Walmart Sales Forecasting

Team Members

Parimala Anjanappa, Sree Surya Sreemanth Yedidhi, Vivek Chandra, Yagna praseeda Atmuri

INTRODUCTION

In the ever-competitive retail industry, accurate sales forecasting plays a pivotal role in ensuring business success. Walmart, being a multinational retail corporation with a vast network of stores, faces the intricate challenge of predicting sales patterns across its diverse locations. Effective sales forecasting not only aids in optimizing inventory levels and reducing overstocking or understocking situations but also enables strategic decision-making processes related to marketing campaigns, pricing strategies, and resource allocation.

The sales performance of a retail store is influenced by a multitude of factors, both internal and external. Internal factors may include store characteristics, such as size, layout, and product assortment, as well as operational aspects like staffing levels and promotional activities. External factors, on the other hand, encompass demographic and economic indicators, weather conditions, fuel prices, and consumer sentiment, among others. Doing EDA of external factors' impact on sales can unveil valuable insights for developing effective marketing strategies and adapting to changing market conditions. For instance, understanding the influence of promotional markdowns or holiday periods on sales can guide targeted campaigns and inventory planning. Similarly, insights into the effects of economic indicators and fuel prices can inform pricing decisions and regional strategies.

This project aims to develop robust predictive models capable of forecasting weekly sales for each Walmart store, taking into account these diverse factors. By leveraging historical sales data, store-specific information, and external data sources, the models can capture the intricate relationships between these variables and sales patterns.

PROBLEM DESCRIPTION

Walmart, as a global retail leader, operates thousands of stores across various regions, each with its unique characteristics and market dynamics. Accurately forecasting sales for each of these stores is a complex challenge that requires a comprehensive understanding of the intricate interplay between numerous factors influencing consumer behavior and purchasing patterns.

One of the primary challenges lies in accounting for the diversity of store formats, sizes, and locations.. Each store's sales performance may be influenced by its specific characteristics, such as size, product assortment, and proximity to competitors or residential areas. Furthermore, external factors play a crucial role in shaping consumer demand and purchasing decisions. Holiday periods and promotional markdowns also have a significant impact on sales patterns.

Addressing these multifaceted challenges requires a robust and data-driven approach, this project aims to develop predictive models that can accurately forecast weekly sales for each store. Despite the limited holiday data available, these models must capture the intricate relationships between various factors and sales patterns, enabling Walmart to make informed decisions regarding inventory management, staffing, promotional strategies, and resource allocation.

DESCRIPTION OF DATA

The project utilizes three main datasets from the Kaggle Walmart Sales Forecasting competition (https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/data) to develop predictive models for forecasting weekly sales across Walmart stores.

1. Train Data: This dataset contains historical sales data for 45 Walmart stores across different regions. It includes the following features: `Store` (store number), `Date`, `Weekly_Sales`, `IsHoliday` (indicating whether the week is a holiday week or not). The dataset spans 3 years(2010-2012), allowing for the analysis of sales patterns over this period.

2. Store Data: This dataset provides supplementary information about the 45 Walmart stores included in the train data. It includes features such as `Store` (store number), `Type` (store format or type), `Size` (store size in square feet), and geographical information like `City`, `State`, and `Zip Code`. This dataset aids in understanding the impact of store characteristics on sales performance.

3. Features Data: This dataset contains additional data related to external factors that may influence sales patterns. It includes features such as `Store` (store number), `Date`, `Temperature` (average temperature), `Fuel_Price` (fuel price in the region), `MarkDown1-5` (anonymized promotional markdowns), `CPI` (consumer price index), and `Unemployment` (unemployment rate in the region). These external factors can provide insights into the impact of weather conditions, economic indicators, and promotional activities on sales.

METHODOLOGY

To achieve the goal of developing accurate predictive models for forecasting weekly sales across Walmart stores, a structured and comprehensive methodology is employed. The methodology involves several stages, including data preprocessing, exploratory data analysis, feature engineering, model development, evaluation, and selection. The following steps outline the approach:

Data Preprocessing:

Handling missing values: The features dataset had missing values in several columns, like CPI, Unemployment, and MarkDown1-5. For the CPI and Unemployment columns, missing values were filled with the respective median values, for the MarkDown1-5 columns, which had a significant number of missing values, the missing values were replaced with zeros, under the assumption that missing values

indicate the absence of markdown promotions during those periods. Additionally, a new column named "Total_MarkDown" was created by summing the values of "MarkDown1" to "MarkDown5" for each row.

