Capstone Project

WallMart-Sale-Forecast

AUTHOR-SHUBHAM SINGH

Table of Contents

- Problem Statement
- Project Objective
- Data Description
- Data Pre-processing Steps and Inspiration
- Choosing the Algorithm for the Project
- Problem Motivation and Reasons For Choosing the Algorithm
- Assumptions
- Model Evaluation and Techniques
- Inferences from the Same
- Future Possibilities of the Project
- Conclusion
- References

Problem Statement

A retail store chain with multiple outlets nationwide is encountering challenges in effectively managing its inventory to align supply with demand. As a data scientist, the task is to analyze the available data, extract valuable insights, and develop prediction models to forecast sales for a specific timeframe, be it months or years. The objective is to provide the retail store with actionable information and accurate sales forecasts to optimize inventory management and enhance overall operational efficiency.

Project Objective

The objective of this project is to utilize data analysis and predictive modeling techniques as a data scientist to address the inventory management challenges faced by retail store chain with multiple outlets across the country. The project aims to achieve the following objectives:

Data Analysis: Perform thorough analysis of the available data to gain insights into the demand patterns, sales trends, and inventory levels across different outlets.

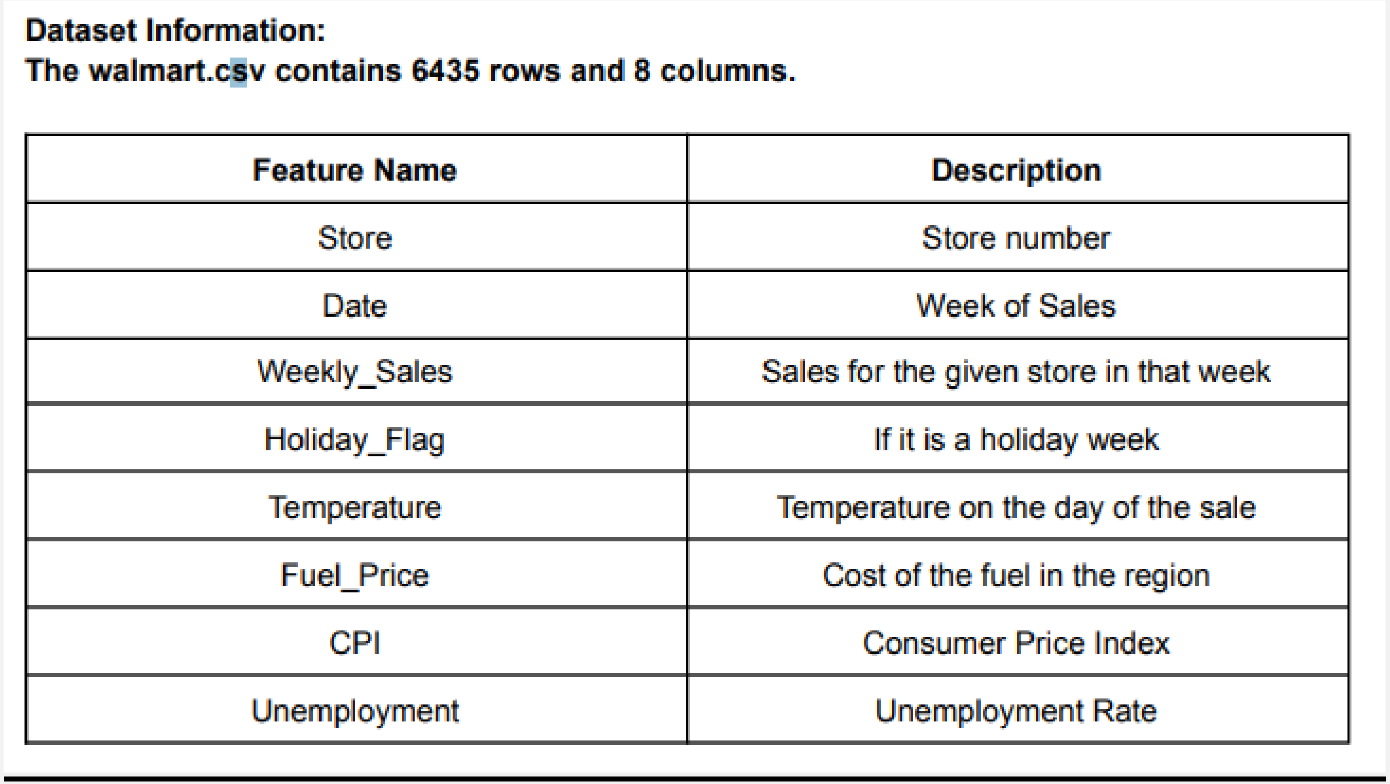
Forecasting: Develop accurate and reliable prediction models to forecast sales for a specified future period, such as months or years. These models will help the retail store anticipate customer demand and align their supply accordingly.

