Application of Data Science in Analyzing and Forecasting Weekly Sales of a Store

Vishal Srivastava
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
Roll No. 19D110025

Aneri Modi
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
Roll No. 190100015

***Abstract*—A necessary task for firm owners is to make informed business decisions regarding their sales, in order to maintain the profitability of their store and survive. To successfully do this it is helpful to have insight regarding past monthly and weekly sales influenced by local parameters, and be able to reliably forecast sales and minimize the opportunity costs. Filling in the gap in existing analyses where there was an absence of a model to predict the overall weekly sales of a store, our project is a result of the motivation to predict it given certain factors.**

The datasets used have been extracted from Kaggle, consisting of the historical sales records of 45 Walmart Stores [1]. Our analysis has been focused on deriving features that directly or indirectly affect weekly sales, for which we extracted parameters like the 'month' and 'week of the month' [2] instead of using absolute date in order to create a more generalized model.

In our diagnostic analysis, we have tried to put forward a broad perspective of what influences a store's weekly sales taking into account parameters like previous week sales, holiday weeks, average fuel price, and unemployment rate which independently affect the consuming power of the local population and logistics to conduct business. We conclude by creating an ultimate machine learning model used to predict the weekly sales of a store, utilizing the results obtained from the analysis performed.

I. Introduction

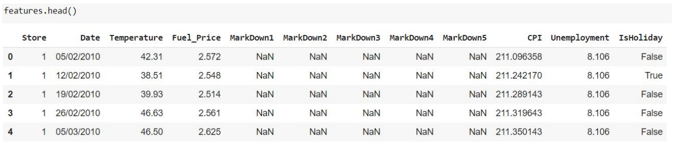
A competitive market is one in which there exist equal opportunities for all producers to produce and sell as per the requirement of the market. But in reality, markets today are far from this ideality. Markets today are dominated by bigger and better stores and firms which makes it a much bigger risk undertaking for store owners to establish and run a firm of any given scale. To address this broader problem, our project aims to analyse the datasets and predict the weekly sales of a store. This is important so that store owners can forecast their sales and accordingly make changes to their market schemes to maximize their profit with a comparative feeling of certainty than before. Does the temperature, fuel price, size of the store affect weekly sales? What would be the sales for the upcoming holiday week? These are some of the questions that our project does its best to answer, with the help of datasets containing records of sales of 45 stores. Exploratory data analysis followed by building models for prediction are the main ideas that the project is based on.

II. Related Work

The data source on Kaggle website [1] has a few notebooks which have worked on a similar task as our project, but with some differences. Most of them had built a model to predict the weekly sales per department of a store rather than the entire store. Noticing this shortcoming, our project is relevant to those who want a broad level observation on the sales that a Walmart store makes per week which could help the company in long term high level predictions, instead of only a departmental level prediction.

III. Datasets

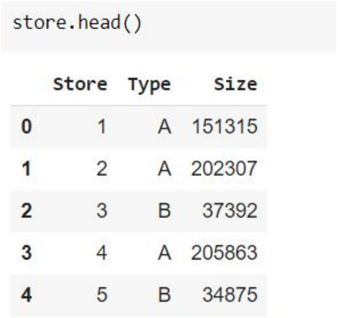
The analysis and prediction of the weekly sales of a Walmart store required the historic sales data of an adequate number of Walmart stores, and the various factors that these sales depended on. The three datasets used in this project are Features, Sales and Store, which are explained in brief below.



```
features.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	05/02/2010	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	12/02/2010	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	19/02/2010	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	26/02/2010	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	05/03/2010	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

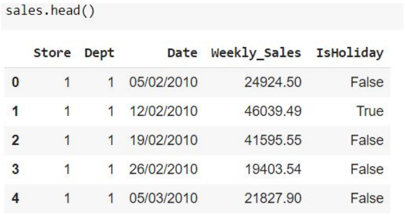
Fig. 1: Features of store and regional activities



