Monterrey Institute of Technology and Higher Studies

State of Mexico Campus

Advanced artificial intelligence for data science I

Feedback Moment: Module 2 Analysis and Report on the performance of the model. (Portfolio Analysis)

Group 101

Gilberto André García Gaytán - A01753176

Teacher:

Jorge Adolfo Ramírez Uresti

Feedback Moment: Module 2 Analysis and Report on the performance of the model. (Portfolio Analysis)

- Report: Analysis of the Random Forest Model in the Soccer Players Dataset 3
- Introduction 3
- Justification for Selection of Dataset 3
- Separation and Evaluation of the Model (Train/Test/Validation) 3
- Training Set Sample: 4
- Validation Set Sample: 4
- Test Set Sample: 5
- Scatter plot of (Train/Test/Validation) 6
- Benefits of Set Separation: 6
- Comparison of Performance Metrics for Different Model Configurations 6
- Diagnosis of the Degree of Bias (Bias) and Variance 7
- 8 Adjustment Level Diagnosis
- Regularization or Parameter Adjustment Techniques 8
- Graphs 9
- Conclusion 12

Report: Analysis of the Random Forest Model in the Soccer Players Dataset

Github repository

Introduction

The objective of this report is to analyze and optimize a regression model based on the Random Forest algorithm, which is responsible for predicting the value of soccer players. Through diagnostic and tuning techniques, we seek to ensure that the model not only performs well on the training data, but also has excellent generalization ability to unseen data.

Justification for Dataset Selection

Dataset Context: The chosen dataset contains detailed records of players from the five main European football leagues. Each entry includes features such as position, club, league, nationality, and more. A crucial characteristic is player value, which is our target variable.

Suitability for the Algorithm: Random Forest is an algorithm that combines the predictions of multiple decision trees. Its ensemble nature makes it robust to variations and capable of handling categorical and numerical characteristics and capturing complex interactions. Given the heterogeneous characteristics and nature of the dataset, a Random Forest is particularly suitable.

Model Separation and Evaluation (Train/Test/Validation)

To ensure a fair and robust evaluation, the dataset was divided into three sets:

Training Set (60%): Used to train the model. This is where the model "learns" the relationships between the features and the target variable.

Validation Set (20%): Serves as an intermediate set to optimize and tune the model without touching the test data.

Test Set (20%): Reserved for final evaluation. Provides an honest metric of model performance on previously unseen data.

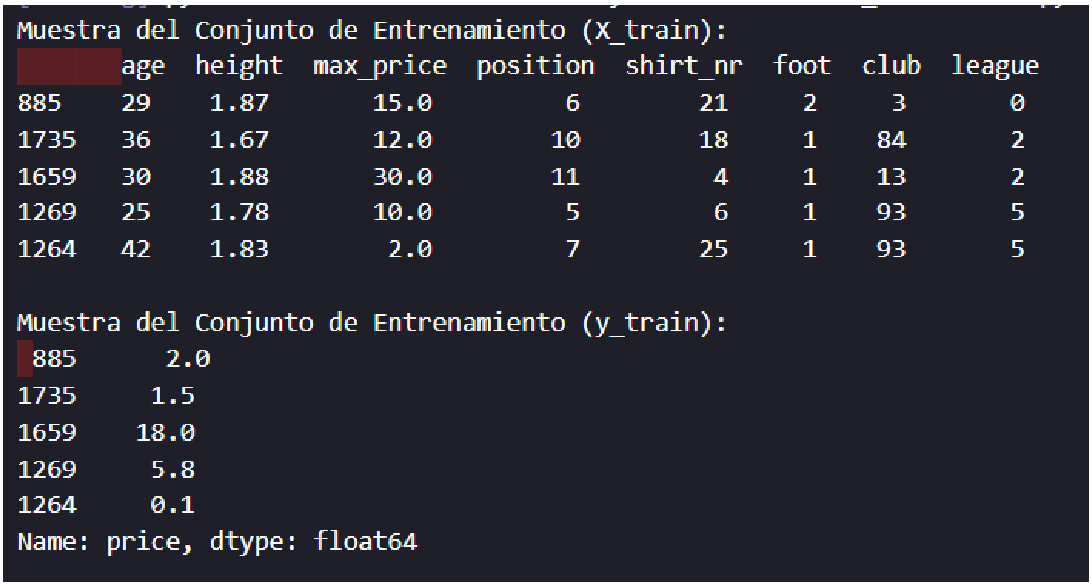
Training Set Sample:

This section shows the first five rows of the training set (X_train) and the corresponding labels **(y_train).**

X_train contains the training characteristics, such as age, height, etc.

y_train contains the corresponding target values (player prices).

