Sales Forecast of CNC Machine Tool Suppliers Based on Ensemble Learning

Kai Zhong, Jiwen Wen* and Jia Zheng

Institute of Economics and Management, Beijing Forestry University, Beijing, 100083, China

*E-mail: wjwlinda@163.com

Abstract. CNC machine tools, whose product sales forecast is a complicated dynamic process, are the basic equipments for manufacturing industry. It is difficult for a single prediction model to achieve the expected effects because product sales can be affected by multiple factors with features like time-varying, non-linear and random. This paper, which takes the sales data of a CNC machine tool supplier as an example, designs an ensemble learning model based on XGBoost and random forest, obtaining lower prediction error through data cleaning, data conversion and parameter optimization. The experimental results show that the generalization ability of this model is excellent, so it can provide decision supports for CNC machine tool suppliers in purchase and sales plans.

1. Introduction

Good sales forecast is an important reference for decision-making in manufacturing operations. Affected by technology, organization, personnel and capital, the effects of sales forecast is often unsatisfactory, which leads to the forecast results have no real reference value and the cost of enterprises increases in vain[1]. Therefore, how to use existing sales data to provide production decision support effectively has become the focus of enterprises.

CNC machine tool industry belongs to manufacturing industry, and its product sales forecast is a complicated dynamic process. Firstly, sales data will be affected by multiple factors, including internal factors and external factors from macroeconomic and industrial environmental. Secondly, these factors are time-varying, non-linear and random. Thirdly, it is difficult for a single prediction model to achieve the ideal prediction effect because complicated non-linear relationship exists between sales data and influencing factors[2].

There are many methods for product sales forecast, such as time series, regression analysis, neural network and ensemble learning. Among them, ARMA (Auto-Regressive Moving Average Model) in time series analysis is often used for linear prediction, but the actual sales data of enterprises are mostly non-linear time series. Neural network is often used for non-linear prediction for its strong non-linear mapping ability, however, it could be easily affected by the complexity of network structure and samples, leading to problems such as over-fitting or low generalization ability[3]. Besides, a single model is not stable, and such instability will have a negative impact on the generalization ability of a single model if the features used in the training set and verification set did not perform well[4]. Nowadays, many studies apply ensemble learning models into solving sales forecast problems. The main advantage of the ensemble model is reflected in its improved generalization capabilities and flexible functional mapping between system variables. The most common ensemble learning frameworks are Bagging[5], Generalized Stacking[6] and Boosting[7]. Among them, XGBoost (eXtreme Gradient Boosting), which
has advantages in overcoming sample missing, preventing model over-fitting and reducing computational difficulty, is applied by Jiang Jinwen[8], Ye Qianyi[9] to manufacturing quality forecast and market sales forecast to achieve accurate prediction; random forest, which is the most representative algorithm in Bagging because of its simplicity, is applied by Pal N[10], Chang Xiaohua[11] to automobile price forecast and medical sales forecast to improve the prediction accuracy for its easy implementation and low computational cost.

Taking the sales data of a CNC machine tool supplier as an example, this paper designs an ensemble learning model by data cleaning, data conversion, parameter optimization and other processing based on XGBoost and random forest. It can overcome the influence of neural networks easily falling into local minimum value, solve the instability of a single model and improve the accuracy and validity of prediction.

2. Sales business and data preparation of a CNC machine tool supplier

2.1. Sales business analysis

The picked CNC machine tool supplier specializes in the production, sales and maintenance of CNC machine tool devices. Its main business is sales. Effective sales forecast helps enterprises to make correct sales decisions and take an advantage in reducing sales costs and scheduling resources.

As shown in Table 1, the sales performance of the picked supplier is affected by many factors which mainly come from customers, competitors, macroeconomic, industry, product and so on.

