2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE 2021)

Machine Learning Model for Sales Forecasting by Using XGBoost

Xie dairu[1],
Chongqing University of Technology,
Chongqing, China,
1990410257@qq.com,

Zhang Shilong[1]
Chongqing University of Technology
Chongqing, China
1392068736@qq.com

2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) | 978-1-7281-8319-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCECE51280.2021.9342304

Abstract—For modern retail corporations operating a huge chain of businesses, exact sales predication is the key in driving corporations development, even success or failure. Sales forecasting allows corporations to efficiently allocate resources including cash flow, production, and make better informed business plan. In this paper, we propose an efficient and accurate sales forecasting model using machine learning. Initially, feature engineering is conducted for extracting features from historical sales data. Furthermore, we used eXtreme Gradient Boosting (XGBoost) to utilize these features for forecasting the future sales amount. The experiment results on a publicly Walmart retail goods dataset provide by Kaggle competition demonstrate our proposed model performs extremely well for sales prediction with less computing time and memory resources.

Index Terms—Sales Forecasting, Feature Engineering, XGBoost, Machine Learning

I. INTRODUCTION

Nowadays, we are undergoing a time of profound transformations powered by digitization, information and communications technology, artificial intelligence and so on. Many researchers in the business and economic fields suggest that this will usher in a new epoch - the Fourth Industrial Revolution. The fundamental shift in this social transformation will be in the decision-making area [1]. This is evident in retail industry that advanced technologies enable corporations to make accurate sales predication. Thus corporations are able to make reliably appropriate decisions and optimize their business resources. Take Walmart as an example, using machine learning for sales forecasting from massive historical data helps optimize their business operation - cash flow, staff, production, financial management. Moreover, it can reduce uncertainty and anticipate change in the market as well as compel to investors.

Recently, machine learning, an artificial intelligence method of discovering knowledge from large amount of data, has achieved pretty good performance in a wide range of problems, especially sales forecasting. Among the machine learning methods used in practice, XGBoost [3] is one technique that shines best. The reasons behind the success of XGBoost include its scalability and efficiency of running ten times faster than existing other machine learning algorithms. In this paper, we aim to design a sales forecasting algorithm based on XGBoost to achieve better performance. To be more specific, we try to discover knowledge of market sales from statistical sales data during the past 1913 days of Walmart Stores across three states and forecast the sales of the these stores in the coming 28 days. Compared to other ensemble learning methods, XGBoost runs ten times faster while uses far fewer resources. In the proposed method, since XGBoost is sensitive to outliers, we first do data preprocessing to filter outliers, then convert them into float data type for saving memory. Then aggregated time feature is extracted from different time periods of sales data from day, week and month to year. Following that is to capture a variety of features including price feature, rolling feature, lag feature and other statistical features. Finally, feature selection is applied to keep those highly relative features for predicating.

To effectively evaluate our algorithm, we extensively conduct various experiments on the public Kaggle competition dataset, which contains sales data collected in 1913 days provided by Walmart corporation. The experimental results demonstrate that our algorithm gives state-of-the-art results on sales forecasting. As we can see from RMSSE score, our proposed method has obtained 0.655, which is 16.3% lower than popular Linear Regression method [22], also 15.4% lower than Ridge Regression method [23] modified from Linear Regression.

The rest of the paper is structure as following. In section 2 we discuss some common machine learning algorithms and applications of XGBoost. Section 3 gives a detailed description of our work, namely feature engineering and XGBoost regression. Section 4 show the experimental results and analyses of our approach compared to other approaches. Finally, we conclude our work and present some points that may be improved in the future.

A. Related Work

Machine learning provides computers with the ability to learn without being explicitly programmed. The machine learning process is similar to the data mining process. Both systems search through data to look for feature patterns. However, instead of extracting data for manual operation, the machine uses the data to learn, to detect hidden patterns, and to adjust program actions according to objective function. In common, machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms learn from a full set of labeled data while training. And supervised learning can be divided into

978-1-7281-8319-0/21/$31.00 ©2021 IEEE

two areas, i.e. classification and regression. On the contrary, unsupervised algorithms handles a dataset without explicit instructions on what to do with it. Depending on the problem at hand, unsupervised learning model can organize the data in three ways: clustering, anomaly detection and autoencoding.

Among numerous machine learning models like Linear regression, Support Vector Machines [2-10], and Naïve Bayes models etc., XGBoost, proposed by Friedman [15], catches the eyes of most researchers and engineers. It gives very good results on a variety of problems. Gumus et al. use XGBoost to learn factors affecting the crude oil prices and make future oil price estimation [18]. In [17], an efficient machine learning method, random forest in combination with XGBoost is used to establish the data-driven wind turbine fault detection framework. In their approach, random forest is used to rank the features by importance, then based on the top-ranking features, XGBoost trains the ensemble classifier for each specific fault. In order to obtain the advantage of both linear and nonlinear models, Gurnani et al. [24] design a hybrid technique with decomposition technique to forecast drug sales of a drug store company, where linear component was forecasted by linear model and nonlinear component was forecasted by nonlinear model. Moreover, Zhong et al. propose a predicting framework based on XGBoost for identifying essential proteins, which includes a SUB-EXPAND-SHRINK method for constructing the composite features with original features and obtaining the better subset of features for essential protein prediction. They also introduce a model fusion method for getting a more effective prediction model [24]. In [21], XGBoost is implemented for predicating PM2.5 concentration through studying the factors influencing PM2.5. By incorporating elimination of unimportant features, the proposed model effectively improves estimation performance.

