Credit Card Fraud Detection

TEAM NO: 202
TEAM MEMBERS: KARAN SAXENA
HARSH BHUTA
NAMAN JAIN
SAMANAYU RAJU
ADREETA GUHA

Table of Contents

Executive Summary:

Credit card fraud is a burden for organizations across the globe. United States is the most credit card fraud prone country in the world. According to Federal Trade Commission's (FTC), instances of credit card fraud in the US increased by 44.6% from 271,927 in 2019 to 393,207 in 2020 which resulted in a loss of approximately 3.3 billion dollars. Credit card fraud also accounted for 393,207 of the nearly 1.4 million (~28.08%) reports of identity thefts in 2020. Through this report, our goal was to build and highlight efficient model to predict fraud. We analyzed a real-world dataset that contained a list of government related credit card transactions over the 2010 calendar year. The data presented a supervised problem as it included a labeled input "Fraud" which indicated if a particular transaction is fraudulent or not. The dataset contained identifying information about each transaction such as the credit card number, merchant, merchant state, etc. The dataset had 96,753 records and 10 data fields. We performed exploratory data analysis and visualized each of the 10 data fields, cleaned the dataset, and imputed missing values. Then we created 1000+ candidate variables and performed feature selection. Finally, we created a variety of models using several algorithms (both linear and nonlinear) and highlighted our results.

Through our data analysis, we found there are nine categorical variables and one numeric variable. We provided a histogram or table for each data field to better understand its distribution. Many of the variables had a right skewed distribution. Overall, the dataset had 1,059 fraudulent transactions of the total 96,753 records (1.09%). The dataset required cleaning before we could move to the variable creation phase. We removed one outlier record due to its extremely high value. Also, we only analyzed records that had a transaction type equal to "P" (purchase) to work with a more focused dataset. Lastly, the "Merchnum", "Merch state", and "Merch zip" data fields contained missing values which required careful data imputation.

Once the dataset was cleaned, we attempted to create as many candidate variables as possible (amount variables, frequency variables, days-since variables, velocity change variables, Benford's Law variables, and a day-of-the-week risk table variable.) After the variable creation process, we performed feature selection via filter and wrapper methods to reduce dimensionality and determine the best variables for use in the development of machine learning models to predict potentially fraudulent activity. Kolmogorov-Smirnov (KS) and Fraud Detection Rate (FDR) at 3% were used to eliminate effective variables as a filtering method, while the average results of three different tree-based methods was used as our wrapper method. Once the final variables were chosen, the data was divided into three sections for model development and analysis: training, testing, and out-of-time (OOT).

To find the best model, the variables were tested in several different models. First, we started with a baseline model Logistic regression. Next, we explored multiple nonlinear models such as Random Forest, Boosted Trees, and Neural Network to compare against the Logistic regression model. Parameter tuning was also performed on each model to get better results. Overall, our best model was Light Gradient Boosting which caught 55.86% fraud at a 3% FDR after hyperparameter tuning.

Description of the Data:

Dataset Name- Card Transaction Data

Dataset Description – Dataset contains information on the actual credit card purchases from a US government organization. It provides information on Credit Card, Merchant, Date, and Amount involved in each transaction. Moreover, it also contains a column called fraud label which tells us whether the transaction is fraudulent or not.

Total Fields – 10

Total Records – 96,753

Time Period - 1st January 2006 – 31st December 2006

Summary Tables:

Field Name	% Populated	Min Val	Max Val	Mean	StdDev	% Zeros
Amount	100%	0.01	3102,045.5	427.9	10,006.1	0

Table 1: Summary Table for Numerical Variables

Field Name	% Populated	# Unique Values	Most Common Value
Recnum	100	96,753	-
Cardnum	100	1,645	51421448452
Date	100	365	2006-02-28
Merchnum	96.5	13,092	930090121224
Merch description	100	13,126	GSA-FSS-ADV
Merch state	98.7	228	TN
Merch zip	95.2	4,568	38118
Transtype	100	4	P
Fraud	100	2	O

Table 2: Summary Table for Categorical Variables

From the above fields, we determined that ‘Date’, ‘Merchnum’, ‘Merchstate’, ‘Cardnum’ and ‘Amount’ are the most important fields. Few of the relevant depictions amongst the critical data fields are provided below.

1. **Date:**
   The 'Date' field gives us information on the date when the transaction took place. The graphs below give us an idea about the total number of transactions and frauds taking place over the year.

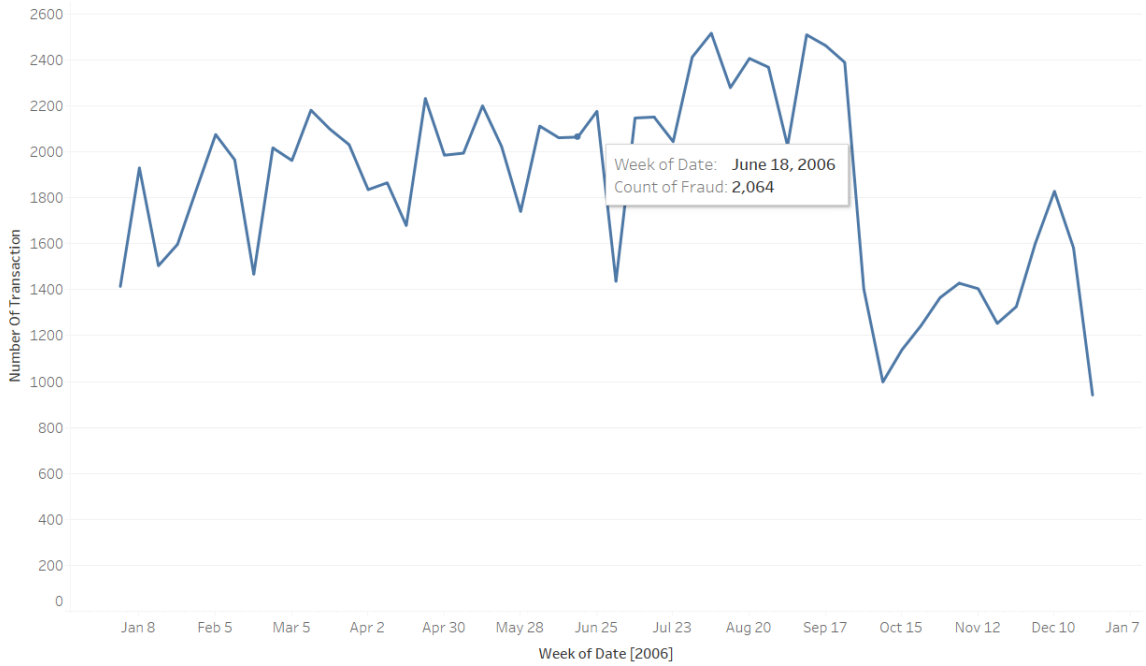


Fig 1. Weekly distribution of Number of Transaction

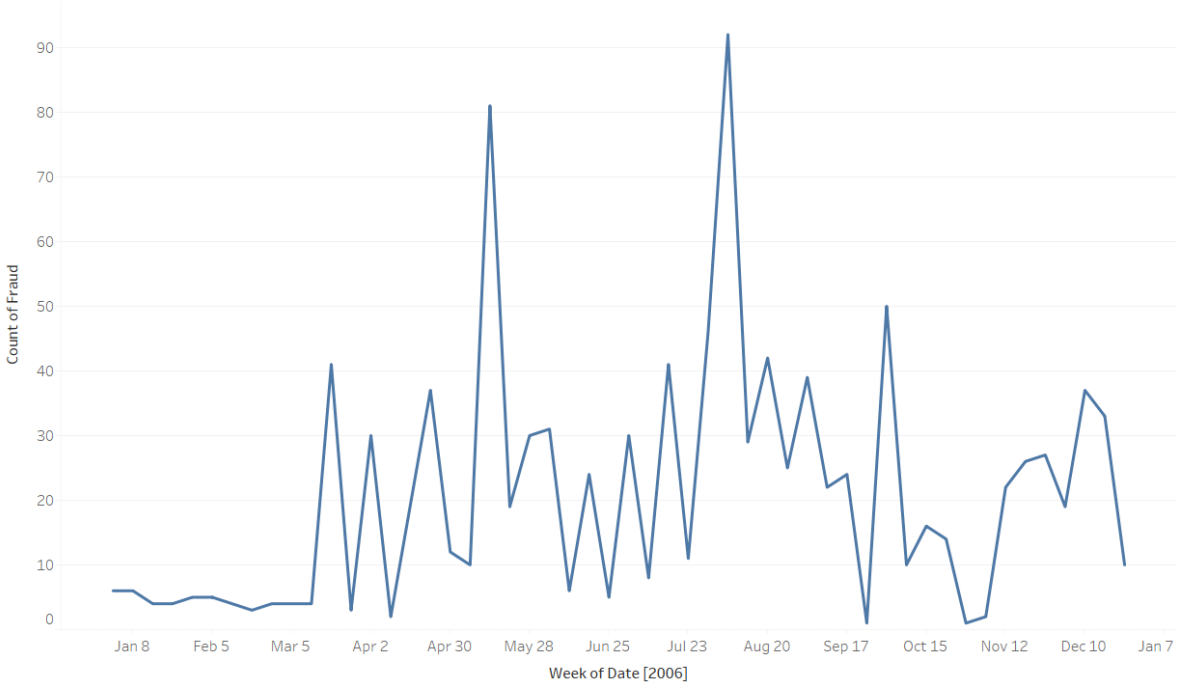


Fig 2. Number of Frauds per Week

2. Merchnum:

The field Merchnum provides information on the Merchnum of the merchant involved in the transaction. There are 13,092 unique addresses, out of which '930090121224' is the most common. The chart below indicates the top 15 most common Merchnum involved in the transactions.

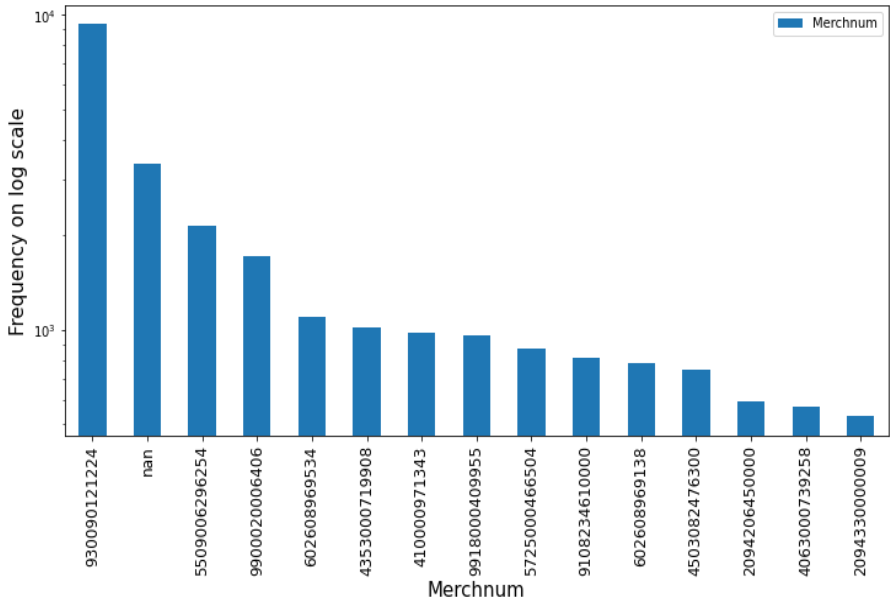


Fig 3. Bar plot displaying distribution of Merchnum

3. Merchsate:

The field Merchstate provides information on the state where the merchant involved in the transaction resides. There are total of 228 unique values for this field. The most common Merchstate being 'TN'.

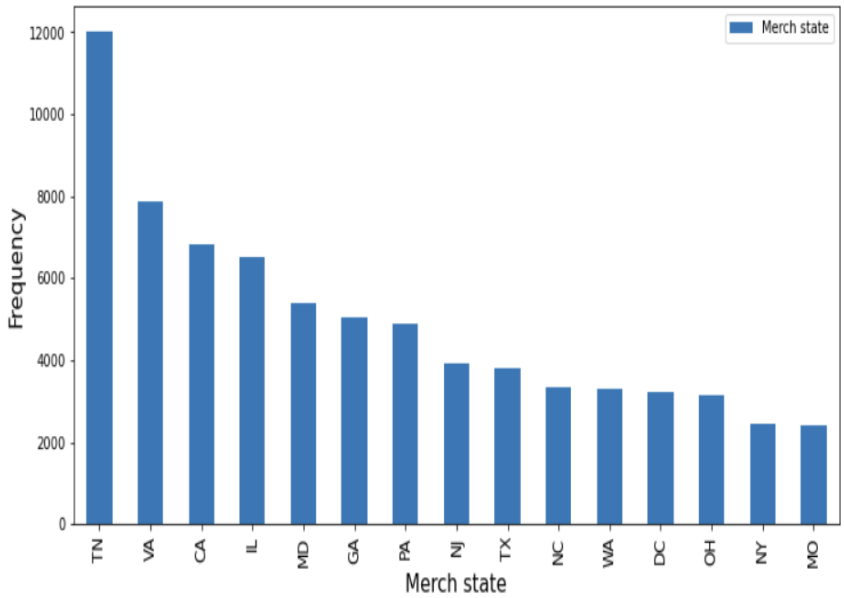


Fig 4. Bar Plot showing distribution of Merchstate

4. **Cardnum:**
The field Cardnum provides information on the cardnum of the customer involved in the transaction. There is a total of 1,654 unique Cardnum used. The most common Cardnum is '51421448452'.

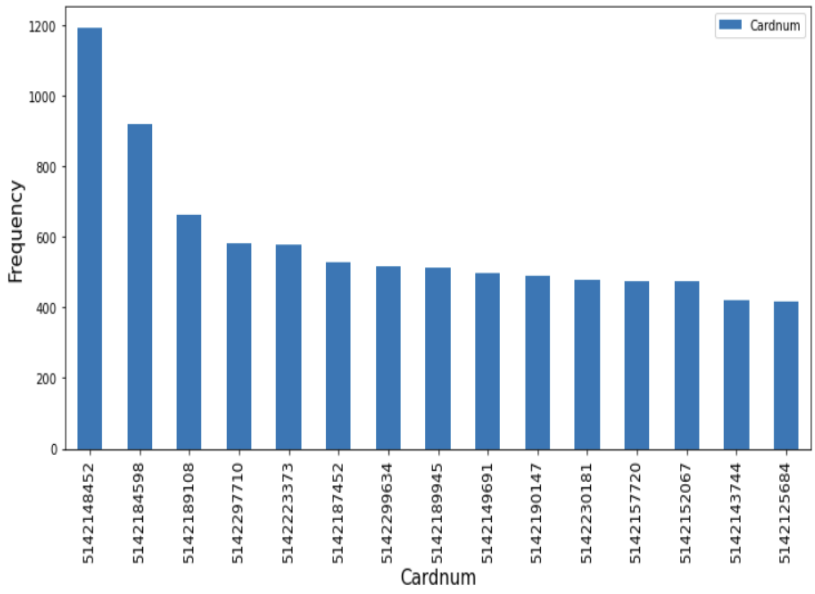


Fig 5. Bar Plot showing distribution of Cardnum

5. **Amount:**
The field Amount provides information on the amount of the money involved in the transaction. The transaction with the maximum amount was of 3102,045.5 dollars. From the graph below it can be seen that a large number of transactions are of low amounts.

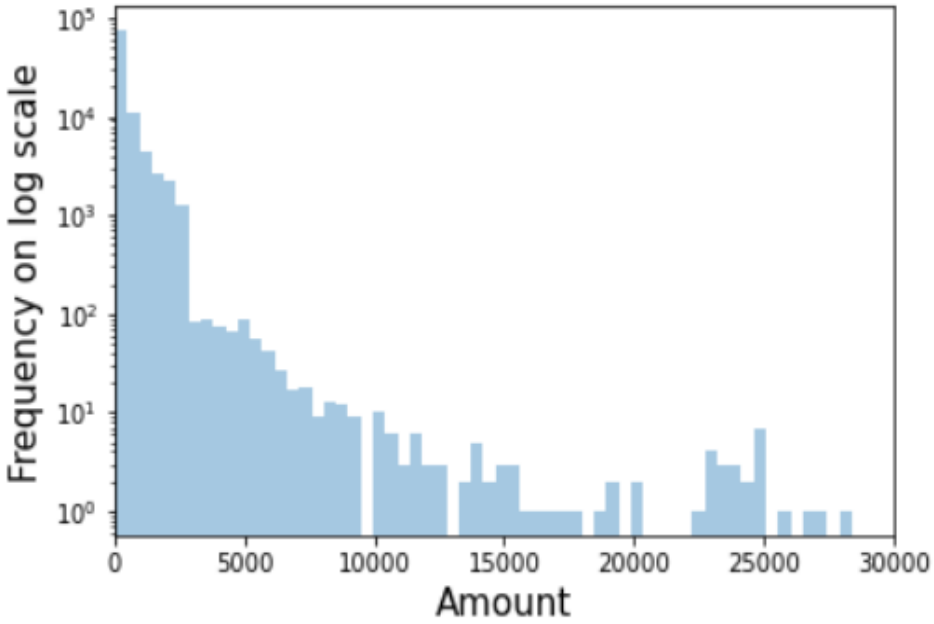


Fig 6. Histogram showing distribution of Amount

Data Cleaning

Removing Exclusions

- We only kept those records where ‘Transtype’ was P.
- We considered 100000$ as the threshold above which the purchase amount would be deemed as outliers. So, we kept the records where ‘Amount’ was less than 100000$ to narrow the decision region for the model.

Field Imputations

- Filling in ‘Merch state’: (98.7% populated)
- For the records that had zip number but missing state, we found the corresponding states to which the zip numbers belong and filled in the missing values by mapping to those states.
- For the rest of the missing states, if the records had ‘Merchnum’ values, we used mode of the states i.e., the state associated maximum number of times with the corresponding Merchnum value. Otherwise, we had used ‘Merch description’ field’s values to fill in the missing states in a similar way.
- For records that were still blank, i.e., where ‘Merch description’ was ‘RETAIL CREDIT ADJUSTMENT’ and ‘RETAIL DEBIT ADJUSTMENT’, we used ‘unknown’ to fill in.
- Filling in ‘Merch zip’: (95.2% populated)
- We filled in the missing zip numbers for records that have ‘Merchnum’ and ‘Merch description’ by mapping with the mode of Merch zip i.e., the zip number that has been associated with the maximum number of corresponding Merchnum or Merch description.
- For records that were still blank, i.e., where ‘Merch description’ was ‘RETAIL CREDIT ADJUSTMENT’ and ‘RETAIL DEBIT ADJUSTMENT’, we used ‘unknown’ to fill in.
- Filling in ‘Merchnum’: (96.5% populated)
- We filled in the missing Merchnum for records that have ‘Merch description’ by mapping with the mode of Merchnum i.e., the zip number that has been associated with the maximum number of corresponding ‘Merch description’ values.
- For records that were still blank, i.e., where ‘Merch description’ was ‘RETAIL CREDIT ADJUSTMENT’ and ‘RETAIL DEBIT ADJUSTMENT’, we used ‘unknown’ to fill in.

