Fraud Detection Report for New York City Property Pricing

Team 3

Peipei Han, Shuhan Zhou, Manwen Hu,

Sheng Ming, Baili Lu, Hao Qu

February 23, 2017

Contents

I. Executive Summary

The purpose of this project is to find unusual items by giving a fraud score to each record in New York Property Data using unsupervised learning models. The dataset was downloaded from new york city open data website showing information about New York City's property tax.

There are some missing values in the dataset, and we replace the numerical ones with the median for other observations and replace the categorical missing value with "NoValue." By adding the different combination of variables in the original dataset, we tried to ferret out some hidden information and relationship among each variable. We used two methods to solve this problem: Autoencoder and Mahalanobis Distance.

For Autoencoder, we standardize the data by Z-scale and reduce the dimension by PCA. We choose 14 PC's which can represent 80% of the data without making the dimension too large. Finally, we get the fraud score by training the Autoencoder model and run it on the entire database; For Mahalanobis, which is similar to what we did in Z-scale and PCA, we get the fraud score by calculating the Mahalanobis distance. We compare the top 10,000 likelihood of fraud observations and found that there's 85% of overlap for the two methods.

According to the fraud score, we get the top 10 unusually observations according. By checking each of them manually, we tried to address the reasons for high fraud score. The reasons can be the huge difference of value for certain records compared with group average, unreasonable missing value or strange ratios of some combination of numerical variables.

We understood, cleaned the data, added new variables and use Z-scale and autoencoder to solve the question. We used unsupervised model in this project and find some meaningful insights which can help us to address unusual items in the dataset and this method can also be applied to other datasets to address the potential fraud.

II. Data Distribution

2.1. Overall Summary

The data contains information about New York City's property tax information with 1,048,576 observations and 29 variables including location, owner, value, volume, and other tax-related variables. There are 14 numerical variables, 13 categorical variables, and two text variables. The original data was downloaded from New York City Open Data in https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8.

2.2. Important Variables

Our team selected 15 important variables and made a Data Quality Report these factors. Figure 1 is the summary of categorical variables. All the important categorical variables except ZIP, which contains 2.51% missing value, are 100% populated.

Field Name	Type	Official Descripton	% Populated
BLOCK	Categorical	block	100%
LOT	Categorical	lot	100%
BLDGCL	Categorical	building class	100%
TAXCLASS	Categorical	tax class	100%
ZIP	Categorical	zip code	97.49%

Figure 1 Summary of categorical variables

Figure 2 is the summary table for numerical data. All the important numerical variables except STORIES, which contains 4.97% missing value, are 100% populated.

Field Name	Type	Official Descripton	% Populated	Min	Max	Mean	Median	Mode	Stdev
LTFRONT	Numerical	lot width	100%	0	9999	36.17	25	0	73.73
LTDEPTH	Numerical	lot depth	100%	0	9999	88.28	100	100	75.45
STORIES	Numerical	number of stories in building	95.03%	1	119	5.06	2	2	8.43
FULLVAL	Numerical	market value	100%	0	6,150,000,000	880487.6579	446,000	0	11702927
AVLAND	Numerical	actual land value	100%	0	2,668,500,000	86000	13,646	0	4100755
AVTOT	Numerical	actual total value	100%	0	4,668,308,947	230800	25,339	0	6951206
EXLAND	Numerical	actual exempt land value	100%	0	2,668,500,000	36810	1,620	0	4024330
EXTOT	Numerical	actual exemtp land total	100%	0	4,668,308,947	92540	1,620	0	6578281
BLDFRONT	Numerical	building width	100%	0	7575	23.02	20	0	35.79
BLDDEPTH	Numerical	building depth	100%	0	9393	40.07	39	0	43.03
AVLAND2	Numerical	transitional land value	26.80%	3	2371000000	246000	20,059	2,408	6199390
AVTOT2	Numerical	transitional total value	26.80%	3	4501000000	716100	80,010	750	11690165
EXLAND2	Numerical	transitional exempt land value	8.27%	1	2371000000	351800	3,053	2,090	10852484
EXTOT2	Numerical	transitional exempt land total	12.40%	7	4501000000	658100	37,116	2,090	16129808

Figure 2 Summary of numeric variables

2.3. Categorical Variable

Below are the distributions of five important categorical variables.

- BLOCK

  BLOCK is a categorical variable. BBLE represents the location of property and BLOCK, LOT and EASEMENT consist the unique parcel identifier. Valid BLOCK ranges by borough are Manhattan 1 to 2255, Bronx 2260 to 5958, Brooklyn 1 to 8955, Queens 1 to 16,350, Staten Island 1 to 8050.

