Multi-Algorithm Fusion Pharmaceutical Sales Forecasting Mode

Huanhuan Jiang[1,a], Yue Fan[2,b], Haoyuan Sun[3,c], Shiqiang Liu[4,d]
[1]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shenyang, China
2Shenyang Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shenyang, China
[3]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shenyang, China
[4]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shenyang, China
[a*] Corresponding author: jianghuanhuan19@mails.ucas.ac.cn
[b]e-mail: fanyue2021@126.com
[c]e-mail: sunhaoyuan19@mails.ucas.ac.cn
[d]e-mail: liushiqiang19@mails.ucas.ac.cn

Abstract

In recent years, China's pharmaceutical e-commerce industry has developed rapidly, and now the industry has entered a critical stage of development. However, when the merchants stationed in the pharmaceutical platform, there are still many problems with the recognition of the pharmaceutical e-commerce and the positioning of the sales target. Many businesses are facing the risk of continuous loss and bankruptcy in the increasingly fierce competition due to their vague positioning and unclear strategic layout. Therefore, this paper mainly analyzes the sales data of merchants staying on the medical platform, and then makes sales forecasts. The sales data involves the number of new users on the day, so first use the time series model ARIMA to predict the number of new users, and then the predicted number of new users is put into the data to predict sales. Time series models ARIMA, LSTM, XGBoost are used as the base models and LightGBM is used as the final prediction model to predict merchant sales using Stacking strategies. This method concentrates the advantages of each model, greatly improves the accuracy of sales forecasts for pharmaceutical merchants, and is of great significance to merchants' decision-making.

1 Introduction

China's huge population has brought huge medical needs, and China's medical and health expenditures are large and maintain steady growth. With the in-depth integration of Internet technology and the medical and health industry, pharmaceutical e-commerce service, as one of the most promising business models of "Internet + Medical Health", is gradually developing into a new normal of consumption. By 2020, the transaction scale of China's pharmaceutical e-commerce market has reached 195.6 billion yuan. Especially during the epidemic period, pharmaceutical e-commerce platforms also provide important medical security for Chinese citizens, and the pharmaceutical e-commerce market has great development potential in China. How to obtain important information containing business opportunities from the huge sales data of pharmaceutical e-commerce so as to make commercial inference beneficial to the merchants has become an important problem to be solved urgently in the research of pharmaceutical e-commerce.

At present, most researches on sales forecast at home and abroad are based on traditional statistical forecasting methods and machine learning forecasting methods. But traditional forecasting methods have poor adaptive ability and low accuracy of forecasting results. With the development of machine learning, problems such as the adaptability and accuracy of prediction methods have also been well resolved, for example, Zheng Yan, Huang Xing [1], et al. based on the time series of commodity forecasting model research and Ge Na, Sun Lianying [2] based on the ARIMA time series model of sales forecast analysis have achieved good results in forecasting sales. In recent years, deep learning has gradually risen. Using deep learning algorithms to integrate models, the prediction effect is generally better, for example, Wang Rongzheng and Liao Xianyi [3] put forward the prediction of blood glucose based on the integrated learning fusion model. By comparing the evaluation indicators, it is found that the error rate of the integrated fusion model is lower than linear regression, random forest and gradient boosting book, and the prediction effect of the fusion model is the best. Wang Hui and Li Changgang [4] put forward that multi-model fusion has better generalization ability and predictive ability than single model in the application of Stacking integrated learning method in sales forecasting.

A large number of literature studies have shown that the fusion of traditional machine learning models and deep learning algorithm models, or the fusion of more than two deep learning models, usually has higher accuracy and better performance than a single algorithm. Therefore, this article also uses the fusion of multiple algorithms to predict the medical sales data.