This separation is done using **train_test_split** with 60% of the data for training.



```
Muestra del Conjunto de Entrenamiento (X_train):
       age  height  max_price  position  shirt_nr  foot  club  league
885     29    1.87       15.0         6        21     2     3       0
1735    36    1.67       12.0        10        18     1    84       2
1659    30    1.88       30.0        11         4     1    13       2
1269    25    1.78       10.0         5         6     1    93       5
1264    42    1.83        2.0         7        25     1    93       5

Muestra del Conjunto de Entrenamiento (y_train):
 885     2.0
1735     1.5
1659    18.0
1269     5.8
1264     0.1
Name: price, dtype: float64
```

Figure 1. Sample training set

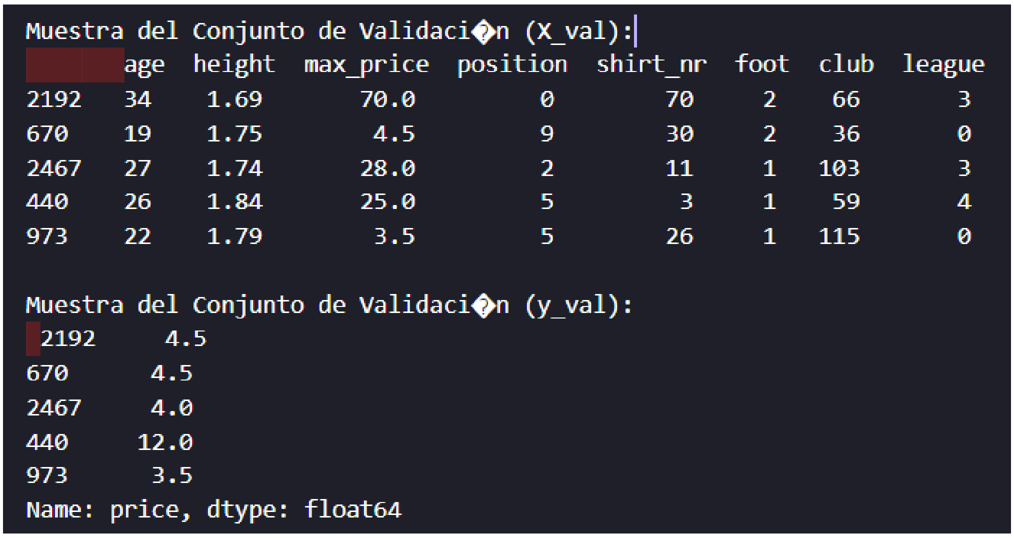
Sample Validation Set:

Presented here are the first five rows of the validation set **(X_val)** and the corresponding labels **(y_val).**

As in the training set, **X_val** contains features and **y_val** contains target values.

This set is used to tune and optimize the model without touching the test data.

The split is done using **train_test_split** again, this time with 20% of the data.



```
Muestra del Conjunto de Validaci�n (X_val):
       age  height  max_price  position  shirt_nr  foot  club  league
2192    34    1.69       70.0         0        70     2    66       3
670     19    1.75        4.5         9        30     2    36       0
2467    27    1.74       28.0         2        11     1   103       3
440     26    1.84       25.0         5         3     1    59       4
973     22    1.79        3.5         5        26     1   115       0

Muestra del Conjunto de Validaci�n (y_val):
2192     4.5
670      4.5
2467     4.0
440     12.0
973      3.5
Name: price, dtype: float64
```

