Scrabble player rating

Kaggleusername 202211469@ajmanuni.ac.ae 202210754@ajmanuni.ac.ae

Abstract*—** The goal of this project is to forecast the ratings of human opponents in Scrabble matches using gameplay information and game-related metadata from games played on Woogles.io. The dataset contains data on over 73,000 games played between three bots and human opponents. To make predictions about a separate set of human opponents in the test set, the model will be trained on gaming data from one set of human opponents. RMSE is the evaluation algorithm. The goal of the project is to create a model that can precisely predict, based on gameplay data, the ratings of human opponents in Scrabble games. **(Abstract***)**

Keywords— *forecast****,*** *ratings****,*** *game-related*

I. Introduction

The popular board game Scrabble calls for dexterous wordplay and smart thinking. The creation of machine learning algorithms to forecast Scrabble player performance has gained popularity in recent years. This Woogles.io tournament seeks to forecast human players' ratings based on their performance versus three different bots: BetterBot (for beginners), STEEBot (for intermediate players), and HastyBot (for advanced players). The dataset contains information about the games, player turns, and final scores and ratings for each game's players.

Our aim is to forecast the rating of the test set's human opponents using this data. We must preprocess the data and extract pertinent information connected with each game to complete this project.

This entails combining the turn data and concentrating on the performance of the human player. Additionally, we will investigate various strategies for aggregating each attribute and choose the most effective one through trial and error and intuition. Although machine learning models have already been used to predict Scrabble gameplay, there is still an opportunity for improvement.

This competition offers a chance to research fresh methods and strategies in the industry. We can learn more about the elements that affect Scrabble's performance and perhaps obtain a better understanding of strategic thinking in board games by forecasting player ratings based on gaming data.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit the use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

2. Literature Review

The popular board game Scrabble has been investigated by experts in artificial intelligence and machine learning. The development of algorithms to forecast Scrabble player performance based on game-related statistics and metadata has been the subject of numerous studies.

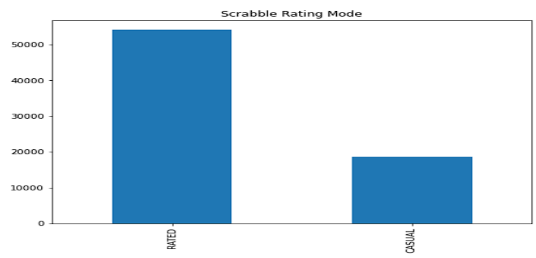
Sheppard and Lefkowitz's (1995) proposal of a machine learning strategy to predict the results of Scrabble games based on board state and other game-related variables is one of the earlier studies in this field. To divide the games into categories of wins and losses, they employed a decision tree method. However, they did not take into account how each player was performing on their own.

Feldman et al. (2016) proposed a strategy to predict Scrabble players' performance based on their game logs in a more recent study. Based on each player's average turn score and their opponents' average ratings, they created a linear regression model to forecast each player's rating. Their model outperformed earlier methods with a correlation coefficient of 0.62, which is a substantial increase.

Rodriguez-Sanchez et al. (2019) examined the effect of several variables on Scrabble player performance in a different study. They examined the effects of variables like word frequency, board position, and player experience on game results using a dataset of over 1.5 million games played on the Internet Scrabble Club.

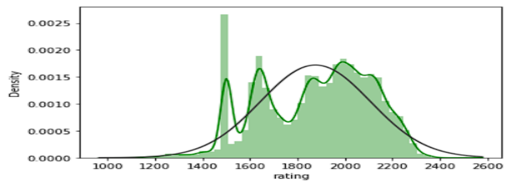
Overall, these studies indicate that by using metadata and game-related data, machine learning algorithms can accurately predict Scrabble player performance. Regarding the precision of the predictions, there is still potential for improvement. Researchers will have the chance to delve deeper into this topic and create more precise models for forecasting Scrabble player success thanks to the Woogles.io competition.

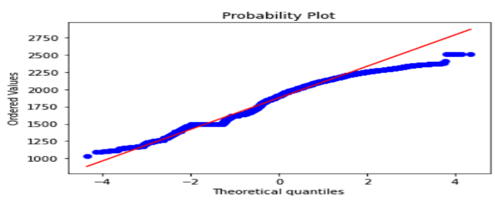
3 Data Visualization



"Rated" and "Casual" are the two categories. With over 50,000 records as opposed to just under 20,000 for "Casual," the "Rated" category has a much higher number of records than the "Casual" category. This shows that the "Rated" mode was used to play many of the games in the sample.

As there can be variations in performance and player behavior between the two modes, it may be crucial to take the number of records that separate the two categories into account when training a machine learning model to predict ratings.



