Detecting High Risk Property Taxpayers Using a New Business Intelligence Model: A Case of New York City Property Tax

S Joshua Johnson
Department Of Computer Science & Engineering
Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra Pradesh, India
Email- joshua.sirasapalli@gmail.com

M Ramakrishna Murty
Department Of Computer Science & Engineering
Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra Pradesh, India
Email- ramakrishna.malla@gmail.com

J Hyma
Department Of Computer Science & Engineering
Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra Pradesh, India
Email- jhyma.cse@gmail.com

Abstract- Many countries generate large volumes of financial data that needs complex mechanisms to extract useful tax information. A lot of fraudulent taxpayers may exist in this generated tax data. The behavior of these fraudulent taxpayers has a greater negative impact on the resources that are available with the financial public services, thus creating inequality among the honest taxpayers. Justice, in this regard, must be provided by the government intervention to bring equality among all the taxpayers of the country. This paper addresses the above-foresaid problem. One important category of taxes is the Property tax. We choose to identify the fraudulent Property taxpayers. The study took place by analyzing the existing machine learning techniques impact on the data, such as SVM and Tukey outlier algorithms to detect the fraudulent taxpayers. We found that the results obtained from these algorithms do not agree with the actual fraudulent taxpayers. So, due to this reason, we tried constructing expert variables using Feature Engineering. Later, Analysis took place by applying Autoencoder and Mahalanobis algorithms individually on these variables, run the test for different fraud score percentages. At the end, the fraud score percentage for which there is maximum overlap is considered as fraud, thereby identifying the fraudulent taxpayers. Results generated by the proposed model improves the accuracy and takes less time in order to detect under and overpayments as outliers when compared to the existing methods.

Keywords – Business Intelligence, Fraud detection, Outliers, PropertyTax

I. INTRODUCTION

All over the world, [Mahmood Mohammadi, Shoreh Yazdani, Mohammadhamed Khanmohammadi, Keyhan Maham,2020] tax authorities are currently experiencing an increasingly high pressure to collect extra tax revenues, to discover under taxpayers, and predict the irregular behavior of the persons who do not pay the tax. Also, the tax authorities has to collect the data from different independent sources and should perform data matching and checking with other sources to find the actual non-compliant cases. As a result, without information technology tools, tax evasion detection performance would be rather limited. The main aim of this paper is to develop a business intelligent model for Property tax to detect under payers of tax, who are further considered as fraudulent taxpayers. We tend to find the potential fraudulent taxpayers from the taken property tax data which is bulk tax with one

million records. It will help the government to take actions and go through the results produced from our study. We consider the New York City property tax data. The New York Property tax data set, which we considered for our study is publicly available in [https://data.cityofnewyork.us 2017].Nowadays, along with the regular taxes, additional restrictions on the taxpayers tend to produce large volumes of data stored in the databases, which indeed need to be processed, stored and provide users or the government with the information obtained from it for further actions. According to tax politics, especially value-added tax, the rate of tax fraud is now increasing. Based on the investigations, recent researchers tend to use similar and standard methods to detect tax fraud, which includes, association rules, clustering, neural networks, decision trees, Bayesian networks, regression and genetic algorithms. Due to large volumes of income tax data, most of the methods we study about fraud detection are computationally intensive. We propose a data mining model as a Business Intelligence (BI) tool to detect fraudulent taxpayers on Property tax data for the City of New York. The paper is organized as follows; Section 1.0 as Introduction, Section 2.0 with literature Survey, and Section 3.0 is the Related Work, Section 4.0 is the Methodology. Section 5.0 gives the implementation results whilst Section 6.0 gives the conclusion.

