E-commerce Sales Forecast Based on Ensemble Learning

Choujun Zhan
School of Information Science and Technology,
Xiamen University Tan Kah Kee College
Zhangzhou, China
zchoujun2@gmail.com

Jianbin Li
School of Electrical and Computer Engineering,
Nanfang College of Sun Yat-Sen University
Guangdong 510970, China

Wei Jiang
School of Electrical and Computer Engineering,
Nanfang College of Sun Yat-Sen University
Guangdong 510970, China
jwwweee0115@gmail.com

Wei Sha
School of Electrical and Computer Engineering,
Nanfang College of Sun Yat-Sen University
Guangdong 510970, China

Yijing Guo
School of Information Science and Technology,
Xiamen University Tan Kah Kee College
Zhangzhou, China

2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN) | 978-1-6654-1461-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISPCE-CN51288.2020.9321858

***Abstract*—E-commerce, driven by the technological advances of the semiconductor industry, becomes one of the fastest-growing industries and has become the largest sector of the electronics industry in recent years. Compared with the traditional offline business, E-commerce business eliminates geographical limitations by allowing consumers to purchase anywhere. Additionally, E-commerce only requires a web-based platform to start a business, resulting in a lower investment or business capital. Based on more than 75,000,000 transactions of one E-commerce company from January 2013 to June 2019, we analyze the economic impact of the holiday season, the relationship between price and sales, etc. Furthermore, we utilize Artificial Intelligence (AI) techniques to predict the sales of products. Here, assemble learning models, including Catboost, GBDT, LightGBM, and XGBoost, are adopted for regression prediction. The results show that Cat boost performs better than other methods.**

***Keywords*-E-commerce; time series prediction; Sales forecast; Big data analysis; ensemble learning;**

I. Introduction

E-commerce, which is shortened from internet commerce or “electronic commerce”, ” applies online transactions of goods and services and executes money and data via the internet. With the advantage of the mobile Internet era, more and more people have smartphones and can connect to the internet anytime and anywhere. Hence, the costumers of E-commerce can purchase goods and services anywhere on the international market. E-commerce can target customers all over the world and significantly eliminate geographical limitations. Additionally, E-commerce is cost-effective due to the fact that it requires lower investment or business capital as compared with traditional offline businesses. E-commerce has gradually changed human life in the last decades, and online shopping is becoming more and more popular [1]. E-commerce outgrows traditional offline business by 13 times from 2015 to 2018 [2] and supplies a huge amount of potential costumers for retail businesses. In 2018, E-commerce sales accounted for 11.9% of all retail sales worldwide and expected to reach 17.5% by 2021 [3].

A vast of literature have been devoted to investing how different features affect the merchandise sales [4]–[11] and how to forecast e-commerce sales [12]–[14]. Correlation analysis and Ordinary least squares (OLS) regression are utilized to investigate the relationship between E-commerce sales and online advertising, including search ads, classified ads, and display ads. Results indicate that search ads and classified ads positively influence European E-commerce sales, while display advertising negatively influences sales. Additionally, the observation time of goods and the participation of the product to users can also affect the sales of products [5]. Holiday season is also an important factor influencing the sales of goods [15], [16]. Chintagunta et al. found that the volume of e-word-of-mouth (eWOM) positively affects movie revenues. Searching information can also be a predictor of sales. The more the potential costumers searching for goods, the more attention the product received [7]. Studies also show that two-thirds of web shoppers abandoned their shopping carts before they finally made an online purchase [17]. Based on the 214 online shoppers survey, researchers found that four key factors of the Businesses-to-costumer (B2C) website, including information content, design, and security, would influence consumers' decisions. In contrast, security and privacy have a greater impact on consumers' purchase intentions [9]. Scholars found that online critics, comments and views plays an important role in predicting sales [18]–[21]. Convolutional neural networks are adopted to automatically extract the intrinsic links in the historical sales data of products as features to predict the total sales of goods in the next week [13]. In Ref [14], a deep learning model(ANN) that considers the L_1 regularization achieved a sale forecasting accuracy rate of 86%.

