Identity Fraud Detection Report

Team 5

Rui Xin Wu, Zhuolin Ouyang, Xiaoxuan Feng, Feifei Shao, Zihao Geng, Zimeng Cao, Zijian Wang

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report examines the Application dataset to detect the fraud activity using supervised machine learning methods. Out team used R as the major tool, “Kolmogorov-Smirnov” as the main feature selection method, and Support Vector Machine, Gradient Boosting Decision Trees, Logistic Regression and Neural Network as featured algorithms to build the predictive model.

The steps we took were:

1. Variable construction and data cleaning
2. Feature selection using “Kolmogorov-Smirnov”
3. Fraud detection model building using SVM, Gradient Boosting Decision Trees, Logistic Regression and Neural Network

The original dataset contains 94,866 records from 1/1/2016 to 12/31/2016 with details about the applicants’ personal information including name, SSN, date of birth, home phone, address, and zip code.

First, we built 160 expert variables with looking at similar characteristics or/and time windows.

Then, we separated the dataset into training, testing and out-of-time dataset, and performed feature selection on the training set. We ranked all variables based on the “Kolmogorov-Smirnov” score and selected 30 most important variables. We used the training dataset to train our model first. Then, we tested the model on both training dataset and testing dataset. After that, we built the final model using both training and testing dataset and tested the model on the out-of-time dataset to calculate and compare the final fraud detection rate (FDR).

Using the 30 variables and supervised machine learning algorithms, we found that Gradient Boosting Decision Trees has the best performance, with FDR of 16.94% at 10% of the population. More detailed model performance information is demonstrated in the table below.

	FDR @ 10%		
Model	**Training**	**Testing**	**Out of Time**
SVM	12.80%	11.29%	13.79%
Gradient Boosting	20.35%	20.31%	16.94%
Neural Net	19.89%	20.05%	16.57%
Logistic Regression	19.94%	20.19%	15.68%

DESCRIPTION OF DATA

Here we provide an overall description of the dataset we analyzed on in this project.

Dataset Name: Applications Data
File Name: Applications.csv

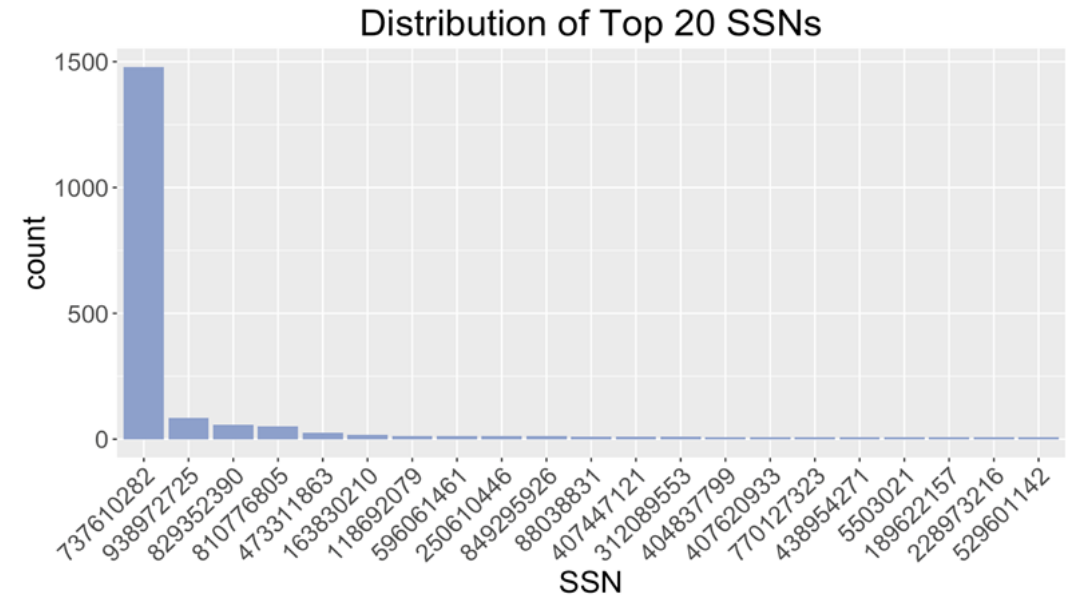
Dataset Overview:

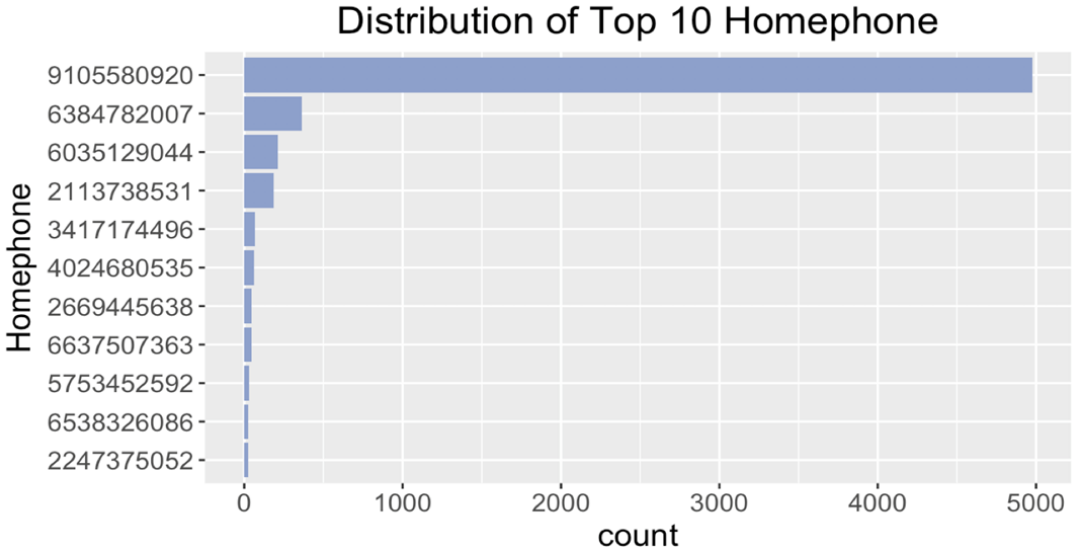
- Category: Identity Fraud
- Dimensions: 94,866 records and 10 variables
    - Categorical Variables: 8
    - Numeric Variables: 0
    - Date Variables: 2
- Features Names: Record, Date, SSN, Firstname, Lastname, Address, Zip5, DOB, Homephone
- Response: fraud

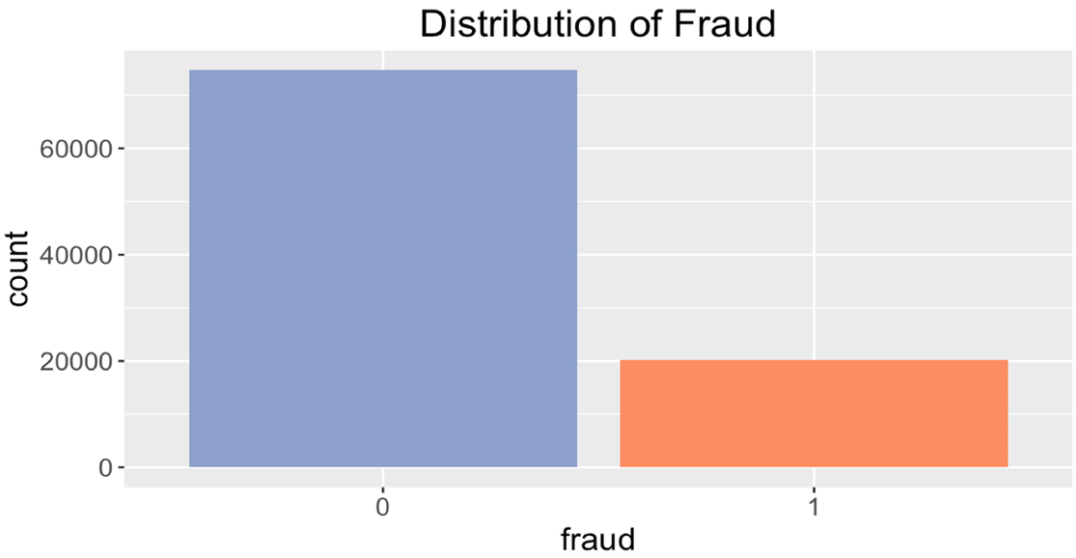
Most important variables in consideration:

Variables	Type	Description
Record	categorical	Unique identifier of each record
Date	date	Date of the application
SSN	categorical	The identifier of the applicant
Firstname	categorical	First name of the applicant
Lastname	categorical	Last name of the applicant
Address	categorical	Address of the applicant
Zip5	categorical	Zip code of the applicant
DOB	date	Date of birth of the applicant
Homephone	categorical	Homephone of the applicant
Fraud	categorical	Whether the application is fraud or not.

Below are the distributions of variables “ssn”, “homephone”, “fraud”: (the full data quality report that includes detailed description and distribution of each variable is attached to the appendix)


Distribution of Top 20 SSNs
count
1500
1000
500
0
737610282
938972725
829352390
810776805
473311863
163830210
118692079
596061461
250610446
849295926
88038831
407447121
312089553
404837799
407620933
770127323
438954271
5503021
189622157
228973216
529601142
SSN


Distribution of Top 10 Homephone
Homephone
9105580920
6384782007
6035129044
2113738531
3417174496
4024680535
2669445638
6637507363
5753452592
6538326086
2247375052
0
1000
2000
3000
4000
5000
count


Distribution of Fraud
count
60000
40000
20000
0
0
1
fraud

VARIABLE CREATION

We created altogether 160 variables in this project. We considered how a fraudster would do when committing fraud crime and the main logic is stated as follows:

- Fraud records tend to have similar characteristics or similar combination of characteristics.
- The fraudster tends to commit fraud in a short time window. We set up 5 groups of time windows including the same day, 3 days, 7 days, 14 days and 30 days.

With this fundamental logic, we derived expert variables using the following two methods:

- **Count records having similar characteristics or similar combination of characteristics as the current record appeared within the last one or several days.**

  To address this aspect, we look back a certain period of time and count the number of previous records that have the same attribute as the current record, and whether the previous record appears with a different combination of attributes.

  For example, variable ***"same_ssn_diff_address_3"*** means to check how many records has the same SSN but different address as the current record within the last 3 days.

- **Calculate the number of days passed since the last occurrence of the same attributes as the current record.**

  To address this aspect, we check the time of the last appearance of the same attribute as the current record.

  For example, variable ***"last_ssn"*** means to count the number of days passed since the last appearance of the same SSN as the current record.

The following chart provides a detailed description of all the 160 variables we created in this project:

No.	Variable Name	Variable Description
1	same_ssn_1	The number of same ssn as the current record appeared on the same day. It only counts the records that appear before the current record.
2	same_ssn_3	The number of same ssn as the current record appeared within the last 3 days.
3	same_ssn_7	The number of same ssn as the current record appeared within the last 7 days.
4	same_ssn_14	The number of same ssn as the current record appeared within the last 14 days.
5	same_ssn_30	The number of same ssn as the current record appeared within the last 30 days.

