Predictive Modeling for Weekly Sales Performance in Walmart Stores

Thang Truong, Kimin Nguyen, Naomi Norton, Priyansh Dhandha

Introduction

In the retail industry, accurately predicting weekly sales is crucial for effective inventory management, resource allocation, and overall business strategy. Walmart – one of the world's largest retailers, constantly tries to optimize its sales forecasting methods to ensure efficient operations across its vast network of stores.

In this study, we aim to leverage a comprehensive dataset encompassing various attributes related to Walmart stores to develop a predictive model for weekly sales. To accomplish this, we used a Kaggle dataset including previous sales data from February 5, 2010 to October 26, 2012, across 45 different US locations. We also modified this data for a more efficient analysis. This includes combining "features.csv" and "train.csv" by the Store and Date columns, as well as combining this merged dataset with "store.csv" by the Store column. Additionally, we extracted Day, Week, Month, and Year from the Date column for each observation, utilized one hot encoding for the Type variable to represent it with numerical values, and standardized logical variables as 0 and 1.

It's noted that in the original dataset, there are approximately 320,000 missing values in the markdown columns. This absence is attributed to the unavailability of markdown data prior to November 2011, as well as irregular reporting of markdown data by some stores post-November 2011. Additionally, variations in markdown periods among stores contribute to the inconsistencies. However, these missing data would be crucial for subsequent analysis. Hence, we addressed the missing values by imputing them with 0 and consolidated all markdown columns into a single entity, representing the average of all markdowns.

After these data modifications, the final dataset utilized in this investigation has 421,570 observations for 17 variables. There were 16 attributes used to predict the response variable, weekly sales. This included crucial factors such as store number, store size, department information, day, week, month, year, and store types (Type A, B, and C), as well as external factors like temperature and fuel prices. Additionally, anonymized data on promotional markdowns, consumer price index (CPI), unemployment rate, and holiday weeks are provided.

From this data, this study will aim to investigate the following questions: 1) How do sales performance metrics vary across different store locations? 2) What factors contribute most to weekly sales performance? 3) Can we accurately predict future sales based on historical data and external factors? These inquiries will guide the exploration of patterns in this dataset that can establish an accurate sales prediction model by applying comprehensive analysis and modeling methods. Such a model could lead to more strategic decision-making at the corporate level while increasing operational efficiency within individual stores.

Data Analysis

1. How do sales performance metrics vary across different store locations and departments?

We compared sales performance metrics across different store locations to identify top-performing stores and potential areas for improvement in underperforming stores.

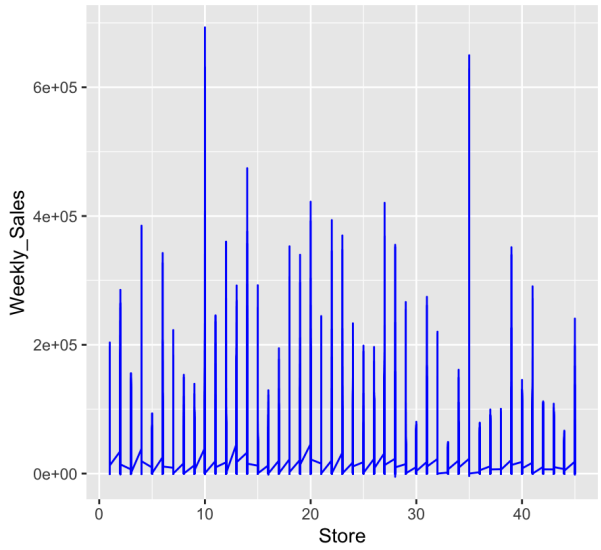


Figure 1: Weekly sales across different stores

From Figure 1, it's apparent that Stores 10 and 45 have the highest Weekly Sales figures. However, we've noticed an uneven distribution among the number of stores. Therefore, our next step was to thoroughly consider other factors, such as time, that could potentially influence sales performance.

