CAPSTONE PROJECT

ANALYSIS OF WALMART DATASET

Brendan C Rosario

TABLE OF CONTENTS

PROBLEM STATEMENT

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

PROBLEM OBJECTIVE

The goal of this project is to analyze historical sales data to understand demand patterns and develop machine learning models to predict future sales. By leveraging advanced predictive analytics, the project aims to:

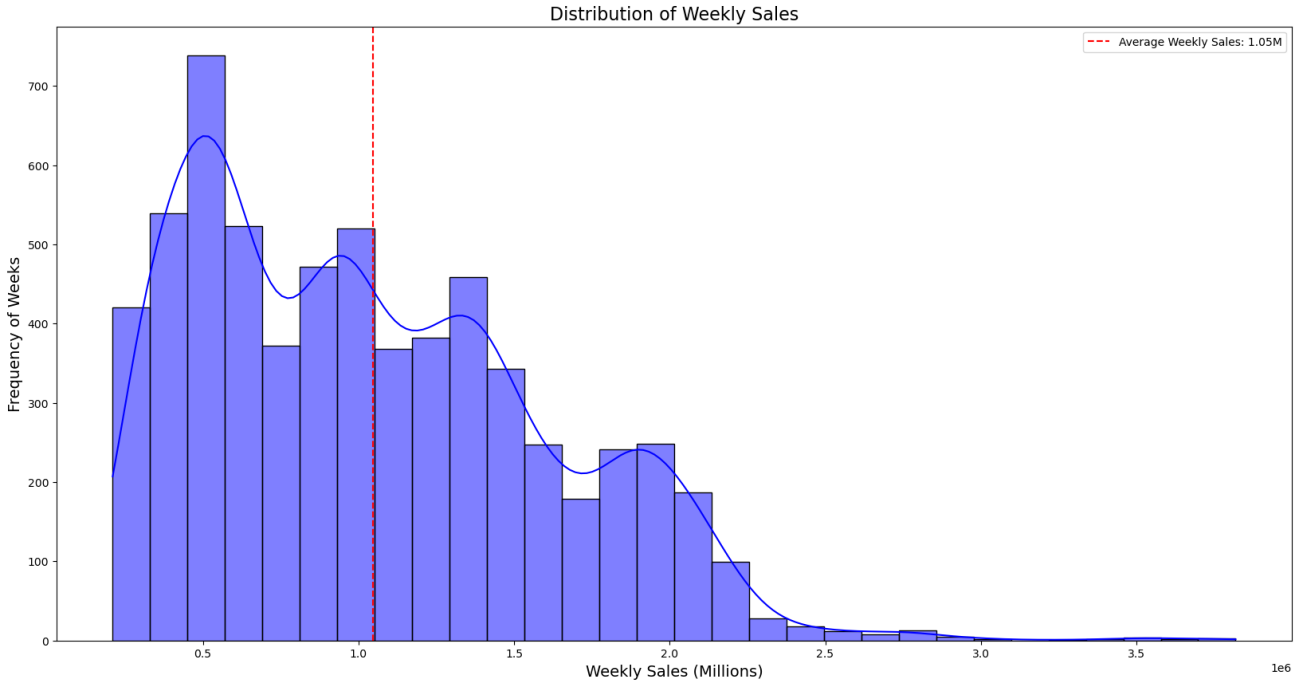
- Identify key factors influencing sales fluctuations, such as holidays, economic conditions, and regional trends.
- Develop accurate forecasting models to optimize inventory management, ensuring that stock levels align with consumer demand and reducing instances of overstocking or stockouts.
- Provide actionable insights for decision-making, enabling store managers and supply chain planners to proactively adjust their strategies.
- Enhance business efficiency by minimizing revenue loss due to poor demand prediction, improving customer satisfaction by ensuring product availability, and optimizing operational costs.
- Compare different machine learning techniques to determine the most effective approach for predicting retail sales with high accuracy.

