Comparision of Multiple Linear Regression model and Random Forest althorithm for predicting Walmart Weekly Sales

Sushmitha Tharla & Shravya Rani Damarapelli

2022-11-08

Introduction

In the running decades trade market has gained immense attention due to economical growth around the World. There have been several retail companies competing in the market for enhancing their sale, but there are several controlling factors such as economic condition, holidays,weather of the area where that retail shop belongs etc. Thus for a retail company it is very important to study the areal factors and predict expected business for a better profit and service, and developing a suitable model including all the required parameter in most convenient and easy way to accomplish. Here in this project we will compare performance of multiple linear regression(MLR) approach and Random forest(RF) approach for predicting our desired outcome, Walmart weekly sale using predictors like Holiday flag, Fuel-price, prevailing consumer price index, prevailing unemployment rate of the area and temperature based on available historical data(2010-2013). Multiple linear regression is a simple way to build a relationship between dependent and independent variables. Random forest is a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree.

Related Works

Enormous number of studies have been conducted till date for predicting Walmart sales data depending on historical data available along with studies based on comparison of different predictive models. Prediction of next 39 weeks Walmart sales was conducted by Harsoor & Patil, 2015 using Holt's winter algorithm, in their study they used same data used in this study. Recently a study "Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017 inspected about which factors affect the sales of Walmart the most. Michael Crown (Crown, 2016) analyzed the same data set to forecast weekly sales of a year using ARIMA model and they evaluated root-mean-square error (RMSE) for model performance measurement.

Rather than only being used in trading or business related fields, Machine learning modeling and related statistical analysis now are expanding to different fields like Medical science, market analysis, weather and climate predictions etcetera. Noi et al.,2017 compared different models like Multiple Linear Regression(MLR), cubist regression and Random Forest algorithms for estimating surface air temperature. They concluded Random Forest algorithm with better R^2 value and lower RMSE, indicate the best fit model among all three. Zhang et al.,2017; Xuefeng et al.,2021 applied machine learning to compare multiple linear regression model and Random Forest algorithm to predict different organic properties of soil and explored that RF perform better than MLR. Huang et al.,2022 conducted a comparative study between MLR and RF machine learning algorithm in medical field to predict Diabetic Urine Albumin–Creatinine Ratio in a 4-Year Follow-Up and revealed that RF may advanced with ability to model highly nonlinear dimensional relationships than other statistical modeling methods.

The review clearly indicates that a notable number of studies has been conducted on using different statistical tools and methods and their combination to built and compare different predictive models from their respective available dataset in order to determine the best-fit model across all.

A comparative study is important to determine the best-fit model for predicting Sales data. Here MLR and RF algorithm models were tested and compared to check their tendency to predict accurately.

Methods

In this study several aspects of data analysis tools were explored to find behavior of each features of the data. However this section comprises of detailed data description, techniques used to predict sales data as well as discusses which model can better explain the prediction. For purpose of this study a MLR was built manually using Matrix method and for exploring RF algorithm in-built package "RandomForest" of R was used.

About the dataset: Walmart Sales data from (https://www.kaggle.com/datasets/aditya6196/retail-analysis-with-walmart-data?resource=download) has been used for this study. The data is about weekly sales data of 45 Walmart stores along with regional factors like Temperature, Fuel_Price, CPI, Unemployment, holiday flag etc.

Holiday flag comprises of two factors 1 or 0. If the week contains any of the holidays mentioned below then the flag is set to 1 otherwise 0.

Holiday Name	Date 1	Date 2	Date 3
Super Bowl	12-Feb-10	11-Feb-11	10-Feb-12
Labor Day	10-Sep-10	9-Sep-11	7-Sep-12
Thanksgiving	26-Nov-10	25-Nov-11	23-Nov-12
Christmas	31-Dec-10	30-Dec-11	28-Dec-12

A summary of different variables of the data set is shown in the following image.

```
summary(dat)
```

```
##      Store          Date            Weekly_Sales     Holiday_Flag
##  1      : 143   Length:6435        Min.   : 209986   0:5985
##  2      : 143   Class :character   1st Qu.: 553350   1: 450
##  3      : 143   Mode  :character   Median : 960746
##  4      : 143                      Mean   :1046965
##  5      : 143                      3rd Qu.:1420159
##  6      : 143                      Max.   :3818686
##  (Other):5577
##   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
##
```

Fig 1:Summary of data.