Target Encoding

When we have labeled data, i.e., for supervised learning, we can do **Target Encoding** of categorical fields. For each possible category assign a numerical value. Several choices to use for the value. We usually use the average (median in case of highly skewed values for that category) of the target or dependent variables for all records in that category, which in this case is basically the average number of frauds per category of that field.
Advantage - direct encoding to what you’re trying to predict, no increase in dimensionality.
Disadvantage - loss of interaction information, danger of overfitting.

To avoid overfitting:

- Before applying Target encoding, we had split the data into train-test or modeling data and OOT data (November and December). We used only training or “past” data for target encoding.

- We target encoded only DOW and Merchant State, as these two fields had >2 cardinality, yet they had enough number of records per category to be shown as examples to the model
- We also used exponential smoothing, i.e., applied an exponential function to smoothly transition between the overall target mean (y_low in the formula) and the target mean (y_high) for each individual category. N is the number of records of the corresponding category, c is the rate of transition. We chose c as 4 and nmid as 20.

$$\text{Value} = Y_{\text{low}} + \frac{Y_{\text{high}} - Y_{\text{low}}}{1 + e^{-(n - n_{mid})/c}}$$

Risk Tables

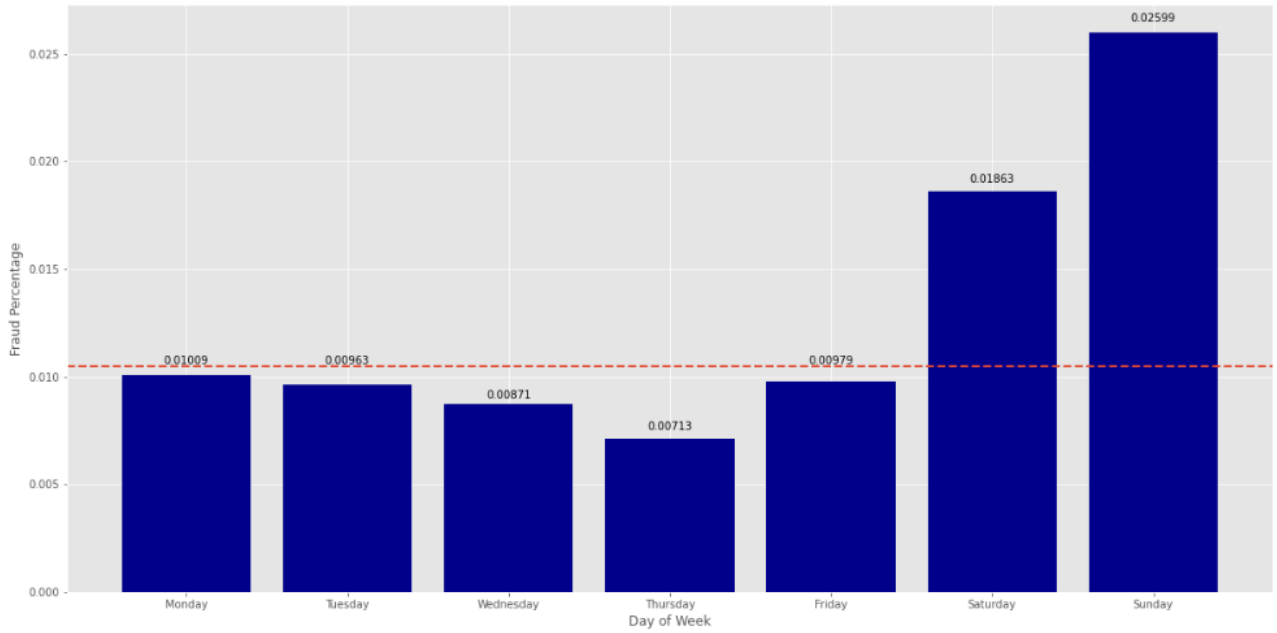


Fig 7. showing percentage of frauds per day of week

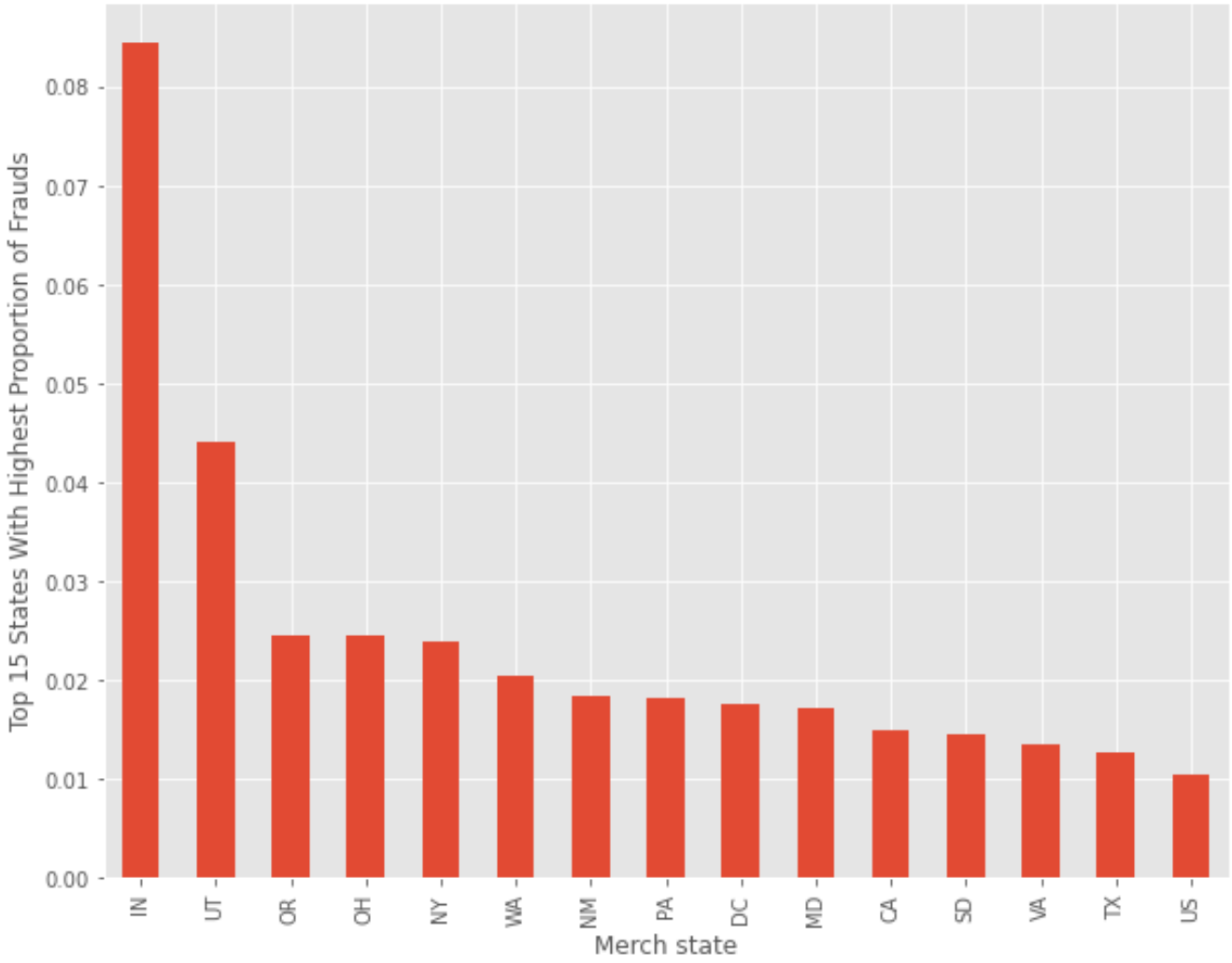


Fig 8. Percentage of frauds per day of Merchant State

Feature Engineering

Features can be defined as any measurable inputs to the model. Important information can get lost amidst the noise and competing signals in a large feature space. Feature engineering allows us to create and measure additional variables that can potentially improve model performance and better emphasize the trends in the data. Feature Engineering is the process of preparing, transforming, and extracting features from raw data by using domain knowledge and various modeling techniques to provide the best inputs to the model. Often, variables become more informative by combining or decomposing into two or more variables. After all, the quality of any model is limited by the quality of the data you feed into it.

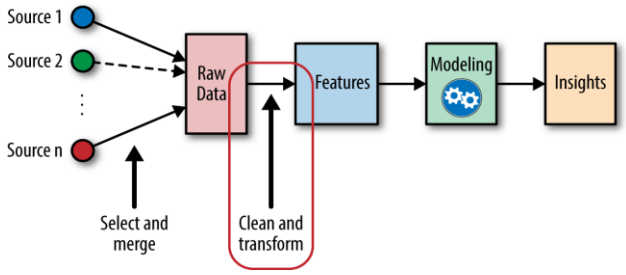


Fig 9. Feature Engineering phase in the overall insights process

Hence, we created the following types of variables:

a. Categorical variables

Additional categorical variables were created before creating candidate variables for our analysis and models. These categorical variables would be used to create candidate predictor variables by using domain knowledge and different combinations of existing fields. We created eight categorical variables for our project, as shown below. Check Appendix B

	Variable Name	Description
1	Cardnum_Merchnum	Card number + Merchant Number
2	Cardnum_MerchState	Card number + Merchant State
3	Cardnum_MerchZip	Card number + Merchant Zip code
4	Cardnum_MerchDescription	Card number + Merchant Description
5	Merchnum_MerchZip	Merchant Number + Merchant Zip code
6	Merchnum_MerchState	Merchant Number + Merchant State
7	Cardnum_Merchnum_MerchState	Card number + Merchant Number + Merchant State
8	Cardnum_Merchnum_MerchZip	Card number + Merchant Number + Merchant Zip code

Table 3: Categorical Variables

Hence, the final set of candidate variables or entities are as follows.

	Entitiy Name
1	Cardnum
2	Merchnum
3	MerchState
4	MerchZip
5	MerchDescription
6	Cardnum_Merchnum
7	Cardnum_MerchState
8	Cardnum_MerchZip
9	Cardnum_MerchDescription
10	Merchnum_MerchZip
11	Merchnum_MerchState
12	Cardnum_Merchnum_MerchState
13	Cardnum_Merchnum_MerchZip

Table 4: Entities

b. Amount Variables

Amount variables contain a statistical summary of the amount with the entities ranging over eight time periods. The statistical summary refers to the average, maximum, median, total, actual by average, actual by maximum, actual by median, and actual by total. The eight time periods are 0, 1, 3, 7, 14, 30, 60, and 90. Thus a total of 832 amount variables were created.

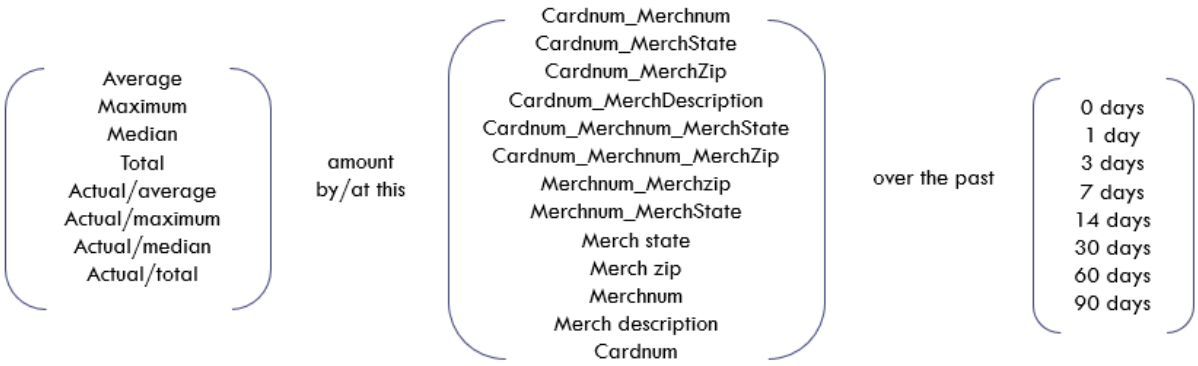


Fig 10. Amount variables combinations map

	Variable Name
1	Cardnum_Merchnum_avg_30
2	Cardnum_Merchnum_max_30
3	Cardnum_Merchnum_med_30
4	Cardnum_Merchnum_total_30
5	Cardnum_Merchnum_actual/avg_30
6	Cardnum_Merchnum_actual/max_30
7	Cardnum_Merchnum_actual/med_30
8	Cardnum_Merchnum_actual/total_30

Table 5: Sample Amount Variables for Cardnum_Merchnum entity

c. Frequency Variables

Frequency variables tracks the number of times the entity was encountered over the eight time periods. The speed at which these transactions appear in our dataset is a way of detecting and identifying potentially fraudulent transactions. A higher frequency value would indicate a greater likelihood of a fraudulent transaction. A total of 102 velocity frequency variables were created.

Fig 11. Frequency variables combinations map.

	Variable Name
1	Cardnum_Merchnum_count_0
2	Cardnum_Merchnum_count_1
3	Cardnum_Merchnum_count_3
4	Cardnum_Merchnum_count_7
5	Cardnum_Merchnum_count_14

Table 6: Sample Frequency Variables for Cardnum_Merchnum entity

d. Days-Since Variables

Days-since, tracks the number of days since the entity was last seen. Days-since returns a whole number for the number of days since last seen. If the entity occurs more than once on the same date, then 1 is returned in the days-since the field of that entity for that record. A total of 13 days-since variables were created using the combination mentioned in the below figure.

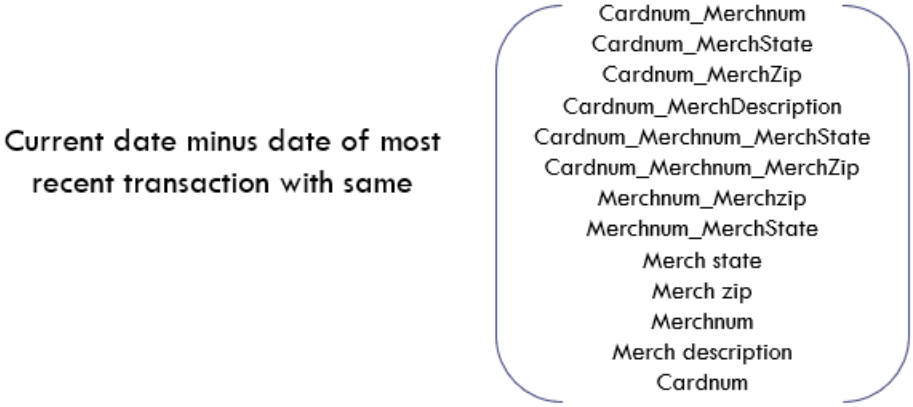


Fig 12. Days-Since variables combinations map.

	Variable Name
1	Cardnum_day_since
2	Merch state_day_since
3	Merchnum_day_since

Table 7: Sample Days-Since Variables entities

e. Velocity Change Variables

Velocity change refers to the speed at which an entity is seen in the dataset for a particular transaction record over a short period of time (0 - 1 days) in relation to how often the same entity is seen over a longer period (3 – 90 days). The speed at which these transactions happen in a shorter timeframe versus a longer time frame is a way to detect and identify potentially fraudulent transactions. A higher value of velocity change would indicate a greater likelihood of a fraudulent transaction. A total of 156 velocity change variables were created using the below mention formula.

$$\frac{\text{Number/amount of transactions with same [Entities] over past } \{0, 1\} \text{ days}}{\text{Average Number/amount of transactions with same [Entities] over past } \{3, 7, 14, 30, 60, 90\} \text{ days}}$$

Variable Name
Cardnum_Merchnum_MerchState_count_0_by_3
Cardnum_Merchnum_MerchState_count_0_by_7
Cardnum_Merchnum_MerchState_count_0_by_14
Cardnum_Merchnum_MerchState_count_0_by_30
Cardnum_Merchnum_MerchState_count_0_by_60
Cardnum_Merchnum_MerchState_count_0_by_90
Cardnum_Merchnum_MerchState_count_1_by_3
Cardnum_Merchnum_MerchState_count_1_by_7
Cardnum_Merchnum_MerchState_count_1_by_14
Cardnum_Merchnum_MerchState_count_1_by_30
Cardnum_Merchnum_MerchState_count_1_by_60
Cardnum_Merchnum_MerchState_count_1_by_90

Table 8: Sample Velocity Change Variables for Cardnum_Merchnum_MerchState entity

f. Benford's Law Variables

Benford's law, also known as the Newcomb–Benford law, the law of anomalous numbers, or the first-digit law, is an observation that in many real-life sets of numerical data, the leading digit is likely to be a small number, e.g., 1 or 2. In sets that obey the law, the number 1 appears as the significant leading digit about 30 % of the time, while 9 appears as the significant leading digit less than 5 % of the time.

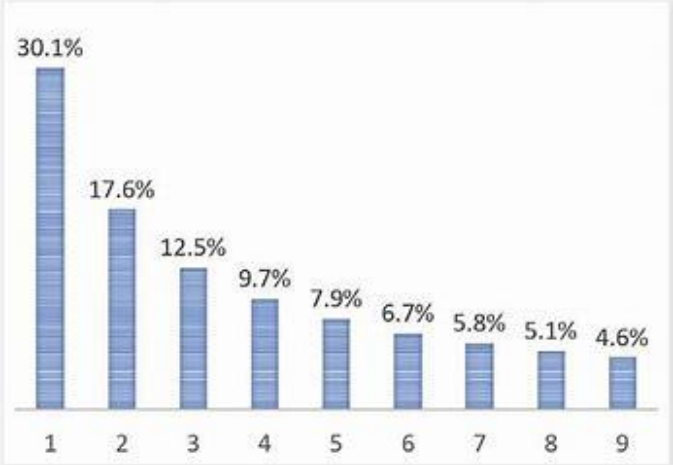


Fig 13. Benford's Law leading digit frequency of occurrence trend.

We checked the amount distribution for each Card number and Merchant number to see if they violate the Benford's law by following the steps below.

1. Removed the transactions from FedEx, as they were unusual.
2. Grouped the transactions by cardnum and merchnum.
3. Measure the unusualness of the first digit of the amount (We looked for 1s or 2s).
4. Smoothing.

By following the above steps, we created 2 Benford's law variables.

#	Variable
1	Cardnum_U*
2	Merchnum_U*

Table 9: Benford Law Variables

Additionally, we also created Day of week variable and risk per merchant state variable to enhance our understanding of the data. Finally, at the end of feature engineering we had a total of 1130 variables.

Feature Selection Process

Feature selection is the process in statistics that aims to reduce the number of features (dimensions) of input variables. Feature selection is performed in such a way so that the most statistically important features are kept from the original input. Also, features that are either highly correlated with others or not significant to perform an accurate prediction are ignored. Feature Selection is important to reduce dimensionality.

Benefits of feature Selection:

- Enables the machine learning algorithm to train faster
- Reduce the complexity
- Improves the accuracy
- Reduces overfitting

There are three main methods of feature selection:

Filter Methods: Filter methods use statistical measures to evaluate the relationship (correlation) of two distributions and measure the correlation between the distribution of each of the classes of each feature and the dependent variable. The features that are chosen are the ones with the highest correlation with the dependent variable.

