Case Study on Top 10 Properties that have highest Fraud Scores

RECORD	917942	956520	658933	1059883	139726	665158	116647	684704	12076	333412
OWNER	LOGAN PROPERTY, INC.	TROMPETA RIZALINA	WAN CHIU CHEUNG		BRADHURST EQUITIES, L	ST JOHNS CEMETERY	MF ASSOCIATES OF NEW	W RUFERT	15 WORTH STREET PROPE	SPOONER ALSTON
LTFRONT	4910	25	25	5	4	1412	25	2	74	17
LTDEPTH	0	91	100	5	5	2532	75	2	150	85
STORIES	3	3	3			1	35		1	3
FULLVAL	374,019,883.00	348,200.00	776,000.00	-	-	29,355,000.00	161,000,000.00	-	2,610,000.00	9060
AVLAND	1,792,808,947.00	15,600.00	26,940.00	-	-	13,140,000.00	19,215,000.00	-	1,170,000.00	3874
AVTOT	4,668,308,947.00	20,892.00	46,560.00	-	-	13,209,750.00	72,450,000.00	-	1,174,500.00	4077
STADDR	154-68 BROOKVILLE BOULEVARD	12 ONEIDA AVENUE	54-76 83 STREET	SAGONA COURT	BRADHURST AVENUE	80-01A METROPOLITAN AVENUE	1849 2 AVENUE	69 STREET	170 WEST BROADWAY	37 MONROE STREET
ZIP	11422	10301	11373			11379	10128		10013	11238
BLDFRONT	0	1812	2500	0	0	12	70	0	5	4017
BLDDEPTH	0	5020	5600	0	0	18	456	0	5	42
r1	0.8	-0.1	0.2	205.0	256.3	-0.4	162.1	176.6	0.0	-0.4
r2	68.2	-0.5	-0.5	0.0	0.0	133.3	4.4	-0.4	102.3	-0.5
r3	47.4	-0.5	-0.5	-0.3	-0.3	278.8	-0.2	-0.4	214.1	-0.5
r4	42.2	-0.1	0.0	253.7	317.2	-0.1	147.5	32.9	1.4	-0.1
r5	896.6	-0.1	-0.1	0.2	0.2	162.8	1.6	-0.1	125.3	-0.1
r6	643.3	-0.1	-0.1	0.0	0.0	350.6	0.0	-0.1	269.7	-0.1
r7	37.8	-0.1	0.0	251.4	314.2	-0.1	191.5	11.3	0.4	-0.1
r8	937.3	-0.1	-0.1	0.2	0.2	65.7	2.4	-0.1	50.5	-0.1
r9	898.6	-0.1	-0.1	0.1	0.1	189.0	0.1	-0.1	145.2	-0.1
r1inv	-0.1	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	-0.1	-0.1	0.1
r2inv	-0.1	27.0	19.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	20.4
r3inv	-0.2	14.0	11.2	-0.1	-0.1	-0.2	-0.1	-0.1	-0.2	11.4
r4inv	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
r5inv	-0.2	25.6	25.1	-0.2	-0.2	-0.2	-0.2	-0.1	-0.2	9.0
r6inv	-0.3	9.1	9.1	-0.3	-0.3	-0.3	-0.3	-0.2	-0.3	5.3
r7inv	-0.2	-0.1	-0.1	-0.2	-0.2	0.0	-0.2	-0.2	-0.2	0.0
r8inv	-0.1	41.4	38.2	-0.1	-0.1	-0.1	-0.1	0.1	-0.1	14.8
r9inv	-0.2	17.3	16.8	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	10.3
r1_zip5	2.2	0.2	0.2	342.0	176.0	-0.5	112.1	202.7	-0.1	-0.5
r2_zip5	66.6	-0.5	-0.5	0.0	0.1	93.1	4.2	-0.4	120.7	-0.5
r3_zip5	31.4	-0.4	-0.4	-0.2	-0.1	136.4	0.1	-0.3	354.7	-0.4
r4_zip5	102.4	0.0	0.1	492.4	180.5	-0.1	80.1	67.3	0.6	-0.1
r5_zip5	635.1	-0.1	-0.1	0.3	0.6	239.3	2.1	-0.1	139.7	-0.1
r6_zip5	429.9	-0.1	-0.1	0.0	0.3	272.1	0.1	-0.1	245.2	-0.1
r7_zip5	146.2	-0.1	0.0	560.5	140.7	-0.1	88.2	26.0	0.1	-0.2
r8_zip5	767.9	-0.1	-0.1	0.8	0.5	172.5	3.1	-0.1	59.4	-0.1
r9_zip5	631.2	-0.1	-0.1	0.2	0.3	243.7	0.4	-0.1	192.4	-0.1
r1inv_zip5	-0.1	-0.1	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.1	0.1