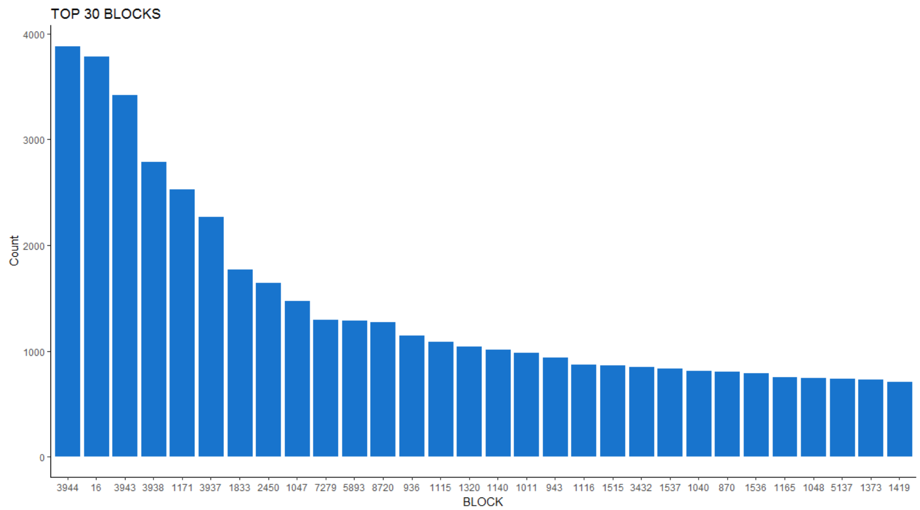


Figure 3 BLOCK Distribution**Figure 3 BLOCK Distribution**

- LOT

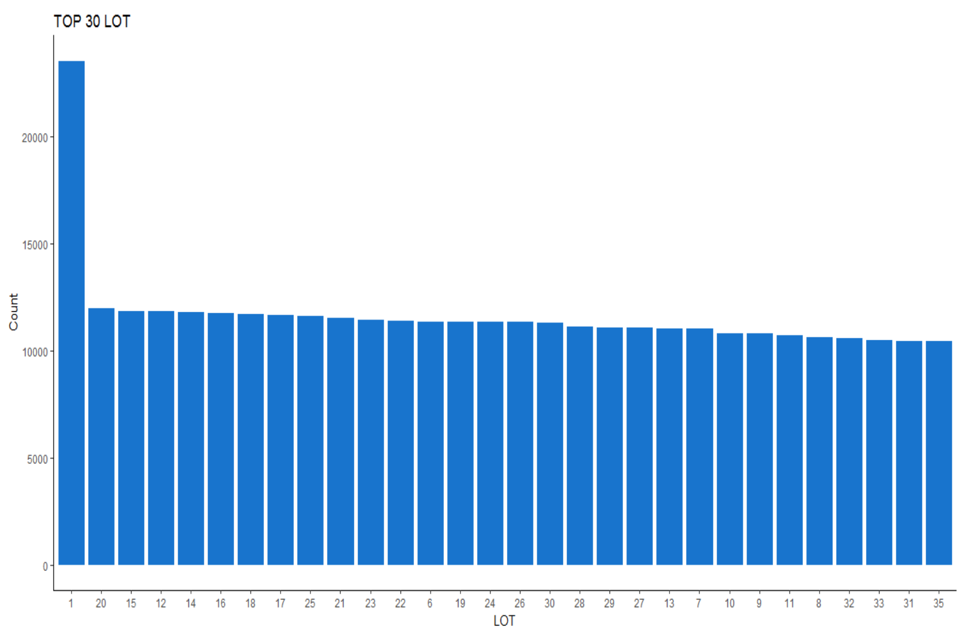


Figure 4 LOT Distribution

- TAXCLASS

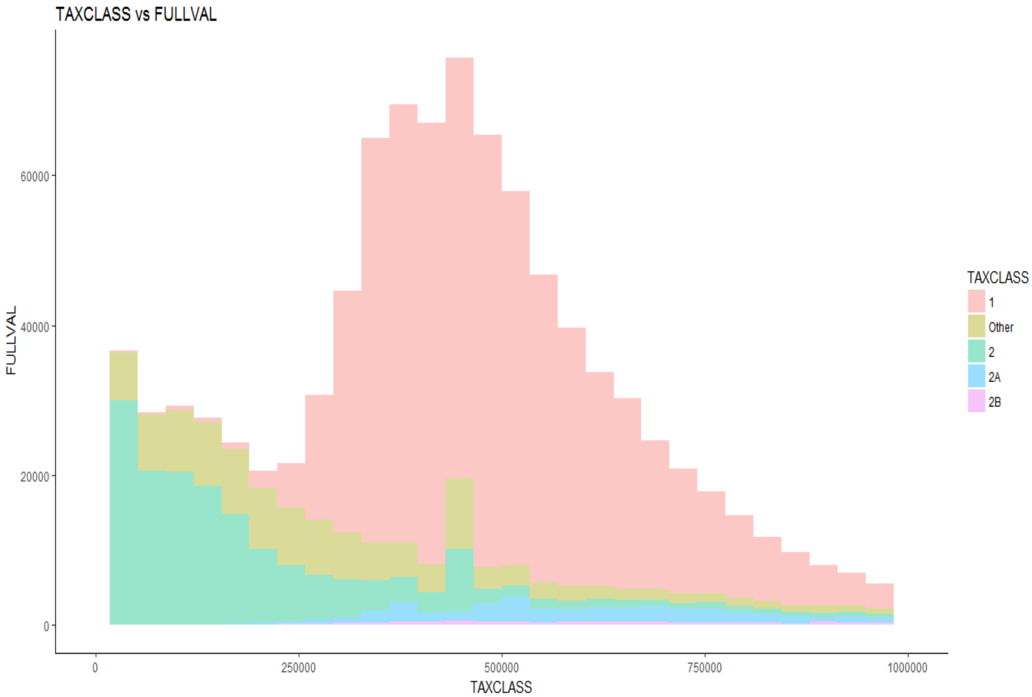


Figure 5 TAXCLASS Distribution

- BLDGCL

BLDGCL represents Building Class. The encoding method is alpha in the first position and numeric in the second position. There is a direct correlation between the Building Class and the Tax Class. If the Building Class