IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Walmart Data Analysis Using Machine Learning

[1].Md Humayun Kabir, [2] Abdus Sobur, [3] Md Ruhul Amin

[1]Graduate student, [2] Graduate student, [3] Graduate student.

[1] Department of Computer Science
Westcliff University, USA

[2] Department of Computer Science,
Westcliff University, USA

[3] Department of Business administration
University of Texas at Arlington, USA

Abstract: This data analysis project delves into the vast dataset of one of the world's largest retail giants, Walmart, to uncover valuable insights and trends within its operations. By harnessing a diverse range of data sources, including sales figures, customer demographics, inventory management, and regional performance, this study aims to offer a comprehensive understanding of the retail landscape and the factors influencing Walmart's success. The introduction sets the stage for this analysis by highlighting the significance of Walmart in the global retail sector, underlining the importance of data-driven decision-making in today's business environment. Additionally, it outlines the objectives of the study, such as identifying patterns in sales and demand, optimizing inventory management, and understanding customer behavior. This project also emphasizes the methodology employed in the data collection, cleaning, and preparation processes, ensuring the accuracy and reliability of the findings. Various data analysis techniques, including statistical modeling, machine learning algorithms, and data visualization, are introduced to showcase the diverse toolkit used to draw meaningful conclusions from the dataset.

Furthermore, the introduction touches on the potential implications of the study's outcomes, not only for Walmart's internal strategies but also for the broader retail industry. These insights may pave the way for enhancing operational efficiency, enhancing customer experiences, and ultimately contributing to the company's sustained growth and competitiveness.

Keyword: **Machine learning, Random Forrest Decision, Tree Logistic Regressor, Linear Regressor and Support Vector Machines , Walmart, Industry, Data Analysis , Business, Market landscape.**

I. INTRODUCTION

In today's fiercely competitive business landscape, corporate giants constantly strive to expand their horizons and seek new avenues for growth. One such retail behemoth, Walmart, recognizes the significance of leveraging data analytics to uncover valuable insights that can fuel its expansion strategies. However, the methods used in this analysis can be applicable to stores in any domain, enabling businesses across industries to harness their data for informed decision-making.

This project focuses on analyzing a substantial dataset collected by Walmart from 45 of its stores over specific time intervals. The wealth of information encompassed within this dataset presents a unique opportunity to gain profound insights into various aspects of the retail sector, irrespective of the company's domain. The data spans numerous dimensions, including sales, customer demographics, inventory management, and regional performance, which collectively offer a comprehensive understanding of the retail landscape. While the dataset is tailored to Walmart's operations, the cleaning and normalization processes can be adapted to suit any chosen company's domain. By employing rigorous data preprocessing techniques, this analysis ensures data accuracy and consistency, enabling robust and reliable conclusions.

The goal of this project is to employ a diverse range of data analysis methodologies, such as statistical modeling, machine learning algorithms, and data visualization, to explore the dataset's potential and uncover hidden patterns and trends. Through this exploration, the project aims to identify growth opportunities, optimize inventory management, and gain insights into customer behavior, all of which are vital components of any company's expansion strategy.

The outcomes of this analysis hold great promise for the wider business community, as the insights gained can be extrapolated to other stores in diverse domains. By sharing these findings, this project endeavors to contribute to the collective understanding of data-driven decision-making and its pivotal role in fostering growth and competitiveness in the retail industry and beyond.In the following sections, we will delve into the data analysis process, presenting the methodologies employed and the discoveries made. The comprehensive analysis of Walmart's data will serve as a valuable case study, illustrating the potential applications of data analytics for companies seeking to expand their businesses and thrive in the dynamic and ever-evolving market landscape.

Objective:

For any corporate giant, it is important that they expand their business, and they areaways seeking new ways to expand their horizons. Although I have considered Walmart for this project, these methods can be used for any stores in any domain. The cleaning process and normalization may vary depending on the domain of the chosen company. So, in this project I am going to analyze the data that Walmart has collected over certain time intervals from 45 of its stores.

