STORE SALES FORECASTING

ADITYA MUKHOPADHYAY, ANKIT KUMAR, SANJAY KUMAR

Abstract

Sales forecasting is a critical function in the dynamic and competitive retail industry, particularly for a global leader like Walmart. As one of the largest retail chains in the world, Walmart faces substantial challenges in effectively managing inventory, ensuring operational efficiency, and meeting fluctuating customer demands. The complexity of these challenges is further amplified by seasonal trends, holidays, promotions, and broader economic conditions.

Accurate sales forecasting plays a pivotal role in navigating these complexities. By ensuring a precise balance between supply and demand, it prevents overstocking, which leads to increased holding costs and markdowns, and avoids stockouts, which result in lost sales and dissatisfied customers. Furthermore, effective forecasting aids in strategic decision-making, workforce planning, and resource allocation while contributing to cost savings and enhanced customer satisfaction.

The objective of this report is to explore how data science techniques can address these challenges by accurately predicting weekly sales, identifying key influencing factors, optimizing inventory levels, and improving operational efficiency. Through the application of advanced analytical models, this study seeks to provide actionable insights that ensure product availability during peak periods, enhance customer satisfaction, and drive strategic growth for Walmart in an increasingly competitive retail landscape.

2. Introduction

Sales forecasting is a cornerstone of effective retail operations, providing valuable insights that guide inventory management, operational planning, and strategic decision-making. For a global retail giant like Walmart, which operates on an immense scale, accurate forecasting is not just an operational necessity but a competitive imperative. Walmart faces challenges unique to its vast network, including balancing inventory across thousands of stores, managing diverse product categories, and responding to rapidly changing customer demands. External factors such as seasonal trends, holiday sales, promotions, and broader economic conditions add further complexity to this task.

This report focuses on predicting sales data for Walmart stores, a vital task that directly impacts key performance metrics such as cost efficiency, customer satisfaction, and revenue generation. Initially, the scope of this study aimed at long-term forecasting of sales trends using time series data. However, due to inaccuracies in the data over extended periods, the scope was refined to focus on short-term sales prediction. This adjustment enabled a more detailed exploration of influencing factors and improved the reliability of the predictions.

The prediction models employed in this study analysed Walmart's sales data in two distinct ways:

1. **Reduced Parameter Scope**: Limited to essential attributes such as store number, department, and week number.
2. **Complete Parameter Scope**: Incorporating a broader range of 15 parameters, including store type, size, weekly sales, holiday indicators, promotional discounts, and various economic and environmental factors like

temperature, fuel price, CPI, and unemployment rates.

Several predictive models were tested during this analysis, including Decision Trees, Random Forest Regressors, GluonTS, LSTM, and other advanced machine learning techniques. Despite the ambition of employing sophisticated models like LSTM and GluonTS, constraints such as limited GPU resources and insufficient data quality hindered their effectiveness. Ultimately, the Decision Tree and Random Forest Regressor models emerged as the most viable solutions, offering robust and accurate predictions.

This report delves into the methodologies, results, and insights gleaned from this endeavour, aiming to demonstrate the potential of data science to optimize sales forecasting for a retail ecosystem of Walmart's scale and complexity.

3. Materials and Methods

3.1 Dataset Description

The Walmart Sales Dataset, sourced from Kaggle, provides historical sales data from various Walmart stores along with contextual details such as promotions, holiday schedules, and economic indicators. The dataset comprises 421,570 rows and 17 columns, with a mix of numerical and categorical variables. Key features include Store and Dept identifiers, weekly sales figures (Weekly_Sales), store size (Size), promotional markdowns (MarkDown1-5), and external factors like Temperature, Fuel_Price, CPI, and Unemployment. Additionally, the dataset includes categorical variables such as Type (store type) and Boolean indicators like IsHoliday.

While the dataset is rich in features, approximately 50% of the data in the markdown columns is missing, requiring careful preprocessing. Despite this limitation, the dataset remains suitable for the study, offering the breadth and depth needed to explore sales prediction across varying parameter scopes.

3.2 Evaluation Parameters

The performance of our sales prediction models is evaluated using the **Root Mean Squared Error (RMSE)**. RMSE is a widely used metric for regression tasks as it provides a direct measure of the average magnitude of prediction errors in the same units as the target variable, making it intuitive and interpretable.

