USC Marshall
School of Business

Fraud Analytics
DSO562 Project 2

Instructor: Stephen Coggeshall

Team 7

Team Members
Ruishan Jiang, Nianzhe Li, Yiqing Song
Yin Wang, Lixuan Lin, Luding Zhang

March 2020

Table of Contents

1.Executive Summary

Identity fraud is a common problem when it comes to credit risk. It is an act of misrepresenting which person you are to improperly get products or services (financial services, consumer products…) or to stay “unidentified” while behaving badly (money laundering, terrorists…). The behavior is generally classified as three modes: Identity Theft, Identity Manipulation and Synthetic Identity.

The purpose of this Application (Identify) Fraud project is to look for anomalies among records of “application data” file using a real time fraud algorithm (supervised models) normally applied to deal with identity fraud. Our goal is to build and tune a model that has the highest FDR (fraud detection rate) when examining 3% of the record population. And we believe such a model will help find out the optimal division point to target suspicious applications from a business perspective.

The original data set represents product application data from 2016-01-01 to 2016-12-31, contains 1,000,000 records and has 10 fields describing the details of the application, like application date, ssn, address, name, zip code and date of birth of the applicant. Among all the fields, 8 fields are categorical variables, 1 field is date, and 1 field is index.

Followed by the detailed exploration of the dataset, we did data cleaning by treating the frivolous fields (SSN, Address, Phone and DOB), created (applying concepts of application velocity and time interval) and scaled all variables (including risk table created by records before Nov 1st) and separated data into modeling (January to October) and Out of Time (OOT) (November to December),

Then we used the modeling data to select the best variables. Two feature selection steps filter (calculate the univariate KS and univariate FDR at 3%) and wrapper (backward stepwise selection) were employed to decrease the number of variables from over 600 to 19.

As for the modeling process, we first built a baseline linear model, Logistic Regression model. Then we built more complex non-linear models including SVM, Random Forest, Boosted Trees and Deep Neural Network. Among all the models, Random Forest performed the best with a 0.525985 fraud detection rate for OOT dataset. Referring to this model’s prediction result, product providers might be able to find an optimal cutoff point for detecting fraudulent applications or to alter their existing fraud detection model.

2.Description of Data

2.1 Data overview

This dataset contains 1000000 records of applications from 2016-01-01 to 2016-12-31, and 14393 of them are fraud applications as their fraud_lable values are marked as 1. There are 10 fields in the dataset: record, date, ssn, firstname, lastname, address, zip5, dob, homephone, fraud_lable. Among all the fields, 8 fields are categorical variables, 1 field is date, and 1 field is index.

A Data Quality Report in the appendix was attached for further details.

2.2 Summary of fields

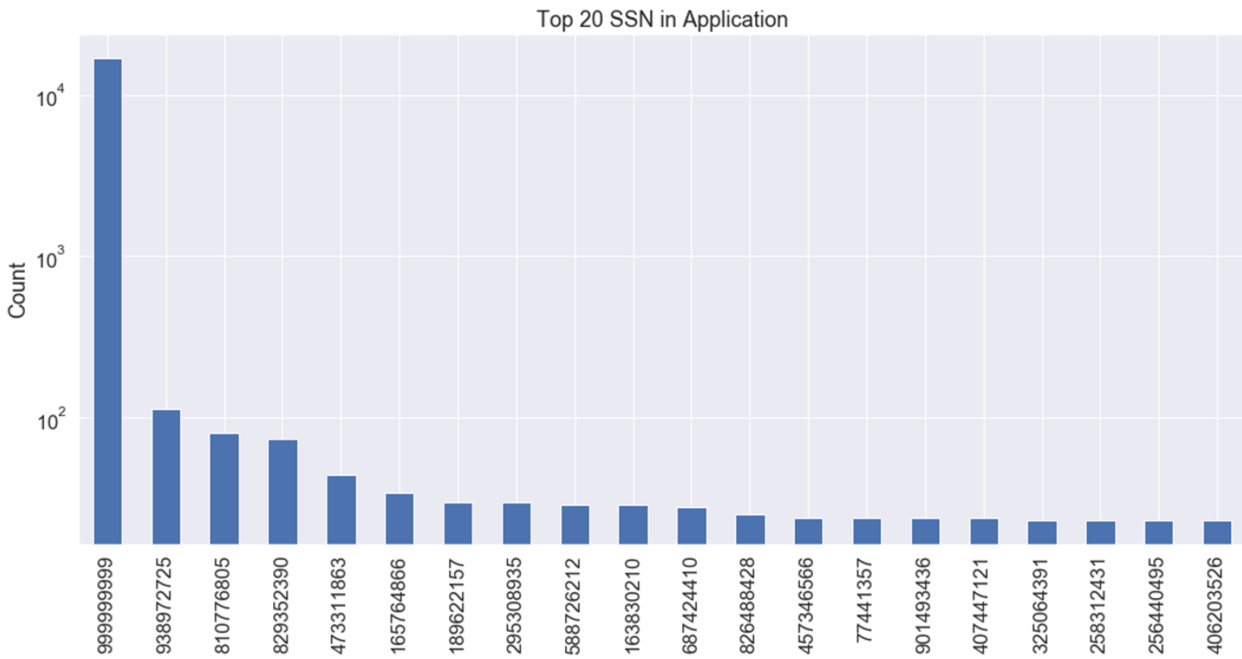
2.2.1 Categorical values

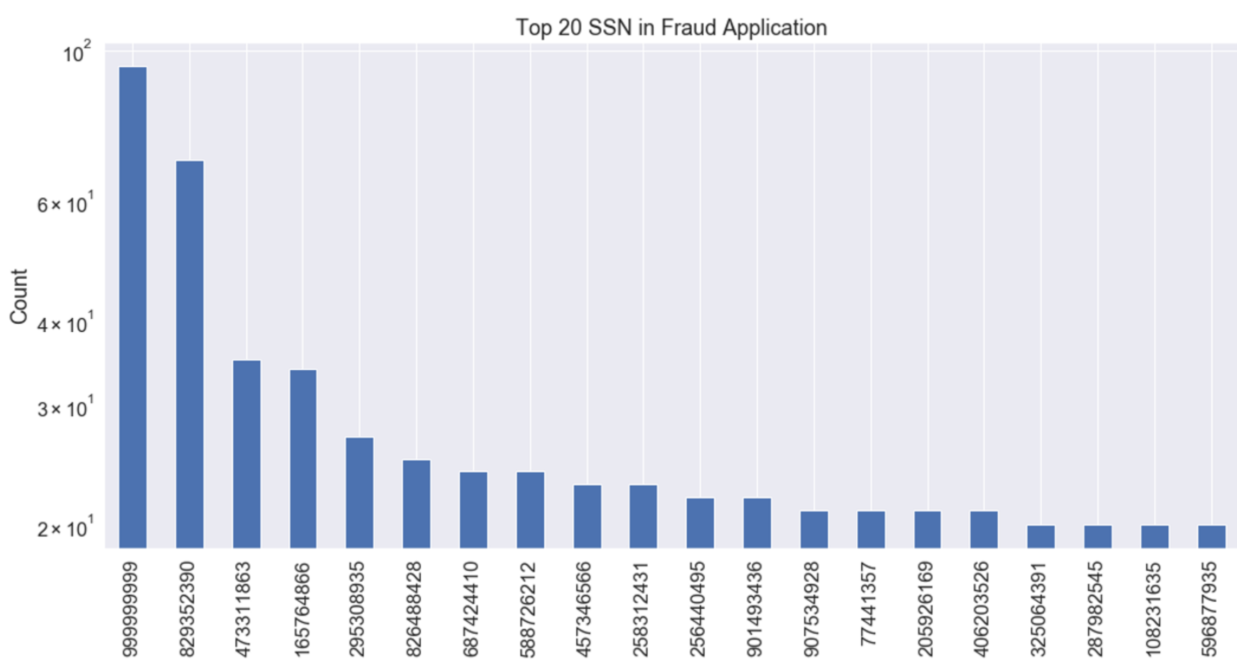
Field Name	**# records that have a value**	**% populated**	**# unique values**	**most common field value**	**# most common value**
Record	1000000	100.00%	1000000	/	/
Date	1000000	100.00%	365	20160816	2877
SSN	1000000	100.00%	835819	999999999	16935
firstname	1000000	100.00%	78136	EAMSTRMT	12658
lastname	1000000	100.00%	177001	ERJSAXA	8580
Address	1000000	100.00%	828774	123 MAIN ST	1079
Zip5	1000000	100.00%	26370	68138	823
Dob	1000000	100.00%	42673	19070626	126568
homephone	1000000	100.00%	28244	9999999999	78512
fraud_label	1000000	100.00%	2	0	985607

2.2.2 Important Fields Distribution

Field: ssn

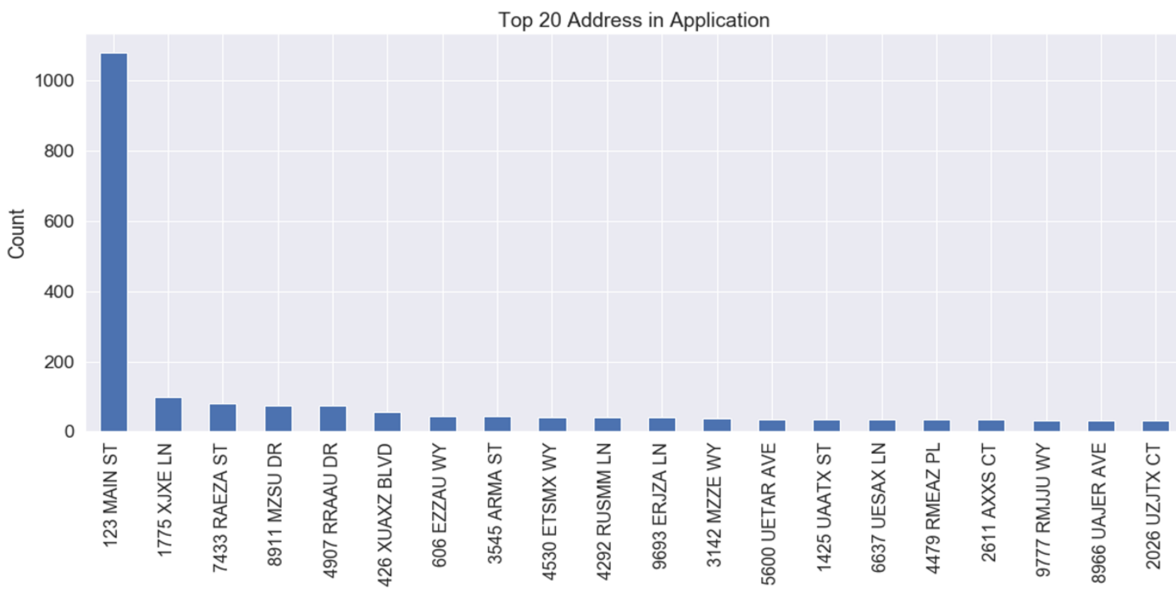
SSN represents the social security number for each applicant. This field has 835819 unique values. The following two graphs show the top 20 SSN values among all applications and fraud applications. 999999999 is the most common ssn, which may indicate that 999999999 is a frivolous value.





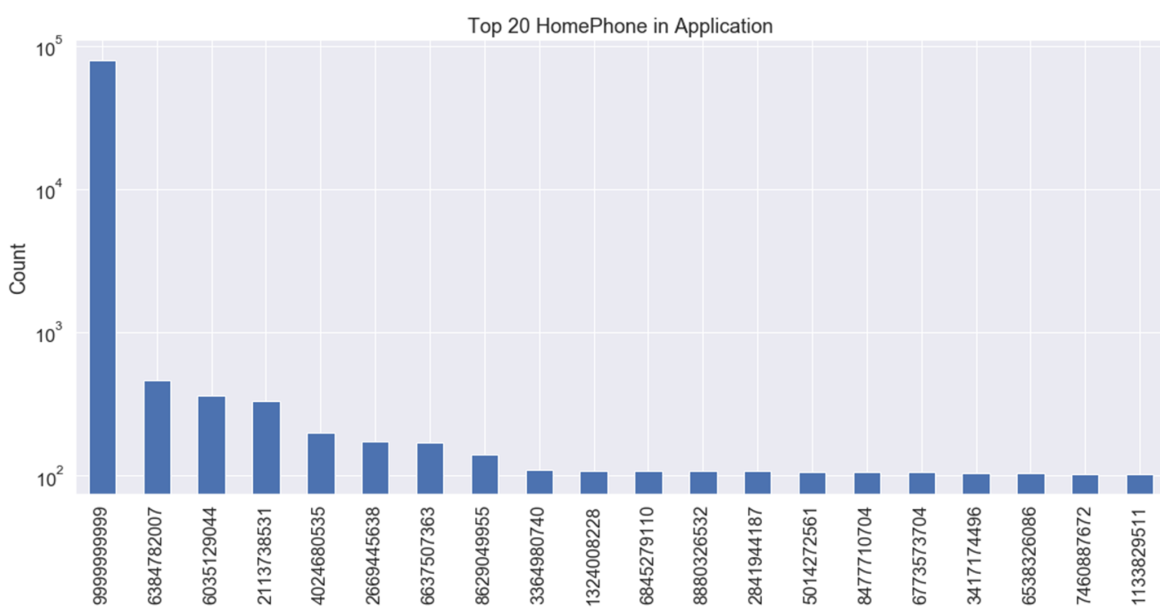
Field: address

This field represents the address of each applicant and has 828774 unique values. The following graph shows the top 20 addresses among all applications. ‘123 MAIN ST’ is the most common address, which may indicate that ‘123 MAIN ST’ is a frivolous value.



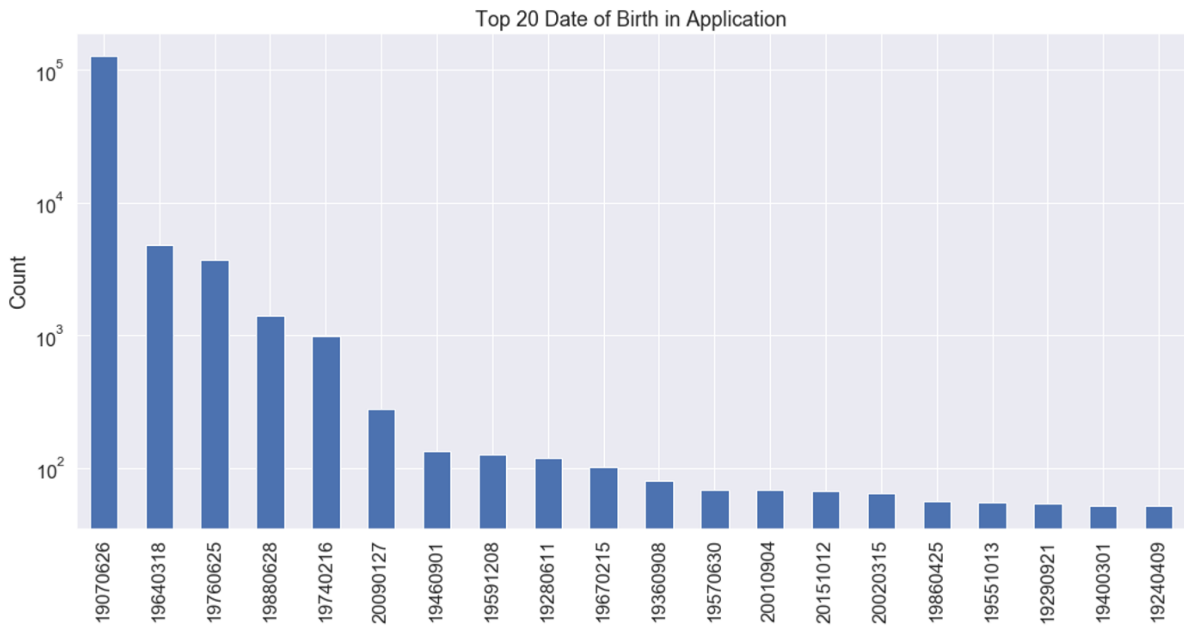
Field name: homephone

This field represents the home phone number of each applicant and has 28244 unique values. ‘9999999999’ is the most common home phone number, which may indicate that ‘9999999999’ is a frivolous value.



Field name: dob

This field represents the date of birth of each applicant and has 42673 unique values. The following graph shows the top 20 dates of birth among all applications, and '19070626' is the most common date of birth. Since the person who was born in 1907 would be more than 100 years old in 2016, it may indicate that '19070626' is a frivolous value.



3. Data Cleaning

3.1 Number of digits

In the dataset, we found that the fields “zip5”, “ssn” and “homephone” contain different numbers of digits. Since we would concatenate entities to build our variables, the format of these fields should keep uniform. We added “0” before those values whose length is less than 5 digits (for zip5), 9 digits (for ssn), and 10 digits (for homephone) to ensure that each value of “zip5”, “ssn”, and “homephone” column is 5-digit-long, 9-digit-long and 10-digit-long.

3.2 Frivolous Values

We also noticed there are several extremely frequent-appeared values in some fields. We examined these values and found out they did not appear frequently among fraudulent applications. Leaving these values unhandled would cause serious problems. As the times of a certain identity used among the past applications would affect the identity’s risky level. A certain identity with higher frequency will be more likely related to a fraud application. Additionally, for example, “999999999” for ssn is obviously without meaning.

We found the following fillings (in the table) appeared so many times and having abnormal values, which meant that these fillings could be made up or randomly typed when applicants filled in the forms. To make sure frivolous values will not have great influences when building models, we replaced these values with the record number of the corresponding application. In this way, each record would have a unique value to substitute the frivolous value, and would not have any linkage with other records.

