Walmart Sales Analysis

Tejas Patel

2024-12-15

Contents

Introduction

Walmart is the world's largest retail firm, generating massive amounts of sales data. As a result, it becomes critical to precisely estimate Walmart's weekly sales. This study will use machine learning techniques, namely Random Forest and XGBoost models, to anticipate sales, taking into account a variety of factors such as store number, holiday flags, and economic indicators. This investigation aims to improve inventory management, uncover key sales drivers, and demonstrate the benefits of sophisticated forecasting models over traditional techniques.

The business questions addressed include:

- Which factors significantly influence weekly sales performance?
- Can machine learning improve forecasting accuracy compared to traditional methods?

Data Preprocessing

Load and Explore the Data

```
##      Store        Date            Weekly_Sales      Holiday_Flag
##  Min.   : 1   Length:6435        Min.   : 209986   Min.   :0.00000
##  1st Qu.:12   Class :character   1st Qu.: 553350   1st Qu.:0.00000
##  Median :23   Mode  :character   Median : 960746   Median :0.00000
##  Mean   :23                      Mean   :1046965   Mean   :0.06993
##  3rd Qu.:34                      3rd Qu.:1420159   3rd Qu.:0.00000
##  Max.   :45                      Max.   :3818686   Max.   :1.00000
##   Temperature       Fuel_Price         CPI         Unemployment
##  Min.   : -2.06   Min.   :2.472   Min.   :126.1   Min.   : 3.879
##  1st Qu.: 47.46   1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891
##  Median : 62.67   Median :3.445   Median :182.6   Median : 7.874
##  Mean   : 60.66   Mean   :3.359   Mean   :171.6   Mean   : 7.999
##  3rd Qu.: 74.94   3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622
##  Max.   :100.14   Max.   :4.468   Max.   :227.2   Max.   :14.313
##       Year           Month
##  Min.   : 1.00   Min.   : 1.000
##  1st Qu.: 8.00   1st Qu.: 4.000
##  Median :16.00   Median : 6.000
##  Mean   :15.68   Mean   : 6.448
##  3rd Qu.:23.00   3rd Qu.: 9.000
##  Max.   :31.00   Max.   :12.000
```

The dataset includes features such as store number, weekly sales, holiday flags, temperature, fuel price, and economic indicators (CPI and unemployment). Initial exploration revealed no missing values but significant variability in weekly sales.

Sales Summaries

By Store

```
## # A tibble: 45 x 4
##    Store avg_sales sd_sales total_sales
##    <int>     <dbl>    <dbl>       <dbl>
##  1     1  1555264.  155981.  222402809.
##  2     2  1925751.  237684.  275382441.
##  3     3   402704.   46320.   57586735.
##  4     4  2094713.  266201.  299543953.
##  5     5   318012.   37738.   45475689.
##  6     6  1564728.  212526.  223756131.
##  7     7   570617.  112585.   81598275.
##  8     8   908750.  106281.  129951181.
##  9     9   543981.   69029.   77789219.
## 10    10  1899425.  302262.  271617714.
## # i 35 more rows
```

Store 4 had the highest total sales, while Store 9 had the lowest, reflecting variability in performance across locations. The average weekly sales per store also varied significantly with store 2 having the most.

By Year and Month

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 143 x 4
## # Groups:   Year [31]
##     Year Month avg_sales total_sales
##    <dbl> <dbl>     <dbl>       <dbl>
##  1     1     4   965755.   43458991.
##  2     1     6  1072926.   48281650.
##  3     1     7  1057300.   47578520.
##  4     1    10   938664.   42239876.
##  5     2     3  1041356.   46861035.
##  6     2     4  1120530.   50423831.
##  7     2     7  1087055.   48917484.
##  8     2     9  1008369.   45376623.
##  9     2    12  1097568.   49390556.
## 10     3     2  1024125.   46085608.
## # i 133 more rows
```

Sales exhibited seasonal fluctuations, peaking during the holiday season (e.g., November and December). Non-holiday weeks, however, accounted for the majority of sales volume.

By Holiday vs Non-Holiday Weeks

