THE FRAUD DETECTION OF CREDIT CARD TRANSACTION

Team 3: Shanjiao Jiang, Yunze Li, Yishi Lu,

Shihao Ma, Xin Xia, Yeyang Tang

Date: March 28, 2019

Contents

1. Executive Summary

This report provides an analysis of the 2010 Government Organization Credit Card Transaction Data, which contains 96,753 credit card transactions that occurred from January 1, 2010 to December 31, 2010. A total of 10 fields (variables) are provided by the original dataset to describe each transaction further. The goal of the analysis is to train and evaluate several supervised machine learning algorithms that are able to effectively predict and detect credit card fraud.

This report provides a detailed explanation of the steps that we went through to generate the six supervised machine learning algorithms. The summary of the seven steps is as follow:

1. Data cleaning, which fills in the necessary missing fields with reasonable numbers.
2. Variable creation and Z-scaling. 371 new variables are created and Z-scaled.
3. Training/testing/OOT. The entire dataset is split into training, testing and out-of-time (OOT) set.
4. Univariate Kolmogorov-Smirnov score (KS) and Fraud Detection Rates (FDR) calculation. For the 371 new variables, the univariate KS and univariate FDR at 3% are calculated. The 371 new variables are then sorted based on their KS and FDR at 3%.
5. Feature selection. The 186 new variables that have the high KS and FDR are first selected. A forward selection is then performed on the 186 variables to come up with 20 variables.
6. Machine learning algorithms. Six machine learning algorithms are trained by using the selected variables. The six machine learning algorithms are Logistic Regression, Neural Network, Random Forest, Boosted Trees, Support Vector Machine (SVM), and K-Nearest Neighbors (K-NN).
7. Machine learning algorithms evaluation. For each of the six machine learning algorithms, the effectiveness of catching credit card fraud is measured through the calculation of the FDR at 3%.

The evaluation of the six machine learning algorithms reveals that the random forest algorithm with 400 trees and 8 variables sampled for splitting is the most effective fraud detection algorithm. On the other hand, K-Nearest Neighbors seems not to be a suitable fraud detection algorithm for this fraud detection project.

2. Data Description

The original dataset used for analysis is the 2010 Government Organization Credit Card Transaction Data. It contains 96,753 credit card transactions that occurred from January 1, 2010 to December 31, 2010. A total of 10 fields (variables) are included in the dataset to provide further information and description of the credit card transactions.

The dataset was created in February 2011, but not published for confidentiality reasons. The data was collected, recorded, and managed by various government departments, including the Department of Credit Risk.

All 10 fields of the original dataset could be further divided into eight categorical fields and two numerical fields. The two tables in this section of the report will summarize the information provided through the 10 fields.

2.1 Numerical Field Table

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Mean	STD	Min	Max
1	Amount	Numerical	96753	100.00%	34909	0	4.28E+02	10006.1403	1.00E-02	3.10E+06
2	Date	Categorical	96753	100.00%	365	0	2/28/10	-	-	12/31/10

Table 1.

2.2 Categorical Field Table

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Most Common Field Name
1	Recnum	Categorical	96753	100.00%	96753	0	All Different
2	Cardnum	Categorical	96753	100.00%	1645	0	5142148452
3	Merchnum	Categorical	93378	96.51%	13092	0	930090121224
4	Merch description	Categorical	96753	100.00%	13126	0	GSA-FSS-ADV
5	Merch state	Categorical	95558	98.76%	228	0	TN
6	Merch zip	Categorical	92097	95.19%	4568	0	38118
7	Date	Categorical	96753	100.00%	365	0	2/28/10
8	Fraud	Categorical	96753	100.00%	2	95694	0

Table 2.

2.3 A Brief Introduction of the Ten Fields

In this section of the report, we will demonstrate the graphs of some of the important variables from the original dataset. A more detailed introduction of the 10 variables will be provided in the Appendix 1, where a data quality report (DQR) is attached.

Field 2

- Field Name: Cardnum
- Field Type: Categorical

- Description: Credit card number

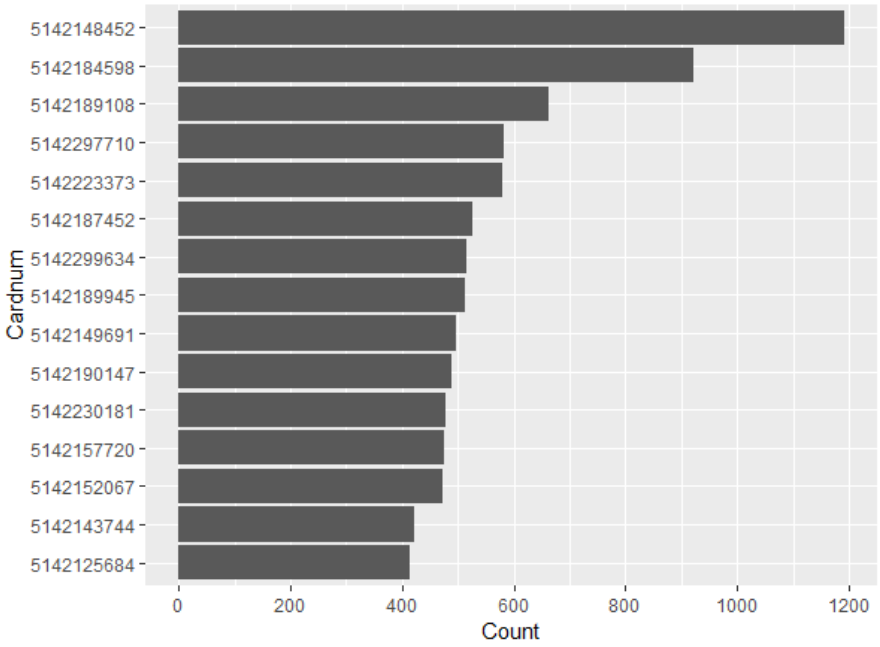


Figure 1.

Field 3

- Field Name: Date
- Field Type: Numerical
- Description: The date that transaction occurs

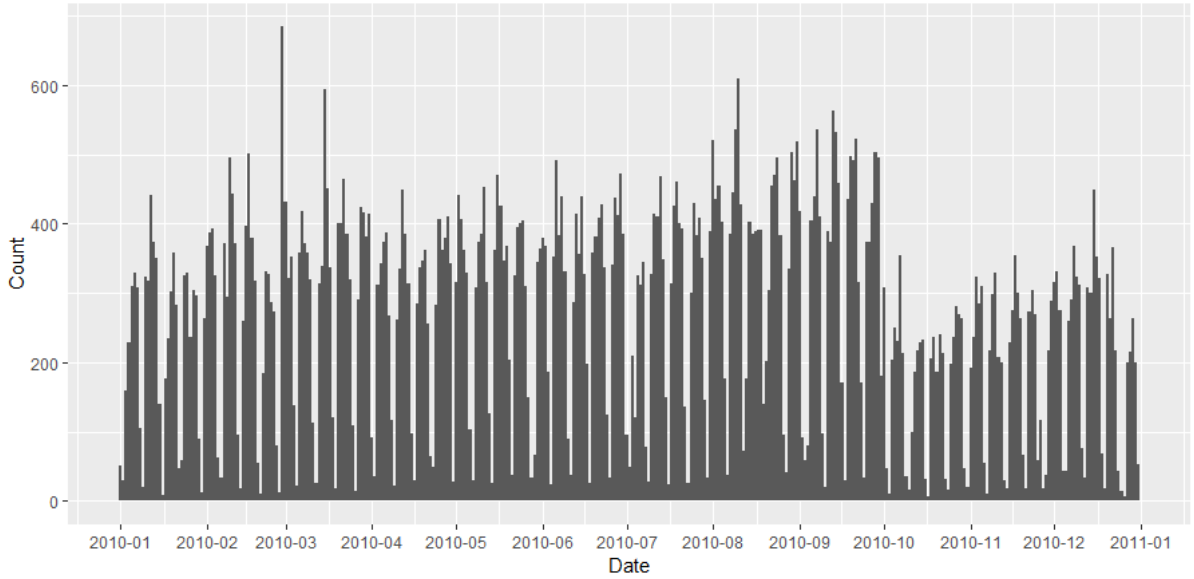


Figure 2.

Field 4

- Field Name: Merchnum
- Field Type: Categorical
- Description: Merchant number

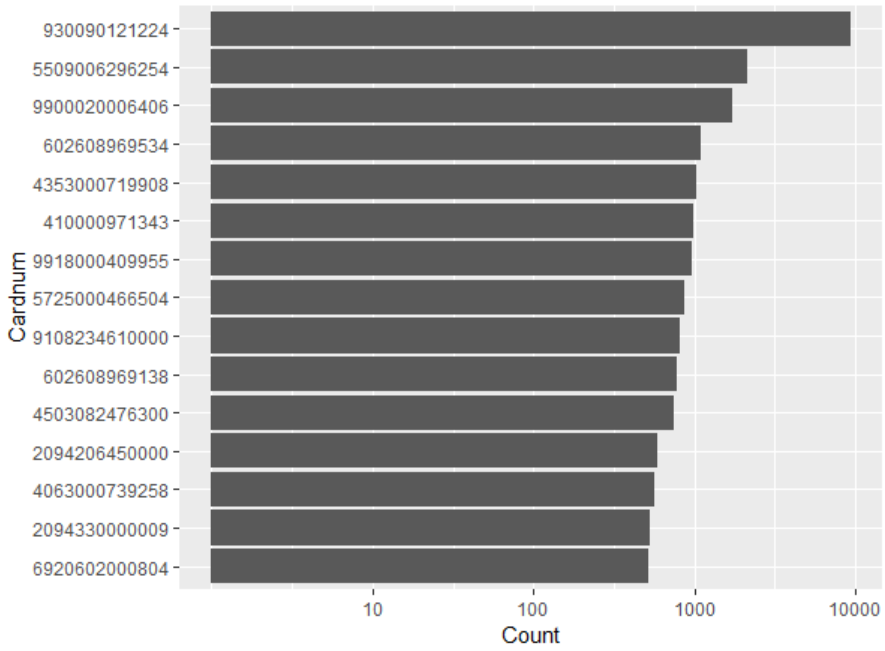


Figure 3.

Field 9

- Field Name: Amount
- Field Type: Numerical
- Description: Credit card transaction dollar amount

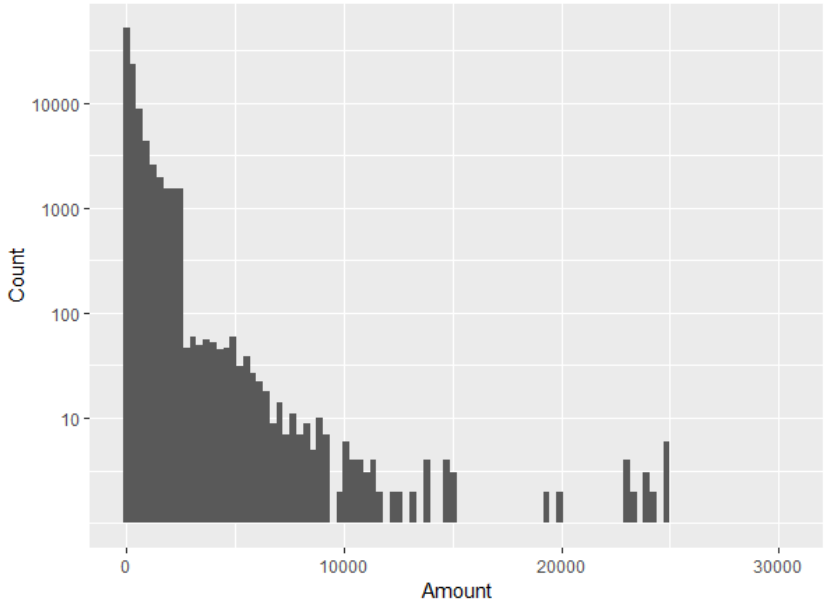


Figure 4.

