“SUPERMARKET SALES FORECASTING”

Submitted
By

Rajashri Chitti (1032212101)

List of Tables

List of Figures

ABSTRACT

In the retail industry, sales forecasting is a critical task. In this way, the use of machine-learning models for sales forecasting was investigated. The idea is to look at the scientific literature and see whether there are any advantages over traditional statistical methods. This study has performed a rigorous investigation and examination of predictive models in order to improve future sales projections. Millions of evaluations are written every day, making it difficult for a consumer to decide whether or not to purchase a product. For product makers, sifting through such a large number of viewpoints is difficult and time-consuming. This study looks into the topic of categorizing reviews based on their overall semantic content (positive or negative) The suggested approach in this work employs three machine learning techniques, namely Linear Regression, Random Forest model, K-neighbours regressor, XGboost Model. Implementation is performed on Walmart store sales dataset. We have implemented custom deep learning neural network and compared with various machine learning models to check the performance of model. We have achieved an accuracy of 97.120% for custom deep learning neural network. It has been observed that Random Forest regressor model, XGboost model and Deep learning neural network performed better than other models and gave almost same accuracy.

CHAPTER 1: INTRODUCTION

Forecasting sales has always been a major focus. For all vendors to maintain the efficacy of marketing organisations, an efficient and ideal forecasting method has become necessary. Manual infestation of this activity might result in significant errors, resulting in poor organisation management, and it would also be time consuming, which is not acceptable in today's fast-paced environment. The business sectors, which are actually required to create suitable quantities of items to meet the overall needs, account for a large portion of the global economy.

The main objective of business sectors is to target the market audience. It is therefore important that the company has been able to achieve this objective by employing a system of forecasting. The process of forecasting involves analyzing data from various sources such as market trends, consumer behavior and other factors. This analysis would also help the companies to manage the financial resources effectively. The forecasting process can be used for many purposes, including: predicting the future demand of the products or service, predicting how much of the product will be sold in a given period.

This is an area where machine learning can be quite useful. Machine learning is the field in which computers learn to outperform humans at specific tasks. They are utilized to perform a specific task in a logical manner in order to achieve superior results for the present society's advancement. Machine learning is built on the foundation of mathematics, which allows numerous paradigms to be established in order to achieve the best results. Machine learning has also proven to be beneficial in the area of sales forecasting. It aids in more precise forecasting of future sales.

In this project, we have implemented Custom Deep Learning Neural Network which is used for sales forecasting on Walmart store sales dataset which is available on Kaggle.

The goal is to predict the pattern of weakly sales and the quantities of products to be sold based on some key characteristics gleaned from the raw data and to compare it with other models to see which model gives better results. To acquire a thorough understanding of the data, analysis and exploration of the collected data were also performed. At each critical stage of marketing strategy, analysis would assist business organizations in making a probabilistic conclusion.

CHAPTER 2: OBJECTIVE

- ✓ To analyze the content of Walmart dataset that lists which historical sales data of around 45 shops.
- ✓ To predict the pattern of weakly sales and the quantities of products to be sold based on some key characteristics gleaned from the raw data and to compare it with other models to see which model gives better results.
- ✓ To acquire a thorough understanding of the data, analysis and exploration of the collected data were also performed. At each critical stage of marketing strategy, analysis would assist business organizations in making a probabilistic conclusion.

CHAPTER 3: IMPLEMENTATION DETAILS

CHAPTER 3.1: DATASET DETAILS

Data frame Dimensions: (406829,8)

This data collection contains historical sales data for 45 Walmart shops around the country. Participants must forecast sales for each department in each store, which has several departments.

