Determination of Winning Probabilities for Both Teams in NBA Basketball Games Project Name: Hoops Gonna Win

Thomas Vanderzande
Email: thomas.vanderzande@gmail.com

The literature has studied numerous aspects of the NBA using Machine Learning techniques but very few studies trying to determine the winning probability of both teams for NBA games have been found. Some algorithms giving those probabilities have been deployed either freely (ESPN's Basketball Power Index) or as a paying service (BetQL). But very few information is shared on their methodology giving the impression to see results coming from a black box. The purpose of this study is to propose a tool which displays winning probabilities of both teams but also the statistics used to make the computation. A series of base classifiers calibrated with Platt Scaling has been developed by optimizing the associated hyperparameters, which has been then used to screen Ensemble classifiers in order to improve the precision of the winning probabilities output. The metrics of the developed models are on-par with the literature with an accuracy close to 70%. Most of them were displaying superior model metrics and results in a betting simulation compared to ESPN's Basketball Power Index. The most promising model is being deployed into production in a Web application.

Nomenclature

3P 3-Point Goals
AST Assists
ATR Assist-to-Turnover Ratio
DRB Defensive Rebounds
FG Field Goals (both 2-point and 3-point goals)
FGA Field Goals Attempts
FT Free Throws
ORB Offensive Rebounds
TOV Turnovers
WS Win Shares
WSI Win Shares of Injured Players
WSR Win Shares of Remaining Players
WST Win Shares of the Whole Team

1 Introduction

The NBA (National Basketball Association) has been extensively studied and analyzed through Machine Learning models, whether it be to predict the winner of single games [1, 2] or predict which players will make it to the All-Stars Game [3]. These are only examples of the plethora of applications of Machine Learning linked to the NBA. One of the applications where very little to no literature has been found concerns the computation of winning probabilities of both teams for single games. Some models have been deployed and are accessible either freely (ESPN Basketball Power Index) or as a paying service (BetQL for example). But very little information is given regarding their methodology.

Most of classification models can output these probabilities natively, but some of them tend to provide skewed values that does not precisely represent the reality. For example, boosted trees generally output values far from the extremes (0 and 100%) while the Naives Bayes classifier, on the contrary, will excessively go towards these extremes. Support Vector Machine is an example of classifier that can not calculate probability natively. Therefore, these models need to go through a calibration process in order to get relevant results. Two of the most well-known calibration methods are Platt Scaling [4] and Isotonic Regression [5]. While Isotonic Regression is seen as more powerful, being able to generalize to most classification algorithms, it also tends to over-fit if the calibration dataset is not large enough [6].

Another possibility to maximize the performance on a classification problem consists in taking several classifiers, and combine their prediction to output a single new prediction. The process is called Ensemble learning and the idea is that an Ensemble classifier can out-perform the classifiers it contains taken apart, by combining their respective strengths. It has found numerous applications with, for example, biology [7] or even for the calibration of probabilities computed [8].

The present study aims at computing well-calibrated winning probabilities for NBA games. A series of single classifiers and Ensemble classifiers has been made to identify the most promising combination for this problem both in terms of Brier Score, Area Under ROC Curve but also using a betting simulation in a more ludic way. Most of the developed models out-performed a reference model that is already publicly accessible (ESPN Basketball Power Index). The most promising model is being deployed in a Web ap-

plication where users can check the analysis of the upcoming games, but also update the inactive players lists according to the last news, and check the updated probabilities. This work can also serve bookmakers to optimize their algorithms.

2 Methodology

For the following sections, the web-scraping, data analysis and model training/validation have been performed with BeautifulSoup, Pandas and Scikit-Learn libraries, respectively. Correlation and normality tests have been made with SciPy.

2.1 Data

Data of interest have been extracted on Basketball Reference. When the script used was making requests on multiple URLs, a sleeper time of one second between each request has been set up to mitigate the risk of adversely impact site performance or access (section 5.i of Sports Reference Terms of Use).

A first dataset of 2424 games from seasons 2015-2016 to 2018-2019 is used for models training and validation, using cross-validation.

A second dataset of 735 games from 2021 (season 2020-2021 and start of season 2021-2022) is used for comparison of the developed models with ESPN's Basketball Power Index.

Games have been picked with both teams having played at least 20 games in order to get significant teams statistics. For the first dataset, teams statistics, as of the date when the analyzed game has been played, are used for feature computation. Inactive players are extracted from games report and teams schedule from the teams statistics page.

