Big Mart Sales Prediction

Kush Patel
Btech Computer Science
(Ahmedabad University)
kush.p3@ahduni.edu.in

Meet Patel
Btech Computer Science
(Ahmedabad University)
meet.p6@ahduni.edu.in

Dev Patel
Btech Computer Science
(Ahmedabad University)
dev.p1@ahduni.edu.in

Vatsal Shah
Btech Computer Science
(Ahmedabad University)
vatsal.s5@ahduni.edu.in

***Abstract*— This ML project aimed at forecasting the sales of a retail location. It involves data gathering, cleansing, feature engineering, model selection, and evaluation, using machine learning techniques such as linear regression, XGBOOST, Random Forest regression. The main objective of the project is to create a predictive model that can estimate product sales in a shop with accuracy, allowing store managers to make informed decisions about inventory control, sales forecasting, and marketing tactics. The report also provides a literature survey that compares various machine learning algorithms used for sales forecasting, such as linear regression, random forest, and XGBoost, and discusses their strengths and weaknesses. The implementation of the project involves data collection, data preprocessing, data visualization, model training and evaluation, and model comparison. The report explains each step of the implementation in detail, including data cleaning, handling missing values, and feature engineering. Exploratory data analysis is performed using the pandas profiling library and the Klib library. The report concludes by highlighting the importance of sales forecasting in business planning and decision-making and emphasizes that the selection of a machine learning algorithm will depend on the specific needs and goals of the business. The Big Mart Sales Prediction project is an excellent illustration of how machine learning can be used to solve real-world business problems and improve inventory management, boost profitability, and make better decisions about marketing and promotional tactics.**

Keywords*— *Sales Prediction, Linear regression, XGBOOST, Random Forest regression, Klib, Pandas.

I. Introduction

The Big Mart Sales Prediction uses past data to forecast the sales of a retail location. In this project, data pertaining to the sales of the products from various outlets is applied using machine learning techniques. In order to forecast the sales of a certain shop, we will use a total of 12 attributes from the dataset. The major goal of this project is to create a predictive model that can estimate product sales in the shop with accuracy, allowing store managers to make deft choices regarding inventory control, sales forecasting, and marketing tactics. Data gathering, data cleansing, feature engineering, model selection, and evaluation are just a few of the processes that the project goes through. Dataset can be gathered from a variety of sources, including as the point of sale system in the business or outside market research firms. The dataset must be cleaned after collection to get rid of any errors or missing information. This calls for activities like data standardisation, data transformation, and data imputation. The next step is to choose the pertinent characteristics that will be utilised to train the machine learning model after the dataset has been cleaned. This entails determining the critical elements that are most likely to affect sales, such as the kind of goods, the price, and any promotional offers. The next stage is to choose the best machine learning algorithm that can be used to predict sales after the features have been found. For this endeavour, methods like linear regression, XGBoost, and Random Forest regression are used. The model must be tested after training to ascertain its correctness and dependability. Comparing the model's predictions to the actual sales data includes testing the model on a holdout set of data. The Big Mart Sales Prediction project, in its whole, is a superb illustration of how machine learning can be utilised to address actual business issues. Retailers may enhance inventory management, boost profitability, and make better decisions regarding their marketing and promotional tactics by properly projecting sales.

II. Literature Survey

Sales forecasting is a critical component of business planning and decision-making, enabling companies to effectively allocate resources, manage their workforce, and optimize cash flow. To achieve accurate sales forecasts, businesses can turn to various machine learning algorithms, such as linear regression, random forest, decision trees, and XGBoost.

Linear regression models assume a linear relationship between inputs and outputs, which may not always be the case in real-world scenarios. On the other hand, random forest and XGBoost are tree-based algorithms that utilize multiple decision trees to make predictions. Decision trees can create a hierarchy of significant features.

XGBoost is an ensemble learning algorithm that combines several decision trees to make predictions. It is based on the concept of gradient boosting, where the algorithm trains weak learners in a sequential manner to improve the overall performance of the model. It also allows for flexible optimization objectives, which enables it to handle different types of regression problems. XGBoost has been used to predict stock market prices, machine failures, and future sales.

