Final Project: Machine Learning for Big Data

Walmart Sales Forecasting

—----------

Akshobhya Yelanduru, Anirudh Penmatcha,
Aathirai Senthilkumar Thamaraiselvi, Rhea Matlapudi

Abstract

In this comprehensive analysis of Walmart's weekly sales dataset spanning 45 stores and 81 departments over several years, our team delved into the intricate web of factors influencing weekly sales. The primary focus was on unraveling the impact of time-based and space-based elements, with a particular emphasis on understanding how holidays influence store sales. The pivotal question guiding our investigation was: How do time-related factors, store-specific characteristics, and the inclusion of holidays collectively shape the weekly sales patterns across Walmart's extensive retail network?

The analysis revealed intriguing insights into the dynamics that contribute to weekly sales, offering valuable information for optimizing inventory management, forecasting demand, and enhancing overall operational efficiency. The identification of patterns and outliers in sales data enables Walmart to make informed decisions, ensuring effective resource allocation and a shopping experience tailored to customer needs. One notable aspect of the analysis was the observation of departmental variability. Department 72 emerged with higher weekly sales values, while department 92 stood out as the average performer. The influence of specific departments, especially during holidays like Thanksgiving, underscored the seasonal nature of sales, providing crucial information for targeted marketing and inventory planning. Store dynamics played a crucial role in shaping weekly sales patterns. Stores 20 and 4 consistently topped the charts in average weekly sales, highlighting the existence of areas with higher seasonal demands.

Further classification of stores into types A, B, and C based on size revealed distinct sales variations, offering valuable insights for tailoring strategies to different store types. The impact of holidays on weekly sales was pronounced, with Thanksgiving outshining Christmas, challenging conventional expectations. The absence of November and December sales data for

2012 added a noteworthy detail influencing yearly comparisons. Seasonal peaks were identified, with weeks 51 and 47, corresponding to Christmas and Thanksgiving, respectively, standing out with the highest sales. Surprisingly, the 22nd week, coinciding with the end of May, emerged as the fifth most important period, possibly linked to preparations for holidays and the general end of the school year. Monthly trends were also analyzed, revealing a dip in sales in January, echoing the aftermath of the high-sales months of November and December. Interestingly, factors such as Consumer Price Index (CPI), temperature, unemployment rates, and fuel prices exhibited no discernible patterns in relation to weekly sales, emphasizing the complex interplay of diverse factors in shaping consumer behavior and purchasing patterns.

In conclusion, this analysis provides a multifaceted understanding of the factors influencing weekly sales at Walmart, offering actionable insights for optimizing retail operations and enhancing the overall customer experience. The incorporation of departmental variability, store dynamics, holiday impacts, and seasonal peaks adds depth to the exploration of time-based and space-based elements in the context of machine learning applications.

Introduction

In the realm of retail analytics, our team embarked on a comprehensive exploration of Walmart's weekly sales dataset, spanning 45 stores and 81 departments over several years. The overarching objective of our research was to unravel the intricate web of factors that shape weekly sales dynamics within Walmart's extensive retail network. With a keen focus on both time-based and space-based elements, our analysis delved into the multifaceted interplay of variables, emphasizing the profound impact of holidays on store sales.

As we delved into the data, our team set out to answer a pivotal question that lies at the heart of this research endeavor: How do time-related factors, store-specific characteristics, and

the inclusion of holidays collectively mold the weekly sales patterns across Walmart's diverse retail landscape? The significance of this inquiry extends beyond mere curiosity, underlining its direct relevance to critical aspects of retail management.

Understanding the dynamics that contribute to weekly sales is paramount for optimizing inventory management, forecasting demand, and enhancing overall operational efficiency. This research not only sheds light on the intricate patterns within the dataset but also provides actionable insights for Walmart to make informed decisions. By identifying patterns and outliers in sales data, our analysis empowers Walmart to allocate resources effectively, ensuring that the shopping experience is maximally tailored to customer needs.

Our investigation unveils the nuanced relationships between various factors, ranging from the seasonal nature of sales, as highlighted by departmental variability, to the specific dynamics of individual stores and the overarching influence of holidays. The outcomes of this research hold the potential to revolutionize retail strategy by leveraging machine learning techniques to predict and understand consumer behavior in a dynamic and ever-evolving retail landscape.

This research project not only contributes to the academic discourse surrounding machine learning applications in retail analytics but also holds practical implications for one of the world's largest and most influential retailers. As we embark on this journey of exploration, our goal is not only to uncover insights but also to pave the way for actionable solutions that can propel Walmart's retail operations to new heights of efficiency and customer satisfaction.

Data Description

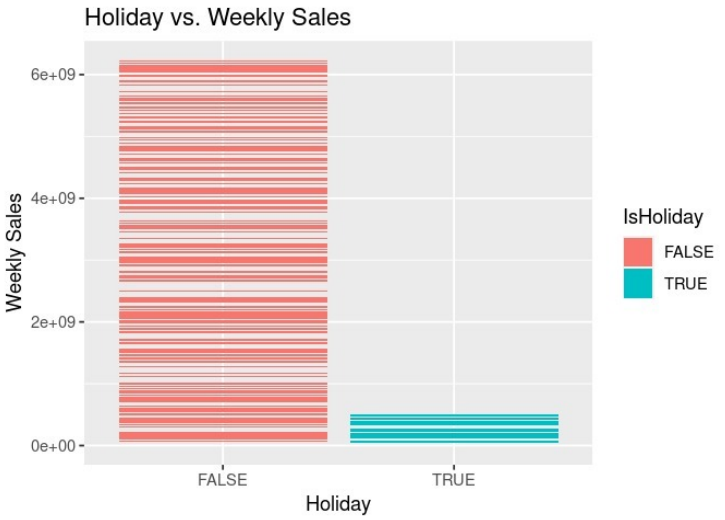
Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
Store	1.0	11.0	22.0	22.2	33.0	45.0	
Dept	1.00	18.00	37.00	44.24	74.00	99.00	
Date	2010-02-05	2010-10-08	2011-06-17	2011-08-18	2012-02-24	2012-10-26	
Weekly_Sales	0	2120	7662	16033	20271	693099	
Temperature	-2.06	46.68	92.09	60.09	74.28	100.14	
Fuel_Price	2.472	2.933	3.452	3.361	3.738	4.468	
MarkDown1	0.27	2240.27	5347.45	7247.82	9210.90	88646.76	270031
MarkDown2	-265.8	41.6	192.0	3330.2	1926.9	104519.5	309308
MarkDown3	-29.1	5.1	24.6	1441.7	104.0	141630.6	283561
MarkDown4	0.22	504.22	1481.31	3384.78	3595.04	67474.85	285694
MarkDown5	135.2	1878.4	3359.4	4629.5	5563.8	108510.3	269283
CPI	126.1	132.0	182.4	171.2	212.4	227.2	
Unemployment	3.879	6.891	7.866	7.960	8.567	14.313	
Size	34875	93638	140167	136750	202505	219622	

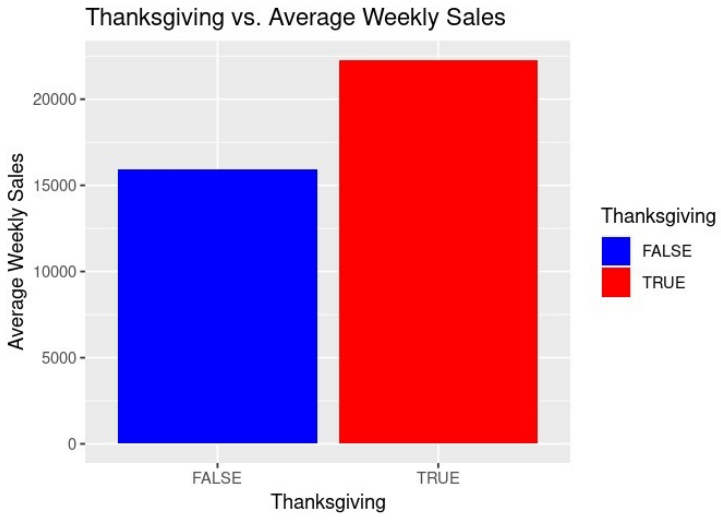
Holiday	Mode	False	True
IsHoliday.y	logical	390652	29560
Super_Bowl	logical	411339	8873
Thanksgiving	logical	414266	5946
Labor_Day	logical	411380	8832
Christmas	logical	414303	5909

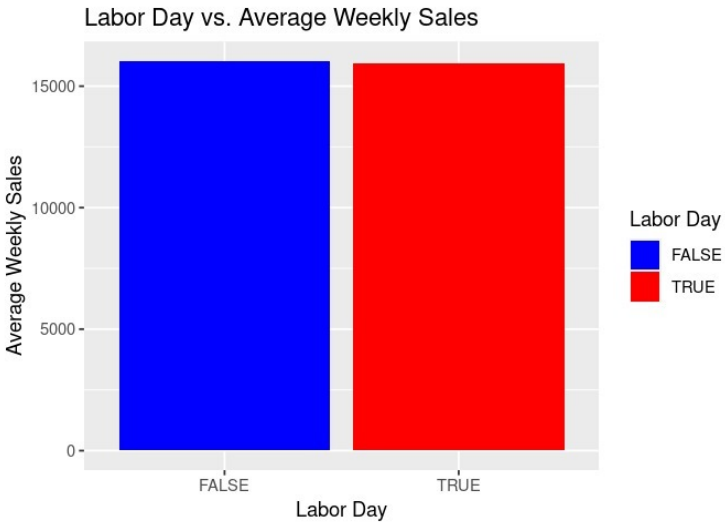
Exploratory Data Analysis

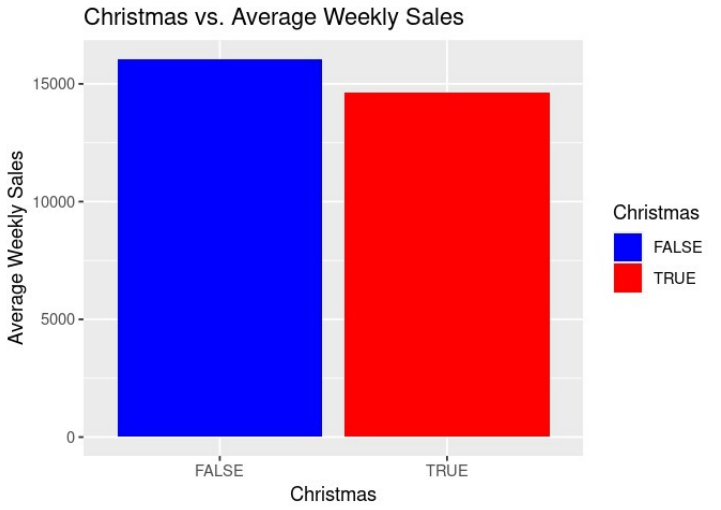
- **Department Variability:** While department 72 boasts higher weekly sales values, department 92 emerges as the average performer. The influence of certain departments, especially during holidays like Thanksgiving, highlights the seasonal nature of sales.
- **Store Dynamics:** Stores 20 and 4 consistently top the charts in average weekly sales, emphasizing that certain areas experience higher seasonal demands. Additionally, the classification of stores into types A, B, and C, based on size, reveals distinct sales variations.
- **Holiday Impact:** Unsurprisingly, holidays significantly boost average sales. Thanksgiving outshines Christmas, challenging conventional expectations. The absence of November and December sales data for 2012 is a noteworthy detail which influences yearly comparisons.
- **Seasonal Peaks:** Weeks 51 and 47, marked by Christmas and Thanksgiving, respectively, stand out with the highest sales. Unexpectedly, the 22nd week, coinciding with the end of May, emerges as the fifth most important period, possibly linked to preparations for holidays and the general end of the school year.
- **Monthly Trends:** January witnessed a dip in sales, echoing the aftermath of the high-sales months of November and December. Notably, CPI, temperature, unemployment rates, and fuel prices exhibit no discernible patterns in relation to weekly sales.

Exploratory Data Analysis Cont.









Econometric Methods and Model Description

This study analyzes the effect that a set of different variables has on the sales in Walmart stores. The sales of the various goods in the Walmart are determined by the eighteen variables that are being considered. These include mainly- the day of sales, holiday, months and also the unemployment of the people along with the Fuel prices in the market. This study combines a diverse set of variables and uses the techniques of simple regression and checks using the LASSO method. The model also uses the random forest method for further analysis. For this however, we had to modify the dataset and drop a few variables to prevent overfitting and other computational errors. The R squared value that we obtained is not very different from that of the linear regression model.

In the dataset, a few variables are dropped in order to improve the efficiency of the model and also making the system tighter in its application. The CPI(Consumer Price Index) variable has been dropped from the model as,the CPI variable typically represents the cost of a typical basket of consumer goods in the market, which does not have a strong correlation with the sales of one firm (in this case Walmart). The other variables however have been considered as they tend to have theoretically a stronger correlation with the sales of the commodities.

The Linear Regression model is a simple yet robust model that can be used for a wide range of applications. Especially, where there might be a linear relation between the dependent and independent variable and their relation can be established with the help of a coefficient. So the sales are kept as the dependent variable and the rest are the independent variables. The coefficient statistics has been used to study the impact that an independent variable has on the dependent variables and based on the value that is obtained we can make our judgements. The variables taken have both positive and negative coefficients of various values and that shall be

discussed in the following section. In addition to these are the R square statistics and the LASSO values.

