



**SCHOOL OF INFORMATION SCIENCES
& TECHNOLOGY**

DEPARTMENT OF STATISTICS
POSTGRADUATE PROGRAM

**Predicting European basketball transfers using
Statistical and Machine learning methods**

By

Dimitrios Vas. Eleftheriou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece

May 2020





**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Πρόβλεψη Ευρωπαϊκών μετεγγραφών στο Μπάσκετ
με τη χρήση στατιστικών μεθόδων και της
μηχανικής μάθησης

Δημήτριος Βασ. Ελευθερίου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Μάιος 2020





Contents

1	Basketball Analytics and Data Science	1
1.1	History of Sport Analytics	1
1.2	Data Science	3
1.3	Basketball Analytics	4
1.3.1	Basketball Evaluation Concepts	4
1.3.1.1	Player Evaluation Metrics	4
1.3.1.1.1	Simple Player Evaluation Metrics	5
1.3.1.1.2	Advanced Player Evaluation Metrics	5
1.3.1.2	Team Evaluation Metrics	11
1.4	Prediction	14
1.4.1	Player Performance Prediction	15
1.4.2	Team Performance Prediction	18
1.4.3	Player Movement Prediction	19
1.4.4	Sport Injuries Prediction	19
1.5	Summary	20
2	Introduction	21
2.1	Initial Data	22
2.2	Data Augmentation	24
2.3	Final Data	29
2.4	Exploratory Data Analysis	31



3	Statistical Modelling	42
3.1	Logistic Regression	42
3.2	Logistic Regression per Year	47
3.3	Bayesian GLMM	54
3.4	Examples	70
4	Machine Learning Modelling	75
4.1	Machine Learning Classifiers	75
4.1.1	Support Vector Machines (SVM)	75
4.1.2	Decision Trees, Bagging and Random Forests	76
4.1.2.1	Decision Trees	77
4.1.2.2	Bagging and Random Forests	77
4.2	Random Forests implementation for transfer prediction in Eurocup Basketball Data	79
4.2.1	Random Forests per Year	81
4.2.2	Random Forests in Case-Control study	84
5	Prediction and Evaluation of transfers	86
5.1	Prediction of transfers	86
5.2	Evaluation of Transfers	89
6	Conclusions Discussion	92
7	Bibliography	122



List of Figures

2.1	Differences in Points per game among years	32
2.2	Differences in Assists per game among years	33
2.3	Differences in Turnovers per game among years	34
2.4	Differences in Offensive Rebounds per game among years	35
2.5	Differences in Offensive Rating per game among years . .	35
2.6	Differences in Defensive Rating per game among years . .	36
2.7	Differences in Minutes per game among years	36
2.8	Differences in Games among years	37
2.9	Linear correlations between variables of the dataset under consideration	40
3.1	Coefficient estimates and 95% confidence intervals for the Intercept	51
3.2	Coefficient estimates and 95% confidence intervals for the Points Produced	52
3.3	Coefficient estimates and 95% confidence intervals for the Net Points	53
3.4	Posterior Distribution of ScPoss	60
3.5	Posterior Distribution of Floor%	60
3.6	Posterior Distribution of Eurocup and Euroleague experience	61
3.7	Posterior Distribution of BLK and OREB	61
3.8	Posterior Distributions of NetPoints and FTA	62



3.9	Diagnostics of MCMC convergence	62
3.10	Difference in mean of ScPoss from the average player of the year between groups	65
3.11	Difference in mean of NetPoints from the average player of the year between groups	65
3.12	Posterior vs Prior for the model with random Intercept . . .	68
3.13	ROC curve for the model with random Intercept	69
4.1	Variable Importance Plot for Random Forests in the com- plete data set	80
4.2	Variable Importance Plot for Random Forests model in the case group	84
5.1	Differences in means of TPI	91



List of Tables

2.1	Transpose of the original data	23
2.2	Description of Variables	25
2.2	Description of Variables (continued 2/4)	26
2.2	Description of Variables (continued 3/4)	27
2.2	Description of Variables (continued 4/4)	28
2.3	Per Game Statistics of outlier observations in figure 2.8	38
3.1	Table of Coefficients for the full logistic regression model	44
3.1	Table of Coefficients for the full logistic regression model (continued 2/2)	45
3.2	Coefficients of 'best' models according to AIC (1/2)	49
3.2	Coefficients of 'best' models (continued 2/2)	50
3.3	The values of $elpd_{loo}$ for nine different models	59
3.4	Error terms estimation for the model with random intercept and random slopes	67
3.5	Quantiles of fitted values for players that made a transcription	69
3.6	Quantiles of fitted values for players that did not make a transcription	69
3.7	Per game Statistics differences from the average player of Dairis Bertans	71
3.8	Per game Statistics differences from the average player of Dairis Bertans	72



3.9	Per game Statistics differences from the average player of Brian Randle	73
3.10	Per game Statistics differences from the average player of Brian Randle	74
4.1	Predictions of Random Forests Model	81
4.2	Mean Decrease Gini Index over years for Variable Importance	83
5.1	Predictions of 2018 transfers	87
5.2	Feature statistics of misclassified players	88
6.1	Successful transfers according to TPI	94
6.1	Successful transfers according to TPI (continued 2/28) . .	95
6.1	Successful transfers according to TPI (continued 3/28) . .	96
6.1	Successful transfers according to TPI (continued 4/28) . .	97
6.1	Successful transfers according to TPI (continued 5/28) . .	98
6.1	Successful transfers according to TPI (continued 6/28) . .	99
6.1	Successful transfers according to TPI (continued 7/28) . .	100
6.1	Successful transfers according to TPI (continued 8/28) . .	101
6.1	Successful transfers according to TPI (continued 9/28) . .	102
6.1	Successful transfers according to TPI (continued 10/28) . .	103
6.1	Successful transfers according to TPI (continued 11/28) . .	104
6.1	Successful transfers according to TPI (continued 12/28) . .	105
6.1	Successful transfers according to TPI (continued 13/28) . .	106
6.1	Successful transfers according to TPI (continued 14/28) . .	107
6.1	Successful transfers according to TPI (continued 15/28) . .	108
6.1	Successful transfers according to TPI (continued 16/28) . .	109
6.1	Successful transfers according to TPI (continued 17/28) . .	110
6.1	Successful transfers according to TPI (continued 18/28) . .	111
6.1	Successful transfers according to TPI (continued 19/28) . .	112



6.1	Successful transfers according to TPI (continued 20/28) . .	113
6.1	Successful transfers according to TPI (continued 21/28) . .	114
6.1	Successful transfers according to TPI (continued 22/28) . .	115
6.1	Successful transfers according to TPI (continued 23/28) . .	116
6.1	Successful transfers according to TPI (continued 24/28) . .	117
6.1	Successful transfers according to TPI (continued 25/28) . .	118
6.1	Successful transfers according to TPI (continued 26/28) . .	119
6.1	Successful transfers according to TPI (continued 27/28) . .	120
6.1	Successful transfers according to TPI (continued 28/28) . .	121





Chapter 1

Basketball Analytics and Data Science

The 21st century is undeniably the era of technology and data collection as the growth rate for both increases rapidly. Most big companies collect data in order to improve their performance and nowadays some of the biggest companies in the world are sports companies, many of which are associated with sports teams. The development of advanced technology and the data increase, have contributed very effectively to the development of sports teams, players as well as the formulation of coaching strategies since detailed data are recorded not only for each player, but also for entire teams and games. As a result, analysts have plenty of play by play data as well as classic box scores and more complex data like court coordinates of the players and the ball.

1.1 History of Sport Analytics

Data from a variety of sports has been collected for decades to document the history and progress of sports, players and teams. As years go by,



more and more information is available. Over the last years, statistical and computer science methods are used more and more for the analysis of sports.

One of the most famous example of these people is Billy Beane (William Lamar Beane III), executive vice president of the Oakland Athletics of MLB (Major League Baseball) team. During the 2001-2002 season, Billy Beane used statistical methods to evaluate baseball players as a general manager of his team and managed with a small roster of players, compared to other teams, to achieve some remarkable achievements such as series of 20 consecutive wins. Since then, there have been a few general managers who have followed his tactic and started using similar or the same, player evaluation statistics.

Nowadays, baseball is the sport where sport analysis has gained its greatest growth and recognition. Billy Bean was the pioneer of sports analysis, with his accomplishments written in a book by Michael Lewis (2003) entitled 'MoneyBall'. However, in reality the first person who introduced statistics in baseball was Bill James (James,1977), using advanced metrics to evaluate the in-game productivity and efficiency of a baseball player.

Nevertheless, Sport analysis has a wider scope than just player evaluation, such as fitness assessment, injury prediction, match scheduling etc. More details will be introduced below.

It was apparent that the rest of the sports community could not be left unaffected. Dean Oliver (statistician) and John Hollinger (ESPN author) were the first who popularized the use of advanced metrics in basketball. They were also credited with the promotion of sabermetrics methods adopted in their books "Basketball on Paper" (Oliver,2004) and "Pro Basketball Forecast" (Hollinger, 2005), respectively. Daryl Morey, general manager of the Houston Rockets NBA team, brought the revolution to the NBA. His methods showed the enormous contribution of the players'



3-point efforts. By his analysis, he concluded that the more 3-point efforts the team has, the higher the probability of winning a game it was. This fact was evident not only by the increased effort of three-pointers by the Houston Rockets, but by the entire NBA community.

Specifically, the three-point effort rate recorded an increase of 11.4% in less than a decade changing from 22.3% for the 2008-2009 season to 33.7% for the 2016-2017 season. In fact, this figure is still constantly increasing.

1.2 Data Science

Data science is a multidisciplinary field which utilizes scientific methods and algorithms to derive information and observations from structured and unstructured data. It unifies different data related scientific areas such as statistics, data analysis and machine learning with the aim to predict and/or understand and interpret real life phenomena. Someone might argue that data science is no different from statistics, but in fact, as the years go by, the difference gets clearer (Gutierrez, 2019). Undoubtedly, data science is one of the most popular fields of science and according to the 2012 Harvard Business Review (Davenport and Patil, 2012), it is the sexiest job of the 21st century.

As it has already been mentioned, the huge contribution of data science is not just limited to sport analysis. Recognizing a disease through specific symptoms in medicine, predicting a stock market performance or predicting and interpreting climate change on the planet are really just a few of examples where data science has proven to be extremely useful.



1.3 Basketball Analytics

Basketball analytics is nothing more than data based analysis that focuses on basketball statistics. The majority of basketball fans and non-fans believe that basketball analytics refers only to typical player and team statistics, figures, indexes and obtained from a game or a season. Nevertheless, the reality is more than just these numbers.

1.3.1 Basketball Evaluation Concepts

It is very important for each team (or for a business in general) to be able to evaluate their staff and its active components. In particular, a sports team is interested to evaluate the performance of their players (both individually and at the team level) and their fitness. The same might also be for other active members of the team such as coaches. For this reason, a great deal of research has been done on the development of such measures, some of which are presented in this section.

1.3.1.1 Player Evaluation Metrics

Player evaluation metrics refer to the quantification of overall performance of a player. By using such players performance metrics it is possible to make comparisons between players of the same or different teams. The bibliography of such measures is quite large with a great variety of different methods and indicators. This refers to different measures, some of them being intuitive and arbitrary while others are more based on scientific methods. For this reason, the separation of performance indicators in two categories is made. These categories are (a) simple evaluation metrics and (b) advanced evaluation metrics.



1.3.1.1.1 Simple Player Evaluation Metrics

Efficiency (EFF): This metric, is introduced by Martin Manley (Manley, 1987). It is just the sum of the positive accomplishments minus the negative ones.

$$\begin{aligned} \text{EFF} = & \text{Points} + \text{Rebounds} + \text{Assists} + \text{Steals} + \text{Blocks} \\ & - \text{Missed Field Goals} - \text{Missed Free Throws} - \text{Turnovers} \end{aligned} \quad (1.1)$$

The major advantage of this metric is that it's very simple to calculate. Is considered as the first player evaluation metric which indicates player's linear efficiency. Nowadays, it is considered outdated and it is rarely used in practice.

Performance Index Rating (PIR): This metric currently used by the Euroleague Basketball Company's first and second tier competitions (the Euroleague and the EuroCup) as well as various European national domestic and regional leagues. It's not the same as the NBA's EFF (Efficiency), but it's the same easy to calculate. PIR is given by:

$$\begin{aligned} \text{PIR} = & (\text{Points} + \text{Rebounds} + \text{Assists} + \text{Steals} + \text{FoulsDrawn}) \\ & - (\text{Missed Field Goals} + \text{Missed Free Throws} + \text{TurnOvers} \\ & + \text{Shots Rejected} + \text{Fouls Committed}) \end{aligned} \quad (1.2)$$

1.3.1.1.2 Advanced Player Evaluation Metrics

Player Efficiency Rating (PER): This metric is the John Hollinger's (Hollinger, 2005) attempt to evaluate the total contribution of a player



in a single number by adding and subtracting the positive and negative accomplishments of a player, respectively, through a statistical point value system. It's the most commonly used alternative to EFF. The formula presented below was written with abbreviations of the statistics that were used. The specification of those statistics can be found at Table 2.2 'Description of Variables'.

First we calculate unadjusted uPER given by:

$$\begin{aligned}
 \text{uPER} = & \frac{1}{\text{min}} \times (3P + \left[\frac{2}{3} \times \text{AST} \right] + \left[\left(2 - \text{factor} \times \frac{\text{tmAST}}{\text{tmFG}} \right) \times \text{FG} \right] \\
 & + \left[0.5 \times \text{FT} \times \left(2 - \frac{1}{3} \times \frac{\text{tmAST}}{\text{tmFG}} \right) \right] - [\text{VOP} \times \text{TO}] \\
 & - [\text{VOP} \times \text{DRBP} \times (\text{FGA} - \text{FG})] \\
 & - [\text{VOP} \times 0.44 \times (0.44 + 0.56 \times \text{DRBP}) \times (\text{FTA} - \text{FT})] \\
 & + [\text{VOP} \times (1 - \text{BRBP}) \times (\text{TRB} - \text{ORB})] \\
 & + [\text{VOP} \times \text{BRBP} \times \text{ORB}] \\
 & + [\text{VOP} \times \text{STL}] + [\text{VOP} \times \text{DRBP} \times \text{BLK}] \\
 & - \left[\text{PF} \times \left(\frac{\text{lgFT}}{\text{lgPF}} - 0.44 \times \frac{\text{lgFTA}}{\text{lgPF}} \times \text{VOP} \right) \right]
 \end{aligned}
 \tag{1.3}$$

Equation (1.3) can be also written as:



$$\begin{aligned}
uPER & \frac{1}{min} \times \left(3P - \frac{PF \times lgFT}{lgPF} + \left[\frac{FT}{2} \times \left(2 - \frac{tmAST}{3 \times tmFG} \right) \right] \right. \\
& \left. + \left[FG \times \left(2 - \frac{factor \times tmAST}{tmFG} \right) \right] + \frac{2 \times AST}{3} \right. \\
& + VOP \times [DRBP \times (2 \times ORB + BLK - 0.2464 \times [FTA - FT]) \\
& - [FGA - FG] - TRB \\
& + \frac{0.44 \times lgFTA \times PF}{lgPF} - (TO + ORB) + STL + TRB \\
& \left. - 0.1936 \times (FTA - FT) \right]
\end{aligned} \tag{1.4}$$

where,

$$factor = \frac{2}{3} - \left[\left(0.5 \times \frac{lgAST}{lgFG} \right) \div \left(2 \times \frac{lgFG}{lgFT} \right) \right],$$

$$VOP = \frac{lgPTS}{lgFGA - lgORB + lgTO + 0.44lgFTA},$$

$$DRBP = \frac{lgTRB - lgORB}{lgTRB}.$$

The unadjusted uPER must be then adjusted for the team pace and then normalized by the league in order to become PER. Hence it is calculated as:

$$PER = \left(uPER \times \frac{lgPace}{tmPace} \right) \times \frac{15}{lg uPER}$$

This final step eliminates the advantage held by players whose teams play a fastbreak style and then adjust by the league average which is set to 15.00.



It should be noted that PER is per minute statistic, that is, it quantifies a player's contribution per minute. Moreover, this measure gives an advantage in favor of aggressive players. As a result 'unskilled', but aggressive players score higher PER scores than those who are considered to be more 'skilled', but non-aggressive.

Win Shares (WS): Win shares is an advanced metric which estimates the player's contribution to the team's win total. It was developed by the American writer and statistician Bill James in his book 'Win Shares' (James and Henzler, 2002). In this book James explains how to apply the methods of sabermetrics, which assess the impact of a baseball player performance, to his team performance. The formula to calculate WS for a player is too complex but a brief description is:

(1) Credit offensive win shares to the players by calculating player's marginal offence from his points produced¹ and offensive possessions and dividing it by the marginal points per win.

(2) Credit defensive win shares are credited by computing a player's marginal defense from his defensive rating² and dividing it by the marginal points per win.

(3) Then, simply add offensive and defensive win shares together to get total win shares.

Due to the fact that this metric is designed to estimate a player's contribution in terms of wins, it is expected that the sum of player WS for a particular team be closely to the total number of wins of this team.

¹Points Produced is an advanced statistic which measures how many points a player produces. "Basketball on Paper" by Dean Oliver (2004)

²Defensive Rating is an advanced statistic which estimates the points allowed per 100 possessions for a player. "Basketball on Paper" by Dean Oliver (2004)



Plus/Minus (+/-): Plus/Minus is a metric, first used in ice hockey, which keeps track of the net changes in score when a particular player is either on or off the court. Plus/Minus does not account for the impact of teammates or opponents, so an improvement of this measure is the Adjusted Plus/Minus (APM), which reflects the impact of each player on his team's scoring margin after controlling for the strength of every teammate and opponent during each minute he's on the court.

Adjusted plus/minus: is estimated via linear regression. So, the formula is given by:

Obtain the OLS estimates for b_i for the set of equations

$$Y_1 = b_1 H_{1,1} + b_2 H_{2,1} + \dots + b_i H_{i,1} + b_{i+1} A_{i+1,1} + b_{i+2} A_{i+2,1} + \dots + b_{i+s} A_{i+s,1},$$

$$Y_2 = b_1 H_{1,2} + b_2 H_{2,2} + \dots + b_i H_{i,2} + b_{i+1} A_{i+1,2} + b_{i+2} A_{i+2,2} + \dots + b_{i+s} A_{i+s,2},$$

⋮

$$Y_p = b_1 H_{1,p} + b_2 H_{2,p} + \dots + b_i H_{i,p} + b_{i+1} A_{i+1,p} + b_{i+2} A_{i+2,p} + \dots + b_{i+s} A_{i+s,p}.$$

where,

Y_j is the average point differential of home team over away team per 100 possessions

$H_{i,p}$ indicates if player i on team H (Home team) plays in the section of time p , which group of players are on the court.

$A_{i,p}$ indicates if player i on team A (Away team) plays in the section of time p , which group of players are on the court.



The beta values are the weights of contribution that players give towards the difference in score, so the estimates of betas are the adjusted plus/minus measures for the players in a single game.

Due to the multicollinearity and noise of data, betas have high variance so the estimates are unstable. An improvement is the Regularized Adjusted Plus/Minus (RAPM).

Regularized Adjusted Plus/Minus (RAPM): is similar to the adjusted plus/minus but, the estimates of betas are obtained via Ridge Regression.

Ridge Regression is regularization method, first introduced by Hoerl and Kennard (1970), which deals with the problem of multicollinearity by adding a penalty term λ , in the minimization of residual sum of squares. It's equivalent to assume a normal prior distribution for the betas of OLS estimates with mean zero and variance equal to σ^2/λ . The main advantage of this approach is the reduction of standard errors.

There are also other versions of the plus/minus approach such as the Real Plus/Minus (RPM), the Box Plus/Minus (BPM), the Statistical Plus/Minus (SPM) which in fact describes the source (box scores statistics, etc.) used to estimate the weights of each index.

Another interesting way of measuring player performance is via statistical networks. Through statistical networks modelling, the impact of teammates can be adjusted and also it can be seen if the contribution of a player in a team came as unexpected. The measure for a player's contribution to the performance of his team is a centrality score, so the player's statistical contribution is determined by the frequency with which that player is visited in a random walk on the network. For more details see Piette et.al (2009).



1.3.1.2 Team Evaluation Metrics

Similarly to the player's performance evaluation measures, is also a wide variety of team performance indicators. Indicatively, we present four of the most important ones:

Four Factors: Four Factors are the box score derived metrics that correlate most closely with winning percentage. These factors are presented by Kubatko, Oliver, Pelton & Rosenbaum (2007). These factors are the Effective field goal percentage with a weight of 40%, the Turnovers per possession with a weight of 25%, the offensive rebounding percentage with a weight of 20% and the free throw rate with a weight of 15%.

Efficiency Differential: efficiency differential indicator is the numerical gap between a team's offensive and defensive efficiencies in entire season. The formula is:

$$Efficiency\ Differential = \frac{PTS\ Scored}{Poss} \times 100 - \frac{PTS\ Allowed}{Poss} \times 100,$$

where PTS and Poss are just the abbreviations of Points and possessions, respectively.

Pythagorean Winning Percentage: This method gives an expected winning percentage using the ratio of a team's wins and losses. It is based on James (1977) formula which was originally developed for baseball. It is related to the number of points scored and allowed, and it is given by:

$$Expected\ Winning\ Percentage = \frac{PTS\ Scored^{16.5}}{PTS\ Scored^{16.5} + PTS\ Allowed^{16.5}}$$



According to Oliver (2004), when the formula is used in NBA, the exponents of the formula are varied, from 11 to 17, depending on when they were estimated. During the higher pace days, the value was higher. The variability of the estimated exponents is the main disadvantage of the method. The advantage is its simplicity.

Logistic Regression Markov Chain (LRMC): This metric is constructed by Kvam and Sokol (2006). It is a college basketball ranking system, designed to use only basic scoreboard data, including which teams played, which team had home court advantage and the margin of victory.

Specifically, they define $x(g)$ be the difference between the home team's score and the visiting (road) team's score in game g , r_x^H to be the probability that a team that outscores its opponent by x points at home is better than its opponent, and $r_x^R = 1 - r_x^H$ to be the probability that a team that is outscored on the road by x points is better than its opponent. They allowed x to be negative to indicate that the home team lost the game. Now, assuming that each outcome is a state of a Markov chain and denoting each game by an ordered pair (i, j) of teams with the visiting team listed first, the state transition probabilities for each team i are:

$$t_{ij} = \frac{1}{N_i} \left[\sum_{g=(i,j)} (1 - r_{x(g)}^R) + \sum_{g=(j,i)} (1 - r_{x(g)}^H) \right], \quad \text{for all } j \neq i$$

and,

$$t_{ii} = \frac{1}{N_i} \left[\sum_j \sum_{g=(i,j)} r_{x(g)}^R + \sum_j \sum_{g=(j,i)} r_{x(g)}^H \right]$$

where N_i is the total of games that team i played.

