Product order-demand prediction model based on random forest

Hao Wang*, #, He Zhang#, Jia Zhao, Xinyi Liu, Xinyue Feng, Yinuo Sun

Institute of Finance and Economics, Qinghai University, Xining, Qinghai, 810016

*Corresponding author: 15855659939@139.com

#These authors contributed equally.

Abstract. This study aims to develop a decision support model based on product order data analysis and demand forecasting. By analyzing the shipment data of a large manufacturing enterprise from September 2015 to December 2018, we establish an accurate prediction model for the demand in the next three months of a large manufacturing enterprise. Quarterly and monthly variables capture trends and seasonal variation by adjusting hyperparameters and cross-validation using a random forest algorithm. The results show that the mean absolute error (MAE) on the test set is 8.965, the root mean square error (RMSE) is 11.369, the relative mean absolute error (MAPE) is 8.256%, and the coefficient of determination (R^2) is 0.826. These indicators confirm that the model can accurately predict the target variable, with little difference from the true value, and show good predictive power and fit. The monthly model has high accuracy and stability and can effectively support production and supply chain planning to meet future needs. This study confirms the potential of product order data analysis and demand prediction models to improve the efficiency and competitiveness of enterprises and provides a valuable reference for the research and practice in related fields.

Keywords: random forest; correlation analysis; difference analysis.

1. Introduction

In recent years, the rapid development of the global economy and the intensification of market competition, enterprises are facing many challenges and opportunities, accurate prediction of product order demand has become one of the key factors for the success of enterprises [1]. Improving the accuracy of order demand forecast is conducive to the scientific formulation of order demand plan, can guide the ordering of raw materials or commodities, reduce the impact of business fluctuations, improve the company's internal sales, procurement, financial budget and other scientific decisions, reduce the enterprise inventory cost [2].

At present, many research institutions and enterprises are beginning to try to improve the accuracy and stability of demand prediction [3] by analyzing large-scale data sets, mining their hidden patterns and associations, and constructing prediction models. However, due to the market changes, the diversity of consumer behavior and the complexity of the supply chain and other influencing factors, the order demand prediction still has certain difficulties and challenges, and the existing methods often fail to fully capture the complex market situation and the diversity of consumer behavior, and the data quality and accessibility need to be improved.

Based on this background, this study will use the machine learning method to analyze the data set, and innovatively use the random forest regression model to predict [4] from multiple dimensions. Through in-depth exploration and empirical analysis, this study aims to provide reliable and innovative order demand prediction solutions for enterprises and promote the sustainable development of enterprises.

2. Model establishment and solution

2.1. Basic principle of the regression model for the random forest

Random forest regression model is a commonly used ensemble learning method with excellent prediction performance and strong generalization ability. It solves the regression problem by building multiple decision tree models and integrating their results. Specifically, the decision tree is a prediction model based on the tree structure and predicts by recursively dividing the feature space. During construction, the best features and cut points are selected to divide the dataset. In order to increase the model diversity and avoid overfitting, the random forest introduces two kinds of randomness: drop-back sampling and random selection of partial features. This reduces the model variance and better adapts to different data distributions.

Random forest obtains the final prediction result through ensemble learning, and in the regression problem, the predicted value of each decision tree is averaged or weighted average. This ensemble approach can reduce model bias and improve the accuracy and stability of prediction [5]. In addition, the random forest can also assess the feature importance, as measured by calculating the accuracy improvement of each feature in dividing the nodes of the decision tree. This is very important for understanding data and feature selection, and can help researchers find the most informative features.

In conclusion, the random forest regression model is a powerful prediction tool for a variety of regression problems. It has the advantages of handling nonlinear relationship, generalization ability, overfitting resistance ability, and evaluating the importance of features [6]. In scientific research, the application of random forest regression model can provide reliable and accurate results for experimental data analysis, prediction modeling, etc.

The prediction process of random forest is as follows: assume that there are common N samples and M feature attributes. N samples were randomly sampled from the original dataset as the training set of the decision tree. Meanwhile, a part of the features are randomly selected from M feature attributes to build a decision tree. The randomness here guarantees the diversity between the different trees. N samples are randomly sampled from the original data set with a random set of features from M feature properties to build a decision tree. The randomness here guarantees the diversity between the different trees. Multiple decision trees were subsequently constructed. When a prediction is required, the sample to be predicted is fed into each decision tree constructed to obtain the respective prediction results. By combining the prediction results of multiple decision trees, we can use the average value as the final prediction result of the regression problem, or use the voting method to determine the final prediction result of the classification problem.