For the second dataset, every game has been analyzed (team statistics, injury reports, team schedules, ESPN predictions extraction, feature computation and probability computation) several hours before the start of the game due to jet lag.

2.2 Feature Selection

2.2.1 Shooting

The Effective Field Goal Percentage is used to evaluate the ability of a team or a player to shoot efficiently. The formula proposed by Oliver [9] (Equation 1) takes into account the difference in reward between 3-point field goals and 2-point field goals. This is particularly important in modern basketball where 3-point field goals are taking an increasingly significant part of the game.

$$eFG = \frac{FG + 0.5 * 3P}{FGA} \tag{1}$$

2.2.2 Rebounds

The formulas proposed by Oliver [9] allows to evaluate the ability of teams to get rebounds as much as possible on both sides of the field (Equations 2 and 3). Offensive rebounds are important to keep the possession alive after a missed shot for example, and eventually increases the likelihood of scoring points.

$$OffRBD = \frac{ORB}{ORB + Opp.DRB} \tag{2}$$

Defensive rebounds are important to shut the opponent possession off and prevent them from scoring points.

$$DefRBD = \frac{DRB}{DRB + Opp.ORB} \tag{3}$$

2.2.3 Turnovers

The proposed formula by Oliver [9] (Equation 4) looks to define the ability of a team to avoid ball losses (turnovers).

$$TOVRate = \frac{TOV}{TOV + FGA + 0.44 * FTA} \tag{4}$$

2.2.4 Free Throws

This parameter is a measurement of how often a team is able to get free throw attempts and how often these attempts are successful. According to Oliver [9]:

$$FTH = \frac{FT}{FGA} \tag{5}$$

2.2.5 Injuries

Injuries are a key factor in the outcome of a basketball game and should be looked at when giving the probability of winning for both teams. Basketball Reference came up with a metric called Win Shares and is derived from Bill James' system developed for baseball. This metric tries to measure the contribution of a player in terms of wins. Intuitively, best players will have larger Win Shares and players with lower impact will have lower Win Shares. The full details of the calculation can be read here.

Therefore, to evaluate the impact of injuries in a given team at a given time, it has been proposed to calculate the sum of Win Shares from injuries players (WSI) and the sum of Win Shares of the entire team (WST), and arrange those numbers into the following equation :

$$WSR = 1 - \frac{WSI}{WST} \tag{6}$$

This allows to measure the contribution of the remaining players. The higher the number, the healthier the team which contributes positively to their chance of winning their upcoming game.

2.2.6 Calendar Effect

The calendar is also an important factor in the outcome of a basketball game. In fact, teams do not have a perfectly

regular schedule and can sometimes play 2 days in a row or even 3 games in a 4-day span. This ultimately impacts their chances as they can lack freshness in these scenarios.

The method tested consists in looking at the 5 preceding days from when the analyzed game is played. A score is initialized to zero. For each day without playing during these 5 days, the score is increased by 0.2. If the analyzed game is part of back-to-back games, a penalty of 0.1 is applied. For the specific (and physically most demanding) scenario where it is the third game in a 4-day span, the score is directly set to 0.25 (or 0.15 if the back-to-back penalty is applicable).

This way, the team with more rest will have a better score.

2.2.7 Parameters Computation

To compute the parameter dedicated to the rebound sector, a team score is defined for a given team which regroups both the offensive (Off RBD) and defensive rebounds (Def RBD). As the best teams are expected to get high values for both ends of the field, the global indicator is calculated as follows :

$$RBD(Team) = \frac{OffRBD}{1 - DefRBD} \tag{7}$$

For the parameters linked to the shooting, turnover and free throw sectors, for a given team, formulas described above are used to calculate the performance of the team against their opponents and the performance of their opponents against them.

For example, for the shooting sector, we calculate the Effective Field Goal Percentage of a given team against their opponents :

$$eFG(Team/Opp) = \frac{FG + 0,5.3P}{FGA} \tag{8}$$

We then compute the Effective Field Goal Percentage of their opponents against them.

$$eFG(Opp/Team) = \frac{Opp.FG + 0,5.Opp.3P}{Opp.FGA} \tag{9}$$

The global indicator for the team becomes:

$$eFG(Team) = \frac{eFG(Team/Opp)}{eFG(Opp/Team)} \tag{10}$$

A team score greater than 1 indicates that the team tends to dominate their opponents in the corresponding sector. On the contrary, a team score lower than 1 indicates that the team tends to be out-played in the corresponding sector.

