ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper | Open Access

Machine Learning for Sales Prediction in Big Mart

Aakanksha Ramesh Jadhav[1] and Ramesh D Jadhav[2*]

[1]E&TC-Sinhgad Institute of Technology and Science, Pune, India. E-mail: aakankshajadhav100201@gmail.com
[2]Sinhgad Institute of Management, Pune, India. E-mail: rameshdjadhav@gmail.com

Article Info

Volume 4, Issue 1, May 2024
Received : 22 November 2023
Accepted : 31 March 2024
Published : 05 May 2024
doi: 10.51483/IJDSBDA.4.1.2024.58-62

Abstract

The retail industry, particularly the grocery sector, generates massive amounts of data daily. Predicting sales accurately is essential for effective inventory management, supply chain optimization, and revenue maximization. In this research paper, we propose a comprehensive study of various machine learning regression algorithms to predict Big Mart sales. We compare the performance of Linear Regression, Decision Trees, Random Forests, Gradient Boosting, and Neural Networks to identify the most accurate model for sales prediction.

Keywords: *Machine learning, Gradient boosting, Predictive analytics, Sales prediction*

1. Introduction

Sales prediction is a vital aspect of the retail industry, playing a pivotal role in optimizing inventory management, supply chain operations, and maximizing revenue. Accurate sales forecasting allows retailers, such as Big Mart, to anticipate consumer demand and make informed decisions regarding stock replenishment, marketing strategies, and pricing. In recent years, the application of machine learning techniques has shown promising results in enhancing sales prediction accuracy. This research paper presents a comprehensive study on sales prediction in Big Mart using machine learning algorithms. By leveraging a diverse set of regression algorithms, including Linear Regression, Decision Trees, Random Forests, Gradient Boosting, and Neural Networks, we aim to identify the most precise and robust model for sales prediction.

Data preprocessing is a crucial step in any predictive modeling task, and this study addresses various aspects of data cleaning and transformation to ensure the quality and reliability of the dataset. Additionally, feature selection techniques are applied to identify the most significant predictors for sales prediction, which helps in reducing computational complexity and improving model performance. To evaluate the performance of each algorithm, we employ standard evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics provide a quantitative assessment of the model's accuracy and predictive capabilities. By comparing the results of each algorithm, we can determine which

* *Corresponding author: Ramesh D Jadhav, Sinhgad Institute of Management, Pune, India. E-mail: rameshdjadhav@gmail.com*

approach is most suitable for accurate sales prediction in the Big Mart context. The practical implications of our research findings are of great significance to the retail industry. Retailers can utilize the insights gained from this study to optimize their business operations, improve inventory management, and enhance overall performance. Additionally, the implementation of accurate sales prediction models can lead to better resource allocation and customer satisfaction, ultimately contributing to increased profitability and competitive advantage in the retail market (Chen *et al.*, 2018; Aakanksha Ramesh and Aditya Ramesh, 2023).

2. Purpose

The paper aims to review and analyze various Machine Learning algorithms and methodologies that can be applied to predict sales in a retail setting. It will explore regression, time series analysis, and other relevant techniques to identify the most suitable approach for sales prediction in Big Mart.