is known the Tax Class can be generated. The corresponding relationship between TAXCLASS and BLDCLASS is:

TAXCLASS	BLDGCLASS
1	A0 - A9, B1 - B9, C0, G0, R3, R6, R7, S0 - S2, V0, V2, V3, Z0
2	C1 - C9, D0 - D9, R0, R1, R2, R4, R8, R9, S3, S4, S5, S9
3	U1 - U2, U4 - U9
4	ALL OTHER

Figure 6 shows the top 15 BLDGCL with high FULLVAL.

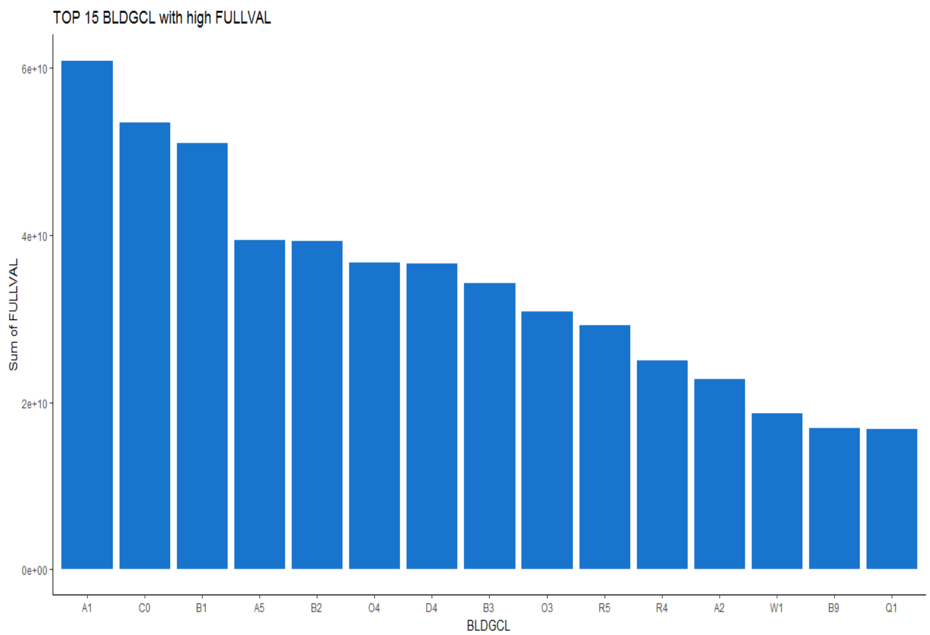


Figure 6 BLDGCL Distribution

- ZIP

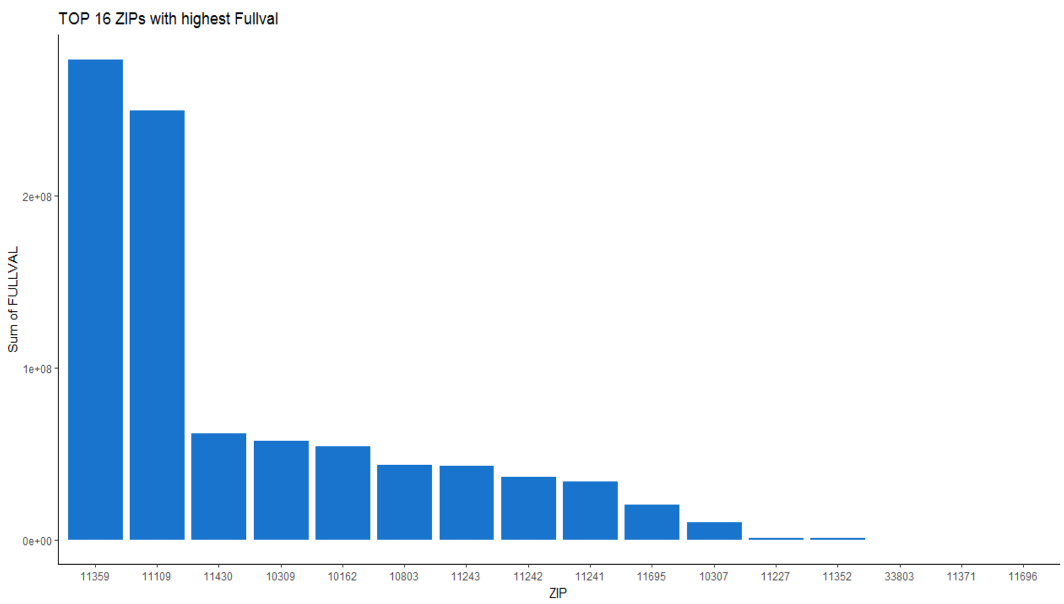
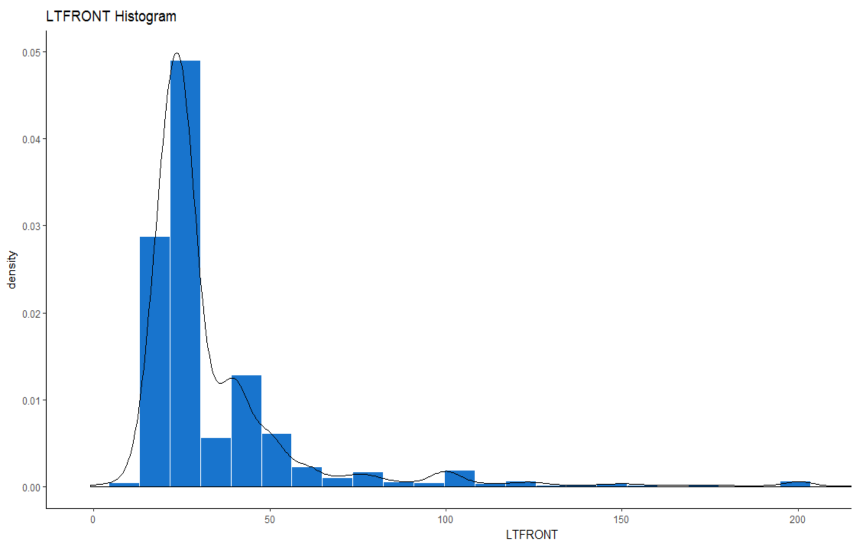


