JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Forecasting Big Mart Sales Using Machine Learning

[1]K. Arun Babu , [2]P. Dimdy Vignesh, [3]T. Chaitanya, [4]T.Tharun, [5]SK.Afrid

[1]Asst. Professor, Department of Computer Science, Bapatla Engineering College, Bapatla, India
[2]CSE, Department of Computer Science, Bapatla Engineering College, Bapatla, India
[3]CSE, Department of Computer Science, Bapatla Engineering College, Bapatla, India
[4]CSE, Department of Computer Science, Bapatla Engineering College, Bapatla, India
[5]CSE, Department of Computer Science, Bapatla Engineering College, Bapatla, India

Abstract : The analysis of Big Mart sales using machine learning algorithms involves studying past sales data to predict future trends. By employing ensemble models such as XGBoost Regressor ,Random Forest regressor and XGBRF Regressor. we aim to uncover patterns that can help anticipate sales fluctuations. These predictive models offer valuable insights for optimizing inventory and ultimately enhancing overall business performance in the retail sector.

IndexTerms **- XGBoost Regression , Random Forest Regression, XGBRF Regression.**

I. INTRODUCTION

The retail industry, characterized by its dynamic nature and evolving consumer preferences, constantly seeks innovative approaches to optimize operations and enhance profitability. we explore the application of machine learning algorithms to analyze historical sales data from Big Mart stores.By employing ensemble models such as XGBoost Regressor ,Random Forest regressor and XGBRF Regressor, our goal is to develop accurate forecasting models that can predict future sales trends. That is helpful, both in developing and improving marketing strategies for the marketplace.

II. RELATED WORK

Statistical Approaches: Various statistical methods have been employed for sales forecasting, including regression, Auto-Regressive Integrated Moving Average (ARIMA), and Auto-Regressive Moving Average (ARMA). These methods have been utilized to establish sales forecasting standards. However, sales forecasting is complex due to the influence of both external and internal factors.

Hybrid Models: A hybrid approach combining a mixture seasonal quantum regression method and ARIMA has been suggested for daily food sales forecasting. This hybrid model outperformed individual models.

Genetic Fuzzy Systems (GFS): E. Hadavandi incorporated GFS and data mining techniques to forecast the sales of printed circuit boards. This approach involved using K-means clustering to group data records, followed by independent handling of each cluster with a database tuning and rule-based extraction capability.

III. DATA SET

The dataset was obtained from Kaggle.com, consisting of both training and testing datasets. The training dataset contains 8000 entries, while the testing dataset contains 5000 entries. The attributes in the dataset are described as follows:

1. Item_Identifier: Unique product ID number.
2. Item_Weight: Weight of the product.
3. Item_Fat_Content: Indicates whether the item is low in fat or not.

4. Item_Visibility: Percentage of the overall viewing area assigned to the item in the shop.
5. Item_Type: Group to which the commodity belongs.
6. Item_MRP: Price of the product.
7. Outlet_Identifier: Unique slot number for the outlet.
8. Outlet_Establishment_Year: Year when the shop was first opened.
9. Outlet_Size: Total area occupied by the supermarket.
10. Outlet_Location_Type: Type of town where the store is situated.
11. Outlet_Type: Whether the shop is a supermarket or a grocery store.
12. Item_Outlet_Sales: Sales of the item in the original shop.

This dataset will be utilized for developing and testing predictive models to forecast item sales in the retail setting.

IV . METHODOLOGY

We have employed three regression algorithms: XGBoost Regression, Random Forest Regression, and XGBRF Regression. We are calculating the MAE, MSE, RMSE ,R2 Score and final concluding the best yield algorithm.

