**A close up of a sign

Description automatically generated**

**DSO 562 FRAUD ANALYTICS**

Project Report

**Identify Fraud Detection**

**Group 4**

Siqi Bian

Siyuan Fang

Smita Bhat

Yiming Ding

Xuejie Guo

Lu Xia

Submitted to:

**Prof. Stephen Coggeshall**

**TABLE OF CONTENTS:**

1. Executive Summary…………………………………………………………………………………………3
2. Data Description……………………………………………………………………………………………..4
3. Data Cleaning………………………………………………………………………………………………….8
4. Variable Creation…………………………………………………………………………………………….9
   1. Velocity………………………………………………………………………………………………….10
   2. Day Since Variables……………………………………………………………………………..….10
   3. Relative Velocity………………………………………………………………………………………11
   4. Random Variable………………………………………………………………………………….….11
5. Feature Selection…………………………………………………………………………………….…….12
   1. Filters………………………………………………………………………………………………………12
   2. Wrappers……………………………………………………………………………………….……….13
6. Algorithms…………………………………………………………………………………………..….…….19
   1. Logistic Regression…………………………………………………………………….…………..19
   2. Random Forest ………………………………………………………………..……..……………..20
   3. Neural Network …………………………………………………………………………..…………21
   4. Support Vector Machine……………………………………………….…………………………22
   5. Boosted trees………………………………………………………………………….………………22
7. Model Summary……………………………………………………………………………………………24
8. Results……………………………………………………………………………………………………..…..26
9. Conclusion…………………………………………………………………………………………………….29

APPENDIX 1: Data Quality Report……………………………………………………………………….30

APPENDIX 2: Variables ……………………………………………………………………………………….40

1. **EXECUTIVE SUMMARY:**

Identity theft occurs when a person gains unauthorized access to another person’s personally identifying information such as name, Social Security Number (SSN), Date of Birth and uses it to commit fraud or other crimes. The motivation of this project is to find these fraudulent activities using supervised machine learning algorithms on product applications dataset.

In order to understand the data well, we examined each variable closely to get some actionable business insights. A thorough process of creating variables, finding the best features has been carried out before building 5 supervised learning models – logistic regression, random forest, boosted trees, support vector machine and neural networks to calculate the likelihood of each transaction being fraudulent. The models were trained over a subset of data and then validated over testing set and out-of-time set to check how the model works over the data they have never seen before. A detailed section of variable creation, feature selection and model building can be found in section 4,5 and 6. A data quality report has been supplemented in the appendix for reference.

Based on our analysis we noted that fraudsters exhibited one or both of the following characteristics:

1. A burst of application within short period of time.
2. Fraudster used his/her own personal information to fill address and phone number.

The summary of all the model’s performance is documented in section 7. The model’s performance is evaluated using Fraud Detection Rate at 3% which is the total number of actual frauds caught in top 3 % after sorting the predicted scores in descending order divided by total number of frauds in a given dataset (train, test, OOT). Based on FDR @ 3% metric Boosted Trees was our champion model. The table which summarizes the statistics of boosted trees separately for training, test and OOT dataset can be found in section 8.

1. **DESCRIPTION OF DATA**

**Description of Dataset**

The product application dataset records personal information of each applicant and a fraud label indicating whether it’s a fraudulent transaction or not. The dataset has 1000000 rows and 10 fields. The fields are record, date – the date when the transaction took place, social security number, first name, last name, address, zip code, date of birth, home phone and fraud labels. The date column begins from 2016-01-01 and ends at 2016-12-31. Looking at the data, we can clearly say that the project falls under real-time fraud detection where time is crucial.

Histograms and tables for some of the variables are described below:

1. Date

Description: Field which includes the year, month and date of each record. Starting date is 2016-01-01 and the ending date is 2016-12-31. The following table is an overview of the top 5 days that have the greatest number of applications.



1. SSN

Description: Social Security Number of the applicant.

A picture containing drawing

Description automatically generated

1. address

Description: Address of the applicant.

A screenshot of a cell phone

Description automatically generated

1. zip5

Description: Postal code of the applicant’s address.

A close up of a device

Description automatically generated

1. dob

Description: Date of birth of the applicant.

A circuit board

Description automatically generated

1. fraud\_labels

Description: This binary field indicates whether a certain record is a fraud or not. Fraud record is encoded in 1. The following table shows the fraud\_label’s distribution. We notice that 1.46% of the data is labeled as fraudulent activity.



To get a better sense of the distribution of fraud records over time, we grouped fraud\_label by date range and divided it by total application per date range. We plot daily fraud rates distribution and weekly fraud rates distribution. As shown in the following two figures, the non-fraud rate is relatively stable across the whole year. However, the fraud rate fluctuates dramatically. The daily fraud rate hits the peak at around the beginning of February, while the weekly fraud rate in the middle of July and is the highest for the entire year.

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Fig. Daily distribution of fraud rates Fig. Weekly distribution of fraud rates

1. **DATA CLEANING**

The dataset has no missing values but has many frivolous records. Frivolous data are usually artificially made by humans to make it easy to record. From the primary data analysis, we can observe that the frivolous values are always the most common value in the fields, which will influence the outcome. As a result, we need to replace them to a distinct value, so that each of them is as unique as possible and gives us reliable outputs.

**Frivolous field values:**



**Handling Frivolous Values:**

* Social Security Number: We replaced the ‘999999999’ value with the record number.
* Address: We replaced the ‘123 Main St’ address with the record number
* Date of Birth: We replaced the ‘19070626’ date of birth with the record number
* Home phone: We replaced the ‘9999999999’ value with the record number

1. **VARIABLE CREATION:**

Variables were created on overall behavioral patterns of the fraudsters. We created 547 expert variables using the original 8 columns. Initially, we created 26 new attributes by concatenating different variables. The following table lists all the 26 new attributes. These attributes were then used to produce in total 546 new variables. The method is described in the next section.

|  |  |  |
| --- | --- | --- |
| ssn | address | homephone |
| firstname\_lastname\_dob | address\_zip5 | ssn\_firstname |
| ssn\_lastname | ssn\_firstname\_lastname | ssn\_address |
| ssn\_zip5 | ssn\_dob | ssn\_homephone |
| ssn\_firstname\_lastname\_dob | ssn\_address\_zip5 | address\_firstname\_lastname\_dob |
| zip5\_firstname\_lastname\_dob | address\_zip5\_firstname\_lastname | homephone\_firstname\_lastname |
| homephone\_zip | homephone\_address | homephone\_firstname\_lastname\_dob |
| homephone\_address\_zip5 | firstnames\_lastname\_ssn\_homephone\_dob\_address\_zip5 | firstname\_lastname\_ssn\_homephone\_dob |
| firstname\_lastname\_ssn\_address\_zip5 | homephone\_sdob |  |

**Velocity:**

208 out of 546 variables were created using this method. This variable tells us how many times we have seen the attribute in the past 0, 1,3,7,14,30,90 and 180 days. It also shows us how frequently an individual had submitted the application.

A close up of a map

Description automatically generated

We have in total 26 attributes and 8 different timeframes to calculate the velocity. So, 26\*8 will yield us 208 new variables.

**Day Since Variable:**

26 out of 546 variables were created using this method. This variable gives us a number which corresponds to the number of days since we last saw an attribute.

A picture containing bird

Description automatically generated

**Relative Velocity:**

312 out of 546 variables were created using this method. Variables using this method are created by calculating the number of applications with an entity seen in 0 or 1 day divided by the number of applications with the same entity seen in the past 3,7,14,30,90,180 days.

