[P-1]

BUSINESS INTELLIGENCE PROJECT REPORT ON

“ IDENTIFYING DATA MINING TASKS AND PERFORMING THEM ON THE GIVEN DATASET “

SUBMITTED TO THE SAVTRIBAI PHULE PUNE UNIVERSITY,
PUNEIN THE PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF THE
DEGREE OF

FINAL YEAR (COMPUTER ENGINEERING)

BY

Student Name	Roll No:
HITESH MUNOT	C43338
CHINMAYEE KHARWADE	C43331
SHIVAM NIKAM	C43340

Under the guidance of

Prof. Priyanka Kinage

DEPARTMENT OF COMPUTER ENGINEERING

STES’S SMT. KASHIBAI NAVALE COLLEGE OF ENGINEERING

VADGAON BK, OFF SINHGAD ROAD, PUNE 411041

SAVTRIBAI PHULE PUNE UNIVERSITY 2022 – 23

[P-2]

DEPARTMENT OF COMPUTER ENGINEERING

STES'S SMT. KASHIBAI NAVALE COLLEGE OF ENGINEERING

VADGAON BK, OFF SINHGAD ROAD, PUNE 411041

CERTIFICATE

This is to certify that the project report entitles

" IDENTIFYING DATA MINING TASKS AND PERFORMING THEM ON THE GIVEN DATASET "

Submitted by

Student Name:	Roll No:
HITESH MUNOT	C43339
CHINMAYEE KHARWADE	C43331
SHIVAM NIKAM	C43340

is a bonafide work carried out by them under the supervision of **Prof. Priyanka Kinage** and it is submitted towards the partial fulfillment of the requirement of SAVTRIBAI PHULE PUNE UNIVERSITY, Pune for the award of the degree (Computer Engineering)

(Prof. Priyanka Kinage)
Guide
Department of Computer Engineering

(Prof. R. H. Borhade)
Head,
Department of Computer Engineering

Seal/Stamp of the College

(Dr. A. V. Deshpande)
Principal,
Smt. Kashibai Navale College of
Engineering
Pune – 41

Place: Pune

Date:

[P-3]

ACKNOWLEDGEMENT

We would like to express our very great appreciation to **Prof. Priyanka Kinage** for her valuable and constructive suggestions during the planning and development of this research work. Her willingness to give his time so generously has been very much appreciated. Her patient guidance, enthusiastic encouragement, useful critiques and also his assistance in keeping our progresson schedule throughout our semester will be acknowledged.

HITESH MUNOT
CHINMAYEE KHARWADE
SHIVAM NIKAM
NAME OF THE STUDENT

[P-4]

ABSTRACT

Data is one of the most essential commodities for any organization in the 21st century. Harnessing data and utilizing it to create effective marketing strategies and making better decisions is extremely essential for organizations. For a conglomerate as big as Walmart, it is necessary to organize and analyze the large volumes of data generated to make sense of existing performance and identify growth potential. The main goal of this project is to understand how different factors affect the sales for this conglomerate and how these findings could be used to create more efficient plans and strategies directed at increasing revenue.

This paper explores the performance of a subset of Walmart stores and forecasts fu- ture 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 implementa- tion 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.

TABLE OF CONTENTS

1. 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 over 27 countries, employing over 2.2 million associates.

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 pre- dict customer buying patterns, future sales, and promotional plans and creating new and innovative in-store technologies. The employment of modern technological ap- proaches 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 under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving 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 organi- zation 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.

1.1 TOOLS AND TECHNOLOGIES APPLIED

The analysis for this study have been performed using some main tools: R, Python, and Power BI. The models and Exploratory Data Analysis have been executed using development tools like R Studio and PyCharm.

Several packages have been used to perform the initial and final outcome EDA for the analysis. For the initial EDA, a combination of R and Python libraries like inspectdf, ggplot2, plotly, caret, matplotlib, seaborn, etc have been implemented. Packages like numpy, pandas, tidyverse, etc. have been used for data wrangling and manipulation. For the models that have been created, several packages like 'scikit-learn', 'xgboost', etc have been applied.

2. PROBLEM STATEMENT

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 markdowns 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 Wal- mart dataset 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

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 has been created and also includes the new predictions to collectively visual ize the outcomes of this research and present them to amateur users more effectively.

3. METHODOLOGY

The project comprises of several different components that explore various aspects of the 45 Walmart stores used in this study. 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.

