ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 116

International Conference on Education Science and Economic Development (ICESED 2019)

Variables Creation in Fraud Detection-Based on New York Property Data

Yutong Yao
No.17 Tsinghua East Road, HaiDian District, Beijing, China
angela@cas-harbour.org

Key Words: Fraud detection, Machine learning, Fraud analysis, Variables.

Abstract. With the rapid development of technology, fraud has become an extremely serious problem in the US society, therefore makes fraud detection an urgent need to improve the situation as much as possible. Fraud detection refers to the process of finding anomalies in a huge bunch of data by building fraud algorithms and models that could predict the possible behaviour of the real world situation. This paper will mainly focus on the 'Variables Creation' process in fraud detection with the New York property data and discuss how to analyze a fraud problem and build variables based on New York property data. The whole process of fraud detection will render practical help to solve the property fraud problem.

1. Introduction

In the US, people can easily commit fraud through various ways. Some steal others' identities to open up multiple accounts, while others intentionally make up some numbers in property data to avoid high tax. In order to catch these frauds, two different kinds of model should be built: unsupervised model and supervised model[1]. These two types of models target two kinds of problem: forensic accounting and real time fraud algorithms. And this paper will only revolve around the Forensic accounting problem. In forensic accounting, the algorithm and variables can be created from all the data, regardless of time flow. The purpose in this case is to find the unusual records in the fixed data in order to catch possible frauds and leave them to the investigator to be checked later[4]. Since the standard to evaluate the goodness of a model is through a combined equation of error and complexity of the model (the lower the equation result is, the better the model creates), in the case, it refers to the fraud detection rate (FDR) which equals the number of actual frauds over the numbers of examined records. To solve this problem, unsupervised modeling is often chosen as the best method due to the absence of labels. Without the previous example of fraud, creating new standards in order to determine which types of records should be regarded as anomalies is of great necessity. Therefore, this paper will cover the whole process of building unsupervised model of 'New York property data', with more detailed analysis to the variable creation part. The conclusion will lead to certain range of data with high fraud scores, which are regarded as our best guess for fraud.

2. Analysis

2.1 Data processing

2.1.1 Data description

The data used in this case is a whole collection of New York property provided by an unknown government organization in 2010. The raw data contains 1,070,994 records and 32 fields, which can be classified into numerical field and categorical field.

Here are some of explanation of the main field names:

LTFRONT: lot front
LTDEPTH: lot depth
FULLVAL: full value of the building
AVLAND: average value of the land
AVTOT: average value of the total area
BLDFRONT: building front

BLDDEPTH: building depth

Table 1. Numerical field

FIELD NAME	#RECORDS VAULE	%POPULATED	#Unique Values	#Value with zero	MEAN	Standard Deviation	Min	Max
LTFRONT	1,070,994	100	1,297	169,108	36.64	74.03	0	9999
LTDEPTH	1,070,994	100	1,370	170,128	88.86	76.40	0	9999
STORIES	1,014,730	94.75	112	NA	5.01	8,37	1	119
FULLVAL	1,070,994	100	109,324	13,007	874,264.51	11,582,431.00	0	6,150,000,000
AVLAND	1,070,994	100	70,921	13,009	85,067.92	4,057,260.00	0	2,668,500,000
AVTOT	1,070,994	100	112,914	13,007	227,238.17	6,877,529.00	0	4,668,000,000
EXLAND	1,070,994	100	33,419	491,699	36,423.89	3,981,576.00	0	2,668,500,000
EXTOT	1,070,994	100	64,255	432,572	91,186.98	6,508,403.00	0	4,668,300,000
BLDFRONT	1,070,994	100	612	228,815	23.04	35.60	0	7,575
BLDDEPTH	1,070,994	100	621	228,853	39.92	42.71	0	9,393
AVLAND2	282,726	26.4	58,592	NA	246,235.72	6,178,963.00	3	2,371,000,000
AVTOT2	282,732	27.33	111,361	NA	713,911.44	11,700,000.00	3	4,500,000,000
EXLAND2	87,449	8.17	22,196	NA	351,235.68	10,800,000.00	1	2,371,000,000
EXTOT2	130,828	12.22	48,349	NA	656,768.28	16,100,000.00	7	4,500,000,000

