International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 3, No. 2, June 2024, pp. 34~41
DOI: 10.59461/ijdiic.v3i2.114

Performance Analysis of Voting Regression-Based Ensemble Learning Methods for Food Demand Forecasting

Denis R[1], Keerthana D[1]
[1]Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, Karnataka, India, 560052

Article Info

Article history:

Received April 02, 2024
Revised May 20, 2024
Accepted May 30, 2024

Keywords:

Machine learning
Demand forecasting
Regression
Ensemble technique
Voting Regression

ABSTRACT

Accurate demand forecasting has become very significant, especially in the food sector, since many products have a limited lifespan, and improper management can cause the organization to incur enormous waste and loss. This research focuses on the problem of analyzing accurate food demand and its prediction through the application of machine learning techniques. An ensemble technique such as voting regression is employed, leveraging Random Forest Regressor and Gradient Boosting Regressor, which were the top-performing models. By integrating these two techniques using voting regression, we can leverage their complementary strengths to enhance prediction accuracy. The ensemble aggregates the predictions of both models, typically by averaging, to produce a final prediction. This technique can assist in reducing overfitting and capturing complex relationships in the data, resulting in more robust and accurate forecasts of food demand. The outcomes of the R2-score, Root Mean Square Error (RMSE) and Mean Average Error (MAE) are 0.99, 0.01, and 0.00, respectively.

Corresponding Author: Denis R (e-mail: denisatshc@gmail.com)

1. INTRODUCTION

Food waste represents a significant worldwide issue with profound implications for society, the environment, and the economy. The Food and Agriculture Organization estimates that one-third of food produced for human use, totalling around 1.3 billion tons, is wasted annually. This loss exacerbates hunger, depletes natural resources, increases greenhouse gas emissions, and intensifies food security issues [1]. Addressing this critical issue, this research endeavours to elevate food waste mitigation efforts by developing an integrated solution, focusing on the enhancement of predicting food demand through machine learning techniques. We have forecasted the requirement for various food categories using regression models like the Random Forest Regressor and Gradient Boosting Regressor. Furthermore, we have combined the Random Forest Regressor and Gradient Boosting Regressor models to construct an ensemble model.

In recent times, many countries have come up with different modern and advanced technologies that forecast the demand for products with high accuracy. These tools and techniques help the modern society. However, they take a long time and are not cost-effective. Currently, machine learning algorithms are rapidly evolving to aid individuals in solving their problems more efficiently by precisely predicting food demand. The research obtained more accurate results in a much less amount of time. Machine learning is one example of how technology is developing quickly every day and demonstrating its significance in many spheres of human endeavour. The study [2] aims to uncover how machine learning algorithms ascertain the demand for food. It emphasizes the importance of three critical steps: the collection of datasets, the preparation of datasets, and the development of models. The dataset was gathered from several sources and contains 4,56,548 rows of different categories of various requirements of the product, i.e., beverages, rice bowls, starters, pasta, sandwiches, biryani, soup, salad, fish, etc.

The core objective is to diminish food waste by implementing ensemble learning techniques, such as voting regression, which synergistically combines gradient-boosting regression and random forest regression algorithms. By melding these machine learning strategies, the project seeks to deliver more precise food sales predictions, thus enabling more efficient inventory management.

1.1. Organization of the paper

The remaining sections of the paper are as follows: Section 2 addresses the literature survey, followed by Section 3 on research objectives, and Section 4 on methodology. The findings obtained, as well as the analysis and results, are discussed in section 5. section 6 ends with the conclusion.

2. LITERATURE REVIEW

This research intends to propose more precise food prediction. In this section, a few important studies on related methods are examined, along with each of their advantages and disadvantages. Proposed a novel approach for demand forecasting in a university refectory, using models for machine learning that consider the calendar effect and meal ingredients. Eighteen prediction models were developed, including Artificial Neural Networks, Gaussian Process Regression, Support Vector Regression, Regression Tree, and Ensemble Decision Tree models. The assessment metrics employed to evaluate the model performance include MSE and MAE. The Ensemble Decision tree-boosted model performed the best, with MSE, MAE, and R values of 0.51, 0.50, and 0.96, respectively. The study's contribution lies in its concentration on the meal ingredients and the diversity of models used. Additionally, prior research employing deep learning and machine learning has demonstrated encouraging outcomes. techniques for demand forecasting in the food industry[3][4].

