DSO 562 Fraud Analytics
Project 2: Application Fraud Detection

Advisor: Prof. Stephen Coggeshall

Chinmayi
Bengaluru Prakash

March 24, 2022

TABLE OF CONTENTS

1. Executive Summary

Application fraud is a key concern in the phone applications and credit card industry where applicants apply for a product under a false identity which were either stolen or manipulated, resulting in identity fraud. In the credit card industry alone, it significantly impacts businesses affecting at least 5-10 million customers yearly in the US causing at least $10 billion fraud losses.

This report describes in detail the processes used to evaluate application data to capture fraud using supervised learning techniques. The raw dataset consists of 1 million rows and 10 columns. The process involved exploring and visualizing each variable in the dataset followed by cleaning the raw data to handle missing and frivolous entries. Further, we created additional variables by extracting features from the raw data resulting in a total of 1016 candidate variables. We followed this by performing feature selection using KS (Kolmogorov-Smirnov) score, first by selecting top 100 features across different groups of variables, and then narrowing down to 30 variables to feed into the models. We used random forest algorithm for wrapping to arrive at the top 30 features.

We considered logistic regression as the baseline model. We followed this up with other modeling techniques including decision trees, random forest, gradient boosting, and neural networks. Our dataset ranged from January to December 2016. We set the last two months of data (November and December 2016) for validation which made up 16.65% of the total data. The remaining 83.85% of data was divided into train and test datasets in 70:30 ratio.

Overall, we arrived at 2 models which had the highest fraud detection rate (FDR) of ~53.76% at 3% FDR. The neural network model (with 5 features, 2 layers each having 50 nodes) had a slightly better performance (0.04% more FDR) in comparison to the light gradient boosting (LGB) model (with 5 features, 31 max leaves). We decided to choose the LGB model as it is a more interpretable, less complex, and faster modeling technique.

2. Description of the data

The data used in this project consists of personal identifying information (PII) for 1,000,000 (1M) cell phone and credit card applicants from the year 2016. The data set has a total of 10 fields - eight categorical fields and two numerical fields. All the data contained in this data set is synthetic but retains the properties of real PII data.

2.1 Categorical summary table

Field Name	# Non-null records	% Populated	# Unique Values	Most Common Value
record	1,000,000	100.00%	1,000,000	NA
ssn	1,000,000	100.00%	835,819	999999999
firstname	1,000,000	100.00%	78,136	EAMSTRMT
lastname	1,000,000	100.00%	177,001	ERJSAXA
address	1,000,000	100.00%	828,774	123 MAIN ST
zip5	1,000,000	100.00%	26,370	68138
homephone	1,000,000	100.00%	28,244	9999999999
fraud_label	1,000,000	100.00%	2	'0'

2.2 Numerical summary table

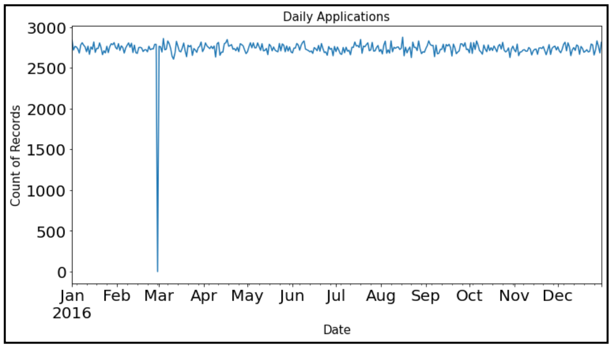
Field Name	# Non-null records	% Populated	Min	Max	Mean	Standard Deviation	% Zero
date	1,000,000	100.00%	2016-01-01	2016-12-31	-	-	0.00%
dob	1,000,000	100.00%	1900-01-01	2016-10-31	-	-	0.00%

2.3 Fields of Importance

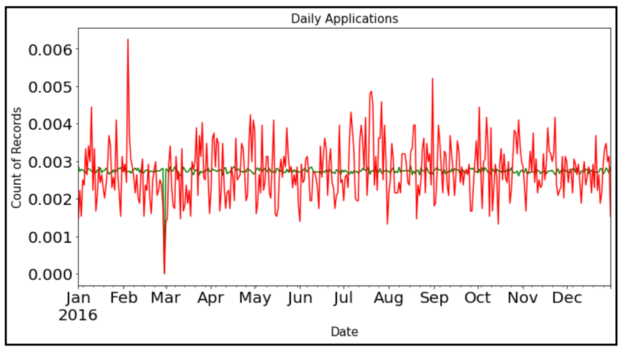
This section provides further details regarding some fields displayed inconsistencies and concerning characteristics. The full data quality report with field descriptions and descriptive graphs can be found in the appendix.

2.3.1 Date ('date')

The ‘date’ field is a numerical field and represents the date and time for when the application was made or submitted. Our analysis of this field revealed missing data. This can also be seen in the two graphs below.



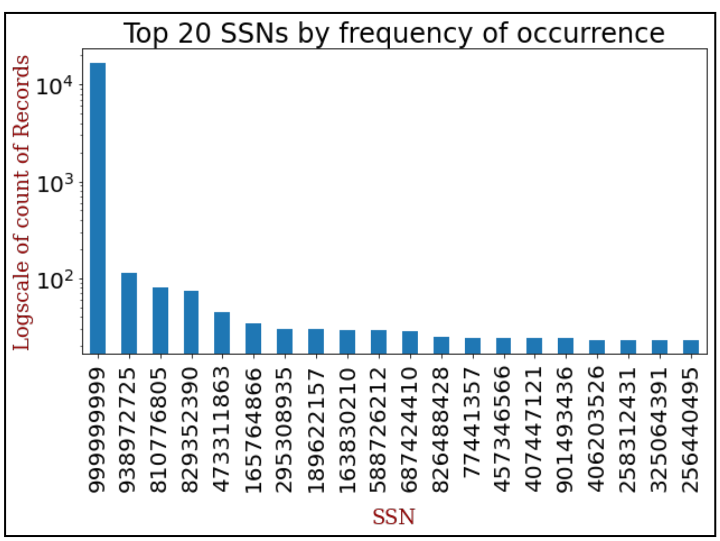
This graph plots the count of records (y-axis) and the date/month the application was submitted (x-axis). There is a notable drop in the number of applications in late February. This drop is a result of missing data in this field.



The graph above plots the count of records (y-axis) and the date/month the application was submitted (x-axis) and bifurcates the count based on the fraud label of the record. Here red represents fraudulent records while green represents authentic records.

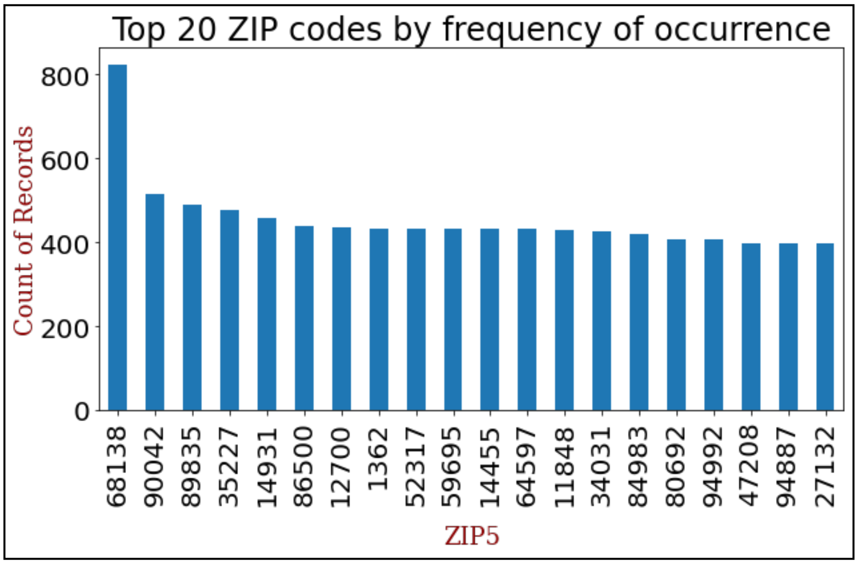
2.3.2 Social Security Number ('ssn')

The ‘ssn’ field is a categorical field that contains the social security numbers (SSN) used for the applications. Our analysis showed that the most used SSN was ‘999999999’ which accounts for approximately 1.69% of the total number of records for this field. This discrepancy is visualized in the graph below:



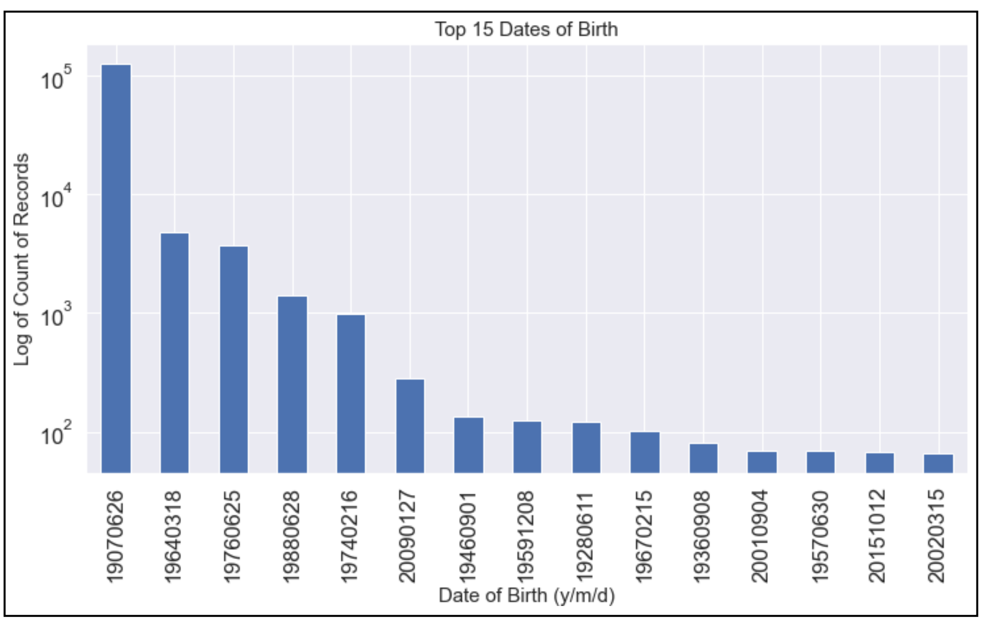
2.3.3 Zip Codes ('zip5')

The 'zip5' field is a categorical field that contains the five-digit zip code associated with the address provided in the application. The zip code '68138' was the most used zip code in the entire data set and accounts for approximately 0.08% of the data in this field. This discrepancy is highlighted in the graph below:



2.3.4 Date of Birth ('dob')

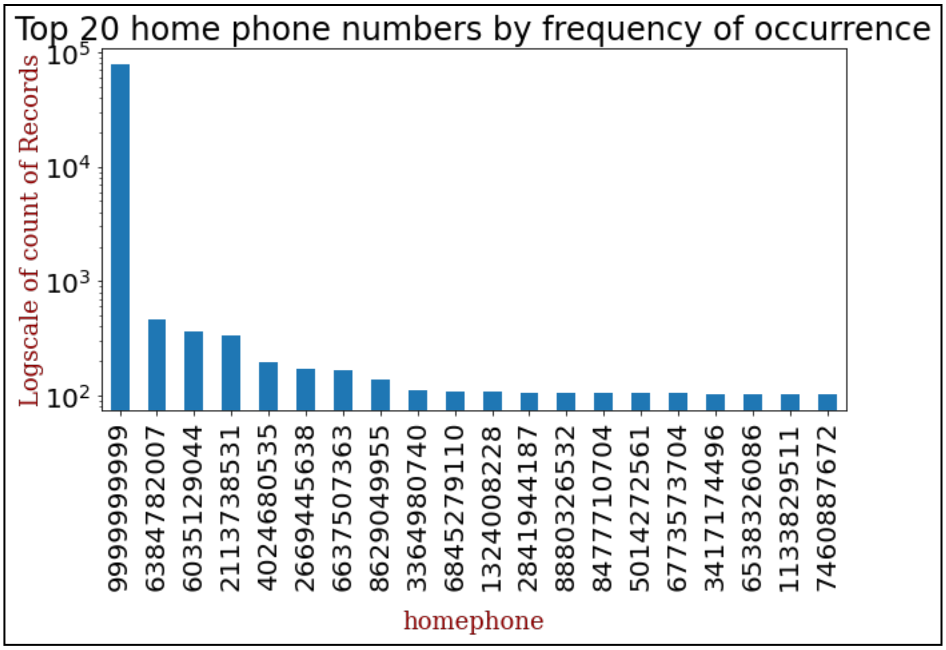
The 'dob' field is a numerical field that contains the date of birth provided by the applicants in their applications. The most common date of birth used for applications recorded in the data set was '06/26/1907', accounting for approximately 12.66% of the total records in this field. This discrepancy is visualized in the graph below:



This graph highlights the year 1907 to be the most common year used for applications.

2.3.5 Home Phone Number ('homephone')

The 'homephone' field is a categorical field that contains the home phone number used for the applications. Our analysis showed that the most used home phone number was '9999999999' which accounts for approximately 7.85% of the total number of records for this field. This discrepancy is highlighted in the graph below:



3. Data Cleaning:

Prior to fitting any machine learning model to a dataset, it is necessary to clean the data by handling missing/null values, frivolous values, corrupted data, irrelevant and inaccurate data. These data points need to be replaced or modified or deleted before proceeding ahead.

3.1 Handling Frivolous Values

The fields zip5, ssn, homephone, address and dob had frivolous values. These were corrected as follows:

1. zip5: Certain zip codes such as 1362 had less than 5 digits. These were handled by front padding the entries with zeros
2. ssn:
    a. There were 16,935 entries of SSN as 999999999 which we assumed to be potentially missing. These were replaced by negative of the corresponding RECORD number to ensure they don't affect the analysis
    b. Like zip5, SSNs which were shorter than 9 digits were handled by front padding with zeros
3. homephone:
    a. There were 78,512 occurrences of homephone as 9999999999. These were replaced with the negative of the corresponding RECORD number
    b. Values with length less than 10 digits were front padded with zeros
4. address: Entries for address '123 MAIN ST' with 1079 entries were assumed to be missing. These addresses were replaced with a concatenated string of record number and adding the string 'RECORD' as suffix
5. dob: The date of birth 19070626 occurs 126,568 times. We assumed that this date was filled as a default value for missing/erroneous data. These were replaced by the negative of the RECORD column

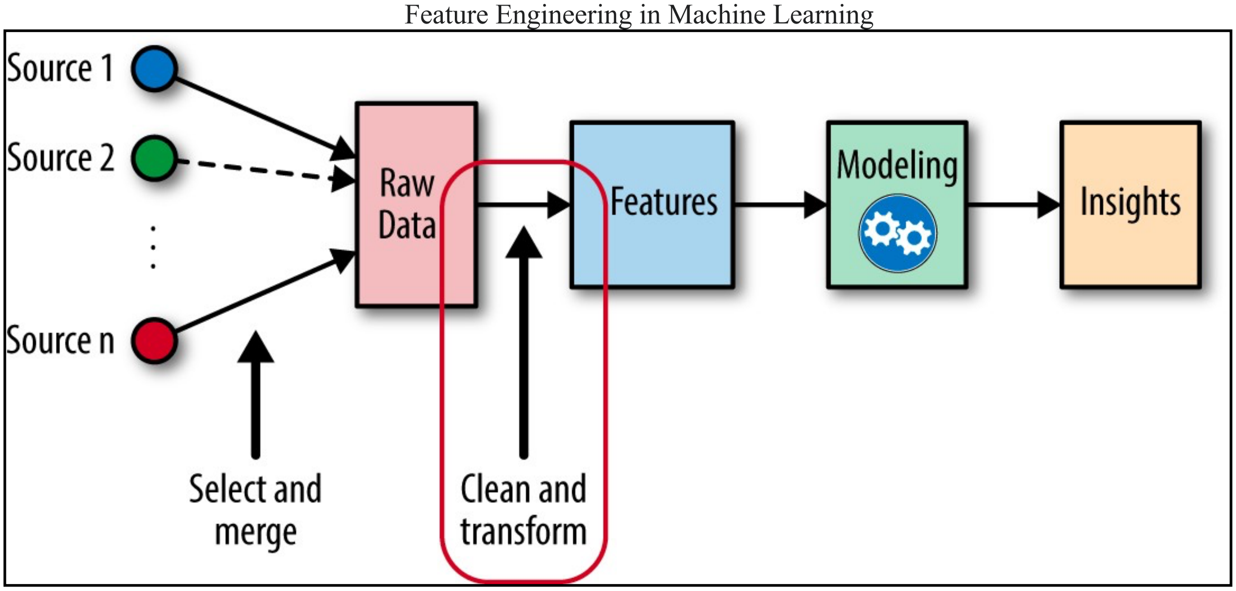
Note: The dataset did not have any null values

4. Candidate Variables:

4.1 Feature Engineering:
Feature engineering (or feature extraction) is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data. New variables that are not present in the training set are created using this machine learning technique. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy. A feature is any measurable input that can be used in a predictive model.