Checking whether there are NA values.

```
library(inspectdf)
y<-inspect_na(dat)
show_plot(y)
```

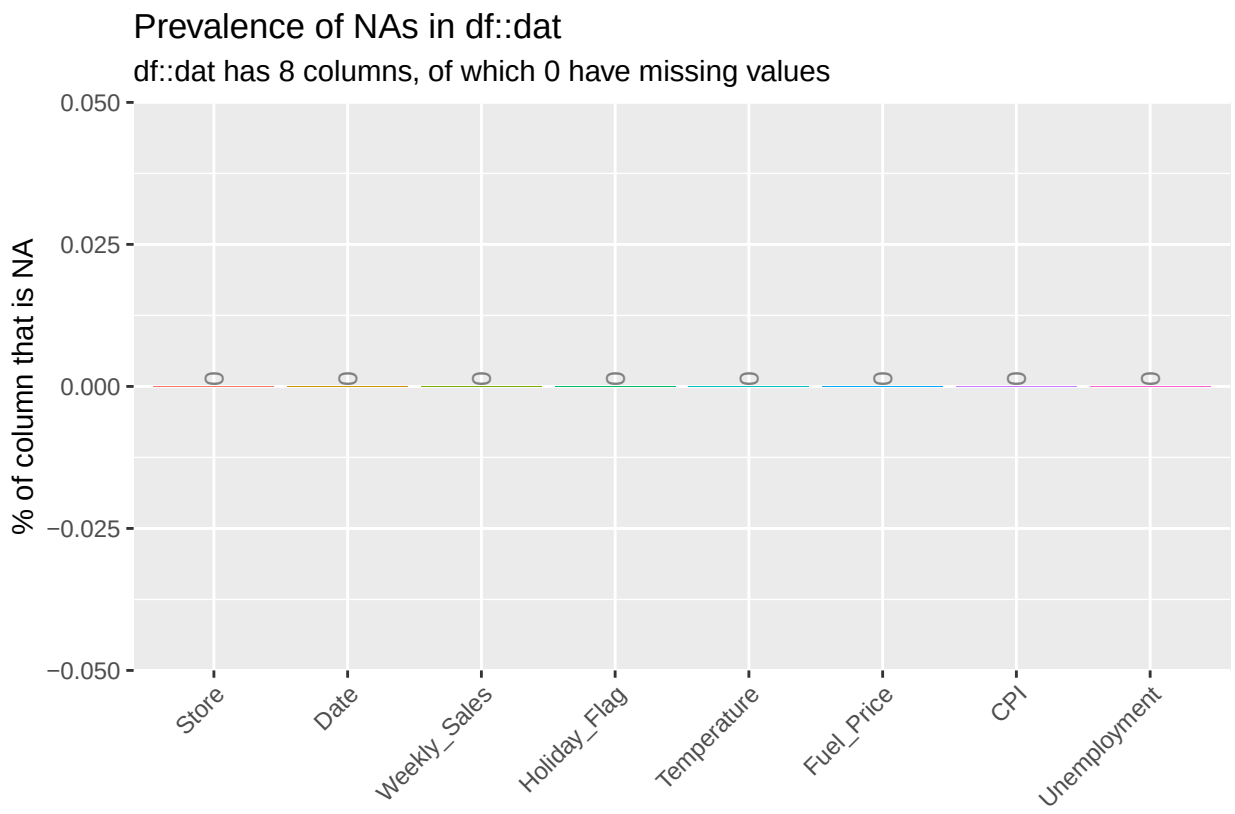

Fig 2: NA or missing values in the data.

Thus no data is missing in the data.

Exploring types of the data:

```
x<-inspect_types(dat)
show_plot(x)
```

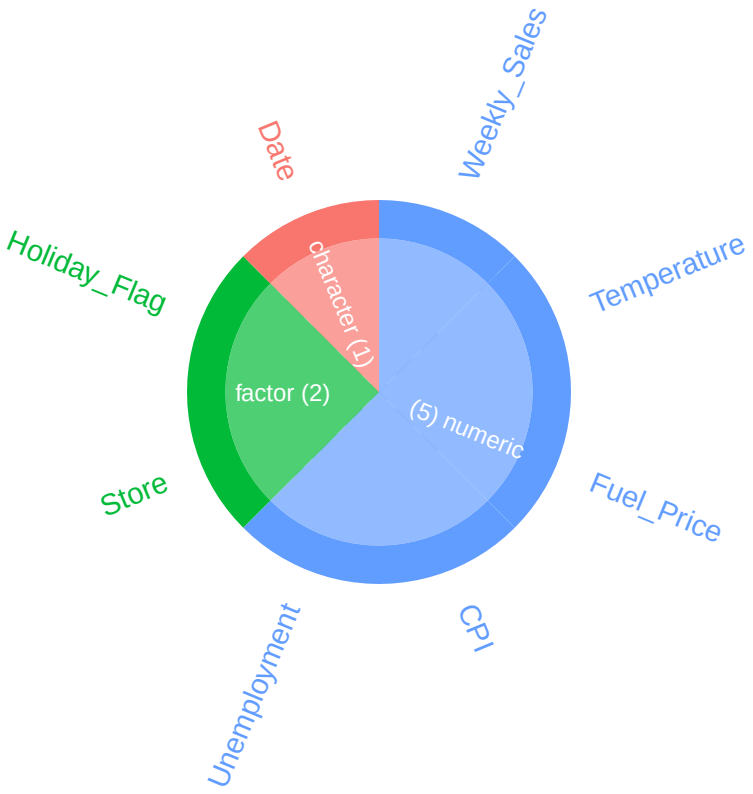

Fig 3: Data types.