r2inv_zip5	-0.1	26.5	15.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	25.9
r3inv_zip5	-0.1	7.8	8.3	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	13.1
r4inv_zip5	-0.3	-0.3	-0.3	-0.3	-0.3	-0.1	-0.3	-0.3	-0.3	-0.2
r5inv_zip5	-0.2	23.3	15.3	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	4.4
r6inv_zip5	-0.2	7.5	5.3	-0.2	-0.2	-0.2	-0.2	0.2	-0.2	2.7
r7inv_zip5	-0.2	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	-0.2	-0.2	0.1
r8inv_zip5	-0.1	30.1	27.7	-0.1	-0.1	-0.1	-0.1	0.2	-0.1	15.6
r9inv_zip5	-0.2	11.0	12.5	-0.1	-0.1	-0.2	-0.2	0.5	-0.2	10.6
r1_taxclass	0.5	-0.1	0.1	142.3	178.0	-0.3	347.5	383.3	0.0	-0.3
r2_taxclass	47.6	-0.3	-0.3	0.1	0.1	92.9	11.2	0.0	71.3	-0.3
r3_taxclass	46.6	-0.3	-0.3	-0.1	-0.1	273.4	2.0	0.0	210.0	-0.3
r4_taxclass	20.0	0.1	0.3	121.6	152.1	-0.3	329.6	387.6	0.4	-0.3
r5_taxclass	670.2	-0.2	-0.2	0.0	0.0	121.6	6.9	0.0	93.5	-0.2
r6_taxclass	514.7	-0.2	-0.2	-0.1	-0.1	280.5	1.2	0.0	215.7	-0.2
r7_taxclass	21.7	0.0	0.3	145.4	181.8	-0.3	344.5	435.8	0.0	-0.3
r8_taxclass	766.1	-0.2	-0.2	0.0	0.0	53.6	6.8	0.0	41.1	-0.2
r9_taxclass	755.9	-0.2	-0.2	-0.1	-0.1	158.9	1.2	0.0	122.0	-0.2
r1inv_taxclass	-0.2	0.0	-0.1	-0.2	-0.2	-0.1	-0.2	-0.2	-0.2	7.1
r2inv_taxclass	-0.1	762.7	562.8	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	286.7
r3inv_taxclass	-0.1	775.1	619.5	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	198.6
r4inv_taxclass	-0.5	-0.2	-0.3	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.3
r5inv_taxclass	-0.4	441.4	433.7	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4	19.9
r6inv_taxclass	-0.5	348.3	346.0	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	15.0
r7inv_taxclass	-0.4	-0.1	-0.2	-0.4	-0.4	-0.2	-0.4	-0.4	-0.3	3.4
r8inv_taxclass	-0.2	472.6	436.1	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	305.7
r9inv_taxclass	-0.4	422.8	409.8	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4	255.7
value_ratio	3.7	-0.1	-0.1	0.3	0.3	0.4	0.2	-0.1	0.4	0.4
Record	917942	956520	658933	1059883	139726	665158	116647	684704	12076	333412
max of all z score variable	937.3	775.1	619.5	560.5	317.2	350.6	347.5	435.8	354.7	305.7
Field with strange value	AVTOT and AVLAND unreasonably high	BLDFRONT, BLDDEPTH unusually high compared to Lot size	BLDFRONT, BLDDEPTH unusually high compared to Lot size	LTDEPTH and LTFRONT is very low and building size is 0	LTDEPTH and LTDEPTH is very low	Dollar value for building size is unsuallly high (BLDFRONT, BLDDEPTH unsually LOW)	Dollar valye for lot size is unsually high (Lot size is small as compared to Buliding size.)	Lot size (LTDEPTH & LTFRONT) unsually low and also BLDEPTH & BLDFRONT is 0	BLDFRONT, BLDDEPTH unusually LOW compared to Lot size	BLDFRONT is unsually large(very large)

