Predicting Weekly Sales at WalMart Stores

August 31, 2015

- 1. Executive Summary
- 2. Introduction
    - 2.1 About the Solution Environment
    - 2.2 About the Data
        - 2.2.1 The Challenge
    - 2.3 Getting the Data
        - 2.3.1 The Data Files
        - 2.3.2 Ingesting the Data
    - 2.4 Libraries Used
- 3. Stage 1: Data Exploration and Preparation
    - 3.1 Summary Statistics
        - 3.1.1 The Training Dataset (train)
        - 3.1.2 The Stores Dataset (stores)
        - 3.1.3 The Features Dataset (features)
        - 3.1.4 The Test Dataset (test)
    - 3.2 Data Preparation - Merging the Datasets
        - 3.2.1 Merging Train and Stores Datasets
        - 3.2.2 Merging Train, Stores and Features Datasets
        - 3.2.3 Merging Test, Stores and Features Datasets
    - 3.3 Data Exploration
        - 3.3.1 Total Sales Vs. Store Size
        - 3.3.2 Store Sales - Time Series
- 4. Stage 2: Formal Statistical Inferences
    - 4.1 Do Holiday Weeks Account for Higher Sales?
    - 4.2 Are Sales
- 5. Stage 3: Linear Regression: Predicting Weekly_Sales
    - 5.1 Predicting Store Weekly_Sales
    - 5.2 Predicting Store-Department Weekly_Sales

1. Executive Summary

Retail stores need to be able to predict sales forecasts for the future and study the effect how strategic offers affect sales, especially during holiday season. Since the number of days in holidays are limited, it becomes more challenging to be able to accruately predict how different aspects affect sales.

The report will attempt to create a predictive model for WalMart a store’s Weekly Sales and store department-wise Weekly Sales.

2. Introduction

2.1 About the Solution Environment

The authors implemented this solution in R. We have used R Markdown Report to create this document. First we explore and prepare the data set before carrying out formal statistical inferences on the dataset. We wrap the report by building a model to predict Weekly sales for the 45 stores in this dataset.

2.2 About the Data

The dataset under consideration is taken from a recruitment competition WalMart ran on Kaggle between February-May 2014. The participants were supposed to create a model to be able to predict Weekly Sales for 45 Stores located in different regions. Each store has multiple departments and the end requirement is to be able to predict the sales for individual departments of each store.

2.2.1 The Challenge

The challenge is to be able to predict how different holiday price markdowns affect the various departments in the store, to model extent of impact of these markdowns.

2.3 Getting the Data

The data was download from Kaggle.

URL to the Kaggle Competition Site: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting (https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting)

The files available are the following:

Data Files

File Name	Available Formats
features.csv	.zip (157.91 kb)
sampleSubmission.csv	.zip (220.25 kb)
stores	.csv (532 b)
test.csv	.zip (235.29 kb)
train.csv	.zip (2.47 mb)

2.3.1 The Data Files

Here we discuss the various CSV Files that are given by WalMart.

2.3.1.1 stores.csv

Contains size and type of 45 stores (45 records).

2.3.1.2 train.csv

Weekly sales dataset from Februray 05, 2010 to November 11, 2012. It contains the following fields:

- Store: store number
- Dept: the department number
- Date: week date
- Weekly_Sales: sales for the given department in the given store
- IsHoliday: whether the week is a special holiday week

2.3.1.3 test.csv

The dataset with similar fields as train.csv, except without Weekly_Sales. This will be used to test the model with unseen data and can be evaulated by uploading the dataset to Kaggle.

2.3.1.4 features.csv

This data file contains additional relevant information relating to the physical and business environment around the store. The fields are as follows:

- Store: store number
- Date: the week date
- Temperature: the average temperature in the region
- Fuel_Price: cost of fuel in the region
- MarkDown1-5: data related to the markdowns that Walmart is running. Markdown data is only available after November 2011 and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI - the Consumer Price Index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

The four holidays fall inthe following weeks in the dataset:

- Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
- Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
- Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
- Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

2.3.2 Ingesting the Data

```
## Ingesting the data from the Data folder
train <- read.csv("Data/train.csv")
stores <- read.csv("Data/stores.csv")
features <- read.csv("Data/features.csv")
test <- read.csv("Data/test.csv")
```

2.4 Libraries Used

The following libraries are used in this report:

```
# Grammar of Graphics Plotting Library
library(ggplot2)
```

3. Stage 1: Data Exploration and Preparation

3.1 Summary Statistics

3.1.1 The Training Dataset (train)

```
str(train)
```

```
## 'data.frame':    421570 obs. of  5 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Dept        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Factor w/ 143 levels "2010-02-05","2010-02-12",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ Weekly_Sales: num  24924 46039 41596 19404 21828 ...
##  $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

`Date` is ingested as factor (as opposed to being ingested as date type). There are 143 dates in total.