2.2. Structure and determination of the product order demand prediction model based on random forest

The random forest-based product order demand prediction model is a commonly used data analysis and prediction method, widely used in areas such as supply chain management and production planning. The model cleans, sorts and transforms historical order data to eliminate noise, outliers and missing data and improve the consistency and comparability of the data. Features related to the order requirements were selected by the feature selection method and used to construct the random forest model [7]. In addition, appropriate feature selection can improve the predictive ability and interpretability of the model and reduce the risk of overfitting.

The random forest model integrates multiple decision trees, introduces randomness, and each decision tree is trained based on a different subset of data and randomly selects features to reduce the variance of the model and improve the generalization ability. The final prediction comes from the composite vote or average of all decision trees. Model evaluation is an important link to verify and evaluate model performance. Commonused evaluation methods include cross validation, error analysis and index evaluation. By evaluating the stability and generalization ability of the model through cross-validation, the error analysis can help to identify the source of the model prediction error, which can guide the improvement of the model. Business indicators such as average absolute

percentage error (MAPE) and root mean square error (RMSE) can be used to evaluate model performance. After using the established random forest model to predict the product order demand, enterprises can formulate corresponding production plans and supply chain management strategies to arrange reasonable production resources and inventory to meet the market demand and reduce costs. In addition, models can help companies identify potential sales opportunities and risks, and optimize product pricing and marketing strategies.

In conclusion, the random forest-based product order demand prediction model is a powerful data analysis tool to provide accurate demand prediction for enterprises, thus supporting decision making and supply chain management optimization. It plays an important role in improving enterprise operational efficiency and customer satisfaction, and achieving sustainable development.

2.3. Evaluation method

Random forest regression evaluated the model based on the MSE, RMSE, MAE, MAPE, and R^2 metrics.

Mean square error (MSE) is the square of the average difference between actual and predicted values. Its calculation formula is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

y_i $\hat{y}_i$ n Where, is the actual value, is the predicted value, is the number of samples.

The root mean square error (RMSE) represents the mean difference between the actual and predicted values. Its calculation formula is as follows:

$$RMSE = \sqrt{MSE} \tag{2}$$

The mean absolute error (MAE) is the absolute value used to measure the mean difference between the actual value and the predicted value. Its calculation formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3}$$

The mean absolute percentage error (MAPE) is the percentage of prediction error relative to the actual value. Its calculation formula is as follows:

$$MAPE = \sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times \frac{100}{n} \tag{4}$$

The coefficient of determination (R^2) is used to measure the interpretation of the variability of the observed values, ranging from 0 to 1. The closer to 1, the better the fitting ability of the model. Its calculation formula is as follows:

$$R^2 = 1 - (\frac{SSR}{SST}) \tag{5}$$

2.4. Analysis steps

This study followed data analysis. First, the training set data are trained to build an efficient random forest regression model and obtain detailed model parameter results. Subsequently, the importance of the features was studied in depth, assessing the extent to which each feature contributed to the model prediction outcome to determine its significance. This process provides valuable information about the importance of the features. Subsequently, the constructed random forest regression model was applied to the training set and test set data, and multiple operations were performed to ensure the reliability of the results. Finally, the model performance is comprehensively

evaluated by evaluating the prediction accuracy of the test data and provides a basis for further research and decision.

In order to improve the prediction efficiency and eliminate the influence of randomness on the results, it is recommended to save the random forest model obtained by the current training. When data prediction is needed, we only need to input the data to be predicted into the saved model to obtain accurate and reliable prediction results [8]. The data used in this study are derived from the forecast data shipped to distributors by a large domestic manufacturing enterprise from September 1,2015 to December 20,2018, as shown in Table 1. The data of the enterprise's shipments to dealers from September 1,2015 to December 20,2018 reflects the price and demand of the enterprise's products in different sales regions. Specifically, the following contents: order date, sales area code, product code, product category code, product category code, sales channel name, product price and order demand.