Figure 2: Line graph of Weekly Sales for each store over time

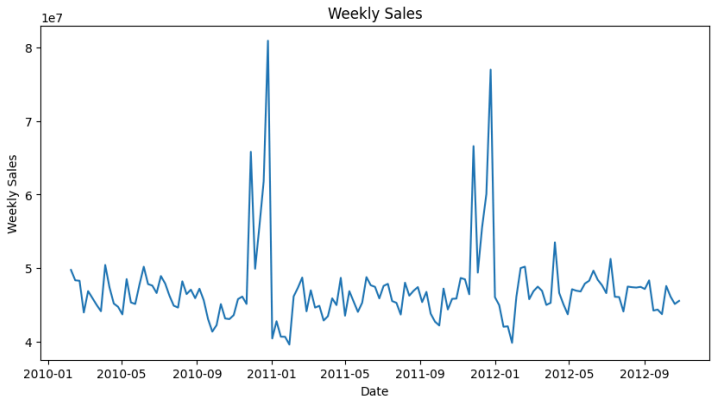


Figure 3: Line graph of the Weekly Sales trend over time for all stores

As seen in Figures 2 and 3, sales vary depending on the time of year. In particular, Figure 2 showcases that nearly all stores have a large decrease in total weekly sales at Week 44 of the year, which follows Halloween in early November. Additionally, there is a large increase in total Weekly Sales at Week 51 of the year, the week of Christmas at the end of December. This is consistent with Figure 3, which captures the spikes in Weekly Sales at the same times in the year for 2011 and 2012.

Figure 4: Bar graph of mean weekly sales by store

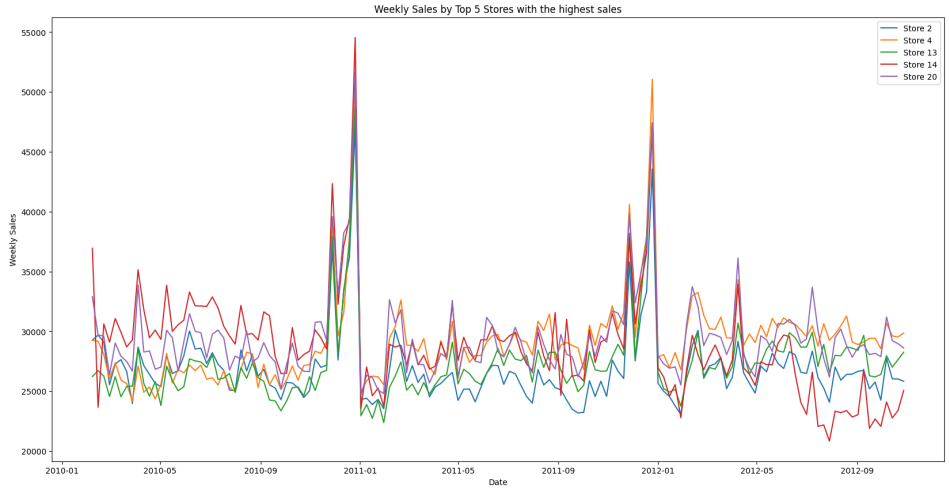


Figure 5: Weekly sales for the top five highest-performing stores

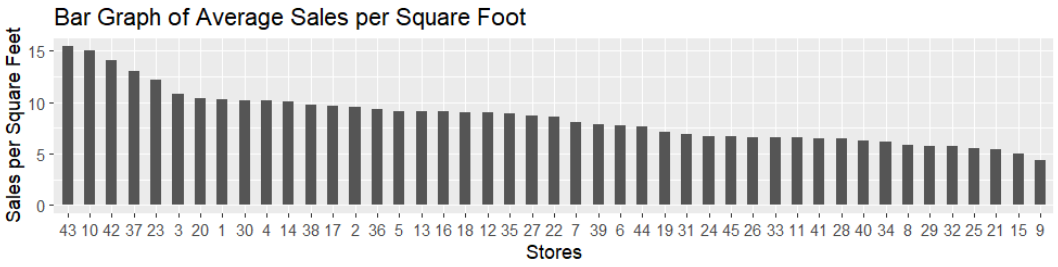


Figure 6: Bar graph of sales per square foot by store

With this aforementioned pattern in mind, Stores 20, 4, and 14 have the highest mean weekly sales (Figure 4). These top stores share a similar trend pattern (Figure 5) to that outlined in Figure 3. However, Stores 43, 10, and 42 have the largest average weekly sales per square foot despite being smaller in size compared to most other locations (Figure 6). Additionally, none of these stores have the highest mean weekly sales, but this indicates that they are the most efficient in generating sales within their location's given space.