Field name	most common field value	# most common value
ssn	999999999	16935
address	123 MAIN ST	1079
dob	19070626	126568
homephone	9999999999	78512

4. Candidate Variables

4.1 Basic Entities for Linking

First, we selected 6 basic entities which we believe might help us build meaningful combination groups. We used full name here, because we thought that full name could represent an applicant better than first name or last name. And we also treated “zip5” as a basic entity, since we believed that zip code could have some meanings when combining with other entities. For synthetic identities, “ssn-zip5” or “homephone-zip5” might be a fraud indicator when they were chosen or combined randomly.

The 6 basic entities are listed in the table below:

Entities	Description
ssn	The social security number for each applicant.
address	The address for each applicant.
homephone	The home phone number for each applicant.
name	full name (the concatenation of 'firstname' and 'lastname') for each applicant.
zip5	The ZIP code for each address.
dob	The date of birth for each applicant.

4.2 Selected Entities and Combination Groups

To build the variables which could help detect fraud applications, we first needed to select related entities or combination groups whose frequency and other related attributes had strong indicative information. Considering that identity theft or identity synthesis might use different combinations of entities when submitting a new application, we used iterations to find all meaningful combinations consisting of 2 or 3 entities. Additionally, we also included some groups consisting of 4 or 5 entities which we believed represent full information of an applicant, like “name-ssn-homephone-dob” and ‘name-ssn-homephone-dob-address-zip5”.

Based on these 6 basic entities, we selected and built 42 entities and combination groups (listed in the table below) which could be used in the following candidate variables creation process. In the table, "-" means concatenating. (e.g. ssn-dob is the concatenation of ssn value and dob value of one record.)

Entities Names		
ssn	address-name	ssn-name-zip5
address	address-zip5	dob-address-homephone
homephone	homephone-name	dob-address-name
name	homephone-zip5	dob-address-zip5
ssn-dob	name-zip5	dob-homephone-name
ssn-address	ssn-dob-address	dob-homephone-zip5
ssn-homephone	ssn-dob-homephone	dob-name-zip5
ssn-name	ssn-dob-name	address-homephone-name
ssn-zip5	ssn-dob-zip5	address-homephone-zip5
dob-address	ssn-address-homephone	address-name-zip5
dob-homephone	ssn-address-name	homephone-name-zip5
dob-name	ssn-address-zip5	name-ssn-homephone-dob-address-zip5
dob-zip5	ssn-homephone-name	name-ssn-homephone-dob
address-homephone	ssn-homephone-zip5	name-ssn-address-zip5

4.3 Variables Creation

Based on the above 42 selected entities and combination groups, we built the following 5 kinds of variables. Since we created 605 candidate variables and the full list was so long, we used a general name to represent each type of variables here and put the full list in Appendix 2.

Because this dataset only used date as the application time, we treated the record number as time sequence when creating time related variables. We assumed that a smaller record happened before a larger record. Then for each of the 42 selected entities and combination groups, we calculated the following:

Entity_count(n)_date (Velocity variable):

The times a certain entity or a combination group appears during the past n days. We chose the time period "n" as 1,3,7,30,90,180. For fraudsters, they may use made-up entities or randomly select some entities when submitting an application. So if a certain entity or combination group appears frequently, it may have a higher risk to be a fraud application. For this reason, the large value of this variable represents that this entity or combination group is more likely related to a fraud application.

Entity_0_count(n)_count_ratio (Relative velocity variable):

The times an entity or a combination group appears in the same day divided by the times the same entity or combination group appears in the past n days. We chose the time period "n" as 1,3,7,30,90,180. This variable represents how frequent an entity or a combination group would appear in more recent past.

Entity_pastday (Days since last seen variable):

The days since we last saw the same entity or combination group. A larger value represents that this entity or combination group would be less likely related to a fraud application. For those records' values which appear for the first time, we assign the variable value "365", meaning that this entity or combination group is at a lowest risk.

Entity1_unique_entity2 (Unique counts of seen for entity2 in entity1 group):

For those identity thieves, they may use different identities but use their own address or phone number when filling in the application. Under this circumstance, we built variables to calculate how many unique identities are correlated with another certain entity value. These variables represent that for each entity1 value, in the past history, how many unique entity2 values are related. For example, if a certain address record has 5 unique ssn related to it in the past, then for this application, address_unique_ssn is 5. A higher value represents that this application is more likely to be a fraud one.

Day_week_risk (Risk table variable):

For each day of a week (from Monday to Sunday), we calculated the proportion of fraud_label of the first 10 months for each group, then assigned these group proportions to each record to get the risk table value. Even though from the exploratory data analysis we did not find out seasonality in day_week_risk for the fraud application, we still wanted to make sure that we took seasonality into consideration during variable creation and didn't ignore any information during the first steps.

After creating all the variables, we z-scaled them for future utilization.

5. Feature Selection Process

Feature selection is the process of selecting a subset of relevant features for model construction. It is used to enhance model training efficiency as well as to avoid the curse of dimensionality. In this process, we used two methods (i.e. filter and wrapper) to find 19 most representative variables among the 605 candidate variables.

We didn't use all the data but extracted those from January to October for this process. Also, we kept the values in the first few weeks in January due to our "365" value assignment method.

5.1 Filter

5.1.1 Kolmogorov-Smirnov test (KS test)

KS test is an efficient way to determine how well two distributions are separated by examining the maximum difference between two cumulative distributions. The larger the difference is, the more separated the two distributions are. Thus, univariate KS test is said to be an outstanding feature selection method for binary classification and can be applied to find out how powerful each variable is in separating fraud and non-fraud applications. The higher KS score a variable has, the more effective it is to predict fraud applications.

Below are the steps of computing KS score:

Step1: Separate the January-October data into two subsets according to the fraud label.

Step2: Conduct the KS test to examine how well each candidate variable separates the fraud and non-fraud applications.

Step3: Re-rank candidate variables based on KS Score.

5.1.2 Fraud Detection Rate (FDR)

Fraud Detection Rate (FDR) is a common measurement of goodness for fraud models. It is the percentage of the frauds caught at a score cutoff. In this case, we used top 3% as a cutoff point. We evaluated each variable's performance by calculating the percentage of frauds it helped catch when examining the top 3% of the population (both in ascending and descending orders). The higher FDR score a variable has, the more effective it is to predict fraud applications.

Below are the steps of computing FDR score:

Step1: Sort each variable in both ascending order and descending orders.

Step2: Calculate the fraud detection rate at 3% for each variable under both orders.

Step3: Comparing the two rates and selecting the larger one as the final FDR score.

Step4: Re-rank candidate variables on final FDR score.

5.1.3 The Final Result with KS and FDR

After we got two rankings according to KS score and FDR score, we calculated a final ranking score with the average of these two rankings. Using the final ranking score, we managed to filter out more than 500 variables and saved 80 variables for further feature selection. The chart below shows 16 candidate variables which have the highest average rank score.

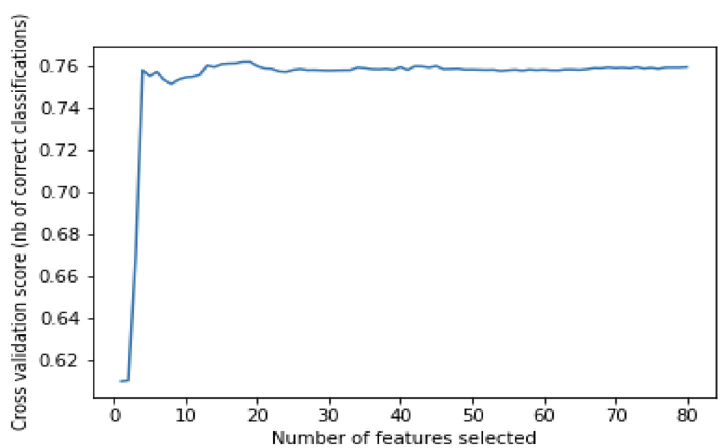
field	KS	FDR	rank_KS	rank_FDR	average_rank
fraud_label	1.000	1.000	607	607	607
address_pastday	0.330	0.352	606	605	605.5
address-zip5_pastday	0.329	0.354	605	606	605.5
address_count30_date	0.328	0.348	604	603	603.5
address-zip5_count30_date	0.328	0.351	603	604	603.5
address_count90_date	0.316	0.342	602	602	602
address-zip5_count90_date	0.315	0.337	601	600	600.5
address_count180_date	0.314	0.339	600	601	600.5
address-zip5_count180_date	0.313	0.337	599	599	599
address_count7_date	0.299	0.322	598	598	598
address-zip5_count7_date	0.298	0.320	597	597	597
address-zip5_unique_ssn	0.297	0.319	596	596	596
address-zip5_unique_name	0.296	0.318	595	595	595
address-zip5_unique_dob	0.289	0.315	594	594	594
address_0_count30_count_ratio	0.289	0.261	593	586	589.5
address-zip5_0_count30_count_ratio	0.287	0.264	592	587	589.5

5.2 Wrapper

A wrapper method proceeds feature selection by using a specific machine learning algorithm. It evaluates all the possible combinations of features against the evaluation criterion using a greedy search approach. The evaluation criterion measures the model performance under certain combinations of features. For regression problems, the evaluation criterion can be p-values, R-squared, Adjusted R-squared, etc. Likewise, for classification problems, the evaluation criterion can be accuracy, ROC score, f1-score, etc. Lastly, It selects the combination of features that gives the optimal results for the specified machine learning algorithm.

There are three commonly used techniques under wrapper methods: forward selection, backward selection, and bi-directional selection. In this project, we chose to use logistic regression as the algorithm to do backward feature selection, as logistic regression is one of the simplest while most powerful models for binary classification. Using the backward feature selection method, all the variables will be used to build the first model, and then the algorithm will examine all variables one by one to decide which one should first be removed. A new model will be built using the remaining variables and the second variable will be removed. Such degradation process will continue until removing any variable can significantly lower the model performance.

In our case, we used “roc_auc” as the evaluation criterion and 19 variables were selected after implementing the backward selection method. The following plot depicts the relationship between the model’s performance (the area under the receiver operating characteristic curve) and the number of features selected. We can see there is a slight performance deterioration when using a combination of 18 variables to predict fraud, while a significant deterioration when using less than 5 variables. Trying to keep around 20 - 30 candidate variables, we decided the cutoff point is 19.



Below are the 19 variables chosen:

address-homephone-zip5_count180_date	address-homephone-zip5_count7_date	address-homephone-zip5_pastday	address-homephone_count180_date
address-homephone_pastday	address-zip5_0_count180_count_ratio	address-zip5_0_count30_count_ratio	address-zip5_0_count7_count_ratio
address-zip5_count30_date	address-zip5_pastday	dob-name_count30_date	dob-name_pastday
homephone-zip5_count30_date	ssn-dob_pastday	ssn-name_count180_date	ssn-name_count7_date
ssn-name_pastday	ssn_count180_date	ssn_unique_address-zip5	

6. Model Algorithms

We implemented five different machine learning algorithms (Logistics Regression, Random Forest, XGBoost, Neural Network, SVM) to build several fraud detection models. We trained the model using three train and test split dataset. Train and test datasets were data from January to October, and the last two months were treated as out of time (OOT) dataset to help us validate our models. We then tuned each model with important parameters. And we calculated the fraud detection rate of each model to evaluate these models' performance.

All the FDRs in the following tables are the average results after three times model training.

6.1 Logistic Regression (baseline linear model)

Logistic regression is the baseline model we built to help us get the baseline performance of our models, and the performance of more complex models we built later should be better than this model.

Logistic regression is actually a binary classifier. Although it can be used for classification, its essence is linear regression. On the basis of linear regression, it adds a layer of sigmoid function (non-linear) mapping the features to the results. In other words, the features are summed linearly first, and then the sigmoid function is used to calculate the final result. However, it is this simple logical function that makes logistic regression model a shining star in the field of machine learning.

After building this model, it was tuned by the solver parameter, the FDR results are shown as below.

solver	**train_fdr**	**test_fdr**	**oot_fdr**
newton-cg	0.407100	0.400759	0.383068
lbfgs	0.405854	0.400759	0.383487
liblinear	0.407100	0.400759	0.383068
sag	0.409280	0.400759	0.383068
saga	0.404816	0.400337	0.383068

6.2 Random Forest

Random forest is an algorithm that integrates multiple strong decision trees through the idea of integrated learning. It is one of the most popular ensemble learning models. There are two key words in the name of random forest, one is 'random', the other is 'forest'.

‘Forest’ is easy to understand. If one tree is called a tree, then hundreds or thousands of trees can be called a forest. This analogy is very appropriate. As for ‘random’, the training samples for each decision tree are randomly selected from the whole training set. If the feature dimension of each sample is M, specify a constant m<<M, randomly select m feature subsets from M features, and select the optimal feature from M features each time the tree is split.

In fact, from an intuitive perspective, every decision tree is a classifier, so for an input sample, N trees may have N classification results. The random forest integrates all the classification voting results and specifies the category with the most votes as the final output, this is the simplest Bagging idea.

After building this model, it was tuned by the parameters n_estimators, max_depth, and min_samples_leaf. FDR results with corresponding parameters are shown as below.

n_estimators	max_depth	min_samples_leaf	train_fdr	test_fdr	oot_fdr
20	10	5	0.538925	0.536452	0.512992
20	10	10	0.539028	0.536452	0.523051
20	20	5	0.551381	0.528445	0.506287
20	20	10	0.550446	0.531816	0.515926
50	10	5	0.539547	0.537716	0.521375
50	10	10	0.540170	0.537295	0.525985
50	20	5	0.551692	0.530552	0.517184
50	20	10	0.550758	0.532238	0.517603
80	10	5	0.539547	0.537295	0.522213
80	10	10	0.539547	0.536452	0.522632
80	20	5	0.551484	0.530552	0.516764
80	20	10	0.550758	0.530973	0.518860

6.3 XGBoost

The full name for XGBoost is extreme gradient boosting. XGBoost is an optimized distributed gradient enhancement library designed to be efficient, flexible, and portable. It implements the machine learning algorithm under the Gradient Boosting framework. It is one of the boosting algorithms. Boosting algorithm's idea is to integrate many weak classifiers to form a strong classifier. Because XGBoost is an ascending tree model, it is a combination of many tree models to form a strong classifier. The tree model used is the CART regression tree model.

The idea of this algorithm is to keep adding trees, keep doing feature splitting to grow a tree, adding one tree at a time, actually learning a new function to fit the residual predicted last time. When we complete the training and get k trees, we need to predict the score of a sample. In fact, according to the characteristics of the sample, the corresponding leaf node will fall in each tree, and each leaf node will correspond to a score. Finally, we just need to add up corresponding scores of each tree to be the predicted value of the sample.

After building this model, it was tuned by the parameters n_estimators, max_depth, and learning_rate. FDR results are shown in the below table.

n_estimators	**max_depth**	**learning_rate**	**train_fdr**	**test_fdr**	**oot_fdr**
100	2	0.1	0.478306	0.480826	0.346186
800	2	0.1	0.492942	0.493890	0.304694
100	2	0.01	0.480071	0.483776	0.466890
800	2	0.01	0.478202	0.480826	0.369656
100	2	0.001	0.480071	0.483776	0.466890
800	2	0.001	0.480071	0.483776	0.466890
100	5	0.1	0.500623	0.500632	0.315172
800	5	0.1	0.503322	0.493468	0.295474
100	5	0.01	0.493564	0.498104	0.341576
800	5	0.01	0.500727	0.501896	0.312657
100	5	0.001	0.498235	0.500632	0.298826
800	5	0.001	0.496263	0.499789	0.341576

6.4 Deep Neural Network (DNN)

Deep neural network trains several layers based on the input variables. Each input variable x_i is treated as a single neuron of the input layer. Then from the input layer, DNN adds several hidden layers between input layer and output layer. Each input neuron of input layer is connected to the first hidden layer neuron with a weight w_i, which represents the strength of the interconnections between neurons in a neural network. The aggregated signal $\sum_{i=1}^{n} w_i x_i$ is then passed through an activation function to get the output, which is treated as the neuron for the following hidden layer. And then for each neuron in the second hidden layer, it is connected to the first hidden layer also with a weight, etc. For hidden layers, we used the activation function “ReLU”: $\sigma(x) = max(0, x)$.

For our binary classification question, the output layer is a single value which represents the probability to be the fraud label. For this output layer we chose the activation function "Sigmoid": $\sigma(x) = \frac{e^{x\beta}}{1+e^{x\beta}}$

We have 19 input variables, so our input shape is 19. Then we add 2 hidden layers and one output layer with the corresponding activation function. The tuning parameters are the number of neurons for each hidden layer, epoch number and batch size. Among them epoch number will change the number of iterations and batch size will influence the input sample size in each step. The table below shows results for deep neural network models.

first_layer	**second_layer**	**epoch**	**batchsize**	**train_fdr**	**test_fdr**	**oot_fdr**
16	12	20	6400	0.53900	0.51755	0.51760
16	12	20	3200	0.53806	0.51924	0.51970
16	12	30	6400	0.53827	0.51797	0.51551
16	12	30	3200	0.53848	0.51839	0.52221
16	7	20	6400	0.53360	0.51290	0.51132
16	7	20	3200	0.53661	0.51501	0.51551
16	7	30	6400	0.53630	0.51797	0.52263
16	7	30	3200	0.53879	0.51966	0.51718
10	12	20	6400	0.53765	0.51882	0.52305
10	12	20	3200	0.53734	0.51839	0.51635
10	12	30	6400	0.53796	0.51670	0.51593
10	12	30	3200	0.53754	0.51882	0.52263
10	7	20	6400	0.53495	0.51586	0.51676
10	7	20	3200	0.53827	0.51755	0.52263
10	7	30	6400	0.53205	0.50994	0.50838
10	7	30	3200	0.53890	0.51797	0.52305

6.5 SVM

The support vector machine (SVM) is a classification method used for classification problems. It effectively detects a non-linear decision boundary between classes by using the kernels functions to enlarge the feature space.The kernel functions provide comprehensive ways to calculate the inner product of the variables and thus expanding the space. To find the optimal boundary (a hyperplane) in the enlarged space, the maximal margin classifier is employed. The maximal margin classifier will calculate the euclidean distance between all the data points and the hyperplane in order to find the parameters that maximize the gap between two classes. The support vectors are selected during such calculation process. They are vectors that are located close to the boundary and help the most to find the best hyperplane. During the process of selecting the support vectors, the SVM algorithm is actually doing an embedded feature selection to squashed down the dimension and just save several important ones.

