Executive summary

DSO 599 Fraud Analytics

Members: Danniel Winarto, Shengjie Qiu, Ning Zhu, Beichen Lin

Project goals:

After examining the quality of dataset, we create several new variables to complement original dataset, select appropriate entities, which contain entity levels that make sense. We also focus on some basic stats of each entity level (percentile, distribution, histogram…), then build models to figure out usual points/records or outliers in the multivariate environment. In general, we would like execute existing analysis method on data which may be related to “Fraud”.

Methods & tools:

In order to figure out “usual points”, we will be working on functions and packages in R. (mahalanobis $ K-means). In addition to the packages, we mainly rely on z-scaling method to generate a “comprehensive” score for each record. During this process, we select typical entity levels to study on how to deal with outliers.

Dealing with data:

Five new variables are:

X1= AVTOT/(LTFRONT\*LTDEPTH);

X2= AVLAND/(LTFRONT\*LTDEPTH)

X3= AVTOT/(BLDFRONT\*BLDDEPTH)

X4= AVLAND/STORIES

X5= FULLVAL/(LTFRONT\*LTDEPTH)

And the entities applied to separate records are:

BLOCK, LOT, BLDGCL, TAXCLASS, ZIP

Results:

Among all the records, we use 98% as a cutoff value, while we figure out 57,060 outliers out of around 800,000 records after removing the non-values. Moreover, we list those records having “extreme” high score (over 1000). In total we have 12 records that are not in accordance with other records. This result is more convincing than 57,060 because it fits well with exact fraud proportion in daily settings.

Results to be improved:

In this task, we mainly focus on z-scaling method, and only briefly study the algorithm of k-means clustering and mahalanobis distance method. If further elaboration is needed, we can compare the differences and distinct advatanges/disadvantages of each method.

Data Summary

The database is acquired online showing the property record in New York State for the period of 2010/11. There are 1048575 rows of records and 29 different field in the dataset. The following is the data quality report for both numerical and categorical fields. There are 17 numerical variables, and the following pictures are the description of those variables.







The following pictures are the data description of 12 categorical variables.





There are three potential outliers based on the data quality report and initial analysis.

The first unusual data in the datasheet is the zip code. There are totally 196 zip code in the data, and 195 zip code belongs to New York state. However, there is one zip code 33803 which belong to Florida and it appears 3 times in the datasheet.



The second unusual data is the stories column. Usually, by researching online, the data in stories should be the integer number or the number has decimal of .5. However, in the column, there are a lot of number has decimals, for example, 2.1 that is not valid.

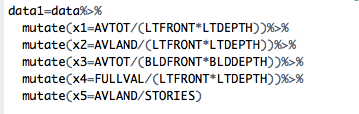


The third unusual data happens in the owner column. Most of data in this column is the address or company name, and also there is some name in this column. Other than that, it has strange data entries such as “/DECLARANT” and “0”

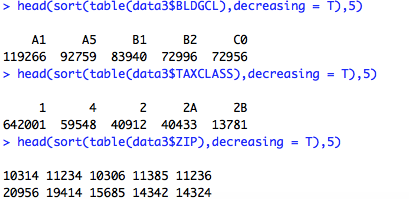


Entity levels and Variables:

We choose LTFRONT times LTDEPTH and BLDFRONT times BLDDEPTH to calculate two types of area, then we create four new variables based on two types of area. We set x1= AVTOT/(LTFRONT\*LTDEPTH) meaning the total value of property per sqft, x2=AVLAND/(LTFRONT\*LTDEPTH) meaning the value of land per sqft, x3=AVTOT/(BLDFRONT\*BLDDEPTH) meaning the total value of property per sqft, and x4= FULLVAL/(LTFRONT\*LTDEPTH) meaning the full value of property per sqft. We also set another new variable x5= AVLAND/STORIES meaning the value of land per story. Therefore, we create totally five new variables for the further analysis. The following is code we create for our five new variables.

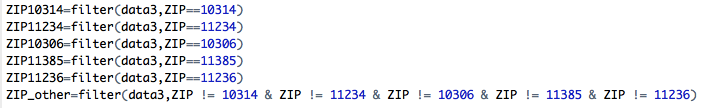


After we calculate five variables, we choose five entities levels that are block, lot, bldglc, taxclass, and zip. For the further convenience in the calculation, we decide to use top five frequencies sub-entity level in the each of entity, and put all others into another sub-entity level. The following is the example code we use to choose the top five sub-entity level.

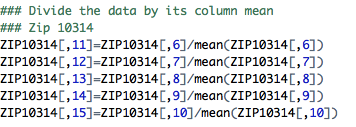


Model algorithm:

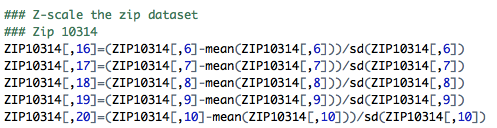
In this session, we will use zip code as an example of our model algorithm. Before we start our model, we first omit all missing values in the row, so the final dataset has 76.72% of original data records. The first step in our model is to subset six new tables based on top five sub-entity levels and remaining sub-entity levels. The following is the code we use to subset six tables.



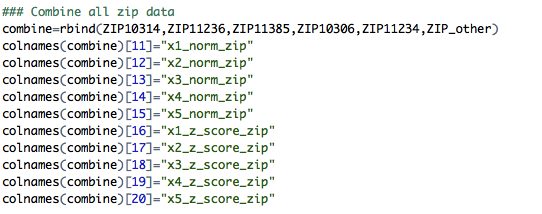
The second step is that we calculate the five new variables by normalize x1, x2, x3, x4, and x5. We divide the records in each column by its corresponding column mean.



