REPORT

BUILDING SUPERVISED MODEL FOR CARD TRANSACTIONS

Prepared for:

Professor Stephen Coggeshall

Prepared by:

CHEN, Siqi

GUO, Qianfei

HUNG, Kai-Ling

JIA, Xuerong

LIN, Xiaoyan

TSAI, Pei Yu

March 28, 2019

Table of Content

1 Executive Summary

1.1 Project Goal

The goal of the project is to build and test different machine learning algorithms based on card transaction data between January 1, 2010 and December 31, 2010 and find out the optimal model to predict future fraud.

1.2 Steps Performed

To build and select an optimal supervised fraud model based on the card transaction data, we first cleaned the data, and then built 269 variables incorporating amount information, card information and transaction location information. We used feature selection methods to reduce the number of variables to 20. We split the card transaction data and used the data in November and December 2010 as out out-of-time data. 75% of the transactions from January to October 2010 were randomly selected to be used as training data and the remaining 25% used as testing data. After that, we built four supervised fraud models using four different types of machine learning algorithms: Logistic Regression, Neural Nets, Random Forests, and Boosting Trees, and found that Random Forests outperforms the rest three.

1.3 Conclusions

We concluded that Random Forest is superior to the rest three models we built. At 3% population cut-off, the result is FDR@3% of 99.0% for the training dataset and FDR@3% of 85.3% for the testing dataset. For the OOT dataset, we got FDR@3% of 39.1%.

Due to limited time and resources, we were only able to build four different machine learning models. Given more time, we would try other machine learning techniques such as Decision Trees, Bagging, Support Vector Machine, etc. We might find a model that is a better fit than Random Forest to predict future fraud.

2 Description of Data

2.1 Data Source

The Card Transactions Data was provided by Professor Stephen Coggeshall on February 21, 2019. This dataset contains 10 fields and 96,753 records. Each record in this dataset represents a card payment between January 1, 2010 and December 31, 2010. The dataset includes information such as card numbers, dates of transactions, merchant numbers, merchant descriptions, merchant state, merchant zip, transaction type, payment amount in US dollars and fraud scores.

2.2 Summary of Numerical Fields

In this dataset there is only one numerical field. Refer to Table 1 below for basic statistics for this field.

Table 1

field	count	% Populated	mean	std	min	25%	50%	75%	max
Amount	96,753	100.00%	427.9	10,006	0.01	33.48	137.98	428.2	3,102,046

2.3 Summary of Categorical Fields

There are eight categorical fields in this data set. Refer to Table 2 below for basic statistics for this field.

Table 2

Field	Count	% Populated	Unique Value	Most Common
Recnum	96,753	100.00%	96,753	Uniform Distributed
Cardnum	96,753	100.00%	1,645	5142148452
Date	96,753	100.00%	365	2/28/10
Merchnum	93,378	96.51%	13,092	9.3009E+11
Merch description	96,753	100.00%	13,126	GSA-FSS-ADV
Merch state	95,558	98.76%	228	TN
Merch zip	92,097	95.19%	4,568	38118
Transtype	96,753	100.00%	4	P

2.4 Response Field

The column labelled “Fraud” represents a response field in this dataset. There are two types, “0” and “1”. Type “0” categorized as non-fraud transaction and Type “1” represents fraud transaction. Summary information is shown as Table 3.

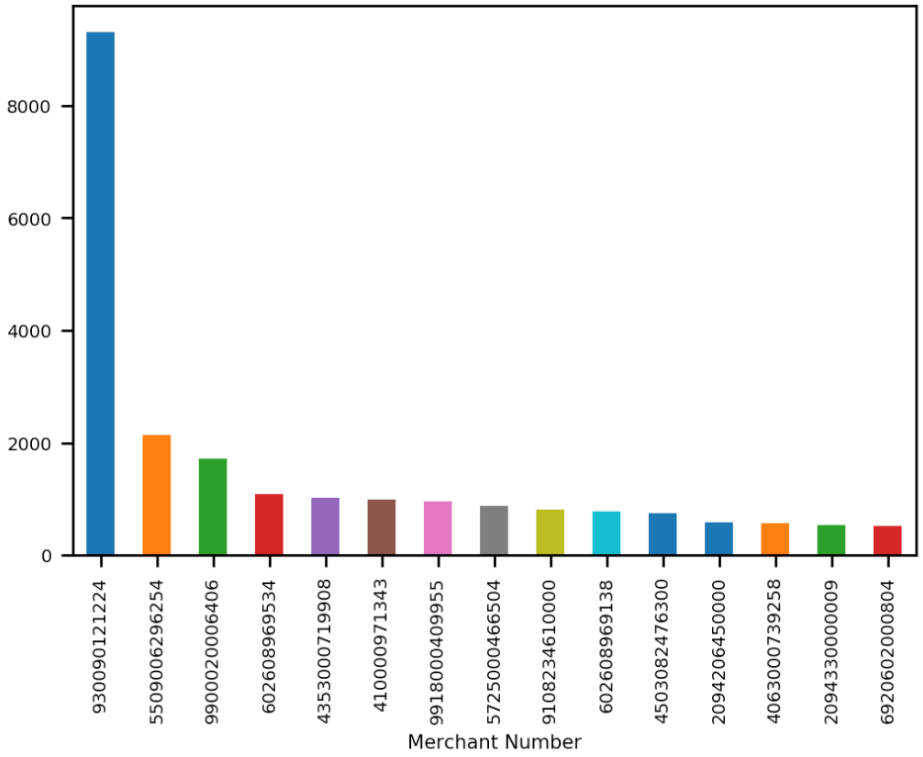
Table 3

Field	records	% Populated	Non-Fraud	Fraud
Fraud	96,753	100.00%	95694 (98.9%)	1059 (1.1%)

2.5 Overview of Card Transactions Data

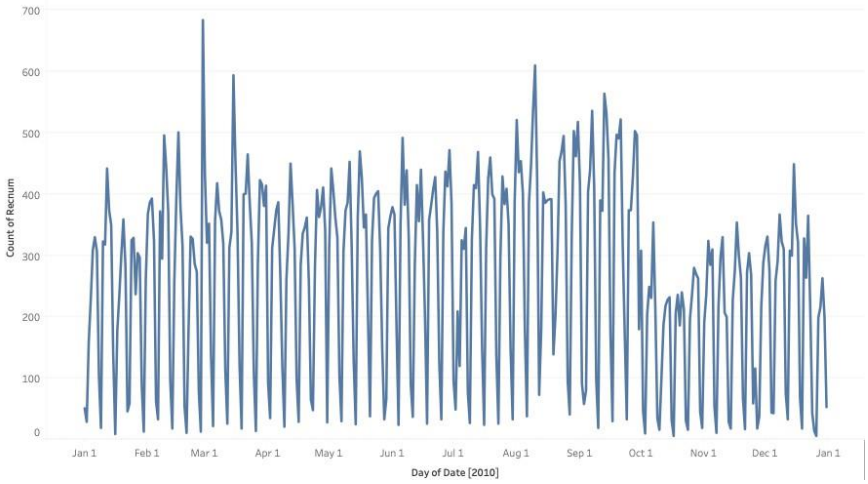
The line charts and bar charts below give a high-level understanding of some important features of Card Transaction Data. Please refer to the Appendix 2 for a complete Data Quality Review Report.

1. Merchnum
   Merchnum represents merchant number. The graph below shows the count of the top 15 most common merchant numbers.

