Walmart Store Sales Prediction

Authors: Yash Pasar[1], Ritika Shetty[1], Shweta Sathisan[1]
School of Information Studies, Syracuse University, Syracuse, New York, USA - 13210
{yspasar, rshett02, ssathisa}@syr.edu

Abstract

A challenge of modeling retail data is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line. We leveraged machine learning models to predict weekly sales of different Walmart stores. These predictions are made taking into consideration Markdowns throughout the year

Introduction

Predicting sales for a store is very important to estimate the quantity for each product, to avoid overstocking or understocking. Our aim is to apply Machine learning algorithms on Walmart's historical sales data for 45 stores present, to predict department wide sales for each data.

Key Terminology

Markdown - Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas.
The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks for accurate predictions. The four holidays fall within the 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

Dataset Description

The dataset that we have consists of historical data of 45 Walmart stores. The historical training data covers a period from 2010-02-05 to 2012-11-01. All available attributes in the dataset with their description for deeper understanding.

Attribute Name	Description
Store	the store number
Dept	the department number
Date	the week
Weekly_Sales	sales for the given department in the given store
IsHoliday	whether the week is a special holiday week
Temperature	average temperature in the region
Fuel_price	cost of fuel in the region
Markdown 1-5	anonymized data related to promotional markdowns that Walmart is running.
CPI	the consumer price index
Unemployment	the unemployment rate

Data Pre-processing and Feature Engineering

Initially, our data consisted of over 42k rows and 16 columns. After generating the correlation matrix and analyzing the similarity between independent variables, we dropped column such as unemployment, fuel prize, date, CPI, etc. We also recognized certain markdown values and weekly sales to have negative values, which were replaced by 0s. Post this we performed one hot encoding on the data.
The dependent variable data i.e. weekly sales was heavily skewed, due to outliers that accounted for about 2.97% of our data. We performed Quantile transformation by first dropping the negative values and then applying transformation, to convert the data to normal distribution.

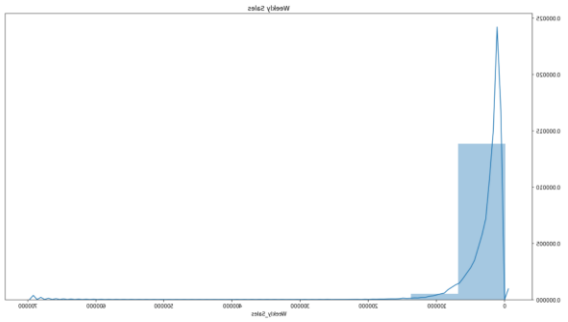


Fig 1.a. Target variable Data Distribution before transformation

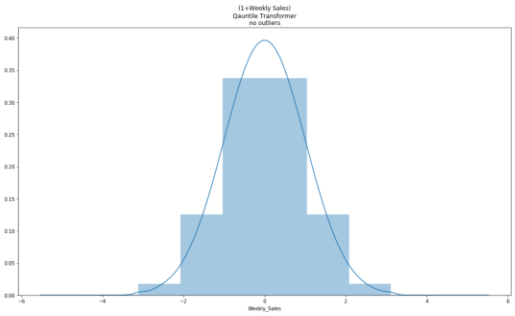


Fig 1.a. Target variable Data Distribution after Transformation

Exploratory Data Analysis

Through the exploratory data analysis conducted, we discovered Store A to be the biggest is size as well as sales that held about 48.9% of the overall sales of Walmart, followed by Store B and Store C with 37.8% and 13.3% of the market share.

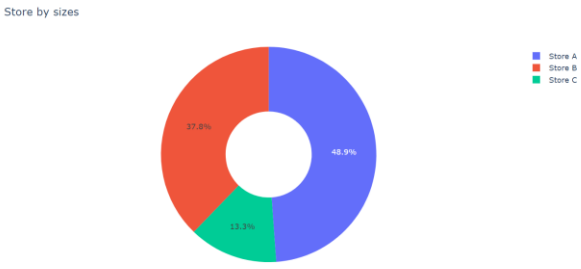


Fig 2.a. Donut chart representing the store categories by their sizes

Fig 2.b. Box Plots representing store categories by their sales

We also recognized the peaks of sales that occurred seasonally during times of Markdown sales.

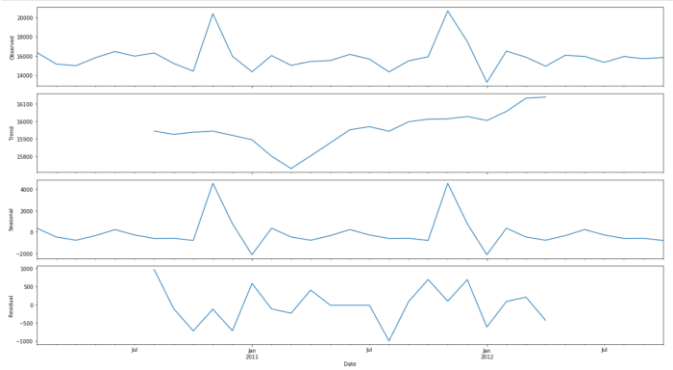


Fig 3. Seasonality of Sales

To determine and quantify the effect of holidays on the overall sales of Walmart by the departments, we projected the weekly sales for each of the departments of Walmart for 2 conditions, when Holiday = True and when Holiday = False. The plot helped us further narrow our focus on stores that projected an increase of 55% or more, such that the management can target these stores to pre-stock before the holidays. We identified 15 departments that had a significant increase in sales during the holiday seasons.