Random Forest is a popular ensemble learning algorithm that has been widely used for regression tasks in machine learning. The algorithm builds multiple decision trees and combines their outputs to make predictions. Random Forest is known for its ability to handle non-linear relationships between input features and the output variable. It is robust to outliers and missing data, which makes it a suitable algorithm for datasets with missing or noisy data. Additionally, Random Forest can handle high-dimensional feature spaces and is relatively computationally efficient, making it useful for large datasets.

When comparing these techniques, it is essential to consider the complexity of the problem, the dataset, and the desired accuracy. Ultimately, the selection of an algorithm will depend on the specific needs and goals of the business.

III. Implementation

The implementation involves a series of steps that can be broadly categorized into data collection, data preprocessing, data visualization, model training and evaluation, model comparison.

Data Collection: The first step is to collect the data required for the project. This may include data related to sales, products, stores, and other relevant variables. The data can be obtained from various sources, such as online repositories or directly from the organization's databases. We have used the Big Mart Sales dataset available on Kaggle, which has the 2013 sales data of 1559 products from 10 stores in different cities. There are a total of 11 features in the dataset: Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Identifier, Outlet Establishment Year, Outlet Size, Outlet Location Type, and Outlet Type. There are 7 categorical features: Item Identifier, Item Fat Content, Item Type, Outlet Identifier, Outlet Size, Outlet Location Type, Outlet Type and 4 numeric features: Item Weight, Item Visibility, Item MRP, and Outlet Establishment Year.

Data Preprocessing: Once the data has been collected, the next step is to preprocess it. This involves handling missing values, outliers, duplicates, and any other data quality issues. Additionally, feature engineering techniques can be used to extract useful information from the data and create new features if necessary. We first checked for the missing values in any of the inputs. We found missing values in Item Weight and Outlet Size columns. Item Weight is a numeric feature, so we replaced the missing values in Outlet Weight with the mean of the values. Outlet Size is a categorical column so we replaced the missing value with the mode of the Outlet Size data.

Data Visualization: After preprocessing, the data can be visualized using graphs and plots to gain insights into the data and identify any patterns or relationships between variables. This step can help identify potential predictors of sales and guide feature engineering. We performed Exploratory Data Analysis (EDA) using the pandas profiling library and Klib library on our dataset. Pandas profiling creates a profile report of the entire dataset, including the data analysis of each feature, plotting distributions of the input features, and correlation matrix. The Klib library is used to plot the distributions of the input features.

Splitting Data: After visualization, the data can be split into training and testing sets. The training set will be used to train the models, while the testing set will be used to evaluate the performance of the models. We split the data into training and testing sets by keeping 80% for training and 20% for testing set.

Model training and evaluation: We trained three models: Linear regression, XGBoost regression, and Random Forest regression. The models are trained using the training set. The models are evaluated using metrics such as R^2 value, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). We have used the inbuilt models available in the sci-kit-learn library. We trained the models with our train set and then tested the model with the test set and obtained the R^2 value, mean absolute error, and root mean squared error to compare the performance of the linear regression, XGBoost regression, and random forest regression models.

IV. Results

The performance of the linear regression, XGBoost regression, and random forest regression models can be compared using metrics such as R^2, mean absolute error, and root mean squared error. The initial results of the models is as shown:

On training set:

Model	R^2 value	MAE	RMSE
Linear regression	0.51	886.56	1181.92
XGBoost	0.86	454.94	631.71
Random Forest	0.93	290.92	419.37

On Test set:

Model	R^2 value	MAE	RMSE
Linear regression	0.51	915.1	1223.8
XGBoost	0.54	826.64	1195.64
Random Forest	0.55	825.14	1178.02

Hyperparameter tuning: From the initial implementation results, we can say that the XGBoost and Random Forest models were overfitting on the training set. Also, the test R^2 value of XGBoost and Random Forest models is very low compared to the train R^2 value. To overcome these problems, we did hyperparameter tuning to improve the performance and used regularization to prevent the models from overfitting on the training set. We used grid search method for hyperparameter tuning to find the best combination of the hyperparameter values for the XGBoost and Random Forest models. The hyperparameters tuned for XGBoost model are: 'n_estimators' (number of trees), 'max_depth' (maximum tree depth), learning rate, 'min_child_weight', 'reg_lambda' (L2 regularization), and Gamma. The hyperparameters tuned for Random Forest model are: 'n_estimators' (number of trees), 'max_depth' (maximum tree depth), 'max_features', 'min_samples_split', 'min_samples_leaf', and 'ccp_alpha'. The performance after doing hyperparameter tuning:

On training set:

Model	R^2 value	MAE	RMSE
Linear regression	0.51	886.56	1181.92
XGBoost	0.61	747.05	1049.47
Random Forest	0.62	726.07	1033.52

On Test set:

Model	R^2 value	MAE	RMSE
Linear regression	0.51	915.1	1223.8
XGBoost	0.60	776.72	1105.73
Random Forest	0.60	767.96	1107.99

V. Conclusion

In conclusion, the Big Mart Sales Prediction project demonstrated how machine learning techniques such as linear regression, XGBoost, and Random Forest regression can be used to predict product sales in a shop with accuracy. The project involved various processes such as data gathering, cleansing, feature engineering, model selection, and evaluation. The report also highlighted the importance of sales forecasting in business planning and decision-making and provided a literature survey comparing various machine learning algorithms used for sales forecasting. The implementation of the project was explained in detail, including data cleaning, handling missing values, and feature engineering. The project showcases how machine learning can be used to solve real-world business problems such as improving inventory management, boosting profitability, and making better decisions about marketing and promotional tactics.

Based on the results, we can conclude that the XGBoost regression and Random Forest regression models give better performance than the linear regression model for the given dataset.

On the training set, XGBoost and random forest regression have higher and nearly the same R^2 value of 0.61 and 0.62, respectively, indicating a better fit to the data compared to linear regression, which has an R^2 value of 0.51. Moreover, XGBoost and Random Forest models have a lower mean absolute error (MAE) and root mean squared error (RMSE) than linear regression, indicating that it has a lower prediction error.

On the test set, both the XGBoost and Random Forest models give an equal and higher R^2 value of 0.60 than linear regression's R^2 value of 0.51.

Therefore, based on the given results, we can conclude that the XGBoost regression and Random Forest regression models are a better choice for the given dataset compared to the linear regression model. Ultimately, the selection of an algorithm will depend on the specific needs and goals of the business.

References

[1] BIG MART SALES PREDICTION USING MACHINE LEARNING. (n.d.). IJCRT from https://ijcrt.org/papers/IJCRT2105404.pdf

[2] Big Mart Sales Prediction Using Machine Learning Techniques. (n.d.). IJSREDfromhttp://www.ijsred.com/volume3/issue4/IJSRED-V3I4P81.pdf

[3] Big Mart Sales Prediction Using Machine Learning Techniques. (n.d.). SVRECfromhttp://www.svrec.ac.in/docs/cse/publications/BIGMART.pdf

[4] PREDICTION OF BIG MART SALES USING MACHINE LEARNING.(2021,September9).IRJMETSfromhttps://www.irjmets.com/uploadedfiles/paper/volume_3/issue_9_september_2021/16025/final/fin_irjmets1630829142.pdf

[5] TIME SERIES ANALYSIS: FORECASTING WITH SARIMAX MODEL AND STATIONARY CONCEPT. (n.d.). Jetir.Org from, https://www.jetir.org/papers/JETIREJ06034.pdf

[6] Akande, Y. F., Idowu, J., Misra, A., Misra, S., OLUWATOBI, A. N., & Ahuja, R. (2022, October 30). Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset. ResearchGate. https://www.researchgate.net/publication/361549465_Application_of_XGBoost_Algorithm_for_Sales_Forecasting_Using_Walmart_Dataset

[7] Hamad, S., Alice, K., & Srivastava, A. (2022, November 7). Sales Forecasting using XGBoost. TechRxiv. https://www.techrxiv.org/articles/preprint/Sales_Forecasting_using_XGBoost/21444129

[8] Kadam, H., Shevade, R., Ketkar, D., & Rajguru, S. (n.d.). A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. Ijedr.org. https://www.ijedr.org/papers/IJEDR1804010.pdf

GitHub Repository Link:

https://github.com/dp913/CSE523_Machine_Learning_2023_Group-10