University of Trieste

Introduction to Machine Learning Course

Academic Year 2024 – 2025

Predicting the winning team of an NBA match

Valeria De Stasio[1], Vittorio Lanzini[2], and Giovanni Lucarelli[3]

[1]problem statement, solution design, solution development, writing
[2]problem statement, solution design, solution development, writing
[3]problem statement, solution design, solution development, writing

1 Problem statement

In this project, we address the challenge of predicting whether the home team in a NBA match will win or not. This is a much discussed problem in Machine Learning (ML) literature [2, 5, 6], because human experts have not been able to give highly accurate predictions. For this reason, we decide to design, implement and assess a ML system for the problem. Since we have many examples at hand—almost twenty years of recorded NBA games, for a total of around 26000 observations—we use supervised learning techniques.

For each match, we consider several features that summarize the last five matches played by each of the two teams.

The features are those listed in table 1: all of them are collected for both the home team and away team, for a total of fourteen features. The Cartesian product of their domains is denoted by X. The response variable y indicates if the home team has won ($y = 1$, positive case) or it has lost ($y = 0$, negative case). Its domain is $Y = \{0, 1\}$, because in basketball there are no draws. The goal is to learn some models $m \in M$, where M denotes the set of all possible models, using the function f'_{learn} and then apply $f'_{predict}$ to $(\boldsymbol{x}, m)$ to predict

Variable name	Domain	Description
WC_last5	$\{0, 1, 2, 3, 4, 5\}$	Victories in last 5 games
D_PTS_last5[1]	$\mathbb{R}$	Avg. score difference in last 5 games
D_REB_last5[1]	$\mathbb{R}$	Avg. rebounds difference in last 5 games
D_AST_last5[1]	$\mathbb{R}$	Avg. assists difference in last 5 games
FT_PCT_last5	$[0, 1]$	Avg. free throw % in last 5 games
FG_PCT_last5	$[0, 1]$	Avg. field goal % in last 5 games
FG3_PCT_last5	$[0, 1]$	Avg. 3-point % in last 5 games

Table 1: The descriptions and the domains of the attributes.

the response variable. The two functions are defined as

$$f'_{learn} : P^*(X \times Y) \to M$$
$$f'_{predict} : X \times M \to Y.$$

where $P^*(X \times Y)$ denotes the set of all possible multisets of $X \times Y$.

2 Assessment and performance indexes

Since our task consists in a binary classification problem we choose to evaluate the effectiveness of the learned f'_{predict} using several classification metrics. We use the Area Under the Receiver Operating Characteristic Curve (ROC AUC), False Positive Rate (FPR), False Negative Rate (FNR), and Classification Accuracy (Acc). For a detailed description of these metrics and many others, see [7].

The ROC AUC is also used to assess and compare the different learning techniques. It is worth noting that in this binary classification problem, false positive error and false negative error are both equally undesirable, so we decided to mainly consider the value of ROC AUC for the assessment.

3 Proposed solution

After examining and pre-processing the dataset, a total of five supervised learning techniques are trained for the prediction: (i) dummy classifier, (ii) Random Forest (RF), (iii) Bernoulli Naive Bayes (BNB), (iv) Support Vector Machines (SVM), (v) k-Nearest Neighbors (kNN). We select the dummy classifier, which outputs the most frequent class, as the baseline due to its ability to establish a reliable lower bound for model performance evaluation. For each of the techniques we also apply hyperparameter tuning to select the hyperparameters that correspond to the best effectiveness, in particular the best ROC AUC.

[1]For simplicity, we defined the domain as the set of real numbers, even though the actual values are limited to a smaller interval in practice.

4 Experimental evaluation

4.1 Data retrieval

The data have been collected by Nathan Lauga and are publicly available on the Kaggle website: we used the dataset `games.csv`, which encompasses all NBA games from October 2003 to December 2022. Each row of the dataset represents a single game and its statistics. In the dataset only 0.5% of the observations contained missing values: we decided to drop them altogether, otherwise some of the ML techniques would not have worked. Some of the observations presented the same game identifier: we kept only one line chosen arbitrarily. Moreover, two columns were the duplicate of two others, but with different names, so we decided to preserve only a pair of them.

4.2 Procedure

We implemented and applied a pre-processing function in order to obtain the features defined in section 1 for each game. We dropped the rows that refer to games for which there were less than five games played by the home and away team, since it was not possible to compute the desired features. In addition to these data, for each game we also kept the date in which the game was played. This was not needed for the prediction, hence why it was not written in the problem statement, but only for the test/train division of the dataset. To preserve the time-dependent nature of our data, we opted for a *static* train/test split. Specifically, the training set includes all games played between 2003 and 2021, while the test set comprises all games from the 2022 season.

For each model, we defined a dictionary of potential hyperparameters, trained the models and selected the hyperparameter values that yielded the highest ROC AUC value on the static test set. Due to the temporal structure of the data, *k-fold Cross-Validation* was not feasible, as it would have resulted in using future data to predict past events (see [3] for details).