After building this model, it was tuned by the parameters kernel, C, and gamma. The kernel parameter has the value of linear, poly and rbf, representing three kernel functions. The radial basis function (rbf) is said to be the most popular kernel function. And the C parameter represents the error budget in each classification set. The gamma parameter is a parameter that appears in the rbf kernel function.

kernel	**c**	**gamma**	**train_fdr**	**test_fdr**	**ott_fdr**
linear	0.1	-	0.50447544	0.50187578	0.48700754
linear	1	-	0.50697336	0.50062526	0.48742666
linear	10	-	0.50437136	0.50104210	0.48365465
rbf	0.01	auto	0.52612406	0.52438516	0.50125733
rbf	0.01	scale	0.52633222	0.52396832	0.50209556
rbf	1	auto	0.52029559	0.51979992	0.50419111
rbf	1	scale	0.52144047	0.52105044	0.50502934
rbf	10	auto	0.51394671	0.51729887	0.50335289
rbf	10	scale	0.51582015	0.52105044	0.50502934

7. Results

We picked Random Forest as our final model using parameters n_estimators = 50, max_depth = 10, min_samples_leaf = 10. This model performed the best with FDR (fraud detection rate) at 3% as 0.54017, 0.537295 and 0.525985 for train, test and OOT dataset respectively (average FDR after three times of model training).

The cumulative statistics showed as following (the first 20 rows of the three tables, the full tables are attached as appendix3).

Training	#Records	#Goods	#Bads	Fraud Rate								
	666805	657171	9634	0.014448002								
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total #Records	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	6668	1703	4965	25.54%	74.46%	6668	1703	4965	0.26%	51.54%	51.28	0.50
2	6668	6479	189	97.17%	2.83%	13336	8182	5154	1.25%	53.50%	52.25	2.33
3	6668	6623	45	99.33%	0.67%	20004	14805	5199	2.25%	53.97%	51.71	4.17
4	6668	6625	43	99.36%	0.64%	26672	21430	5242	3.26%	54.41%	51.15	5.99
5	6668	6609	59	99.12%	0.88%	33340	28039	5301	4.27%	55.02%	50.76	7.75
6	6668	6622	46	99.31%	0.69%	40008	34661	5347	5.27%	55.50%	50.23	9.50
7	6668	6616	52	99.22%	0.78%	46676	41277	5399	6.28%	56.04%	49.76	11.21
8	6668	6626	42	99.37%	0.63%	53344	47903	5441	7.29%	56.48%	49.19	12.91
9	6668	6626	42	99.37%	0.63%	60012	54529	5483	8.30%	56.91%	48.62	14.58
10	6668	6630	38	99.43%	0.57%	66680	61159	5521	9.31%	57.31%	48.00	16.24
11	6668	6629	39	99.42%	0.58%	73348	67788	5560	10.32%	57.71%	47.40	17.87
12	6668	6615	53	99.21%	0.79%	80016	74403	5613	11.32%	58.26%	46.94	19.43
13	6668	6620	48	99.28%	0.72%	86684	81023	5661	12.33%	58.76%	46.43	20.98
14	6668	6622	46	99.31%	0.69%	93352	87645	5707	13.34%	59.24%	45.90	22.51
15	6668	6625	43	99.36%	0.64%	100020	94270	5750	14.34%	59.68%	45.34	24.03
16	6668	6614	54	99.19%	0.81%	106688	100884	5804	15.35%	60.24%	44.89	25.48
17	6668	6623	45	99.33%	0.67%	113356	107507	5849	16.36%	60.71%	44.35	26.95
18	6668	6631	37	99.45%	0.55%	120024	114138	5886	17.37%	61.10%	43.73	28.43
19	6668	6628	40	99.40%	0.60%	126692	120766	5926	18.38%	61.51%	43.13	29.88
20	6669	6615	54	99.19%	0.81%	133361	127381	5980	19.38%	62.07%	42.69	31.23

Testing	#Records	#Goods	#Bads	Fraud Rate								
	166702	164329	2373	0.014234982								
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total #Records	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	1667	446	1221	26.75%	73.25%	1667	446	1221	0.27%	51.45%	51.18	0.53
2	1667	1630	37	97.78%	2.22%	3334	2076	1258	1.26%	53.01%	51.75	2.38
3	1667	1654	13	99.22%	0.78%	5001	3730	1271	2.27%	53.56%	51.29	4.24
4	1667	1656	11	99.34%	0.66%	6668	5386	1282	3.28%	54.02%	50.75	6.07
5	1667	1656	11	99.34%	0.66%	8335	7042	1293	4.29%	54.49%	50.2	7.86
6	1667	1653	14	99.16%	0.84%	10002	8695	1307	5.29%	55.08%	49.79	9.61
7	1667	1651	16	99.04%	0.96%	11669	10346	1323	6.30%	55.75%	49.46	11.29
8	1667	1656	11	99.34%	0.66%	13336	12002	1334	7.30%	56.22%	48.91	12.99
9	1667	1659	8	99.52%	0.48%	15003	13661	1342	8.31%	56.55%	48.24	14.7
10	1667	1657	10	99.40%	0.60%	16670	15318	1352	9.32%	56.97%	47.65	16.36
11	1667	1658	9	99.46%	0.54%	18337	16976	1361	10.33%	57.35%	47.02	18.01
12	1667	1656	11	99.34%	0.66%	20004	18632	1372	11.34%	57.82%	46.48	19.61
13	1667	1651	16	99.04%	0.96%	21671	20283	1388	12.34%	58.49%	46.15	21.1
14	1667	1657	10	99.40%	0.60%	23338	21940	1398	13.35%	58.91%	45.56	22.66
15	1667	1659	8	99.52%	0.48%	25005	23599	1406	14.36%	59.25%	44.89	24.24
16	1667	1657	10	99.40%	0.60%	26672	25256	1416	15.37%	59.67%	44.3	25.76
17	1667	1650	17	98.98%	1.02%	28339	26906	1433	16.37%	60.39%	44.01	27.11
18	1667	1654	13	99.22%	0.78%	30006	28560	1446	17.38%	60.94%	43.56	28.52
19	1667	1652	15	99.10%	0.90%	31673	30212	1461	18.39%	61.57%	43.18	29.86
20	1667	1657	10	99.40%	0.60%	33340	31869	1471	19.39%	61.99%	42.6	31.29

OOT	#Records	#Goods	#Bads	Fraud Rate								
	166493	164107	2386	0.014330933								
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total #Records	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	1664	545	1119	32.75%	67.25%	1664	545	1119	0.33%	46.90%	46.57	0.71
2	1665	1560	105	93.69%	6.31%	3329	2105	1224	1.28%	51.30%	50.02	2.50
3	1665	1640	25	98.50%	1.50%	4994	3745	1249	2.28%	52.35%	50.06	4.36
4	1665	1648	17	98.98%	1.02%	6659	5393	1266	3.29%	53.06%	49.77	6.19
5	1665	1653	12	99.28%	0.72%	8324	7046	1278	4.29%	53.56%	49.27	8.02
6	1665	1654	11	99.34%	0.66%	9989	8700	1289	5.30%	54.02%	48.72	9.81
7	1665	1651	14	99.16%	0.84%	11654	10351	1303	6.31%	54.61%	48.30	11.55
8	1665	1653	12	99.28%	0.72%	13319	12004	1315	7.31%	55.11%	47.80	13.27
9	1665	1657	8	99.52%	0.48%	14984	13661	1323	8.32%	55.45%	47.12	15.01
10	1665	1651	14	99.16%	0.84%	16649	15312	1337	9.33%	56.04%	46.70	16.65
11	1665	1657	8	99.52%	0.48%	18314	16969	1345	10.34%	56.37%	46.03	18.34
12	1665	1652	13	99.22%	0.78%	19979	18621	1358	11.35%	56.92%	45.57	19.94
13	1665	1650	15	99.10%	0.90%	21644	20271	1373	12.35%	57.54%	45.19	21.47
14	1665	1657	8	99.52%	0.48%	23309	21928	1381	13.36%	57.88%	44.52	23.09
15	1664	1653	11	99.34%	0.66%	24973	23581	1392	14.37%	58.34%	43.97	24.63
16	1665	1655	10	99.40%	0.60%	26638	25236	1402	15.38%	58.76%	43.38	26.17
17	1665	1653	12	99.28%	0.72%	28303	26889	1414	16.39%	59.26%	42.88	27.65
18	1665	1655	10	99.40%	0.60%	29968	28544	1424	17.39%	59.68%	42.29	29.14
19	1665	1657	8	99.52%	0.48%	31633	30201	1432	18.40%	60.02%	41.61	30.66
20	1665	1657	8	99.52%	0.48%	33298	31858	1440	19.41%	60.35%	40.94	32.17

We also checked the feature importance of our best Random Forest model and the results are shown in the following table.

Variable	importance
address-zip5_count30_date	0.599635
dob-name_count30_date	0.313043
address-zip5_pastday	0.019255
ssn-dob_pastday	0.009939
ssn_unique_address-zip5	0.009816
ssn-name_count7_date	0.008048
ssn-name_pastday	0.008038
ssn_count180_date	0.007251
dob-name_pastday	0.006639
address-homephone_pastday	0.004204
ssn-name_count180_date	0.002636
address-zip5_0_count180_count_ratio	0.002245
address-zip5_0_count7_count_ratio	0.001981
homephone-zip5_count30_date	0.001948
address-homephone-zip5_pastday	0.001802
address-zip5_0_count30_count_ratio	0.001720
address-homephone-zip5_count7_date	0.000699
address-homephone_count180_date	0.000669
address-homephone-zip5_count180_date	0.000422

8. Conclusions

For this project, we built supervised machine learning models for detecting fraud applications in a real time problem. We started from a data quality report which helped us get better understanding about our data. Then after data cleaning, we built 605 candidate variables based on different entities and combination groups. Next, we used several feature selection methods to narrow down the number of input variables for model building. Among all the machine learning models we trained, random forest performed the best with fraud detection rate 0.525985 for OOT dataset. We had three parts for our conclusions: business application, feature importance and future improvement.

8.1 Business application

After getting the result tables, we can try two kinds of analysis: finding the optimal model cutoff point, and speculating the ratio of gain in one fraud caught and loss in one false positive.

Finding the optimal cutoff point: Based on the statistics of our project, we assume that each fraud caught will help product providers avoid losing 7000$, while each false positive detection will cause 100$ loss. Then looking back to the OOT result table: when the population bin% is smaller than or equal to 3, the loss avoided will be larger than or equal to 175000$, the loss caused will be less than or equal to 164000$, which means bringing profits; when the population bin% is larger than or equal to 4, the loss avoided will be less than or equal to 119000$, the loss caused will be larger than or equal to 164800$, which means making losses. Thus, the optimal cutoff point is 3%, which ensures the maximum profits.

This analysis is suitable for the relatively new-born business and helps the product provider and analysts to find the optimal fraud cutoff point (as expert point for future use) at the very beginning stage. Also, for the grown-up business, if facing the rapid value variation for fraud caught and false positive detection, they could apply this analysis to alter the expert point for fraud detection.

Speculating the ratio of gain in one fraud caught and loss in one false positive: based on our OOT table statistics, we assume that the optimal fraud cutoff point is 3%, which is given by the business expert. Then we know that the product provider gets maximum profits at 3% and makes losses when altering the point to 4%. Thus, we could speculate the ratio of gain in one fraud caught and loss in one false positive ranges from 65.6 (1640/25) to 96.9 (1648/17). This ratio tells that if the loss caused by one false positive is

100$, then the loss avoided by one fraud caught in this business might range from 6560$ to 9690$.

This analysis could be applied by analysts on the basis of knowing the expert cutoff point, which could be treated as an indicator of the relationship between fraud caught value and false positive detection value in the specific business.

8.2 Feature importance

From our best random forest model, we extracted the feature importance for the 19 variables used. As we could see, the top 2 most important variables are "address-zip5_count30_date" and "dob-name_count30_date" with corresponding importance 0.5996 and 0.3130. These top 2 features were much stronger predictors than others. From another perspective, these top 2 features were built from small entity groups which consists of 2 basic entities. "Address-zip5" represents the full address for an applicant and "dob-name" represents the simplest full information which can be used to identify a person. We could imply from this result that fraudsters will not frequently use a specific more complicated entities combination but use the basic entities more frequently, which was a smart choice and might be less likely to be detected. If an applicant frequently uses his or her address and name information during the past one month, then the related applications may be more risky than others.

What's more, we can see that "ssn_unique_address-zip5" is among the top 5 most important variables. Then we could come to the conclusion that our unique count variables did provide useful information when detecting fraud. If a certain SSN was used with many different addresses, this SSN could be marked as risky and related applications could be a fraud.

8.3 Future improvement

8.3.1 Time ambiguity

In this project we assumed that record number represented time sequence and larger record happened after smaller record. If we could get the exact time of day for each application, we would build more accurate variables and hence would improve the model accuracy.

8.3.2 Finite dataset

For days since variables, we assigned the value "365" for those entities which appeared for the first time. This was an issue for finite dataset and if we could get more historical information, days since variables could catch more information and get more predictive power.

8.3.3 Computing cost for feature selection

In this step, we used recursive feature selection which used the backward stepwise algorithm. When applying this method in cross validation, it took a huge amount of time to start from 80 candidate variables. If we could try to apply other algorithms and save time in feature selection, then we could improve our efficiency and find a better candidate variables list. Additionally, our results showed that the random forest is the best model, but we used logistic regression as the estimator in feature selection when considering the time consuming. If we could use a more complicated model as the estimator here, the feature selection process might be more effective.

8.3.4 Computing cost for SVM

SVM's model complexity would increase quadratically with the increase of sample size. The normal sample size for the SVM model was nearly 10000, but we had 100 times this normal size for our dataset. For some kernel options, it might cost more than 80000 seconds to run an ensemble SVM classifier, let alone a single SVM classifier using all the training dataset. We might need a more efficient way to train the SVM model.

9. Appendix1: Data Quality Report

9.1 Data Overview

This dataset contains 1000000 records of applications from 2016-01-01 to 2016-12-31, and 14393 of them are fraud applications as their fraud_lable values are marked as 1. There are 10 fields in the dataset: record, date, ssn, firstname, lastname, address, zip5, dob, homephone, fraud_lable. Among all the fields, 8 fields are categorical variables, 1 field is date, and 1 field is index.

9.2 Summary of fields

Field Name	# records that have a value	% populated	# unique values	most common field value	# most common value
Record	1000000	100.00%	1000000	/	/
Date	1000000	100.00%	365	20160816	2877
SSN	1000000	100.00%	835819	999999999	16935
firstname	1000000	100.00%	78136	EAMSTRMT	12658
lastname	1000000	100.00%	177001	ERJSAXA	8580
Address	1000000	100.00%	828774	123 MAIN ST	1079
Zip5	1000000	100.00%	26370	68138	823
Dob	1000000	100.00%	42673	19070626	126568
homephone	1000000	100.00%	28244	9999999999	78512
fraud_label	1000000	100.00%	2	0	985607

9.3 Distribution of fields

9.3.1 record

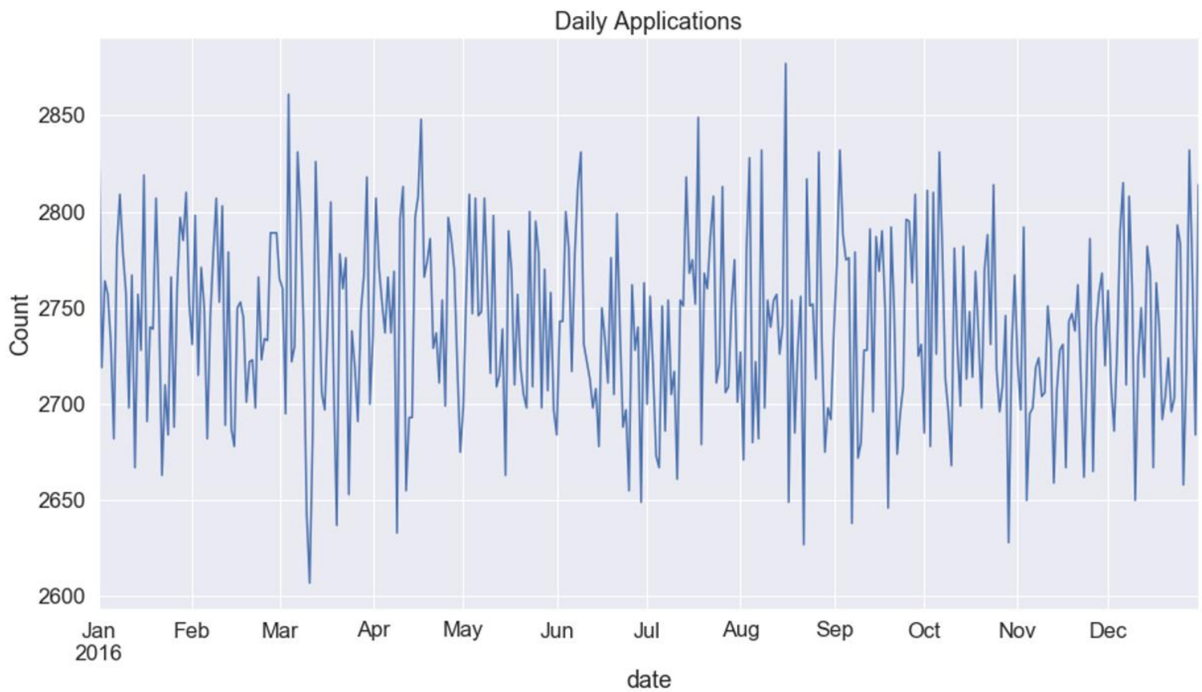
Record represents the index of this dataset, ranging from 1 to 1000000.

