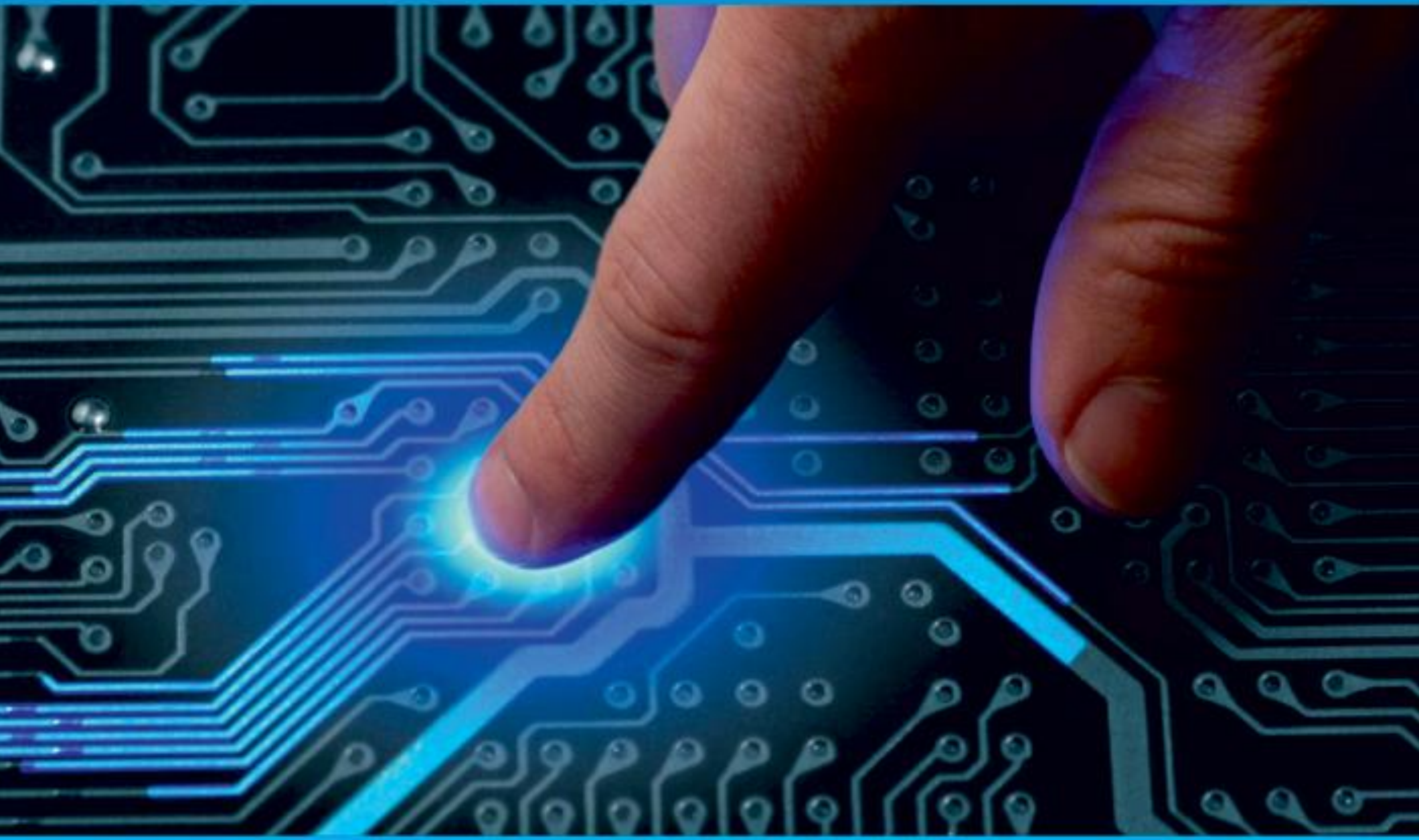
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

9940 572 462 6381 907 438 ijircce@gmail.com www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107039 |

Big Mart Sales Prediction Using Random Forest Algorithm

Soundarya H N[1], Dr.Raghavendra S P[2]

PG Student, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India[1]

Assistant Professor, Dept. of Master of Computer Applications, Jawaharlal Nehru New College of Engineering, Shivamogga, India[2]

ABSTRACT: In recent years, the number of supermarkets and their franchises has greatly increased. They currently need to forecast item sales if they want to improve their sales. By doing this, they can make money and protect themselves from losses. So, it will take a lot of time and effort to complete this examination. So, in order to anticipate the sales of the items, we created a machine learning model that will employ the random forest algorithm. By doing this, retailers may plan their hiring strategy, anticipate problems before they arise, inspire their sales staff, forecast sales, support future marketing initiatives, and in many other ways.

