IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FORECASTING FUTURE SALES THROUGH THE APPLICATION OF MACHINE LEARNIG ALGORITHEMS

[1]Hemant Ganapati Devadig, [2]Janani G Hegde, [3]Jayasurya, [4]S. Jeenita

[1] Student, [2] Student, [3] Student, [4] Student
[1]Computer Science and Engineering
[1]K V G College of Engineering, Sullia, India

Abstract: This report discusses the use of machine learning to predict future sales for different products in various retailers. Currently, supermarket run-centers, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using XGBoost, Linear regression, Decision Tree , Random Forest, Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.

Index Terms* – sales prediction, machine learning, XGBoost, *Linear regression

1. INTRODUCTION

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems. In this paper, we are addressing the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc, are used for prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex.

2. Need of The Study

Everyday competitiveness between various shopping centers as and as huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market seeks to offer personalized and limited-time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. "To find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales." In order to help Big Mart, achieve this goal, a predictive model can be built to find out for every store, the key factors that can increase their sales and what changes could be made to the product or store's characteristics

3. Related Work

Machine learning has been a subject undergoing intense study across many different industries and fortunately, companies are becoming gradually more aware of the various machine learning approaches to solve their problems. However, in order to fully harvest the potential of different machine learning models and to achieve efficient results, one needs to have a good understanding of the application of the models and of the nature of data. This approach is used to analyze historical sales data to identify trends and patterns that can be used to predict future sales. It involves building a model that predicts future sales based on a set of independent variables such as product features, pricing, and marketing campaigns. The proposed approach involves building complex neural networks with multiple layers to analyze large amounts of data and make accurate predictions. These approaches have been widely used in research and industry for predicting future sales using machine learning techniques.

4. Tools and Framework used

The proposed framework for future sales prediction requires a processor of Intel core i3 or higher, a RAM of 4 GB or higher, and a hard disk drive of 20 GB. Peripheral devices such as monitor, mouse, and keyboard are also required. In terms of software requirements, the framework requires an operating system of Windows 8/10, an IDE tool of Jupyter Notebook, and coding language of Python 3.6. Additionally, APIs such as Numpy, Pandas, and Matplotlib are also required. These software and hardware systems are essential for building an effective machine learning model for future sales prediction. However, it is important to note that the choice of software and hardware systems may vary depending on the specific requirements of the task and available resources.

5. Dataset Selection

Dataset selection is the process of choosing a relevant and representative dataset for a specific machine learning task. In the context of future sales prediction using machine learning, dataset selection involves identifying and collecting historical sales data from various sources such as point-of-sale systems, customer databases, and marketing campaigns. The quality and quantity of the dataset can significantly impact the accuracy of future sales predictions. Therefore, it is essential to select a dataset that is large enough to capture different patterns and trends in sales data while also being representative of the target population or market. Big Mart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities.

Table 1. Attribute Information

Attribute	Description
Item_Identifer	It is the unique product Id number.
Item Weight	It will include the product's weight.
Item_Fat_Content	It will mean whether the item is low in fat or not.
Item -Visibility	The percentage of the overall viewing area assigned to the particular item from all items in the shop.
Item -Type	To which group does the commodity belong
Item-MRP	The product's price list
Outlet-Identifier	a distinct slot number
Outlet-Establishment Year	The year that the shop first opened its doors.
Outlet-Size	The sum of total area occupied by a supermarket.
Outlet-Location	The kind of town where the store is situated.
Outlet-Type	The shop is merely a supermarket or a grocery store.
Item-Outlet-Sales	The item's sales in the original shop

6. Methodology

The proposed approach for developing effective machine learning models for future sales prediction comprises four algorithms, a stacking ensemble technique, and a particular approach to feature selection. The methodology encompasses crucial steps of data pre-processing and feature engineering and is outlined in a flowchart with seven steps. The first step entails loading the data into the environment, followed by data pre-processing in step two. In step three, feature selection is executed to identify important features. In step four, the data is divided into train and test sets. Step five involves training data using different classifiers, while step six passes the test data to all trained classifiers. Finally, in step seven, evaluation is performed using a confusion matrix measure accuracy and precision.

