Project 3 - Predicting the NBA draft

Andy Tan
May 6, 2020

Executive Summary:

The aim of this project was to build a classification model to predict NBA draft status for collegiate basketball players.

- Random Forest classifier provided the best model for predicting NBA draft round.
- Features with high importance include boxscore plus-minus and year in school

Background:

The goal for every elite basketball player is to play in the NBA and every year, 60 players achieve this through the NBA draft. While being drafted comes with millions in financial gain, failing to be drafted has pitfalls as well. This is especially true for underclassmen who must forgo their remaining college eligibility when they enter the draft. The aim of this project is to build a classification model to predict which round a player is likely to be drafted in order to aid in their decision making.

Methodology:

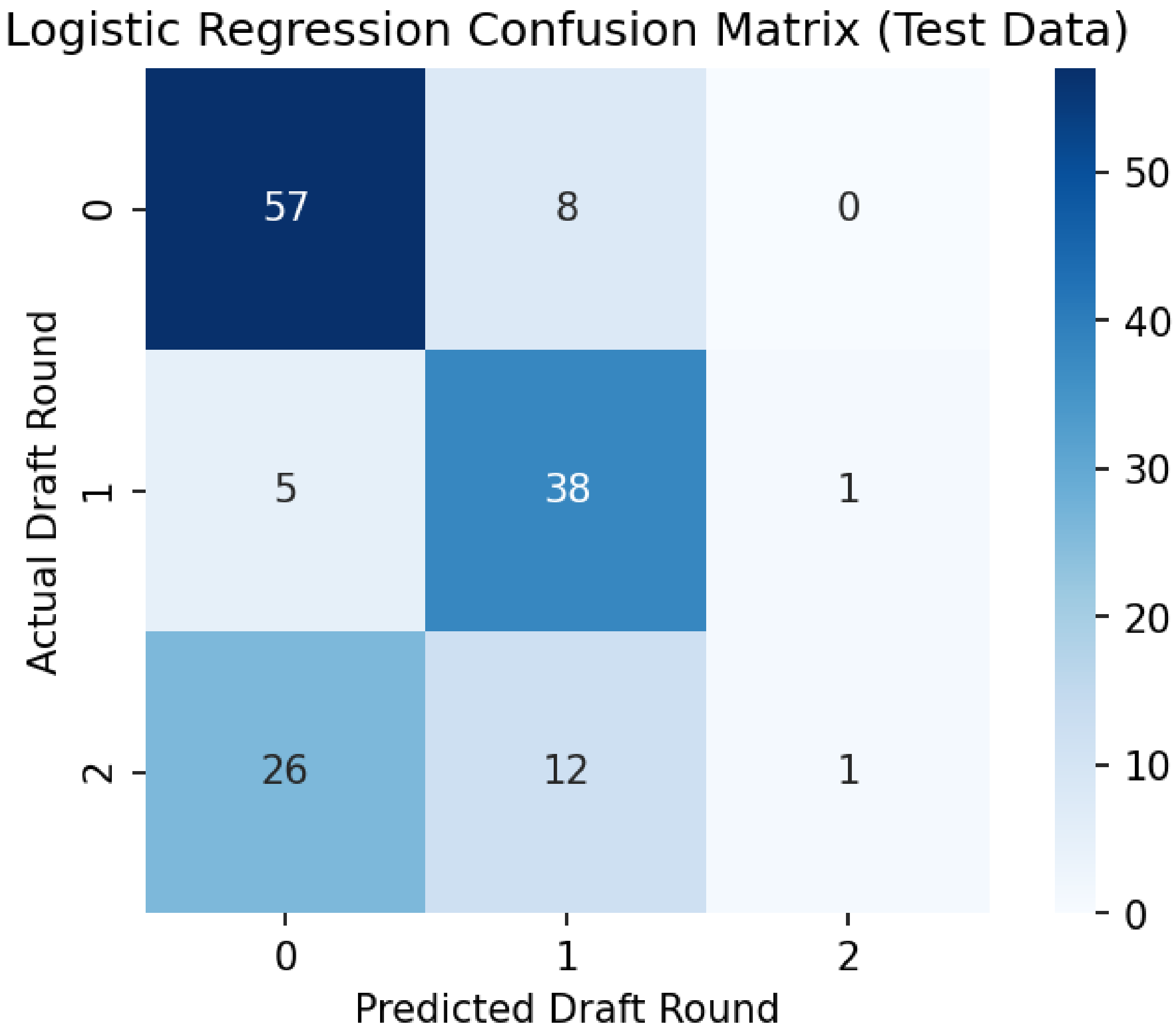
The primary dataset for this project was obtained through web-scraping of sports-reference.com through the use of their API. This includes yearly and career stats for every division 1 basketball player from the years 2010 to 2019. This dataset was subsequently uploaded to SQL in postgresql. To narrow the scope of the data to players with a reasonable change of actually being drafted, a dataset of prospects was created using information scraped from previous NBA draft results, NBA draft combine data, lists of previous early declaration players, and pre-draft rankings from nbadraft.net.

Analysis and Results:

After data cleaning and merging, the dataset consisted of 740 players who were eligible for the draft between 2011 and 2019. Features included basic collegiate statistics (points, assists, rebounds, etc.) transformed to a per-game measure. Advanced metrics such as boxscore plus-minus were then added. Physical attributes obtained from the NBA draft combine were also added. These measures were adjusted relative to mean measurements by position.

Using GridsearchCV with 10 fold cross-validation, the data was fit with 3 models - KNN, logistic regression (LR), and random forest (RF) classifier. The overall accuracy was similar between LR and RF with a score of .645 for LR and 0.659 for RF. On further examination of one-vs-all error metrics (F1, precision, recall), the LR model was found to have higher recall (0.877) for outcome 0 (going undrafted) and higher F1 score (0.745) for outcome 1 (drafted in 1st round).

As a result it was selected over RF despite a slightly lower overall accuracy score. A confusion matrix of the model on test data is as follows:



Conclusion:

A classification model was developed using logistic regression to predict NBA draft round status for collegiate basketball players. Important features include boxscore plus-minus and year in school. The overall accuracy on test data was 0.649. Other error metrics suggest this model is weighted towards recall of outcome 0, ie identifying those that are most likely to go undrafted. It also balances precision/recall for being drafted in the 1st round.