Supervised Fraud Risk Modeling

Team 9

Faculty Advisor
Dr. Stephen Coggeshall

Table of Content

1. Executive Summary

The objective of this project is developing an end-to-end risk scoring system that screens product applications real-time for Client A. The prototype fraud risk scoring system has been successfully built and fine-tuned to maximize the client's overall fraud savings in a year. The prototype fraud risk model is expected to predict relative risk of frauds. This proposal serves as a technical documentation where we detailed the process through which the fraud scoring system was developed.

The implementation of the risk scoring system requires an automated pipeline from application submission through back-end data imputation to risk scoring. In this proposal, we will walk you through the process in the following sequence:

- data quality report
- data imputation
- candidate fraud risk variable creation
- fraud risk variable selection
- fraud risk model development
- results

The final fraud risk scoring model is selected from 5 candidate models trained using algorithms including Logistic Regression, Neural Network, Adaptive Boosting, Gradient Boosting, and Random Forest. Resampling and hyperparameter tuning were performed to ensure optimal prediction accuracy in terms of fraud detection rate (FDR). Compared to other models, the Adaptive Boosting model was selected to be the final fraud risk model. Based on the final model predictions, we estimate that the fraud detection rate would be around 55.1% and false positive rate 54.6% if Client A rejects the top risky %2 applications.

2. Data Description

The dataset received from Client A contains applications successfully submitted in calendar year 2016. For each application, applicant identity and contact information is recorded. A fraud label is also present, indicating whether an application is fraudulent or legitimate in retrospect.

- Dimension: 1,000,000 records, 10 fields
- Time frame: January 1, 2016 - December 31, 2016

Field Name	Description
record	Index uniquely identifying an application
date	The date on which the application was submitted
ssn	9-digit social security number
firstname	Applicant first name
lastname	Applicant last name
address	Applicant billing street address
zip5	Applicant billing address 5-digit zip code
dob	Applicant date of birth
homephone	10-digit applicant home phone number
fraud_label	Binary indicator (1 = fraudulent application, 0 = legitimate application)

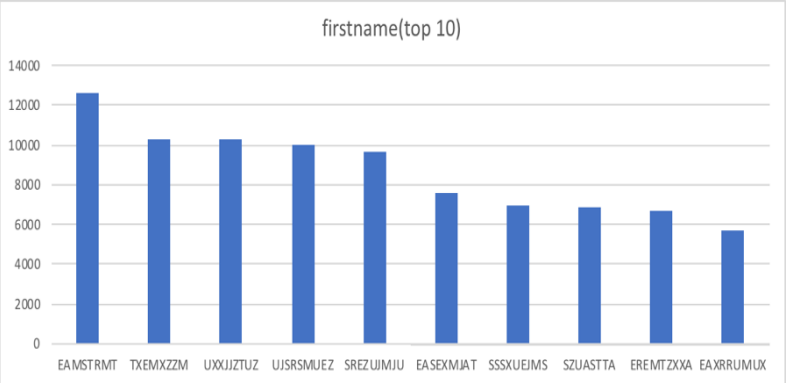
2.1 Summary Statistics

Field	Type	# Missing	% Populated	# Unique
ssn	Categorical	0	100%	835819
firstname	Categorical	0	100%	78136
lastname	Categorical	0	100%	177001
address	Categorical	0	100%	828774
zip5	Categorical	0	100%	26370
homephone	Categorical	0	100%	28244
date	Datetime	0	100%	365
dob	Datetime	0	100%	42673
fraud_label	Categorical	0	100%	2

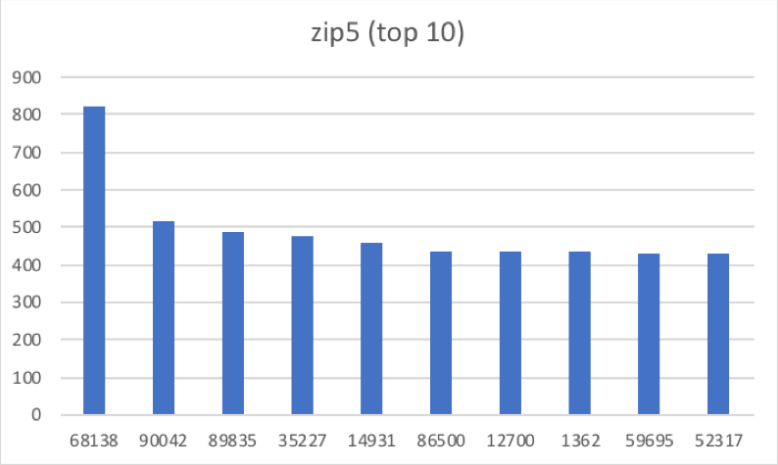
2.2 Field Information

As detailed summary on each field is provided in the data quality report (Appendix), we quickly visualize some of fields here as an overview.

[firstname]



[zip5]



[**Fraud_label**]

Value	Count	% Records
0	985607	98.56
1	14393	1.44

3. Data Imputation

3.1 Filling in Placeholder Values

As detailed in the data quality report attached in Appendix, 4 fields (**ssn**, **homephone**, **address**, **dob**) in the dataset contain placeholder values in place of empty values. Possible reasons for such values include:

1) Applicant does not have certain information
2) Applicant fails to provide certain information
3) Data quality issues

Under all circumstances, these placeholder values convey no real information about the applicant. Furthermore, the same placeholder value for all missing cells in a field wrongfully implies that many applications have the same such values in real life.

To avoid such misinterpretation, for each application, we replaced the placeholder value with record number, since record number is unique to each application. This way, we avoid unnecessary linking between records – a critical procedure in variable creation in the following phase.

Field name	Description	Placeholder	# Records	Replacement
ssn	Social security #	999999999	16935	**Record #**
homephone	Home phone #	9999999999	78512	**Record #**
address	Billing address	123 MAIN ST	1079	**Record #**
dob	Date of birth	19070626	126568	**Record #**

4. Candidate Variables

A total of 99 variables are created from the original 10 fields. We believe these 99 variables best characterise 2 types of fraud modes we know are prevalent in the business:

1. **Individual fraudster** - individual fraudsters copying multiple victims' identity (name, dob, ssn) with his/her own contact information (address, zip);
2. **Multiple fraudsters**- multiple fraudsters copying 1 victim's identity.

The following section details the reasoning behind the creation of such variables and exact formulas to calculate them. Overall, 3 categories of variables are created:

- **Recency variables** (Qty: 14)
- **Velocity variables** (Qty: 84)
- **Risk-table variable** (Qty: 1)

** **Recency** and **velocity variables** are created from the entire dataset; the **risk-table variable** is created from data excluding records after November 11, 2016.

** **Recency** and **velocity variables** are created based on link entities listed below:

14 Link Entities

- ssn
- namedob
- fulladdress
- homephone
- ssn, namedob
- ssn, fulladdress
- ssn, homephone
- namedob, fulladdress
- namedob, homephone,
- fulladdress, homephone
- ssn, namedob, fulladdress
- ssn, fulladdress, homephone
- ssn, namedob, homephone,
- namedob, fulladdress, homephone

Note:

- Namedob = firstname_lastname_dob
- fulladdress = address_zip5

4.1 Recency Variables

A set of 14 variables were created to measure the recency of certain link entities showing up. For every record, we calculate the length of days between certain link entities were last seen until the date of that particular record. The relationship between these recency variables and fraud modes are the following:

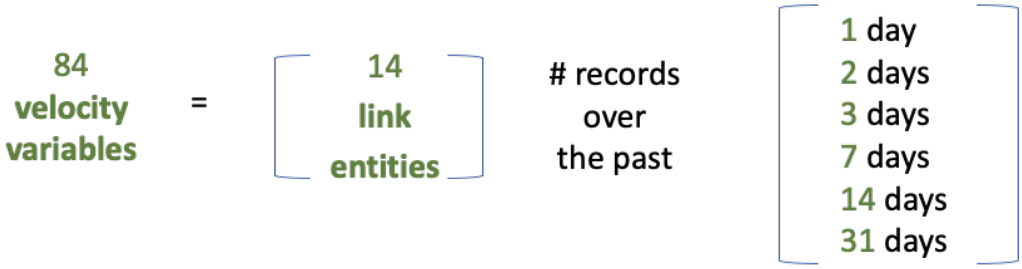
- Small value of recency variables ~ high risk of fraud mode 1
- Small value of recency variables ~ high risk of fraud mode 2

