Walmart Sales Forecast

Authors:

Min Wei Teoh (Jimmy)	Le Ngoc Canh (James)
Guolin Oo	Ray Tng Rui Kiat

1. Objective/Problem Statement

Retail is a highly competitive and dynamic industry with ever-changing patterns in customer preferences, economic conditions, and market trends. Understanding customer needs and demands is crucial for a retailer such as Walmart. Walmart, like many other retailers, has a primary objective of enhancing its profitability, achieved through a combination of reducing costs and increasing revenue. Accurate demand and sales forecasts not only allow them to achieve goals but affect its logistic operations, financial planning, and its competitiveness as well.

2. Business Justification/Impacts

Accurate demand planning is critical as it helps retailers determine the suitable quantity of products to stock. By predicting customer demand with precision, companies can avoid overstocking, which essentially ties up the capital and potentially leads to discounts or markdowns to clear stocks. Additionally understocking results in missed sale opportunities. The optimization reduces financial risks and costs associated with excess and low-demand inventory. It also helps reduce waste of perishable products. It can be part of sustainability efforts to protect the environment and to reduce waste.

Financially, accurate demand planning also optimizes the supply chain and costs. The supermarket can redistribute financial resources from underperformed products to high demand ones. Therefore, Walmart can maximize its returns on investment and overall profitability. In marketing's implications, supermarkets can have optimal plans for promotions, discounts, and marketing campaigns based on precise insights into customers' demand. It results in a win-win situation with potential increases in sales and improvements on customer satisfaction.

3. Research Questions/Hypotheses

- Is the future sales volume for Walmart stores predictable based on historical data?
    - Hypotheses: The future sales volume for Walmart would be predictable based on seasonality.
- What predictors would influence the volume / sales of a Walmart store?
    - Hypotheses: Holiday indicator and markdowns would influence the volume / sales of a Walmart store.
- Which predictors are more influential than other predictors?
    - Hypotheses: CPI and Unemployment would have less influence than other predictors.

4. Literature Review

Demand forecasting is not new for retailers and there are many studies that are done to predict demand. Demand typically follows a seasonal pattern where during festive seasons like holidays or celebrations (Christmas), the typical demand no longer holds. In such scenarios, machine learning methods such as ANN, RNN and gradient-boosted regression trees may be better compared to traditional time series methods (Huber & Stuckenschmidt, 2020) which separates special periods compared to normal periods. Hence, in our study, we may consider splitting different models for different periods (normal period vs holiday period).

When forecasting for demand, different granularity of forecasting must be employed, whether forecasting is done on product (SKU) level, store level or perhaps on channel level (Fildes et al., 2022). Firdevs et al. also showed that marketing mix and promotions do play a part in driving demand. This is typically modeled via multi-variable regression models.

Time series methods are also used in demand forecasting and can also be done with a hybrid auto-regressive SARIMAX model (Arunraj & Ahrens, 2015). The SARIMAX model is a combination of a ARIMA time series model combined with external variables. By combining external variables (beyond time series), the results are promising compared to traditional time series models.

Marketing promotions will cause complexity to sales volume. Andrade & Cunha (2023) showed that by using Gradient Boosting Machines, the prediction of the model performs better than the standard base-lift model which utilizes simple exponential smoothing model.

5. Methodology and Approaches

The initial step in the exploratory data analysis is data exploration and cleaning. Data is typically required to be cleaned by handling missing or out-of-range values. Additionally, scaling might be required for some models that predictors' value could significantly affect the model performance. Next step is joining all data tables by relevant mutual columns. We noticed that holiday predictors can be further drilled down by encoding the name of the holidays. The periods before holidays are also taken into consideration whether these periods should be treated as holiday period as well.

In the project, we are going to use a mix of typical and advanced models such as time series analysis, gradient boosting, decision tree, linear regression, and random forest regressions. In each model, we will finetune the hyperparameters to find the best performance model of each methodology. Advantages and disadvantages of each method are taken into consideration with the data set and business context.

Each methodology has specific metrics to evaluate the performance with different hyperparameters. Therefore, in order to compare results between different methods, metrics such as mean absolute error (MAE), root mean squared error (RMSE), and symmetric mean absolute percentage error (SMAPE) are used and analyzed. Higher penalties might be given to errors on holiday's period to simulate the business context of missed sale opportunities during peak periods. The trained models will be validated against an unseen test set to get a sense of their accuracies.

6. Dataset

6.1. Dataset description

The data set comprises 4 csv files including features.csv, stores.csv, test.csv, and train.csv. The data source is from Kaggle.

Variable	Description/Unit	Value/Format
Store	The store number	1-45
Dept	Department number	1-99
Date	Friday of the week	YYYY-MM-DD
Weekly_Sales	Weekly sales	Numeric value
IsHoliday	Whether the week is a special holiday	TRUE/FALSE
Type	Store type	A/B/Other
Size	Store size	Numeric value
Temperature	Average temperature in the region	Numeric value
Fuel_Price	Cost of fuel in the region	Numeric value
MarkDown	Promotional markdowns	Numeric value/NA
CPI	Consumer price index	Numeric value/NA

Table 1 Predictors and their descriptions, values in data set

Figure 1,2, and 3 show the screenshots of the dataset.

Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<lgl>
1	2010-02-05	42.31	2.572	NA	NA	NA	NA	NA	211.0964	8.106	FALSE
1	2010-02-12	38.51	2.548	NA	NA	NA	NA	NA	211.2422	8.106	TRUE
1	2010-02-19	39.93	2.514	NA	NA	NA	NA	NA	211.2891	8.106	FALSE

Figure 1 Data table schema of features.csv

Store	Dept	Date	Weekly_Sales	IsHoliday
<int>	<int>	<chr>	<dbl>	<lgl>
1	1	2010-02-05	24924.50	FALSE
1	1	2010-02-12	46039.49	TRUE
1	1	2010-02-19	41595.55	FALSE

Figure 2 Data table schema of train.csv and test.csv

Store	Type	Size
<int>	<chr>	<int>
1	A	151315
2	A	202307

Figure 3 Data table schema of stores.csv

6.2. Data exploration

6.2.1. Exploration on holiday of data sets

Currently, there are 4 major holidays across the years on train set such as Labor Day (2010-09-10 and 2011-09-09), Thanksgiving (2010-11-26 and 2011-11-25), Christmas (2010-12-31 and 2011-12-30), and Super Bowl (2011-02-11 and 2012-02-10). Holiday weeks will be further drilled down into specific holidays. From the train set, Christmas surprisingly has the lowest average weekly sales compared to other holidays and normal weeks. However, the week before Christmas is the peak of sales across the year, store, and department. Based on this observation, the week before Christmas is marked as a holiday period as people tend to shop before the biggest festive period of the year.

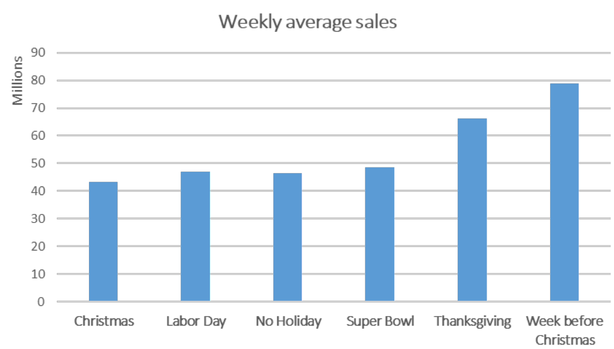


Figure 4 Weekly average sales across different periods of holiday and non-holiday

6.2.2. Negative Values in Weekly_Sales

The lowest Weekly_Sales was identified to be from Store 28 Dept 6 with -\$4988.94 on 2010-10-08 (no holiday). Further analysis reveals that there are a total of 1285 negative values for the Weekly_Sales variable within the train dataset. While the proportion of negative values (0.3%) may be minimal, there is a need to impute them so as to not skew the modeling. A moving average of the month's weekly sales was used to impute these negative values. This could be due to product returns.

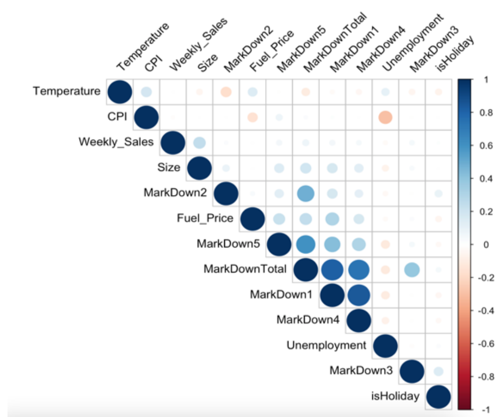


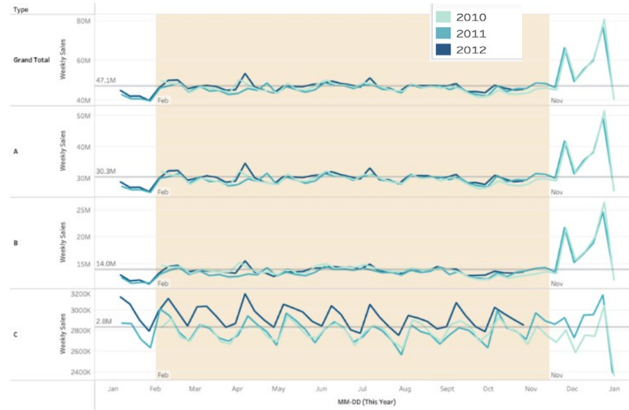
Figure 5 Correlation Plot of Numeric Predictors

6.2.3. Correlation Study on Numeric Predictors

A correlation study was conducted on numeric predictors (Figure 5). It is revealed that the MarkDowns are positively correlated to the MarkDownTotal; this is expected given that the MarkDownTotal is a sum of the respective MarkDowns. It would be advisable to remove either one to avoid multicollinearity. MarkDown2 and MarkDown3 are correlated positively to IsHoliday while CPI and Unemployment are correlated negatively.