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

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

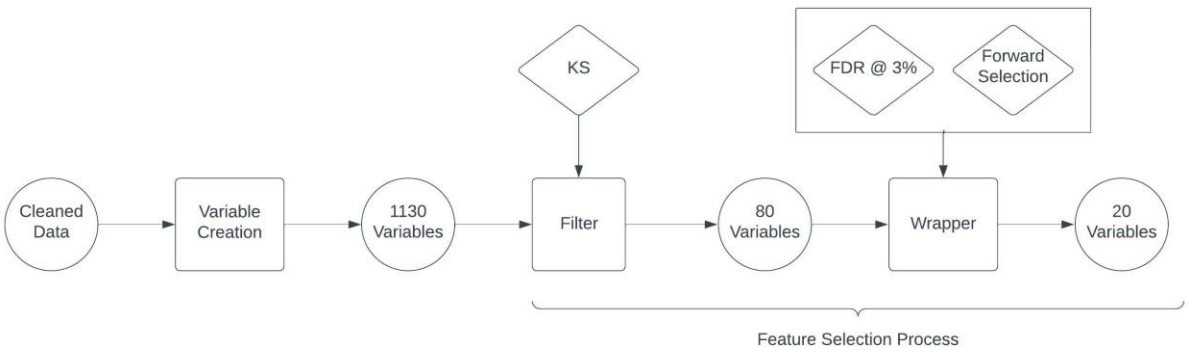


Fig 14. Feature Selection Flowchart

Filter Methods

Purpose of Filter method is to determine how relevant each variable is in predicting dependent variable on its own. For our analysis, we performed feature selection using the Kolmogorov-Smirnov distance and fraud detection rate.

Kolmogorov–Smirnov (KS) distance:

- KS is a statistical measure of how well two normalized distributions are separated into good label and a bad label population.
- We calculated the difference in KS between these two distributions. More difference means the variable is more important.

For this analysis, we used the KS distance metric by calculating the univariate KS value as a filtering method to aid in determining which features provide a better separation between the “Fraud” values of 1 and 0. Meaning, for each numerical candidate variable, we generated the distribution of the two classes (1 and 0) based on the dependent variable ("Fraud” data field). Subsequently, we measured the KS distance between the distributions of the two classes for each of the numerical candidate variables.

More formally:

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

We then rank ordered the KS distance value from high to low for each of the numerical candidate variables. We only kept only 80 variables with the highest score since they were the most significant.

Fraud Detection Rate (FDR) at 3%:

The second metric we used for filtering the features was the Fraud Detection Rate (FDR). In general, the FDR is the percentage of all the frauds that are detected up to a particular cutoff point. In the context of analysis for feature selection, we used a cutoff threshold of 3% and calculated the univariate FDR for each numerical candidate variable. The FDR at 3% was determined by first sorting the numerical candidate variables in descending order, and then computing the percentage of frauds in the top 3%. We then assigned a rank for each of the numerical candidate variables and used this ranking to evaluate the importance of each variable.

In our analysis, we obtained the univariate KS distance and the univariate FDR at 3% for each numerical candidate variable and used their average rank to serve as a final score for the importance of each candidate variables. Subsequently, we discarded the lowest 260 ranked numerical candidate variables, reducing the total number of features from 340 to 80. In figure 14 we present the KS, FDR and final ranking score normalized from 0 to 1 in descending order (higher ranking score indicates higher feature importance). The score threshold is shown by the green dashed line and indicates the score cutoff of the top 80 features. Also, note that the “Fraud” column had KS and FDR scores of 1 which validates our filter method.

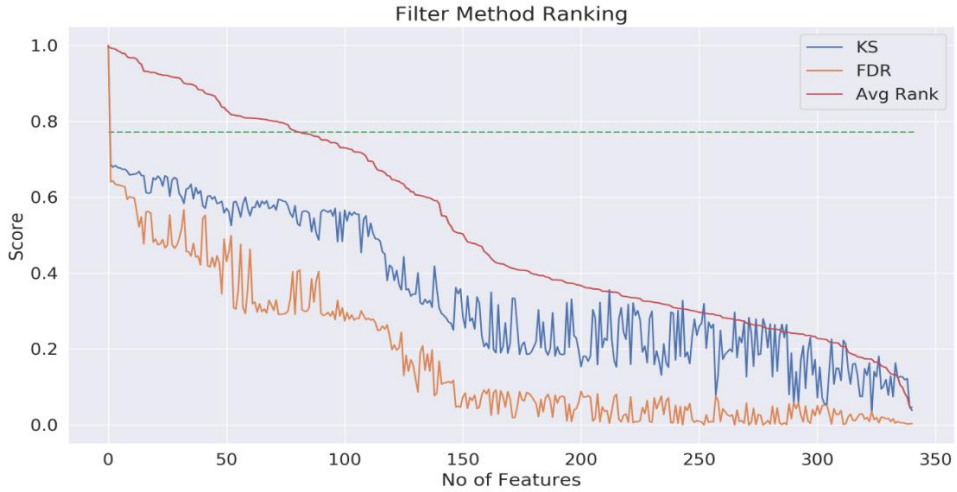


Fig 15. Filter Method Ranking

Wrapper Methods

We used Decision Trees, Random Forest, and Boosted Trees as our models for the wrapper method for feature (candidate variable) selection. These models are simple to implement with a relatively low computational cost for their default parameters. Common measures for feature importance in these models are the mean decrease in accuracy (misclassification) or the mean decrease in node impurity (Gini index), which represents how well the trees split the data. We evaluated feature importance as the decrease in node impurity weighted by the probability of reaching that node in each model. Like the filter method, we then ranked the feature importance score for each model and averaged the ranking to obtain a final rank for each feature. With our final ranking we then kept the top 20 features for our models.

Forward Selection

- We used forward selection method that starts with no features in the model
- In each Iteration, we keep adding the feature which best improves our model till no new variable could improve the model performance.

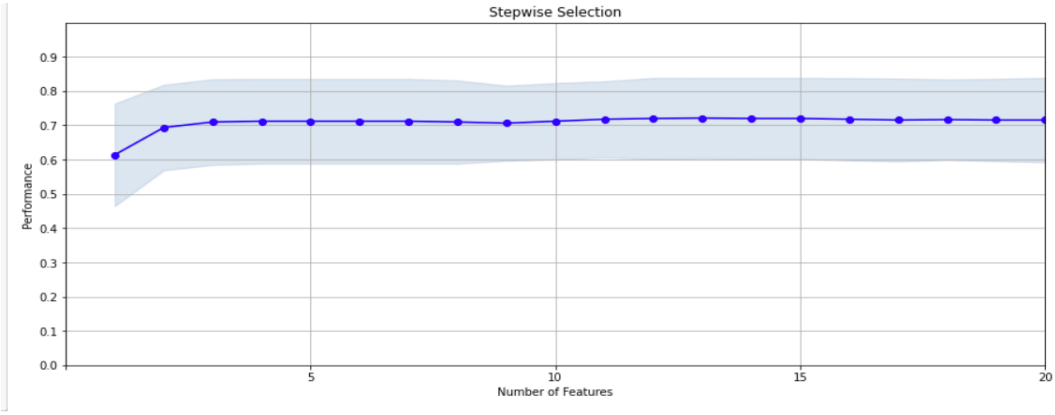


Fig 16. Wrapper graph

- The graph above, shows the changes in model performance as the number of feature increase during forward selection.
- If we read the graph, the performance after five features do not have significant improvement in performance when the number of features increase.

Once we decided which independent variables to use from this graph and we prepare those for analysis on first 10 months of our training data set.

Rank	Variable name	avg_score
1	Cardnum_MerchDescription_total_3	0.614
2	Cardnum_MerchState_max_14	0.694
3	Cardnum_MerchDescription_max_30	0.710
4	Cardnum_Merchnum_MerchZip_max_60	0.711
5	Cardnum_MerchZip_total_90	0.713
6	Cardnum_MerchDescription_total_1	0.718
7	Cardnum_Merchnum_MerchZip_max_14	0.719
8	Cardnum_MerchDescription_max_1	0.724
9	Cardnum_Merchnum_MerchZip_max_30	0.724
10	Cardnum_MerchZip_max_14	0.722
11	Cardnum_MerchDescription_max_3	0.720
12	Cardnum_Merchnum_max_1	0.721
13	Cardnum_MerchZip_max_3	0.722
14	Cardnum_Merchnum_MerchZip_max_3	0.722
15	Cardnum_MerchState_max_1	0.718
16	Cardnum_Merchnum_MerchState_max_1	0.718
17	Cardnum_MerchZip_max_1	0.718
18	Cardnum_Merchnum_max_14	0.717
19	Cardnum_MerchDescription_max_7	0.716
20	Cardnum_Merchnum_MerchState_max_7	0.717

Table 10: 20 Final Variables Sorted by Wrapper

- The score of these variables is quite different and not in order, because algorithm chooses the best features to maximize model performance.
- Although one variable itself may not be as significant as some variables not selected during forwarding selection. It can best contribute to the model performance when it is added to a specific feature set.

Model Algorithms:

After selecting the top 20 best variables, with each having a wrapper score above 0.5. We start with a logistic regression to get a base line model and then test Decision Tree, Random Forest, Boosted Tree, and Neural Network models with varying hyperparameters to choose the best models by comparing the Fraud Detection Rate (FDR) at 3% for the train, test and out of time (OOT) datasets. The following are the details of each test conducted.

Logistics Regression:

The logistic regression is one of the most popular classification algorithms. In logistic regression, a linear output is converted into a probability between 0 & 1 using the sigmoid function.

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}$$

In the equation above, X is the set of predictor features and b is the corresponding vector of weights. Computing S(x) above produces a probability that indicates if an observation should be classified as “1” (if the calculated probability is at least 0.5), and “0” otherwise. It’s an S-shaped curve that can take a real-valued number and map it into a value between 0 and 1, but never

exactly at those limits.

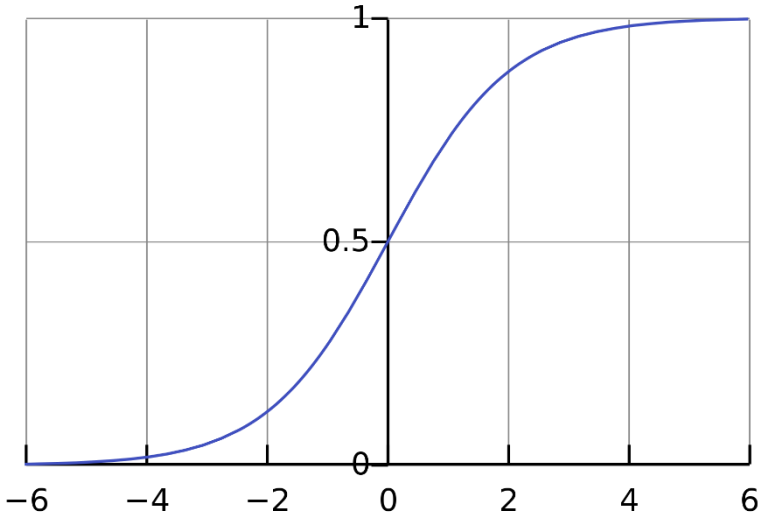


Fig 17. Logistic regression

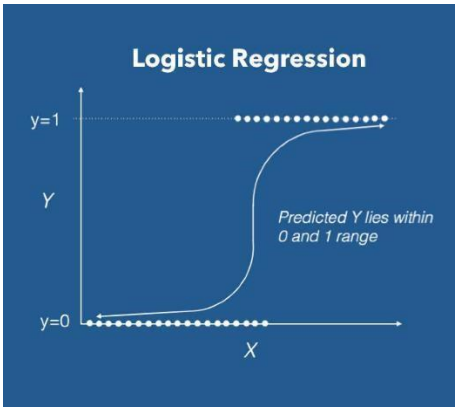


Fig 18. Sample Logistic regression curve

Below is an example logistic regression equation:

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$

Where y is the predicted output, b_0 is the bias or intercept term and b_1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The intercept term controls the location of the midpoint of the curve and b_1 controls the raise of the curve.
More simply put, here is how the Logistic Regression equation for Machine Learning looks like: $\text{logit}(p) = \ln(p/(1-p)) = h_0+h_1X1+h_2X_2+h_3X_3\ldots.+h_kX_k$

p = probability of the occurrence of the feature

$x_1, x_2, \ldots, x_k$ = set of input features $h_1, h_2, \ldots h_k$ = parametric values to be estimated in the Logistic Regression equation.

Syntax: sklearn.linear_model.LogisticRegression()

For this project's fraud analysis, five versions of logistic regression were created by changing the solver, penalty and c hyperparameters and training the model with the identified 30 best variables. To understand the used hyperparameters further:

- Solver: This parameter represents which algorithm to use in the optimization problem.
    - liblinear − It is a good choice for small datasets. It also handles L1 penalty. For multiclass problems, it is limited to one-versus-rest schemes.
    - lbfgs − For multiclass problems, it handles multinomial loss. It also handles only L2 penalty.
- Default is 'lbfgs'.
- Penalty: Penalized logistic regression imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contributive variables toward zero. This is also known as regularization. L1 is therefore useful for feature selection, as we can drop any variables associated with coefficients that go to zero. L2, on the other hand, is useful when you have collinear/codependent features. Default is 'L2'
- C: It represents the inverse of regularization strength, which must always be a positive float. Smaller values specify stronger regularization. The results of the logistic regression are shown below.

Model		Parameters				Average FDR at 3%		
		solver	penalty	C	max_iter	Train	Test	OOT
Logistic Regression		liblinear	l1	0.01	100	0.64	0.609	0.344
		saga	l1	0.01	100	0.626	0.629	0.332
		lbfgs	l2	0.01	100	0.631	0.617	0.346
		lbfgs	l2	0.1	100	0.629	0.641	0.315
		saga	l1	0.1	100	0.637	0.629	0.337
		liblinear	l1	0.1	100	0.632	0.649	0.345
		saga	l1	0.1	1000	0.641	0.631	0.342

Table 11: Logistic Regression Model Results

The best results were given when solver = liblinear, penalty = l2, c = 1.

Decision Tree:

In decision analysis, a decision tree can be used to represent decisions and decision making visually and explicitly. The main goal of Decision Trees is to create a model predicting target variable value by learning simple decision rules deduced from the data features. Decision trees have two main entities; one is root node, where the data splits, and other is decision nodes or leaves, where we got final output. In three-dimensional view, decision trees approximate the surface into y = f(x) with steps or platforms. These steps form boxes and each box contains the average of the dependent variable y for its range.

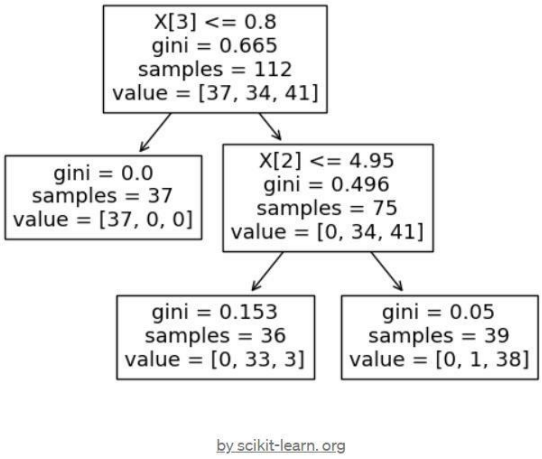


Fig 19. Sample Decision Tree

The decision trees decide the cut point of these boxes by measuring the impurity of the resulting boxes to calculate the goodness of the candidate. Common measures of impurity are variance, Gini index and Entropy. Best cut point has the lowest impurity.
Syntax: sklearn.tree.DecisionTreeClassifier()
For this project's fraud analysis, four versions of logistic regression were created by changing the max_depth, min_sample_leaf and min_samples_split hyperparameters and training the model with the identified 30 best variables. To understand the used hyperparameters further:

- max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Default is None
- min_sample_leaf: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf

training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression. Default is 1.

- min_samples_split: The minimum number of samples required to split an internal node. Default is 2.

The results of the Decision Trees are shown below.

		max_depth	min_sample_leaf	min_samples_split	min_weight_fraction_leaf	Train	Test	OOT
Decision Tree		10	1	2	-	0.522	0.515	0.495
		20	60	300	-	0.524	0.511	0.502
		20	60	320	-	0.525	0.514	0.5
		25	80	600	-	0.516	0.533	0.498
		10	70	20	0.0001	0.82	0.762	0.436
		100	70	100	0.0001	0.767	0.767	0.447
		10	70	20	0.0001	0.826	0.773	0.434
		150	100	200	0.0001	0.806	0.744	0.414

Table 12: Decision Tree model results

The best results were given when max_depth = 25, min_sample_leaf = 80, min_samples_split = 600.

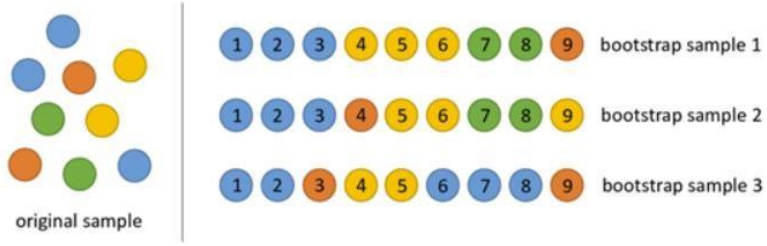
Random Forest

Random forest is an ensemble of many decision trees. Random forests are built using a method called bagging in which each decision trees are used as parallel estimators. If used for a classification problem, the result is based on average prediction from each decision tree. For regression, the prediction of a leaf node is the mean value of the target values in that leaf.

Random forest regression takes mean value of the results from decision trees.

Random forests reduce the risk of overfitting and accuracy is much higher than a single decision tree. Furthermore, decision trees in a random forest run in parallel so that the time does not become a bottleneck.

The success of a random forest highly depends on using uncorrelated decision trees. If we use same or very similar trees, overall result will not be much different than the result of a single decision tree. Random forests achieve to have uncorrelated decision trees by bootstrapping and feature randomness.



Random Forests follow the same basic principle as that of Decision Trees, however, Random Forests train multiple Decision Trees by a randomly chosen subset of variables or

records for each tree and each split of the tree. It then gets the prediction from each of them and finally selects the best solution by means of voting (for classification type data) or by averaging (for regression type data) . It can be used for both classification as well as regression tasks.

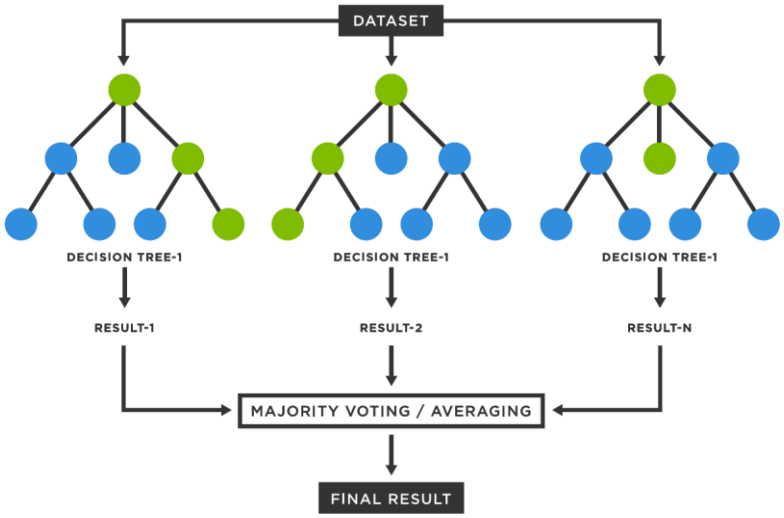


Fig 20. Sample working principle of Random Forest Model

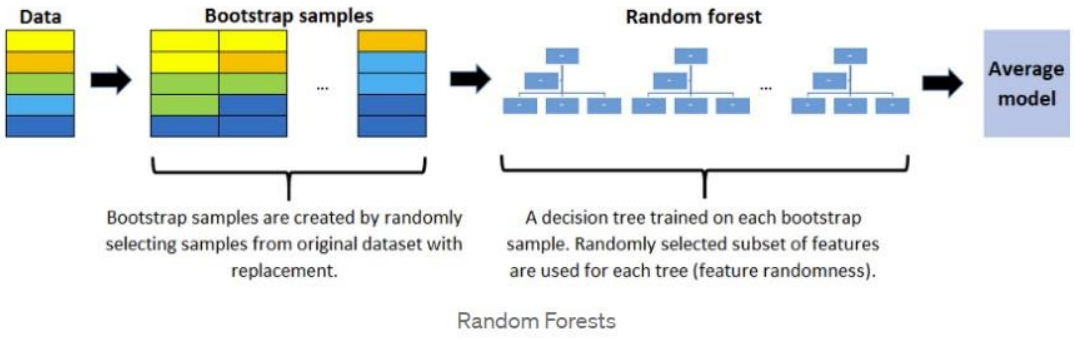


Fig 21. Random Forest Model

Syntax: sklearn.ensemble.RandomForestClassifier() For this project's fraud analysis, four versions of Random Forests were created by changing the n_estimators, max_depth, max_features, min_samples_leaf, and min_samples_split hyperparameters and training the model with the identified 30 best variables. To understand the used hyperparameters further:

- n_estimators: The number of trees in the forest. default=100
- max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- default=None
- max_features: The number of features to consider when looking for the best split.

