PROJECT REPORT

Forecasting Weekly Sales at Walmart Using Regression Algorithm

Neha Cemerla, Varsha Komalla
Sujith Kondreddy, Durgesh Mukkamala
Venkata Renuka Meda

Abstract:

This project focuses on forecasting weekly sales at Walmart using regression algorithms, addressing challenges such as stockouts due to unpredictable demand. The dataset includes historical sales data from 45 Walmart stores, incorporating various events and holidays. The study employs Lasso regression, Decision Tree regression, and K-nearest neighbors (KNN) regression to construct predictive models. Data preprocessing involves cleaning, exploring, and treating outliers, while model performance is assessed using metrics like RMSE and R2 scores. Results indicate that the KNN regressor outperforms Lasso and Decision Tree, achieving an R2 score of 0.90. A 5-fold cross-validation on KNN further confirms its efficacy, making it the recommended model for accurate Walmart sales prediction.

Introduction:

The dataset concerns Walmart, a renowned retail corporation in the United States, and contains information on several events and holidays that impact daily sales. The dataset comprises sales data from 45 Walmart stores. The company faces difficulties, such as stockouts caused by unforeseen high demand, which can be linked to a lack of anticipation and readiness. Hence, there is an urgent want for a proficient machine learning algorithm that can precisely forecast sales and demand, considering factors such as economic conditions (CPI, Unemployment Index, etc.).

Walmart organizes several promotional markdown events each year, intentionally timed to coincide with major holidays such as the Super Bowl, Labor Day, Thanksgiving, and Christmas. Weeks that include these holidays are given a weight that is five times more in the evaluation compared to weeks without holidays. An essential obstacle in this rivalry is to accurately predict the effect of price reductions during holiday periods, even when there is limited or imperfect previous data available.

The dataset consists of past sales data from 45 Walmart locations located in various regions. The primary objective of this work is to utilize regression techniques to develop a resilient regression model using the available data. The main objective is to assist Walmart's business by forecasting weekly sales in advance, allowing retail outlets to appropriately prepare with ample inventory and manpower.

Methods:

The methodology employed in this project encompasses a multi-step approach to forecasting weekly sales at Walmart using regression algorithms. The initial phase involved extensive data cleaning and exploration, where patterns and trends within the dataset were analyzed through visualizations such as data distributions, pie charts for holiday flags, and box plots for outlier detection. Subsequently, the dataset was split into training and test sets for model evaluation. Lasso regression was applied, leveraging the glmnet library for its feature selection and multicollinearity handling capabilities. The Decision Tree algorithm, implemented with the rpart library, captured complex relationships in the data. The K-Nearest Neighbors (KNN) regressor, utilizing the caret library, emerged as the most effective model with superior performance, as evidenced by the highest R-squared score. To validate the KNN model further, a 5-fold cross-validation was conducted, exploring different tuning parameters and assessing performance metrics such as RMSE, R2, and MAE. Overall, this methodology provided a robust framework for enhancing understanding and decision-making in the context of Walmart's weekly sales prediction.

Results:

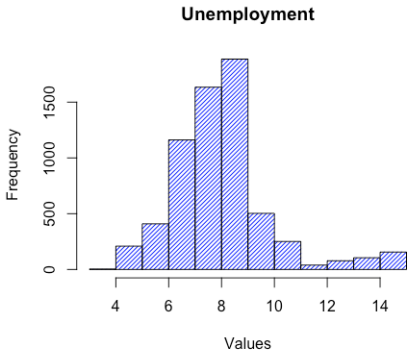
1. Data Loading, Cleaning and Exploration:

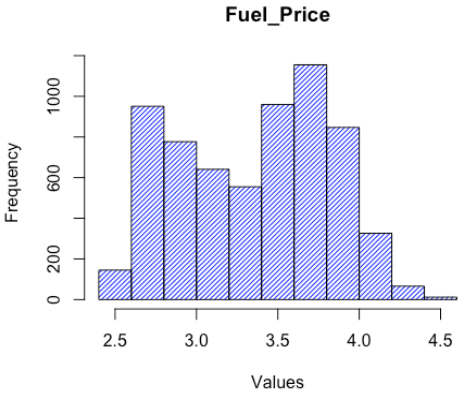
   Conducting an in-depth exploration of the Walmart dataset to identify underlying patterns and trends within the data.