Optimization: Optimize inventory management by utilizing the sales forecasts to determine optimal stocking levels, identify potential stakeouts or excess inventory situations, and streamline the supply chain processes.

Actionable Insights: Provide the retail store management with actionable insights and recommendations based on the data analysis and predictive modeling results. These insights will support informed decision-making, such as adjusting inventory levels, planning promotions, or identifying underperforming outlets.

By achieving these project objectives, the aim is to enable the retail store chain to efficiently manage their inventory, reduce stakeouts and excess inventory, improve customer satisfaction, and ultimately drive profitability across their outlets nationwide.

Data Description

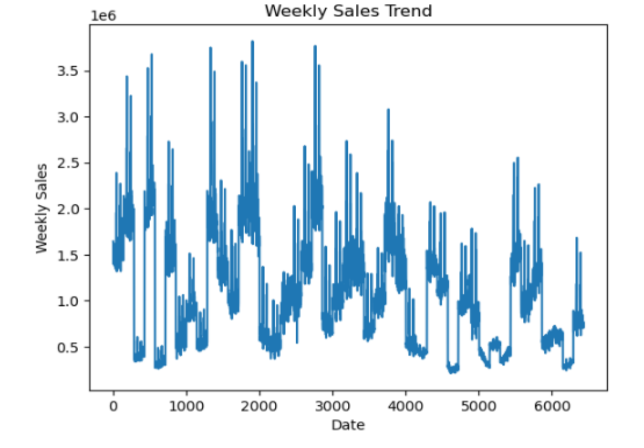


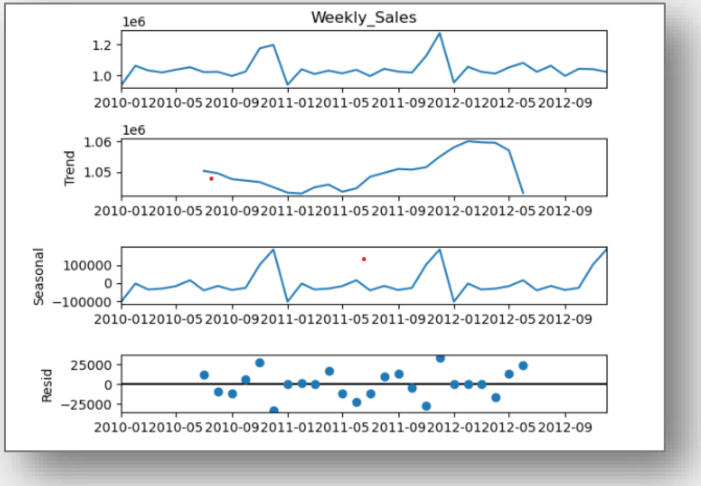
Dataset Information:
The walmart.csv contains 6435 rows and 8 columns.

Feature Name	Description
Store	Store number
Date	Week of Sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

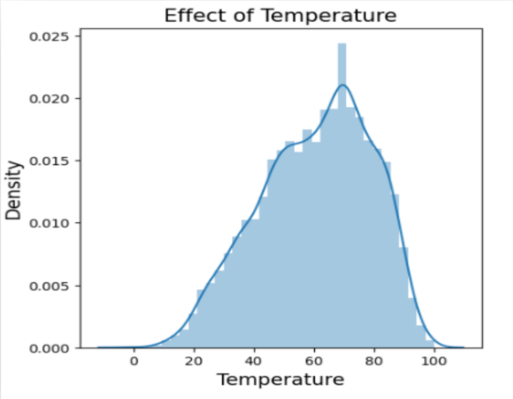
Exploratory Data Analysis

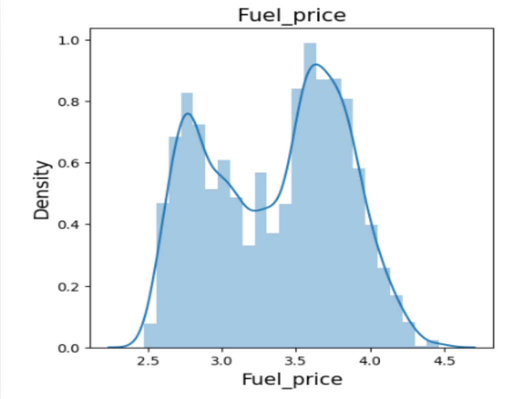
SALES WEEKLY TRENDS

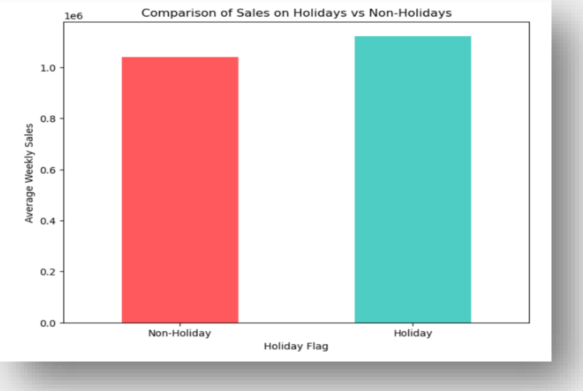

Weekly Sales Trend
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Weekly Sales
0
1000
2000
3000
4000
5000
6000
Date