The second step is to standardize x1, x2, x3, x4, and x5. We create another five variables by using z-score formula that (observed data in the column-mean of the column)/standard deviation of the column.



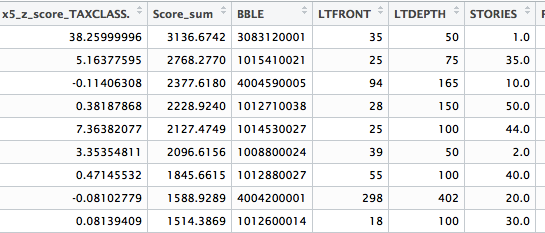
We repeat these steps for six zip tables, and then combine them together to a new dataset.



Therefore, we totally create 10 new variables for the zip entity levels. After we calculate the norm value and z-score for zip entity levels, we use combine table to repeat all previous steps for block, lot, bldgcl, and taxclass. At the end, we create totally 50 new variables in the final data table.

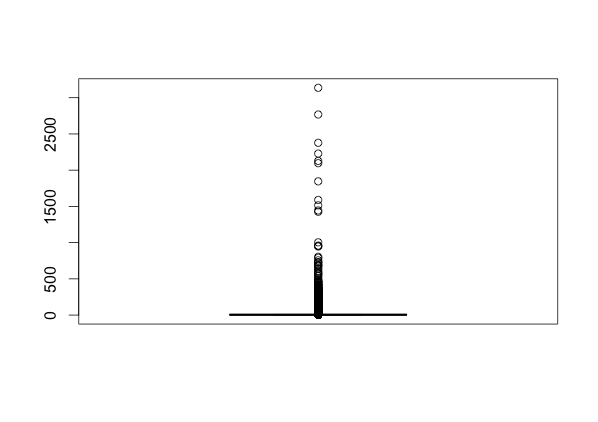
../../../../Pictures/Snip20160309_7.png

Then, we subset a new table called Score that contains all entity levels and z-scores of x1, x2, x3, x4, and x5 for five entity levels. The next step is that we calculate the score by summing up 25 z-scores in absolute value for each record and order the records by decreasing value in score\_sum. This is the screen shot of our final table.



Results interpretation:

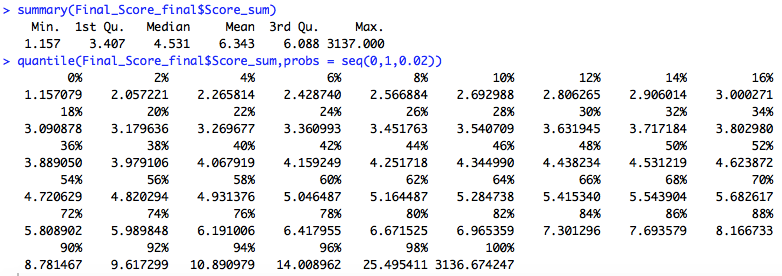
We draw a boxplot for the column score\_sum to detect the outliers.



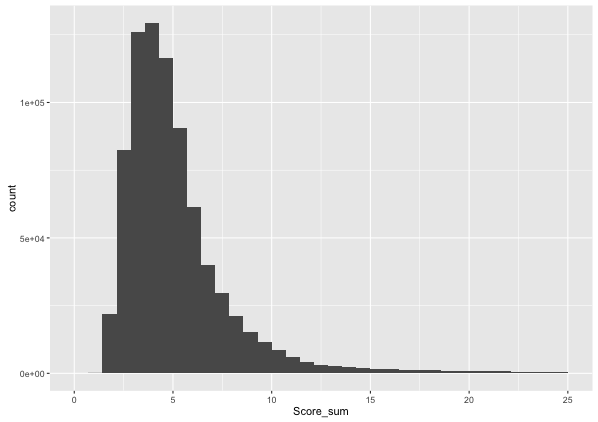
Based on the graph, it hard to tell the exact number of outliers, then we count the outliers using the code.

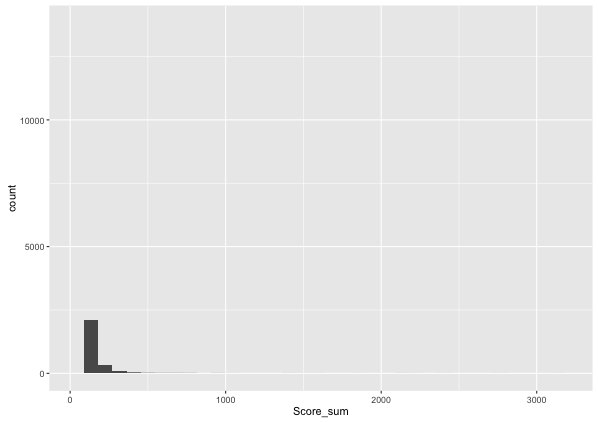
../../../../Pictures/Snip20160309_9.png

There are 57060 out of 804471 records are outliers based on our model. Then, we check the data summary of score\_sum column and create 50 bins for that column.



98% of records in the score\_sum column is between 1.157 and 25.495, so we create two separate histograms. The first histogram is the distribution of score\_sum between 1.157 and 25, and the second histogram is the distribution beyond 25.





These are score\_sum records that are greater than 1000. We think these 12 records are the extremely suspect records in the whole dataset.