Table 2. Categorical field

Field Name	#Records with value	#populated	#Unique Value	Most common field value
RECORD	1,070,994	100	1,070,994	NA
BBLE	1,070,994	100	1,066,541	NA
BLOCK	1,070,994	100	13,984	3944
B	1,070,994	100	5	4
LOT	1,070,994	100	6,366	1
EASEMENT	4,636	0.43	13	E
OWNER	1,039,249	97.04	863,348	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100	200	R4
TAXCLASS	1,070,994	100	11	1
EXT	354,305	33.08	4	G
EXCD1	638,488	59.61	130	1017
EXCD2	92,948	8.68	61	1017
STADDR	1,070,318	99.93	839,281	501 SURF AVENUE
ZIP	1,041,104	97.21	197	10314
EXMPTCL	15,579	1.45	15	X1
PERIOD	1,070,994	100	1	FINAL
VALTYPE	1,070,994	100	1	AC-TR
YEAR	1,070,994	100	1	40483

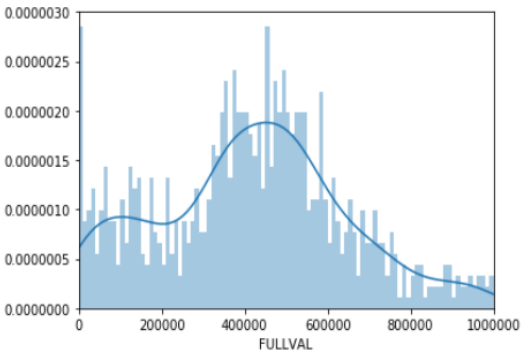

Fig. 1. Distribution of full value

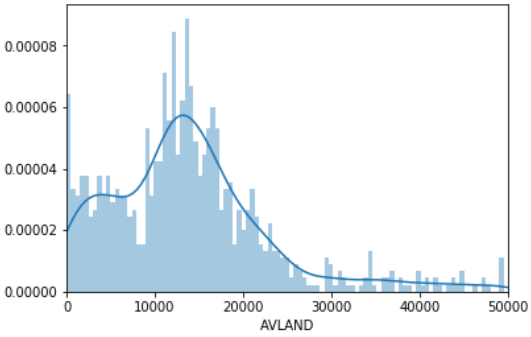

Fig. 2. Distribution of average value

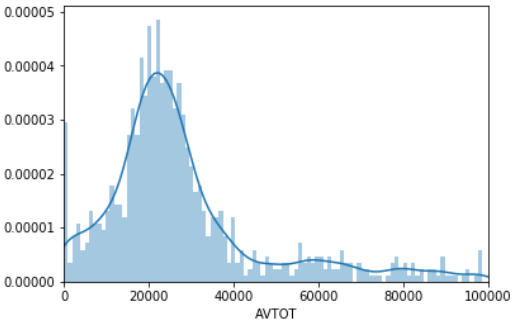

Fig. 3. Distribution of average value of total building

It is clear to see from these distributions that most of the data settles in the normal range and only a small part of it settles far from the highly distributed value, which are exactly the outliers that the algorithms need to recognize.

2.1.2 Data cleaning

Right now the data is the raw data collected directly from the official website. The next important process is to do the data cleaning to make the raw data into better form in order to prepare for creating variables from it.

Normally, data cleaning process is mainly used to fill the missing value, which can be achieved from two different ways:

① Use the average or most common value of that field over all records to fill the missing fields.

② Select one or more other fields that are important in deciding the missing field and group them into categories, and replace the missing field with the average or most common value for its appropriate group.

In this case, both methods are applied to minimize the possibility of measurement bias.

2.2 Variable creation

2.2.1 Variable chosen

Among all the processes of fraud detection, one of the most essential parts definitely is the variable creation part. This is because the final result, also called the fraud score, mostly depending on the goodness of the variables. If the variables are chosen well enough to let the model distinguish the normal data from the strange data, then the algorithms is very likely to be a great success. In contrast, if the variables are not well-chosen, then most likely, no matter how perfect the model is, the result cannot really satisfy the initial expectation.