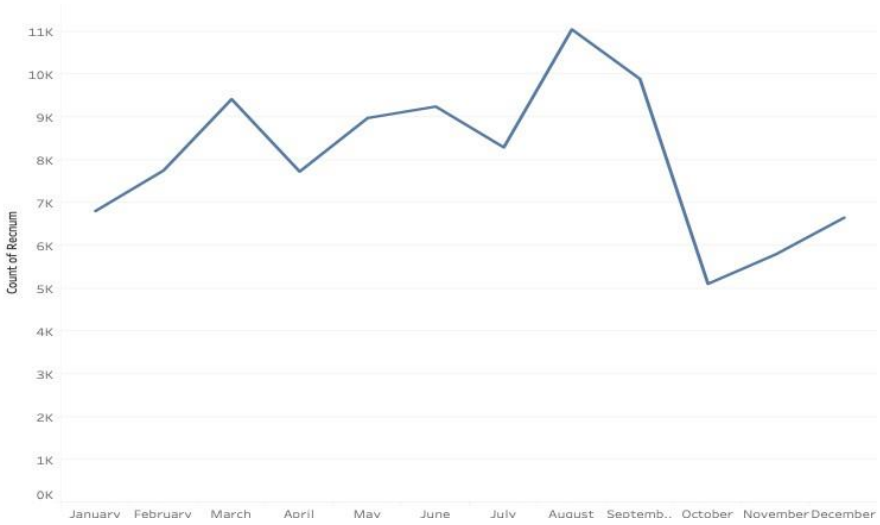


2. Date

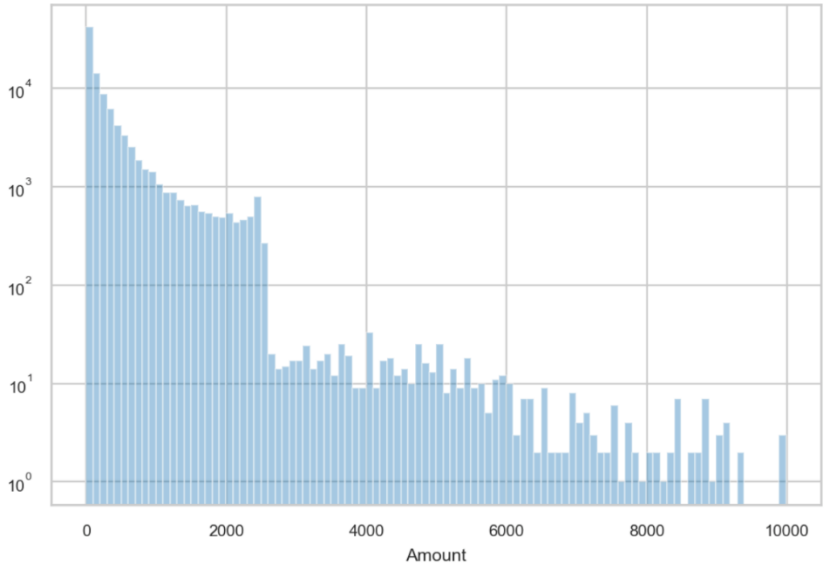
   The Date field contains 365 unique values, which are 365 different dates from January 1, 2010 to December 31, 2010. The line chart below represents number of payment transactions by date. We can see the weekly seasonality from this chart.



   The line chart below represents number of transactions by month. A sharp decrease is noted in October.



3. Amount
   The graph below shows the log scale distribution of transaction amount with amount limited to 10,000.



3 Data Cleaning

3.1 Removing Outlier & Filter Transtype

As a first step of data cleaning, we excluded one extremely large value in the Amount Field: 3,102,045.53. We further removed the column named Transtype because there is only one value “P” in this column representing purchase.

3.2 Filling Missing Fields

Next, we filled missing values in the three fields, Merch state, Merch zip and Merchnum, with innocuous values according to following rules.

1. Merch state
   - Grouped by Merch zip, filled the missing values with most frequent Merch state of the group; and
   - Filled the remaining missing values with TN, the most frequent Merch state in the dataset.

2. Merch zip
   - Grouped by Merchnum, filled the missing values with most frequent Merch zip in the group;
   - Grouped by Merch description, filled the missing values with the most frequent Merch zip in the group;
   - Grouped by Merch state, filled the missing values with the most frequent Merch zip in the group; and
   - Filled the remaining missing values with the most frequent Merch zip in the whole dataset.

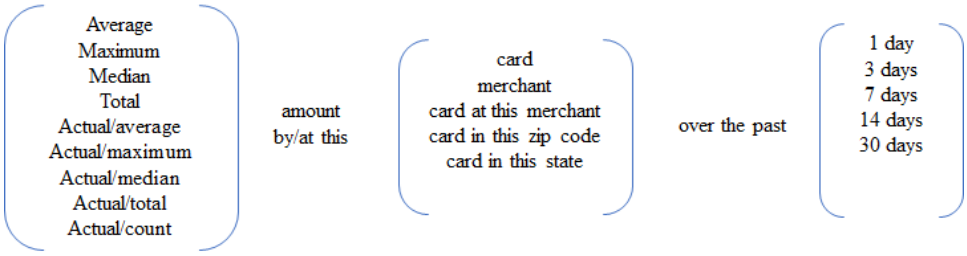
3. Merchnum
   - Grouped by Merch description, filled the missing values with the most frequent Merchnum in the group;
   - Grouped by Merch zip, filled the missing values with most frequent Merchnum of the group; and
   - Grouped Merch state, filled the remaining missing values with the most frequent Merchnum in the group.

4 Candidate Variables

The goal is to build a supervised fraud model on the card transaction data. To that end, we built 269 variables and used feature selection methods illustrated in the Section 5 to reduce the number of variables. We built variables based on transaction amount, transaction frequency, days since last transaction and velocity change of transaction.

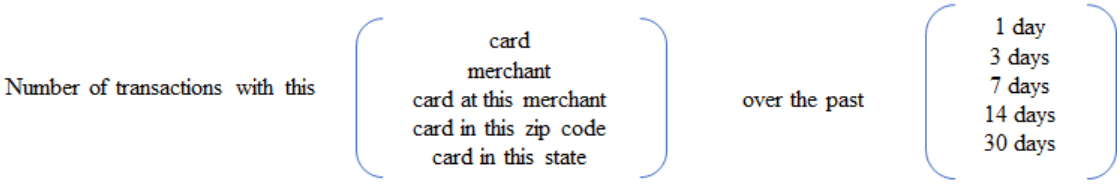
4.1 Amount Variables

One indication of credit card fraud is unusual transaction amount. For example, unusually large payment amounts for the same or different merchants. To build quality amount variables, we wanted to build variables incorporating amount information, card information and transaction location information over a certain period. The following chart illustrates how we created the amount variables. For example, one of the variables is created by grouping all records by card number and calculating the average transaction amount during the past 1 day. Another example would be grouping all records by card number and merchant number and calculating the total transaction amount of this group over the past month. As shown below, we created 9 x 5 x 5 = 225 variables.



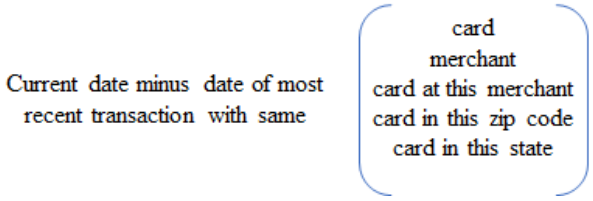
4.2 Frequency Variables

Another indication of credit card fraud is unusual transaction frequency. For example, bursts of activities at different merchants or transactions at merchants where the card had not been used before. To build quality frequency variables, we want to build variables that incorporate frequency information, card information and transaction location information over a certain period. The following chart illustrates how we created the frequency variables. For example, one of the variables is created by grouping all records by card number, counting the number of transactions during the past 1 day. Another example would be grouping all records by card number and merchant number, counting the number of transactions over the past month. As shown below, we created 5 x 5 = 25 variables.



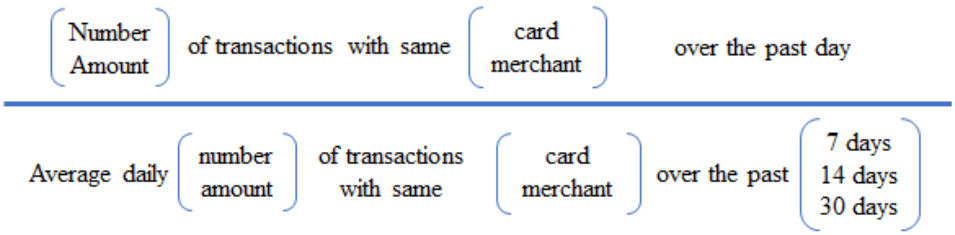
4.3 Days Variables

Unusually long or short duration of time since the last transaction could also be a signal of credit card fraud. To build quality Days variables, we want to build variables that incorporate the number of days since the last transaction, card information and transaction location information. The following chart illustrates how we created the Days variables. We created the variable by subtracting the date of most recent transaction from the current date for the same card number. As shown below, we created 5 variables.



4.4 Velocity Change Variables

We also wanted to compare the frequency and amount of transactions with the same card or at the same merchant over the past day to the frequency and amount over the past seven, 14 and 30 days, respectively. The following chart illustrates how we created the Velocity Change variables. For example, we grouped all records by card number and calculated the total transaction amount during the past day. Then we divided the result by the daily average amount for the same card number over the past seven days. As shown below, we created 2 x 2 x 3 = 12 variables.



A list of all candidate variables is included in Appendix 1 – Candidate Variables.

5 Feature Selection Process

After creating 269 variables, we conducted filter method through KS and FDR, and wrapper method for feature selection to reduce dimensionality, correlation. We reduced the number of variables to 20.

5.1 Filter method by KS and FDR

1. KS is a robust measure of how well good and bad distribution are separated. We used KS as a guide to evaluate how well a certain variable can separate Fraud and non-Fraud records. The higher the KS, the better the variable serve as an important feature.
2. Fraud Detection Rate (FDR) represents what percent of frauds are caught at a certain examination cutoff location. Here we used 3% FDR, meaning that after rank order records based on a particular variable, what percent of Fraud we can detect by looking at the top 3% of the population. The higher the FDR, the better the variable serve as an important feature.
3. After we got KS and 3% FDR for each variable, we first ranked all variables according to their KS ranking among 269 variables and got KS ranking number. Then we ranked all variables again according to their FDR and got FDR ranking number. For each record, we calculated the average of KS ranking number and FDR ranking number to have a combined score.
4. We sorted all variables based on the combined score and select only top 50% of the variables (135 variables).

