REPORT

BUILDING UNSUPERVISED MODEL FOR PROPERTY FRAUD IN NEW YORK CITY

Prepared for:

Professor Stephen Coggeshall

Prepared by:

CHEN, Siqi

GUO, Qianfei

HUNG, Kai-Ling

JIA, Xuerong

LIN, Xiaoyan

TSAI, Pei Yu

February 20, 2019

Table of Content

1 Executive Summary

1.1 Work Performed

The goal of our project is to build an unsupervised fraud model to analyze NYC property data to determine whether there is indication of fraud. We obtained the Property Valuation and Assessment Data of New York City from the NYC Open Data website, and then performed the following steps.

1. Filled the missing values for nine key fields;
2. Created 45 new variables;
3. Z-scaled the data so that they are on the same footing;
4. Conducted PCA to reduce dimensionality of the data to seven PCs;
5. Z-scaled the seven PCs again;
6. Combined the Z-scores with a heuristic algorithm to arrive at Score 1;
7. Train an autoencoder on the seven Z-scaled PCs to reproduce seven new PCs. Score 2 would be the difference between the original input records and the autoencoder output records;
8. Combine Score 1 and Score 2 using weighted average rank orders to get the final score;
9. Sorted the final score on a descending order and further investigated into records with top 10 highest fraud scores.

1.2 Results

Our analysis found that there are mainly three causes for the anomalies we noted:

1. Unusually small building or lot sizes. For instance, building front and building depth with the value of zero or one foot;
2. Unusually high values of lot sizes; and
3. Unusually high values of the property in comparison with other properties in the same borough, zip code or tax class. The unusual values are typically in FULLVAL, AVLAND and AVTOT fields.

Further research and inquiries are needed to verify whether the anomalies we noticed were caused by human error, plausible explanations[1], or fraudulent activities.

[1] For instance, Parks of Recreation's lot size could be unusually big because a park would understandably have a big parking lot.

2 Description of Data

2.1 Data Source

The NYC Property Valuation and Assessment Data used to build our fraud model was obtained from the NYC Open Data website[2]. The data was originally provided by Department of Finance on September 2, 2011, and the most recently revised on September 10, 2018. This dataset has 32 fields and 1,070,994 records. Dataset represents NYC properties assessments for purpose of calculating Property Tax, Grant eligible properties Exemptions and/or Abatements. Data was collected and entered into the system by City employees including Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc.

2.2 Summary of Numerical Fields

This dataset has 14 numerical fields, which are summarized in Table 1 below. Fields that are marked red are fields of low percentage populated.

Table 1 – Summary of Numerical Fields

Field Name	Count	Unique Values	Value with Zero	% Populated	Mean	Max	Min	SD
LTFRONT	1,070,994	1,297	169,108	100.00%	3.66E+01	1.00E+04	0.00E+00	7.40E+01
LTDEPTH	1,070,994	1,370	170,128	100.00%	8.89E+01	1.00E+04	0.00E+00	7.64E+01
STORIES	1,014,730	112	n/a	94.75%	5.01E+00	1.19E+02	1.00E+00	8.37E+00
FULLVAL	1,070,994	109,324	13,007	100.00%	8.74E+05	6.15E+09	0.00E+00	1.16E+07
AVLAND	1,070,994	70,921	13,009	100.00%	8.51E+04	2.67E+09	0.00E+00	4.06E+06
AVTOT	1,070,994	112,914	13,007	100.00%	2.27E+05	4.67E+09	0.00E+00	6.88E+06
EXLAND	1,070,994	33,419	491,699	100.00%	3.64E+04	4.67E+09	0.00E+00	3.98E+06
EXTOT	1,070,994	64,255	432,572	100.00%	9.12E+04	2.67E+09	0.00E+00	6.51E+06
BLDFRONT	1,070,994	612	228,815	100.00%	2.30E+01	7.58E+03	0.00E+00	3.56E+01
BLDDEPTH	1,070,994	621	228,853	100.00%	3.99E+01	9.39E+03	0.00E+00	4.27E+01
AVLAND2	282,726	58,592	n/a	26.40%	2.46E+05	2.37E+09	3.00E+00	6.18E+06
AVTOT2	282,732	111,361	n/a	26.40%	7.14E+05	4.50E+09	3.00E+00	1.17E+07
EXLAND2	87,449	22,196	n/a	8.17%	3.51E+05	2.37E+09	1.00E+00	1.08E+07
EXTOT2	130,828	48,349	n/a	12.22%	6.57E+05	4.50E+09	7.00E+00	1.61E+07

2.3 Summary of Categorical Fields

This dataset has 18 categorical fields, which are summarized in Table 2 below. Fields that are marked red are fields of low percentage populated.

