Walmart Store Sales Prediction



**INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT AKURDI,**

**PUNE**

Documentation On

**“Walmart Store Sales Perdiction”**

PG-DBDA SEP 2020

Submitted By:

**Group No: G-17**

Shivani Saratkar - 1349

Ankita ugale -1354

Mr.Parshant Karhale Mr.Akshay Tilekar

**Centre Co-ordinate Project Guide**

**1.Acknowledgement**

First and Foremost we thank to almighty God for giving us the support, strength, positive spirit and talent to do this project.

“The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible”.

“This work is the result of inspiration, motivation, knowledge, interest, support, guidance, cooperation and efforts by many people at different levels. We are indebted to all of them”.

We would also like to take this opportunity to acknowledge the valuable contributions made by our family members by supporting and motivating us in every walk of life.

We are thankful to **Mr. Prashant Karhale**, Centre Co-ordinator **IACSD CDAC, Akurdi** Pune for providing the opportunity, infrastructure and facilities for entire work.

We would like to express our great appreciation to our project Guide **Mr. Akshay Tilekar**, Internal Project Guides **Mr. Rahul**, and **Mr. Manish** for their valuable and constructive suggestions during the planning and development of this project.

We also thank all staff members from the IACSD who in some way or other Helped us in completion of this project.

We cannot conclude our acknowledgement without expressing our thanks to our friends who helped us directly or indirectly during the course of this project.

Feedback for improving the contents of the report would be more than welcome**TABLE OF CONTENTS**

1. Introduction

1.1. Introduction And Objectives

1.2. Why this problem needs To be Solved?

1.3. Dataset Information

2. Problem Definition and Algorithm

2.1 Problem Definition

2.2 Algorithm Definition

3. Experimental Evaluation

3.1 Methodology/Model

3.2 Exploratory Data Analysis

3.3 Flow Diagram

4. Results And Discussion

5. GUI

6. Future Work And Conclusion

6.1 Future Work

6.2 Conclusion

**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.

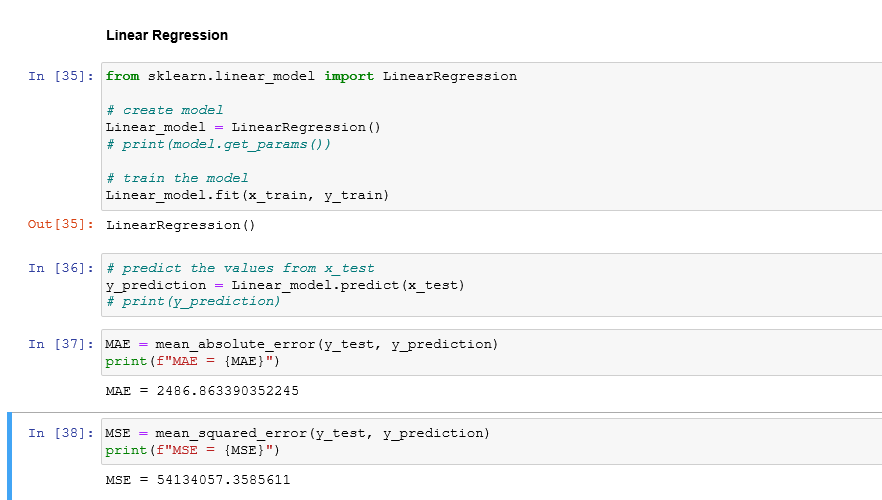
**2. Problem Definition and Algorithm:**

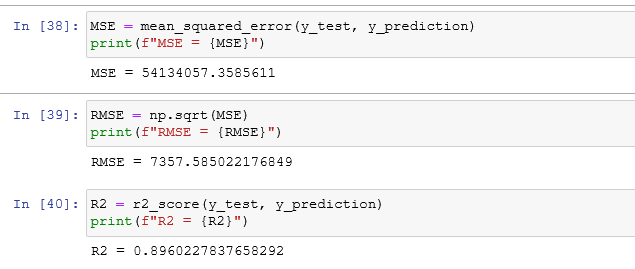
**2.1 Problem Definition**

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**

**1.Linear regression:** Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explotary variables (also known as dependent and independent variables). In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