I. II. LITERATURE SURVEY

Financial sector frauds, including tax administration sector, [Pinak Patel, Siddharth Mal, Yash Mhaske,2019] is increasingly becoming a very serious problem. As a result, this influences negatively the incomes available to public services as well as creating harm on the honest taxpayers. Therefore [Jalindar Gandal, Dr. R.G.Pawar,2020], no country can conclude that it is fraud free. Governments, irrespective of whether they are public or private, local or multinational, huge or small, they are affected by this reality of fraud, which seriously undermines the principles of harmony and fairness of citizens before the law and threatens business. Currently, more and more companies are using BI tools to analyze sales and other related transactional data to detect fraud. This section will therefore look at application of BI and data mining in different regions of the world. The fight against tax fraud and evasion by the European Union is also initiated and different measures to stop this evasion can be found in [Loredana Andreea Cristea, Alina Daniela Vodă, Bianca Ciocanea, and Mihaela Luca,2019]. Tax evasion fraud is an issue faced by all governments in the world and one way to improve its detection is the application of big data technologies[Priya Mehta, Jithin Mathews, Sandeep Kumar, K. Suryamukhi, Ch. Sobhan Babu , S. V. Kasi Visweswara Rao, Vishal Shivapujimath, and Dikshant Bisht,2019].The countries that topped the list in chasing tax dodgers the hardest are listed as follows: Spain, Argentina, Germany, Brazil, Russia, UK, Australia, Canada, Indonesia, New Zealand [Alison Steed 2015].

In India, The GST Network project (GSTN), a unique and complex IT initiative, has been set for implementing GST, major reason of this design is to check the tax evasion and thereby fix cheatings related to tax by everyone at various levels during tax administration [https://www.deccanchronicle.com 2017].Hence, this research study focused on the development of a data mining model as a Business Intelligence tool for detection of fraud on Property tax data and analysis of bulk data for City of New York Property Valuation and Assessment Data.

III. RELATED WORK

Detecting Outliers has been studied broadly by the statistics community [TalesMatos, JoseAntonioMacedo, FrancescoLettich, Jose MariaMonteiro,ChiaraRenso,RaffaelePerego,Franco MariaNardini], where the objects are modeled as distribution, and objects are marked as outliers depending on their deviation from this distribution. The database and data mining communities has addressed the problem of outlier detection .Higher education universities include the courses related to BI and Data Mining which are inevitable [Cheng M 2012]. Data mining techniques, being interesting, helps in increasing the retention rate of students and increase the learning outcome of the Students [O. D. ,. Bala M 2012]. Thus, data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationship which help in effective decision making [O. D. ,. Bala M 2012].

IV. METHODOLOGY

4.1 Proposed Approach for Tax Fraud Detection –

In this section, we try to establish a new approach as shown in Figure 1, which helps to support audit planning. We are realizing a case study that explains our approach basing on the techniques of DATA MINING on R Software. As already explained, we are studying the New York City Data Set.

1. Property tax Data set for New York City has been collected.
2. Preprocessing the data for any missing values, and later application of smoothening functions.
3. Applying Data Mining models such as SVM and Tukey on the obtained data.

4. Variable construction using feature engineering for identifying the most important variables, later applying PCA for dimensionality reduction.
5. Applying Autoencoder and Mahalanobis algorithms to get fraud score-1 and fraud score-2 respectively.
6. Displaying the optimal overlapped fraud score as final result in the form of graphs and tables for identifying the fraudulent tax payers.

4.2 Data Set Input –

The input dataset has more than one million records and thirty attributes. The dataset consists of both nominal and categorical variables. Following is description of the variables we consider to be the most important and studied are listed-
Record,Bble,Easement,Bldgcl,Taxclass,Ltfront,Ltdepth,Stories,Fullval,Avland,Avtot,Exland,Extot,Zip,Bldfront,Bldd epth.

Figure1. Architecture of the Proposed Data Mining model

Only some variables are explained in brief.Variable Name: EASEMENT. EASEMENT is a nominal categorical variable representing the property's easement type (Using some others property under lease). It has 13 levels – "", "E", "F", "G", "H", "I", "J", "K", "L", "M", "N", "P", "U". The null value indicates the property does not have any special easement type. No missing values exist.
Variable Name: BLDGCL
BLDGCL is a nominal categorical variable indicating the building class-types of buildings. It has 200 unique levels. Each level has 2 digits – the first digit is a character from A to Z, the second digit is a number from 0 to 9. No missing values exist.
Variable Name: BLDFRONT
BLDFRONT is a numeric variable representing the length of building frontage in feet. It has 610 unique values ranging from 0 to 7575. No missing values exist. However, there are 224,661 records with value 0, which could be in fact missing values.
Variable Name: ZIP
ZIP is a categorical variable, recording the zip code of the property. ZIP has 197 unique values and 26,356 missing values. There are three obvious anomalous records with ZIP of 33803, which should be in Florida.
Variable Name: BLDDEPTH
BLDDEPTH is a numeric variable representing the length of building depth in feet. It has 620 unique values ranging from 0 to 9393. No missing values exist. However, there are 224,699 records with value 0, which could be in fact missing values.
4.3 Data Cleaning–