To address the issue of uneven food production and increased food demand brought on by the world's expanding population. The project trains a learning model to forecast different food demand requirements, which can help suppliers detect food demand requirements at an early stage of production and prevent food wastage and financial losses. The model achieves an accuracy of 94.36% upon training a dataset of 4,56,548 rows and various features of different demand categories. By accurately forecasting food demand requirements, learning models have the potential to greatly enhance food demand forecasting and reduce food waste, benefiting both suppliers and the environment[2].

The study [1] focused on demand forecasting in the food industry, where precise forecasts are necessary due to the short shelf life of products. The study uses the Genpact Food Demand Forecasting dataset and compares the impact of various factors on demand. Seven regressors, including Random Forest, GBR, Light GBM, XG Boost, Cat Boost, LSTM, and Bi LSTM, are used for forecasting. The results indicate that deep learning models, particularly LSTM, perform better than other algorithms. The evaluation metrics used are RMSLE, RMSE, MAPE, and MAE, with values of 0.28, 18.83, 6.56%, and 14.18, respectively. The literature survey highlights the significance of ML and DL techniques in demand forecasting and the possibility of LSTM in improving accuracy.

Suggested [5][6] a model for food using machine learning to predict sales methods to achieve its first goal. The second goal involves comparing two datasets: one characterized by high correlation among its features and the other by low correlation. Multiple machine learning algorithms were employed in prediction in the second goal, to determine the top three algorithms that give the most accurate predictions. Conversely, when employing the second dataset, the best three algorithms are gradient boosting, random forest regressor, and decision tree. These conclusions are drawn from metrics such as RMSE and MSE.

The researcher [7] proposed an alternative method for predicting the stress intensity factor (SIF) of propagating fractures. They suggested using the gradient boosting regressor (GBR) as a substitute for the finite element technique (FEM) traditionally used for this purpose. According to the authors, the primary drawbacks of FEM-based SIF prediction include high calculation costs and significant time commitment. In their study, the authors trained the GBR using values of SIF derived via FEM, with 70% of the information utilized for training and 30% for validation. Throughout the validation procedure, the authors found a coefficient of correlation (P) of 0.977 between the SIF values produced by FEM and those predicted by GBR. This high correlation shows a high degree of agreement between the two techniques.

3. RESEARCH OBJECTIVES

To propose predicting meal orders across diverse centres by employing the dataset, preprocessing, feature engineering, model selection, and ensemble techniques to optimize predictive performance.

To implement data preprocessing (merging, encoding, dropping, scaling), feature engineering (creating 'percentage_checkout_price' and possibly PCA), model exploration (Gradient Boosting, Random Forest), and ensemble construction (Voting Regressor) to enhance predictive accuracy.

To apply the implemented solution by partitioning the dataset for training and testing sets, then training individual models as well as the ensemble based on the practice data. Lastly, the testing information is forecasted to assess performance.

To test model functionality and ensemble on unknown data (testing set). Utilize measures to assess their performance, such as RMSE, R-squared, and MAE. This stage assists in determining how effectively the models generalize to new data.

To prove the efficiency of the suggested solution by analyzing the results obtained during testing. Comparing the performances of individual models with the ensemble model. If the ensemble model outperforms individual models, it demonstrates the efficiency of ensemble learning in this context.

4. METHODOLOGY

In this research, the forecasting method employed is machine learning, a technique designed to identify inherent data patterns through iterative learning. By continually applying new information to the patterns that were learned, machine learning permits the forecasting of upcoming patterns determined by these patterns. Although there are numerous methods within machine learning, this research specifically utilizes random forest regression and gradient-boosting regression[8].

