DETECTING FRAUD

In the Productions Application Dataset

DSO 562 - Team 4

Chong Li, Jie Chen, Raman Deep Singh, Xiaowen Zhang, Yu Dong

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This report provides an analysis and evaluation of Application Data for detecting fraud using unsupervised machine learning methods. The tools used are R and Tableau, and methods for analysis include Principal Component Analysis and Autoencoder.

The original data set contains 100,000 product application records with 9 variables of applicants' personal information. The general process of analysis follows building expert variables with different time windows, standardization and dimensionality reduction, applying fraud algorithm, calculating fraud score, and identifying potential fraud.

Using heuristic fraud algorithm and autoencoder, a fraud score is calculated for each record. Records with high scores are determined to be potentially fraudulent. On a closer look, we find the high score records of two unsupervised machine learning methods are overlapped, and determines that the overlapped part of the top records from both methods are highly likely to be fraudulent.

Detailed examination of the most suspicious records indicates that potentially fraudulent applicants usually apply multiple times within days using different names, addresses and social security number, but not necessarily using different phone numbers. Further, examination of the top 10 most suspicious records shows that a same phone number is generally being used multiple times for applications with various names, addresses and social security numbers.

2. DATA OVERVIEW

Production Application Data is a dataset containing the records of 100,000 product applications. It includes information about the date the application was made, and the SSN, name, address, zip code, date of birth and home phone number that each applicant self-reported. All variables were masked so they appear to be different from real data.

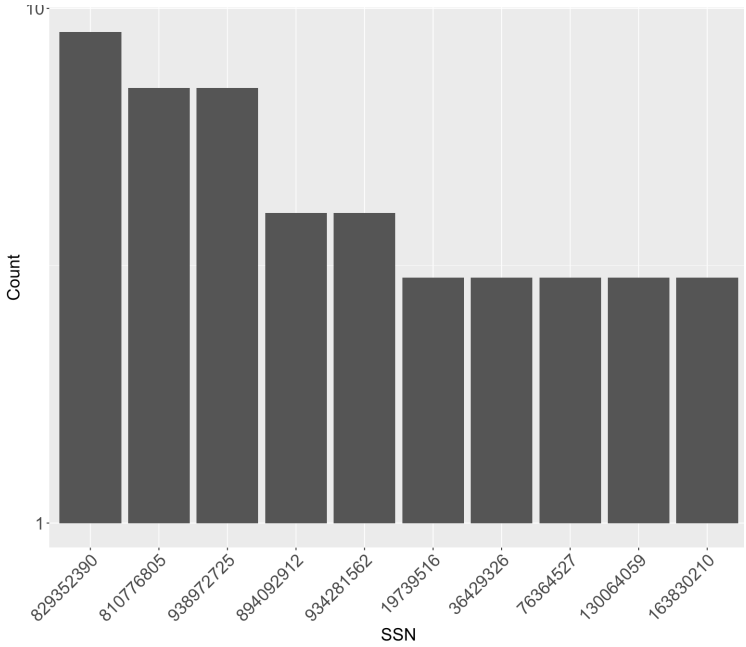
Following is description of the variables we consider to be the most important. The complete Data Quality Report can be found in appendix.

DESCRIPTION OF IMPORTANT VARIABLES

Field Name: ssn
"ssn" is a categorical variable, indicating the applicant's ssn number he or she self-reported. 100% populated with 96,535 unique values. Values in this field have lengths from 4 to 9. A ssn shorter than 9 digits indicates there are leading 0s in the ssn record. For example, ssn record "2503" is actually "000002503". The top 10 frequent appeared ssn records are shown below. It is obvious that the ssn record "737610282" is a frivolous record, which appears 173 times more than the second highest appeared record.
The top 10 most frequent records are listed below:

SSN	Frequency
737610282	1.73%
938972725	0.01%
810776805	0.01%
829352390	0.01%
473311863	0.00%
189622157	0.00%
163830210	0.00%
407447121	0.00%
118692079	0.00%
849295926	0.00%

The distribution of the top 10 records (excluding the most frequent record "737610282"):



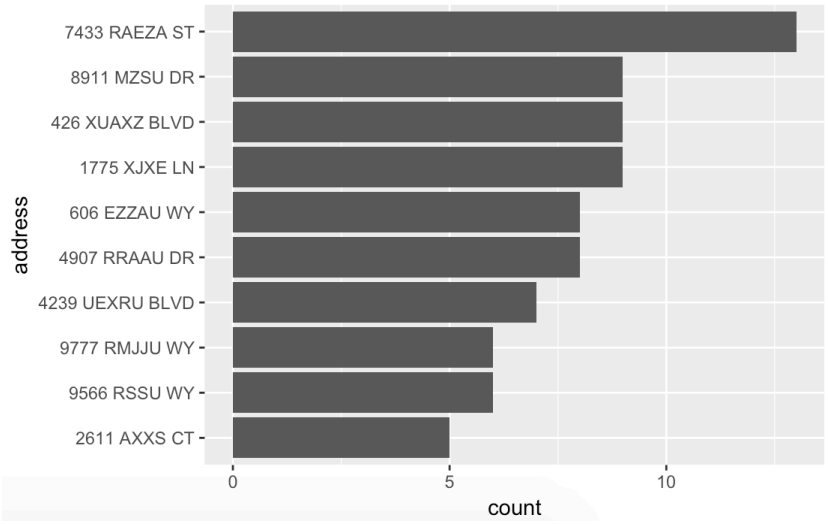
Field Name: address

"address" is a text variable that contains inputs from applicants of their address. 100% populated with 97,563 unique values from 100,000 records. The most frequent address "2602 AJTJ AVE" appeared 117 times, which accounts for 0.12% of all records. This address is likely to be a frivolous address.

The top 10 most frequent records are listed below:

address	Frequency
2602 AJTJ AVE	0.12%
7433 RAEZA ST	0.01%
1775 XJXE LN	0.01%
426 XUAXZ BLVD	0.01%
8911 MZSU DR	0.01%
4907 RRAAU DR	0.01%
606 EZZAU WY	0.01%
4239 UEXRU BLVD	0.01%
9566 RSSU WY	0.01%
9777 RMJJU WY	0.01%

The distribution of the top 10 records (excluding the most frequent record "2602 AJTJ AVE"):



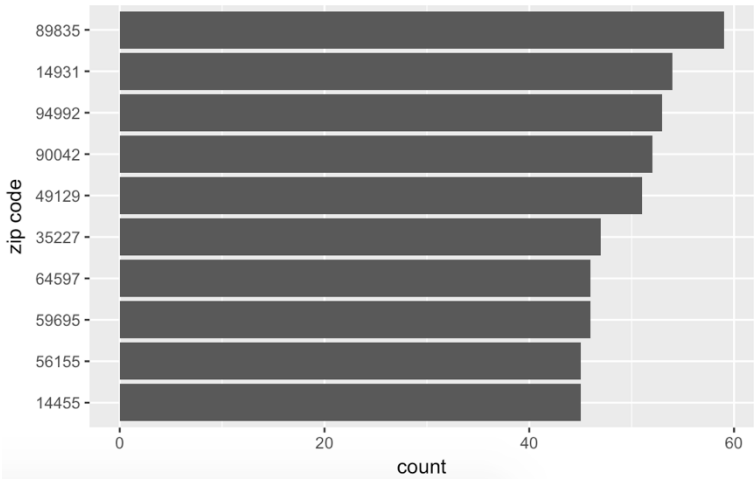
Field Name: zip5

"zip5" is a categorical variable that contains inputs from applicants of their zip codes. 100% populated with 16,547 unique values from 100,000 records. No missing values. The most frequent zip code "68138" appeared 823 times, which accounts for 0.09% of all records. This zip code is likely to be a frivolous value.