II. FEATURE ENGINEERING

Basically, all machine learning algorithms use some features extracted from input data to create outputs. In order to prepare the proper input data, which needs to be compatible with the machine learning algorithm requirements, and improve the performance of machine learning models. We apply some common feature engineering techniques for our algorithm.

First of all, to avoid outliers affect the performance, we detect the outliers by standard deviation method and filter these outliers to get clean data. Follow on upon with that is memory compression by converting data type into float type, which requires less memory. On account of our dataset size, memory compression can significantly reduce memory and speed up training.

Second, time is an important information for sales forecasting. For example, it can be easily noticed in Figure 1 that sales amount changes contain certain periodicity. So we attempt to capture different sales amount features across day, week, month and year, then aggregate them into joint time features.

Figure 1. Daily sales amount reports in a single month in 10 Walmart Stores.

Besides time features, some statistical features are necessary for analyzing sales as illustrate in Figure 2. Several deliberate features are calculated includes: rolling feature, lag feature, selling price feature as well as min, max, median features. Then we combine these features together.

Finally, feature selection is used to select rather relevant, important features and remove some redundant, useless features for creating our algorithm. This can not only reduce overfitting, but also enable our algorithm to train faster and cost less computing resources.

Once have obtained important features from raw data via feature engineering, we can feed them into our XGBoost algorithm for training.

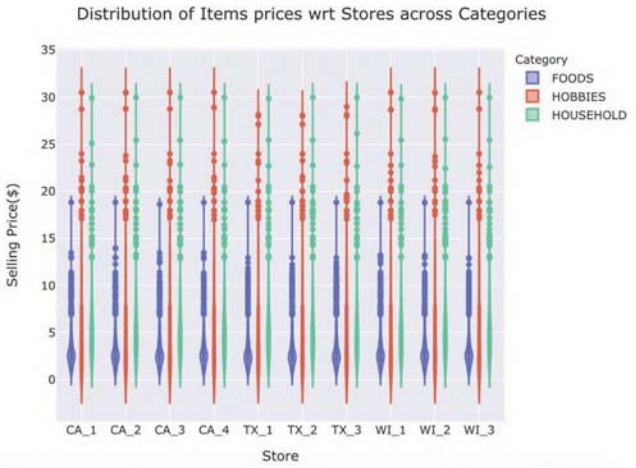

Figure 2. The distributions of commodity price of Foods, Hobbies, Household in 10 Walmart Stores.

III. SALES FORECASTING MODEL

In this section, we will introduce the proposed sales forecasting model based on XGBoost. XGBoost is a regression tree that has the same decision rules as the classic decision tree. In the regression tree, each inside node represents values for attributes test and leaf node with scores represents a decision. The output is the sum of all scores predicted by K trees, as shown below.

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

Where x_i is the i-th training sample of sales, f_k is the score for k-th tree and F is the space of function containing all the regression trees. XGBoost adopts the same gradient boosting as the Gradient Boosting Machine (GBM) [11], but makes a small improvement on the regularized objective, which penalizes the complexity of the model.

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)$$

Here L is the total objective function, l is a differentiable convex loss function that measures the distance between the prediction $\hat{y}_i$ and ground-truth y_i. The Ω is a regularization term defined as following.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (3)$$

Where γ and λ are constant coefficients, T controls number of tree leaves, w controls the score of each leaf. Different from GBM, XGBoost applies second order approximation to extend the loss function and removes the constant term to get the simplest goal, which help quickly optimize the objective in the general setting. Thus XGBoost can adapt to a wide range of problems.

Besides the mentioned regularized objective, XGBoost applies two additional techniques to further reduce over-fitting. The first is shrinkage [14], which reduces the influence of each individual tree and leaves space for future trees to improve the model by scaling newly added weights by a factor coefficient after each step of tree boosting. The second technique is column sub-sampling, which prevents over-fitting more than the traditional row sub-sampling.

Furthermore, in order to make model run efficiently, XGBoost introduces several powerful techniques includes: an approximate algorithm for exact greedy algorithm; storing the data in in-memory units for parallel learning; leveraging a cache-aware prefetching algorithm; enabling out-of-core computation. Through above methods, XGBoost is able to handle larger dataset and run much faster.

IV. EXPERIMENTS

To verify the performance superiority of the XGBoost-based sales forecasting model, we carry out extensive comparison experiments on a publicly Walmart retail goods dataset provide by Kaggle competition. The dataset contains hierarchical sales data during 1913 days from Walmart Stores in three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events.

