Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, July : 2023

AN EXPLORATORY DATA ANALYSIS OF PREDICTING WALMART SALES USING MACHINE LEARNING

Praveen Kumar, MCA, DCA, DVR & Dr. Hima Shekar MIC College of Technology, A.P., India.
B. MURALI KRISHNA, Assistant Professor, Dept. Of AI & IT, DVR & Dr. Hima Shekar MIC college of Technology, A.P., India.

Abstract— This paper explores the performance of a subset of Walmart stores and forecasts future weekly sales for these stores based on several models including linear and lasso re-gression, random forest, and gradient boosting. An exploratory data analysis has been performed on the dataset to explore the effects of different factors like holidays, fuel price, and temperature on Walmart's weekly sales. Additionally, a dashboard high-lighting information about predicted sales for each of the stores and departments has been created in Power BI and provides an overview of the overall predicted sales.

Through the analysis, it was observed that the gradient boosting model provided the most accurate sales predictions and slight relationships were observed between factors like store size, holidays, unemployment, and weekly sales. Through the implementation of interaction effects, as part of the linear models, relationship between a combi-nation of variables like temperature, CPI, and unemployment was observed and had a direct impact on the sales for Walmart stores.

INTRODUCTION

The 21st century has seen an outburst of data that is being generated as a result of the continuous use of growing technology. Retail giants like Walmart consider this data as their biggest asset as this helps them predict future sales and customers and helps them lay out plans to generate profits and compete with other organizations. Walmart is an American multinational retail corporation that

has almost 11,000 stores in over27 countries, employingover2.2 million associates (Wikipedia)

Catering to their customers with the promise of ‘everyday low prices’, the range of products sold by Walmart draws its yearly revenue to almost 500 billion dollars thus making it extremely crucial for the company to utilize extensive techniques to forecast future sales and consequent profits. The world’s largest company by revenue, Walmart, sells everything from groceries, home furnishings, body care products to electronics, clothing, etc. and generates a large amount of consumer data that it utilizes to predict customer buying patterns, future sales, and promotional plans and creating new and innovative in-store technologies. The employment of modern technological approaches is crucial for the organization to survive in today’s cutting-edge global market and create products and services that distinguish them from its competitors.

The main focus of this research is to predict Walmart’s sales based on the avail-able historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores understudy. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanks giving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue.

Walmart runs several promotional markdown sales throughout the year on days immediately following the prominent holidays in the United States; it becomes crucial for the organization to determine the impact of these promotional offerings on weekly sales to drive resources towards such key strategic initiatives. It is also essential for Walmart to understand user requirements and user buying patterns to create higher customer retention, increasing their demand adding to their profits. The findings from this study can help the organization understand market conditions at various times of the year and allocate resources according to regional demand and profitability.

Additionally, the application of big data analytics will help analyze past data efficiently to generate insights and observations and help identify stores that might be at risk, help predict as well as increase future sales and profits and evaluate if the organization is on the right track.

The analysis for this study has been done using SQL, R, Python, and Power BI on the dataset provided by Walmart Recruiting on Kaggle ("Walmart Recruiting – Store Sales Forecasting," 2014). The modeling, as well as the exploratory data analysis for the research, have been performed in R and Python, aggregation and querying will be

Performed using SQL and the final dashboard has been created using Power BI.

LITERATURE REVIEW

In2015, Harsoor and Patil (Harsoor & Patil, 2015) worked on forecasting Sales of Walmart Stores using big data applications: Hadoop, MapReduce, and Hive so that resources are managed efficiently. This paper used the same sales dataset that has been used for analysis in this study, however, they forecasted the sales for the upcoming 39 weeks using Holt's winter algorithm. The forecasted sales are visually represented in Tableau using bubble charts.

Michael Crown (Crown, 2016), a data scientist, performed analysis on a similar dataset but instead focused on the usage of time series forecasting and non-seasonal ARIMA models to make his predictions. He worked on ARIMA modeling to create one year of weekly forecasts using 2.75 years of sales data, with features of the store, department, date, weekly sales, and holiday data. Performance was measured using normalized root-mean-square error (NRMSE).