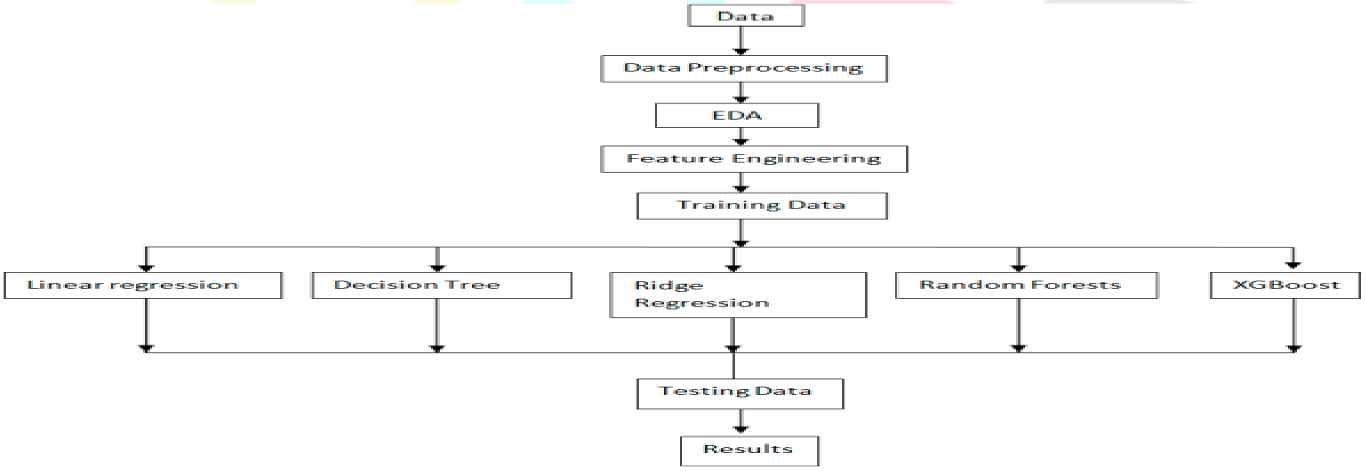


Fig.1 Proposed Architecture Diagram

6.1 Data Acquisition

The data acquisition step is the process of collecting and gathering data from various sources for future sales prediction using machine learning. In this step, the data is obtained from different sources such as online databases, surveys, or other relevant sources. The collected data may include information such as product attributes, sales figures, customer demographics, and other relevant variables that can impact future sales. The quality of the data is crucial in this step as it directly affects the accuracy and reliability of the machine learning model. Therefore, it is important to ensure that the collected data is accurate, complete, and relevant to the task at hand. Once the data has been acquired, it needs to be pre-processed and cleaned to remove any inconsistencies or errors before being used for further analysis. This step lays the foundation for building an effective machine learning model for future sales prediction by providing a reliable source of information on which the model can be trained.

