written_report

Question

Can we predict a player's performance/worth in an upcoming season based on their previous performance stats and other metrics?

Project 1 Data Wrangling

Choose the question to consider for the entire project. Find data for the project and clean the data to get it ready for modeling. Simply run data_wrangling.py to clean the raw data.

After merging the Baseball Reference and Statcast datasets from the 2015-2024 seasons into one giant dataset, we are now able to start modeling.

Project 2 Exploratory Data Analysis

Univariate, Bivariate, and Multivariate data analysis in eda.qmd I hand-picked 8 features that I felt were most important to predict WAR.

Main Takeaways: - All data features looked to be distributed somewhat normally, with some skewed slightly. - Rbat+ had the most correlation with WAR , which makes sense because Rbat+ measures offensive run contributions, which directly affects a player's value. - Player age has very little to no correlation with WAR. This is likely due to baseball being a sport where longevity is much more of a norm compared to other sports.

Project 3 Modeling

LSTM and Random Forest Models. project3.qmd contains all the code and analysis, modeling.py is all the code contained in project3.qmd

From Recurrent Neural Networks (RNN, specifically Long Short-Term Memory), we got a root mean squared error (RMSE) of around 1.7 and a r^2 value of around 0.2. These numbers

represent a model that is showing some signs of picking up patterns within the data but has lots of area for improvement. We could probably improve this model to make it more accurate in predictions but this might not be as good of a model as I initially thought for this project because not all players have the same developments; some players might be great in their late 20s and fall off dramatically after while others are consistently good throughout a long career but never great.

Pivoting to Random Forest Regressor model, we were able to get a RMSE of around 1.2 and r^2 of around 0.6. These numbers are significant improvements over the numbers we got from the LSTM model. There can be a few explanations for this phenomenon. Random Forests deal with noisy sequences better than LSTMs do. The data we have can be noisy because of a few factors: our data includes the shortened 2020 COVID season in which there were only 60 games played instead of the usual 162, which can lead to more outliers; some players may experience injury, slumps, or another player at the same position may take over many of their play-time; and we do not have enough player data for an LSTM model to reach its full potential.

Project 4 Monte Carlo Simulation

Data generation, model training, and analysis of each model (Random Forest, Linear and Ridge Regression). project4_simulation_sakai_sui.qmd contains all the code and analysis of the Monte Carlo simulation.

Ridge Regression performs slightly better than Linear Regression due to its regularization. Random Forest performs the worst overall but drastically increases accuracy as correlation increases while the two linear regression models have little to no difference between the differing correlations.

Conclusion

Final Verdict: It is possible to get a fairly accurate prediction of a player's WAR given their previous stats but in real life there are so many factors including but not limited to: injuries, off-the-field matters, nerves, slumps, etc. which make it hard to accurately predict how a player would perform.

Through this project, I was able to learn how to clean real-world data and test models to determine what the best model is for the situation. I also learned the importance of making any work reproducible and how much of a hassle that can be, both from my own work and from looking at the work of others.