Case Study on Top 10 Properties that have highest Fraud Scores

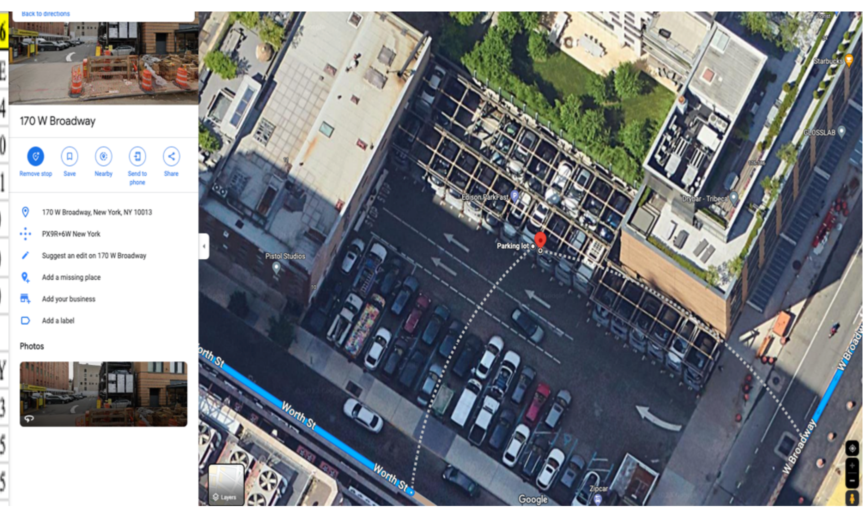
Some examples of properties that have high fraud scores:

BLDFRONT is unusually large as compared to another dimension, and the FULLVAL, AVLAND, AVTOT is very low, for the residential property which indicate there might be some fraud.

RECORD	333412
BBLE	3019850059
BORO	3
BLOCK	1985
LOT	59
EASEMENT	
OWNER	SPOONER ALSTON
LTFRONT	17
LTDEPTH	85
STORIES	3
FULLVAL	9060
AVLAND	3874
AVTOT	4077
STADDR	37 MONROE STREET
ZIP	11238
BLDFRONT	4017
BLDDEPTH	42

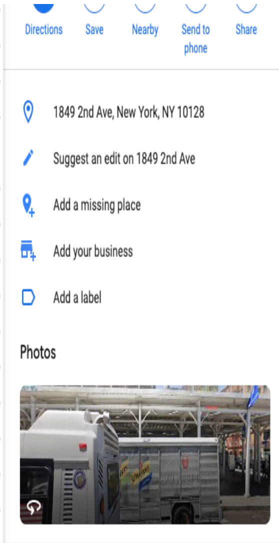
BLDFRONT, BLDDEPTH unusually LOW as compared to Lot size, for this property.

