3/25/2021

DSO 562 Project 2

Identity Fraud in the Credit Card Dataset

Team Member:

Xue Gao,
Tianze Yang,
Yang Jiao,
Yunlong Wang,
Huiyi Zhang,
Xinzhu Zhang

University of Southern California
MARSHALL SCHOOL OF BUSINESS

Table of Contents

Executive Summary

As technology advances rapidly evolve around methods of payments, also presented are potential exposures to various risks. As the most common type of identity theft, Credit card fraud is widely prevalent in the U.S. According to research, of the 1.5 billion credit cards issued in the United States, millions fall victim to this unscrupulous tactic each year.

In addition to performing unauthorized transactions using stolen or lost credit cards, it has been discovered that fraudsters often apply for a credit card in someone else's name. Basic information such as legal name, date of birth, address, and social security numbers are rudimentary to this scheme; overtime, the criminals also adopted less conventional methods to extract relevant info from supporting documents to fly under the radar.

In this project, we aim to build a real-time fraud detection model to predict if a credit card applicant uses someone else's in combination with made-up information to commit fraud.

The credit card application dataset we use contains one million rows of records each with ten columns, including date, ssn, first name, last name, address, zip5, dob, homephone, and fraud label.

In order to develop this predictive model, we went through several steps. At first, we did data cleaning and sorting to adjust the data type and replace frivolous values. Then, we did feature engineering by creating combination group variables and day since, velocity, and relative velocity variables for each combination group.

After generating candidate variables, we operated feature selection to select a number of best potential variables to train the models. By having a dataset split in training, testing, and out of time validation, we trained models by applying different machine learning algorithms and compared the performance by calculating average fraud detection rate at 3% to select the best model to be our final model.

As a result, we found that the Gradient Boosting Tree with Learning_rate 0.01, n_estimator 1100, and max_depth of 5 has the highest average FDR at 3% in the test set. After that, we used that model to calculate the bins and cumulative goods, bads and built the tables in the results section.

Description of Data

Applications Data is a dataset containing records of 1,000,000 applications. It includes fields such as date of application, SSN, first and last name, address, zip code, date of birth, home phone number, and fraud label of each applicant.

File Name: applications data.csv

Data Source: An identity fraud prevention company

Time Period: Jan 1st 2016 – Dec 31st 2016

Number of Records: 1,000,000 records

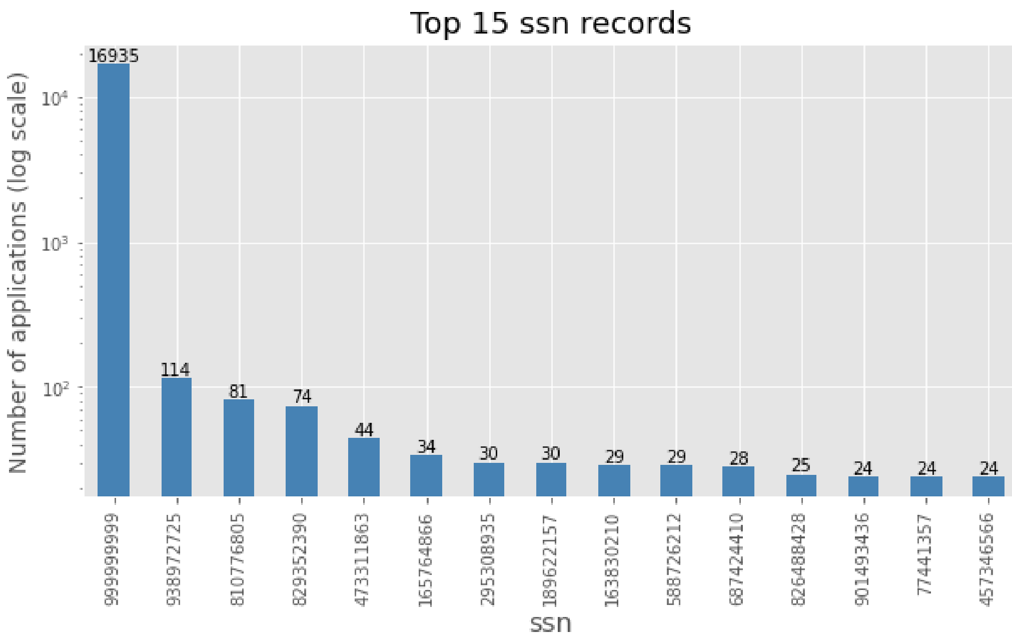
Number of Fields: 9 variables in total: 7 categorical variables, 2 date variables

Field Name	Data Type	%Populated	Unique number
date	Date variable	100%	365
ssn	Categorical variable	100%	835819
firstname	Categorical variable	100%	78136
lastname	Categorical variable	100%	177001
address	Categorical variable	100%	828774
zip5	Categorical variable	100%	26370
dob	Date variable	100%	42673

homephone	Categorical variable	100%	28244
Fraud_label	Categorical variable	100%	2

date: The date when the credit card application was filled. It ranges from Jan 1st 2016 – Dec 31st 2016.

ssn: The SSN used for that particular credit card application.



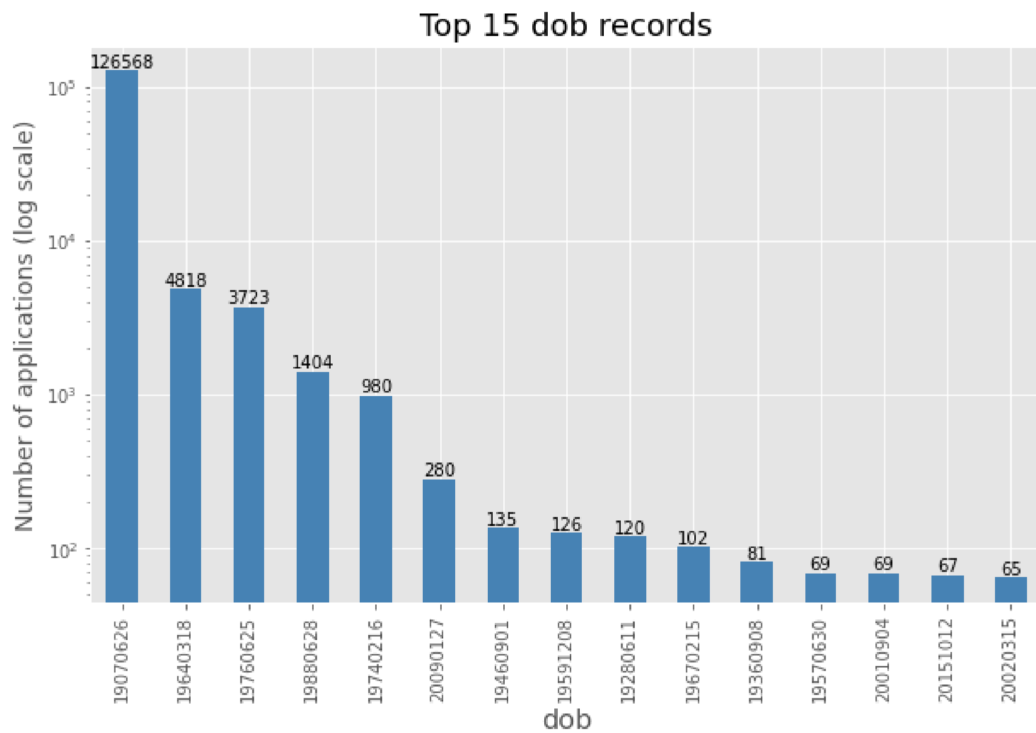
firstname: The first name used for that credit card application.

lastname: The last name used for that credit card application.

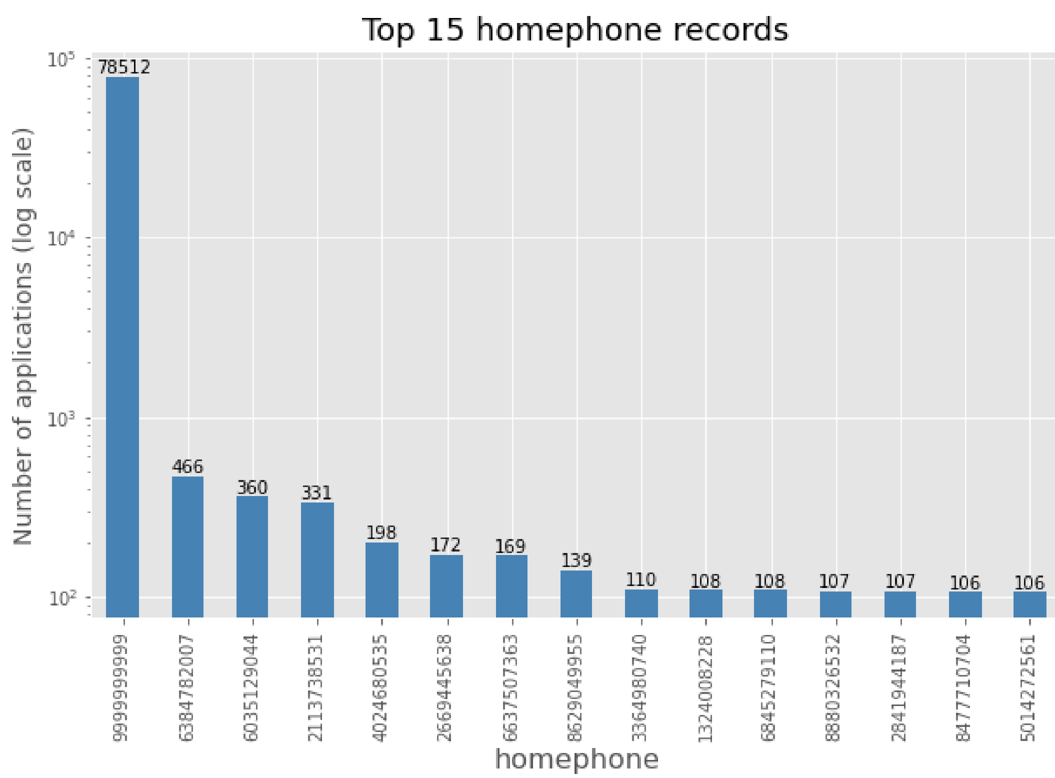
address: The address used for that credit card application, which includes street number and street name.

zip5: The 5-digit zip code used for that credit card application.

dob: The date of birth that the applicants used for application, the format is YYYY-MM-DD.



homephone: The phone number that the applicants used for credit card application.



fraud_lable: Whether the record is considered fraud.

Data Cleaning

The original dataset has a total of 9 fields. Each of the fields is 100% populated. There are no missing values in any of the fields.

The date field was originally in the type of int64. We firstly changed the data type to a string, added dashes in between to separate the year, month and day, and finally converted it to a datetime variable using the pandas to_datetime function.

The zip5 field has some values in 4 digits and others in 5 digits. To unify the number of digits in each entry, we formatted the 4-digit ones by adding a 0 in front of each entry.

The most common value in the ssn field was 9999999999, which was clearly a frivolous value. To fix this problem, we replaced the frivolous value with a field that would not link -- the negative of the record number and then formatted the entries into 9 digits.

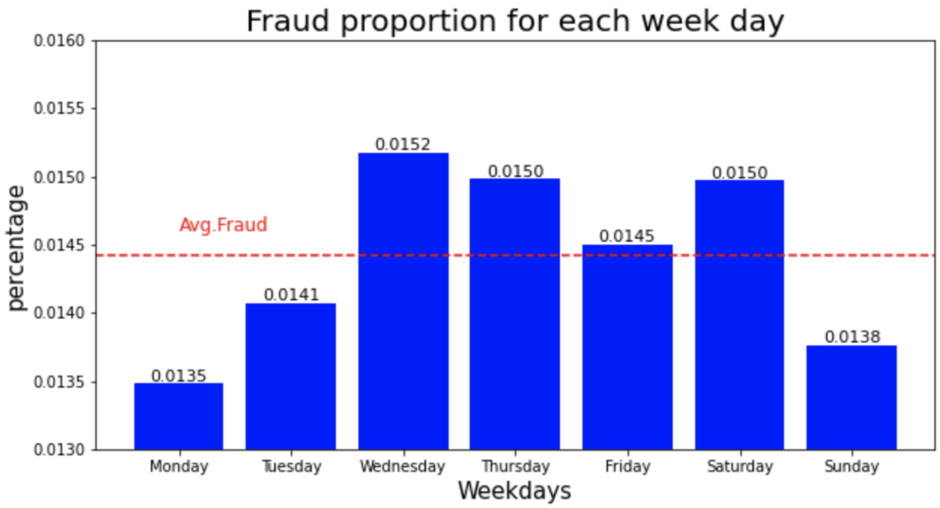
Similarly, the most common value in the homephone field was 9999999999. We replaced the frivolous value with the negative of the record number and formatted it into a 10-digit number.