For a given game, the team score is calculated for both the home team and away team and then is computed the ratio of these two values to get the parameter that will be used for model training/validation. For example, for the shooting sector :

$$eFG(Game) = \frac{eFG(Team = HomeTeam)}{eFG(Team = AwayTeam)} \tag{11}$$

A parameter greater than 1 will indicate that the home team is more likely to dominate the away team in the corresponding sector. On the contrary, a parameter lower than 1 indicates that the away team is more likely to out-play the home team in the corresponding sector. For the turnover sector, as a high value is detrimental for any team, the inverse ratio is computed.

For the parameters linked to injuries and calendar effect, the ratio of the home team and the away team related scores is calculated to get the parameter that will be used for model training and validation. For example, for injuries :

$$INJ = \frac{WSR(Team = HomeTeam)}{WSR(Team = AwayTeam)} \tag{12}$$

Using this parameter computation method, the different models will then analyze each game with a total of six parameters (shooting, rebounds, turnovers, free throws, injuries and calendar effect). A Spearman's correlation has been run for each possible pair of parameters. There was no to very low correlation found for each pair ($-0.21 < r_s < 0.21$, n = 2424, $p < 0.001$), meaning there is no redundancy among the parameters.

2.3 Model Screening

Models selected for the screening are the following :

- An Ensemble of Decision Trees selected between :
  - Random Forest (RND)
  - Gradient Boost Classifier (GBC)
  - AdaBoost Classifier (ABC)
- Logistic Regression (LR)
- Support Vector Classifier (SVC)
- Naive Bayes Classifier (NVB)
- Multi-Layer Perceptron (MLP)

Each model has been trained and validated on the first dataset using a 3-fold and 5-fold cross-validation. Brier Score Loss is used for parameters selection, as it is linked to the difference between the predicted probability and the actual outcome. The mean value obtained on the test part of all splits of the cross-validation is considered. All models have been tested with and without Platt Scaling which has proved to improve the probability outputs of several types of classifier when they give distorted probability distribution [6]. First tests with Isotonic Regression have not been conclusive. The models were outputting extreme values (0% and 100%) with Isotonic Regression, which is not realistic.

All models have been compared on the second dataset with ESPN's Basketball Power Index predictions using Brier

Score Loss and Area Under ROC Curve as metrics. A screening of Ensemble classifiers with the selected classifiers has also been made as Ensemble can improve the accuracy of the probabilities output by classifiers [8].

2.4 Betting Simulation

All models have also been tested in a betting simulation. The hypothesis is that optimal models should be able to make profit by using bookmaker's errors. Games from the second dataset are used as they have not been used to train the models, allowing an apple-to-apple comparison with the Basketball Power Index. As games of the second dataset have been analyzed before they actually happen, it also reproduces more accurately the real-life conditions of betting.

Odds for each team for a given game has been picked in an online bookmaker right after the analysis by the tested models. The odds picked concern the final result of the game (overtime included). This type of bet is especially popular in Europe and is directly linked to the probability of each team of winning a game. For example, a team having an odd of 2 means that they should be winning more than 50% of times so betting on them is profitable. Therefore, this is the ideal way to test the probabilities output by the models.

For a given game, a decision is made (betting or not) by looking at the probability output for each team by the different models. The minimal acceptable odd for a given team is defined as the inverse of the winning probability computed by the model. If the odd proposed by the bookmaker for a team is higher than its minimal acceptable odd, the model places a bet on them (with a fictional wager of 1 "unit").

All models have also been compared with sub-optimal betting strategies which do not require a thorough analysis of games. The picked strategies are the following :

- Always bet on the Home Team
- Always bet on the Away Team
- Always bet on the Favorite Team
- Always bet on the Underdog Team
- Pure random pick

The hypothesis is that the developed models should be making a better profit than these strategies, confirming the value of analyzing games.

3 Results

3.1 Model Training

The overall performance of the tested models on the testing part of the first dataset is shown in table below.

K-Fold	Brier Score Loss	AUC
3	$0,201 \pm 0,001$	$0,742 \pm 0,002$
5	$0,200 \pm 0,002$	$0,744 \pm 0,002$

We observe that models trained and validated with a 3-Fold cross-validation have a higher mean value of Brier Score Loss but also a lower standard deviation. This is thought to be due to the K value used for cross-validation as a lower K value implies a smaller training dataset (higher bias) and a larger testing dataset (lower variance). All models have also a Brier Score Loss lower than 0.25, meaning that their classification skills are better than a theoretical model which would simply output 50%/50% for every game.