6	same_ssn_diff_address_1	The number of same ssn but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
7	same_ssn_diff_address_3	The number of same ssn but different address as the current record appeared within the last 3 days.
8	same_ssn_diff_address_7	The number of same ssn but different address as the current record appeared within the last 7 days.
9	same_ssn_diff_address_14	The number of same ssn but different address as the current record appeared within the last 14 days.
10	same_ssn_diff_address_30	The number of same ssn but different address as the current record appeared within the last 30 days.
11	same_ssn_diff_phone_1	The number of same ssn but different homephone as the current record appeared on the same day. It only counts the records that appear before the current record.
12	same_ssn_diff_phone_3	The number of same ssn but different homephone as the current record appeared within the last 3 days.
13	same_ssn_diff_phone_7	The number of same ssn but different homephone as the current record appeared within the last 7 days.
14	same_ssn_diff_phone_14	The number of same ssn but different homephone as the current record appeared within the last 14 days.
15	same_ssn_diff_phone_30	The number of same ssn but different homephone as the current record appeared within the last 30 days.
16	same_ssn_diff_bdname_1	The number of same ssn but different birthday, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
17	same_ssn_diff_bdname_3	The number of same ssn but different birthday, first name and last name as the current record appeared within the last 3 days.
18	same_ssn_diff_bdname_7	The number of same ssn but different birthday, first name and last name as the current record appeared within the last 7 days.
19	same_ssn_diff_bdname_14	The number of same ssn but different birthday, first name and last name as the current record appeared within the last 14 days.
20	same_ssn_diff_bdname_30	The number of same ssn but different birthday, first name and last name as the current record appeared within the last 30 days.
21	same_ssn_diff_zipname_1	The number of same ssn but different zip code as the current record appeared on the same day. It only counts the records that appear before the current record.
22	same_ssn_diff_zipname_3	The number of same ssn but different zip code as the current record appeared within the last 3 days.
23	same_ssn_diff_zipname_7	The number of same ssn but different zip code as the current record appeared within the last 7 days.
24	same_ssn_diff_zipname_14	The number of same ssn but different zip code as the current record appeared within the last 14 days.
25	same_ssn_diff_zipname_30	The number of same ssn but different zip code as the current record appeared within the last 30 days.
26	same_name_1	The number of same applicant name as the current record appeared on the same day. It only counts the records that appear before the current record.

27	same_name_3	The number of same applicant name as the current record appeared within the last 3 days.
28	same_name_7	The number of same applicant name as the current record appeared within the last 7 days.
29	same_name_14	The number of same applicant name as the current record appeared within the last 14 days.
30	same_name_30	The number of same applicant name as the current record appeared within the last 30 days.
31	same_name_diff_address_1	The number of same applicant name but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
32	same_name_diff_address_3	The number of same applicant name but different address as the current record appeared within the last 3 days.
33	same_name_diff_address_7	The number of same applicant name but different address as the current record appeared within the last 7 days.
34	same_name_diff_address_14	The number of same applicant name but different address as the current record appeared within the last 14 days.
35	same_name_diff_address_30	The number of same applicant name but different address as the current record appeared within the last 30 days.
36	same_name_diff_zip_1	The number of same applicant name but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
37	same_name_diff_zip_3	The number of same applicant name but different address as the current record appeared within the last 3 days.
38	same_name_diff_zip_7	The number of same applicant name but different address as the current record appeared within the last 7 days.
39	same_name_diff_zip_14	The number of same applicant name but different address as the current record appeared within the last 14 days.
40	same_name_diff_zip_30	The number of same applicant name but different address as the current record appeared within the last 30 days.
41	same_name_diff_bd_1	The number of same applicant name but different birthday as the current record appeared on the same day. It only counts the records that appear before the current record.
42	same_name_diff_bd_3	The number of same applicant name but different birthday as the current record appeared within the last 3 days.
43	same_name_diff_bd_7	The number of same applicant name but different birthday as the current record appeared within the last 7 days.
44	same_name_diff_bd_14	The number of same applicant name but different birthday as the current record appeared within the last 14 days.
45	same_name_diff_bd_30	The number of same applicant name but different birthday as the current record appeared within the last 30 days.
46	same_name_diff_ssn_1	The number of same applicant name but different ssn as the current record appeared on the same day. It only counts the records that appear before the current record.
47	same_name_diff_ssn_3	The number of same applicant name but different ssn as the current record appeared within the last 3 days.
48	same_name_diff_ssn_7	The number of same applicant name but different ssn as the current record appeared within the last 7 days.

49	same_name_diff_ssn_14	The number of same applicant name but different ssn as the current record appeared within the last 14 days.
50	same_name_diff_ssn_30	The number of same applicant name but different ssn as the current record appeared within the last 30 days.
51	same_name_diff_phone_1	The number of same applicant name but different homephone as the current record appeared on the same day. It only counts the records that appear before the current record.
52	same_name_diff_phone_3	The number of same applicant name but different homephone as the current record appeared within the last 3 days.
53	same_name_diff_phone_7	The number of same applicant name but different homephone as the current record appeared within the last 7 days.
54	same_name_diff_phone_14	The number of same applicant name but different homephone as the current record appeared within the last 14 days.
55	same_name_diff_phone_30	The number of same applicant name but different homephone as the current record appeared within the last 30 days.
56	same_address_1	The number of same address as the current record appeared on the same day. It only counts the records that appear before the current record.
57	same_address_3	The number of same address as the current record appeared within the last 3 days.
58	same_address_7	The number of same address as the current record appeared within the last 7 days.
59	same_address_14	The number of same address as the current record appeared within the last 14 days.
60	same_address_30	The number of same address as the current record appeared within the last 30 days.
61	same_address_diff_zip_1	The number of same address but different zip code as the current record appeared on the same day. It only counts the records that appear before the current record.
62	same_address_diff_zip_3	The number of same address but different zip code as the current record appeared within the last 3 days.
63	same_address_diff_zip_7	The number of same address but different zip code as the current record appeared within the last 7 days.
64	same_address_diff_zip_14	The number of same address but different zip code as the current record appeared within the last 14 days.
65	same_address_diff_zip_30	The number of same address but different zip code as the current record appeared within the last 30 days.
66	same_address_diff_phone_1	The number of same address but different homephone code as the current record appeared on the same day. It only counts the records that appear before the current record.
67	same_address_diff_phone_3	The number of same address but different homephone code as the current record appeared within the last 3 days.
68	same_address_diff_phone_7	The number of same address but different homephone code as the current record appeared within the last 7 days.
69	same_address_diff_phone_14	The number of same address but different homephone code as the current record appeared within the last 14 days.
70	same_address_diff_phone_30	The number of same address but different homephone code as the current record appeared within the last 30 days.

71	same_address_diff_bdname_1	The number of same address but different birthday, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
72	same_address_diff_bdname_3	The number of same address but different birthday, first name and last name as the current record appeared within the last 3 days.
73	same_address_diff_bdname_7	The number of same address but different birthday, first name and last name as the current record appeared within the last 7 days.
74	same_address_diff_bdname_14	The number of same address but different birthday, first name and last name as the current record appeared within the last 14 days.
75	same_address_diff_bdname_30	The number of same address but different birthday, first name and last name as the current record appeared within the last 30 days.
76	same_address_diff_ssnname_1	The number of same address but different ssn, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
77	same_address_diff_ssnname_3	The number of same address but different ssn, first name and last name as the current record appeared within the 3 days.
78	same_address_diff_ssnname_7	The number of same address but different ssn, first name and last name as the current record appeared within the 7 days.
79	same_address_diff_ssnname_14	The number of same address but different ssn, first name and last name as the current record appeared within the 14 days.
80	same_address_diff_ssnname_30	The number of same address but different ssn, first name and last name as the current record appeared within the 30 days.
81	same_zip_1	The number of same zip code as the current record appeared on the same day. It only counts the records that appear before the current record.
82	same_zip_3	The number of same zip code as the current record appeared within the 3 days.
83	same_zip_7	The number of same zip code as the current record appeared within the 7 days.
84	same_zip_14	The number of same zip code as the current record appeared within the 14 days.
85	same_zip_30	The number of same zip code as the current record appeared within the 30 days.
86	same_zip_diff_address_1	The number of same zip code but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
87	same_zip_diff_address_3	The number of same zip code but different address as the current record appeared within the 3 days.
88	same_zip_diff_address_7	The number of same zip code but different address as the current record appeared within the 7 days.
89	same_zip_diff_address_14	The number of same zip code but different address as the current record appeared within the 14 days.
90	same_zip_diff_address_30	The number of same zip code but different address as the current record appeared within the 30 days.
91	same_zip_diff_phone_1	The number of same zip code but different homephone as the current record appeared on the same day. It only counts the records that appear before the current record.

92	same_zip_diff_phone_3	The number of same zip code but different homephone as the current record appeared within the last 3 days.
93	same_zip_diff_phone_7	The number of same zip code but different homephone as the current record appeared within the last 7 days.
94	same_zip_diff_phone_14	The number of same zip code but different homephone as the current record appeared within the last 14 days.
95	same_zip_diff_phone_30	The number of same zip code but different homephone as the current record appeared within the last 30 days.
96	same_zip_diff_bdname_1	The number of same zip code but different birthday, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
97	same_zip_diff_bdname_3	The number of same zip code but different birthday, first name and last name as the current record appeared within the last 3 days.
98	same_zip_diff_bdname_7	The number of same zip code but different birthday, first name and last name as the current record appeared within the last 7 days.
99	same_zip_diff_bdname_14	The number of same zip code but different birthday, first name and last name as the current record appeared within the last 14 days.
100	same_zip_diff_bdname_30	The number of same zip code but different birthday, first name and last name as the current record appeared within the last 30 days.
101	same_zip_diff_ssnname_1	The number of same zip code but different ssn and name as the current record appeared on the same day. It only counts the records that appear before the current record.
102	same_zip_diff_ssnname_3	The number of same zip code but different ssn and name as the current record appeared within the last 3 days.
103	same_zip_diff_ssnname_7	The number of same zip code but different ssn and name as the current record appeared within the last 7 days.
104	same_zip_diff_ssnname_14	The number of same zip code but different ssn and name as the current record appeared within the last 14 days.
105	same_zip_diff_ssnname_30	The number of same zip code but different ssn and name as the current record appeared within the last 30 days.
106	same_homephone_1	The number of same homephone as the current record appeared on the same day. It only counts the records that appear before the current record.
107	same_homephone_3	The number of same homephone as the current record appeared within the 3 days.
108	same_homephone_7	The number of same homephone as the current record appeared within the 7 days.
109	same_homephone_14	The number of same homephone as the current record appeared within the 14 days.
110	same_homephone_30	The number of same homephone as the current record appeared within the 30 days.
111	same_homephone_diff_address_1	The number of same homephone but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
112	same_homephone_diff_address_3	The number of same homephone but different address as the current record appeared within the last 3 days.

113	same_homephone_diff_address_7	The number of same homephone but different address as the current record appeared within the last 7 days.
114	same_homephone_diff_address_14	The number of same homephone but different address as the current record appeared within the last 14 days.
115	same_homephone_diff_address_30	The number of same homephone but different address as the current record appeared within the last 30 days.
116	same_homephone_diff_zip_1	The number of same homephone but different zip code as the current record appeared on the same day. It only counts the records that appear before the current record.
117	same_homephone_diff_zip_3	The number of same homephone but different zip code as the current record appeared within the last 3 days.
118	same_homephone_diff_zip_7	The number of same homephone but different zip code as the current record appeared within the last 7 days.
119	same_homephone_diff_zip_14	The number of same homephone but different zip code as the current record appeared within the last 14 days.
120	same_homephone_diff_zip_30	The number of same homephone but different zip code as the current record appeared within the last 30 days.
121	same_homephone_diff_bdname_1	The number of same homephone but different birthday, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
122	same_homephone_diff_bdname_3	The number of same homephone but different birthday, first name and last name as the current record appeared within the last 3 days.
123	same_homephone_diff_bdname_7	The number of same homephone but different birthday, first name and last name as the current record appeared within the last 7 days.
124	same_homephone_diff_bdname_14	The number of same homephone but different birthday, first name and last name as the current record appeared within the last 14 days.
125	same_homephone_diff_bdname_30	The number of same homephone but different birthday, first name and last name as the current record appeared within the last 30 days.
126	same_homephone_diff_ssnname_1	The number of same homephone but different ssn, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
127	same_homephone_diff_ssnname_3	The number of same homephone but different ssn, first name and last name as the current record appeared within the last 3 days.
128	same_homephone_diff_ssnname_7	The number of same homephone but different ssn, first name and last name as the current record appeared within the last 7 days.
129	same_homephone_diff_ssnname_14	The number of same homephone but different ssn, first name and last name as the current record appeared within the last 14 days.
130	same_homephone_diff_ssnname_30	The number of same homephone but different ssn, first name and last name as the current record appeared within the last 30 days.
131	last_ssn	The number of days passed since the appearance of the last same ssn as the current record.

