Alicia Qu

Final Report
Walmart Sales Forecast

Problem Statement:

In today's retail landscape, accurate sales forecasting is crucial for effective inventory management and operational planning. This project will develop Time series models to forecast the weekly sales of Walmart stores using historical sales data. This project aims to provide Walmart with a robust Time series model for sales forecasting, enabling better decision-making and resource allocation. The insights gained from this project could significantly improve the efficiency of inventory management and contribute to overall business success.

The potential client for this project could be the management team of Walmart, who want to forecast the weekly sales of different stores to help them to optimize their inventory management, logistics management and promotion strategy.

Data Wrangling

There are four datasets utilized in this project. Firstly, the "features" dataset encompasses all impactful features of the store, comprising promotion data, unemployment rates, CPI (Consumer Price Index), fuel prices, among others. Secondly, the "Stores" dataset details the store types categorized by store size, as well as the actual store sizes themselves. Thirdly, the "train" dataset contains information on weekly sales by department across various stores, alongside the corresponding sales collection dates. Lastly, the "test" dataset mirrors the structure of the "train" set but lacks sales data.

To streamline the analysis, I merged the "features," "Stores," and "train" datasets into a unified dataset. For this project, I'll exclusively focus on the "train" set, which I'll subsequently divide into training and validation sets.

This data set is clean, but still I did some minor wrangling on it. When investigating the categorical features, I noticed that there are three categories of stores based on their size. Category A comprises large stores, constituting approximately half of the total. Category B consists of medium-sized stores, while Category C includes relatively small-sized stores. Generally, larger stores tend to have higher sales. It is noteworthy that the minimum values for all three categories are 30,000+, which raises some curiosity and warrants further investigation. After look into three types separately, I find out that except for stores 33 and 36, all stores in Type A surpass the 150,000 marks in terms of size. This suggests a potential misclassification for these two specific stores. Two stores within Type B exhibit a size below 50,000, whereas all Type C stores fall below this threshold. This leads to a reasonable inference that any store with a size less than 50,000 should be categorized as Type C.

There are not too much missing values in this dataset, only the MarkDonw columns have missing values. MarkDonws are promotional markdowns that Walmart is running. So, it means that there is no promotion running at that time if the MarkDown values is missing. As a result, I will fill all the missing values with 0.

As I said before, this dataset is almost clean. After data wrangling, I keep most of the da ta and have a clean dataset which contains 421570 rows and 16 columns.

Exploratory Data Analysis

The analysis of store data reveals three store types based on size: Type A (large), Type B (medium), and Type C (small). Larger stores consistently have higher sales (Figure 1), and Type C stores exhibit a distinct pattern, with sales dropping sharply post-holidays (Figure 2).

