2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9389949

# Supply Chain Fraud Prediction Based On XGBoost Method

Yichun Zhou[1],
New York University,
New York, China,
yz6176@nyu.edu,

Xuyang Song[1]
Hubei University Of Technology
Hubei, China
276439774@qq.com

Mengyuan Zhou[2],
Fuzhou University,
Fuzhou, China,
2210549527@qq.com

***Abstract*—It is very meaningful to build a model based on supply chain data to determine whether there is fraud in the product transaction process. It can help merchants in the supply chain avoid fraud, default and credit risks, and improve market order. In this paper, I propose a fraud prediction model based on XGBoost. The data set required to build the model comes from the supply chain data provided by DataGo. Compared with the model based on Logistic regression and the model of Gausian Naive bayes, the model proposed in this paper shows better classification ability. Specifically, the F1 score based on the Logistic regression model is 98.96, the F1 score based on the Gausian Naive bayes model is 71.95, and the F1 score value of the XGBoost-based model proposed in this paper is 99.31 in the experiment.**

***Index Terms*—Fraud prediction, XGBoost, Machine learning algorithms, Data Mining,**

## I. INTRODUCTION

The supply chain of many products has the characteristics of low degree of informationization, scattered market participants, large number of enterprises and small scale, rapid price changes, and insufficient information on industrial and commercial, taxation, and historical transaction records of buyers and sellers. The current market situation is not fully understood, and fraud, default and credit risks are often encountered in the supply chain. Machine learning and data mining technology can analyze the previous supply chain data to determine whether there is fraud in the circulation of a product. Avoid fraud, default and credit risks, and make the market order more perfect.

With the rapid development of computer technology and machine learning, machine learning algorithms have been widely used in various industries to solve practical problems. In recent years, machine learning has received extensive attention from researchers in judging credit default risks, which can help the market establish a more transparent and honest trading system. Therefore, this paper aims to use the supply chain data provided by DataGo to establish a model for judging whether the product is fraudulent.

In the experimental phase, I first performed feature engineering processing, including memory compression, outlier processing, statistical features, and selection of important features. Experimental results show that the fraud prediction model based on XGBoost is better than other machine learning models, such as Logistic regression algorithm and Gausian Naive bayes algorithm. Specifically, the F1 Score value of our XGBoost model is 99.31, while the F1 Score of the model based on the Logistic regression algorithm is 98.96, and the F1 Score of the Gausian Naive bayes algorithm is 71.95. In addition, I observed the structural characteristics of some products based on statistical data, and obtained some important findings to guide future work. The experiment proved that the idea of using the XGBoost algorithm is effective for judging whether the products in the supply chain are fraudulent.

In summary, the contributions of this paper are as follows:

1) I proposed a model based on XGBoost algorithm to determine whether there is fraud in product transactions. This model can effectively mine the characteristics of products in different dimensions in the supply chain.

2) During the experiment, I performed detailed feature engineering processing, such as memory compression, abnormal data processing, statistical features, and important feature selection. Experiments show that feature engineering is effective for training models.

3) I used the supply chain data provided by DataGo to evaluate the XGBoost method. Experiments show that the model based on XGBoost algorithm is better than Logistic regression and Gausian Naive bayes algorithm. At the same time, the structural characteristics of the product in different dimensions show some interesting conclusions that can guide future work.

### A. Related Work

Based on the supply chain data provided by DataGo, a classification model is built to determine whether the product is fraudulent in the transaction process. The classification model is to build a predictable model by analyzing and processing previous data information. In the modeling process, feature extraction and application of data sets are very important. I should find a suitable model to describe the relationship between features. Thus, the products are classified according to the characteristics of the data. In the field of machine learning,

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

many scientists have made outstanding contributions, and their work inspired the experiments in this paper.

In [1], it mainly introduces the gradient-enhanced decision tree, a machine learning algorithm. It can estimate all possible segmentation points of information acquisition by scanning all data instances, but this method has high time complexity, small gradient and combined with mutual exclusion characteristics. Regression models are introduced in [2], [3] and [4]. The goal of function-to-function regression is to establish a mapping from function predictor variables to function response. A functional regression model based on pattern sparse regularization is proposed, and the modal sparse regularization method is used to automatically filter irrelevant functions. [5] [6] [7] introduced very classic machine learning methods, among which the linear regression, logistic regression and random forest methods mentioned in this paper have good performance and interpretability, and their application scenarios are also very good

In [8], three machine learning models of GBDT, Xgboost and LightGBM were used to predict monthly house rents. By comparing and analyzing the prediction results under different data sets training, it is found that the Xgboost and LightGBM models are better than the traditional GBDT model.

In [9] and [10], it mainly introduces the grid search algorithm. Two focused grid search schemes are proposed in the paper. By repeatedly zooming into a more concentrated discrete grid point set in the parameter search space, this method is faster than the standard search efficiency. In [11], it mainly introduces the Bayesian linear regression model, which uses the BLR model to predict data points, and uses a multi-processor design to improve computing power, and finally outputs the predicted data points after learning.