132	last_phone	The number of days passed since the appearance of the last same homephone as the current record.
133	last_fullname	The number of days passed since the appearance of the last same first name and last name as the current record.
134	last_zip5	The number of days passed since the appearance of the last same zip code as the current record.
135	last_address	The number of days passed since the appearance of the last same address as the current record.
136	same_dbname_1	The number of same birthday, first name and last name as the current record appeared on the same day. It only counts the records that appear before the current record.
137	same_dbname_3	The number of same birthday, first name and last name as the current record appeared within the last 3 days.
138	same_dbname_7	The number of same birthday, first name and last name as the current record appeared within the last 7 days.
139	same_dbname_14	The number of same birthday, first name and last name as the current record appeared within the last 14 days.
140	same_dbname_30	The number of same birthday, first name and last name as the current record appeared within the last 30 days.
141	same_dbname_diff_address_1	The number of same birthday, first name and last name but different address as the current record appeared on the same day. It only counts the records that appear before the current record.
142	same_dbname_diff_address_3	The number of same birthday, first name and last name but different address as the current record appeared within the last 3 days.
143	same_dbname_diff_address_7	The number of same birthday, first name and last name but different address as the current record appeared within the last 7 days.
144	same_dbname_diff_address_14	The number of same birthday, first name and last name but different address as the current record appeared within the last 14 days.
145	same_dbname_diff_address_30	The number of same birthday, first name and last name but different address as the current record appeared within the last 30 days.
146	same_dbname_diff_zip_1	The number of same birthday, first name and last name but different zip code as the current record appeared on the same day. It only counts the records that appear before the current record.
147	same_dbname_diff_zip_3	The number of same birthday, first name and last name but different zip code as the current record appeared within the last 3 days.
148	same_dbname_diff_zip_7	The number of same birthday, first name and last name but different zip code as the current record appeared within the last 7 days.
149	same_dbname_diff_zip_14	The number of same birthday, first name and last name but different zip code as the current record appeared within the last 14 days.
150	same_dbname_diff_zip_30	The number of same birthday, first name and last name but different zip code as the current record appeared within the last 30

		days.
151	same_dbname_diff_ssn_1	The number of same birthday, first name and last name but different ssn as the current record appeared on the same day. It only counts the records that appear before the current record.
152	same_dbname_diff_ssn_3	The number of same birthday, first name and last name but different ssn as the current record appeared within the last 3 days.
153	same_dbname_diff_ssn_7	The number of same birthday, first name and last name but different ssn as the current record appeared within the last 7 days.
154	same_dbname_diff_ssn_14	The number of same birthday, first name and last name but different ssn as the current record appeared within the last 14 days.
155	same_dbname_diff_ssn_30	The number of same birthday, first name and last name but different ssn as the current record appeared within the last 30 days.
156	same_dbname_diff_phone_1	The number of same birthday, first name and last name but different homephone as the current record appeared on the same day. It only counts the records that appear before the current record.
157	same_dbname_diff_phone_3	The number of same birthday, first name and last name but different homephone as the current record appeared within in the last 3 days.
158	same_dbname_diff_phone_7	The number of same birthday, first name and last name but different homephone as the current record appeared within in the last 7 days.
159	same_dbname_diff_phone_14	The number of same birthday, first name and last name but different homephone as the current record appeared within in the last 14 days.
160	same_dbname_diff_phone_30	The number of same birthday, first name and last name but different homephone as the current record appeared within in the last 30 days.

DATA CLEANING

From the data quality report, we can see that there are frivolous values occurred in variable ***"ssn",*** ***"homephone***" and ***"dob".*** For example, there are 4974 applications with home phone 9105580920. The frivolous value in ***"dob"*** did not affect our analysis because we built variables based on ***"dob",*** ***"firstname",*** and ***"lastname"*** instead of just ***"dob"*** for identification of an applicant.

To eliminate the effects posed by frivolous values in ***"ssn"*** and ***"homephone"***, we replaced the affected records in a new variable with the mean of all other records (excluding records that are affected by frivolous values) in that specific variable. For example, for the variable ***"last_ssn",*** we calculated the average number of days passed since the appearance of the last same SSN as the current record, excluding all records with SSN "737610282" (the frivolous SSN). We then replaced the values of ***"last_ssn"*** of those records with frivolous SSN by that average.

We also deleted two records with date of birth 1900/02/29, which is not a valid date.

FEATURE SELECTION PROCESS

We decided to use a “Filter” method before implementing our algorithms. The filter that we chose is to calculate the “Kolmogorov-Smirnov” score (KS Score) for each unique feature.

The Kolmogorov-Smirnov test is to measure how separate is one distribution from the other distribution.

For applying to feature selection:

- Firstly, we derived two distributions from each variable: “Fraud=1” Distribution and “Fraud=0” Distribution.
- Then we used Kolmogorov-Smirnov test to see how well these two distributions are separated. The higher the KS score, the better performance of the variable in separating “goods” and “bads.”

We calculated the KS scores for all 160 variables on the training dataset, and we picked the top 30 variables for our algorithms. The chart below lists the top 30 variables and their KS Score:

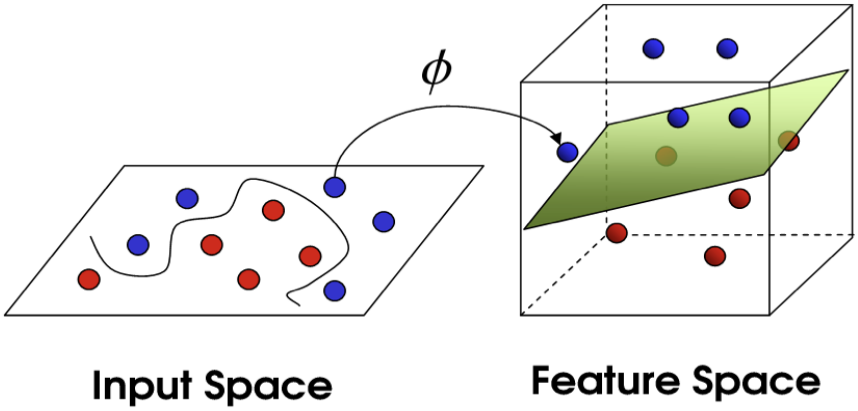
	Variable Names	KS Score
1	last_ssn	11.82934
2	last_address	11.72848
3	same_name_1	5.797611
4	same_homephone_3	4.41055
5	same_homephone_7	4.41055
6	same_homephone_14	4.41055
7	same_homephone_30	4.41055
8	same_name_diff_ssn_1	3.977709
9	same_name_diff_zip_1	3.973869
10	same_name_diff_address_1	3.9658
11	same_name_diff_phone_1	3.932676
12	same_name_diff_bd_1	3.858358
13	same_homephone_diff_ssnname_3	2.672979
14	same_homephone_diff_ssnname_7	2.672979
15	same_homephone_diff_ssnname_14	2.672979

16	same_homephone_diff_ssnname_30	2.672979
17	same_homephone_diff_bdname_3	2.665105
18	same_homephone_diff_bdname_7	2.665105
19	same_homephone_diff_bdname_14	2.665105
20	same_homephone_diff_bdname_30	2.665105
21	same_homephone_diff_address_3	2.665105
22	same_homephone_diff_address_7	2.665105
23	same_homephone_diff_address_14	2.665105
24	same_homephone_diff_address_30	2.665105
25	same_homephone_diff_zip_3	2.651373
26	same_homephone_diff_zip_7	2.651373
27	same_homephone_diff_zip_14	2.651373
28	same_homephone_diff_zip_30	2.651373
29	same_homephone_diff_address_1	2.184209
30	same_homephone_diff_zip_1	2.180175

ALGORITHMS

SVM (Support Vector Machine)

We chose to use SVM as one of our classification algorithms. SVM is a good supervised learning algorithm for both classification and regression problems. As a classifier, SVM can form a non-linear decision boundary by projecting input observations to higher dimension spaces and split them as wide as possible. SVM could have a good performance when the actual decision boundary is not linear. However, in this applications dataset, SVM is not a strong model as expected.



We used 30 variables that we selected from feature selection stage to fit the SVM model, and we used the default parameters settings that the function returned. SVM can give each observation a probability for each level, and we treat the probability to be classified as “1” (fraud) as the score to sort our populations. We got FDR@10% of 12.80% of the training dataset. We got FDR@10% of 11.29% for the testing dataset, which is slightly lower than the training FDR as expected. Amazingly, for OOT dataset, we saw a slight increase in FDR@10% which is 13.79%. However, in general, SVM is not a very strong model for this dataset, and it is possible that the performance is volatile and deteriorated due to the random noise.

Gradient Boosting Decision Trees

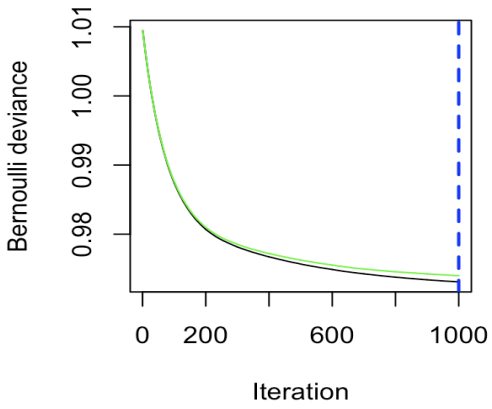
Gradient Boosting Decision Trees (GBDT) is an iterative functional gradient descent machine learning algorithm for regression and classification problems and is also the model with the best performance in this classification problem. The algorithm creates a series of weak models, where the final prediction model is then a linear combination of this series of weak models, typically decision trees.

Generally speaking, the next model in the series is trained on a weighted data set where the records with the largest error so far are more heavily weighted. GBDT builds the model in a stage-wise

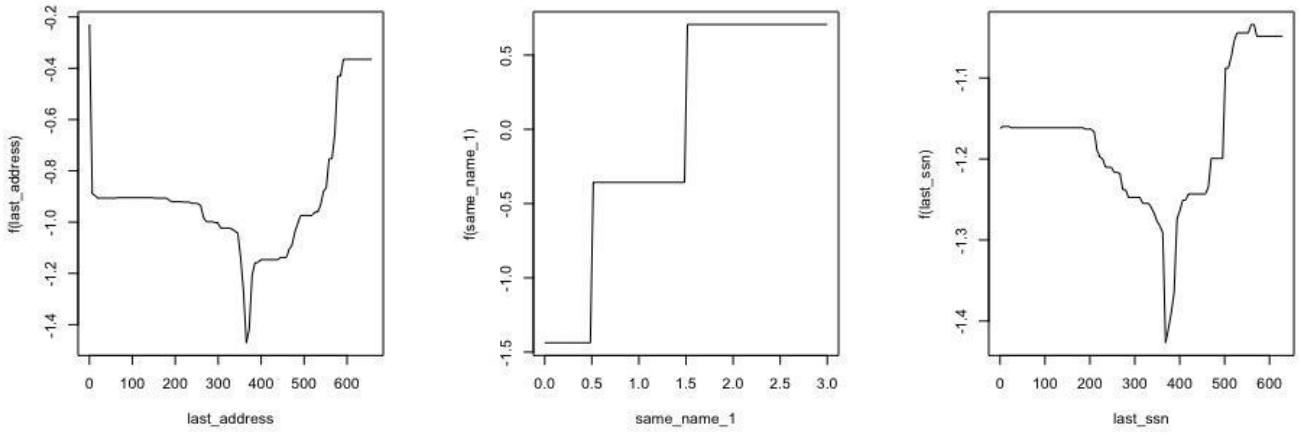
fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

In our model, we used gbm() function in R to train the model. We set the number of trees at 1000, shrinkage at 0.01 and the distribution of loss function as “Bernoulli” since this is a binary classification problem. To ensure the model’s accuracy, we set a 5-fold cross-validation.

