Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231335

Combinatorial Model-Based Demand Forecast Analysis of E-Commerce Agricultural Products

Jianglong LIU[a,1], Tianhao WU[b,2], Junhui WU[c,3], Zixi CHEN[a,4], Jue GONG[a,5], Hongyi CHI[a,6]
[a]*School of Electronic and Information Engineering Tongji University Shanghai, China*
[b]*College of Transportation Engineering Tongji University Shanghai, China*
[c]*National Engineering Research Center of Protected Agriculture Tongji University Shanghai, China*

Abstract. This paper constructs the e-commerce agricultural products feature data set by using technical means such as web crawler and carries out preprocessing operations such as data filling and cleaning. K-Means algorithms are utilized to perform cluster analysis on the e-commerce agricultural products feature data set to realize multi-dimensional and multi-angle summarization of sales factors. Taking the dataset as the data support of the prediction model and based on the limitations of a single prediction model in the scope of application and accuracy, ARIMA, PSO-LSTM model and other combined models are constructed by using the arithmetic weight method and the integrated learning method respectively. According to the weights calculated out in the different arithmetic methods, the weights of the model are assigned to synthesize the prediction results. Among them, the best prediction results were obtained from the combined model using the inverse error weighted average method to calculate the weight composition.

Keywords. Clustering, E-commerce produce, Predictive modeling, Data mining

1. Introduction

In the era of rapid development of information technology, the e-commerce industry is also ushering in a development boom. For e-commerce agricultural products, accurately grasping the specific characteristics and trends of online sales of agricultural products helps to gain a deeper understanding of the demand for agricultural products. The work

[1] Jianglong LIU, School of Electronic and Information Engineering Tongji University Shanghai, China; email: 2233022@tongji.edu.cn

[2] Tianhao WU, College of Transportation Engineering Tongji University Shanghai, China; email: 2253926@tongji.edu.cn

[3] Corresponding author: Junhui WU, National Engineering Research Center of Protected Agriculture Tongji University Shanghai, China; email: junhui_wu@163.com

[4] Zixi CHEN, School of Electronic and Information Engineering Tongji University Shanghai, China; email: 2897231674@qq.com

[5] Jue GONG, School of Electronic and Information Engineering Tongji University Shanghai, China; email: 874321116@qq.com

[6] Hongyi CHI, School of Electronic and Information Engineering Tongji University Shanghai, China; email: 2460327840@qq.com

of the current online agricultural products transaction information processing focuses on the analysis and prediction of a large amount of data to achieve an accurate grasp of the market in this field. Nowadays, many industries cannot be without the participation of prediction models[1]. The current prediction models can be divided into single model prediction and combined model prediction. Although the improvement of the single model will improve the accuracy, it cannot completely solve the defects of the model[2]. Therefore, today's scholars are more likely to choose to combine multiple models in an integrated learning approach to improve the adaptability and accuracy of the model.

Combined modelling refers to the use of different predictive models for the same research object and combining them according to certain weights or integrated learning methods. The theory of combinatorial prediction was first proposed by Bates and Granger[3], and combinatorial prediction has now been applied to the study of various prediction problems[4]. VanCalte et al[5] predicted the sales volume of Coca-Cola by using a combination of heuristic search and genetic algorithms, and the results proved that the combinatorial prediction method is more accurate; He Wei et al[6] predicted the sales volume of Coca-Cola based on the combination of XGBoost and LSTM neural network combination method to predict the sales of a supermarket. The results show that the predicted value is close to the true value; He Zhang et al[7] predicted the short-term sales of vegetables based on the combination of LGBM and LSTM neural network method; Kalagotla et al[8] fused three models, namely, MLP, SVM (Support Vector Machine) and LR, based on the Stacking fusion model to build a diabetes prediction model; Guannan Li et al[9] built a sensor fault detection based on improved Stacking combinatorial model, which greatly reduces the false alarm rate; Chengshi Tian et al[10] proposed a nonlinear combinatorial prediction system for short-term load forecasting and validation of half-hourly electricity load data in New South Wales; Luo Long and Li Lianghuan[11] combined time series and deep learning based on ARIMA and LSTM, which predicted both linear and nonlinear parts well and obtained low prediction errors; Feng Chen and Chen Zhide[12] et al. divided the object into linear and nonlinear parts, as the linear part is processed by ARIMA model, and the nonlinear part is processed by the weighted combination of XGBoost and LSTM, which proves that the prediction performance of the combination model is good; Ning Zhang[13] built a convenience store sales prediction system based on the combination of two neural network algorithms, NN and DBN, as well as SVR.