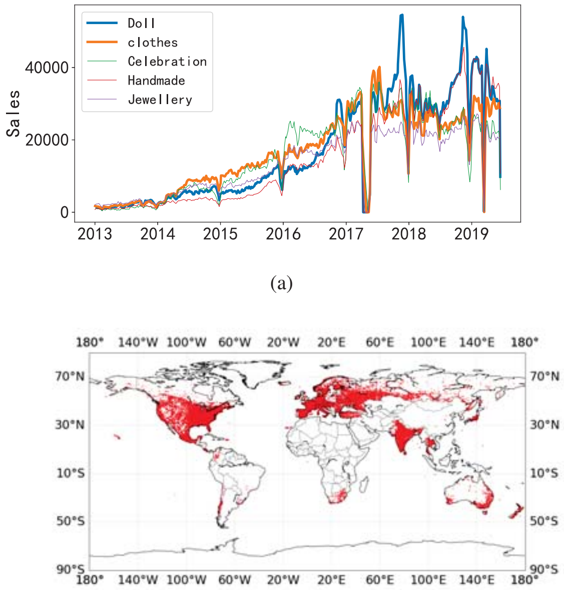
An e-commerce company provides our data with several hundred accounts on eBay. The period of the dataset is from January 1, 2013 to June 18, 2019. There are 75,680,610 orders, 802,110 kinds, and 31 categories of goods. There is a period of data missing in a few months in 2017 and 2019. Each order contains the basic information of the user's purchase of goods, mailing information, and user's consumption information. After 2017, the company's development has gradually stabilized. Based on these data, we investigate the factors that influence the sales of products. Additionally, we utilized machine learning methods to develop prediction models to optimize the procurement strategy, which is important for E-commerce companies. A good procurement strategy or model can reduce the procurement costs of a company and the operating cost. In this paper, due to the large time span of the data, we only choose a slice of data in a short period for prediction. We use ensemble learning algorithms to predict the cumulative sales of the goods in the later period, including GBDT , Catboost, Xgboost, and Lightgbm.

II. Online shopping data

An e-commerce company provides the dataset utilized in our paper with annual sales of more than 200 million US dollars. This company has registered several hundreds of accounts on eBay to sell more than 802,110 different products, which can be generally classified into 31 categories. Figure 1(a) shows the sales of several types of lady's goods of the e-commerce company. Figure 1(b) shows the distribution of order receiving locations in this period, while the British postcode data is missing. Because this is based on the user's mailing information, it isn't easy to search the order receiving location information. There will be some errors, but it does not affect the accuracy of these data.

III. Methodology

Ensemble learning methods in machine learning combine the decisions from multiple learning algorithms to form strong learners to improve the overall performance. Hence, the performance of ensemble learning methods is always better than performance obtained from any of the constituent learning algorithm alone. Here we use GBDT (Gradient Boosting Decision Tree), Catboost (Categorical Boosting), Xgboost(eXtreme Gradient Boosting), and Lightgbm (Light Gradient Boosting Machine) models for predicting online sales.



(a)

(b)

Figure 1. (a) Monthly sales of several goods from 2013 to June 2019; (b)Location map of consumers from 2017 to 2018.

Table I
Features Information

Feature	Introducion
Accumulated Sales($x_s(t)$)	$\sum_{j=1}^{t}$ before t day's sales
Sales revenue($x_s(t)$)	$\sum_{j=1}^{t}$ before t day's sales revenue
Average Price($x_a(t)$)	$\frac{x_s(t)}{x_q(t)}$

First of all, we input data, e-commerce sales data. After getting the new data $X(t)$and $Y(i)$through calculation and screening, we standardize the data of $X(t)$and finally get the reconstructed data. Then, we cut the data set into a training set and test set, use the training set to train the model, and use the test set to evaluate the performance of the model and output the prediction data $\hat{Y}(i)$. The detailed algorithm and steps can be found in Algorithm 1 and Figure III.

