1. Executive Summary



This report provides an analysis and evaluation of card transaction data for detecting fraud using supervised machine learning methods. The original data set contains 96,708 records of card transactions and has 10 variables of transaction details such as card number, date of transaction, merchant ID and merchant description, transaction type and transaction amount.

**Exhibit 1: Process Overview**

The general process of analysis follows data cleaning and manipulation, building expert variable, selecting important variables, applying fraud algorithms, calculating fraud scores and evaluating results. We divided the dataset into training, testing and out-of-time to evaluate the fraud score for each of these brackets.

The tools used include R, Python, JMP and Excel, and some of the algorithms used for analysis are Bootstrap Forest (aka Random Forest), Boosted Trees, Neural Networks, Naïve Bayes and Logistic Regression. These five models were built and tested a combination of data analytics tools and their respective performances were recorded. Among all the models built, Bootstrap Forest with 150 trees showed the best results. The optimal ROI using this model was $231,710 at a cut off 16.5%.

Using this best model, we got Fraud Detection Rates (FDR) of 66.5%, 58.9% and 52.4% for training, testing and out-of-time datasets. Through our analysis we could further identify that there were several expert variables that were strong predictors of fraud. Detailed examination of the important features selected from different algorithms indicate that fraudulent records typically have unusual geographical appearance.

1. Data Overview

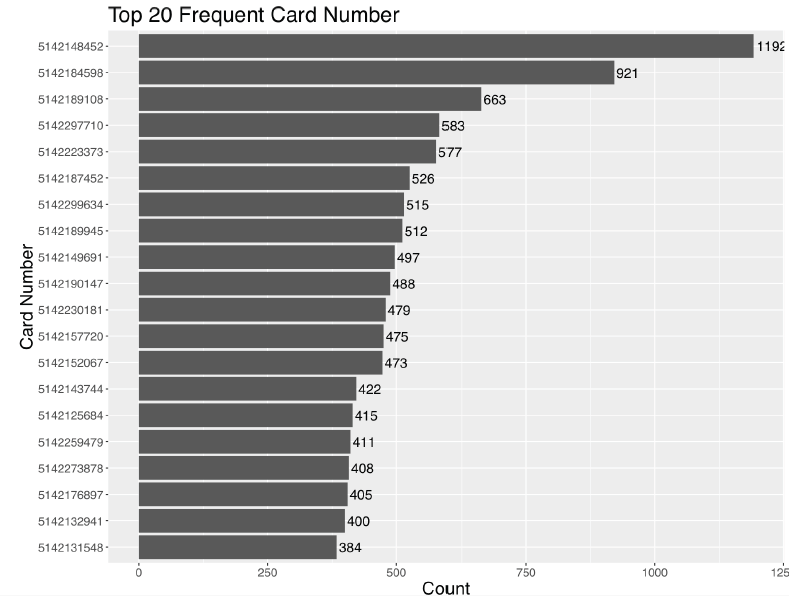
Card payment data is a dataset containing 96,708 records of card transactions from 2010-01-01 to 2010-12-31. It includes information about card number, date, merchant number, description, state, ZIP code, transaction type and amount. Every record is labeled whether it was fraud or not. In total, there are 1014 labeled fraudulent records.

**10 variables in total** – 1 numeric, 7 categorical, 1 text, 1 date  
**Numeric**: amount  
**Categorical**: recordnum, cardnum, merchnum, merch.state, merch.zip, transtype, and fraud  
**Text**: merch.description  
**Date**: date

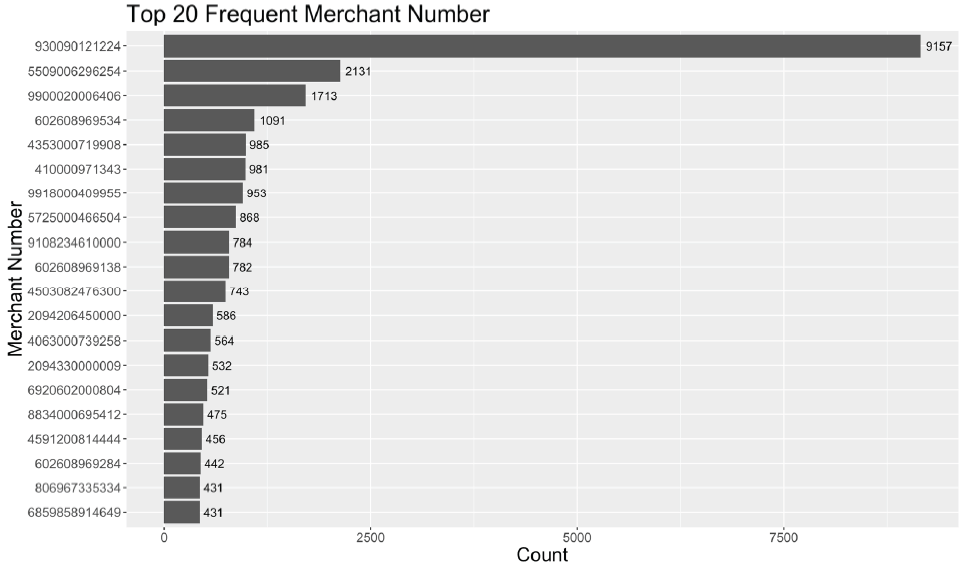
Following is the description of the variables we consider to be the most important. The complete Data Quality Report can be found in Appendix.

# **2.1 Description of important variables**

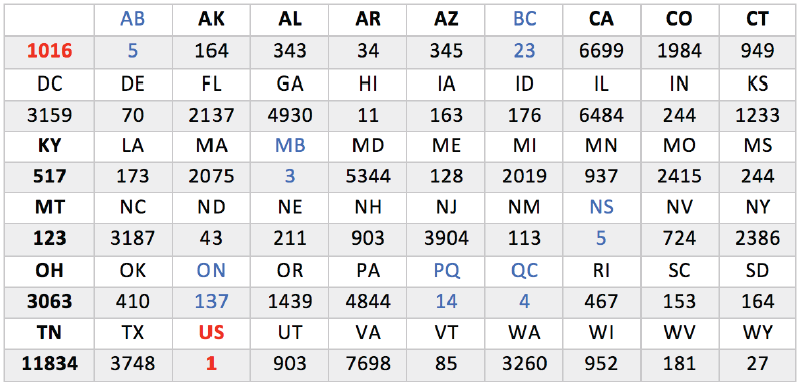
**2.1.1 cardnum**  
*cardnum* is a categorical variable. It is the number of the card used for payment. The variable is 100% populated and has 1644 unique values. The following plot shows top 20 frequent card numbers.

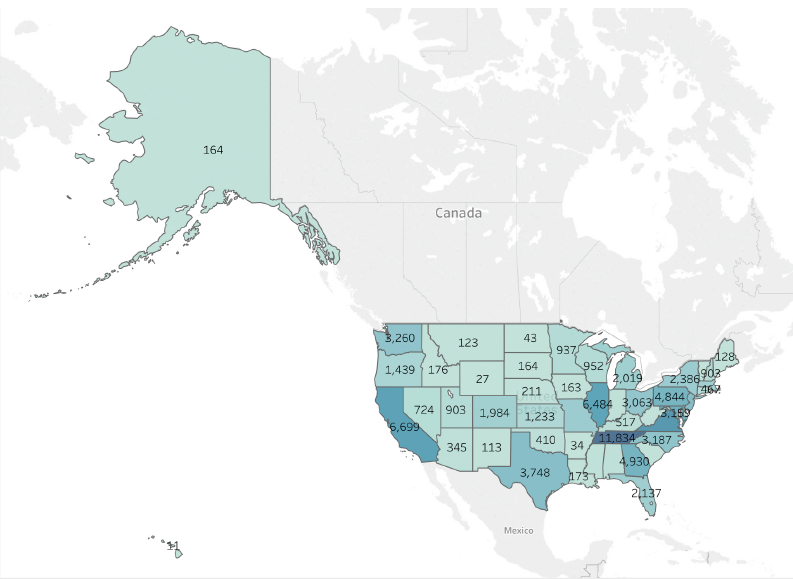


**2.1.2 merchnum***merchnum* is a categorical variable. It is the merchant number involved in the payment.The variable is 96.5% populated and has 13,090 unique values. There are 3,375 missing values, which includes the number of 0’s. The following plot shows top 20 frequent merchant numbers.