Given that the chosen models are Decision Tree and Random Forest Regressors, RMSE effectively captures how well the models predict weekly sales by penalizing larger errors more heavily than smaller ones. This ensures that the models prioritize accuracy across the entire dataset, particularly for outlier predictions where high accuracy is critical. A lower RMSE value indicates better model performance and more reliable predictions.

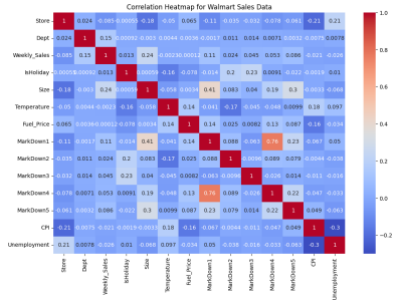
3.3 EDA

The goal of this EDA is to gain an understanding of the dataset that will be used for forecasting store sales. By analysing the data, we aim to uncover patterns, relationships, and insights that can inform the model-building phase. The datasets involved in this analysis include sales, store, and feature information, which are merged to form a comprehensive dataset for analysis.

3.3.1 Data Loading and Merging

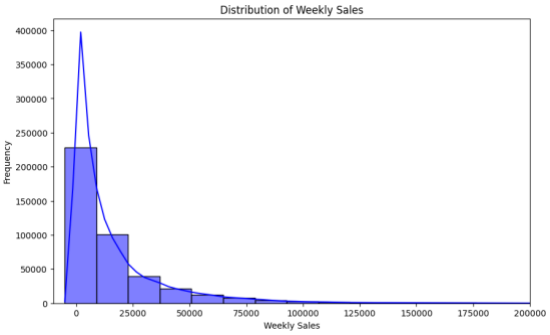
Initially, we load multiple datasets: train.csv, features.csv, stores.csv, test.csv, and sampleSubmission.csv. These datasets are essential components for our analysis. The train.csv and features.csv datasets are merged with stores.csv using a left join to ensure all store-related data is included, leading to the creation of the primary dataset for analysis.

3.3.2 Correlation Analysis



A correlation heatmap is generated to reveal the relationships between numerical features within the dataset. This visualization aids in identifying potential predictors for sales forecasting by highlighting strong correlations.

3.3.3 Distribution of Weekly Sales

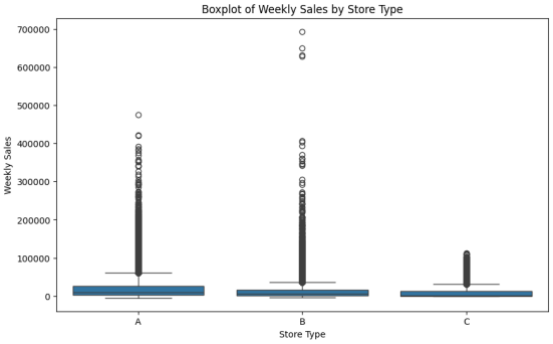


The distribution of the target variable, Weekly_Sales, is explored through a histogram complemented by a kernel density estimate (KDE). This plot provides insights into the spread and frequency of sales values, helping to identify any skewness or outliers.

3.3.4 Sales Over Time

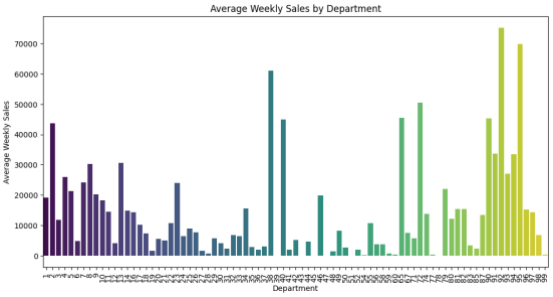
A line plot visualizes Weekly_Sales over time, with a distinction between holiday and non-holiday periods. This analysis assesses the temporal trends and the impact of holidays on sales, offering a glimpse into seasonal patterns.