9.3.2 fraud_label

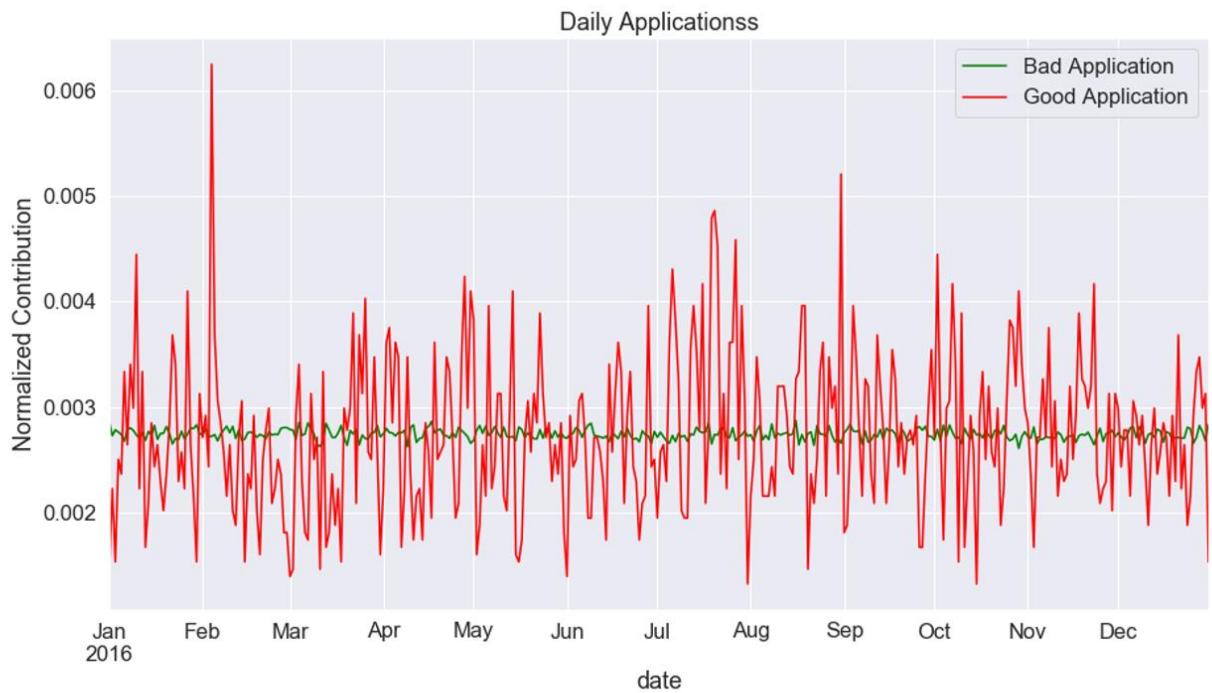
This field represents the fraud label for each application. Among all the 1 million application records, 14393 records are fraud applications and 985607 records are good applications. The percentage of fraud is about 1.46%, which points that this dataset is a very unbalanced dataset.

9.3.3 date

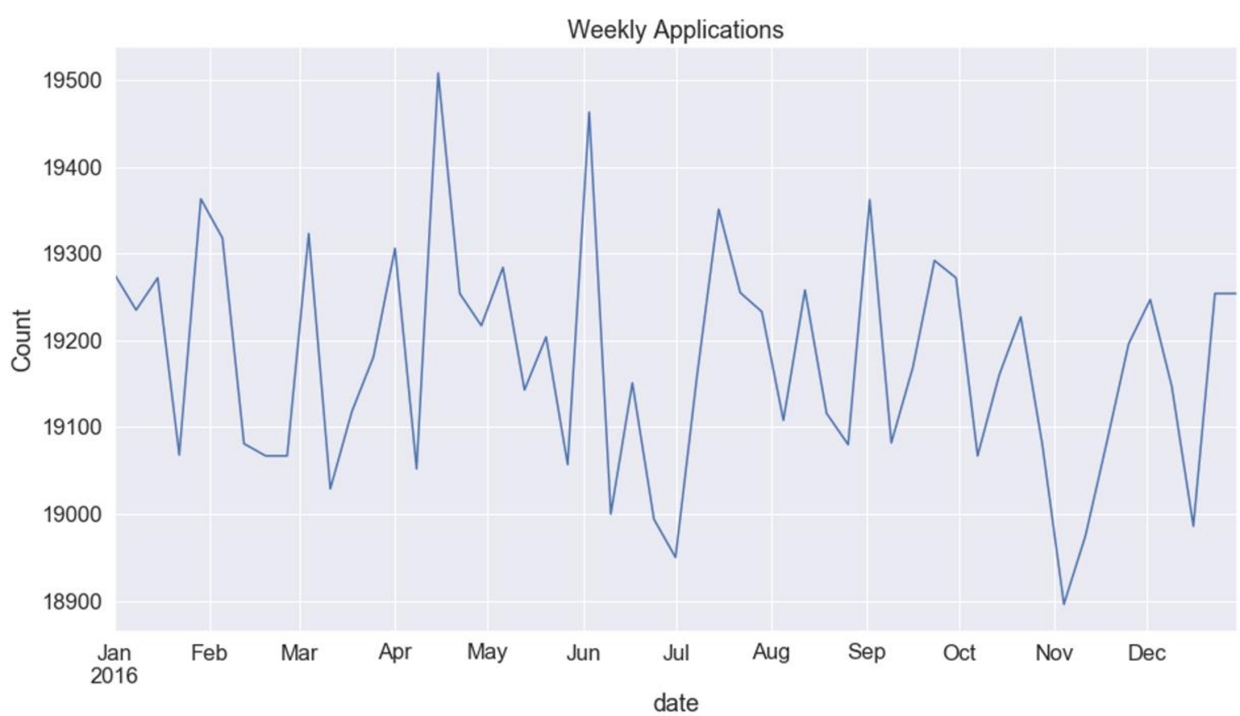
This field represents the date of each product application, ranging from 2016-01-01 to 2016-12-31. The following picture shows the count of daily applications, we can see that the applications are randomly distributed and have no seasonal patterns.



The following graph shows the count of daily fraud applications (bad application with green line) and good applications (red line). We can see that the distribution of fraud applications is kind of stable.

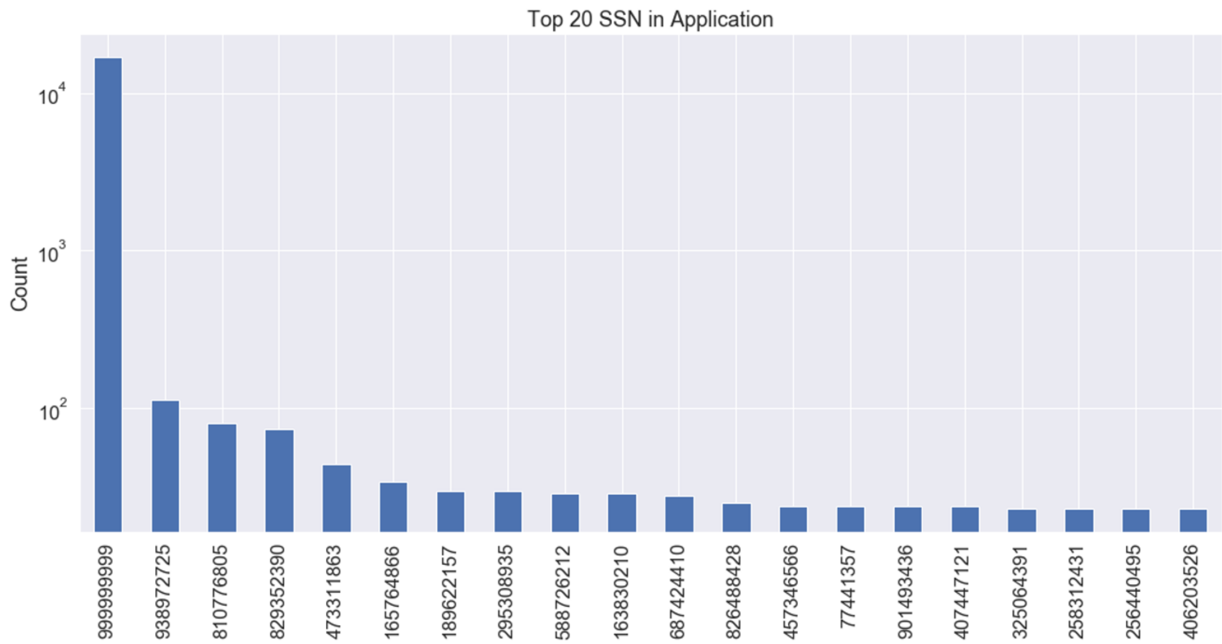


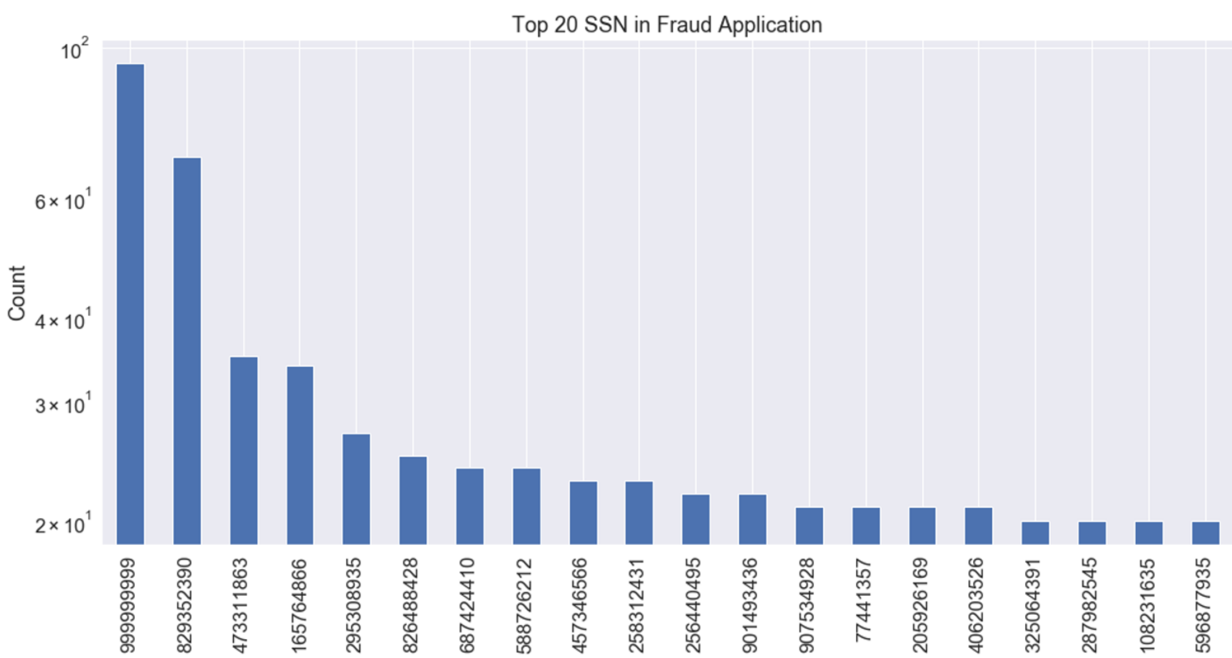
The following graph shows weekly distribution of daily applications count, which also does not have seasonal patterns.



9.3.4 ssn

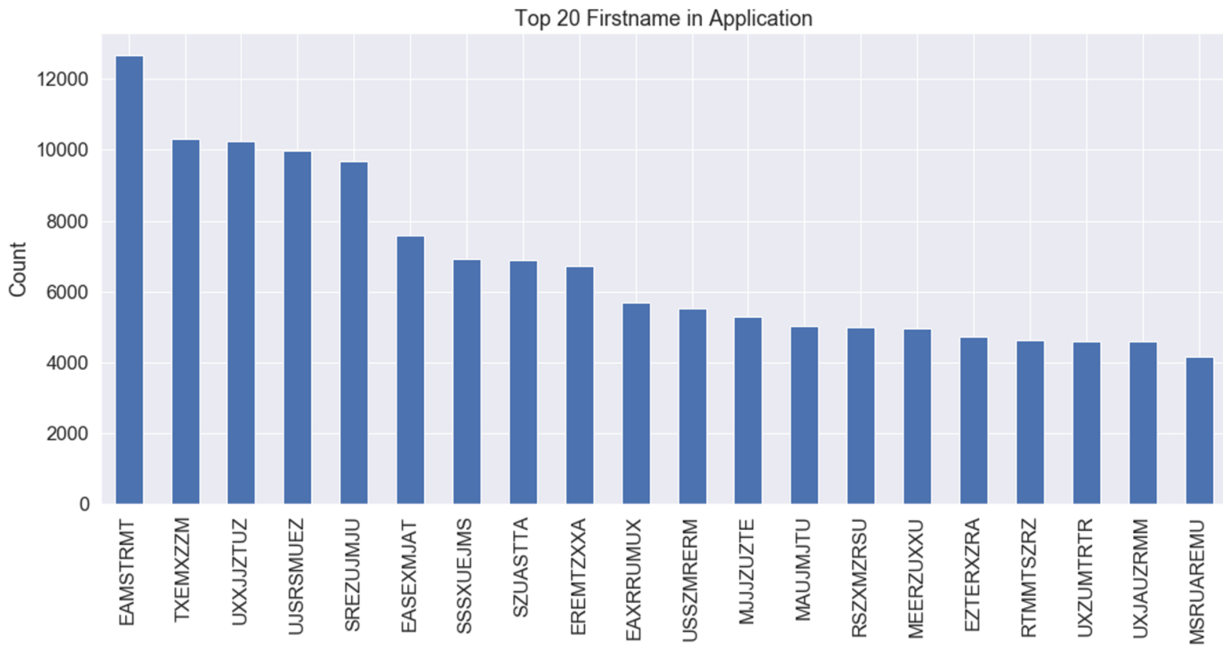
Ssn represents the social security number for each applicant. This field has 835819 unique values. The following two graphs show the top 20 SSN in all applications and fraud applications. 999999999 is the most common ssn, which may indicate that 999999999 is a frivolous value.

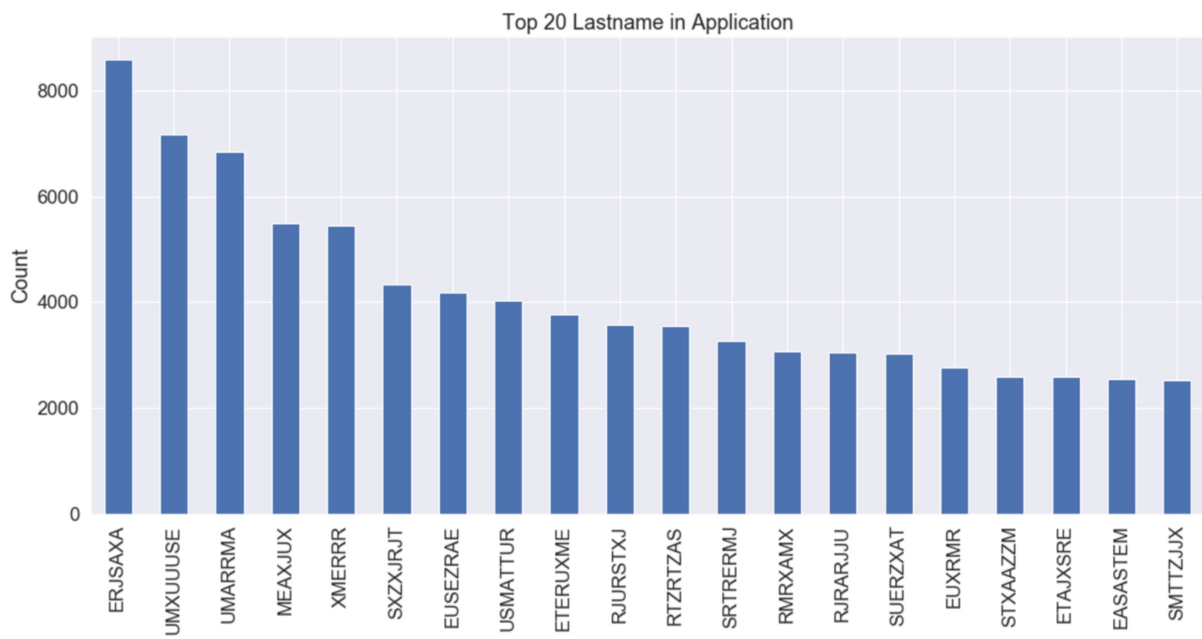




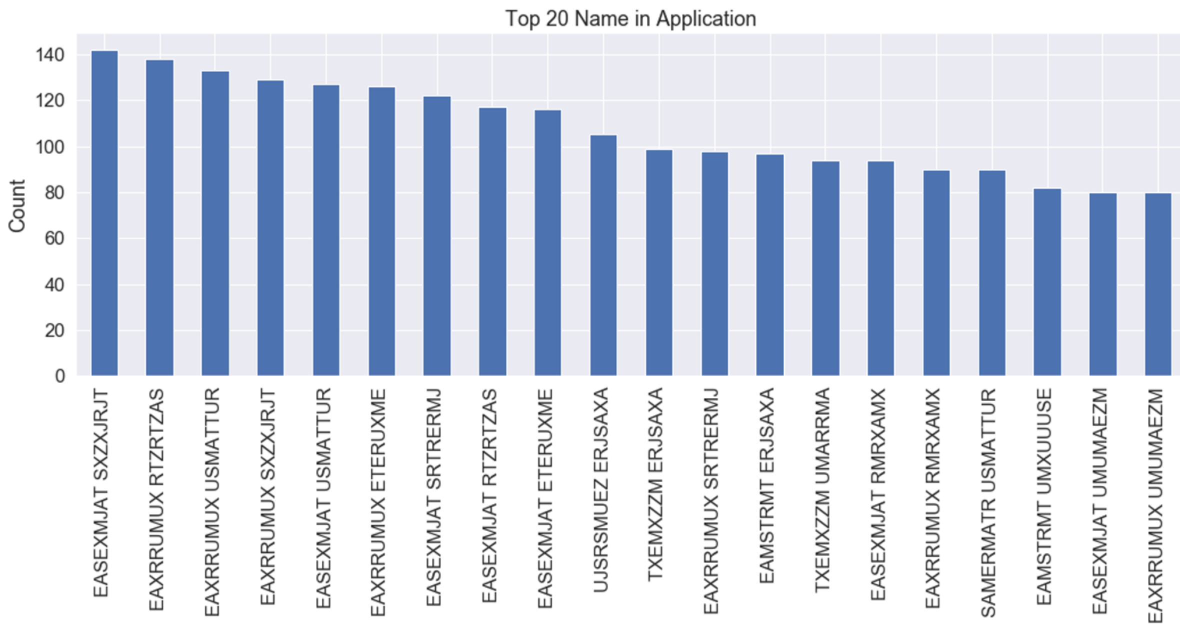
9.3.5 firstname and lastname

The field firstname has 78136 unique values and lastname has 177001 unique values. The following two graphs show the top 20 firstname and lastname among all the applications. It is obvious that 'EAMSTRMT' is the most popular first name and 'ERJSAXA' is the most popular last name.