IV. Data Pretreatment

Here, we adopt the sales of 433,535 goods, including 22,498,884 orders from Oct. 2017 to Jan. 2019, to develop forecasting algorithms. Figure 3 shows the probability density distribution of cumulative sales of the adopted 433,535 goods. During this period, the cumulative sales of most products (non-best-selling products) were less than 1,000 pieces, and daily sales profiles are irregular. Only 3,675 of

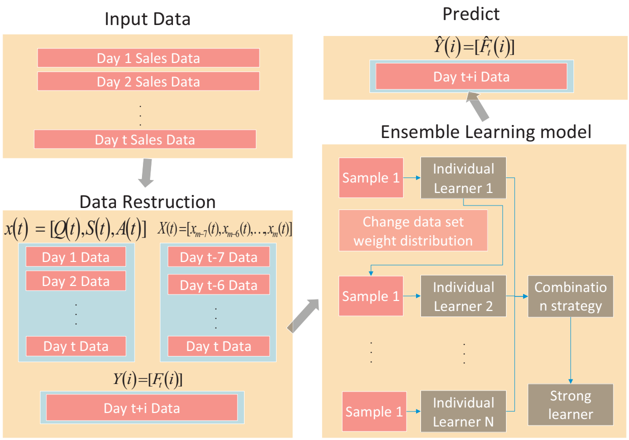


Figure 2. Framework of methodology

```
Algorithm 1 Algorithm for forecasting e-commerce sales through ensemble model
Input: Historical training dataset.
Output: The cumulative sales of the next $i$ day $\hat{Y}(i)$.
function DATA RECONSTRUCTION PROCESS($t$,$i$)
    Combined with the data of the $t$ day previous and the next $i$ day, the accumulated sales $x_c(t)$, sales revenue $x_s(t)$, average price $x_a(t)$ of the $t$ day previous and the cumulative sales $Y(t)$ of the next week to form the training dataset,
        $x(t) = [x_c(t), x_s(t), x_a(t)],$
        $Y(t) = \sum_{i=1}^{7} x_c(t+i-1).$    (1)
    Time serialization of $X(t)$ for $M$ days,
        $X(t) = [x(t-M), x(t-M+1), \dots, x(t-1)].$    (2)
    Data standardization for $X(t)$;
    return $[X(t), Y(t)]$;
end function

function PREDICTION PROCESS(TrainData,TestData)
    Input TrainData,TestData;
    Model training and evaluation;
    Calculate the cumulative sales of the next 7 days,
        $\hat{Y}(i) = [\hat{F}_t(i)]$    (3)
    return $\hat{Y}(i)$
end function
```

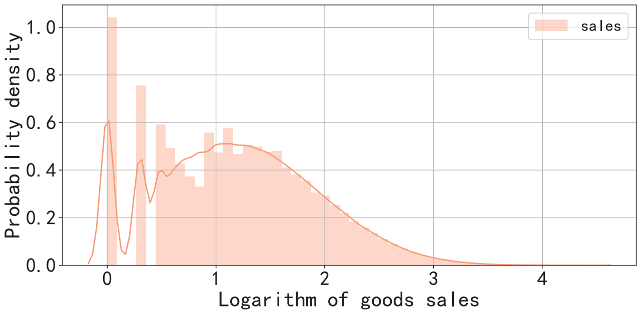


Figure 3. Probability density of goods

the products have accumulated sales exceeding 1,000 pieces, accounting for 24.8% of the total. In this study, we develop forecasting methods to predict the sales of hot-selling goods.

Then, we attempt to utilized historical sales data to predict the coming week's cumulative sales. In other words, the prediction of

$$\begin{aligned} y(t_i) &= \sum_{t=t_i}^{t_i+6} x(t_i) \\ &= f\,(x(t_i-M), x(t_i-M+1), \cdots .x(t_i-1), \\ &\qquad x_p(t_i-M), x_p(t_i-M+1), \cdots, x_p(t_i-1),) \end{aligned} \tag{4}$$

where $x(t_i)$ represents the daily sales of a good; $x_p(t_i)$stands for the price of the good; $y(t_i)$ is the cumulative sales in the next week. The prediction problem is

$$\min \sum_{j=M}^{K} \|y(t_i) - \hat{y}(t_i)\|_2^2. \tag{5}$$

Find an optimal model will be our major work.

V. Evaluation And Regression Result

Table II
THE AVERAGE VALUE OF THE EVALUATION INDEX OF THE MODEL IN THE SITUATION WITH $M = 14$.