Field 10

- Field Name: Fraud
- Field Type: Categorical
- Description: Whether this is a fraudulent transaction

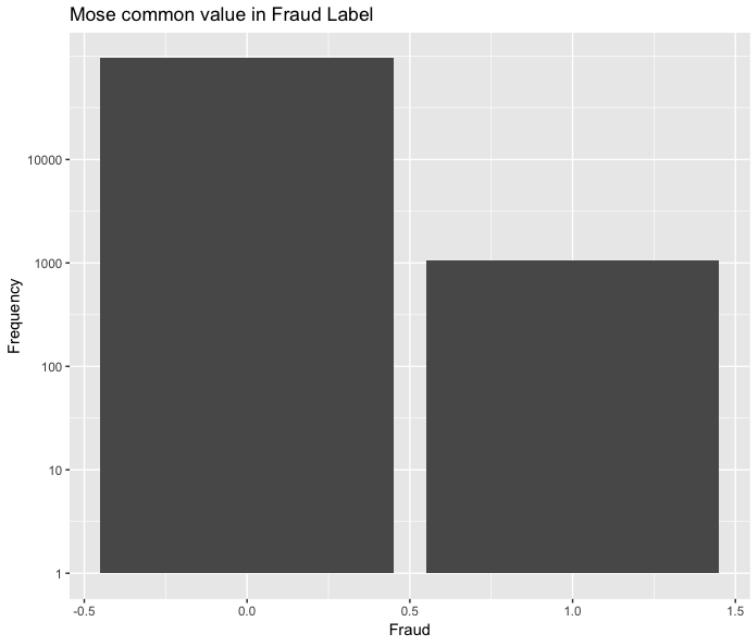


Figure 5.

A more detailed introduction of the 10 fields will be provide in the data quality report (DQR) attached in Appendix.

The 371 new variables are created based on the original 10 fields in the dataset. Section 4 of this report will provide a breakdown of the new variables and how they are created.

3. Data Cleaning

In this section of the report, the methods used to fill in the missing values associated with the 10 fields and 96,753 records in the original dataset will be introduced.

The data cleaning methods can be summarized as a three-step process: (1) removing outliers, (2) filtering data records, and (3) filling missing value. After the three steps, we kept 10 fields and 96,397 records in the "cleaned" dataset

- **Step1: Remove outliers**
  A credit card transaction that had a $3,102,045.53 transaction amount was removed as an outlier. Plus, the transaction was occurred in Mexico, instead of the United States.
- **Step2: Filter records**
  Only the records with transtype "P" are kept. Transtype "P" stands for purchase.
- **Step3: Fill missing value**
  The missing values for the three fields were filled. The three fields are Merch state, Merch zip, and Merchnum.

Here are how the missing values of the three fields are filled.

- **Merch state**
  Group by Zip and fill by mode of state in each group of Zip. If still missing, fill NA with the most frequent state "TN".
- **Merch zip**
  Group by Cardnum and fill by mode of zip in each group of Cardnum. If still missing, group by Merch state and fill by mode of zip in each group of Merch state.
- **Merchnum**
  Group by Cardnum and fill by mode of Merchnum in each group of Cardnum. If still missing, group by Merch state and fill by mode of Merchnum in each group of Merch state.

4. Variable Creation

In this section of the report, the creation of the 371 variables and the reasons why the 371 variables are needed to detect fraudulent credit card transactions will be explained.

4.1 The Logics Behind Variable Creation:

Generally speaking, the detection of fraudulent credit card transactions equates the detection of anomalous credit card transaction records. The most effective way to measure a transaction's abnormality is through its various "dimensionalities", which refers to the different features of the transaction record.

However, we do not know what kind of variables should be focused on or demonstrate the most related information, so the best strategy is to create as many variables (that may relate) as we can, and select the most suitable variables (see Section 5).

Before building the new variables, a quick review of the information (data) in the original dataset seem to be very helpful. The 10 fields, 96,753 credit card transaction data can be divided into:

- **Statistical data:** including average number, maximum number, median number and total (sum) number. Besides, the differences between actual data (the number of a specific transaction itself) and these four number (average, maximum, median and total) can also be great variables, here we choose the quotient of actual data and these four variables to show the difference.
- **Group data:** group data means the data that we use to group and calculate statistical data. Basically, considering the original fields, we have five groups: card, merchant, card at this merchant, card in this zip code and card in this state.
- **Time data:** when we identify transaction fraud, we are using historical data to measure its reliability. Therefore, time is an important factor to consider. In total, we have one year of credit card transaction records, so we choose 0, 1, 3, 7, 14, 30 days as intervals to create variables.

Based on the logic stated above, we have the following four parts of candidate variables:

1. Amount Variables:

Average	amout by/at this		over the past	
Maximum		card		0 days
Median		merchant		1 day
Total		card at this merchant		3 days
Actual/average		card in this zip code		7 days
Actual/maximum		card in this state		14 days
Actual/median				30 days
Actual/total				

Table 3.

2. Frequency Variables:

Number of transactions with this	card	over the past	0 days
	merchant		1 day
	card at this merchant		3 days
	card in this zip code		7 days
	card in this state		14 days
			30 days

Table 4.

3. Days since Variables:

Current date minus date of most recent transaction with same	card
	merchant
	card at this merchant
	card in this zip code
	card in this state

Table 5.

4. Velocity change Variables:

Number	of transactions with same		card	over the past	0 days
Amount			merchant		1 day
Divided by					
Average daily	number	of transactions with same	card	over the past	7 days
	amount		merchant		14 days
					30 days

Table 6.

4.2 The Detailed List of All Variables

Considering the length of the detailed list of all variables (18 pages), we attached it at the end of the report (Appendix).

5. Feature Selection Process

In this section of the report, the methods we used for feature selection will be explained.

Generally speaking, there are three ways to categorize feature selection. The three ways are (1) Filter, (2) Wrapper, and (3) Embedded. To complete feature selection, we used both Filter and Wrapper methods.

To select the variables used to build the machine learning algorithm, we first calculated the univariate Kolmogorov-Smirnov (KS) and univariate Fraud Detection Rate (FDR) at 3% for each one of the 371 new variables. The 371 new variables are then sorted based on their KS and FDR at 3%. The 186 variables that have the high KS and FDR are selected. A forward selection is then performed on the 186 variables to reduce the number of candidate variables for our machine learning algorithm to 20.

In machine learning, univariate KS and univariate FDR are examples of Filter methods. Forward selection and other stepwise selection are examples of Wrapper methods.

A more detailed explanation of univariate KS and univariate FDR are provided next.

Filter is a method that is independent of any modeling. It includes univariate model performance measure of every single variable. KS is a robust measure of how well the distributions of goods and bads are separated and is the maximum of the difference of the cumulative goods and bads. The formulas of KS are shown as below:

$$KS = \max_{x} \int_{x_{min}}^{x} [P_{\text{good}} - P_{\text{bad}}]\, dx$$

$$KS = \max_{x} \sum_{x_{min}}^{x} [P_{\text{good}} - P_{\text{bad}}]$$

Specifically, in our dataset, fraud label of 0 represents good and fraud label of 1 represents bad. For each variable, we calculated the probability distribution function (PDF) as well as cumulative distribution function (CDF) for both goods as bads to get the maximum difference between cumulative goods and bads. Figure 6 is a visual illustration describing how KS is calculated, and the dashed lines are plotted based on cumulative goods and bads.

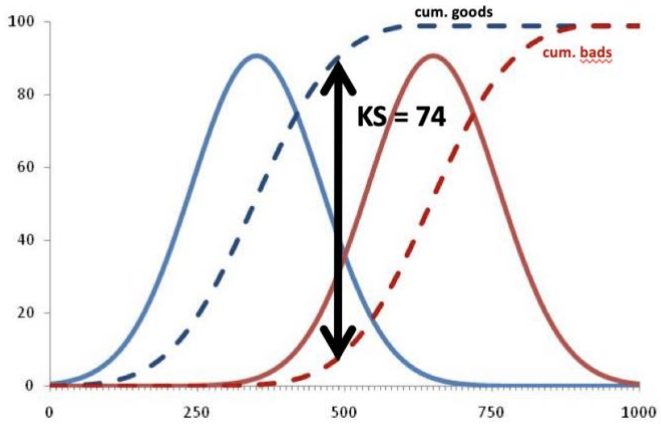


Figure 6: Cumulative Percentage of Goods and Bads in KS

FDR measures what percentage of all the frauds are caught at a particular examination cutoff location. It is the number of fraud caught divided by total number of fraud for each subpopulation. Specifically, for each variable, we first ranked our data based on its value in both descending and ascending order, and then calculated the percentage of fraud caught at the cutoff point we set, which is the top 3% of records. Therefore, for each variable, we got two FDRs. We chose the greater of the two FDR's to be the FDR of each variable.

After calculating KS and FDR for each variable, we ranked variables in descending order and took the average of these two rank orders (KS and FDR) in order to select the first half of variables out of the 371 variables. On top of univariate measures, we also applied wrapper method, which indicates that a model "wrapped" around the process, to measure how well several variables work together. Therefore, after we reduced the dimensionality in the previous step, we used stepwise logistic regression and forward selection to select 20 variables from the 186 variables.

Forward selection is a bottom up method. It builds one-dimensional models for all variables and adds one variable each time.

tot_cs_3	Total amount by/at this card in this state over the past 3 days
tot_card_0	Total amount by/at this card over the past 0 days
tot_merch_1	Total amount by/at this merchant over the past 1 days
max_merch_1	Maximum amount by/at this merchant over the past 1 days
tot_cm_30	Total amount by/at this card at this merchant over the past 30 days
vcv_ac1_nc14	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 14 days
qat_cs_3	Actual/total amount by/at this card in this state over the past 3 days
qamed_card_30	Actual/median amount by/at this card over the past 30 days
max_card_30	Maximum amount by/at this card over the past 30 days
max_card_0	Maximum amount by/at this card over the past 0 days
tot_card_14	Total amount by/at this card over the past 14 days

tot_cs_30	Total amount by/at this card in this state over the past 30 days
tot_card_1	Total amount by/at this card over the past 1 days
qat_cm_3	Actual/total amount by/at this card at this merchant over the past 3 days
avg_cs_0	Average amount by/at this card in this state over the past 0 days
tot_card_30	Total amount by/at this card over the past 30 days
med_cs_14	Median amount by/at this card in this state over the past 14 days
med_cm_7	Median amount by/at this card at this merchant over the past 7 days
max_merch_14	Maximum amount by/at this merchant over the past 14 days
max_merch_7	Maximum amount by/at this merchant over the past 7 days

Table 7.

6. Machine Learning Algorithms

We built our machine learning models for the transactions happened before November 1, 2010 considering a historical data window, and we then wanted to test the model consistency against the transactions happened after November 1, 2010 as our out-of-time validation. Thus, we split our datasets into three parts before building our models, which were training and testing sets based on transactions before November 1, 2010, and out-of-time set (OOT). We prepared our training and testing sets followed a 70/30 split rule. We measured the goodness for fraud using FDR at 3%, which was obtained by finding out the number of true frauds (using the actual fraud label) in the top 3% records after sorting the data based on predicted probability. For each model we built, we sampled 10 times randomly to get different training and testing sets and averaged the FDRs.

6.1 Logistic Regression

Logistic regression is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as good and bad. In our case, the output should be the indicator variable telling whether the record is a fraud or not. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables; Each independent variable can be a binary variable or a continuous variable.

When using Logistic regression model, it is necessary to consider "collinearity" which stands for the problem whether variables are correlated to each other. Before fitting the logistic regression model, we checked "collinearity" for all 20 variables after feature selection. We deleted 3 variables which were highly correlated to other variables and 2 variables which were not statistically significant in the model. With 15 variables left, we built our final full model of regression.