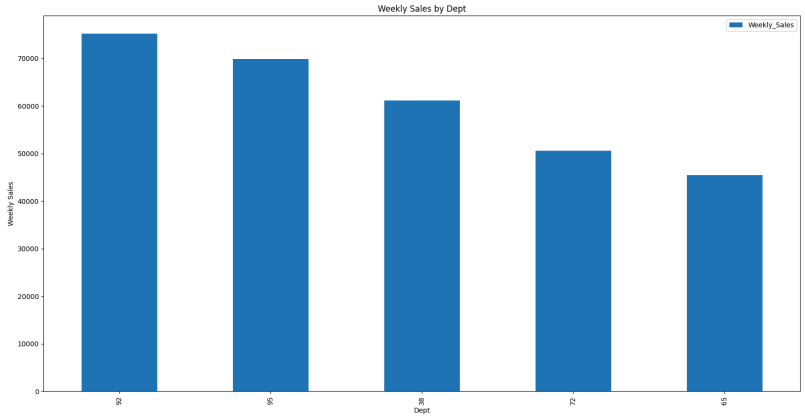


Figure 7: Bar graph of Weekly Sales by Department

Figure 7, meanwhile, shows that Departments 92, 95, 38, 72, and 90 at Walmart have the highest weekly sales, corresponding to Grocery, Directed Store Delivery Grocery, Pharmacy Rx, Electronics, and Dairy, respectively.

2. What factors contribute most to weekly sales performance?

We tried to determine the factors that impact Weekly Sales performance the greatest. We initially used the pairs function to visualize the pattern of each variable. However, the data set was too big, which made the graph too dense to visualize. Instead, we used the correlation matrix to visualize the interaction between the variables. Based on Figure 8b, Size has the most linear relationship to the Weekly_Sales variable, followed by Type and Department.

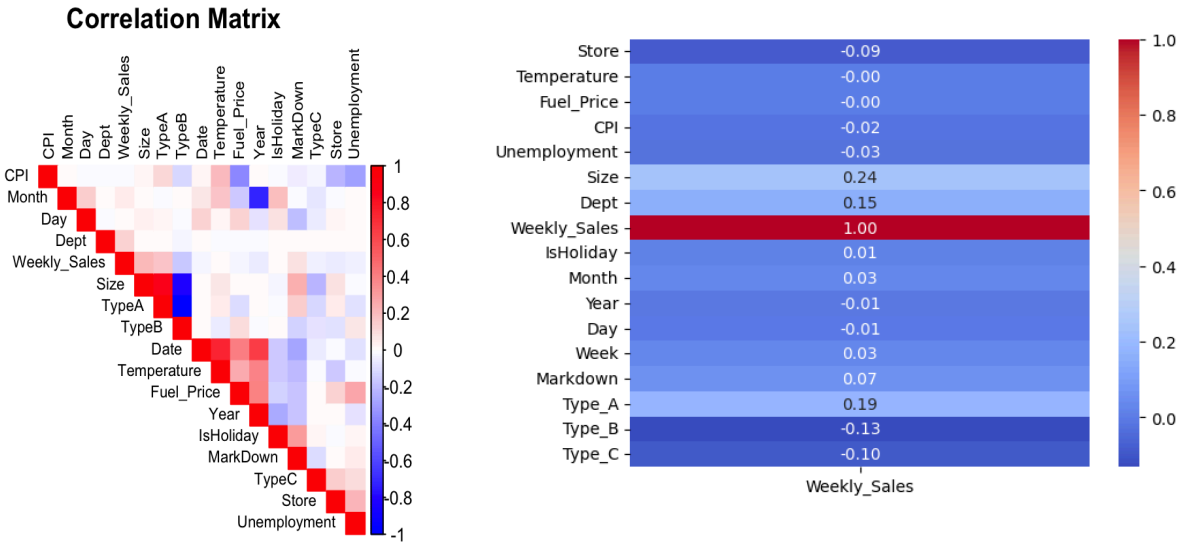


Figure 8a & 8b: Visualization of the correlation between each variable in the data set

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.793e+05	6.464e+04	-12.055	< 2e-16 ***
Date	-1.048e+00	8.961e-02	-11.700	< 2e-16 ***
Temperature	-2.488e-01	3.266e-03	-76.175	< 2e-16 ***
Fuel_Price	1.183e+00	1.677e-01	7.053	1.77e-12 ***
CPI	-2.373e-02	1.059e-03	-22.418	< 2e-16 ***
Unemployment	1.365e+00	2.313e-02	59.000	< 2e-16 ***
TypeB	7.562e-01	1.473e-01	5.134	2.84e-07 ***
TypeC	2.203e+01	4.399e-01	50.083	
Size	4.648e-05	1.469e-06	31.647	
Dept	4.699e-03	1.187e-03	3.959	
Weekly_Sales	-3.336e-05	1.485e-06	-22.462	
Month	3.308e+01	2.733e+00	12.106	
Year	3.953e+02	3.281e+01	12.048	
Day	1.050e+00	8.971e-02	11.701	
IsHolidayTRUE	-3.780e-01	1.320e-01	-2.863	
MarkDown	-9.689e-06	1.970e-06	-4.917	

```
TypeC          < 2e-16 ***
Size           < 2e-16 ***
Dept          7.52e-05 ***
Weekly_Sales   < 2e-16 ***
Month          < 2e-16 ***
Year           < 2e-16 ***
Day            < 2e-16 ***
IsHolidayTRUE  0.00419 **
MarkDown      8.79e-07 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
  ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 121.6444)

    Null deviance: 14064449  on 97055  degrees of freedom
Residual deviance: 11804376  on 97040  degrees of freedom
  (324514 observations deleted due to missingness)
AIC: 741427

Number of Fisher Scoring iterations: 2
```

