Forecasting Retail Store Sales and Determining Key Factors for Increasing Sales Volumes

Thomas Jung

Ryerson University

Table of Contents

Introduction

In today's data-saturated world of retail, the ability to sort through millions of records to identify key predictors for sales volumes is paramount to economic viability. While retail managers have historically relied on their experience and a series of simple linear regression plots to estimate future sales, modern databases with massive amounts of retail data pose a real challenge in generating these same estimates. Unlike sales forecasting in the past, which required humans to manually choose the most relevant factors, machine learning models today are able not only to determine the most important factors but also make more accurate predictions using much larger sets of variables (de Mézerac et al., n.d.).

This project will explore whether retail sales volumes are correlated with factors such as the time of the year, holidays, weather, historical sales volumes, as well as economic indicators such as fuel prices and unemployment rates. These factors will then be used to predict sales volumes for specific stores throughout the year. It is hypothesized that events with department-specific effects such as inclement weather and holidays increase sales in certain departments while decreasing them in others, whereas general economic factors such as fuel prices and unemployment affect sales in all departments in a similar fashion. The Walmart Recruiting - Store Sales Forecasting dataset from Kaggle was chosen for this project. Methodology will include exploratory data analysis, feature engineering, as well as the assessment of three machine learning algorithms for their predictive abilities: multiple linear regression, XGBoost, and Facebook Prophet.

Literature Review

Retail sales forecasting, which is the prediction of future sales using past sales data, allows for informed decision-making with regards to the balance between marketing and sales efforts and supply chain planning. Hill and Orrebrant have found that inaccuracies in forecasting are not necessarily due to the technique itself but may be a result of an unorganized forecasting process and inefficient data flows (2014). Improvements in forecasting require not only the flow of internal information, but also the availability of the data on the overall supply chain. Some of the factors affecting sales volumes include the political and economic environment of the retail store (Ziyadat, 2019, pp. 44). Ziyadat has found that the most important economic factor affecting daily sales is the inflation rate, to the extent that stores must change their pricing strategies in response to changes in the inflation rate. Weather is a complex predictor of sales as the magnitude and direction of the effect are dependent on both the location of the store and the items carried by the specific sales department (Badorf & Hoberg, 2020, pp. 1). The weather's impact on daily sales with regards to different store locations may be up to 23.1%, whereas its impact based on the specific department may be up to 40.7%. The availability of weather forecast data was found to improve the forecasting of daily sales up to seven days in advance, albeit with diminishing accuracies further in time (Badorf & Hoberg, 2020, pp. 1).

Dataset

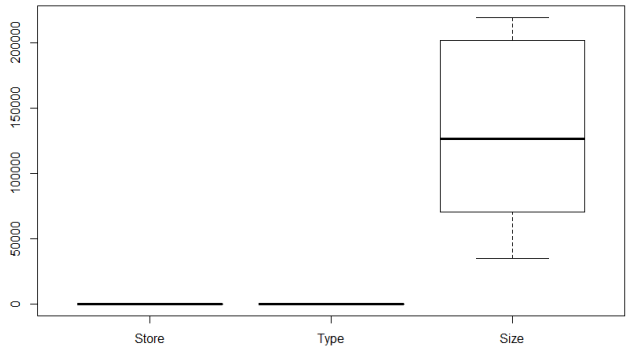
The dataset for this investigation was chosen from a Kaggle Competition entitled “Walmart Recruiting - Store Sales Forecasting”. It contains the following files:

File Name	Description	# of Attributes	# of Records
stores.csv	types and sizes of 45 Walmart stores	3	45
train.csv	weekly sales volumes between 2010-02-05 and 2012-11-01	5	421570
test.csv	test set for predicting sales volumes between 2012-11-02 and 2013-07-26	4	115064
features.csv	additional data related to store, department, and regional activity for the given dates including temperature, fuel price, markdowns, consumer price index, unemployment rate, and whether it is a holiday	12	8190

stores.csv

```
'data.frame':   45 obs. of  3 variables:
 $ Store: Factor w/ 45 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Type : Factor w/ 3 levels "A","B","C": 1 1 2 1 2 1 2 1 2 2 ...
 $ Size : int  151315 202307 37392 205863 34875 202505 70713 155078 125833 126512 ...
```

```
    Store    Type        Size
1       : 1   A:22   Min.    : 34875
2       : 1   B:17   1st Qu.: 70713
3       : 1   C: 6   Median :126512
4       : 1          Mean    :130288
5       : 1          3rd Qu.:202307
6       : 1          Max.    :219622
(Other):39
```

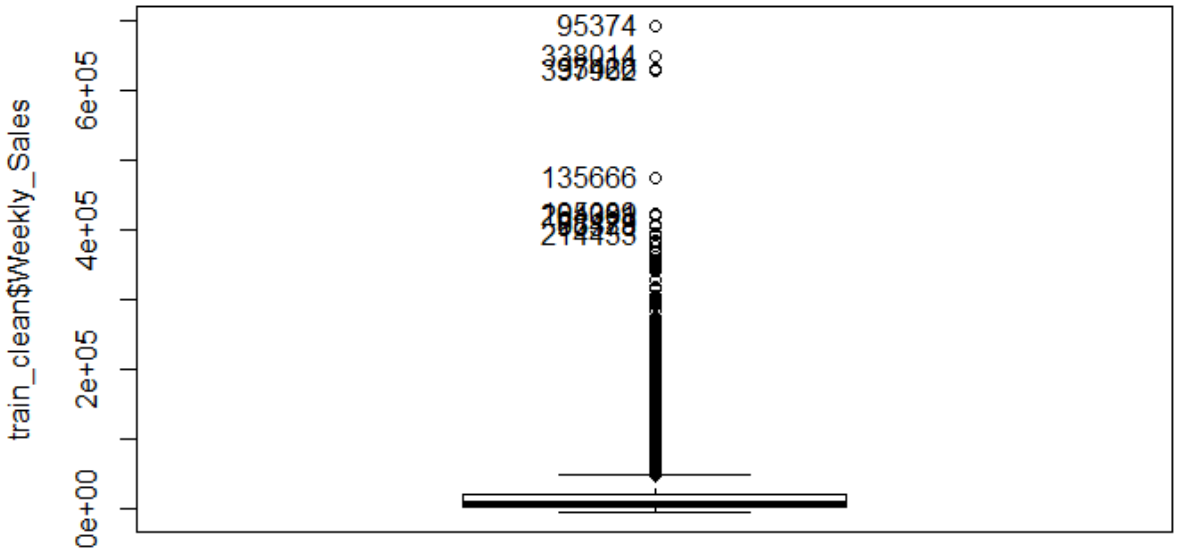


“Stores.csv” contains store type and size data for 45 unique stores, with 3 store types (A, B, C) and sizes ranging from 34,875 to 219,622 sqft, with a mean of 130,288 sqft. There are no missing values nor outliers.

