Green University of Bangladesh

Department of Computer Science and Engineering (CSE)
Semester: (Fall, Year: 2023), B.Sc. in CSE (Day)

Sales Data Analysis and Prediction System for Big Mart Using Machine Learning Algorithms

---

Course Title: Data Mining Lab
Course Code: CSE 436
Section: D5 201

Students Details

Name	ID
MD Sakhawat Hossain Rabbi	201002311
Md Mesbahul Bari	201002385

Course Teacher's Name: Mr. Mozdaher Abdul Quader, Lecturer Dept. of CSE

[For teachers use only: Don't write anything inside this box]

Lab Project Status	
Marks:	**Signature:**
Comments:	**Date:**

Contents

Chapter 1

Introduction

1.1 Abstract

In the present era, supermarket operations, including those of Big Marts, extensively monitor the sales data of each individual item. This meticulous tracking enables them to anticipate potential consumer demand and efficiently manage inventory. Through data mining techniques applied to the data warehouse, anomalies and general trends are unveiled.For retail giants like Big Mart, this wealth of data serves as a valuable resource for predicting future sales volumes. Various cutting-edge machine learning techniques, such as RandomForestRegressor and Linear Regression, have been employed to develop a predictive model. The outcome of this endeavor has demonstrated superior performance compared to existing models.This predictive model plays a pivotal role in forecasting the sales trajectory of businesses like Big Mart, providing valuable insights for strategic decision-making. The integration of advanced forecasting techniques has proven to be a game-changer in optimizing inventory management and staying ahead in the competitive retail landscape.

Keywords

Machine learning, Sales forecasting, Random forest, Regression

1.2 Introduction

In the dynamic landscape of retail, the ability to predict sales accurately is a pivotal aspect of strategic decision-making for businesses. The Big Mart Sales Prediction Project ventures into the realm of machine learning to harness the power of data-driven insights. BigMart, a prominent retail chain, collected extensive sales data across diverse products and outlets, setting the stage for a transformative predictive modeling initiative.This project revolves around a dataset meticulously gathered in 2013, encompassing 1559 products distributed across 10 stores in varied cities. The objective is clear: construct a predictive model capable of unraveling the intricacies of product sales at specific outlets. As businesses navigate an era of intense competition and evolving consumer behavior, the foresight derived from accurate sales predictions becomes a strategic advantage. The dataset unfolds with a comprehensive set of attributes, including product weights, fat content, visibility, and categorical classifications. Each entry encapsulates critical information such as Maximum Retail Price (MRP), establishment years of outlets, and distinctive identifiers. Leveraging this wealth of information, the predictive model aims not only to forecast sales but also to unearth the underlying factors influencing purchasing patterns.In the face of real-world challenges, this dataset is not without its imperfections. Technical glitches have led to missing values, demanding meticulous treatment to ensure the integrity of the predictive model. The challenge, therefore, lies not only in constructing an accurate model but also in navigating the complexities of real-world data.

The significance of this endeavor lies in BigMart's quest to enhance its understanding of product and outlet dynamics, ultimately leading to optimized sales and inventory management strategies. By embracing advanced machine learning techniques, this project aspires to go beyond traditional methods, exploring algorithms like XGBoost, Random Forest, and Linear Regression.

As the journey unfolds, the project aims to contribute valuable insights to the retail sector's perennial quest for precision in sales forecasting. The predictive model developed here may well serve as a guiding beacon for businesses seeking a competitive edge in an ever-evolving marketplace. In the quest for actionable insights, the Big Mart Sales Prediction Project stands poised to unveil the untapped potential of machine learning in shaping the future of retail.

1.3 Objective & Motivation

- **Objectives of Big Mart Sales Prediction Using Machine Learning:**

1. **Sales Forecasting Accuracy:** The primary objective is to enhance the accuracy of sales forecasting for Big Mart. Machine learning models can analyze historical data, identify patterns, and make predictions, leading to more precise estimations of future sales.

2. **Inventory Optimization:** By predicting sales accurately, Big Mart aims to optimize its inventory management. This involves maintaining an optimal level of stock to meet customer demand, avoiding overstocking or under-stocking situations.

3. **Operational Efficiency:** Machine learning models contribute to operational efficiency by automating the sales prediction process. This reduces the reliance on manual forecasting methods, allowing the retail chain to allocate resources more efficiently.

4. **Marketing Strategy Enhancement:** Accurate sales predictions empower Big Mart to refine its marketing strategies. Understanding consumer behavior and predicting popular products help tailor marketing campaigns, promotions, and discounts effectively.

5. **Consumer Satisfaction:** The ability to anticipate and meet customer demand contributes to improved customer satisfaction. When products are available when and where customers need them, it enhances their shopping experience.

6. **Competitive Advantage:** Utilizing advanced machine learning techniques provides Big Mart with a competitive edge. The ability to adapt to market trends swiftly and optimize operations positions the retailer ahead of competitors.

- **Motivation Behind Big Mart Sales Prediction Using Machine Learning:**

1. **Data-Driven Decision-Making:** The retail industry generates vast amounts of data. The motivation lies in harnessing this data to make informed decisions. Machine learning algorithms can uncover insights, patterns, and trends that might not be apparent through traditional analysis.

