PREDICTIVE ANALYSIS OF BIG MART SALES USING XGBOOST

Sanjaykumar S

ABSTRACT

Big Mart Sales prediction is an integral part of the retail industry, enabling companies to optimize inventory management, plan effective marketing strategies and make informed decisions. This study proposes a sales prediction system for Big Mart using the Extreme Gradient Boosting (XGBoost) algorithm, a powerful ensemble learning technique. Using historical sales data, product attributes, customer data and external factors, the system aims to accurately predict future sales volumes. The implementation process involves several steps, including data processing, feature selection, train test splitting, XGBoost model training, and performance evaluation. Historical sales data is carefully processed to ensure data integrity, and key features are designed to capture important patterns and trends. Feature selection techniques are used to identify the most informative attributes that contribute to accurate predictions. The XGBoost algorithm is then used to build a predictive model by iteratively combining weak decision tree models to form a robust ensemble. The model is trained on historical sales data and learns from mistakes in previous iterations to continuously improve forecast accuracy. Hyperparameter tuning is performed to optimize model performance. To evaluate the effectiveness of the system, a separate test series is used to evaluate the accuracy of sales forecasts. Common evaluation metrics such as mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) are used to measure model performance against actual sales values. The accuracy of XGBoost algorithm is 91.14 % and the accuracy of Big Mart's prediction is 61.14 %

Keywords: *Big Mart, Inventory management, Forecast sales, Hyperparameters, XGBoost.*

1 INTRODUCTION

1.1 Background

The daily competition between various commercial centers and giant markets is becoming tighter, more violent because of the rapid development of global commercial centers also in electronic commerce. Each market aims to offer personalized and limited-time offers to attract a large number of customers based on the period, so that the sales volume of each product can be estimated for the storage, transportation and logistics services of the organization. The current machine learning (ML) algorithm is very advanced and provides methods to forecast or predict sales for all kinds of organizations, which is very useful to overcome the low price used in forecasting. Better forecasting is helpful for developing and improving marketing strategies. which is also extremely useful. The ML techniques predict the sales of different products in different stores of a retail chain called "Big Mart". The dataset used in this method contains information about the products, stores and sales of the retail chain. The goal is to develop a forecasting model that can accurately predict the sales of each product in each store in the coming year. A retailer can use this to predict future market demand and adjust their inventory levels accordingly. The accuracy of these forecasts determines whether the trader makes a profit or a loss. Once the model is trained, it can be used to predict new data. Input data is fed into a trained model that applies the learned patterns to predict outcomes. It is possible to predict the sales of products in Big Mart stores with great accuracy. Retailers can use this information to optimize inventory management and improve their ability to forecast sales

Objectives

The goal of Big Mart Sales is to develop an accurate and reliable model that can predict Big Mart sales or revenue based on various factors such as store size, location, product price, promotions and customer behaviour. By predicting sales or revenue, ML models can help retailers optimize their inventory levels. This can ensure the store has the right products in stock at the right time, increasing customer satisfaction and reducing waste. The XGBoost algorithm is known for its accuracy and speed and has been used successfully in many applications, including sales forecasting. One of the main advantages of using XGBoost is that it can handle both categorical and numerical data, making it good for sales forecasting where both types of data may be present. In addition, XGBoost can be used to analyse the impact of various factors on sales and identify opportunities for growth. It can be used, for example, to identify which products or categories drive sales, which offers are most effective, and which factors may affect seasonal sales fluctuations.

1.2 Constraints

XGBoost, an optimized gradient boosting algorithm, is widely used for sales prediction in Big Mart or retail scenarios. Although XGBoost is a powerful predictive modelling tool, it also has some limitations that must be considered when forecasting Big Mart sales. The challenges of proper data processing, feature design, model tuning and domain knowledge can help to create more accurate and robust sales forecasting models

