Fraud Analysis Using Machine Learning Algorithms

Molei Qin
Peiyang Park Campus, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, China

Abstract. In the previous studies done by others, the methods of picking abnormal numbers and data have been developed. This paper will focus on using machine-learning method to find abnormal value. Machine learning has been applied in all kinds of fields, including picture identifying and regression which is widely used in finding anomaly. The author finds some of the anomalies during the investigation and the fraud detection rate is higher than 80 percent. This paper not only aims at using traditional way of finding abnormal value, but also aims at using the modern computer science to deal with huge amount of data. Traditionally, when it comes to the topic of finding anomalies, people tend to think about t-checking or f-checking[1,2], which shares the same limitation that the target has to be fixed when being measured. However, a more realistic problem is that the target and the measuring environment is always changing. To fix the problem, this paper uses machine learning to do regressions of the environment and the target. Not only can it help us find the relationship between our target and ‘environment’, but also it can be written into computer elegantly so that it could be able to deal with huge mount of data.

1. Introduction
Economic behaviors, such as paying taxes for a particular item or a credit card transaction, are playing a more and more important role in people’s daily life, which might cause frauds such as telling a lie when asked about the price of a poverty, or stealing people’s credit card to buy a lot of things before the owner of the card knowing about it. Actually, these kinds of frauds are more prevalent than ever in that people can even make a great life by doing this kind of dirty work. Therefore, a method must be established to prevent this kind of things from happening. It is not easy, though, as fraud is only a small part of the huge amount of data, which makes check each one of them by human hand impossible. Computers, however, can do vastly considerable mount of calculation in a rather short time, making it a perfect tool for catching fraud. All kinds of frauds are for sure anomaly compared with regular data, which means that the fraud record may disobey some certain rules that other regular data obeys. Catching fraud, in this case, is like catching anomaly in a huge amount of data.

2. Methods and Analysis

2.1. Collecting data and data cleaning
First, enough mount of data are needed and the author should make a data quality report to make sure that there are few mistakes in our data set. Next step is to clean the data, which means that the author should fill in the blank with certain number to make every records usable in the following data analysis. The principal here is to fill in the blanks with normal number, making sure that the record will not appear strange when it is completed.
2.2. Variables-creating

The following part is creating variables, which is an essential part before building the model. Whether an algorithm is good or not can be largely dependent on the variables involved. As many variables as possible should be created in case that some important variables which could make the model work better might be left. Then the number of the variables should increase from tens to hundreds. In this process, the columns filled with text are deleted and the author only uses numerical variables to describe the records.

2.3. Supervised algorithms

Supervised model means that whether a record in a learning group is normal or not is known, as long as examples of frauds are given, what the author has to do is to develop an algorithm that can help separate the normal from abnormal, so this model can be used to predict whether it is fraud or not.

In the supervised model, the number of the dimensions should be reduced because it will take an absolutely long time to run the whole model with so many variables, and when the number of the dimensions is too high, it will be much harder for the models to find the nonlinear relationships among so many variables. Thus it is necessary to reduce dimensions. What has to be done here is to pick those variables that can make a difference to the result. There are three general methods to complete this process, including filter, wrapper and embedded. Filter is basically a method for picking certain variables, which can separate the fraud record from the regular ones when the author is dealing with binary classification issue. Fraud detection rate and the KS distance between the distributions of anomaly and the normal ones are great criterion in this process.

\[ KS = \max_{x} \int_{x_{min}}^{x_{max}} |P(x) - P(x')| dx' \]  

(1)

This method can work extremely fast but it does not work very well because the ability of variables to separate fraud from the regular data varies little, so in this paper the author only uses this kind of method to get away the first half of the variables. Wrapper is a process using simple models such as logistic regression or linear regression which involves different subsets of variables to get rid of the variables containing the models which work badly. It can reduce one variable at a time and can repeat the process again and again until getting a proper number of variables which is based on how many records the author has in all. Then embedded in the algorithms can be used and will be discussed in detail in the algorithms building part.

The supervised model means that the author has the examples of fraud, and there are two general algorithms that can be used, including neural network and decision tree.