The probability that team A will beat team B on B 's court given that A has beat B by x points on A 's court, are estimated via the logistic



regression model

$$\log \frac{s_{x(g)}^H}{1 - s_{x(g)}^H} = b_0 + b_1 * x,$$

where $s_{x(g)}^H$ are the observed win probabilities. $s_{x(g)}^H$ answers the following question: “Given that Team A beat Team B by x points on A 's home court, what is the probability that A beat B on B 's home court?”

Finally, under the assumption that the effect of home court advantage is additive, they conclude to the result $r_x^H = s_{x+h}^H$, where h is the home-court advantage.

Although this metric is quite demanding in its construction, it outperforms other common methods such as tournament seedings, the AP and ESPN/USA Today polls, the RPI, and the Sagarin and Massey ratings.



1.4 Prediction

Forecasting is one of the most interesting areas in sport and in basketball, specifically. From the ordinary spectators to the big sports clubs, everyone is troubled with questions such as which team will be the winner in a match, what will be the final score, or even more complex ones such as how many years a player will be effective or is there a way to anticipate an upcoming injury.

Accurate forecasts are of the utmost importance for the strategy pursued by each individual regardless of the reason of interest.

Starting from the simple spectators who bet on the outcome of a game, the final score, the points a player will score and so on we have a number of multiple predictions from both spectators and betting companies. Betting companies offer a wide variety of possible betting options at player level, team level, match level and season level with odds based on the likelihood of betting on any event. It is clear that both the betting industry and the betters are interested to predict correctly the outcome of each game in order to maximize their profits.

But, not only sports fans and betting enthusiasts are interested in prediction. Every professional sports club is a company that strives to maximize its profits mainly through the team finishing in the league or championships in which it participates. It is important, therefore, to be able to make predictions about team and player performance.

The methods used to make predictions are based on the game data of each team which are constantly increasing in volume and complexity due to the evolution of technology.

Since 2013 basketball teams have at their disposal tracking data for players and the ball by SportVU. SportVU is a camera based system that collects data at a rate of 25 times per second and follows the ball and



every player on the court delivering the spatial coordinates of them. Before SportVU technology, play-by-play data has been the main source of information gathered. A typical play-by-play data set provides information about the time of the possession, the player who initiated the possession, the opposing player who initiated the possession but also information about the shot distances and player coordinates for a shot.

1.4.1 Player Performance Prediction

There have been numerous and varied in their methodology attempts to model a player's performance in order to predict it and for the concept of a player's performance that researchers try to reproduce.

For example, Hwang (2012) tries to predict the points per game using a time series model. He modeled the points per game for each player as response using a Weibull Distribution with co-varied hazard rate functions and a gamma mixing distribution. To be more specific the model is given by:

$$F(t) = 1 - e^{\lambda \cdot D(t)},$$

where,

$$D(t) = \sum_{i=1}^t [i^c - (i-1)^c] \cdot e^{\beta' \cdot x(i)}$$

Moreover the gamma mixing distribution is given by

$$g(\lambda) = \frac{a^r \lambda^{r-1} e^{-a\lambda}}{\Gamma(r)}.$$

As a result, we have

$$P(T \leq t) = \int_0^{\infty} (1 - e^{\lambda \cdot D(t)}) \frac{a^r \lambda^{r-1} e^{-a\lambda}}{\Gamma(r)} d\lambda = 1 - \left(\frac{a}{a + D(t)} \right)^r.$$



Another attempt to predict the points a player scores is made by Casals and Martinez (2013), using generalized linear mixed models (GLMM). They try to model and predict not only the points a player scores, but also the player's win score, using the same independent variables (as predictors) for both models. The win score is a player evaluation metric and is a simplification of the 'win shares'. It is very interesting to study the differences that arise both between the importance of the variables in predicting the points and the win score and in the magnitude of their influence.

Due to sparse and irrelevant data that most teams have, more advanced methods could be more useful for prediction. Vinue and Epifanio (2019), try to estimate Box Plus/Minus and WinShares advanced statistics via ROPES (Regularized Optimization for Prediction and Estimation with Sparse data), a metric which is obtained by the optimization criterion:

$$(U, V) = \underset{U, V}{\operatorname{argmin}} (\|W \odot (Y - UV^T)\|^2 + \|U\|^2 + \|DIFF_2(m, \lambda_2)V\|^2 + \|DIFF_1(m, \lambda_1)V\|^2 + \|DIFF_0(m, \lambda_0)V\|^2) \quad (1.5)$$

where,

- Y is a $n \times m$ matrix
- U is a $n \times k$ matrix of 'scores' ('coefficients'), $k = \min(n, m)$
- V is a $m \times k$ matrix of 'features' ('shapes')
- \odot is the element-wise matrix multiplication
- W is a $n \times m$ 'masking matrix' of weights



- λ_0, λ_1 and λ_2 are smoothing parameters

and PACE (Principal Component Analysis through Conditional Expectation), where the prediction for the trajectory $X_i(t)$ for the i -th subject, using the p ϕ_q eigenfunctions, is:

$$\hat{X}_i^p = \hat{\mu} + \sum_{q=1}^p \hat{\xi}_{iq} \hat{\phi}_q(t)$$

ROPES and PACE seem to predict more accurately the two metrics mentioned above (Box Plus/Minus and Win Shares) in contrast to other methods. Complete details are given in the paper.

As already mentioned, the Plus / Minus statistic and its derivatives are player performance metrics that try to include in their estimates the correlation between players in the same team and those of the opponent. Simply put, a player's overall contribution to his team during the match depends not only on himself but also on the entire network of people around him, namely his teammates and opponents. Beneath this idea, Piette, Pham, and Anand (2011) modeled the performance of players through statistical networks. More specifically, let y_{ij} denote the number of points scored (or allowed, when analyzing defense) by unit i for possession j after adjusting for home court effects. Then,

$$Y_{ij} \sim Normal(\theta_i, \sigma^2)$$

with prior distribution of θ to be $\theta_i \sim Normal(\mu, \tau^2)$, where μ represents the league-mean efficiency and τ^2 is the corresponding variance.

Assuming that two players share an interaction if they played together in a ve-man unit, they evaluate the importance of a player using his eigenvector centrality.



1.4.2 Team Performance Prediction

When referring to team performance the first thing that comes in mind is the winning percentage of the team. Maybe a team can be improved by new offensive or defensive strategies or by transcripts, but it's useful to know what is the risk from a change in team structure. It might be a good idea to explain the risk factor by these assumptions, by modelling the probability of a win.

A nice attempt to predict the outcome of a basketball game is described by Hu and Zidek (2004). The idea of this method is to add information to the likelihood which results by assuming a Bernoulli random variable $Y_{AB}(h)$ which takes the values of one when the home team wins and zero otherwise. Similarly, they define $Y_{BA}(r)$ for home team B. The same weights are chosen in the likelihood factor corresponding to each of the games A played against teams other than B, irrespective of the opponent. The log-likelihood with weights is given by:

$$\sum_{i=1}^{k_{AB}} \log f(y_{AB}(h), p_{AB}(h)) + \alpha_{AB}(h) \sum_{A(B)} \log f(y_{A(B)}, p_{AB}(h)) \\ + \beta_{AB}(h) \sum_{(A)B} \log f(y_{(A)B}, p_{AB}(h))$$

Although, this method can be enhanced with improvements like the quality of weights as they could be a pre-game information, it provides guidelines for the development of a prediction strategy. Also, the Weighted Likelihood idea has much wider applicability inside as well outside the domain of sports.



The most recent methods for analytics performance use artificial neural networks (ANN) and machine learning (ML) algorithms, as they are not too 'overloaded' with assumptions and parameters. Also, they seem to be pretty accurate in their predictions.

Examples of these methods are shown by Giuliadori (2017) and Giasemidis (2020) where the object of the study is the final result of the game. That is, considering the outcome of a victory or a defeat for the team playing at home ground, the problem becomes a classification problem. Results are presented from various methods such as random forests, artificial neural networks and support vector machines. With regard to the advantages of each method for predicting effects in different sports, useful information is provided by Langaroudi and Yamaghani (2019).

1.4.3 Player Movement Prediction

With the SportUV technology applied in basketball, it would be a paradox not to simulate and predict the movements of players during a game. We can learn a lot from such data as for example the man offensive and defensive movements of each player (Wu and Born, 2017), their movements after a shot or after a rebound. One way to do this, is to think of a player's move as a stochastic process and calculate the probabilities for every possible move. It's one of the most attractive achievement both in sports and machine learning technology. For further information we refer the interested reader to the TED talk of Rajiv Maheswaran.

1.4.4 Sport Injuries Prediction

As mentioned earlier another interesting aspect of prediction is the occurrence of an injury. In fact, most athletes and sports fans believe that injuries occur by pure bad luck, but that's not always the truth as Stephen



Smith, CEO of Kitman Labs, says. If we can predict that the probability of injury for a specific player is high, then it might be better for the player and the team not to participate in games for a small period. There are many examples in all sports where players' careers were destroyed due to an injury. Although this is the worst case scenario for an athlete, this also affects his team given that injuries are quite expensive and some times specific players can not be replaced easily. For instance, knee injuries costed 358\$ million dollars in 2014 for the teams in NBA and the overall cost of injuries in Major League Baseball was 1.4\$ billion for the same year.

Nowadays, there are numerous companies involved in sport science and sport analysis, aiming to predict an injury by finding features and variables which are responsible for sport injuries. Data science, statistics and machine learning algorithms seem to be the right way to predict an upcoming injury. Neural networks have also been used for predicting such events, but they are insufficient when out-of-sample data are presented to the network, which yields limited generalization capability. A very useful guide to understand the underlying procedures of modelling sports injuries is presented by Ruddy et.al (2019).

1.5 Summary

The aim of this chapter is to introduce those interested in the world of sport analysis as well as its main areas. The most important sector, of course, seems to be the forecast, since it is also the sector with the largest scientific and technological research.



Chapter 2

Introduction

The purpose of this thesis is to analyze, comprehend and even replicate the process behind the basketball transcription area. Specifically, the aim is to implement a methodology in order to forecast transcriptions from the EuroCup Basketball League to EuroLeague Basketball League. A secondary aim is then to evaluate the success and the performance contribution of such transcriptions.

It is important to understand that this type of transcriptions can be considered as a big achievement for the career of individual players individually and also a measure for calculating a player's performance. In fact, EuroLeague is the most important European Basketball event, so it is reasonable that every player would eventually aim to participate in this tournament. Therefore, examining the impact of box score statistics and other advanced performance measures on the probability of a player's transfer success.



2.1 Initial Data

The data set we consider in this thesis was provided by newstats.eu. The data include the complete individual box score statistics from 2010 to 2018 and part of the corresponding measures for 2019, for all EuroCup and EuroLeague players. More specifically, the data were composed of 7790 observations of 20 box score statistics per game each year for every player, that is the total box statistics divided by the number of games for each player. So the data after changing names to most of statistics, were consisted of players observations for whom we had records of the following average individual per game stats:

The name of the player, his team, the League (EuroCup or Euroleague) in which the team participates, the year of statistics, the number of games, the number of minutes, the number of points made, the two-point field goals made, the two-point field goals attempted, the two-point field goals percentage, the three-point field goals made, the three-point field goals attempted, the three-point field goals percentage, the free-throws made, the free throws attempted, the free-throw percentage, the offensive rebounds, the defensive rebounds, the total rebounds, the assists, the fouls committed, the blocks made, the steals made and the turnovers made.

During the data preparation we firstly categorized observations by player and year. Part of the data structure after some initial data management is provided at Table 2.1.

Note that for outer illustration this table presents the transpose of the operating dataset. working table which means that rows are the columns and inverse. As it can be observed, the names of variables are just abbreviations. For the complete description of the abbreviations used see Table 2.2 in page 29.



Table 2.1: Transpose of the original data

	1	2	3	4
player	aaron-cel	aaron-cel	aaron-cel	aaron-cel
league	eurocup	eurocup	euroleague	eurocup
year	2012	2014	2014	2015
team	turow-zgorzelec	stelmet-zielona-gora	stelmet-zielona-gora	stelmet-zielona-gora
games	6	6	8	10
MIN	17.92	24.74	20.84	22.45
PTS	6.83	8.83	6.50	7.00
FG2M	2.33	2.67	2.13	2.00
FG2A	4.33	4.33	3.75	4.20
FG3M	0.50	1.17	0.75	0.90
FG3A	1.17	2.17	2.25	2.70
FTM	0.66	0.00	0.00	0.30
FTA	1.00	0.33	0.00	0.40
OREB	1.33	1.33	1.38	0.70
DREB	2.83	3.50	3.00	3.80
TREB	4.17	4.83	4.38	4.50
AST	1.00	1.17	0.88	2.30
PF	2.50	1.33	2.13	2.80
BLK	0.17	0.17	0.25	0.10
STL	0.50	0.67	0.50	1.20
TOV	1.00	1.33	0.75	1.30
FG2%	0.54	0.62	0.57	0.48
FG3%	0.43	0.54	0.33	0.33
FT%	0.67	0.00	0.00	0.75



2.2 Data Augmentation

Table 2.1 contains the basic statistics that are recorded in every European Basketball League. Further measures were also calculated which are trying to quantify the ‘value’ of a player in a single number or evaluate player’s offensive or defensive skills, unfortunately do not exist in European Basketball except PIR (see section 1.3,1,1). PIR will not be very helpful in our models since it is a simple linear combination of the rest of our covariates.

As Oliver (2004), classic box statistics are not enough to evaluate the performance of a player. So it is clear that more advanced metrics to evaluate players are needed. The construction of advanced measures used by Oliver (2004) was not as simple as the European PIR, but they to quantify better the performance of players.

In order to calculate these advanced measures we need the total per year were needed, instead of average statistics per game addition of the total team statistics per year. That was a problem, because in Oliver’s formulas statistics by opposing teams are used, so these statistics have to be estimated or approximated. The final data used are listed at Table 2.2.



Table 2.2: Description of Variables

<u>Variables</u>	<u>Description</u>
Player	<i>Name of player</i>
League	<i>The league the player and his team participates</i>
Year	<i>The year of the statistics recorded</i>
Team	<i>The player's team</i>
Games	<i>The number of games played</i>
MIN	<i>The number of minutes played</i>
PTS	<i>The number of points made</i>
FG2M	<i>The number of two-point field goals made</i>
FG2A	<i>The number of two-point field goal attempts</i>
FG3M	<i>The number of three-point field goals made</i>
FG3A	<i>The number of three point fiel goal attempts</i>
FTM	<i>The number of free-throws made</i>
FTA	<i>The number of three-throw attempts</i>
OREB	<i>The number of Offensive Rebound made</i>
DREB	<i>The number of Deffensive Rebound made</i>
TREB	<i>The number of total Rebounds</i>
AST	<i>The number of Assists made</i>
PF	<i>The number of personal fouls commited</i>
BLK	<i>The number of Blocks made</i>
STL	<i>The number of Steals</i>
TOV	<i>The number of Turnovers made</i>
FG2%	<i>Two-point field goals percentage</i>
FG3%	<i>Three-point field goals percentage</i>
FT%	<i>Free-throw percentage</i>
FGM	<i>The number of field goals made</i>

Table 2.2: Description of Variables (continued 2/4)

<u>Variables</u>	<u>Description</u>
FGA	<i>The number of field goals attempted</i>
FG%	<i>Field-goals percentage</i>
TS%	<i>True shooting percentage</i>
eFG%	<i>Effective field-goal percentage</i>
ORtg	<i>Points Produced divided by total possessions times 100</i>
DRtg	<i>Points allowed divided by total possessions times 100</i>
Floor%	<i>Scoring possessions divided by possessions</i>
ScPoss	<i>Scoring possessions</i>
Poss	<i>Total possessions</i>
Stop%	<i>Stops per possession</i>
Stops	<i>The number of stops made</i>
PtsPerScPoss	<i>Points produced per scoring possession</i>
TMFG3%	<i>Team three-point field goal percentage</i>
TMFG%	<i>Team field-goal percentage</i>
TMDREB	<i>Team defensive rebounds</i>
TMBLK	<i>Team blocks</i>
TMFT%	<i>Team free-throw percentage</i>
TMFGM	<i>Team field-goals made</i>
TMFGA	<i>Team field-goals attempts</i>
TMFG3M	<i>Team three-point field goal made</i>
TMFG3A	<i>Team three-point free goal attempts</i>
TMPF	<i>Team personal fouls committed</i>
TMOREB	<i>Team offensive rebounds</i>
TMMIN	<i>Team total minutes played</i>
TMFTM	<i>Team free-throw made</i>



Table 2.2: Description of Variables (continued 3/4)

<u>Variables</u>	<u>Description</u>
TMFTA	<i>Team free-throw attempts</i>
TMposs	<i>Team total possessions</i>
TMplay	<i>Team plays. The technical definition at Dean Oliver's book is the period between when one team gains control of the ball and when they lose control of the ball, either when the opposing team gains control or when a shot goes up</i>
TMOREB%	<i>Team offensive rebounds percentage</i>
TMplay%	<i>Team play percentage. Play percentage is the fraction of a team's plays on which it produces a scoring possession</i>
TMTOV	<i>Team turnovers</i>
TMSTL	<i>Team steals</i>
TMScPoss	<i>Team scoring possessions</i>
TMPTS	<i>Team points made</i>
TMptsPerScPoss	<i>Team points produced per scoring possessions</i>
TMORtg	<i>Team points produced divided by team total possessions times 100</i>
TMDRtg	<i>Team points allowed divided by team total possessions times 100</i>
DPtsPerScPoss	<i>Points produced per scoring possession by opposing teams</i>
DFGA	<i>Field goal attempts by opposing teams</i>
DOREB%	<i>Offensive rebound percentage by opposing teams</i>
DDREB	<i>Defensive rebounds by opposing teams</i>
TMDFT%	<i>Free-throw percentage by opposing teams</i>
TMDFTA	<i>Free-throw attempts by opposing teams</i>
TMDTOV	<i>Turnovers by opposing teams</i>
TMDFGM	<i>Field goals made by opposing teams</i>



Table 2.2: Description of Variables (continued 4/4)

<u>Variables</u>	<u>Description</u>
TMFloor%	<i>Team scoring possessions divided by possessions</i>
PointsProduced	<i>The credit an individual receives for the points his team generates on the offensive end</i>
NetPoints	<i>The difference between points produced and points allowed</i>
Win%	<i>Individual winning percentage</i>
Eurocup_exp	<i>The number of years a player played in Eurocup</i>
Euroleague_exp	<i>The number of years a player played in Euroleague</i>

The formulas for constructing Wining Percentage, Points Produced, Stops, Offensive and Defensive Rating, Possessions and Scoring Possessions, Floor and Play Percentages are given by Oliver (2004).

In these formulas we had to use the opponent's team statistics. As for the opponent's field goals made, field goal attempts, field goal percentage, free throw percentage, free throw attempts, offensive rebounds and offensive rebound percentage, defensive rating and points produced per scoring possessions we approximate them simply by the average stats of the league by not taking into account the team for which the statistics were made. Thus, the interpretation of these statistics changed from the statistics of the opposing team to the statistics of the league average team. They also show what our team could achieve if it was praised again and again with a team whose statistics would be the same as those of the middle league team. Maybe, the new interpretation of these statistics is a bit misleading because that team does not exist, but it was a discount that had to be made in order to calculate those statistics.

Also, instead of approximating the opposing defensive rebounds again



by the average of the league, a more accurate way was opted to use. The opponent's defensive rebounds are proportions of our team's missed field goal attempts and missed free throw attempts. So, by taking sample of 50 basketball games in both Eurocup and Euroleague competitions and assuming a normal distribution for the opponent's defensive rebound, the formula was:

$$DDREB = 0.86 \cdot missedFG + 0.7 \cdot missedFT - 0.81 \cdot OREB$$

with $R^2 = 0.72$.

The values of DDREB are always integers and the formula given above does not give results in integers numbers, so a Poisson distribution might be more appropriate, But, a simple histogram of the DDREB showed that a normality assumption was rationale. It is interesting that the normality assumption in many different variables in basketball seems to be reasonable assumption. For example, Oliver's formula for estimating the possessions of a team came under the normality assumption of possessions.

Also, the eurocup_exp and euroleague_exp variables which are indicators of how many years a player takes part at Eurocup league and Euroleague league, respectively, have been measured since 2010. So, a player who had been playing in Eurocup league, for example, from 2008 to 2012 and has 4 years of Eurocup experience, here was considered to have only two years.

2.3 Final Data

The dataset was almost finalised after the calculation of advanced measures. the next step was to identify our main response variable which was whether a player managed to transfer from Eurocup to Euroleague and the year



that the transfer took place. A binary indicator variable with the name "transcript" was defined as:

$$transcript = \begin{cases} 1, & \text{if a Eurocup player made a transcription to Euroleague} \\ 0, & \text{otherwise} \end{cases}$$

This variable was meaningful only for Eurocup players. The transcription indicator takes the value of one at the end of the year the transcription is made. For example, if a player was a Eurocup player at 2012 season and a Euroleague player at 2013 season, then the value of transcript will take the value of one for 2012 year and it can not be specified for 2013 as the player was a member of Euroleague at that year. Also, there are many cases where teams, immediately after a good season, were upgraded and took part in the Euroleague Championships instead of the Eurocup Championships where they were the previous year. Therefore, such cases where the whole team 'jumped' from Eurocup to Euroleague were not taken into account as a transcription. So, in order the transcription variable to take the value of one for a player, the individual player must 'move' from Eurocup to Euroleague in two subsequent years by also changing team.

Finally, all individual player statistics centered around the year specific mean of the tournament. Hence, all variables express the player's difference from the league average, for each year of registration, at the game level. Furthermore, the observations of players that participated only in Euroleague were removed since they do not offer any information for the response of interest

In the following section we present exploratory analysis of our data. By this way we will learn about peculiarities and special characteristics of the problem we deal with. This will help us to build more sensible prediction models in the next step of our analysis. Data processing in this thesis was



conducted with RStudio version (3.6.1).