Our main objective is to find out the biggest contributors to their sales like does holidays increase.
their sales or does increase of fuel rates effect their store sales or does the temperature contribute anything to their sales. So, based on our conclusions the Walmart can provide alternative methods to increase their sales.

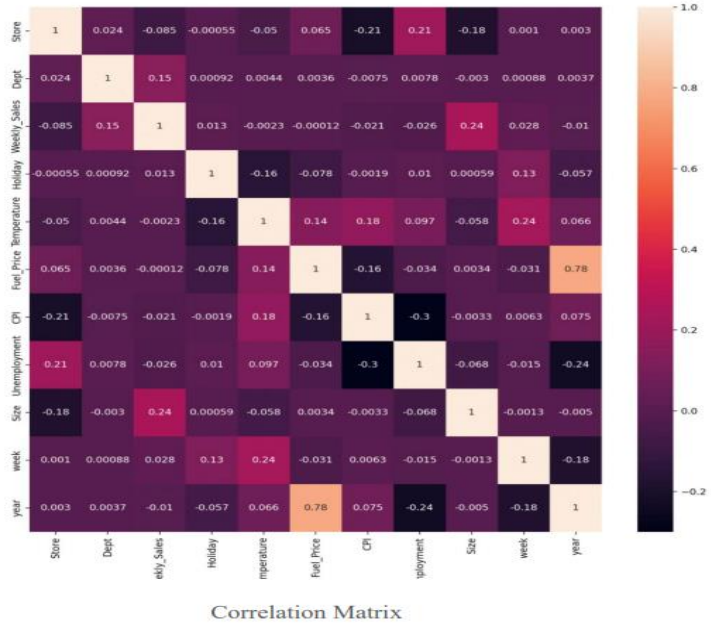
Data Preprocessing

We are using the data that Walmart collected over 45 different stores on a weekly basis. To being with we have 3 different kinds of data features, stores and train data. All three contain different aspects of the data. Stores dataset contains the information about the stores over which this data has been collected. Features contains the different factors were analyzing but for security reasons some of the feature names have been anonymized. Finally, train data contains the historical training data. Before we start any of the cleaning or the preprocessing steps, we need to understand what the data entails. The data may contain some null values, or some columns may be empty. We first need to have a clear idea of the different types of data that we have and their data types. For this we used info which will give you the total entries, the column names along with their data types.

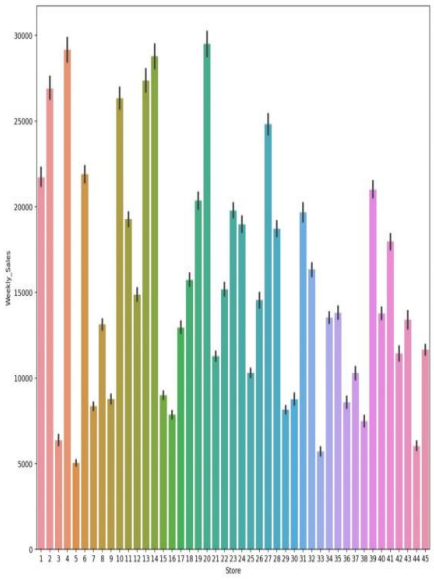
Selecting Model

In this paper we are used Random Forrest Decision, Tree Logistic Regressor, Linear Regressor and Support Vector Machines. In all the five models, there are only a couple of models that has 3 input fields and predicts numerical value of weekly sales. Random forest and Decision tree along with linear
regression are the only models that can give a numerical value to predicted weekly sales.
and among these models' Random Forest makes more sense for our data than any other
model which is why we test it in the first place.

