E-commerce Sales Prediction using Machine Learning Techniques

Vijayshri Khedkar[1], Anuja Bokhare[2], Sileshi Girmaw Miretie[3], Tushar Laad[4]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University),Pune, India,vijayshrik17@gmail.com

[2]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University),Atur Centre, Gokhale Cross Road, -Model Colony, Pune, India, anuja.bokhare@gmail.com

[3]Department of Computer Science Kotebe University of Education. Addis Ababa ,Ethiopia, sgirmaw@gmail.com

[4]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University),Atur Centre, Gokhale Cross Road, -Model Colony, Pune, India, thl2021115@sicsr.ac.in

Abstract.

In the current digital world, people are using online shopping for purchasing goods extensively with first preference. Reviews and ratings on the sites will play a vital role in improving global communications among the customers and it has the potential to influence consumer buying patterns as well. Many E-commerce giants like Amazon, Flipkart are providing a platform that empowers users to share their real-time experiences and provide real-time insights about the performance of the product to future buyers. Sentimental Analysis and classification of the reviews as positive and negative will help us in understanding the voice of the consumers. So current study works on the reviews whose ratings are not provided by the customers. Machine learning models like Naïve Bayes, Multinomial NB, Bernoulli NB and Logistic Regression are used during the study. The result predicted the sentiments of those reviews.

Keywords. Machine Learning; Sentiment Analysis; Naïve Bayes classifiers; Feature extraction; Predictions.

1. INTRODUCTION

Everyone have arrived on the e-commerce globe. Users frequently come across a plethora of different retailers on the internet. Inventors made it possible to trade with anyone and from anywhere. User can purchase items without leaving the house, compare prices in other stores in seconds, and see what user want rather than settling for the first or least appropriate supply. It would be highly eye-catching to appear in this environment through the data it generates. Prediction or forecasting of the sales of E-commerce Company is very crucial now days. The prediction gives directions to especially smaller companies to increase their sales and reach the desired target. Amazon, for example, will continue to aggressively grab market share and outperform many small e-commerce firms. Smaller e-commerce firms are

among the losers, partially because they lack more sophisticated ways of reaching out to potential customers. Larger companies have recently begun to use advanced data analytics, machine learning, and artificial intelligence to predict the sales [1]. Also the sales prediction is helpful in inventory management. The prediction of product sales is an important component in inventory optimization. Because, as each one know, some e-commerce businesses have its own exclusive products that they sell online. As a result, that type of E-commerce platform must usually maintain a close eye on their inventory [2]. Sales prediction has some benefits like allocation of budget, setting the goals, targeting audience, assessing sales performance and many more. Similarly, in any E-commerce Company, the forecasting or prediction of sales plays a major role, as it shows the performance of the sales department; it will help in handling the budget and taking decisions according to it; to cornerstone the target audience etc. So, this paper focus on experiment with e-commerce figures and checks out to gain a better understanding of it.

2. PREVIOUS STUDY

The global e-commerce market is expected to reach $66,932.1 billion by 2030, growing at a rate of 13.5 percent per year from 2020 to 2030, owing to increased online buying and digital transactions in the wake of the COVID-19 epidemic. Value e-commerce in India, which is currently valued at $4 billion, is predicted to rise significantly, reaching $20 billion by 2026 and $40 billion by 2030, a 10x increase in ten years, according to the report. In E-commerce, sales prediction is a necessary process that has a significant influence in creating educated business decisions. It can assist us in managing our personnel, cash flow, and resources, as well as optimizing manufacturer supply chains, among other things. Sales prediction is a difficult subject because sales are influenced by a variety of factors such as promotional activities, pricing adjustments, and user preferences, among others [3]. In addition to linear regression, random forest, and decision tree, a study was conducted on sales prediction utilizing diverse models such as artificial neural networks and long short-term memory approaches [4]. The K-means algorithm, the Market Basket model, and the Vector Distance model is used in another study to establish a collection of important variables that would reflect group qualities. The silhouette index was calculated to assess whether or not these clusters are compact [5, 11, 12]. A study demonstrates the flaws and problems of standard online purchasing behaviour prediction approaches and offers a network shopping behaviour analysis and prediction system [6]. The authors of a research looked at the problem of demand forecasting on an e-commerce website and concluded that their method will predict significantly better when more data is used. Because the difference between the suggested model and random forest is not statistically significant, the proposed method can be utilized to forecast demand due to its accuracy with fewer data [7]. The ability to predict real-time, hourly order arrivals has been inadequate [8]. Prediction is more of a regression problem than a time series problem. It has been predicted that the regression methodology trends in historical data would repeat themselves in the future. Lasso regression may be employed in the future [9]. The researchers came to the conclusion that in order for businesses to handle massive amounts of data, they need an intelligent sales predictive model [10].