Table 1: Shipping data

Order date	Sales region code	Item code	First cate code	Second cate code	Sales chan name	Item price	Ord qty
2015/9/1	104	22069	307	403	offline	1114	19
2015/9/1	104	20028	301	405	offline	1012	12

Where "order date" is the date of a demand; a "product category code" will correspond to multiple "product category codes"; "sales channel name" is divided into online and offline, "online" refers to e-commerce platforms such as Taobao and Jingdong, and "offline" refers to offline entity dealers. The forecast data provides the sales area code, product codes, product categories and product categories of the products to be predicted, as shown in Table 2.

Table 2: Data samples of the product required to be predicted

Sales region	Item code	first cate code	Second cate code
101	20002	303	406
101	20003	301	405

3. Results

3.1. Model parameters and training duration

To evaluate the model performance and validate its effectiveness, a random forest algorithm was used for model training, and the following parameter configurations and model training duration were recorded. Table 3 shows the specific model parameter values.

Table 3: Model parameters

P arameter N ame	Parameter Value
Training time	0.66s
Data segmentation	0.7
Data shuffling	False
cross validation	False
Evaluation criteria for node splitting	mse
Maximum proportion of features to consider for partitioning	None
Minimum number of samples for internal node split	2
Minimum number of samples at a leaf node	1
Minimum weight of samples in leaf nodes	0
The maximum depth of the tree	10
Maximum number of leaf nodes	50
Threshold for node partition impurity	0
Number of decision trees	100
There's put back sampling	true
Out-of-bag data testing	false

Training time: 0.66s shows that this parameter indicates the short time spent in the model training, which helps to improve the training efficiency. Data segmentation: 0.7 The data segmentation set to 0.7 indicates that 70% of the data is used for training and 30% for testing. Usually, reasonable data segmentation ratio can provide sufficient training data to verify the generalization ability of the model. Evaluation criteria for node splitting: The evaluation criteria for mse node splitting is mean variance (Mean Squared Error). Mean variance is a common measure of the model fitting error, and more accurate prediction results can be obtained by minimizing the mean variance. Number of decision trees: 100 indicates the number of decision trees. The random forest model consists of multiple decision trees to improve the model accuracy and stability by merging the results of multiple decision trees. There's s put back sampling: True means that the selected sample is reproducible each time it is selected from the dataset. Putting back sampling helps to increase the diversity of the sample and improve the model stability.

For model training, the full dataset was used in this study and divided into 70% to build the decision tree, and the remaining 30% to evaluate the model performance. The mean square error (MSE) was selected to measure the prediction error when division. To control for the model complexity and the risk of overfitting, this study limits the decision tree to the maximum depth of 10 and the maximum number of leaf nodes of 50. Furthermore, this study requires that internal nodes have at least 2 samples, leaf nodes have at least 1 sample, and all features were considered for division. To ensure the model generalization ability, the study used the drop-back sampling technique to sample the training samples and did not test the performance with out-of-bag data. Note that the model training time was only 0.66 seconds, indicating that the model training speed is faster. These parameters were chosen based on experimental experience and professional principles and adjusted with the characteristics of the dataset, aiming to obtain models with high accuracy and good generalization ability.

