Analysis on financial fraud cases by Python

Yuan Tao
No.1 Zhenghe Avenue, Huishan new town, Wuxi City, Jiangsu Province, China
angela@cas-harbour.org

Key words: Fraud cases; Python; Financial.

Abstract. Financial frauds, which is a common phenomenon in our daily life, may take various forms. Based on real estate fraud cases in New York in 2010, this paper illustrated the process of anomaly detection by computer. Also, this paper introduced techniques to detect financial frauds, including Benford's law and fuzzy matching and their applications in fraud detection. Through the analysis of New York property in 2010, the author also proposed suggestions by using Python.

1.Introduction of Python-based analysis of financial frauds

The New York property fraud in 2010, a real-world case of fraud, is selected in this study to make the analysis more convincing. This is a good example of fraud detection because real estate is a field closely connected with people's daily life. Python is adopted for analysis because it is more accessible to people without basic knowledge of programming. The software provides users with explanations of each procedure to facilitate understanding. However, Python is an interpreted language, which means it runs slower than most compiled languages.

2.Data analysis of New York property in 2010

2.1 Procedures

People who are aware of being cheated by faking tax and numbers would turn to institutions dealing with financial fraud. There is a job called tax preparer, who takes care of people's tax refund after receiving annual bills. Two kinds of frauds could take place in the process. On one hand, there is the possibility of faking bills to create inaccurate family situations. They need to analyze these data which are different from the normal percentage of each family. Furthermore, there is also a kind of fraud which takes place at the hospital, where the doctor can exaggerate the patient's situation by claiming that he has cured five fingers of the patient whereas in fact, the patient just broke one finger. In this situation, the institution needs to check with the doctor for confirmation.

Fraud could take place at company or even government level, taking the 2010 New Your property as one example[1]. As shown in the excel, the data is categorized into two types, including the numeric number.

Table 1 Numeric number

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

The other is category number.

Table 2 Category number

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

Now, the author has 7 key factors, ‘AVLAND’, ‘AVTOT’, ‘FULLVAL’, ‘BLDFRONT’, ‘BLDEPTH’, ‘LIFRONT’, ‘LIDEPTH’. Then the Python notebook is used to analyze the distribution of each data.