Figure 9: Linear Regression Result of Predicting Weekly Sales

In addition, we applied linear regression to narrow down which factors were significant for Weekly Sales performance (Figure 9). Unexpectedly, it’s evident that all of the features in the dataset are significant predictors of Weekly_Sales, as seen by each having a p-value of less than alpha = 0.05. Two potential factors can be attributed to this result: multicollinearity and outliers

Multicollinearity, arising from high correlations among predictor variables, poses challenges for linear regression by obscuring the unique contribution of each predictor. With 421,570 variables in the dataset, multicollinearity becomes inevitable, leading to inflated standard errors and reduced power in identifying significant predictors. Outliers further complicate regression analysis by disproportionately influencing coefficient estimation, potentially yielding unreliable results. Linear regression's sensitivity to outliers can obscure the significance of predictors, affecting the overall model's accuracy and reliability. Beyond statistical considerations, the importance of each feature may vary depending on the specific context and dataset. Factors such as store location, demographics, promotions, weather, holidays, and economic indicators like fuel prices and unemployment rates contribute to sales dynamics. Store-specific attributes, including size, marketing strategies, price markdowns, and store type, further shape sales patterns by influencing product range, customer attraction, and competition dynamics. Understanding and accounting for these diverse factors are crucial for accurate predictions of weekly sales.

While the linear regression model fails to address the significant predictors for the "Weekly_Sales", we applied regression analysis or machine learning algorithms to provide further insight. Specifically, we applied the Random Forest and LASSO models and found that the RMSEs for Random Forest and LASSO regression were 14053 and 21639.27 respectively.

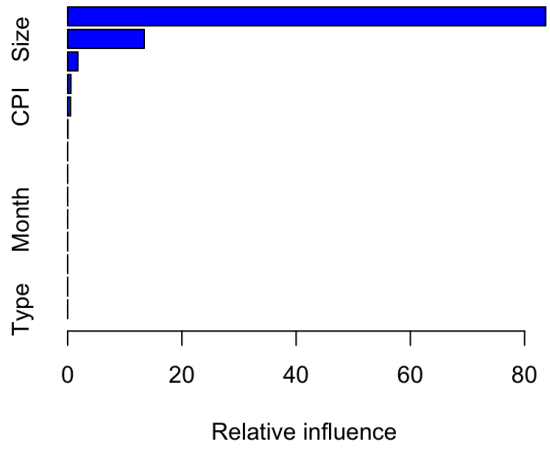
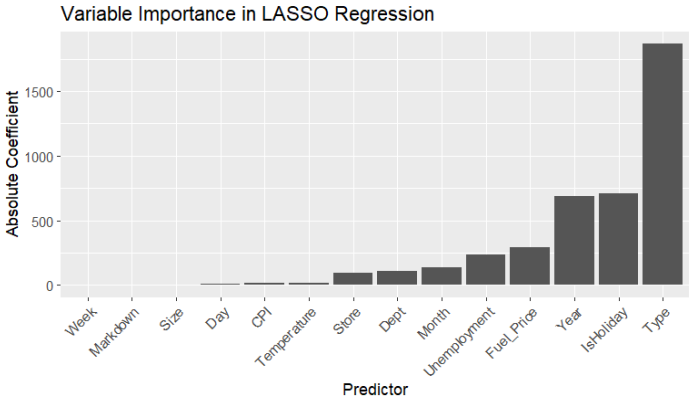


Figure 10: Bar graph of important features to predict Weekly Sales by Random Forest in the training data

Figure 11: Scatterplot of important features to predict Weekly Sales by Random Forest in the whole data set

We applied ExtraTreeRegressor to determine that "Department" is the most important of the features (Figures 10 and 11). A reason for this trend is product differentiation, in which different departments typically offer distinct products or services, catering to varying customer needs and preferences. Certain departments might sell high-demand items or have a broader range of products, leading to higher sales volumes. Another explanation for this trend may be customer segmentation. Customers often visit specific departments based on their interests, demographics, or shopping habits. Departments serving popular or essential goods might attract more foot traffic, resulting in increased sales compared to less frequented departments. Lastly, seasonal variation may result in some departments experiencing fluctuations in sales based on certain trends or events.