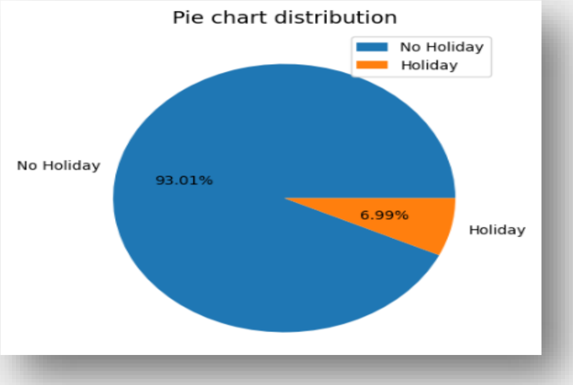

Weekly_Sales
1e6
1.2
1.0
2010-012010-052010-092011-012011-052011-092012-012012-052012-09
1e6
1.06
1.05
Trend
2010-012010-052010-092011-012011-052011-092012-012012-052012-09
Seasonal
100000
0
−100000
2010-012010-052010-092011-012011-052011-092012-012012-052012-09
Resid
25000
0
−25000
2010-012010-052010-092011-012011-052011-092012-012012-052012-09

FEATURE DISTRIBUSTION

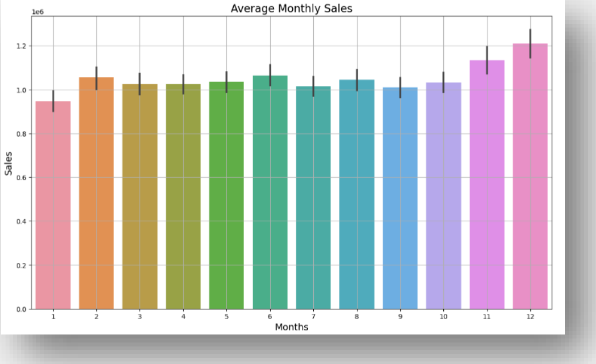


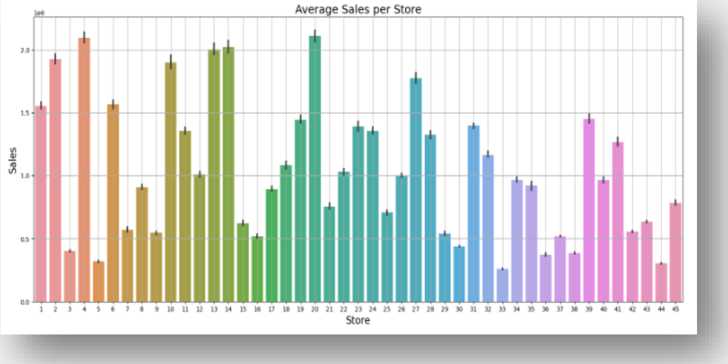






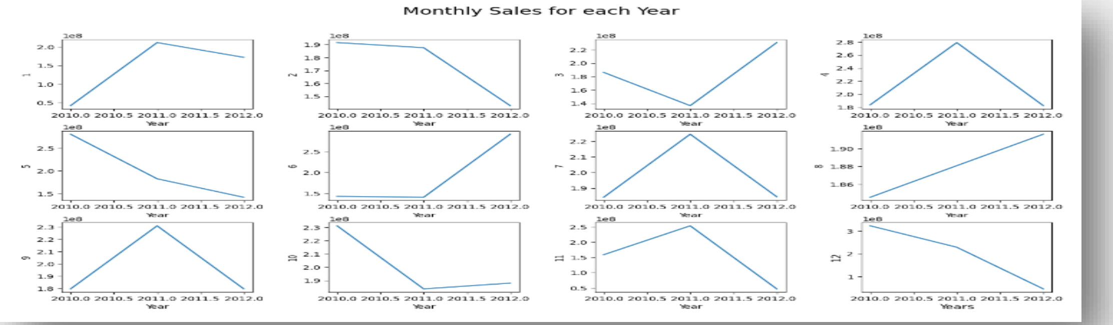
EDA UNIVARIATE ANALYSIS





Average Monthly Sales

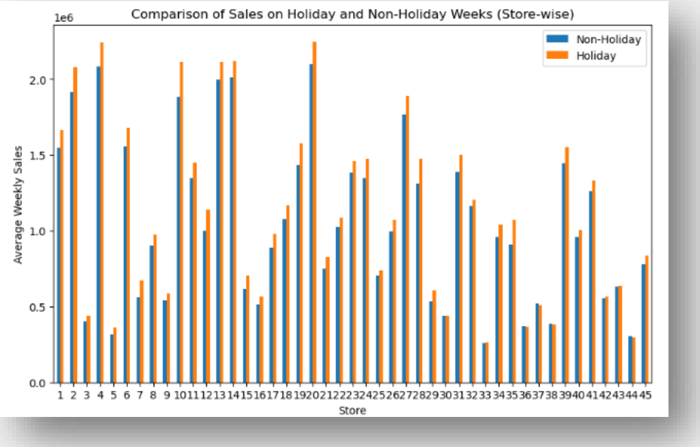
Average Sales per Store

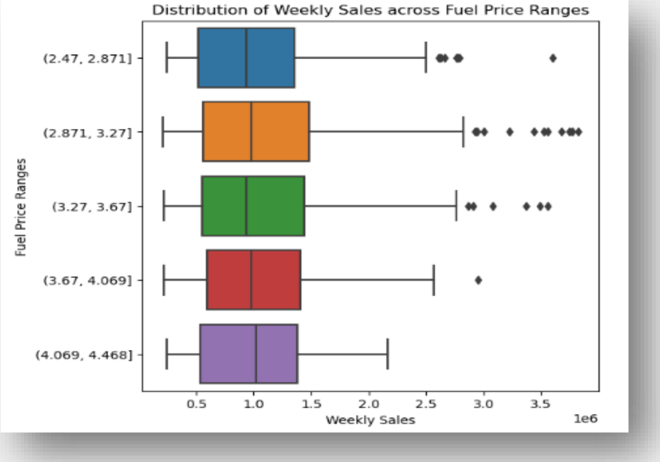


Effect of Features on Sales

Sales on holiday

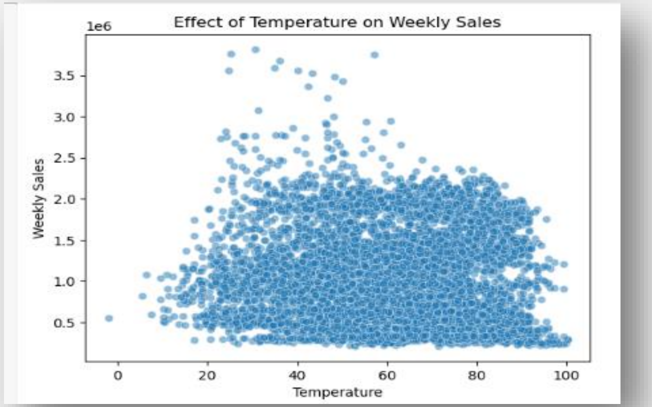
Sales with fuel Price

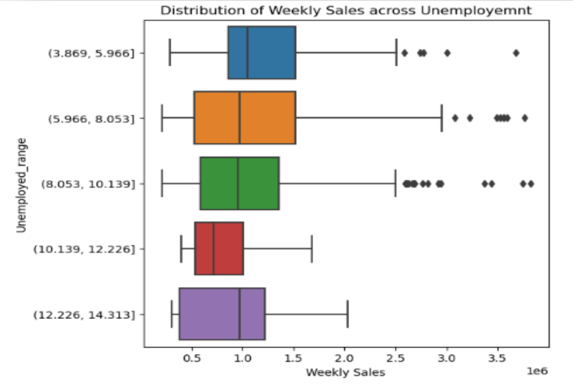


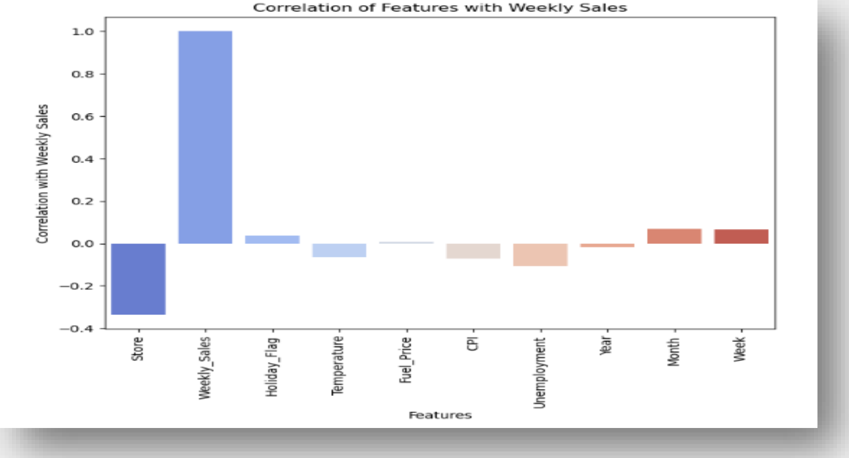


Sales with different Temperature

Sales with different Unemployment


Effect of Temperature on Weekly Sales
1e6
Weekly Sales
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
20
40
60
80
100
Temperature


Distribution of Weekly Sales across Unemployemnt
