**% Goods**	**% Bads**	**Total # Records**	**Cumulative Goods**	**Cumulative Bads**	**% Goods_cum**	**% Bads (FDR)**	**KS**	**FPR**
1	121	39	82	32.23	67.77	121	39	82	0.32	45.81	45.48	0.47

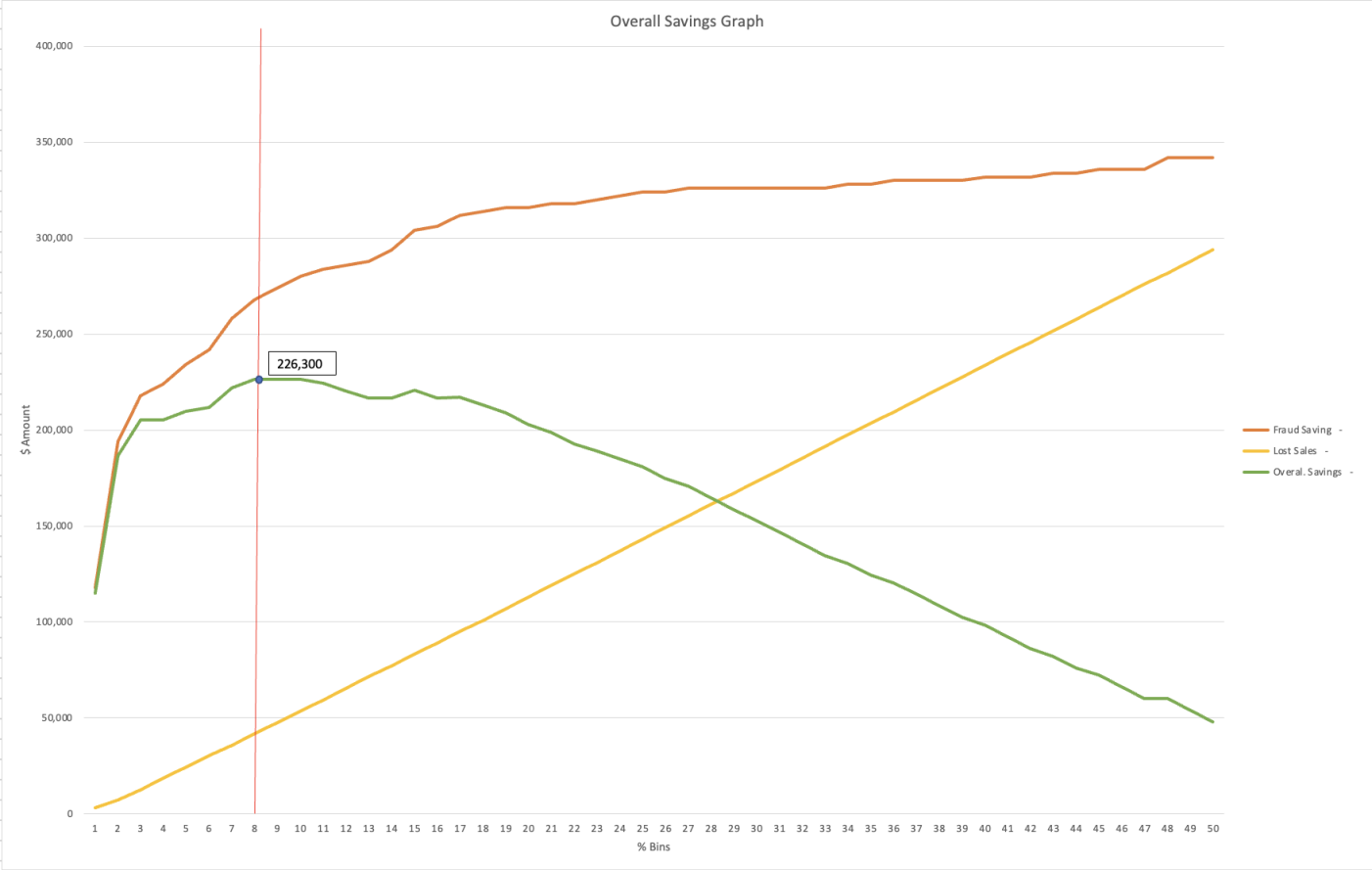
2	121	102	19	84.30	15.70	242	141	101	1.18	56.42	55.24	1.40
3	121	114	7	94.21	5.78	363	255	108	2.14	60.33	58.20	2.36
4	121	113	8	93.39	6.61	484	368	116	3.09	64.80	61.72	3.17
5	121	121	0	100	0	605	489	116	4.10	64.80	60.70	4.21
6	121	112	9	92.57	7.44	726	601	125	5.04	69.83	64.79	4.81
7	121	114	7	94.21	5.78	847	715	132	6.00	73.74	67.74	5.42
8	121	120	1	99.17	0.83	968	835	133	7.01	74.30	67.29	6.28
9	121	119	2	98.35	1.65	1089	954	135	8.00	75.42	67.41	7.07
10	121	118	3	97.52	2.48	1210	1072	138	8.99	77.09	68.10	7.77
11	121	120	1	99.17	0.83	1331	1192	139	10.00	77.65	67.65	8.57
12	121	118	3	97.52	2.48	1452	1310	142	10.99	79.33	68.34	9.22
13	121	118	3	97.52	2.49	1573	1428	145	11.98	81.00	69.02	9.85
14	121	118	3	97.52	2.48	1694	1546	148	12.97	82.68	69.71	10.44
15	121	119	2	98.35	1.65	1815	1665	150	13.97	83.80	69.83	11.1
16	121	119	2	98.35	1.65	1936	1784	152	14.97	84.91	69.95	11.73
17	120	117	3	97.5	2.5	2056	1901	155	15.95	86.59	70.64	12.26
18	121	120	1	99.17	0.83	2177	2021	156	16.96	87.15	70.19	12.95
19	121	121	0	100	0	2298	2142	156	17.93	87.15	69.18	13.73
20	121	120	1	99.17	0.83	2419	2262	157	18.98	87.70	68.73	14.40

Fraud Savings Calculations

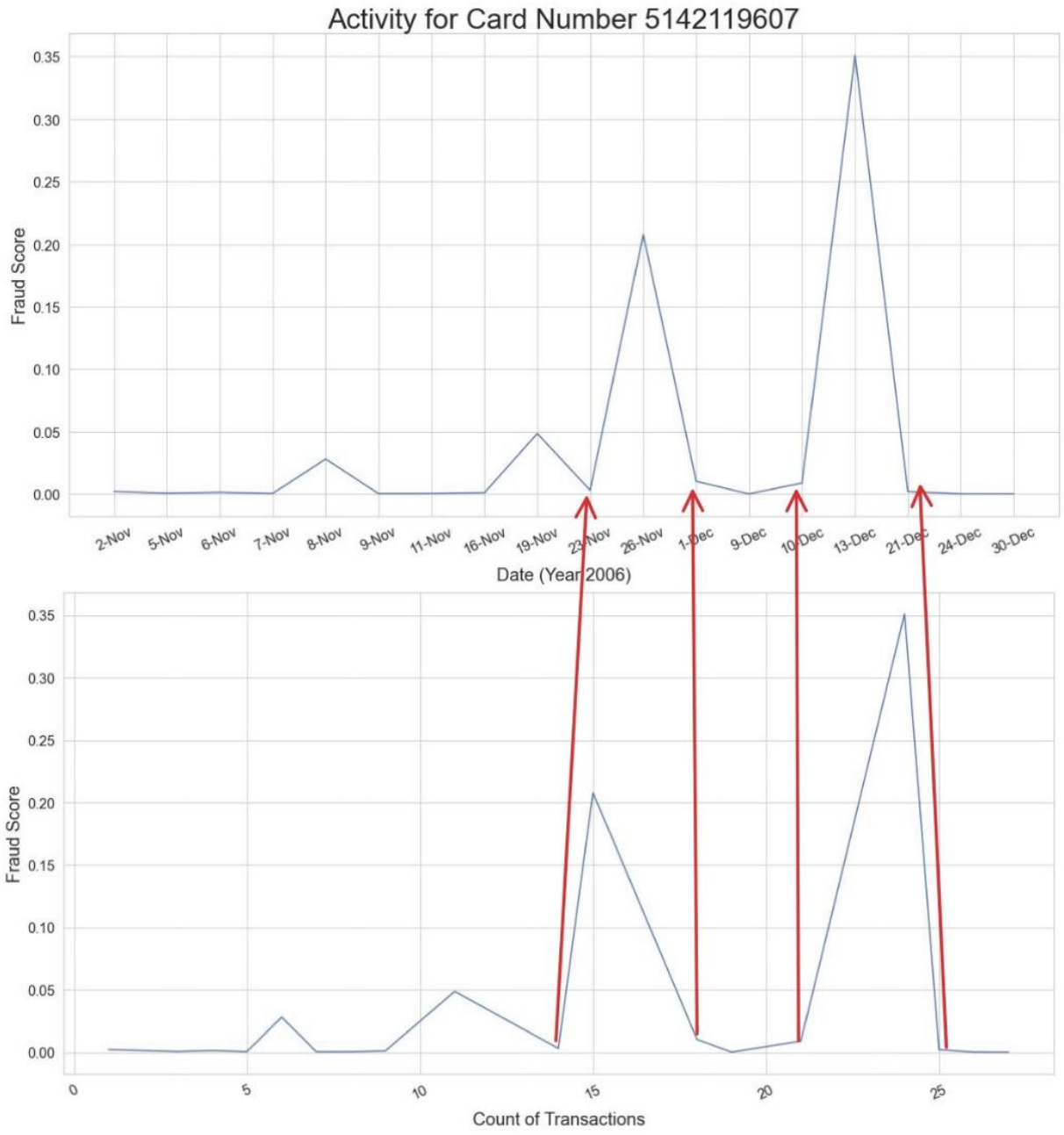
In order to apply our model into business, we created a plot to provide a recommendation for the FDR cutoff point, i.e. any transaction would be marked as fraudulent if it scored above the threshold at the indicated cutoff point.We plotted the Fraud Savings (orange), the Lost Sales (yellow), and the Overall Savings (green) for the out-of-time dataset assuming a gain of $2,000 per fraud detected (true positive, TP) and a loss of $50 per non-fraud tagged as fraud (false positive, FP).