The address field had around a thousand frivolous values of “123 MAIN ST”. We created a value by concatenating the record number with the word “RECORD” and used it to replace each frivolous value under the address field.

The most common field value in the dob field was 19070626, which was an impossible number. Again, we replaced it with the negative of the record number and formatted it into an 8-digit number.

Candidate Variables

First of all, we created a variable called "day of the week" to figure out the risk of each day in a week. From the risk table for day of week, it was obvious that Monday had the lowest fraud risk and Wednesday had the highest risk.



Then we linked the original fields and created new entities:

name = firstname + lastname
fulladdress = address + zip5
name_dob = name + dob
name_fulladdress = name + fulladdress
name_homephone = name + homephone
fulladdress_dob=fulladdress+dob
fulladdress_homephone=fulladdress+homephone
dob_homephone=dob+homephone
homephone_name_dob=homephone+name_dob

Then we linked applicants' SSN to all these entities described above. In total, we created 28 entities.

Next, we built 'velocity' variables for all these attributes, which means the number of records with the same attributes over the last 0, 1, 3, 7, 14, 30 days.

Another group 'day since' variables mean that how many days has passed since the last time the attributes appeared.

In the end, we built ‘relative velocity’ group variables, and they represented the proportion of the number of times we have seen that entity in the past days comes from the recent past.

Variable groups	Variable name format
Velocity	attributes _count_xx
Day since	attibutes_since
Relative velocity	attributes_count_yy_by_xx

xx: 0, 1, 3, 7, 14 days ; yy: recent days

Feature Selection Process

During the feature selection stage, we calculated the Kolmogorov–Smirnov (KS) and Fraud Detection Rate (FDR) value for each variable and used backward stepwise methods to select 30 variables for our final models.

- Kolmogorov–Smirnov (KS)
  Kolmogorov–Smirnov is the measurement of how well two distributions are separated. The larger the KS, the more separate the two distributions. We used KS to measure the differences between fraud records and non-fraud records for each variable created. Specifically, for each variable, we gathered a list of fields corresponding to fraud records and the other list of random numbers between 0 and 1. Then we applied stats.ks.2samp function to compute KS for all variables.

- Fraud Detection Rate (FDR)
  Fraud Detection Rate is the percentage of all the fraud found at a score cutoff. 3% of FDR means the calculation of how many fraud records in the top 3% of all records. Specifically, we sort the data and compute the number of bad records in the top 3% of the records, then divided by the total number of fraud data.

Wrapped Method:

Backward step-by-step selection involves starting with all candidate variables, testing the deletion of each variable using the selection model to meet the criteria and repeating this process until no further variable can be eliminated. In the wrapped method, we use FDR score as the score in the function RFECV in order to have a better result. Below is a list of the 30 variables chosen by backward selection and are ones we used in our final models:

Variable	Description	Variable	Description
address_count_0	Number of same address seen in the past 0 days	fulladdress_count_0	Number of same address plus zip code seen in the past 0 days

address_count_0_by_3	Number of same address seen in the past 0 days divided by the same group in 3 days.	fulladdress_count_0_by_14	Number of same full address and dob seen in the past 0 days divided by the same group in 14 days.
address_count_0_by_7	Number of same address seen in the past 0 days divided by the same group in 7 days.	fulladdress_count_0_by_3	Number of same full address and dob seen in the past 0 days divided by the same group in 3 days.
address_count_1	Number of same address seen in the past 1 days	fulladdress_count_0_by_7	Number of same full address and dob seen in the past 0 days divided by the same group in 7 days.
address_count_1_by_7	Number of same address seen in the past 1 days divided by the same group in 7 days.	fulladdress_count_1	Number of same address plus zip code seen in the past 1 days
address_count_3	Number of same address seen in the past 3 days	fulladdress_count_3	Number of same address plus zip code seen in the past 3 days
address_count_30	Number of same address seen in the past 30 days	fulladdress_count_30	Number of same address plus zip code seen in the past 30 days

homephone_count_3	Number of same phone number seen in the past 3 days	fulladdress_homephone_count_30	Number of same address plus zip plus home phone number code seen in the past 30 days
name_dob_count_0_by_14	Number of same name and dob seen in the past 0 days divided by the same group in 14 days.	fulladdress_homephone_count_7	Number of same address plus zip plus home phone number code seen in the past 7 days
name_dob_count_14	Number of same address plus zip code seen in the past 30 days	ssn_count_0_by_30	Number of same ssn in the past 0 days divided by the same group in 30 days.
name_dob_count_30	Number of same address plus zip code seen in the past 30 days	ssn_count_30	Number of same ssn code seen in the past 30 days
name_dob_count_7	Number of same address plus zip code seen in the past 30 days	ssn_count_7	Number of same ssn code seen in the past 7 days
ssn_dob_count_0_by_14	Number of same ssn and dob seen in the past 0 days divided by the same group in 14 days.	ssn_dob_count_30	Number of same SSN and date of birth seen in the past 30 days

ssn_dob_count_0_by_30	Number of same ssn and dob seen in the past 0 days divided by the same group in 30 days.	ssn_dob_count_7	Number of same SSN and date of birth seen in the past 7 days
ssn_dob_count_14	Number of same address plus zip code seen in the past 30 days	ssn_firstname_count_14	Number of same SSN and first name seen in the past 14 days

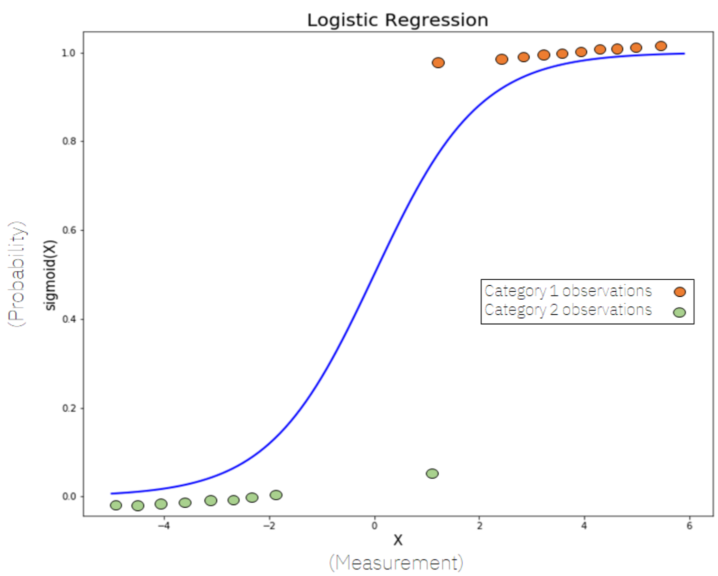
Model Algorithms

After feature selection, we applied different machine learning algorithms to train models and calculated the average FDR at 3% for the train, test, and OOT data for each model. We applied 4 machine learning algorithms as below.

Logistic Regression:

Logistic regression analysis studies the relationship between a categorical dependent variable and a set of independent variables. Logistic regression can predict the probability of an outcome that has two values (i.e., 0 and 1).

In logistic regression, we don't directly fit a straight line to our data like in linear regression. Instead, we fit a S-shaped curve, called Sigmoid, to our observations.

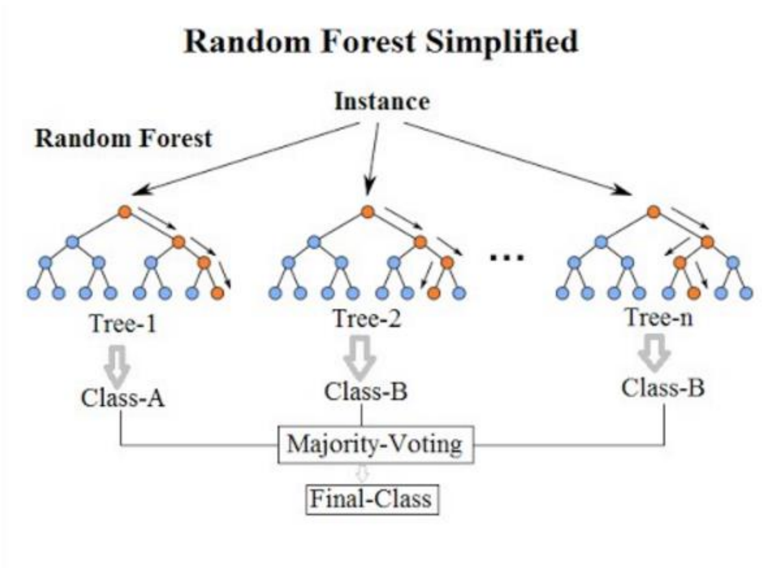


In the logistic regression the constant moves the curve left and right and the slope defines the steepness of the curve. By simple transformation, the logistic regression equation can be written as: $p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_n X_n}}$

With the fraud label being the dependent variable and the 30 selected features being the independent variables, we trained the logistic regression model with parameter $cv = 10$and $Cs = 0.01$. We then used the model to predict the probability of fraud on the train, test and oot dataset separately and get the best average FDR rate of 0.5445, 0.5364 and 0.5207 for each of the dataset.

Model	Parameter		Average FDR(%) at 3%		
Logistic Regression	Total variables	# of variables selected	Train	Test	OOT
1	30	10	0.3577	0.3572	0.3206
2	30	20	0.5231	0.5213	0.5031
3	30	25	0.5436	0.5358	0.5194
4	30	30	0.5445	0.5364	0.5207

Random Forests:

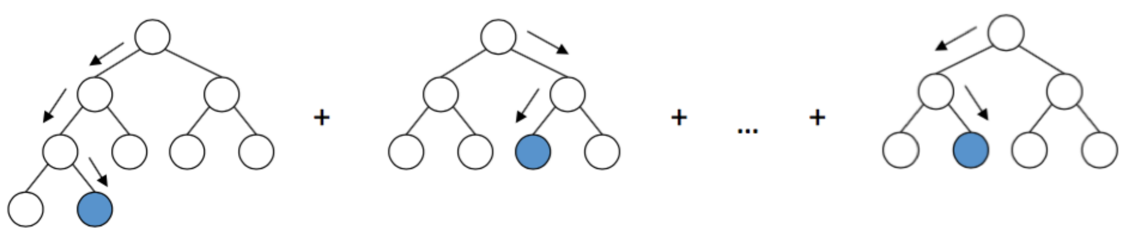


Random forest is an integrated learning method based on classification and regression. It constructs a large number of decision trees during training and outputs the pattern of classes (classification) or average/average prediction of a single tree (regression). We used the randomForest package in the Python sklearn package. When we trained the random forest model, we tried generating 100, 200 and 300 decision trees with max_depth of 60, 70, 80.

Random Forests	# of Vars	N_estimator	Max_depth	max_features	Train	Test	OOT
1	30	100	60	7	0.567	0.553	0.535

2	30	100	70	7	0.568	0.553	0.535
3	30	100	80	7	0.567	0.553	0.536
4	30	200	60	7	0.567	0.553	0.537
5	30	200	70	7	0.567	0.552	0.536
6	30	200	80	7	0.567	0.553	0.535
7	30	300	60	7	0.567	0.553	0.537
8	30	300	70	7	0.567	0.552	0.536
9	30	300	80	7	0.567	0.552	0.535

Gradient Boosting Trees:



Like other boosting methods, gradient boosting combines weak 'learners' into a single strong 'learner'. The goal of this model is to predict values by minimizing the mean squared error. The model in supervised learning usually refers to the mathematical structure by which the prediction y_i is made from the input x_i. Gradient boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor, i.e., each succeeding one attempts to fit the new predictor to the residual errors made by the previous one. In this project we use this model to be a classification model to determine whether it is fraud.

We use the GradientBoostingClassifier from the Scikit-learn package to predict the correct classification of fraud case and use the predict_proba method to rank the predictions by their likelihood of being fraud before tallying the results and compare with the true count of fraud case in the test and oot set. As mentioned previously, this will be defined as the top 3% FDR rate and effectiveness of model selection will rely heavily on this metric.

In the quest for a best fitting model in this ensemble from Sci-kit Learn, 3 hyperparameters are adjusted to find the best possible outcome, namely,

- max_depth: the maximum length from a root to a node in a decision tree.
- n_estimators: the number of trees in the ensemble.
- learning_rate: a value to adjust the amount of information retained from each tree to compensate against overfitting.