Table 2 – Summary of Categorical Fields

Field Name	Count	Unique Values	Most Common	% Populated
RECORD	1,070,994	1,070,994	Uniform Distributed	100.00%
BBLE	1,070,994	1,070,994	Uniform Distributed	100.00%
B	1,070,994	5	4(33.43%)	100.00%
BLOCK	1,070,994	13,984	3944(0.36%)	100.00%

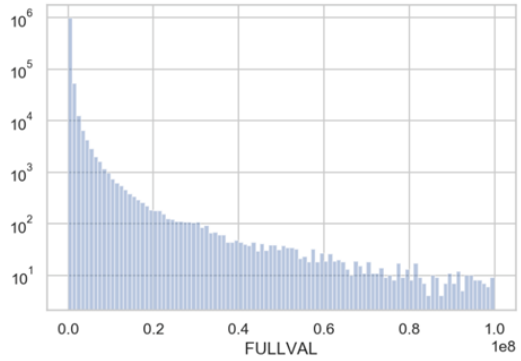
[2] External link: https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8

Field Name	Count	Unique Values	Most Common	% Populated
LOT	1,070,994	6,366	1(2.28%)	100.00%
EASEMENT	4,636	13	E(89.47%)	0.43%
OWNER	1,039,249	863,348	PARKCHESTER PRESERVAT (0.58%)	97.04%
BLDGCL	1,070,994	200	R4(13.06%)	100.00%
TAXCLASS	1,070,994	11	1(61.69%)	100.00%
EXT	354,305	4	G(75.35%)	33.08%
EXCD1	638,488	130	1017(0.1%)	59.62%
STADDR	1,070,318	839,281	501 SURF AVENUE(0.08%)	99.94%
ZIP	1,041,104	197	10314(2.36%)	97.21%
EXMPTCL	15,579	15	X1(44.37%)	1.45%
EXCD2	92,948	61	65777(6.1%)	8.68%
PERIOD	1,070,994	1	FINAL	100.00%
YEAR	1,070,994	1	2010/11	100.00%
VALTYPE	1,070,994	1	AC-TR	100.00%

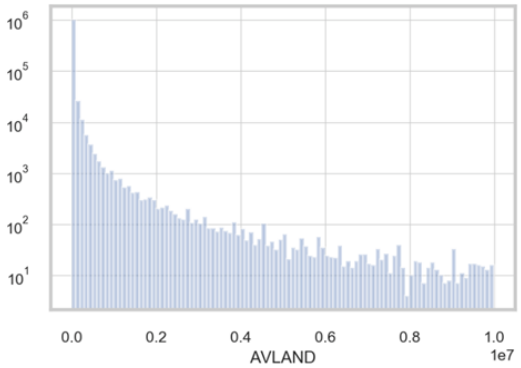
2.4 Overview of NYC Property

The histograms and bar charts below give a high-level understanding of the sizes and prices of the NYC property in our dataset. Please refer to the Appendix for a full DQR.

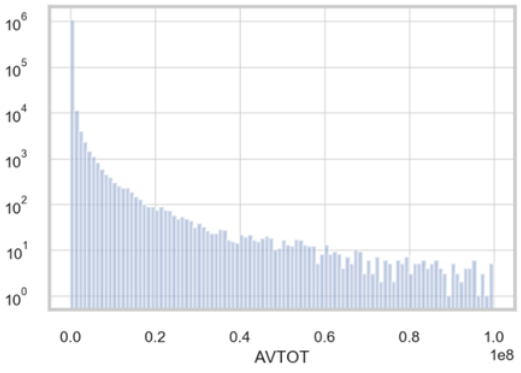
1. FULLVAL – the total market value of the property



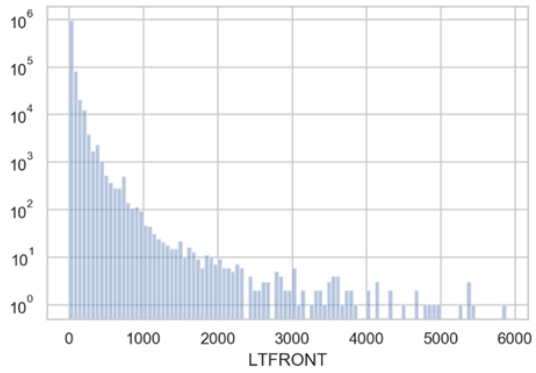
2. AVLAND – the market value of the land



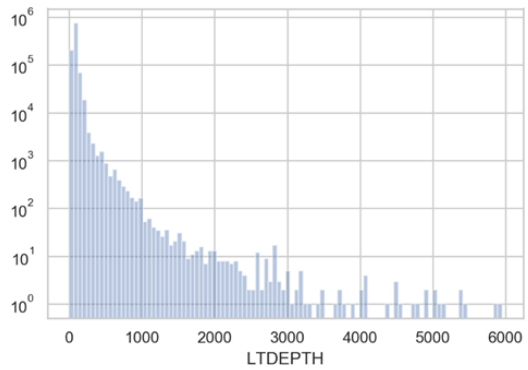
3. AVTOT – assessed value of the property