2. E-commerce Agricultural Product Feature Dataset Construction

Aspects of data access, we comprehensively considered e-commerce websites such as JD.com, Tmall, Pinduoduo and Taobao, and finally chose Taobao, which has the largest volume of e-commerce agricultural products, as the data crawling website. Based on Taobao, the data of 15 types of e-commerce agricultural products, such as grapefruit, oranges, apples and so on, were crawled in the past five years. When choosing the forecasting research object, we also considered the complexity and availability of data, as well as the sustainability, influence and seasonal available research value of agricultural products sold on e-commerce platforms. Finally, pomelo, whose sales account for 15% of the entire share of agricultural products in e-commerce, was finally chosen as the forecasting research object. A total of 186,743 data on various attribute characteristics of pomelo sales in the past 5 years were obtained through web crawlers and specific website queries.

Agricultural products as a special food category, its influence fluctuation factors are different from other products, and need to be considered in conjunction with the characteristics of agricultural products when conducting research. Fresh agricultural products are generally characterized by regional, cyclical, perishable quality, and variable demand, which are different from other items. So, it is necessary to integrate the unique attributes that agricultural products have when conducting research and analysis. It can be summarized as four major characteristics of agricultural products' own attribute characteristics, store characteristics, weather factor characteristics, and socio-economic and environmental factor characteristics, which cover a variety of aspects such as agricultural products' attributes, e-commerce platform characteristics, and social and environmental factors.

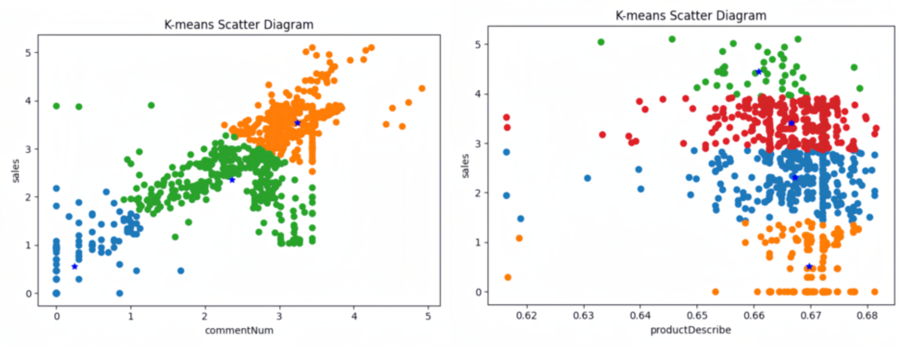
To solve the problems of noise, missing values, and data quality of the original data, it is necessary to pre-process the original data. In this study, the preprocessing of raw data for the original e-commerce agricultural products feature dataset raw data mainly include data validity screening, missing value processing, and normalization processing. Among them, the missing value processing selects the sliding average and LaGrange interpolation method for completing.

3. K-Means Cluster Analysis

Cluster analysis is a typical unsupervised learning method. Through the cluster analysis method of data mining exploration, it can produce the characteristics of the product, consumer demand, sales patterns and trends, market competition and many other useful information. The principle of cluster analysis is the process of dividing data into different subsets through static classification methods, in which data with similar nature characteristics are divided into a subset.

