Analyst Intern, Data Science & Solutions Project

Project Outline

Forecasting All-NBA selections

1. Describe, generally, from start to finish how you approached, executed, and completed the project. Include all relevant materials (ex: code) through a cloud storage provider (ex: Dropbox, Google Drive).

**A1**. The first major hurdle that I had to cross was to understand what parameters or statistics prove how good a player is, for this I had to understand the game to a greater detail. After researching this, I began to formulate my plan and began to think of datasets. I was looking for specific kinds of datasets that would enable me to answer the questions posted in this assessment. Though the NBA site has vast amounts of data, I couldn't find all data there with the specific parameters that I was looking for. I procured multiple datasets and put them together to get my final dataset. This involved manipulation on Excel and python through code. I finally had my dataset ready! I created a new feature called picks which is the number of times a player was in the All NBA team up until now.

The next step was to clean the data, while doing text comparison of a dataset with the statistics with a dataset with the All NBA players, a simple apostrophe(') or a hyphen(-) can cause big problems as player names wouldn't match with them, so I had to clean up these inconsistencies(Text processing). After this I moved on to data exploration to see just how much each of these features varied and what caused them to vary, I did a correlation analysis and then realized that the data was skewed. To overcome this, I normalized the data in order to reduce bias in the model. I then ran a regression on 2 datasets, once on which I first applied dimension reduction and another without. The accuracy with dimension reduction was 35% and without it was 55%. Seeing that the accuracy wasn't that great I used a 3rd model with a focus on this (XGBOOST) which gave me 65% accuracy which means that it could predict the number of times a player would be in the All NBA team much more accurately.

1. Include and describe a visualization from the project. The visualization should highlight a feature or insight from your model. Explain the decisions you made in constructing the visualization.

(A2) One interesting visualization that I would like to talk about is the one below which is the variation points over the career of a player. The visualization below shows the variation of points of Karl Anthony-Towns, Kyrie Irving & Stephen Curry from 2010 to 2017, form this we can discern the performance levels of each player for each year. We can also look for substantial drops in points in a particular year which would usually occur due to a drop-in form or injury etc. We can also see the years for which they are/were active or playing. This kind of comparison help’s us estimate with a glance amongst, the players under consideration, whose performance has been consistently good or bad etc. With each feature that we visualize, we will have different insights which will help us understand player performance.



1. Using your model, estimate for the players listed below: (i) the total number of All-NBA selections remaining in each of these player’s careers, and (ii) the likelihood of each player having the greatest number of All-NBA selections remaining among this group.
   * Luka Doncic
   * Karl Anthony-Towns
   * Kyrie Irving
   * Stephen Curry

(A3) I have used 3 models to predict the total number of points of each of these players. Each model predicts this with a different level of accuracy.

The Regression model had an accuracy of 55.2% and the XGBOOST model had an accuracy of 65%.

The predicted total number of All NBA selections for these players are as follows:

|  |  |  |
| --- | --- | --- |
| **Player** | **Regression** | **XGBOOST** |
| Luka Donic | 0 | 6 |
| Karl Anthony-Towns | 8 | 6 |
| Kyrie Irving | 6 | 15 |
| Stephen Curry | 20 | 6 |

**References:**

<https://www.nba.com/players>

<https://www.basketball-reference.com/>