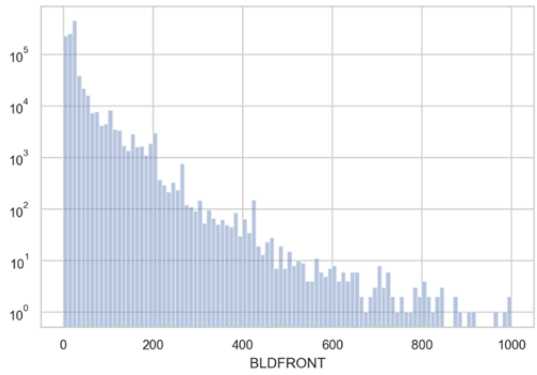
4. LTFRONT – lot frontage in feet



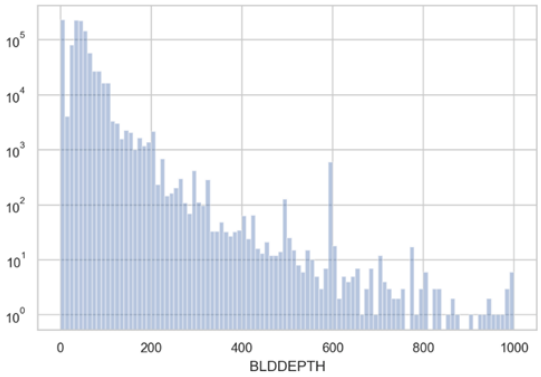
5. LTDEPTH – lot depth in feet



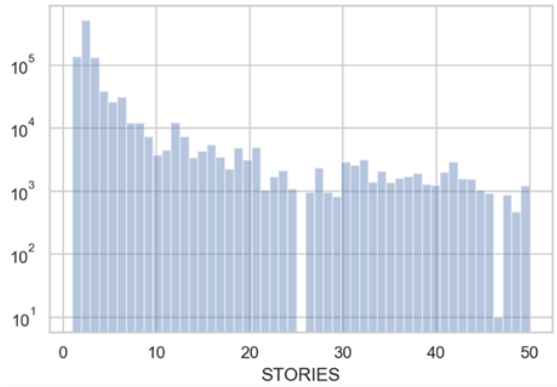
6. BLDFRONT – building frontage in feet



7. BLDDEPTH – building depth in feet



8. STORIES – number of floors in each building



3 Data Cleaning - Filling Missing Values

3.1 Missing Fields

To perform analysis on the NY property data, we need to fill in missing values for fields listed below. The goal is to fill in these fields with innocuous values that would not set off the alarm.

- ZIP
- STORIES
- FULLVAL, AVLAND, AVTOT
- LTFRONT，LTDEPTH，BLDFRONT, BLDDEPTH

3.2 Steps for ZIP

1. Grouped by B and BLOCK, filled the missing values with the most frequent zip of the group. If there are less than 5 records in the group, left it as is;
2. Grouped by B and STADDR, filled the missing values with the most frequent zip of the group. If there are less than 5 records in the group, left it as is; and

3. Grouped by B, filled the missing values with the most frequent zip of the group.

3.3 Steps for STORIES

1. Grouped by TAXCLASS and ZIP, filled the missing values with the average STORIES of the group. If there are less than 10 records in the group, left it as is;
2. Grouped by BLOCK, filled the missing values with the average STORIES of the group. If there are less than 5 records in the group, left it as is; and
3. Grouped by ZIP, filled the missing values with the average STORIES of the group.

3.4 Steps for FULLVAL, AVLAND, AVTOT

1. Grouped by ZIP, STORIES, TAXCLASS, Lot Area (LOTFRONT times LOTDEPTH) when Lot Area value is larger than 4 and smaller than 10,000; and then grouped by Building Area (BLDFRONT times BLDDEPTH) when Building Area is larger than 4 and smaller than 5,000, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
2. Grouped by ZIP, STORIES, TAXCLASS, Lot Area, and filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
3. Grouped by ZIP, STORIES, TAXCLASS, Building Area, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
4. Grouped by ZIP, TAXCLASS, Lot Area, and Building Area, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
5. Grouped by ZIP, TAXCLASS, and Building Area, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
6. Grouped by ZIP, TAXCLASS, and Lot Area, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
7. Grouped by TAXCLASS, ZIP, and STORIES, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
8. Grouped by ZIP, and STORIES, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
9. Grouped by ZIP, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is; and
10. Grouped by STORIES, filled the missing values with the average value of the group.

3.5 Steps for LTFRONT，LTDEPTH，BLDFRONT, BLDDEPTH

1. Created a new variable, defined it as the average value of FULLVAL, AVLAND and AVTOT, and transformed it into reasonable bin number;
2. Grouped by ZIP, STORIES, TAXCLASS and the bin number, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
3. Grouped by STORIES, TAXCLASS and the bin number, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is.
4. Grouped by TAXCLASS and the bin number, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is;
5. Grouped by STORIES and the bin number, filled the missing values with the average value of the group. If there are less than 10 records in the group, left it as is; and
6. Grouped by TAXCLASS, filled the missing values with the average value of the group.