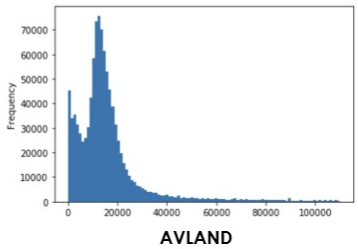

Figure 1

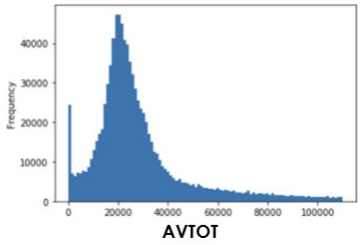

Figure 2

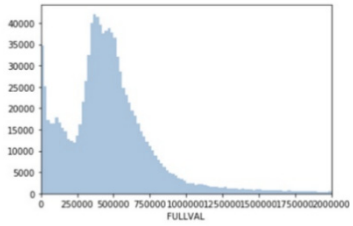

Figure 3

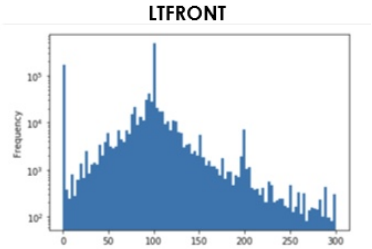

Figure 4

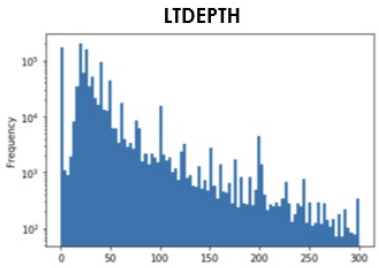

Figure 5

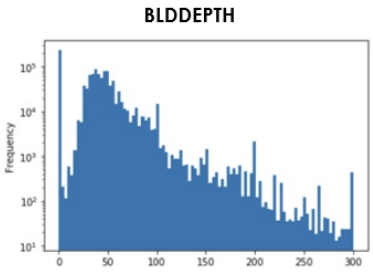

Figure 6

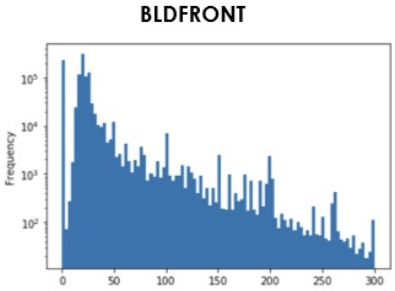

Figure 7

The data would be clearly seen that every piece of data has a deviant number beyond the average level, on which the author focuses for analysis. Each record has some missing values. In order to analyze all the records, the author needs to fill in the missing values by making up ones close to the average level. There are two ways to fill the gap. First, the average or typical values could be referred to. Besides, it also works to select one or more other important fields closely related to the missing field and then group them into different categories. The author adopts the first method to prepare for data. When a strange value is detected, it is compared to the standard level to see if there is a need for modification. And the unite value equals to the value/area or volume. The author uses V1 to present the FULLVAL, V2 to present the AVLAND, V3 to present the AVTOT, and S1 to present LITFRONT*LTDEPTH, S2 to present BLDFRONT*BLDDEPTH(equals to S2) and S3 to present S2*STORIES(equals to S3). The author groups them into 9 ratios, namely r1=V1/S1, r2=V1/S2, r3=V1/S3, r4=V2/S1, r5=V2/S2, r6=V2/S3, r7=V3/S1, r8=V3/S2, r9=V3/S3. Also, 5 groups has already been existed, including zip5, zip3, TAXCLASS, and borough. Then, these five groups, after being categorized and are used for calculation. In the end, we have 45 variables.

The next step is to prepare for data. The z scale could be prepared through Python with a mean of 0. The standard deviation equals to 1. Any number far from 1 on the z scale is considered as anomaly. Principal component analysis should be performed so as to remove linear correlations and reduce the dimensions by rotating the axis and screening out insignificant direction, namely the small variance. The author does the z scale again to ensure all the PCs are in the same scale. After data preparation, the author has a record of points forming spherical clouds in eight dimensions. The purpose is to find a criterion to determine whether a particular dot in the space is normal or not through algorithms that can quantify the anomaly. It can be seen that they are all distributed in search of something unusual, which is quite different from the rest. There are many methods to detect singular points by linear method. The author chooses Hamilton distance, which is, n=1, as the score 1. Normal points always stay close to the origin and only the abnormal one is far away. The second algorithm uses the autoencoder. It is a neural network for reproducing the data in normal conditions. In case of abnormality, huge losses can take place. The author uses the error as score 2. The transfer functions are nonlinear, which means that they can provide nonlinear terms for the entire neural network process, indicating nonlinear output. The structure of the network consists of four layers, with 12, 7, 10 and 8 nodes and they are functioned as tanh, sigmoid, relu and sigmoid respectively. The good data tends to have a higher score of 1 while the bad one has a lower score of 1.

2.2 Reasons

As it is a government-sponsored project, most of the companies try to get profits from it, indicating the possibility of financial fraud. Sometimes government officials are negligent and do not make adequate preparations or supervise the operation of the program.

2.3 Results

Companies would exaggerate the figures to make more profits. The governments could lose a lot of money if they fail to detect the fraud. Therefore, the government need expertise in this regard. At present, there are obvious defects in laws governing financial fraud, such as disguising its harm to the society and failing to deal with risks that come along. Also, there is a need for fully-developed criminal law provisions. Therefore, it is extremely urgent to revise the criminal standards of financial fraud in the era of big data and to construct a diversified criminal standard for this purpose. Specifically, when revising the criminal standard of financial fraud, we should also pay attention to its influence scope and the existence of other factors that would aggravate the circumstances in addition to the amount of money involved in the crime. At the same time, the revised standard should serve as a starting point to deal with newly-emerging forms of crimes in the era of big data. To do this, it is necessary to make further revisions against financial fraud to protect the interests of the public.

3. Possible methods to prevent financial fraud

3.1 Benford's Law

There is Benford's law for people to detect the fraud. This law is associated with the calculus and is based on people's daily life. Its scope of application is unusually wide, and almost no statistical data of artificial rules in everyday life can satisfy this law. For example, the population, land area, books, physical and chemical constants, answers behind textbooks of mathematics and physics, half-life of radioactivity and other data of the world are all conformed to Benford's law. This law indicates a nonintuitive fact that the first digit of many measurements is not evenly distributed. The first digit '1' appears about 30% (the picture of the distribution to see figure 2).

Figure 8 Digits

The discoveries can be used to detect the human-made numbers. Because it is a law just few people know and only the expertise can use this law. If the cheaters want to make a fraud, they need to make up some exaggerated number to gain more money. But the intentionally made number would violate the law. In addition, this law has a long history. In 1881, an astronomer, Simon Newberber, found that the pages which started with '1' was worse than the rest. But at that time, no one was convinced necause they all thought it was just an accidental event. Actually the statistics have a formula when the sample size is large enough. The distribution of first digit with '1' is not 1/9, but 30.1%. The distribution of first digit with '2' is less than the 20%, only 17.6%. Then the distribution of the rest is less than the former.