Type	Model	MAE	MSE	RMSE	R^2
$Y(i=7)$	CatBoost	10.646226	302.096289	17.380678	0.595281
	GBDT	10.519757	309.019032	17.576578	0.586327
	LightGBM	11.514019	350.460613	18.473398	0.530851
	XGBoost	10.607647	315.147400	17.751255	0.578123
$Y(i=14)$	CatBoost	21.617309	1234.067940	35.128471	0.543462
	GBDT	21.289542	1219.863209	34.921631	0.548717
	LightGBM	21.695061	1261.058138	35.507385	0.533478
	XGBoost	21.435667	1246.917471	35.307231	0.538709

We adopt MAE, MSE, RMSE, R-square(R^2) as indicators to evaluate the performance of the model. Here, $y_i$is the real value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the average of the data. Models with various hyperparameters are adopted. Table II shows the average value generated by models with various hyperparamters.

Table III
THE BEST VALUE OF THE EVALUATION INDEX OF THE MODEL IN THE SITUATION WITH $M = 14$.

Type	Model	MAE	MSE	RMSE	R^2
$Y(i = 7)$	Catboost	10.627013	297.294175	17.242221	0.602023
	GBDT	10.659582	321.570968	17.932400	0.574683
	LightGBM	10.607454	298.678476	17.282317	0.600170
	XGBoost	10.288501	300.984801	17.348914	0.597082
$Y(i = 14)$	Catboost	21.612092	1212.889122	34.826558	0.551297
	GBDT	21.750402	1236.241194	35.160222	0.517874
	LightGBM	21.571649	1216.740462	34.881807	0.549873
	XGBoost	20.413600	1164.846735	34.129851	0.569070

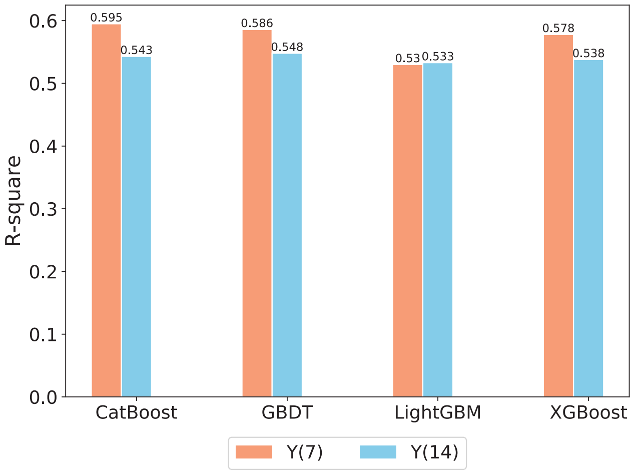


Figure 4. Evaluation of R-square

The experiment chooses to forecast the sales of commodities under the company's stable development and reconstruct the data. After reconstruction, Pearson correlation analysis is used to analyze the previous few days' data and the cumulative sales of commodities in the next few days. The features with the most considerable correlation are adopted for training prediction models.

The hyperparameters of the model are set by permutation and combination. After many experiments, combined with all the models trained, Table II shows the average value of all models, while Table III shows the best performance of each model. Experimental results indicate that the predictive performance of ensemble learning methods is not satisfied. Figure 4 shows that the predictive performance with $Y(i = 7)$ is slightly better than $Y(i = 14)$. We found that CatBoost has the best performance and the most stable performance. Table II and III shows The average value of R^2 is 0.595, the best is 0.602, and it keeps relatively low level in other evaluation indexes. The experimental results show that although the performance of the ensemble learning algorithms is not very good, it is stable on the whole. For the goods with high sales, the forecast will be more accurate, while for the goods with relatively low sales, the sales profile is irregular, which is not easy to predict.

VI. CONCLUSION

This paper analyzes more than 75,000,000 transactions of one E-commerce company from Jan. 2013 to June 2019. Additionally, we use ensemble learning algorithms to forecast the sales of e-commerce. The performance of the Catboost model is best, and the evaluation index R^2 reaches 0.595. With the rapid development of the times, users' online shopping has become a habit. It is an opportunity and challenge for e-commerce, especially for large-scale e-commerce. It is also an important part of the development of e-commerce, whether the sales of e-commerce can be accurately predicted through the consumption data and behavioral data of users. We will collect user's comment data, user click-through rate, collection status, and other user behavior data in future work. These characteristics are closely related to the sales of goods and play an essential role in predicting goods. Additionally, we can classify similar goods into one category for analysis, and then use deep learning methods for prediction.