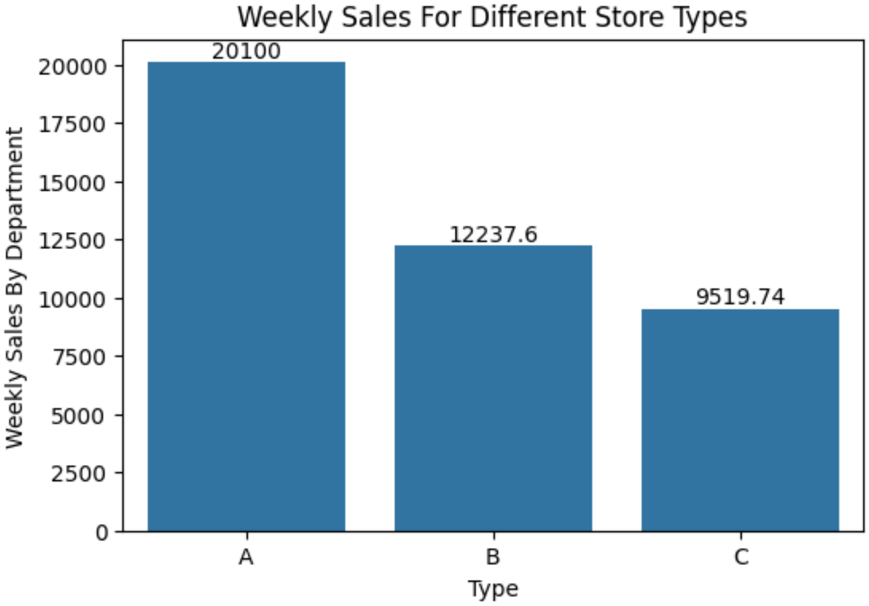

Figure 1 Average weekly sales by store type

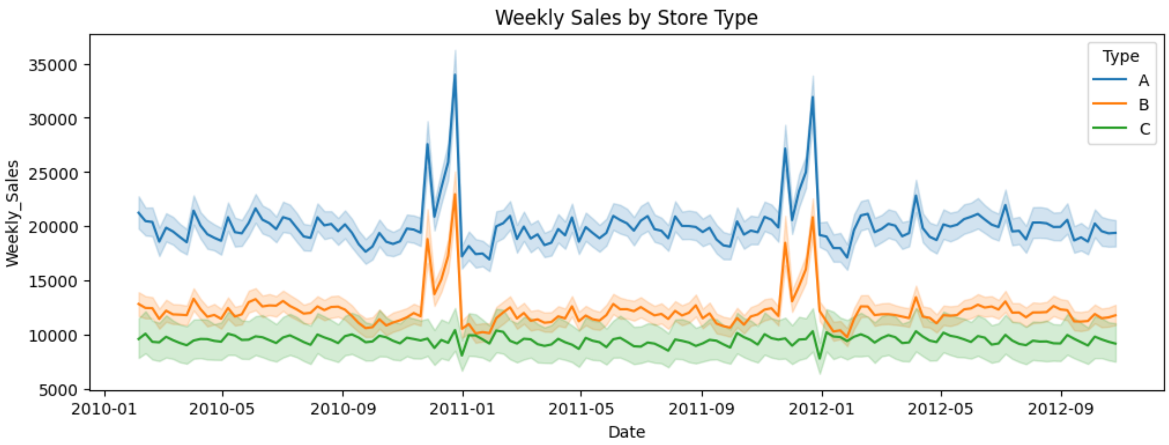

Figure 2

By Investigating weekly sales patterns, I observed that holidays, specifically Thanksgiving and Christmas, significantly impact sales (Figure 3). Some departments show spikes in sales during specific weeks, such as Super Bowl weeks.

