Antoine Dangeard
Thomas Inman

MAIS 202 Baseball Stats Model Deliverable 3

Final Training Result

Our final training result is the output below. Compared to our initial training, this is a pretty good improvement (with 2-3% better accuracy and MSE dropped by a factor of over 18). During training we were able to get higher accuracies (~60%) but those results often had extremely high MSE and so we decided to prioritise lowering the MSE. Our final result was achieved with Ridge regression, implemented using sklearn, as well as the addition of PCA, individual player stats as opposed to just team stats (so that starting lineup affects the outcome as well), hyperparameter tuning (notably alpha, the number of principal components for PCA and max. number of iterations), and other small fixes that made our model more reliable.

Ridge:
Mean Squared Error: 0.98442179
RMSE: 0.99218032131
Accuracy: 57.769701267618125%
Parameters: {'alpha': 0.11280000000000001, 'copy_X': True, 'fit_intercept': True, 'max_iter': 5000, 'normalize': 'deprecated', 'positive': False, 'random_state': None, 'solver': 'auto', 'tol': 1e-05}
Mean of predicted: -0.0014583339177951255
Mean of true: -0.008814254780432431
Hyperparameters: num_components: 500, max_samples: -1, startYear: 1990, endYear: 2020

Since the last deliverable, we also changed the data selection so that games that took place within the first 20% of the regular season would take the player stats from the previous year, as overall season stats may not represent that player as well if the season is just starting. Moreover, we tried other things like reversing each game sample to increase our number of samples, (so putting the same game in but the home team and visiting team's stats are swapped and the score difference is reversed), using different models (Random Forests, Support Vector Machines, Lasso, K-Neighbours and Decision Trees) but these changes ultimately did not help our model's performance and made computation too long, so we decided on Ridge and Linear regression to fine tune and use for the final version.

Final Demonstration Proposal:

For our final project, we are going to create a web app that would guess the outcome of a baseball game. The user would enter a starting lineup for both teams out of a list of all players on file (9 players per team). Then, our webapp scrapes the MLB stats website for those players' current stats, runs them through the model, and displays the predicted score for each team. If time allows, we could also make it automatically try to guess the outcome of upcoming MLB games and show them on the landing page when the webapp starts.