5.2 Wrapper method by RFECV

After conducting Filter Method to select 135 candidate variables, we performed Feature Elimination Cross-Validation (RFECV) to select the best 20 features. RFECV is a feature selection method that removes the weakest features using cross-validation recursively to find the best features. Through the selection process, it also eliminates correlation that might exist between variables. We conducted the following steps:

1 Performed RFECV to find important features using 5-fold cross-validation. Since the number of records for credit card transaction data is relatively small, instead of directly using the best 14 features generated by this algorithm;
2 Ranked all variables, and kept the top 45 variables;
3 Performed RFECV again on these 45 variables. Once again, instead of directly using the best 12 features generated by the algorithm; and

4 Ranked the 45 variables, kept the top 20 as our final 20 features.

Top 20 variables	
Actual/mean_Cardnum_30d	Actual/mean_Merchnum_14d
Actual/mean_Merchnum_30d	Actual/mean_Merchnum_7d
Actual/median_Merchnum_7d	Actual/sum_Cardnum_Merchzip_3d
Actual/sum_Cardnum_Merchnum_3d	Days_since_per_Cardnum
Days_since_per_Cardnum_Merchstate	count_Cardnum_1d
count_Cardnum_7d	count_Cardnum_Merchstate_3d
count_Cardnum_Merchstate_7d	count_Cardnum_Merchzip_3d
count_Cardnum_Merchnum_3d	count_Cardnum_Merchnum_7d
mean_Cardnum_1d	Actual/count_Cardnum_1d
Actual/count_Cardnum_30d	mean_Cardnum_30d

6 Model Algorithms

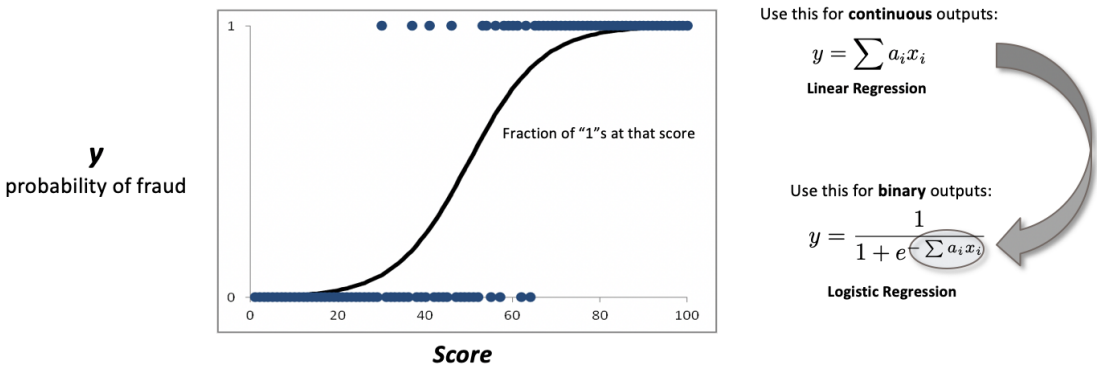
We split the dataset into three parts before training the models: training data, testing data, and out-of-time data. In our case, we took the last two months, between November 1, 2010 and December 31, 2010, as our out-of-time data. For the transactions between January 1, 2010 and October 31, 2010, we randomly selected 75% of them to be our training data and the remaining 25% was used as testing data. In our models, we used training set to fit the model, and then tune the model according to the testing set. Finally, we tested our model with out-of-time data.

We tried four different machine learning algorithms to build supervised models. In each model, we used the probability for data belonging to class 1 as fraud score, and then we sorted all the records by scores in descending order. After getting the top 3% transactions for each model, we summarized the fraud detection rate shown as below:

	FDR @ 3%		
	Training	Testing	Out of Time
Logistic Regression	59.3	56.1	23.9
Neural Net	71.4	65.1	31.6
Random Forest	99.0	85.3	39.1
Boosted Trees	84.0	76.4	33.0

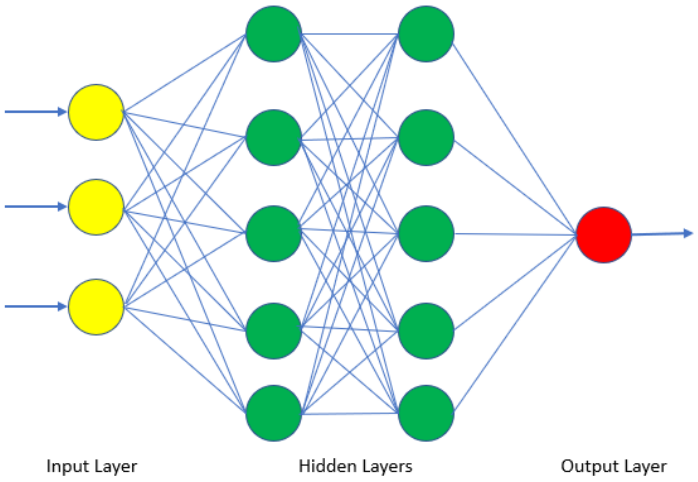
6.1 Logistic Regression

Logistic regression is an algorithm suitable to apply for data with binary output. Based on the input information, it transforms model output using the logistic sigmoid function to return a probability value which can then be mapped to the binary classes. For this card transaction project, we ranked order all the records based on the probability for that record to be fraudulent and evaluate FDR for top 3% of the listed records. We ran logistic regression ten times and take the average FDR@3%. The result is FDR@3% of 59.3% for the training dataset and FDR@3% of 56.1% for the testing dataset. For the OOT dataset, we got FDR@3% of 23.9%.



6.2 Neural Nets

A neural network has an input layer, one or more hidden layers, and an output layer. For this model, each node in the input layer consists the information of all selected variables for each record. The nodes in the output layer are either 0 or 1; 0 represents non-fraudulent transaction, and 1 represents fraud. For the hidden layer, we chose 1 hidden layer with 5 nodes at the beginning. Then we slightly increased the size of hidden layers and hidden nodes to 2 layers with 5 nodes in each layer.



We ran ten times and took the average FDR@3%. The result is FDR@3% of 71.4% for the training dataset and FDR@3% of 65.1% for the testing dataset. For the OOT dataset, we got FDR@3% of 31.6%.

6.3 Random Forests

Random forests provide an improvement over bagged trees by way of a random small tweak that decorrelates the trees. As in bagging, we build a number forest of decision trees on bootstrapped training samples. After that, different from bagging, each time a split in a tree is considered, a random sample of predictors is chosen as split candidates from the full set of predictors. Finally, we combine all the results by averaging or voting.

We ran ten times using different parameters and achieved the best performance on FDR@3% using the following parameters:

- n.trees = 20
  n.trees is the total number of trees to fit. This is equivalent to the number of iterations and the number of basic functions in the additive expansion.
- Min.sample.leaf= 5
  It is the minimum number of samples in newly created leaves. A split is discarded if after the split, one of the leaves would contain less then min.samples.leaf samples.
- interaction.depth = 15
  Interaction. depth parameter as a number of splits it has to perform on a tree (starting from a single node).

We got FDR@3% of 99.0% for the training dataset and FDR@3% of 85.3% for the testing dataset. For the OOT dataset, we got FDR@3% of 39.1%.

6.4 Boosting Trees

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
We ran ten times using different parameters and achieved the best performance on FDR@3% using the following parameters:

- n.trees = 190.
  n.trees is the total number of trees to fit. This is equivalent to the number of iterations and the number of basic functions in the additive expansion.
- shrinkage = 0.001.
  It is a shrinkage parameter applied to each tree in the expansion. Also known as the learning rate or step-size reduction.
- interaction.depth = 5
  Interaction. depth parameter as a number of splits it has to perform on a tree (starting from a single node).

Using Gradient boosting, we got FDR@3% of 84.0% for the training dataset and FDR@3% of 76.4% for the testing dataset. For the OOT dataset, we got FDR@3% of 33.0%.

7 Results

According to the analysis above, we concluded that Random Forest is superior to the rest three models we built. The following tables summarize both bin statistics and cumulative statistics when random forest is applied to detect fraud from 1% to 20% population in training, testing and Out-of-time datasets. Each table contains the information about the number of goods, bads, percentage of goods, percentage of bads, cumulative goods, cumulative bads, FDR, KS and false positive ratio. At 3% population cut-off, the result is FDR@3% of 99.0% for the training dataset and FDR@3% of 85.3% for the testing dataset. For the OOT dataset, we got FDR@3% of 39.1%.