According to our analysis, we recommend a cutoff point at 8% as this threshold maximizes the profits (P) calculated as follows:

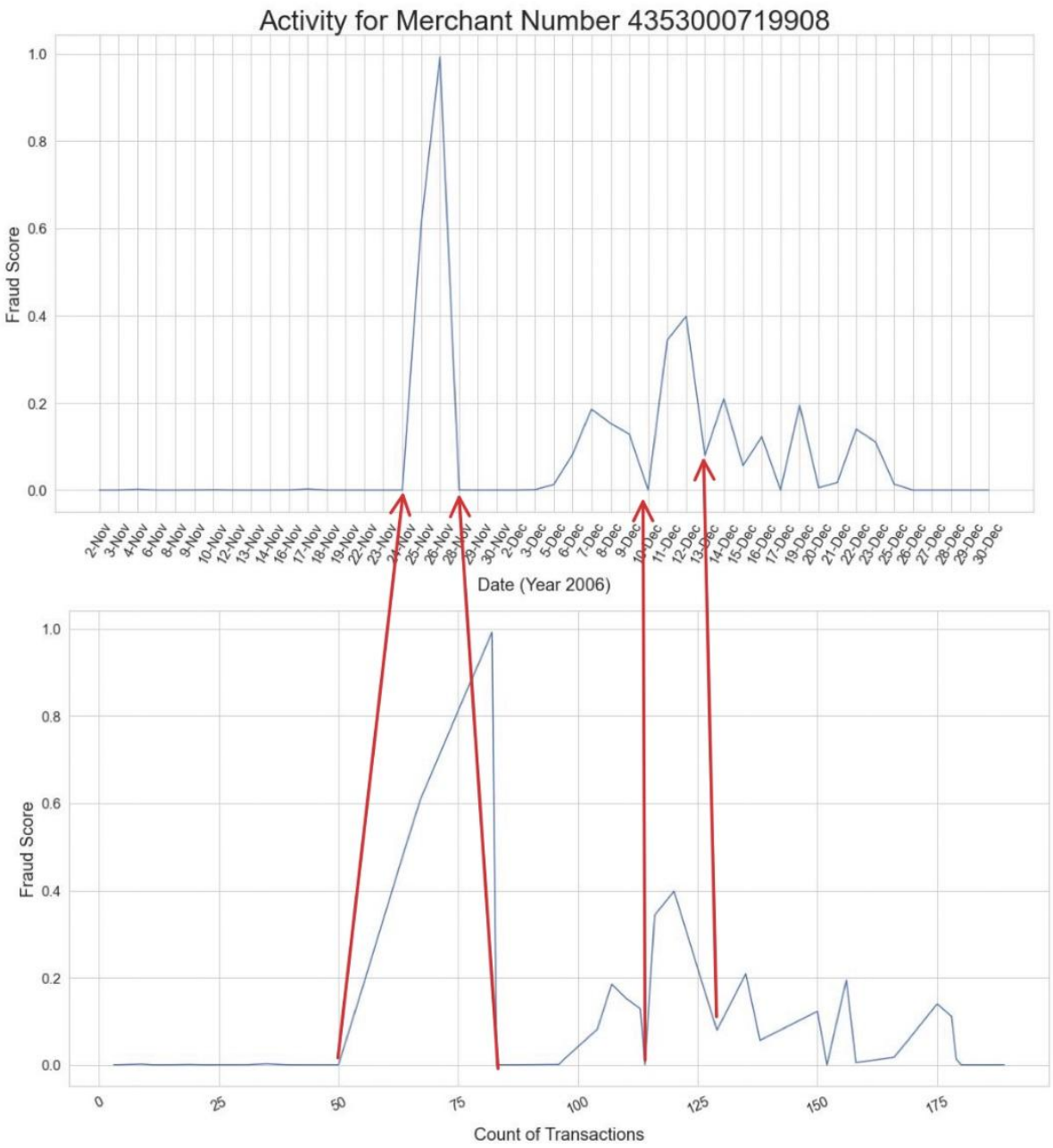
$$P = 2000 * TP - 50 * FP$$



Fraud Score Dynamics



From our analysis, we found there was a correlation between transaction counts and fraud scores. When there is a burst of transaction activity, the fraud score rises rapidly. We illustrate this with the “Cardnum” value of “5142119607”. There was a burst of activity on 11/23/2006 containing 3 transactions and an additional 5 transactions between 12/10/2006 to 12/21/2006. From the graph above, we can see the fraud score rose steeply in accordance with the time period.



Here is another example showing potentially fraudulent activity from the "Merchnum" data field containing the value of “4353000719908”. There was a burst of activity with 34 transactions from 11/24/2006 to 11/28/2006 and another burst of activity with 7 transactions from 12/10/2006 to 12/13/2006. Similar to the “Cardnum” example, we can see in the graph above that the fraud score rose steeply in accordance with the time period.

Conclusions

A comprehensive analysis of credit card transaction fraud cases has been conducted.In the first step, the data was cleaned and all missing values were filled logically.We then created approximately 1030 candidate variables and performed feature selection (filters and wrappers) to choose the best variables.Various models were then run using these variables, including logistic regression, boosted trees, random forests, and neural networks.We found the best model to predict fraud to be Neural Network , which resulted in a ~60% FDR at 3% of the OOT data i.e it was able to capture around 60% of the fraud in the top 3% of the population.

Our best algorithm was used to assign fraud scores to all records, and then we explored and generalized the fraud score dynamics across cardholders and merchants with sharp increases in activity.Our team then introduced and analyzed a hypothetical scenario to suggest to our potential client a score threshold of 8% that maximizes the overall fraud savings by marking all transactions below that threshold as fraud.

During future phases of this analysis, there are several areas where we would want to improve.First, we would like to consult experts so that we can better understand the significance of each variable in the dataset, thereby allowing for a more robust data cleaning process.As of now, we removed some values or used random numbers to clean up the database, based on our understanding.Consulting the experts would also assist us in developing a more comprehensive list of candidate variables. And finally, since this dataset was small and only captured a small amount of fraud activity, we will need more data to perform a robust analysis.

Appendix

Candidate Variables

Recnum
Cardnum
Date
Merchnum
Merch description
Merch state
Merch zip
Transtype
Amount
Fraud
dow
dow_risk
state_risk
state_zip
Cardnum_Merchnum
Cardnum_state
Cardnum_zip
Merchnum_state
Merchnum_zip
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
Cardnum_count_60
Cardnum_avg_60
Cardnum_max_60
Cardnum_med_60
Cardnum_total_60
Cardnum_actual/avg_60
Cardnum_actual/max_60
Cardnum_actual/med_60
Cardnum_actual/toal_60
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
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
Merchnum_count_60
Merchnum_avg_60
Merchnum_max_60
Merchnum_med_60
Merchnum_total_60
Merchnum_actual/avg_60
Merchnum_actual/max_60
Merchnum_actual/med_60
Merchnum_actual/toal_60
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
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
Merch zip_count_60
Merch zip_avg_60
Merch zip_max_60
Merch zip_med_60
Merch zip_total_60
Merch zip_actual/avg_60
Merch zip_actual/max_60
Merch zip_actual/med_60
Merch zip_actual/toal_60
state_zip_day_since
state_zip_count_0
state_zip_avg_0
state_zip_max_0
state_zip_med_0
state_zip_total_0
state_zip_actual/avg_0
state_zip_actual/max_0
state_zip_actual/med_0
state_zip_actual/toal_0
state_zip_count_1

state_zip_avg_1
state_zip_max_1
state_zip_med_1
state_zip_total_1
state_zip_actual/avg_1
state_zip_actual/max_1
state_zip_actual/med_1
state_zip_actual/toal_1
state_zip_count_3
state_zip_avg_3
state_zip_max_3
state_zip_med_3
state_zip_total_3
state_zip_actual/avg_3
state_zip_actual/max_3
state_zip_actual/med_3
state_zip_actual/toal_3
state_zip_count_7
state_zip_avg_7
state_zip_max_7
state_zip_med_7
state_zip_total_7
state_zip_actual/avg_7
state_zip_actual/max_7
state_zip_actual/med_7
state_zip_actual/toal_7
state_zip_count_14
state_zip_avg_14
state_zip_max_14
state_zip_med_14
state_zip_total_14
state_zip_actual/avg_14
state_zip_actual/max_14
state_zip_actual/med_14
state_zip_actual/toal_14
state_zip_count_30
state_zip_avg_30
state_zip_max_30
state_zip_med_30
state_zip_total_30
state_zip_actual/avg_30
state_zip_actual/max_30
state_zip_actual/med_30
state_zip_actual/toal_30
state_zip_count_60
state_zip_avg_60