- Customer factors. The target customer of the picked supplier is industrial enterprises, mainly considering the size of the enterprise, the type of industry (segment market), the geographical location, etc.
- Competitive factors. The main competitors of the picked supplier are Siemens, Mitsubishi, Yaskawa, etc.
- Macroeconomic factors. Macroeconomic factors will affect the sales trend of the machine tool industry, generally including China’s GDP (trillion yuan), total investment (billion yuan), etc.
- Industry factors. CNC machine tool suppliers belong to industrial manufacturing industry, whose industry factors include CPI index, PMI index. The picked supplier specializes in production, sales and maintenance of CNC machine tools, so the overall development trend of the machine tool industry, whose industry factors include the output of CNC forming machine tools (10,000 units), the import of CNC machine tools (10,000 units) and so on, has a greater impact on it. In addition, whether the end-user’s industry of the picked supplier is prosperous or not affects its purchasing power and demand, thus considering the industry factors of market segment is needed. Among them, the automobile manufacturing industry is quite important, the main reference factors include passenger car output (10,000 units), passenger car sales volume (10,000 units) and so on. In 2014, the electronics industry represented by smart phones has a strong demand for CNC machine tool drilling centers and vertical machining centers. Therefore, the influence of the electronics industry needs to be taken into account. The main influencing factors include smartphones (10,000 units), non-smart phones (10,000 units), domestic mobile phones (10,000 units) and imported mobile phones (10,000 units).
- Product factors. The competitiveness of the product itself, including performance, price, quality, service and so on, affects the demand for products in the future market directly.

Table 1. Summary of factors affecting the picked supplier’s sales volume.

| Classification          | Attributes                                |
|------------------------|-------------------------------------------|
| Customer factors       | customer name, city, market segment       |
| Competitive factors    | SIEMENS: sales volume of SIEMENS, sales of SIEMENS |
|                       | MITSUBISHI: sales volume of MITSUBISHI, sales of MITSUBISHI |
|                       | GSK: sales volume of GSK, sales of GSK    |
|                       | SYNTPEC: sales volume of SYNTPEC, sales of SYNTPEC |
Total of major competitors: Total sales of major competitors, Total sales of major competitors

Macroeconomic factors

- Economic data: China's GDP (trillion yuan), foreign exchange reserve balance (billion dollars), total exports (thousand dollars), total imports (thousand dollars), private fixed assets investment (billion yuan), national fixed assets investment (billion yuan), total investment (billion yuan), total consumption (billion yuan)

Industry factors

- Industrial manufacturing industry: CPI index (industrial added value), PMI index (prosperity index of manufacturing industry)
- CNC machine tool industry: common forming machine tool output (10,000 units), common metal cutting machines output (10,000 units), CNC forming machine tool output (10,000 units), CNC machine tool import (10,000 units), CNC metal cutting machines output (10,000 units), CNC metal cutting machines output (10,000 units)
- Automobile manufacturing industry: passenger car production (10,000 units), passenger car sales (10,000 units), commercial vehicle production (10,000 units), commercial vehicle sales (10,000 units), new energy vehicle production (10,000 units), new energy vehicle sales (10,000 units)
- Electronics industry: smart phones (10,000 units), non-smart phones (10,000 units), domestic mobile phones (10,000 units), imported mobile phones (10,000 units)

Product factors

- Product name, machine tool type, sales organization, sales order price, order number, order amount

2.2. Sales data preprocessing

2.2.1. Data collection. Based on the above analysis of influencing factors, this paper collects the sales performance data of the picked supplier and its competitors, as well as the macroeconomic data and industry data of the CNC machine tool industry and its related industries.

Sales performance: data of all sales orders from January 2016 to March 2018, including month, sales order number, city, customer, product, machine tool type, market segment, sales organization, sales order price, order number and order amount, a total of 9 features, 33,696 data.

Competitor’s Performance: From January 2016 to March 2018, monthly sales volume and sales of competitors, etc., a total of 10 features.

Macroeconomic and industry: From January 2016 to March 2018, the macroeconomic and industry indicators for each month, such as month, CNC machine cutting machine exports (10,000 units), etc., a total of 26 features.

2.2.2. Data description. The object of study is mainly for the picked supplier with more than 200 orders, including models as followed: 0I-MF, 0I-TF, XSBJ, 0I M-MD(5), 0I-MD, 0I-TD, 31I-B, 32I -B, 0I M-TD (5) and 0I-TF (G).

The data set is large, with a total span of 33,696 data from 2016 to 2018, and predicts for a variety of products with considerable scale in both time and space.

Problems are as followed: existing many dirty data; collecting industry and macroeconomic data in some years; containing missing values and cross-border values in some features; filling and eliminating some data is needed.