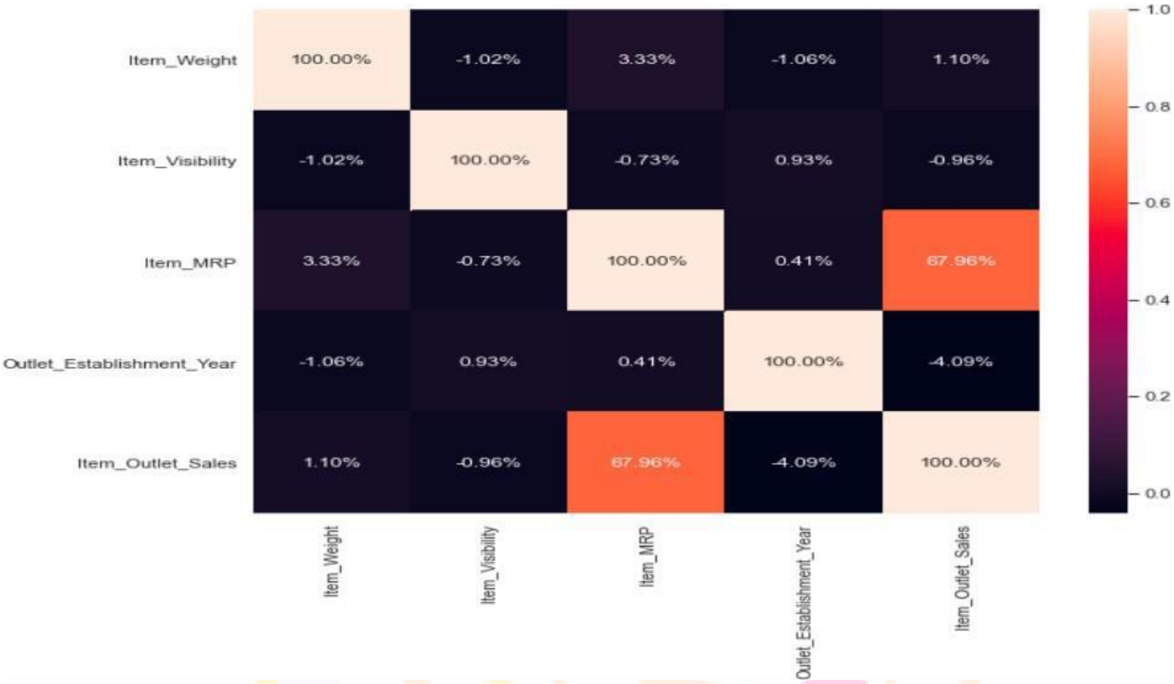


Fig.2 Heatmap showing the correlation between attributes

6.2 Data Preprocessing

Data preprocessing is a crucial step in the methodology future sales prediction using machine learning. In this step, the collected data is cleaned, transformed, and prepared for further analysis. The data may contain inconsistencies, missing values, or outliers that can affect the accuracy of the machine learning model. Therefore, it is important to preprocess the data to ensure that it is consistent and reliable. The preprocessing step involves several sub-steps such as data cleaning, feature scaling, and handling missing values. In this study, some columns such as Item_Identifier and Outlet_Identifier were removed as they did not contribute to attaining the results of the algorithm. Additionally, missing values were handled by imputing them with mean or mode values depending on their distribution. Feature scaling was also performed to ensure that all features are on a similar scale and have equal importance in the model. This step helps to improve the accuracy of the machine learning model by ensuring that it is trained on clean and consistent data.

6.3 Feature engineering

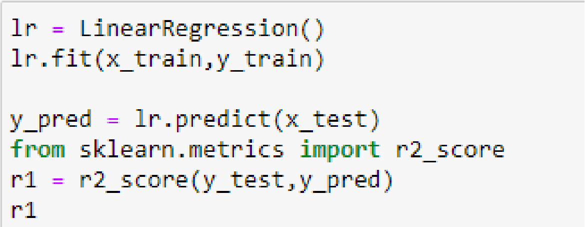
In Feature engineering, new features are created from the existing ones to improve the accuracy and performance of the machine learning model. Feature engineering involves several techniques such as feature scaling, outlier removal, and handling missing values. For instance, in this study, feature scaling was performed to convert data into a precise and adaptable size to improve accuracy and reduce error. Outliers were also removed or excluded from the dataset for better performance. Additionally, missing values were replaced by taking the mean value of the column. The goal of feature engineering is to build raw data features that help facilitate the machine learning process and enhance its predictive capability. By creating new features that capture important information about the data, feature engineering can help improve the accuracy of machine learning models and make them more effective in predicting future sales.

6.4 Algorithm Selection

Algorithm selection is an important step in the methodology presented in the PDF for future sales prediction using machine learning. In this step, an appropriate algorithm is selected based on the characteristics of the data and the problem at hand. The selection process involves several factors such as the size of the dataset, the number of features, and the type of problem being solved. he selection process involves evaluating different algorithms based on their performance metrics such as accuracy, precision, recall, and F1 score. Once an appropriate algorithm has been selected, it can be used to build a predictive model that accurately predicts future sales based on new data inputs.

