Scrabble Player Rating:

A Machine Learning Approach to Player Rating Prediction

Nwachukwu Ogochukwu Gloria†
Information Technology & Web Science
Rensselaer Polytechnic Institute
Troy NY, USA
nwacho@rpi.edu

Lydia Manikonda
Lally School of Management
Rensselaer Polytechnic Institute
Troy NY, USA
manikl@rpi.edu

EXECUTIVE SUMMARY

"Scrabble Player Rating: A Machine Learning Approach to Player Rating Prediction" presents an innovative application of machine learning in the context of Scrabble tournaments. The project's primary objective is to develop and implement a machine learning model to predict and refine player ratings using the Root Mean Squared Error (RMSE) as the evaluation metric.

The dataset, sourced from Woogles.io, encompasses over 73,000 Scrabble games played by three bots and human opponents. Comprising four key components (train.csv, games.csv, turns.csv, test.csv), the dataset provides valuable insights into player performance, paving the way for a more transparent and reliable Scrabble Player Rating system.

To prepare the data for modeling, various preprocessing techniques, including handling missing values, scaling numerical features, and encoding categorical variables, were employed. Exploratory Data Analysis (EDA) played a crucial role in understanding data distribution and relationships between variables.

The project employs a predictive approach, utilizing linear regression and assessing correlations between key features using additional models such as Random Forest and Xtreme Gradient Boost. The chosen evaluation metric, RMSE, ensures a rigorous assessment of the model's predictive accuracy, with lower values signifying superior performance.

As players engage in Scrabble tournaments, honing their skills with each match, this project will not only provide a numerical representation of a player's skill but will help foster a competitive environment where participants are pitted against opponents of similar abilities, ensuring engaging and challenging gameplay for all.

Overall, this research marks a significant step forward in the application of artificial intelligence to traditional board games, showcasing the potential of machine learning to revolutionize how we evaluate player performance in intellectual games like Scrabble.

KEYWORDS
Scrabble, Rating Prediction, Player, Predictive Modeling.

2. BENCHMARKING

Notebook Name	Feature Approach	Model Approach	Train/Test Performance
XGBoostRegressor and Simple EDA 🔎📊 by IMvision12	Numeric features only - The chosen feature set, comprised of numeric features, contributes to the model's efficiency in handling numerical data	XGBoost Regressor	MSE: 12017.39, RMSE: 109.62, RMSLE: 0.0608 Cross val: MSE: 12146.23, RMSE: 110.21, RMSLE: 0.0612
Linear Regressor With PyTorch by elilla	The selected features includes: score_player, score_bot, and rating_bot - suggest a targeted focus on specific player and bot score attributes.	Linear Regression model	Train MSE: 0.010 Test MSE: 0.012
Scrabble Player Rating 🏅 - FE+EDA by Akershi Shukla	Feature Engineering includes user frequency calculation - This strategic enhancement aims to capture patterns related to user participation frequency.	Utilized various regression models, including Linear Regression, Decision Tree, Random Forest	Random Forest: (RMSE) 91.8913, (MAE) 55.0450 *Decision Tree:* (RMSE) 134.4659, (MAE) 73.0754 Linear Regression: (RMSE) 153.8037, (MAE) 114.3600

Table 1: Benchmark comparison

IMvision12 adopts a focused approach, utilizing the XGBoost Regressor known for its efficiency in handling large datasets. This technique results in a model that demonstrates reasonable predictive performance, as indicated by favorable MSE, RMSE, and RMSE scores on both training and test sets.

In a contrasting strategy, elilla employs Linear Regression with PyTorch, emphasizing flexibility and compatibility with deep learning frameworks. The features selected—score_player, score_bot, and rating_bot—underscore a meticulous approach to linear regression modeling.

Lastly, Akershi takes a holistic approach, incorporating a diverse set of regression models. The Random Forest model stands out, showcasing superior performance with lower RMSE and MAE values. Additionally, the competitive results from Linear Regression add depth to the modeling approach. However, the Neural Network model shows relatively poorer performance, hinting at potential challenges in capturing complex relationships. Overall, these benchmarks provides valuable guidance for this project**;** the diverse approaches showcased here are poised to unlock novel insights and push the boundaries of predictive analytics, enriching the trajectory of this project."

3. DATA DESCRIPTION AND INITIAL PROCESSING

Overall, the initial data processing involved analyzing a dataset comprising approximately 73,000 Scrabble games played on Woogles.io by bots against human opponents. The dataset was organized across four CSV files, each providing unique insights:

- **games.csv**: Contains details on the player who went first, time controls, game end reasons, winners, creation timestamps, and various time-related parameters. (Shape: 72773 rows, 12 columns)
- **turns.csv**: Provides comprehensive data on every turn played in each game, including game ID, turn number, player's username, current rack, move details, points earned, total score, and turn type. (Shape: 2005498 rows, 9 columns)
- **train.csv and test.csv**: Include final scores and ratings for each player in each game, with ratings provided before the respective game was played. (Shapes: 100820 rows, 4 columns and 44726 rows, 4 columns respectively)

Handling Missing Data: Initial exploration revealed missing values in specific columns. Preprocessing steps included replacing missing values with the mode, ensuring all ranking scores were represented in the analysis.