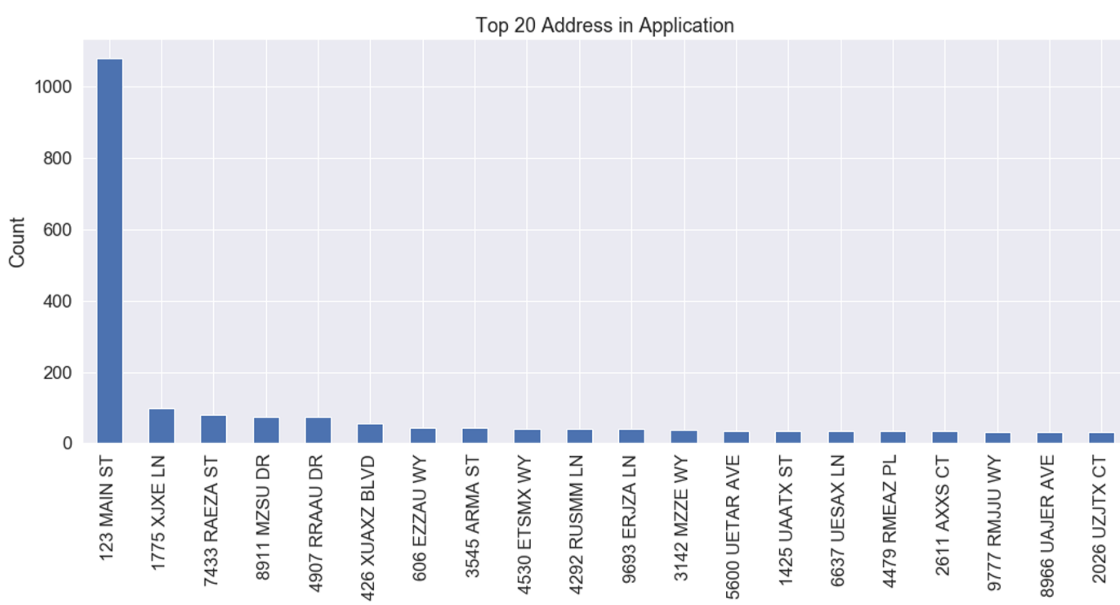


When combining first name and last name together, we can get the field ‘name’. This name field has 717126 unique values. The following graph shows the top 20 full names among all the applications. ‘EAMSTRMT SXZXJRJT’ is the most popular full name.



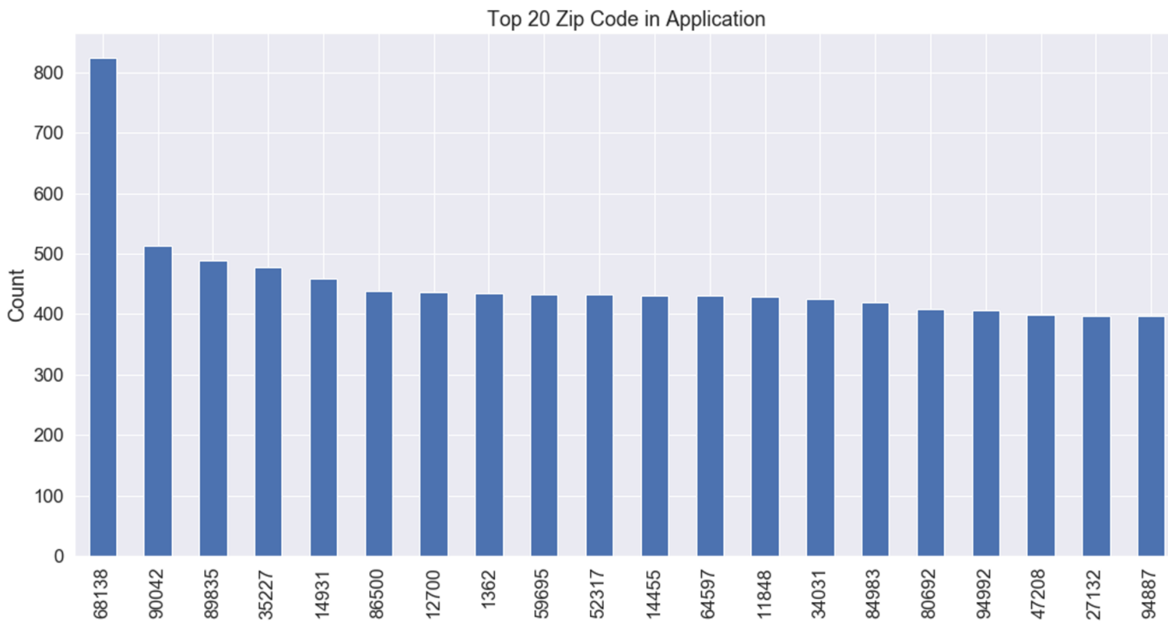
9.3.6 address

This field represents the address of each applicant and has 828774 unique values. The following graph shows the top 20 addresses among all applications. ‘123 MAIN ST’ is the most common address, which may indicate that ‘123 MAIN ST’ is a frivolous value.



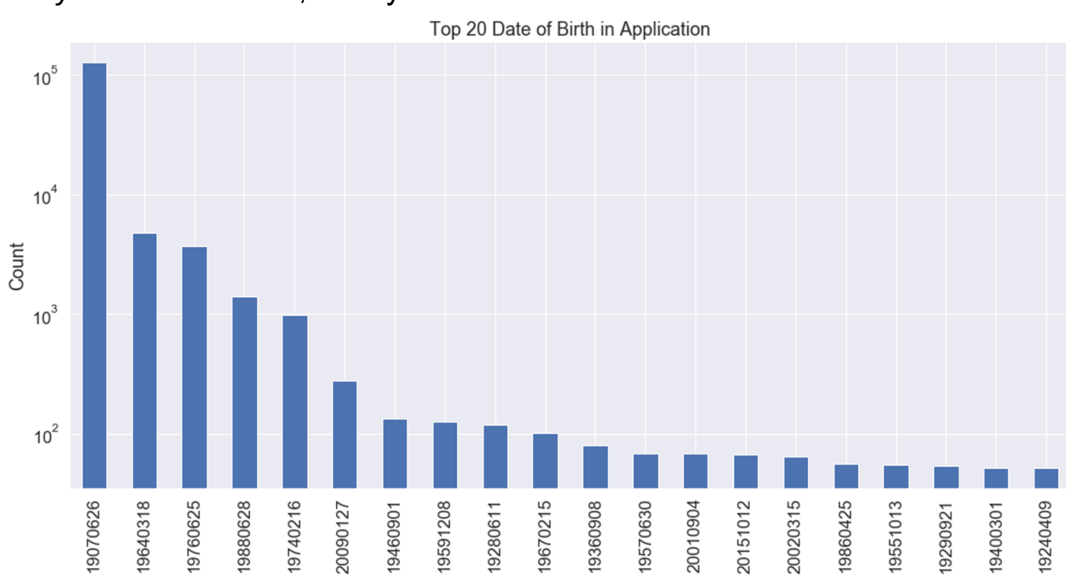
9.3.7 zip5

This field represents the zip code of each applicant and has 26370 unique values. The following graph shows the top 20 zip codes among all the applications.



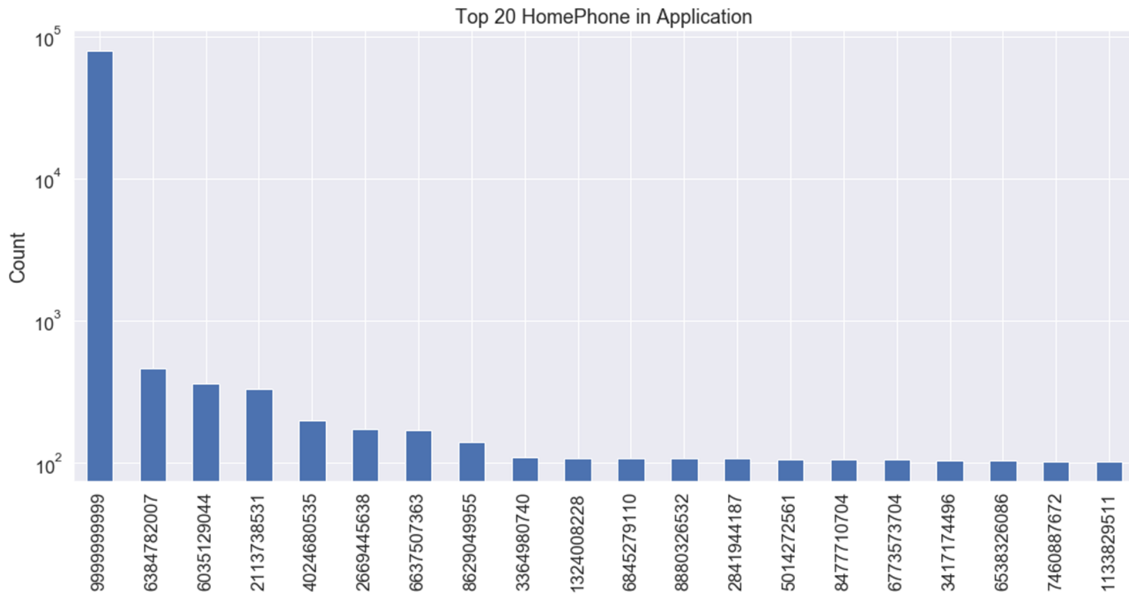
9.3.8 dob

This field represents the date of birth of each applicant and has 42673 unique values. The following graph shows the top 20 date of birth among all applications, and ‘19070626’ is the most common date of birth. Since the person who was born in 1907 would be more than 100 years old in 2016, it may indicate that ‘19070626’ is a frivolous value.



9.3.9 homephone

This field represents the home phone number of each applicant and has 28244 unique values. '9999999999' is the most common home phone number, which may indicate that '9999999999' is a frivolous value.



10. Appendix2: List of All Variables

Variables Name	
ssn_count0_date	address_0_count7_count_ratio
ssn_count1_date	address_0_count30_count_ratio
ssn_count3_date	address_0_count90_count_ratio
ssn_count7_date	address_0_count180_count_ratio
ssn_count30_date	homephone_0_count1_count_ratio
ssn_count90_date	homephone_0_count3_count_ratio
ssn_count180_date	homephone_0_count7_count_ratio
address_count0_date	homephone_0_count30_count_ratio
address_count1_date	homephone_0_count90_count_ratio
address_count3_date	homephone_0_count180_count_ratio
address_count7_date	name_0_count1_count_ratio
address_count30_date	name_0_count3_count_ratio
address_count90_date	name_0_count7_count_ratio
address_count180_date	name_0_count30_count_ratio
homephone_count0_date	name_0_count90_count_ratio
homephone_count1_date	name_0_count180_count_ratio
homephone_count3_date	ssn-dob_0_count1_count_ratio
homephone_count7_date	ssn-dob_0_count3_count_ratio
homephone_count30_date	ssn-dob_0_count7_count_ratio
homephone_count90_date	ssn-dob_0_count30_count_ratio
homephone_count180_date	ssn-dob_0_count90_count_ratio
name_count0_date	ssn-dob_0_count180_count_ratio
name_count1_date	ssn-address_0_count1_count_ratio
name_count3_date	ssn-address_0_count3_count_ratio
name_count7_date	ssn-address_0_count7_count_ratio
name_count30_date	ssn-address_0_count30_count_ratio
name_count90_date	ssn-address_0_count90_count_ratio
name_count180_date	ssn-address_0_count180_count_ratio
ssn-dob_count0_date	ssn-homephone_0_count1_count_ratio
ssn-dob_count1_date	ssn-homephone_0_count3_count_ratio
ssn-dob_count3_date	ssn-homephone_0_count7_count_ratio
ssn-dob_count7_date	ssn-homephone_0_count30_count_ratio
ssn-dob_count30_date	ssn-homephone_0_count90_count_ratio
ssn-dob_count90_date	ssn-homephone_0_count180_count_ratio
ssn-dob_count180_date	ssn-name_0_count1_count_ratio
ssn-address_count0_date	ssn-name_0_count3_count_ratio
ssn-address_count1_date	ssn-name_0_count7_count_ratio
ssn-address_count3_date	ssn-name_0_count30_count_ratio
ssn-address_count7_date	ssn-name_0_count90_count_ratio
ssn-address_count30_date	ssn-name_0_count180_count_ratio
ssn-address_count90_date	ssn-zip5_0_count1_count_ratio
ssn-address_count180_date	ssn-zip5_0_count3_count_ratio
ssn-homephone_count0_date	ssn-zip5_0_count7_count_ratio
ssn-homephone_count1_date	ssn-zip5_0_count30_count_ratio
ssn-homephone_count3_date	ssn-zip5_0_count90_count_ratio
ssn-homephone_count7_date	ssn-zip5_0_count180_count_ratio
ssn-homephone_count30_date	dob-address_0_count1_count_ratio
ssn-homephone_count90_date	dob-address_0_count3_count_ratio
ssn-homephone_count180_date	dob-address_0_count7_count_ratio

ssn-name_count0_date	dob-address_0_count30_count_ratio
ssn-name_count1_date	dob-address_0_count90_count_ratio
ssn-name_count3_date	dob-address_0_count180_count_ratio
ssn-name_count7_date	dob-homephone_0_count1_count_ratio
ssn-name_count30_date	dob-homephone_0_count3_count_ratio
ssn-name_count90_date	dob-homephone_0_count7_count_ratio
ssn-name_count180_date	dob-homephone_0_count30_count_ratio
ssn-zip5_count0_date	dob-homephone_0_count90_count_ratio
ssn-zip5_count1_date	dob-homephone_0_count180_count_ratio
ssn-zip5_count3_date	dob-name_0_count1_count_ratio
ssn-zip5_count7_date	dob-name_0_count3_count_ratio
ssn-zip5_count30_date	dob-name_0_count7_count_ratio
ssn-zip5_count90_date	dob-name_0_count30_count_ratio
ssn-zip5_count180_date	dob-name_0_count90_count_ratio
dob-address_count0_date	dob-name_0_count180_count_ratio
dob-address_count1_date	dob-zip5_0_count1_count_ratio
dob-address_count3_date	dob-zip5_0_count3_count_ratio
dob-address_count7_date	dob-zip5_0_count7_count_ratio
dob-address_count30_date	dob-zip5_0_count30_count_ratio
dob-address_count90_date	dob-zip5_0_count90_count_ratio
dob-address_count180_date	dob-zip5_0_count180_count_ratio
dob-homephone_count0_date	address-homephone_0_count1_count_ratio
dob-homephone_count1_date	address-homephone_0_count3_count_ratio
dob-homephone_count3_date	address-homephone_0_count7_count_ratio
dob-homephone_count7_date	address-homephone_0_count30_count_ratio
dob-homephone_count30_date	address-homephone_0_count90_count_ratio
dob-homephone_count90_date	address-homephone_0_count180_count_ratio
dob-homephone_count180_date	address-name_0_count1_count_ratio
dob-name_count0_date	address-name_0_count3_count_ratio
dob-name_count1_date	address-name_0_count7_count_ratio
dob-name_count3_date	address-name_0_count30_count_ratio
dob-name_count7_date	address-name_0_count90_count_ratio
dob-name_count30_date	address-name_0_count180_count_ratio
dob-name_count90_date	address-zip5_0_count1_count_ratio
dob-name_count180_date	address-zip5_0_count3_count_ratio
dob-zip5_count0_date	address-zip5_0_count7_count_ratio
dob-zip5_count1_date	address-zip5_0_count30_count_ratio
dob-zip5_count3_date	address-zip5_0_count90_count_ratio
dob-zip5_count7_date	address-zip5_0_count180_count_ratio
dob-zip5_count30_date	homephone-name_0_count1_count_ratio
dob-zip5_count90_date	homephone-name_0_count3_count_ratio
dob-zip5_count180_date	homephone-name_0_count7_count_ratio
address-homephone_count0_date	homephone-name_0_count30_count_ratio
address-homephone_count1_date	homephone-name_0_count90_count_ratio
address-homephone_count3_date	homephone-name_0_count180_count_ratio
address-homephone_count7_date	homephone-zip5_0_count1_count_ratio
address-homephone_count30_date	homephone-zip5_0_count3_count_ratio
address-homephone_count90_date	homephone-zip5_0_count7_count_ratio
address-homephone_count180_date	homephone-zip5_0_count30_count_ratio
address-name_count0_date	homephone-zip5_0_count90_count_ratio