For a clean and accurate analysis it is very important to detect and remove outliers from the data. For detecting the outliers boxplot is a powerful tool. First we will check for outliers in the numerical dependent variables.

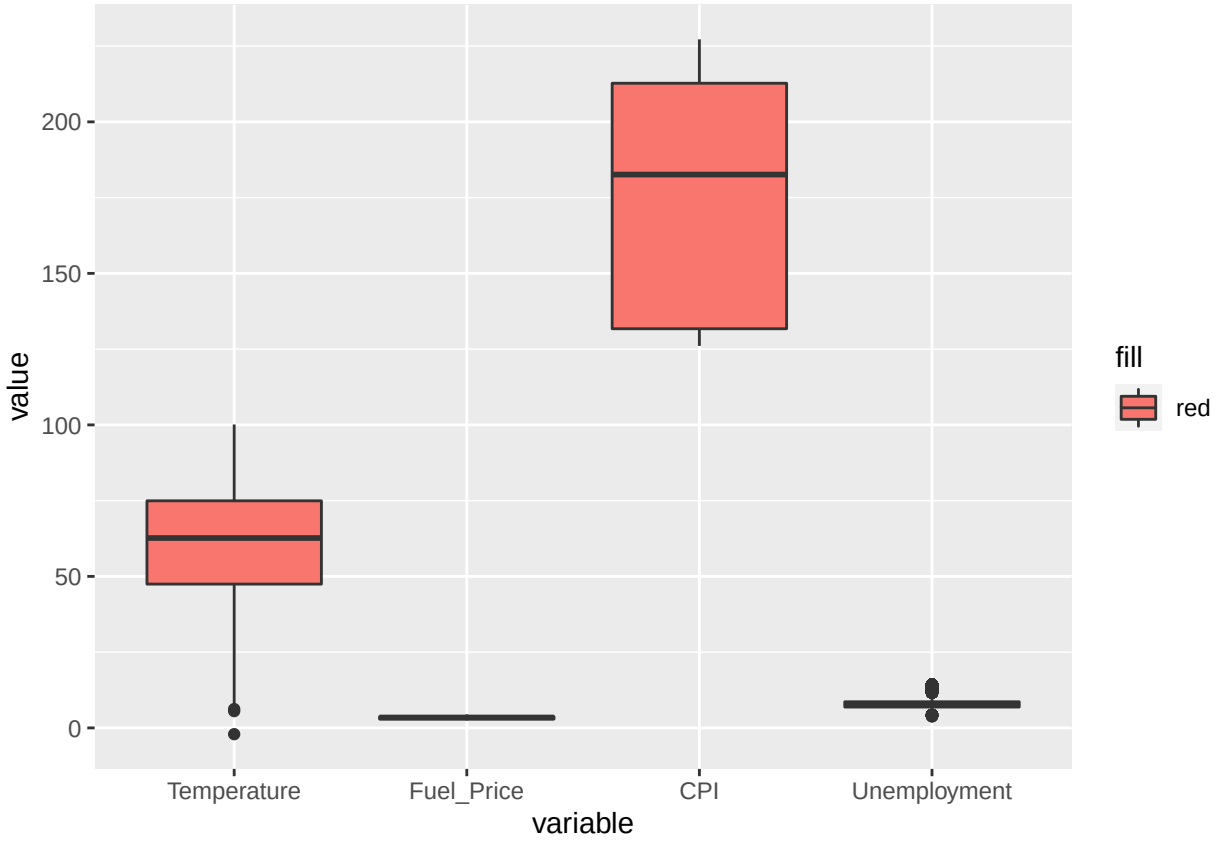

Fig 4: Boxplots showing outliers in the numerical predictors.