3. Process Flow

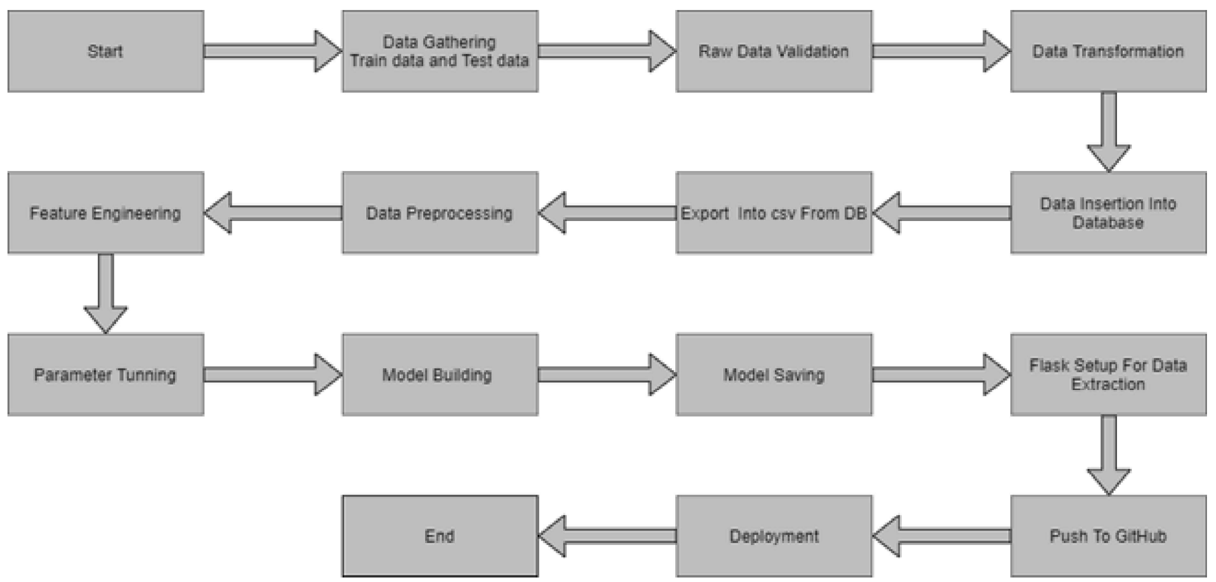

Figure 1: Process Flow

4. Introduction to Dataset

For this research paper, the dataset used for sales prediction in Big Mart was obtained from Kaggle, a popular platform for data science and machine learning competitions. The dataset is typically provided as a CSV (Comma-Separated Values) file and contains historical sales data and related information for various products in the Big Mart stores (Figure 2) (Barreto *et al.*, 2019).

t[5]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establisl
0	FDA15	9.300	Low Fat	0.016047	Dairy	249.8092	OUT049	
1	DRC01	5.920	Regular	0.019278	Soft Drinks	48.2692	OUT018	
2	FDN15	17.500	Low Fat	0.016760	Meat	141.6180	OUT049	
3	FDX07	19.200	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	
4	NCD19	8.930	Low Fat	0.000000	Household	53.8614	OUT013	
5	FDP36	10.395	Regular	0.000000	Baking Goods	51.4008	OUT018	
6	FDO10	13.650	Regular	0.012741	Snack Foods	57.6588	OUT013	

Figure 2: Dataset Description

5. Data Cleaning

Data cleaning involves the essential task of preparing data for analysis by addressing issues like incomplete, incorrect, irrelevant, duplicated, or improperly formatted data. In the training data set, , in the data, there are 976 rows with missing "Item Weight" values and 1606 rows with missing "Outlet_Size" values, as shown in Figure 3. To handle these missing values appropriately, a method has been applied where numerical columns are filled with their respective mean values, while categorical columns are filled with their respective mode values. This ensures that the missing values are replaced with representative values based on the column type, maintaining the integrity of the data for further analysis (Aakanksha Ramesh and Aditya Ramesh, 2023; Barreto *et al.*, 2019).

```
In [8]:  df_test.isnull().sum()

Out[8]:  Item_Identifier                 0
         Item_Weight                   976
         Item_Fat_Content                0
         Item_Visibility                 0
         Item_Type                       0
         Item_MRP                        0
         Outlet_Identifier               0
         Outlet_Establishment_Year       0
         Outlet_Size                  1606
         Outlet_Location_Type            0
         Outlet_Type                     0
         dtype: int64
```

Figure 3: Data Cleaning

Data Cleaning using Klib Library

```
In [10]:  klib.data_cleaning(df_train)
          klib.data_cleaning(df_test)

Shape of cleaned data: (8523, 12)Remaining NAs: 3873

Changes:
Dropped rows: 0
     of which 0 duplicates. (Rows: [])
Dropped columns: 0
     of which 0 single valued.     Columns: []
Dropped missing values: 0
Reduced memory by at least: 0.52 MB (-66.67%)

Shape of cleaned data: (5681, 11)Remaining NAs: 2582

Changes:
Dropped rows: 0
     of which 0 duplicates. (Rows: [])
Dropped columns: 0
     of which 0 single valued.     Columns: []
Dropped missing values: 0
Reduced memory by at least: 0.33 MB (-68.75%)
```

Figure 4: Data Cleaning Using Klib Library

6. Feature Engineering

Feature engineering is a critical step in the data analysis process where new features are created or existing ones are transformed to improve the performance of machine learning models. This step helps in extracting valuable information from the data and making it more suitable for the modeling process (Aakanksha Ramesh and Aditya Ramesh, 2023; Jain *et al.*, 2018).

7. Model Building

Model building is the process of creating a predictive machine learning model using a prepared dataset. It involves selecting an appropriate algorithm, splitting the data into training and testing sets, training the model on the training data, evaluating its performance on the testing data, and optimizing its hyperparameters for better results (Figure 5). The goal is to build an accurate and reliable model that can make predictions on new, unseen data with good generalization capabilities.

```
In [42]: from sklearn.ensemble import GradientBoostingRegressor

In [43]: from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

In [44]: model = GradientBoostingRegressor()
         model.fit(X_train_full,y_train)
         preds = model.predict(X_valid_full)
         print('score of', x)
         print('mae :',mean_absolute_error(y_valid, preds))
         print('R2 :',r2_score(y_valid, preds))
         print('rmse :',np.sqrt(mean_squared_error(y_valid, preds)))
         print()

         score of outlet_type
         mae : 773.7203584017598
         R2 : 0.594925548374112
         rmse : 1088.8507647270947
```