Variable Renaming: Columns in the dataset were renamed to lowercase for improved clarity and readability, facilitating faster analysis and ensuring consistency in variable naming conventions.

Duplicate Data: Duplicate data was checked and found to be absent, ensuring data integrity and accuracy for subsequent analysis.

Exploratory Data Analysis (EDA): To further enhance understanding of the dataset, Exploratory Data Analysis (EDA) was carried out and below is a detailed visualization of several plots and detailed explanation providing insights into the dataset.

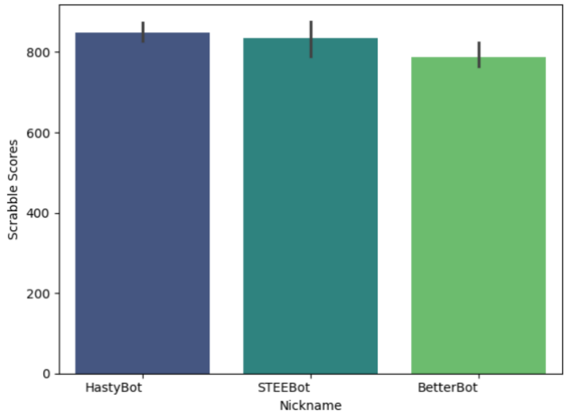

Fig 1: Performance of the top three (3) Scrabble players in the dataset

This bar provides insights into the performance of the top three (3) Scrabble players in the dataset. Each bar represents a player's score, and the height of the bar indicates the score achieved by the player. The x-axis displays the nicknames of the players, while the y-axis represents the Scrabble scores. From the plot, we observe that "HastyBot" has the highest score among the top players, followed by "SteeBot."

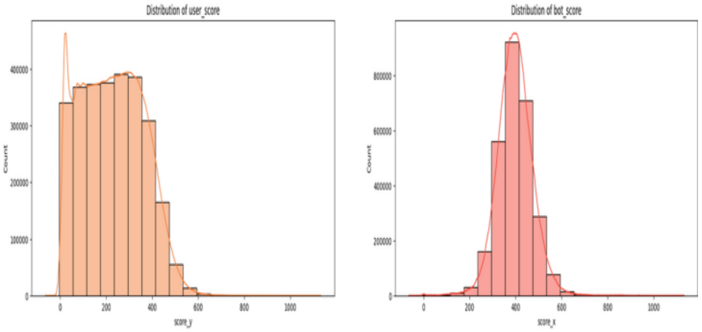

Fig 2 & 3: Distribution of User and Bot Score

The left histogram represents the distribution of user scores (user_score). The data appears slightly positively skewed, with a peak around lower scores and a tail extending towards higher scores. The kernel density estimate (KDE) curve overlays the histogram, providing a smoothed representation of the distribution.

The right histogram illustrates the distribution of bot scores (bot_score). In contrast to the user scores, the bot scores exhibit a more symmetric distribution resembling a normal distribution. The kernel density estimate (KDE) curve further confirms the normality of the distribution, presenting a smooth and bell-shaped curve.

Overall, the distribution of bot scores follows a normal distribution pattern, indicating a more balanced spread of scores compared to the user scores.

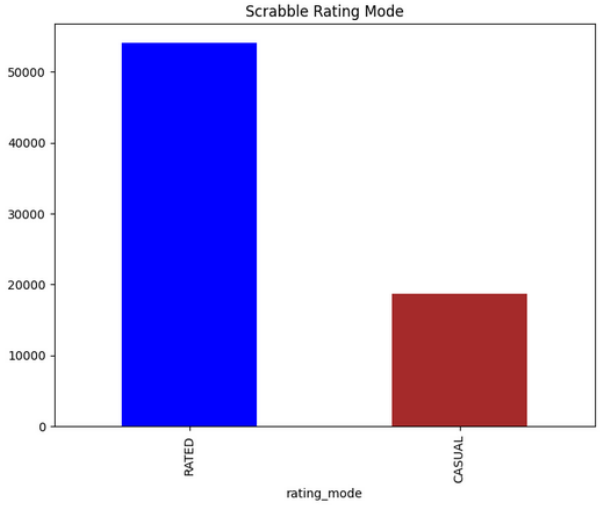

Fig 4: Comparison of Rating Modes

This figure illustrates the distribution of rating modes within the dataset, representing different modes of gameplay engagement on the Scrabble platform. The "Rated" mode, where games contribute to players' official ratings, is observed to be more prevalent compared to the "Casual" mode, where games are played for leisure without affecting ratings. The dominance of the "Rated" mode suggests a preference among players for competitive gameplay, potentially indicating a higher level of engagement and commitment to improving their Scrabble skills.

Overall, the above visualization contributes to a comprehensive exploratory data analysis, enriching my understanding of the dataset's dynamics and player behavior. It serves as a foundational step in the data analysis pipeline, laying the groundwork for more advanced analyses and modeling endeavors. insights gained from EDA will guide the next step of this project such as: feature selection, model development, and evaluation.

