Forecasting Retail Sales Via the Use of Stacking Model

Che Sun

Sino-European school of Technology of Shanghai, Shanghai University, Shanghai, China, 200444
Email: sunche_28@shu.edu.cn

ABSTRACT

Nowadays, the march of machine learning brings about the improvements of companies' ability to respond the changes in the marketplace and enables them to balance more easily the supply and demand. Thus, predicting based on historical data is getting more and more prevalent. There are numerous approaches applied to attain better results in this research area. The data in this research is from Kaggle and is genuine data provided by 1C company. This paper adopts six models, i.e., Linear Regression, Ridge regression, Random Forest, GBDT, XGBOOST and Stacking to forecast the future sales of retail products based on the historical data. The root mean square error between the real and anticipated data is utilized as performance evaluation. And the results show that the stacking method presents the best performance.

Keywords: *machine learning, predict, models, stacking*

1. INTRODUCTION

The sales trend is always changing over time. It is necessary to make the prediction based on the past data to find out the supply and demand to avoid the scarcity or oversupplying from this. Many giant firms have started to use the historical data to forecast the future sales to adjust their marketing strategies. Prediction by machine learning based on the historical data has been increasingly common nowadays.[1] Mitchell mentioned that machine learning is one of the most rapidly growing technical fields today and data-intensive machine-learning methods are applied to many walks of life including financial modeling, marketing and so on. And finding the novel and suitable learning algorithms and theory propels the machine learning forward. Besides,[2] Hasan presented in his paper that analyzing the gigantic data with machine learning to solve business problem such as predicting future sales and future demand is very crucial to every company and prediction has been an integral and key part of the business value chain. Thus, applications of machine learning to sales prediction are proving to be promising and pretty important to any organization as a valuable reference. And finding a best model to address the specific problem is also worth thinking.

There are some papers focusing on the comparison of different time series forecasting methods but not as much as expected and numerous papers employ the complex models to solve the time series problem, such as Neural network, LSTM and ARIMA.[3] Gupta used multiple methods, for example, Random Forest, Neural network and SVM to solve a time series problem about the covid-19 in his paper. He compared the several methods and found that Random Forest is the best way to solve that specific problem. [4] Pan used the XGBOOST model to predict hourly PM2.5 concentrations in China. And the regression models of random forest algorithm, multiple linear regression, decision tree regression and support vector machine are applied as well to compare with XGBOOST. The results show that the XGBOOST algorithm outperforms other data mining methods. [5] Using an ensemble learning methodology, Ashkan employed the random forest (RF) model to forecast both the exact values and the class labels of 24hourly prices in the California Independent System Operator's (CAISO) day-ahead energy market. It has been discovered that the proposed data mining approach performs well in both forecasting the precise value and classifying prices as low, medium, or high. It can be found that in different cases, the performance of each model will also change. Only experiment as much as possible in different situations and then people can better explore the applicability, strengths, weaknesses and potential of different models. Moreover, it has also been found that there are not many papers mentioned the stacking method in this time series prediction. Therefore, in this paper, several basic methods will be applied to make the prediction of sales of the retail stores and the comparisons between different methods will be made to find a best model for this specific problem. The stacking methods is presented in the paper as well.

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 405–411, 2022.
https://doi.org/10.2991/978-94-6463-036-7_59

There are several steps in the experiment, including data pre-processing, feature engineering and selection, model selection, and model parameterization. These stages contribute to the generalizability of the final prediction model, and each of these procedures is critical. During the experiment, some better methods came up according to the deficiency of the prior method. The stacking method is employed as well. First, an approximate preview of the data is made, and missing data is processed. Next, the data is analyzed visually. Then hidden information in the data is mined and new features are created. Finally, the different models are used to make the prediction. Linear Regression, Ridge Regression are first used to make a prediction. Random Forest Ensemble, GBDT Ensemble and XGBOOST Ensemble are used respectively to make the prediction in order to increase the complexity of stacking. The selection of models and order of use also can be regarded to be a replay of the process of model optimization. Different models yielded totally different results and each model also has its own strengths and weaknesses. In order to integrate the results of these models, the stacking method can be used to train the model again on the basis of the original prediction results. In this experiment, root mean squared error between actual and predicted data works as the evaluation of the performance. Then, through the comparison of different aspects, it has been found that the most accurate models for this specific problem in the paper is stacking method. It indeed gives the lowest RMSE.