We selected several subsets of all 15 variables to build our logistic regression model through forward selection to get our best logistic regression model with all 15 or a subset of 15 variables. We applied Fraud Detection Rate on the top 3% of overall dataset as a benchmark to see how well a model is. For each selection, we ran the model for ten times to collect for ten times and summarized the model performance in terms of the average value of FDRs over these ten times modeling. The results of all logistic regression models we fitted and corresponding FDR is shown as following and we got the optimal FDR with 15 variables in the fitted logistic regression model:

Logistic Regression: FDR @ 3%			
	TRAIN	**TEST**	**OOT**
3V	64.11%	61.71%	30.39%
6V	66.57%	65.97%	30.17%
9V	67.47%	68.24%	32.74%
12V	68.14%	68.18%	36.37%
15V	68.84%	68.98%	36.98%

Table 8: FDR at 3% from 6 Logistic Regression Models

6.2 Neural Networks

Neural network is an architecture that builds an approximate mathematical function by fitting data points, and it mimics the way that human brain operates. A neural network contains layers of interconnected nodes, including an input layer, an output layer, and one or several hidden layers as shown in Figure 7. Each node takes an input, applies a function (often nonlinear) to it and then passes the output on to the next layer. Generally, the networks are defined to be feedforward: a node feeds its output to all the nodes on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one node to another. Below is an illustration of how Neural Networks works:

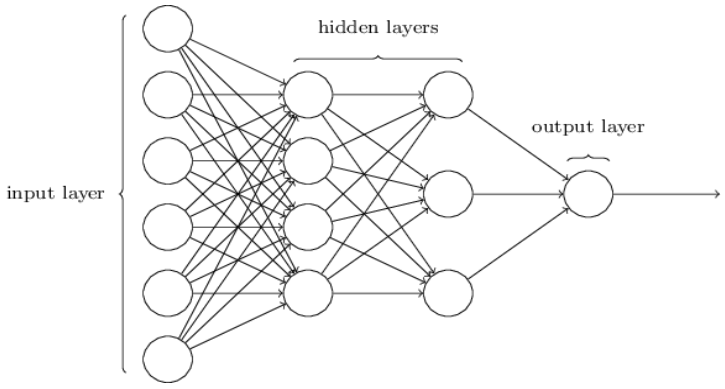


Figure 7: Neural Network Algorithm

We chose one hidden layer, which is sufficient for the large majority of problems, and set the solver for weight optimization as 'adam' for it works pretty well on relatively large datasets in terms of both training time and validation score. We tuned the number of nodes (range from 1 to 20) with the rectified linear unit function as activation function for the hidden layer. The following table shows the FDR at 3% calculated for training, testing and OOT datasets for each model.

Neural Networks: FDR @ 3%			
Nodes	**TRAIN**	**TEST**	**OOT**
4	74.56%	71.14%	34.52%
8	80.36%	76.24%	40.11%
12	82.49%	76.94%	39.66%
16	85.40%	77.61%	43.63%
20	85.26%	77.41%	40.95%

Table 9: FDR at 3% from 5 Neural Networks Models

6.3 Boosted Trees

Tree boosting is an ensemble method that seeks to create a strong classifier based on "weak" classifiers. In this context, weak and strong refer to a measure of the correlation between the learners and the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or reproduced by the model. Gradient Boosting also comprises an ensemble method that sequentially adds predictors and corrects previous models. However, instead of assigning different weights to the classifiers after every iteration, this method fits the new model to new residuals of the previous prediction and then minimizes the loss when adding the latest prediction. In the end, the model is updated using gradient descent. XGBoost implements this algorithm with an additional custom regularization term in the objective function to control overfitting.

In our variable domain, we only had 20 variables after feature selection. We used these 20 variables or subset of these variables to develop boosted tree models. We applied FDR as a benchmark to see how well a model is. For each selection, we ran the model for ten times to collect for ten times and then summarized the model's performance in terms of the average value of Fraud Detection Rates over these ten times modeling. The results all boosted tree models we fitted and corresponding Fraud Detection Rate is shown as following and we got the optimal FDR for OOT when we have 20 variables, 400 trees and 10 splits:

Boosted Tree: FDR @ 3%			
	TRAIN	**TEST**	**OOT**
20V, 10, 300	89.13%	82.99%	51.40%
20V, 10, 400	89.44%	83.26%	52.96%
20V, 10, 500	89.42%	84.13%	50.45%
20V, 20, 400	97.06%	91.91%	51.17%
17V, 10, 400	89.08%	84.34%	51.17%
15V, 10, 400	89.03%	83.18%	52.18%

Table 10: FDR at 3% from 6 Boosted Tree Models

6.4 Random Forest

Similar to Boosted Trees, random forest is an ensemble learning method and predict regression or classification by combining the outputs from individual trees. However, the order and the way results combined are different from Boosted Trees. It trains each tree independently, using a random sample of the data, which makes the model more robust than a single decision tree and less likely to overfit on the training data. Overall, it builds multiple decision trees and amalgamate them together to get a more accurate and stable prediction. Some primary parameters in Random Forest include number of trees, number of predictors sampled for splitting at each node, minimum node size and maximum tree depth. Figure 8 is an illustration of how Random Forest works:

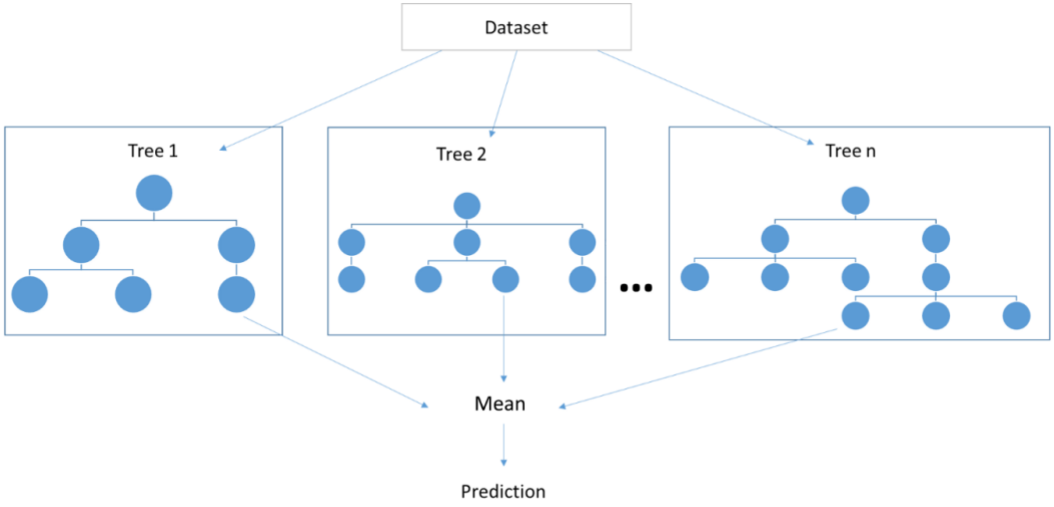


Figure 8: Random Forest Algorithm

When using Random Forest for prediction, collinearity is not an issue since the model is able to learn when some features are highly correlated, thus we started building our models with 20 variables. We set minimum node size at 1 and unlimited tree depth, and tuned number of trees and number of predictors sampled to build 13 models. The following table shows the FDR at 3% calculated for training, testing and OOT datasets for each model. We get the optimal FDR for OOT of 54.02% when we have 20 variables, 400 trees and 8 variables sampled for splitting.

Random Forest: FDR @ 3%			
	TRAIN	**TEST**	**OOT**
20V 500, 14	93.34%	93.46%	51.34%
20V 400, 14	93.38%	93.68%	51.90%
20V 400, 12	93.21%	94.23%	52.57%
20V 400, 10	93.43%	94.17%	52.57%
20V 400, 8	93.58%	93.34%	54.02%
20V 400, 6	93.53%	93.52%	53.52%
20V 300, 8	93.36%	92.81%	53.51%
17V 500, 8	92.81%	92.81%	51.23%
17V 500, 6	93.13%	93.13%	51.51%
17V 400, 8	92.76%	93.92%	51.17%
15V 500, 8	93.30%	92.64%	50.17%
15V 500, 6	92.89%	92.73%	52.12%
15V 300, 6	92.84%	92.68%	50.89%

Table 11: FDR at 3% from 13 Random Forest Models with Highlighted Row as the Optimal Model

6.5 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. It tries to project observations to higher dimension to find a split boundary. SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong. For instance, in two-dimensional space a hyperplane is a line dividing a plane in two parts where in each class lay in either side.

For SVM model, we tried different combination of kernel types and penalty parameter C of the error term. The following table shows the FDR at 3% calculated for training, testing and OOT datasets for each model.

SVM: FDR @ 3%			
Kernel, C	**TRAIN**	**TEST**	**OOT**
poly, 1	78.93%	71.70%	45.47%
sigmoid, 1	11.28%	10.94%	14.08%
rbf, 1	81.04%	77.87%	45.75%
rbf, 5	84.33%	79.76%	36.20%
rbf, 10	85.13%	80.79%	34.53%

Table 12: FDR at 3% from 5 SVM Models

6.6 K-Nearest Neighbors

Nearest neighbor is another popular machine learning algorithm. The most commonly seen nearest neighbor algorithm is the K-Nearest Neighbors (K-NN) algorithm. To run the K-NN algorithm, we first need to determine the value of K. Given the predetermined K and an observation x in the training data, the K-NN algorithm identifies the K points that are close to x. The algorithm then calculates the conditional probability for each class as the fraction of the K points.

When training the K-NN algorithm, we tried different values of K. Table 13 below summarizes the training, testing, and OOT FDR at 3%. It seems that K-NN algorithm is not a good choice to detect fraudulent credit card transaction because of its overfitting issue.

K-NN: FDR @ 3%			
K	TRAIN	TEST	OOT
4	100%	3.10%	0.56%
6	100%	3.22%	0.56%
8	96.19%	2.89%	0.56%
10	96.47%	2.93%	0.56%
100	81.82%	2.91%	0.56%

Table 13: FDR at 3% from 5 K-NN Algorithms

6.7 Results

The six candidate models we used in this project cover the linear and non-linear models with top popularity and accuracy. And the results in detecting top 3% fraud scores are listed in the below table.

Model	FDR @ 3%		
	Train	Test	OOT
Logistic Regression	64.84%	68.98%	36.98%
Neural Networks	85.40%	77.61%	43.63%
Boosted Trees	89.13%	83.26%	52.96%
Random Forests	93.58%	93.34%	54.02%
Support Vector Machine	81.04%	77.87%	45.75%
K-Nearest Neighbors	100%	3.22%	0.56%

Table 14.

Random Forest gives the most outstanding general performance and produces the highest Fraud Detection Rate on both train and test set, so we choose Random Forest with 400 trees and 8 variables sampled for splitting as our best model.

Important Features selected by Random Forest are the following:

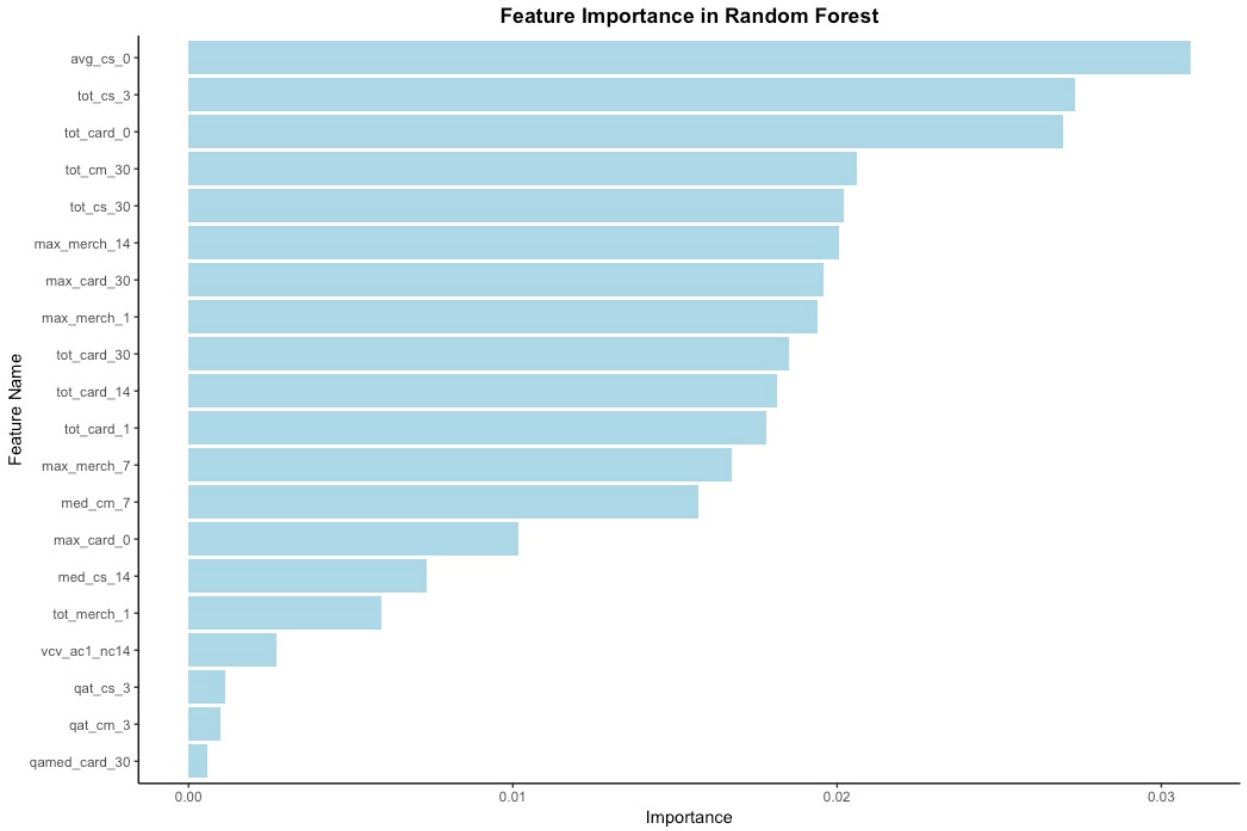


Figure 9.