ACKNOWLEDGMENT

This work was supported by National Science Foundation of China Project 61703355 and Science and Technology Program of Guangzhou, China (201904010224 and 201804010292), and Natural Science Foundation of Guangdong Province, China (2020A1515010761).

REFERENCES

[1] K Das and Afreen Ara. Growth of e-commerce in india. *Growth, available at: http://ijcem. in/wp-content/uploads/2015/08/Growth_of_E_Commerce_ in_India. pdf (accessed 9 December 2015)*, 2015.

[2] Azizi Othman. National ecommerce strategic roadmap overview, 12 2018.

[3] Statista. E-commerce share of total global retail sales from 2015 to 2023, 2020.

[4] Osama Harfoushi, Bader Alfawwaz, Bader Obeidat, Ruba Obiedat, Hossam Faris, et al. *Journal of Software Engineering and Applications*, 6(11):564, 2013.

[5] Lifang Peng, Weiguo Zhang, Xiaorong Wang, and Shuyi Liang. Moderating effects of time pressure on the relationship between perceived value and purchase intention in social e-commerce sales promotion: Considering the impact of product involvement. *Information & Management*, 56(2):317–328, 2019.

[6] Wenjing Duan, Bin Gu, and Andrew B. Whinston. Do online reviews matter? — an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008.

[7] Hui Yuan, Wei Xu, and Mingming Wang. Can online user behavior improve the performance of sales prediction in e-commerce? In *2014 IEEE International Conference on Systems, Man and Cybernetics - SMC*, 2014.

[8] Samuel W. K. Chan and James Franklin. A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1):189–198, 2012.

[9] C. Ranganathan and Shobha Ganapathy. Key dimensions of business-to-consumer web sites. *Information & Management*, 39(6):457 – 465, 2002.

[10] Choujun Zhan and K Tse Chi. A model for growth of markets of products or services having hierarchical dependence. *IEEE Transactions on Network Science and Engineering*, 6(3):198–209, 2018.

[11] Choujun Zhan, Bing Li, Xiaoting Zhong, Hu Min, and Zhengdong Wu. A model for collective behaviour propagation: a case study of video game industry. *Neural Computing and Applications*, 32(9):4507–4517, 2020.

[12] Pablo M Pincheira and Nicolas Hardy. Forecasting base metal prices with commodity currencies. *Available at SSRN 3095448*, 2018.

[13] Kui Zhao and Can Wang. Sales forecast in e-commerce using convolutional neural network. *arXiv preprint arXiv:1708.07946*, 2017.

[14] Yuta Kaneko and Katsutoshi Yada. A deep learning approach for the prediction of retail store sales. In *IEEE International Conference on Data Mining Workshops*, 2016.

[15] Choujun Zhan, Fujian Wu, Zhenhua Huang, Wei Jiang, and Qizhi Zhang. Analysis of collective action propagation with multiple recurrences. *Neural Computing and Applications*, pages 1–14, 2020.

[16] Quansi Wen, Choujun Zhan, Ying Gao, Xiping Hu, Edith Ngai, and Bin Hu. Modeling human activity with seasonality bursty dynamics. *IEEE Transactions on Industrial Informatics*, 16(2):1130–1139, 2019.

[17] Ming Cheung. Copyright challenges facing the website design industry: A survey with creative directors in hong kong. *The Design Journal*, 17(2):291–313, 2014.

[18] Hui Yuan, Wei Xu, Qian Li, and Raymond Lau. Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*, 270(1-2):553–576, 2018.

[19] Joonhyuk Yang, Wonjoon Kim, Naveen Amblee, and Jaeseung Jeong. The heterogeneous effect of wom on product sales: why the effect of wom valence is mixed? *European Journal of Marketing*, 2012.

[20] Geng Cui, Hon-Kwong Lui, and Xiaoning Guo. The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, 17(1):39–58, 2012.

[21] Christy M.K. Cheung and Dimple R. Thadani. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems*, 54(1):461 – 470, 2012.