Machine Learning and Data Mining project: Basketball

Baurice Nafack

Course of AY 2021-2022 - Introduction to Machine Learning and Evolutionary Robotics, Prof Eric Medvet

1 Problem statement

The National Basketball Association (NBA) league, founded in 1946, is the world's most popular basketball league. Since gambling companies have financial assets at stake, fans and potential bidders are all interested in estimating the odds of a game in advance. Many participants place their chances subjectively based on their personal team preferences without any scientific basis, resulting in extremely poor predictions. Thus, accurately predicting the outcome of NBA games based on past statistics of sports competition (games, teams, players, etc) is a difficult problem that both researchers and the general public are interested in tackling [1]. Any team has a good chance of winning each game because of the high degree of uncertainty involved [4].

NBA league consists of 30 teams divided into the Eastern and Western conferences. Except in 2020, when the season was cut short by the Covid pandemic, it has always had 82 regular-season games, each team will play 41 away games, and 41 home games. The top eight teams from each conference (Eastern and Western) are chosen to compete for the championship during the playoffs (The playoffs adopted a new format in the 2020-21 season, and are a little more competitive than before, but the changes are minimal). The rankings are determined by the number of teams winning games in the regular season. The teams then compete against each other, with the first-place team facing the eighth-place team, the second-place team facing the seventh place, etc. Each game will be a best-of-seven series, with teams rotating between home and away [5]. The project aims to predict the winner of a basketball game, the ranking of the NBA Playoffs, the winner of each conference, as well as the winner of the NBA. Predicting basketball games makes it easier for bettors to make informed decisions, as it may increase their confidence in betting. It is also interesting to see which factors are closely related to team success. Teams can gain an advantage in winning games through this type of analysis.

2 Proposed solution

2.1 Logistic Regression Algorithm

Logistic regression is a type of supervised learning method that predicts class membership. A model is trained to predict the probability (p) of new data falling into each class. Typically, new data are assigned to the class to which they are most likely to belong. Data are converted into log-odds (logits), which are then converted into odds and probabilities of belonging to the ”positive” class [3]. Cases are assigned to the positive class if their probability exceeds a predetermined threshold (0.5 by default).

2.2 Decision Trees Algorithm

The basic idea behind tree-based classification algorithms is that they learn which questions identify cases in different classes. Each case will be sent down the left or right branch based on which criteria it meets. Based on metrics like information gain, it performs automatic feature selection for decision tree induction. Multicollinearity is easily handled, even of higher orders. After learning, the model can be represented graphically as a tree.

2.3 SVM Algorithm

The SVM algorithm identifies a linear hyperplane that separates classes. The dimensions of hyperplanes are less than those of variables in a dataset. A hyperplane is a straight line in a two-dimensional space and a surface in a three-dimensional space. They are hard to visualize in a four-dimensional or higher-dimensional feature space, but their concept is the same: they are surfaces that cut through feature space [3].

3 Experimental evaluation

3.1 Data

The dataset we used is from Kaggle's datasets [2] include five data frames . *games.csv*: containing all games from the 2004 season to the last update with the date, teams, and some details like a number of points, etc. *games_details.csv*: details of games dataset, all statistics of players for a given game. *players.csv*: players details (name). *ranking.csv*: ranking of NBA given a day (split into west and east on conference column. *teams.csv*: all teams of NBA

We used the *games.csv* data frame for our ML prediction. It contains as variables: $Games\_date$; PTS : Percent of Team's Points. AST: Percent of Team's Assists. FT_PCT : Free Throw Percentage. REB: Rebounds . FG_PCT : field goal percentage . $FG3\_PCT$: 3 Point Field Goal Percentage. $HOME\_TEAMS\_win$ information: win or lose. $SEASON$ of the games.

Home and Visitor Teams names

This dataset contains 25024 rows, 44 duplicated games, and 99 missing values for the following variables: *PTS_home*, *FG_PCT_home*, FT_PCT_home, FG3_PCT_home, REB_home, PTS_away, FG_PCT_away, FT_PCT_away,

FG3_PCT_home, AST_away, REB_away.

3.2 Experimental Procedure

3.2.1 Data cleaning (Remove duplicated games and Handle missing values)

Duplicates games that can affect ranking predictions were removed. Two options are available for handling missing values:

- Simply exclude cases with missing data from the analysis, this will end up with droping games which will be a problem when predicted the post-playoff ranking.
- Apply an imputation mechanism to fill in the gaps

Because each game is important for ranking prediction to select a winning team for the season, deleting missing values will have an effect on the final regular-season prediction ranking. Therefore, locating the correct missing value is critical. This process was done using the imputation method from the R caret library. I then split the dataset into two-part, one to predict the winner of a game and the second for the playoff ranking predictions starting from 2016 to 2020. Using the caret R package, the first set was divided into 85% training sets and 15% testing sets.

