Property Tax Fraud Algorithm - Final Report

Executive Summary

In this project, we aimed to develop a fraud detection algorithm for property tax data using anomaly detection techniques. The dataset used for this analysis consisted of 1,070,994 records, each representing a property within the city. We performed extensive data cleaning, including removing government-owned properties and imputing missing values. After data cleaning, we created 61 variables to measure the unusualness of property values and sizes. These variables included ratios comparing market value and land value to lot size and building size, as well as grouped averages based on ZIP code and tax class.

For dimensionality reduction, we applied a technique to reduce the number of variables while preserving the maximum amount of information. Principal Component Analysis (PCA) was used, resulting in 10 principal components that explained over 80% of the data's variance. Two anomaly detection methods were employed: Z-Score Outliers and Autoencoder. Z-Score Outliers measured the distance of each record from the origin using standardized principal components, while Autoencoder assessed the error in reproducing the data. We combined the scores from these algorithms using a weighted average rank order to obtain a final anomaly score for each property. The results of our analysis provide a way to identify potentially fraudulent properties that deviate significantly from expected values. Users can examine properties with the highest anomaly scores and investigate them further. We finally present five interesting case studies showcasing properties with unusual characteristics.

Data Description

1. Data Description

The Department of Finance collects Property Valuation and Assessment Data every year to determine the value of properties within the city for the purpose of calculating property tax bills. Each record in this dataset represents a single property within the city and includes information such as the property's location, size, assessed value, and any exemptions or abatements. The dataset includes a collection of 1,070,994 records. There are in total 32 fields, including both numeric and categorical ones. This dataset will be used to identify patterns and trends in these properties to detect tax fraud and to develop a fraud detection algorithm that can accurately identify and prevent it.

2. Summary Tables

a. Numerical Table

Field	% Populated	Min	Max	Mean	Standard Deviation	%Zero
LTFRONT	100%	0	9,999	36.64	74.03	15.79%
LTDEPTH	100%	0	9,999	88.86	76.40	15.89%
STORIES	94.75%	1	119	5	8.37	0%
FULLVAL	100%	0	6,150,000,000	874,265	11,582,431	1.21%
AVLAND	100%	0	2,668,500,000	85,067.9	4,057,260	1.21%
AVTOT	100%	0	4,668,308,947	227,238	6,877,529	1.21%
EXLAND	100%	0	2,668,500,000	36,423.89	3,981,576	45.91%
EXTOT	100%	0	4,668,308,974	91,187	6,508,403	40.39%
BLDFRONT	100%	0	7,575	23.04	35.58	21.36%
BLDDEPTH	100%	0	9,393	39.92	42.71	21.37%
AVALAND2	26.4%	3	2,371,005,000	246,236	6,178,963	0%
AVTOT2	26.4%	3	4,501,180,000	713,911	11,652,530	0%
EXLAND2	8.17%	1	2,371,005,000	351,236	10,802,21	0%
EXTOT2	12.22%	7	4,501,180,000	656,768	16,072,510	0%

b. Categorical table

Field	% Populated	# Blank	# Zeros	# Unique Values	Most Common Value
RECORD	100%	0	0	1,070,994	N/A
BBLE	100%	0	0	1,070,994	N/A
BORO	100%	0	0	5	4
BLOCK	100%	0	0	13,984	3944
LOT	100%	0	0	6,366	1
EASEMENT	0.43%	1,066,358	0	12	E

OWNER	97.04%	31,745	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	100%	0	0	200	R4
TAXCLASS	100%	0	0	11	1
EXT	33.08%	716,689	0	3	G
EXCD1	59.62%	432,506	0	129	1017.0
STADDR	99.94%	676	0	839,280	501 SURF AVENUE
ZIP	97.21%	29,890	0	196	10314.0
EXMPTCL	1.45%	1,055,415	0	14	X1
EXCD2	8.68%	978,046	0	60	1017.0
PERIOD	100%	0	0	1	2010/11
YEAR	100%	0	0	1	FINAL
VALTYPE	100%	0	0	1	AC-TR

Data Cleaning

1. Remove Exclusions
    - Remove all the owners that appears to be government. It is not necessary to look into government-owned properties.
    - Look at the most common owners and modify the remove list