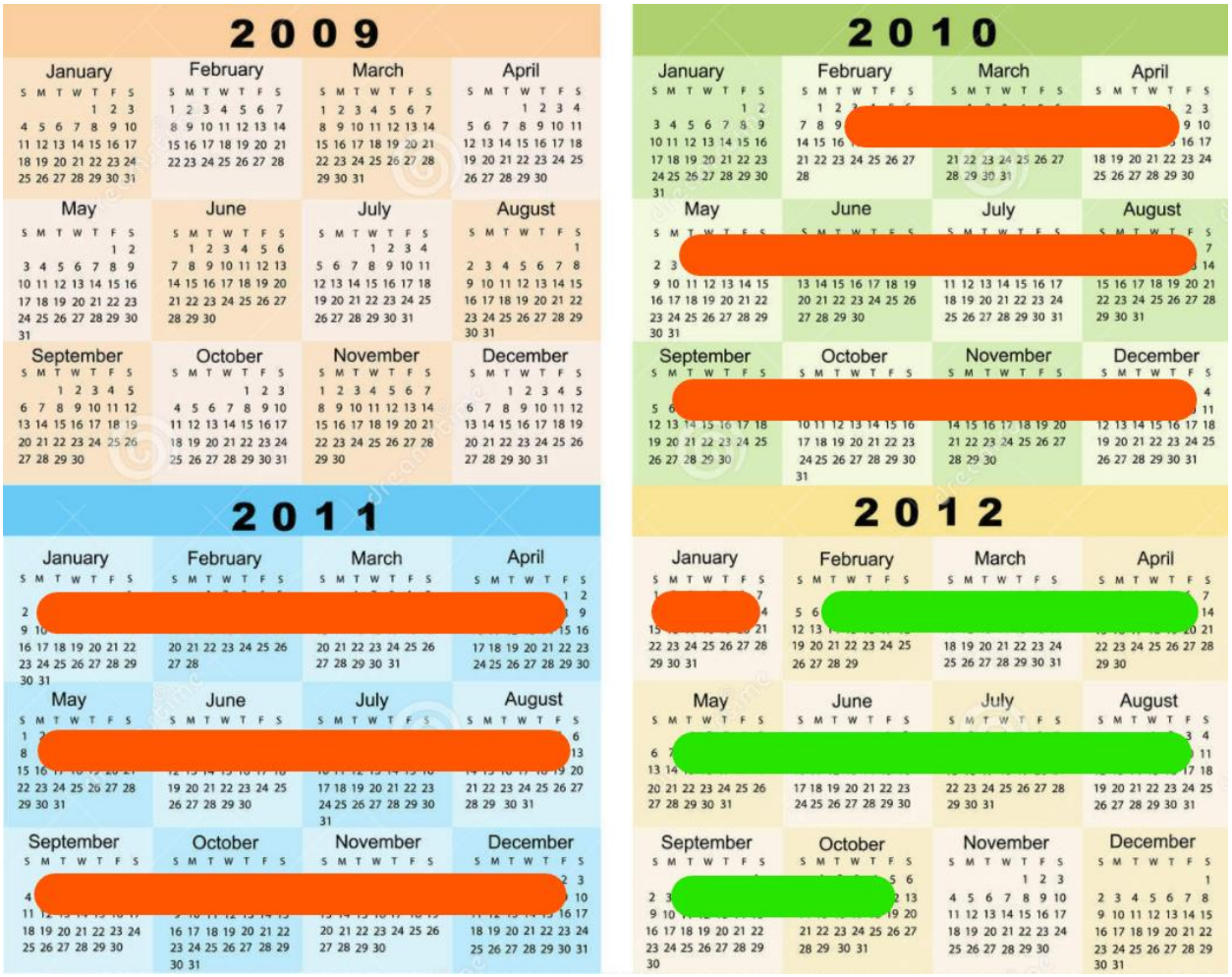
train.csv

```
'data.frame':   421570 obs. of  5 variables:
 $ Store       : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Dept        : Factor w/ 81 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Date        : Date, format: "1970-01-01" "1970-01-01" "1970-01-01" "1970-01-01" ...
 $ Weekly_Sales: num  24925 46039 41596 19404 21828 ...
 $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

```
     Store            Dept             Date              Weekly_Sales      IsHoliday
 13     : 10474   1      :  6435   Min.   :2010-02-05   Min.   : -4989   Mode :logical
 10     : 10315   2      :  6435   1st Qu.:2010-10-08   1st Qu.:  2080   FALSE:391909
 4      : 10272   3      :  6435   Median :2011-06-17   Median :  7612   TRUE :29661
 1      : 10244   4      :  6435   Mean   :2011-06-18   Mean   : 15981
 2      : 10238   7      :  6435   3rd Qu.:2012-02-24   3rd Qu.: 20206
 24     : 10228   8      :  6435   Max.   :2012-10-26   Max.   :693099
 (Other):359799   (Other):382960
```



"Train.csv" contains inputs for machine learning such as the store number, department number, date, weekly sales amounts, and a holiday attribute. There are 45 unique stores and 81 unique departments and every possible combination of the two attributes is iterated over 143 different dates beginning on 2010-02-05 and ending on 2012-10-26 for a total of 421,570 observations. Dates increase by 1 week and are always on Fridays with no gaps. There are no missing data. Outliers are extreme sales volumes surrounding Christmas, Black Friday, and Super Bowl and are not removed since they provide valuable information on holiday sales, and because additional data is captured by the "IsHoliday" attribute. The week number (of the year) will be used to transform the Date column to a numeric variable compatible with machine learning algorithms.

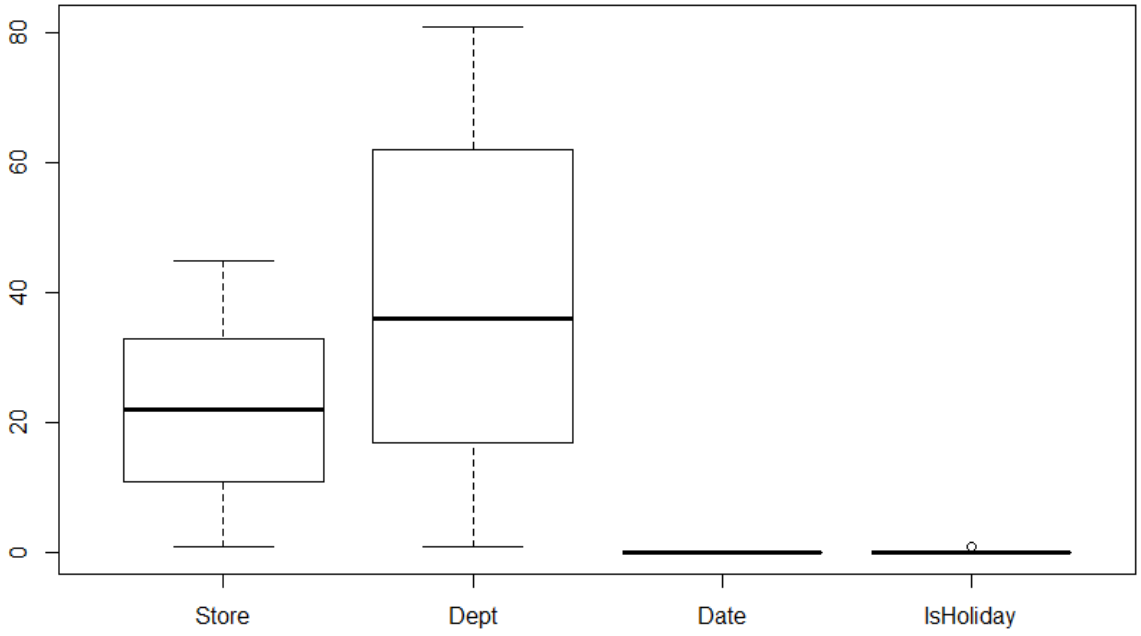


This dataset contains 2 years and 9 months' worth of data. Machine learning models will be trained on 2 years of data (indicated by orange lines above) and tested on 9 months of data (indicated by green lines above).

test.csv

```
'data.frame':   115064 obs. of  4 variables:
 $ Store    : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Dept     : Factor w/ 81 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Date     : Date, format: "1970-01-01" "1970-01-01" "1970-01-01" "1970-01-01" ...
 $ IsHoliday: logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
```