4 Creating Variables

4.1 Creating 45 New Variables

The goal is to build fraud model to identify anomalous values in the fields such as FULLVAL, AVTOT and AVLAND. To that end, we need to create variables and normalize the variables so that each record can be compared with similar properties on a scaled basis. We created 45 variables for each record and the steps for creating these 45 variables are as follows:

1. Created three measurements of the square footage of each property;
   • lotarea = LTFRONT * LTDEPTH
   • bldarea = BLDFRONT * BLDDEPTH
   • bldvol = bldarea* STORIES
2. Normalized FULLVALUE, AVTOT, AVLAND, respectively, by the lotarea, bldarea and bldvol for each record;
3. Calculated nine ratios of value per square foot for each record;
4. Grouped ratios by zip5, zip3, taxclass, borough and all, in order to compare the ratios with properties under similar condition or classification;
5. Calculated the averages of the nine ratios for each group; and
6. Divided each of the nine ratios by the average ratio in that group based on zip5, zip3, taxclass, borough.

4.2 Definitions of the 45 Variables

Each of the 45 variables is an indicator of how the record's value per square foot deviated from the average of properties within its group. Refer to Table 3 below for the definition of the new 45 variables created.

Table 3 – Definitions of the 45 Variables

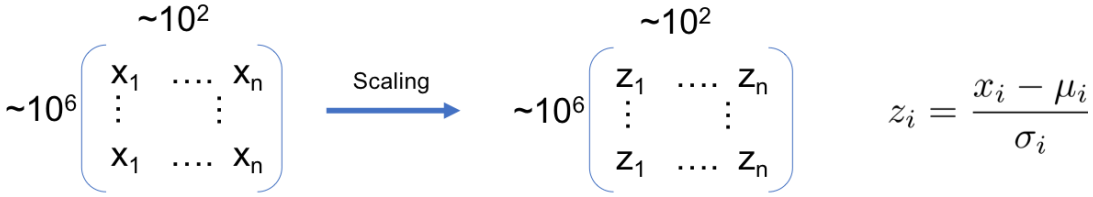
Variable	Definition
V1	FULLVAL per square footage of LOT compared to the average ratio in same zip5.
V2	FULLVAL per square footage of LOT compared to the average ratio in same zip3.
V3	FULLVAL per square footage of LOT compared to the average ratio in same taxclass.
V4	FULLVAL per square footage of LOT compared to the average ratio in same borough.
V5	FULLVAL per square footage of LOT compared to the average ratio.
V6	FULLVAL per square footage of Building compared to the average ratio in same zip5.
V7	FULLVAL per square footage of Building compared to the average ratio in same zip3.
V8	FULLVAL per square footage of Building compared to the average ratio in same taxclass.
V9	FULLVAL per square footage of Building compared to the average ratio in same borough.
V10	FULLVAL per square footage of Building compared to the average ratio.
V11	FULLVAL per square footage of Total Building compared to the average ratio in same zip5.
V12	FULLVAL per square footage of Total Building compared to the average ratio in same zip3.
V13	FULLVAL per square footage of Total Building compared to the average ratio in same taxclass.
V14	FULLVAL per square footage of Total Building compared to the average ratio in same borough.
V15	FULLVAL per square footage of Total Building compared to the average ratio.
V16	AVLAND per square footage of LOT compared to the average ratio in same zip5.
V17	AVLAND per square footage of LOT compared to the average ratio in same zip3.
V18	AVLAND per square footage of LOT compared to the average ratio in same taxclass.
V19	AVLAND per square footage of LOT compared to the average ratio in same borough.
V20	AVLAND per square footage of LOT compared to the average ratio.
V21	AVLAND per square footage of Building compared to the average ratio in same zip5.
V22	AVLAND per square footage of Building compared to the average ratio in same zip3.

Variable	Definition
V23	AVLAND per square footage of Building compared to the average ratio in same taxclass.
V24	AVLAND per square footage of Building compared to the average ratio in same borough.
V25	AVLAND per square footage of Building compared to the average ratio.
V26	AVLAND per square footage of Total Building compared to the average ratio in same zip5.
V27	AVLAND per square footage of Total Building compared to the average ratio in same zip3.
V28	AVLAND per square footage of Total Building compared to the average ratio in same taxclass.
V29	AVLAND per square footage of Total Building compared to the average ratio in same borough.
V30	AVLAND per square footage of Total Building compared to the average ratio.
V31	AVTOT per square footage of LOT compared to the average ratio in same zip5.
V32	AVTOT per square footage of LOT compared to the average ratio in same zip3.
V33	AVTOT per square footage of LOT compared to the average ratio in same taxclass.
V34	AVTOT per square footage of LOT compared to the average ratio in same borough.
V35	AVTOT per square footage of LOT compared to the average ratio.
V36	AVTOT per square footage of Building compared to the average ratio in same zip5.
V37	AVTOT per square footage of Building compared to the average ratio in same zip3.
V38	AVTOT per square footage of Building compared to the average ratio in same taxclass.
V39	AVTOT per square footage of Building compared to the average ratio in same borough.
V40	AVTOT per square footage of Building compared to the average ratio.
V41	AVTOT per square footage of Total Building compared to the average ratio in same zip5.
V42	AVTOT per square footage of Total Building compared to the average ratio in same zip3.
V43	AVTOT per square footage of Total Building compared to the average ratio in same taxclass.
V44	AVTOT per square footage of Total Building compared to the average ratio in same borough.
V45	AVTOT per square footage of Total Building compared to the average ratio.

