VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI-590018

A Internship Project Report

on

Big Mart Sales Prediction using ML

Submitted in partial fulfillment of the requirements for the VI semester

Computer Science and Engineering of Visvesvaraya Technological University, Belagavi

Submitted by:

Swaroop KS 1RN20CS161

Under the Guidance of:

Mrs.Chethana H R
Associate Professor
Dept. of CSE

Department of Computer Science and Engineering
(Accredited by NBA up to 30/6/2025)
RNS Institute of Technology
Channasandra, Dr.Vishnuvardhan Road, Bengaluru-560 098
2023

RNS INSTITUTE OF TECHNOLOGY

Channasandra, Dr.Vishnuvardhan Road, Bengaluru-560098

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

(Accredited by NBA up to 30/6/2025)

CERTIFICATE

This is to certify that the mini project work entitled **Big Mart Sales Prediction using ML** has been successfully carried out by **Swaroop K S** bearing USN **1RN20CS161**, bonafide student of **RNS Institute of Technology** in partial fulfillment of the requirements for the 7th semester **Computer Science and Engineering of Visvesvaraya Technological University"**, Belagavi, during academic year 2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The internship project report has been approved as it satisfies the internship requirements of 7th semester BE, CSE.

Signature of the Guide	Signature of the HoD	Signature of the Principal
Mrs.Chethana H R	**Dr. Kiran P**	**Dr. Ramesh Babu H S**
Associate Professor	Professor & Head	Principal
Dept. of CSE	Dept. of CSE	

External Viva:

Name of the Examiners — Signature with Date

1.

2.

Acknowledgement

The successful completion of any achievement is not solely dependent on individual efforts but also on the guidance, encouragement, and cooperation of intellectuals, elders, and friends. We would like to take this opportunity to express our heartfelt gratitude to all those who have contributed to the successful execution of this project.

First and foremost, we extend our profound thanks to **Sri. Satish R Shetty**, Managing Trustee of R N Shetty Trust and Chairman of RNS Group of Institutions, and **Sri. Karan S Shetty** , CEO of RNS Group of Institutions, Bengaluru, for providing a conducive environment that facilitated the successful completion of this project.

We would also like to express our sincere appreciation to our esteemed Director, **Dr. M K Venkatesha**, for providing us with the necessary facilities and support throughout the duration of this work.

Our heartfelt thanks go to our respected Principal, **Dr. Ramesh Babu H S**, for his unwavering support, guidance, and encouragement that played a vital role in the completion of this project.

We would like to extend our wholehearted gratitude to our HOD, **Dr. Kiran P**, Professor, and Head of the Department of Computer Science & Engineering, RNSIT, Bangalore, for his valuable suggestions and expert advice, which greatly contributed to the success of this endeavor.

A special word of thanks is due to our project guide,**Mrs.Chethana H R**, Associate Professor in the Department of CSE, RNSIT, Bangalore, for her exceptional guidance, constant encouragement, and unwavering assistance throughout the project.

We would also like to express our sincere appreciation to all the teaching and non-teaching staff of the Department of Computer Science & Engineering, RNSIT, for their consistent support and encouragement.

Once again, we express our deepest gratitude to everyone involved, as their support and cooperation were instrumental in the successful completion of this project.

Abstract

The aim of a ”Big Mart Sales Prediction using Machine Learning” project is to develop a predictive model that can accurately forecast sales for a retail chain like Big Mart. This project seeks to leverage advanced data analytics and machine learning techniques to enhance the operational efficiency and profitability of the Big Mart retail chain. By harnessing historical sales data, external factors, and innovative predictive models, this project aims to accurately forecast sales, providing valuable insights and strategic advantages.

In pursuit of this objective, the project encompasses data collection, preprocessing, and feature engineering to ensure the availability of high-quality data for analysis. Various machine learning algorithms, including regression models and time series forecasting techniques, are employed and fine-tuned to develop precise sales prediction models.

The outcomes of this project extend beyond mere sales forecasts. They include optimized inventory management, demand forecasting, pricing strategies, cost reduction, customer satisfaction improvement, and the ability to adapt to dynamic market trends. Additionally, the project promotes data-driven decision-making and facilitates the identification of opportunities for competitive advantage in the retail industry.