**Code Snippet:** 



**Feature and Advantages of Linear Regression:**

* Linear Regression is simple to implement and easier to interpret the output coefficients.
* Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

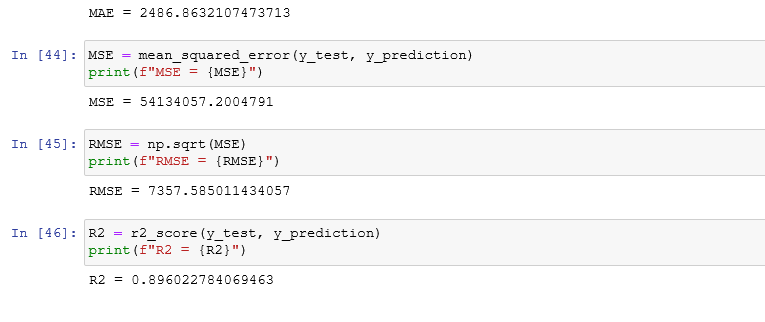
**Disadvantages of Linear Regression:**

* On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
* Linear regression looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

**2.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.

**Code snippets :**





**Advantages of Ridge Regression**

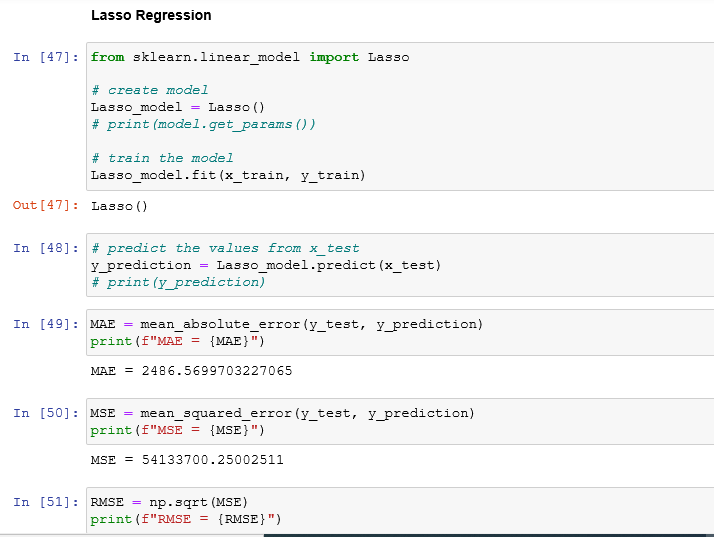
* Ridge regression solves the problem of overfitting , as just regular squared error regression fails to recognize the less important features and uses all of them , leading to overfitting.
* Ridge Regression adds a slight bias, to fit the model according to the true values of the data
* In datasets where we have the number of features(n) larger than the number of training examples(m), ridge regression becomes crucial, as it performs significantly better than regular sum of squares method.
* The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

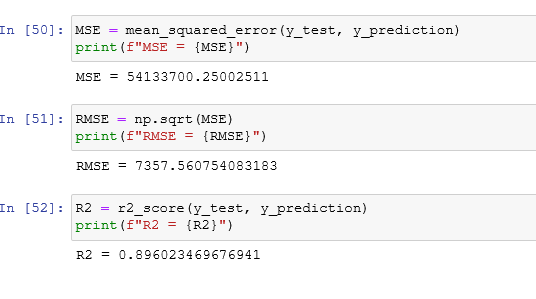
**Disadvantages of Ridge Rgression**

* Ridge regression , although improving the test accuracy, uses all the input featuresin the dataset, unlike step-wise methods that only select a few important features for regression.
* Ridge Regression reduces the coefficient theta to very low values if the feature is not important, but it won’t completely make then zero, hence still using the feature in our model .
* Lasso regression overcomes this drawbacks.

**3.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 procedureencourages 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.