After training the model, we used gbm.pref() function to determine the optimal number of iterations which is 1000. The picture below shows the relationship between Bernoulli deviance and number of iterations.



Then, we used summary.gbm( ) function to identify the importance of variables. From 30 selected variables, the model itself further selected 23 ones, and the top 3 most important variables are ***last_address, same_name_1*** and ***last_ssn***. The plots below show marginal effect of these three variables.



More specific explanation on the top 3 variables are as follows:

1. ***last_address:*** The number of days passed since the previous occurrence of the same address to the record (range from 0 to 364, if not exists, the value is 365). The relative influence takes up 42.85%.
2. ***same_name_1:*** Within one day, the frequency of occurrence of exactly the same name (occurred before the record). The relative influence takes up 36.63%.
3. ***last_ssn:*** Same logic as last_address. The relative influence takes up 13.31%.

We got FDR@10% of 20.35% for the training dataset and FDR@10% of 20.31% for the testing dataset. For the OOT dataset, we got FDR@10% of 16.94%.

Logistic Regression

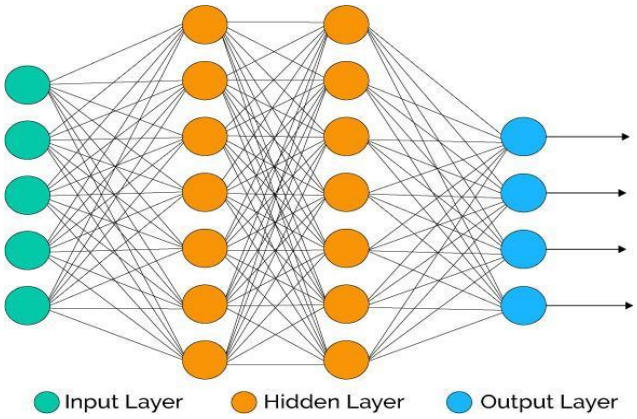
Logistic regression uses maximum likelihood to fit the logistic function, and gives outputs of the probability of fraud. Mathematically, logistic regression model gives a linear fit to log-odds, and estimates the coefficients to yield a number close to 1 for each individual who committed fraud, and a number close to 0 for each individual who did not commit fraud.

To reduce dimensionality in the model and hence reduce the model variance, we used forward stepwise logistic regression to select important variables in predicting fraud. The forward stepwise logistic regression method starts modeling with nothing, then test the addition of each variable according to AIC, and add the variables whose inclusion gives the most statistically significant improvement of the fit. It will repeat this process until none improves the model to a statistically significant extent. By implementing stepwise logistic regression, we selected 6 out of the 30 variables filtered by KS. Then we used these 6 variables to build logistic regression. The 6 variables selected include: 'last_address', 'same_name_1', 'same_homephone_diff_address_1', 'last_ssn', 'same_homephone_3', 'same_homephone_diff_ssnname_3'.

Using logistic regression, we got FDR@10% of 19.94% for the training dataset and FDR@10% of 20.19% for the testing dataset. For the OOT dataset, we got FDR@10% of 15.68%.

Neural Network

A neural network has an input layer, one or more hidden layers, and an output layer. For this model, each node in the input layer consists the information of 30 selected variables for each record. The nodes in the output layer are either 0 or 1; 0 represents non-fraudulent application and 1 represents fraud. For the hidden layer, we chose 1 hidden layer with 5 nodes. We trained our neural network with python package ‘sklearn.nerual_network’, setting hidden layer to have size 1 and 5 nodes.



Using neural network algorithm, we got FDR@10% of 19.89% for the training dataset and FDR@10% of 20.05% for the testing dataset. For the OOT dataset, we got FDR@10% of 16.57%.

The results of all above algorithms are shown as follows:

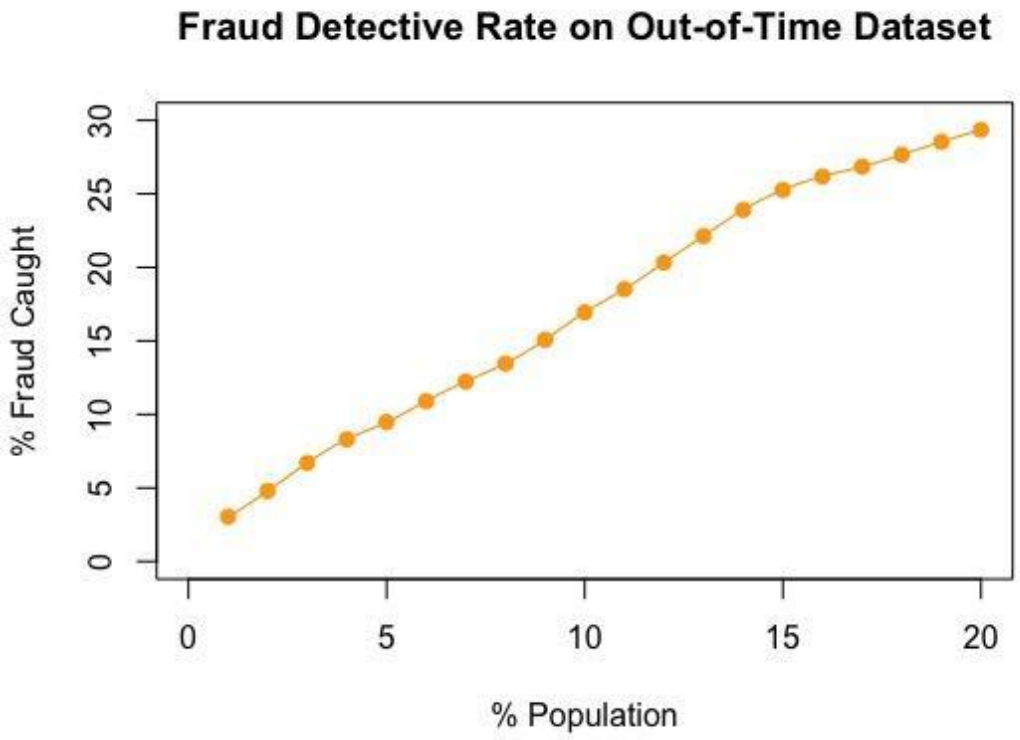
	FDR @ 10%		
Model	**Training**	**Testing**	**Out of Time**
SVM	12.80%	11.29%	13.79%
Gradient Boosting	20.35%	20.31%	16.94%
Neural Net	19.89%	20.05%	16.57%
Logistic Regression	19.94%	20.19%	15.68%

RESULTS

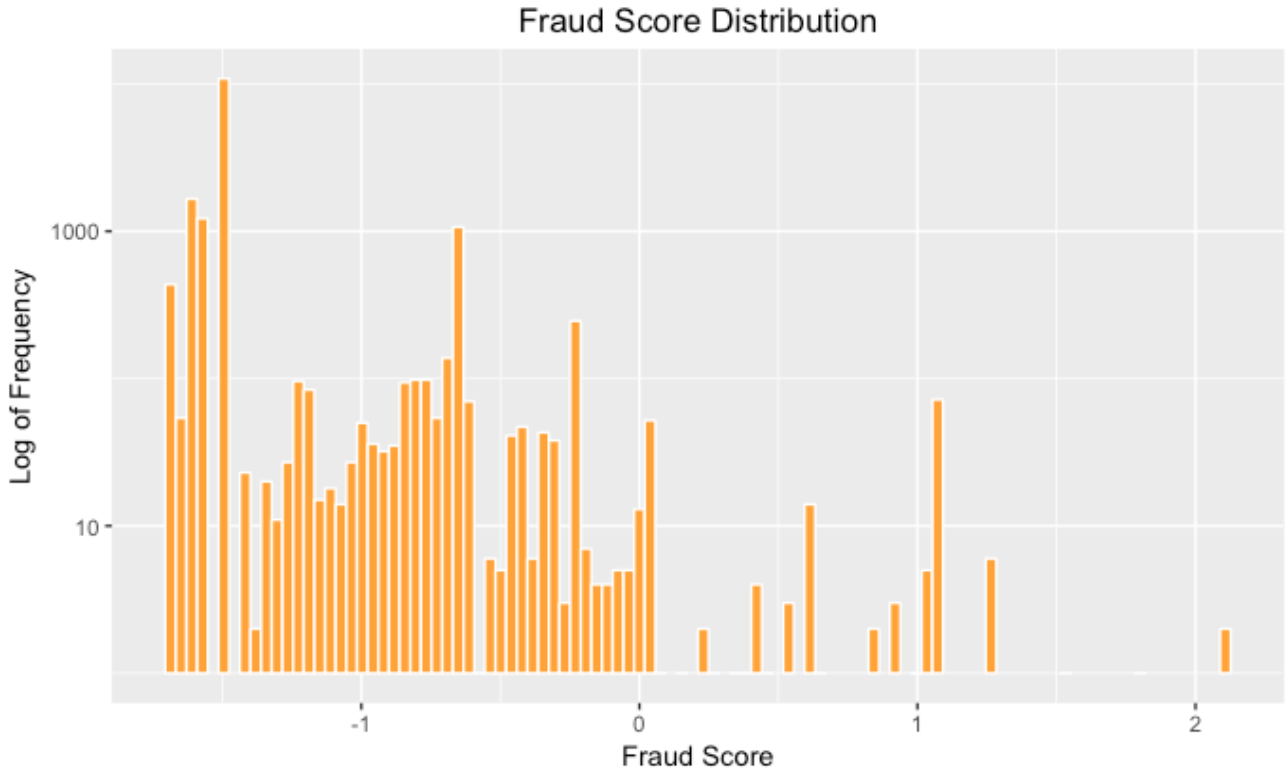
Gradient Boosting Decision Trees algorithm has the best performance among all the algorithm, and we got FDR@10% of 16.94%. The table below shows top 20% bins of the population. The Gradient Boosting Decision Trees model caught 29.35% fraud at 20% cutoff.