Before constructing expert variables, we performed data cleaning to prepare the dataset for subsequent analysis.
Adjusting and combining existing variables: For the variables BBLE, we extracted its first digit and changed the variable name to "BORO", indicating the borough where the property land located. For the variable BLDGCL, since there used to be 200 unique levels in the form of "[A-Z] [0-9]", and some of the categories had very few records, we only kept the first digit -the character digit of BLDGCL variable. Therefore, there are only 26 unique levels after the transformation. We defined the product of the variables LTFRONT and LTDEPTH as a new variable LOT_AREA, indicating the lot size of each property. We multiplied the values of BLDFRONT, BLDDEPTH and STORIES, and defined the output as a new variable BLD_VOLUME, indicating the volume of each building.
Removing variables: We removed three types of variables: variables with less information, less populated variables, and already aggregated variables.
We found 7 less informative variables - STADDR, OWNER, BLOCK, LOT, PERIOD, YEAR and VALTYPE. The reasons are, some variables have too many levels to feed into our fraud detection model, some were not unique, some variables contain same values in all their rows. During the data cleaning process, we removed all of the above 7 variables with less information.
There are 7 less populated variables-EXCD1, EXMPTCL, AVLAND2, AVTOT2, EXLAND2, EXTOT2 and EXCD2. They could not serve as strong indicators of fraud considering their actual meaning. Therefore, we removed these variables from the dataset.
Aggregated variables:
Since we created the variable LOT_AREA based on LTFRONT and LTDEPTH, and we created the variable BLD_VOLUME based on BLDFRONT, BLDDEPTH and STORIES, we decided to remove LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH and STORIES from our dataset.
Filling in the missing values: For the variable EASEMENT, 99.6% of the properties in this dataset were left blank, indicating that they did not have an easement type.
Since we considered that EASEMENT is an important indicator for fraud, we filled in the missing values with a newly-created category "NO".
For the variable STORIES, there were 5% records with missing values. We filled in the missing values with the average STORIES in their own TAXCLASS.
For the variable ZIP, there were 2.5% records with missing values. We filled in the missing values with "00000".
After the data cleaning process, we kept 13 variables in the dataset viz., RECORD, FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT,
BORO, EASEMENT, BLDGCL, TAXCLASS, ZIP, LOT_AREA, BLD_VOLUME.
After the data cleaning process, we kept 13 variables in the dataset as shown in Fig 4.2.