It is the second most important step in the entire life cycle of model building after designing a solution approach.

Feature Engineering in Machine Learning



In our project, we have carried out feature extraction process across the following steps:

- Target encoding
- Statistical smoothing
- Creation of attributes
- Creation of 'velocity' and 'days since' variables
- Creation of Relative Velocity variables
- Creation of Entity Count variables
- Creation of Interesting variables

4.2 Target Encoding:

Encoding categorical variables is a very important step in feature engineering. Generally, encoding of categorical variables is a procedure of replacing categorical variables with one or more numeric variables, so that the resulting data set may be used in the statistical and machine learning algorithms that expect numeric variables. There are many encoding techniques including one-hot encoding, ordinal encoding, target encoding and Bayesian target encoding.

We compute the mean of the target variable for each possible category and encode that category with the target mean. Hence, each categorical field becomes a numeric variable. This technique works for both binary classification and regression. For multiclass classification a similar technique is applied, where we encode the categorical variable with (m-1) new variables, where m is the number of classes.

A good practice of target encoding is to build a table with values for each category. For instance, consider a categorical field, say xx, that has possible values aa, bb, cc, dd

- Calculate $v_{aa} = \langle y \rangle | xx=aa$, $v_{bb} = \langle y \rangle | xx=bb$, $v_{cc} = \langle y \rangle | xx=cc$, $v_{dd} = \langle y \rangle | xx=dd$
- Then encode xx as:

$$xx \Rightarrow xx_{new} = \begin{cases} v_{aa} & \text{when } xx=aa \\ v_{bb} & \text{when } xx=bb \\ v_{cc} & \text{when } xx=cc \\ v_{dd} & \text{when } xx=dd \end{cases}$$

In our project, we carry out target encoding (aka Risk Tables) as the cardinality (number of categories) is more than 2. This method ensures no dimensionality expansion, direct encoding of what we are trying to predict and is the easiest for the model to figure out the relationship y = f(x). However, there is still a problem of losing interaction information and overfitting.

Overfitting can be avoided when target encoding is used by ensuring the following:

- Training data to be used when calculating the table values
- Statistically sufficient sample is present in each category, say, at least several dozen records
- If there are not enough records in a category, then use expert knowledge to group categories, for example, combine A and B together into a single category
- Use a smoothing formula (described in section 1.3)
- If there is still overfitting after the above methods are ensured, then systematically remove any of these table variables and observe the result. This will identify any table variable that is overfitting.

In this project, we do target encoding for day of week to create a ‘dow’ variable. Train test data is separated from the validation data (data with date above ‘2016-11-01’)

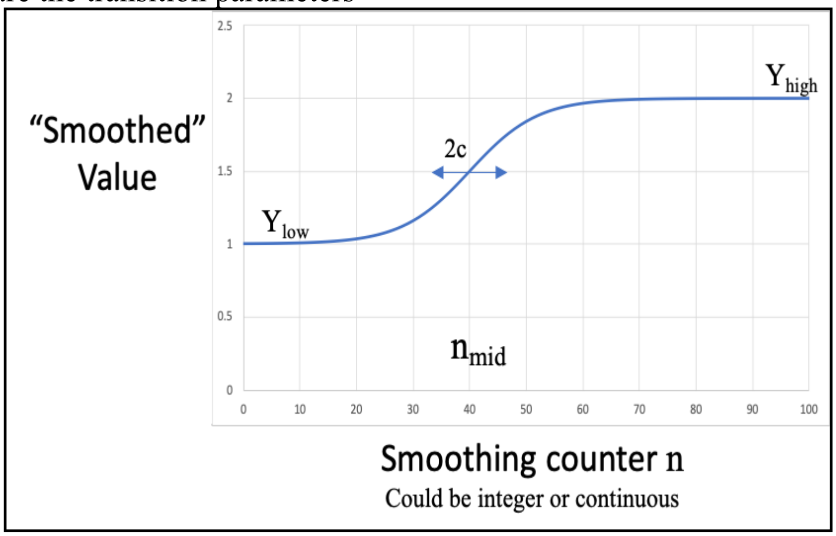
4.3 Statistical Smoothing:

A smoothing formula is used to smoothly transition a value from one number to another. It is used in target encoding to overcome the problem of overfitting. We have used the following logistic formula as our smoothing formula in our project.

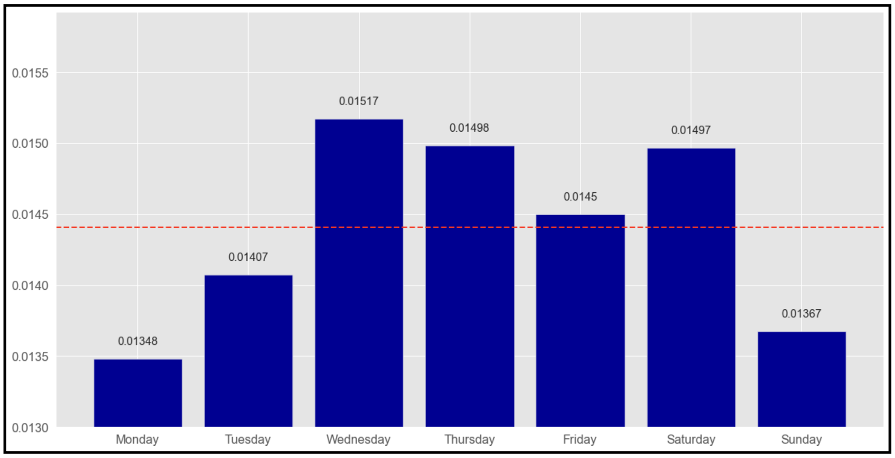
$$\boxed{\text{Value} = Y_{\text{low}} + \frac{Y_{\text{high}} - Y_{\text{low}}}{1 + e^{-(n - n_{mid})/c}}}$$

where: Y_{low} is one number
Y_{high} is the other number
n_{mid} is the value of n where the smoothed value is halfway between Y_{low} and Y_{high}
c is a measure of how quickly it transitions
n is the smoothing counter (integer/continuous)

Here, n_{mid} and c are the transition parameters



In our project, we have taken n_{mid} = 20 and c = 4



4.4 Rationale behind creation of variables:

An individual fraudster might be applying to various products or services with many stolen identities. This can be confirmed when he/she uses the same address or phone number in many applications containing different SSNs, Names and DOBs.

Individual Fraudster
(Address, Phone Number)

Victim 1 (SSN1, Name1, DOB1)
Victim 2 (SSN2, Name2, DOB2)
Victim 3 (SSN3, Name3, DOB3)
Victim 4 (SSN4, Name4, DOB4)
…

On the other hand, one victim's SSN, Name and DOB may be used by several fraudsters using these in multiple applications having different addresses and phone numbers.

Victim
(SSN, Name, DOB)

Fraudster1 (Address1, Phone number1)
Fraudster2 (Address2, Phone number2)
Fraudster3 (Address3, Phone number3)
Fraudster4 (Address4, Phone number4)
…

Therefore, we decided to create our candidate variables linking these features with one another. For example, we considered the number of applications with the same address, same phone number, number of SSNs with the same phone number or address etc. to capture all the possible combinations and in the end select the top 30 expert variables by feature selection techniques.

4.5 Creation of combination groups:

We created new features like 'name' by combining 'firstname' and 'lastname', 'fulladdress' by combining 'address' and 'zip5' and so on.

Here is a list of initial combination groups that we created out of linking entities like ssn, address, dob, first name, last name, fulladress and name features which are further used to create variables such as velocity variables, days-since variables, and relative velocity variables.

Combination groups
name_dob
name_fulladdress
name_homephone
fulladdress_dob
fulladdress_homephone
dob_homephone
homephone_name_dob
ssnfirstname
ssnlastname
ssnaddress
ssnzip5
ssndob
ssnhomephone
ssnname

4.6 Creation of variables:

For each entity or combination group, we create a set of following candidate/expert variables.

- Velocity variables
- Days-since variables
- Relative velocity variables
- Entity count variables
- Interesting variables

4.6.1 Velocity variables:

The logic behind creating velocity variables is to understand how many times we encountered each entity or combination group over the past 'n' days where n = 0, 1, 3, 7, 14 and 30.

4.6.2 Days-since variables:

The idea behind the creation of days-since variables is to check how many days has it been since we last encountered each entity or combination group.

4.6.3 Relative velocity variables:

Relative velocity variables were created to determine the ratio of short-term velocity to a longer term averaged velocity.

Relative velocity:

$$\frac{\text{\# applications with that } \textbf{\textit{entity/combination group}} \text{ seen in the recent past}}{\text{\# applications with that } \textbf{\textit{same entity/combination group}} \text{ seen in the past } \{0, 1, 3, 7, 14, 30\} \text{ days}}$$

4.6.4 Entity count variables:

Entity count variables were created to calculate the occurrences of unique entities/combination groups for a particular entity/combination group over the past n days where n = 1, 3, 7, 14, 30 and 60.

4.6.5 Interesting variables:

We created an 'SNPD' variable which is the combination of all four PIIs (personal identity information) such as SSN, Name, Phone number and DOB.

Furthermore, we created a 'zip3' variable by considering only the last three digits of a traditional zip5 (5-digit zip code) variable. We also used this as an entity to further link it with other entities to form combination groups and hence the above-mentioned variables.

In addition, we created a 'short address' variable which includes only the street name and door number from the given address variable to zoom in further and investigate in our analysis to detect fraud. This was in turn used as an entity for linking with other combination groups and finally, used to create velocity, days-since, relative velocity and entity count variables.

4.6.6 Summary of variables:

<table>
<tr><th>Name</th><th>Description of Variables</th><th># Variables created</th></tr>
<tr><td>VELOCITY AND DAYS SINCE VARIABLES</td><td>Velocity variables: # Applications at that entity over the past n days.

Days Since variables: # Days since that entity has been last seen

Values of n are:
{0, 1, 3, 7, 14, 30}

Entities are:
1. 'address',
2. 'dob',
3. 'homephone',
4. 'name',
5. 'fulladdress',
6. 'name_dob',
7. 'name_fulladdress',
8. 'name_homephone',
9. 'fulladdress_dob',
10. 'fulladdress_homephone',
11. 'dob_homephone',
12. 'homephone_name_dob',
13. 'ssn_firstname',
14. 'ssn_lastname',
15. 'ssn_address',
16. 'ssn_zip5',
17. 'ssn_dob',
18. 'ssn_homephone',
19. 'ssn_name',
20. 'ssn_fulladdress',
21. 'ssn_name_dob'</td><td>154</td></tr>
<tr><td>RELATIVE VELOCITY VARIABLES</td><td>Relative velocity variables (ratio)

Applications with that entity seen in the recent past / # Applications with that same entity seen in the past {0, 1, 3, 7, 14, 30} days</td><td>176</td></tr>
</table>

Name	Description of Variables	# Variables created
	Entities are: 1. 'address', 2. 'dob', 3. 'homephone', 4. 'name', 5. 'fulladdress', 6. 'name_dob', 7. 'name_fulladdress', 8. 'name_homephone', 9. 'fulladdress_dob', 10. 'fulladdress_homephone', 11. 'dob_homephone', 12. 'homephone_name_dob', 13. 'ssn_firstname', 14. 'ssn_lastname', 15. 'ssn_address', 16. 'ssn_zip5', 17. 'ssn_dob', 18. 'ssn_homephone', 19. 'ssn_name', 20. 'ssn_fulladdress', 21. 'ssn_name_dob'	
ENTITY COUNT VARIABLES (UNIQUE)	**Entity Count Variables (Unique)** # unique entities for that particular entities over the past n days: **Values of n are:** {1, 3, 7, 14, 30, 60} **Entities are:** 1. 'ssn', 2. 'fulladdress', 3. 'name_dob', 4. 'name_fulladdress', 5. 'fulladdress_dob', 6. 'dob_homephone', 7. 'ssn_lastname', 8. 'ssn_zip5', 9. 'ssn_name', 10. 'ssn_fulladdress', 11. 'ssn_name_dob'	**660**

5. Feature Selection

After creating new variables, we perform a variable selection process for dimensionality reduction. We perform this by two step approach:

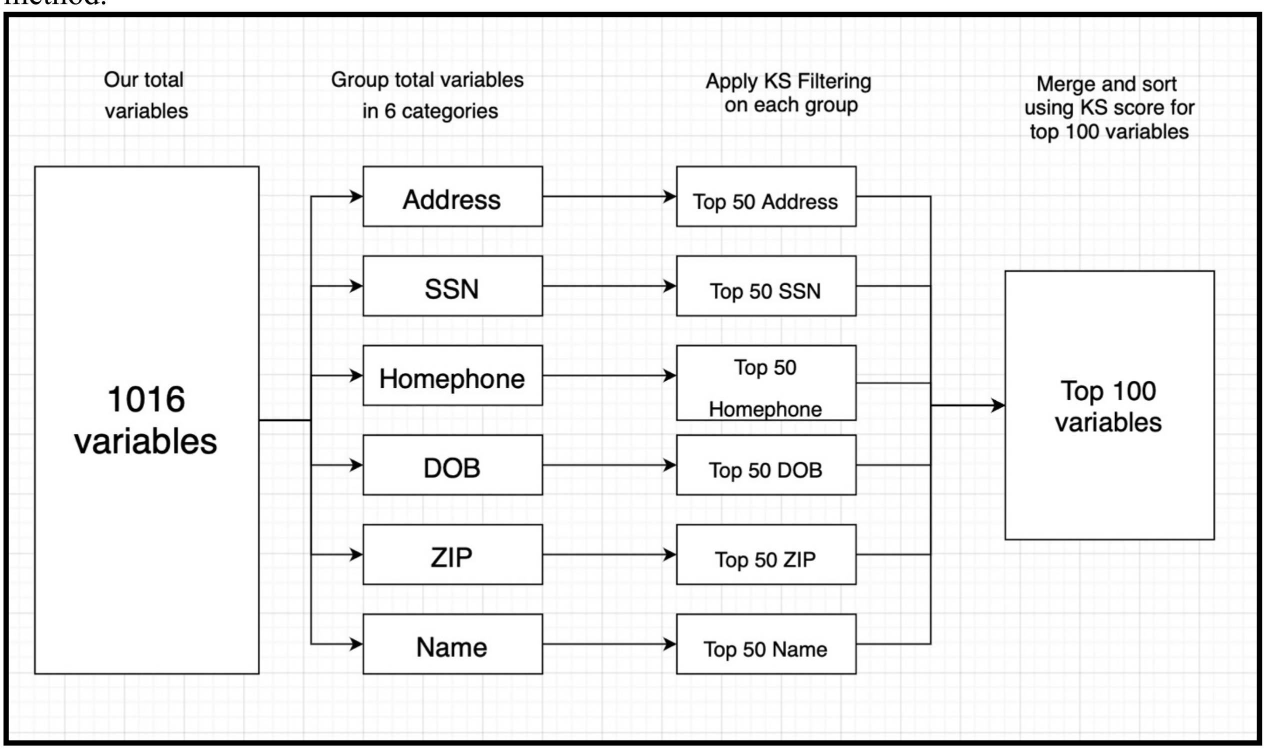
1- Filtering
2- Wrapping

5.1 Filtering

Feature selection is generally performed as a pre-processing step. The selection of features is independent of any machine learning algorithm. Instead, features are selected based on their scores in statistical scores for their correlation with the outcome variable. Depending on type of features and response variable different scores are used.

Feature / Response	**Continuous**	**Categorical**
Continuous	Pearson's Coeff or KS	LDA
Categorical	Anova	Chi-Squared

Since our feature and response variables are continuous, we used KS (Kolmogorov-Smirnov) method.



For the filtering process, following steps are performed:

- First, we break out data (1020 variables) into multiple groups based on different kinds of information they provide, for example, a set of variables containing address, or SSN etc. To make sure that all types of information is covered.
- Second, we perform KS filtering on each group to select top 100 features from each group
- Third, we merge all the selected features and sort them again based on KS score
- Lastly, we select the top 30 features

Note: This might not work for all problems but here it worked to increase the OOT accuracy since we were able to include SSN and homephone variables in the final set of variables.

