Yeshwantrao Chavan College of Engineering

(An Autonomous Institution affiliated to RashtrasantTukadoji Maharaj Nagpur University)
Hingna Road, Wanadongri, Nagpur - 441 110
Ph.: 07104-237919, 234623, 329249, 329250 Fax: 07104-232376, Website: www.ycce.edu

Department of Computer Technology

Session 2022-23(ODD)

7th Semester Section A & B

Mapped Course Outcome:

CO4	**Implement** various machine learning algorithms on a given dataset using modern toolS and write a report

Project Report on Retail Analysis with Walmart Data

Course Code:	**CT2426**	**Course Name:**	**Machine Learning**

Team Name:

Roll No of Team Member	**Team Member Names**
121	**Shubhalakshmee Warutkar**
122	**Snehal Nasare**
142	**Atharv Kevalram**
159	**Nikhil Tamrakar**

Introduction:

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 Walmart stores. It's very difficult to predict the demand of any retail store as there are certain events and holidays which impact sales each day. We have sales data available for 45 Walmart stores.

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.

Interpretation of data and relevance to real life problem:

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

- 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

Design methodology and Implementation:

1. **Import required libraries and dataset—**
   - Numpy
   - Pandas
   - Matplotlib
   - Seaborn
   - Sklearn
   - Warnings

2. **Changing the data type of the 'Date' column —**
   - We are changing the data type of the 'Date' column because it is an object type.
   - Here, the dataset does not have any null values.

3. **Statistical Tasks —**
   a. **Which store has maximum sales?**
   In order to find out the maximum sales, we can create a new variable called 'total_sales'. Then group by stores and find the sum of the weekly sales of each store. This will give us the maximum sales.

   b. **Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation.**

To find out the maximum standard deviation, create a new variable and then group it by stores and find the standard deviation.

c. **Which store/s has a good quarterly growth rate in Q3'2012?**
First, find the Q2 sales and then Q3 sales, take out the difference and then find the growth rate.

d. **Some holidays have a negative impact on sales. Find out holidays that have higher sales than the mean sales in the non-holiday season for all stores together.**
We have 4 Holiday Events, (1) Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13, (2) Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13, (3) Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13, (4) Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13. We can calculate the holiday event sales of each of the events and then find the non-holiday sales.

e. **Provide a monthly and semester view of sales in units and give insights.**
We can plot a month-wise bar graph for weekly sales to get an idea about which month has the maximum sales, and then we can plot a year-wise bar graph for weekly sales to know which year has the highest weekly sales.

Model Building

First, define dependent and independent variables.

Here, store, fuel price, CPI, unemployment, day, month, and year are the independent variables and weekly sales is the dependent variable.

Now, To train the model, Import train_test_spit from sklearn.model_selection and train 80% of the data and test on the rest 20% of the data.

We need to standardize the data because we want to bring down all the features to a common scale without distorting the differences in the range of the values.

We have used 4 different algorithms to know which model to use to predict the weekly sales.

1. Linear Regression
2. Random Forest Regressor
3. Decision Tree Regressor
4. KNearest Neighbors

Result and Analysis:

- We have done cross-validation, which is a method of assessing ML models that involves training numerous ML models on subsets of the available input data and evaluating them on the complementary subset of data. Cross-validation can be used to detect overfitting, or the failure to generalize a pattern.
- Here, we have used 4 different algorithms to know which model to use to predict the weekly sales. Linear Regression is not an appropriate model to use as accuracy is very low. However, Random Forest Regression gives an accuracy of almost 95%. so, it is the best model to forecast weekly sales.

Conclusion:

Here, we have used 4 different algorithms to know which model to use to predict the weekly sales. Linear Regression is not an appropriate model to use as accuracy is very low. However, Random Forest Regression gives accuracy of almost 95% . so, it is the best model to forecast weekly sales.

References:

https://www.kaggle.com/code/dhruvalpatel30/retail-analysis-with-walmart-data/notebook

https://github.com/ZarahShibli/Retail-Analysis-with-Walmart-Data/blob/main/Retail%20Analysis%20with%20Walmart%20Data.ipynb