Training	# Records	# Goods	# Bads	Fraud Rate
	62977	62316	661	0.0105

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	629	81	548	0.1	82.9	629	81	548	0.1	82.9	82.8	0.0
2	630	543	87	0.9	13.2	1259	624	635	1.0	96.1	95.1	0.0
3	630	611	19	1.0	2.9	1889	1235	654	2.0	98.9	97.0	0.0
4	630	625	5	1.0	0.8	2519	1860	659	3.0	99.7	96.7	0.0
5	629	629	0	1.0	0.0	3148	2489	659	4.0	99.7	95.7	0.0
6	630	628	2	1.0	0.3	3778	3117	661	5.0	100.0	95.0	0.1
7	630	630	0	1.0	0.0	4408	3747	661	6.0	100.0	94.0	0.1
8	630	630	0	1.0	0.0	5038	4377	661	7.0	100.0	93.0	0.1
9	629	629	0	1.0	0.0	5667	5006	661	8.0	100.0	92.0	0.1
10	630	630	0	1.0	0.0	6297	5636	661	9.0	100.0	91.0	0.1
11	630	630	0	1.0	0.0	6927	6266	661	10.1	100.0	89.9	0.1
12	630	630	0	1.0	0.0	7557	6896	661	11.1	100.0	88.9	0.1
13	630	630	0	1.0	0.0	8187	7526	661	12.1	100.0	87.9	0.1
14	629	629	0	1.0	0.0	8816	8155	661	13.1	100.0	86.9	0.1
15	630	630	0	1.0	0.0	9446	8785	661	14.1	100.0	85.9	0.1
16	630	630	0	1.0	0.0	10076	9415	661	15.1	100.0	84.9	0.2
17	630	630	0	1.0	0.0	10706	10045	661	16.1	100.0	83.9	0.2
18	629	629	0	1.0	0.0	11335	10674	661	17.1	100.0	82.9	0.2
19	630	630	0	1.0	0.0	11965	11304	661	18.1	100.0	81.9	0.2
20	630	630	0	1.0	0.0	12595	11934	661	19.2	100.0	80.8	0.2

Testing	# Records	# Goods	# Bads	Fraud Rate
	20993	20774	219	0.0104

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	209	63	146	0.3	66.7	209	63	146	0.3	66.7	66.4	0.0
2	210	175	35	0.8	16.0	419	238	181	1.1	82.6	81.5	0.0
3	210	204	6	1.0	2.7	629	442	187	2.1	85.4	83.3	0.0
4	210	209	1	1.0	0.5	839	651	188	3.1	85.8	82.7	0.0
5	210	206	4	1.0	1.8	1049	857	192	4.1	87.7	83.5	0.0
6	210	207	3	1.0	1.4	1259	1064	195	5.1	89.0	83.9	0.1
7	210	208	2	1.0	0.9	1469	1272	197	6.1	90.0	83.8	0.1
8	210	210	0	1.0	0.0	1679	1482	197	7.1	90.0	82.8	0.1
9	210	209	1	1.0	0.5	1889	1691	198	8.1	90.4	82.3	0.1
10	210	207	3	1.0	1.4	2099	1898	201	9.1	91.8	82.6	0.1
11	210	210	0	1.0	0.0	2309	2108	201	10.1	91.8	81.6	0.1
12	210	209	1	1.0	0.5	2519	2317	202	11.2	92.2	81.1	0.1
13	210	210	0	1.0	0.0	2729	2527	202	12.2	92.2	80.1	0.1
14	210	208	2	1.0	0.9	2939	2735	204	13.2	93.2	80.0	0.1
15	209	207	2	1.0	0.9	3148	2942	206	14.2	94.1	79.9	0.2
16	210	209	1	1.0	0.5	3358	3151	207	15.2	94.5	79.4	0.2
17	210	210	0	1.0	0.0	3568	3361	207	16.2	94.5	78.3	0.2
18	210	210	0	1.0	0.0	3778	3571	207	17.2	94.5	77.3	0.2
19	210	209	1	1.0	0.5	3988	3780	208	18.2	95.0	76.8	0.2
20	210	209	1	1.0	0.5	4198	3989	209	19.2	95.4	76.2	0.2

Out of Time	# Records	# Goods	# Bads	Fraud Rate
	12427	12248	179	0.0144

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	FPR
1	124	79	45	0.6	25.1	124	79	45	0.6	25.1	24.5	0.0
2	124	112	12	0.9	6.7	248	191	57	1.6	31.8	30.3	0.0
3	124	111	13	0.9	7.3	372	302	70	2.5	39.1	36.6	0.1
4	125	118	7	1.0	3.9	497	420	77	3.4	43.0	39.6	0.1
5	124	120	4	1.0	2.2	621	540	81	4.4	45.3	40.8	0.1
6	124	123	1	1.0	0.6	745	663	82	5.4	45.8	40.4	0.1
7	124	120	4	1.0	2.2	869	783	86	6.4	48.0	41.7	0.1
8	125	118	7	1.0	3.9	994	901	93	7.4	52.0	44.6	0.1
9	124	119	5	1.0	2.8	1118	1020	98	8.3	54.7	46.4	0.2
10	124	119	5	1.0	2.8	1242	1139	103	9.3	57.5	48.2	0.2
11	124	121	3	1.0	1.7	1366	1260	106	10.3	59.2	48.9	0.2
12	125	121	4	1.0	2.2	1491	1381	110	11.3	61.5	50.2	0.2
13	124	123	1	1.0	0.6	1615	1504	111	12.3	62.0	49.7	0.2
14	124	119	5	1.0	2.8	1739	1623	116	13.3	64.8	51.6	0.2
15	125	119	6	1.0	3.4	1864	1742	122	14.2	68.2	53.9	0.2
16	124	122	2	1.0	1.1	1988	1864	124	15.2	69.3	54.1	0.2
17	124	124	0	1.0	0.0	2112	1988	124	16.2	69.3	53.0	0.2
18	124	124	0	1.0	0.0	2236	2112	124	17.2	69.3	52.0	0.2
19	125	124	1	1.0	0.6	2361	2236	125	18.3	69.8	51.6	0.3
20	124	122	2	1.0	1.1	2485	2358	127	19.3	70.9	51.7	0.3

Fraud Savings Calculation

- Loss Sales = $50 * number of false positive

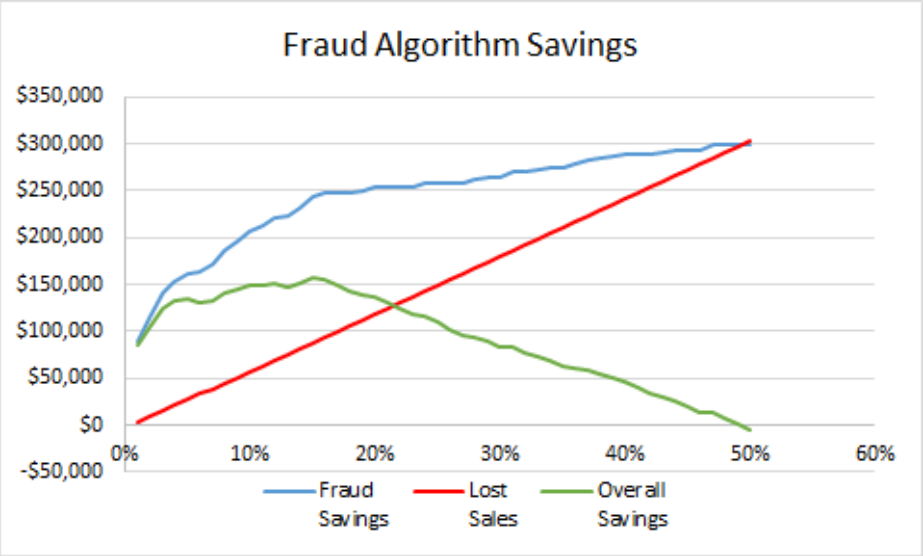
When we flag a good transaction as a fraud, we anger the customer, and some of them leave. On average, we assume a $50 loss for each false positives.

- Fraud Saving = $2000 * number of true positive.

We save $2000 for every fraud we catch with our algorithm.

- Overall savings = Fraud Saving - Loss Sales

Based on our out-of-time data, we could calculate our model's business value. The following graph shows the fraud algorithm savings.



The blue line shows our business value on how much we saved by catching the right fraudulent transactions. The lost sales red line represents the wrongly caught transaction values. Combining these two lines together, we could get the net saving green line. At 15% cut-off, the loss sale is $87,100, and the fraud saving is $244,000. Therefore, the net savings of applying the random forest model in out-of-time dataset achieves the maximum ($156,900) at 15% population cutoff.