Figure 2. Sample validation set

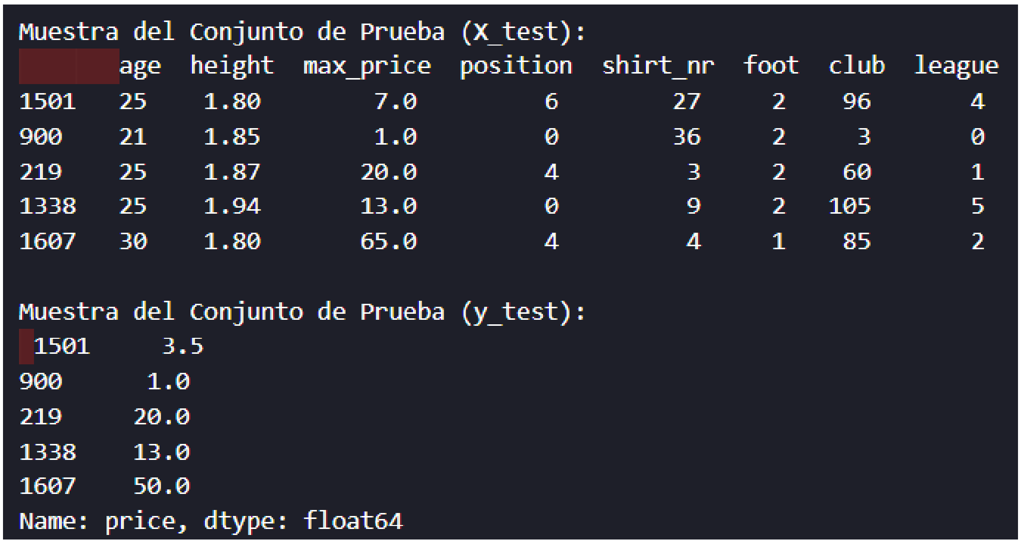
Test Set Sample:

This section displays the first five rows of the test set **(X_test)** and the corresponding labels **(y_test).**

X_test contains the test characteristics, and **y_test** contains the test target values.

The test set is used to evaluate the final performance of the model after it has been trained and tuned.

Train_test_split is also used , with 20% of the data.



```
Muestra del Conjunto de Prueba (X_test):
       age  height  max_price  position  shirt_nr  foot  club  league
1501    25    1.80        7.0         6        27     2    96       4
900     21    1.85        1.0         0        36     2     3       0
219     25    1.87       20.0         4         3     2    60       1
1338    25    1.94       13.0         0         9     2   105       5
1607    30    1.80       65.0         4         4     1    85       2

Muestra del Conjunto de Prueba (y_test):
1501     3.5
900      1.0
219     20.0
1338    13.0
1607    50.0
Name: price, dtype: float64
```

Figure 3. Sample test set

Scatter plot of (Train/Test/Validation)

They serve to visualize the separation of data between the training, validation and test sets.

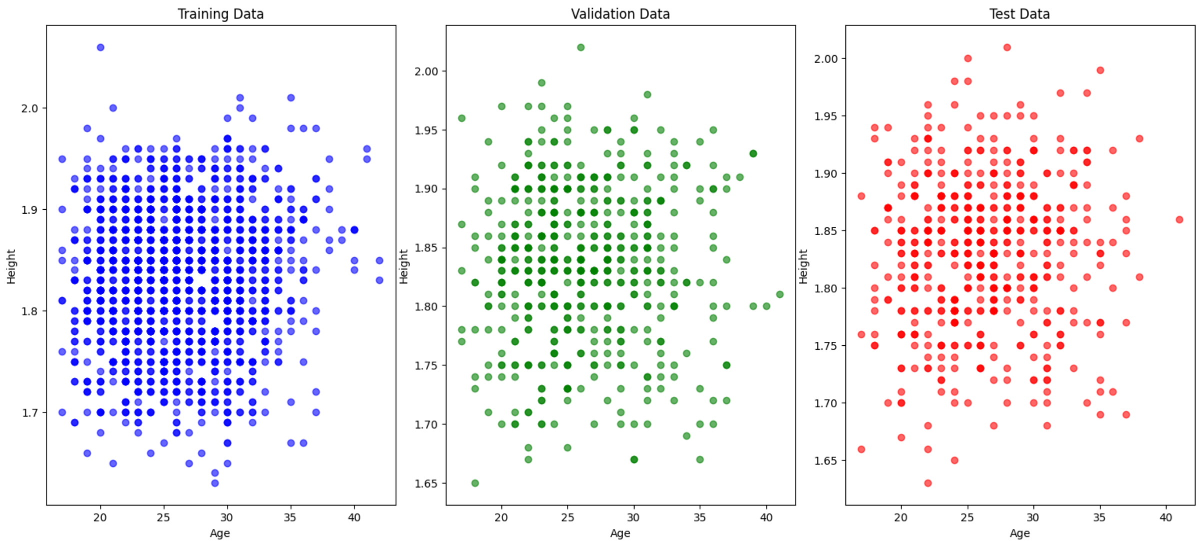


Figure 4. Train/Test/Validation visualization

Benefits of Set Separation:

Separating data into training, validation, and test sets allows you to fairly evaluate model performance and avoid overfitting.