We then realized that the LASSO method can also be used to determine the most significant variables. This is because LASSO shrinks the parameter weight of non-important features, thereby automatically performing feature selection and allowing us to determine which variables are the most relevant.



	s0
(Intercept)	1.381941e+06
Store	-9.304388e+01
Temperature	2.030100e+01
Fuel_Price	2.940055e+02
CPI	-1.951283e+01
Unemployment	-2.341650e+02
Size	1.015284e-01
Dept	1.112370e+02
IsHoliday	7.091778e+02
Month	1.385339e+02
Year	-6.880746e+02
Day	-1.220550e+01
Week	.
Markdown	4.078547e-02
Type	1.867881e+03

Figure 12: Histogram of important variables by Lasso Regression in Training Data

To perform the LASSO, we ran a 5-fold cross-validation and then used a lambda value of $\lambda = 6.81$ to calculate the RMSE as mentioned previously, which was 21639.27. We expected the same trend as the Random Forest method by running the LASSO Model. However, it's evident from the LASSO regression that the Type predictor was the most influential (Figure 12). This is inconsistent with the Random Forest model results, which suggest that "Department" is the most important (Figure 11). However, it's noted that LASSO assumes a linear relationship between the predictor and response variables when it may be non-linear. Additionally, the Linear Regression model previously indicated that all predictor variables are statistically significant, which means more complex non-linear relationships may have been overlooked. The Random Forest model better captures these non-linear relationships, as this model does not hold the same linear relationship assumption as the LASSO model. This may also explain why the LASSO model had a much larger RMSE than the Random Forest model. Since the Random Forest can capture non-linear variables and has a lower RMSE, we decided to follow the result of the Random Forest which Department is the most important feature to predict Linear Regression.

3. Can we accurately predict future sales based on historical data and external factors?

We developed predictive models using machine learning algorithms to forecast future sales based on historical sales data, promotional activities, and external factors like weather and economic indicators.

First, we implemented Autoregressive Integrated Moving Averages (ARIMA) to make predictions on the weekly sales purely based on the date and the weekly sales column. It utilizes the lag values and lag prediction errors to make future predictions of the sales. Since the problem is based on time-series forecasting, we chose the ARIMA model, as it is one of the most popular

statistical analysis models that utilizes past values to predict future values. It'a noted that in ARIMA, there are three important values that we need to adjust and specify to improve the interpretability and performance of the model. The parameters include *p* (the number of lags), *d* (the degree of differencing), and *q* (the number of lag errors included).

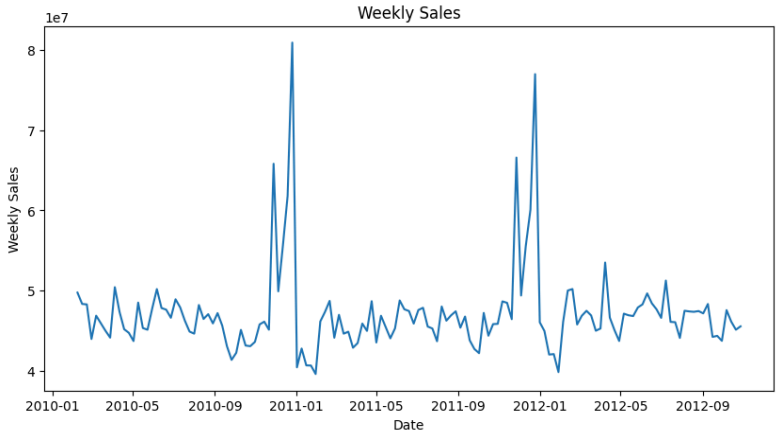


Figure 13: Weekly Sales Across All Stores for each week

Moreover, to enhance the model's interpretation, we grouped the Weekly Sales on the mean values in the week across all stores and departments to make the graph show only one value at a specific timestamp (Figure 13). We also split the dataset into training and testing with 80% of the dataset as the training set and 20% as testing. Therefore, the model will try to predict from April 2012 to December 2012. As such, experimenting with p, d, and q, we arrive at several models that perform well with the training set.