8 Conclusions

8.1 Steps Performed

To build and select an optimal supervised fraud model based on the card transaction data, we first cleaned the data, and then built 269 variables incorporating amount information, card information and transaction location information. We used feature selection methods to reduce the number of variables to 20. We split the card transaction data and used the data in November and December 2010 as out out-of-time data. 75% of the transactions from January to October 2010 were randomly selected to be used as training data and the remaining 25% used as testing data. After that, we built four supervised fraud models using four different types of machine learning algorithms: Logistic Regression, Neural Nets, Random Forests, and Boosting Trees, and found that Random Forests outperforms the rest three.

8.2 Recommendations

Due to limited time and resources, we were only able to perform the work described in previous sections of this report. Given more time, we could use other machine learning techniques such as Decision Trees, Bagging, Support Vector Machine, etc. We might find a model that is a better fit than Random Forest to predict future fraud.

9 Appendix

9.1 Appendix 1 - Candidate Variables

1	Fraud	135	Actual/sum_Cardnum_Merchnum_3d
2	mean_Cardnum_1d	136	sum_Cardnum_Merchnum_7d
3	Actual/mean_Cardnum_1d	137	Actual/sum_Cardnum_Merchnum_7d
4	mean_Cardnum_3d	138	sum_Cardnum_Merchnum_14d
5	Actual/mean_Cardnum_3d	139	Actual/sum_Cardnum_Merchnum_14d
6	mean_Cardnum_7d	140	sum_Cardnum_Merchnum_30d
7	Actual/mean_Cardnum_7d	141	Actual/sum_Cardnum_Merchnum_30d
8	mean_Cardnum_14d	142	count_Cardnum_Merchnum_1d
9	Actual/mean_Cardnum_14d	143	Actual/count_Cardnum_Merchnum_1d
10	mean_Cardnum_30d	144	count_Cardnum_Merchnum_3d
11	Actual/mean_Cardnum_30d	145	Actual/count_Cardnum_Merchnum_3d
12	max_Cardnum_1d	146	count_Cardnum_Merchnum_7d
13	Actual/max_Cardnum_1d	147	Actual/count_Cardnum_Merchnum_7d
14	max_Cardnum_3d	148	count_Cardnum_Merchnum_14d
15	Actual/max_Cardnum_3d	149	Actual/count_Cardnum_Merchnum_14d
16	max_Cardnum_7d	150	count_Cardnum_Merchnum_30d
17	Actual/max_Cardnum_7d	151	Actual/count_Cardnum_Merchnum_30d
18	max_Cardnum_14d	152	mean_Cardnum_Merch zip_1d
19	Actual/max_Cardnum_14d	153	Actual/mean_Cardnum_Merch zip_1d
20	max_Cardnum_30d	154	mean_Cardnum_Merch zip_3d
21	Actual/max_Cardnum_30d	155	Actual/mean_Cardnum_Merch zip_3d
22	median_Cardnum_1d	156	mean_Cardnum_Merch zip_7d
23	Actual/median_Cardnum_1d	157	Actual/mean_Cardnum_Merch zip_7d
24	median_Cardnum_3d	158	mean_Cardnum_Merch zip_14d
25	Actual/median_Cardnum_3d	159	Actual/mean_Cardnum_Merch zip_14d
26	median_Cardnum_7d	160	mean_Cardnum_Merch zip_30d
27	Actual/median_Cardnum_7d	161	Actual/mean_Cardnum_Merch zip_30d
28	median_Cardnum_14d	162	max_Cardnum_Merch zip_1d
29	Actual/median_Cardnum_14d	163	Actual/max_Cardnum_Merch zip_1d
30	median_Cardnum_30d	164	max_Cardnum_Merch zip_3d
31	Actual/median_Cardnum_30d	165	Actual/max_Cardnum_Merch zip_3d
32	sum_Cardnum_1d	166	max_Cardnum_Merch zip_7d
33	Actual/sum_Cardnum_1d	167	Actual/max_Cardnum_Merch zip_7d
34	sum_Cardnum_3d	168	max_Cardnum_Merch zip_14d
35	Actual/sum_Cardnum_3d	169	Actual/max_Cardnum_Merch zip_14d
36	sum_Cardnum_7d	170	max_Cardnum_Merch zip_30d
37	Actual/sum_Cardnum_7d	171	Actual/max_Cardnum_Merch zip_30d
38	sum_Cardnum_14d	172	median_Cardnum_Merch zip_1d
39	Actual/sum_Cardnum_14d	173	Actual/median_Cardnum_Merch zip_1d
40	sum_Cardnum_30d	174	median_Cardnum_Merch zip_3d
41	Actual/sum_Cardnum_30d	175	Actual/median_Cardnum_Merch zip_3d
42	count_Cardnum_1d	176	median_Cardnum_Merch zip_7d
43	Actual/count_Cardnum_1d	177	Actual/median_Cardnum_Merch zip_7d
44	count_Cardnum_3d	178	median_Cardnum_Merch zip_14d
45	Actual/count_Cardnum_3d	179	Actual/median_Cardnum_Merch zip_14d

46	count_Cardnum_7d	180	median_Cardnum_Merch zip_30d
47	Actual/count_Cardnum_7d	181	Actual/median_Cardnum_Merch zip_30d
48	count_Cardnum_14d	182	sum_Cardnum_Merch zip_1d
49	Actual/count_Cardnum_14d	183	Actual/sum_Cardnum_Merch zip_1d
50	count_Cardnum_30d	184	sum_Cardnum_Merch zip_3d
51	Actual/count_Cardnum_30d	185	Actual/sum_Cardnum_Merch zip_3d
52	mean_Merchnum_1d	186	sum_Cardnum_Merch zip_7d
53	Actual/mean_Merchnum_1d	187	Actual/sum_Cardnum_Merch zip_7d
54	mean_Merchnum_3d	188	sum_Cardnum_Merch zip_14d
55	Actual/mean_Merchnum_3d	189	Actual/sum_Cardnum_Merch zip_14d
56	mean_Merchnum_7d	190	sum_Cardnum_Merch zip_30d
57	Actual/mean_Merchnum_7d	191	Actual/sum_Cardnum_Merch zip_30d
58	mean_Merchnum_14d	192	count_Cardnum_Merch zip_1d
59	Actual/mean_Merchnum_14d	193	Actual/count_Cardnum_Merch zip_1d
60	mean_Merchnum_30d	194	count_Cardnum_Merch zip_3d
61	Actual/mean_Merchnum_30d	195	Actual/count_Cardnum_Merch zip_3d
62	max_Merchnum_1d	196	count_Cardnum_Merch zip_7d
63	Actual/max_Merchnum_1d	197	Actual/count_Cardnum_Merch zip_7d
64	max_Merchnum_3d	198	count_Cardnum_Merch zip_14d
65	Actual/max_Merchnum_3d	199	Actual/count_Cardnum_Merch zip_14d
66	max_Merchnum_7d	200	count_Cardnum_Merch zip_30d
67	Actual/max_Merchnum_7d	201	Actual/count_Cardnum_Merch zip_30d
68	max_Merchnum_14d	202	mean_Cardnum_Merch state_1d
69	Actual/max_Merchnum_14d	203	Actual/mean_Cardnum_Merch state_1d
70	max_Merchnum_30d	204	mean_Cardnum_Merch state_3d
71	Actual/max_Merchnum_30d	205	Actual/mean_Cardnum_Merch state_3d
72	median_Merchnum_1d	206	mean_Cardnum_Merch state_7d
73	Actual/median_Merchnum_1d	207	Actual/mean_Cardnum_Merch state_7d
74	median_Merchnum_3d	208	mean_Cardnum_Merch state_14d
75	Actual/median_Merchnum_3d	209	Actual/mean_Cardnum_Merch state_14d
76	median_Merchnum_7d	210	mean_Cardnum_Merch state_30d
77	Actual/median_Merchnum_7d	211	Actual/mean_Cardnum_Merch state_30d
78	median_Merchnum_14d	212	max_Cardnum_Merch state_1d
79	Actual/median_Merchnum_14d	213	Actual/max_Cardnum_Merch state_1d
80	median_Merchnum_30d	214	max_Cardnum_Merch state_3d
81	Actual/median_Merchnum_30d	215	Actual/max_Cardnum_Merch state_3d
82	sum_Merchnum_1d	216	max_Cardnum_Merch state_7d
83	Actual/sum_Merchnum_1d	217	Actual/max_Cardnum_Merch state_7d
84	sum_Merchnum_3d	218	max_Cardnum_Merch state_14d
85	Actual/sum_Merchnum_3d	219	Actual/max_Cardnum_Merch state_14d
86	sum_Merchnum_7d	220	max_Cardnum_Merch state_30d
87	Actual/sum_Merchnum_7d	221	Actual/max_Cardnum_Merch state_30d
88	sum_Merchnum_14d	222	median_Cardnum_Merch state_1d
89	Actual/sum_Merchnum_14d	223	Actual/median_Cardnum_Merch state_1d
90	sum_Merchnum_30d	224	median_Cardnum_Merch state_3d
91	Actual/sum_Merchnum_30d	225	Actual/median_Cardnum_Merch state_3d
92	count_Merchnum_1d	226	median_Cardnum_Merch state_7d
93	Actual/count_Merchnum_1d	227	Actual/median_Cardnum_Merch state_7d
94	count_Merchnum_3d	228	median_Cardnum_Merch state_14d
95	Actual/count_Merchnum_3d	229	Actual/median_Cardnum_Merch state_14d