```
## Changing the Date from "Format" type to "Date" Type
train$Date <- as.Date(train$Date)
## Getting the summary of the Data
summary(train)
```

```
##      Store           Dept            Date              Weekly_Sales
##  Min.   : 1.0   Min.   : 1.00   Min.   :2010-02-05   Min.   : -4989
##  1st Qu.:11.0   1st Qu.:18.00   1st Qu.:2010-10-08   1st Qu.:  2080
##  Median :22.0   Median :37.00   Median :2011-06-17   Median :  7612
##  Mean   :22.2   Mean   :44.26   Mean   :2011-06-18   Mean   : 15981
##  3rd Qu.:33.0   3rd Qu.:74.00   3rd Qu.:2012-02-24   3rd Qu.: 20206
##  Max.   :45.0   Max.   :99.00   Max.   :2012-10-26   Max.   :693099
##  IsHoliday
##  Mode :logical
##  FALSE:391909
##  TRUE :29661
##  NA's :0
##
##
```

There is no missing data in the dataset.

As discussed in the Introduction, this report contains data of 45 stores - represented by Store. There are a total of 99 stores in all.

The starting date for training dataset is `2010-02-05`. It starts on a `Friday`. The last date recorded in the dataset is `2012-10-26`, which is also a `Friday`. There are `994` days between them - so the data consists of a total of `143` weeks of data.

It is interesting to note that for some departments the `Weekly_Sales` are negative. Returns and special offers cause these negative sales figures.

There are no missing values in this dataset.

3.1.2 The Stores Dataset (stores)

```
## Structure of Stores Dataset
str(stores)
```

```
## 'data.frame':    45 obs. of  3 variables:
##  $ Store: int  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 1
26512 ...
```

```
## summary Statistics of Stores dataset
summary(stores)
```

```
##      Store    Type        Size
##  Min.   : 1   A:22   Min.   : 34875
##  1st Qu.:12   B:17   1st Qu.: 70713
##  Median :23   C: 6   Median :126512
##  Mean   :23          Mean   :130288
##  3rd Qu.:34          3rd Qu.:202307
##  Max.   :45          Max.   :219622
```

No missing data.

3.1.3 The Features Dataset (features)

```
## Structure of features dataset
str(features)
```

```
## 'data.frame':    8190 obs. of  12 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Factor w/ 182 levels "2010-02-05","2010-02-12",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ 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 ...
```

`Date` is ingested as factor (as opposed to being ingested as date type). There are 182 dates in total. This dataset is relevant for both the `train` and the `test` dataset.

```
## Changing the Date from "Format" type to "Date" Type
features$Date <- as.Date(features$Date)
## Summary Statistics of Features Dataset
summary(features)
```

```
##      Store          Date              Temperature       Fuel_Price
##  Min.   : 1   Min.   :2010-02-05   Min.   : -7.29   Min.   :2.472
##  1st Qu.:12   1st Qu.:2010-12-17   1st Qu.: 45.90   1st Qu.:3.041
##  Median :23   Median :2011-10-31   Median : 60.71   Median :3.513
##  Mean   :23   Mean   :2011-10-31   Mean   : 59.36   Mean   :3.406
##  3rd Qu.:34   3rd Qu.:2012-09-14   3rd Qu.: 73.88   3rd Qu.:3.743
##  Max.   :45   Max.   :2013-07-26   Max.   :101.95   Max.   :4.468
##
##    MarkDown1         MarkDown2            MarkDown3           MarkDown4
##  Min.   : -2781   Min.   :  -265.76   Min.   :  -179.26   Min.   :    0.22
##  1st Qu.:  1578   1st Qu.:    68.88   1st Qu.:     6.60   1st Qu.:  304.69
##  Median :  4744   Median :   364.57   Median :    36.26   Median : 1176.42
##  Mean   :  7032   Mean   :  3384.18   Mean   :  1760.10   Mean   : 3292.94
##  3rd Qu.:  8923   3rd Qu.:  2153.35   3rd Qu.:   163.15   3rd Qu.: 3310.01
##  Max.   :103185   Max.   :104519.54   Max.   :149483.31   Max.   :67474.85
##  NA's   :4158     NA's   :5269        NA's   :4577        NA's   :4726
##    MarkDown5             CPI         Unemployment    IsHoliday
##  Min.   :  -185.2   Min.   :126.1   Min.   : 3.684   Mode :logical
##  1st Qu.:  1440.8   1st Qu.:132.4   1st Qu.: 6.634   FALSE:7605
##  Median :  2727.1   Median :182.8   Median : 7.806   TRUE :585
##  Mean   :  4132.2   Mean   :172.5   Mean   : 7.827   NA's :0
##  3rd Qu.:  4832.6   3rd Qu.:213.9   3rd Qu.: 8.567
##  Max.   :771448.1   Max.   :229.0   Max.   :14.313
##  NA's   :4140       NA's   :585     NA's   :585
```

The `features` dataset has missing variables for `Markdown1-5`, `CPI` & `Unemployment`.

3.1.4 The Test Dataset (test)