2. Fill in missing zip codes
    - Store all the missing (NaN) values in the ‘missing_zips’ variables
    - Assuming the data is sorted by zip, if a zip is missing, and the before and after zips are similar, fill in the zip with that value.
    - For any remaining missing values, fill in the Zip value from the previous row

3. Fill in missing AVTOT, AVLAND, FULLVAL
    - Fill value 0 in these fields with NA because 0 doesn’t make sense
    - Find the average number of these fields grouped by TAXCLASS
    - Fill in the missing in these fields with the calculated average matching with the TAXCLASS value

4. Fill in missing STORIES
    - Store all the missing (NaN) values in a temporary dataframe

- Replace the null values in STORIES with 0
- Calculate the mean of STORIES for each unique value in TAXCLASS
- Replace 0 values with the corresponding mean value calculated

5. Fill in missing LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT
   - Fill value 0 and 1 in these fields with NA because they don't make sense
   - Find the average number of these fields grouped by TAXCLASS
   - Fill in the missing in these fields with the calculated average matching with the TAXCLASS value

In summary, imputation for this dataset has been done in 2 ways:

- For Zip, we fill in by neighboring values
- For the other fields, we fill in by averaging the values of each column by TAXCLASS and replace the missing value with the average value. Each field is done differently since missing values are under different forms for each field (NaN or 0 or 1)

Variables Generation

Variables	Count	Description
ltsize	1	Area of the lot (sqft)
bldsize	1	Area of a floor of the building
bldvol	1	Total floor area of the building
Ratio (r1, r2,..., r9)	9	$ market value, actual land value, and actual total value per sqft of the land, building footprint, and total floor area. This metric provides insights how the values of the property compare to the size and volume. Unusual values of these ratios could help detect any abnormality of evaluation and property's specifications, which indicates potential fraud.
ratio_inv (r1_inv, r2_inv, ..., r9_inv)	9	Sqft of the land, building footprint, and total floor area per dollar of market value, actual land value, and actual total value. This metric provides insights how the size and volumn of the property compare to the values. Unusual values of these ratios could help detect any abnormality of evaluation and property's specifications, which indicates potential fraud.
ratio_zip5 (r1_zip5, ..., r9_zip5)	9	$ market value, actual land value, and actual total value per sqft of the land, building footprint, and total floor area compared to its average in the similar zip code. This metric can help detect $ per SQFT abnormality of properties within the same neighborhood

ratio_inv_zip5 (r1_inv_zip5,..., r9_inv_zip5)	9	Sqft of the land, building footprint, and total floor area per dollar of market value, actual land value, and actual total value compared to its average in the similar zip code. This metric can help detect SQFT per $ abnormality of properties within the same neighborhood
ratio_taxclass (r1_taxclass,..., r9_taxclass)	9	$ market value, actual land value, and actual total value per sqft of the land, building footprint, and total floor area compared to its average of the similar taxclass. This metric can help detect $ per SQFT abnormality of properties within the same taxclass
ratio_inv_taxclass (r1_inv_taxclass,..., r9_inv_taxclass)	9	Sqft of the land, building footprint, and total floor area per dollar of market value, actual land value, and actual total value compared to its average of the similar taxclass. This metric can help detect SQFT per $ abnormality of properties within the same taxclass
value_ratio	1	The ratio of the ratio of market value over the sum of actual land value and actual total value compared to the average ratio. This metric can help detect abnormality of the ratio of market value to the actual value among all properties in the data
(new) **zs_val_taxclass**	3	Z-score of market value, actual land value, and actual total value of the property within its taxclass. This measures how many standard deviations a data is away from the mean. High z-scores indicate unusually low or high values compared to the belonging tax class, which indicates potential anomalies in land and values.
(new) **zs_val_zip5**	3	Z-score of market value, actual land value, and actual total value of the property within the same neighborhood. This measures how many standard deviations a data is away from the mean. High z-scores indicate unusually low or high values compared to the belonging zipcode, which indicates potential anomalies in land and values.
Total variables	61	(excluding ltsize, bldsize, bldvol)

Dimensionality Reduction

Principle Component Analysis (PCA) can be useful in reducing the dimensionality of the input data by identifying the most important features or patterns in the data. In this case of a fraud model with 61 generated variables, there may be redundant or highly correlated features that can be removed without sacrificing too much information. This can help to simplify the model and make it more efficient. To implement PCA, we followed these steps:

1. **Standardize the data:** The PCA method is sensitive to the scale of the variables since it is based on variance. Therefore, it is crucial to standardize the data to ensure that all variables have the same scale. In this project, we use z-scale methodology to standardize all features to have a mean of 0 and a standard deviation of 1.