96	count_Merchnum_7d	230	median_Cardnum_Merch state_30d
97	Actual/count_Merchnum_7d	231	Actual/median_Cardnum_Merch state_30d
98	count_Merchnum_14d	232	sum_Cardnum_Merch state_1d
99	Actual/count_Merchnum_14d	233	Actual/sum_Cardnum_Merch state_1d
100	count_Merchnum_30d	234	sum_Cardnum_Merch state_3d
101	Actual/count_Merchnum_30d	235	Actual/sum_Cardnum_Merch state_3d
102	mean_Cardnum_Merchnum_1d	236	sum_Cardnum_Merch state_7d
103	Actual/mean_Cardnum_Merchnum_1d	237	Actual/sum_Cardnum_Merch state_7d
104	mean_Cardnum_Merchnum_3d	238	sum_Cardnum_Merch state_14d
105	Actual/mean_Cardnum_Merchnum_3d	239	Actual/sum_Cardnum_Merch state_14d
106	mean_Cardnum_Merchnum_7d	240	sum_Cardnum_Merch state_30d
107	Actual/mean_Cardnum_Merchnum_7d	241	Actual/sum_Cardnum_Merch state_30d
108	mean_Cardnum_Merchnum_14d	242	count_Cardnum_Merch state_1d
109	Actual/mean_Cardnum_Merchnum_14d	243	Actual/count_Cardnum_Merch state_1d
110	mean_Cardnum_Merchnum_30d	244	count_Cardnum_Merch state_3d
111	Actual/mean_Cardnum_Merchnum_30d	245	Actual/count_Cardnum_Merch state_3d
112	max_Cardnum_Merchnum_1d	246	count_Cardnum_Merch state_7d
113	Actual/max_Cardnum_Merchnum_1d	247	Actual/count_Cardnum_Merch state_7d
114	max_Cardnum_Merchnum_3d	248	count_Cardnum_Merch state_14d
115	Actual/max_Cardnum_Merchnum_3d	249	Actual/count_Cardnum_Merch state_14d
116	max_Cardnum_Merchnum_7d	250	count_Cardnum_Merch state_30d
117	Actual/max_Cardnum_Merchnum_7d	251	Actual/count_Cardnum_Merch state_30d
118	max_Cardnum_Merchnum_14d	252	Days_since_per_Cardnum
119	Actual/max_Cardnum_Merchnum_14d	253	Days_since_per_Merchnum
120	max_Cardnum_Merchnum_30d	254	Days_since_per_Cardnum_Merchnum
121	Actual/max_Cardnum_Merchnum_30d	255	Days_since_per_Cardnum_Merch zip
122	median_Cardnum_Merchnum_1d	256	Days_since_per_Cardnum_Merch state
123	Actual/median_Cardnum_Merchnum_1d	257	['Cardnum']_7dvelo_sum
124	median_Cardnum_Merchnum_3d	258	['Merchnum']_7dvelo_sum
125	Actual/median_Cardnum_Merchnum_3d	259	['Cardnum']_14dvelo_sum
126	median_Cardnum_Merchnum_7d	260	['Merchnum']_14dvelo_sum
127	Actual/median_Cardnum_Merchnum_7d	261	['Cardnum']_30dvelo_sum
128	median_Cardnum_Merchnum_14d	262	['Merchnum']_30dvelo_sum
129	Actual/median_Cardnum_Merchnum_14d	263	['Cardnum']_7dvelo_count
130	median_Cardnum_Merchnum_30d	264	['Merchnum']_7dvelo_count
131	Actual/median_Cardnum_Merchnum_30d	265	['Cardnum']_14dvelo_count
132	sum_Cardnum_Merchnum_1d	266	['Merchnum']_14dvelo_count
133	Actual/sum_Cardnum_Merchnum_1d	267	['Cardnum']_30dvelo_count
134	sum_Cardnum_Merchnum_3d	268	['Merchnum']_30dvelo_count
		269	RANDOM

9.2 Appendix 2 – Data Quality Report

I. Dataset Summary

This credit card transaction dataset has **96,753 records**, **9 independent fields**, and **one dependent field,** which represent whether the record is fraud or not. This dataset is gathered from state of Tennessee. The time period is from 01/01/2010 to 12/31/2010.

II. Field Summary

Categorical Field

There are 8 categorical fields in this data set. Summary information for each categorical fields is shown as Table 1.

Field	Count	% Populated	Unique Value	Most Common
Recnum	96,753	100.00%	96,753	Uniform Distributed
Cardnum	96,753	100.00%	1,645	5142148452
Date	96,753	100.00%	365	2/28/10
Merchnum	93,378	96.51%	13,092	9.3009E+11
Merch description	96,753	100.00%	13,126	GSA-FSS-ADV
Merch state	95,558	98.76%	228	TN
Merch zip	92,097	95.19%	4,568	38118
Transtype	96,753	100.00%	4	P

Table 1.

Numerical Field

In this dataset there is only 1 numerical field. Table 2 represents some basic statistic summary for this field.

field	count	%Populated	mean	std	min	25%	50%	75%	max
Amount	96,753	100.00%	427.9	10,006	0.01	33.48	137.98	428.2	3102046

Table 2.

Response Field

"Fraud" is a response field in this dataset. There are two types, "0" and "1". Type "0" categorized as non-fraud transaction and Type "1" represents fraud transaction. Summary information is shown as Table 3.

Field	records	% Populated	Non-Fraud	Fraud
Fraud	96,753	100.00%	95694 (98.9%)	1059 (1.1%)

Table 3.

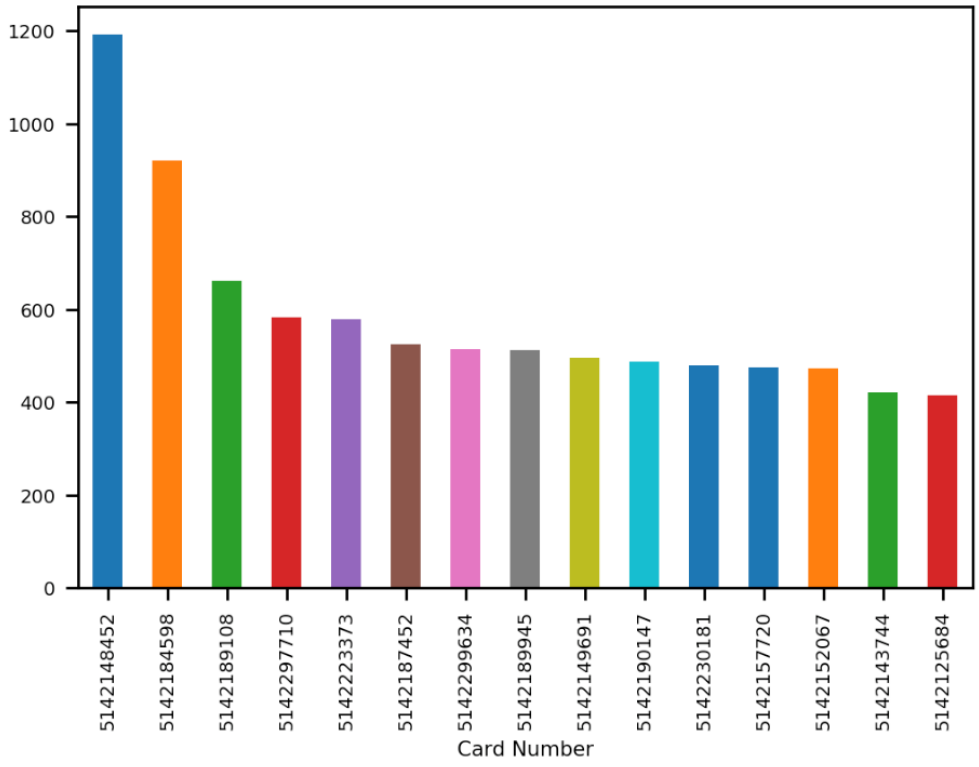
III. Fields Description

1. **Recnum**

   Recnum represents the numbering for each record. There are 96753 unique values in the datase t which means each record has one corresponding **Recnum.**