3.2. Characteristic importance of daily prediction

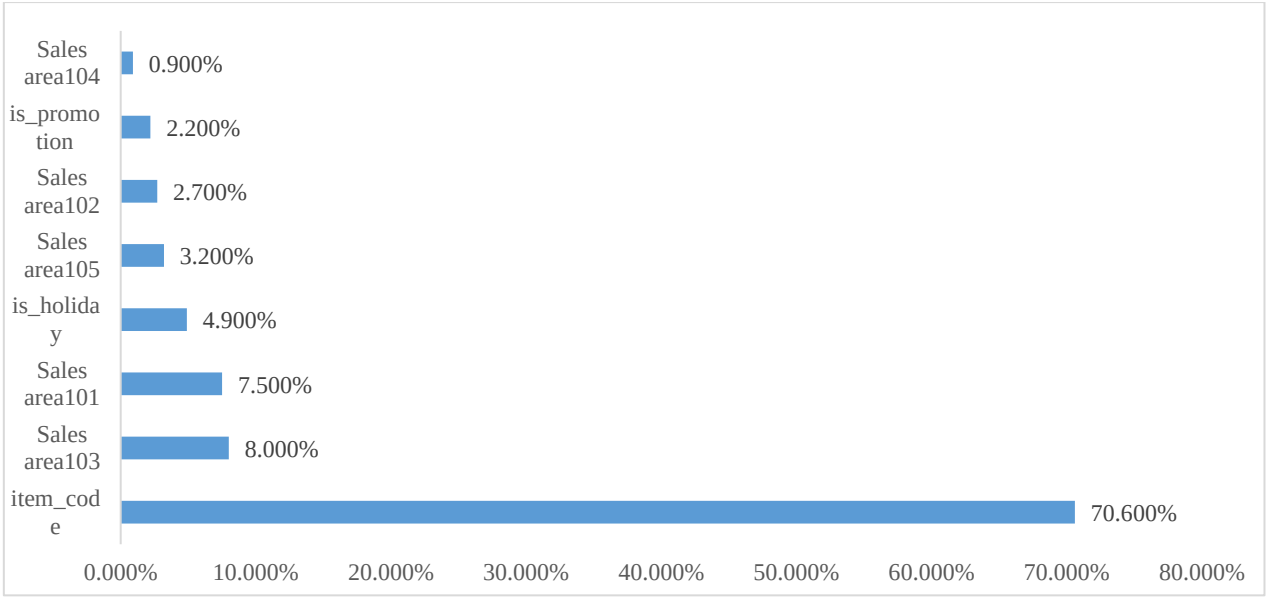

Figure 1 Feature importance predicted by day

The feature importance of daily prediction is shown in Figure 1, and this study found item _ code to be the most important feature and had the greatest impact on daily forecast order demand. Other features such as Sales area103, Sales area101, and is_holiday also contribute to the prediction results.

Table 4: Characteristic significance of daily prediction

	MSE	RMSE	MAE	MAPE	R^2
Training set	38.2695	12.1356	8.342	8.833	0.779
Test set	35.2369	13.2654	8.203	8.392	0.767

As shown in the evaluation indexes in Table 4, the model showed good predictive performance and generalization ability. These results can be used to optimize the prediction model and improve the efficiency of supply chain management.