2. **Run PCA and look at the cumulative variance plot:** Using Python Scikit-learn package, we generate a set of principal components, each representing a linear combination of the original variables, that can retain 99% of the variance. We created a plot of the cumulative variance explained by each principal component, showing how much of the variance is explained by the first principal component, the first two principal components, and so on.
3. **Select the desired number of components:** Based on the cumulative variance plot, we choose the number of principal components to retain. This involves a trade-off between retaining enough components to capture most of the variance in the data and reducing the dimensionality of the data. In this case, the plot shows that the model can explain over 80% of the variance in the data with 9 principal components. As a result, we decide to select 10 components to move on.
4. **Transform the data:** After determining the number of principal components to retain, we use them to transform the data into a lower-dimensional space. This results in a new set of variables (principle component) that can be used for the unsupervised fraud model.

As the result, we are able to reduce the dimensionality from 61 original features down to 10 principle components that could capture more than 80% of the variance.

Anomaly Detection Algorithms

To generate the fraud score for each record, we employ two scoring methods and combine the two scores into a final fraud score:

- **Method 1: Z-Score Outliers for score 1**
  - Measures distance to the origin for each point
  - With the transformed data of 10 principle components obtained from the PCA process, we apply z-scale again to standardize each principal component to make all the dimensions equally important.
  - To calculate the abnomaly score, we use the general MinKowski distance measure using the following formula:

$$s_i = \left(\sum_n |z_n^i|^p \right)^{1/p}$$

where: p=2, z_n^i: z-score corresponding to record the i^{th} and principal component n^{th}

- **Method 2: Autoencoder for score 2**
  - The error from an autoencoder as a measure of unusualness of a record
  - With the transformed data of 10 principle components obtained from the PCA process, we apply z-scale again to standardize each principal component to make all the dimensions equally important.

- 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 records that aren't reproduced well are unusual records.
- To calculate the abnomaly score, we use the general MinKowski distance measure for the autoencoder error using the following formula:

$$s_i = \left(\sum_n |z_n'^i - z_n^i|^p \right)^{1/p}$$

where: p=2, $z_n'^i$: predicted z-score corresponding to record the i^{th} and principal component n^{th}, and z_n^i: the actual z-score corresponding to record the i^{th} and principal component n^{th}

- **Combining score 1 and score 2**

  To combine the 2 scores into a final fraud score, we use the weighted average rank orders using this formula:

$$Final\_score_i = w_1 * rank(score\_1_i) + w_2 * rank(score\_2_i)$$

where: the higher the score, the higher the rank of that score

Results

The higher final score means that the record has either or both high z-score outlier and high autoencoder error, increasing the likelihood of abnormality. Using the final score, we select the top 10,000 properties with the highest score for further investigation. We examine each property, starting with the one with the highest score. For each property, we look at the characteristics of the property (building/land size, # of stories), its reported valuation, and 61 z-scaled variables from the features generation. We also use external source like Google Maps to verify the property's true characteristics or Zillow to see its estimated value. In general, abnormal records are the ones having very large z-scores, counterintuitive facts like the building area is a lot bigger than the land area or very cheap valuation, etc. Below is the 5 abnormal properties that we investigate:

Record: 333412
Owner: SPOONER ALSTON
Address: 37 MONROE STREET, NY 11283

Information			
BLDGCL	C5	**TAXCLASS**	2B
LTFRONT	17	**LTDEPTH**	58
BLDFRONT	4,017	**BLDDEPTH**	42
STORIES	3	**FULLVAL**	$9,060
AVLAND	$3,874	**AVTOT**	$4,077

Abnormalities				
	r2inv_taxclass	**r3inv_taxclass**	**r8inv_taxclass**	**r9inv_taxclass**
z-score	286.1757	198.6333	305.7233	255.7101