The top 10 most frequent records are listed below:

Zip5	Frequency
68138	0.09%
89835	0.06%
14931	0.05%
94992	0.05%
90042	0.05%
49129	0.05%
35227	0.05%
59695	0.05%
64597	0.05%
14455	0.05%

The distribution of the top 10 records (excluding the most frequent record "68138"):



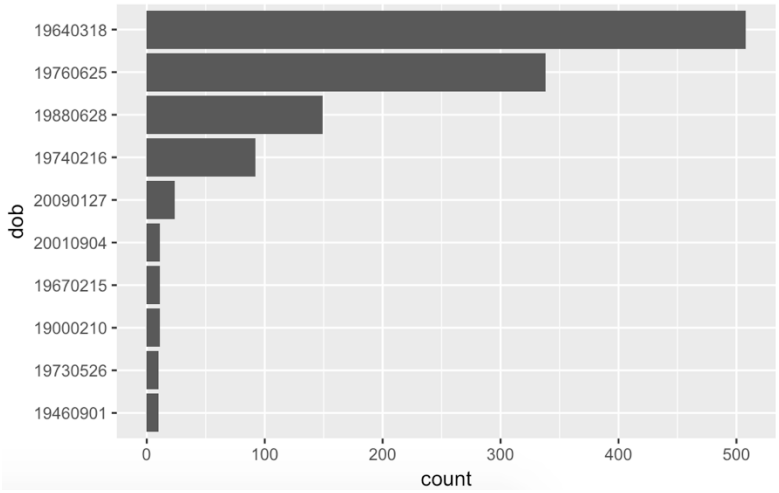
Field Name: dob

"dob" is a date variable that contains inputs from applicants of their date of birth. 100% populated with 36,816 unique values from 100,000 records. No missing values. The most frequent date of birth "19070626" appeared 12,488 times, which accounts for 12.49% of all records. This date of birth is likely to be a frivolous value.

The top 10 most frequent records are listed below:

dob	Frequency
19070626	12.49%
19640318	0.51%
19760625	0.34%
19880628	0.15%
19740216	0.09%
20090127	0.02%
19000210	0.01%
19670215	0.01%
20010904	0.01%
19460901	0.01%

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



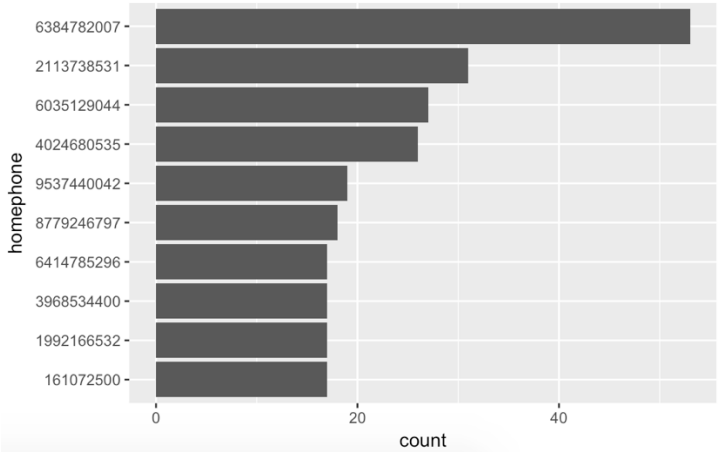
Field Name: homephone

"homephone" is a categorical variable that contains inputs from applicants of their home phone number. 100 populated with 22,181 unique values from 100,000 records. No missing values. The most frequent home phone number "9105580920" appeared 7735 times, which accounts for 7.74% of all records. This homephone number is likely to be a frivolous value.

The top 10 most frequent records are listed below:

homephone	Frequency
9105580920	7.74%
6384782007	0.05%
2113738531	0.03%
6035129044	0.03%
4024680535	0.03%
9537440042	0.02%
8779246797	0.02%
161072500	0.02%
1992166532	0.02%
3968534400	0.02%

The distribution of the top 10 records (excluding the most frequent record "9105580920"):



3. VARIABLE CONSTRUCTION

To make the analysis more in depth and meaningful we modified some of the existing variables before we built our expert variables.

1. We built a new variable "**PERSON**", which is the concatenation of the existing variables "firstname", "lastname" and "dob".

2. We built a new variable "**ADD_ZIP**", which is the concatenation of the existing variables "address" and "zip5".

Since this analysis involves time, we chose two different time windows, **3 days and 7 days**. The rationale is to capture more fraudulent applications that might be different in those time windows.

We then built two types of variables:

1) **Type I variables** are the count of one entity associated with specific value of another entity in a particular time frame. The count of Type I variables is dependent on the time window and hence they are created separately for 3-day and 7-day time window. There are a total of 12 Type I variables.

Example: PERSON_SSN tells the number of PERSONs relate with one SSN.

2) **Type II variables** are the count of an entity in a time frame. The count of Type II variables is also dependent on the time window and hence they are also created separately for 3-day and 7-day time window. There are a total of 4 Type II variables.

Example: TIME_PERSON tells the number of times a particular person ("PERSON") appears in a 3-day time window or a 7-day time window.

TYPE I VARIABLES

1. **PERSON_SSN**: the number of people ("PERSON") related with one particular ssn in a 3-day time window or a 7-day time window.

2. **PERSON_PHONE**: the number of people ("PERSON") related with one particular phone number in a 3-day time window or a 7-day time window.

3. **PERSON_ADD**: the number of people (“PERSON”) related with one particular address and zip combination (“ADD_ZIP”) in a 3-day time window or a 7-day time window.

4. **SSN_PERSON**: the number of ssns related with one particular person (“PERSON”) in a 3-day time window or a 7-day time window.

5. **SSN_PHONE**: the number of ssns related with one particular phone number in a 3-day time window or a 7-day time window.

6. **SSN_ADD**: the number of ssns related with one particular address and zip combination (“ADD_ZIP”) in a 3-day time window or a 7-day time window.

7. **PHONE_PERSON**: the number of phone numbers related with one particular person (“PERSON”) in a 3-day time window or a 7-day time window.

8. **PHONE_SSN**: the number of phone numbers related with one particular ssn in a 3-day time window or a 7-day time window.

9. **PHONE_ADD**: the number of phone numbers related with one particular address and zip combination (“ADD_ZIP”) in a 3-day time window or a 7-day time window.

10. **ADD_PERSON**: the number of address and zip combinations (“ADD_ZIP”) related with one particular person (“PERSON”) in a 3-day time window or a 7-day time window.

11. **ADD_SSN**: the number of address and zip combinations (“ADD_ZIP”) related with one particular ssn in a 3-day time window or a 7-day time window.

12. **ADD_PHONE**: the number of address and zip combinations (“ADD_ZIP”) related with one particular phone number in a 3-day time window or a 7-day time window.

TYPE II VARIABLES

1. **TIME_PERSON**: the number of times a particular person (“PERSON”) appears in a 3-day time window or a 7-day time window.

2. **TIME_SSN**: the number of times a particular ssn appears in a 3-day time window or a 7-day time window.

3. **TIME_PHONE**: the number of times a particular phone number appears in a 3-day time window or a 7-day time window.

4. **TIME_ADD_ZIP**: the number of times a particular address and zip combination (“ADD_ZIP”) appears in a 3-day time window or a 7-day time window.

We wrote “for loops” in R to count the number of times a variable appeared in a time window and wrote functions to repeat the process for all the expert variables.

HANDLING FRIVOLOUS VALUES

The expert variables related to "**ssn**", "**homephone**", and "**ZIP_ADD**" were affected by frivolous values. The frivolous value in "**dob**" did not affect our analysis because we built a variable "**PERSON**" instead of "**dob**" for identification of an applicant. In this light, "**PERSON**" is not affected by frivolous anymore though it contains frivolous '**dob**'.