Variable employment has several outliers. So for removing the outliers 25th and 75 quantiles are calculated and the data points lying outside this quantile range are considered as outliers.

```
quartiles <- quantile(dat$Unemployment, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(dat$Unemployment)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

data_no_outlier <- subset(dat, dat$Unemployment > Lower & dat$Unemployment < Upper)
```

Dependent variable also has a large number of outliers. So it is also crucial to remove them

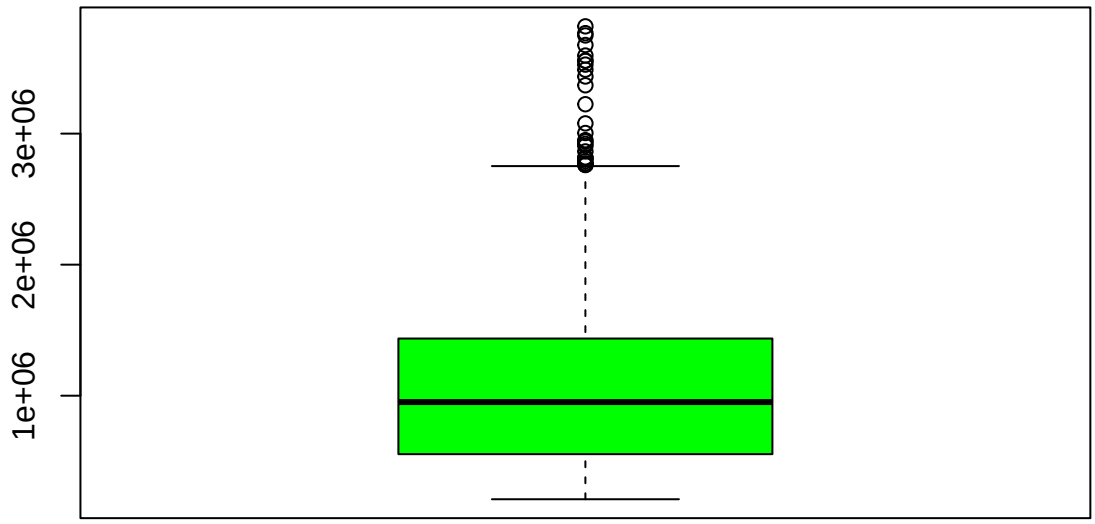

Fig 5: Boxplots showing outliers in the dependent variable.

Here the dependent variable is upper skewed , thus it is very important to remove the outliers for getting a less error model.

```
quartiles <- quantile(data_no_outlier$Weekly_Sales, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(data_no_outlier$Weekly_Sales)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR
data_no_outlier<-subset(data_no_outlier,data_no_outlier$Weekly_Sales>Lower &data_no_outlier$Weekly_Sales
dat<-as.data.frame(data_no_outlier)
```

Below plot shows distribution of numerical predictors in the data.

