Retail Analysis with Walmart Data

Course-end Project 2

Description

One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

Dataset Description

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file Walmart_Store_sales. Within this file you will find the following fields:

Store - the store number
Date - the week of sales
Weekly_Sales - sales for the given store
Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
Temperature - Temperature on the day of sale
Fuel_Price - Cost of fuel in the region
CPI – Prevailing consumer price index
Unemployment - Prevailing unemployment rate

Holiday Events

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labour 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

Analysis Tasks

Basic Statistics tasks
Which store has maximum sales
Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation
Which store/s has good quarterly growth rate in Q3'2012
Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together
Provide a monthly and semester view of sales in units and give insights

Statistical Model

For Store 1 – Build prediction models to forecast demand

Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.
Change dates into days by creating new variable.
Select the model which gives best accuracy.

ANALYSIS SUMMARY

1. Maximum Sales and Standard Deviation:

- Store 20 has the maximum sales, with a value of $301,397,792.00.
- Store 14 has the maximum standard deviation, indicating high variability in sales.

2. Quarterly Growth Rate in Q3'2012:

- Stores 7, 16, 35, 26, 39, 41, 44, 24, 40, and 23 have positive growth rates in Q3'2012, ranging from 0.83% to 13.33%.
- Store 7 has the highest growth rate at 13.3%.

3. Impact of Holidays on Sales:

- Thanksgiving has higher sales than the mean sales in the non-holiday season for all stores combined. Thanksgiving sales are $1,471,273.43 compared to non-holiday sales of $1,041,256.38.

4. Monthly and Semester View of Sales:

- December has the highest weekly sales, and Semester 2 has the highest weekly sales overall.

5. Statistical Model:

- Linear Regression model was used to forecast demand, considering variables like date, CPI, unemployment, and fuel price.
- Random Forest Regression model provided better accuracy compared to Linear Regression, with an accuracy of almost 82%.

Overall, the analysis provides insights into sales trends, the impact of holidays, and the effectiveness of different prediction models for forecasting demand. It helps understand the factors influencing sales and suggests strategies for improving sales forecasting accuracy.