Unemployed_range
(3.869, 5.966]
(5.966, 8.053]
(8.053, 10.139]
(10.139, 12.226]
(12.226, 14.313]
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e6
Weekly Sales


Correlation of Features with Weekly Sales
Correlation with Weekly Sales
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Store
Weekly_Sales
Holiday_Flag
Temperature
Fuel_Price
CPI
Unemployment
Year
Month
Week
Features

Insights from EDA :

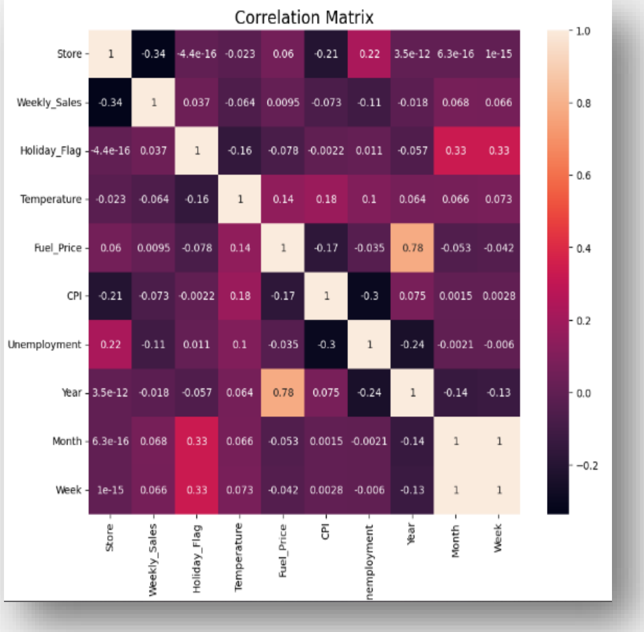
Based on the analysis of the Walmart sales data, the following key inferences can be drawn:

1. The weekly sales exhibit a stationary trend over the period of three years, indicating that there is no significant upward or downward trend observed during this time frame.
2. The average monthly sales display an increasing trend during the last three months of the year, suggesting a positive seasonal effect towards the end of each year.
3. When considering average sales store-wise, there is a wide variation in business performance among stores. Some stores are experiencing below-average sales, while others are performing above-average.