```
     Store            Dept             Date             IsHoliday
13      : 2836   1       :  1755   Min.   :1970-01-01   Mode :logical
4       : 2803   2       :  1755   1st Qu.:1970-01-01   FALSE:106136
19      : 2799   3       :  1755   Median :1970-01-01   TRUE :8928
2       : 2797   4       :  1755   Mean   :1970-01-01
27      : 2791   7       :  1755   3rd Qu.:1970-01-01
24      : 2790   8       :  1755   Max.   :1970-01-01
(Other):98248   (Other):104534
```

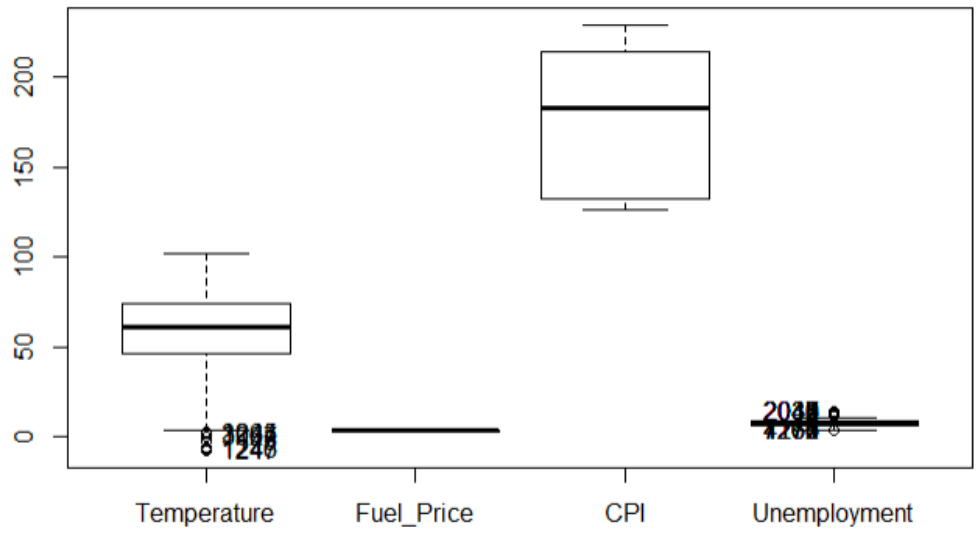


The file “Test.csv” lacks the dependent variable, “Weekly_Sales”, as this dataset was originally used in a Kaggle competition. Since there would be no way to test machine learning models using data missing the dependent variable, this file will be excluded.

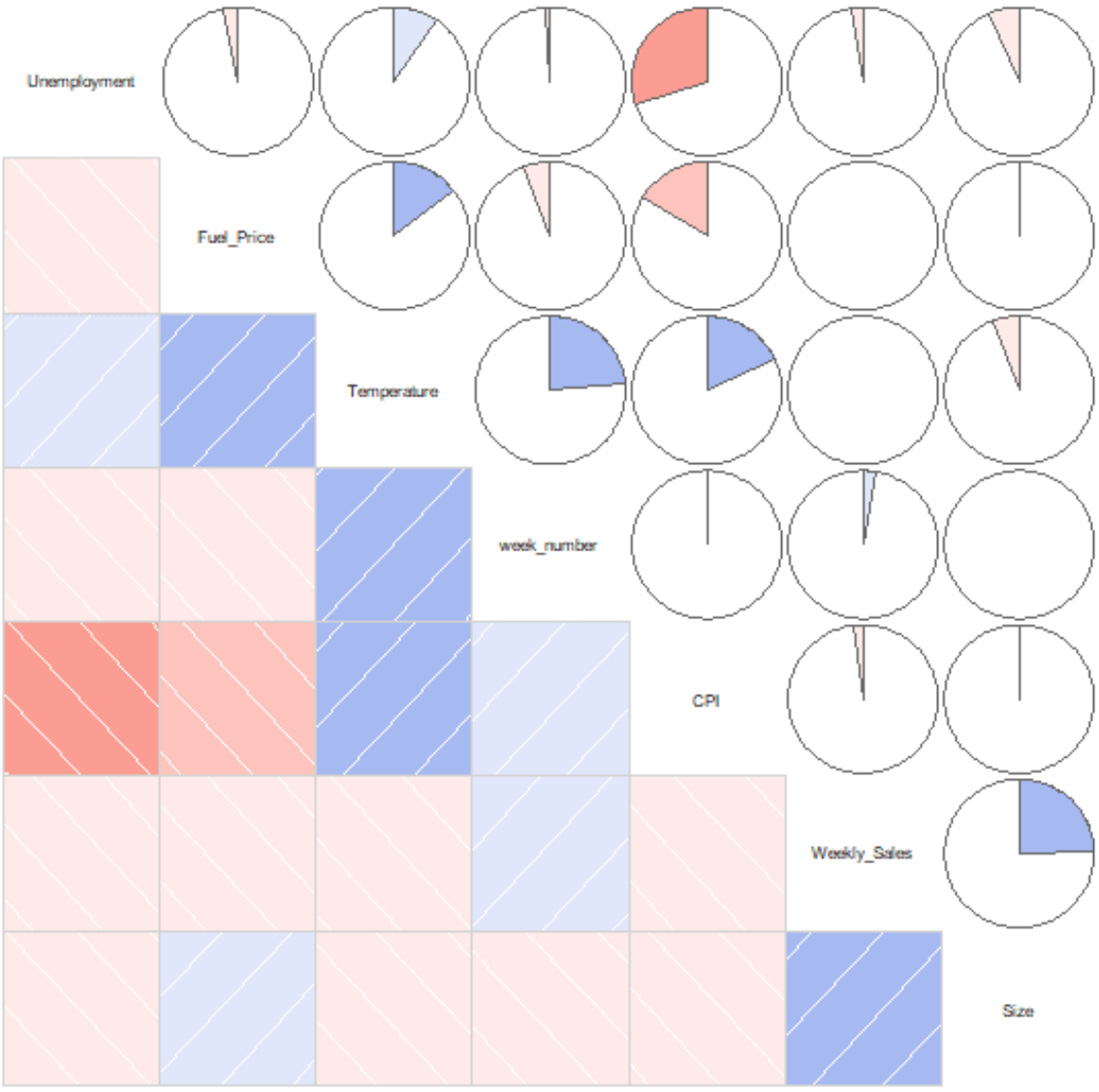
features.csv

```
     Store            Date             Temperature      Fuel_Price     MarkDown1        MarkDown2
1       : 182   Min.   :2010-02-05   Min.   : -7.3   Min.   :2.47   Min.   : -2781   Min.   :  -266
2       : 182   1st Qu.:2010-12-17   1st Qu.: 45.9   1st Qu.:3.04   1st Qu.:  1578   1st Qu.:    69
3       : 182   Median :2011-10-31   Median : 60.7   Median :3.51   Median :  4744   Median :   365
4       : 182   Mean   :2011-10-31   Mean   : 59.4   Mean   :3.41   Mean   :  7032   Mean   :  3384
5       : 182   3rd Qu.:2012-09-14   3rd Qu.: 73.9   3rd Qu.:3.74   3rd Qu.:  8923   3rd Qu.:  2153
6       : 182   Max.   :2013-07-26   Max.   :102.0   Max.   :4.47   Max.   :103185   Max.   :104520
(Other):7098                                                        NA's   :4158     NA's   :5269
   MarkDown3         MarkDown4         MarkDown5           CPI        Unemployment   IsHoliday
Min.   :  -179   Min.   :    0   Min.   :  -185   Min.   :126   Min.   : 4    Mode :logical
1st Qu.:     7   1st Qu.:  305   1st Qu.:  1441   1st Qu.:132   1st Qu.: 7    FALSE:7605
Median :    36   Median : 1176   Median :  2727   Median :183   Median : 8    TRUE :585
Mean   :  1760   Mean   : 3293   Mean   :  4132   Mean   :172   Mean   : 8
3rd Qu.:   163   3rd Qu.: 3310   3rd Qu.:  4833   3rd Qu.:214   3rd Qu.: 9
Max.   :149483   Max.   :67475   Max.   :771448   Max.   :229   Max.   :14
NA's   :4577     NA's   :4726    NA's   :4140     NA's   :585   NA's   :585
```

```
'data.frame':   8190 obs. of  12 variables:
 $ Store       : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Date        : Date, format: "2010-02-05" "2010-02-12" "2010-02-19" "2010-02-26" ...
 $ Temperature : num  42.3 38.5 39.9 46.6 46.5 ...
 $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
 $ MarkDown1   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ MarkDown2   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ MarkDown3   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ MarkDown4   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ MarkDown5   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ CPI         : num  211 211 211 211 211 ...
 $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
 $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```