DATASET DESCRIPTION

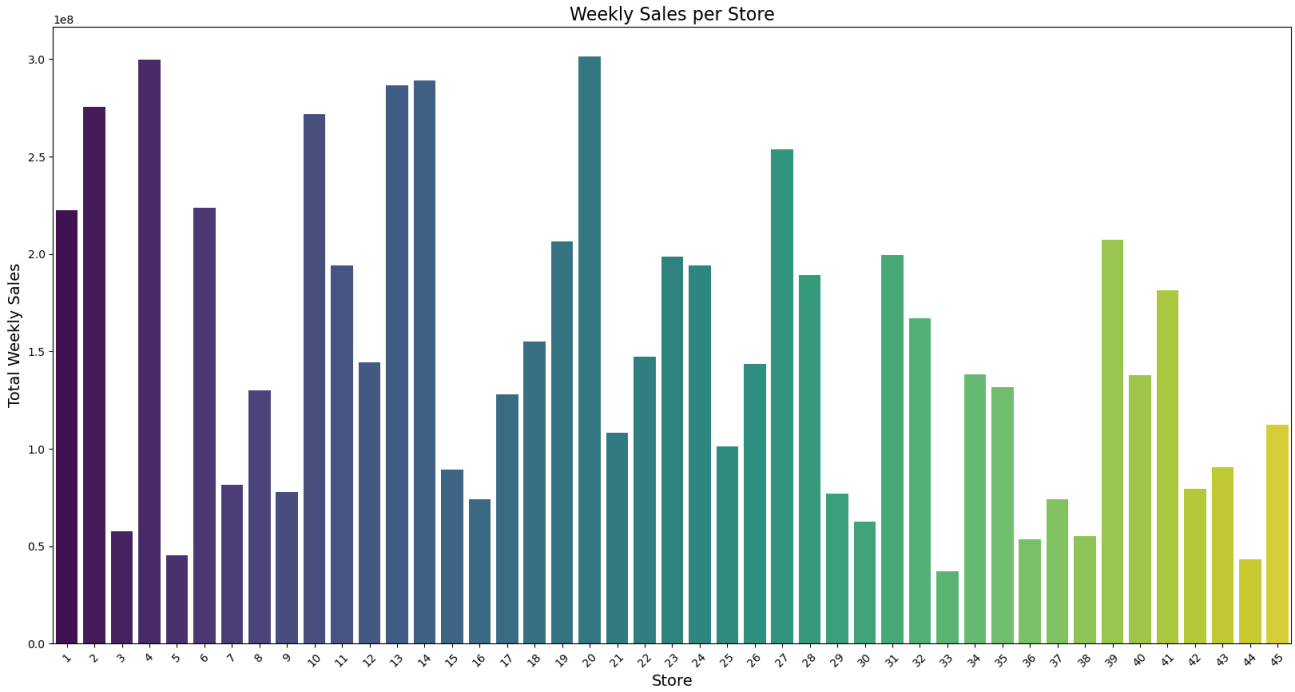
- **Store**: The store number.
- **Date**: The date of the sales record.
- **Weekly_Sales**: The total sales for a given store and date.
- **Holiday_Flag**: Indicates whether the week includes a holiday (1) or not (0).
- **Temperature**: Average temperature for the store's location.
- **Fuel_Price**: Fuel price in the region of the store.
- **CPI (Consumer Price Index)**: Economic indicator measuring the change in price levels.
- **Unemployment**: The unemployment rate in the region.
- **Year, Month, Day**: Extracted from the Date column for trend analysis.