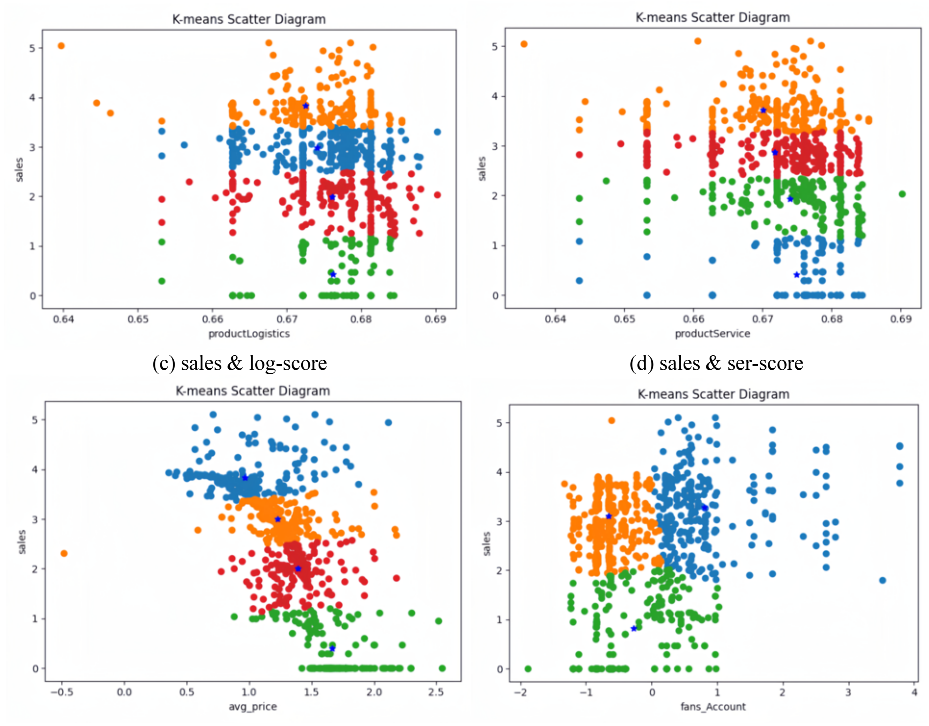
The core idea of K-means algorithm is to calculate the distance between all samples and the K proposed center points. For each sample data, the data point is divided into the representative cluster closest to the center point of k point, and then the mean value of all data points in each cluster is calculated as the new cluster center point until the optimal cluster center is found or the number of iterations is reached.

It based on the obtained dataset to do cluster analysis of the pomelo sales merchants’ characteristic attributes that are easy to adjust, uncontrollable factors for the time being. It explores the relationship between grapefruit sales and other attribute factors. Because the sales volume and some attribute factors have large values, log method is selected in the choice of normalization method, that is, taking the logarithm of the data.



(a) sales & comment

(b) sales & des-score



(c) sales & log-score

(d) sales & ser-score

(e) sales & price

(f) sales & fans-count

Figure 1. K-Means clustering plot of sales versus each factor.

As shown in Figure 1. the K-Means method is used to cluster sales with the number of reviews, product description score, product logistics score, product service score and price respectively. The results and analysis of the clustering show that if merchants want to increase product sales, they need to pay extra attention to the number of reviews of the product and the price and strive to make consumers shop and then comment on the product. And they also need to take the price factor into consideration for the development of the program under the premise of considering the service, description, and logistics. They are not supposed to blindly focus on the pursuit of the score, which may lead to a decline in customer traffic, and pay extra attention to the description score of the product in the three aspects of the product, the Consumers are more concerned about the product itself. At the same time, it is also necessary to build the store's own brand, observing the high fan volume of the store, it can be learned that the goods are local specialties. So, merchants should also take the local characteristics into consideration.

4. Predictive Modelling Scheme Based on Combination of Adjustment Weights

4.1 Formatting Author Affiliations

This paper proposes the adjustment weight method for single model combination, the flow chart of this method is shown in Figure 2. This flow uses ARIMA, PSO-LSTM and improved Random Forest model, utilizes the arithmetic weight assignment method to construct the combination model and finally predicts the e-commerce agricultural products data and calculates the error. This process combines the methods of time series analysis, neural networks, and random forests to each the final prediction results through the assignment of weights.