2.4 Exploratory Data Analysis

A simple descriptive analysis of our data showed that there are players who played less than 10 minutes in a whole Eurocup season. An interesting point under investigation was that we identified players who managed to transfer even when they had less than five minutes of participation in a season in total. In fact, two such players were found. The first one was Antonios Koniaris who made a transfer from Paok BC to Panathinaikos BC in 2014 with a total of 4 minutes played in Eurocup games. The second one, was Dimitrios Agravanis who made a transcription from Panionios BC to Olympiakos BC in 2013 with a total of 1 minute played in a single game. Note that, players with less than seven minutes in total in a single game in a Eurocup season and even players than played less than ten minutes in total with at least two games in the competition were excluded from our analysis.

Note that in order to account for participation time of each player, we have used the total time played (in minutes) instead of the average time as in other game metrics. The reason for this is that there were players with an average time of about 5 minutes per game, but they had played more than 5 games in total. If the average statistics had been used, then they should have been excluded, but a player who played in five games and not for too long might have a bigger impact on his team than another one who had played ten minutes in total in only two games.

The next step was to investigate which variables discriminate between players who transferred and players who did not transfer. The simplest way to do so is to visualize both the statistics of these two groups of players. In all figures of this section, statistics shown refer to the differences from the



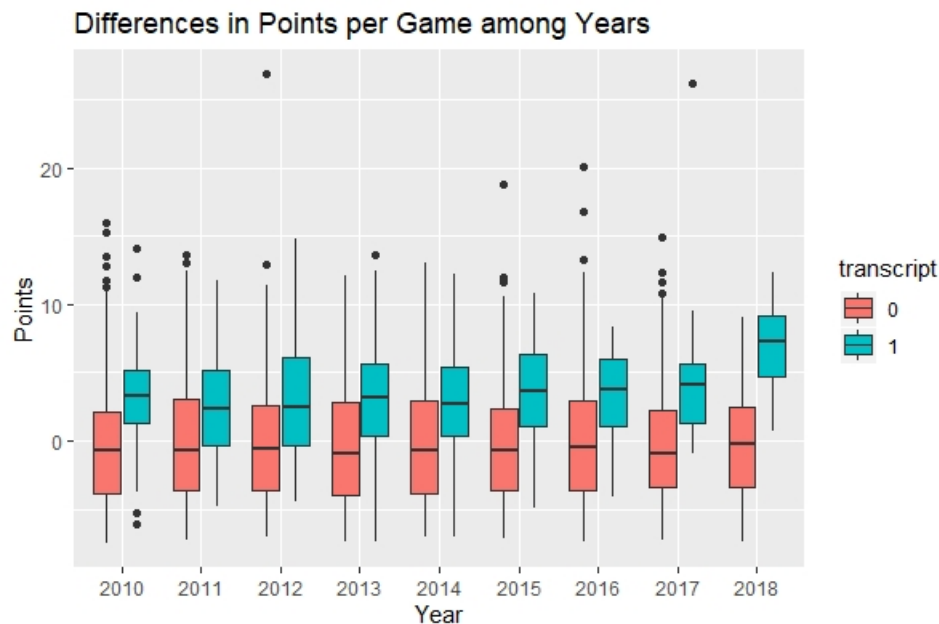


Figure 2.1: Differences in Points per game among years

average Eurocup player for each season.

From Figure 2.1 it is clear that players who made a transcription tend to score more points per game than players without a transfer (t student's test, $p\text{-value} \ll 0.01$). In year 2018 there is a significant raise in the points per game occurred for the players who made a transcription. An interesting point in this figure however, is that most of the best players, according to points per game statistic, did not manage to transfer (see the outliers of the red boxplots).

Assists in Figure 2.2 do not seem to discriminate the two player groups despite the fact that players got transferred tend to have higher number of assists (student t-test, $p\text{-value} \ll 0.01$). In addition, according to assists per game, in almost half seasons, players that got transferred were not better than the average.

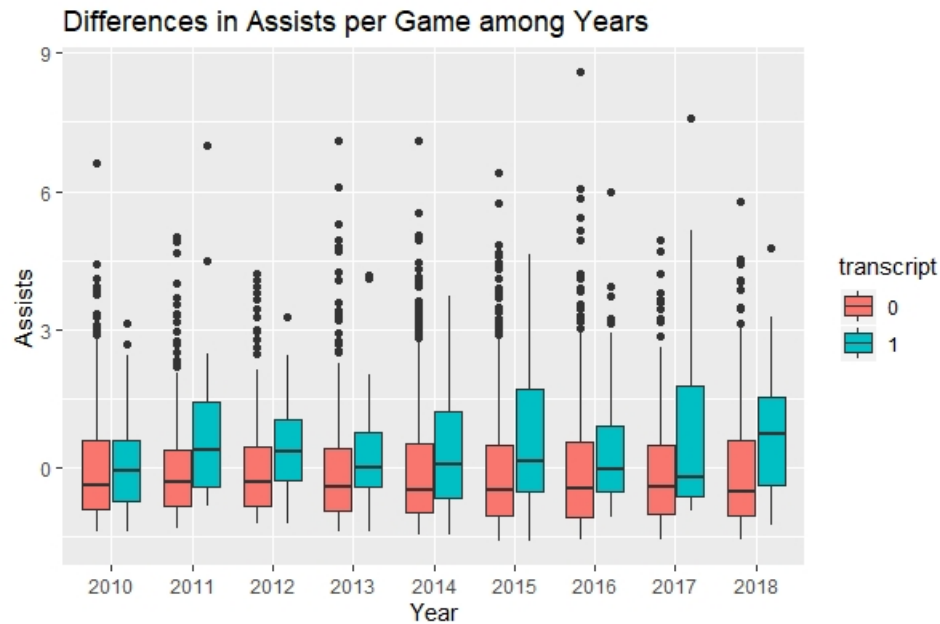


Figure 2.2: Differences in Assists per game among years

It is also very interesting (and quite misleading at first) the fact that players who made a transcription, had more turnovers than both the average and the players who didn't make a transcription (student t-test, p -value $\ll 0.001$); see Figure 2.3. An explanation might be that players who score more points, they also have more individual efforts for scoring, so they also have more turnovers. This thought is also supported from the correlogram in Figure 2.9 in which we can see high linear correlation between turnovers and points.

From Figure 2.4, it can be concluded that players got transfer achieved a higher number of offensive rebounds (student t-test, p -value $\ll 0.001$). On the other, offensive rebounds do not differ significantly for the two groups.

From Figures 2.5 and 2.6, we can conclude that offense is the main

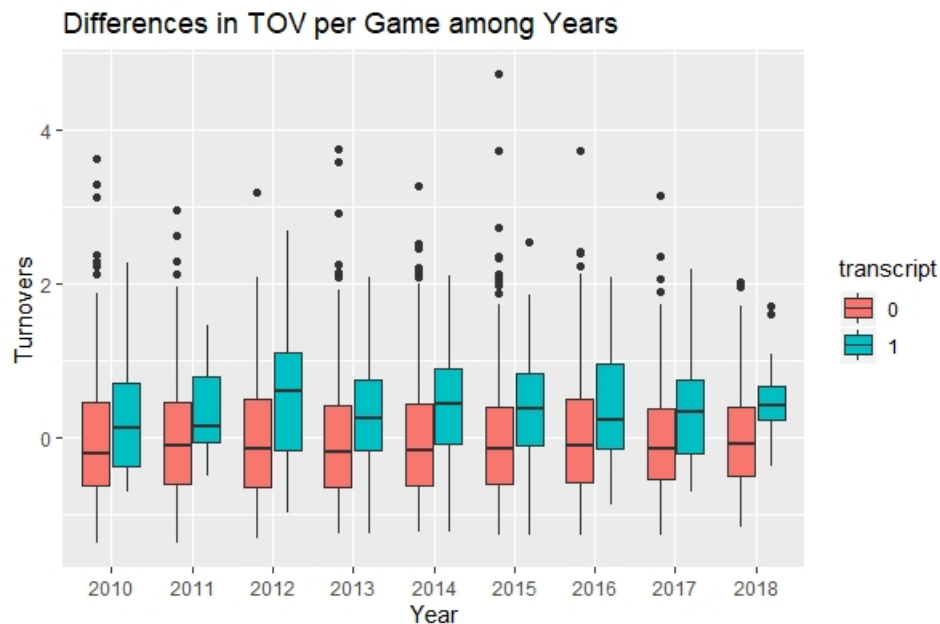


Figure 2.3: Differences in Turnovers per game among years

reason for getting a a successful transfer. Players who got transfered are better in offense than others (student t-test, p-value $\ll 0.001$) and better offensive players than the defensive ones because the hypothesis that both players who made a transcription and players who don't perform pretty much the same in defense, couldn't be rejected (student t-test, p-value $\ll 0.001$).

From figure 2.7 we can confirm that got transfered participated in games with a significantly higher number of minutes (student t-test, p-value $\ll 0.001$). In fact, the number of minutes played show the value of each player for his team. The more he plays, the bigger his playing time value. Also, for as long a player appears in the court, more data about his performance are available can be extracted and therefore scouts can have a clearer picture about each player's skills .

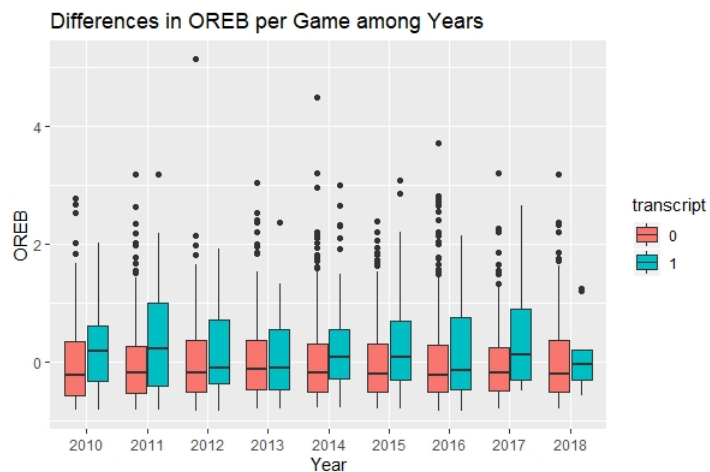


Figure 2.4: Differences in Offensive Rebounds per game among years

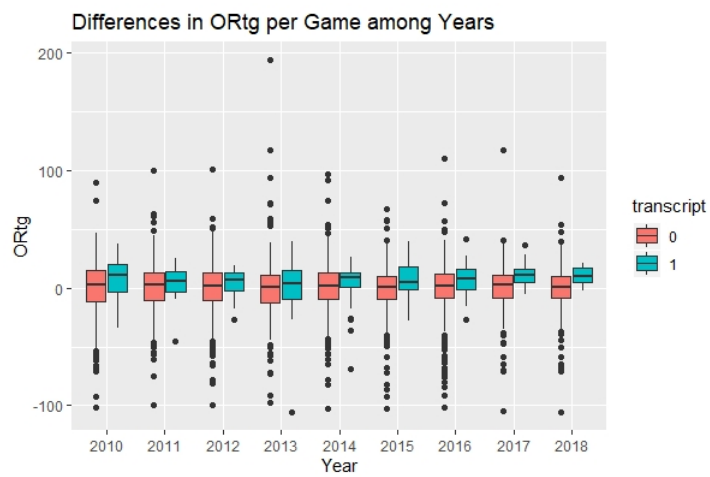


Figure 2.5: Differences in Offensive Rating per game among years



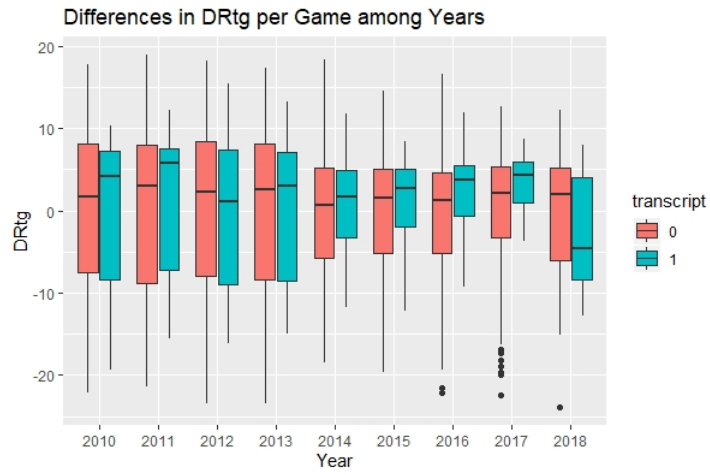


Figure 2.6: Differences in Defensive Rating per game among years

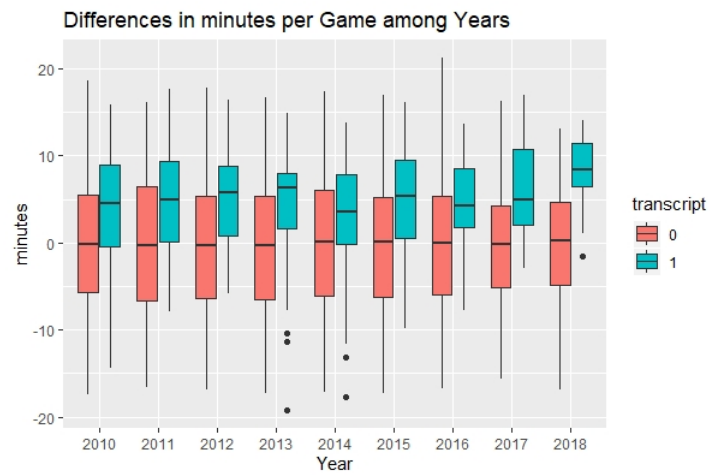


Figure 2.7: Differences in Minutes per game among years



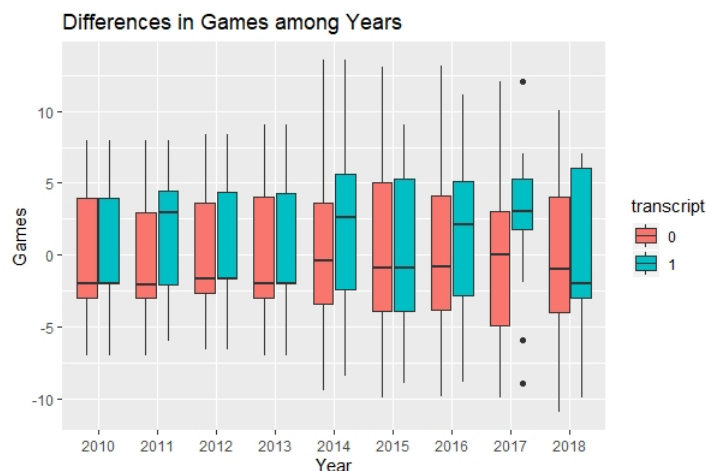


Figure 2.8: Differences in Games among years

Finally, from Figure 2.8, we reach the conclusion that the number of games a player participated is not so informative about the probability of getting transferred. The reason for this is the high variance of this measure. Nevertheless the difference between the two groups was found to be significant (student t-test, $p\text{-value} \ll 0.001$), implying that players that got transferred play on average more than players that remained in Eurocup

Three players which appeared as outliers in year 2017 are quite interesting. The first was Pierre Oriola, whose performance was incredible at that year. He also won the Spanish League 2016–17 season championship with Valencia Basket. His rights, after the end of the season, were sold to Barcelona. The second was Kenny Gabriel. On August 8, 2016, Gabriel signed a two-year deal with Russian club Lokomotiv Kuban. On November 24, 2016, he left Lokomotiv and signed with Greek club Panathinaikos for the rest of the season. The third was Pierre Jackson. Pierre played only for one month in Cedevita because he was reacquired by his older NBA G League team, Texas Legends. On July 14, 2017, he signed a one-year deal, with an option for another season, with Maccabi Tel Aviv. Their statistics

are shown in Table 2.3.

Table 2.3: Per Game Statistics of outlier observations in figure 2.8

	pierre-jackson	kenny-gabriel	pierre-oriola
league	eurocup	eurocup	eurocup
year	2017	2017	2017
team	cedevita	lokomotiv-kuban	valencia-basket
games	2	5	23
MIN	16.84	34.14	15.32
PTS	6.80	33.50	8.26
FG2M	1.6	5.5	3.0
FG2A	2.60	10.0	4.78
FG3M	1.00	4.50	0.22
FG3A	3.00	7.50	0.43
FTM	0.60	9.00	1.61
FTA	1.00	11.00	2.09
OREB	0.40	1.00	1.17
DREB	1.80	2.50	1.48
TREB	2.20	3.50	2.65
AST	1.20	6.00	0.74
PF	2.60	3.00	2.91
BLK	0.40	0.00	0.30
STL	1.20	1.00	0.35
TOV	0.60	3.00	0.78
FG2%	0.62	0.55	0.63
FG3%	0.33	0.60	0.50
FT%	0.60	0.82	0.77



Finally, Figure 2.9 presents a correlogram for the linear correlations between all variables we consider here. Many of the variables were highly positive correlated, except for the defensive rating which had a strong negative correlation with Stop Percentage. this relation is reasonable since the formula for constructing the individual DRtg is a function of $(1-\text{Stop}\%)$.

The previous analysis is made in order to identify useful characteristics of players who succeed in transferring. It is clear that players who made a transfer tend to score more points and play more minutes than players who did not make a transfer. Also, despite the fact that players who get transferred are better in Offensive Rating, the best players according to Offensive Rating were not players who got transferred. This indicates that a good performance in offense alone is not enough for get transferred.



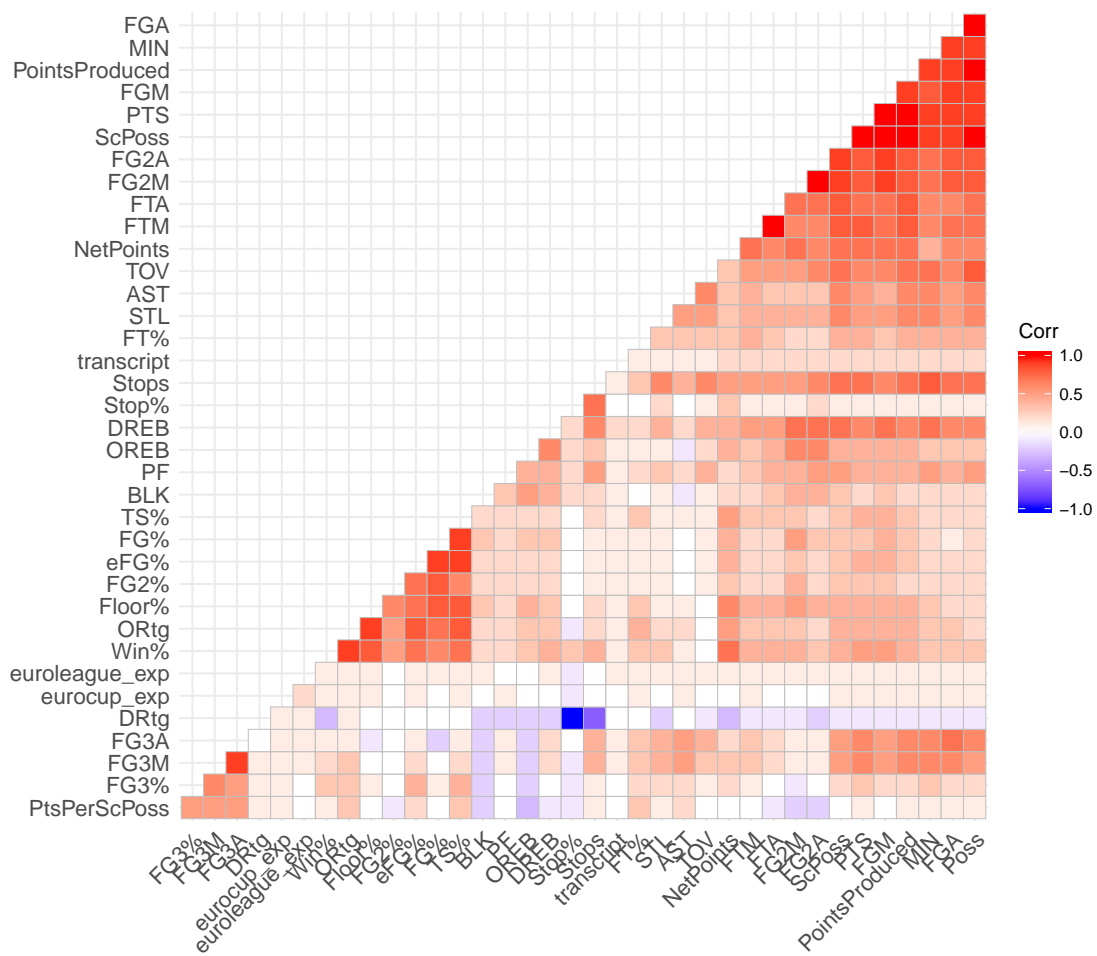


Figure 2.9: Linear correlations between variables of the dataset under consideration





Chapter 3

Statistical Modelling

The purpose of this thesis is to quantify the impact of Eurocup player statistics on a successful transfer to a team participating in Euroleague. But, first we need to clarify and specify what should be considered as a successful transfer. It's not an easy task to evaluate the success of a transfer, because in order to do so we need to consider the reason of a transfer, the quality of the team and the role of the player. So here we focus on the prediction of a transfer and then in the prediction of the 'quality' of it.

3.1 Logistic Regression

As the problem was related to the classification of the players who made a transfer, the initial idea was to use a simple logistic regression to model the probability of transfer. Logistic regression models are the most common models for binary response variables and classification problems. So, as in generalised linear models (GLM) terminology, we consider as a response variable a binary indicator of 'transfer', which indicates when a player made a transfer or not, taking the values 1 and 0, respectively. More



formally:

Define Y_i as,

$$Y_i = \begin{cases} 1, & \text{if the } i\text{-th player was transferred to a Euroleague team} \\ 0, & \text{otherwise} \end{cases}$$

and π_i is the success probability of Y_i . So the model is,

$$Y_i \sim Be(\pi_i)$$

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p}$$

where $X_{i,1}, X_{i,2}, \dots, X_{i,p}$ are called the linear predictors and they were the per game statistics deviation from the average player of season statistics for the i -th player. The summary of the model is presented in table 3.1.