Figure 7 ZIP Distribution

2.4. Numerical Variable

Below are the distributions of ten important numerical variables.

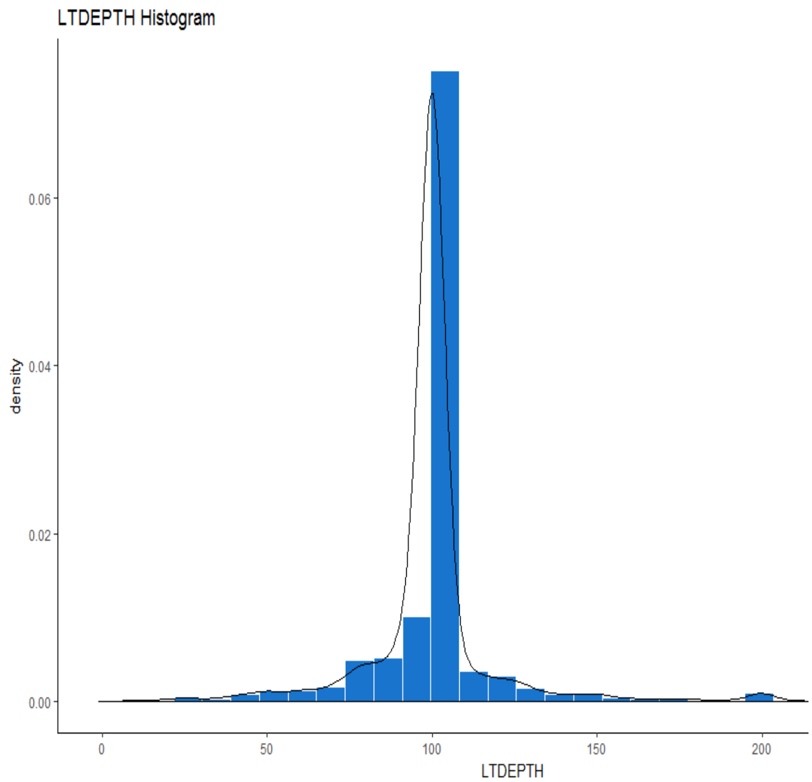
- LTFRONT



Min	1st Qu.	Median
1	21	25
Max	**3rd Qu.**	**Mean**
9999	40	40.2

Figure 8 LTFRONT Distribution

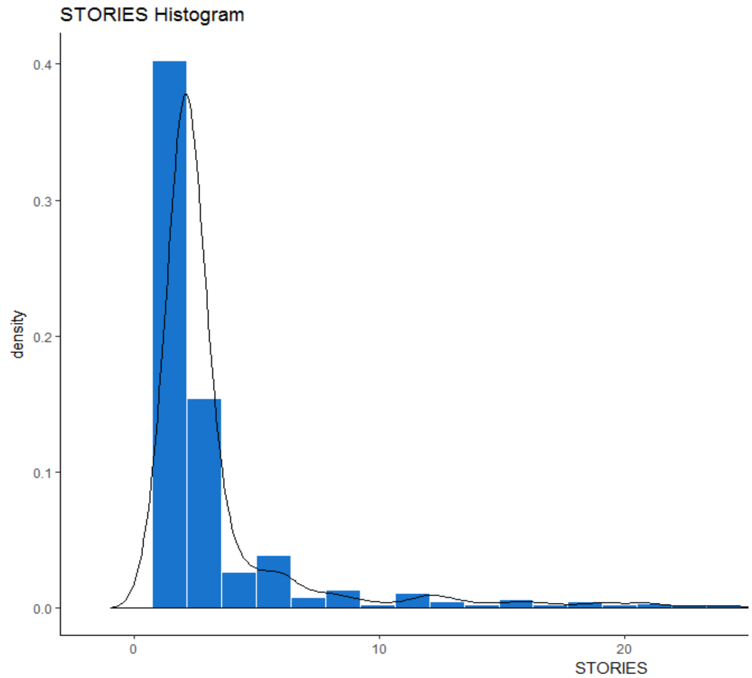
- LTDEPTH



Min	1st Qu.	Median
1	100	100
Max	**3rd Qu.**	**Mean**
9999	100	104.5

Figure 9 LTDEPTH Distribution

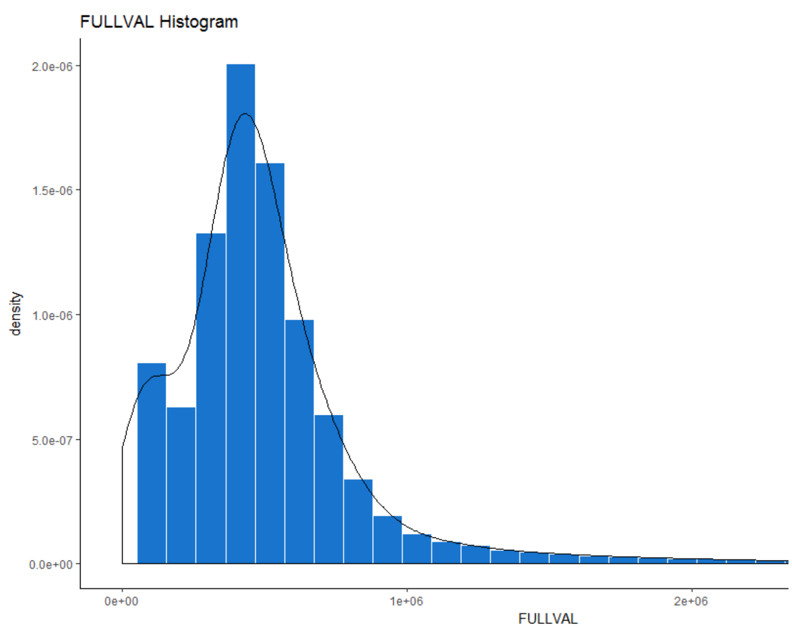
- STORIES



Min	1st Qu.	Median
1	2	2
Max	**3rd Qu.**	**Mean**
119	3	4.911

Figure 10 STORIES Distribution

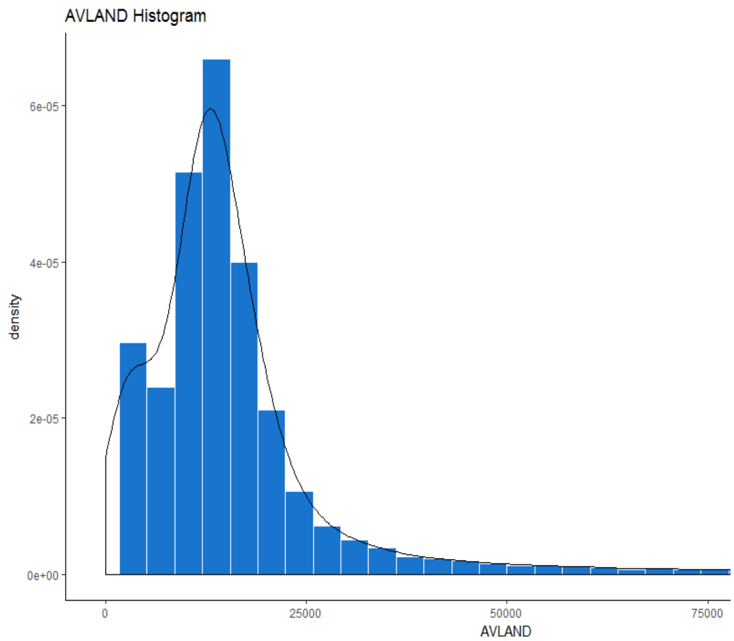
- FULLVAL



Min	1st Qu.	Median
4	313000	45000
Max	**3rd Qu.**	**Mean**
6150000000	619000	886000

Figure 11 FULLVAL Distribution

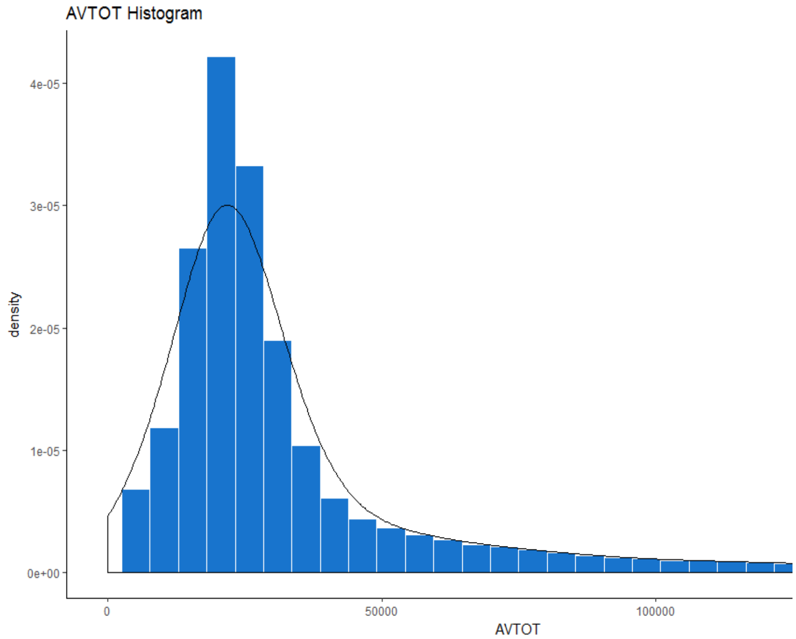
- AVLAND



Min	**1st Qu.**	**Median**
1	9513	13750
Max	**3rd Qu.**	**Mean**
2668000000	19710	86160

Figure 12 AVLAND Distribution

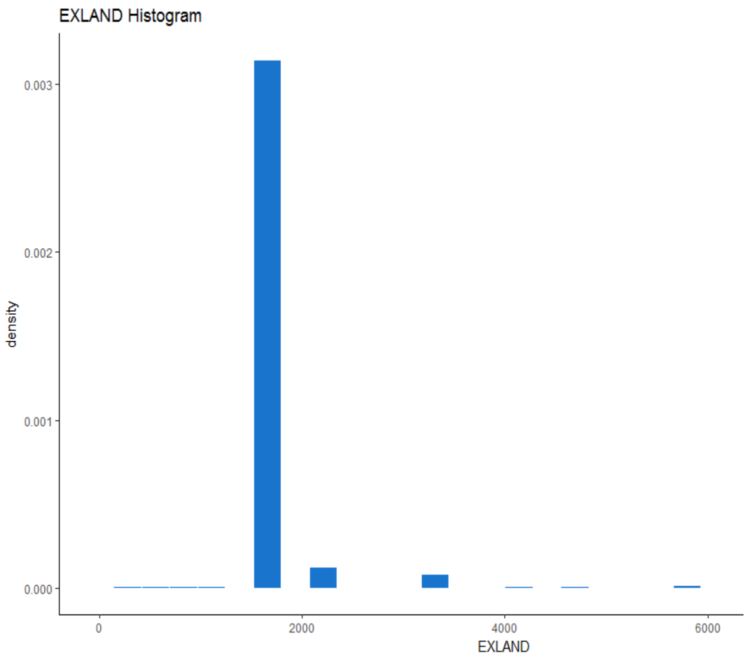
- AVTOT



Min	**1st Qu.**	**Median**
1	18740	25560
Max	**3rd Qu.**	**Mean**
4668000000	46100	231100