**2.1.3 merch.state***merch*.*state* is a categorical variable indicating the abbreviations of US states corresponding to each merchant. The variable is 98.7% populated and has 60 unique values. 51 values stand for regions in the United States – 50 states and District of Columbia; 7 values stand for Canada. Apart from these, there are 1,199 missing values, and 1 invalid value. Distribution of states is shown in the following table. States in Canada are marked in blue. Invalid and missing values are marked in red. Regions in the United States are in black.



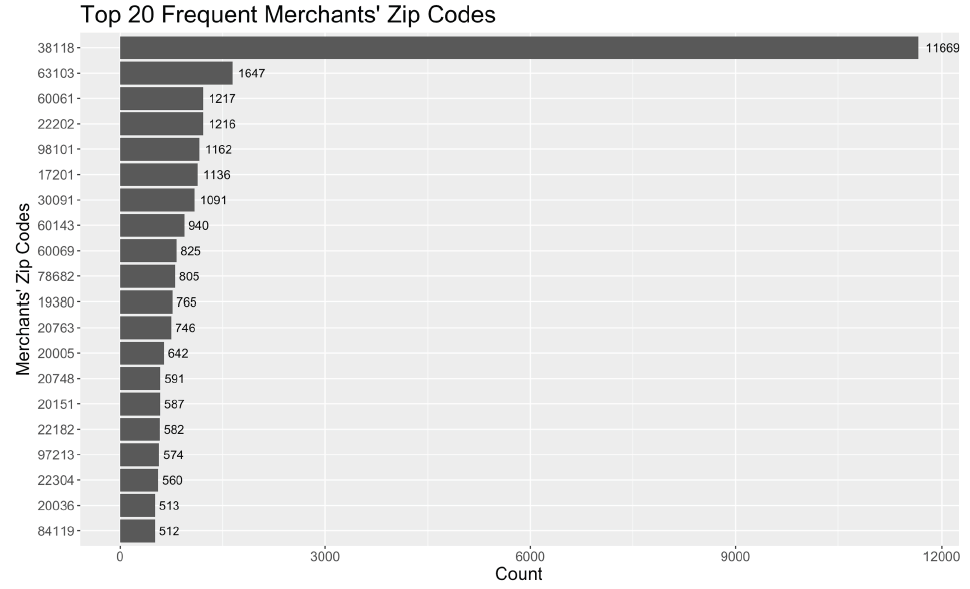
The following illustration shows the distribution of merchants’ states in the United States. We can clearly see that TN has the largest number of records (11,834). However, California and Washington feature high in terms of fraudulent card transactions.

**Details for merch.state with colors showing Number of Records**

**2.1.4 merch.zip**

*merch.zip* is a categorical variable indicating the ZIP code of the merchant. The variable is 95.2% populated and has 4567 unique values. 86.99% of the values have invalid 5-digit ZIP code. The

following plot shows top 20 ZIP codes. There is one ZIP code in Tennessee which has abnormally high frequency. Since, Tennessee is the most frequently occurring merchant state, we do not treat this ZIP code as a frivolous value.



1. Handling Missing Values

According to data quality analysis, three variables – *merchnum, merch.state* and *merch.zip* are not 100% populated and have missing values and/or 0s. We assigned values to these fields based on *merch.description*. We considered records with different *merch.description* as different merchants, and assumed that each merchant should have a unique merchant number, be in one state, and only have one ZIP code. Since all these three variables are categorical, we assigned unique values in these fields according to *merch.description*, and those values were designed to be quite different from other existing values, therefore easy to recognize. Missing values and 0s in *merchnum* were replaced by strings from ‘M1’ to ‘M771’, missing values in *merch.state* were substituted by numbers from ‘1’ to ‘150’, and NA’s in *merch.zip* were replaced by strings from ‘M1’ to ‘M644’.

1. Variable Creation

We created 130 expert variables for the training set. For the purposes of variable creation, we did not use *Fraud* variable since, using class to predict class isn’t an ideal case. Having knowledge about class may give us perfect models but that model might not make business sense.

For the purposes of variable creation, we counted linkages across five different time windows, namely, 1, 3, 7, 15 and 30 days. The rationale is to capture more (and different types of) fraudulent records that might be detected in those time windows.

After this, using Kolmogorov-Smirnov method, we selected a total of 40 variables. By carrying out Lasso, a shrinkage method, we were able to bring down the number of variables to 25. We built our models using those variables.