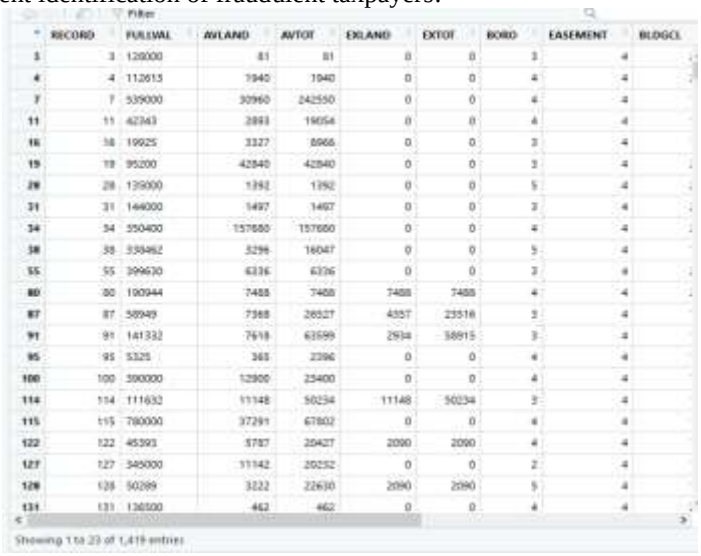
Filter

	RECORD	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	BORO	EASEMENT	BLDGCL	TAXCLASS	ZIP	LOT_AREA	BLD_VOLUME
1	1	467000	12337	19537	1620	1620	3	NO	B	1	11203	1680	1296.0
2	2	415800	13301	21312	1620	1620	5	NO	A	1	10306	2500	2142.0
3	3	126000	81	81	0	0	3	NO	V	1B	00000	384	0.0
4	4	112613	1940	1940	0	0	4	NO	V	1B	00000	1575	0.0
5	5	0	0	0	0	0	1	E	U	3	00000	0	0.0
6	6	582000	17902	29859	0	0	4	NO	A	1	11375	2000	1480.0
7	7	529800	30960	242550	0	0	4	NO	R	4	11355	0	0.0
8	8	416000	13966	22345	0	0	3	NO	B	1	11236	2400	1760.0
9	9	680000	14418	38064	0	0	4	NO	A	1	11358	3840	2058.0
10	10	702000	18091	29672	1620	1620	3	NO	C	1	11223	2480	2340.0
11	11	42343	2993	19054	0	0	4	NO	R	2	11373	0	0.0
12	12	624000	15369	27999	1620	1620	3	NO	A	1	11223	2400	1600.0
13	13	331000	8876	14473	1620	1620	4	NO	A	1	11693	2500	2016.0
14	14	440000	12631	22593	0	0	5	NO	A	1	10314	4000	1200.0
15	15	572000	11318	28552	1620	1620	4	NO	B	1	11421	2000	2200.0
16	16	19825	3327	8966	0	0	3	NO	R	2	11224	0	0.0
17	17	341100	14501	20189	1620	1620	3	NO	A	1	11203	1976	1476.0
18	18	1750000	315000	787500	0	274950	4	NO	K	4	11368	28050	10824.0
19	19	95200	42840	42840	0	0	3	NO	V	4	11207	1190	0.0
20	20	531000	15085	24055	1620	1620	4	NO	B	1	11370	2200	1692.0
21	21	289800	52200	94050	0	0	2	NO	K	4	10461	4320	3125.0

Figure 2. The final 13 Variables

4.4 Applying SVM and Tukey Outlier Algorithms–

We applied both SVM and Tukey algorithms and observed that the Tukey outlier detection leverages the interquartile range to detect outliers in dataset and also the efficiency obtained in this algorithm is more than SVM. Now, we overlapped the results of both SVM and Tukey algorithms, found that only 1419 records out of 1048575 records have been marked as fraud as shown in Figure 3.We found that the overlapped score of these two algorithms is not satisfactory, and also, clearly tells us that these algorithms cannot deeply identify the actual reason for a tax payer to be declared as a fraudulent taxpayer. As a result, we now apply feature engineering on these variables for further analysis and efficient identification of fraudulent taxpayers.



	RECORD	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	BORO	EASEMENT	BLDGCL
3	3	120000	81	81	0	0	3	4	
4	4	112613	1940	1940	0	0	4	4	
7	7	539000	30960	242550	0	0	4	4	
11	11	42343	2893	19054	0	0	4	4	
16	16	19925	3327	8966	0	0	3	4	
19	19	95200	42840	42840	0	0	3	4	
28	28	139000	1392	1392	0	0	5	4	
31	31	144000	1497	1497	0	0	3	4	
34	34	550400	157680	157680	0	0	4	4	
38	38	338462	3298	16047	0	0	5	4	
55	55	399630	6336	6336	0	0	3	4	
80	80	190944	7488	7488	7488	7488	4	4	
87	87	58949	7368	28527	4557	23516	3	4	
91	91	141332	7618	63599	2934	58915	3	4	
95	95	5325	365	2396	0	0	4	4	
100	100	390000	12900	25400	0	0	4	4	
114	114	111632	11148	50234	11148	50234	3	4	
115	115	780000	37291	67802	0	0	4	4	
122	122	45393	5787	20427	2090	2090	4	4	
127	127	345000	11142	20232	0	0	2	4	
129	129	50289	3222	22630	2090	2090	5	4	
131	131	130500	462	462	0	0	4	4	

Showing 1 to 23 of 1,419 entries

Figure 3. Overlapped Results of both SVM and Tukey Algorithms

4.5 Feature Engineering–

To begin with, we divided the original variables into two sets, 9 numerators and 6 denominators, before constructing expert variables. The 9 numerators variables are:

1.FULLVAL: full value of building,
2.AVLAND: assessed value of land,
3.AVTOT: assessed value of property,
4.EXLAND: exemption value of land,
5.EXTOT: exemption value of property
6.FULLVAL / AVTOT: the ratio of full value of building to assessed value of property
7.AVTOT / EXTOT: the ratio of assessed value to exemption value of property
8.AVLAND / EXLAND: the ratio of assessed value to exemption value of land.
9.FULLVAL / EXTOT: the ratio of full value of building to exemption value of property

All the numerators are numeric variables, which closely relate to the monetary value of properties. 1-5 were from the original dataset, while 6-9 were created by us to capture the relationship between full values, assessed values, and exemption values.