Figure 1. Steps involved

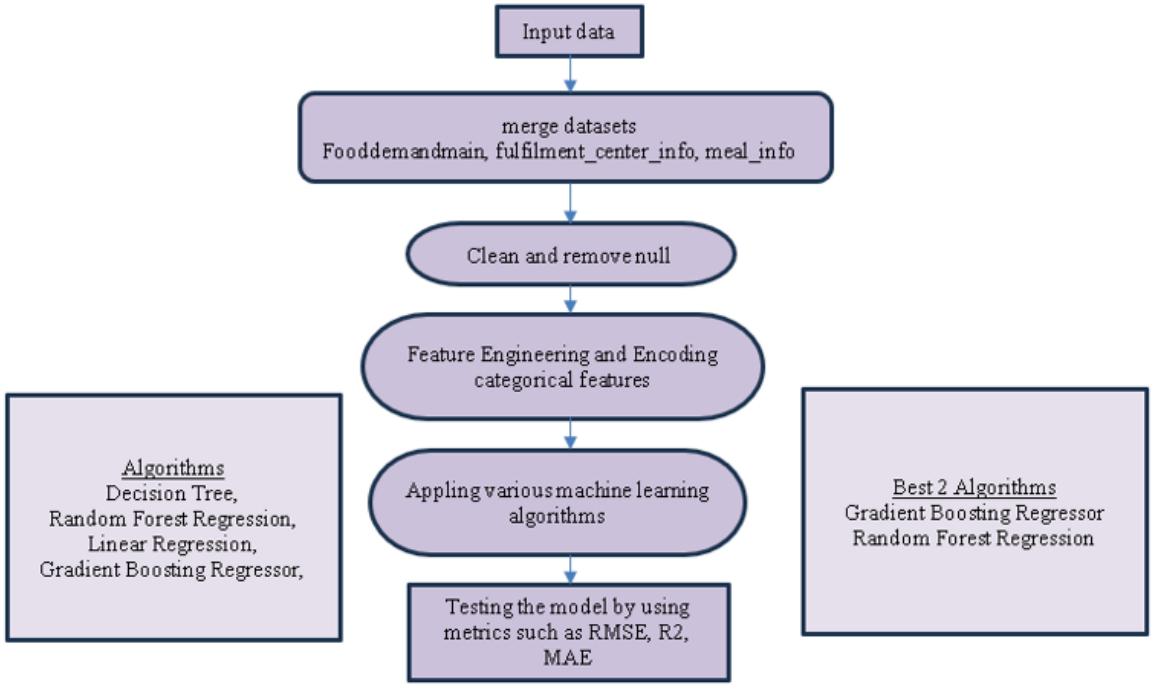

Figure 2. Architectural layout of study

Figure 1 Shows that steps to involved to build the machine learning model for the proposed problem. Figure 2 shows from the first steps to feed the input data, and cleaning the noisy in data preprocessing techniques , applying different machine models to identify the best model among them.

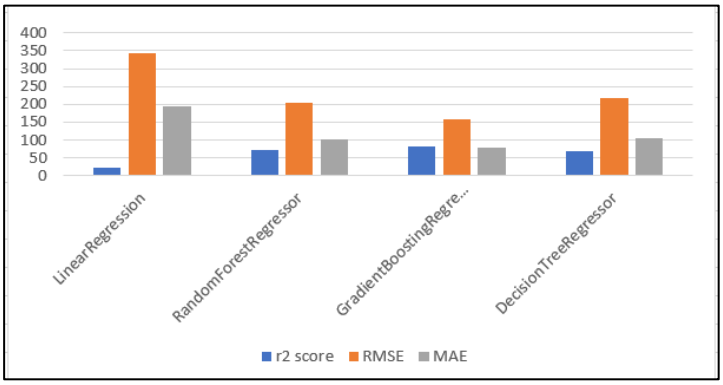

Figure 3. Comparing results of different algorithms

In this research, Figure 3, the evaluation is based on four algorithms for predictive modelling: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and Decision Tree Regressor, assessing their performance based on R2-score, RMSE, and MAE. Notably, the Gradient Boosting Regressor and Random Forest Regressor exhibit the highest R2 scores of 83 and 73, respectively, indicating superior predictive capability. Leveraging ensemble techniques like voting regression, this research aims to merge the top-performing models to enhance predictive accuracy further. By merging the strengths of these algorithms, the research anticipates achieving excellent results, warranting further exploration and validation.

4.1. Dataset description

An American professional services company called Genpact provided the "Food Demand Forecasting" dataset for machine learning[1]. The dataset consists of three separate datasheets: meal_info, which offers information about various meals; fulfilment_center, which includes details about each fulfilment centre; and food demand, which includes historical information on demand for all centres. Together, these three datasets total 4,56,548 and 15 features[1].