Before variable chosen, what should first be considered is that the purpose of this algorithm, which is the final goal to be achieved. In this case, the goal is to find strange values, so we need to do comparison and sort the unusual ones. Since the data is about property, ‘unit value’ can be used to accomplish the comparison. According to the fields of the cleaned data, the following formula goes like this: **Unit value = Value/(Area or Volume)**

Here are the fields that involve in Value:

V_1 = FULLVAL
V_2 = AVLAND
V_3 = AVTOT

Here are the fields that involve in Area or Volume:

S_1 = LTFRONT * LTDEPTH
S_2 = BLDFRONT * BLDDEPTH
S_3 = S_2 * STORIES

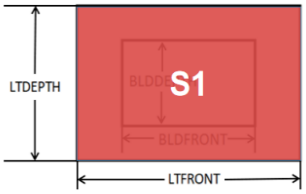

Fig. 4. S1 Land area

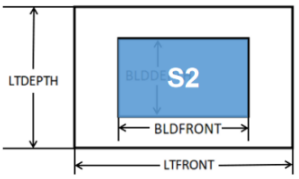

Fig. 5. S2 Building area

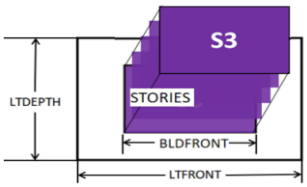

Fig. 6. S3 Building volume

Then combine them and create unit values as variables:

For each record we create 9 ratios:

$$r_1 = \frac{V_1}{S_1} \quad r_4 = \frac{V_2}{S_1} \quad r_7 = \frac{V_3}{S_1}$$
$$r_2 = \frac{V_1}{S_2} \quad r_5 = \frac{V_2}{S_2} \quad r_8 = \frac{V_3}{S_2}$$
$$r_3 = \frac{V_1}{S_3} \quad r_6 = \frac{V_2}{S_3} \quad r_9 = \frac{V_3}{S_3}$$

However, only nine ratios are not enough for variables. Variables should first be created as much as better in order to take all factors into consideration. There is no need to worry about the complexity of the input of the model because later many low-correlated dimensions will be reduced in the later steps. Therefore, here the influence of geographical and logical factors should be taken into consideration, which refers to zip code and tax class. For example, if a building settles in Manhattan, which has the most expensive land price in New York, then it will be very strange if this building has a low land value compared to the average land value of this area with the same zip code. In addition, the reason why to include tax class is that it is generally acknowledged that properties, which pay the same amount of tax should have about the same value. Thus, in this case, we will separately group records into 5 groups: zip5 (the first five numbers of zip codes), zip3 (the first three numbers of zip codes), TAXCLASS, borough, all.

With the 9 ratios and 5 groups, 45 variables can eventually be created through mathematical calculations:

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \dots \quad \frac{r_9}{< r_9 >_g} \qquad g = 1, \dots, 5$$

2.2.2 Z scale

According to Mahalanobis distance theory, right now the data shown as variables is like the figure 7, unevenly distributed[3]. However, data need to be transferred into the same scale because it is easier to see the distance of from the center to the point. The result expected is like in figure 8, in which all the data are distributed in a circle that has the same scale.

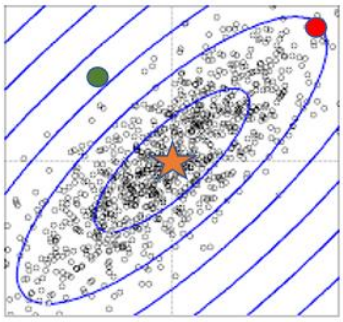

Fig. 7. Original data

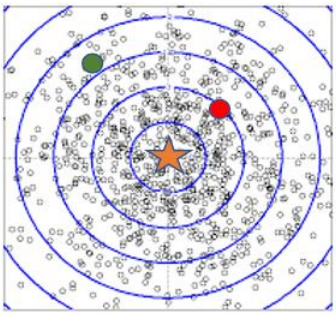

Fig. 8. Data after z scaling

To achieve this goal, the most commonly used method is 'z scale', which refers to the following formula:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

After the transfer of all the records through this formula, it is automatically to find that the mean value of the data is 0, and the standard deviation becomes 1. This makes the data easier to look at and to do further analysis.