6.2.4. Seasonal Patterns of Orders

Sales volume of the stores follow a seasonal pattern where sales remain fixed at around 47 million weekly from Feb to Mid-Nov as seen from Figure 6 (orange band) below across 2010, 2011 and 2012. This is also true for Store Types A and B where across 2010 to 2012, the average weekly orders remain fairly consistent at ~30M for Type A and ~14M for Type B. Type C however sees an uplift in weekly orders in 2012 compared to the weekly order of 3M in previous years.



January sees a lower weekly order for Type A and B stores while Type C stores see similar levels of orders.

6.3. Data cleaning and manipulation

6.3.1. Missing Data

6.3.1.1. Markdown Fields

In the dataset, the Markdown fields do not have any entry prior to 2011-11 as seen from Figure 7 below. Therefore, in our models, we will be separating the models to account for markdown and not include markdown which will only be from 2011-11 onwards.

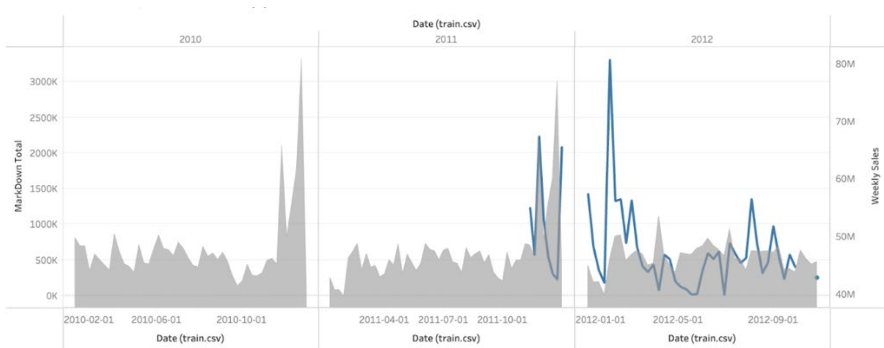


Figure 7 Weekly Markdown Spending by Year

6.3.1.2. Missing CPI feature in 2013

For our prediction / test set that span across 2013, there is no CPI data which is one of the predictors that will be used for prediction. To account for this, a proxy national CPI data will be used instead. However, the national CPI data (Figure 9) has a different range of values compared to the values in the dataset which also varies on store level (Figure 8). Since the range of data is different, the national CPI value will be scaled down to the range of our dataset CPI by store.

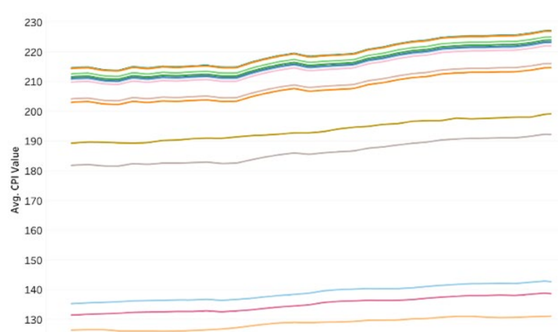


Figure 8 8 Average CPI by Store Type from Training Set

Figure 9 9 National Average CPI values from 2011 to 2013

7. Preliminary Modeling

7.1. Multivariable linear regression

A simple linear regression is done to determine the effects of the various predictors on the weekly sales. The results can be seen below in Figure 10 and Figure 11:

```
Call:
lm(formula = Weekly_Sales ~ ., data = combined_cleaned %>% select(-c(Date,
    Store, Dept, MarkDown1, MarkDown2, MarkDown3, MarkDown4,
    MarkDown5, MarkDownTotal, Type)))

Residuals:
   Min     1Q Median     3Q    Max
-26847 -12843  -6358   4364 676231

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.578e+03  3.989e+02  21.504  < 2e-16 ***
Size          9.074e-02  5.580e-04 162.607  < 2e-16 ***
Temperature   2.915e+01  1.956e+00  14.905  < 2e-16 ***
Fuel_Price   -4.542e+02  7.684e+01  -5.912 3.39e-09 ***
CPI          -1.880e+01  9.554e-01 -19.680  < 2e-16 ***
Unemployment -2.648e+02  1.951e+01 -13.571  < 2e-16 ***
isHoliday     1.400e+03  1.345e+02  10.412  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22010 on 421563 degrees of freedom
Multiple R-squared:  0.06079,	Adjusted R-squared:  0.06078
F-statistic:  4548 on 6 and 421563 DF,  p-value: < 2.2e-16
```

Figure 10 Linear Regression Summary without Markdown

```
Call:
lm(formula = Weekly_Sales ~ ., data = combined_with_markdown %>%
    select(-c(Date, Store, Dept, Type)))

Residuals:
   Min     1Q Median     3Q    Max
-38132 -13035  -6397   4493 629180

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.476e+04  1.031e+03  14.319  < 2e-16 ***
Size          8.874e-02  1.076e-03  82.443  < 2e-16 ***
Temperature   3.546e+01  3.696e+00   9.595  < 2e-16 ***
Fuel_Price   -1.773e+03  2.463e+02  -7.197 6.18e-13 ***
MarkDown1     1.411e-02  1.310e-02   1.077    0.281
MarkDown2     1.512e-02  8.164e-03   1.852    0.064 .
MarkDown3     1.297e-01  7.227e-03  17.945  < 2e-16 ***
MarkDown4     1.406e-02  1.728e-02   0.813    0.416
MarkDown5     7.315e-02  1.035e-02   7.068 1.58e-12 ***
CPI          -2.340e+01  1.681e+00 -13.920  < 2e-16 ***
Unemployment -3.912e+02  3.564e+01 -10.978  < 2e-16 ***
isHoliday    -2.503e+02  2.559e+02  -0.978    0.328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22470 on 151420 degrees of freedom
Multiple R-squared:  0.06464,	Adjusted R-squared:  0.06457
F-statistic: 951.3 on 11 and 151420 DF,  p-value: < 2.2e-16
```