The complete dataset is divided into three parts: -

- train.csv — This is the historical training data, which covers to 2010–02–05 to 2012–11–01.
- features.csv — This file contains additional data related to the store, department, and regional activity for the given dates.
- stores.csv — This file contains anonymized information about the 45 stores, indicating the type and size of the store.

train.csv - CSV Data file containing following attributes

Store

Dept

Date

Weekly_Sales

IsHoliday

115064 Data rows

stores.csv - CSV Data File containing following attributes

Store

Type

Size

45 Data rows

features.csv - CSV Data file containing following attributes

Store

Date

Temperature

Fuel_Price

MarkDown1

MarkDown2

MarkDown3

MarkDown4

MarkDown5

CPI

Unemployment

IsHoliday

8190 Data rows

CHAPTER 3.2: IMPLEMENTATION STEPS

The steps of implementation are as given below:

1. Data Preprocessing
2. Data Exploration
3. Data Normalization
4. Correlation between features
5. Feature Importance
6. Data Splitting to train and test
7. Fitting the different machine learning models
    a. Linear Regression model
    b. Random Forest Regressor model
    c. K nearest neighbour Regressor model
    d. XGBoost model
    e. Custom Deep Learning Neural Network model

Data pre-processing:

In data pre-processing step, we are finding the missing values (filled all the null and missing values with medians of the respective columns. The we merged the datasets into one final dataset.

So, we got total data rows: 421570 and have 15 attributes in total. Then we have identified the Outliers and abnormalities in the data set.

In the outlier detection step, markdowns were summed into Total_markdown. Also outlier was removed using z-score. After outlier's removal, there were 375438 rows and 20 columns in the data set. Plot o of Negative and Zero weekly sales have been observed in the following Fig 3.2.1

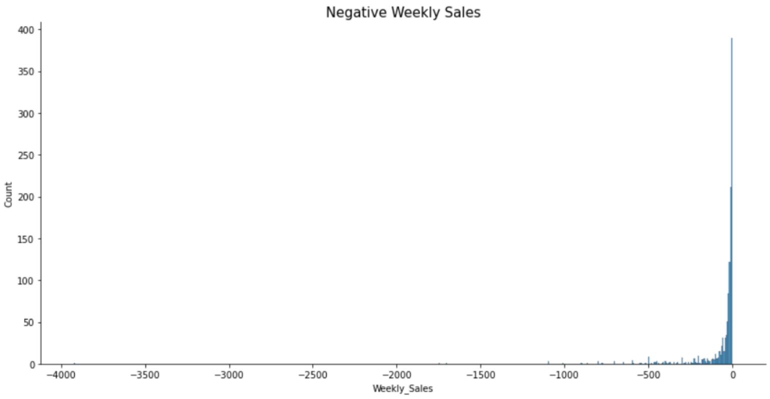


Figure 3.2.1 Plot of Negative and Zero weekly sales

Data Exploration:

In data exploration, graphs have been plotted like average monthly sale, monthly sale for each year, average weekly sale store wise, average weekly sale per department for better analysis. Also effect of temperature and distribution of holiday was observed using various plots as given below in Fig. 3.2.2, Fig. 3.3.3, Fig. 3.2.4, Fig. 3.2.5, Fig. 3.2.6, Fig. 3.2.6, Fig. 3.2.7.