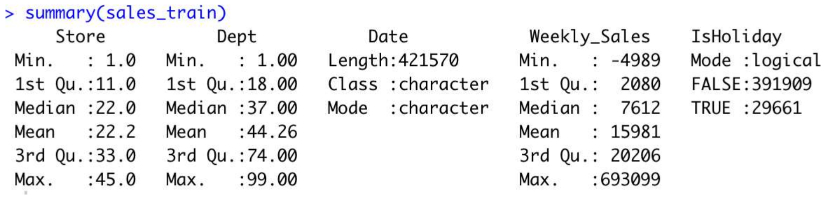
4. ABOUT THE DATASET

The dataset for this study has been acquired from a past Kaggle competition hosted by Walmart, this can be found here: https://www.kaggle.com/c/ walmart-recruiting-store-sales-forecasting/data. It contains historic weekly sales in- formation about 45 Walmart stores across different regions in the country along with department-wide information for these stores.

The ’test.csv’ data file that is a part of this dataset is only being used to predict values derived from the model with the lowest WMAE score. Because, the dataset contains no target variable, in our case ’Weekly Sales’, it cannot be used for testing for this analysis. Instead, the training dataset ’train.csv’ is being split into training and validation datasets for this study.

The main goal of this study is going to be to predict the department-wide weekly sales for each of these stores.

The dataset is already divided into separate training and testing data; the testing data is identical to the training dataset apart from the weekly sales information. The training dataset contains weekly sales information from 2010-02-05 to 2012-11-01 about the stores and departments. It also contains a column that suggests whether a particu- lar date falls on a holiday or not. In total, there are 4,21,570 rows in the training dataset and 1,15,064 rows in the testing dataset. (Figure 1)



```
> summary(sales_train)
     Store           Dept           Date            Weekly_Sales     IsHoliday
 Min.   : 1.0   Min.   : 1.00   Length:421570      Min.   : -4989   Mode :logical
 1st Qu.:11.0   1st Qu.:18.00   Class :character   1st Qu.:  2080   FALSE:391909
 Median :22.0   Median :37.00   Mode  :character   Median :  7612   TRUE :29661
 Mean   :22.2   Mean   :44.26                      Mean   : 15981
 3rd Qu.:33.0   3rd Qu.:74.00                      3rd Qu.: 20206
 Max.   :45.0   Max.   :99.00                      Max.   :693099
```

Figure 1. A summary of the Training dataset

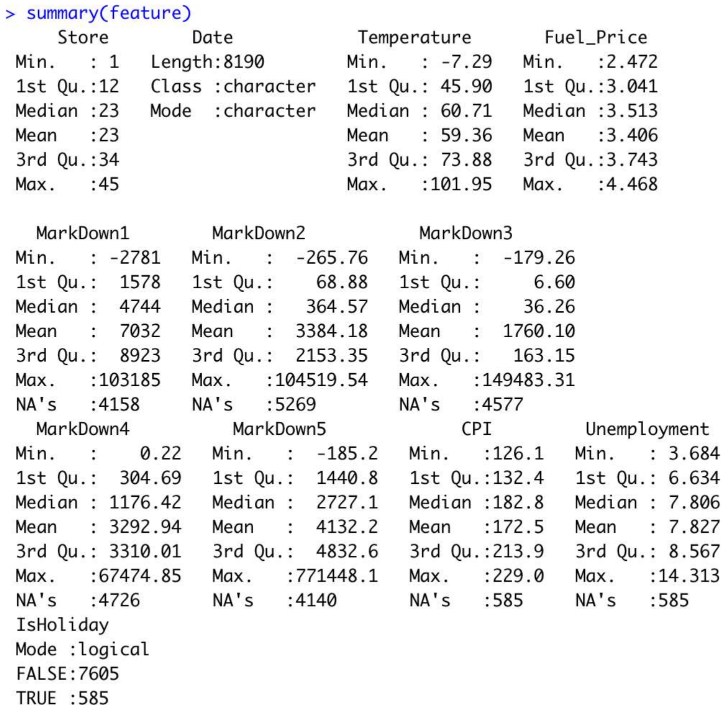
There is another dataset called ‘stores.csv’ that contains some more detailed information about the type and size of these 45 stores used in this study.

Another big aspect of this study is to determine whether there is an increase in the weekly store sales because of changes in temperature, fuel prices, holidays, mark- downs, unemployment rate, and fluctuations in consumer price indexes, The file ‘fea- tures.csv’ contains all necessary information about these factors and is used in the anal- ysis to study their impact on sale performances.

The holiday information listed in the study is

Holiday Name	Date 1	Date 2	Date 3	Date 4
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

A summary of the features dataset is displayed in the image below. (Figure 2)



```
> summary(feature)
     Store         Date            Temperature       Fuel_Price   
 Min.   : 1   Length:8190        Min.   : -7.29   Min.   :2.472  
 1st Qu.:12   Class :character   1st Qu.: 45.90   1st Qu.:3.041  
 Median :23   Mode  :character   Median : 60.71   Median :3.513  
 Mean   :23                      Mean   : 59.36   Mean   :3.406  
 3rd Qu.:34                      3rd Qu.: 73.88   3rd Qu.:3.743  
 Max.   :45                      Max.   :101.95   Max.   :4.468  
                                                                 
   MarkDown1        MarkDown2           MarkDown3        
 Min.   : -2781   Min.   :  -265.76   Min.   :  -179.26  
 1st Qu.:  1578   1st Qu.:    68.88   1st Qu.:     6.60  
 Median :  4744   Median :   364.57   Median :    36.26  
 Mean   :  7032   Mean   :  3384.18   Mean   :  1760.10  
 3rd Qu.:  8923   3rd Qu.:  2153.35   3rd Qu.:   163.15  
 Max.   :103185   Max.   :104519.54   Max.   :149483.31  
 NA's   :4158     NA's   :5269        NA's   :4577       
   MarkDown4          MarkDown5             CPI         Unemployment   
 Min.   :    0.22   Min.   :  -185.2   Min.   :126.1   Min.   : 3.684  
 1st Qu.:  304.69   1st Qu.:  1440.8   1st Qu.:132.4   1st Qu.: 6.634  
 Median : 1176.42   Median :  2727.1   Median :182.8   Median : 7.806  
 Mean   : 3292.94   Mean   :  4132.2   Mean   :172.5   Mean   : 7.827  
 3rd Qu.: 3310.01   3rd Qu.:  4832.6   3rd Qu.:213.9   3rd Qu.: 8.567  
 Max.   :67474.85   Max.   :771448.1   Max.   :229.0   Max.   :14.313  
 NA's   :4726       NA's   :4140       NA's   :585     NA's   :585     
 IsHoliday      
 Mode :logical  
 FALSE:7605     
 TRUE :585      
```

The final file called ‘sampleSubmission.csv’ contains two main columns: dates for each of the weeks in the study as well as a blank column that should be utilized to record predicted sales for that week based on the different models and techniques applied.

The results of the most accurate and efficient model have been recorded in this file and the final Power BI dashboard has been created based on these predicted values, in conformity with the ‘stores’ and ‘features’ dataset.

4.1 EXPLORATORY DATA ANALYSIS

It is crucial to have an in-depth understanding of the dataset that is used in this analysis to understand the models that would give the most accurate prediction. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough exam- ination is necessary to understand the underlying structure of the dataset and to draw conclusions or insight about the validity of our analysis.

The study is going to begin with a brief analysis of the available dataset to get a sense of the main characteristics and components that are relevant to the research. An exploratory data analysis is crucial to this study considering the numerous attributes that are a part of the dataset that will be essential when trying to draw insights and making predictions. As part of the exploratory data analysis, several visualizations have been created that will help us understand what it is that we are trying to achieve and to keep in mind the various attributes that we can use to improve results.

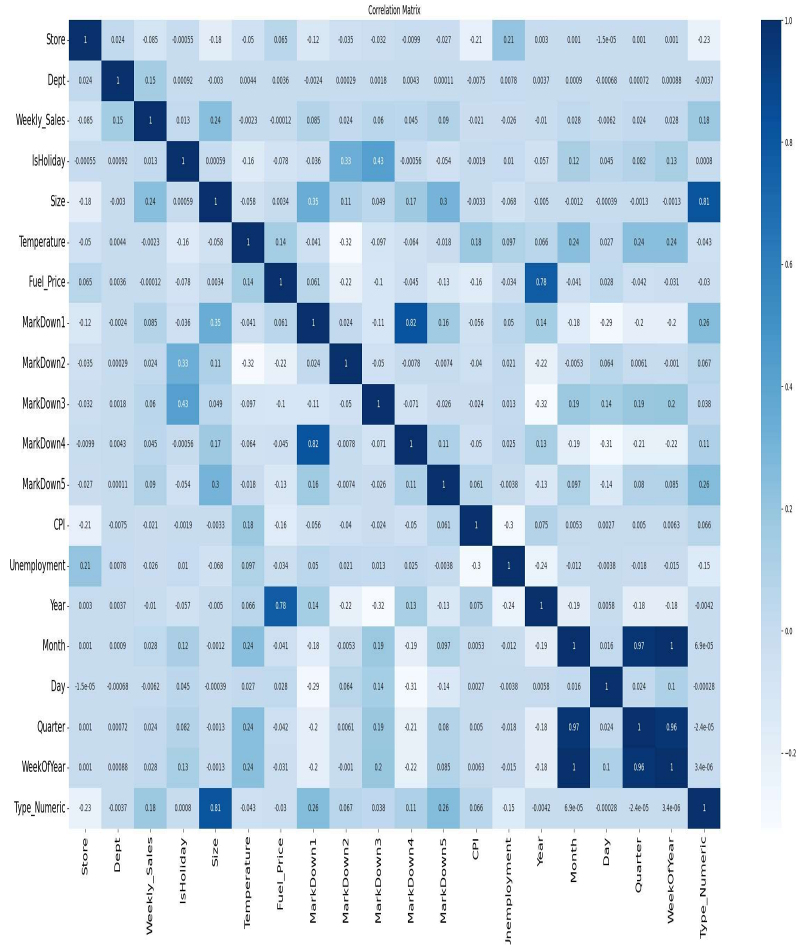
The EDA is like a primary investigation and tries to look at the relationships and nature of the different columns available to us. As part of this, the ‘inspectdf’ pack- age (Ellis, 2019) and the ‘glimpse’ package (Sullivan, 2019) have been used and imple- mented in R that will answer questions related to the number and nature of columns and rows in the dataset, missing values, distribution of numeric and categorical vari- ables, correlation coefficients, etc.

Several other packages like ‘ggplot2’, ‘matplotlib’, ‘seaborn’, and ‘plotly’ have also been used in this study to create visualizations that provide information about weekly sales by store and department, weekly sales on holidays versus on normal days, weekly sales based on region, store type and store size, average sales per year, change in sales as a result of factors like CPI, fuel price, temperature, and unemployment, etc in the form of heatmaps, correlation matrix (Kedia et al., 2013), histograms, scatterplots and several more. These visualizations are accompanied by brief descriptions that will discuss the findings and scope for potential modeling that will be performed in the next stages of this project.

4.2 CORELATION MATRIX

A correlation matrix describes the correlation between the various variables of a dataset. Each variable in the table is correlated to each of the other variables in the table and helps in understanding which variables are more closely related to each other (Glen, 2016).

With the numerous variables available through this dataset, it became imperative to study correlations between some of them. By default, this matrix also calculates correlation through Pearson's Correlation Coefficient (Johnson, 2021) that calculates the linear relationship between two variables, within a range of -1 to $+1$. The closer the correlation to $|1|$, the higher the linear relationship between the variables and vice versa.



	Store	Dept	Weekly_Sales	IsHoliday	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Year	Month	Day	Quarter	WeekOfYear	Type_Numeric
Store	1	0.024	-0.085	-0.00055	-0.18	-0.05	0.065	-0.12	-0.035	-0.032	-0.0099	-0.027	-0.21	0.21	0.003	0.001	-1.5e-05	0.001	0.001	-0.23
Dept	0.024	1	0.15	0.00092	-0.003	0.0044	0.0036	-0.0024	0.00029	0.0018	0.0043	0.00011	-0.0075	0.0078	0.0037	0.0009	-0.00068	0.00072	0.00088	-0.0037
Weekly_Sales	-0.085	0.15	1	0.013	0.24	-0.0023	-0.00012	0.085	0.024	0.06	0.045	0.09	-0.021	-0.026	-0.01	0.028	-0.0062	0.024	0.028	0.18
IsHoliday	-0.00055	0.00092	0.013	1	0.00059	-0.16	-0.078	-0.036	0.33	0.43	-0.00056	-0.054	-0.0019	0.01	-0.057	0.12	0.045	0.082	0.13	0.0008
Size	-0.18	-0.003	0.24	0.00059	1	-0.058	0.0034	0.35	0.11	0.049	0.17	0.3	-0.0033	-0.068	-0.005	-0.0012	-0.00039	-0.0013	-0.0013	0.81
Temperature	-0.05	0.0044	-0.0023	-0.16	-0.058	1	0.14	-0.041	-0.32	-0.097	-0.064	-0.018	0.18	0.097	0.066	0.24	0.027	0.24	0.24	-0.043
Fuel_Price	0.065	0.0036	-0.00012	-0.078	0.0034	0.14	1	0.061	-0.22	-0.1	-0.045	-0.13	-0.16	-0.034	0.78	-0.041	0.028	-0.042	-0.031	-0.03
MarkDown1	-0.12	-0.0024	0.085	-0.036	0.35	-0.041	0.061	1	0.024	-0.11	0.82	0.16	-0.056	0.05	0.14	-0.18	-0.29	-0.2	-0.2	0.26
MarkDown2	-0.035	0.00029	0.024	0.33	0.11	-0.32	-0.22	0.024	1	-0.05	-0.0078	-0.0074	-0.04	0.021	-0.22	-0.0053	0.064	0.0061	-0.001	0.067
MarkDown3	-0.032	0.0018	0.06	0.43	0.049	-0.097	-0.1	-0.11	-0.05	1	-0.071	-0.026	-0.024	0.013	-0.32	0.19	0.14	0.19	0.2	0.038
MarkDown4	-0.0099	0.0043	0.045	-0.00056	0.17	-0.064	-0.045	0.82	-0.0078	-0.071	1	0.11	-0.05	0.025	0.13	-0.19	-0.31	-0.21	-0.22	0.11
MarkDown5	-0.027	0.00011	0.09	-0.054	0.3	-0.018	-0.13	0.16	-0.0074	-0.026	0.11	1	0.061	-0.0038	-0.13	0.097	-0.14	0.08	0.085	0.26
CPI	-0.21	-0.0075	-0.021	-0.0019	-0.0033	0.18	-0.16	-0.056	-0.04	-0.024	-0.05	0.061	1	-0.3	0.075	0.0053	0.0027	0.005	0.0063	0.066
Unemployment	0.21	0.0078	-0.026	0.01	-0.068	0.097	-0.034	0.05	0.021	0.013	0.025	-0.0038	-0.3	1	-0.24	-0.012	-0.0038	-0.018	-0.015	-0.15
Year	0.003	0.0037	-0.01	-0.057	-0.005	0.066	0.78	0.14	-0.22	-0.32	0.13	-0.13	0.075	-0.24	1	-0.19	0.0058	-0.18	-0.18	-0.0042
Month	0.001	0.0009	0.028	0.12	-0.0012	0.24	-0.041	-0.18	-0.0053	0.19	-0.19	0.097	0.0053	-0.012	-0.19	1	0.016	0.97	1	6.9e-05
Day	-1.5e-05	-0.00068	-0.0062	0.045	-0.00039	0.027	0.028	-0.29	0.064	0.14	-0.31	-0.14	0.0027	-0.0038	0.0058	0.016	1	0.024	0.1	-0.00028
Quarter	0.001	0.00072	0.024	0.082	-0.0013	0.24	-0.042	-0.2	0.0061	0.19	-0.21	0.08	0.005	-0.018	-0.18	0.97	0.024	1	0.96	-2.4e-05
WeekOfYear	0.001	0.00088	0.028	0.13	-0.0013	0.24	-0.031	-0.2	-0.001	0.2	-0.22	0.085	0.0063	-0.015	-0.18	1	0.1	0.96	1	3.4e-06
Type_Numeric	-0.23	-0.0037	0.18	0.0008	0.81	-0.043	-0.03	0.26	0.067	0.038	0.11	0.26	0.066	-0.15	-0.0042	6.9e-05	-0.00028	-2.4e-05	3.4e-06	1

The heatmap/correlation matrix in Figure 22, created using the seaborn library in Python (Szabo, 2020) gives the following information:

- There is a slight correlation between weekly sales and store size, type, and de- partment
- There seems to be a negative correlation between weekly sales and temperature, unemployment, CPI, and fuel price. This could suggest that sales are not im- pacted by

changes in these factors

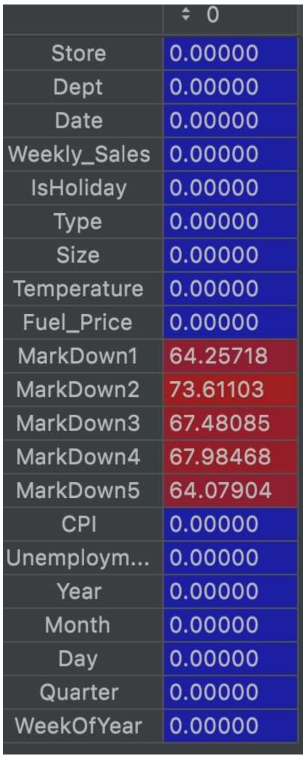
- Markdowns 1-5 also seem to have no distinct correlation with weekly sales, thus they are not as important a factor in the study

5. DATA CLEANING AND PREPROSESSING

The data contains 421,570 rows, with some store-specific departments missing a few too many weeks of sales. As observed in Figure 4, some columns in the features dataset contain missing values, however, after the features dataset is merged with the training dataset, the only missing values that exist are in the Markdown columns (as shown in figure 23).

After the extensive EDA, it was determined that these five markdown files, with missing values, have barely any correlation to the weekly sales for Walmart, hence these five columns have been eliminated from the subsequent training and testing dataset. Because the source already provides training and testing datasets, there is no need to create them for our study.

Because the main focus of this study is to accurately predict weekly sales for different Walmart stores, the previously modified ‘Date’, ‘Month’, ‘Quarter’, and ‘Day’ columns have been dropped and only the ‘Week of Year’ column has been used in the upcoming models.



	0
Store	0.00000
Dept	0.00000
Date	0.00000
Weekly_Sales	0.00000
IsHoliday	0.00000
Type	0.00000
Size	0.00000
Temperature	0.00000
Fuel_Price	0.00000
MarkDown1	64.25718
MarkDown2	73.61103
MarkDown3	67.48085
MarkDown4	67.98468
MarkDown5	64.07904
CPI	0.00000
Unemploym...	0.00000
Year	0.00000
Month	0.00000
Day	0.00000
Quarter	0.00000
WeekOfYear	0.00000

Data has been checked for inaccuracies, missing or out of range values using the ‘inspectdf’ package in R as part of the initial EDA. Columns with missing values have been dropped. The dataset contains information about weekly sales which was ini- tially broken down to acquire information about monthly as well as quarterly sales for our analysis, however, that information is not going to be utilized during the modeling process.

The boolean ‘isHoliday’ column in the dataset contains information about whether the weekly date was a holiday week or not. As observed in the EDA above, sales have been higher during the holiday season as compared to non-holiday season sales, hence the ‘isHoliday’ column has been used for further analysis.

Furthermore, as part of this data preprocessing step, I have also created input and target data frames along with the training and validation datasets that help accurately measure the performance of applied models. In addition, as part of this data prepro- cessing, feature scaling (Vashisht, 2021) has been applied to normalize different data attributes. This has primarily been done to unionize the independent variables in the training and testing datasets so that these variables will be centered around the same range (0,1) and provide more accuracy.

Also referred to as normalization, this method uses a simple min-max scaling tech- nique (implemented in Python using the Scikit-learn (Sklearn) library. The Weighted Mean

Absolute Error is one of the most common metrics used to measure accuracy for continuous variables (JJ, 2016).

A WMAE function has been created that provides a measure of success for the dif- ferent models applied. It is the average of errors between prediction and actual ob- servations, with a weighting factor. In conclusion, the smaller the WMAE, the more efficient the model.

6. MODEL SELECTION AND IMPLEMENTATION

Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the anal- ysis that has to be performed on the data (UNSW, 2020).

Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models have been used:

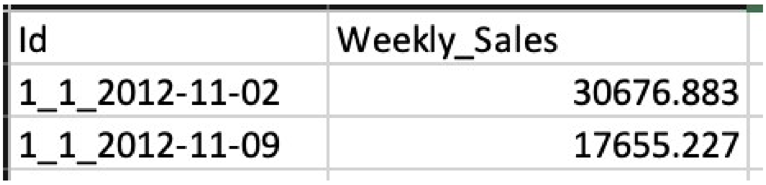
- Linear Regression
- Lasso Regression
- Gradient Boosting Machine
- Random Forest

Each of these methods have been discussed briefly in the upcoming report. For each of the models, why they were chosen, their implementation and their success rate (through WMAE) have been included.

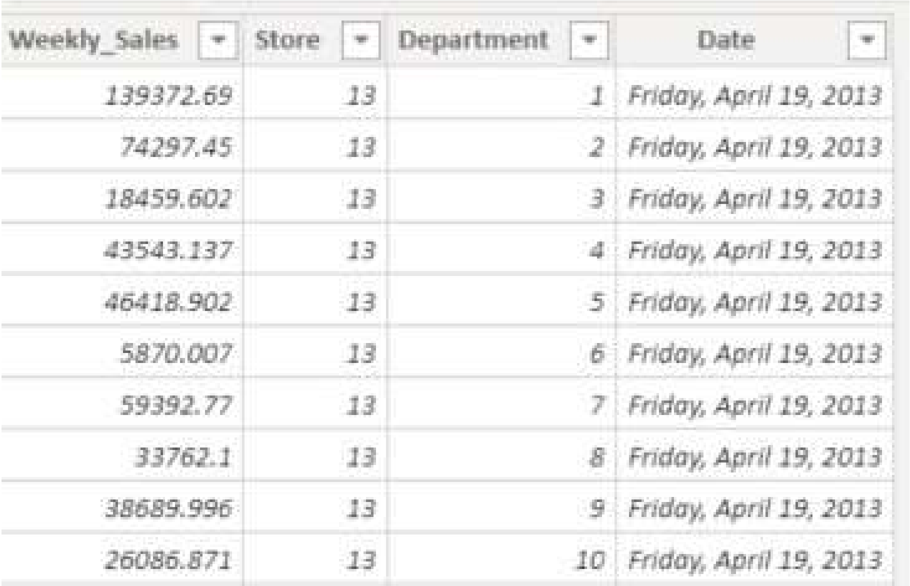
7. BUILDING BI DASHBOARD

As an end product, this Power BI dashboard is going to serve as the final product of this research. The dashboard contains detailed information about the original data related to the 45 Walmart stores as well as displays their respective predicted weekly sales. Most of the explorations that have been performed as part of the EDA will be included in this dashboard in the form of a story and users can filter data based on their requirements in the dashboard.

After the final predicted weekly sales are exported in the ‘sampleSubmissionFinal’ file, the id column is split to separate the store, department, and date information into different columns through Power BI data transformations (as shown in the figures be- low).



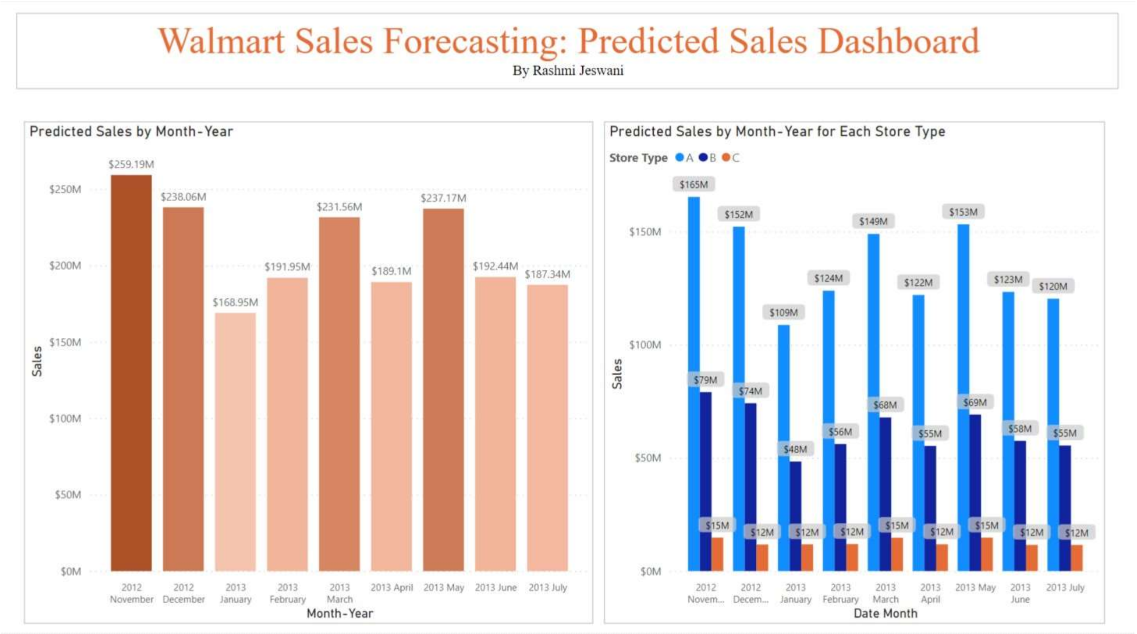
Id	Weekly_Sales
1_1_2012-11-02	30676.883
1_1_2012-11-09	17655.227