**Code Snippet :**





**Advantages of lasso regression**

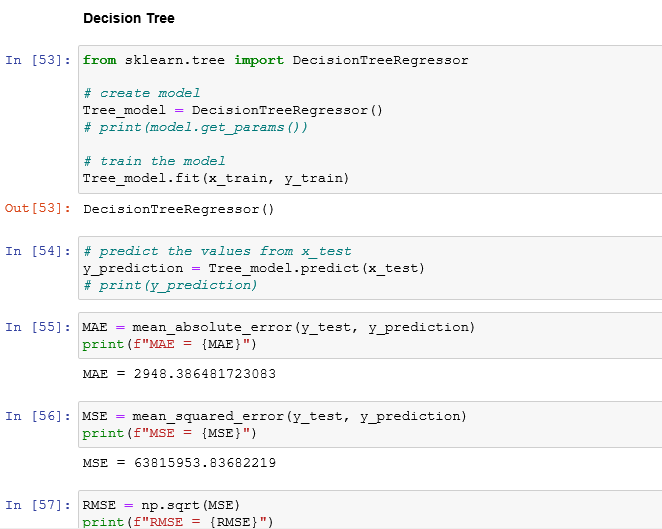
* As any regularization method, it can avoid overfitting . it can be applied even when number of features is larger than number of data.
* It can do features selction .
* It is faast in terms of inference and fitting .

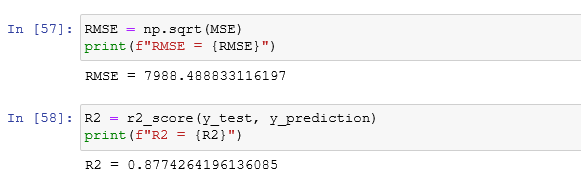
**Disadvantages of lasso regression**

* The model selected by lasso is not stable. For example, on different bootstrapped are selected can be very different.
* The model selection result is not intuitive to interpret. For example, why lasso select a feature.
* When there are highly correlated features, lasso may randomly select one of them of part of them.

**4.Decision Tree:** This 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.

**Code Snippet :**





**Advantages of Decision Tree :**

* Decision Tree can be used for both classification and regression problems.
* Decision Tree can handle both **continuous and categorical variables**.
* Decision Tree can automatically **handle missing values**.
* Decision Tree is usually **robust to outliers** and can handle them automatically.

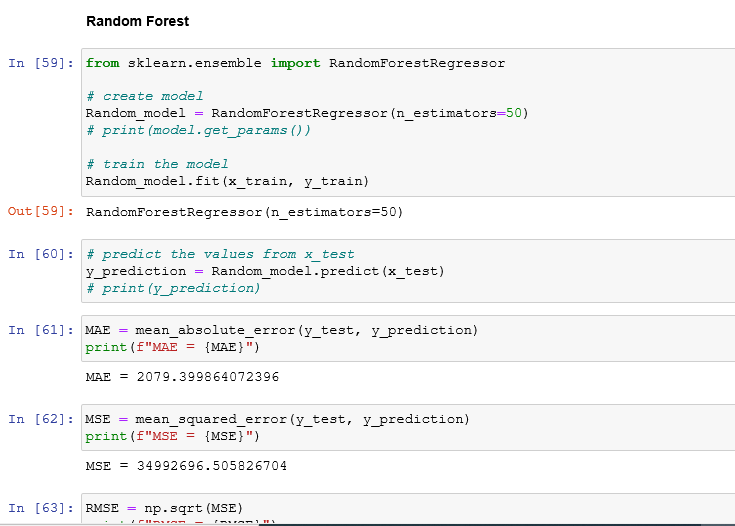
Disadvantages of DT :

* **Overfitting:** This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.

**5.Random Forest :**

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

Code snippets :





**Advantages :**

* It can perform both regression and classification tasks.
* A random forest produces good predictions that can be understood easily.
* It can handle large datasets efficiently.
* The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