3. METHODOLOGY

The purpose of this study is to discover that surprises and gifts sell diversely depending on the season: peak sales, then a sharp dip the next year, then a steady increase until the next peak. Here authors have use Grid Search and Cross-Validation to test Linear Regression, Decision Tree and Random Forest Regression as shown in the Figure 1. Authors had calculated the mean square error and mean absolute error, then compare its results to see which one best fits the regression.

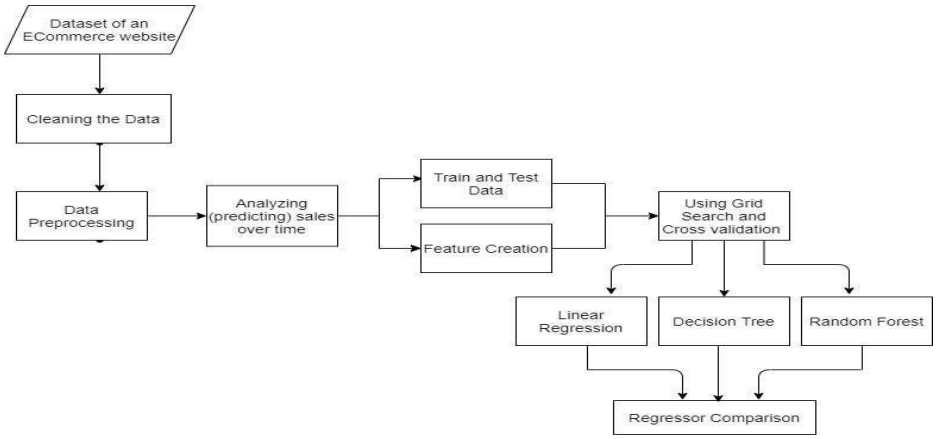

Figure 1.E-commerce Sales Prediction Model

4. DATASET AND EXPERIMENT DISCUSSION

The dataset which has been taken for this study has 8 columns InvoiceNo, StockCode, description, Quantity, Invoice Date, UnitPrice, CustomerID, Country. Pandas, numpy, matplotlib, scikit-learn and seaborn python libraries are used in this study. Using pandas dataframe, the records were read and added to the dataframe. After that the data was visualized using seaborn.

4.1 Experiment Steps

To generate fascinating visual findings, authors begin by setting up our environment and loading relevant libraries such as numpy and pandas. We also require data visualisation tools matplotlib and seaborn.

1. Read *csv()* method used to read our data. Then, for visuals and numerical, author conducted exploratory data analysis. A quick statistical overview is done to see negative quantities and unit prices.
2. Dealing with the data types and null values. Examining the columns one by one to determine the buying patterns.
3. Using plots and calculations, it was discovered that the vast majority of sales were made in the United Kingdom, with only 8.49 percent going abroad.
4. Detecting the outliers by plotting scatter plot and removing it.
5. After cleaning, removing invalid records and removing outliers visually check the distribution of numeric features.
6. Analysing the sales over time by resampling time data, and observing the patterns.

7. Creating features like quantity per invoice, quantity range, price range and month to improve the data for modelling.
8. Scaling 'QuantityInv' that is quantity per invoice feature to bring it in the range 0-1 like others. Splitting the data into train and test data.
9. Testing and validating three types of regressors: Linear, DecisionTree and RandomForest by using GridSearch and CrossValidation.
10. Create a bar plot for comparing the three types of regressors.

5. RESULT ANALYSIS

In Figure 2 a graph can be seen. The graph shows the items that were bought more often. Here White Hanging T-Light Holder is bought more often. Regency Cake stand 3 Tier and Jumbo Bag Red Retro spot are also bought often. The count for these three of the things is more than 2000. Also, Jam Making Set with Jars is the item which has not bought more often as compare to the others. Same is the case with Natural Slate Heart Chalkboard, Postage, Jumbo Bag Pink Polka dot and Heart of Wicker Small, that these are also not bought often.

