Future Sales Prediction

Harish Kumar Pallimoni

20/06/2022

## Summary

Companies in the traditional days has been operating in production without considering the number of sales and demand for the products they produce. Therefore, due to increased competition among businesses, there has been the need to have the most effective and accurate technique that can help in decision-making for businesses in terms of the number of goods that need to be produced depending on sales and demand. Therefore, this brings in the need to have a sales prediction model that will be used for this case. With sales prediction, it is helpful to the companies as it provides them with a method where they can gain valuable insight from their sales and predict future sales. There will be proper stock management and production planning for businesses that decide to conduct sales predictions. Therefore, this paper has implemented machine learning techniques multilinear regression model that can be used for future sales predictions. The project utilizes a clean\_sales\_data retrieved from the Kaggle.com website for an online store named Walmart stores. From the results of this project, it is clear that the created multilinear regression machine learning algorithm is suggested to be the most suitable model that can be used in future sales predictions. However, there needs to have more research and implementation of random forest and decision tree models as recommended in this paper for sales prediction.

## Introduction

This project’s main focus is to implement a reliable and efficient sales prediction mechanism through data mining techniques to ensure businesses achieve the highest possible profits and have a competitive advantage over their competitors. The project utilizes data from an eCommerce business as this is the most competitive industry that needs prediction models with the highest possible reliability and accuracy for future sales predictions. Future sales prediction for many businesses is an essential aspect of planning and decision-making. Future sales prediction allows businesses to plan their business strategies effectively. With the sales prediction, businesses can have insight into how they manage their resources, workforce, and cash flow.

To achieve its purpose, the paper is divided into different sections. The first section is the introduction. The next section, the literature review, states the review of various literature associated with future sales predictions. The theory section provides some hypotheses for the project. For the data section, the data cleaning process is highlighted. The fourth section is the methodology. The fifth section includes results obtained for the project. The next section is the implications section, where the recommendations have been provided. Lastly is the conclusion section.

## Literature Review

There is a constant search for a better model in the sales predictions for many businesses to ensure that they have higher profits and remain competitive in the market. Businesses face several challenges in searching for the best and most accurate technique for predicting sales. This challenge comes in due to the rapid growth of huge volumes of data that eCommerce uses in their transaction. Therefore, in their comparative study, Shrivatava & Arya (2012) have explained the different clustering techniques for sales data. In the study by Rajagopal (2011), their analysis illustrated that in the decision-making process, it is crucial to classify data. In their paper, Mann & Kaur (2013) indicated that any data size could be transformed into a suitable format with the appropriate choice of a data mining technique. The authors also identified that effective decision-making could be done when there is an appropriate sales prediction technique. As identified in the reviewed literature, it is clear that there needs to have the most accurate and efficient sales prediction technique for businesses that can handle any size of data. Therefore, the need to perform this project is identified to help develop a multilinear regression model, a machine learning algorithm that can be used for sales predictions.

## Theory

H1: Holidays have an impact on total sales for particular stores.

## Data

The data used for this project is for the Walmart stores that contain sales transactions between 2010 and 2012. This clean\_sales\_data has been retrieved from

##

## Attaching package: 'scales'

## The following objects are masked from 'package:formattable':

##

## comma, percent, scientific

##

## Attaching package: 'zoo'

## The following objects are masked from 'package:base':

##

## as.Date, as.Date.numeric

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ tibble 3.1.7 ✔ stringr 1.4.0

## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ✔ purrr 0.3.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──

## ✖ readr::col\_factor() masks scales::col\_factor()

## ✖ purrr::discard() masks scales::discard()

## ✖ dplyr::filter() masks stats::filter()

## ✖ dplyr::lag() masks stats::lag()

##

## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':

##

## date, intersect, setdiff, union

## Loading required package: carData

##

## Attaching package: 'car'

## The following object is masked from 'package:purrr':

##

## some

## The following object is masked from 'package:dplyr':

##

## recode

## corrplot 0.92 loaded

##

## Attaching package: 'MLmetrics'

## The following object is masked from 'package:base':

##

## Recall

## Store Date Weekly\_Sales Holiday\_Flag Temperature Fuel\_Price CPI

## 1 1 05-02-2010 1643691 0 42.31 2.572 211.0964

## 2 1 12-02-2010 1641957 1 38.51 2.548 211.2422

## 3 1 19-02-2010 1611968 0 39.93 2.514 211.2891

## 4 1 26-02-2010 1409728 0 46.63 2.561 211.3196

## 5 1 05-03-2010 1554807 0 46.50 2.625 211.3501

## 6 1 12-03-2010 1439542 0 57.79 2.667 211.3806

## Unemployment

## 1 8.106

## 2 8.106

## 3 8.106

## 4 8.106

## 5 8.106

## 6 8.106

Check for missing values in the column values.