Figure 13 AVTOT Distribution

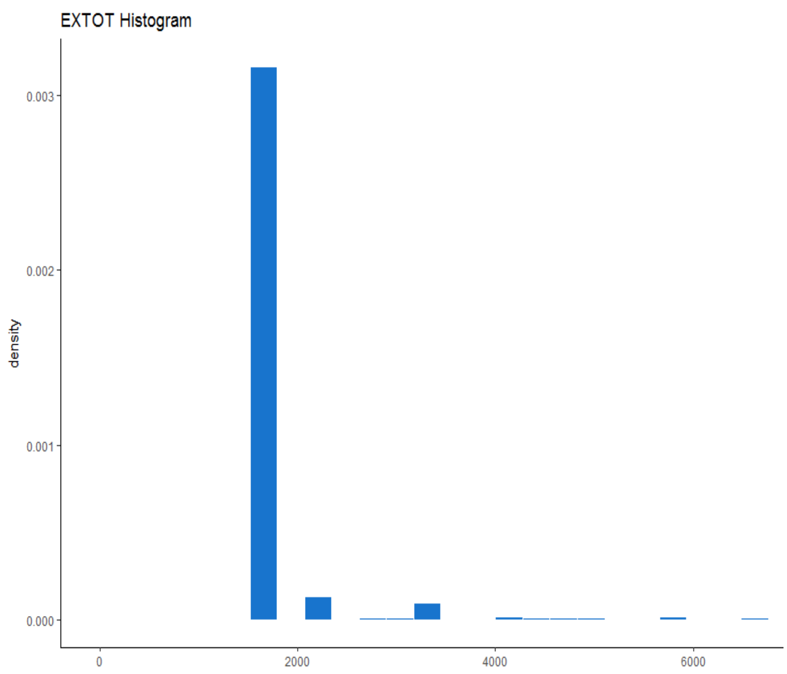
- EXLAND



Min	1st Qu.	Median
1	1620	1620
Max	**3rd Qu.**	**Mean**
2668000000	1620	37560

Figure 14 EXLAND Distribution

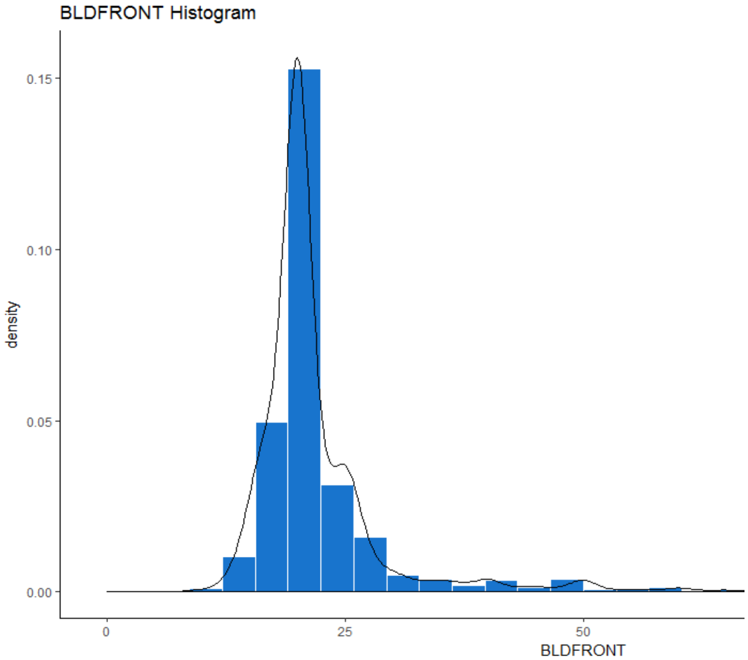
- EXTOT



Min	1st Qu.	Median
1	1620	1620
Max	**3rd Qu.**	**Mean**
4668000000	2090	93200

Figure 15 EXTOT Distribution

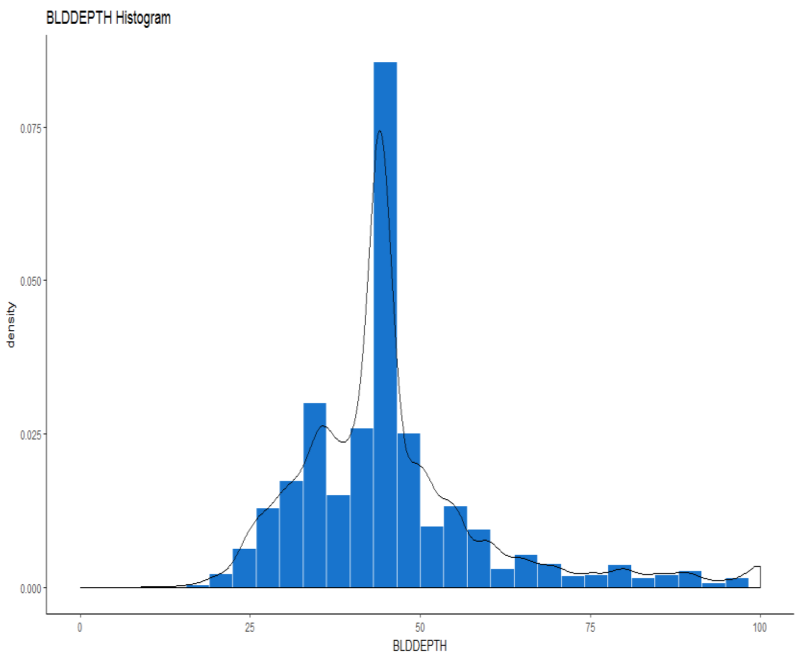
- BLDFRONT



Min	1st Qu.	Median
1	20	20
Max	**3rd Qu.**	**Mean**
7575	24	27.3

Figure 16 BLDFRONT Distribution

- BLDDEPTH



Min	1st Qu.	Median
1	37	44
Max	**3rd Qu.**	**Mean**
9393	51	49.5

Figure 17 BLDDEPTH Distribution

III. Variables Manipulation

3.1. Create some new numerical variables as predictors

3.1.1. Product of existing variables

The product of some existing variables might have practical implications which cannot represent by the linear model. For the product of LTFRONT and LTDEPTH, it indicates the lot area of each property. The product of BLDFRONT, BLDDEPTH and STORIES represents the spatial volume of each property.

- $\text{LotArea} = \text{LTFRONT} * \text{LTDEPTH}$
- $BuildVolume = BLDFRONT * BLDDEPTH * STORIES$

3.1.2. Fraction of existing variables

Adding variables such as FULLVAL divided by BuildVolume, it displays the average value per cubic meter of each building. We also add some ratios of assessed values, such as actual land value divided by actual exempt land value and transitional total value divided by transitional exempt land value which represents the portion of land value of the property. The parking spaces might also affect the unit price of each property. In order to measure the unit price under different lot size, we add another fraction of actual land value divided by the lot area, which stands for the land value per lot area.

- $\text{FULLVAL}/\text{BuildVolume}$
- $AVLAND/EXLAND$
- $AVTOT/EXTOT$
- $AVLAND/LotArea$

3.1.3. The difference between actual and calculated assesses value

Based on NYC gov. website, for tax class 1, the assessment ratio is 6%; for tax class 2,3,4, the assessment ratio is 45%. So we can calculate the assessed value by multiplying FULLVAL by the assessment ratio. Given the assess ratios of different Tax Class, we build the variable “AssessDiffRatio” to examine how significant the difference is between the AVTOT and the appropriate assessed value of each record ($Market\ Value * assess\ ratio$). If the value of “AssessDiffRatio” is noticeably different from zero, the corresponding record could be probably identified as a fraud.

- $AssessDiffRatio = \frac{AVTOT - FULLVAL*assessment\_ratio}{AVTOT}$

3.2. Create Intermediary Variables

Group some numerical variables by different categorical variables, and then calculate the average value of each group as intermediary variables.

3.2.1. Grouping by TAXCLASS

The current system, which was enacted in 1981 over a gubernatorial veto, classifies all real estate parcels into four classes, as follows:

Tax Class 1 indicates the following types of primarily residential property;

Tax Class 2 is for all other primarily residential properties, including any residential condominiums not in Class 1;

Tax Class 3 includes real estate of utility corporations and special franchise properties, excluding land and certain buildings;

Tax Class 4 is all commercial real estate. It includes all other properties, such as stores, warehouses, hotels, and any vacant land not classified as Class 1.

The tax rates have a downward trend going from Class 1 to Class 4, so this difference in tax rate provides us motivation to group every property into its tax class. And we calculate average assessed value level within each group.

- $TXCmeanAVLANDtoEXLAND = mean\left(\frac{AVLAND}{EXLAND}\right)$
- $TXCmeanAVTOTtoEXTOT = mean\left(\frac{AVTOT}{EXTOT}\right)$
- $TXCmeanFVtoBLDVOL = mean\left(\frac{FULLVAL}{BuildVolume}\right)$
- $TXCmeanAVtoFULL = mean(AVtoFULL)$
- $TXCmeanAVLANDtoLotArea = mean\left(\frac{AVLAND}{LotArea}\right)$
- $TXCmeanSTORIES = mean(STORIES)$
- $TXCmeanAVLAND = mean(AVLAND)$
- $TXCmeanLotArea = mean(LotArea)$
- $TXCmeanBLDVOL = mean(BuildVolume)$

3.2.2. Grouping by AERA

AREA is a new categorical variable to describe sub neighborhoods of the city, and it is composed of the serial digits in BLOCK and LOT, which divides the city into hundreds of different geographical regions. Given the fact that there is a geographical difference on the scope of real estate price and land price, we group variables associating with price by AREA.