Overall Bad Rate is 25.49%	Bin Statistics					Cumulative Statistics					
Population Bin %	Total # records	# Good	# Bad	% Good	% Bad	Cumulative Good	Cumulative Bad	% Good	% Bad (FDR)	KS	False Pos. Ratio
1	170	38	132	22.35	77.65	38	132	0.30	3.04	2.74	0.29
2	170	94	76	55.29	44.71	132	208	1.04	4.79	3.75	0.63
3	170	87	83	51.18	48.82	219	291	1.73	6.71	4.98	0.75
4	171	102	69	59.65	40.35	321	360	2.53	8.30	5.77	0.89
5	170	119	51	70.00	30.00	440	411	3.47	9.47	6.00	1.07
6	170	108	62	63.53	36.47	548	473	4.32	10.90	6.58	1.16
7	170	112	58	65.88	34.12	660	531	5.21	12.24	7.03	1.24
8	170	117	53	68.82	31.18	777	584	6.13	13.46	7.33	1.33
9	170	100	70	58.82	41.18	877	654	6.92	15.08	8.16	1.34
10	171	90	81	52.63	47.37	967	735	7.63	16.94	9.32	1.32
11	170	102	68	60.00	40.00	1069	803	8.43	18.51	10.08	1.33
12	170	91	79	53.53	46.47	1160	882	9.15	20.33	11.18	1.32
13	170	92	78	54.12	45.88	1252	960	9.88	22.13	12.25	1.30
14	170	93	77	54.71	45.29	1345	1037	10.61	23.91	13.30	1.30
15	170	111	59	65.29	34.71	1456	1096	11.48	25.27	13.78	1.33
16	171	131	40	76.61	23.39	1587	1136	12.52	26.19	13.67	1.40
17	170	141	29	82.94	17.06	1728	1165	13.63	26.86	13.23	1.48
18	170	135	35	79.41	20.59	1863	1200	14.69	27.66	12.97	1.55
19	170	132	38	77.65	22.35	1995	1238	15.74	28.54	12.80	1.61
20	170	135	35	79.41	20.59	2130	1273	16.80	29.35	12.54	1.67

The picture below shows the relationship between fraud detection rate and % population. We could see that fraud detection rate increases steadily with % Population. After 15% population, the fraud detected rate rises more slowly compared to before.



The picture below shows the distribution of predicted fraud score on the out-of-time dataset. The distribution is highly right-skewed. Most records have predicted fraud score below zero.



CONCLUSIONS

In this fraud detection project, we first created 160 new variables in two ways:

1) counts of the records with similar characteristics;

2) counts of days of the last appearance of same attributes as the current record.

Then, we eliminated frivolous values in variables: “ssn”, “homephone”, “dob” and replaced with the mean of all other records from the same variable; we also deleted two records with a non-existing date. Next, we separated the dataset into training, testing and out-of-time sets, and did feature selection on the training dataset using KS and picked the top 30 variables with the highest KS scores. Last, we used SVM, Gradient Boosting Decision Trees, Logistic Regression, and Neural Networks to test individual algorithm’s performance on the dataset. As a result, Gradient Boosting Decision Trees algorithm has the best performance with FDR @ 10% of 16.94% on the Out of Time dataset.

Overall, the model that we built has weak performance, and there still remains much room for us to improve our methodology. With more time, we plan to:

- Create more features to characterize what a fraudster would do when committing fraud crime.
- Some classification models suffer from the class imbalance problem, and the classes of our dataset are not presented equally. About 80% of the applications are in the “Not-Fraud” class, and 20% are in the “Fraud” class. To fix this problem, we can use an oversampling method such us SMOTE to create synthetic samples from the minor class.
- Use k-fold cross-validation to determine the optimal parameters for our models.
- Try more feature selection methods and classification models.

APPENDIX

Summary Statistics

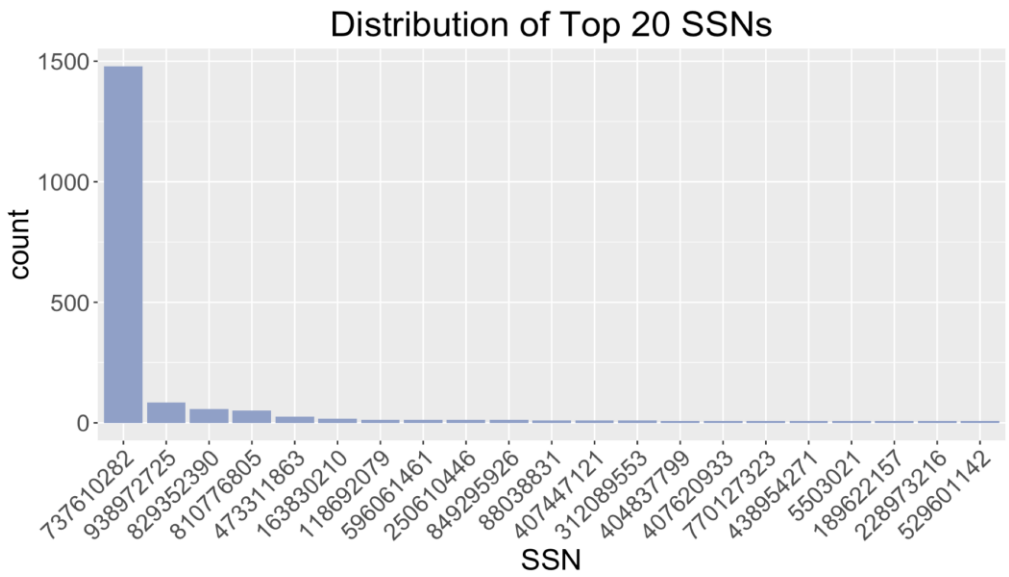
There are altogether 94,866 observations and 10 variables in this dataset. For all the 10 variables, we provide a summary table including the data type, the number of unique values, the number of missing values and the percentage populated in the field to describe the data uniqueness and completeness in this dataset.

Variable Name	Data Type	Number of Unique Values	Number of NAs	Data Completeness (%)
record	Categorical	94,866	0	100.00
date	Date	365	0	100.00
ssn	Categorical	86,771	0	100.00
firstname	Categorical	14,626	0	100.00
lastname	Categorical	31,513	0	100.00
address	Text	88,167	0	100.00
zip5	Categorical	15,855	0	100.00
dob	Date	30,599	0	100.00
homephone	Categorical	20,762	0	100.00
fraud	Categorical	2	0	100.00

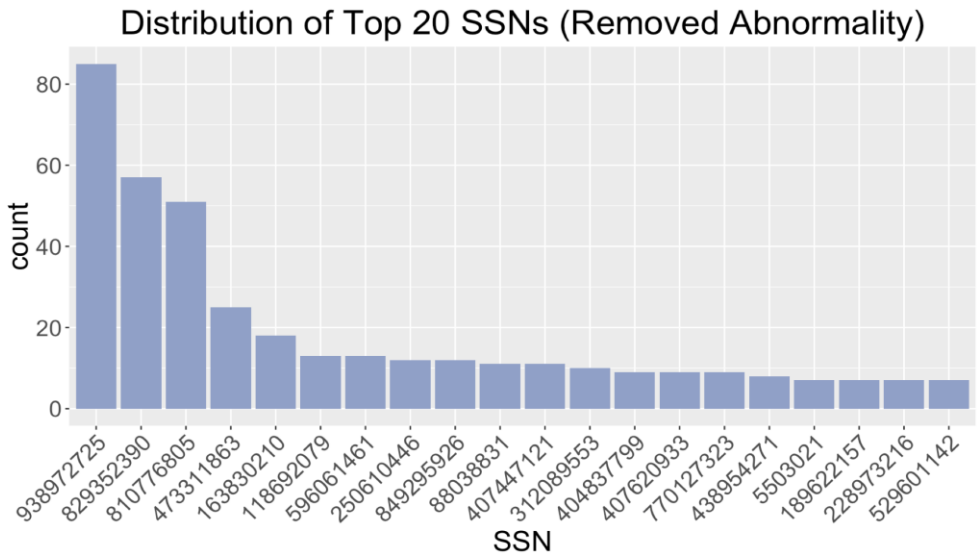
Detail Description of Variables

Field Name	Description
record	Categorical with no metrics. Unique ordinal reference number for each application record.

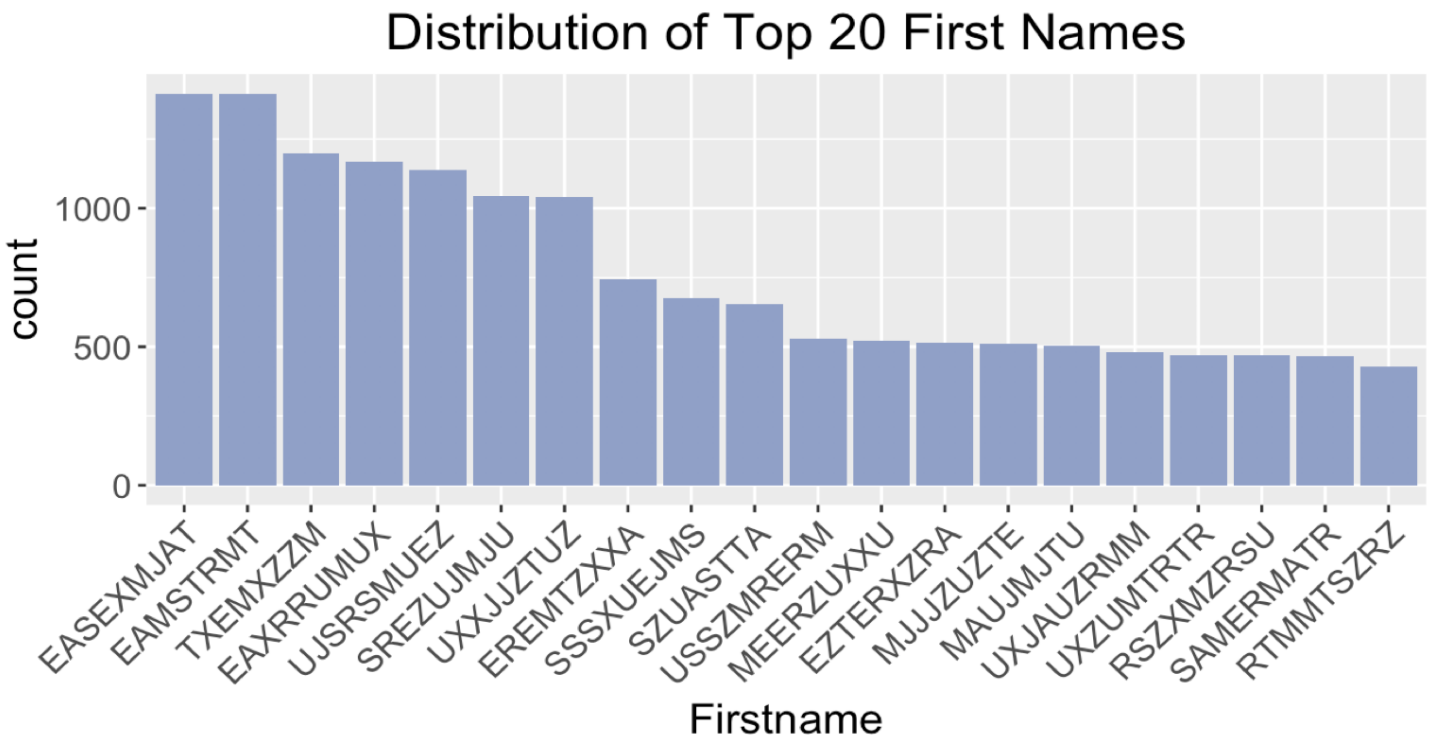
Field Name	Description
ssn	Categorical with no metrics. Applicant's SSN number. Most frequent record 737610282 removed.



