MACHINE LEARNING- (22AIE213)

Topic:

Board Game Review Prediction

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

CSE(AI)

Centre for Computational Engineering and Networking

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112 (INDIA)

JUNE– 2024

A Project

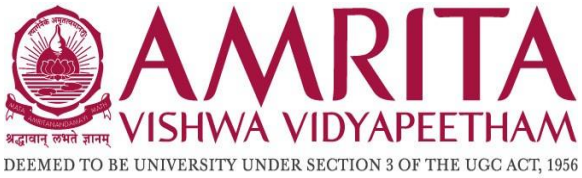
Submitted by

M.C. DHNAUSH -CB.EN.U4AIE22130

AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112



BONAFIDE CERTIFICATE

This is to certify that the Project submitted by M.C. DHNAUSH - CB.EN.U4AIE22130 for the award of the Degree of Bachelor of Technology in the "CSE(AI) " is a bonafide record of the work carried out by her under our guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

Dr.Abhishek

Project Guid

Dr. K.P.SOMAN

Professor and Dean AI

Submitted for the University examination held on 11-06-2024

TABLE OF CONTENTS

DECLARATION

Myself M.C.DHNAUSH - CB.EN.U4AIE22130 hereby declare that this project is the record of the original work done by me under the guidance of Dr.Abhishek, Professor, Centre for Computational Engineering and Networking, Amrita School of Artificial Intelligence, Coimbatore. To the best of my knowledge this work has not formed the basis for the award of any degree/diploma/ associate ship/fellowship/or a similar award to any candidate in any University.

Place: Coimbatore

Date: 11-05-2024 **Signature of the Students**

Acknowledgement

I would like to express our special thanks of gratitude to our teacher (Dr.Abhishek), who gave me the golden opportunity to do this wonderful project, which also helped me in doing a lot of exploration and i came to know about so many new things. I am thankful for the opportunity given.

PART-1
DIGIT CLASSIFICATION USING ADAM OPTIMIZER

1 INTRODUCTION

BoardGameGeek is a very popular site where different types of board games are discussed and reviewed.

In this project, we have a dataset containing 80,000 board games and their corresponding review scores.

These data was scraped from BoardGameGeek. The data contains rows and columns, Each row represnets a single board game and has statiistics about the board game as well as review information. Some of the columns are:

Name: The name of the board game

Playingtime: the playing time (given by the manufacturer).

minplaytime: the minimum playing time (given by the manufacturer).

maxplaytime: the maximum playing time (given by the manufacturer).

minage: the minimum recommended age to play.

users_rated: the number of users who rated the game.

average_rating: the average rating given to the game by users. (0-10)

total_weights:Number of weights given by users. Weight is a subjective measure that is made up by BoardGameGeek. It describes how "deep" or involved a game is.

average_weight: the average of all the subjective weights (0-5).

Size: The dataset contains a significant number of entries with several features describing each board game

2 AIM

In machine learning project the AIM is about predicting the average rating of board games.

3 DATA PREPROCESSING

Handling Missing Values: Missing values in the dataset were handled by [specific method, e.g., removing rows with missing values, filling with mean/median].

Encoding Categorical Variables: Categorical variables such as game name were encoded using [specific technique, e.g., one-hot encoding].

Feature Selection: Features were selected based on their relevance and correlation with the target variable, 'average_rating'. This involved dropping irrelevant features and retaining the most impactful ones.