- $AERAmeanAVLAND = mean(AVLAND)$
- $AERAmeanFULLVALtoBuildVolume = mean\left(\frac{FV}{BuildVolume}\right)$
- $AERAmeanAVTOT = mean(AVTOT)$
- $AERAmeanAVLANDtoFULLVAL = mean\left(\frac{AVLAND}{FV}\right)$
- $AERAmeanFULLVAL = mean(FULLVAL)$
- $AVLANDtolotArea = mean\left(\frac{AVLAND}{lotArea}\right)$

3.2.3. Group by BLGGCL (Building Class)

As it is implicated from New York City Website, http://nycprop.nyc.gov/nycproperty/help/hlpbldgcode.html#D

The classifications of buildings are assigned with value staring from A to Z by size and usage, so we are able to grasp the intrinsic difference between building classes by looking at some examples of BLDGCL values, such as: Building code starting with letter 'D' represents Elevator Apartment, and 'F' represents Factory.

Moreover, the tax exemption class (EXMPTL) also relates to building class. Considering the fact that more than 90% of the EXMPTL field is missing, it is not appropriate if calculating group average of EXLAND, EXTOT, AVLAND/EXLAND, AVTOT/EXTOT by EXMPTL. But observing the situation that properties built with municipal purposes (eg. government buildings) often enjoy the highest tax exemption, while buildings of commercial use are often associated with lower exemption. In this case, building class (BLDGAL) could serve as an effective substitute of exemption class (EXMPTCL).

In consequence, a logical inference could be made: There is discernable pattern on the values of many numerical variables in each building class. Based on this inference, we calculate the mean of the below variables by BLDGCL:

- $BLDCmeanAVLANDtoEXLAND = mean\left(\frac{AVLAND}{EXLAND}\right)$
- $BLDCmeanAVTOTtoEXTOT = mean\left(\frac{AVTOT}{EXTOT}\right)$
- $BLDCmeanFVtoBLDVOL = mean\left(\frac{FULLVAL}{BuildVolume}\right)$
- $BLDCmeanAVtoFULL = mean(AVtoFULL)$
- $BLDCmeanAVLANDtoLotArea = mean\left(\frac{AVLAND}{LotArea}\right)$
- $BLDCmeanSTORIES = mean(STORIES)$
- $BLDCmeanLotArea = mean(LotArea)$
- $BLDCmeanBLDVOL = mean(BuildVolume)$

3.2.4. Grouping by ZIP

zip code is another effective method to categorize the city into sub groups. In illustration, in each borough, there is a bunch of ZIP codes correspond with small neighborhoods within that borough. Taking Manhattan as an example, neighborhood East Harlem's zip codes are 10029, 10035, while neighborhood Greenwich Village's zip codes are 10012, 10013, 10014. As the land price, real estate price, and the real estate price per square foot could be drastically different in different regions and neighborhoods according to statistical reports. So, we calculate the mean value of all the numerical variables relating to price and unit price:

- $ZIPmeanFULLVAL = mean(FULLVAL)$
- $ZIPmeanAVLAND = mean(AVLAND)$
- $ZIPmeanAVTOT = mean(AVTOT)$
- $ZIPmeanEXLAND = mean(EXLAND)$

- $ZIPmeanEXTOT = mean(EXTOT)$
- $ZIPmeanAVLANDtoEXLAND = mean\left(\frac{AVLAND}{EXLAND}\right)$
- $ZIPmeanAVTOTtoEXTOT = mean\left(\frac{AVTOT}{EXTOT}\right)$
- $ZIPmeanFVtoBLDVOL = mean\left(\frac{FULLVAL}{BuildVolume}\right)$
- $ZIPmeanAVtoFULL = mean(AVtoFULL)$
- $ZIPmeanAVLANDtoLotArea = mean(\frac{AVLAND}{LotArea})$

3.3. Variables based on intermediary variables from above

Create a series of expert variables which involved with those intermediary variables in 3.2., which is simply by comparing the value of variables to the group mean calculated in 3.2.

3.3.1. Transformation of AVLAND and AVTOT

Comparing the value of AVLAND (Property Assessed Value) in each record to the average value within its group, following the grouping rule of tax class, building class, zip code, area. And listing the names as below: AVLAND_TXC, AVLAND_BLDC, AVLAND_ZIP, AVLAND_AREA. Complying with the logic in the transformation of AVLAND, we build new variables derived from AVTOT: AVTOT_TXC, AVTOT_BLDC, AVTOT_ZIP, AVTOT_AREA.

3.3.2. Transformation of AVLAND/EXLAND, AVTOT/EXTOT, EXLAND, EXTOT

Group by BLDC, TXC

3.3.3. Transformation of BuildVolume, LotArea

Group by TXC, BLDCL

3.3.4. Transformation of AVLAND/FULLVAL, AVLAND/LotArea, FULLVAL/BuildVolume

Group by BLDC, TXC, ZIP,

3.3.5. Transformation of STORIES

Group by TXC, BLDCL

3.4. Summary of Expert Variables

There are 50 expert variables included into the Principle Component Analysis procedure, consisting of 9 numerical variables in the original fields, 8 variables derived from the original fields, and 33 variables generated by calculating the ratio of the previous variables and their group means in various categories.

IV. Methods and Techniques

4.1. Standard normalization and PCA

As mentioned above, we totally have 50 numeric variables to build model and calculate the fraud score. But high dimension could be problematic. It means high computation cost and leads to overfitting. There could also be high correlation among the variables.

Dimensionality reduction addresses these problems, while preserving most of the relevant information in the data needed to learn accurate, predictive models. The axes of the reduced subspace typically correspond to latent features that remove noise, abstract, compress and in general better describe the correlations and interactions among the original set of features - thus enabling learning algorithms to perform better.

Principal component analysis is the main linear technique for dimensionality reduction. It performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

However, before implementing PCA, we should do $ZScale = \frac{x-\mu}{\sigma}$ standardization first in order to adjust values measured on different scales to a notionally common scale, as PCA is a variance maximizing exercise.

In practice, the covariance matrix of the data is constructed and the eigen vectors on this matrix are computed. The eigen vectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigen vectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space has been reduced to the space spanned by a few eigenvectors. One common criterion is to include all those PCs up to a predetermined total percent variance explained, such as 80%.

PCA can be done using *prcomp()* function in R. The variable standard deviations are stored in the attribute scale and scores are in the attribute x. After PCA, we selected 14 variables as the corresponding cumulative eigen value (variance) reaching 80%. Figure 18 shows that there is a decline at PC_{14} and behind PC_{14} there is less information contributed to dataset. Therefore, the first 14 variables are chosen to be the input of the fraud score algorithm to calculate fraud score. Figure 19 illustrates the variables which the most significant ones are for PC_1.