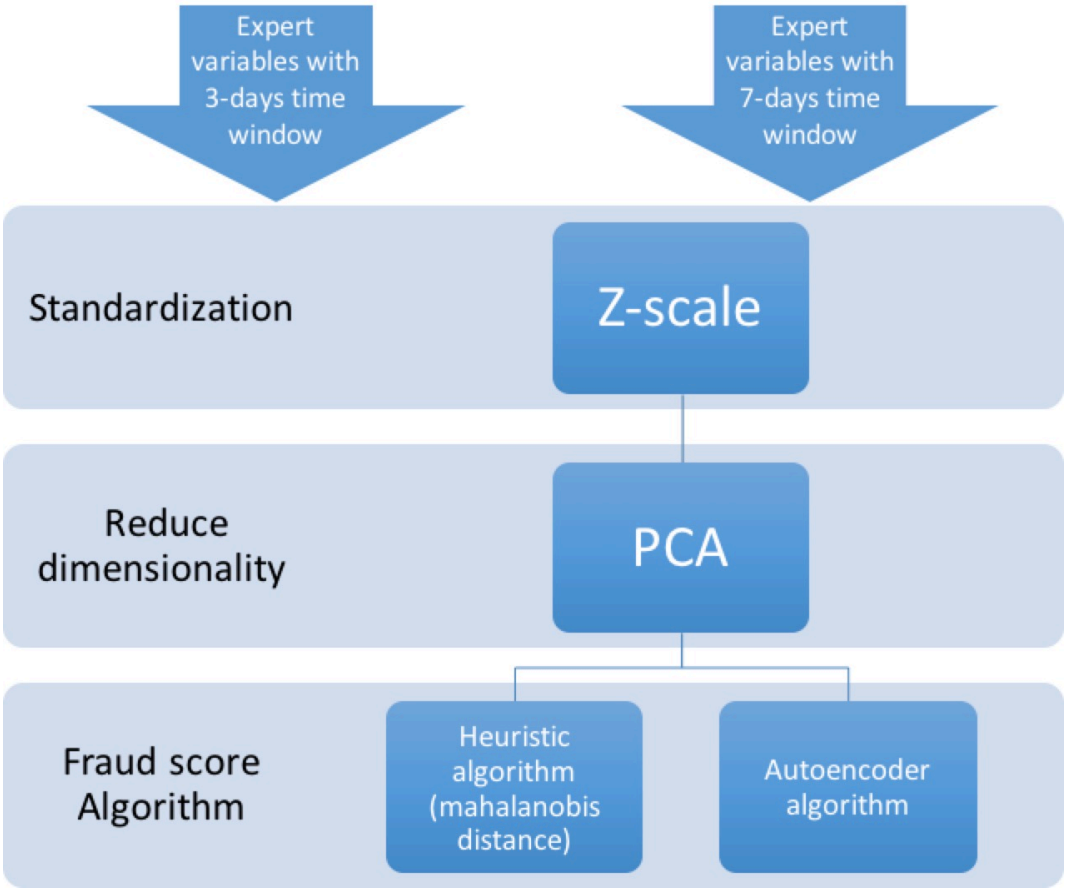
Corrective Action: To eliminate the effects posed by frivolous values, we replaced the affected records in the expert variable with the mean of all the records (excepting records affected by frivolous values) in that specific expert variable.

For example, for the expert variable "**PERSON_SSN**" we calculated the average counts of records whose SSNs were not "737610282" (the frivolous SSN). We then replaced the values of "PERSON_SSN" of those records with frivolous SSN by that average.

4. FRAUD ALGORITHM

PRINCIPAL COMPONENT ANALYSIS (PCA)

After creating expert variables, we did standardization and dimensionality reduction. To do this, we ran the PCA algorithm with both 3-day time window and 7-day time window respectively.



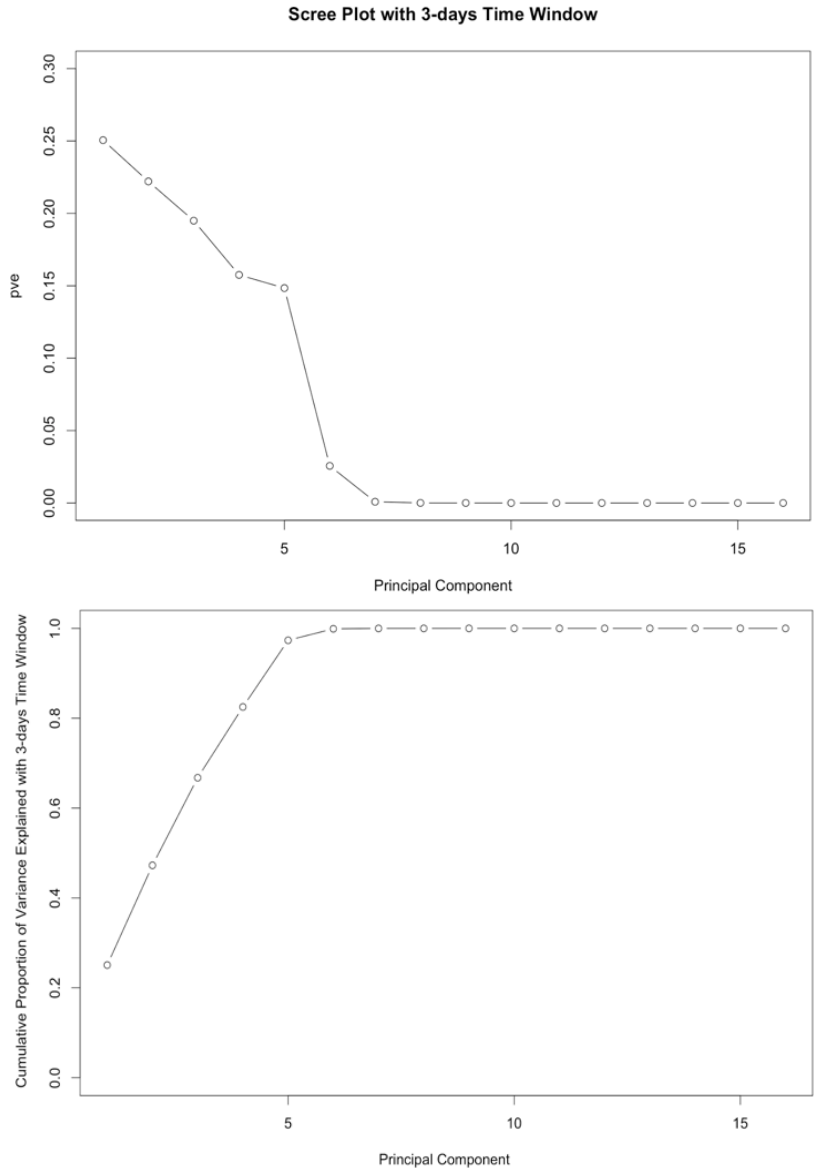
We performed principal components analysis using the **prcomp**() function, which is one of several functions in R that perform PCA. By default, the **prcomp**() function centers the variables to have mean zero. By using the option **scale = TRUE**, we scaled the variables to have standard deviation of 1. The '**center'** and '**scale**' components correspond to the means and standard deviations of the variables that were used for standardization prior to implementing PCA. Additionally, this z-scaling made all those variables associated with frivolous values into the value zero.