6.4.1 Linear Regression

One of the most essential and commonly used regression techniques is linear regression. It's one of the most basic regression techniques. In this Random Error Regardless of how well the model is trained, tested, and validated, there will always be a variation between observed and predicted, which is irreducible error, so we cannot rely entirely on the learning algorithm's predicted results. Data must meet several conditions for a successful linear regression model. One of them is the lack of multiple linear regression, which means that the independent variables should be correlated.

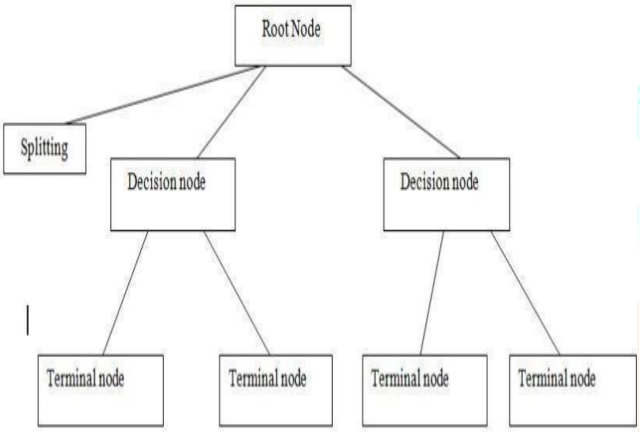


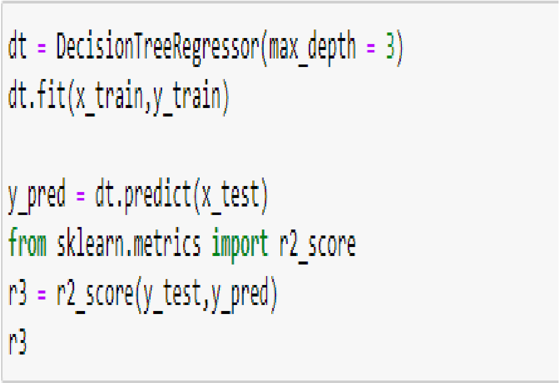
```
lr = LinearRegression()
lr.fit(x_train,y_train)

y_pred = lr.predict(x_test)
from sklearn.metrics import r2_score
r1 = r2_score(y_test,y_pred)
r1
```

Fig.3 Testing with linear regression

6.4.2 Decision Tree

A decision tree is a machine learning algorithm used for classification and regression analysis. It consists of nodes and branches that represent conditions and outcomes, respectively. The algorithm recursively splits data into subsets based on the most significant attribute or condition, resulting in a tree-like structure that can be used for prediction. Decision trees can handle both categorical and numerical data, are easy to interpret and visualize, and require little data preparation. However, they can overfit if not properly pruned or regularized. Decision trees are useful for explaining complex models to non-technical stakeholders.





```
dt = DecisionTreeRegressor(max_depth = 3)
dt.fit(x_train,y_train)

y_pred = dt.predict(x_test)
from sklearn.metrics import r2_score
r3 = r2_score(y_test,y_pred)
r3
```

Fig.4 Structure of Decision Tree

Fig.5Testing with Decision Tree

6.4.3 Ridge Regression

Ridge regression is a linear regression algorithm that adds a penalty term to the cost function of the model to prevent overfitting in high-dimensional data. It reduces the magnitude of coefficients by adding an L2 regularization term that penalizes large coefficients and shrinks them towards zero, thus reducing their impact on predictions. Ridge regression is particularly useful in identifying relevant features in high-dimensional data by reducing the impact of less important features. It does not result in sparse models and retains all features while reducing their impact on final predictions. Overall, ridge regression is a powerful technique for preventing overfitting in linear regression models with high-dimensional data.