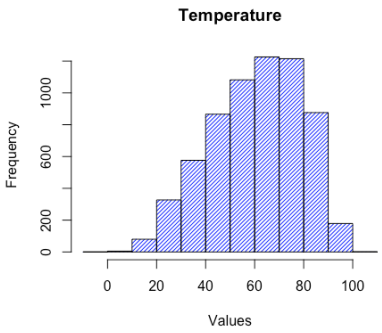
```
> descriptive_stats
     Store        Date           Weekly_Sales      Holiday_Flag      Temperature
 Min.   : 1   Length:6435        Min.   : 209986   Min.   :0.00000   Min.   : -2.06
 1st Qu.:12   Class :character   1st Qu.: 553350   1st Qu.:0.00000   1st Qu.: 47.46
 Median :23   Mode  :character   Median : 960746   Median :0.00000   Median : 62.67
 Mean   :23                      Mean   :1046965   Mean   :0.06993   Mean   : 60.66
 3rd Qu.:34                      3rd Qu.:1420159   3rd Qu.:0.00000   3rd Qu.: 74.94
 Max.   :45                      Max.   :3818686   Max.   :1.00000   Max.   :100.14
   Fuel_Price         CPI         Unemployment
 Min.   :2.472   Min.   :126.1   Min.   : 3.879
 1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891
 Median :3.445   Median :182.6   Median : 7.874
 Mean   :3.359   Mean   :171.6   Mean   : 7.999
 3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622
 Max.   :4.468   Max.   :227.2   Max.   :14.313
```

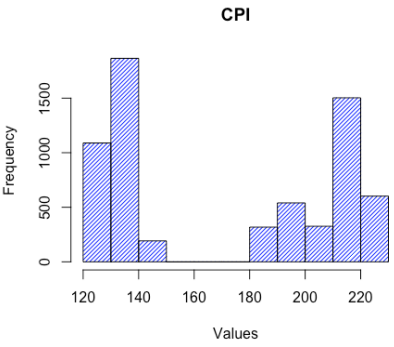
```
> # Display the total number of rows and columns
> cat("Number of Rows:", dimensions[1], "\n")
Number of Rows: 6435
> cat("Number of Columns:", dimensions[2], "\n")
Number of Columns: 8
>
```

Key variables like unemployment, fuel price, temps, CPI, and weekly sales were histogram. The fuel price histogram was bimodal, but unemployment was normal. Although temperatures were normal, the CPI histogram revealed bimodality. Weekly sales were right-skewed, indicating occasional strong sales. These visualizations reveal distribution features, helping predict Walmart's weekly sales by revealing dataset patterns and trends.









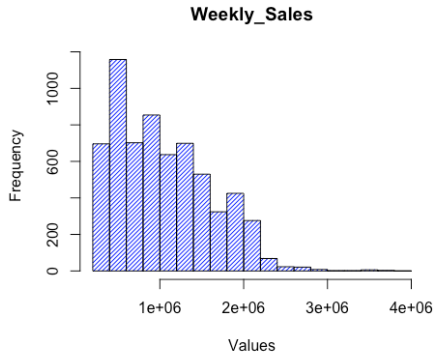


Figure: Data distribution represented

Observation after viewing the distribution.

1. The Weekly_Sales exhibit a right-skewed distribution, which is expected as sales may experience occasional high values.
2. Temperature and Unemployment follow a normal distribution,
3. CPI and Fuel_Price display a bimodal distribution.

B. Analysing Holiday flag in the data

It is evident from the pie chart that 93 percent is non-holiday, and the rest is holiday.

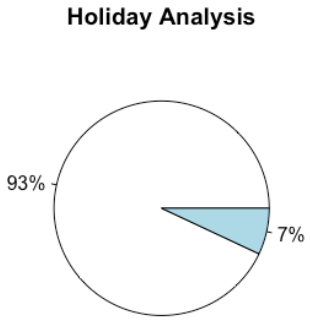


Figure: Pie chart showing the percentage of holidays

C. Detecting outliers in the dataset.

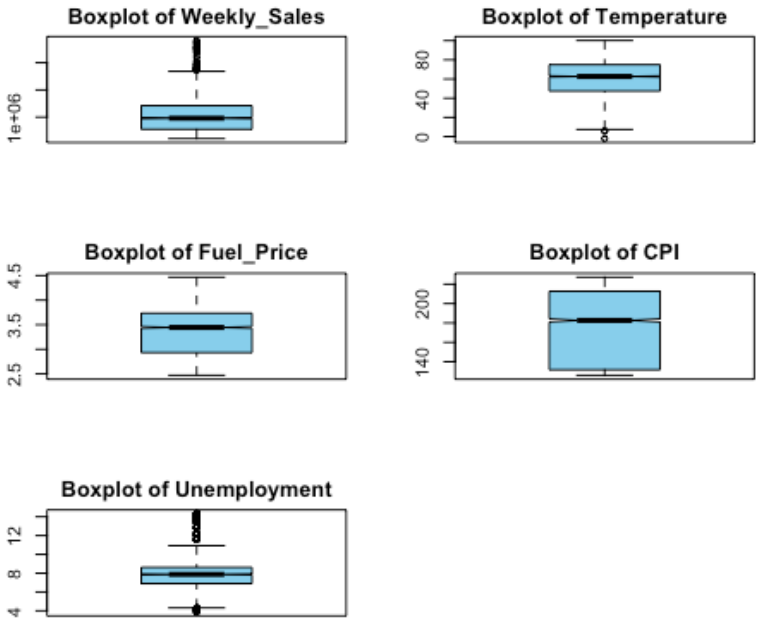


Figure: Outliers detection via box plotting

From the boxplot, it can be observed that Weekly_Sales","Temperature",

and "Unemployment" have an outlier. Thus, outliers were then treated using the quantile feature.

