Machine Learning and Data Mining project: Basketball project

Anna Pederzani[1] and Thomas Verardo[2]

[1,2] problem statement, solution design, solution development, data gathering, writing

Course of AA 2021-2022 - Data Science and Scientific Computing

1 Goal of the analysis

Basketball is one of the most popular sports in the USA. Its competitions are organised under the National Basketball Association (NBA) which is one of the most followed and viewed professional leagues in the world. The NBA generates extremely high revenues and it is the third top sports league in terms of turnover, generating revenues of about $8 billion per season in the last few years.

Nike is the official merchandising brand of the NBA producing the jerseys of all players. Furthermore, many popular players are sponsored by Nike. The brand is also willing to sponsor the Most Valuable Player (MVP) of the NBA, the award given to the best performing player of the regular season, and it needs to make an agreement with the winner before he will be elected in order to save money. Thus, Nike has to predict the next winner and therefore it decided to hire a team of experts for this task. Moreover, the brand has decided to wait until the end of the regular season before making the analysis in order to obtain a more confident prediction.

The goal of our analysis is to predict who will be the MVP of 2021 using information on players and matches of the entire regular season.

2 Data collection

Provided that the *Kaggle.com*'s datasets on the NBA do not contain crucial information needed for the MVP prediction, it has been decided to directly use the nba_api from which the *Kaggle.com*'s data were extracted. This API client allows access to all the NBA's data included in the official *stats.nba.com* website. To achieve the goal of our analysis, different information from different datasets has been used. The four datasets that have been selected to perform the study are:

- **commonallplayers** which contains the list of all NBA players and the period when they have played;
- **playercareerstats** which records all the data (such as points scored, number of matches, minutes played, field goal made and free throw made) of all players for each season;
- **playerawards** which lists the awards that each player has won in each season;
- **teamyearbyyearstats** which contains, for each regular season, both the Eastern and the Western conference standings.

3 Data description and data preparation

In order to use the data presents in the four datasets, the different information needs to be joined together. Before the merge, data cleaning and data filtering have occurred.

Our study has been performed using the data of each regular season from the NBA Season 2003-2004 to 2020-2021. Since the aim of our analysis is to predict the 2020-2021 MVP, we disregarded the *playerawards* dataset for this season. The period of our analysis has been selected due to the fact that *Kaggle.com*'s datasets describe only the NBA games from 2004 to 2021 and since the methods of awarding of the MVP have changed over the years; taking into consideration older data would have caused our study to be less accurate. Thus the four databases have been filtered to obtain all the data of all the players who had played at least one game since 2003-2004.

Afterwards, the *playerawards* dataset has been reduced by keeping only the observations in which the award won was the MVP and then the merge between the four datasets has been performed. Moreover, a binary variable (MVP) was created to which has been assigned value 1 if the player won the MVP award in the given year and the value 0 otherwise.

The result of this process is a dataset containing all the necessary attributes for each player in each season under study. The final dataset contains 9111 observations made of 22 features (such as games played, field goals made, rebounds, assists, steals, rank of the team), the MVP label and three identifiers (player_id, team_id and season_id) that together uniquely identify each row of the database.

Considering the aim of our analysis, the response variable will be the binary variable indicating if the player has won or not the MVP. The type of problem to be solved is a binary classification problem.

The training set has been set to 70% of the data, while the remaining 30% has been used as test set to check the quality of the prediction (using the function of the Python library *sklearn train_test_split()* that randomly split the dataset). The proportion has been chosen due to the relatively small dataset, as suggested by many studies. [1] [3]

Since different seasons are contained in the dataset and since the features have different scales between seasons, we have decided to standardise both training set and test set per season with *StandardScaler()*.

Moreover, a feature selection analysis was performed through the calculation of the correlation (using the function *corr()*) between the independent variables. This decision has been taken due to the assumption of absence of multicollinearity in the logistic regression [6]. For every pair of features, the correlation has been calculated and one of the two variables was removed if the correlation was greater than 0.9. The removed features are: *'Field goal attempted'*, *'3 points field goal attempted'*, *'Tree throws attempted'*, *'Points'* and *'Rebounds'*.

Due to the nature of the data, our dataset is very unbalanced with respect to the dependent variable chosen. Indeed, each year there is only one winner of the MVP award. The percentage of MVP=1 is 0.18% of the observations. The unbalanceness of the dataset may produce problems due to the fact that the classification algorithms are constructed for balanced data. Since a modification of the dataset is needed, an oversampling technique has been implemented on the training set. The technique chosen is the SMOTE [5][2]. It synthesises new examples from the minority class. After this modification the new percentage of players with the MVP =1 is 33.33%.

