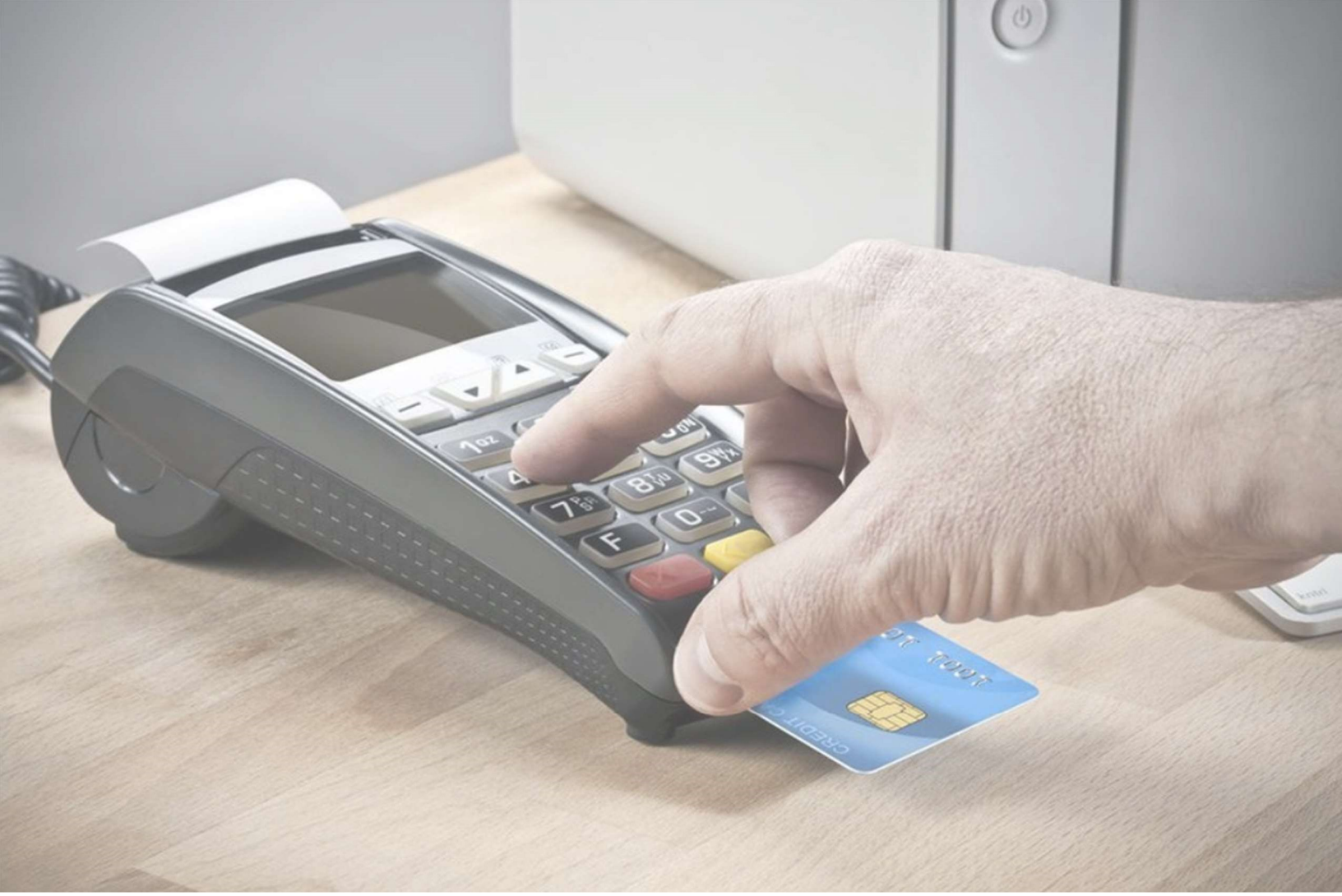


Fraud Detection on Credit Card Transaction Data

DSO562 Fraud Analytics

Team 2:
Bei Wang
Gregy Thomas
Huan-Wei Chang
Xinyu Bao
Yi-Chen Lin
YingHong Lin

2019.03.26

Table of Contents

Part I. Executive Summary

This report detects fraud events in the credit card transaction dataset using supervised machine learning algorithms, including logistic regression, boosted tree, random forest and neural network.

The dataset contains 96,754 records of credit card transaction, where 10 fields are related to credit card and merchant information. After data exploration and data cleaning, we first built 278 candidate variables with different time windows through data manipulation (see Part IX: Appendix). Then we evaluated the importance of each variable using KS distance and fraud detection rate at 3% of records. Variables are ranked and 20 most influential variables were selected through recursive feature elimination method.

By splitting the data into training, testing and out-of-time dataset, we trained models using training dataset and tested the model on the testing dataset and out-of-time dataset using fraud detection rate (FDR). For each algorithm, we figured out the optimal combination of variables and tuned the parameters to achieve the best results.

Overall, the best model is boosted tree with 15 variables selected, a maximum depth of 3, a learning rate of 0.1, gamma of 1 and 500 trees, using which we achieve a fraud detection rate of 49.16% on out-of-time data in the top 3% highest scored records.

	FDR@3%		
Model	**Training**	**Testing**	**OOT**
Logistic Regression	65.51%	65.24%	37.54%
Boosted Tree	99.05%	91.77%	49.16%
Random Forest	97.03%	87.96%	47.07%
Neural Network	61.65%	62.51%	24.58%

Part II. Description of Data

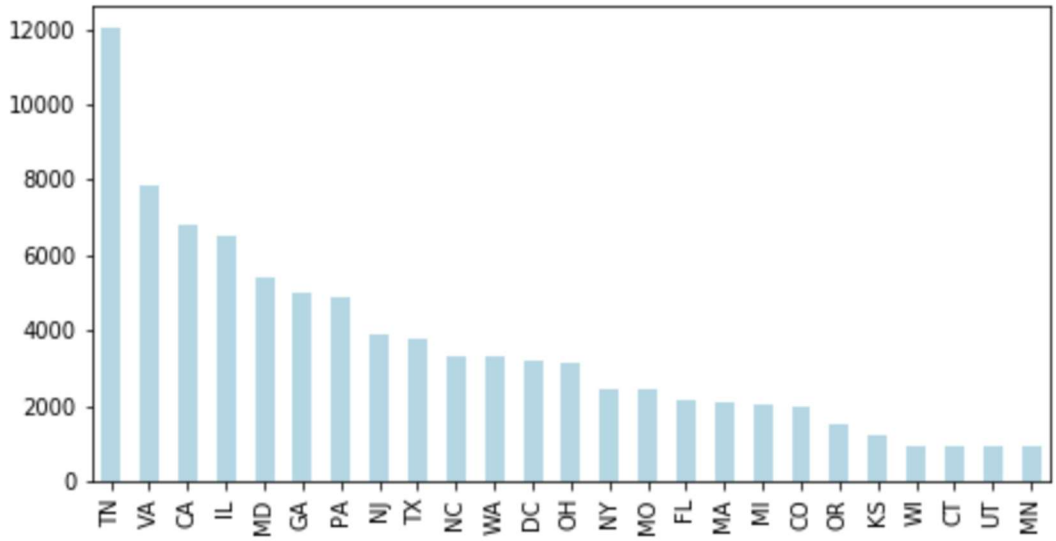
The credit card transaction dataset has the following properties:

- Number of records: 96,753
- Number of fields: 10
- Time period: 2010-01-01 to 2010-12-31

Here is a description of the variables we consider to be the most important.

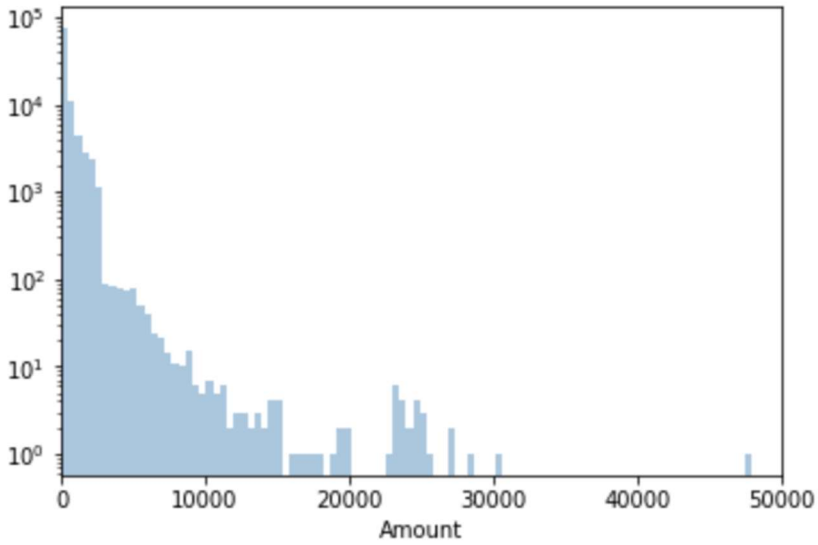
Merch state

This refers to the state in which each merchant is located. Below is a plot shows the count of transactions which happened in each state. This field is 98.76% populated and has 227 unique values. As we can see, a large proportion of transactions happened in TN.



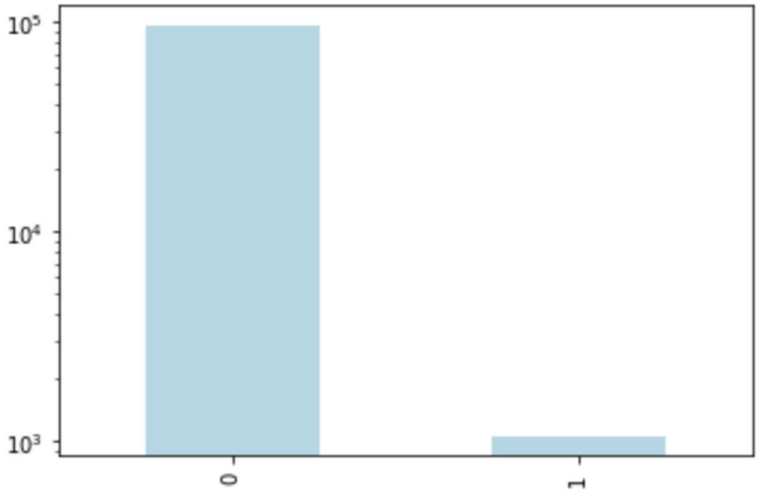
Amount

This field refers to the amount of transaction, which is 100% populated and has 34,909 unique values. The maximum value of transaction amount exceeds 3,000,000, which is quite far away from the center of the data. Thus, in the following plot the distribution of transaction amount below 50,000 is shown on a log scale. The majority of transactions are below 20,000.



Fraud

This field is the response variable, which refers whether the transaction is fraud or not. As we can see, this field is quite imbalanced. There are 95,694 transaction records which are not fraud.



Part III. Data Cleaning

From the data quality analysis, we observed that there was one outlier with extraordinary high value in “Amount”, and there are also many missing values. We removed the outlier first and keep data with transaction type “P”, which is the main transaction type we focused on in this project.

For further analysis, we filled in missing fields with innocuous values that would not change the distribution of variables dramatically across the data, and also would not introduce any anomalies. The approaches we took for data cleaning for each variable are as described below:

Merch state

We first filled in missing values for Merch state among all the variables. Since Merch state is geographical data, we decided to aggregate it by Zip and fill in the most common Merch state for each zip code. After doing this, we filled in most of the missing values in Merch state. However, there are still missing values after aggregate by Zip, which means there are some records with both null values in Merch state and Zip. Consequently, we decided to fill “TN”, which is the most common value of Merch state across the whole dataset, in the rest of data that has null value of Merch state.

Zip

Considering different preferences of merchandise in each zip code region, we aggregated Zip by Merch number and filled in the most frequent Zip for each Merch number. After this, we found some data with both Zip and Merch number missing. We then grouped Zip by Merch description and filled in the most common Zip for each Merch description. Finally, there were still some data with Zip value missing. We decided to aggregate Zip by Merch state and fill in the most common Zip for each Merch state.

Merchnum

Similar to the approach mentioned above, we first grouped Merchnum by Merch description and filled in the most common Merchnum for each Merch description. After this, we aggregated Merchnum by Zip and filled in the most frequent Merchnum for different Zip. Lastly, we grouped Merchnum by Merch state and filled in the most common Merchnum for different Merch state.

Part IV. Candidate Variables Creation

Since this dataset only has limited number of variables on transaction data, we would like to create some expert variables that capture behavioral pattern over a course of time. Thus, we created 278 variables based on transaction amount and frequency over different time windows.

Amount Variables

There are 200 new variables associated with amount. We grouped amount by each objective over different period of time. We focused on summary statistics such as average, maximum, median, total, and ratio on actual over those summary statistics. One example would be the average amount spent by a particular card number at the zip code over the past seven days, which is denoted by Card_Zip_Amount_mean_7D. Considering there might be multiple transactions by the same entity in a given date, we preprocessed the transactions based on total amount spent in one date by grouping the entity (card/merchant/card at this merchant/card in this zip code/card in this state). Then we applied rolling method to calculate summary statistics for each group over the past few intervals (1 day, 3 days, 7 days, 14 days, and 30 days).

Frequency Variables

Using the same concept, we calculated number of transactions by a particular entity over the past several days. We also first grouped transaction by each entity and extracted count. This kind of variable is denoted by entity_Count_day. For example, Merch_Count_1D represents number of transactions with this merchant over the past 1 day.

Days Since Variables

In this category, we calculated how many days it has been since the last transaction for different objectives. We grouped date by Cardnum, Merchnum, same Cardnum at the same Merchnum, same Cardnum at the same Zip, and same Cardnum at the same Merch state. 5 new variables were created through this approach.