ISBN 978-3-8007-5739-8

2 Algorithm Model

2.1 ARIMA Algorithm

First, Autoregressive moving average model ARIMA [5] is a time series analysis and forecasting method. The idea of the ARIMA model is that the current time series value has a linear relationship with the past time series value and the amount of external interference, which is suitable for various time series data. The ARIMA model assumes a stationary time series. If the time series does not meet the stationarity condition, it needs to be converted into a stationary time series by difference before using the ARIMA model. ARIMA model (p, d, q) is a combination of moving average model, autoregressive model and difference method [6]. As shown in Formula (1):

$$y_t = c + \emptyset_1 \cdot y_{t-1} + L + \emptyset_p \cdot y_{t-p} + \xi_t - \theta_1 \cdot \xi_{t-1} - L - \theta_q \cdot \xi_{t-q} \quad (1)$$

Where p is the autoregressive order, q is the moving average order, φ_1 ,… φ_p are the autoregressive coefficients, θ_1,…θ_q are the moving average coefficients, and $\{\xi\}$ is the white noise sequence.

2.2 LightGBM Algorithm

The LightGBM model is an ensemble learning algorithm model based on Gradient Facilitated Decision Tree (GBDT) proposed by Microsoft [7] in 2017. The LightGBM model is optimized in the original GBDT algorithm to solve the problem of GBDT time-consuming and low generalization ability. The traditional algorithm based on GBDT such as XGBoost [8] uses a pre-sorting algorithm. First, all node feature values are pre-sorted, and the best segmentation point on the feature is found by traversing all data sets, but when the amount of data is large, this will greatly increase the time-consuming. Compared with the XGBoost algorithm, the LightGBM algorithm uses histogram algorithm (Histogram), leaf-wise leaf growth strategy with depth limitation, and gradient-based unilateral sampling and mutually exclusive feature bundling algorithm ideas, while ensuring efficiency overfitting.

2.3 Stacking Algorithm

Stacking is an ensemble learning machine learning method proposed by Wolpert [9] in 1992. Stacking algorithm is a combination of several learners. The individual learner trained in the first layer is the primary learner, and the classifier that relearns the predicted value generated by the primary classifier as the input feature is called the secondary learner. The training set of the secondary classifier of the Stacking algorithm is generated by the primary learner. If the prediction results of the primary learner are used directly, the risk of overfitting may occur. K-fold cross-validation is used to predict the primary learner, which not only improves the generalization ability of the model, but also prevents overfitting.

The process of five fold cross validation is shown in Figure 1 below.

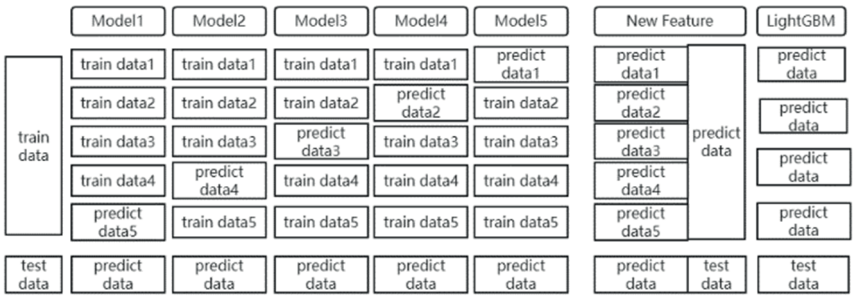

Figure 1 Process of five fold cross validation.

The process is as follows:

- Randomly divide the training set after data processing into 5 parts of equal data size. For each model, the 4 sub-sets are used as the training set in turn, and the remaining 1 is the prediction set.
- Each model trains the data separately and outputs the prediction result of its own prediction set as a new feature of the training set.
- Each model makes predictions on the test data, and averages the 5 columns of results it produces to obtain a new feature of the test set.
- Select the LSTM, XGBoost model in turn, then repeat the above four steps.
- Finally, use LightGBM as the final model to train the training set generated in the third step and the new feature fusion.
- Make predictions on the test set generated in the fourth step, and get the final prediction result.

2.4 Fusion Algorithm Framework

This article chooses time series models ARIMA, LSTM, and XGBoost as the first layer of primary learners. ARIMA is a time series model that can predict linear changes well, and the data in this article is based on time-varying data. LSTM is a kind of recurrent neural network, which mainly solves the phenomenon of gradient disappearance and gradient explosion in the back propagation process, and is widely used in the field of business intelligence [10]. The

ISBN 978-3-8007-5739-8

XGBoost is a Boosting integrated learning method based on decision trees, which has achieved good results in actual application scenarios. The secondary learner must ensure high generalization ability and correct the deviation of each algorithm, so the LightGBM algorithm is used as the final model. The algorithm framework is shown in Figure 2:

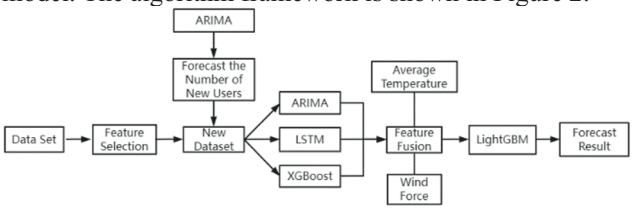

Figure 2 Algorithm framework diagram

The specific training process is as follows:

- Clean missing values and abnormal values of the data.
- Perform Pearson correlation coefficient calculation on the features in the data set and delete redundant features.
- Use time series models ARIMA, LSTM, and XGBoost as primary learners to train feature vectors.
- Feature fusion of the average temperature, wind, and the output of the primary learner, as the input of the final LightGBM model.
- Finally, use LightGBM to train and predict the data.

2.5 Feature Selection

Feature selection is the process of automatically selecting the most important feature subsets for the problem. Commonly used methods for feature selection include Filter, Wrapper, and Embedded. Pearson's correlation coefficient filtering method is one of the simplest methods that can reflect the linear correlation between features and predicted values. The value range of the result is [-1, +1], -1 means a complete negative correlation between the feature and the predicted value, +1 means a complete positive correlation, and 0 means no linear correlation. The calculation formula is shown in Formula (2):

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}} \qquad (2)$$

Where $\bar{x}$,$\bar{y}$is the mean value of the elements in each vector.

3 Experiment and Analysis

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

3.1 Data Description

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations. The data set used in this article is the actual sales data of a medical e-commerce company. This data is the sales data of a certain merchant in 2020. It contains a total of 248 drugs, 283952 pieces of data, involving 16 attributes, namely: product number, transaction amount, number of orders, number of payment users, number of products, unit price, number of new user, number of old user, time, maximum temperature, minimum temperature, average temperature, daytime weather, nighttime weather, wind direction, wind force, etc. The data attribute format is shown in Table I.

Table I Data Attribute Table

Name	Describe
product_id	product number
transaction_money	transaction amount
order_number	number of orders
paid_user_number	number of payment users
product_number	number of products
user_unit_price	unit price
new_user_number	number of new user
old_user_number	number of old user
order_time	time
high_temperature	maximum temperature
low_temperature	minimum temperature
average_temperature	average temperature
day_weather	daytime weather
night_weather	nighttime weather
wind_direction	wind direction
wind_power	wind force

3.2 Data Processing

- Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.
- Some products in the data are not sold on the day, and the sales data is empty, and 0 is used for filling.
- Involving the outliers in the sales data, use the mean to fill in.
- Add the sales data of the goods on the same day, and remove the characteristics such as the customer unit price and the product number.

ISBN 978-3-8007-5739-8

3.3 Experiment Procedure

Use Pearson correlation coefficient to filter out relatively independent features, delete redundant features, and finally use attributes such as transaction amount, number of orders, number of products, number of new customers, number of old customers, time, average temperature, and wind force.

3.4 Feature Selection

3.4.1 Evaluation Index:

RMSE root mean square error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

MAPE mean absolute percentage error:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{(\hat{y}_i - y_i)^2}{y_i}\right| \quad (4)$$

Where n is the number of samples, i is the i-th sample, y is the true value of sample sales, $\hat{y}$ is the predicted value of sales.

3.4.2 Forecast of the Number of New Users

The sales data for 2020 is divided into a training set and a forecast set, and the time series model ARIMA is used for prediction, Use pandas to draw a new user sequence diagram from January to December 2020 as shown in Figure 3:

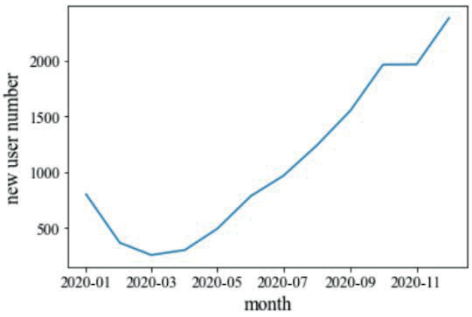

Figure 3 Number of new users trend.