# apps with that ***group*** seen in the recent past

# apps with that ***same group*** seen in the past {3,7,14,30} days

The variables were first created considering ‘0’ days in the numerator as recent past and 6 different timeframes in the denominator. This produces 26\*6 = 156 variables. Next ‘1’ day was considered as the recent past in the numerator and 6 different timeframes in the denominator. This produces 26\*6 = 156 variables. In total, 312(156+156) variables were created using this method.

**Random Variable:**

In addition to these 546 variables, we created another variable called “Random Variable” which assigns a random value between 0 and 1 to every single record. The motivation behind this variable is to see if the frauds can be strongly predicted simply by a random guess.

The table containing the minimum value, maximum value, the mean and standard deviation of all the 546 variables is supplemented in the appendix for reference.

1. **FEATURE SELECTION**

The 547 variables created using the variable creation process cannot be directly fed into the model because of high dimensions. The results produced by a higher dimension dataset will be non-intuitive, and it is highly likely that the data points will become outliers in one or more dimensions. The phenomenon is called the curse of dimensionality in data science glossary. We also run the risk of massively overfitting our model, which would generally result in terrible out of the model performance. We adopted two steps to pick our features for model building. The first step is filtering, and the next step is the wrapper.

**Filters:**

For each variable we make two separate distribution (goods and bads). The amount of separation between these distributions is the importance of the variable. We used Kolmogorov -Smirnov score and Fraud Detection Rate.

**Kolmogorov- Smirnov (KS) Score:**

The KS score measures the maximum separation between two distributions. The more different the curves, the better the variable for separating, and thus more important the variable is. The two distribution of interest here is Fraudulent Event and Non-Fraudulent Event. In this situation, KS acts as a robust measure to assess the variable importance. The formula to calculate KS score is described below.

A screenshot of a cell phone

Description automatically generated

**Fraud Detection Rate:**

To calculate Fraud Detection Rate @ 3% we first sort the predicted values in descending order of the magnitude. Then we calculate the total number of fraudulent records in the top 3% and divide it by the total number of fraudulent records in a given data set, i.e. training, testing and out-of-time (OOT). For example: If the top 3% consists of 150 fraudulent records and the training dataset consists of 300 fraudulent records the FDR @ 3% is 150/300 = 50%.

Now after calculating these two metrics for every single variable, we rank all the variables based on KS and FDR values and assign each variable two ranks respectively in an increasing order of magnitude. Each variable now has one KS Rank and one FDR Rank. We take an average of these two ranks as a final score metric and then again rank order all the variables by their final scores in a decreasing fashion.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| var\_name | ks | FDR | rank\_ks | rank\_FDR | avg\_rank |
| **fraud\_label** | 1 | 1 | 548 | 548 | 548 |
| **address\_count30\_date** | 0.333 | 0.353 | 547 | 546 | 546.5 |
| **address\_zip5\_count30\_date** | 0.332 | 0.355 | 546 | 547 | 546.5 |
| **address\_pastday** | 0.325 | 0.35 | 545 | 544 | 544.5 |
| **address\_zip5\_pastday** | 0.324 | 0.35 | 544 | 544 | 544 |
| **address\_count14\_date** | 0.322 | 0.346 | 543 | 542 | 542.5 |
| **address\_zip5\_count14\_date** | 0.322 | 0.342 | 542 | 538 | 540 |
| **address\_count90\_date** | 0.321 | 0.347 | 541 | 543 | 542 |
| **address\_zip5\_count90\_date** | 0.32 | 0.345 | 540 | 541 | 540.5 |
| **address\_count180\_date** | 0.319 | 0.344 | 539 | 540 | 539.5 |

After this process, we choose the top 279 variables in terms of their ranking, and the rest will be discarded.

**Wrapper:**

The next step in feature selection is wrapper. We used Recursive Feature Elimination with Cross Validation (RFECV) to select our final 30 and 20 variables. RFECV works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes and combinations of attributes contribute the most in predicting the target attribute. The process eliminates dependencies and collinearities between variables by keeping only one of the two variables that might be highly correlated. A description of the process is as follows:

1. The estimator is trained using the initial set of features, and the importance of each feature is obtained either through a coefficient attribute or a feature importance attribute such as p-value.
2. The least important features are pruned based on their importance.
3. The process is repeated recursively on the pruned set until the desired number of features has been reached.

RFE with cross-validation has a function of automatically tuning the number of features to be selected with cross validation to help determine what the ideal number of features would be. Here, we applied RFE with 3-fold cross validation using a logistic regression function as typically for this process.

We applied two iterations of cross validation for the following reasons:

1. On the first iteration, we find 55 variables as the optimal choice. However, 55 still results in high dimensionality and thus a second iteration is necessary. A graph indicating the CV scores based on the number of features is shown below:

A close up of a logo

Description automatically generated

1. Now on the second iteration, we identified the top 34 variables and decided to keep the top 30 variables as the strongest indicators to identify fraud. A graph indicating the CV scores based on the number of features is shown below:

A close up of a person

Description automatically generated

**Top 30 Selected Variables are:**

1. **address\_zip5\_0\_count30\_count\_ratio**: The ratio of number of applications that have the same address and Zip code in the recent past divided by the number of applications in this group in the past 30 days.
2. **address\_count3\_date**: The number of applications that have the same address over the past 3 days.
3. **address\_zip5\_0\_count180\_count\_ratio:** The ratio of number of applications that have the same address and Zip code in the recent past divided by the number of applications in this group in the past 180 days.
4. **address\_0\_count7\_count\_ratio:** The ratio of number of applications that have the same address in the recent past divided by the number of applications in this group in the past 7 days.
5. **homephone\_address\_count30\_date**: The number of applications that have the same address and home phone number over the past 30 days.
6. **address\_zip5\_0\_count3\_count\_ratio:**The ratio of number of applications that have the same address and Zip code in the recent past divided by the number of applications in this group in the past 3 days.
7. **homephone\_address\_pastday**: The number of days since that home phone and address appears.
8. **homephone\_address\_zip5\_count90\_date:** The number of applications that have the same address, home phone number and Zip code over the past 90 days.
9. **ssn\_dob\_count90\_date:** The number of applications that have the same SSN, and date of birth over the past 90 days.
10. **address\_1\_count90\_count\_ratio:** The ratio of number of applications that have the same address in the recent past divided by the number of applications in this group in the past 90 days.
11. **ssn\_dob\_0\_count30\_count\_ratio:** The ratio of number of applications that have the same SSN and date of birth in the recent past divided by the number of applications in this group in the past 30 days.
12. **firstname\_lastname\_dob\_0\_count30\_count\_ratio**: The ratio of number of applications that have the full-name and identical birthday in the recent past divided by the number of applications in this group in the past 30 days.
13. **ssn\_0\_count30\_count\_ratio:** The ratio of number of applications that have the same SSN in the recent past divided by the number of applications in this group in the past 30 days.
14. **ssn\_firstname\_lastname\_dob\_0\_count30\_count\_ratio:**The ratio of number of applications that have the same SSN, full name and date of birth in the recent past divided by the number of applications in this group in the past 30 days
15. **ssn\_firstname\_0\_count30\_count\_ratio:** The ratio of number of applications that have the same SSN and first name in the recent past divided by the number of applications in this group in the past 30 days.
16. **address\_zip5\_1\_count90\_count\_ratio:** The ratio of number of applications that have the same address and Zip code in the recent past divided by the number of applications in this group in the past 90 days.
17. **ssn\_firstname\_count7\_date:** The number of applications that have the same SSN, and first name over the past 7 days.
18. **address\_count0\_date:** The number of applications that have the same address over the past day.
19. **homephone\_address\_0\_count180\_count\_ratio:** The ratio of number of applications that have the same home phone number and address in the recent past divided by the number of applications in this group in the past 180 days.
20. **address\_zip5\_1\_count7\_count\_ratio:** The ratio of number of applications that have the same address and Zip code in the recent past divided by the number of applications in this group in the past 7 days.
21. **firstname\_lastname\_dob\_0\_count7\_count\_ratio:** The ratio of number of applications that have the same full name and date of birth in the recent past divided by the number of applications in this group in the past 7 days.
22. **ssn\_dob\_1\_count30\_count\_ratio:** The ratio of number of applications that have the same SSN and date of birth in the recent past divided by the number of applications in this group in the past 30 days.
23. **homephone\_zip\_count1\_date:** The number of applications that have the same home phone number and Zip code over the past day.
24. **ssn\_firstname\_1\_count180\_count\_ratio:** The ratio of number of applications that have the same SSN and first name in the recent past divided by the number of applications in this group in the past 180 days.
25. **homephone\_address\_zip5\_1\_count7\_count\_ratio:** The ratio of number of applications that have the same home phone number, address and Zip code in the recent past divided by the number of applications in this group in the past 7 days.
26. **homephone\_count180\_date:** The number of applications that have the same home phone number over the past 180 days.
27. **address\_firstname\_lastname\_dob\_count30\_date:** The number of applications that have the same address, full name and identical date of birth over the past 30 days.
28. **zip5\_firstname\_lastname\_dob\_count30\_date:** The number of applications that have the same Zip code, full name and identical date of birth over the past 30 days.
29. **homephone\_firstname\_lastname\_count30\_date:** The number of applications that have the same home phone number and full name over the past 30 days.
30. **zip5\_firstname\_lastname\_dob\_0\_count30\_count\_ratio:** The ratio of number of applications that have the same Zip code, full name and identical date of birth in the recent past divided by the number of applications in this group in the past 30 days.
31. **ALGORITHMS:**

