Unveiling the future of retail with advanced Machine Learning models for sales prognostication - A case study on Walmart's dynamic dataset

Name: Mounika Marella

1. Abstract

In these modern retail corporations, it is the paramount concern to predict the sales with utmost accuracy to increase the company's profit and maintain effective inventory. So, in this paper, author proposed a comparative analysis among different machine learning algorithms like Linear and Ridge Regression, K-Nearest Neighbour, Decision Tree, Random Forest, XGBoost (Xtreme Gradient Boosting) and a deep learning algorithm LSTM (Long Short-Term Memory) along with feature engineering. Leveraging the Kaggle's Walmart dataset that contains different aspects like holidays dates, an unemployment rate, CPI and more along with weekly sales, an in-depth analysis is performed to provide valuable insights to the retailers that helps in making proper resource allocation and business plans. The methodology starts with data exploration, preprocessing, feature engineering, visualization, model testing with metrics accuracy and R2-score. The results showed that LSTM outperformed all other algorithms with an R2-score of "0.99" and stands as a best fit algorithm. With this research, the sales prediction will be more accurate and help in planning efficient retail operations for the maximum possible profit.

2. Introduction

The "Sales Prediction" is a dynamic and continuously evolving sector which needs a lot of research to increase the effectiveness in inventory management, decision-making, resource allocation, and profit maximization. The conventional approaches like using mathematics, statistics, and time series to analyze the sales data often fail to determine the insights. Whereas, with machine learning algorithms it is comparatively easy to handle the challenges with those large sales datasets to identify the insights and patterns to improve the precision of the Walmart sales forecasting by performing comparative analysis among different algorithms which is the objective of this research.

Many other researchers have explored different approaches in the concept of sales forecasting. Research paper by S. B. Laha et al. [1] focus on predicting the sales of Walmart's large datasets by utilizing the different regression machine learning (ML) algorithms, especially the advanced regressor "eXtreme Gradient Boosting (XGBoost or XGB)" in which XGBoost outperformed with an accuracy of 95.6 percent. In the similar context of accurate prediction to Walmart sales, X. dairu et al. [2] proposed a study on how the precise sales prediction helps for the optimized resource allocation and strategic decision-making in Walmart. The XGBoost surpassed other conventional regression techniques like Logistic and Ridge regressor with comparatively lower "Root Mean Square Scaled Error (RMMSE)" of "0.655". With the similar problem statement and Kaggle's Walmart sales dataset as X. dairu, author Y. Niu et al. [3] also used same Logistic and Ridge regressor's but with a bit difference in meticulous feature engineering that resulted in a even more lower "Root Mean Squared Scaled Error (RMSSE)" of "0.652". Also reviewed the study by H. Li et al. [4] that explored the role of ensemble learning technique XGB approach in improving the accuracy of credit rating. The authors show the advantages of the XGB model over Linear Regressor model.

The significance of this research is to optimize the inventory management, improve decision-making, increase the profit of retail corporations and providing customer satisfaction by exploring the different machine learning and deep learning techniques especially XGBoost and LSTM algorithms for precise sales forecasting for Walmart. This research contributes to the Walmart business by doing comparative study among the machine learning algorithms like Linear Regression, Ridge Regressor, DT, KNN, RF, XGB and the deep learning algorithm LSTM with feature engineering on a Walmart sales dataset containing 6000 records with multiple dependent numerical and categorical features that helps for better training and testing of models.

The scope of this study lies in improving the accuracy in predicting the Walmart weekly sales by doing data analytics on the Kaggle's Walmart sales dataset by mainly focusing on the XGBoost and LSTM algorithms. It also provided different insights and trends in the sales based on features like month, CPI, unemployment, temperature, and other features. Whereas the limitations of this study are making inferences solely on this dataset is not a good approach. In addition, there is a lack of interpretability of models along with inability to handle even larger or continuous data.
Navigating through the research, the methodology is meticulously outlined. Data preprocessing involves addressing

missing values, label encoding categorical variables, treating outliers, and normalizing data. The dataset's informative nature limits extensive feature engineering. Data visualization, executed through Seaborn and Matplotlib, aids in comprehending patterns and revealing insights into factors influencing credit approval.