Then we applied this model to see how it performs in Fraud Detection in the first 1%~20% of dataset we put on prediction. The following three tables are results of the chosen random forest model performs in Fraud Detection based on training and testing sets that we sampled only once, and out-of-time set (OOT) prediction. The column of KS is the difference between the detection rate of "bads" and "goods", indicating how well the scores of these two groups are differentiated. False positive ratio (FPR) is the number of goods caught divided by the number of "bads" caught. For each set, a plot comes with the corresponding table to show how much our selected fraud algorithm help in saving money in estimation:

Training	# Records	# Goods	# Bads	Fraud Rate							
	58779	58153	626	0.010650062							
	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	588	92	496	15.65%	84.35%	588	92	496	15.65%	79.23%	63.59%	0.19
2	588	520	68	88.44%	11.56%	1176	612	564	52.04%	90.10%	38.06%	1.09
3	587	570	17	97.10%	2.90%	1763	1182	581	67.04%	92.81%	25.77%	2.03
4	588	585	3	99.49%	0.51%	2351	1767	584	75.16%	93.29%	18.13%	3.03
5	588	585	3	99.49%	0.51%	2939	2352	587	80.03%	93.77%	13.74%	4.01
6	588	584	4	99.32%	0.68%	3527	2936	591	83.24%	94.41%	11.17%	4.97
7	588	587	1	99.83%	0.17%	4115	3523	592	85.61%	94.57%	8.96%	5.95
8	587	584	3	99.49%	0.51%	4702	4107	595	87.35%	95.05%	7.70%	6.90
9	588	585	3	99.49%	0.51%	5290	4692	598	88.70%	95.53%	6.83%	7.85
10	588	587	1	99.83%	0.17%	5878	5279	599	89.81%	95.69%	5.88%	8.81
11	588	587	1	99.83%	0.17%	6466	5866	600	90.72%	95.85%	5.13%	9.78
12	587	586	1	99.83%	0.17%	7053	6452	601	91.48%	96.01%	4.53%	10.74
13	588	588	0	100.00%	0.00%	7641	7040	601	92.13%	96.01%	3.87%	11.71
14	588	587	1	99.83%	0.17%	8229	7627	602	92.68%	96.17%	3.48%	12.67

15	588	587	1	99.83%	0.17%	8817	8214	603	93.16%	96.33%	3.16%	13.62
16	588	586	2	99.66%	0.34%	9405	8800	605	93.57%	96.65%	3.08%	14.55
17	587	587	0	100.00%	0.00%	9992	9387	605	93.95%	96.65%	2.70%	15.52
18	588	588	0	100.00%	0.00%	10580	9975	605	94.28%	96.65%	2.36%	16.49
19	588	588	0	100.00%	0.00%	11168	10563	605	94.58%	96.65%	2.06%	17.46
20	588	588	0	100.00%	0.00%	11756	11151	605	94.85%	96.65%	1.79%	18.43

Table 15.

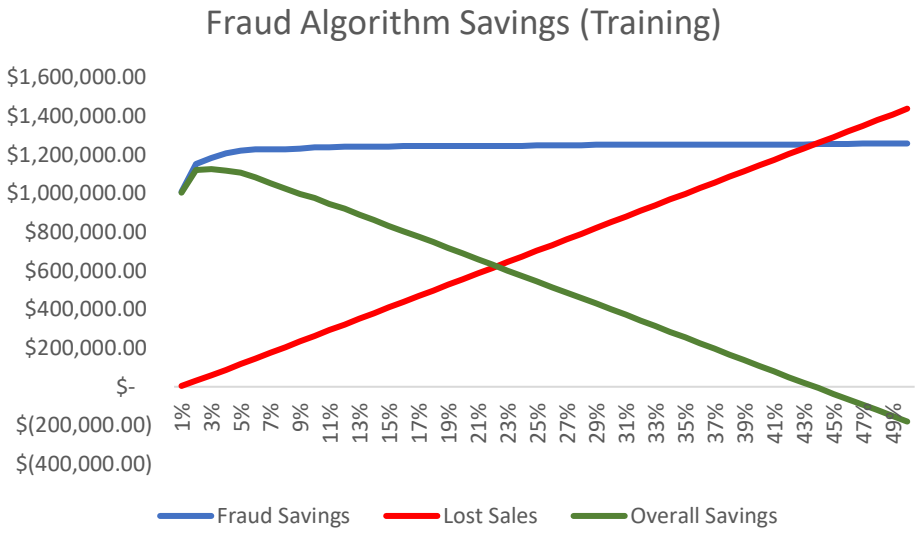


Figure 10.

Test	**# Records**		**# Goods**		**# Bads**	**Fraud Rate**						
	25191		**24937**		**254**	**0.010082966**						
	Bin Statistics					**Cumulative Statistics**						
Population Bin %	**# Records**	**# Goods**	**# Bads**	**% Goods**	**% Bads**	**Total # Records**	**Cumulative Goods**	**Cumulative Bads**	**% Goods**	**% Bads (FDR)**	**KS**	**FPR**
1	252	43	209	17.06%	82.94%	252	43	209	17.06%	82.28%	65.22%	0.21
2	252	224	28	88.89%	11.11%	504	267	237	52.98%	93.31%	40.33%	1.13
3	252	247	5	98.02%	1.98%	756	514	242	67.99%	95.28%	27.29%	2.12
4	252	247	5	98.02%	1.98%	1008	761	247	75.50%	97.24%	21.75%	3.08
5	252	251	1	99.60%	0.40%	1260	1012	248	80.32%	97.64%	17.32%	4.08
6	251	251	0	100.00%	0.00%	1511	1263	248	83.59%	97.64%	14.05%	5.09
7	252	252	0	100.00%	0.00%	1763	1515	248	85.93%	97.64%	11.70%	6.11
8	252	251	1	99.60%	0.40%	2015	1766	249	87.64%	98.03%	10.39%	7.09
9	252	252	0	100.00%	0.00%	2267	2018	249	89.02%	98.03%	9.02%	8.10
10	252	251	1	99.60%	0.40%	2519	2269	250	90.08%	98.43%	8.35%	9.08
11	252	252	0	100.00%	0.00%	2771	2521	250	90.98%	98.43%	7.45%	10.08
12	252	252	0	100.00%	0.00%	3023	2773	250	91.73%	98.43%	6.70%	11.09
13	252	252	0	100.00%	0.00%	3275	3025	250	92.37%	98.43%	6.06%	12.10
14	252	252	0	100.00%	0.00%	3527	3277	250	92.91%	98.43%	5.51%	13.11
15	252	251	1	99.60%	0.40%	3779	3528	251	93.36%	98.82%	5.46%	14.06
16	252	252	0	100.00%	0.00%	4031	3780	251	93.77%	98.82%	5.05%	15.06
17	251	251	0	100.00%	0.00%	4282	4031	251	94.14%	98.82%	4.68%	16.06
18	252	252	0	100.00%	0.00%	4534	4283	251	94.46%	98.82%	4.35%	17.06
19	252	252	0	100.00%	0.00%	4786	4535	251	94.76%	98.82%	4.06%	18.07
20	252	252	0	100.00%	0.00%	5038	4787	251	95.02%	98.82%	3.80%	19.07

Table 16.

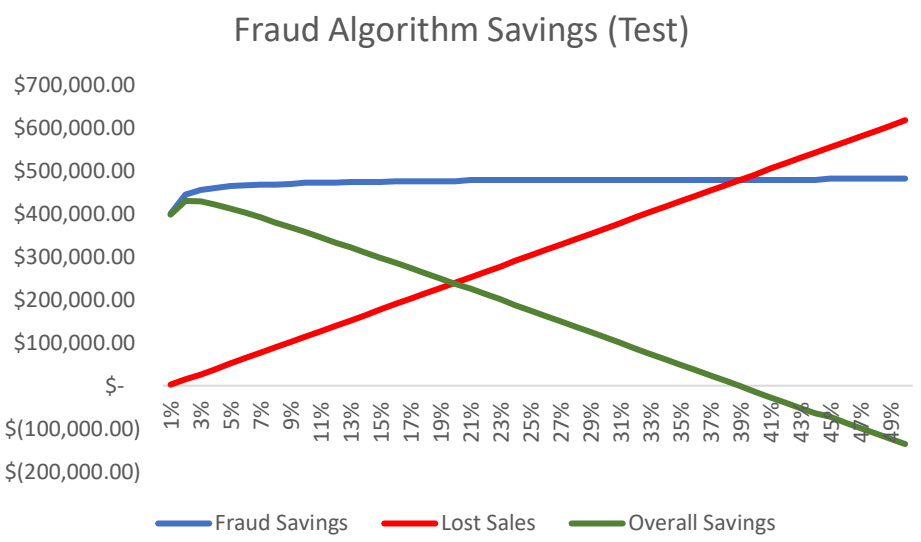


Figure 11.

Out of Time	# Records	# Goods	# Bads	Fraud Rate
	12427	12248	179	0.01440412

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	124	70	54	56.45%	43.55%	124	70	54	56.45%	30.17%	26.28%	1.30
2	125	93	32	74.40%	25.60%	249	163	86	65.46%	48.04%	17.42%	1.90
3	124	114	10	91.94%	8.06%	373	277	96	74.26%	53.63%	20.63%	2.89
4	124	114	10	91.94%	8.06%	497	391	106	78.67%	59.22%	19.45%	3.69
5	124	119	5	95.97%	4.03%	621	510	111	82.13%	62.01%	20.11%	4.59
6	125	120	5	96.00%	4.00%	746	630	116	84.45%	64.80%	19.65%	5.43
7	124	123	1	99.19%	0.81%	870	753	117	86.55%	65.36%	21.19%	6.44
8	124	122	2	98.39%	1.61%	994	875	119	88.03%	66.48%	21.55%	7.35
9	124	122	2	98.39%	1.61%	1118	997	121	89.18%	67.60%	21.58%	8.24
10	125	122	3	97.60%	2.40%	1243	1119	124	90.02%	69.27%	20.75%	9.02
11	124	114	10	91.94%	8.06%	1367	1233	134	90.20%	74.86%	15.34%	9.20
12	124	119	5	95.97%	4.03%	1491	1352	139	90.68%	77.65%	13.02%	9.73
13	125	124	1	99.20%	0.80%	1616	1476	140	91.34%	78.21%	13.12%	10.54
14	124	123	1	99.19%	0.81%	1740	1599	141	91.90%	78.77%	13.13%	11.34
15	124	120	4	96.77%	3.23%	1864	1719	145	92.22%	81.01%	11.22%	11.86
16	124	123	1	99.19%	0.81%	1988	1842	146	92.66%	81.56%	11.09%	12.62
17	125	125	0	100.00%	0.00%	2113	1967	146	93.09%	81.56%	11.53%	13.47
18	124	124	0	100.00%	0.00%	2237	2091	146	93.47%	81.56%	11.91%	14.32
19	124	123	1	99.19%	0.81%	2361	2214	147	93.77%	82.12%	11.65%	15.06
20	124	124	0	100.00%	0.00%	2485	2338	147	94.08%	82.12%	11.96%	15.90

Table 17.