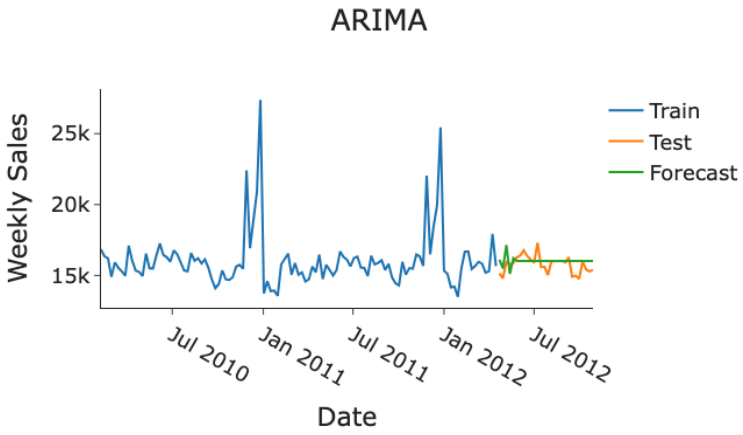


Figure 14: Predicting weekly sales across all stores with ARIMA (0,0,5)

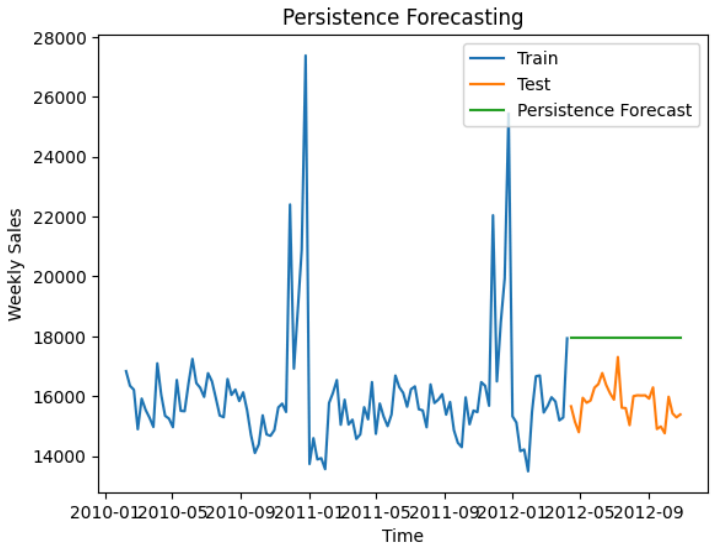


Figure 15: Baseline of persistent forecast across all stores

Using 0, 0, and 5 as p, d, and q, we have the graph as shown in Figure 14. It has an RMSE of 670.95, an R^2 score of -0.27, and an AIC of 2046.9. The value of RMSE is promising because it is less than the persistent forecast's (Figure 15) RMSE of 2232. However, the R^2 indicates that it performs poorly and is worse than predicting the mean. The AIC value can be used for model selection purposes. In terms of interpretability, the model does not produce meaningful results because it outputs a straight line from May 27, 2012 to December 2012.

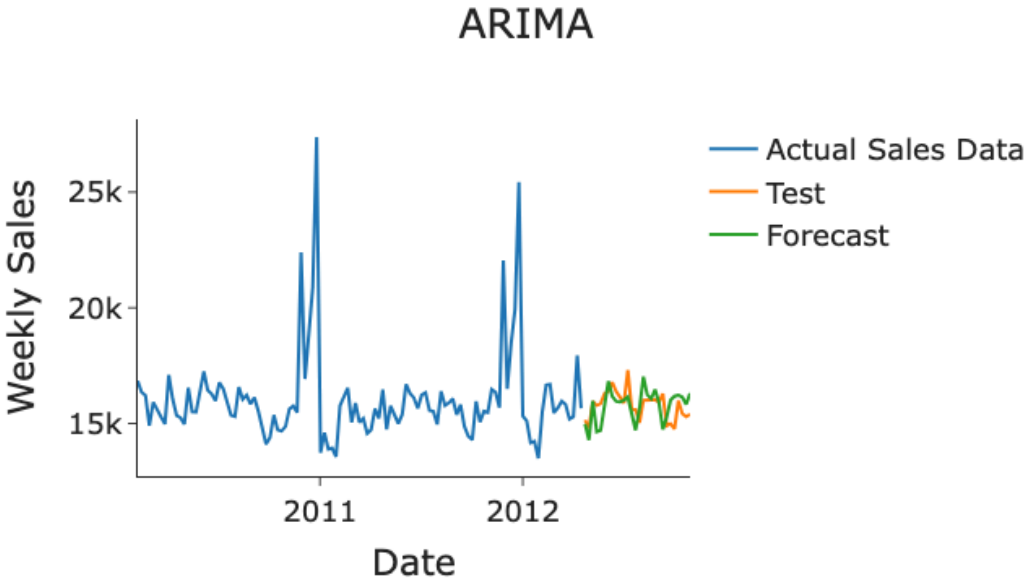


Figure 16: Predicting weekly sales across all stores with ARIMA (15,1,15)

Applying 15, 1, and 15 for p, d, and q, we have Figure 16 with an RMSE of 733.70, an R^2 score of -0.52, and an AIC of 2085.92. All of the evaluation metrics are worse than the 0, 0, 5 combination; however, the graph shows that the model actually learns to detect the seasonality trend and does not solely output a linear line.