2. **Dynamic Market Conditions:** Markets are dynamic, and consumer preferences can change rapidly. Machine learning allows Big Mart to adapt to these changes swiftly by providing real-time insights and predictions based on the latest data.

3. **Optimizing Resources:** Efficient resource allocation is crucial in retail. Predicting sales with accuracy ensures that resources such as staff, storage, and transportation are utilized optimally, minimizing wastage and costs.

4. **Meeting Consumer Expectations:** In the era of e-commerce and fast-paced retail, consumers expect seamless shopping experiences. Accurate sales predictions contribute to the availability of products, meeting consumer expectations and fostering loyalty.

5. **Strategic Planning:** Machine learning aids in strategic planning by providing a forward-looking view of sales. This is instrumental in setting long-term goals, planning marketing campaigns, and aligning resources with anticipated demand.

6. **Continuous Improvement:** The motivation for deploying machine learning in sales prediction is rooted in a culture of continuous improvement. By learning from past data and refining models over time, Big Mart can continuously enhance the accuracy and effectiveness of its sales predictions.

In essence, the integration of machine learning in sales prediction aligns with the broader goals of improving efficiency, meeting customer expectations, and staying competitive in a rapidly evolving retail landscape.

1.4 Literature Review

In the ever-evolving landscape of Big Mart sales prediction, the quest for precision has steered researchers towards a diverse array of machine learning algorithms. Kadam et al. scrutinized conventional methodologies, such as Random Forest and Linear Regression, unveiling their suboptimal accuracy. This revelation sparked the adoption of the heralded XGBoost algorithm, celebrated for its prowess in refining sales prognostications with unparalleled accuracy and efficiency [1].

Makridakis et al. delved into the intricacies of predictive methodologies, grappling with challenges posed by data scarcity and truncated life cycles in applications. Their discernment underscored the pivotal role of historical data, particularly in markets attuned to dynamic consumer needs, as a linchpin in accurate outcome predictions [2].

In the comparative scrutiny led by C. M. Wu et al., a spectrum of machine learning algorithms for Multiple Regression on Black Friday Sales Data underwent evaluation. While neural networks tantalized as an option, their intricate nature and relative inefficiency prompted the advocacy for simpler algorithms tailored for precision in predictions. This insight spotlights the judicious selection of algorithms, striking a balance between accuracy and computational efficiency [3].

Das et al. embarked on a foray into predicting retail sales of footwear, employing the sophistication of Recurrent Neural Networks (RNN) and feed-forward neural networks. Despite the allure of neural networks, inefficiencies were unraveled in this intricate approach. This realization seamlessly aligns with the recommendation to opt for more straightforward algorithms, exemplified by the efficiency powerhouse, XGBoost [3].

S. Cheriyan et al. navigated the landscape by implementing three machine learning algorithms on a designated dataset. Rigorous testing culminated in the ascendancy of the gradient boosting algorithm as the epitome of accuracy, magnifying the role of adept algorithmic selection in optimizing predictive outcomes [4].

A. Krishna et al. conducted an exhaustive study, juxtaposing traditional regression against the dynamism of boosting algorithms. Their revelations echoed the symphony of results favoring boosting algorithms, a testament to the belief that intricate methodologies wield the scepter of superiority in the arena of sales forecasting for entities akin to the retail titan, Big Mart [5].

Various studies have delved into the application of machine learning algorithms for sales prediction, shedding light on the comparative efficiency of different methodologies. [6] In the exploration by C. M. Wu et al., a comprehensive comparison of diverse machine learning algorithms was conducted for multiple regression on Black Friday Sales Data. Notably, the study incorporated the intricacies of neural networks, highlighting their conceptual complexity and inherent inefficiency. The conclusion drawn emphasized the preference for simpler algorithms over neural networks, underscoring the need for efficiency in predictive models [7].

In a distinct domain, Das et al. explored the prediction of retail sales, specifically focusing on footwear. Their approach involved the utilization of recurrent Neural Networks (RNN) and feed-forward neural networks. Despite the potential sophistication of neural networks, the study revealed inefficiencies in using this complex approach for sales prediction. As a remedy, the suggestion was put forth to leverage the efficiency of the XGBoost algorithm as a more suitable alternative [8] .

S. Cheriyan et al. contributed to the discourse by implementing three machine learning algorithms on a given dataset. The study aimed to evaluate the performance of these algorithms, employing rigorous testing procedures. The pivotal outcome was the identification of the gradient boosting algorithm as the most accurate among the tested models. This finding underscores the importance of meticulous evaluation to select algorithms that yield optimal accuracy for sales prediction [**?**].

A. Krishna et al. undertook a comprehensive study that involved the implementation of both normal regression and boosting algorithms. Their findings illuminated the superiority of boosting algorithms over regular ones in the context of sales prediction. This observation further strengthens the argument for adopting sophisticated techniques to achieve enhanced predictive results in sales forecasting scenarios [9].

In amalgamating these studies, a common thread emerges – the quest for accuracy and efficiency in sales prediction models. The literature collectively accentuates the importance of algorithm selection based on a nuanced understanding of the dataset and the specific predictive goals, fostering a pragmatic approach in the realm of sales forecasting [1].

1.5 Dataset

The dataset for the Big Mart Sales Prediction project is sourced from Kaggle and focuses on sales data collected by BigMart in 2013. This dataset includes information on 1559 products across 10 stores situated in various cities. The goal is to develop a predictive model to forecast sales for each product at specific outlets, aiming to uncover factors influencing sales.