The training and validation sets are used to train and tune the model, while the test set measures its ability to generalize to unseen data.

This explanation provides context on how the original dataset was divided into the different sets and why it is done this way to properly evaluate and tune the model. You can add this section to the report under the "Model Separation and Evaluation" section to complete your report.

Comparison of Performance Metrics for Different Model Configurations

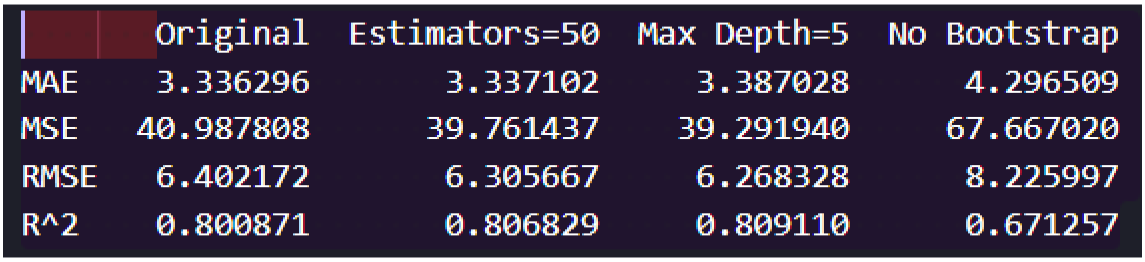
Evaluation Metrics:

MAE (Mean Absolute Error): This metric measures the average of the absolute differences between model predictions and actual values. The lower the MAE, the better the model is in terms of accuracy. Among the configurations, the model with “Estimators=50” has the lowest MAE (3.337102), indicating that it has the smallest average discrepancy between predictions and actual values in the validation set.

MSE (Mean Squared Error): MSE measures the average of the squares of the differences between predictions and actual values. Like MAE, a lower MSE indicates better performance. Again, the model with "Estimators=50" has the lowest MSE (39.761437), suggesting that it has a better fit to the validation data compared to the other configurations.

RMSE (Root Mean Square Error): The RMSE is simply the square root of the MSE and is interpreted in the same way as the MSE. A lower RMSE indicates a model with less error in its predictions. In this case, the “Estimators=50” setting has the lowest RMSE (6.305667), which again supports its good performance.

R^2 (Coefficient of Determination): The R^2 measures the proportion of the variance in the dependent variable (player value) that is predictable from the independent variables (player characteristics). A value of R^2 closer to 1 indicates a better fit of the model to the data. In this case, the "Max Depth=5" setting has the highest R^2 (0.809110), suggesting that it is the best at explaining variability in player prices.



```
       Original  Estimators=50  Max Depth=5  No Bootstrap
MAE    3.336296       3.337102     3.387028      4.296509
MSE   40.987808      39.761437    39.291940     67.667020
RMSE   6.402172       6.305667     6.268328      8.225997
R^2    0.800871       0.806829     0.809110      0.671257
```

Figure 5. Evaluation Metrics:

Diagnosis of the Degree of Bias (Bias) and Variance

Learning curves are powerful tools for diagnosing model performance:

Bias: If the model does not perform well on the training set, we can deduce that it has a high bias. This means that the model is too simple and does not capture the complexity of the data.

Variance: Observing the gap between performance on the training and validation set tells us the variance. A wide gap suggests that the model is overfitting the training data and has high variance.

Model Adjustment Level Diagnosis

The level of fit refers to how the model performs relative to its generalizability:

Underfitting: If the model performs poorly on both the training and validation sets, it is underfitting.

Overfitting: If the model has high performance on the training set but significantly lower performance on the validation set, it is overfitting.

Good Fit: If the model performs well on both sets and the performance gap is minimal, it has a good fit.

Regularization or Parameter Adjustment Techniques

Several techniques were implemented to optimize the model:

Adjusting the Number of Estimators: Reducing the number of trees in the forest can prevent overfitting and improve generalization.

Adjusting Maximum Depth: Limiting the depth of trees can help prevent the model from capturing noise and improve its generalization ability.

Using Bootstrap Sampling Techniques: Adjusting bootstrap sampling can affect the diversity of trees in the forest, which in turn can affect yield.