After selecting the top 30 features as mentioned from the previous section, the following section summarizes all the supervised learning methods that we used to build our models. The five supervised learning methods are Logistic Regression, Random Forest, Neural Network, Support Vector Machine (SVM), and Boosted Trees.

This process starts by removing the first two weeks of data to train the model effectively because the information that we get from the first two weeks is not significant. The remaining dataset is separated into three parts: training data, test data and out-of-time (OOT) data. The OOT data included all the records for the last two months in 2016. All other records after the first two weeks of 2016 and before November 2016 are randomly separated into train/test data, while 75% of the data is used to train the model.

For each method, different models have been trained with different parameter values, such as the number of estimators, number of trees, learning rate, and so on. From the top 30 features, we selected the top 20 features and built multiple models. All models are evaluated using FDR @ 3%. The explanation of each model is penned below.

**Logistic Regression:**

Logistic regression is a classification technique which is considered by most of the Machine Learning Engineers and Data Scientists as a baseline model. In our Identity Fraud Detection project, we have used logistic regression as the baseline model. The assumption that the logistic regression will make is that the classes are almost or perfectly linearly separable. The task is to find hyperplane which is best in separating the classes i.e., positive class or negative class.

Machine learning models tend to fit the training data set almost perfectly and yield poor performance for test data and out of time data. This is caused by overfitting, and to circumvent this problem, we used L1 regularization. L1 regularization is also called a lasso, which means Least Absolute Shrinkage and Selection Operator. Lasso adds an absolute value of the magnitude of coefficient as penalty term to the loss function. Lasso shrinks the less important feature’s coefficient to zero; thus, removing some features altogether. So, this also works well for feature selection.

We built several models with the top 30 and top 20 variables selected using feature selection. The table with the parameters and results are described in the next section.

**Random Forest:**

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from a randomly selected subset of the training set. It aggregates the votes from different decision trees to decide the final class of the test object. Ensemble algorithms are those which combine more than one algorithm of a same or different kind for classifying objects. There are several advantages of the Random Forest.

1. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won’t overfit the model.
2. The classifier of Random Forest can handle missing values
3. Random Forest classifier can be modeled for categorical values.

Parameters of Random Forest:

* n\_estimators: Number of trees in the forest.
* max\_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min\_samples\_split samples.
* min\_samples\_leaf: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min\_samples\_leaf training samples in each of the left and right branches.
* max\_features: The number of features to consider when looking for the best split

We used grid search to find the best n\_estimators, max\_features, min\_samples\_leaf, max\_depth and used those parameters to build the models. We tried n\_estimators = 400, 550,700, max\_features= 6,7,14, min\_samples\_leaf=3,5,8, max\_depth= 5,7,10. The table describing the models and the parameters with their respective FDR@3% is reported in the next section.

**Neural Network:**

Neural nets take inspiration from the learning process occurring in human brains. They consist of an artificial network of functions, called parameters, which allows the system to learn, and to fine-tune itself, by analyzing new data. Each parameter, sometimes also referred to as neurons, is a function that produces an output, after receiving one or multiple inputs. Those outputs are then passed to the next layer of neurons, which use them as inputs of their function and produce further outputs. Those outputs are then passed on to the next layer of neurons, and so it continues until every layer of neurons has been considered, and the terminal neurons have received their input. Those terminal neurons then output the final result for the model.

The typical architecture of a neural network consists of the following key elements:

* Input Layer: The input layer is the net’s layer that receives the data. In other words, the input layer contains all x variables.
* Hidden Layers: These are the layers between the input and output layers. There may or may not be more than one layers based on the input given by the solver. At each of the layer, nodes receive weighted signals from the previous layer and transform them to an output through the assigned activation function.
* Output Layer: It is last layer that outputs the dependent variables based on the transformations at each layer.

In our project, we considered two hidden layers and used ReLU activation. The output layer had sigmoid activation. The output layer gave us the probability of being a fraudulent event. To compile the model, a loss function must be defined, which will be used to evaluate the model’s performance. We used binary cross-entropy as a loss function and Adam optimizer to optimize the weights and the bias of each node. The FDR @ 3% for each model is reported in the next section.

**Boosted Trees:**

Boosted Tree is one of the non-linear models used to build predictors in supervised learning. It converts weak learners to become strong and powerful learners. In boosting, every new tree is used to fit on the data set. By computing a sequence of decision trees, each sequential decision tree is built upon the prediction residual of the previous one. After training and evaluating the first tree where each observation is assigned equal weights, the algorithm increases the weights of those variables that are difficult to classify and reduces the weights for those that are easier to classify.

Different models have been tried using different number of features or giving different parameter values. For example, the models are fine-tuned by setting number of trees to 500 or 1000; setting learning rate to 0.01 or 0.1; setting max depth to 1 or 2; setting number of estimators to 500 or 1000. Corresponding Fraud Detection Rates (FDR @ 3%) for train/test/OOT data have been reported in the next section.

**Support Vector Machine (SVM):**

An SVM works for linear and non-linear data. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space where N is the number of features that distinctly classify the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Several models have been trained using the different number of variables, kernels, regularization parameters (gamma), and specification of the loss function. We have experimented SVM with linear, polynomial, sigmoid and radial basis function (rbf) function. Better results are obtained from linearSVM and radial basis function. All detailed model information and corresponding Fraud Detection Rates (FDR @ 3%) for train/test/OOT data have been reported in the next section.

1. **MODEL SUMMARY**

After building all models using the above five supervised learning models, this section focuses on reporting key metrics to evaluate the performance of each method. For each of the five methods, there are several models with different numbers of variables, different parameter values, penalty methods, and so on. Each model outputs predicted fraud score, which is a continuous number that ranges from 0 to 1, and model is evaluated by calculating the FDR at 3%. This rate is derived by sorting predicted fraud scores in descending order. Pushing all the predicted fraud records to the top, the FDR is calculated by dividing the sum of the actual fraud score in the 3% population bin by the sum of total actual fraud labels. This quantitative measurement is calculated for train, test, and OOT dataset correspondingly. Here is a summarized table for all models. The best models from each method have been highlighted in color.