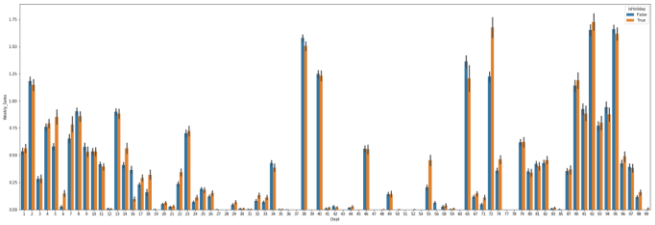


Fig 4. Department wide sales on Holidays and non-holidays

	Dept	Weekly_Sales	Weekly_Holiday_Sales	Percent_Increase
5	6	0.02824	0.14988	430.736544
16	18	0.15804	0.32089	103.043533
22	24	0.06947	0.10936	57.420469
30	32	0.08029	0.12997	61.875701
31	33	0.06858	0.11020	60.688247
39	41	0.00956	0.01507	57.635983
42	44	0.01517	0.02758	81.806196
46	48	0.00000	0.00109	inf
48	50	0.00015	0.00148	886.666667
50	52	0.00007	0.00208	2871.428571
52	55	0.20409	0.45428	122.588074
55	59	0.00237	0.00863	264.135021
59	71	0.04844	0.11113	129.417836
68	83	0.01035	0.01671	61.449275
80	99	0.00000	0.00644	inf

Aim

The aim of this project is to accurately predict Department wide Weekly Sales for the given Walmart store. Along with forecasting sales we are modelling the effect of Markdowns on Holiday weeks.

Objective

We ran various prediction models on the dataset and our major objective is to achieve the minimum prediction error in these prediction models. We calculate the error using WMAE (Weighted Mean Absolute Error)

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

where

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

Machine Learning Algorithms

A) KNN

Data pre-processing: The prediction of our query instance is based on the simple majority of the category of nearest neighbors, and for our dataset, the prediction is heavily influenced by the size of the store.

Model: We used K Nearest Neighbors which is instance-based learning to make data points closer to each other tend to behave similarly. We tuned the after tuning KNN model for a higher accuracy with WMAE value to be 0.1093.

B) SVM

Data pre-processing: First, we prepare the data by scaling the data to the centre.
Model: We implemented support vector machine as a black box algorithm to find the most optimal decision boundary that maximizes the distance from the nearest data points of all classes. We achieved a WMAE value of 0.1096.

C) Random Forest

Model: We trained large numbers of attribute data to reduce bias. The accuracy of our model, determined by the WMAE value was about 0.06, which has been the most efficient so far.

param_max_features	auto	log2	sqrt
param_n_estimators			
100	0.013	0.053	0.054
250	0.013	0.053	0.054
500	0.013	0.053	0.054

The following table shows the important variables. We found Department, Size, Store and Temperature to be the most significant attributes. To further validate the results obtained from Random Forest, we decided the run Linear Regression on the training dataset.

	importance
Dept	0.705009
Size	0.085554
Store	0.044799
Temperature	0.042463
Week	0.026631
Day	0.025454
Type_B	0.023253
Month	0.014262
MarkDown3	0.006772
Type_C	0.006021
Year	0.004483
MarkDown1	0.004457
MarkDown5	0.004214
MarkDown2	0.003947
IsHoliday	0.002681

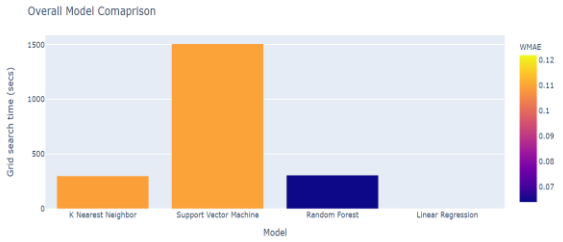
D) **Linear Regression**

Model: As a comparison model to verify the results obtained in RF, we implemented Linear Regression that gave us a WMAE value of 0.122, which is the highest, indicating a lower accuracy. The following table shows the week, size and department to be the most significant attributes in predicting Weekly Sales.

	column	Coefficients
11	Week	1.68
3	Size	0.16
1	Dept	0.07
14	Type_C	0.05
7	MarkDown3	0.03
5	MarkDown1	0.01
8	MarkDown5	0.01
4	Temperature	0.00
13	Type_B	0.00
6	MarkDown2	-0.00
2	IsHoliday	-0.00
9	Year	-0.02
0	Store	-0.03
12	Day	-0.13
10	Month	-1.63

Model Evaluation

In the last part of the modelling, we compared the models that we tried previously. From the analysis we drew based on the WMAE and the implementation time, we observed that SVM gave us the least accurate results with very high implementation time of about 1507 seconds and second to that is K Nearest Neighbors with an implementation time of 300 seconds. Linear Regression provided us with a higher WMAE factor of 0.122 but the least amount of implementation time of only 0.8 seconds.



	Model	Initial model Time (secs)	Grid search time (secs)	WMAE
0	K Nearest Neighbor	8.800	300.0	0.1093
1	Support Vector Machine	180.310	1507.0	0.1096
2	Random Forest	2.620	306.0	0.0640
3	Linear Regression	0.039	0.8	0.1220

Conclusion

After working on different algorithms and comparing their equivalent search criterions we concluded that Random Forest gave us the most accurate results with a WMAE of about 0.06 and an optimal implementation time of 2.6 seconds.

Dash App Link

https://predict-sales-walmart1.herokuapp.com

Acknowledgment

The authors would like to thank Prof. Ying Lin for his teachings and guidance which helped throughout this project from IST 707-Data Analytics.

References

1. https://www.kaggle.com/andredornas/tp2-walmart-sales-forecast
2. https://www.kaggle.com/nsawal/walmart-baseline-sales-forecasting-lstm