Table 3.1: Table of Coefficients for the full logistic regression model

	Estimate	Std. Er- ror	z value	Pr(> z)
(Intercept)	-2.37	0.18	-13.21	0.00*
games	0.03	0.02	1.65	0.10
MIN	0.05	0.08	0.65	0.52
PTS	1.08	2.44	0.44	0.66
FG2M	-3.57	4.91	-0.73	0.47
FG2A	1.35	0.37	3.60	0.00*
FG3M	-5.01	7.34	-0.68	0.49
FG3A	1.13	0.37	3.06	0.00*
FTM	-2.44	2.47	-0.99	0.32
FTA	1.05	0.41	2.54	0.01*
OREB	-0.20	0.20	-1.01	0.31
DREB	0.07	0.07	0.93	0.35
AST	-0.15	0.18	-0.83	0.41
PF	0.06	0.10	0.64	0.52
BLK	0.42	0.19	2.20	0.03*
STL	0.11	0.17	0.65	0.52
TOV	1.72	0.50	3.44	0.00*
'FG2%'	0.50	1.11	0.45	0.65
'FG3%'	-1.15	0.57	-2.02	0.04*
'FT%'	0.44	0.53	0.83	0.40
'FG%'	-4.17	7.85	-0.53	0.60
'TS%'	-2.71	4.14	-0.65	0.51
'eFG%'	6.51	7.83	0.83	0.41
ORtg	-0.04	0.06	-0.68	0.50

Table 3.1: Table of Coefficients for the full logistic regression model (continued 2/2)

	Estimate	Std. Er- ror	z value	Pr(> z)
DRtg	0.63	0.31	2.00	0.05*
‘Floor%’	2.42	10.33	0.23	0.82
ScPoss	-0.49	1.97	-0.25	0.80
Poss	-1.54	0.54	-2.85	0.00*
‘Stop%’	27.01	13.85	1.95	0.05*
Stops	-0.21	0.17	-1.19	0.23
PtsPerScPoss	0.12	0.68	0.18	0.86
PointsProduced	1.04	1.03	1.01	0.31
NetPoints	0.28	0.29	0.95	0.34
‘Win%’	2.50	2.06	1.21	0.22
eurocup_exp	0.10	0.06	1.62	0.11
euroleague_exp	0.17	0.05	3.35	0.00*
year`2011	-0.70	0.28	-2.49	0.01*
year`2012	-0.47	0.26	-1.80	0.07
year`2013	-0.16	0.25	-0.64	0.52
year`2014	-0.29	0.22	-1.30	0.19
year`2015	-0.38	0.24	-1.60	0.11
year`2016	-0.81	0.26	-3.17	0.00*
year`2017	-0.64	0.31	-2.09	0.04*
year`2018	-1.15	0.34	-3.37	0.00*

Due to multicollinearity issues, the algorithm for this model couldn't define the coefficients of TREB, FGM and FGA. Also, multicollinearity issue, resulted to high standard errors for the estimates of coefficients. Finally, the Deviance of this model was $D = 1931.7$ and the Akaike's Information Criterion was $AIC = 2019.7$

A quick look at the estimates of this model can be misleading and very confusing as well. For example, the estimated coefficient of two-pointers (FG2M) was -3.57, which indicates a negative effect to the probability of making a transcription. Precisely, if a player in 2010, at the end of that season, was better in FG2M by one two-pointer from an average player of that year, then the probability for the average player to make a transcription was about 9% while for the player with the additional two-pointer FG2M is 0.3%.

This might be explained by the fact that two-pointers result in more attempts. Therefore, more turnovers and points will be made. All these three variables for example had positive estimates of their coefficients which means that they have an overall positive impact to the probability of transfer as expected. However, the correlation coefficient of turnovers and points is equal to 0.63, which indicates that despite the fact that turnovers are a negative element for the performance of a player in basketball, they represent an important part of a player's effort to score more points.

The above problems arise from the fact that many performance indicators. A remedy for the multicollinearity issue is LASSO (Least Absolute Shrinkage and Selection Operator) by Tibshirani (2011) which is both a regularization and a variable screening method. By shrinkage of regression coefficients to zero, LASSO can be also used for variable selection. By this way, it enhances the prediction accuracy of the model it produces. LASSO indicated that the solution to this problem was the null model (the model with only intercept as covariate). But, a model with non zero explanatory



variables was preferable as we strongly believe that statistics influence the probability of transcription.

Further solution of the multicollinearity issue was not proceeded, as this model violates another basic assumption that is made in order to perform a GLM analysis.

Common assumption when fitting a GLM is that observations should be independent identically distributed (Dobson and Barnett, 2008), which was clearly not the case in this survey. The sample was consisted of observations of the same individuals and these observations can't be independent. Imagine the case, that we have statistics for a player of Eurocup for some years and every year that player performs better and better. When a player has a good performance for one year, it is not expected to perform bad the following year. Although, fluctuations in performance due to factors like psychology issues or injuries are expected, the expectation of a huge change in performance is not contracted, especially in players who are not very young and enough data for them exist. The reason for this discussion is that the correlation in these observations couldn't be ignored.

Another problem was the assumption of fixed coefficients across seasons. A quick look at Figures 2.1 to 2.8 at the previous section also indicated that there were fluctuations from a common value for all coefficients. There is high uncertainty around the impact of each statistic across different seasons and this uncertainty is not captured from this model, but further analysis was needed in order to gain more precision about this fact.

3.2 Logistic Regression per Year

In order to solve the problem of dependency of observations we analyze the data for each season separately. So, a logistic regression in every year separately of the data was fitted and in that way each data can be



assumed as independent observations of eurocup players, avoiding by this way across season correlation. Variable selection by AIC (Akaike's Information Criterion) by Akaike (1974) was used to identify significant predictors.

AIC estimates the relative amount of information lost by a given model, so the smaller the AIC is the better the model becomes. Table 3.2 present covariates selected by the minimum AIC procedure alone with their estimates. At the bottom of this table, the values of the Deviance measure for the null model of that season, the full model and for the minimum AIC model, are also presented. Performing a hypothesis testing that AIC model is the same as the full model using a chi-square goodness of fit test, ended up with the conclusion that this hypothesis couldn't be rejected for all years. The p-values of the tests can be found at the bottom of Table 4 as well as the AIC values of the selected models.

It's very interesting that none of the variables was selected as explanatory variable for every model. This fact indicates that the mechanism behind the transcriptions may differ between years, at least as concerns the player's statistics. The most important variables seem to be Points Produced and Net Points because they were found to be import in 6 seasons out of 9 and they were always statistically significant. As you can see, the coefficients of Points Produced were negative for three out of six seasons. This was not due to the change of the effect among years from positive to negative, but from the high correlation between explanatory variables, which was not be solved through the variable selection method was used. Patterns in variable selection procedure are observed, as the variables FG2A, FG3A, FTM and TOV were selected either all or none. Also, each time these variables were selected, Points Produced and Net Points were also selected. All of these variables are high linear correlated.



Table 3.2: Coefficients of 'best' models according to AIC (1/2)

	2010	2011	2012	2013	2014	2015	2016	2017	2018
Intercept	-2.84*	-3.24*	-3.33*	-2.93*	-2.83*	-2.94*	-3.53*	-4.34*	-7.00*
AST	-	-	-	0.99	-	-	-0.72*	-	-1.08*
ORtg	-0.27*	-	-	-	-	0.08*	-	-	-0.12*
PTS	-	-0.49*	-2.43*	-	8.62	-	-	-	1.52*
PF	-	-	-	0.49	-	-	-	-	-1.89*
STL	-	-	-	-	-0.67	-	-	-	2.28
'Win%'	-	-	-	-	-	-6.94*	-	-	8.97
NetPts	-	-	1.79*	2.13*	0.57*	0.45*	-3.66*	6.19*	-
MIN	-	-	0.37*	0.59*	-	0.10*	-0.85*	2.12*	-
Pts-Produced	-	0.67*	-5.89*	-3.95*	1.49*	-	4.14*	-5.93*	-
Stops	-	-	-	-0.87	-	-	1.03*	-3.30*	-
games	0.07	0.10	-	-	0.14*	-	-	-0.13	-
Poss	-	-	-4.84*	-	-3.93*	-	-	2.31*	-
eurocup-exp	-	1.14*	0.93*	-	-	-	-	0.37	-
ScPoss	-	-	15.60*	-	-	-	-	1.64	-
FG3%	-2.92	-	-	-	-	-2.46*	-3.22	-	-
FG2A	-	-	4.00*	1.54*	3.25*	-	-1.55*	-	-
FG3A	-	-	4.17*	1.82*	2.58*	-	-1.89*	-	-
FTM	-	-	1.88*	2.33*	-10.6*	-	-1.26*	-	-
TOV	-	-	6.17*	1.36*	4.68*	-	-2.22*	-	-



Table 3.2: Coefficients of 'best' models (continued 2/2)

	2010	2011	2012	2013	2014	2015	2016	2017	2018
PtsPer- ScPoss	12.86*	-	-	-3.73*	-	-	3.95	-	-
OREB	-	-	-	0.97	-	-	-0.81	-	-
euro- league- exp	-	-	0.91*	-	0.22*	0.22	-	-	-
FG3M	0.90*	-	5.33*	2.95	-28.4*	-	-	-	-
Stop%	-	-	-4.28	-217*	77.25	-	-	-	-
FG2%	3.14*	-	-	-	3.63*	-	-	-	-
FG2M	-	-	-	1.99	-19.9*	-	-	-	-
DRtg	-	-	-	-4.93*	1.73	-	-	-	-
DREB	0.38*	-	-0.43	-	-	-	-	-	-
Floor%	61.63*	-	-	-	-	-	-	-	-
FTA	-	-	-	-	1.72*	-	-	-	-
TS%	-	-	-	-	-5.50*	-	-	-	-
BLK	-	-	-	-	-	-	-	1.34	-
FGA	-	-	-	-	-	-	-	-	-0.94
Null	266.88	178.38	223.68	257.06	432.74	328.15	257.27	143.80	108.42
Dev.									
Full	217.60	131.39	161.98	191.71	344.15	260.59	194.84	77.85	-
Dev.									
AIC	228.67	150.39	170.02	196.61	352.41	275.41	205.15	86.78	50.79
Dev.									
P(X^2)	0.99	0.94	0.99	1.00	0.96	0.95	0.95	0.99	-
AIC	246.67	160.39	202.02	246.67	390.41	289.41	231.15	106.78	66.79



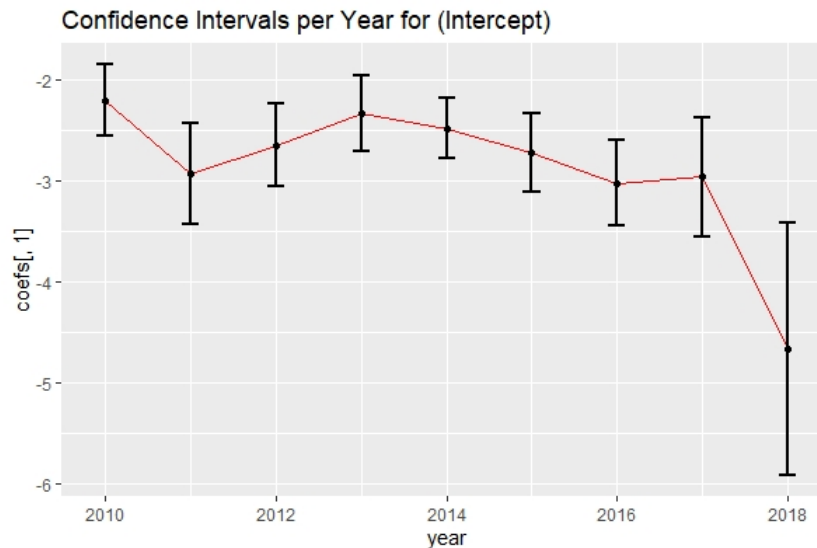


Figure 3.1: Coefficient estimates and 95% confidence intervals for the Intercept

Different coefficient values of all variables, even for the intercept, were resulted. But these models are different from each other so we can't compare the coefficients of them in order to search for a year effect at the coefficients. In order to check where a year effect exists, we fitted the same logistic regression model for each season. Using only Points Produced as explanatory variable at first and Net Points after, the results are summarized in Figures 3.1, 3.2 and 3.3. The intercepts across seasons for both models were identical. From the Figures, it can be assumed that there is an effect of the year observed. This effect may not only have the effect of a statistic or an evaluation measure. It is possible that as the years went by, the average Eurocup player would not have the same acceptance of Euroleague teams. It is reasonable that Euroleague teams, which search for players in lower categories like Eurocup, are interested for real talents rather than average players. It should be kept in mind that the constant of

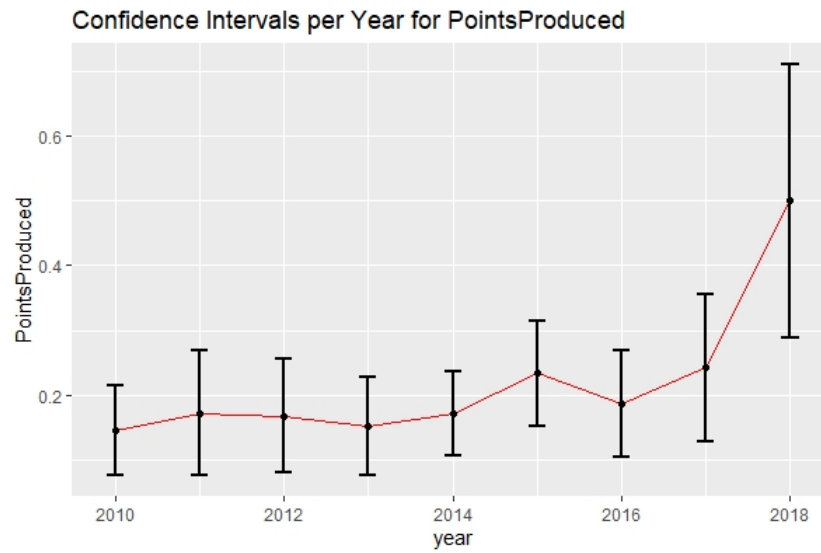


Figure 3.2: Coefficient estimates and 95% confidence intervals for the Points Produced

these models represents, to some extent, the average's player probability for transcription.

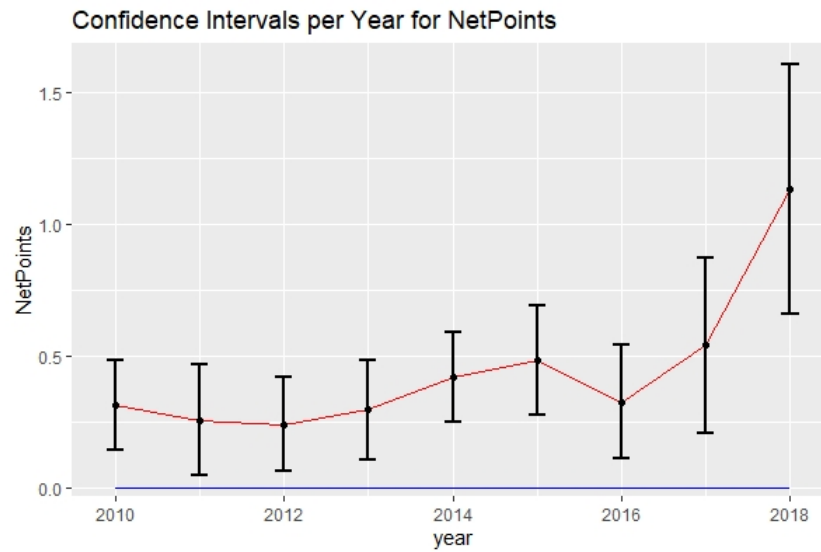


Figure 3.3: Coefficient estimates and 95% confidence intervals for the Net Points

After the construction of these models the interest was in evaluating them, by their predictive accuracy. When it comes to prediction, these models performed very poorly. Each model was used for prediction of the probability of transcription for the next year observations and the results were not very coherent.

Several problems occurred by this method. Firstly, it was difficult to identify a threshold of probability, that discriminate players that made a transcription and players who didn't. Using ROC curves to identify that threshold, straight lines and not curves were emerged. By splitting the data into seasons, important loss was occurred for each player. Despite solving the problem of correlation between observations, those models did not take into account the fact that data were consisted of repeated measurements of the same players. As a result, they performed extremely poorly in predicting the next year's transcription probability.

3.3 Bayesian GLMM

One way to solve the above-mentioned problems in modeling the probability of successful transfer using existing data was to construct a Bayesian Logistic Linear Mixed Model. By using such a model, the correlation between observations of the same subject could be taken into consideration and also it allows to investigate for differences between player's dynamism and year effects. Moreover, by using Bayesian statistics, it was assumed that model's quantities were now random variables and not fixed but unknown constants as were in the previous models. In this way, prior knowledge or ignorance for model's quantities could be defined by using the appropriate prior distributions for the random variables. As a result an entire distribution could be constructed to specify any expectation about model parameters.



Another important issue was that only a few players provide a positive outcome (transfer). Specifically, only 6% of the observations succeed in making a transfer from a Eurocup team to Euroleague team. A slightly higher proportion equal to 10% were made transcriptions from a Eurocup team to another Eurocup team. These type of transcriptions were not the subject under investigation, but they could be helpful for the comparison of model's predictions. It should be mentioned that a transcription from Eurocup to Eurocup could be a degradation for a player.

Due to this sparsity of the data a choice to continue with a case-control study was made, where the case group was the players who made a transfer from Eurocup to Euroleague, while the control group was the rest of the players. In this way, players who succeeded in making a transcription were studied in order to understand the factors that contributed to this outcome, as a first step. The next step, was to compare these factors between the case and control groups.

So, assume that Y_{ijt} is the indicator variable of the success for i -th player in his j -th observation which was at year t , where success is assumed to be the observation in which the player got transferred. The model is:

$$Y_{ijt} \sim \text{Bernoulli}(\pi_{ijt}),$$

$$\log\left(\frac{\pi_{ijt}}{1 - \pi_{ijt}}\right) = \beta_0 + b_i + c_t + \sum_k \beta_k X_{ijk}$$

with prior distributions $\beta_0 \sim N(0, 10)$, $\beta_k \sim N(0, 2.5)$, $b_i \sim N(0, \sigma_b^2)$ and $c_t \sim N(0, \sigma_c^2)$, for $i = 1, \dots, N$, $j = 1, \dots, P$, $t = 2010, \dots, 2018$. N denotes the number of players and P the number of observations of each player.

Using this model, it is assumed that the probability of transfer is determined by the variables X_{ijk} which are the statistics measurements of



player i , in his subsequent observation j , at year t . The probability of making a transcription fluctuates around a constant value depending of the variables X_1, X_2, \dots, X_p but some extra error terms were added in the model. To be more specific, a random effect b_i was used, which captured the between players ability. But, also another random effect c_t was used, which captured the between years variability of the constant term of the model. Remember that the constant term of the model represents by some way the probability of average player in Eurocup to make a transcription to Euroleague. Using this model, it was assumed that a percentage of the total variability of the data was explained by the difference in the capability of each player, but also over time which influences the probability of the player being transcribed. The assumption of independent error terms was also made.

When we talk about binary responses, correlation coefficient is not a natural measure of dependency between observations. So in order to express this correlation, the latent variable approach was used as described by Ntzoufras (2012). So by assuming a latent variable Z_{ijt} exists such that:

$$Z_{ijt} \sim \text{Logistic}(\mu_{ijt}, 1)$$

with

$$\mu_{ijt} = \beta_0 + b_i + c_t + \sum_k \beta_k X_{ijk}$$

and $Y_{ijt} = 1$ if $Z_{ijt} > 0$ and $Y_{ijt} = 0$ otherwise.

By this way, everything could be expressed by the latent variable Z_{ijt} which results to within-player correlation of the latent measurements equal to:

$$r_Z = \frac{\sigma_b^2 + \sigma_c^2}{\sigma_b^2 + \sigma_c^2 + \frac{\pi^2}{3}}$$

where σ_b^2 and σ_c^2 are the variances of the random effects b_i and c_t respectively, which have to be estimated.



With regards to the explanatory variables used in this model, not all of them were used, because it wasn't expected that all of them would be informative and due to the high linear correlation between them. So a variable selection was made according to the maximization of $elpd_{loo}$, which is the expected log pointwise predictive density for a new data set using leave-one-out cross-validation as described by Vehatri et al. (2016). Specifically, their modification to the $elpd_{loo}$ was used which is the $elpd_{PSIS-loo}$ which is an improvement to the LOO estimate. The reason for choosing this criterion was because the evaluation of the candidate models by their predictive accuracy was preferred. Formally it corresponds to:

$$elpd_{loo} = \sum_{i=1}^n \log(y_i|y_{-i}),$$

where

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$

is the leave-one-out predictive density given the data without the i-th data point. The way, the predictive density given the data without the i-th data point is evaluated, is by draws θ^s from the full posterior $p(\theta|y)$ using importance ratios

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \propto \frac{p(\theta^s|y_{-i})}{p(\theta^s|y_i)},$$

Under the assumption that the n data points are conditionally independent as noted by Gelfand, Dey, and Chang (1992). The result was the Pareto Smoothed Importance sampling which gave the:

$$e\hat{l}pd_{PSIS-loo} = \sum_{i=1}^n \log \left(\frac{\sum_{s=1}^S w_i^s p(y_i|\theta^s)}{\sum_{s=1}^S w_i^s} \right)$$

where w_i^s were the importance weights.

Finally, after 16000 iterations of MCMC (Markov Chain Monte Carlo) simulations, the outcome was the model with eight explanatory variables



according to maximization of $elpd_{loo}$ and posterior distributions for coefficients as shown in the Figures 3.4-3.8. The convergence of the algorithm to the posterior distributions is presented in Figure 3.9.

The variable selection was made by a step-wise procedure. All variables were tested for their performance according to the $elpd_{loo}$. Table 3.3 shows the different values of $elpd_{loo}$ for the models with different explanatory variables that maximized the $elpd_{loo}$. These models are defined by the variables that were used by the model as follows:

Model 1: Intercept

Model 2: Intercept, Scoring Possessions

Model 3: Intercept, Scoring Possessions, Free throw Attempts

Model 4: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds

Model 5: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds, Blocks

Model 6: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds, Blocks, Net Points

Model 7: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds, Blocks, Net Points, Floor percentage

Model 8: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds, Blocks, Net Points, Floor percentage, Euroleague experience

Model 9: Intercept, Scoring Possessions, Free throw Attempts, Offensive Rebounds, Blocks, Net Points, Floor percentage, Euroleague experience, Eurocup experience

From the posterior distributions of coefficients and the normality assumption was made as prior beliefs, the conclusion was that except FTA and OREB variables, all the other variables affect the probability of making a transcription with probability of 80%. All of the variables used in this



Table 3.3: The values of $elpd_{loo}$ for nine different models

Model	$elpd_{loo}$
Model 1	-21039
Model 2	-498.5
Model 3	-498
Model 4	-497.8
Model 5	-497.5
Model 6	-497.1
Model 7	-492.8
Model 8	-477.2
Model 9	-464.8

model, were uncorellated and Floor%, Eurocup experience and Euroleague experience have a negative impact to the probability of making a transcription. It may seem inconvenient at first, but one should not forget that an increase in experience in the euroleague and the eurocup also means an increase in age. Also the negative effect of Floor% can be explained like this:

Floor% is defined as $Floor\% = \frac{ScoringPossessions}{Possessions}$. A raise to Floor% is more likely to happen due to a decrease in possessions rather an increase to Scoring Possessions because the standard deviations of them are $sd(Poss) = 3.32$ and $sd(ScPoss) = 1.74$, and probably teams are not fond of this.