When calculating those ratios, we encountered many 0s in some of the numerical variables. Our calculation gave back infinity if we calculated their averages and put them in the denominator position. Value 0s themselves could be signs of fraud, while sometimes they could be valid and reasonable as well. For example, 0 in EXLAND meant the property did not have exemptions. Therefore, replacing them with the median value might not be reasonable. Our decision was to substitute these 0s with 1s. Since those values were large enough (usually in thousands), 1s would still be small enough for us to detect anomaly without causing calculation problems.

The 6 denominator variables are:

BORO: borough code, **EASEMENT**: easement is a non-possessory right to use and/or enter onto the real property of another without possessing it, **BLDGCL**(1st): building class, **TAXCLASS**: tax class, **LOT_AREA** =LOTFRONT * LOTDEPTH: measurement of lot area, **BLD_VOLUME** = STORIES * BLDP * BLFT: measurement of building volume, **ZIP**: zip code

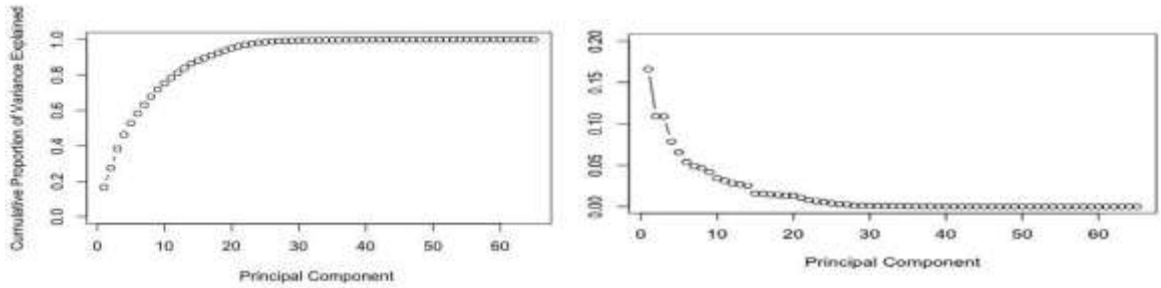
All the denominator variables (except LOT_AREA and BLD_VOLUME) are used to classify numerators. That is, we divided all those numerators by these denominator variables, calculated median of numerical variables in each

group, and divided numerical variables by the median of each group. In total, we created 61 expert variables, the sample screen is shown in Figure 4.

Figure 4. Variables obtained after feature engineering

4.6 Dimensionality reduction using PCA–

After we had all the expert variables and their respective values at hand, we started the process of standardization and dimensionality reduction for further analysis. We performed principal components analysis using the prcomp() function, which was one of several functions in R that could perform PCA. By default, the prcomp() function centers the variables to have mean zero. By using the option scale = TRUE, we scaled the variables to have standard deviation of 1. The 'center' and 'scale' components correspond to the means and standard deviations of the variables that were used for standardization prior to implementing PCA. The rotation matrix provided the principal component loadings, each column of pr.out$rotation contained the corresponding principal component loading vector. To compute the proportion of variance explained by each principal component, we simply divided the variance explained by each PC by the total variance explained by all 61 PCs. We made the screen plot and cumulative plot to determine which PCs to keep as shown in Figures 5(a) and 5(b) respectively. We would like to use the smallest number of PCs required to get a good understanding of the data. By examining the scree plot below, we discovered that there is a significant drop between PC14 and PC15. Thus, we decided to keep PC1 through PC14, which explained approximately 90% of the entire dataset. Finally, we reduced the dimension of the original dataset by multiplying the PCA matrix and the original data matrix to get the final dataset for further calculation of the fraud scores. We used two different ways to calculate fraud scores. The first one being Autoencoder, and the second one was a Mahalanobis algorithm.