5.2 Wrapping

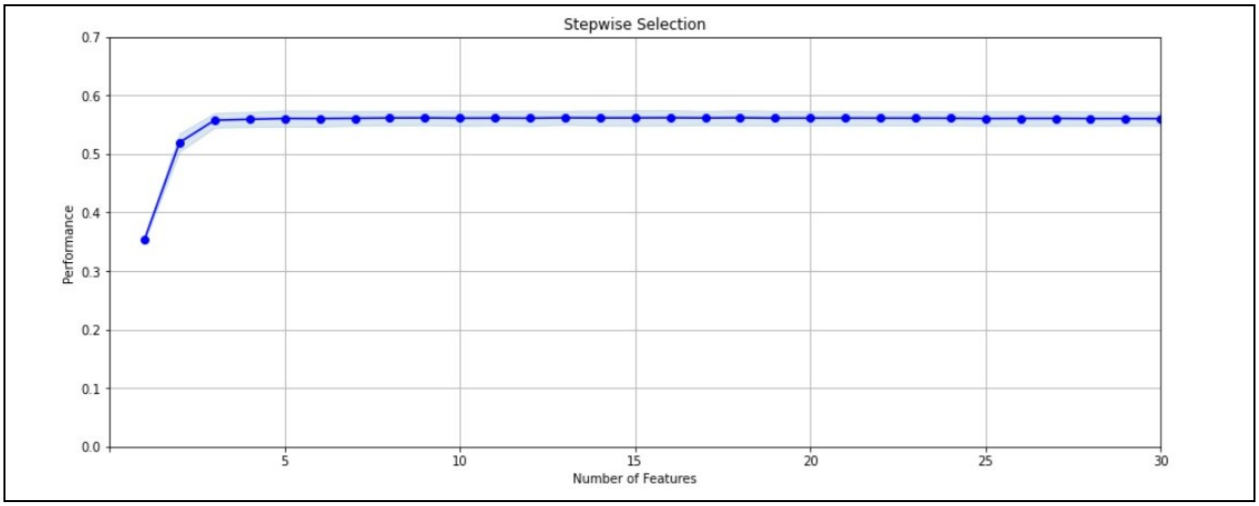
In the wrapper method, we take a subset of features and train the model on them. Based on the performance, we add or remove features to improve the model performance. In our project we use a forward feature selection method for wrapping.

5.2.1 Forward Feature Selection:

- Here we start with the best performing feature and then add only those features which increase the model performance.
- We used Random Forest as the machine learning algorithm for wrapping on the resultant variables from the filtering phase.
- After performing these two steps we came up with the following 30 variables:

	New Feature	**Average _score after wrapping**	**Individual KS Score**
1	fulladdress_count_30	0.3543	0.3320
2	ssn_count_30	0.5195	0.2270
3	homephone_count_3	0.5575	0.1949
4	fulladdress_unique_count_for_dob_homephone_60	0.5590	0.2883
5	fulladdress_unique_count_for_ssn_30	0.5605	0.2818
6	fulladdress_unique_count_for_name_dob_30	0.5603	0.2788
7	ssn_dob_count_30	0.5608	0.2285
8	fulladdress_unique_count_for_dob_homephone_30	0.5613	0.2828
9	ssn_name_count_0_by_30	0.5615	0.2043
10	fulladdress_unique_count_for_ssn_zip5_30	0.5609	0.2818
11	fulladdress_unique_count_for_ssn_name_dob_30	0.5611	0.2810

12	ssn_firstname_count_0_by_30	0.5609	0.2053
13	address_count_30	0.5616	0.3327
14	ssn_dob_count_14	0.5614	0.2149
15	ssn_firstname_count_30	0.5615	0.2261
16	name_dob_count_0_by_30	0.5617	0.2070
17	ssn_lastname_count_0_by_14	0.5613	0.1928
18	ssn_firstname_count_14	0.5616	0.2138
19	fulladdress_unique_count_for_ssn_name_dob_60	0.5610	0.2846
20	fulladdress_unique_count_for_ssn_zip5_60	0.5611	0.2866
21	fulladdress_unique_count_for_ssn_fulladdress_60	0.5611	0.2866
22	ssn_dob_count_0_by_30	0.5610	0.2077
23	ssn_name_dob_count_7	0.5609	0.1925
24	ssn_lastname_count_7	0.5609	0.1926
25	fulladdress_unique_count_for_name_dob_60	0.5604	0.2829
26	ssn_lastname_count_0_by_30	0.5606	0.2053
27	ssn_lastname_count_30	0.5606	0.2260
28	fulladdress_unique_count_for_fulladdress_dob_30	0.5603	0.2787
29	fulladdress_unique_count_for_name_fulladdress_60	0.5603	0.2844
30	ssn_count_0_by_14	0.5602	0.1938



The above chart indicates the performance achieved by the forward stepwise selection wrapper we have used for filtering our expert variables.

6. **Model Algorithms**

After getting the final variables we explore different machine learning algorithms to predict fraud. We have used the following algorithms:

- Logistic Regression
- Decision Tree
- Random forest
- Gradient boosting
- Neural Network

This section gives a brief idea about each one of these algorithms and then the results we got after using them with different hyperparameters.

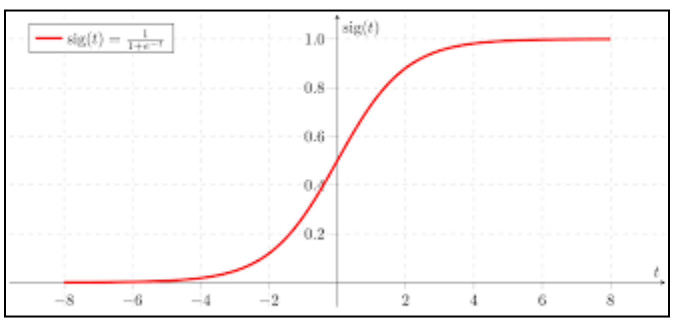
6.1 Logistic Regression

For this project we have used logistic regression as a baseline model since it's simple and linear. This method is used for classification and models the probability of response variable given independent variables using a logistic function. It can be summarized mathematically as follows:

$$\boxed{P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

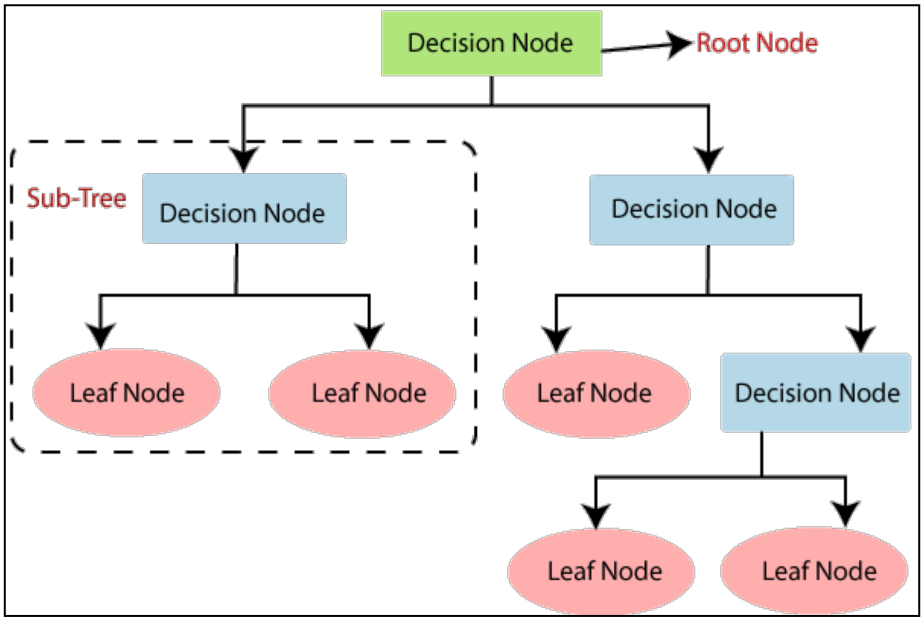
where P is Probability (Y = 1 | X = x)

The logistic or sigmoid function ensures that the modeled probability lies between 0 and 1.



6.2 Decision Tree

This is a non-linear classification method (could be used for regression problems as well) where the algorithm partitions the feature space till it achieves sufficient purity in each partition. In other words, it learns simple decision rules inferred from the data features to predict the value of Y. Decision trees are easy to interpret. Following picture shows the different components of the decision tree.



Decision trees can easily overfit the data hence some hyperparameters must be kept in mind to reduce overfitting.

Some common hyperparameters are:

- Criterion: Criterion to calculate impurity of nodes. Common values are gini and entropy. Gini impurity is the weighted average of the impurities of each node in a split done by a feature. Impurity of node is calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

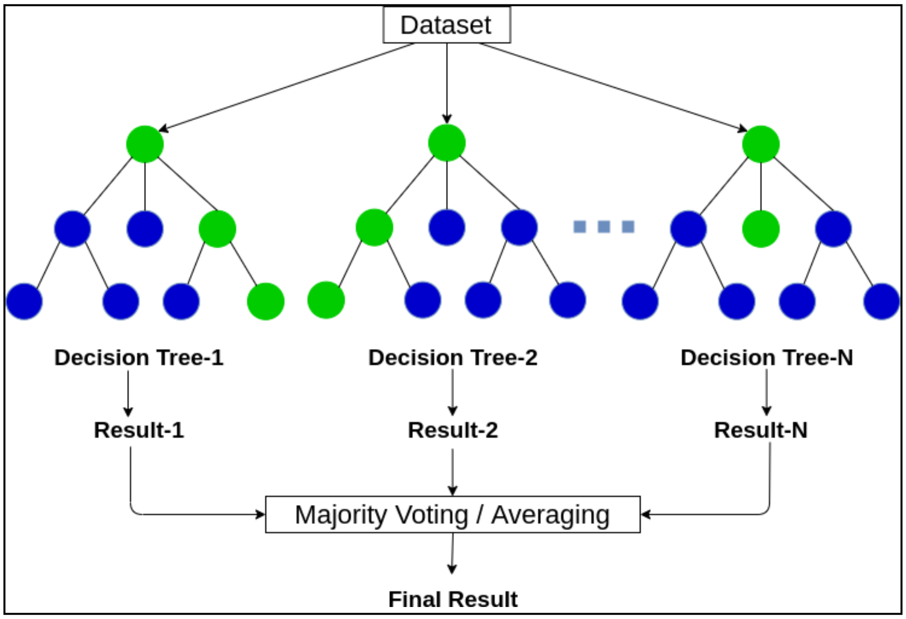
- Entropy is defined as follows and measures the impurity of each node

$$H[X] = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

- Max_depth: Maximum number of levels in the trained tree
- Min_samples_split: Minimum number of samples to be present at the node for it to be considered for splitting
- min_samples _leaf: Minimum number of samples to be present in the leaf nodes if the split were to happen

6.3 Random Forest

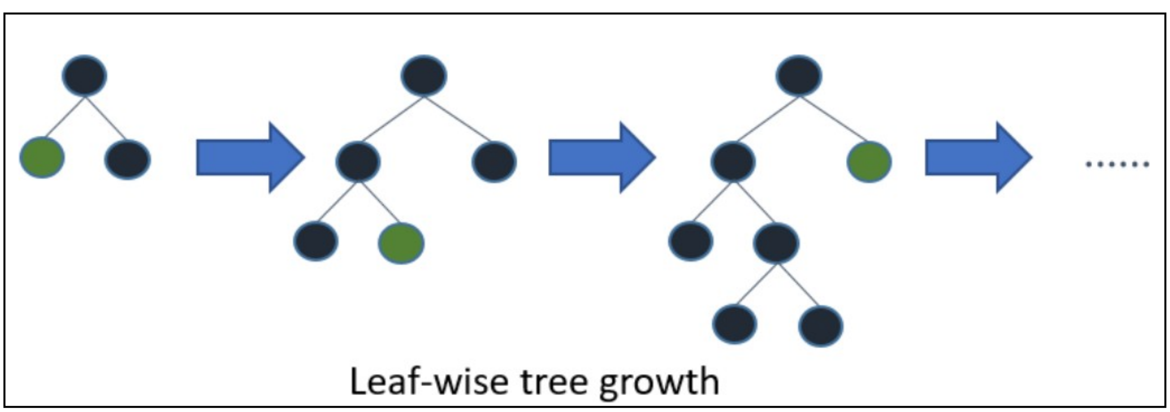
This is an ensemble of decision trees used for classification or regression problems. It creates multiple decision trees. Each individual tree is created through some sort of randomness for example using only a randomly chosen subset of variables or records for each tree and/or for each split iteration within a tree. Results from all trees are combined using voting or average.



Random forests improve over decision trees in terms of overfitting and stability. Hyperparameters for random forests that have been explored are the same as decision tree hyperparameters.

6.4 Light Gradient Boosting

Boosted trees are another way of improving on the drawback of decision trees. These are different from random forests because construction of trees is sequential and not random. Light GBM is a fast, distributed, high-performance gradient boosting framework based on a decision tree algorithm. Unlike other boosting algorithms, light GBM grows leaf wise. It also uses less memory and is faster than other boosting algorithms.

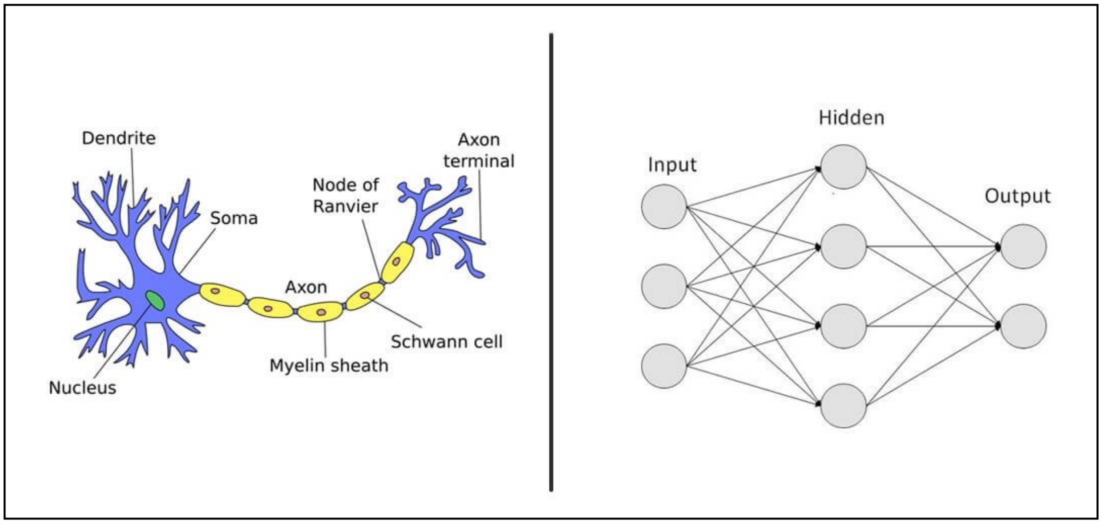


Following hyperparameters have been explored in this project:

- n_estimators: number of boosted trees to fit
- num_leaves: maximum tree leaves for base learners
- learning_rate: boosting learning rate
- boosting_type: Gradient Boosting Decision Tree (GBDT) or Dropouts meet Multiple Additive Regression Trees (DART)

6.5 Neural Network

Artificial neural networks (ANNs) are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. This Mechanism mimics how the human brain works.



Some hyperparameters:

- N_hidden: number of hidden layers
- hidden_layer_sizes: number of neurons in hidden layers
- activation: Activation function for the hidden layer
- solver: solver for weight optimization. Sgd implies stochastic gradient descent and adam implies different version of stochastic gradient descent
- alpha: L2 (regularization) parameter
- learning_rate: kind of learning. ‘constant’ is a constant learning rate given by ‘learning_rate_init’. ‘invscaling’ gradually decreases the learning rate at each time step ‘t’. ‘adaptive’ keeps the learning rate constant to ‘learning_rate_init’ as long as training loss keeps decreasing.
- learning_rate_init: The initial learning rate used

6.6 Model performance across algorithms

The following table shows the different algorithms and hyperparameters that were explored and the corresponding results. We tried a range of hyperparameters that would potentially underfit, fit properly and overfit the data.