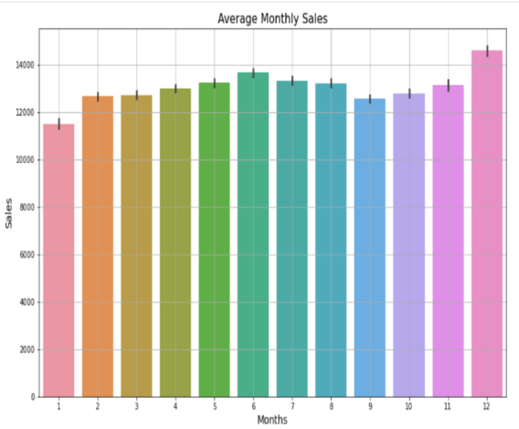


Figure 3.2.2 Average monthly sales

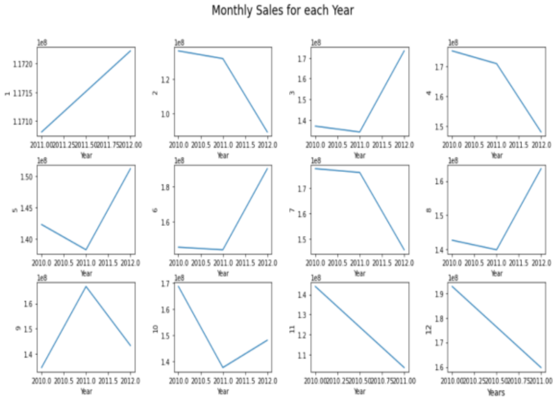


Figure 3.2.3 Monthly sales for each year

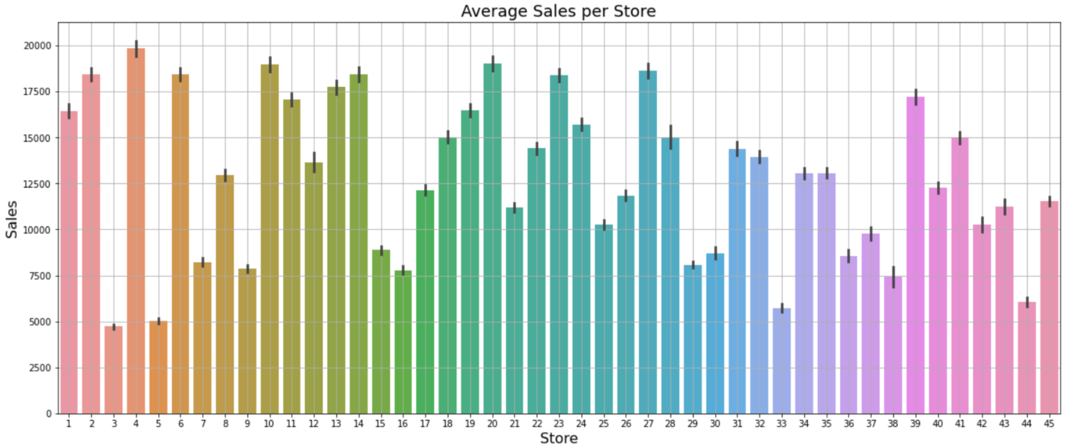


Figure 3.2.4 Average sales per store

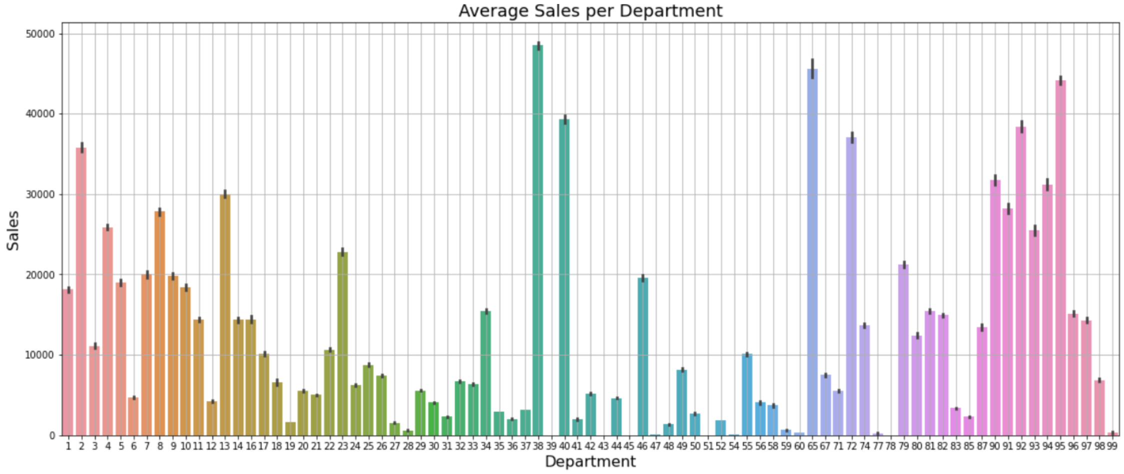
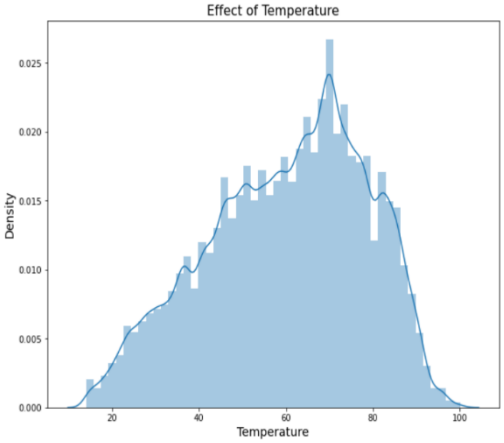