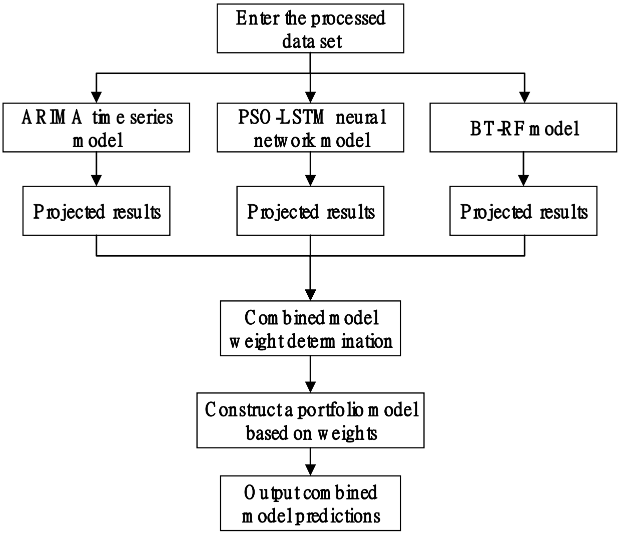

Enter the processed data set
ARIMA time series model
PSO-LSTM neural network model
BT-RF model
Projected results
Projected results
Projected results
Combined model weight determination
Construct a portfolio model based on weights
Output combined model predictions

Figure 2. Adjustment weight combination modeling process

4.2 Adjustment Weight Combination Model Results

The key to the construction process of the combined model is the selection of the weights of a single model, to ensure that the appropriate weights can be taken to improve the adaptability of the model. The number of models in the equal weight method is 3, so each model weight is 1/3. The weights obtained by the error variance weighted average method are obtained by calculating the variance, which is ranked according to the variance from high to low for the time series model, the LSTM neural network model, and the random forest model, respectively. According to the ranking can be derived from the error variance weighted average method of the weights obtained: ARIMA take 1/6, LSTM take 1/3, RF take 1/2; error inverse weighted average method to RMSE error as an indicator, according to the three models of the EMSE error results to take the inverse of the average can be derived from the weights accounted for by the different models. The ARIMA model is 0.32, the LSTM model is 0.33, and the RF model is 0.35.

According to the weights obtained above, the predicted values are brought into different prediction models according to the weight calculation to get the predicted values of the combined model, and then the prediction results are compared and analysed by the four evaluation indexes, namely, RMSE, MAE, MAPE, and SUM_Re. The SUM_Re error index refers to the sum of the relative errors between the predicted values and the real values in the next seven days. The errors of the combined model obtained by calculating these three arithmetic weighting methods according to the error indicators are shown in TABLE 1.

Table 1 Error results of comparing the real value of sales with the forecast value for the next seven days

Evaluation Indicators	Equal Weighting Method	Error Variance Weighted Average	Inverse Error Weighted Average Method
RMSE	13.623	12.113	7.8813
MAE	10.839	9.596	5.974
MAPE	0.0032	0.0020	0.0013
SUM_Re	0.0189	0.0146	0.0091

The comparison of the error indicators derived from the above table clearly shows that the best performance among the three arithmetic weight allocation methods is the inverse error weighted average method, whose error is smaller than that derived from the other two methods, indicating that the error between the real and predicted values is minimized and the prediction effect performs better.

5. Stacking-based Integrated Learning Predictive Modelling Program

5.1 Integrated Learning Modeling Process

Stacking method usually contains two layers. In this paper, in constructing the Stacking integrated learning model, ARIMA, Improved Random Forest Model and PSO-LSTM model are selected in the initial model of the first layer. The first layer models are also parameter optimized to be able to reduce the error as much as possible, the ARIMA model can have a good extraction of linear relationships, the random forest model can enhance the model generalization ability, and the neural network model can reduce the data correlation. Therefore, these three models were used as the original models for the first layer. For the second layer of meta-models, a simple linear regression model was selected, and these two layers of models constructed the Stacking model. The fusion flowchart of the Stacking model is shown in Figure 3.

