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 set to balanced resulted in a model with a max depth of 1, a minimum samples leaf of 1, and minimum samples split of 5 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 a max depth of 1, a minimum sample leaf of 10, a minimum samples split of 10, and 10 estimators. Comparing the best scores from each model (which is measured by accuracy), the forest model had a higher score, and was chosen as the winning model, with an accuracy 93.03%.

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 the parameter options instead included criterion options, number of estimators, max depth, minimum samples lead, bootstrap, and an out of bag score. The optimal model ended up with friedman\_mse for the criterion, a max depth of 4, a minimum samples leaf of 1, bootstrap set to true, 100 estimators, and out of bag score set to true. Comparing the tree and forest models revealed that the random forest model had a higher best score (which was measured in negative mean squared error), so that was chosen as the winning model. Using the validation data on the winning model revealed an R^2 of 5.44%.