Machine Learning and Data Mining project: Basketball

Noemi Ippolito

Course of AA 2022-2023 - Data Science & Scientific Computing

1 Problem statement

The aim of this project was to settle (at least) one learning problem with a dataset containing information about NBA games from several recent seasons (from 2004 on), at the level of the single player contribution to each game. The chosen learning problem to settle was predicting the outcome of a match. The goal was to obtain a model able to predict win or loss (of the home team) based on the 3-game moving average statistics of the game.
The input variable is a single game. A detailed description of the input variable is provided later.
The output is a label $y \in \{1$ for "home team wins", 0 for "away team wins"$\}$; the problem is thus intended as a binary classification problem.

2 Assessment and performance indexes

The assessment of the proposed solutions will be done by splitting the dataset: 80% of it will be the training set and the remaining 20% will be the test set. Different models will be considered and their performance will be assessed using the following indexes (calculated for the standard threshold of 0.5);

- $Accuracy = (TP + TN)/(P + N)$
- $True\ Positive\ Rate = TP/P$
- $True\ Negative\ Rate = TN/N$

Finally, to get a general overview of the model's performance with different thresholds, the *ROC curve* and the AUC value will be computed. In order to understand how the models behave their performance indexes will be confronted with each other and with the one obtained by the Dummy classifier, that will be used as a comparison baseline.
Please do consider that the positive class is "home team wins" while the negative one is "away team wins".

3 Proposed solution

The proposed solution is to use different supervised learning techniques to build the model that will be used for the predictions. Two phases have been involved in the process.

3.1 Learning phase

The first necessary thing to do in order to build the model is to exactly define the "game", i.e. the input variable. After a phase of data cleaning the final features selected to describe each game (some of which have been feature engineered) were decided and are reported in Table 1. For each of the final features the reported value refers to the difference between the home team's feature and the away team's feature in the considered game.
The output variable was "HOME_TEAM_WINS", the binary variable whose values, 0 and 1, represent, respectively, whether the home team lost or won the game.
The different machine learning models chosen to build this tool are Logistic Regression, Random Forest Classifier, Naïve Bayes, Support Vector Machine with linear kernel and K Nearest Neighbors Classifier.

Feature	**Class**	**Description**
FG_PCT	num	Field goal shooting percentage
FT_PCT	num	Free throw shooting percentage
FG3_PCT	num	Three-point shooting percentage
TS_PCT	num	True shooting percentage
AST	int	Number of assists
REB	int	Number of rebounds
TUR	int	Number of turnovers
STL	int	Number of steals
BLK	int	Number of blocks
W_PCT	num	Winning percentage over the last 10 games
ELO_before	num	ELO rating before the game

Table 1: Features used to create the model.

3.2 Prediction phase

Once the models were trained they were ready to make predictions on the test set, which would be used for the evaluation of each model's performance.
In order to know who won a game, all the features of that game would have to be

inserted as inputs and the model would present an output $y \in \{0, 1\}$ indicating whether the home team won the game $\{1\}$ or not $\{0\}$; this will be the prediction that the model has made on the game.
It is worth noting that all the features of the actual game are unknown, as the model is trying to predict the outcome of a game that has not been played yet and for this reason the value of each feature is given by the average value of that feature in the last 3 games (which are known) played by the considered teams.

4 Experimental evaluation

4.1 Data cleaning and preparing

The dataset used for the final project was obtained by combining the dataset containg informations about the team's performances in different games with some of the features of a different dataset containing informations at the level of the single player's contribution to each game. This was done in order to get the values of some of the team's features that were not specified in the original dataset but were present in the dataset containing the player's information. This operation resulted in the collection of the team's numbers of turnovers, blocks, steals, attempted free throws and attempted field goals of each game. The last two features were not included in the final dataset but they were used to calculate each team's true shooting percentage, given by the following formula:

$$TS = \frac{0.5 * PTS}{FGA + 0.44 * FTA}$$

For each of the considered features were then calculated the values based on the 3-game moving average, which was chosen as a method of estimation for the in-game statistics. All the rows that contained teams that had not previously played at least 3 games were dropped from the database, as their number was small compared to the size of the database and their exclusion would not affect the analysis.
Two other important features that were added to the final dataset are the winning percentage of a team over their last 10 games and the ELO ratings of the team before and after the game. The first was calculated to give the model an understanding as to how the team had been performing recently. The second is a metric used to gauge team strength and performance and the formula that has been used to obtain it is the same presented in Nate Silver's and Reuben Fischer-Baum's article [**1**].
For creating the model, all non-numeric columns and columns linearly related to the outcome of the game (such as the ELO rating of the two teams after the game) were dropped from the dataset, and the other columns were combined to get the difference between the home team's perfomance and the away team's performance. The final dataset was then cleaned and all the rows containing NA values were dropped.
At the end of this process the dataset was made of 26500 observations of 12

variables; however, the dataset was unbalanced, with 15607 "home team wins" observations and 10893 "away team wins" observations. To fix this problem both oversampling of the minority class and undersampling of the majority class were used in the training set. With this method, the training data had 10729 "home team wins" games and 10471 "away team wins" games.

4.2 Procedure and results

The training set has been used to train the different models. The test set has been used to assess the models' performance; in particular, the four performance indexes described earlier were computed for each model. The results are reported in Table 2.

Model	*Accuracy*	TPR	TNR	AUC	*mean*
Logistic Regression	0.644	0.656	0.627	0.692	0.655
Random Forest	0.641	0.693	0.567	0.679	0.645
Naïve bayes	0.635	0.660	0.601	0.677	0.643
SVM	0.643	0.657	0.625	0.691	0.654
KNN	0.630	0.647	0.607	0.674	0.640
Dummy classifier	0.584	1.000	0.000	0.500	0.521

Table 2: Performance indexes of the different models.

4.3 Results and discussion

Different models were developed to estimate the probability of a team winning a game. The models were validated and they all worked well, especially compared to the dummy classifier. The model that seemed to work better was the Logistic Regression model, which obtained a mean accuracy of 0.655, meaning that it was able to correctly predict the outcome of a game 65.5% of the times.
Predicting the outcome of sport games is not easy as there is a level of unpredictability. In basketball, any team can win on any given day and it is extremely difficult to quantify all statistics of the game. A model that can perform and successfully predict the outcome of a game better than the dummy classifier is already an achievement. The best of the proposed models successfully predicted 65.5% of NBA games in the dataset so it is a successful model.
Some changes that might be made to the model in the future are to explore other variables that could improve its performance, such as the plus-minus, the home court advantage, fatigue (back-to-back games) and travel (distance traveled). Unfortunately these variables were not available in the considered dataset and there was no way to obtain them.

References

[1] Nate Silver, Reuben Fischer-Baum *How We Calculate NBA Elo Ratings*