4. The distribution of temperature shows a left-skewed pattern, whereas fuel price, CPI, and unemployment follow bimodal distributions, indicating different clusters of data points.
5. Approximately 6.7% of the data corresponds to holiday periods, and during these times, the stores tend to exhibit better sales performance, suggesting a positive impact on sales during holidays.
6. Lower fuel prices have a positive effect on the weekly sales of stores, potentially influencing consumer spending habits.
7. Moderate temperatures are associated with the highest weekly sales, indicating a favorable temperature range for consumer activities.
8. A medium unemployment rate appears to correlate with the highest weekly sales, implying that a balance in the employment market positively affects consumer spending.
9. Weekly sales demonstrate a negative correlation with all the features in the dataset, implying that fluctuations in these features influence the sales trends, albeit inversely.

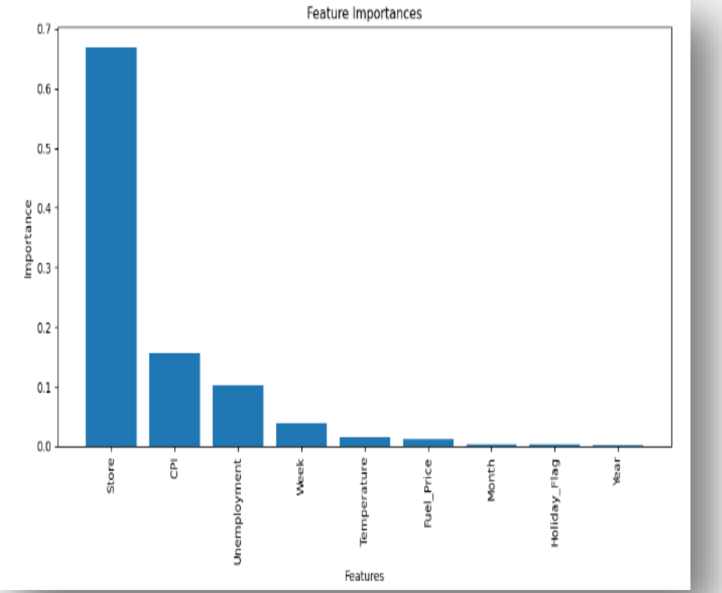
These findings provide valuable insights into the sales patterns and performance of Walmart stores, facilitating data-driven decision-making and targeted strategies to optimize business outcomes.