Benford's law can also be used in the election fraud. For example, based on this law, the scientists found that the fraud in the US presidential election in Florida in 2004, in Venezuela in 2004 and in Mexico in 2006. Mathematicians have found that the frequency of header numbers in corporate books

were coincided with Benford's Law. If the person who falsifies the accounts changes in the real data in the books, the frequency of header numbers in the books will change and thus deviate from Benford's Law. More interestingly, mathematicians have changed the frequency of header numbers in the books. They also found that among those fake accounts, figures 5 and 6 are the most common seen starting numbers, not the law-abiding number 1, which indicates that forgers try to "hide" data in the middle of the accounts.

In the United States, the law has been applied in some practical fields. In economic terms, accountants can use this law to analyze the company's annual accounts and find forged data. The most typical example is Enron's "fake account" incident. Shortly after September 11, 2001, Enron, who was once the largest U.S. energy trader and the seventh largest global 500 company, suddenly declared bankruptcy without any warning and was exposed the scandal of suspected fake accounts. Afterwards, it was found that the earnings per share, published by Enron from 2001 to 2002, was totally inconsistent with Benford's Law, which proved that senior leaders of Enron did change the data[2].

3.2 Fuzzy Matching

Computer-based fuzzy matching is sometimes used to screen out false data. This is because there are people who would use fuzzy numbers to make the original data vague. How do people automatically search for word pairs of fuzzy matching? First, we assume the presence of word pairs before the searching. And next, we classify these words into two groups: function words and content words. Since the number of function words is limited, it is easy to distinguish function words from content words by screening them in a list of function words. By comparing the content words, it is assumed that all the content words can be matched with. Thus, we will have a lot of word pairs. The deletion algorithm is used to filter the candidate matchings, where the remaining points are fuzzy matchings. Finally, the similarity of each word pair could be calculated.

4.Discussion

For decades, the financial system has engaged in abusive practices or fraud. Our goal is to compare the prevalence of psychological distress and levels of health-related quality of life, based on exposure to financial fraud and its economic impact on household finances. The Madrid City Health Survey 2017 included specific questions on exposure to financial fraud, where half of the participants (n=4425) claimed to have experienced frauds. Furthermore, mental health needs were defined as having scores greater than two on 12 items on the Goldberg health questionnaire. The Darmouth Coop functional health assessment form /WONCA (Coop /WONCA) was used to assess health-related quality of life, where the incidence of financial fraud was 10.8%. The prevalence of severe economic impact was because fraud was 1.62 (95%, CI 1.17-2.25 compared with respondents without experience of fraud) after adjustment of age, sex, social class, and immigrant status. Men are more likely to be influenced by financial fraud compared with women. The current study contributes to a growing body of literature showing the effects of economic shocks on health as a result of financial fraud[3]. The two approaches above could help prevent further losses for companies. Among numerous daily transactions in the United States, credit card fraud is a common form of financial fraud as people could change their name without much restriction, it is easy to get a credit card. When people apply for a credit card, they need to provide personal information, including name, address, social number, telephone number, etc. However, the bank would not check whether the credibility of these information could enable exploitation of the loophole in the system. Therefore management needs to be enhanced to prevent cheating. Meanwhile, there are people who would fake information in order to possess more credit cards by changing the name, address, or telephone number. In the future, it is advisable to detect more types of financial fraud based on Python.

5. Conclusion

The author uses this paper to illustrate how to detect fraud and uses the Python as an instruction to

analyze the data in real life. The author also describes some means to prevent the fraud, such as fuzzy matching, and Benford' law. As this paper does not mention variable financial fraud, related studies can be proposed in the future.

References

[1] "2010 Property Maintenance Code of NY". [Online]. Available:https://up.codes/viewer/new_york/ny-property-maintenance-code-2010.[Accessed on Aug. 26, 2019]

[2] Bratton, William W. "Does Corporate Law Protect the Interests of Shareholders and Other Stakeholders? Enron and the Dark Side of Shareholder Value". Tulane Law Review. New Orleans: Tulane University Law School. May, 2002.

[3] Int. J. Environ. Res. Public Health 2019, 16(18), p.3276. [Online]. Available: https://doi.org/10.3390/ijerph16183276[Accessed on Aug. 26, 2019]