```
## Structure of test dataset
str(test)
```

```
## 'data.frame':    115064 obs. of  4 variables:
##  $ Store    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Dept     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date     : Factor w/ 39 levels "2012-11-02","2012-11-09",..: 1 2 3 4 5 6
7 8 9 10 ...
##  $ IsHoliday: logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
```

`Date` is ingested as factor (as opposed to being ingested as date type). There are 39 dates in total.

```
## Changing the Date from "Format" type to "Date" Type
test$Date <- as.Date(test$Date)
## Summary Statistics of test Dataset
summary(test)
```

```
##      Store            Dept            Date            IsHoliday
##  Min.   : 1.00   Min.   : 1.00   Min.   :2012-11-02   Mode :logical
##  1st Qu.:11.00   1st Qu.:18.00   1st Qu.:2013-01-04   FALSE:106136
##  Median :22.00   Median :37.00   Median :2013-03-15   TRUE :8928
##  Mean   :22.24   Mean   :44.34   Mean   :2013-03-14   NA's :0
##  3rd Qu.:33.00   3rd Qu.:74.00   3rd Qu.:2013-05-24
##  Max.   :45.00   Max.   :99.00   Max.   :2013-07-26
```

3.2 Data Preparation - Merging the Datasets

3.2.1 Merging Train and Stores Datasets

Since the `Type` & `Size` variables may influence the Weekly Sales, we are merging the `train` & datasets. We merge the data by `Store` .

```
## Merging train and stores by Store
trainStoresMerge <- merge(train , stores , by = "Store")
```

3.2.2 Merging Train, Stores and Features Datasets

Since `Markdown1-5` and other variables could play an important role at predicting `Weekly_Sales` , this should be merged with the `trainStoresMerge` dataset. We merge the data by `Store` & `Date` .