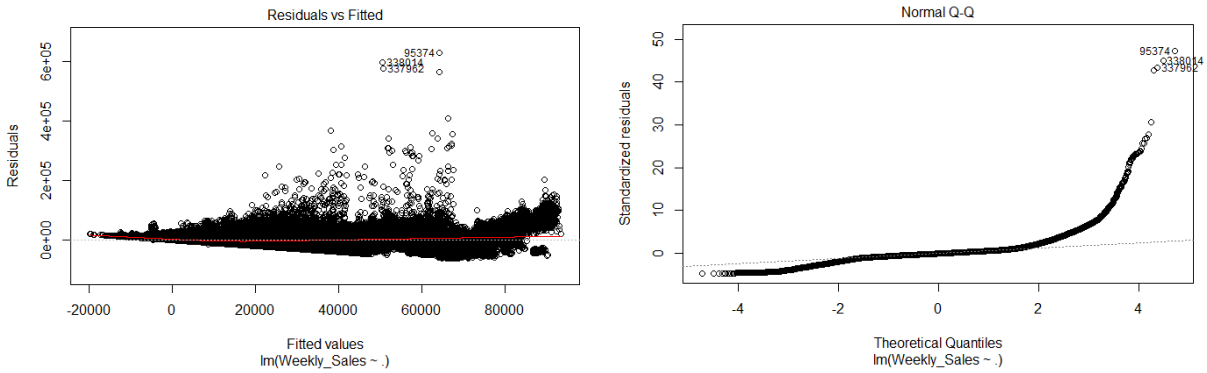
"Features.csv" contains 8,190 observations of additional variables describing each store region's temperature, fuel price, consumer price index, unemployment rate, whether the day is a holiday, and five price markdown event markers. It contains additional dates compared to "train.csv", however CPI and unemployment attributes have NAs in those dates. Since these extra weeks have no associated sales figures for predictive purposes, they will be disregarded. Over half of each of the 5 markdown attributes are filled with NAs, so these attributes will be discarded.



There appears to be a weak, positive correlation between the following attribute pairs:

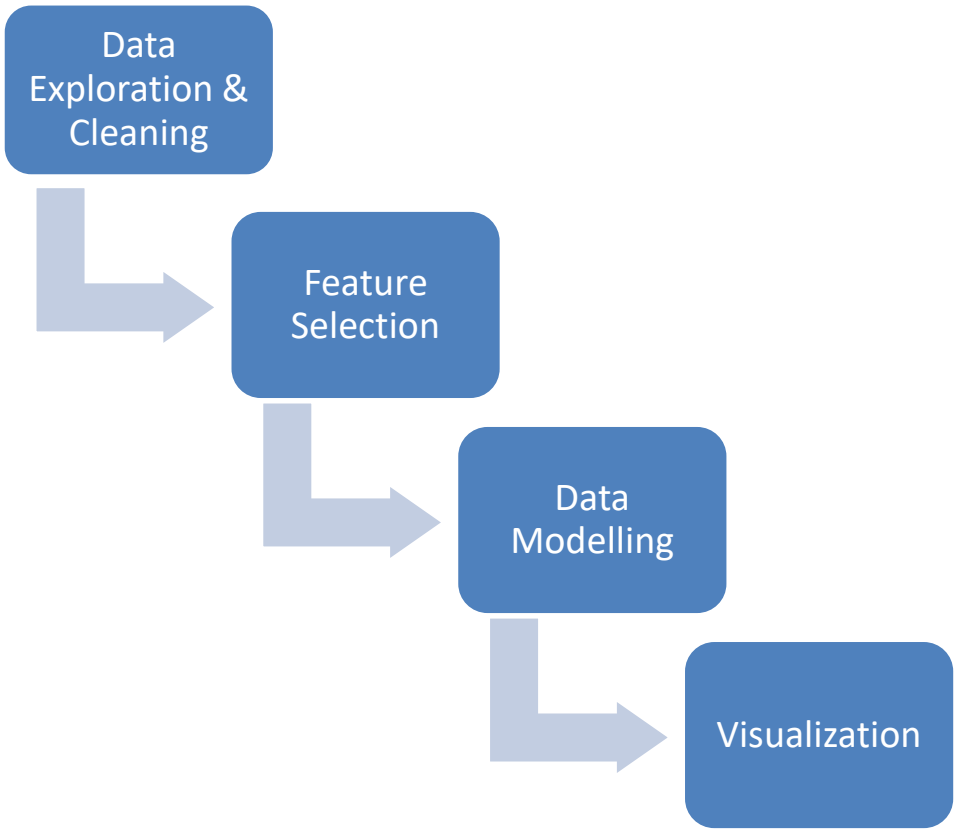
- Fuel price & temperature
- Temperature & week number (time of year)
- Temperature & CPI
- Weekly sales & store size

There appears to be a moderate, negative correlation between unemployment & CPI.



When fitted to a multiple linear regression model, the residuals vs fitted graph shows a funnel-shaped pattern indicative of non-constant variance (heteroskedasticity) among the residuals. The normal Q-Q plot shows a deviation from the dotted straight line, meaning that the residuals have a non-normal distribution. For this dataset, the multiple linear regression algorithm would not provide a suitable fit, so there will be more focus on non-parametric models such as XGBoost and Prophet.

Approach



Step 1: Data Exploration

Import database csv files and explore numbers of observations and attributes, and the data structure and summary. Check for NAs, outliers, and look for relationships between tables. Join training and store tables by store identifier, then join the resulting table with the features table by both store identifier and date.

Step 2: Feature Selection

Perform independent rounds of feature selection for multiple regression, XGBoost, and Prophet. Since multiple regression is a parametric algorithm, check to see if assumptions are met using skewness and normality tests, and check for non-constant variance (heteroskedasticity) and correlations between independent variables (multicollinearity).

Step 3: Data Modelling

Split data into train & test sets (24 months and 9 months respectively) to allow for 2 years' worth of data for training. This method was chosen over the random sampling method since the goal is to use past sales data to predict future data. For the 3 algorithms multiple regression, XGBoost, and Prophet, train models using only the training set to prevent fitting on the test set, and then test the models by making predictions using the test set. Compare predicted and actual values to find the root mean squared error and the percentage of cases with less than 30% error. For XGBoost, use 5-fold cross validation and multiple rounds of modelling for parameter tuning.