The paper is organized as follows. A description of the data is given in Section 2. Section 3 concerns about the process of exploratory data analysis and data cleaning. The methods for modelling these data are introduced in Section 4. In Section 5, the results of different modelling methods and a brief comparison of these methods will be presented. In section 6 and 7 are in regard to discussing the results and making the conclusion, respectively.

2. DATA

1C Company is a Russian software development and publishing company that is completely independent of any other organization. This company is widely regarded as the market leader in commercial software on the Russian internal market.

It launched a competition on the Kaggle platform (https://www.kaggle.com) titled "Predict future sales," which is concerned with time series prediction in order to discover a robust model that can handle this problem and allow them to alter their marketing plans as necessary. The company offered historical sales data for every store and item that is sold on a daily basis on the Kaggle. The daily historical data ranges from 2 January, 2013 to 31 October, 2015. Forecasting total sales for each product and store for the upcoming month is a requirement for competitors.

Prior to initiating the data analysis process, it is essential to conduct a preliminary investigation into the data collection's distribution. In order to ensure the accuracy of the prediction accuracy, it is necessary to check the missing information from train sets and test sets. By organizing and aggregating the data, the pandas function can generate a table similar to Table 1 that contains an overall summary of the data. The Table 1 contains information on the daily transaction volume, total transaction amount, and price of the product. As demonstrated by the table, there is a significant difference between the minimum and maximum value of “mean_price”, respectively 0.09 and 5000. This implies that there may be outliers.

Table 1. descriptive statistics of the train set data

	Date_block_num	Item_cnt_month	Mean_cnt_day	Mean_price
mean	15.16	1.81	0.86	748.00
min	0.00	-22.00	-22.00	0.09
max	33.00	1474.00	500.00	5000.00

Afterwards, using Matplotlib and Pandas, conduct visual analysis to identify data trends, detect abnormal in data, and prepare the data for further processing before continuing. In Figure 1 and Figure 3, it is clear to see the outliers that need to be addressed. Dealing with anomalous data can aid in the model's generalizability by allowing it to become more general. Aside from that, the visualization process can uncover some previously undiscovered information, allowing for the identification of features that have an impact on the prediction quantity to be discovered. These characteristics can also stimulate the development of fresh ideas for the company's marketing strategy. For example, Figures 2 and 3 illustrate the sales trend over time and the difference in sales between items. This kind of graphs assists in identifying hidden trends in sales and delving further into some of the aspects that may affect sales, hence improving prediction accuracy.