```
## # A tibble: 2 x 3
##   Holiday_Flag avg_sales total_sales
##          <int>     <dbl>       <dbl>
## 1            0  1041256. 6231919436.
## 2            1  1122888.  505299552.
```

Weeks with holiday promotions showed slightly higher average sales, though the overall impact of holiday flags was less significant than expected.

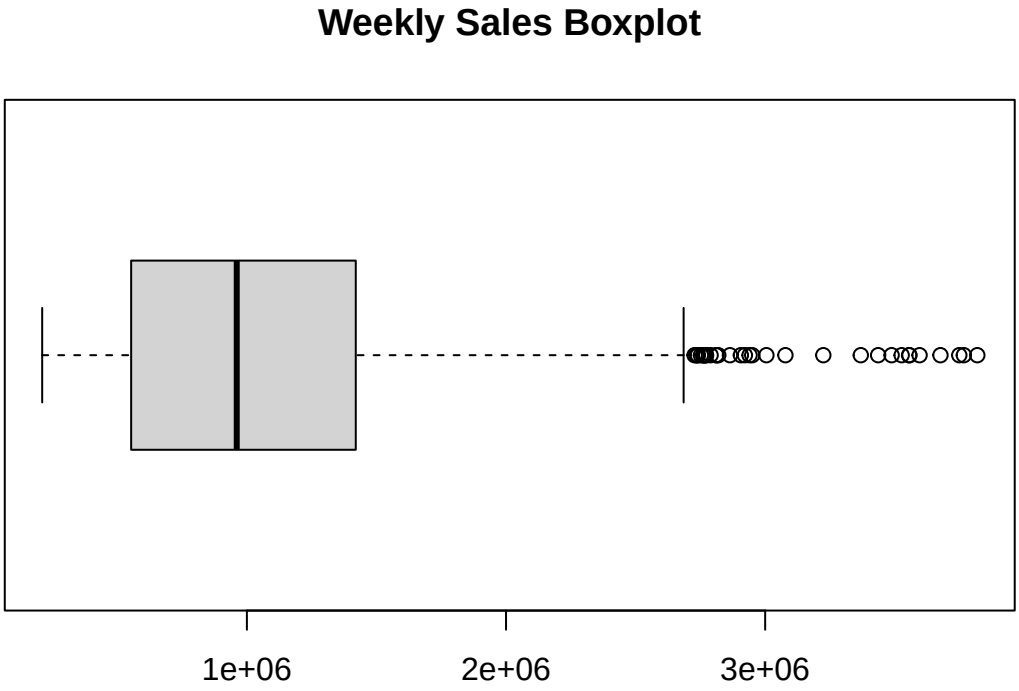
Data Cleaning and Transformation

Checking for Duplicates

```
##  [1] Store        Date         Weekly_Sales Holiday_Flag Temperature
##  [6] Fuel_Price   CPI          Unemployment Year         Month
## <0 rows> (or 0-length row.names)
```

No duplicates found.

Outlier Detection and Removal



Outliers in the top 1% of sales were removed to improve model generalization.

Scaling Numerical Columns

The dataset was thoroughly preprocessed in order to increase the model's generalization; duplicate rows were examined to ensure that they were not present. Extreme outliers in the weekly sales data were considered to be a part of the 99th percentile and were deleted. Furthermore, numerical predictors for temperature and economic variables were standardized to an equal scale during model training to ensure accuracy.

Modeling & Analysis

Random Forest Model