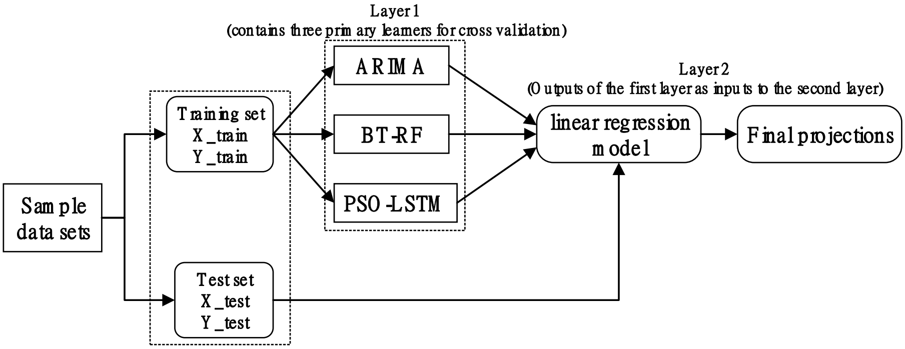


Figure 3. Flowchart of fusion based on Stacking model

The pre-processed dataset was used to randomly select data samples using a random number generator, of which 70% was used as a training set and 30% as a test set. In the process of Stacking model construction, the data set is firstly divided by using the tri-fold crossover method. After that, the first layer base learner is trained and learned by using the tri-fold crossover method to get the new training set and test set of the second layer input, and the new feature vectors obtained are fed into the meta-learner of the second layer to continue the training, and finally the constructed Stacking model is validated and evaluated.

5.2 Integrated Learning Model Results

It can be obtained from the above learning of Stacking model that the main idea of Stacking model is to combine some prediction performance situations that have low impact on the fluctuation of product sales and to utilize the combined effect of its combined prediction ability of poorer models to improve the expected effect on the increase and decrease of sales. After the implementation of the Stacking fusion model, the test set prediction result curves were obtained as shown in Figure 4.

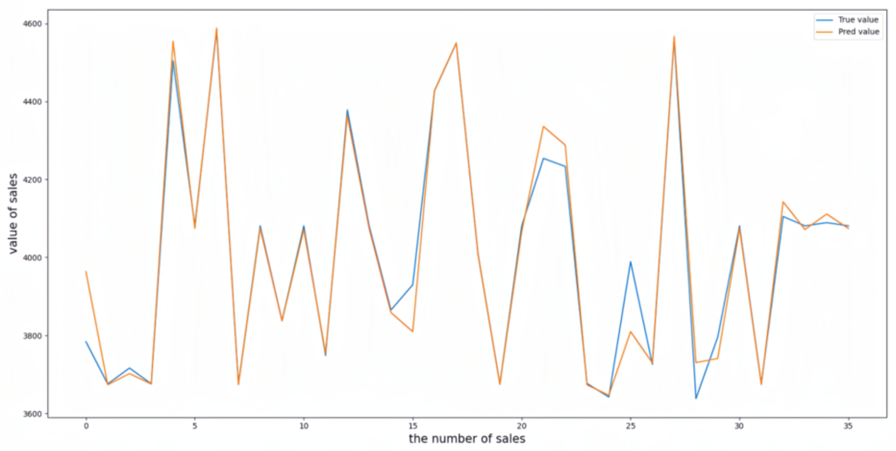


Figure 4. Integrated Learning Prediction Curve

Based on the prediction accuracy, the model score is calculated as: 0.864, which the score proves that the model has a good fit and can process and analyse the dataset well. The error results of the integrated learning model are compared by different error evaluation criteria RMSE, MAE and MAPE. The error results obtained from different models are demonstrated as shown in TABLE 2.