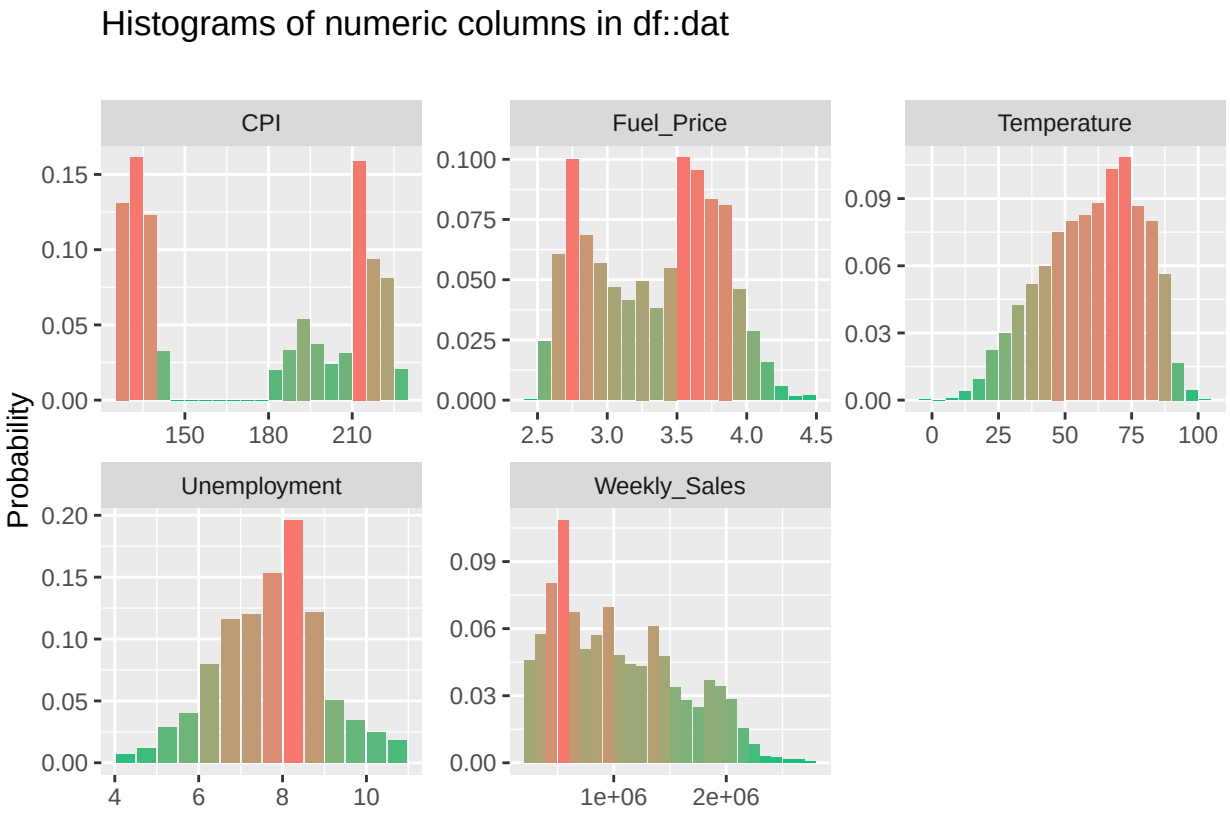

Fig 6: Histogram showing data distribution of numerical variables.

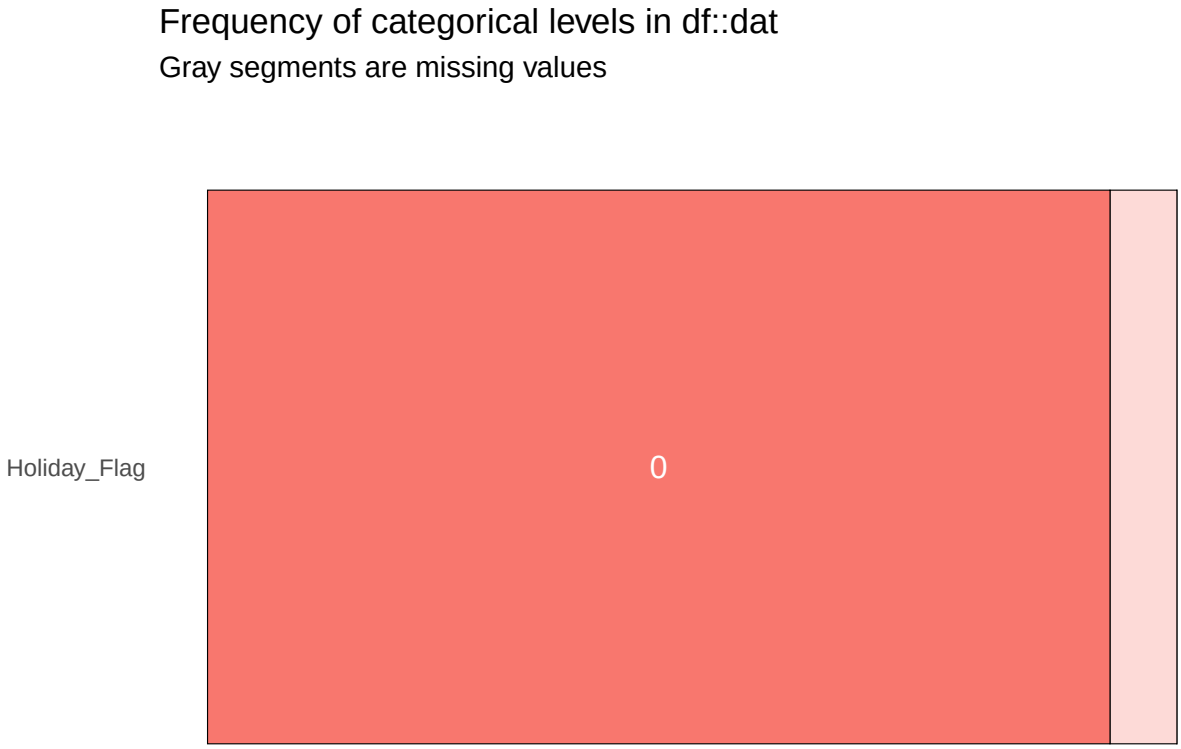

Fig 7: Distribution of Categorical Variable holiday_Flag.