Figure 11 Linear Regression Summary with Markdown Data

Both models yield a low adjusted R^2 value of 0.06. However, in the model without Markdown, all predictors are statistically significant at alpha = 0.001. However, in Markdown model (Figure 11), isHoliday is no longer statistically significant after Markdown fields are added. This could indicate that adding Markdown has introduced multi-collinearity.

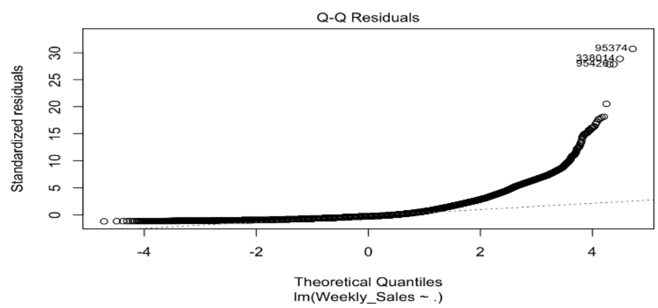


Figure 12 Q-Q Residual Plots of No-Markdown Regression Model

Looking at the Q-Q plot of the fitted regression to the residuals, the regression model seems to generate higher residuals at higher quantiles. This could indicate that the problem may not be linear. Running the model again with Store and Dept predictors yields an R^2 of 0.666. This is because the model now accounts for the baseline for each store and department.

7.1.1. **Elastic Net Regression**

By running an elastic net regression, we can determine if there are any significant predictors that can be removed (LASSO-like) or scale down the importance of parameters (ridge-like). When running the model, the accuracy of the model does not differ significantly compared to a normal regression with R^2 of around 0.66 (with store and dept predictors fitted).

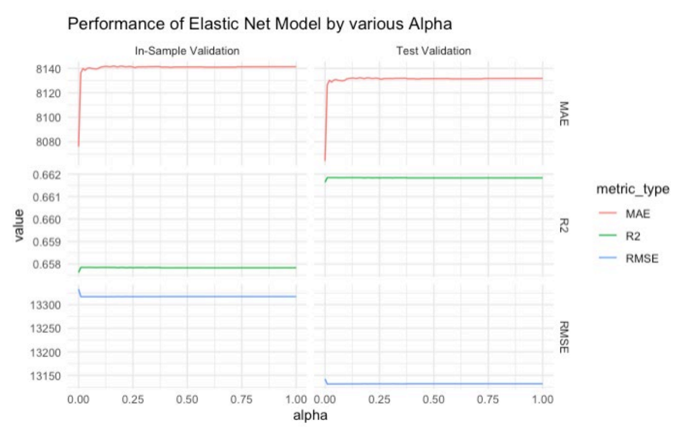


Figure 13 Model Quality across different elastic net weights

The performance does not change much across different weights of LASSO vs Ridge (alpha) as seen from Figure 13 where across all alpha values, the performance seems similar. Neither LASSO nor Ridge would improve the model. Therefore, problem could not be modelled using a Regression model. The regression models could serve as a baseline in other models' accuracies (which should beat an R^2 of 0.66).

7.2. Decision Tree Regression

The Classification and Regression Tree (CART) technique predicts outcomes based on a set of predictor factors (Sharma, Shweta., 2022). Regression trees would be appropriate in this situation given the need for prediction of Weekly Sales data. The cleaned dataset was split into train-test (70-30). A decision tree regression was performed on the train dataset. Size was identified to be the determining factor (Figure 14 left) The tree model was used to predict the values for the test dataset. The tree model classifies the stores into 3 groupings – Large (>171112), Medium (>98689.5 and < 171112) and Small (<98659.5). The model has a MAE of 14531 and RMSE of 22002; the R^2 value is 0.062. Due to the simple estimation using Size alone, it is logical to observe low R^2 value of the tree model and high errors in the predicted values.

To further refine the analysis, the dataset was aggregated on the various Stores with mean of the Weekly Sales factor. A tree regression was performed on the dataset; the model was tested on the test dataset which was aggregated in a similar manner. The resultant model with 9 terminal nodes has a MAE of 4069 and RMSE of 22002 (Figure 14 right); the improved R^2 is 0.0557. It was observed that pruning of tree to lesser terminal nodes resulted in less ideal R^2 values. Figure 15 shows the plot of actual values (black) against predicted values (red).