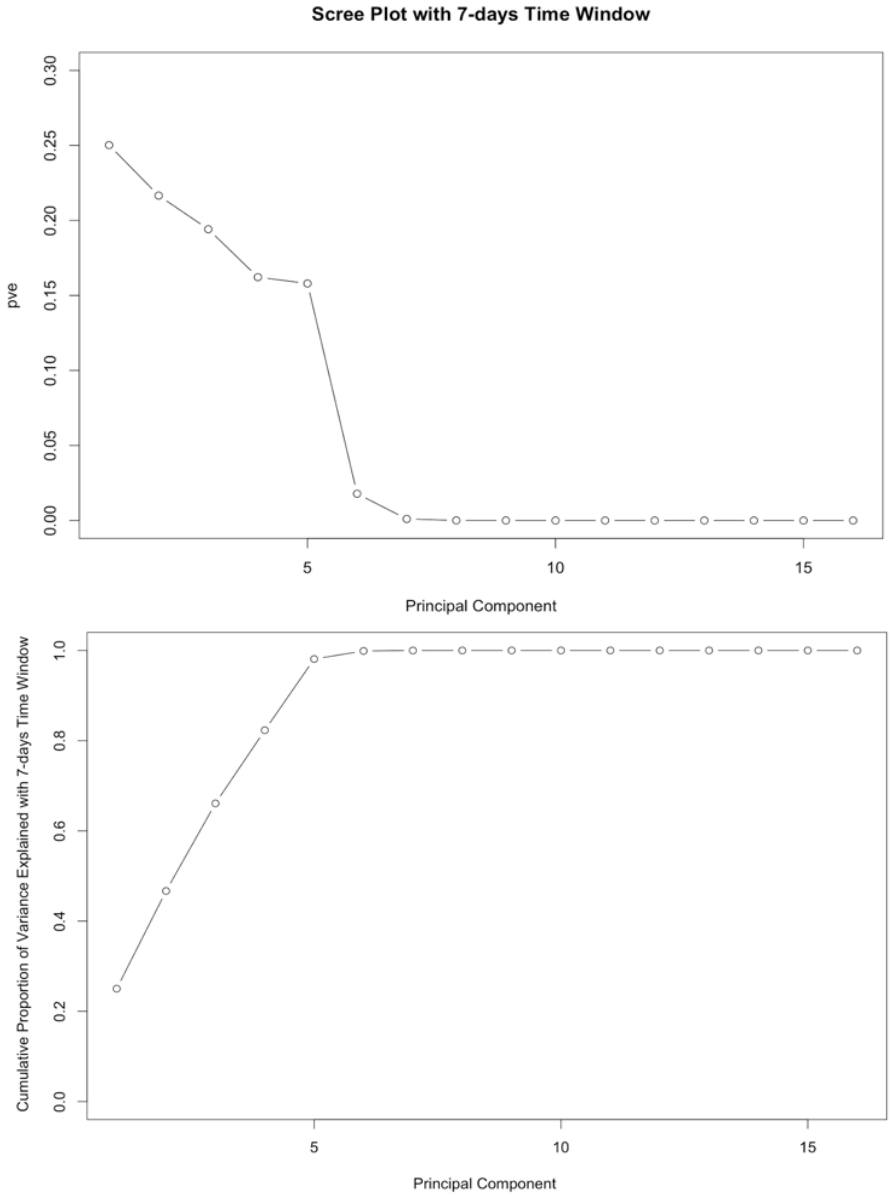
The **rotation** matrix provided the principal component loadings, with each column of **pr.out$rotation** containing the corresponding principal component loading vector.

To compute the proportion of variance explained by each principal component, we divided the variance explained by each PC by the total variance. We made the scree plot and cumulative plot to determine which PCs to keep. We would like to use the smallest number of PCs required to get a good understanding of the data.

For the 3-day time window analysis, we discovered that there is a significant drop(‘**elbow**’) between PC6 and PC7 according to the scree plot. Thus, we decided to keep PC1 through PC6, which explained approximately 99.9% of variance of the entire dataset.

For the 7-day time window analysis, we discovered that it has similar pattern with 3-day time window: a significant drop(‘**elbow’**) between PC6 and PC7. Thus, we decided to keep PC1 through PC6, accounting for approximately 99.9% of the variance of the entire dataset.





AUTOENCODER

We auto-encoded our PCs using an R package called "**h2o**". Then, we called the **deep learning** function with parameter "autoencoder" set to TRUE. This function took the dataset after PCA and auto-encoded it. The neural network in our algorithm contained **4 hidden layers** with **100 neurons** in each hidden layer. We then called the **h2o.anomaly** function to reconstruct the original dataset using the reduced set of features and calculated a mean squared error between both. We set the "**per_feature**" parameter to TRUE because we wanted a reconstruction mean error based on individual features. We saved the reconstruction error in a dataset called "**error**".

In the end, we summed the reconstruction error values of all the PCs to get a single score, which became our fraud score from autoencoder, for each of the record. Also, we repeated the process for both 3-day time window and 7-day time window separately.

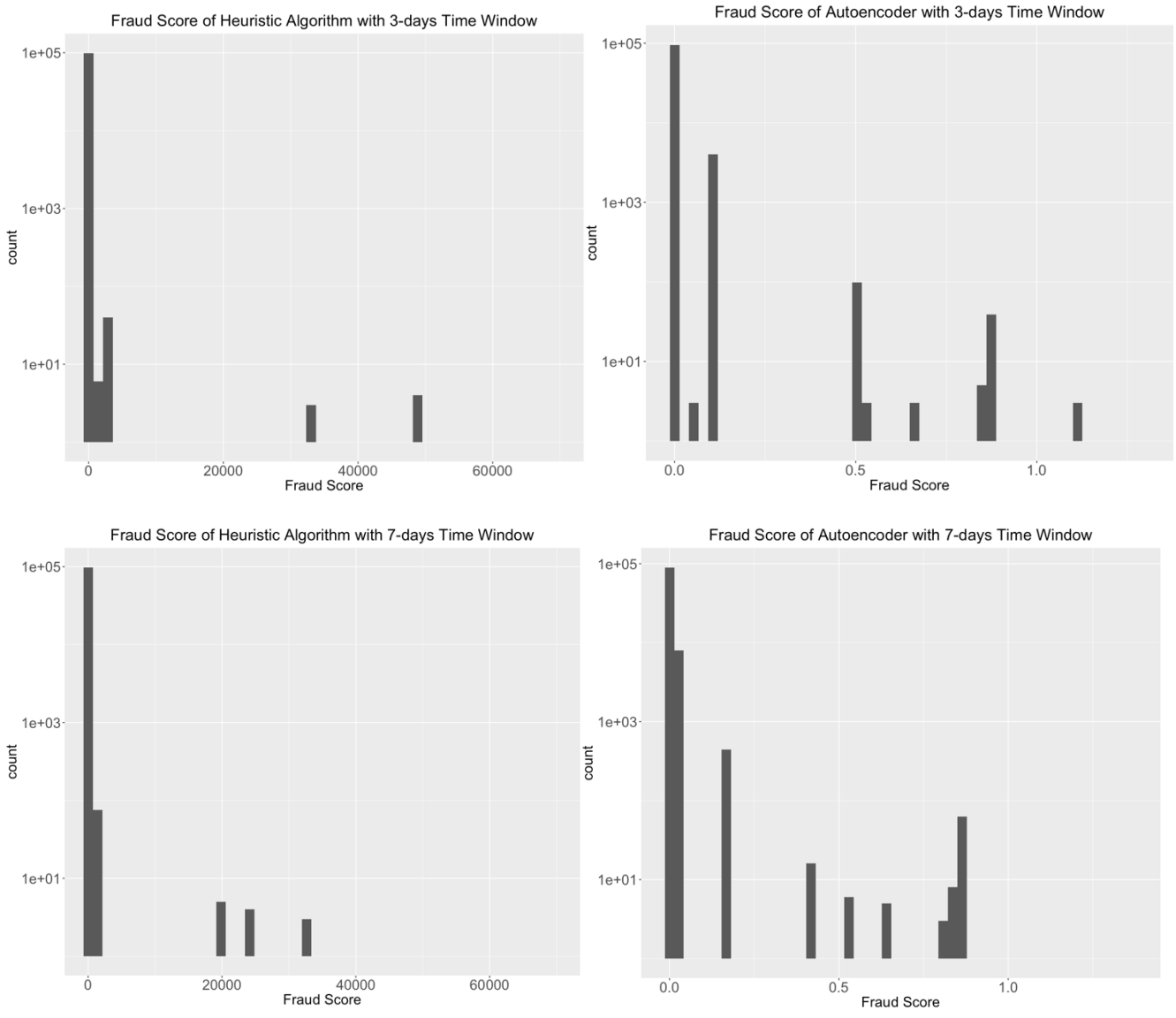
HEURISTIC ALGORITHM

Another way to calculate the fraud score is by the heuristic algorithm. We calculated the **mahalanobis distance** between each record and the (mean, covariance) of records within each particular PC. The mahalanobis distance was our fraud score for each record. It was calculated using the function '**mahalanobis**' in R. Again, we ran the algorithm for the 3-day time window and the 7-day time window separately.

5. RESULTS

We sorted the records according to fraud scores from both **autoencoder** and **heuristic algorithm**. The majority of records had low fraud scores while a small proportion of the records had high fraud scores.

Below is an overview of what the distribution of fraud scores looks like from both methods with both time windows:



All the fraud score distribution appeared kind of **discrete** with several small peaks. This is because all the expert variables were counts in nature, which only takes integers from around 1 to 10. Also, many records had identical values, since their values of expert variables were simply "1 1 1 1 1 …". Therefore, after the standardization and PCA algorithm, these records still had the same value.