address-name_count1_date	homephone-zip5_0_count180_count_ratio
address-name_count3_date	name-zip5_0_count1_count_ratio
address-name_count7_date	name-zip5_0_count3_count_ratio
address-name_count30_date	name-zip5_0_count7_count_ratio
address-name_count90_date	name-zip5_0_count30_count_ratio
address-name_count180_date	name-zip5_0_count90_count_ratio
address-zip5_count0_date	name-zip5_0_count180_count_ratio
address-zip5_count1_date	ssn-dob-address_0_count1_count_ratio
address-zip5_count3_date	ssn-dob-address_0_count3_count_ratio
address-zip5_count7_date	ssn-dob-address_0_count7_count_ratio
address-zip5_count30_date	ssn-dob-address_0_count30_count_ratio
address-zip5_count90_date	ssn-dob-address_0_count90_count_ratio
address-zip5_count180_date	ssn-dob-address_0_count180_count_ratio
homephone-name_count0_date	ssn-dob-homephone_0_count1_count_ratio
homephone-name_count1_date	ssn-dob-homephone_0_count3_count_ratio
homephone-name_count3_date	ssn-dob-homephone_0_count7_count_ratio
homephone-name_count7_date	ssn-dob-homephone_0_count30_count_ratio
homephone-name_count30_date	ssn-dob-homephone_0_count90_count_ratio
homephone-name_count90_date	ssn-dob-homephone_0_count180_count_ratio
homephone-name_count180_date	ssn-dob-name_0_count1_count_ratio
homephone-zip5_count0_date	ssn-dob-name_0_count3_count_ratio
homephone-zip5_count1_date	ssn-dob-name_0_count7_count_ratio
homephone-zip5_count3_date	ssn-dob-name_0_count30_count_ratio
homephone-zip5_count7_date	ssn-dob-name_0_count90_count_ratio
homephone-zip5_count30_date	ssn-dob-name_0_count180_count_ratio
homephone-zip5_count90_date	ssn-dob-zip5_0_count1_count_ratio
homephone-zip5_count180_date	ssn-dob-zip5_0_count3_count_ratio
name-zip5_count0_date	ssn-dob-zip5_0_count7_count_ratio
name-zip5_count1_date	ssn-dob-zip5_0_count30_count_ratio
name-zip5_count3_date	ssn-dob-zip5_0_count90_count_ratio
name-zip5_count7_date	ssn-dob-zip5_0_count180_count_ratio
name-zip5_count30_date	ssn-address-homephone_0_count1_count_ratio
name-zip5_count90_date	ssn-address-homephone_0_count3_count_ratio
name-zip5_count180_date	ssn-address-homephone_0_count7_count_ratio
ssn-dob-address_count0_date	ssn-address-homephone_0_count30_count_ratio
ssn-dob-address_count1_date	ssn-address-homephone_0_count90_count_ratio
ssn-dob-address_count3_date	ssn-address-homephone_0_count180_count_ratio
ssn-dob-address_count7_date	ssn-address-name_0_count1_count_ratio
ssn-dob-address_count30_date	ssn-address-name_0_count3_count_ratio
ssn-dob-address_count90_date	ssn-address-name_0_count7_count_ratio
ssn-dob-address_count180_date	ssn-address-name_0_count30_count_ratio
ssn-dob-homephone_count0_date	ssn-address-name_0_count90_count_ratio
ssn-dob-homephone_count1_date	ssn-address-name_0_count180_count_ratio
ssn-dob-homephone_count3_date	ssn-address-zip5_0_count1_count_ratio
ssn-dob-homephone_count7_date	ssn-address-zip5_0_count3_count_ratio
ssn-dob-homephone_count30_date	ssn-address-zip5_0_count7_count_ratio
ssn-dob-homephone_count90_date	ssn-address-zip5_0_count30_count_ratio
ssn-dob-homephone_count180_date	ssn-address-zip5_0_count90_count_ratio
ssn-dob-name_count0_date	ssn-address-zip5_0_count180_count_ratio
ssn-dob-name_count1_date	ssn-homephone-name_0_count1_count_ratio

ssn-dob-name_count3_date	ssn-homephone-name_0_count3_count_ratio
ssn-dob-name_count7_date	ssn-homephone-name_0_count7_count_ratio
ssn-dob-name_count30_date	ssn-homephone-name_0_count30_count_ratio
ssn-dob-name_count90_date	ssn-homephone-name_0_count90_count_ratio
ssn-dob-name_count180_date	ssn-homephone-name_0_count180_count_ratio
ssn-dob-zip5_count0_date	ssn-homephone-zip5_0_count1_count_ratio
ssn-dob-zip5_count1_date	ssn-homephone-zip5_0_count3_count_ratio
ssn-dob-zip5_count3_date	ssn-homephone-zip5_0_count7_count_ratio
ssn-dob-zip5_count7_date	ssn-homephone-zip5_0_count30_count_ratio
ssn-dob-zip5_count30_date	ssn-homephone-zip5_0_count90_count_ratio
ssn-dob-zip5_count90_date	ssn-homephone-zip5_0_count180_count_ratio
ssn-dob-zip5_count180_date	ssn-name-zip5_0_count1_count_ratio
ssn-address-homephone_count0_date	ssn-name-zip5_0_count3_count_ratio
ssn-address-homephone_count1_date	ssn-name-zip5_0_count7_count_ratio
ssn-address-homephone_count3_date	ssn-name-zip5_0_count30_count_ratio
ssn-address-homephone_count7_date	ssn-name-zip5_0_count90_count_ratio
ssn-address-homephone_count30_date	ssn-name-zip5_0_count180_count_ratio
ssn-address-homephone_count90_date	dob-address-homephone_0_count1_count_ratio
ssn-address-homephone_count180_date	dob-address-homephone_0_count3_count_ratio
ssn-address-name_count0_date	dob-address-homephone_0_count7_count_ratio
ssn-address-name_count1_date	dob-address-homephone_0_count30_count_ratio
ssn-address-name_count3_date	dob-address-homephone_0_count90_count_ratio
ssn-address-name_count7_date	dob-address-homephone_0_count180_count_ratio
ssn-address-name_count30_date	dob-address-name_0_count1_count_ratio
ssn-address-name_count90_date	dob-address-name_0_count3_count_ratio
ssn-address-name_count180_date	dob-address-name_0_count7_count_ratio
ssn-address-zip5_count0_date	dob-address-name_0_count30_count_ratio
ssn-address-zip5_count1_date	dob-address-name_0_count90_count_ratio
ssn-address-zip5_count3_date	dob-address-name_0_count180_count_ratio
ssn-address-zip5_count7_date	dob-address-zip5_0_count1_count_ratio
ssn-address-zip5_count30_date	dob-address-zip5_0_count3_count_ratio
ssn-address-zip5_count90_date	dob-address-zip5_0_count7_count_ratio
ssn-address-zip5_count180_date	dob-address-zip5_0_count30_count_ratio
ssn-homephone-name_count0_date	dob-address-zip5_0_count90_count_ratio
ssn-homephone-name_count1_date	dob-address-zip5_0_count180_count_ratio
ssn-homephone-name_count3_date	dob-homephone-name_0_count1_count_ratio
ssn-homephone-name_count7_date	dob-homephone-name_0_count3_count_ratio
ssn-homephone-name_count30_date	dob-homephone-name_0_count7_count_ratio
ssn-homephone-name_count90_date	dob-homephone-name_0_count30_count_ratio
ssn-homephone-name_count180_date	dob-homephone-name_0_count90_count_ratio
ssn-homephone-zip5_count0_date	dob-homephone-name_0_count180_count_ratio
ssn-homephone-zip5_count1_date	dob-homephone-zip5_0_count1_count_ratio
ssn-homephone-zip5_count3_date	dob-homephone-zip5_0_count3_count_ratio
ssn-homephone-zip5_count7_date	dob-homephone-zip5_0_count7_count_ratio
ssn-homephone-zip5_count30_date	dob-homephone-zip5_0_count30_count_ratio
ssn-homephone-zip5_count90_date	dob-homephone-zip5_0_count90_count_ratio
ssn-homephone-zip5_count180_date	dob-homephone-zip5_0_count180_count_ratio
ssn-name-zip5_count0_date	dob-name-zip5_0_count1_count_ratio
ssn-name-zip5_count1_date	dob-name-zip5_0_count3_count_ratio
ssn-name-zip5_count3_date	dob-name-zip5_0_count7_count_ratio

ssn-name-zip5_count7_date	dob-name-zip5_0_count30_count_ratio
ssn-name-zip5_count30_date	dob-name-zip5_0_count90_count_ratio
ssn-name-zip5_count90_date	dob-name-zip5_0_count180_count_ratio
ssn-name-zip5_count180_date	address-homephone-name_0_count1_count_ratio
dob-address-homephone_count0_date	address-homephone-name_0_count3_count_ratio
dob-address-homephone_count1_date	address-homephone-name_0_count7_count_ratio
dob-address-homephone_count3_date	address-homephone-name_0_count30_count_ratio
dob-address-homephone_count7_date	address-homephone-name_0_count90_count_ratio
dob-address-homephone_count30_date	address-homephone-name_0_count180_count_ratio
dob-address-homephone_count90_date	address-homephone-zip5_0_count1_count_ratio
dob-address-homephone_count180_date	address-homephone-zip5_0_count3_count_ratio
dob-address-name_count0_date	address-homephone-zip5_0_count7_count_ratio
dob-address-name_count1_date	address-homephone-zip5_0_count30_count_ratio
dob-address-name_count3_date	address-homephone-zip5_0_count90_count_ratio
dob-address-name_count7_date	address-homephone-zip5_0_count180_count_ratio
dob-address-name_count30_date	address-name-zip5_0_count1_count_ratio
dob-address-name_count90_date	address-name-zip5_0_count3_count_ratio
dob-address-name_count180_date	address-name-zip5_0_count7_count_ratio
dob-address-zip5_count0_date	address-name-zip5_0_count30_count_ratio
dob-address-zip5_count1_date	address-name-zip5_0_count90_count_ratio
dob-address-zip5_count3_date	address-name-zip5_0_count180_count_ratio
dob-address-zip5_count7_date	homephone-name-zip5_0_count1_count_ratio
dob-address-zip5_count30_date	homephone-name-zip5_0_count3_count_ratio
dob-address-zip5_count90_date	homephone-name-zip5_0_count7_count_ratio
dob-address-zip5_count180_date	homephone-name-zip5_0_count30_count_ratio
dob-homephone-name_count0_date	homephone-name-zip5_0_count90_count_ratio
dob-homephone-name_count1_date	homephone-name-zip5_0_count180_count_ratio
dob-homephone-name_count3_date	name-ssn-homephone-dob-address-zip5_0_count1_count_ratio
dob-homephone-name_count7_date	name-ssn-homephone-dob-address-zip5_0_count3_count_ratio
dob-homephone-name_count30_date	name-ssn-homephone-dob-address-zip5_0_count7_count_ratio
dob-homephone-name_count90_date	name-ssn-homephone-dob-address-zip5_0_count30_count_ratio
dob-homephone-name_count180_date	name-ssn-homephone-dob-address-zip5_0_count90_count_ratio
dob-homephone-zip5_count0_date	name-ssn-homephone-dob-address-zip5_0_count180_count_ratio
dob-homephone-zip5_count1_date	name-ssn-homephone-dob_0_count1_count_ratio
dob-homephone-zip5_count3_date	name-ssn-homephone-dob_0_count3_count_ratio
dob-homephone-zip5_count7_date	name-ssn-homephone-dob_0_count7_count_ratio
dob-homephone-zip5_count30_date	name-ssn-homephone-dob_0_count30_count_ratio
dob-homephone-zip5_count90_date	name-ssn-homephone-dob_0_count90_count_ratio
dob-homephone-zip5_count180_date	name-ssn-homephone-dob_0_count180_count_ratio
dob-name-zip5_count0_date	name-ssn-address-zip5_0_count1_count_ratio
dob-name-zip5_count1_date	name-ssn-address-zip5_0_count3_count_ratio
dob-name-zip5_count3_date	name-ssn-address-zip5_0_count7_count_ratio
dob-name-zip5_count7_date	name-ssn-address-zip5_0_count30_count_ratio
dob-name-zip5_count30_date	name-ssn-address-zip5_0_count90_count_ratio
dob-name-zip5_count90_date	name-ssn-address-zip5_0_count180_count_ratio
dob-name-zip5_count180_date	ssn_pastday
address-homephone-name_count0_date	address_pastday
address-homephone-name_count1_date	homephone_pastday
address-homephone-name_count3_date	name_pastday
address-homephone-name_count7_date	ssn-dob_pastday

address-homephone-name_count30_date	ssn-address_pastday
address-homephone-name_count90_date	ssn-homephone_pastday
address-homephone-name_count180_date	ssn-name_pastday
address-homephone-zip5_count0_date	ssn-zip5_pastday
address-homephone-zip5_count1_date	dob-address_pastday
address-homephone-zip5_count3_date	dob-homephone_pastday
address-homephone-zip5_count7_date	dob-name_pastday
address-homephone-zip5_count30_date	dob-zip5_pastday
address-homephone-zip5_count90_date	address-homephone_pastday
address-homephone-zip5_count180_date	address-name_pastday
address-name-zip5_count0_date	address-zip5_pastday
address-name-zip5_count1_date	homephone-name_pastday
address-name-zip5_count3_date	homephone-zip5_pastday
address-name-zip5_count7_date	name-zip5_pastday
address-name-zip5_count30_date	ssn-dob-address_pastday
address-name-zip5_count90_date	ssn-dob-homephone_pastday
address-name-zip5_count180_date	ssn-dob-name_pastday
homephone-name-zip5_count0_date	ssn-dob-zip5_pastday
homephone-name-zip5_count1_date	ssn-address-homephone_pastday
homephone-name-zip5_count3_date	ssn-address-name_pastday
homephone-name-zip5_count7_date	ssn-address-zip5_pastday
homephone-name-zip5_count30_date	ssn-homephone-name_pastday
homephone-name-zip5_count90_date	ssn-homephone-zip5_pastday
homephone-name-zip5_count180_date	ssn-name-zip5_pastday
name-ssn-homephone-dob-address-zip5_count0_date	dob-address-homephone_pastday
name-ssn-homephone-dob-address-zip5_count1_date	dob-address-name_pastday
name-ssn-homephone-dob-address-zip5_count3_date	dob-address-zip5_pastday
name-ssn-homephone-dob-address-zip5_count7_date	dob-homephone-name_pastday
name-ssn-homephone-dob-address-zip5_count30_date	dob-homephone-zip5_pastday
name-ssn-homephone-dob-address-zip5_count90_date	dob-name-zip5_pastday
name-ssn-homephone-dob-address-zip5_count180_date	address-homephone-name_pastday
name-ssn-homephone-dob_count0_date	address-homephone-zip5_pastday
name-ssn-homephone-dob_count1_date	address-name-zip5_pastday
name-ssn-homephone-dob_count3_date	homephone-name-zip5_pastday
name-ssn-homephone-dob_count7_date	name-ssn-homephone-dob-address-zip5_pastday
name-ssn-homephone-dob_count30_date	name-ssn-homephone-dob_pastday
name-ssn-homephone-dob_count90_date	name-ssn-address-zip5_pastday
name-ssn-homephone-dob_count180_date	ssn_unique_dob
name-ssn-address-zip5_count0_date	ssn_unique_name
name-ssn-address-zip5_count1_date	ssn_unique_homephone
name-ssn-address-zip5_count3_date	ssn_unique_address-zip5
name-ssn-address-zip5_count7_date	name_unique_ssn
name-ssn-address-zip5_count30_date	name_unique_dob
name-ssn-address-zip5_count90_date	name_unique_homephone
name-ssn-address-zip5_count180_date	name_unique_address-zip5
ssn_0_count1_count_ratio	homephone_unique_ssn
ssn_0_count3_count_ratio	homephone_unique_dob
ssn_0_count7_count_ratio	homephone_unique_name
ssn_0_count30_count_ratio	homephone_unique_address-zip5
ssn_0_count90_count_ratio	address-zip5_unique_ssn
ssn_0_count180_count_ratio	address-zip5_unique_dob
address_0_count1_count_ratio	address-zip5_unique_name
address_0_count3_count_ratio	address-zip5_unique_homephone
	day_week_risk