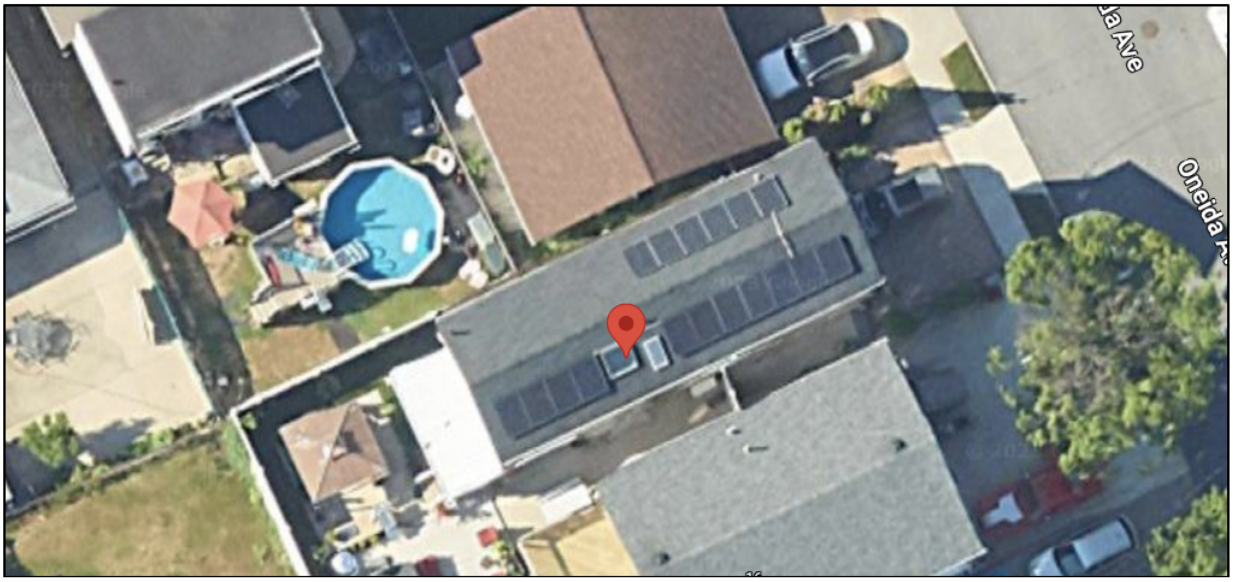
For property 333412, r2inv_taxclass (LTFRONT*LTDEPTH/FULLVAL), r3inv_taxclass (LTFRONT*LTDEPTH*STORIES/FULLVAL), r8inv_taxclass (LTFRONT*LTDEPTH /AVTOT), and r9inv_taxclass (LTFRONT*LTDEPTH*STORIES/AVTOT), are unusually high. This is due to the value of FULLVAL, AVLAND, and AVTOT are ridiculously cheap compared to the general property value in New York. According to Zillow, a property on this street is estimated to be over $1M. Also, it doesn't make sense that the building specifications are a lot higher than the lot specifications. Therefore, we have enough evidence to label this property as unusual record which has a high risk of potential property recording fraud.

Record: 956520
Owner: TROMPETA RIZALINA
Address: 12 ONEIDA AVENUE, NY 10301

Information			
BLDGCL	A1	**TAXCLASS**	1
LTFRONT	25	**LTDEPTH**	91
BLDFRONT	1812	**BLDDEPTH**	5020
STORIES	3	**FULLVAL**	$348,200
AVLAND	$15,600	**AVTOT**	$20,892

Abnormalities

	r2inv_taxclass	**r3inv_taxclass**	**r5inv_taxclass**	**r6inv_taxclass**	**r8inv_taxclass**	**r9inv_taxclass**
z-score	762.6817	775.1231	441.4327	348.3292	472.6452	422.8196