- **Train Dataset:** Contains 8523 entries with both input and output variables for model training.
- **Test Dataset:** Comprises 5681 entries with input variables, and predictions are required for sales.

Train File Variable Descriptions

Attribute	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of the total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which the store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in a particular store. This is the outcome variable to be predicted

Test File Variable Descriptions

Attribute	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of the total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which the store store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket

By exploring, analyzing, and modeling this dataset, the project aims to enhance BigMart's understanding of product and outlet properties influencing sales, ultimately leading to improved business strategies.

1.6 METHODOLOGY

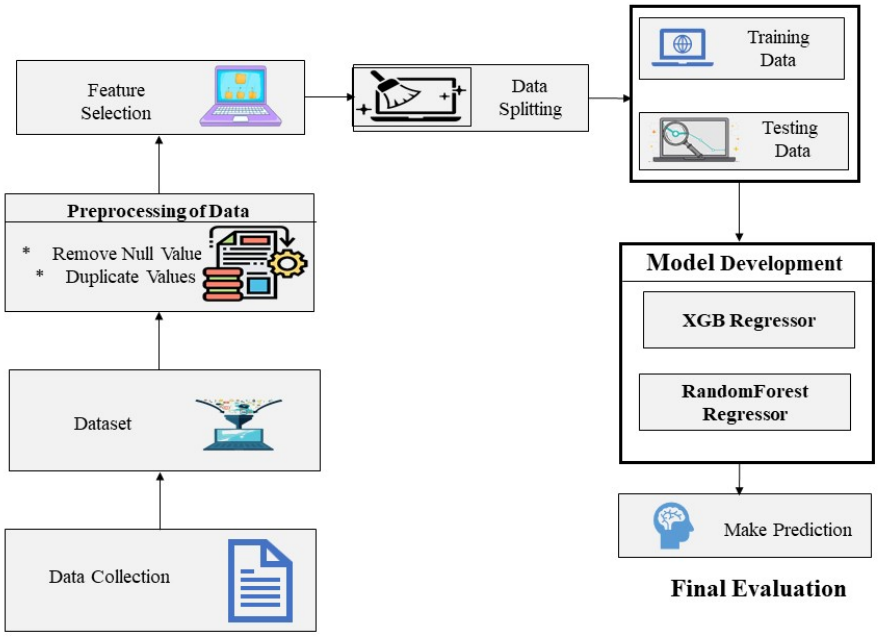

Figure 1.1: Our Proposed Methodology for Big Mart sales Prediction

The diagram in fig. 1.1 shows the sequence of steps that the dataset of Big Mart sales goes through to build up the proposed model to produce accurate results. There are a total of seven steps and each step plays a crucial role to build up the proposed model.

Data Prepossessing

Getting rid of extraneous data from the dataset is known as pre-processing. To create a dataset with a machine learning-friendly structure, pre-processing data transformation techniques are applied. The dataset is made more effective at this step by being cleaned of any inaccurate or superfluous data that can reduce the dataset's accuracy. Remove missing data: In this process, the null values, which including missing values and Nan values, are changed to 0. Data was cleared of any errors and missing values as well as duplicates. Encoding Data that may be categorised: Variables with a finite set of label values are regarded as categorical data. Most machine learning algorithms prefer numerical input and output variables.