Figure 5: Model Training

8. Model Saving

Model saving is the process of persisting a trained machine learning model to a file so that it can be reused for making predictions on new data without having to retrain the model every time. This is essential for deploying the model in real-world applications. The model is typically saved in a standardized format that allows easy loading and use in different programming environments (Figure 6).

```
import joblib

filename = 'model_final2.sav'
joblib.dump(model, filename)
```

Figure 6: Model Saving

9. Results

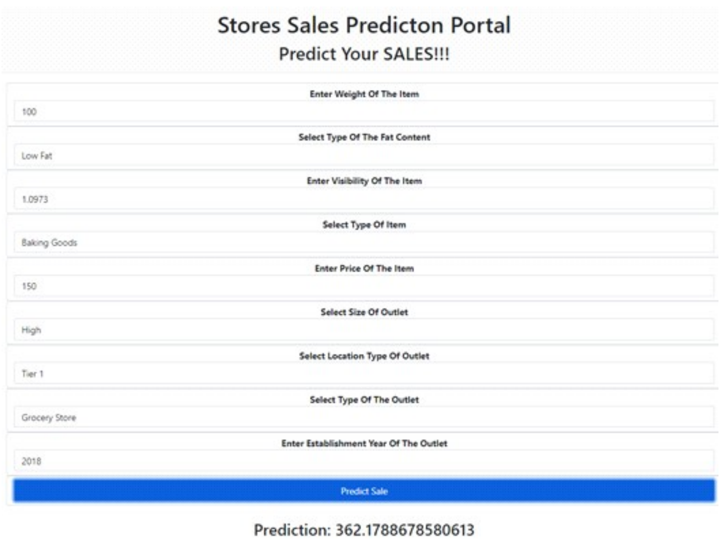

Figure 7: Website

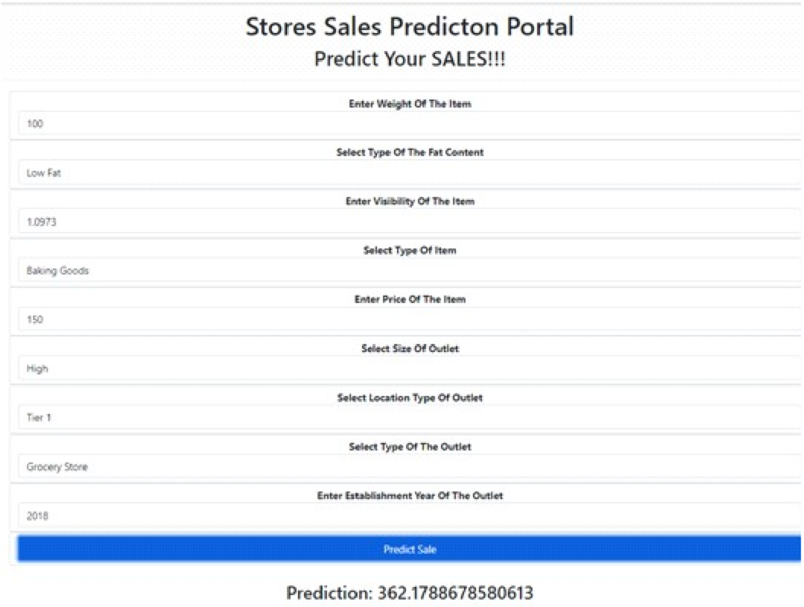

Figure 8: Final Prediction

10. Conclusion

This research paper explored the application of Machine Learning for sales prediction in Big Mart, aiming to address the increasing demand for accurate forecasting in retail environments.

References

Aakanksha Ramesh Jadhav, Aditya Ramesh Jadhav and Dr Ramesh, D. (2023). Association Rule Mining in Retail: Exploring Market Basket Analysis with Apriori Algorithm (SSRN, May 27, 2023).

Barreto, A.M., Pedro, A. and Pereira, M. (2019). Sales Forecasting in Retail Using Machine Learning Algorithms: A Case Study. In Proceedings of the International Conference on Enterprise Information Systems (ICEIS), Vol. 1, pp. 440-447.

Chen, X., Li, X., Wang, Y., Liu, X. and Qian, X. (2018). Sales Prediction in Retail Industry Using Machine Learning Algorithms. In Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC), pp. 155-159.

Jain, R., Kumar, P., Sharma, R. and Sharma, A. (2018). Predictive Modeling of Sales Data Using Machine Learning Algorithms. In Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS), pp. 220-225.

Patel, K., Patel, M. and Shah, S. (2019). Predicting Sales Using Machine Learning Algorithms: A Case Study in the Retail Industry. In Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 158-163.

Zhang, Y., Liu, C., Chen, S. and Wu, W. (2017). Deep Learning for Sales Forecasting in Retail. In Proceedings of the IEEE International Conference on Big Data (Big Data), pp. 2574-2581.

Cite this article as: Aakanksha Ramesh Jadhav and Dr Ramesh D Jadhav (2024). Machine Learning for Sales Prediction in Big Mart. *International Journal of Data Science and Big Data Analytics*, 4(1), 58-62. doi: 10.51483/IJDSBDA.4.1.2024.58-62.