The overall performance of the tested models on the second dataset is shown in the table below.

Name	Brier Score Loss	AUC
Reference		
ESPN BBPI	0.226	0.672
Tested Models		
All Models	0.222 ± 0.002	0.702 ± 0.005
Best Model NVB-MLP CV3	0.218	0.706

We can observe a decrease in the performance compared to the first dataset, as the prediction was made several hours prior to the start of the games, which can increase the bias. For example, some players can be declared injured or returning between the analysis and the actual start of the game.

To get a more practical view of the performance, the accuracy (considering a 50% threshold) has been plotted against the Brier Score Loss (Figure 1)

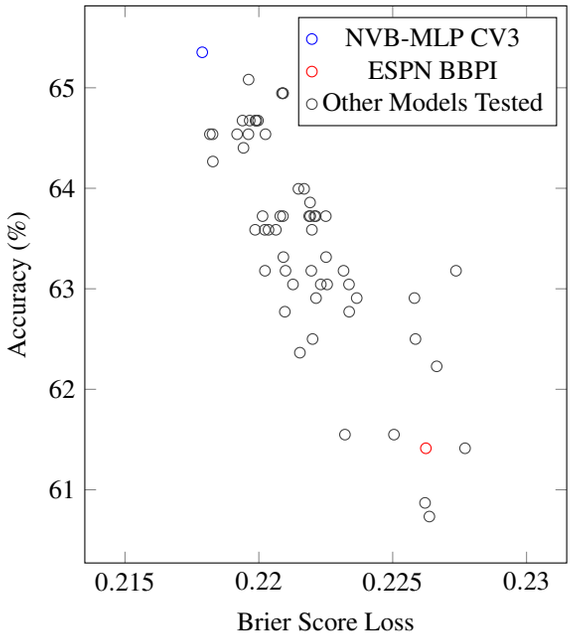

Fig. 1: Scatterplot of Accuracy vs. Brier Score Loss (second dataset)

While the scatterplot suggests there might be a linear correlation between the Brier Score Loss and the Accuracy variables, a Pearson's correlation could not be run as the Accuracy variable does not come from a normal distribution according to D'Agostino and Pearson's normality test [10]. Therefore, a Spearman's correlation was run on 59 pairs of Brier Score Losses and Accuracies. There was a strong, negative monotonic correlation between Brier Score Loss and Accuracy (r_s = -0.79, n = 59, $p < 0.001$).

3.2 Betting Simulation

The results of all models on the betting simulation is described in the table below.

Name	Profit	ROI (%)
Reference		
ESPN BBPI	-26.96	-4.6
Tested Models		
All Models	12.68 +/- 16.86	2.1 +/- 2.8
Best Model	54.59	9.3
Sub-Optimal Betting Strategies		
Home Team	-39.64	-5.4
Away Team	-12.54	-1.7
Favorite Team	-31.82	-4.3
Underdog Team	-20.36	-2.8
Random Pick	-0.47 +/- 0.55	-0.06 +/- 0.07

Most of models were able to make a profit after several hundred games and surpass sub-optimal betting strategies, which are not profitable. Only the random pick strategy was barely on the even line. Probability outputs from the Basketball Power Index did not allow to make any profit.

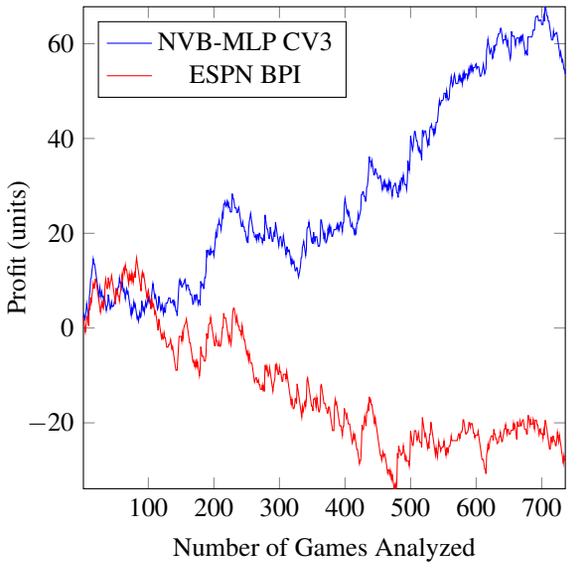

Fig. 2: Profit curve of ESPN BBPI (red) and the best performing model developed (blue)

Looking at the profit curves above, we can see that the optimal model could not avoid the typical ups and downs which are inevitable consequences of the variance. This is also due to the fact that the model is most of times betting on underdog teams which are thought to be under-estimated. This implies a rather low winrate (compensated with bigger earnings) and a higher likeliness to lose on several games in a row.

