A PROJECT ON

"WALMART STORES SALES PREDICTION"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYSIS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY
'Plot no R/2', Market yard road,
Behind hotel Fulera,Gultekdi
Pune – 411037.
MH-INDIA

SUBMITTED BY:

Dattatray Hake (49732)

Udit Deshmukh (49519)

UNDER THE GUIDENCE OF:
Mr. Girish Gaikwad
Faculty Member
Sunbeam Institute of Information Technology, PUNE.

CERTIFICATE

This is to certify that the project work under the title 'Walmart Stores Sales Prediction' is done by Dattatray Hake & Udit Deshmukh in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Girish Gaikwad
Project Guide

Mrs. Pradnya Dindorkar
Course Co-ordinator

Date:

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Pradnya Dindorikar (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Girish Gaikwad.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Dattatray Hake
DBDA May 2021 Batch,
SIIT Pune

Udit Deshmukh
DBDA May 2021 Batch,
SIIT Pune

.

TABLE OF CONTENTS

1. Introduction

1.1 Introduction And Objectives:

Walmart is a renowned retail corporation that operates a chain of hypermarkets. Here, Walmart has provided a data combining of 45 stores including store information and monthly sales. The data is provided on weekly basis. Walmart tries to find the impact of holidays on the sales of store. For which it has included four holidays weeks into the dataset which are Christmas, Thanksgiving, Super bowl, Labor day. Here we are owing to Analyze the dataset given. before doing that , let me point out the objective of this analysis. Our Main Objective is to predict sales of store in a week.

1.2 Why this problem needs To be Solved?

Holidays can create a huge impact on sales. So, if there is a good prediction on Sales then Walmart can calculate how much product to order during Holiday time. It will help in predicting which products needs to be purchased during the holiday time. As customers planning to buy something expects the products to be available immediately. And through prediction they can figure out which product will require at what time . Thus soar the trust of Customer on Walmart. This problem can also solve the issue of Marketing Campaigns. As Forecasting is often used to adjust ads and marketing campaigns and can influence the number of sales. Walmart runs several markdown events throughout the year. And these markdown event precede to the prominent holidays. So to solve the issue Walmart can organize such events more efficiently.

1.3 Dataset Information.

Stores.csv:

It has three columns.

Store: stores numbered from 1 to 45

Size : stores size has provided

type : store type has been provided ,there are 3 types — A,B and C .

Train.csv

It has five columns.

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

Test.csv: is same as train.csv except it does not have 'IsHoliday' Column.

Features.csv
It has eleven columns.
Store: the store number
Date: the week
Temperature: average temperature in the region.
Fuel_Price: cost of fuel in the region
MarkDown1–5: anonymized data related to promotional markdowns that Walmart is running. Markdown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA. Selected holiday markdown events are included in the dataset. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact.
CPI — the consumer price index
Unemployment : the unemployment rate.
IsHoliday: whether the week is a special holiday week

2. Problem Definition and Algorithm:

2.1 Problem Definition

The problem is quite straightforward. Data from Walmart stores accross the US is given, and it is up to us to forecast their weekly sales. The data is already split into a training and a test set, and we want to fit a model to the training data that is able to forecast those weeks sales as accurately as possible. In fact, our metric of interest will be the Mean Absolute Error and R2 score value.The metric is not very complicated. The further away from the actual outcome our forecast is, the harder it will be punished. Optimally, we exactly predict the weekly sales. This of course is highly unlikely, but we must try to get as close as possible.

2.2 Algorithm Definition

Linear regression: is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Ridge Regression: Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

Lasso Regression: Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.Lasso Regression uses L1 regularization technique,It is used when we have more number of features because it automatically performs feature selection.

Random forest: is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Decision Tree: algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

XGBoost: or extreme gradient boosting is one of the well-known gradient boosting techniques (ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. It is open-source software. Earlier only python and R packages were built for XGBoost but now it has extended to Java, Scala, Julia and other languages as well.

3.Experimental Evaluation:

3.1 Methodology:

The objective of this project is to predict the weekly sales of wallmart in US. The data set is contained from Kaggle and has 3 csv files namely features, stores and train. The data is merged to obtain one master datafile and then the data preprocessing is carried out.