Step 4: Visualization

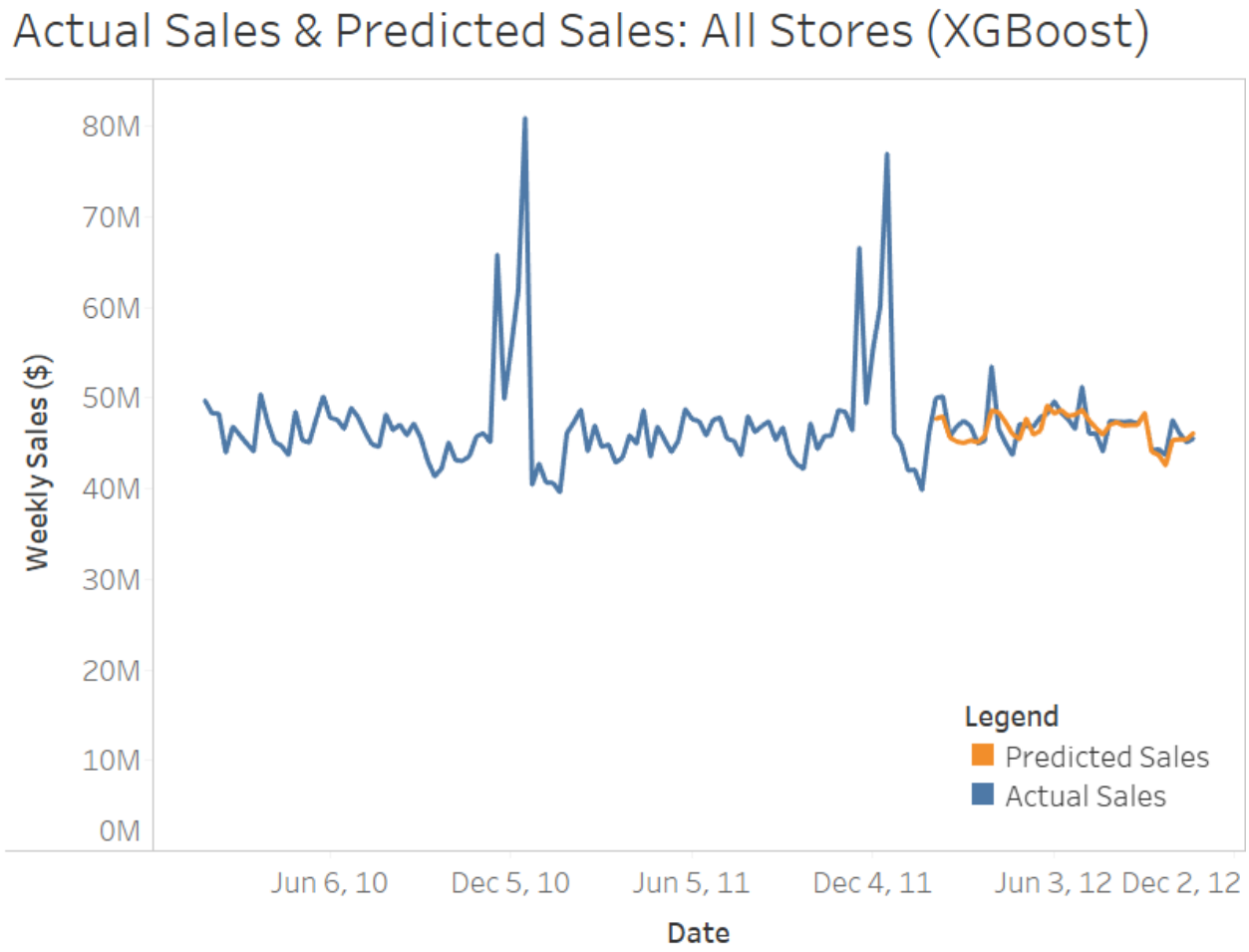
Export resulting predictions and actual, historical values as CSV and plot in Tableau. For XGBoost, also plot decision trees and importance matrix.

Results

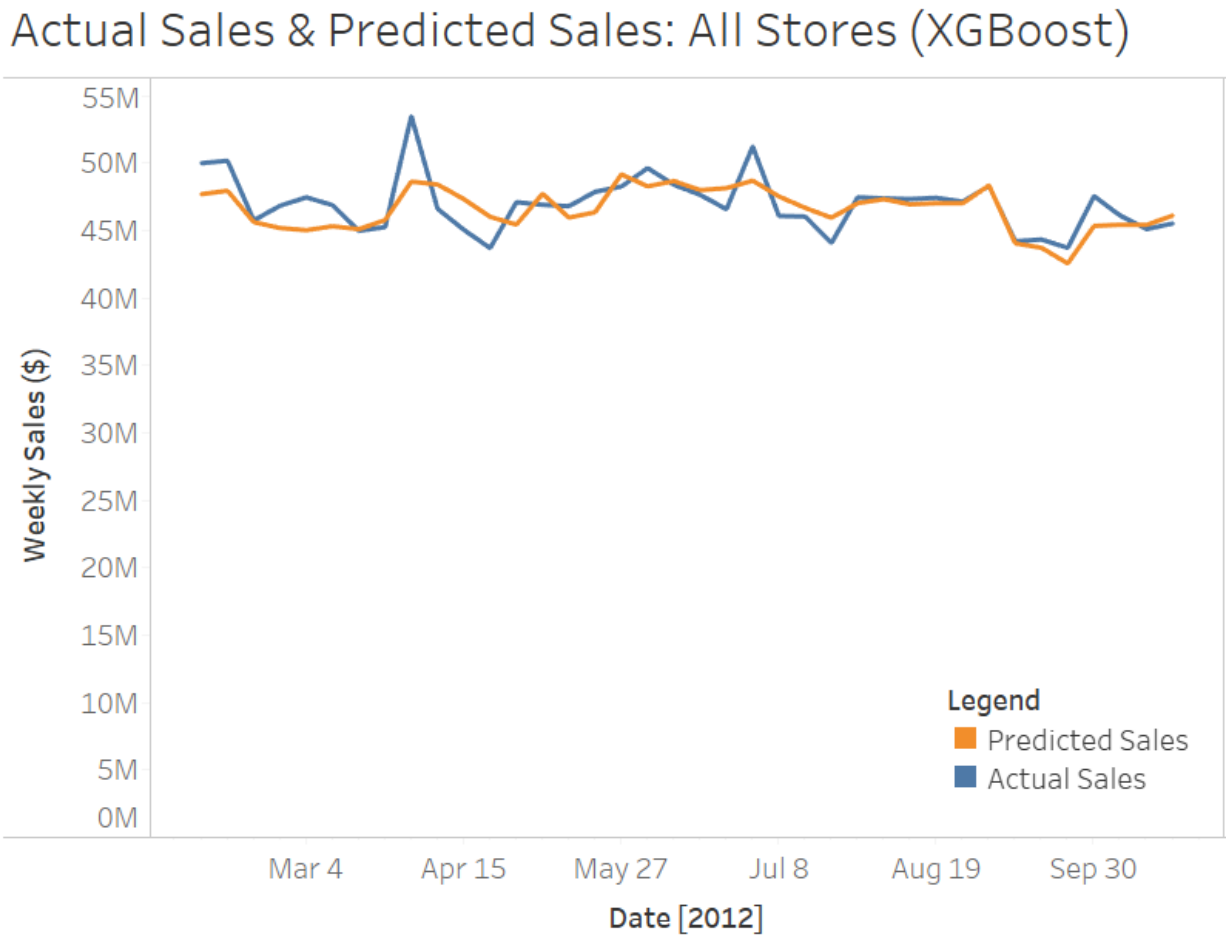
Out of the three machine learning algorithms used, multiple regression was the simplest to implement and had the shortest runtime but was the most restrictive. Since it is a parametric algorithm, there are many assumptions and requirements such as linear relationships between the dependent variable and each independent variable, the normality and homoscedasticity (constant variance) of the residual errors, no correlation between independent variables (lack of multicollinearity). XGboost did not require these characteristics in the training data since it is non-parametric, but it required the most time to implement and had the longest runtime. Prophet is also non-parametric and did not require as much time, but it assumes that there is a seasonality in the time series training data and provides support for only a few regressor (predictive) variables besides the time data and the dependent variable itself.

