Sales Prediction Based on Machine Learning Approach: Support Vector Machine, Xgboost and Random Forest

Yixuan Jin *

Guanghua Cambridge International School, Shanghai, China

* Corresponding Author Email: nonghuan@ldy.edu.rs

Abstract. Living in the age of digitalize, the introduction of the Internet and e-commerce have brought consumers with convenience and choices. Since more people are used to shopping on the Internet, e-commerce enterprises should be able to predict demand of consumers when facing the fierce competition in the rapidly growing market. Therefore, the background of highly competitive and rapidly changing market has proved the significance sales prediction is for an enterprise. An accurate result of sales prediction can support enterprises with business decision basis. Previous studies on machine learning approaches used for sales prediction include techniques such as data mining, deep learning and time series analysis, so as to improve the performance of prediction models. Besides, researchers develop prediction models against specific industries, regions, etc. Through continuous study and innovation, with the support of more data source, sales prediction based on machine learning will promote the efficiency and value for enterprises. This paper aims at studying the significance of sales prediction based on machine learning, as well as analyzing the principle, advantages and limitations of three common machine learning approaches: support vector machine, XGBoost and random forest.

Keywords: Machine learning; sales prediction; support vector machine; XGBoost; random forest.

1. Introduction

With the increasing demand of people shopping online in today's world, the e-commerce industry is growing rapidly and attracted lots of customers. Indeed, with the development of e-commerce platforms and the increasing reliance on the Internet in people's lives, the importance of enterprises having an accurate sales volume prediction is becoming more crucial. In the field of supply chain management and sales decision, sales prediction with high accuracy is one of the most essential topics [1, 2]. To sustain the status in the industry, enterprises can apply sales prediction to increase productivity and better meet the challenges in the fierce competition [3]. Sales prediction helps enterprises in their operation in several aspects. Enterprises can use the result of prediction to allocate resources, manage stock, build a sustainable supply chain as well as to make sales decisions [4]. In summary, an accurate sales volume prediction is increasingly essential in the e-commerce industry due to its advantage of providing market information, optimize resource allocation and operational management, guide sales decisions, reduce risks, and improve supply chain efficiency for enterprises. It plays an important role in the development and competitiveness of enterprises.

The early research of sales prediction can be dated back to a few decades ago when people use artificial statistical measures such as moving averages, exponential smoothing, and trend analysis. These methods sure provide predictions to some extent, but they are highly inaccurate in a complex and rapidly changing market. The advancements of technology and techniques had significantly promoted the sales prediction methods, showing the limitation of the original statistical ones. To overcome these problems, researchers start turning to more advanced technologies and approaches. The emergence of computers has made it easier for researchers to process a large amount of sales data and run more complicated models. In the late 20th century, the time series analysis became the most widely use approach of sales prediction. The ARIMA model and exponential smoothing method are two of the most popular ones among all algorithms [3, 5]. This method involved analyzing historical sales patterns, trends, seasonality, and other underlying factors to make a prediction of the future sales. Today, people use machine learning and powerful algorithms to deal with such problems.

By applying machine learning models such as XGBoost, random forest and support vector machines, one can predict the sales value in a more accurate way and obtain refined results [6]. There exists a large number of machine learning approach for solving the problem of sales prediction. In order to maximize the advantage of sales prediction, it is wise to choose appropriate models for different scenes [7]. The rest of the paper will illustrate the principle of three most widely used ones, the support vector machine, XGBoost and random forest.

2. Basic Descriptions

The machine learning sales prediction is based on machine learning algorithms. The models are trained with several factors related to sales. Sales related data can be classified by different ways into several categories. Usually, the factors can be divided into internal and external factors. Internal factors, like stock, productivity and sales volume, are factors about specific enterprises [8]. The external factors such as economic environment, sales of competitors, are factors about the macro environment [4]. Besides, these variables can also be classified according to the principal machine learning works, the variables can be categorized into two types (discrete data and continuous data). Discrete data, including categories of products, seasons and dates, etc., are variables with a limited number of possible values. The continuous data is infinite in its possible values, such as price, sales value, stock volume, price reduction, etc. [9]. The training and predicting process of the machine learning models involves a large amount of these variables, and the suitable machine learning approach varies with the type of data. A large amount of data or multiple kinds of variables can be helpful for sales prediction, increasing accuracy and reliability, while spending more time and energy training the model and doing predictions [3].