Merging datasets: The main dataset containing historical sales data was merged with the stores dataset, which provided additional information about each store, including its type and size. Subsequently, the resulting dataset was merged with the features dataset, which contained external factors such as temperature, fuel price, CPI, unemployment rate, and markdown promotions.

Date Column Processing: Firstly, the Date column was converted into the datetime data type, then the date attribute was set as the index of the combined dataset, facilitating time-based indexing and slicing operations on the data. Furthermore, the Date column was split into separate Year, Month, and Week components, extracting these temporal features explicitly.

Outlier detection and treatment: Outliers in the sales data were identified and removed using the Z-score method to ensure the robustness of the analysis. Negative weekly sales values, which may indicate data anomalies or errors, were also removed from the dataset to maintain data quality.

Exploratory Data Analysis (EDA):

In the exploratory data analysis, we delve into understanding the distribution and characteristics of sales across departments and stores within Walmart. On the department front, department 92 emerged as the top performer, followed by department 95. These departments exhibit substantially higher sales compared to others, such as department 78, 43, 47. This wide range of average sales among departments underscores the importance of understanding department-specific dynamics and tailoring strategies accordingly.

Similarly, when examining average sales by store, we observe notable disparities in performance across different Walmart locations. Store 20 is the top-performing store followed by stores 4 and 14. In contrast, store 47 registers negative average weekly sales, indicating a potential area for improvement or further investigation. These findings highlight the variability in sales performance among stores, emphasizing the need for targeted interventions and optimizations at the store level.

In the next part of EDA we saw how various factors such as Temperature, Unemployment, CPI (Consumer Price Index), Fuel Price, and IsHoliday impacts sales. We found out that warmer temperatures between 40-80°F promote higher sales, while extremely cold or hot weather deters shoppers. Higher unemployment rates correlate with decreased sales, especially for store types A and B when the unemployment index exceeds 11. While no clear relationship emerges between CPI changes and weekly sales, store type B exhibits sales peaks when CPI is low. Lower fuel prices ranging from $2.75 to $3.75 encourage higher sales and holiday weeks generate significantly higher sales despite representing a small percentage of the year, underscoring the importance of holiday sales events. By understanding these factor-specific impacts, Walmart can optimize inventory management, marketing strategies, and resource allocation to drive sales performance throughout the year, tailoring approaches to weather conditions, economic indicators, and peak demand periods.

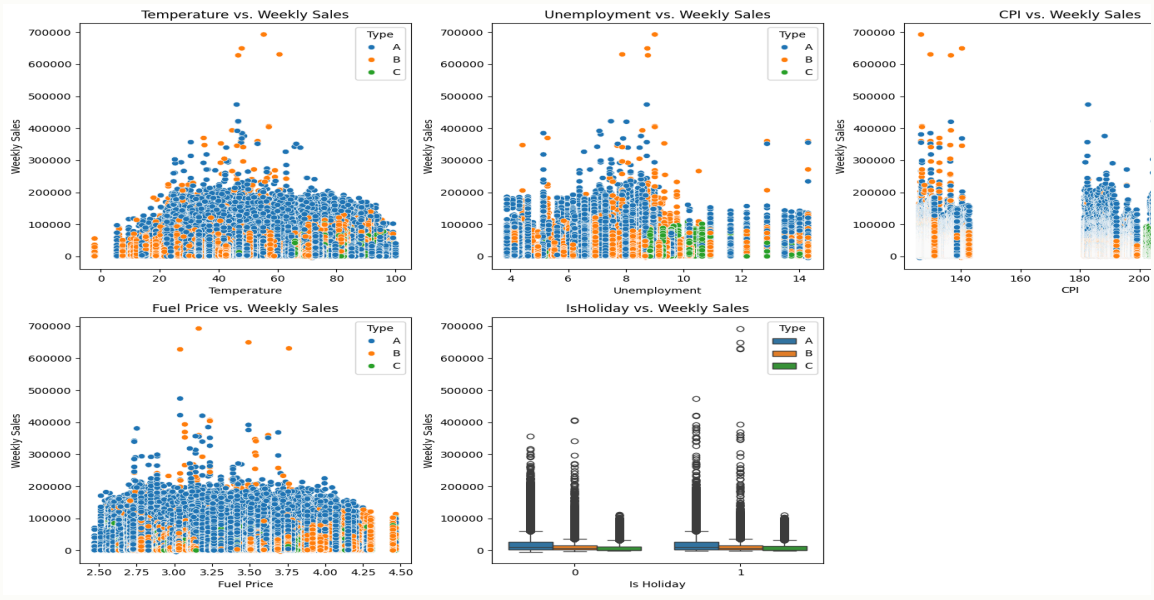

Figure 1: EDA across various columns of the dataset

Feature Engineering:

One-Hot-Encoding:

Categorical columns such as Store, Department (Dept), and Type were one-hot-encoded using the get_dummies method to convert them into numerical representations suitable for machine learning algorithms. This transformation expanded the categorical columns into binary columns representing the presence or absence of each category, increasing the dimensionality of the dataset.