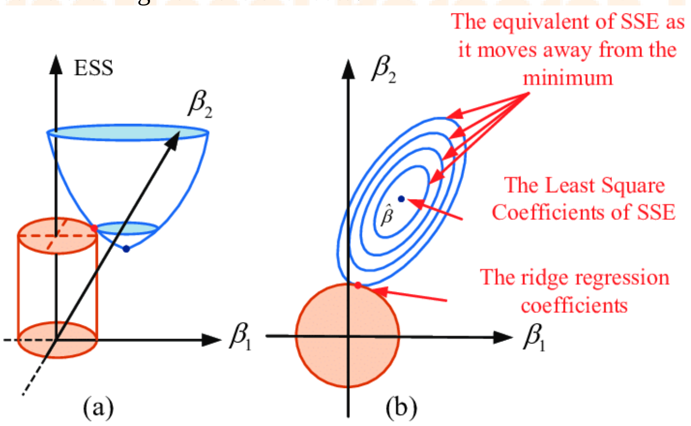

ESS
β_2
β_1
(a)
β_2
$\hat{\beta}$
β_1
(b)
The equivalent of SSE as it moves away from the minimum
The Least Square Coefficients of SSE
The ridge regression coefficients

Fig.6 Ridge Regression Algorithm

6.4.4 Random Forest

Random forest is an ensemble learning algorithm used for classification, regression, and other tasks. It creates many decision trees, each trained on a random subset of the data and features, to improve accuracy and robustness in machine learning models. During prediction, each tree independently predicts the outcome based on its own set of rules, and the final prediction is made by aggregating the predictions from all trees in the forest. Random forests are less prone to overfitting and can handle missing values and noisy data well. They are widely used in various applications, especially in high-dimensional data.

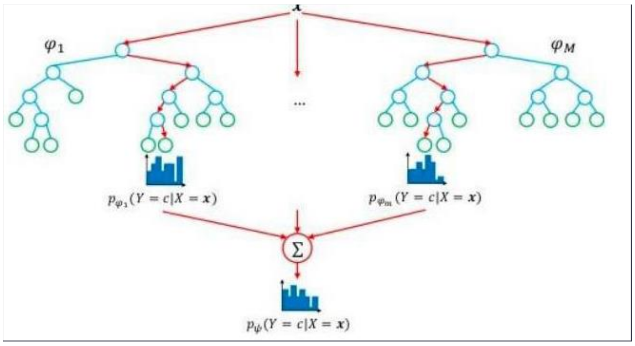

x
φ_1
φ_M
...
$p_{\varphi_1}(Y = c|X = x)$
$p_{\varphi_m}(Y = c|X = x)$
Σ
$p_{\varphi}(Y = c|X = x)$

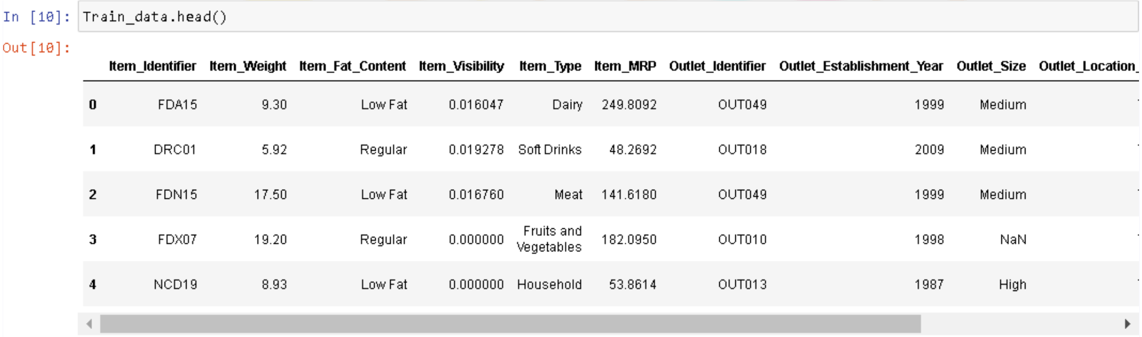
Fig.7 Random Forest

6.4.5 XGBoost Regression

XGBoost Regression is a machine learning algorithm that combines multiple decision trees using gradient boosting to improve the accuracy and robustness of the model. It can be used for regression and classification tasks and is known for its ability to handle missing values and noisy data well. It includes regularization techniques to prevent overfitting and achieves state-of-the-art performance on many benchmark datasets. Overall, XGBoost Regression is a powerful tool for structured data problems where there are many features or predictors.