|  |  |
| --- | --- |
| **Variable Name** | **Description/Formula** |
| cardnum\_amount\_avg\_30 | Average of consumption amount with same card number within 30 days |
| cardnum\_amount\_avg\_15 | Average of consumption amount with same card number within 15 days |
| cardnum\_amount\_avg\_7 | Average of consumption amount with same card number within 7 days |
| cardnum\_amount\_avg\_3 | Average of consumption amount with same card number within 3 days |
| cardnum\_amount\_avg\_1 | Average of consumption amount with same card number within 1 days |
| cardnum\_amount\_max\_30 | Maximum of consumption amount with same card number within 30 days |
| cardnum\_amount\_max\_15 | Maximum of consumption amount with same card number within 15 days |
| cardnum\_amount\_max\_7 | Maximum of consumption amount with same card number within 7 days |
| cardnum\_amount\_max\_3 | Maximum of consumption amount with same card number within 3 days |
| cardnum\_amount\_max\_1 | Maximum of consumption amount with same card number within 1 days |
| cardnum\_amount\_sum\_30 | Sum of consumption amount with same card number within 30 days |
| cardnum\_amount\_sum\_15 | Sum of consumption amount with same card number within 15 days |
| cardnum\_amount\_sum\_7 | Sum of consumption amount with same card number within 7 days |
| cardnum\_amount\_sum\_3 | Sum of consumption amount with same card number within 3 days |
| cardnum\_amount\_sum\_1 | Sum of consumption amount with same card number within 1 days |
| cardnum\_amount\_week\_diff\_avg\_30 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 30 days |
| cardnum\_amount\_week\_diff\_avg\_15 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 15 days |
| cardnum\_amount\_week\_diff\_avg\_7 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 7 days |
| cardnum\_amount\_week\_diff\_avg\_3 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 3 days |
| cardnum\_amount\_week\_diff\_avg\_1 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 1 days |
| cardnum\_amount\_week\_diff\_max\_30 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 30 days |
| cardnum\_amount\_week\_diff\_max\_15 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 15 days |
| cardnum\_amount\_week\_diff\_max\_7 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 7 days |
| cardnum\_amount\_week\_diff\_max\_3 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 3 days |
| cardnum\_amount\_week\_diff\_max\_1 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 1 days |
| cardnum\_amount\_week\_diff\_sum\_30 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 30 days |
| cardnum\_amount\_week\_diff\_sum\_15 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 15 days |
| cardnum\_amount\_week\_diff\_sum\_7 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 7 days |
| cardnum\_amount\_week\_diff\_sum\_3 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 3 days |
| cardnum\_amount\_week\_diff\_sum\_1 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same card number within 1 days |
| cardnum\_amount\_month\_diff\_avg\_30 | Average of the difference between personal card consumption and average of comsumption amount for different month with same card number within 30 days |
| cardnum\_amount\_month\_diff\_avg\_15 | Average of the difference between personal card consumption and average of comsumption amount for different month with same card number within 15 days |
| cardnum\_amount\_month\_diff\_avg\_7 | Average of the difference between personal card consumption and average of comsumption amount for different month with same card number within 7 days |
| cardnum\_amount\_month\_diff\_avg\_3 | Average of the difference between personal card consumption and average of comsumption amount for different month with same card number within 3 days |
| cardnum\_amount\_month\_diff\_avg\_1 | Average of the difference between personal card consumption and average of comsumption amount for different month with same card number within 1 days |
| cardnum\_amount\_month\_diff\_max\_30 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 30 days |
| cardnum\_amount\_month\_diff\_max\_15 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 15 days |
| cardnum\_amount\_month\_diff\_max\_7 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 7 days |
| cardnum\_amount\_month\_diff\_max\_3 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 3 days |
| cardnum\_amount\_month\_diff\_max\_1 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 1 days |
| cardnum\_amount\_month\_diff\_sum\_30 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 30 days |
| cardnum\_amount\_month\_diff\_sum\_15 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 15 days |
| cardnum\_amount\_month\_diff\_sum\_7 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 7 days |
| cardnum\_amount\_month\_diff\_sum\_3 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 3 days |
| cardnum\_amount\_month\_diff\_sum\_1 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same card number within 1 days |
| cardnum\_30 | Count of records with same card number within 30 days before orginal record |
| cardnum\_15 | Count of records with same card number within 15 days before orginal record |
| cardnum\_7 | Count of records with same card number within 7 days before orginal record |
| cardnum\_3 | Count of records with same card number within 3 days before orginal record |
| cardnum\_1 | Count of records with same card number within 1 days before orginal record |
| cardnum\_merchantnum\_30 | Count of different merchant number with same card number within 30 days |
| cardnum\_merchantnum\_15 | Count of different merchant number with same card number within 15 days |
| cardnum\_merchantnum\_7 | Count of different merchant number with same card number within 7 days |
| cardnum\_merchantnum\_3 | Count of different merchant number with same card number within 3 days |
| cardnum\_merchantnum\_1 | Count of different merchant number with same card number within 1 days |
| cardnum\_zip\_30 | Count of different zip with same card number within 30 days |
| cardnum\_zip\_15 | Count of different zip with same card number within 15 days |
| cardnum\_zip\_7 | Count of different zip with same card number within 7 days |
| cardnum\_zip\_3 | Count of different zip with same card number within 3 days |
| cardnum\_zip\_1 | Count of different zip with same card number within 1 days |
| cardnum\_state\_30 | Count of different state with same card number within 30 days |
| cardnum\_state\_15 | Count of different state with same card number within 15 days |
| cardnum\_state\_7 | Count of different state with same card number within 7 days |
| cardnum\_state\_3 | Count of different state with same card number within 3 days |
| cardnum\_state\_1 | Count of different state with same card number within 1 days |
| merchantnum\_amount\_avg\_30 | Average of consumption amount with same merchant number within 30 days |
| merchantnum\_amount\_avg\_15 | Average of consumption amount with same merchant number within 15 days |
| merchantnum\_amount\_avg\_7 | Average of consumption amount with same merchant number within 7 days |
| merchantnum\_amount\_avg\_3 | Average of consumption amount with same merchant number within 3 days |
| merchantnum\_amount\_avg\_1 | Average of consumption amount with same merchant number within 1 days |
| merchantnum\_amount\_max\_30 | Maximum of consumption amount with same merchant number within 30 days |
| merchantnum\_amount\_max\_15 | Maximum of consumption amount with same merchant number within 15 days |
| merchantnum\_amount\_max\_7 | Maximum of consumption amount with same merchant number within 7 days |
| merchantnum\_amount\_max\_3 | Maximum of consumption amount with same merchant number within 3 days |
| merchantnum\_amount\_max\_1 | Maximum of consumption amount with same merchant number within 1 days |
| merchantnum\_amount\_sum\_30 | Sum of consumption amount with same merchant number within 30 days |
| merchantnum\_amount\_sum\_15 | Sum of consumption amount with same merchant number within 15 days |
| merchantnum\_amount\_sum\_7 | Sum of consumption amount with same merchant number within 7 days |
| merchantnum\_amount\_sum\_3 | Sum of consumption amount with same merchant number within 3 days |
| merchantnum\_amount\_sum\_1 | Sum of consumption amount with same merchant number within 1 days |
| merchantnum\_amount\_week\_diff\_avg\_30 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 30 days |
| merchantnum\_amount\_week\_diff\_avg\_15 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 15 days |
| merchantnum\_amount\_week\_diff\_avg\_7 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 7 days |
| merchantnum\_amount\_week\_diff\_avg\_3 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 3 days |
| merchantnum\_amount\_week\_diff\_avg\_1 | Average of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 1 days |
| merchantnum\_amount\_week\_diff\_max\_30 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 30 days |
| merchantnum\_amount\_week\_diff\_max\_15 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 15 days |
| merchantnum\_amount\_week\_diff\_max\_7 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 7 days |
| merchantnum\_amount\_week\_diff\_max\_3 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 3 days |
| merchantnum\_amount\_week\_diff\_max\_1 | Maximum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 1 days |
| merchantnum\_amount\_week\_diff\_sum\_30 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 30 days |
| merchantnum\_amount\_week\_diff\_sum\_15 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 15 days |
| merchantnum\_amount\_week\_diff\_sum\_7 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 7 days |
| merchantnum\_amount\_week\_diff\_sum\_3 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 3 days |
| merchantnum\_amount\_week\_diff\_sum\_1 | Sum of the difference between personal card consumption and average of comsumption amount for different day of week with same merchant number within 1 days |
| merchantnum\_amount\_month\_diff\_avg\_30 | Average of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 30 days |
| merchantnum\_amount\_month\_diff\_avg\_15 | Average of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 15 days |
| merchantnum\_amount\_month\_diff\_avg\_7 | Average of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 7 days |
| merchantnum\_amount\_month\_diff\_avg\_3 | Average of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 3 days |
| merchantnum\_amount\_month\_diff\_avg\_1 | Average of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 1 days |
| merchantnum\_amount\_month\_diff\_max\_30 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 30 days |
| merchantnum\_amount\_month\_diff\_max\_15 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 15 days |
| merchantnum\_amount\_month\_diff\_max\_7 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 7 days |
| merchantnum\_amount\_month\_diff\_max\_3 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 3 days |
| merchantnum\_amount\_month\_diff\_max\_1 | Maximum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 1 days |
| merchantnum\_amount\_month\_diff\_sum\_30 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 30 days |
| merchantnum\_amount\_month\_diff\_sum\_15 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 15 days |
| merchantnum\_amount\_month\_diff\_sum\_7 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 7 days |
| merchantnum\_amount\_month\_diff\_sum\_3 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 3 days |
| merchantnum\_amount\_month\_diff\_sum\_1 | Sum of the difference between personal card consumption and average of comsumption amount for different month with same merchant number within 1 days |
| merchantnum\_30 | Count of records with same merchant number within 30 days before orginal record |
| merchantnum\_15 | Count of records with same merchant number within 15 days before orginal record |
| merchantnum\_7 | Count of records with same merchant number within 7 days before orginal record |
| merchantnum\_3 | Count of records with same merchant number within 3 days before orginal record |
| merchantnum\_1 | Count of records with same merchant number within 1 days before orginal record |
| merchantnum\_cardnum\_30 | Count of different card number with same merchant number within 30 days |
| merchantnum\_cardnum\_15 | Count of different card number with same merchant number within 15 days |
| merchantnum\_cardnum\_7 | Count of different card number with same merchant number within 7 days |
| merchantnum\_cardnum\_3 | Count of different card number with same merchant number within 3 days |
| merchantnum\_cardnum\_1 | Count of different card number with same merchant number within 1 days |
| merchantnum\_zip\_30 | Count of different zip with same merchant number within 30 days |
| merchantnum\_zip\_15 | Count of different zip with same merchant number within 15 days |
| merchantnum\_zip\_7 | Count of different zip with same merchant number within 7 days |
| merchantnum\_zip\_3 | Count of different zip with same merchant number within 3 days |
| merchantnum\_zip\_1 | Count of different zip with same merchant number within 1 days |
| merchantnum\_state\_30 | Count of different state with same merchant number within 30 days |
| merchantnum\_state\_15 | Count of different state with same merchant number within 15 days |
| merchantnum\_state\_7 | Count of different state with same merchant number within 7 days |
| merchantnum\_state\_3 | Count of different state with same merchant number within 3 days |
| merchantnum\_state\_1 | Count of different state with same merchant number within 1 days |

We then built four types of variables using R (sqldf package):

**1. Type I** variables are intended to capture unusual amounts of transaction, both at the card level and the merchant level.