Velocity Change Variables

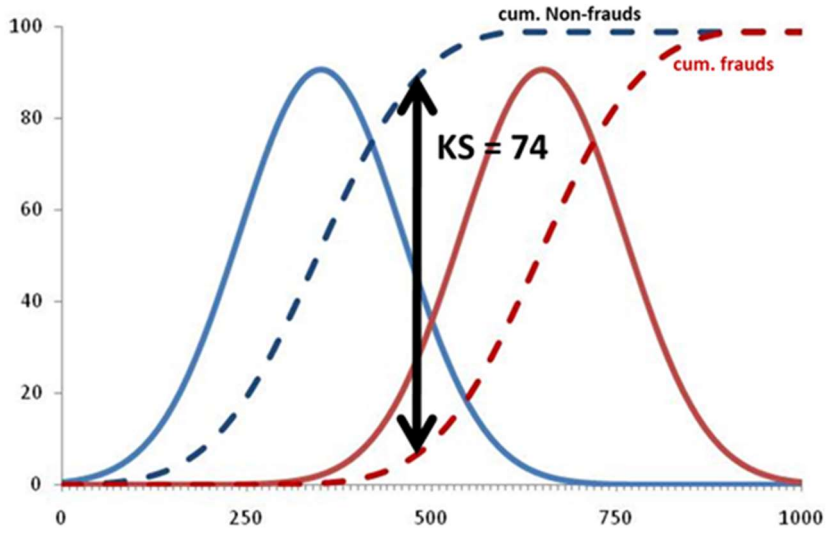
To capture the purchasing velocity of transactions, we looked back on the number and amount of transactions with the same Cardnum or Merchnum over the past one day, and divided that by the average number and amount of transactions with the same Cardnum or Merchnum over the past seven, fourteen, and thirty days. 48 new variables were created through this approach.

Part V. Feature Selection

Step1: KS distance and Fraud Detection Rate (FDR)

After creating 278 candidate variables, we calculated KS distance and Fraud Detection Rate (FDR) at 3% of records for all variables to evaluate the importance of each variable. After calculation, we sorted the variables by the values of both measures and provided rankings for each variable. Then we averaged the rankings of two measures as a final measurement to rank all candidate variables. Finally, we dropped variables of low rankings and kept 111 variables for further feature selection.

-KS Distance
KS distance refers to the maximum of the difference of the cumulative fraud records and non-fraud records. It measures how well the distributions of fraud and non-fraud records are separated. The value of KS distance is always between 0 and 100. The higher KS distance of a variable is, the more likely a variable is a good indicator of detecting frauds. Below is a graph denoting the concept of KS distance.



-Fraud Detection Rate(FDR)
FDR refers to the percentage of fraud records that each variable captures after sorted in descending or ascending order, depending on which value is higher. We calculated FDR at 3% of the population, which is commonly used in business applications. The higher the FDR of a variable is, the more likely a variable is a good indicator of detecting frauds.

Below graph shows the top 20 variables sorted by the averaged ranking of KS distance and FDR @ 3%

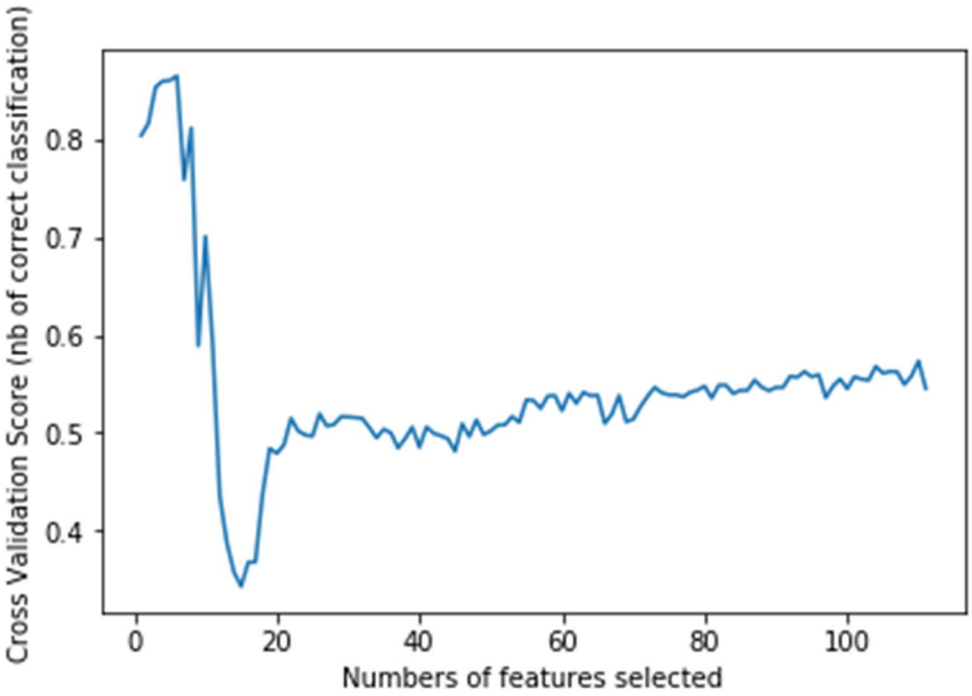
Field	KS	FDR	Rank_ks	Rank_FDR	Average_Rank
Card_Zip_Amount_total_7D	0.6988	0.64318	279	280	279.5
Card_Merch_Amount_total_7D	0.6956	0.63636	276	279	277.5
Card_Merch_Amount_total_14D	0.69367	0.62841	274	277	275.5
Card_Zip_Amount_max_14D	0.69965	0.6125	280	271	275.5
Card_Merch_Amount_max_14D	0.69625	0.6125	277	271	274
Card_Zip_Amount_total_3D	0.69211	0.62727	272	275.5	273.75
Card_Merch_Amount_max_7D	0.69247	0.61932	273	274	273.5
Card_Zip_Amount_total_14D	0.69377	0.6125	275	271	273
Card_Zip_Amount_max_7D	0.69778	0.61023	278	268	273
Card_Merch_Amount_total_3D	0.68608	0.62727	266	275.5	270.75
Card_State_Amount_max_7D	0.67793	0.62955	263	278	270.5
Card_Merch_Amount_max_3D	0.68163	0.61591	265	273	269
Card_State_Amount_total_7D	0.68648	0.61023	267	268	267.5
Card_Merch_Amount_max_30D	0.68986	0.60795	269	266	267.5
Card_Zip_Amount_max_3D	0.68839	0.60682	268	263.5	265.75
Card_State_Amount_total_3D	0.67756	0.61023	262	268	265
Card_State_Amount_max_3D	0.67698	0.60682	261	263.5	262.25
Card_Zip_Amount_max_30D	0.69088	0.6	271	253	262
Card_Merch_Amount_total_1D	0.66499	0.60455	251.5	257.5	254.5

-Step2: Recursive Feature Elimination (RFE)

We kept the top 111 variables based on KS distance & FDR ranking. However, the number of candidate variables is still high, so we conducted recursive feature elimination (RFE) to rank the importance of each variable and calculated the optimal number of features for model building.

We used logistics regression as our fitting estimator to provide information about feature importance. Then we used recursive feature elimination (RFE) to select features by recursively considering smaller and smaller sets of features, and to decide the importance of each feature according to coefficients of logistics regression. After that, the least important features were pruned from current set of features. The procedure was recursively repeated on the pruned set until the desired number of features to select was eventually reached.

The optimal number of features from our RFE result was six. The below graph shows when selecting six features, the percentage of correct classification reaches the highest.



Although the optimal number of feature selection was six, we didn't use it as our final number of feature selection. We used the cross-validation score from RFE as a reference and selected the top 20 variables based on the ranking of score. Below is the list of 20 variables we selected, which would be used for model building in the next section.

1	Card_Amount_Act_median_30D
2	Card_Merch_Amount_max_7D
3	v_num_card_amount_card14
4	v_num_card_amount_card7
5	v_num_merch_amount_card7
6	v_num_merch_amount_merch14
7	Card_Merch_Amount_max_30D
8	Card_State_Amount_max_30D
9	Card_State_Amount_max_14D
10	Card_Amount_mean_30D
11	Card_Zip_Amount_total_3D
12	Card_Zip_Amount_max_3D

13	Card_Zip_Amount_median_3D
14	Card_Zip_Amount_mean_3D
15	Card_Amount_median_14D
16	Card_Merch_Amount_median_7D
17	Card_State_Amount_mean_30D
18	Card_State_Amount_mean_30D
19	v_amount_card_num_card7
20	Card_Merch_Amount_mean_14D

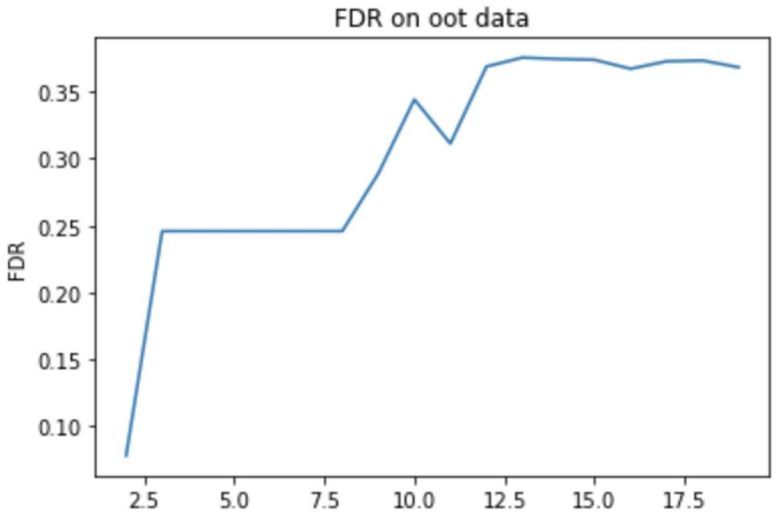
Part VI. Fraud Detection Algorithms

In this part, we performed four supervised machine learning algorithms. We chose logistic regression as a baseline model, and then built boosted tree, random forest and neural network models in order to achieve a better result. Before feeding data to models, we used z-scale to rescale the data so that the different units of variables would not affect the model performance.

Model 1: Logistic regression

Logistic regression is often used to explore the relationship between one dependent binary variable and independent variables. As a baseline model, logistic regression uses maximum likelihood to fit the model and gives the probability of each data record being fraud event. After fitting the model, it would classify each data record into 'fraud' or 'normal' category by estimating the probability based on the independent variables.

To reduce dimensionality and reduce the model variance, variables are added sequentially to the model according to their rankings from recursive feature elimination. To evaluate the performance, we calculated the FDR@3% on out of time data every time. Higher FDR@3% means that a higher proportion of fraud events can be caught in the top 3% data records with high probability of being fraudulent.



FDR@3% for logistic regression

Train	Test	OOT
65.5%	65.2%	37.5%

This way we selected the optimal number of variables which succeeded in achieving the highest FDR@3%. Below is a graph summarize the selected 13 variables.

1	Card_Amount_Act_median_30D
2	Card_Merch_Amount_max_7D
3	v_num_card_amount_card14
4	v_num_card_amount_card7
5	v_num_merch_amount_card7
6	v_num_merch_amount_merch14
7	Card_Merch_Amount_max_30D
8	Card_State_Amount_max_30D
9	Card_State_Amount_max_14D
10	Card_Amount_mean_30D
11	Card_Zip_Amount_total_3D
12	Card_Zip_Amount_max_3D
13	Card_Zip_Amount_median_3D

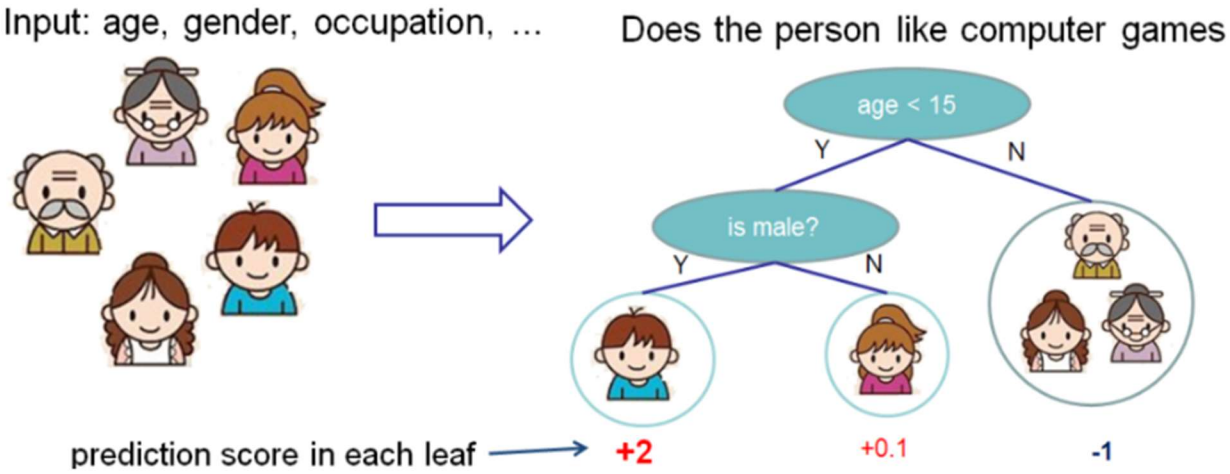
Using logistic regression, we got FDR@3% of 65.5% for the training dataset, 65.2% for the testing dataset and 37.5% for the out of time dataset. As for the compute area under the receiver operating characteristic curve, this model achieves 59.19% AUC score.

Next, we would try to beat the baseline model with three different algorithm - boosted tree, random forest and neural nets.