Figure 3.2.5 Average sales per department



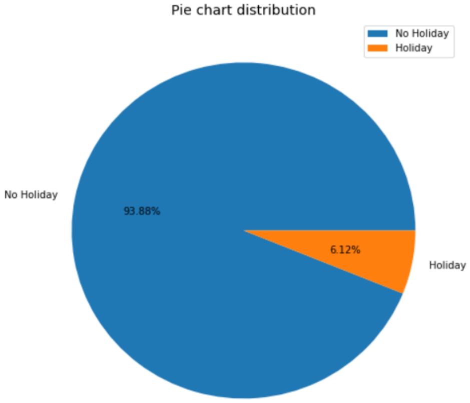


Figure 3.2.6 Effect of temperature

Figure 3.2.7 Pie chart distribution

The representation of categorical data can be more expressive with a one-hot encoding therefore One-hot encoding is performed for Store, Dept, Type columns. Many machine learning algorithms are unable to operate directly with categorical data. The categories must be numerically transformed.

Data normalization:

Data Normalization is a data preparation technique that is frequently used in machine learning. Normalization is the process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values. Every dataset does not need to be normalised for machine learning. So here MinMaxScalar is used to scale the values between 0 to 1.

Correlation between features:

To find correlation between features in the dataset, correlation matrix can be represented as can be seen below in Fig. 3.2.8.

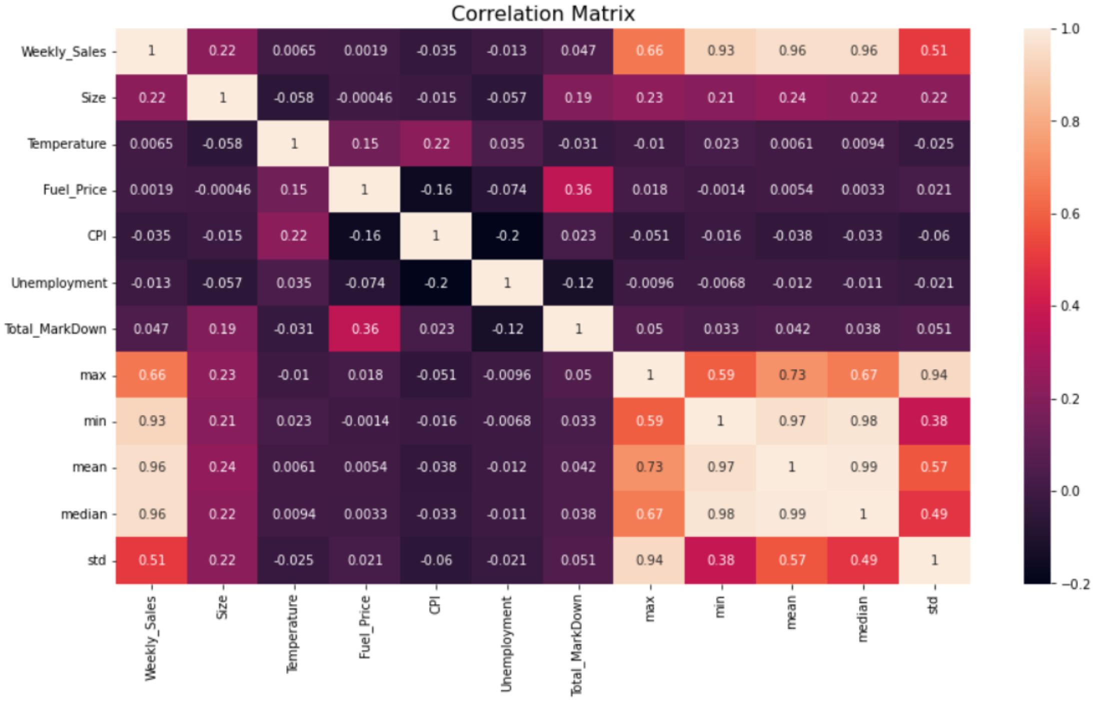
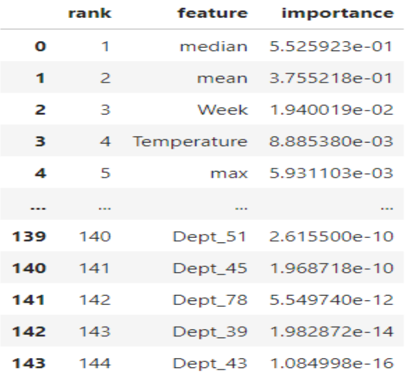