5 Dimensionality Reduction

5.1 Principal Component Analysis (PCA)

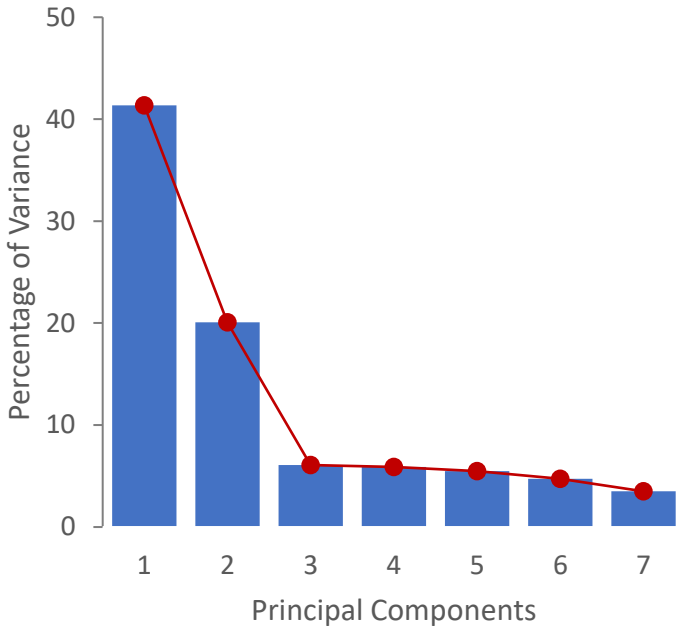
The 45 newly created variables are on different scale and correlated with each other to certain extent. Therefore, we performed Principal Component Analysis to reduce dimensionality and correlations before calculating the final fraud score. Z-scaling can convert variables to same scale and same unit to make the correlation matrix for applying PCA.



$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

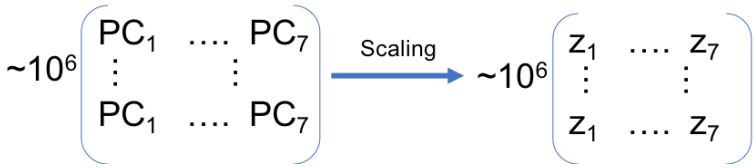
PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

The PCA produced a reduced number of variables to seven Principal Components ('PCs'). These seven PCs are able to explain 87% of variance.



5.2 Z-scaling

After reducing 45 variables to 7 PCs, we scaled them again so that the outlier records would get unusually large z-scores and thus stand out from the population.



$$z_i = \frac{PC_i - \mu_{PC_i}}{\sigma_{PC_i}}$$

6 Algorithms

6.1 Manhattan Distance Score

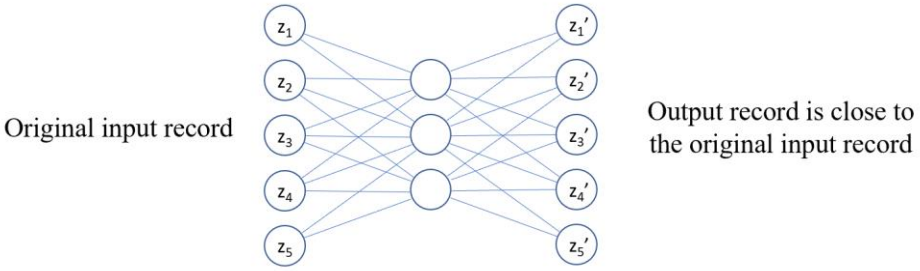
After Dimensionality Reduction, all values for all records are z-scaled. We call these z-scaled variables z-scores. We add up these z-scores of each record, without letting them cancel each other out.

$$s_i = \left(\sum_k |z_k^i|^n \right)^{1/n}$$, n = 1

When a value is unusual, its z-score would be unusual, so the sum of z-sores shows the distance of the value from the origin.

6.2 Autoencoder Score

An autoencoder is a model trained to output the original vector input. We train an autoencoder on the entire data set. The model will learn to reproduce the data records as well as possible, and will learn the nature of the bulk of the data.