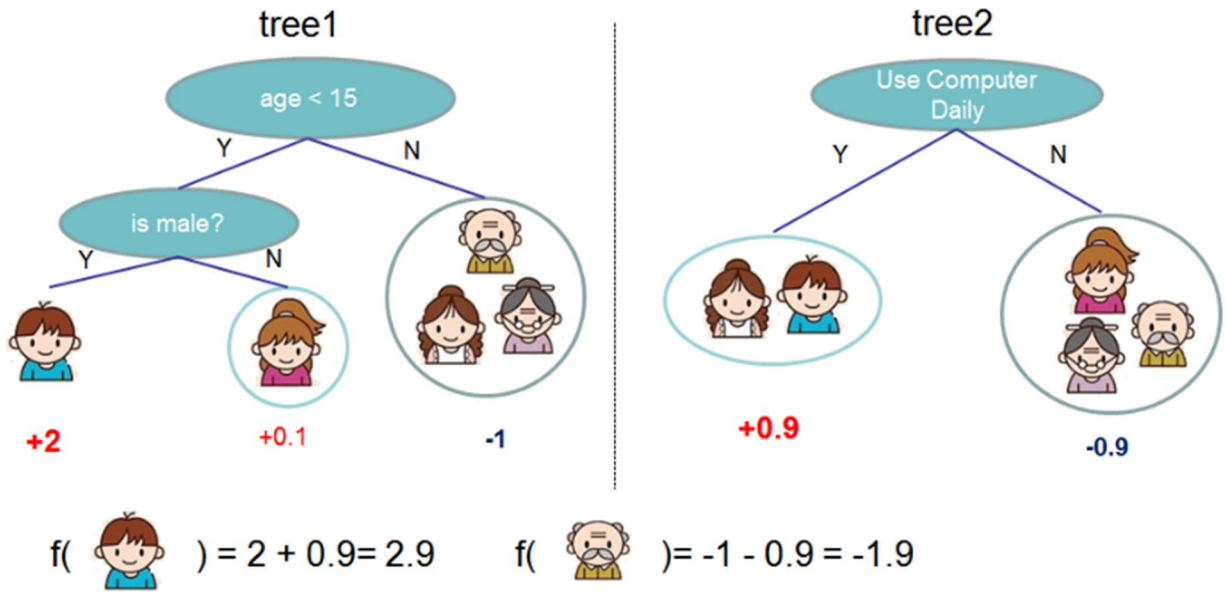
Model 2: Boosted Tree

Boosted tree is a classification technique to construct a strong learner by combining several weak learners. Usually one single tree is not comprehensive enough to capture all information, so we turned to an ensemble model. The following is a simple illustration of a boosted tree.

If one wants to predict whether an individual like computer games with input such as age, gender, occupation and other, a CART model would divide the observations into groups based on age and gender criteria. Then the observation would be assigned to a score based on which leaf or node it falls into. For example, the son with age less than 15 and gender as male would have a score of 2 based on the model.



On the other hand, there would be two trees under a tree ensemble model. While the tree1 is the same as a simple CART model, tree2 contains more information by incorporating whether the individual uses computer daily. Then an observation would be assigned a score as the weighted average of scores from both trees. To illustrate, the son gets 2 points from tree1 and 0.9 from tree2, so the final score would be 2.9.



We built boosted tree models using the XGboost package in python. Other than the selected parameters, all other parameters are left as default.

Below is a summary of the user-defined parameters:

max_depth [3,5,10]

Maximum depth of a tree, and a larger number would make the model more complex and overfit. Typical value range from 3 to 10.

learning _rate [0.01,0.05,0.1]

The weight of new features on each step to prevent overfitting. A high learning rate keeps the model more conservative.

n_estimators [100,500,1000]

Number of boosted trees to fit. Usually 1000 would be for a small dataset.

subsample=0.8

Subsample ratio of the training instance.0.8 means the model only extract 80% of the observation randomly for each tree. Lower value makes the model more conservative and thus control overfitting. Typical values range from 0.5 to 1.

gamma=[0,1,5]

Regularization term. Minimum loss reduction required for further partition. Regularization increases as gamma increases.

We tried different combination of parameters along with different number of variables [6,10,15,20] and recorded FDR at 3% for training, testing, and out of time data for 5 randomly selected training and testing sets. Below is a chart summarizing our final parameter combination.

Final Parameter Combination

max_depth	3
learning_rate	0.1
n_estimators	500
gamma	1
number of variables	15

As we can see, the boosted tree model achieved a 99.05% FDR@3% on training dataset, 91.77% FDR@3% on testing dataset and 49.16% FDR@3% on out of time dataset.

FDR@3% for boosted tree

Train	Test	OOT
99.05%	91.77%	49.16%

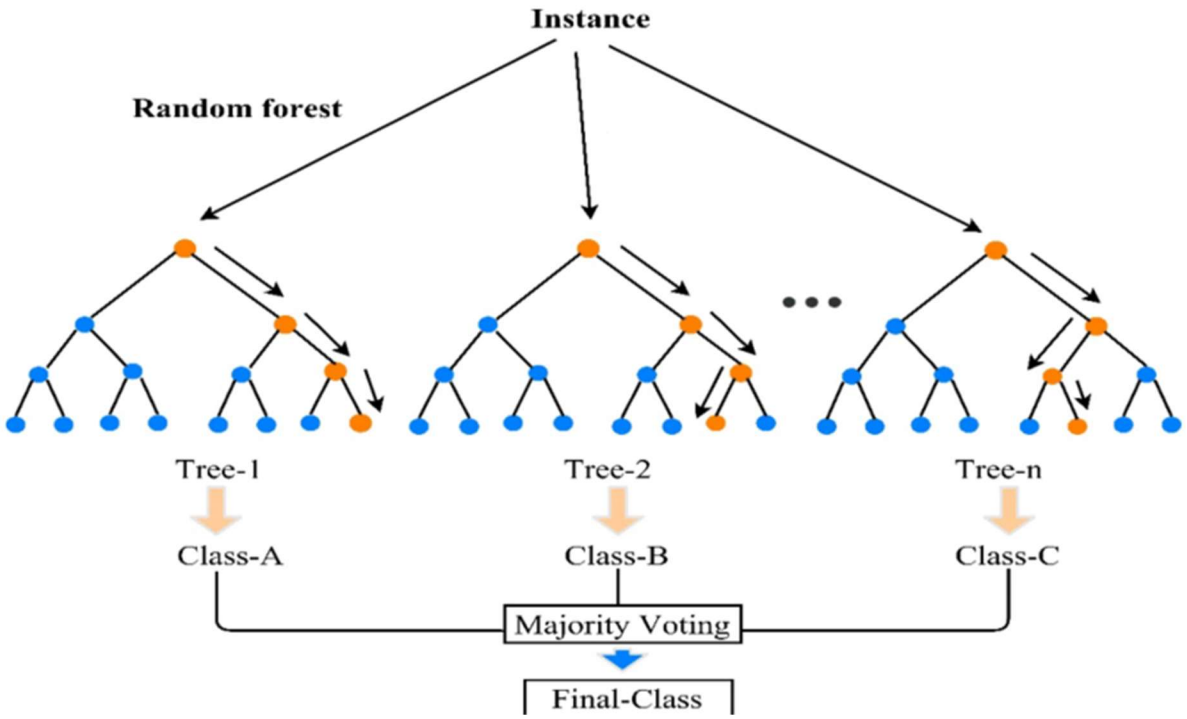
Model 3: Random Forest

Random forests algorithm is an ensemble classifier and works as a large collection of decorrelated decision trees. It combines the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy. The final class is the mode of the class's output by individual trees.

Each decision tree can be created using the following steps:

Step 1: Create a bootstrapped dataset. To create a bootstrapped dataset that is the same size as the original dataset, we just randomly select samples from the original dataset. The important detail is that we can pick the same record more than once.
Step 2: Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each level.
Step 3: Go back to step 1 and repeat and we get a lot of trees in the end.

Notes: Step 1 and step 2 result in a wide variety of trees. The variety is what makes random forests more effective than individual decision trees.



We shall get a random forest after combining all individual decision trees created from step 1 to step 3. We chose this algorithm since random forest has some advantages over other unsupervised machine learning algorithms. For example, it can help avoid overfitting by averaging several trees; the randomness will make the model robust, not suffering too much noise; plus, it does not require using standardized data and there are not too many parameters need to be tuned.

We tried different combination of parameters along with different number of variables and evaluated their performance based on FDR@3% on out of time data for 20 randomly selected training and testing sets. Below is a chart summarizing our final parameter combination.

Final Parameter Combination

n_estimators	200
max_depth	13
min_sample_split	5
number of variables	15

Overall the random forests have a good performance over most of our other algorithms, which achieves a 47.07% of FDR@3% on the out of time data.

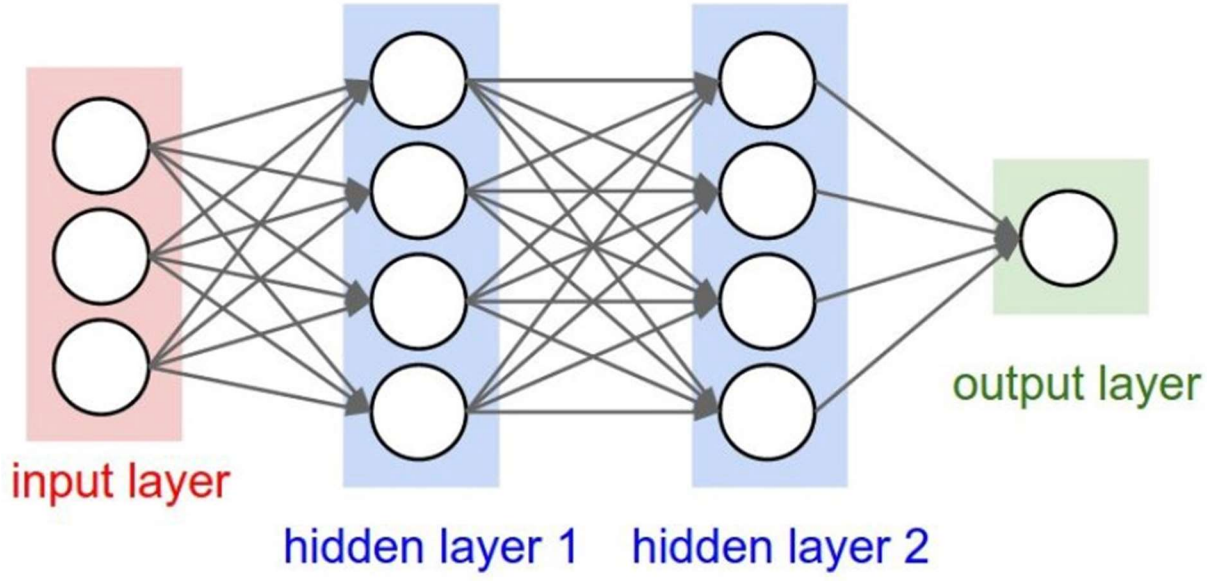
FDR@3% for random forest

Train	Test	OOT
97.03%	87.96%	47.07%

Model 4: Neural Network

An Artificial Neural Network (ANN or NN) consists of an interconnected group of artificial neurons. The principle of neural network is motivated by the functions of the brain especially *pattern recognition* and *associative memory*. The neural network recognizes similar patterns, predicts future values or events based upon the associative memory of the patterns it was learned. It is widely applied in classification and clustering. The advantages of neural networks over other techniques are that these models are able to learn from the past and thus, improve results as time passes. They can also extract rules and predict future activity based on the current situation. By employing neural networks, effectively, banks can detect fraudulent use of a card, faster and more efficiently.

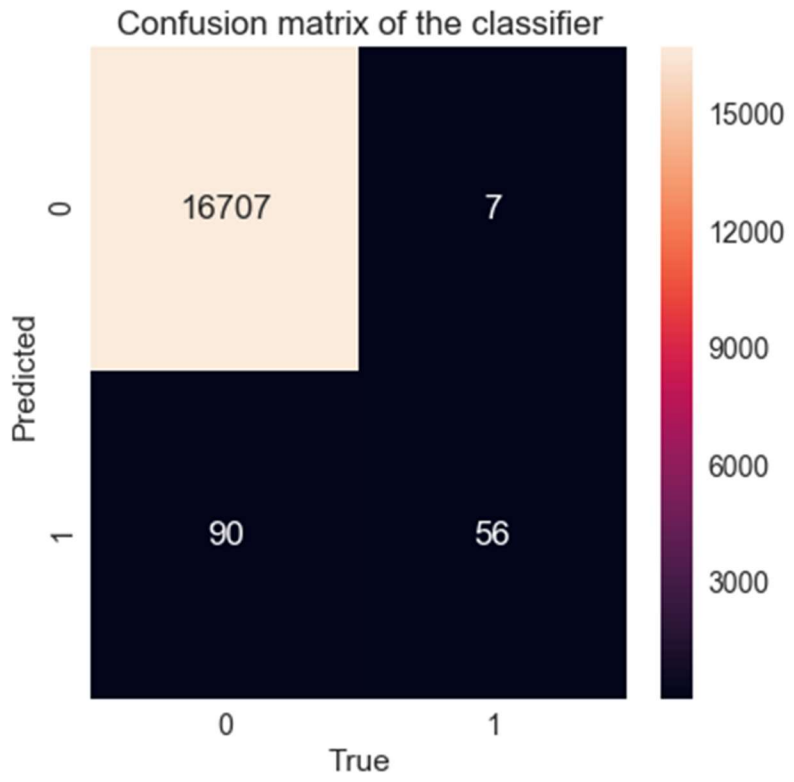
Most of the reported credit card fraud studies has focused on using neural networks. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. There are two phases in neural network: training and recognition. NNs can produce best result for only large transaction dataset. And they need a long training dataset



As shown in the schematic diagram above, the multi-layer perceptron (MLP) neural network was that we employed in our analysis consists of 20 input nodes, two densely connected hidden layer with 12 nodes with S-shaped rectified linear activation and 20% dropout, and 1 output node with a sigmoid activation. The models were fit using batches of 32 observations for up to 10 epochs although validation loss (binary cross entropy) is monitored to permit early stopping. Stochastic optimization was performed using Adam. The final model was an ensemble of the stratified k-fold neural networks, constructed by averaging the model predictions. The models were implemented in Python using Keras and TensorFlow as the backend,

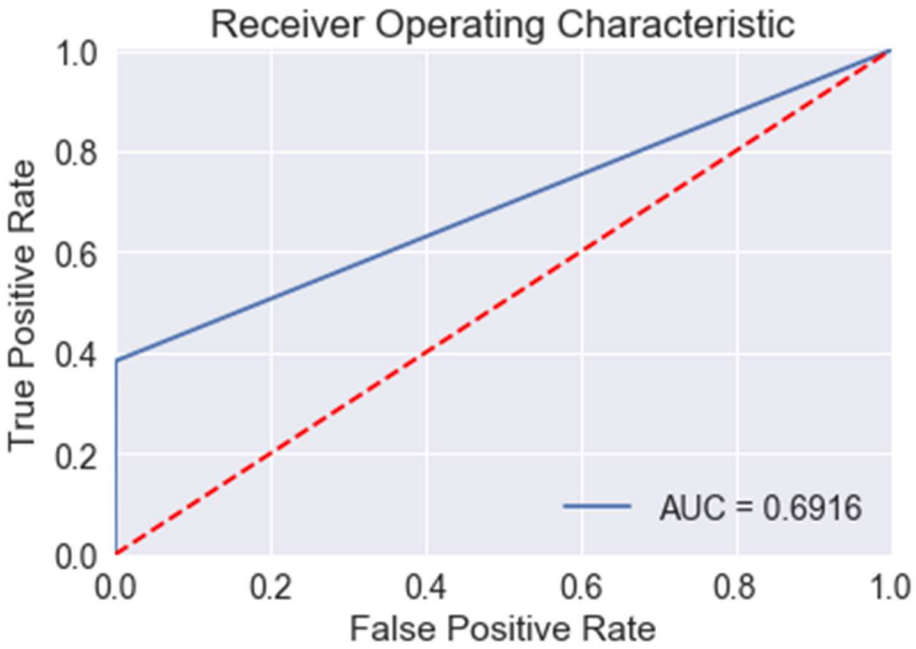
The prediction made on the test data after the training of model yielded the following confusion matrix.

```
Test Data Accuracy: 0.9942
```



	precision	recall	f1-score	support
0	0.99	1.00	1.00	16714
1	0.89	0.38	0.54	146
avg / total	0.99	0.99	0.99	16860

The final model achieved an overall f1 score of 0.99 with 38% sensitivity (recall) and 89% precision for the positive class. That is, the model correctly identified 38% of the fraud cases (true positives) but about 90% of the transactions predicted as fraudulent were actually fraudulent, leaving the rest 10% as false classification on fraud. In other words, even though the model caught only 35% of the fraudulent cases — it identified more cases of fraud with a high precision.