Through its multifaceted approach, the Big Mart Sales Prediction project addresses the complex challenges faced by modern retailers, offering a comprehensive solution to improve decision-making, operational efficiency, and ultimately, the bottom line

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions). The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production

instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically with all the subtlety this allows in edge cases, instead of traditional rule-based methods that require rigid theories and a full comprehension of problems.

1.2 Problem Statement

Big Mart, a prominent retail chain, grapples with several operational challenges that hinder its ability to maximize profitability and operational efficiency. The primary obstacle is the absence of precise sales forecasting, which impacts various facets of its business operations. The specific challenges include:

Inefficient Inventory Management: Big Mart frequently faces the dilemma of maintaining optimal inventory levels. Overstocking leads to excessive capital being tied up in inventory and the wastage of perishable goods, while understocking results in missed sales opportunities and dissatisfied customers.

Variable Demand Patterns: The sales patterns for different products and stores exhibit considerable variability due to a multitude of factors, including seasonality, promotional activities, geographical location, and evolving consumer preferences. Accurately predicting these patterns is crucial for planning inventory and resources effectively.

Pricing and Promotion Strategy: Big Mart currently lacks a data-driven approach to pricing and promotions. Accurate sales predictions are essential for setting competitive prices and designing effective promotional campaigns, ultimately impacting revenue and customer engagement.

Resource Allocation: Efficient allocation of resources, including workforce scheduling, shelf space management, and advertising efforts, remains a challenge. Accurate sales forecasts are instrumental in optimizing these allocations, leading to reduced operational costs and enhanced customer satisfaction.

Customer Satisfaction: The inconsistency in product availability stemming from inaccurate sales forecasts leads to customer dissatisfaction, eroding customer loyalty and potentially driving customers to competitors.

Chapter 2

Resource Requirements

2.1 Hardware Requirements

The Hardware requirements are very minimal and the program can be run on most of the machines.Table 2.1 gives details of hardware requirements.

Table 2.1: Hardware Requirements

Processor	Intel Core i3 processor
Processor Speed	1.70 GHz
RAM	4 GB
Storage Space	40 GB
Monitor Resolution	1024*768 or 1336*768 or 1280*1024

2.2 Software Requirements

The software requirements are description of features and functionalities of the system.Table 2.2 gives details of software requirements.

Table 2.2: Software Requirements

Operating System	Windows 8.1
IDE	Anaconda
Libraries	Pandas,NumPy,Streamlit,Matplotlib,Seaborn.

2.2.1 Anaconda

Anaconda is the birthplace of Python data science. It is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. Anaconda distribution comes with over 250 packages automatically installed, and over 7,500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command-line interface.

2.2.2 Jupyter Notebook

Jupyter Notebook (formerly IPython Notebook) is a web-based interactive computational environment for creating notebook documents. Jupyter Notebook is built using several open- source libraries, including IPython, ZeroMQ, Tornado, jQuery, Bootstrap, and MathJax. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension. Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathematica in the 1980s. Jupyter interest overtook the popularity of the Mathematica notebook interface in early 2018.

Chapter 3

Design

The description of all the phases performed in the project is given below:

- **Data Collection:**

  Gather relevant data from various sources. This may include structured data from databases, spreadsheets, or APIs, as well as unstructured data like text or images.

- **Data Preprocessing:**

  Prepare and clean the data to make it suitable for analysis. This involves tasks such as: Handling missing values. Removing duplicates. Encoding categorical variables. Scaling or normalizing numeric features. Handling outliers. Feature engineering to create new informative features.

- **Data Splitting:**

  Divide your dataset into training, validation, and testing sets. The training set is used to train the model, the validation set helps tune hyperparameters, and the testing set is used to evaluate model performance.

- **Feature Selection/Extraction:**

  Choose the most relevant features for your predictive model. Feature selection helps reduce dimensionality, while feature extraction may involve creating new features from existing ones.

- **Model Selection:**

  Choose an appropriate machine learning algorithm or model for your problem. Consider factors like data type (classification or regression), model complexity, interpretability, and the size of your dataset.