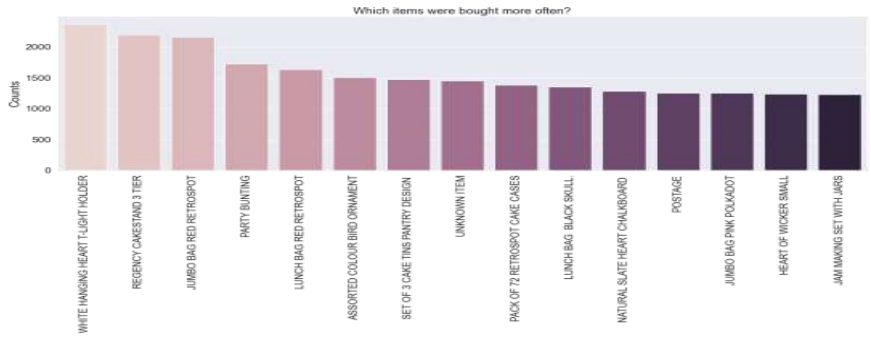

Figure 2. Items bought more often

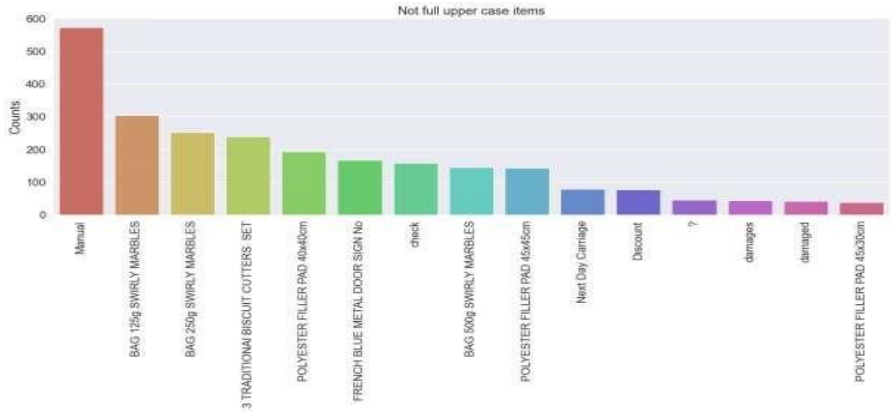

Figure 3. Not full upper case items

Figure 3 indicates the items which are not full upper case. Checking the case of letters in a sentence is a good idea. According to the description, certain units have lower case letters

in its names, and lower case records are for cancelled products. Authors can see that data management in the store can be enhanced in this case.

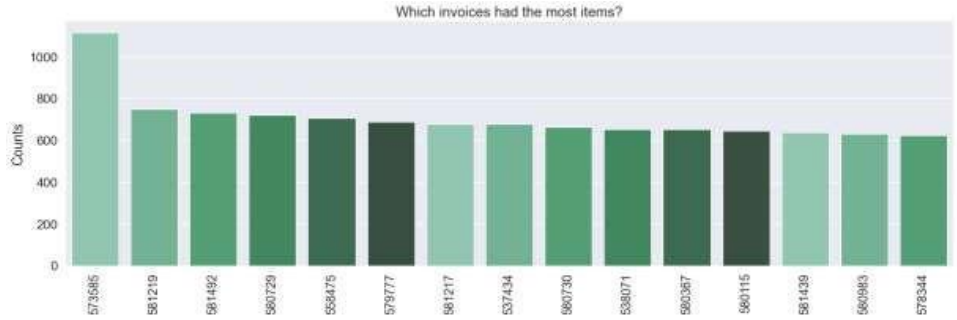

Figure 4. Invoice having four items

The figure 4 shows which invoices had the most items. Here authors can see that Invoice no. 573585 has most items. The items for Invoice no. 573585 have exceeded 1000 counts as shown in the graph. While looking at Figure 5 we can look at stock codes. Authors can see that 85123A has exceeded the 2000 counts. It appears that they are strongly linked to descriptions, that is Figure 1, which makes sense.

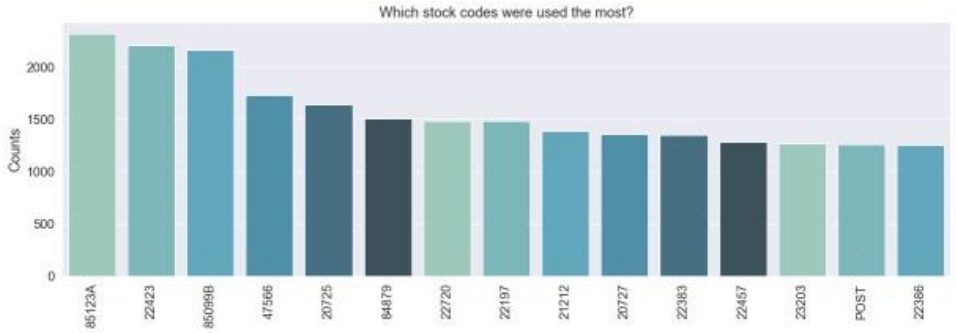

Figure 5. Most used stock codes

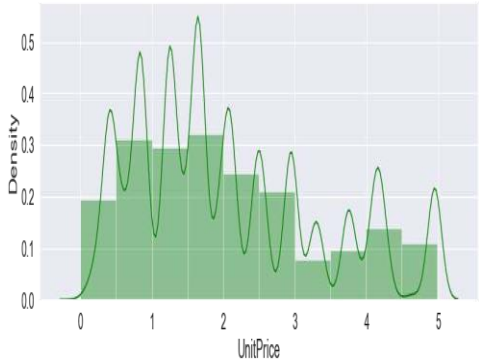

Figure 6. UnitPrice plot

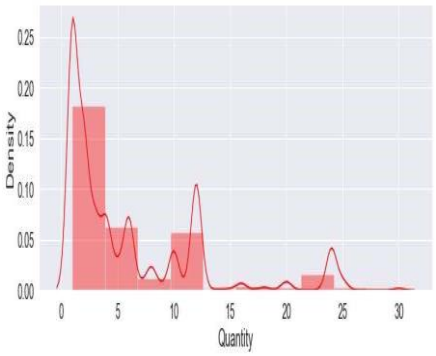

Figure 7. Quantity plot

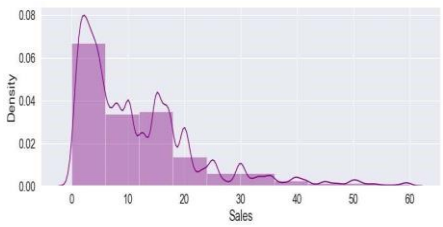

Figure 8. Sales plot

Figure 9. Sales over time

Authors are visually checking the numeric features distribution in the Figure 6, 7 and 8. Figure 6 shows unit price dist plot where authors can see that the majority of the things sold at this store are inexpensive, ranging from 0 to 3 pounds. Figure 7 shows that customers typically purchase 1-5 products. In the case of a sale, perhaps 10-12 items. Although, figure 8 shows that the majority of sales per order were between 1 and 15 pounds each.

Table 1. Sales as per Invoice Date

Invoice Date	Sales
12/5/2010	185427.8
12/12/2010	329936.8
12/19/2010	216012.2
12/26/2010	92369.3
1/2/2011	0
1/9/2011	133658.2
1/16/2011	193362
1/23/2011	138349.6
1/30/2011	125170.8

Table 2. Statistical Measures

	Linear Regression	Decision Tree	Random Forest
Best Score	0.1742	0.3727	0.4400
Mean Absolute Error	15.1135	6.7723	6.7276
Mean Squared Error	3918.8825	2101.3317	1951.5702
R2 score	0.15630	0.5476	**0.5798**

Figure 9 shows the analysis of sales over time. Authors can see here that in particular week of month January the sales look 0. The sales in first week of December have increasing rapidly from 200,000 to around 340,000 pounds. However, suddenly after that it reaching towards the month of January the sales decreased and became 0. After that it again started increasing and went on increasing till December next year. Well, Table 1 confirms the fact that there are 0 sales in January.

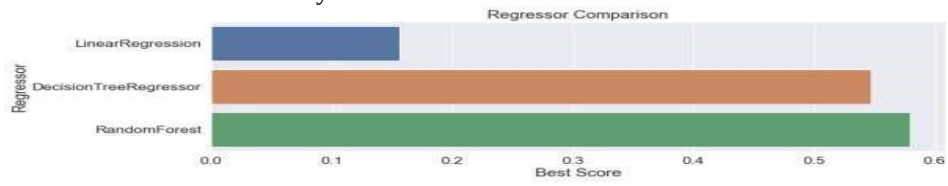

Figure 10. Comparing the Regressor

As shown in Figure 10 and Table 2 authors can observe that the Best score for Random Forest is high compare to Decision Tree Regressor and Linear Regression.

6. CONCLUSION

The sales were predicted using linear regression, decision trees, and random forests. With linear regression, the mean squared error was 0.15, and the coefficient of determination was 1. Linear Regression's fundamental flaw is that it only considers the dependent variable's mean. This is a simple analysis of a transaction dataset using a sales prediction model. It is observed that a sales high, followed by a sharp drop the next year in January, and then a steady increase until the next year's end. Authors were able to successfully estimate E-commerce sales using machine-learning algorithms, and observed that Random is the greatest fit for prediction. The best score obtained with Random Forest is 0.6. As a result, for future analysis, random forest must be applied rather than linear regression or decision tree.

7. FUTURE SCOPE

In future work, it would be good to focus on the variable choices rather than the algorithm. Many aspects can still be improved, such as cluster analysis and feature creation.

8. REFERENCES

[1] Thobani, Shaheen. Improving e-Commerce sales using machine learning. PhD diss., Massachusetts Institute of Technology, 2018.

[2] Li, Jiahua, Tao Wang, Zhengshi Chen, and Guoqiang Luo. Machine learning algorithm generated sales prediction for inventory optimization in cross-border E-commerce. International Journal of Frontiers in Engineering Technology 1, no. 1 (2019).

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

[4] Dong, Wenxiang, Qingming Li, and H. Vicky Zhao. Statistical and Machine Learning based E-commerce Sales Forecasting. In Proceedings of the 4th International Conference on Crowd Science and Engineering, pp. 110-117. 2019.

[5] Janićijević, Stefana, Đorđe Petrović, and Miodrag Stefanović. Sales prediction on e-commerce platform, by using data mining model. Serbian Journal of Engineering Management 5, no. 2 (2020): 60-76.

[6] Liu, Cheng-Ju, Tien-Shou Huang, Ping-Tsan Ho, Jui-Chan Huang, and Ching-Tang Hsieh.Machine learning-based e-commerce platform repurchase customer prediction model. Plos one 15, no. 12 (2020): e0243105.

[7] Chandel, Archisha, Akanksha Dubey, Saurabh Dhawale, and Madhuri Ghuge. Sales prediction system using machine learning. Int. J. Sci. Res. Eng. Dev 2, no. 2 (2019): 667-670.

[8] Leung, Ka Ho, Daniel Y. Mo, George TS Ho, Chun-Ho Wu, and George Q. Huang. Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology. Industrial Management & Data Systems (2020).

[9] Pavlyshenko, Bohdan M. Machine-learning models for sales time series forecasting. Data 4, no. 1 (2019): 15.

[10] Cheriyan, Sunitha, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa. Intelligent Sales Prediction Using Machine Learning Techniques. In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58. IEEE, 2018.

[11] Ingale, V. R., & Phursule, R. N. Product Feature-based Ratings for Opinion Summarization of E-Commerce Feedback Comments. International Journal of Computer Applications, 975, 8887.

[12] Ingale, V. R., & Phursule, R. N. (2014). Sentiment Analysis by Visual Inspection of User Data from Social Sites-A Review on Opinion Mining. International Journal of Science and Research, 3(12), 2188-2191

Biographies

Vijayshri Khedkar is an Assistant Professor working at Symbiosis Institute of Technology and skilled in NLP, Applied Machine Learning, Data Analytics, Information Retrieval & Deep Learning. A life-long learner with a strong educational background holding two Master's Degrees (M.B.A. & M.E.) and pursuing Ph.D. in Computer Engineering (NLP) from Symbiosis International University, India. She is a life member of IAENG and Senior IEEE member. She has published around 30 papers at various reputed conferences and journals. She is active reviewer for IJECE journal indexed in Scopus and reviewed papers for various conferences and journals.

Dr. Anuja Bokhare is working as Assistant Professor in the department of Computer Science at Symbiosis Institute of Computer Studies and Research, Pune, Maharashtra India. She received M.Phil. (Computer Science) at Y.C.M.O.U, Nasik, India and completed PhD from Symbiosis International (Deemed University) in faculty of Computer Studies. She has 20 years of experience in the field of academics. Her research interest includes software engineering, applications of Artificial Intelligence, machine Learning and soft computing. She had published 22 research papers in international journal and conferences along with one book in her account.

Sileshi Girmaw Miretie has been a lecturer of Computer Science in Debre Markos University and Kotebe University of Education. Addis Ababa, Ethiopia since 2013. He has Master's in Computer Science and Engineering from Symbiosis International University, India in 2019.His research interests are, Natural Language Processing, data, mining, data science, machine learning, Artificial intelligence, Internet of things (IOT).

Tushar Laad is a final year Bachelor of Computer Application student at Symbiosis Institute of Computer Studies and Research. His electives during his course included Machine Learning and Data Visualization. His profound interest in Machine Learning is supported by his work as a Data Engineer at ABMFrogs Technologies Pvt. Ltd. His research interests include Data Mining, Data Scraping, and Data Analysis using Machine Learning. Apart from this, he has also established his startup "BuildMySite.in" which provides, Website Development Services.