Forecasting has not been limited to just business enhancement. Several researchers have tried to utilize machine learning and statistical analysis to build predictive mod-els that can accurately predict the weather, monitor stock prices and analyze market trends, predict illnesses in a patient, etc.

Likewise, in 2017, Chouskey and Chauhan (Chouksey& Chauhan, 2017) worked on creating a weather forecasting model that accurately predicts the weather and sends out weather warnings for people and businesses o that they can better prepare for the unforeseeable weather. The authors make use of MapReduce and Spark to create their models and gather data from various weather sensors; weather forecasts can be essentially important as they influence all human aspects and the authors have made use of various parameters like temperature, humidity, pressure, wind speed, etc.to make better predictions.

Another approach followed by Rajat Panchotia (Panchotia, 2020) to create a pre-dictive model using linear regression throws light on the various regression techniques and the metrics that need to be defined when creating such models. He talks about the importance of defining techniques that should be considered, like studying the num-ber of independent variables and type of dependent variables, determining the best fit, etc., based on the nature of data and the most accurate regression model that should be selected based on results obtained. In his article, he also emphasizes on the use of regression coefficients, p-values, variable selection, and residual analysis to study the performance of regression models. While Panchotia only focuses on studying the direct relationship

between the independent and dependent variables of the dataset, another theory by James Jaccard and Robert Turrisi (Jaccard&Turrisi,2018)involves observ-ing the change in the relationship between an independent and dependent variable as a result of the presence of a third variable, called the moderator variable.

RELATED WORK

The paper comprises of several different components that explore various aspects ofthe45Walmartstoresusedinthisstudy.The methodology section is broken down into several sub-sections that follow a 'top-down' approach of the process that is followed in this analysis.

This section contains detailed information about the dataset, the exact techniques that have been used in forecasting weekly sales and the last section talks about how this study is significant in predicting the weekly sales for Walmart stores. It will also discuss the success of the applied models in identifying the effect of different factors on such weekly sales.

PROPOSED WORK

The purpose of this study is to predict the weekly sales for Walmart based on available historical data (collected between 2010 to 2013) from 45 stores located in different re-gions around the country. Each store contains a number of departments and the main deliverable is to predict the weekly sales for all such departments.

The data has been collected from Kaggle and contains the weekly sales for 45 stores, the size and type of store, department information for each of those stores, the amount of weekly sales, and whether the week is a holiday week or not.There is additional information in the dataset about the factors that might influence the sales of a particular week. Factors like Consumer Price Index (CPI), temperature, fuel price, promotional mark downs for the week, and unemployment rate have

been recorded for each week to try and understand if there is a correlation between the sales of each week and their determinant factors.

Correlation testing has been performed to understand if there is a correlation be-tween the individual factors and weekly sales and whether such factors have any impact on sales made by Walmart.

This study also includes an extensive exploratory data analysis on the provided Walmart data set to understand the following:

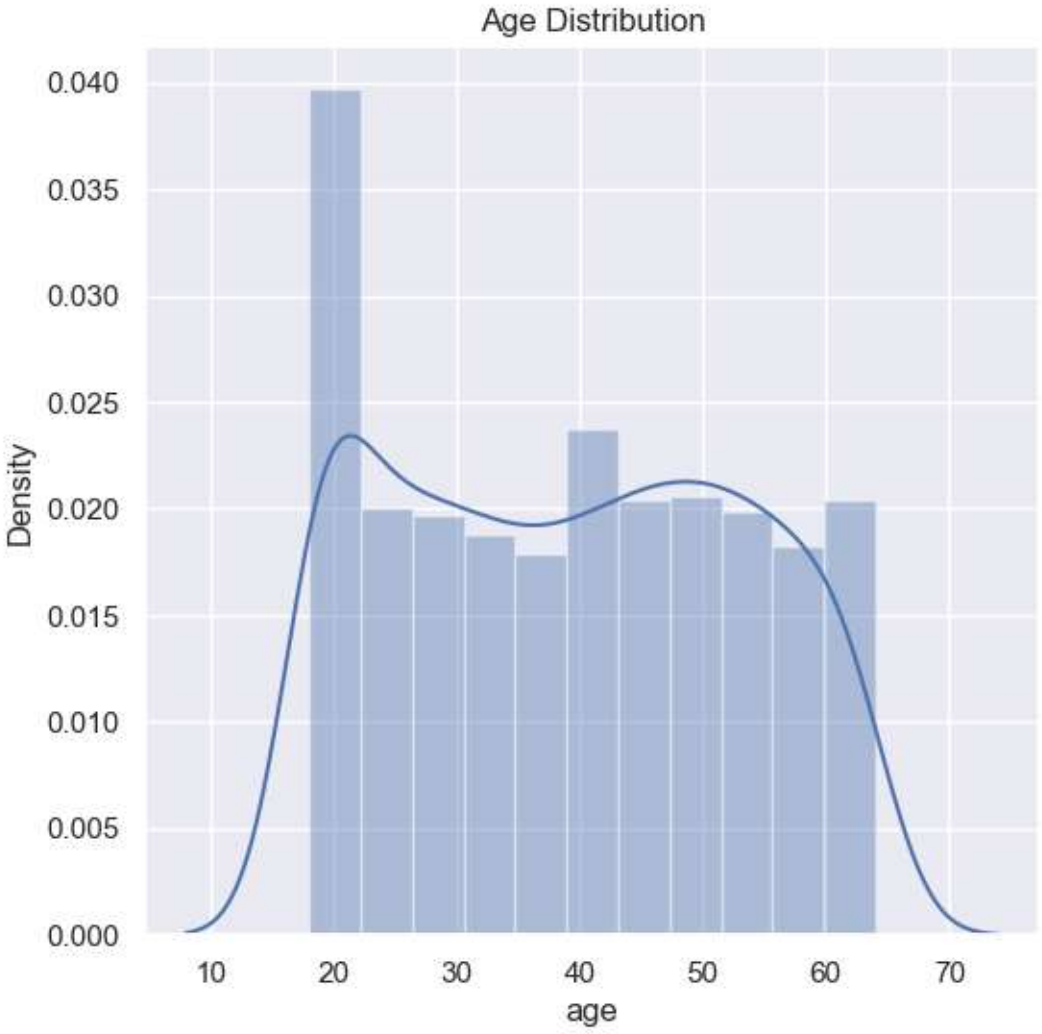
- Identifying store as well as department-wide sales in Walmart
- Identifying sales based on store size and type
- Identifying how much sales increase during holidays
- Correlation between the different factors that affect sales
- Average sales per year
- Weekly sales as per region temperature, CPI, fuel price, unemployment

Apart from identifying these direct relationships between independent and depen-dent variables, some interaction effects have also been studied as part of the Multiple Linear Regression model to understand if a certain combination of the factors under study can directly impact the weekly sales for Walmart.