3. Models

3.1. Support Vector Machine

Support vector machine is one of the most common machine learning approaches used in sales prediction. The operational principle of support vector machine comes as follows, including prepare of raw data, training the model, optimization and prediction. One first inputs all the data of sales history, mainly sales value and including external factors such as economic environment, sales of competitors and price of the product [4]. In the second step, one can preprocess the raw data collected, reorganizing the missing data, outliers and duplicate data so that the data set can be trained and used for further prediction. Thirdly, one can extract the features of data through case folding, stemming, tokenization according to the type of data selected [2]. Subsequently, the data set is separated into training data set and testing data set. A large proportion of the data is used for training the model and the rest is prepared for evaluation and validation in the future [2]. After that all the data is ready and the training can get started. A kernel function (linear, polynomial, radial basis function, etc.) is involved when generating the best fitting hyper-parameter, and that set constructs a hyper-plane which best separate different kinds of data, based on structural risk minimization and Vapnik-Chervonenkis [1]. In other words, the hyper-plain maximizes the margin between two kinds of data (seen from Fig. 1).

By using the feature and sales in the training data set, the model is trained through finding the best decision boundary. After the long process, the model is assessed according to accuracy, recall or root mean square error (RSME) to estimate its performance [2]. If the model turns out to be inaccurate, the parameters and kernel function should be changed to improve the performance of the model. After all these preparations, the model is able to predict sales. By entering the latest sales value, one can get the evaluated sales value and category. In addition, the new sales data should also be update regularly to retrain the previous model so that it can better adapt the changing marketing environment. Support vector machine has the advantages of non-linearity, robustness and high accuracy when

predicting sales. However, it requires complex calculation and is sensitive to parameters and wrong raw data [1].

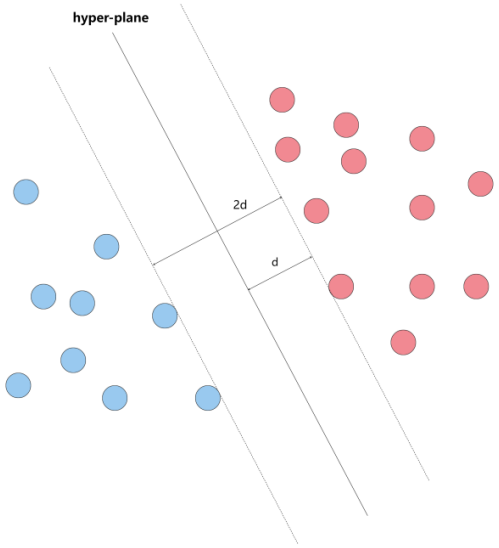

Fig. 1 Support vector machine generating a hyper-plane.

3.2. XGBoost

XGBoost stands for extreme gradient boosting package [10]. It is based on the principle of gradient boosting decision tree but much more efficient [3]. It performs well in prediction and regression problems. It can be used to predict continuous target variables such as sales volume, user behavior, stock prices, housing prices, etc., and accurately predict based on historical data and other characteristics [8]. XGBoost is a gradient boosting algorithm [3]. It can deal with both continuous variables (sales volume and value, etc.) and discrete variables (such as category of product, etc.) [3]. Through combining multiple weak learners (such as decision trees) into one strong learner, prediction performance is improved. Each weak learner is trained on the basis of the previous weak learner, attempting to correct the errors of the previous model and gradually improve overall performance. Gradient boosting is a way of realizing tree boosting [8]. In each iteration, XGBoost calculates the gradient of the loss function over the predicted value and uses a weak learner to estimate the negative gradient, reducing the loss function. This allows the model to make more precise corrections to errors in each iteration, gradually improving model performance. A sketch is shown in Fig. 2.