4.2. Dataset preprocessing

Data preprocessing assists in converting unprocessed data into a format that is usable. To change the original data into a usable form for this article, we employed exploratory data analysis, creating feature engineering and data cleaning approaches.

We use the correlation matrix to examine the correlations between the target variable; from Figure 4 'num_orders' and other properties within the collected dataset, we can identify which traits have substantial connections with the target variable and with each other by calculating the respective correlation coefficients of variables.

In essence, the correlation matrix here is an essential tool for understanding the complexities present within the collected dataset and for reaching well-informed conclusions during the modelling and data preprocessing stages.

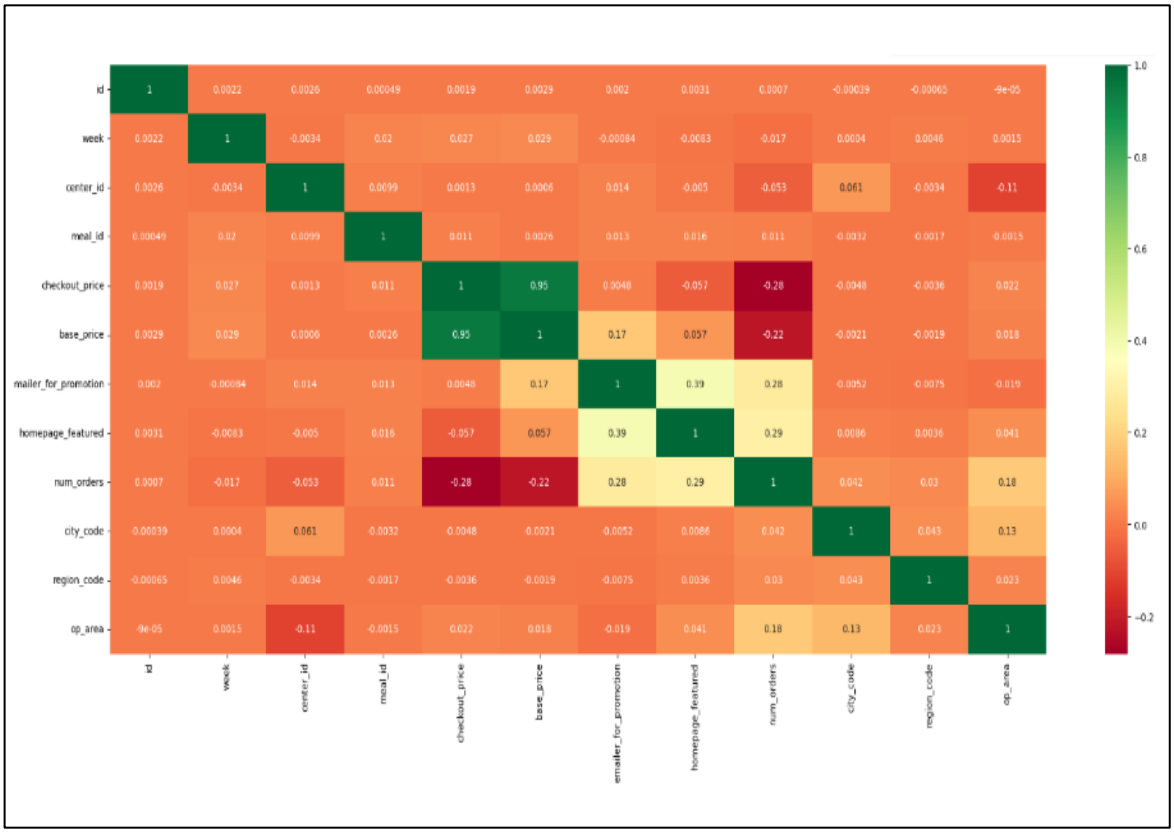

Figure 4. Heat map for the given dataset

Figure 5 shows the scatter matrix employed to analyze how several variables are related to one another. It results in a deeper comprehension of the processes and can be beneficial in making informed

decisions and identifying issues. A scatter matrix is used to analyze the connections between many variables, which improves process comprehension, aids in decision-making, and helps identify issues.