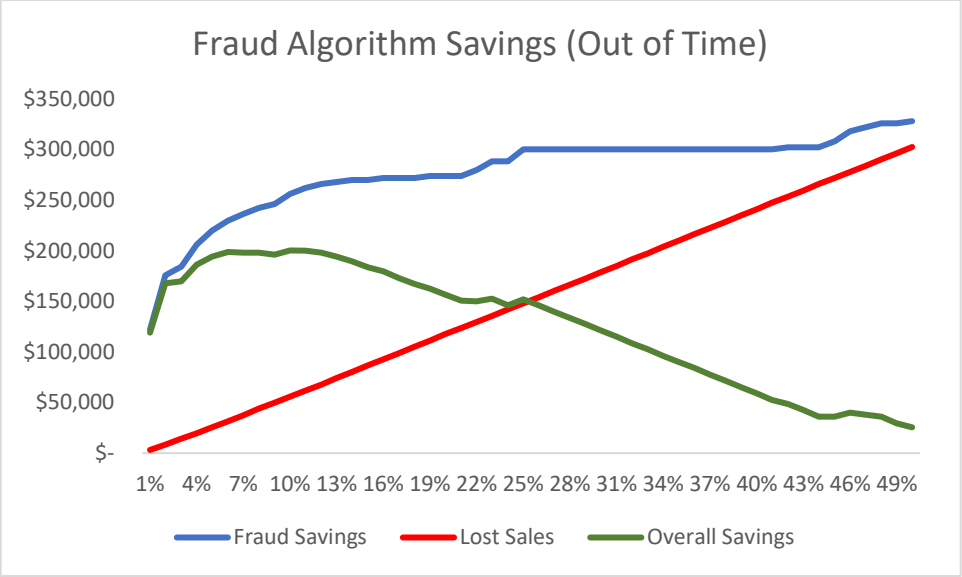

Fraud Algorithm Savings (Out of Time)
$350,000
$300,000
$250,000
$200,000
$150,000
$100,000
$50,000
$-
1% 4% 7% 10% 13% 16% 19% 22% 25% 28% 31% 34% 37% 40% 43% 46% 49%
Fraud Savings
Lost Sales
Overall Savings

Figure 12.

7. Conclusion

The 2010 Government Organization Credit Card Transaction Data provides the information of 96,753 credit card transactions that occurred from January 1, 2010 to December 31, 2010. This report provides a thorough analysis of the 96,753 credit card transactions. The goal of the analysis was to train and evaluate six machine learning algorithms that can help us effectively detect credit card transaction fraud.

To train the excellent machine learning algorithms, a total of 371 new variables were built and Z-scaled. The entire dataset was split into training, testing, and out-of-time (OOT) subset. Both Filter and Wrapper methods were performed on the new variables to select the 20 variables that were ready to be used to build the machine learning models.

The six machine learning algorithms that we built were (1) Logistic Regression, (2) Neural Network, (3) Random Forest, (4) Boosted Trees, (5) Support Vector Machine (SVM), and (6) K-Nearest Neighbors (K-NN). For each one of the six models, the effectiveness of catching credit card fraud was measured through the calculation of the FDR at 3%.

Our Random Forest algorithm with 400 trees and 8 variables sampled for splitting appears to be the most effective fraud detection algorithm. The K-Nearest Neighbors model, however, seems not to be a smart choice for this fraud detection project.

If we could have more time to work on this project, we probably would do the following to further improve the six machine learning algorithms.

As mentioned above, before building the six machine models, we split the 96,753 credit card transactions into training, testing, and out-of-time (OOT) subset. Any transactions that occurred on and after 11/1/2010 were included in the OOT set. The transactions before 11/1/2010 were included in the training and testing sets.

When we built the six models, for each model, we tried different combinations of parameters. For example, for the Random Forest model, we tried 300, 400, and 500 trees, together with various number of variables sampled for splitting. For each combination of the parameters, we reshuffled the training and testing sets 10 times to build 10 slightly different models.

Now, thinking back, we believe that using the 10-fold cross-validation to train and test 10 slightly different models for each combination of the parameters is a better way than simply doing the reshuffle. The rationale behind this claim is that using the10-fold cross-validation can make sure that each transaction record of the dataset would be used to train and test the models. Reshuffling, on the other hand, might use most, but not all records (because we just set the 70% to 30% ratio and let the computer randomly select the records to form the training and testing subsets).

When building the 371 new variables, we used six different time windows. However, if we could have more time, we would like to try more time windows. For example, we can use week and month. We can also use other time windows such as 5 days, 10 days, 15 days, etc.

One more thought about our analysis is that when we did the data cleaning, we tossed away the transaction that took place in Mexico. We treated it as an outlier. This is fine, if we wanted to only focus on the credit card transactions that occurred in the US, or in other words, to train the machine learning algorithms that only detect credit card fraud in the US. However, due to the globalization, international transactions become more and more common. Research also has revealed that international transactions that occurred in different time zone and used in different currency is a huge source of credit card fraud. Therefore, if we could have more time, we would like to closely examine the transaction that we tossed away as an outlier.

8. Appendix

8.1 The Detailed List of All Variables

Following is the list of all variables:

No.	Variables	Description (0 day here means 'today')
1	amount	The amount of transaction
2	avg_card_0	Average amount by/at this card over the past 0 days
3	avg_card_1	Average amount by/at this card over the past 1 days
4	avg_card_3	Average amount by/at this card over the past 3 days
5	avg_card_7	Average amount by/at this card over the past 7 days
6	avg_card_14	Average amount by/at this card over the past 14 days
7	avg_card_30	Average amount by/at this card over the past 30 days
8	max_card_0	Maximum amount by/at this card over the past 0 days
9	max_card_1	Maximum amount by/at this card over the past 1 days
10	max_card_3	Maximum amount by/at this card over the past 3 days
11	max_card_7	Maximum amount by/at this card over the past 7 days
12	max_card_14	Maximum amount by/at this card over the past 14 days
13	max_card_30	Maximum amount by/at this card over the past 30 days
14	med_card_0	Median amount by/at this card over the past 0 days
15	med_card_1	Median amount by/at this card over the past 1 days
16	med_card_3	Median amount by/at this card over the past 3 days
17	med_card_7	Median amount by/at this card over the past 7 days
18	med_card_14	Median amount by/at this card over the past 14 days
19	med_card_30	Median amount by/at this card over the past 30 days
20	tot_card_0	Total amount by/at this card over the past 0 days
21	tot_card_1	Total amount by/at this card over the past 1 days
22	tot_card_3	Total amount by/at this card over the past 3 days
23	tot_card_7	Total amount by/at this card over the past 7 days
24	tot_card_14	Total amount by/at this card over the past 14 days
25	tot_card_30	Total amount by/at this card over the past 30 days
26	qaa_card_0	Actual/average amount by/at this card over the past 0 days
27	qaa_card_1	Actual/average amount by/at this card over the past 1 days
28	qaa_card_3	Actual/average amount by/at this card over the past 3 days
29	qaa_card_7	Actual/average amount by/at this card over the past 7 days
30	qaa_card_14	Actual/average amount by/at this card over the past 14 days
31	qaa_card_30	Actual/average amount by/at this card over the past 30 days
32	qamax_card_0	Actual/maximum amount by/at this card over the past 0 days
33	qamax_card_1	Actual/maximum amount by/at this card over the past 1 days
34	qamax_card_3	Actual/maximum amount by/at this card over the past 3 days
35	qamax_card_7	Actual/maximum amount by/at this card over the past 7 days
36	qamax_card_14	Actual/maximum amount by/at this card over the past 14 days

37	qamax_card_30	Actual/maximum amount by/at this card over the past 30 days
38	qamed_card_0	Actual/median amount by/at this card over the past 0 days
39	qamed_card_1	Actual/median amount by/at this card over the past 1 days
40	qamed_card_3	Actual/median amount by/at this card over the past 3 days
41	qamed_card_7	Actual/median amount by/at this card over the past 7 days
42	qamed_card_14	Actual/median amount by/at this card over the past 14 days
43	qamed_card_30	Actual/median amount by/at this card over the past 30 days
44	qat_card_0	Actual/total amount by/at this card over the past 0 days
45	qat_card_1	Actual/total amount by/at this card over the past 1 days
46	qat_card_3	Actual/total amount by/at this card over the past 3 days
47	qat_card_7	Actual/total amount by/at this card over the past 7 days
48	qat_card_14	Actual/total amount by/at this card over the past 14 days
49	qat_card_30	Actual/total amount by/at this card over the past 30 days
50	freq_card_0	Number of transactions with this card over the past 0 days
51	freq_card_1	Number of transactions with this card over the past 1 days
52	freq_card_3	Number of transactions with this card over the past 3 days
53	freq_card_7	Number of transactions with this card over the past 7 days
54	freq_card_14	Number of transactions with this card over the past 14 days
55	freq_card_30	Number of transactions with this card over the past 30 days
56	days_since_card	Current date minus date of most recent transaction with same card
57	avg_merch_0	Average amount by/at this merchant over the past 0 days
58	avg_merch_1	Average amount by/at this merchant over the past 1 days
59	avg_merch_3	Average amount by/at this merchant over the past 3 days
60	avg_merch_7	Average amount by/at this merchant over the past 7 days
61	avg_merch_14	Average amount by/at this merchant over the past 14 days
62	avg_merch_30	Average amount by/at this merchant over the past 30 days
63	max_merch_0	Maximum amount by/at this merchant over the past 0 days
64	max_merch_1	Maximum amount by/at this merchant over the past 1 days
65	max_merch_3	Maximum amount by/at this merchant over the past 3 days
66	max_merch_7	Maximum amount by/at this merchant over the past 7 days
67	max_merch_14	Maximum amount by/at this merchant over the past 14 days
68	max_merch_30	Maximum amount by/at this merchant over the past 30 days
69	med_merch_0	Median amount by/at this merchant over the past 0 days
70	med_merch_1	Median amount by/at this merchant over the past 1 days
71	med_merch_3	Median amount by/at this merchant over the past 3 days
72	med_merch_7	Median amount by/at this merchant over the past 7 days
73	med_merch_14	Median amount by/at this merchant over the past 14 days
74	med_merch_30	Median amount by/at this merchant over the past 30 days
75	tot_merch_0	Total amount by/at this merchant over the past 0 days
76	tot_merch_1	Total amount by/at this merchant over the past 1 days
77	tot_merch_3	Total amount by/at this merchant over the past 3 days
78	tot_merch_7	Total amount by/at this merchant over the past 7 days

79	tot_merch_14	Total amount by/at this merchant over the past 14 days
80	tot_merch_30	Total amount by/at this merchant over the past 30 days
81	qaa_merch_0	Actual/average amount by/at this merchant over the past 0 days
82	qaa_merch_1	Actual/average amount by/at this merchant over the past 1 days
83	qaa_merch_3	Actual/average amount by/at this merchant over the past 3 days
84	qaa_merch_7	Actual/average amount by/at this merchant over the past 7 days
85	qaa_merch_14	Actual/average amount by/at this merchant over the past 14 days
86	qaa_merch_30	Actual/average amount by/at this merchant over the past 30 days
87	qam_merch_0	Actual/maximum amount by/at this merchant over the past 0 days
88	qam_merch_1	Actual/maximum amount by/at this merchant over the past 1 days
89	qam_merch_3	Actual/maximum amount by/at this merchant over the past 3 days
90	qam_merch_7	Actual/maximum amount by/at this merchant over the past 7 days
91	qam_merch_14	Actual/maximum amount by/at this merchant over the past 14 days
92	qam_merch_30	Actual/maximum amount by/at this merchant over the past 30 days
93	qamed_merch_0	Actual/median amount by/at this merchant over the past 0 days
94	qamed_merch_1	Actual/median amount by/at this merchant over the past 1 days
95	qamed_merch_3	Actual/median amount by/at this merchant over the past 3 days
96	qamed_merch_7	Actual/median amount by/at this merchant over the past 7 days
97	qamed_merch_14	Actual/median amount by/at this merchant over the past 14 days
98	qamed_merch_30	Actual/median amount by/at this merchant over the past 30 days
99	qat_merch_0	Actual/total amount by/at this merchant over the past 0 days
100	qat_merch_1	Actual/total amount by/at this merchant over the past 1 days
101	qat_merch_3	Actual/total amount by/at this merchant over the past 3 days
102	qat_merch_7	Actual/total amount by/at this merchant over the past 7 days
103	qat_merch_14	Actual/total amount by/at this merchant over the past 14 days
104	qat_merch_30	Actual/total amount by/at this merchant over the past 30 days
105	fre_merch_0	Number of transactions with this merchant over the past 0 days
106	fre_merch_1	Number of transactions with this merchant over the past 1 days
107	fre_merch_3	Number of transactions with this merchant over the past 3 days
108	fre_merch_7	Number of transactions with this merchant over the past 7 days
109	fre_merch_14	Number of transactions with this merchant over the past 14 days
110	fre_merch_30	Number of transactions with this merchant over the past 30 days
111	days_since_merch	Current date minus date of most recent transaction with same merchant