2. **Cardnum**

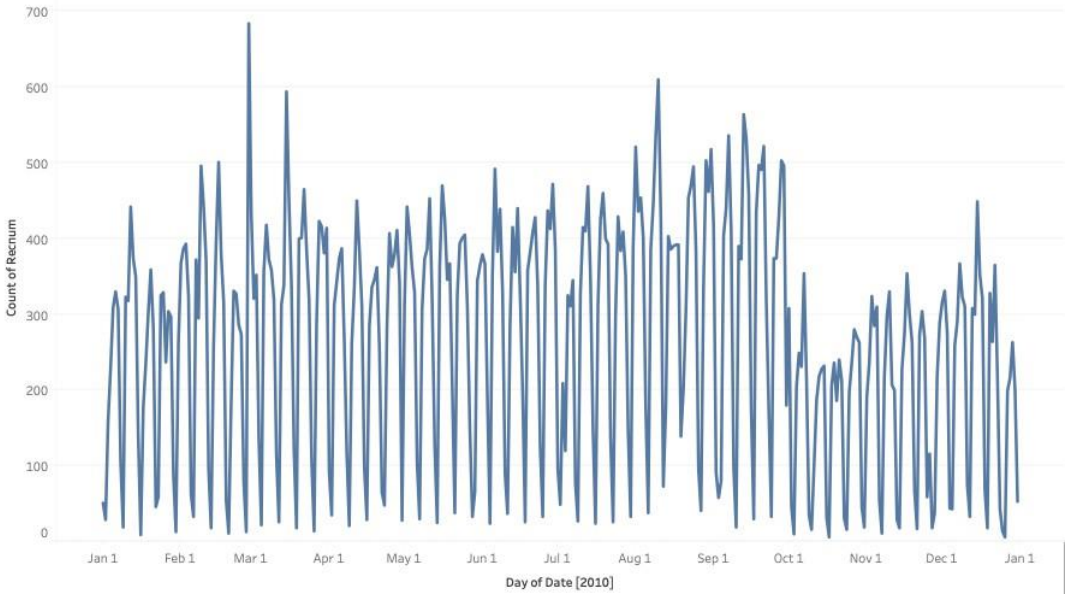
   Cardnum stands for credit card number. The following graph shows the transaction count for the top 15 most common credit card numbers.



3. Date

The Date field contains 365 unique values, which are date from 1/1/2010 to 12/31/2010.

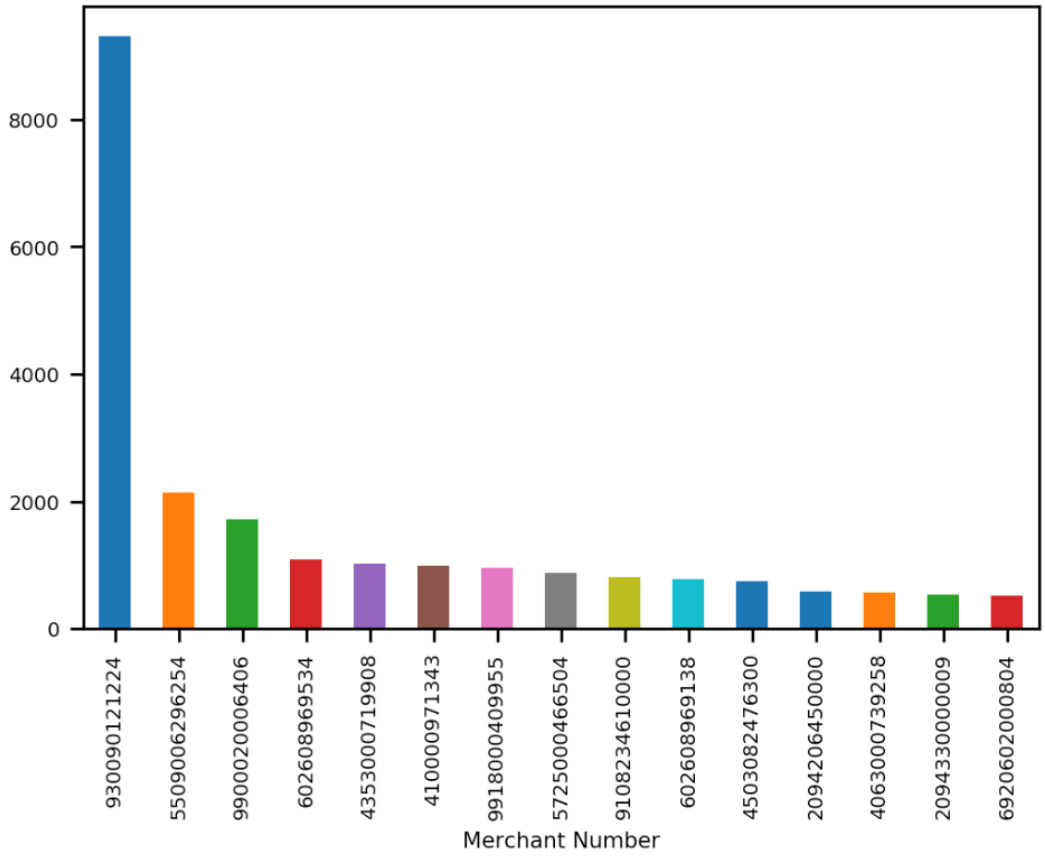
Count of Transactions by month



Count of Transactions by date

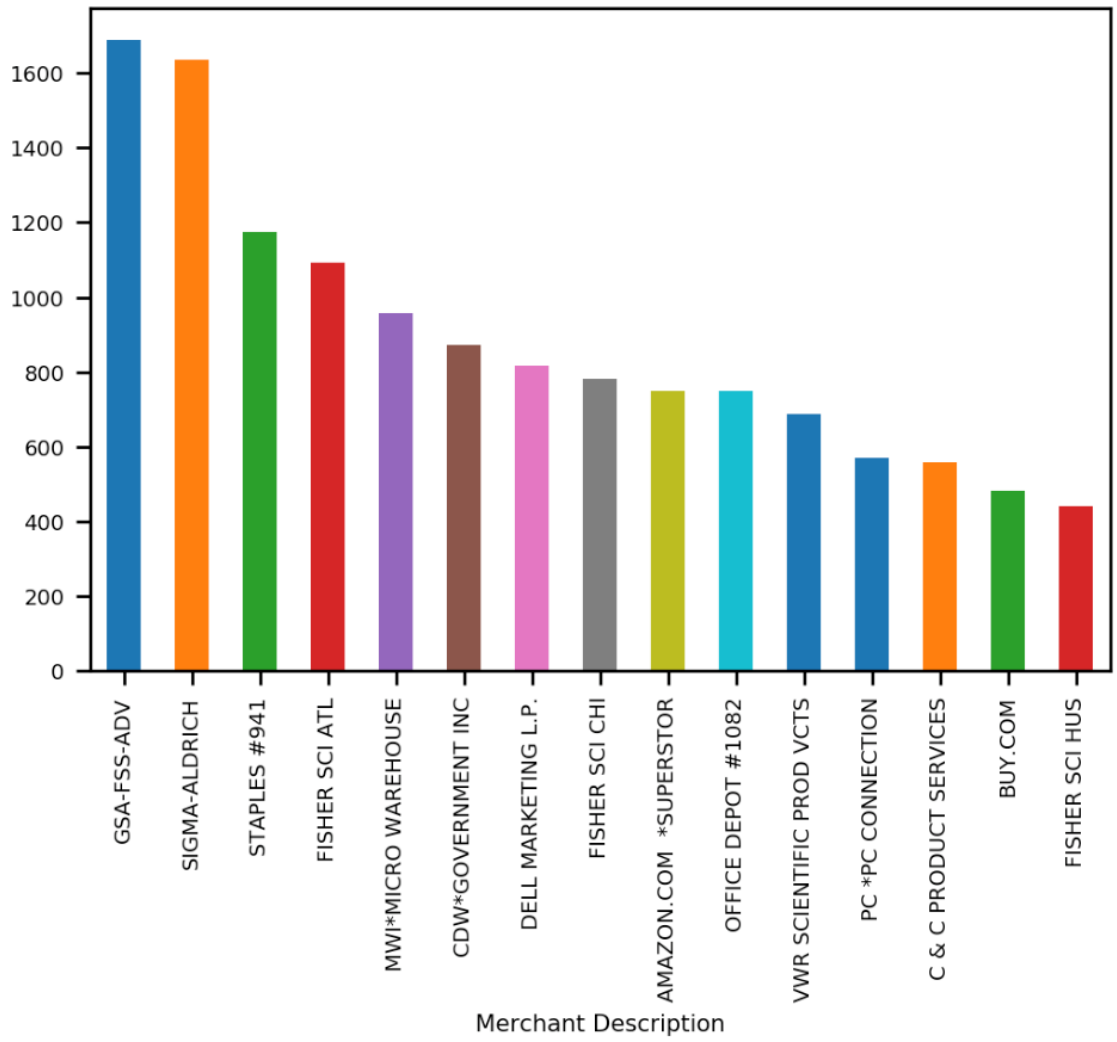
4. Merchnum

Merchnum represents merchant number. The graph below shows the count for the top 15 most common merchant numbers.