state_zip_max_60
state_zip_med_60
state_zip_total_60
state_zip_actual/avg_60
state_zip_actual/max_60
state_zip_actual/med_60
state_zip_actual/toal_60
Cardnum_Merchnum_day_since
Cardnum_Merchnum_count_0
Cardnum_Merchnum_avg_0
Cardnum_Merchnum_max_0
Cardnum_Merchnum_med_0
Cardnum_Merchnum_total_0
Cardnum_Merchnum_actual/avg_0
Cardnum_Merchnum_actual/max_0
Cardnum_Merchnum_actual/med_0
Cardnum_Merchnum_actual/toal_0
Cardnum_Merchnum_count_1
Cardnum_Merchnum_avg_1
Cardnum_Merchnum_max_1
Cardnum_Merchnum_med_1
Cardnum_Merchnum_total_1
Cardnum_Merchnum_actual/avg_1
Cardnum_Merchnum_actual/max_1
Cardnum_Merchnum_actual/med_1
Cardnum_Merchnum_actual/toal_1
Cardnum_Merchnum_count_3
Cardnum_Merchnum_avg_3
Cardnum_Merchnum_max_3
Cardnum_Merchnum_med_3
Cardnum_Merchnum_total_3
Cardnum_Merchnum_actual/avg_3
Cardnum_Merchnum_actual/max_3
Cardnum_Merchnum_actual/med_3
Cardnum_Merchnum_actual/toal_3
Cardnum_Merchnum_count_7
Cardnum_Merchnum_avg_7
Cardnum_Merchnum_max_7
Cardnum_Merchnum_med_7
Cardnum_Merchnum_total_7
Cardnum_Merchnum_actual/avg_7
Cardnum_Merchnum_actual/max_7
Cardnum_Merchnum_actual/med_7
Cardnum_Merchnum_actual/toal_7
Cardnum_Merchnum_count_14
Cardnum_Merchnum_avg_14

Cardnum_Merchnum_max_14
Cardnum_Merchnum_med_14
Cardnum_Merchnum_total_14
Cardnum_Merchnum_actual/avg_14
Cardnum_Merchnum_actual/max_14
Cardnum_Merchnum_actual/med_14
Cardnum_Merchnum_actual/toal_14
Cardnum_Merchnum_count_30
Cardnum_Merchnum_avg_30
Cardnum_Merchnum_max_30
Cardnum_Merchnum_med_30
Cardnum_Merchnum_total_30
Cardnum_Merchnum_actual/avg_30
Cardnum_Merchnum_actual/max_30
Cardnum_Merchnum_actual/med_30
Cardnum_Merchnum_actual/toal_30
Cardnum_Merchnum_count_60
Cardnum_Merchnum_avg_60
Cardnum_Merchnum_max_60
Cardnum_Merchnum_med_60
Cardnum_Merchnum_total_60
Cardnum_Merchnum_actual/avg_60
Cardnum_Merchnum_actual/max_60
Cardnum_Merchnum_actual/med_60
Cardnum_Merchnum_actual/toal_60
Cardnum_state_day_since
Cardnum_state_count_0
Cardnum_state_avg_0
Cardnum_state_max_0
Cardnum_state_med_0
Cardnum_state_total_0
Cardnum_state_actual/avg_0
Cardnum_state_actual/max_0
Cardnum_state_actual/med_0
Cardnum_state_actual/toal_0
Cardnum_state_count_1
Cardnum_state_avg_1
Cardnum_state_max_1
Cardnum_state_med_1
Cardnum_state_total_1
Cardnum_state_actual/avg_1
Cardnum_state_actual/max_1
Cardnum_state_actual/med_1
Cardnum_state_actual/toal_1
Cardnum_state_count_3
Cardnum_state_avg_3

Cardnum_state_max_3
Cardnum_state_med_3
Cardnum_state_total_3
Cardnum_state_actual/avg_3
Cardnum_state_actual/max_3
Cardnum_state_actual/med_3
Cardnum_state_actual/toal_3
Cardnum_state_count_7
Cardnum_state_avg_7
Cardnum_state_max_7
Cardnum_state_med_7
Cardnum_state_total_7
Cardnum_state_actual/avg_7
Cardnum_state_actual/max_7
Cardnum_state_actual/med_7
Cardnum_state_actual/toal_7
Cardnum_state_count_14
Cardnum_state_avg_14
Cardnum_state_max_14
Cardnum_state_med_14
Cardnum_state_total_14
Cardnum_state_actual/avg_14
Cardnum_state_actual/max_14
Cardnum_state_actual/med_14
Cardnum_state_actual/toal_14
Cardnum_state_count_30
Cardnum_state_avg_30
Cardnum_state_max_30
Cardnum_state_med_30
Cardnum_state_total_30
Cardnum_state_actual/avg_30
Cardnum_state_actual/max_30
Cardnum_state_actual/med_30
Cardnum_state_actual/toal_30
Cardnum_state_count_60
Cardnum_state_avg_60
Cardnum_state_max_60
Cardnum_state_med_60
Cardnum_state_total_60
Cardnum_state_actual/avg_60
Cardnum_state_actual/max_60
Cardnum_state_actual/med_60
Cardnum_state_actual/toal_60
Cardnum_zip_day_since
Cardnum_zip_count_0
Cardnum_zip_avg_0

Cardnum_zip_max_0
Cardnum_zip_med_0
Cardnum_zip_total_0
Cardnum_zip_actual/avg_0
Cardnum_zip_actual/max_0
Cardnum_zip_actual/med_0
Cardnum_zip_actual/toal_0
Cardnum_zip_count_1
Cardnum_zip_avg_1
Cardnum_zip_max_1
Cardnum_zip_med_1
Cardnum_zip_total_1
Cardnum_zip_actual/avg_1
Cardnum_zip_actual/max_1
Cardnum_zip_actual/med_1
Cardnum_zip_actual/toal_1
Cardnum_zip_count_3
Cardnum_zip_avg_3
Cardnum_zip_max_3
Cardnum_zip_med_3
Cardnum_zip_total_3
Cardnum_zip_actual/avg_3
Cardnum_zip_actual/max_3
Cardnum_zip_actual/med_3
Cardnum_zip_actual/toal_3
Cardnum_zip_count_7
Cardnum_zip_avg_7
Cardnum_zip_max_7
Cardnum_zip_med_7
Cardnum_zip_total_7
Cardnum_zip_actual/avg_7
Cardnum_zip_actual/max_7
Cardnum_zip_actual/med_7
Cardnum_zip_actual/toal_7
Cardnum_zip_count_14
Cardnum_zip_avg_14
Cardnum_zip_max_14
Cardnum_zip_med_14
Cardnum_zip_total_14
Cardnum_zip_actual/avg_14
Cardnum_zip_actual/max_14
Cardnum_zip_actual/med_14
Cardnum_zip_actual/toal_14
Cardnum_zip_count_30
Cardnum_zip_avg_30
Cardnum_zip_max_30

Cardnum_zip_med_30
Cardnum_zip_total_30
Cardnum_zip_actual/avg_30
Cardnum_zip_actual/max_30
Cardnum_zip_actual/med_30
Cardnum_zip_actual/toal_30
Cardnum_zip_count_60
Cardnum_zip_avg_60
Cardnum_zip_max_60
Cardnum_zip_med_60
Cardnum_zip_total_60
Cardnum_zip_actual/avg_60
Cardnum_zip_actual/max_60
Cardnum_zip_actual/med_60
Cardnum_zip_actual/toal_60
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_state_count_60
Merchnum_state_avg_60
Merchnum_state_max_60
Merchnum_state_med_60
Merchnum_state_total_60
Merchnum_state_actual/avg_60
Merchnum_state_actual/max_60
Merchnum_state_actual/med_60
Merchnum_state_actual/toal_60
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
Merchnum_zip_count_60
Merchnum_zip_avg_60
Merchnum_zip_max_60
Merchnum_zip_med_60

Merchnum_zip_total_60
Merchnum_zip_actual/avg_60
Merchnum_zip_actual/max_60
Merchnum_zip_actual/med_60
Merchnum_zip_actual/toal_60
Cardnum_count_0_by_7
Cardnum_count_0_by_14
Cardnum_count_0_by_30
Cardnum_count_0_by_60
Cardnum_count_1_by_7
Cardnum_count_1_by_14
Cardnum_count_1_by_30
Cardnum_count_1_by_60
Merchnum_count_0_by_7
Merchnum_count_0_by_14
Merchnum_count_0_by_30
Merchnum_count_0_by_60
Merchnum_count_1_by_7
Merchnum_count_1_by_14
Merchnum_count_1_by_30
Merchnum_count_1_by_60
Merch zip_count_0_by_7
Merch zip_count_0_by_14
Merch zip_count_0_by_30
Merch zip_count_0_by_60
Merch zip_count_1_by_7
Merch zip_count_1_by_14
Merch zip_count_1_by_30
Merch zip_count_1_by_60
state_zip_count_0_by_7
state_zip_count_0_by_14
state_zip_count_0_by_30
state_zip_count_0_by_60
state_zip_count_1_by_7
state_zip_count_1_by_14
state_zip_count_1_by_30
state_zip_count_1_by_60
Cardnum_Merchnum_count_0_by_7
Cardnum_Merchnum_count_0_by_14
Cardnum_Merchnum_count_0_by_30
Cardnum_Merchnum_count_0_by_60
Cardnum_Merchnum_count_1_by_7
Cardnum_Merchnum_count_1_by_14
Cardnum_Merchnum_count_1_by_30
Cardnum_Merchnum_count_1_by_60
Cardnum_state_count_0_by_7