112	avg_cm_0	Average amount by/at this card at this merchant over the past 0 days
113	avg_cm_1	Average amount by/at this card at this merchant over the past 1 days
114	avg_cm_3	Average amount by/at this card at this merchant over the past 3 days
115	avg_cm_7	Average amount by/at this card at this merchant over the past 7 days
116	avg_cm_14	Average amount by/at this card at this merchant over the past 14 days
117	avg_cm_30	Average amount by/at this card at this merchant over the past 30 days
118	max_cm_0	Maximum amount by/at this card at this merchant over the past 0 days
119	max_cm_1	Maximum amount by/at this card at this merchant over the past 1 days
120	max_cm_3	Maximum amount by/at this card at this merchant over the past 3 days
121	max_cm_7	Maximum amount by/at this card at this merchant over the past 7 days
122	max_cm_14	Maximum amount by/at this card at this merchant over the past 14 days
123	max_cm_30	Maximum amount by/at this card at this merchant over the past 30 days
124	med_cm_0	Median amount by/at this card at this merchant over the past 0 days
125	med_cm_1	Median amount by/at this card at this merchant over the past 1 days
126	med_cm_3	Median amount by/at this card at this merchant over the past 3 days
127	med_cm_7	Median amount by/at this card at this merchant over the past 7 days
128	med_cm_14	Median amount by/at this card at this merchant over the past 14 days
129	med_cm_30	Median amount by/at this card at this merchant over the past 30 days
130	tot_cm_0	Total amount by/at this card at this merchant over the past 0 days
131	tot_cm_1	Total amount by/at this card at this merchant over the past 1 days
132	tot_cm_3	Total amount by/at this card at this merchant over the past 3 days
133	tot_cm_7	Total amount by/at this card at this merchant over the past 7 days

134	tot_cm_14	Total amount by/at this card at this merchant over the past 14 days
135	tot_cm_30	Total amount by/at this card at this merchant over the past 30 days
136	qaa_cm_0	Actual/average amount by/at this card at this merchant over the past 0 days
137	qaa_cm_1	Actual/average amount by/at this card at this merchant over the past 1 days
138	qaa_cm_3	Actual/average amount by/at this card at this merchant over the past 3 days
139	qaa_cm_7	Actual/average amount by/at this card at this merchant over the past 7 days
140	qaa_cm_14	Actual/average amount by/at this card at this merchant over the past 14 days
141	qaa_cm_30	Actual/average amount by/at this card at this merchant over the past 30 days
142	qamax_cm_0	Actual/maximum amount by/at this card at this merchant over the past 0 days
143	qamax_cm_1	Actual/maximum amount by/at this card at this merchant over the past 1 days
144	qamax_cm_3	Actual/maximum amount by/at this card at this merchant over the past 3 days
145	qamax_cm_7	Actual/maximum amount by/at this card at this merchant over the past 7 days
146	qamax_cm_14	Actual/maximum amount by/at this card at this merchant over the past 14 days
147	qamax_cm_30	Actual/maximum amount by/at this card at this merchant over the past 30 days
148	qamed_cm_0	Actual/median amount by/at this card at this merchant over the past 0 days
149	qamed_cm_1	Actual/median amount by/at this card at this merchant over the past 1 days
150	qamed_cm_3	Actual/median amount by/at this card at this merchant over the past 3 days
151	qamed_cm_7	Actual/median amount by/at this card at this merchant over the past 7 days
152	qamed_cm_14	Actual/median amount by/at this card at this merchant over the past 14 days
153	qamed_cm_30	Actual/median amount by/at this card at this merchant over the past 30 days
154	qat_cm_0	Actual/total amount by/at this card at this merchant over the past 0 days
155	qat_cm_1	Actual/total amount by/at this card at this merchant over the past 1 days

156	qat_cm_3	Actual/total amount by/at this card at this merchant over the past 3 days
157	qat_cm_7	Actual/total amount by/at this card at this merchant over the past 7 days
158	qat_cm_14	Actual/total amount by/at this card at this merchant over the past 14 days
159	qat_cm_30	Actual/total amount by/at this card at this merchant over the past 30 days
160	freq_cm_0	Number of transactions with this card at this merchant over the past 0 days
161	freq_cm_1	Number of transactions with this card at this merchant over the past 1 days
162	freq_cm_3	Number of transactions with this card at this merchant over the past 3 days
163	freq_cm_7	Number of transactions with this card at this merchant over the past 7 days
164	freq_cm_14	Number of transactions with this card at this merchant over the past 14 days
165	freq_cm_30	Number of transactions with this card at this merchant over the past 30 days
166	days_since_cm	Current date minus date of most recent transaction with same card at this merchant
167	avg_cz_0	Average amount by/at this card in this zip code over the past 0 days
168	avg_cz_1	Average amount by/at this card in this zip code over the past 1 days
169	avg_cz_3	Average amount by/at this card in this zip code over the past 3 days
170	avg_cz_7	Average amount by/at this card in this zip code over the past 7 days
171	avg_cz_14	Average amount by/at this card in this zip code over the past 14 days
172	avg_cz_30	Average amount by/at this card in this zip code over the past 30 days
173	max_cz_0	Maximum amount by/at this card in this zip code over the past 0 days
174	max_cz_1	Maximum amount by/at this card in this zip code over the past 1 days
175	max_cz_3	Maximum amount by/at this card in this zip code over the past 3 days
176	max_cz_7	Maximum amount by/at this card in this zip code over the past 7 days
177	max_cz_14	Maximum amount by/at this card in this zip code over the past 14 days

178	max_cz_30	Maximum amount by/at this card in this zip code over the past 30 days
179	med_cz_0	Median amount by/at this card in this zip code over the past 0 days
180	med_cz_1	Median amount by/at this card in this zip code over the past 1 days
181	med_cz_3	Median amount by/at this card in this zip code over the past 3 days
182	med_cz_7	Median amount by/at this card in this zip code over the past 7 days
183	med_cz_14	Median amount by/at this card in this zip code over the past 14 days
184	med_cz_30	Median amount by/at this card in this zip code over the past 30 days
185	tot_cz_0	Total amount by/at this card in this zip code over the past 0 days
186	tot_cz_1	Total amount by/at this card in this zip code over the past 1 days
187	tot_cz_3	Total amount by/at this card in this zip code over the past 3 days
188	tot_cz_7	Total amount by/at this card in this zip code over the past 7 days
189	tot_cz_14	Total amount by/at this card in this zip code over the past 14 days
190	tot_cz_30	Total amount by/at this card in this zip code over the past 30 days
191	qaa_cz_0	Actual/average amount by/at this card in this zip code over the past 0 days
192	qaa_cz_1	Actual/average amount by/at this card in this zip code over the past 1 days
193	qaa_cz_3	Actual/average amount by/at this card in this zip code over the past 3 days
194	qaa_cz_7	Actual/average amount by/at this card in this zip code over the past 7 days
195	qaa_cz_14	Actual/average amount by/at this card in this zip code over the past 14 days
196	qaa_cz_30	Actual/average amount by/at this card in this zip code over the past 30 days
197	qam_cz_0	Actual/maximum amount by/at this card in this zip code over the past 0 days
198	qam_cz_1	Actual/maximum amount by/at this card in this zip code over the past 1 days
199	qam_cz_3	Actual/maximum amount by/at this card in this zip code over the past 3 days
200	qam_cz_7	Actual/maximum amount by/at this card in this zip code over the past 7 days
201	qam_cz_14	Actual/maximum amount by/at this card in this zip code over the past 14 days

202	qam_cz_30	Actual/maximum amount by/at this card in this zip code over the past 30 days
203	qamed_cz_0	Actual/median amount by/at this card in this zip code over the past 0 days
204	qamed_cz_1	Actual/median amount by/at this card in this zip code over the past 1 days
205	qamed_cz_3	Actual/median amount by/at this card in this zip code over the past 3 days
206	qamed_cz_7	Actual/median amount by/at this card in this zip code over the past 7 days
207	qamed_cz_14	Actual/median amount by/at this card in this zip code over the past 14 days
208	qamed_cz_30	Actual/median amount by/at this card in this zip code over the past 30 days
209	qat_cz_0	Actual/total amount by/at this card in this zip code over the past 0 days
210	qat_cz_1	Actual/total amount by/at this card in this zip code over the past 1 days
211	qat_cz_3	Actual/total amount by/at this card in this zip code over the past 3 days
212	qat_cz_7	Actual/total amount by/at this card in this zip code over the past 7 days
213	qat_cz_14	Actual/total amount by/at this card in this zip code over the past 14 days
214	qat_cz_30	Actual/total amount by/at this card in this zip code over the past 30 days
215	fre_cz_0	Number of transactions with this card in this zip code over the past 0 days
216	fre_cz_1	Number of transactions with this card in this zip code over the past 1 days
217	fre_cz_3	Number of transactions with this card in this zip code over the past 3 days
218	fre_cz_7	Number of transactions with this card in this zip code over the past 7 days
219	fre_cz_14	Number of transactions with this card in this zip code over the past 14 days
220	fre_cz_30	Number of transactions with this card in this zip code over the past 30 days
221	days_since_cz	Current date minus date of most recent transaction with same card in this zip code
222	avg_cs_0	Average amount by/at this card in this state over the past 0 days
223	avg_cs_1	Average amount by/at this card in this state over the past 1 days
224	avg_cs_3	Average amount by/at this card in this state over the past 3 days
225	avg_cs_7	Average amount by/at this card in this state over the past 7 days

226	avg_cs_14	Average amount by/at this card in this state over the past 14 days
227	avg_cs_30	Average amount by/at this card in this state over the past 30 days
228	max_cs_0	Maximum amount by/at this card in this state over the past 0 days
229	max_cs_1	Maximum amount by/at this card in this state over the past 1 days
230	max_cs_3	Maximum amount by/at this card in this state over the past 3 days
231	max_cs_7	Maximum amount by/at this card in this state over the past 7 days
232	max_cs_14	Maximum amount by/at this card in this state over the past 14 days
233	max_cs_30	Maximum amount by/at this card in this state over the past 30 days
234	med_cs_0	Median amount by/at this card in this state over the past 0 days
235	med_cs_1	Median amount by/at this card in this state over the past 1 days
236	med_cs_3	Median amount by/at this card in this state over the past 3 days
237	med_cs_7	Median amount by/at this card in this state over the past 7 days
238	med_cs_14	Median amount by/at this card in this state over the past 14 days
239	med_cs_30	Median amount by/at this card in this state over the past 30 days
240	tot_cs_0	Total amount by/at this card in this state over the past 0 days
241	tot_cs_1	Total amount by/at this card in this state over the past 1 days
242	tot_cs_3	Total amount by/at this card in this state over the past 3 days
243	tot_cs_7	Total amount by/at this card in this state over the past 7 days
244	tot_cs_14	Total amount by/at this card in this state over the past 14 days
245	tot_cs_30	Total amount by/at this card in this state over the past 30 days
246	qaa_cs_0	Actual/average amount by/at this card in this state over the past 0 days
247	qaa_cs_1	Actual/average amount by/at this card in this state over the past 1 days
248	qaa_cs_3	Actual/average amount by/at this card in this state over the past 3 days
249	qaa_cs_7	Actual/average amount by/at this card in this state over the past 7 days
250	qaa_cs_14	Actual/average amount by/at this card in this state over the past 14 days
251	qaa_cs_30	Actual/average amount by/at this card in this state over the past 30 days
252	qamax_cs_0	Actual/maximum amount by/at this card in this state over the past 0 days
253	qamax_cs_1	Actual/maximum amount by/at this card in this state over the past 1 days