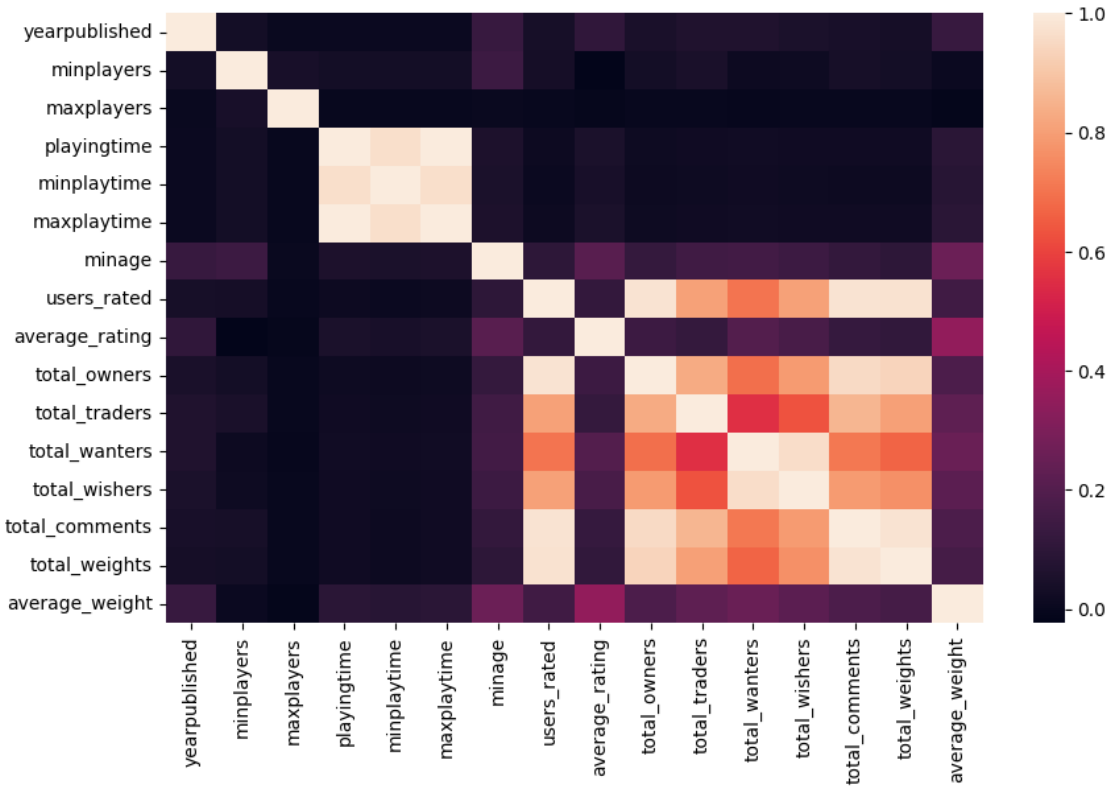
Data Splitting: The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance accurately.

3.1 PREPROCESSING STEPS INVOLVED

1. The columns [‘id’, ‘type’, ‘name’, ‘bayes_average_rating’] are to be dropped.
2. The rows with missing values are to be dropped.
3. The rows with ‘users_rated’ = 0 are to be dropped.
4. Swapping is to be done for rows with ‘minplayers’ > ‘maxplayers’ and ‘minplaytime’ > ‘maxplaytime’

4 VISUALIZING THE DATA

Check the correlation between variables. Correlation is any statistical relationship (causal or not) between 2 random variables though it commonly refers to the degree to which a pair of variables are linearly related.