From the table above, the best model from each supervised learning method has been selected. By referring to the FDR @ 3% from the table above. The 5 best models from each method do not have much difference. One may notice Random Forest and Boosted Trees predicted better than others, and they derive about 55%, 54% and 51% Fraud Detection Rate @ 3% for training, testing and out-of-time data respectively. The final model is selected as Boosted Trees, which has the best performance in the out-of-time dataset.

1. **RESULTS**

Based on the model algorithms from the previous section, the following three tables summarize statistical details for training, testing, and out-of-time data separately. In each table and each dataset, all records are arranged by predicted fraud score given by our final model in descending order. By dividing each dataset into equally weighted bins, i.e., 1% of all populations, each table calculates the following statistics:

* Number of Record: total number of records in a certain bin
* Number of Good: total number of actual non-fraud records
* Number of Bad: total number of actual fraud records
* % Good: the portion of non-fraud records in a certain bin
* % Bad: the portion of fraud records in a certain bin
* Total Number of Records: cumulated total number of records in certain bins
* Cumulative Good: cumulative total number of actual non-fraud records
* Cumulative Bad: cumulative total number of actual fraud records
* % Good: the portion of non-fraud records in cumulative bins
* % Bad (FDR): the portion of fraud records in cumulative bins
* KS: the difference between % Bad and % Good
* FPR: the difference between Cumulative Good and Cumulative Bad

**Training Dataset:**



**Testing Dataset:**

****

**Out-Of-Time Dataset:**

****

1. **CONCLUSION**

The entire process performed rigorous fraud detection using a vast database that contains personal identity information. By carefully looking into the data, the report starts with providing qualitative and illustrative visualizations of data and dealing with frivolous field values. After ensuring the data is cleaned enough, our team came up with 547 important candidate variables by calculating velocity, relative velocity, day since, and random variable from the combination of variables and attributes. By choosing different timeframes, new candidate variables are created.

The 547 variables are filtered and reduced to 30 strongest variables using filter and wrapper method as mentioned in detail from the previous section. To build supervised learning models on those 30 powerful candidate variables, our team tried Logistic Regression, Random Forest, Support Vector Machine, Boosted Trees, and Neural Network. Models are trained using training dataset and fitted into test and out-of-time dataset correspondingly. For each method, several models have been selected in order to find the best-fine-tuned models. The finalized model is Boosted Tree, where a model is evaluated based on the performance of the Fraud Detection Rate at 3%. Choosing Boosted Trees as our final model, the final FDR at 3% is 55.98%, 54.28%, and 52% for training, testing, out-of-time datasets correspondingly. Detailed statistical summary for train, test, and out-of-time dataset has been included in this report.

In the future, our team wants to take some time to create more meaningful and innovative variables. In addition to those 547 variables, we want to perform feature selections using different methods, for example, Kulback-Leibler, Information Value, Mutual Information, etc. Either voting for selected variables or combing the selected variables into a big pool, our team wants to test if there is any improvement in final Fraud Detection Rate at 3%. Having all the skills and methods mentioned above, our team also intends to implement this process into different fields by identifying different fraudulent events/transactions/identities.

**APPENDIX 1: DATA QUALITY REPORT FOR APPLICATIONS DATA**

**DATA DESCRIPTION:**

Dataset Name: Applications Data

Dataset Purpose: Superficial dataset created to serve academic purpose of tackling identity frauds.

Number of Variables: 10

Number of records: 1000000

**SUMMARY TABLE:**

Categorical Variables:



**DATA FIELD EXPLORATION:**

Field 1:

Name: Record

Description: Unique Identifier for each entry in the data.

Field 2:

Name: Date

Description: Field which includes the year, month and date of each record. Starting date is 2016-01-01 and the ending date is 2016-12-31.



Field 3:

Name: SSN

Description: Social Security Number of the applicant.

A picture containing drawing

Description automatically generated

Field 4:

Name: firstname

Description: First Name of the applicant.

A picture containing comb

Description automatically generated

Field 5:

Name: lastname

Description: Last Name of the applicant.

A screenshot of a cell phone

Description automatically generated

Field 6:

Name: address

Description: Address of the applicant.

A screenshot of a cell phone

Description automatically generated

Field 7:

Name: zip5

Description: Postal code of the applicant’s address.

A close up of a device

Description automatically generated

Field 8:

Name: dob

Description: Date of birth of the applicant.

A circuit board

Description automatically generated

Field 9:

Name: homephone

Description: Home phone number of the applicant.

A circuit board

Description automatically generated

Field 10:

Name: fraud\_label

Description: 1 = Fraud and 0 = Not fraud.



Time Series Plots:

Since the variables in this dataset can be treated as time dependent, I plotted the number of applications on daily basis and weekly basis.

Daily Applications:

Number of applications on 2016-02-29 was ‘0’ and hence there was a steep drop in the graph. To avoid this, I assigned 2016-02-29 value to be equal to 2016-02-28 value.

The plot is show below.

A picture containing brush

Description automatically generated

Weekly Applications:

There was a steep drop in 9th week and the last week of 2016. This was because there were no record values on 2019-02-29 and the last week considered 2019-12-30 and 2019-12-31. These two days had 5498 applications in total but the average applications on weekly basis was 19175.32.

Hence the values of previous week i.e for the 9th week value corresponding to 8th week and for the 60th week value corresponding to 59th week was assigned.

The plot is show below:

A close up of a piece of paper

Description automatically generated

**APPENDIX 2: VARIABLE CREATION**

1. Risk Table & Graph

* The likelihood of fraud for CERTAIN day of a week
* After implementing the smoothing formula, we have found that the likelihood of fraud does not vary too much given certain day of a week. Wednesday, Thursday and Saturday are most likely to have fraud labels
* Note: the following graph and table used OOT dataset, i.e. ignored last two months data from original dataset

A screenshot of a cell phone

Description automatically generated

|  |  |
| --- | --- |
| weekday | Risk\_Percentage |
| Monday | 0.0135 |
| Tuesday | 0.0141 |
| Wednesday | 0.0152 |
| Thursday | 0.0150 |
| Friday | 0.0145 |
| Saturday | 0.0150 |
| Sunday | 0.0137 |