INSIGHTS FROM THE DATA

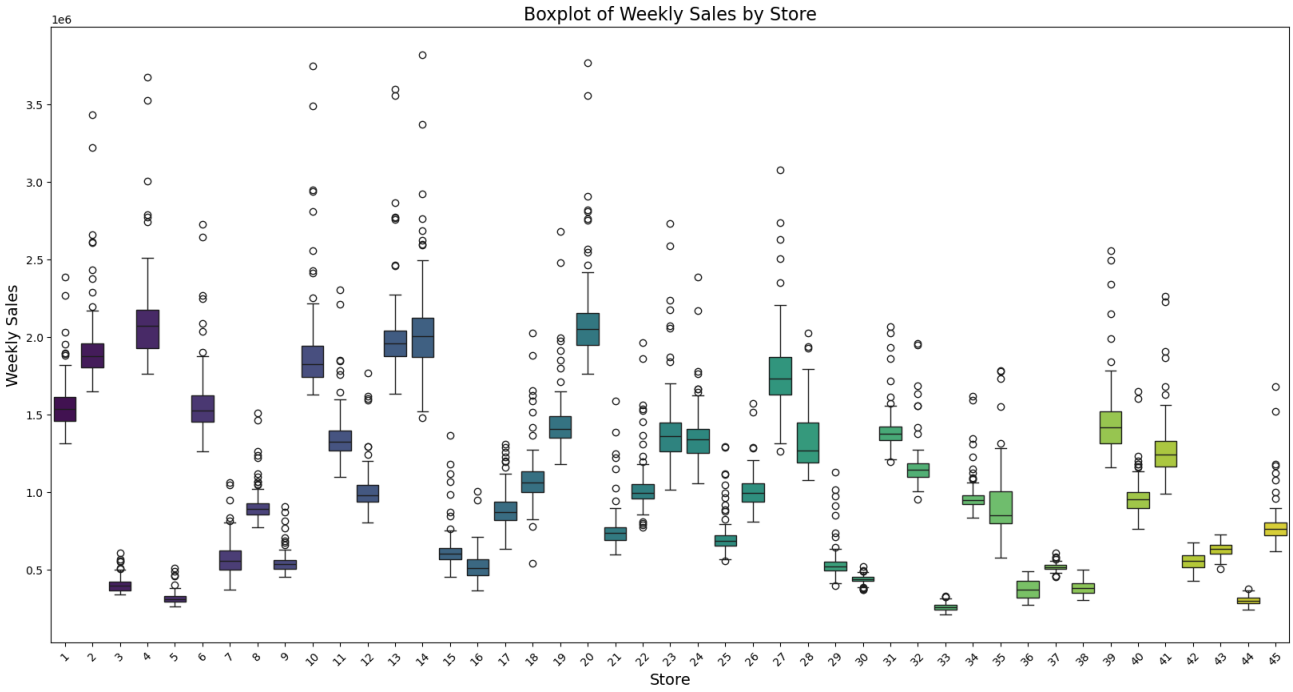
- **Sales Distribution**: The histogram of weekly sales shows a right-skewed distribution, with sales values ranging between $0.25 million and $2.2 million. The majority of stores have sales around the mean value of approximately $1.1 million.



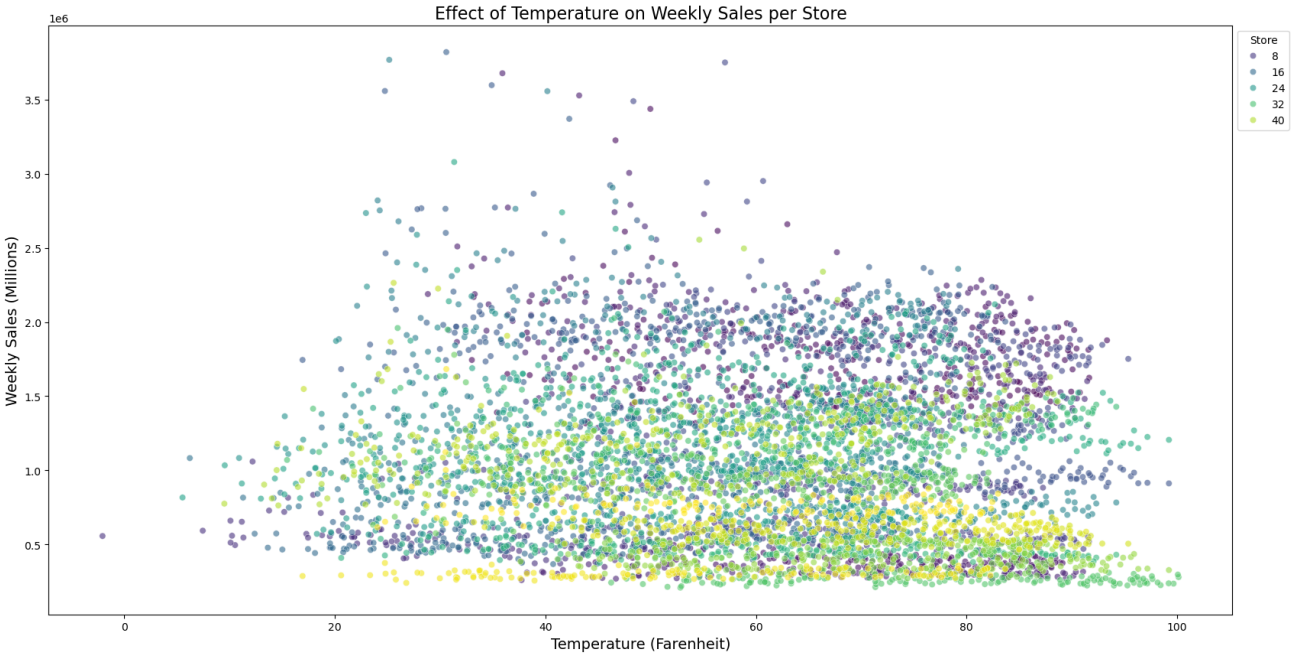
- **Store-wise Sales**: The bar chart comparing total weekly sales across different stores reveals that certain stores consistently outperform others in sales volume, highlighting differences in demand based on store location or size.



