**IDS561 Project Report**

**Walmart Sales Forecasting**

**Group- 10 Members:**

Akhilesh Sunil Joshi UIN - 669332717

Harsh Mehta UIN - 659163943

Kranti P. Yeole UIN – 655817383

**Problem Setting:**

Sales forecasting is essential for businesses to optimize their inventory, project revenue, and inform investment decisions. For a large company like Walmart, understanding seasonal variations in sales is particularly critical. Accurate forecasts help the company meet its financial targets from the start of each season, positively influencing stock prices and investor confidence. Conversely, failing to meet sales projections can lead to significant declines in stock value. Therefore, effective sales forecasting not only supports operational efficiency but also impacts financial stability and market perception.

The problem we are solving with the Walmart Sales Forecasting project involves predicting future sales across various Walmart stores and departments using historical data available from a Kaggle dataset. The primary objective is to develop a model or set of models that can accurately forecast weekly sales, considering factors such as seasonality, holidays, promotional events, and possibly external economic conditions. This problem is fundamentally about leveraging data analytics and machine learning techniques to analyze past sales patterns and use these insights to predict future sales performance.

**Data Description:**

In the Walmart Sales Forecasting project, we utilize a comprehensive dataset sourced from Kaggle, designed to enable accurate sales predictions across various Walmart stores. The dataset is divided into four distinct subsets: Features, Stores, Train, and Test. Each subset plays a crucial role in understanding the factors that influence sales and in building robust forecasting models.

A) Features Dataset

The Features dataset includes data such as temperature, fuel price, promotional markdowns, CPI (Consumer Price Index), and unemployment rates. These variables are recorded weekly and are specific to each store and date. The inclusion of these external factors allows for an analysis of how economic and environmental conditions affect sales.

B) Stores Dataset

The Stores dataset provides fundamental information about each of the Walmart stores, including the type of store and its size. This dataset is crucial for segmenting the stores in the analysis, allowing for differentiated strategies in sales forecasting based on store characteristics.

C) Train Dataset

The Train dataset contains historical sales data for 98 products across 45 Walmart stores, covering dates from 2010 to 2012. This dataset includes weekly sales figures and is linked to specific stores and departments, providing the training data for our forecasting models.

D) Test Dataset

The Test dataset mirrors the structure of the Train dataset but without the weekly sales figures, which are to be predicted. This dataset spans 39 weeks following the period covered in the Train dataset. The performance of our forecasting models is evaluated based on how accurately they predict sales in this dataset.

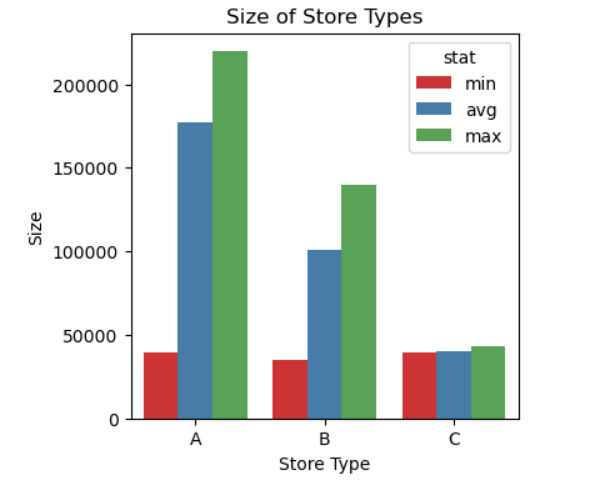
**Data Cleaning & Exploratory Data Analysis:**

Our EDA commenced with data cleaning operations, specifically targeting the handling of missing values in the markdown columns of the Features dataset. Recognizing the significance of accurate data representation, we converted the 'NA' string values to nulls, which enables us to treat these absent entries more appropriately in our subsequent analyses.

Additionally, we converted economic indicators such as CPI (Consumer Price Index) and the unemployment rate to a numeric data type, ensuring the data's compatibility with our analytical models. This meticulous data preparation facilitates the examination of sales trends and the influence of economic factors on sales performance.