- default=”auto”
- min_samples_leaf: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression. Default is 1.
- min_samples_split: The minimum number of samples required to split an internal node. Default is 2.

The results of the Random Forest models are shown below.

		n_estimators	max_depth	max_features	min_samples_leaf	min_samples_split	Train	Test	OOT
Random Forest		100	10	5	1	2	1	0.81	0.439
		300	10	-	20	100	0.838	0.808	0.474
		150	-	7	30	15	0.865	0.838	0.513
		150	3	7	30	15	0.664	0.667	0.368
		200	-	7	30	15	0.874	0.754	0.48

Table 13: Random Forest model results

The best results were given when n_estimators = 50, max_depth = 20, max_features = 5, min_samples_leaf = 30 and min_samples_split = 500.

Boosted Tree

Gradient boosting algorithm sequentially combines weak learners in way that each new learner fits to the residuals from the previous step so that the model improves. The final model aggregates the results from each step and a strong learner is achieved. Gradient boosted decision trees algorithm uses decision trees as week learners. A loss function is used to detect the residuals. For instance, mean squared error (MSE) can be used for a regression task and logarithmic loss (log loss) can be used for classification tasks. It is worth noting that existing trees in the model do not change when a new tree is added. The added decision tree fits the residuals from the current model. The steps are as follows:

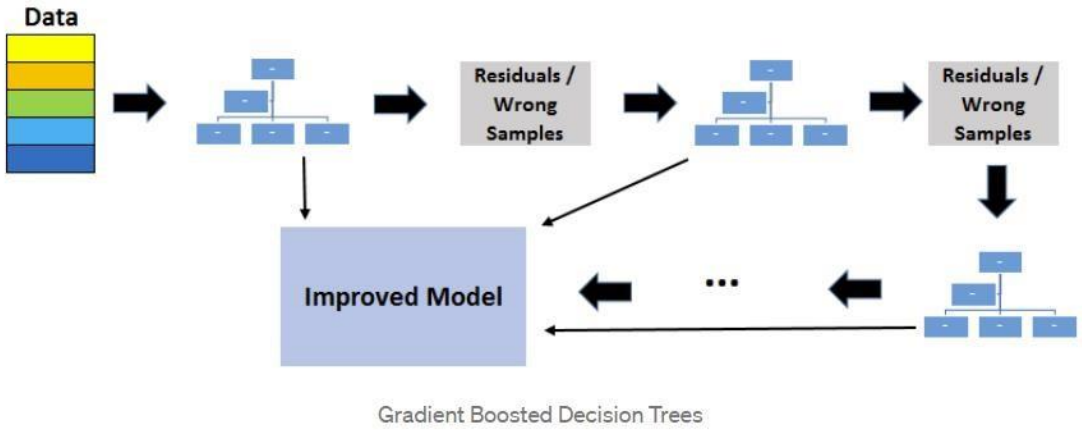


Fig 22. Sample working principle of Boosted Tree Model

		learning_rate	n_estimators	max_depth	num_leaves	Train	Test	OOT
Boosted Tree		0.1	100	3	20	0.841	0.805	0.505
		0.1	50	4	30	0.854	0.804	0.4871
		0.1	50	5	20	0.891	0.788	0.479
		0.1	100	5	20	0.873	0.814	0.526
		0.001	**300**	**-1**	**31**	**0.842**	**0.772**	**0.564**
		0.001	4000	-1	25	0.912	0.491	0.491
		0.001	300	-1	30	0.905	0.525	0.525
		0.0001	100	10	30	0.717	0.717	0.469
		0.0001	200	-100	50	0.781	0.781	0.462

Table 14: Boosted Tree model results

Learning rate and n_estimators

Hyperparameters are key parts of learning algorithms which effect the performance and accuracy of a model. Learning rate and n_estimators are two critical hyperparameters for gradient boosting decision trees. Learning rate, denoted as α, simply means how fast the model learns. Each tree added modifies the overall model. The magnitude of the modification is controlled by learning rate. The steps of gradient boosted decision tree algorithms with learning rate introduced:

- $f_1(x) \approx y$
- The residual is $y - \alpha f_1(x)$
- $f_2(x) \approx y - \alpha f_1(x)$
- The residual is $y - \alpha f_1(x) - \alpha f_2(x)$
- $f_3(x) \approx y - \alpha f_1(x) - \alpha f_2(x)$

Gradient boosted decision tree algorithm with learning rate (α)

The lower the learning rate, the slower the model learns. The advantage of slower learning rate is that the model becomes more robust and generalized. In statistical learning, models that learn slowly perform better. However, learning slowly comes at a cost. It takes more time to train the model which brings us to the other significant hyperparameter. n_estimator is the number of trees used in the model. If the learning rate is low, we need more trees to train the model. However, we need to be very careful at selecting the number of trees. It creates a high risk of overfitting to use too many trees.

Neural Network

Neural network function in a way like biological neurons. They take in various inputs (at input layers), weight these inputs, and then combine the weighted inputs through a linear combination (much like linear regression). If the combined weighted output is past some thresholds set by an activation function, the output is then set out to other layers. The base unit is generally referred to as a perceptron. Perceptron's are combined to form neural networks, which is why they are also called as multi-layer perceptron's (MLPs).

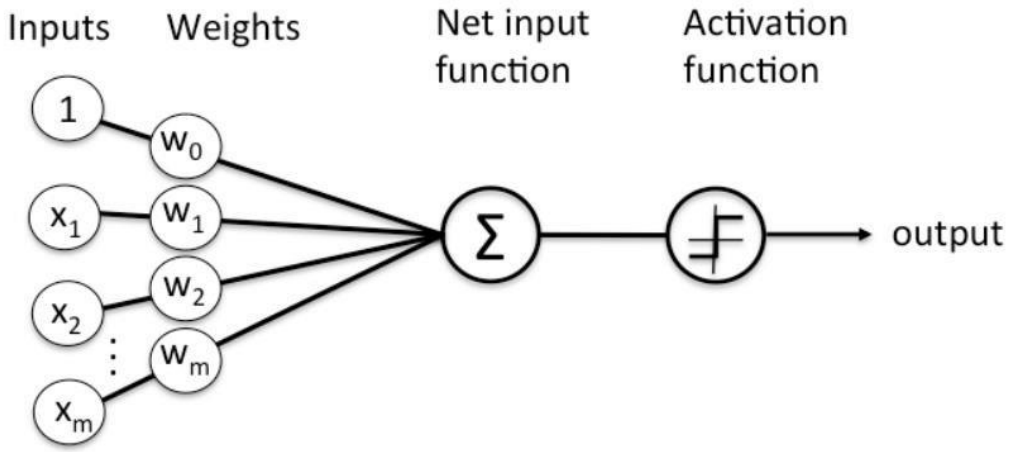


Fig 23. Sample working principle of Neural Network Model

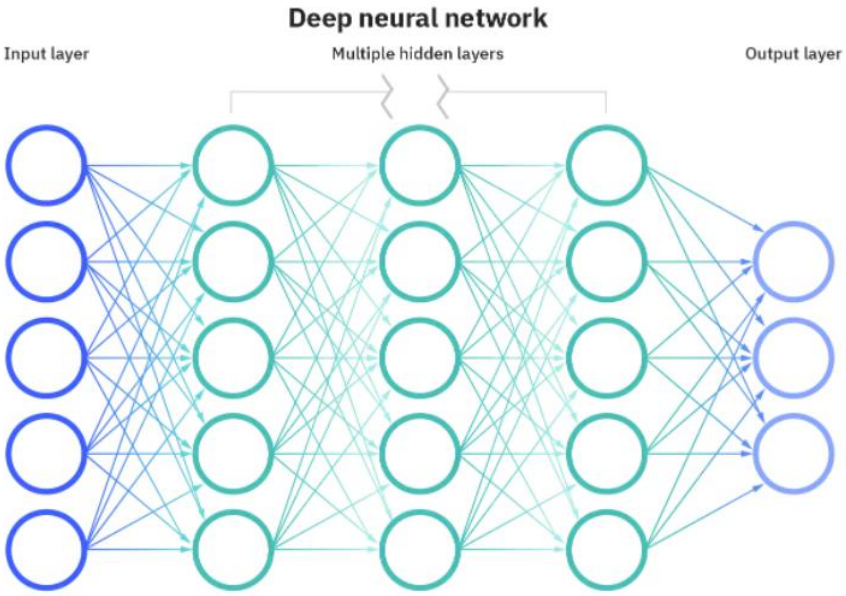


Fig 24. Neural Network Model with multiple hidden layers

The process of receiving inputs and generating an output continues until an output layer is reached. This is generally done in a forward manner, meaning that layer's process incoming data in a sequential forward way. The layers of neurons that are not input or output layers are called the hidden layers. Hidden layers allow for a specific transformation of the data within each layer. Each hidden layer can be specialized to produce a particular output.
The learning process for neural networks is called backpropagation. This technique modifies the weights of neural network iteratively through the calculation of deltas between predicted and expected outputs. After this calculation, the weights are updated backwards through earlier layers via stochastic gradient descent. The process continues until the weights that minimize the loss function is found.

		activation	max_iter	learning_rate	alpha	solver	hidden layer sizes	Train	Test	OOT
Neural Network		relu	140	constant	0.001	adam	20,20	0.7381	0.727	0.53
		relu	140	constant	0.001	adam	100,	0.71	0.705	0.52
		relu	1000	constant	0.001	adam	100,	0.732	0.705	0.519
		relu	-	constant	0.001	adam	-	0.727	0.718	0.47
		relu	1000	adaptive	0.001	adam	100,	0.749	0.704	0.547
		relu	500	constant	0.001	lbfgs	100,	0.741	0.742	0.463
		tanh	1000	adaptive	0.001	adam	100,	0.716	0.704	0.506
		tanh	1000	adaptive	0.01	adam	100,	0.709	0.679	0.464
		tanh	1000	constant	0.01	adam	100,	0.698	0.694	0.435
		tanh	1000	adaptive	0.001	sgd	50,	0.624	0.605	0.341

Table 15: Neural Network model results

Results:

The below diagram gives us the results from all models that were tested to predict fraud.

Model		Parameters						Average FDR at 3%		
Logistic Regression		solver	penalty	C		max_iter		Train	Test	OOT
		liblinear	l1	0.01		100		0.64	0.609	0.344
		saga	l1	0.01		100		0.626	0.629	0.332
		lbfgs	l2	0.01		100		0.631	0.617	0.346
		lbfgs	l2	0.1		100		0.629	0.641	0.315
		saga	l1	0.1		100		0.637	0.629	0.337
		liblinear	l1	0.1		100		0.632	0.649	0.345
		saga	l1	0.1		1000		0.641	0.631	0.342
Decision Tree		max_depth	min_sample_leaf	min_samples_split		min_weight_fraction_leaf		Train	Test	OOT
		10	1	**2**		-		0.522	0.515	0.495
		20	60	**300**		-		0.524	0.511	0.502
		20	60	**320**		-		0.525	0.514	0.5
		25	80	**600**		-		0.516	0.533	0.498
		10	70	20		0.0001		0.82	0.762	0.436
		100	70	100		0.0001		0.767	0.767	0.447
		10	70	20		0.0001		0.826	0.773	0.434
		150	100	200		0.0001		0.806	0.744	0.414
Random Forest		n_estimators	max_depth	max_features	min_samples_leaf	min_samples_split		Train	Test	OOT
		100	10	5	1	**2**		1	0.81	0.439
		300	10	-	20	**100**		0.838	0.808	0.474
		150	-	7	30	**15**		0.865	0.838	0.513
		150	3	7	30	**15**		0.664	0.667	0.368
		200	-	7	30	**15**		0.874	0.754	0.48
Boosted Tree		learning_rate	n_estimators	max_depth	num_leaves			Train	Test	OOT
		0.1	100	3	**20**			0.841	0.805	0.505
		0.1	50	4	**30**			0.854	0.804	0.4871
		0.1	50	5	**20**			0.891	0.788	0.479
		0.1	100	5	**20**			0.873	0.814	0.526
		0.001	**300**	**-1**	**31**			**0.842**	**0.772**	**0.564**
		0.001	4000	-1	25			0.912	0.491	0.491
		0.001	300	-1	30			0.905	0.525	0.525
		0.0001	100	10	30			0.717	0.717	0.469
		0.0001	200	-100	50			0.781	0.781	0.462
Neural Network		activation	max_iter	learning_rate	alpha	solver	hidden layer sizes	Train	Test	OOT
		relu	140	constant	0.001	adam	20,20	0.7381	0.727	0.53
		relu	140	constant	0.001	adam	100,	0.71	0.705	0.52
		relu	1000	constant	0.001	adam	100,	0.732	0.705	0.519
		relu	-	constant	0.001	adam	-	0.727	0.718	0.47
		relu	1000	adaptive	0.001	adam	100,	0.749	0.704	0.547
		relu	500	constant	0.001	lbfgs	100,	0.741	0.742	0.463
		tanh	1000	adaptive	0.001	adam	100,	0.716	0.704	0.506
		tanh	1000	adaptive	0.01	adam	100,	0.709	0.679	0.464
		tanh	1000	constant	0.01	adam	100,	0.698	0.694	0.435
		tanh	1000	adaptive	0.001	sgd	50,	0.624	0.605	0.341

Table 16: Final Summary of Models

Final Model: Light Gradient Boosting

After comparing the results between logistics regression, boosted trees, random forest, and a neural network, we determined that Light Gradient Boosting performed the best. Random Forest outperformed other models for both testing and out of time validation datasets with 77.2% and 56.4% respectively.

		learning_rate	n_estimators	max_depth	num_leaves	Train	Test	OOT
Boosted Tree		0.1	100	3	**20**	0.841	0.805	0.505
		0.1	50	4	**30**	0.854	0.804	0.4871
		0.1	50	5	**20**	0.891	0.788	0.479
		0.1	100	5	**20**	0.873	0.814	0.526
		0.001	**300**	**-1**	**31**	**0.842**	**0.772**	**0.564**
		0.001	4000	-1	25	0.912	0.491	0.491
		0.001	300	-1	30	0.905	0.525	0.525
		0.0001	100	10	30	0.717	0.717	0.469
		0.0001	200	-100	50	0.781	0.781	0.462

The chosen hyperparameters are:

- learning_rate: 0.001
- n_estimators: 300
- max_depth: -1
- num_leaves: 31

Runs	Training	Testing	OOT
0	0.742	0.742	0.464
1	0.829	0.814	0.559
2	0.725	0.717	0.430
3	0.724	0.710	0.464
4	0.736	0.707	0.453
5	0.743	0.724	0.559
6	0.730	0.694	0.441
7	0.750	0.705	0.547
8	0.731	0.696	0.492
9	0.728	0.684	0.520

Table 17: Run Summary of Light Gradient Boosting

Final Model Result are Light Gradient Boosting with OOT result of 55.9% with FDR @ 3%

Best Model Results:

The following results are the in-depth analysis of the final Gradient Boosting model for training, testing, and out-of-time datasets:

Training	#Records	#Goods	#Bads
	59010	58367	643

Bin Statistics						Cumulative Statistics						
Bin %	# Records	# Goods	# Bads	% Good	% Bad	Total Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	590	182	408	30.847	69.153	590	182	408	0.312	63.453	63.141	0.446
2	590	482	108	81.695	18.305	1180	664	516	1.138	80.249	79.111	1.287
3	590	573	17	97.119	2.881	1770	1237	533	2.119	82.893	80.773	2.321
4	590	583	7	98.814	1.186	2360	1820	540	3.118	83.981	80.863	3.370
5	590	584	6	98.983	1.017	2950	2404	546	4.119	84.914	80.796	4.403
6	591	588	3	99.492	0.508	3541	2992	549	5.126	85.381	80.255	5.450
7	590	585	5	99.153	0.847	4131	3577	554	6.128	86.159	80.030	6.457
8	590	588	2	99.661	0.339	4721	4165	556	7.136	86.470	79.334	7.491
9	590	589	1	99.831	0.169	5311	4754	557	8.145	86.625	78.480	8.535
10	590	587	3	99.492	0.508	5901	5341	560	9.151	87.092	77.941	9.538
11	590	587	3	99.492	0.508	6491	5928	563	10.156	87.558	77.402	10.529
12	590	582	8	98.644	1.356	7081	6510	571	11.154	88.802	77.649	11.401
13	590	588	2	99.661	0.339	7671	7098	573	12.161	89.114	76.953	12.387
14	590	588	2	99.661	0.339	8261	7686	575	13.168	89.425	76.256	13.367
15	591	588	3	99.492	0.508	8852	8274	578	14.176	89.891	75.715	14.315
16	590	588	2	99.661	0.339	9442	8862	580	15.183	90.202	75.019	15.279
17	590	587	3	99.492	0.508	10032	9449	583	16.189	90.669	74.480	16.208
18	590	588	2	99.661	0.339	10622	10037	585	17.196	90.980	73.783	17.157
19	590	587	3	99.492	0.508	11212	10624	588	18.202	91.446	73.244	18.068
20	590	587	3	99.492	0.508	11802	11211	591	19.208	91.913	72.705	18.970

