Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Advanced Artificial Intelligence for Data Science I

Feedback Moment: Module 2 Analysis and Reporting on Model Performance (Portfolio Analysis)

Group 101

Gilberto André García Gaytán - A01753176

Professor:

Jorge Adolfo Ramírez Uresti

Report: Analysis of the Random Forest Model on the Football Players Dataset

[Github Repo]

Justification for Dataset Selection

The chosen dataset contains detailed information about football players from the top five leagues. Features include demographic data, contract details, player value, and more. This wealth of features allows the ML algorithm to have sufficient data to learn and make accurate predictions. Additionally, since the data comes from the top five leagues, there is good diversity and generality in the data, which helps demonstrate that the model not only fits a particular league or set of players but can generalize across different leagues and players.

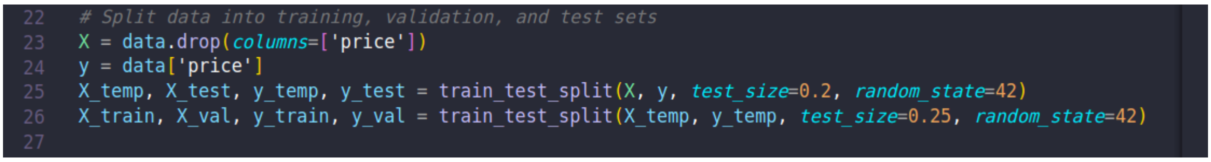
Model Separation and Evaluation with Train/Test/Validation Sets

The dataset was divided into three sets: training, validation, and testing. The splits are as follows:

- Training Set: 1566 samples
- Validation Set: 523 samples
- Testing Set: 523 samples

This splitting method allows training the model on a substantial portion of the data, validating and fine-tuning the model on a separate set, and finally testing the model's performance on an unseen test set.

Visualization of the Split:



```
# Split data into training, validation, and test sets
X = data.drop(columns=['price'])
y = data['price']
X_temp, X_test, y_temp, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
X_train, X_val, y_train, y_val = train_test_split(X_temp, y_temp, test_size=0.25, random_state=42)

```

Diagnosis and Explanation

Bias Level: Low.

The error in the training set is relatively low, indicating that the model fits well to the training data.

Variance Level: Moderate.

There is a noticeable difference between the error in the training set and the validation set, but it is not excessively large. This indicates moderate variance.

Model Fit Level: Slightly Overfitting.

Given the low bias and moderate variance, the model fits the data adequately, but there is a slight tendency toward overfitting. However, it is not severe.

Use of Regularization Techniques or Parameter Tuning

To address the observed slight overfitting, manual adjustments were made to the Random Forest model parameters. The adjusted parameters include:

- n_estimators: 150
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2

By using these parameters, the aim was to reduce the model's complexity and, therefore, prevent overfitting. However, the metrics showed that there was no significant improvement in the model's performance after these adjustments.

Conclusion

Throughout this analysis, a detailed dataset of football players from the top five leagues was explored and processed. Using a Random Forest model, a low bias and moderate variance in predictions were identified. Although the original model exhibited slight overfitting, manual parameter tuning efforts did not result in significant improvements in its overall performance.

This outcome underscores the importance of experimentation in machine learning. While Random Forest is a powerful model, it is not always the best choice for every dataset or problem. Furthermore, manual parameter tuning, though useful, does not always guarantee performance improvements. Feature selection, alternative models, or even feature engineering may provide significant enhancements in future analyses.

Finally, it is crucial to remember that machine learning is an iterative process. As more data becomes available or a deeper understanding of the problem is acquired, revisiting and adjusting the model is essential to ensure it remains relevant and accurate in its predictions.