Walmart

WEIGHTED SALES FORECAST

Sohan Reddy Ganampeta

Introduction

Walmart Sales kaggle Contest

- Dataset Overview: Weekly sales data from 45 Walmart stores, covering 81 departments from February 5, 2010, to December 31, 2013.
- Holiday Sales Analysis: Data includes sales during Walmart's four major promotional holidays: Thanksgiving, Labor Day, Super Bowl, and Christmas.
- Data Processing: The dataset comprises three CSV files containing detailed information on dates, store configurations, and external factors affecting sales.
- Model Training and Testing:
  - Training Data: Data until September 7, 2012.
  - Testing Data: Used to evaluate model performance post-September 7, 2012.

Problem Statement

Problem Statement: Importance of Sales Forecasting for Walmart

- Strategic Planning: Essential for aligning business operations with market demand.

- Inventory and Revenue Management: Critical for optimizing stock levels and projecting future earnings.

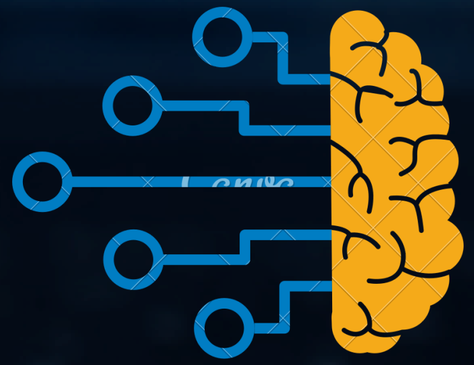
- Impact on Stock Prices: Direct influence on investor confidence and market valuation.

- Scope of Operations: Crucial for overseeing more than 10,000 global stores, ensuring efficient management and decision-making.

Objective

Exploring Walmart Data

Conduct extensive exploratory data analysis (EDA), preprocess data, visualize trends, and derive insights to understand underlying patterns and behaviors in Walmart's sales data.



Build a Robust Forecasting Model

Develop a sophisticated forecasting model that not only predicts sales but also incorporates external variables such as unemployment rates, fuel prices, and holiday effects to improve prediction accuracy.

Apply Time Series Analytics

Utilize time series analysis techniques to better understand market fluctuations, helping Walmart to anticipate future demands and adjust strategies accordingly.

Data Description

Overview of Walmart Store Sales Dataset

- Timeframe Covered: Sales data from February 5, 2010, to November 1, 2012.

Fields Included:

1. Store: Identifies the store number.
2. Date: Represents the week the sales data was recorded.
3. Weekly_Sales: Total sales for the specified store during the given week.
4. Holiday_Flag: Indicates whether the week included a major U.S. holiday (1 if Yes, 0 if No).
5. Temperature: Ambient temperature on the day of sales in degrees Fahrenheit.
6. Fuel_Price: Cost of fuel in the region on the day of sales.
7. CPI: Consumer Price Index, indicating the economic inflation or deflation for goods and services.
8. Unemployment: Percentage of unemployment in the region during the date of sales.

Key Holidays Considered:

- Super Bowl: Feb 12, 2010; Feb 11, 2011; Feb 10, 2012; Feb 8, 2013
- Labor Day: Sep 10, 2010; Sep 9, 2011; Sep 7, 2012; Sep 6, 2013
- Thanksgiving: Nov 26, 2010; Nov 25, 2011; Nov 23, 2012; Nov 29, 2013
- Christmas: Dec 31, 2010; Dec 30, 2011; Dec 28, 2012; Dec 27, 2013

Exploratory Data Analysis

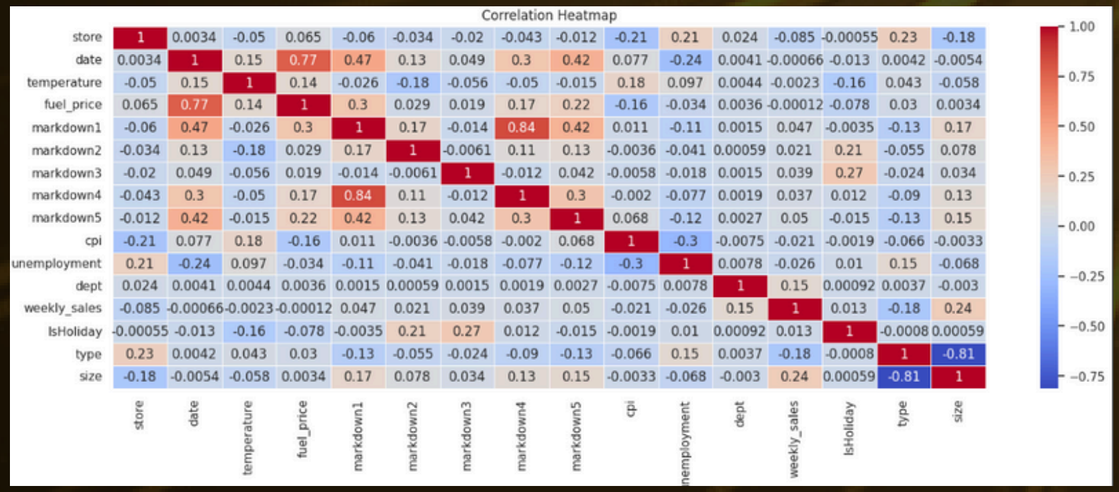
- Sales Variability: Weekly sales vary significantly across different stores and departments, indicating the influence of store-specific factors and local demographics on sales performance.