Figure 14 Tree Regression Plot (with & without aggregate)

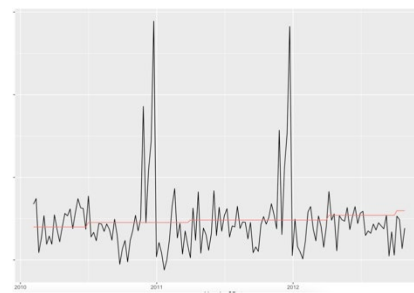


Figure 15 Tree Regression (Actual-Predicted)

7.3. Random Forest

Random forests aggregate the outcome of multiple decision trees where each tree is based on a random subset of the data and a random subset of the features. As the number of trees in the forest increases, the overall error rate decreases. (Breiman, L. ,2001). A random forest analysis was performed on the train dataset at various levels; Overall, Store level, Store-Department (SD) level. The ranger package was used for analysis as the normal random forest package was too slow and couldn't handle predictors with more than 53 variables.

At the overall level, store and department were the main predictors (Figure 16). At the overall and store level, the model does a good job at predicting the Weekly Sales. At the Store-Department level, the model does the best job predicting the Weekly Sales for each Store-Department as can be seen from is having the best metrics (R^2 = 0.898, see Section 8 for further details). However, when we split to such a granular level, not every Store-Department was represented in both the train/test dataset. These were left out in the evaluation which led to a lower-than-expected Weekly Sales when charted at the overall level (Figure 16).

Two different models were used, one with the markdown data and one without. They both performed similarly in the random forest model as can be seen from the metrics (R^2: Overall level = 0.869 v.s. 0.914, Store level = 0.860 v.s. 0.902, Store-Department level = 0.898 v.s. 0.911). In fact the model without markdown data seems to perform slightly better. Since the test set from Kaggle had markdown data, the markdown model was used to predict its values. Interestingly, when using the models to predict the Weekly Sales from the supplied test dataset from Kaggle, we can see that at the various levels, they were unable to predict the double peak at the end of the year. This is possibly due to only 1 year of double peak in the model with markdown data.

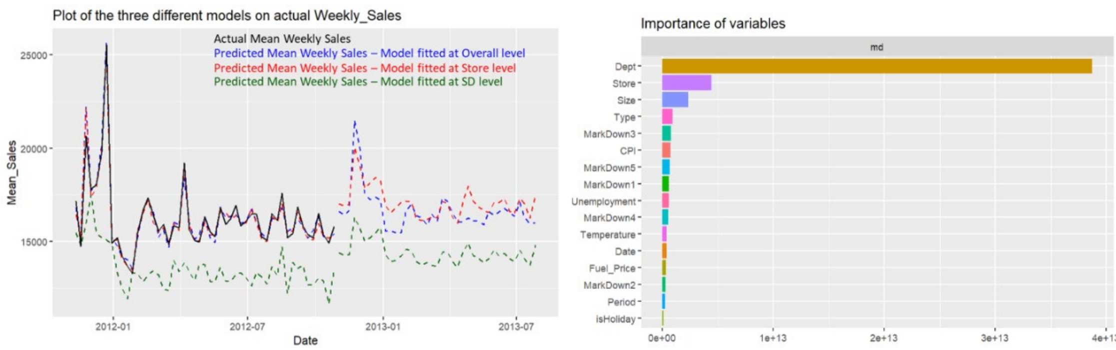


Figure 16 Predicted versus Actual using random forest models at the Overall, Store and Store-Department level and Barchart depicting the importance of predictors

7.4. XGBoost Regression

7.4.1. With versus without Store & Dept variables

The model performs significantly better by including the store and department predictors. This is because each store, department has a baseline order volume. This can be seen in the table above where MAE, RMSE and R-squared are significantly better when including those predictors (without Store & Dept: R^2 = 0.07; with: R^2 = 0.94)

Size is an important feature in the XGBoost model (Figure 17) when fitted without Store and Dept categorical predictors. It far outpaces other predictors like Unemployment and CPI. When fitted with Store and Dept predictors (Figure 18), Dept seems to be a significant predictor as well with Dept 92 having the next largest influence in predictive power for XGBoost.

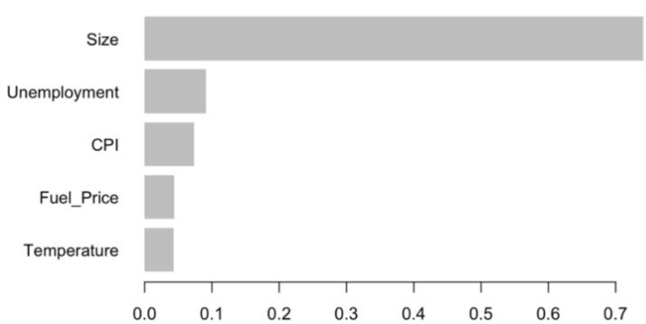


Figure 17 17 Feature Importance of XGBoost (without Store and Dept features fitted)