Cardnum_state_count_0_by_14
Cardnum_state_count_0_by_30
Cardnum_state_count_0_by_60
Cardnum_state_count_1_by_7
Cardnum_state_count_1_by_14
Cardnum_state_count_1_by_30
Cardnum_state_count_1_by_60
Cardnum_zip_count_0_by_7
Cardnum_zip_count_0_by_14
Cardnum_zip_count_0_by_30
Cardnum_zip_count_0_by_60
Cardnum_zip_count_1_by_7
Cardnum_zip_count_1_by_14
Cardnum_zip_count_1_by_30
Cardnum_zip_count_1_by_60
Merchnum_state_count_0_by_7
Merchnum_state_count_0_by_14
Merchnum_state_count_0_by_30
Merchnum_state_count_0_by_60
Merchnum_state_count_1_by_7
Merchnum_state_count_1_by_14
Merchnum_state_count_1_by_30
Merchnum_state_count_1_by_60
Merchnum_zip_count_0_by_7
Merchnum_zip_count_0_by_14
Merchnum_zip_count_0_by_30
Merchnum_zip_count_0_by_60
Merchnum_zip_count_1_by_7
Merchnum_zip_count_1_by_14
Merchnum_zip_count_1_by_30
Merchnum_zip_count_1_by_60
Cardnum_vdratio_0
Cardnum_vdratio_1
Merchnum_vdratio_0
Merchnum_vdratio_1
Merch zip_vdratio_0
Merch zip_vdratio_1
state_zip_vdratio_0
state_zip_vdratio_1
Cardnum_Merchnum_vdratio_0
Cardnum_Merchnum_vdratio_1
Cardnum_state_vdratio_0
Cardnum_state_vdratio_1
Cardnum_zip_vdratio_0
Cardnum_zip_vdratio_1
Merchnum_state_vdratio_0

Merchnum_state_vdratio_1
Merchnum_zip_vdratio_0
Merchnum_zip_vdratio_1
Cardnum_Merchnum_nunique
Cardnum_Merch state_nunique
Cardnum_Merch zip_nunique
Cardnum_Transtype_nunique
Cardnum_state_zip_nunique
Cardnum_Cardnum_Merchnum_nunique
Cardnum_Cardnum_state_nunique
Cardnum_Cardnum_zip_nunique
Cardnum_Merchnum_state_nunique
Cardnum_Merchnum_zip_nunique
Merchnum_Cardnum_nunique
Merchnum_Merch state_nunique
Merchnum_Merch zip_nunique
Merchnum_Transtype_nunique
Merchnum_state_zip_nunique
Merchnum_Cardnum_Merchnum_nunique
Merchnum_Cardnum_state_nunique
Merchnum_Cardnum_zip_nunique
Merchnum_Merchnum_state_nunique
Merchnum_Merchnum_zip_nunique
Merch state_Cardnum_nunique
Merch state_Merchnum_nunique
Merch state_Merch zip_nunique
Merch state_Transtype_nunique
Merch state_state_zip_nunique
Merch state_Cardnum_Merchnum_nunique
Merch state_Cardnum_state_nunique
Merch state_Cardnum_zip_nunique
Merch state_Merchnum_state_nunique
Merch state_Merchnum_zip_nunique
Merch zip_Cardnum_nunique
Merch zip_Merchnum_nunique
Merch zip_Merch state_nunique
Merch zip_Transtype_nunique
Merch zip_state_zip_nunique
Merch zip_Cardnum_Merchnum_nunique
Merch zip_Cardnum_state_nunique
Merch zip_Cardnum_zip_nunique
Merch zip_Merchnum_state_nunique
Merch zip_Merchnum_zip_nunique
Transtype_Cardnum_nunique
Transtype_Merchnum_nunique
Transtype_Merch state_nunique

Transtype_Merch zip_nunique
Transtype_state_zip_nunique
Transtype_Cardnum_Merchnum_nunique
Transtype_Cardnum_state_nunique
Transtype_Cardnum_zip_nunique
Transtype_Merchnum_state_nunique
Transtype_Merchnum_zip_nunique
state_zip_Cardnum_nunique
state_zip_Merchnum_nunique
state_zip_Merch state_nunique
state_zip_Merch zip_nunique
state_zip_Transtype_nunique
state_zip_Cardnum_Merchnum_nunique
state_zip_Cardnum_state_nunique
state_zip_Cardnum_zip_nunique
state_zip_Merchnum_state_nunique
state_zip_Merchnum_zip_nunique
Cardnum_Merchnum_Cardnum_nunique
Cardnum_Merchnum_Merchnum_nunique
Cardnum_Merchnum_Merch state_nunique
Cardnum_Merchnum_Merch zip_nunique
Cardnum_Merchnum_Transtype_nunique
Cardnum_Merchnum_state_zip_nunique
Cardnum_Merchnum_Cardnum_state_nunique
Cardnum_Merchnum_Cardnum_zip_nunique
Cardnum_Merchnum_Merchnum_state_nunique
Cardnum_Merchnum_Merchnum_zip_nunique
Cardnum_state_Cardnum_nunique
Cardnum_state_Merchnum_nunique
Cardnum_state_Merch state_nunique
Cardnum_state_Merch zip_nunique
Cardnum_state_Transtype_nunique
Cardnum_state_state_zip_nunique
Cardnum_state_Cardnum_Merchnum_nunique
Cardnum_state_Cardnum_zip_nunique
Cardnum_state_Merchnum_state_nunique
Cardnum_state_Merchnum_zip_nunique
Cardnum_zip_Cardnum_nunique
Cardnum_zip_Merchnum_nunique
Cardnum_zip_Merch state_nunique
Cardnum_zip_Merch zip_nunique
Cardnum_zip_Transtype_nunique
Cardnum_zip_state_zip_nunique
Cardnum_zip_Cardnum_Merchnum_nunique
Cardnum_zip_Cardnum_state_nunique
Cardnum_zip_Merchnum_state_nunique

Cardnum_zip_Merchnum_zip_nunique
Merchnum_state_Cardnum_nunique
Merchnum_state_Merchnum_nunique
Merchnum_state_Merch state_nunique
Merchnum_state_Merch zip_nunique
Merchnum_state_Transtype_nunique
Merchnum_state_state_zip_nunique
Merchnum_state_Cardnum_Merchnum_nunique
Merchnum_state_Cardnum_state_nunique
Merchnum_state_Cardnum_zip_nunique
Merchnum_state_Merchnum_zip_nunique
Merchnum_zip_Cardnum_nunique
Merchnum_zip_Merchnum_nunique
Merchnum_zip_Merch state_nunique
Merchnum_zip_Merch zip_nunique
Merchnum_zip_Transtype_nunique
Merchnum_zip_state_zip_nunique
Merchnum_zip_Cardnum_Merchnum_nunique
Merchnum_zip_Cardnum_state_nunique
Merchnum_zip_Cardnum_zip_nunique
Merchnum_zip_Merchnum_state_nunique
Cardnum_count_0_by_7_sq
Cardnum_count_0_by_14_sq
Cardnum_count_0_by_30_sq
Cardnum_count_0_by_60_sq
Cardnum_count_1_by_7_sq
Cardnum_count_1_by_14_sq
Cardnum_count_1_by_30_sq
Cardnum_count_1_by_60_sq
Merchnum_count_0_by_7_sq
Merchnum_count_0_by_14_sq
Merchnum_count_0_by_30_sq
Merchnum_count_0_by_60_sq
Merchnum_count_1_by_7_sq
Merchnum_count_1_by_14_sq
Merchnum_count_1_by_30_sq
Merchnum_count_1_by_60_sq
Merch zip_count_0_by_7_sq
Merch zip_count_0_by_14_sq
Merch zip_count_0_by_30_sq
Merch zip_count_0_by_60_sq
Merch zip_count_1_by_7_sq
Merch zip_count_1_by_14_sq
Merch zip_count_1_by_30_sq
Merch zip_count_1_by_60_sq
state_zip_count_0_by_7_sq