  .

- **Model Training:**

  Train the selected model using the training dataset. The model learns patterns and relationships within the data to make predictions.

- **Hyperparameter Tuning:**

  Optimize the model's hyperparameters to improve its performance. Techniques like grid search or random search can be used to find the best combination of hyperparameters.

- **Model Evaluation:**

  Assess the model's performance using the validation dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, and mean squared error (depending on the type of problem).

- **Model Testing:**

  Evaluate the final model on the testing dataset to assess its real-world performance. This step provides an estimate of how well the model will perform on new, unseen data.

- **Deployment:**

  Deploy the trained model in a production environment where it can make real-time predictions. This may involve creating APIs, integrating with databases, or embedding the model in software applications.

Chapter 4

Implementation

- **Step 1:Importing the necessary libraries**

  import numpy as np #numerical python

  import pandas as pd #loading,processing and analysis of data

  import matplotlib.pyplot as plt #graphical visulization

  import seaborn as sns #graphical visulization with corr,trends,patterns

  import warnings

  warnings.filterwarnings('ignore')

- **Step 2:Load the Dataset¶**

  df_train=pd.read_csv("train.csv")

  df_test=pd.read_csv("test.csv")

  df_train.head()

  df_test.head()

- **Step 3:Exploratory Data Analysis(EDA)**

  Data checking

  df_train.shape

  df_train.size

  df_train.dtypes

  Data Cleaning

  #Check and drop Duplicates

```
df_train=df_train.drop_duplicates()
df_train.shape

#Check for missing values
df_train.isnull().sum()

#replacing 0 values in dataset
df_train['Item_Weight'].fillna(df_train['Item_Weight'].mean(),inplace=True)
df_test['Item_Weight'].fillna(df_test['Item_Weight'].mean(),inplace=True)
df_train['Outlet_Size'].fillna(df_train['Outlet_Size'].mode()[0],inplace=True)
df_test['Outlet_Size'].fillna(df_test['Outlet_Size'].mode()[0],inplace=True)
```

- **Step 4:Data Visualization**

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df_train, x='Item_Visibility', y='Item_Outlet_Sales')
plt.title('Relationship between Item Visibility and Sales')
plt.xlabel('Item Visibility')
plt.ylabel('Item Outlet Sales')
plt.show()

# Checking how Item Outlet sales varies with Item Visibility
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_train, x='Outlet_Establishment_Year', y='Item_Outlet_Sales')
plt.title('Relationship between Outlet Establishment Year and Sales')
plt.xlabel('Outlet Establishment Year')
plt.ylabel('Item Outlet Sales')
plt.xticks(rotation=45)
plt.show()