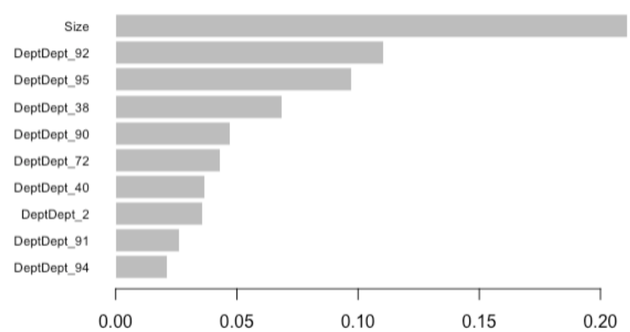


Figure 18 16 Feature Importance of XGBoost (With Store and Dept features fitted)

Both models seem to generalize well when looking holistically as seen from Figure 19. However, inspecting the fit on the department level yields a different story where prediction is very far off from the true value in Figure 20 below. Our hypothesis is that the department did not have further details regarding its characteristics. The model generalizes without looking at department information. An interesting observation is that the model with store and department fitted acts like an average trend on Dept level as seen from the orange line vs black line on Figure 20.

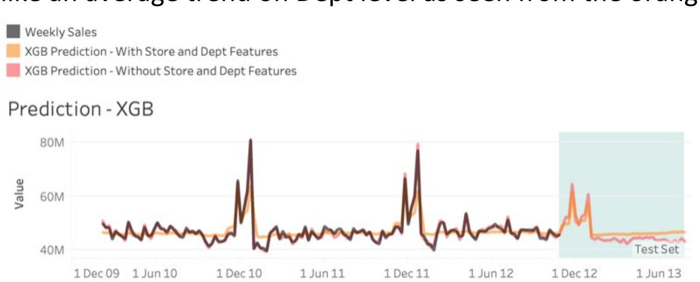


Figure 19 Predicted vs Actual summed at Total level comparing model with store and dept fitted vs without.

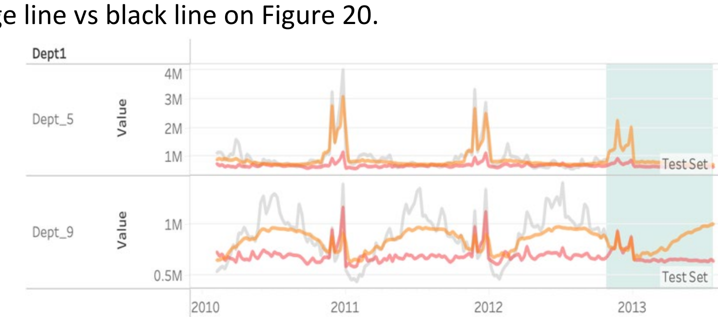


Figure 20 Predicted vs Actual viewed on Dept Level comparing model with store and dept fitted vs without

7.4.2. With versus without Markdown information

The quality metrics do not differ much when Markdown predictors are added into the model (without Markdown: R^2 = 0.943, with Markdown R^2 = 0.947). However, inspecting the trendline shows that the Markdown model performed better during the 2 peaks in Nov – Dec period. The peaks are closer to the actual data compared to the model without Markdown (orange line) in Figure 21. Even during non-peak periods, Markdown model (blue) shows more variance and is closer to the actual sales. This is true even on the Store level as seen from the right chart in Figure 22.

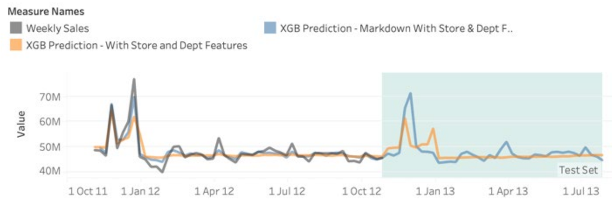


Figure 20 19 Predicted vs Actual on total level with Markdown Fitted

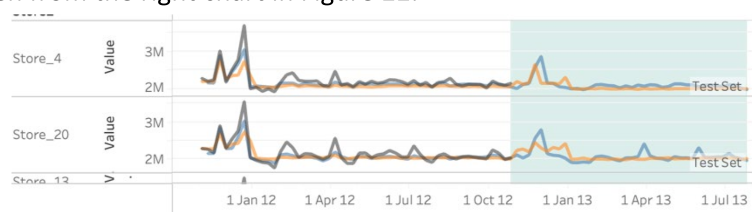


Figure 18 Predicted vs Actual on Store with Markdown Fitted

Generally, Markdown model performs when inspecting trendline although the quality metrics show similar performance.

7.4.3. Prediction on other levels of granularity (Store vs Walmart level)

Another granularity of prediction we can run would be the Store granularity (instead of store-department granularity that we've ran above). When predicting a store granularity, the fit seems on the actual sales data seems pretty good.