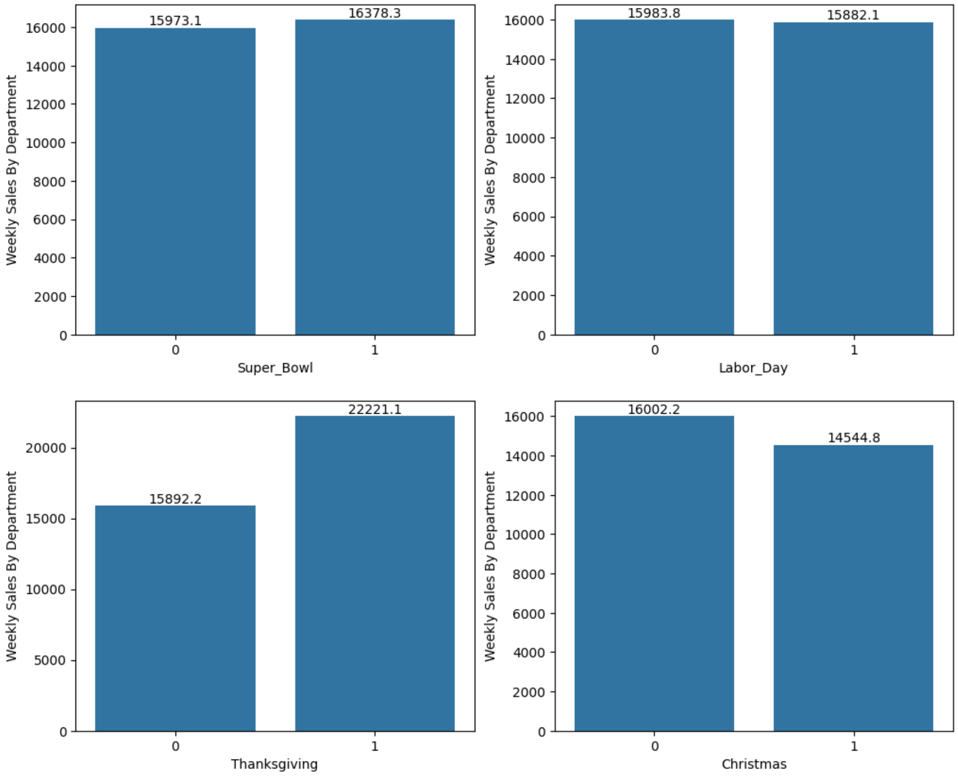

Figure 3 Holiday impact on weekly sales

I also examined the impact of promotions on sales is examined. Negative minimum values for weekly sales and markdowns are identified as potential typos and corrected to positive values. While promotions generally boost sales, it is noted that promotion data for all stores starts after November 11, 2011. Despite concerns about missing data, these features are retained for further analysis. (Figure 4)

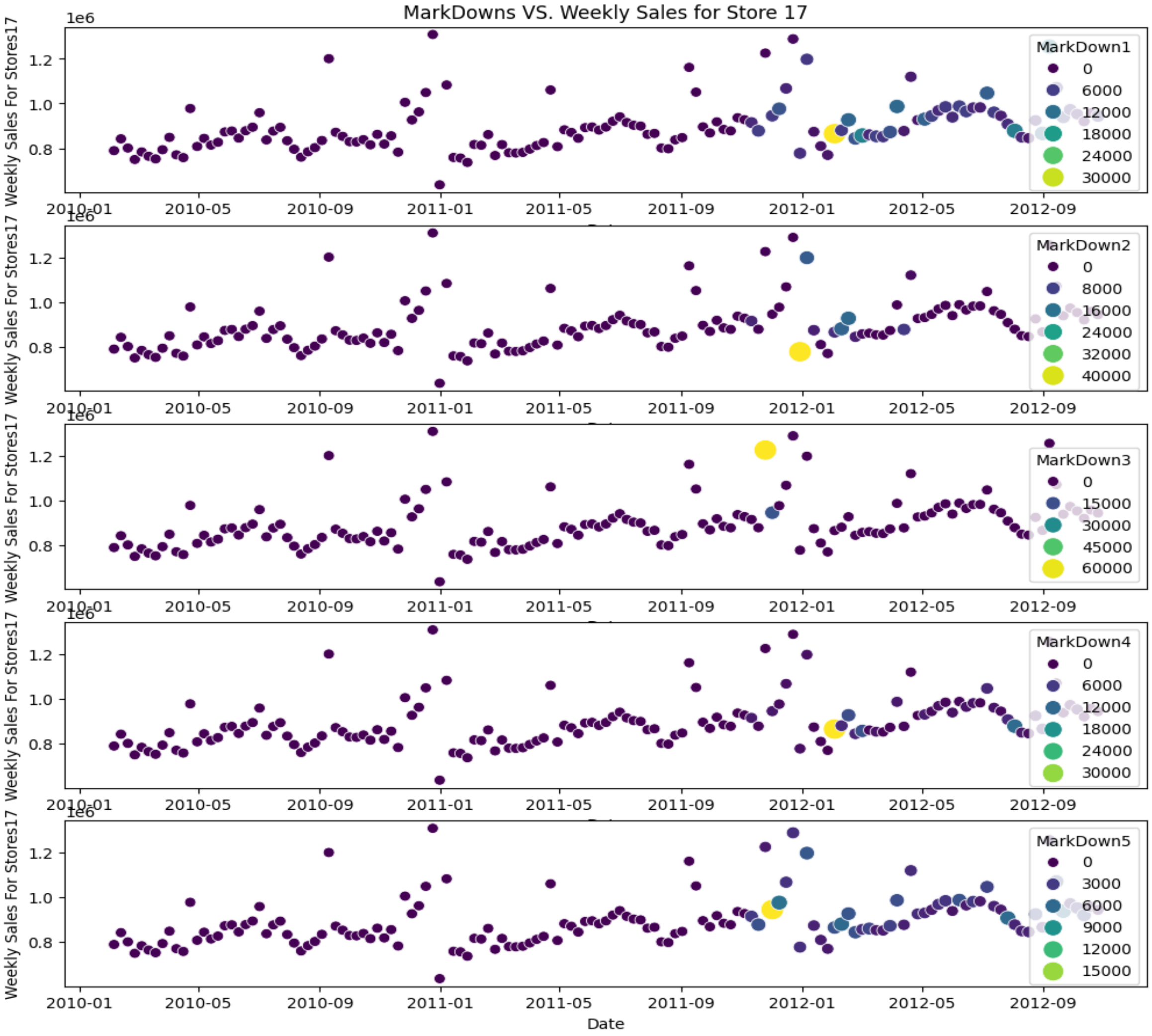

Figure 4 Markdowns Vs Weekly sales for Store 17

Further exploration involves analyzing weekly sales over time, revealing clear seasonal patterns. Temperature and CPI showing less correlation with sales. Fuel prices, despite rising, do not exhibit a clear correlation with weekly sales. Unemployment rates also do not appear to influence sales.