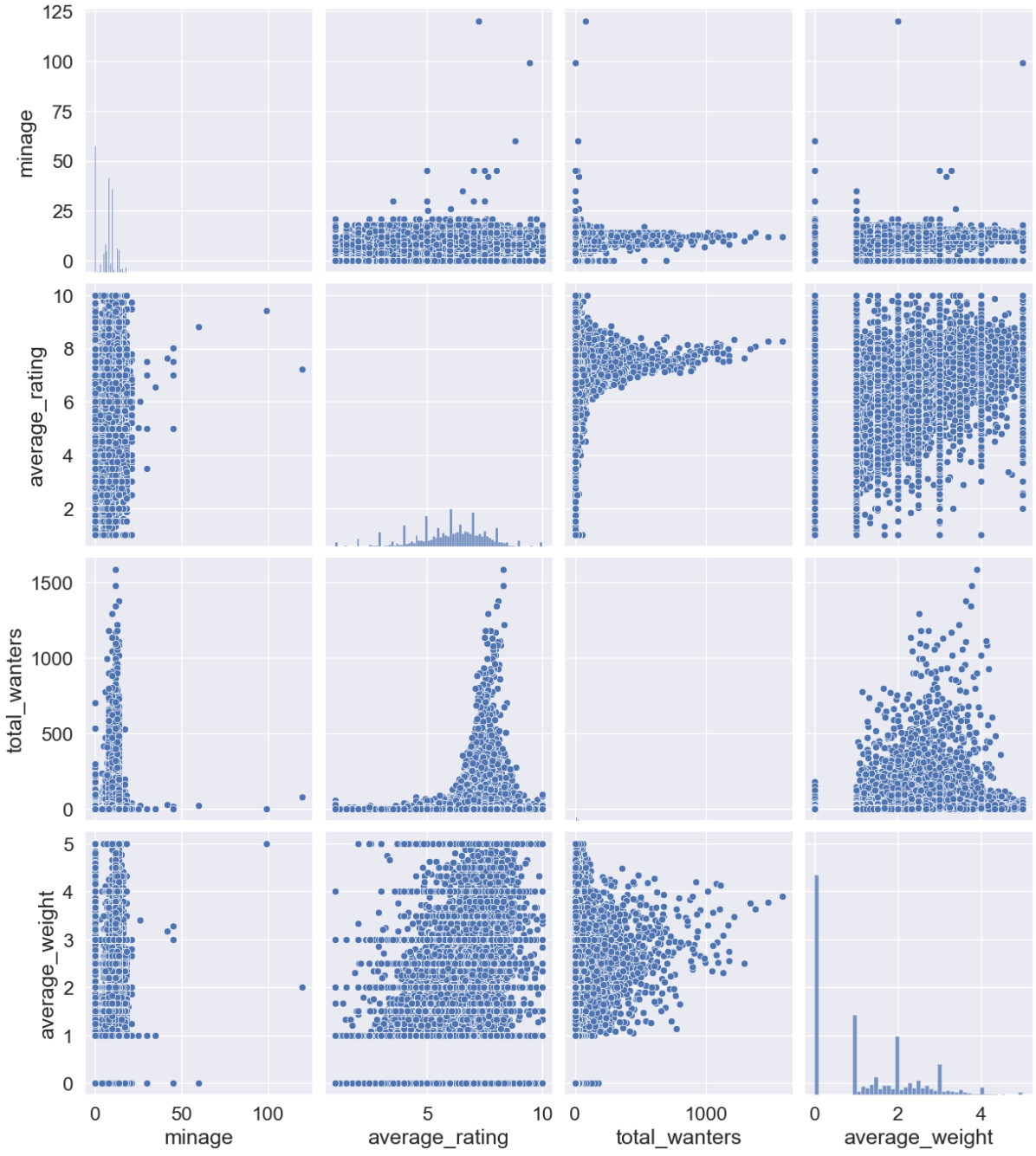


The light shaded areas are highly correlated. It can be observed that 'average_rating' has relatively high correlation with 'minage', 'total_wanters' and 'average_weight'.

'total_owners', 'total_traders', 'total_wanters', 'total_wishers', 'total_comments' and 'total_weights' have good correlation among themselves which is expected as each of these variables are directly proportional to demand value of a board game.