KEYWORDS: Machine learning, sales prediction, random forest algorithm

I.INTRODUCTION

Any retail organization must predict sales since it is essential for setting sales goals, optimizing inventory control, and making wise business decisions. The Big Mart dataset offers a great chance to investigate the use of the Random Forest algorithm for sales prediction in this situation.

A potent machine learning tool that makes use of the idea of ensemble learning is the Random Forest algorithm. To build a strong predictive model, it mixes various decision trees. A random subset of the training data and a random subset of the input features are used to build each decision tree in the forest. This randomization aids in decreasing overfitting and improving the model's capacity for generalization.

The Big Mart dataset should contain historical data regarding a variety of parameters, such as item visibility, item weight, item kind, shop location, promotion, etc., in order to forecast sales using the Random Forest algorithm. The goal variable that we want to predict is the sales column.

II.RELATED WORK

Aaditi Narkhede, Mitali Awari, Prof. Amrapal Mhaisgawali, and Suvarna Gawali. et al[1]Using machine learning techniques, "BigMart Sales Prediction" Vinayaga Sundharam R and Naveenraj R. This offers information on the precision and efficacy of the used machine learning models, as well as potential repercussions and future research possibilities.

Bohdan M. Pavlyshenko et al. [2] put forward the perspective of machine generalization. This comes into play mostly when fewer data is available in the system, perhaps with the introduction of new products or new outlets. The special technique of stacking was implemented to build the regression.

Inedi. Theresa, Dr.Venkata Reddy Medikonda, K.V. Narasimha Reddy et al.[3] discusses sales prediction by using the methodology of Exploratory Machine Learning. They carried out the whole process by figuring out proper steps that included a collection of data, thesis generation to efficiently understand bugs, further cleaning and processing the data.

Kadam, H., Shevade, R., Ketkar, P. and Rajguru et al.[4] proposed a model that works effectively with multiple linear regression and a random forest algorithm. This model was utilised to forecast big mart sales prediction and with that, a certain data set.

Kumari Punam, Rajendra Pamula and Praphula Kumar Jain in et al.[5] A Two-Level Statistical Model for Big Mart Sales Prediction have devised a two-level approach to predict sales of products that promise to yield better efficiency. It involves stacking up of algorithms wherein the top layer consists of just one learning algorithm and the bottom layer has one or more algorithms placed. This methodology of two-level modeling outperforms the single model predictive technique and results in better predictions of sales.

Ranjitha P and Spandana M et al. [6] Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms have implemented Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting sales of big mart.

Rohit Sav,Pratiksha Shinde, and Saurabh Gaikwad et al. [7] .The authors discuss the implications of their findings and the potential impact of machine learning-based sales prediction in Big Mart.

Venkata Arun Kumar Dasari, B. Ramasubbaiah, Ayesha Syed, Asha Jyothi Kalluri,Venkateswara Reddy Pocha (2020, FEB) et al.[8]. The phrase "BIGMART SALES USING MACHINE LEARNING WITH DATA ANALYSIS" JES, Volume 11, Issue 2. The paper describes the study's methodology, data collection, preprocessing, and machine learning algorithms for sales prediction, focusing on data analysis techniques.

III.PROBLEM STATEMENT

Due to the rapid growth of international malls and online shopping, competition between different shopping centres and large marts is becoming more heated and violent every day. Every store wants to know what their customers desire in advance in today's digitally connected world so that they don't run out of sales items during a specific season. For every retailing organization, like Big Mart or Mall, anticipating sales before actual sales plays a crucial role in maintaining a profitable business. This work's manual material handling may lead to serious errors, bad organization management, and, more importantly, it would be time-consuming, resource-intensive, and less accurate—none of which are desirable in the fast-paced climate of today. The resulting data can be utilized by retailers like Big Mart to anticipate future sales volume using machine learning techniques like regression models to ascertain the likelihood of sales.