Graphics

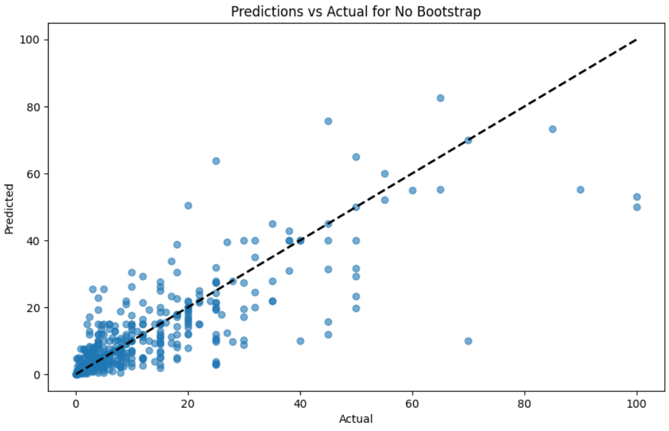


Figure 6. Predictions vs Actual for no bootstrap

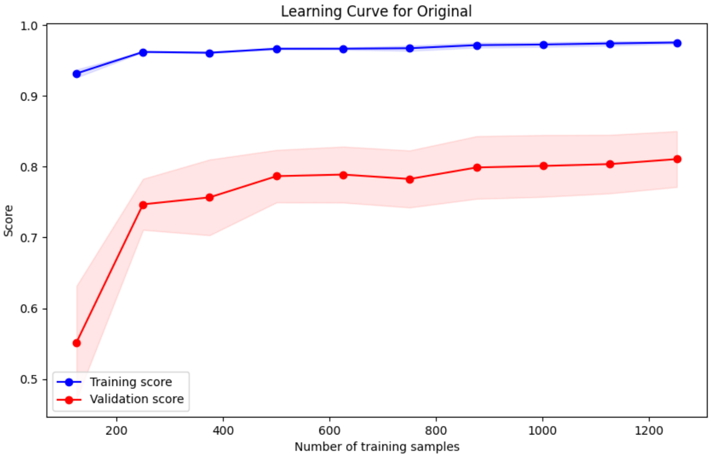


Figure 7. Learning Curve for Original

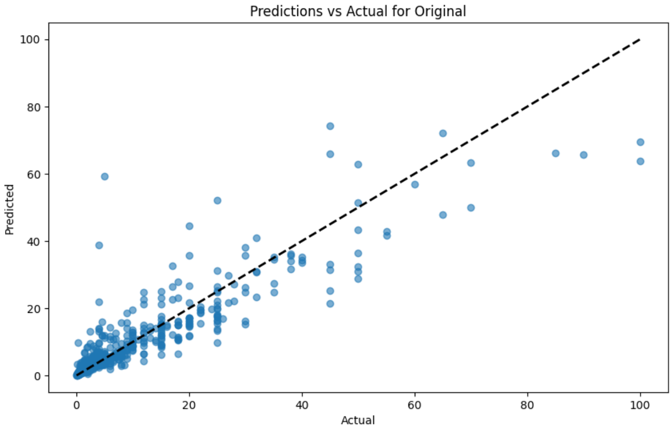


Figure 8. Predictions vs Actual for Original

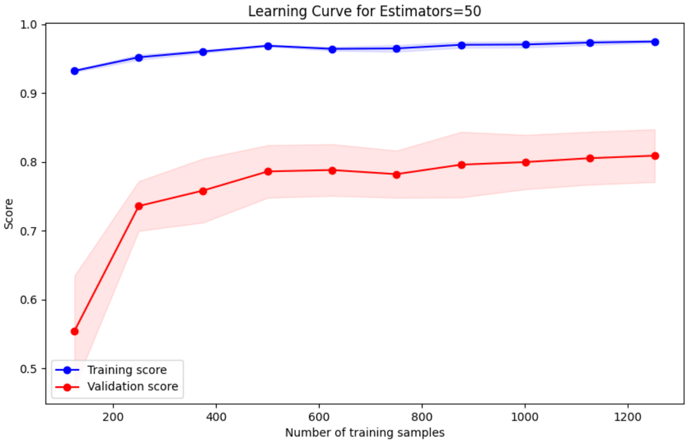


Figure 9. Learning Curve for Estimators = 50

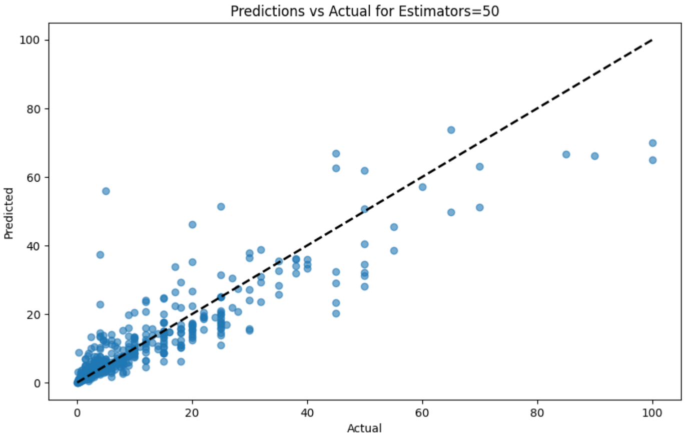


Figure 10. Predictions vs Actual for Estimators = 50

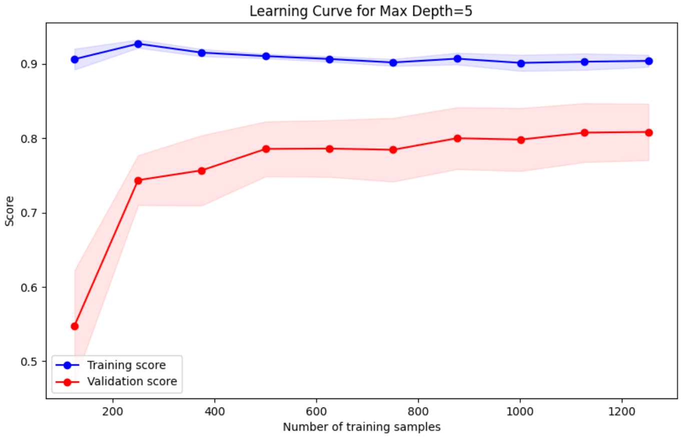


Figure 11.Learning Curve for Max Depth=5

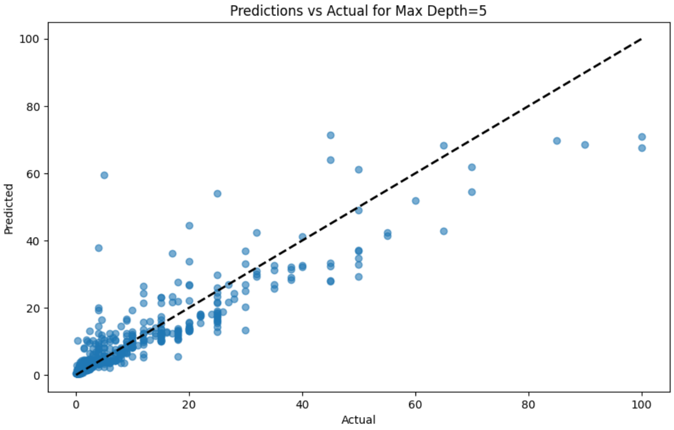


Figure 12. Predictions vs Actual for Max Depth = 50