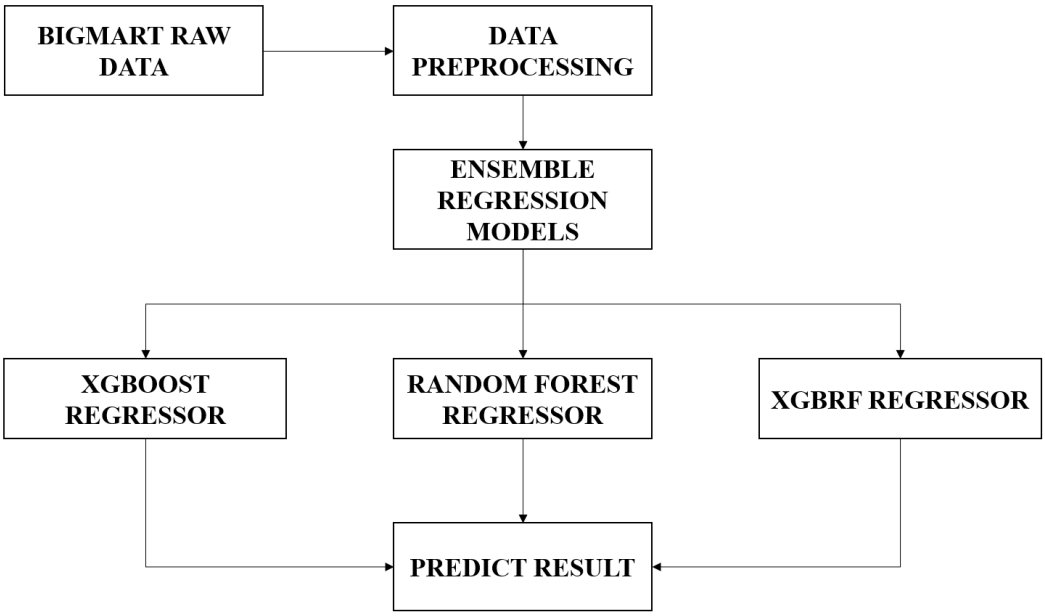

Figure1 : Shows the proposed Architecture Diagram

1. XGBoost Regressor:

XGBoost stands for Extreme Gradient Boosting, a powerful algorithm known for its efficiency and effectiveness in handling structured data.

Working Procedure:

1.1 Initialization: Start by initializing the model with default hyperparameters or customized parameters based on the dataset characteristics.
1.2 Iteration: Iteratively build an ensemble of weak learners (decision trees) by minimizing a loss function and adding trees that reduce the residual error.
1.3 Gradient Boosting: At each iteration, the model calculates the gradient of the loss function and adjusts the predictions to minimize the loss.
1.4 Regularization: Apply regularization techniques like L1 and L2 regularization to control overfitting and improve generalization.
1.5 Prediction: Finally, make predictions by aggregating the predictions of all the weak learners in the ensemble.

2. Random Forest Regressor:

Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the mean prediction of the individual trees for regression tasks.

Working Procedure:

2.1 Bootstrapping: Randomly select samples with replacement from the dataset to create multiple training datasets.
2.2 Feature Selection: At each node of the decision tree, randomly select a subset of features to split on.
2.3 Tree Construction: Build decision trees using the selected features and splitting criteria (e.g., Gini impurity for classification, MSE for regression).
2.4 Ensemble Formation: Aggregate the predictions of all the decision trees in the forest by averaging (for regression) or voting (for classification).
2.5 Prediction: Make predictions using the aggregated output of all the trees.

3. XGBRF Regressor:

XGBRF stands for Extreme Gradient Boosting Random Forest, which combines the principles of both XGBoost and Random Forest.

Working Procedure:

3.1 Initialization: Similar to XGBoost, initialize the model with default or customized hyperparameters.
3.2 Random Feature Sampling: Unlike XGBoost, XGBRF randomly samples a subset of features at each node split, similar to Random Forest.
3.3 Gradient Boosting: Build an ensemble of decision trees using gradient boosting principles, where each tree corrects the errors of the previous trees.
3.4 Regularization: Apply regularization techniques to control overfitting and improve model generalization.
3.5 Prediction: Generate predictions by aggregating the outputs of all the trees in the ensemble.