The characteristics of the data set include time, product, competitors’ information, macroeconomic information, industry information and so on. The correlation degree between these features is low, and the impact of many features on sales is difficult to express by linear model. An example of how monthly sales of different products change over time is shown in Figure 1. Considering the influence of time and product on sales volume, we can see that the sales volume of different products varies significantly with
time, and some products even fluctuate greatly. The prediction accuracy is not high only from the perspective of time, and it is difficult to describe linearly.

Figure 1. Sales trend graph of different products over time.

2.2.3. Data cleaning. Missing value processing: counting the missing values in the data set, we can find that the output of CNC forming machine tool and ordinary forming machine account for a very large proportion, so these two features are eliminated. For the missing data of macroeconomic and industry information, the average value is used for replacement. The missing data of product and customer information, including market segments, machine tool types, cities, are processed by filling forward or backward.

Exception processing: In sales orders, sales volume might be empty. Since the sales volume is equal to unit price multiplied by the sales volume, so when the sales volume is empty, the order is invalid and should be deleted directly. When the unit price is empty, it can be recalculated based on sales and sales volume. When 2 or 3 features are empty, it is considered to be an invalid order and should be deleted directly.

2.2.4. Data conversion and normalization. After data cleaning, we need to consider whether the current data form can be used for modelling. XGboost and random forest are only suitable for dealing with numerical vectors, but in the original data set, product types, cities and other features are numerically independent. In this paper, One-hot Encoding is used to flatten these features and extend them to a new Boolean dimension. Taking product type conversion as an example, as shown in Table 2.

Table 2. One-hot encoding after product type processing.

| product type | 31I-B | XSBJ | 32I-B | 01-TD | 01-M- | 01-MF | 01- | 01-M- | 01-TF | 01-M- |
|--------------|-------|------|-------|------|-------|------|-----|-------|-------|-------|
|              | 31I-B | XSBJ | 32I-B | 01-TD | 01-M- | 01-MF | 01- | 01-M- | 01-TF | 01-M- |
| 31I-B        | 1     | 0    | 0     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |
| XSBJ         | 0     | 1    | 0     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |
| 32I-B        | 0     | 0    | 1     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |
| 01-TD        | 0     | 0    | 0     | 1     | 0     | 0     | 0   | 0     | 0     | 0     |
| 01-M-TD(5)   | 0     | 0    | 0     | 0     | 1     | 0     | 0   | 0     | 0     | 0     |
| 01-MF        | 0     | 0    | 0     | 0     | 0     | 1     | 0   | 0     | 0     | 0     |
| 01-MD        | 0     | 0    | 0     | 0     | 0     | 0     | 1   | 0     | 0     | 0     |
| 01-TF        | 0     | 0    | 0     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |
| 01-M-MD(5)   | 0     | 0    | 0     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |
| 01-TF(G)     | 0     | 0    | 0     | 0     | 0     | 0     | 0   | 0     | 0     | 0     |

3. Sales forecast model based on ensemble learning

3.1. Ensemble learning algorithm

Bagging[5] is a parallel ensemble learning method, and random forest[12] is an extended variant of Bagging. Random forest is simple, easy to implement and has low computational cost, but it can show powerful performance in many practical tasks. In view of the shortcomings that decision trees are prone
to over-fitting, random forest uses voting mechanisms of multiple decision trees to improve decision trees. Random forest initiation performance is often relatively poor, but with the increasing of base learners, generalization error will be lower, thus its training efficiency is often better than that of base Bagging[13].

Boosting[7] is a common serial ensemble learning method. By changing the weight of samples in the training process, the optimal learner is finally obtained. On the basis of traditional Boosting, XGboost[14] uses CPU multi-threading, introducing regularization items and pruning, which reduces the complexity of the model. The advantages are as follows: XGboost can learn splitting direction automatically when there are missing values in samples; the cost function of XGboost introduces regularization to prevent model over-fitting and reduce model complexity; XGboost allocates learning rate of leaf nodes after each iteration, reduces the weight of each tree and provides better learning space.