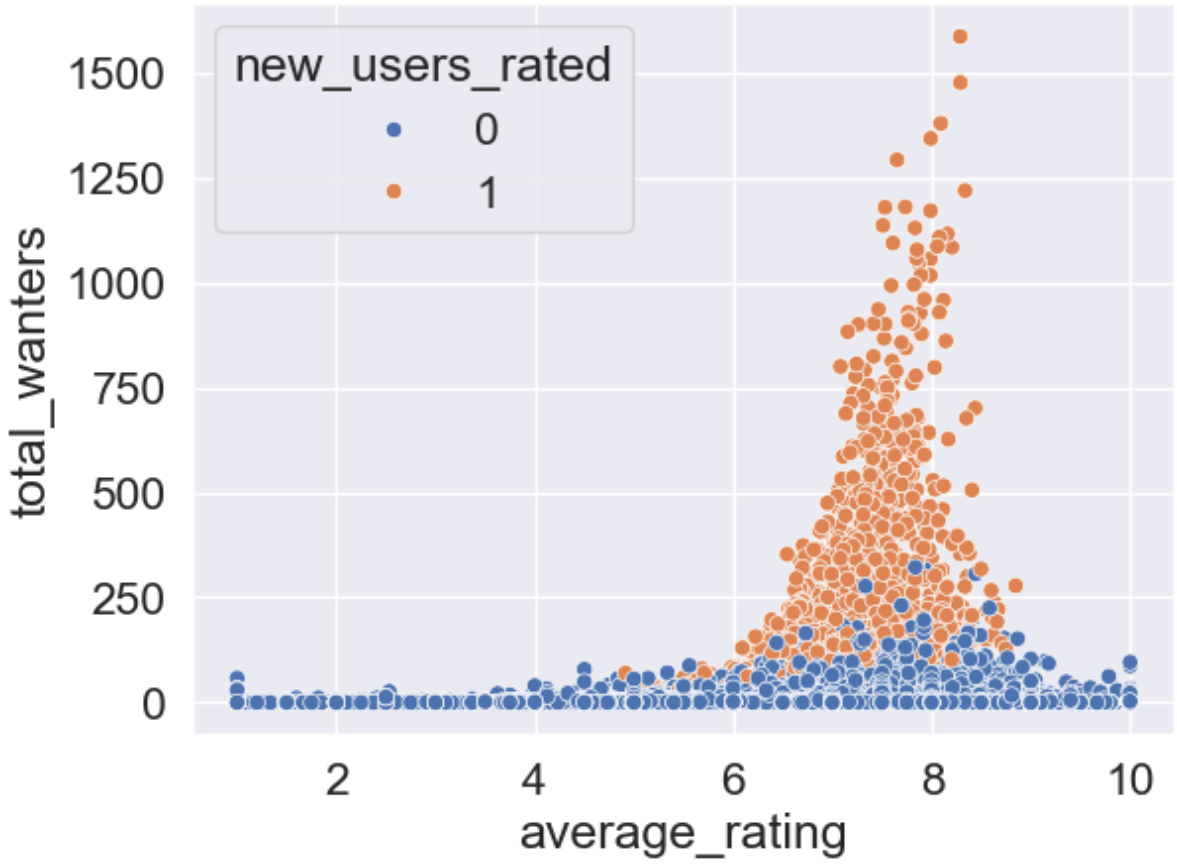
'playtime', 'minplaytime' and 'maxplaytime' have good correlation among themselves which is expected as each of these variables are related to playing time of a board game.Since 'average_rating' has relatively high correlation

with ‘minage’, ‘total_wanters’ and ‘average_weight’, we would focus on these variables.



In ‘total_wanters’ vs ‘minage’ graph, it can be observed that ‘total_wanters’ is high for board games with ‘minage’ between 10 and 20. This implies many people prefer board games designed for teens.

A new column called ‘new_users_rated’ is defined. ‘new_users_rated’ = 1 if ‘users_rated’ > df[‘users_rated’].mean() and 0 otherwise.