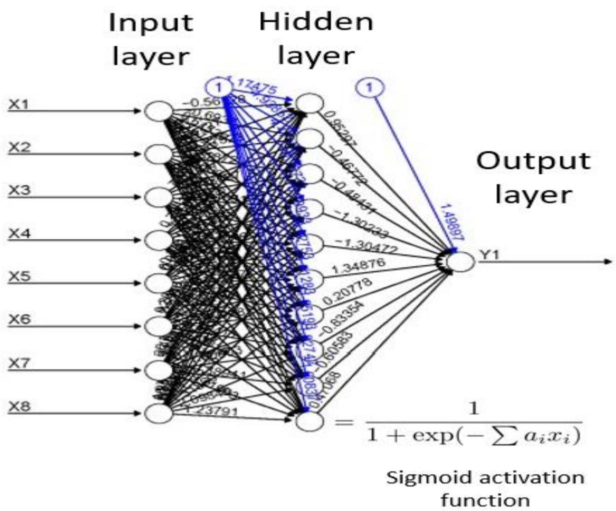
We have tried 60 combinations of different hyperparameters. The max depths are 1,5,6, the numbers of trees are 700,800,1100,1400 and the learning rates goes from 0.01 to 0.1. We picked 5 models from the 60 models with the top 5 testing FDR.

In the following results, we found that Gradient Boosting Tree 1 with max_depth = 5, N_estimators = 1100, and leaning_rate = 0.01 provided the most promising result where top 3%FDR equals 0.5592 and 0.5369 for the test and oot datasets, respectively.

Gradient Boosting Trees	Learning rate	n_estimators	max_depth	Training FDR	Testing FDR	OOT FDR
Gradient Boosting Tree 1	0.01	1100	5	0.5644	0.5592	0.5369
Gradient Boosting Tree 2	0.01	800	6	0.5647	0.559	0.536
Gradient Boosting Tree 3	0.01	800	5	0.564	0.5589	0.5369
Gradient Boosting Tree 4	0.01	700	5	0.5637	0.5586	0.5369

Gradient Boosting Tree 5	0.03	1400	1	0.5551	0.5523	0.5313

Neural Net:



Neural network is a machine learning algorithm that they make use of architecture that mimics how the neurons work in the brain. For example, a brain neuron receives an input and based on that input, fires off an output that is used by another neuron. The neural network simulates this behavior in learning about collecting the data and predicting outcomes.

In the above graph, we can see that a typical neural net consists of an input layer, hidden layers, and an output layer. The input layer is formed by all the independent variables. Each hidden layer is a set of nodes (neurons). Each neuron in the hidden layer receives weighted signals from all the nodes in the previous and transforms the linear combination of signals. The transform/activation function can be a logistic function(sigmoid) or something else. Finally, the output layer is the dependent variable.

We trained six neural net models with different hyperparameters as in the following table and listed all average FDRs at 3% on training, testing, and oot data. According to each model's performance, we didn't observe overfitting so we could trust the results of these models. The best model is the highlighted Neural Net 6 with average FDR of 0.5339 on oot data.

Neural Net	Nodes in layer 1	Activation function layer 1	Nodes in layer 2	Activation function layer 2	Activation function output layer	Optimizer	Epoch	Batch size	Training FDR	Testing FDR	OOT FDR
Neural net 1	10	relu	None	None	sigmoid	adam	20	10	0.556	0.5488	0.5264
Neural net 2	15	relu	None	None	sigmoid	adam	25	10	0.5449	0.5386	0.5201
Neural net 3	25	relu	None	None	sigmoid	adam	30	15	0.5433	0.5386	0.5184
Neural net 4	10	relu	10	relu	sigmoid	adam	20	15	0.5607	0.5548	0.5314
Neural net 5	15	relu	10	relu	sigmoid	adam	25	10	0.5608	0.5528	0.5331
Neural net 6	15	relu	15	relu	sigmoid	adam	30	20	0.5602	0.5519	0.5339

Results

By training several models using different machine learning algorithms and adjusting hyperparameters for each particular machine learning algorithm, we compared the model performance based on average FDR at 3% on testing data and found that the best model is the gradient boosting tree as highlighted in the following graph. Of the attempted methods, a common peak value for the FDR at 3% is reached for the OOT dataset at around 53.5%. Since out of time data is considered unknown information in reality, we have little choice but to turn to the test set results. Random forest and Gradient Boosting Tree came in neck to neck and the latter won by a slight lead of around 0.5%. It is worth mentioning that the Gradient Boosting Tree is a more cultivated model in terms of number of trees and learning rate, making this choice both more time consuming and computationally demanding. Random Forest in comparison provided similar results with much less trees but deeper depth, demonstrating a diminishing marginal return on the complexity of model structure. We ultimately opted to use test set average FDR at 3% to be the guideline for the choice of model since algorithm efficiency and time limit is outside the scope of this project.

Model	Parameters							Average FDR at 3%		
Logistic Regression	# variables							Train	Test	OOT
	10							0.3577	0.3572	0.3206
	20							0.5231	0.5213	0.5031
	25							0.5436	0.5358	0.5194
	30							0.5445	0.5364	0.5207
Random Forest	# variabl	n_estimato	max_depth	max_features				Train	Test	OOT
	30	100	60	7				0.567	0.553	0.535
	30	100	70	7				0.568	0.553	0.535
	30	100	80	7				0.567	0.553	0.536
	30	200	60	7				0.567	0.553	0.537
	30	200	70	7				0.567	0.552	0.536
	30	200	80	7				0.567	0.553	0.535
	30	300	60	7				0.567	0.553	0.537
	30	300	70	7				0.567	0.552	0.536
	30	300	80	7				0.567	0.552	0.535
Gradient Boosting Tree	# variabl	n_estimato	max_depth	learning rate				Train	Test	OOT
	30	1100	5	0.01				0.5644	0.5592	0.5369
	30	800	6	0.01				0.5647	0.559	0.536
	30	800	5	0.01				0.564	0.5589	0.5369
	30	700	5	0.01				0.5637	0.5586	0.5369
	30	1400	1	0.03				0.5551	0.5523	0.5313
Neural Net	# variable	layer 1 noc	layer 1 activa	layer 2 nod	layer 2 activa	epoch	batch size	Train	Test	OOT
	30	10	relu	0	none	20	10	0.556	0.5488	0.5264
	30	15	relu	0	none	25	10	0.5449	0.5386	0.5201
	30	25	relu	0	none	30	15	0.5433	0.5386	0.5184
	30	10	relu	10	relu	20	15	0.5607	0.5548	0.5314
	30	15	relu	10	relu	25	10	0.5608	0.5528	0.5331
	30	15	relu	15	relu	30	20	0.5602	0.5519	0.5339

After selecting the first gradient boosting tree as our final model, we ran the model on training data again and reevaluated the model on both testing and oot data. We took a closer look at the structure of the predicted results on all three datasets by computing critical statistics by individual percentiles and then generated the tables below to examine how our final model performed on the three datasets when detecting fraudulent applications.

As a brief review, the oot data represented approximately 16.65% of the entire dataset, the remaining are split into training and testing by 75% and 25%, respectively. The underlying fraud rate is consistent for all three sets at around 1.43%, or 1 fraud case in about 70 applications.

More specifically, the first percentile of data ranked by our model of choice uniformly contained the largest percentage in true frauds identified, up to 76.7% in training, 76.56% in testing and 72.71% in OOT. This means that approximately 3 out of 4 applications that we deny arbitrarily without additional information in the first percentile is a true fraud case. This is an immediate choice for any financial institution where no other alternative is present. The gains in fraud detection percentage quickly diminishes in subsequent percentiles, however; before we penetrate through the first 4% of all applications, the true percentage of fraud in each additional percentile of data falls below 1%. As a comparison, the company now looks at a proposal much less tempting: give up 5% of all applications in the hope of catching up to 57.28% of all frauds, still good if the institution's profit margin prevails but nowhere near as good if the model's effectiveness holds through.

Model Performance on Training

Train	# of Records	# Good		# of bads		Fraud rate						
	556498	548443		8054		0.014472649						
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	5565	1297	4268	23.30638	76.69362	5565	1297	4268	0.236488	52.992302	52.75581	0.303889
1	5565	5384	181	96.74753	3.252471	11130	6681	4449	1.218176	55.239632	54.02146	1.501686
2	5565	5489	76	98.63432	1.365678	16695	12170	4525	2.219009	56.183263	53.96425	2.689503
3	5565	5512	53	99.04762	0.952381	22260	17682	4578	3.224036	56.841321	53.61729	3.862385
4	5565	5516	49	99.1195	0.880503	27825	23198	4627	4.229792	57.449714	53.21992	5.013616
5	5565	5511	54	99.02965	0.97035	33390	28709	4681	5.234637	58.120189	52.88555	6.133091
6	5565	5521	44	99.20934	0.790656	38955	34230	4725	6.241305	58.666501	52.4252	7.244444
7	5565	5524	41	99.26325	0.736748	44520	39754	4766	7.24852	59.175565	51.92705	8.341167
8	5565	5520	45	99.19137	0.808625	50085	45274	4811	8.255006	59.734294	51.47929	9.410518
9	5565	5525	40	99.28122	0.718778	55650	50799	4851	9.262403	60.230941	50.96854	10.47186
10	5565	5528	37	99.33513	0.66487	61215	56327	4888	10.27035	60.69034	50.41999	11.52353
11	5565	5525	40	99.28122	0.718778	66780	61852	4928	11.27774	61.186988	49.90924	12.55114
12	5565	5524	41	99.26325	0.736748	72345	67376	4969	12.28496	61.696052	49.41109	13.55927
13	5565	5526	39	99.29919	0.700809	77910	72902	5008	13.29254	62.180283	48.88774	14.55711
14	5565	5533	32	99.42498	0.575022	83475	78435	5040	14.3014	62.577601	48.27621	15.5625
15	5565	5522	43	99.22731	0.772686	89040	83957	5083	15.30825	63.111497	47.80325	16.51721
16	5564	5527	37	99.33501	0.664989	94604	89484	5120	16.31601	63.570896	47.25489	17.47734
17	5565	5532	33	99.40701	0.592992	100169	95016	5153	17.32468	63.980631	46.65595	18.43897
18	5565	5522	43	99.22731	0.772686	105734	100538	5196	18.33153	64.514527	46.183	19.34911
19	5565	5542	23	99.5867	0.413297	111299	106080	5219	19.34203	64.800099	45.45807	20.32573

Model Performance on Testing

Test	# Record		# goods		# bads		Fraud Rate					
	238499		235067		3432		0.014389997					
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
0	2385	559	1826	23.43816	76.56184	2385	559	1826	0.237805	53.2051282	52.96732	0.306134
1	2385	2311	74	96.89727	3.102725	4770	2870	1900	1.220929	55.3613054	54.14038	1.510526
2	2385	2357	28	98.826	1.174004	7155	5227	1928	2.223621	56.1771562	53.95353	2.7111
3	2385	2366	19	99.20335	0.796646	9540	7593	1947	3.230143	56.7307692	53.50063	3.899846
4	2385	2366	19	99.20335	0.796646	11925	9959	1966	4.236664	57.2843823	53.04772	5.065615
5	2385	2361	24	98.99371	1.006289	14310	12320	1990	5.241059	57.983683	52.74262	6.190955
6	2385	2369	16	99.32914	0.67086	16695	14689	2006	6.248857	58.4498834	52.20103	7.322532
7	2385	2358	27	98.86792	1.132075	19080	17047	2033	7.251975	59.2365967	51.98462	8.385145
8	2385	2362	23	99.03564	0.964361	21465	19409	2056	8.256795	59.9067599	51.64997	9.440175
9	2385	2364	21	99.1195	0.880503	23850	21773	2077	9.262466	60.518648	51.25618	10.48291
10	2385	2370	15	99.37107	0.628931	26235	24143	2092	10.27069	60.955711	50.68502	11.54063
11	2385	2368	17	99.28721	0.712788	28620	26511	2109	11.27806	61.451049	50.17299	12.57041
12	2385	2369	16	99.32914	0.67086	31005	28880	2125	12.28586	61.9172494	49.63139	13.59059
13	2385	2372	13	99.45493	0.545073	33390	31252	2138	13.29493	62.2960373	49.0011	14.6174
14	2385	2364	21	99.1195	0.880503	35775	33616	2159	14.3006	62.9079254	48.60732	15.57017
15	2385	2369	16	99.32914	0.67086	38160	35985	2175	15.3084	63.3741259	48.06572	16.54483
16	2385	2376	9	99.62264	0.377358	40545	38361	2184	16.31918	63.6363636	47.31719	17.56456
17	2385	2367	18	99.24528	0.754717	42930	40728	2202	17.32612	64.1608392	46.83472	18.49591
18	2385	2371	14	99.413	0.587002	45315	43099	2216	18.33477	64.5687646	46.23399	19.44901
19	2385	2367	18	99.24528	0.754717	47700	45466	2234	19.34172	65.0932401	45.75152	20.35184