3.3.5 Sales by Store Type



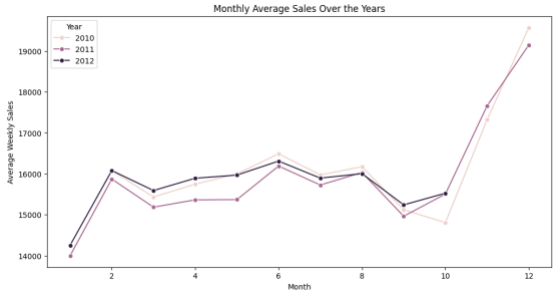
A boxplot examines the distribution of sales across different store types. This visualization helps in understanding how store characteristics influence sales performance, revealing variations in sales based on store type. Although the data might look like to be housing a lot of outliers, these values hold critical information about exceeding Sale values because of the influence of several markdowns, Holidays and key features in dataset. Hence a reasonable decision to withhold them was taken.

3.3.6 Average Sales by Department



The average weekly sales per department are depicted using a bar plot, which assists in identifying the most revenue-generating departments. This insight can inform strategies for department-specific promotions and inventory management.

3.3.7 Grouped Sales by Year and Month



Sales data grouped by year and month is illustrated through a line plot, showcasing monthly average sales over several years. This visualization helps detect long-term trends and seasonal fluctuations, which are crucial for accurate forecasting. Here we can visualise the impact of Christmas and year end offers impacting Sales which earlier looked like outliers.

3.3.8 Data Manipulation & Preparation

Further data manipulation includes filling missing values in markdown-related columns with zeros and extracting additional temporal features, such as the week number from the date, to enhance the dataset's richness.

3.3.9 Data Imputation and Scaling

Numeric columns undergo imputation and scaling processes to ensure standardized data, which is vital for model performance. This step involves the use of SimpleImputer and MinMaxScaler.

3.3.10 One-Hot Encoding

Categorical features are transformed through One-Hot Encoding, converting them into a machine learning-friendly format and facilitating model training.

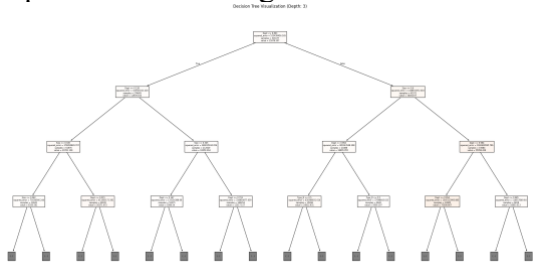
3.4 Data Splitting

The dataset is partitioned into training and validation sets, setting the stage for effective model training and evaluation. With 75% retained in the training dataset and remaining 25% moved into a testing dataset.

3.5 Models Implemented

3.5.1 Decision Tree

In this project, a Decision Tree Regressor from the scikit-learn library is employed to predict weekly sales figures. Decision Trees are non-parametric models that can capture non-linear relationships between features and the target variable. They are intuitive and easy to interpret, making them a popular choice for regression tasks.



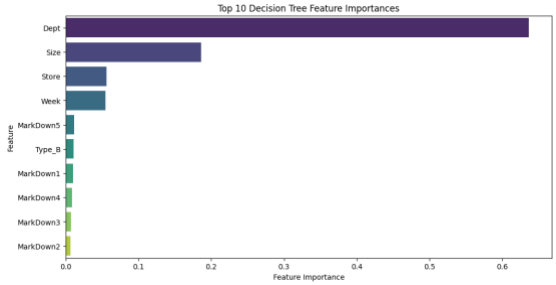
Model Initialization and Training

The Decision Tree Regressor is initialized with a specified random state for reproducibility. The model is then trained on the provided training dataset (train_inputs and train_targets). This process involves recursively splitting the data according to feature values that result in the greatest reduction in variance, effectively learning patterns from the data.

Model Evaluation

- **Training Evaluation**: The model's performance on the training data is evaluated using the Root Mean Squared Error (RMSE), which measures the average magnitude of the prediction error. The training RMSE is extremely low (4.4732427018867185e-17), suggesting that the model fits the training data almost perfectly.
- **Validation Evaluation**: The model's performance on unseen validation data is also assessed using RMSE. This metric indicates how well the model generalizes to new data. The validation RMSE provides a more realistic estimate of model performance in practice (5280.93931407318).