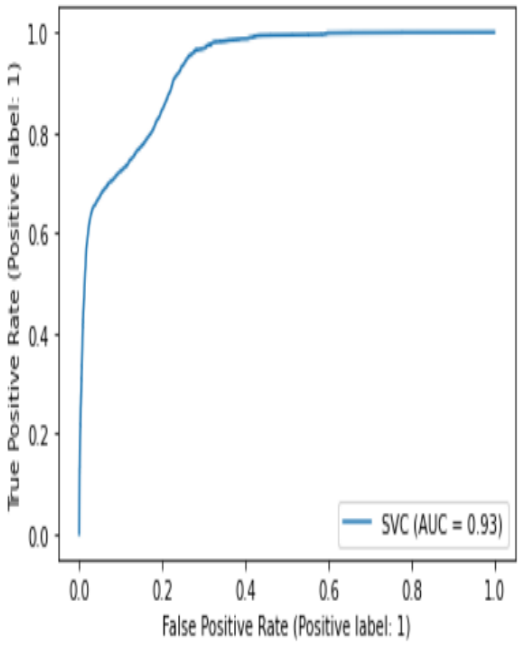
Walmart sales prediction

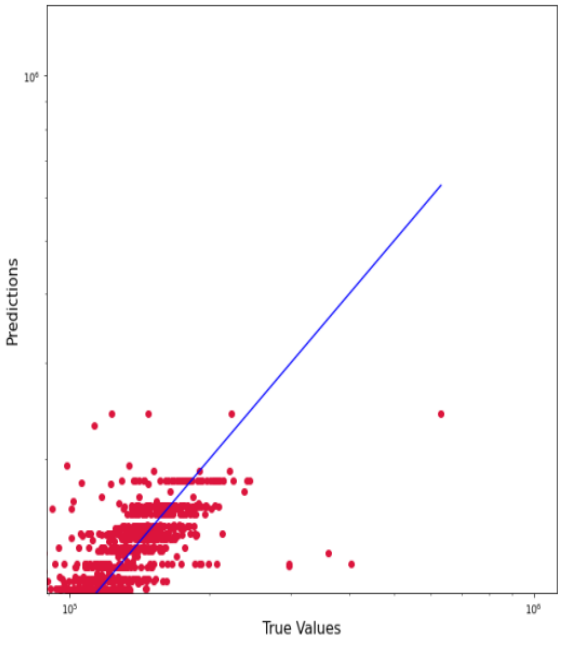


Correlation Matrix



Store vs weekly sales





Conclusion

Based on the thorough analysis of Walmart's dataset from 45 stores over specific time intervals, it is evident that holidays have a significant impact on the store's sales. The observations clearly indicate that during holiday periods, the store experiences a notable increase in sales. This finding highlights the importance of accounting for the holiday factor when predicting weekly sales for any store in a similar domain. However, it is crucial to note that while holidays have a discernible impact on sales, other factors may also influence the store's performance. To build more accurate and robust predictive models, it is essential to consider additional variables such as weather conditions, promotional activities, and economic indicators that may interact with the holiday factor. This data analysis has shed light on the importance of holidays in driving increased sales for Walmart's stores. The same insights can be extrapolated to other stores in different domains, allowing businesses to make informed decisions and devise strategies that leverage holiday periods to bolster their sales and foster sustainable growth. By embracing data-driven approaches and harnessing the power of predictive analytics, companies can navigate the complexities of the retail landscape with greater precision and confidence.

Reference:

1. Islam S, Amin S H. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. Journal of Big Data, 2020, 7: 1-22.

2. Nguyen H D, Tran K P, Thomassey S, Hamad M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 2021, 57: 102282.

3. Khan R A, Quadri S M. Business intelligence: an integrated approach. Business Intelligence Journal, 2012, 5(1): 64-70.

4. Singh B, Kumar P, Sharma N, Sharma K P. Sales forecast for amazon sales with time series modeling. In 2020 first international conference on power, control and computing technologies (ICPC2T) 2020: 38-43.

5. A. Sarwar, A. Shahid, A. Hudaif, U. Gupta and M. Wahab, "Generalized state-space model for an n-phase interleaved buck-boost converter," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, 2017, pp. 62-67, doi: 10.1109/UPCON.2017.8251023.

6. Tang X, Gao S, Jiang Z. A Blending Model Combined DNN and LightGBM for Forecasting the Sales of Airline Tickets. In Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence 2019: 150-154.