The ROC curve plotted the true positive rate versus the false positive rate over different threshold values. Basically, we wanted the blue line to be as close as possible to the upper left corner. While the ROC results could have been better, we had to keep in mind of the imbalanced or skewed nature of our dataset.

Upon doing the FDR@3% along with varying the parameters, we achieved the following metrics for NN:

FDR@3% for neural network

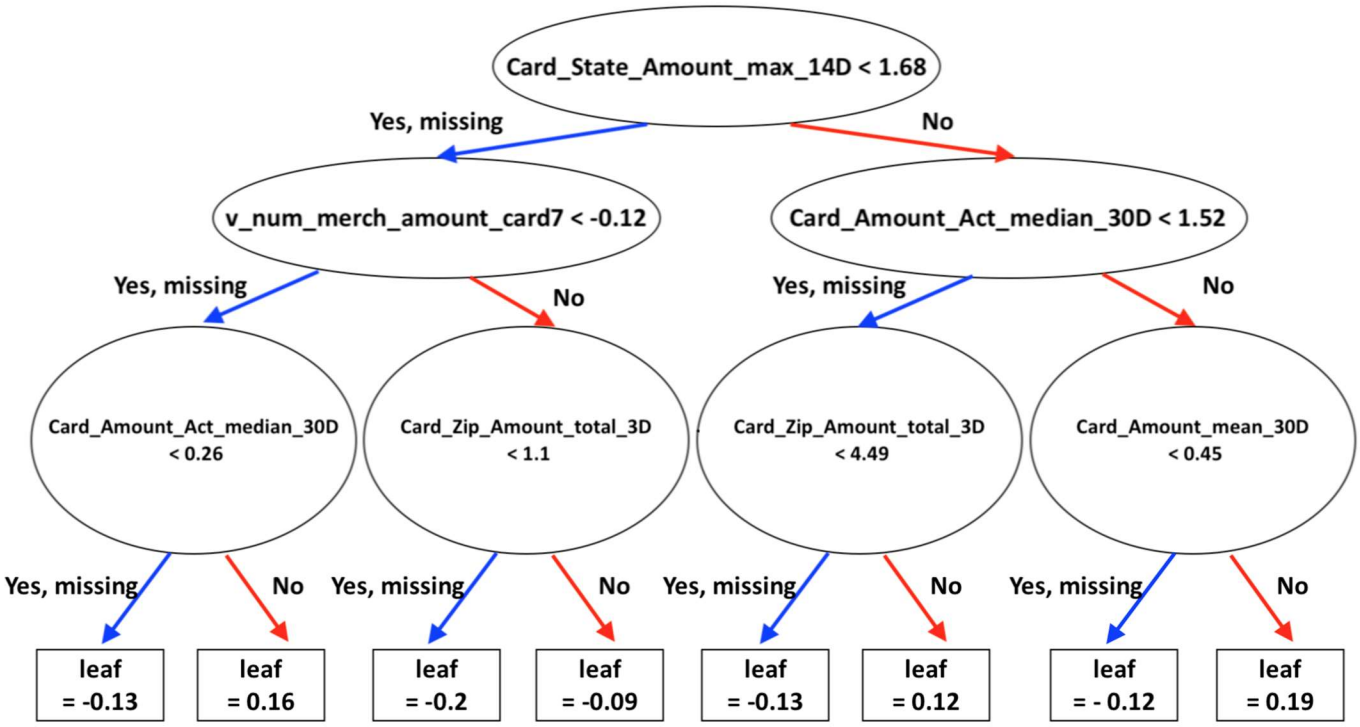
Train	Test	OOT
0.616476	0.625187	0.24581

Part VII. Results

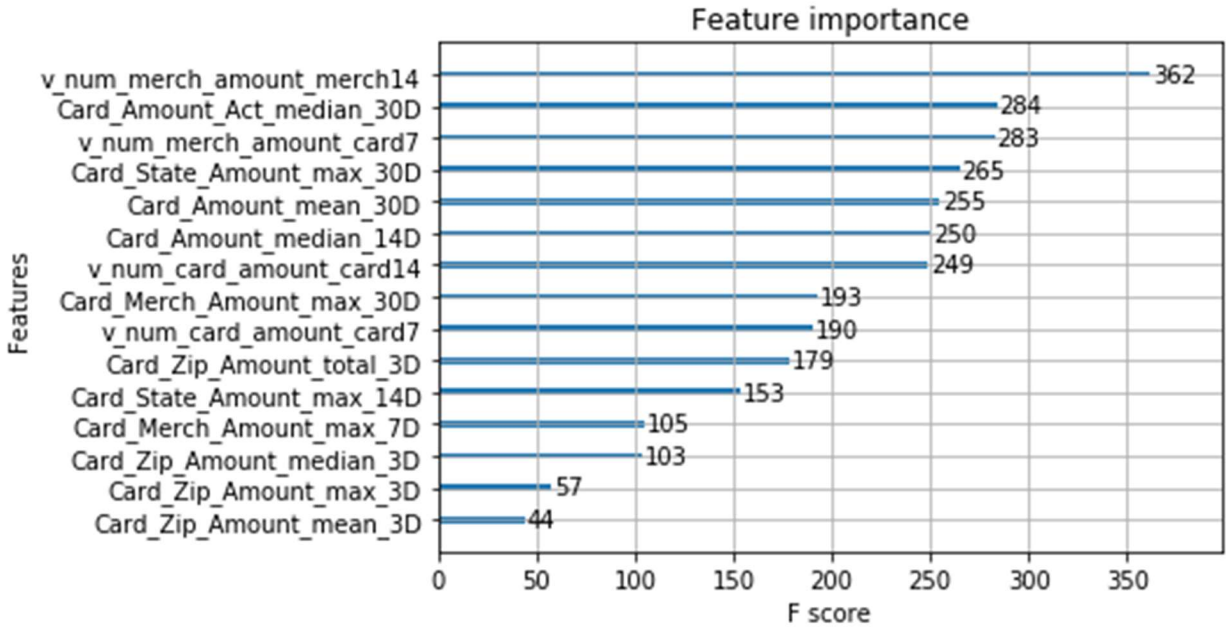
Our final result was based on random sampling with 20% testing data with a random state of 0. The final boosted tree model (with parameters in max_depth=3, learning_rate=0.1, n_estimators=500, gamma=1) contained the following 15 variables:

Variable Name	Description
Card_Amount_Act_median_30D	Actual over median amount spent by this card over the past 30 days
Card_Merch_Amount_max_7D	Maximum amount spent by this card at this merchant over the past 7 days
v_num_card_amount_card14	Amount spent with the same merchant over average daily amount by the same card in the past 14 days
v_num_card_amount_card7	Amount spent with the same merchant over average daily amount by the same card in the past 7 days
v_num_merch_amount_card7	Number of transactions with the same merchant over average daily amount by the same card in the past 7 days
v_num_merch_amount_merch14	Number of transactions with the same merchant over average daily amount with the same merchant in the past 14 days
Card_Merch_Amount_max_30D	Maximum amount spent by this card at this merchant over the past 30 days
Card_State_Amount_max_30D	Maximum amount spent by this card at this state over the past 30 days
Card_State_Amount_max_14D	Maximum amount spent by this card at this state over the past 14 days
Card_Amount_mean_30D	Amount spent by this card over the past 30 days
Card_Zip_Amount_total_3D	Total amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_max_3D	Maximum amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_median_3D	Median amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_mean_3D	Mean amount spent by this card at this zip code over the past 3 days
Card_Amount_median_14D	Amount spent by this card over the past 14 days

The following diagram illustrated one of 500 trees:



The following graph illustrated the importance of each variable. We can see that v_num_merch_amount_merch14 was the most important variable.



Below are three tables containing information on the first 20 bins of training, testing, and out of time dataset to illustrate our model performance. The ultimate goal for constructing a model is to catch as many as fraudulent activities with a small fraction of sample.

This table contains two parts: bin statistics and cumulative statistics. Bin statistics outline how many additional fraudulent instances each bin contribute. On the other hand, cumulative statistics add up all bin statistics from the previous bins so that users can decide which bin to cutoff.

The higher the percentage of bads and cumulative bads, the more efficient our model is at capturing fraud transactions. Usually the top bins would capture the most fraud instance and the number of fraud record would decrease when we increase number of bins.

Summary statistics on training set:

Training	# Records			# Goods		# Bads			Fraud Rate		
	67,439			66,707		734			0.011		
	Bin Statistics					Cumulative Statistics					
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	675	5	670	0.74	99.26	675	5	670	0.01	91.28	0.91	0.01
2	674	622	52	92.28	7.72	1349	627	722	0.94	98.37	0.97	0.87
3	675	672	3	99.56	0.44	2024	1299	725	1.95	98.77	0.97	1.79
4	674	672	2	99.70	0.30	2698	1971	727	2.95	99.05	0.96	2.71
5	674	671	3	99.55	0.45	3372	2642	730	3.96	99.46	0.95	3.62
6	675	672	3	99.56	0.44	4047	3314	733	4.97	99.86	0.95	4.52
7	674	674	0	100.00	0.00	4721	3988	733	5.98	99.86	0.94	5.44
8	675	674	1	99.85	0.15	5396	4662	734	6.99	100.00	0.93	6.35
9	674	674	0	100.00	0.00	6070	5336	734	8.00	100.00	0.92	7.27
10	674	674	0	100.00	0.00	6744	6010	734	9.01	100.00	0.91	8.19
11	675	675	0	100.00	0.00	7419	6685	734	10.02	100.00	0.90	9.11
12	674	674	0	100.00	0.00	8093	7359	734	11.03	100.00	0.89	10.03
13	674	674	0	100.00	0.00	8767	8033	734	12.04	100.00	0.88	10.94
14	675	675	0	100.00	0.00	9442	8708	734	13.05	100.00	0.87	11.86
15	674	674	0	100.00	0.00	10116	9382	734	14.06	100.00	0.86	12.78
16	675	675	0	100.00	0.00	10791	10057	734	15.08	100.00	0.85	13.70
17	674	674	0	100.00	0.00	11465	10731	734	16.09	100.00	0.84	14.62
18	674	674	0	100.00	0.00	12139	11405	734	17.10	100.00	0.83	15.54
19	675	675	0	100.00	0.00	12814	12080	734	18.11	100.00	0.82	16.46
20	674	674	0	100.00	0.00	13488	12754	734	19.12	100.00	0.81	17.38

Summary statistics on testing set:

Testing	# Records	# Goods	# Bads	Fraud Rate
	16,860	16,714	146	0.009

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	169	46	123	27.22	72.78	169	46	123	0.28	84.25	0.84	0.37
2	169	160	9	94.67	5.33	338	206	132	1.23	90.41	0.89	1.56
3	168	163	5	97.02	2.98	506	369	137	2.21	93.84	0.92	2.69
4	169	169	0	100.00	0.00	675	538	137	3.22	93.84	0.91	3.93
5	168	166	2	98.81	1.19	843	704	139	4.21	95.21	0.91	5.06
6	169	168	1	99.41	0.59	1012	872	140	5.22	95.89	0.91	6.23
7	169	169	0	100.00	0.00	1181	1041	140	6.23	95.89	0.90	7.44
8	168	168	0	100.00	0.00	1349	1209	140	7.23	95.89	0.89	8.64
9	169	168	1	99.41	0.59	1518	1377	141	8.24	96.58	0.88	9.77
10	168	168	0	100.00	0.00	1686	1545	141	9.24	96.58	0.87	10.96
11	169	169	0	100.00	0.00	1855	1714	141	10.25	96.58	0.86	12.16
12	169	169	0	100.00	0.00	2024	1883	141	11.27	96.58	0.85	13.35
13	168	167	1	99.40	0.60	2192	2050	142	12.27	97.26	0.85	14.44
14	169	169	0	100.00	0.00	2361	2219	142	13.28	97.26	0.84	15.63
15	168	167	1	99.40	0.60	2529	2386	143	14.28	97.95	0.84	16.69
16	169	169	0	100.00	0.00	2698	2555	143	15.29	97.95	0.83	17.87
17	169	169	0	100.00	0.00	2867	2724	143	16.30	97.95	0.82	19.05
18	168	168	0	100.00	0.00	3035	2892	143	17.30	97.95	0.81	20.22
19	169	169	0	100.00	0.00	3204	3061	143	18.31	97.95	0.80	21.41
20	168	167	1	99.40	0.60	3372	3228	144	19.31	98.63	0.79	22.42

Summary statistics on out-of-time set:

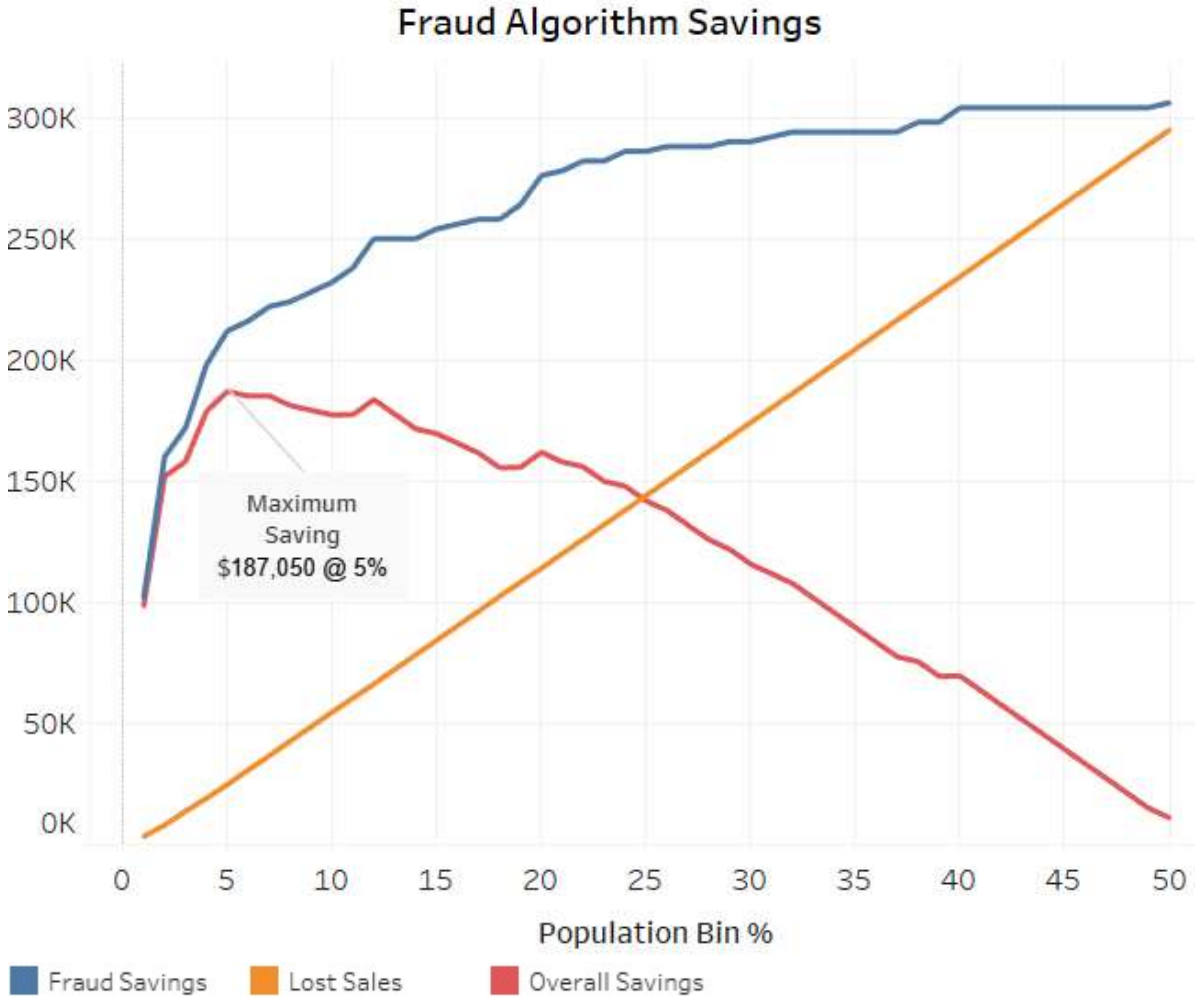
Out Of Time	# Records	# Goods	# Bads	Fraud Rate
	12,097	11,918	179	0.0147

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	121	70	51	57.85	42.15	121	70	51	0.59	28.49	0.28	1.37
2	121	92	29	76.03	23.97	242	162	80	1.36	44.69	0.43	2.03
3	121	115	6	95.04	4.96	363	277	86	2.32	48.04	0.46	3.22
4	121	108	13	89.26	10.74	484	385	99	3.23	55.31	0.52	3.89
5	121	114	7	94.21	5.79	605	499	106	4.19	59.22	0.55	4.71
6	121	119	2	98.35	1.65	726	618	108	5.19	60.34	0.55	5.72
7	121	118	3	97.52	2.48	847	736	111	6.18	62.01	0.56	6.63
8	121	120	1	99.17	0.83	968	856	112	7.18	62.57	0.55	7.64
9	121	119	2	98.35	1.65	1089	975	114	8.18	63.69	0.56	8.55
10	121	119	2	98.35	1.65	1210	1094	116	9.18	64.80	0.56	9.43
11	121	118	3	97.52	2.48	1331	1212	119	10.17	66.48	0.56	10.18
12	121	115	6	95.04	4.96	1452	1327	125	11.13	69.83	0.59	10.62
13	121	121	0	100.00	0.00	1573	1448	125	12.15	69.83	0.58	11.58
14	121	121	0	100.00	0.00	1694	1569	125	13.17	69.83	0.57	12.55
15	121	119	2	98.35	1.65	1815	1688	127	14.16	70.95	0.57	13.29
16	121	120	1	99.17	0.83	1936	1808	128	15.17	71.51	0.56	14.13
17	121	120	1	99.17	0.83	2057	1928	129	16.18	72.07	0.56	14.95
18	121	121	0	100.00	0.00	2178	2049	129	17.19	72.07	0.55	15.88
19	121	118	3	97.52	2.48	2299	2167	132	18.18	73.74	0.56	16.42
20	121	115	6	95.04	4.96	2420	2282	138	19.15	77.10	0.58	16.54

Our boosted tree has excellent performance on training and testing set by capturing over 90% of fraud records within 2% bins. In terms of out of time set, FDR score reaches to nearly 50% at first three bins.

Besides bin statistics and FDR at 3%, we are also interested in the business implication on our boosted tree. Assume we earin $2000 for every fraud we catch and incur $50 loss for false positive transaction. Ideally, we want to catch as many as fraud but not wrongly classify transaction as fraud as customers might be frustrated.

Below is a graph summarizing the performance of fraud savings, lost sales and overall saving. Similar to how we obtain records of goods and bads in the previous tables, we apply the same code to get number of fraud and non-fraud on out of time data set. Fraud savings is calculated by multiplying 2000 by number of fraud records from current and all previous bins. Lost sales is calculated by multiplying 50 by number of non-fraud records from current and all previous bins. While overall saving is calculated by the difference between fraud savings and lost sales.



We can see that at fraud savings and overall savings both have a positive slope. This can be explained by the increase number of both kind of transactions along with increasing number of bins. However, fraud saving increases at a rapid rate in the early stage and slowly increase at a steady pace because there are less fraudulent transactions as predicted probability decreases. On the other hand, lost sales increase at a constant rate due to the fact that the majority of the sample are legitimate. Under the combination effect of fraud savings and lost sales, overall savings peak at first then decreases at a negative trend when lost sales outweigh fraud savings. Maximum saving is at 5% of the sample with an overall savings of $187,050.

Part VIII. Conclusions

After data quality analysis, we removed the outlier and only kept the data with transaction type "P". Then we filled in missing values with innocuous values after aggregating other fields that are related to the missing fields. This is to avoid dramatically changing the value and distribution of the data.

We leveraged domain experts for creating 278 candidate variables. The majority of the newly created variables was associated with amount on different entities over different time windows. For example, one variable measured maximum amount spent by a card on a particular zip code over the past 7 days. Another similar variable was based on frequency, which captures the times certain card visited at a particular entity over different time windows.

These variables were ranged based on importance from recursive feature elimination. We tested four machine learning algorithms with the top 20 variables we obtained from feature selection step. During the tuning stage, we tested each model by inputting different number of variables and tried different combination of parameters. Based on the performance of FDR@3%, we selected the models with optimal number of variables. For example, we selected the top 13 variables for the logistic regression model and got FDR@3% of 37.5% for the out of time dataset. We treated logistic regression model as the baseline model and ran three other non-linear algorithm--boosted tree, random forest, neural network--for better performance.

For boosted tree, we figured that parameters like number of trees and depth of a tree affected model performance, so we tried different combination. Even though more

complicated trees yielded better performance, we went with a smaller tree for model simplicity. For instance, FDR at 3% did not vary much for max_depth of 3 compared to 5 and n_estimators of 500 and 1000, so we selected the former. At the same time, we tried using 6, 10, 15, and 20 variables.

For random forest, we also tried different combinations of parameters to get a better result. As we increased the depth of the tree, we found the model performance improved a lot. However, the problem of overfitting surfaced at the same time. Therefore, we adjusted this parameter carefully while testing the model.

For neural network, different parameter combinations were used to run multiple runs in order to get best results. The parameters that were mainly controlled were the number of variables, the random samples, and the number of neurons of the hidden layers. However the best FDR at 3% for the OOT data was about 29%. This can be mainly attributed to the compromised performance of NN with highly skewed data and the lack of generalization. This can be addressed by reducing the batch size, but this can demand higher computation power and potential overfitting.

After running several machine learning models and compare FDR at 3%, we finalized on boosted tree with max_depth=3, learning_rate=0.1, n_estimators=500, and gamma=1. This model achieved the highest FDR of 49% on out of time data set. When assuming a gain of $2000 for each fraud transaction we catch successfully and a loss of $50 for a false positive incidence, we concluded from the fraud saving plot that at a cutoff of 5%, we will achieve the optimal saving of $187,050.

If we have more time, we would like to explore more about each machine learning model, documentation of libraries, and parameters. In order to tune the model for best result, we need a solid understanding of the mechanism behind each model and how each parameter affects the result. Surprisingly, there are a lot more parameters than the most common or well-known ones. For example, min_child_weight in xgboost. Grid Search is a helpful package to help use tune the model, but we do need to input a set of parameters, which requires extensive research.

Part IX. Appendix

Data Quality Report

Overview

This Data Quality report examines the data available on the card transaction data. Exhaustive data analysis has been done to perform quality checks on the data. The purpose of this data analysis is to expand the cleaned data to build a supervised fraud model on the card transaction data as part of DSO 562 Project 2. The data is collected from the Department of Finance in New York City government. The report is divided into three parts:

- Data Description
- Summary Statistics of fields
- Exploratory Data Analysis of individual fields

I) Data Description

The data collected has the following properties:

- Number of records : 96753
- Number of fields : 10
- Time period : 2010-01-01 to 2010-12-31

II) Summary Statistics of fields

Below table is a snapshot of the basic summary statics of all the fields in their order of appearance in the card transactions data:

Field Name	Data Type	Field Type	# of Records with Value	% Populated	# Unique Values	# Records with Value Zero	Mean	SD	Min	Max	Most Common Field
Recnum	int64	Categorical	96753	100	96753	0	-	-	-	-	2047
Cardnum	int64	Categorical	96753	100	1645	0	-	-	-	-	5142148452
Date	datetime64[ns]	Time	96753	100	365	0	-	-	-	-	2/28/2010 0:00
Merchnum	object	Categorical	93378	96.51	13091	0	-	-	-	-	9.3009E+11
Merch description	object	Text	96753	100	13126	0	-	-	-	-	GSA-FSS-ADV
Merch state	object	Categorical	95558	98.76	227	0	-	-	-	-	TN
Merch zip	float64	Categorical	92097	95.19	4567	0	-	-	-	-	38118
Transtype	object	Categorical	96753	100	4	0	-	-	-	-	P
Amount	float64	Numeric	96753	100	34909	0	427.89	10006.09	0.01	3E+06	-
Fraud	int64	Categorical	96753	100	2	95694	-	-	-	-	0

Numerical Variable:

Field Name	Data Type	Field Type	Mean	SD	Min	Max
Amount	float64	Numeric	427.89	10006.1	0.01	3102045.53

Categorical Variables:

Field Name	Data Type	Field Type	Most Common Field
Recnum	int64	Categorical	2047
Cardnum	int64	Categorical	5142148452
Date	datetime64[ns]	Time	2/28/2010 0:00
Merchnum	object	Categorical	9.3009E+11
Merch description	object	Text	GSA-FSS-ADV
Merch state	object	Categorical	TN
Merch zip	float64	Categorical	38118
Transtype	object	Categorical	P
Fraud	int64	Categorical	0

III) Exploratory Data Analysis of individual fields

1) Recum

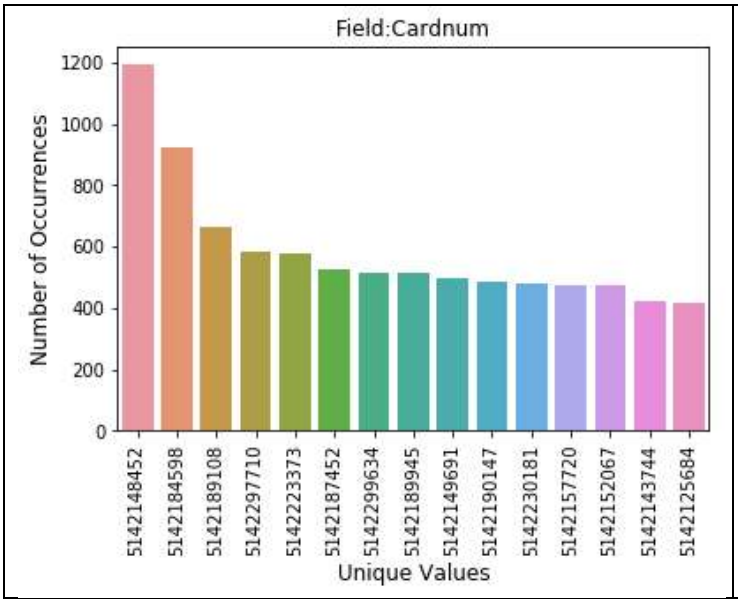
Recnum serves as a unique identification code ‘Record Number’ for every row.

```
Column Name                                =  Recnum
Data type                                  =  int64
Number of Missing Values                   =  0
Percentage of rows populated               =  100.00%
Number of Unique values                    =  96753
Percentage of Unique values in population =  100.00%
```

