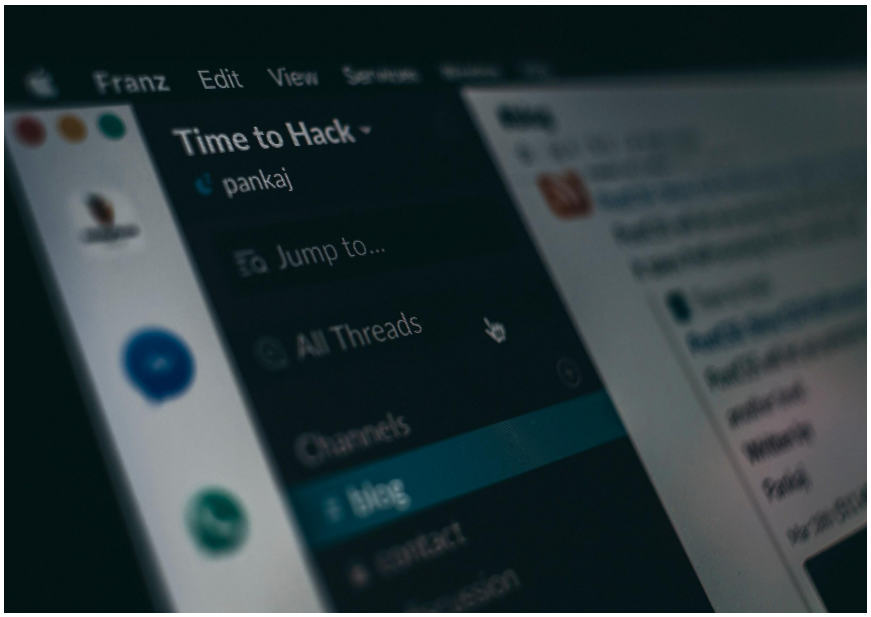
DSO 562 Project 2

USC Marshall
School of Business

Fraud Detection on Product Application Data



UNDER THE GUIDANCE OF

Professor Stephen Coggeshall

3/15/2020

JENNY SHANG |

JENNY WANG |

LINGROU WANG |

RORY WANG |

Table of Contents

Part I. Executive Summary

This report is a summary of the analysis of the Product Application dataset. The dataset contains 1,000,000 records with ten attributes of basic information for product applications, and it is collected for the purpose of detecting product application fraud. The objective of the report is to identify fraud events using supervised machine learning algorithms, including logistic regression, decision trees, neural networks, random forests, and boosted trees to produce the best model to predict fraudulent applications.

After data exploration and data cleaning, 305 candidate variables were built through data manipulation. Then the importance of each variable was evaluated using KS distance and fraud detection rate at a 3% cutoff for all records. Variables were ranked, and the 30 most influential variables were selected through the recursive feature elimination method.

To help alleviate unbalanced data, a 3% cutoff and SMOTE method were adopted to help simulate more minority cases without losing existing information from the data. By splitting the data into training, testing, and out-of-time datasets, models were trained using the training dataset, and the model effectiveness was examined on the testing dataset. Finally, the fraud detection rate (FDR) at 3% for the out of time dataset was used as a final test assessment to identify the best model. For each algorithm, the optimal combination of variables was achieved by tuning the parameters to get the best results. This complete process is summarized in Figure 1.1 below.

Figure 1.1 - Overall Process of Project

Overall, the best model with the highest prediction accuracy of fraudulent applications is the neural net model with two layers of 15 nodes each, epochs of 50, and a learning rate of 0.001, which achieved a fraud detection rate of 52.85% on out-of-time data in the 3% highest scored records. All other models have relatively close prediction accuracy, and details can be found in Table 1.1 below.

	FDR @ 3%		
Model	**Training**	**Testing**	**OOT**
Logistic Regression	52.69%	53.94%	51.93%
Decision Tree	53.56%	53.53%	51.39%
Neural Net	54.25%	54.59%	52.85%
Random Forest	55.79%	55.13%	52.64%
Gradient Boosting Trees	55.52%	55.13%	52.72%

Table 1.1 Overall Result with Different Models

Part II. Description of Data

Overview of Data

The "applications data" is a dataset that is generated using an algorithm developed by one of Professor Coggeshall's colleagues. This dataset simulates a real-life business scenario where a company tracks and identifies fraudulent applications amongst regular, non-fraudulent applications. The data is generated and provided by Professor Coggeshall for academic purposes only.

In this dataset, the value for all fields are randomly generated using the algorithm and thus contains no real information. In summary, this dataset includes a total of 10 fields and 1,000,000 records. All records are generated with dates in 2016 in a chronological order, from January 1, 2016 to December 31, 2016. Since in a real-life scenario, it is impossible to know the fraud label of applications that occur in the future, it is vital to keep in mind the order of applications when constructing supervised fraud detection models.

A data quality report is constructed with details in Appendix One. The report begins with a summary of the variables in the dataset, follows with an analysis of each of the variables, and concludes with additional notes on the dataset. A snippet of essential features from the data quality report is included in the sections below.

Summary Table

The dataset contains one variable, "record", that acts as a unique identifier of each application. A date field tracks the year, month, and date of the application in the format "yyyymmdd". Since all information is personal identification information, there is no numerical field in this dataset. The remaining eight variables are all categorical variables. The numbers in the remaining fields are for personal identification purposes and thus do not exhibit meaning based on the absolute size of these numbers.

Table 2.1 Variables Summary Table

Field Name	# Records with Value	% Populated	# Unique Values	# Records with Value Zero	Most Common Value
record	1000000	100%	100000	0	
date	1000000	100%	365	0	20160816
ssn	1000000	100%	835819	0	999999999
firstname	1000000	100%	78136	0	EAMSTRMT
lastname	1000000	100%	177001	0	ERJSAXA
address	1000000	100%	828774	0	123 MAIN ST
zip5	1000000	100%	26370	0	68138
dob	1000000	100%	42673	0	19070626
homephone	1000000	100%	28244	0	9999999999
fraud_label	1000000	100%	2	985607	0

Visualizations for Variables

Figure 2.1 suggested that there are more applications on August 16th compared to other days, followed by March 4th and July 18th. This aligns with the spikes in daily applications as depicted in Figure 2.2.

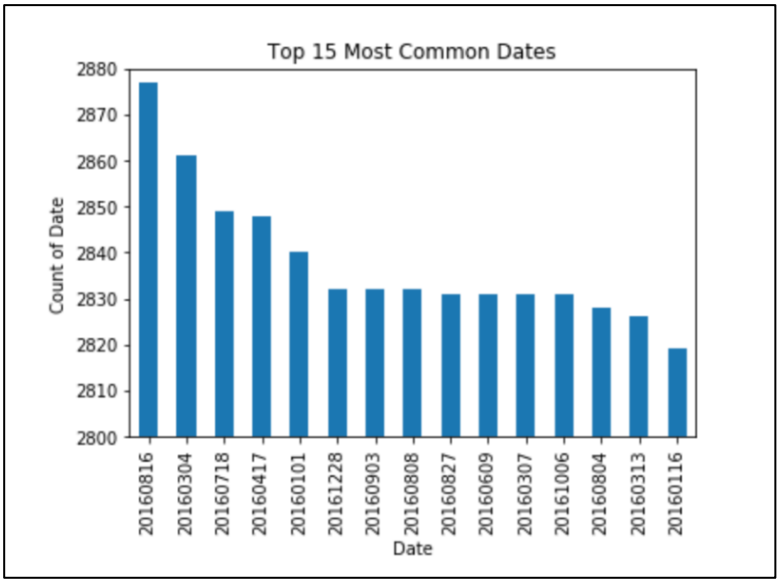


Figure 2.1 – "date" Distribution Graph

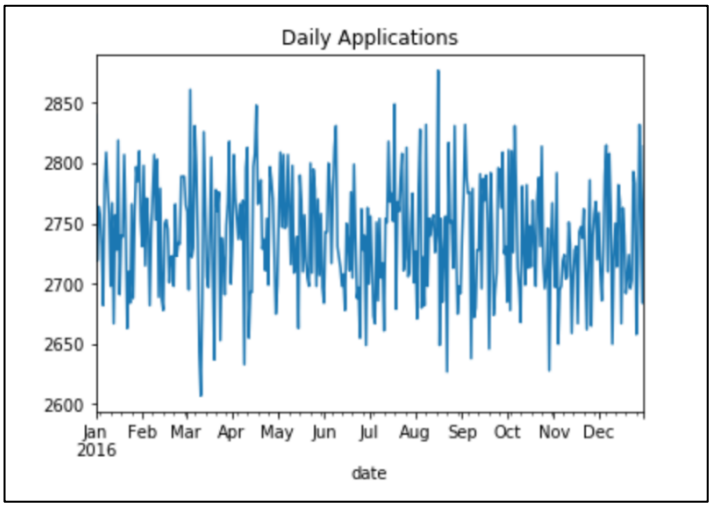


Figure 2.2 – Number of Applications (Daily)

After examining the distribution of fraudulent applications across time, a noticeable pattern is that the percentage of non-fraudulent applications per day over the year is at a steady rate of 0.03% of total non-fraudulent applications (as indicated by the green line in Figure 2.3). The percentage of fraudulent applications per day has ups and downs with big spikes in February and during the summer months of July to September (as indicated by the red line in Figure 2.3). Figure 2.3 below shows the breakdown of fraudulent vs. non-fraudulent applications by day, and Figure 2.4 shows the same breakdown by week.

These two graphs show that there is a particular day in February that has relatively more fraudulent applications than usual (likely 2016-02-04 with 90 fraudulent applications), while July and August consistently have more fraudulent applications than other times of the year.

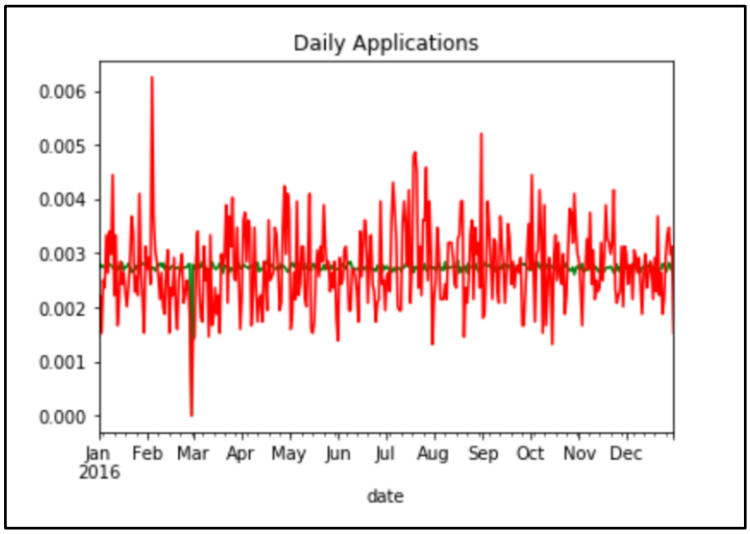


Figure 2.3 – Breakdown of Fraudulent vs. Non-Fraudulent Applications by Day

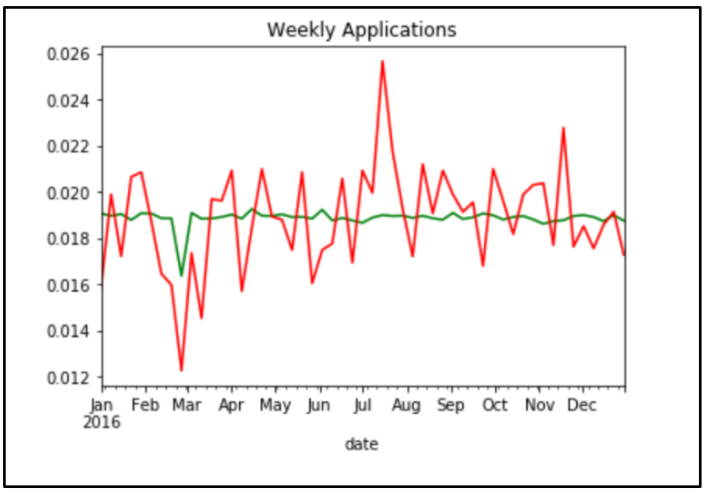


Figure 2.4 – Breakdown of Fraudulent vs. Non-Fraudulent Applications by Week

Further investigation of the data also reveals that there are frivolous values populated for the "ssn", "address", "dob", and "homephone" fields. Frivolous values are placeholder values populated for the sake of completion rather than being the actual values related to the applicant. A summary of the frivolous values is provided below.

Table 2.2 Summary of Frivolous Values

Field	Possible Frivolous Value	# of Frivolous Values that are Fraud	% of Frivolous Values that are Fraud
ssn	999999999	95	0.56%
address	123 MAIN ST	6	0.56%
dob	19070626	709	0.56%
homephone	9999999999	388	0.49%

Part III. Data Cleaning

The dataset does not have any null values among the total of 10 variables. But frivolous values seem to be the main concern for the data quality. Before performing feature selection and building models, the frivolous variables listed above had to be cleaned so that the data is usable. Without appropriate handling, frivolous values will have a huge impact on the counting of repetitive records if the two records have the same frivolous value. For example, even though they may have the same SSN at 9999999999, two records are likely to be unrelated. Therefore, the frivolous variables value was replaced with a number that was entirely unique and not seen before in the data. Specifically, a negative value of the record number is used.

The four variables that had data cleaning performed on were “ssn”, “address”, “homephone, and “dob”. These variables were deemed to be frivolous as they occurred more times than realistically possible in a given sample of the population of applicants. After identifying the different frivolous values in each of these four variables, they were replaced with the negative value of their record number. This ensured that it was a unique number for each case, and by making it negative, it would guarantee that it would not be an actual phone number or date of birth.

Therefore, after performing this replacement an applicant with the frivolous SSN of 999999999 would have that SSN number replaced with the negative value of their record number. No other data cleaning was performed as the team began to build as many candidate variables as possible before feature selection.