The most interesting results of this model were the variables $\hat{\sigma}_b^2 = 0.12^2$ and $\hat{\sigma}_c^2 = 0.31^2$, which led to a correlation estimate of the latent variable Z equal to $r_Z = 0.03$. This means that the random effects did not improve the model as it was confirmed by the $elpd_{loo}$.



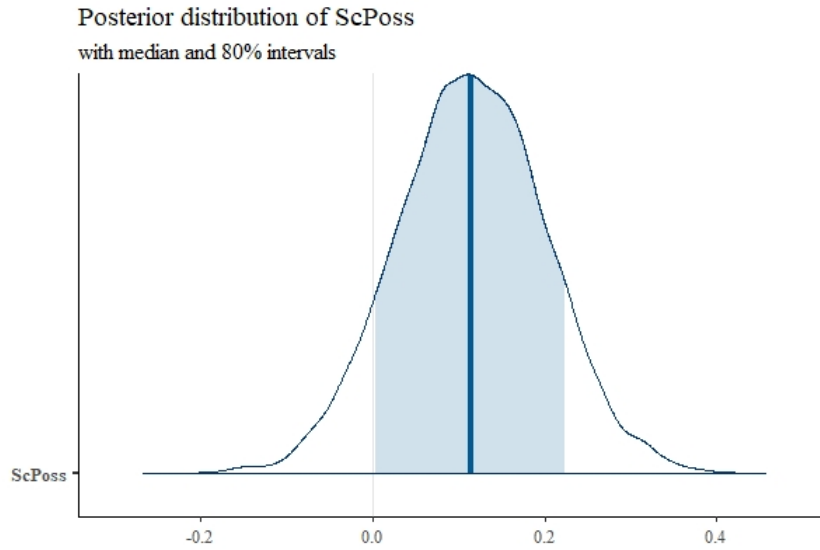


Figure 3.4: Posterior Distribution of ScPoss

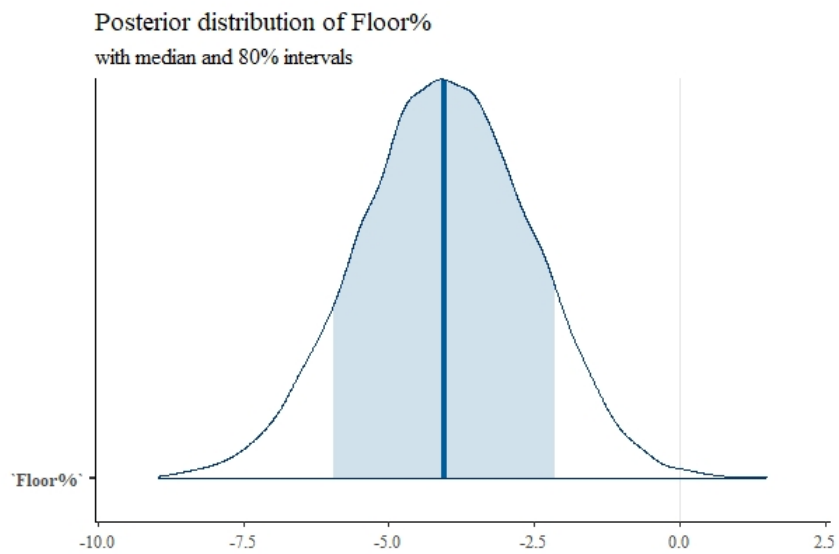


Figure 3.5: Posterior Distribution of Floor%



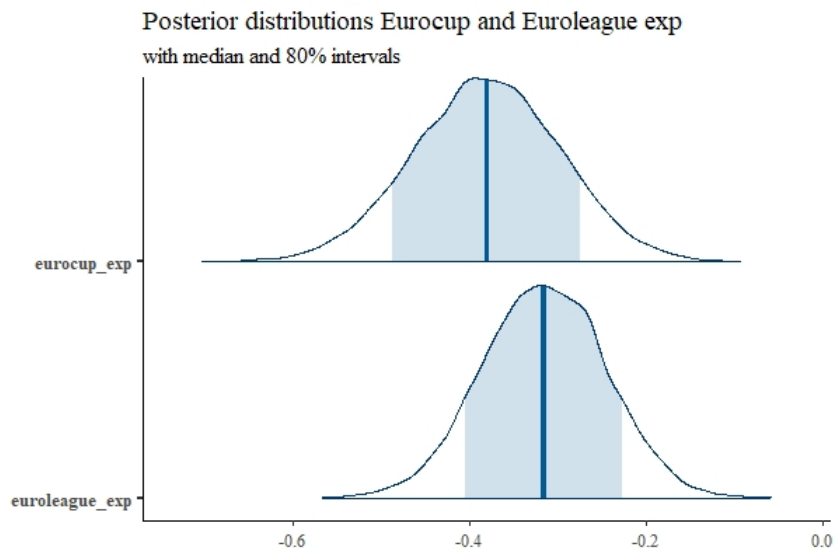


Figure 3.6: Posterior Distribution of Eurocup and Euroleague experience

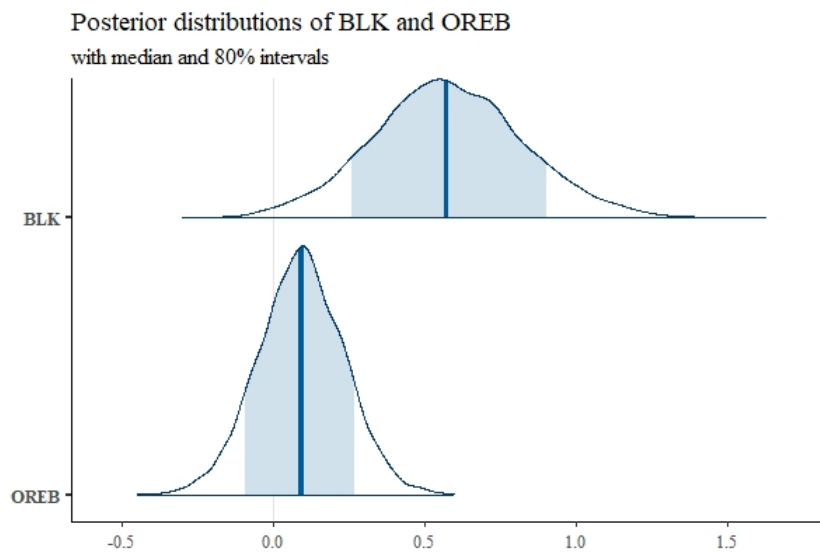


Figure 3.7: Posterior Distribution of BLK and OREB



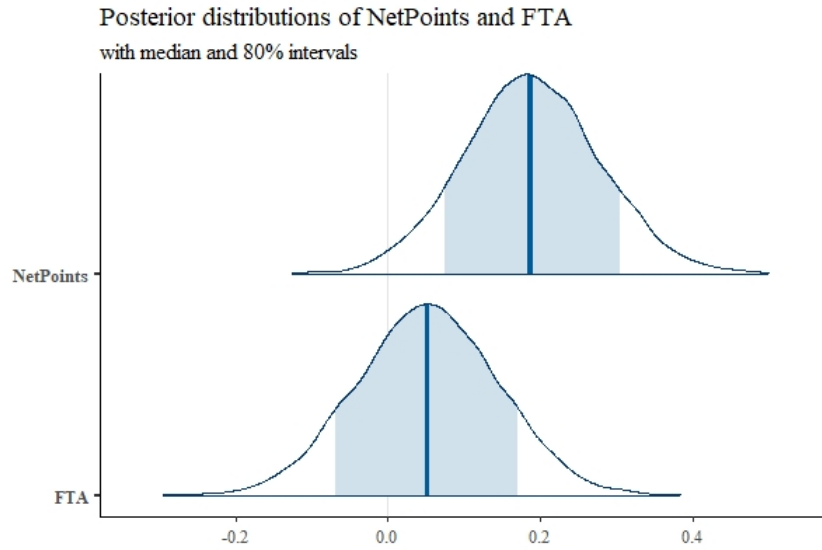


Figure 3.8: Posterior Distributions of NetPoints and FTA

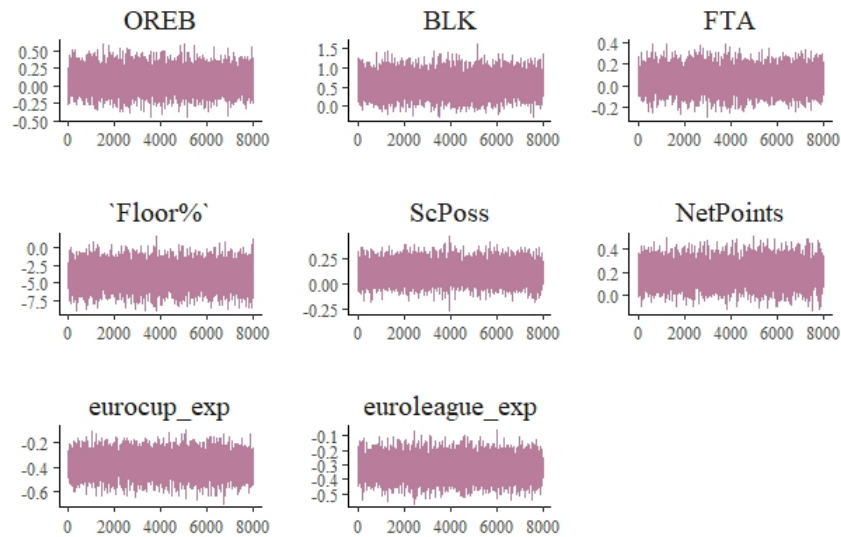


Figure 3.9: Diagnostics of MCMC convergence



In order to understand better the estimated coefficients of the last model three groups were defined, as follows:

Group-1:= Players who made a transfer at this year

Group-2:= Players who made a transfer but not at this year

Group-3:= Players who didn't make a transfer

By considering these groups in the data, Figures 3.10 and 3.11 demonstrate the impact of each variable in the probability of a transfer. A clear difference between the three groups appears in these Figures. Specifically, the third group of players with no transfer performs worse on average than the average player of the league and than the players of the other groups. There is also a big transfer difference between groups 1 and 2 that differ only in the year. What these mean is that in order to make a transfer, being above the average of the league it's not enough. So this is a clear indication that players should considerably outperform. It seems that a difference in skills between players who are going to transfer and those are not exists.

Thus, by taking into account the difference in abilities between the 2 groups of players that made a transfer and players who did not the model is given by:

$$Y_{ijs} \sim Be(\pi_{ijs})$$

$$\log\left(\frac{\pi_{ijs}}{1 - \pi_{ijs}}\right) = \beta_0 + b_s + \sum_k \beta_k X_{ijk}$$

with prior distributions $\beta_0 \sim N(0, 10)$, $\beta_j \sim N(0, 2.5)$ and $b_s \sim N(0, \sigma_b^2)$ and now Y_{ijs} is the j-th observation of i-th subject in group s.

The random intercept b_s represents the random effect between the average players of two group. The explanatory variables used are the same as those of the last model, since the best predictors of the probability of transcribing players who succeeded at transcription at some point emerged. Now, the intraclass correlation or within-groups correlation can be also



expressed by the latent variable approach and is equal to $r_Z = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi^2}{3}}$.

The difference by the last model is that now this model assumes that not only correlation within player measurements exists but in the whole group of this class of players and is equal to r_Z . By fitting this model to complete data we ended up with the same estimates of the covariates and the same posterior distributions except FTA, which is now centered in zero. But, the most interesting estimate is $\sigma_b^2 = 4.3^2$, which leads to $r_Z = 0.85$. By this estimate, the fact resulting was that about 85% of the total variability of the data could be explained from the between groups variability.

The meaning of these results was that the average player who was going to make a transcription at some time of his career was very different from the average player who will never make a transcription.



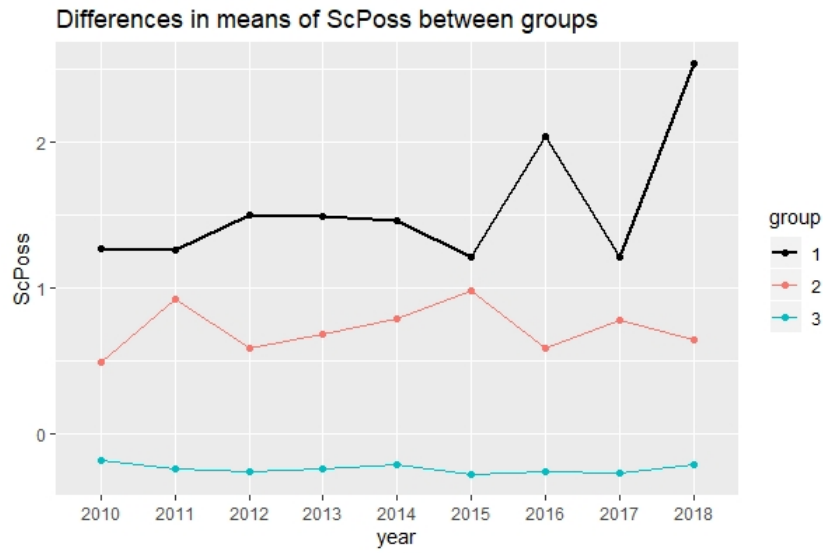


Figure 3.10: Difference in mean of ScPoss from the average player of the year between groups

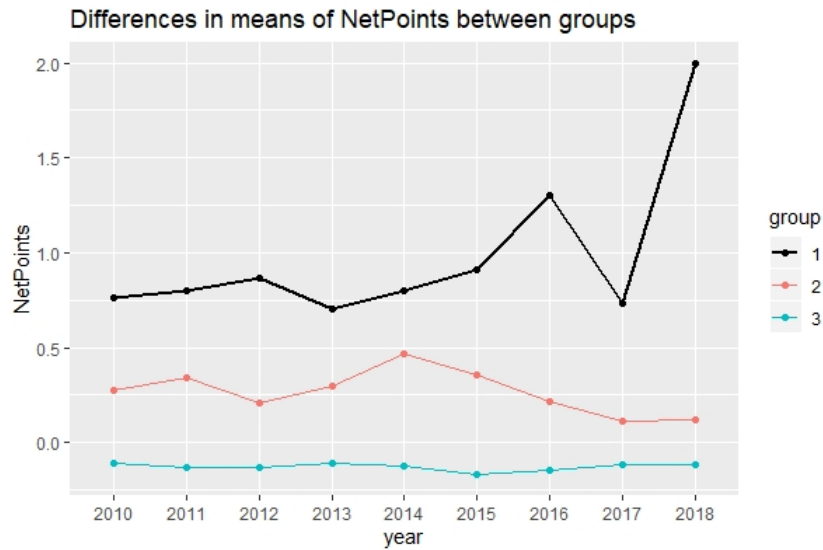


Figure 3.11: Difference in mean of NetPoints from the average player of the year between groups



This difference refers to the general notion of a player's talent. Indeed, there are strong indications that players who succeed at some point in their career have greater overall talent than others. It may be important, however, to take into account everyone's personal abilities in statistics. Maybe, some stats do not have the same impact in player's transcription probability which is very reasonable according to his personal skills.

For example, consider the case of a defender who scores more points than other defenders. It's reasonable that for this player the points he scores has a bigger impact to the transcription probability under the logical assumption that every team and every coach wants defenders with good offensive skills. In order to examine this hypothesis, the same model as before was fitted with difference that random slopes to the explanatory variables were added, so the model formulation was as follows:

$$Y_{ijs} \sim Be(\pi_{ijs})$$

$$\log\left(\frac{\pi_{ijs}}{1 - \pi_{ijs}}\right) = \beta_0 + b_s + \gamma_{i0} + \sum_k (\beta_k + \gamma_{ik}) X_{ijk}$$

with prior distributions $\beta_0 \sim N(0, 10)$, $\beta_k \sim N(0, 2.5)$, $b_s \sim N(0, \sigma_b^2)$, and $\gamma_{.k} \sim N(0, \sigma_{\gamma_k}^2)$.

The estimates of $\sigma_{\gamma_k}^2$ were less than 0.1^2 which indicates that the assumption of different personal impact in the abilities of some players was not true.

After the previous considerations for personal abilities, the hypothesis of some special abilities between groups existed according to their stats, was examined. In order to did so, we fitted a model with random slopes to the explanatory variables, but this time these random slopes indicated different abilities between groups and not between players. So the model was:

$$Y_{ijs} \sim Be(\pi_{ijs})$$



$$\log\left(\frac{\pi_{ijs}}{1 - \pi_{ijs}}\right) = \beta_0 + b_s + \sum_k (\beta_k + \gamma_{sk}) X_{ijk}$$

with prior distributions $\beta_0 \sim N(0, 10)$, $\beta_k \sim N(0, 2.5)$, $b_s \sim N(0, \sigma_b^2)$, and $\gamma_{.k} \sim N(0, \sigma_{\gamma_k}^2)$.

The results can be found at table 3.3, where it appeared that random slopes may improve even better the random intercept model, because the estimated standard deviations of random slopes were high. Therefore, this indicated a difference between the groups' abilities in terms of specific statistics, which can be very important. However, as the goal was to find a model with the best possible predictive accuracy, it was preferable to model with the random constant as the random effect, which was shown to be the best according to elpd' loo.

Table 3.4: Error terms estimation for the model with random intercept and random slopes

Error terms:								
Groups	Name	Standard Deviations	Correlations					
group	(Intercept)	4.8						
	ScPoss	1.9	0.02					
	'Floor%'	2.1	-0.02	-0.01				
	NetPoints	2.0	0.01	-0.01	-0.01			
	BLK	2.1	0.05	0.00	-0.01	0.00		
	OREB	2.0	0.05	-0.01	0.00	0.01	0.00	
	FTA	2.0	0.04	-0.01	-0.01	0.01	0.00	0.00

Returning to the model with only random intercept, some plots are provided in order to understand the power of data into the assumptions. In figure 3.12, it shown how the data influence the prior thoughts about model parameters by providing means and variances both a priori and a posteriori. It's clear that all model coefficients have moved from zero a priori estimates (even a little bit) with low variances which is beneficial.



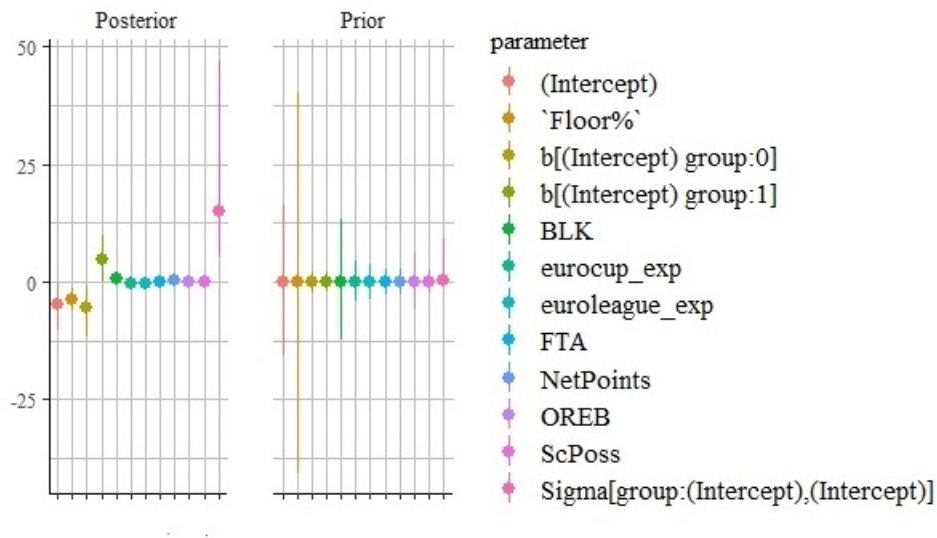


Figure 3.12: Posterior vs Prior for the model with random Intercept

As it was discussed previously, the purpose isn't just to estimate the probability of a transfer, but also to classify players according to their prediction. So, a threshold for the probability of transfer was needed to be specified in order to perform this classification. Table 3.4 and Table 3.5 show that a threshold value equal to 0.36 can capture more than 75% of players who transferred and misclassify the 5.7% of the players who did not manage to get a transfer.

ROC (Receiver Operating Characteristic) curves are widely used at classification problems as an evaluation measure of the classifier. According to Hanley and McNeil (1982), there are graphical plots that illustrate the diagnostic ability of a binary classifier as its discrimination threshold is varied. They are created by plotting the true positive rate (sensitivity) at y-axis and false positive rate (1-specificity) at x-axis. It is clear from Figure 3.13 that a ROC curve recommend a good predictive mechanism, if it is as near as possible to the upper left corner of the box.

Table 3.5: Quantiles of fitted values for players that made a transcription

Quantiles	0%	25%	50%	75%	100%
Probability	0.05	0.37	0.50	0.63	0.97

Table 3.6: Quantiles of fitted values for players that did not make a transcription

Quantiles	0%	25%	50%	75%	94.3%	100%
Probability	0.00	0.00	0.00	0.00	0.36	0.91

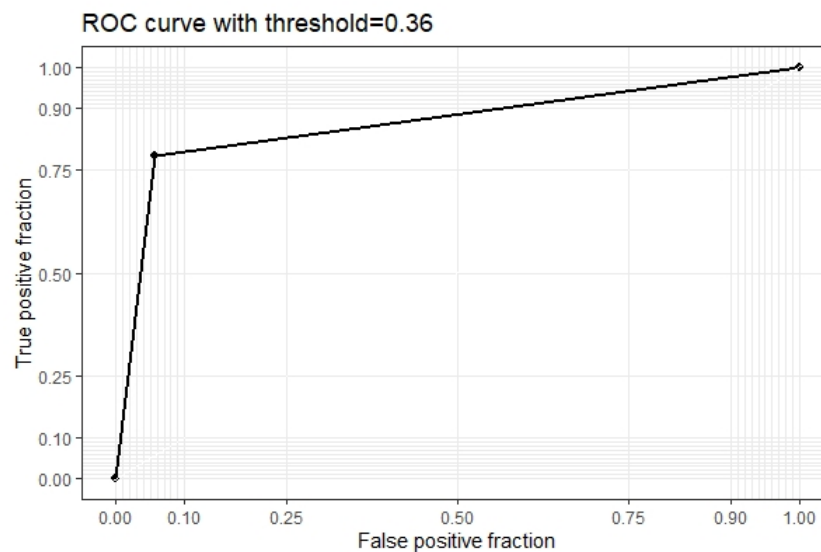


Figure 3.13: ROC curve for the model with random Intercept

3.4 Examples

An example of the model's fitted values for Dairis Bertans from 2010 season until 2018 is given at Tables 3.6, 3.7 and 3.8 where his stats are shown. These statistics had been chosen as predictors for probability of transfer from the best model, according to $elpd_{loo}$.