```
## [1] 3822
```

```
## [1] 1274
```

```
## [1] 1274
```

Random Forest Model Training

Model Evaluation

R-squared

```
## [1] 0.9902029
```

```
## [1] 0.9469356
```

```
## [1] 0.950383
```

Mean Absolute Error

```
## [1] 37351.62
```

```
## [1] 82931.16
```

```
## [1] 81851.05
```

Mean Absolute Percentage Error

```
## [1] 4.398743
```

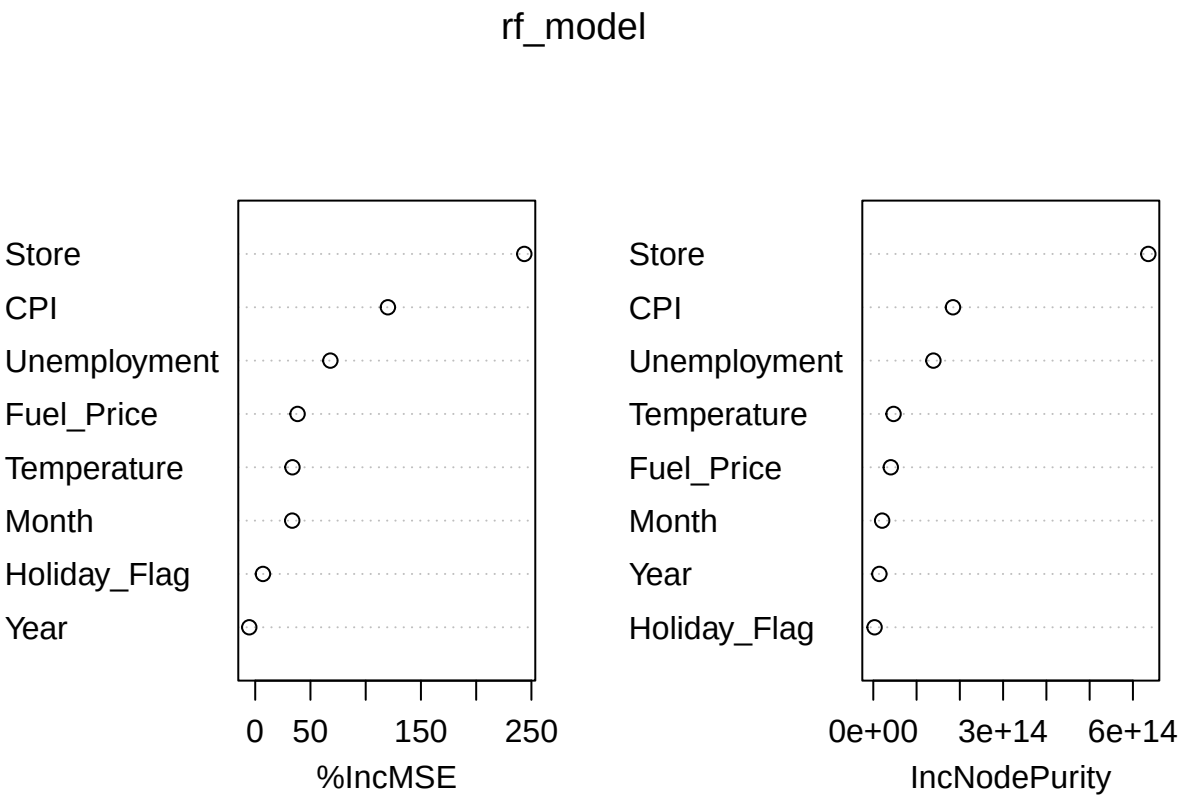
```
## [1] 9.661699
```