The below figure shows the box plot after removing the outliers.

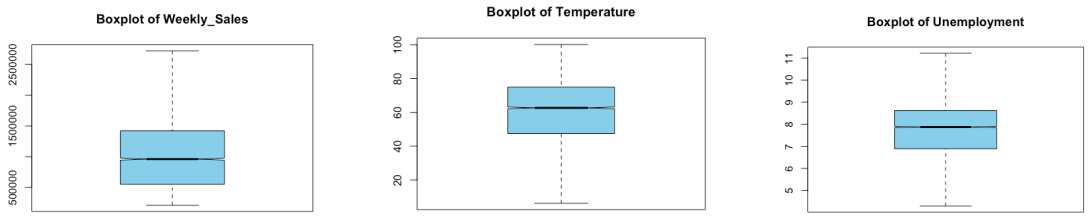


Figure: Boxplots after removing outliers

2. Split of training and test data

Used Sampling method to split train and test dataset from the original data.

Seeds were set to 42 for reproducibility.

And post that X_train, Y_train, X_test and Y_test was created.

3. Lasso Regression Application

Utilizing the Lasso regression algorithm to predict sales. Lasso regression is chosen for its ability to perform feature selection and handle potential multicollinearity among predictor variables.

Glmnet library was used to implement the lasso regression model. While model building the X_train was passed as a matrix and alpha was selected as 1. The Root Mean Square for Lasso Regression came as “511647.015920635". Even the R2 score was calculated for the model, and it turned as only “0.171417631183401”.

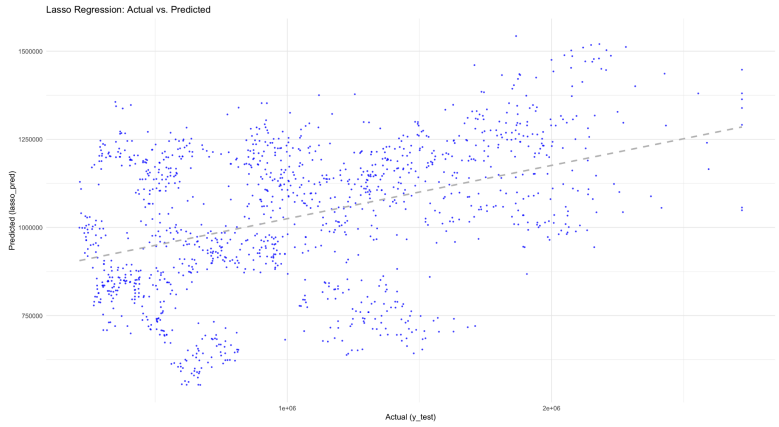


Figure: Actual v/s Predicted for Lasso Regression

4. Implementation of Decision Tree

Applying the Decision Tree algorithm as another predictive modelling technique. Decision trees are beneficial for capturing complex relationships in the data and are interpretable, allowing for insights into the decision-making process.

Rpart library was used for implementing the Decision tree regressor, and the method utilized for implementing was “anova”.

The root mean square and r2 score both was calculated and they came out as "RMSE for Decision Tree: 211038.267205617" and "R2 score for Decision Tree: 0.858472908977968" respectively.

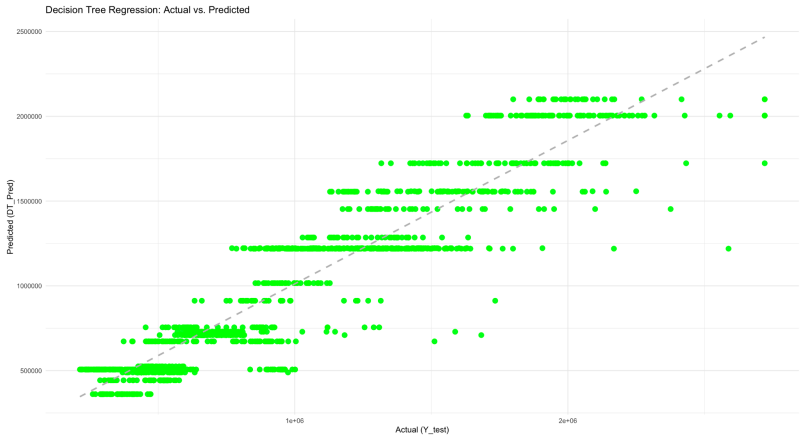


Figure: Actual v/s Predicted for Decision Tree Regressor

5. Application of KNN Regressor

Employing the K-Nearest Neighbours (KNN) regression algorithm on the dataset. KNN is a non-parametric method that makes predictions based on the similarity of data points, making it useful for capturing localized patterns. For this library caret was utilized for training the knn model and the method used was “knn”. This turned out to be the best amongst all the other three models. The RMSE and R2 score calculated for this model was "RMSE for KNN regressor: 170594.876241763" and "R2 score for KNN: 0.907987954336066".