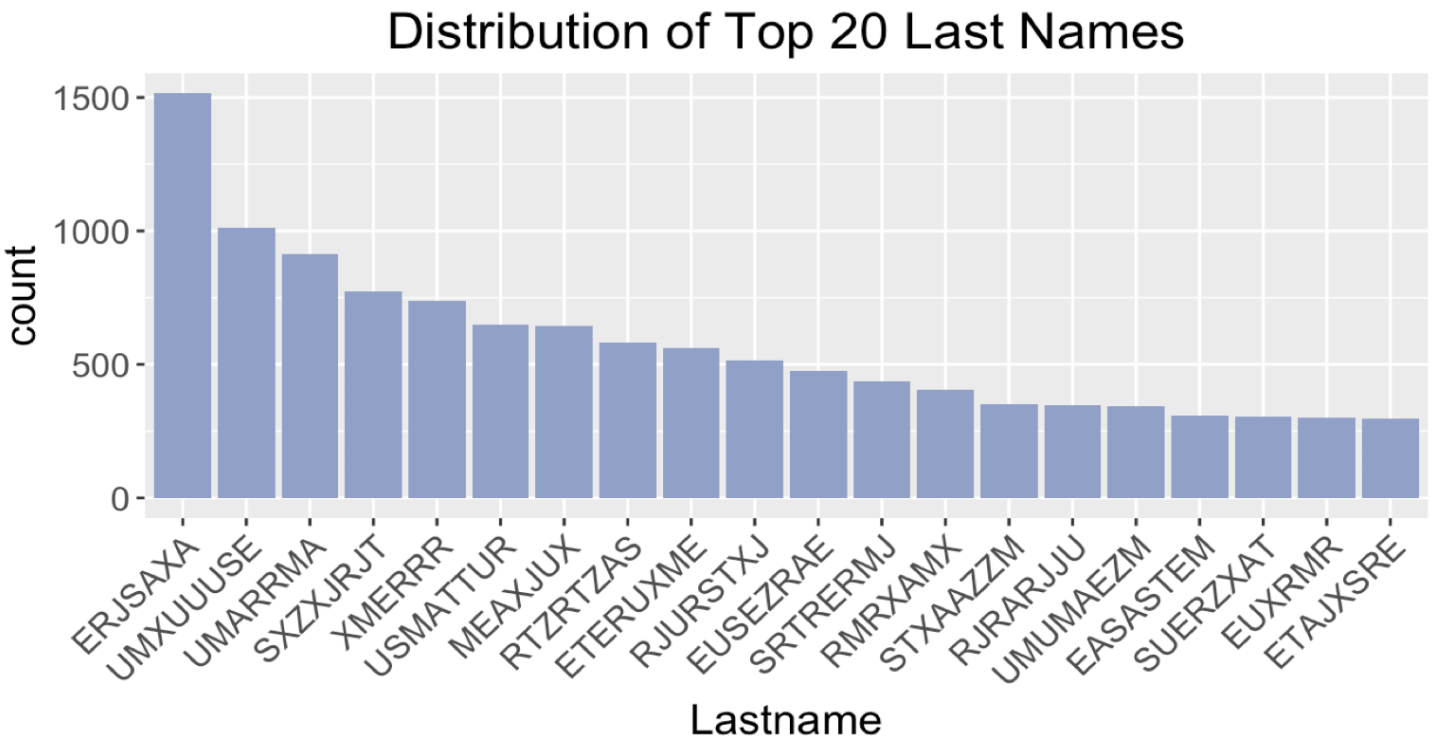
After plotting the variable "ssn", we find there is a SSN (737610282) that shows abnormally frequently in records. We believe this is a strong sign of erroneous data. Therefore, we decide to remove this abnormality in our future discussion of this dataset. The following graph is the distribution of the top 20 SSNs after removing 737610282.



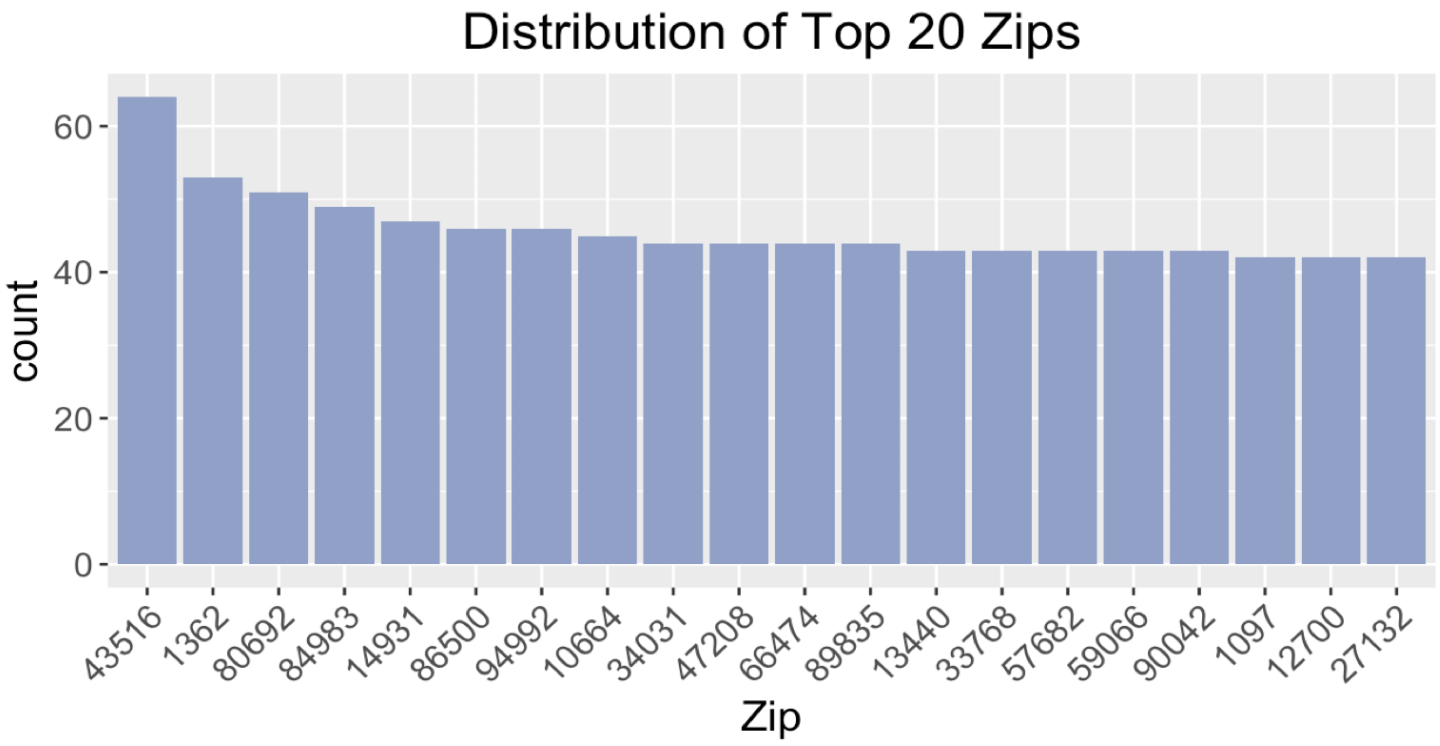
Field Name	Description
firstname	Categorical with no metrics. Applicant's first name.



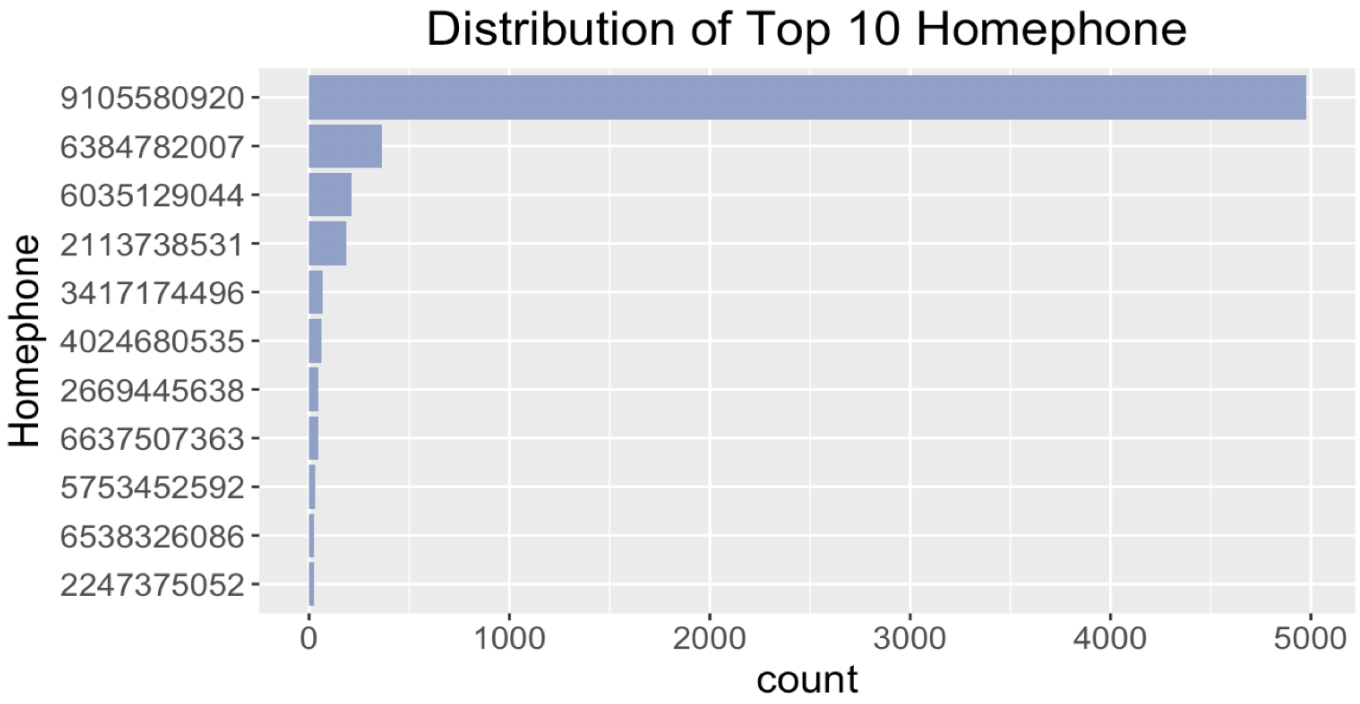
Field Name	Description
lastname	Categorical with no metrics. Applicant's last name.



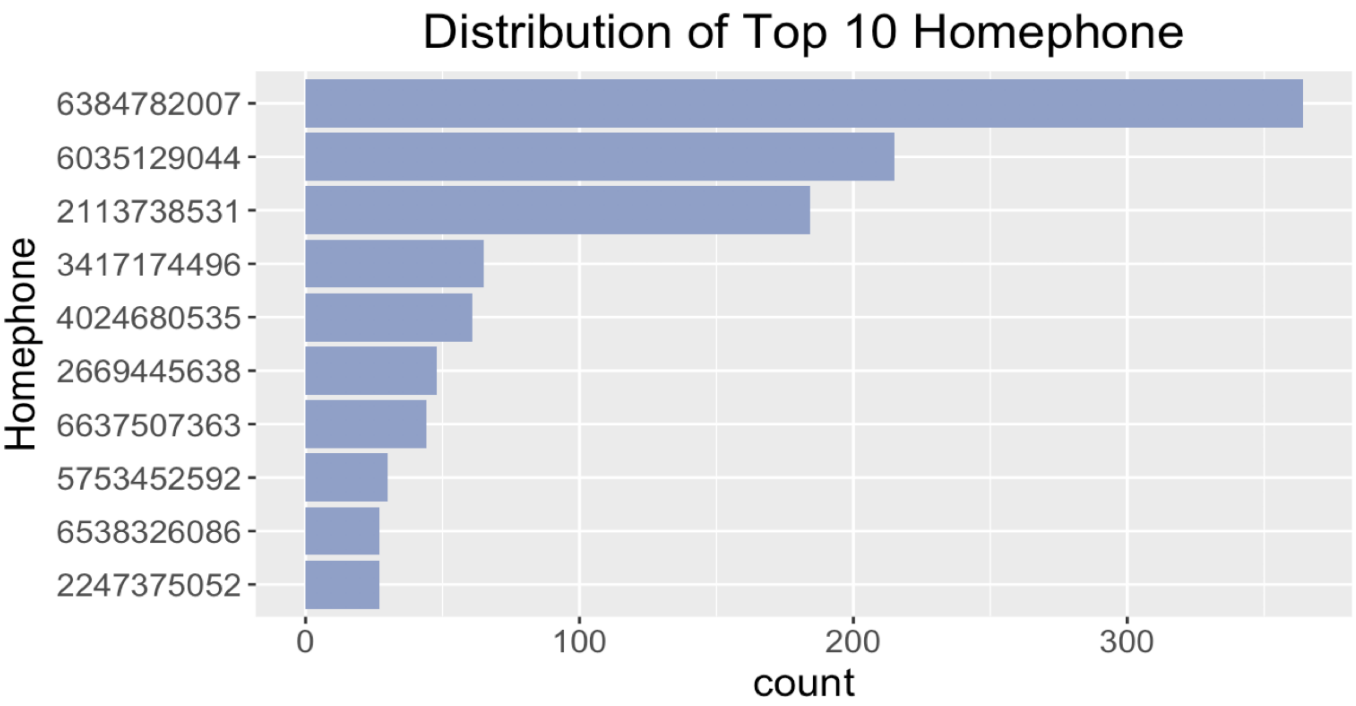
Field Name	Description
zip5	Categorical with no metrics. Applicant's ZIP code.



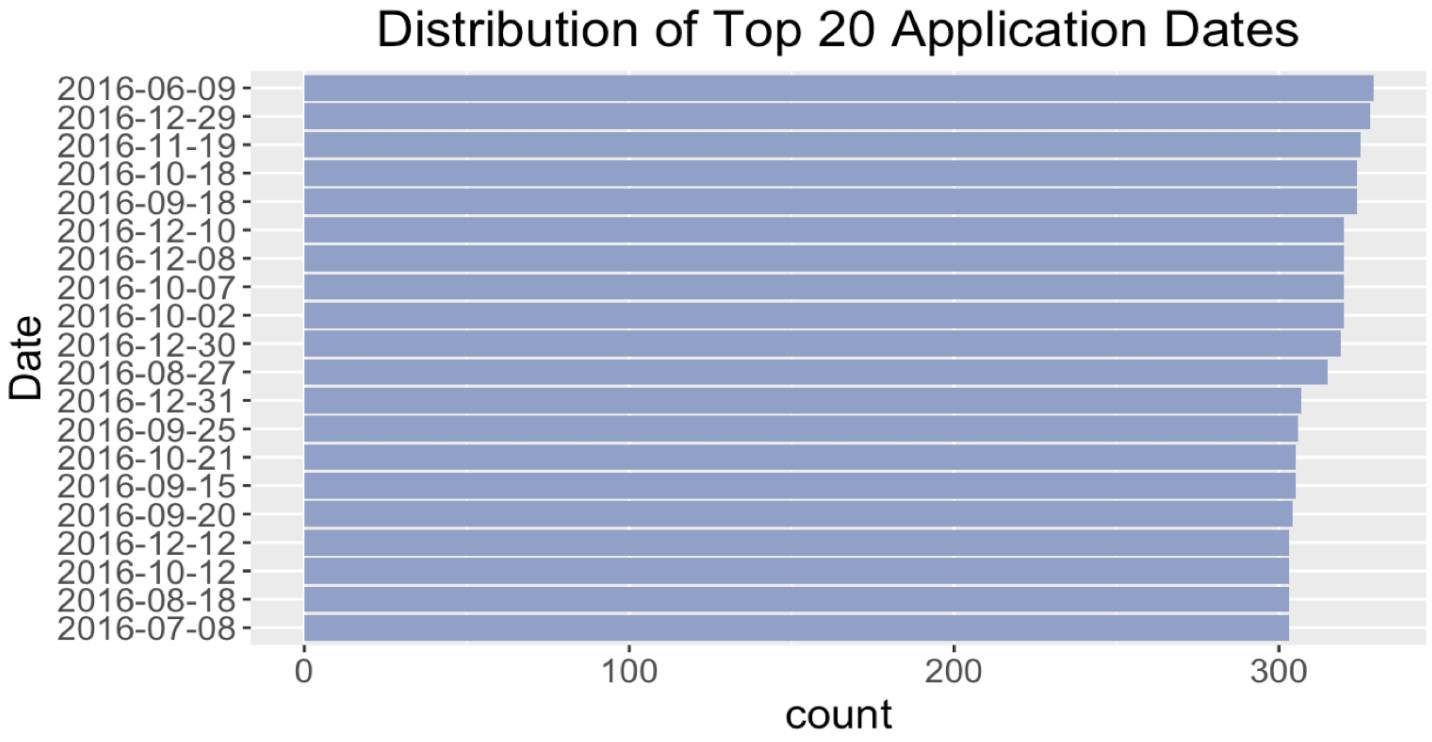
Field Name	Description
homephone	Categorical with no metrics. Applicant's homephone number. Most frequent record 9105580920 removed.

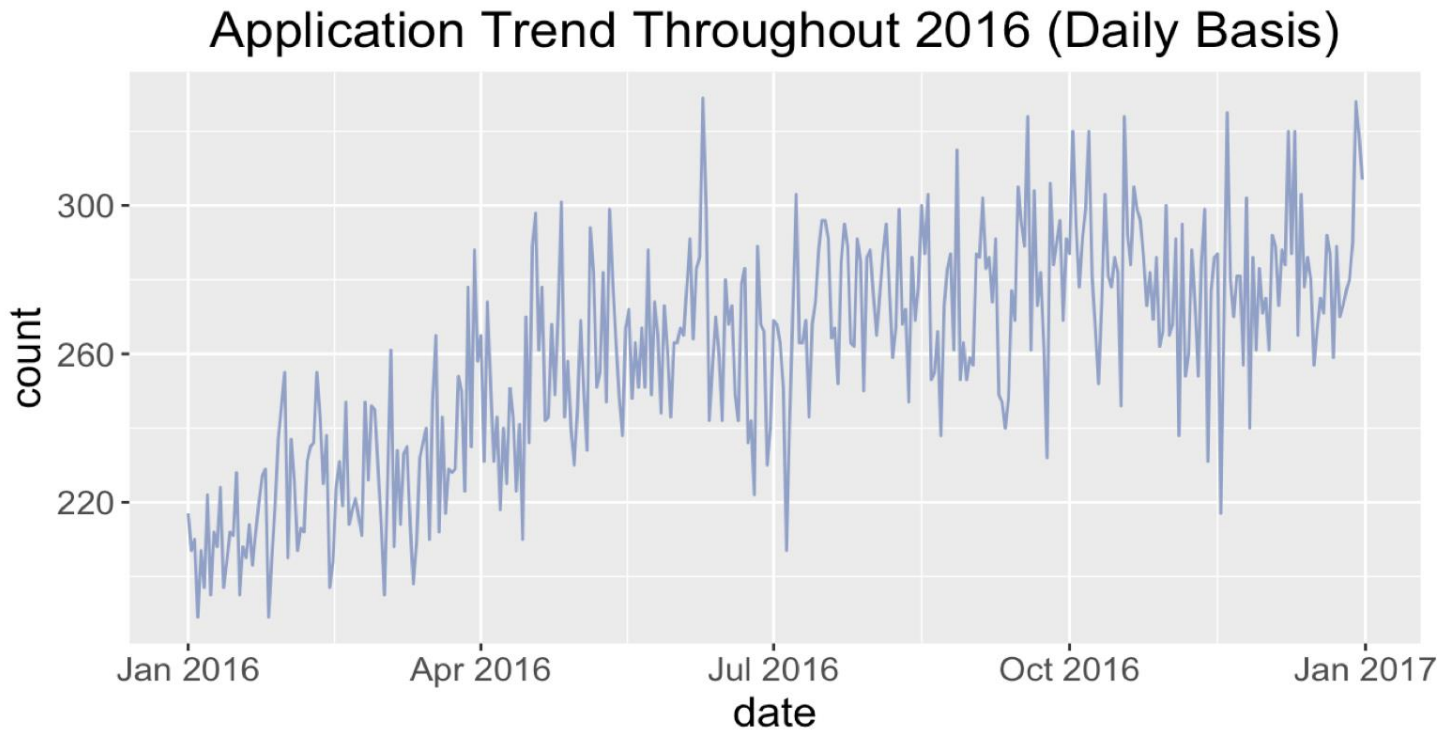


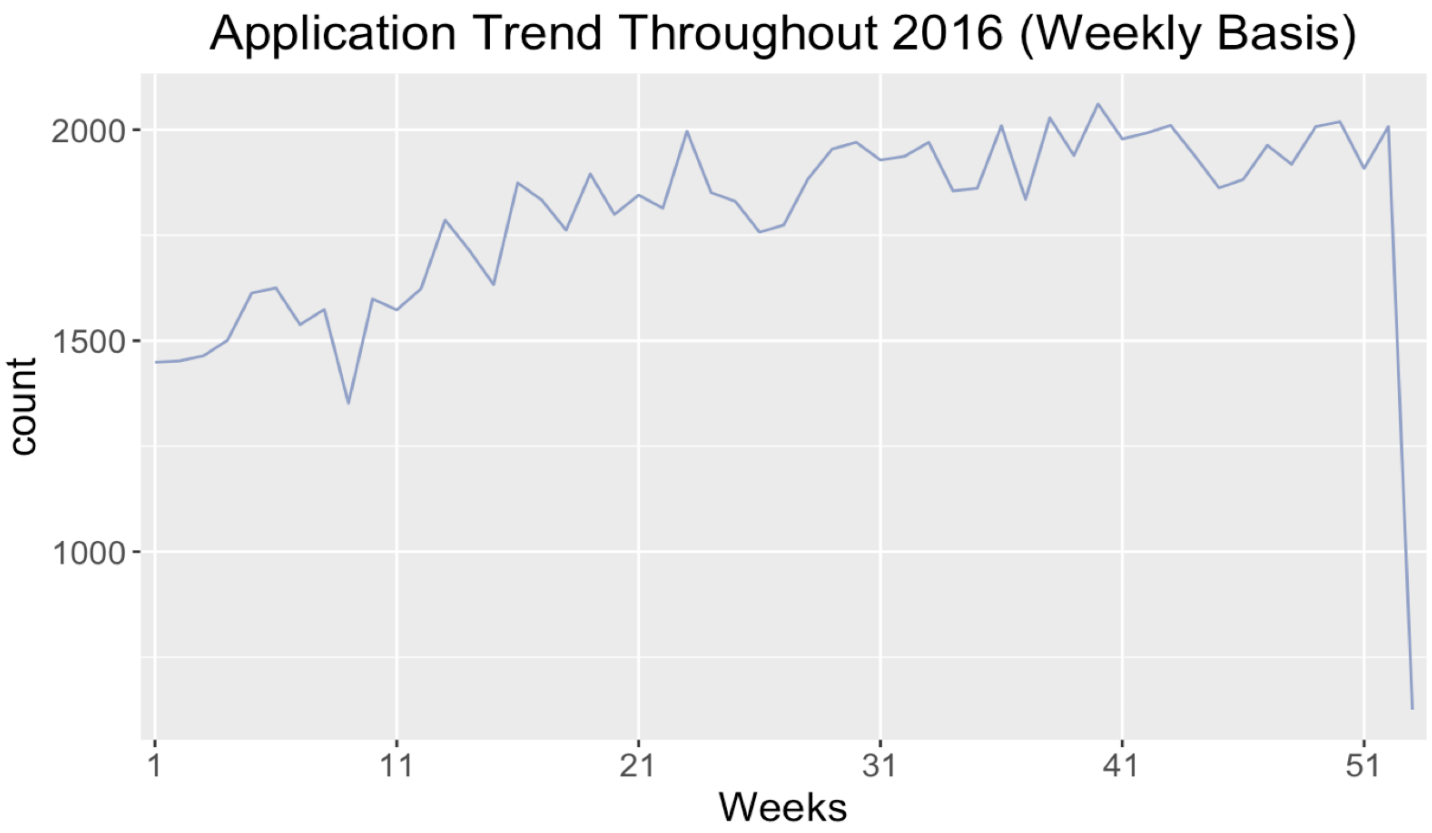
After plotting the variable "homephone", we find there is a home phone number (9105580920) that shows abnormally frequently in records. We believe this is a strong sign of erroneous data. Therefore, we decide to remove this abnormality in our future discussion of this dataset. The following graph is the distribution of the top 20 home phone numbers after removing 9105580920.

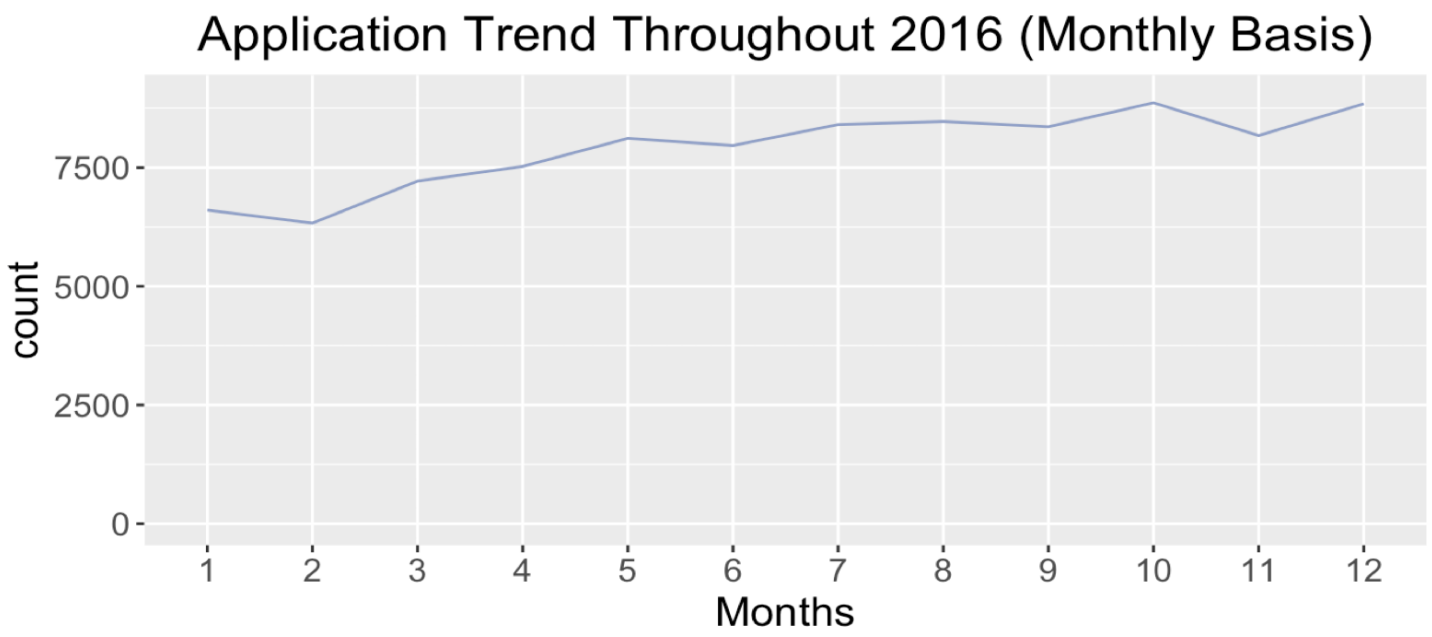


Field Name	Description
date	Categorical with no metrics. The date that the application was created.

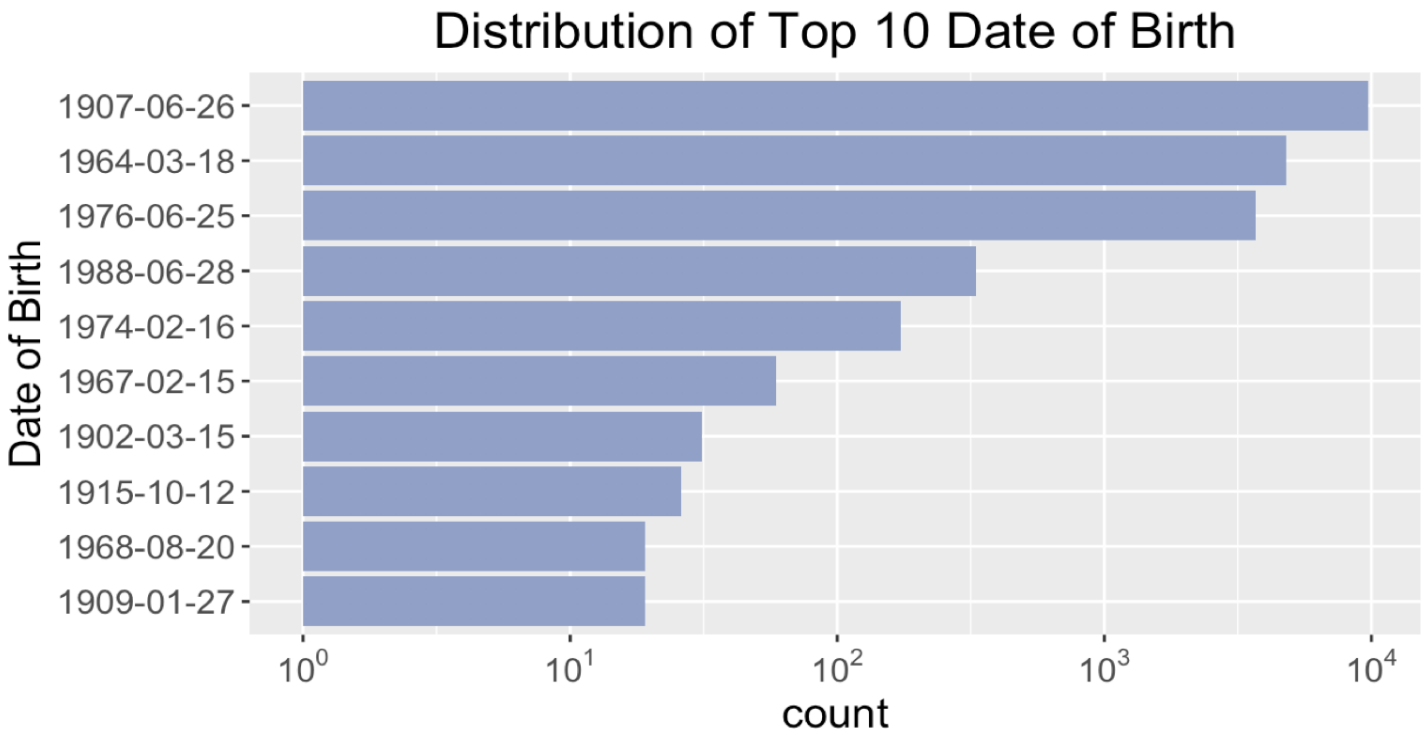




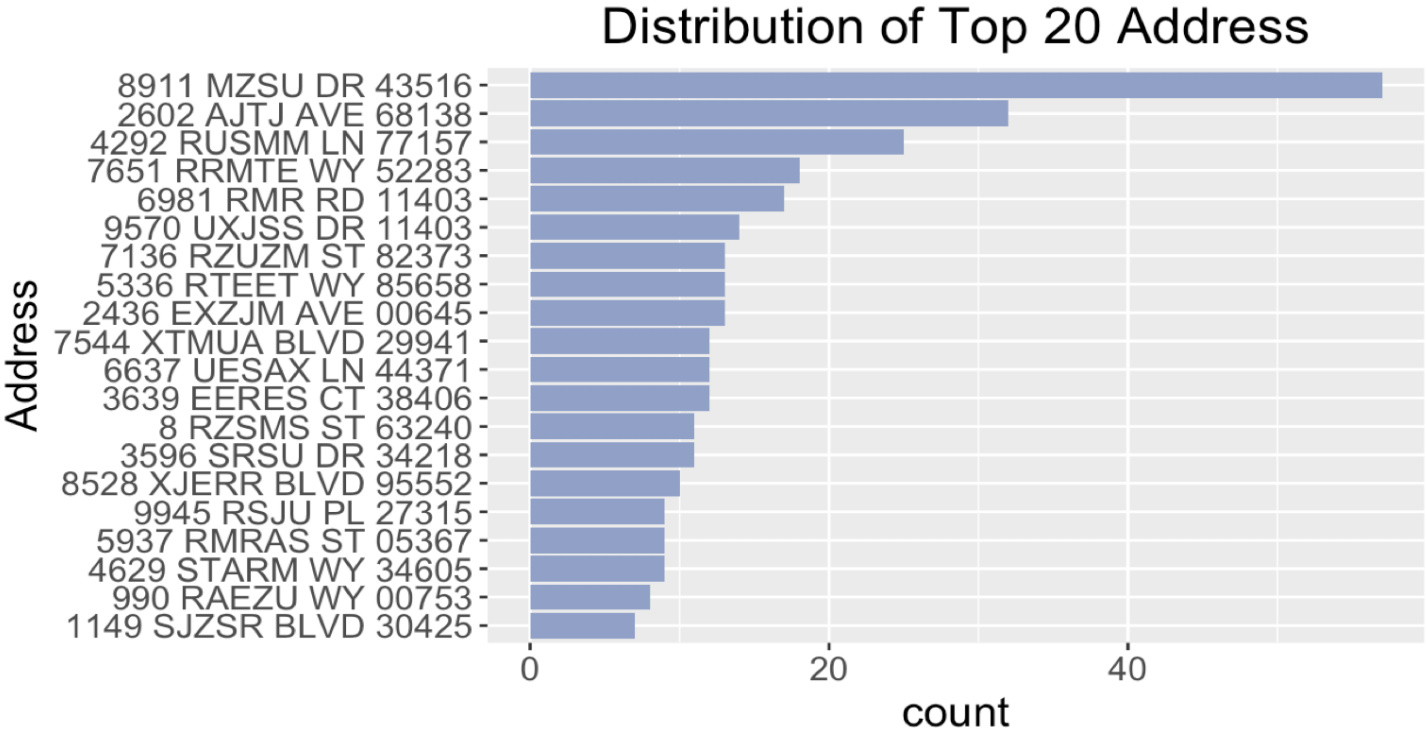

Application Trend Throughout 2016 (Weekly Basis)
count
2000
1500
1000
1
11
21
31
41
51
Weeks


Application Trend Throughout 2016 (Monthly Basis)
count
7500
5000
2500
0
1
2
3
4
5
6
7
8
9
10
11
12
Months

Field Name	Description
dob	Categorical with no metrics. The applicant's date of birth.



Field Name	Description
address	Categorical with no metrics. The applicant's address.



Field Name	Description
fraud	Categorical with metrics. If the transaction was recorded as a fraud, the field shows “1”; if the transaction was recorded as normal, the field shows “0”. There are altogether 74702 (78.74%) normal cases and 20164 (21.26%) fraud cases in this dataset.