(a) (b)

Figure 5. (a) Cumulative plot of variables (b) Screen plot of variables

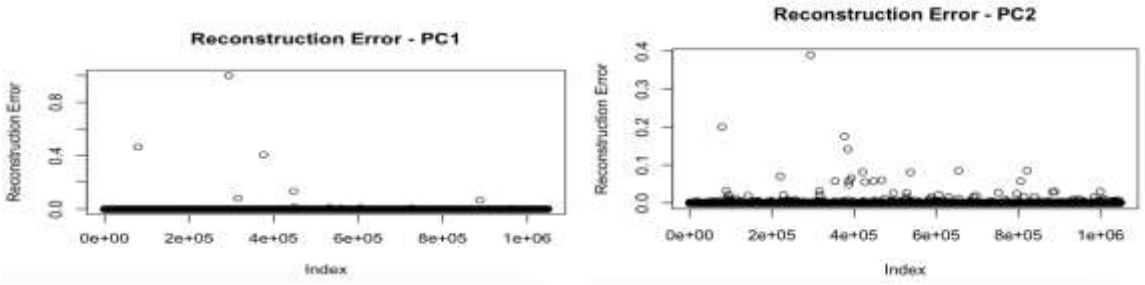
4.6 Applying Autoencoder and Mahalanobis algorithms–

Autoencoder:

First we tried to autoencode our PCs using an R package called "h2o". Then, we called the deep learning function with parameter "autoencoder" set to TRUE. This function took the original dataset with all the PCs and autoencoded it. We then called the h2o.anomaly function to reconstruct the original dataset using the reduced set of features and calculated a mean squared error between both. We set the "per_feature" parameter to TRUE because we wanted a reconstruction mean error based on individual features. We saved the reconstruction error in a dataset called "error". From plotting of the "error" dataset, we could see that there were some abnormal values, which might indicate fraud. Below are examples of the distribution of reconstruction errors for PC1 and PC2 as shown in Figures 6(a) and 6(b) respectively. In the end we summed the reconstruction error values of all the PCs to get a single score, which would be our fraud score from Autoencoder, for each of the record.

Mahalanobis Algorithm

Our algorithm calculates the Mahalanobis distance between each record and the mean, covariance of records within each particular PC. The Mahalanobis distance will be our fraud score for each record. It is calculated using the function 'Mahalanobis' in R.

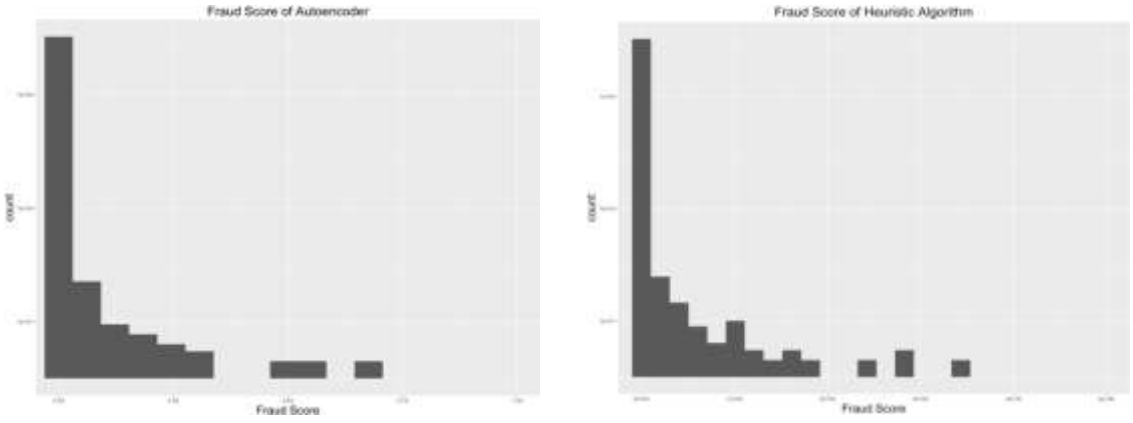


(a) (b)

