To find the best possible model to predict if a baseball player will win a golden glove or not based on their fielding statistics, two different models were compared. Both models were created using grid searches, meant to find the most useful parameters for a model. For the golden glove dataset, the data was split into training and validation or test datasets. Data from 2023 was withheld as the validation data, and the model was trained on data from 2013 – 2022. Using a decision tree classification, and giving the parameter options between max depth, minimum samples split, minimum samples leaf, and class weight resulted in a model with no class weight, no max depth, 1 minimum samples leaf, and 5 minimum samples split for the final model. This process was repeated for a random forest classification model, but number of estimators was also included in the possible parameter options. The optimal model ended up with no class weight, no max depth, 1 minimum sample lead, 5 minimum samples split, and 10 estimators. Comparing the tree and forest models revealed that they had the same best scores, meaning that the models can equally predict. Since the models are equivalent, I chose the random forest model as my winning model, and tested the validation data on this model, getting 99.1% accuracy.

To find the best possible model the salary of a baseball player based on their batting, pitching, and fielding statistics, two different models were compared. Both models were created using grid searches, meant to find the most useful parameters for a model. For the salary dataset, the data was split into training and validation or test datasets. Data from 2016 was withheld as the validation data, and the model was trained on data from 2006 – 2015. Using a decision tree classification, and giving the parameter options between max depth, minimum samples split, minimum samples leaf, and criterion resulted in a model with squared\_error for the criterion, a max depth of 10, minimum samples leaf of 5, and a minimum samples split of 2 for the final model. This process was repeated for a random forest classification model, but number of estimators was also included in the possible parameter options. The optimal model ended up with friedman\_mse for the criterion, a max depth of 4, 5 minimum samples leaf, 10 minimum samples split, and 100 estimators. Comparing the tree and forest models revealed that the random forest model had a higher best score, so that was chosen as the winning model. Using the validation data on the winning model revealed an R^2 of 5%.

#different param\_grids?

#what to include in param\_grids

#how to have get\_metrics include ROC and AUC

#same results for classification models

#use neg\_mean\_squared\_error for regression grid searches?