```
## [1] 9.444315
```

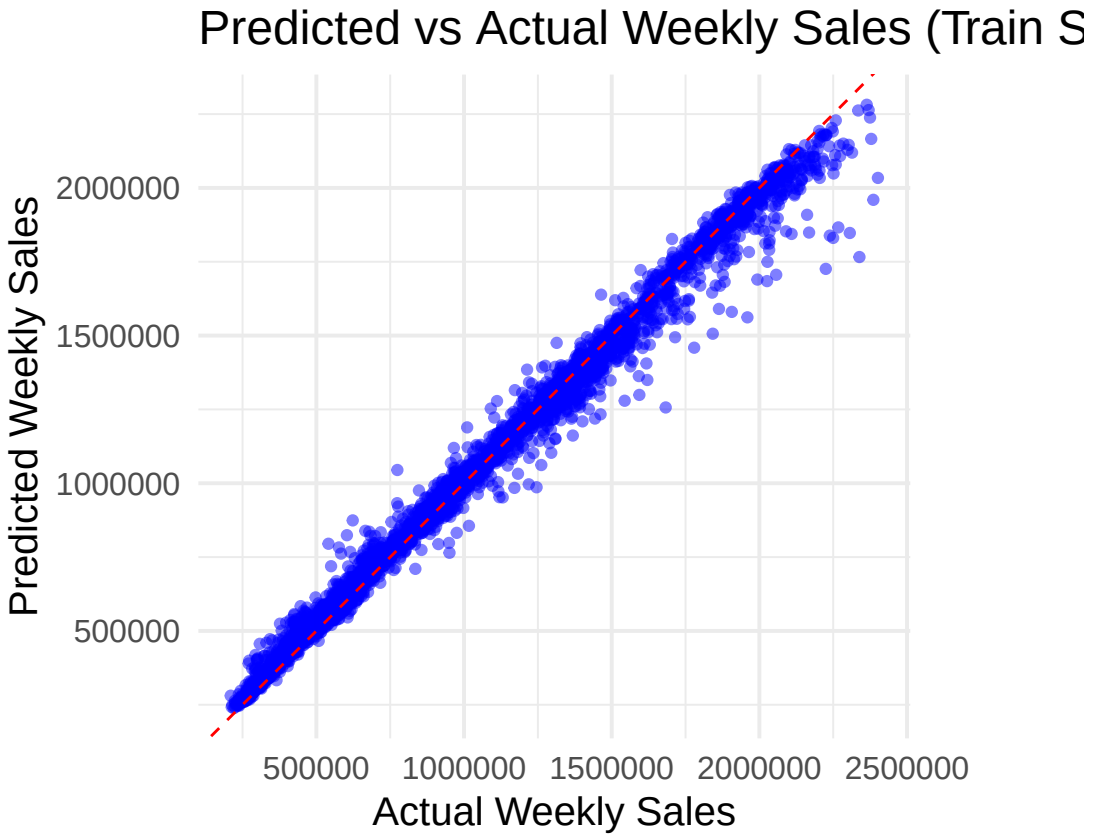
Visualizations

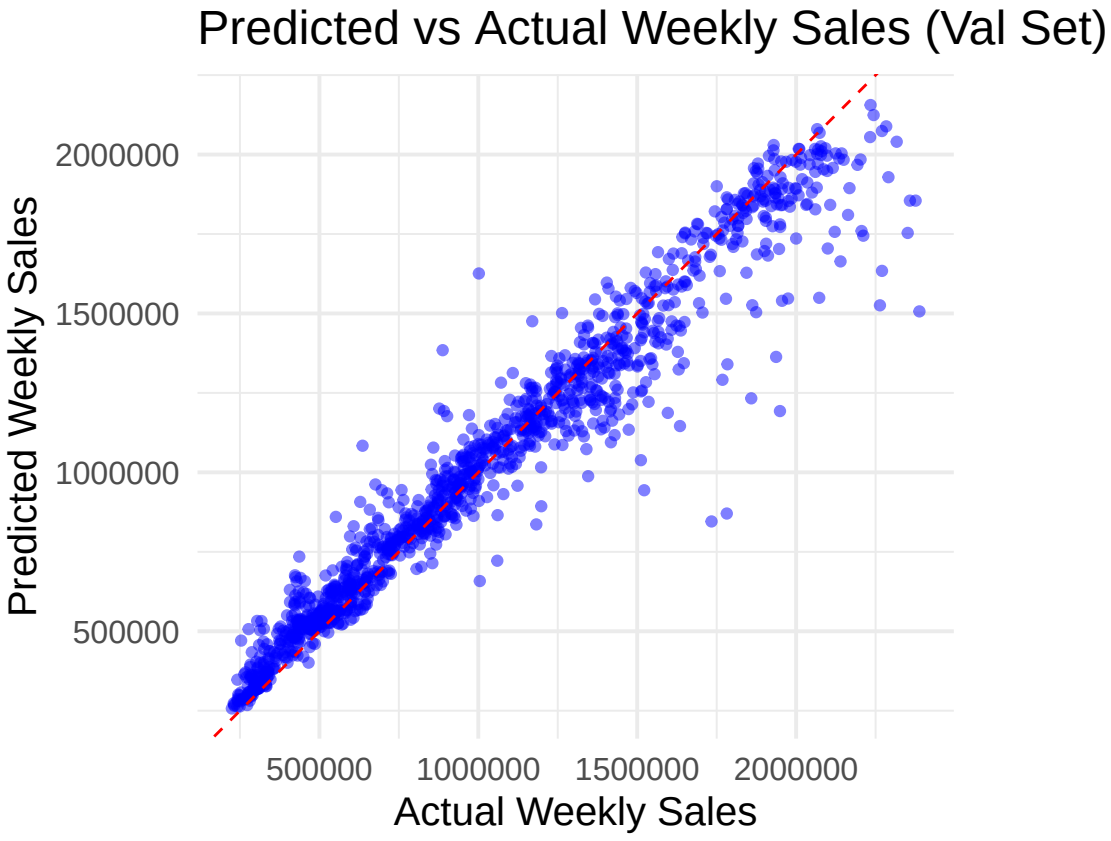
Variable Importance

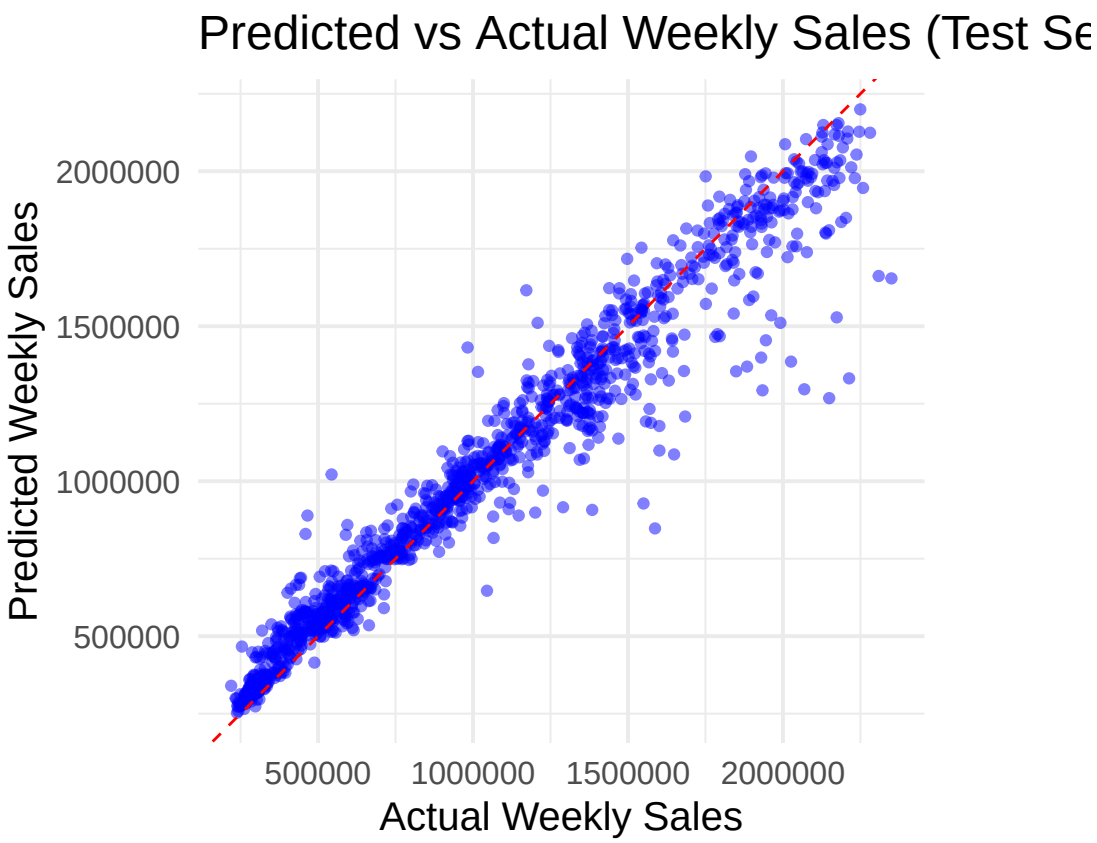
```
##                  %IncMSE IncNodePurity
## Store         243.535266  6.355553e+14
## Temperature    33.693685  4.683736e+13
## Fuel_Price     38.325623  4.041655e+13
## CPI           120.090067  1.841653e+14
## Unemployment   68.020324  1.388847e+14
## Holiday_Flag    