A successful transfer for this player appeared at year 2016, where Bertans transferred from Bilbao Basket to Darussafaka Basketbol.

In August 2010, Bertans signed a two-year contract with VEF Rīga, after a two seasons with Ventspils. VEF Rīga's coach at the time, Rimas Kurtinaitis, saw a great potential in Bertāns as a point guard despite primarily being a shooting guard. So, during the 2010–11 season, Bertāns developed point guard skills to a different level, and was one of the key factors in VEF Rīga's first championship. In Table 3.6 you can see a big increase in his scoring possessions. In July 2012, he re-signed with VEF Rīga on a three-year deal.

In June 2013, Bertāns parted ways with VEF Rīga to sign a three-year deal with Bilbao Basket of the Liga ACB. In one of his first games with Bilbao he scored 19 points in a preseason game against Philadelphia 76ers. In July 2014, Bertāns joined the Boston Celtics for the 2014 NBA Summer League. In his second season with Bilbao, Bertāns helped the team to reach the 2015 ACB Playoffs as the fifth seed, but they eventually lost to Valencia in the Quarterfinals. In July 2015, Bertāns joined the San Antonio Spurs for the 2015 NBA Summer League, where he averaged 11.3 points, 3 rebounds and 1.6 in three games played for the Spurs. On January 20, 2016, Bertāns recorded a season-high 27 points, shooting 6-of-11 from three-point range, along with four rebounds and two assists in a 76-78 loss to Bayern Munich.

On July 13, 2016, Bertāns signed a 1+1-year deal with Turkish club



Darüşşafaka Doğuş under head coach David Blatt. On January 12, 2017, Bertāns recorded a career-high 29 points, shooting 10-of-13 from the field, along with five assists in a 98–89 win over Baskonia. Bertāns helped the team to reach the 2017 EuroLeague Playoffs as the eighth seed, but they eventually were eliminated by Real Madrid in the Quarterfinals. On July 10, 2017, Bertāns signed with Italian club Olimpia Milano. In his first season with Milano, Bertāns helped Milano to win the 2018 Italian League championship. On June 29, 2018, Bertāns re-signed with Milano for the 2018–19 season. However, on March 1, 2019, Bertāns parted ways with Milano so he can continue the season in the NBA.

Table 3.7: Per game Statistics differences from the average player of Dairis Bertans

	Year	2010	2011	2012	2013
player	dairis-bertans	dairis-bertans	dairis-bertans	dairis-bertans	dairis-bertans
league	eurocup	eurocup	eurocup	eurocup	eurocup
team	bk-ventspils	bc-vef-riga	bc-vef-riga	bc-vef-riga	bc-vef-riga
FTA	0.94	-0.09	0.39	0.30	0.30
OREB	-0.32	-0.65	0.16	-0.46	-0.46
BLK	-0.23	-0.23	-0.26	-0.20	-0.20
Floor%	-0.10	-0.07	-0.05	-0.02	-0.02
ScPoss	0.61	1.18	1.10	1.04	1.04
eurocup_exp	0.00	0.74	1.56	2.31	2.31
euroleague_exp	-0.01	-0.13	-0.26	-0.48	-0.48
transcript	no	no	no	no	no
tr_eurocup	yes	no	no	yes	yes
prob	0.42	0.38	0.30	0.32	0.32



Table 3.8: Per game Statistics differences from the average player of Dairis

Bertans

Year	2014	2016	2017	2018
player	dairis-bertans	dairis-bertans	dairis-bertans	dairis-bertans
league	eurocup	eurocup	euroleague	euroleague
team	retabet-bilbao-basket	retabet-bilbao-basket	darussafaka-basketbol-istanbul	ea7-emporio-armani-milano
FTA	1.32	0.95	0.34	-0.52
OREB	-0.50	-0.22	-0.41	-0.40
BLK	-0.15	-0.04	-0.12	-0.13
Floor%	0.00	0.12	0.06	0.00
ScPoss	1.14	2.20	-0.25	-0.31
eurocup_exp	3.33	3.78	4.27	4.01
euroleague_exp	-0.58	-1.07	-2.47	-1.55
transcript	no	yes		
tr_eurocup	no	no	no	no
prob	0.21	0.29		



Another very interesting example is Brian Randle, for whom the model gave one of the highest probabilities for transfer. Brian Randle signed with Hapoel Jerusalem on a two years contract at 2010, so he could not made a transcription in 2011.

Table 3.9: Per game Statistics differences from the average player of Brian Randle

	2011	2012	2013	2014
player	brian-randle	brian-randle	brian-randle	brian-randle
league	eurocup	eurocup	euroleague	eurocup
team	hapoel-jerusalem	hapoel-jerusalem	alba-berlin	maccabi-bazan-haifa
FTA	3.41	0.62	0.53	0.75
OREB	2.02	0.66	0.27	0.64
BLK	2.11	1.24	0.30	1.21
Floor%	0.12	0.09	-0.10	0.06
ScPoss	3.57	2.94	-1.06	2.25
eurocup_exp	-0.26	0.56	1.28	1.33
euroleague_exp	-0.13	-0.26	-1.05	0.42
transcript	no	yes		yes
tr_eurocup	no	no		no
prob	0.91	0.74		0.52



Table 3.10: Per game Statistics differences from the average player of **Brian Randle**

	2015	2016	2016	2017
player	brian-randle	brian-randle	brian-randle	brian-randle
league	euroleague	euroleague	eurocup	eurocup
team	maccabi-tel-aviv	maccabi-tel-aviv	maccabi-tel-aviv	hapoel-jerusalem
FTA	0.84	1.01	-0.55	-0.17
OREB	1.17	0.15	0.41	0.21
BLK	1.01	0.54	0.52	0.45
Floor%	0.12	0.09	0.22	0.13
ScPoss	2.81	0.84	-0.31	-0.03
eurocup_exp	1.72	1.34	1.78	2.55
euroleague_exp	-0.37	0.24	1.93	1.83
transcript			no	no
tr_eurocup			yes	no
prob			0.14	0.14



Chapter 4

Machine Learning Modelling

The purpose of this section is firstly to introduce some basic algorithms for classification and secondly to implement a machine learning algorithm for classification of players according to their transfer status. A variety of classification methods are available within the machine learning context. In statistics the most known techniques for classification are the logistic regression and linear or quadratic discriminant analysis. Some of the most popular techniques in machine learning are presented in the next section.

4.1 Machine Learning Classifiers

4.1.1 Support Vector Machines (SVM)

One of the most known classifier in machine learning is the Support Vector Machine (SVM). The SVM is an extension of the support vector classifier algorithm. SVM additionally deals with classification problem, where classes can not be separated by linear boundaries. This is achieved by enlarging the feature space in a specific way, using kernels. Specifically the SVM deal with the maximization problem:



$$\begin{aligned}
 & \underset{\beta_{ji}, e_i, \forall i, j}{\text{maximize}} && M \\
 & \text{subject to} && y_i f(x_i) \geq M(1 - e_i), \\
 & && \sum_{i=1}^n e_i \leq C, e_i \geq 0
 \end{aligned}$$

where $f(x) = \beta_0 + \sum_i \alpha_i K(x_{i'}, x_i)$ and $K(x_{i'}, x_i)$ is the kernel function. Some popular choices of the kernel are:

linear kernel: $K(x_{i'}, x_i) = \sum_j x_{i'j} x_{ij}$

polynomial kernel: $K(x_{i'}, x_i) = \sum_j (1 + x_{i'j} x_{ij})^d$

radial kernel: $K(x_{i'}, x_i) = \exp(-\gamma \sum_j (x_{ij} - x_{i'j})^2)$

For more details see James et al. (2013).

SVM algorithm has been used in various problems and occasions in basketball analytics; see for example Pai, ChangLiao and Lin (2016). They used a hybrid model of SVM and decision trees approaches (HSVMDT) for predicting the outcome of NBA games.

4.1.2 Decision Trees, Bagging and Random Forests

There is a large number of classification and regression algorithms in machine learning based on decision trees and their expansions, that is Random Forests, Bagging and Boosting. Random Forests are maybe the most common ML algorithm for both regression and classification problems. These algorithms are well known for their predictive accuracy and have many applications in basketball analytics. For example King (2017), used random forests in order to predict the NBA game attendance using random forests.

In the following sections, a brief explanation of these algorithms is provided.



4.1.2.1 Decision Trees

In machine learning the simplest algorithms for classification problem are decision trees. A decision tree algorithm divide the predictor space in distinct and not-overlapping regions according to a specified criterion and then predict the same outcome for all observations that fall to the same region. The minimization of classification error rate is a natural criterion for splitting the predictor space, but in practise two other measures are preferable due to their better properties which are: Assuming that \hat{p}_{mk} represents the proportion of observations in the m-th region that are from k-th class. In our problem we have two classes, that of players who made a transcription and players who didn't. Then the Gini Index is defined by

$$G = \sum_1^K \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

and is a measure of total variance among the K classes and its alternative measure Cross-Entropy which is defined by

$$D = - \sum_1^K \hat{p}_{mk} \log \hat{p}_{mk}$$

In fact, both measures are quite similar numerically.

4.1.2.2 Bagging and Random Forests

The most important problem with decision trees is that they suffer from high variance, which means that if we split the data at random and perform a decision tree modelling to every new dataset, the results can be quite different. In order to solve this problem, we can generate by bootstrapping multiple training sets by taking repeated samples from the existing training set, fit a decision tree model to every dataset and finally take the average of all these expectations. This is called bagging or bootstrap aggregation.



Bagging improves dramatically the predictions over decision trees, but it yields to serious correlation between the bootstrap trees in the manner that the same predictors and the same sequence of those will be used to the construction of every tree. So, at most times we end up to very similar trees. Random forests overcome this problem by taking a random sample of predictors in order to fit a decision tree. The most common number of predictors used in every tree is \sqrt{p} , assuming that the total number of predictors we have is p .

Also, after the termination of the algorithm, we can compare the explanatory variables as for their accuracy in discriminating the groups by the variable importance measures. The importance of each variable is inversely proportional to the mean decrease in the Gini Index if Gini Index is used for tree grown. For more details see the paper 'Reinforcement Learning Trees' by Zhu R et al.

It must be clear that it is not the purpose here to compare statistical and machine learning techniques, but to evaluate their accuracy in problem tested. Both of them have pros and cons and it was desirable that the most suitable method for this problem would be chosen. There are plenty studies comparing these two approaches in several fields of science; see for example in Makridakis et al. (2017).



4.2 Random Forests implementation for transfer prediction in Eurocup Basketball Data

Initially a random forest model was used and data were splitted into train and test subsamples. The training subsample used to construct the model while the test subsample was used to evaluate its predictive ability.

Due to the lack of any assumption about the independence of observations, or any relationship of interdependence as repeated measurements, a completely random sample of size equal to $2/3$ of the total data was used as train data. The only thing that was taken into consideration in the separation of the data into train and test was the percentage of transfers to be about the same in both subsamples.

Figure 4.1 demonstrates the variable importance plot for this model. What it could be understood from this plot is that Scoring Possessions has the biggest importance for classification of players according to their transcription situation. After Scoring Possessions the most important variables are Blocks, two-pointer field goal attempts, points and Points Produced which means that the most important thing for a player in order to make a transcription is how he performs in scoring himself or in helping his team in scoring points, because Scoring Possessions, Points, Points Produced and two-pointer field goal attempts are high linear correlated.



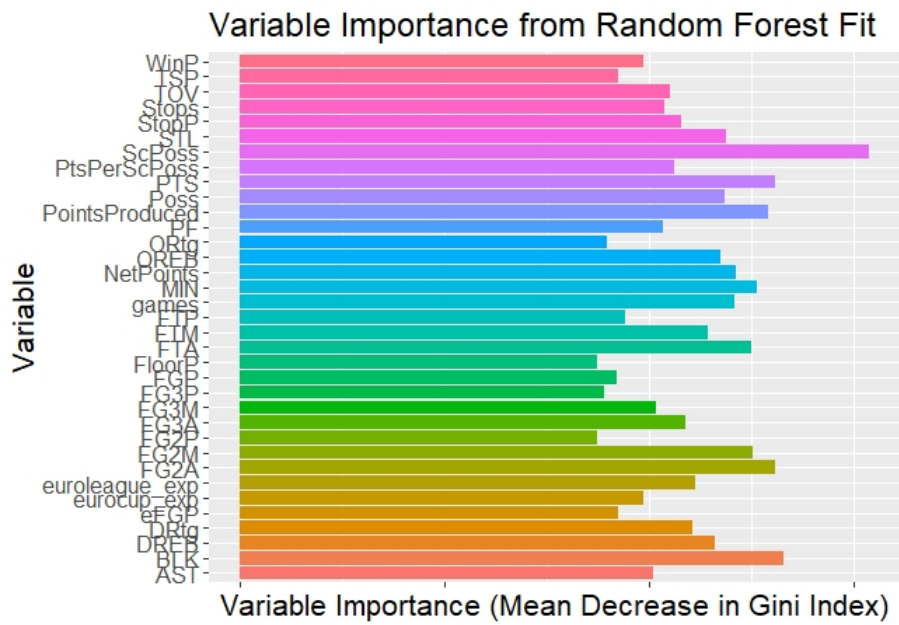


Figure 4.1: Variable Importance Plot for Random Forests in the complete data set



In order to evaluate this model, it was used to predict the observations in the test data. The results are in the next table:

Table 4.1: Predictions of Random Forests Model

Predictions	Actual	
	<i>No Trascription</i>	<i>Trascription</i>
<i>No Trascription</i>	1211	110
<i>Trascription</i>	0	0

This model predicts that none of the test observations will make a transfer which is clearly not the case. In comparison to logistic regression model, described in section 3.1, the random forest model performs worse than the logistic regression model. Assuming inherently a threshold of 0.5 for the probabilities predicted in order to make a transcription, the logistic regression model predicts that no transfers are going to be made except three which are predicted correct. According to these findings, it can be concluded that the logistic regression model is slightly better than the random forests model.

4.2.1 Random Forests per Year

We now fit random forests model to the data of every year in order to understand better how the factor year affects the importance of variables to the probability of transfer. Table 4.2 depicts the most important variables by their effect in the mean decrease of Gini index, for discriminating players according to their transcription success.

The most important variables for each year are colloured with yellow. It is interesting that almost every year the most important variable is affected. For two years the most important variable was defensive rebounds, which was not indicated as so important from the full data analysis.



The misclassification error of the random forest implemented of the annual may not improved.



Table 4.2: Mean Decrease Gini Index over years for Variable Importance

	2010	2011	2012	2013	2014	2015	2016	2017	2018
games	1.01	0.89	0.70	0.84	3.85	1.84	0.63	0.69	0.11
MIN	1.23	1.34	1.36	2.54	2.66	3.03	1.27	0.66	0.70
PTS	2.03	1.21	1.68	2.01	3.11	2.66	1.24	0.53	0.91
FG2M	1.25	0.78	1.13	2.45	4.00	2.50	1.04	1.82	0.32
FG2A	1.33	1.05	1.37	1.85	3.11	3.33	1.33	1.17	0.33
FG3M	1.83	1.47	0.95	1.19	1.53	1.46	0.97	0.28	0.83
FG3A	1.48	1.35	0.91	1.34	1.58	1.70	1.02	0.40	0.67
FTM	1.14	1.53	1.16	1.53	1.94	1.37	0.76	0.83	0.88
FTA	1.11	1.71	1.19	1.55	2.06	1.56	0.93	1.08	1.23
OREB	0.98	1.39	1.04	0.78	3.01	3.21	1.52	0.77	0.19
DREB	1.56	1.15	1.26	1.30	4.14	2.32	1.76	0.84	0.15
AST	1.29	1.04	1.27	1.38	2.10	2.07	1.26	0.66	0.27
PF	2.04	1.09	0.96	1.32	2.00	1.81	1.03	0.25	0.34
BLK	1.27	0.42	1.07	0.99	3.06	1.97	1.06	0.79	0.17
STL	0.92	1.24	1.10	1.19	2.40	2.18	1.26	0.63	0.38
TOV	0.98	0.80	1.49	0.94	2.53	1.97	0.81	0.48	0.27
FG2P	1.41	0.62	1.09	1.49	2.01	1.69	0.75	0.30	0.27
FG3P	1.19	0.72	1.38	1.01	1.38	1.17	0.92	0.21	0.17
FTP	1.11	1.06	0.77	0.96	2.04	1.63	0.92	0.42	0.25
FGP	1.26	0.91	0.90	2.71	1.88	1.63	0.92	0.24	0.19
TSP	1.11	0.78	0.87	1.60	2.07	1.52	0.83	0.32	0.26
eFGP	1.35	0.78	0.94	1.83	2.37	1.60	0.88	0.31	0.16
ORtg	1.01	0.69	1.09	1.35	2.21	1.90	0.84	0.42	0.18
DRtg	1.47	1.07	1.27	1.03	2.66	1.76	1.15	0.60	0.43
FloorP	0.94	0.75	1.06	1.65	2.34	1.75	0.77	0.54	0.15
ScPoss	2.06	1.14	1.48	1.48	3.31	2.55	1.14	1.09	0.44
Poss	1.34	1.48	2.22	1.80	2.23	2.57	1.37	0.91	0.47
StopP	1.52	1.10	1.27	1.17	2.45	1.74	1.19	0.47	0.42
Stops	1.07	1.03	1.91	1.66	2.23	1.92	1.12	0.34	0.40
PtsPerScPoss	1.85	0.86	1.33	0.99	1.89	2.11	1.24	0.44	0.61
PointsProduced	1.57	1.14	1.48	1.70	2.93	2.66	1.41	0.83	0.67
NetPoints	1.94	1.10	1.76	1.72	3.13	1.96	1.48	0.78	0.41
WinP	1.73	0.69	0.99	3.27	2.34	1.77	0.93	0.26	0.18
eurocup_exp	0.00	0.46	0.87	0.38	1.27	0.63	1.02	0.19	0.10
euroleague_exp	0.25	0.29	0.59	0.79	1.00	2.30	0.91	0.22	0.22



4.2.2 Random Forests in Case-Control study

The purpose here is to fit a random forests model only in the case group and then to evaluate this model in the control group. By this way, it is believed that a better variable importance decision is going to be made. Figure 4.2, the variable importance plot for this model can be observed, which makes apparent that Euroleague and Eurocup experience are the most important variables for discrimination.

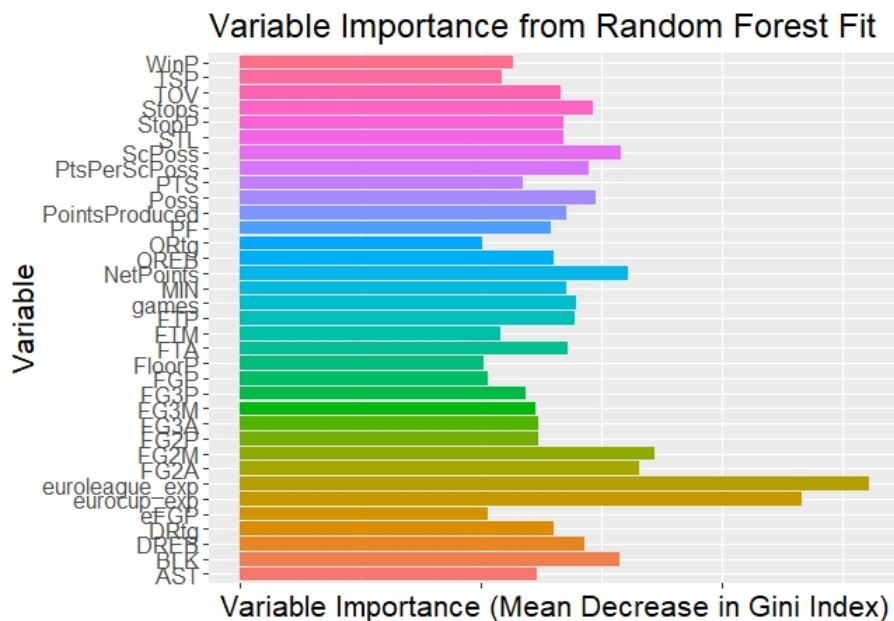


Figure 4.2: Variable Importance Plot for Random Forests model in the case group

That model was selected to predict the transcription success in control group, where it is known that none player made a transcription. The model predicted 754 observations to be transcribed out of a total of 3212 observations, which led to a classification error of 23%.

For the GLMM, eight variables were used which were Scoring Pos-

sessions, eurocup experience, euroleague experience, Floor Percentage, Net Points, Blocks, Offensive rebounds and Free Throw Attempts. The eight most important variables for the random forest model, according to the mean decrease in Gini Index, were euroleague experience, eurocup experience, Scoring Possessions, Net Points, Offensive Rebounds, Free Throw Attempts, two-point field goal attempts and two-point field goal made. So, in comparison to the variables selected for these two different methods, the difference was that in the Random Forest model two-point field goal attempts and two-point field goals made were showed up as more important than Floor Percentage and Blocks.



Chapter 5

Prediction and Evaluation of transfers

5.1 Prediction of transfers

The purpose of this chapter is the comparison of the predictive accuracy of different methods. Every model was fitted in the data from year 2010 until 2017 in order to get trained and after that their predictive accuracy were evaluated in year's 2018 transfers. Table 16 presents the predictions for 17 players. These players were all transferred or misclassified as transfers for 2018. Here we compare the GLMM, the RF (Random Forests) model and the SVM (Support Vector Machine) model. The predictions of these models are the sm_preds , rf_preds and svm_preds , respectively.