state_zip_count_0_by_14_sq
state_zip_count_0_by_30_sq
state_zip_count_0_by_60_sq
state_zip_count_1_by_7_sq
state_zip_count_1_by_14_sq
state_zip_count_1_by_30_sq
state_zip_count_1_by_60_sq
Cardnum_Merchnum_count_0_by_7_sq
Cardnum_Merchnum_count_0_by_14_sq
Cardnum_Merchnum_count_0_by_30_sq
Cardnum_Merchnum_count_0_by_60_sq
Cardnum_Merchnum_count_1_by_7_sq
Cardnum_Merchnum_count_1_by_14_sq
Cardnum_Merchnum_count_1_by_30_sq
Cardnum_Merchnum_count_1_by_60_sq
Cardnum_state_count_0_by_7_sq
Cardnum_state_count_0_by_14_sq
Cardnum_state_count_0_by_30_sq
Cardnum_state_count_0_by_60_sq
Cardnum_state_count_1_by_7_sq
Cardnum_state_count_1_by_14_sq
Cardnum_state_count_1_by_30_sq
Cardnum_state_count_1_by_60_sq
Cardnum_zip_count_0_by_7_sq
Cardnum_zip_count_0_by_14_sq
Cardnum_zip_count_0_by_30_sq
Cardnum_zip_count_0_by_60_sq
Cardnum_zip_count_1_by_7_sq
Cardnum_zip_count_1_by_14_sq
Cardnum_zip_count_1_by_30_sq
Cardnum_zip_count_1_by_60_sq
Merchnum_state_count_0_by_7_sq
Merchnum_state_count_0_by_14_sq
Merchnum_state_count_0_by_30_sq
Merchnum_state_count_0_by_60_sq
Merchnum_state_count_1_by_7_sq
Merchnum_state_count_1_by_14_sq
Merchnum_state_count_1_by_30_sq
Merchnum_state_count_1_by_60_sq
Merchnum_zip_count_0_by_7_sq
Merchnum_zip_count_0_by_14_sq
Merchnum_zip_count_0_by_30_sq
Merchnum_zip_count_0_by_60_sq
Merchnum_zip_count_1_by_7_sq
Merchnum_zip_count_1_by_14_sq
Merchnum_zip_count_1_by_30_sq

Merchnum_zip_count_1_by_60_sq
Cardnum_variability_avg_0
Cardnum_variability_max_0
Cardnum_variability_med_0
Cardnum_variability_avg_1
Cardnum_variability_max_1
Cardnum_variability_med_1
Cardnum_variability_avg_3
Cardnum_variability_max_3
Cardnum_variability_med_3
Cardnum_variability_avg_7
Cardnum_variability_max_7
Cardnum_variability_med_7
Cardnum_variability_avg_14
Cardnum_variability_max_14
Cardnum_variability_med_14
Cardnum_variability_avg_30
Cardnum_variability_max_30
Cardnum_variability_med_30
Merchnum_variability_avg_0
Merchnum_variability_max_0
Merchnum_variability_med_0
Merchnum_variability_avg_1
Merchnum_variability_max_1
Merchnum_variability_med_1
Merchnum_variability_avg_3
Merchnum_variability_max_3
Merchnum_variability_med_3
Merchnum_variability_avg_7
Merchnum_variability_max_7
Merchnum_variability_med_7
Merchnum_variability_avg_14
Merchnum_variability_max_14
Merchnum_variability_med_14
Merchnum_variability_avg_30
Merchnum_variability_max_30
Merchnum_variability_med_30
Merch zip_variability_avg_0
Merch zip_variability_max_0
Merch zip_variability_med_0
Merch zip_variability_avg_1
Merch zip_variability_max_1
Merch zip_variability_med_1
Merch zip_variability_avg_3
Merch zip_variability_max_3
Merch zip_variability_med_3

Merch zip_variability_avg_7
Merch zip_variability_max_7
Merch zip_variability_med_7
Merch zip_variability_avg_14
Merch zip_variability_max_14
Merch zip_variability_med_14
Merch zip_variability_avg_30
Merch zip_variability_max_30
Merch zip_variability_med_30
state_zip_variability_avg_0
state_zip_variability_max_0
state_zip_variability_med_0
state_zip_variability_avg_1
state_zip_variability_max_1
state_zip_variability_med_1
state_zip_variability_avg_3
state_zip_variability_max_3
state_zip_variability_med_3
state_zip_variability_avg_7
state_zip_variability_max_7
state_zip_variability_med_7
state_zip_variability_avg_14
state_zip_variability_max_14
state_zip_variability_med_14
state_zip_variability_avg_30
state_zip_variability_max_30
state_zip_variability_med_30
Cardnum_Merchnum_variability_avg_0
Cardnum_Merchnum_variability_max_0
Cardnum_Merchnum_variability_med_0
Cardnum_Merchnum_variability_avg_1
Cardnum_Merchnum_variability_max_1
Cardnum_Merchnum_variability_med_1
Cardnum_Merchnum_variability_avg_3
Cardnum_Merchnum_variability_max_3
Cardnum_Merchnum_variability_med_3
Cardnum_Merchnum_variability_avg_7
Cardnum_Merchnum_variability_max_7
Cardnum_Merchnum_variability_med_7
Cardnum_Merchnum_variability_avg_14
Cardnum_Merchnum_variability_max_14
Cardnum_Merchnum_variability_med_14
Cardnum_Merchnum_variability_avg_30
Cardnum_Merchnum_variability_max_30
Cardnum_Merchnum_variability_med_30
Cardnum_state_variability_avg_0

Cardnum_state_variability_max_0
Cardnum_state_variability_med_0
Cardnum_state_variability_avg_1
Cardnum_state_variability_max_1
Cardnum_state_variability_med_1
Cardnum_state_variability_avg_3
Cardnum_state_variability_max_3
Cardnum_state_variability_med_3
Cardnum_state_variability_avg_7
Cardnum_state_variability_max_7
Cardnum_state_variability_med_7
Cardnum_state_variability_avg_14
Cardnum_state_variability_max_14
Cardnum_state_variability_med_14
Cardnum_state_variability_avg_30
Cardnum_state_variability_max_30
Cardnum_state_variability_med_30
Cardnum_zip_variability_avg_0
Cardnum_zip_variability_max_0
Cardnum_zip_variability_med_0
Cardnum_zip_variability_avg_1
Cardnum_zip_variability_max_1
Cardnum_zip_variability_med_1
Cardnum_zip_variability_avg_3
Cardnum_zip_variability_max_3
Cardnum_zip_variability_med_3
Cardnum_zip_variability_avg_7
Cardnum_zip_variability_max_7
Cardnum_zip_variability_med_7
Cardnum_zip_variability_avg_14
Cardnum_zip_variability_max_14
Cardnum_zip_variability_med_14
Cardnum_zip_variability_avg_30
Cardnum_zip_variability_max_30
Cardnum_zip_variability_med_30
Merchnum_state_variability_avg_0
Merchnum_state_variability_max_0
Merchnum_state_variability_med_0
Merchnum_state_variability_avg_1
Merchnum_state_variability_max_1
Merchnum_state_variability_med_1
Merchnum_state_variability_avg_3
Merchnum_state_variability_max_3
Merchnum_state_variability_med_3
Merchnum_state_variability_avg_7
Merchnum_state_variability_max_7

Merchnum_state_variability_med_7
Merchnum_state_variability_avg_14
Merchnum_state_variability_max_14
Merchnum_state_variability_med_14
Merchnum_state_variability_avg_30
Merchnum_state_variability_max_30
Merchnum_state_variability_med_30
Merchnum_zip_variability_avg_0
Merchnum_zip_variability_max_0
Merchnum_zip_variability_med_0
Merchnum_zip_variability_avg_1
Merchnum_zip_variability_max_1
Merchnum_zip_variability_med_1
Merchnum_zip_variability_avg_3
Merchnum_zip_variability_max_3
Merchnum_zip_variability_med_3
Merchnum_zip_variability_avg_7
Merchnum_zip_variability_max_7
Merchnum_zip_variability_med_7
Merchnum_zip_variability_avg_14
Merchnum_zip_variability_max_14
Merchnum_zip_variability_med_14
Merchnum_zip_variability_avg_30
Merchnum_zip_variability_max_30
Merchnum_zip_variability_med_30

DQR

a. Numerical Fields

Field Name	# Records that have a value	% Populated	% zero	Min	Max	Mean	Std Dev.
Amount	96,753	100%	0	0.01$	3102045.53$	427.8857$	10006.14$

b. Categorical Fields

Field Name	# Records that have a value	% Populated	# Unique values	Most common value
Recnum	96,753	100%	96,753	NA
Cardnum	96,753	100%	1,645	514214852
Merchnum	93,378	96.51%	13,092	930090121224
Merch Description	96,753	100%	13,126	'G5A-FSS-ADV'
Merch state	95,558	98.76%	228	'TN
Merch zip	92,097	95.19%	4,568	38118
Transtype	96,753	100%	4	'P'
Fraud	96,753	100%	2	0

c. Date Fields

Field Name	# records that have a value	% populated	# unique values	Most common value
Date	96,753	100%	365	2006-02-28

Description

This DQR summarizes the “card transactions.csv” dataset of US card transactions conducted from 2006-01-10 till 2006-12-31. The dataset has 10 fields with 96,753 records. The recorded transactions are made by government representatives, and the products and/or services obtained were related to government operational work.