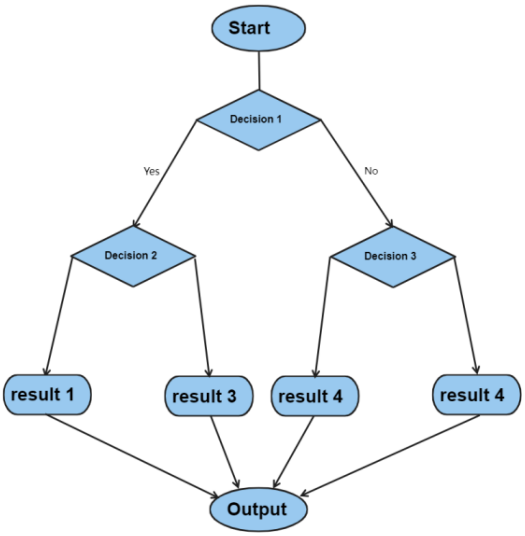

Fig. 2 Display of an XGBoost decision tree

XGBoost first collects sales related data, including historical sales data, product features, prices, promotional activities, competitor data, etc. and ensure the quality and integrity of data, followed by pre-processing and feature engineering of data, including handling missing and abnormal values, standardizing and normalizing data [11]. At the same time, feature selection and feature construction are carried out to extract the most informative and relevant features. To estimate the performance of the model in the future, the data set is divided into training and testing data sets. In most cases, data is divided in chronological order to ensure that the training set contains earlier data, while the test set contains latest data. The parameters are then set to optimize the model for further training. There are some common parameters used in XGBoost. Learning rate is used as a weight multiplier in generating the decision tree to prevent over fitting [8]. The maximum depth of the tree can be applied to control the complexity of the model. A tree with higher depth can capture complicated relationships, while also can cause over fitting [8]. In the following step, the model is trained based on distribution of target variables and it obeys loss minimization function. Next, by using techniques such as cross validation to tune the model. The optimal hyper-parameter combination can be found through grid search or random search [8] The model is then estimated according to indicators such as root mean square error and mean absolute error and brought into use. The model can predict the future sales through entering related features of sales data.

The results achieved by XGBoost in sales forecasting are usually satisfactory. Its advantage lies in its ability to handle high-dimensional sparse data and complex feature relationships, and its strong generalization ability. XGBoost can automatically handle missing values, support feature selection and feature importance assessment, which is very helpful for explaining the model and identifying key features [8].

3.3. Random forest

Random forest is an ensemble learning algorithm based on decision trees. It conducts comprehensive prediction by integrating the prediction results of multiple decision trees. Firstly, it randomly selects multiple samples from existing sales data to form a randomly sampled training set and a random subset of multiple features. Then, a decision tree model is constructed for each sample and feature subset. Decision trees divide data into different categories (or regression values) by dividing the feature space [3, 6]. When making sales predictions, the model input new sales samples into each decision tree. Each decision tree will be classified based on the characteristics of the samples and decision rules to obtain a prediction result. Finally, by summarizing the prediction results of each decision tree, such as taking majority votes (classification problems) or averaging (regression problems), the final sales prediction result is obtained [3]. Random forests can effectively handle noise and nonlinear relationships in sales data through the integration of decision trees, improving prediction accuracy and robustness. At the same time, by introducing randomness, each decision tree is established using different samples and feature subsets, reducing the variance of the model and the risk of over fitting [12]. A sketch of the random forest is given in Fig. 3.

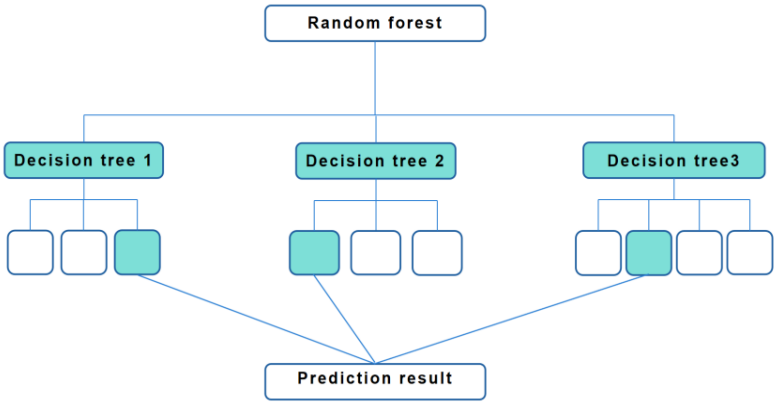

Fig. 3 Display of the Random Forest.

