CS 559: Machine Learning Fundamentals and Applications

Instructor: Suk Jang

From:

Praneeth Gubba

Sahithh Gudiyella

Hemil Patel

Rocco Vaccone

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

INDEX

INTRODUCTION:

Our team competed in the second edition of the Scrabble Player Rating competition. The data was extracted from Woogles.io, an online volunteer non-profit game service. The website hosts Scrabble games where users can play against bots to improve the ranking associated with their accounts. Scrabble Player Rating took inspiration from the first competition on Woogles.io data, where competitors were tasked with predicting the point value of the 20th turn of a given Scrabble game. Kaggle challenged our team to predict the ranking of an account, given their performance during various games.

Kaggle used a root mean square error (RMSE) evaluation method against team submissions because the ratings were continuous values and not discrete. Our team, Project 2 Team 3, scored an RMSE weight of 175.9548, ranking 200 on the private leaderboard on the day of our submission on the 13th. Our team still ranks relatively strong currently, with a ranking of 233rd.

DATASET:

The total dataset contained information from around seventy-three thousand Scrabble games played by three different bots on Woogles.io. The recorded games occurred between regular registered users and a given bot. The Kaggle hosts used metadata from the individual games, including data extracted from each turn. Competitors trained models on distinct datasets on three different CSV files: train.csv, games.csv, and turns.csv. The first CSV contained data from each game, including which player won and who went first. The second file had information regarding the turns played by each player in each game, featuring their points earned and the type of turn played. The final dataset consisted of final scores and ratings from before a given game for each player in each game.

- Games.csv
  - Metadata about games (e.g., who went first, time controls)
  - Used few columns from this dataset for training the model.
- Turns.csv
  - Data about turns played by each player
  - All turns from start to finish of each game
  - Done preprocessing over the dataset and used for training the model.
- Train.csv
  - Final scores and ratings for each player in each game
  - Ratings for each player are as of BEFORE the game was played
  - Used for training the model.
- Test.csv
  - Final scores and ratings for each player in each game
  - Ratings for each player are as of BEFORE the game was played
  - Used for testing the model built.

OBSERVATION AND STRATEGY:

Our team first noticed that a large portion of the data was categorical. However, we noted that most numerical columns contained few NAN values. Looking further into the datasets revealed that there were many games for the same player. At the same time, the total collected data occurred over a short period, significantly indicating that player ratings did not have enough time to change drastically. At the same time, a large number of players only played casual mode, so predicting their rating was a guessing game.

Our team took an individualistic approach in assessing how we should perform preprocessing. Each member attempted preprocessing separately, and then we implemented the best method by combining different aspects of each's.

MODELS IMPLEMENTED:

- **BAYESIAN RIDGE :**

  Bayesian Ridge Regression is a type of regularized linear regression model that uses Bayesian techniques to estimate the model parameters. The model is based on the assumption that the target variable follows a Gaussian distribution and that the features are independent given the target variable.

  Overall, Bayesian Ridge Regression can be a useful tool for linear regression tasks where it is important to estimate the uncertainty of the model parameters or to select a sparse set of features.

  RMSE SCORE obtained: 138.71

- **LINEAR REGRESSION:**

  Linear regression is a simple and widely used statistical model that can be used for a variety of prediction tasks, including forecasting, trend analysis, and causal inference. However, it is important to note that linear regression assumes a linear relationship between the dependent and independent variables, which may not always be appropriate.

  RMSE SCORE obtained: 129.07

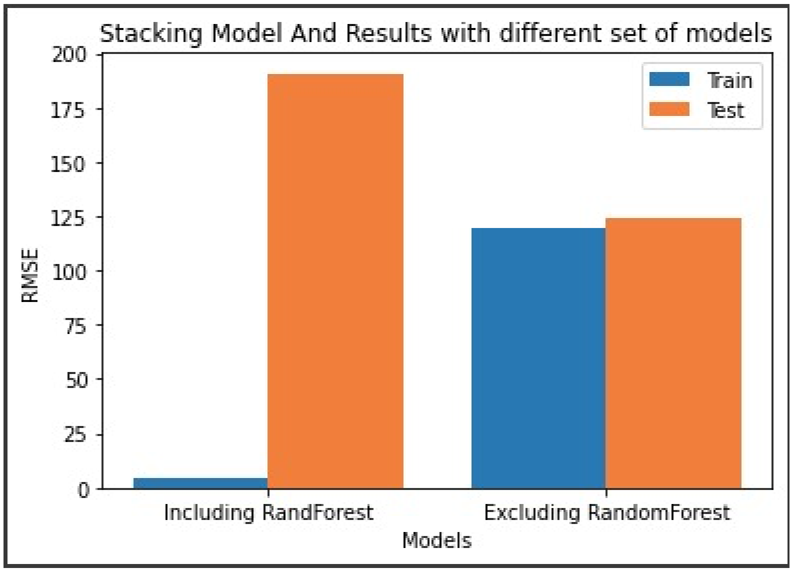
- **LINEAR REGRESSION WITH STACKING:**

  Linear regression with stacking is an ensemble learning method that combines multiple linear regression models to make predictions. Stacking is a method of ensembling where the predictions from multiple models are combined using a second-level model.