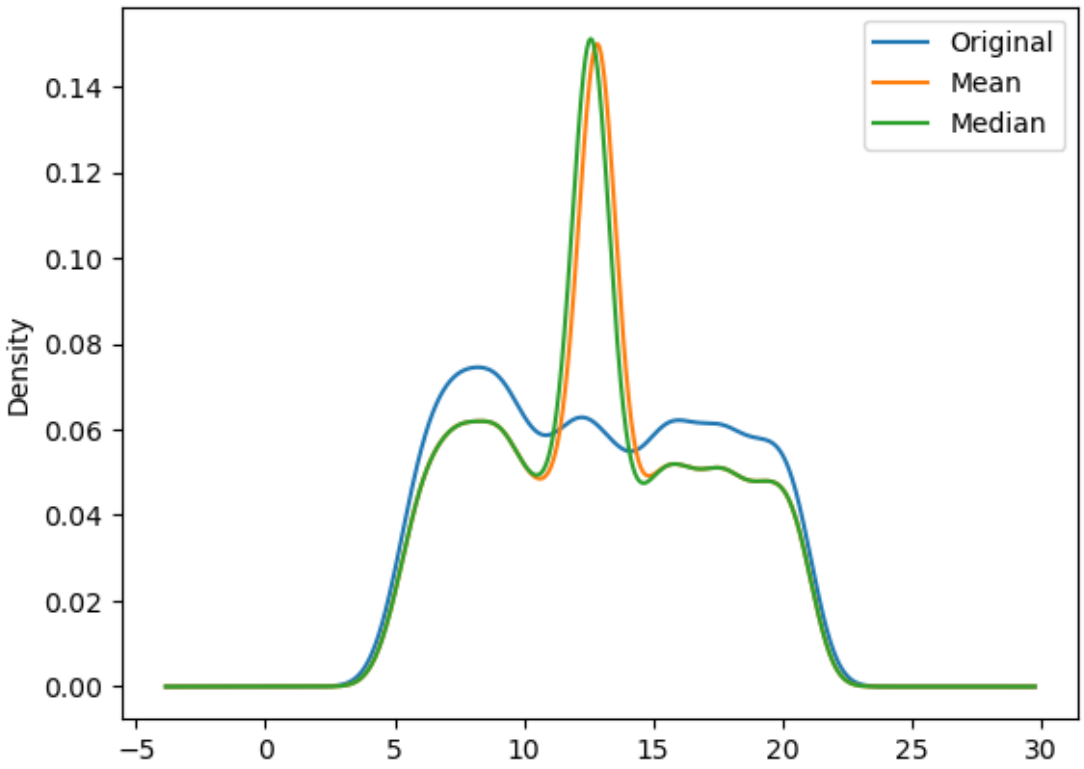

Figure 1.2: Item Weight variance after mean & Mediian imputation

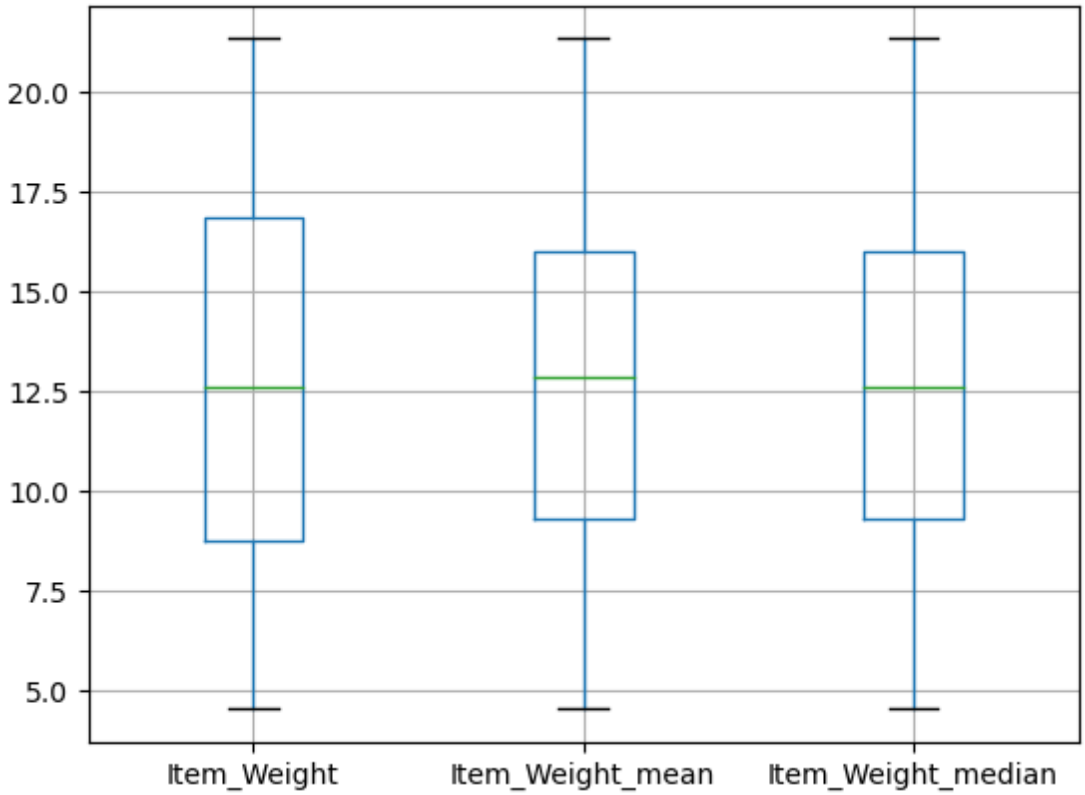

Figure 1.3: Item Weight variance after mean & Mediian imputation

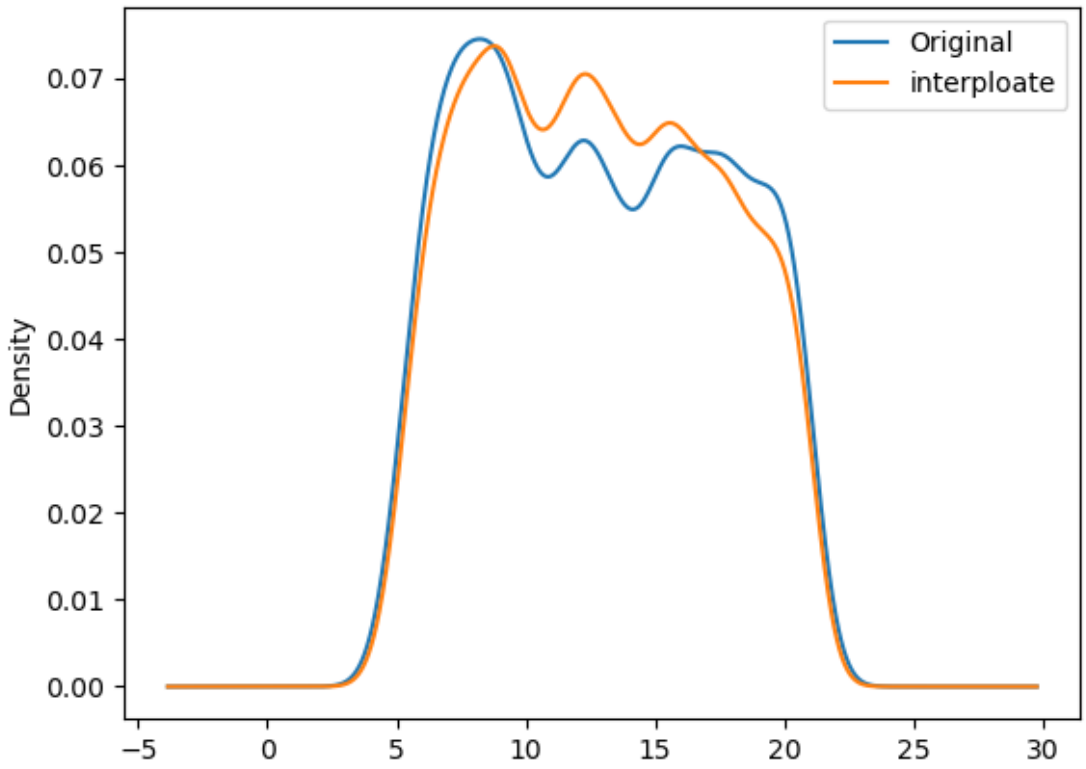

Figure 1.4: After Linear interpolating the Item Weight

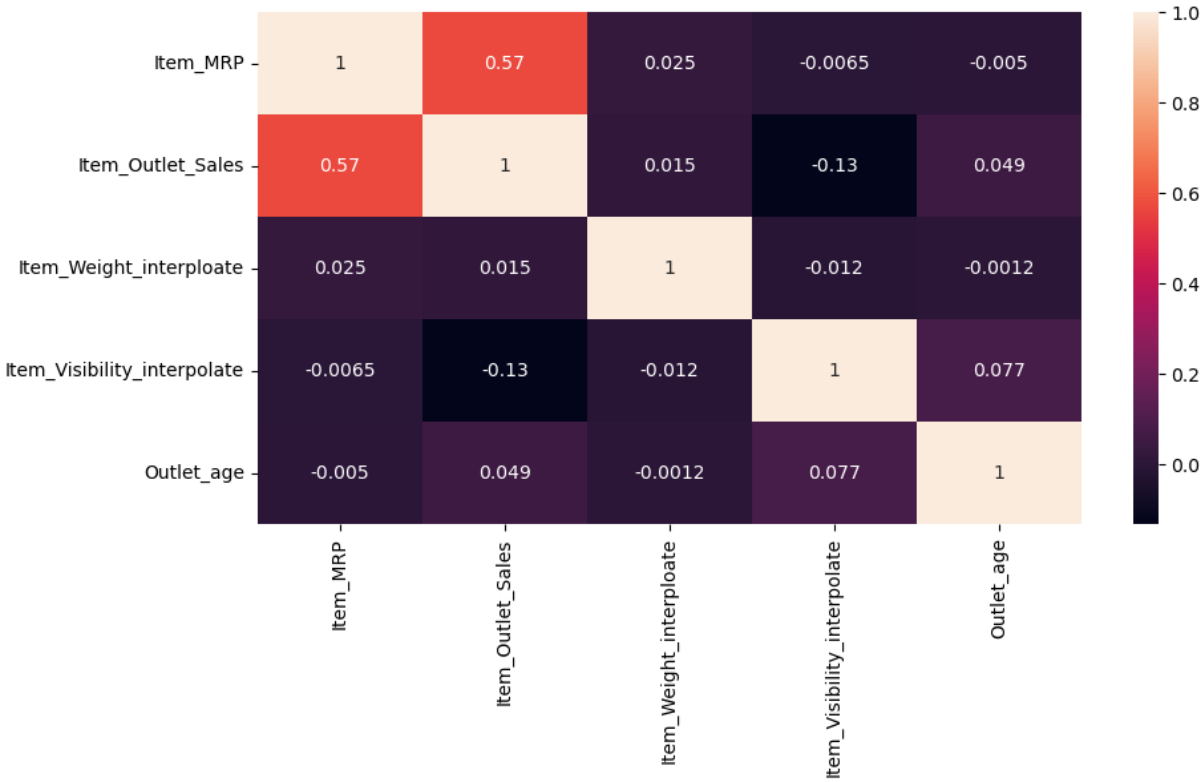

Figure 1.5: Cross Validation Matrix

Data Splitting

For machine learning to be successful, data must be available. Test data are required in additionto the training data in order to evaluate how efficiently the algorithm works, although in this case, the training and testing dataset are distinct. We must separate training and testing in our process into x train, y train, x test, and y test. The process of breaking accessible data into two pieces, often for cross-validator needs, referred to as data splitting. A portion of the data is applied to create a prediction model, while another portion is utilised to assess the effectiveness of the model.

Regression Algorithms

We must incorporate machine learning algorithms into our method.