Part IV. Variable Creation

The variable creation process is summarized in two steps:

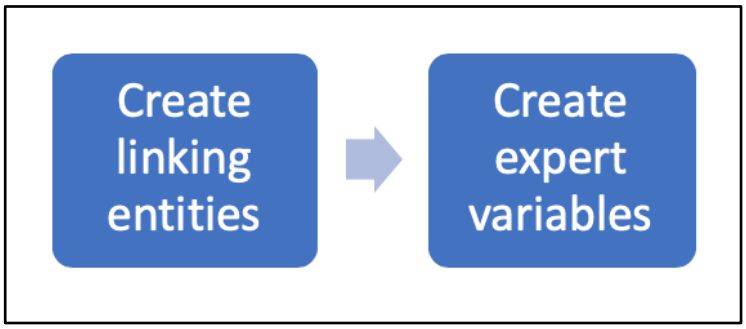


Figure 4.1 Variable Creation Process

The first step in the variable creation process is to create linking entities using the original fields present in the dataset. The purpose of creating the linking entities is to better distinguish each individual application such that there is a greater contrast between fraudulent and non-fraudulent records. This is an important step in the successful identification of the three types of fraud.

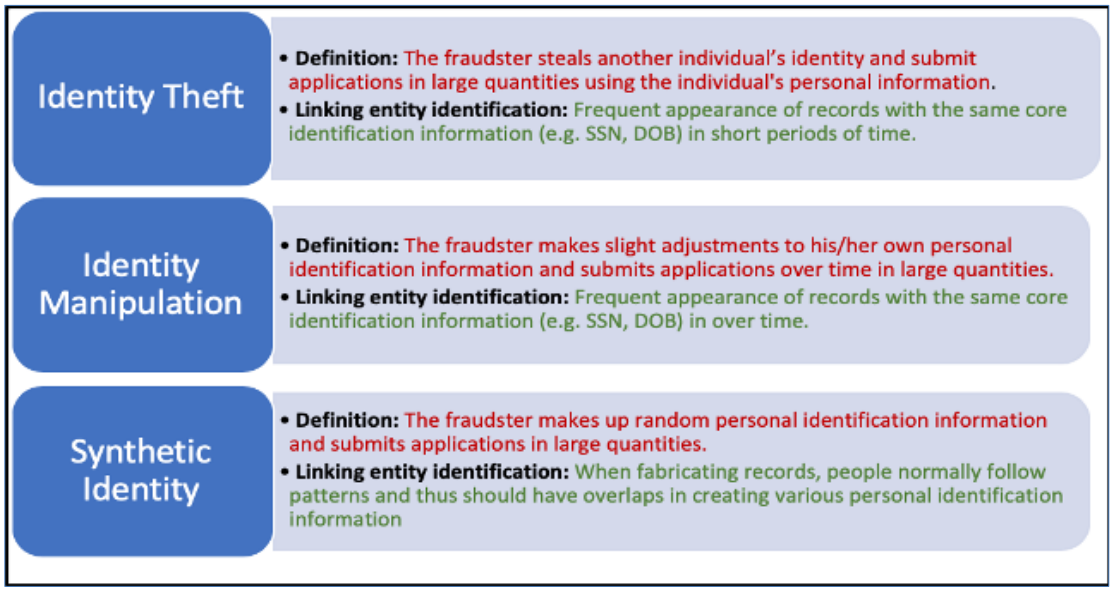


Figure 4.2 Different Types of Fraud

A total of 18 linking entities are created in addition to the existing six personal identification fields (see Table 4.1 below).

Table 4.1 Summary of Frivolous Values

Existing Entities	Newly Created Linking Entities		
ssn	namedob	ssnzip5	namefulladdress
firstname	fulladdress	ssndob	fulladdressdob
lastname	ssnfname	ssnhomephone	fulladdresshomephone
address	ssnlname	ssnnamedob	fulladdressnamedob
dob	ssnfullname	ssnfulladdress	dobhomephone
homephone	ssnaddress	namehomephone	homephonenamedob

After creating the linking entities, a total of 305 expert variables are created, which includes velocity variables, relative velocity variables, days since last seen variables, cross entity variables, and one risk variable. From the above Figure, it is evident that the three types of fraud have one aspect in common, which is the frequent appearance of the same identity information across different time periods. This unique characteristic is the core focus in creating the different types of expert variables.

Velocity Variables

Velocity variables reflect the number of occurrences for each value seen over the past certain number of days. A total of 154 velocity variables are created based on different combinations of entities and number of day counts. As an example, for each record's SSN, we can calculate how many times this particular SSN appears in the previous 7 days. Since there could be multiple individuals with the same first name and last name, these fields were omitted from being used as reference variables for velocity. The remaining 22 entities were considered and the number of occurrences for each value across the 22 entities in the previous 0, 1, 3, 7, 14, 30, and 180 days are calculated.

A key consideration in creating the velocity variables is the chronological sequence in which these records are generated. Since the data represents a time series and it is technically impossible to predict the future, the number of occurrences only factors in prior records with the same value. For example, if a particular SSN appeared three times in the same day, the 0-day velocity variable would have different values of 1, 2, and 3 because at the time of first occurrence, the latter two records have not taken place.

Relative Velocity Variables

Relative velocity variables are calculated by taking the ratio of the previously calculated velocity variables. The numerator is the velocity variable for short durations, such as 1 or 3

days, whereas the denominator is the same velocity variable but for longer periods. The relative velocity variables are particularly good indicators of fraudulent activities in cases where fraudsters send large amounts of applications in a short time frame. A total of 88 relative velocity variables are created.

Days Since Last Seen Variables

Days since last seen variables calculate the number of days since a particular value was last seen, and it is computed for 22 entities. The calculation first identifies the dates when a particular entity value appears and computes the difference between the record date and the previous date. This type of variable is a good indicator of fraudulent activities where fraudsters submit many applications consistently across time.

Similar to the velocity variables, the chronological occurrence of each record is taken into account. For values that appear multiple times in the same day, the latter occurrences except for the first will all have a value of 0. In this dataset there are many applications with values that only appear once. For these cases, a value of 366 is inputted as a placeholder value.

The variable created from above method has an indication of the higher the value, the lower the risk of that specific variable. For example, the higher the days since last seen a full address, the less likely it is to be linked to other applications with the same full address. In order to make this variable the same scale as all other ones that have been created, where the higher the value, the higher risk of abnormality, a transformation of using 366 minus the original days since value was adopted. Now with the new calculation, the higher the values in the specific variable, the riskier the application is.

Cross-Entity Variables

Cross-entity variables look at two different entities across a specific time period. It shows for a particular value in one entity, how many distinct values of the other entity appeared in a set number of days. To calculate these variables, five entities are used to create different pairs that are examined for the periods of one day and three days. Again, the order of applications is taken into consideration when calculating the value for these cross-entity variables. Using the pair of “ssn” and “homephone” entities as an example, one cross-entity variable looks at for each SSN, how many unique home phone numbers appeared in the last three days. A total of 40 cross-entity variables are generated.

Risk Variable

Risk variables are used to assign a value to categorical variables. In this case, a risk variable is created based on the day of week each application is created. For each day of week, the average fraud label value is computed for all records that occur on that particular day of week. This value then becomes the record’s risk variable value. Note that the averages are only computed using training data, which are records that occurred before November 1, 2016.

In addition, a smoothing formula is applied such that the value is less variant and less affected by values calculated based on statistically insufficient observations in each category.

A list of all 305 variables is included in Appendix 2.

Part V. Feature Selection Process

After creating the 305 expert variables, an additional two variables were added as a check to ensure the appropriateness of the feature selection process: 1) actual fraud label, and 2) an independent random variable. If the feature selection process is accurate, the fraud label would be the top variable to be considered since it is the best variable for indicating fraudulent records, while the random variable should be relatively low in consideration. Then, all variables are standardized by Z-scaling before feature selection to make sure all variables are on a comparable scale.

With 307 variables created and standardized, the next step is to choose the top variables that explain the majority of the data variance to reduce dimensionality. This step benefits algorithm development in later stages to enable models to be trained faster, reduces the likelihood of overfitting, and tries to make the balance between complexity of variables and the total variance of data explained by the variables. Two measures are adopted and all variables are ranked accordingly:

1. **Kolmogorov-Smirnov (KS)**, a measure of how well two distributions are separated (in this case, fraud and regular applications). The higher the KS, the more separate the two distributions and the better the variable is as a predictor.
2. **Fraud detection rate (FDR)**, which is the relative percentage of all fraud identified at a specific cutoff location. In this case, a 3% cutoff was suggested by the subject matter expert, which means that the top 3% of rows will be used to calculate the fraud detection rate of that variable.

A detailed ranking of the variables can be found in Appendix 3. The rankings served as a primary filter for the feature selection process. Feature ranking with recursive feature elimination and cross-validated selection of the best number of features using logistic regression was used to narrow down what variables to include for model building. There is one purposefully created variable, which is a random number generated.

There are two variables that confirm the ranking rationale: the first one is the fraud_label, which is the dependent variable in the future model. In both KS and FDR rankings, it is the number one variable with perfect indication. The other one is the purposefully created artificial random number that does not relate to fraudulent application at all. This variable is ranked at 277th, which is relatively bottom to the list of the ranking. The two variables resulted in expected ranking, and based on the average ranking obtained from two measurements, it filtered out variables that did not explain the data variance as well.

The top 80 variables with the highest average KS and FDR ranking are used for further feature selection. A wrapper method called recursive feature elimination is used to find the best subset of variables to build algorithms. This is a process of repeatedly creating models and keeping the best or the worst performing feature in each iteration. Then it builds the next

model with the features left until all the features are exhausted. Finally, it will rank the features based on the order of their elimination. This selection process using the top ranked 80 variables is initially used as input for a desired top 30 variables.

However, the initial selected variables are far more than 30 variables, and with multiple selection processes, each selection yields slightly different results. This is due to the fact that among 307 variables that are created, some variables are highly correlated (for example, full address seen in the past 3 days and full address with the same date of birth in the past 3 days). Each time when there are highly correlated variables, the selection algorithm randomly selects one variable out of the correlated group. Depending on what prior variables are chosen, it will give different results, which would result in a drastically different variable list. A sample first feature selection result could be found in the appendix.

After the first feature selection, all variables that are ranked “1” indicate they are critical variables that explain the data variance. In the first recursive feature selection process, a total of 58 variables are identified to be important with the ranking at 1. In this case, it will be hard to tell from the 58 variables, which ones are more important than others.

Therefore, an additional recursive feature elimination was again used to find among the best previously identified 58 variables, and try to find which ones seemed to be more critical (have a ranking of “1”). If the number of top variables with the rank of “1” is more than the desired number of variables, 30, the recursive feature elimination is applied again until ranking provides an ideal number of variables.

After running recursive feature elimination three times, the final 30 features that are selected is listed below:

Table 5.1 Top 30 Feature Selected from Candidate Variables

Ranking	Variable
1	fulladdress_count1_date
1	fulladdress_count30_date
1	fulladdress_days_since_last_seen
1	fulladdress_nunique1_dob
1	fulladdresshomephone_count30_date
1	fulladdresshomephone_days_since_last_seen
1	namedob_count14_date
1	namedob_count30_date
1	namedob_days_since_last_seen
1	ssn_days_since_last_seen
1	ssndob_count30_date
1	ssndob_days_since_last_seen
1	ssnfname_days_since_last_seen
1	ssnfullname_count30_date
1	ssnfullname_days_since_last_seen
1	ssnlname_count30_date
1	ssnlname_days_since_last_seen
1	ssnnamedob_count30_date
1	ssnnamedob_days_since_last_seen
2	ssnlname_count180_date
3	fulladdresshomephone_count14_date
4	fulladdress_nunique3_ssn
5	address_days_since_last_seen
6	ssnfname_count14_date
7	ssnlname_count14_date
8	ssnfname_count30_date
9	ssnfullname_count180_date
10	ssnfname_count180_date
11	namedob_count180_date
12	ssnnamedob_count180_date

As mentioned before, each run of recursive feature elimination gives out different results each run. Another 20 features were selected with the same process to serve as another set of features in algorithm development state. It might be the case that a model with 20 variables could perform better than a model with 30 variables. A list of the selected 20 features is listed below:

Table 5.2 Another Set of Top 20 Feature Selected from Candidate Variables

Rank	Variable Name
1	fulladdress_count1_date
1	fulladdress_count30_date
1	fulladdress_days_since_last_seen
1	fulladdress_nunique1_dob
1	fulladdresshomephone_count30_date
1	fulladdresshomephone_days_since_last_seen
1	namedob_count14_date
1	namedob_count30_date
1	namedob_days_since_last_seen
1	ssn_days_since_last_seen
1	ssndob_count30_date
1	ssndob_days_since_last_seen
1	ssnfname_days_since_last_seen
1	ssnfullname_count30_date
1	ssnfullname_days_since_last_seen
1	ssnlname_count30_date
1	ssnlname_days_since_last_seen
1	ssnnamedob_count30_date
2	ssnnamedob_days_since_last_seen
3	ssnlname_count180_date

In either case, from the below Figure, it can be seen that majority of the data variance can be captured by 20-30 variables. Even if more variables were added, not much additional variance was explained. Therefore, the feature selection process effectively reduced dimension and only chose the ones that are vital.

