Predicting Weekly Sales at WalMart Stores

August 31, 2015

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 stores department-wise Weekly Sales.

Introduction

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.

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.

The Challenge

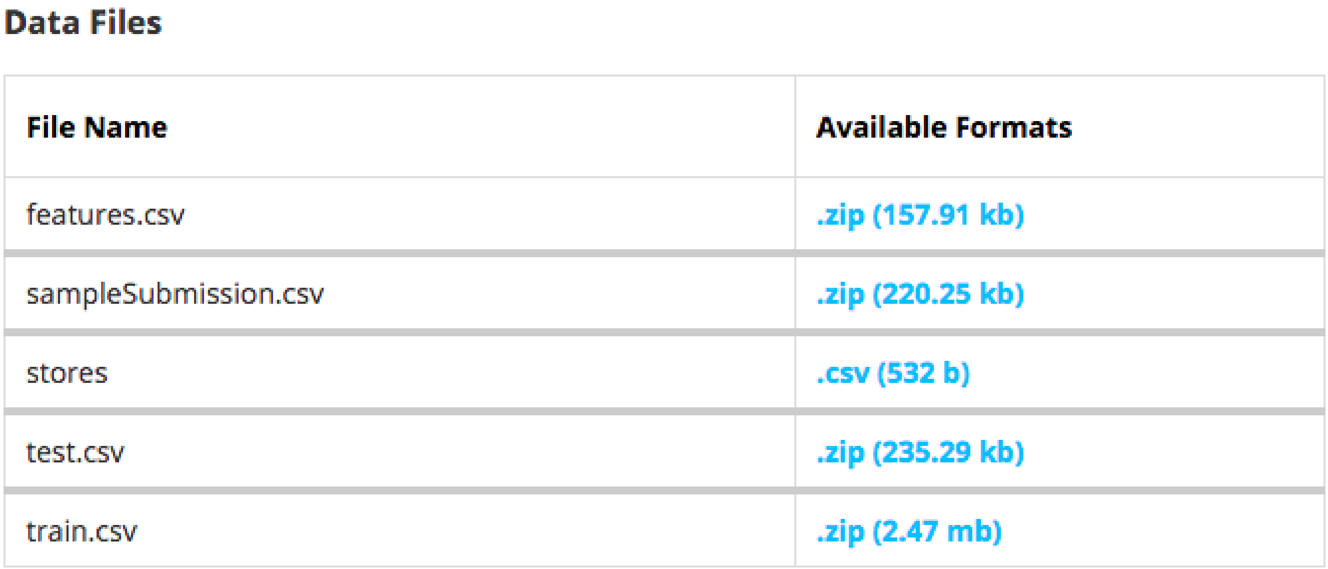
The challenge is to be able to predict how different holiday price markdowns affect the various departments in the store, to model extent of impact of these markdowns.

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)

The Data Files

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

stores.csv

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

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

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.

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

Ingesting the Data

```r
## 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")
```

Data Exploration and Preparation

Inital Exploration

Summarizing The Training Dataset (train)

```r
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 ...
```

We see the structure of the training dataset. Dates are ingested as factors (as opposed to being ingested as dates). This needs to be corrected to date format. we see that 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    IsHoli
day
##  Min.   : 1.0   Min.   : 1.00   Min.   :2010-02-05   Min.   : -4989   Mode
:logical
##  1st Qu.:11.0   1st Qu.:18.00   1st Qu.:2010-10-08   1st Qu.:  2080   FALS
E:391909
##  Median :22.0   Median :37.00   Median :2011-06-17   Median :  7612   TRUE
:29661
##  Mean   :22.2   Mean   :44.26   Mean   :2011-06-18   Mean   : 15981   NA's
:0
##  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
```

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.

There are ```