The Autoencoder focuses on the majority of the data. When a value is abnormal, the autoencoder does a poor job reproducing the value. The records that aren't reproduced well are what we're looking for. After the model is trained, the difference between the original input vector and the model output vector is the fraud score for that record.

$$s_i = \left(\sum_k |z_k'^i - z_k^i|^n \right)^{1/n}$$, n = 1

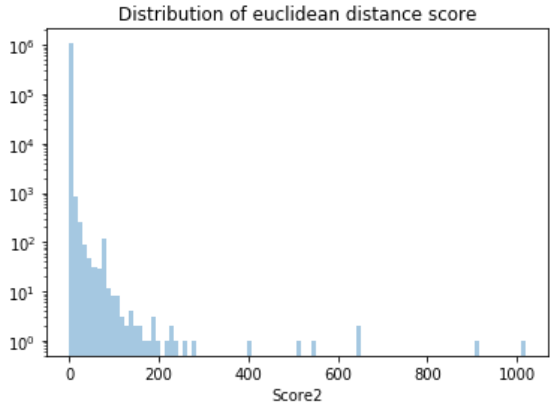
6.3 Weighted Score

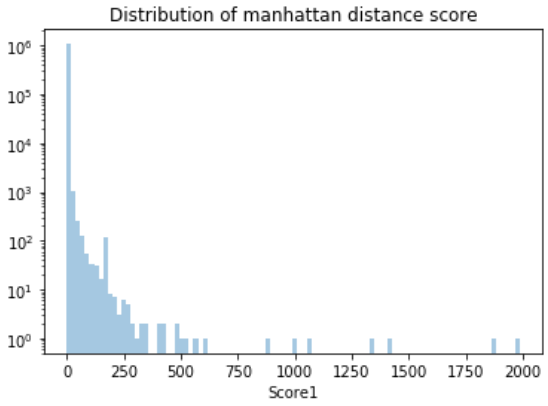
To scale two scores to the same level, we use quantile binning method. For each score, we replace the score with the record's rank order after sorting by the score. Based on the distribution of scores, we decided to set our weighted score a linear combination of the two binned scores mentioned above.

Weighted Score = 0.5 * Manhattan Distance Score + 0.5 * Autoencoder Score

7 Results

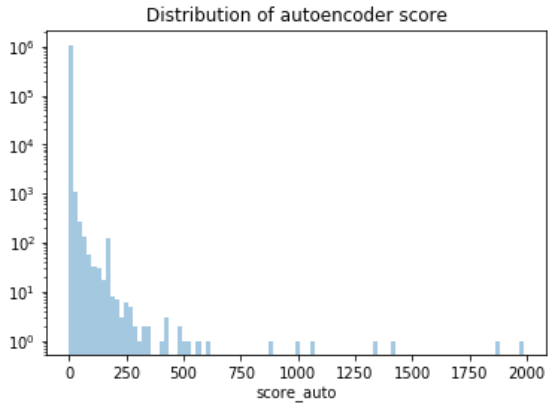
We sorted the records by fraud score in a descending order and selected the records with top 10 highest scores. Below is an overview of the distribution of fraud scores from both methods. Fraud score calculated based on z-scores algorithm exhibited right-skewed distribution (refer to the illustrations below). We examined the score from the perspective of both Manhattan distance and Euclidean distance. As expected, most of the records had low fraud scores and there were a few outliers relative to the number of total records.

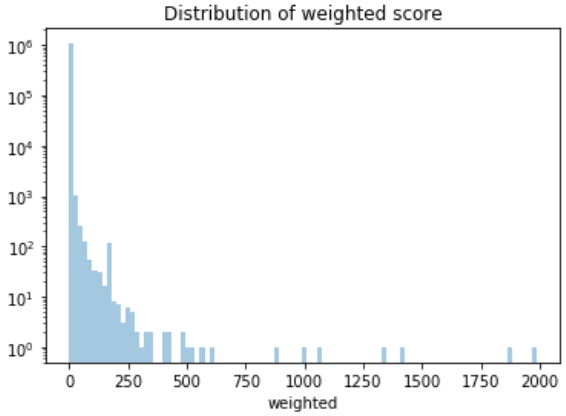




The distribution of fraud scores calculated using autoencoder algorithm also had a right skew (refer to the illustrations below).

For weighted average score, we allotted 50% and 50% weights to scores obtained via autoencoder and Z-scaling, respectively. Distribution of weighted score fraud score has the approximately same pattern as the three above.





We manually examined the records with high fraud scores. Some properties are owned by government properties and universities. Although these properties are anomalies on several fronts, we know that government properties, parks and universities have very low risk of tax fraud. The top candidates for potential tax fraud have been listed below.

