Research on sales forecasting of fresh chicken based on ensemble learning for a broiler processing enterprise

Huiting Xia[a], Yu Cao*[a], Xu Cheng[b]

[a]School of Information and Control Engineering, Liaoning Petrochemical University, Fushun, Liaoning,China
[b]College of Economics and Management, Shenyang Agricultural University,Shenyang, Liaoning,China

* Corresponding author: caoyu@lnpu.edu.cn

Abstract. In order to ensure the freshness of fresh food and reduce the inventory of enterprises, enterprises need to predict the sales volume of tomorrow before ordering fresh chickens. The effect of the traditional model on sales forecast is not ideal, so a sales forecast based on stacking model is proposed. Random forest and XGBoost are used as basic learners, and the artificial neural network model is used as meta-learner. Compared with the single model, the overlay fusion model is superior to other methods, with an average absolute error of 10.97, a MAPE of 83.30%, and a determination coefficient of 0.7634, which is helpful for enterprise decision-making.

Keywords: Sales forecasting, Random forest, Artificial neural network, XGBoost, Stacking

1 Introduction

With the development of the social economy, people's requirements for quality of life are getting higher and higher. Pay more attention to the freshness of food when purchasing ingredients. Therefore, ensuring freshness has become the primary consideration for food processing enterprises.

The production process of poultry processing enterprises is generally as follows: The generation of real sales orders lags behind the processing and production links. Following their experience the previous day, the salesperson ordered a certain number of live birds and sent them to the processing workshop before the next day. Among the poultry production procedures are slaughters, splits, grades, packs, refrigerates / freezes, and then delivery to wholesalers and retailers in the evening. It is important to note that if poultry production is less than sales, some customers' needs cannot be met and this results in loss of profits and customers. Conversely, if poultry production exceeds sales, excess fresh meat is frozen for cryopreservation and sent to the cold storage. Since frozen meat is cheaper than fresh meat, and nutrients are lost in the process. Furthermore, frozen meat incurs additional costs in terms of storage, refrig-

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_14

eration, and management. Merchants will have to occupy more cold storage warehouses in order to store excess frozen poultry inventory. This will consume a lot of electricity, increases the pressure on the cash flow of enterprises, and reduces the profit margin. As per the statistics for the first half of 2020, the total area of cold storage in China exceeds 9,522,600 square meters, and the annual electricity consumption of cold storage is approximately 15 billion kilowatt-hours. In addition, the warehouse also needs cold storage workers to manage it, which will consume a lot of manpower, material resources and financial resources. Meanwhile, the backlog of stale fresh goods will lead to unsaleable and discarded fresh goods, which will negatively affect the company's reputation. Therefore, accurate prediction of sales volume is necessary for poultry processing enterprises.

There are two types of traditional sales forecasting methods: qualitative forecasting and quantitative forecasting. Qualitative prediction is subjective, and it is difficult to draw reliable conclusions from it [1]. Thus, most scholars use quantitative forecasting methods to study sales forecasting. In quantitative prediction, time series analysis, causal analysis, grey theory, artificial neural networks, and machine learning algorithms are most commonly used [2]. Literature [3] and others use the RF-XGBoost-LR model to forecast the sales of retail companies located in the United States, and the performance of the hybrid RF-XGBoost-LR model is better than that of the single model. Literature [4] puts forward that the best and worst method and the SalesKBR model with k-Means integrated into RFM can be used to predict retail sales and help enterprises make scientific and valuable management decisions. Literature [5] uses the neural network, CHAID, C&RT, Gen Lin, and Ensemble algorithms to forecast the sales of golf clothing brand "A", and the results show that the integrated model has the best forecasting ability. Literature [6] uses random forest, SVM, XGBoost, ANN and Stacking algorithms to predict food sales, the number of customers and the number of raw materials in Siddhartha restaurants. The results show that the superposition algorithm is superior to other algorithms. Literature [7] uses self-organizing mapping and K- means clustering method to group goods from 10 locations in different cities. Then the sales volume of each group is predicted by gradient lifting, decision tree, and generalized linear regression algorithm. The results show that the combination of K- means clustering and gradient lifting method is better.

To improve the accuracy of fresh sales forecasts, this paper proposes a stacking strategy based on random forest and XGBoost, which takes the artificial neural network as the final forecast model. Using the data of a broiler processing enterprise, the prediction was made and compared with the experimental results obtained from the random forest, XGBoost and artificial neural network model. The prediction model combines the advantages of random forest, XGBoost and artificial neural network model. It improves the forecast accuracy of fresh sales. Therefore, the model proposed in this paper is of great significance for the use of scientific methods to predict fresh chicken sales.