Data Pre-processing Steps and Inspiration

Correlastion



Feature-Importance



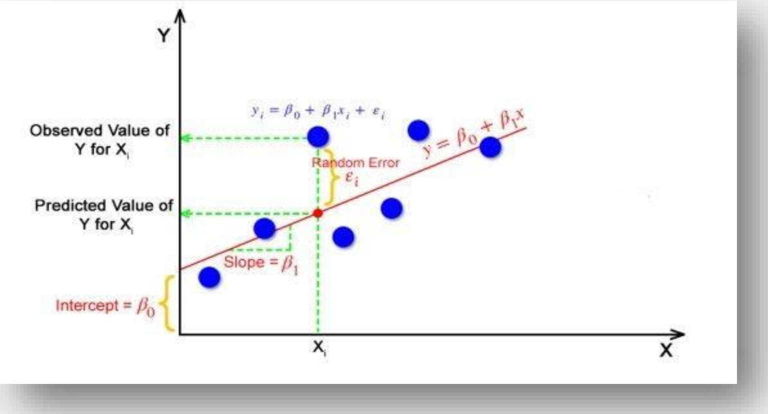
After conducting a thorough analysis of the correlation and using random forest feature importance, I made the strategic decision to drop certain features from the dataset. The features removed include "holiday," "CPI," and "temperature." The rationale behind this decision is to streamline the dataset and focus on the most influential factors that have a significant impact on the weekly sales.

By removing these features, we aim to simplify the model and potentially improve its performance by reducing noise and increasing the interpretability of the results. This process allows us to prioritize the key drivers of weekly sales and focus on the most relevant factors that directly affect the business performance.

This data refinement should enable us to build a more accurate and efficient model, making it easier to draw actionable insights and make well-informed decisions based on the remaining crucial features. The streamlined dataset will provide a clearer understanding of the relationships between these essential variables and the weekly sales, allowing for more targeted strategies and optimizations to enhance overall sales performance.

Choosing the Algorithm for the Project:

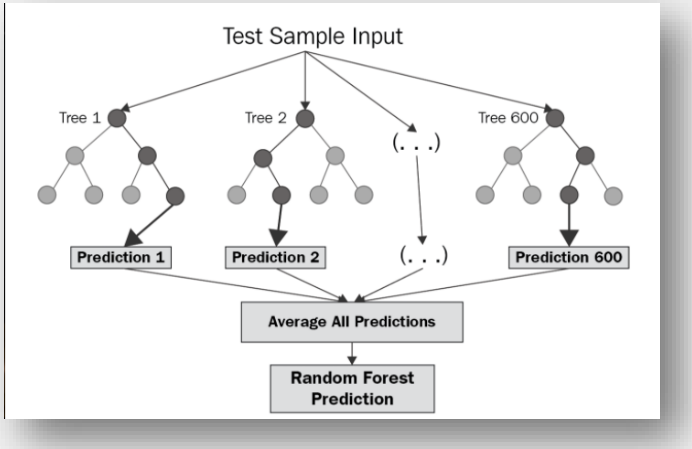
Linear Regression


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
$y = \beta_0 + \beta_1 x$
Random Error ε_i
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

Linear regression is a statistical technique used to model the relationship between the weekly sales and predictor variables, enabling accurate predictions based on historical data.

Using linear regression, we can identify the underlying linear pattern in the historical sales data and make reliable predictions for future weekly sales. By leveraging this predictive model, businesses can optimize inventory, plan marketing strategies, and enhance decision-making to drive sales growth and improve overall performance.