V . RESULT AND DISCUSSION

1. Xgboost Regressor :

TABLE 1 : Shows the Xgboost regressor result on the various parameter.

Parameter	Value
MSE	0.3459
MAE	0.4497
RMSE	0.5881

2. Random Forest Regressor :

TABLE 2 : Shows the Random Forest regressor result on the various parameter.

Parameter	Value
MSE	0.3213
MAE	0.4370
RMSE	0.5668

3. XGBRF Regressor :

TABLE 3 : Shows the XGBRF regressor result on the various parameter.

Parameter	Value
MSE	0.2888
MAE	0.4116
RMSE	0.5374

TABLE 4 : Comparison of MAE, MSE, RMSE with the Models.

Model	MSE	MAE	RMSE	R2_SCORE
XGBoost Regressor	0.3459	0.4497	0.5881	0.6797
Random Forest Regressor	0.3213	0.4370	0.5668	0.7025
XGBRF Regressor	0.2888	0.4116	0.5374	0.7325

VI . CONCLUSION

In this study, we utilized three regression algorithms - XGBoost Regression, Random Forest Regression, and XGBRF Regression - to forecast sales trends in the Big Mart retail stores. Our analysis revealed insightful results regarding the performance of each model in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) Score. we conclude that the XGBRF Regressor outperforms both XGBoost and Random Forest regressors in terms of forecasting Big Mart sales. It demonstrated superior predictive accuracy and generalization ability, making it the most suitable model for sales prediction in this context.

REFERENCES

[1] Halo BI. (2018). Sales forecasting: Five uses. Retrieved from https://halobi.com/blog/sales-forecasting-five-uses/
[2] Suma, V., & Hills, S. M. (2020). Data mining based prediction of demand in Indian market for refurbished electronics. Journal of Soft Computing Paradigm (JSCP), 2(02), 101-110.
[3] Wang, H. (2019). Sustainable development and management in consumer electronics using soft computation. Journal of Soft Computing Paradigm (JSCP), 1(01), 56-62.
[4] Chu, C. W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal Production Economics, 86, 217-231.
[5] Lin, Z. C., & Wu, W. J. (1999). Multiple linear regression analysis of the overlay accuracy model zone. IEEE Transactions on Semiconductor Manufacturing, 12(2), 229-237.
[6] Ajao, O. I., Abdullahi, A. A., & Raji, I. I. (2012). Polynomial regression model of making cost prediction in mixed cost analysis. International Journal on Mathematical Theory and Modeling, 2(2), 14-23.
[7] Nunnari, G., & Nunnari, V. (2017). Forecasting monthly sales retail time series: A case study. In Proceedings of IEEE Conference on Business Informatics (CBI).
[8] Suma, V., & Hills, S. M. (2020). Data mining based prediction of demand in Indian market for refurbished electronics. Journal of Soft Computing Paradigm (JSCP), 2(02), 101-110.
[9] Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In IEEE Transactions on Information Theory, 56(7), 3561.
[10] Shu, X., & Wang, P. (2015). An improved Adaboost algorithm based on uncertain functions. Proceedings of International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration.
[11] Kuo, R. J., Hu, T. L., & Chen, Z. Y. (2009). Application of radial basis function neural networks for sales forecasting. In Proceedings of International Asian Conference on Informatics in Control, Automation, and Robotics, 325-328.
[12] Majhi, R., Panda, G., & Sahoo, G. (2008). On the development of Improved Adaptive Models for Efficient Prediction of Stock Indices using Clonal-PSO (CPSO) and PSO Techniques. International Journal of Business Forecasting and Market Intelligence, 1(1), 50-67.
[13] Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics, 170, 321-335.
[14] Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. International Journal of Production Economics, 170, 97-135.
[15] Yu, X., Qi, Z., & Zhao, Y. (2013). Support Vector Regression for Newspaper/Magazine Sales Forecasting. Procedia Computer Science, 17, 1055-1062.