1. Logistic regression: Since it is a baseline model, we did not explore many hyperparameters for this model. We altered the number of variables to see how the model performance changed.
2. Decision Tree: In scenario 6, we limit the depth to 2 to get a very simple tree and take 10 features to make that tree. This results in a poor performance on all sets due to underfitting. In scenario 8, we increased the depth to 10000 and number of features to 30 to try overfitting but it was not that evident.
    - Optimal hyperparameter set is scenario 7 with 0.536 FDR in validation set
3. Random Forest: scenario 10 produces a relatively simple model than other RF models with only 5 levels and 5 features in constituent trees. It does not perform that bad. Tried overfitting with *max_depth* of 1000 but again is not that evident.
    - Optimal hyperparameter set is scenario 8 with 0.537 FDR in validation set
4. Light Gradient Boosting: In scenario 7, an attempt to underfit with 5 leaves and 5 variables has been made but it still performs fine on the validation set. Increasing the number of leaves to 10,000 also did not overfit the model.
    - Optimal hyperparameter set is scenario 1 with 0.538 FDR in validation set
5. Neural Networks: The lowest performance is 0.214 FDR which is through an extremely simple model with 1 layer and 1 hidden node with *sgd* solver. The last scenario is comparable to logistic regression output with 20 variables.
    - Optimal hyperparameter set is scenario 3 with 0.538 FDR in validation set

Model Performance Across Algorithms

Model	Parameters	Average FDR at 3%		
Logistic Regression				
Scenario No	**# Variables**	**Train**	**Test**	**OOT**
1	5	0.515	0.515	0.497
2	10	0.533	0.536	0.515
3	15	0.532	0.536	0.517
4	20	0.542	0.549	0.524
5	30	0.545	0.549	0.525

Decision Tree									
Scenario No	**# Variables**	**criterion**	**max_depth**	**min_samples_split**	**max_features**	**min_samples_leaf**	**Train**	**Test**	**OOT**
1(default)	5	gini	None	2	5	1	0.563	0.545	0.533
2	15	gini	None	2	15	1	0.559	0.558	0.533
3	15	entropy	None	2	15	1	0.562	0.549	0.532
4	15	gini	4	2	15	1	0.524	0.527	0.502
5	15	entropy	1000	2	15	1	0.561	0.554	0.533
6	30	entropy	2	2	10	1	0.482	0.472	0.466
7	5	gini	None	100	5	50	0.556	0.563	0.536
8	30	gini	10000	2	30	1	0.561	0.552	0.529

Random Forest									
Scenario No	**# Variables**	**criterion**	**max_depth**	**min_samples_split**	**max_features**	**min_samples_leaf**	**Train**	**Test**	**OOT**
1(default)	5	gini	None	2	5	1	0.557	0.562	0.537
2	15	gini	None	2	15	1	0.558	0.565	0.536
3	15	entropy	None	2	15	1	0.562	0.554	0.537
4	15	gini	20	2	15	1	0.557	0.567	0.536
5	15	entropy	20	2	15	1	0.556	0.569	0.536
6	30	entropy	None	2	10	1	0.562	0.556	0.536
7	3	gini	None	2	3	1	0.557	0.553	0.536
8	30	gini	100	2	30	1	0.561	0.559	0.537
9	5	gini	None	100	5	50	0.557	0.562	0.536
10	5	entropy	5	2	5	1	0.543	0.547	0.522
11	5	entropy	1000	2	5	1	0.557	0.562	0.536

Light Gradient Boosting							
Scenario No	**# Variables**	**num_leaves**	**learning_rate**	**boosting_type**	**Train**	**Test**	**OOT**
1(default)	5	31	0.1	gbdt	0.559	0.557	**0.538**
2	15	31	0.1	gbdt	0.562	0.555	0.537
3	15	31	0.01	gbdt	0.561	0.556	0.536
4	15	31	0.1	dart	0.558	0.563	0.537
5	15	40	0.1	gbdt	0.561	0.558	0.537
6	30	40	0.01	dart	0.560	0.554	0.536
7	5	5	0.001	gbdt	0.529	0.534	0.505
8	5	10000	0.001	gbdt	0.558	0.559	0.537

Neural Network											
Scenario No	**# Variables**	**n_hidden**	**hidden_layer_sizes**	**activation**	**solver**	**alpha**	**learning_rate**	**learning_rate_init**	**Train**	**Test**	**OOT**
1	5	1	100	relu	adam	0.0001	constant	0.001	0.558	0.557	0.537
2	5	1	10	relu	adam	0.0001	invscaling	0.001	0.551	0.550	0.528
3	5	2	50	relu	adam	0.001	adaptive	0.001	0.561	0.552	**0.538**
4	5	2	50	logistic	adam	0.0001	constant	0.025	0.559	0.554	0.537
5	5	3	100	logistic	sgd	0.0001	constant	0.5	0.557	0.561	0.538
6	5	3	25	logistic	sgd	0.0001	constant	0.05	0.543	0.555	0.525
7	5	1	1	relu	adam	0.0001	constant	0.001	0.493	0.497	0.479
8	5	1	1	relu	sgd	0.0001	constant	0.001	0.212	0.209	0.214
8	5	1	1	logistic	adam	0.0001	constant	0.001	0.544	0.541	0.521

7. Results

Based on the various results in the previous section we have the two best models:

1. NN: Neural Network with 5 variables, two hidden layers with 50 nodes in each with other hyperparameters unaltered
2. LGB: Light Gradient boosting with 5 variables, 31 leaves (default), 0.1 learning rate and GBDT as the boosting type with other hyperparameters unaltered

Out of the two options we have chosen a less complex model of LGB to be our final model with 0.538 or 53.76% FDR in the top 3% population in the out of time validation set.
The following tables show how the FDR increases as we increase the percentage of the top population in training, test and validation set:

Training Set:

Training	#Records			#Goods			#Bads			Fraud Rate		
	583454			575081			8373			0.014350746		
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	5835	1551	4284	26.58	73.42	5835	1551	4284	0.27	51.16	50.89	0.36
2	5834	5523	311	94.67	5.33	11669	7074	4595	1.23	54.88	53.65	1.54
3	5835	5762	73	98.75	1.25	17504	12836	4668	2.23	55.75	53.52	2.75
4	5834	5769	65	98.89	1.11	23338	18605	4733	3.24	56.53	53.29	3.93
5	5835	5784	51	99.13	0.87	29173	24389	4784	4.24	57.14	52.90	5.10
6	5834	5789	45	99.23	0.77	35007	30178	4829	5.25	57.67	52.43	6.25
7	5835	5785	50	99.14	0.86	40842	35963	4879	6.25	58.27	52.02	7.37
8	5834	5793	41	99.30	0.70	46676	41756	4920	7.26	58.76	51.50	8.49
9	5835	5788	47	99.19	0.81	52511	47544	4967	8.27	59.32	51.05	9.57
10	5834	5790	44	99.25	0.75	58345	53334	5011	9.27	59.85	50.57	10.64
11	5835	5786	49	99.16	0.84	64180	59120	5060	10.28	60.43	50.15	11.68
12	5834	5793	41	99.30	0.70	70014	64913	5101	11.29	60.92	49.63	12.73
13	5835	5792	43	99.26	0.74	75849	70705	5144	12.29	61.44	49.14	13.75
14	5835	5805	30	99.49	0.51	81684	76510	5174	13.30	61.79	48.49	14.79
15	5834	5800	34	99.42	0.58	87518	82310	5208	14.31	62.20	47.89	15.80
16	5835	5791	44	99.25	0.75	93353	88101	5252	15.32	62.73	47.41	16.77
17	5834	5789	45	99.23	0.77	99187	93890	5297	16.33	63.26	46.94	17.73
18	5835	5793	42	99.28	0.72	105022	99683	5339	17.33	63.76	46.43	18.67
19	5834	5798	36	99.38	0.62	110856	105481	5375	18.34	64.19	45.85	19.62
20	5835	5802	33	99.43	0.57	116691	111283	5408	19.35	64.59	45.24	20.58

Test Set:

Test	#Records			#Goods			#Bads			Fraud Rate		
	250053			246419			3634			0.014532919		
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	2501	624	1877	24.95	75.05	2501	624	1877	0.25	51.65	51.40	0.33
2	2500	2378	122	95.12	4.88	5001	3002	1999	1.22	55.01	53.79	1.50
3	2501	2462	39	98.44	1.56	7502	5464	2038	2.22	56.08	53.86	2.68
4	2500	2482	18	99.28	0.72	10002	7946	2056	3.22	56.58	53.35	3.86
5	2501	2485	16	99.36	0.64	12503	10431	2072	4.23	57.02	52.78	5.03
6	2500	2479	21	99.16	0.84	15003	12910	2093	5.24	57.59	52.36	6.17
7	2501	2471	30	98.80	1.20	17504	15381	2123	6.24	58.42	52.18	7.24
8	2500	2481	19	99.24	0.76	20004	17862	2142	7.25	58.94	51.69	8.34
9	2501	2473	28	98.88	1.12	22505	20335	2170	8.25	59.71	51.46	9.37
10	2500	2491	9	99.64	0.36	25005	22826	2179	9.26	59.96	50.70	10.48
11	2501	2485	16	99.36	0.64	27506	25311	2195	10.27	60.40	50.13	11.53
12	2500	2486	14	99.44	0.56	30006	27797	2209	11.28	60.79	49.51	12.58
13	2501	2487	14	99.44	0.56	32507	30284	2223	12.29	61.17	48.88	13.62
14	2500	2485	15	99.40	0.60	35007	32769	2238	13.30	61.59	48.29	14.64
15	2501	2485	16	99.36	0.64	37508	35254	2254	14.31	62.03	47.72	15.64
16	2500	2482	18	99.28	0.72	40008	37736	2272	15.31	62.52	47.21	16.61
17	2501	2485	16	99.36	0.64	42509	40221	2288	16.32	62.96	46.64	17.58
18	2501	2482	19	99.24	0.76	45010	42703	2307	17.33	63.48	46.15	18.51
19	2500	2485	15	99.40	0.60	47510	45188	2322	18.34	63.90	45.56	19.46
20	2501	2488	13	99.48	0.52	50011	47676	2335	19.35	64.25	44.91	20.42

Validation Set:

Validation	#Records			#Goods			#Bads			Fraud Rate		
	166493			164107			2386			0.014330933		
	Bin Statistics					Cumulative Statistics						
Population Bin%	#Records	#Goods	#Bads	%Goods	%Bads	Total	Cumulative Goods	Cumulative Bads	%Goods	%Bads (FDR)	KS	FPR
1	1665	496	1169	29.79	70.21	1665	496	1169	0.30	48.99	48.69	0.42
2	1665	1577	88	94.71	5.29	3330	2073	1257	1.26	52.68	51.42	1.65
3	1665	1639	26	98.44	1.56	4995	3712	1283	2.26	53.77	51.51	2.89
4	1665	1653	12	99.28	0.72	6660	5365	1295	3.27	54.27	51.01	4.14
5	1665	1645	20	98.80	1.20	8325	7010	1315	4.27	55.11	50.84	5.33
6	1665	1647	18	98.92	1.08	9990	8657	1333	5.28	55.87	50.59	6.49
7	1665	1655	10	99.40	0.60	11655	10312	1343	6.28	56.29	50.00	7.68
8	1664	1646	18	98.92	1.08	13319	11958	1361	7.29	57.04	49.75	8.79
9	1665	1648	17	98.98	1.02	14984	13606	1378	8.29	57.75	49.46	9.87
10	1665	1658	7	99.58	0.42	16649	15264	1385	9.30	58.05	48.75	11.02
11	1665	1653	12	99.28	0.72	18314	16917	1397	10.31	58.55	48.24	12.11
12	1665	1659	6	99.64	0.36	19979	18576	1403	11.32	58.80	47.48	13.24
13	1665	1652	13	99.22	0.78	21644	20228	1416	12.33	59.35	47.02	14.29
14	1665	1652	13	99.22	0.78	23309	21880	1429	13.33	59.89	46.56	15.31
15	1665	1653	12	99.28	0.72	24974	23533	1441	14.34	60.39	46.05	16.33
16	1665	1649	16	99.04	0.96	26639	25182	1457	15.34	61.06	45.72	17.28
17	1665	1652	13	99.22	0.78	28304	26834	1470	16.35	61.61	45.26	18.25
18	1665	1654	11	99.34	0.66	29969	28488	1481	17.36	62.07	44.71	19.24
19	1665	1650	15	99.10	0.90	31634	30138	1496	18.36	62.70	44.33	20.15
20	1665	1654	11	99.34	0.66	33299	31792	1507	19.37	63.16	43.79	21.10

8. Conclusion

In this project, we went through the process of finding fraudulent credit application using machine learning. We started by initial exploratory data analysis and generating a Data Quality Report which outlines the overall distribution of data. Following which we cleaned the data, looking for any missing or outlier values.

After that we proceeded towards the most important step of the project, feature engineering. We believe that this the most important step, because with right set of features we can get great results even with linear models. Based on suggestions from our advisor Prof. Coggeshall, we created around one thousand new variables based on multiple combinations of given variables. Thereafter, we worked to reduce the dimensionality of our data from 1016 variables to top 30 features, based on techniques such as filtering (using KS method), and wrapping (forward feature selection using random forest).

Going forward we ran several linear and non-linear models on our data, using logistic regression as the base model, followed by decision trees, random forest, neural networks, light GBM etc. We chose light GBM as our final method which gave us the Fraud Detection Rate (FDR) of 53.76% at 3% of population, which means it can detect more than half of the fraud applications given only the top 3% of data sorted by our fraud algorithm score.

Every project has its own time limitations, but given more time, we would suggest the following improvements to our project:

- Interview more experts in this field to create more candidate variables
- If possible, get more data with fraud labels as 1
- Try more combinations of hyperparameters in tuning our models

9. Appendix:

9.1 DATA QUALITY REPORT

9.1.1 High Level Description:

- **File name:** "applications data.csv"
- **File description:** The dataset contains information of applicants applying for a product. It includes fields such as social security number, name, phone, date of birth, etc., with the PII data morphed. It also has a fraud label categorizing the entry as good or bad.
- **Time Period:** Jan 1 – Dec 31, 2016
- **Granularity:** One record for each application
- **Data Volume:** The raw dataset has
  - 10 fields (8 Categorical, 2 Numeric)
  - 1,000,000 records

9.1.2 Summary Table:

- The following 8 columns are **categorical**

Field Name	# Non-null records	% Populated	# Unique Values	Most Common Value
record	1,000,000	100.00%	1,000,000	NA
ssn	1,000,000	100.00%	835,819	999999999
firstname	1,000,000	100.00%	78,136	EAMSTRMT
lastname	1,000,000	100.00%	177,001	ERJSAXA
address	1,000,000	100.00%	828,774	123 MAIN ST
zip5	1,000,000	100.00%	26,370	68138
homephone	1,000,000	100.00%	28,244	9999999999
fraud_label	1,000,000	100.00%	2	'0'

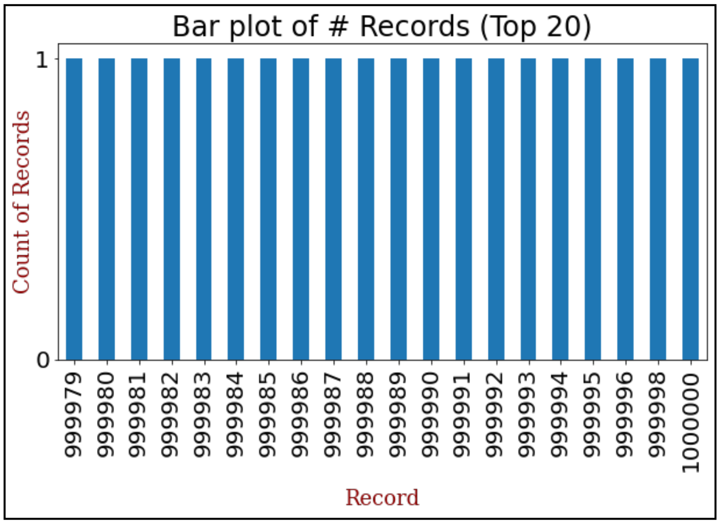
- The following 2 columns are **numeric**

FieldName	# Non-null records	% Populated	Min	Max	Mean	Standard Deviation	% Zero
date	1,000,000	100.00%	2016-01-01	2016-12-31	-	-	0.00%
dob	1,000,000	100.00%	1900-01-01	2016-10-31	-	-	0.00%

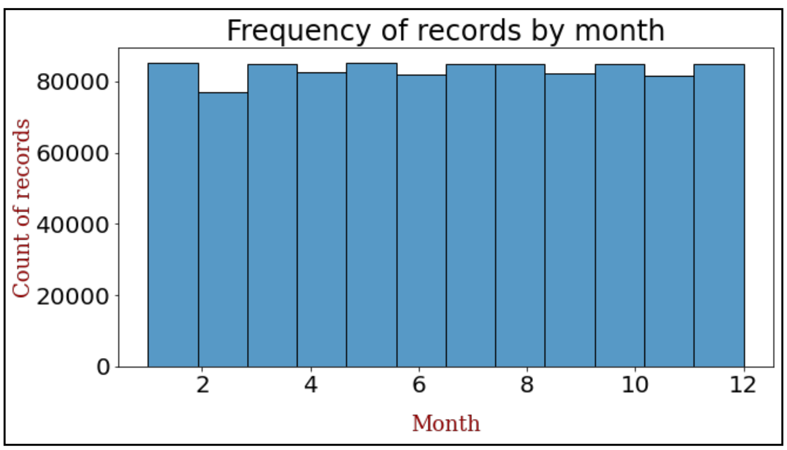
9.1.3 Description of Fields:

1. record (Categorical)
   - The record field is a unique ID assigned for each entry in the dataset.