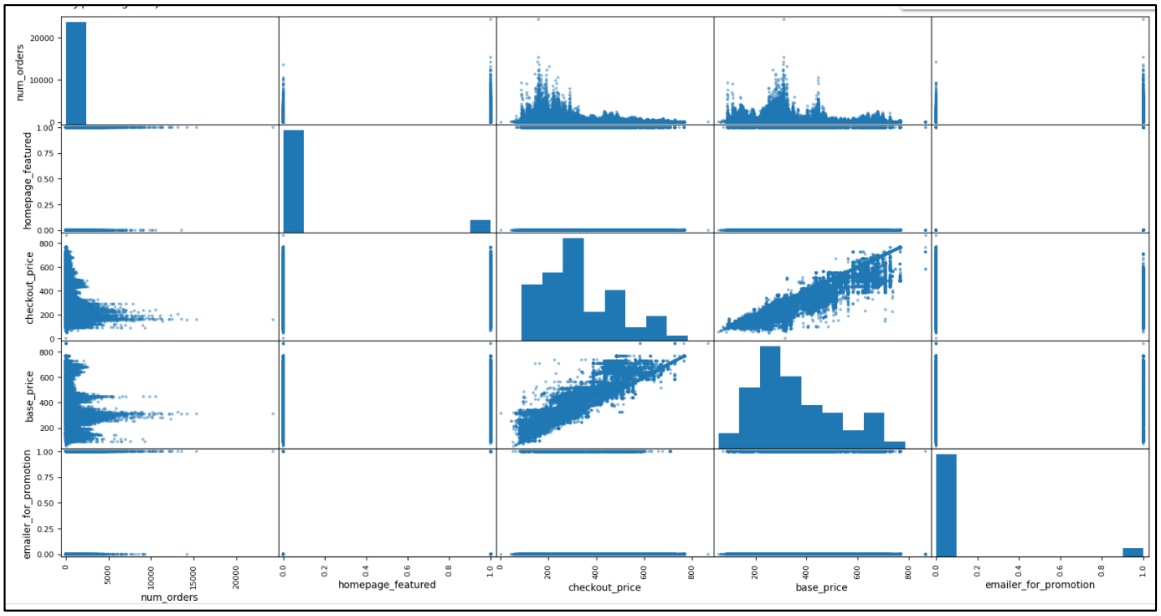

Figure 5. Scatter matrix

4.3. Data cleaning and Feature engineering

The code begins by merging three datasets based on common columns. Unnecessary columns are dropped to streamline the dataset. Categorical variables are encoded for numerical processing. Feature engineering is conducted to create a new variable capturing the percentage difference between checkout and base prices. PCA is utilized in the feature space to decrease dimensionality. Two regression models, Gradient Boosting and Random Forest, are trained and evaluated. Finally, an ensemble model combining both regressors is constructed and assessed for predictive performance. Overall, the code encompasses data consolidation, preprocessing, feature engineering, model training, and ensemble modelling, aiming to predict the number of orders effectively.

4.4. Model Building

4.4.1. Random Forest Regression

The code begins by merging three datasets based on common columns. Unnecessary columns are dropped to streamline the dataset. Categorical variables are encoded for numerical processing. Feature engineering is conducted to create a new variable capturing the percentage difference between checkout and base prices. PCA is utilized in the feature space to decrease dimensionality. Two regression models, Gradient Boosting and Random Forest, are trained and evaluated. Finally, an ensemble model combining both regressors is constructed and assessed for predictive performance. Overall, the code encompasses data consolidation, preprocessing, feature engineering, model training, and ensemble modelling, aiming to predict the number of orders effectively.

4.4.2. Gradient Boosting Regression

Gradient Boosting Regression (GBR) is an effective algorithm for ensemble machine learning that has gained significant popularity for its effectiveness in tackling regression predictive modelling tasks. GBR iteratively improves based on the forecasts produced by the previous models. This iterative process focuses on minimizing the residual errors between the values that were predicted and those of actual values. Gradient Boosting regression played a central role in this project, contributing to accurate prediction of food demand based on various features, and it is compared with other models to assess its effectiveness[9].

4.4.3. Voting Regression

In a voting ensemble Figure 6, we utilize multiple approaches instead of relying solely on one model. This methodology significantly improves system efficiency for issues involving both regression and

classification. For regression tasks, we aggregate the estimations from each model by averaging them to derive a final estimate; these ensembles are called voting regressors (VRs)[10].