Feature Importance Analysis



The Decision Tree model provides insights into feature importance, which indicates the contribution of each feature to the model's predictions. The top 5 features identified are:

1. **Dept**: With the highest importance score, department is the most influential feature, suggesting that sales predictions heavily rely on department-specific patterns.
2. **Size**: The size of the store is the second most important feature, indicating its significant impact on sales.
3. **Store**: Store identifiers contribute to the model but to a lesser extent than department and size.
4. **Week**: The week number is also influential, reflecting potential seasonal trends.
5. **Markdown Feature 5**: Several markdown features contribute to the model, although Markdown 5 promotional activity have a noticeable, though less dominant, impact on sales over majority data.

3.5.2 Random Forest Regressor

In this analysis, a Random Forest Regressor is employed to predict weekly sales figures. Random Forest, an ensemble learning method, constructs multiple decision trees during training and outputs the average prediction of the individual trees, reducing overfitting and improving generalization compared to a single decision tree.

Model Initialization and Training

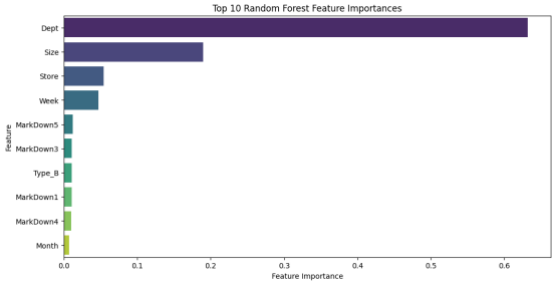
The Random Forest Regressor is initialized with the following parameters:

- **Random State**: Set to 42 for reproducibility.
- **Number of Trees (n_estimators)**: 100 trees are used, providing a balance between computational efficiency and model accuracy.
- **Maximum Depth**: Not limited, allowing trees to expand fully, which helped capturing complex patterns in data.

Model Evaluation

- **Training Evaluation**: The model's performance on the training data is assessed using the Root Mean Squared Error (RMSE). A training RMSE of approximately 1429.53 indicates a strong fit to the training data, suggesting that the model captures a significant amount of the underlying data patterns.
- **Validation Evaluation**: The validation RMSE is approximately 3883.02, suggesting that while the model generalizes reasonably well to unseen data, there is room for improvement, potentially through hyperparameter tuning which we will come to shortly.

Feature Importance Analysis

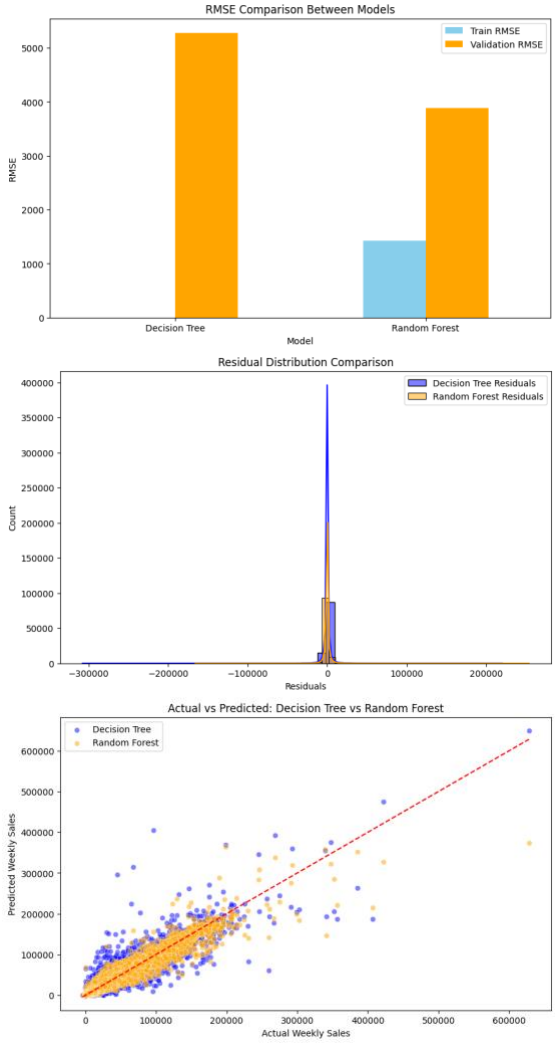


The Random Forest model provides insights into feature importance, quantifying each feature's contribution to the model's predictions:

1. **Dept**: The department feature is the most influential, consistent with the observations from the Decision Tree analysis, indicating department-specific sales patterns.