IV.DESIGN AND IMPLEMENTATION

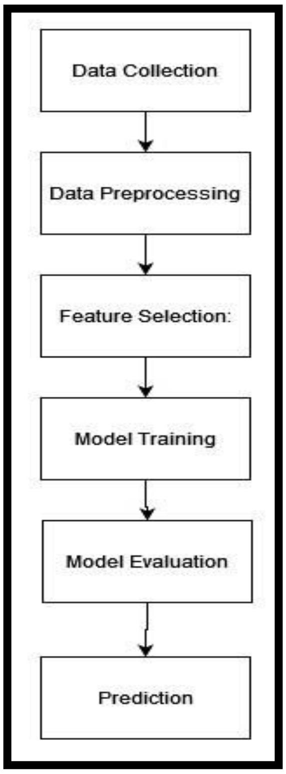


Figure1: Flow chart of the system

Data Collection: Collect historical sales data and other relevant factors from various sources. This may involve obtaining data from government agencies, commodity exchanges, and private data providers.

Data preprocessing: Examine and organize the dataset first. If necessary, handle outliers, missing values, and feature engineering. In this stage, the dataset is checked to make sure it is suitable for the Random Forest model's training.

Feature selection: Determine relevant features that have a big impact on sales while choosing features. This can be accomplished through statistical analysis, correlation, or subject-matter knowledge. Only including features that are useful improves model performance and lowers computational complexity.

Training the Random Forest Model: Apply the training dataset to the Random Forest model. Between the input features and the sales target, the model will discover patterns and relationships. The number of trees, the maximum tree depth, and the number of features taken into account at each split are examples of hyperparameters for Random Forests. The performance of the model can be enhanced by tuning these hyperparameters.

Model evaluation: Evaluate the trained model using the test dataset. Common evaluation metrics for regression tasks include mean squared error (MSE), root mean squared error, and R-squared value. These metrics help assess how well the model predicts sales values. The accuracy of the model's sales value predictions can be evaluated using these criteria.

Sales forecasting: After the model has been trained and assessed, it may be used to forecast sales based on fresh, unused data. The trained Random Forest model will produce the anticipated sales numbers when given the pertinent input features.

V.RESULT ANALYSIS

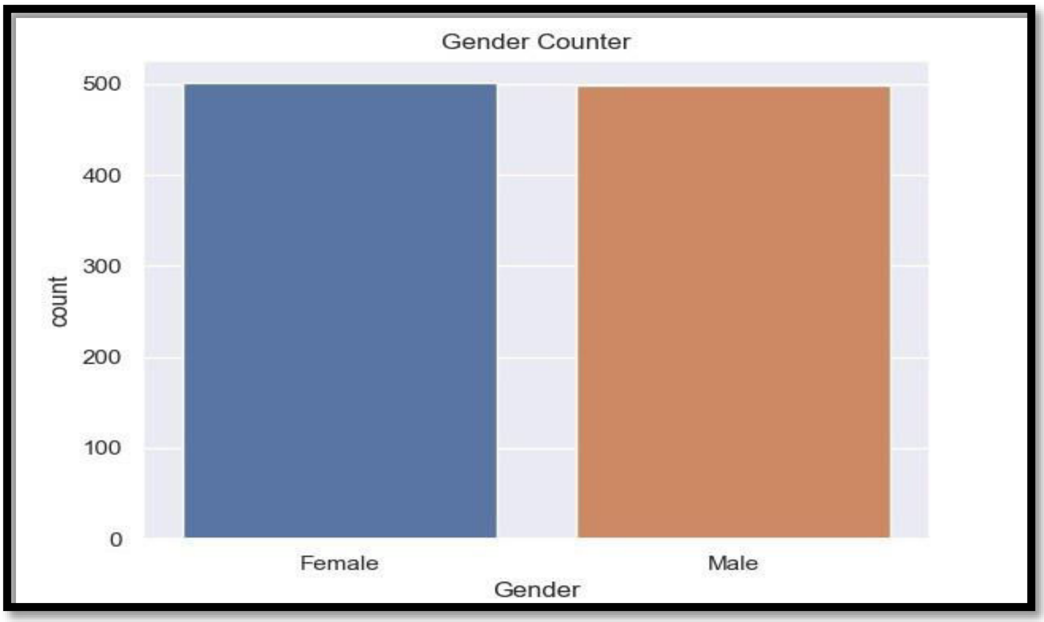


Figure 2: A visual representation of males and females in the dataset.