Only few days are holidays.

```
## # A tibble: 4 x 7
##   col_1        col_2          corr  p_value    lower   upper pcnt_nna
##   <chr>        <chr>         <dbl>    <dbl>    <dbl>   <dbl>    <dbl>
## 1 Weekly_Sales CPI          -0.0834 1.42e-10 -0.109   -0.0580      100
## 2 Weekly_Sales Unemployment -0.0745 9.92e- 9 -0.0998  -0.0491      100
## 3 Weekly_Sales Temperature  -0.0439 7.38e- 4 -0.0692  -0.0184      100
## 4 Weekly_Sales Fuel_Price    0.0185 1.55e- 1 -0.00701  0.0439      100
```

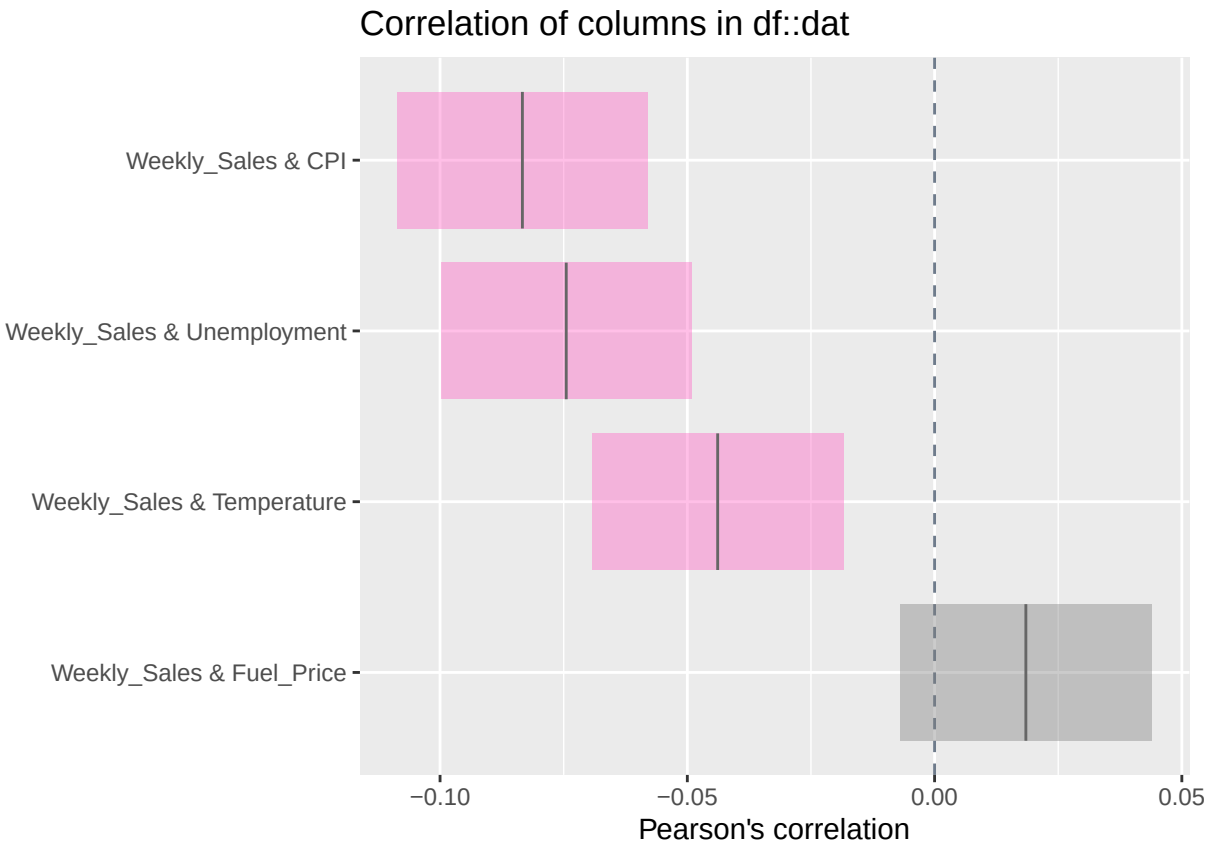

Fig 8: Plots showing correlation between dependent and numerical independent predictors.

Plot shows that weekly_sales has a significant negative correlation with Unemployment, CPI, Temperature while Fuel-price does not show any significant correlation.

Sales were averaged over week per year to check how Weekly Sales varies depending on holidays