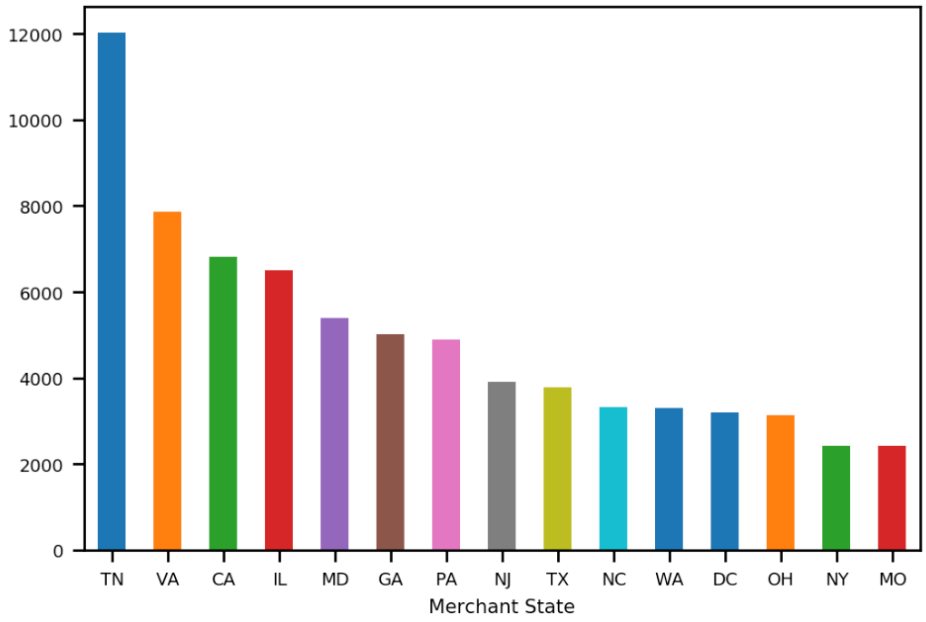
5. **Merch description**

Merch description is merchant description. Notice that there are 13,126 unique values for merchant description, which is more than that of merchant numbers. Below graph shows the count for the top 15 most common merchant description.



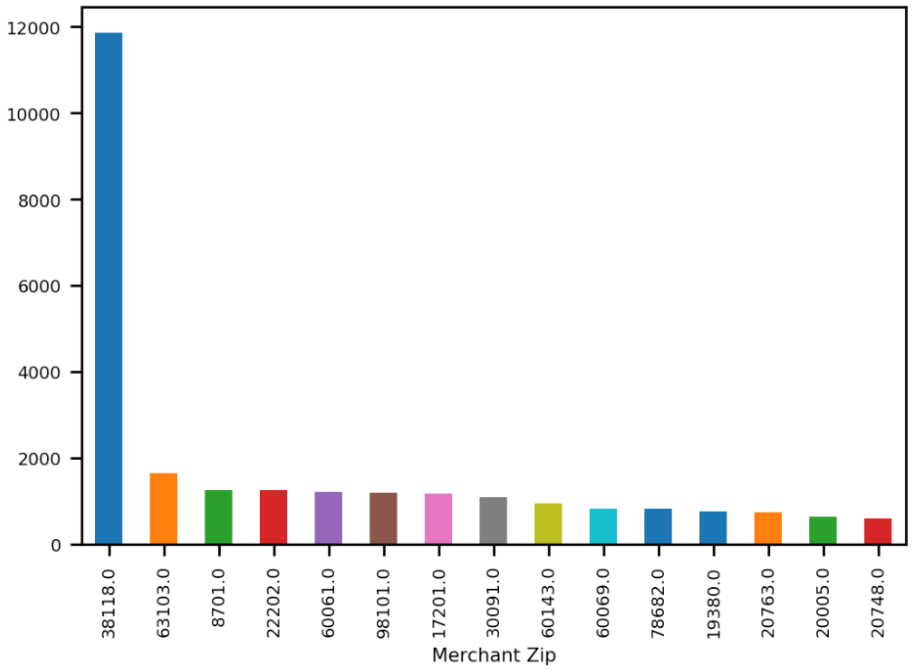
6. **Merch state**

Merch state represents the state where the merchant located. The following graph shows the count for the top 15 most common merchant states.



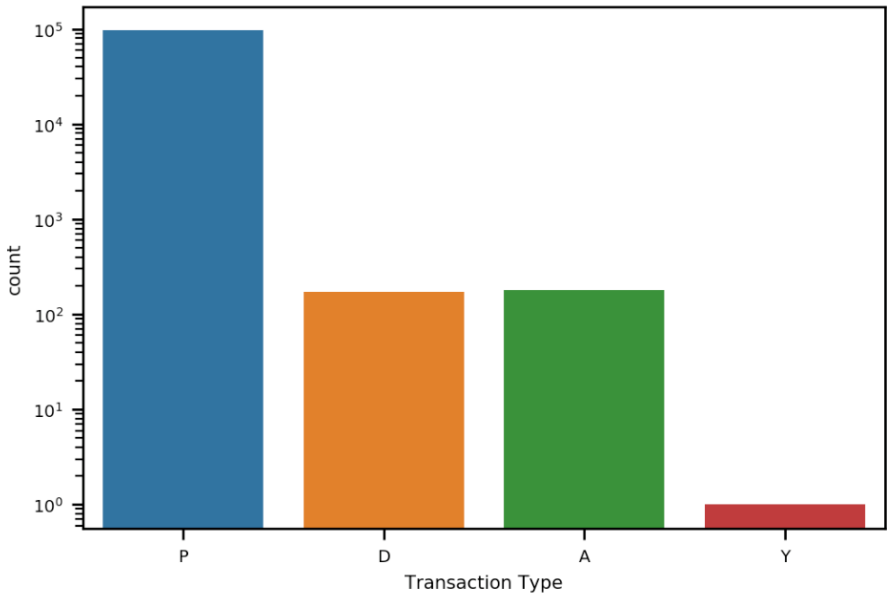
7. **Merch zip**

Merch zip represents the zip code where the merchant located. The following graph shows the count for the top 15 most common merchant zip code.



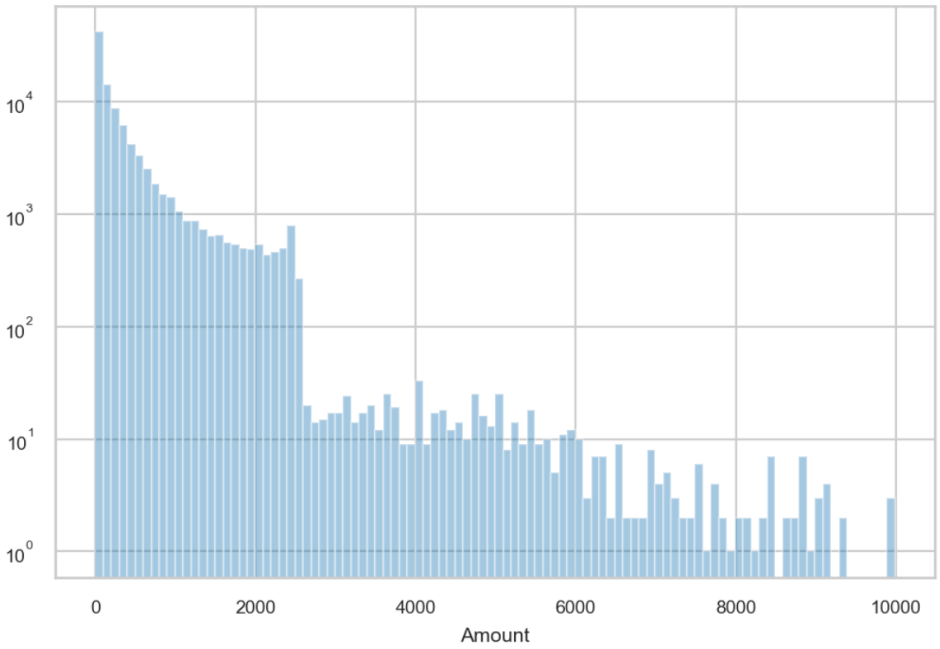
8. Transtype

Transtype stands for transaction types. There are 4 types of transaction in this dataset. “P”, which means “Purchase”, is the most common type in this filed. The following graph show the count for all types of transaction.



9. Amount

Amount is the transaction amount. Below graph shows the log scale distribution of transaction amount with amount limited to 10,000.



10. Fraud

Fraud is a response field in this dataset. There are two types, “0” and “1”. Type “0” categorized as non-fraud transaction and Type “1” represents fraud transaction. The fraud records account for about 1.1% of the whole dataset. The following is the log scale count of Fraud type.