Figure 6. (a) Reconstruction error for PC1 (b) Reconstruction error for PC2

V. RESULTS AND DISCUSSIONS

Having fraud scores ready, we sorted the records according to fraud scores from both Autoencoder and Mahalanobis algorithm outcomes. Not surprisingly, the majority of records had low fraud scores while a small proportion of the records had typically high fraud scores. Below is an overview of what the distribution of fraud scores look like from both methods as shown in Figures 7(a) and 7(b)



(a) (b)

Figure 6. (a) Fraud score of Autoencoder (b)Fraud score of Mahalanobis Algorithm(Heuristic)

5.1 General Trends

We have taken the fraud score percentages and compared them based on these two models as shown in Figure 7

Autoencoder	Heuristic	Overlap %
10%	10%	65%
1%	1%	72%
0.5%	0.5%	63%

Figure 7 Comparison of overlapping fraud scores

We considered the fraud score percentages of 10% ,1%,0.5% from both Mahalanobis and autoencoder, then overlapped them to see which one is best to consider for frauds, whose overlapping percentages are 65%,72%,63% respectively. We decided to look at the overlapping part of the top 1% high score records from autoencoder output and the top 1% high score records from Mahalanobis algorithm output as about 70% of the records from these two algorithms matched, further selecting these overlapped records as the best candidates for potential fraud. We found some general trends on the overlapped part of the top 1% high score records. The following table compares the mean, median and standard deviation of "Top 1%" records with the mean, median and standard deviation of complete data as shown in the Figure 8.

	Complete Data				Top 1% records			
	Number	mean	Stdev	Median	Number	mean	Stdev	Median
LTFRONT	1048575	36	74	25	7217	221	457	125
LTDEPTH	1048575	88	75	100	7217	238	379	131
STORIES	996433	5	8	2	6754	12	13	6
BLDFRONT	1048575	23	36	20	7217	114	143	92
BLDDEPTH	1048575	40	43	39	7217	121	114	99
LOT_AREA	1048575	5902	154727	2400	7217	153885	1754442	17574
BLD_VOLUME	1048575	19043	2315821	1520	7217	232034	969582	60000
FULLVAL	1048575	880488	11702927	446000	7217	37590715	134865357	12230000
AVLAND	1048575	85995	4100755	13646	7217	6614340	48918875	1413000
AVTOT	1048575	230758	6951206	25339	7217	16935641	81733286	5175000
EXLAND	1048575	36812	4024330	1620	7217	3565607	48339361	0
EXTOT	1048575	92544	6578281	1620	7217	8156874	78753026	0
FV_AT	1048575	22	429	18	7217	40	1587	2
AT_ET	1048575	95286	1942435	17	7217	7532982	22081850	2169000
FV_ET	1048575	348064	4339495	352	7217	17883478	48834099	7310000
AL_EL	1048575	39253	777335	12	7217	3022125	8811510	607500

Figure 8. Comparison of mean, median and standard deviation of original data and top 1% data

We see that the potential fraud properties have significantly higher mean, median and standard deviation values of variables (The top seven variables in above table) compared to the complete data. This means these are usually the big buildings of the city.

The bottom 9 variables have significantly higher mean, median and standard deviation values of variables in Top 1% records when compared to the complete data.

5.2 Top 10 Records

Figure 9 shows the analysis of the Top 10 Records in detail.We examined these records one by one(Analysis of only 3 records are shown:

1) RECORD No. 53359 - We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.

2) RECORD No. 55152 – Although its FULLVAL with respect to LOTAREA and BLDGCL seems fine, it has unusual values of FULLVAL with respect to BORO, BASEMENT, TAXCLASS, ZIP. The ratios are higher than 50 and surely indicate that there may be some kind of fraud.

3) RECORD No. 88293 –We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 1000 and is a must pick record for investigation purposes.