Table 18: Best Model Light Gradient Boosting – Training results

Testing	#Records	#Goods	#Bads
	25290	25053	237

Bin Statistics						Cumulative Statistics						
Bin %	# Records	# Goods	# Bads	% Good	% Bad	Total Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	253	117	136	46.245	53.755	253	117	136	0.467	57.384	56.917	0.860
2	253	208	45	82.213	17.787	506	325	181	1.297	76.371	75.074	1.796
3	253	241	12	95.257	4.743	759	566	193	2.259	81.435	79.175	2.933
4	253	251	2	99.209	0.791	1012	817	195	3.261	82.278	79.017	4.190
5	252	249	3	98.810	1.190	1264	1066	198	4.255	83.544	79.289	5.384
6	253	252	1	99.605	0.395	1517	1318	199	5.261	83.966	78.705	6.623
7	253	253	0	100.000	0.000	1770	1571	199	6.271	83.966	77.696	7.894
8	253	251	2	99.209	0.791	2023	1822	201	7.273	84.810	77.538	9.065
9	253	251	2	99.209	0.791	2276	2073	203	8.274	85.654	77.380	10.212
10	253	251	2	99.209	0.791	2529	2324	205	9.276	86.498	77.222	11.337
11	253	251	2	99.209	0.791	2782	2575	207	10.278	87.342	77.064	12.440
12	253	248	5	98.024	1.976	3035	2823	212	11.268	89.451	78.183	13.316
13	253	253	0	100.000	0.000	3288	3076	212	12.278	89.451	77.174	14.509
14	253	253	0	100.000	0.000	3541	3329	212	13.288	89.451	76.164	15.703
15	253	252	1	99.605	0.395	3794	3581	213	14.294	89.873	75.580	16.812
16	252	252	0	100.000	0.000	4046	3833	213	15.300	89.873	74.574	17.995
17	253	249	4	98.419	1.581	4299	4082	217	16.293	91.561	75.268	18.811
18	253	251	2	99.209	0.791	4552	4333	219	17.295	92.405	75.110	19.785
19	253	253	0	100.000	0.000	4805	4586	219	18.305	92.405	74.100	20.941
20	253	253	0	100.000	0.000	5058	4839	219	19.315	92.405	73.090	22.096

Table 19: Best Model Light Gradient Boosting – Testing results

OOT	#Records	#Goods	#Bads
	2419	2294	125

Bin Statistics						Cumulative Statistics						
Bin %	# Records	# Goods	# Bads	% Good	% Bad	Total Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	0	0	0	0	0	0	0	0	0	0	0	0
1	121	73	48	60.331	39.669	121	73	48	0.61	26.816	26.203	1.521
2	121	83	38	68.595	31.405	242	156	86	1.31	48.045	46.736	1.814
3	121	107	14	88.430	11.570	363	263	100	2.21	55.866	53.659	2.630
4	121	118	3	97.521	2.479	484	381	103	3.20	57.542	54.345	3.699
5	121	118	3	97.521	2.479	605	499	106	4.19	59.218	55.031	4.708
6	121	119	2	98.347	1.653	726	618	108	5.19	60.335	55.150	5.722
7	121	120	1	99.174	0.826	847	738	109	6.19	60.894	54.702	6.771
8	121	121	0	100.000	0.000	968	859	109	7.21	60.894	53.686	7.881
9	121	120	1	99.174	0.826	1089	979	110	8.21	61.453	53.238	8.900
10	121	119	2	98.347	1.653	1210	1098	112	9.21	62.570	53.357	9.804
11	121	118	3	97.521	2.479	1331	1216	115	10.20	64.246	54.043	10.574
12	121	119	2	98.347	1.653	1452	1335	117	11.20	65.363	54.162	11.410
13	121	120	1	99.174	0.826	1573	1455	118	12.21	65.922	53.713	12.331
14	121	119	2	98.347	1.653	1694	1574	120	13.21	67.039	53.832	13.117
15	121	119	2	98.347	1.653	1815	1693	122	14.21	68.156	53.951	13.877
16	121	120	1	99.174	0.826	1936	1813	123	15.21	68.715	53.503	14.740
17	120	120	0	100.000	0.000	2056	1933	123	16.22	68.715	52.496	15.715
18	121	121	0	100.000	0.000	2177	2054	123	17.23	68.715	51.481	16.699
19	121	121	0	100.000	0.000	2298	2175	123	18.25	68.715	50.465	17.683
20	121	119	2	98.347	1.653	2419	2294	125	19.25	69.832	50.584	18.352

Table 20: Best Model Light Gradient Boosting – OOT results

Cutoff Plot:

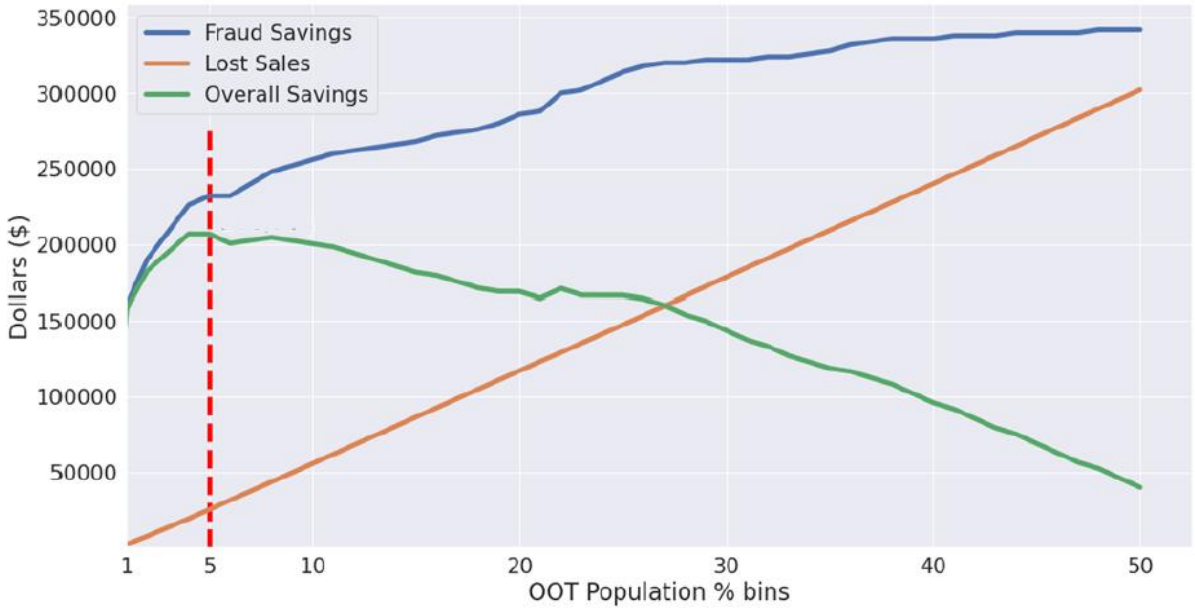


Fig 25. FDR Cutoff

To put our model into business practice, we generated the above plot to provide a recommendation for the FDR cutoff point. Any transaction with a score above the threshold at that cutoff point would be classified as fraudulent.

Assuming $2,000 gain for every fraud that is detected and $50 loss for every non-fraud that is flagged as a fraud, we plotted the Fraud Savings (blue), the Lost Sales (orange) and the Fraud Savings (green) for the out of time dataset. After analysis we recommend a cutoff point at 5% as this threshold maximizes the profits (P) calculated as follows:

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

Time Plot:

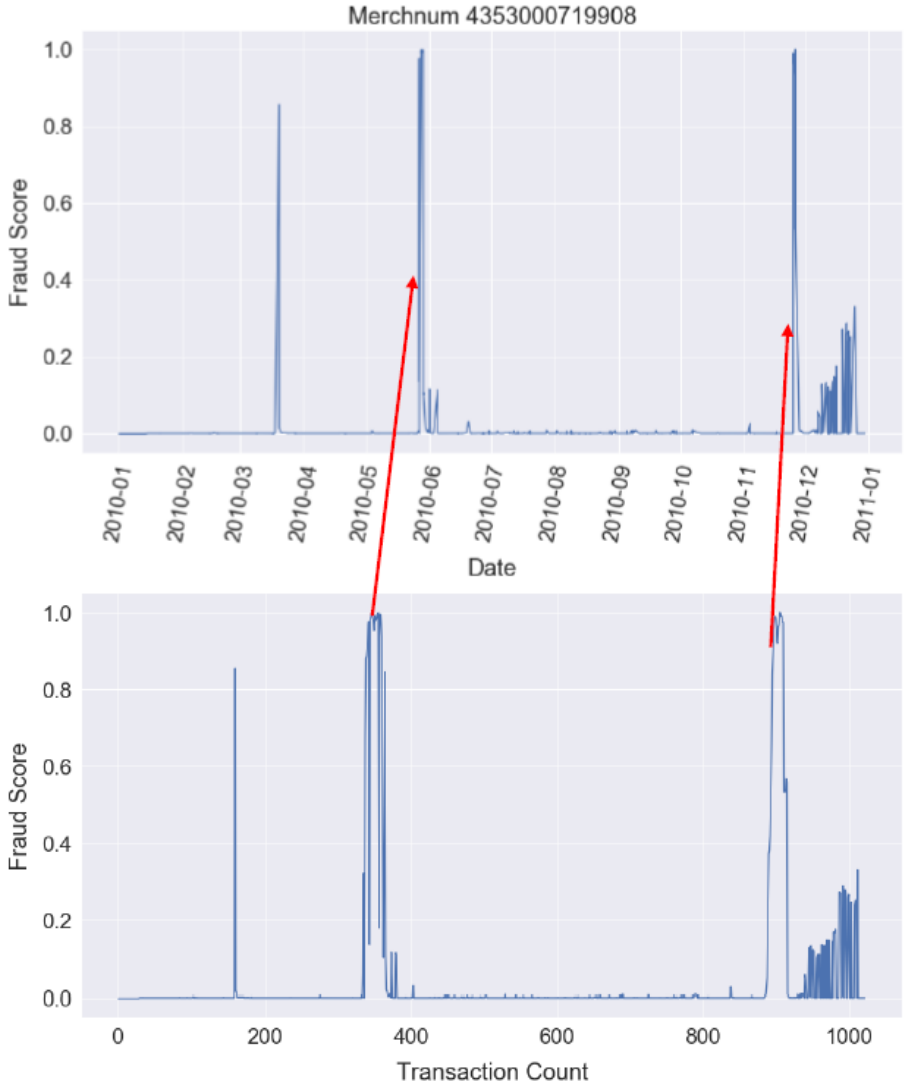


Fig 26. Transaction Counts Vs Fraud Score (MerchNum)

The above graph shows an example of potentially fraudulent activity from the "Merchnum" data field containing the value of “4353000719908”. There was a burst of activity with 34 transactions from 05/27/2010 to 05/29/2010 and another burst of activity with 32 transactions from 11/25/2010 to 11/26/2010. Through the graph we can see that the fraud score rose steeply in accordance with the time-period.

Conclusion:

A comprehensive analysis of credit card transaction fraud cases was performed. First, we performed exploratory data analysis to explore important fields and understand the distribution of the data. This was followed by data cleaning, outlier removals and missing value imputation. Then, over 1000 candidate variables were created and feature selection was performed (filter and wrapper methods) to pick the best variables. The variables were used across several models: logistic regression, boosted trees, random forest, and neural networks. Our best model to predict fraud was Gradient boosting which resulted in an 81.4% FDR at 3% for the testing dataset and a 56.4% FDR at 3% for the OOT dataset.

For future steps of this analysis, we would further explore different techniques to improve our fraud detection performance. In our model building, we found that compared to other algorithms, gradient boosting performed better in reducing variance in the out of time results. We would investigate why this is the case. Furthermore, in order to manage the imbalanced dataset, we applied weights to each class and down sampled the majority class. In the future, we would use bootstrapping (under sampling and over sampling). The other most popular which we can use to handle imbalance is called SMOTE (synthetic minority oversampling technique), which creates synthetic samples of the rare class rather than pure copies by selecting various instances. Moreover, we found it was quite challenging to work with a small dataset containing only a small fraction of fraudulent activity. Thus, we would attempt to collect more data for a more robust analysis. Finally, we hope to further consult experts to generate a more comprehensive collection of candidate variables.

Appendix A

Data Quality Report

1. Dataset Overview

Dataset Name- Card Transaction Data

Dataset Description – Dataset contains information on the actual credit card purchases from a US government organization. It provides information on Credit Card, Merchant, Date, and Amount involved in each transaction. Moreover, it also contains a column called fraud label which tells us whether the transaction is fraudulent or not.

Total Fields – 10

Total Records – 96,753

Time Period - 1st January 2006 – 31st December 2006

2. Summary

Numerical:

Field Name	% Populated	Min Val	Max Val	Mean	StdDev	% Zeros
Amount	**100%**	**0.01**	**3102,045.5**	**427.9**	**10,006.1**	**0**

Table 1: Summary Table for Numerical Variables

Categorical:

Field Name	% Populated	# Unique Values	Most Common Value
Recnum	**100**	**96,753**	**-**
Cardnum	**100**	**1,645**	**51421448452**
Date	**100**	**365**	**2006-02-28**
Merchnum	**96.5**	**13,092**	**930090121224**
Merch description	**100**	**13,126**	**GSA-FSS-ADV**
Merch state	**98.7**	**228**	**TN**
Merch zip	**95.2**	**4,568**	**38118**
Transtype	**100**	**4**	**P**
Fraud	**100**	**2**	**O**

Table 2: Summary Table for Numerical Variables

3. Field Exploration

Field 1 – Recnum
Description – A categorical field containing unique number for each record.

Field 2 – Cardnum
Description – A categorical field containing the card number used in each transaction

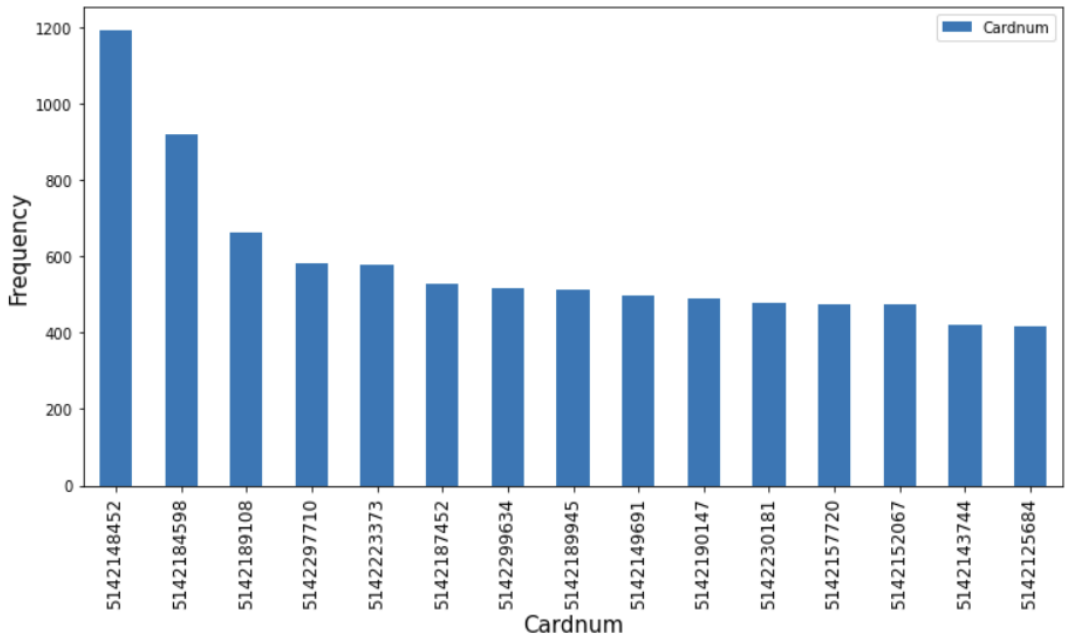


Fig 1. Distribution of field "Cardnum" for top 15 values

Field 3- Date
Description- Field containing the date of each transaction

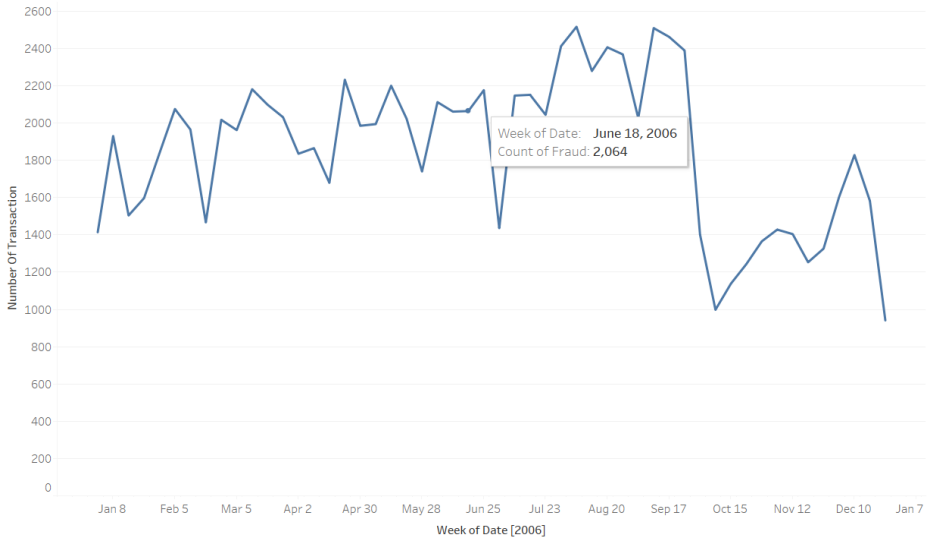


Fig 2. Weekly Distribution of Number of Transactions

Field 4 – Merchnum
Description – Field containing the merchant number of each transaction.

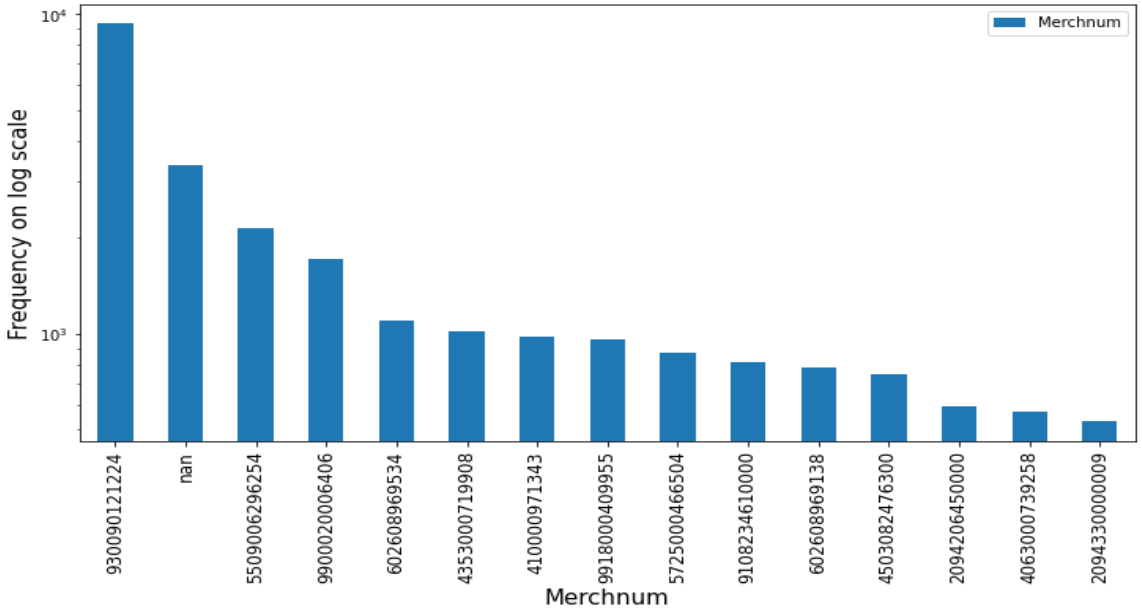


Fig 3. Bar Plot showing distribution of Field "Merchnum"

Field 5 – Merch description
Description – Field containing merchant description of each transaction.