Table 2 Error Results for Comparison of Data Set Test Set

Model	RMSE	MAE	MAPE
PSO-LSTM	**44.971**	**34.326**	**0.0084**
LSTM	**59.541**	**53.692**	**0.0129**
RF	**44.424**	**26.895**	**0.0067**
GD-RF	**40.178**	**24.040**	**0.006**
BT-RF	**38.325**	**21.524**	**0.0051**
Stacking（pomelo）	**33.254**	**15.731**	**0.0039**

From TABLE 2. for the grapefruit dataset, the error between the predicted value and the true value are derived from model combination by using Stacking integration method. The root mean square error (RMSE) is 33.254, the mean absolute error (MAE) is 15.731, and the mean absolute percentage error (MAPE) is 0.39%; from the perspective of the three error indicators, all of them are improved compared to the accuracy of a single model, and compared to the single model with the best performance of the Compared with the BT-RF model, which is the best performing single model, the

RMSE error is lower than 5.071, the MAE is lower than 5.793, and the MAPE is 0.12%, which proves that adopting the integrated learning method can effectively reduce the prediction error and improve the prediction accuracy. It also indicates that the Stacking integrated learning method can merge the advantages of multiple base learners with better prediction robustness. And the K-fold cross-validation method can effectively improve the model generalization ability.

6. Conclusion

In this paper, we construct e-commerce agricultural products feature data set by using technical means such as web crawler, K-Means algorithm to perform clustering analysis after data preprocessing and finally summarize the sales factors from multi-dimensional and multi-angle perspectives. At the same time, ARIMA, PSO-LSTM model and other combination models are constructed by using arithmetic weighting method and integrated learning method, and weights are assigned to the models according to the weights derived from different arithmetic methods. According to the prediction results, it is concluded that the combination model composed of weights by using the inverse weighted average of error method has the best performance effect. In addition, the integrated learning of the three models is carried out to build a combined prediction model based on Stacking fusion, which proves that the use of the combined model method can effectively improve the accuracy of the model. Moreover, after analysing and comparing the two combination models, this paper concludes that the ensemble learning for combination can predict the demand of agricultural products more accurately while considering the advantages of different single prediction models.

Acknowledgments

Development and demonstration of online service trading platform for agricultural materials and agricultural products. (Grant Number: 2020YFD1100603))

References

[1] Zeng Yu. Research on inventory decision-making of e-commerce retail enterprises based on improved random forest prediction [D]. Sichuan:Southwest Jiaotong University,2020.
[2] Ferreira K J, Lee B H A, Simchi-Levi D. Analytics for an online retailer: Demand forecasting and price optimization[J]. Manufacturing & service operations management, 2016, 18(1): 69-88.
[3] Bates J M, Granger C. The combination of forecasts[J],Oper. Res.Q. 2001(1969):451- 468.
[4] Qian Huilan. Merchandise sales portfolio forecasting based on MI algorithm [D]. Nanjing University,2019.
[5] Van Calster T, Baesens B, Lemahieu W. ProfARIMA: A profit-driven order identification algorithm for ARIMA models in sales forecasting[J]. Applied Soft Computing, 2017, 60: 775-785.
[6] Wei H, Zeng Q T. Research on sales Forecast based on XGBoost-LSTM algorithm Model[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1754(1): 012191.
[7] He Z, Yu S. Application of LightGBM and LSTM combined model in vegetable sales forecast[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1693(1): 012110.
[8] Kalagotla S K, Gangashetty S V, Giridhar K. A novel stacking technique for prediction of diabetes[J]. Computers in Biology and Medicine, 2021, 135: 104554.

[9] Li G, Zheng Y, Liu J, et al. An improved stacking ensemble learning-based sensor fault detection method for building energy systems using fault-discrimination information[J]. Journal of Building Engineering, 2021, 43: 102812.

[10] Tian C, Hao Y. A novel nonlinear combined forecasting system for short-term load forecasting[J]. Energies, 2018, 11(4): 712.

[11] LUO Long, LI Lianghuan, WANG Chengyang, LU Peidong, YANG Peishi. An insulator state data mining method based on ARIMA-LSTM[J]. Journal of Electric Power Science and Technology,2017,32(04):38-43.

[12] Feng Chen,Chen Zhide. Application of weighted combination model based on XGBoost and LSTM for sales forecasting[J]. Computer System Applications,2019,28(10):226-232.

[13] Zhang N. Research and application of deep learning-based sales prediction for chain convenience stores [D]. Beijing University of Technology,2019.