This paper starts with brief overview on the research and contains background information and context of the research. Moreover, author have added objective of the research and its significance in answering the objective. In addition, author included the scope and limitation on high level. Next, it continued with methodology which covers aspects like data preprocessing that includes outlier handling, encoding categorical features, scaling the numerical features. Followed by data visualization that leverages matplotlib and seaborn to find insights and patterns of data. Finally, it ends with results, discussion, and conclusion of the study.

3. Methodology

3.1 Collecting and describing the dataset:

From the Kaggle website, the Walmart sales dataset is collected in the csv format using “read_csv” of pandas that has 6300 records and eight customer features. The list of features is "Store", "Date", "Weekly_Sales", "Holiday_Flag", "Temperature", "Fuel_Price", "CPI", "Unemployment". There is only one categorical data that is date and all other are numerical features. A unique identifier is assigned to each store and there are a total of 45, and there is a date for each sales record. In the “Weekly_Sales” column the total sales for a week is recorded in USD. The “Holiday_Flag” has a binary value as 0 or 1 which indicates “No holiday” and “Holiday” on that day. Then the “Fuel_Price” indicates the cost of fuel for that week, which potentially affects sales due to high transportation cost. The “CPI (Consumer Price Index)”, is the change in the price paid by urban consumers for goods and services, providing insights of economic situations. Finally, “Unemployment” is the rate of unemployment during that week which reduces sales.

3.2 Data Cleaning and Preprocessing:

Cleaning of dataset starts with a check for missing values, duplicate records, and outliers. There were no missing values and duplicated but has around 8 percent of outliers in which most of them are in “Weekly_Sales” and “Unemployment” that were handled using “Winsorization” because it will preserve the overall distribution of features and replace extremes with nearest extremes which is always valid. As part of data preprocessing, the author had performed feature engineering by adding more features like “Month”, “Year” and “Season”. Later as the data is not in same range which affects the model performance, it is scaled using “Standard Scaler” because the features followed normal distribution. The categorical features like “Store” and “Season” are encoded using “Binary Encoder”. Then the dataset is split into train and test set of 80 and 20 percent. Finally with all these steps data is ready to train to the models.

3.3 Data Visualization:

The author leveraged “matplotlib” and “seaborn” libraries for visualizing the data. Used “Boxplot” to identify the outliers and found in “Weekly_Sales” and “unemployment”. The count plot of stores shows that records were collected equally from each store. The bar plot visualization of “Month”, “Season”, “Year” features with respect to “Weekly_Sales” shows that winter season has highest sales. Whereas the scatter plot of “Unemployment” and “CPI” shows they are strongly related to “Weekly_Sales” and the same is shown from correlation heatmap. Moreover, the feature importance graph show that among all the features, "CPI", "Unemployment" and stores 1, 4, 0 are playing crucial role in predicting the "Weekly_Sales".

3.4 Selection of Models:

In this phase, multiple machine learning models like Linear Regression, Decision Tree, Ridge Regression, Random Forest, XGBoost, K Nearest Neighbor, and deep learning algorithm LSTM were considered for this comparative analysis. For Linear and Ridge Regressor, the parameter “polynomialfeatures__degree” is considered. Whereas KNN has ‘n_neighbors’, ‘weights’, ‘metric’, ‘algorithm’, and ‘leaf_size’. Then for DT ‘max_depth’, ‘min_samples_split’, ‘min_samples_leaf’. For the RF and XGBoost ‘n_estimators’ and ‘max_depth’ were considered. At first the author considered 5 parameters for DT, then it took a lot of time, so considered only main parameters for further models.