Figure 3.2.8 Correlation Heat map

We can see that; every attribute is highly correlated with each other. Some are positively correlated with each other and some attributes are negatively correlated with each other.

Feature Importance:

Important features can be observed by using correlation matrix as heat map and recursive features has been eliminated and feature importance can be observed as shown in Fig. 3.2.9.



	rank	feature	importance
0	1	median	5.525923e-01
1	2	mean	3.755218e-01
2	3	Week	1.940019e-02
3	4	Temperature	8.885380e-03
4	5	max	5.931103e-03
...	...	...	...
139	140	Dept_51	2.615500e-10
140	141	Dept_45	1.968718e-10
141	142	Dept_78	5.549740e-12
142	143	Dept_39	1.982872e-14
143	144	Dept_43	1.084998e-16

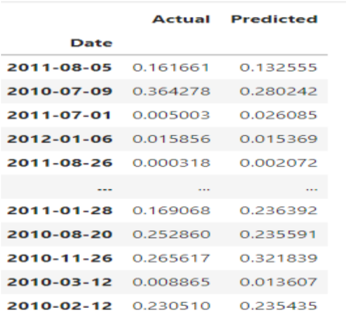
Figure 3.2.9 Feature importance

Data Splitting to train and test:

The data set was split into 80% for training and 20% for testing and we have considered Weely_Sales as target feature.

Fitting the different machine learning models:

a. Linear Regression model



	Actual	Predicted
Date		
2011-08-05	0.161661	0.132555
2010-07-09	0.364278	0.280242
2011-07-01	0.005003	0.026085
2012-01-06	0.015856	0.015369
2011-08-26	0.000318	0.002072
...	...	...
2011-01-28	0.169068	0.236392
2010-08-20	0.252860	0.235591
2010-11-26	0.265617	0.321839
2010-03-12	0.008865	0.013607
2010-02-12	0.230510	0.235435