```
## Merging trainStoresMerge and features datasets
trainStoresFeaturesMerge <- merge( trainStoresMerge , features , by = c( "Stor
e" , "Date" ) )
## Clearing memory - removing intermediate datasets
rm(trainStoresMerge , train)
## Fixing the name of the Column
colnames(trainStoresFeaturesMerge)[5] <- "IsHoliday"
trainStoresFeaturesMerge$IsHoliday.y <- NULL
```

3.2.3 Merging Test, Stores and Features Datasets

We similarily merge the `test` , `stores` & `features` to create the `testStoresFeaturesMerge` dataset.

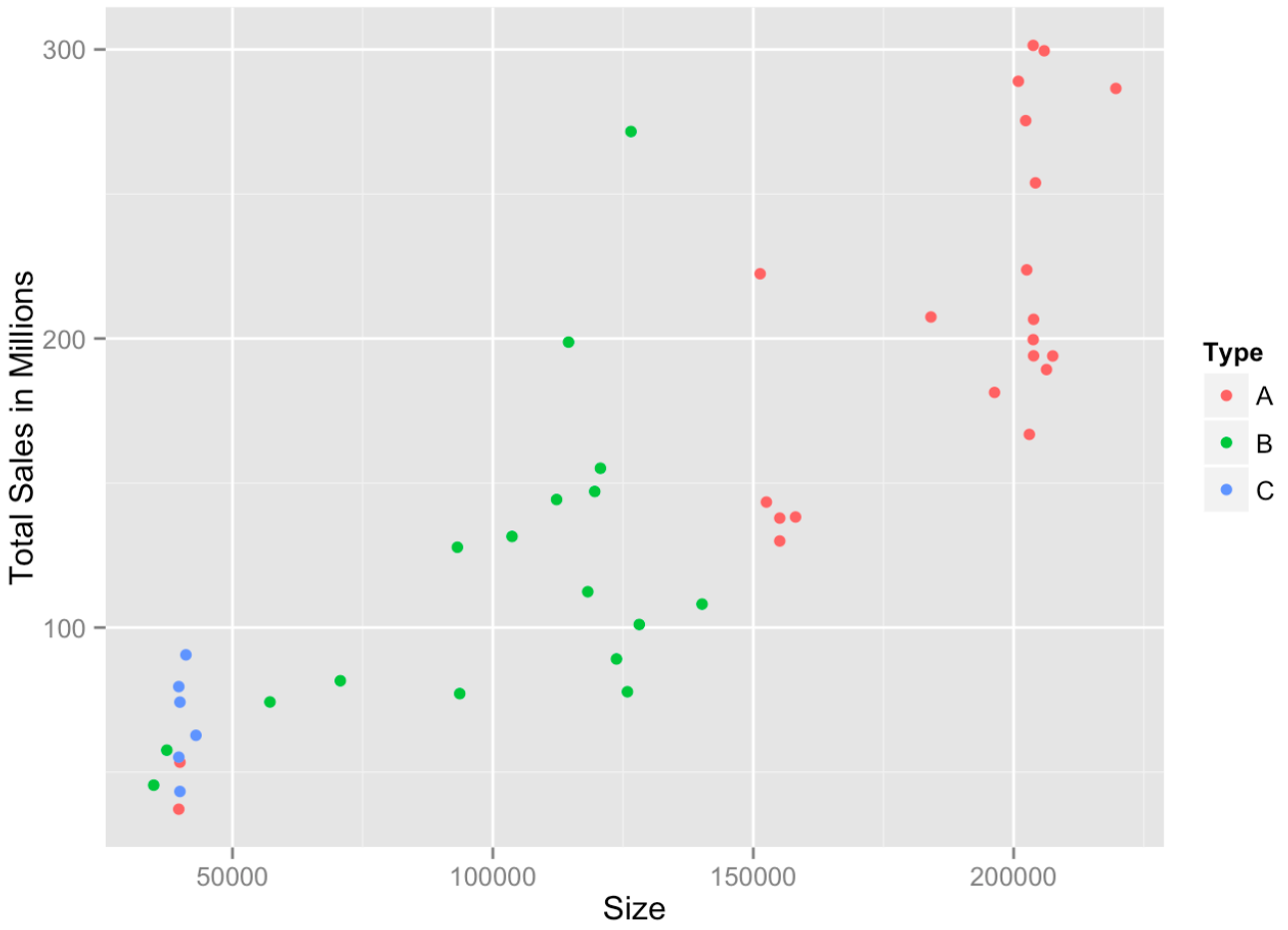
```
## Merging test and stores by Store
testStoresMerge <- merge(test , stores , by = "Store")
## Merging testStoresMerge and features datasets
testStoresFeaturesMerge <- merge( testStoresMerge , features , by = c( "Store"
, "Date" ) )
## Clearing Memory - removing intermediate Datasets
rm( test , testStoresMerge , features )
## Fixing the name of the Column
colnames(testStoresFeaturesMerge)[5] <- "IsHoliday"
testStoresFeaturesMerge$IsHoliday.y <- NULL
```

3.3 Data Exploration

3.3.1 Total Sales Vs. Store Size

Plotting the total sales of a store vs. Store Size. We first calculate the total sales per Store and plot it as a response (y-axis) to the Store size (x-axis) to understand the relationship between them.

```
## Total Sales vs. Store Size - plotting the relationship
## calculating the sum of all the store sales
StoreTotalSales <- tapply(trainStoresFeaturesMerge$Weekly_Sales, trainStoresFea
turesMerge$Store, FUN = sum)
## converting the table to a DataFrame
stores$TotalSales <- StoreTotalSales
stores$TotalSalesInMillion <- stores$TotalSales/1000000
## Plotting the Total Sales vs. Store Size
ggplot( stores , aes(x=Size , y=TotalSalesInMillion , color = Type ) ) +
  geom_point() +
  scale_y_continuous(name="Total Sales in Millions" )
```



This plot indicates that there is a postive relationship between the size of the store and total sales. Also Type ‘A’ Stores are mostly larger stores with bigger sales and Type ‘C’ Stores are small with lower sales.

3.3.2 Store Sales - Time Series

4. Stage 2: Formal Statistical Inferences

4.1 Do Holiday Weeks Account for Higher Sales?

4.2 Are Sales

5. Stage 3: Linear Regression: Predicting Weekly_Sales

5.1 Predicting Store Weekly_Sales

5.2 Predicting Store-Department Weekly_Sales