I resample the data into monthly, quarterly, and yearly intervals, which provides insights into different temporal resolutions. The weekly sales exhibit a notable surge starting from October, reaching their peak in December. Quarter 4 consistently shows higher weekly sales for 2010 and 2011, aligning with the holiday season. However, a decline in 2012 could be attributed to the lack of data after 10/26/2012, suggesting that people might delay shopping before Thanksgiving and Christmas, anticipating promotions. (Figure 5)

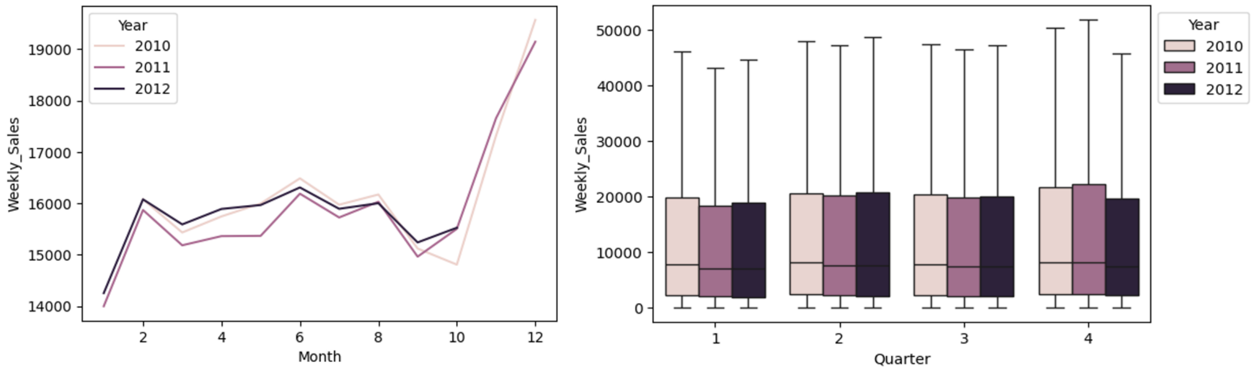

Figure 5 Average Weekly sales by Month and Quarter

In summary, the analysis uncovers store classifications, seasonal sales patterns, and the impact of holidays on sales. Challenges such as potential misclassifications and missing data are noted, providing a foundation for further investigation and modeling.

Pre-Processing & Modeling

In this project, I experimented with five different models.

Initially, I utilized the Linear Regression model, Random Forest Regressor and XGBoost Regressor for the machine learning aspect. The linear model served as a benchmark, providing a baseline with modest predictive capabilities. In contrast, the other two ensemble models exhibited promising results, significantly enhancing predictive power, particularly the Random Forest Regressor, which achieved an exceptional R2 score of 0.996 on the training set. I used all 22 features for all these three models.

Following this, I explored a traditional statistical time series approach. Based on the exploratory data analysis (EDA) findings, which revealed seasonality in the department's weekly sales and consistent patterns across different stores for the same department, I constructed a SARIMA model using the average department sales across stores. Although this model yielded a lower RMSE of 896, it had a diminished R2 score of 0.6. Before modeling I applied Dickey-Fuller Test and plot (Figure 6) the transformed data to confirm its’ stationarity. Also, I did a white noise to verify the data is not just white noise after removing seasonality and trend.