254	qamax_cs_3	Actual/maximum amount by/at this card in this state over the past 3 days
255	qamax_cs_7	Actual/maximum amount by/at this card in this state over the past 7 days
256	qamax_cs_14	Actual/maximum amount by/at this card in this state over the past 14 days
257	qamax_cs_30	Actual/maximum amount by/at this card in this state over the past 30 days
258	qamed_cs_0	Actual/median amount by/at this card in this state over the past 0 days
259	qamed_cs_1	Actual/median amount by/at this card in this state over the past 1 days
260	qamed_cs_3	Actual/median amount by/at this card in this state over the past 3 days
261	qamed_cs_7	Actual/median amount by/at this card in this state over the past 7 days
262	qamed_cs_14	Actual/median amount by/at this card in this state over the past 14 days
263	qamed_cs_30	Actual/median amount by/at this card in this state over the past 30 days
264	qat_cs_0	Actual/total amount by/at this card in this state over the past 0 days
265	qat_cs_1	Actual/total amount by/at this card in this state over the past 1 days
266	qat_cs_3	Actual/total amount by/at this card in this state over the past 3 days
267	qat_cs_7	Actual/total amount by/at this card in this state over the past 7 days
268	qat_cs_14	Actual/total amount by/at this card in this state over the past 14 days
269	qat_cs_30	Actual/total amount by/at this card in this state over the past 30 days
270	freq_cs_0	Number of transactions with this card in this state over the past 0 days
271	freq_cs_1	Number of transactions with this card in this state over the past 1 days
272	freq_cs_3	Number of transactions with this card in this state over the past 3 days
273	freq_cs_7	Number of transactions with this card in this state over the past 7 days
274	freq_cs_14	Number of transactions with this card in this state over the past 14 days
275	freq_cs_30	Number of transactions with this card in this state over the past 30 days

276	days_since_cs	Current date minus date of most recent transaction with same card in this state
277	vcv_nc0_nc7	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 7 days
278	vcv_nc0_nc14	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 14 days
279	vcv_nc0_nc30	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 30 days
280	vcv_nc0_nm7	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 7 days
281	vcv_nc0_nm14	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 14 days
282	vcv_nc0_nm30	Number of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 30 days
283	vcv_nc0_ac7	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 7 days
284	vcv_nc0_ac14	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 14 days
285	vcv_nc0_ac30	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 30 days
286	vcv_nc0_am7	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 7 days
287	vcv_nc0_am14	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 14 days
288	vcv_nc0_am30	Number of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 30 days
289	vcv_nc1_nc7	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 7 days
290	vcv_nc1_nc14	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 14 days

291	vcv_nc1_nc30	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 30 days
292	vcv_nc1_nm7	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 7 days
293	vcv_nc1_nm14	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 14 days
294	vcv_nc1_nm30	Number of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 30 days
295	vcv_nc1_ac7	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 7 days
296	vcv_nc1_ac14	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 14 days
297	vcv_nc1_ac30	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 30 days
298	vcv_nc1_am7	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 7 days
299	vcv_nc1_am14	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 14 days
300	vcv_nc1_am30	Number of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 30 days
301	vcv_nm0_nc7	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 7 days
302	vcv_nm0_nc14	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 14 days
303	vcv_nm0_nc30	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 30 days
304	vcv_nm0_nm7	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 7 days

305	vcv_nm0_nm14	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 14 days
306	vcv_nm0_nm30	Number of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 30 days
307	vcv_nm0_ac7	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 7 days
308	vcv_nm0_ac14	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 14 days
309	vcv_nm0_ac30	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 30 days
310	vcv_nm0_am7	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 7 days
311	vcv_nm0_am14	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 14 days
312	vcv_nm0_am30	Number of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 30 days
313	vcv_nm1_nc7	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 7 days
314	vcv_nm1_nc14	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 14 days
315	vcv_nm1_nc30	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 30 days
316	vcv_nm1_nm7	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 7 days
317	vcv_nm1_nm14	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 14 days
318	vcv_nm1_nm30	Number of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 30 days

319	vcv_nm1_ac7	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 7 days
320	vcv_nm1_ac14	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 14 days
321	vcv_nm1_ac30	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 30 days
322	vcv_nm1_am7	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 7 days
323	vcv_nm1_am14	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 14 days
324	vcv_nm1_am30	Number of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 30 days
325	vcv_ac0_nc7	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 7 days
326	vcv_ac0_nc14	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 14 days
327	vcv_ac0_nc30	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same card over the past 30 days
328	vcv_ac0_nm7	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 7 days
329	vcv_ac0_nm14	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 14 days
330	vcv_ac0_nm30	Amount of transactions with same card over the past 0 days divided by average daily number of transactions with same merchant over the past 30 days
331	vcv_ac0_ac7	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 7 days
332	vcv_ac0_ac14	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 14 days

333	vcv_ac0_ac30	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same card over the past 30 days
334	vcv_ac0_am7	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 7 days
335	vcv_ac0_am14	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 14 days
336	vcv_ac0_am30	Amount of transactions with same card over the past 0 days divided by average daily amount of transactions with same merchant over the past 30 days
337	vcv_ac1_nc7	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 7 days
338	vcv_ac1_nc14	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 14 days
339	vcv_ac1_nc30	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same card over the past 30 days
340	vcv_ac1_nm7	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 7 days
341	vcv_ac1_nm14	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 14 days
342	vcv_ac1_nm30	Amount of transactions with same card over the past 1 days divided by average daily number of transactions with same merchant over the past 30 days
343	vcv_ac1_ac7	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 7 days
344	vcv_ac1_ac14	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 14 days
345	vcv_ac1_ac30	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same card over the past 30 days
346	vcv_ac1_am7	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 7 days

347	vcv_ac1_am14	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 14 days
348	vcv_ac1_am30	Amount of transactions with same card over the past 1 days divided by average daily amount of transactions with same merchant over the past 30 days
349	vcv_am0_nc7	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 7 days
350	vcv_am0_nc14	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 14 days
351	vcv_am0_nc30	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same card over the past 30 days
352	vcv_am0_nm7	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 7 days
353	vcv_am0_nm14	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 14 days
354	vcv_am0_nm30	Amount of transactions with same merchant over the past 0 days divided by average daily number of transactions with same merchant over the past 30 days
355	vcv_am0_ac7	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 7 days
356	vcv_am0_ac14	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 14 days
357	vcv_am0_ac30	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same card over the past 30 days
358	vcv_am0_am7	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 7 days
359	vcv_am0_am14	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 14 days
360	vcv_am0_am30	Amount of transactions with same merchant over the past 0 days divided by average daily amount of transactions with same merchant over the past 30 days

361	vcv_am1_nc7	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 7 days
362	vcv_am1_nc14	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 14 days
363	vcv_am1_nc30	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same card over the past 30 days
364	vcv_am1_nm7	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 7 days
365	vcv_am1_nm14	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 14 days
366	vcv_am1_nm30	Amount of transactions with same merchant over the past 1 days divided by average daily number of transactions with same merchant over the past 30 days
367	vcv_am1_ac7	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 7 days
368	vcv_am1_ac14	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 14 days
369	vcv_am1_ac30	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same card over the past 30 days
370	vcv_am1_am7	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 7 days
371	vcv_am1_am14	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 14 days
372	vcv_am1_am30	Amount of transactions with same merchant over the past 1 days divided by average daily amount of transactions with same merchant over the past 30 days

8.2 Data Quality Report (DQR)

DATASET DESCRIPTION

Name of Dataset: Card Transactions
Description: This dataset gives information on real credit card transactions, including card number, merchant information and transaction information consisting of 10 fields and 96,753 records. It is provided by a government organization purchasing cards.

SUMMARY OF ALL FIELDS

Numerical Fields (2 Fields)

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Mean	STD	Min	Max
1	Amount	Numerical	96753	100.00%	34909	0	4.28E+02	10006.1403	1.00E-02	3.10E+06
2	Date	Categorical	96753	100.00%	365	0	2/28/10	-	-	12/31/10

Categorical Fields (8 Fields)

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Most Common Field Name
1	Recnum	Categorical	96753	100.00%	96753	0	All Different
2	Cardnum	Categorical	96753	100.00%	1645	0	5142148452
3	Merchnum	Categorical	93378	96.51%	13092	0	930090121224
4	Merch description	Categorical	96753	100.00%	13126	0	GSA-FSS-ADV
5	Merch state	Categorical	95558	98.76%	228	0	TN
6	Merch zip	Categorical	92097	95.19%	4568	0	38118
7	Date	Categorical	96753	100.00%	365	0	2/28/10
8	Fraud	Categorical	96753	100.00%	2	95694	0

FIELD DESCRIPTION

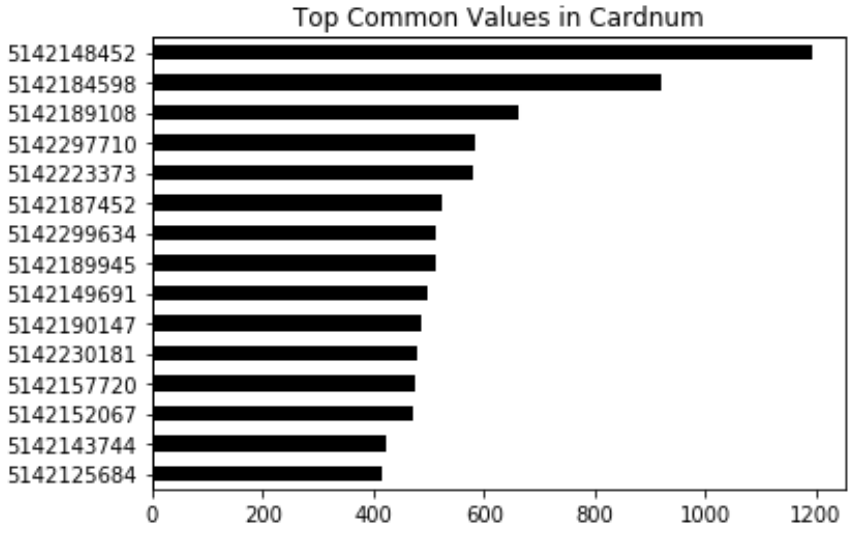
Field 1 Recnum

Field Name: Recnum
Field Type: Categorical
Description: Track data order
Most Common Values: No missing value and all values are different

Field 2 Cardnum

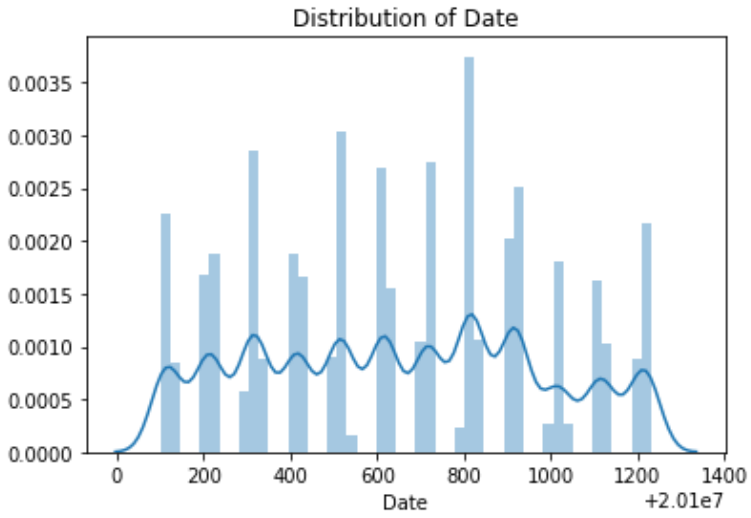
Field Name: Cardnum
Field Type: Categorical
Description: Credit card number
Most Common Values:

5142148452	1192
5142184598	921
5142189108	663
5142297710	583
5142223373	579
5142187452	526
5142299634	515
5142189945	512
5142149691	497
5142190147	488
5142230181	479
5142157720	475
5142152067	473
5142143744	422
5142125684	415