4 Proposed solution

To predict our response variable, three binary classifiers have been implemented using the *sklearn* library: a logistic regression using *LogisticRegression()*, a Support Vector Machine using *SVC()* and a Random Forest using *RandomForestClassifier()*. To find the best parameters for each classifier, leave-one-out cross-validations have been used as auto-tuning through the function *RandomizedSearchCV()*. To evaluate the performance of the cross-validated models *scoring='balanced_accuracy'* has been selected. The parameters that have been auto-tuned for the three models are:

- **Logistic regression**: the inverse of regularisation strength and the algorithm used to solve the optimization problem of the regression. As regularisation term we have chosen the L2 term since the feature selection was already part of the data preparation;
- **Support vector machine**: the inverse of regularisation strength and the type of kernel function;
- **Random forest**: the criterion to measure the quality of a split, the maximum depth of the trees and the number of trees in the forest. As number of independent variables selected in the learning phase by each tree, we decided to use the square root of the total number of features since we are facing a classification problem.

Once we have found the parameters for each classifier, the three models have been trained. In the learning phase we have set, in all the three models, *class_weight="balanced"* that automatically adjust weights inversely proportional to class frequencies in the training set, due to the fact that the database is still unbalanced.

5 Evaluation of the performances

The test set has been used to evaluate the performances of the three classifiers and thus to understand which one is the most appropriate to predict the MVP. Once the predictions on the test set were performed, the balanced accuracy of the three different models have been calculated. This measurement of quality has been selected since it is suitable and appropriate for unbalanced datasets. Indeed, contrary to the conventional accuracy that counts the correct classifications out of the total number of predictions, the balanced accuracy is the average between sensitivity and specificity [4]. The balanced accuracy for the three models are:

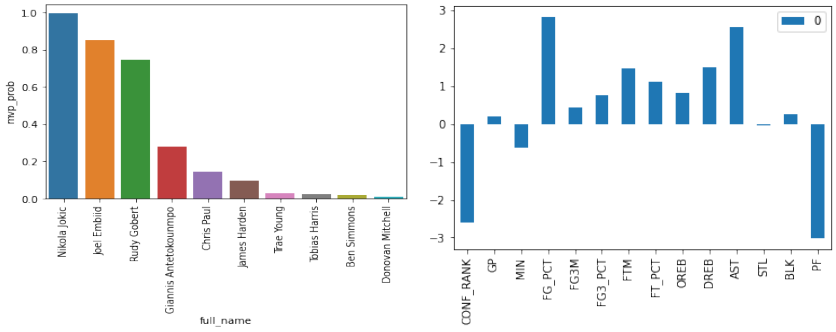
- Logistic regression: 0.999, with sensitivity = 1 and specificity = 0.997;
- Support vector machine: 0.9, with sensitivity = 0.8 and specificity = 0.998;
- Random forest: 0.799, with sensitivity = 0.6 and specificity = 0.999.

The logistic regression is thus the most appropriate model for our study, since it has the highest balanced accuracy. By looking at the confusion matrices of the three classifiers, we reach the same conclusion. Indeed, the logistic regression has very few false positives (players that have been classified as winners but that in reality they haven't won) and it is the only classifier that has zero false negatives (players that have been classified as non winners but that in reality they have won the MVP).

6 Prediction of season 2020-2021

To complete our study, we have used the trained logistic regression to predict the MVP of season 2020-2021. The first graph shows the ten players who have the highest probability of winning the award. Nicola Jokic is the favourite with 99.5% probability of winning (he is the actual MVP of season 2020-2021).

The second graph shows the different impact that the features had in the prediction of the response variable. The features with the highest influence are: *'Final confederation rank'*, *'Field goal percentage'*, *'Assists'* and *'Personal fouls'*.



References

[1] Olga Kosheleva Afshin Gholamy, Vladik Kreinovich. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation.* University of Texas at El Paso, 2018.

[2] Jennifer Shang Gu Mingyun Huang Yuanyue Gong Bing Guo Haixiang, Li Yijing. *Learning from class-imbalanced data: Review of methods and applications.* 2017.

[3] Isabelle Guyon. *A scaling law for the validation-set training-set size ratio.* ATT Bell Laboratories, Berkeley, California.

[4] Dr.Taklit Akrouf Alitouche Mohamed Bekkar, Dr.Hassiba Kheliouane Djemaa. *Evaluation Measures for Models Assessment over Imbalanced Data Sets.* 2013.

[5] Lawrence O. Hall W. Philip Kegelmeyer Nitesh V. Chawla, Kevin W. Bowyer. *SMOTE: Synthetic Minority Over-sampling Technique.* Morgan Kaufmann, 2002.

[6] Deanna Schreiber-Gregory. *Logistic and Linear Regression Assumptions: Violation Recognition and Control.* Henry M Jackson Foundation.