The previous results in this field have given me a lot of inspiration. In the task of this paper, I proposed the use of a classification model based on XGboost to achieve product fraud prediction.

## II. FEATURE ENGINEERING

Feature engineering is the most important step in all machine learning. It can be divided into the following steps: coding, sorting, automatic feature selection, etc. Our data set comes from the supply chain data provided by DataGo.

First, I encode the data set, because some of the functions in the table are discrete values. Continuous values are easy to view the laws of data, but discrete values are difficult to participate in calculations, so coding is required. There are many coding methods. The well-known one is one-hot encoding, which means that if a feature takes a certain value, the corresponding "has" feature takes the value 1, and other features take the value 0.

Feature selection is also important in the development of iterative selection. There are three basic strategies to judge the importance of each function: univariate statistics, model-based selection and iterative selection. All these methods are supervised methods, that is, they need labels to train the model.

In the machine learning model, the parameters that need to be manually selected are called hyperparameters. For example, the number of decision trees in the random forest, the number of hidden layers and nodes in each layer of the artificial neural network model, and the constant size of the regular items need to be set before use. If the hyperparameters are not selected correctly, the problem of underfitting or overfitting will occur. When selecting hyperparameters, there are two methods: one is to fine-tune based on experience. The other is to select parameters of different sizes and bring them into the model to select the parameters with the best performance. One method of fine-tuning is to manually modulate the hyperparameters until a good combination of hyperparameters is found, which may require a lot of work, so a grid search can be used to search.

I also used grid search in the experiment and optimized the model by adjusting the maximum depth, learning rate and other parameters.

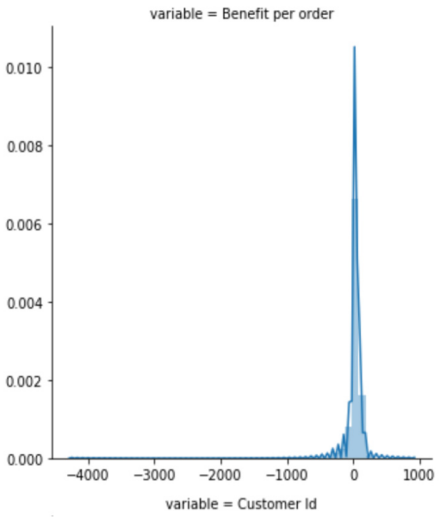


Figure 1. Profit distribution of different customers

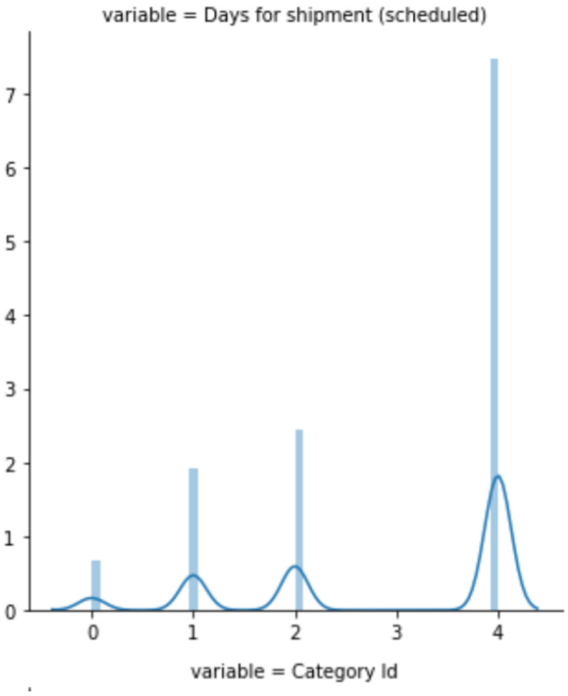


Figure 2. Distribution of mailing time of different products

## III. XGBOOST

I select three models for experiments, and compared and analyzed the experimental results. Compared with Logistic regression model and Gausian Naive bayes model, XGBoost model performs best. Its F1 Score is 99.31, and its prediction accuracy far exceeds 98.96 of Logistic regression model and 71.95 of Gausian Naive bayes.

XGBoost is a tree ensemble model that adds the results of K (the number of trees) as the final predicted value. The key problem of tree model learning is how to find the best split point. The first method is called the basic exact greedy algorithm, which enumerates all possible segmentation of all features and finds the best segmentation point. The algorithm needs to enumerate all possible segmentation of continuous features, which is very demanding on the computer. In order to perform this operation effectively, XGBoost first sorts the elements, and then accesses the data sequentially, thereby accumulating gradient statistics in terms of loss reduction.