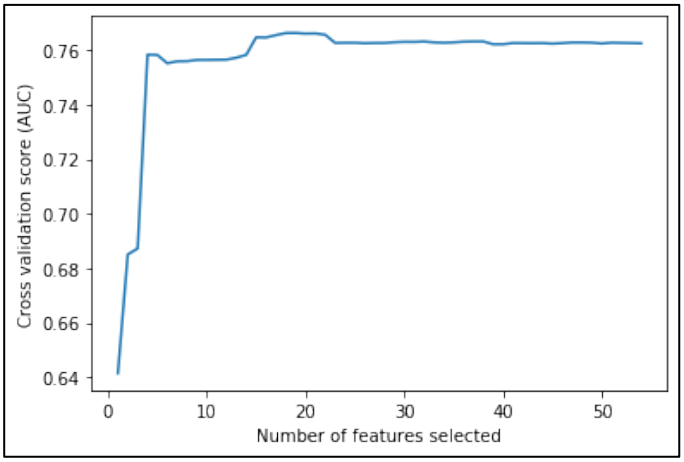


Figure 5.1 Variance Captured by Variables

Methods to Alleviate Unbalanced Data

Fraudulent applications tend to be a rare event within the dataset, since only 1.44% of the applications are labeled fraudulent. Therefore, regular model development might seem to be overly lenient when classifying fraud applications to achieve a higher accuracy score. To help alleviate this issue, two specific methods are adopted:

a. **Fraud Detection Rate (FDR) at 3%**

The model was used to predict the probability of a record to be fraudulent instead of giving out either 0 or 1 predictions (to classify if the records are fraud or not). The probability list was used to rank the data, from the records with the highest probability to the lowest. A suggested cutoff point was provided by the domain expert at 3%, and a count of fraud applications before the cutoff was used to calculate as a percentage of fraud caught over total fraud in the dataset.

b. **Synthetic Minority Oversampling (SMOTE)**

SMOTE is a statistical method to help simulate more minority cases, in this situation, the fraudulent records, while it does not change the number of majority cases. After re-sampling with SMOTE, the dataset achieved 1 fraud application to 1 regular application and is no longer imbalanced, and no information is lost during this process.

Part VI. Algorithm Development

With all the features selected, the team used a standardized method to develop possible algorithms. Out of time (OTT) data is excluded in development stage, which are the records after 10/31/2016. Only applications received from 1/15/2016 to 10/31/2016 were used to build models. After models have been developed, the FDR at 3% of the OTT data is used as an evaluation set to examine model performance.

An overview of a summary of the five algorithms used is listed below. All models are relatively similar on OOT data for FDR performance. Detailed individual model examination is listed after.

Table 6.1 Model Results Summary Table

Algorithm Name	Parameters	Training FDR at 3%	Testing FDR at 3%	OOT FDR at 3%
Logistic Regression (Baseline)	30 Variables	52.69%	53.94%	51.93%
Decision Tree	max_depth=6.5, min_samples_split=0.625	52.69%	53.53%	51.39%
Neural Net	30 Variables, 2 layers with 15 nodes each, Epochs =50, learning rate = 0.1	54.25%	54.59%	52.85%
Random Forest	25 Variables, 100 trees, depth=20	55.79%	55.13%	52.64%
Gradient Boosting Trees	30 Variables, 800 trees, depth=5	55.52%	55.13%	52.72%

Logistic Regression

The logistic regression model serves as a baseline model, and this model uses a logistic function to model the binary dependent variable, which is if the application is fraudulent or not. Different subsets are selected and used to try to find the best and minimal features that provide the best performance. By building the regression model with the selected 30 features, the average model performance has a 3% FDR at 52.69% on training data, 53.94% on test data, and out of sample FDR of 51.93%. A total of 8 sets of parameters are applied, and the default parameter has the highest FDR rate.

As a baseline model, the performance of logistic regression seems reasonable, as it was able to catch on average 51.93% of fraudulent applications at a 3% level. When considering other models with higher complexity that might result in a higher fraud detection rate, it is worth

noting that the ultimate best model should be the one that minimizes errors as well as complexity.

Table 6.2 Logistic Regression Result

Model	**Parameters**		**Average FDR (%) at 3%**		
Logistic Regression	Total Variables	Variables Selected	Train	Test	OOT
1	20	5 Variables	35.41%	37.09%	32.10%
2	20	10 Variables	50.93%	52.24%	48.99%
3	20	15 Variables	52.14%	53.50%	50.29%
4	20	20 Variables	52.23%	53.59%	50.38%
5	30	15 Variables	52.12%	53.55%	50.21%
6	30	20 Variables	52.65%	53.90%	50.63%
7	30	25 Variables	52.73%	53.90%	50.84%
8	30	30 Variables	52.69%	53.94%	**51.93%**

Decision Tree

A decision tree model builds a classification model in the form of a tree structure where smaller and smaller subsets are broken down with the final result being a tree with decision nodes and leaf nodes. After trying a total of 9 sets of parameters, the best performance model appears to be with the 30 pre-selected features. It was able to achieve an out of time data fraud detection rate of 51.39% at a 3% cutoff.

By comparing this model with the logistic model, it seems that logistic regression outperforms the decision tree model both by having a higher fraud detection rate and also a simpler model.

Table 6.3 Decision Tree Result

Model	**Parameters**		**Average FDR (%) at 3%**		
Decision Tree	Total Variables	Number of Variables Selected	Train	Test	OOT
1	20	Default, 10 Variables	52.41%	49.91%	47.78%
2	20	Default,15 Variables	53.62%	50.13%	48.15%
3	20	Default,20 Variables	53.70%	50.24%	48.22%
4	30	Default,15 Variables	53.62%	50.10%	48.09%
5	30	Default,20 Variables	53.69%	50.54%	48.15%
6	30	Default,30 Variables	54.10%	50.42%	47.56%
7	20	max_depth=6.5, min_samples_split=0.625	51.12%	51.01%	48.68%
8	30	max_depth=6.5, min_samples_split=0.625	53.56%	53.53%	**51.39%**
9	30	max_depth=6, min_samples_split=0.625	53.62%	53.45%	51.11%

Neural Net

A neural net is a model that consists of an input layer, a distinct number of hidden layers, and a final output layer. The hidden layers are sets of nodes that take the weighted signals of nodes from the layers before it and perform a transform on the linear combination of these signals.

The neural net model was performed on both the top 20 variables, as well as the top 30 variables. After analyzing from different sets of parameters, the best performing neural net model was the one that used the top 30 variables with two hidden layers of 15 nodes each, 50 epochs, and a learning rate of 0.001. This model has an average FDR at the top 3% of 52.85% for the out of time data.

By comparing this model with the logistic model and the decision tree model, it seems that the neural net model has a higher fraud detection rate, thereby outperforming them.

Table 6.4 Neural Net Result

Model	**Parameters**					**Average FDR (%) at 3%**		
NN	Total Variables	Layers	Nodes	Epochs	Learning Rate	Train	Test	OOT
1	20	1	6	N/A	N/A	50.31%	50.14%	48.01%
2	20	1	12	N/A	N/A	48.29%	47.92%	46.41%
3	20	1	20	N/A	N/A	51.53%	52.30%	49.00%
4	30	1	6	20	N/A	52.99%	52.95%	50.99%
5	30	1	20	50	N/A	54.20%	53.20%	52.55%
6	30	1	30	50	N/A	53.92%	55.02%	52.64%
7	30	2	(3,6)	20	0.1	54.42%	53.06%	52.37%
8	30	2	(15,15)	50	0.001	54.25%	54.59%	**52.85%**

Random Forest

A random forest model constructs multitudes of decision trees at training time and outputs the class that is the mode of the classes or mean prediction of the individual result of decision trees. After trying a total of 10 sets of parameters, the best performance model appears to be with the 30 pre-selected features. It was able to achieve an out of time data fraud detection rate of 52.64% at a 3% cutoff.

By comparing this model with the logistic model and the decision tree model, it seems that the random forest model has a higher fraud detection rate, as well as a higher model complexity.

Table 6.5 Random Forest Result

Model	**Parameters**		**Average FDR (%) at 3%**		
Random Forest	Total Variables	Number of Variables Selected	Train	Test	OOT
1	20	10 Variables, 50 trees, depth=20	55.31%	54.81%	51.84%
2	20	15 Variables, 50 trees, depth=20	55.39%	54.59%	51.80%
3	20	20 Variables, 50 trees, depth=20	55.47%	54.72%	52.23%
4	20	15 Variables, 100 trees, depth=20	55.62%	55.07%	52.37%
5	20	20 Variables, 100 trees, depth=20	55.79%	54.87%	52.45%
6	30	20 Variables, 50 trees, depth=20	55.20%	54.89%	52.33%
7	30	25 Variables, 50 trees, depth=20	55.32%	54.96%	52.31%
8	30	30 Variables, 50 trees, depth=20	55.56%	55.02%	52.41%
9	30	25 Variables, 100 trees, depth=20	55.79%	55.13%	**52.64%**
10	30	30 Variables, 100 trees, depth=20	55.53%	55.07%	52.48%

Gradient Boosting Trees

A gradient boosting tree model produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion and generalizes them by allowing optimization of an arbitrary differentiable loss function. After trying a total of 8 sets of parameters, the best performance model appears to be with the 30 pre-selected features. It was able to achieve an out of time data fraud detection rate of 52.72% at a 3% cutoff.

By comparing this model with all the models above, it seems that gradient boosting tree outperforms most of the models above, excluding the neural net model, by having a higher fraud detection rate at a 3% cutoff.

Table 6.6 Gradient Boosting Tree Result

Model	**Parameters**		**Average FDR (%) at 3%**		
Boosted Trees	Total Variables	Number of Variables Selected	Train	Test	OOT
1	20	600 trees, depth=4	54.67%	54.92%	52.37%
2	20	800 trees, depth=4	54.83%	55.03%	52.39%
3	20	600 trees, depth=5	55.12%	54.84%	52.67%
4	20	800 trees, depth=5	55.39%	54.96%	52.85%
5	30	600 trees, depth=4	54.75%	54.94%	52.13%
6	30	800 trees, depth=4	54.93%	55.12%	52.69%
7	30	600 trees, depth=5	55.23%	55.09%	52.63%
8	30	800 trees, depth=5	55.52%	55.13%	**52.72%**

Part VII. Results

The FDR performance at 3% on the out of time data was used to pick the best model with the highest percentage of fraud applications caught. Since neural network returned the highest average FDR out of all the models, this model is selected to be the final model with 30 variables as input, two hidden layers each with 15 nodes, and running at a learning rate of 0.001 with 50 epochs. The overall cumulative statistics for training, testing, and out of time data is presented below.

Table 7.1 Final Result on Training Dataset

Train	# Records	# Goods	# Bads	Fraud Rate
	635996	626807	9189	1.44%

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # of Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	6360	1639	4721	25.77%	74.23%	6360	1639	4721	0.26%	51.38%	51.12	0.35
2	6360	6163	197	96.90%	3.10%	12720	7802	4918	1.24%	53.52%	52.28	1.59
3	6360	6292	68	98.93%	1.07%	19080	14094	4986	2.25%	54.26%	52.01	2.83
4	6360	6310	50	99.21%	0.79%	25440	20404	5036	3.26%	54.80%	51.55	4.05
5	6360	6309	51	99.20%	0.80%	31800	26713	5087	4.26%	55.36%	51.10	5.25
6	6360	6318	42	99.34%	0.66%	38160	33031	5129	5.27%	55.82%	50.55	6.44
7	6360	6315	45	99.29%	0.71%	44520	39346	5174	6.28%	56.31%	50.03	7.60
8	6360	6303	57	99.10%	0.90%	50880	45649	5231	7.28%	56.93%	49.64	8.73
9	6360	6328	32	99.50%	0.50%	57240	51977	5263	8.29%	57.28%	48.98	9.88
10	6360	6317	43	99.32%	0.68%	63600	58294	5306	9.30%	57.74%	48.44	10.99
11	6360	6313	47	99.26%	0.74%	69960	64607	5353	10.31%	58.25%	47.95	12.07
12	6360	6320	40	99.37%	0.63%	76320	70927	5393	11.32%	58.69%	47.37	13.15
13	6360	6309	51	99.20%	0.80%	82680	77236	5444	12.32%	59.24%	46.92	14.19
14	6360	6314	46	99.28%	0.72%	89040	83550	5490	13.33%	59.75%	46.42	15.22
15	6360	6327	33	99.48%	0.52%	95400	89877	5523	14.34%	60.10%	45.77	16.27
16	6360	6322	38	99.40%	0.60%	101760	96199	5561	15.35%	60.52%	45.17	17.30
17	6360	6313	47	99.26%	0.74%	108120	102512	5608	16.35%	61.03%	44.67	18.28
18	6360	6315	45	99.29%	0.71%	114480	108827	5653	17.36%	61.52%	44.16	19.25
19	6360	6314	46	99.28%	0.72%	120840	115141	5699	18.37%	62.02%	43.65	20.20
20	6360	6319	41	99.36%	0.64%	127200	121460	5740	19.38%	62.47%	43.09	21.16

Table 7.2 Final Result on Testing Dataset

Test	# Records	# Goods	# Bads	Fraud Rate
	159000	156703	2297	1.44%

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # of Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	1590	401	1189	25.22%	74.78%	1590	401	1189	0.26%	51.76%	51.51	0.34
2	1590	1540	50	96.86%	3.14%	3180	1941	1239	1.24%	53.94%	52.70	1.57
3	1590	1575	15	99.06%	0.94%	4770	3516	1254	2.24%	54.59%	52.35	2.80
4	1590	1571	19	98.81%	1.19%	6360	5087	1273	3.25%	55.42%	52.17	4.00
5	1590	1580	10	99.37%	0.63%	7950	6667	1283	4.25%	55.86%	51.60	5.20
6	1590	1575	15	99.06%	0.94%	9540	8242	1298	5.26%	56.51%	51.25	6.35
7	1590	1575	15	99.06%	0.94%	11130	9817	1313	6.26%	57.16%	50.90	7.48
8	1590	1579	11	99.31%	0.69%	12720	11396	1324	7.27%	57.64%	50.37	8.61
9	1590	1576	14	99.12%	0.88%	14310	12972	1338	8.28%	58.25%	49.97	9.70
10	1590	1581	9	99.43%	0.57%	15900	14553	1347	9.29%	58.64%	49.35	10.80
11	1590	1580	10	99.37%	0.63%	17490	16133	1357	10.30%	59.08%	48.78	11.89
12	1590	1573	17	98.93%	1.07%	19080	17706	1374	11.30%	59.82%	48.52	12.89
13	1590	1578	12	99.25%	0.75%	20670	19284	1386	12.31%	60.34%	48.03	13.91
14	1590	1576	14	99.12%	0.88%	22260	20860	1400	13.31%	60.95%	47.64	14.90
15	1590	1578	12	99.25%	0.75%	23850	22438	1412	14.32%	61.47%	47.15	15.89
16	1590	1573	17	98.93%	1.07%	25440	24011	1429	15.32%	62.21%	46.89	16.80
17	1590	1580	10	99.37%	0.63%	27030	25591	1439	16.33%	62.65%	46.32	17.78
18	1590	1582	8	99.50%	0.50%	28620	27173	1447	17.34%	63.00%	45.65	18.78
19	1590	1580	10	99.37%	0.63%	30210	28753	1457	18.35%	63.43%	45.08	19.73
20	1590	1586	4	99.75%	0.25%	31800	30339	1461	19.36%	63.60%	44.24	20.77

Table 7.3 Final Result on OOT Dataset

OOT	# Records	# Goods	# Bads	Fraud Rate
	166493	164107	2386	1.43%

	Bin Statistics					Cumulative Statistics						
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # of Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	1665	495	1170	29.73%	70.27%	1665	495	1170	0.30%	49.04%	48.73	0.42
2	1665	1589	76	95.44%	4.56%	3330	2084	1246	1.27%	52.22%	50.95	1.67
3	1665	1650	15	99.10%	0.90%	4995	3734	1261	2.28%	52.85%	50.57	2.96
4	1665	1652	13	99.22%	0.78%	6660	5386	1274	3.28%	53.39%	50.11	4.23
5	1665	1657	8	99.52%	0.48%	8325	7043	1282	4.29%	53.73%	49.44	5.49
6	1665	1656	9	99.46%	0.54%	9990	8699	1291	5.30%	54.11%	48.81	6.74
7	1665	1650	15	99.10%	0.90%	11655	10349	1306	6.31%	54.74%	48.43	7.92
8	1665	1649	16	99.04%	0.96%	13320	11998	1322	7.31%	55.41%	48.10	9.08
9	1665	1653	12	99.28%	0.72%	14985	13651	1334	8.32%	55.91%	47.59	10.23
10	1665	1658	7	99.58%	0.42%	16650	15309	1341	9.33%	56.20%	46.87	11.42
11	1665	1650	15	99.10%	0.90%	18315	16959	1356	10.33%	56.83%	46.50	12.51
12	1665	1654	11	99.34%	0.66%	19980	18613	1367	11.34%	57.29%	45.95	13.62
13	1665	1655	10	99.40%	0.60%	21645	20268	1377	12.35%	57.71%	45.36	14.72
14	1665	1655	10	99.40%	0.60%	23310	21923	1387	13.36%	58.13%	44.77	15.81
15	1665	1651	14	99.16%	0.84%	24975	23574	1401	14.37%	58.72%	44.35	16.83
16	1665	1656	9	99.46%	0.54%	26640	25230	1410	15.37%	59.09%	43.72	17.89
17	1665	1654	11	99.34%	0.66%	28305	26884	1421	16.38%	59.56%	43.17	18.92
18	1665	1650	15	99.10%	0.90%	29970	28534	1436	17.39%	60.18%	42.80	19.87
19	1665	1654	11	99.34%	0.66%	31635	30188	1447	18.40%	60.65%	42.25	20.86
20	1665	1646	19	98.86%	1.14%	33300	31834	1466	19.40%	61.44%	42.04	21.71

Part VIII. Conclusion

In conclusion, the final model that was selected for fraud detection was a neural network model using 30 variables, two layers of 15 nodes each, 50 epochs, and a learning rate of 0.001. The steps that were taken to arrive at this final model were data cleaning, creating candidate variables, feature selection, creating various models, and then finally model selection.

The first step was data cleaning, in which frivolous data was replaced with the negative record number that corresponded to each application. This ensured that the data was completely unique and had never appeared before, while also replacing repetitive data that would negatively impact future work. Afterwards, candidate variables were created, which is a process that included first creating linking variables which were then used to create candidate variables. A total of 18 linking variables were created to append to the existing six fields, and from that 305 candidate variables were created. These variables included velocity variables, relative velocity variables, days since last seen variables, cross entity variables, and one risk variable.

With these 305 candidate variables, as well as the addition of a fraud label and random variable appended to analyze ranking accuracy, feature selection was then performed to find the best variables for the modeling. The Kolmogorov-Smirnov (KS) method and the Fraud Detection Rate (FDR) method were used to rank all of the 307 variables. Feature ranking with recursive feature elimination and cross-validated selection of the best number of features using logistic regression was used to narrow down to the final 30 variables that would be used for building models was created.

From these 30 features, various models were built. The baseline logistic regression model was created and was used later on to use as comparison against the other more complex models. With an average FDR of 51.93% at 3% level, the logistic regression model seems reasonable. Decision tree models were then developed, and the best parameters was able to achieve an out of time data fraud detection rate of 51.39%. Neural net with two layers of 15 nodes each, 50 epochs, and a learning rate of 0.001 achieved the highest out of time data fraud detection rate of 52.85%. Random forest models were also created, and the one that achieved the highest out of time data fraud detection rate of 52.64% with 100 trees and a depth of 20. Finally, gradient boosting tree with 800 trees and a depth of five was able to achieve an out of time data fraud detection rate of 52.72% at the same cutoff.

Therefore, a final neural net model was selected that had one with two layers of 15 nodes each, 50 epochs, and a learning rate of 0.001, since this model had the highest out of time fraud detection rate. While not as simple as a logistic regression model, this neural net model gives the highest fraud detection rate at a 3% cutoff, which is a worthwhile trade in exchange for having the simplest model possible.

It's worth noting that the current evaluation of the best models only looks at the prediction accuracy of fraudulent applications on out of time data. Because the final prediction results among all models seem relatively similar, it is advised that if the analysis team was given more time, a cost benefit analysis should be conducted to see which model will achieve the best result under the cost constraints that may be in place.

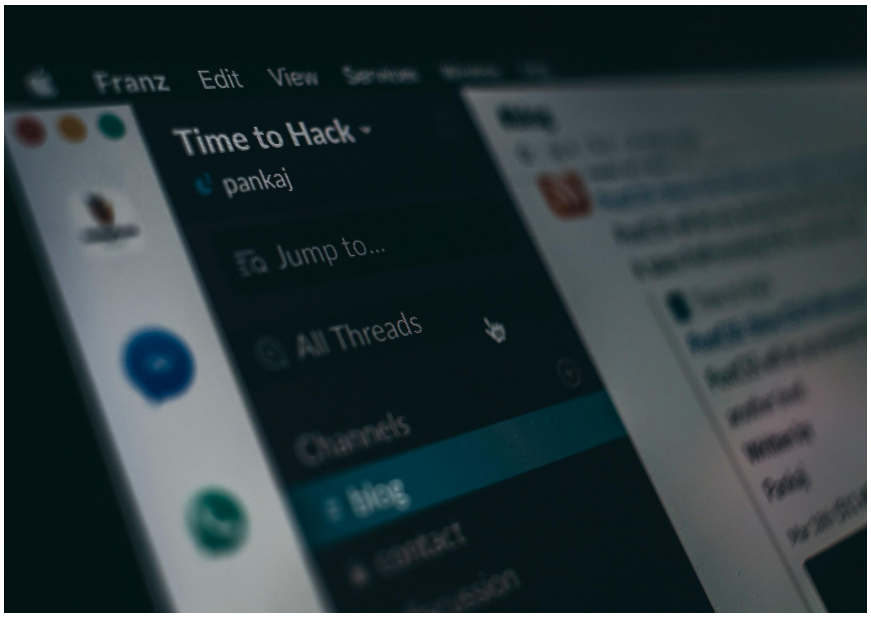
If more time was given, more models would be created to truly analyze which one will provide the best fraud detection rate at a 3% cutoff. For example, support vector machine (SVM) models, and Gaussian Naive Bayes models, could be possible options. In addition, only a rough search of best parameters for each model has been done when building models. Thus, it would be beneficial to perform detailed grid search to identify the bets parameters for each model. Finally, if given more time and access to domain experts, more expert variables could be created that better depict real life, as there are some expert variables that can only be created if one is or knows an expert on the subject.

Despite the limitations, the process of developing the most accurate algorithm with the product application data has been successful, as the team was able to come up with the final neural net model that achieves the fraud detection rate at 3% that was able to catch 52.85% of total fraudulent applications.

Appendix

USC Marshall
School of Business

Fraud Detection on Product Application Data



UNDER THE GUIDANCE OF

Professor Stephen Coggeshall

3/15/2020

JENNY SHANG |

JENNY WANG |

LINGROU WANG |

RORY WANG |

Table of Contents

Appendix 1. Data Quality Report

Description of Data

This report provides a background analysis of the "applications data" dataset, including summary statistics and visualizations of each of the fields contained in the dataset. The "applications data" is a dataset that is generated using an algorithm developed by one of Professor Coggeshall's colleagues. This dataset simulates a real-life business scenario where a company tracks and identifies fraudulent applications amongst normal, non-fraudulent applications. The label of whether a record is fraudulent or not helps in developing a supervised fraud detection model, which will be built in future stages. The data is generated and provided by Professor Coggeshall for academic purpose only.

In this dataset, the value for all fields are randomly generated using the algorithm and thus contains no real information. In summary, this dataset contains a total of 10 fields and 1,000,000 records. All records are generated with dates in 2016, from January 1, 2016 to December 31, 2016.

Summary of Fields

1. **Numerical Fields**

There are no numerical fields in this dataset.

All information is personal identification information, which usually is not continuous nor ordinal. The numbers present in certain fields, such as "ssn", are solely to identify individuals and have no meaning based on the size of these numbers.

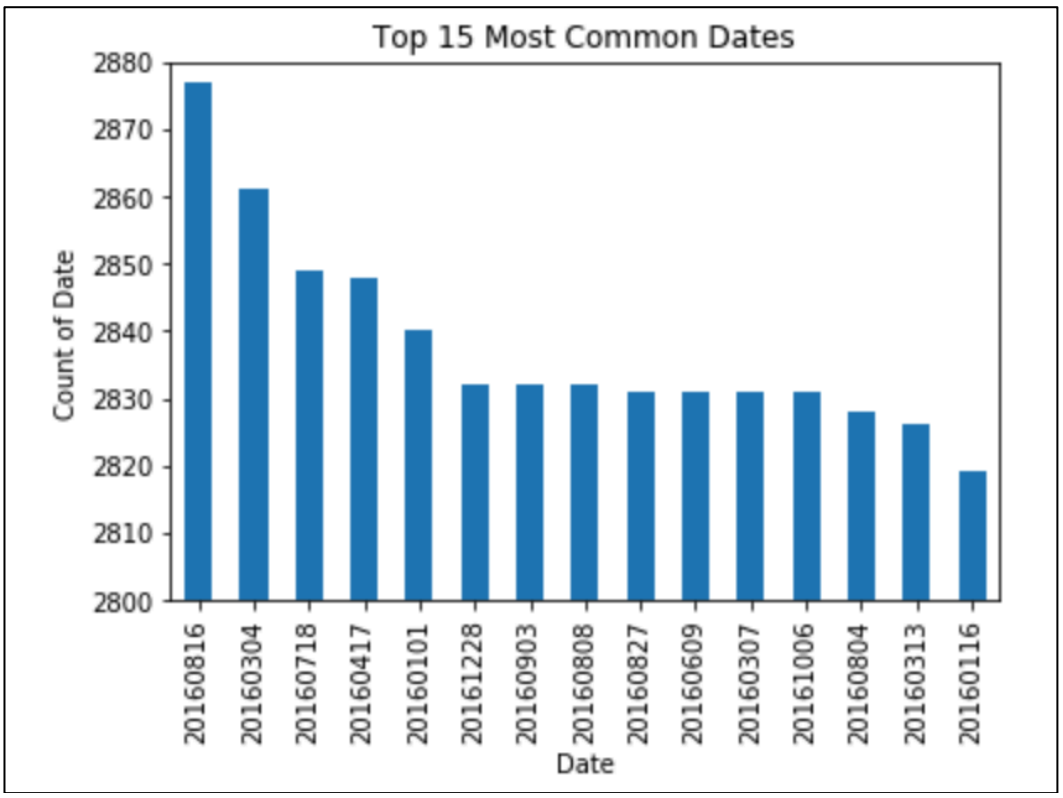
2. **Categorical Fields**

Field Name	# Records with Value	% Populated	# Unique Values	# Records with Value Zero	Most Common Value
record	1000000	100%	100000	0	
date	1000000	100%	365	0	20160816
ssn	1000000	100%	835819	0	999999999
firstname	1000000	100%	78136	0	EAMSTRMT
lastname	1000000	100%	177001	0	ERJSAXA
address	1000000	100%	828774	0	123 MAIN ST
zip5	1000000	100%	26370	0	68138
dob	1000000	100%	42673	0	19070626
homephone	1000000	100%	28244	0	9999999999
fraud_label	1000000	100%	2	985607	0

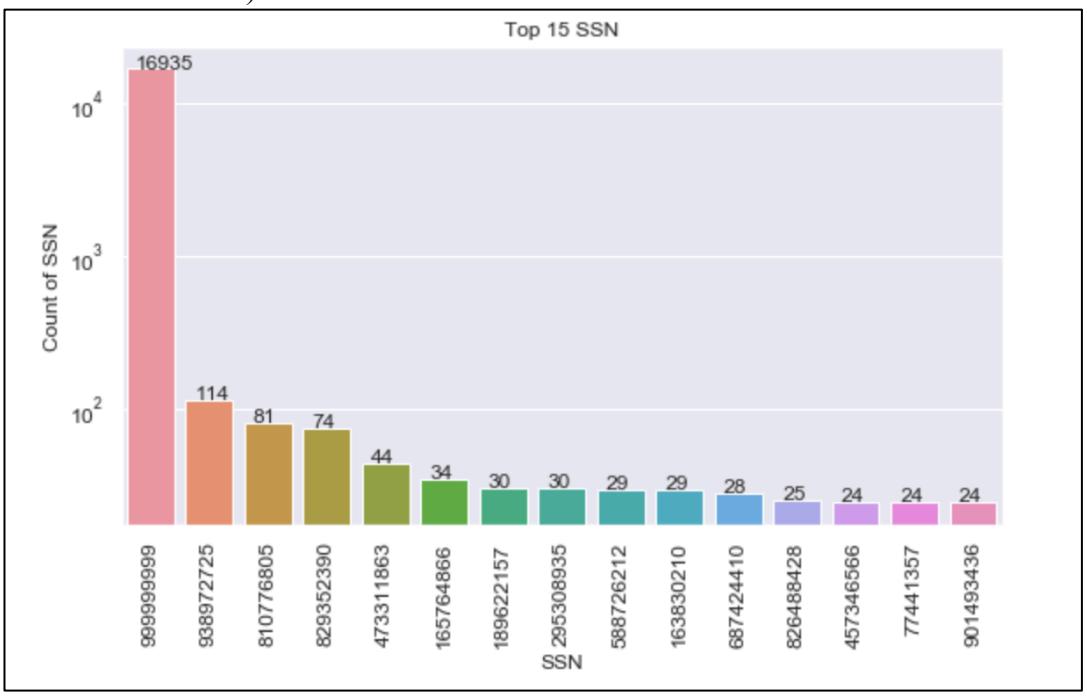
Description of Fields

1. **Field Name**: record
   Description: A unique identifier (index) given to each record entry. Since this is an identifier field, there is no need to create visualization on the distribution since each value will have a frequency count of 1.

