PURDUE
KRANNERT
SCHOOL OF MANAGEMENT

Using R for Analytics

R-Shiny DSS App Final Project

Summer 2021 MGMT 59000

Date: 08/14/2021

Jackson Bronkema	jbronke@purdue.edu
Ayush Maheshwari	mahesh15@purdue.edu
Dakor Gupta	gupta615@purdue.edu
Harsh Vardhan Kalra	hkalra@purdue.edu
Sandeep Mukhopadhyay	mukhopa6@purdue.edu

Links

Application: https://amaheshw.shinyapps.io/NBA_Team6/
GitHub Codes: https://github.com/am11917/NBA_Prediction_Analytics
Video Presentation: Team 6 - NBA Prediction & Analytics - YouTube

Overview

Basketball is one of the most popular sports in the US, and we have seen how sports have been revolutionized in the past 10 years, thanks to the use of analytics. This is especially evident in American sports leagues like the National Basketball Association (NBA), a league that generates billions of dollars in revenue each season. We analyzed the last 5 seasons' data of NBA matches, sourced from Kaggle, and performed a set of different analyses ranging from EDA, clustering analysis and predictive analytics, using this data. Such analyses helped answer questions like – "Top performing Teams", "effective offense", "Players with similar impact", generating a value for the end user.

Business Problem

- Provide a hub for coaches and management to perform team performance analysis and develop winning strategies
- Aid in player selection for fantasy league players and help them understand the true value of the players they select
- Enable sports bettors to make decisions based on predictive insights
- Develop a one-stop-shop application for all the users who seek to perform such analyses
- Prediction modeling to determine winner of head-to-head matches

Analytics Problem

We analyzed the performance of teams in each game, whether it was a home or away game, and used metrics such as free throws, number of field goals attempted, 3 pointers, assists, and rebounds that help define the offensive and defensive strengths of a team. These strengths impact the probability of winning future matches if the teams continue with the same strategies and effectiveness. To do the prediction, we spent a good amount of time exploring the team statistics and individual player performances. It was formulated as a classification problem where the target variable was the probability of the home team winning or losing the game. We predicted which team would win the game based on their performance in previous home and away games. We used AUC to compare different models. To tackle the assumption of constant performance, we used metrics like 10-gamemoving averages which would take into consideration the performance average in the last 10 games to incorporate changing team structures.

Data

We used NBA matches and players data from Kaggle and included the salary information from Basketball-Reference.com. We took the games data and used that for our prediction modeling and team performance analysis. We analyzed the players data to do EDA on individual performances. We found our target variable to be a binary vector which would predict if the home team would win or lose the game.

Methodology Selection

We used two approaches for our analysis:

1. Descriptive Analysis & Cluster Analysis: We used the last 5 seasons of players and team (games) data to analyze respective team and player performances
2. Predictive Analytics: We used 9 variables to forecast the winning team.
3. One of the key problems with the data was that all the variables are produced during the game, and no variables were present in the dataset that can be used without data wrangling and feature selection. We decided to take the 10-game moving average of all the variables and added new variables like overall winning percentage, winning percentage in road games and losing percentage while playing at home. This was done to have data which would predict chances of an upset occurring, something that frequently happens in NBA.

Model Building

Four different models were created on the modulated data to predict the probability of the home team winning the game. Logistic Regression, Neural Networks, XgBoost and Random Forest were trained on 70% of the dataset with 30% as the test dataset. Following are the ROC- AUC values for each of the models.

Model	Logistic Regression	Neural Network	XgBoost	Random Forest
ROC	0.82	0.82	0.82	0.81

All the candidate models have similar ROC values. Therefore, we have chosen logistic regression because it is a more interpretable model.

```
Coefficients:
 (Intercept)   FG_PCT_home  FG3_PCT_home     REB_home   FG_PCT_away  FG3_PCT_away     REB_away  home_loose_o    away_win_o    W_pct_away
      4.9983       -0.7974       -0.6541      -0.6383        0.9868        0.6222        1.0486       -5.7301       -5.3261       -0.2892

Degrees of Freedom: 16389 Total (i.e. Null);  16380 Residual
Null Deviance:      22160
Residual Deviance: 16390        AIC: 16410
```

From the summary we can see that *home_loose_o* and *away_win_o* are the most telling or indicative variables. This model was used for the prediction on the home team and away team based on their current statistics. Therefore, if these two teams were to matchup in a game now, the model will predict the chance of the home team winning the game.

Functionality

Our DSS is broken down into six different tabs, each with a different functionality. We have shown team logos, videos, and team statistics across seasons in interactive plots. We also created offensive and defensive ranking of players, performed clustering analysis on all players, and even have twitter reactions and twitter feeds in our application. We used packages like *TwitteR* and *shinyhelper*, but the most useful packages were *ggplot2* for the EDA analysis, *factoextra* for the clustering and *caret* for predictive analytics.

GUI Design and Functionality

- We kept the application aesthetically simple and pleasing with three main parts – Header, Sidebar, and Functional Plots
- Users can select team or player names from reactive dropdowns or type team/player names and all data is populated for the end user
- Dashboard is intuitive – anywhere we assumed the user would need guidance, we provided that with a help button to give the user additional details or insights

Conclusion

- Current Model: We concluded that out of all the prediction models, logistic regression is the best model based on the metric we used (AUC). Also, it is very easily interpretable for analysis.
- Future Scope: We would like to add features like average age, experience, and how frequently players are injured, which could help us garner an even more reliable prediction.

References

- UI Design Ideas – https://github.com/manyadadarya/IPL-EDA-and-Predictions
- Prediction Modelling – https://www.r-bloggers.com/2021/03/how-to-build-a-predictive-model-for-nba-games/
- Matches Data – https://www.kaggle.com/nathanlauga/nba-games
- Players Data - https://www.kaggle.com/justinas/nba-players-data
- Roster, Salary, Individual Player Data – Basketball Statistics and History | Basketball-Reference.com
- Clustering-https://towardsdatascience.com/which-nba-players-are-most-similar-machine-learning-provides-the-answers-r-project-b903f9b2fe1f