RandomForest-Regressor

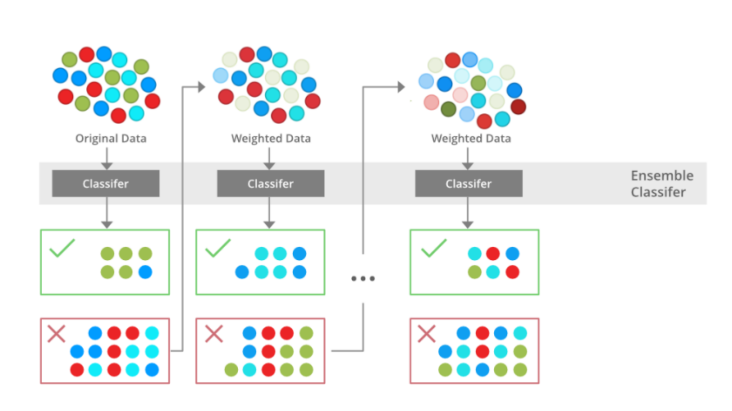


Random Forest Regressor is a powerful machine learning algorithm that utilizes an ensemble of decision trees to predict weekly sales based on historical data.

By aggregating the predictions of multiple decision trees, the Random Forest Regressor provides robust and accurate forecasts for future weekly sales, enabling data-driven decision-making and optimized business strategies.

Leveraging the Random Forest Regressor, companies can gain valuable insights into sales patterns, identify influential factors, and implement targeted approaches to boost sales performance and maximize profitability.

XGBOOST REGRESSOR



XGBoost Regressor excels at handling large datasets and overcoming overfitting, resulting in robust and generalizable predictions for weekly sales.

Its ability to capture nonlinear relationships between features allows businesses to identify key drivers and optimize pricing, promotions, and inventory to maximize revenue and enhance overall sales performance. Leveraging XGBoost Regressor empowers companies with actionable insights to stay competitive and make data-driven decisions in a dynamic marketplace.

Motivation and Reasons For Choosing the Algorithm

Using XGBoost Regressor as the final model was a strategic choice based on its superior performance over other algorithms like Random Forest and Linear Regression. The decision was driven by its ability to achieve the best Root Mean Square Error (RMSE) score and significantly reduce prediction errors.

XGBoost's strength lies in its ensemble approach, which combines multiple decision trees to create a more accurate and robust predictive model. Through gradient boosting and advanced regularization techniques, XGBoost optimizes the model's performance, effectively handling complex data patterns and capturing nonlinear relationships between features.

The model's ability to minimize overfitting ensures generalizability to new data, making it reliable for weekly sales predictions. By selecting XGBoost Regressor, the organization gains a powerful tool to drive data-driven decisions, optimize inventory, and enhance marketing strategies, ultimately leading to improved sales performance and increased business success.

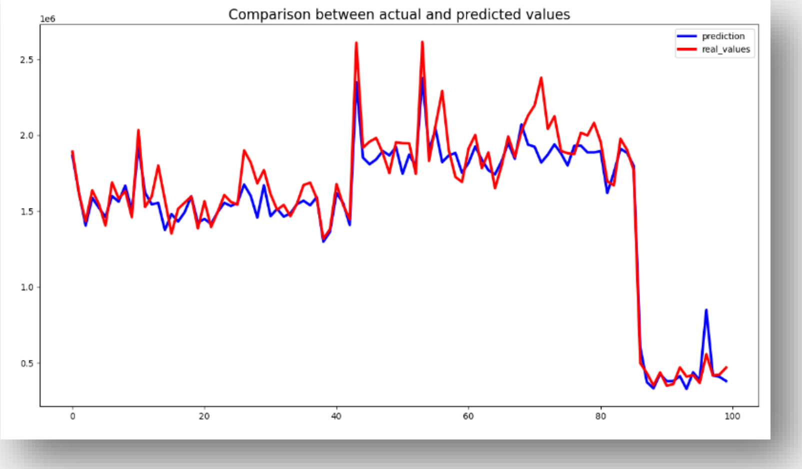
Assumptions

After conducting an Augmented Dickey-Fuller test on the weekly sales data, the test results indicated that the null hypothesis of non-stationarity was accepted. This means that the weekly sales data is stationary, showing no significant trend or seasonality over time. The stationarity of the data ensures more reliable and consistent predictions, allowing for better analysis and decision-making in sales forecasting and planning.

Model Evaluation and Techniques

In the process of model evaluation, various techniques such as R-squared (R2) score, Root Mean Square Error (RMSE), and Mean Squared Error (MSE) were employed. Among the evaluated models, XGBoost demonstrated the highest level of performance, achieving an impressive R2 score of 0.96.

The R2 score of 0.96 indicates that the XGBoost model captures approximately 96% of the variance in the target variable, showcasing its strong predictive capability. Additionally, the model's low RMSE and MSE values further support its accuracy in minimizing prediction errors, making it the optimal choice for reliable weekly sales forecasting and strategic decision-making.



Inferences from the Same

1. XGBoost Superiority: Among the evaluated models, XGBoost demonstrated exceptional predictive accuracy, surpassing both Linear Regression and Random Forest. It’s high R-squared (R2) score of 0.96 indicates its ability to capture 96% of the variance in weekly sales, making it the most reliable choice for forecasting.