1. Statistics Summary for All Final Fields

* Note: Python code resource is from Blackboard – modified the control of time span
* Including `record` and `fraud\_label`, it has the total of 548 variables
* Velocity Numerical Meanings:
  + Look over past 0, 1,3,7,14,30,90 and 180 days
* Relative Velocity Numerical Meanings
  + Recent Past is assessed by: # of apps that group seen in 0 or 1 day
  + Divided the number above by # of apps that same group in the past 3,7,14,30,90 and 180 days.
  + Day Since Variable – naming convention is: Attributes/Combination Group Past Days (note: fast python code resources from classmate)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field Name | min | max | mean | stdev |
| record | 1 | 1000000 | 500000.500 | 288675.279 |
| ssn\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_count3\_date | 1 | 34 | 1.020 | 0.423 |
| ssn\_count7\_date | 1 | 34 | 1.026 | 0.454 |
| ssn\_count14\_date | 1 | 34 | 1.035 | 0.478 |
| ssn\_count30\_date | 1 | 34 | 1.051 | 0.513 |
| ssn\_count90\_date | 1 | 35 | 1.104 | 0.660 |
| ssn\_count180\_date | 1 | 63 | 1.165 | 0.886 |
| address\_count0\_date | 1 | 24 | 1.012 | 0.283 |
| address\_count1\_date | 1 | 30 | 1.025 | 0.483 |
| address\_count3\_date | 1 | 30 | 1.033 | 0.544 |
| address\_count7\_date | 1 | 30 | 1.043 | 0.586 |
| address\_count14\_date | 1 | 30 | 1.055 | 0.618 |
| address\_count30\_date | 1 | 30 | 1.078 | 0.664 |
| address\_count90\_date | 1 | 30 | 1.155 | 0.850 |
| address\_count180\_date | 1 | 53 | 1.241 | 1.148 |
| homephone\_count0\_date | 1 | 22 | 1.071 | 0.365 |
| homephone\_count1\_date | 1 | 31 | 1.202 | 0.627 |
| homephone\_count3\_date | 1 | 32 | 1.449 | 0.862 |
| homephone\_count7\_date | 1 | 33 | 1.933 | 1.213 |
| homephone\_count14\_date | 1 | 36 | 2.764 | 1.744 |
| homephone\_count30\_date | 1 | 50 | 4.598 | 2.864 |
| homephone\_count90\_date | 1 | 128 | 10.719 | 6.941 |
| homephone\_count180\_date | 1 | 240 | 17.651 | 12.739 |
| firstname\_lastname\_dob\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| firstname\_lastname\_dob\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| firstname\_lastname\_dob\_count3\_date | 1 | 34 | 1.020 | 0.422 |
| firstname\_lastname\_dob\_count7\_date | 1 | 34 | 1.025 | 0.451 |
| firstname\_lastname\_dob\_count14\_date | 1 | 34 | 1.032 | 0.472 |
| firstname\_lastname\_dob\_count30\_date | 1 | 34 | 1.046 | 0.497 |
| firstname\_lastname\_dob\_count90\_date | 1 | 34 | 1.092 | 0.573 |
| firstname\_lastname\_dob\_count180\_date | 1 | 39 | 1.144 | 0.677 |
| address\_zip5\_count0\_date | 1 | 24 | 1.012 | 0.283 |
| address\_zip5\_count1\_date | 1 | 30 | 1.024 | 0.482 |
| address\_zip5\_count3\_date | 1 | 30 | 1.031 | 0.542 |
| address\_zip5\_count7\_date | 1 | 30 | 1.039 | 0.581 |
| address\_zip5\_count14\_date | 1 | 30 | 1.048 | 0.607 |
| address\_zip5\_count30\_date | 1 | 30 | 1.065 | 0.634 |
| address\_zip5\_count90\_date | 1 | 30 | 1.118 | 0.704 |
| address\_zip5\_count180\_date | 1 | 39 | 1.179 | 0.802 |
| ssn\_firstname\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_firstname\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_firstname\_count3\_date | 1 | 34 | 1.020 | 0.423 |
| ssn\_firstname\_count7\_date | 1 | 34 | 1.026 | 0.452 |
| ssn\_firstname\_count14\_date | 1 | 34 | 1.034 | 0.473 |
| ssn\_firstname\_count30\_date | 1 | 34 | 1.049 | 0.499 |
| ssn\_firstname\_count90\_date | 1 | 34 | 1.099 | 0.577 |
| ssn\_firstname\_count180\_date | 1 | 39 | 1.156 | 0.680 |
| ssn\_lastname\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_lastname\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_lastname\_count3\_date | 1 | 34 | 1.020 | 0.423 |
| ssn\_lastname\_count7\_date | 1 | 34 | 1.026 | 0.452 |
| ssn\_lastname\_count14\_date | 1 | 34 | 1.034 | 0.473 |
| ssn\_lastname\_count30\_date | 1 | 34 | 1.049 | 0.499 |
| ssn\_lastname\_count90\_date | 1 | 34 | 1.099 | 0.577 |
| ssn\_lastname\_count180\_date | 1 | 39 | 1.156 | 0.680 |
| ssn\_firstname\_lastname\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_firstname\_lastname\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_firstname\_lastname\_count3\_date | 1 | 34 | 1.020 | 0.423 |
| ssn\_firstname\_lastname\_count7\_date | 1 | 34 | 1.026 | 0.452 |
| ssn\_firstname\_lastname\_count14\_date | 1 | 34 | 1.034 | 0.473 |
| ssn\_firstname\_lastname\_count30\_date | 1 | 34 | 1.049 | 0.499 |
| ssn\_firstname\_lastname\_count90\_date | 1 | 34 | 1.099 | 0.576 |
| ssn\_firstname\_lastname\_count180\_date | 1 | 39 | 1.155 | 0.680 |
| ssn\_address\_count0\_date | 1 | 3 | 1.001 | 0.023 |
| ssn\_address\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| ssn\_address\_count3\_date | 1 | 5 | 1.003 | 0.059 |
| ssn\_address\_count7\_date | 1 | 6 | 1.007 | 0.087 |
| ssn\_address\_count14\_date | 1 | 8 | 1.014 | 0.122 |
| ssn\_address\_count30\_date | 1 | 11 | 1.028 | 0.182 |
| ssn\_address\_count90\_date | 1 | 26 | 1.078 | 0.336 |
| ssn\_address\_count180\_date | 1 | 39 | 1.133 | 0.490 |
| ssn\_zip5\_count0\_date | 1 | 3 | 1.001 | 0.023 |
| ssn\_zip5\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| ssn\_zip5\_count3\_date | 1 | 5 | 1.003 | 0.060 |
| ssn\_zip5\_count7\_date | 1 | 6 | 1.007 | 0.088 |
| ssn\_zip5\_count14\_date | 1 | 8 | 1.014 | 0.122 |
| ssn\_zip5\_count30\_date | 1 | 11 | 1.028 | 0.182 |
| ssn\_zip5\_count90\_date | 1 | 26 | 1.078 | 0.337 |
| ssn\_zip5\_count180\_date | 1 | 39 | 1.133 | 0.490 |
| ssn\_dob\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_dob\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_dob\_count3\_date | 1 | 34 | 1.020 | 0.422 |
| ssn\_dob\_count7\_date | 1 | 34 | 1.025 | 0.451 |
| ssn\_dob\_count14\_date | 1 | 34 | 1.032 | 0.472 |
| ssn\_dob\_count30\_date | 1 | 34 | 1.046 | 0.496 |
| ssn\_dob\_count90\_date | 1 | 34 | 1.092 | 0.570 |
| ssn\_dob\_count180\_date | 1 | 39 | 1.143 | 0.670 |
| ssn\_homephone\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| ssn\_homephone\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| ssn\_homephone\_count3\_date | 1 | 5 | 1.003 | 0.058 |
| ssn\_homephone\_count7\_date | 1 | 6 | 1.007 | 0.086 |
| ssn\_homephone\_count14\_date | 1 | 8 | 1.013 | 0.120 |
| ssn\_homephone\_count30\_date | 1 | 11 | 1.027 | 0.178 |
| ssn\_homephone\_count90\_date | 1 | 26 | 1.074 | 0.332 |
| ssn\_homephone\_count180\_date | 1 | 39 | 1.127 | 0.486 |
| ssn\_firstname\_lastname\_dob\_count0\_date | 1 | 21 | 1.007 | 0.223 |
| ssn\_firstname\_lastname\_dob\_count1\_date | 1 | 34 | 1.015 | 0.381 |
| ssn\_firstname\_lastname\_dob\_count3\_date | 1 | 34 | 1.020 | 0.422 |
| ssn\_firstname\_lastname\_dob\_count7\_date | 1 | 34 | 1.025 | 0.451 |
| ssn\_firstname\_lastname\_dob\_count14\_date | 1 | 34 | 1.032 | 0.471 |
| ssn\_firstname\_lastname\_dob\_count30\_date | 1 | 34 | 1.046 | 0.496 |
| ssn\_firstname\_lastname\_dob\_count90\_date | 1 | 34 | 1.091 | 0.569 |
| ssn\_firstname\_lastname\_dob\_count180\_date | 1 | 39 | 1.142 | 0.669 |
| ssn\_address\_zip5\_count0\_date | 1 | 3 | 1.001 | 0.023 |
| ssn\_address\_zip5\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| ssn\_address\_zip5\_count3\_date | 1 | 5 | 1.003 | 0.059 |
| ssn\_address\_zip5\_count7\_date | 1 | 6 | 1.007 | 0.087 |
| ssn\_address\_zip5\_count14\_date | 1 | 8 | 1.014 | 0.122 |
| ssn\_address\_zip5\_count30\_date | 1 | 11 | 1.028 | 0.181 |
| ssn\_address\_zip5\_count90\_date | 1 | 26 | 1.077 | 0.336 |
| ssn\_address\_zip5\_count180\_date | 1 | 39 | 1.132 | 0.489 |
| address\_firstname\_lastname\_dob\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| address\_firstname\_lastname\_dob\_count1\_date | 1 | 3 | 1.001 | 0.037 |
| address\_firstname\_lastname\_dob\_count3\_date | 1 | 5 | 1.003 | 0.057 |
| address\_firstname\_lastname\_dob\_count7\_date | 1 | 6 | 1.007 | 0.084 |
| address\_firstname\_lastname\_dob\_count14\_date | 1 | 8 | 1.013 | 0.117 |
| address\_firstname\_lastname\_dob\_count30\_date | 1 | 11 | 1.026 | 0.175 |
| address\_firstname\_lastname\_dob\_count90\_date | 1 | 26 | 1.071 | 0.327 |
| address\_firstname\_lastname\_dob\_count180\_date | 1 | 39 | 1.121 | 0.479 |
| zip5\_firstname\_lastname\_dob\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| zip5\_firstname\_lastname\_dob\_count1\_date | 1 | 3 | 1.001 | 0.037 |
| zip5\_firstname\_lastname\_dob\_count3\_date | 1 | 5 | 1.003 | 0.057 |
| zip5\_firstname\_lastname\_dob\_count7\_date | 1 | 6 | 1.007 | 0.084 |
| zip5\_firstname\_lastname\_dob\_count14\_date | 1 | 8 | 1.013 | 0.118 |
| zip5\_firstname\_lastname\_dob\_count30\_date | 1 | 11 | 1.026 | 0.176 |
| zip5\_firstname\_lastname\_dob\_count90\_date | 1 | 26 | 1.071 | 0.330 |
| zip5\_firstname\_lastname\_dob\_count180\_date | 1 | 39 | 1.122 | 0.484 |
| address\_zip5\_firstname\_lastname\_count0\_date | 1 | 3 | 1.001 | 0.023 |