Experimenting more with p, d, and q, we arrive at a good combination of these values, as seen in Table 1.

Table 1: Table of various p, d, and q combinations with their evaluation metrics

	(0, 0, 5)	(15, 1, 15)	(10, 1, 10)	(20, 1, 20)	(10, 2, 10)
RMSE	670.95	733.70	**636.23**	1055.13	1307.70
R^2 score	-0.27	-0.52	**-0.14**	-2.136	-3.81
AIC	**2046.9**	2085.92	2074.57	2110.15	2092.45
Detect trends	NO	YES	YES	NO	NO

In addition to ARIMA, we applied a Long Short-Term Memory (LSTM) network. Similarly to ARIMA, the motivation for employing the LSTM network is based on its proficiency with time series data. Given the detection of seasonal trends in our weekly sales data, we chose to utilize an LSTM model to enhance our predictive capabilities. To ensure consistency, the data preprocessing steps were the same as that of ARIMA, the weekly sales were averaged for a given week, across all stores and departments. Given that there were only 143 distinct average weekly sales values, we felt it was prudent for our network to only have one layer of 50 units to reduce the risk of overfitting. In the training phase, we utilized four weeks of historical data as input to predict sales for the upcoming week. While fine-tuning the parameters of the model, we noticed that the number of epochs had a significant influence on our results. This is the number of times the network processes the data in the training phase. After trying various iterations, 200 epochs rendered the best results in terms of minimizing the RMSE and maximizing R^2. The results were steady and for a network with 200 epochs, the test RMSE was 560.34 and the R^2 value was 0.115. This RMSE number makes more sense when putting the range of average sales under context. The range for average sales is approximately 13,885 and our RMSE is 0.4% of this range. An R^2 value of 0.115 suggests that approximately 11.5% of the variance in the weekly sales data is predictable from the features and patterns the LSTM model has learned. While this value is not very high, indicating that the model has limited predictive accuracy, it still provides some level of explanation beyond what would be expected by chance.

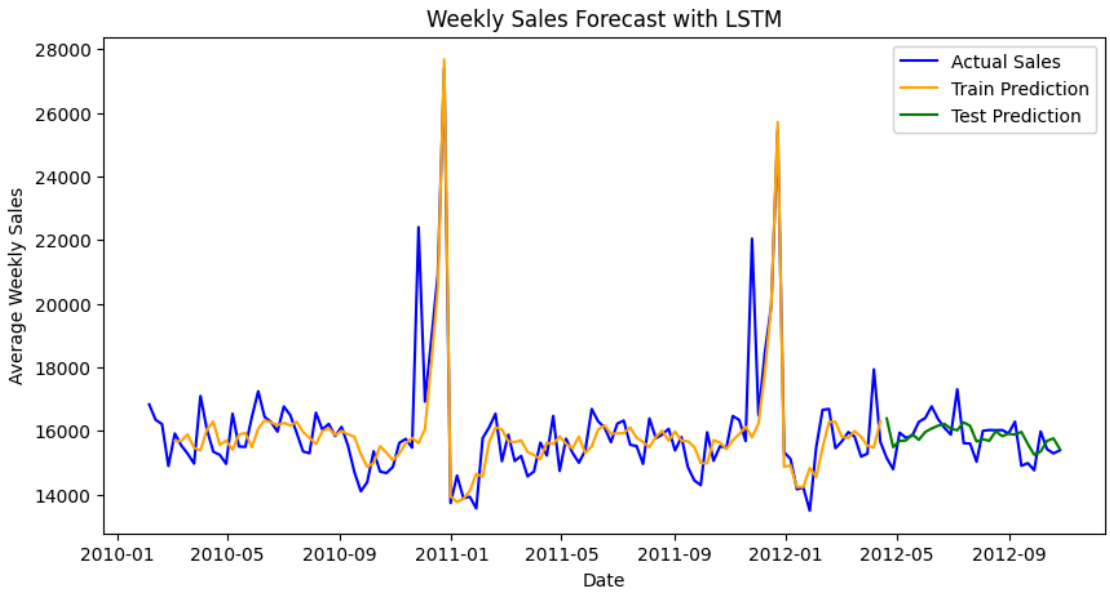


Figure 17: Weekly forecast with LSTM (Single Layer, 200 epochs)