Example: avg\_amount\_cardnum\_30 tells the average of consumption amount with same card number within 30 days**.**

**2. Type II** variables are intended to capture unusual transaction frequency during a set period of time, both at the card level and the merchant level.

Example: cardnum\_30 describes the transaction frequency for each specific card number in the past 30 days

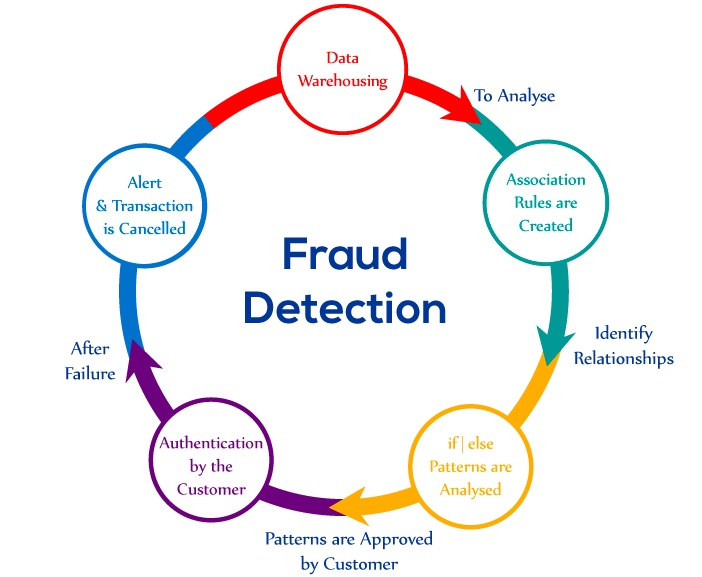
**3. Type III** variables are location related variables, which are intended to capture merchants with different ZIP codes and states in a given period of time.

Example: zip\_merchantnum\_30 tells how many zip codes are related to a particular merchant in the past 30 days.

**4.** **Type IV** variables are intended to catch card appearance pattern, either for a merchant or

for a card holder.

Example: cardnum\_merchantnum\_30 tells how many different cards a merchant was used for transaction within the past 30 days.

1. Supervised Learning Fraud Algorithms

# **5.1 Training/Test/OOT Set Split**

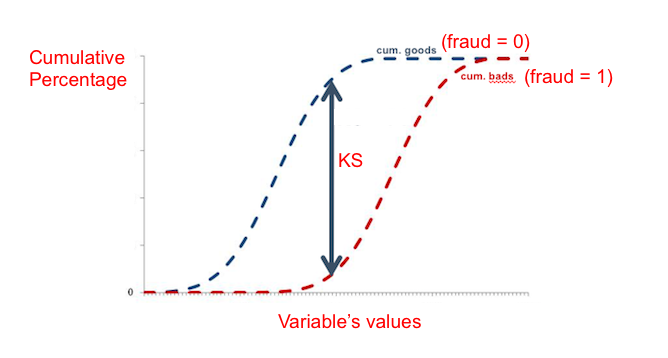
Before starting with the process of building our models, we split first 10-month worth of data randomly into 3:2 ratio to get our training and testing dataset, respectively. The last two months of data was set aside for out-of-time validation.

**5.2 Kolmogorov-Smirnov**

Kolmogorov–Smirnov score can be used in filter feature selection to measure the ability of one variable to distinguish classes. In [statistics](https://en.wikipedia.org/wiki/Statistics), the Kolmogorov–Smirnov test (K–S test) is a [nonparametric test](https://en.wikipedia.org/wiki/Nonparametric_statistics) of the equality of continuous, one-dimensional [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) that can be used to compare a [sample](https://en.wikipedia.org/wiki/Random_sample) with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

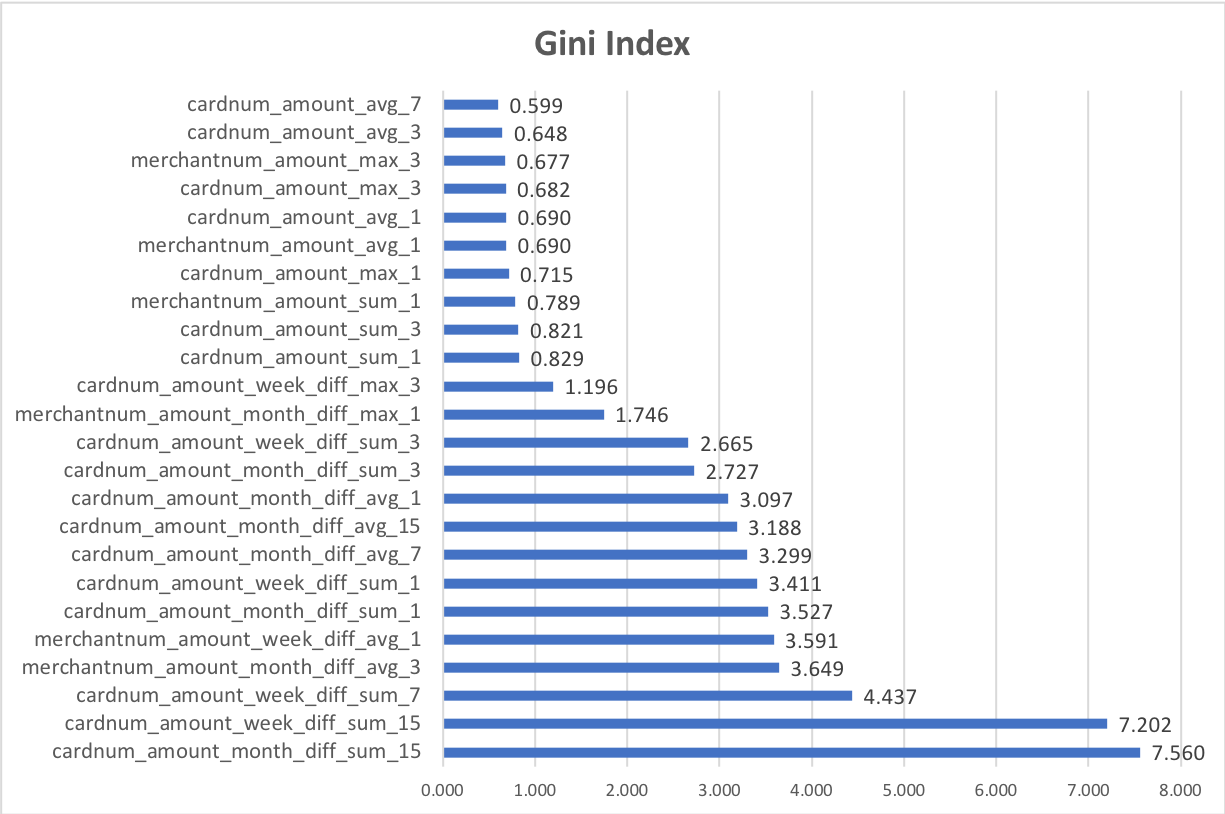
We implemented Kolmogorov–Smirnov using Python in the following steps:

* For each variable, we calculated the cumulative percentage on every variable's class i.e., good (*fraud* = 0) and bad (*fraud* = 1) separately, so we get two cumulative distributions.
* For each variable, we used *stats.ks\_2samp()* function from *spicy* package in Python to calculate the K-S between distributions of good (*fraud* = 0) and bad (*fraud* = 1) records.
* Finally, we sorted variables by decreasing K-S and chose the first 40 variables for the next step.

An illustration of K-S is as under.

This step reduced the number of variables from 130 to 40.