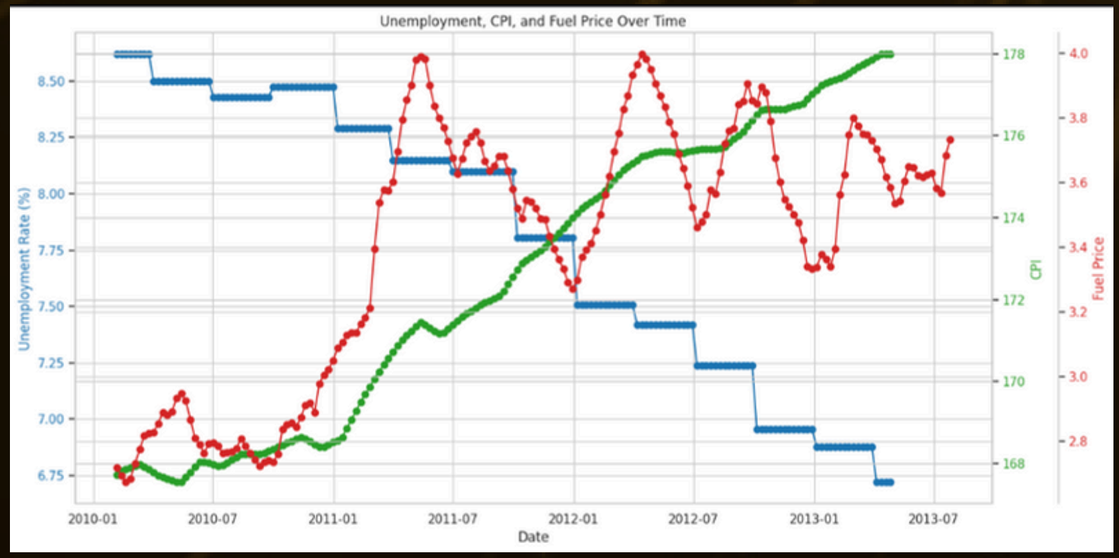
- Impact of Holidays: Sales spikes during holiday weeks (e.g., Thanksgiving, Christmas) suggest that holidays have a significant impact on sales, necessitating holiday-specific forecasting models.

- External Economic Factors: Correlation analysis shows that fuel prices, unemployment rates, and the Consumer Price Index (CPI) are moderately correlated with weekly sales, underscoring the importance of including these economic indicators in the sales forecasting model.

- Temperature Influence: There is a notable variability in sales with changes in temperature, suggesting weather-dependent buying patterns, particularly in specific departments like Gardening or Apparel.

- Markdown Effects: Markdowns, especially Markdowns 1 and 3, show a potential relationship with increased sales, indicating that price reductions play a crucial role in driving sales volume.





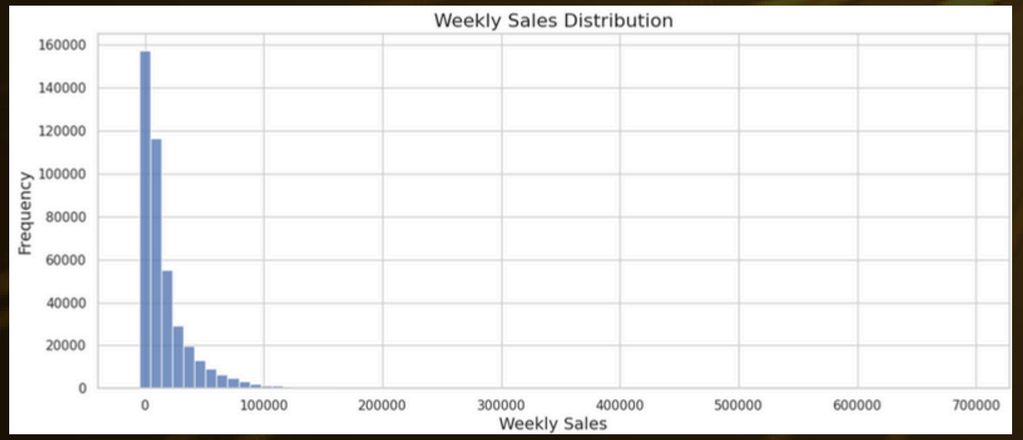
Exploratory Data Analysis

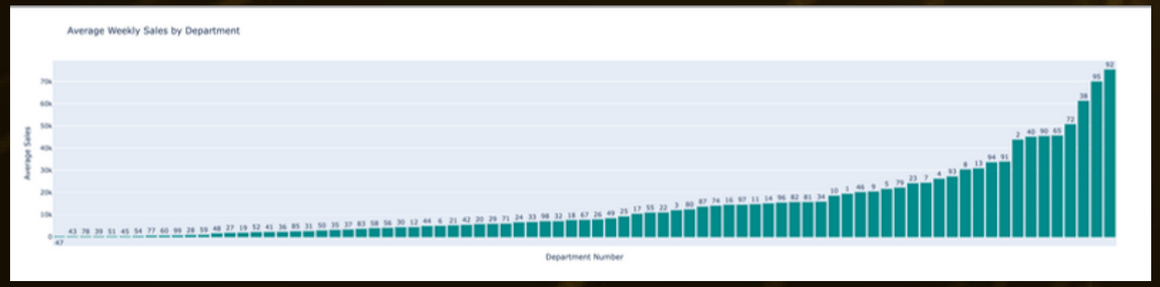
- Data Sparsity in Markdowns: Large amounts of missing data in markdown fields could affect the model's accuracy, highlighting the need for imputation strategies or sensitivity analysis regarding these inputs.