11. Appendix3: Full Results Tables

Training	#Records	#Goods	#Bads	raud Rate								
	666805	657171	9634	0.014448								
		Bin Statistics						Cumulative Statistics				
ulation B	#Records	#Goods	#Bads	%Goods	%Bads	tal#Recor	lative G	lative B	%Goods	Bads(FDR)	KS	FPR
1	6668	1703	4965	25.54%	74.46%	6668	1703	4965	0.26%	51.54%	51.28	0.50
2	6668	6479	189	97.17%	2.83%	13336	8182	5154	1.25%	53.50%	52.25	2.33
3	6668	6623	45	99.33%	0.67%	20004	14805	5199	2.25%	53.97%	51.71	4.17
4	6668	6625	43	99.36%	0.64%	26672	21430	5242	3.26%	54.41%	51.15	5.99
5	6668	6609	59	99.12%	0.88%	33340	28039	5301	4.27%	55.02%	50.76	7.75
6	6668	6622	46	99.31%	0.69%	40008	34661	5347	5.27%	55.50%	50.23	9.50
7	6668	6616	52	99.22%	0.78%	46676	41277	5399	6.28%	56.04%	49.76	11.21
8	6668	6626	42	99.37%	0.63%	53344	47903	5441	7.29%	56.48%	49.19	12.91
9	6668	6626	42	99.37%	0.63%	60012	54529	5483	8.30%	56.91%	48.62	14.58
10	6668	6630	38	99.43%	0.57%	66680	61159	5521	9.31%	57.31%	48.00	16.24
11	6668	6629	39	99.42%	0.58%	73348	67788	5560	10.32%	57.71%	47.40	17.87
12	6668	6615	53	99.21%	0.79%	80016	74403	5613	11.32%	58.26%	46.94	19.43
13	6668	6620	48	99.28%	0.72%	86684	81023	5661	12.33%	58.76%	46.43	20.98
14	6668	6622	46	99.31%	0.69%	93352	87645	5707	13.34%	59.24%	45.90	22.51
15	6668	6625	43	99.36%	0.64%	100020	94270	5750	14.34%	59.68%	45.34	24.03
16	6668	6614	54	99.19%	0.81%	106688	100884	5804	15.35%	60.24%	44.89	25.48
17	6668	6623	45	99.33%	0.67%	113356	107507	5849	16.36%	60.71%	44.35	26.95
18	6668	6631	37	99.45%	0.55%	120024	114138	5886	17.37%	61.10%	43.73	28.43
19	6668	6628	40	99.40%	0.60%	126692	120766	5926	18.38%	61.51%	43.13	29.88
20	6669	6615	54	99.19%	0.81%	133361	127381	5980	19.38%	62.07%	42.69	31.23
21	6668	6621	47	99.30%	0.70%	140029	134002	6027	20.39%	62.56%	42.17	32.59
22	6668	6620	48	99.28%	0.72%	146697	140622	6075	21.40%	63.06%	41.66	33.93
23	6668	6622	46	99.31%	0.69%	153365	147244	6121	22.41%	63.54%	41.13	35.26
24	6668	6614	54	99.19%	0.81%	160033	153858	6175	23.41%	64.10%	40.68	36.53
25	6668	6647	21	99.69%	0.31%	166701	160505	6196	24.42%	64.31%	39.89	37.98
26	6668	6631	37	99.45%	0.55%	173369	167136	6233	25.43%	64.70%	39.27	39.31
27	6668	6620	48	99.28%	0.72%	180037	173756	6281	26.44%	65.20%	38.76	40.55
28	6668	6614	54	99.19%	0.81%	186705	180370	6335	27.45%	65.76%	38.31	41.74
29	6668	6635	33	99.51%	0.49%	193373	187005	6368	28.46%	66.10%	37.64	43.05
30	6668	6622	46	99.31%	0.69%	200041	193627	6414	29.46%	66.58%	37.11	44.26
31	6668	6611	57	99.15%	0.85%	206709	200238	6471	30.47%	67.17%	36.70	45.36
32	6668	6622	46	99.31%	0.69%	213377	206860	6517	31.48%	67.65%	36.17	46.53
33	6668	6627	41	99.39%	0.61%	220045	213487	6558	32.49%	68.07%	35.59	47.72
34	6668	6627	41	99.39%	0.61%	226713	220114	6599	33.49%	68.50%	35.00	48.90
35	6668	6613	55	99.18%	0.82%	233381	226727	6654	34.50%	69.07%	34.57	49.95
36	6668	6623	45	99.33%	0.67%	240049	233350	6699	35.51%	69.53%	34.03	51.07
37	6668	6619	49	99.27%	0.73%	246717	239969	6748	36.52%	70.04%	33.53	52.13
38	6668	6620	48	99.28%	0.72%	253385	246589	6796	37.52%	70.54%	33.02	53.19
39	6668	6614	54	99.19%	0.81%	260053	253203	6850	38.53%	71.10%	32.57	54.19
40	6669	6612	57	99.15%	0.85%	266722	259815	6907	39.54%	71.69%	32.16	55.14
41	6668	6606	62	99.07%	0.93%	273390	266421	6969	40.54%	72.34%	31.80	56.04
42	6668	6616	52	99.22%	0.78%	280058	273037	7021	41.55%	72.88%	31.33	57.01
43	6668	6621	47	99.30%	0.70%	286726	279658	7068	42.55%	73.37%	30.81	58.00
44	6668	6622	46	99.31%	0.69%	293394	286280	7114	43.56%	73.84%	30.28	58.99
45	6668	6611	57	99.15%	0.85%	300062	292891	7171	44.57%	74.43%	29.87	59.88
46	6668	6611	57	99.15%	0.85%	306730	299502	7228	45.57%	75.03%	29.45	60.74
47	6668	6624	44	99.34%	0.66%	313398	306126	7272	46.58%	75.48%	28.90	61.71
48	6668	6619	49	99.27%	0.73%	320066	312745	7321	47.59%	75.99%	28.40	62.63
49	6668	6621	47	99.30%	0.70%	326734	319366	7368	48.60%	76.48%	27.88	63.54
50	6668	6619	49	99.27%	0.73%	333402	325985	7417	49.60%	76.99%	27.38	64.43
51	6668	6614	54	99.19%	0.81%	340070	332599	7471	50.61%	77.55%	26.94	65.26
52	6668	6628	40	99.40%	0.60%	346738	339227	7511	51.62%	77.96%	26.34	66.21
53	6668	6624	44	99.34%	0.66%	353406	345851	7555	52.63%	78.42%	25.79	67.11
54	6668	6616	52	99.22%	0.78%	360074	352467	7607	53.63%	78.96%	25.33	67.93
55	6668	6628	40	99.40%	0.60%	366742	359095	7647	54.64%	79.38%	24.73	68.84
56	6668	6628	40	99.40%	0.60%	373410	365723	7687	55.65%	79.79%	24.14	69.75
57	6668	6626	42	99.37%	0.63%	380078	372349	7729	56.66%	80.23%	23.57	70.62
58	6668	6617	51	99.24%	0.76%	386746	378966	7780	57.67%	80.76%	23.09	71.41
59	6668	6623	45	99.33%	0.67%	393414	385589	7825	58.67%	81.22%	22.55	72.24
60	6669	6615	54	99.19%	0.81%	400083	392204	7879	59.68%	81.78%	22.10	72.97
61	6668	6621	47	99.30%	0.70%	406751	398825	7926	60.69%	82.27%	21.58	73.77
62	6668	6621	47	99.30%	0.70%	413419	405446	7973	61.70%	82.76%	21.06	74.55
63	6668	6611	57	99.15%	0.85%	420087	412057	8030	62.70%	83.35%	20.65	75.23
64	6668	6610	58	99.13%	0.87%	426755	418667	8088	63.71%	83.95%	20.25	75.88
65	6668	6624	44	99.34%	0.66%	433423	425291	8132	64.72%	84.41%	19.69	76.67
66	6668	6610	58	99.13%	0.87%	440091	431901	8190	65.72%	85.01%	19.29	77.31

67	6668	6613	55	99.18%	0.82%	446759	438514	8245	66.73%	85.58%	18.85	77.97
68	6668	6619	49	99.27%	0.73%	453427	445133	8294	67.73%	86.09%	18.36	78.68
69	6668	6604	64	99.04%	0.96%	460095	451737	8358	68.74%	86.76%	18.02	79.23
70	6668	6632	36	99.46%	0.54%	466763	458369	8394	69.75%	87.13%	17.38	80.05
71	6668	6619	49	99.27%	0.73%	473431	464988	8443	70.76%	87.64%	16.88	80.74
72	6668	6628	40	99.40%	0.60%	480099	471616	8483	71.76%	88.05%	16.29	81.50
73	6668	6622	46	99.31%	0.69%	486767	478238	8529	72.77%	88.53%	15.76	82.20
74	6668	6616	52	99.22%	0.78%	493435	484854	8581	73.78%	89.07%	15.29	82.83
75	6668	6623	45	99.33%	0.67%	500103	491477	8626	74.79%	89.54%	14.75	83.53
76	6668	6623	45	99.33%	0.67%	506771	498100	8671	75.79%	90.00%	14.21	84.21
77	6668	6616	52	99.22%	0.78%	513439	504716	8723	76.80%	90.54%	13.74	84.82
78	6668	6624	44	99.34%	0.66%	520107	511340	8767	77.81%	91.00%	13.19	85.50
79	6668	6619	49	99.27%	0.73%	526775	517959	8816	78.82%	91.51%	12.69	86.13
80	6669	6620	49	99.27%	0.73%	533444	524579	8865	79.82%	92.02%	12.19	86.75
81	6668	6597	71	98.94%	1.06%	540112	531176	8936	80.83%	92.75%	11.93	87.14
82	6668	6631	37	99.45%	0.55%	546780	537807	8973	81.84%	93.14%	11.30	87.87
83	6668	6619	49	99.27%	0.73%	553448	544426	9022	82.84%	93.65%	10.80	88.46
84	6668	6621	47	99.30%	0.70%	560116	551047	9069	83.85%	94.14%	10.28	89.08
85	6668	6616	52	99.22%	0.78%	566784	557663	9121	84.86%	94.68%	9.82	89.63
86	6668	6615	53	99.21%	0.79%	573452	564278	9174	85.86%	95.23%	9.36	90.17
87	6668	6611	57	99.15%	0.85%	580120	570889	9231	86.87%	95.82%	8.95	90.66
88	6668	6600	68	98.98%	1.02%	586788	577489	9299	87.87%	96.52%	8.65	91.04
89	6668	6620	48	99.28%	0.72%	593456	584109	9347	88.88%	97.02%	8.14	91.61
90	6668	6628	40	99.40%	0.60%	600124	590737	9387	89.89%	97.44%	7.55	92.26
91	6668	6637	31	99.54%	0.46%	606792	597374	9418	90.90%	97.76%	6.86	92.99
92	6668	6635	33	99.51%	0.49%	613460	604009	9451	91.91%	98.10%	6.19	93.69
93	6668	6628	40	99.40%	0.60%	620128	610637	9491	92.92%	98.52%	5.60	94.32
94	6668	6632	36	99.46%	0.54%	626796	617269	9527	93.93%	98.89%	4.96	94.98
95	6668	6638	30	99.55%	0.45%	633464	623907	9557	94.94%	99.20%	4.26	95.70
96	6668	6646	22	99.67%	0.33%	640132	630553	9579	95.95%	99.43%	3.48	96.50
97	6668	6644	24	99.64%	0.36%	646800	637197	9603	96.96%	99.68%	2.72	97.27
98	6668	6657	11	99.84%	0.16%	653468	643854	9614	97.97%	99.79%	1.82	98.18
99	6668	6654	14	99.79%	0.21%	660136	650508	9628	98.99%	99.94%	0.95	99.05
100	6669	6663	6	99.91%	0.09%	666805	657171	9634	100.00%	100.00%	0.00	100.00

Testing	#Records	#Goods	#Bads	'raud Rate								
	166702	164329	2373	0.014235								
			Bin Statistics					Cumulative Statistics				
ulation B	#Records	#Goods	#Bads	%Goods	%Bads	tal#Recor	lative G	ulative B	%Goods	Bads(FDR)	KS	FPR
1	1667	446	1221	26.75%	73.25%	1667	446	1221	0.27%	51.45%	51.18	0.53
2	1667	1630	37	97.78%	2.22%	3334	2076	1258	1.26%	53.01%	51.75	2.38
3	1667	1654	13	99.22%	0.78%	5001	3730	1271	2.27%	53.56%	51.29	4.24
4	1667	1656	11	99.34%	0.66%	6668	5386	1282	3.28%	54.02%	50.75	6.07
5	1667	1656	11	99.34%	0.66%	8335	7042	1293	4.29%	54.49%	50.20	7.86
6	1667	1653	14	99.16%	0.84%	10002	8695	1307	5.29%	55.08%	49.79	9.61
7	1667	1651	16	99.04%	0.96%	11669	10346	1323	6.30%	55.75%	49.46	11.29
8	1667	1656	11	99.34%	0.66%	13336	12002	1334	7.30%	56.22%	48.91	12.99
9	1667	1659	8	99.52%	0.48%	15003	13661	1342	8.31%	56.55%	48.24	14.70
10	1667	1657	10	99.40%	0.60%	16670	15318	1352	9.32%	56.97%	47.65	16.36
11	1667	1658	9	99.46%	0.54%	18337	16976	1361	10.33%	57.35%	47.02	18.01
12	1667	1656	11	99.34%	0.66%	20004	18632	1372	11.34%	57.82%	46.48	19.61
13	1667	1651	16	99.04%	0.96%	21671	20283	1388	12.34%	58.49%	46.15	21.10
14	1667	1657	10	99.40%	0.60%	23338	21940	1398	13.35%	58.91%	45.56	22.66
15	1667	1659	8	99.52%	0.48%	25005	23599	1406	14.36%	59.25%	44.89	24.24
16	1667	1657	10	99.40%	0.60%	26672	25256	1416	15.37%	59.67%	44.30	25.76
17	1667	1650	17	98.98%	1.02%	28339	26906	1433	16.37%	60.39%	44.01	27.11
18	1667	1654	13	99.22%	0.78%	30006	28560	1446	17.38%	60.94%	43.56	28.52
19	1667	1652	15	99.10%	0.90%	31673	30212	1461	18.39%	61.57%	43.18	29.86
20	1667	1657	10	99.40%	0.60%	33340	31869	1471	19.39%	61.99%	42.60	31.29
21	1667	1653	14	99.16%	0.84%	35007	33522	1485	20.40%	62.58%	42.18	32.60
22	1667	1646	21	98.74%	1.26%	36674	35168	1506	21.40%	63.46%	42.06	33.72
23	1667	1653	14	99.16%	0.84%	38341	36821	1520	22.41%	64.05%	41.65	34.98
24	1667	1655	12	99.28%	0.72%	40008	38476	1532	23.41%	64.56%	41.15	36.27
25	1667	1648	19	98.86%	1.14%	41675	40124	1551	24.42%	65.36%	40.94	37.36
26	1667	1652	15	99.10%	0.90%	43342	41776	1566	25.42%	65.99%	40.57	38.52
27	1667	1654	13	99.22%	0.78%	45009	43430	1579	26.43%	66.54%	40.11	39.72
28	1667	1654	13	99.22%	0.78%	46676	45084	1592	27.44%	67.09%	39.65	40.89
29	1667	1656	11	99.34%	0.66%	48343	46740	1603	28.44%	67.55%	39.11	42.11
30	1667	1650	17	98.98%	1.02%	50010	48390	1620	29.45%	68.27%	38.82	43.13
31	1667	1655	12	99.28%	0.72%	51677	50045	1632	30.45%	68.77%	38.32	44.28
32	1667	1660	7	99.58%	0.42%	53344	51705	1639	31.46%	69.07%	37.60	45.56
33	1667	1656	11	99.34%	0.66%	55011	53361	1650	32.47%	69.53%	37.06	46.70
34	1667	1659	8	99.52%	0.48%	56678	55020	1658	33.48%	69.87%	36.39	47.92
35	1667	1651	16	99.04%	0.96%	58345	56671	1674	34.49%	70.54%	36.06	48.89
36	1667	1653	14	99.16%	0.84%	60012	58324	1688	35.49%	71.13%	35.64	49.90
37	1667	1660	7	99.58%	0.42%	61679	59984	1695	36.50%	71.43%	34.93	51.10
38	1667	1657	10	99.40%	0.60%	63346	61641	1705	37.51%	71.85%	34.34	52.21
39	1667	1651	16	99.04%	0.96%	65013	63292	1721	38.52%	72.52%	34.01	53.11
40	1667	1653	14	99.16%	0.84%	66680	64945	1735	39.52%	73.11%	33.59	54.05
41	1667	1654	13	99.22%	0.78%	68347	66599	1748	40.53%	73.66%	33.13	55.02
42	1667	1658	9	99.46%	0.54%	70014	68257	1757	41.54%	74.04%	32.50	56.10
43	1667	1649	18	98.92%	1.08%	71681	69906	1775	42.54%	74.80%	32.26	56.87
44	1667	1649	18	98.92%	1.08%	73348	71555	1793	43.54%	75.56%	32.01	57.63
45	1667	1658	9	99.46%	0.54%	75015	73213	1802	44.55%	75.94%	31.38	58.67
46	1667	1658	9	99.46%	0.54%	76682	74871	1811	45.56%	76.32%	30.76	59.70
47	1667	1661	6	99.64%	0.36%	78349	76532	1817	46.57%	76.57%	30.00	60.82
48	1667	1658	9	99.46%	0.54%	80016	78190	1826	47.58%	76.95%	29.37	61.83
49	1667	1656	11	99.34%	0.66%	81683	79846	1837	48.59%	77.41%	28.82	62.77
50	1668	1656	12	99.28%	0.72%	83351	81502	1849	49.60%	77.92%	28.32	63.65
51	1667	1655	12	99.28%	0.72%	85018	83157	1861	50.60%	78.42%	27.82	64.53
52	1667	1655	12	99.28%	0.72%	86685	84812	1873	51.61%	78.93%	27.32	65.39
53	1667	1649	18	98.92%	1.08%	88352	86461	1891	52.61%	79.69%	27.07	66.03
54	1667	1656	11	99.34%	0.66%	90019	88117	1902	53.62%	80.15%	26.53	66.90
55	1667	1660	7	99.58%	0.42%	91686	89777	1909	54.63%	80.45%	25.81	67.91
56	1667	1658	9	99.46%	0.54%	93353	91435	1918	55.64%	80.83%	25.18	68.84
57	1667	1655	12	99.28%	0.72%	95020	93090	1930	56.65%	81.33%	24.68	69.65
58	1667	1656	11	99.34%	0.66%	96687	94746	1941	57.66%	81.80%	24.14	70.49
59	1667	1657	10	99.40%	0.60%	98354	96403	1951	58.66%	82.22%	23.55	71.35
60	1667	1660	7	99.58%	0.42%	100021	98063	1958	59.67%	82.51%	22.84	72.32
61	1667	1650	17	98.98%	1.02%	101688	99713	1975	60.68%	83.23%	22.55	72.91
62	1667	1653	14	99.16%	0.84%	103355	101366	1989	61.68%	83.82%	22.13	73.59
63	1667	1657	10	99.40%	0.60%	105022	103023	1999	62.69%	84.24%	21.55	74.42
64	1667	1650	17	98.98%	1.02%	106689	104673	2016	63.70%	84.96%	21.26	74.98
65	1667	1656	11	99.34%	0.66%	108356	106329	2027	64.70%	85.42%	20.71	75.75
66	1667	1657	10	99.40%	0.60%	110023	107986	2037	65.71%	85.84%	20.13	76.55