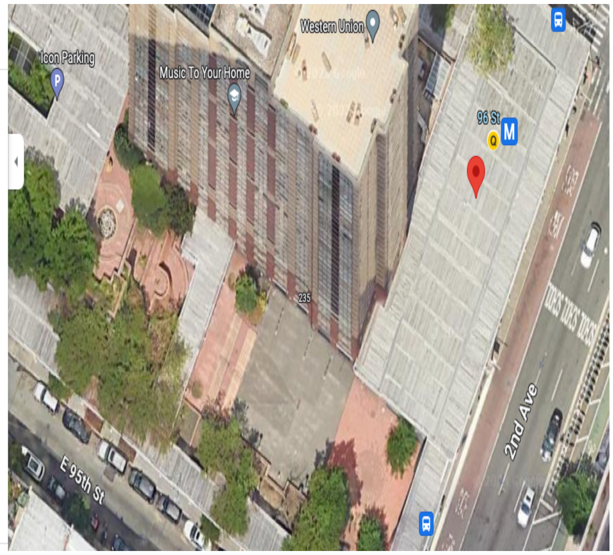
RECORD	12076
OWNER	15 WORTH STREET PROPE
LTFRONT	74
LTDEPTH	150
STORIES	1
FULLVAL	2,610,000.00
AVLAND	1,170,000.00
AVTOT	1,174,500.00
EXLAND	-
STADDR	170 WEST BROADWAY
ZIP	10013
BLDFRONT	5
BLDDEPTH	5



Dollar value for lot size is unusually high (Lot size is small as compared to Building size.)

RECORD	116647
BLOCK	1541
LOT	21
OWNER	MF ASSOCIATES OF NEW
LTFRONT	25
LTDEPTH	75
STORIES	35
FULLVAL	161,000,000.00
AVLAND	19,215,000.00
AVTOT	72,450,000.00
STADDR	1849 2 AVENUE
ZIP	10128
BLDFRONT	70
BLDDEPTH	456





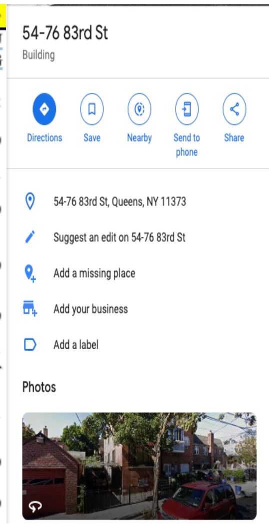
Information about address from net : This is a business registration address for five companies. These are some of the names: Dahlgren, Lauren and Rite Aid. Mf Associates of New York LLC owns this real estate property. 1986 is the year the property was built. The property is 37 years old, which is 69 years younger than the average age of a building in New York of 106 years. Estimated value for the property $106,415,550. The property features 968,264 sqft of living area. The size of the land lot is 92,927 sqft. The building has 35 floors. (https://clustrmaps.com/a/2gf1mv/)

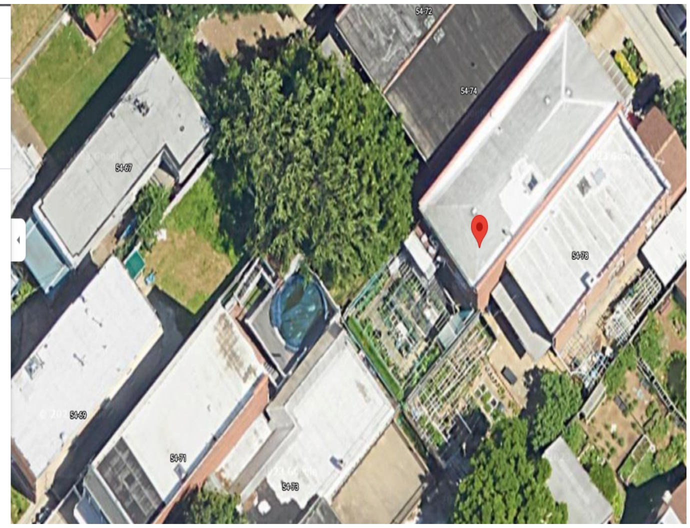
Dollar value for building size is unsuallly high (BLDFRONT, BLDDEPTH unusually LOW)

RECORD	665158
OWNER	ST JOHNS CEMETERY
LTFRONT	1412
LTDEPTH	2532
STORIES	1
FULLVAL	29,355,000.00
AVLAND	13,140,000.00
AVTOT	13,209,750.00
STADDR	80-01A METROPOLITAN AVENUE
ZIP	11379
BLDFRONT	12
BLDDEPTH	18