Our aim is to forecast daily sales for the next 28 days using the historical sales data. And we choose RMSSE metric to measure model performance. The RMSSE stands for Root Mean Squared Scaled Error a variant of the original Mean Absolute Scaled Error (MASE) metric.

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \hat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(Y_t - Y_{t-1}\right)^2}}$$

Where h is prediction period, n is the time period of training, Y_t is the sales amount on the t-th day, correspondingly, $\hat{Y}_t$ is the estimated sales amount on the t-th day.

From the experimental results shown in Table 1, we can easily observe that our model has the superior performance over the traditional machine learning models. For RMSSE score, XGBoost model obtains lowest 0.655, while the classic Linear Regression model obtains 0.783, 19.5% higher than XGBoost model. And Ridge Regression model get 0.774, also 13.6% higher than XGBoost model. Through the detailed experimental results, we can see that the proposed XGBoost-based model has achieved much better performance than other regression models in sales forecasting.

Table 1. The experimental results of our XGBoost models compared to other popular models.

Models	**RMSSE**
Linear Regression	0.783
Ridge	0.774
XGBoost	**0.655**

V. CONCLUSIONS

In this paper, we systematically investigate the knowledge behind the historical sales data from Walmart stores. By leveraging feature engineering, we are able to capture the most useful and important features relative to sales and feed them to XGBoost model for forecasting future sales.

XGBoost, an optimized distributed gradient boosting tree model designed to be highly efficient and flexible, is good at classification and regression problem. The experimental results demonstrate that the proposed model based on XGBoost has good performance in improving the running rate and the prediction accuracy. And we achieve better sales forecasting compared to other popular machine learning methods.

In the future, we plan to incorporate other machine learning algorithms with XGBoost to construct a more powerful model, which can handle more hierarchy feature representations. Meanwhile, it is possible to explore more efficient features in feature engineering to further improve forecasting performance.

ACKNOWLEDGEMENT

First, we sincerely thanks to the Kaggle competition platform. The author Yiyang Niu thanks Zhihui Cai for the contribution of the idea and thanks Zhihui Cai the help in experiments. In addition, we also thanks Zhihui Cai for useful discussions.

REFERENCES

[1] Syam N, Sharma A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice[J]. Industrial Marketing Management, 2018, 69: 135-146.

[2] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[3] Suykens J A K , Vandewalle J . Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.

[4] Furey T S , Cristianini N , Duffy N , et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906-14.

[5] Tong S , Koller D . Support vector machine active learning with applications to text classification[C]// JMLR.org, 2002:999-1006.

[6] Furey, T, S, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. BIOINFORMATICS -OXFORD-, 2000.

[7] Tong, Simon, Koller, et al. Support Vector Machine Active Learning with Applications to Text Classification.[J]. Journal of Machine Learning Research, 2002.

[8] NGUYEN DUNG DUC MATSUMOTO KAZUNORI TAKISHIMA YASUHIRO. Re-learning method for support vector machine[J]. 2009.

[9] Heumann, Benjamin, W. An Object-Based Classification of Mangroves Using a Hybrid Decision Tree-Support Vector Machine Approach.[J]. Remote Sensing, 2011.

[10] Dong-Xiao N , Yong-Li W , Xiao-Yong M A . Optimization of support vector machine power load forecasting model based on data mining and Lyapunov exponents[J]. Journal of Central South University of Technology, 2010, 17(002):406-412.

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

[12] Friedman J H. Stochastic gradient boosting[J]. Computational statistics & data analysis, 2002, 38(4): 367-378.

[13] Torlay L , Perrone-Bertolotti M , Thomas E , et al. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017.

[14] Ji X , Tong W , Liu Z , et al. Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost[J]. Frontiers in Genetics, 2013, 10.

[15] Yin Y , Sun Y , Zhao F , et al. Improved XGBoost model based on genetic algorithm[J]. International Journal of Computer Applications in Technology, 2020, 62(3):240.

[16] Ren X , Guo H , Li S , et al. A Novel Image Classification Method with CNN-XGBoost Model[C]// International Workshop on Digital Watermarking. 2017.

[17] 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.

[18] Gumus M, Kiran M S. Crude oil price forecasting using XGBoost[C]//2017 International Conference on Computer Science and Engineering (UBMK). IEEE, 2017: 1100-1103.

[19] 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. 2016: 785-794.

[20] Zhong J, Sun Y, Peng W, et al. XGBFEMF: An XGBoost-based framework for essential protein prediction[J]. IEEE Transactions on NanoBioscience, 2018, 17(3): 243-250.

[21] Zamani Joharestani M, Cao C, Ni X, et al. PM2. 5 Prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data[J]. Atmosphere, 2019, 10(7): 373.

[22] Montgomery D C, Peck E A, Vining G G. Introduction to linear regression analysis[M]. John Wiley & Sons, 2012.

[23] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[24] Gurnani M, Korke Y, Shah P, et al. Forecasting of sales by using fusion of machine learning techniques[C]//2017 International Conference on Data Management, Analytics and Innovation (ICDMAI). IEEE, 2017: 93-101.