# **5.3 Lasso**

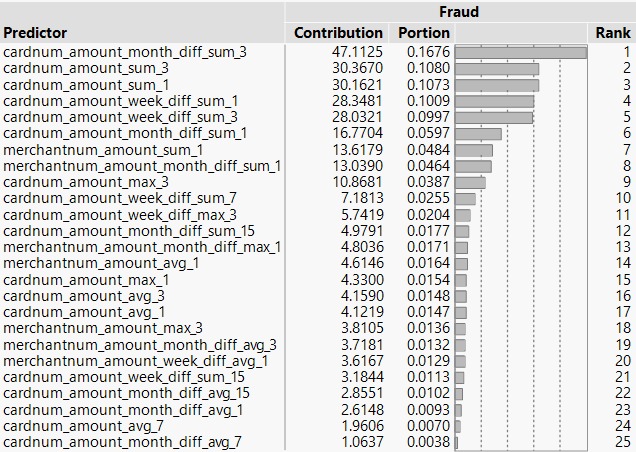
****In statistics and machine learning, lasso (least absolute shrinkage and selection operator) (also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. Lasso regression essentially shrinks the coefficient estimates towards zero. It tends to give sparser models. By applying this shrinkage method on the variables obtained after K-S, we were able to trim down to 25 variables. To get an idea about the inequality (and/or uniformity) of the variables, we calculated their Gini indices. It can be seen as a proxy for variable importance. The rank-ordered plot has been shown below.

**5.4 Supervised Machine Learning Models**

**5.4.1 Naïve Bayes**

Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

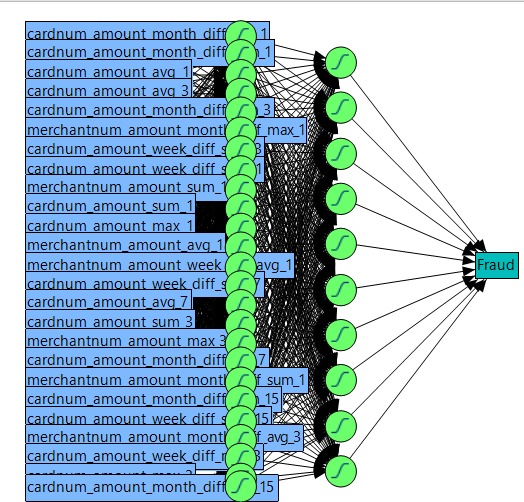
**5.4.2 Bootstrap Forest**

The Bootstrap Forest fits an ensemble model by averaging many decision trees each of which is fit to a bootstrap sample of the training data. Each split in each tree considers a random subset of the predictors. In this way, many weak models are combined to produce a more powerful model. The final prediction for an observation is the average of the predicted values for that observation over all the decision trees. Besides building the model, we also evaluated the contribution of predictors on the response using Bootstrap Forest partitioning (refer to the table given below).

**5.4.3 Boosted Tree**

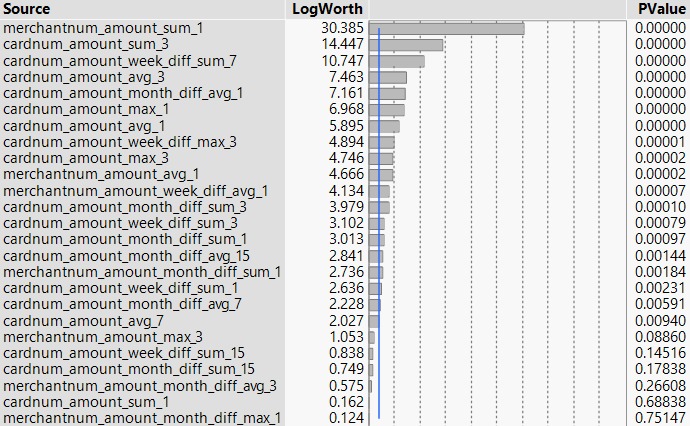
Boosting is the process of building a large, additive decision tree by fitting a sequence of smaller decision trees, called layers. The tree at each layer consists of a small number of splits. The tree is fit based on the residuals of the previous layers, which allows each layer to correct the fit for bad fitting data from the previous layers. The final prediction for an observation is the sum of the predictions for that observation over all of the layers. For our model, we used 150 layers, 3 splits per tree and a learning rate of 0.1.

**5.4.4 Neural Network**

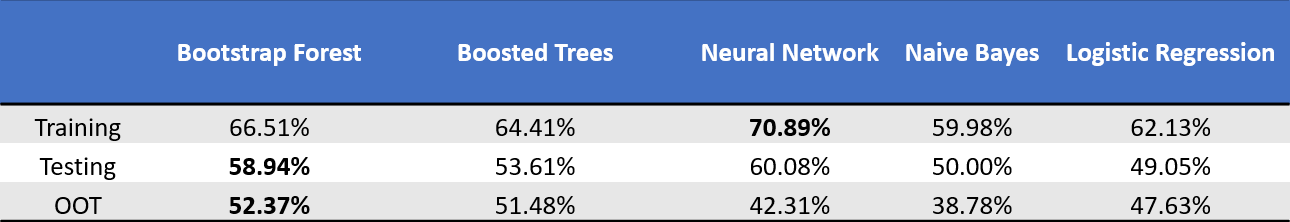
It is a technique in which while learning, one of the input patterns is given to the net's input layer. This pattern is propagated through the net (independent of its structure) to the net's output layer. The output layer generates an output pattern which is then compared to the target pattern. A depiction of our neural network model has been shown in this section. It has 25 input layer nodes and 10 hidden layer nodes.

**5.4.5 Logistic Regression**

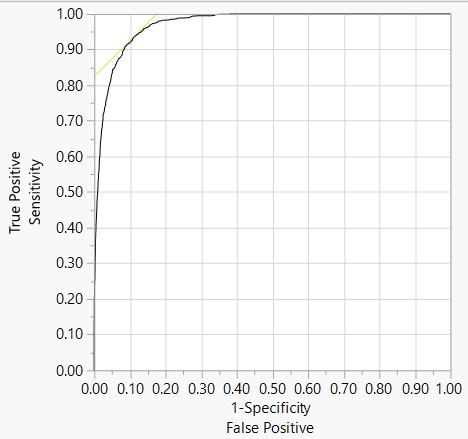
Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

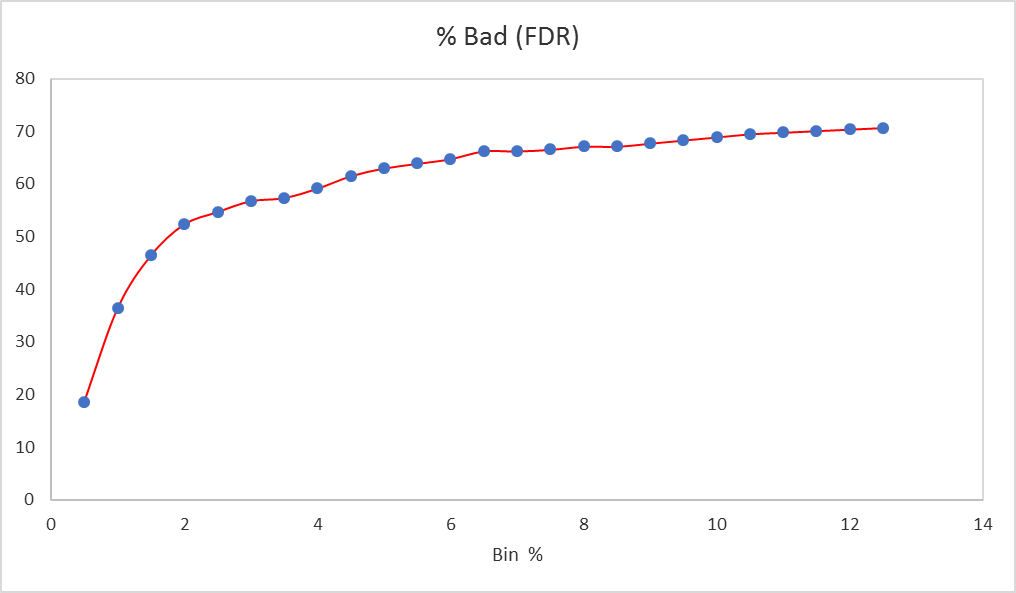
Variable importance (aka predictor contribution towards class) analysis based on logistic regression gave us the following table.