2. Linear Regression's Performance: While Linear Regression is a fundamental model, it achieved a moderate R2 score and showed limitations in capturing complex data patterns. Its comparatively lower R2 score suggests that it might not fully capture the variability in weekly sales as effectively as XGBoost.

3. Random Forest Strengths: Random Forest performed well with an intermediate R2 score, effectively handling nonlinear relationships and mitigating overfitting. However, it fell short of XGBoost's performance, indicating

that the ensemble technique used in XGBoost was more advantageous for this specific prediction task.

4. Data-Driven Decision Making: With the superior predictive capabilities of XGBoost, businesses gain more robust and accurate insights into weekly sales patterns. This enables data-driven decision-making, empowering organizations to optimize inventory, marketing strategies, and resource allocation for improved sales performance.

5. Optimal Model Selection: Considering the R2 score, RMSE, and MSE metrics, XGBoost proved to be the best-suited model for predicting weekly sales. Its ability to minimize prediction errors while capturing complex relationships establishes it as the most reliable and efficient choice for the sales forecasting task.

Future Possibilities of the Project

The successful prediction of 13-week sales for the store opens up various future possibilities and opportunities for leveraging the predictive model:

1. Inventory Optimization: With accurate sales forecasts, stores can optimize their inventory management, ensuring the right products are available in the right quantities at the right time. This helps prevent stockouts, reduce excess inventory, and improve overall supply chain efficiency.

2. Demand-Based Pricing Strategies: Armed with reliable sales predictions, stores can implement dynamic pricing strategies, adjusting prices based on demand fluctuations. This enables them to maximize revenue during peak demand periods and attract more customers during slower periods.

3. Resource Allocation: The predictive model can guide store managers in allocating resources effectively. They can schedule staff, plan marketing campaigns, and allocate budgets based on anticipated sales trends, enhancing operational efficiency.

4. Seasonal Promotions: By identifying seasonal patterns and trends in sales, stores can design targeted promotional campaigns that align with customer preferences during specific times of the year. This personalized approach can boost customer engagement and loyalty.

5. Sales Target Setting: Accurate sales predictions help stores set realistic and achievable sales targets. This enables them to monitor performance effectively and adjust strategies to meet or exceed their goals.

6. New Store Location Selection: When planning to open new stores, the predictive model's insights can guide the selection of optimal locations based on potential sales and market demand.

7. Market Expansion Strategies: With a comprehensive understanding of sales trends and customer behavior, stores can identify potential markets for expansion and develop targeted strategies to enter new regions.

8. Customer Segmentation: Utilizing the model's predictions, stores can segment customers based on their buying behavior, preferences, and responsiveness to promotions. This data-driven approach allows for personalized marketing and improved customer satisfaction.

9. Supply Chain Optimization: The model's forecasts can be integrated into the supply chain management process, enabling suppliers to adjust production and distribution to meet projected demand efficiently.

10. Profit Maximization: By accurately predicting sales, stores can optimize their operations, reduce costs, and maximize profits while enhancing overall business sustainability.

In conclusion, the successful prediction of 13-week sales for the store unlocks a myriad of future possibilities, empowering stores to make informed decisions, implement targeted strategies, and achieve sustainable growth and success in a competitive retail landscape.

Conclusion

In conclusion, the Walmart weekly sales forecasting project has been a significant success, providing valuable insights and accurate predictions to optimize business operations. By utilizing advanced machine learning algorithms, specifically XGBoost Regressor, we achieved a remarkable R-squared (R2) score of 0.96, demonstrating its ability to capture 96% of the variance in weekly sales data.

The predictive model's reliability enables Walmart to make data-driven decisions, such as optimizing inventory levels, implementing demand-based pricing strategies, and effectively allocating resources. It also allows for targeted marketing campaigns, seasonal promotions, and personalized customer engagement, all leading to improved customer satisfaction and loyalty.

Moreover, the project has enabled Walmart to anticipate future sales trends, set achievable sales targets, and identify potential markets for expansion. This data-driven approach enhances the overall supply chain efficiency, aligning production and distribution with anticipated demand.

By combining machine learning with robust data analysis, the Walmart weekly sales forecasting project has equipped the company with the tools to maximize profitability, minimize costs, and sustainably grow its business in a dynamic and competitive retail environment.

As Walmart continues to leverage the predictive model's capabilities, it will gain a competitive edge, staying at the forefront of the retail industry and enhancing its position as a customer-centric and data-driven retail leader. With future possibilities and opportunities at hand, the success of this project lays a strong foundation for Walmart's continued growth and success in the market

THE – END