Simple average: $P_{voting} = \frac{1}{x}\sum_{n=1}^{x} P_n$ (1)

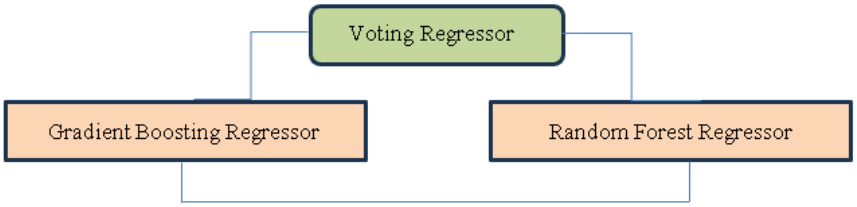

Figure 6. Voting regressor ensemble of gradient boosting and random forest

4.5. Performance prediction measures

In discussing regression accuracy, it is often noted that several metrics play key roles: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R2-Score. MAE, for instance, is seen as reflecting the magnitude, essentially showing how far off the predictions are on average. MAE, on the other hand, provides information on the average relative error and provides a sense of accuracy in comparison to the actual numbers. Since RMSE is the square root of the average squared error, it can be directly compared to the units of the predicted variable, making it a popular choice when it comes to error magnitude. In contrast, the R2-Score quantifies the proportion of the variation in the dependent variable that can be predicted from the independent variables. When combined, these metrics provide an in-depth understanding of the effectiveness of the regression model [11][12].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{1=1}^{n}\left(\begin{matrix} n \\ y_i - y_i \end{matrix}\right)^2} \quad (2)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i-1}^{n} |\, y_i - \hat{y}_i| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

5. RESULTS AND DISCUSSION

This research compares the performance of the Random Forest Regressor and Gradient Boosting Regressor in predicting food demand, with both models achieving high accuracy. The Gradient Boosting Regressor has a lower average prediction error than the Random Forest Regressor. An ensemble model combining both models using voting regression achieves an R2-score of 0.99, RMSE of 0.01, and MAE of 0.00, indicating high prediction accuracy. The improvement in food demand forecasting can contribute to more efficient and sustainable food supply chains, ultimately benefiting both businesses and the environment. The results demonstrate the potential of machine learning techniques in addressing the critical issue of inaccurate demand forecasting in the food sector.

5.1. Random Forest Regression

The RMSE value of 0.010 means an average prediction error of approximately 0.01 units. The R2 value of 0.99 indicates the Random Forest model can explain approximately 99% of the variance in actual grocery sales. The MAE value of 0.004 represents the mean absolute difference between the actual and expected food sales values.

5.2. Gradient Boosting Regression

In Figure 7, the RMSE value of the test is 0.002, indicating a lower average prediction error compared to Random Forest Regression. The R2 value of 0.99 suggests that approximately 99% of the variance in actual grocery sales can be explained by the Gradient Boosting Regression model. The MAE value of 0.001 represents a lower average absolute deviation comparing the expected and actual food sales values.

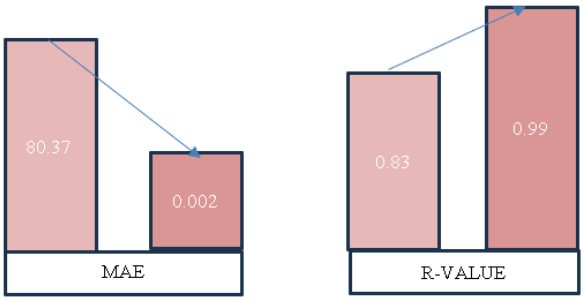

Figure 7. Comparison with the previous studies

6. CONCLUSION

In this research, we applied Voting Regression to predict food demand, combining Random Forest Regressor and Gradient Boosting Regressor. The results suggested that the ensemble model outperformed the distinct models, with RMSE, R2 score, and Mean Average Error (MAE) attaining 0.01, 0.99, and 0.00, respectively. These findings suggest that using Voting Regression to predict food demand can result in reliable and accurate forecasts; in conclusion, Machine Learning Algorithms have been shown to be effective in predicting food demand, and ensemble techniques like Voting Regression present a valuable approach to enhancing prediction accuracy. Future studies can focus on applying these techniques to different contexts and exploring the application of other ensemble methods to further improve prediction accuracy.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest in this work.