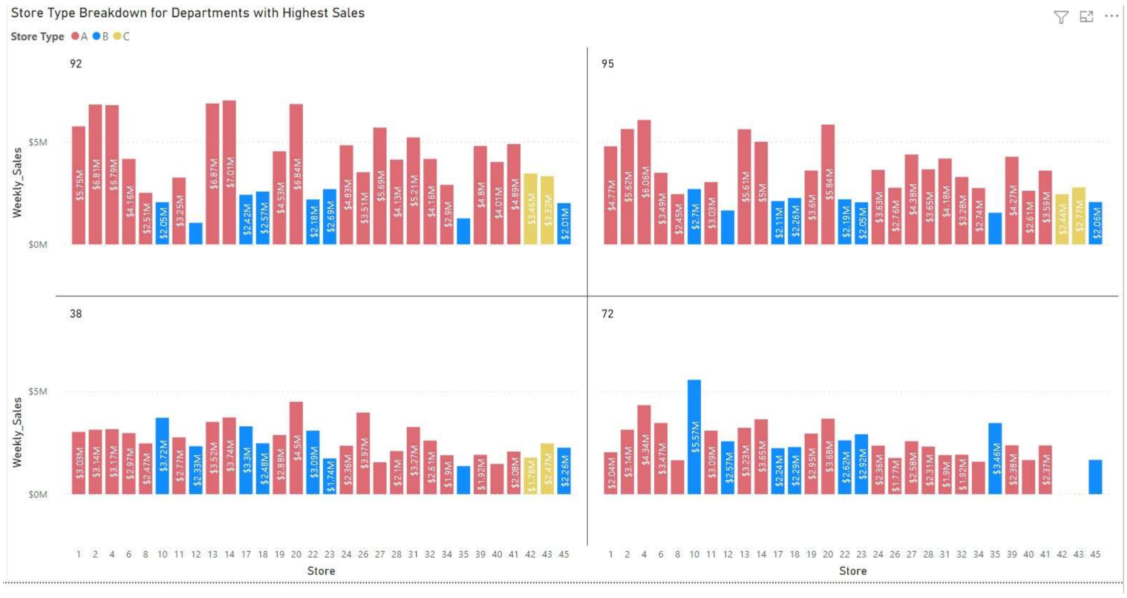


Weekly_Sales	Store	Department	Date
139372.69	13	1	Friday, April 19, 2013
74297.45	13	2	Friday, April 19, 2013
18459.602	13	3	Friday, April 19, 2013
43543.137	13	4	Friday, April 19, 2013
46418.902	13	5	Friday, April 19, 2013
5870.007	13	6	Friday, April 19, 2013
59392.77	13	7	Friday, April 19, 2013
33762.1	13	8	Friday, April 19, 2013
38689.996	13	9	Friday, April 19, 2013
26086.871	13	10	Friday, April 19, 2013

This file is then merged with the ‘stores’ file that contains information about the type and size of the store as well as holiday information. All these columns will be used to create several visualizations that track weekly predicted sales for various stores and departments, sales based on store size and type, etc. The dashboard also provides detailed information about stores and departments that generate the highest revenue and their respective store types. The PDF file contains brief information about all the visualizations created in the dashboard.

The dashboard can be found in the final submitted folder. If a user does not have access to Power BI, a PDF export of the entire dashboard is included along with the
.pbix file that contains all of the created visualizations and reports in the dashboard.
Some views of the dashboard created are included below:





8. CONCLUSION

The main purpose of this study was to predict Walmart's sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This study also aims to under- stand 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.

As observed through the exploratory data analysis, store size and holidays have a direct relationship with high Walmart sales. It was also observed that out of all the store types, Type A stores gathered the most sales for Walmart. Additionally, departments 92, 95, 38, and 72 accumulate the most sales for Walmart stores across all three store types; for all of the 45 stores, the presence of these departments in a store ensures higher sales. Pertaining to the specific factors provided in the study (temperature, unemploy- ment, CPI, and fuel price), it was observed that sales do tend to go up slightly during favorable climate conditions as well as when the prices of fuel are adequate. However, it is difficult to make a strong claim about this assumption considering the limited scope of the training dataset provided as part of this study. By the observations in the ex- ploratory data analysis, sales also tend to be relatively higher when the unemployment level is lower. Additionally, with the dataset provided for this study, there does not seem to be a relationship between sales and the CPI index. Again, it is hard to make a substantial claim about these findings without the presence of a larger training dataset with additional information available.

Interaction effects were studied as part of the linear regression model to identify if a combination of different factors could influence the weekly sales for Walmart. This was necessary because of the presence of a high number of predictor variables in the dataset. While the interaction effects were tested on a combination of significant vari- ables, a statistically significant relationship was only observed between the indepen- dent variables of temperature, CPI and unemployment, and weekly sales (predictor variable). However, this is not definite because of the limitation of training data.

9. REFERENCES

1. Bakshi, C. (2020). Random forest regression. https : / / levelup . gitconnected . com / random-forest-regression-209c0f354c84
2. Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to utilize linear regressions in predictive analytics. https://www.dummies.com/programming/big-data/data-science/ how-to-utilize-linear-regressions-in-predictive-analytics/
3. Baum, D. (2011). How higher gas prices affect consumer behavior. https : / / www . sciencedaily.com/releases/2011/05/110512132426.htm
4. Brownlee, J. (2016). Feature importance and feature selection with xgboost in python. https : / / machinelearningmastery . com / feature - importance - and - feature - selection-with-xgboost-in-python/
5. Chouksey, P., & Chauhan, A. S. (2017). A review of weather data analytics using big data. International Journal of Advanced Research in Computer and Communica- tion Engineering,https://doi.org/https://ijarcce.com/upload/2017/january-17/IJARCCE%2072.pdf
6. Crown, M. (2016). Weekly sales forecasts using non-seasonal arima models. http : // mxcrown.com/walmart-sales-forecasting/
7. Editor, M. B. (2013). Regression analysis: How do i interpret r-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
8. Ellis, L. (2019). Simple eda in r with inspectdf. https://www.r-bloggers.com/2019/05/ part-2-simple-eda-in-r-with-inspectdf/
9. Frost, J. (2021). Regression coefficients- statistics by jim. https://statisticsbyjim.com/ glossary/regression-coefficient/
10. Glen, S. (2016). Elementary statistics for the rest of us. https://www.statisticshowto. com/correlation-matrix/