2) Cradnum

Cardnum serves as a unique identification code for each card. One cardnum can have many transactions. Card number 5142148452 is the most frequent card number, which appears almost 1,200 times.

```
Column Name                                =  Cardnum
Data type                                  =  int64
Number of Missing Values                   =  0
Percentage of rows populated               =  100.00%
Number of Unique values                    =  1645
Percentage of Unique values in population =  1.70%
```



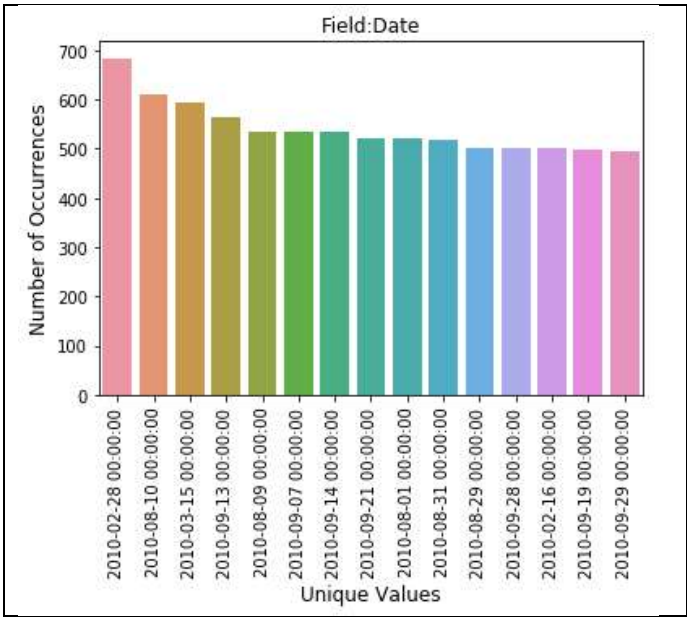
	Value	Count	Percent(%)
0	5142148452	1192	1.23
1	5142184598	921	0.95
2	5142189108	663	0.69
3	5142297710	583	0.60
4	5142223373	579	0.60
5	5142187452	526	0.54
6	5142299634	515	0.53
7	5142189945	512	0.53
8	5142149691	497	0.51
9	5142190147	488	0.50
10	5142230181	479	0.50
11	5142157720	475	0.49
12	5142152067	473	0.49
13	5142143744	422	0.44
14	5142125684	415	0.43

3) Date

Date is the date when a transaction occurs.

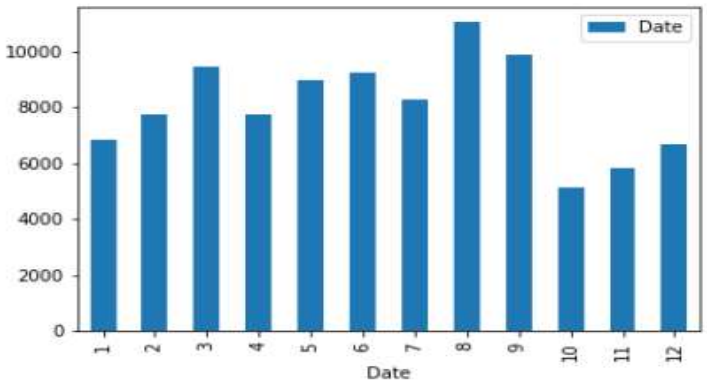
```
Column Name                              =  Date
Data type                                =  datetime64[ns]
Number of Missing Values                 =  0
Percentage of rows populated             =  100.00%
Number of Unique values                  =  365
Percentage of Unique values in population =  0.38%
```

The distribution of top 15 are as follows:



	Value	Count	Percent(%)
0	2010-02-28	684	0.71
1	2010-08-10	610	0.63
2	2010-03-15	594	0.61
3	2010-09-13	564	0.58
4	2010-08-09	536	0.55
5	2010-09-07	536	0.55
6	2010-09-14	533	0.55
7	2010-09-21	522	0.54
8	2010-08-01	521	0.54
9	2010-08-31	518	0.54
10	2010-08-29	503	0.52
11	2010-09-28	503	0.52
12	2010-02-16	501	0.52
13	2010-09-19	497	0.51
14	2010-09-29	496	0.51

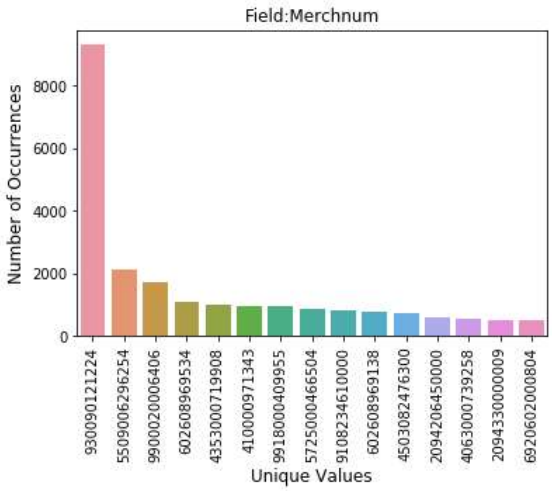
When aggregated by month (as shown below), August (month 8) appears most often while December is the least frequent month.



4) Merchnum

Merchnum serves as a unique identification code for each merchant. Merchnum 930090121224 is the most frequently visited merchant.

```
Column Name                                =  Merchnum
Data type                                  =  object
Number of Missing Values                   =  3375
Percentage of rows populated               =  96.51%
Number of Unique values                    =  13091
Percentage of Unique values in population =  14.02%
```

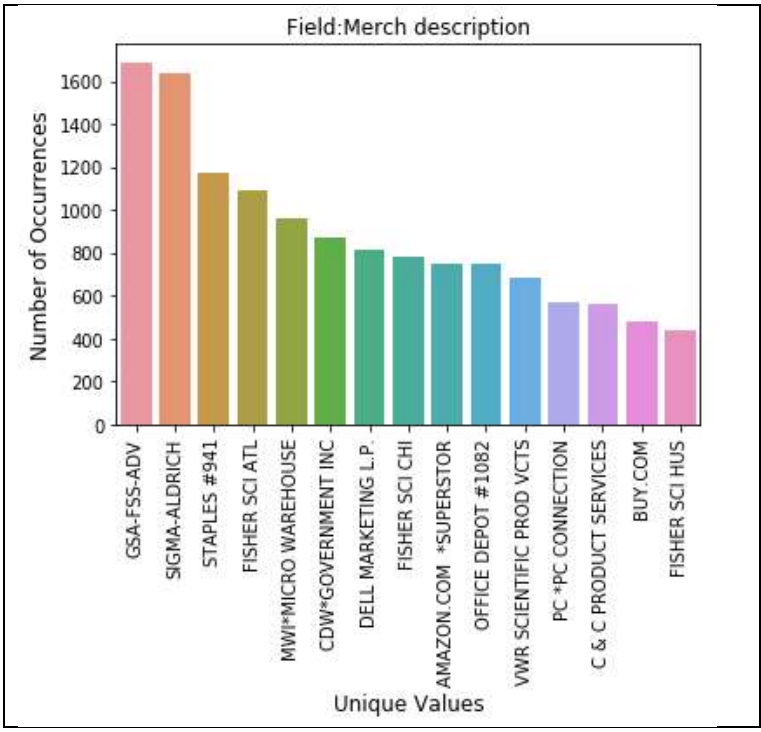


	Value	Count	Percent(%)
0	930090121224	9310	9.97
1	5509006296254	2131	2.28
2	9900020006406	1714	1.84
3	602608969534	1092	1.17
4	4353000719908	1020	1.09
5	410000971343	982	1.05
6	9918000409955	956	1.02
7	5725000466504	872	0.93
8	9108234610000	817	0.87
9	602608969138	783	0.84
10	4503082476300	746	0.80
11	2094206450000	590	0.63
12	4063000739258	568	0.61
13	2094330000009	533	0.57
14	6920602000804	523	0.56

5) Merch Description

Merch description describes what the merchant is along with number of that merchant store. GSA-FSS-ADV is the most frequent merchant description.

```
Column Name                                =  Merch description
Data type                                  =  object
Number of Missing Values                   =  0
Percentage of rows populated               =  100.00%
Number of Unique values                    =  13126
Percentage of Unique values in population =  13.57%
```

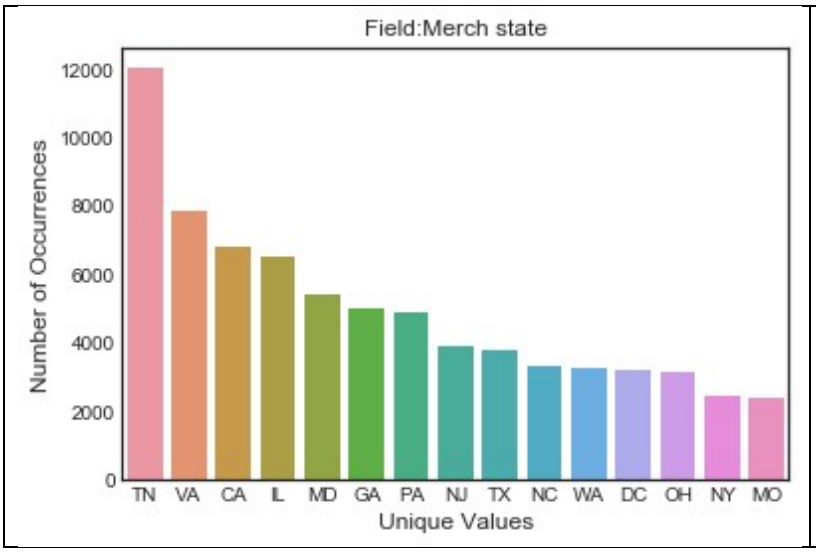


	Value	Count	Percent(%)
0	GSA-FSS-ADV	1688	1.74
1	SIGMA-ALDRICH	1635	1.69
2	STAPLES #941	1174	1.21
3	FISHER SCI ATL	1093	1.13
4	MWI*MICRO WAREHOUSE	958	0.99
5	CDW*GOVERNMENT INC	872	0.90
6	DELL MARKETING L.P.	816	0.84
7	FISHER SCI CHI	783	0.81
8	AMAZON.COM *SUPERSTOR	750	0.78
9	OFFICE DEPOT #1082	748	0.77
10	VWR SCIENTIFIC PROD VCTS	688	0.71
11	PC *PC CONNECTION	570	0.59
12	C & C PRODUCT SERVICES	558	0.58
13	BUY.COM	481	0.50
14	FISHER SCI HUS	442	0.46

6) Merch state

Merch state is where the merchant is located in. Besides for state abbreviation such as CA, there are several numbers. The most frequent merchant state is TN.

```
Column Name                          =  Merch state
Data type                            =  object
Number of Missing Values             =  1195
Percentage of rows populated         =  98.76%
Number of Unique values              =  227
Percentage of Unique values in population =  0.24%
```

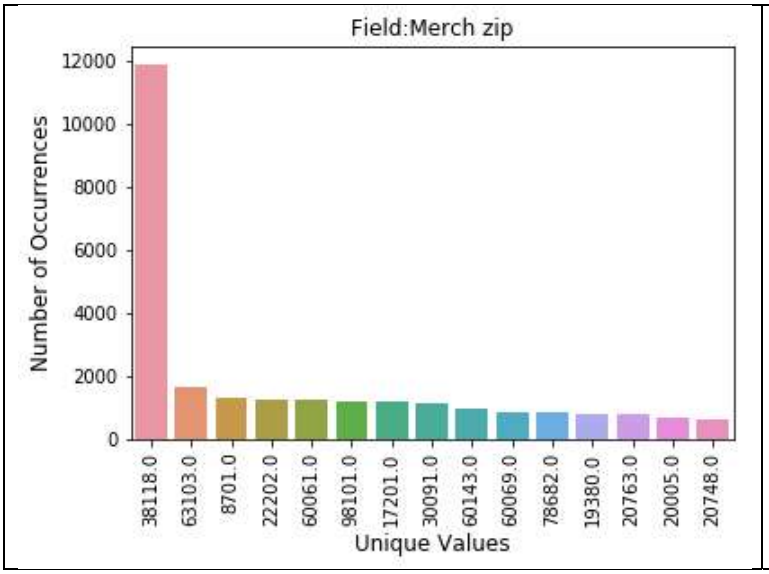


	Value	Count	Percent(%)
0	TN	12035	12.59
1	VA	7872	8.24
2	CA	6817	7.13
3	IL	6508	6.81
4	MD	5398	5.65
5	GA	5025	5.26
6	PA	4899	5.13
7	NJ	3912	4.09
8	TX	3790	3.97
9	NC	3322	3.48
10	WA	3300	3.45
11	DC	3208	3.36
12	OH	3131	3.28
13	NY	2430	2.54
14	MO	2420	2.53

7) Merch zip

Merch zip is the zip code where the merchant is located in. The most frequent merchant zip is 38118.

```
Column Name                          =  Merch zip
Data type                            =  float64
Number of Missing Values             =  4656
Percentage of rows populated         =  95.19%
Number of Unique values              =  4567
Percentage of Unique values in population =  4.96%
```

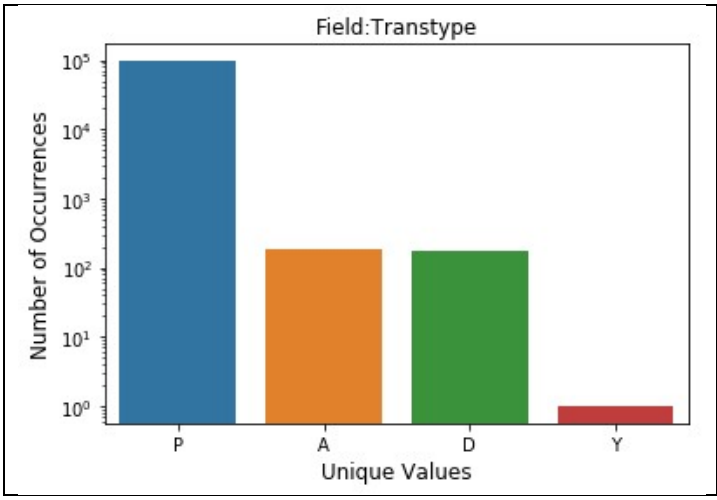


	Value	Count	Percent(%)
0	38118.0	11868	12.89
1	63103.0	1650	1.79
2	8701.0	1267	1.38
3	22202.0	1250	1.36
4	60061.0	1221	1.33
5	98101.0	1197	1.30
6	17201.0	1180	1.28
7	30091.0	1092	1.19
8	60143.0	942	1.02
9	60069.0	826	0.90
10	78682.0	817	0.89
11	19380.0	769	0.83
12	20763.0	749	0.81
13	20005.0	648	0.70
14	20748.0	592	0.64

8) Transtype

Transtype represents the type of transaction. Transtype has four values: P,D,A,Y. P is the most frequent transaction type while Y is the least frequent.