2. Tables

Table 1. Table of Categorical Variables

Field Name	Num. Records Populated	% Records Populated	Num Unique Values	Most Common Value	Frequency	%
Recnum	96753	100	96753	-	-	-
Cardnum	96753	100	1645	5142148452	1192	1.23
Date	96753	100	365	2006-02-28	684	0.71
Merchnum	93378	96.51	13091	930090121224	9310	9.62
Merch description	96753	100	13126	GSA-FSS-ADV	1688	1.74
Merch state	95558	98.76	227	TN	12035	12.44
Merch zip	92097	95.18	4567	38118.0	11868	12.27
Transtype	96753	100	4	P	96398	99.63
Fraud	96753	100	2	0	95694	98.91

Table 2 : Table of Numerical Variables

Field Name	Num Records Populated	% Records Populated	Num Unique Values	Most Common Value	Min	Max	Mean	SD
Amount	96753	100	34909	3.62	0.01	3102045.53	427.88	10006.14

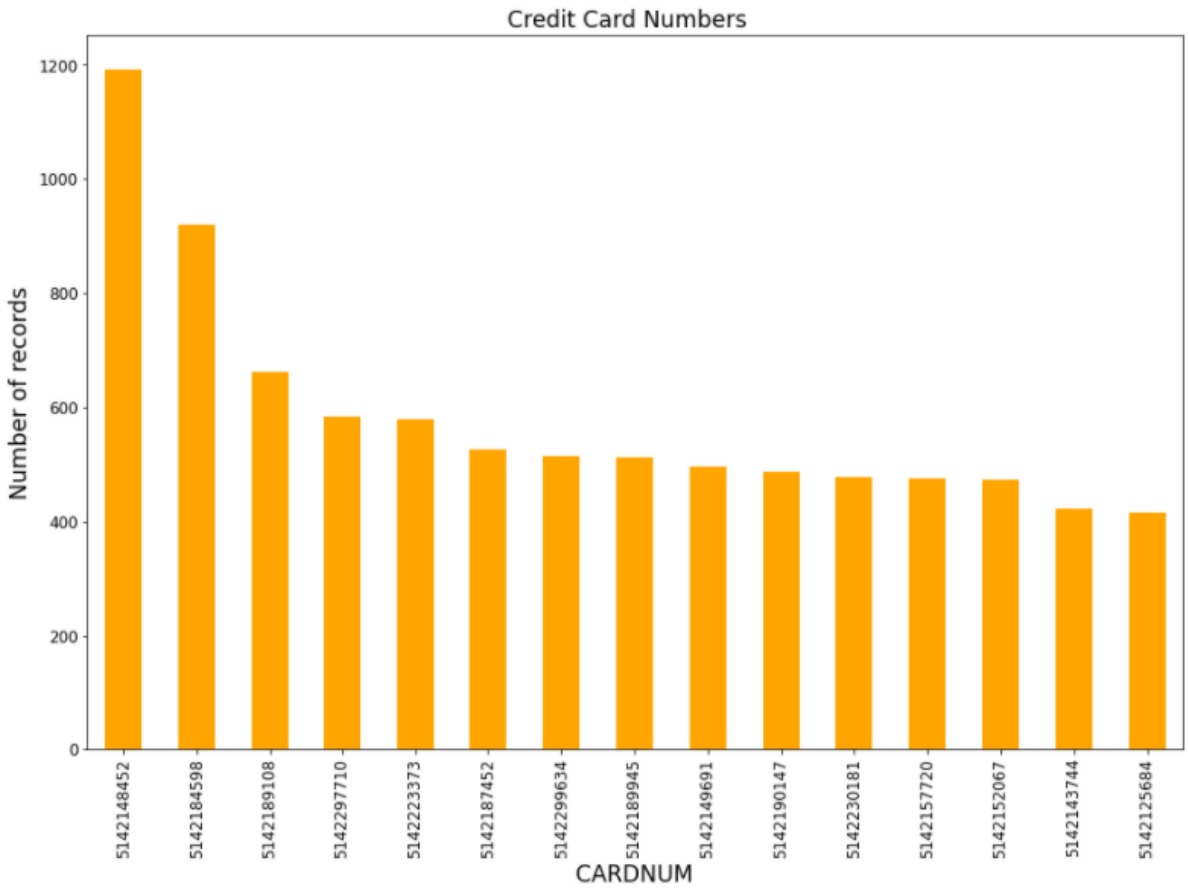
3. Field Descriptions

3.1. RECNUM

“RECNUM” is a unique identifier of each transaction in the dataset. Its primary function is to label the observations. It has 96,753 unique entries, with no missing values.

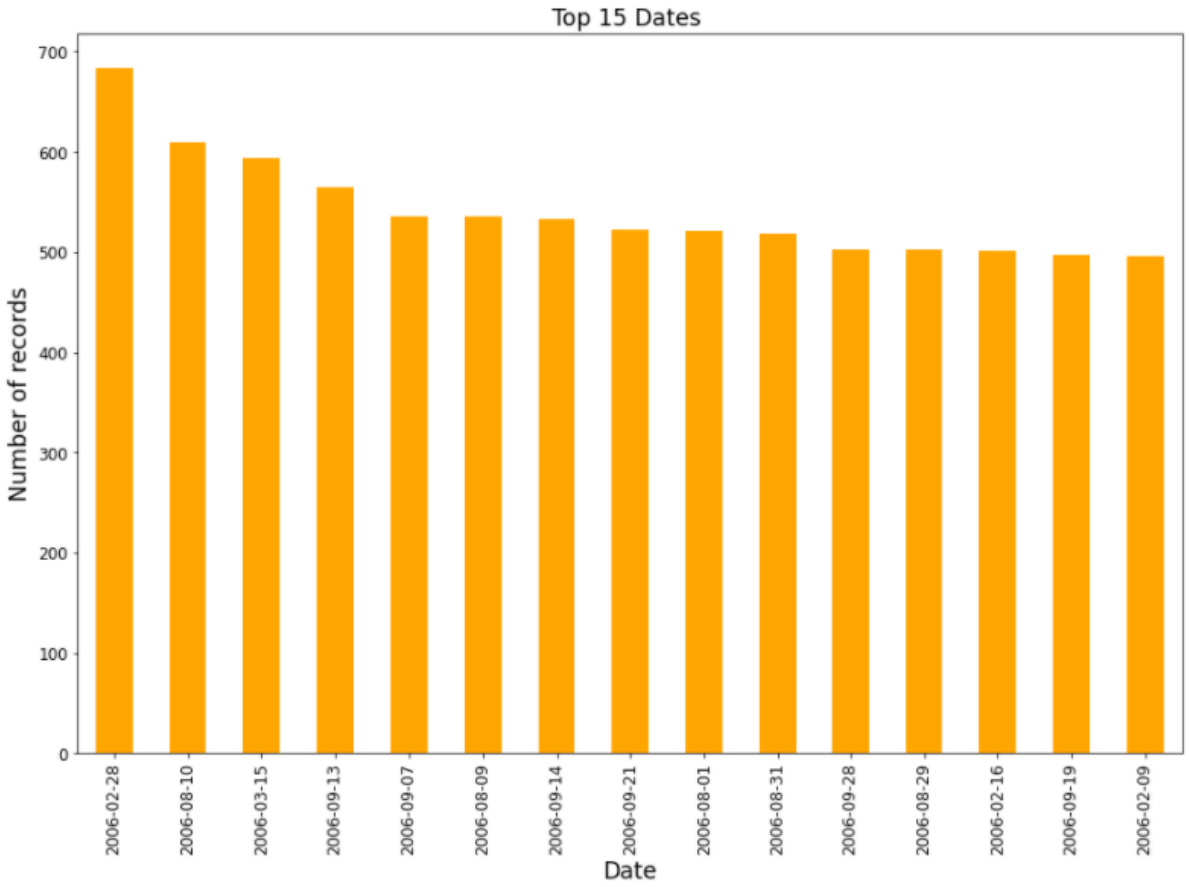
3.2. CARDNUM

CARDNUM is the credit card number according to each transaction. Below is the bar chart of the top 15 credit card numbers and their distribution.