.2 RELATED WORKS

T.K. Thivakaran, et al. (2022) [2] proposed about the Bigmart dataset and identify patterns and trends that can be used to predict future sales. The problem statement of the paper is to leverage machine learning techniques to analyse and predict sales data for Bigmart, with the aim of improving sales performance and maximizing profitability. The supervised learning algorithm used in the paper is Random Forest Regressor. This algorithm is a popular machine learning algorithm that can be used for both regression and classification tasks. It works by creating an ensemble of decision trees, where each tree is trained on a random subset of the features and data points. The output of the algorithm is then obtained by taking the average of the outputs of all the trees. Li, Yuanjiang, et al. (2021) [4] reported about the Clothing Sale Forecasting by a Composite Gated Recurrent Unit (GRU)-Prophet Model with an Attention Mechanism and the problem statement of this paper is to develop an accurate and robust forecasting model that can help the clothing industry to optimize their business strategies and improve their bottom line. The proposed composite GRU-Prophet model with an attention mechanism aims to address the challenges of clothing sales forecasting by leveraging the strengths of different algorithms and identifying the most relevant features for prediction, the composite GRU-Prophet model with an attention mechanism is a powerful algorithm for clothing sale forecasting that leverages the strengths of both the GRU and Prophet algorithms. The GRU captures the temporal dependencies in the data, while Prophet captures the seasonality and cyclical patterns. The attention mechanism helps to give more weight to the most relevant features, which leads to more accurate predictions The accuracy of the algorithm is 54.32\% this is the results of the composite GRU-Prophet model with an attention mechanism. Aguilar-Palacios, Caros, et al. (2019) [7] examined about the Forecasting Promotional sales within the Neighbourhood and the problem of the paper is to develop a predictive model for forecasting promotional sales within a specific neighbourhood. The paper focuses on the retail industry, where businesses use promotional sales to increase foot traffic, attract new customers, and increase sales revenue. The algorithm used in K-Nearest Neighbours (KNN) method that produces explainable and easy to modify predictions. Also, wanted to couple feature selection and prediction, which accomplished through a cost function that is minimized with a non-negative least squares (NNLS) solver. Our method is founded on online learning, so the predictions are calculated with the latest data available, thus avoiding to retrain the model. And the accuracy of the KNN algorithm is 62.23%

3 PROPOSED SYSTEM

The Bigmart sales forecasting using XGBoost involves the development of a ML model that can predict the sales of various products in Big Mart stores. The system uses historical sales data and other relevant variables to train the model and predict future sales. ML models are based on the assumption that patterns observed in historical data will persist into the future. The accuracy of the predictions is affected by the quality of the training data, the selected features, the complexity of the model and other factors. Regular monitoring of the model and retraining with updated data can help to maintain predictive performance over time.

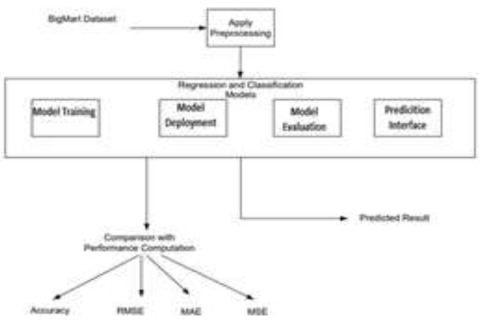


Fig.1. Architecture design of proposed system

Architecture diagram show the key components involved in Big Mart Sales Prediction Using XGBoost and their Relationships. The comparison with performance computation the accuracy is based on the predict of the Big Mart sales and the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE) is help to find the comparing of the accuracy

4 METHODOLOGY

In general, steps XGBoost uses gradient boosting to iteratively improve the predictions of the model. Gradient boosting involves creating a sequence of models, where each subsequent model is trained to improve upon the errors of the previous model. The models are combined using a weighted sum, where the weights are determined by their performance on the training data

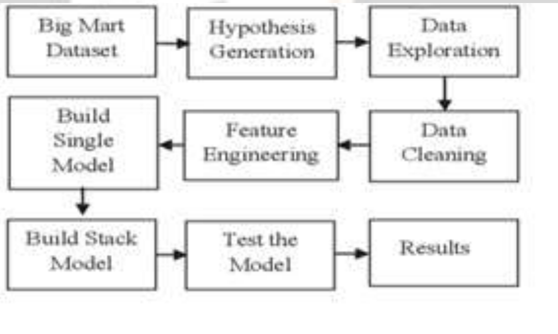
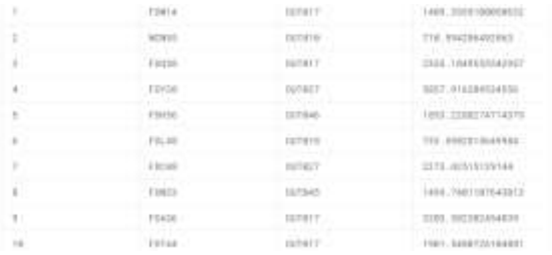


Fig.2. Project flow diagram

The main steps involved in Big Mart Sales Prediction using XGBoost, from data collection to model deployment and prediction. By the Hypothesis generation the dataset processed to the data exploration based on the feature engineering and finally the stack model is tested and get a result.

1.3 Dataset Collection

The primary focus of data collection is to gather historical sales data from Big Mart. This includes information such as the date of sale, store ID, product ID, sales volume, and possibly other relevant attributes like price, promotions, and discounts. Collecting a sufficient amount of high-quality sales data is essential to train the XGBoost model effectively.