6. Results

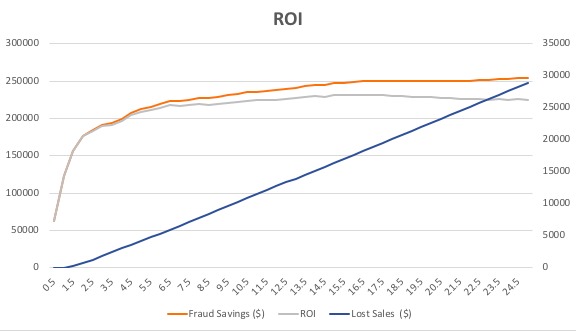
Among the five supervised learning models we built, Bootstrap Forest gave us the best results. There was a certain level of overfitting because of Type I variables, nonetheless, the results were up to the mark. Boosted Trees’ performance on out-of-time dataset came at a close second to Bootstrap Forest, which was the overall best model. The table shown below summarizes the FDR at 2% of all the models.

For our best model, Bootstrap Forest, the area under curve for ROC plot was 97.128%, which is very good. The corresponding plot has been shown below.

****A bin-wise summary (bin size = 0.5%) of the predictions of fraudulent card transactions on the out-of-time dataset (based on our best model, Bootstrap Forest) has been shown in the following table.

The following is a plot depicting the number of frauds caught in a given proportion of the total number of transactions.

To calculate the return on investment realized from our best model, we assumed a loss of $1000 for every fraud that’s not caught and a $10 loss for every good that’s flagged as bad (false positive). We found that a maximum return of $231,710 could be realized at a cut off 16.5%. Clearly, this cut off does not make a lot of business sense since a company will lose business if it’s set so high. A suitable balance based on ROI and customer convenience should be sought for. The plot given below depicts the savings realized against the percentage of transaction records. It can be seen that the fraud savings curve flattens out at 16.5% (slope becomes zero).



7. Conclusions

By comparing all the models and their performances, we found that Bootstrap Forest with 25 variables selected using K-S and Lasso performed the best.

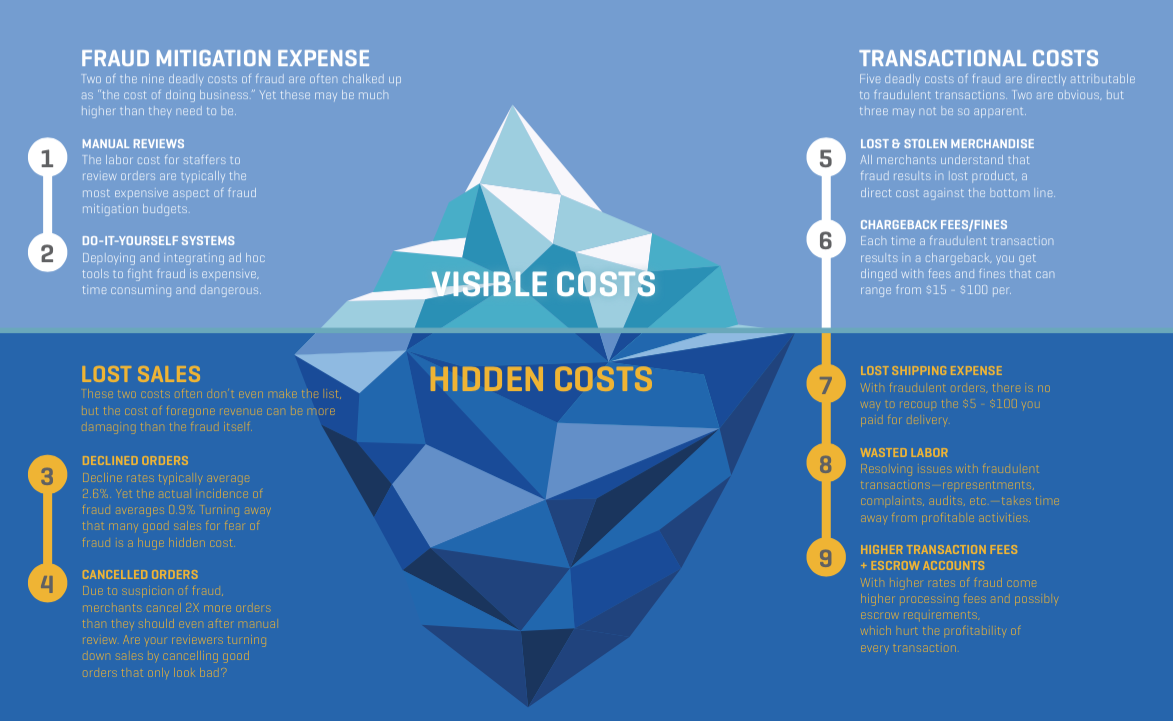
As seen from the ROI curve, flagging top 16.5% of all the records will create an estimated opportunity of $231,710. However, it’s important to note that this opportunity is not the same as true net savings because there are hidden costs associated with the declining legitimate card transactions. Customers tend to get annoyed, for example, if their card is declined when they are shopping for groceries in a different ZIP code.

One way to counter this risk is to assign different tiers to potentially fraudulent transactions based on the transaction amount. Purchases above a certain price could immediately be declined. While purchases below a certain threshold might only warrant a call (this illustrates the point of so called *soft margin*). Another factor to consider is that different customer segments, might be more sensitive to card declines. In order to properly assign tiers, we need to conduct a follow-up study with the marketing team to determine how sensitive certain customer segments are to card transaction declines at different levels.

A complete analysis of potential savings realized through our best model by 0.5% of the population has been summarized below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Population Bin(%)** | **Fraud Savings ($)** | **Lost Sales ($)** | **ROI** |
| 0.5 | 63000 | 0 | 63000 |
| 1 | 123000 | 30 | 122970 |
| 1.5 | 157000 | 320 | 156680 |
| 2 | 177000 | 750 | 176250 |
| 2.5 | 185000 | 1300 | 183700 |
| 3 | 192000 | 1860 | 190140 |
| 3.5 | 194000 | 2470 | 191530 |
| 4 | 200000 | 3040 | 196960 |
| 4.5 | 208000 | 3590 | 204410 |
| 5 | 213000 | 4170 | 208830 |
| 5.5 | 216000 | 4770 | 211230 |
| 6 | 219000 | 5370 | 213630 |
| 6.5 | 224000 | 5950 | 218050 |
| 7 | 224000 | 6580 | 217420 |
| 7.5 | 225000 | 7200 | 217800 |
| 8 | 227000 | 7810 | 219190 |
| 8.5 | 227000 | 8440 | 218560 |
| 9 | 229000 | 9050 | 219950 |
| 9.5 | 231000 | 9660 | 221340 |
| 10 | 233000 | 10270 | 222730 |
| 10.5 | 235000 | 10880 | 224120 |
| 11 | 236000 | 11500 | 224500 |
| 11.5 | 237000 | 12120 | 224880 |
| 12 | 238000 | 12740 | 225260 |
| 12.5 | 239000 | 13360 | 225640 |
| 13 | 241000 | 13970 | 227030 |
| 13.5 | 243000 | 14580 | 228420 |
| 14 | 245000 | 15190 | 229810 |
| 14.5 | 245000 | 15820 | 229180 |
| 15 | 248000 | 16420 | 231580 |
| 15.5 | 248000 | 17050 | 230950 |
| 16 | 249000 | 17670 | 231330 |
| **16.5** | **250000** | **18290** | **231710** |
| 17 | 250000 | 18920 | 231080 |
| 17.5 | 251000 | 19540 | 231460 |
| 18 | 251000 | 20170 | 230830 |
| 18.5 | 251000 | 20800 | 230200 |
| 19 | 251000 | 21430 | 229570 |
| 19.5 | 251000 | 22060 | 228940 |
| 20 | 251000 | 22690 | 228310 |
| 20.5 | 251000 | 23320 | 227680 |
| 21 | 251000 | 23950 | 227050 |
| 21.5 | 251000 | 24580 | 226420 |
| 22 | 251000 | 25210 | 225790 |
| 22.5 | 252000 | 25830 | 226170 |
| 23 | 252000 | 26460 | 225540 |
| 23.5 | 253000 | 27080 | 225920 |
| 24 | 253000 | 27710 | 225290 |
| 24.5 | 254000 | 28330 | 225670 |
| 25 | 254000 | 28960 | 225040 |
|  |  | **max ROI** | **231710** |

**7.1 Business Insights and Real-World Scenario**

****

Appendix

Data Quality Report

**Dataset:** Card Transaction Data

**Description:** This data is a simulated representation of the 96708 card transaction details applications during 2010. A typical observation contains information about the card and merchant details, along with geographical location parameters and dollar amount transacted.