	RMSE ($)	PRED(30) (% of cases with less than 30% error)
Multiple Regression	12,229	30.7%
XGBoost	4,086	65.9%
Prophet	507,580	42.1%

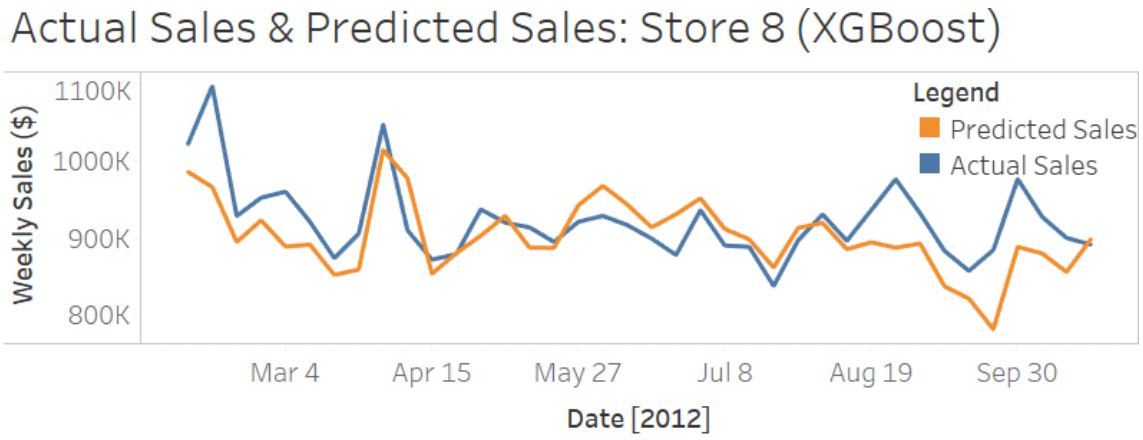
Multiple regression was the worst performing model with a high error metric (12,229) and the lowest % of predicted cases with less than 30% error (30.7%). This is likely because the dataset violates many of the assumptions for linear regression. XGBoost was the best performer, having the lowest error (4,086) and the highest % of predicted cases with less than 30% error (65.9%). This is expected of a gradient boosting algorithm that is robust to non-normal data with non-linear relationships and which can combine many weak models into one accurate model. Prophet had the highest error metric (507,580) but fared well in predicting sales, with 42.1% of predictions having less than 30% error.

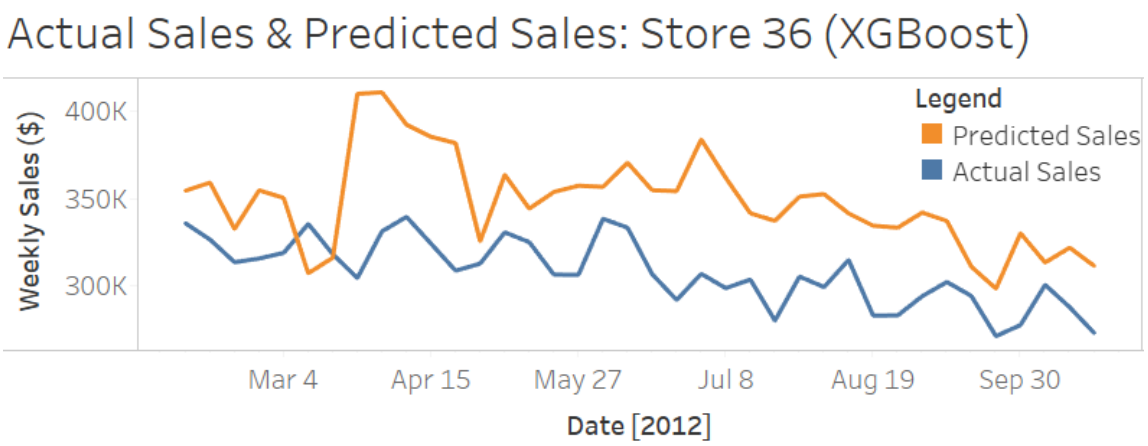


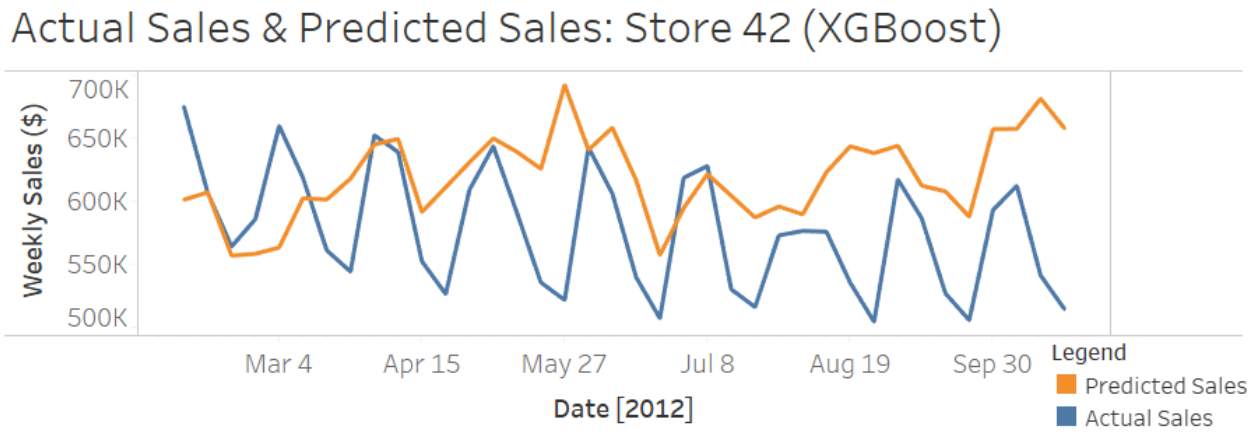
The above graph shows an overview of Walmart's sales in all its stores and all its departments between 2010-02-05 and 2012-02-04. XGBoost modelling was done using 24 months of data starting 2010-02-05, and testing & prediction were done on the following 9 months of data, between 2012-02-05 and 2012-10-26. The four large spikes in the actual sales were from the weeks of Thanksgiving and Christmas over two years, however there was not enough data in the test set to predict sales in November and December. The predicted sales trend (orange) appears to fit the actual sales trend (blue) well, although the spikes in sales were not fully captured.



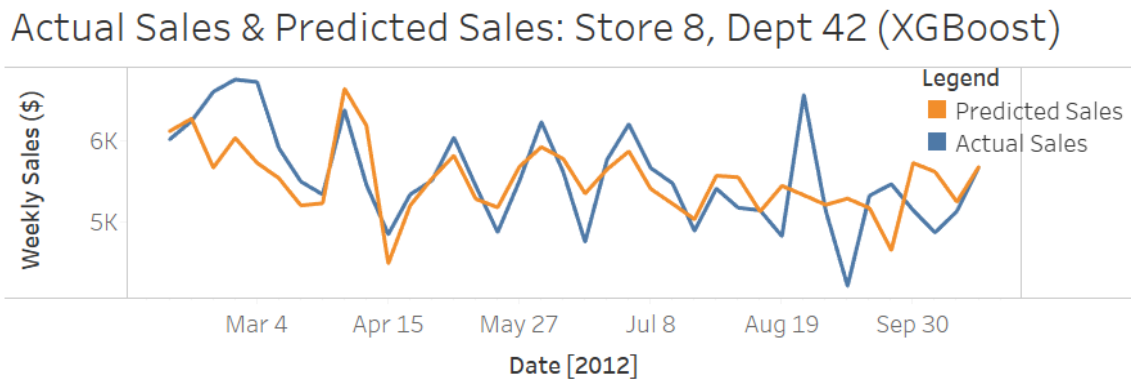
The above graph shows a close-up of the sales data between 2012-02-05 and 2012-10-26 for all Walmart stores. Using XGBoost, most of the peaks and troughs were captured on the right dates, but the larger peaks and troughs were not fully captured in terms of magnitude.