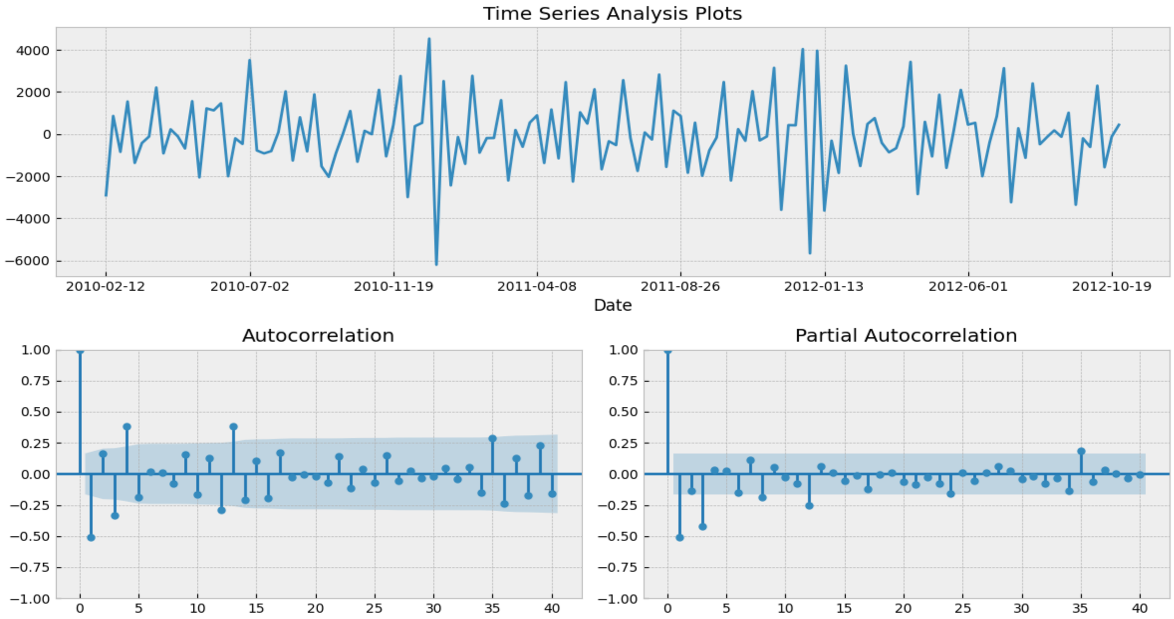

Figure 6

Seeking further improvements, I experimented with a deep learning model called LMST, a type of Recurrent Neural Network (RNN) tailored for handling short/long-term time series data. Unfortunately, the results from applying this model to the department's data were disappointing.

Model Selecting

For model selection in this project, three metrics were employed: R2 score, RMSE, and MAE. Among all the models, the Random Forest Regressor demonstrated the highest R2 score. However, the LSMT model produced a negative R2 score, rendering it unsuitable for consideration as the final model. Although the SARIMA model exhibited a lower RMSE and MAE compared to all other models, its R2 score hovered around 0.6 (Table 1).

Performance Metrics				
Metrics	**Linear Regression**	**Random forest Regressor**	**XGBoost Regressor**	**SARIMA**
Train RMSE:	21737.601	1517.267	5316.638	
Test RMSE:	21851.049	4033.612	5594.730	896.360
Train MAE:	14582.799	578.454	3026.045	
Test MAE:	14589.474	1575.786	3093.823	677.566
Train R2:	0.081	0.996	0.945	
Test R2:	0.084	0.969	0.940	0.597

Table 1 Performance Metrics Table

In addition to statistical metrics, I also considered how well the prediction data aligned with the original data to guide my model selection process. Upon examining the plots for both the Random Forest (Figure 7) and SARIMA (Figure 8) models, it became evident that the Random Forest predictions closely matched the true values. As a result, I concluded that the Random Forest model outperformed the SARIMA model for this project. Although the Random Forest model exhibited relatively higher RMSE and MAE, it's important to note that in this project, our focus is on predicting weekly sales for a single department, which typically involves large numerical values. Therefore, the RMSE remains acceptable in this context.