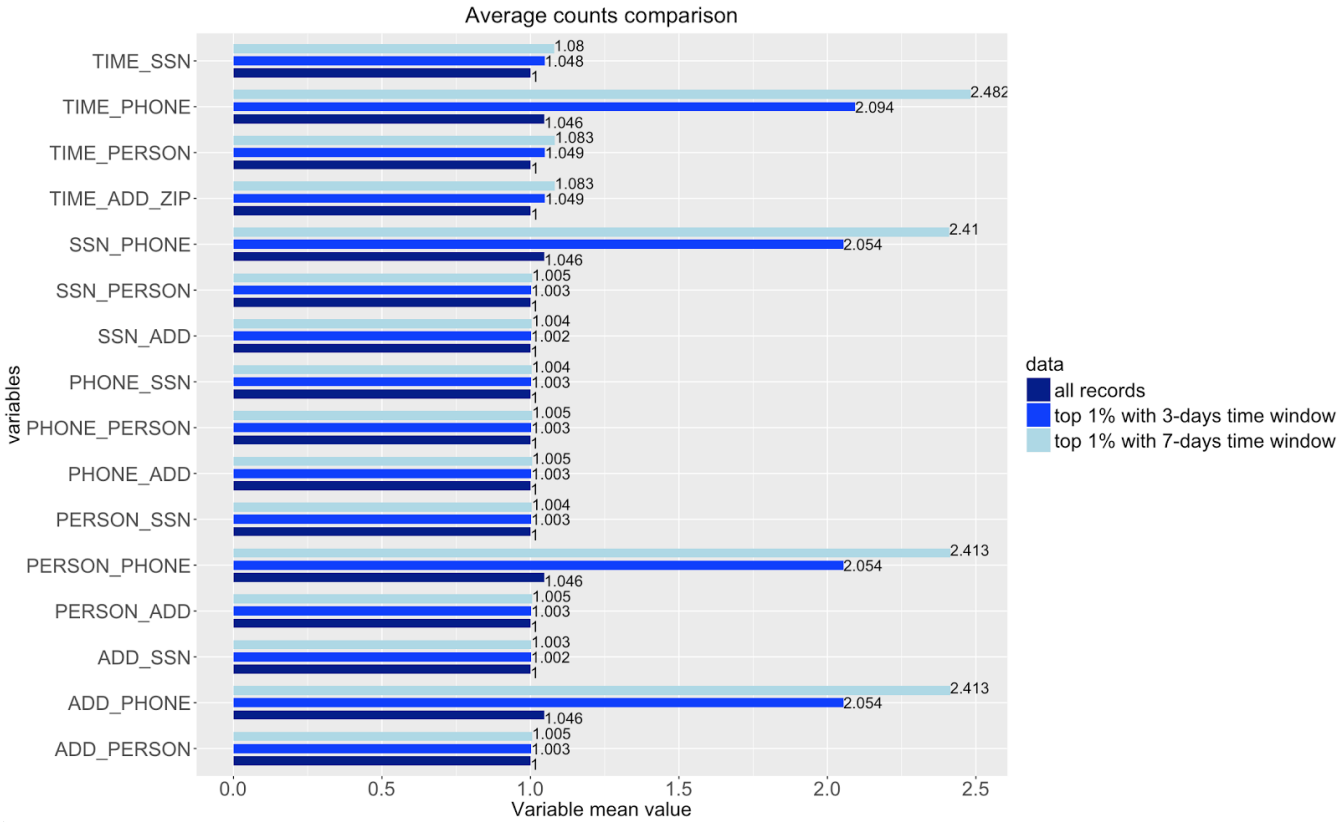
We decided to look at the **overlapping part** of the **top 1%** high scoring records from autoencoder output and the **top 1%** high scoring records from heuristic algorithm output. For both 3-days' time window outputs and 7-days' time window outputs, almost **100%** of the records from the two algorithms match. This could also be explained by the above distribution plot -- a certain small group of the records was highly suspicious and had high scores.

INSIGHTS

	all records			top 1% records for 3-days time window			top 1% records for 7-days time window		
Expert variables	mean	stdev	median	mean	stdev	median	mean	stdev	median
PERSON_SSN	1.000	0.005	1	1.003	0.055	1	1.004	0.064	1
PERSON_PHONE	1.046	0.206	1	2.054	0.399	2	2.413	0.680	2
PERSON_ADD	1.000	0.005	1	1.003	0.055	1	1.005	0.071	1
SSN_PERSON	1.000	0.005	1	1.003	0.055	1	1.005	0.071	1
SSN_PHONE	1.046	0.206	1	2.054	0.399	2	2.410	0.678	2
SSN_ADD	1.000	0.004	1	1.002	0.045	1	1.004	0.064	1
PHONE_PERSON	1.000	0.005	1	1.003	0.055	1	1.005	0.071	1
PHONE_SSN	1.000	0.005	1	1.003	0.055	1	1.004	0.064	1
PHONE_ADD	1.000	0.005	1	1.003	0.055	1	1.005	0.071	1
ADD_PERSON	1.000	0.005	1	1.003	0.055	1	1.005	0.071	1
ADD_SSN	1.000	0.004	1	1.002	0.045	1	1.003	0.055	1
ADD_PHONE	1.046	0.206	1	2.054	0.399	2	2.413	0.680	2
TIME_PERSON	1.000	0.022	1	1.049	0.217	1	1.083	0.275	1
TIME_SSN	1.000	0.022	1	1.048	0.215	1	1.080	0.271	1
TIME_PHONE	1.046	0.207	1	2.094	0.339	2	2.482	0.585	2
TIME_ADD_ZIP	1.000	0.022	1	1.049	0.217	1	1.083	0.275	1

In the above summary descriptive data table, we compared the mean, standard deviation and median among the whole dataset, top 1% high scoring records with 3-day time window, and top 1% high scoring records with 7-day time window. We can see that there are obvious differences between the whole dataset and the top 1% records. Typically, top 1% records have higher average values for all the expert variables we created.

We also noticed that the average values of expert variables with 3-day time window are always lower than those with 7-day time window. This is reasonable because with 7-day time window we can accumulate more counts for any expert variable.



When we take a closer look at all the mean values for expert variables as plotted above, we can see that there is larger deviation between the whole dataset and the top records for four of the expert variables -- **TIME_PHONE, SSN_PHONE, PERSON_PHONE and ADD_PHONE.**

TIME_PHONE is the frequency a certain phone number appears within the time window, while the other three represent the number of different SSN number, or personal information or address information associated with a certain phone number. Large values in these variables indicate a single phone number is used multiple times with different inputs in other fields.

We came up with two explanations for this pattern. The first explanation is that although people could use fake information to apply for a product, they still need a correct phone number to stay in contact. Thus, some people would make up different SSNs, addresses and names but still use their real phone numbers. The second explanation is that there could be members of illegal groups sharing a single phone number to conduct fraudulent activities.

IN-DEPTH ANALYSIS OF TOP 10 RECORDS

The below table shows the top 10 records for both time windows. Four records (record # in red) appeared in both top 10 lists, indicating high probability to be fraudulent. Besides, when we look at the application information, we can see that few records with frivolous values are included. This is because we have replaced all values involving frivolous values

with the average of other records without frivolous values before we calculated the fraud scores.