67	1667	1658	9	99.46%	0.54%	111690	109644	2046	66.72%	86.22%	19.50	77.39
68	1667	1654	13	99.22%	0.78%	113357	111298	2059	67.73%	86.77%	19.04	78.06
69	1667	1656	11	99.34%	0.66%	115024	112954	2070	68.74%	87.23%	18.49	78.80
70	1667	1658	9	99.46%	0.54%	116691	114612	2079	69.75%	87.61%	17.87	79.61
71	1667	1658	9	99.46%	0.54%	118358	116270	2088	70.75%	87.99%	17.24	80.41
72	1667	1658	9	99.46%	0.54%	120025	117928	2097	71.76%	88.37%	16.61	81.21
73	1667	1651	16	99.04%	0.96%	121692	119579	2113	72.77%	89.04%	16.28	81.72
74	1667	1652	15	99.10%	0.90%	123359	121231	2128	73.77%	89.68%	15.90	82.27
75	1667	1659	8	99.52%	0.48%	125026	122890	2136	74.78%	90.01%	15.23	83.08
76	1667	1657	10	99.40%	0.60%	126693	124547	2146	75.79%	90.43%	14.64	83.81
77	1667	1656	11	99.34%	0.66%	128360	126203	2157	76.80%	90.90%	14.10	84.49
78	1667	1656	11	99.34%	0.66%	130027	127859	2168	77.81%	91.36%	13.55	85.16
79	1667	1658	9	99.46%	0.54%	131694	129517	2177	78.82%	91.74%	12.92	85.91
80	1667	1658	9	99.46%	0.54%	133361	131175	2186	79.82%	92.12%	12.30	86.65
81	1667	1657	10	99.40%	0.60%	135028	132832	2196	80.83%	92.54%	11.71	87.35
82	1667	1657	10	99.40%	0.60%	136695	134489	2206	81.84%	92.96%	11.12	88.04
83	1667	1658	9	99.46%	0.54%	138362	136147	2215	82.85%	93.34%	10.49	88.76
84	1667	1649	18	98.92%	1.08%	140029	137796	2233	83.85%	94.10%	10.25	89.11
85	1667	1652	15	99.10%	0.90%	141696	139448	2248	84.86%	94.73%	9.87	89.58
86	1667	1658	9	99.46%	0.54%	143363	141106	2257	85.87%	95.11%	9.24	90.28
87	1667	1658	9	99.46%	0.54%	145030	142764	2266	86.88%	95.49%	8.61	90.98
88	1667	1660	7	99.58%	0.42%	146697	144424	2273	87.89%	95.79%	7.90	91.75
89	1667	1653	14	99.16%	0.84%	148364	146077	2287	88.89%	96.38%	7.48	92.24
90	1667	1655	12	99.28%	0.72%	150031	147732	2299	89.90%	96.88%	6.98	92.79
91	1667	1660	7	99.58%	0.42%	151698	149392	2306	90.91%	97.18%	6.27	93.55
92	1667	1660	7	99.58%	0.42%	153365	151052	2313	91.92%	97.47%	5.55	94.30
93	1667	1659	8	99.52%	0.48%	155032	152711	2321	92.93%	97.81%	4.88	95.01
94	1667	1662	5	99.70%	0.30%	156699	154373	2326	93.94%	98.02%	4.08	95.84
95	1667	1665	2	99.88%	0.12%	158366	156038	2328	94.95%	98.10%	3.15	96.79
96	1667	1657	10	99.40%	0.60%	160033	157695	2338	95.96%	98.53%	2.56	97.40
97	1667	1656	11	99.34%	0.66%	161700	159351	2349	96.97%	98.99%	2.02	97.96
98	1667	1663	4	99.76%	0.24%	163367	161014	2353	97.98%	99.16%	1.17	98.82
99	1667	1658	9	99.46%	0.54%	165034	162672	2362	98.99%	99.54%	0.54	99.45
100	1668	1657	11	99.34%	0.66%	166702	164329	2373	100.00%	100.00%	0.00	100.00

OOT	#Records	#Goods	#Bads	Fraud Rate
	166493	164107	2386	0.0143309

	Bin Statistics					Cumulative Statistics						
pulationBi	#Records	#Goods	#Bads	%Goods	%Bads	tal#Recor	nulativeGoc	mulativeBa	%Goods	%Bads(FDR)	KS	FPR
1	1664	545	1119	32.75%	67.25%	1664	545	1119	0.33%	46.90%	46.57	0.71
2	1665	1560	105	93.69%	6.31%	3329	2105	1224	1.28%	51.30%	50.02	2.50
3	1665	1640	25	98.50%	1.50%	4994	3745	1249	2.28%	52.35%	50.06	4.36
4	1665	1648	17	98.98%	1.02%	6659	5393	1266	3.29%	53.06%	49.77	6.19
5	1665	1653	12	99.28%	0.72%	8324	7046	1278	4.29%	53.56%	49.27	8.02
6	1665	1654	11	99.34%	0.66%	9989	8700	1289	5.30%	54.02%	48.72	9.81
7	1665	1651	14	99.16%	0.84%	11654	10351	1303	6.31%	54.61%	48.30	11.55
8	1665	1653	12	99.28%	0.72%	13319	12004	1315	7.31%	55.11%	47.80	13.27
9	1665	1657	8	99.52%	0.48%	14984	13661	1323	8.32%	55.45%	47.12	15.01
10	1665	1651	14	99.16%	0.84%	16649	15312	1337	9.33%	56.04%	46.70	16.65
11	1665	1657	8	99.52%	0.48%	18314	16969	1345	10.34%	56.37%	46.03	18.34
12	1665	1652	13	99.22%	0.78%	19979	18621	1358	11.35%	56.92%	45.57	19.94
13	1665	1650	15	99.10%	0.90%	21644	20271	1373	12.35%	57.54%	45.19	21.47
14	1665	1657	8	99.52%	0.48%	23309	21928	1381	13.36%	57.88%	44.52	23.09
15	1664	1653	11	99.34%	0.66%	24973	23581	1392	14.37%	58.34%	43.97	24.63
16	1665	1655	10	99.40%	0.60%	26638	25236	1402	15.38%	58.76%	43.38	26.17
17	1665	1653	12	99.28%	0.72%	28303	26889	1414	16.39%	59.26%	42.88	27.65
18	1665	1655	10	99.40%	0.60%	29968	28544	1424	17.39%	59.68%	42.29	29.14
19	1665	1657	8	99.52%	0.48%	31633	30201	1432	18.40%	60.02%	41.61	30.66
20	1665	1657	8	99.52%	0.48%	33298	31858	1440	19.41%	60.35%	40.94	32.17
21	1665	1655	10	99.40%	0.60%	34963	33513	1450	20.42%	60.77%	40.35	33.60
22	1665	1649	16	99.04%	0.96%	36628	35162	1466	21.43%	61.44%	40.02	34.87
23	1665	1646	19	98.86%	1.14%	38293	36808	1485	22.43%	62.24%	39.81	36.04
24	1665	1657	8	99.52%	0.48%	39958	38465	1493	23.44%	62.57%	39.13	37.46
25	1665	1652	13	99.22%	0.78%	41623	40117	1506	24.45%	63.12%	38.67	38.73
26	1665	1652	13	99.22%	0.78%	43288	41769	1519	25.45%	63.66%	38.21	39.98
27	1665	1657	8	99.52%	0.48%	44953	43426	1527	26.46%	64.00%	37.54	41.35
28	1665	1650	15	99.10%	0.90%	46618	45076	1542	27.47%	64.63%	37.16	42.50
29	1664	1652	12	99.28%	0.72%	48282	46728	1554	28.47%	65.13%	36.66	43.72
30	1665	1648	17	98.98%	1.02%	49947	48376	1571	29.48%	65.84%	36.36	44.77
31	1665	1653	12	99.28%	0.72%	51612	50029	1583	30.49%	66.35%	35.86	45.95
32	1665	1658	7	99.58%	0.42%	53277	51687	1590	31.50%	66.64%	35.14	47.26
33	1665	1658	7	99.58%	0.42%	54942	53345	1597	32.51%	66.93%	34.43	48.57
34	1665	1654	11	99.34%	0.66%	56607	54999	1608	33.51%	67.39%	33.88	49.73
35	1665	1653	12	99.28%	0.72%	58272	56652	1620	34.52%	67.90%	33.37	50.84
36	1665	1656	9	99.46%	0.54%	59937	58308	1629	35.53%	68.27%	32.74	52.04
37	1665	1654	11	99.34%	0.66%	61602	59962	1640	36.54%	68.73%	32.20	53.16
38	1665	1650	15	99.10%	0.90%	63267	61612	1655	37.54%	69.36%	31.82	54.13
39	1665	1653	12	99.28%	0.72%	64932	63265	1667	38.55%	69.87%	31.31	55.18
40	1665	1650	15	99.10%	0.90%	66597	64915	1682	39.56%	70.49%	30.94	56.11
41	1665	1650	15	99.10%	0.90%	68262	66565	1697	40.56%	71.12%	30.56	57.03
42	1665	1651	14	99.16%	0.84%	69927	68216	1711	41.57%	71.71%	30.14	57.97
43	1664	1652	12	99.28%	0.72%	71591	69868	1723	42.57%	72.21%	29.64	58.96
44	1665	1651	14	99.16%	0.84%	73256	71519	1737	43.58%	72.80%	29.22	59.86
45	1665	1653	12	99.28%	0.72%	74921	73172	1749	44.59%	73.30%	28.71	60.83
46	1665	1652	13	99.22%	0.78%	76586	74824	1762	45.59%	73.85%	28.25	61.74
47	1665	1652	13	99.22%	0.78%	78251	76476	1775	46.60%	74.39%	27.79	62.64
48	1665	1654	11	99.34%	0.66%	79916	78130	1786	47.61%	74.85%	27.24	63.60
49	1665	1654	11	99.34%	0.66%	81581	79784	1797	48.62%	75.31%	26.70	64.55
50	1665	1647	18	98.92%	1.08%	83246	81431	1815	49.62%	76.07%	26.45	65.23
51	1665	1654	11	99.34%	0.66%	84911	83085	1826	50.63%	76.53%	25.90	66.16
52	1665	1655	10	99.40%	0.60%	86576	84740	1836	51.64%	76.95%	25.31	67.11
53	1665	1656	9	99.46%	0.54%	88241	86396	1845	52.65%	77.33%	24.68	68.08
54	1665	1654	11	99.34%	0.66%	89906	88050	1856	53.65%	77.79%	24.13	68.98
55	1665	1659	6	99.64%	0.36%	91571	89709	1862	54.66%	78.04%	23.37	70.05
56	1665	1650	15	99.10%	0.90%	93236	91359	1877	55.67%	78.67%	23.00	70.77
57	1665	1654	11	99.34%	0.66%	94901	93013	1888	56.68%	79.13%	22.45	71.63
58	1664	1642	22	98.68%	1.32%	96565	94655	1910	57.68%	80.05%	22.37	72.05
59	1665	1654	11	99.34%	0.66%	98230	96309	1921	58.69%	80.51%	21.82	72.89
60	1665	1644	21	98.74%	1.26%	99895	97953	1942	59.69%	81.39%	21.70	73.34
61	1665	1655	10	99.40%	0.60%	101560	99608	1952	60.70%	81.81%	21.11	74.19
62	1665	1656	9	99.46%	0.54%	103225	101264	1961	61.71%	82.19%	20.48	75.08
63	1665	1659	6	99.64%	0.36%	104890	102923	1967	62.72%	82.44%	19.72	76.08
64	1665	1648	17	98.98%	1.02%	106555	104571	1984	63.72%	83.15%	19.43	76.63
65	1665	1646	19	98.86%	1.14%	108220	106217	2003	64.72%	83.95%	19.22	77.10
66	1665	1657	8	99.52%	0.48%	109885	107874	2011	65.73%	84.28%	18.55	77.99

67	1665	1656	9	99.46%	0.54%	111550	109530	2020	66.74%	84.66%	17.92	78.84
68	1665	1654	11	99.34%	0.66%	113215	111184	2031	67.75%	85.12%	17.37	79.59
69	1665	1648	17	98.98%	1.02%	114880	112832	2048	68.76%	85.83%	17.08	80.10
70	1665	1650	15	99.10%	0.90%	116545	114482	2063	69.76%	86.46%	16.70	80.68
71	1665	1650	15	99.10%	0.90%	118210	116132	2078	70.77%	87.09%	16.33	81.25
72	1664	1649	15	99.10%	0.90%	119874	117781	2093	71.77%	87.72%	15.95	81.82
73	1665	1652	13	99.22%	0.78%	121539	119433	2106	72.78%	88.26%	15.49	82.45
74	1665	1657	8	99.52%	0.48%	123204	121090	2114	73.79%	88.60%	14.81	83.28
75	1665	1654	11	99.34%	0.66%	124869	122744	2125	74.80%	89.06%	14.27	83.98
76	1665	1657	8	99.52%	0.48%	126534	124401	2133	75.80%	89.40%	13.59	84.80
77	1665	1651	14	99.16%	0.84%	128199	126052	2147	76.81%	89.98%	13.17	85.36
78	1665	1651	14	99.16%	0.84%	129864	127703	2161	77.82%	90.57%	12.75	85.92
79	1665	1647	18	98.92%	1.08%	131529	129350	2179	78.82%	91.32%	12.50	86.31
80	1665	1649	16	99.04%	0.96%	133194	130999	2195	79.83%	91.99%	12.17	86.77
81	1665	1652	13	99.22%	0.78%	134859	132651	2208	80.83%	92.54%	11.71	87.35
82	1665	1658	7	99.58%	0.42%	136524	134309	2215	81.84%	92.83%	10.99	88.16
83	1665	1652	13	99.22%	0.78%	138189	135961	2228	82.85%	93.38%	10.53	88.72
84	1665	1654	11	99.34%	0.66%	139854	137615	2239	83.86%	93.84%	9.98	89.36
85	1665	1656	9	99.46%	0.54%	141519	139271	2248	84.87%	94.22%	9.35	90.08
86	1664	1654	10	99.40%	0.60%	143183	140925	2258	85.87%	94.64%	8.76	90.74
87	1665	1651	14	99.16%	0.84%	144848	142576	2272	86.88%	95.22%	8.34	91.24
88	1665	1658	7	99.58%	0.42%	146513	144234	2279	87.89%	95.52%	7.63	92.02
89	1665	1651	14	99.16%	0.84%	148178	145885	2293	88.90%	96.10%	7.21	92.50
90	1665	1657	8	99.52%	0.48%	149843	147542	2301	89.91%	96.44%	6.53	93.23
91	1665	1654	11	99.34%	0.66%	151508	149196	2312	90.91%	96.90%	5.98	93.82
92	1665	1652	13	99.22%	0.78%	153173	150848	2325	91.92%	97.44%	5.52	94.33
93	1665	1659	6	99.64%	0.36%	154838	152507	2331	92.93%	97.69%	4.76	95.12
94	1665	1654	11	99.34%	0.66%	156503	154161	2342	93.94%	98.16%	4.22	95.70
95	1665	1661	4	99.76%	0.24%	158168	155822	2346	94.95%	98.32%	3.37	96.57
96	1665	1660	5	99.70%	0.30%	159833	157482	2351	95.96%	98.53%	2.57	97.39
97	1665	1661	4	99.76%	0.24%	161498	159143	2355	96.98%	98.70%	1.73	98.25
98	1665	1657	8	99.52%	0.48%	163163	160800	2363	97.98%	99.04%	1.05	98.94
99	1665	1652	13	99.22%	0.78%	164828	162452	2376	98.99%	99.58%	0.59	99.41
100	1665	1655	10	99.40%	0.60%	166493	164107	2386	100.00%	100.00%	0.00	100.00