2. **Size**: Store size remains a significant predictor, reflecting its impact on sales volume.
3. **Store**: The store identifier contributes meaningfully but less than the department and size.
4. **Week**: Time-related features such as the week number are important, suggesting temporal effects on sales.
5. **Markdown Features**: Markdown features
   like MarkDown5 and MarkDown3 show their relevance, likely due to their role in promotional strategies.
6. **Month**: The inclusion of the month as an important feature highlights potential seasonal trends affecting sales.

3.6 Comparison & Inference



The comparison between Decision Tree and Random Forest models reveals that Random Forest outperforms the Decision Tree in multiple aspects:

1. **Prediction Accuracy**: Random Forest predictions are more closely aligned with actual values, as evidenced by tighter clustering around the perfect prediction line, while the Decision Tree shows greater variability and outliers.
2. **Residual Distribution**: Random Forest residuals are more concentrated near zero, indicating smaller prediction errors, whereas the Decision Tree residuals exhibit a wider spread, reflecting lower accuracy and higher variability.
3. **Error Metrics**: Random Forest achieves significantly lower RMSE values for validation data while Decision Tree was a prey to Overfitting during training and resulted in less accurate predictions on test data, highlighting Random Forest's superior generalization ability and robustness to overfitting compared to the Decision Tree.

Overall, Random Forest's ensemble learning approach makes it the preferred model for this experiment, ensuring accurate and reliable sales predictions critical for effective inventory and operational decision-making.

3.7 Hyper Parameter Tuning

In this analysis, hyperparameter tuning is conducted on a Random Forest Regressor to enhance its predictive performance on the sales prediction task. The tuning process involves exploring various combinations of hyperparameters to identify the optimal configuration that minimizes prediction error.

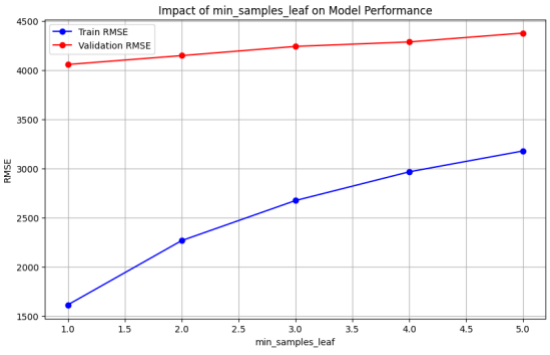
3.7.1 Hyperparameter Tuning using Random Search

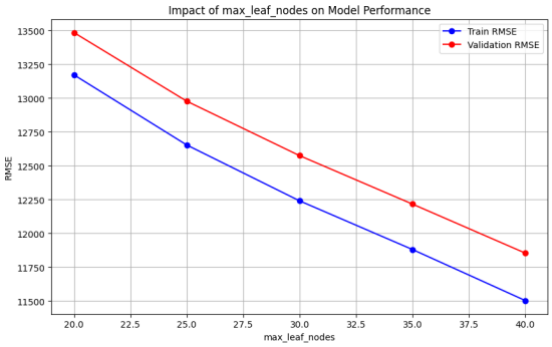
1. **Model Definition**: A Random Forest Regressor is initialized with default settings, using all available

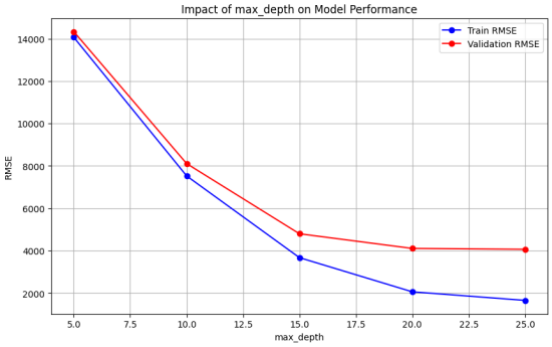
CPU cores (n_jobs=-1) for efficient computation.