Figure 3.2.10 Actual vs predicted values for LR model

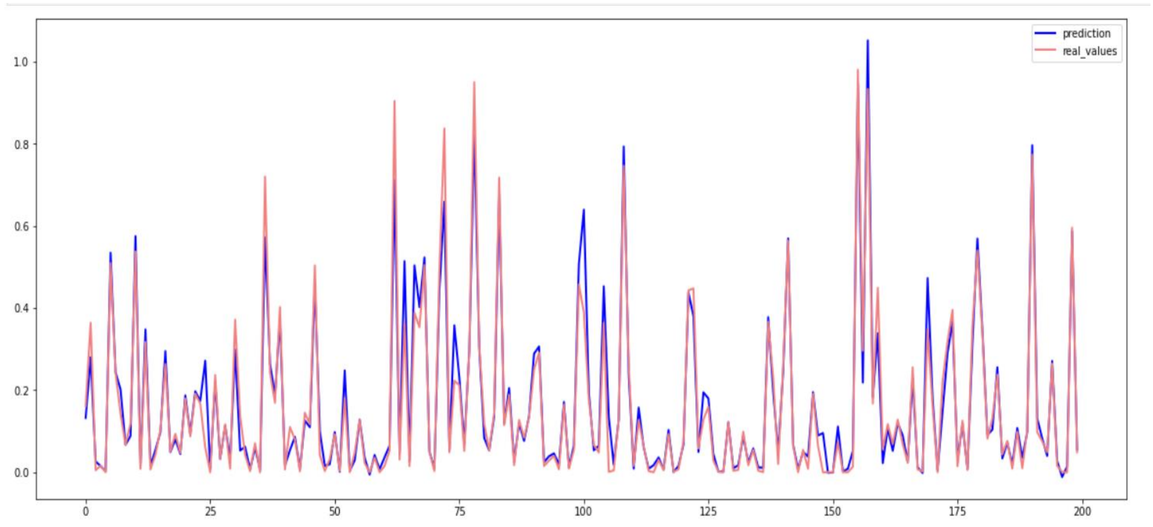


Figure 3.2.11 Graph of Actual and predicted values for LR model

Accuracy 92.28%

b. Random Forest Regressor model

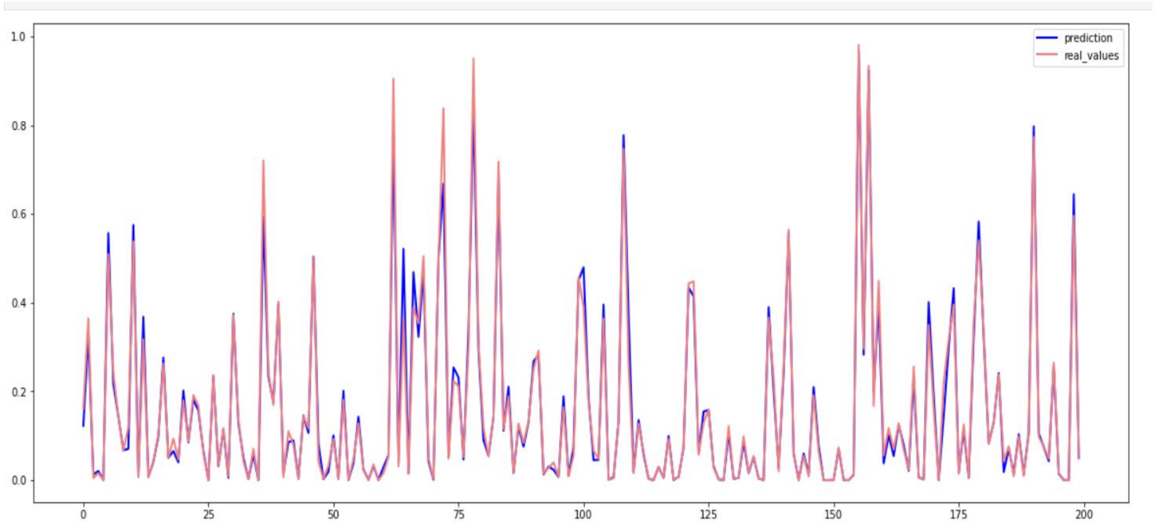


Figure 3.2.12 Graph of Actual and predicted values for Random Forest Regressor model

Accuracy 97.89%

c. K nearest neighbour Regressor model

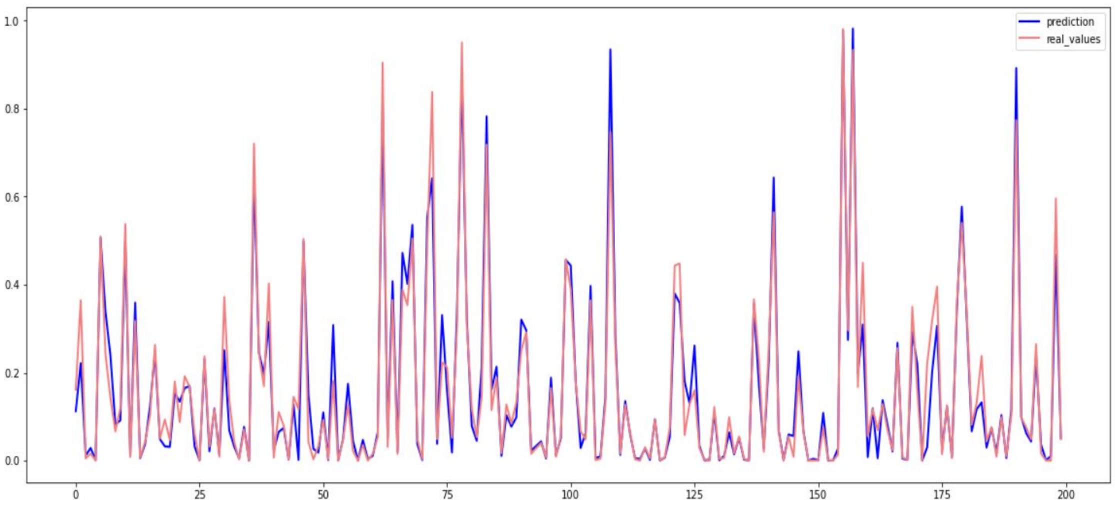


Figure 3.2.13 Graph of Actual and predicted values for K-Neighbors Regressor model