Data Normalization:

Numerical columns were normalized using MinMaxScaler to scale the values to a range between 0 and 1. Normalization ensures that all features contribute equally to the model fitting process and prevents features with larger magnitudes from dominating the analysis.

Recursive Feature Elimination:

A Random Forest Regressor model with 23 estimators was used to calculate feature ranks and importance. Recursive Feature Elimination (RFE) was performed to select the most important features based on their contribution to predicting weekly sales. Features selected for retention included mean, median, Week, Temperature, maximum, CPI, Fuel_Price, minimum, standard deviation, Unemployment, Month, Total_MarkDown, and various Department (Dept) categories.

Machine Learning models:

In selecting multiple regression algorithms for the analysis, we aimed to leverage the strengths of each model while compensating for their respective weaknesses. By incorporating algorithms such as Linear

Regression, Random Forest Regressor, Decision Tree Regressor, and Gradient Boosted Tree Regressor, we explored the diverse range of modeling approaches available and assessed their performance in predicting weekly sales for Walmart stores. This multi-algorithm approach allowed us to evaluate the trade-offs between model complexity, interpretability, and predictive accuracy, ultimately informing our selection of the most suitable algorithm(s) for the task at hand.

Linear Regression:

Linear regression models the relationship between the independent variables (features) and the dependent variable (target) by fitting a linear equation to the observed data. The model assumes a linear relationship between the predictors and the target variable.The results indicate that the RMSE value is approximately 0.0731, the MAE value is around 0.0407, and the R2 Score is approximately 0.9086, indicating that the model explains approximately 90.86% of the variance in the target variable.

Random Forest Regressor:

Random forest regression is an ensemble learning method that builds multiple decision trees during training. The random forest regressor achieved an R2 score of approximately 0.912, indicating a slightly higher explanatory power than linear regression. The RMSE and MAE were 0.072 and 0.040, respectively. This suggests that the model explains approximately 91.23% of the variance in the target variable, with relatively low errors in prediction.

Decision Tree Regressor:

Decision tree regression is a tree-based model that recursively splits the data into subsets based on the value of the features. Each split is chosen to minimize the variance of the target variable within each subset. This model outperformed other models with an R2 score of approximately 0.936, indicating superior explanatory power. The RMSE and MAE were the lowest among all models, at 0.061 and 0.035, respectively.

Gradient Boosted Tree Regressor:

Gradient boosted tree regression is another ensemble learning technique that builds a sequence of decision trees iteratively. Each tree corrects the errors of the previous ones, with the model optimizing a loss function to minimize prediction errors. This model outperforms Linear Regression, Random Forest Regressor, and Decision Tree Regressor in terms of R2 Score, MAE, and RMSE. It achieved the highest R2 Score of 93.60%, indicating a better fit to the data compared to other models. The MAE and RMSE values of the Gradient Boosted Tree Regressor are also lower, indicating more accurate predictions and smaller errors compared to other models.

Deep Neural Network (DNN):

Deep Neural Network using Keras was employed to conduct regression analysis here, data was initially split into features and the target variable, 'Weekly_Sales', followed by further division into training and testing sets. The DNN model architecture comprised three fully connected layers with 64, 32, and 1 neurons, respectively, incorporating the ReLU activation function in the first layer and linear activation in subsequent layers. The DNN model achieved an accuracy of approximately 84.51%, and exhibited promising capabilities in predicting weekly sales based on the provided features.

RESULTS

The Gradient Boosted Tree Regressor exhibited the highest predictive performance, achieving an R^2 score of 93.61% and the lowest RMSE and MAE values, outperforming the other models evaluated. The Decision Tree Regressor also demonstrated strong performance, with a high R^2 score of 90.99% and low error metrics. Linear Regression and Random Forest Regressor showed moderate performance, while the Deep Neural Network achieved competitive results but slightly lower than some of the other models. Notably, external factors such as temperature, unemployment rate, fuel price, and markdown promotions were identified as significant predictors of weekly sales, highlighting their importance in the forecasting models. Additionally, weather conditions, economic indicators, and holiday periods were observed to have a notable impact on sales patterns, influencing consumer behavior and purchasing decisions. These findings underscore the importance of incorporating relevant external factors and understanding their influence on sales patterns to develop accurate and reliable forecasting models for Walmart's weekly sales.