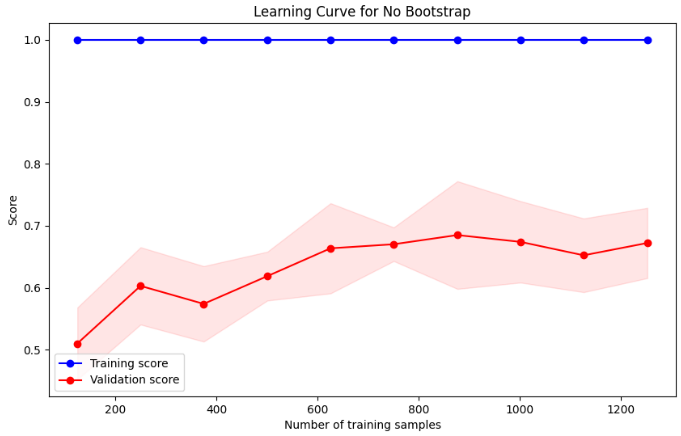


Figure 13. Learning Curve for No Bootstrap

Conclusion

The process of optimizing a Machine Learning model goes beyond simply training and testing. It involves in-depth diagnosis, constant adjustments and validations to ensure its generalizability. Through this report, we have demonstrated the critical steps and essential techniques to ensure that our Random Forest model is robust and accurate in predicting the value of soccer players. With the adjustments and techniques applied, the model is now better prepared to make accurate and consistent predictions in unseen data.