REFERENCES

[1] S. K. Panda and S. N. Mohanty, "Time Series Forecasting and Modeling of Food Demand Supply Chain Based on Regressors Analysis," *IEEE Access*, vol. 11, pp. 42679–42700, 2023, doi: 10.1109/ACCESS.2023.3266275.

[2] V. Kumar and V. V. K. Kumar, "Food Demand Prediction using Deep Learning," *IRE Journals*, vol. 6, no. 10, pp. 1–5, 2023.

[3] M. Aci and D. Yergök, "Demand Forecasting for Food Production Using Machine Learning Algorithms: A Case Study of University Refectory," *Teh. Vjesn.*, vol. 30, no. 6, pp. 1683–1691, Dec. 2023, doi: 10.17559/TV-20230117000232.

[4] H. . P. and K. . S. Kruthika V, Lavanya H.R, Mahalakshmi E.H, Ranju P.S.R, "Integrated Approach for Food Donation System, Restaurant Food Demanding forecasting using Machine Learning, and Global Food Waste Analysis," *Int. Res. J. Mod. Eng. Technol. Sci.*, Jul. 2023, doi: 10.56726/irjmets42802.

[5] H. M. Merdas and A. H. Mousa, "Food sales prediction model using machine learning techniques," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 6, p. 6578, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6578-6585.

[6] A. Keprate and R. M. C. Ratnayake, "Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, Dec. 2017, pp. 1331–1336. doi: 10.1109/IEEM.2017.8290109.

[7] T. Tanizaki, T. Hoshino, T. Shimmura, and T. Takenaka, "Restaurants store management based on demand forecasting," *Procedia CIRP*, vol. 88, pp. 580–583, 2020, doi: 10.1016/j.procir.2020.05.101.

[8] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments," *Energies*, vol. 14, no. 16, p. 5196, Aug. 2021, doi: 10.3390/en14165196.

[9] B. Erdebilli and B. Devrim-İçtenbaş, "Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey," *Mathematics*, vol. 10, no. 14, p. 2466, Jul. 2022, doi: 10.3390/math10142466.

[10] A. Arjomandi-Nezhad, A. Ahmadi, S. Taheri, M. Fotuhi-Firuzabad, M. Moeini-Aghtaie, and M. Lehtonen, "Pandemic-Aware Day-Ahead Demand Forecasting Using Ensemble Learning," *IEEE Access*, vol. 10, pp. 7098–7106, 2022, doi: 10.1109/ACCESS.2022.3142351.

[11] K. Posch, C. Truden, P. Hungerländer, and J. Pilz, "A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants," *Int. J. Forecast.*, vol. 38, no. 1, pp. 321–338, Jan. 2022, doi: 10.1016/j.ijforecast.2021.06.001.

[12] A. Keprate and R. M. Chandima Ratnayake, "Remaining Fatigue Life Prediction of Topside Piping Using Response Surface Models," in *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, Dec. 2018, pp. 237–241. doi: 10.1109/IEEM.2018.8607831.

BIOGRAPHIES OF AUTHORS

Dr. Denis, MCA., M.Phil., Ph.D., works as an Assistant Professor at Mount Carmel College. Received a Bachelor of Science (B.Sc.) degree from Loyola College (Autonomous), Madras University, Chennai, TN, India, and Master of Computer Applications (MCA) from Sacred Heart College (Autonomous), Tirupattur in the years 2006 and 2009, respectively. He obtained a PhD in Computer Science at Periyar University, Salem, TN, India. He can be contacted at email: denisatshc@gmail.com

Keerthana D has Obtained a Bachelor of Science (B.Sc.) degree from St Anne's first-grade College for Women, miller's Road, Bengaluru, India. Currently, She is pursuing a Master of Science (M.Sc.) degree from Mount Carmel College (Autonomous). Keerthana has demonstrated an unquenchable thirst for knowledge, an unwavering dedication to thorough research, and a profound desire to make a significant impact in both the academic and social domains. She can be contacted at email: keerthanadbnkp@gmail.com