Model Performance on OOT

OOT	# Record		# goods		# bads		Fraud rate					
	163772		161427		2345		0.01432					
	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FD	KS	FPR
0	1638	447	1191	27.28938	72.71062	1638	447	1191	0.276905	50.78891	50.51201	0.375315
1	1637	1588	49	97.00672	2.99328	3275	2035	1240	1.260632	52.87846	51.61783	1.641129
2	1638	1620	18	98.9011	1.098901	4913	3655	1258	2.264181	53.64606	51.38187	2.905405
3	1638	1624	14	99.1453	0.854701	6551	5279	1272	3.270209	54.24307	50.97286	4.150157
4	1638	1625	13	99.20635	0.793651	8189	6904	1285	4.276856	54.79744	50.52059	5.372763
5	1637	1625	12	99.26695	0.733048	9826	8529	1297	5.283503	55.30917	50.02567	6.575944
6	1638	1615	23	98.59585	1.404151	11464	10144	1320	6.283955	56.28998	50.00602	7.684848
7	1638	1618	20	98.779	1.221001	13102	11762	1340	7.286266	57.14286	49.85659	8.777612
8	1637	1628	9	99.45021	0.549786	14739	13390	1349	8.294771	57.52665	49.23188	9.925871
9	1638	1629	9	99.45055	0.549451	16377	15019	1358	9.303896	57.91045	48.60655	11.05965
10	1638	1626	12	99.2674	0.732601	18015	16645	1370	10.31116	58.42217	48.11101	12.14964
11	1638	1620	18	98.9011	1.098901	19653	18265	1388	11.31471	59.18977	47.87505	13.15922
12	1637	1628	9	99.45021	0.549786	21290	19893	1397	12.32322	59.57356	47.25034	14.2398
13	1638	1627	11	99.32845	0.671551	22928	21520	1408	13.3311	60.04264	46.71154	15.28409
14	1638	1626	12	99.2674	0.732601	24566	23146	1420	14.33837	60.55437	46.216	16.3
15	1638	1620	18	98.9011	1.098901	26204	24766	1438	15.34192	61.32196	45.98004	17.22253
16	1637	1619	18	98.90043	1.099572	27841	26385	1456	16.34485	62.08955	45.7447	18.12157
17	1638	1629	9	99.45055	0.549451	29479	28014	1465	17.35397	62.47335	45.11937	19.12218
18	1638	1632	6	99.6337	0.3663	31117	29646	1471	18.36496	62.72921	44.36425	20.15364
19	1637	1622	15	99.08369	0.91631	32754	31268	1486	19.36975	63.36887	43.99912	21.04172

Conclusions

The plethora of methods through which frauds are committed are unfathomable to recount; however, there are traces that remain to be picked up by the vigilant and learned. In this project, we attempted to predict the underlying fraudulent activity by observing none other than the information filled out on the applications. Via well-traversed methodologies developed by our seniors and predecessors, we were able to peek at the subtleties involuntarily revealed by the schemers, magnify their actions with tools based in both statistics and machine learning methods, and acutely label the potentially guilty. hile the predictions are far from perfect, these methods cemented a handle on a problem otherwise extremely difficult to tackle. Moreover, we duly believe that with more experience on the subject matter and techniques, we could significantly improve the end results by creating better expert variables and implementing more appropriate data cleaning techniques and machine learning algorithms in a more polished manner. Also observed was that machine learning algorithms tend to perform better with datasets large in both volume and diversity and with models that churn through the numbers laboriously. These are invaluable experiences that we will be utilizing faithfully in the upcoming projects in the hope of a more streamlined process and more indicative results.

Appendix

Data Quality Report

File Description:

Applications Data is a dataset containing records of 1,000,000 applications. It includes fields such as date of application, SSN, first and last name, address, zip code, date of birth, home phone number, and fraud label of each applicant.

File Name: applications data.csv

Number of Records: 1,000,000 records

Number of Fields: 9 variables in total: 7 categorical variables, 2 date variables

Field Name	Data Type	% Populated	Unique number
date	Date variable	100%	365
ssn	Categorical variable	100%	835819
firstname	Categorical variable	100%	78136
lastname	Categorical variable	100%	177001
address	Categorical variable	100%	828774
zip5	Categorical variable	100%	26370
dob	Date variable	100%	42673

homephone	Categorical variable	100%	28244
fraud_label	Categorical variable	100%	2

Time of Records: Jan 1st 2016 – Dec 31st 2016

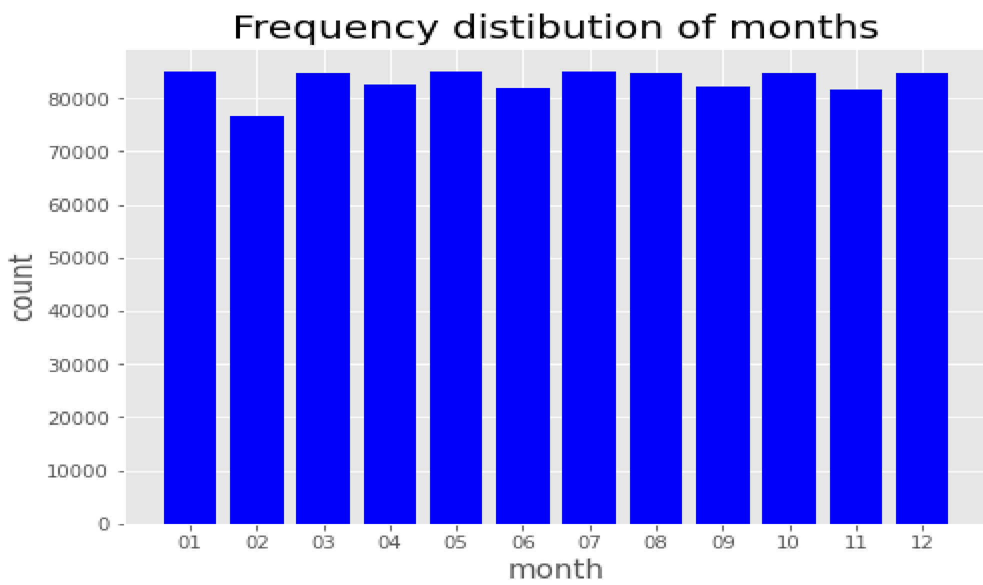
Field 1

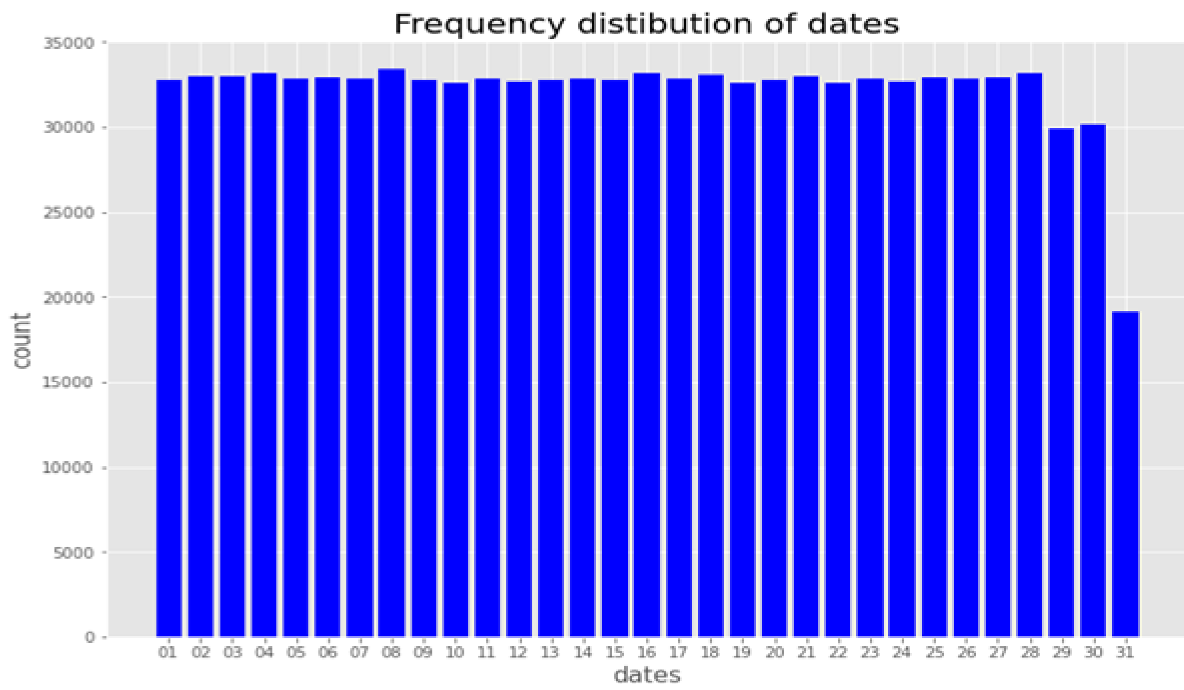
Field Name: date

Description:“date” is a date variable, including the application dates.

	month	count	percentage
0	01	85199	0.085199
11	02	76792	0.076792
4	03	84871	0.084871
7	04	82515	0.082515
1	05	85083	0.085083
9	06	82035	0.082035
2	07	84943	0.084943
6	08	84830	0.084830
8	09	82374	0.082374
5	10	84865	0.084865
10	11	81602	0.081602
3	12	84891	0.084891

Plot the distribution of the months and dates as below, the overall frequency is consistent.





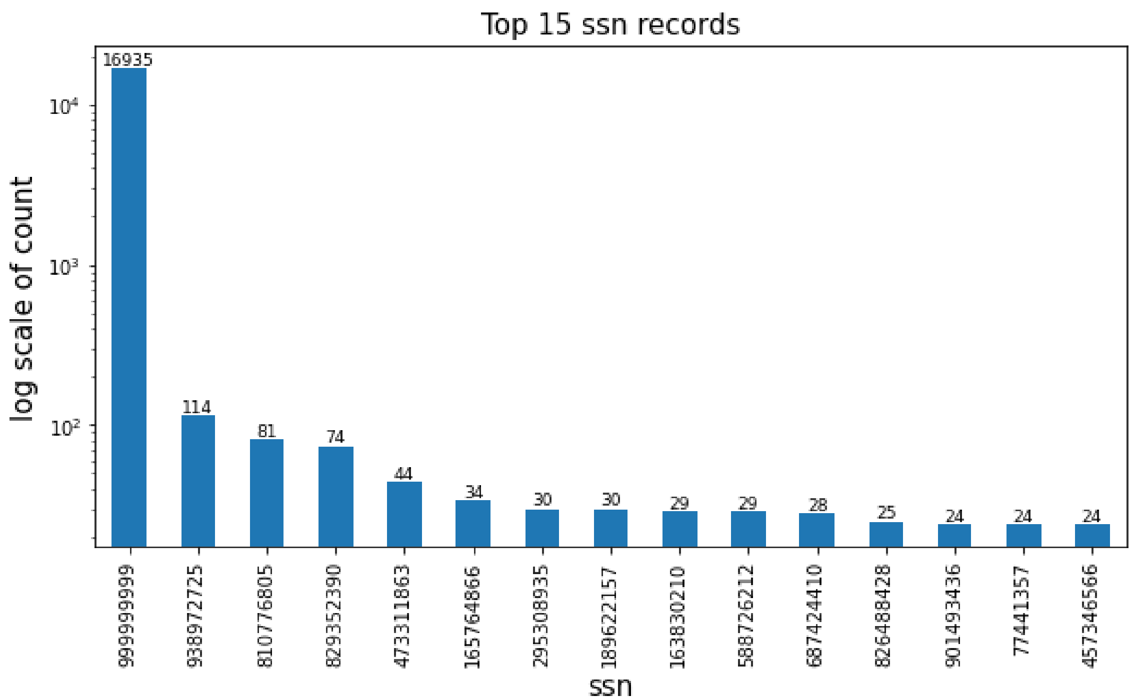
Field 2

Field Name: ssn

Description:“ssn” is a categorical variable, indicating the applicant’s ssn number.