6.5 Model Training

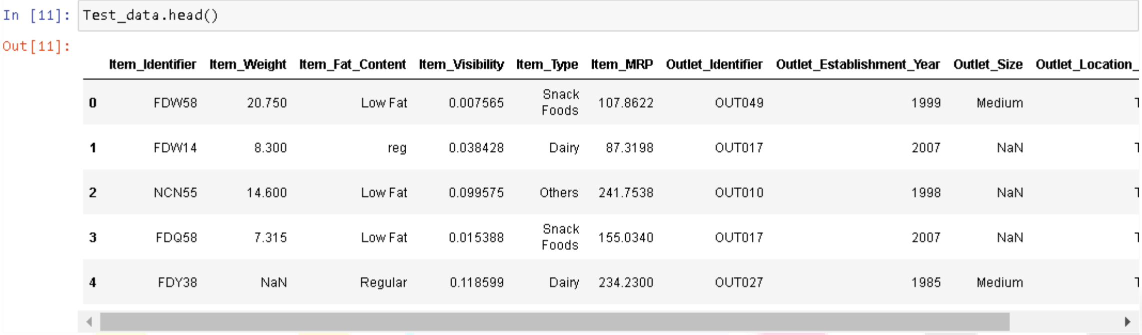
Once an algorithm has been selected, we can train a model on the training set using that algorithm. During training, we adjust the parameters of the algorithm to minimize a loss function that measures how well the model fits the training data. Model training and evaluation are important steps in the process of future sales prediction using machine learning. In this context, model training refers to the process of building a predictive model using historical sales data, while model evaluation refers to the process of assessing the performance of the model on new, unseen data.



```
In [10]: Train_data.head()
```

Out[10]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	



```
In [11]: Test_data.head()
```

Out[11]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_
0	FDW58	20.750	Low Fat	0.007565	Snack Foods	107.8622	OUT049	1999	Medium	T
1	FDW14	8.300	reg	0.038428	Dairy	87.3198	OUT017	2007	NaN	T
2	NCN55	14.600	Low Fat	0.099575	Others	241.7538	OUT010	1998	NaN	T
3	FDQ58	7.315	Low Fat	0.015388	Snack Foods	155.0340	OUT017	2007	NaN	T
4	FDY38	NaN	Regular	0.118599	Dairy	234.2300	OUT027	1985	Medium	T

Fig.8 Train and Test the Data

- The first step in model training is to prepare the data by cleaning and preprocessing it. This may involve removing missing values, scaling, or normalizing features, and encoding categorical variables. Once the data is prepared, it can be split into training and validation sets.

- The next step is to select an appropriate machine learning algorithm for the task at hand. In this case, we might consider using algorithms such as random forest regression or XGBoost regression that are well-suited for predicting sales based on historical data.
- After training is complete, we evaluate the performance of our model on a separate validation set. This allows us to assess how well our model generalizes to new, unseen data. We might use metrics such as mean squared error or R-squared to evaluate our model's performance. If our model performs well on the validation set, we can then use it to make predictions on new sales data. However, it's important to note that even a well-performing model may not always make accurate predictions due to changes in market conditions or other factors outside of our control.

6.6 Predictions and Evaluation

After developing predictive models based on different iterations outlined in the experiment plan, the next step is to use these models for prediction. Using the different versions of developed models, Item Outlet Sales are predicted. Each prediction function produces a vector of predicted sales using a particular model. For the prediction process, the fitted model object is the first argument, and the set of predictor variables used to predict sales is the second argument. In this work, the original train dataset was split into a validation set and a testing set. Therefore, for each model, two different predicted values were obtained - one with the predictors of the validation set and another with the predictors of the testing set. This is why each performance evaluation metric has two values for a given model.

The accuracy of future sales prediction using machine learning depends on several factors such as the quality and quantity of historical sales data, the selection of appropriate features, the choice of machine learning algorithm, and the accuracy of the model evaluation process. If these factors are carefully considered and optimized, it is possible to achieve high levels of accuracy in future sales prediction. The proposed model very well performed on XGBoost regression we got the accuracy 59.17%.Therefore, it's important to regularly monitor and update the predictive model to ensure its continued accuracy over time. It's important to note that predicting future sales production is not an exact science and there are many factors that can influence demand such as changes in consumer preferences or unexpected events like natural disasters. Therefore, it's important to regularly monitor our predictive models and update them as needed based on new information or changing market conditions.