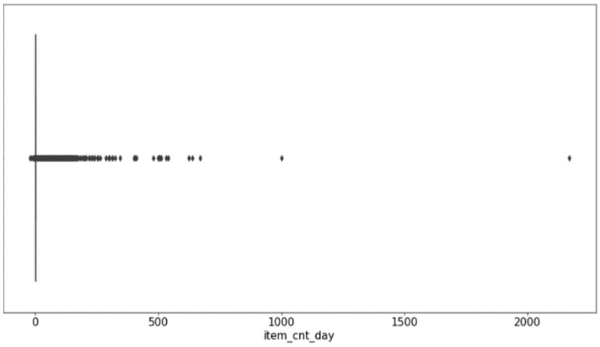

Figure 1 Daily sales value of the items

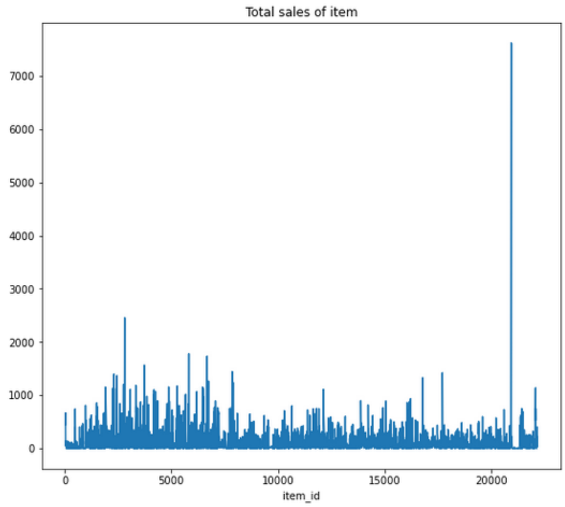

Figure 2 Sales of different product

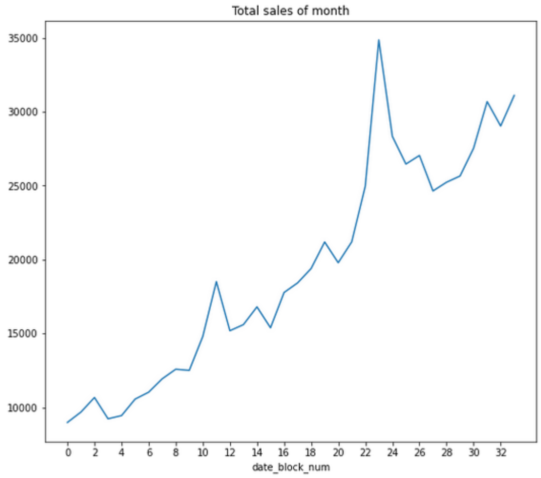

Figure 3 The trend of the sales by month

After data reconstruction and outlier processing, hidden information can be mined from the data to create new features to enhance the effectiveness of the model. The link between various characteristics and monthly sales must be examined in order to trim the data, lower the variance of the data in order to ensure the generalizability of the model, and avoid overfitting when the model is applied to the data in the following step.

The following feature engineering process was carried out using a variety of feature-picking strategies based on the model that was chosen: correlation analysis and Ensemble's own feature selection tool. This method has the potential to significantly increase the performance of the models through experimentation. What's more it has been observed that using time series to build a variety of trend features can significantly improve the accuracy of the resulting model. Thus, moving averages utilized in the trials include moving averages of three, six, and twelve months, which were used to reflect the short, medium, and long-term levels of month-to-month sales, as well as to generate monthly indicators.

3.METHODOLOGY

3.1 Linear Regression

Linear regression is a statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables that are interdependent and is widely used in different prediction problems. This algorithm is a most basic regression model which has a fast calculation speed without parameter adjustment.

In general, linear regression can be solved by least squares, and a straight line can be calculated:

$$\boldsymbol{y = a + bx} \qquad \textbf{(1)}$$

Where x refers to the explanatory variable, y is the dependent variable, a is the intercept (the value of y when x = 0), and b is the slope of the line.

However, there is often more than one factor influencing y. Therefore, a straight line like this is required to fit our data:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon} \qquad \textbf{(2)}$$

where $\boldsymbol{\beta_0, \beta_1, \dots, \beta_k}$ refers to the parameters, and $\boldsymbol{\varepsilon}$ is the random term.

3.2 Ridge Regression

Linear regression is based primarily on the least square method. The limitations of the ordinary least squares method make it impossible to use it directly for linear regression fitting in many cases. In order to solve the problems that arise in both cases, Ridge Regression came into being. The ridge regression algorithm emerged with the goal of finding one such solution among the underfitting and overfitting algorithms. The ridge regression algorithm will regularize the over-fitted higher power function. Regularization is the operation of reducing the higher power of a higher power function to a lower power. Regularization is divided into L1 regularization and L2 regularization. L1 regularization is the elimination of higher powers by directly setting the higher power coefficients to 0. L2 regularization means replacing the higher power coefficient with a coefficient infinitely close to 0, thus reducing the impact of the higher power on the whole linear regression process. Ridge regression can be regarded to be a modified least squares estimation method that effectively prevents the model from overfitting by adding the L2 regular term (2-norm) to the loss function and helps to solve the problem of inverse difficulties under non-full rank conditions, thereby improving the robustness of the model. [6] Kannard mentioned in his paper that the probability that Ridge regression produces a smaller squared error than least squares is greater than 0.50.

In the standard equation method, an equation is derived to solve for the regression coefficient:

$$\boldsymbol{\omega} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} \tag{3}$$

The following formula can be obtained by adding the regular term:

$$\boldsymbol{\theta} = (\boldsymbol{X^T X} - \boldsymbol{\lambda E})^{-1}\boldsymbol{X^T y} \tag{4}$$

3.3 Random Forest

The Random Forest is an ensemble technique of multiple decision trees. For the classification problem, the final classification result is decided by voting on a multi-tree classifier. For the regression problem, the mean value of the predicted values of the multiple trees determines the final prediction result. [7] The training samples for each tree are random. The set of training features for each tree is also randomly drawn from all features. The introduction of two randomness is crucial to the classification performance of random forests. Due to their introduction, the random forest is less likely to fall into overfitting and has good noise immunity.

During bootstrapping, some data may not be selected, these data are called out-of-bag (OOB) examples. The calculation formula is as followed:

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n} \approx \frac{1}{e} \tag{5}$$

Where n refers to number of samples

When n is large enough, approximately 36.8 percent of the training samples are not selected. In random forests, with out-of-bag (OOB) examples, we do not need to take out a part of the data, out-of-bag (OOB) examples is that part of the data that is not used, we can even directly use it as a validation set. [8]

[9] Zhang used this model in his paper concerning about the time series prediction and it presented the best performance.