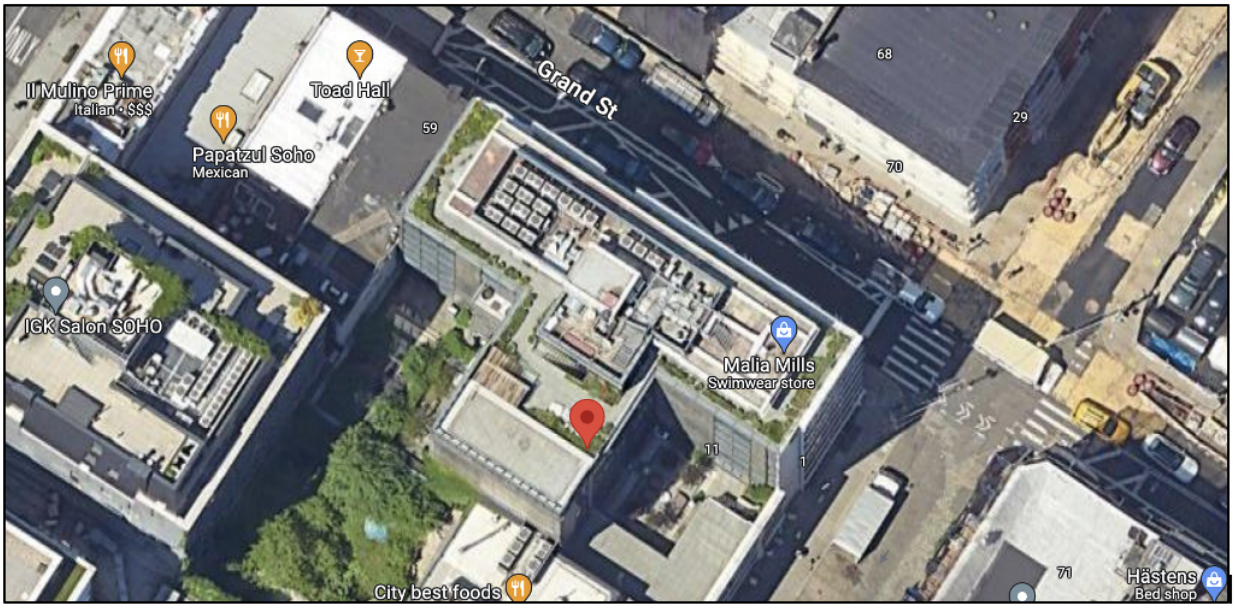


For property 956520, r2inv_taxclass (LTFRONT*LTDEPTH/FULLVAL), r3inv_taxclass (LTFRONT*LTDEPTH*STORIES/FULLVAL), r5inv_taxclass (BLDFRONT*BLDDEPTH/AVLAND), r6inv_taxclass (BLDFRONT* BLDDEPTH*STORIES/AVLAND), r8inv_taxclass (LTFRONT*LTDEPTH /AVTOT), and r9inv_taxclass (LTFRONT*LTDEPTH*STORIES/AVTOT) are unusually high. This is due to the value of AVLAND, and AVTOT are lower while BLDFRONT and BLDDEPTH are unusually big. In fact, BLDFRONT and BLDDEPTH are much higher than LTFRONT and LTDEPTH, which is not reasonable. According to Zillow, this property is worth $700K (as of now) so the FULLVAL doesn't seem too suspicious. LTFRONT and LTDEPTH look normal according to Google Maps. It's possible that BLDFRONT and BLDDEPTH were exaggerated, while it should be less than LTFRONT and LTDEPTH. Therefore, we have enough evidence to label this property as unusual record which has a high risk of potential property reporting fraud.

Record: 14979
Owner: ENJAY ASSOCIATES
Address: 69 GRAND STREET, NY 10013

Information			
BLDGCL	G6	**TAXCLASS**	4
LTFRONT	114	**LTDEPTH**	80
BLDFRONT	8	**BLDDEPTH**	6
STORIES	1	**FULLVAL**	$2,680,000
AVLAND	$1,201,500	**AVTOT**	$1,206,000

Abnormalities					
r3	**r6**	**r3_zip5**	**r6_zip5**	**r3_taxclass**	**r3_taxclass**
z-score 114.2442	144.2230	189.5386	131.1207	112.1535	115.3073



For property 14979, r3 (FULLVAL/BLDVOL) and r6 (AVLAND/BLDVOL) are unusually high not only in total but also when grouping by tax class and zip code. This is due to BLDFRONT and BLDDEPTH are unusually small and not true. Also, according to Google Maps, LTFRONT and LTDEPTH look normal. This property should have more than 1 story as well. It's possible that BLDFRONT, BLDDEPTH, and STORIES were reported wrong. Therefore, we have enough evidence to label this property as unusual record which has a high risk of potential property reporting fraud.