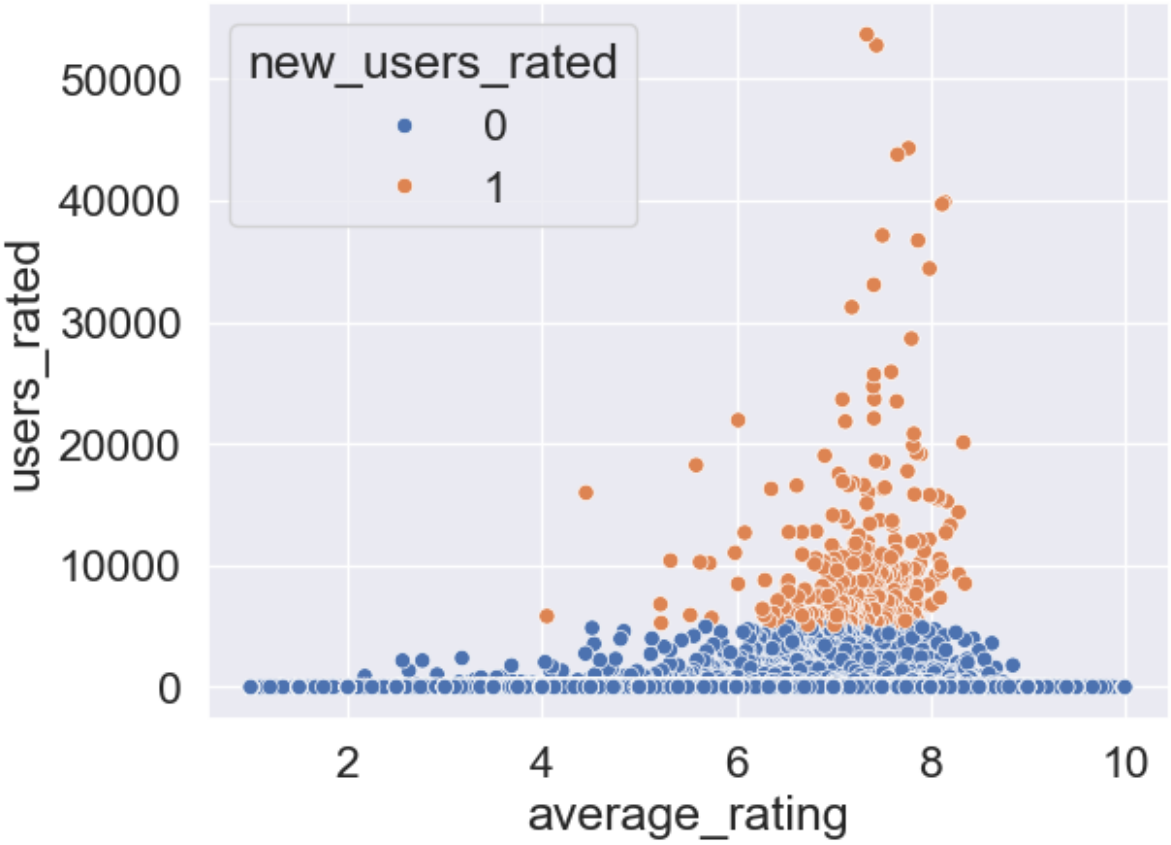


‘users_rated’ is high for board games with ‘average_rating’ between 6 and 9. ‘total_wanters’ is high for board games with ‘average_rating’ around 8 which is sensible as people would desire to have board games which have high ratings.

It is not right to compare board games based on ratings alone.

The ‘users_rated’ must also be considered before making a decision. It may be that board games with high ‘average_rating’ had less number of ‘users_rated’ which makes the ‘average_rating’ biased.

A new column called ‘new_users_rated’ is defined. ‘new_users_rated’ = 1 if ‘users_rated’ > 5000 and 0 otherwise



The preferred board game would be the one with highest ‘average_rating’ in the group of orange points(‘users_rated’ > 5000).

5 TRAINIGN THE MODEL

The data set has been preprocessed and is ready to be trained. A comparison is made between 2 models :-

1) Linear Regression
2) Random Forest Regression.

6 COMPARISION OF REGRESSION ALGORITMS

The comparison is made based on the cross validation score. The given training set is divided into 2 sets :-

1) 'Train_set'
2) 'Test_set'.

The model is trained using a portion of 'Train_set'.

Cross validation score is calculated based on performance of trained model in remaining portion of 'Train_set'.

There are various cross validation techniques which will be discussed later. Here K-Fold cross validation technique is used.

The training set is split into 'Train_set' and 'Test_set'.

7 MODEL SELECTION

Linear Regression: Simple model assuming a linear relationship. Random Forest Regressor: An ensemble method that uses multiple decision trees.

Random Forest Regressor: Parameters such as number of trees, maximum depth, etc., were tuned for optimal performance.The model was trained on the training dataset using these parameters.

8 MODEL EVALUATION

The Mean Squared Error (MSE) for Linear Regression was 2.03 The Mean Squared Error (MSE) for Random Forest Regressor was 1.47, indicating better performance.

As expected the performance of Random Forest Regressor model is better than Linear Regression model for the given data set. This is because the board game data set is very large and it is difficult for the Linear Regression model to fit the data by a straight line.

9 RESULT

Predictions: The predicted average ratings for two selected board games were [7.7380353, 7.6425094]. The actual average ratings for these games were [7.74070, 7.65777]. This close match shows the effectiveness of the Random Forest Regressor.

10 CONCLUSION

The Random Forest Regressor outperformed the Linear Regression in predicting board game ratings.

Accurate predictions can help in understanding the factors that make a board game popular and successful.

11 FUTURE WORK

Incorporate additional features such as user reviews and game categories.

Explore other machine learning models like Gradient Boosting or Neural Networks.

Perform hyperparameter tuning to further improve model accuracy.

12 REFERANCE

How to drop empty rows from a Pandas dataframe in Python (adamsmith.haus)

How To Filter Pandas Dataframe By Values of Column? - Python and R Tips (cmdlinetips.com)

board-game-rating-prediction-based-on-reviews (mintu07ruet.github.io)