Loading in raw data

```
features_df = pd.read_csv("features.csv")
stores_df = pd.read_csv("stores.csv")
walmart_df = pd.read_csv("train.csv")
master_df =walmart_df.merge(stores_df, how='left').merge(features_df, how='left')
print(master_df.shape)
master_df.head()
```

Preprocessing:

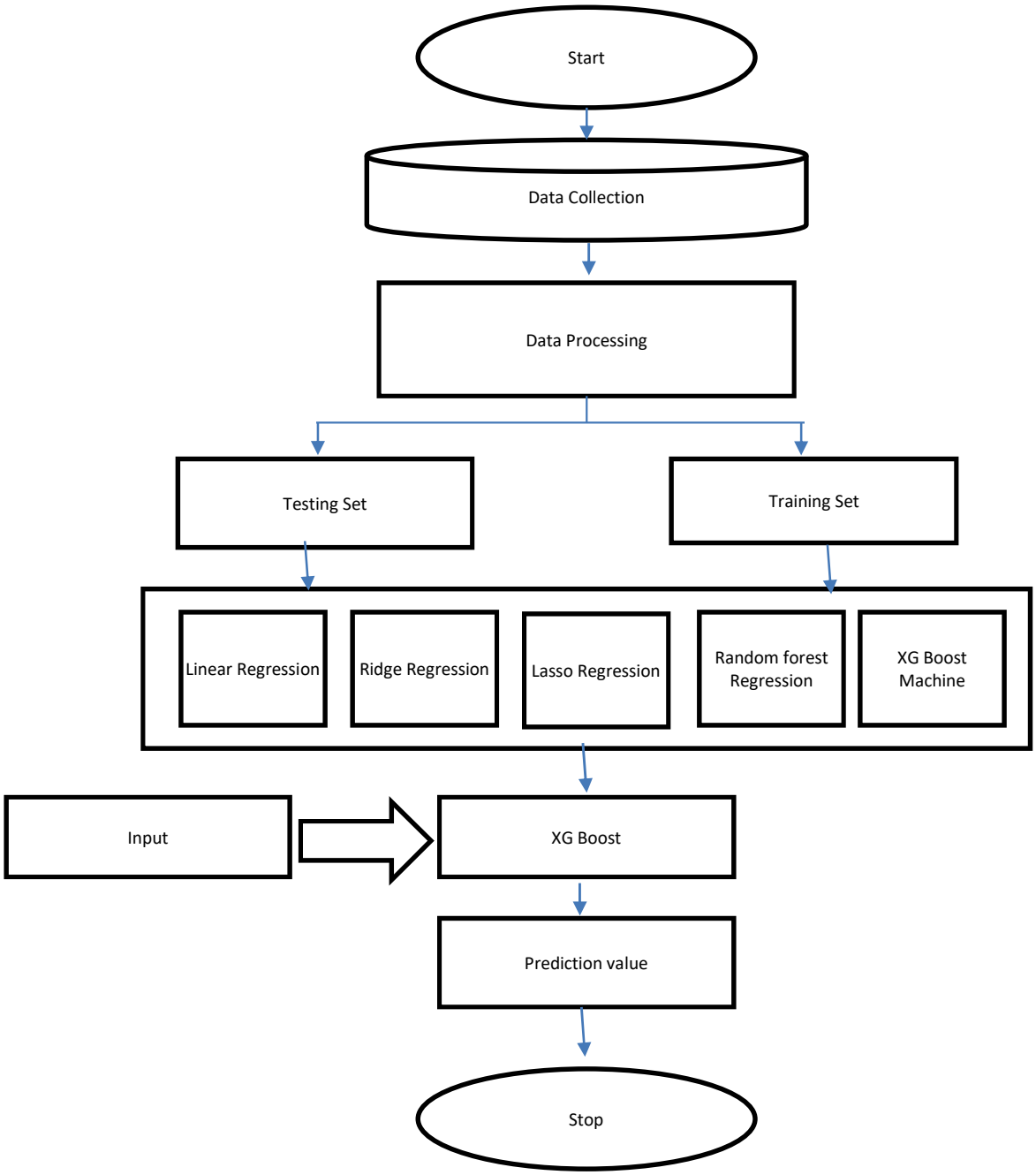
The sales are given for Years 2012-2012 on weekly basis. This data was split to extract information for year, month and week.

```
master_df['Date'] = pd.to_datetime(master_df['Date'], format='%Y-%m-%d')
master_df['Week_Number'] = master_df['Date'].dt.week
master_df['Month'] = master_df['Date'].dt.month
master_df["Year"] = master_df["Date"].dt.year
```