XGBoost Regression

XGBoost, short for Extreme Gradient Boosting, stands as a powerful and versatile machine learning algorithm that has gained immense popularity in the data science community. It falls under the category of gradient boosting frameworks and is particularly well-suited for regression and classification tasks. What sets XGBoost apart is its ability to deliver high performance and efficiency. The algorithm achieves this through a combination of gradient boosting principles, regularization techniques, and a focus on model simplicity. XGBoost is designed to handle large datasets efficiently, making it a robust choice for real-world applications.One notable feature of XGBoost is its regularization approach, which helps prevent overfitting and enhances the model's generalization capabilities. It employs parallel and distributed computing strategies, making it scalable and suitable for large-scale datasets. The algorithm also incorporates tree pruning techniques to remove unnecessary components of the model, optimizing its structure. Additionally, XGBoost provides valuable insights into feature importance, enabling practitioners to identify key factors influencing predictions. With its well-balanced combination of accuracy, interpretability, and efficiency, XGBoost has become a cornerstone in various machine learning projects and competitions.

Random Forest Regressor

The Random Forest Regressor is a robust and versatile machine learning algorithm used for regression tasks. As an ensemble learning method, it operates by constructing a multitude of decision trees during training and outputs the average prediction of the individual trees for regression purposes. This algorithm is particularly effective in mitigating overfitting and improving accuracy compared to a single decision tree.The key idea behind the Random Forest Regressor is to introduce randomness in the tree-building process. Instead of relying on a single decision tree, it creates an ensemble of trees with each tree trained on a random subset of the data and using a random subset of features. This randomness helps in capturing diverse patterns within the dataset and prevents the

model from becoming too specialized.The Random Forest Regressor is known for its flexibility, ease of use, and capability to handle both numerical and categorical data. It is less sensitive to hyperparameters and requires minimal tuning compared to some other algorithms. Additionally, the algorithm provides a feature importance ranking, aiding in the interpretation of the model.

1.7 Result Analysis

Big marts' sales prediction is conducted by using many algorithms. We use machine learning algorithms to solve our dataset. Initially, we want to predict the sales of the mart by studying the sales of different marts with specific attributes, so that is why we set the "Price" attribute with the dependent variable, and we see above there are more than 10 attributes that we use as some independent variables. Our dataset contains two different train and test files; we concatenate our files to understand the data better. A hypothesis is necessary to check the possible attributes of the data. Also, it gives the understanding between the data scientist and the prediction. Therefore, we created some hypotheses and then compared them with the existing data. We saw a little bit of difference b/w the hypothesis and the data, and then we adjusted the data with our hypothesis attributes. Next, we moved to explore the data. In this part, we check the basic statistics of the dataset and missing values in the data. We found that three attributes have many missing values, and we will resolve these missing values in the coming section. Moved to the nominal variables, we checked the unique values in the data and found that there are 4-5 categorical variables. After, we need to impute the missing values further. Attributes named 'Product-weight' and 'Outlet-Size' filled the missing values, 2439 and 4016, respectively. We use mean and mode for the missing values product weight and outlet size, respectively. Feature engineering is the process of creating some new attributes for a better understanding of the data. Once we were ready with the data, we had to make a model. We used three machine learning models to predict sales and compare the results.When working with machine learning (ML) models, it is good to split the files into train and test, but using the built-in function of sci-kit learn lib is a good idea. The advantages of using a split function are avoiding over-fitting and under-fitting, so use a split function before using any machine learning algorithm.