```
store.head()
```

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

Fig. 2: Details of the store



```
sales.head()
```

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	05/02/2010	24924.50	False
1	1	1	12/02/2010	46039.49	True
2	1	1	19/02/2010	41595.55	False
3	1	1	26/02/2010	19403.54	False
4	1	1	05/03/2010	21827.90	False

Fig. 3: Details pertaining to Weekly sales of the stores

A. *Features*

This dataset contained numerous variables that the weekly sales of a store depended on. There were 12 columns headed by the features, which included Store number, Date, Average Temperature, Fuel Price in the region, Markdown1, Markdown2, Markdown3, Markdown4, Markdown5, Consumer Price Index (CPI), Unemployment rate, and whether the week was a holiday week or not. It consisted of 8190 records, however due to a majority of the data entries for the Markdown columns being Not Available and the weekly sales not recorded in the Sales dataset for the time period 02/11/2012 onwards, select columns and rows had to be dropped.

B. *Sales*

This dataset included our target variable which was the Weekly sales, and other features. It contained 5 columns namely Store number, Department, Date, Weekly Sales for that department and whether the week was a holiday or not. There were a total of 421570 records.

C. *Store*

The characteristics that described a Walmart store were given in this dataset, with the 3 columns being Store number, Type of Store indicative of its size, and the numerical size of the store. It had 45 records, each corresponding to a single store.
To create one comprehensive dataset to work on, the cleaned Features dataset was merged with the Sales and Store datasets using the groupby function to sum up the sales of all departments of a store on a particular date.

IV. Analysis Pipeline

The overall approach taken while performing an exploratory analysis on the datasets, was to grasp a general understanding of the data, to seek any existing correlations that may occur between the target total weekly sales and its features, and to draw sensible conclusions that would help in our ultimate task of predicting the sales. Apart from the general trivial analysis that is usually done, the plotting of a variety of graphs helped in the clear visualization of our data and its distribution. Steps included in the analysis are as follows:

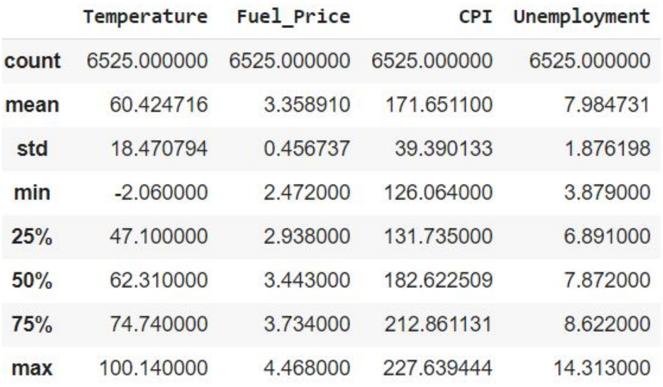
- The number of columns and rows in each of the three datasets, and the data type, number of missing and unique values for each feature. This helped in the initial grasping of the data.
- The list of unique values in each column. This helped recognize whether all three datasets were cohesive or there are extra elements added in some.
- The mean, standard deviation, maximum, minimum and skew for all continuous variables, along with their respective QQ, histogram, and Box and whiskers plots. Doing this helped in visualizing the distributions and their statistical significance.
- A heatmap which would show the correlation between variables.
- Bar graph between weekly sales and whether the week is a holiday, to see how that factor affects the sales.
- Scatter plot between the weekly sales and temperature for the top 100 highest sales, to visualize the effect it has on sales.
- Bar graph between weekly sales and the type of store, to see how sales are distributed amongst the types of stores.
- Line chart between weekly sales and date, to see how the sales depend on the period of the year.
- Line chart between weekly sales and week, to see how the sales depend only on the week of any month
- Line chart between sales and month-week combination in order to understand the underlying general behaviour of weekly sales irrepective of year

These steps took us closer to our goal of prediction of sales, by giving us a clear idea as to how the features and target are correlated and their extent of dependence on each other.

V. Results

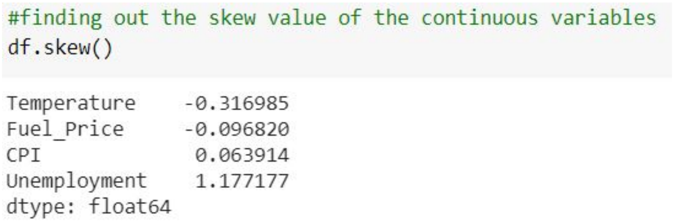
After the analysis was conducted, insightful results were obtained and the key observations made are as given below, corresponding to each point mentioned in the pipeline previously:

- The preliminary details about each of our datasets and their respective features helped us in feature selection as we could eliminate those features which had too little information. We could infer whether a variable is continuous or categorical and in turn make a balanced final dataset without losing much relevant information.
- Along with giving us the range of the features, the lists of unique values conveyed the fact that the Sales dataset contained weekly sales from dates 05/02/2010 to 26/10/2012 only, whereas the Features dataset had records from 05/02/2010 till 26/07/2013.
- The skew values indicate that Temperature and Fuel Price are left skewed and CPI and Unemployment are right skewed, both of which can be visually depicted in the QQ plots and histograms with 20 bins.



	Temperature	Fuel_Price	CPI	Unemployment
count	6525.000000	6525.000000	6525.000000	6525.000000
mean	60.424716	3.358910	171.651100	7.984731
std	18.470794	0.456737	39.390133	1.876198
min	-2.060000	2.472000	126.064000	3.879000
25%	47.100000	2.938000	131.735000	6.891000
50%	62.310000	3.443000	182.622509	7.872000
75%	74.740000	3.734000	212.861131	8.622000
max	100.140000	4.468000	227.639444	14.313000

Fig. 4: Statistical inferences from continuous features