- The graph below represents the top 20 values all having a minimum and maximum value of 1

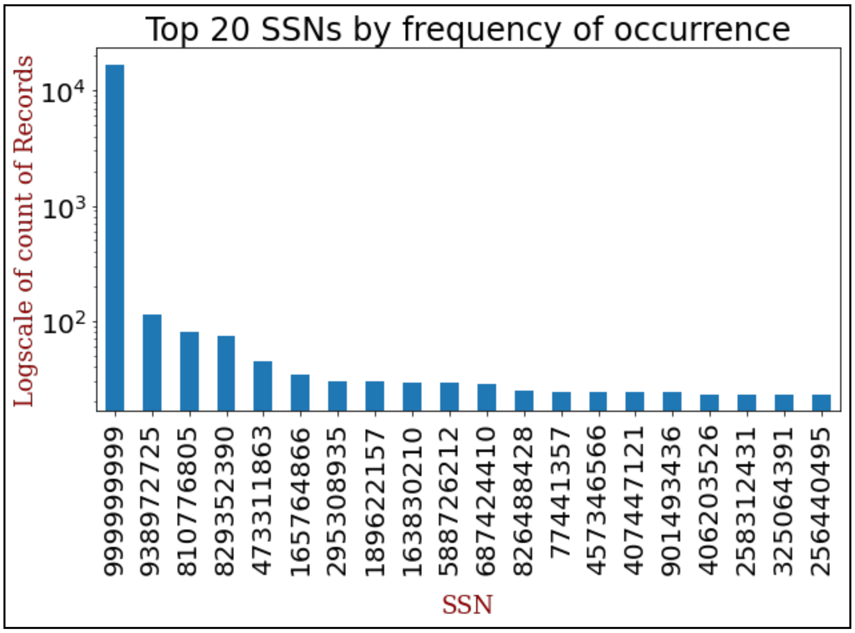


2. date (Numeric)
   - Date is the date of filing of the application
   - Month of the date of application is plotted
   - The table on the right shows the top 10 dates by frequency of occurrence

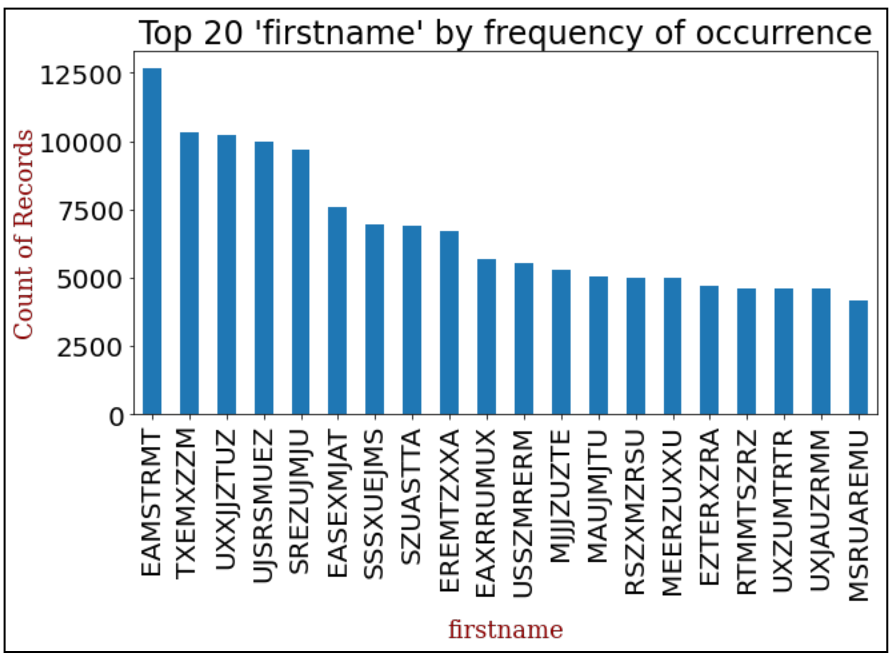


Date	# Records
2016-08-16	2877
2016-03-04	2861
2016-07-18	2849
2016-04-17	2848
2016-01-01	2840
2016-09-03	2832
2016-08-08	2832
2016-12-28	2832
2016-08-27	2831
2016-10-06	2831

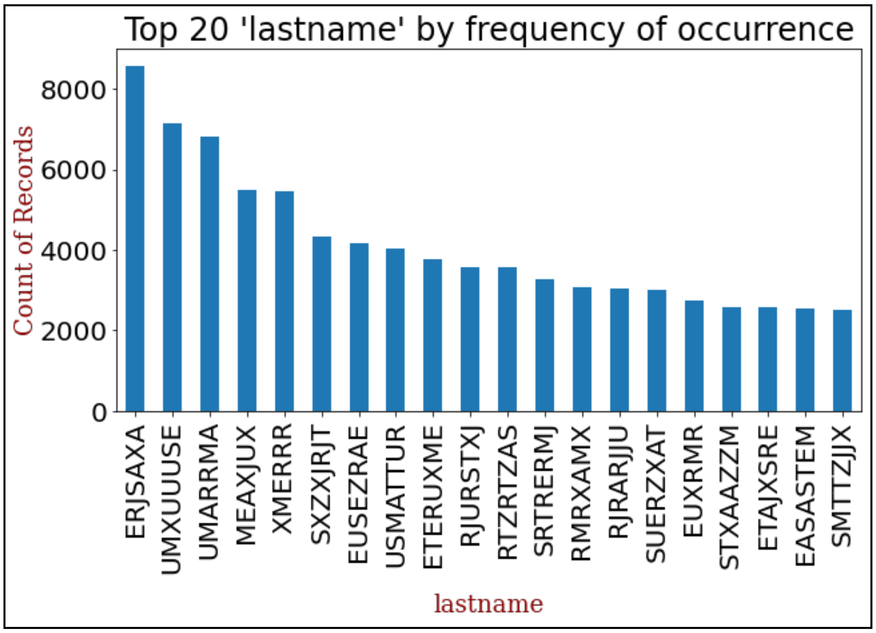
3. ssn (Categorical)
   - SSN is the 9-digit social security number associated with the applicant of each application
   - The below chart displays the frequency of top 20 SSN occurrences



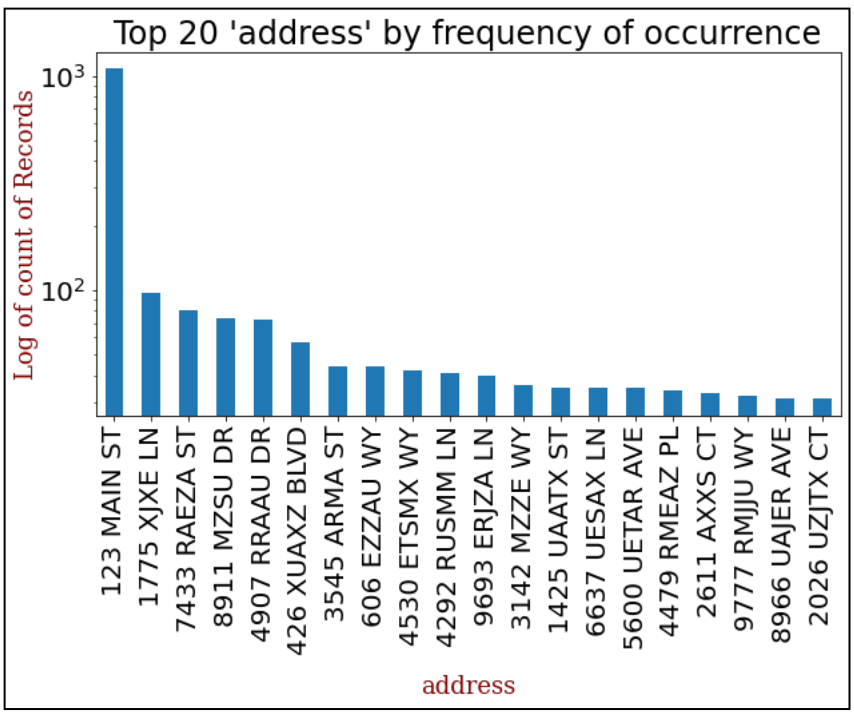
4. firstname (Categorical)
    - First name variable refers to the first name of the applicant
    - The below chart displays the frequency of top 20 first name occurrences



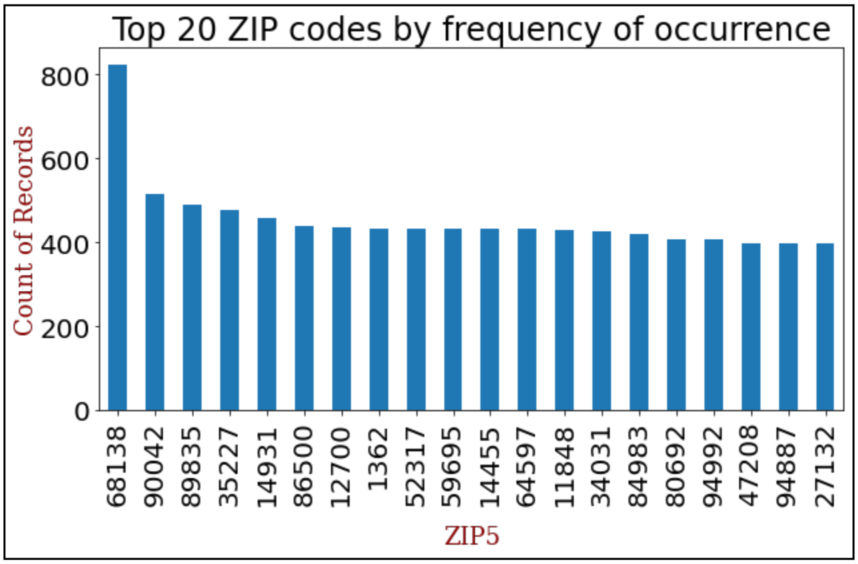
5. lastname (Categorical)
    - Last name variable refers to the last name of the applicant
    - The below chart displays the frequency of top 20 last name occurrences



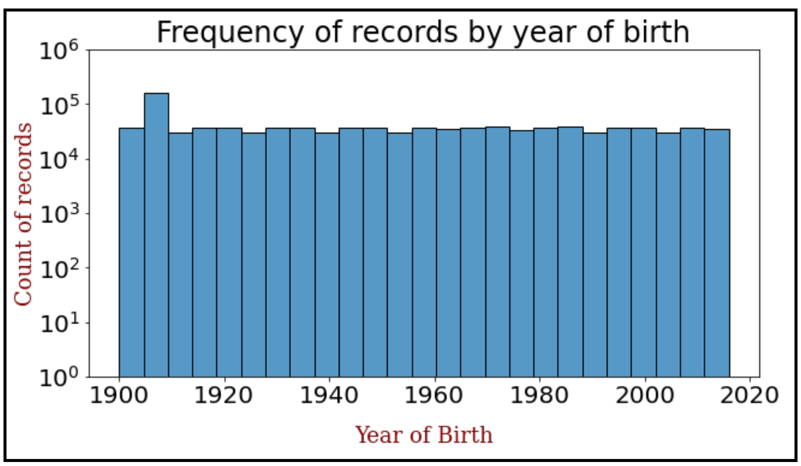
6. address (Categorical)
    - Address variable refers to the current address of the applicant while filling the application
    - The below chart displays the frequency of top 20 addresses



7. zip5 (Categorical)
    - This variable refers to the 5-digit zip code of the address associated with the application
    - The below chart displays the frequency of top 20 zip codes

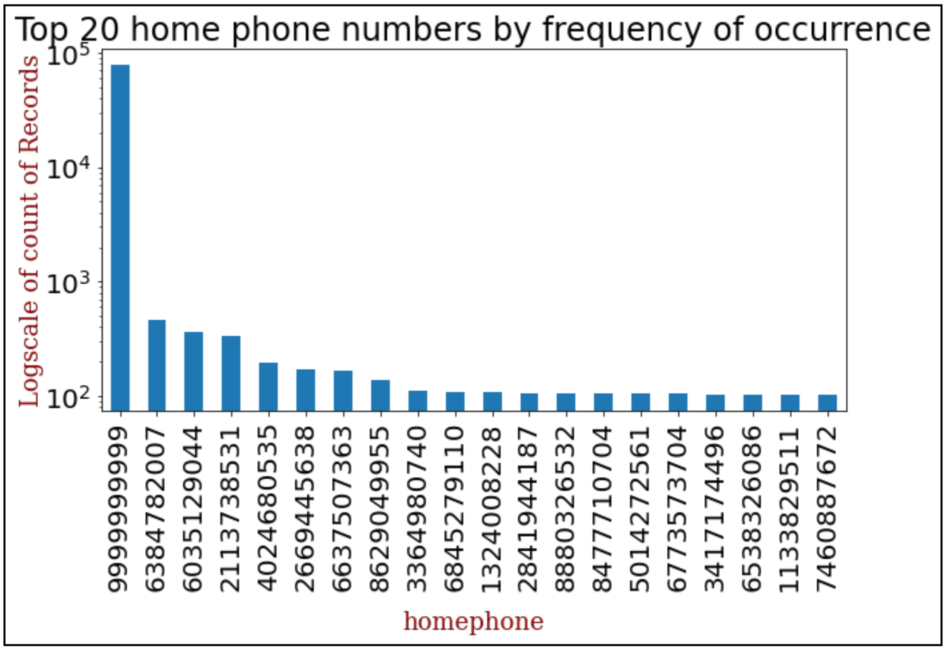


8. dob (Numeric)
   - The dob variable is the date of filing of the application
   - Year of the date of birth is plotted
   - The table on the right shows the top 10 dates of birth by frequency of occurrence



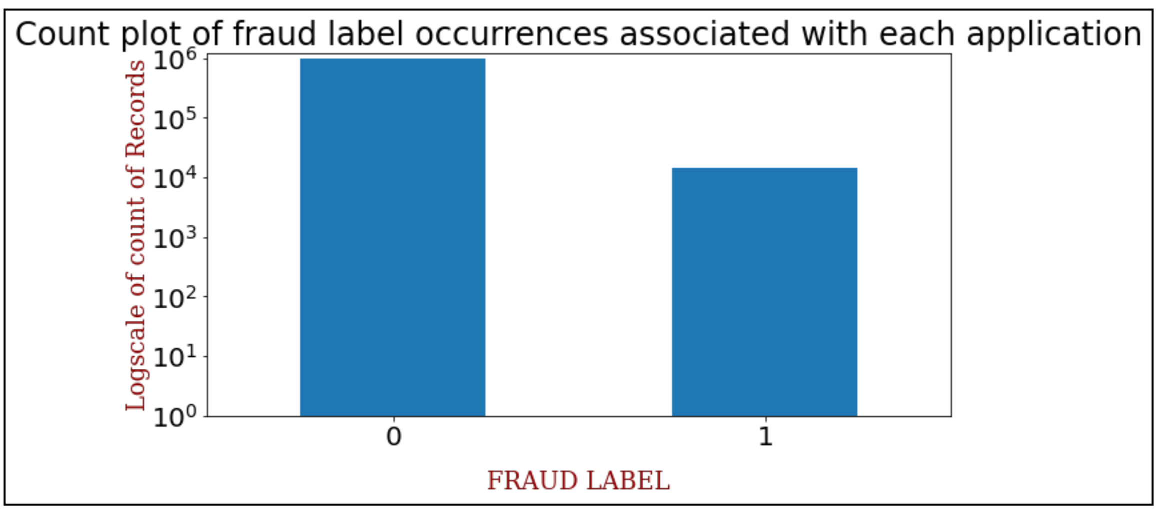
Date	# Records
1907-06-26	126568
1964-03-18	4818
1976-06-25	3723
1988-06-28	1404
1974-02-16	980
2009-01-27	280
1946-09-01	135
1959-12-08	126
1928-06-11	120
1967-02-15	102

9. homephone (Categorical)
   - The home phone variable refers to the 10-digit phone number of the applicant
   - The below chart displays the frequency of occurrence of top 20 phone numbers



10. fraud_label (Categorical)

- The fraud label refers to the fraud tag associated with the application
- The below graph displays to the counts of each fraud label



9.2 Variables List

1	ssn	51	homephone_count_3
2	firstname	52	homephone_count_7
3	lastname	53	homephone_count_14
4	address	54	homephone_count_30
5	zip5	55	name_day_since
6	dob	56	name_count_0
7	homephone	57	name_count_1
8	dow_risk	58	name_count_3
9	name	59	name_count_7
10	fulladdress	60	name_count_14
11	name_dob	61	name_count_30
12	name_fulladdress	62	fulladdress_day_since
13	name_homephone	63	fulladdress_count_0
14	fulladdress_dob	64	fulladdress_count_1
15	fulladdress_homephone	65	fulladdress_count_3
16	dob_homephone	66	fulladdress_count_7
17	homephone_name_dob	67	fulladdress_count_14
18	ssn_firstname	68	fulladdress_count_30
19	ssn_lastname	69	name_dob_day_since
20	ssn_address	70	name_dob_count_0
21	ssn_zip5	71	name_dob_count_1
22	ssn_dob	72	name_dob_count_3
23	ssn_homephone	73	name_dob_count_7
24	ssn_name	74	name_dob_count_14
25	ssn_fulladdress	75	name_dob_count_30
26	ssn_name_dob	76	name_fulladdress_day_since
27	ssn_day_since	77	name_fulladdress_count_0
28	ssn_count_0	78	name_fulladdress_count_1
29	ssn_count_1	79	name_fulladdress_count_3
30	ssn_count_3	80	name_fulladdress_count_7
31	ssn_count_7	81	name_fulladdress_count_14
32	ssn_count_14	82	name_fulladdress_count_30
33	ssn_count_30	83	name_homephone_day_since
34	address_day_since	84	name_homephone_count_0
35	address_count_0	85	name_homephone_count_1
36	address_count_1	86	name_homephone_count_3
37	address_count_3	87	name_homephone_count_7
38	address_count_7	88	name_homephone_count_14
39	address_count_14	89	name_homephone_count_30
40	address_count_30	90	fulladdress_dob_day_since
41	dob_day_since	91	fulladdress_dob_count_0
42	dob_count_0	92	fulladdress_dob_count_1
43	dob_count_1	93	fulladdress_dob_count_3
44	dob_count_3	94	fulladdress_dob_count_7
45	dob_count_7	95	fulladdress_dob_count_14
46	dob_count_14	96	fulladdress_dob_count_30
47	dob_count_30	97	fulladdress_homephone_day_since
48	homephone_day_since	98	fulladdress_homephone_count_0
49	homephone_count_0	99	fulladdress_homephone_count_1
50	homephone_count_1	100	fulladdress_homephone_count_3