The code creates a graphical representation of the gender distribution in the dataset, with two bars representing males and females. The height of each bar indicates the number of occurrences of each gender category. This count plot helps understand customer demographics, segment target audience, and understand gender-specific trends. It is clear, concise, and easy to interpret, enabling quick visual comparisons of males and females, highlighting imbalances or significant differences in gender representation. Overall, the code is a valuable tool for understanding and supporting gender-related analysis and decision-making processes.

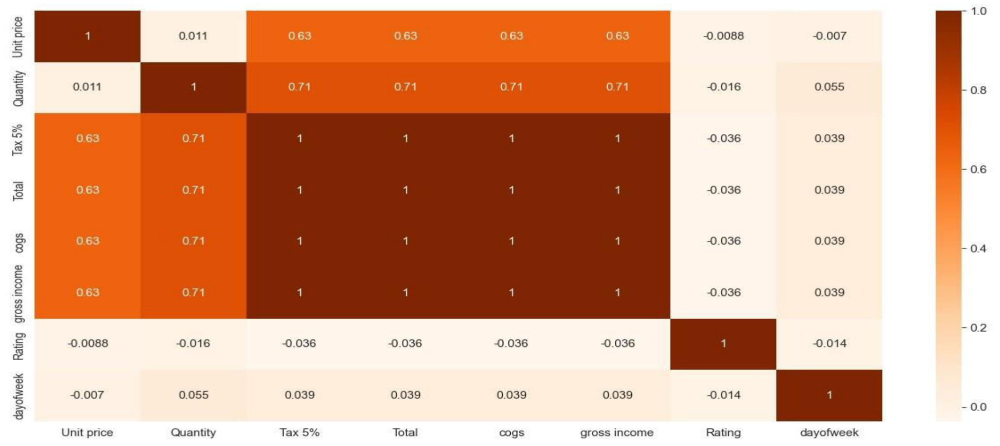


Figure 3: The correlation between different variables in the dataset.

The code generates a heatmap that visually represents the correlation between different variables in the dataset. Each cell corresponds to the correlation between two variables, with the color representing the strength and direction of the correlation. The annotation within each cell displays the correlation coefficient, indicating the numerical value of the correlation. This heatmap helps in understanding relationships and dependencies between variables, identifying patterns, associations, and potential multicollinearity. This information is valuable for feature selection, identifying influential variables, and understanding the impact of one variable on another.

The heatmap provides a comprehensive and intuitive exploration of correlations, making it easier to identify significant relationships and make data-driven decisions. It is particularly useful in identifying strong correlations, which provide insights into variables that may have a high impact on the target variable or each other. Overall, the heatmap supports data exploration, feature selection, and decision-making processes by providing a concise and visually appealing summary of the correlation structure within the dataset. This project gives 97% of accuracy on predicting the total requirements.

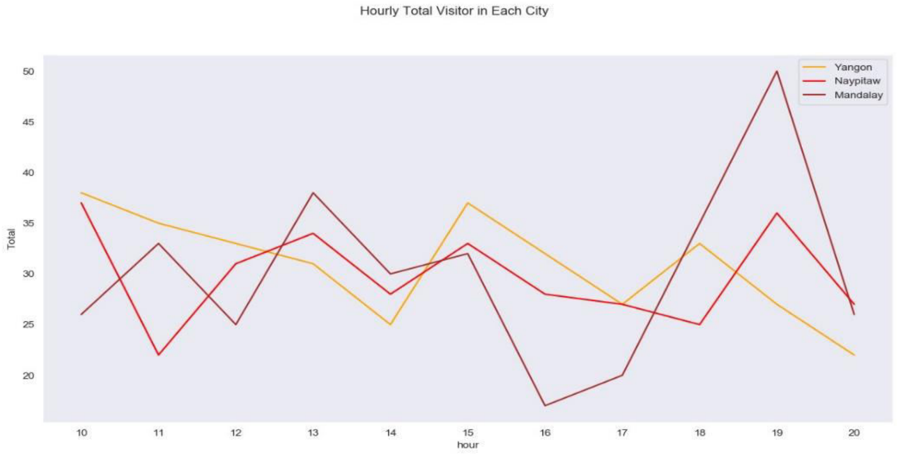


Figure 4: The line plot that visualizes the hourly total visitors in each city.