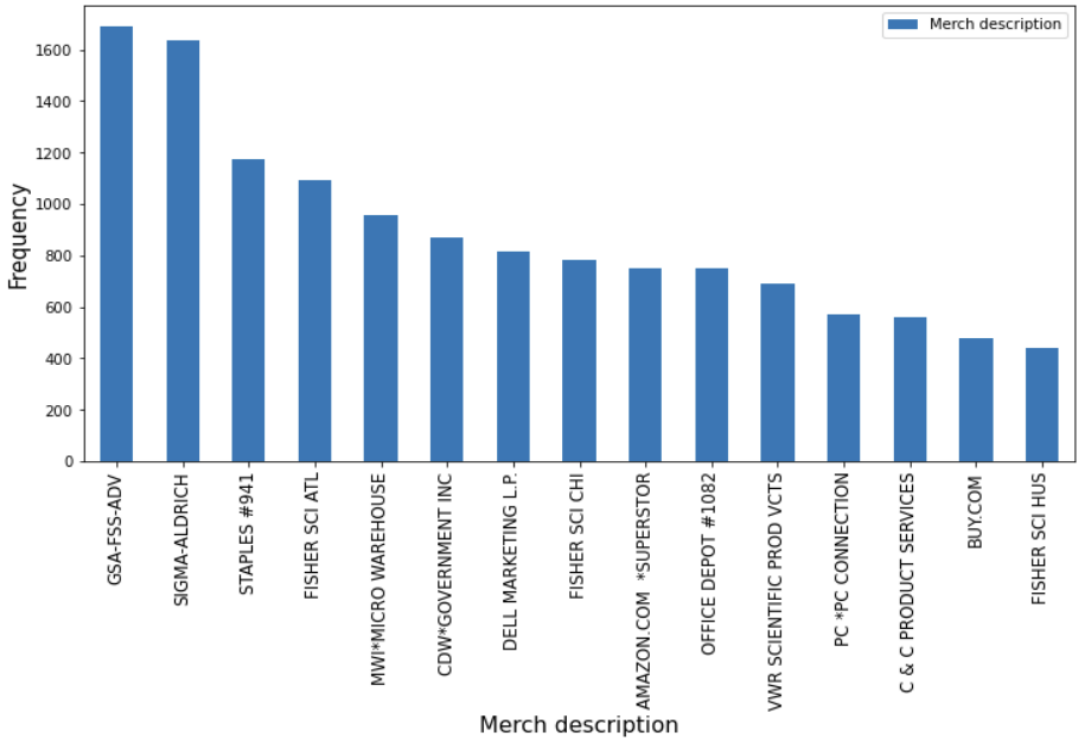


Fig 4. Bar Plot showing distribution of Field "Merch description"

Field 6 – Merch State

Description – A field which describes the state in which the merchant resides.

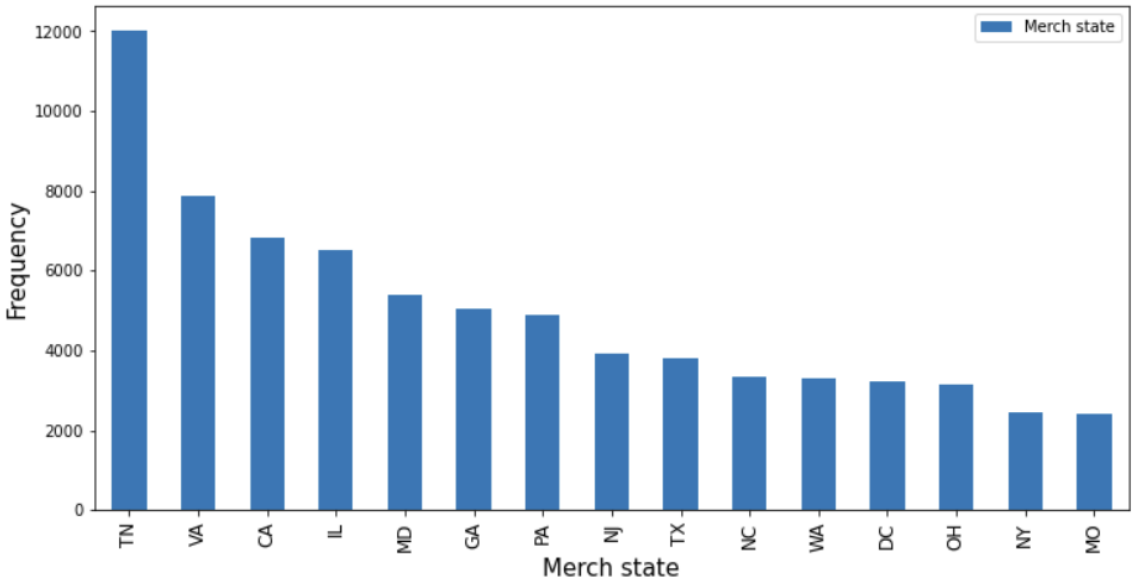


Fig 5. Bar Plot showing distribution of Field "Merch State"

Field 7 – Merch zip
Description – Field containing Zip code of merchant location.

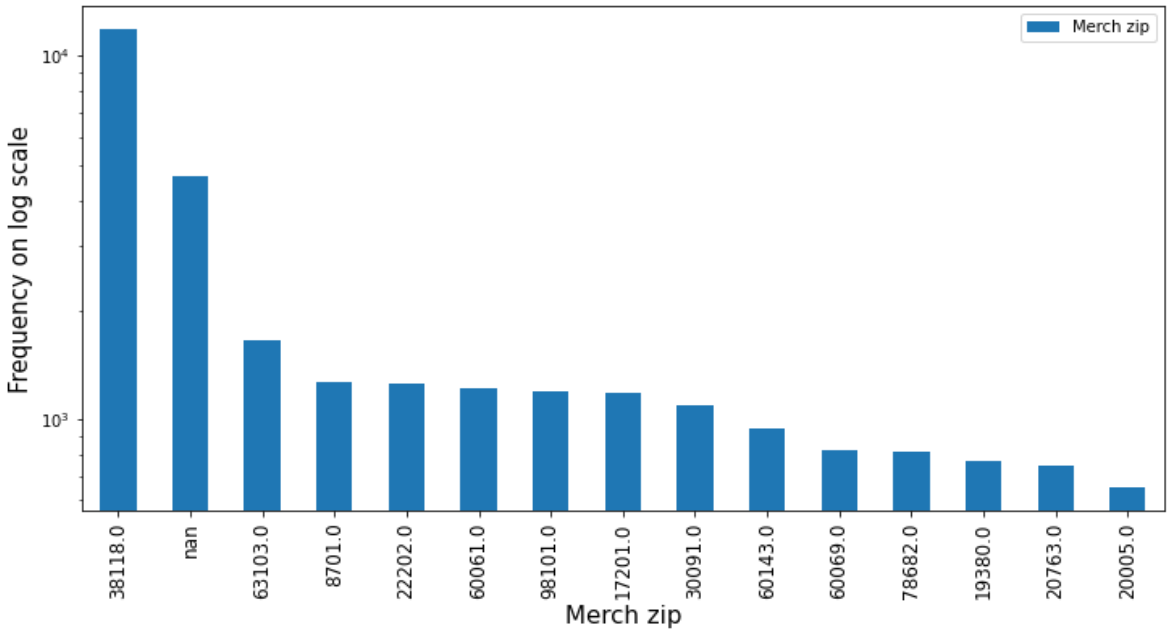


Fig 6. Bar Plot showing distribution of Field "Merch Zip"

Field 8 – Transtype
Description – Field containing types of transaction. It has 4 categories "P", "A", "D", "Y".

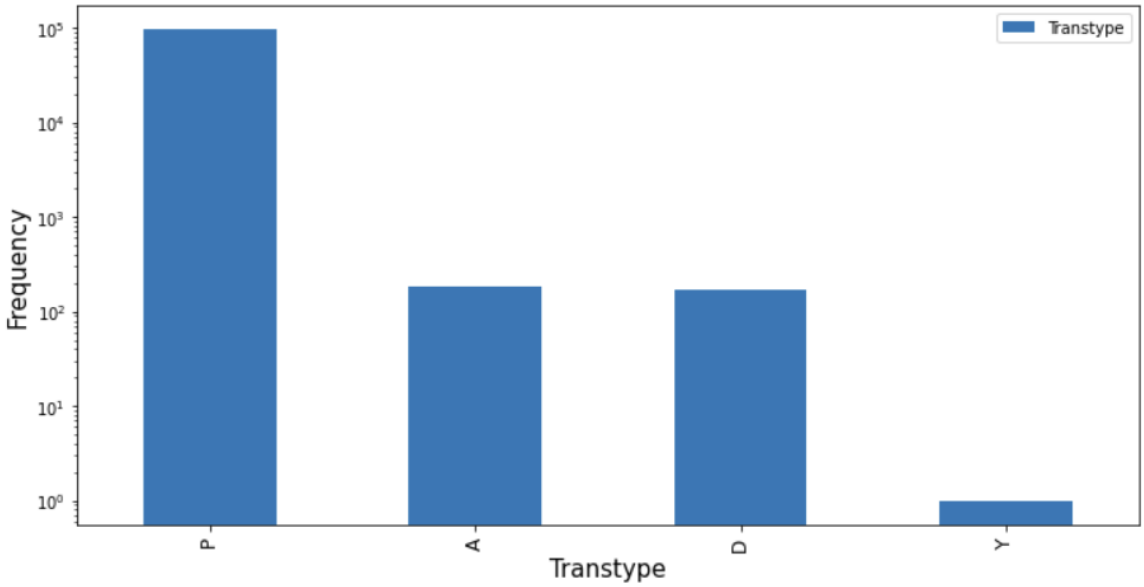


Fig 7. Bar Plot showing distribution of Field "Transtype"

Field 9 – Amount
Description – Numerical Field which containing the amount of each transaction

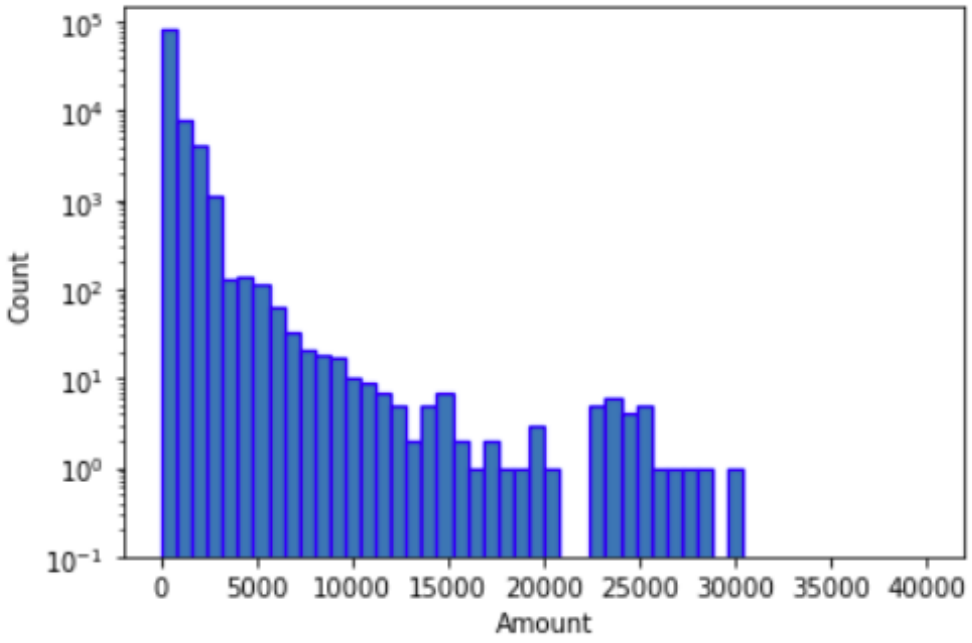


Fig 8. Histogram showing distribution of Field "Amount" on logarithmic scale

Field 10 – Fraud
Description – Field containing 2 categories. 0 indicates a good application and 1 indicates a bad application.

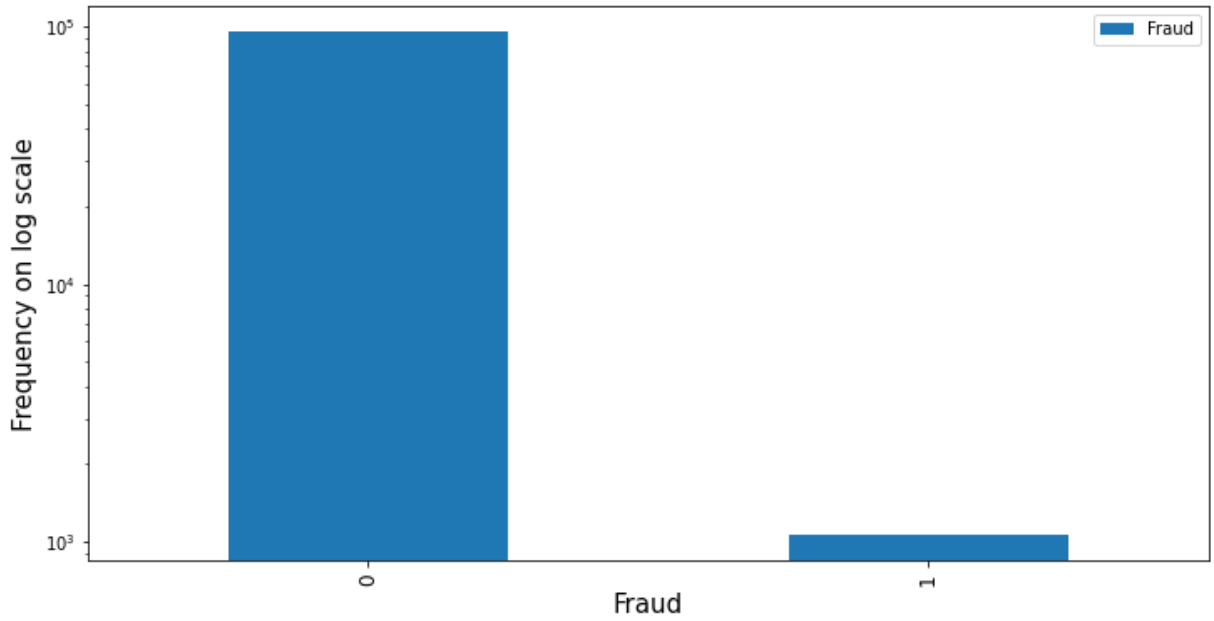


Fig 9. Bar Plot showing distribution of Fraud

Appendix B

Recnum	Merchnum_Merchzip_med_7
Cardnum_x	Merchnum_Merchzip_total_7
Date	Merchnum_Merchzip_actual/avg_7
Merchnum_x	Merchnum_Merchzip_actual/max_7
Merch description	Merchnum_Merchzip_actual/med_7
Merch state	Merchnum_Merchzip_actual/total_7
Merch zip	Merchnum_Merchzip_count_14
Transtype	Merchnum_Merchzip_avg_14
Amount	Merchnum_Merchzip_max_14
Fraud	Merchnum_Merchzip_med_14
DOW	Merchnum_Merchzip_total_14
DOW_Risk	Merchnum_Merchzip_actual/avg_14
MS_Risk	Merchnum_Merchzip_actual/max_14
Cardnum_Merchnum	Merchnum_Merchzip_actual/med_14
Cardnum_MerchState	Merchnum_Merchzip_actual/total_14
Cardnum_MerchZip	Merchnum_Merchzip_count_30
Cardnum_MerchDescription	Merchnum_Merchzip_avg_30
Merchnum_Merchzip	Merchnum_Merchzip_max_30
Merchnum_MerchState	Merchnum_Merchzip_med_30
Cardnum_Merchnum_MerchState	Merchnum_Merchzip_total_30
Cardnum_Merchnum_MerchZip	Merchnum_Merchzip_actual/avg_30
check_date	Merchnum_Merchzip_actual/max_30

check_record	Merchnum_Merchzip_actual/med_30
Cardnum_Merchnum_day_since	Merchnum_Merchzip_actual/total_30
Cardnum_Merchnum_count_0	Merchnum_Merchzip_count_60
Cardnum_Merchnum_avg_0	Merchnum_Merchzip_avg_60
Cardnum_Merchnum_max_0	Merchnum_Merchzip_max_60
Cardnum_Merchnum_med_0	Merchnum_Merchzip_med_60
Cardnum_Merchnum_total_0	Merchnum_Merchzip_total_60
Cardnum_Merchnum_actual/avg_0	Merchnum_Merchzip_actual/avg_60
Cardnum_Merchnum_actual/max_0	Merchnum_Merchzip_actual/max_60
Cardnum_Merchnum_actual/med_0	Merchnum_Merchzip_actual/med_60
Cardnum_Merchnum_actual/total_0	Merchnum_Merchzip_actual/total_60
Cardnum_Merchnum_count_1	Merchnum_Merchzip_count_90
Cardnum_Merchnum_avg_1	Merchnum_Merchzip_avg_90
Cardnum_Merchnum_max_1	Merchnum_Merchzip_max_90
Cardnum_Merchnum_med_1	Merchnum_Merchzip_med_90
Cardnum_Merchnum_total_1	Merchnum_Merchzip_total_90
Cardnum_Merchnum_actual/avg_1	Merchnum_Merchzip_actual/avg_90
Cardnum_Merchnum_actual/max_1	Merchnum_Merchzip_actual/max_90
Cardnum_Merchnum_actual/med_1	Merchnum_Merchzip_actual/med_90
Cardnum_Merchnum_actual/total_1	Merchnum_Merchzip_actual/total_90
Cardnum_Merchnum_count_3	Merchnum_MerchState_day_since
Cardnum_Merchnum_avg_3	Merchnum_MerchState_count_0
Cardnum_Merchnum_max_3	Merchnum_MerchState_avg_0
Cardnum_Merchnum_med_3	Merchnum_MerchState_max_0
Cardnum_Merchnum_total_3	Merchnum_MerchState_med_0
Cardnum_Merchnum_actual/avg_3	Merchnum_MerchState_total_0
Cardnum_Merchnum_actual/max_3	Merchnum_MerchState_actual/avg_0
Cardnum_Merchnum_actual/med_3	Merchnum_MerchState_actual/max_0
Cardnum_Merchnum_actual/total_3	Merchnum_MerchState_actual/med_0
Cardnum_Merchnum_count_7	Merchnum_MerchState_actual/total_0
Cardnum_Merchnum_avg_7	Merchnum_MerchState_count_1
Cardnum_Merchnum_max_7	Merchnum_MerchState_avg_1
Cardnum_Merchnum_med_7	Merchnum_MerchState_max_1
Cardnum_Merchnum_total_7	Merchnum_MerchState_med_1
Cardnum_Merchnum_actual/avg_7	Merchnum_MerchState_total_1
Cardnum_Merchnum_actual/max_7	Merchnum_MerchState_actual/avg_1
Cardnum_Merchnum_actual/med_7	Merchnum_MerchState_actual/max_1
Cardnum_Merchnum_actual/total_7	Merchnum_MerchState_actual/med_1
Cardnum_Merchnum_count_14	Merchnum_MerchState_actual/total_1
Cardnum_Merchnum_avg_14	Merchnum_MerchState_count_3
Cardnum_Merchnum_max_14	Merchnum_MerchState_avg_3
Cardnum_Merchnum_med_14	Merchnum_MerchState_max_3
Cardnum_Merchnum_total_14	Merchnum_MerchState_med_3

Cardnum_Merchnum_actual/avg_14	Merchnum_MerchState_total_3
Cardnum_Merchnum_actual/max_14	Merchnum_MerchState_actual/avg_3
Cardnum_Merchnum_actual/med_14	Merchnum_MerchState_actual/max_3
Cardnum_Merchnum_actual/total_14	Merchnum_MerchState_actual/med_3
Cardnum_Merchnum_count_30	Merchnum_MerchState_actual/total_3
Cardnum_Merchnum_avg_30	Merchnum_MerchState_count_7
Cardnum_Merchnum_max_30	Merchnum_MerchState_avg_7
Cardnum_Merchnum_med_30	Merchnum_MerchState_max_7
Cardnum_Merchnum_total_30	Merchnum_MerchState_med_7
Cardnum_Merchnum_actual/avg_30	Merchnum_MerchState_total_7
Cardnum_Merchnum_actual/max_30	Merchnum_MerchState_actual/avg_7
Cardnum_Merchnum_actual/med_30	Merchnum_MerchState_actual/max_7
Cardnum_Merchnum_actual/total_30	Merchnum_MerchState_actual/med_7
Cardnum_Merchnum_count_60	Merchnum_MerchState_actual/total_7
Cardnum_Merchnum_avg_60	Merchnum_MerchState_count_14
Cardnum_Merchnum_max_60	Merchnum_MerchState_avg_14
Cardnum_Merchnum_med_60	Merchnum_MerchState_max_14
Cardnum_Merchnum_total_60	Merchnum_MerchState_med_14
Cardnum_Merchnum_actual/avg_60	Merchnum_MerchState_total_14
Cardnum_Merchnum_actual/max_60	Merchnum_MerchState_actual/avg_14
Cardnum_Merchnum_actual/med_60	Merchnum_MerchState_actual/max_14
Cardnum_Merchnum_actual/total_60	Merchnum_MerchState_actual/med_14
Cardnum_Merchnum_count_90	Merchnum_MerchState_actual/total_14
Cardnum_Merchnum_avg_90	Merchnum_MerchState_count_30
Cardnum_Merchnum_max_90	Merchnum_MerchState_avg_30
Cardnum_Merchnum_med_90	Merchnum_MerchState_max_30
Cardnum_Merchnum_total_90	Merchnum_MerchState_med_30
Cardnum_Merchnum_actual/avg_90	Merchnum_MerchState_total_30
Cardnum_Merchnum_actual/max_90	Merchnum_MerchState_actual/avg_30
Cardnum_Merchnum_actual/med_90	Merchnum_MerchState_actual/max_30
Cardnum_Merchnum_actual/total_90	Merchnum_MerchState_actual/med_30
Cardnum_day_since	Merchnum_MerchState_actual/total_30
Cardnum_count_0	Merchnum_MerchState_count_60
Cardnum_avg_0	Merchnum_MerchState_avg_60
Cardnum_max_0	Merchnum_MerchState_max_60
Cardnum_med_0	Merchnum_MerchState_med_60
Cardnum_total_0	Merchnum_MerchState_total_60
Cardnum_actual/avg_0	Merchnum_MerchState_actual/avg_60
Cardnum_actual/max_0	Merchnum_MerchState_actual/max_60
Cardnum_actual/med_0	Merchnum_MerchState_actual/med_60
Cardnum_actual/total_0	Merchnum_MerchState_actual/total_60
Cardnum_count_1	Merchnum_MerchState_count_90
Cardnum_avg_1	Merchnum_MerchState_avg_90

Cardnum_max_1	Merchnum_MerchState_max_90
Cardnum_med_1	Merchnum_MerchState_med_90
Cardnum_total_1	Merchnum_MerchState_total_90
Cardnum_actual/avg_1	Merchnum_MerchState_actual/avg_90
Cardnum_actual/max_1	Merchnum_MerchState_actual/max_90
Cardnum_actual/med_1	Merchnum_MerchState_actual/med_90
Cardnum_actual/total_1	Merchnum_MerchState_actual/total_90
Cardnum_count_3	Merch state_day_since
Cardnum_avg_3	Merch state_count_0
Cardnum_max_3	Merch state_avg_0
Cardnum_med_3	Merch state_max_0
Cardnum_total_3	Merch state_med_0
Cardnum_actual/avg_3	Merch state_total_0
Cardnum_actual/max_3	Merch state_actual/avg_0
Cardnum_actual/med_3	Merch state_actual/max_0
Cardnum_actual/total_3	Merch state_actual/med_0
Cardnum_count_7	Merch state_actual/total_0
Cardnum_avg_7	Merch state_count_1
Cardnum_max_7	Merch state_avg_1
Cardnum_med_7	Merch state_max_1
Cardnum_total_7	Merch state_med_1
Cardnum_actual/avg_7	Merch state_total_1
Cardnum_actual/max_7	Merch state_actual/avg_1
Cardnum_actual/med_7	Merch state_actual/max_1
Cardnum_actual/total_7	Merch state_actual/med_1
Cardnum_count_14	Merch state_actual/total_1
Cardnum_avg_14	Merch state_count_3
Cardnum_max_14	Merch state_avg_3
Cardnum_med_14	Merch state_max_3
Cardnum_total_14	Merch state_med_3
Cardnum_actual/avg_14	Merch state_total_3
Cardnum_actual/max_14	Merch state_actual/avg_3
Cardnum_actual/med_14	Merch state_actual/max_3
Cardnum_actual/total_14	Merch state_actual/med_3
Cardnum_count_30	Merch state_actual/total_3
Cardnum_avg_30	Merch state_count_7
Cardnum_max_30	Merch state_avg_7
Cardnum_med_30	Merch state_max_7
Cardnum_total_30	Merch state_med_7
Cardnum_actual/avg_30	Merch state_total_7
Cardnum_actual/max_30	Merch state_actual/avg_7
Cardnum_actual/med_30	Merch state_actual/max_7
Cardnum_actual/total_30	Merch state_actual/med_7

Cardnum_count_60	Merch state_actual/total_7
Cardnum_avg_60	Merch state_count_14
Cardnum_max_60	Merch state_avg_14
Cardnum_med_60	Merch state_max_14
Cardnum_total_60	Merch state_med_14
Cardnum_actual/avg_60	Merch state_total_14
Cardnum_actual/max_60	Merch state_actual/avg_14
Cardnum_actual/med_60	Merch state_actual/max_14
Cardnum_actual/total_60	Merch state_actual/med_14
Cardnum_count_90	Merch state_actual/total_14
Cardnum_avg_90	Merch state_count_30
Cardnum_max_90	Merch state_avg_30
Cardnum_med_90	Merch state_max_30
Cardnum_total_90	Merch state_med_30
Cardnum_actual/avg_90	Merch state_total_30
Cardnum_actual/max_90	Merch state_actual/avg_30
Cardnum_actual/med_90	Merch state_actual/max_30
Cardnum_actual/total_90	Merch state_actual/med_30
Merchnum_day_since	Merch state_actual/total_30
Merchnum_count_0	Merch state_count_60
Merchnum_avg_0	Merch state_avg_60
Merchnum_max_0	Merch state_max_60
Merchnum_med_0	Merch state_med_60
Merchnum_total_0	Merch state_total_60
Merchnum_actual/avg_0	Merch state_actual/avg_60
Merchnum_actual/max_0	Merch state_actual/max_60
Merchnum_actual/med_0	Merch state_actual/med_60
Merchnum_actual/total_0	Merch state_actual/total_60
Merchnum_count_1	Merch state_count_90
Merchnum_avg_1	Merch state_avg_90
Merchnum_max_1	Merch state_max_90
Merchnum_med_1	Merch state_med_90
Merchnum_total_1	Merch state_total_90
Merchnum_actual/avg_1	Merch state_actual/avg_90
Merchnum_actual/max_1	Merch state_actual/max_90
Merchnum_actual/med_1	Merch state_actual/med_90
Merchnum_actual/total_1	Merch state_actual/total_90
Merchnum_count_3	Merch zip_day_since
Merchnum_avg_3	Merch zip_count_0
Merchnum_max_3	Merch zip_avg_0
Merchnum_med_3	Merch zip_max_0
Merchnum_total_3	Merch zip_med_0
Merchnum_actual/avg_3	Merch zip_total_0

Merchnum_actual/max_3	Merch zip_actual/avg_0
Merchnum_actual/med_3	Merch zip_actual/max_0
Merchnum_actual/total_3	Merch zip_actual/med_0
Merchnum_count_7	Merch zip_actual/total_0
Merchnum_avg_7	Merch zip_count_1
Merchnum_max_7	Merch zip_avg_1
Merchnum_med_7	Merch zip_max_1
Merchnum_total_7	Merch zip_med_1
Merchnum_actual/avg_7	Merch zip_total_1
Merchnum_actual/max_7	Merch zip_actual/avg_1
Merchnum_actual/med_7	Merch zip_actual/max_1
Merchnum_actual/total_7	Merch zip_actual/med_1
Merchnum_count_14	Merch zip_actual/total_1
Merchnum_avg_14	Merch zip_count_3
Merchnum_max_14	Merch zip_avg_3
Merchnum_med_14	Merch zip_max_3
Merchnum_total_14	Merch zip_med_3
Merchnum_actual/avg_14	Merch zip_total_3
Merchnum_actual/max_14	Merch zip_actual/avg_3
Merchnum_actual/med_14	Merch zip_actual/max_3
Merchnum_actual/total_14	Merch zip_actual/med_3
Merchnum_count_30	Merch zip_actual/total_3
Merchnum_avg_30	Merch zip_count_7
Merchnum_max_30	Merch zip_avg_7
Merchnum_med_30	Merch zip_max_7
Merchnum_total_30	Merch zip_med_7
Merchnum_actual/avg_30	Merch zip_total_7
Merchnum_actual/max_30	Merch zip_actual/avg_7
Merchnum_actual/med_30	Merch zip_actual/max_7
Merchnum_actual/total_30	Merch zip_actual/med_7
Merchnum_count_60	Merch zip_actual/total_7
Merchnum_avg_60	Merch zip_count_14
Merchnum_max_60	Merch zip_avg_14
Merchnum_med_60	Merch zip_max_14
Merchnum_total_60	Merch zip_med_14
Merchnum_actual/avg_60	Merch zip_total_14
Merchnum_actual/max_60	Merch zip_actual/avg_14
Merchnum_actual/med_60	Merch zip_actual/max_14
Merchnum_actual/total_60	Merch zip_actual/med_14
Merchnum_count_90	Merch zip_actual/total_14
Merchnum_avg_90	Merch zip_count_30
Merchnum_max_90	Merch zip_avg_30
Merchnum_med_90	Merch zip_max_30

Merchnum_total_90	Merch zip_med_30
Merchnum_actual/avg_90	Merch zip_total_30
Merchnum_actual/max_90	Merch zip_actual/avg_30
Merchnum_actual/med_90	Merch zip_actual/max_30
Merchnum_actual/total_90	Merch zip_actual/med_30
Merch description_day_since	Merch zip_actual/total_30
Merch description_count_0	Merch zip_count_60
Merch description_avg_0	Merch zip_avg_60
Merch description_max_0	Merch zip_max_60
Merch description_med_0	Merch zip_med_60
Merch description_total_0	Merch zip_total_60
Merch description_actual/avg_0	Merch zip_actual/avg_60
Merch description_actual/max_0	Merch zip_actual/max_60
Merch description_actual/med_0	Merch zip_actual/med_60
Merch description_actual/total_0	Merch zip_actual/total_60
Merch description_count_1	Merch zip_count_90
Merch description_avg_1	Merch zip_avg_90
Merch description_max_1	Merch zip_max_90
Merch description_med_1	Merch zip_med_90
Merch description_total_1	Merch zip_total_90
Merch description_actual/avg_1	Merch zip_actual/avg_90
Merch description_actual/max_1	Merch zip_actual/max_90
Merch description_actual/med_1	Merch zip_actual/med_90
Merch description_actual/total_1	Merch zip_actual/total_90
Merch description_count_3	Cardnum_Merchnum_MerchState_day_since
Merch description_avg_3	Cardnum_Merchnum_MerchState_count_0
Merch description_max_3	Cardnum_Merchnum_MerchState_avg_0
Merch description_med_3	Cardnum_Merchnum_MerchState_max_0
Merch description_total_3	Cardnum_Merchnum_MerchState_med_0
Merch description_actual/avg_3	Cardnum_Merchnum_MerchState_total_0
Merch description_actual/max_3	Cardnum_Merchnum_MerchState_actual/avg_0
Merch description_actual/med_3	Cardnum_Merchnum_MerchState_actual/max_0
Merch description_actual/total_3	Cardnum_Merchnum_MerchState_actual/med_0
Merch description_count_7	Cardnum_Merchnum_MerchState_actual/total_0
Merch description_avg_7	Cardnum_Merchnum_MerchState_count_1
Merch description_max_7	Cardnum_Merchnum_MerchState_avg_1
Merch description_med_7	Cardnum_Merchnum_MerchState_max_1
Merch description_total_7	Cardnum_Merchnum_MerchState_med_1
Merch description_actual/avg_7	Cardnum_Merchnum_MerchState_total_1
Merch description_actual/max_7	Cardnum_Merchnum_MerchState_actual/avg_1
Merch description_actual/med_7	Cardnum_Merchnum_MerchState_actual/max_1
Merch description_actual/total_7	Cardnum_Merchnum_MerchState_actual/med_1
Merch description_count_14	Cardnum_Merchnum_MerchState_actual/total_1

Merch description_avg_14	Cardnum_Merchnum_MerchState_count_3
Merch description_max_14	Cardnum_Merchnum_MerchState_avg_3
Merch description_med_14	Cardnum_Merchnum_MerchState_max_3
Merch description_total_14	Cardnum_Merchnum_MerchState_med_3
Merch description_actual/avg_14	Cardnum_Merchnum_MerchState_total_3
Merch description_actual/max_14	Cardnum_Merchnum_MerchState_actual/avg_3
Merch description_actual/med_14	Cardnum_Merchnum_MerchState_actual/max_3
Merch description_actual/total_14	Cardnum_Merchnum_MerchState_actual/med_3
Merch description_count_30	Cardnum_Merchnum_MerchState_actual/total_3
Merch description_avg_30	Cardnum_Merchnum_MerchState_count_7
Merch description_max_30	Cardnum_Merchnum_MerchState_avg_7
Merch description_med_30	Cardnum_Merchnum_MerchState_max_7
Merch description_total_30	Cardnum_Merchnum_MerchState_med_7
Merch description_actual/avg_30	Cardnum_Merchnum_MerchState_total_7
Merch description_actual/max_30	Cardnum_Merchnum_MerchState_actual/avg_7
Merch description_actual/med_30	Cardnum_Merchnum_MerchState_actual/max_7
Merch description_actual/total_30	Cardnum_Merchnum_MerchState_actual/med_7
Merch description_count_60	Cardnum_Merchnum_MerchState_actual/total_7
Merch description_avg_60	Cardnum_Merchnum_MerchState_count_14
Merch description_max_60	Cardnum_Merchnum_MerchState_avg_14
Merch description_med_60	Cardnum_Merchnum_MerchState_max_14
Merch description_total_60	Cardnum_Merchnum_MerchState_med_14
Merch description_actual/avg_60	Cardnum_Merchnum_MerchState_total_14
Merch description_actual/max_60	Cardnum_Merchnum_MerchState_actual/avg_14
Merch description_actual/med_60	Cardnum_Merchnum_MerchState_actual/max_14
Merch description_actual/total_60	Cardnum_Merchnum_MerchState_actual/med_14
Merch description_count_90	Cardnum_Merchnum_MerchState_actual/total_14
Merch description_avg_90	Cardnum_Merchnum_MerchState_count_30
Merch description_max_90	Cardnum_Merchnum_MerchState_avg_30
Merch description_med_90	Cardnum_Merchnum_MerchState_max_30
Merch description_total_90	Cardnum_Merchnum_MerchState_med_30
Merch description_actual/avg_90	Cardnum_Merchnum_MerchState_total_30
Merch description_actual/max_90	Cardnum_Merchnum_MerchState_actual/avg_30
Merch description_actual/med_90	Cardnum_Merchnum_MerchState_actual/max_30
Merch description_actual/total_90	Cardnum_Merchnum_MerchState_actual/med_30
Cardnum_MerchState_day_since	Cardnum_Merchnum_MerchState_actual/total_30
Cardnum_MerchState_count_0	Cardnum_Merchnum_MerchState_count_60
Cardnum_MerchState_avg_0	Cardnum_Merchnum_MerchState_avg_60
Cardnum_MerchState_max_0	Cardnum_Merchnum_MerchState_max_60
Cardnum_MerchState_med_0	Cardnum_Merchnum_MerchState_med_60
Cardnum_MerchState_total_0	Cardnum_Merchnum_MerchState_total_60
Cardnum_MerchState_actual/avg_0	Cardnum_Merchnum_MerchState_actual/avg_60
Cardnum_MerchState_actual/max_0	Cardnum_Merchnum_MerchState_actual/max_60

Cardnum_MerchState_actual/med_0	Cardnum_Merchnum_MerchState_actual/med_60
Cardnum_MerchState_actual/total_0	Cardnum_Merchnum_MerchState_actual/total_60
Cardnum_MerchState_count_1	Cardnum_Merchnum_MerchState_count_90
Cardnum_MerchState_avg_1	Cardnum_Merchnum_MerchState_avg_90
Cardnum_MerchState_max_1	Cardnum_Merchnum_MerchState_max_90
Cardnum_MerchState_med_1	Cardnum_Merchnum_MerchState_med_90
Cardnum_MerchState_total_1	Cardnum_Merchnum_MerchState_total_90
Cardnum_MerchState_actual/avg_1	Cardnum_Merchnum_MerchState_actual/avg_90
Cardnum_MerchState_actual/max_1	Cardnum_Merchnum_MerchState_actual/max_90
Cardnum_MerchState_actual/med_1	Cardnum_Merchnum_MerchState_actual/med_90
Cardnum_MerchState_actual/total_1	Cardnum_Merchnum_MerchState_actual/total_90
Cardnum_MerchState_count_3	Cardnum_Merchnum_MerchZip_day_since
Cardnum_MerchState_avg_3	Cardnum_Merchnum_MerchZip_count_0
Cardnum_MerchState_max_3	Cardnum_Merchnum_MerchZip_avg_0
Cardnum_MerchState_med_3	Cardnum_Merchnum_MerchZip_max_0
Cardnum_MerchState_total_3	Cardnum_Merchnum_MerchZip_med_0
Cardnum_MerchState_actual/avg_3	Cardnum_Merchnum_MerchZip_total_0
Cardnum_MerchState_actual/max_3	Cardnum_Merchnum_MerchZip_actual/avg_0
Cardnum_MerchState_actual/med_3	Cardnum_Merchnum_MerchZip_actual/max_0
Cardnum_MerchState_actual/total_3	Cardnum_Merchnum_MerchZip_actual/med_0
Cardnum_MerchState_count_7	Cardnum_Merchnum_MerchZip_actual/total_0
Cardnum_MerchState_avg_7	Cardnum_Merchnum_MerchZip_count_1
Cardnum_MerchState_max_7	Cardnum_Merchnum_MerchZip_avg_1
Cardnum_MerchState_med_7	Cardnum_Merchnum_MerchZip_max_1
Cardnum_MerchState_total_7	Cardnum_Merchnum_MerchZip_med_1
Cardnum_MerchState_actual/avg_7	Cardnum_Merchnum_MerchZip_total_1
Cardnum_MerchState_actual/max_7	Cardnum_Merchnum_MerchZip_actual/avg_1
Cardnum_MerchState_actual/med_7	Cardnum_Merchnum_MerchZip_actual/max_1
Cardnum_MerchState_actual/total_7	Cardnum_Merchnum_MerchZip_actual/med_1
Cardnum_MerchState_count_14	Cardnum_Merchnum_MerchZip_actual/total_1
Cardnum_MerchState_avg_14	Cardnum_Merchnum_MerchZip_count_3
Cardnum_MerchState_max_14	Cardnum_Merchnum_MerchZip_avg_3
Cardnum_MerchState_med_14	Cardnum_Merchnum_MerchZip_max_3
Cardnum_MerchState_total_14	Cardnum_Merchnum_MerchZip_med_3
Cardnum_MerchState_actual/avg_14	Cardnum_Merchnum_MerchZip_total_3
Cardnum_MerchState_actual/max_14	Cardnum_Merchnum_MerchZip_actual/avg_3
Cardnum_MerchState_actual/med_14	Cardnum_Merchnum_MerchZip_actual/max_3
Cardnum_MerchState_actual/total_14	Cardnum_Merchnum_MerchZip_actual/med_3
Cardnum_MerchState_count_30	Cardnum_Merchnum_MerchZip_actual/total_3
Cardnum_MerchState_avg_30	Cardnum_Merchnum_MerchZip_count_7
Cardnum_MerchState_max_30	Cardnum_Merchnum_MerchZip_avg_7
Cardnum_MerchState_med_30	Cardnum_Merchnum_MerchZip_max_7
Cardnum_MerchState_total_30	Cardnum_Merchnum_MerchZip_med_7

Cardnum_MerchState_actual/avg_30	Cardnum_Merchnum_MerchZip_total_7
Cardnum_MerchState_actual/max_30	Cardnum_Merchnum_MerchZip_actual/avg_7
Cardnum_MerchState_actual/med_30	Cardnum_Merchnum_MerchZip_actual/max_7
Cardnum_MerchState_actual/total_30	Cardnum_Merchnum_MerchZip_actual/med_7
Cardnum_MerchState_count_60	Cardnum_Merchnum_MerchZip_actual/total_7
Cardnum_MerchState_avg_60	Cardnum_Merchnum_MerchZip_count_14
Cardnum_MerchState_max_60	Cardnum_Merchnum_MerchZip_avg_14
Cardnum_MerchState_med_60	Cardnum_Merchnum_MerchZip_max_14
Cardnum_MerchState_total_60	Cardnum_Merchnum_MerchZip_med_14
Cardnum_MerchState_actual/avg_60	Cardnum_Merchnum_MerchZip_total_14
Cardnum_MerchState_actual/max_60	Cardnum_Merchnum_MerchZip_actual/avg_14
Cardnum_MerchState_actual/med_60	Cardnum_Merchnum_MerchZip_actual/max_14
Cardnum_MerchState_actual/total_60	Cardnum_Merchnum_MerchZip_actual/med_14
Cardnum_MerchState_count_90	Cardnum_Merchnum_MerchZip_actual/total_14
Cardnum_MerchState_avg_90	Cardnum_Merchnum_MerchZip_count_30
Cardnum_MerchState_max_90	Cardnum_Merchnum_MerchZip_avg_30
Cardnum_MerchState_med_90	Cardnum_Merchnum_MerchZip_max_30
Cardnum_MerchState_total_90	Cardnum_Merchnum_MerchZip_med_30
Cardnum_MerchState_actual/avg_90	Cardnum_Merchnum_MerchZip_total_30
Cardnum_MerchState_actual/max_90	Cardnum_Merchnum_MerchZip_actual/avg_30
Cardnum_MerchState_actual/med_90	Cardnum_Merchnum_MerchZip_actual/max_30
Cardnum_MerchState_actual/total_90	Cardnum_Merchnum_MerchZip_actual/med_30
Cardnum_MerchZip_day_since	Cardnum_Merchnum_MerchZip_actual/total_30
Cardnum_MerchZip_count_0	Cardnum_Merchnum_MerchZip_count_60
Cardnum_MerchZip_avg_0	Cardnum_Merchnum_MerchZip_avg_60
Cardnum_MerchZip_max_0	Cardnum_Merchnum_MerchZip_max_60
Cardnum_MerchZip_med_0	Cardnum_Merchnum_MerchZip_med_60
Cardnum_MerchZip_total_0	Cardnum_Merchnum_MerchZip_total_60
Cardnum_MerchZip_actual/avg_0	Cardnum_Merchnum_MerchZip_actual/avg_60
Cardnum_MerchZip_actual/max_0	Cardnum_Merchnum_MerchZip_actual/max_60
Cardnum_MerchZip_actual/med_0	Cardnum_Merchnum_MerchZip_actual/med_60
Cardnum_MerchZip_actual/total_0	Cardnum_Merchnum_MerchZip_actual/total_60
Cardnum_MerchZip_count_1	Cardnum_Merchnum_MerchZip_count_90
Cardnum_MerchZip_avg_1	Cardnum_Merchnum_MerchZip_avg_90
Cardnum_MerchZip_max_1	Cardnum_Merchnum_MerchZip_max_90
Cardnum_MerchZip_med_1	Cardnum_Merchnum_MerchZip_med_90
Cardnum_MerchZip_total_1	Cardnum_Merchnum_MerchZip_total_90
Cardnum_MerchZip_actual/avg_1	Cardnum_Merchnum_MerchZip_actual/avg_90
Cardnum_MerchZip_actual/max_1	Cardnum_Merchnum_MerchZip_actual/max_90
Cardnum_MerchZip_actual/med_1	Cardnum_Merchnum_MerchZip_actual/med_90
Cardnum_MerchZip_actual/total_1	Cardnum_Merchnum_MerchZip_actual/total_90
Cardnum_MerchZip_count_3	Cardnum_Merchnum_count_0_by_3
Cardnum_MerchZip_avg_3	Cardnum_Merchnum_count_0_by_7

Cardnum_MerchZip_max_3	Cardnum_Merchnum_count_0_by_14
Cardnum_MerchZip_med_3	Cardnum_Merchnum_count_0_by_30
Cardnum_MerchZip_total_3	Cardnum_Merchnum_count_0_by_60
Cardnum_MerchZip_actual/avg_3	Cardnum_Merchnum_count_0_by_90
Cardnum_MerchZip_actual/max_3	Cardnum_Merchnum_count_1_by_3
Cardnum_MerchZip_actual/med_3	Cardnum_Merchnum_count_1_by_7
Cardnum_MerchZip_actual/total_3	Cardnum_Merchnum_count_1_by_14
Cardnum_MerchZip_count_7	Cardnum_Merchnum_count_1_by_30
Cardnum_MerchZip_avg_7	Cardnum_Merchnum_count_1_by_60
Cardnum_MerchZip_max_7	Cardnum_Merchnum_count_1_by_90
Cardnum_MerchZip_med_7	Cardnum_count_0_by_3
Cardnum_MerchZip_total_7	Cardnum_count_0_by_7
Cardnum_MerchZip_actual/avg_7	Cardnum_count_0_by_14
Cardnum_MerchZip_actual/max_7	Cardnum_count_0_by_30
Cardnum_MerchZip_actual/med_7	Cardnum_count_0_by_60
Cardnum_MerchZip_actual/total_7	Cardnum_count_0_by_90
Cardnum_MerchZip_count_14	Cardnum_count_1_by_3
Cardnum_MerchZip_avg_14	Cardnum_count_1_by_7
Cardnum_MerchZip_max_14	Cardnum_count_1_by_14
Cardnum_MerchZip_med_14	Cardnum_count_1_by_30
Cardnum_MerchZip_total_14	Cardnum_count_1_by_60
Cardnum_MerchZip_actual/avg_14	Cardnum_count_1_by_90
Cardnum_MerchZip_actual/max_14	Merchnum_count_0_by_3
Cardnum_MerchZip_actual/med_14	Merchnum_count_0_by_7
Cardnum_MerchZip_actual/total_14	Merchnum_count_0_by_14
Cardnum_MerchZip_count_30	Merchnum_count_0_by_30
Cardnum_MerchZip_avg_30	Merchnum_count_0_by_60
Cardnum_MerchZip_max_30	Merchnum_count_0_by_90
Cardnum_MerchZip_med_30	Merchnum_count_1_by_3
Cardnum_MerchZip_total_30	Merchnum_count_1_by_7
Cardnum_MerchZip_actual/avg_30	Merchnum_count_1_by_14
Cardnum_MerchZip_actual/max_30	Merchnum_count_1_by_30
Cardnum_MerchZip_actual/med_30	Merchnum_count_1_by_60
Cardnum_MerchZip_actual/total_30	Merchnum_count_1_by_90
Cardnum_MerchZip_count_60	Merch description_count_0_by_3
Cardnum_MerchZip_avg_60	Merch description_count_0_by_7
Cardnum_MerchZip_max_60	Merch description_count_0_by_14
Cardnum_MerchZip_med_60	Merch description_count_0_by_30
Cardnum_MerchZip_total_60	Merch description_count_0_by_60
Cardnum_MerchZip_actual/avg_60	Merch description_count_0_by_90
Cardnum_MerchZip_actual/max_60	Merch description_count_1_by_3
Cardnum_MerchZip_actual/med_60	Merch description_count_1_by_7
Cardnum_MerchZip_actual/total_60	Merch description_count_1_by_14

Cardnum_MerchZip_count_90	Merch description_count_1_by_30
Cardnum_MerchZip_avg_90	Merch description_count_1_by_60
Cardnum_MerchZip_max_90	Merch description_count_1_by_90
Cardnum_MerchZip_med_90	Cardnum_MerchState_count_0_by_3
Cardnum_MerchZip_total_90	Cardnum_MerchState_count_0_by_7
Cardnum_MerchZip_actual/avg_90	Cardnum_MerchState_count_0_by_14
Cardnum_MerchZip_actual/max_90	Cardnum_MerchState_count_0_by_30
Cardnum_MerchZip_actual/med_90	Cardnum_MerchState_count_0_by_60
Cardnum_MerchZip_actual/total_90	Cardnum_MerchState_count_0_by_90
Cardnum_MerchDescription_day_since	Cardnum_MerchState_count_1_by_3
Cardnum_MerchDescription_count_0	Cardnum_MerchState_count_1_by_7
Cardnum_MerchDescription_avg_0	Cardnum_MerchState_count_1_by_14
Cardnum_MerchDescription_max_0	Cardnum_MerchState_count_1_by_30
Cardnum_MerchDescription_med_0	Cardnum_MerchState_count_1_by_60
Cardnum_MerchDescription_total_0	Cardnum_MerchState_count_1_by_90
Cardnum_MerchDescription_actual/avg_0	Cardnum_MerchZip_count_0_by_3
Cardnum_MerchDescription_actual/max_0	Cardnum_MerchZip_count_0_by_7
Cardnum_MerchDescription_actual/med_0	Cardnum_MerchZip_count_0_by_14
Cardnum_MerchDescription_actual/total_0	Cardnum_MerchZip_count_0_by_30
Cardnum_MerchDescription_count_1	Cardnum_MerchZip_count_0_by_60
Cardnum_MerchDescription_avg_1	Cardnum_MerchZip_count_0_by_90
Cardnum_MerchDescription_max_1	Cardnum_MerchZip_count_1_by_3
Cardnum_MerchDescription_med_1	Cardnum_MerchZip_count_1_by_7
Cardnum_MerchDescription_total_1	Cardnum_MerchZip_count_1_by_14
Cardnum_MerchDescription_actual/avg_1	Cardnum_MerchZip_count_1_by_30
Cardnum_MerchDescription_actual/max_1	Cardnum_MerchZip_count_1_by_60
Cardnum_MerchDescription_actual/med_1	Cardnum_MerchZip_count_1_by_90
Cardnum_MerchDescription_actual/total_1	Cardnum_MerchDescription_count_0_by_3
Cardnum_MerchDescription_count_3	Cardnum_MerchDescription_count_0_by_7
Cardnum_MerchDescription_avg_3	Cardnum_MerchDescription_count_0_by_14
Cardnum_MerchDescription_max_3	Cardnum_MerchDescription_count_0_by_30
Cardnum_MerchDescription_med_3	Cardnum_MerchDescription_count_0_by_60
Cardnum_MerchDescription_total_3	Cardnum_MerchDescription_count_0_by_90
Cardnum_MerchDescription_actual/avg_3	Cardnum_MerchDescription_count_1_by_3
Cardnum_MerchDescription_actual/max_3	Cardnum_MerchDescription_count_1_by_7
Cardnum_MerchDescription_actual/med_3	Cardnum_MerchDescription_count_1_by_14
Cardnum_MerchDescription_actual/total_3	Cardnum_MerchDescription_count_1_by_30
Cardnum_MerchDescription_count_7	Cardnum_MerchDescription_count_1_by_60
Cardnum_MerchDescription_avg_7	Cardnum_MerchDescription_count_1_by_90
Cardnum_MerchDescription_max_7	Merchnum_Merchzip_count_0_by_3
Cardnum_MerchDescription_med_7	Merchnum_Merchzip_count_0_by_7
Cardnum_MerchDescription_total_7	Merchnum_Merchzip_count_0_by_14
Cardnum_MerchDescription_actual/avg_7	Merchnum_Merchzip_count_0_by_30

Cardnum_MerchDescription_actual/max_7	Merchnum_Merchzip_count_0_by_60
Cardnum_MerchDescription_actual/med_7	Merchnum_Merchzip_count_0_by_90
Cardnum_MerchDescription_actual/total_7	Merchnum_Merchzip_count_1_by_3
Cardnum_MerchDescription_count_14	Merchnum_Merchzip_count_1_by_7
Cardnum_MerchDescription_avg_14	Merchnum_Merchzip_count_1_by_14
Cardnum_MerchDescription_max_14	Merchnum_Merchzip_count_1_by_30
Cardnum_MerchDescription_med_14	Merchnum_Merchzip_count_1_by_60
Cardnum_MerchDescription_total_14	Merchnum_Merchzip_count_1_by_90
Cardnum_MerchDescription_actual/avg_14	Merchnum_MerchState_count_0_by_3
Cardnum_MerchDescription_actual/max_14	Merchnum_MerchState_count_0_by_7
Cardnum_MerchDescription_actual/med_14	Merchnum_MerchState_count_0_by_14
Cardnum_MerchDescription_actual/total_14	Merchnum_MerchState_count_0_by_30
Cardnum_MerchDescription_count_30	Merchnum_MerchState_count_0_by_60
Cardnum_MerchDescription_avg_30	Merchnum_MerchState_count_0_by_90
Cardnum_MerchDescription_max_30	Merchnum_MerchState_count_1_by_3
Cardnum_MerchDescription_med_30	Merchnum_MerchState_count_1_by_7
Cardnum_MerchDescription_total_30	Merchnum_MerchState_count_1_by_14
Cardnum_MerchDescription_actual/avg_30	Merchnum_MerchState_count_1_by_30
Cardnum_MerchDescription_actual/max_30	Merchnum_MerchState_count_1_by_60
Cardnum_MerchDescription_actual/med_30	Merchnum_MerchState_count_1_by_90
Cardnum_MerchDescription_actual/total_30	Merch state_count_0_by_3
Cardnum_MerchDescription_count_60	Merch state_count_0_by_7
Cardnum_MerchDescription_avg_60	Merch state_count_0_by_14
Cardnum_MerchDescription_max_60	Merch state_count_0_by_30
Cardnum_MerchDescription_med_60	Merch state_count_0_by_60
Cardnum_MerchDescription_total_60	Merch state_count_0_by_90
Cardnum_MerchDescription_actual/avg_60	Merch state_count_1_by_3
Cardnum_MerchDescription_actual/max_60	Merch state_count_1_by_7
Cardnum_MerchDescription_actual/med_60	Merch state_count_1_by_14
Cardnum_MerchDescription_actual/total_60	Merch state_count_1_by_30
Cardnum_MerchDescription_count_90	Merch state_count_1_by_60
Cardnum_MerchDescription_avg_90	Merch state_count_1_by_90
Cardnum_MerchDescription_max_90	Merch zip_count_0_by_3
Cardnum_MerchDescription_med_90	Merch zip_count_0_by_7
Cardnum_MerchDescription_total_90	Merch zip_count_0_by_14
Cardnum_MerchDescription_actual/avg_90	Merch zip_count_0_by_30
Cardnum_MerchDescription_actual/max_90	Merch zip_count_0_by_60
Cardnum_MerchDescription_actual/med_90	Merch zip_count_0_by_90
Cardnum_MerchDescription_actual/total_90	Merch zip_count_1_by_3
Merchnum_Merchzip_day_since	Merch zip_count_1_by_7
Merchnum_Merchzip_count_0	Merch zip_count_1_by_14
Merchnum_Merchzip_avg_0	Merch zip_count_1_by_30
Merchnum_Merchzip_max_0	Merch zip_count_1_by_60

Merchnum_Merchzip_med_0	Merch zip_count_1_by_90
Merchnum_Merchzip_total_0	Cardnum_Merchnum_MerchState_count_0_by_3
Merchnum_Merchzip_actual/avg_0	Cardnum_Merchnum_MerchState_count_0_by_7
Merchnum_Merchzip_actual/max_0	Cardnum_Merchnum_MerchState_count_0_by_14
Merchnum_Merchzip_actual/med_0	Cardnum_Merchnum_MerchState_count_0_by_30
Merchnum_Merchzip_actual/total_0	Cardnum_Merchnum_MerchState_count_0_by_60
Merchnum_Merchzip_count_1	Cardnum_Merchnum_MerchState_count_0_by_90
Merchnum_Merchzip_avg_1	Cardnum_Merchnum_MerchState_count_1_by_3
Merchnum_Merchzip_max_1	Cardnum_Merchnum_MerchState_count_1_by_7
Merchnum_Merchzip_med_1	Cardnum_Merchnum_MerchState_count_1_by_14
Merchnum_Merchzip_total_1	Cardnum_Merchnum_MerchState_count_1_by_30
Merchnum_Merchzip_actual/avg_1	Cardnum_Merchnum_MerchState_count_1_by_60
Merchnum_Merchzip_actual/max_1	Cardnum_Merchnum_MerchState_count_1_by_90
Merchnum_Merchzip_actual/med_1	Cardnum_Merchnum_MerchZip_count_0_by_3
Merchnum_Merchzip_actual/total_1	Cardnum_Merchnum_MerchZip_count_0_by_7
Merchnum_Merchzip_count_3	Cardnum_Merchnum_MerchZip_count_0_by_14
Merchnum_Merchzip_avg_3	Cardnum_Merchnum_MerchZip_count_0_by_30
Merchnum_Merchzip_max_3	Cardnum_Merchnum_MerchZip_count_0_by_60
Merchnum_Merchzip_med_3	Cardnum_Merchnum_MerchZip_count_0_by_90
Merchnum_Merchzip_total_3	Cardnum_Merchnum_MerchZip_count_1_by_3
Merchnum_Merchzip_actual/avg_3	Cardnum_Merchnum_MerchZip_count_1_by_7
Merchnum_Merchzip_actual/max_3	Cardnum_Merchnum_MerchZip_count_1_by_14
Merchnum_Merchzip_actual/med_3	Cardnum_Merchnum_MerchZip_count_1_by_30
Merchnum_Merchzip_actual/total_3	Cardnum_Merchnum_MerchZip_count_1_by_60
Merchnum_Merchzip_count_7	Cardnum_Merchnum_MerchZip_count_1_by_90
Merchnum_Merchzip_avg_7	Cardnum_U*
Merchnum_Merchzip_max_7	Merchnum_U*