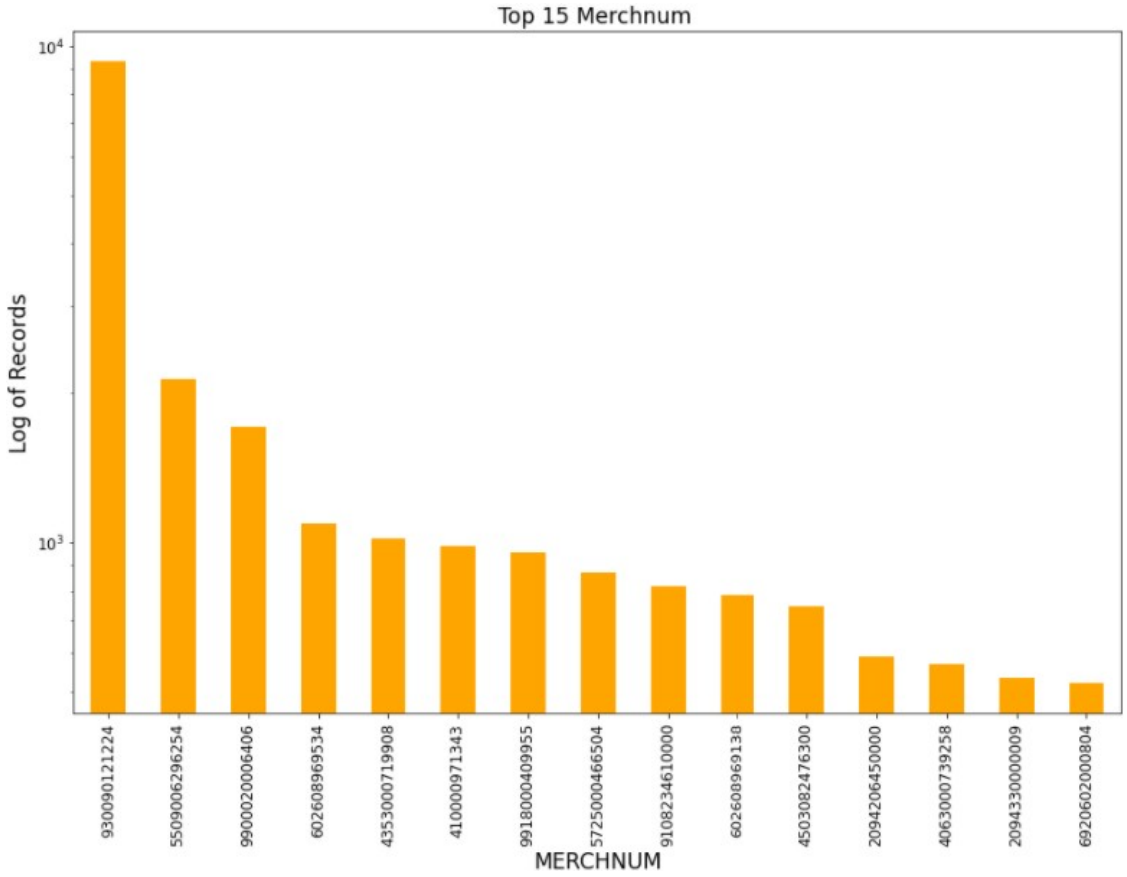
3.3. DATE

DATE is the day when the according transactions were conducted. The following bar chart represents the top 15 days in 2006 with the highest number of transactions registered.



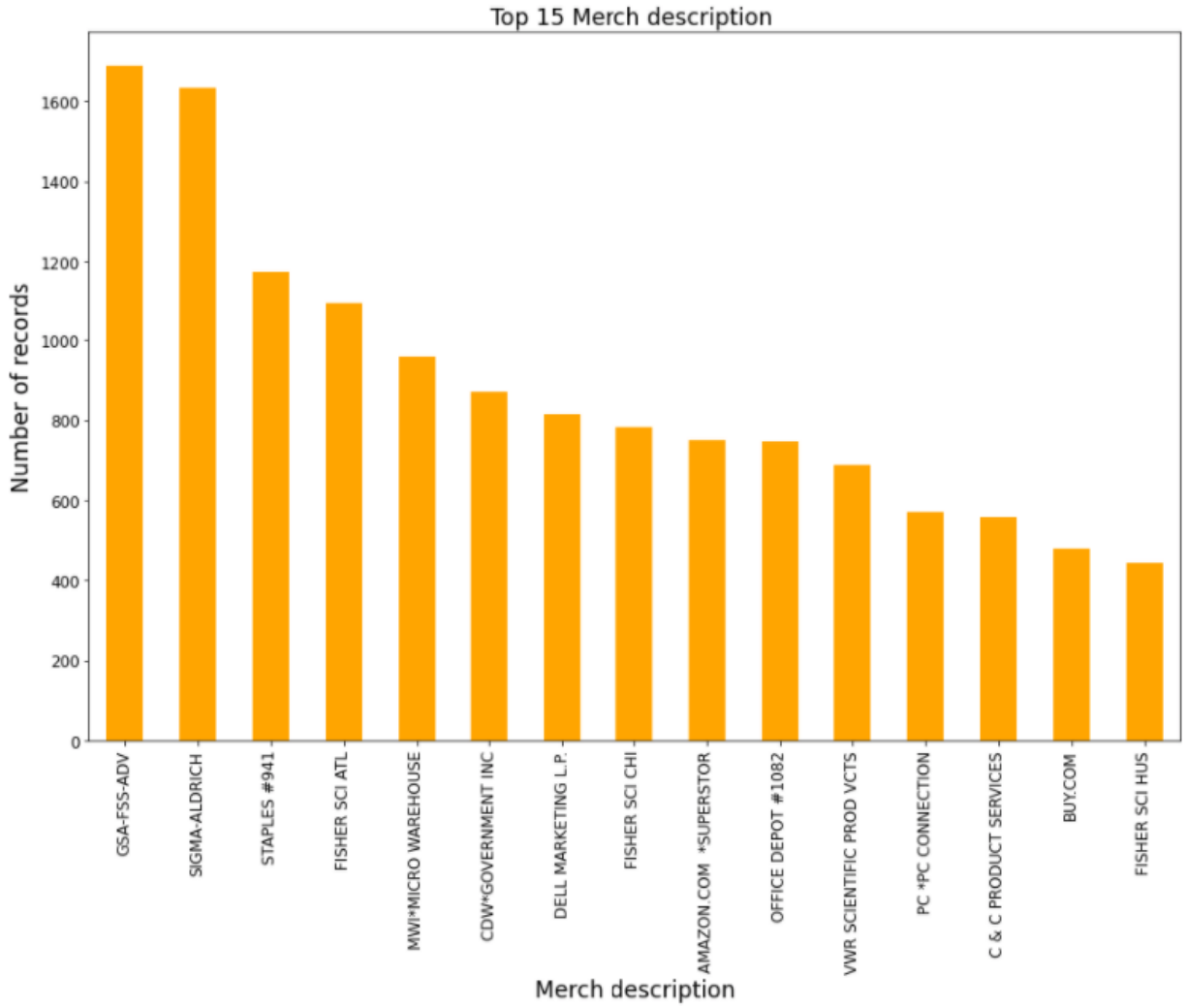
3.4. MERCHNUM

MERCHNUM is the identification number for each merchant. The bar chart shows the top 15 Marchnum on a log scale of records on the Y axis.



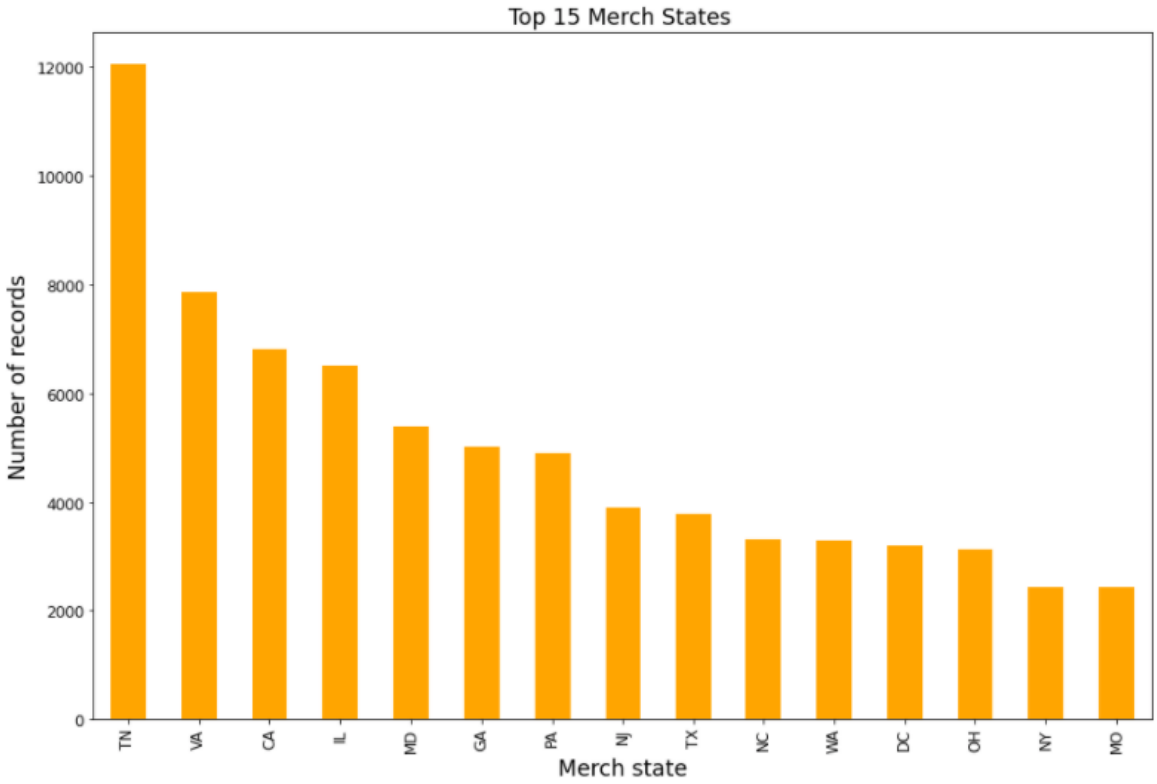
3.5. MERCH DESCRIPTION

MERCH DESCRIPTION gives a brief explanation on the products or services associated with the corresponding credit card transaction. Below is the bar chart of the top 15 Merchant descriptions.



