Fraud Analytics Project 1 Unsupervised learning on New York Property valuation

Prepared by:

Rupantar Rana

MS Business Analytics Student
Marshall School of Business
University of Southern California
rupantar@marshall.usc.edu

MARCH 2016

TABLE OF CONTENTS

Project Introduction and Motivation

The objective of this project is to build unsupervised model on the NY property evaluation data to identify properties that have been fraudulently evaluated. The data set consists of 1048575 rows and 26 features. The dataset contains the property valuations for the fiscal year 2010/2011.

Initial feature set:

BBLE
BLOCK
LOT
EASEMENT
OWNER
BLDGCL
TAXCLASS
LTFRONT
LTDEPTH
STORIES
FULLVAL
AVLAND
AVTOT
EXLAND
EXTOT
EXCD1
STADDR
ZIP
EXMPTCL
BLDFRONT
BLDDEPTH
AVLAND2
AVTOT2
EXLAND2
EXTOT2
EXCD2
PERIOD
YEAR
VALTYPE

Variable Description / Data Dictionary

BBLE
Length 11 alphanumeric
Concatenation of AV_BORO, AV_BLOCK, AV_LOT, AV_EASEMENT, descriptions of which follow.

BLOCK
Length 5 numeric
VALID BLOCK RANGES BY BORO
MANHATTAN 1 TO 2,255
BRONX 2,260 TO 5,958
BROOKLYN 1 TO 8,955
QUEENS 1 TO 16,350
STATEN ISLAND 1 TO 8,050

LOT
Length 4 numeric
UNIQUE # WITHIN BORO/BLOCK.

EASE
Length 1 alpha
IS A FIELD THAT IS USED TO DESCRIBE EASEMENT.
SPACE Indicates the lot has no Easement.
'A' Indicates the portion of the Lot that has an Air Easement
'B' Indicates Non-Air Rights.
'E' Indicates the portion of the lot that has a Land Easement
'F' THRU 'M' Are duplicates of 'E'.
'N' Indicates Non-Transit Easement
'P' Indicates Piers.
R' Indicates Railroads.
'S' Indicates Street
'U' Indicates U.S. Government

YEAR
4 Length 4 Numeric
Four-digit year of the file. For example: if the year4 = 2001
the current values are for the Fiscal year 2001/2002 assessments.
The Tentative and Final value contain the predicted values for the 2002/2003
fiscal year.

TAX-CLASS
Length 2 Character
Current Property Tax Class Code (NYS Classification)
VALID VALUES -
TAX CLASS 1 = 1-3 UNIT RESIDENCES
TAX CLASS 1A = 1-3 STORY CONDOMINIUMS
ORIGINALLY A CONDO
TAX CLASS 1B = RESIDENTIAL VACANT LAND
TAX CLASS 1C = 1-3 UNIT CONDOMINUMS
ORIGINALLY TAX CLASS 1
TAX CLASS 1D = SELECT BUNGALOW COLONIES
TAX CLASS 2 = APARTMENTS
TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS
TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS
TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS
TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR)
TAX CLASS 4A = UTILITIES - CEILING RAILROADS
TAX CLASS 4 = ALL OTHERS

OWNER
Length 21 Character
The Owner's Name.

ZIP
Length 5 numeric (no decimals)
Postal Zip code of the property

STADDR
Length 21 Character
The street address

LTFRONT
DEC Length 7 Numeric (9999.99)
Lot Frontage in feet.

LOTDEP
DEC Length 7 Numeric (9999.99)
Lot Depth in feet.

BLDFRONT
DEC Length 7 Numeric (9999.99)
Building Frontage in feet.

BLDDEPTH-DEC Length 7 Numeric (9999.99)
Lot Depth in feet.

MARKET VALUES

AVLAND
FULLVAL-LAND
Length 11 numeric (no decimals)
If not zero, Current year's total market value of the land

AVTOT
FULLVAL-TOTAL
Length 11 numeric (no decimals)
If not zero, Current year's total market value

FULLVAL
Length 11 numeric (no decimals)
If not zero, New Total Market Value of property

Data Cleaning

Step 1:

Remove the features that have no predictive power. These features remained constant throughout the data set.

Variables Removed:

Period
Year
Valtype

We removed these variables by assigning them Null values.