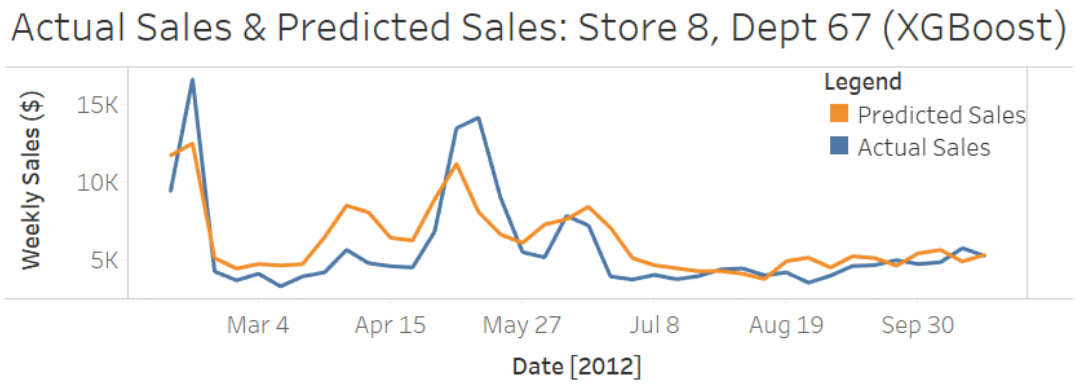


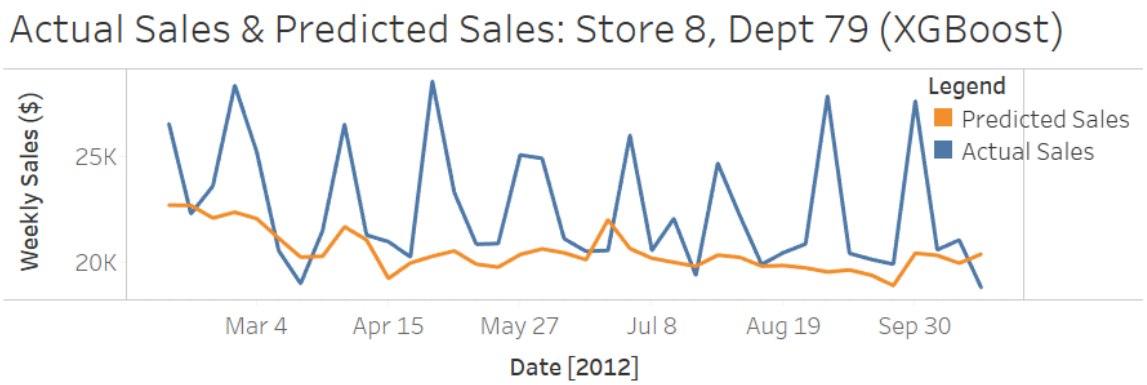




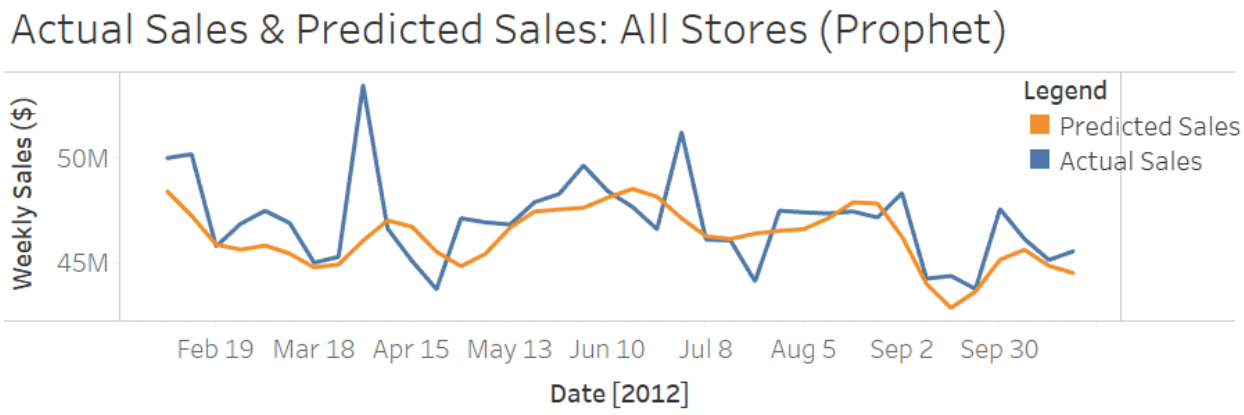
The above graphs show the actual and predicted store-wide sales for 3 of the 45 Walmart stores. Note that the y-axis scales are not the same. Predictions for many of the larger stores such as Store 8 fit well with the actual data, whereas those for the smaller stores such as Store 36 did not fit as well. Although it would be ideal for the model to perform good predictions on smaller store sales also, incorrect predictions would result in smaller sales volume differences in terms of magnitude for such stores. The model had less accurate predictions on stores with unique sales volume patterns such as Store 42 since there were not enough relevant data for training.