Time window	record #	date	ssn	firstname	lastname	address	zip5	dob	homephone
3 days	45115	6/13/15	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
	10982	2/9/15	3313643	XAEEAUJUS	UEMMSJMS	721 SAMZA DR	8106	19630429	2113738531
	88136	11/18/15	246788144	XJZJSAEXJ	RMZUSJR	6544 XJZXJ ST	15215	19070626	1069975274
	91874	12/2/15	705091123	EZXUMUUMR	ERRASRTX	6888 XJUXR LN	55194	19931208	6384782007
	53251	7/13/15	4178024	XSMZZJMZT	SZEZUUMX	7484 SSXUM LN	49101	19070626	7633541686
	2957	1/11/15	451558906	MSRUAREMU	EZATSJSU	8626 XXSTE PL	82999	19981120	1033418292
	6452	1/24/15	883562211	XRXTAMURS	AXUJZEE	893 RXZJE AVE	66108	19480108	9707343639
	6473	1/24/15	997039893	XEJMESRU	STSMJRUM	5994 SATTM CT	19031	20030521	1791599483
	7922	1/29/15	235296366	REZETXRXE	EEXSRMRT	9804 XRAMT ST	50573	19570812	1629142547
	12517	2/15/15	323723506	EAMSTRMT	EARJRZAE	2511 REMMU RD	63201	19430308	8598779413
7 days	45115	6/13/15	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
	93119	12/6/15	192546579	MERMESSXT	RATTUTXT	6125 XXSRS CT	46601	19070626	6384782007
	82618	10/28/15	51030882	ZZURRTU	SZSUTMRR	1597 UMAUE AVE	63473	19370410	3310797315
	4838	1/18/15	657877548	RXTSZJATS	RJURERXZ	3121 ERJZ WY	94535	19791116	3773220142
	6452	1/24/15	883562211	XRXTAMURS	AXUJZEE	893 RXZJE AVE	66108	19480108	9707343639
	6467	1/24/15	20782895	SMRAUMMMZ	UUEZSUTZ	7947 SUTUJ LN	1139	19950524	4584161412
	6473	1/24/15	997039893	XEJMESRU	STSMJRUM	5994 SATTM CT	19031	20030521	1791599483
	6698	1/25/15	271497366	EZMEJASZX	SRTAEMRJ	100 XXRAZ BLVD	52283	19570730	7182053816
	7922	1/29/15	235296366	REZETXRXE	EEXSRMRT	9804 XRAMT ST	50573	19570812	1629142547
	11127	2/10/15	618197293	RJSXUAJEE	SXZXJRJT	862 RJZ PL	55344	19860607	7346124998

Then we looked at the values of expert variables of these top records in the following table. We can see that many of the expert variable values are greater than 1.

Time window	record #	PERSON_SSN	PERSON_PHONE	PERSON_ADD	SSN_PERSON	SSN_PHONE	SSN_ADD	PHONE_PERSON	PHONE_SSN	PHONE_ADD	ADD_PERSON	ADD_SSN	ADD_PHONE	TIME_PERSON	TIME_SSN	TIME_PHONE	TIME_ADD_ZIP
3 days	45115	2	2	2	1	2	1	1	2	2	1	1	2	1	2	2	2
	10982	1	4	1	1	4	1	1	1	1	1	1	4	1	1	4	1
	88136	1	4	1	1	4	1	1	1	1	1	1	4	1	1	4	1
	91874	1	4	1	1	4	1	1	1	1	1	1	4	1	1	4	1
	53251	1	2	1	1	2	1	1	1	1	1	1	2	2	2	3	2
	2957	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6452	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6473	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	7922	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	12517	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
7 days	45115	2	3	2	1	3	1	1	2	2	1	1	3	1	2	3	2
	93119	1	6	1	1	6	1	1	1	1	1	1	6	1	1	6	1
	82618	1	3	1	1	3	1	1	1	1	1	1	3	2	2	4	2
	4838	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6452	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6467	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6473	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	6698	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	7922	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
	11127	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2

We also examined the above records one by one:

1. **Record number 45115:**

   The SSN of this application has been used twice with different personal information in past 3 or 7 days, and the phone number and address also have this pattern. This record is highly likely to be fraudulent, as someone fraudulently applying for products with different combinations of real and fake information.

2. **Record number 10982**

   The phone number in this application record has been used 4 times with different personal information, SSN and address. This could be caused by an illegal group of 4 sharing the same phone number, or someone trying to make application with his/her real phone number and different sets of fake information.

3. **Record number 88136**

   Just like in record number 10982, the phone number in this application record has been used 4 times with different personal information, SSN and address. This could be caused by an illegal group of 4 sharing the same phone number, or someone trying to apply with his/her real phone number and different sets of fake information.

4. **Record number 91784**

   Just like in record number 10982 and 88136, the phone number in this application record has been used 4 times with different personal information, SSN and address. This could be caused by an illegal group of 4 sharing the same phone number, or someone trying to apply with his/her real phone number and different sets of fake information.

   The three phone numbers in record 10982, 88136 and 91784 are very likely to be related to fraud in applications.

5. **Record number 53251**

   The phone number in this record has been associated with two different people, SSN and address in the past 3 days. Meanwhile, the personal information, SSN and address has also been seen twice in the past 3 days. This record is highly likely to be fraudulent, as someone fraudulently applying for products with different combinations of real and fake information.

6. **Record number 2957, 6452, 6473, 7922, 12517:**

   In each of these five records, the applicant used the same set of information (name, date of birth, SSN, phone number, address and zip code) and applied twice in the past 3 days. However, this may not be a signal of fraud. Maybe he or she didn't receive the application results and decided to apply again the next day. Or it could be due to Internet connection error.

7. **Record number 93119:**

   The phone number in this application record has been associated with six different people and SSN in the past seven days. But at the same time, it has only one address associated with it. This could be caused by a group of 6 (maybe a family or roommates) sharing the same home phone number, or someone trying to apply with his/her real phone number and different sets of fake information.

8. **Record number 82618:**

   The phone number in this application record has been associated with 3 different people and SSNs in the past 7 days. However, it has only one address associated with it. Meanwhile, the person, SSN and address information have all appeared twice in the past 7 days. The phone number has appeared four times. This is very likely to be caused by someone applying multiple times with different sets of fake information.

9. **Record number 4838, 6467, 6698, 11127:**

   In each of these four records, the applicant used the same set of information (name, date of birth, SSN, phone number, address and zip code) and applied twice in past 7 days. However, this may not be a signal of fraud. Maybe he or she just didn't receive the application results and decided to apply again a few days later. Or it could be due to Internet connection error.

6. APPENDIX

Product Applications Data

Data Quality Report

03/02/2017

Summary

File Description:

Production Application Data is a dataset containing the records of 100,000 product applications. It includes information about the date the application was made, and the SSN number, name, address, zip code, data of birth and home phone number that each applicant self-reports.

File Name:

Applications 100k.xlsx

Data Source:

This dataset is the simulated data based on real product application records.

Number of Records:

100,000 records

Number of Fields:

9 variables in total – 6 categorical variables, 1 text variable, 2 date variables

Field Name	Data Type
record	Categorical variable
date	Date variable
ssn	Categorical variable
firstname	Categorical variable
lastname	Categorical variable
address	Text variable
zip5	Categorical variable
dob	Date variable
homephone	Categorical variable

Time of Records:

Jan 1st 2015 – Dec 31st 2015

Fields Explanation

Field 1

Field Name: record

Description:
"record" is a categorical variable. It works as the ordinal reference number for each application record.

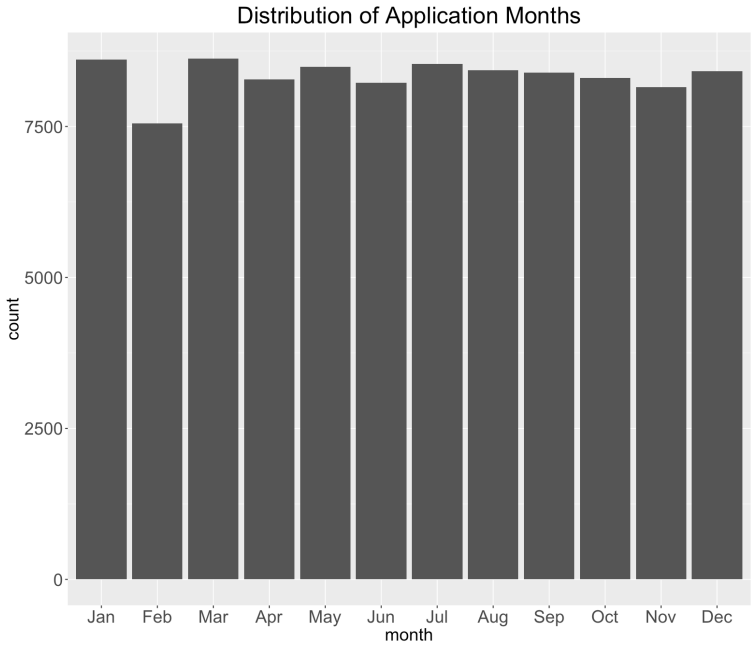
Unique Values:
100% populated with 100,000 unique values, ranging from 1 to 100,000. No repeated values or missing values exist.