Step 2:

Calculate the % of data that is missing for each of the feature:

	percent_missing	type
BBLE	0.000	id
BLOCK	0.000	cat
LOT	0.000	id
EASEMENT	0.996	cat
OWNER	0.030	name
BLDGCL	0.000	cat
TAXCLASS	0.000	cat
LTFRONT	0.000	num
LTDEPTH	0.000	num
STORIES	0.050	num
FULLVAL	0.000	cat
AVLAND	0.000	num
AVTOT	0.000	num
EXLAND	0.000	num
EXTOT	0.000	num
EXCD1	0.406	num
STADDR	0.001	num
ZIP	0.025	cat
EXMPTCL	0.986	num
BLDFRONT	0.000	num
BLDDEPTH	0.000	num
AVLAND2	0.732	num
AVTOT2	0.732	num
EXLAND2	0.917	num
EXTOT2	0.876	num
EXCD2	0.913	num

We then use 70 % as the threshold for removing variables that more than 70% missing values. The following variables were removed from the dataset.

EASEMENT
EXMPTCL
EXCD2
EXLAND2
EXTOT2
AVLAND2
AVTOT2

Building Variables:

Using the existing variables in the dataset we start to build new variables that include the following.

LTAREA	=	LTFRONT * LTDEPTH
BLDAREA	=	BLDFRONT * BLDDEPTH
Full value per LtArea	=	Fullvalue / LtArea
Full value per Bld Area	=	Fullvalue / BldArea
AvtotPerAvLand	=	Avtotal / Avland

ENTITY LEVELS:

Entity levels used for each of these variables include:

BLDGCL
TAXCLASS
STORIES
ZIP

For each of the entities we calculate the mean and append it to the dataset then we divide the feature by the mean to build entity level variables.

The new dataset has 44 variables with 20 numerical variables that can be used for model building. These variables include:

Variables name : ValPerLtArea

ValPerLtAreaByBLDGCL
ValPerLtAreaByTaxcls
ValPerLtAreaByZipcls
ValPerLtAreaBystoriescls

Variable name : ValPerBldArea

ValPerBldAreaByBLDGCL
ValPerBldAreaByTaxcls
ValPerBldAreaByZipcls
ValPerBldAreaBystoriescls

Variable name : AvtotPerAvLand

AvtotPerAvLandByBLDGCL
AvtotPerAvLandByTaxclass
AvtotPerAvLandByZipcls
AvtotPerAvLandByStories

Variable name : LTAREA

LTAREAByBLDGCL
LTAREAByTaxcls
LTAREAByZip
LTAREAByStories

Variable name : BLDAREA

BLDAREAByBLDGCL
BLDAREAByTaxcls
BLDAREAByZip
BLDAREAByStories

We use these 20 new variables to build our model.