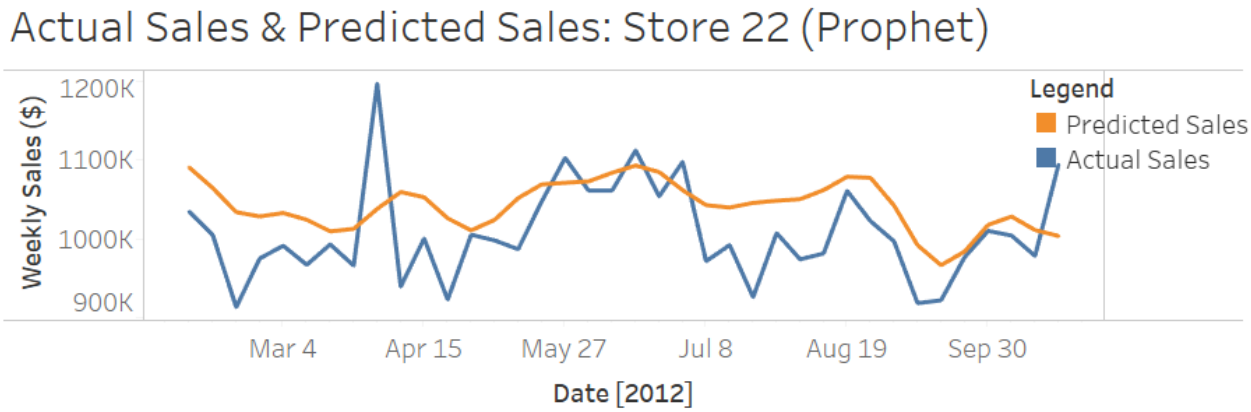


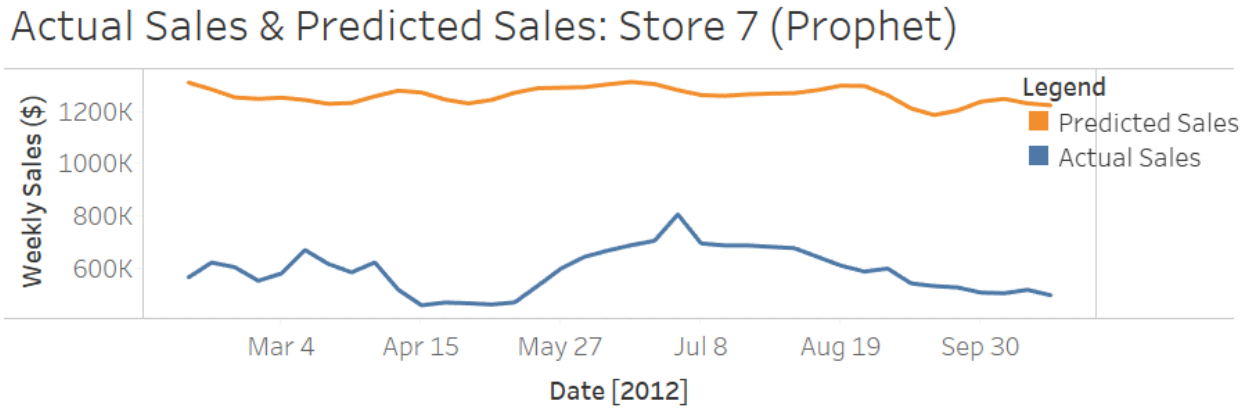




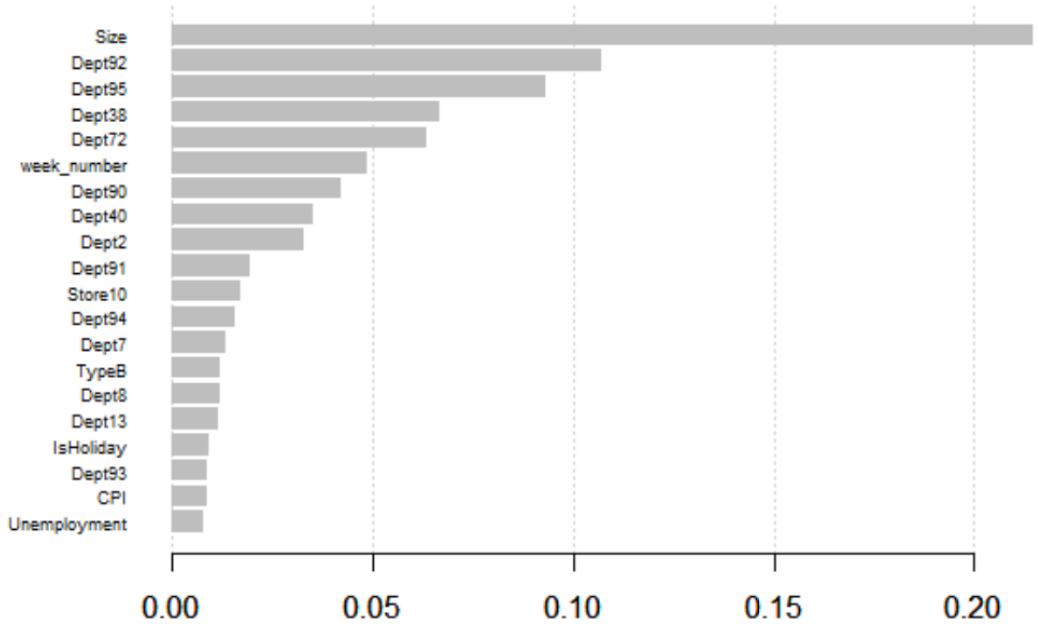
The above graphs show the actual and predicted department-wide sales for some of the 81 departments in store 8. Note that the y-axis scales are not the same. Predictions for departments with small to moderate sales volumes such as Dept 42 & Dept 67 were generally accurate but sudden spikes in sales were not fully captured. The model was less accurate with larger departments with many extreme sales spikes such as Dept 79.



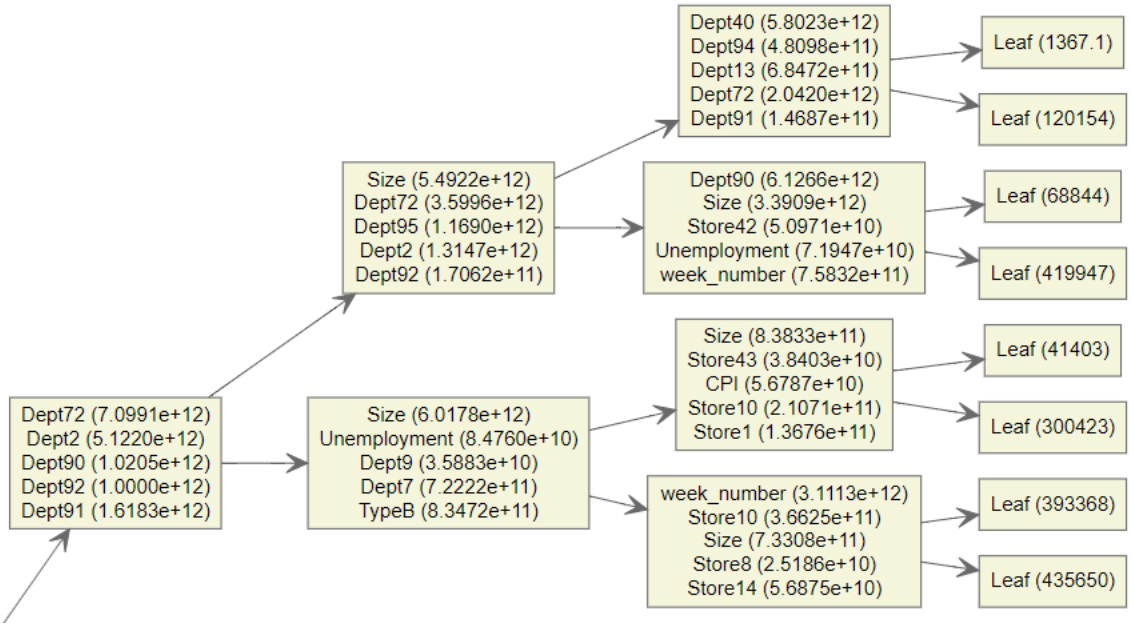




Prophet resulted in a fairly accurate and similar prediction curve as XGBoost for the aggregate sales of all 45 Walmart stores. Note that the y-axis scales are not the same. Unlike XGBoost which fits a different prediction curve for each store and department, Prophet provided the same curve for every store. This resulted in accurate store-wide sales predictions for stores with similar sales curves as the all-store aggregate (Store 22), but very inaccurate predictions for stores such as Store 7, with vastly different sales patterns. The apparent accuracy of the all-store aggregate sales predictions may be due to similar numbers of overestimated and underestimated prediction curves for the individual stores averaging out to the correct values. This is corroborated by the fact that the RMSE for Prophet was much higher than for the other algorithms.



XGboost also provides an importance matrix showing the relative importance of each of the independent variables on the prediction. The above graph shows that the size of the store was the most important attribute for predicting sales volumes, followed by the department number and the week number. Attributes such as holidays, consumer price index, unemployment rates did not play a big role in the predictions, as well as other attributes with even smaller measures of importance such as weather and fuel price.



The above chart is a subset of the aggregate decision tree provided by XGBoost, showing where the nodes were split for the majority of the trees in the model.

Conclusions

This investigation revealed that XGBoost was the best algorithm for predicting weekly sales volumes using the Walmart store sales forecasting dataset, despite taking the most time for modelling. The predictions are fairly accurate at the all-stores aggregate level, with 65.9% of the predicted cases being less than 30% different than the actual cases. At the store level, XGBoost provided more accurate predictions for larger stores than for smaller stores, whereas at the department level, XGBoost performed the best on small- to medium-sized departments. In all cases, large spikes in sales were captured most of the time in terms of the specific date, though not fully in terms of magnitude.

Although it was hypothesized that attributes such as weather, holidays, fuel prices, and unemployment rates can be used as predictors for sales volumes, the results of this investigation showed that these play a minor role. The size of the store, store number, department number, and week number provided the most predictive power. Since these attributes provide information on the past, actual sales curves, it appears that having information on the specific store, department, and the point in time of the prediction is sufficient for generating accurate predictions. This is validated by the fact that the Prophet algorithm generated a similar, fairly accurate prediction curve despite not being provided with any variable other than the dates, weekly sales, and the store number. For this dataset and other similar datasets, the hypothesis that weather, holidays, fuel prices, and unemployment rates provide predictive capabilities for machine learning algorithms is rejected. Having historical sales data and specific store and department numbers is sufficient for making sales predictions for specific stores and departments.

References

Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services*, *52*, pp. 1-13.

de Mézerac, E., Ladoux, E., Blaclard, V., & Ilin, A. (n.d.). *MACHINE LEARNING FOR RETAIL*. Oliver Wyman Digital. https://labs.oliverwyman.com/latest/machine-learning-for-retail.html

Hill, A., & Orrebrant, R. (2014). Increasing sales forecast accuracy with technique adoption in the forecasting process. *Bachelor Thesis, Industrial Engineering and Management, Logistics and Management, School of Engineering, Jonkoping University*, pp. I - 70.

Khaled, A. (n.d.). *Optimizing Retailer Revenue with Sales Forecasting AI*. Toptal. https://www.toptal.com/artificial-intelligence/retail-sales-forecasting-ai

Krishna, A., V, A., Aich, A., & Hegde, C. (2018). Sales-forecasting of Retail Stores using Machine Learning Techniques. *3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, pp. 160-166.

Ziyadat, A. (2019). The Effect of External Environment on Marketing Performance of Retail Stores: Applied Study on Amman City of Jordan. *International Journal of Marketing Studies*, *11*(3).