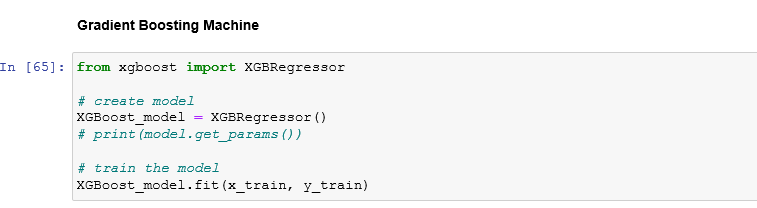
**Disadvantages :**

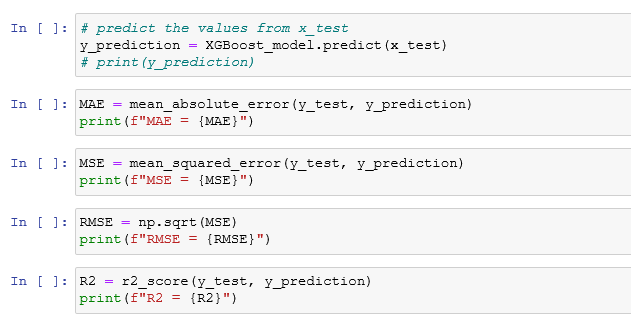
* When using a random forest, more resources are required for computation.
* It consumes more time compared to a decision tree algorithm

**5.XGBoost or extreme gradient boosting:**  is one of the well-known gradient boosting techniques (ensemble) having enhanced performance and speed in treebased (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 opensource 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.

Code Snippets :





**Advantages :**

**Regularization:** XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why, XGBoost is also called regularized form of GBM (Gradient Boosting Machine).  
  
While using Scikit Learn libarary, we pass two hyper-parameters (**alpha**and **lambda**) to XGBoost related to regularization. **alpha**is used for L1 regularization and **lambda**is used for L2 regularization.  
  
**Parallel Processing:** XGBoost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model.  
  
While using Scikit Learn libarary, **nthread**hyper-parameter is used for parallel processing. **nthread**represents number of CPU cores to be used. If you want to use all the available cores, don't mention any value for **nthread**and the algorithm will detect automatically.  
  
**Handling Missing Values:** XGBoost has an in-built capability to handle missing values. When XGBoost encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node. It then does the same when working on the testing data.

**Disadvantages :**

**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 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()

**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.

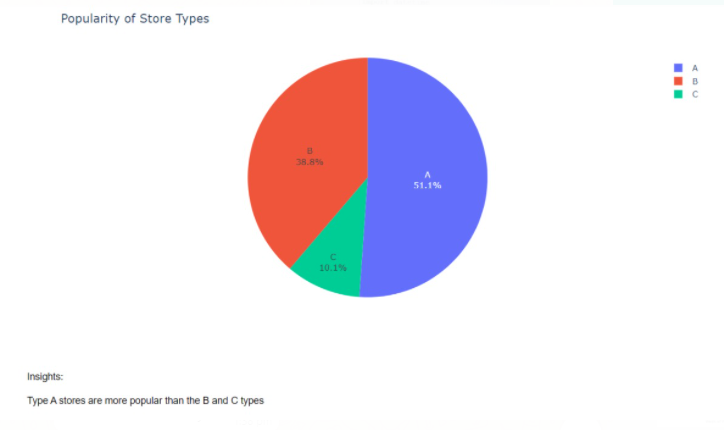


Fig 2. Pie chart- Showing store- type wise popularity

The average sale for each store- type is visualized using bar plot(fig 3.) from the fig 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 provide

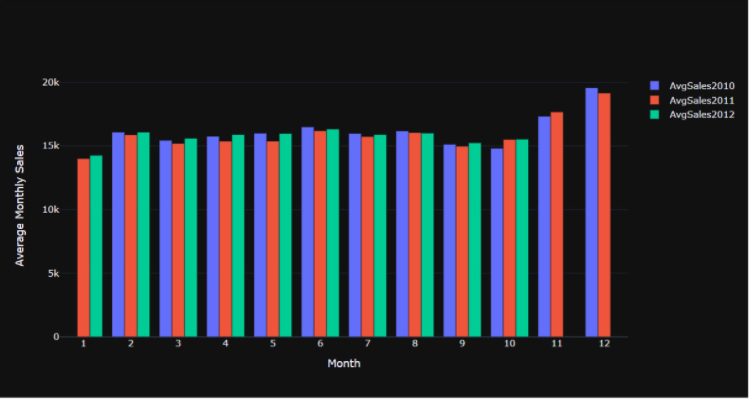


Fig 4 . Average monthly sale per year

The average weekly sale per year is plotted using scatter and line plot (fig 5). Month of January witnessed thw lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data from feburary till October the weekly salesy 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