BLDFRONT, BLDDEPTH unusually high compared to Lot size

RECORD	658933
OWNER	WAN CHIU CHEUNG
LTFRONT	25
LTDEPTH	100
STORIES	3
FULLVAL	776,000.00
AVLAND	26,940.00
AVTOT	46,560.00
STADDR	54-76 83 STREET
ZIP	11373
BLDFRONT	2500
BLDDEPTH	5600





Observations after investigation of five fraud cases:

In our analysis, we identified these reasons for the anomalies we observed in the data:

- The dollar value for building size is unusually high when the BLDFRONT and BLDDEPTH are unusually low. (Compared to other properties in the same borough, zip code, or tax class, which are typically found in the FULLVAL, AVLAND, and AVTOT fields.)
- Some properties have BLDFRONT and BLDDEPTH values that are unusually low compared to the lot size, which may indicate incorrect or incomplete building size data.
- There are usually high values for lot sizes for very small building front and depth.

Observations and steps followed:

To examine the top properties with the highest fraud scores, we sort the records by the final fraud score in descending order and focus on the top 100 records. Additionally, a heatmap of the variable z-scores is used to identify which variables are driving the high score for these top properties. This information helped us better to understand the reasons behind the high fraud score and guide further investigation into potential fraudulent activities.

Steps followed for using fraud score after feature Engineering:

1. **Data cleaning**: Clean the data by removing irrelevant information and handling missing data using imputation techniques. The missing values for nine key fields were filled.

2. **Feature engineering:** Create new variables that are designed to look for the kinds of anomalies or fraud you are interested in. In this case, variables that help to identify unusual property valuations were created.

3. **Dimensionality reduction**: Since the dataset has a high number of features, it's important to reduce the number of features to avoid overfitting. One way to do this is to remove correlations and reduce dimensionality using principal component analysis (PCA). Z-scale the data so that they are on the same footing before conducting PCA to reduce dimensionality of the data to 5 principal components (PCs). Z-scale the 5 PCs again to make each retained PC equally important.

4. **Anomaly detection**: Use two different anomaly detection algorithms to identify potential fraud cases. The first method looks for outliers in the final scaled PC space using a **Minkowski distance** from the origin. The second method involves building a simple autoencoder, and the fraud score is then the reproduction error.

5. **Combining scores**: Since we have two scores, we average them to obtain a final fraud score. Replace the score with its rank order, and then average the rank-ordered scores for the final score.

6. **Sort the data:** Sort the records by the final score and explore the top records to investigate potential fraud cases. A heat map of the variable z-scores can help identify which variables are driving the top scores.

Result:

The fraud scores produced by both the autoencoder, and z-scores were found to be skewed to the right. This was expected because many even after removing some of the government properties and parks many were left that have missing values, including missing building fronts and depths, and they have high property value with fewer stories. Upon closer analysis of the data, we found that some properties had exceptionally high or low values, while others had incomplete property data, such as missing values in the lot depth or lot front fields. And also, some of the properties whose lot size was very small, but their total values were very high. Additionally, the full value of the property per building area was found to be either excessively high or low in some cases (compared to other properties in the same borough, zip code, or tax class).

To improve the fraud model and investigate potential property fraud further, we recommend taking the following steps:

1. Explore additional data cleaning methods: While we have used various methods to fill in missing values, we could explore additional techniques such as imputation methods to identify and clean the data more effectively.

2. Seek expert opinions: Subject matter experts in the field of fraud examination and real estate industry could provide valuable insights into the potential causes of the anomalies in the data and help form hypotheses regarding fraudulent activities.

3. Verify anomalous data against third-party sources: Conducting site visits or internet research could help verify the integrity of anomalous data and strengthen the fraud detection model.

4. Incorporate additional data sources: Incorporating additional data sources such as property ownership, crime rates, and average income in the ZIP codes' areas could provide a more comprehensive picture of the properties and owners, and ultimately improve the model's accuracy.