The data had several missing values and needed to be cleaned. The missing values in ‘Markdown1-5’needed to be cleaned. Since the number of missing values were significant, they were not removed but were replaced with zero.

```
print(master_df.isna().sum())
missing_values = master_df.isna().sum()
master_df['MarkDown1'] = master_df['MarkDown1'].fillna(0)
master_df['MarkDown2'] = master_df['MarkDown2'].fillna(0)
master_df['MarkDown3'] = master_df['MarkDown3'].fillna(0)
master_df['MarkDown4'] = master_df['MarkDown4'].fillna(0)
master_df['MarkDown5'] = master_df['MarkDown5'].fillna(0)
master_df.isna().sum()
```

Flow Diagram :



3.2 Exploratory Data Analysis

The popularity of each store is plot with the help of a pie chart (fig 2). From the figure we can infer that type A store has the highest popularity followed by type B store and type C store has the least popularity.

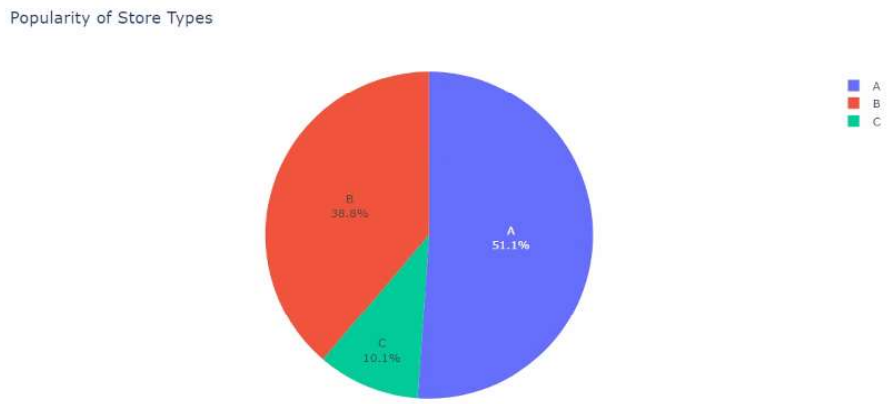


Fig 2: Pie- chart showing store- type wise popularity

The average sale for each store- type is visualized using bar plot (fig 3). From the figure we can infer that type A store has the highest average sales followed by type B store. The type C store has least average sale among the three.