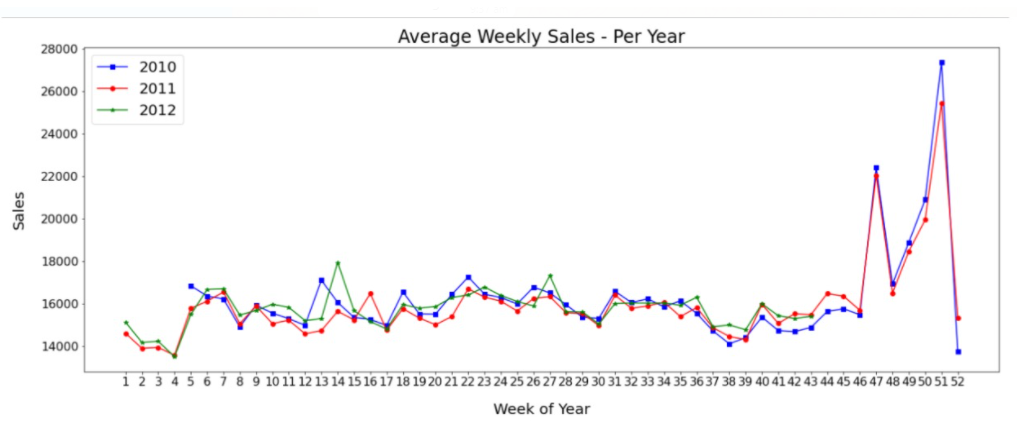


Fig 5. The average weekly sale per year in plotted using scatter and line plot

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 provide week of thanks 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 provide week of thanks giving among the 45 stores which have highest average sales

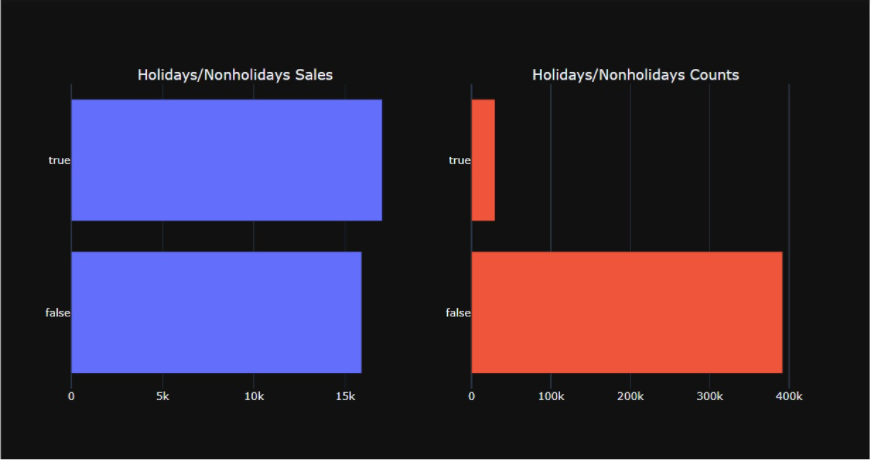


Fig 6. Analysis of sales on holidays and working days

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 improves the store sales. Only 7 percent of the weeks in the data are the holidays weeks despite being the less percentage of holiday weeks the sales in the holidays week are on the average higher than in the non-holiday weeks