Utilizing the Stores dataset, we aggregated the size data to find the minimum, average, and maximum store sizes for each type. This analysis is crucial as it provides insights into the scale of operations and potential sales volume that different store types may achieve.

Our findings are visualized in a bar plot, which succinctly illustrates the comparative size metrics across store types. From the bar plot, it is evident that Type A stores are the largest on average, with a considerable variation in size as indicated by the significant difference between the minimum and maximum values. Type B stores show less variation in size compared to Type A, with their average size being noticeably smaller. Type C stores are the smallest, with the least variation in size, suggesting a more uniform store size within this category.



To understand the temporal dynamics of Walmart's sales, we constructed a time series plot depicting weekly sales from February 2010 to October 2012. This visualization plays a pivotal role in identifying patterns, trends, and potential anomalies in the sales data over time.

The line plot generated reveals the following observations:

Seasonality: There are conspicuous peaks suggesting seasonal trends in sales, which could correspond to national holidays, promotional events, or seasonal shopping periods like back-to-school or Christmas.

Sales Average: A horizontal dashed line indicates the average weekly sales over the entire period, providing a benchmark against which to compare weekly performance. The average line facilitates the identification of weeks with significantly higher or lower sales than usual.

Volatility: The time series displays volatility in sales, with some weeks showing substantial spikes, while others are more consistent with the average. This variation points to the need for a forecasting model that accounts for irregularities and has the flexibility to adapt to such fluctuations.