Figure 4

As shown in the figure above, the number of new users of this merchant shows a trend of first decline and then a significant increase. Therefore, the sequence is a non-stationary sequence, and the data needs to be first-differentiated to obtain a stationary sequence. The first-order and second-order difference diagrams are shown in Figure 4:

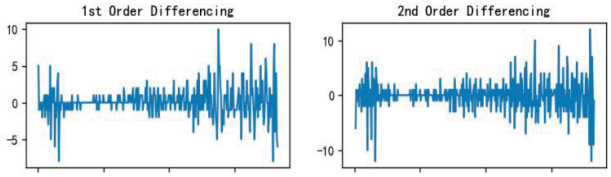

Figure 5 Number of new users trend.

me series is close to a stationary series when it is first-order difference, and the d value is 1. According to ACF, PACF, and information criterion function method, the order of the model is determined. Finally, when p is 8 and q is 1, the residual sum of squares is the smallest.

3.4.3 Fusion Model Prediction

After dividing the data into a test set and a training set, put the predicted number of new users in the test set of the data, use ARIMA, LSTM, and XGBoost as the primary learner for training and testing, and finally use LightGBM as the final model for the primary learner The training set generated in the fusion of the new feature is trained, and the test set is tested.

3.5 Result Analysis

The prediction results of each model are compared using RMSE and MAPE indicators, and the results are shown in Table II below.

Table II MSE, MAPE results Table

Model	RMSE	MAPE%
ARIMA	153.63	5.68
LSTM	112.75	5.19
XGBoost	127.61	5.42
Stacking	93.82	4.87

It is found that the stacking-based integrated model can enhance the ability of the primary model in sales forecasting, and can integrate the advantages of each model. Shows stronger generalization ability and predictive ability.

4 Conclusion

In order to solve the problem of a single model's poor predictability and generalization ability for sample data, this paper proposes to use a multi-algorithm fusion model to effectively predict the sales of pharmaceutical businesses. Using stacking to integrate algorithms such as ARIMA, LSTM, XGBoost, LightGBM, etc, reduces the average percentage error of the single model ARIMA and LSTM by 0.81% and 0.32%, and can predict sales more accurately. But the Stacking model framework is more complicated, and the training time is longer after combining each model. In the future, distributed computing can be used to further optimize this.

References

[1] Zheng Yan, Huang Xing, Xiao Yujie, Research on commodity demand forecasting model based on time

ISBN 978-3-8007-5739-8

series Chongqing University of Technology (Natural Science), 2019, 33(9): 217-222.

[2] GE N,SUN L Y.,ZHAO P,et al. Sales forecast analysis based on ARIMA time series model [J]. Journal of Beijing Union University:Natural Science, 2018, 32(4):27-33

[3] Wang Rongzheng,Liao Xianyi,Chen Xiangping,et al.Blood Glucose Prediction Based on Integrated Learning Fusion Model [J].Journal of Medical Informatics, 2019,40(1):59-62

[4] Wang Hui, Li Changgang. The application of Stacking integrated learning method in sales forecasting[J]. Computer Applications and Software, 2020, 37(08): 85-90.

[5] LIS,LI R. Comparison of forecasting energy consumption in Shandong, China using the ARIMA model, GM model,and ARIMA-GM model[J]. Sustainability, 2017, 9(7):1181.

[6] Yuan Yuan, Guo Tiantian. Research on Sales Forecast of ARIMA-RF Combination Model[J]. Software Guide, 2021, 20(09): 33-38.

[7] KE G,MENG O,FINLEY T,et al.LightGBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 2017 Annual Conference on Neural Information Processing Systems.New York:Curran Associates Inc,2017 :3146-3154.

[8] CHEN T,GUESTRIN C.XGBoost:a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NewYork: ACM, 2016: 785-794.

[9] Wolpert D H.Stacked Generalization [J].Neural Networks, 1992,5(2):241-259.

[10] Qiu Jun, Zhang Ruilin. Application of recurrent neural network based on genetic algorithm in sales forecasting[J]. Journal of Zhejiang Sci-Tech University, 2007(03):266-270

ISBN 978-3-8007-5739-8