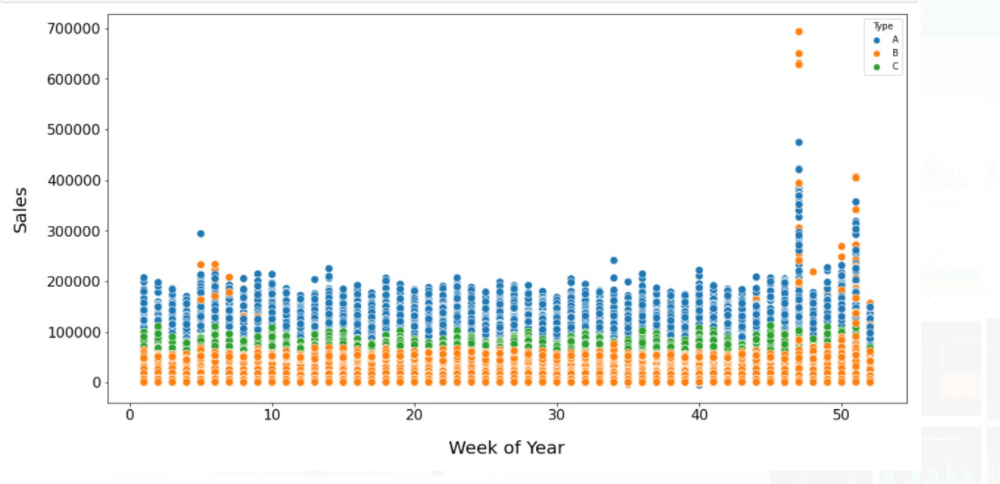
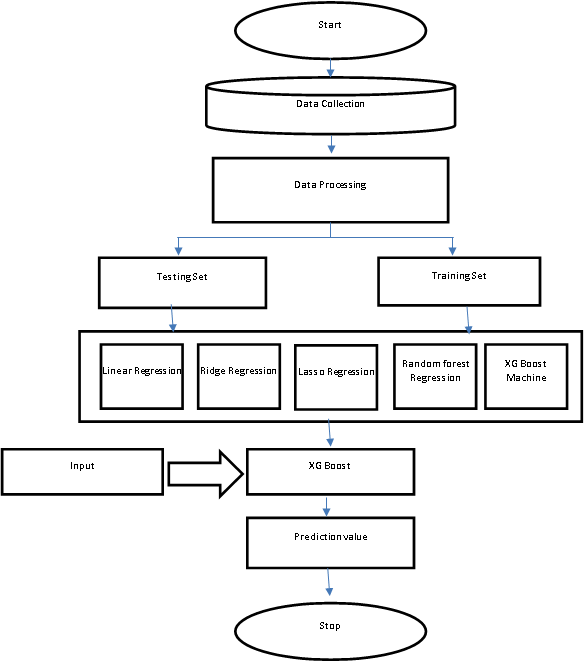


Fig 7. Week of the year vs sales

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.

**3.3 Flow Diagram :**

****

**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}")

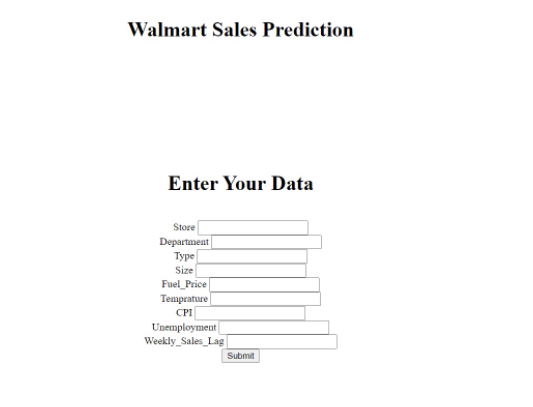
MAE : 1940.99

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**

**6.1 Future Work**

* Modifying date feature into days, month, weeks.
* The dataset includes special occasions i.e Christmas, pre-Christmas, black Friday, Labour day, etc. On these days people tend to shop more than usual days. So adding these as a feature to data will also improve accuracy to a great extent.
* Also there are a missing value gap between training data and test data with 2 features i.e. CPI and Unemployment. If that gap is reduced then also performance can be improved.

**6.2 Conclusion**

1. Type 'A' stores are more popular than 'B' and 'C' types
2. Type 'A' stores outclass the 'B' and 'C' types in terms of size and the avergae weekly sales
3. 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
4. Size of the store is a major contributing factor in the weekly sales
5. Sales are also dependent on the department of the store as different departments showed different levels of weekly sales
6. Among the trained models for predicting the future sales, Grdient Boosting Machine with tuned hyperparameters performs the best