	Variable Name	*Description*
1	ssn_days	# days since ssn was last seen
2	fulladdress_days	# days since fulladdress was last seen
3	namedob_days	# days since namedob was last seen
4	homephone_days	# days since homephone was last seen
5	ssn,fulladdress_days	# days since ssn,fulladdress was last seen
6	ssn,namedob_days	# days since ssn,namedob was last seen
7	ssn,homephone_days	# days since ssn,homephone was last seen
8	fulladdress,namedob_days	# days since fulladdress,namedob was last seen
9	fulladdress,homephone_days	# days since fulladdress,homephone was last seen
10	namedob,homephone_days	# days since namedob,homephone was last seen
11	ssn,fulladdress,namedob_days	# days since ssn,fulladdress,namedob was last seen
12	ssn,fulladdress,homephone_days	# days since ssn,fulladdress,homephone was last seen
13	ssn,namedob,homephone_days	# days since ssn,namedob,homephone was last seen
14	fulladdress,namedob,homephone_days	# days since fulladdress,namedob,homephone was last seen

4.2 Velocity variables

A set of 84 variables were created to measure the rate of certain link entities showing up. The relationship between velocity variables and fraud modes are as follows:

- Large value of velocity variables ~ high risk of “individual fraudster” fraud
- Large value of velocity variables ~ high risk of “multiple fraudster” fraud



	Variable Name	Description
15	ssn1	# records seen over the past 1 day reporting the same ssn
16	ssn2	# records seen over the past 2 days reporting the same ssn
17	ssn3	# records seen over the past 3 days reporting the same ssn
18	ssn7	# records seen over the past 7 days reporting the same ssn
19	ssn14	# records seen over the past 14 days reporting the same ssn
20	ssn31	# records seen over the past 31 days reporting the same ssn
21	fulladdress1	# records seen over the past 1 day reporting the same fulladdress
22	fulladdress2	# records seen over the past 2 days reporting the same fulladdress
23	fulladdress3	# records seen over the past 3 days reporting the same fulladdress
24	fulladdress7	# records seen over the past 7 days reporting the same fulladdress
25	fulladdress14	# records seen over the past 14 days reporting the same fulladdress
26	fulladdress31	# records seen over the past 31 days reporting the same fulladdress
27	namedob1	# records seen over the past 1 day reporting the same namedob
28	namedob2	# records seen over the past 2 days reporting the same namedob

29	namedob3	# records seen over the past 3 days reporting the same namedob
30	namedob7	# records seen over the past 7 days reporting the same namedob
31	namedob14	# records seen over the past 14 days reporting the same namedob
32	namedob31	# records seen over the past 31 days reporting the same namedob
33	homephone1	# records seen over the past 1 day reporting the same homephone
34	homephone2	# records seen over the past 2 days reporting the same homephone
35	homephone3	# records seen over the past 3 days reporting the same homephone
36	homephone7	# records seen over the past 7 days reporting the same homephone
37	homephone14	# records seen over the past 14 days reporting the same homephone
38	homephone31	# records seen over the past 31 days reporting the same homephone
39	ssn,fulladdress1	# records seen over the past 1 day reporting the same ssn,fulladdress
40	ssn,fulladdress2	# records seen over the past 2 days reporting the same ssn,fulladdress
41	ssn,fulladdress3	# records seen over the past 3 days reporting the same ssn,fulladdress
42	ssn,fulladdress7	# records seen over the past 7 days reporting the same ssn,fulladdress
43	ssn,fulladdress14	# records seen over the past 14 days reporting the same ssn,fulladdress
44	ssn,fulladdress31	# records seen over the past 31 days reporting the same ssn,fulladdress
45	ssn,namedob1	# records seen over the past 1 day reporting the same ssn,namedob
46	ssn,namedob2	# records seen over the past 2 days reporting the same ssn,namedob
47	ssn,namedob3	# records seen over the past 3 days reporting the same ssn,namedob
48	ssn,namedob7	# records seen over the past 7 days reporting the same ssn,namedob
49	ssn,namedob14	# records seen over the past 14 days reporting the same ssn,namedob
50	ssn,namedob31	# records seen over the past 31 days reporting the same ssn,namedob
51	ssn,homephone1	# records seen over the past 1 day reporting the same ssn,homephone
52	ssn,homephone2	# records seen over the past 2 days reporting the same ssn,homephone
53	ssn,homephone3	# records seen over the past 3 days reporting the same ssn,homephone
54	ssn,homephone7	# records seen over the past 7 days reporting the same ssn,homephone
55	ssn,homephone14	# records seen over the past 14 days reporting the same ssn,homephone
56	ssn,homephone31	# records seen over the past 31 days reporting the same ssn,homephone

57	fulladdress,namedob1	# records seen over the past 1 day reporting the same fulladdress,namedob
58	fulladdress,namedob2	# records seen over the past 2 days reporting the same fulladdress,namedob
59	fulladdress,namedob3	# records seen over the past 3 days reporting the same fulladdress,namedob
60	fulladdress,namedob7	# records seen over the past 7 days reporting the same fulladdress,namedob
61	fulladdress,namedob14	# records seen over the past 14 days reporting the same fulladdress,namedob
62	fulladdress,namedob31	# records seen over the past 31 days reporting the same fulladdress,namedob
63	fulladdress,homephone1	# records seen over the past 1 day reporting the same fulladdress,homephone
64	fulladdress,homephone2	# records seen over the past 2 days reporting the same fulladdress,homephone
65	fulladdress,homephone3	# records seen over the past 3 days reporting the same fulladdress,homephone
66	fulladdress,homephone7	# records seen over the past 7 days reporting the same fulladdress,homephone
67	fulladdress,homephone14	# records seen over the past 14 days reporting the same fulladdress,homephone
68	fulladdress,homephone31	# records seen over the past 31 days reporting the same fulladdress,homephone
69	namedob,homephone1	# records seen over the past 1 day reporting the same namedob,homephone
70	namedob,homephone2	# records seen over the past 2 days reporting the same namedob,homephone
71	namedob,homephone3	# records seen over the past 3 days reporting the same namedob,homephone
72	namedob,homephone7	# records seen over the past 7 days reporting the same namedob,homephone
73	namedob,homephone14	# records seen over the past 14 days reporting the same namedob,homephone
74	namedob,homephone31	# records seen over the past 31 days reporting the same namedob,homephone
75	ssn,fulladdress,namedob1	# records seen over the past 1 day reporting the same ssn,fulladdress,namedob
76	ssn,fulladdress,namedob2	# records seen over the past 2 days reporting the same ssn,fulladdress,namedob

77	ssn,fulladdress,namedob3	# records seen over the past 3 days reporting the same ssn,fulladdress,namedob
78	ssn,fulladdress,namedob7	# records seen over the past 7 days reporting the same ssn,fulladdress,namedob
79	ssn,fulladdress,namedob14	# records seen over the past 14 days reporting the same ssn,fulladdress,namedob
80	ssn,fulladdress,namedob31	# records seen over the past 31 days reporting the same ssn,fulladdress,namedob
81	ssn,fulladdress,homephone1	# records seen over the past 1 day reporting the same ssn,fulladdress,homephone
82	ssn,fulladdress,homephone2	# records seen over the past 2 days reporting the same ssn,fulladdress,homephone
83	ssn,fulladdress,homephone3	# records seen over the past 3 days reporting the same ssn,fulladdress,homephone
84	ssn,fulladdress,homephone7	# records seen over the past 7 days reporting the same ssn,fulladdress,homephone
85	ssn,fulladdress,homephone14	# records seen over the past 14 days reporting the same ssn,fulladdress,homephone
86	ssn,fulladdress,homephone31	# records seen over the past 31 days reporting the same ssn,fulladdress,homephone
87	ssn,namedob,homephone1	# records seen over the past 1 day reporting the same ssn,namedob,homephone
88	ssn,namedob,homephone2	# records seen over the past 2 days reporting the same ssn,namedob,homephone
89	ssn,namedob,homephone3	# records seen over the past 3 days reporting the same ssn,namedob,homephone
90	ssn,namedob,homephone7	# records seen over the past 7 days reporting the same ssn,namedob,homephone
91	ssn,namedob,homephone14	# records seen over the past 14 days reporting the same ssn,namedob,homephone
92	ssn,namedob,homephone31	# records seen over the past 31 days reporting the same ssn,namedob,homephone
93	Fulladdress, namedob, homephone1	# records seen over the past 1 day reporting the same fulladdress,namedob,homephone
94	Fulladdress,namedob, homephone2	# records seen over the past 2 days reporting the same fulladdress,namedob,homephone
95	Fulladdress,namedob, homephone3	# records seen over the past 3 days reporting the same fulladdress,namedob,homephone

96	Fulladdress,namedob, homephone7	# records seen over the past 7 days reporting the same fulladdress,namedob,homephone
97	Fulladdress,namedob, homephone14	# records seen over the past 14 days reporting the same fulladdress,namedob,homephone
98	Fulladdress,namedob, homephone31	# records seen over the past 31 days reporting the same fulladdress,namedob,homephone

4.3 Weekday (Mean-encoding)

A mean-encoded variable was created, indicating relative fraud risk on a given weekday when the application was submitted. The theory was that fraudulent applications might occur more frequently on some days of the week than other days.