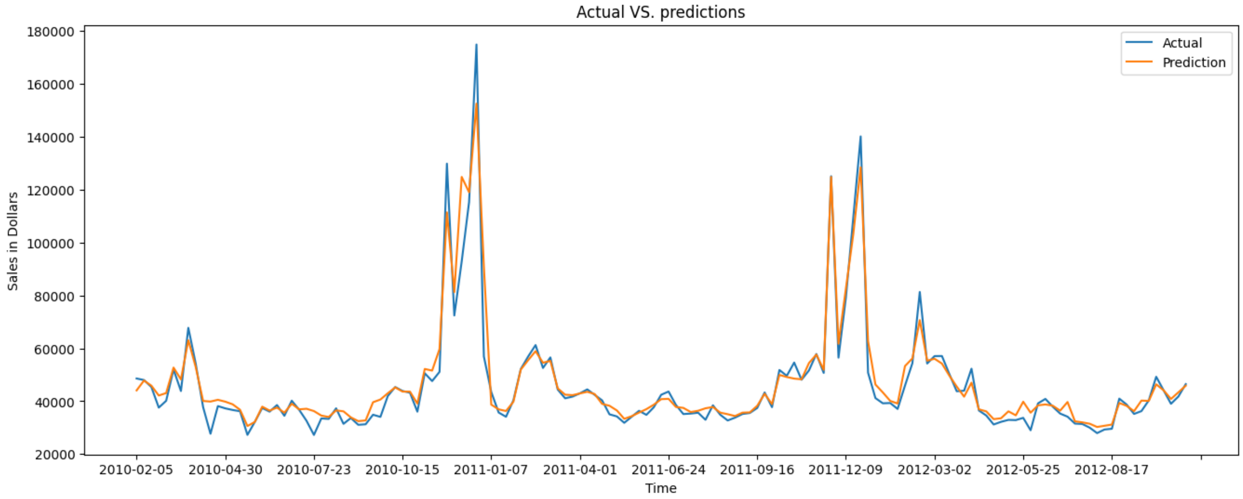

Figure 7 Prediction Plot of the Random Forest Regressor

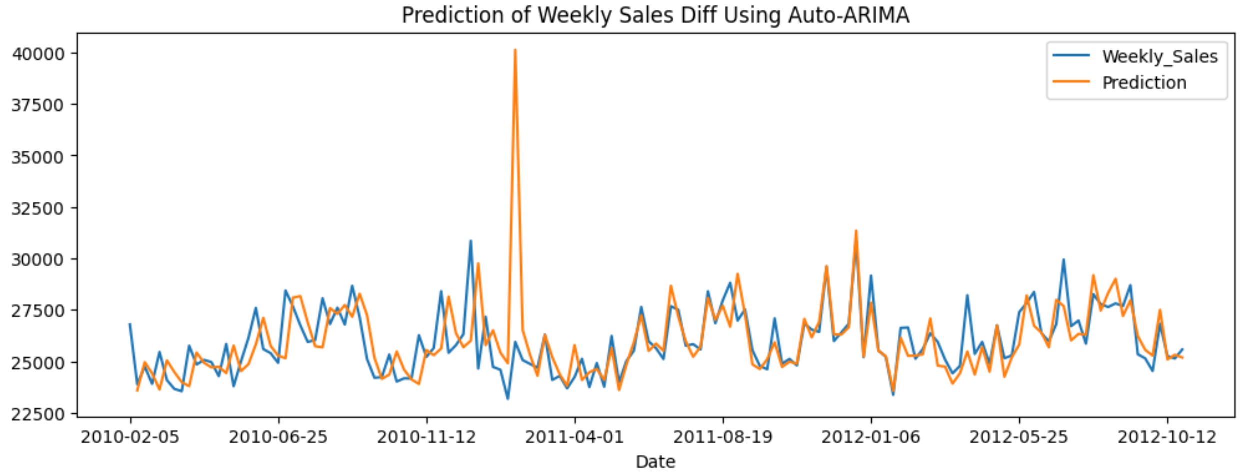

Figure 8 Prediction Plot of the SARIMA Model

Given the Random Forest model's notably high R2 score and its clear alignment with the validation data, it emerged as the preferred choice for achieving accurate sales forecasting in this project.

Takeaways

1. My model has a very high predictive power.
   My final model has a high test R2 score of 0.969, and the prediction plot aligned with the true data perfectly.

2. Department is the crucial predictor for the weekly sales prediction.
   The feature importance plot for the Random Forest Regressor highlights that the department itself contributes significantly to the model, accounting for 62.5% of the overall importance. Subsequently, the size of the store and the specific store also exhibit notable contributions. It's understandable that sales vary across different departments, making departmental differences a crucial factor in predicting our target: weekly sales by department. Additionally, the size of the store is expected to impact departmental weekly sales, as larger stores typically attract more customers, leading to higher overall sales for each department.