Top 10 High Fraud Score Records (First half)					
RECORD	85886	565392	565398	585118	585120
BBLE	1.01E+09	3.09E+09	3.09E+09	4.00E+09	4.00E+09
B	1	3	3	4	4
BLOCK	1254	8590	8591	420	420
LOT	10	700	100	1	101
EASEMENT					
OWNER	PARKS AND RECREATION	U S GOVERNMENT OWNRD	DEPT OF GENERAL SERVI	NEW YORK CITY ECONOMI	
BLDGCL	Q1	V9	V9	O3	O3
TAXCLASS	4	4	4	4	4
LTFRONT	4000	117	466	298	139
LTDEPTH	150	108	1009	402	342
EXT					
STORIES	1			20	20
FULLVAL	7.02E+07	4.33E+09	2.31E+09	3.44E+06	2.15E+06
AVLAND	3.15E+07	1.95E+09	1.04E+09	1.55E+06	9.68E+05
AVTOT	3.16E+07	1.95E+09	1.04E+09	1.55E+06	9.68E+05
EXLAND	3.15E+07	1.95E+09	1.04E+09	0.00E+00	0.00E+00
EXTOT	3.16E+07	1.95E+09	1.04E+09	0.00E+00	0.00E+00
EXCD1	2231	2231	2191		
STADDR	JOE DIMAGGIO HIGHWAY	FLATBUSH AVENUE	FLATBUSH AVENUE	28-10 QUEENS PLAZA SOUTH	28 STREET
ZIP				11101	
EXMPTCL	X1	X1	X1	X1	
BLDFRONT	8	0	0	1	1
BLDDEPTH	8	0	0	1	1
AVLAND2	2.81E+07	8.48E+08	4.35E+08	1.59E+06	9.75E+05
AVTOT2	2.83E+07	8.48E+08	4.35E+08	1.59E+06	9.75E+05
EXLAND2	2.81E+07	8.48E+08	4.35E+08		
EXTOT2	2.83E+07	8.48E+08	4.35E+08		
EXCD2					
PERIOD	FINAL	FINAL	FINAL	FINAL	FINAL
YEAR	2010/11/1	2010/11/1	2010/11/1	2010/11/1	2010/11/1
VALTYPE	AC-TR	AC-TR	AC-TR	AC-TR	AC-TR

Top 10 High Fraud Score Records (Second half)					
RECORD	585439	632816	920628	935158	1067360
BBLE	4.00E+09	4.02E+09	4.16E+09	5.00E+09	5.08E+09
B	4	4	4	5	5
BLOCK	459	1842	15577	13	7853
LOT	5	1	29	60	85
EASEMENT					
OWNER	11-01 43RD AVENUE REA	864163 REALTY, LLC	PLUCHENIK, YAAKOV	RICH-NICH REALTY,LLC	
BLDGCL	H9	D9	A1	D3	B2
TAXCLASS	4	2	1	2	1
LTFRONT	94	157	91	136	1
LTDEPTH	165	95	100	132	1
EXT					
STORIES	10	1	2	8	2
FULLVAL	3.71E+06	2.93E+06	1.90E+06	1.04E+06	8.36E+05
AVLAND	2.52E+05	1.32E+06	9.76E+03	2.36E+05	2.88E+04
AVTOT	1.67E+06	1.32E+06	7.58E+04	4.68E+05	5.02E+04
EXLAND	0.00E+00	0.00E+00	0.00E+00	2.21E+05	0.00E+00
EXTOT	1.42E+06	0.00E+00	0.00E+00	4.53E+05	0.00E+00
EXCD1	1986			5113	
STADDR	11-01 43 AVENUE	86-55 BROADWAY	7-06 ELVIRA AVENUE	224 RICHMOND TERRACE	20 EMILY COURT
ZIP	11101	11373	11691	10301	10307
EXMPTCL					
BLDFRONT	1	1	1	1	36
BLDDEPTH	1	1	1	1	45
AVLAND2		1.20E+06		2.10E+05	
AVTOT2		1.20E+06		7.48E+05	
EXLAND2				1.95E+05	
EXTOT2				7.33E+05	
EXCD2					
PERIOD	FINAL	FINAL	FINAL	FINAL	FINAL
YEAR	2010/11/1	2010/11/1	2010/11/1	2010/11/1	2010/11/1
VALTYPE	AC-TR	AC-TR	AC-TR	AC-TR	AC-TR

Clearly, for some of these records, the full value is exceptionally high. Further scrutiny reveals that such records have no information about lot front, lot depth, etc. For some other records, the full value of the property per build area is either excessively high or low. Since, a lot of these properties are owned by real estate firms, we can infer that either the properties owned by them are significantly different from an average property or they might be exploiting loopholes (and/or committing potential tax fraud) in the property tax law.