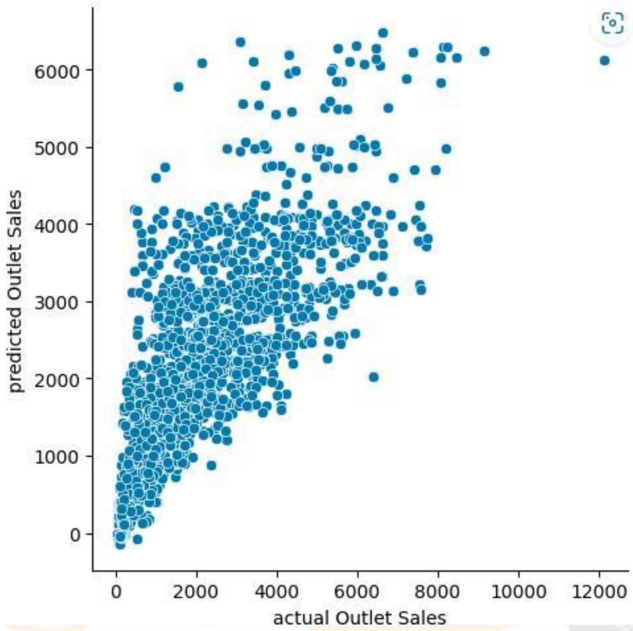


Fig. 9 Resultant Graph

7. Conclusion and future work

This methodology is primarily used by shopping marts, groceries, Brand outlets etc. The data analysis applied to the predictive machine learning models provides a very effective way to manage sales, it also generously contributes to better decisions and plan strategies based on future demands. This approach is very much encouraged in today's world since it aids many companies, enterprises, researchers and brands for outcomes that lead to management of their profits, sales, inventory management, data research and customer demand. In this proposed model We tested five algorithms XGBoost, Random Forest, Linear Regression, Decision Tree, Ridge regressor. From the results, we can conclude that among all the five algorithms XGBoost Gradient Regressor has the highest accuracy of 59.17% when distinguished together. Hence, we can say that XGBoost Gradient Regressor is the better algorithm for efficient sales analysis. This

With traditional methods not being of much help to the business organizations in revenue growth, use of Machine Learning approaches proves to be an important aspect for shaping business strategies keeping into consideration the purchase patterns of the consumers. Prediction of sales with respect to various factors including the sales of previous years helps businesses adopt suitable strategies for increasing sales and set their foot undaunted in the competitive world. Predicting future sales production is a complex process that requires careful consideration of various factors such as historical sales data, market trends, and other relevant features. Machine learning algorithms provide a powerful tool for analyzing large amounts of data and identifying patterns and trends that can be used to make accurate predictions. By using machine learning to predict future sales production, businesses can optimize their production schedules, manage inventory levels more effectively, and allocate resources more efficiently. However, it's important to note that predictive models are not perfect and may require regular monitoring and updating to ensure continued accuracy over time. Overall, the use of machine learning for predicting future sales production has the potential to significantly improve business operations and increase profitability.

ACKNOWLEDGMENT

The successful fulfilment of this paper depends on many people whom I would like to take a moment to thank. First, I would like to thank Department of Computer science and engineering KVG College of Engineering Sullia, Manglore. My heartfelt gratitude is Co-author Department of Computer science and engineering KVG College of Engineering Sullia, Manglore.For their guidance, motivation and for sparing his extremely precious time for me.

REFERENCES

[1] Singh Manpreet, Bhawick Ghutla, Reuben Lilo Jnr, Aesaan FS Mohammed, and Mahmood A. Rashid. "Walmart's Sales Data Analysis-A Big Data Analytics Perspective." In 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 114-119. IEEE, 2017.

[2] Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019).

[3] Panjwani, Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar, and Rupali Hande. Sales Prediction System Using Machine Learning. No. 3243. EasyChair, 2020.