- **Holiday Impact**: A boxplot analysis comparing holiday and non-holiday weeks shows a clear increase in sales during holiday weeks, confirming the seasonal impact on sales trends.



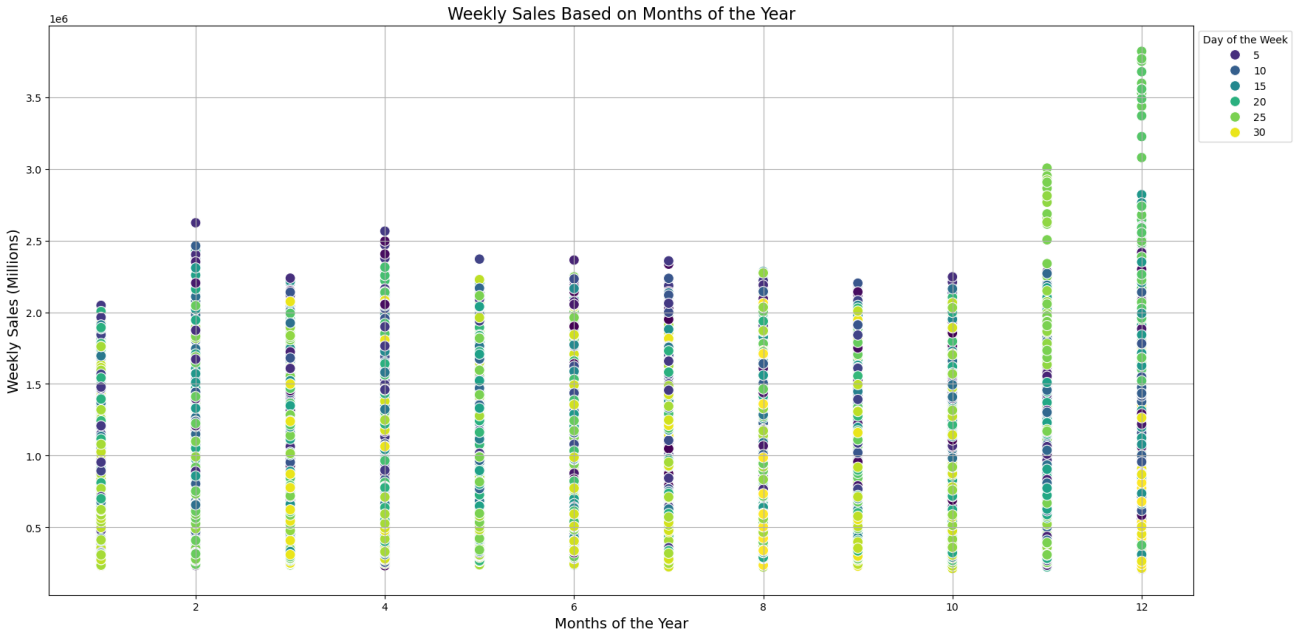
- **Temperature Influence**: A scatter plot between temperature and weekly sales indicates a weak negative correlation, suggesting that extreme temperatures may reduce sales in certain regions.