# Checking how Item Outlet sales varies with Outlet Establishment Year
plt.figure(figsize=(10,5))
sns.heatmap(df_train.corr(),annot=True)
```

plt.show()

Checking the correlation of features

- **Step 5:Splitting Data into X and Y (predictors and outcome)**

x = df_train.drop('item_outlet_sales',axis=1)
y = df_train['item_outlet_sales']

- **Step 6:Feature Scaling**

from sklearn.preprocessing import StandardScaler from sklearn.preprocessing import StandardScaler sc = StandardScaler() x_train_std = sc.fit_transform(x_train) x_test_std = sc.fit_transform(x_test)

X=pd.DataFrame(X_scaled, columns=X.columns)
X.head()

- **Step 7:Train Test Split**

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,random_state=101,test_size=0.2)

- **Step 8:Building the Regression Algorithm**
  - **Linear Regression**
  - **Random Forest**
- **Step 9:Hyperparameter tuning using GridSearchCV**

This helps in finding the best parameters to get the best r2_score.¶

Chapter 5

5.1 Results & Snapshots

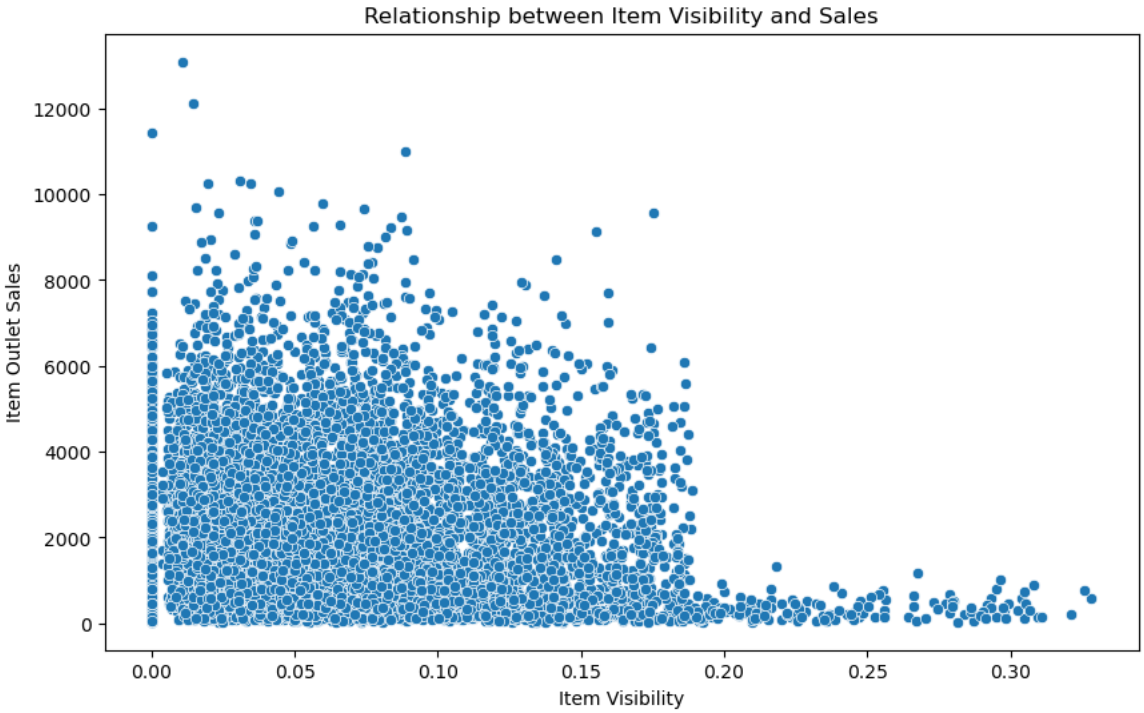

Figure 5.1: Checking relation between Item Visibility and Item Outlet Sales

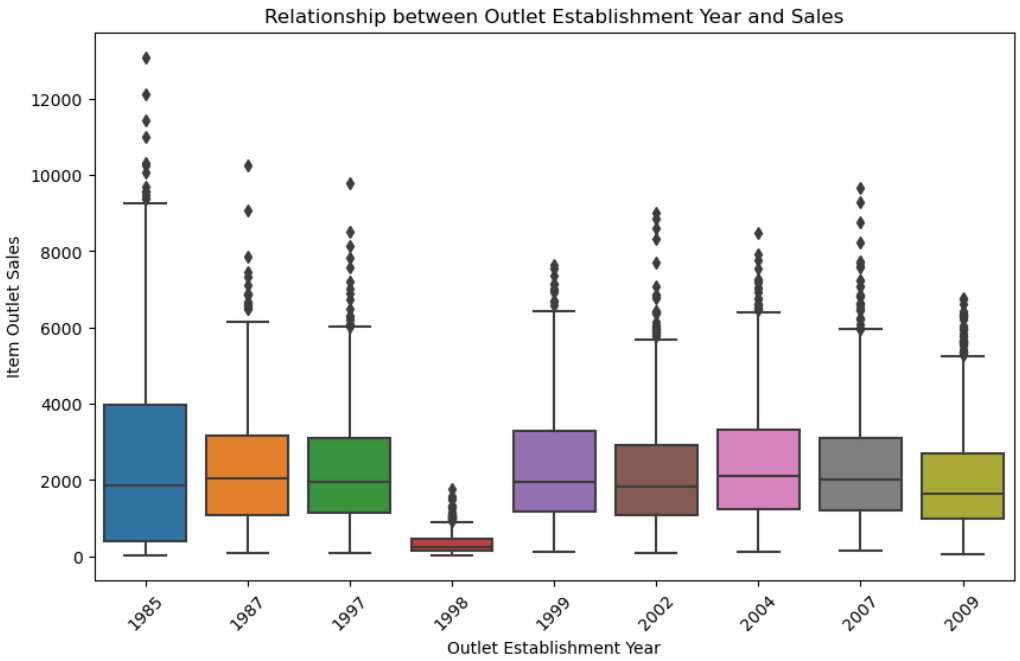

Figure 5.2: Box plot of Outlet sales and Establishment year

?? Correlation Matrix

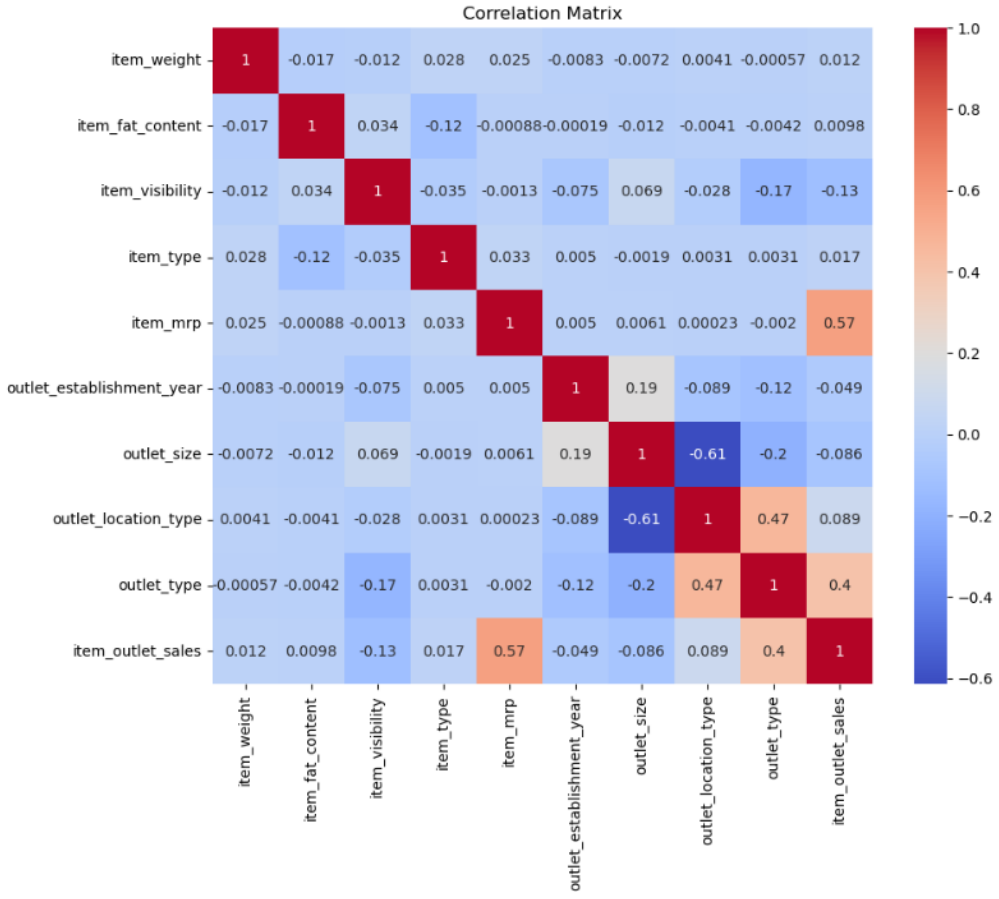

Figure 5.3: Correlation between features to drop the columns

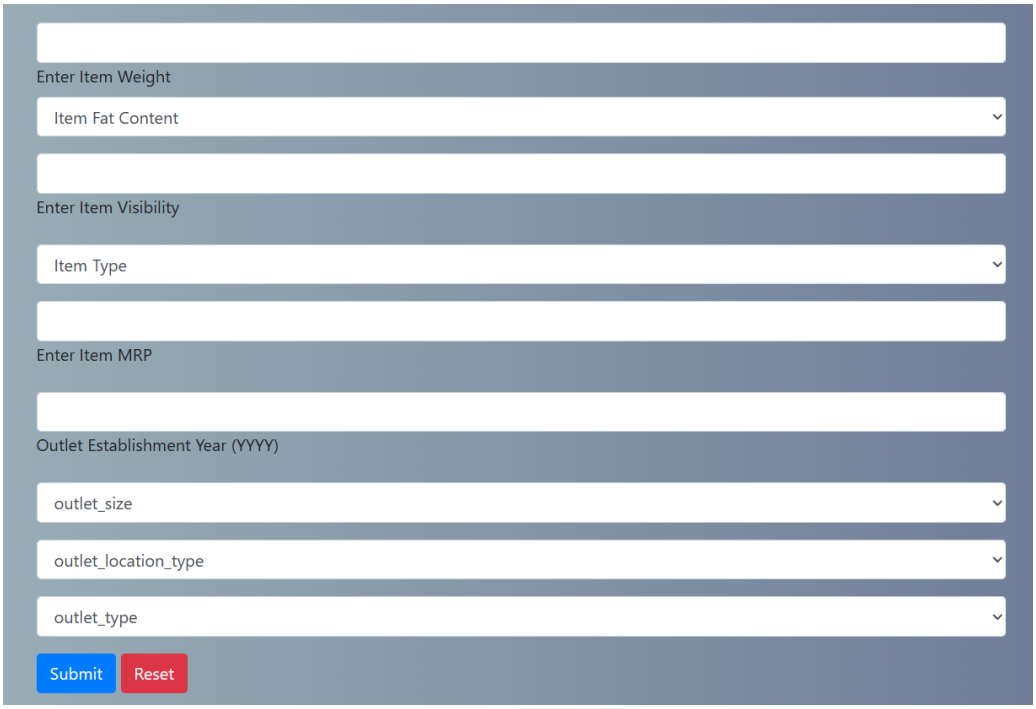

Figure 5.4: Prediction of values

Chapter 6

Conclusion & Future Enhancements

6.1 Conclusion

In conclusion, the use of machine learning in predicting sales for Big Mart, or any retail establishment, offers tremendous potential for optimizing operations and maximizing profitability. Through the analysis of historical sales data, factors such as product attributes, store location, seasonal trends, and customer behavior can be leveraged to build accurate predictive models.

By employing machine learning algorithms, Big Mart can achieve several benefits:

Improved Inventory Management: Accurate sales predictions enable better inventory management, reducing overstocking or understocking issues. This leads to cost savings and ensures that customers can find the products they want.

Enhanced Marketing Strategies: ML models can identify which products are likely to sell well during specific seasons or events. This information can be used to create targeted marketing campaigns, promotions, and product placements.

Optimized Pricing: Machine learning can help in dynamic pricing, adjusting prices based on demand and other factors. This can maximize revenue while remaining competitive in the market.

Personalized Customer Experience: ML models can analyze customer data to provide personalized product recommendations and shopping experiences, increasing customer loyalty and satisfaction.

Reduced Operational Costs: Efficient inventory management, optimized staffing, and improved resource allocation can lead to reduced operational costs and higher profit margins.

Data-Driven Decision-Making: ML-based sales predictions provide valuable insights for strategic decision-making, allowing Big Mart to adapt and respond quickly to changing market conditions

6.2 Future Enhancements

Future enhancements of Big Mart's sales prediction using ML include:

Advanced AI models

Real-time predictions

Customer segmentation

Supply chain optimization

Dynamic pricing optimization

Demand forecasting for new products

Enhanced inventory management

Predictive maintenance

Customer experience enhancement

References

[1] Edward Angel, *"Interactive Computer Graphics A Top-Down Approach With OpenGL"* 5th Edition, Addison-Wesley, 2008.

[2] F.S. Hill,*"Computer Graphics Using OpenGL"*, 2nd Edition, Pearson Education, 2001.

[3] James D.Foley, Andries Van Dam, Steven K. Feiner, John F Hughes, *"Computer Graphics"*, Second Edition, Addison-Wesley Professional, August 14,1995.

[4] @online OpenGL Official ,`https://www.opengl.org/`

[5] @online OpenGL Overview , `https://https://www.khronos.org/opengl/`