Model Building

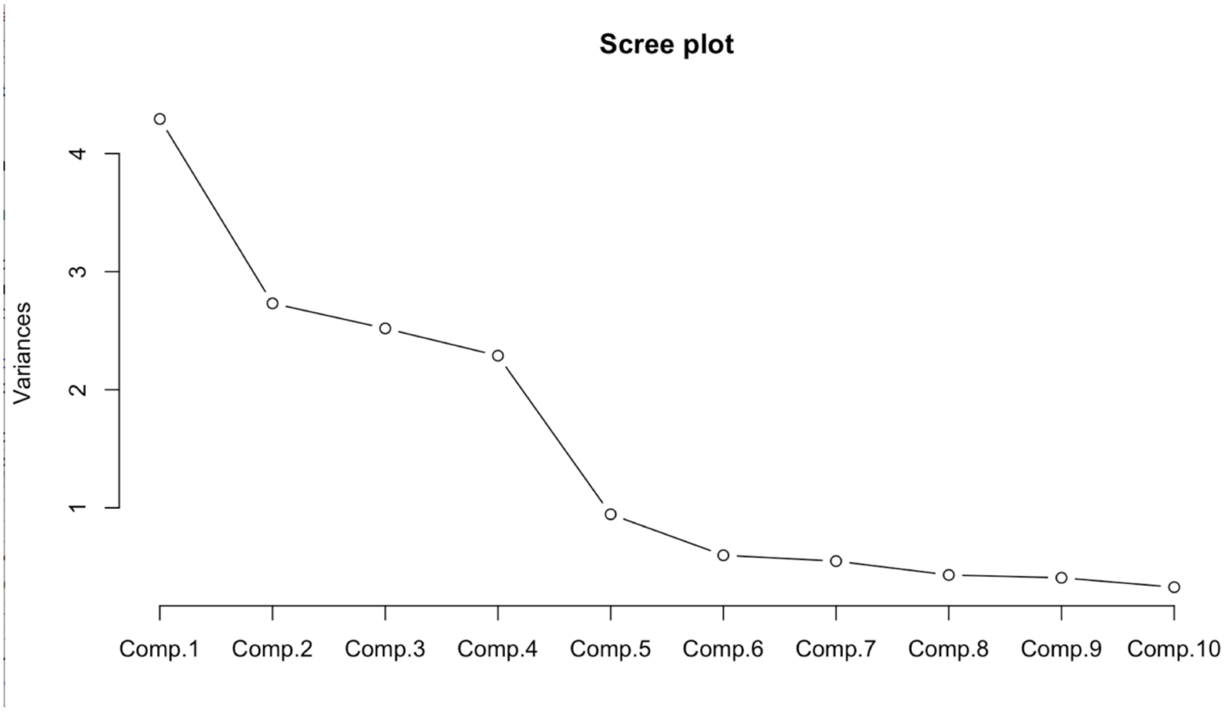
Principal Components Model

Scaling

We use the scale function in R to scale the features so that the it becomes easier for us to calculate the covariance matrix when we build the PCA model. We also removing 'NA' and 'Inf' valued rows.

Now we have 8,78,471 rows and 20 columns of scaled data to build the PCA model.

Principal components Analysis



```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     2.0721060 1.6527583 1.5872868 1.5127671 0.97220118
Proportion of Variance 0.2699012 0.1717116 0.1583768 0.1438550 0.05941459
Cumulative Proportion  0.2699012 0.4416127 0.5999895 0.7438445 0.80325912
                           Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     0.77317075 0.74068104 0.65643344 0.63820681
Proportion of Variance 0.03757782 0.03448603 0.02708708 0.02560376
Cumulative Proportion  0.84083694 0.87532298 0.90241006 0.92801381
                          Comp.10    Comp.11    Comp.12    Comp.13
Standard deviation     0.57237734 0.46504073 0.41494225 0.33563578
Proportion of Variance 0.02059424 0.01359449 0.01082321 0.00708137
Cumulative Proportion  0.94860805 0.96220253 0.97302574 0.98010711
                           Comp.14     Comp.15     Comp.16     Comp.17
Standard deviation     0.311474295 0.285471689 0.234791494 0.215432693
Proportion of Variance 0.006098531 0.005122794 0.003465337 0.002917454
Cumulative Proportion  0.986205645 0.991328439 0.994793777 0.997711231
                           Comp.18      Comp.19
Standard deviation     0.149820307 0.1181690408
Proportion of Variance 0.001410984 0.0008777851
Cumulative Proportion  0.999122215 1.0000000000
```

From the above figure we find that almost 80% of the variation in the data is explained by the first 5 components. For the most important components we choose the components that have eigen values (square of the standard deviation) greater than 1. Based on the eigen values, we choose the first 4 components for outlier detection. Now we have successfully reduced the dimension of the dataset from 20 features to 4 principal components.

Principal components dataset

	BBLE	Comp.1	Comp.2	Comp.3	Comp.4
1	3066081006	-0.28924104	-0.5088977...	3.585915e-01	-0.059690795
2	3082470011	-0.09813901	0.065758110	5.084171e-02	0.084481312
3	2027680188	-0.24126793	-0.0026131...	-3.492824e-01	0.121465269
4	5007280062	-0.08284536	0.032122495	2.347867e-04	0.021196981
5	5000210001	0.33297383	-0.0538612...	-5.281705e-01	-0.310577101
6	2054700036	-0.02246981	0.100951954	-4.501225e-02	0.131039025
7	5040700123	-0.09080476	0.012317662	8.216179e-02	-0.001925657
8	4010540025	-0.03606683	0.052295454	9.311666e-02	0.078654017
9	4065850029	-0.03067913	-0.0166321...	-1.492036e-01	-0.008326561
10	5036570005	-0.20780187	-0.0273138...	1.517692e-01	0.114399058