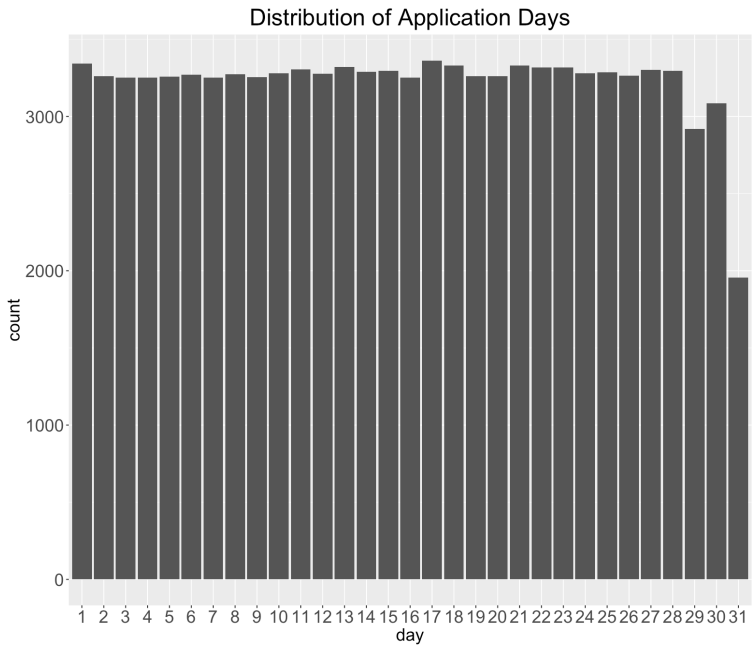
Field 2

Field Name: date

Description:
"date" is a date variable, indicating the date that the application was made.

Unique Values:
100% populated with 365 unique values, i.e. every day in year 2015. Plot the distribution of the months and dates as below, we can see that the two distributions follow the natural occurrence distribution.


Distribution of Application Months
count
7500
5000
2500
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


Distribution of Application Days
count
3000
2000
1000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day

Field 3

Field Name: ssn

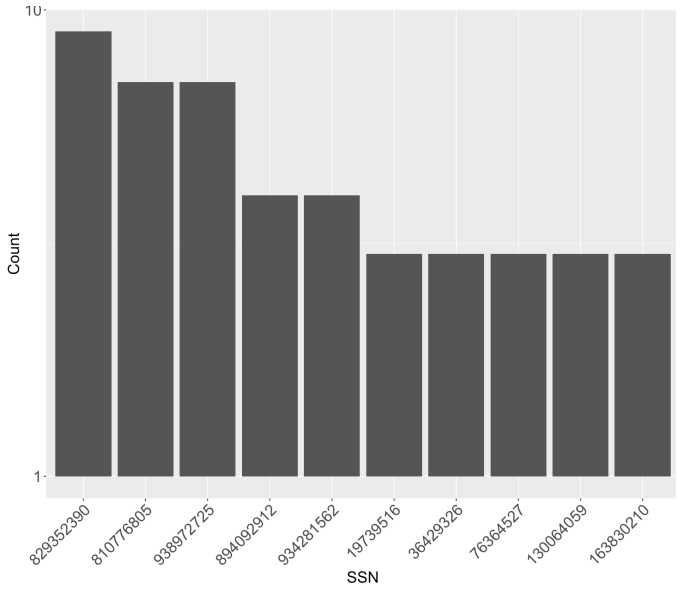
Description:
"ssn" is a categorical variable, indicating the applicant's ssn number he or she self-reported.

Unique Values:
100% populated with 96,535 unique values. Values in this field have lengths from 4 to 9. A ssn shorter than 9 digits indicates there are leading 0s in the ssn record. For example, ssn record "2503" is actually "000002503". The top 10 frequent appeared ssn records are shown below. It is obvious that the ssn record "737610282" is a frivolous record, which appears 173 times more than the second highest appeared record.

The top 10 most frequent records are listed below:

SSN	Frequency
737610282	1.73%
938972725	0.01%
810776805	0.01%
829352390	0.01%
473311863	0.00%
189622157	0.00%
163830210	0.00%
407447121	0.00%
118692079	0.00%
849295926	0.00%

The distribution of the top 10 records (excluding the most frequent record "737610282"):

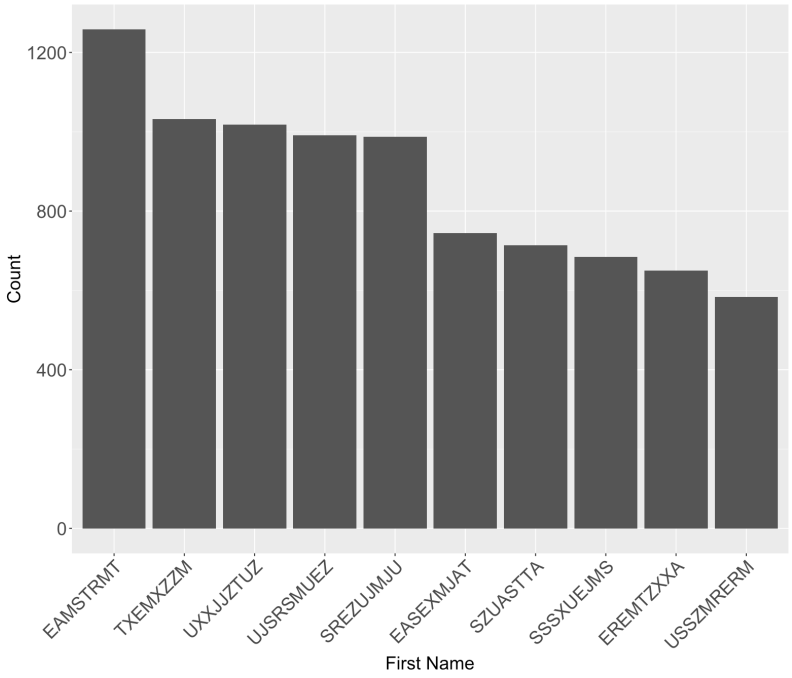


Field 4

Field Name: firstname

Description:
"firstname" is a categorical variable, indicating the applicant's first name he or she self-reported.

Unique Values:
100% populated with 16,576 unique values. The top 10 frequent appeared first name records are shown below.



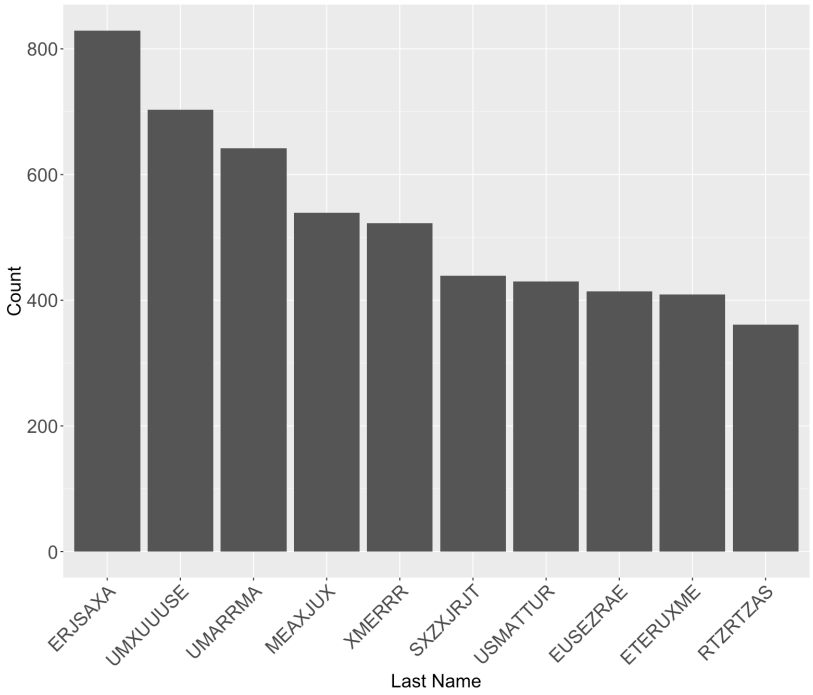
First Name	Frequency
EAMSTRMT	1.26%
TXEMXZZM	1.03%
UXXJJZTUZ	1.02%
UJSRSMUEZ	1.00%
SREZUJMJU	1.00%
EASEXMJAT	0.08%
SZUASTTA	0.07%
SSSXUEJMS	0.07%
EREMTZXXA	0.07%
USSZMRERM	0.06%

Field 5

Field Name: lastname

Description:
"lastname" is a categorical variable, indicating the applicant's last name he or she self-reported.