3.4 GBDT

This model is an iterative decision tree algorithm. GBDT consists of multiple decision trees and finally the residual values derived from each tree are accumulated to make a decision. This is an algorithm with high generalization capability. Since it is the gradient values that GBDT has to fit in each iteration, the decision tree used by GBDT is a CART regression tree. Like other strengthening methods, build the model in the form of stage, and by allowing the optimization of the loss function of arbitrary separable variables to a generalized model. [10] Yang used this method in his paper to predict the stock and GBDT was the most accurate model for their problem.

This paper makes use of the GBDT regression technique. Assuming a training set sample $T = (x, y_1), (x, y_2), \ldots, (x, y_m)$, a maximum number of iterations T, a loss function L, and the output is a strong learner f(x). The f(x) can be shown as below:

$$\boldsymbol{f(x) = f_T(x) = f_0(x) + \sum_{t=1}^{T}\sum_{j=1}^{J} c_{tj}, I(x \in R_{tj})} \tag{6}$$

3.5 XGBoost

The XGBoost algorithm is often used in competition. In many papers, this method is always applied to deal with complex prediction problems and get a high degree of accuracy. Hence, it seems that using this method usually makes the accuracy of the prediction a big improvement. Gradient boosting is the XGBoost's original model, which iteratively combines weak base learning models to create a stronger learner.[11] The critical distinction between the boosting class of algorithms is the way gain is defined in each round of residual tree fitting. The gain definition used by XGBOOST is the structure score before splitting minus the structure score after splitting, and the splitting point with the highest gain is chosen as the optimal splitting point, i.e. the splitting point that results in the greatest reduction in model loss relative to the loss before splitting. XGBOOST customizes the gain splitting to optimize the decrease of the model's loss function in each iteration, hence speeding up the optimization. And XGBOOST adds a penalty term, which is mostly composed of the number of leaf nodes and their values, which minimizes the model's variance and prevents overfitting. [12] Swami employed this method in his paper and it presented the best performance in his experiment.

The objective function of XGBOOST is expressed as:

$$\boldsymbol{Obj^{(t)} = \sum_{j=1}^{T}\left[G_j\omega_j + \frac{1}{2}(H_j + \lambda)\omega_j{}^2\right] + \gamma T} \tag{7}$$

Where $\boldsymbol{G_j}$ refers to the sum of the first-order partial derivatives of samples contained in leaf node j (a constant), $\boldsymbol{\omega_j}$ is the value of the jth leaf node, $\boldsymbol{H_j}$ is the sum of the second-order partial derivatives of samples contained in leaf node j (a constant) and T is the number of leaves.

3.6 Stacking

Stacking regression is an integration learning technique that combines multiple regression models by means of a meta-regressor. Aside from that, each base regression model is trained with the entire training set, and the output of each base regression model is used as input for the meta-regressor as meta-features during the integration learning process. The meta-regressor then combines multiple models by fitting the meta-features to each base regression model. The following is a brief diagram of the stacking method as demonstrated on [12] CSDN.

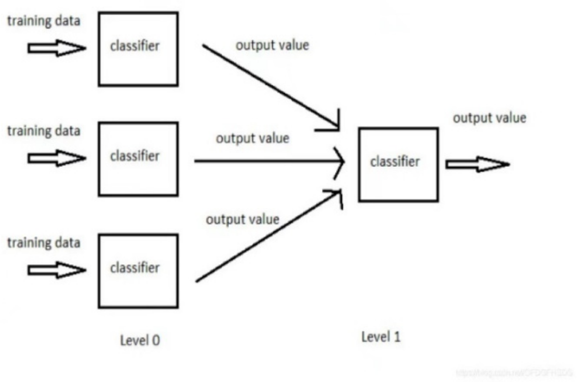

Figure 4 Concept diagram of stacking

4. RESULTS

A growing number of algorithms with improved performance have emerged in recent years, as machine learning continues to develop. Several model selection methods have been tried, including SVM, neural networks, and KNN methods. However, due to the large amount of data, the training speed of these models is extremely slow, tuning the parameters is difficult, and the actual efficiency of the model is not as good as some of the models applied in this paper.

In this experiment, the data set is divided into a training set, a validation set and a test set and root mean squared error (RMSE) between actual and anticipated data serves as a means of evaluating performance. RMSE is the square root of the mean of the squared differences between predicted and actual observations. It is a measure of the observed value's divergence from the real value.