Record: 934793
Owner: BREEZY POINT COOPERAT
Address: 217-02 BREEZY POINT BLVD, NY 11697

Information			
BLDGCL	A8	**TAXCLASS**	1D
LTFRONT	2,798	**LTDEPTH**	997
BLDFRONT	30	**BLDDEPTH**	40
STORIES	1	**FULLVAL**	$273,000,000
AVLAND	$10,920,000	**AVTOT**	$16,380,000

Abnormalities				
	r2	**r3**	**r6**	**r9**
z-score	223.4458	467.0439	52.3950	42.1458

For property 934793, r2 (FULLVAL per floor area SQFT), r3 (FULLVAL per total floor area SQFT) are unusually high. This is due to the value of FULLVAL being too high comparing to the AVTOT. According to Zillow, properties in that area are estimated within $500K - $1M. The reported FULLVAL of $273M is completely non-sense for such a small house. Moreover, the building dimensions are reasonably correct while the lot dimensions are extremely unreasonably high according to Google Maps. Therefore, we have enough evidence to question this property.

Record: 39770
Owner: GREENHORN DEVELOPMENT
Address: 142 WEST 23 STREET, NY 10011

Information			
BLDGCL	D9	**TAXCLASS**	2
LTFRONT	75	**LTDEPTH**	98
BLDFRONT	8	**BLDDEPTH**	8
STORIES	13	**FULLVAL**	$10,200,000
AVLAND	$1,741,500	**AVTOT**	$4,590,000

Abnormalities						
	r2_taxclass	**r3_taxclass**	**r5_taxclass**	**r6_taxclass**	**r8_taxclass**	**r9_taxclass**
z-score	362.67	198.2783	317.1930	160.7314	219.5274	120.3756

For property 39770, r2_taxclass (FULLVAL per floor area SQFT by TAXCLASS), r3_taxclass (FULLVAL per total floor area SQFT by TAXCLASS), r5_taxclass (AVLAND per floor area SQFT by TAXCLASS), r6_taxclass (AVLAND per total floor area SQFT by TAXCLASS), r8_taxclass (AVTOT per floor area SQFT by TAXCLASS), and r9inv_taxclass (total floor area SQFT per $ of AVTOT) are unusually high. This is due to the very small and unrealistic reported building dimensions of 8x8, which is misleading according to Google Maps. The property is a towering condominium complex with 13 stories, which is consistent with Google Maps. Consequently, there needs to be more investigation on this valuation, particularly regarding the BLDFRONT and BLDDEPTH of the property.

Conclusion

In this project, we analyzed a dataset of property valuation and assessment data to detect potential tax fraud. We started by cleaning the data, which involved removing government-owned properties, filling in missing values for zip codes and various property attributes, and generating additional variables related to property size and value ratios. We then performed dimensionality reduction using PCA to reduce the number of features from 61 to 10 while retaining over 80% of the variance.

To identify potential anomalies, we employed two anomaly detection algorithms: Z-Score Outliers and Autoencoder. The algorithms generated scores for each property, indicating the likelihood of abnormality. We combined these scores using a weighted average and obtained a final fraud score for each property. Based on the final scores, we selected the top 10,000 properties for further investigation.

Using the final fraud scores, we examined the characteristics of the properties, their reported valuations, and the generated variables. We also utilized external sources such as Google Maps and Zillow to verify property characteristics and estimated values. We identified five interesting case studies, including properties with extremely high z-scores, counterintuitive valuation and size relationships, and other abnormal patterns.

To adjust the algorithm with expert feedback, modifications can be made to variables and exclusions. By analyzing the characteristics and patterns of flagged properties and comparing them with expert knowledge, we can determine which variables are most relevant and should be included or modified. Additionally, exclusions can be adjusted based on feedback to refine the algorithm's accuracy and focus on specific property types or owners that are more likely to exhibit fraudulent behavior. Continuous iteration and collaboration with experts will improve the algorithm's effectiveness in detecting tax fraud and help identify new patterns or indicators of abnormality.