2.2.3 Principal component analysis (PCA)

Now that the data is perfectly placed in the same scale, but with too many fields, or dimensions, the model can not be perfectly trained. So now reducing dimensions is needed to be done, which often refers to PCA in unsupervised modeling. PCA, principal component analysis, is a kind of mathematical method mainly to reduce dimensions and remove linear correlations[6]. First, assume that every record represents a point in a high dimensional space, so the data will be distributed like this (three dimensions):

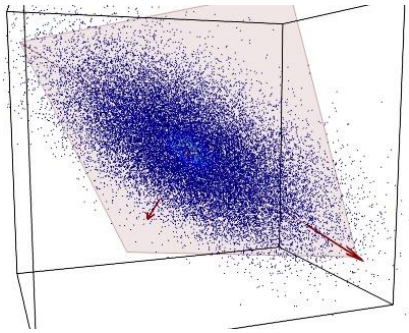

Fig. 9. Data distribution in three-dimensional space

The data spreads differently in each direction, which means that it has different variance on each dimension. The larger the variance is, the more influential the dimension is to the model. The next step is to rank the data with its variance on each dimension, and the result would be like the following:

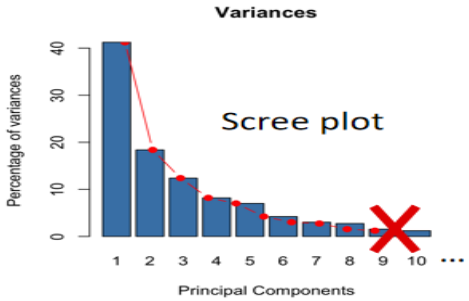

Fig. 10. Data variance in different dimensions

Afterwards, a decision should be made on how many dimensions to keep and how many to throw away. In this case, we choose to keep eight dimensions that have the high variance.

2.2.4 Z scale again

After some dimensions are thrown away, the data now is no longer in the same scale. To treat all remaining PCs the same, it is better to do the z scale again to benefit future steps. The method is just the same as the first. Finally, the data is fully prepared for training models.

2.3 Modeling

2.3.1 Algorithms building

With the prepared data, we can now start to build algorithms. To clarify before actual modeling, one thing should be sure is that the process to build a model is actually the process of finding the best surface in a high dimensional space to fit as much data as possible. When we operate the model with the given testing data, the anomalies can be easily distinguished from the normal ones. And later when the model is put into practical use, the same process would work again to find strange values which are predicted to be fraud scores.

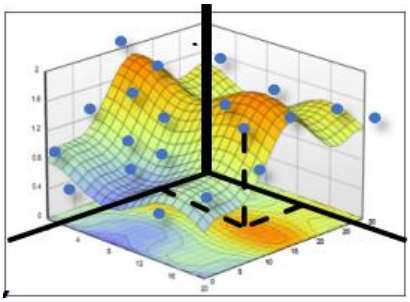

Fig. 11. Modeling simulation

As for unsupervised model, there are mainly two methods to score the data, Heuristic Function of the z scores and autoencoder. Heuristic function of the z scores is a basic method that is using some fixed formula to calculate the distance from the center to each point (record), and employ this distance to the fraud score. The general formula is as follows:

$$s_i = \left(\sum_k |z_k^i|^n \right)^{1/n}, \quad n \text{ anything} \tag{2}$$

When n=1, it is the Manhattan distance: D =|z1|+|z2|+...+|zn|

When n=2, it is the Euclidean distance: D= $\sqrt{z_1^2 + z_2^2 + ... + z_n^2}$

In this case, we choose to use the Manhattan distance and record it as score 1. But in most situations, this method cannot achieve good results, so we further employ a more advanced method, Autoencoder.