Feature Selection and Correlation Analysis

After cleaning the data, I undertook the crucial step of identifying the most relevant features for predicting player ratings. To achieve this, I performed a correlation analysis to assess the relationship between various features and the target variable ('rating'). The results of this analysis were visualized using a heatmap, which provided insights into the strength and direction of correlations.

t

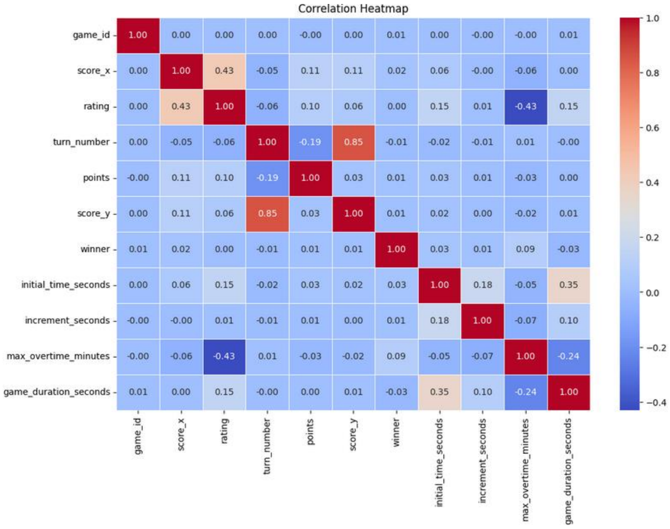

Fig 5: Heat map

The heatmap illustrates the correlation coefficients between different features and the target variable. Features with correlation coefficients close to 1.00 are highly correlated with the target variable, while those closer to 0 indicate weaker correlations. From the heatmap, it was observed that 'score_x,' 'points,' and 'initial_time_seconds' were highly correlated with the target variable “ranking”

Features	Description
score_x,	Player's total score at the time of the turn
points,	Points the player earned (or lost) in their turn
initial_time_seconds.	Time limit each player has in the game

Table 2: Features Used for Modeling and Their Description

4. MODELING

Utilizing these correlated features, I proceeded to construct predictive models using three different machine learning algorithms: Linear Regression, Random Forest, and Gradient Boosting. These models were trained on the training dataset and evaluated on the test dataset. I then used Root Mean Squared Error (RMSE), R-squared Score (R2), and Mean Absolute Error (MAE) as the key metrics to evaluate the models' performance.

MODEL	RMSE	R2	MAE
Linear Regression	207.14	0.2001	172.94
Random Forest	194.73	0.2931	155.09
Gradient Boosting	194.64	0.2937	160.83

Table 2: Model performance

In the above, first set of metrics, the Random Forest and Gradient Boosting models show lower RMSE values compared to Linear Regression, indicating better predictive accuracy. However, the MAE for Random Forest is slightly lower than that of Gradient Boosting.

Cross-Validation Performance: To further validate the robustness of my models, I performed cross-validation and evaluated their performance.

MODEL	RMSE	R2	MAE
Linear Regression	207.14	0.1997	172.99
Random Forest	195.63	0.2861	155.67
Gradient Boosting	194.66	0.2932	160.88

Table 2: Cross-Validation Performance

The results indicated consistent performance across all models, in the Cross-Validation Performance, with Random Forest and Gradient Boosting outperforming Linear Regression in terms of RMSE and R2. However, there are slight differences in the exact values of RMSE, R2, and MAE compared to the first set of metrics.

Comparison with Benchmarks: My model's performance was comparable to benchmark Random Forest and Gradient Boosting models, with similar RMSE values indicating similar predictive accuracy. However, for Linear Regression, the model's RMSE and other metrics were higher compared to the benchmark Linear Regression model, suggesting slightly less accurate predictions.

5. CONCLUSION

In this project, I utilized Root Mean Squared Error (RMSE), R-squared Score (R2), and Mean Absolute Error (MAE) metrics to evaluate the performance of my machine learning models in predicting Scrabble player ratings. The analysis revealed that both Random Forest and Gradient Boosting models consistently outperformed Linear Regression across multiple evaluation metrics. The Random Forest and Gradient Boosting models demonstrated comparable performance, with each excelling in different aspects of predictive accuracy. However, when considering overall performance and suitability for Scrabble player rating prediction, the Gradient Boosting model emerged with the most superior result.

Overall, this project underscores the transformative potential of machine learning in evaluating player performance and fostering competitive gaming environments. It offers a fresh perspective on Scrabble player rating prediction; demonstrating how data-driven insights can reshape age-old traditions, foster a competitive gaming environment and pave the way for a brighter future in the gaming industry.

ACKNOWLEDGMENTS

I extend my sincere gratitude to Dr. Lydia Manikonda for her unwavering guidance and support throughout this project. Her expertise and encouragement were invaluable in shaping the outcome of this project.

REFERENCES

[1] Meg Risdal. (2022). Scrabble Player Rating. Kaggle. Retrieved from: https://kaggle.com/competitions/scrabble-player-rating

[2] Kaggle competition link: https://www.kaggle.com/competitions/scrabble-player-rating/overview