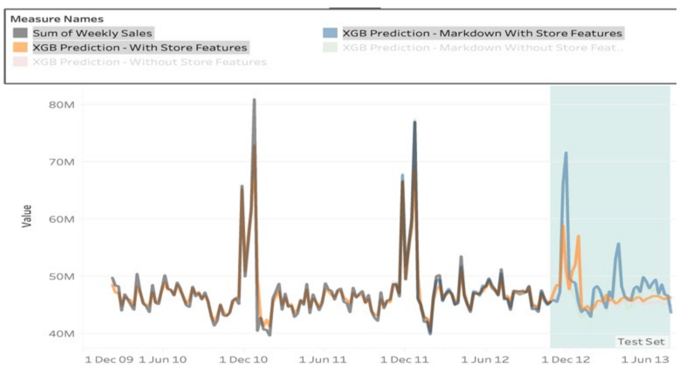


Figure 22 20 XGB Prediction Fitted on Store Level

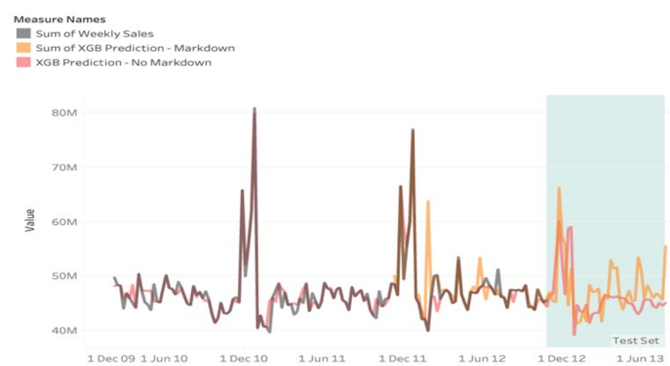


Figure 21 XGB Prediction fitted on Total Aggregated Walmart

For Store level (Figure 22), on unseen data (blue region), the predicted volume from XGB with markdown can predict a spike on Nov 25 but no spike on Dec 23. The markdown model also sees significant variation compared to non-markdown model. On the overall Walmart level (Figure 23), the model deteriorates significantly where predictions for both with Markdown and without. This is because we lose the size of store information which was a significant predictor in the XGB model.

7.5. Time Series Models

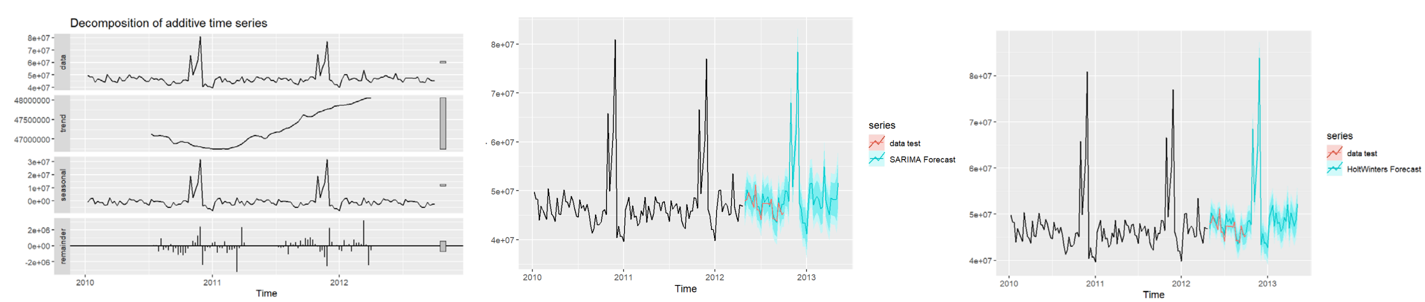


Figure 25 22 Decomposition of additive time series, model performance of SARIMA and HoltWinters model

Recently, classical time series model such as Seasonal AutoRegressive Integrated Moving Average (SARIMA) and Triple Exponential Smoothing are applied for prediction model on sales of retail stores (Yasaman Ensafi.,2022). Figure 25 presents the decomposition of additive time series analysis on the train data set. The data is broken down into trend, seasonal, and remainder factors. Generally, there was increasing trend of sales from 2010 to 2012. From seasonal chart, there was a clear spike in sales during the winter period from October to December with Thanksgiving and Christmas holiday season. Surprisingly, the residual part of sale happened during the period from July to March each year while the sale during the rest of the year were consistent over course of two years.

In time series model with SARIMA algorithm, there are 3 approaches with different granularity in train and test sets: group, store, and department level. In department level model, each department of each store has separate time series model while each store has specific model in store level model. Generally, each model fits well into the data set and makes predictions with relatively good performance. Model on group and store level have significant higher error metrics because the size of data points on these granularity levels are significantly bigger than lower level. The prediction on group level of SARIMA model is presented in figure 25. The model can predict the fluctuation of sales and peak season during year end festivals. In addition, Holt Winters model was also utilized to compare with SARIMA model. In figure 25, Holt Winters also performed well on predictions of dataset and capturing well the peak of sales during holiday season. In comparison, SARIMA model performed better in term of error metrics compared to Holt Winters model.