1	FDW14	OUT017	1469.[illegible]
2	NCN55	OUT010	716.[illegible]
3	FDQ58	OUT017	2368.[illegible]
4	FDY38	OUT027	5057.[illegible]
5	FDH56	OUT046	1853.[illegible]
6	FDL48	OUT018	719.[illegible]
7	FDC48	OUT027	2173.[illegible]
8	FDN33	OUT045	1466.[illegible]
9	FDA36	OUT017	2100.[illegible]
10	FDT44	OUT017	1981.[illegible]

Fig.3. Dataset of Bigmart sales

Dataset is created by collecting the good pixel weapon images and making them ready for the creation of the dataset.

1.4 Data Pre-processing

Pre-processing the dataset is an essential step before training an XGBoost model for sales prediction. It involves cleaning the data, handling missing values, transforming features, and scaling the data appropriately, the dataset prepares it for effective model training and ensures that the data is in a suitable format for XGBoost. It helps address data quality issues, handles missing values, transforms features appropriately, and prepares the data for the subsequent steps of model training and evaluation.

4.3 Data Exploration

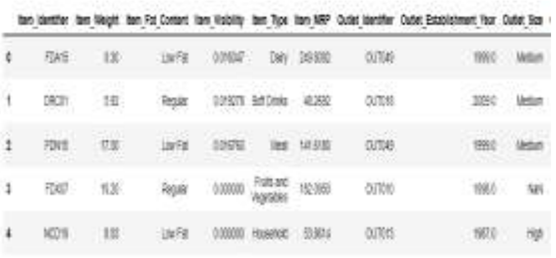
It helps to understanding the dataset's structure, characteristics, and relationships, enabling you to make informed decisions during preprocessing and model development. It assists in feature selection, identifying outliers, understanding patterns, and leveraging domain knowledge to build a more accurate and robust sales prediction model using XGBoost. Calculate basic descriptive statistics such as mean, median, standard deviation, minimum, maximum, and quartiles for numerical variables. This provides a summary of the distribution and variability of the data.

4.4 XGBoost Algorithm

XGBoost is ability to handle complex relationships, handle missing data, and provide feature importance analysis makes it a powerful algorithm for sales prediction in Big Mart. By fine-tuning the model and leveraging its strengths, it is possible to develop accurate and robust sales prediction models that assist in optimizing inventory management, pricing strategies, and decision-making in the retail domain.

4.5 Model Training

Configure the chosen model with appropriate hyperparameters. Train the model using the training data. The model will learn the patterns and relationships between the input features and the sales target variable. Evaluate the trained model using the testing set. Calculate relevant evaluation metrics such as root mean squared error (RMSE), mean absolute error (MAE), or R-squared to assess the model's performance in predicting sales. Compare the model's performance against baseline models or previous approaches.



	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999.0	Medium
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009.0	Medium
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999.0	Medium
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998.0	NaN
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987.0	High

Fig.4. Training the dataset

4.6 Model Testing

Testing the XGBoost model on unseen data allows you to assess its predictive capabilities and gain confidence in its ability to generalize to real-world scenarios. By evaluating the model's performance, you can make informed decisions about its suitability for sales prediction in the Big Mart setting and identify opportunities for further improvement. XGBoost. Analyse the evaluation metrics to understand the performance of the XGBoost model on the testing data. Assess whether the model meets the desired level of accuracy and reliability for sales prediction in the Big Mart context. Compare the results with the performance achieved during the training and validation stages.

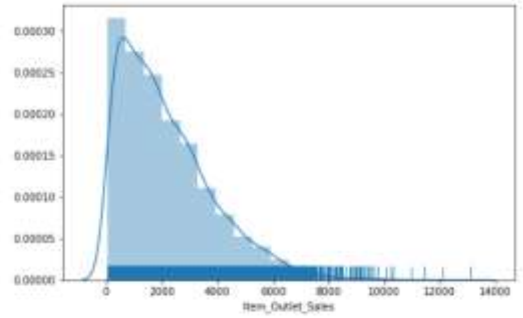


Fig.5. Sales Item Outlet Sales

4.7 Internal Design

The internal design of XGBoost combines techniques such as gradient boosting, regularization, parallel processing, and optimized algorithms to achieve high accuracy, scalability, and efficiency in training predictive models. These design choices contribute to XGBoost is popularity and its effectiveness in various domains, including sales prediction in Big Mart.

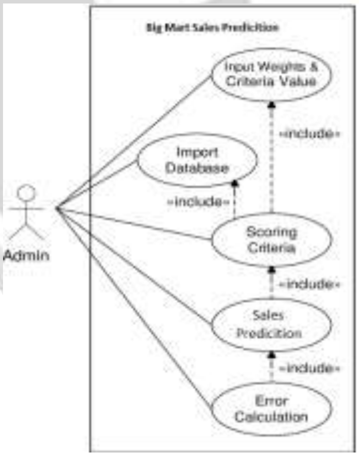


Fig.7. Use case of the system

5 RESULT

The accuracy of the model, the performance of the hardware used for training and prediction, and the complexity of the problem being solved. The Proposed system can be made more efficient by optimizing the hyperparameters of the XGBoost model, such as the number of tree, learning rate, and regularization parameters. This can help to reduce the training time and improve the accuracy of the model dimensionality to improve the presentation.