colSums(is.na(salesdata))

## Store Date Weekly\_Sales Holiday\_Flag Temperature Fuel\_Price

## 0 0 0 0 0 0

## CPI Unemployment

## 0 0

There are no missing values. Therefore, next we check for any duplicate values in the clean\_sales\_data.

#Check anyDuplicate Values

all(duplicated(salesdata) == TRUE)

## [1] FALSE

The clean\_sales\_data has no duplicate values identified.

## Methodology

The clean\_sales\_data sales data is clean; therefore, some analysis can be performed. First, we analyze to identify the stores with the most sales.

# Combine data by 'Store' and Find of 'Weekly\_Sales.'

sales\_for\_stores<- aggregate(Weekly\_Sales ~ Store, data = sales data, sum)

#Change the column name of sales

colnames(sales\_for\_stores)[2] <- "Total\_Sales\_by\_Store"

#Find the Store that has the highest sales

#Sort the Stores by Sales in descending order

sales\_for\_stores <-arrange(sales\_for\_stores, desc(Total\_Sales\_by\_Store))

#Choosing the first store that comes in this order

sales\_for\_stores[1,]

## Store Total\_Sales\_by\_Store

## 1 20 301397792

Perform an analysis on the impact of holidays on sales.

Check the difference in sales between hoidays and non-holidays.

#Filter the holiday dates and find mean of Weekly Sales

Dates\_Holiday <- filter(salesdata2,Holiday\_Flag ==1)

Dates\_Holiday\_Sales<-summarise(group\_by(Dates\_Holiday,Date),mean(Weekly\_Sales))

#Caluclating mean of Weekly Sales for non holidays

mean\_non\_holiday\_sales <- mean(filter(salesdata2,Holiday\_Flag ==0)$Weekly\_Sales)

Dates\_Holiday\_Sales$higher\_than\_non\_holiday <- Dates\_Holiday\_Sales[,2] > mean\_non\_holiday\_sales

Next involved creating a multilinear regression model to provide the sales predictions for Walmart stores. First, we create a data frame with the columns essential to create the model.

#make a copy of the original data for alterations

salesdata3 <- sales data

#select only first store.

salesdata3<- dplyr::filter(salesdata3, Store ==1)

#changing date column to date format

salesdata3$Date <- lubridate::dmy(salesdata3$Date)

salesdata3 <- dplyr::arrange(salesdata3,Date)

#Create a week number,month,quarter column in dataframe

salesdata3$Week\_Number <- seq(1:length(unique(salesdata3$Date)))

#add quarter and month columns

salesdata3$month <- lubridate::month(salesdata3$Date)

salesdata3$quarter <- lubridate::quarter(salesdata3$Date)

##Create a holiday event type data frame

Dates\_Holiday <- c("12-02-2010", "11-02-2011", "10-02-2012", "08-02-2013","10-09-2010", "09-09-2011", "07-09-2012", "06-09-2013","26-11-2010", "25-11-2011", "23-11-2012", "29-11-2013","31-12-2010", "30-12-2011", "28-12-2012", "27-12-2013")

#assigning date format to Dates\_Holiday vector

Dates\_Holiday <- lubridate::dmy(Dates\_Holiday)

#Creating holiday\_event vector

holiday\_event <-c(rep("Super Bowl", 4), rep("Labour Day", 4),rep("Thanksgiving", 4), rep("Christmas", 4))

#Creating data frame with holiday\_event and date

Holidays\_Data <- data.frame(holiday\_event,Dates\_Holiday)

#merging both dataframes

salesdata3<-merge(salesdata3,Holidays\_Data, by.x= "Date", by.y="Dates\_Holiday", all.x = TRUE)

#Replacing null values in Event with No\_Holiday

salesdata3$holiday\_event = as.character(salesdata3$holiday\_event)

salesdata3$holiday\_event[is.na(salesdata3$holiday\_event)]= "No\_Holiday"

Remove unnecessary columns from the clean\_sales\_data and change structure of the holiday\_event column.