The code generates a line plot to display hourly total visitors in each city, with each city having a separate line representing the count of visitors for each hour. This helps in understanding visitor variations across different hours of the day, enabling quick comparisons of visitor patterns and identifying peak and off-peak hours. This information is valuable for staffing, resource allocation, and identifying busiest times in each city. The output of the code enables businesses to optimize their services based on varying demand throughout the day, facilitating the analysis of visitor patterns and making informed decisions related to operations and customer service.

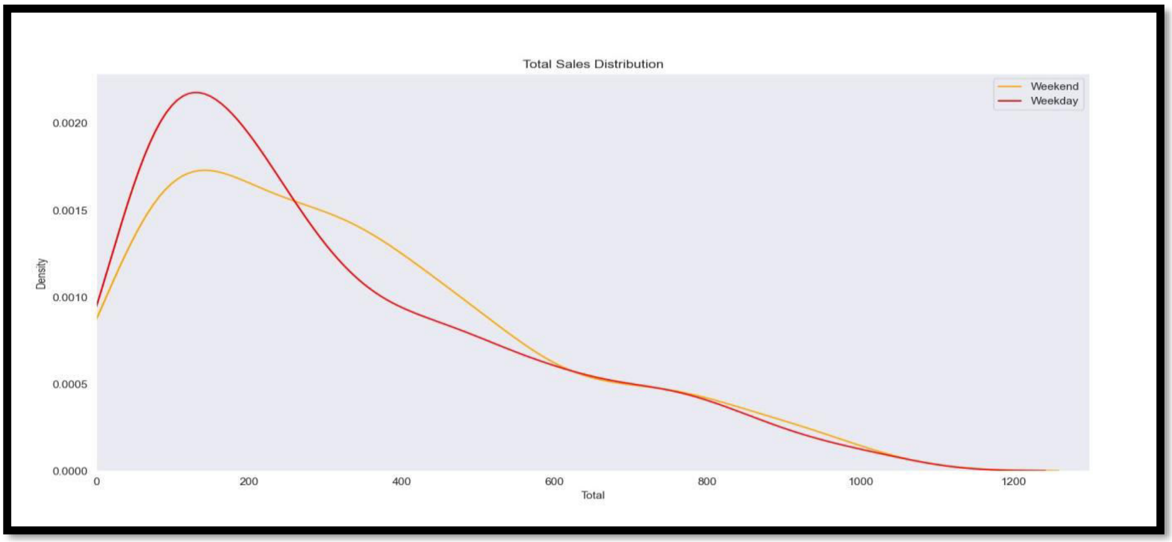
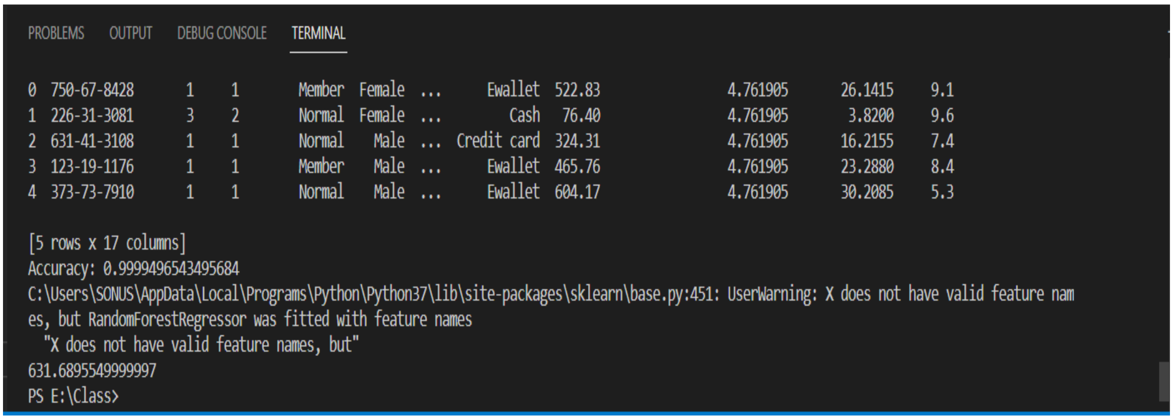


Figure 5: The distribution of total sales on weekdays and weekends.