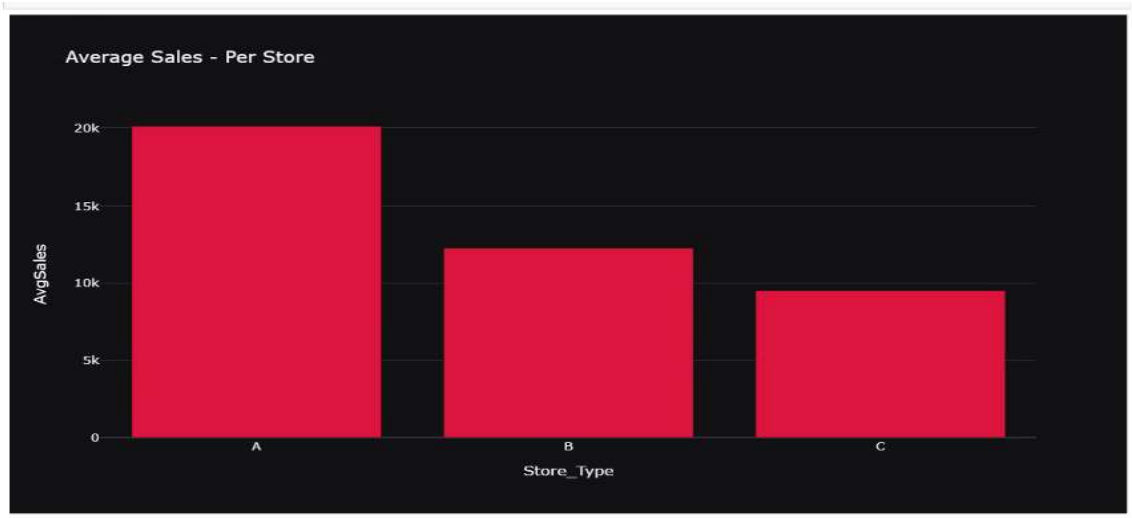


Fig 3: Store- type Vs Average sales

The average sale for each store- type is plotted for each year (fig 4). The plot

shows that Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided

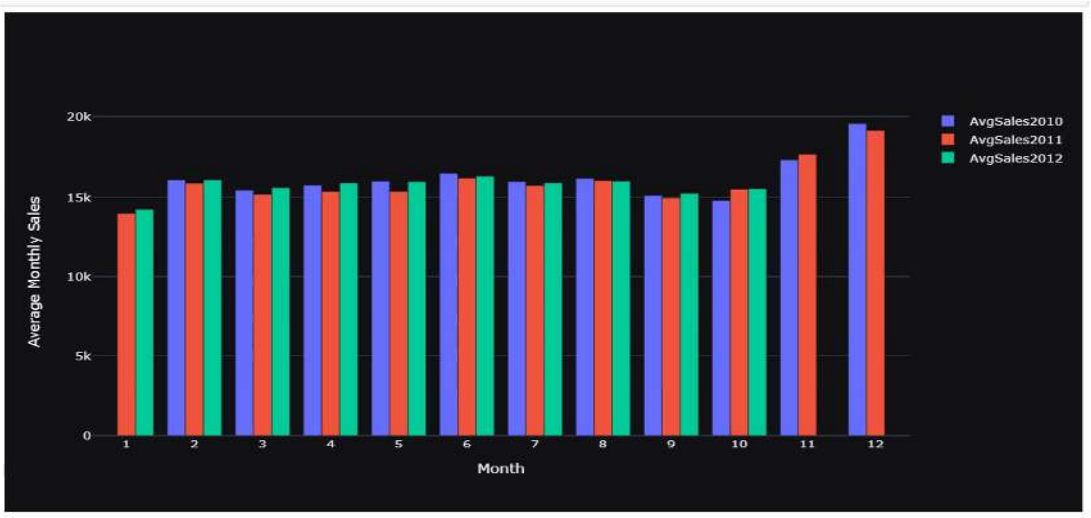


Fig 4: Average monthly sale per year

The average weekly sale per year is plotted using scatter and line plot (fig 5). Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided or any special even.

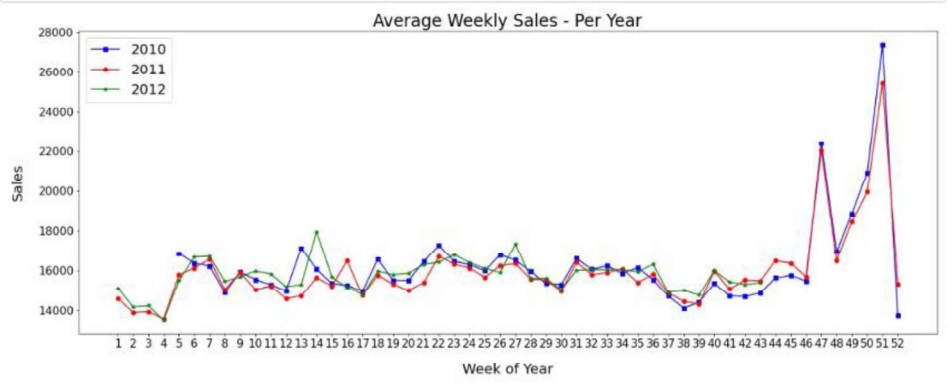


Fig 5: The average weekly sale per year is plotted using scatter and line plot

The average store sales per year is plotted. Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided week of Thanks giving the highest sales in all the 3 years