| address\_zip5\_firstname\_lastname\_count1\_date | 1 | 3 | 1.001 | 0.039 |
| address\_zip5\_firstname\_lastname\_count3\_date | 1 | 5 | 1.003 | 0.060 |
| address\_zip5\_firstname\_lastname\_count7\_date | 1 | 6 | 1.007 | 0.088 |
| address\_zip5\_firstname\_lastname\_count14\_date | 1 | 8 | 1.014 | 0.123 |
| address\_zip5\_firstname\_lastname\_count30\_date | 1 | 11 | 1.029 | 0.183 |
| address\_zip5\_firstname\_lastname\_count90\_date | 1 | 26 | 1.078 | 0.339 |
| address\_zip5\_firstname\_lastname\_count180\_date | 1 | 39 | 1.134 | 0.495 |
| homephone\_firstname\_lastname\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| homephone\_firstname\_lastname\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| homephone\_firstname\_lastname\_count3\_date | 1 | 5 | 1.003 | 0.058 |
| homephone\_firstname\_lastname\_count7\_date | 1 | 6 | 1.007 | 0.086 |
| homephone\_firstname\_lastname\_count14\_date | 1 | 8 | 1.013 | 0.120 |
| homephone\_firstname\_lastname\_count30\_date | 1 | 11 | 1.027 | 0.178 |
| homephone\_firstname\_lastname\_count90\_date | 1 | 26 | 1.075 | 0.333 |
| homephone\_firstname\_lastname\_count180\_date | 1 | 39 | 1.128 | 0.488 |
| homephone\_zip\_count0\_date | 1 | 21 | 1.007 | 0.224 |
| homephone\_zip\_count1\_date | 1 | 30 | 1.015 | 0.378 |
| homephone\_zip\_count3\_date | 1 | 30 | 1.021 | 0.426 |
| homephone\_zip\_count7\_date | 1 | 30 | 1.026 | 0.459 |
| homephone\_zip\_count14\_date | 1 | 30 | 1.034 | 0.482 |
| homephone\_zip\_count30\_date | 1 | 30 | 1.049 | 0.508 |
| homephone\_zip\_count90\_date | 1 | 30 | 1.097 | 0.585 |
| homephone\_zip\_count180\_date | 1 | 39 | 1.152 | 0.687 |
| homephone\_dob\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| homephone\_dob\_count1\_date | 1 | 3 | 1.001 | 0.037 |
| homephone\_dob\_count3\_date | 1 | 5 | 1.003 | 0.057 |
| homephone\_dob\_count7\_date | 1 | 6 | 1.007 | 0.083 |
| homephone\_dob\_count14\_date | 1 | 8 | 1.013 | 0.117 |
| homephone\_dob\_count30\_date | 1 | 11 | 1.026 | 0.174 |
| homephone\_dob\_count90\_date | 1 | 26 | 1.070 | 0.327 |
| homephone\_dob\_count180\_date | 1 | 39 | 1.121 | 0.480 |
| homephone\_address\_count0\_date | 1 | 21 | 1.007 | 0.224 |
| homephone\_address\_count1\_date | 1 | 30 | 1.015 | 0.378 |
| homephone\_address\_count3\_date | 1 | 30 | 1.020 | 0.426 |
| homephone\_address\_count7\_date | 1 | 30 | 1.026 | 0.459 |
| homephone\_address\_count14\_date | 1 | 30 | 1.034 | 0.482 |
| homephone\_address\_count30\_date | 1 | 30 | 1.048 | 0.508 |
| homephone\_address\_count90\_date | 1 | 30 | 1.096 | 0.583 |
| homephone\_address\_count180\_date | 1 | 39 | 1.149 | 0.685 |
| homephone\_firstname\_lastname\_dob\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| homephone\_firstname\_lastname\_dob\_count1\_date | 1 | 3 | 1.001 | 0.036 |
| homephone\_firstname\_lastname\_dob\_count3\_date | 1 | 5 | 1.003 | 0.056 |
| homephone\_firstname\_lastname\_dob\_count7\_date | 1 | 6 | 1.007 | 0.083 |
| homephone\_firstname\_lastname\_dob\_count14\_date | 1 | 8 | 1.012 | 0.116 |
| homephone\_firstname\_lastname\_dob\_count30\_date | 1 | 11 | 1.025 | 0.173 |
| homephone\_firstname\_lastname\_dob\_count90\_date | 1 | 26 | 1.069 | 0.325 |
| homephone\_firstname\_lastname\_dob\_count180\_date | 1 | 39 | 1.119 | 0.477 |
| homephone\_address\_zip5\_count0\_date | 1 | 21 | 1.007 | 0.224 |
| homephone\_address\_zip5\_count1\_date | 1 | 30 | 1.015 | 0.378 |
| homephone\_address\_zip5\_count3\_date | 1 | 30 | 1.020 | 0.426 |
| homephone\_address\_zip5\_count7\_date | 1 | 30 | 1.026 | 0.459 |
| homephone\_address\_zip5\_count14\_date | 1 | 30 | 1.034 | 0.482 |
| homephone\_address\_zip5\_count30\_date | 1 | 30 | 1.048 | 0.508 |
| homephone\_address\_zip5\_count90\_date | 1 | 30 | 1.096 | 0.583 |
| homephone\_address\_zip5\_count180\_date | 1 | 39 | 1.149 | 0.684 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count1\_date | 1 | 3 | 1.001 | 0.037 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count3\_date | 1 | 5 | 1.003 | 0.058 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count7\_date | 1 | 6 | 1.007 | 0.085 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count14\_date | 1 | 8 | 1.013 | 0.119 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count30\_date | 1 | 11 | 1.027 | 0.176 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count90\_date | 1 | 26 | 1.073 | 0.329 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_count180\_date | 1 | 39 | 1.125 | 0.480 |
| firstname\_lastname\_ssn\_homephone\_dob\_count0\_date | 1 | 3 | 1.000 | 0.022 |
| firstname\_lastname\_ssn\_homephone\_dob\_count1\_date | 1 | 3 | 1.001 | 0.036 |
| firstname\_lastname\_ssn\_homephone\_dob\_count3\_date | 1 | 5 | 1.003 | 0.056 |
| firstname\_lastname\_ssn\_homephone\_dob\_count7\_date | 1 | 6 | 1.006 | 0.083 |
| firstname\_lastname\_ssn\_homephone\_dob\_count14\_date | 1 | 8 | 1.012 | 0.116 |
| firstname\_lastname\_ssn\_homephone\_dob\_count30\_date | 1 | 11 | 1.025 | 0.172 |
| firstname\_lastname\_ssn\_homephone\_dob\_count90\_date | 1 | 26 | 1.069 | 0.323 |
| firstname\_lastname\_ssn\_homephone\_dob\_count180\_date | 1 | 39 | 1.118 | 0.474 |
| firstname\_lastname\_ssn\_address\_zip5\_count0\_date | 1 | 3 | 1.001 | 0.023 |
| firstname\_lastname\_ssn\_address\_zip5\_count1\_date | 1 | 3 | 1.001 | 0.038 |
| firstname\_lastname\_ssn\_address\_zip5\_count3\_date | 1 | 5 | 1.003 | 0.059 |
| firstname\_lastname\_ssn\_address\_zip5\_count7\_date | 1 | 6 | 1.007 | 0.087 |
| firstname\_lastname\_ssn\_address\_zip5\_count14\_date | 1 | 8 | 1.014 | 0.122 |
| firstname\_lastname\_ssn\_address\_zip5\_count30\_date | 1 | 11 | 1.028 | 0.181 |
| firstname\_lastname\_ssn\_address\_zip5\_count90\_date | 1 | 26 | 1.077 | 0.335 |
| firstname\_lastname\_ssn\_address\_zip5\_count180\_date | 1 | 39 | 1.132 | 0.488 |
| ssn\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.008 |
| ssn\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| ssn\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_0\_count3\_count\_ratio | 0.020 | 0.333 | 0.332 | 0.017 |
| address\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.009 |
| address\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.006 |
| address\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.004 |
| address\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| address\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.011 |
| address\_1\_count7\_count\_ratio | 0.012 | 0.143 | 0.142 | 0.008 |
| address\_1\_count14\_count\_ratio | 0.005 | 0.071 | 0.071 | 0.005 |
| address\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| address\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| address\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_0\_count3\_count\_ratio | 0.010 | 0.333 | 0.282 | 0.081 |
| homephone\_0\_count7\_count\_ratio | 0.004 | 0.143 | 0.100 | 0.042 |
| homephone\_0\_count14\_count\_ratio | 0.002 | 0.071 | 0.038 | 0.022 |
| homephone\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.012 | 0.010 |
| homephone\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.003 | 0.003 |
| homephone\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.001 | 0.002 |
| homephone\_1\_count3\_count\_ratio | 0.013 | 0.333 | 0.299 | 0.068 |
| homephone\_1\_count7\_count\_ratio | 0.004 | 0.143 | 0.106 | 0.040 |
| homephone\_1\_count14\_count\_ratio | 0.002 | 0.071 | 0.041 | 0.022 |
| homephone\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.013 | 0.010 |
| homephone\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.003 | 0.003 |
| homephone\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.001 | 0.002 |
| firstname\_lastname\_dob\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| firstname\_lastname\_dob\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| firstname\_lastname\_dob\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| firstname\_lastname\_dob\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_dob\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| firstname\_lastname\_dob\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| firstname\_lastname\_dob\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_dob\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_zip5\_0\_count3\_count\_ratio | 0.020 | 0.333 | 0.332 | 0.016 |
| address\_zip5\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.008 |
| address\_zip5\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| address\_zip5\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| address\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| address\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_zip5\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.010 |
| address\_zip5\_1\_count7\_count\_ratio | 0.012 | 0.143 | 0.142 | 0.007 |