Record No. 632816: This property only has a one-story building and its full value is about $3M. This results in a very high full value per unit building volume, which indicate that there might be some fraud.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
632816	864163 REALTY, LLC	D9	2	157	95	1	1	1	2930000	1318500	1318500	11373

Record No. 1067360: The building front and depth of this property are both 1 and it has very high assessed value of land per unit lot area. It's value of land per unit lot area is unusually high compared with other properties which have same TAXCLASS in the same zip code. Besides the property has missing owner name, which increase its probability of fraud.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
1067360		B2	1	1	1	36	45	2	836000	28800	50160	10307

Record No. 85886: This property is owned by New York government department of parks and recreation. The building front and depth of this property are both 8 but the lot area is 6e+05. The full value per unit lot area is unusually high.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
85886	PARKS AND RECREATION	Q1	4	4000	150	8	8	1	70214000	31455000	31596300	

Record No. 565392: This property is owned by New York government. This record has no BLDFRONT, BLDDEPTH, zip and stories data. Besides, FULLVAL, AVLAND and AVTOT are billions of dollars, unusually high with respect to its TAXCLASS, LTFRONT and LTDDEPTH.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
565392	U S GOVERNMENT OWNRD	V9	4	117	108	0	0		4.326E+09	1.947E+09	1.95E+09	

Record No. 565398: This property is owned by New York government department of general service. FULLVAL, AVLAND and AVTOT are billions of dollars, unusually high with respect to its TAXCLASS, LTFRONT and LTDDEPTH.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
565398	DEPT OF GENERAL SERVI	V9	4	466	1009	0	0		2.311E+09	1.04E+09	1.04E+09	

Record No. 585118: This property is owned by New York government department of economics. For this property, average land per unit building volume and full value per unit building volume seems very high.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
585118	NEW YORK CITY ECONOMI	O3	4	298	402	1	1	20	3443400	1549530	1549530	11101

Record No. 585439: For this property, average land per unit building volume and full value per unit building volume seems very high.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
585439	11-01 43RD AVENUE REA	H9	4	94	165	1	1	10	3712000	252000	1670400	11101

Record No. 585120: The front and depth of this building is 1, which is too small as compared to its value $2.1m and 20 stories. Besides, average land per unit building volume and full value per unit building volume seems very high.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
585120		O3	4	139	342	1	1	20	2151600	968220	968220	

Record No. 920628: The front and depth of this building is 1. Small front and depth, only 2 stories and the $1.9m full value does not match perfectly. While the total market value and land value are too low. Besides, contrary to the small front and depth, the building has at least 1,902 ft² volume according to other resource.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
920628	PLUCHENIK, YAAKOV	A1	1	91	100	1	1	2	1900000	9763	75763	11691

Record No. 935158: The front and depth of this building is 1, which is too small as compared to its value $10.4m, 136 lot front, 132 lot depth and 8 stories. Probably because the old data was not updated since the building was built in 2012.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
935158	RICH-NICH REALTY,LLC	D3	2	136	132	1	1	8	1040000	236250	468000	10301

8 Conclusions

8.1 Steps Performed

1. **DQR and Data Cleaning**: We started with exploratory analysis of the data which included descriptive analysis, data visualization, correlation analysis, missing data analysis, etc. We then cleaned up the data by filling in missing values;
2. **Creating 45 Variables**: We created 45 new variables mainly based on seven fields: FULLVAL, AVLAND, AVTOT, BLDFRONT, BLDDEPTH, LTFRONT, LTDEPTH, etc;
3. **Z-scaling**: We scaled the 45 new variables;
4. **Principal Component Analysis**: PCA was conducted to reduce the 45 variables to seven PCs, which explained more than 87% of the variance. The PCA reduced dimensionally and removed correlation between different variables;
5. **Z-scaling (again)**: We z-scaled again on the seven PCs;
6. **Heuristic algorithm:** we combined model variables with a heuristic algorithm that utilizes the sum of absolute z-scores; and
7. **Autoencoder:** We used autoencoder to train the model into reproduce the original data. the reproduction error being a measure of the record's unusualness and thus, a fraud score. Two approaches were used for calculating fraud scores – zscore and autoencoder algorithms.

8.2 Results

Fraud scores from both autoencoder and z-scores were skewed to the right. The algorithm produced high fraud scores to a lot of government properties and parks, which is expected because these properties generally have some missing values including missing building fronts and depths, have fewer stories and at the same time have high property value. Closer analysis of these records reveals that the full value of some of these properties is exceptionally high/low and they do not have complete property data (e.g., missing values in the lot depth field, lot front field, etc). For some other records, the full value of the property per building area is either excessively high or low.

8.3 Recommendations

If given more time, we recommend doing the following to improve our fraud model and further investigate the potential property fraud:

1. Fill missing data in a different way – There is a lot of missing data in original dataset. If we had more time, we would explore different ways of filling missing values to see if we could identify different anomalous records;
2. Seek opinions from Subject Matter Experts – Expert in fraud examination or in real estate industry could give us more insight in potential causes of the anomalies in the data, and help us form hypothesis with regard to the fraudulent activities;
3. Verify the anomalous data against third-party data – We could conduct site visit or internet research to verify the veracity of the anomalous data; and
4. Refine the model with more data – If we can get more information about property owners, crime rates, average income in the ZIP codes' areas, etc., it may be useful in improving the model further.

9 Appendix