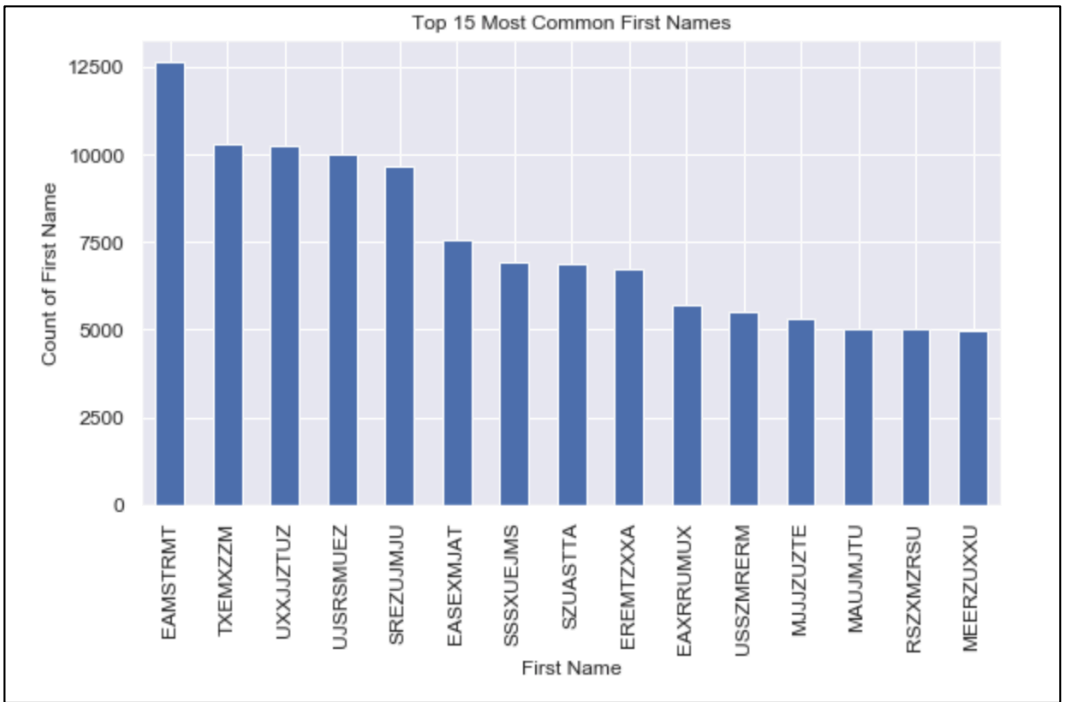
2. **Field Name**: date
   Description: This is a date field that tracks the year, date, and month of when the record is created (i.e. application filing date). The format in this field follows YYYYMMDD, all in numbers.



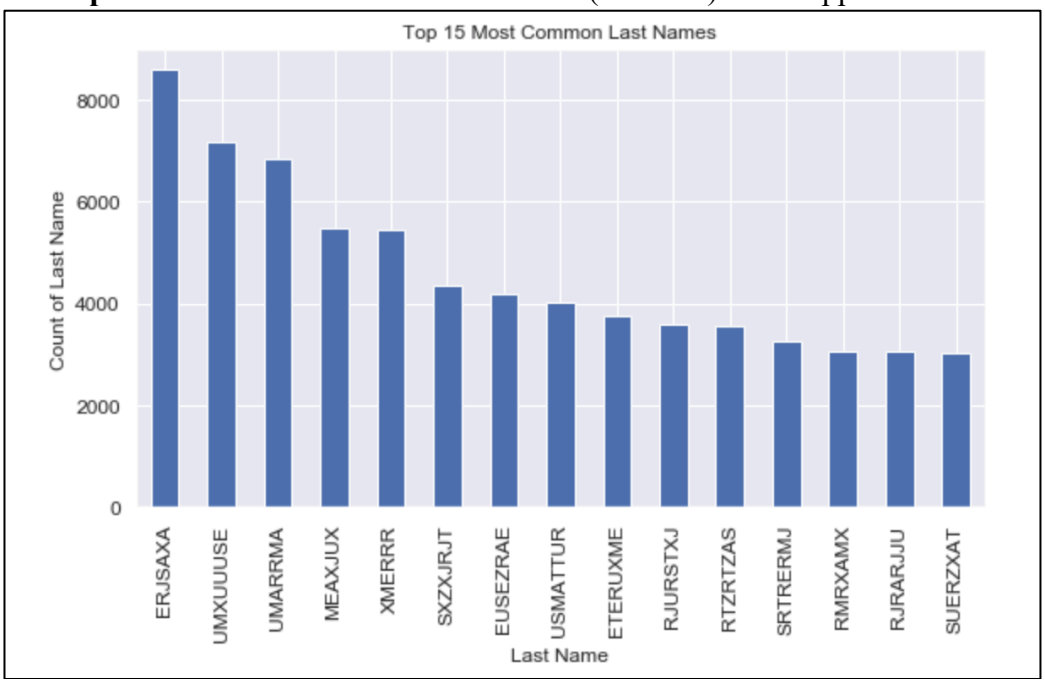
3. **Field Name**: ssn
   Description: This field tracks the Social Security Number (SSN) of the applicant. The SSN is a 9-digit numerical identification that is uniquely assigned by the U.S. federal government to U.S. citizens and some residents (e.g. foreigners who are legally able to work in the U.S.).



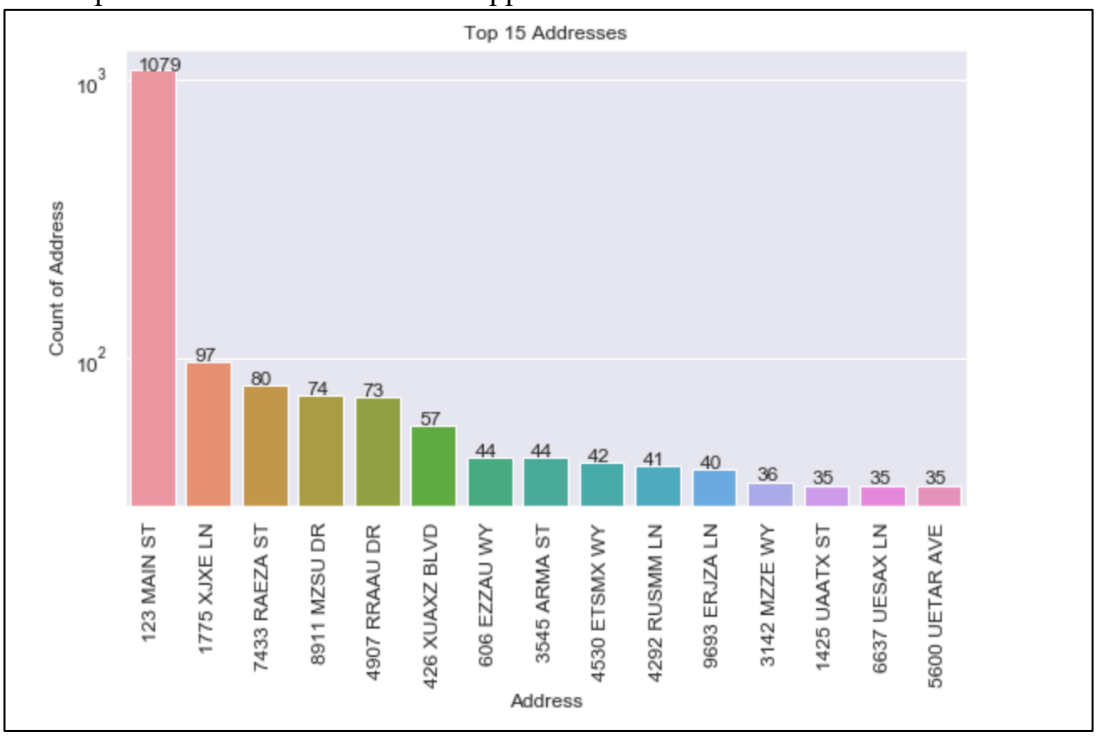
4. **Field Name**: firstname
   Description: This field records the first name (given name) of the applicant.



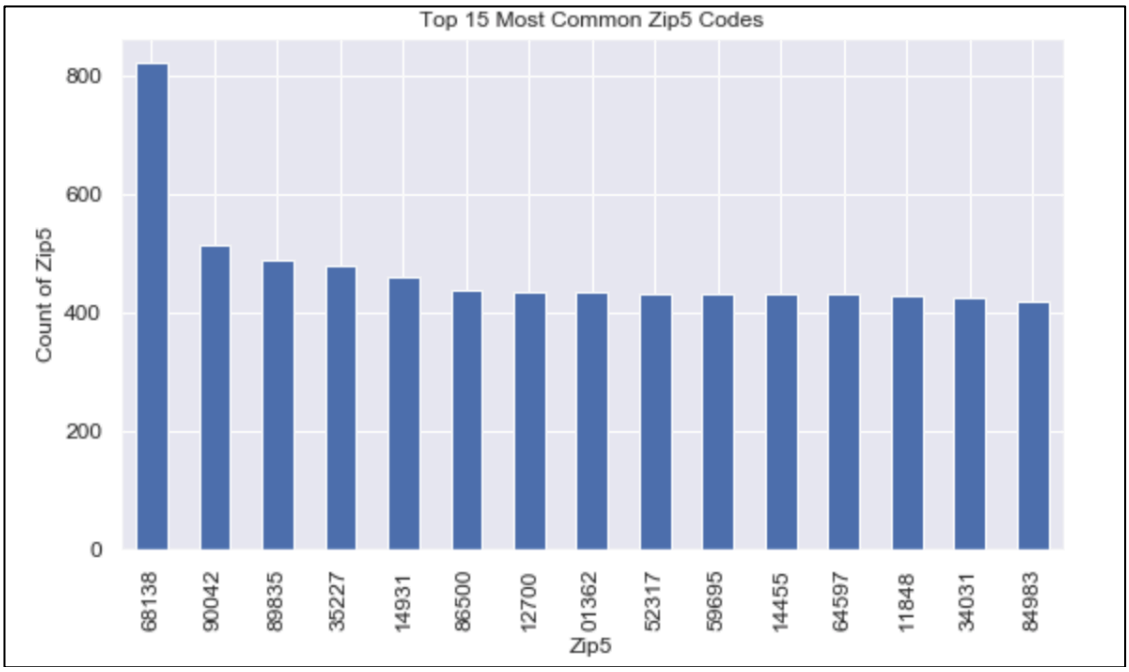
5. **Field Name**: lastname
   Description: This field records the last name (surname) of the applicant.



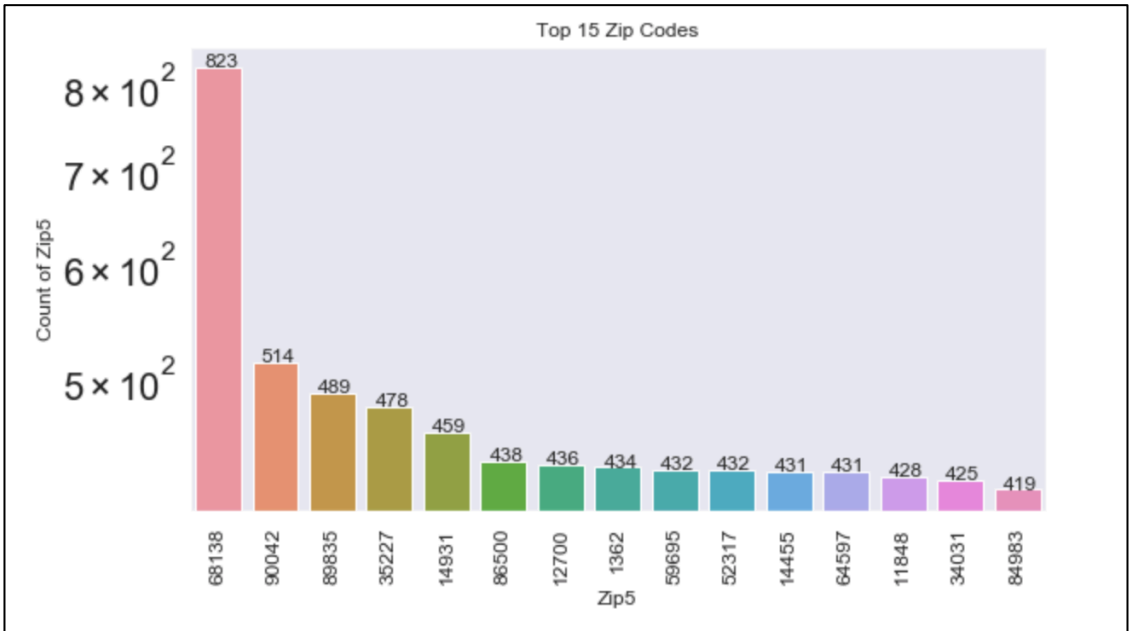
6. **Field Name**: address
   Description: This field tracks the address (street name and house number) of the applicant. Although there is no clear indication, it can be assumed that this field asks for the permanent address where the applicant resides.