Average department wise sale per year is plotted for each year. Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data
From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided week of Thanks giving among the 45 stores which have highest average sales

Fig 6: Analysis of sales on holidays and working days

The data is analysed for sales on holidays and other working days. This shows that the sales are comparatively higher on holidays. This information is useful to further improve the store sales. Only 7 percent of the weeks in the data are the holiday weeks Despite being the less peecentage of holiday weeks the sales in the holidays week are on the average higher than in the non-holiday weeks

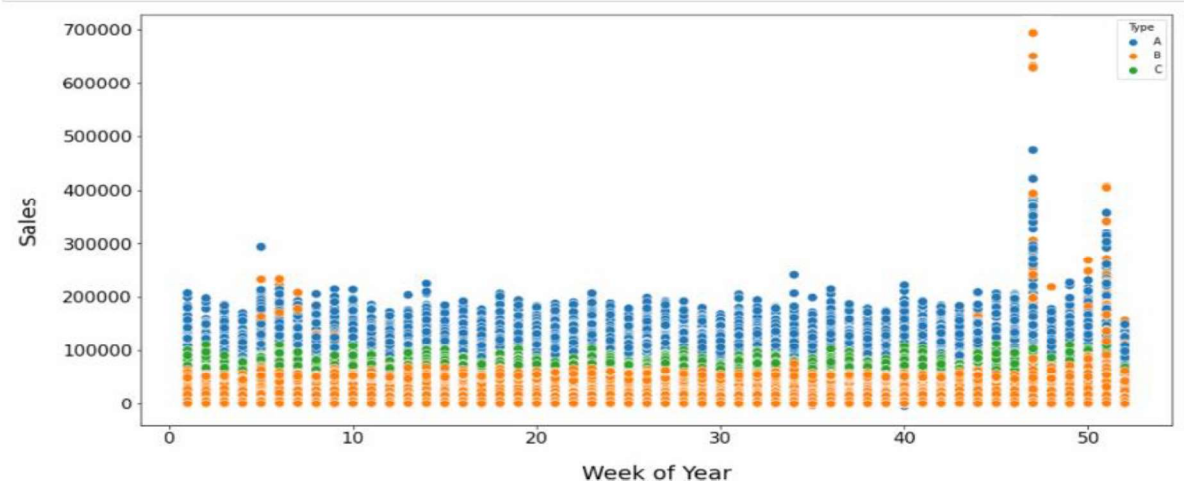


Fig 7: Week of the year vs sales

The sales for each week is plotted for 3 years. This shows only a slight relationship as the weekly sales increased towards the end of the year.

4. Results and discussion:

Linear regression, Lasso regression, ridge regression, random forest, decision tree and gradient boosting machine algorithm were used to predict the weekly saes of wallmart. Among the given algorithms Gradient Boosting Machine algorithm was the best performing one as it provided the highest R2 score of 0.94.

```
from xgboost import XGBRegressor

XGBoost_model = XGBRegressor()
XGBoost_model.fit(x_train, y_train)

y_prediction = XGBoost_model.predict(x_test)

MAE = mean_absolute_error(y_test, y_prediction)
print(f"MAE = {MAE}")

R2 = r2_score(y_test, y_prediction)
print(f"R2 = {R2}")
```

MAE : 1940.99
R2 Score : 0.94

5. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

6. Future work And Conclusion

6.1Future Work:

Walmart can analyze the entire store data across US to arrive at an even more accurate prediction. They can analyze the inventory data as well to optimize their inventory. They can analyze the sales targets and incentives that are given for employees to arrive at achievable sales targets for employees to motivate them better.

6.2 Conclusion:

- Type 'A' stores are more popular than 'B' and 'C' types
- Type 'A' stores outclass the 'B' and 'C' types in terms of size and the avergae weekly sales
- Weekly Sales are effected by the week of year. Holiday weeks witnessed more sales than the non-holiday weeks. Notables are Thanksgiving and Christmas weeks
- Size of the store is a major contributing factor in the weekly sales

- Sales are also dependent on the department of the store as different departments showed different levels of weekly sales

- Among the trained models for predicting the future sales, Gradient Boosting Machine performs the best.