The calculation formula is as followed:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_t)^2} \qquad (8)$$

Where m refers to the number of observations, $\hat{y}_i$ is actual value and y_t is predicted value.

The results of the training and validation sets derived from different models are listed in the Table2.

Table 2. Prediction results of different models

RMSE Models	Train set	Validation set
Linear Regression	0.391	1.248
Ridge regression	0.635	0.864
Random Forest	0.656	0.997
GBDT	0.561	0.988
XGBOOST	0.425	0.956
Stacking	0.657	0.745

As is shown in the table, the RMSE of the linear regression validation set is significantly larger than the RMSE of the training set, indicating that the model is overfitted. The RMSE of the validation set decreases and the degree of overfitting falls in the ridge regression exactly because the regular term is included in the ridge regression. In addition to the first two linear regression models, the following three models are tree models. As can be observed, the fitting impact improves incrementally, and the validation set's root mean square error diminishes.

The first five models obtained varying fits to the training set, and in order to combine their predictions, a model training was done again using their prediction outputs. The stacking model is formed by feeding the second layer of the model the prediction results from each individual model in the first layer of the validation set. In other words, the stacking training set's features are composed of five prediction results. I used the previous best performing XGBOOST model in the stacking method. And it is obvious that some single model predictions were found to be overfit, whereas stacking strategy not only minimized overfitting but also greatly improved model performance. Among these methods, the stacking model has the best fitting effect and performance.

5. CONCLUSION AND FUTURE WORKS

This study employs six models: Linear Regression, Ridge regression, Random Forest, GBDT Ensemble, XGBOOST and stacking to forecast future sales for the store, which are based on previous sales data provided by the company. The results demonstrate that the strategy of stacking effectively combines the results of the many models and delivers more accurate forecasts than other approaches. During the experiment, it has been found that data pre-processing, feature engineering and selection, model selection, and model parameterization all contribute to the generalizability of the final prediction model.

Unfortunately, due to linguistic constraints, some further potential features in the dataset cannot be mined in this time. For instance, whether the store is an online store or an offline store is not categorized as a feature. Additionally, there are not enough experiments to compare the characteristics of these models, and some time series models such as ARIMA are planned to make predictions in future work.

REFERENCES

[1] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[2] Hasan M R, Kabir M A, Shuvro R A, et al. A Comparative Study on Forecasting of Retail Sales[J]. arXiv preprint arXiv:2203.06848, 2022.

[3] Kashte S, Gulbake A, El-Amin III S F, et al. COVID-19 vaccines: rapid development, implications, challenges and future prospects[J]. Human cell, 2021, 34(3): 711-733.

[4] Pan B. Application of XGBoost algorithm in hourly PM2. 5 concentration prediction[C]//IOP conference series: earth and environmental science. IOP publishing, 2018, 113(1): 012127.

[5] Sadeghi-Mobarakeh A, Kohansal M, Papalexakis E E, et al. Data mining based on random forest model to predict the California ISO day-ahead market prices[C]//2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE, 2017: 1-5.

[6] Hoerl A E, Kannard R W, Baldwin K F. Ridge regression: some simulations[J]. Communications in Statistics-Theory and Methods, 1975, 4(2): 105-123.

[7] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[8] Zhang D, Qian L, Mao B, et al. A data-driven design for fault detection of wind turbines using random forests and XGboost[J]. Ieee Access, 2018, 6: 21020-21031.

[9] Zhang Y, Wu X, Gu C, et al. Predict Future Sales using Ensembled Random Forests[J]. arXiv preprint arXiv:1904.09031, 2019.

[10] Yang J S, Zhao C Y, Yu H T, et al. Use GBDT to predict the stock market[J]. Procedia Computer Science, 2020, 174: 161-171.

[11] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[12] Swami D, Shah A D, Ray S K B. Predicting Future Sales of Retail Products using Machine Learning[J]. arXiv preprint arXiv:2008.07779, 2020.

[13] CSDN. URL: https://blog.csdn.net/GFDGFHSDS/article/details/105324621

[12] H. Barringer, C.S. Pasareanu, D. Giannakopolou, Proof rules for automated compositional verification through learning, in Proc. of the 2nd International Workshop on Specification and Verification of Component Based Systems, 2003.

[13] M.G. Bobaru, C.S. Pasareanu, D. Giannakopoulou, Automated assume-guarantee reasoning by abstraction refinement, in: A. Gupta, S. Malik (Eds.), Proceedings of the Computer Aided Verification, Springer, Berlin, Heidelberg, 2008, pp. 135–148. DOI: https://doi.org/10.1007/978-3-540-70545-1_14