3.5 Performance Metrics for Models:

The author had considered mainly two metrics for all the machine learning models like Linear Regression, Random Forest, Ridge Regression, Decision Tree, K Nearest Neighbor, XGBoost. Whereas for the deep learning model LSTM, 'mean_squared_error (mse)' was considered and then calculated R2 score from that 'mse'. The R2-score is used to evaluate how well the regression models fit the dataset. As R2-score goes towards 1, then it shows that the model was a good fit for that data and vice versa if it goes towards 0. The accuracy is the closeness of predicted value to actual value. In simple words, it tells how well the model predicts the sales. A higher accurate model fits better than all other models. Whereas the MSE average squared difference between actual and predicted values. Lower MSE will be a good fit model.

4. Results and Discussion:

In this study, the performance of the models is calculated using metrics accuracy and R2-score. With a phenomenal R2-score of "0.99", LSTM outperformed all other machine learning models mentioned earlier. The LSTM R2-score says it is perfect fit for this Walmart sales dataset. XGBoost stands next to LSTM with an R2-score of "0.97" with the highest training and test accuracies compared to other ML models. Ridge and Linear regressors stands in third position with an R2-score of "0.96" and test accuracy of "97%", while the RF and DT stand next with test accuracy of "96%" and "94.5%" orderly.

Table 1: Performance metrics of all the algorithms

Algorithm (model)	R-Squared score	Training data accuracy (%)	Test data accuracy (%)
KNN	0.92	100%	94.5%
Decision Tree	0.955	98%	95.3%
Random Forest	0.963	99%	95.9%
Linear Regression	0.969	98%	97.03%
Ridge Regressor	0.969	98%	97.03%
XGBoost	0.972	99%	96.91%
LSTM	0.99	-	-