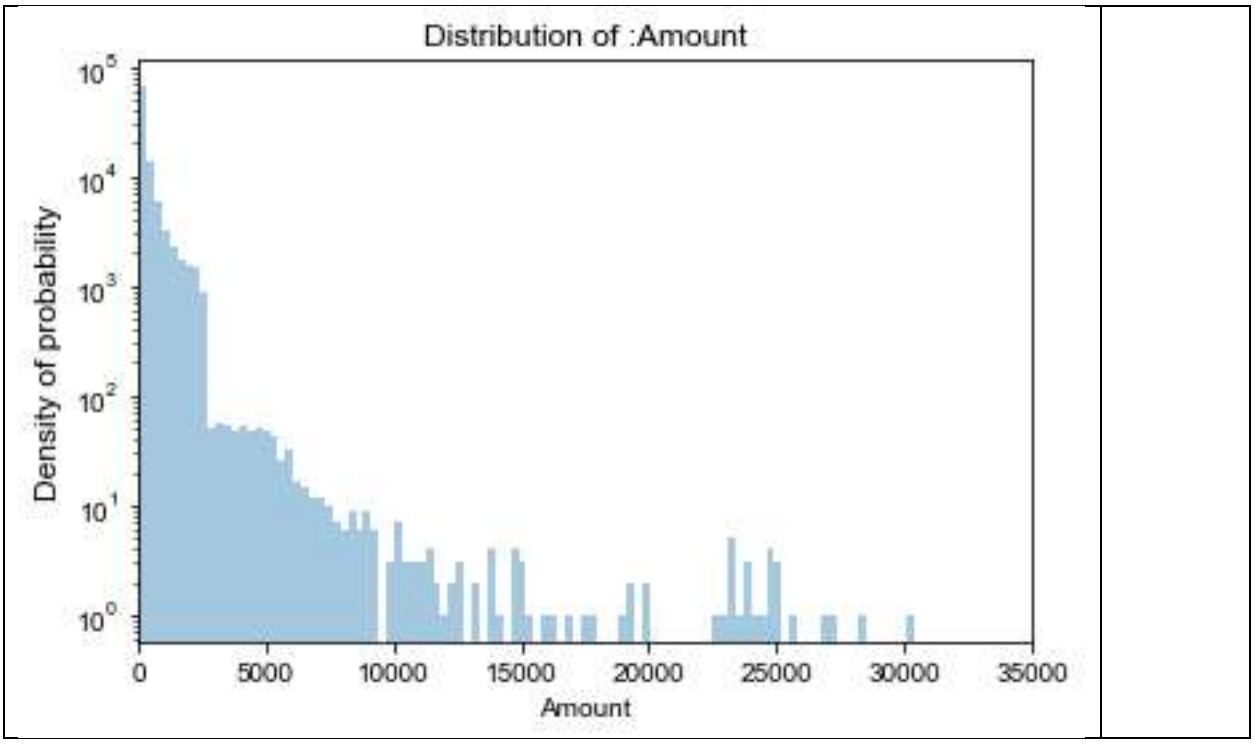
```
Column Name                              =  Transtype
Data type                                =  object
Number of Missing Values                 =  0
Percentage of rows populated             =  100.00%
Number of Unique values                  =  4
Percentage of Unique values in population =  0.00%
```



	Value	Count	Percent(%)
0	P	96398	99.63
1	A	181	0.19
2	D	173	0.18
3	Y	1	0.00

9) Amount

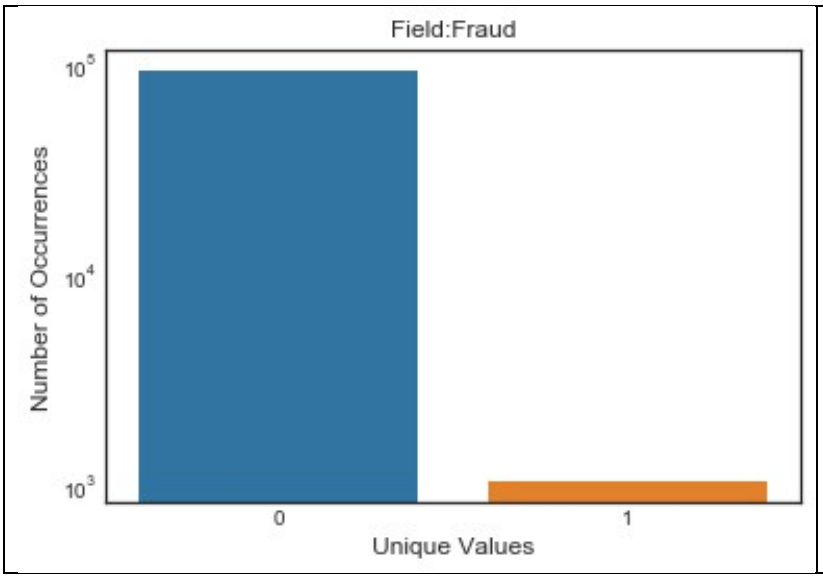
Amount indicates the dollar value of that transaction. Below is the distribution of amount after filtering out an outlier ($3,102,045.53) and cap amount under $35,000 on a log scale. Most transaction are in small amount.



10)Fraud

Fraud indicates whether that transaction is a fraudulent transaction or not. This will serve as the label on our algorithm later. Fraud has two levels—0 represent legitimate transactions while 1 represents fraudulent transaction. We can see this variable is extremely imbalanced.

```
Column Name                          =  Fraud
Data type                            =  int64
Number of Missing Values             =  0
Percentage of rows populated         =  100.00%
Number of Unique values              =  2
Percentage of Unique values in population =  0.00%
```



	Value	Count	Percent(%)
0	0	95694	98.91
1	1	1059	1.09

List of Variables

Amount Variables:

Variable Name	Description
Card_Amount_mean_1D	Amount spent by this card over the past 1 day
Card_Amount_max_1D	Amount spent by this card over the past 1 day
Card_Amount_median_1D	Amount spent by this card over the past 1 day
Card_Amount_total_1D	Amount spent by this card over the past 1 day
Card_Amount_Act_mean_1D	Actual over mean amount spent by this card over the past 1 day
Card_Amount_Act_max_1D	Actual over maximum amount spent by this card over the past 30 days

Card_Amount_Act_median_1D	Actual over median amount spent by this card over the past 1 day
Card_Amount_Act_total_1D	Actual over total amount spent by this card over the past 1 day
Card_Amount_mean_3D	Amount spent by this card over the past 3 days
Card_Amount_max_3D	Amount spent by this card over the past 3 days
Card_Amount_median_3D	Amount spent by this card over the past 3 days
Card_Amount_total_3D	Amount spent by this card over the past 3 days
Card_Amount_Act_mean_3D	Actual over mean amount spent by this card over the past 3 days
Card_Amount_Act_max_3D	Actual over maximum amount spent by this card over the past 3 days
Card_Amount_Act_median_3D	Actual over median amount spent by this card over the past 3 days
Card_Amount_Act_total_3D	Actual over total amount spent by this card over the past 3 days
Card_Amount_mean_7D	Amount spent by this card over the past 7 days
Card_Amount_max_7D	Amount spent by this card over the past 7 days
Card_Amount_median_7D	Amount spent by this card over the past 7 days
Card_Amount_total_7D	Amount spent by this card over the past 7 days

Card_Amount_Act_mean_7D	Actual over mean amount spent by this card over the past 7 days
Card_Amount_Act_max_7D	Actual over maximum amount spent by this card over the past 7 days
Card_Amount_Act_median_7D	Actual over median amount spent by this card over the past 7 days
Card_Amount_Act_total_7D	Actual over total amount spent by this card over the past 7 days
Card_Amount_mean_14D	Amount spent by this card over the past 14 days
Card_Amount_max_14D	Amount spent by this card over the past 14 days
Card_Amount_median_14D	Amount spent by this card over the past 14 days
Card_Amount_total_14D	Amount spent by this card over the past 14 days
Card_Amount_Act_mean_14D	Actual over mean amount spent by this card over the past 14 days
Card_Amount_Act_max_14D	Actual over maximum amount spent by this card over the past 14 days
Card_Amount_Act_median_14D	Actual over median amount spent by this card over the past 14 days
Card_Amount_Act_total_14D	Actual over total amount spent by this card over the past 14 days
Card_Amount_mean_30D	Amount spent by this card over the past 30 days
Card_Amount_max_30D	Amount spent by this card over the past 30 days

Card_Amount_median_30D	Amount spent by this card over the past 30 days
Card_Amount_total_30D	Amount spent by this card over the past 30 days
Card_Amount_Act_mean_30D	Actual over mean amount spent by this card over the past 30 days
Card_Amount_Act_max_30D	Actual over maximum amount spent by this card over the past 30 days
Card_Amount_Act_median_30D	Actual over median amount spent by this card over the past 30 days
Card_Amount_Act_total_30D	Actual over total amount spent by this card over the past 30 days
Merch_Amount_mean_1D	Amount spent at this merchant over the past 1 day
Merch_Amount_max_1D	Amount spent at this merchant over the past 1 day
Merch_Amount_median_1D	Amount spent at this merchant over the past 1 day
Merch_Amount_total_1D	Amount spent at this merchant over the past 1 day
Merch_Amount_Act_mean_1D	Actual over mean amount spent at this merchant over the past 1 day
Merch_Amount_Act_max_1D	Actual over maximum amount spent at this merchant over the past 1 day
Merch_Amount_Act_median_1D	Actual over median amount spent at this merchant over the past 1 day

Merch_Amount_Act_total_1D	Actual over total amount spent at this merchant over the past 1 day
Merch_Amount_mean_3D	Amount spent at this merchant over the past 3 days
Merch_Amount_max_3D	Amount spent at this merchant over the past 3 days
Merch_Amount_median_3D	Amount spent at this merchant over the past 3 days
Merch_Amount_total_3D	Amount spent at this merchant over the past 3 days
Merch_Amount_Act_mean_3D	Actual over mean amount spent at this merchant over the past 3 days
Merch_Amount_Act_max_3D	Actual over maximum amount spent at this merchant over the past 3 days
Merch_Amount_Act_median_3D	Actual over median amount spent at this merchant over the past 3 days
Merch_Amount_Act_total_3D	Actual over total amount spent at this merchant over the past 3 days
Merch_Amount_mean_7D	Amount spent at this merchant over the past 7 days
Merch_Amount_max_7D	Amount spent at this merchant over the past 7 days
Merch_Amount_median_7D	Amount spent at this merchant over the past 7 days
Merch_Amount_total_7D	Amount spent at this merchant over the past 7 days

Merch_Amount_Act_mean_7D	Actual over mean amount spent at this merchant over the past 7 days
Merch_Amount_Act_max_7D	Actual over maximum amount spent at this merchant over the past 7 days
Merch_Amount_Act_median_7D	Actual over median amount spent at this merchant over the past 7 days
Merch_Amount_Act_total_7D	Actual over total amount spent at this merchant over the past 7 days
Merch_Amount_mean_14D	Amount spent at this merchant over the past 14 days
Merch_Amount_max_14D	Amount spent at this merchant over the past 14 days
Merch_Amount_median_14D	Amount spent at this merchant over the past 14 days
Merch_Amount_total_14D	Amount spent at this merchant over the past 1 day
Merch_Amount_Act_mean_14D	Actual over mean amount spent at this merchant over the past 14 days
Merch_Amount_Act_max_14D	Actual over maximum amount spent at this merchant over the past 14 days
Merch_Amount_Act_median_14D	Actual over median amount spent at this merchant over the past 14 days
Merch_Amount_Act_total_14D	Actual over total amount spent at this merchant over the past 14 days
Merch_Amount_mean_30D	Amount spent at this merchant over the past 30 days

Merch_Amount_max_30D	Amount spent at this merchant over the past 30 days
Merch_Amount_median_30D	Amount spent at this merchant over the past 30 days
Merch_Amount_total_30D	Amount spent at this merchant over the past 30 days
Merch_Amount_Act_mean_30D	Actual over mean amount spent at this merchant over the past 30 days
Merch_Amount_Act_max_30D	Actual over maximum amount spent at this merchant over the past 30 days
Merch_Amount_Act_median_30D	Actual over median amount spent at this merchant over the past 30 days
Merch_Amount_Act_total_30D	Actual over total amount spent at this merchant over the past 30 days
Card_Merch_Amount_mean_1D	Mean amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_max_1D	Maximum amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_median_1D	Median amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_total_1D	Total amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_Act_mean_1D	Actual over mean amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_Act_max_1D	Actual over maximum amount spent by this card at this merchant over the past 1 day

Card_Merch_Amount_Act_median_1D	Actual over median amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_Act_total_1D	Actual over total amount spent by this card at this merchant over the past 1 day
Card_Merch_Amount_mean_3D	Mean amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_max_3D	Maximum amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_median_3D	Median amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_total_3D	Total amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_Act_mean_3D	Actual over mean amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_Act_max_3D	Actual over maximum amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_Act_median_3D	Actual over median amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_Act_total_3D	Actual over total amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_mean_7D	Mean amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_max_7D	Maximum amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_median_7D	Median amount spent by this card at this merchant over the past 7 days