101	fulladdress_homephone_count_7	151	ssn_dob_count_14
102	fulladdress_homephone_count_14	152	ssn_dob_count_30
103	fulladdress_homephone_count_30	153	ssn_homephone_day_since
104	dob_homephone_day_since	154	ssn_homephone_count_0
105	dob_homephone_count_0	155	ssn_homephone_count_1
106	dob_homephone_count_1	156	ssn_homephone_count_3
107	dob_homephone_count_3	157	ssn_homephone_count_7
108	dob_homephone_count_7	158	ssn_homephone_count_14
109	dob_homephone_count_14	159	ssn_homephone_count_30
110	dob_homephone_count_30	160	ssn_name_day_since
111	homephone_name_dob_day_since	161	ssn_name_count_0
112	homephone_name_dob_count_0	162	ssn_name_count_1
113	homephone_name_dob_count_1	163	ssn_name_count_3
114	homephone_name_dob_count_3	164	ssn_name_count_7
115	homephone_name_dob_count_7	165	ssn_name_count_14
116	homephone_name_dob_count_14	166	ssn_name_count_30
117	homephone_name_dob_count_30	167	ssn_fulladdress_day_since
118	ssn_firstname_day_since	168	ssn_fulladdress_count_0
119	ssn_firstname_count_0	169	ssn_fulladdress_count_1
120	ssn_firstname_count_1	170	ssn_fulladdress_count_3
121	ssn_firstname_count_3	171	ssn_fulladdress_count_7
122	ssn_firstname_count_7	172	ssn_fulladdress_count_14
123	ssn_firstname_count_14	173	ssn_fulladdress_count_30
124	ssn_firstname_count_30	174	ssn_name_dob_day_since
125	ssn_lastname_day_since	175	ssn_name_dob_count_0
126	ssn_lastname_count_0	176	ssn_name_dob_count_1
127	ssn_lastname_count_1	177	ssn_name_dob_count_3
128	ssn_lastname_count_3	178	ssn_name_dob_count_7
129	ssn_lastname_count_7	179	ssn_name_dob_count_14
130	ssn_lastname_count_14	180	ssn_name_dob_count_30
131	ssn_lastname_count_30	181	ssn_count_0_by_3
132	ssn_address_day_since	182	ssn_count_0_by_7
133	ssn_address_count_0	183	ssn_count_0_by_14
134	ssn_address_count_1	184	ssn_count_0_by_30
135	ssn_address_count_3	185	ssn_count_1_by_3
136	ssn_address_count_7	186	ssn_count_1_by_7
137	ssn_address_count_14	187	ssn_count_1_by_14
138	ssn_address_count_30	188	ssn_count_1_by_30
139	ssn_zip5_day_since	189	address_count_0_by_3
140	ssn_zip5_count_0	190	address_count_0_by_7
141	ssn_zip5_count_1	191	address_count_0_by_14
142	ssn_zip5_count_3	192	address_count_0_by_30
143	ssn_zip5_count_7	193	address_count_1_by_3
144	ssn_zip5_count_14	194	address_count_1_by_7
145	ssn_zip5_count_30	195	address_count_1_by_14
146	ssn_dob_day_since	196	address_count_1_by_30
147	ssn_dob_count_0	197	dob_count_0_by_3
148	ssn_dob_count_1	198	dob_count_0_by_7
149	ssn_dob_count_3	199	dob_count_0_by_14
150	ssn_dob_count_7	200	dob_count_0_by_30

201	dob_count_1_by_3	251	name_homephone_count_1_by_14
202	dob_count_1_by_7	252	name_homephone_count_1_by_30
203	dob_count_1_by_14	253	fulladdress_dob_count_0_by_3
204	dob_count_1_by_30	254	fulladdress_dob_count_0_by_7
205	homephone_count_0_by_3	255	fulladdress_dob_count_0_by_14
206	homephone_count_0_by_7	256	fulladdress_dob_count_0_by_30
207	homephone_count_0_by_14	257	fulladdress_dob_count_1_by_3
208	homephone_count_0_by_30	258	fulladdress_dob_count_1_by_7
209	homephone_count_1_by_3	259	fulladdress_dob_count_1_by_14
210	homephone_count_1_by_7	260	fulladdress_dob_count_1_by_30
211	homephone_count_1_by_14	261	fulladdress_homephone_count_0_by_3
212	homephone_count_1_by_30	262	fulladdress_homephone_count_0_by_7
213	name_count_0_by_3	263	fulladdress_homephone_count_0_by_14
214	name_count_0_by_7	264	fulladdress_homephone_count_0_by_30
215	name_count_0_by_14	265	fulladdress_homephone_count_1_by_3
216	name_count_0_by_30	266	fulladdress_homephone_count_1_by_7
217	name_count_1_by_3	267	fulladdress_homephone_count_1_by_14
218	name_count_1_by_7	268	fulladdress_homephone_count_1_by_30
219	name_count_1_by_14	269	dob_homephone_count_0_by_3
220	name_count_1_by_30	270	dob_homephone_count_0_by_7
221	fulladdress_count_0_by_3	271	dob_homephone_count_0_by_14
222	fulladdress_count_0_by_7	272	dob_homephone_count_0_by_30
223	fulladdress_count_0_by_14	273	dob_homephone_count_1_by_3
224	fulladdress_count_0_by_30	274	dob_homephone_count_1_by_7
225	fulladdress_count_1_by_3	275	dob_homephone_count_1_by_14
226	fulladdress_count_1_by_7	276	dob_homephone_count_1_by_30
227	fulladdress_count_1_by_14	277	homephone_name_dob_count_0_by_3
228	fulladdress_count_1_by_30	278	homephone_name_dob_count_0_by_7
229	name_dob_count_0_by_3	279	homephone_name_dob_count_0_by_14
230	name_dob_count_0_by_7	280	homephone_name_dob_count_0_by_30
231	name_dob_count_0_by_14	281	homephone_name_dob_count_1_by_3
232	name_dob_count_0_by_30	282	homephone_name_dob_count_1_by_7
233	name_dob_count_1_by_3	283	homephone_name_dob_count_1_by_14
234	name_dob_count_1_by_7	284	homephone_name_dob_count_1_by_30
235	name_dob_count_1_by_14	285	ssn_firstname_count_0_by_3
236	name_dob_count_1_by_30	286	ssn_firstname_count_0_by_7
237	name_fulladdress_count_0_by_3	287	ssn_firstname_count_0_by_14
238	name_fulladdress_count_0_by_7	288	ssn_firstname_count_0_by_30
239	name_fulladdress_count_0_by_14	289	ssn_firstname_count_1_by_3
240	name_fulladdress_count_0_by_30	290	ssn_firstname_count_1_by_7
241	name_fulladdress_count_1_by_3	291	ssn_firstname_count_1_by_14
242	name_fulladdress_count_1_by_7	292	ssn_firstname_count_1_by_30
243	name_fulladdress_count_1_by_14	293	ssn_lastname_count_0_by_3
244	name_fulladdress_count_1_by_30	294	ssn_lastname_count_0_by_7
245	name_homephone_count_0_by_3	295	ssn_lastname_count_0_by_14
246	name_homephone_count_0_by_7	296	ssn_lastname_count_0_by_30
247	name_homephone_count_0_by_14	297	ssn_lastname_count_1_by_3
248	name_homephone_count_0_by_30	298	ssn_lastname_count_1_by_7
249	name_homephone_count_1_by_3	299	ssn_lastname_count_1_by_14
250	name_homephone_count_1_by_7	300	ssn_lastname_count_1_by_30

301	ssn_address_count_0_by_3	351	ssn_name_dob_count_0_by_14
302	ssn_address_count_0_by_7	352	ssn_name_dob_count_0_by_30
303	ssn_address_count_0_by_14	353	ssn_name_dob_count_1_by_3
304	ssn_address_count_0_by_30	354	ssn_name_dob_count_1_by_7
305	ssn_address_count_1_by_3	355	ssn_name_dob_count_1_by_14
306	ssn_address_count_1_by_7	356	ssn_name_dob_count_1_by_30
307	ssn_address_count_1_by_14	357	ssn_unique_count_for_fulladdress_1
308	ssn_address_count_1_by_30	358	ssn_unique_count_for_fulladdress_3
309	ssn_zip5_count_0_by_3	359	ssn_unique_count_for_fulladdress_7
310	ssn_zip5_count_0_by_7	360	ssn_unique_count_for_fulladdress_14
311	ssn_zip5_count_0_by_14	361	ssn_unique_count_for_fulladdress_30
312	ssn_zip5_count_0_by_30	362	ssn_unique_count_for_fulladdress_60
313	ssn_zip5_count_1_by_3	363	ssn_unique_count_for_name_dob_1
314	ssn_zip5_count_1_by_7	364	ssn_unique_count_for_name_dob_3
315	ssn_zip5_count_1_by_14	365	ssn_unique_count_for_name_dob_7
316	ssn_zip5_count_1_by_30	366	ssn_unique_count_for_name_dob_14
317	ssn_dob_count_0_by_3	367	ssn_unique_count_for_name_dob_30
318	ssn_dob_count_0_by_7	368	ssn_unique_count_for_name_dob_60
319	ssn_dob_count_0_by_14	369	ssn_unique_count_for_name_fulladdress_1
320	ssn_dob_count_0_by_30	370	ssn_unique_count_for_name_fulladdress_3
321	ssn_dob_count_1_by_3	371	ssn_unique_count_for_name_fulladdress_7
322	ssn_dob_count_1_by_7	372	ssn_unique_count_for_name_fulladdress_14
323	ssn_dob_count_1_by_14	373	ssn_unique_count_for_name_fulladdress_30
324	ssn_dob_count_1_by_30	374	ssn_unique_count_for_name_fulladdress_60
325	ssn_homephone_count_0_by_3	375	ssn_unique_count_for_fulladdress_dob_1
326	ssn_homephone_count_0_by_7	376	ssn_unique_count_for_fulladdress_dob_3
327	ssn_homephone_count_0_by_14	377	ssn_unique_count_for_fulladdress_dob_7
328	ssn_homephone_count_0_by_30	378	ssn_unique_count_for_fulladdress_dob_14
329	ssn_homephone_count_1_by_3	379	ssn_unique_count_for_fulladdress_dob_30
330	ssn_homephone_count_1_by_7	380	ssn_unique_count_for_fulladdress_dob_60
331	ssn_homephone_count_1_by_14	381	ssn_unique_count_for_dob_homephone_1
332	ssn_homephone_count_1_by_30	382	ssn_unique_count_for_dob_homephone_3
333	ssn_name_count_0_by_3	383	ssn_unique_count_for_dob_homephone_7
334	ssn_name_count_0_by_7	384	ssn_unique_count_for_dob_homephone_14
335	ssn_name_count_0_by_14	385	ssn_unique_count_for_dob_homephone_30
336	ssn_name_count_0_by_30	386	ssn_unique_count_for_dob_homephone_60
337	ssn_name_count_1_by_3	387	ssn_unique_count_for_ssn_lastname_1
338	ssn_name_count_1_by_7	388	ssn_unique_count_for_ssn_lastname_3
339	ssn_name_count_1_by_14	389	ssn_unique_count_for_ssn_lastname_7
340	ssn_name_count_1_by_30	390	ssn_unique_count_for_ssn_lastname_14
341	ssn_fulladdress_count_0_by_3	391	ssn_unique_count_for_ssn_lastname_30
342	ssn_fulladdress_count_0_by_7	392	ssn_unique_count_for_ssn_lastname_60
343	ssn_fulladdress_count_0_by_14	393	ssn_unique_count_for_ssn_zip5_1
344	ssn_fulladdress_count_0_by_30	394	ssn_unique_count_for_ssn_zip5_3
345	ssn_fulladdress_count_1_by_3	395	ssn_unique_count_for_ssn_zip5_7
346	ssn_fulladdress_count_1_by_7	396	ssn_unique_count_for_ssn_zip5_14
347	ssn_fulladdress_count_1_by_14	397	ssn_unique_count_for_ssn_zip5_30
348	ssn_fulladdress_count_1_by_30	398	ssn_unique_count_for_ssn_zip5_60
349	ssn_name_dob_count_0_by_3	399	ssn_unique_count_for_ssn_name_1
350	ssn_name_dob_count_0_by_7	400	ssn_unique_count_for_ssn_name_3

401	ssn_unique_count_for_ssn_name_7	451	fulladdress_unique_count_for_ssn_lastname_30
402	ssn_unique_count_for_ssn_name_14	452	fulladdress_unique_count_for_ssn_lastname_60
403	ssn_unique_count_for_ssn_name_30	453	fulladdress_unique_count_for_ssn_zip5_1
404	ssn_unique_count_for_ssn_name_60	454	fulladdress_unique_count_for_ssn_zip5_3
405	ssn_unique_count_for_ssn_fulladdress_1	455	fulladdress_unique_count_for_ssn_zip5_7
406	ssn_unique_count_for_ssn_fulladdress_3	456	fulladdress_unique_count_for_ssn_zip5_14
407	ssn_unique_count_for_ssn_fulladdress_7	457	fulladdress_unique_count_for_ssn_zip5_30
408	ssn_unique_count_for_ssn_fulladdress_14	458	fulladdress_unique_count_for_ssn_zip5_60
409	ssn_unique_count_for_ssn_fulladdress_30	459	fulladdress_unique_count_for_ssn_name_1
410	ssn_unique_count_for_ssn_fulladdress_60	460	fulladdress_unique_count_for_ssn_name_3
411	ssn_unique_count_for_ssn_name_dob_1	461	fulladdress_unique_count_for_ssn_name_7
412	ssn_unique_count_for_ssn_name_dob_3	462	fulladdress_unique_count_for_ssn_name_14
413	ssn_unique_count_for_ssn_name_dob_7	463	fulladdress_unique_count_for_ssn_name_30
414	ssn_unique_count_for_ssn_name_dob_14	464	fulladdress_unique_count_for_ssn_name_60
415	ssn_unique_count_for_ssn_name_dob_30	465	fulladdress_unique_count_for_ssn_fulladdress_1
416	ssn_unique_count_for_ssn_name_dob_60	466	fulladdress_unique_count_for_ssn_fulladdress_3
417	fulladdress_unique_count_for_ssn_1	467	fulladdress_unique_count_for_ssn_fulladdress_7
418	fulladdress_unique_count_for_ssn_3	468	fulladdress_unique_count_for_ssn_fulladdress_14
419	fulladdress_unique_count_for_ssn_7	469	fulladdress_unique_count_for_ssn_fulladdress_30
420	fulladdress_unique_count_for_ssn_14	470	fulladdress_unique_count_for_ssn_fulladdress_60
421	fulladdress_unique_count_for_ssn_30	471	fulladdress_unique_count_for_ssn_name_dob_1
422	fulladdress_unique_count_for_ssn_60	472	fulladdress_unique_count_for_ssn_name_dob_3
423	fulladdress_unique_count_for_name_dob_1	473	fulladdress_unique_count_for_ssn_name_dob_7
424	fulladdress_unique_count_for_name_dob_3	474	fulladdress_unique_count_for_ssn_name_dob_14
425	fulladdress_unique_count_for_name_dob_7	475	fulladdress_unique_count_for_ssn_name_dob_30
426	fulladdress_unique_count_for_name_dob_14	476	fulladdress_unique_count_for_ssn_name_dob_60
427	fulladdress_unique_count_for_name_dob_30	477	name_dob_unique_count_for_ssn_1
428	fulladdress_unique_count_for_name_dob_60	478	name_dob_unique_count_for_ssn_3
429	fulladdress_unique_count_for_name_fulladdress_1	479	name_dob_unique_count_for_ssn_7
430	fulladdress_unique_count_for_name_fulladdress_3	480	name_dob_unique_count_for_ssn_14
431	fulladdress_unique_count_for_name_fulladdress_7	481	name_dob_unique_count_for_ssn_30
432	fulladdress_unique_count_for_name_fulladdress_14	482	name_dob_unique_count_for_ssn_60
433	fulladdress_unique_count_for_name_fulladdress_30	483	name_dob_unique_count_for_fulladdress_1
434	fulladdress_unique_count_for_name_fulladdress_60	484	name_dob_unique_count_for_fulladdress_3
435	fulladdress_unique_count_for_fulladdress_dob_1	485	name_dob_unique_count_for_fulladdress_7
436	fulladdress_unique_count_for_fulladdress_dob_3	486	name_dob_unique_count_for_fulladdress_14
437	fulladdress_unique_count_for_fulladdress_dob_7	487	name_dob_unique_count_for_fulladdress_30
438	fulladdress_unique_count_for_fulladdress_dob_14	488	name_dob_unique_count_for_fulladdress_60
439	fulladdress_unique_count_for_fulladdress_dob_30	489	name_dob_unique_count_for_name_fulladdress_1
440	fulladdress_unique_count_for_fulladdress_dob_60	490	name_dob_unique_count_for_name_fulladdress_3
441	fulladdress_unique_count_for_dob_homephone_1	491	name_dob_unique_count_for_name_fulladdress_7
442	fulladdress_unique_count_for_dob_homephone_3	492	name_dob_unique_count_for_name_fulladdress_14
443	fulladdress_unique_count_for_dob_homephone_7	493	name_dob_unique_count_for_name_fulladdress_30
444	fulladdress_unique_count_for_dob_homephone_14	494	name_dob_unique_count_for_name_fulladdress_60
445	fulladdress_unique_count_for_dob_homephone_30	495	name_dob_unique_count_for_fulladdress_dob_1
446	fulladdress_unique_count_for_dob_homephone_60	496	name_dob_unique_count_for_fulladdress_dob_3
447	fulladdress_unique_count_for_ssn_lastname_1	497	name_dob_unique_count_for_fulladdress_dob_7
448	fulladdress_unique_count_for_ssn_lastname_3	498	name_dob_unique_count_for_fulladdress_dob_14
449	fulladdress_unique_count_for_ssn_lastname_7	499	name_dob_unique_count_for_fulladdress_dob_30
450	fulladdress_unique_count_for_ssn_lastname_14	500	name_dob_unique_count_for_fulladdress_dob_60