  Linear regression with stacking can be a powerful method for improving the accuracy of linear regression models, especially when the first-level models are diverse and accurate.

However, it can also be more computationally intensive than training a single linear regression model, as it requires training and evaluating multiple models.

RMSE SCORE obtained: 127.58



- **XGBoost:**

XGBoost (eXtreme Gradient Boosting) is an open-source software library that provides a gradient boosting framework for training and deploying efficient machine learning models. XGBoost is an implementation of gradient boosting that is designed to be efficient, flexible, and portable.

XGBoost uses decision trees as the base learners, and it applies several techniques to make the training process more efficient, such as weight pruning and column subsampling. It also provides a range of hyperparameters that can be tuned to control the complexity and behavior of the model.

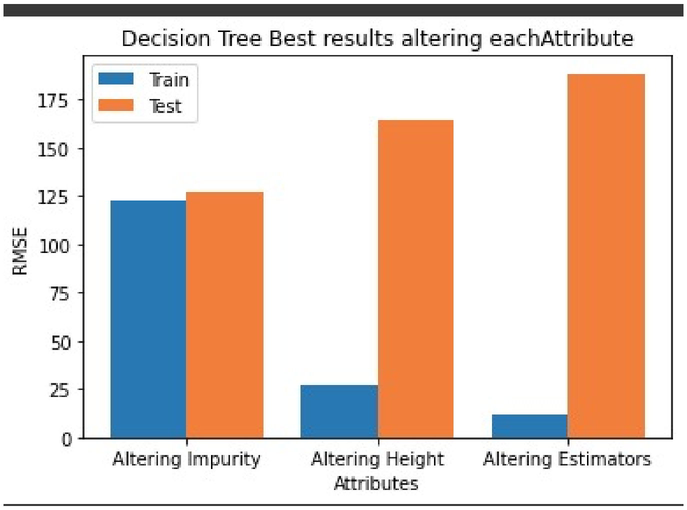
RMSE SCORE obtained: 127.01

- **DECISION TREE REGRESSOR:**

In decision tree regression, the model builds a tree of decisions based on the input features and the target variable. Each internal node in the tree represents a decision based on one of the input features, and each leaf node represents a predicted value for the target variable. The model makes predictions by traversing the tree from the root node to a leaf node based on the values of the input features.

Decision tree regression can be a useful tool for regression tasks where it is important to understand the relationships between the input features and the target variable. It can also be combined with other techniques, such as bagging or boosting, to improve the prediction accuracy.

RMSE SCORE obtained: 126.2644

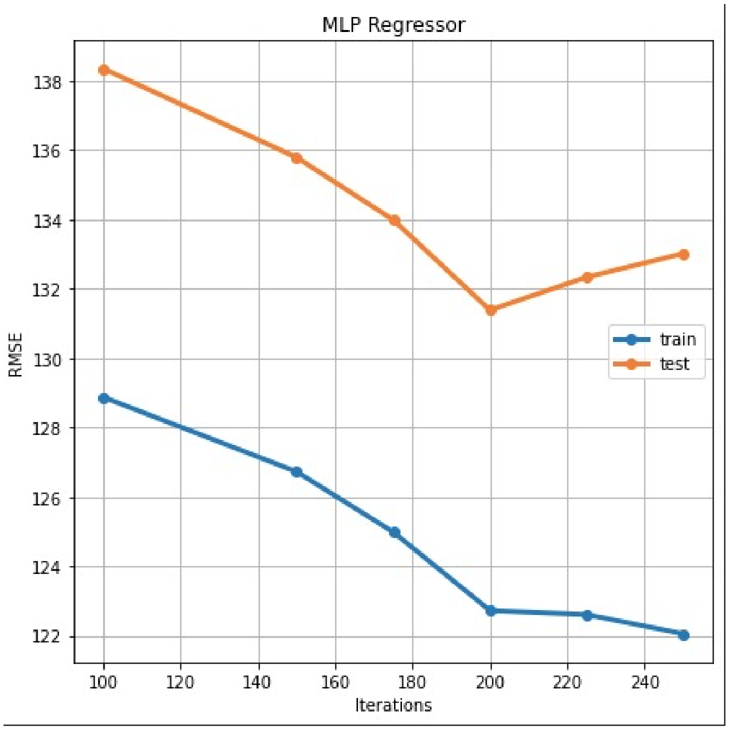


- **MULTI-LAYER PERCEPTRON CLASSIFIER:**

Multi-Layer Perceptron (MLP) is a type of artificial neural network that is used for supervised learning tasks, such as classification and regression. It is called a multi-layer perceptron because it consists of multiple layers of artificial neurons (also called units or nodes) that are interconnected.