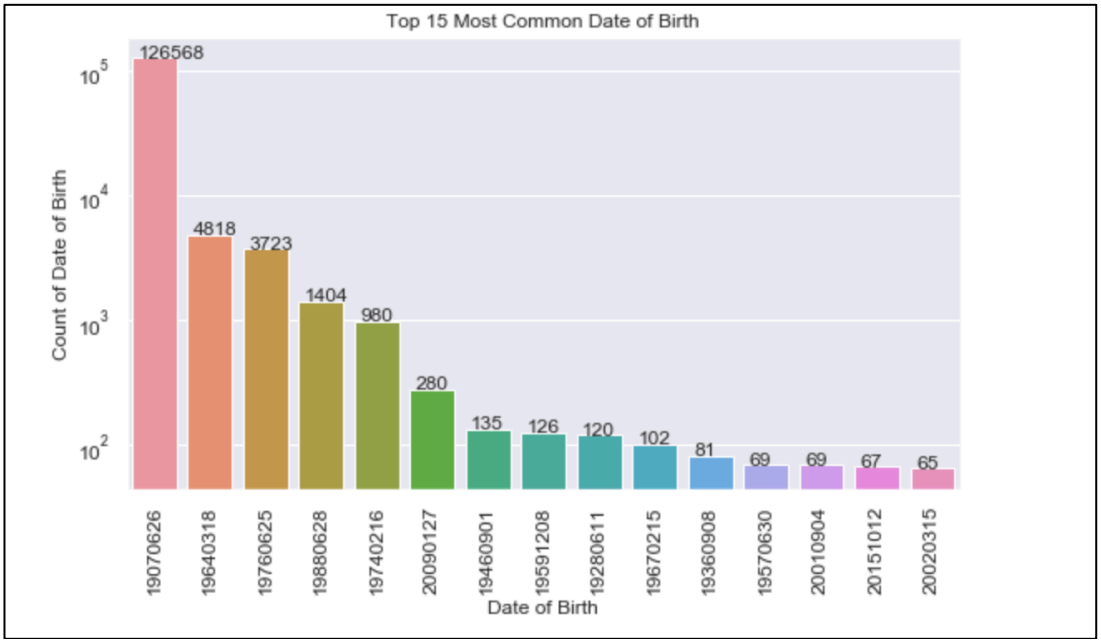
7. **Field Name**: zip5
Description: This field is the 5-digit zip code associated with the address that is entered in the previous field. In the original data, some zip5 values have less than 5 digits (e.g. 1362). These are zip codes that start with a "0". For the purpose of consistency, a leading 0 is added to these zip codes.



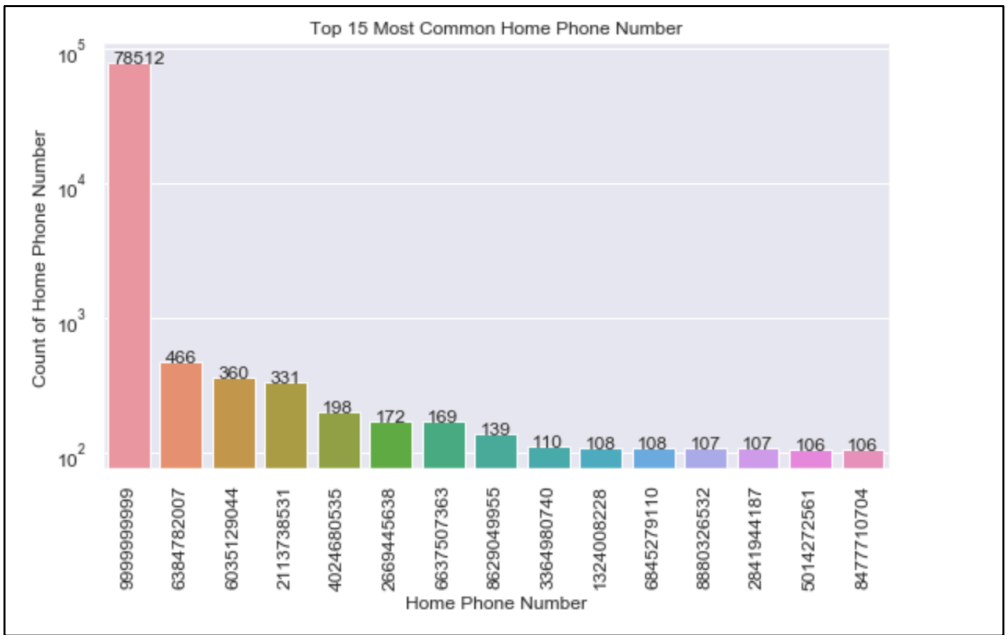
To better show the number of records for the top 15 zip codes:



8. **Field Name**: dob
Description: This field records the date of birth of the applicant. The format of values in the field follows YYYYMMDD, all in numbers.



9. **Field Name**: homephone
Description: This field records the home phone number of the applicant. The phone number is a 10-digit numeric number. The first 3 digits are the area code corresponding to a geographic area, the next 3 digits are the exchange that is associated with the carrier, and the last 4 are random numbers.



10. **Field Name**: fraud_label
    Description: This field identifies whether a record is a fraudulent record or not. A value of 0 indicates that the record is not a fraudulent record; whereas a value of 1 means that it is a fraudulent record.

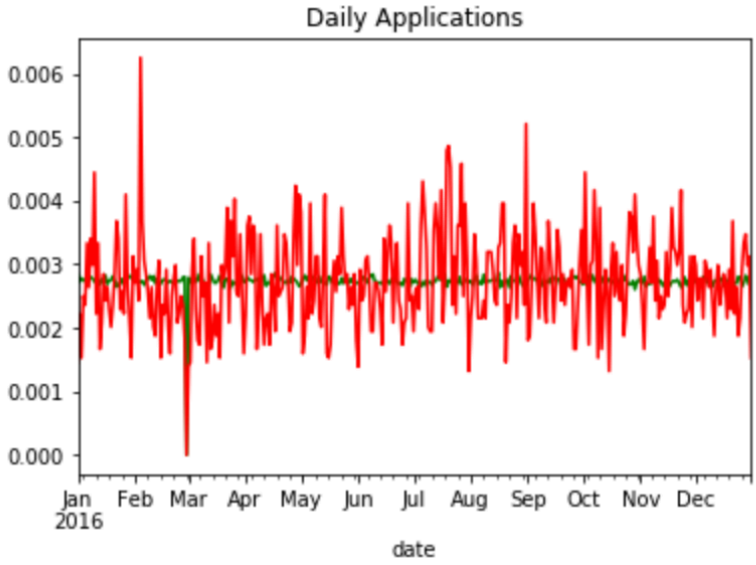
Fraud Label (0 = not fraud, 1 = fraud)	Count	% of Total
0	985,607	98.56%
1	14,393	1.44%

Additional Notes

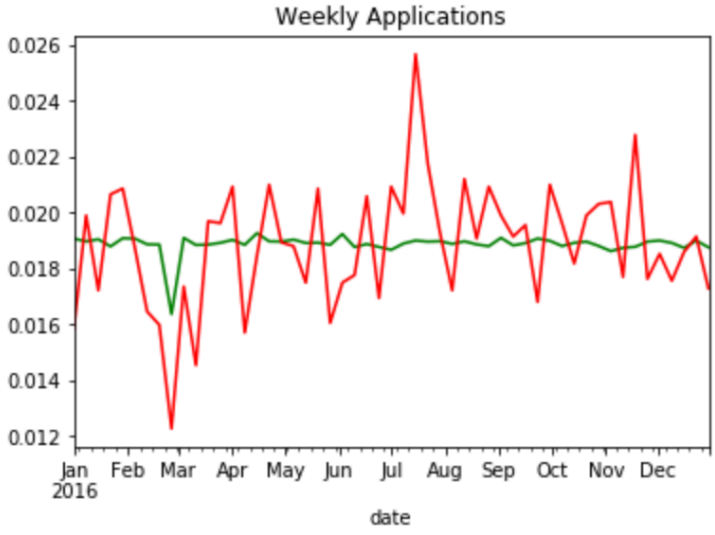
Based on the graphs, there are certain values that appear a lot more frequently than others. These values are likely to be frivolous values that were populated for the sake of completion rather than being the actual values populated by the applicant. One possible explanation may be that these are not mandatory fields and that the applicant left them blank. A summary of the frivolous values is provided below.

Field	Possible Frivolous Value	# of Frivolous Values that are Fraud	% of Frivolous Values that are Fraud
ssn	999999999	95	0.56%
address	123 MAIN ST	6	0.56%
dob	19070626	709	0.56%
homephone	9999999999	388	0.49%

In addition, the number of fraudulent applications vary by day and by week. From the below graphs, it is evident that the percentage of non-fraudulent applications per day over the year is at a steady rate of 0.03% of total non-fraudulent applications (as indicated by the green line); whereas the percentage of fraudulent applications per day has ups and downs with big spikes in February and during the summer months of July to September (as indicated by the red line). The sharp decrease near March is as a result of no data for February 29th.

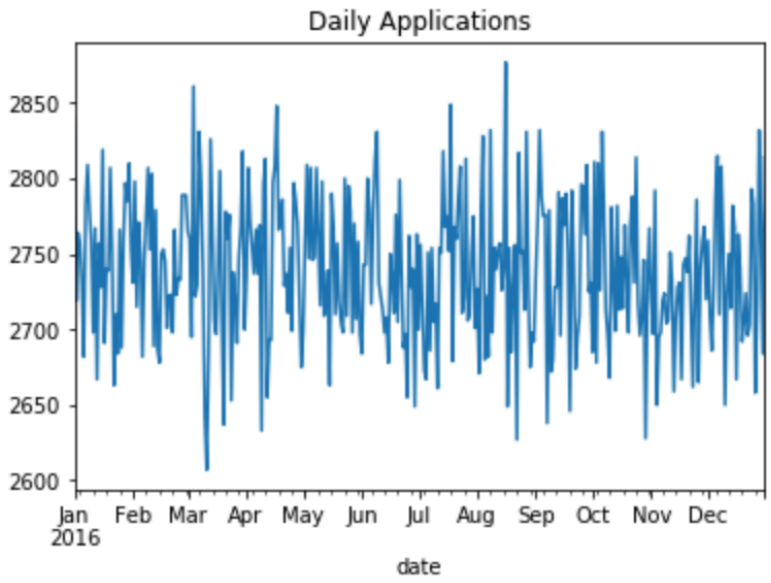


Looking at the percentage of fraudulent and non-fraudulent applications by week, the large spike in February is no longer evident but the spike in the summer months (July and August) persists. This implies that there is a particular day in February that has relatively more fraudulent applications than usual (likely 2016-02-04 with 90 fraudulent applications), while July and August consistently have more fraudulent applications than other times of the year. Note that since the last week of December only has two days, the data is modified to assume that it received the same number of applications as the week prior.



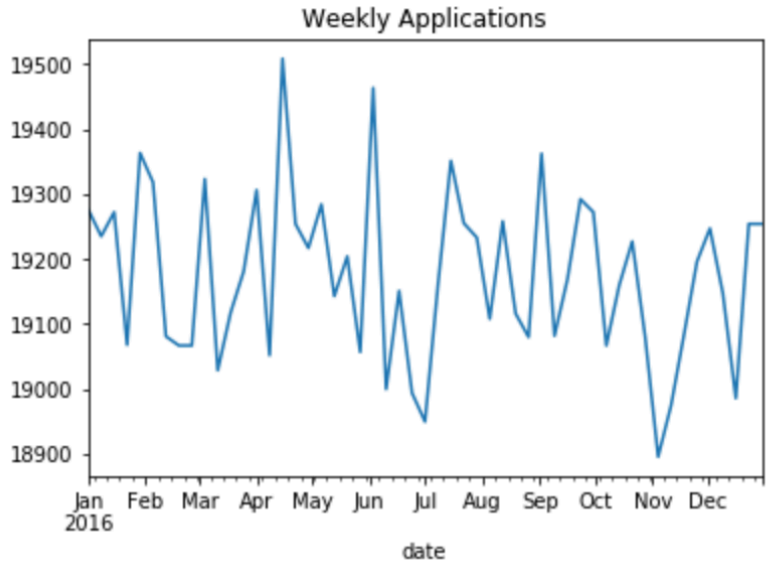
Number of applications (daily)

*To smooth curve, assume the number of applications for 2016-02-29 is the same as 2016-02-28.



Number of applications (weekly)

*To smooth curve, assume the number of applications for the week of 2016-12-30 is the same as the week of 2016-12-23.