Anomaly score: Mahalanobis distance

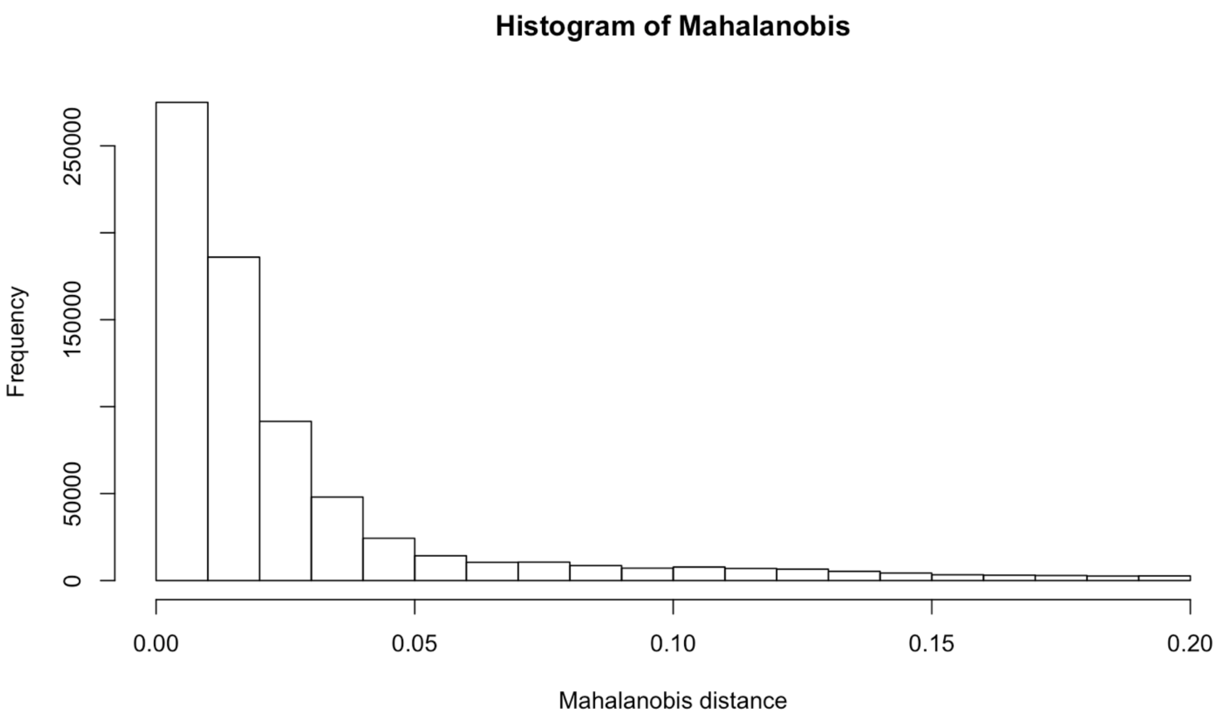
The multivariate model uses the mahalanobis distance to calculate the anomaly score for each observation in the dataset. The mahalanobis distance calculation is done based on the 4 principal component features derived earlier. After calculation of the mahalanobis distance, the BBLE was sorted based on the distance metric that was calculated.

	BBLE	Mahanolobis_Distance
1	3066081006	0.166886406
2	3082470011	0.007970825
3	2027680188	0.068428786
4	5007280062	0.002172602
5	5000210001	0.179756809
6	2054700036	0.012156010
7	5040700123	0.004656913
8	4010540025	0.007448919
9	4065850029	0.009186601
10	5036570005	0.025191299

Mahanalobis Quantiles

0%	25%	50%	75%	100%
0.000	0.007	0.016	0.040	475637.189

Please note here that 75 % of the mahalanobis distance is below 0.04. There are about 111 data points that have mahalanobis distance greater than 1000. The maximum distance is for the property owned by the Port of New York.



Mahanobilis outliers

	BBLE	Mahanolobis_Distance
1	3001990126P	475637.189
2	4029060054	279437.712
3	4022090010	264518.231
4	4018420001	208984.808
5	5006590012	177923.969
6	3070730101	164986.966
7	3080360001	162082.563
8	4092370001	158456.833
9	4155770029	153727.968
10	1015101092	56758.749
11	4089460045	49212.231
12	5050670001	40058.572
13	3009020001	39339.402
14	4004590005	38642.382
15	4004200001	34996.224
16	1015110001	31150.052
17	4090550033	31014.366
18	5020400001	30050.627
19	4089460047	25427.482
20	5059000500	22257.849
21	3013430005	21329.529
22	3084950041	20397.957
23	3034750001	19299.519
24	5000130060	18535.237
25	5014000001	18187.479

Executive Summary

Feature engineering was used to create new entity level features that were used as the input feature set for the Principal components analysis. Out of the PCAs calculated for each of the features only 4 PCAs with maximum proportion of variation explained were chosen. For Anomaly detection, Mahalanobis distance was calculated using the 4 PCAs. Data set clearly showed around 111 outliers based on the mahalanobis distance. Although the method used above is based on classical statistics, machine learning based models such as autoencoders could be used for outlier detection.