8. Quality of models and Findings

Model	MAE	RMSE	R-squared
Time Series Models			
HoltWinters (alpha = 0.03938526, beta = 0.07670469, gamma = 0)	41,172.2	46,610.2	-0.29
SARIMA Model on Group level	60,011.5	68,523.2	0.64
SARIMA Model on Store level	16,093.6	49,163.3	N/A*
SARIMA Model on Department level	903	3,371	N/A*
Multivariable Linear Regression			
With Markdown & Store + Dept categorical variables	8260	13280	0.666
Without Markdown Features & Store + Dept categorical variables	8139	13132	0.662
Elastic Net with Store + Dept categorical variables	8140	13150	0.662
Regression Trees (Two tree models were created, one for the overall data and one with aggregated mean by Store			
Simple Regression Tree – Overall	14531	22002	0.062
Simple Regression Tree – Store Level	4068	4172	0.557
Random Forest (Two different models were created at each level, one with markdown features and one without)			
With (Without) Markdown features – Overall level	4402 (3253)	8352 (6651)	0.869 (0.914)
With (Without) Markdown features – Aggregate by store	4328 (3386)	8629 (7115)	0.860 (0.902)
With (Without) Markdown features – Aggregate by store and department	1933 (1681)	5917 (5441)	0.898 (0.911)
XGBoost Regression (Two different models were created at each level, one with markdown features and one without)			
With (Without) Markdown Features & Store + Dept categorical variables	3109 (3013)	5444 (5519)	0.947 (0.943)
With (Without) Markdown Feature & Without Store + Dept categorical variables	14644 (14381)	22141 (21739)	0.0754 (0.0765)

**The performance is calculated based on many submodels ran on individual stores / store-dept. Therefore, R-squared is not available for these models.*

Fitting the store and department as categorical variables would result in a better-quality model as seen from the table above. For instance, model with store and department has an R^2 of 0.94 while without is 0.07. This could be because the model would fit a baseline weekly order volume per store and department.

Adding the Markdown features do not improve the quality of the models much as seen from the table above. For instance, random forest with Markdown at Store-Department level sees R^2 of 0.898 while without is 0.911. The accuracy of both do not deviate too much from one another. For the XGBoost model, the quality of Markdown model also do not deviate much from without ($Without: R^2 = 0.943, With: R^2 = 0.947$).

Time series model such as SARIMA on Store-Department performs better than the factor-based counterparts with RMSE of 3371 versus 5000-6000 from XGBoost and Random Forest on store-department level.

Hyperparameter tuning varies across different models. For instance, in the elastic-net model, exploring across various weights of LASSO – Ridge yields similar accuracy. As for XGBoost, a tree-based model 'gbtree' or linear regression 'gblinear' was attempted and the tree-based model yielded better accuracy. The original depth and iterations of 15 and 25 respectively were used because accuracy yielded was satisfactory.

9. Limitation, Future works, and Recommendations

Time series models would perform better if the data were collected properly for all stores and departments. Moreover, time series models would be more reliable if the data were collected for more than 3-5 years. The limitation of test set without real sale data made the performance evaluation harder. Furthermore, after splitting the data into those with markdown data and those without markdown data, the markdown data effectively only consisted of about a year's worth of data which limits the ability of the model to predict a double peak in the Nov-Dec period.

The tree regression model had limitations to the type of variables used – there was a maximum of 32 factors allowed within the function. As "Store" and "Department" have more than the maximum number of factors after one hot encoding, these variables must be excluded from the modelling.

For future works, we could do a combination of auto-regressive time series model and factor-based model which would improve the accuracy of forecasting. Furthermore, we should continuously retrain the model via a feedback loop to ensure that the model remains relevant. Additionally, hybrid models with combination of traditional statistical methods and machine learning, deep learning approaches would improve the forecasting accuracy.

In the future, should data be collected with more information such as product sales, quantity, and cost, the model would eventually be able to forecast the demand and sales for each product. These advantages are essential to maintain a well-stocked warehouse and logistics management.

10. Conclusions & Research Questions

Going back to the original research questions which are addressed below:

a. Is the future sales volume for Walmart stores predictable based on historical data?
b. What predictors would influence the volume / sales of a Walmart store?
c. Which predictors are more influential than other predictors?

In general, sales of Walmart stores are predictable based on limited historical data. Prediction seems fairly reliable for both factor-based models like Random Forest and XGBoost as well as time-series model like SARIMA. The SARIMA model actually performs better than factor-based model and should be used as for forecasting.

A drawback of time-series model like SARIMA is that there is no way to explain the data other than seasonality. However, from the factor-based models, Size of a store and the department of the stores play the most significant factor in driving the sales of a Walmart store.

Macroeconomic factors like CPI, Unemployment, Temperature and Fuel Price do not influence the forecast as much as store size and department. This could indicate that Walmart is resilient to economic shocks like increased unemployment and high inflation as they are a necessity to Americans.

This project was fulfilling for all four of us as an introductory module for OMSA as we got to apply the techniques we learned from this course. In addition to that, we were challenged to explore beyond the syllabus by doing research on possible models and techniques that industry veterans employ for forecasting retail chains.

11. References

Andrade, L. A. C., & Cunha, C. B. (2023). Disaggregated retail forecasting: A gradient boosting approach. *Applied Soft Computing*, *141*, 110283.

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321-335.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. International Journal of Forecasting, 38(4), 1283-1318.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420-1438.

Sharma, Shweta. (2022). Classification and Regression Trees: The use and significance of Trees in analytics.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – a comparative analysis. *International Journal of Information Management Data Insights,* 2(1), 100058. https://doi.org/10.1016/j.jjimei.2022.100058