Field 3 Date

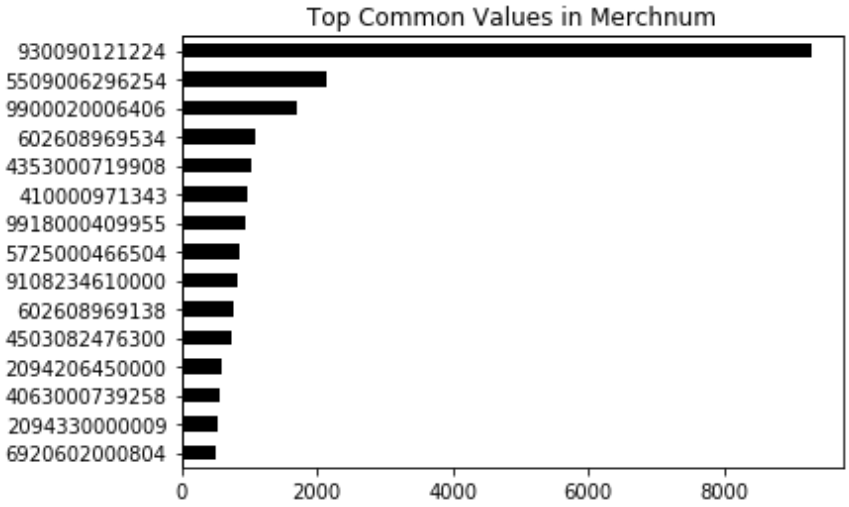
Field Name: Date
Field Type: Numerical, continuous
Description: Date of transaction
Distribution:



Field 4 Merchnum

Field Name: Merchnum
Field Type: Categorical
Description: Merchant number
Most Common Values:

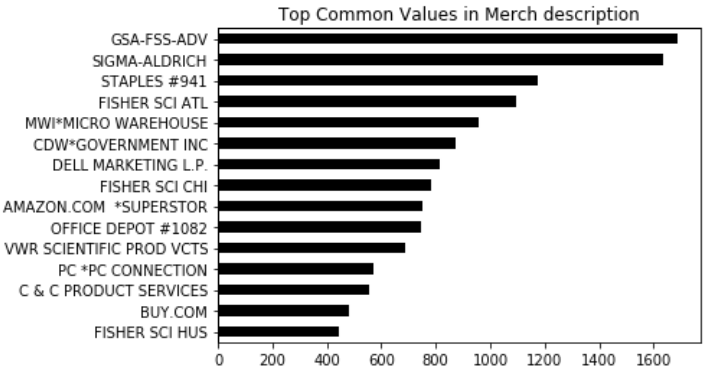
930090121224	9310
5509006296254	2131
9900020006406	1714
602608969534	1092
4353000719908	1020
410000971343	982
9918000409955	956
5725000466504	872
9108234610000	817
602608969138	783
4503082476300	746
2094206450000	590
4063000739258	568
2094330000009	533
6920602000804	523



Field 5 Merch description

Field Name: Merch description
Field Type: Categorical
Description: Merchant description
Most Common Values:

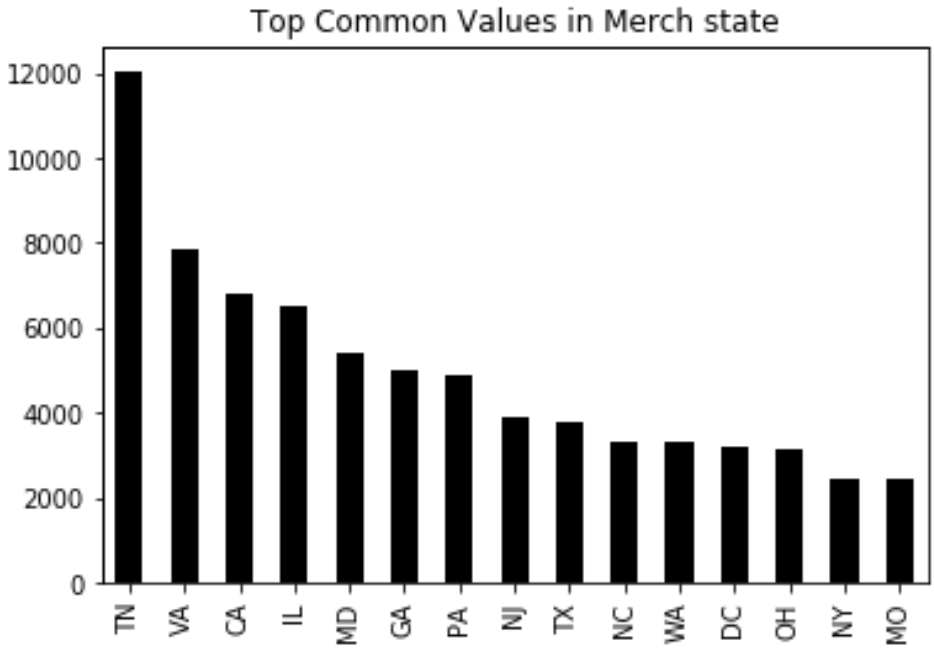
GSA-FSS-ADV	1688
SIGMA-ALDRICH	1635
STAPLES #941	1174
FISHER SCI ATL	1093
MWI*MICRO WAREHOUSE	958
CDW*GOVERNMENT INC	872
DELL MARKETING L.P.	816
FISHER SCI CHI	783
AMAZON.COM *SUPERSTOR	750
OFFICE DEPOT #1082	748
VWR SCIENTIFIC PROD VCTS	688
PC *PC CONNECTION	570
C & C PRODUCT SERVICES	558
BUY.COM	481
FISHER SCI HUS	442



Field 6 Merch state

Field Name: Merch state
Field Type: Categorical
Description: State that merchant is from
Most Common Values:

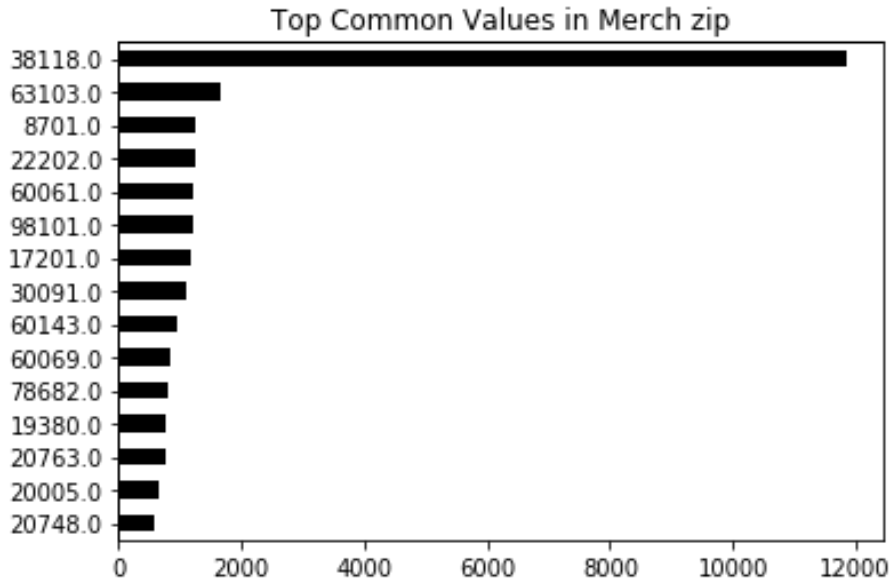
TN	12035
VA	7872
CA	6817
IL	6508
MD	5398
GA	5025
PA	4899
NJ	3912
TX	3790
NC	3322
WA	3300
DC	3208
OH	3131
NY	2430
MO	2420



Field 7 Merch zip

Field Name: Merch zip
Field Type: Categorical
Description: Zip code of merchant's location
Most Common Values:

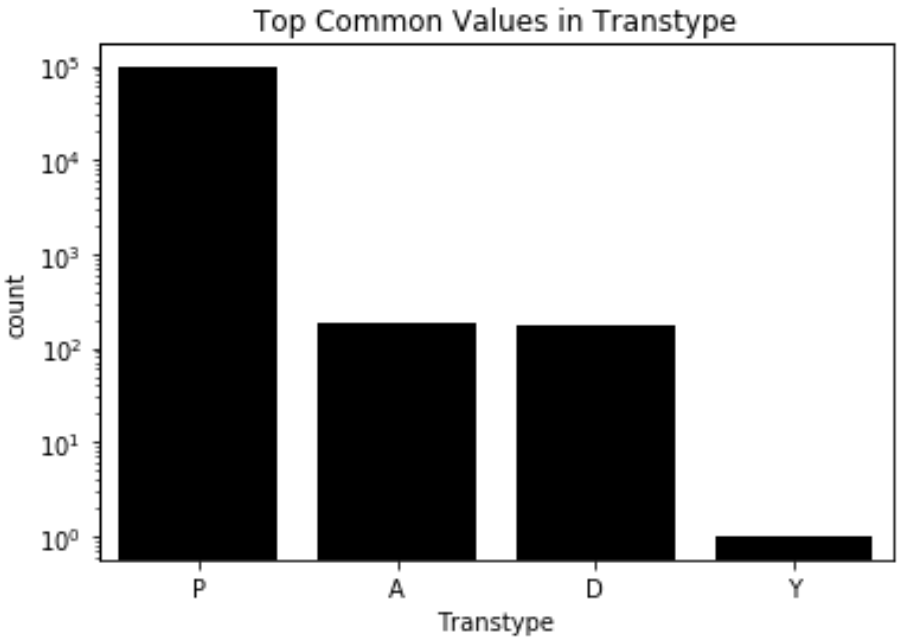
```
38118.0    11868
63103.0     1650
8701.0      1267
22202.0     1250
60061.0     1221
98101.0     1197
17201.0     1180
30091.0     1092
60143.0      942
60069.0      826
78682.0      817
19380.0      769
20763.0      749
20005.0      648
20748.0      592
```



Field 8 Transtype

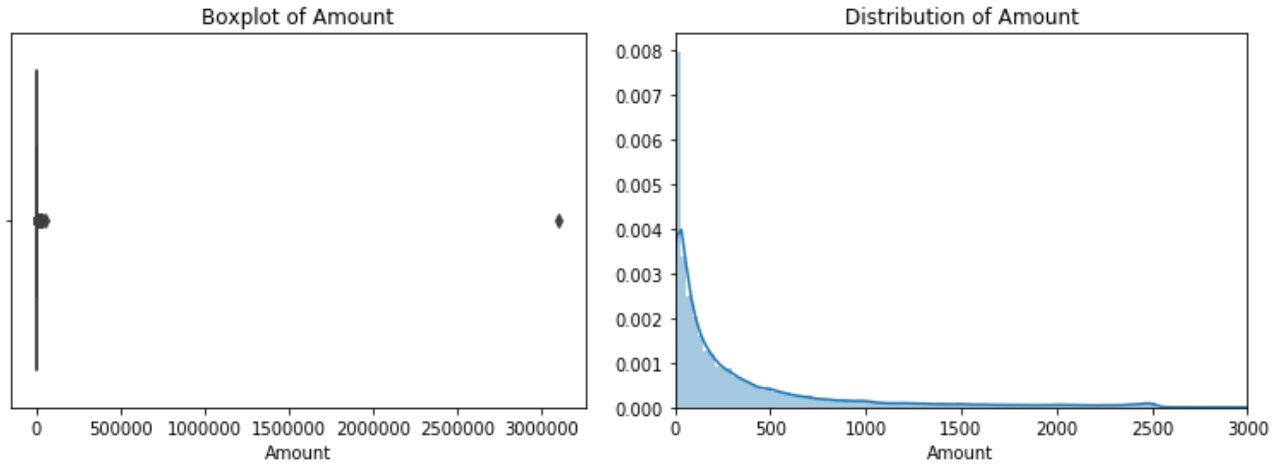
Field Name: Transtype
Field Type: Categorical
Description: Transaction type
Most Common Values:

```
P    96398
A      181
D      173
Y        1
```



Field 9 Amount

Field Name: Amount
Field Type: Numerical, continuous
Description: Transaction Amount
Distribution:



Field 10 Fraud

Field Name: Fraud
Field Type: Categorical
Description: Whether the record is a fraud
Most Common Values:

```
0    95694
1     1059
```