```
#finding out the skew value of the continuous variables
df.skew()
```

```
Temperature    -0.316985
Fuel_Price     -0.096820
CPI             0.063914
Unemployment    1.177177
dtype: float64
```

Fig. 5: Skew

- The following correlation chart helped us understand how much each of our feature directly influenced our target variable.

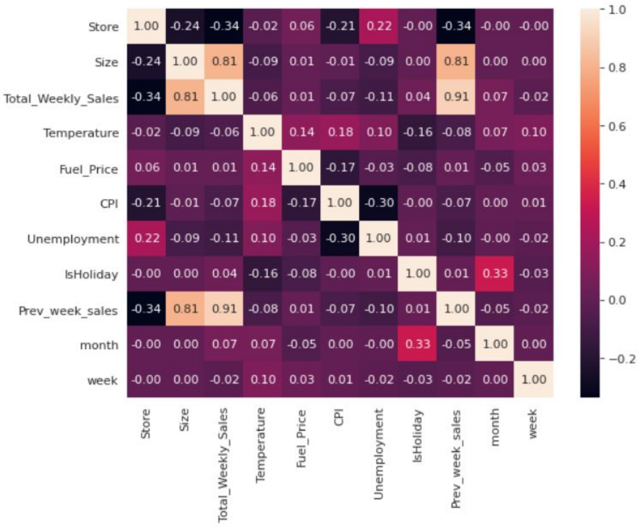


Fig. 6: Correlation chart for continuous variables

- The following bar graph depicts how weekly sales of a store increase during holiday week.

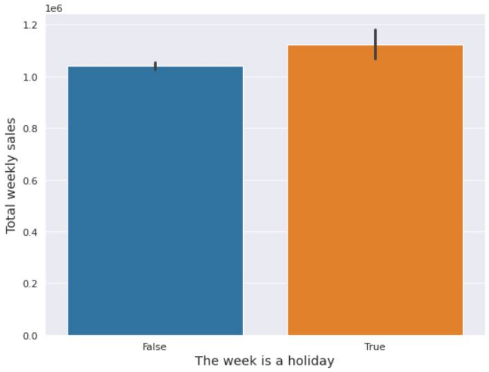


Fig. 7: Effect of holiday week on weekly sales

- Scatter plot for Weekly Sales v/s Temperature

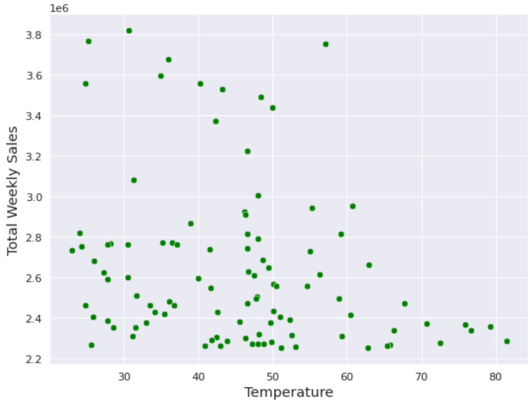


Fig. 8: Effect of temperature on weekly sales

- Bar graph for Weekly Sales v/s Type of store

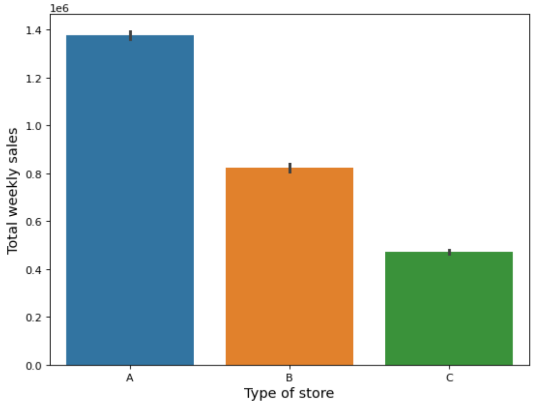


Fig. 9: Effect of type of store on weekly sales

- Variation of sales over time

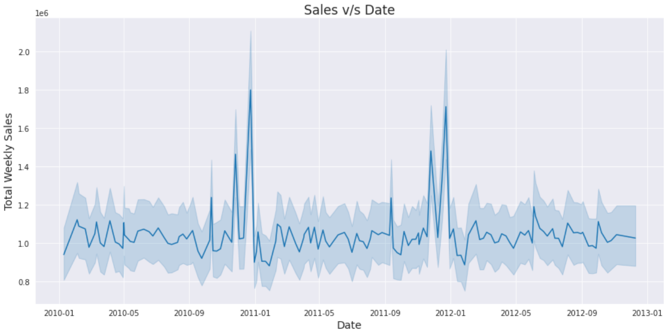


Fig. 10: Sales v/s Date

- Influence of week no. on weekly sales

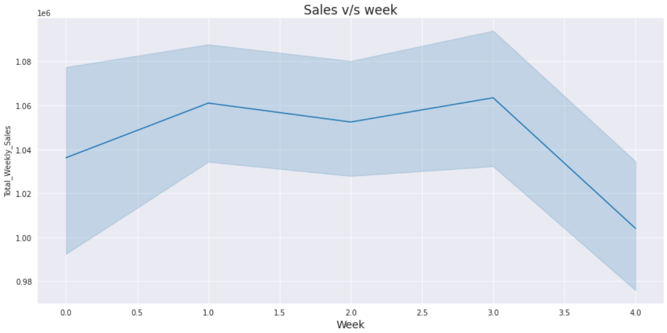


Fig. 11: Sales v/s week

- Simultaneous effect of month and week on weekly sales [2]

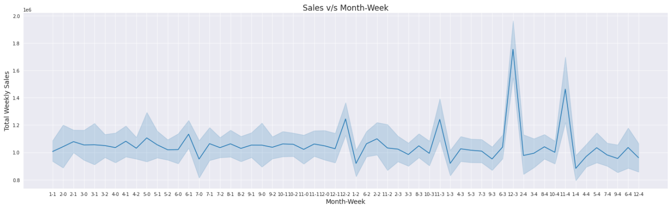


Fig. 12: Sales v/s Month-Week

For the prediction of the total weekly sales, the dataset was split into validation, training and testing sets in a **70:15:15** ratio. Three models were constructed with the purpose of finding out which model would be the most accurate in terms of metrics like mean square error and R2 score.
Results obtained from our prediction models after hyper parameter tuning:

- Random Forest Regression:
  Coefficient of determination for Validation Set-**0.9218808142481352**
  Mean Square Error for Validation Ser-**0.07375261316580636**
- Ridge Regression:
  Coefficient of determination for Validation Set-**0.8371815128571186**
  Mean Square Error for Validation Set-**0.17781348329269878**
- Support Vector Machine Regression (SVM-R):
  Coefficient of determination for Validation Set-**0.9099669885802577**
  Mean Square Error for Validation Set-**0.08786674007267384**

VI. Discussion

The graphical results can be interpreted qualitatively as follows:

- The heatmap shows a strong correlation between total weekly sales and size, previous week's sales and store, implying that these features are necessary to include in our models.
- From the bar graph it is observed that the week being a holiday week increases the sales of a store which can be attributed to the fact that customers tend to buy more during festivals like Christmas, Thanksgiving, etc.
- The scatter plot contains more samples with higher sales concentrated at lower temperatures, with decreasing sales as the temperature rises. Customers hence tend to not go shopping during extreme weather conditions.
- As evident from the bar graph, stores of Type A contribute the most to total sales of the franchise, followed by B and C, in that order.
- The line chart depicts the flat but slightly fluctuating trend in sales as time passes, with sharp rises occurring during holiday seasons of December due to Christmas,etc. Sharp declines before and after a holiday too occur, which is usually a period of saving for customers.
- It can be inferred that sales increase in the first days of every month and decrease by the last days which is representative of the fact that most people receive their income at the beginning of every month so they tend to buy more and decrease their consumption to lower expenditure.
- A big spike can be seen in the 12th month-3rd week and the 11th month-4th week which means that evidently

sales plunge upwards in the Christmas and Thanksgiving weeks respectively.

On observing the model performances, it can be concluded that in this case, Random Forest Regressor is the most accurate, followed by SVM-R and then Ridge Regression. When this best model is applied on the test set, the metrics obtained are:
Coefficient of determination for Test Set- **0.9410911381175502**
Mean Square Error for Test Set- **05847353469771061**

A. *Strengths and Limitations*

Strengths of our analysis include doing an in-depth exploration into the data sets, visually representing important relations and distributions, and testing multiple machine learning models on the validation set to supply the user/store-owner with the most accurate predictions.
Limitations of our approach would be it being unable to accept a date and values of other features for prediction, and the model not being deployed.

B. *Gain from analysis*

From this analysis, store owners could benefit from knowing how their sales vary across weeks. Using the model built once deployed, they could forecast their future sales and plan schemes to maximize their profit accordingly.

C. *What else can be explored*

From this analysis, store owners could benefit from knowing how their sales vary across weeks. Using the model built once deployed, they could forecast their future sales and plan schemes to maximize their profit accordingly.

VII. Acknowledgements

We would like to express our sincere gratitude and appreciation to our DS 203 course instructors:

Professor Amit Sethi
Professor Manjesh K Hanawal
Professor Sunita Sarawagi
Professor S. Sudarshan

for giving us this opportunity to do this project and also for their continued support and guidance throughout the course.
We would also like to thank our Teaching Assistants, Parikshit Bansal and Mohit Gupta for their help during the tutorials and also beyond the class hours, and our classmate Ritwik Gupta for his support to help understand various concepts in the course.

VIII. References

[1] The Kaggle data source from which we extracted the required dataset records to enable us to predict the weekly sales: Dataset
[2] Code for extracting week of the month from any given date: Stackoverflow