2 Methods

Stacking divides the sample training set into k subsets (Figure 1). The k-1 subset is taken as the training set, and it is brought into the first-level basic learner for training to predict the result of the remaining one. In this way, after K rounds, the prediction results of K rounds are spliced into new data sets, and the new data sets output by each base learner are spliced together as training sets, which are brought into the second layer meta-learner for training. The sample test set is predicted every round, and the average of the predicted results is taken as the new test set. Then it is brought into the second layer meta-learner for prediction.

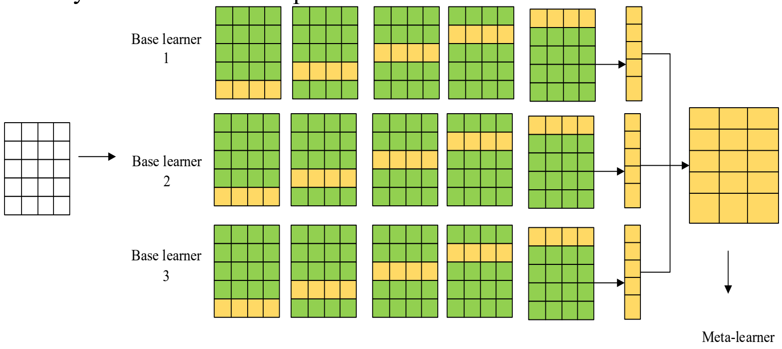


Fig. 1. Stacking model framework

In this paper, XGBoost, random forest is used as the base learner of the first layer, and the artificial neural network is used as the meta-learner of the second layer to model and analyze the data set to get the final prediction result.

(1) Divide the processed training set into 5 copies on average

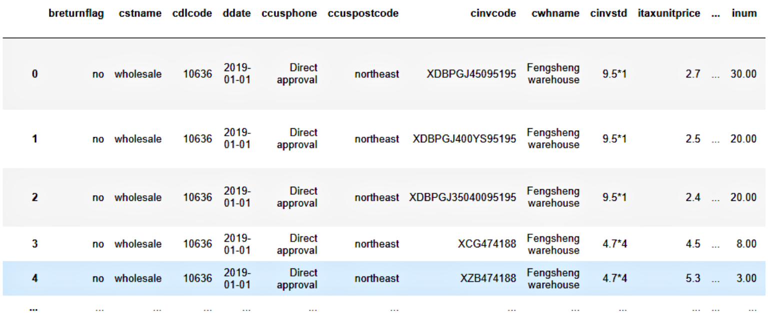
(2) 50% off cross-validation, taking four copies as a training set each time and bringing in the last one as a test set respectively. Go to the random forest, XGBoost, get two values equal to the data quantity of the training set, and splice them into the artificial neural network for training.

(3) Bring the test set into the random forest and use XGBoost to make predictions. Average the obtained results, and bring them into the artificial neural network for prediction.

3 Experiment and result analysis

3.1 Experimental data

The experimental data comes from a broiler processing enterprise (Figure 2). There are 95,117 historical fresh sales data from January 2019 to October 2019. The attributes mainly include goods number, delivery date, types of specifications, tax price, number of pieces, etc.



	breturnflag	cstname	cdlcode	ddate	ccusphone	ccuspostcode	cinvcode	cwhname	cinvstd	itaxunitprice	...	inum
0	no	wholesale	10636	2019-01-01	Direct approval	northeast	XDBPGJ45095195	Fengsheng warehouse	9.5*1	2.7	...	30.00
1	no	wholesale	10636	2019-01-01	Direct approval	northeast	XDBPGJ400YS95195	Fengsheng warehouse	9.5*1	2.5	...	20.00
2	no	wholesale	10636	2019-01-01	Direct approval	northeast	XDBPGJ35040095195	Fengsheng warehouse	9.5*1	2.4	...	20.00
3	no	wholesale	10636	2019-01-01	Direct approval	northeast	XCG474188	Fengsheng warehouse	4.7*4	4.5	...	8.00
4	no	wholesale	10636	2019-01-01	Direct approval	northeast	XZB474188	Fengsheng warehouse	4.7*4	5.3	...	3.00
...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 2. Original bill sales data table

3.2 Evaluation indicators

(1)Determinable Coefficient

The determinable coefficient R^2 is a statistic to measure the fitting degree between the regression line and the true value. The calculation formula is as follows:

$$\mathrm{R}^2 = 1\text{-}\frac{\sum (Y\_actual - Y\_predict)^2}{\sum (Y\_actual - Y\_mean)^2} \tag{1}$$

The larger the R^2, the better the fitting effect. The optimal value is 1.