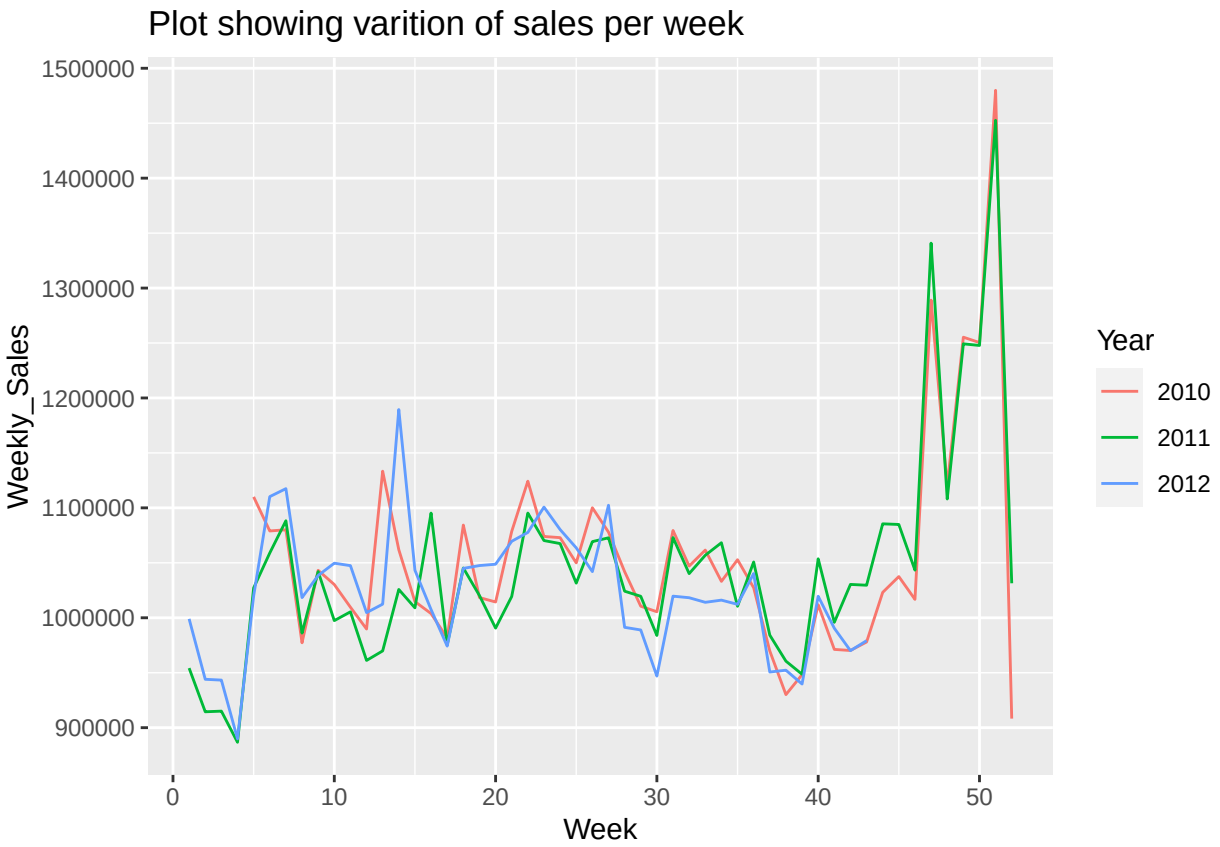

Fig 9: Variation of sales with week per year.

Figure 9 reveals that each year there is a hike in sales in holiday weeks specially between 47-52. Thus holiday flags impact sales positively.

In the later sections we will discuss about comparing different prediction models.

Fitting Multiple linear regression model

The most common Multiple regression model equation is :

$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + ... + \beta_k \times X_k + \varepsilon$

where k=number of regressors or predictors=5 for our model. p=number of parameters. p=k+1=6 n=total number of observations=total number of rows in the data=5954. ε = error terms or uncertainty in predictions.

This equation can be written in matrix form as:

$Y = \beta X + \varepsilon$

where,

$$Y = \begin{bmatrix} y1 \\ y2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

Observed outcomes.

and

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & . & . & .x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & . & . & .x_{2k} \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & . & . & .x_{nk} \end{bmatrix}$$

predictor matrix.

random error matrix is

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

and our coefficient matrix is

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix}$$

Coefficient matrix

Here noticeable part in matrix X is that first column is containing all 1s.The importance of this is that when we multiply our coefficient matrix β with predictor matrix X then first term of each equation formed will be β_0. Therefore the coefficient of X we multiply with β_0 is simply 1. So it is crucially important that we create a column of 1s as first column in model matrix X.

In order to get a best fit multiple linear regression model we must minimize the sum of squared error.

Here the sum of squared error L is:

$$L = \sum_{i=1}^{n} {\varepsilon_i}^2 = (y - X\beta)^{'}(y - X\beta)$$

The least square estimator $\hat{\beta}$ is the solution for β in the equation:

$\frac{\partial L}{\partial \beta} = 0.$

Here $\hat{\beta}$ is a vector that contains all our estimates for the parameters in our model.

The minimization of L in calculus leads to the normalization equation: $X^{'}X\hat{\beta} = X^{'}Y$

Solution of this normal equation leads to solution for $\hat{\beta}$ as:

$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$

$\hat{\beta}$ estimates for multiple linear model are:

Coefficients	Estimates
$Intercept(\beta_0)$	1606805.50303
β_1	44065.48933

Coefficients	Estimates
β_2	715.73962
β_3	-2613.74345
β_4	66.54202
β_5	-39945.86093

Table 1: Coefficient estimates.