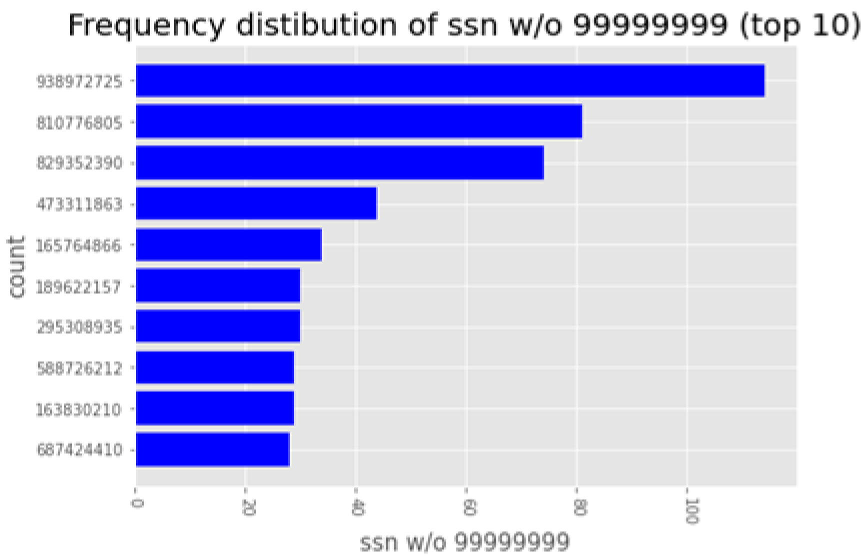
The top 10 most frequent records are listed below:

	ssn	count	percentage
0	999999999	16935	0.016935
1	938972725	114	0.000114
2	810776805	81	0.000081
3	829352390	74	0.000074
4	473311863	44	0.000044
5	165764866	34	0.000034
6	189622157	30	0.000030
7	295308935	30	0.000030
8	588726212	29	0.000029
9	163830210	29	0.000029
10	687424410	28	0.000028



Since the number of 999999999 takes most of the count and it means the record is frivolous, we plot the graph which exclude the record of 99999999.

The distribution of the top 10 records (excluding the most frequent record “999999999”):



Field 3

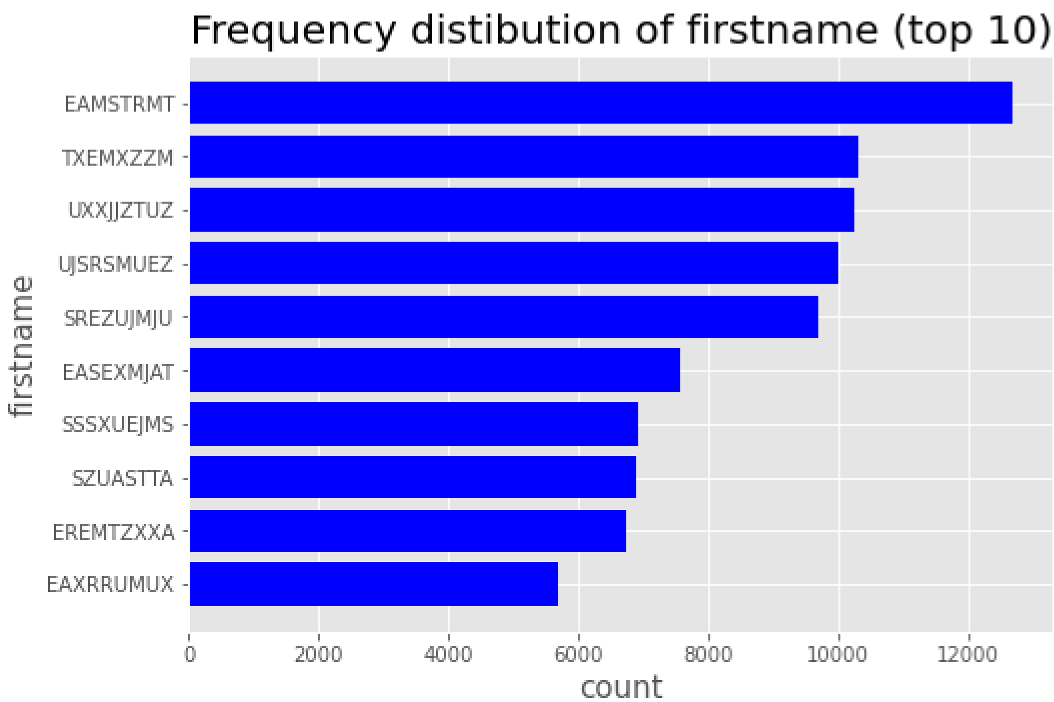
Field Name: firstname

Description:

"firstname" is a categorical variable, indicating the applicant's first name.

	firstname	count	percentage
0	EAMSTRMT	12658	0.012658
1	TXEMXZZM	10297	0.010297
2	UXXJJZTUZ	10235	0.010235
3	UJSRSMUEZ	9994	0.009994
4	SREZUJMJU	9688	0.009688
5	EASEXMJAT	7576	0.007576
6	SSSXUEJMS	6923	0.006923
7	SZUASTTA	6878	0.006878
8	EREMTZXXA	6717	0.006717
9	EAXRRUMUX	5686	0.005686

The top 10 most frequent records are listed below:



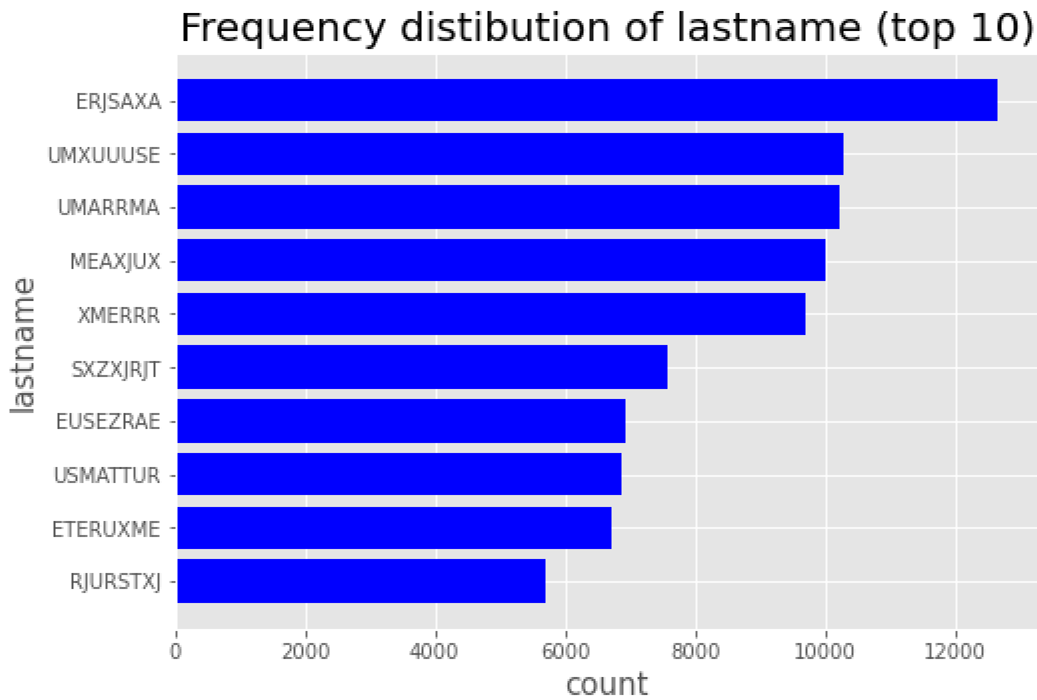
Field 4

Field Name: lastname

Description: “lastname” is a categorical variable, indicating the applicant’s last name.

The top 10 most frequent records are listed below:

	lastname	count	percentage
0	ERJSAXA	8580	0.008580
1	UMXUUUSE	7156	0.007156
2	UMARRMA	6832	0.006832
3	MEAXJUX	5492	0.005492
4	XMERRR	5451	0.005451
5	SXZXJRJT	4340	0.004340
6	EUSEZRAE	4173	0.004173
7	USMATTUR	4036	0.004036
8	ETERUXME	3762	0.003762
9	RJURSTXJ	3575	0.003575



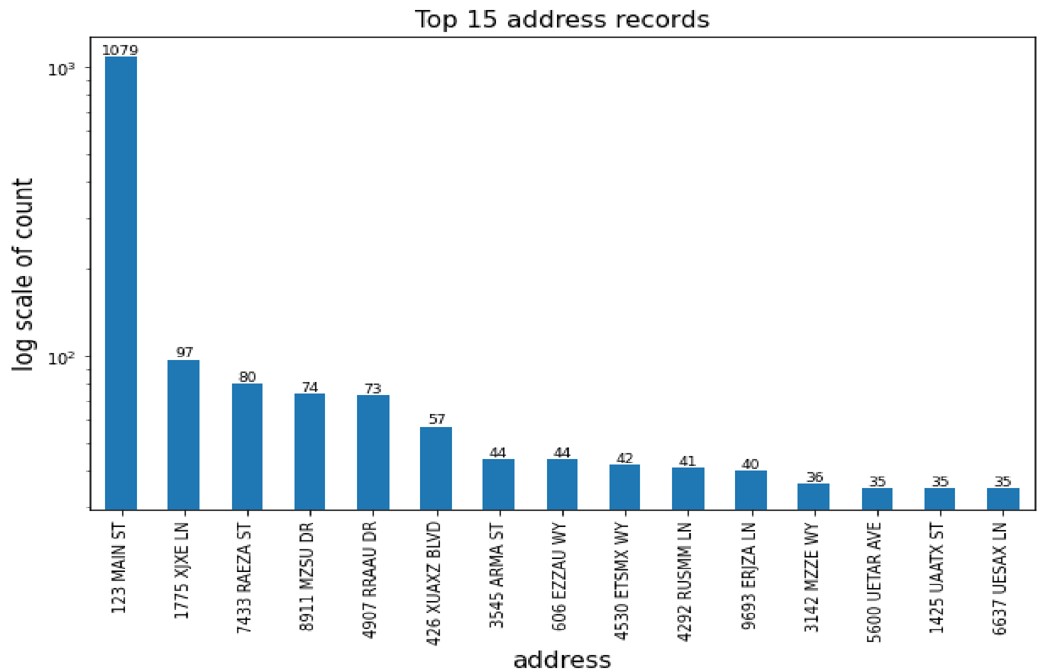
Field 5

Field Name: address

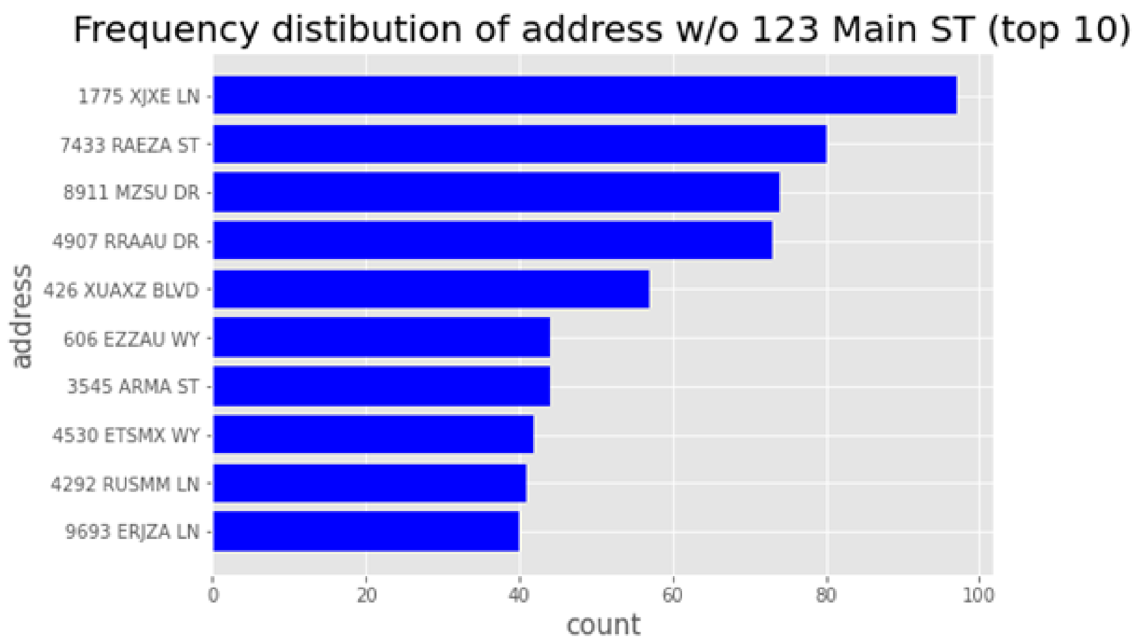
Description: “address” is a categorical variable that contains inputs from applicants of their address.