Accuracy 91.97%

d. XGBoost model

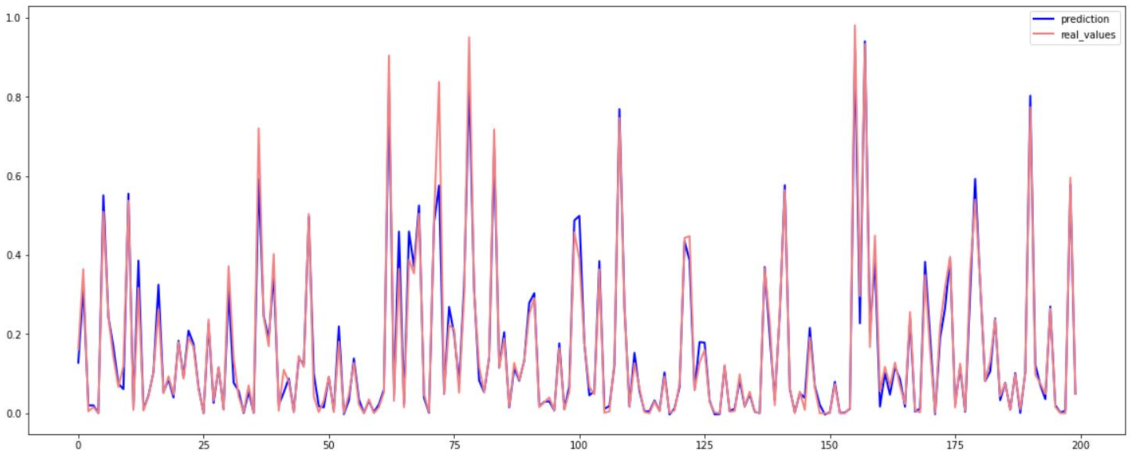


Figure 3.2.14 Graph of Actual and predicted values for XGBoost model

Accuracy 97.22%

e. Custom Deep Learning Neural Network model

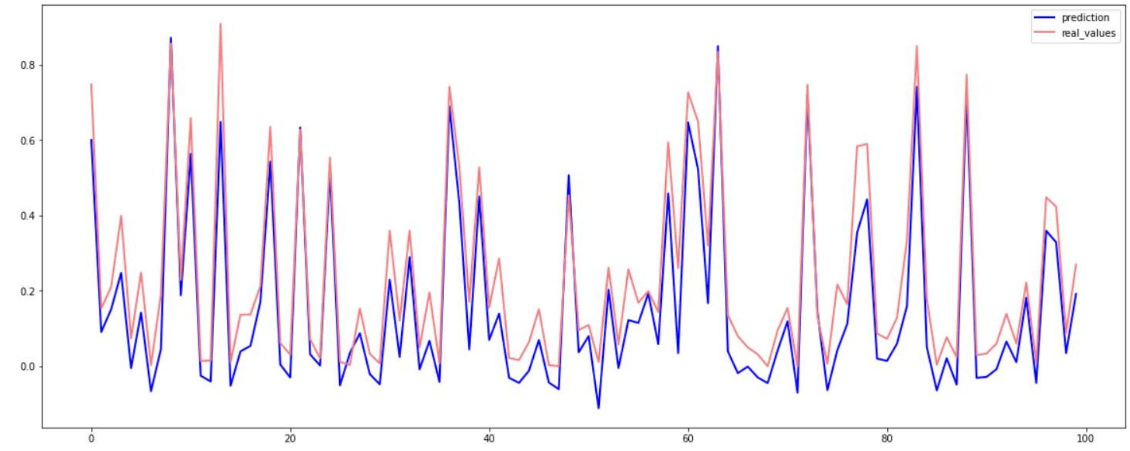


Figure 3.2.15 Graph of Actual and predicted values for DNN model

Accuracy 97.12%

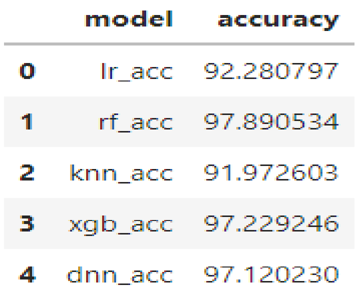
CHAPTER 4: RESULT

Comparative analysis is performed and values have been observed for different accuracy metrics as shown as in table 4.1.

Model used	*Mean Absolute Error*	*Mean Squared Error*	*Root Mean Square Error*	*R2*
Linear Regression model	*0.030057*	*0.0034851*	*0.059*	*0.9228*
Random Forest Regression Model	*0.015522*	*0.000953*	*0.3087*	*0.9788*
K Neighbours Regression Model	*0.331221*	*0.0036242*	*0.60202*	*0.91992*
XGBoost Regression Model	*0.0267718*	*0.0026134*	*0.05112*	*0.94211*
Custom Deep Learning Neural Network	*0.033255*	*0.003867*	*0.06218*	*0.9144106*

Table 4.1 Comparison of models

Model accuracy is calculated and plotted using boxplot as in Fig. 4.1 and Fig. 4.2. Random forest regressor model, XGboost model, and Deep learning neural network were found to perform better than other models and provide almost identical accuracy.



	model	accuracy
0	lr_acc	92.280797
1	rf_acc	97.890534
2	knn_acc	91.972603
3	xgb_acc	97.229246
4	dnn_acc	97.120230

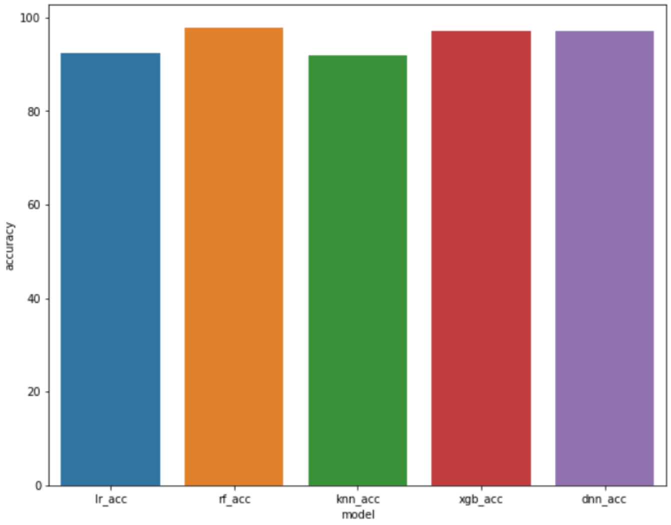


Figure 4.1 Accuracy of different models

Figure 4.2 Plot for comparison of model

CHAPTER 4: CONCLUSION & APPLICABILITY

In today's digitally connected world, every shopping mall wishes to anticipate customer wants in order to minimize seasonal shortages of sale items. Companies and shopping malls are getting better at anticipating product sales and consumer requests on a daily basis. For precise sales forecasting, extensive research is being conducted at the organization level. In this project, implementation is performed on Walmart store sales dataset. We have implemented custom deep learning neural network and compared with various machine learning models to check the performance of model.

We have achieved an accuracy of 97.120% for custom deep learning neural network. It has been observed that Random Forest regressor model, XGboost model and Deep learning neural network performed better than other models and gave almost similar accuracies. Out of all, the best performing model has come out as a Random Forest with 97.89% followed by XGB and DNN.

In practice, this seasonal effect may cause the categories defined over a 10-month period to be quite different from those extrapolated from the last two months. In order to correct such bias, it would be beneficial to have data that would cover a longer period of time. Through customer segmentation and personalized marketing campaigns, you can reduce the risk of showing ads to uninterested consumers. This clearly increases the efficiency of the campaign and hence produces ROI-producing efforts.