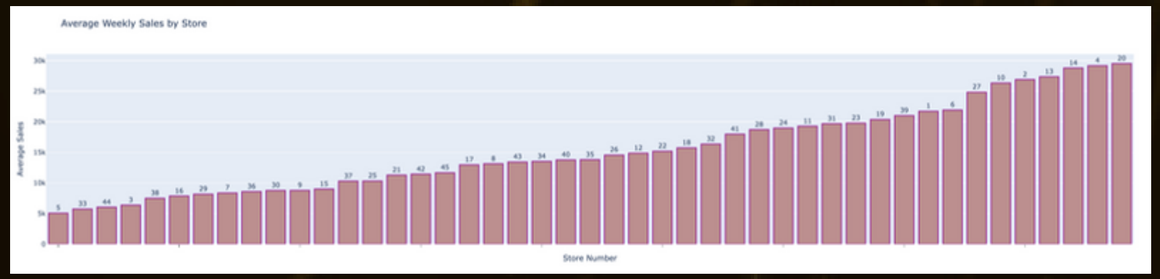
- Store Size and Type: Larger stores and specific store types (e.g., Type A) tend to have higher sales, suggesting that store size and type are significant predictors that should be incorporated into the forecasting models.

- Seasonal Patterns: Sales demonstrate clear seasonal trends, with peaks typically around major holidays and during certain seasons, which can guide inventory and staffing decisions.

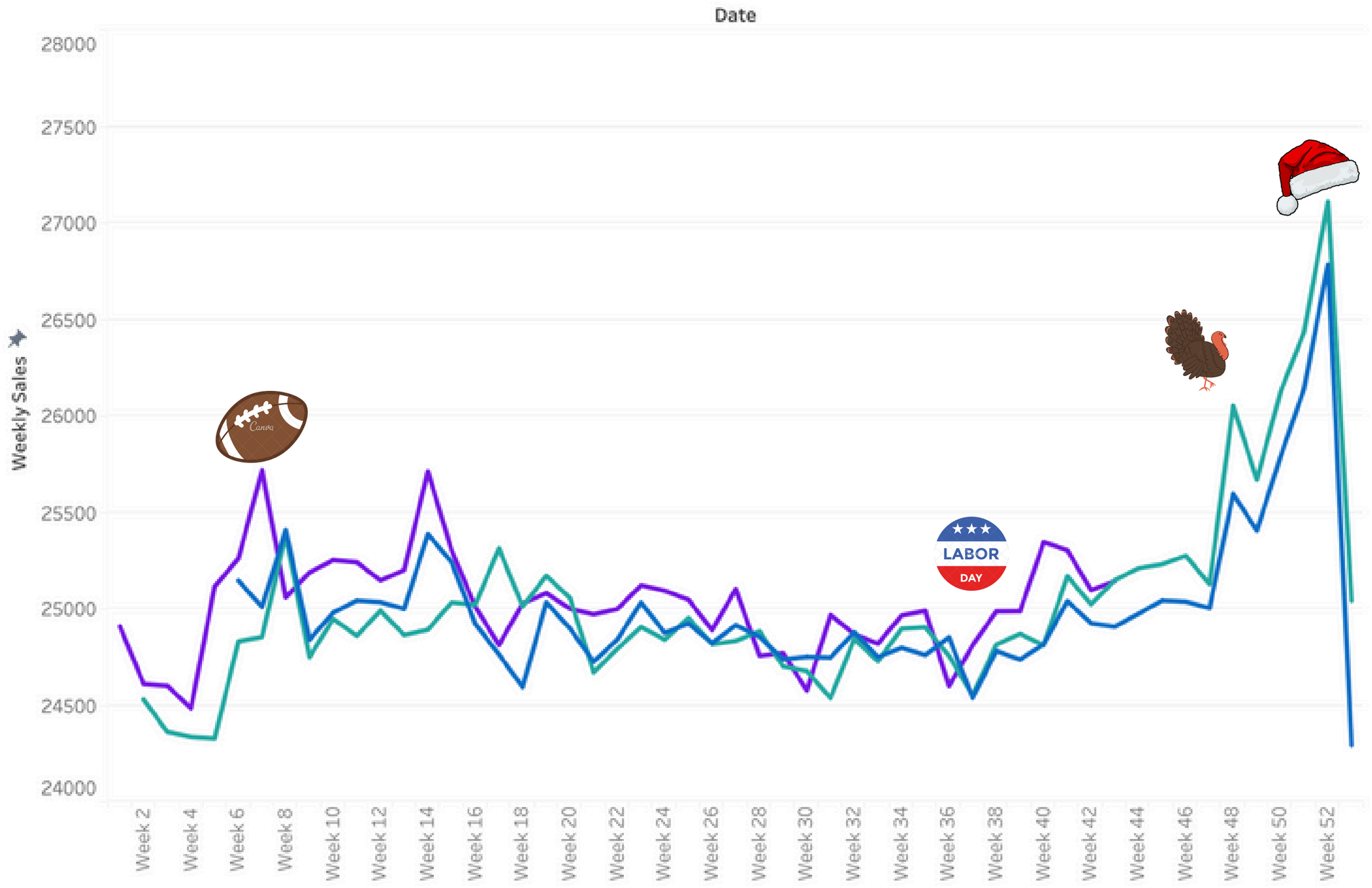
- Long-Term Trends: CPI and unemployment rates show long-term trends over the data collection period, which could be predictive of broader economic impacts on consumer spending behavior.



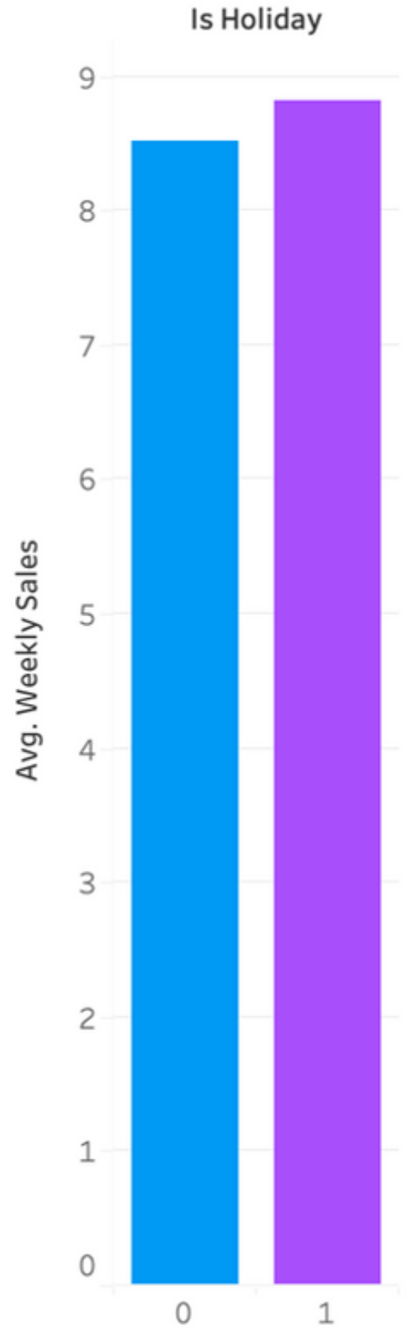




Exploratory Data Analysis



The trend of sum of Weekly Sales for Date Week. Color shows details about Date Year.



Machine Learning Models Comparison

- Understanding Model Limitations:
  - Models with high MSE and low R^2 scores are less capable, indicating issues like overfitting or underfitting, and are not recommended for future use.
  - Prioritize models that capture a comprehensive range of data patterns effectively.

- Models to Advance With:
  - Random Forest (100 estimators): Reliable for consistent high performance. Best fit for general forecasting needs with a balance of accuracy and computation efficiency.
  - XGB Regressor (250 estimators): Best predictive performance with the highest R^2, suitable for complex patterns.
  - Gradient Boosting (500 estimators): Good for scenarios where model interpretability is crucial despite slightly lower performance than XGB.

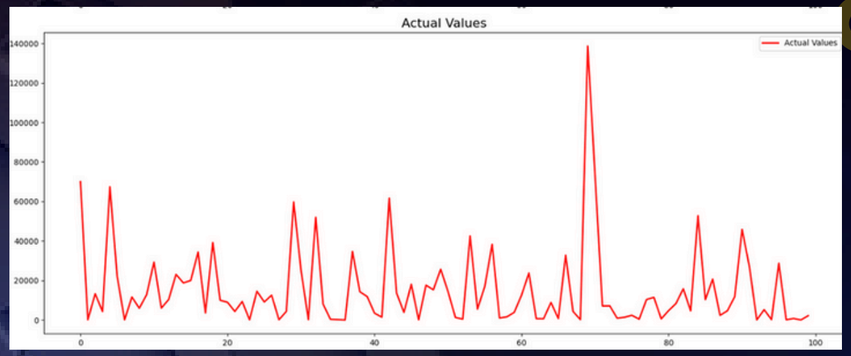
Model	Estimators	Mean Squared Error (MSE)	R^2 Score	Moving Forward
Linear Regression	N/A	470,345,208.58	0.088	No - Inadequate performance.
Ridge Regression	N/A	455,905,800	0.116	No - Outperformed by ensemble methods.
Lasso Regression	N/A	456,138,500	0.115	No - Outperformed by ensemble methods.
Random Forest	10	13,061,010	0.975	Yes - Excellent accuracy.
Random Forest	50	12,745,890	0.975	Yes - Slightly improved accuracy with more estimators.
Random Forest	100	12,745,890	0.975	Yes - Consistent high performance with increased estimators.
Gradient Boosting	100	179,390,500	0.652	No - Lower performance, consider increasing estimators.
Gradient Boosting	250	132,145,400	0.744	Yes - Improving with more estimators.
Gradient Boosting	500	101,713,300	0.803	Yes - Best results within Gradient Boosting models.
XGB Regressor	100	40,205,660	0.922	Yes - Highly effective with substantial R^2 score.
XGB Regressor	250	21,241,970	0.959	Yes - Superior performance, optimal choice.
XGB Regressor	500	21,241,970	0.959	Yes - Stable performance with increased estimators.

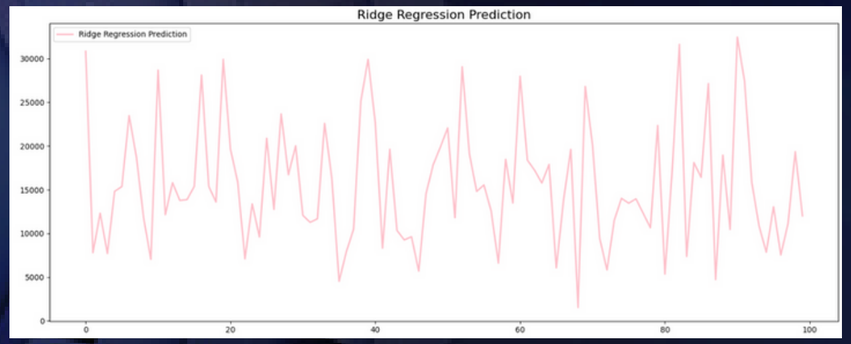
Linear Models: Lasso vs. Ridge Regression

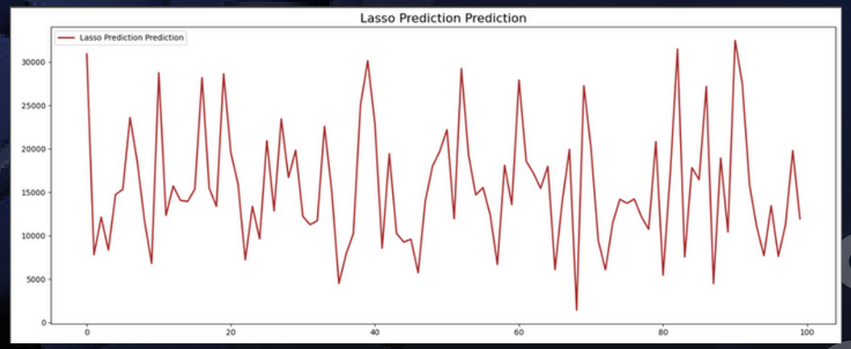
- **Model Overview:**
  - Lasso and Ridge Regression are linear models with regularization. Lasso uses L1 regularization which can zero out feature coefficients, while Ridge uses L2 which does not zero out coefficients but minimizes their effect.

- **Performance Analysis:**
  - Visual comparison of prediction results shows both models struggle with high variability in data, failing to capture complex patterns.
  - Both models exhibit high Mean Squared Error and low R^2 scores, indicating poor fit to the data.

- **Conclusion:**
  - Linear models are insufficient for handling the complexity and scale of Walmart's sales data due to their limitations in modeling non-linear relationships.





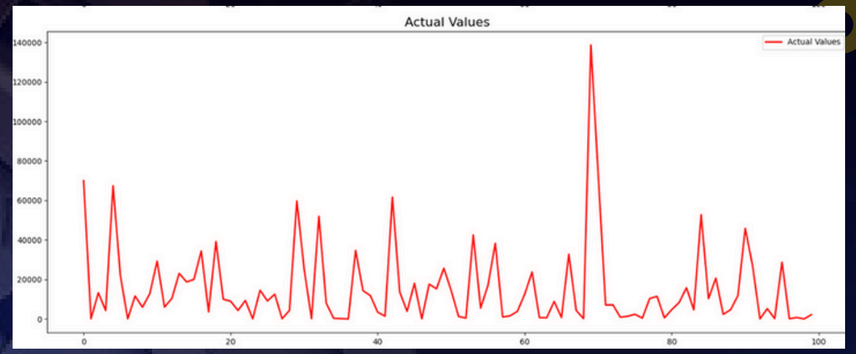


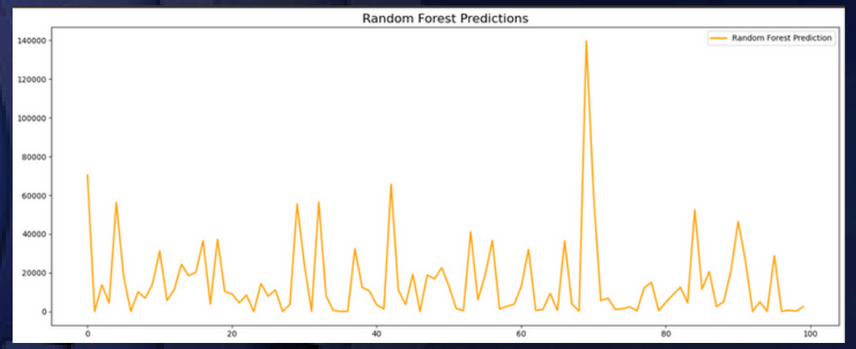
Tree-Based Models: Random Forest vs. Gradient Boosting

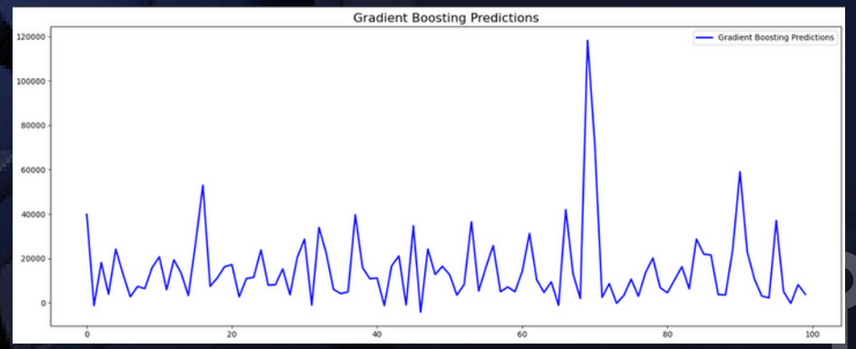
- **Model Overview:**
  - Random Forest and Gradient Boosting are both ensemble methods that build multiple trees to make decisions. Random Forest reduces variance by averaging multiple deep decision trees, while Gradient Boosting builds trees sequentially to correct previous trees' errors.

- **Performance Analysis:**
  - Both models show improved performance over linear models with significantly better R^2 scores and lower MSE, indicating a better grasp of data complexities.
  - Gradient Boosting tends to perform slightly better in handling extreme values and non-linear patterns compared to Random Forest.

- **Conclusion:**
  - Tree-based models prove more effective for our dataset, capturing more complex relationships and interactions between features.





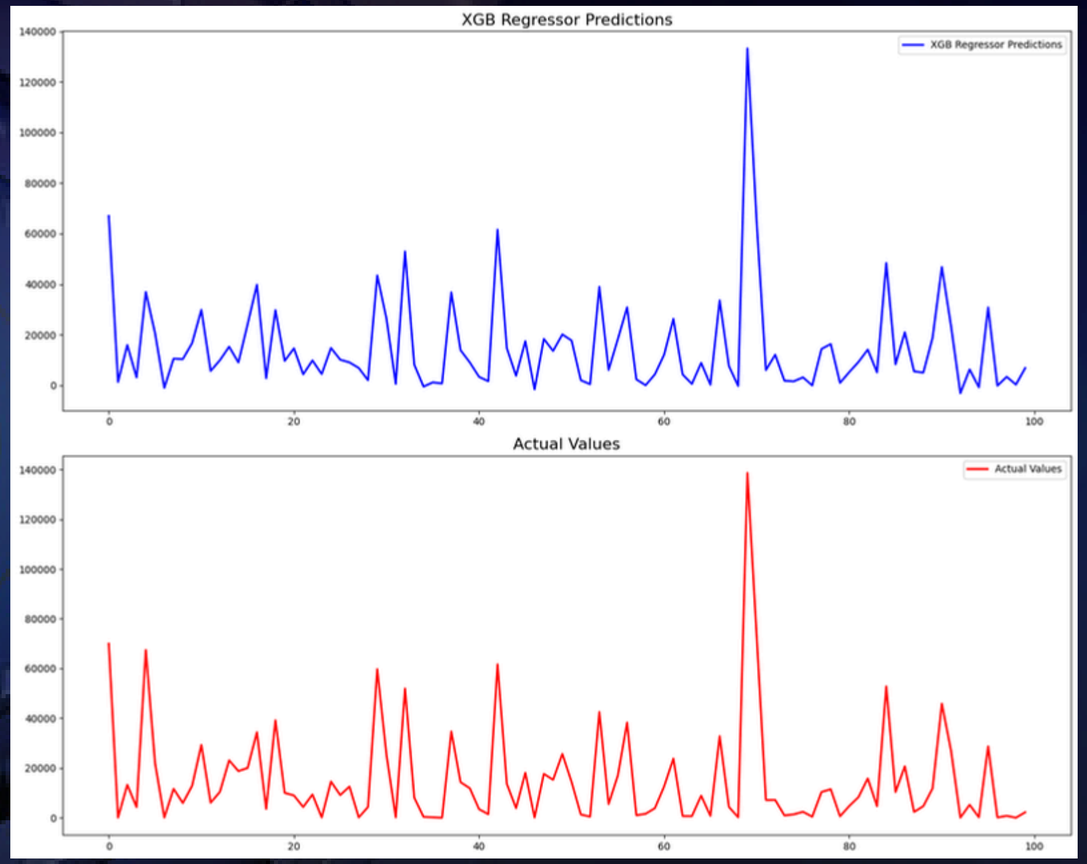


Optimal Performance with XGBoost

- Model Overview:
  - XGBoost, an advanced implementation of gradient boosting, offers fine-tuning of parameters and handles large data sets efficiently.

- Performance Analysis:
  - XGBoost outperforms all previous models, achieving the highest R^2 values and the lowest MSE, effectively capturing both trends and anomalies in sales data.
  - The model's ability to handle multidimensional data and provide actionable insights through feature importance scores makes it particularly valuable.

- Conclusion:
  - Given its superior predictive power and flexibility, XGBoost is selected as the best model for forecasting Walmart's weekly sales, providing a robust foundation for strategic decision-making.



Time Series Analysis with ARIMA and SARIMA

Optimal ARIMA Configuration:

- Best ARIMA Order: (0, 1, 5) for non-seasonal patterns.
- Achieved AIC: 3050.93, indicating a strong model fit.
- Model RMSE: 6,263,405.38, quantifying forecast error magnitude.

SARIMA Model Performance:

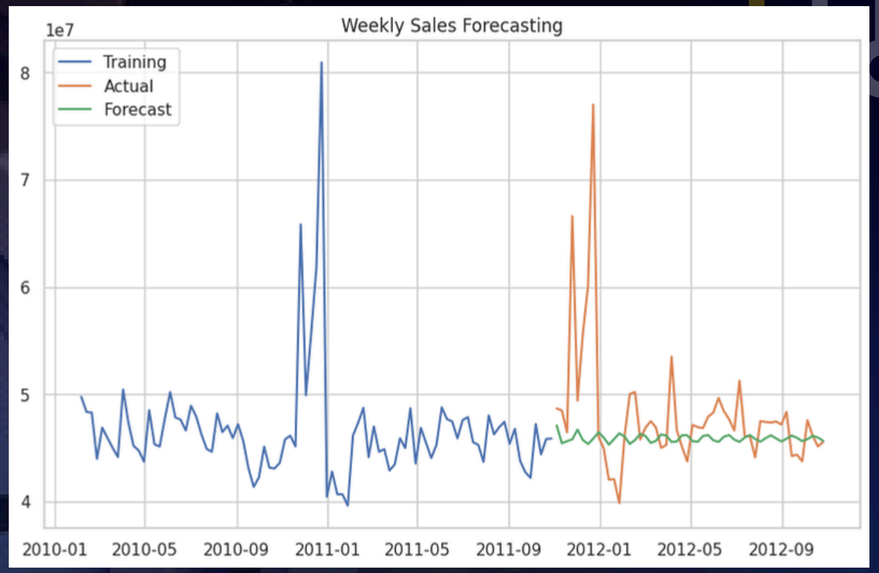
- Best SARIMA Order: Adjusted for seasonal variations based on data.
- Model RMSE: 2,128,341.29, significantly lower than ARIMA, reflecting better accuracy with seasonal adjustments.

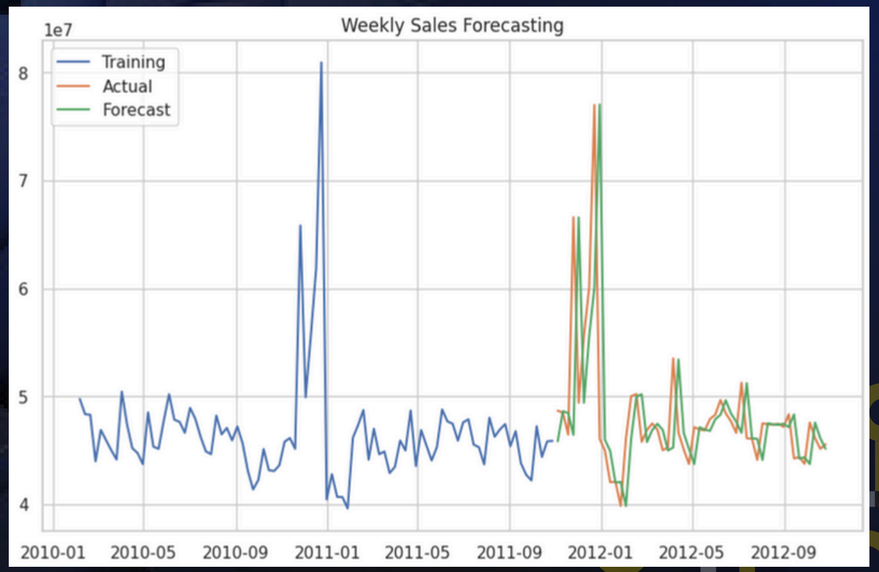
Insights and Model Implications:

- ARIMA's high RMSE suggests limited capability in handling high variability or seasonal spikes.
- SARIMA's lower RMSE underscores its effectiveness in capturing seasonal trends and forecasting more accurately for retail sales cycles.

Recommendations for Improvement:

- Investigate deeper into specific seasonal patterns & promotional impacts.
- Consider refining the ARIMA model with additional explanatory variables or switching to more complex models like SARIMA for all forecasting needs due to its enhanced accuracy.



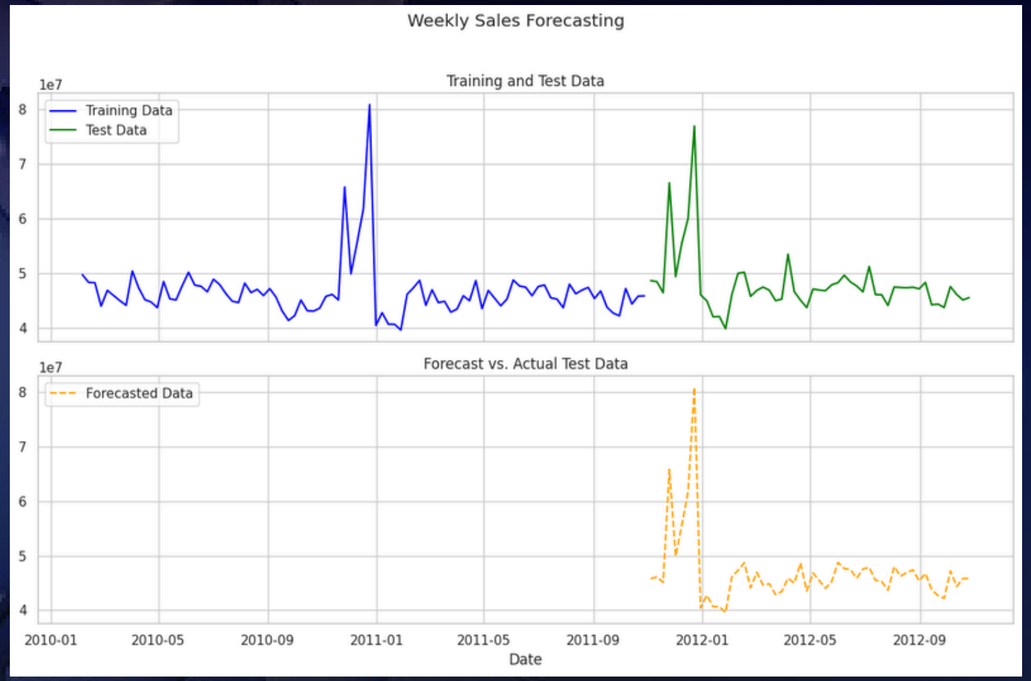


Model Accuracy and Performance Metrics

1. Key Accuracy Metrics:
   - Mean Absolute Error (MAE): 1,597,388.98 - Reflects average errors in prediction, highlighting prediction deviations.
   - Root Mean Squared Error (RMSE): 2,128,341.29 - Indicates the standard deviation of the prediction errors.
   - Mean Absolute Percentage Error (MAPE): 10.14% - Shows errors as a percentage of actual values, useful for comparability.

2. Overall Model Performance:
   - Accuracy Rate: 89.86% - Measures the proportion of accurate predictions, useful for assessing model reliability.
   - High MAE and RMSE point to notable variances, particularly in peak sales periods, suggesting room for model refinement.

3. Strategic Implications and Recommendations:
   - Given the significant error rates, consider enhancing the model's accuracy by incorporating more detailed data, such as specific promotional activities and holiday impacts.
   - Focus on refining prediction models to better capture high variability periods, which could improve strategic decision-making and forecasting reliability.




```
Mean Absolute Error (MAE): 1597388.9843533235
Root Mean Squared Error (RMSE): 2128341.295087557
Mean Absolute Percentage Error (MAPE): 10.139488649093506%
Accuracy: 89.86051135090649%
```

Limitations

- The amount of data lagged systems, especially during Grid-search CV, better hyperparameters could have been obtained with more tuning.
- Despite that, the data is still limited to 45 stores. As of 2022, there are 4742 Walmart stores as of 2022. As a result the models we built are greatly constrained which could lead to reduced accuracy
- It becomes difficult to predict consumer purchasing behavior over changing periods of time, a model built from the data in 2011 cannot accurately predict in 2013 without understanding external factors
- Difficult to observe whether our observations meet industry benchmarks without comparing it to competitors like Target, Amazon, etc.

Conclusion

- Our analysis identifies key factors influencing sales variability, such as promotions, holidays, and economic indicators like fuel prices and unemployment rates. Recognizing these drivers allows for targeted strategies to mitigate risks and capitalize on opportunities.

- Dynamic Pricing Strategies: Implement dynamic pricing based on predictive insights to adjust prices in response to external economic factors and consumer demand trends.

- Prioritize accuracy in forecasting for high-sales periods identified in the analysis, such as holiday seasons and promotional events, to ensure optimal stock levels and staffing.

- Continuously refine forecasting models by integrating real-time data and feedback from sales performance to improve accuracy. Consider exploring more sophisticated machine learning models that can dynamically adapt to new patterns.

- Encourage the sales and marketing teams to use the insights from the predictive models to plan better and make informed decisions that drive sales growth and customer satisfaction.

- Invest in advanced analytics platforms and training for staff to leverage these tools, ensuring the sales team can interpret model outputs and make data-driven decisions effectively.

Future Work + Recommendations

- Customer segmentation is an area Walmart can focus on. Understanding consumer buying behavior at a department and store level could prove profitable for increasing revenue.

- The average sales are marginally higher during holidays than non-holidays, this can be a huge area of improvement. Early access deals combined with marketing can be utilized to improve holiday sales

- Year over year forecasted sales should be compared with the actual sales to understand shifts in consumer behavior and also maintain an updated customer database to ensure accurate forecasts

- Walmart needs to focus on other strategies along with cost-leadership to avoid thin profit margins and can have a better e-commerce online presence

Thank You!

Walmart

Questions?