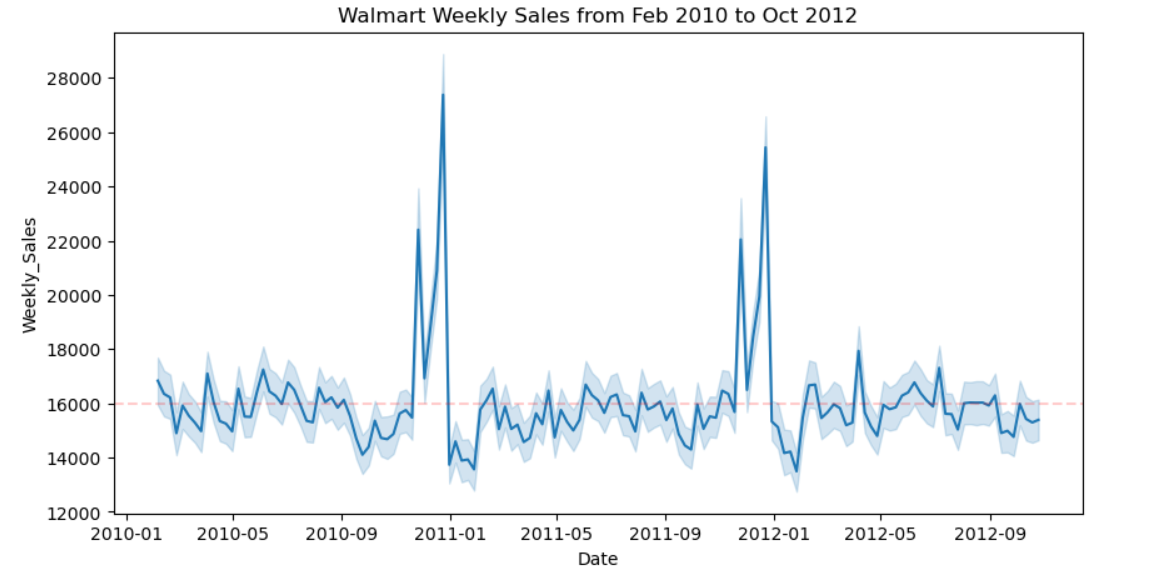


Fig: Sales from Feb 2010 - Oct 2012

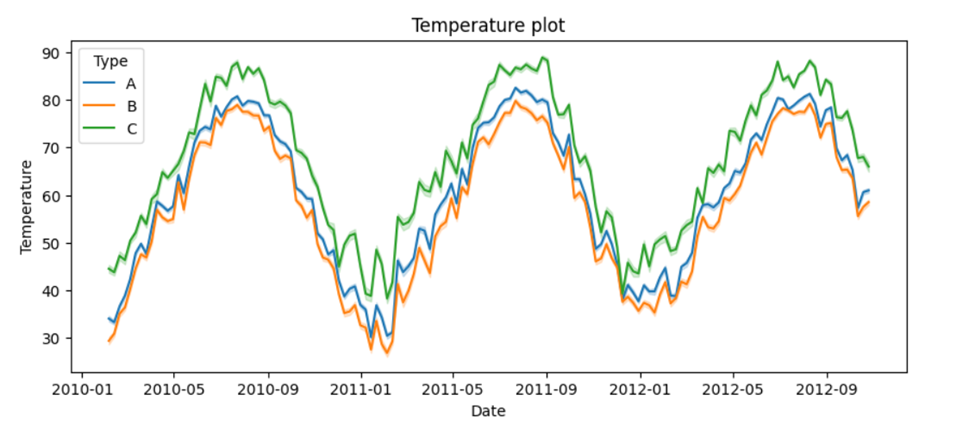
To understand the environmental factors that may influence shopping behavior and, consequently, sales performance, we conducted an analysis of temperature trends across different types of Walmart stores.

The temperature data, plotted over time from January 2010 to September 2012, reveals discernible patterns correlated with seasonal changes. The line plot includes three temperature trend lines, each representing one of the three store types: A, B, and C.

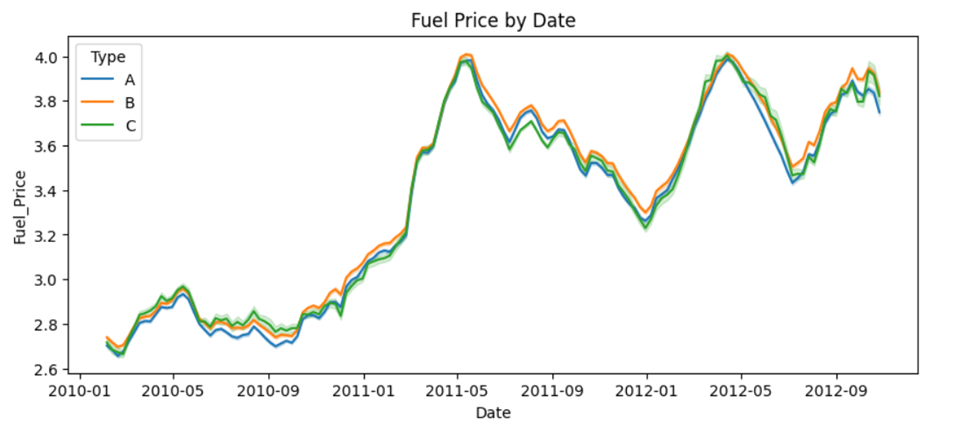
Observations from the plot suggest the following:

Seasonal Patterns: All store types exhibit a similar pattern, following the natural progression of seasons with temperatures peaking during the summer months and dipping during the winter months.

Store Type Correlation: While the temperature trends for store types A, B, and C follow a similar seasonal pattern, there are noticeable differences in temperature levels. This might indicate geographical differences in store locations, with some store types being situated in warmer or cooler climates.



We also analyzed the fuel patterns and how it affects the sales of Walmart. The trends for all store types appear to converge, suggesting that fuel prices are consistent across the different geographical locations of the stores or that the stores are similarly affected by nationwide fuel price changes. There are periodic peaks and troughs which may correspond with seasonal travel patterns, such as increased travel during holiday seasons, potentially influencing the cost of goods sold and consumer shopping behavior.