Card_Merch_Amount_total_7D	Total amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_Act_mean_7D	Actual over mean amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_Act_max_7D	Actual over maximum amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_Act_median_7D	Actual over median amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_Act_total_7D	Actual over total amount spent by this card at this merchant over the past 7 days
Card_Merch_Amount_mean_14D	Mean amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_max_14D	Maximum amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_median_14D	Median amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_total_14D	Total amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_Act_mean_14D	Actual over mean amount spent by this card at this merchant over the past 3 days
Card_Merch_Amount_Act_max_14D	Actual over maximum amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_Act_median_14D	Actual over median amount spent by this card at this merchant over the past 14 days
Card_Merch_Amount_Act_total_14D	Actual over total amount spent by this card at this merchant over the past 14 days

Card_Merch_Amount_mean_30D	Mean amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_max_30D	Maximum amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_median_30D	Median amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_total_30D	Total amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_Act_mean_30D	Actual over mean amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_Act_max_30D	Actual over maximum amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_Act_median_30D	Actual over median amount spent by this card at this merchant over the past 30 days
Card_Merch_Amount_Act_total_30D	Actual over total amount spent by this card at this merchant over the past 30 days
Card_Zip_Amount_mean_1D	Mean amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_max_1D	Maximum amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_median_1D	Median amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_total_1D	Total amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_Act_mean_1D	Actual over mean amount spent by this card at this zip code over the past 1 day

Card_Zip_Amount_Act_max_1D	Actual over maximum amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_Act_median_1D	Actual over median amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_Act_total_1D	Actual over total amount spent by this card at this zip code over the past 1 day
Card_Zip_Amount_mean_3D	Mean amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_max_3D	Maximum amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_median_3D	Median amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_total_3D	Total amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_Act_mean_3D	Actual over mean amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_Act_max_3D	Actual over maximum amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_Act_median_3D	Actual over median amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_Act_total_3D	Actual over total amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_mean_7D	Mean amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_max_7D	Maximum amount spent by this card at this zip code over the past 7 days

Card_Zip_Amount_median_7D	Median amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_total_7D	Total amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_Act_mean_7D	Actual over mean amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_Act_max_7D	Actual over maximum amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_Act_median_7D	Actual over median amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_Act_total_7D	Actual over total amount spent by this card at this zip code over the past 7 days
Card_Zip_Amount_mean_14D	Mean amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_max_14D	Maximum amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_median_14D	Median amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_total_14D	Total amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_Act_mean_14D	Actual over mean amount spent by this card at this zip code over the past 3 days
Card_Zip_Amount_Act_max_14D	Actual over maximum amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_Act_median_14D	Actual over median amount spent by this card at this zip code over the past 14 days

Card_Zip_Amount_Act_total_14D	Actual over total amount spent by this card at this zip code over the past 14 days
Card_Zip_Amount_mean_30D	Mean amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_max_30D	Maximum amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_median_30D	Median amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_total_30D	Total amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_Act_mean_30D	Actual over mean amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_Act_max_30D	Actual over maximum amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_Act_median_30D	Actual over median amount spent by this card at this zip code over the past 30 days
Card_Zip_Amount_Act_total_30D	Actual over total amount spent by this card at this zip code over the past 30 days
Card_State_Amount_mean_1D	Mean amount spent by this card at this state over the past 1 day
Card_State_Amount_max_1D	Maximum amount spent by this card at this state over the past 1 day
Card_State_Amount_median_1D	Median amount spent by this card at this state over the past 1 day
Card_State_Amount_total_1D	Total amount spent by this card at this state over the past 1 day

Card_State_Amount_Act_mean_1D	Actual over mean amount spent by this card at this state over the past 1 day
Card_State_Amount_Act_max_1D	Actual over maximum amount spent by this card at this state over the past 1 day
Card_State_Amount_Act_median_1D	Actual over median amount spent by this card at this state over the past 1 day
Card_State_Amount_Act_total_1D	Actual over total amount spent by this card at this state over the past 1 day
Card_State_Amount_mean_3D	Mean amount spent by this card at this state over the past 3 days
Card_State_Amount_max_3D	Maximum amount spent by this card at this state over the past 3 days
Card_State_Amount_median_3D	Median amount spent by this card at this state over the past 3 days
Card_State_Amount_total_3D	Total amount spent by this card at this state over the past 3 days
Card_State_Amount_Act_mean_3D	Actual over mean amount spent by this card at this state over the past 3 days
Card_State_Amount_Act_max_3D	Actual over maximum amount spent by this card at this state over the past 3 days
Card_State_Amount_Act_median_3D	Actual over median amount spent by this card at this state over the past 3 days
Card_State_Amount_Act_total_3D	Actual over total amount spent by this card at this state over the past 3 days
Card_State_Amount_mean_7D	Mean amount spent by this card at this state over the past 7 days

Card_State_Amount_max_7D	Maximum amount spent by this card at this state over the past 7 days
Card_State_Amount_median_7D	Median amount spent by this card at this state over the past 7 days
Card_State_Amount_total_7D	Total amount spent by this card at this state over the past 7 days
Card_State_Amount_Act_mean_7D	Actual over mean amount spent by this card at this state over the past 7 days
Card_State_Amount_Act_max_7D	Actual over maximum amount spent by this card at this state over the past 7 days
Card_State_Amount_Act_median_7D	Actual over median amount spent by this card at this state over the past 7 days
Card_State_Amount_Act_total_7D	Actual over total amount spent by this card at this state over the past 7 days
Card_State_Amount_mean_14D	Mean amount spent by this card at this state over the past 14 days
Card_State_Amount_max_14D	Maximum amount spent by this card at this state over the past 14 days
Card_State_Amount_median_14D	Median amount spent by this card at this state over the past 14 days
Card_State_Amount_total_14D	Total amount spent by this card at this state over the past 14 days
Card_State_Amount_Act_mean_14D	Actual over mean amount spent by this card at this state over the past 3 days
Card_State_Amount_Act_max_14D	Actual over maximum amount spent by this card at this state over the past 14 days

Card_State_Amount_Act_median_14D	Actual over median amount spent by this card at this state over the past 14 days
Card_State_Amount_Act_total_14D	Actual over total amount spent by this card at this state over the past 3 days
Card_State_Amount_mean_30D	Mean amount spent by this card at this state over the past 30 days
Card_State_Amount_max_30D	Maximum amount spent by this card at this state over the past 30 days
Card_State_Amount_median_30D	Median amount spent by this card at this state over the past 30 days
Card_State_Amount_total_30D	Total amount spent by this card at this state over the past 30 days
Card_State_Amount_Act_mean_30D	Actual over mean amount spent by this card at this state over the past 30 days
Card_State_Amount_Act_max_30D	Actual over maximum amount spent by this card at this state over the past 30 days
Card_State_Amount_Act_median_30D	Actual over median amount spent by this card at this state over the past 30 days
Card_State_Amount_Act_total_30D	Actual over total amount spent by this card at this state over the past 30 days

Frequency Variables:

Variable Name	**Description**
Card_Count_1D	Number of transactions by this card over the past 1 day

Card_Count_3D	Number of transactions by this card over the past 3 days
Card_Count_7D	Number of transactions by this card over the past 7 days
Card_Count_14D	Number of transactions by this card over the past 14 days
Card_Count_30D	Number of transactions by this card over the past 30 days
Merch_Count_1D	Number of transactions with this merchant over the past 1 day
Merch_Count_3D	Number of transactions with this merchant over the past 3 days
Merch_Count_7D	Number of transactions with this merchant over the past 7 days
Merch_Count_14D	Number of transactions with this merchant over the past 14 days
Merch_Count_30D	Number of transactions with this merchant over the past 30 days
CardMerch_Count_1D	Number of transactions by this card at this merchant over the past 1 day
CardMerch_Count_3D	Number of transactions by this card at this merchant over the past 3 days
CardMerch_Count_7D	Number of transactions by this card at this merchant over the past 7 days
CardMerch_Count_14D	Number of transactions by this card at this merchant over the past 14 days

CardMerch_Count_30D	Number of transactions by this card at this merchant over the past 30 days
CardZip_Count_1D	Number of transactions by this card at this zip over the past 1 day
CardZip_Count_3D	Number of transactions by this card at this zip over the past 3 days
CardZip_Count_7D	Number of transactions by this card at this zip over the past 7 days
CardZip_Count_14D	Number of transactions by this card at this zip over the past 14 days
CardZip_Count_30D	Number of transactions by this card at this zip over the past 30 days
CardState_Count_1D	Number of transactions by this card at this state over the past 1 day
CardState_Count_3D	Number of transactions by this card at this state over the past 3 days
CardState_Count_7D	Number of transactions by this card at this state over the past 7 days
CardState_Count_14D	Number of transactions by this card at this state over the past 14 days
CardState_Count_30D	Number of transactions by this card at this state over the past 30 days

Days Since Variables:

Variable Name	**Description**
Days_since_per_Cardnum	Days since the last transaction by the same

	card
Days_since_per_Merchnum	Days since the last transaction at the same merchant
Days_since_per_Cardnum_Merchnum	Days since the last transaction by the same card at the same merchant
Days_since_per_Cardnum_Merchzip	Days since the last transaction by the same card at the same merchant zip code
Days_since_per_Cardnum_Merchstate	Days since the last transaction by the same card at the same merchant state

Velocity Variables:

Variable Name	**Description**
v_num_card_num_card7	Number of transactions by the same card over average daily number of transactions by the same card in the past 7 days
v_num_card_num_card14	Number of transactions by the same card over average daily number of transactions by the same card in the past 14 days
v_num_card_num_card30	Number of transactions by the same card over average daily number of transactions by the same card in the past 30 days
v_num_merch_num_merch7	Number of transactions with the same merchant over average daily number of transactions with the same merchant in the past 7 days
v_num_merch_num_merch14	Number of transactions with the same merchant over average daily number of transactions with the same merchant in the past 14 days

v_num_merch_num_merch30	Number of transactions with the same merchant over average daily number of transactions with the same merchant in the past 30 days
v_num_card_num_merch7	Number of transactions by the same card over average daily number of transactions with the same merchant in the past 7 days
v_num_card_num_merch14	Number of transactions by the same card over average daily number of transactions with the same merchant in the past 14 days
v_num_card_num_merch30	Number of transactions by the same card over average daily number of transactions with the same merchant in the past 30 days
v_num_merch_num_card7	Number of transactions with the same merchant over average daily number of transactions by the same card in the past 7 days
v_num_merch_num_card14	Number of transactions with the same merchant over average daily number of transactions by the same card in the past 14 days
v_num_merch_num_card30	Number of transactions with the same merchant over average daily number of transactions by the same card in the past 30 days
v_amount_merch_amount_merch7	Amount spent with the same merchant over average daily amount with the same merchant in the past 7 days
v_amount_merch_amount_merch14	Amount spent with the same merchant over average daily amount with the same merchant in the past 14 days
v_amount_merch_amount_merch30	Amount spent with the same merchant over average daily amount with the same merchant in the past 30 days

v_amount_merch_amount_card7	Amount spent with the same merchant over average daily amount by the same card in the past 7 days
v_amount_merch_amount_card14	Amount spent with the same merchant over average daily amount by the same card in the past 14 days
v_amount_merch_amount_card30	Amount spent with the same merchant over average daily amount by the same card in the past 30 days
v_amount_card_amount_merch7	Amount spent by the same card over average daily amount with the same merchant in the past 7 days
v_amount_card_amount_merch14	Amount spent by the same card over average daily amount with the same merchant in the past 14 days
v_amount_card_amount_merch30	Amount spent by the same card over average daily amount with the same merchant in the past 30 days
v_amount_card_amount_card7	Amount spent by the same card over average daily amount by the same card in the past 7 days
v_amount_card_amount_card14	Amount spent by the same card over average daily amount by the same card in the past 14 days
v_amount_card_amount_card30	Amount spent by the same card over average daily amount by the same card in the past 30 days
v_num_merch_amount_merch7	Number of transactions with the same merchant over average daily amount with the same merchant in the past 7 days

v_num_merch_amount_merch14	Number of transactions with the same merchant over average daily amount with the same merchant in the past 14 days
v_num_merch_amount_merch30	Number of transactions with the same merchant over average daily amount with the same merchant in the past 30 days
v_num_merch_amount_card7	Number of transactions with the same merchant over average daily amount by the same card in the past 7 days
v_num_merch_amount_card14	Number of transactions with the same merchant over average daily amount by the same card in the past 14 days
v_num_merch_amount_card30	Number of transactions with the same merchant over average daily amount by the same card in the past 30 days
v_num_card_amount_merch7	Amount spent with the same merchant over average daily amount with the same merchant in the past 7 days
v_num_card_amount_merch14	Amount spent with the same merchant over average daily amount with the same merchant in the past 14 days
v_num_card_amount_merch30	Amount spent with the same merchant over average daily amount with the same merchant in the past 30 days
v_num_card_amount_card7	Amount spent with the same merchant over average daily amount by the same card in the past 7 days
v_num_card_amount_card14	Amount spent with the same merchant over average daily amount by the same card in the past 14 days

v_num_card_amount_card30	Amount spent with the same merchant over average daily amount by the same card in the past 30 days
v_amount_merch_num_merch7	Amount spent with the same merchant over average daily number of transactions with the same merchant in the past 7 days
v_amount_merch_num_merch14	Amount spent with the same merchant over average daily number of transactions with the same merchant in the past 14 days
v_amount_merch_num_merch30	Amount spent with the same merchant over average daily number of transactions with the same merchant in the past 30 days
v_amount_merch_num_card7	Amount spent with the same merchant over average daily number of transactions by the same card in the past 7 days
v_amount_merch_num_card14	Amount spent with the same merchant over average daily number of transactions by the same card in the past 14 days
v_amount_merch_num_card30	Amount spent with the same merchant over average daily number of transactions by the same card in the past 30 days
v_amount_card_num_merch7	Amount spent by the same card over average daily number of transactions with the same merchant in the past 7 days
v_amount_card_num_merch14	Amount spent by the same card over average daily number of transactions with the same merchant in the past 14 days
v_amount_card_num_merch30	Amount spent by the same card over average daily number of transactions with the same merchant in the past 30 days

v_amount_card_num_card7	Amount spent by the same card over average daily number of transactions by the same card in the past 7 days
v_amount_card_num_card14	Amount spent by the same card over average daily number of transactions by the same card in the past 14 days
v_amount_card_num_card30	Amount spent by the same card over average daily number of transactions by the same card in the past 30 days