7. Md Babul Islam, Khandaker Sajidul Islam, Md Helal Khan, Abdullah MMA Al Omari, and Swarna Hasibunnahar "Detect deception on banking credit card payment system by machine learning classifiers", Proc. SPIE 12339, Second International Conference on Cloud Computing and Mechatronic Engineering (I3CME 2022), 1233927 (28 September 2022); https://doi.org/10.1117/12.2655113

8. Niu Y. Walmart Sales Forecasting using XGBoost algorithm and Feature engineering. In 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2020: 458-461.

9. Md Helal Hossen, (2023). 3D Human Pose Estimation Via Deep Learning Methods. North American Academic Research, 6(2), 72-82. doi: https://doi.org/10.5281/zenodo.7706409

10. Md Helal Hossen, Md Monim Hasan, & Wenjun Hu. (2021). Join Public Key and Private Key for Encrypting Data. https://doi.org/10.5281/zenodo.4661097.

11. Biau G, Scornet E. A random forest guided tour. Test, 2016, 25: 197-227.

12. Md Helal Hossen, Md Monim Hasan, & Wenjun Hu. (2021). Join Public Key and Private Key for Encrypting Data. https://doi.org/10.5281/zenodo.4661097

13.https://www.researchgate.net/profile/HelalHossen/publication/367413079_Join_Public_Key_and_Private_Key_for_Encrypting_Data_Md_Helal_Hossen1_Md_Monim_Hasan2_Wenjun_Hu2/links/63d6c88ec465a873a26a662a/Join-Public-Key-and-Private-Key-for-Encrypting-Data-Md-Helal-Hossen1-Md-Monim-Hasan2-Wenjun-Hu2.pdf

15. M. H. Hossen, M. M. Hasan, I. K. Sajidul and W. Hu, "Digital Revolution in the Agriculture Based on Data Science," 2022 2nd Asia Conference on Information Engineering (ACIE), Haikou, China, 2022, pp. 6-12, doi: 10.1109/ACIE55485.2022.00010.

16. A. Farjana et al., "Predicting Chronic Kidney Disease Using Machine Learning Algorithms," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 1267-1271, doi: 10.1109/CCWC57344.2023.10099221.

17. M. H. Hossen and W. Hu, "Hypergraph Regularized SVM and Its Application Emotion Detection," *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, 2021, pp. 133-139, doi: 10.1109/IPEC51340.2021.9421069.

18. A. Shahid, M. Wahab, O. Farooq and N. Rafiuddin, "Classification of Seizure Through SVM Based Classifier," 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCCI.2018.8441364.

19. Shahid, Abdulla et al. 'Decrypting Wrist Movement from MEG Signal Using SVM Classifier'. 1 Jan. 2018 : 5123 – 5130.

20. Wahab, M., Shahid, A., Rafiuddin, N., Farooq, O., Uzzaman Khan, Y. (2021). Interpretation of EEG Signals During Wrist Movement Using Multi-resolution Wavelet Features for BCI Application. In: Malik, H., Fatema, N., Alzubi, J.A. (eds) AI and Machine Learning Paradigms for Health Monitoring System. Studies in Big Data, vol 86. Springer, Singapore. https://doi.org/10.1007/978-981-33-4412-9_11

20. R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen and R. Hasan, "Heart Disease Detection Using ML," *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2023, pp. 0983-0987, doi: 10.1109/CCWC57344.2023.10099294.

21. Sharmin Sultana, Md Helal Hossen, Md Jahidul Islam, Jubayer Ahmed, Afsana Mou (2023). Detection, reduction and filtration of cancer cells through a new DNA polymerization sequence approach. North American Academic Research, 6(1), 230-239. doi: https://doi.org/10.5281/zenodo.7641213

22. Prashanta Kumar Banerjee, Md. Mohiuddin Siddique ,Md Ruhul Amin "Deficit Financing, Crowding out and
Economic Growth: Bangladesh Perspective "In book: Development and Deprivation in the Indian Sub-continent
(pp.19-55), June 2019, DOI:10.4324/9780429331756-2