Unique Values:
100% populated with 36,312 unique values. The top 10 frequent appeared last name records are shown below.



Last Name	Frequency
ERJSAXA	0.83%
UMXUUUSE	0.70%
UMARRMA	0.64%
MEAXJUX	0.54%
XMERRR	0.52%
SXZXJRJT	0.44%
USMATTUR	0.43%
EUSEZRAE	0.41%
ETERUXME	0.41%
RTZRTZAS	0.36%

Field 6

Field Name: address

Description:

"address" is a text variable that contains inputs from applicants of their address.

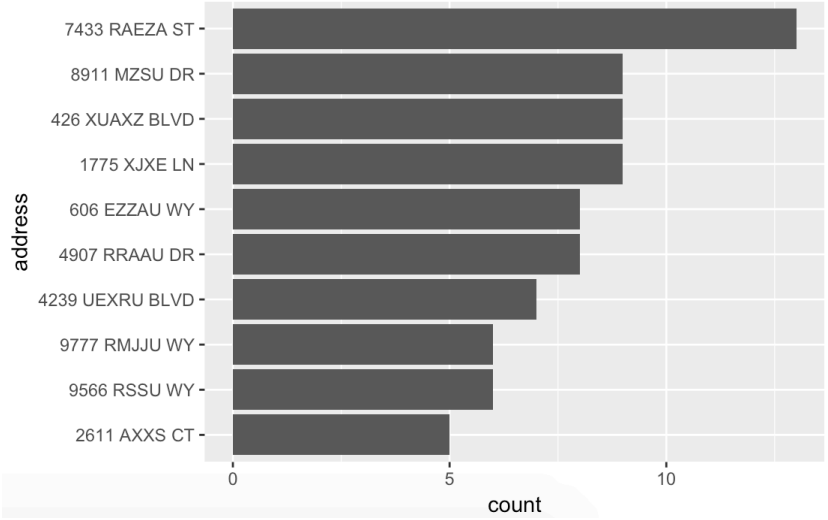
Unique Values:

100% populated with 97,563 unique values from 100,000 records. The most frequent address "2602 AJTJ AVE" appeared 117 times, which accounts for 0.12% of all records. This address is likely to be a frivolous address.

The top 10 most frequent records are listed below:

address	Frequency
2602 AJTJ AVE	0.12%
7433 RAEZA ST	0.01%
1775 XJXE LN	0.01%
426 XUAXZ BLVD	0.01%
8911 MZSU DR	0.01%
4907 RRAAU DR	0.01%
606 EZZAU WY	0.01%
4239 UEXRU BLVD	0.01%
9566 RSSU WY	0.01%
9777 RMJJU WY	0.01%

The distribution of the top 10 records (excluding the most frequent record "2602 AJTJ AVE"):



Field 7

Field Name: zip5

Description:

"zip5" is a categorical variable that contains inputs from applicants of their zip codes.

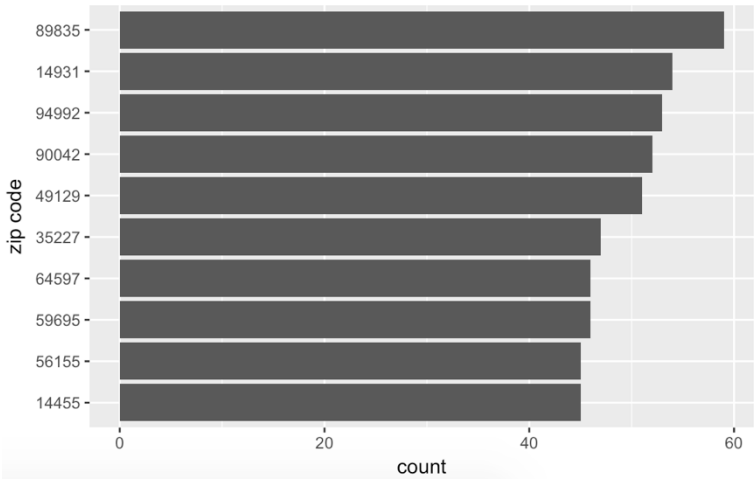
Unique Values:

100% populated with 16,547 unique values from 100,000 records. No missing values. The most frequent zip code "68138" appeared 823 times, which accounts for 0.09% of all records. This zip code is likely to be a frivolous value.

The top 10 most frequent records are listed below:

Zip5	Frequency
68138	0.09%
89835	0.06%
14931	0.05%
94992	0.05%
90042	0.05%
49129	0.05%
35227	0.05%
59695	0.05%
64597	0.05%
14455	0.05%

The distribution of the top 10 records (excluding the most frequent record "68138"):



Field 8

Field Name: dob

Description:

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

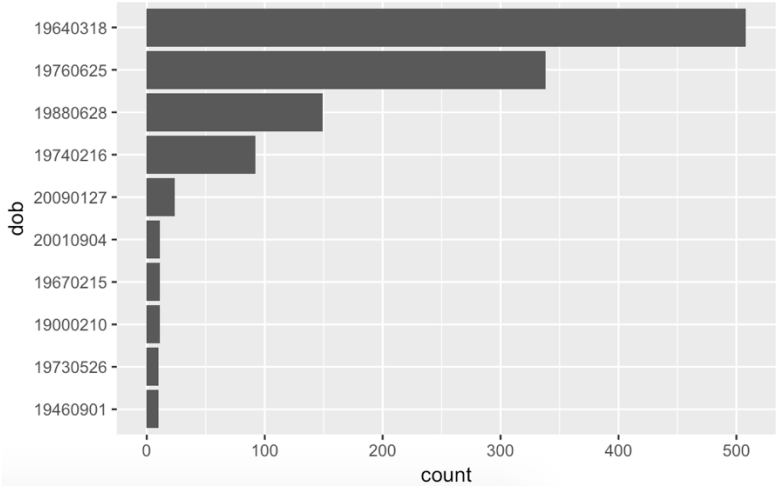
Unique Values:

100% populated with 36,816 unique values from 100,000 records. No missing values. The most frequent date of birth "19070626" appeared 12,488 times, which accounts for 12.49% of all records. This date of birth is likely to be a frivolous value.

The top 10 most frequent records are listed below:

dob	Frequency
19070626	12.49%
19640318	0.51%
19760625	0.34%
19880628	0.15%
19740216	0.09%
20090127	0.02%
19000210	0.01%
19670215	0.01%
20010904	0.01%
19460901	0.01%

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



Field 9

Field Name: homephone

Description:

“homephone” is a categorical variable that contains inputs from applicants of their home phone number.

Unique Values:

100 populated with 22,181 unique values from 100,000 records. No missing values. The most frequent home phone number “9105580920” appeared 7735 times, which accounts for 7.74% of all records. This homephone number is likely to be a frivolous value.

The top 10 most frequent records are listed below:

homephone	Frequency
9105580920	7.74%
6384782007	0.05%
2113738531	0.03%
6035129044	0.03%
4024680535	0.03%
9537440042	0.02%
8779246797	0.02%
161072500	0.02%
1992166532	0.02%
3968534400	0.02%

The distribution of the top 10 records (excluding the most frequent record “9105580920”):