The code's result will be a distribution plot that shows how the total sales are distributed between weekdays and weekends. Two lines will be displayed on the plot, one reflecting how the total sales are distributed on weekends (in orange) and the other how they are distributed during the workweek (in red). Understanding the fluctuation in total sales between weekdays and weekends is made easier by the distribution graphic. Insights into the sales trends on various days of the week may be gained by comparing the distributions' form, spread, and central tendency. In conclusion, the code's output is a distribution plot that shows how the total sales are distributed across weekdays and weekends. This plot makes it easier to analyze sales trends and assists in identifying.

RESULT:



```
0  750-67-8428      1     1       Member  Female  ...      Ewallet  522.83              4.761905      26.1415     9.1
1  226-31-3081      3     2       Normal  Female  ...         Cash   76.40              4.761905       3.8200     9.6
2  631-41-3108      1     1       Normal    Male  ...  Credit card  324.31              4.761905      16.2155     7.4
3  123-19-1176      1     1       Member    Male  ...      Ewallet  465.76              4.761905      23.2880     8.4
4  373-73-7910      1     1       Normal    Male  ...      Ewallet  604.17              4.761905      30.2085     5.3

[5 rows x 17 columns]
Accuracy: 0.9999496543495684
C:\Users\SONUS\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\base.py:451: UserWarning: X does not have valid feature nam
es, but RandomForestRegressor was fitted with feature names
  "X does not have valid feature names, but"
631.6895549999997
PS E:\Class>
```

Figure 6: snapshot of result

The above snapshot represents the result of the sales prediction

VI.CONCLUSION AND FUTURE WORK

In conclusion, the application of the Random Forest algorithm for Big Mart sales prediction has yielded promising results. Through the use of this ensemble learning technique, we have been able to effectively model and forecast sales patterns with a high degree of accuracy. The algorithm's ability to handle both numerical and categorical data, as well as its robustness to overfitting, has proven to be advantageous in extracting meaningful insights from the dataset. However, there is still room for improvement and future work. Further research could focus on exploring feature engineering techniques to enhance model performance, incorporating additional external data sources to capture more dynamic market trends, and experimenting with hyperparameter tuning to optimize the algorithm's performance. Additionally, exploring other advanced machine learning algorithms and comparing their predictive capabilities with the Random Forest model could also be a valuable avenue for future investigation, ultimately contributing to more accurate and reliable sales predictions for Big Mart.

REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A comparative study of linear and nonlinear models for aggregate retails sales forecasting”, Int. Journal Production Economics, vol. 86, pp. 217- 231, 2003.

[2] Gopal Behera1 and Neeta Nain2 1,2Department Of Computer Science and Engineering 1,2Malaviya National Institute of Technology Jaipur, India 12019rcp9002@mnit.acn.

[3] Heramb Kadam, Rahul Shevade, Prof. Deven Ketkar , Mr. Sufiyan Rajguru (2018). A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. (IJEDR).

[4] Inedi. Theresa, Dr. Venkata Reddy Medikonda,K.V.Narasimha Reddy. (2020, March). Prediction of Big Mart Sales using Exploratory Machine Learning Techniques 020 International Journal of Advanced Science and Technology (IJAST).

[5] Kumari Punam, Rajendra Pamula, Praphula Kumar Jain (2018, September 28-29). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018 International conference on on Computing, Power and Communication Technologies.

[6] Linda Camilla Boldt, Vinothan Vinayagamoorthy, Florian Winder, Melanie Schnittger, Mats Ekran, Raghava Rao Mukkamala, Niels Buus Lassen, Benjamin Flesch, Abid Hussain1 and Ravi Vatrapu1,2 1Centre for Business Data Analytics, Copenhagen Oslo Business School, Denmark Westerdal, Denmark Comm & Tech, Norway.

[7] Ranjitha P, Spandana M. (2021). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms.Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021).

[8] 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, no. 02 (2020): 101- 110.

[9] Xiaodan Yua,b, Zhiquan Qib ,Yuanmeng Zhaoc.” support Vector Regression for Newspaper/Magazine Sales Forecasting” Published by Elsevier B.V. 2013 International Conference on Information Technology and Quantitative Management Open access under CC BY-NC-N. DOI: 10.1016/j.procs.2013.05.134.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING