Classification Project by Med Billian

Predicting NBA All-Star Players using NBA Statistics

Abstract

The goal of this project is to create a model that will help predict NBA All-Star players in a specific season. I scraped data from https://www.basketball-reference.com and joined NBA players statistics and teams statistics from 2000 to 2021. I then used this dataset to train the classification model using a Random Forest algorithm. Final model shows that 'Points per Game', 'Games Started', 'Assists per Game' and 'Team Standing' are the most important features in classifying an All-Star player.

Design

Predicting the roster of NBA All-Star players in a season through a machine learning model would greatly help the All-Star coaches pick players in their teams more efficiently.

I scraped the players statistics data and the team statistics data from the site mentioned above. I joined the two datasets by matching the team name and season fields.

Data Features include: Year, Position, Age, Games Played, Points per Game, Total Rebound per Game, Assists per Game, Field Goal per Game, Field Goal Percent, Effective Field Goal, Block per Game, Turnovers per Game, Personal Foul per Game, Minutes Played, Games Started, Steal per Game, Free Throws per Game, All-Star Player, Team Standing.

Algorithms

Feature Selection and Engineering in Final Dataset

1. Categorical feature was transformed to a binary feature
2. Missing numeric values in some fields were imputed by averaging the available field values.
3. Logit Model and VIF were used to narrow down features of the Training data.
4. Data Standardization was necessary in developing the Logistic Regression as the baseline model.

5. SMOTE was used to address the imbalance in the Training data.

Models:

Logistic Regression, Random Forest, GBM, and XGB classification models were developed using the Training data. The Random Forest model returned with the highest F1 and AUC scores using the Validation data, hence was decided to be the final classification model for predicting All-Star players.

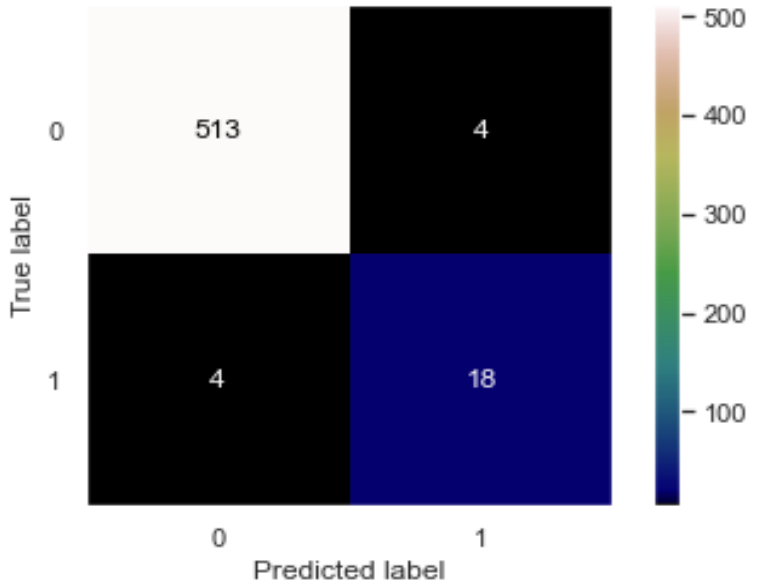
9,283 records were used to train models using NBA statistics from 2000 to 2019; 528 records were used to validate models using NBA statistics from 2020, and 539 records were used to test the final model using NBA statistics from 2021.

The metrics used to evaluate the final model were F1, AUC and Confusion Matrix. While the features_importances property of the Random Forest model was used to help determine that the 'Points per game', 'Games Started', 'Assists per Game', and 'Team Standing' are the features with the best predictive power in classifying an All-Star player.

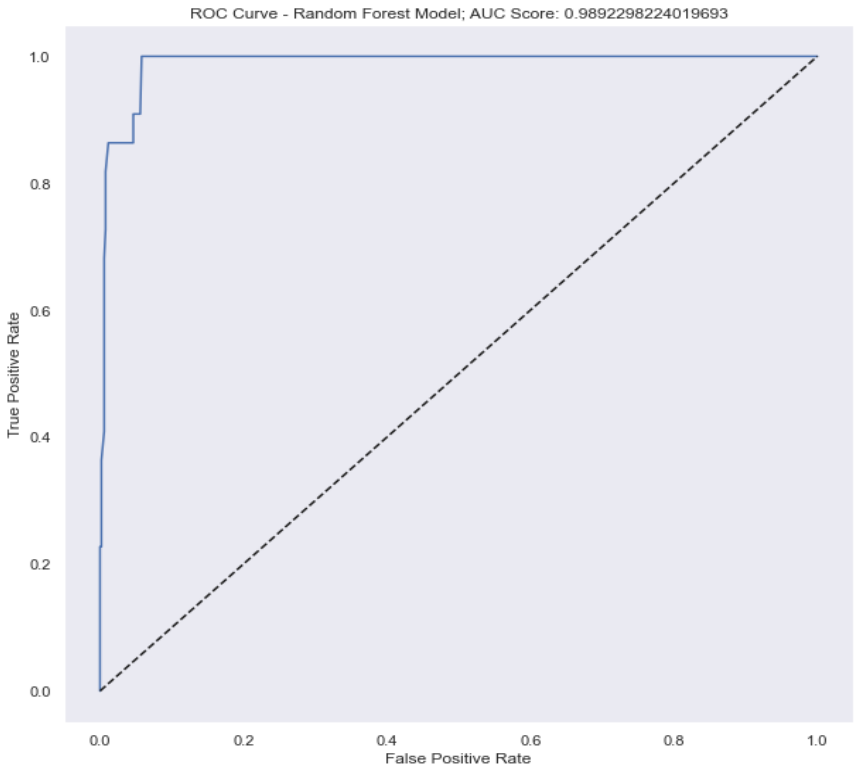
The Random Forest model scored well using the validation data and better using the test data. Addressing the imbalance in the training data and using the Random Forest algorithm greatly helped in improving the results from the Baseline model's F1 score=0.666 to the final model's F1 score=0.818. Finally, having many data points (~10K rows) also helped create a fairly accurate classification model.

Logistic Regression (Baseline) Model Metrics on Validation data: Accuracy= 0.963; F1=0.666; AUC=0.881

Random Forest (Final) Model Metrics on Test data: Accuracy= 0.985; F1=0.818; AUC=0.989



Random Forest - Confusion Matrix on Test Data



Random Forest - ROC Curve and AUC Score

Future Solution

The data feature ‘Games Started’ in a player’s statistics seems to be correlated with the player’s popularity, and thus was helpful in creating the final model. Other popularity metrics should be considered in the future to improve the performance of the Classification model.

Tools

- Numpy and Pandas for data manipulation
- Google Sheets for data storing and matching
- BeautifulSoup and requests for web scraping
- Statsmodels and Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Tableau for EDA and visualization

Communication

Slides and Visual graphs are created for presentation purposes. A Tableau Dashboard was created to help understand the relationship of the Features with best predictive power: https://public.tableau.com/app/profile/med.billian/viz/PPGvsTmSrs/NBAAll-Star