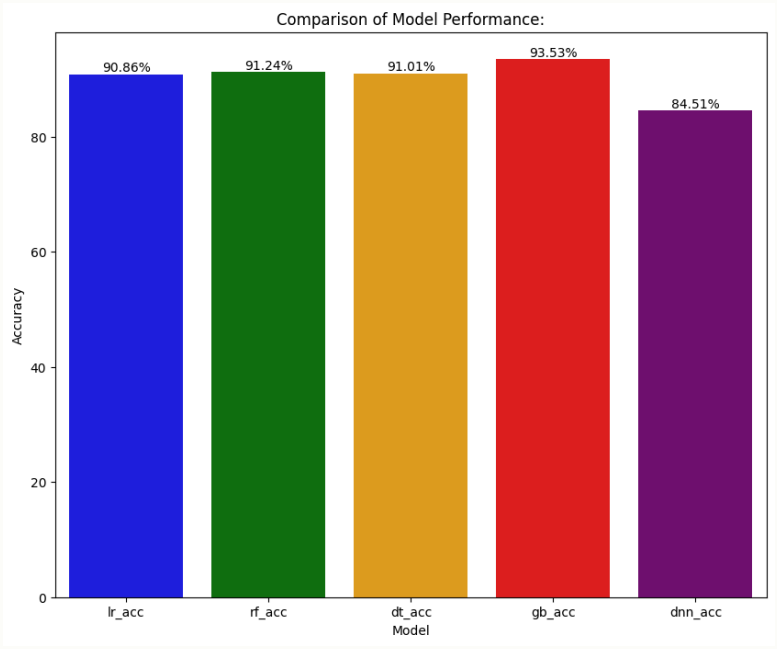

Figure 2: Accuracy across different models

Discussion: Challenges faced

Outlier Removal: The inability to use a predefined Z-score function in Spark DataFrame led to the use of box plot method for outlier removal.

Limited ML Algorithm Support: PySpark's limited support for certain machine learning algorithms, such as KNN and XGBoost, posed a challenge in building models using these techniques. To overcome this limitation, alternative algorithms like decision tree, gradient boosting algorithms available in PySpark were explored and utilized.

REFERENCES

1. Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Bangkok, Thailand, 2020, pp. 458-461, doi: 10.1109/ICBASE51474.2020.00103.
2. Walmart sales prediction based on Machine learning, 2023, Siming Yi, https://doi.org/10.54097/hset.v47i.8170
3. Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: a comparison of artificial intelligence and statistical techniques. Journal of Retailing and Consumer Services, 8(3), 147-156.
4. Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88.
5. Gür Ali, Ö., & Peker, S. (2018). A novel approach to forecasting retail sales with external factors: An application of ensemble machine learning. International Journal of Forecasting, 34(4), 759-772.
6. Grewal, R., Dharwadkar, R., Gotfried, D., & Brown, P. (2017). Forecasting product demand using deep learning. arXiv preprint arXiv:1711.11039.
7. Huang, T., Bergman, D., & Eklund, J. (2019). Forecasting retail sales with dynamic Bayesian machine learning models. International Journal of Forecasting, 35(4), 1494-1508.

Appendix 1: Contributions from each member

I. Surya took charge of data collection from the Kaggle - Walmart Sales Forecasting competition and data preprocessing tasks. He handled missing values, removed outliers, transformed features. He also contributed to evaluating machine learning algorithms.

II. Vivek evaluated various machine learning algorithms like regression, decision trees, random forests, XGBoost, and neural networks for sales forecasting. He implemented the selected algorithms, performed feature engineering, tuned hyperparameters, and trained the models using cross-validation techniques.

III. Praseeda defined and implemented evaluation metrics like MAE, RMSE, and R-squared to assess model performance. She conducted comprehensive model evaluation using k-fold cross-validation, analyzed the impact of external factors on sales, and optimized the best-performing models based on results.

IV. Parimala conducted exploratory data analysis to gain insights, contributed to evaluating machine learning algorithms and prepared detailed documentation covering the methodology, preprocessing, model selection, training, evaluation, and project findings.