3.2.2 Model Implementation

All methods in this study were implemented in R. The mlr package was used to create a task, learner (specifying ”classif.logreg” for logistic regression, ”classif.rpart” for the decision tree, and ”classif.svm” for SVM), and model. I then train and cross-validate the Logistic Regression model to predict how it will perform. I first define a resampling method with makeResampleDesc(), then apply stratified, 5-fold cross-validation to the wrapped learner 50 times. The cross-validation is then performed using resample().

I tuned the algorithm's hyperparameters for the SVM using the following hyperparameter space: kernel (values = kernels), degree (lower = 1, upper = 3), cost (lower = 0.1, upper = 10), and gamma (lower = 0.1, upper = 10); for decision tree, minsplit (lower = 5, upper = 20), minbucket (lower = 3, upper = 10), cp (lower = 0.01, upper = 0.1), maxdepth (lower = 3, upper = 10). Because the hyperparameter space is so large, I chose a random search over a grid search with 5 iterations. I also define a cross-validation strategy for tuning that uses 5-fold. To accelerate things by using parallelMap and the parallel R library, I begin parallelization by calling parallelStartSocket() and setting the number of CPUs to the number of CPUs I have available (16). The tuneParams() function is then used to begin the tuning process. When it's finished, I stop parallelization and use the setHyperPars() function to create a learner with the tuned hyperparameters, after which I train a model. The tree is then cross-validated with hyperparameter tuning.

3.2.3 Feature Selection and importance

Using Buruta R package, 11 iterations were performed to ensure all of our features are considered important. PTS home and away were identified as highly important for winning predictions based on the plot. To win a game, betters and teams should therefore focus on those variables.

3.3 Results and discussion

The team playing at home has a clear advantage over its opponents. Home win target was used to construct our model based on that fact.

We cross-validate our models to get the following accuracy, 0.92 for SVM,0.97 for the decision tree, and 0.99 for logistic regression to correctly classify games as win or loss. The respective runtimes were 2522.04, 47.84, and 84.77.

The 95% confidence interval in which we expect the prediction accuracy to decline appears high in table 1. Although our data were imbalanced, we still have a high Kappa value. With one of our models, NBA point spread predictions for live betting could have a high probability of giving the better an advantage. Using real-time data, he could simulate his own for prediction purposes. All provided models could be used for all the season game.

For post-playoff predictions, 2 is the inter-range of winning games for each team. The true value is predicted by the model with 75 percent of teams having 3 more or less winning games difference. The first quartile value is one, the second quartile value is three, and the upper adjacent value is six. Therefore, this model could be used for real-time playoff ranking which would motivate teams to win the next game. Post playoff ranking prediction holds to the real ranking with slight changes for the number of winning games.

Model	prediction	Acc	95 % CI	Sensitivity	Specificity	Kappa
SVM	Test data	0.98	(0.9746 ; 0.9854)	0.984	0.9781	0.96
	Season >= 2016	0.979	(0.9753 ; 0.9823)	0.983	0.976	0.957
	feature selection	0.9996	(0.998 ; 1)	0.999	1	0.999
Decision Tree	Test data	0.9596	(0.9515 ; 0.9667)	0.9362	0.9762	0.916
	Season >= 2016	0.9283	(0.9219 ; 0.9343)	0.8773	0.9670	0.8524
logistic regression	Test data/feature	0.9996	(0.998 ; 1)	0.9991	1	0.9992
	Season >= 2016	1	(0.9995 ; 1)	1	1	1

Table 1: Statistical of model prediction on single basketball.

References

[1] Ge Cheng, Zhenyu Zhang, Moses Ntanda Kyebambe, and Nasser Kimbugwe. Predicting the outcome of nba playoffs based on the maximum entropy principle. *Entropy*, 18(12), 2016.

[2] Nathan Lauga. NBA games data. `https://www.kaggle.com/nathanlauga/nba-games`, 2019. [Online; accessed 19-01-2022].

[3] Hefin I. Rhys. *Machine Learning with R, the tidyverse, and mlr*, volume 7. Manning Publications, 2020.

[4] 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, March 2019.

[5] Richard Tovar. NBA 2021-22 playoff format: How will the basketball postseason work? `https://bolavip.com/en/nba/nba-2021-22-playoff-format-how-will-the-basketball-postseason-work-20211018-0003.html`, 2021. [Online; accessed 19-01-2022].