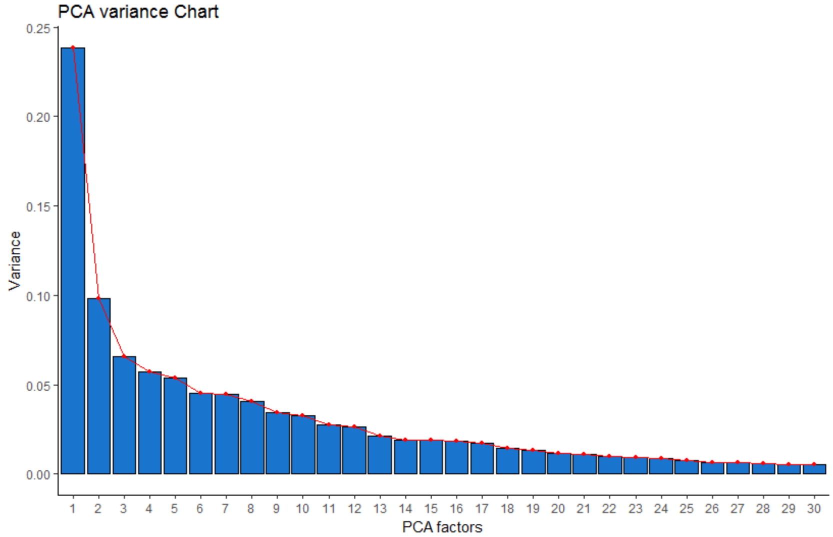


Figure 18 PCA Variance Chart

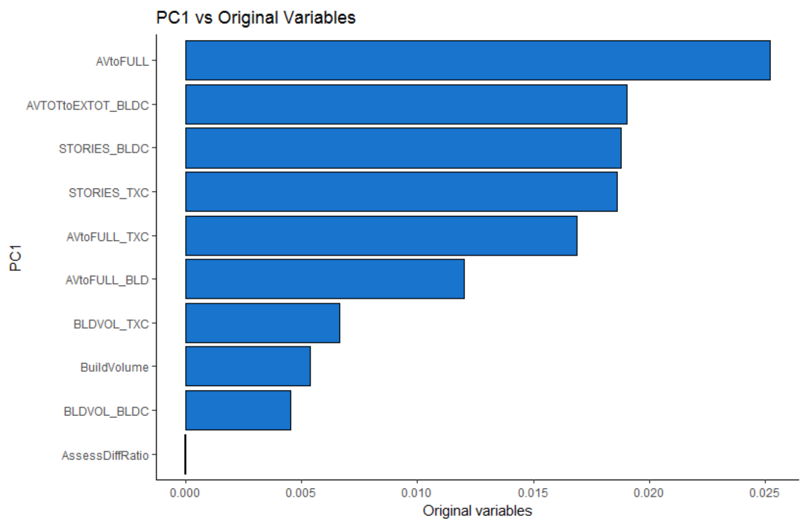


Figure 19 Original variables contributed to PC_1

4.2. Heuristic Modeling

In order to generate fraud score, we used Heuristic algorithm and Autoencoder individually. In terms of the Heuristic algorithm, we manipulated and modeled all the principal components in the following procedures:

1. Z-scale all the PCs.
   Although we have already z-scaled all the original and expert variables before we did PCA, we are now trying to investigate the deviation of each observation within each principal component. Since each principal component have different mean value and different variance, a z-scaling on all the PCs is a must before we do any calculation on the PCs.

2. Sum up the absolute values of all z-scaled PCs and take the cube root.
   Since now we have the z-scaled PCs, one of the most straight-forward ways to measure the total deviation is to sum up all the absolute values of the z-scaled PCs. We choose to take the cube root on the sum for a less skewed distribution and more comparable result to the Autoencoder score (will mention below). So we get:

$$Score.Heuristic = (\sum^{i} |PC_i|)^{\frac{1}{3}}$$

3. Scale the Score.Heuristic to [0,1].
   To make our score more comparable, we scale the scores to [0,1] using (score-min)/(max-min).

The distribution of $Score.Heuristic$ is shown in Figure 20, which illustrates that the majority of the Heuristic fraud score are concentrated around 0.1 and there is an obvious skew and long tail after 0.25.

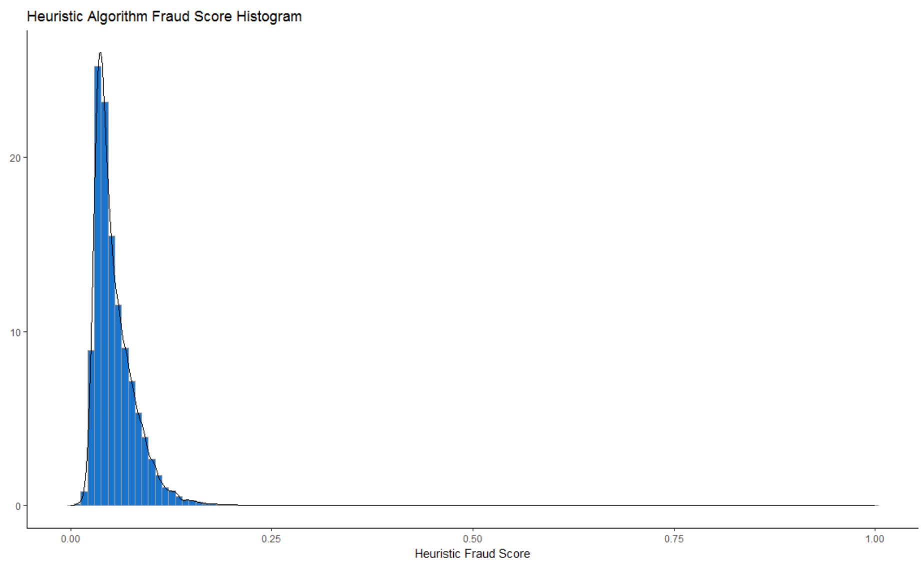


Figure 20 Score. Heuristic Histogram

4.3. Autoencoder

An autoencoder neural network is an unsupervised learning algorithm that applies back-propagation, setting the target values to be equal to the inputs. The Autoencoder tries to learn a function $h_{w,b}(x) \approx x$. In other words, it is trying to learn an approximation to the identity function, so as to output x^ that is similar to x and the fraud score is given according to the reconstruction error. The general process is shown in Figure 21.

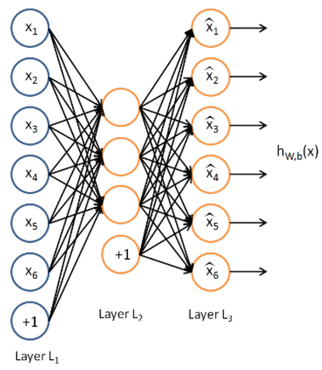


Figure 21 Autoencoder Algorithm

In practice, we used the h2o library and tuned the autoencoder model with different parameters such as the number of hidden layers, the number of neurons within each layer and the number of iterations. After a number of trials, we find one hidden layer with 14 neurons (which is the same as the number of our PCs) is an efficient neural network for our training and 50 iterations is good enough to make the result converge to an optimal value.

Besides, we take the sixth root of the reconstruction error and scale the result to [0,1] to make the Autoencode score more comparable to the Heuristic score since we are about to combine them both linearly. So we get:

$$Score.Autoencoder = (Reconstruction.MSE)^{\frac{1}{6}}$$

The distribution of $Score.Autoencoder$ is shown in Figure 22, which shows a similar pattern compared with the distribution of Heuristic score.

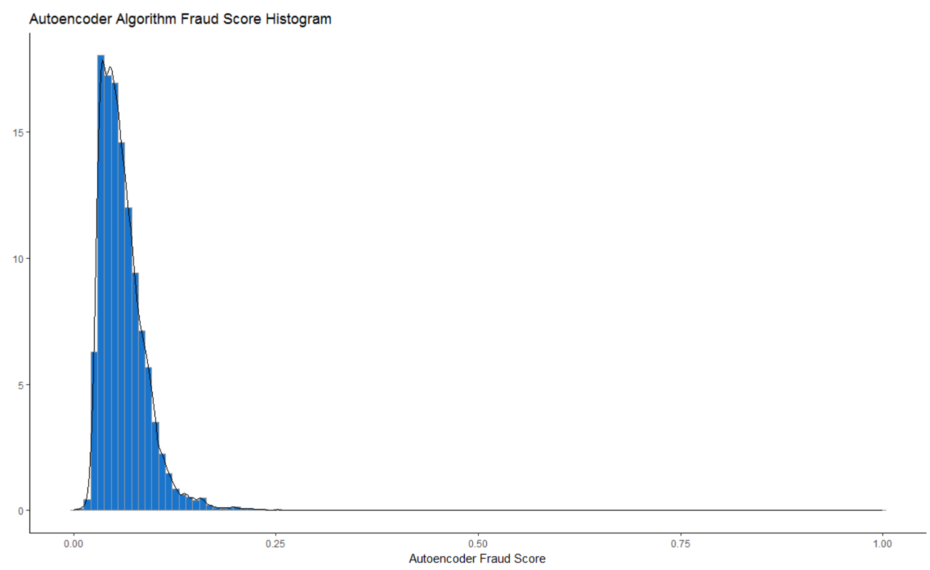


Figure 22 Score. AutoencoderHistogram

We sorted the records by Heuristic.Score and Autoencoder.Score respectively and selected the top 10,000 observations from each result. As it turns out, 8,453 of the records appear in both two top 10,000 records! That's a remarkable result and a very good sign of the effectiveness of both the two models.