Stacking[6] is a typical representative of learning methods. In the training stage, the secondary training set is generated by the primary learner. If the training set of the primary learner is used to generate the secondary training set directly, the over-fitting risk will be quite large. Therefore, the training samples of the secondary learner are usually generated by using cross-validation or leave-one method, using the unused samples of the primary learner. There are many advantages of using combination strategy[15]. It can reduce the risk of poor generalization performance caused by wrong selection of single learner, reduce the risk of falling into bad local minima, and improve the generalization performance corresponding to local minima.

3.2. Model analysis
Based on the characteristics of the picked supplier’s sales data set, the prediction model needs to meet the following basic requirements: have low prediction error on non-linear regression problems; adapt to the data sets with diversified data categories; perform well in most cases and have good generalization ability; be easy to explain and assist sales decision-making.

Both the random forest and XGboost meet the above requirements. First of all, both of them are based on tree model and perform well in non-linear regression problems. Secondly, both of them belong to combination model, and have higher precision and better generalization ability than single model. Thirdly, both of them don’t have too much requirement for input data, and their training process does not involve the calculation of feature values, which make them applicable in most application scenarios. Finally, after the training process, XGboost can count the important indicators of all features, which can guide the subsequent sales decisions.

Random forest and XGboost have different advantages and disadvantages. Random forest has good generalization ability, fast training speed, and performs well in many problems, while XGboost training process is serial iteration, slow training speed, high fitting degree and higher precision under certain circumstances. In general, random forest shows better ability to resist over-fitting when processing in moderate precision range, training results with certain accuracy quickly, XGboost can converge to the high accuracy range through continuous iteration, and the fitting effect is good in the training of high epochs.

Therefore, this paper uses random forest and XGboost to train the corresponding models respectively, and then carries on combination forecasting with Stacking method. After Comparing the training speed and evaluation index of each model and making excellent adjustment of parameters, the optimal model will be found to further improve the accuracy and generalization ability of the prediction model.

3.3. Experimental process
This paper uses Python language programming to train, validate and test the model. The training set is 80%, the validation set is 10%, and the test set is 10%.

3.3.1. Parameter optimization. XGboost model. The processed data are trained by XGboost model, and the model parameters are optimized by Grid Search. The adjusted parameters are shown in Table 3.
Table 3. XGboost model parameters.

| parameter        | value | parameter        | value |
|------------------|-------|------------------|-------|
| max_depth        | 10    | learning_rate    | 0.1   |
| n_estimators     | 500   | seed             | 42    |
| subsample        | 0.8   | objective        | reg: linear |
| colsample_bytree | 0.85  | nthrea           | 6     |

Random forest model. The processed data are trained by random forest model, and the model parameters are optimized by Grid Search. The adjusted parameters are shown in Table 4.

Table 4. Random forest model parameters.

| parameter          | value     | parameter            | value |
|--------------------|-----------|----------------------|-------|
| n_estimators       | 1000      | min_samples_split    | 2     |
| criterion          | mse       | max_depth            | 50    |
| random_state       | 1         | min_weight_fraction_leaf | 0   |
| n_jobs             | -1        |                       |       |

Sales forecast model based on Staking. The mlxtend library is used to complete the stacking of the model, and the feature outputs generated by the previous classifier are used as the input data of the last total meta-classifier. The basic parameters are shown in Table 5.

Table 5. Stacking model parameters.

| parameter            | value | parameter                        | value     |
|----------------------|-------|----------------------------------|-----------|
| classifiers          | Xgboost, rf | average_probas                | False    |
| meta_classifier      | ridge  | verbose                         | 1         |
| use_probas           | False  | use_features_in_secondary       | False    |

3.3.2. Analysis of the importance of factors. Figure 2 shows the importance ranking of influencing factors in XGboost model. It can be found that features related to the CNC industry who have the greatest impact are as followed: the export volume of CNC metal cutting machines (10,000 units), the import volume of CNC machine tools (10,000 units), the output of CNC metal machines cutting machine (10,000 units), and the output of common metal cutting machines tool (10,000 units) and PMI index, which means indicates that the overall sales trend of the CNC industry has a greater impact on the sales of a certain company.