**Techniques:**

We employed three powerful machine learning techniques to develop accurate sales forecasting models: Decision Tree, Random Forest, and XGBoost.

Decision Tree is a tree-based model that recursively partitions the data based on the feature values, creating a tree-like structure of decisions and their possible consequences. It is a non-parametric supervised learning algorithm that can handle both regression and classification tasks.

Random Forest is an ensemble learning technique that constructs multiple decision trees on different subsets of the data and combines their predictions. It addresses the high variance problem of decision trees and improves the overall predictive performance.

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting, which is a powerful ensemble technique. It iteratively builds weak prediction models and combines them to create a strong predictive model. XGBoost is highly efficient, flexible, and has built-in regularization to prevent overfitting.

**Results:**

The performance metrics for each model are as follows:

Decision Tree Model:

- Accuracy: 87.81%

- RMSE (Training): 7662.26

- R-squared (Training): 0.8800

- RMSE (Testing): 7744.60

- R-squared (Testing): 0.8782

The Decision Tree model demonstrated moderate accuracy and a reasonable fit to the training data, but its generalization to unseen data was slightly weaker, as indicated by the slightly higher test RMSE and lower test R-squared.

Random Forest Model:

- Accuracy: 98.021%

- RMSE: 3121.822

- R-squared: 0.98021

The Random Forest model exhibited high accuracy and a strong fit to the data, with an R-squared value of 0.98021, indicating that the model explained 98.021% of the variance in the target variable. However, the presence of outliers may have affected the model's predictions, as suggested by the higher RMSE compared to the MAE (not reported).

XGBoost Model:

- Accuracy: 98.348%

- RMSE: 2852.098

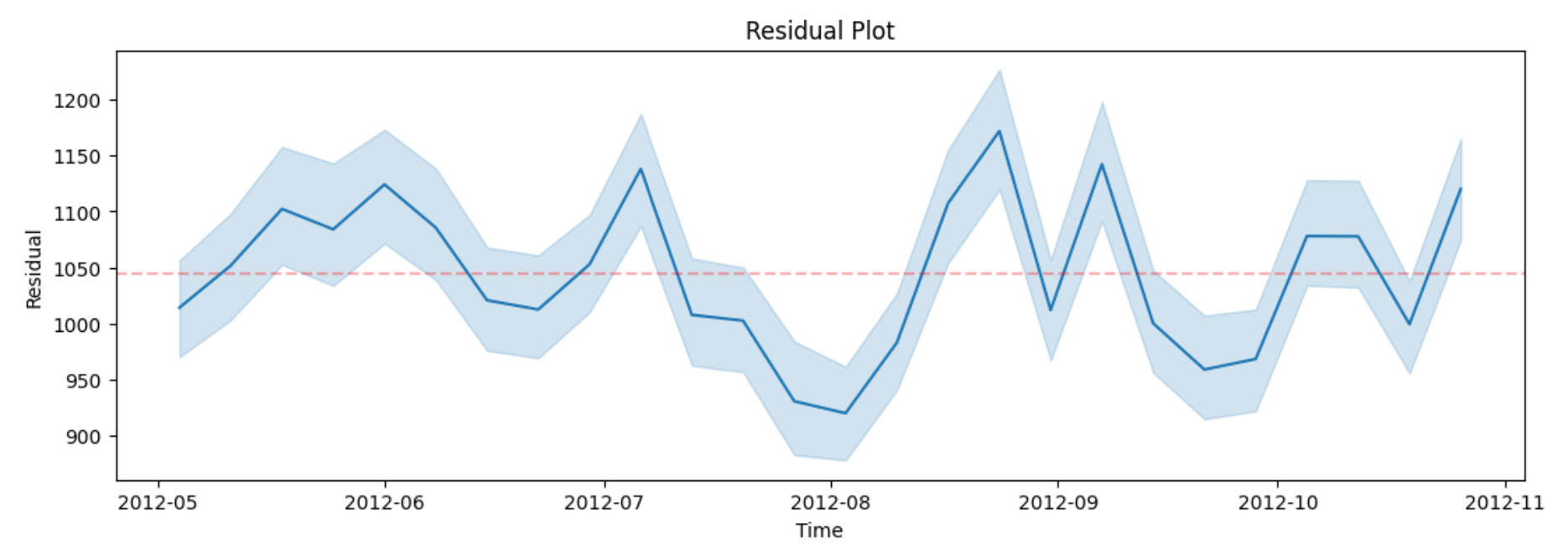
- R-squared: 0.98348

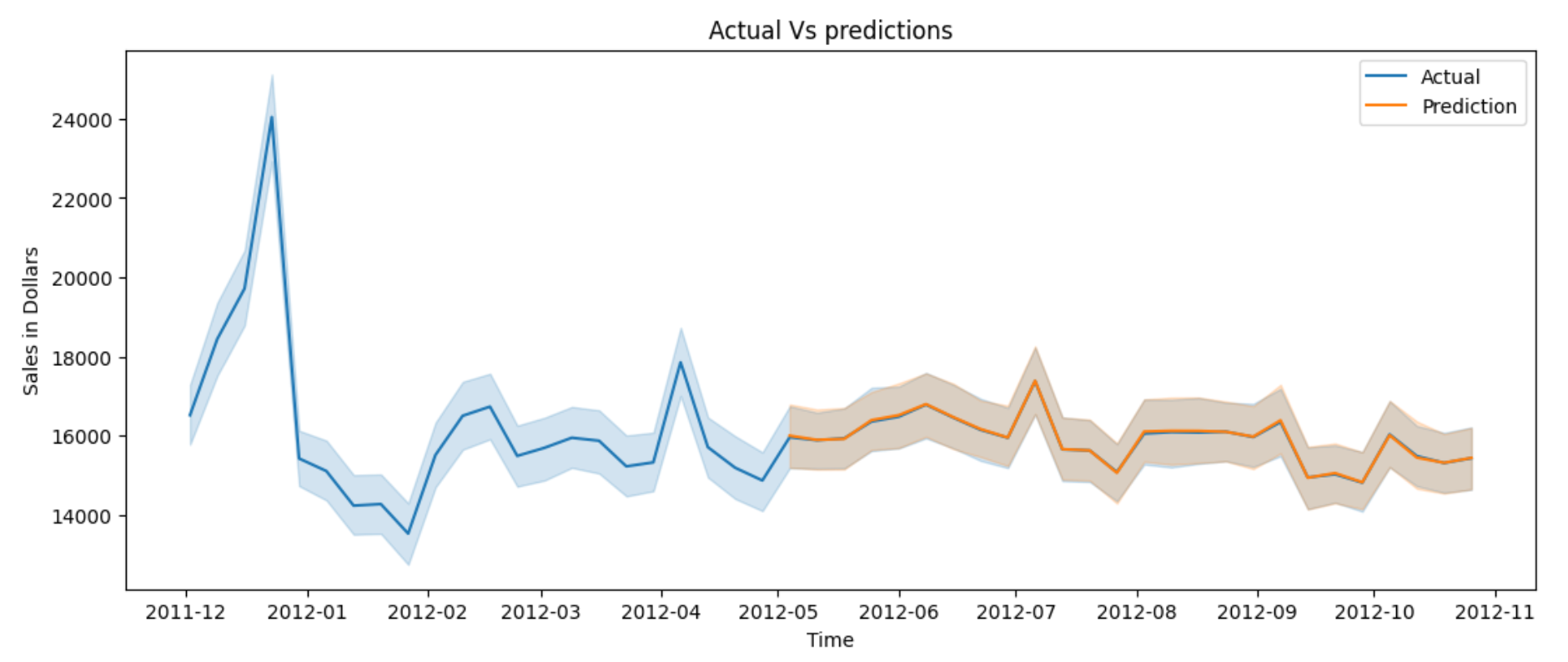
The XGBoost model outperformed the other two models, achieving the highest accuracy of 98.348% and the lowest RMSE of 2852.098. With an R-squared value of 0.98348, the model explained 98.348% of the variability in the sales data, indicating an excellent fit.