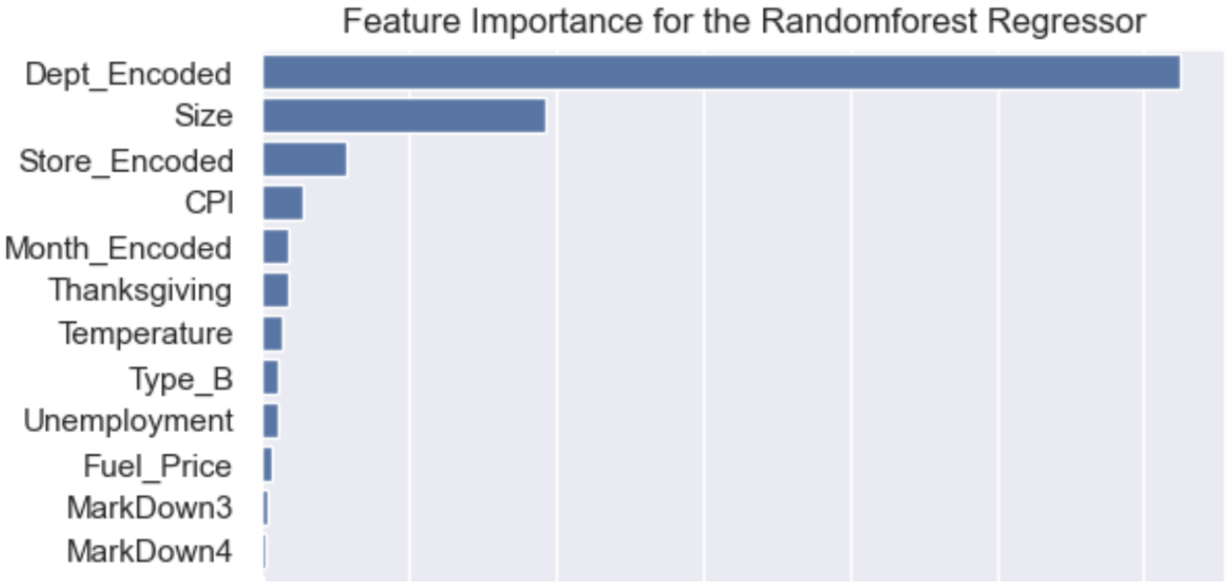

Figure 8 Feature importance (Top 12)

3. Clients can optimize their inventory management and logistics management by utilizing this model
   By leveraging the model developed in this project, the Walmart management team can enhance the accuracy of weekly sales predictions for each department across different stores. This approach surpasses relying solely on business intuition or historical sales data from recent weeks. Consequently, Walmart can optimize its inventory and logistics management processes, leading to reduced operational costs.

Moreover, the predictive capabilities of the model enable the management team to identify optimal timing for promotions, thereby boosting sales and maximizing revenue generation. Overall, the implementation of this model empowers Walmart to make informed decisions, optimize operations, and drive business growth in a competitive retail landscape.

Future Research

I want more data.

Firstly, the store data set for this project only contain the store size information. But I believe incorporating additional information about the store's location and surrounding demographics can greatly enhance the model's predictive capabilities. Factors such as whether the store is situated in a densely populated area, its proximity to affluent or poor communities, the demographic composition of the local population (including age distribution and household composition), can all significantly impact store sales and consequently, departmental sales.

For instance, stores located in densely populated areas might experience higher foot traffic and sales volumes. Similarly, stores in affluent neighborhoods might cater to a different consumer demographic with potentially higher purchasing power and different shopping preferences compared to stores in less affluent areas. Additionally, understanding the age distribution and household composition of the local population can provide insights into consumer behavior and shopping patterns.

Additionally, based on the exploratory data analysis (EDA), the absence of complete promotion markdown data could adversely affect model performance. Therefore, I will include the complete markdown data in the model to improve its predictive capabilities.

Furthermore, the dataset captures weekly sales data from February 5, 2010, to October 26, 2012, encompassing only 143 time steps. Given the observed yearly pattern with a cycle of 52 steps, it's evident that there are only three complete cycles present in the data. Incorporating additional historical data beyond this timeframe may enhance the model's performance by providing a more comprehensive understanding of seasonal trends and patterns.