To observe the relationship between model metrics and the profit generated, the scatterplot of the Brier Score Loss against the Return on Investment for all models has been plotted (Figure 2).

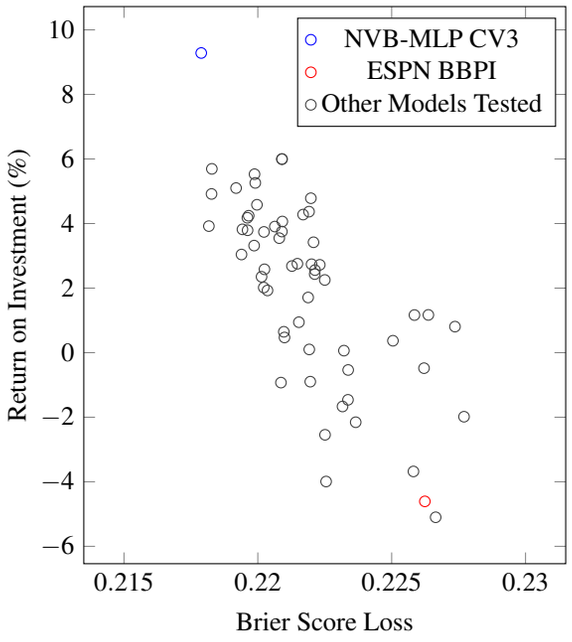

Fig. 3: Scatterplot of the Return of Investment vs. Brier Score Loss

While the scatterplot suggests there might be a linear correlation between the Brier Score Loss and the Return on Investment variables, a Pearson's correlation could not be run as the Return on Investment variable does not come from a normal distribution according to D'Agostino and Pearson's normality test [10]. Therefore, a Spearman's correlation was run on 59 pairs of Brier Score Losses and Returns on Investment. There was a strong, negative monotonic correlation between Brier Score Loss and Return on Investment (r_s = -0.74, n = 59, $p < 0.001$).

4 Discussions

Different classifiers have been trained and validated through cross-validation using the first dataset. As the literature is mostly focusing on predicting the winner of NBA games and not computing the winning probabilities for each team, no Brier Score Loss value has been found for comparison. The Area under ROC Curve and Accuracy achieved with the different classifiers are on-par with the state-of-the-art [1, 2]. The accuracy is also on-par with the reported accuracy achieved by NBA experts [11]. However, it should be noted that the present study used games with both teams having played at least 20 games while the literature were using whole seasons for the training and validation. The developed models are expected to perform slightly worse using whole seasons as they would be analyzing games with team statistics that might not be robust enough at the start of said seasons.

The pre-trained classifiers have been then tested with a second dataset, whose games are exclusive from the first dataset. They have been compared to a predictive model that is already deployed in production and accessible to the public, using metrics and a betting simulation. Most of the de-

veloped models were performing better than the reference model with a improvement of Brier Score Loss by up to 3.5%. While the improvement does not look spectacular at first sight, it has been shown that a small improvement of the Brier Score Loss leads to a significant improvement on both the accuracy of the model and the ability to generate betting profit. This is thought to be due to the highly unpredictable nature of NBA games, the average winning probability for the home team being around 50-60% across the developed models on the second dataset. This implies that the Brier Score Loss will be close to 0.25 and getting very small values is not realistic. As this study is the first one to author's knowledge to test predictive models on games before they happened, it is not feasible to make an apple-to-apple comparison with the literature. Base classifiers (GBC, LR, SVC, NVB, MLP) were giving satisfactory results with Platt Scaling but using Ensemble classifiers allowed to optimize further the performance.

The fact that most models were making profit through betting in real-life conditions is seen as a particularly encouraging result, given that the odds proposed by bookmakers are not fair. For example, for a game with both teams having similar skills, their odds will be equal to 1.9 or 1.95 instead of 2 implying a margin of 2.5 or 5% for the bookmaker [12]. However, even if the betting simulation was run on several hundred games, models should be tested on even more games as NBA games are highly unpredictable, generating important variance. Until it is done, it is not recommended to use the developed models to make real-life betting decisions.