501	name_dob_unique_count_for_dob_homephone_1	551	name_fulladdress_unique_count_for_name_dob_7
502	name_dob_unique_count_for_dob_homephone_3	552	name_fulladdress_unique_count_for_name_dob_14
503	name_dob_unique_count_for_dob_homephone_7	553	name_fulladdress_unique_count_for_name_dob_30
504	name_dob_unique_count_for_dob_homephone_14	554	name_fulladdress_unique_count_for_name_dob_60
505	name_dob_unique_count_for_dob_homephone_30	555	name_fulladdress_unique_count_for_fulladdress_dob_1
506	name_dob_unique_count_for_dob_homephone_60	556	name_fulladdress_unique_count_for_fulladdress_dob_3
507	name_dob_unique_count_for_ssn_lastname_1	557	name_fulladdress_unique_count_for_fulladdress_dob_7
508	name_dob_unique_count_for_ssn_lastname_3	558	name_fulladdress_unique_count_for_fulladdress_dob_14
509	name_dob_unique_count_for_ssn_lastname_7	559	name_fulladdress_unique_count_for_fulladdress_dob_30
510	name_dob_unique_count_for_ssn_lastname_14	560	name_fulladdress_unique_count_for_fulladdress_dob_60
511	name_dob_unique_count_for_ssn_lastname_30	561	name_fulladdress_unique_count_for_dob_homephone_1
512	name_dob_unique_count_for_ssn_lastname_60	562	name_fulladdress_unique_count_for_dob_homephone_3
513	name_dob_unique_count_for_ssn_zip5_1	563	name_fulladdress_unique_count_for_dob_homephone_7
514	name_dob_unique_count_for_ssn_zip5_3	564	name_fulladdress_unique_count_for_dob_homephone_14
515	name_dob_unique_count_for_ssn_zip5_7	565	name_fulladdress_unique_count_for_dob_homephone_30
516	name_dob_unique_count_for_ssn_zip5_14	566	name_fulladdress_unique_count_for_dob_homephone_60
517	name_dob_unique_count_for_ssn_zip5_30	567	name_fulladdress_unique_count_for_ssn_lastname_1
518	name_dob_unique_count_for_ssn_zip5_60	568	name_fulladdress_unique_count_for_ssn_lastname_3
519	name_dob_unique_count_for_ssn_name_1	569	name_fulladdress_unique_count_for_ssn_lastname_7
520	name_dob_unique_count_for_ssn_name_3	570	name_fulladdress_unique_count_for_ssn_lastname_14
521	name_dob_unique_count_for_ssn_name_7	571	name_fulladdress_unique_count_for_ssn_lastname_30
522	name_dob_unique_count_for_ssn_name_14	572	name_fulladdress_unique_count_for_ssn_lastname_60
523	name_dob_unique_count_for_ssn_name_30	573	name_fulladdress_unique_count_for_ssn_zip5_1
524	name_dob_unique_count_for_ssn_name_60	574	name_fulladdress_unique_count_for_ssn_zip5_3
525	name_dob_unique_count_for_ssn_fulladdress_1	575	name_fulladdress_unique_count_for_ssn_zip5_7
526	name_dob_unique_count_for_ssn_fulladdress_3	576	name_fulladdress_unique_count_for_ssn_zip5_14
527	name_dob_unique_count_for_ssn_fulladdress_7	577	name_fulladdress_unique_count_for_ssn_zip5_30
528	name_dob_unique_count_for_ssn_fulladdress_14	578	name_fulladdress_unique_count_for_ssn_zip5_60
529	name_dob_unique_count_for_ssn_fulladdress_30	579	name_fulladdress_unique_count_for_ssn_name_1
530	name_dob_unique_count_for_ssn_fulladdress_60	580	name_fulladdress_unique_count_for_ssn_name_3
531	name_dob_unique_count_for_ssn_name_dob_1	581	name_fulladdress_unique_count_for_ssn_name_7
532	name_dob_unique_count_for_ssn_name_dob_3	582	name_fulladdress_unique_count_for_ssn_name_14
533	name_dob_unique_count_for_ssn_name_dob_7	583	name_fulladdress_unique_count_for_ssn_name_30
534	name_dob_unique_count_for_ssn_name_dob_14	584	name_fulladdress_unique_count_for_ssn_name_60
535	name_dob_unique_count_for_ssn_name_dob_30	585	name_fulladdress_unique_count_for_ssn_fulladdress_1
536	name_dob_unique_count_for_ssn_name_dob_60	586	name_fulladdress_unique_count_for_ssn_fulladdress_3
537	name_fulladdress_unique_count_for_ssn_1	587	name_fulladdress_unique_count_for_ssn_fulladdress_7
538	name_fulladdress_unique_count_for_ssn_3	588	name_fulladdress_unique_count_for_ssn_fulladdress_14
539	name_fulladdress_unique_count_for_ssn_7	589	name_fulladdress_unique_count_for_ssn_fulladdress_30
540	name_fulladdress_unique_count_for_ssn_14	590	name_fulladdress_unique_count_for_ssn_fulladdress_60
541	name_fulladdress_unique_count_for_ssn_30	591	name_fulladdress_unique_count_for_ssn_name_dob_1
542	name_fulladdress_unique_count_for_ssn_60	592	name_fulladdress_unique_count_for_ssn_name_dob_3
543	name_fulladdress_unique_count_for_fulladdress_1	593	name_fulladdress_unique_count_for_ssn_name_dob_7
544	name_fulladdress_unique_count_for_fulladdress_3	594	name_fulladdress_unique_count_for_ssn_name_dob_14
545	name_fulladdress_unique_count_for_fulladdress_7	595	name_fulladdress_unique_count_for_ssn_name_dob_30
546	name_fulladdress_unique_count_for_fulladdress_14	596	name_fulladdress_unique_count_for_ssn_name_dob_60
547	name_fulladdress_unique_count_for_fulladdress_30	597	fulladdress_dob_unique_count_for_ssn_1
548	name_fulladdress_unique_count_for_fulladdress_60	598	fulladdress_dob_unique_count_for_ssn_3
549	name_fulladdress_unique_count_for_name_dob_1	599	fulladdress_dob_unique_count_for_ssn_7
550	name_fulladdress_unique_count_for_name_dob_3	600	fulladdress_dob_unique_count_for_ssn_14

601	fulladdress_dob_unique_count_for_ssn_30	651	fulladdress_dob_unique_count_for_ssn_name_dob_1
602	fulladdress_dob_unique_count_for_ssn_60	652	fulladdress_dob_unique_count_for_ssn_name_dob_3
603	fulladdress_dob_unique_count_for_fulladdress_1	653	fulladdress_dob_unique_count_for_ssn_name_dob_7
604	fulladdress_dob_unique_count_for_fulladdress_3	654	fulladdress_dob_unique_count_for_ssn_name_dob_14
605	fulladdress_dob_unique_count_for_fulladdress_7	655	fulladdress_dob_unique_count_for_ssn_name_dob_30
606	fulladdress_dob_unique_count_for_fulladdress_14	656	fulladdress_dob_unique_count_for_ssn_name_dob_60
607	fulladdress_dob_unique_count_for_fulladdress_30	657	dob_homephone_unique_count_for_ssn_1
608	fulladdress_dob_unique_count_for_fulladdress_60	658	dob_homephone_unique_count_for_ssn_3
609	fulladdress_dob_unique_count_for_name_dob_1	659	dob_homephone_unique_count_for_ssn_7
610	fulladdress_dob_unique_count_for_name_dob_3	660	dob_homephone_unique_count_for_ssn_14
611	fulladdress_dob_unique_count_for_name_dob_7	661	dob_homephone_unique_count_for_ssn_30
612	fulladdress_dob_unique_count_for_name_dob_14	662	dob_homephone_unique_count_for_ssn_60
613	fulladdress_dob_unique_count_for_name_dob_30	663	dob_homephone_unique_count_for_fulladdress_1
614	fulladdress_dob_unique_count_for_name_dob_60	664	dob_homephone_unique_count_for_fulladdress_3
615	fulladdress_dob_unique_count_for_name_fulladdress_1	665	dob_homephone_unique_count_for_fulladdress_7
616	fulladdress_dob_unique_count_for_name_fulladdress_3	666	dob_homephone_unique_count_for_fulladdress_14
617	fulladdress_dob_unique_count_for_name_fulladdress_7	667	dob_homephone_unique_count_for_fulladdress_30
618	fulladdress_dob_unique_count_for_name_fulladdress_14	668	dob_homephone_unique_count_for_fulladdress_60
619	fulladdress_dob_unique_count_for_name_fulladdress_30	669	dob_homephone_unique_count_for_name_dob_1
620	fulladdress_dob_unique_count_for_name_fulladdress_60	670	dob_homephone_unique_count_for_name_dob_3
621	fulladdress_dob_unique_count_for_dob_homephone_1	671	dob_homephone_unique_count_for_name_dob_7
622	fulladdress_dob_unique_count_for_dob_homephone_3	672	dob_homephone_unique_count_for_name_dob_14
623	fulladdress_dob_unique_count_for_dob_homephone_7	673	dob_homephone_unique_count_for_name_dob_30
624	fulladdress_dob_unique_count_for_dob_homephone_14	674	dob_homephone_unique_count_for_name_dob_60
625	fulladdress_dob_unique_count_for_dob_homephone_30	675	dob_homephone_unique_count_for_name_fulladdress_1
626	fulladdress_dob_unique_count_for_dob_homephone_60	676	dob_homephone_unique_count_for_name_fulladdress_3
627	fulladdress_dob_unique_count_for_ssn_lastname_1	677	dob_homephone_unique_count_for_name_fulladdress_7
628	fulladdress_dob_unique_count_for_ssn_lastname_3	678	dob_homephone_unique_count_for_name_fulladdress_14
629	fulladdress_dob_unique_count_for_ssn_lastname_7	679	dob_homephone_unique_count_for_name_fulladdress_30
630	fulladdress_dob_unique_count_for_ssn_lastname_14	680	dob_homephone_unique_count_for_name_fulladdress_60
631	fulladdress_dob_unique_count_for_ssn_lastname_30	681	dob_homephone_unique_count_for_fulladdress_dob_1
632	fulladdress_dob_unique_count_for_ssn_lastname_60	682	dob_homephone_unique_count_for_fulladdress_dob_3
633	fulladdress_dob_unique_count_for_ssn_zip5_1	683	dob_homephone_unique_count_for_fulladdress_dob_7
634	fulladdress_dob_unique_count_for_ssn_zip5_3	684	dob_homephone_unique_count_for_fulladdress_dob_14
635	fulladdress_dob_unique_count_for_ssn_zip5_7	685	dob_homephone_unique_count_for_fulladdress_dob_30
636	fulladdress_dob_unique_count_for_ssn_zip5_14	686	dob_homephone_unique_count_for_fulladdress_dob_60
637	fulladdress_dob_unique_count_for_ssn_zip5_30	687	dob_homephone_unique_count_for_ssn_lastname_1
638	fulladdress_dob_unique_count_for_ssn_zip5_60	688	dob_homephone_unique_count_for_ssn_lastname_3
639	fulladdress_dob_unique_count_for_ssn_name_1	689	dob_homephone_unique_count_for_ssn_lastname_7
640	fulladdress_dob_unique_count_for_ssn_name_3	690	dob_homephone_unique_count_for_ssn_lastname_14
641	fulladdress_dob_unique_count_for_ssn_name_7	691	dob_homephone_unique_count_for_ssn_lastname_30
642	fulladdress_dob_unique_count_for_ssn_name_14	692	dob_homephone_unique_count_for_ssn_lastname_60
643	fulladdress_dob_unique_count_for_ssn_name_30	693	dob_homephone_unique_count_for_ssn_zip5_1
644	fulladdress_dob_unique_count_for_ssn_name_60	694	dob_homephone_unique_count_for_ssn_zip5_3
645	fulladdress_dob_unique_count_for_ssn_fulladdress_1	695	dob_homephone_unique_count_for_ssn_zip5_7
646	fulladdress_dob_unique_count_for_ssn_fulladdress_3	696	dob_homephone_unique_count_for_ssn_zip5_14
647	fulladdress_dob_unique_count_for_ssn_fulladdress_7	697	dob_homephone_unique_count_for_ssn_zip5_30
648	fulladdress_dob_unique_count_for_ssn_fulladdress_14	698	dob_homephone_unique_count_for_ssn_zip5_60
649	fulladdress_dob_unique_count_for_ssn_fulladdress_30	699	dob_homephone_unique_count_for_ssn_name_1
650	fulladdress_dob_unique_count_for_ssn_fulladdress_60	700	dob_homephone_unique_count_for_ssn_name_3

701	dob_homephone_unique_count_for_ssn_name_7	751	ssn_lastname_unique_count_for_dob_homephone_30
702	dob_homephone_unique_count_for_ssn_name_14	752	ssn_lastname_unique_count_for_dob_homephone_60
703	dob_homephone_unique_count_for_ssn_name_30	753	ssn_lastname_unique_count_for_ssn_zip5_1
704	dob_homephone_unique_count_for_ssn_name_60	754	ssn_lastname_unique_count_for_ssn_zip5_3
705	dob_homephone_unique_count_for_ssn_fulladdress_1	755	ssn_lastname_unique_count_for_ssn_zip5_7
706	dob_homephone_unique_count_for_ssn_fulladdress_3	756	ssn_lastname_unique_count_for_ssn_zip5_14
707	dob_homephone_unique_count_for_ssn_fulladdress_7	757	ssn_lastname_unique_count_for_ssn_zip5_30
708	dob_homephone_unique_count_for_ssn_fulladdress_14	758	ssn_lastname_unique_count_for_ssn_zip5_60
709	dob_homephone_unique_count_for_ssn_fulladdress_30	759	ssn_lastname_unique_count_for_ssn_name_1
710	dob_homephone_unique_count_for_ssn_fulladdress_60	760	ssn_lastname_unique_count_for_ssn_name_3
711	dob_homephone_unique_count_for_ssn_name_dob_1	761	ssn_lastname_unique_count_for_ssn_name_7
712	dob_homephone_unique_count_for_ssn_name_dob_3	762	ssn_lastname_unique_count_for_ssn_name_14
713	dob_homephone_unique_count_for_ssn_name_dob_7	763	ssn_lastname_unique_count_for_ssn_name_30
714	dob_homephone_unique_count_for_ssn_name_dob_14	764	ssn_lastname_unique_count_for_ssn_name_60
715	dob_homephone_unique_count_for_ssn_name_dob_30	765	ssn_lastname_unique_count_for_ssn_fulladdress_1
716	dob_homephone_unique_count_for_ssn_name_dob_60	766	ssn_lastname_unique_count_for_ssn_fulladdress_3
717	ssn_lastname_unique_count_for_ssn_1	767	ssn_lastname_unique_count_for_ssn_fulladdress_7
718	ssn_lastname_unique_count_for_ssn_3	768	ssn_lastname_unique_count_for_ssn_fulladdress_14
719	ssn_lastname_unique_count_for_ssn_7	769	ssn_lastname_unique_count_for_ssn_fulladdress_30
720	ssn_lastname_unique_count_for_ssn_14	770	ssn_lastname_unique_count_for_ssn_fulladdress_60
721	ssn_lastname_unique_count_for_ssn_30	771	ssn_lastname_unique_count_for_ssn_name_dob_1
722	ssn_lastname_unique_count_for_ssn_60	772	ssn_lastname_unique_count_for_ssn_name_dob_3
723	ssn_lastname_unique_count_for_fulladdress_1	773	ssn_lastname_unique_count_for_ssn_name_dob_7
724	ssn_lastname_unique_count_for_fulladdress_3	774	ssn_lastname_unique_count_for_ssn_name_dob_14
725	ssn_lastname_unique_count_for_fulladdress_7	775	ssn_lastname_unique_count_for_ssn_name_dob_30
726	ssn_lastname_unique_count_for_fulladdress_14	776	ssn_lastname_unique_count_for_ssn_name_dob_60
727	ssn_lastname_unique_count_for_fulladdress_30	777	ssn_zip5_unique_count_for_ssn_1
728	ssn_lastname_unique_count_for_fulladdress_60	778	ssn_zip5_unique_count_for_ssn_3
729	ssn_lastname_unique_count_for_name_dob_1	779	ssn_zip5_unique_count_for_ssn_7
730	ssn_lastname_unique_count_for_name_dob_3	780	ssn_zip5_unique_count_for_ssn_14
731	ssn_lastname_unique_count_for_name_dob_7	781	ssn_zip5_unique_count_for_ssn_30
732	ssn_lastname_unique_count_for_name_dob_14	782	ssn_zip5_unique_count_for_ssn_60
733	ssn_lastname_unique_count_for_name_dob_30	783	ssn_zip5_unique_count_for_fulladdress_1
734	ssn_lastname_unique_count_for_name_dob_60	784	ssn_zip5_unique_count_for_fulladdress_3
735	ssn_lastname_unique_count_for_name_fulladdress_1	785	ssn_zip5_unique_count_for_fulladdress_7
736	ssn_lastname_unique_count_for_name_fulladdress_3	786	ssn_zip5_unique_count_for_fulladdress_14
737	ssn_lastname_unique_count_for_name_fulladdress_7	787	ssn_zip5_unique_count_for_fulladdress_30
738	ssn_lastname_unique_count_for_name_fulladdress_14	788	ssn_zip5_unique_count_for_fulladdress_60
739	ssn_lastname_unique_count_for_name_fulladdress_30	789	ssn_zip5_unique_count_for_name_dob_1
740	ssn_lastname_unique_count_for_name_fulladdress_60	790	ssn_zip5_unique_count_for_name_dob_3
741	ssn_lastname_unique_count_for_fulladdress_dob_1	791	ssn_zip5_unique_count_for_name_dob_7
742	ssn_lastname_unique_count_for_fulladdress_dob_3	792	ssn_zip5_unique_count_for_name_dob_14
743	ssn_lastname_unique_count_for_fulladdress_dob_7	793	ssn_zip5_unique_count_for_name_dob_30
744	ssn_lastname_unique_count_for_fulladdress_dob_14	794	ssn_zip5_unique_count_for_name_dob_60
745	ssn_lastname_unique_count_for_fulladdress_dob_30	795	ssn_zip5_unique_count_for_name_fulladdress_1
746	ssn_lastname_unique_count_for_fulladdress_dob_60	796	ssn_zip5_unique_count_for_name_fulladdress_3
747	ssn_lastname_unique_count_for_dob_homephone_1	797	ssn_zip5_unique_count_for_name_fulladdress_7
748	ssn_lastname_unique_count_for_dob_homephone_3	798	ssn_zip5_unique_count_for_name_fulladdress_14
749	ssn_lastname_unique_count_for_dob_homephone_7	799	ssn_zip5_unique_count_for_name_fulladdress_30
750	ssn_lastname_unique_count_for_dob_homephone_14	800	ssn_zip5_unique_count_for_name_fulladdress_60