For each weekday, we calculate the % of frauds and use this value as risk measurement (excluding records in the last 60 days in 2016).

	Variable Name	Description
99	wday_risk	Weekday fraud risk • Monday risk: 0.013480 • Tuesday risk: 0.014070 • Wednesday risk: 0.014981 • Thursday risk: 0.014499 • Friday risk: 0.014968 • Saturday risk: 0.013674 • Sunday risk: 0.013480

5. Feature Selection

Previously in the variable creation phase, we created as many as 99 candidate variables to characterize frauds. In the way they are created, many of them are strongly correlated with one another; it is also true that some of them are superior to others in terms of predictive power. Thus, we sought to select a much smaller batch of variables that are the most indicative of frauds. In this phase, we built a feature selection pipeline using 2 techniques to distill the best 20 predictors to develop fraud detection models in the later phases:

1) "Filter": composite ranking by KS and FDR 3%
2) "Wrapper": recursive feature elimination

5.1 Composite Ranking by Kolmogorov–Smirnov (KS) & Fraud Detection Rate (FDR)

Out of 99 candidate variables, those that individually separate fraudulent and legitimate applications well are great predictors on their own. Thus, we sought to rank 99 candidate variables based on such individual predictive power through Kolmogorov–Smirnov and fraud detection rate:

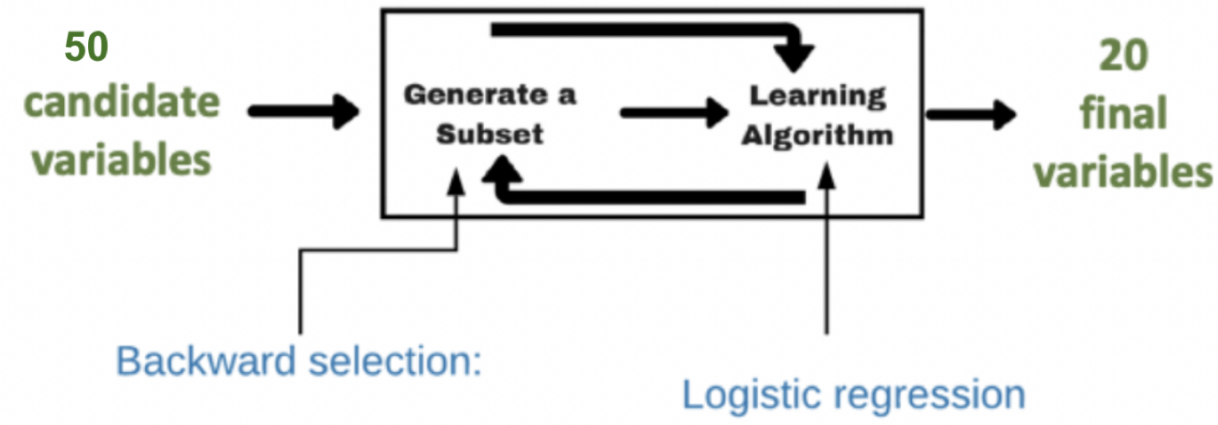
1) **Kolmogorov–Smirnov** test measures how well a numerical variable separates different classes. In our case, a good predictor should be able to separate goods and bads well enough.

$$KS = \max_{x} \int_{x_{min}}^{x} [P_{\text{good}} - P_{\text{bad}}]\, dx$$

2) **Fraud detection rate (3%)**: If we sort all records by a numerical variable descending in terms of extremeness, and identify the top 3% records as frauds, FRD 3% calculates the % of frauds successfully caught of all frauds in the data. In our case, a good predictor should have a high fraud detection rate relative to other variables.
3) **Composite rank = KS rank + FDR @ 3% rank.** After sorted by KS score and FDR @ 3%, 50 candidate variables that top both the KS and FDR ranking were selected to the move down the pipeline.

5.2 “Wrapper” with Logistic Regression and Backward Feature Selection

Through the filtering technique, 50 variables came out on top; however, modeling on such high dimensionality still proves computationally inefficient and theoretically unsound. The goal is to select 20 variables in total to develop fraud risk models. Because we screened variables by individual predictive power, we were eager to find out which 20 of these 50 variables collectively hold the most predictive power. Hence, we resorted to a “wrapper method” to find such optimal set of 20 variables.



1) the estimator is trained on the initial set of **50** features and the estimator performance is obtained;

2) The estimator loses 1 variable such that model performance increases the most;

3) The procedure is recursively repeated on the pruned set until 20 variables are eventually reached

	Final 20 Variable Name
1	fulladdress31
2	fulladdress7
3	fulladdress2
4	fulladdress1
5	namedob31
6	ssn,namedob31
7	fulladdress,homephone14
8	fulladdress,homephone7
9	homephone2
10	ssn,namedob3
11	fulladdress,homephone2
12	fulladdress,homephone1
13	namedob1
14	fulladdress,namedob31
15	namedob,homephone31
16	ssn,homephone31
17	fulladdress,namedob,homephone31
18	ssn,fulladdress31
19	ssn,fulladdress,homephone31
20	ssn,fulladdress,namedob14

6. Fraud Detection Modeling

6.1 Data Partition

After selecting 20 features, we split the whole dataset into three parts before training the models: training, testing, and out-of-time set. We separated records in the last 60 days in 2016 as out-of-time and then randomly split the rest into training and testing in 7:3 proportion.

Dataset	# Records	# Frauds	% Fraud
Training	583454	8400	1.44
Testing	250053	3607	1.44
Out-Of-Time	166493	2386	1.43
Total	1000000	14393	1.44

During modeling, we fitted the model on training set, fine-tune the model according to the testing set, and finally tested our model on out-of-time set.

6.2 Model Development

Out of all trained 5 models, the **Adaptive Boosting model** was selected to be the final fraud risk model. We summarized the fraud detection rate at 3% for each model below. The Gradient boosting model is the most performant in terms of FDR on the out-of-time set; it also produces relatively steady results across 10 trials.

Model	*Training*	*Testing*	*Out of Time*
Logistic Regression (baseline)	53.88%	51.57%	51.93%
Random Forest	57.97%	55.58%	55.00%
Gradient Boosting	57.56%	55.17%	54.04%
Adaptive Boosting (Final)	**57.48%**	**57.08%**	**55.06%**
Neural Network	58.01%	55.68%	54.91%

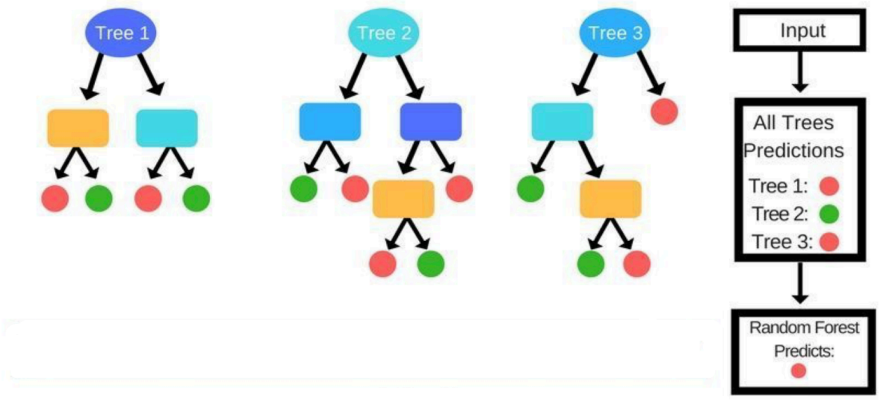
6.2.1 Logistic Regression (baseline model)

Logistic Regression is a classification algorithm that predicts the probability of binary outcomes using sigmoid function. As our baseline model, logistic regression model got more than 51% FDR @ 3% in out-o- time dataset.

FDR @3%	*Train*	*Test*	*OOT*
	53.88%	51.57%	51.93%

6.2.2 Random Forest

Random Forest is a supervised learning algorithm. It is an ensemble of many decision trees through the "bagging" approach. The general idea of the bagging method is that a combination of learning models increases the overall result. Random Forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This creates de-correlated trees that generally results in superior performance. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. We can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree).



Final parameters

- # trees: 200
- # of candidate variables for splitting: 4

FDR@3%	*Train*	*Test*	*OOT*
1	57.96%	55.59%	54.99%
2	57.94%	55.61%	54.99%
3	57.94%	55.59%	54.99%
4	57.96%	55.56%	55.03%
5	58.00%	55.53%	55.03%
6	57.96%	55.61%	54.99%
7	57.92%	55.59%	54.99%
8	58.05%	55.53%	55.03%
9	58.00%	55.59%	54.99%
10	57.94%	55.56%	54.99%
Average	**57.97%**	**55.58%**	**55.00%**

6.2.3 Adaptive Boosting

Boosted Trees seeks to build a robust model by combining several weak learners such as naive decision trees. The AdaBoost algorithm trains all records and the next generation will focus on previously misclassified records and assign stronger weight. Optimizing model by changing important parameters including the depth of the tree, the number of weak learners, and the loss function. In our case, we used a linear combination when updating the weights after each boosting iteration since this set has good fit through the previous logistic regression model.

Final parameters

- # trees: 1000
- Max depth: 10
- Loss function: linear

FDR 3%	*Train*	*Test*	*OOT*
1	57.35%	57.43%	55.01%
2	57.01%	56.02%	55.76%
3	57.53%	55.79%	54.66%
4	59.49%	56.41%	54.33%
5	56.95%	56.13%	55.41%
6	55.31%	58.91%	56.97%
7	57.96%	56.14%	55.49%
8	57.24%	57.17%	53.39%
9	57.70%	57.50%	54.03%
10	58.11%	57.40%	55.50%
Average	**57.48%**	**57.08%**	**55.06%**

6.2.3 Gradient Boosting

The Gradient Bosting is another type of Boosted Trees. GB builds an additive model in a forward stage-wise fashion. It allows for the optimization of arbitrary differentiable loss functions. In each stage a tree is fit on the negative gradient of the given loss function. The main differences are that Gradient Boosting is a generic algorithm to find approximate solutions to the additive modeling problem, while AdaBoost can be seen as a special case with a particular loss function. Therefore, gradient boosting is much more flexible and could capture more frauds in this case.

Final parameters

- # trees: 1000
- Max depth: 10
- Loss function: linear

FDR 3%	*Train*	*Test*	*OOT*
1	58.29%	55.22%	54.50%
2	58.09%	56.96%	53.54%
3	59.09%	54.09%	53.61%
4	56.49%	55.10%	54.54%
5	56.77%	54.53%	53.35%

6	56.75%	56.71%	55.46%
7	56.95%	53.90%	54.08%
8	57.44%	55.23%	53.72%
9	58.46%	55.26%	54.33%
10	57.36%	53.89%	54.16%
Average	**57.56%**	**55.17%**	**54.04%**

6.2.4 Neural Network

Neural Network is a highly adaptive learning algorithm and excellent alternative to other techniques mentioned above.

Final parameters

- # input layer nodes: 25
- # hidden layer nodes: 10
- Transformation function: sigmoid

FDR 3%	*Train*	*Test*	*OOT*
1	58.00%	55.75%	54.99%
2	58.08%	55.61%	54.90%
3	58.02%	55.72%	54.90%
4	58.06%	55.56%	55.07%
5	58.07%	55.61%	54.95%
6	57.94%	55.56%	54.90%
7	57.88%	55.61%	54.74%
8	58.14%	55.89%	54.99%
9	57.88%	55.64%	54.74%
10	57.98%	55.84%	54.95%
Average	58.01%	55.68%	54.91%

7. Results

As we have decided on the Adaptive Boosting model to be the final model for implementation, here we report a more in-depth view of its performance on training set, test set and out-of-time dataset separately.

The following tables summarize the top 1% to 20% records of high fraud risk. It is clear that the final model does a decent job on catching frauds on the training set and the performance degrades on the test set and further on the out-of-time set.

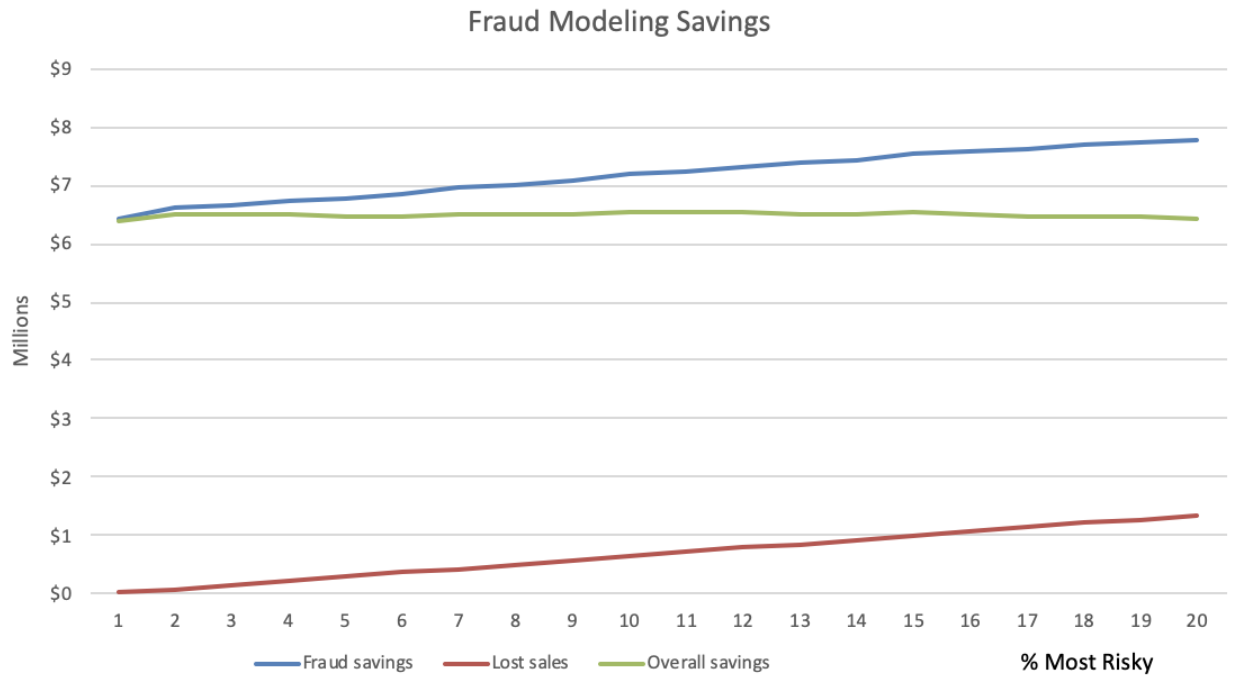
Training	# Records 600534		# Goods 591935		# Bads 8599	Fraud Rate 0.0143					
	Bin Statistics					Cumulative Statistics					
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Good	% Bad (FDR)	KS	FPR
1	6005	1303	4702	21.7	78.3	6005	1303	4702	0.2	54.7	54.5	0.3
2	6006	5874	132	97.8	2.2	12011	7177	4834	1.2	56.2	55.0	1.5
3	6005	5926	79	98.7	1.3	18016	13103	4913	2.2	57.1	54.9	2.7
4	6005	5962	43	99.3	0.7	24021	19065	4956	3.2	57.6	54.4	3.8
5	6006	5959	47	99.2	0.8	30027	25024	5003	4.2	58.2	54.0	5.0
6	6005	5960	45	99.3	0.7	36032	30984	5048	5.2	58.7	53.5	6.1
7	6005	5954	51	99.2	0.8	42037	36938	5099	6.2	59.3	53.1	7.2
8	6006	5938	68	98.9	1.1	48043	42876	5167	7.2	60.1	52.8	8.3
9	6005	5971	34	99.4	0.6	54048	48847	5201	8.3	60.5	52.2	9.4
10	6005	5954	51	99.2	0.8	60053	54801	5252	9.3	61.1	51.8	10.4
11	6006	5971	35	99.4	0.6	66059	60772	5287	10.3	61.5	51.2	11.5
12	6005	5967	38	99.4	0.6	72064	66739	5325	11.3	61.9	50.7	12.5
13	6005	5960	45	99.3	0.7	78069	72699	5370	12.3	62.4	50.2	13.5
14	6006	5969	37	99.4	0.6	84075	78668	5407	13.3	62.9	49.6	14.5
15	6005	5972	33	99.5	0.5	90080	84640	5440	14.3	63.3	49.0	15.6
16	6005	5977	28	99.5	0.5	96085	90617	5468	15.3	63.6	48.3	16.6
17	6006	5973	33	99.5	0.5	102091	96590	5501	16.3	64.0	47.7	17.6
18	6005	5965	40	99.3	0.7	108096	102555	5541	17.3	64.4	47.1	18.5
19	6005	5963	42	99.3	0.7	114101	108518	5583	18.3	64.9	46.6	19.4
20	6006	5975	31	99.5	0.5	120107	114493	5614	19.3	65.3	45.9	20.4

Testing	# Records 257373		# Goods 253613		# Bads 3760	Fraud Rate 0.0146					
	Bin Statistics					Cumulative Statistics					
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Good	% Bad (FDR)	KS	FPR
1	2574	564	2010	21.9	78.1	2574	564	2010	0.2	53.5	53.2	0.3
2	2573	2512	61	97.6	2.4	5147	3076	2071	1.2	55.1	53.9	1.5
3	2574	2539	35	98.6	1.4	7721	5615	2106	2.2	56.0	53.8	2.7
4	2574	2549	25	99.0	1.0	10295	8164	2131	3.2	56.7	53.5	3.8
5	2574	2551	23	99.1	0.9	12869	10715	2154	4.2	57.3	53.1	5.0
6	2573	2551	22	99.1	0.9	15442	13266	2176	5.2	57.9	52.6	6.1
7	2574	2553	21	99.2	0.8	18016	15819	2197	6.2	58.4	52.2	7.2
8	2574	2554	20	99.2	0.8	20590	18373	2217	7.2	59.0	51.7	8.3
9	2574	2551	23	99.1	0.9	23164	20924	2240	8.3	59.6	51.3	9.3
10	2573	2560	13	99.5	0.5	25737	23484	2253	9.3	59.9	50.7	10.4
11	2574	2558	16	99.4	0.6	28311	26042	2269	10.3	60.3	50.1	11.5
12	2574	2558	16	99.4	0.6	30885	28600	2285	11.3	60.8	49.5	12.5
13	2573	2561	12	99.5	0.5	33458	31161	2297	12.3	61.1	48.8	13.6
14	2574	2560	14	99.5	0.5	36032	33721	2311	13.3	61.5	48.2	14.6
15	2574	2555	19	99.3	0.7	38606	36276	2330	14.3	62.0	47.7	15.6
16	2574	2559	15	99.4	0.6	41180	38835	2345	15.3	62.4	47.1	16.6
17	2573	2557	16	99.4	0.6	43753	41392	2361	16.3	62.8	46.5	17.5
18	2574	2554	20	99.2	0.8	46327	43946	2381	17.3	63.3	46.0	18.5
19	2574	2560	14	99.5	0.5	48901	46506	2395	18.3	63.7	45.4	19.4
20	2574	2555	19	99.3	0.7	51475	49061	2414	19.3	64.2	44.9	20.3

Out-Of-Time	# Records 142093		# Goods 140059		# Bads 2034	Fraud Rate 0.0143					
	Bin Statistics					Cumulative Statistics					
Population Bin %	# Records	# Goods	# Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Good	% Bad (FDR)	KS	FPR
1	1421	352	1069	24.8	75.2	1421	352	1069	0.3	52.6	52.3	0.3
2	1421	1389	32	97.7	2.3	2842	1741	1101	1.2	54.1	52.9	1.6
3	1421	1412	9	99.4	0.6	4263	3153	1110	2.3	54.6	52.3	2.8
4	1421	1410	11	99.2	0.8	5684	4563	1121	3.3	55.1	51.9	4.1
5	1421	1416	5	99.6	0.4	7105	5979	1126	4.3	55.4	51.1	5.3
6	1421	1407	14	99.0	1.0	8526	7386	1140	5.3	56.0	50.8	6.5
7	1421	1403	18	98.7	1.3	9947	8789	1158	6.3	56.9	50.7	7.6
8	1420	1408	12	99.2	0.8	11367	10197	1170	7.3	57.5	50.2	8.7
9	1421	1410	11	99.2	0.8	12788	11607	1181	8.3	58.1	49.8	9.8
10	1421	1405	16	98.9	1.1	14209	13012	1197	9.3	58.8	49.6	10.9
11	1421	1410	11	99.2	0.8	15630	14422	1208	10.3	59.4	49.1	11.9
12	1421	1410	11	99.2	0.8	17051	15832	1219	11.3	59.9	48.6	13.0
13	1421	1412	9	99.4	0.6	18472	17244	1228	12.3	60.4	48.1	14.0
14	1421	1409	12	99.2	0.8	19893	18653	1240	13.3	61.0	47.6	15.0
15	1421	1407	14	99.0	1.0	21314	20060	1254	14.3	61.7	47.3	16.0
16	1421	1412	9	99.4	0.6	22735	21472	1263	15.3	62.1	46.8	17.0
17	1421	1413	8	99.4	0.6	24156	22885	1271	16.3	62.5	46.1	18.0
18	1421	1410	11	99.2	0.8	25577	24295	1282	17.3	63.0	45.7	19.0
19	1421	1411	10	99.3	0.7	26998	25706	1292	18.4	63.5	45.2	19.9
20	1421	1416	5	99.6	0.4	28419	27122	1297	19.4	63.8	44.4	20.9

Recommendation

As a result, we recommend that Client A rejects top 2% risky applications based on predictions from the final model. Assuming $6000 cost savings for each fraud successfully caught and $50 loss for each false positive, we plotted the overall savings at cut-off from 1% to 20%. At 2%, overall savings is maximized at $ 6,518,950 and the client does not risk rejecting too many applicants.



8 Conclusion

In this section, we briefly reflect on the steps we took to arrive at the results above.

Process flow

- **Data exploration**. Exploratory data analysis was done and Data Quality Report produced.
- **Data cleaning**. Outliers were removed and missing values filled in. Placeholder values were replaced by record number.
- **Candidate variable creation**. We created 99 candidate variables out the original 10 fields.
- **Feature selection**. We selected the top 20 candidate variables in terms of predicted power.
- **Fraud model development**. An Adaptive Boosting model was chosen to be the final model among 5 candidate models
- **Results / recommended actions**. A 2% cut-off on predicted fraud likelihood was recommended for fraud classification.

Limitations

- Some of the frauds are always missed out even at loose thresholds around 20%. This is a possible indication that our expert variables do not characterize some of the fraud modes present in the data.
- The wrapper feature selection process was conducted on logistic regression, which is a linear algorithm. Some important non-linear relationships might not be sufficiently captured by wrapping with logistic regression.

Appendix

Data Quality Report

Data Description

This data contains credit card application information in 2016, for purpose to calculate the frequency of credit card fraud, collect characteristic of fraud applications and predict future fraud activities. This dataset provides data by date, with 9 meaningful fields and 1,000,000 records.

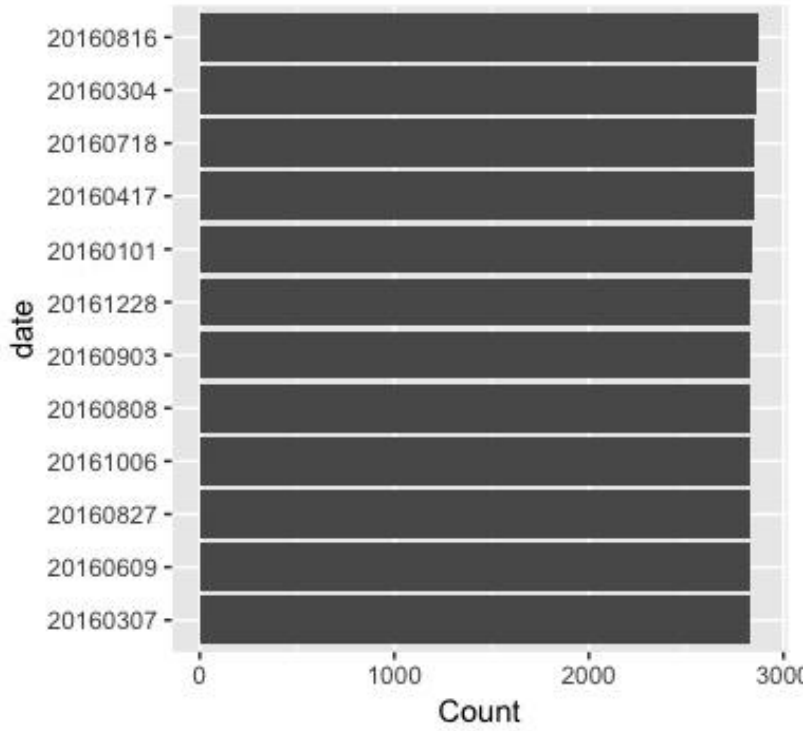
Table of Fields

	Field Type	Count	% Populated	# Unique	Top	Freq
date	Categorical	1 million	100%	365	20160816	2877
ssn	Categorical	1 million	100%	835819	999999999	16935
firstname	Categorical	1 million	100%	78136	EAMSTRMT	12658
lastname	Categorical	1 million	100%	177001	ERJSAXA	8580
address	Categorical	1 million	100%	828774	123 MAIN ST	1079
zip5	Categorical	1 million	100%	26370	68138	823
dob	Categorical	1 million	100%	42673	19070626	126568
homephone	Categorical	1 million	100%	28244	9999999999	78512
fraud_label	Categorical	1 million	100%	2	0	985607

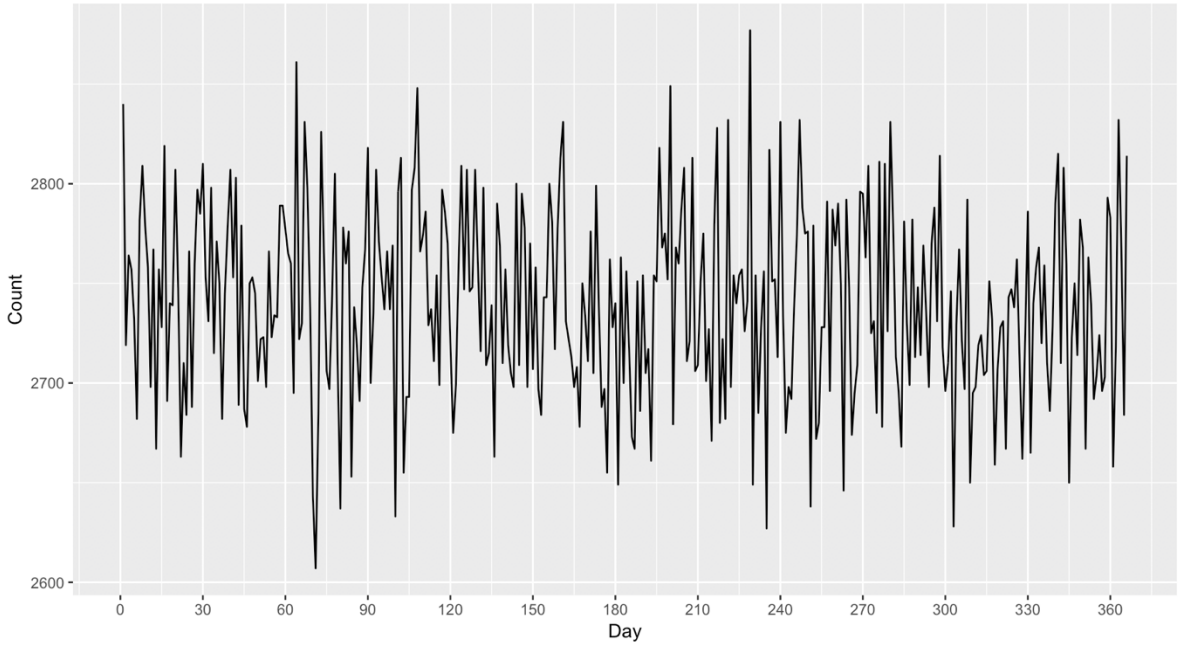
Field Description

Date

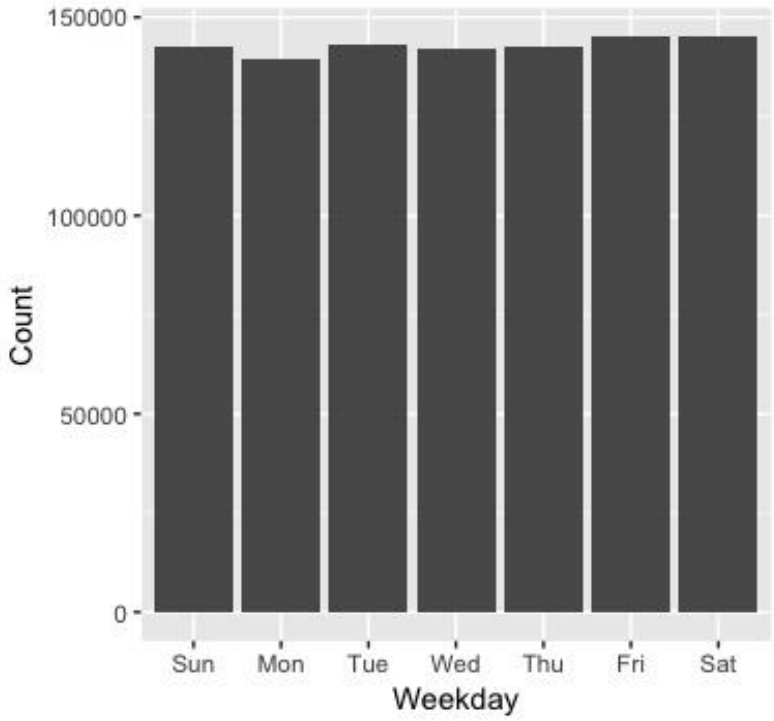
This field is categorical, each record of it has a value. It has 365 unique values. It represents applications in all 365 days. The most common date is 20160816, and the frequency is 2,877.



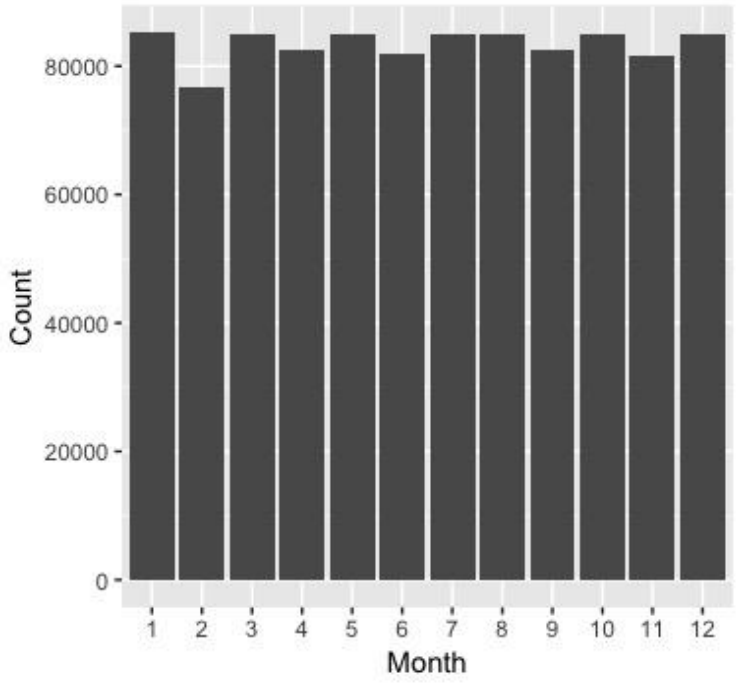
Number of applications by day.



Number of applications by weekday.