Discussion of Results

The Model provides us with logical yet fascinating results. The values of the coefficients in the variables show how much of an impact it has on the sales. From the observations we have obtained the following results

a) **Holidays-** There is a strong relationship between the holidays and the sales of the goods. This makes sense logically as depending on whether there is a holiday or not people would want to buy more or less of certain goods. The coefficient that is seen in the holiday variables does express a correlation, but when the holiday variable is taken out of the regression, we notice that there is not much change in the sales variable with a change in the holidays.

   We also notice from the data that the impact that a holiday has on the sales depends on the type of holiday that is being talked about. While the data shows a positive relationship of average sales per week with the holiday on a Thanksgiving break and a neutral relationship on Labor day, there is an unusually negative relationship between the sales during the Christmas week. This brings us to the point that the sales variable also depends on the type and basket of goods that is being bought, which does vary with the season and the type of holiday. To give a simple example- the sales of Turkey go up on Christmas and the sales of Wine and Christmas trees during Christmas. If one measures the sales of winter clothes during summer break, there is bound to be a negative correlation.

So, we can conclude that this relationship between the holiday and the sales may not be completely defined as the data available limits us only to the general sales of all goods without any specifications.

b) **Months of the Year-** There are multiple variables that are dedicated to this in the regression model, where each one is dedicated to one month of the calendar year. The strong correlation that is established from the data shows that the month of the year that is being considered is very important for the level of sales. This is to do with various other subfactors like-Income, holidays and the seasons all of which have an impact on the sales. This once again brings us to the types of goods that are being sold, but is unavailable with this data.

c) **The Unemployment level-** The data in the model shows a strong negative correlation between the unemployment rates and the sales of the goods in the market. This is correct theoretically as when there is a greater level of unemployment, the purchasing power of the consumer plummets causing a fall in the demand and hence the sales. But Once again all of this depends on the type of goods that are being considered in the basket. The luxury goods are more elastic than the essentials as they happened to be first to be compromised when there is unemployment.

d) **The Fuel Prices-** The fuel prices affect the transport and hence would theoretically impact the sales as well. Even from the model the coefficient that is seen is relatively higher than the other variables and hence there does seem to be a

strong impact that the fuel prices have on the sales. But the relationship seems to be a positive one.

Robustness Checks

The robustness of the model as mentioned before is estimated using the R square statistics and the LASSO estimate. The R square method shows us the effect that an unseen data has on the model that has been estimated, or more generally, it estimates the variance of the dependent variable based on the variance in an independent variable. This is different from correlation in the way that, in correlation, the change in one variable is measured by the change in other, while R squared is a metric by which we can decide the proportion of impact one variance has on the other. The

The LASSO Estimate is a test from which one can test a model for overfitting and unwanted variables. The better Lasso score can potentially ensure a better model performance

```
18 x 1 sparse Matrix of class "dgCMatrix"
                          s1
(Intercept)      1563665.476
...1                       .
Fuel_Price         43038.527
Unemployment      -25230.876
Week_Number                .
EventsLabour.Day   28368.318
EventsNo_Holiday  -19717.254
month2            126756.179
month3             61953.960
month4             11811.069
month5             47940.739
month6             80055.141
month7             -1806.824
month8             47424.373
month9                     .
month10                    .
month11            28754.045
month12                    .
```

The Above statistics shows us the LASSO test for the given variables in the model.

Conclusion

From this model and study, that is based on a simple analysis of the sales statistics in Walmart and how that is impacted by the various factors that are taken in the model. This study is brilliant from the perspective of economics and data science alike. From the perspective of economics we directly get to witness some of the most fundamental concepts of the discipline like- Demand, supply, unemployment, direct and indirect costs and the change in the demand due to other factors, like holidays. From the perspective of Data Science, we got to learn and examine the application of the various models that we have studied in the class. From Simple Regression to

Random foresting , from LASSO to the R squared test all were used in this model. Even the various graphical and modeling methods that we studied- Bar graphs, the heat maps and also the pie charts all of which were used in the analysis of this data. The dataset was slightly confusing in the sense that it was hard to manage and hence it became difficult for us to manage with the random forest model. In addition, as mentioned in the earlier sections, the data for the sales are not specific to any product or a set of products and hence don't tell the entire story, as the independent variables that are being considered impact different goods in different ways and if we knew the type of goods that were bought, one could perfectly understand the markets. But Overall, I think that this assignment was truly enriching and was a great learning experience for all of us. It taught us everything from team work, collaboration, and many other skills all the way to complex econometrics concepts. This class was indeed an experience for a lifetime.

****