801	ssn_zip5_unique_count_for_fulladdress_dob_1	851	ssn_name_unique_count_for_name_dob_7
802	ssn_zip5_unique_count_for_fulladdress_dob_3	852	ssn_name_unique_count_for_name_dob_14
803	ssn_zip5_unique_count_for_fulladdress_dob_7	853	ssn_name_unique_count_for_name_dob_30
804	ssn_zip5_unique_count_for_fulladdress_dob_14	854	ssn_name_unique_count_for_name_dob_60
805	ssn_zip5_unique_count_for_fulladdress_dob_30	855	ssn_name_unique_count_for_name_fulladdress_1
806	ssn_zip5_unique_count_for_fulladdress_dob_60	856	ssn_name_unique_count_for_name_fulladdress_3
807	ssn_zip5_unique_count_for_dob_homephone_1	857	ssn_name_unique_count_for_name_fulladdress_7
808	ssn_zip5_unique_count_for_dob_homephone_3	858	ssn_name_unique_count_for_name_fulladdress_14
809	ssn_zip5_unique_count_for_dob_homephone_7	859	ssn_name_unique_count_for_name_fulladdress_30
810	ssn_zip5_unique_count_for_dob_homephone_14	860	ssn_name_unique_count_for_name_fulladdress_60
811	ssn_zip5_unique_count_for_dob_homephone_30	861	ssn_name_unique_count_for_fulladdress_dob_1
812	ssn_zip5_unique_count_for_dob_homephone_60	862	ssn_name_unique_count_for_fulladdress_dob_3
813	ssn_zip5_unique_count_for_ssn_lastname_1	863	ssn_name_unique_count_for_fulladdress_dob_7
814	ssn_zip5_unique_count_for_ssn_lastname_3	864	ssn_name_unique_count_for_fulladdress_dob_14
815	ssn_zip5_unique_count_for_ssn_lastname_7	865	ssn_name_unique_count_for_fulladdress_dob_30
816	ssn_zip5_unique_count_for_ssn_lastname_14	866	ssn_name_unique_count_for_fulladdress_dob_60
817	ssn_zip5_unique_count_for_ssn_lastname_30	867	ssn_name_unique_count_for_dob_homephone_1
818	ssn_zip5_unique_count_for_ssn_lastname_60	868	ssn_name_unique_count_for_dob_homephone_3
819	ssn_zip5_unique_count_for_ssn_name_1	869	ssn_name_unique_count_for_dob_homephone_7
820	ssn_zip5_unique_count_for_ssn_name_3	870	ssn_name_unique_count_for_dob_homephone_14
821	ssn_zip5_unique_count_for_ssn_name_7	871	ssn_name_unique_count_for_dob_homephone_30
822	ssn_zip5_unique_count_for_ssn_name_14	872	ssn_name_unique_count_for_dob_homephone_60
823	ssn_zip5_unique_count_for_ssn_name_30	873	ssn_name_unique_count_for_ssn_lastname_1
824	ssn_zip5_unique_count_for_ssn_name_60	874	ssn_name_unique_count_for_ssn_lastname_3
825	ssn_zip5_unique_count_for_ssn_fulladdress_1	875	ssn_name_unique_count_for_ssn_lastname_7
826	ssn_zip5_unique_count_for_ssn_fulladdress_3	876	ssn_name_unique_count_for_ssn_lastname_14
827	ssn_zip5_unique_count_for_ssn_fulladdress_7	877	ssn_name_unique_count_for_ssn_lastname_30
828	ssn_zip5_unique_count_for_ssn_fulladdress_14	878	ssn_name_unique_count_for_ssn_lastname_60
829	ssn_zip5_unique_count_for_ssn_fulladdress_30	879	ssn_name_unique_count_for_ssn_zip5_1
830	ssn_zip5_unique_count_for_ssn_fulladdress_60	880	ssn_name_unique_count_for_ssn_zip5_3
831	ssn_zip5_unique_count_for_ssn_name_dob_1	881	ssn_name_unique_count_for_ssn_zip5_7
832	ssn_zip5_unique_count_for_ssn_name_dob_3	882	ssn_name_unique_count_for_ssn_zip5_14
833	ssn_zip5_unique_count_for_ssn_name_dob_7	883	ssn_name_unique_count_for_ssn_zip5_30
834	ssn_zip5_unique_count_for_ssn_name_dob_14	884	ssn_name_unique_count_for_ssn_zip5_60
835	ssn_zip5_unique_count_for_ssn_name_dob_30	885	ssn_name_unique_count_for_ssn_fulladdress_1
836	ssn_zip5_unique_count_for_ssn_name_dob_60	886	ssn_name_unique_count_for_ssn_fulladdress_3
837	ssn_name_unique_count_for_ssn_1	887	ssn_name_unique_count_for_ssn_fulladdress_7
838	ssn_name_unique_count_for_ssn_3	888	ssn_name_unique_count_for_ssn_fulladdress_14
839	ssn_name_unique_count_for_ssn_7	889	ssn_name_unique_count_for_ssn_fulladdress_30
840	ssn_name_unique_count_for_ssn_14	890	ssn_name_unique_count_for_ssn_fulladdress_60
841	ssn_name_unique_count_for_ssn_30	891	ssn_name_unique_count_for_ssn_name_dob_1
842	ssn_name_unique_count_for_ssn_60	892	ssn_name_unique_count_for_ssn_name_dob_3
843	ssn_name_unique_count_for_fulladdress_1	893	ssn_name_unique_count_for_ssn_name_dob_7
844	ssn_name_unique_count_for_fulladdress_3	894	ssn_name_unique_count_for_ssn_name_dob_14
845	ssn_name_unique_count_for_fulladdress_7	895	ssn_name_unique_count_for_ssn_name_dob_30
846	ssn_name_unique_count_for_fulladdress_14	896	ssn_name_unique_count_for_ssn_name_dob_60
847	ssn_name_unique_count_for_fulladdress_30	897	ssn_fulladdress_unique_count_for_ssn_1
848	ssn_name_unique_count_for_fulladdress_60	898	ssn_fulladdress_unique_count_for_ssn_3
849	ssn_name_unique_count_for_name_dob_1	899	ssn_fulladdress_unique_count_for_ssn_7
850	ssn_name_unique_count_for_name_dob_3	900	ssn_fulladdress_unique_count_for_ssn_14

901	ssn_fulladdress_unique_count_for_ssn_30	951	ssn_fulladdress_unique_count_for_ssn_name_dob_1
902	ssn_fulladdress_unique_count_for_ssn_60	952	ssn_fulladdress_unique_count_for_ssn_name_dob_3
903	ssn_fulladdress_unique_count_for_fulladdress_1	953	ssn_fulladdress_unique_count_for_ssn_name_dob_7
904	ssn_fulladdress_unique_count_for_fulladdress_3	954	ssn_fulladdress_unique_count_for_ssn_name_dob_14
905	ssn_fulladdress_unique_count_for_fulladdress_7	955	ssn_fulladdress_unique_count_for_ssn_name_dob_30
906	ssn_fulladdress_unique_count_for_fulladdress_14	956	ssn_fulladdress_unique_count_for_ssn_name_dob_60
907	ssn_fulladdress_unique_count_for_fulladdress_30	957	ssn_name_dob_unique_count_for_ssn_1
908	ssn_fulladdress_unique_count_for_fulladdress_60	958	ssn_name_dob_unique_count_for_ssn_3
909	ssn_fulladdress_unique_count_for_name_dob_1	959	ssn_name_dob_unique_count_for_ssn_7
910	ssn_fulladdress_unique_count_for_name_dob_3	960	ssn_name_dob_unique_count_for_ssn_14
911	ssn_fulladdress_unique_count_for_name_dob_7	961	ssn_name_dob_unique_count_for_ssn_30
912	ssn_fulladdress_unique_count_for_name_dob_14	962	ssn_name_dob_unique_count_for_ssn_60
913	ssn_fulladdress_unique_count_for_name_dob_30	963	ssn_name_dob_unique_count_for_fulladdress_1
914	ssn_fulladdress_unique_count_for_name_dob_60	964	ssn_name_dob_unique_count_for_fulladdress_3
915	ssn_fulladdress_unique_count_for_name_fulladdress_1	965	ssn_name_dob_unique_count_for_fulladdress_7
916	ssn_fulladdress_unique_count_for_name_fulladdress_3	966	ssn_name_dob_unique_count_for_fulladdress_14
917	ssn_fulladdress_unique_count_for_name_fulladdress_7	967	ssn_name_dob_unique_count_for_fulladdress_30
918	ssn_fulladdress_unique_count_for_name_fulladdress_14	968	ssn_name_dob_unique_count_for_fulladdress_60
919	ssn_fulladdress_unique_count_for_name_fulladdress_30	969	ssn_name_dob_unique_count_for_name_dob_1
920	ssn_fulladdress_unique_count_for_name_fulladdress_60	970	ssn_name_dob_unique_count_for_name_dob_3
921	ssn_fulladdress_unique_count_for_fulladdress_dob_1	971	ssn_name_dob_unique_count_for_name_dob_7
922	ssn_fulladdress_unique_count_for_fulladdress_dob_3	972	ssn_name_dob_unique_count_for_name_dob_14
923	ssn_fulladdress_unique_count_for_fulladdress_dob_7	973	ssn_name_dob_unique_count_for_name_dob_30
924	ssn_fulladdress_unique_count_for_fulladdress_dob_14	974	ssn_name_dob_unique_count_for_name_dob_60
925	ssn_fulladdress_unique_count_for_fulladdress_dob_30	975	ssn_name_dob_unique_count_for_name_fulladdress_1
926	ssn_fulladdress_unique_count_for_fulladdress_dob_60	976	ssn_name_dob_unique_count_for_name_fulladdress_3
927	ssn_fulladdress_unique_count_for_dob_homephone_1	977	ssn_name_dob_unique_count_for_name_fulladdress_7
928	ssn_fulladdress_unique_count_for_dob_homephone_3	978	ssn_name_dob_unique_count_for_name_fulladdress_14
929	ssn_fulladdress_unique_count_for_dob_homephone_7	979	ssn_name_dob_unique_count_for_name_fulladdress_30
930	ssn_fulladdress_unique_count_for_dob_homephone_14	980	ssn_name_dob_unique_count_for_name_fulladdress_60
931	ssn_fulladdress_unique_count_for_dob_homephone_30	981	ssn_name_dob_unique_count_for_fulladdress_dob_1
932	ssn_fulladdress_unique_count_for_dob_homephone_60	982	ssn_name_dob_unique_count_for_fulladdress_dob_3
933	ssn_fulladdress_unique_count_for_ssn_lastname_1	983	ssn_name_dob_unique_count_for_fulladdress_dob_7
934	ssn_fulladdress_unique_count_for_ssn_lastname_3	984	ssn_name_dob_unique_count_for_fulladdress_dob_14
935	ssn_fulladdress_unique_count_for_ssn_lastname_7	985	ssn_name_dob_unique_count_for_fulladdress_dob_30
936	ssn_fulladdress_unique_count_for_ssn_lastname_14	986	ssn_name_dob_unique_count_for_fulladdress_dob_60
937	ssn_fulladdress_unique_count_for_ssn_lastname_30	987	ssn_name_dob_unique_count_for_dob_homephone_1
938	ssn_fulladdress_unique_count_for_ssn_lastname_60	988	ssn_name_dob_unique_count_for_dob_homephone_3
939	ssn_fulladdress_unique_count_for_ssn_zip5_1	989	ssn_name_dob_unique_count_for_dob_homephone_7
940	ssn_fulladdress_unique_count_for_ssn_zip5_3	990	ssn_name_dob_unique_count_for_dob_homephone_14
941	ssn_fulladdress_unique_count_for_ssn_zip5_7	991	ssn_name_dob_unique_count_for_dob_homephone_30
942	ssn_fulladdress_unique_count_for_ssn_zip5_14	992	ssn_name_dob_unique_count_for_dob_homephone_60
943	ssn_fulladdress_unique_count_for_ssn_zip5_30	993	ssn_name_dob_unique_count_for_ssn_lastname_1
944	ssn_fulladdress_unique_count_for_ssn_zip5_60	994	ssn_name_dob_unique_count_for_ssn_lastname_3
945	ssn_fulladdress_unique_count_for_ssn_name_1	995	ssn_name_dob_unique_count_for_ssn_lastname_7
946	ssn_fulladdress_unique_count_for_ssn_name_3	996	ssn_name_dob_unique_count_for_ssn_lastname_14
947	ssn_fulladdress_unique_count_for_ssn_name_7	997	ssn_name_dob_unique_count_for_ssn_lastname_30
948	ssn_fulladdress_unique_count_for_ssn_name_14	998	ssn_name_dob_unique_count_for_ssn_lastname_60
949	ssn_fulladdress_unique_count_for_ssn_name_30	999	ssn_name_dob_unique_count_for_ssn_zip5_1
950	ssn_fulladdress_unique_count_for_ssn_name_60	1000	ssn_name_dob_unique_count_for_ssn_zip5_3

1001	ssn_name_dob_unique_count_for_ssn_zip5_7	1009	ssn_name_dob_unique_count_for_ssn_name_30
1002	ssn_name_dob_unique_count_for_ssn_zip5_14	1010	ssn_name_dob_unique_count_for_ssn_name_60
1003	ssn_name_dob_unique_count_for_ssn_zip5_30	1011	ssn_name_dob_unique_count_for_ssn_fulladdress_1
1004	ssn_name_dob_unique_count_for_ssn_zip5_60	1012	ssn_name_dob_unique_count_for_ssn_fulladdress_3
1005	ssn_name_dob_unique_count_for_ssn_name_1	1013	ssn_name_dob_unique_count_for_ssn_fulladdress_7
1006	ssn_name_dob_unique_count_for_ssn_name_3	1014	ssn_name_dob_unique_count_for_ssn_fulladdress_14
1007	ssn_name_dob_unique_count_for_ssn_name_7	1015	ssn_name_dob_unique_count_for_ssn_fulladdress_30
1008	ssn_name_dob_unique_count_for_ssn_name_14	1016	ssn_name_dob_unique_count_for_ssn_fulladdress_60