A screenshot of a phone

Description automatically generated

The XGBoost models feature importance analysis revealed that store size, department, store type, and location were the most significant factors influencing sales. Economic factors, such as CPI and unemployment rate, had a moderate to low impact.





The analysis of residuals reveals a stable performance of the model across various time periods, as indicated by their consistent fluctuation within a band width. Absence of discernible trends in the residuals suggests the model does not exhibit systematic underprediction or overprediction tendencies within specific periods. Notably, the consistency around the red dashed line, likely representing the zero line, implies minimal bias in the model. Moreover, the close alignment of the prediction band with actual sales patterns underscores the model's commendable accuracy in capturing sales trends. However, the variability in the width of prediction intervals hints at fluctuations in the model's confidence levels in its predictions.

A graph of a train

Description automatically generated

The rapid decrease followed by stabilization of RMSE values for both train and test datasets signify efficient learning of the model with minimal iterations required. Furthermore, the convergence of the train and test lines indicates that the model is not succumbing to overfitting tendencies, as evidenced by their close proximity. This suggests that the model's performance generalizes well beyond the training data, demonstrating its robustness and reliability in real-world applications.

**Conclusion:**

The project aimed to develop an accurate sales forecasting model for Walmart, considering various factors such as seasonality, holidays, promotional events, and economic conditions. After exploring the data and evaluating multiple machine learning techniques, the XGBoost model emerged as the best-performing model, achieving an accuracy of 98.348% and an RMSE of 2852.098.

The XGBoost models feature importance analysis provided valuable insights into the key factors influencing Walmart's sales, such as store size, department, store type, and location. This knowledge can inform strategic decision-making and optimization efforts for Walmart's operations and marketing strategies.

While the XGBoost model demonstrated excellent performance, there is still room for improvement. Further enhancements could include more detailed feature engineering, incorporating additional data sources (e.g., holiday-specific sales data, competitor sales data), and conducting analyses to identify high-demand and low-demand items for targeted inventory management and marketing campaigns.

Overall, the project successfully developed a robust sales forecasting model that can support Walmart's decision-making processes, enabling efficient inventory management, targeted marketing campaigns, and sustained business growth.

**References:**

* Walmart Sales Data. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast
* Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
* Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

**Role of Team Members:**

Kranti P. Yeole (655817383) took on a leadership role, guiding the team through the project lifecycle. Her strategic thinking and problem-solving abilities were instrumental in defining the project's objectives, developing a model to forecast weekly sales, and addressing challenges related to incorporating factors such as seasonality, holidays, promotional events, and economic conditions. Kranti's effective communication skills facilitated seamless collaboration within the team.

Akhilesh Sunil Joshi (669332717) brought invaluable technical expertise to the project. He is responsible for data preparation, including creating test and train datasets from the Walmart Sales Data and integrating relevant information from the Store Data and external sources. Akhilesh's attention to detail and proficiency in data analysis techniques played a crucial role in the exploratory data analysis phase, where they identified key patterns and insights related to store types, sales trends, and the impact of holidays and economic factors.

Harsh Mehta (659163943) contributed his creativity and adaptability to the project. He was instrumental in evaluating and implementing the machine learning models, including Decision Tree, Random Forest, and XGBoost. Harsh's innovative approach and collaborative spirit were vital in fine-tuning the models, interpreting the results, and conducting feature importance analysis to understand the key drivers of Walmart's sales. His fresh perspectives and open-mindedness facilitated the team's exploration of potential improvements and future enhancements.

Together, we formed a cohesive unit, leveraging our complementary skills and expertise. Our seamless collaboration and commitment to the project's success enabled the development of an accurate sales forecasting model that can support Walmart's decision-making processes, inventory management, and marketing strategies.