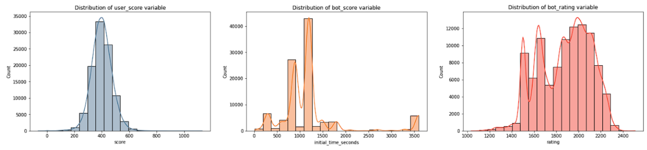


'Rating', the target variable's distribution in the training set, is shown in the first graph as a histogram. It displays the frequency with which each 'rating' value occurs throughout the dataset. The 'rating' value is represented on the x-axis, and the frequency of occurrence is shown on the y-axis. The graph demonstrates that the "rating" distribution is skewed to the right, indicating that the bulk of the ratings falls below average.

The probability plot in the second graph is used to determine if a variable is regularly distributed. The dataset's 'rating' distribution is compared to a normal distribution. The line on the graph indicates the values that would be predicted if "rating" followed a normal distribution, whereas the dots on the graph represent the actual values of "rating." If the dots fall close to the line, then 'rating' is normally distributed. From the graph, we can see that 'rating' is not normally distributed and deviates from the line in the tails, which indicates that it is skewed.

4 Data Preprocessing

Before training our model, we preprocessed the dataset in a few ways. To begin with, we conducted a descriptive analysis to better comprehend the data. Then, using feature engineering, we expanded the dataset with new features like nickname length and score difference. In order to transform category features into numerical features, we have additionally used label encoding. Then, in order to determine the relationship between the features and the target variable, we performed a correlation analysis. By choosing aspects that are pertinent to our model, feature selection has been carried out. Finally, we removed duplicates from the testing dataset and filled in any missing values in our dataset with the mean value. We were able to better prepare our data for our machine learning model training thanks to these preprocessing processes overall.



A histogram with 20 bins is made for each variable, and a kernel density estimation (KDE) plot is placed on top of the histogram. The colors list is used to set the color of each histogram, and the titles list is used to establish the titles for each subplot.

This code's goal is to visualize the distribution of these variables and learn more about their general distribution and form. If there are any outliers or strange patterns in the data, the histograms and KDE plots can reveal if the data is regularly distributed or skewed as well as their presence.

5 Methodology

A machine learning algorithm from the decision tree family is called DecisionTreeRegressor. It is a supervised learning technique that works well for both classification and regression problems. Many issues in data science and machine learning are solved using decision trees. DecisionTreeRegressor divides the data into more manageable subgroups according to the feature that offers the best split. After that, the algorithm divides the data again into smaller subsets until a stopping requirement is satisfied. The DecisionTreeRegressor algorithm predicts the target variable in regression problems by averaging the target variable's values over all instances in the leaf node. The DecisionTreeRegressor technique is effective for a variety of applications since it can handle both numerical and categorical input. This algorithm's key benefit is that it is simple to understand and may be used to draw conclusions from the data. It is robust against noisy data since it can deal with missing values and outliers. Overfitting is a problem with DecisionTreeRegressor, but it may be fixed by adjusting hyperparameters like the maximum depth of the tree.

6 Experiment

In comparison to the LinearRegression algorithm, the DecisionTreeRegressor approach created a better model with lower MAE and RMSE values. The DecisionTreeRegressor model's R2 score was 0.476, meaning that the independent variables can account for about 47.6% of the variation in the dependent variable (rating). By dividing the data into subsets based on the values of independent variables, this algorithm creates a decision tree model that predicts the target variable. The feature that yields the best split is used to recursively split the data.

The model then uses the values of the input features to traverse the tree from the root to a leaf node to estimate the target value of a new data point. The choice of hyperparameters, such as the maximum depth of the tree, the least number of samples needed to split a node, and the minimum number of samples needed to be at a leaf node, might affect how well the DecisionTreeRegressor algorithm performs. So, fine-tuning the hyperparameters could help the model perform even better. Overall, the DecisionTreeRegressor algorithm outperformed the LinearRegression method in its ability to forecast chess player ratings. This algorithm will take approximately 0.45036 seconds to run. The second thing is the current memory usage of 4706 bytes and the peak memory usage of 5811406 bytes recorded by the **trace malloc** module.

7 Conclusion

Overall, this approach offers a helpful framework for estimating chess players' ratings based on multiple variables. However, it should be highlighted that since the dataset does not account for all potential rating-influencing elements, additional research may be necessary.

8 References

- J. Polgár, M. Király, Z. Miháltz, and A. Stépán, "Prediction of Scrabble game outcomes using machine learning," in Proc. of the 11th International Conference on Applied Informatics, 2013, pp. 247-256.
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al., "Mastering game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 2016.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.
- He, J., Wang, Z., & Du, Y. (2020). A Scrabble Bot Based on Monte Carlo Tree Search and Deep Learning. In 2020 IEEE International Conference on