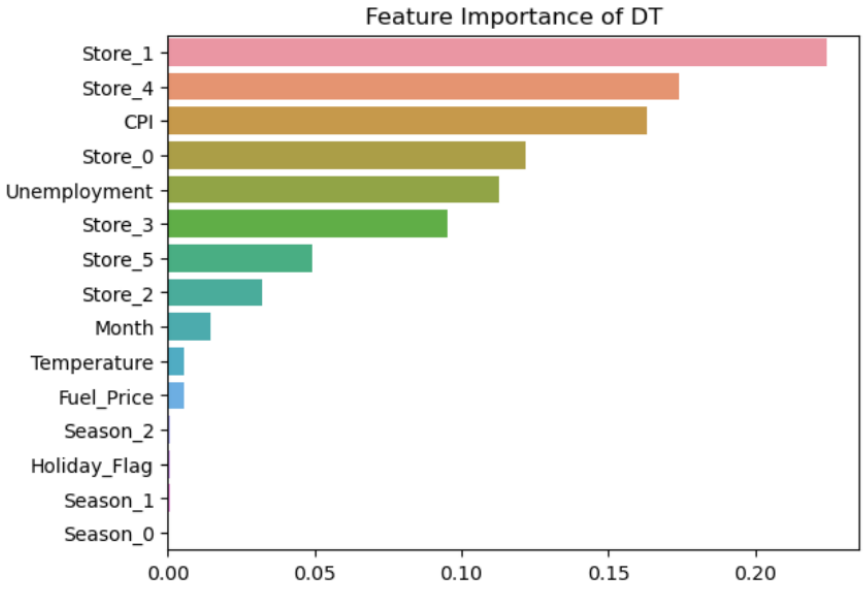

Figure 1: Feature importance of DT

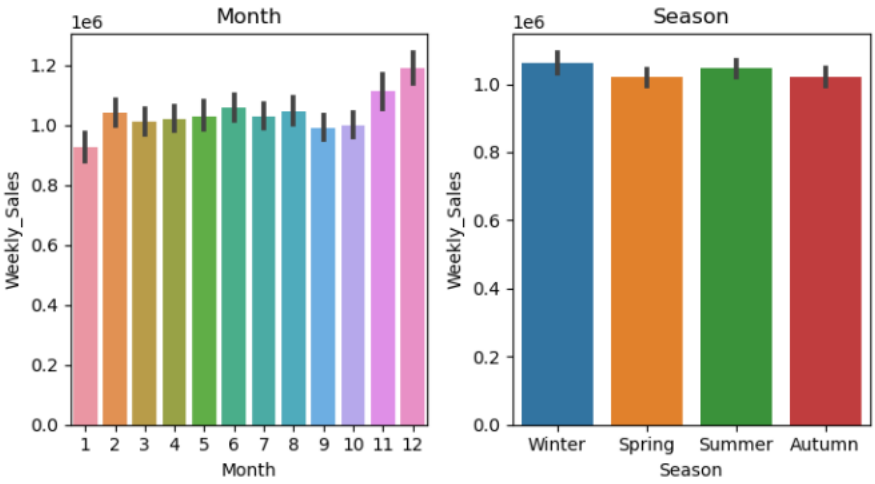

Figure 2: Weekly sales during Months and seasons

Results interpreted that, being complex algorithms, LSTM and XGBoost excelled compared to generic regressor machine learning algorithms like Decision Tree, Random Forest, Ridge and Linear Regressor with an R2-score of "0.99" and "0.97" orderly. It is because the LSTM has a strong memory of remembering a non-linear time series or sequential numerical data and can identify the hidden patterns. With this wonderful LSTM model, it is easy for Walmart to predict sales accurately and helps their retail business to increase profit and optimize inventory management.

The findings from this study show that "LSTM" is the best deep learning algorithm compared to all the machine learning algorithms and predominantly the XGBoost that every other research paper mentioned as best fit for this dataset. These insights have answered the objective of this study i.e. to improve the accuracy of sales prediction. Which further helps the Walmart business to increase the effectiveness in inventory management, decision-making, resource allocation, and profit maximization**.**

5. Conclusion and Future work

In conclusion, this study has proved that the deep learning algorithm "LSTM" has performed way better than the regressor machine learning models, especially the XGBoost with an R2-score of "0.99" unlike the findings of the below reference papers. However, XGBoost outperforms all other machine learning algorithms [1]-[4]. Moreover, while data visualizations author had seen multiple insights and hidden trends in the dataset. The first one is even though there are more records (data points) belonging to summer and spring, the highest weekly sales were recorded during winter from figure 2, that is "November" and "December" which are the months of "Black Friday" and "Christmas" [5]. Secondly, from figure 1, feature importance graphs show that among all the features, "CPI", "Unemployment" and stores 1, 4, 0 are playing crucial role in predicting the "Weekly_Sales" [5].

The limitations of this study are making inferences or concluding the best algorithm solely based on this dataset is not a good approach. In addition, there is a lack of interpretability of models along with inability to handle even larger or continuous data [4]. In future, testing using integrated or ensembled algorithms may increase performance of sales prediction. Moreover, considering different datasets of wide range of features and large sizes to check scalability and robustness of models. Finally, adding explainability will provide transparency of the model resulting in trustworthiness.

6. References

[1] Latha, S. B., Dastagiraiah, C., Kiran, A., Asif, S., Elangovan, D., & Reddy, P. C. S. (2023). *An adaptive machine learning model for walmart sales prediction*. IEEE. https://doi.org/10.1109/iccpct58313.2023.10245029

[2] Dairu, X., & Shilong, Z. (2021). *Machine learning model for sales forecasting by using XGBoost*. IEEE. https://doi.org/10.1109/iccece51280.2021.9342304

[3] Niu, Y. (2020). *Walmart sales forecasting using XGBoost algorithm and feature engineering*. IEEE. https://doi.org/10.1109/icbase51474.2020.00103

[4] Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). *XGBoost model and its application to personal credit evaluation*. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mis.2020.2972533

[5] Ramasami, M. V., Thangaraj, R., Manoj Kumar, S., & Eswaran, S. (2023). *Exploratory data analysis of walmart outlets sales using data analytics techniques*. IEEE. https://doi.org/10.1109/icdate58146.2023.10248586

[6] https://www.kaggle.com/datasets/yasserh/walmart-dataset/data