In order to assess the different learning techniques, considering the time problems mentioned above, we used the *Rolling-Origin Cross-Validation*[2] (ROCV). In this approach, each test set consists of observations from a single year, while the corresponding training set includes only data from years preceding those in the test set. This ensures that no future observations are used in constructing the forecast. This process provided a statistical basis for obtaining different values for the performance indexes in order to compare the different learning techniques.

4.3 Results and discussion

The results obtained for the individual trained models, evaluated over the last available year (namely 2022) can be seen in table 2 and figure 1 (on the left).

[2]Note that in literature the procedure is also known as Forward-Chaining Cross-Validation or rolling-origin-recalibration [1].

Model	ROC AUC	FPR	FNR	Acc
baseline				
dummy	0.5	1.0	0.0	0.566
RF	**0.609**	0.686	0.160	**0.612**
BNB	0.590	**0.636**	0.252	0.581
SVM	**0.609**	0.788	**0.103**	0.600
kNN	0.605	0.683	0.165	0.611

Table 2: Evaluation metrics for 2022 prediction. Best results are shown in bold.

All models provided only marginal improvements over the baseline (dummy classifier), both in terms of AUC and Acc. RF and SVM achieved the highest AUC scores (~ 0.61), but had limited discriminatory power. RF may be considered the best in terms of Acc for the predictions of the last year, but just slightly over the kNN upshot. Nevertheless, the BNB and the SVM achieved the lowest value of respectively FPR and FNR.

The boxplots[3] in the figure 1 (on the right) represent the variability of the ROC AUC scores across the multiple runs obtained with the ROCV for each model. Looking at the plot as a whole, no single model stood out as significantly better than the others. Nonetheless, the median ROC AUC for RF and kNN was slightly higher (~ 0.63) compared to NB and SVM (~ 0.61), indicating that these models performed better on average. However, the wide InterQuartile Range (IQR) indicated that the performance of RF was more variable compared to the other models.

Another interesting detail that emerged from the analysis was the sharp drop in all effectiveness indexes in 2020, *e.g.*, ROC AUC ~ 0.56 for all models, likely due to the COVID-19 virus pandemic that reduced the number of games played.

5 Conclusions

The model that gave the best results in terms of effectiveness for predicting if a home team will win or not an upcoming NBA match was the Random Forest, even if only by a small margin with respect to the other learned models. In general, the results indicated that none of the models showed strong predictive performance, with only marginal improvements over the baseline and limited discriminatory power. This may be attributed to the limited number of features or the fact that the chosen features (averaged over the last 5 games) might not have fully captured the complexity of NBA match outcomes.

Further improvements could be achieved by considering more complex models or by considering additional features, *e.g.*, statistics regarding each player of the two competing teams. It must also be noted that good results have been obtained in literature by applying Artificial Neural Network to the problem [4,6].

[3]The boxplot for the dummy was not reported since it was constantly equal to 0.5.

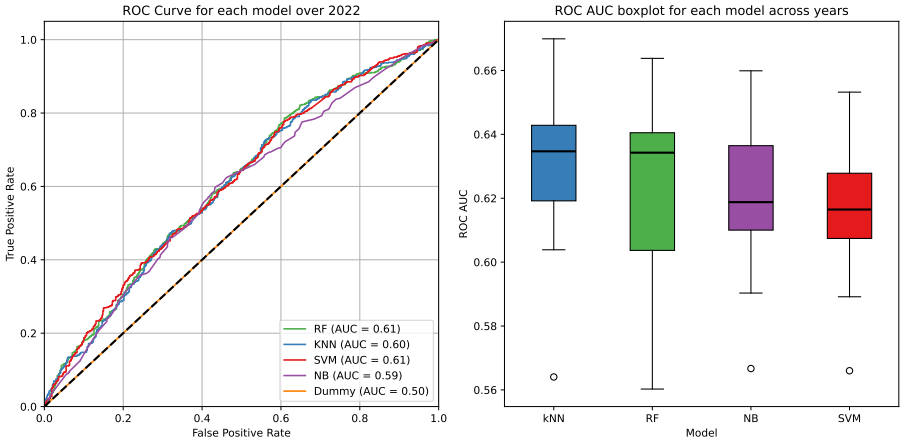

Figure 1: Left: ROC AUC Curves of the models tested over the 2022. Right: ROC AUC boxplots obtained through ROCV.

References

[1] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[2] Chenjie Cao. *Sports data mining technology used in basketball outcome prediction*. PhD thesis, Technological University Dublin, 2012.

[3] RJ Hyndman. *Forecasting: principles and practice*. OTexts, 2018.

[4] Bernard Loeffelholz, Earl Bednar, and Kenneth W Bauer. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[5] Dragan Miljković, Ljubiša Gajić, Aleksandar Kovačević, and Zora Konjović. The use of data mining for basketball matches outcomes prediction. In *IEEE 8th international symposium on intelligent systems and informatics*, pages 309–312. IEEE, 2010.

[6] Fadi Thabtah, Li Zhang, and Neda Abdelhamid. Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116, 2019.

[7] Željko Đ. Vujović. Classification model evaluation metrics. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 12, 2021.