3.3. Characteristic importance predicted by week

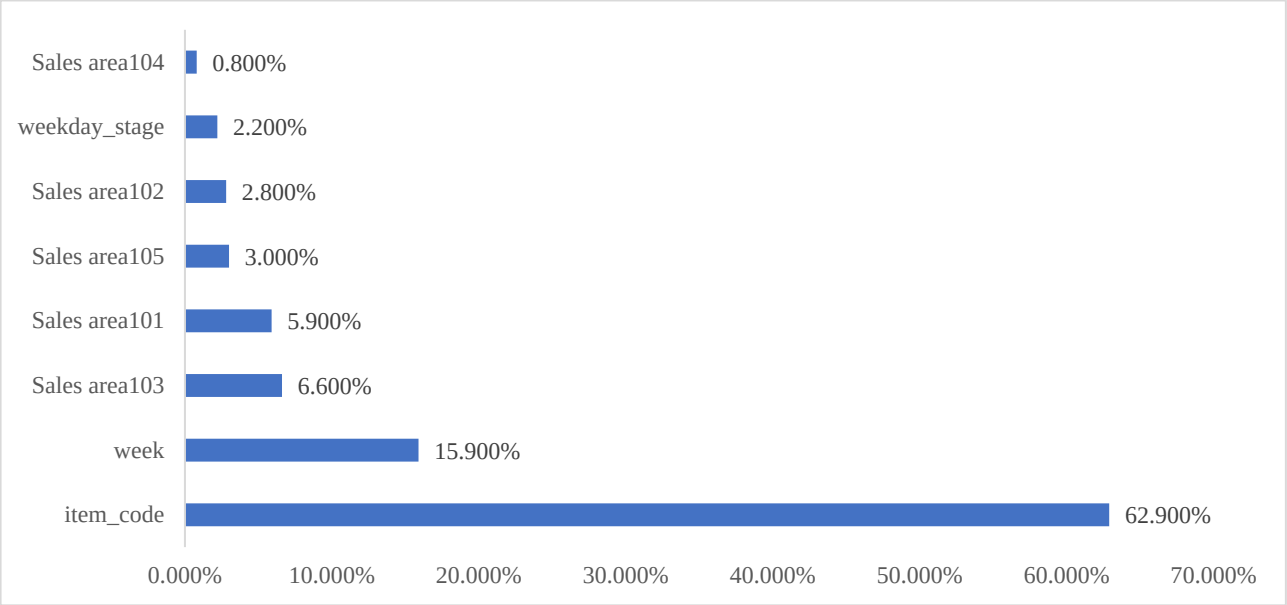

Figure 2 The predicted feature importance by weeks

Table 5: Features importance predicted by week

	MSE	RMSE	MAE	MAPE	R^2
Training set	34.0748	13.289	8.342	8.233	0.772
Test set	31.2356	12.639	8.203	8.361	0.756

Based on the results of the data analysis in Figures 2 and Table 5, some important conclusions can be drawn from this study. First, when predicting order demand weekly, the weekly feature is more important than the weekday feature. This means that the week plays a key role in predicting the order demand, and different weeks may have different effects on the forecast results. In addition, a moderate selection of time granularity is a key factor to ensure a relatively stable prediction results. These conclusions play a vital reference role in guiding supply chain management decisions. Based on the importance of week characteristics, supply chain managers can pay more attention to adopt different regulation strategies in different weeks to better meet the order demand. At the same time, by controlling the time granularity in a moderate range, the prediction error can be reduced, and the accuracy and stability of the model can be improved. Moreover, the model performed well on the training and test sets, with a lower error and a higher coefficient of determination. This indicates that the model can fit the training data well and show good predictive power on the unseen test data. This provides strong support for the present study regarding the reliability and validity of the model. The results show that when predicting order demand weekly, the importance of weekly characteristics is higher than that of working day characteristics, and the time granularity is moderate, so the prediction results are relatively stable. This provides important reference information for guiding supply chain management decisions, and the model performs well on the training and test sets, with low error and high coefficient of determination.

3.4. Characteristic importance predicted monthly

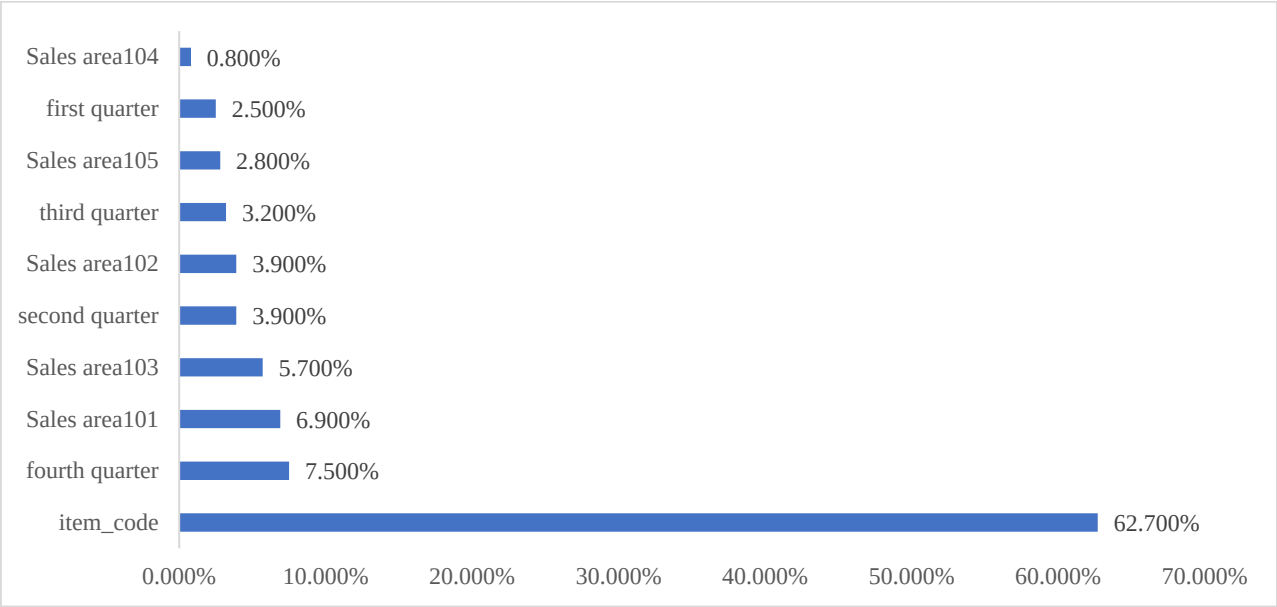

Figure 3 Features importance predicted by month

Table 6: Features importance predicted by month

	MSE	RMSE	MAE	MAPE	R^2
Training set	31.2695	11.289	7.345	7.953	0.802
Test set	30.2569	11.369	8.965	8.256	0.826