Appendix 2. List of Variables Created

Variable Name	mean	std	min	max
ssn_count0_date	1.0073	0.2234	1.0000	21.0000
ssn_count1_date	1.0149	0.3812	1.0000	34.0000
ssn_count3_date	1.0201	0.4232	1.0000	34.0000
ssn_count7_date	1.0263	0.4536	1.0000	34.0000
ssn_count14_date	1.0346	0.4776	1.0000	34.0000
ssn_count30_date	1.0508	0.5133	1.0000	34.0000
ssn_count180_date	1.1647	0.8863	1.0000	63.0000
address_count0_date	1.0119	0.2833	1.0000	24.0000
address_count1_date	1.0247	0.4830	1.0000	30.0000
address_count3_date	1.0331	0.5443	1.0000	30.0000
address_count7_date	1.0425	0.5858	1.0000	30.0000
address_count14_date	1.0547	0.6180	1.0000	30.0000
address_count30_date	1.0782	0.6641	1.0000	30.0000
address_count180_date	1.2414	1.1481	1.0000	53.0000
dob_count0_date	1.0857	0.7307	1.0000	23.0000
dob_count1_date	1.2496	1.7895	1.0000	40.0000
dob_count3_date	1.5681	3.9687	1.0000	72.0000
dob_count7_date	2.1922	8.3315	1.0000	129.0000
dob_count14_date	3.2662	15.9025	1.0000	236.0000
dob_count30_date	5.6334	32.7863	1.0000	446.0000
dob_count180_date	22.4562	162.7984	1.0000	2439.0000
homephone_count0_date	1.0711	0.3648	1.0000	22.0000
homephone_count1_date	1.2017	0.6266	1.0000	31.0000
homephone_count3_date	1.4493	0.8616	1.0000	32.0000
homephone_count7_date	1.9331	1.2131	1.0000	33.0000
homephone_count14_date	2.7638	1.7441	1.0000	36.0000
homephone_count30_date	4.5975	2.8636	1.0000	50.0000
homephone_count180_date	17.6510	12.7395	1.0000	240.0000
namedob_count0_date	1.0072	0.2232	1.0000	21.0000
namedob_count1_date	1.0147	0.3809	1.0000	34.0000
namedob_count3_date	1.0196	0.4223	1.0000	34.0000
namedob_count7_date	1.0252	0.4513	1.0000	34.0000
namedob_count14_date	1.0324	0.4718	1.0000	34.0000
namedob_count30_date	1.0463	0.4969	1.0000	34.0000
namedob_count180_date	1.1439	0.6765	1.0000	39.0000
fulladdress_count0_date	1.0117	0.2828	1.0000	24.0000
fulladdress_count1_date	1.0239	0.4822	1.0000	30.0000
fulladdress_count3_date	1.0314	0.5424	1.0000	30.0000
fulladdress_count7_date	1.0390	0.5812	1.0000	30.0000
fulladdress_count14_date	1.0480	0.6069	1.0000	30.0000
fulladdress_count30_date	1.0646	0.6338	1.0000	30.0000
fulladdress_count180_date	1.1785	0.8015	1.0000	39.0000
ssnfname_count0_date	1.0073	0.2232	1.0000	21.0000
ssnfname_count1_date	1.0148	0.3810	1.0000	34.0000
ssnfname_count3_date	1.0199	0.4226	1.0000	34.0000
ssnfname_count7_date	1.0258	0.4519	1.0000	34.0000
ssnfname_count14_date	1.0336	0.4729	1.0000	34.0000
ssnfname_count30_date	1.0489	0.4990	1.0000	34.0000
ssnfname_count180_date	1.1559	0.6804	1.0000	39.0000
ssnlname_count0_date	1.0073	0.2232	1.0000	21.0000
ssnlname_count1_date	1.0148	0.3810	1.0000	34.0000
ssnlname_count3_date	1.0199	0.4226	1.0000	34.0000
ssnlname_count7_date	1.0258	0.4519	1.0000	34.0000
ssnlname_count14_date	1.0336	0.4729	1.0000	34.0000
ssnlname_count30_date	1.0489	0.4990	1.0000	34.0000
ssnlname_count180_date	1.1558	0.6803	1.0000	39.0000

ssnfullname_count0_date	1.0073	0.2232	1.0000	21.0000
ssnfullname_count1_date	1.0148	0.3810	1.0000	34.0000
ssnfullname_count3_date	1.0199	0.4226	1.0000	34.0000
ssnfullname_count7_date	1.0258	0.4519	1.0000	34.0000
ssnfullname_count14_date	1.0336	0.4729	1.0000	34.0000
ssnfullname_count30_date	1.0488	0.4989	1.0000	34.0000
ssnfullname_count180_date	1.1554	0.6796	1.0000	39.0000
ssnaddress_count0_date	1.0005	0.0227	1.0000	3.0000
ssnaddress_count1_date	1.0015	0.0384	1.0000	3.0000
ssnaddress_count3_date	1.0035	0.0595	1.0000	5.0000
ssnaddress_count7_date	1.0073	0.0874	1.0000	6.0000
ssnaddress_count14_date	1.0139	0.1222	1.0000	8.0000
ssnaddress_count30_date	1.0284	0.1815	1.0000	11.0000
ssnaddress_count180_date	1.1328	0.4897	1.0000	39.0000
ssnzip5_count0_date	1.0005	0.0227	1.0000	3.0000
ssnzip5_count1_date	1.0015	0.0384	1.0000	3.0000
ssnzip5_count3_date	1.0035	0.0595	1.0000	5.0000
ssnzip5_count7_date	1.0073	0.0875	1.0000	6.0000
ssnzip5_count14_date	1.0139	0.1223	1.0000	8.0000
ssnzip5_count30_date	1.0285	0.1817	1.0000	11.0000
ssnzip5_count180_date	1.1332	0.4900	1.0000	39.0000
ssndob_count0_date	1.0072	0.2232	1.0000	21.0000
ssndob_count1_date	1.0147	0.3809	1.0000	34.0000
ssndob_count3_date	1.0196	0.4223	1.0000	34.0000
ssndob_count7_date	1.0251	0.4512	1.0000	34.0000
ssndob_count14_date	1.0323	0.4716	1.0000	34.0000
ssndob_count30_date	1.0461	0.4963	1.0000	34.0000
ssndob_count180_date	1.1428	0.6702	1.0000	39.0000
ssnhomephone_count0_date	1.0005	0.0222	1.0000	3.0000
ssnhomephone_count1_date	1.0014	0.0376	1.0000	3.0000
ssnhomephone_count3_date	1.0033	0.0582	1.0000	5.0000
ssnhomephone_count7_date	1.0070	0.0856	1.0000	6.0000
ssnhomephone_count14_date	1.0133	0.1198	1.0000	8.0000
ssnhomephone_count30_date	1.0272	0.1782	1.0000	11.0000
ssnhomephone_count180_date	1.1273	0.4856	1.0000	39.0000
ssnnamedob_count0_date	1.0072	0.2232	1.0000	21.0000
ssnnamedob_count1_date	1.0147	0.3808	1.0000	34.0000
ssnnamedob_count3_date	1.0196	0.4222	1.0000	34.0000
ssnnamedob_count7_date	1.0251	0.4511	1.0000	34.0000
ssnnamedob_count14_date	1.0322	0.4715	1.0000	34.0000
ssnnamedob_count30_date	1.0459	0.4961	1.0000	34.0000
ssnnamedob_count180_date	1.1421	0.6691	1.0000	39.0000
ssnfulladdress_count0_date	1.0005	0.0226	1.0000	3.0000
ssnfulladdress_count1_date	1.0015	0.0384	1.0000	3.0000
ssnfulladdress_count3_date	1.0035	0.0594	1.0000	5.0000
ssnfulladdress_count7_date	1.0073	0.0873	1.0000	6.0000
ssnfulladdress_count14_date	1.0139	0.1220	1.0000	8.0000
ssnfulladdress_count30_date	1.0283	0.1813	1.0000	11.0000
ssnfulladdress_count180_date	1.1325	0.4891	1.0000	39.0000
namehomephone_count0_date	1.0005	0.0224	1.0000	3.0000
namehomephone_count1_date	1.0014	0.0377	1.0000	3.0000
namehomephone_count3_date	1.0033	0.0584	1.0000	5.0000
namehomephone_count7_date	1.0070	0.0858	1.0000	6.0000
namehomephone_count14_date	1.0134	0.1200	1.0000	8.0000
namehomephone_count30_date	1.0273	0.1784	1.0000	11.0000
namehomephone_count180_date	1.1281	0.4880	1.0000	39.0000
namefulladdress_count0_date	1.0005	0.0228	1.0000	3.0000

namefulladdress_count1_date	1.0015	0.0385	1.0000	3.0000
namefulladdress_count3_date	1.0035	0.0597	1.0000	5.0000
namefulladdress_count7_date	1.0073	0.0877	1.0000	6.0000
namefulladdress_count14_date	1.0140	0.1227	1.0000	8.0000
namefulladdress_count30_date	1.0287	0.1825	1.0000	11.0000
namefulladdress_count180_date	1.1342	0.4953	1.0000	39.0000
fulladdressdob_count0_date	1.0005	0.0218	1.0000	3.0000
fulladdressdob_count1_date	1.0013	0.0368	1.0000	3.0000
fulladdressdob_count3_date	1.0032	0.0570	1.0000	5.0000
fulladdressdob_count7_date	1.0067	0.0838	1.0000	6.0000
fulladdressdob_count14_date	1.0127	0.1175	1.0000	8.0000
fulladdressdob_count30_date	1.0260	0.1751	1.0000	11.0000
fulladdressdob_count180_date	1.1218	0.4803	1.0000	39.0000
fulladdresshomephone_count0_date	1.0074	0.2237	1.0000	21.0000
fulladdresshomephone_count1_date	1.0152	0.3783	1.0000	30.0000
fulladdresshomephone_count3_date	1.0205	0.4257	1.0000	30.0000
fulladdresshomephone_count7_date	1.0263	0.4591	1.0000	30.0000
fulladdresshomephone_count14_date	1.0338	0.4820	1.0000	30.0000
fulladdresshomephone_count30_date	1.0482	0.5077	1.0000	30.0000
fulladdresshomephone_count180_date	1.1491	0.6843	1.0000	39.0000
fulladdressnamedob_count0_date	1.0005	0.0218	1.0000	3.0000
fulladdressnamedob_count1_date	1.0013	0.0366	1.0000	3.0000
fulladdressnamedob_count3_date	1.0032	0.0569	1.0000	5.0000
fulladdressnamedob_count7_date	1.0066	0.0836	1.0000	6.0000
fulladdressnamedob_count14_date	1.0127	0.1171	1.0000	8.0000
fulladdressnamedob_count30_date	1.0259	0.1745	1.0000	11.0000
fulladdressnamedob_count180_date	1.1213	0.4792	1.0000	39.0000
dobhomephone_count0_date	1.0005	0.0218	1.0000	3.0000
dobhomephone_count1_date	1.0013	0.0367	1.0000	3.0000
dobhomephone_count3_date	1.0032	0.0569	1.0000	5.0000
dobhomephone_count7_date	1.0066	0.0835	1.0000	6.0000
dobhomephone_count14_date	1.0126	0.1169	1.0000	8.0000
dobhomephone_count30_date	1.0258	0.1741	1.0000	11.0000
dobhomephone_count180_date	1.1209	0.4798	1.0000	39.0000
homephonenamedob_count0_date	1.0005	0.0216	1.0000	3.0000
homephonenamedob_count1_date	1.0013	0.0363	1.0000	3.0000
homephonenamedob_count3_date	1.0031	0.0564	1.0000	5.0000
homephonenamedob_count7_date	1.0065	0.0828	1.0000	6.0000
homephonenamedob_count14_date	1.0124	0.1160	1.0000	8.0000
homephonenamedob_count30_date	1.0253	0.1729	1.0000	11.0000
homephonenamedob_count180_date	1.1190	0.4770	1.0000	39.0000
ssn_1_count_ssn_3_count_Ave	2.9953	0.0837	0.4286	3.0000
ssn_1_count_ssn_7_count_Ave	6.9734	0.3097	0.5000	7.0000
ssn_1_count_ssn_14_count_Ave	13.8989	0.8532	0.8750	14.0000
ssn_1_count_ssn_30_count_Ave	29.5693	2.5568	1.7647	30.0000
address_1_count_address_3_count_Ave	2.9929	0.1022	0.4286	3.0000
address_1_count_address_7_count_Ave	6.9612	0.3749	0.5833	7.0000
address_1_count_address_14_count_Ave	13.8564	1.0227	0.9333	14.0000
address_1_count_address_30_count_Ave	29.4135	3.0193	1.3043	30.0000
dob_1_count_dob_3_count_Ave	2.8443	0.4631	0.0811	3.0000
dob_1_count_dob_7_count_Ave	6.0604	1.6438	0.0921	7.0000
dob_1_count_dob_14_count_Ave	10.5213	4.0143	0.0859	14.0000
dob_1_count_dob_30_count_Ave	17.4221	9.2911	0.0763	30.0000
homephone_1_count_homephone_3_count_Ave	2.6916	0.6121	0.1200	3.0000
homephone_1_count_homephone_7_count_Ave	5.1955	1.9671	0.2121	7.0000
homephone_1_count_homephone_14_count_Ave	7.9811	4.2182	0.4118	14.0000
homephone_1_count_homephone_30_count_Ave	11.6649	8.7628	0.6522	30.0000