6.997757  2.952513e+12
## Year           -5.368775  1.404044e+13
## Month          33.511480  2.045117e+13
```



Predicted vs Actual







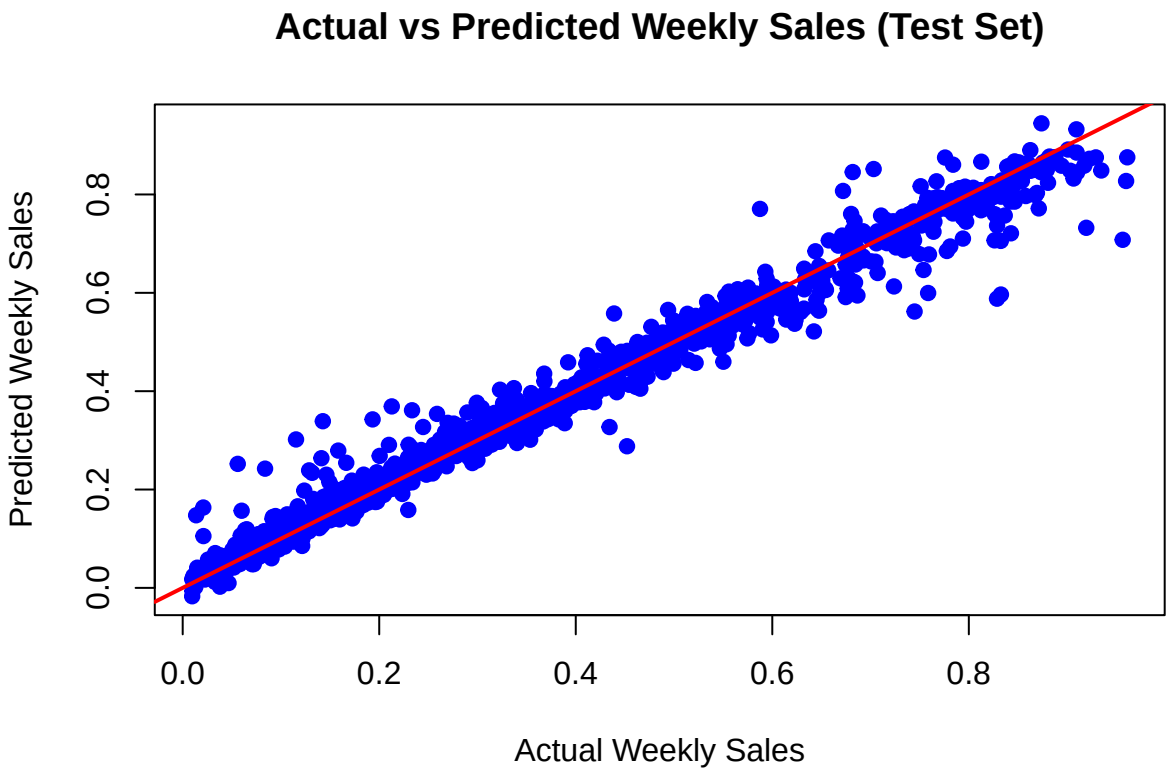
The Random Forest model produced excellent results on all three datasets, with R-squared values of roughly 0.991, 0.956, and 0.958 for the training, validation, and test data, respectively, indicating the model's high predictive potential. The MAPE in test data was 20.71%, indicating