In the process of predicting sales, random forest is carried out through the following steps. Firstly, one collects relevant feature data such as historical sales data, product features, market information, etc., and then clean, process, and feature engineer the data to convert it into a format that can be used by random forest models. Next, construct a random forest model. A random forest is composed of multiple decision trees, each trained based on a randomly sampled data set and feature subset [3]. This randomness helps to improve the stability and generalization ability of the model [12]. Each decision tree is gradually constructed based on the segmentation points of features, ultimately forming a powerful integrated model. Then, use the trained random forest model to predict sales data. For each sales sample, the model classifies or regresses the samples on each decision tree to obtain the prediction results for each tree. The common method is to use majority voting (classification problem) or averaging (regression problem) to summarize the prediction results of each tree and obtain the final prediction [12]. Finally, evaluate the prediction results and the performance of the model. Some indicators such as mean square error (MSE) or accuracy can be used to evaluate the predictive performance of the model, and cross validation techniques can be used to verify the robustness and generalization ability of the model. Overall, the process of random forest prediction sales includes data collection and preparation, model construction, prediction, and evaluation. Random forests can effectively process sales data and provide accurate sales prediction results through the integration and randomness of decision trees.

In summary, random forests integrate multiple decision trees and utilize randomness for sales forecasting. It can process complex sales data, provide accurate prediction results, and has good robustness and generalization ability [6]. In the case of a single sample, ensemble model is less likely to make mistakes than a certain classifier [3].

4. Conclusion

Support vector machines (SVM), XGBoost and random forest are three of the most widely used sales prediction models with high efficiency. Their advantages and limitations vary from each other. Support vector machines have great performance when processing high-dimensional data set and small sample data set. Moreover, support vector machines can successfully handle complex sorting problems through nonlinear mapping by using kernel functions in the case of nonlinear data set. However, it has limitations such as long training periods when processing large scale data set, as well as over fitting when dealing with data full of noise. XGboost has advantages of high training efficiency under large scale data set, and it has strong generalization ability and accuracy dealing with nonlinear relationship, while it also has the shortage of over fitting and require some time and experience tuning hyper-parameters. Random forest is the least likely model that will over fit with data set full of noise and outliers and it is highly extensible. Nevertheless, it provides a less accurate result in the prediction due to its randomly chosen sample set and quality of adjustment of its parameters. Therefore, in practical applications, when selecting suitable algorithms, it is necessary to evaluate and compare them based on specific problems and data characteristics. The performance and accuracy of the model can be improved by using cross validation, parameter tuning, and integration methods. In addition, feature engineering, data quality, and quantity also play an important role in the quality of prediction results. This paper showed the advantages and limitations of three common approaches for machine learning-based sales prediction. Entrepreneurs should balance the significance of efficiency, accuracy, robustness and interpret ability when choosing a suitable method for specific sales prediction. However, it is important to note that the success of machine learning-based sales prediction relies on quality data, model training, and continuous monitoring and refinement. Enterprises should invest in data collection, data quality assurance, and model development to ensure reliable predictions and achieve optimal results. Furthermore, the machine learning field is developing rapidly which new models and technique are introduced. It’s worth time and energy studying the latest machine learning approach in order to make full use of sales prediction.

References

[1] Yue L, Yafeng Y, Junjun G, et al. Demand forecasting by using support vector machine. Third International Conference on Natural Computation (ICNC 2007). IEEE, 2007, 3: 272-276

[2] Lutfi A A, Permanasari A E, Fauziati S. Sentiment analysis in the sales review of Indonesian marketplace by utilizing Support Vector Machine. Journal of Information Systems Engineering and Business Intelligence, 2018, 4(1): 57-64.

[3] Huo Z. Sales prediction based on machine learning. 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 2021: 410-415.

[4] Ningyu T, Runyou F, Zhiyu Z, et al. Application of Regression Analysis in Sales Forecast. China Township Enterprises Accounting, 2019 (12): 107-109.

[5] Tsoumakas G. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 2019, 52(1): 441-447.

[6] Bohanec M, Borštnar M K, Robnik-Šikonja M. Explaining machine learning models in sales predictions. Expert Systems with Applications, 2017, 71: 416-428.

[7] Cheriyan S, Ibrahim S, Mohanan S, et al. Intelligent sales prediction using machine learning techniques. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 2018: 53-58.

[8] Qianyi Y, Hong R, Mingshu J. Commercial Sales Forecast Based on Xgboost. Journal of Nanchang University: Science Edition, 2017, 41(3): 275-281.

[9] Pavlyshenko B M. Machine-learning models for sales time series forecasting. Data, 2019, 4(1): 15.

[10] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015, 1(4): 1-4.

[11] Zhang L, Bian W, Qu W, et al. Time series forecast of sales volume based on XGBoost. Journal of Physics: Conference Series. IOP Publishing, 2021, 1873(1): 012067.

[12] Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random Forest. Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3. Springer Berlin Heidelberg, 2012: 246-252s.