MLPs are trained using an optimization algorithm, such as stochastic gradient descent, to minimize the loss function of the model. During training, the model adjusts the weights of the connections between the units to minimize the error between the predicted output and the true output.

RMSE SCORE obtained: 122.7242

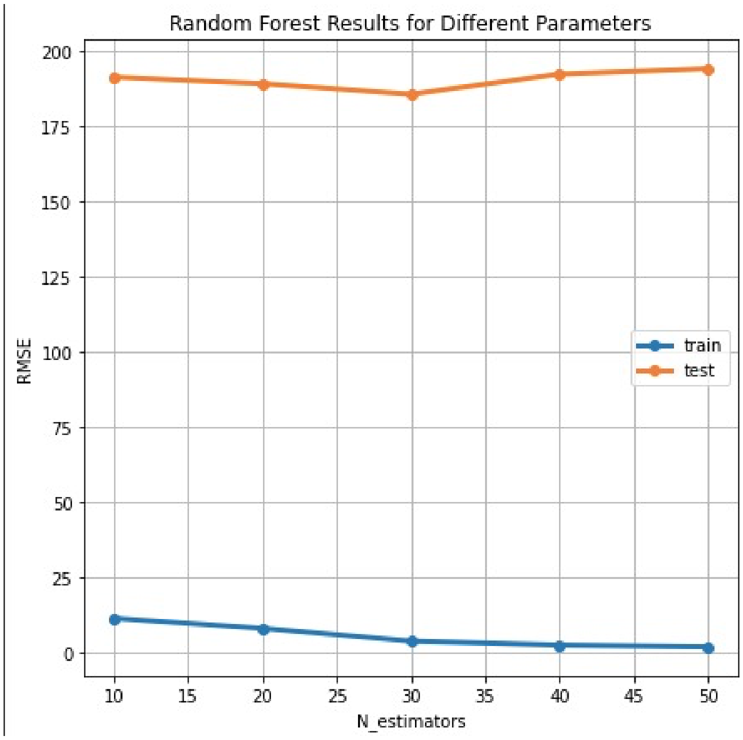


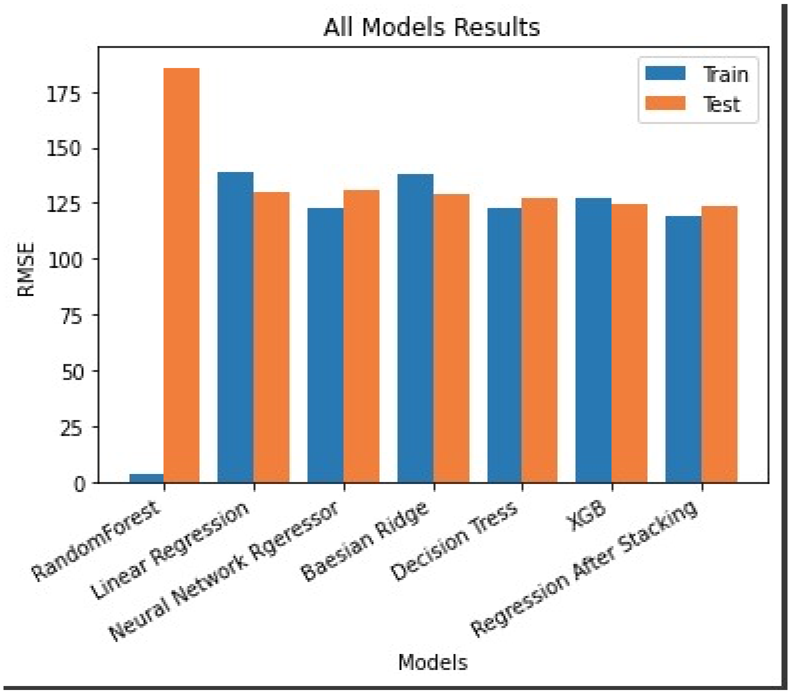
- **<u>RANDOM FOREST CLASSIFIER:</u>**

Random Forest is a type of ensemble learning algorithm that is used for supervised learning tasks, such as classification and regression. It is called a random forest because it is composed of a collection (forest) of decision trees, where each tree is trained on a random subset of the data.

One of the main advantages of random forests is that they can handle high-dimensional data and a large number of features, and they are relatively resistant to overfitting. They are also easy to interpret and visualize, as the decision trees can be visualized individually or as a whole.

RMSE SCORE obtained: 185.8313





CONCLUSION:

The competition demonstrated the effectiveness of Linear Regression with Stacking as an algorithm for regression analysis.

Our team, Project 2 Team 3, scored 175.9548, ranking 200 on the private leaderboard on the day of our submission on the 13th. Our team still ranks relatively strong currently, with a ranking of 233rd.