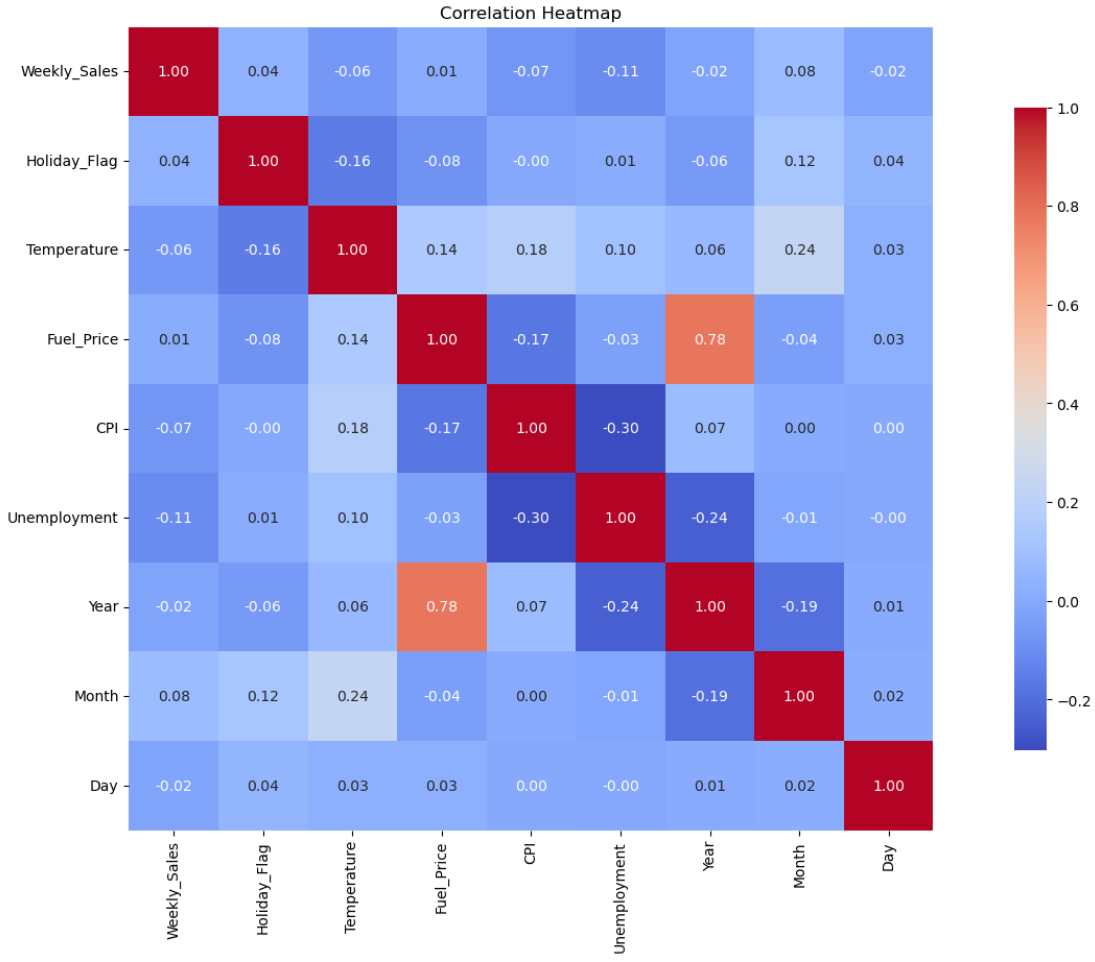
- **Fuel Price & CPI**: The scatter plots show that fuel price variations do not have a strong direct impact on sales, while CPI changes exhibit a moderate correlation, indicating that economic conditions influence purchasing power.

- **Unemployment Rate**: The scatter plot of unemployment vs. weekly sales shows a slight negative correlation, suggesting that higher unemployment may lead to lower sales volumes.

- **Monthly Trends**: A line plot showing average weekly sales over different months indicates that sales peak in November and December, aligning with holiday shopping trends.



- **Correlation Heatmap**: A heatmap depicting the correlation between all numerical features suggests that Weekly Sales have notable correlations with CPI, Unemployment, and Holiday Flags, indicating key influencing factors.



DATA PRE-PROCESSING STEPS AND INSPIRATION

To prepare the data for model training, the following preprocessing steps were applied:

- **Datetime Conversion**: The Date column was converted into a datetime format to enable time-series analysis and feature extraction.
- **Feature Engineering**: Extracted Year, Month, and Day from the Date column to capture seasonal trends and sales variations over time.
- **Handling Missing Values**: Checked for missing values in the dataset and confirmed there were none, ensuring data consistency.
- **Data Normalization and Scaling**: While tree-based models like Random Forest and XGBoost do not require feature scaling, standardization was considered for potential improvements in model interpretability.
- **Outlier Detection**: Identified extreme values in Weekly_Sales to assess potential data anomalies and ensure robust model performance.
- **Exploratory Data Analysis (EDA)**: Conducted comprehensive EDA using visualizations such as histograms, scatter plots, and heatmaps to understand feature relationships and uncover patterns in the data.
- **Data Splitting**: The dataset was split into training and testing sets using an 80-20 ratio to evaluate model performance effectively.