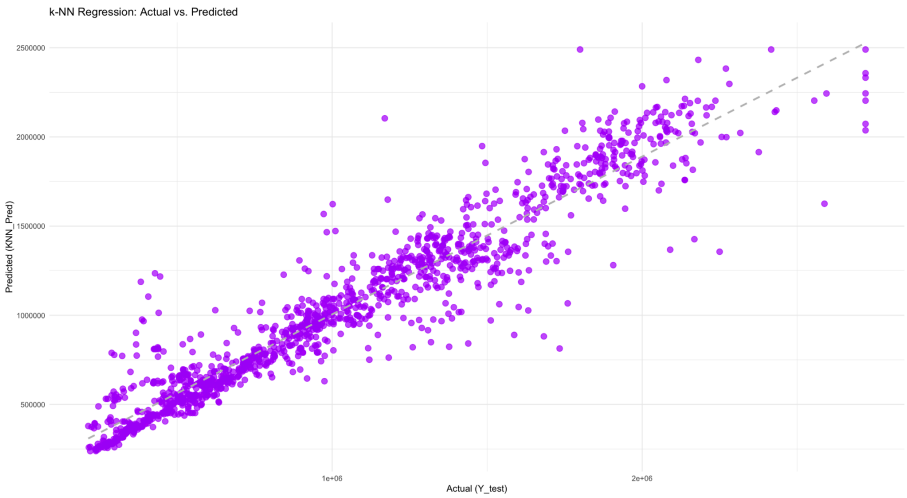


Figure: Actual v/s Predicted for KNN Regressor

6. 5 Fold cross-validation applied for KNN

Since KNN performed the best among all the applied ML regression models thus 5 fold cross-validation was applied on the KNN. A 5 fold cross validation was created with method “cv” as given below

train_control <- trainControl(method = "cv", number = 5)

Then the KNN model was trained by assigning “Weekly_Sales” column as the target variable and the rest other columns as the independent variable as shown below:

model <- train(Weekly_Sales ~., data = data, method = "knn", trControl = train_control)

Below is the output generated:

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 5149, 5148, 5147, 5148, 5148

Resampling results across tuning parameters:

k	*RMSE*	*Rsquared*	*MAE*
5	*176803.9*	*0.9002398*	*111261.3*
7	*179820.7*	*0.8973014*	*113559.4*
9	186710.1	0.8895853	118799.7

Figure: Actual v/s Predicted for 5 fold cross validation KNN Regressor

Performance Comparison:

Comparing the performance of the Lasso regressor, Decision Tree, and KNN regressor models. This comparative analysis aims to identify the strengths and weaknesses of each algorithm in predicting sales for the Walmart dataset.

Overall, this data processing and modeling workflow involve exploring the dataset, applying different regression algorithms, evaluating their performance, and drawing insights from the predictive models to enhance understanding and decision-making in the context of Walmart sales prediction.

Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 5148, 5147, 5149, 5148, 5148
Resampling results across tuning parameters:

k	RMSE	Rsquared	MAE
5	174677.1	0.9027790	110346.3
7	179134.0	0.8981883	113647.8
9	186048.7	0.8903392	118087.4

Conclusion:

Lasso regression, K-nearest neighbors (KNN), and Decision Tree regressor were employed to forecast Walmart's weekly sales. The results revealed that KNN exhibited the highest performance with an R-squared score of 0.90, followed by the Decision Tree regressor with a score of 0.85. Conversely, Lasso regression displayed the least favorable outcome with an R-squared score of only 0.17. Additionally, the Root Mean Squared Error (RMSE) was computed for all three models.

Given the superior performance of KNN, a 5-fold cross-validation was conducted on the KNN model using the 'cv' method, resulting in the best R-squared score of 0.90. Consequently, it can be inferred that the KNN regressor model demonstrated the most effective predictive capability for Walmart sales.

References:

https://www.r-bloggers.com/2021/05/lasso-regression-model-with-r-code/

https://datagy.io/sklearn-decision-tree-classifier/

https://www.r-bloggers.com/2021/05/lasso-regression-model-with-r-code

https://datagy.io/sklearn-decision-tree-classifier/

https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html