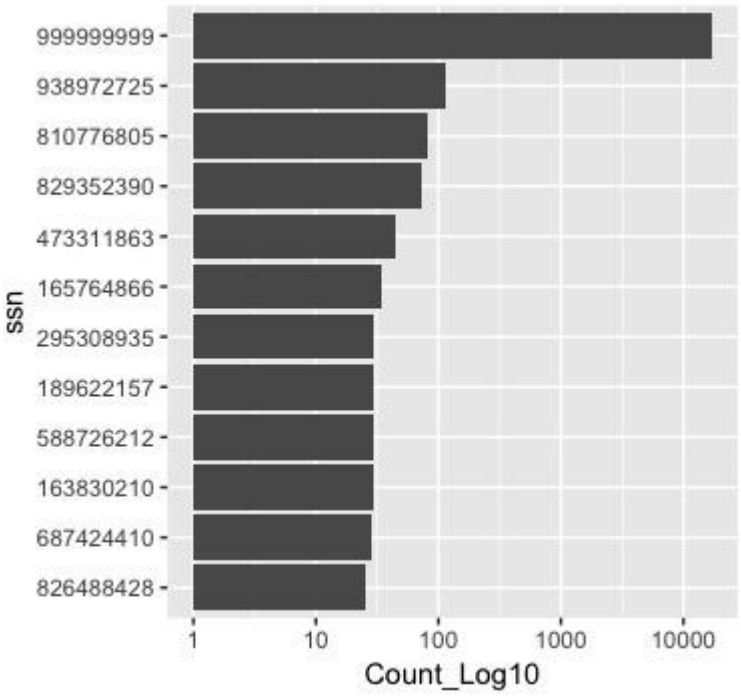
Number of applications by month.



SSN

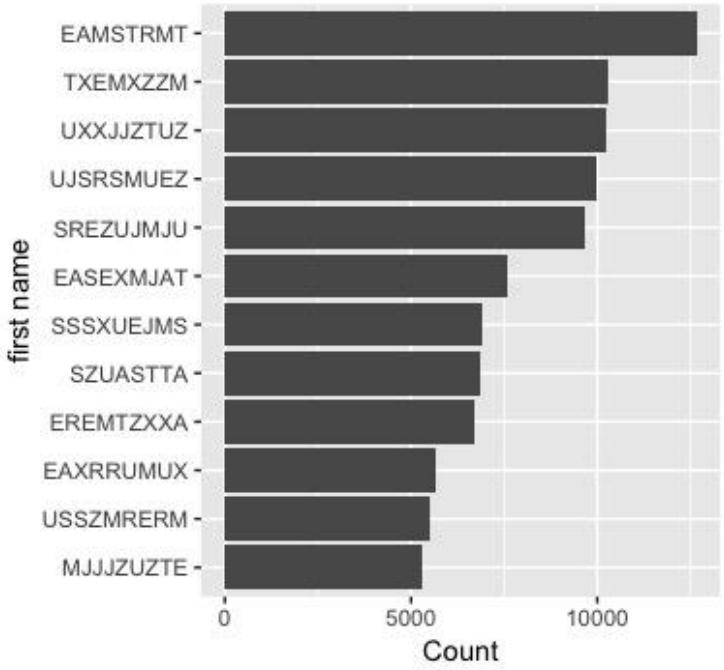
This field is categorical, each record of it has a value. It has 835,819 unique values. There is no missing value is this field. The most common ssn number filling in this field is 999999999, happened 16935 times.

This field could be very useful for fraud detection, especially target on the ones who didn't give the real ssn number.



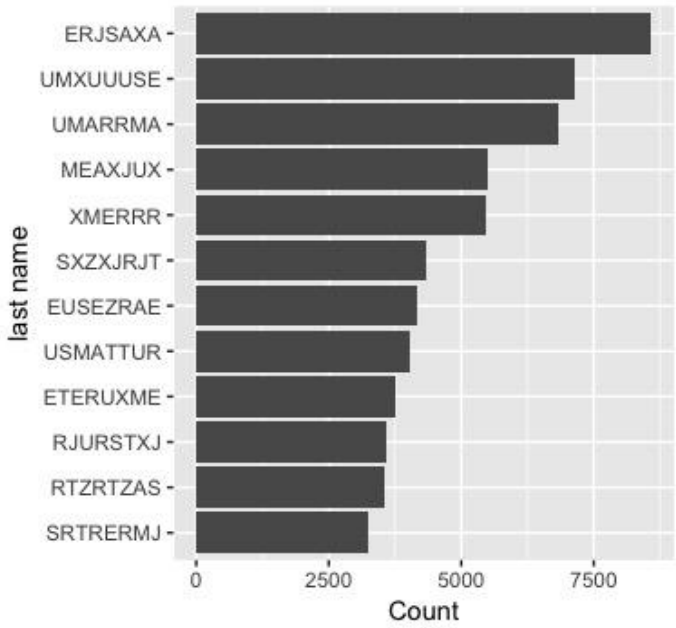
Firstname

This field is categorical, each record of it has a value. It has 78,136 unique values. There is no missing value is this field. The most common name in this field is EAMSTRMT, the frequency is 12,658.



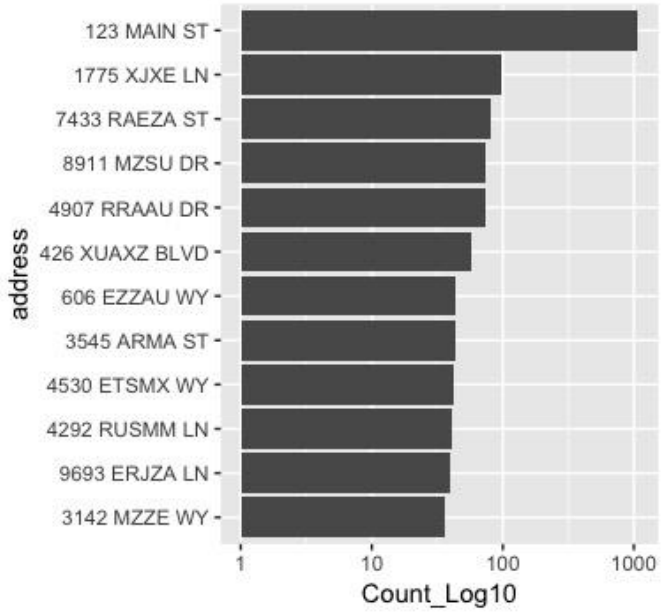
Lastname

This field is categorical, each record of it has a value. It has 177,001 unique values. There is no missing value is this field. The most common name in this field is ERJSAXA, the frequency is 8,580.



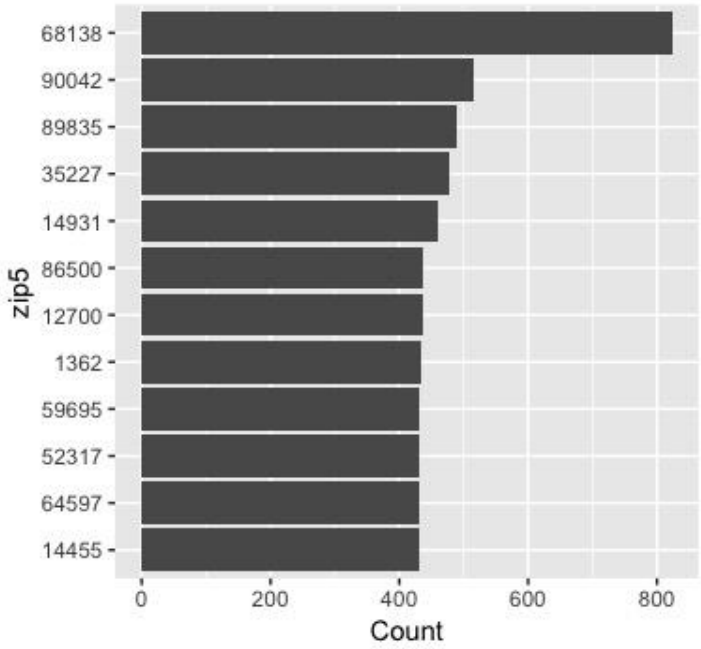
Address

This field is categorical, each record of it has a value. It has 828,774 unique values. There is no missing value is this field. The most common address is 123 MAIN ST, the frequency is 1,079.