**Number of Records:** 96708

**Number of Variables:** 10

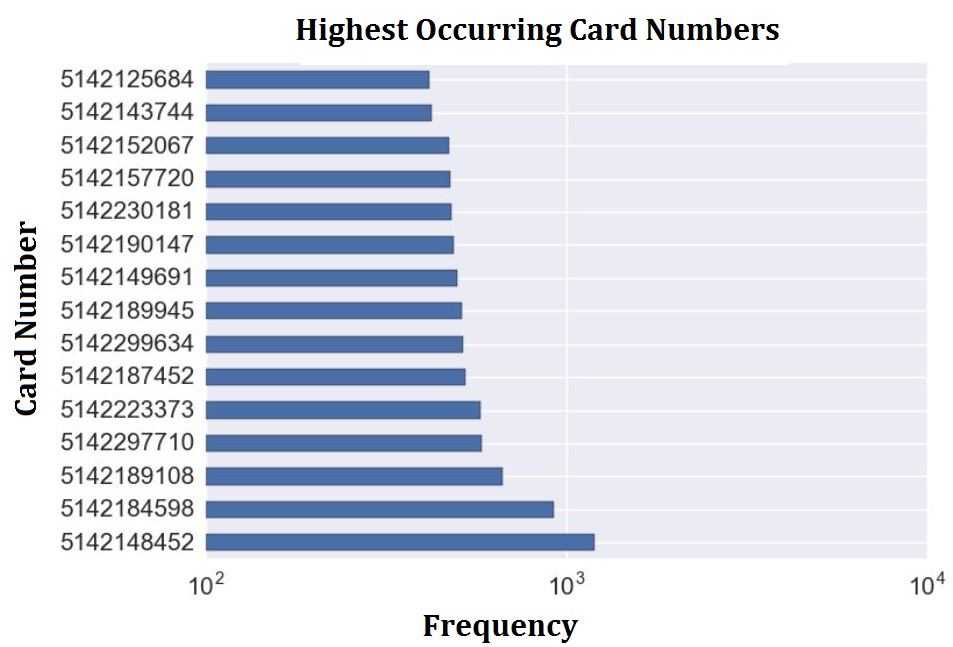
**Summary statistics for numerical variables**



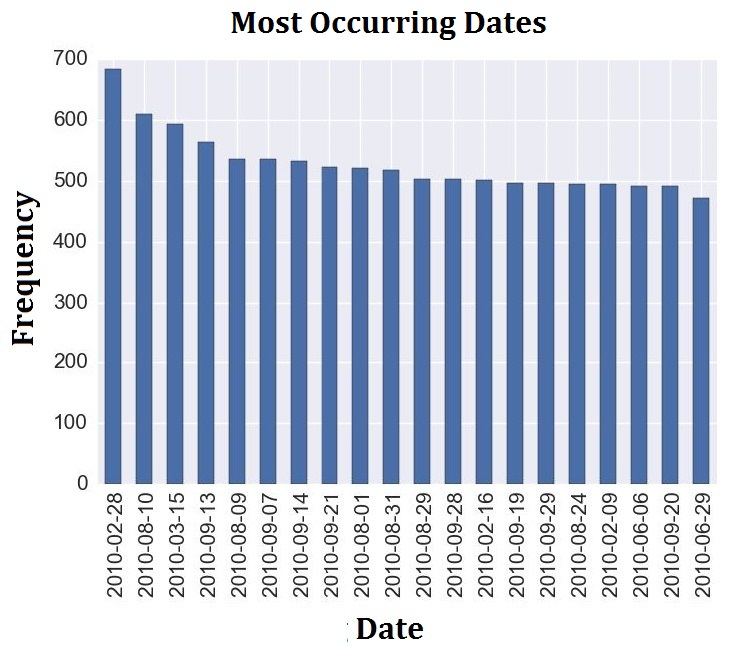
**Description and visualization of data**

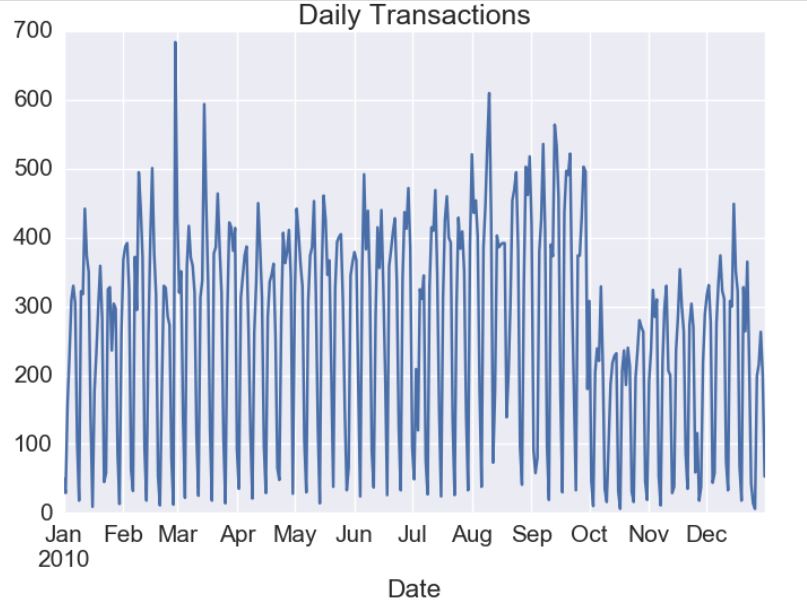
**1. *Recordnum*** is a categorical variable. It works as the ordinal reference number for each property record. There are 96708 records overall. Each row is a unique number/identifier and hence, a visualization is not required.

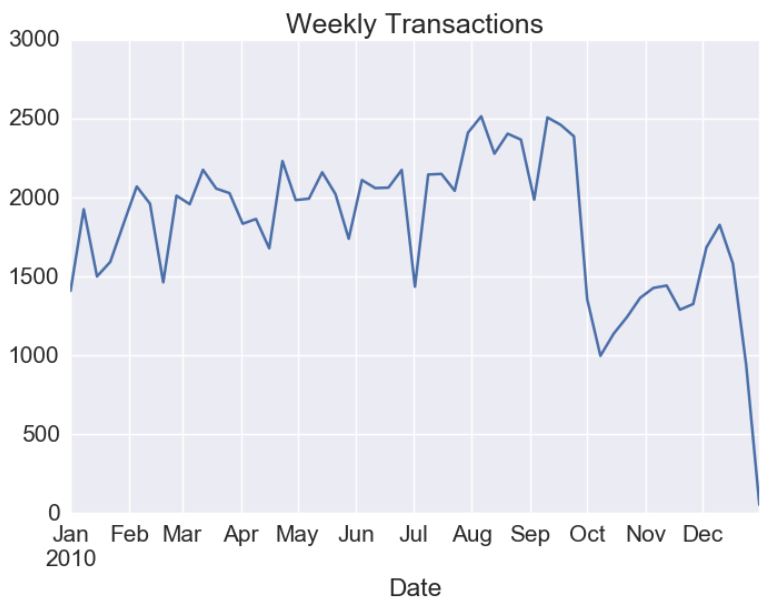
**2. *Cardnum*** refers to the Card Number used to make transaction.

