## Lab 2.3 – Building and Testing MLB leaning models

In this lab, we continue the work of the previous by building and testing various machine learning models for our two primary objectives, provided below.

**Research Questions.**

1. **Classification problem.** How well fielding statistics can be used to predict whether a player wins a gold glove at their position.
2. **Regression problem.** How well batting, pitching, and fielding statistics help predict a player’s salaries.

**Tasks.** Perform each of the following tasks before constructing the two data sets.

1. Perform a grid search for both single trees—as well as Random Forests—to find the best performing model for each problem.
2. In each case, validate the winning model by testing its performance on the validation data.
3. Write up a short report summarizing your findings.

**Link to my repo:**

**Short Report:**

1. **Classification Problem**

* **Decision Tree Classifier**

**A computer screen shot of a number

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

* **Random Forest Classifier**

The initial forest model (0.9933) has the same accuracy scores as the second model (0. 9933) with extra parameters.

**A screenshot of a computer program

Description automatically generated**

**A computer screen shot of a black screen

Description automatically generated**

* **Output**

|  |  |
| --- | --- |
|  | **AUC Scores** |
| **Decision Tree Classifier** | 0.6998 |
| **Random Forest Classifier** | 0.6998 |

In conclusion, based on the AUC Scores, there is no difference in performance between Decision Tree Classifier (0.6998) and Random Forest Classifier (0.6998). Since these models have AUC scores higher than 0.5, we can conclude that these models are not the best, but better than randomly choosing the winner of gold globe award.

The best Decision Tree Classifier has the following parameters:

min\_samples\_split = 5

The best Random Forest Classifier has the following parameters:

min\_samples\_split = 5, n\_estimators = 10

**A graph of a positive rate

Description automatically generated with medium confidence**

**A screenshot of a computer program

Description automatically generated**

1. **Regression Problem**

* **Decision Tree Regressor**

The initial tree (0.2271) works better than the second tree model (0.2068) with extra parameters.

The initial tree has the following parameters:

max\_depth = 5, min\_sample\_leaf = 10, min\_samples\_split =5.

A screenshot of a computer program

Description automatically generated

* **Random Forest Regressor**

The second model (0.3771) performs slightly better than the first model (0.3759).

The second random forest has the following parameters:

min\_samples\_split =5, n\_estimattors = 99.

**A screenshot of a computer program

Description automatically generated**

* **Output**

|  |  |  |
| --- | --- | --- |
|  | **R^2** | **MSE** |
| **Decision Tree Regressor** | 0.2180 | 23573377589891 |
| **Random Forest Regressor** | 0.2800 | 21497208211048 |

In conclusion, the second random forest model (min\_samples\_split =5, n\_estimattors = 99) performs the best among the other 3 models (R^2 = 0.2800). However, based on the low R^2 score, this model is not reliable to predict the salaries based on batting, pitching, and fielding statistics.

**Decision Tree Regressor**

**A screenshot of a computer program

Description automatically generated**

**Random Forest Regressor**

**A computer screen shot of a program

Description automatically generated**