In Figure 3 and Table 6, the importance of the second and fourth features is higher than that of the first and third quarter characteristics, with high time granularity, the most stable prediction but less capture details. The prediction evaluation index of the cross-validation set, training set and test set are shown, and the prediction effect of the random forest is measured by the quantitative index. The cross-validation set was evaluated by adjusting the hyperparameters to obtain a reliable and stable model. Feature selection plays an important role in choosing to predict time granularity [9]. For example, short-term fluctuations can be captured by adding the variables "is it a holiday" and "is it a promotion day". Add the "weekend or not" variable to the weekly forecast. Adding "quarterly" and "monthly" variables to the monthly forecasts can capture trends and seasonal changes. In conclusion, different time granularity has multiple effects on prediction accuracy and may ignore important details. At lower time granularity, predictions are more sensitive; at higher time granularity, predictions are more stable but capture less detail. Feature selection can also affect the prediction accuracy. Therefore, the appropriate time granularity and characteristics should be selected according to the business requirements and data characteristics. In this study, mathematical models are needed to predict the product demand in the next three months. Before determining which model to use, preliminary analysis [10] to investigate the problem characteristics and applicable model.

4. Conclusion

In this study, a product order demand prediction model based on random forest was successfully developed. An accurate and reliable prediction model was established through in-depth analysis of the influence of various factors on product demand. The model demonstrates its ability to deal with complex problems and large-scale data, and it improves the quality of the prediction results through a complete set of data preprocessing methods. The study found that the monthly prediction model performs the best, providing higher accuracy and stability. This implies that companies can utilize

the model to arrange production plans effectively, optimize inventory management, and enhance customer satisfaction and economic benefits. Overall, this study provides a scientific decision-making method for enterprises to make more informed decisions regarding product order requirements.

References

[1] Zhang Qingshan, Liu Yanfeng, Xu Wei. Multiple product order similarity study under multiple uncertainty requirements [J]. Journal of Shenyang University of Technology (Social Science Edition), 2020,13 (03): 219-225.

[2] G.T.S.H,S.K.C,P.H.T, et al.A forecasting analytics model for assessing forecast error in e-fulfilment performance[J].Industrial Management & Data Systems,2022,122(11).

[3] Hugo V F,Silva D L R C,Cesar A C, et al.Big Data Analytics for Spatio-Temporal Service Orders Demand Forecasting in Electric Distribution Utilities[J].Energies,2021,14(23).

[4] Tunyang G,Tianzhen J,Bingnan L, et al.Prediction of the Tropospheric NO2 Column Concentration and Distribution Using the Time Sequence-Based versus Influencing Factor-Based Random Forest Regression Model[J].Sustainability,2023,15(3).

[5] L.S.L,C.V.C.G,A.L.M, et al.A travelling wave-based fault locator for radial distribution systems using decision trees to mitigate multiple estimations[J].Electric Power Systems Research,2023,223.

[6] Yaowen L,Jianguo Y,C S M, et al.Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model.[J].Environmental science and pollution research international,2022,29(22).

[7] Chen S,Wei Q,Zhu Y, et al.Mediumand long-term runoff forecasting based on a random forest regression model[J].Water Supply,2020,20(8).

[8] Hissou H,Benkirane S,Guezzaz A, et al.A Novel Machine Learning Approach for Solar Radiation Estimation[J].Sustainability,2023,15(13).

[9] Min L,YiTing W,XiaoKang W, et al.A multi-granularity convolutional neural network model with temporal information and attention mechanism for efficient diabetes medical cost prediction.[J].Computers in biology and medicine,2022,151(Pt A).

[10] L.K H,L.M M,Amanda L, et al.In Vitro Induction of Human Regulatory T-Cells (iTregs) Using Conditions of Low Tryptophan Plus Kynurenines[J].Blood,2016,128(22).