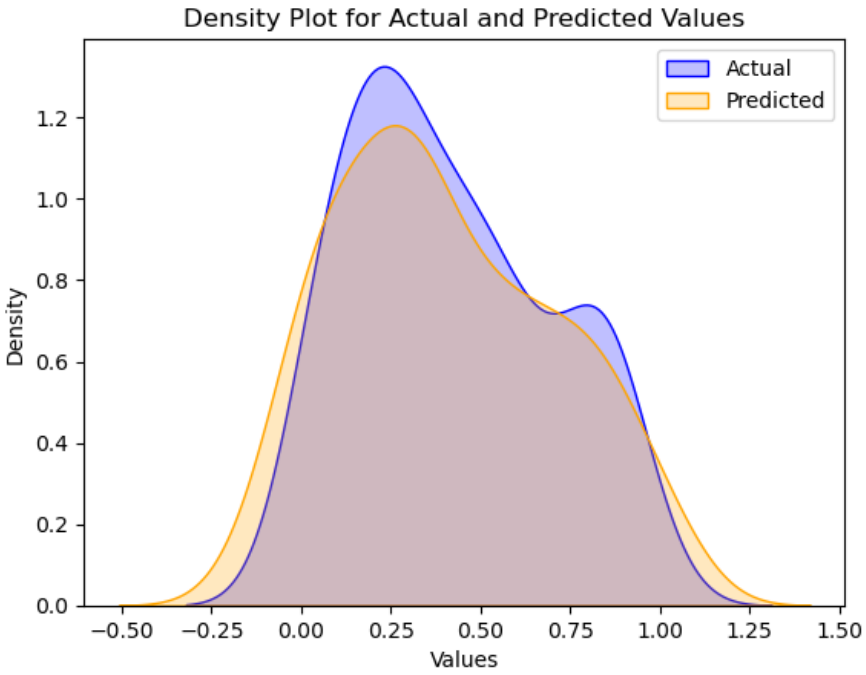

Figure 1.6: Density Plot for Actual Predicted Values

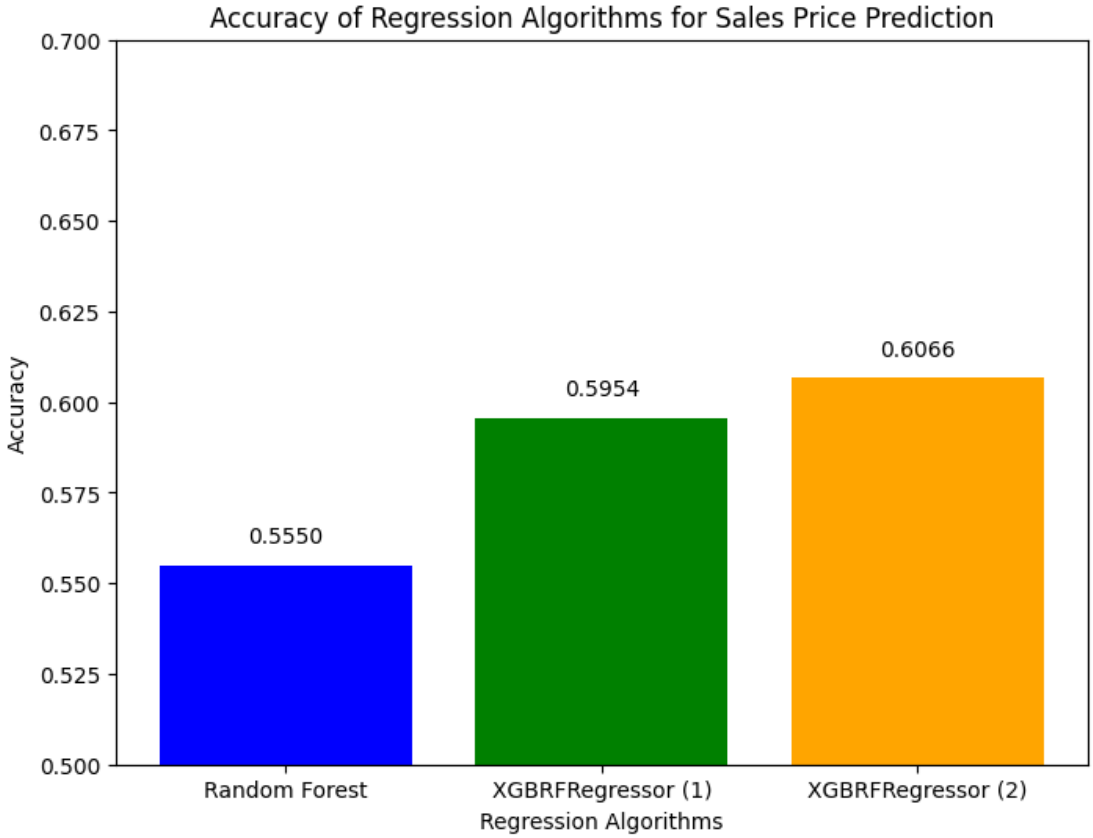

Figure 1.7: Final Accuracy of Our Regression Model

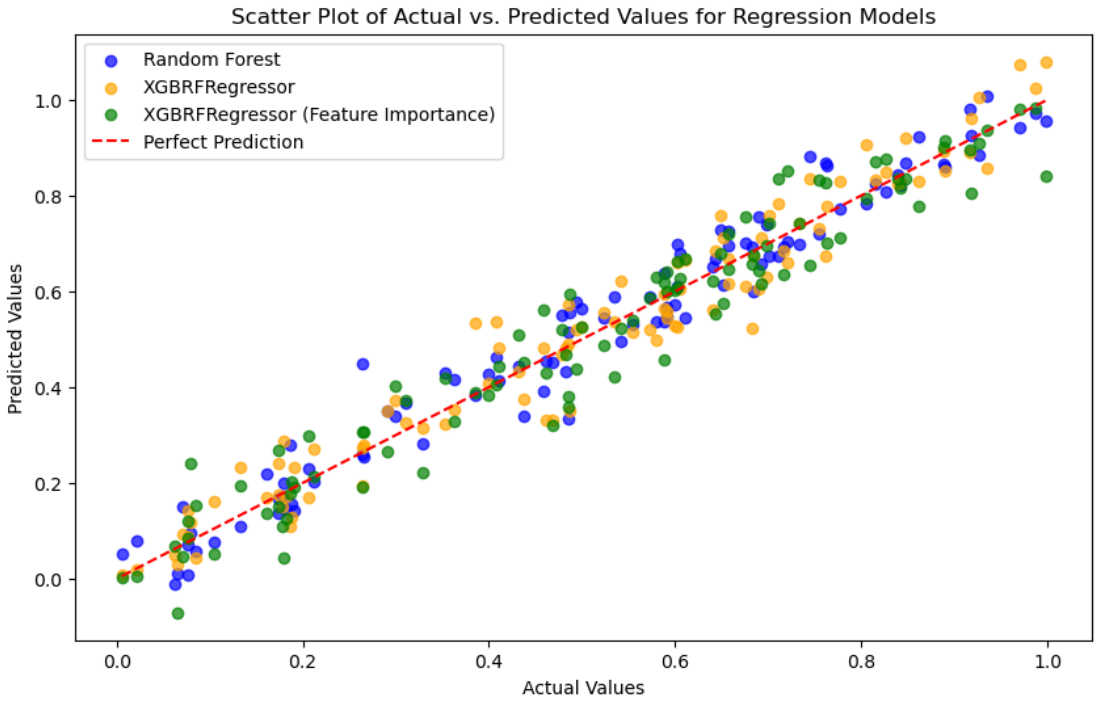

Figure 1.8: Scatter Plot of Actual VS Predicted Values for Regression Model

1.8 Limitations

- In our project we only used one dataset (Big Mart Sales) to evaluate the performance of the machine learning models, which may not be representative of other datasets or domains.
- Wer did not consider the effect of outliers or noise in the data, which may affect the accuracy and robustness of the models.
- We did not compare the machine learning models with other baseline methods or state-of-the-art approaches for sales prediction, which may limit the unreliability and validity of the results