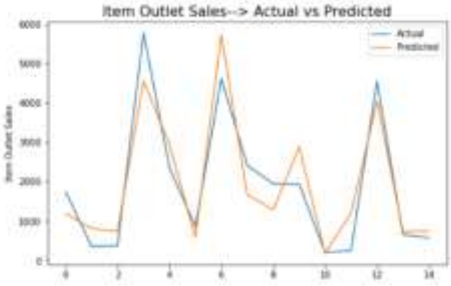


Fig.8. Result of detected Pistols

6 FUTURE WORK

The accuracy of the XGBoost model can be improved by incorporating additional data sources, such as weather patterns, and economic indicators. It involves creating new features from the existing dataset that may improve the predictive performance of the model. This can include transformations of existing features, combining multiple features, or creating entirely new features based on domain knowledge

7 CONCLUSION

The XGBoost algorithm is well-suited for this task, as it is capable of handling large dataset with complex features and relationships, and can be tuned to optimize predictive accuracy. By using XGBoost to predict sales, pricing strategies, and promotional activities. This can lead to increased profitability and customer satisfaction, as well as improved overall business performance. However, as with any machine learning model, it is important to carefully evaluate and tune the XGBoost model to ensure the best possible performance. This includes selecting appropriate features, pre-processing the data, optimizing hyperparameters, and validating the model using appropriate metrics and techniques. The accuracy of the XGBoost algorithm is 91.14% And the highest accuracy of Big Mart sales prediction is 61.14%

8 REFERENCES

[1] Anwer, Mussab Osamah, and Sureyya Akyuz. "Sales Forecasting of a Hypermarket: Case Study in Baghdad Using Machine Learning. "2022 30th Signal Processing and Communications Applications Conference (SIU). IEEE,2022.

[2] Thivakaran T.K., M. Ramesh, "Exploratory Data analysis and forecasting of Big Mart dataset using supervised and ANN algorithm" Measurement: Sensors23 2022.

[3] Aguilar-Palacios, Carlos, Sergio Munoz-Romero, and Jose Luis Rojo-Alvarez. "Casual Quantification of Cannibalization 2021.

[4] Li, Yuanjiang, et al. "Clothing sale forecasting by a composite GRU-Prophet model with an attention mechanism." IEEE Transactions on Industrial Informatics 17.12 2021.

[5] Naveenraj, R., and R. Vinayaga Sundharam. "Prediction Of Big Mart Sales Using Machine Learning" International Journal of Modernization in Engineering, Technology and sciene, Volume:03, Issue:09 2021.

[6] Suma, V., and Shavige Malleshwara Hills. "Data mining based prediction of demand in Indian market for refurbished electronics." Journal of Soft Computing Paradigm (JSCP)2.02 2020.

[7] Aguilar-Palacios, Caros, et al. "Forecasting Promotional Sales within the Neighbourhood." Ieee Access 7 2019.

[8] Sun, Shaolong, et al. "A clustering-based nonlinear ensemble approach for exchange rates forecasting." IEEE Transactions on Systems, Man, and Cybernetics: Systems 50.6 2018.

[9] Wang, Yi, et al. "An ensemble forecasting method for the aggregated load with subprofiles." IEEE Transactions on Smart Grid 9.4 2018.

[10] Baldan, Francisco et al. "A forecasting methodology for workload forecasting in cloud systems." IEEE Transactions on Cloud Computing 6.4 2016.

[11] Xavier, E.M., et al. "Requirements to Leverage the Electricity Distributors Sales and Revenues in the Brazilian Free Market." IEEE Latin America Transactions 14.10 2016.

[12] Behesti-Kashi, Samaneh, et al. "A survey on retail sales forecasting and prediction in fashion markets." Systems Science & Control Engineering 3.1 2015.

[13] Duan, Zhaoyang, Brittni Gutierrez, and Lizhi Wang "Forecasting plug-in electric vehicle sales and the diurnal recharging load curve." IEEE Transactions on Smart Grid 5.1 2014.

[14] Gil-Alana, Luis Alberiko, Carlos Pestana Barros, and Albert Assaf. "Retail sales: persistence in the short-term and long-term dynamics." IMA Journal of Management Mathematics 25.3 2014.

[15] Ren, Shuyun, Tsan-Minf Choi, and Na Liu. "Fashion sales forecasting with a panel data-based particle-filter model."IEEE Transactions on Systems, Man, and Cybernetics: System 45.3 2014.