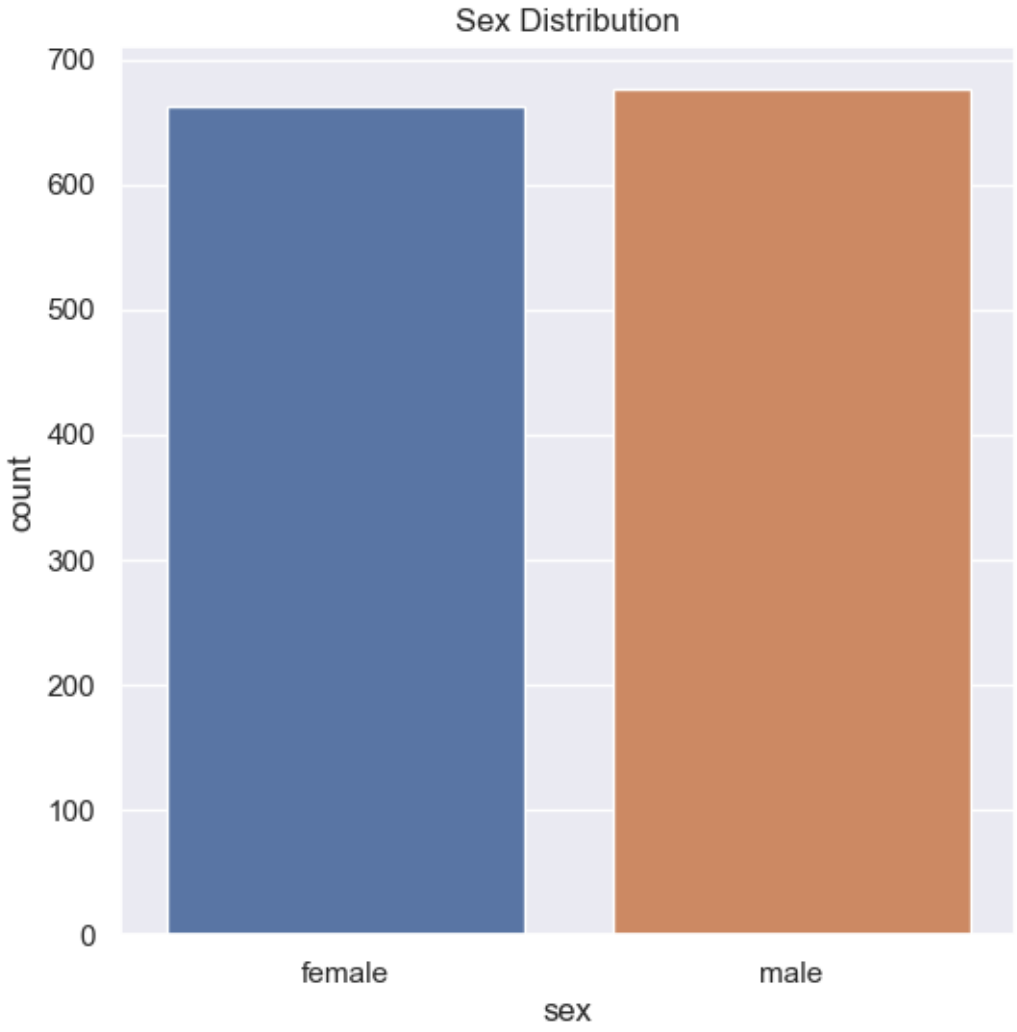
After employing different algorithms to predict future sales and correlation between factors for the retail store, a dashboard that tracks the above-mentioned outcomes has been created (in Power BI) and also includes then ew predictions to collectively visual-ize the out comes of this research and present them to amateur users more effectively.