For the SVM model a radial kernel was chosen and the value of the parameters were $C = 2$ and $\gamma = 0.01$. These values were selected after a 10-fold-cross-validation.

From table 5.1, we can conclude that is extracted is that the statistical model outperforms the other two methods (random forest and the support vector machine), as it classified correct in the most of the observations for



2018.

Table 5.1: Predictions of 2018 transfers

Player	Transcription	GLMM	R.F	SVM
alen-omic	yes	no (0.17)	no	no
amedeo-della-valle	yes	yes (0.70)	yes	yes
curtis-jerrells	yes	no (0.22)	no	no
dmitry-kulagin	no	yes (0.56)	no	yes
frank-elegar	no	yes (0.58)	no	yes
jaka-blazic	yes	no (0.25)	no	no
james-bell	no	yes (0.58)	no	no
klemen-prepelic	yes	yes (0.50)	no	no
kyle-kuric	yes	yes (0.51)	no	no
marius-grigonis	yes	yes (0.69)	no	yes
nicolas-laprovittola	no	yes (0.53)	no	yes
nigel-williams-goss	yes	yes (0.84)	yes	yes
scottie-wilbekin	yes	yes (0.72)	yes	yes
shavon-shields	yes	yes (0.68)	no	yes
stephane-lasme	yes	yes (0.69)	yes	yes
tony-crocker	yes	yes (0.74)	no	yes
zanis-peiners	yes	yes (0.68)	no	yes



In Table 5.2, the stats that were used from statistical model are presented. The yellow lines indicate players that made a transcription in year 2018 but the model didn't predict it. The reason for this was that these observations had high eurocup and euroleague experience which indicated that probably are old. The age of Alen Omic, Curtis Jerrells and Jaka Blazic was 26, 31 and 28, respectively. The green lines indicate players that didn't make a transcription that year but the statistical model predicted the opposite. The reason was that all of them had low values in eurocup and euroleague experience.

Table 5.2: Feature statistics of misclassified players

player	FTA	OREB	BLK	FloorP	ScPoss	NetPoints	eurocup_exp	euroleague_exp
alen-omic	2.13	1.20	-0.21	0.06	0.75	0.76	2.50	2.04
amedeo-della-valle	5.29	-0.49	-0.05	0.03	4.16	2.20	0.50	-0.96
curtis-jerrells	-0.87	0.20	-0.21	0.01	2.84	3.18	1.50	4.04
dmitry-kulagin	2.13	-0.08	0.35	0.02	2.06	0.84	-0.50	1.04
frank-elegar	2.71	0.82	0.41	0.12	1.20	0.99	-0.50	0.04
jaka-blazic	2.23	0.10	-0.21	0.03	1.66	2.59	-0.50	5.04
james-bell	-0.34	-0.21	0.09	-0.04	0.05	-1.27	-0.50	-0.96
klemen-prepelic	3.58	-0.58	-0.21	-0.04	2.55	1.94	1.50	0.04
kyle-kuric	1.13	-0.12	-0.05	0.01	2.12	1.20	1.50	-0.96
marius-grigonis	2.38	-0.05	-0.21	0.05	1.43	1.35	-1.50	-0.96
nicolas-laprovittola	1.13	-0.52	-0.21	0.08	0.90	1.92	-0.50	0.04
nigel-williams-goss	1.13	-0.13	-0.21	0.04	4.76	4.59	-1.50	-0.96
scottie-wilbekin	2.69	-0.41	-0.21	0.04	4.51	3.56	-1.50	1.04
shavon-shields	0.63	0.06	-0.14	0.04	1.97	0.01	-1.50	-0.96
stephane-lasme	2.71	1.25	2.27	0.09	3.09	1.33	-0.50	3.04
tony-crocker	0.13	-0.30	-0.08	-0.02	0.89	1.33	-1.50	-0.96
zanis-peiners	2.53	0.20	0.09	0.10	2.37	1.92	-0.50	-0.96



5.2 Evaluation of Transfers

Prediction of transfers was the main purpose of this thesis. Even though some features can help to predict a transfer, this does not guarantee that a transfer will be a successful one. It's not easy to evaluate the success of a transfer since the aim of each transfer is unknown. Assuming that it is possible to determine the success of a transfer from Eurocup to Euroleague, this cannot be achieved by only using the player's appearances in the Euroleague matches. Transfer decisions, both by teams and coaches, are made with the aim of improving the overall performance of the team both in national season and in the Euroleague competition.

Despite these problems, let's assume that a transfer is decided only for improving the performance of the team in Euroleague games. As it is unclear how many wins and losses a team has in order to compare its performance before and after a player arrives at the team, a way of measuring the success of this transfer is by a player's evaluation index. Many player evaluation metrics were presented in the first chapter of this thesis, but it was preferred to proceed with a new one, the Total Performance Index (TPI), which is introduced by Marmarinos et al. (2019).

TPI is an advanced player evaluation metric which is an improvement over the simpler PIR (Performance Index Rating), which is used by both Euroleague and Eurocup competitions for player evaluation. The formula for constructing the TPI for a player is explained with many details in the above paper, but some slight modifications were made in order to construct it.

In order to calculate a player's TPI, the statistics Team's Defensive Points per Possessions and Team's Defensive Rebound percentage must be known. But as these stats are not recorded, they must be estimated. A simple way to estimate Team's Defensive Points per Possessions, was by



taking the average Team's Offensive Points per Possessions of all the other teams of the league at the same year. The Team's Defensive Rebound percentage had already been estimated earlier by the similar manner. Although these estimates may not be very accurate, they are quite representative by a small difference in their interpretation.

Thus, each transcription was evaluated by calculating player's TPI with the stats he made in his new team and then by comparing that TPI to the average TPI of the league in that year. So, a transfer was thought as successful, if the player that came in Euroleague was at least equal to the average player of the league at this year, according to their Total Performance Indexes.

The results are shown in Table 6.1, where the transfer year was added, the player's Total Performance Index for the year that achieved transcript written as TPI, the probability of transcript written as Prob, the new team to which the player was transferred is written as newteam, and the player's Total Performance Index for the year he completed with his new team is listed as new TPI. Moreover, figure 5.1 shows the differences in TPI for players, who just came to Euroleague from Eurocup and for players that were already playing in the league.



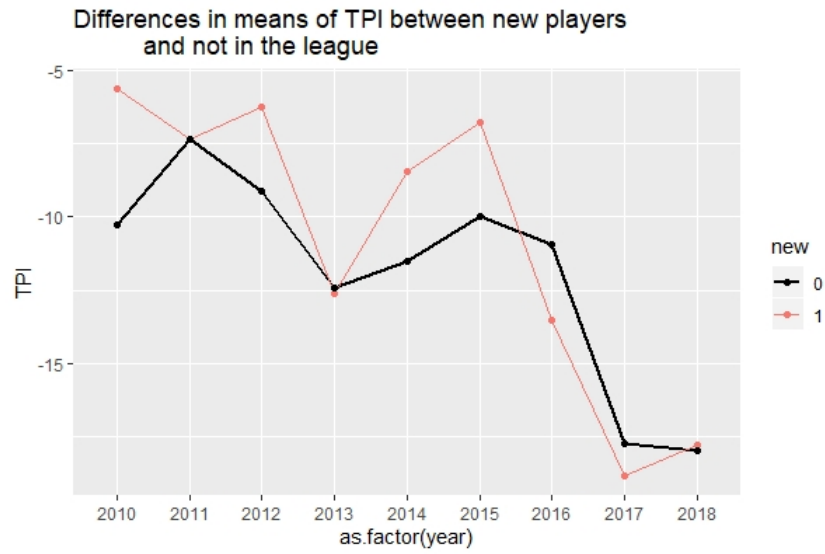


Figure 5.1: Differences in means of TPI



Chapter 6

Conclusions Discussion

Predicting an upcoming transfer is not an easy task. Due to the fact that there are plenty of limitations, lots of information recorded in basketball cannot be used into this study.

For instance, variables such as the age, height and position of the players were not included, although they should have been, as they capture lot of information for a player that other variables cannot. An example of this is the unexpected adverse effects of Eurocup and Euroleague experiences in years that were used. The explanation for this negative effect is the high positive correlation with the age of a player. But, they do not show the real impact of age on the odds of a player being transferred, which is very essential.

As for the height and position of a player, there is no calculation for their influence. There are not any other variable that can capture even a little bit their effect in transcriptions. It's unquestionable that both of these variables affect these probabilities by their marginal and join effects. They are not a few who prefer very tall players for defense but not for attack.

Another very important issue is that a necessity of advanced statistics in European basketball arises in order to quantify as well as possible



player's 'value'. From statistical point of view, it was pointed out that the advanced stats which were used can capture more information about a player's transcription probability and keep in mind that in order to succeed in making a transcription from Eurocup to Euroleague, a player must be good and valuable. Simple stats, like points made are also crucial, but they do not show the player's abilities and potentials. For example, imagine of two players both having a very high number of points made and a coach wishes to select one of them to come to his team. The player with less scoring possessions is probably more effective shooter than the other one, so it is more likely to be the coach's option.

In this study only the statistics from Eurocup and Euroleague competitions were utilized so as to investigate their impact in transcription probability. This can be misleading, because there is no availability of player statistics and performance for their local championships. According to the findings, players who made transcriptions tend to have higher scoring possessions from average in Eurocup competition, but this may differ in their local leagues. Furthermore, there were players with serious lower stats than the average that succeed in making transcription. The question that arises from this fact, and it is important to answer, is whether these players should be regarded as extreme observations or not. Maybe this player performed very well in his local league and that is the reason why he made it or maybe he was injured, or he was recovering from an injury and he could not perform as well as before, but his reputation helped him to make a transcription. Thus, not only are statistics from their local stats necessary, but an indicator that indicates the occurrence or not of an injury as well.



Table 6.1: Successful transfers according to TPI

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
aaron-white	2017	22.30	0.58	zalgeris	48.87	yes
adas-juskevicius	2012	-8.01	0.46	zalgeris	-8.34	yes
adrien-moerman	2016	43.43	0.39	darussafaka-basketbol-istanbul	-6.63	yes
aj-ogilvy	2012	-5.59	0.54	brose-baskets-bamberg	-12.73	no
aj-slaughter	2014	-33.37	0.60	panathi-naikos	-74.38	no
aleksandar-rasic	2012	-11.95	0.33	mens-sana-1871-siena	-5.88	yes
aleksandr-karpukhin	2016	-23.09	0.56	unics-kazan	-1.59	yes
aleksey-zozulin	2012	-2.53	0.32	cska-moscow	-2.71	yes
alen-omic	2016	57.83	0.52	anadolu-efes	-18.03	no
alen-omic	2017	0.31	0.23	kk-crvena-zvezda	11.69	yes
alen-omic	2018	-15.26	0.16	ea7-emporio-armani-milano	-11.00	no



Table 6.1: Successful transfers according to TPI (continued 2/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
alessandro-gentile	2011	-22.78	0.34	ea7-emporio-armani-milano	-30.02	no
alessandro-gentile	2012	-0.80	0.41	ea7-emporio-armani-milano	-22.58	no
alex-acker	2012	-32.90	0.48	asseco-prokom-gdynia	-31.15	no
alex-renfro	2016	8.00	0.61	fc-barcelona-lassa	-12.93	yes
alexei-savrasenko	2010	10.33	0.56	khimki	5.87	yes
alexey-shved	2010	-8.30	0.54	cska-moscow	-7.20	yes
ali-muhammed	2015	-41.42	0.65	fenerbahce-ulker	-58.48	no
ali-traore	2012	14.56	0.62	alba-berlin	-0.08	yes
ali-traore	2015	-29.20	0.49	csp-limoges	-10.81	yes
amedeo-della-valle	2018	-46.87	0.71	ea7-emporio-armani-milano	-4.72	no



Table 6.1: Successful transfers according to TPI (continued 3/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
andrea-cinciarini	2015	-14.73	0.67	ea7-emporio-armani-milano	-12.85	no
andreas-seiferth	2015	-27.20	0.56	bayern-munich	0.00	yes
andrew-goudelock	2014	-87.59	0.75	fenerbahce-ulker	-40.05	no
andrey-zubkov	2017	-17.77	0.28	khimki	-12.52	yes
andrija-zizic	2013	-22.42	0.48	maccabi-fox-tel-aviv	-0.36	yes
antanas-kavaliauskas	2016	-4.64	0.36	zalgiris	-2.08	yes
antonios-koniaris	2014	-0.81	0.32	panathinaikos	1.00	yes
artem-klimenko	2016	2.66	0.65	unics-kazan	-8.31	yes
artsiom-parakhouski	2014	10.38	0.78	nizhny-novgorod	74.82	yes
arturas-gudaitis	2017	9.91	0.84	ea7-emporio-armani-milano	110.92	yes
arturas-milaknis	2013	-7.18	0.37	zalgiris	3.09	yes
arturas-milaknis	2016	-22.55	0.19	zalgiris	-42.37	no
arvydas-siksnis	2010	-8.53	0.72	lietuvos-rytas	2.82	yes



Table 6.1: Successful transfers according to TPI (continued 4/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
augustine-rubit	2017	-5.91	0.71	brose-baskets-bamberg	-12.99	yes
baris-ermis	2012	1.24	0.54	fenerbahce-ulker	-19.36	no
baris-hersek	2015	-0.22	0.37	fenerbahce-ulker	-9.02	yes
benjamin-ortner	2012	-2.86	0.47	mens-sana-1871-siena	11.98	yes
boban-marjanovic	2010	9.67	0.61	cska-moscow	3.90	yes
boris-savovic	2011	-24.21	0.55	galatasaray	-4.52	yes
boris-savovic	2013	-1.63	0.47	bayern-munich	-21.45	no
brad-newley	2010	-0.74	0.57	lietuvos-rytas	-11.08	no
brian-randle	2012	5.22	0.74	alba-berlin	-3.08	yes
brian-randle	2014	3.23	0.52	maccabi-fox-tel-aviv	63.31	yes
bryce-taylor	2013	-8.81	0.37	bayern-munich	19.43	yes
bryce-taylor	2017	-5.59	0.05	brose-baskets-bamberg	4.79	yes
caleb-green-1	2014	17.11	0.65	unicaja-malaga	-13.58	no



Table 6.1: Successful transfers according to TPI (continued 5/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
can-altintig	2014	-41.42	0.47	fenerbahce-ulker	-7.26	yes
cemal-nalga	2010	-7.79	0.35	lietuvos-rytas	18.08	yes
cemal-nalga	2012	-9.68	0.26	besiktas-sompo-japan	-5.19	yes
cevher-ozler	2011	-22.69	0.58	galatasaray	-13.27	no
charles-jenkins	2016	-7.71	0.25	kk-crvena-zvezda	-15.61	yes
cj-wallace	2011	-13.03	0.47	fc-barcelona-lassa	-15.30	no
clay-tucker	2010	-60.68	0.54	real-madrid	-64.47	no
colton-iverson	2014	-20.24	0.67	kirolbet-baskonia-vitoria-gasteiz	46.53	yes
colton-iverson	2016	-0.78	0.53	maccabi-fox-tel-aviv	25.21	yes
curtis-jerrells	2018	1.59	0.21	ea7-emporio-armani-milano	-53.82	no
dairis-bertans	2016	20.33	0.29	darussafaka-basketbol-istanbul	-36.19	no



Table 6.1: Successful transfers according to TPI (continued 6/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
damir-markota	2010	-8.42	0.37	union-olimpija	-34.96	no
daniel-ewing	2012	-16.07	0.53	besiktas-sompo-japan	-20.91	no
danilo-andjusic	2011	-8.24	0.41	kk-partizan	-12.18	no
dario-saric	2014	-9.01	0.68	anadolu-efes	-12.42	no
darius-adams	2015	-38.55	0.78	kirolbet-baskonia-vitoria-gasteiz	-139.48	no
darius-johnson-odom	2015	-55.48	0.66	olympiacos	-20.86	no
darius-songaila	2013	9.29	0.68	lietuvos-rytas	-21.32	no
darius-songaila	2014	10.56	0.37	zalgiris	-29.81	no
darius-washington-1	2010	7.97	0.73	virtus-roma	-29.52	no
darko-planinic	2015	-29.46	0.47	kirolbet-baskonia-vitoria-gasteiz	-31.69	no

Table 6.1: Successful transfers according to TPI (continued 7/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
david-logan	2014	-19.28	0.33	banco-di-sardegna-sassari	-36.12	no
davide-pascolo	2016	55.54	0.79	ea7-emporio-armani-milano	7.12	yes
deandre-kane	2017	-11.62	0.77	maccabi-fox-tel-aviv	-18.63	no
deividad-gailius	2014	-36.61	0.67	neptunas	-8.34	yes
dejan-borovnjak	2015	14.53	0.44	stelmet-zielona-gora	5.82	yes
dejan-musli	2017	21.68	0.54	brose-baskets-bamberg	15.22	yes
demarcus-nelson	2014	-35.36	0.40	panathi-naikos	-45.53	no
demon-dmallet	2010	-29.93	0.44	proximus-spirou-charleroi	-12.60	no
deon-thompson	2016	-0.18	0.41	kk-crvena-zvezda	-0.82	yes
derrick-brown-1	2015	42.13	0.65	anadolu-efes	-26.14	no
deshaun-thomas	2014	-0.50	0.57	fc-barcelona-lassa	-6.99	yes



Table 6.1: Successful transfers according to TPI (continued 8/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
devin-smith	2011	21.96	0.55	maccabi-fox-tel-aviv	11.80	yes
dimitrios-agravanis	2013	0.00	0.80	olympiacos	-1.06	yes
dj-seeley	2016	-2.75	0.47	maccabi-fox-tel-aviv	-56.60	no
dj-strawberry	2015	-41.91	0.36	olympiacos	-8.00	yes
dmitry-kulagin	2015	-31.08	0.63	cska-moscow	-8.81	yes
dmitry-sokolov	2016	-1.79	0.23	khimki	2.10	yes
dominic-waters	2016	-35.44	0.57	olympiacos	-25.51	no
donatas-motiejunas	2011	-38.45	0.50	assecoprokom-gdynia	-12.16	no
donatas-zavackas	2014	0.04	0.36	neptunas	21.45	yes
donnie-mcgrath	2012	-9.68	0.36	zalgiris	-11.08	yes
donnie-mcgrath	2015	-24.12	0.28	anadolu-efes	-8.82	yes
dontaye-draper	2012	-15.08	0.57	real-madrid	-13.64	no
dor-fischer	2013	29.29	0.41	brose-baskets-bamberg	9.41	yes



Table 6.1: Successful transfers according to TPI (continued 9/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
dor-fischer	2014	23.31	0.45	unics-kazan	67.52	yes
doron-perkins	2014	-15.43	0.27	kirolbet-baskonia-vitoria-gasteiz	9.93	yes
dragan-milosavljevic	2017	-6.77	0.25	unicaja-malaga	-49.23	no
edgar-sosa	2014	-43.05	0.65	banco-di-sardegna-sassari	-70.58	no
edgaras-ulanovas	2014	12.08	0.57	zalgoris	-33.09	no
edwin-jackson	2015	-20.92	0.55	fc-barcelona-lassa	-13.39	no
ermal-kuo	2011	-32.43	0.41	anadolu-efes	-12.32	no
ersin-dagli	2010	30.23	0.62	anadolu-efes	-6.66	yes
esteban-batista	2014	59.21	0.69	panathi-naikos	6.73	yes
fabien-causeur	2012	-8.40	0.72	kirolbet-baskonia-vitoria-gasteiz	-36.76	no
facundo-campazzo	2017	-70.20	0.77	real-madrid	-53.90	no

Table 6.1: Successful transfers according to TPI (continued 10/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
filip-videnov	2010	-22.54	0.42	assecoprokom-gdynia	-10.42	no
frank-pilepic	2014	-31.85	0.42	cedevita	2.64	yes
frank-elegar	2014	8.44	0.72	eamilano	2.32	yes
gediminas-orelik	2013	-20.90	0.47	lietuvs-rytas	-1.74	yes
german-gabriel	2014	-35.11	0.57	unicajamalaga	0.77	yes
goran-jagodnik	2010	0.80	0.55	union-olimpija	-8.07	no
goran-suton	2012	-0.25	0.41	cedevita	-6.84	yes
greg-brunner	2011	25.10	0.47	red-october-cantu	-5.88	yes
heiko-schaffartzik	2015	-30.54	0.22	esp-limoges	-24.64	no
hilton-armstrong	2012	15.08	0.69	panathinaikos	-11.13	yes
hrvoje-peric	2011	-11.47	0.32	unicajamalaga	-2.73	yes
ian-vougioukas	2010	-21.24	0.63	panathinaikos	-12.88	no
ian-vougioukas	2014	31.69	0.32	galatasaray	-6.81	yes



Table 6.1: Successful transfers according to TPI (continued 11/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
ian-vougioukas	2017	19.97	0.16	panathinaikos	3.70	yes
ilya-popov	2014	-1.03	0.66	nizhny-novgorod	5.64	yes
ivan-garcia	2010	-7.58	0.47	fc-barcelona-lassa	1.57	yes
ivan-lazarev	2015	-15.86	0.49	cska-moscow	-9.56	yes
ivan-radenovic	2013	-17.84	0.48	kk-crvena-zvezda	-8.43	yes
izzet-turkyilmaz	2013	0.93	0.32	fenbahce-ulker	-5.17	yes
jacob-pullen	2013	-38.26	0.74	fc-barcelona-lassa	-35.12	no
jacob-pullen	2015	-19.37	0.57	cedevita	-131.59	no
jaka-blazic	2018	-0.08	0.26	fc-barcelona-lassa	0.73	no
jamar-smith	2014	-11.34	0.53	csp-limoges	5.46	yes
jamar-smith	2015	-25.31	0.45	unicaja-malaga	-54.28	no
jamel-mclean	2014	-4.36	0.77	alba-berlin	20.25	yes
james-augustine	2010	29.91	0.44	valencia-basket	-0.08	yes



Table 6.1: Successful transfers according to TPI (continued 12/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
james-feldeine	2015	-45.04	0.68	panathinaikos	-74.46	no
james-white-1	2015	20.70	0.48	cedevita	24.90	yes
jamont-gordon	2013	-36.41	0.44	anadolu-efes	-51.30	no
jan-jagla	2014	-3.81	0.27	bayern-munich	1.50	yes
janis-strelnieks	2013	-50.99	0.27	bc-budivel'nik	-4.04	yes
janis-timma	2017	15.03	0.66	kirolbet-baskonia-vitoria-gasteiz	-22.41	no
jannik-freese	2014	5.87	0.33	alba-berlin	-5.94	yes
jaycee-carroll	2011	-22.83	0.63	real-madrid	-5.93	yes
jekel-foster	2012	-9.03	0.44	alba-berlin	-28.51	no
jekel-foster	2013	-31.48	0.31	jsf-nanterre	-13.25	no
jerel-mcneal	2016	-64.23	0.81	brose-baskets-bamberg	-20.50	no
jeremiah-massey	2012	3.52	0.78	brose-baskets-bamberg	-12.64	no
jeremy-richardson	2010	22.92	0.52	valencia-basket	-10.36	no