1.9 Future Scope

Future Work and Applications of Big Mart Sales Prediction

- Time Series Analysis: Considering the temporal nature of sales data, incorporating time series analysis methods can be beneficial. This includes analyzing sales trends over time, identifying seasonality patterns, and developing models that explicitly consider temporal dependencies.
- Geospatial Analysis: Exploring the geographical aspects of store locations and their influence on sales could be a valuable addition. Geospatial analysis can uncover location-specific trends and help optimize inventory management and marketing strategies.
- Dynamic Pricing Strategies: Integration of dynamic pricing strategies based on real-time sales predictions can be an impactful application. Adaptive pricing models can optimize revenue by adjusting prices according to demand fluctuations and market conditions.
- Supply Chain Optimization: Extending the predictive model to optimize supply chain management by forecasting demand for different products can contribute to efficient inventory management. This can help in reducing costs and minimizing stockouts or overstock situations.
- Customer Segmentation: Developing models for customer segmentation based on purchasing behavior can enable targeted marketing efforts. Understanding customer segments can lead to personalized promotions and improved customer satisfaction.
- Integration with IoT and Sensors: Leveraging Internet of Things (IoT) devices and sensors in stores can provide real-time data on factors like foot traffic, product interactions, and environmental conditions. Integrating this data into sales prediction models can offer more granular insights.
- Cross-Channel Sales Prediction: Extending the model to predict sales across different channels, including online and offline, can provide a comprehensive view of overall business performance. This can guide omnichannel marketing and sales strategies.
- Impact on Sustainability: Exploring the environmental impact of sales and optimizing strategies for sustainable retail practices could be a novel avenue. This involves considering factors such as product shelf life, reducing waste, and minimizing the carbon footprint associated with distribution.

the future work on Big Mart sales prediction can focus on adopting advanced techniques, incorporating additional data sources, and expanding applications to optimize various aspects of retail operations and contribute to the broader field of data-driven decision-making in retail.

1.10 Conclusion

This project proposed a framework that predicts mart's sales using a machine learning model and different techniques. This Project uses the data of various marts and then combines and analyses the data so that any mart can check the product's demand and sales overall. This forecasting helps the retailers to set the stock quantity more accurately. Next, we use many models to study the scores outcomes. Considering the outcomes, it is suitable for present data, but extensive data might be unsuitable or change the model selection. For more accuracy, we need a massive amount of data with minimum outliers. Therefore, the authors seek to check the individual product demand in a particular area in future work. Further, in the future, a retailer checks the score of a specific product by entering product attributes and its store's information, like location, and culture. Also, we consider an online App for the costumer's review regarding the stores and specific products for future work. This App works as a ranking App. Customers rank the stores by giving feedback; this helps the other customers to move on towards the stores. Sort of forum allows checking the demand for a specific store's specific product. Also, this kind of portal helps the retailer compare the stocks and scores of the public leader board

References

[1] Joan Nnadi. *Prediction of Changes in Sales Forecast Models During a Crisis*. PhD thesis, The College of St. Scholastica, 2023.

[2] Adilakshmi Konda, Rahul Bandaru, Manish Manchala, Krishna Teja Naraharisetty, and Ashwin S Thankachan. Predictive analysis for big mart sales using machine learning. In *AIP Conference Proceedings*, volume 2754. AIP Publishing, 2023.

[3] Shruti Shivankar, Shardul Mehetar, Neha Darade, Saachi Bhimanpalli, and Dnyanada Dafale. Global superstores sales prediction and data visualization using power bi. *International Journal of Research in Engineering, Science and Management*, 6(4):90–94, 2023.

[4] Kamil Samara and Mark Stanich. Using machine learning to predict grocery sales. In *Science and Information Conference*, pages 936–943. Springer, 2023.

[5] P Srinivasa Rao, N Sanjay, and K Anusha Reddy. Forecasting future sales of bigmarts.

[6] Yutian Shi. Sales prediction optimization via gradient boosting decision tree. *Highlights in Science, Engineering and Technology*, 44:121–128, 2023.

[7] R Sujatha and B Uma Maheswari. Sales prediction and conversion. In *Marketing Analytics*, pages 243–279. Apple Academic Press, 2023.

[8] R Sujatha and B Uma Maheswari. Sales prediction and conversion. In *Marketing Analytics*, pages 243–279. Apple Academic Press, 2023.

[9] Johanes Fernandes Andry, Henny Hartono, Honni Honni, Deny Deny, and Jeffrey Jo. Analysis and prediction supermarket sales with data mining using rapidminer. In *AIP Conference Proceedings*, volume 2693. AIP Publishing, 2023.