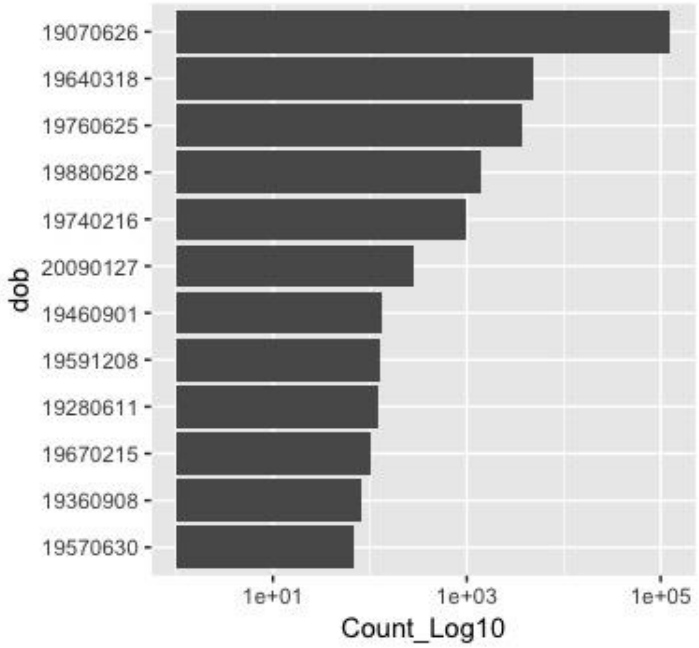
Zip5

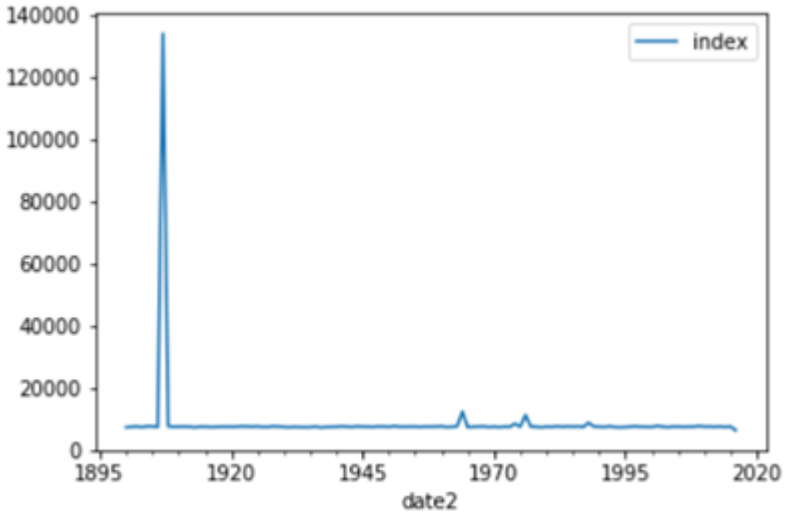
This field is categorical, each record of it has a value. It has 26,370 unique values. There is no missing value is this field. The most common zip code is 68138, the frequency is 823



Dob

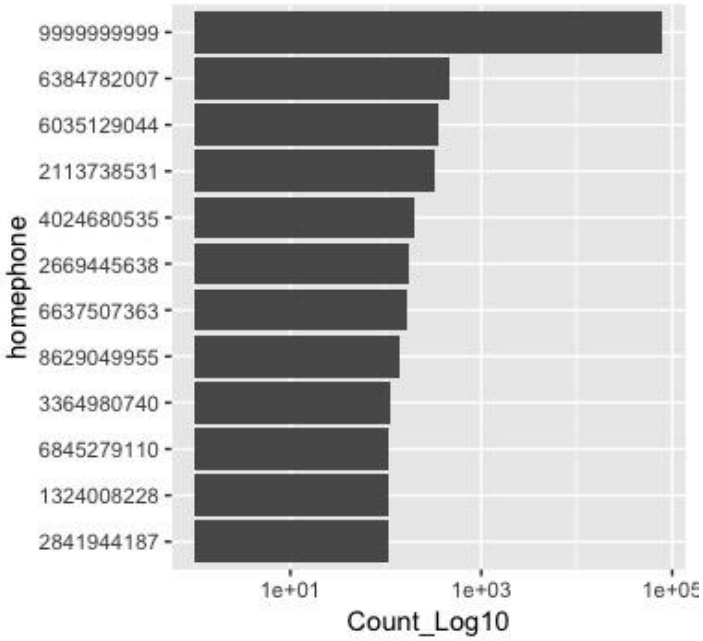
This field is categorical, each record of it has a value. It has 42,673 unique values. There is no missing value is this field. The most common date of birth is 19070626, the frequency is 12,6568, which is abnormal.





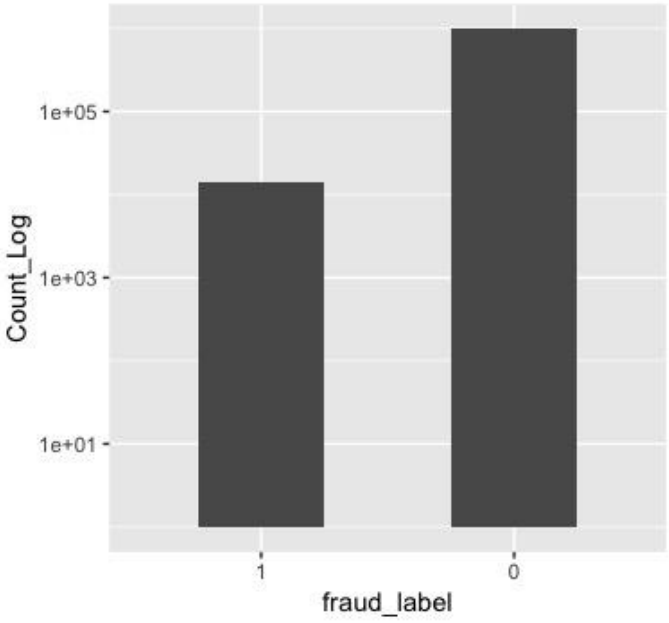
Homephone

This field is categorical, each record of it has a value. It has 28,244 unique values. There is no missing value is this field. The most common phone number is 9999999999, the frequency is 78,512.

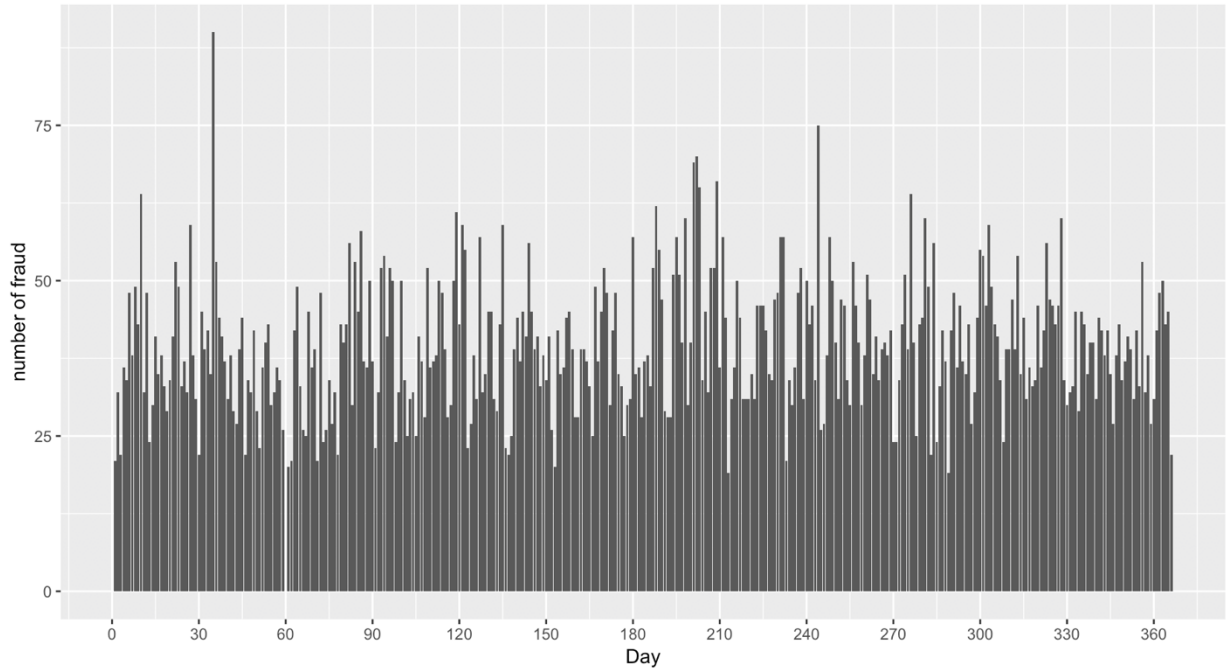


Fraud_label

This field is categorical, each record of it has a value. It has only 2 unique values. There are 14,393 fraud applications. The fraud rate is 1.44%.



Number of fraud applications by day.



Number of fraud applications by month.