The top 10 most frequent records are listed below:

	address	count	percentage
0	123 MAIN ST	1079	0.001079
1	1775 XJXE LN	97	0.000097
2	7433 RAEZA ST	80	0.000080
3	8911 MZSU DR	74	0.000074
4	4907 RRAAU DR	73	0.000073
5	426 XUAXZ BLVD	57	0.000057
6	606 EZZAU WY	44	0.000044
7	3545 ARMA ST	44	0.000044
8	4530 ETSMX WY	42	0.000042
9	4292 RUSMM LN	41	0.000041
10	9693 ERJZA LN	40	0.000040



The distribution of the top 10 records (excluding the most frequent record “123 MAIN ST”):



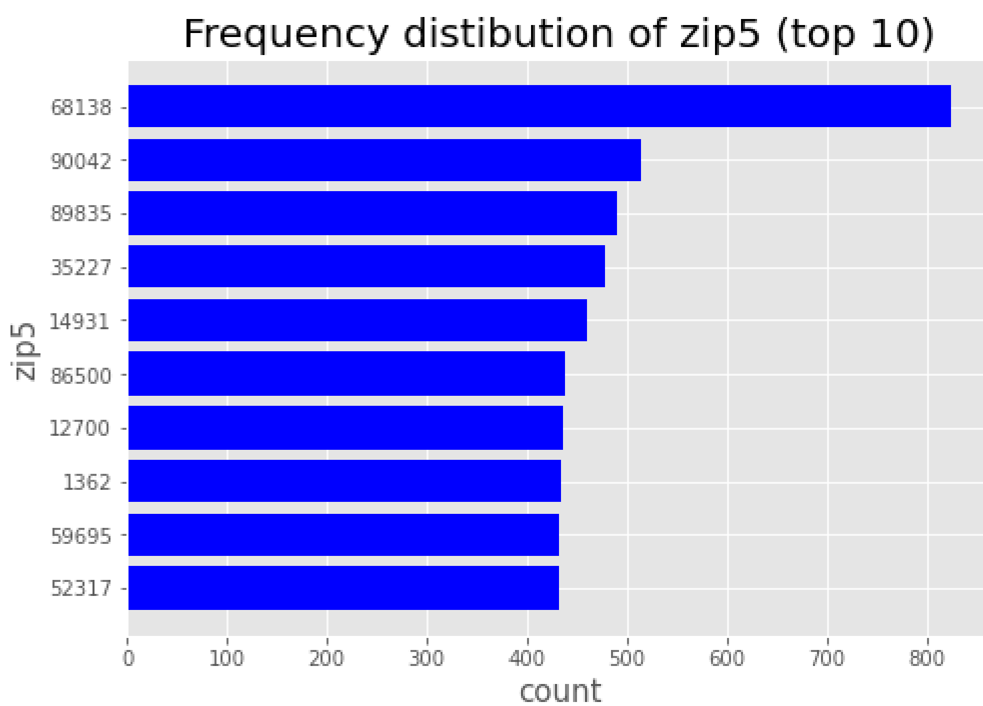
Field 6

Field Name: zip5

Description: “zip5” is a categorical variable that contains inputs from applicants of their zip codes.

	zip5	count	percentage
0	68138	823	0.00823
1	90042	514	0.00514
2	89835	489	0.00489
3	35227	478	0.00478
4	14931	459	0.00459
5	86500	438	0.00438
6	12700	436	0.00436
7	1362	434	0.00434
8	59695	432	0.00432
9	52317	432	0.00432

The distribution of the top 10 records:



Field 7

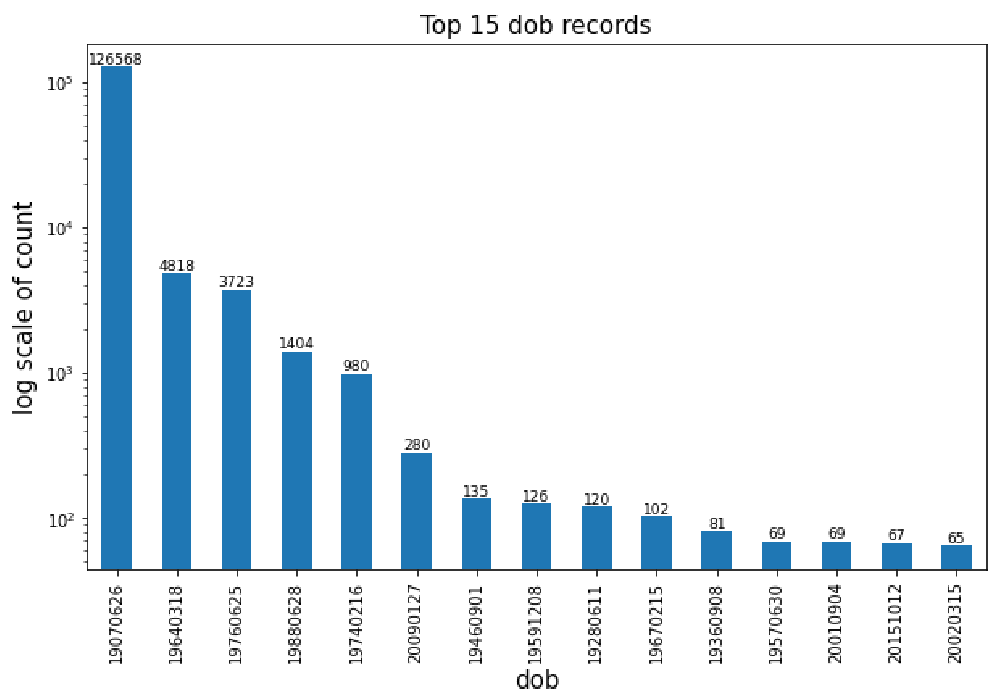
Field Name: dob

Description:

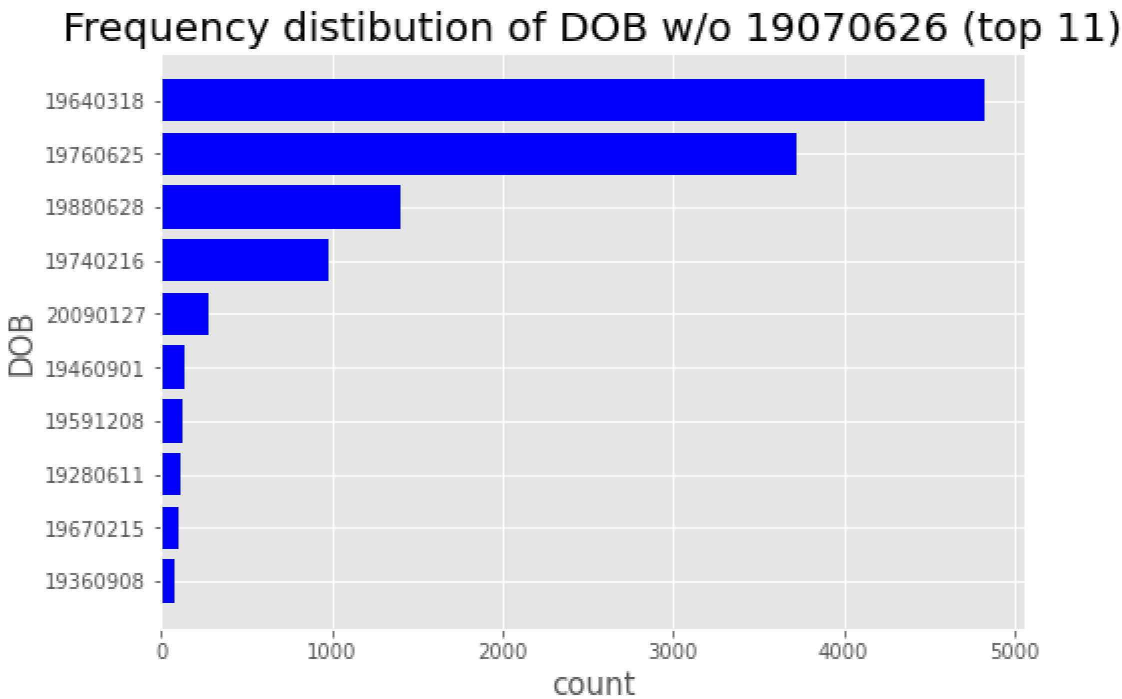
"dob" is a date variable that contains inputs from applicants of their date of birth.

The top 10 most frequent records are listed below:

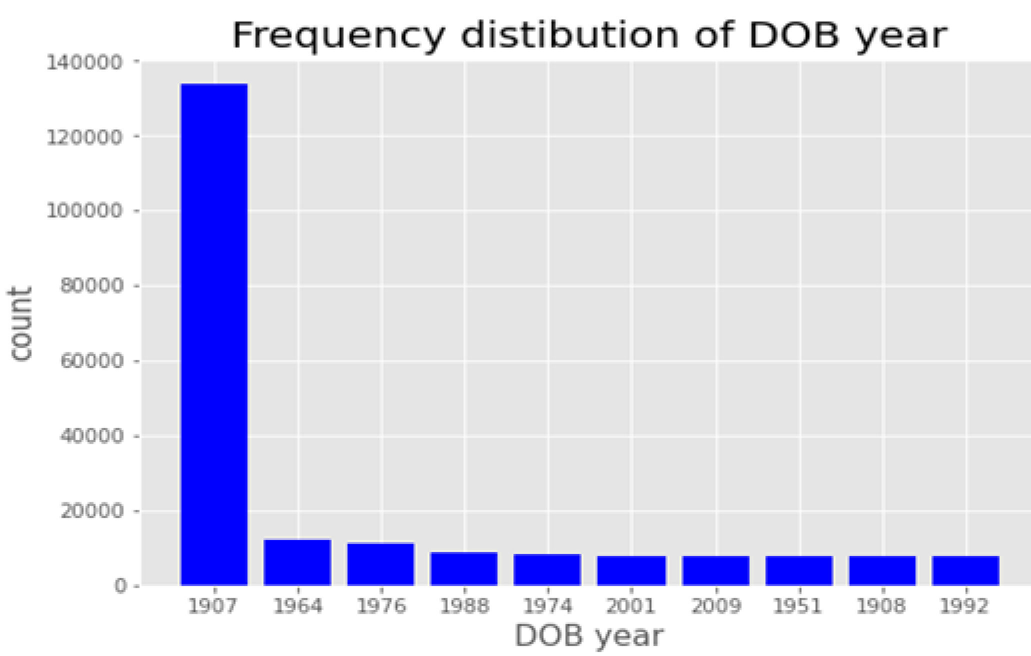
	dob	count	percentage
0	19070626	126568	0.126568
1	19640318	4818	0.004818
2	19760625	3723	0.003723
3	19880628	1404	0.001404
4	19740216	980	0.000980
5	20090127	280	0.000280
6	19460901	135	0.000135
7	19591208	126	0.000126
8	19280611	120	0.000120
9	19670215	102	0.000102
10	19360908	81	0.000081



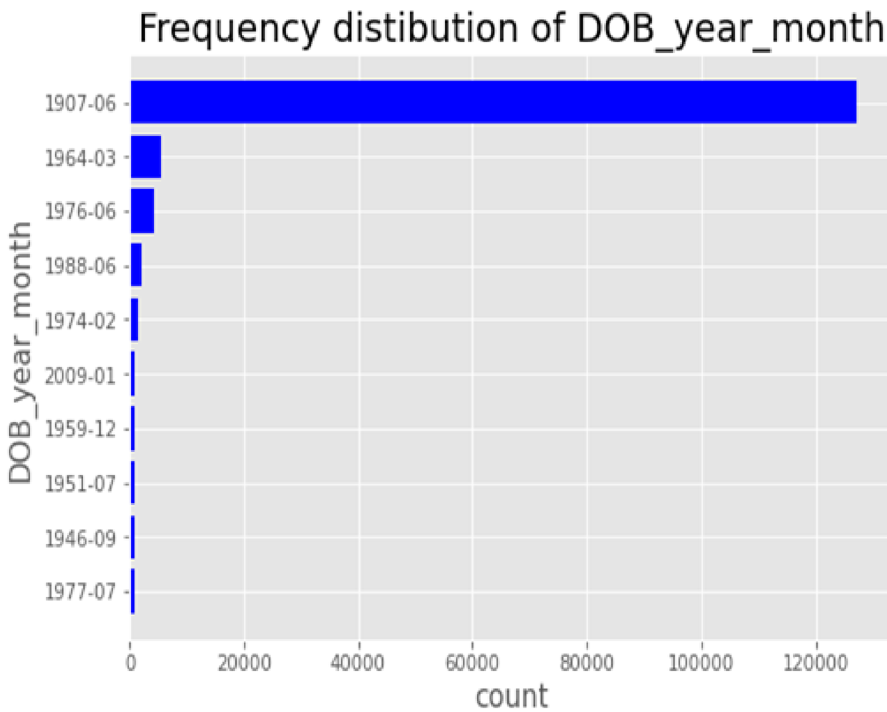
The distribution of the top 10 records (excluding the most frequent record "19070626"):



Also, the distribution of different DOB years is plotted below.