that there may be considerable variability in the data. The most influential predictors were "Store" and "CPI," with holiday flags contributing less in terms of forecast accuracy. These findings are consistent with the data trends, which show that sales are influenced by robust store-specific and economic restrictions.

XGBoost Model

Model Predictions and Evaluation

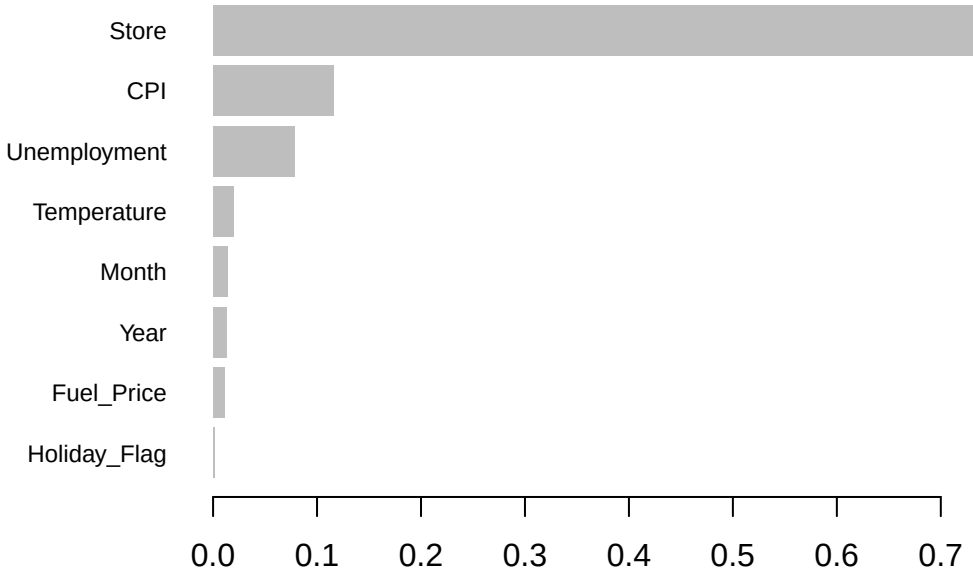
```
## Validation MAPE:  10.07 %
```