The first one is neural network and the author uses the label which can help identify whether it is a fraud or not as the target output. And it is easy to see that the output is binary. So naturally the logistic function is used as the transfer function because the result of it is between 0 and 1. However, it is so delicate of the neural network that it can be easy to be overfitting. The real process of learning can be very meandering thus the machine may stop training when it achieves a local optimum. Fortunately, there are some details that make the neural network more robust. The first thing is that the learning rate can be changed while the model is training. Learning rate is basically the step length for each learning opportunity, which is known as the chance to change the value of the weight matrix. The author can use comparatively large number for learning rate at first and decrease it gradually. This process is like finding the lowest spot in a meandering surface. If you take a ball and let it slide naturally, it will probably find itself in a local optimum. However, if you shake the table, it will be more possible to find the lowest point. The second thing is that the author can optimize the way of changing the weight matrix. The gradient is normally used as the indicator of adjusting the weight matrix. However, there are more things that can be done. For example, it is acceptable to use the gradient as the only part of the indicator and add some adjustment term such as a fixed number to make the model more stable and make the model with only one node and one layer as well as gradually increase the complexity of the model. When the overfitting occurs, it is able to make best decision concerning super coefficients.

The second method for supervised model is the decision tree. The author simply divides the whole
data set and gives the record in the same area with the same label. It is a delicate thing as a computer can have a simple solution as it can simply divide every record. However, it will not work for the testing. So there are some other algorithms based on decision tree performing better due to the specific structure they have. Two of them will be introduced. One is boosting[6] tree and the other is random forest.

Random forest is a combination of hundreds of single decision trees. However, the structure of each tree varies a lot. They use different subset of variables as the decision indicator. If you draw a picture, it means that each decision tree is using different axes. The author combines all the decision tree and finally gets a result. The computer will automatically decide the position of the boundary to minimize the error which is similar to the neural network model’s loss.

Boosting is very similar to the random forest by its definition and they are all combinations of a large number of decision trees. What can separate boosting tree from random forest is that each decision tree has the same structure with the boosting tree. The author produces such series by adding an adjustment term for mistakes. It is more similar to a Taylor series. For instance, the author makes a simple division of records and then makes mistakes. Next, the author adds a decision tree to correct the mistakes and makes fewer mistakes. The process has been repeated again and again before final model become stable.

2.4. **Unsupervised algorithms**

Unsupervised model means that the author does not have any examples of frauds, so here is how the author deals with the problem. The author first scales the variables so they are basically in the same distribution where z-scale is a great example. Z-scale is standardizing the variables by using the subtraction of the real number and the mean divided by the standard deviation, making the mean of every variable 0 and making the standard deviation of every variable 1. And then the author can do PCA, which stands for principal component analysis. During the process, the author could not only remove the linear relationship among variables, but also select variables that have fluctuation. In data analysis, more fluctuations mean more information, so the author naturally selects variables that contain the most information. In this way, the author can reduce the dimensions of the unsupervised.

Now, what the author has for supervised model is records with selected features and what the author has for unsupervised model is records with tens of principle components. The author is going to build a criterion that can tell which record is regular and which is not.

For unsupervised model, the first thing the author needs to do is to z-scale the principal components because they certainly have different distributions and z-scale can make them equally important. Then there are two ways of analyzing the problem. The first way is simple and the author can view the record as a vector in a high-dimension space. Each of the variable can be viewed as a dimension or an axis in the high-dimension space. The author can simply use the distance between the record vector and the origin as the criterion because the author has already z-scaled the record and it should form a spherical cloud around the origin with the radius of one. If it is an anomaly, it should be far away from the origin.

The second method is using a machine-learning algorithm which is called autoencoder. Autoencoder[5] is a specific kind of neural network.