(2)Mean Absolute Error

Mean Absolute Error (MAE) is the absolute error between the true value and the predicted value. The formula is as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} \left| y_i - \hat{y}_i \right| \tag{2}$$

(3)Mean Absolute Percentage Error

A MAPE of 0% indicates a perfect model and a MAPE greater than 100 % indicates a poor model. The formula is as follows:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{3}$$

3.3 Experimental procedure

In this paper, semi-rational methods are used to select features. First, the data irrelevant to sales forecasting is artificially eliminated. Second, we use the filtering method and the F-test method to screen out features that have a strong correlation with sales forecasts. Finally, it is input into the fusion model for sales forecasting.

(1)Data preprocessing. First, delete the fresh sales data of returned goods, and do not study such data. Next, fill in the missing value of fresh sales data. Then, the classification features are encoded using the embedding method. Finally, the characteristics and labels of fresh data are divided, and 'the number of pieces' is used as the label of prediction. Other attributes as features.

(2)Feature Selection. The semi-rational method is used to select features, and 12 attributes that are more important in sales analysis are screened out from 23 features. On this basis, the filtering method is used to screen out features with variance not equal to 0, and 11 attributes are screened out. Then, eight features are screened out using the F-test method, including month, day, access, customer abbreviation, customer address, goods number, specifications and models, unit price including tax.

(3)Dividing data and training model. Take 70% of the data as the training set and 30% as the test set. The training set and test set are brought into random forest, XGBoost, artificial neural network and stacking model for training and prediction, and the effect of the model is evaluated by MAE, MAPE, and R^2.

3.4 Experimental results

The training sets are brought to the random forest, XGBoost, artificial neural network for training, and then predict the test set. Random forest model, XGBoost model and artificial neural network model are average in predicting the sales of fresh chickens. On this basis, an improved superposition model based on random forest and XGBoost and artificial neural network as meta-learner is adopted to improve it. The summary results are shown in Table 1. The effect diagram of stacking predicted value and true value is shown in Figure 3.

Table 1. Forecast Index of Fresh Sales Based on Different Models

model	**Error parameter**		
	R^2	**MAE**	**MAPE**
Random forest	0.7553	11.08	89.77
XGBoost	0.7510	12.09	145.14
Artificial neural network	0.5927	17.90	408.68
Stacking	0.7634	10.97	83.30

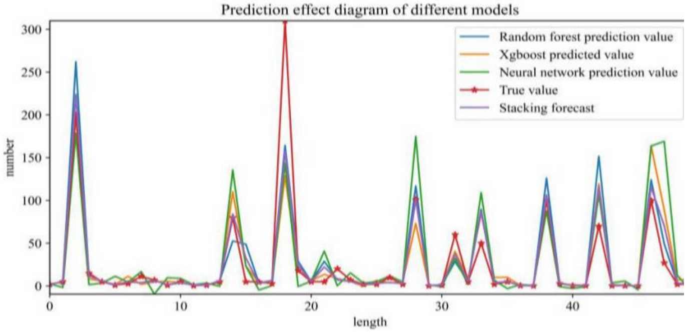


Fig. 3. Prediction effect diagram of different models

4 Conclusion

The performance of the improved stacking model in sales forecasting is obviously better than that of the single model, with the MAE reduced by 0.11, the MAPE reduced by 6.47% and the determinable coefficient reaching 0.7634. The overlay fusion algorithm combines the advantages of random forest and XGBoost, and introduces an artificial neural network model with good interpretability and robustness, which further improves the accuracy of the model. Therefore, the model integrates more accurate prediction of future sales data, which is very important for the development of business production.

However, the stacking model has a complex framework, especially following the integration of multiple models with massive data. It takes a long time to run, and it can be considered to run in a distributed computing environment in the future, so as to reduce its running time.

References

1. Li Xiaoran, Jiang Yi. Comparison and exploration of quantitative and qualitative analysis methods in sales forecast[J]. Commercial Accounting, 2017, (06): 65-67.
2. Zheng Jinfeng, Luo Ronglei. Research progress on quantitative forecast methods of clothing sales[J]. Advanced Textile Technology, 2022, 30(02): 27-35.
3. Mitra Arnab et al. A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach[J]. Operations Research Forum, 2022, 3(4)
4. Gustriansyah Rendra and Ermatita Ermatita and Rini Dian Palupi. An approach for sales forecasting[J]. Expert Systems With Applications, 2022, 207
5. Han Jixiang and Han, Ki Hyang. Using machine learning to predict women's golf clothing sales-centered on meteorological factors and day of the week -[J]. Doubles, 2021, 71(1)
6. K. Harshini, P. K. Madhira, S. Chaitra and G. P. Reddy, "Enhanced Demand Forecasting System For Food and Raw Materials Using Ensemble Learning," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9671005.
7. S. N. Gunjal, D. B. Kshirsagar, B. J. Dange and H. E. Khodke, "Fusing Clustering and Machine Learning Techniques for Big-Mart Sales Predication," 2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), 2022, pp. 1-6, doi: 10.1109/ICBDS53701.2022.9935906.