Compared with the GBDT algorithm, The base learners of the XGBoost algorithm have both tree classifiers (gbtree) and linear classifiers (gblinear). Simultaneously, the XGBoost algorithm executes second-order Taylor expansion on the objective function. And to avoid overfitting, the XGBoost algorithm uses the complexity function of the tree as the constant term of the objective loss function. The detailed formula of the objective loss function of our XGBoost method is as follows:

$$Object(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \Omega(f_t) + C,$$
$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i),$$
$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \parallel w \parallel^2,$$

where C is a constant, $x_i$ is the input vector, $T_t$ represents the number of leaves in the tree, $\gamma$ and $\lambda$ are hyperparameter. $y_i$ and $\hat{y}_i^t$ mean the real value and the predicted value of sales, respectively. $l(\cdot)$ is the square loss function and $f_t(\cdot)$ is a regression tree. However, according to the Taylor's formula, object function $Object(\cdot)$ is approximately expressed as follows:

$$Object(t) \approx \sum_{i=1}^{n}[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) + C,$$

where $g_i$ and $\frac{1}{2}h_i$ are the coefficients of the first and quadratic terms in Taylor expansion, respectively [9].

XGBoost has the following advantages:

1) Xgboost adds a regularization term to the objective function. The regularization term is related to the number of leaf nodes T of the tree and the value of the leaf nodes. The regular term contains the number of leaf nodes of the tree and the square sum of the L2 modulus of the score output on each leaf node. From the perspective of Bias-variance tradeoff, the regular term reduces the variance of the model, makes the learned model simpler, and can effectively prevent overfitting. At the same time, XGBoost introduces a reduction factor in each step to reduce the impact of a single tree on the results, so that subsequent models have more room for optimization, and further prevent overfitting.

2) Xgboost not only uses the first derivative, but also the second derivative, the loss is more accurate, and the loss can be customized. Before training, the model sorts the data in advance and saves it as a block. It is reused in subsequent iterations to reduce calculations. At the same time, when calculating the split point, it can be calculated in parallel. Parallel approximate

histogram algorithm requires tree nodes to calculate the gain of each node when splitting. If the amount of data is large, the features of all nodes need to be sorted, and then the optimal split point is obtained through traversal. This greedy method is very time-consuming. At this time, an approximate histogram algorithm is introduced to generate efficient split points, that is, a certain value after the split is subtracted from a certain value before the split to obtain a gain. In order to limit the growth of the tree, Introduce a threshold, when the gain is greater than the threshold, split.

## IV. EXPERIMENTS

In this paper, I use the F1 Score metric, which is often used in multi-classification problems. The higher the F1 Score value, the better the prediction effect of the model.,

TABLE 1 PERFORMANCE COMPARISON OF MODELS

| **Models** | **F1 Score** |
|---|---|
| Logistic Regression | 98.96 |
| Gausian Naive bayes | 71.95 |
| **XgBoost** | **99.31** |

Table 1 shows the model based on other algorithms and the experimental results based on the XGBoost model. It is easy to find that our XGBoost model has the highest F1 Score value. Experiments show that I use the XGBoost algorithm and feature engineering methods to be effective for the task of judging whether there is fraud in product transactions.

## V. CONCLUSIONS

In this paper, I propose a new method for judging whether there are fraudulent behaviors of goods in the supply chain. I combine feature engineering processing and machine learning algorithms on the supply chain data set provided by DataGo. Experiments show that, compared with Logistic regression algorithm and Gausian Naive bayes algorithm, the model based on XGBoost shows better predictive ability in commodity fraud classification. In addition, I have obtained some structural characteristics of the data in data processing, and he has constructive inspiration for the classification of products in more dimensions.


## ACKNOWLEDGEMENT

I sincerely thank DataGo for the supply chain data set.


## REFERENCES


[1] Thomas Finley.LightGBM: A Highly Efficient Gradient Boosting Decision Tree.2011.

[2] Tibor Kiss,Claudia RochJan,Jan Strunk.A logistic regression model of determiner omission in PPs.2010

[3] Mohiuddeen Khan,Kanishk Srivastava. Regression Model for Better Generalization and Regression Analysis.2016

[4] Autcha Araveeporn, Choojai Kuharatanachai. Comparing Penalized Regression Analysis of Logistic Regression Model with Multicollinearity.2018.

[5] Breiman, Leo. Random Forests. Machine Learning 45 (1), 5-32, 2001.

[6] David W Hosmer, Stanley Lemeshow. Applied logistic regression. Technometrics. 2000.

[7] Breiman, L., Friedman, J. Olshen, R. and Stone C. Classification and Regression Trees, Wadsworth, 1984.

[8] Xiang Wei, Mengzhong Ji, Jun Peng.The application analysis of forecasting housing monthly rent based on Xgboost and LightGBM algorithm.2019.

[9] Basrak,Z.A routine for parameter potimization using an accelerated grid-search method.1987.

[10] Alvaro Barbero Jimenez,Jorge Lopez Lazaro.Finding optimal model parameters by deterministic and annealed focused grid search.2008.

[11] Rifkin R M,Lippert R A.Computer-implemented method of performing linear regression involves applying Bayesian linear regression model to predict data points and outputting predicted data points.2008.