Since we have both the two scores now and they are about of the same scale and very similar distribution. We are planning to derive the final fraud score with a combination of the two scores. After all, we are doing an unsupervised learning modeling for the project. We should not bet all on one model.

4.4. Fraud Score Combination

Since the two scores have extremely similar distribution, we think the one of the most straight-forward ways to balance the two score is to make a simple linear combination. Since the Autoencoder algorithms is more sophisticated, we decided to give a 0.7 weight to Score.Autoencoder and 0.3 to Score.Heuristic, so we get:

$$Score.Combined = 0.7 * Score.Autoencoder + 0.3 * Score.Heuristic.$$

The result of combined fraud score is shown in Figure 23. The long tail indicates that amount of records have high fraud scores and we should get the original abnormal records and go deep for future research of the reasons that led to high fraud scores. Therefore, after the combination, we selected the top 1% highest fraud score records.

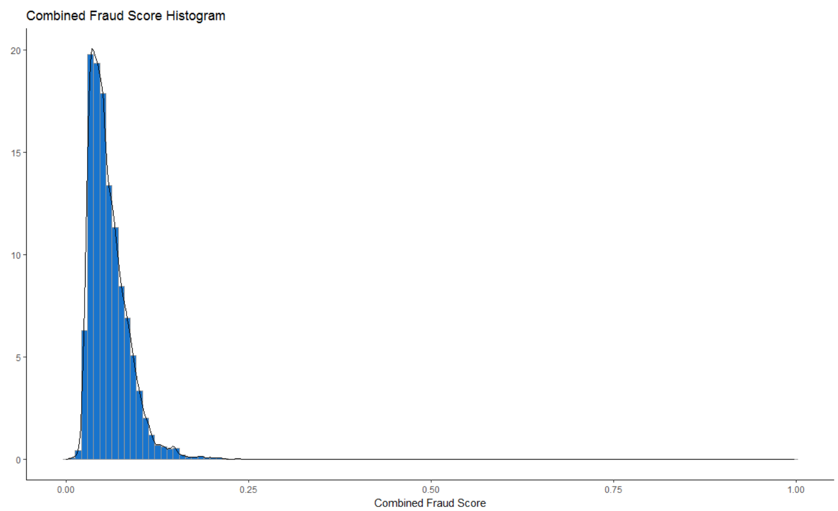


Figure 23 Combined Histogram

V. Result Analysis

As mentioned above, we extracted the top 1% (10,000 rows) records from all the observations according to the combined fraud score. In this part, we will pick the top 10 records to explore and try to explain what's behind the high fraud scores.

RECORD	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	EXCD1
78804	3085900700	8590	700		U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665	1946836665	1946836665	2231
276946	3002451441	245	1441			R4	2	0	0	16	721385	102971	324623	0	1	1920
294061	1011110001	1111	1		CULTURAL AFFAIRS	Q1	4	840	0	NA	6.15E+09	2668500000	2767500000	2668500000	2767500000	2231
6949	1015101092	1510	1092		BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429	0	0	NA
376243	4142600001	14260	1		LOGAN PROPERTY, INC.	T1	4	4910	0	3	374019883	1792808947	4668308947	1792808947	4668308947	2198
901790	4141400001	14140	1		UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610	32408610	32408610	4600
5393	4018420001	1842	1		864163 REALTY, LLC	D9	2	157	95	1	2930000	1318500	1318500	0	0	NA
648675	4066610005E	6661	5	E	M FLAUM	V0	1B	1	1	NA	0	0	0	0	0	NA
902256	2049910126	4991	126			V0	1B	1	1	NA	0	0	0	0	0	NA
486117	3002451419	245	1419			R4	2	0	0	16	408167	58262	183675	0	1	1920

EXCD1	STADDR	ZIP	EXMPTCL	BLDFRONT	BLDDEPTH	AVLAND2	AVTOT2	EXLAND2	EXTOT2	EXCD2	PERIOD	YEAR	VALTYPE	Score.Heur	Score.AE	Score.Combined
2231	FLATBUSH AVENUE	NA	X1	0	0	848484666	848484666	848484666	848484666	NA	FINAL	2010/11	AC-TR	1	0.997522936	0.998266055
1920	360 FURMAN STREET	11201		0	0	79162	406102	NA	NA	NA	FINAL	2010/11	AC-TR	0.948191983	1	0.948057079
2231	1000 5 AVENUE	10028	X1	0	0	2371005000	2465055000	2371005000	2465055000	NA	FINAL	2010/11	AC-TR	0.921307245	0.969348174	0.929718389
NA	1438 3 AVENUE	10028		7575	9393	22896	146183	NA	NA	NA	FINAL	2010/11	AC-TR	0.995704791	0.944294811	0.896264874
2198	154-68 BROOKVILLE BOULEVARD	11422	X4	0	0	1644454002	4501180002	1644454002	4501180002	NA	FINAL	2010/11	AC-TR	0.96256437	0.951500733	0.893976181
4600	CROSS BAY BOULEVARD	11414	X3	0	0	NA	NA	NA	NA	NA	FINAL	2010/11	AC-TR	0.982895799	0.914466536	0.877912426
NA	86-55 BROADWAY	11373		1	1	1201200	1201200	NA	NA	NA	FINAL	2010/11	AC-TR	0.976132756	0.945699893	0.825297203
NA	VLEIGH PLACE	NA		0	0	NA	NA	NA	NA	NA	FINAL	2010/11	AC-TR	0.855994434	0.839281079	0.82205614
NA	BELL AVENUE	NA		0	0	NA	NA	NA	NA	NA	FINAL	2010/11	AC-TR	0.855979407	0.839279273	0.822044122
1920	360 FURMAN STREET	11201		0	0	44789	229773	NA	NA	NA	FINAL	2010/11	AC-TR	0.771009286	0.814426453	0.775288686

Again we can see from the top 10 that the observations with highest Heuristic scores also have highest Autoencoder scores. The high overlapped rate (85%) for the two scores in the top 10,000 records and high consistency among the top 10 provide strong evidence for the effectiveness and robustness of both the two models we built.

For the record 78804, it's strange that the property is totally exempt from taxation and it is such a large amount of tax exemption, which is very suspicious. Besides, there is a huge drop from the original assessed values to the transitional assessed values, which is also very odd. For the record 294061, 376243 and 901790, there are similar issues with the first one. They are totally exempt from taxation and they are both with extremely high market value and assessed values. We noticed that record 78804 and 294061 share the same exemption code X1 and record 376243 and 901790 have the exemption code X4 and X3, we can't find the codebook for those values. These exemptions should only be applicable to certain types of buildings like non-profit ones, but it's still worthy of investigating what's the real reason behind the huge exemption.

For record 276946 and 486177, the ratio of land value over total value are unexpectedly low compared to other observations, which may indicate that the assessed total value is reported to be artificially high for some purposes.

For record 6949, since it has high building width and building depth, it is supposed to have larger market value and assessed value, however, the building has relatively low market values and assessed values, which is noticeable.

Besides, there are some observations having many missing values like record 648675 and 902256, which may also be an indication of potential fraud.

In conclusion, in the real world, the problems are case by case, although the records may have similar fraud scores, the reason behind the scores might be quite different.