RECORD	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES
53359	1009990019	999	19	NA	DOLP 114 PROPERTIES I	O4	4	125	200	25
55152	4120990001	12099	1	NA	RISING SAM DITMARS LLC	H9	4	115	366	12
88293	1013081101	1308	1101	NA	BP 399 PARK AVENUE,	R5	4	200	405	42
89299	1012840007	1284	7	NA	KATO KAGAKU CO LTC	O4	4	124	201	45
91536	1013740014	1374	14	NA	CRP/AAC 650 MADISON O	O3	4	200	245	27
94468	1012960046	1296	46	NA	15042 ESTATE BORROWIN	O3	4	420	197	42
101017	1007819002	781	9002	NA	EU ACQUISITION LLC	O3	4	455	257	31
111320	1013070001	1307	1	NA	375 PARK AVE LP	O4	4	200	302	38
121555	1013100063	1310	63	NA	TOWER 56 REAL ESTATE	O4	4	60	100	33
129615	1012700001	1270	1	NA	SILVER AUTUMN HTL COR	H1	4	100	120	35

Figure 9. Analysis of Top 10 Records

VI. CONCLUSION AND FUTURE WORK

Detailed examination of the most suspicious records indicates that potentially fraudulent properties have significantly higher values in a lot of variables compared to the majority of records. Some properties are also significantly undervalued and hence paying lower taxes than they should. Meanwhile, most of the potentially fraudulent properties are located in Manhattan borough and belong to the tax class 4. Further examination of the top 10 most suspicious records shows that the owners of these properties are mostly real estate agencies and organizations instead of single households. The main purpose of this study is to automate and develop a fraud detection tool which uses the given information and can effectively identify the high risk property tax payers. There is a lot of scope to explore frauds in different tax categories. As the economies grow, people or companies who under pay the tax also grows. This leads to analysis of large amounts of information, and there is a need to build more efficient models and tools to identify the fraudulent tax payers thereby helping government to bring justice, order and equality among all the citizens of the country.

REFERENCES

[1] Mahmood Mohammadi, Shoreh Yazdani, Mohammadhamed Khanmohammadi, Keyhan Maham, “Financial Reporting Fraud Detection: An Analysis of Data Mining Algorithms”, International Journal of Finance and Managerial Accounting, Vol.4, No.16, Winter 2020.

[2] “https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8”. Accessed 02 May 2018.

[3] Pinak Patel, Siddharth Mal, Yash Mhaske, " A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques," International Research Journal of Engineering and Technology (IRJET),2019.

[4] Jalindar Gandal, Dr. R.G.Pawar., " A Study of Advance Fee Fraud Detection using Data Miningand Machine Learning Technique," Our Heritage, January 2020.

[5] Loredana Andreea Cristea, Alina Daniela Vodă, Bianca Ciocanea, and Mihaela Luca**.,** **“**EBEEC 2019 Economies of the Balkan and Eastern European Countries Volume 2019**”**

[6] Priya Mehta, Jithin Mathews, Sandeep Kumar, K. Suryamukhi, Ch. Sobhan Babu , S. V. Kasi Visweswara Rao, Vishal Shivapujimath, and Dikshant Bisht,” Big Data Analytics for Tax Administration”, c Springer Nature Switzerland AG 2019 A. K″o et al. (Eds.): EGOVIS 2019, LNCS 11709, pp. 47–57, 2019. https://doi.org/10.1007/978-3-030-27523-5_4

[7] https://www.telegraph.co.uk/finance/personalfinance/expat-money/11927295/Revealed-the-10-countries-toughest-on-tax-evaders.html by Alison Steed 2015.

[8] https://www.deccanchronicle.com/nation/current-affairs/300617/gst-to-use-data-mining-to-diagnose-tax-fraud.html. Accessed 22 July 2017.

[9] TalesMatos, JoseAntonioMacedo, FrancescoLettich, Jose MariaMonteiro,ChiaraRenso,RaffaelePerego,Franco MariaNardini " Leveraging feature selection to detect potential tax fraudsters”., https://doi.org/10.1016/j.eswa.2019.113128.

[10] Cheng M., "Application of business Intelligence in higher Education sector," 2012.

[11] O. D. ,. Bala M., "Study of applications of Data Mining Techniques in Education," International Journal of Research in Science and Technology, vol. 1, no. 6, pp. 135 - 146, 2012.

[12] M.RamakrishnaMurty, J.V.R.Murthy, Prasad Reddy P.V.G.D “***Text document Classification Based on a Least Square Support Vector Machines with Singular Value Decomposition****”* International journal of Computer Application (IJCA) indexed by DOAJ,Informatics,ProQuest CSA research database,NASA ADS (Hardward university)etc , ISBN 978-93-80864-56-6, DOI 10.5120/3312-4540,[impact factor 0.821, 2012] Vol. 27 –No. 7, August 2011, pp 21-26.