INSPIRATION FOR PREPROCESSING STEPS

- The decision to extract Year, Month, and Day was inspired by the evident seasonality in sales trends observed in the line plot.

- Checking for missing values was essential to ensure the completeness and reliability of the dataset before model training.
- The correlation heatmap suggested that certain features, such as Holiday_Flag and CPI, strongly impact sales trends, leading to further emphasis on feature selection.
- Scatter plots and boxplots guided the decision to explore feature importance and assess which external factors influence sales the most.

CHOOSING THE ALGORITHM FOR THE PROJECT

Three models were selected for the prediction task:

- **Linear Regression**
- **Random Forest Regressor**
- **XGBoost Regressor**

MOTIVATION AND REASONS FOR CHOOSING THE ALGORITHM

- **Linear Regression**: This model was chosen as a baseline to evaluate the linear relationship between independent variables and sales. It is simple and interpretable, allowing us to understand general trends, but may not capture complex patterns in the data.
- **Random Forest Regressor**: This ensemble learning method was selected due to its ability to handle non-linearity and interactions between variables. It performs well on structured data and reduces overfitting by averaging multiple decision trees.
- **XGBoost Regressor**: XGBoost was selected for its high predictive accuracy, efficiency, and ability to handle large datasets. It is an advanced gradient boosting algorithm that captures intricate relationships between variables and provides better generalization than Random Forest in many cases.

ASSUMPTIONS

- Historical sales trends will continue into the future.
- The features used (temperature, fuel price, CPI, and unemployment) influence sales.
- Holiday weeks have a significant impact on sales.
- Data quality and consistency are maintained.

MODEL EVALUATION AND TECHNIQUES

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: MAE calculates the absolute difference between actual and predicted values, then averages them. It provides a direct measure of the average error magnitude in the predictions. A lower MAE indicates a more accurate model.
- **Root Mean Squared Error (RMSE)**: RMSE is the square root of the mean squared differences between actual and predicted values. It penalizes larger errors more than MAE, making it more sensitive to outliers. RMSE is useful when large errors are particularly undesirable.
- **R-squared (R^2)**: The R^2 score measures how well the independent variables explain the variability in the dependent variable. A higher R^2 score (closer to 1) indicates that the model captures the variance in the data well, whereas a lower score suggests a poor fit.

Model	MAE	RMSE	R^2
Linear Regression	4,32,598.3	5,21,597.61	0.1555
Random Forest	58,102.33	1,14,408.15	0.9593
XGBoost	**49,296.78**	**86,031.13**	**0.9770**

INFERENCES DRAWN FROM THE MODEL EVALUATION

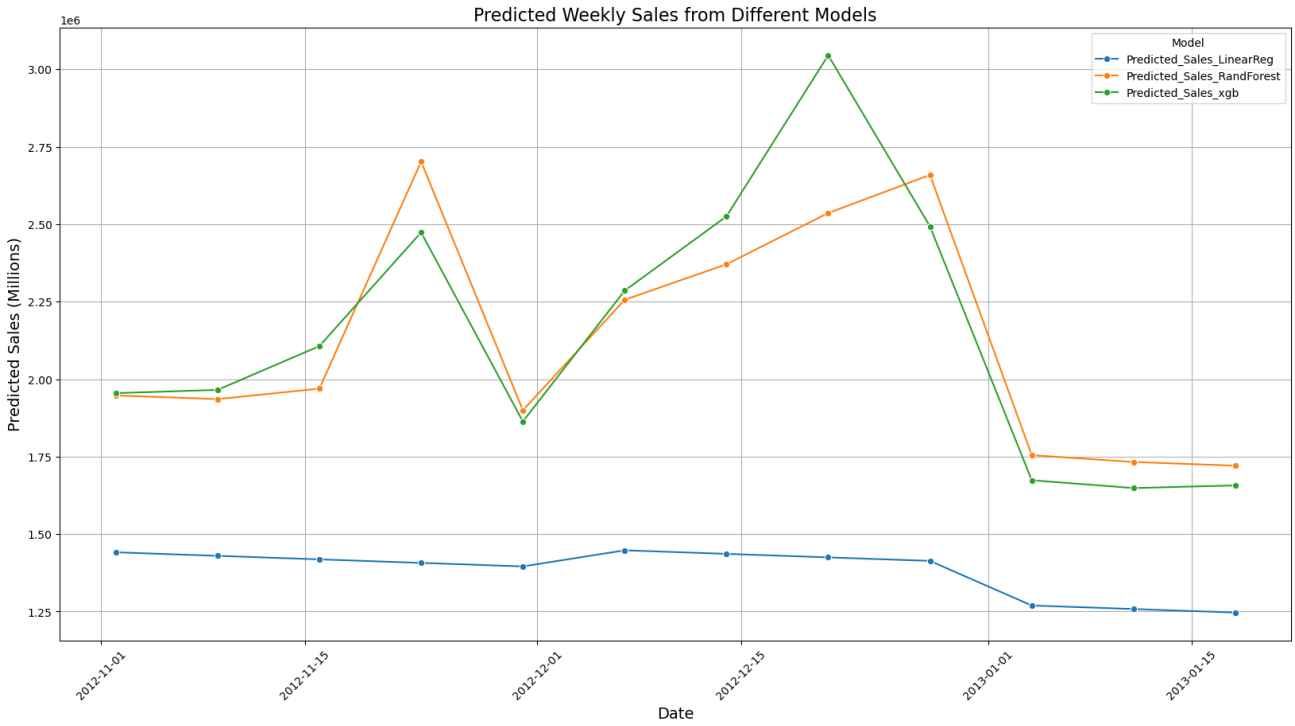
- **XGBoost outperformed all models**, achieving the lowest MAE and RMSE, and the highest R^2 score (97.7%). This suggests that XGBoost effectively captures complex relationships within the dataset and provides highly accurate forecasts.
- **Random Forest also demonstrated strong performance**, with an R^2 of 95.9%, making it a reliable choice for predictions. It is particularly useful for capturing non-linear dependencies but slightly less accurate than XGBoost.
- **Linear Regression underperformed significantly**, with a low R^2 score of 0.1555, indicating that it failed to capture the underlying patterns in the data. This highlights that sales predictions are influenced by complex factors that cannot be adequately modeled by a simple linear approach.

FUTURE POSSIBILITIES OF THE PROJECT

- **Incorporating additional data sources**, such as promotions, competitor pricing, and local events.
- **Implementing real-time sales prediction models** for dynamic inventory management.
- **Developing a recommender system** for inventory restocking based on predicted demand.
- **Exploring deep learning techniques**, such as LSTMs or neural networks, for enhanced forecasting.
- **Enhancing customer segmentation models** to personalize product availability based on regional preferences.
- **Implementing cloud-based solutions** for scalable and real-time analytics, enabling rapid decision-making across multiple store locations.

PREDICTION FOR THE NEXT 12 WEEKS

Using the trained models, sales predictions were generated for the next 12 weeks.



Data in a table format:

Date	Predicted_Sales_LinearReg	Predicted_Sales_RandForest	Predicted_Sales_xgb
2012-11-02	14,41,042	19,47,445	19,54,807.375
2012-11-09	14,29,659	19,35,564	19,65,319.250
2012-11-16	14,18,275	19,69,302	21,06,474.000
2012-11-23	14,06,892	27,01,979	24,73,599.250
2012-11-30	13,95,509	18,99,389	18,63,367.375

Date	Predicted_Sales_LinearReg	Predicted_Sales_RandForest	Predicted_Sales_xgb
2012-12-07	14,47,459	22,56,103	22,85,445.000
2012-12-14	14,36,076	23,70,646	25,25,251.000
2012-12-21	14,24,693	25,36,144	30,45,179.250
2012-12-28	14,13,310	26,58,916	24,92,878.500
2013-01-04	12,69,377	17,54,705	16,73,689.875
2013-01-11	12,57,994	17,32,598	16,48,419.375
2013-01-18	12,46,610	17,20,631	16,56,941.000

Predicted weekly sales values suggest that businesses should plan for demand surges and allocate inventory accordingly. Future improvements may involve integrating real-time data to enhance prediction accuracy.

CONCLUSION

This project analyzed retail sales data and built predictive models using XGBoost, the most effective approach with high accuracy and low error rates.

Insights from this study optimize inventory planning, enhance supply chain efficiency, and reduce stockouts and overstocking. External factors like holidays, economic indicators, and regional demand variations impact sales trends.

Future enhancements include integrating real-time analytics, deep learning, and reinforcement learning to refine predictions and drive data-driven decisions. Adopting these strategies helps businesses stay competitive, improve profitability, and serve customers better.

REFERENCES

- Scikit-learn documentation: Linear Regression, Random Forest Regressor
- XGBoost documentation: XGBoost Regressor