Table 6.1: Successful transfers according to TPI (continued 13/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
jiri-welsch	2011	-17.59	0.32	proximus-spirou-charleroi	-21.35	no
joe-alexander	2016	2.03	0.66	maccabi-fox-tel-aviv	-9.30	yes
joe-ragland	2014	-11.90	0.70	ea7-emporio-armani-milano	-50.50	no
joey-dorsey	2016	23.61	0.33	fc-barcelona-lassa	-2.92	yes
joffrey-lauvergne	2013	9.43	0.38	kk-partizan	27.15	yes
john-bryant	2013	31.42	0.76	bayern-munich	6.08	yes
jon-diebler	2015	17.97	0.60	anadolu-efes	14.61	yes
jon-stefansson	2014	1.09	0.48	unicaja-malaga	-54.68	no
jonathan-tabu	2014	9.56	0.28	alba-berlin	-13.84	no
josh-carter	2013	10.37	0.61	mens-sana-1871-siena	-11.13	no
josh-fisher	2010	-24.14	0.36	real-madrid	17.25	yes
jp-batista	2014	-35.65	0.31	csp-limoges	-21.54	no



Table 6.1: Successful transfers according to TPI (continued 14/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
jr-bremer	2011	-0.20	0.54	ea7-emporio-armani-milano	-1.51	yes
jr-reynolds	2015	-44.15	0.58	stelmet-zielona-gora	-23.33	no
julius-jenkins	2011	-23.87	0.41	brose-baskets-bamberg	-35.81	no
jure-lalic	2010	-9.71	0.49	kk-cibona	4.57	yes
justin-doellman	2014	70.82	0.58	fc-barcelona-lassa	-15.30	no
kc-rivers	2012	4.72	0.26	khimki	-19.12	no
keith-haynes	2013	-34.49	0.52	ea7-emporio-armani-milano	-22.48	no
keith-haynes	2014	-27.15	0.45	maccabi-fox-tel-aviv	-58.52	no
kenan-bajramovic	2010	-5.88	0.43	lietuvos-rytas	-24.06	no
kenny-gabriel	2017	3.57	0.39	panathi-naikos	15.30	yes
kevin-jones	2017	5.86	0.74	kirolbet-baskonia-vitoria-gasteiz	10.50	yes



Table 6.1: Successful transfers according to TPI (continued 15/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
kevin-pangos	2016	-46.05	0.69	zalgoris	-73.69	no
kevinn-pinkney	2010	3.40	0.43	union-olimpija	6.70	yes
klemen-prepelic	2018	-50.75	0.51	real-madrid	-32.66	no
kosta-perovic	2010	38.18	0.51	fc-barcelona-lassa	23.15	yes
kostas-sloukas	2011	4.04	0.35	olympiacos	17.80	yes
kostas-vasiliadis	2013	-8.21	0.59	anadolu-efes	-18.63	no
kresimir-loncar	2010	-2.62	0.55	khimki	-5.98	yes
kresimir-loncar	2014	11.41	0.25	valencia-basket	-15.56	no
krunoslav-simon	2015	26.83	0.47	ea7-emporio-armani-milano	-46.31	no
kyle-kuric	2018	-19.72	0.50	fc-barcelona-lassa	-26.88	no
lamont-hamilton	2013	13.57	0.73	kirolbet-baskonia-vitoria-gasteiz	-30.94	no



Table 6.1: Successful transfers according to TPI (continued 16/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
lamont-hamilton	2015	-9.40	0.67	kirolbet-baskonia-vitoria-gasteiz	-14.28	no
lazaros-papadopoulos	2011	12.07	0.87	olympiacos	-14.01	no
leo-westermann	2012	-26.43	0.35	kk-partizan	-38.43	no
leo-westermann	2016	-16.95	0.18	zalgiris	-59.07	no
leon-radosevic	2016	-5.88	0.18	brose-baskets-bamberg	-3.02	yes
lorenzo-dercole	2015	-5.19	0.34	banco-di-sardegna-sassari	-3.85	yes
loukas-mavrokefalidis	2013	2.00	0.44	panathinaikos	-3.75	yes
lucca-staiger	2015	-2.67	0.13	brose-baskets-bamberg	-16.68	no
luke-harangody	2014	52.42	0.51	valencia-basket	30.18	yes
luke-harangody	2015	29.87	0.43	darussafaka-basketbol-istanbul	50.69	yes
maik-zirbes	2014	-17.32	0.56	kk-crvena-zvezda	24.49	yes

Table 6.1: Successful transfers according to TPI (continued 17/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
malcolm-delaney	2013	-37.02	0.69	bayern-munich	-40.63	no
mantas-kalnietis	2016	-0.92	0.17	zalgiris	-53.51	no
manuchar-markoishvili	2013	-9.98	0.37	red-october-cantu	-37.31	no
marco-cusin	2014	12.13	0.62	banco-di-sardegna-sassari	9.64	yes
mardy-collins	2015	-21.99	0.74	strasbourg-ig	-19.18	no
marius-grigonis	2018	5.85	0.71	zalgiris	-13.96	no
marko-arapovic	2014	-3.15	0.43	cedevita	-8.58	yes
marko-keselj	2010	-4.67	0.44	olympiacos	-2.24	yes
marko-popovic-2	2011	-8.60	0.54	zalgiris	-22.68	no
marko-scekic	2011	2.76	0.55	red-october-cantu	5.64	yes
mateusz-ponitka	2015	-11.18	0.48	stelmet-zielona-gora	3.14	yes
matt-janning	2014	7.15	0.37	anadolu-efes	-57.85	no

Table 6.1: Successful transfers according to TPI (continued 18/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
matt-janning	2016	-17.76	0.28	lokomotiv-kuban	-29.77	no
matt-janning	2017	-47.49	0.26	kirolbet-baskonia-vitoria-gasteiz	-36.03	no
matt-lojeski	2013	38.07	0.73	olympiacos	35.93	yes
matt-nielsen	2010	1.69	0.53	olympiacos	-17.73	no
matt-walsh	2010	-6.71	0.48	union-olimpija	-5.63	yes
michael-bramos	2012	-15.32	0.40	panathinaikos	20.34	yes
michail-tsarelis	2014	-19.99	0.54	olympiacos	2.62	yes
mickael-gelabale	2012	-7.12	0.24	cedevita	22.04	yes
milan-macvan	2010	0.24	0.48	maccabi-fox-tel-aviv	-8.25	no
milan-macvan	2011	11.23	0.48	kk-partizan	45.58	yes
milan-macvan	2015	37.47	0.49	ea7-emporio-armani-milano	6.75	yes
milenko-tepic	2013	-18.50	0.35	lietuvos-rytas	-22.85	no

Table 6.1: Successful transfers according to TPI (continued 19/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
milovan- rakovic	2010	-25.12	0.67	mens- sana- 1871- siena	-18.95	no
mindaugas- kuzminskas	2010	5.43	0.65	zalgiris	-7.03	yes
mirza- begic	2015	0.93	0.75	kirolbet- baskonia- vitoria- gasteiz	21.55	yes
nate- wolters	2016	-22.45	0.85	kk- crvena- zvezda	-43.45	no
nemanja- bjelica	2010	-12.13	0.50	kirolbet- baskonia- vitoria- gasteiz	5.07	yes
nemanja- gordic	2014	-40.23	0.40	cedevita	-35.64	no
nemanja- nedovic	2015	-18.17	0.39	unicaja- malaga	-77.18	no
nicolas- laprovittola	2016	-37.61	0.80	kirolbet- baskonia- vitoria- gasteiz	-42.79	no
nigel- williams- goss	2018	-22.08	0.86	olympiacos	-89.47	no



Table 6.1: Successful transfers according to TPI (continued 20/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
nik-caner-medley	2012	27.90	0.60	maccabi-fox-tel-aviv	5.07	yes
nikita-kurbanov	2015	7.78	0.20	cska-moscow	69.58	yes
nikola-kalinic	2014	7.68	0.60	kk-crvena-zvezda	-41.92	no
nikola-milutinov	2015	-18.82	0.58	olympiacos	6.04	yes
nikos-pappas	2013	-4.06	0.84	panathi-naikos	-13.43	no
nikos-zisis	2014	-34.61	0.17	fenerbahce-ulker	-28.74	no
nolan-smith-1	2014	-70.37	0.74	galatasaray	-7.58	yes
oliver-lafayette	2014	-50.12	0.29	olympiacos	-33.03	no
oliver-stevic	2011	5.84	0.53	retabet-bilbao-basket	-1.03	yes
othello-hunter	2014	-2.69	0.57	olympiacos	46.74	yes
pape-philippe-amagou	2011	-22.62	0.59	nancy-basket	-26.75	no
patrick-christopher	2012	0.49	0.67	besiktas-sompo-japan	-86.03	no



Table 6.1: Successful transfers according to TPI (continued 21/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
pau-ribas	2015	-17.07	0.19	fc-barcelona-lassa	6.74	yes
paul-stoll	2016	-7.34	0.79	unics-kazan	-12.72	yes
pavel-antipov	2016	-8.03	0.19	unics-kazan	29.91	yes
pavel-korobkov	2014	-8.93	0.53	cska-moscow	19.10	yes
petr-gubanov	2014	-2.18	0.32	unics-kazan	-0.77	yes
pierre-jackson	2017	7.07	0.97	maccabi-fox-tel-aviv	-134.17	no
pierre-oriola	2017	28.04	0.52	fc-barcelona-lassa	18.81	yes
pietro-aradori	2010	-6.15	0.63	mens-sana-1871-siena	2.37	yes
pietro-aradori	2014	2.56	0.33	galatasaray	17.11	yes
pj-tucker	2011	6.21	0.65	brose-baskets-bamberg	6.50	yes
predrag-samardziski	2013	-6.94	0.49	lietuvos-rytas	-28.46	no
przemyslaw-zamojski	2013	-22.62	0.46	stelmet-zielona-gora	-44.29	no



Table 6.1: Successful transfers according to TPI (continued 22/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
rakim-sanders	2014	5.67	0.60	banco-di-sardegna-sassari	-32.02	no
ramel-curry	2013	-20.73	0.54	panathi-naikos	-22.22	no
randal-falker	2012	6.08	0.57	besiktas-sompo-japan	6.87	yes
randy-culpepper	2016	-3.76	0.70	csp-limoges	-11.65	no
rasko-katic	2010	6.61	0.39	kk-partizan	-14.99	no
richard-hendrix	2013	23.50	0.56	ea7-emporio-armani-milano	-0.87	yes
richard-hendrix	2015	67.55	0.31	unicaja-malaga	24.69	yes
ricky-minard	2012	-23.76	0.71	besiktas-sompo-japan	4.44	yes
robert-witka	2010	-3.38	0.50	assecoprokom-gdynia	-1.04	yes
roderick-blakney	2010	35.43	0.42	unicaja-malaga	-15.13	no
rodrigue-beaubois	2016	-47.10	0.62	kirolbet-baskonia-vitoria-gasteiz	-88.36	no



Table 6.1: Successful transfers according to TPI (continued 23/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
rok-stipcevic	2010	-12.11	0.52	kk-cibona	-34.60	no
rok-stipcevic	2015	-9.98	0.47	banco-di-sardegna-sassari	-29.17	no
romeo-travis	2015	56.01	0.71	strasbourg-ig	11.89	yes
ryan-broekhoff	2015	55.19	0.51	lokomotiv-kuban	32.41	yes
ryan-thompson	2016	-18.89	0.39	kk-crvena-zvezda	-31.12	no
sarra-camara	2014	19.78	0.54	esp-limoges	-13.74	no
sasha-pavlovic	2015	-6.35	0.54	panathi-naikos	-31.92	no
sasu-salin	2017	8.56	0.18	unicaja-malaga	-71.98	no
scottie-wilbekin	2018	-27.21	0.74	maccabi-fox-tel-aviv	-133.68	no
semen-antonov	2016	19.71	0.64	cska-moscow	16.01	yes
sergei-moniam	2010	-5.51	0.72	khimki	-3.16	yes
sergey-bykov	2010	-54.50	0.80	cska-moscow	7.32	yes
sergey-bykov	2015	-10.60	0.15	lokomotiv-kuban	-42.95	no



Table 6.1: Successful transfers according to TPI (continued 24/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
shane-lawal	2015	29.33	0.82	fc-barcelona-lassa	19.65	yes
shavon-shields	2018	-13.94	0.68	kirolbet-baskonia-vitoria-gasteiz	-34.34	no
siim-sander-vene	2013	-17.35	0.50	zalgiris	2.71	yes
simas-buterlevicius	2012	-16.91	0.46	lietuvos-rytas	-5.50	yes
simas-jasaitis	2010	15.73	0.55	lietuvos-rytas	-13.92	no
sofoklis-schortsanitis	2016	-27.45	0.35	kk-crvena-zvezda	-7.61	yes
stanko-barac	2015	5.38	0.29	ea7-emporio-armani-milano	10.34	yes
steed-tchicamboud	2014	-14.77	0.42	csp-limoges	-11.68	no
stefan-jovic	2014	-4.15	0.47	kk-crvena-zvezda	-5.05	yes
stefan-markovic	2014	-15.97	0.15	unicaja-malaga	-30.22	no
stefan-markovic	2017	-10.47	0.26	khimki	-61.47	no



Table 6.1: Successful transfers according to TPI (continued 25/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
stephane-lasme	2018	35.48	0.70	panathinaikos	1.77	no
steven-smith	2011	-2.26	0.40	panathinaikos	3.92	yes
tadija-dragicevic	2013	-1.58	0.34	anadolu-efes	-9.45	yes
taylor-rochestie	2012	-15.63	0.51	kirolbet-baskonia-vitoria-gasteiz	-7.16	yes
taylor-rochestie	2017	-37.55	0.13	kk-crvena-zvezda	-52.78	no
thomas-kelati	2010	1.39	0.40	khimki	-11.67	no
thomas-kelati	2013	-2.92	0.22	kirolbet-baskonia-vitoria-gasteiz	18.81	yes
tony-crocker	2018	-9.17	0.74	khimki	-91.16	no
trent-meacham	2014	-8.08	0.41	ea7-emporio-armani-milano	-1.07	yes
trent-plaisted	2010	15.08	0.70	zalgiris	1.60	yes
trent-plaisted	2014	33.49	0.36	csp-limoges	1.17	yes



Table 6.1: Successful transfers according to TPI (continued 26/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
trevor- mbakwe	2015	40.09	0.70	maccabi- fox-tel- aviv	6.00	yes
uros- nikolic	2014	-7.44	0.59	turow- zgorzelec	-5.83	yes
uros- tripkovic	2010	-28.62	0.41	unicaja- malaga	-21.29	no
vadim- panin	2010	4.86	0.39	khimki	-0.64	yes
vadim- panin	2014	24.92	0.42	unics- kazan	6.95	yes
vassilis- kavvadas	2013	4.97	0.53	olympiacos	-0.79	yes
vassilis- kavvadas	2014	-5.62	0.46	olympiacos	-15.21	no
victor- rudd	2016	-28.85	0.79	maccabi- fox-tel- aviv	-36.58	no
viktor- sanikidze	2014	19.91	0.52	unics- kazan	6.20	yes
vlad- moldoveanu	2014	-15.37	0.58	turow- zgorzelec	-6.21	yes
vlad- moldoveanu	2015	8.82	0.42	stelmet- zielona- gora	28.10	yes
vladimir- boisa	2010	4.47	0.68	union- olimpija	0.48	yes
vladimir- dragicevic	2013	50.16	0.37	stelmet- zielona- gora	28.33	yes



Table 6.1: Successful transfers according to TPI (continued 27/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
vladimir-golubovic	2014	82.55	0.70	unicaja-malaga	15.92	yes
vladimir-jankovic	2013	-10.35	0.63	panathi-naikos	0.02	yes
vladimir-stimac	2013	-0.19	0.54	unicaja-malaga	35.19	yes
vladimir-stimac	2015	26.46	0.22	kk-crvena-zvezda	25.88	yes
vladimir-veremeenko	2016	18.65	0.14	brose-baskets-bamberg	-17.98	no
vlado-ilijevski	2012	-3.67	0.23	anadolu-efes	-10.26	no
vule-avdalovic	2012	-57.90	0.48	alba-berlin	-10.23	yes
walter-hodge	2013	-22.85	0.80	kirolbet-baskonia-vitoria-gasteiz	-49.80	no
will-daniels	2013	-33.85	0.56	jsf-nanterre	-25.04	no
will-daniels	2014	-2.06	0.39	nizhny-novgorod	-6.11	yes
willy-hernangomez	2015	-5.33	0.63	real-madrid	18.55	yes
yassin-idbihi	2015	-2.86	0.17	brose-baskets-bamberg	-8.91	yes

Table 6.1: Successful transfers according to TPI (continued 28/28)

Player	Year	TPI	Prob	NewTeam	NewTPI	Success
yogev-ohayon	2011	-4.26	0.37	maccabi-fox-tel-aviv	0.25	yes
zack-wright	2013	19.66	0.66	panathinaikos	-20.87	no
zanis-peiners	2018	23.13	0.68	darussafaka-basketbol-istanbul	-12.77	no
zoran-dragic	2012	2.22	0.57	unicaja-malaga	-13.45	no



Chapter 7

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle., *Proceedings of 2nd International Symposium on Information Theory, Academiai Kiado, Budapest, pp. 267-281*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716-723.

Casals, & Martinez, (2013). Modelling player performance in Basketball through mixed models. *International Journal of Performance Analysis in Sports*. 13. 64-82.

Marmarinos C., Bolatoglou T., Karteroliotis K. & Apostolidis N. (2019). Structural validity and reliability of new index for evaluation of high-level basketball players. *International Journal of Performance Analysis in Sport*, 19:4, 624-631, DOI: 10.1080/24748668.2019.1644803.

Davenport, T. H. and Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*.

Dobson, J. Barnett, G. (2008). An Introduction to Generalized Linear Models, Third Edition. ISBN-13: 978-1584889502

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based



methods. *Bayesian Statistics 4*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 147–167. Oxford University Press.

Giasemidis, G. (2020). Descriptive and Predictive Analysis of Euroleague Basketball Games and the Wisdom of Basketball Crowds.

Giuliodori, P. (2017). An artificial Neural Network-based Prediction Model for Underdog teams in NBA MAtches.

Gutierrez, D. (2019) Data Scientists vs Statisticians. *ODSC journal*

Hanley, J.A. Mcneil, B. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 143. 29-36. [10.1148/radiology.143.1.7063747](https://doi.org/10.1148/radiology.143.1.7063747).

Hoerl, A. E. & Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, Vol. 12, No. 1 (Feb., 1970), pp. 55-67

Hollinger, J. (2005). Pro Basketball Forecasts. *U.S: Potomac Books*.

Hu, Feifang & Zidek, James, V. (2004). Forecasting NBA Basketball Playo Outcomes Using the Weighted Likelihood. [10.1214/lnms/1196285406](https://doi.org/10.1214/lnms/1196285406).

Hwang, D. (2012). Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model. *MIT Sloan Sports Analytics Conference 2012*

James, Bill (1977). 1977 Baseball Abstract. *Lawrence, Kansas*

James, Bill & Henzler, Jim (2002). Win Shares. *Stats Inc; 1st edition (March 1, 2002)*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning, with applications in R. *Springer Science+Business Media New York 2013*

Ruddy JD, Stuart J. Cormack1 , Whiteley R., MD. Williams, RG. Timmins & Opar, DA. (2019). Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches. *Front. Physiol*. 10:829. doi: 10.3389/fphys.2019.00829



- King, B. (2017).** Predicting National Basketball Association Game Attendance Using Random Forests. *Journal of Computer Science and Information Technology*.
- Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D.T. (2007).** A starting point for Analyzing Basketball Statistics. *Journal of Quantitative Analysis in Sports, Vol. 3: Iss. 3, Article 1.*
- Kvam, P. Sokol JS. (2006).** A Logistic Regression/Markov Chain Model For NCAA Basketball. *Naval Research Logistics Journal.*
- Lewis, M. (2003).** Moneyball: The art of winning an unfair game. *New York: W.W. Norton.*
- Makridakis, S. Spiliotis, E. Assimakopoulos, V. (2017).** The Accuracy of Machine Learning (ML) Forecasting Methods versus Statistical Ones: Extending the Results of the M3-Competition.
- Manley, M. (1987).** Martin Manley's Basketball Heaven 1987-88. *ISBN-13: 978-0944877005*
- Langaroudi, MK., Yamaghani, MR. (2019).** Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey. *j. adv compeng eng technol, 5(1) Winter2019 : 27-36*
- Ntzoufras, I. (2012).** Bayesian Modeling Using WinBUGS. *wiley series in computational statistics.*
- Oliver, D. (2004).** Basketball on Paper: Rules and Tools for Performance Analysis *U.S: Potomac Books.*
- Piette, J., Pham, L., & Anand, S. (2011).** Evaluating Basketball Player Performance via Statistical Network Modeling. *MIT SLOAN: Sports Analytics Conference*
- Pai PF, ChangLiao LH & Lin, KP (2016).** Analyzing basketball games by a support vector machines with decision tree model. *Neural Computing and Applications. 28. 10.1007/s00521-016-2321-9.*
- Tibshirani, R. (2011).** Regression shrinkage selection via the LASSO.



Journal of the Royal Statistical Society Series B. 73. 273-282. 10.2307/41262671.

Vehtari, A., Gelman, A. & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 2017, Volume 27, Issue 5, pp 1413-1432.

Vinue, G. & Epifanio, I. (2019). Forecasting basketball players' performance using sparse functional data. *Statistical Analysis and Data Mining : The ASA Data Science Journal.* 10.1002/sam.11436.

Wu, S. and Born, L. (2017). Modeling Offensive Player Movement in Professional Basketball.

Zhu, R., Zeng, D. & Kosorok, M. (2015). Reinforcement Learning Trees. *Journal of the American Statistical Association.* 110. 0-0. 10.1080/01621459.2015.1036994.

Zimmermann, A. Moorthy, S. Shi, Z. (2013). Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned.