SAMPLE SCREENSHOTS

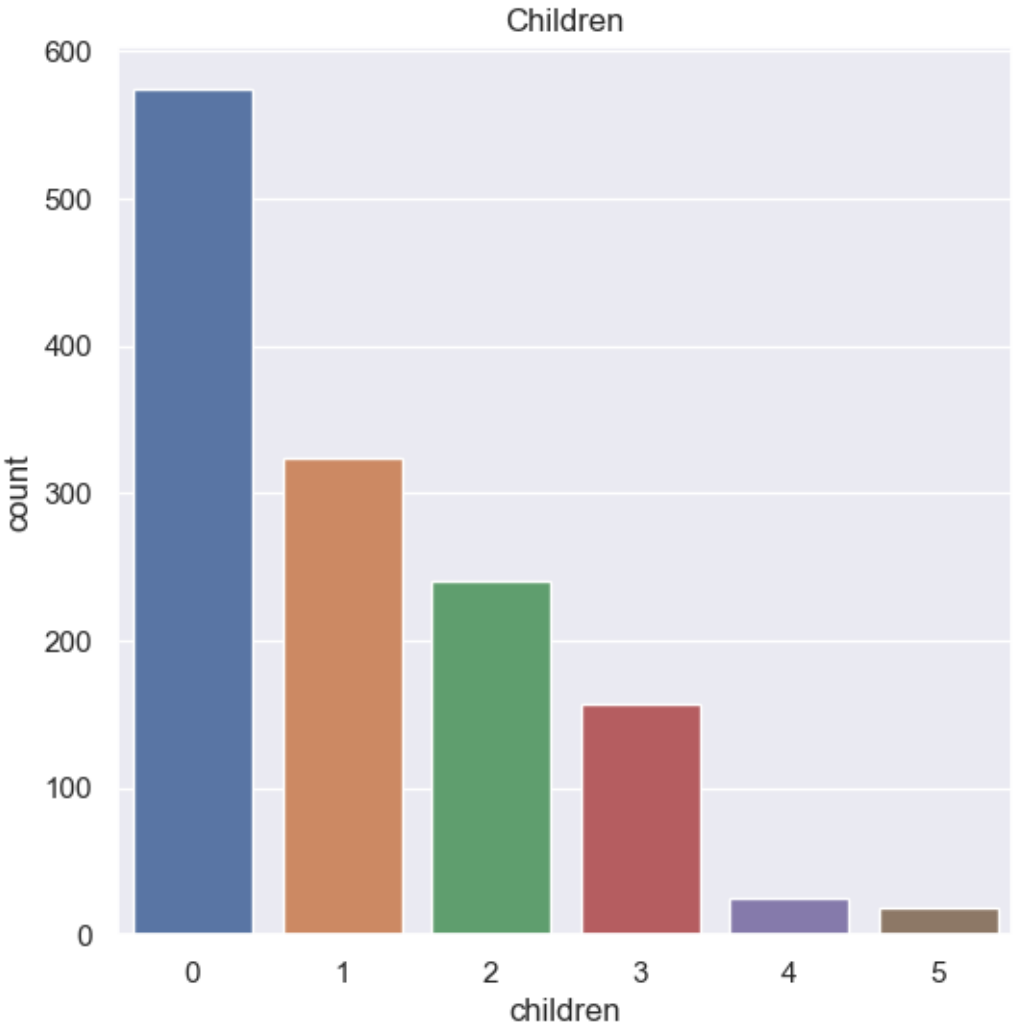
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, July : 2023



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, July : 2023



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 7, July : 2023



CONCLUSION

In this paper, we propose a the models created have been developed based on certain preset as-assumptions and business conditions; it is harder to predict the effects of certain eco-nomic, political, or social policies on the sales recorded by the organization. Also, itis tough to predict how the consumer buying behavior changes over the years or how the policies laid down by the management might affect the company's revenue; these factors can have a direct impact on Walmart sales and it is necessary to constantly study the market trends and compare them with existing performance to create better policies and techniques for increased profits.

References

Bakshi, C. (2020). Random forest regression. https : / / levelup . gitconnected . com /random-forest-regression-209c0f354c84

Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to utilize linear regressions in predictiveanalytics.https://www.dummies.com/programming/big-

data/data-science/how-to-utilize-linear-regressions-in-predictive-analytics/

Baum, D. (2011). How higher gas prices affect consumer behavior. https : / / www .sciencedaily.com/releases/2011/05/110512132426.htm

Brownlee,J.(2016).Feature importance and feature selection with xgboost in python. https : / / machine learning mastery . com / feature - importance - and - feature -selection-with-xgboost-in-python/

Chouksey, P., & Chauhan, A. S. (2017). A review of weather data analytics using bigdata. *International Journal of Advanced Research in Computer and Communica-tionEngineering*,*6*.https://doi.org/https://ijarcce.com/upload/2017/january-17/IJARCCE%2072.pdf

Crown, M. (2016). Weekly sales forecasts using non-seasonal arima models. http : //mxcrown.com/walmart-sales-forecasting/

Editor,M.B.(2013).Regression analysis: How do interpretr-squared and assess the goodness-of-fit?https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Ellis, L. (2019). Simple eda in r with inspectdf. https://www.r-bloggers.com/2019/05/part-2-simple-eda-in-r-with-inspectdf/

.