salesdata3$Date <-NULL

salesdata3$Store <- NULL

salesdata3$holiday\_event <- as.factor(salesdata3$holiday\_event)

str(salesdata3)

## 'data.frame': 143 obs. of 10 variables:

## $ Weekly\_Sales : num 1643691 1641957 1611968 1409728 1554807 ...

## $ Holiday\_Flag : int 0 1 0 0 0 0 0 0 0 0 ...

## $ 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 ...

## $ CPI : num 211 211 211 211 211 ...

## $ Unemployment : num 8.11 8.11 8.11 8.11 8.11 ...

## $ Week\_Number : int 1 2 3 4 5 6 7 8 9 10 ...

## $ month : num 2 2 2 2 3 3 3 3 4 4 ...

## $ quarter : int 1 1 1 1 1 1 1 1 2 2 ...

## $ holiday\_event: Factor w/ 5 levels "Christmas","Labour Day",..: 3 4 3 3 3 3 3 3 3 3 ...

salesdata3$Holiday\_Flag <- as.numeric(salesdata3$Holiday\_Flag)

salesdata3$Week\_Number <- as.numeric(salesdata3$Week\_Number)

salesdata3$quarter <- as.numeric(salesdata3$quarter)

set.seed(123)

library(caTools)

clean\_sales\_data <- salesdata3

#Create a sample split and split testing & training sets in 20-80 ratio respectively

sample = sample.split(clean\_sales\_data, SplitRatio = 0.8)

# Return a vector with T for 80% of data

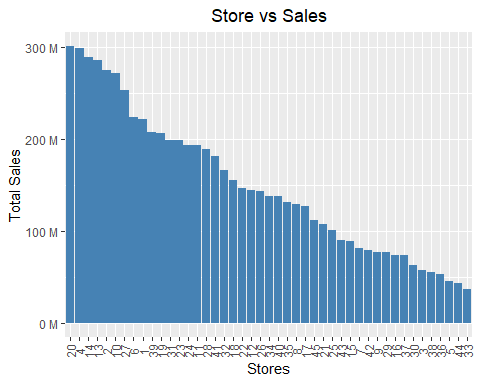
trainingSet = subset(clean\_sales\_data, sample == T)

testingSet = subset(clean\_sales\_data, sample == F)

# Create model

model = lm(formula = Weekly\_Sales ~ . , data = trainingSet)

## Results

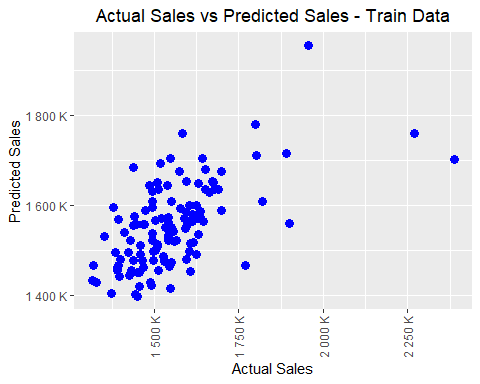
Various results have been identified for this project. First, the analysis for identifying the highest selling store illustrates that store 20 has the highest number of sales. The analysis for sales during holidays and non-holidays also gave some interesting results. First, it was identified that there were higher sales for the Thanksgiving, Labour Day, and Super Bowl holidays than mean sales for Non-Holiday.

view(Dates\_Holiday\_Sales)

The model provided predictions on the sales, as illustrated in the scatterplot below.

## Warning in predict.lm(model, new data = trainingSet): prediction from a rank-

## deficient fit may be misleading



## Implications

The project in this paper can be improvised in various ways. First, further research in the model would consider all store data for sales prediction since, in this project, we only considered the first store. In addition, further work on this project should consider using more advanced machine learning models like Random Forest and Decision Tree. For proper data sampling, I would recommend using K cross-validation techniques for future work.

**Conclusion**

The paper included the analysis of sales from Walmart stores, an eCommerce business. In addition, using the Multilinear regression model, predictions on sales were performed. From the project, it is clear that there is a positive impact on the sales by holidays. Therefore, the theory that holidays impact total sales for particular stores is true.

## References

Rajagopal, D. (2011). Customer data clustering using data mining technique. arXiv preprint arXiv:1112.2663.

Shrivastava, V., & Arya, N. (2012). A study of various clustering algorithms on retail sales data. Int. J. Comput. Commun. Netw, 1(2).