namedob_1_count_namedob_3_count_Ave	2.9957	0.0799	0.4286	3.0000
namedob_1_count_namedob_7_count_Ave	6.9762	0.2933	0.5000	7.0000
namedob_1_count_namedob_14_count_Ave	13.9097	0.8055	0.8750	14.0000
namedob_1_count_namedob_30_count_Ave	29.6169	2.4110	1.7647	30.0000
fulladdress_1_count_fulladdress_3_count_Ave	2.9942	0.0920	0.4286	3.0000
fulladdress_1_count_fulladdress_7_count_Ave	6.9697	0.3313	0.5833	7.0000
fulladdress_1_count_fulladdress_14_count_Ave	13.8892	0.8942	0.9333	14.0000
fulladdress_1_count_fulladdress_30_count_Ave	29.5398	2.6419	1.3043	30.0000
ssnfname_1_count_ssnfname_3_count_Ave	2.9954	0.0822	0.4286	3.0000
ssnfname_1_count_ssnfname_7_count_Ave	6.9743	0.3041	0.5000	7.0000
ssnfname_1_count_ssnfname_14_count_Ave	13.9017	0.8385	0.8750	14.0000
ssnfname_1_count_ssnfname_30_count_Ave	29.5793	2.5191	1.7647	30.0000
ssnlname_1_count_ssnlname_3_count_Ave	2.9954	0.0822	0.4286	3.0000
ssnlname_1_count_ssnlname_7_count_Ave	6.9743	0.3039	0.5000	7.0000
ssnlname_1_count_ssnlname_14_count_Ave	13.9019	0.8380	0.8750	14.0000
ssnlname_1_count_ssnlname_30_count_Ave	29.5794	2.5189	1.7647	30.0000
ssnfullname_1_count_ssnfullname_3_count_Ave	2.9954	0.0821	0.4286	3.0000
ssnfullname_1_count_ssnfullname_7_count_Ave	6.9744	0.3035	0.5000	7.0000
ssnfullname_1_count_ssnfullname_14_count_Ave	13.9021	0.8369	0.8750	14.0000
ssnfullname_1_count_ssnfullname_30_count_Ave	29.5810	2.5143	1.7647	30.0000
ssnaddress_1_count_ssnaddress_3_count_Ave	2.9970	0.0667	0.7500	3.0000
ssnaddress_1_count_ssnaddress_7_count_Ave	6.9799	0.2649	1.1667	7.0000
ssnaddress_1_count_ssnaddress_14_count_Ave	13.9150	0.7707	1.7500	14.0000
ssnaddress_1_count_ssnaddress_30_count_Ave	29.6118	2.4049	2.7273	30.0000
ssnzip5_1_count_ssnzip5_3_count_Ave	2.9970	0.0668	0.7500	3.0000
ssnzip5_1_count_ssnzip5_7_count_Ave	6.9799	0.2654	1.1667	7.0000
ssnzip5_1_count_ssnzip5_14_count_Ave	13.9147	0.7720	1.7500	14.0000
ssnzip5_1_count_ssnzip5_30_count_Ave	29.6107	2.4082	2.7273	30.0000
ssndob_1_count_ssndob_3_count_Ave	2.9957	0.0797	0.4286	3.0000
ssndob_1_count_ssndob_7_count_Ave	6.9763	0.2926	0.5000	7.0000
ssndob_1_count_ssndob_14_count_Ave	13.9102	0.8031	0.8750	14.0000
ssndob_1_count_ssndob_30_count_Ave	29.6186	2.4045	1.7647	30.0000
ssnhomephone_1_count_ssnhomephone_3_count_Ave	2.9971	0.0654	0.7500	3.0000
ssnhomephone_1_count_ssnhomephone_7_count_Ave	6.9808	0.2596	1.1667	7.0000
ssnhomephone_1_count_ssnhomephone_14_count_Ave	13.9186	0.7545	1.7500	14.0000
ssnhomephone_1_count_ssnhomephone_30_count_Ave	29.6292	2.3528	2.7273	30.0000
ssnnamedob_1_count_ssnnamedob_3_count_Ave	2.9957	0.0796	0.4286	3.0000
ssnnamedob_1_count_ssnnamedob_7_count_Ave	6.9764	0.2920	0.5000	7.0000
ssnnamedob_1_count_ssnnamedob_14_count_Ave	13.9106	0.8014	0.8750	14.0000
ssnnamedob_1_count_ssnnamedob_30_count_Ave	29.6208	2.3981	1.7647	30.0000
ssnfulladdress_1_count_ssnfulladdress_3_count_Ave	2.9970	0.0666	0.7500	3.0000
ssnfulladdress_1_count_ssnfulladdress_7_count_Ave	6.9800	0.2646	1.1667	7.0000
ssnfulladdress_1_count_ssnfulladdress_14_count_Ave	13.9152	0.7697	1.7500	14.0000
ssnfulladdress_1_count_ssnfulladdress_30_count_Ave	29.6130	2.4013	2.7273	30.0000
namehomephone_1_count_namehomephone_3_count_Ave	2.9971	0.0655	0.7500	3.0000
namehomephone_1_count_namehomephone_7_count_Ave	6.9807	0.2599	1.1667	7.0000
namehomephone_1_count_namehomephone_14_count_Ave	13.9183	0.7561	1.7500	14.0000
namehomephone_1_count_namehomephone_30_count_Ave	29.6280	2.3565	2.7273	30.0000
namefulladdress_1_count_namefulladdress_3_count_Ave	2.9970	0.0671	0.7500	3.0000
namefulladdress_1_count_namefulladdress_7_count_Ave	6.9798	0.2659	1.1667	7.0000
namefulladdress_1_count_namefulladdress_14_count_Ave	13.9143	0.7740	1.7500	14.0000
namefulladdress_1_count_namefulladdress_30_count_Ave	29.6087	2.4148	2.7273	30.0000
fulladdressdob_1_count_fulladdressdob_3_count_Ave	2.9973	0.0640	0.7500	3.0000
fulladdressdob_1_count_fulladdressdob_7_count_Ave	6.9816	0.2536	1.1667	7.0000
fulladdressdob_1_count_fulladdressdob_14_count_Ave	13.9222	0.7381	1.7500	14.0000
fulladdressdob_1_count_fulladdressdob_30_count_Ave	29.6459	2.3012	2.7273	30.0000
fulladdresshomephone_1_count_fulladdresshomephone_3_count_Ave	2.9955	0.0812	0.5000	3.0000

fulladdresshomephone_1_count_fulladdresshomephone_7_count_Ave	6.9754	0.2978	0.5833	7.0000
fulladdresshomephone_1_count_fulladdresshomephone_14_count_Ave	13.9066	0.8184	0.9333	14.0000
fulladdresshomephone_1_count_fulladdresshomephone_30_count_Ave	29.6029	2.4515	1.3043	30.0000
fulladdressnamedob_1_count_fulladdressnamedob_3_count_Ave	2.9973	0.0640	0.7500	3.0000
fulladdressnamedob_1_count_fulladdressnamedob_7_count_Ave	6.9817	0.2529	1.1667	7.0000
fulladdressnamedob_1_count_fulladdressnamedob_14_count_Ave	13.9226	0.7362	1.7500	14.0000
fulladdressnamedob_1_count_fulladdressnamedob_30_count_Ave	29.6477	2.2952	2.7273	30.0000
dobhomephone_1_count_dobhomephone_3_count_Ave	2.9973	0.0640	0.7500	3.0000
dobhomephone_1_count_dobhomephone_7_count_Ave	6.9818	0.2524	1.1667	7.0000
dobhomephone_1_count_dobhomephone_14_count_Ave	13.9229	0.7347	1.7500	14.0000
dobhomephone_1_count_dobhomephone_30_count_Ave	29.6491	2.2907	2.7273	30.0000
homephonenamedob_1_count_homephonenamedob_3_count_Ave	2.9973	0.0634	0.7500	3.0000
homephonenamedob_1_count_homephonenamedob_7_count_Ave	6.9821	0.2506	1.1667	7.0000
homephonenamedob_1_count_homephonenamedob_14_count_Ave	13.9240	0.7294	1.7500	14.0000
homephonenamedob_1_count_homephonenamedob_30_count_Ave	29.6547	2.2730	2.7273	30.0000
ssn_days_since_last_seen	327.9537	96.9799	0.0000	366.0000
address_days_since_last_seen	320.9858	105.1350	0.0000	366.0000
dob_days_since_last_seen	75.9948	131.8574	0.0000	366.0000
homephone_days_since_last_seen	47.4864	111.0105	0.0000	366.0000
namedob_days_since_last_seen	331.9765	92.4769	0.0000	366.0000
fulladdress_days_since_last_seen	325.9429	99.3807	0.0000	366.0000
ssnfname_days_since_last_seen	328.3735	96.4830	0.0000	366.0000
ssnlname_days_since_last_seen	328.3835	96.4731	0.0000	366.0000
ssnfullname_days_since_last_seen	328.4948	96.3463	0.0000	366.0000
ssnaddress_days_since_last_seen	330.0166	94.2347	0.0000	366.0000
ssnzip5_days_since_last_seen	329.9109	94.3522	0.0000	366.0000
ssndob_days_since_last_seen	332.1524	92.2884	0.0000	366.0000
ssnhomephone_days_since_last_seen	331.7116	92.3395	0.0000	366.0000
ssnnamedob_days_since_last_seen	332.2976	92.1081	0.0000	366.0000
ssnfulladdress_days_since_last_seen	330.1058	94.1296	0.0000	366.0000
namehomephone_days_since_last_seen	331.5303	92.5358	0.0000	366.0000
namefulladdress_days_since_last_seen	329.7125	94.5598	0.0000	366.0000
fulladdressdob_days_since_last_seen	333.3147	90.4426	0.0000	366.0000
fulladdresshomephone_days_since_last_seen	330.6539	94.0002	0.0000	366.0000
fulladdressnamedob_days_since_last_seen	333.4419	90.2807	0.0000	366.0000
dobhomephone_days_since_last_seen	333.6762	90.0446	0.0000	366.0000
homephonenamedob_days_since_last_seen	334.1353	89.4696	0.0000	366.0000
ssn_nunique1_homephone	1.013521	0.37940523	1	34
ssn_nunique1_ssnfname	1.000101	0.01187392	1	4
ssn_nunique1_dob	1.000238	0.01667164	1	4
ssn_nunique1_fulladdress	1.013461	0.37932828	1	34
ssn_nunique3_homephone	1.016819	0.41928069	1	34
ssn_nunique3_ssnfname	1.000245	0.02188471	1	5
ssn_nunique3_dob	1.000559	0.02804796	1	5
ssn_nunique3_fulladdress	1.016685	0.41912383	1	34
homephone_nunique1_ssn	1.20009	0.62502191	1	31
homephone_nunique1_ssnfname	1.200096	0.62503118	1	31
homephone_nunique1_dob	1.200162	0.62510764	1	31
homephone_nunique1_fulladdress	1.185962	0.49358118	1	22
homephone_nunique3_ssn	1.445197	0.85808354	1	32
homephone_nunique3_ssnfname	1.445211	0.85809842	1	32
homephone_nunique3_dob	1.445388	0.85827517	1	32
homephone_nunique3_fulladdress	1.427176	0.73848711	1	25
ssnfname_nunique1_ssn	1	0	1	1
ssnfname_nunique1_homephone	1.013425	0.37922937	1	34
ssnfname_nunique1_dob	1.000142	0.01191554	1	2
ssnfname_nunique1_fulladdress	1.013365	0.37915237	1	34

ssnfname_nunique3_ssn	1	0	1	1
ssnfname_nunique3_homephone	1.016586	0.41872554	1	34
ssnfname_nunique3_dob	1.000328	0.01821792	1	3
ssnfname_nunique3_fulladdress	1.016446	0.41856147	1	34
dob_nunique1_ssn	1.234492	1.74697068	1	40
dob_nunique1_homephone	1.247976	1.78627128	1	40
dob_nunique1_ssnfname	1.234524	1.74740938	1	40
dob_nunique1_fulladdress	1.248106	1.78810811	1	40
dob_nunique3_ssn	1.546981	3.9388684	1	72
dob_nunique3_homephone	1.56361	3.95721309	1	72
dob_nunique3_ssnfname	1.5471	3.94039834	1	72
dob_nunique3_fulladdress	1.56422	3.96458895	1	72
fulladdress_nunique1_ssn	1.022484	0.48070022	1	30
fulladdress_nunique1_homephone	1.008729	0.29880243	1	30
fulladdress_nunique1_ssnfname	1.022489	0.48070519	1	30
fulladdress_nunique1_dob	1.022597	0.48048788	1	30
fulladdress_nunique3_ssn	1.02797	0.53933475	1	30
fulladdress_nunique3_homephone	1.010959	0.3356412	1	30
fulladdress_nunique3_ssnfname	1.027977	0.53934087	1	30
fulladdress_nunique3_dob	1.028236	0.53928195	1	30
dayofweek_risk	0.0144	0.0006	0.0135	0.0152

Appendix 3. Initial Recursive Feature Selection Result

Ranking	Variable
1	address_1_count_address_14_count_Ave
1	address_1_count_address_30_count_Ave
1	address_count14_date
1	address_count1_date
1	address_count30_date
1	address_count3_date
1	address_count7_date
1	address_days_since_last_seen
1	fulladdress_1_count_fulladdress_14_count_Ave
1	fulladdress_1_count_fulladdress_30_count_Ave
1	fulladdress_1_count_fulladdress_7_count_Ave
1	fulladdress_count0_date
1	fulladdress_count14_date
1	fulladdress_count180_date
1	fulladdress_count1_date
1	fulladdress_count30_date
1	fulladdress_count3_date
1	fulladdress_count7_date
1	fulladdress_days_since_last_seen
1	fulladdress_nunique1_ssn
1	fulladdress_nunique1_ssnfname
1	fulladdress_nunique3_dob
1	fulladdress_nunique3_ssn
1	fulladdress_nunique3_ssnfname

Ranking	Variable
1	fulladdresshomephone_count14_date
1	fulladdresshomephone_count30_date
1	fulladdresshomephone_count7_date
1	fulladdresshomephone_days_since_last_seen
1	homephone_count1_date
1	homephone_count3_date
1	homephone_count7_date
1	homephone_nunique3_dob
1	homephone_nunique3_ssn
1	homephone_nunique3_ssnfname
1	namedob_count180_date
1	namedob_count30_date
1	namedob_count3_date
1	namedob_count7_date
1	namedob_days_since_last_seen
1	ssn_count7_date
1	ssn_days_since_last_seen
1	ssndob_count30_date
1	ssndob_days_since_last_seen
1	ssnfname_count14_date
1	ssnfname_count180_date
1	ssnfname_count30_date
1	ssnfname_count3_date
1	ssnfname_count7_date
1	ssnfname_days_since_last_seen

Ranking	Variable
1	ssnfullname_count14_date
1	ssnfullname_count180_date
1	ssnfullname_count30_date
1	ssnfullname_days_since_last_seen
1	ssnlname_count14_date
1	ssnlname_count180_date
1	ssnlname_days_since_last_seen
1	ssnnamedob_count180_date
1	ssnnamedob_count30_date
1	ssnnamedob_days_since_last_seen
2	fulladdresshomephone_count180_date
3	ssnfullname_count7_date
4	ssndob_count7_date
5	ssn_count3_date
6	namedob_count14_date
7	ssnnamedob_count7_date
8	ssnlname_count7_date
9	ssn_count14_date
10	address_count0_date
11	ssnlname_count30_date
12	homephone_nunique1_ssn
13	ssn_count30_date
14	ssn_count180_date
15	address_1_count_address_7_count_Ave
16	address_count180_date

Ranking	Variable
17	fulladdress_nunique1_dob
18	ssnnamedob_count14_date
19	fulladdresshomephone_count3_date
20	homephone_count14_date
21	ssndob_count180_date
22	ssndob_count14_date