| address\_zip5\_1\_count14\_count\_ratio | 0.005 | 0.071 | 0.071 | 0.005 |
| address\_zip5\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| address\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| address\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_firstname\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| ssn\_firstname\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_firstname\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| ssn\_firstname\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_firstname\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_firstname\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_firstname\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_firstname\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_lastname\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_lastname\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| ssn\_lastname\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_lastname\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_lastname\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| ssn\_lastname\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_lastname\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_lastname\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_lastname\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_lastname\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_lastname\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_lastname\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_lastname\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_firstname\_lastname\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| ssn\_firstname\_lastname\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_firstname\_lastname\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_lastname\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.002 |
| ssn\_firstname\_lastname\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_lastname\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_firstname\_lastname\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_firstname\_lastname\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_firstname\_lastname\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_lastname\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_firstname\_lastname\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_address\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| ssn\_address\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| ssn\_address\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_address\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_address\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_address\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_address\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| ssn\_address\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| ssn\_address\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_address\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_address\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_address\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_zip5\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| ssn\_zip5\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| ssn\_zip5\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_zip5\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_zip5\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| ssn\_zip5\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| ssn\_zip5\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_zip5\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_dob\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_dob\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| ssn\_dob\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_dob\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_dob\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_dob\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_dob\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_dob\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_homephone\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| ssn\_homephone\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| ssn\_homephone\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_homephone\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_homephone\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_homephone\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_homephone\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| ssn\_homephone\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| ssn\_homephone\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_homephone\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_homephone\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_homephone\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_lastname\_dob\_0\_count3\_count\_ratio | 0.015 | 0.333 | 0.332 | 0.013 |
| ssn\_firstname\_lastname\_dob\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| ssn\_firstname\_lastname\_dob\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| ssn\_firstname\_lastname\_dob\_0\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_lastname\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_firstname\_lastname\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_firstname\_lastname\_dob\_1\_count3\_count\_ratio | 0.048 | 0.333 | 0.333 | 0.009 |
| ssn\_firstname\_lastname\_dob\_1\_count7\_count\_ratio | 0.010 | 0.143 | 0.142 | 0.006 |
| ssn\_firstname\_lastname\_dob\_1\_count14\_count\_ratio | 0.004 | 0.071 | 0.071 | 0.004 |
| ssn\_firstname\_lastname\_dob\_1\_count30\_count\_ratio | 0.002 | 0.033 | 0.033 | 0.003 |
| ssn\_firstname\_lastname\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_firstname\_lastname\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_address\_zip5\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| ssn\_address\_zip5\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| ssn\_address\_zip5\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_address\_zip5\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_address\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_address\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_address\_zip5\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| ssn\_address\_zip5\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| ssn\_address\_zip5\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| ssn\_address\_zip5\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| ssn\_address\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| ssn\_address\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_firstname\_lastname\_dob\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| address\_firstname\_lastname\_dob\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| address\_firstname\_lastname\_dob\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| address\_firstname\_lastname\_dob\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| address\_firstname\_lastname\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| address\_firstname\_lastname\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_firstname\_lastname\_dob\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| address\_firstname\_lastname\_dob\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| address\_firstname\_lastname\_dob\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| address\_firstname\_lastname\_dob\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| address\_firstname\_lastname\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| address\_firstname\_lastname\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| zip5\_firstname\_lastname\_dob\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| zip5\_firstname\_lastname\_dob\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| zip5\_firstname\_lastname\_dob\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| zip5\_firstname\_lastname\_dob\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| zip5\_firstname\_lastname\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| zip5\_firstname\_lastname\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| zip5\_firstname\_lastname\_dob\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| zip5\_firstname\_lastname\_dob\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| zip5\_firstname\_lastname\_dob\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| zip5\_firstname\_lastname\_dob\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| zip5\_firstname\_lastname\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| zip5\_firstname\_lastname