The distribution of different DOB years and months is plotted below



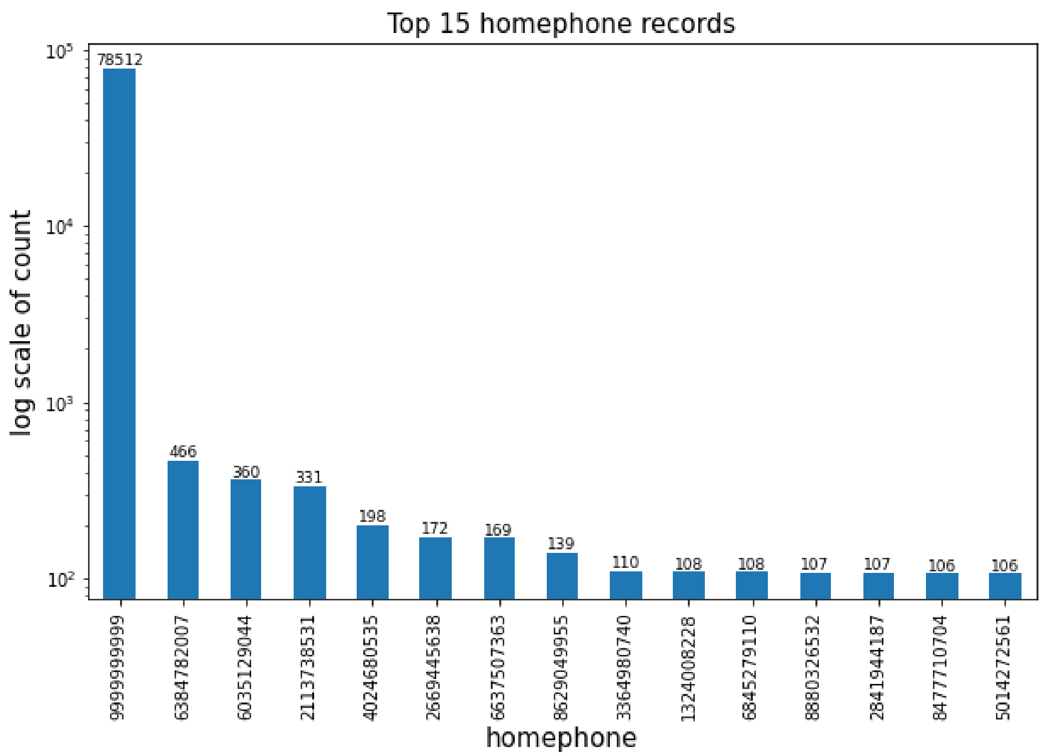
Field 8

Field Name: homephone

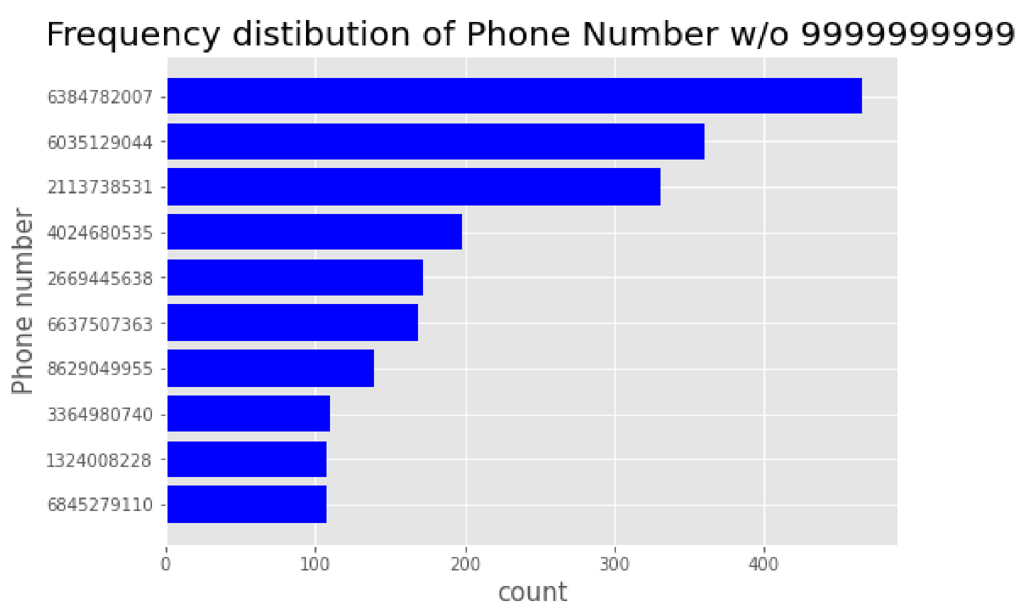
Description: “homephone” is a categorical variable that contains home phone number of each applicant.

The top 10 most frequent records are listed below:

	phone	count	percentage
0	9999999999	78512	0.078512
1	6384782007	466	0.000466
2	6035129044	360	0.000360
3	2113738531	331	0.000331
4	4024680535	198	0.000198
5	2669445638	172	0.000172
6	6637507363	169	0.000169
7	8629049955	139	0.000139
8	3364980740	110	0.000110
9	1324008228	108	0.000108
10	6845279110	108	0.000108



The distribution of the top 10 records (excluding the most frequent record “9999999999”):

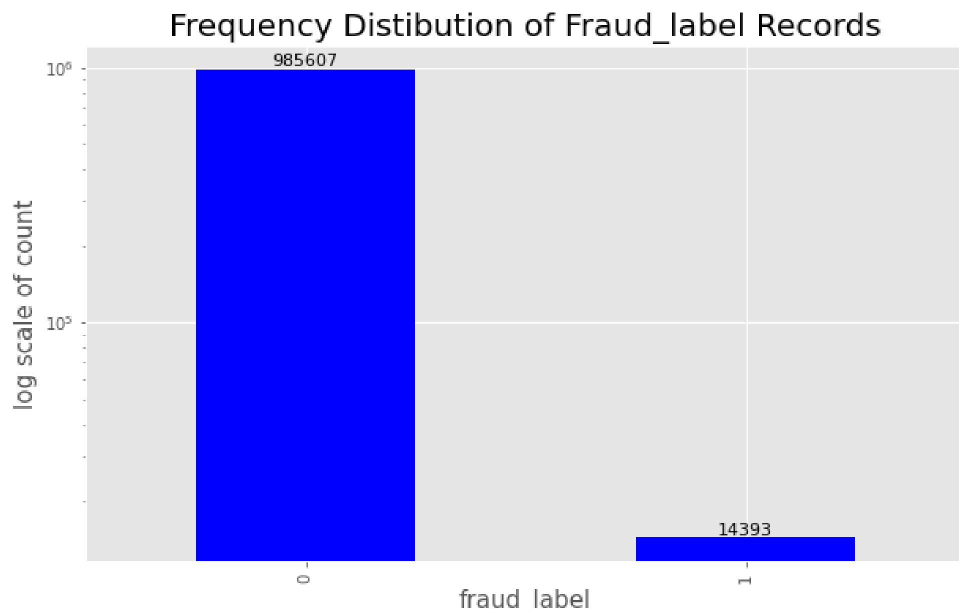


Field 9

Field Name: Fraud Label

Description:“fraud_label” is a categorical variable whether the applicant is fraud.

	fraud	count	percentage
0	0	985607	0.985607
1	1	14393	0.014393



Candidate Variables

Velocity Candidate Variables			
1	ssn_count_0	85	ssn_lastname_count_0
2	ssn_count_1	86	ssn_lastname_count_1
3	ssn_count_3	87	ssn_lastname_count_3
4	ssn_count_7	88	ssn_lastname_count_7
5	ssn_count_14	89	ssn_lastname_count_14
6	ssn_count_30	90	ssn_lastname_count_30
7	address_count_0	91	ssn_address_count_0
8	address_count_1	92	ssn_address_count_1
9	address_count_3	93	ssn_address_count_3
10	address_count_7	94	ssn_address_count_7
11	address_count_14	95	ssn_address_count_14
12	address_count_30	96	ssn_address_count_30
13	dob_count_0	97	ssn_zip5_count_0
14	dob_count_1	98	ssn_zip5_count_1
15	dob_count_3	99	ssn_zip5_count_3

16 dob_count_7
17 dob_count_14
18 dob_count_30
19 homephone_count_0
20 homephone_count_1
21 homephone_count_3
22 homephone_count_7
23 homephone_count_14
24 homephone_count_30
25 name_count_0
26 name_count_1
27 name_count_3
28 name_count_7
29 name_count_14
30 name_count_30
31 fulladdress_count_0
32 fulladdress_count_1
33 fulladdress_count_3
34 fulladdress_count_7
35 fulladdress_count_14
36 fulladdress_count_30
37 name_dob_count_0
38 name_dob_count_1
39 name_dob_count_3
40 name_dob_count_7
41 name_dob_count_14
42 name_dob_count_30
43 name_fulladdress_count_0
44 name_fulladdress_count_1
45 name_fulladdress_count_3
46 name_fulladdress_count_7
47 name_fulladdress_count_14
48 name_fulladdress_count_30
49 name_homephone_count_0
50 name_homephone_count_1
51 name_homephone_count_3
52 name_homephone_count_7
53 name_homephone_count_14
54 name_homephone_count_30
55 fulladdress_dob_count_0

100 ssn_zip5_count_7
101 ssn_zip5_count_14
102 ssn_zip5_count_30
103 ssn_dob_count_0
104 ssn_dob_count_1
105 ssn_dob_count_3
106 ssn_dob_count_7
107 ssn_dob_count_14
108 ssn_dob_count_30
109 ssn_homephone_count_0
110 ssn_homephone_count_1
111 ssn_homephone_count_3
112 ssn_homephone_count_7
113 ssn_homephone_count_14
114 ssn_homephone_count_30
115 ssn_name_count_0
116 ssn_name_count_1
117 ssn_name_count_3
118 ssn_name_count_7
119 ssn_name_count_14
120 ssn_name_count_30
121 ssn_fulladdress_count_0
122 ssn_fulladdress_count_1
123 ssn_fulladdress_count_3
124 ssn_fulladdress_count_7
125 ssn_fulladdress_count_14
126 ssn_fulladdress_count_30
127 ssn_name_dob_count_0
128 ssn_name_dob_count_1
129 ssn_name_dob_count_3
130 ssn_name_dob_count_7
131 ssn_name_dob_count_14
132 ssn_name_dob_count_30
133 ssn_name_fulladdress_count_0
134 ssn_name_fulladdress_count_1
135 ssn_name_fulladdress_count_3
136 ssn_name_fulladdress_count_7
137 ssn_name_fulladdress_count_14
138 ssn_name_fulladdress_count_30
139 ssn_name_homephone_count_0