For evaluating significance and performance of the models statistical properties like F-statistics, R^2, standard error associated would be calculated.

residuals $e = (y - \hat{y})$

Squared Standard error(SSE)

$$\hat{\sigma}_2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

Root Mean Square Error(RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

and

Mean Absolute Error(MAE)

$$MAE = \frac{\sum_{i=1}^{n} |(Y_i - \hat{Y}_i)|}{n} = \frac{\sum e_i}{n}$$

```
##                SSE   RMSE      MAE
## [1,] 1.092385e+27 655051 561616.5
```

Now our regression sum of squares

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

and the total sum of squares is

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Now we will run a F-statistics test for obtaining significance of our overall model.

F-statistics is $\frac{\frac{SSR}{k}}{SSE}$

```
F0<-(SSR)/(SSE)
pf(F0,k,n-p)
```

```
## [1] 1.817886e-42
```

Here the p-value from f-statistics test indicates that model is statistically significant. Now to evaluate model performance we will calculate R^2 value by using the formula. $R^2 = 1 - \frac{SSR}{SST} = 0.0411$ and standard error=$\sqrt{(\frac{SST}{n})} = 553772$

Now same model parameters will be estimated using random forest regression method.

Random forest

```
library(randomForest)
rf<-randomForest(Weekly_Sales~.,dat[,3:8],ntree = 500)
print(rf)
```

```
##
## Call:
##  randomForest(formula = Weekly_Sales ~ ., data = dat[, 3:8], ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 266135636080
##                     % Var explained: 13.22
```

By applying the above formulas the statictical summery table(RMSE, MAE, R^2 and p-value) for Random Forest Model is:

RMSE	MAE	R^2	p-value
472281.1	400426.8	0.1355375	2.095769e-22

Results and Discussion:

Sales per week for different are influenced by several factors, as found in this study there are sales hike during holidays. Unemployment and temperature affect sales negatively whereas CPI influence is positive.

The model comparision summarization table is:

Model Name	RMSE	MAE	R^2	p-value
MLR	655051	561616.5	0.04105873	1.1478e-30
RF	472281.1	400426.8	0.1355375	7.70231e-22

Table 2: Compression table of MLR and RF.

Higher RMSE and MAE for MLR model indicate better performance for RF algorithm. For MLR model a R^2 value of 0.041 indicates that the model can only explain 4% variation in the dependent variable which is not reliable to explain our dependent variable. Similarly for RF R^2 value of 0.14 indicates that 13% variation of dependent variable can be explained by the model which is of course better than MLR. BOth the models are statistically significant as the p-value is much lower than typical 0.05.

Conclusion:

Sales increases during holidays whereas on hot days sales tend to decrease, also if in the area of the store, people unemployment rate is higher sales are likely to go down. Among fitted models LR performed better than MLR with higher R^2 and lower RMSE and MAE.

Data and software availability:

Here Walmart retail data (https://www.kaggle.com/datasets/aditya6196/retail-analysis-with-walmart-data?resource=download) was used for the analysis.

and all the analyses were conducted using R-Studio. Link to download the software is (https://cran.r-project.org/bin/windows/base/R-4.2.2-win.exe).

References

1) Huan Zhang, Pengbao Wu, Aijing Yin, Xiaohui Yang, Ming Zhang, Chao Gao, 2017. "Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model, Science of The Total Environment", Volume 592,2017,Pages 704-713,ISSN 0048-9697,https://doi.org/10.1016/j.scitotenv.2017.02.146.

2) Xuefeng Xie, Tao Wu, Ming Zhu, Guojun Jiang, Yan Xu, Xiaohan Wang, Lijie Pu, 2021. "Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land, Ecological Indicators", Volume 120,2021,106925,ISSN 1470-160X,https://doi.org/10.1016/j.ecolind.2020.106925.

3) Huang, Li-Ying, Fang-Yu Chen, Mao-Jhen Jhou, Chun-Heng Kuo, Chung-Ze Wu, Chieh-Hua Lu, Yen-Lin Chen, Dee Pei, Yu-Fang Cheng, and Chi-Jie Lu. 2022. "Comparing Multiple Linear Regression and Machine Learning in Predicting Diabetic Urine Albumin–Creatinine Ratio in a 4-Year Follow-Up Study" Journal of Clinical Medicine 11, no. 13: 3661. https://doi.org/10.3390/jcm11133661

4) Phan Thanh Noi, ORCID,Jan Degener, Martin Kappas. 2017 "Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data" Remote Sens. 2017, 9(5), 398; https://doi.org/10.3390/rs9050398