\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_zip5\_firstname\_lastname\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| address\_zip5\_firstname\_lastname\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| address\_zip5\_firstname\_lastname\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| address\_zip5\_firstname\_lastname\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| address\_zip5\_firstname\_lastname\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| address\_zip5\_firstname\_lastname\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| address\_zip5\_firstname\_lastname\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| address\_zip5\_firstname\_lastname\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| address\_zip5\_firstname\_lastname\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| address\_zip5\_firstname\_lastname\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| address\_zip5\_firstname\_lastname\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| address\_zip5\_firstname\_lastname\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_firstname\_lastname\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| homephone\_firstname\_lastname\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| homephone\_firstname\_lastname\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_firstname\_lastname\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_firstname\_lastname\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_firstname\_lastname\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_firstname\_lastname\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| homephone\_firstname\_lastname\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| homephone\_firstname\_lastname\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_firstname\_lastname\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_firstname\_lastname\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_firstname\_lastname\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_zip\_0\_count3\_count\_ratio | 0.020 | 0.333 | 0.332 | 0.013 |
| homephone\_zip\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| homephone\_zip\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| homephone\_zip\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_zip\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_zip\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_zip\_1\_count3\_count\_ratio | 0.056 | 0.333 | 0.333 | 0.009 |
| homephone\_zip\_1\_count7\_count\_ratio | 0.012 | 0.143 | 0.142 | 0.006 |
| homephone\_zip\_1\_count14\_count\_ratio | 0.005 | 0.071 | 0.071 | 0.004 |
| homephone\_zip\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_zip\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_zip\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_dob\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| homephone\_dob\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| homephone\_dob\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_dob\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_dob\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| homephone\_dob\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| homephone\_dob\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_dob\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_address\_0\_count3\_count\_ratio | 0.020 | 0.333 | 0.332 | 0.013 |
| homephone\_address\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| homephone\_address\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| homephone\_address\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_address\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_address\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_address\_1\_count3\_count\_ratio | 0.056 | 0.333 | 0.333 | 0.009 |
| homephone\_address\_1\_count7\_count\_ratio | 0.012 | 0.143 | 0.142 | 0.006 |
| homephone\_address\_1\_count14\_count\_ratio | 0.005 | 0.071 | 0.071 | 0.004 |
| homephone\_address\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_address\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_address\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_firstname\_lastname\_dob\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| homephone\_firstname\_lastname\_dob\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| homephone\_firstname\_lastname\_dob\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_firstname\_lastname\_dob\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_firstname\_lastname\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_firstname\_lastname\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_firstname\_lastname\_dob\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| homephone\_firstname\_lastname\_dob\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| homephone\_firstname\_lastname\_dob\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| homephone\_firstname\_lastname\_dob\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| homephone\_firstname\_lastname\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_firstname\_lastname\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_address\_zip5\_0\_count3\_count\_ratio | 0.020 | 0.333 | 0.332 | 0.013 |
| homephone\_address\_zip5\_0\_count7\_count\_ratio | 0.006 | 0.143 | 0.142 | 0.007 |
| homephone\_address\_zip5\_0\_count14\_count\_ratio | 0.003 | 0.071 | 0.071 | 0.005 |
| homephone\_address\_zip5\_0\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_address\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_address\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| homephone\_address\_zip5\_1\_count3\_count\_ratio | 0.056 | 0.333 | 0.333 | 0.009 |
| homephone\_address\_zip5\_1\_count7\_count\_ratio | 0.012 | 0.143 | 0.142 | 0.006 |
| homephone\_address\_zip5\_1\_count14\_count\_ratio | 0.005 | 0.071 | 0.071 | 0.004 |
| homephone\_address\_zip5\_1\_count30\_count\_ratio | 0.001 | 0.033 | 0.033 | 0.003 |
| homephone\_address\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| homephone\_address\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_homephone\_dob\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.009 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.006 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_address\_zip5\_0\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count3\_count\_ratio | 0.083 | 0.333 | 0.333 | 0.007 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count7\_count\_ratio | 0.024 | 0.143 | 0.142 | 0.005 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count14\_count\_ratio | 0.009 | 0.071 | 0.071 | 0.004 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count30\_count\_ratio | 0.003 | 0.033 | 0.033 | 0.003 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count90\_count\_ratio | 0.000 | 0.011 | 0.011 | 0.001 |
| firstname\_lastname\_ssn\_address\_zip5\_1\_count180\_count\_ratio | 0.000 | 0.006 | 0.005 | 0.001 |
| ssn\_pastday | 0 | 366 | 164.555 | 105.243 |
| address\_pastday | 0 | 366 | 161.061 | 105.586 |
| homephone\_pastday | 0 | 366 | 23.815 | 58.102 |
| firstname\_lastname\_dob\_pastday | 0 | 366 | 166.579 | 105.452 |
| address\_zip5\_pastday | 0 | 366 | 163.551 | 105.331 |
| ssn\_firstname\_pastday | 0 | 366 | 164.765 | 105.237 |
| ssn\_lastname\_pastday | 0 | 366 | 164.770 | 105.239 |
| ssn\_firstname\_lastname\_pastday | 0 | 366 | 164.826 | 105.240 |
| ssn\_address\_pastday | 0 | 366 | 165.611 | 105.071 |
| ssn\_zip5\_pastday | 0 | 366 | 165.560 | 105.066 |
| ssn\_dob\_pastday | 0 | 366 | 166.667 | 105.473 |
| ssn\_homephone\_pastday | 0 | 366 | 166.469 | 105.190 |
| ssn\_firstname\_lastname\_dob\_pastday | 0 | 366 | 166.740 | 105.473 |
| ssn\_address\_zip5\_pastday | 0 | 366 | 165.657 | 105.072 |
| address\_firstname\_lastname\_dob\_pastday | 0 | 366 | 167.337 | 105.265 |
| zip5\_firstname\_lastname\_dob\_pastday | 0 | 366 | 167.301 | 105.262 |
| address\_zip5\_firstname\_lastname\_pastday | 0 | 366 | 165.458 | 105.041 |
| homephone\_firstname\_lastname\_pastday | 0 | 366 | 166.375 | 105.167 |
| homephone\_zip\_pastday | 0 | 366 | 165.542 | 105.378 |
| homephone\_dob\_pastday | 0 | 366 | 167.455 | 105.308 |
| homephone\_address\_pastday | 0 | 366 | 165.864 | 105.408 |
| homephone\_firstname\_lastname\_dob\_pastday | 0 | 366 | 167.684 | 105.319 |
| homephone\_address\_zip5\_pastday | 0 | 366 | 165.911 | 105.409 |
| firstname\_lastname\_ssn\_homephone\_dob\_address\_zip5\_pastday | 0 | 366 | 166.774 | 105.207 |
| firstname\_lastname\_ssn\_homephone\_dob\_pastday | 0 | 366 | 167.798 | 105.332 |
| firstname\_lastname\_ssn\_address\_zip5\_pastday | 0 | 366 | 165.692 | 105.072 |
| fraud\_label | 0 | 1 | 0.014 | 0.119 |