![Fig. 1. Structure of a neural net work](image)

Neural network is basically a model which can automatically evolve, and here is how it works. The structure of a neural net work is shown in figure 1. It consists of three parts. The first one is the input layer, where records are put. The second part is the hidden layer, where there are several layers hidden in this part. It is very similar to the way a brain works. There is a matrix in the interval of layers and the output vector multiply this matrix to form the input of the next layer. Through a transfer function, which
can provide some nonlinear terms in this algorithm, the input of the layer transform into the output of the layer. The third part is the output layer, where the output is produced. Then the output and the real result of the input will be compared together and produce a value measuring loss through a loss function. The whole system is self-evolving because the loss is the controller of the the matrix and this process will automatically reduce the loss. The number of ways to control the matrix is more than considered, and here is a basic method about how to realize the control. Since the transfer function, the input and the output are all fixed and the loss function as a function only concerning the matrix can be seen. That is to say, there is a fixed functional relationship between the matrix and the loss. Surely loss can be represented by the matrix and the author can take the partial derivative for the loss by one variable in the matrix. When the result is below 0, the author can make the variable in the matrix a little higher to reduce the loss and make the variable a little bit lower when the partial derivatives is above 0 to reduce the loss.

Fig. 2. Think this way for model building process, algorithm design and selection

Additionally, autoencoder is basically a neural network whose input is exactly the same as the target output. It is primarily used in the dimensional reduction, because when the functional relationship is written down, there is a hidden functional relationship among the variables. Also, some changes to the equation can be done, and finally the one variable can be represented by other variables. Here, the author does not use it to reduce the number of dimensions but uses it to find the hidden functional relationship among variables. When the record is normal, the variables in this record will certainly obey the hidden relationship. As a result, the record can be reproduced well. However, when the record is abnormal, the variables tend not to obey the hidden relationship so the record is reproduced horribly. The autoencoder does not need to be perfect because the error part is needed only. The bigger the error for a record, the more abnormal it tends to be.

The next step is to combine the two results which can be concluded from methods above for unsupervised model. One way is ranking. Due to the varieties of the definitions of the distance and the specific structure of the autoencoder, the absolute value of each record in these two methods means nothing. However, the author can combine the ranks for each record together and make a whole ranking order for them. And now, it is known which record is the most-likely anomaly while others tend not to be.

2.5. Result analysis
For unsupervised model, only ranked suspicious list is provided. The author can simplify and investigate it one by one, from high possibility of fraud to low possibility of fraud. If enough numbers of fraud are found, the algorithms will work well. For supervised model, what the author has is records with the possibility of being a fraud. The author can simply apply the result to a new data set and see how it works. Fraud detection rate is always a great criterion in this part.

3. Conclusion
Given different kind of initial conditions, the problems could be divided into supervised regression and unsupervised problems. After cleaning the data and make some feature selection process, we use machine learning to fix the regression problem and it works well. The machine-learning solution to the
regression problem could be a great way of solving other business problems such as prediction[3] and detection[4].

References
[1] Comparison between linear model and machine learning model in prediction of yak weight, Modern Agricultural Sciences and Technology, 2019. [Online] Available: http://kns.cnki.net/kcms/detail/34.1278.S.20191209.1349.242.html [Accessed on Oct. 15, 2019]
[2] P. C. Yu and L. W. Zhong, Prediction of product testing time based on linear regression, China Water Transport, vol. 18, pp.79-80,83, 2018.
[3] J. P. Pu, The application of machine learning in the prediction of commodity sales, Practical Electronics, vol. 22, pp.81-82,70, 2018.
[4] M. J. Zhang, W. J. Yu, W. B. Li, X. D. Zhu, Design of license plate recognition system based on machine learning [J/OL]. Computer technology and development, 2020 (04): 1-9, [Online] Available: Http://kns.cnki.net/kcms/detail/61.1450.tp.20191218.1113.046.html [Accessed on Dec. 19, 2019]
[5] C. H. Zhang, X. T. Zhou, Y. A. Zhang, J. M. Yao, T. L. Guo, Q. Yan, Research on Application of deep self encoder in data anomaly detection [J/OL], Computer engineering and application: 1-9 [Online] Available: Http://kns.cnki.net/kcms/detail/11.2127.tp.20191213.1448.010.html [Accessed on Dec. 15, 2019]
[6] Rui Sun, Guanyu Wang, Wenyu Zhang, Li-Ta Hsu, Washington Y. Ochieng, A gradient boosting decision tree based GPS signal reception classification algorithm[J], Applied Soft Computing Journal, 2019.