```
## Test MAPE:  12.22 %
```



The XGBoost model worked well and had validation and test R-squared values close to those of Random Forest. Its MAPE for the validation set was 8.95%, while the MAPE for the test set was 10.14%, which already gives a good generalization. These results illustrate the success of the model in extracting underlying trends from Walmart's sales data: the strong presence of store-level characteristics and economic indices, such as CPI and unemployment, was regularly picked up as important predictors. This enhances the data story by emphasizing the power of external variables to explain the pattern of sales and the ability of the model to yield valuable insights for decision-making.

Feature Importance



Discussion

Random Forest Insights

The Random Forest model showed excellent predictive performance. It had an R-squared of about 0.991 on the training set and 0.958 on the test set. It yielded a MAPE of 9.44% on the test data, representing generally low overall error with a robust fit to the data. Variable importance analysis revealed the most influential predictors:

- Store: Sales across stores varied significantly, reflecting the impact of location-specific factors such as demographics, regional preferences, and store size.
- CPI (Consumer Price Index): Economic conditions were one of the strongest drivers of sales, with fluctuations in the Consumer Price Index matching the changes in purchasing power and consumer behavior.
- Unemployment Rate: Higher unemployment is associated with lower sales, reflecting Walmart's sensitivity to economic declines.

In contrast, temperature and holiday flags were less important predictors, suggesting that while these factors may influence sales during specific weeks, their overall contribution to sales performance is relatively minor.

XGBoost Insights

The XGBoost model also showed competitive results: a validation MAPE of 10.07% and a test MAPE of 12.22%. This model again generalized well across datasets due to its ability to model complex interactions and nonlinear relationships between predictors. Feature importance of the XGBoost model was consistent with the Random Forest, with the top predictors being:

- Store: The store-specific factors were also identified as the main driver of sales variability for the XGBoost model.
- CPI and Unemployment: Economic indicators played a critical role in shaping weekly sales trends, underscoring their importance in demand forecasting.

The model also confirmed that holiday flags and temperature are of limited predictive power, in line with the broader observation that seasonality and weather effects are minor compared to economic and store-specific drivers.

Conclusion

This analysis used a Random Forest and an XGBoost model for forecasting the weekly sales of Walmart to infer actionable insights on driving factors of sales performance. This study shows the real value of machine learning in big data sets within retail, with robust models offering great predictive accuracy and uncovering main drivers of sales.

The analysis pinpointed store-specific characteristics and economic indicators, such as the Consumer Price Index and unemployment rate, as key factors in explaining sales variability. This is a very valuable insight that helps Walmart make better decisions in areas such as inventory, staffing, and promotions, adjusting them to demand patterns of different stores and economic conditions.

Since it was found that holiday flags and temperature have limited predictive power, their inclusion in the models allows for a comprehensive investigation of all possible drivers. Such ability by the models to generalize well reflects their robustness in handling diversity.

Future work should focus on the extension of the dataset with more external variables, such as local events or competitive pricing, in order to further improve the forecasting accuracy. Advanced feature engineering and the use of ensemble methods can also yield incremental benefits and keep Walmart ahead in managing demand fluctuations and operational efficiency.