To prove our earlier hypothesis that multiple layers will likely result in overfitting the data, here is what the time series looks like when epochs are held constant at 200 but the number of layers is modified to 2:

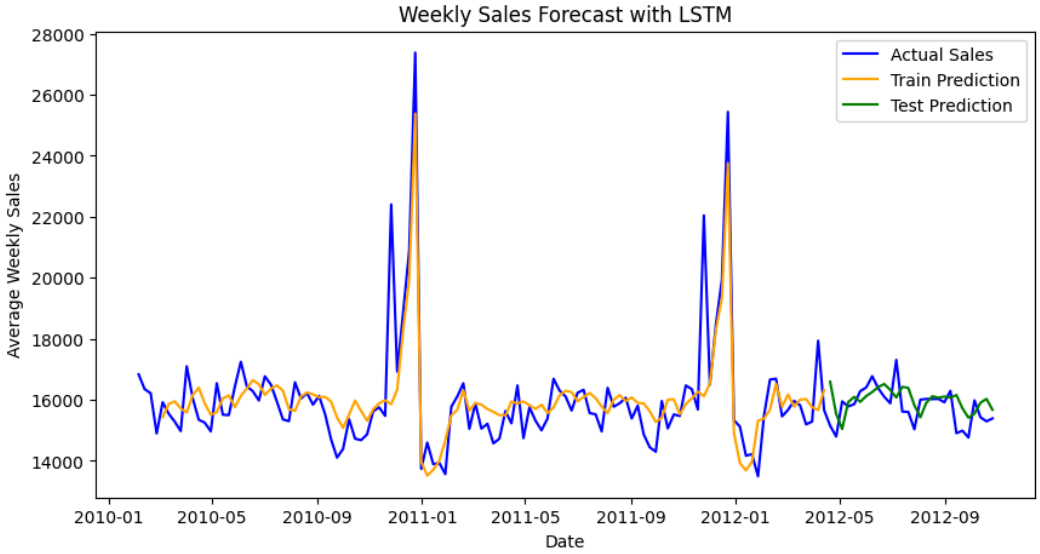


Figure 18: Weekly forecast with LSTM (Double layer, 200 epochs)

Both the resulting test RMSE and test R^2 values for this model were worse than the aforementioned model with just 1 layer, with the former being 648.62 and the latter being -0.185. These numbers suggest that there is overfitting and that a negative R^2 value implies poor model performance. Here are some results of the test metrics with varying parameters.

Table 2: Table displaying the test metrics of different LSTM models

(Layer/s, Epochs)	(1, 50)	(1, 100)	**(1, 200)**	(1, 300)	(2, 200)
RMSE	578.91	573.98	**560.34**	615.32	648.62
R^2 score	0.056	0.072	**0.115**	-0.067	-0.185

Ultimately, these test metrics are important to business decisions, where even the slightest improvements can significantly reduce the costs that Walmart might incur given its massive scale of operations. Lastly, it would be plausible for the company to prefer LSTM over ARIMA, because of its ability to produce a positive R^2 value while minimizing RMSE.

Conclusion

Based on our best-performing model, Random Forest, we found that Department number is the most important factor in predicting sales. In real-world applications, we recommend that Walmart locations should ensure all departments are adequately stocked, especially the most popular ones (such as grocery as seen in Figure 7), to continue encouraging sales.

In the future, if additional work was to be done to predict Walmart sales, we could utilize a dataset that has more consistent markdown data over time. The dataset that we had has many missing values, so our investigation might not have truly captured the extent to which promotional markdowns contribute to weekly sales predictions. Additionally, we could look into the effects of additional regional demographics (eg. average income) on sales, as this is directly related to the consumer base of each store. Moreover, we can apply a transformer neural network, which, as suggested by Professor Ho, may be even more suitable for sequential data predictive analytics than an LSTM model. Lastly, we may also modify the parameters in LSTM to increase R^2 even more while reducing RMSE.

Member Contribution

Kimin	Thang	Priyansh	Naomi
Data Preprocessing & Modification			Attempting SVR and Ridge
Trying Time series to see how accurate the random forest can predict the variables	Build visualization and extract insights for the rise and fall of sales for all store, each store and each department	Trying PCA	Lasso Model, graph, and interpretation
Sketch the correlation/ pattern between each variables an Linear model	Build and evaluate different ARIMA models	Researching and understanding the feasibility of LSTM	Explaining why the Department is significant in tree models
Build Tree and Random Forest models to determine the most significant predictors	Feature Selection for ExtraBoost	LSTM model fitting and parameter optimization	Comparing sales performance metrics and constructing graphs
Writing Project Proposal and Final Project Report			
Presenting and making the slides for the presentation			