56	fulladdress_dob_count_1	140	ssn_name_homephone_count_1
57	fulladdress_dob_count_3	141	ssn_name_homephone_count_3
58	fulladdress_dob_count_7	142	ssn_name_homephone_count_7
59	fulladdress_dob_count_14	143	ssn_name_homephone_count_14
60	fulladdress_dob_count_30	144	ssn_name_homephone_count_30
61	fulladdress_homephone_count_0	145	ssn_fulladdress_dob_count_0
62	fulladdress_homephone_count_1	146	ssn_fulladdress_dob_count_1
63	fulladdress_homephone_count_3	147	ssn_fulladdress_dob_count_3
64	fulladdress_homephone_count_7	148	ssn_fulladdress_dob_count_7
65	fulladdress_homephone_count_14	149	ssn_fulladdress_dob_count_14
66	fulladdress_homephone_count_30	150	ssn_fulladdress_dob_count_30
67	dob_homephone_count_0	151	ssn_fulladdress_homephone_count_0
68	dob_homephone_count_1	152	ssn_fulladdress_homephone_count_1
69	dob_homephone_count_3	153	ssn_fulladdress_homephone_count_3
70	dob_homephone_count_7	154	ssn_fulladdress_homephone_count_7
71	dob_homephone_count_14	155	ssn_fulladdress_homephone_count_14
72	dob_homephone_count_30	156	ssn_fulladdress_homephone_count_30
73	homephone_name_dob_count_0	157	ssn_dob_homephone_count_0
74	homephone_name_dob_count_1	158	ssn_dob_homephone_count_1
75	homephone_name_dob_count_3	159	ssn_dob_homephone_count_3
76	homephone_name_dob_count_7	160	ssn_dob_homephone_count_7
77	homephone_name_dob_count_14	161	ssn_dob_homephone_count_14
78	homephone_name_dob_count_30	162	ssn_dob_homephone_count_30
79	ssn_firstname_count_0	163	ssn_homephone_name_dob_count_0
80	ssn_firstname_count_1	164	ssn_homephone_name_dob_count_1
81	ssn_firstname_count_3	165	ssn_homephone_name_dob_count_3
82	ssn_firstname_count_7	166	ssn_homephone_name_dob_count_7
83	ssn_firstname_count_14	167	ssn_homephone_name_dob_count_14
84	ssn_firstname_count_30	168	ssn_homephone_name_dob_count_30
Relative Velocity Candidate Variables			
169	ssn_count_0_by_3	276	ssn_firstname_count_0_by_30
170	ssn_count_0_by_7	277	ssn_firstname_count_1_by_3
171	ssn_count_0_by_14	278	ssn_firstname_count_1_by_7
172	ssn_count_0_by_30	279	ssn_firstname_count_1_by_14
173	ssn_count_1_by_3	280	ssn_firstname_count_1_by_30
174	ssn_count_1_by_7	281	ssn_lastname_count_0_by_3
175	ssn_count_1_by_14	282	ssn_lastname_count_0_by_7
176	ssn_count_1_by_30	283	ssn_lastname_count_0_by_14
177	address_count_0_by_3	284	ssn_lastname_count_0_by_30
178	address_count_0_by_7	285	ssn_lastname_count_1_by_3

179 address_count_0_by_14
180 address_count_0_by_30
181 address_count_1_by_3
182 address_count_1_by_7
183 address_count_1_by_14
184 address_count_1_by_30
185 dob_count_0_by_3
186 dob_count_0_by_7
187 dob_count_0_by_14
188 dob_count_0_by_30
189 dob_count_1_by_3
190 dob_count_1_by_7
191 dob_count_1_by_14
192 dob_count_1_by_30
193 homephone_count_0_by_3
194 homephone_count_0_by_7
195 homephone_count_0_by_14
196 homephone_count_0_by_30
197 homephone_count_1_by_3
198 homephone_count_1_by_7
199 homephone_count_1_by_14
200 homephone_count_1_by_30
201 name_count_0_by_3
202 name_count_0_by_7
203 name_count_0_by_14
204 name_count_0_by_30
205 name_count_1_by_3
206 name_count_1_by_7
207 name_count_1_by_14
208 name_count_1_by_30
209 fulladdress_count_0_by_3
210 fulladdress_count_0_by_7
211 fulladdress_count_0_by_14
212 fulladdress_count_0_by_30
213 fulladdress_count_1_by_3
214 fulladdress_count_1_by_7
215 fulladdress_count_1_by_14
216 fulladdress_count_1_by_30
217 name_dob_count_0_by_3
218 name_dob_count_0_by_7
286 ssn_lastname_count_1_by_7
287 ssn_lastname_count_1_by_14
288 ssn_lastname_count_1_by_30
289 ssn_address_count_0_by_3
290 ssn_address_count_0_by_7
291 ssn_address_count_0_by_14
292 ssn_address_count_0_by_30
293 ssn_address_count_1_by_3
294 ssn_address_count_1_by_7
295 ssn_address_count_1_by_14
296 ssn_address_count_1_by_30
297 ssn_zip5_count_0_by_3
298 ssn_zip5_count_0_by_7
299 ssn_zip5_count_0_by_14
300 ssn_zip5_count_0_by_30
301 ssn_zip5_count_1_by_3
302 ssn_zip5_count_1_by_7
303 ssn_zip5_count_1_by_14
304 ssn_zip5_count_1_by_30
305 ssn_dob_count_0_by_3
306 ssn_dob_count_0_by_7
307 ssn_dob_count_0_by_14
308 ssn_dob_count_0_by_30
309 ssn_dob_count_1_by_3
310 ssn_dob_count_1_by_7
311 ssn_dob_count_1_by_14
312 ssn_dob_count_1_by_30
313 ssn_homephone_count_0_by_3
314 ssn_homephone_count_0_by_7
315 ssn_homephone_count_0_by_14
316 ssn_homephone_count_0_by_30
317 ssn_homephone_count_1_by_3
318 ssn_homephone_count_1_by_7
319 ssn_homephone_count_1_by_14
320 ssn_homephone_count_1_by_30
321 ssn_name_count_0_by_3
322 ssn_name_count_0_by_7
323 ssn_name_count_0_by_14
324 ssn_name_count_0_by_30
325 ssn_name_count_1_by_3

219 name_dob_count_0_by_14
220 name_dob_count_0_by_30
221 name_dob_count_1_by_3
222 name_dob_count_1_by_7
223 name_dob_count_1_by_14
224 name_dob_count_1_by_30
225 name_fulladdress_count_0_by_3
226 name_fulladdress_count_0_by_7
227 name_fulladdress_count_0_by_14
228 name_fulladdress_count_0_by_30
229 name_fulladdress_count_1_by_3
230 name_fulladdress_count_1_by_7
231 name_fulladdress_count_1_by_14
232 name_fulladdress_count_1_by_30
233 name_homephone_count_0_by_3
234 name_homephone_count_0_by_7
235 name_homephone_count_0_by_14
236 name_homephone_count_0_by_30
237 name_homephone_count_1_by_3
238 name_homephone_count_1_by_7
239 name_homephone_count_1_by_14
240 name_homephone_count_1_by_30
241 fulladdress_dob_count_0_by_3
242 fulladdress_dob_count_0_by_7
243 fulladdress_dob_count_0_by_14
244 fulladdress_dob_count_0_by_30
245 fulladdress_dob_count_1_by_3
246 fulladdress_dob_count_1_by_7
247 fulladdress_dob_count_1_by_14
248 fulladdress_dob_count_1_by_30
249 fulladdress_homephone_count_0_by_3
250 fulladdress_homephone_count_0_by_7
251 fulladdress_homephone_count_0_by_14
252 fulladdress_homephone_count_0_by_30
253 fulladdress_homephone_count_1_by_3
254 fulladdress_homephone_count_1_by_7
255 fulladdress_homephone_count_1_by_14
256 fulladdress_homephone_count_1_by_30
257 dob_homephone_count_0_by_3
258 dob_homephone_count_0_by_7
326 ssn_name_count_1_by_7
327 ssn_name_count_1_by_14
328 ssn_name_count_1_by_30
329 ssn_fulladdress_count_0_by_3
330 ssn_fulladdress_count_0_by_7
331 ssn_fulladdress_count_0_by_14
332 ssn_fulladdress_count_0_by_30
333 ssn_fulladdress_count_1_by_3
334 ssn_fulladdress_count_1_by_7
335 ssn_fulladdress_count_1_by_14
336 ssn_fulladdress_count_1_by_30
337 ssn_name_dob_count_0_by_3
338 ssn_name_dob_count_0_by_7
339 ssn_name_dob_count_0_by_14
340 ssn_name_dob_count_0_by_30
341 ssn_name_dob_count_1_by_3
342 ssn_name_dob_count_1_by_7
343 ssn_name_dob_count_1_by_14
344 ssn_name_dob_count_1_by_30
345 ssn_name_fulladdress_count_0_by_3
346 ssn_name_fulladdress_count_0_by_7
347 ssn_name_fulladdress_count_0_by_14
348 ssn_name_fulladdress_count_0_by_30
349 ssn_name_fulladdress_count_1_by_3
350 ssn_name_fulladdress_count_1_by_7
351 ssn_name_fulladdress_count_1_by_14
352 ssn_name_fulladdress_count_1_by_30
353 ssn_name_homephone_count_0_by_3
354 ssn_name_homephone_count_0_by_7
355 ssn_name_homephone_count_0_by_14
356 ssn_name_homephone_count_0_by_30
357 ssn_name_homephone_count_1_by_3
358 ssn_name_homephone_count_1_by_7
359 ssn_name_homephone_count_1_by_14
360 ssn_name_homephone_count_1_by_30
361 ssn_fulladdress_dob_count_0_by_3
362 ssn_fulladdress_dob_count_0_by_7
363 ssn_fulladdress_dob_count_0_by_14
364 ssn_fulladdress_dob_count_0_by_30
365 ssn_fulladdress_dob_count_1_by_3

259	dob_homephone_count_0_by_14	366	ssn_fulladdress_dob_count_1_by_7
260	dob_homephone_count_0_by_30	367	ssn_fulladdress_dob_count_1_by_14
261	dob_homephone_count_1_by_3	368	ssn_fulladdress_dob_count_1_by_30
262	dob_homephone_count_1_by_7	369	ssn_fulladdress_homephone_count_0_by_3
263	dob_homephone_count_1_by_14	370	ssn_fulladdress_homephone_count_0_by_7
264	dob_homephone_count_1_by_30	371	ssn_fulladdress_homephone_count_0_by_14
265	homephone_name_dob_count_0_by_3	372	ssn_fulladdress_homephone_count_0_by_30
266	homephone_name_dob_count_0_by_7	373	ssn_fulladdress_homephone_count_1_by_3
267	homephone_name_dob_count_0_by_14	374	ssn_fulladdress_homephone_count_1_by_7
268	homephone_name_dob_count_0_by_30	375	ssn_fulladdress_homephone_count_1_by_14
269	homephone_name_dob_count_1_by_3	376	ssn_fulladdress_homephone_count_1_by_30
270	homephone_name_dob_count_1_by_7	377	ssn_dob_homephone_count_0_by_3
271	homephone_name_dob_count_1_by_14	378	ssn_dob_homephone_count_0_by_7
272	homephone_name_dob_count_1_by_30	379	ssn_dob_homephone_count_0_by_14
273	ssn_firstname_count_0_by_3	380	ssn_dob_homephone_count_0_by_30
274	ssn_firstname_count_0_by_7	381	ssn_dob_homephone_count_1_by_3
275	ssn_firstname_count_0_by_14	382	ssn_dob_homephone_count_1_by_7
Day Since Candidate Variables			
383	ssn_day_since	397	ssn_lastname_day_since
384	address_day_since	398	ssn_address_day_since
385	dob_day_since	399	ssn_zip5_day_since
386	homephone_day_since	400	ssn_dob_day_since
387	name_day_since	401	ssn_homephone_day_since
388	fulladdress_day_since	402	ssn_name_day_since
389	name_dob_day_since	403	ssn_fulladdress_day_since
390	name_fulladdress_day_since	404	ssn_name_dob_day_since
391	name_homephone_day_since	405	ssn_name_fulladdress_day_since
392	fulladdress_dob_day_since	406	ssn_name_homephone_day_since
393	fulladdress_homephone_day_since	407	ssn_fulladdress_dob_day_since
394	dob_homephone_day_since	408	ssn_fulladdress_homephone_day_since
395	homephone_name_dob_day_since	409	ssn_dob_homephone_day_since
396	ssn_firstname_day_since	410	ssn_homephone_name_dob_day_since

Reference:

Team, V. (2021, February 25). Machine learning algorithms: What is a neural network? Retrieved March 25, 2021, from https://www.verypossible.com/insights/machine-learning-algorithms-what-is-a-neural-network

Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016. https:web.archive.org/web/20160417030218/http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf

Brownlee, J. (2020, August 14). A gentle introduction to the gradient boosting algorithm for machine learning. Retrieved March 25, 2021, from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Z_ai (2020, February 8). Logistic Regression Explained. Retrieved March 23, 2021 from https://towardsdatascience.com/logistic-regression-explained-9ee73cede081