2. **Hyperparameter Grid**: A set of hyperparameters is defined to explore their impact on model performance:
    - n_estimators: Number of trees in the forest (10, 50, 100, 200, 300).
    - max_depth: Maximum depth of the trees (5, 10, 15, 20, 25, None).
    - min_samples_split: Minimum samples required to split a node (2, 5, 10, 15).
    - min_samples_leaf: Minimum samples required at a leaf node (1, 2, 4, 6).
    - max_features: Number of features to consider for splitting at each node ('sqrt', 'log2', None).
    - bootstrap: Whether to use bootstrap samples when building trees (True, False).
3. **Randomized Search**: A RandomizedSearchCV is employed to search over the parameter grid, evaluating 50 different parameter combinations using 3-fold cross-validation. The scoring metric is negative RMSE, which is converted to positive for interpretation.
4. **Model Fitting**: The hyperparameter search is executed on the training data, resulting in 150 fits.

3.7.2 Results and Inferences

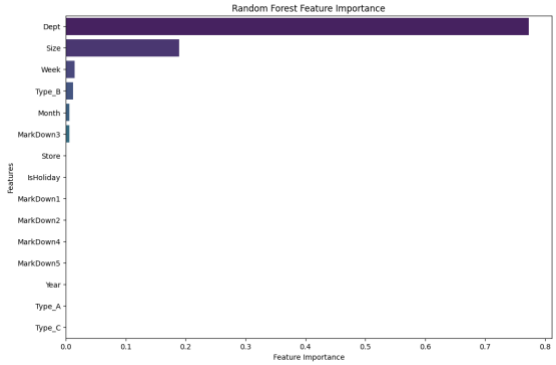




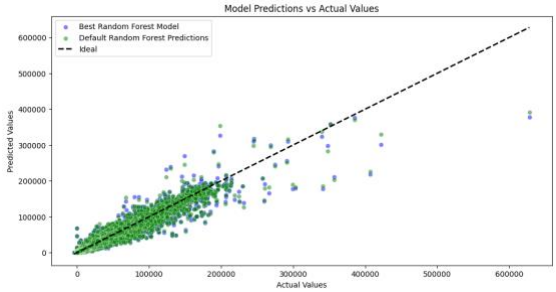


- **Best Parameters**: The optimal parameter combination found includes 200 estimators, a max depth of 25, a minimum of 5 samples to split a node, and 1 sample at a leaf, with no restriction on features per split and bootstrap sampling enabled.
- **Performance**: The best validation RMSE identified during cross-validation is approximately 4092, while the best model achieves a validation RMSE of 3917.16 on the validation set, indicating a reduced performance. A reason could be due to model overfitting or an imbalanced dataset in train and test.
- **Feature Importance**: The analysis reveals that:
    - Dept is the most significant feature, heavily influencing sales predictions.
    - Size also plays a crucial role, followed by temporal features like Week and Month.
    - Surprisingly, features such as Store, IsHoliday, and markdown-related features exhibit negligible importance, suggesting

limited impact in the presence of other strong predictors.



Here is the comparison between the hyperparameter tuned model vs default model



3.8 Conclusion

The implementation and fine-tuning of the Random Forest Regressor for store sales forecasting have resulted in a robust predictive model capable of adapting to different input scopes. This flexibility enables the model to provide reliable sales predictions under varying scenarios, whether utilizing a streamlined set of features or the full spectrum of available data.

1. **Reduced Parameter Scope**: When employing a limited set of key features—store number, department, and week number—the model capitalizes on the most influential factors driving sales. This approach is particularly advantageous for scenarios where data availability is constrained, yet accuracy remains critical. The Random Forest's inherent ability to handle feature interactions ensures that even with fewer inputs, the model maintains a commendable level of precision.
2. **Complete Parameter Scope**: With the inclusion of a comprehensive set of features, the model leverages the full breadth of data to enhance its predictive capabilities. This approach captures intricate patterns and correlations, providing deeper insights and potentially higher accuracy. It is well-suited for environments where detailed data is readily available and the goal is to maximize forecast granularity.

Ultimately, the Random Forest Regressor's adaptability across different input configurations underscores its utility and effectiveness in predicting store sales. By balancing simplicity and complexity, the model delivers tailored predictions that can inform strategic decision-making and drive business success, whether in resource-limited contexts or data-rich environments. This versatility positions the model as a vital tool in the arsenal of data-driven forecasting solutions.