Autoencoder is a special type of Neural Net, one of the machine learning model. Autoencoder is meant for reproducing the data. When the data is normal, it can be reproduced well, but when it is

abnormal, the loss can be huge[2]. As a result, the error (distance) is interpreted as the fraud score.

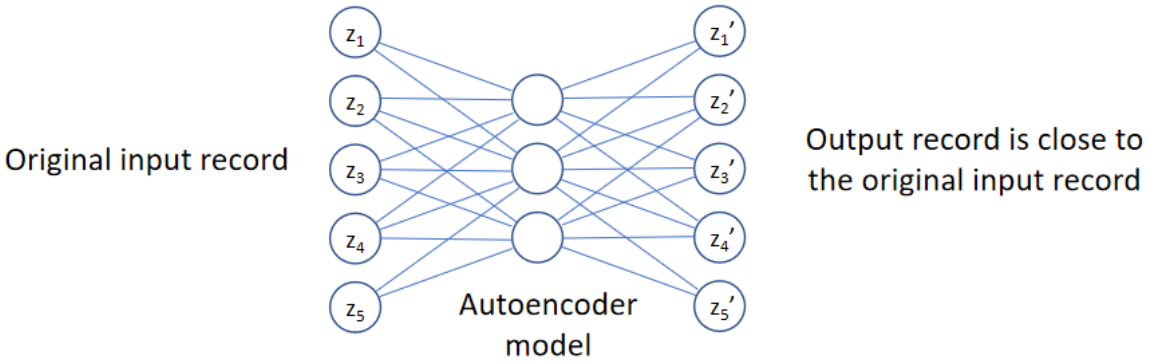

Fig. 12. Autoencoder model

The autoencoder model contains many parameters that can be changed artificially, such as the number of layers, the number of nodes on each layer, and the loop. By adjusting these parameters, the model is trained to better fit our data[5].

Table 3. Training epoch

Epoch	Loss
1	0.7860
2	0.7684
3	0.7669
4	0.7667s
5	0.7666

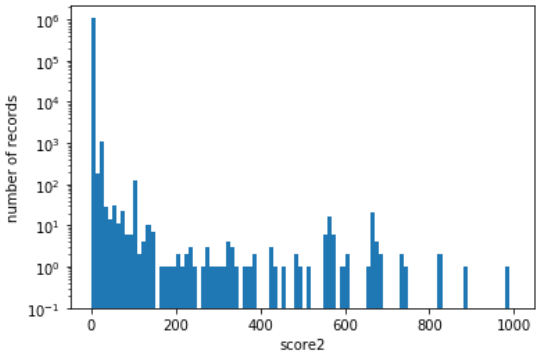

Fig. 13. Score 2 distribution

Figure 13 demonstrates the distribution of the fraud score 2 resulting from the autoencoder.

2.3.2 Score combination

Right now, the scores in hand are score1, which are acquired from linear model, and score 2, which is obtained from none linear model. To achieve a better model, the next step is to carefully consider the reliability of the two kinds of scores and attach different weight to them as different influence that one will have on the final score.

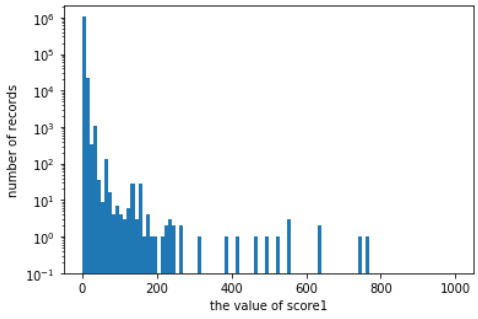

Fig. 14. Score 1 distribution

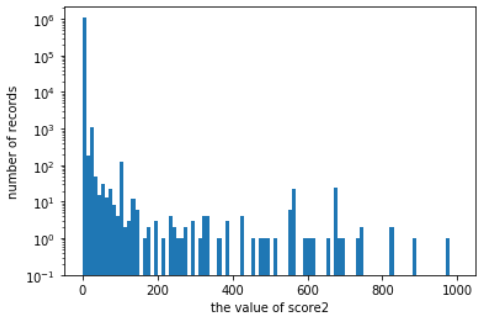

Fig. 15. Score 2 distribution

Lastly, the author combines the two scores using weighted average rank orders to get the final score, from which strange values can be clearly distinguished. And the final score distributed like this:

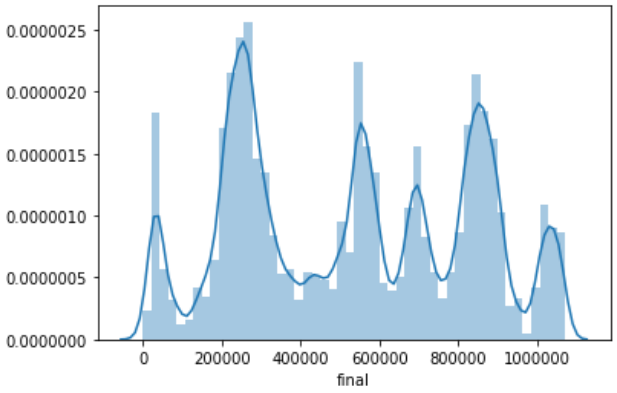

Fig. 16. Final score distribution

2.4 Result analysis

By doing the modeling step by step, the result leads to some statistically strange values which usually are interpreted as the best guess to fraud. However, these values still left to be well-defined as fraud by the investigators from companies or banks who expert in translating the data result into real life fraud detection in real world situation. They will use various methods to figure out why the data looks strange and if it is really fraud, and give data scientist feedback on the correctness of the fraud detection. According to their feedback, further adjustment to the fraud model will be made, and the process is just the same from the beginning.

3. Discussion

This paper discussed about the detailed process of building the unsupervised model for 'NY property data'. However, due to the limitation of page and knowledge, more detailed background knowledge and practical Python codes can't be shown in the paper. For future research, more in-depth analysis of variable creation and wider application of the unsupervised model can be focused on. It is greatly expected that more researches into fraud detection so that the credit system can be strongly improved.

4. Conclusion

This paper discusses the whole process of building unsupervised model for 'New York property data', and pays more attention to the variable creation part. Through this process, it is obvious that creating variables is really essential to the final goodness of the model. When doing this step, careful consideration about the involving factors is in real need, and creating as much variables as possible also exert significance on the accuracy of the further model training. As for choosing models, it is usually better to combine the linear and nonlinear models so that a better final score can be presented without bias.

With all the processes discussed above, the final model will work just fine for detecting the fraud in 'NY property data'. However, more different kinds of data would certainly have different requirements for the result. Hence, more kinds of new models or combination method are fully expected to better solve real world problems.

Acknowledgement

First and foremost, I would like to show my deepest gratitude to my teachers and professors in my university, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank all my friends and roommates for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

References

[1] B. Baesens, Fraud analytics using descriptive, predictive, and social network techniques: A guide to data science for fraud detection, Hoboken, New Jersey: Wiley, 2015.
[2] Y. Cheng, D. Zhao, Y. Wang, and G. Pei, Multi-label learning with kernel extreme learning machine autoencoder, Knowledge-Based Systems, vol. 178, pp.1-10, 2019.
[3] R. Maesschalck, D. De, Jouan-Rimbaud, and D. L. Massart. The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems, vol.50, chapter 1, pp.1-18, 2000.
[4] M. J. Nigrini, Benford's law: Applications for forensic accounting, auditing, and fraud detection, Hoboken, New Jersey: Wiley, 2012.
[5] K.Sun, J. Zhang, C. Zhang, and J. Hu, Generalized extreme learning machine autoencoder and a new deep neural network, Neurocomputing, vol. 230, pp.374-381, 2017.
[6] Z.Xu, J. Liu, X. Luo, Z. Yang, Y. Zhang, T. Zhang, Y. Tang, Software defect prediction based on kernel PCA and weighted extreme learning machine, Information and Software Technology, vol. 106, pp.182-200, 2019.