3.4. Model Evaluation
In order to compare the generalization effect of each model, this paper uses the RMSE value to measure each single model respectively and the combined model between them to evaluate the generalization ability of the model on the test set (as shown in Table 6).
Table 6. RMSE value of each model.

| Model                                      | RMSE   |
|--------------------------------------------|--------|
| Time series model                          | 0.3129 |
| Multiple linear regression model           | 1.8927 |
| XGboost model                              | 0.0472 |
| random forest model                        | 0.0527 |
| XGboost model+ random forest model         | 0.0531 |
| XGboost model+ random forest model + ridge | 0.0459 |

It can be seen that the XGboost single model is better than other models. By combining XGboost model and random forest model with Stacking method, we can obtain the optimization model XGboost model + random forest model + ridge, which improves the RMSPE value but not significantly. The performance of single model good enough, and the generalization performance is optimal.

Figure 3 shows the change process of the model’s RMSE values in the training set. The abscissa is the number of times and the ordinate is the RMSE value. We can find that the RMSE value of the first 60 training times in the training set reduces to 0.05 rapidly, and reduces to 0.04 after 180 training times. It can be seen that the effect in the training set is very good.

Figure 3. RMSE values in the training set.

4. Conclusion and discussion
There are many factors affecting the sales of CNC machine tools, including customer factors, competitive factors, macroeconomic factors, industry factors and product factors, but most of them are difficult to express by linear models. The ensemble learning model performs well in solving non-linear regression problems, which can not only overcome the difficulties of these influencing factors in the prediction process but also process with higher accuracy and generalization ability.

This paper compares the effects of time series model, multiple regression model, XGboost model and random forest model prediction analysis, finding that XGboost has obvious advantages in both convergence speed and RMSE evaluation criteria. In order to further improve the accuracy and generalization ability of the prediction model, this paper combines XGboost and random forest by Stacking method, and obtains the combination optimization model. Experiments show that the combined model is better than the single XGboost prediction model in performance.

The combination model based on XGboost and random forest could be broad from the sales forecast of the picked supplier to other selling practices with similar features. Moreover, it plays an important guiding role in improving the operation mode, inventory management, price management and precision marketing of enterprises.

References
[1] Liu, Y. (2014) Analysis of Commodity Sales Forecast Based on Data Mining. Bulletin of Science and Technology, 7: 140-143.
[2] Wang, D.X., H.Z.Y., Q. L.J.(2015) Product sales forecast based on chaotic fruit fly support vector machine regression. Computer Engineering and Applications, 12: 235-240.
[3] Haykin, S. (1994) Neural Networks: A Comprehensive Foundation. The Knowledge Engineering Review, 13: 409-412.

[4] Chen, H., Cohn, A. G., Yao, X. (2012) Ensemble Machine Learning. Springer, Boston.

[5] Breiman, L. (1996) Bagging Predictors. Machine Learning, 24:123-140.

[6] Wolpert, D. H. (1992) Stacked generalization. Neural Networks, 5:241-259.

[7] Schapire, R. E. (1999) Improved boosting algorithms using confidence-rated predictions. Machine Learning, 37:297-336.

[8] Jiang, J.W., Liu, W.G. (2017) Application of XGBoost Algorithm in Manufacturing Quality Prediction. Journal of Intelligent computers and applications, 6: 58-60.

[9] Ye, Q.Y., Rao, Y., Ji, M.S. (2017) Commercial Sales Forecast Based on Xgboost. Journal of Nanchang University (Science and Technology), 3:275-281.

[10] Pal, N., Arora, P., Sundararaman, D., et al. (2017). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Future of Information and Communications Conference (FICC). London. pp. 1-6.

[11] Chang, X.H., Xiong, W. (2018) Application of Adaboost-based random forest algorithm in medical sales forecasting. Computer Systems Applications, 2:202-206.

[12] Breiman, L. (2001) Random Forests. Machine Learning, 45:5-32.

[13] Zhou, Z.H. (2016) Ensemble Learning. In: Zhou, Z.H. (Eds.), Machine Learning. Tsinghua University Press, Beijing. 171-183.

[14] Chen, T., Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. The 22nd ACM SIGKDD International Conference. San Francisco. pp. 785-794.

[15] Dietterich, T. G. (2000) Ensemble Methods in Machine Learning. Proc International Workshgp on Multiple Classifier Systems, 1857:1-15.