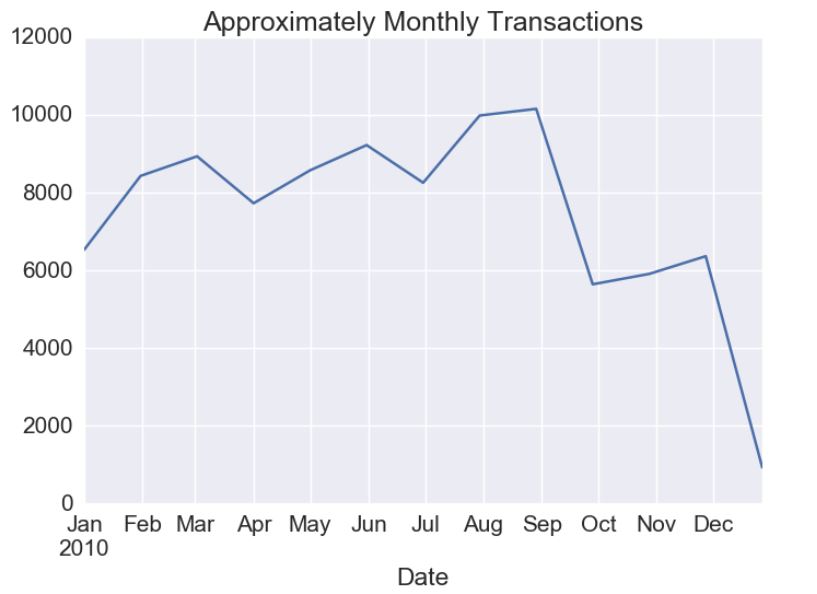


**3. *Date*** refers to the date on which the transaction was made.

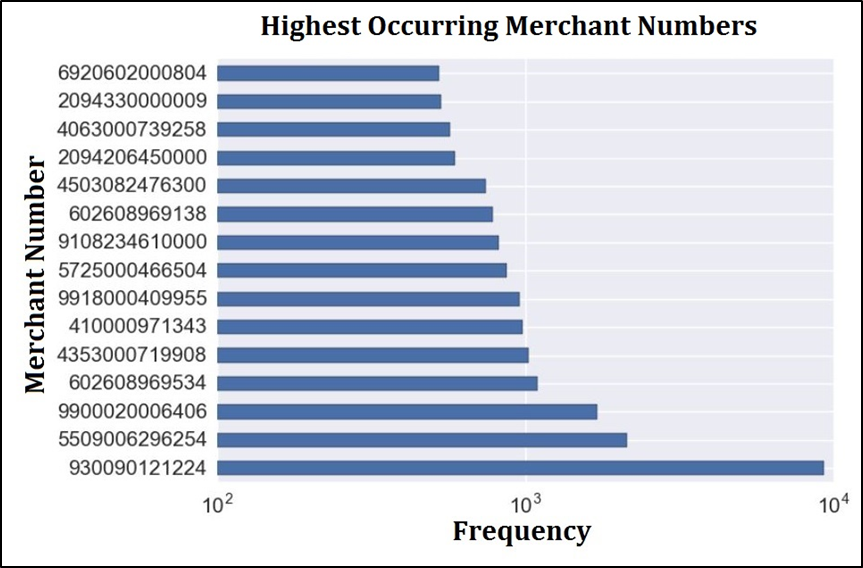
****



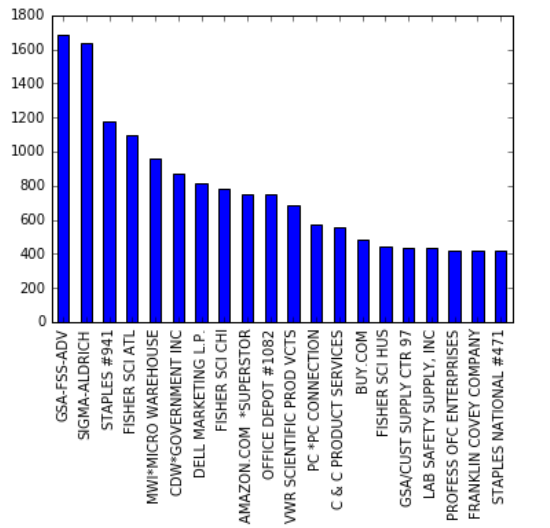




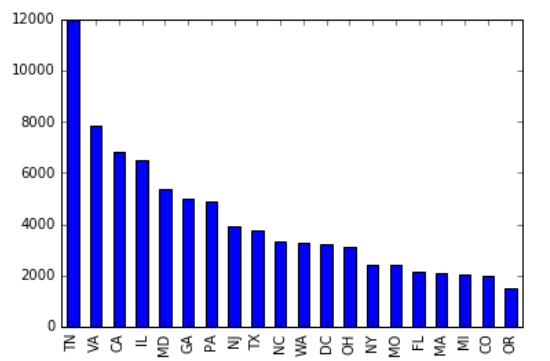
**4. *MerchantNum*** refers to unique Merchant ID associated with each transaction.



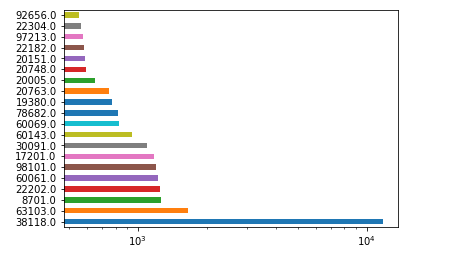
**5. *Merch Description*** refers to the name of the merchant which carried out the transaction. 20 most frequently occurring merchants have been shown below.



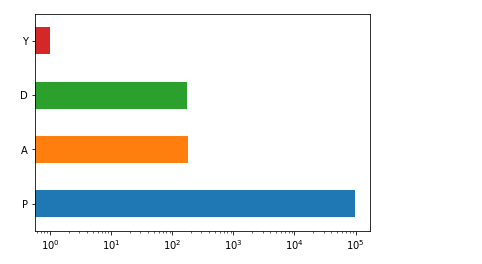
**6. *Merchant State*** refers to the state in which the merchant is located. States are abbreviated using 2-letter codes.



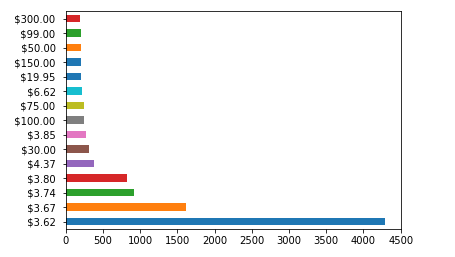
**7. *Merchant Zip*** contains information about each merchant’s ZIP code. Below is a graph of the top 20 most frequent zip codes.



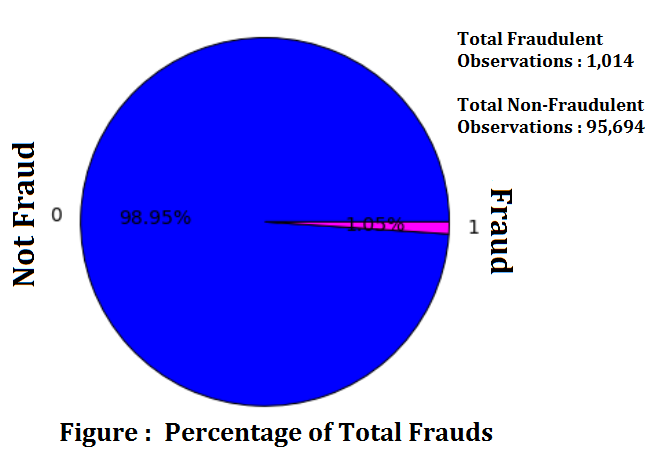
**8. *Transtype*** is a categorical variable referring to the type of transaction. There are four different types of transactions in the dataset with no missing values.



**9. *Amount*** contains information about the dollar amount corresponding to each transaction. Below is a graph of the top 15 most frequent transaction amounts.



**10. *Fraud*** contains information whether observation includes fraudulent data or not. Of the total 96,708 observations, we observed that 95,694 observations were noted as accurate/legitimate, while 1,014 observations were marked as fraudulent.



**Summary**

The Data Quality Report created for the Card Transaction dataset is required for the integrity of the data management by covering gaps of data issues. We not only cleaned and transformed the data, but also rectified inconsistencies and redundancies. Furthermore, we tabulated and visualized all variables, which provided us new insights into the dataset. These reports are valuable when administered on data that has made multiple iterations and additions before that data becomes authorized or stored for enterprise intelligence. As we gather more data to add to our current database, we can construct similar accuracy checks for all data sourced to our class project.

**Future Scope**

This report will aid data governance by monitoring data to find exceptions and anomalies undiscovered by previous data management operations. Further data quality checks may be defined at attribute level to have full control on its remediation steps. This report will serve as a cornerstone for all future phases of the group project.