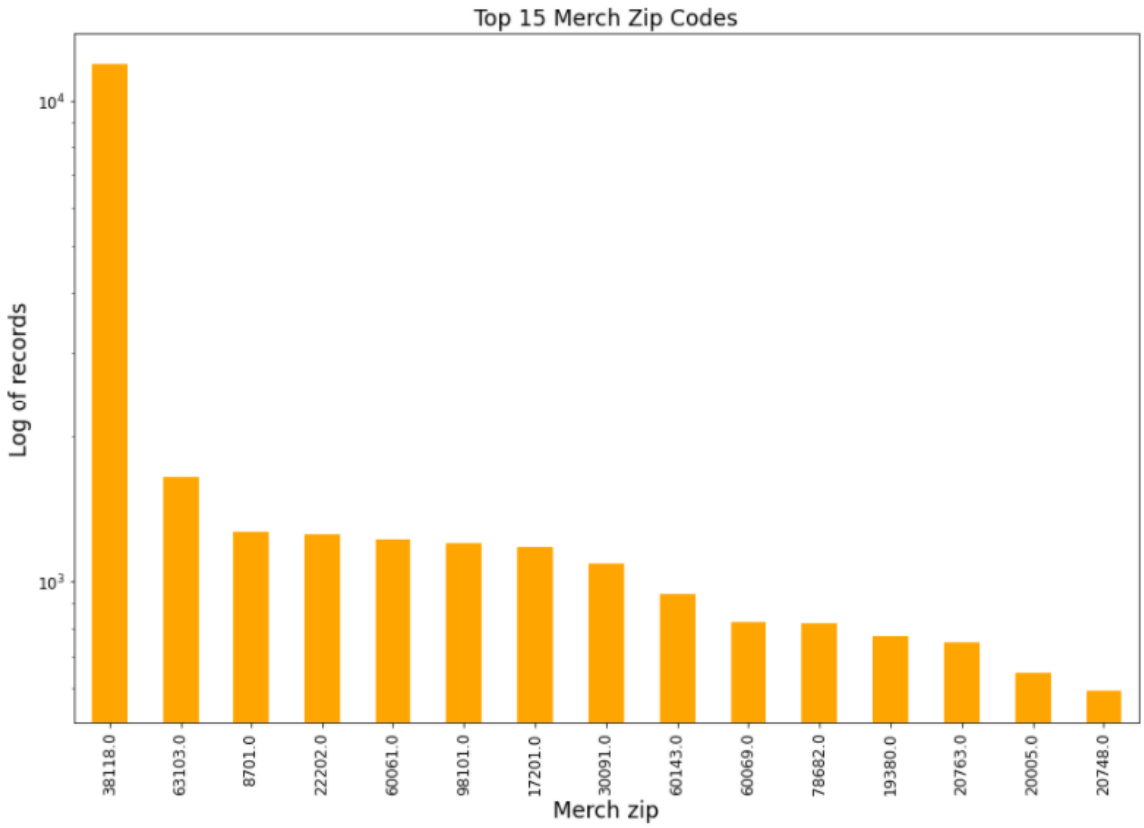
3.6. MERCH STATE

MERCH STATE is the state where the merchant is registered/located.
Note: There are 277 Unique Values in the dataset for States, however, as we know there are only 50 states in US. All the other entries are invalid.



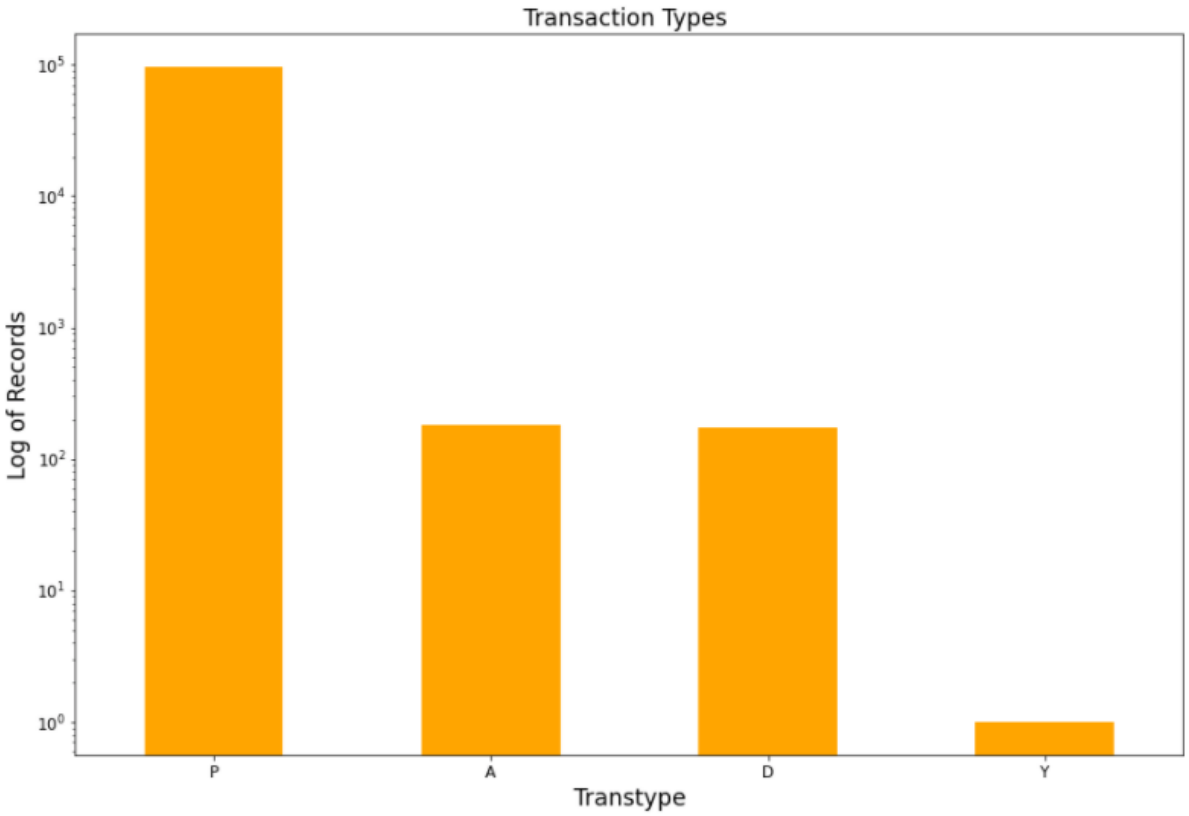
3.7. MERCH ZIP

MERCH ZIP is a 5-digit code of the merchant location. The top 15 occurrences for zip codes are shown above.



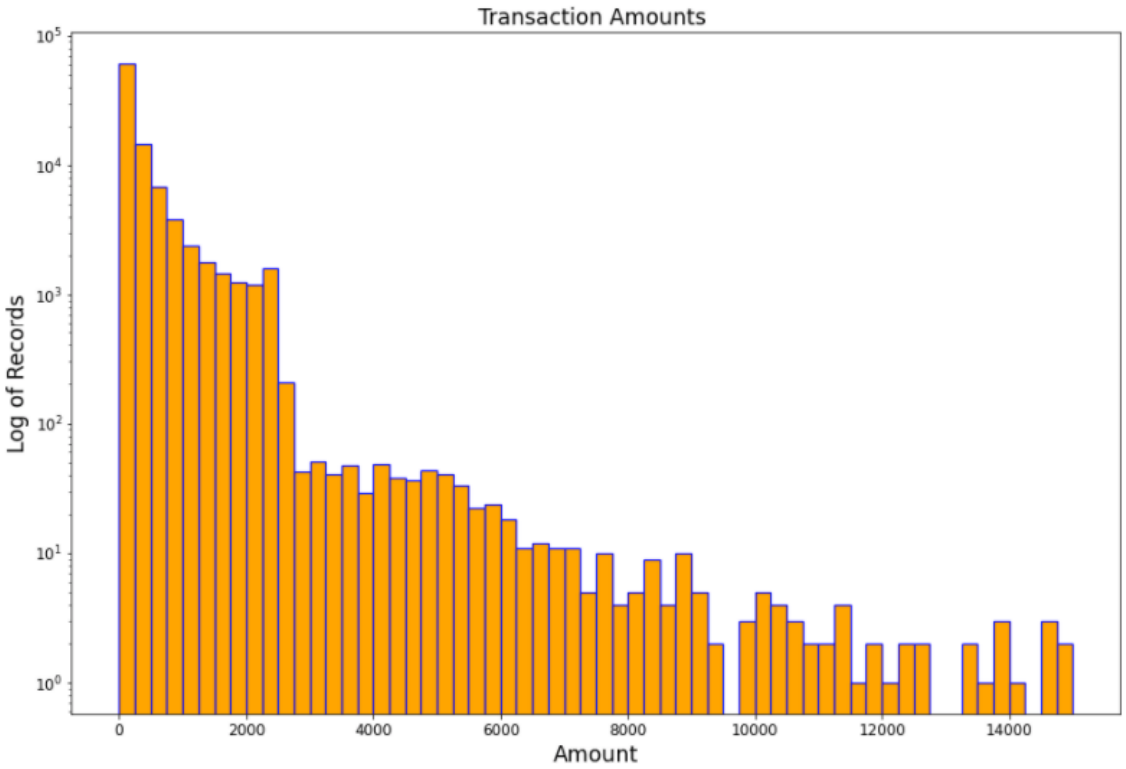
3.8. TRANSTYPE

TRANSTYPE represents which type of a transaction each occurrence belongs to. The bar chart below illustrates the frequency of each 4 types of transactions in a given year.



3.9. AMOUNT

AMOUNT field shows the amount of money that has been spent on each credit card transaction. Y scale shows the log of the records. The range of amount distributed above is up to 15,000 USD.



FRAUD is a dummy variable with two values:

0 : meaning that the transaction record is not

fraudulent 1: meaning that the transaction record is

fraudulent

Below is the distribution of fraudulent and non- fraudulent occurrences on a log scale of records.