SML Project - Sales Prediction Model

Aditya Bagri - 2022029

Suyash Kumar - 2021293

I. Introduction

This is a report for our course project for the course CSE342 - Statistical Machine Learning. This project is basically a sales prediction model which takes the statistics about the store and a specific item and predicts its future sales.

II. Problem Statement

The objective is to construct a predictive model to forecast product sales at specific outlets accurately. This enables e-commerce companies to enhance inventory management, fine-tune marketing tactics, and improve revenue estimates. Through accurate sales predictions, businesses can better align operations with market demands, increasing efficiency and maintaining competitiveness in the fast-paced retail environment.

III. Motivation

Accurate sales prediction is crucial for e-commerce businesses to thrive in a dynamic market environment. By leveraging historical sales data and predictive modelling techniques, businesses can anticipate customer demand, streamline operations, and stay ahead of the competition, driving sustainable growth and success.

IV. Literature Review

Sales prediction models utilize a range of techniques from simple statistical methods to advanced machine learning approaches. Linear Regression is commonly used for its simplicity and interpretability. To address issues of high dimensionality and overfitting, Regularized Linear Regression and Lasso Regression are often employed. Lasso Regression, in particular, has shown superior performance in many studies by effectively zeroing out less important features, thus enhancing model accuracy and robustness. Additionally, ensemble methods like Random Forest and gradient boosting models such as XGBoost are utilized for their ability to capture non-linear relationships and improve predictive accuracy.

V. Dataset Overview

We have picked up the dataset from Kaggle. This dataset, titled 'Big Mart Sales Prediction Datasets', consists of 8,500 rows of simulated sales data for a small e-commerce website. The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. It has the following columns.

A. Fields in dataset

1) Item Identifier: Unique product ID
2) Item Weight: Weight of product
3) Item Fat Content: Whether the product is low fat or not
4) Item Visibility: The Percentage of the total display area of all products in a store allocated to the particular product
5) Item Type: The category to which the product belongs
6) Item MRP: Maximum Retail Price (list price) of the product
7) Outlet Identifier: Unique store ID
8) Outlet Establishment Year: The year in which the store was established
9) Outlet Size: The size of the store in terms of ground area covered
10) Outlet Location Type: The type of city in which the store is located
11) Outlet Type: Whether the outlet is just a grocery store or some sort of supermarket
12) Item Outlet Sales: Sales of the product in a particular store. This is the outcome variable to be predicted.

VI. Distribution plot of each feature

A. Item Weight Distribution

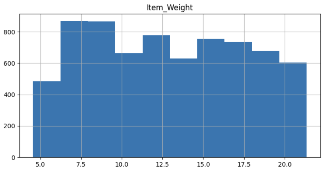

Fig. 1. Item Weight Distribution

B. Item MRP Distribution

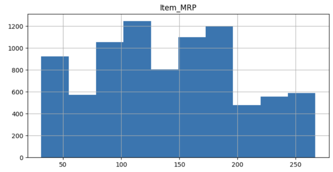

Fig. 2. Item MRP Distribution

C. Item Visibility

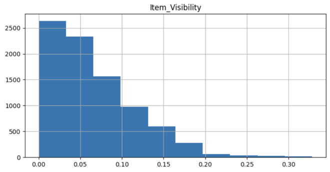

Fig. 3. Item Visibility

D. Outlet Establishment Year Distribution

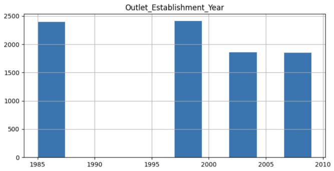

Fig. 4. Outlet Establishment Year Distribution

E. Item Outlet Sales

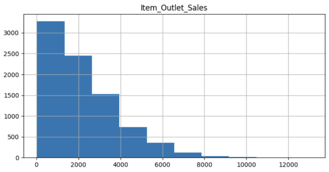

Fig. 5. Item Outlet Sales

VII. Method

The process begins with EDA (Exploratory Data Analysis) on the dataset, followed by data reparation and feature engineering, handling missing data values, and encoding categorical variables to numerical values like Outlet Type. After that, we split the data into 2 training sets (0.4 fractions each) and 1 test set (the remaining 0.2 fractions of the dataset).

A. Data preparation

We begin by checking the information about the data set and then check for the missing values in the dataset, which we plot on a heat map, as shown in Fig. 6.

The features Item Weight and Outlet Size contain the missing values, so we start treating them.

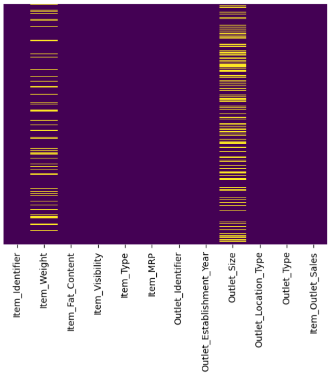

Fig. 6. Heat Map of the missing values

1) Treating Item Weight: We check if we can replace the missing values from mean and median, so we replace the missing values with both mean and median and compare the variance from the original data. The variance of the original data was 21.561, and both mean and median yielded a variance of about 17.86. We then plot both the mean replaced data and the median replaced data against the original data to compare the changes in a plot and verify if any of them give a closer plot to the original dataset. The plot is shown in Fig. 7. The mean and median replaced data give a huge spike from the original data, so we try to fill the missing values using interpolation. Interpolation is a process of determining the unknown values that lie in between the known data points. – Google. We then plot the interpolated values against the original dataset to verify the deviation from the original graph. Shown in Fig. 8. The interpolation seems to give better results, so we use it to fill the missing values of the dataset.

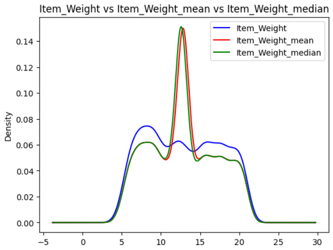

Fig. 7. Plot of replaced data vs original data

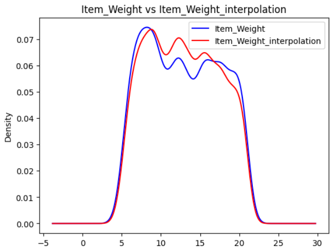

Fig. 8. Interpolation Graph vs Original Graph

2) Treating Outlet Size: Outlet size must be treated differently since it's a categorical feature, so we use the outlet size mode to replace the missing values. We get the following mode values as our answer:

TABLE I
MODE VALUES

Outlet Type	Grocery Store	Supermarket Type 1	Supermarket Type 2	Supermarket Typ
Outlet Size	Small	Small	Medium	Medium

All the missing values of the dataset are treated now, and we can move to our next step.

B. Feature Engineering

We convert the categorical features to give numerical values using Label Encoder from the *sklearn* library.

We plot a scatter plot of each feature against the value we need to predict for better observation. Represented in Fig. 9 to Fig. 17.

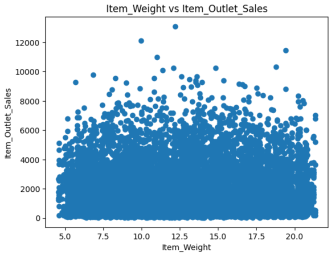

Fig. 9. Scatter plot of Item Weight

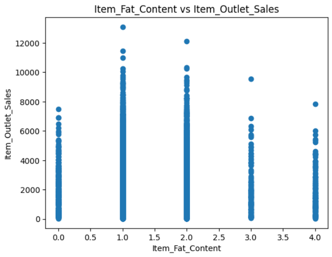

Fig. 10. Scatter plot of Item Fat Content

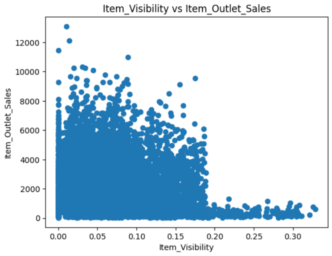

Fig. 11. Scatter plot of Item Visibility

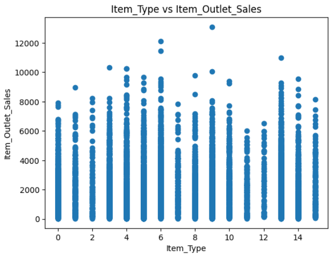

Fig. 12. Scatter plot of Item Type

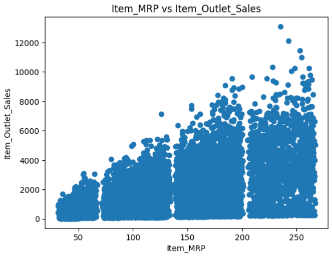

Fig. 13. Scatter plot of Item MRP

We also plot a correlation matrix between all the features to see and understand their correlation. See Fig. 19. We plot a feature importance graph to check the importance of each feature in predicting outlet sales value. We do this using the feature importance functionality from the *sklearn* library. See Fig. 20.

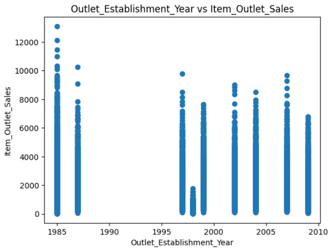

Fig. 14. Scatter plot of Outlet Establishment Year

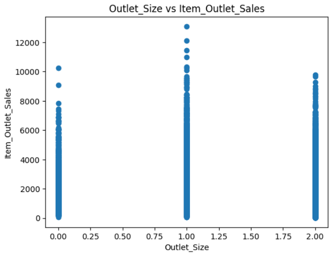

Fig. 15. Scatter plot of Outlet Size

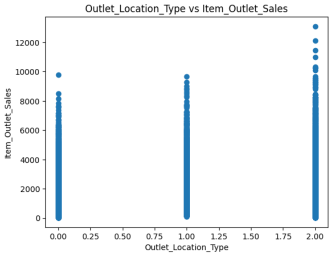

Fig. 16. Scatter plot of Outlet Location Type

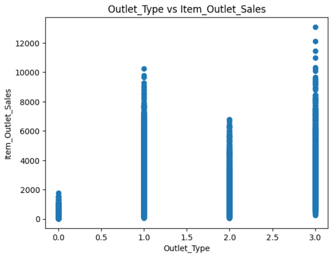

Fig. 17. Scatter plot of Outlet Type

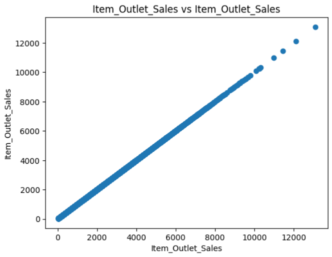

Fig. 18. Scatter plot of Item Outlet Sales

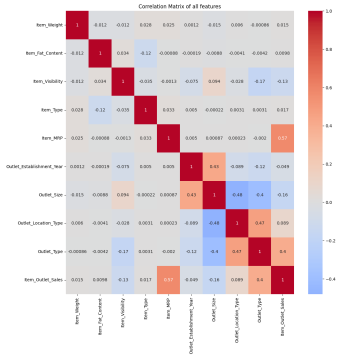

Fig. 19. Correlation Matrix

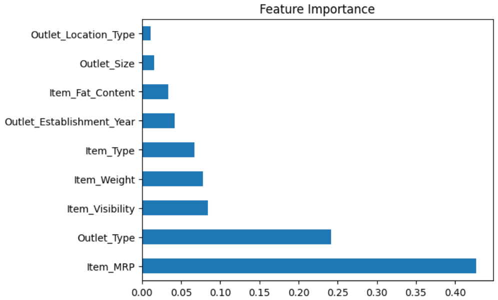

Fig. 20. Feature Importance Graph

C. Predictions

We separate our dataset into training sets and train different models from the *sklearn* library. We use models:

1) Decision Trees
2) Random Forests
3) Linear Regression
4) ADA Boost Regression
5) Gradient Boosting Regression

We predict the values according to both training sets and then take their mean for the final prediction. We use RMSE as our evaluation metric and end up with the following results:

1) Decision Trees: R2 Score - 0.35208195198319214 RMSE - 1368.6808985210655

2) Random Forests: R2 Score - 0.5903925791166651 RMSE - 1088.243484245638

3) Linear Regression: R2 Score - 0.5131412938989109 RMSE - 1186.43414702223

4) ADA Boost: R2 Score - 0.5234647692452334 RMSE - 1173.7880248229767

5) Gradient Boosting: R2 Score - 0.616743810945301 RMSE - 1052.6566923477844

Note: These values might change with each run but are similar to the mentioned values with deviations around 0.05 in the R2 scores.

D. Analysis - 1

We plot the results of each prediction on scatter plots against the true values to visualize them. We also plot all the R2 scores on a histogram. See from Fig. 21 to Fig. 27.

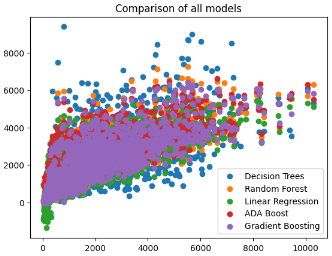

Fig. 21. All Models in a single plot

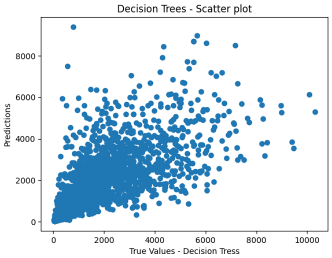

Fig. 22. Decision Trees

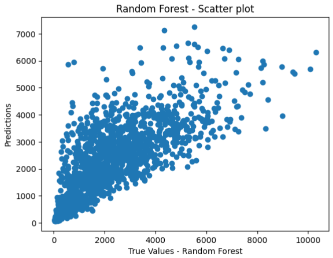

Fig. 23. Random Forest

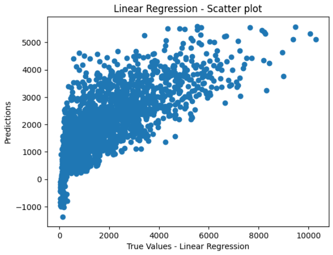

Fig. 24. Linear Regression

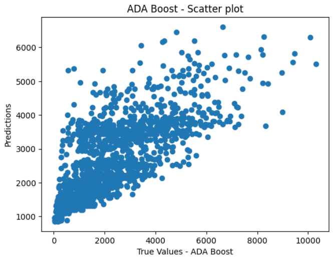

Fig. 25. ADA Boost

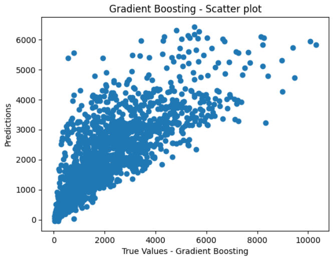

Fig. 26. Gradient Boosting

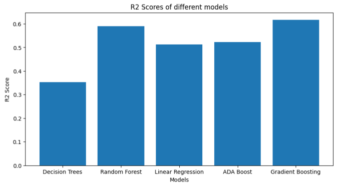

Fig. 27. R2 Scores of each model

E. Feature Scaling and Selection

We try to improve the accuracy of our model by scaling and selecting features according to their importance.

1) Feature scaling is a method used to normalize the range of independent variables or features of data. – Google
2) Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. – Google

We use Standard Scaler and Select From Model functions of *sklearn* library to do the following steps. After feature selection we get the following result: Total features: 9; Selected features: 2; with coefficients shrank to zero: 0. Plotting the selected features against Item Outlet Sales to visualize.

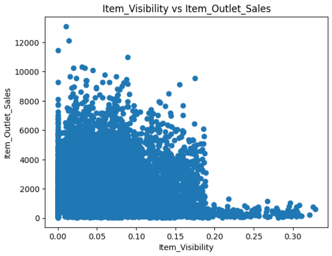

Fig. 28. Selected Feature 1 - Item Visibility

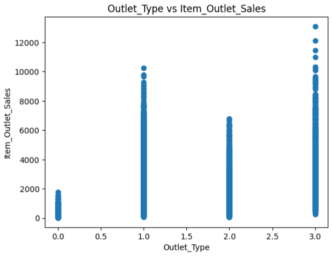

Fig. 29. Selected Feature 2 - Outlet Type

F. Predicting values again

The new scores we get are:

1) Decision Trees: R2 score 0.9333592215183424 RMSE: 457.3453275695202

2) Random Forests: R2 score 0.8139552005013182 RMSE: 764.1576813016744

3) Linear Regression: R2 score 0.1881115172911323 RMSE: 1596.3290093537187

4) ADA Boost: R2 score 0.174788891131893 RMSE: 1609.3731389367358

5) Gradient Boosting: R2 score 0.2940897074958978 RMSE: 1488.5004583494153 Note: We see a significant improvement in the accuracy of Decision Trees and Random Forest but also notice a significant decrease in the accuracy of the other models.

G. Analysis - 2

We again plot the predictions against the true values. We plot histogram of R2 Scores of all models. See Fig. 30 to Fig. 36.

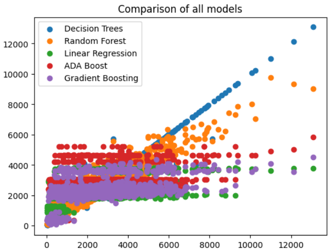

Fig. 30. Comparison of all models

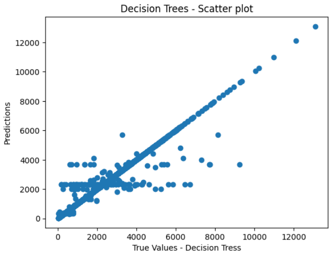

Fig. 31. Decision Trees

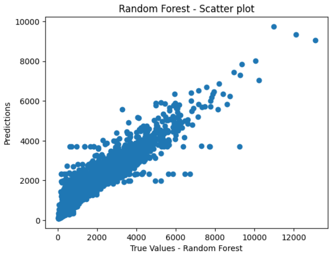

Fig. 32. Random Forests

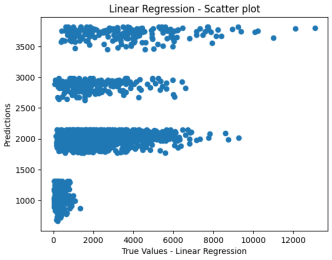

Fig. 33. Linear Regression

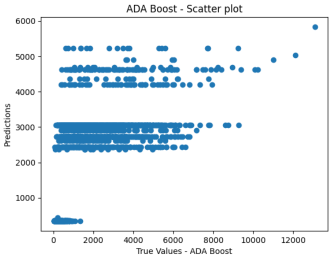

Fig. 34. ADA Boosting

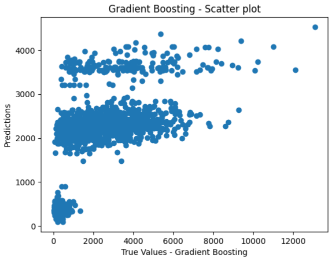

Fig. 35. Gradient Boosting

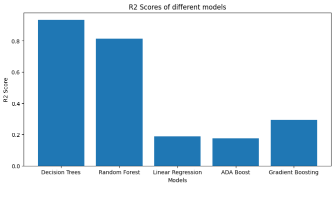

Fig. 36. R2 Scores of all models

VIII. Different Evaluation Metrics

A. Before Feature Scaling and Selection

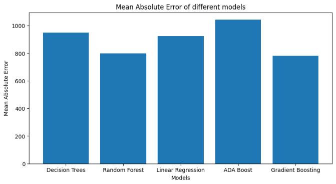

Fig. 37. Mean Absolute Error Histogram

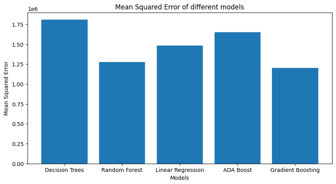

Fig. 38. Mean Squared Error

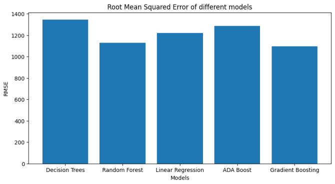

Fig. 39. Root Mean Squared Error

B. After Feature Scaling and Selection

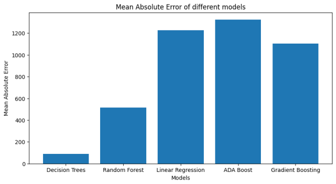

Fig. 40. Mean Absolute Error

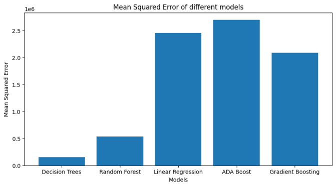

Fig. 41. Mean Squared Error

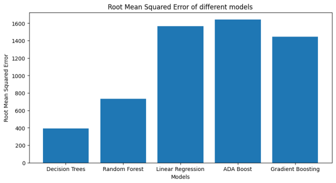

Fig. 42. Root Mean Squared Error

IX. Individual Contributions

- **Aditya Bagri**: Exploratory data analysis, performing various model checks and trying to fit the best one. Improving the model's accuracy by performing feature selection and feature scaling. Making the final report for the project.
- **Suyash Kumar**: Exploratory data analysis, plotting various graphs to visualise the data and making it easier to make observations. Making the final presentation for the project.

X. References

- Kaggle BigMart Dataset
- Matplotlib tutorial by GFG
- SciKit Standard Scaler Documentation
- SciKit SelectFromModel Documentation
- SciKit Ridge Documentation
- Label Encoder - GeeksForGeeks