For the future, it should be interesting to train the models on a bigger dataset. In fact, one of the major advantages of this study is that very few parameters are used to analyze games. This means that the first dataset can be scaled up while keeping the training and validation time reasonable. For instance, the hyperparameter tuning of all base classifiers has been made in a cloud environment without GPU/TPU resources. Another thing that might be worth trying is to implement Sentiment Analysis in the game analysis, using social media and Natural Language Processing, to see how teams are perceived before a game (about streaks, recent injuries or returns, etc.). This has already been successfully implemented in the literature for the prediction of NFL games [13]. At last, the notion of momentum can be implemented with a rolling average for team statistics. This would allow to see if a team is improving or under performing recently.

5 Conclusions

A series of models have been trained in order to output the winning probabilities of both teams for NBA games. The combination of using Platt Scaling and using Ensemble Classifiers allowed to achieve performance on-par with the literature (in terms of accuracy) and slightly superior to the Basketball Power Index. A strong correlation between the Brier Score Loss and both the Accuracy and the ability to make betting profit has been observed, showing the importance of optimizing the Brier Score Loss for probabilistic classifiers. The most promising model (NVB-MLP with a 3-fold cross-validation) is being deployed into production in a web application (beta version accessible here). The web application has been developed with a back-end made in Python (Flask) and a front-end made in HTML, CSS and JavaScript. The user can modify at will the list of inactive players and compute the winning probabilities of both teams accordingly. Datasets and a Jupyter Notebook for profit computation are available on the GitHub of the project (link), so you can try your own models and compare them to the results disclosed in this report. Only transformed data are shared as giving raw data (team statistics) would not comply with Section 5.j of Terms of Use of Sports Reference.

Further work is planed in order to try to improve further the results, the priority being put on increasing the size of the training-validation dataset. This should not only increase the performance of the models but also allow the use of Isotonic Regression in place of Platt Scaling for the probabilities calibration.

References

[1] Miljković, Dragan & Gajić, Ljubisa & Kovacevic, Aleksandar & Konjovic, Zora. (2010). The use of data mining for basketball matches outcomes prediction. 309-312. 10.1109/SISY.2010.5647440.

[2] Migliorati, Manlio. (2021). Features selection in NBA outcome prediction through Deep Learning. (Pre-print)

[3] Albert, A.A.; de Mingo López, L.F.; Allbright, K.; Gomez Blas, N. A Hybrid Machine Learning Model for Predicting USA NBA All-Stars. Electronics (2022),11,97. https://doi.org/10.3390/electronics11010097

[4] Platt, John. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif.. 10.

[5] Zadrozny, Bianca & Elkan, Charles. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 10.1145/775047.775151.

[6] Niculescu-Mizil, Alexandru & Caruana, Rich. (2005). Predicting good probabilities with supervised learning. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. 625-632. 10.1145/1102351.1102430.

[7] Shen, H.-B & Chou, Kuo-Chen. (2007). Using ensemble classifier to identify membrane protein types. Amino acids. 32. 483-8. 10.1007/s00726-006-0439-2.

[8] Zhong, Wenliang & Kwok, James. (2013). Accurate probability calibration for multiple classifiers. IJCAI International Joint Conference on Artificial Intelligence. 1939-1945.

[9] Dean Oliver, Basketball On Paper: Rules And Tools For Performance Analysis (2004), Potomac Books Inc.

[10] Schober, Patrick & Boer, Christa & Schwarte, Lothar. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. 126. 1. 10.1213/ANE.0000000000002864.

[11] Loeffelholz, Bernard & Bednar, Earl & Bauer, Kenneth. (2009). Predicting NBA Games Using Neural Networks. Journal of Quantitative Analysis in Sports. 5. 7-7. 10.2202/1559-0410.1156.

[12] Hubáček, Ondřej & Sourek, Gustav & Železný, Filip. (2019). Exploiting sports-betting market using machine learning. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2019.01.001.

[13] Sinha, Shiladitya & Dyer, Chris & Gimpel, Kevin & Smith, Noah. (2013). Predicting the NFL using Twitter. Proc. ECML/PKDD Workshop on Machine Learning and Data Mining for Sports Analytics.