We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter 5

Using Neural Networks in Preparing and Analysis of Basketball Scouting

Branko Markoski, Zdravko Ivankovic and Miodrag Ivkovic

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48178

1. Introduction

"I don't believe in keeping statistics. Only statistics that is important is the final result."

This used to be true, mostly because opponents also did not keep or use statistics. However, the times have changed. Final result is still the most important thing, but a way in which such result is obtained is also of great importance.

Basketball is the one of most popular sports, in Serbia and in the world. It is a team sport. Actors of a basketball game are the players from two opposing teams, their team officials with coaches, assistant coaches, doctors and officials (commissioner, referees, table officials, statisticians). Every team, depending on a league or competition, may have no more than 12 or 10 players per game, and of those 5 are actively engaged in the game [1]. Basketball game would not start without 5 players from each game on the court. In Europe, regular basketball game is divided into 4 quarters of 10 minutes each, while in NBA every quarter is 2 minutes longer. If a result is draw, after regular time additional time of 5 minutes is played, as many times as necessary to decide a winner of a game.

There is no limitations regarding a number of substitutions of players during a game, but there is a limitation regarding personal fouls. If a player gets fifth personal foul, he must leave the court and may not play any more.

A basket scored gets 2 or 3 points, depending on a distance from which the ball has been thrown. A line discerning two cases is drawn on the court.

In the times past, basketball statistics used to be a luxury, available only to big professional teams [2]. First beginnings of a basketball statistics started in 1969, when it was first introduced at the NBA game. Statisticians were keeping only one- or two- point scores for every player. During next two years, statistics has been developing and had sixteen events

INTECH
open science | open minds

© 2012 Markoski et al., licensee InTech. This is an open access chapter distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

that statisticians had to keep track. Slowly but steadily, it gradually took a central position in analysis and preparation of a game, for individual players and for the team as a whole. Every team in the NBA had four statisticians, keeping track how a team is doing in offence and in defense. In those times, basketball was more oriented to offensive activities than defensive ones. There was even a name for such type of basketball: "Run and gun". For the average coach, statistics was a nightmare: it required a great deal of time and effort, first at collecting statistical data, then in manual computing various joint statistical parameters [3]. For most coaches, statistics was simply not worth the effort.

Having in mind characteristics of information quality, manual keeping of statistics, using a pencil and paper, has several flaws. The most important one is incompleteness, since due to restricted paper area evidence is kept only to most basic statistics parameters, sixteen in this case. Presently, 38 standard statistical parameters are being kept. This demands exceptional knowledge of basketball game. Moreover, it takes time to write a data on statistics sheet, and to compute data afterwards [4]. There is a considerable possibility to calculate data incorrectly, not to mention long time needed to calculate data or to write it on a paper. Since there were special forms for statistics, statisticians had to write data in a precisely defined spots, and therefore probability to omit one or more actions happening on the court was considerable. In addition, if one or more actions were omitted, there is no real information or real data. There is a thin line between win and loss [5] and every bit of information is invaluable preparations for the game, for every player and for team as a whole. Because of this, it is essential that statisticians know basketball very well, and their training is of utmost importance in order to achieve certain speed and to "catch" all actions on the basketball court. Considering the statistical sheet used for keeping statistics manually, it is evident that noted data are (un)intelligible, especially in comparison to those processed on computer and printed, with abundance of computed summary and other parameters [6]. Time and effort that must be invested in manual computing of different summary statistical parameters are enormous. Therefore, summary data may not be available at the moment when they are needed most. Moreover, possibility of human error in this kind of calculation is considerable, so even the accuracy is compromised. Coaches often could not obtain quick and precise information that was necessary in order to react in a right moment to help the team.

It should be noted that, after "manually extracting parameters", statistics had been sent to the Basketball Federation, where all sheets from all courts were again manually processed in order to try to find some regularities. Another manual statistics was kept for needs of referee commission [7]. In other words, there were other types of statistics in need. Regarding manually kept statistics, we must emphasize that this is a very difficult task; for instance, one cannot obtain total shots of a player, or a number of turnovers because time is needed to "count and add" all the data. In addition, there was a possibility to omit something and therefore to obtain incorrect data [8].

Information technologies development and their integration in all areas of social and business life have not bypassed the sports. Increasing professionalism and competition

cause clubs to approach all their activities in more and more systematic way. Considerable progress is being noted in training system and in analysis of opponent teams and players, so-called scouting. Introduction of computers was helpful, releasing assistant coaches from responsibility for keeping statistics, and at the same time providing a number of information that they could only dream of twenty years a go. Moreover, computers and software are now widespread and relatively inexpensive, so information is accessible to everyone.

At the beginning of a season, coaches are mostly interested in using different statistical reports for analyzing and evaluation of individual players. Once they have insight into advantages and flaws of their players, their interest is moving towards the team as a whole. They want to know how good the team is. The team statistics therefore becomes most important. After all, a basketball is still a team sport. Finally, having in mind that different statistic reports may be used also for analyzing opponent's play, coaches' interest is moved during a season towards their opponents. More often than not, well-analyzed opponents' play means a difference between winning and losing.

Statistics is not used solely by coaches. On the contrary, whole population of sport lovers and fans is able, due to mass usage of technology and media such as television and internet, to follow efficiency of individuals and teams. Numerous professions, for instance journalists and commentators, have use of statistics in doing their job, while for some, as sports managers, it is of vital importance for their profession. In last decades, sport became more than a game: it is a large business with considerable amounts of money invested.

The final result is not the only thing important any more...

Computers are used in NBA for a some time now. Question is how the statistics is kept. Until 2000, a very popular program has been used, where actions were entered using the mouse and the keyboard. At every game there had to be five people: two keeping statistics for one team, two for another, and one person was a supervisor, controlling every step. There had to be yet another person at every game, as a substitute. Two statisticians in charge for one team were organized in such a way that one of them follows attack ant other a defense. These two could never switch, for instance taking turns for every game. Therefore, they were offense and defense specialists. Since 2001, statistics is kept using voice. This is a modern technology that slowly comes even to Serbia. In this case, one must “train” the computer to his or her voice, and to be careful to use only those words that are present in a database, or else such action would not be noted in a database. Another aggravating issue for this kind of keeping statistics is noise. Statisticians became skilled and hold microphone very close, in order to minimize noise. This manner of keeping statistics, using voice, considerably facilitates a process because the statistician may observe much more details when he is not busy with a mouse or keyboard. There is another way of keeping statistics, using PDA devices, where actions are entered using a pen. [9]

We will describe only some of differences in keeping statistics. Let us start from assists, which is probably most controversial one, besides rebounds and turnovers. In Europe, assist is noted only if a player passes the ball to another player on a good shooting position. This

other player must score, for two or three points. Until recently, assists were noted only if a player is alone in the paint, but presently they are valid wherever the player is in a position to shoot and score two or three points [10]. In USA, this is quite different, whether in NBA, NSA or WNBA. There the assist is noted if a ball was passed to other player when he/she is alone on a shooting position, even if he/she was fouled after that. Assist is noted since the first player has done everything right. We had quite a discussion with Mr. Aleksandar Đikić regarding this, because we think this is quite logical because it is not assisting player guilt that receiving player did not score. Then again, basketball is a team sport. We have often witnessed situations when assisting player does everything right, and shooting player misses and assist is not noted.

This problem demands deeper analysis in order to find the best solution. In USA it is also noted where assist occurred, whether in the paint, outside the paint or at perimeter. We have a type of report saying which player assists which player the most. This is a good detail to estimate the teams' play.

Another category differing between Europe and USA is difference between steal and defensive rebound. This caused a lot of controversy in our country too. In USA, after every shot, a defensive rebound ensues and there is no problem whatsoever. However, what if after shot ball drops to the ground, or if after a free-throw shot ball bounces to the player's hands? Is this still a rebound or maybe a steal? We have spoken to a number of coaches and our only conclusion is that there is no rule. Opinions differ. In Serbia, when a ball bounces to hands after a free throw it is noted as a steal. In USA, steals are kept differently, every good defense is noted as a steal; for instance, when a good defense forces a player to step out. In such case this ball is stolen. There are several instances when after jump ball, the ball is won and noted as a rebound, and not a steal. Another interesting issue is a turnover. There are nineteen types of turnover in present basketball. Most turnovers during a game are due to bad pass. For statistician it is often hard to determine whether it was bad pass or bad catch [11], especially since statisticians are mostly seated at relatively unfavorable positions, below or around the basket. It would be much better to be placed close to scorers table, where they would have good view and could easily process disputable cases. In addition, coordination with scorers table would be easier if they note that something is not right. One of turnovers is when a player stays in paint for five seconds (three seconds in Europe). This rule was literally introduced for Shaquille O'Neal; as a dominant centre, he often entered "down the paint" and easily scored. Until recently there was another type of turnover: illegal defense. It was up to referees to recognize this. In our country, in younger categories zone defense is illegal; referees, if note this, are to warn the coach. Until two years a go, first time was the warning, and every next time was technical foul for the coach. In USA, first one was the warning, and every next time there was a free throw. There is also another type of scouting close to statistics – a way and type of defense. They must recognize a type of defense and therefore to determine a shot by a player: successful for two or three points in zone or on man [12]. Therefore, in order to keep statistics successfully, a person must know rules very fell. A number of sports organizations have their own information system regarding membership, spectators, caterers, competitors, employees, resources etc. Building an

effective information system is important in order to enable easier accomplishing goals and mission defined in strategic analysis of the sport organization. It is important to collect enough information, but not too much in order to avoid information overload. In such situations, decision makers who should benefit from information gathered are unable to reach right decision due to too much information. Developing a good information system is a dynamic process. First, it should be determined which information is to be collected and how to do it. When data are collected, they must be processed and analyzed. Finally, information must be stored in a way that provides easy use and adaptation of different analysis to fit new situations [13]. System storing information regarding a sports organization is called database. Use of computers and appropriate software (database) provides quick access, analyze and report regarding huge number of important aspects in work and development of the sports organization.

2. Scouting and data mining techniques

The main reason for scouting is to know the opponent in all stages of basketball. Scouting is done at team and individual level. Team level reviews opponents systems of playing the game in offense, defense and transition; how the team acts in all kinds of defense, how it attacks after outs and how it transits from defense to offence. Every stage may be statistically shown as number of tries, lines of offense and percentage. In addition, good and bad sides of team's and individual game may be shown. Individual scouting reviews performance of every individual player in all game stages, his statistical performance, his good and bad sides. For example, from which action he attacks most frequently and most successfully or in which actions he has lower performance, as well as in how he (or she) performs in different kinds of defense (what he defends worst?)

All in all, scouting shows how to attack the opponent in most efficient way and how to handle defense. Therefore, statistics and scouting are important part of every analysis required in order to prepare for future games. A good scouting often requires to follow several games of the opponent team, mostly last four (two at home and two away). This requires exceptional knowledge about basketball and computers as well. Scouts often need several days to prepare players and coach for the next opponent. With present rhythm, playing games twice a week (Wednesday – Sunday), scouts often don't have enough time to cover all opponents, so some teams have two or even more scouts in order to analyze every next opponent. Naturally, there is a question of a mean to shorten the time for the scout, but in such a way that he still obtains good quality information that will provide advantage over the opponent. There is a powerful and good-quality tool today: "data mining in sport" techniques. Data mining techniques in sports, especially in basketball, are in rise recently. Those tools and techniques are developed with the aim to measure performances of individual players and of the team as a whole. Since the sport is one of most profitable industries, these methods, as well as performances of players and teams, attracts much attention of sport clubs and managing companies. Before data mining and its advantages, analysis of opponents, as well as preparation of tactics for the game, was a task of professional scouts. Since the number of games constantly increased, and scouts could not

manage large number of games and corresponding amount of information, new methods were sought in order to extract knowledge from raw data. Every team could choose between two ways. One was to engage professional statisticians who had deep knowledge about concrete sport, in his case basketball, and who would enable the team to reach right decisions. Other way was to find methods that will shorten the time, and still provide precious knowledge; in other words, to start data mining techniques. When appropriately used, data mining techniques may result in better preparation of the game, and better performances of the team and players. This means that players may be prepared for certain events that may occur at the game, using all downsides and flaws of the opponent team.

In order to analyze the opponent, we need information. Knowing individual qualities of players and their habits, in both offence and defense, we may easily predict where advantages or problems will occur in individual situations at offence and defense. Having this in mind, we will pay more or less attention to certain segments of opponent's game, thus reducing the number of information and allowing all players of the team the clear and identical idea how to play that game. When number of misunderstandings decrease, power of team play rises. Therefore for appliance of data mining quality and precise data are necessary. One of definitions in FIBA manual [14] is: Basketball is a complex game between two opposing teams with the aim to score most points and win. During the game, a vast number of events occur, and it is very hard or even impossible to note them all. Besides basic events as shots, assists, turnovers and steals, offensive and defensive rebounds, there are a number of relevant events such as movements of player across the court with and without the ball, a type of defense played by a team, and which player is the weakest point in which defense. We must emphasize that there are measurable and non-measurable part in statistic [15]. Measurable part is the one that may be presented in statistics, but there is also a part that is not noted anywhere, for instance the last good defense, last wrong foul, last turnover or steal…

We must emphasize that during a game every team has its statistical unit that processes data and gives their coaches printed materials with information so that they react during a game and reach the right decision, and to prepare information relevant for next game or the next opponent. Due to amount and complexity of data, basketball is extremely quick and dynamic game and therefore suitable for application of data mining techniques, especially neural networks that enable extraction of conclusions from raw data. Having this in mind, we will pay more or less attention to certain segments of opponent's game, lowering number of information and allowing all players to have a clear and identical idea how to play that particular game. We do not want players to choose one or two information according to their will. We want players to choose and implement two identical information, even if those have only secondary influence on opponent's game. Therefore, it is not a bad idea to test your team during a preparation period and determine how they accept information. By using neural networks, as one of integral and indispensable technologies in data mining used most frequently in basketball data mining, patterns were discerned pointing to influence of different parameters of basketball game. In order to check correctness of results obtained, input data were analyzed using C5.0 decision tree. It is important to say that a basketball court is divided into zones, and basic division is to six fields [16].

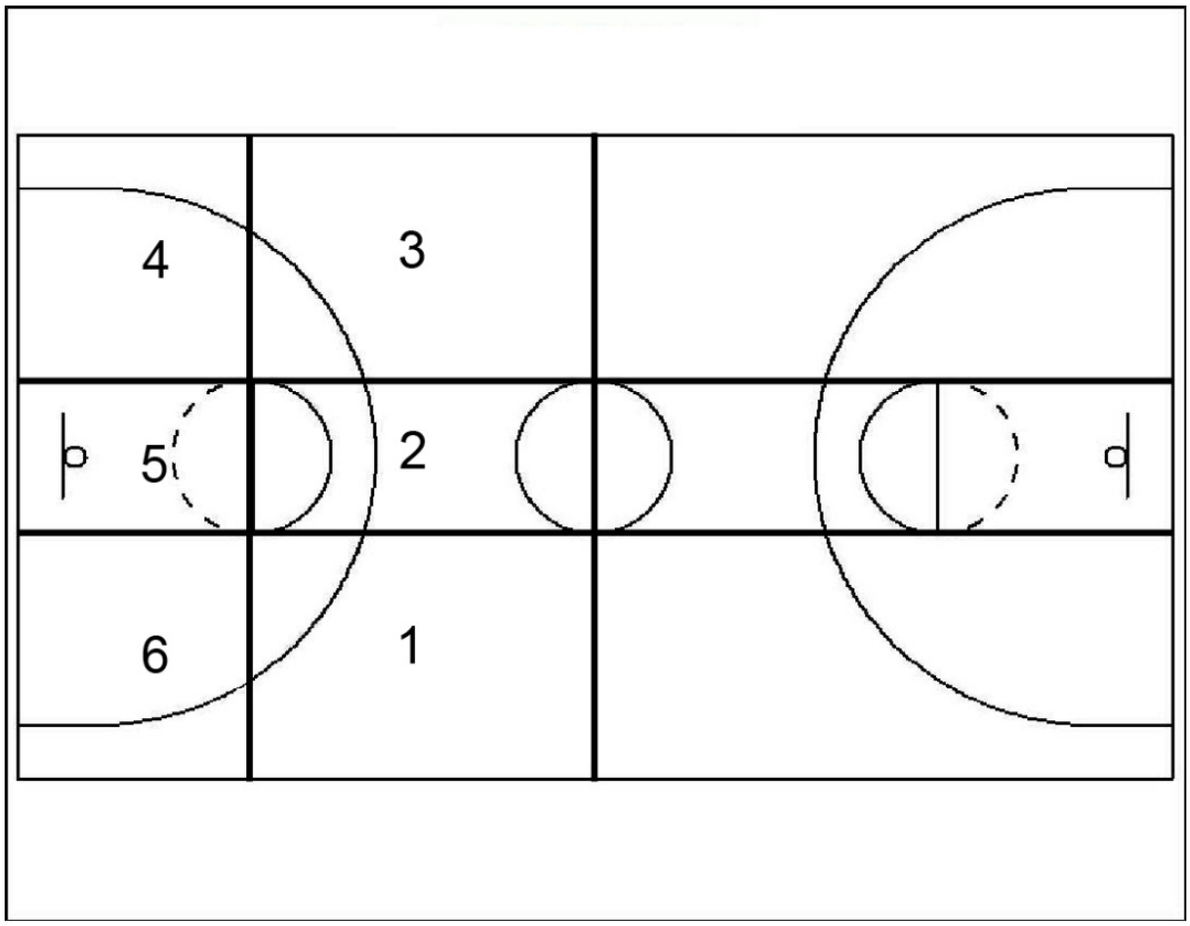


Figure 1. Division of a basketball court

This is of utmost importance since we may know how players, or a team, shoot for two or three points. By using data mining techniques, we consider influence of shot from certain positions in a field and in general. Influence of all gathered parameters is considered (shots for one, two and three points, offensive and defensive rebounds, turnovers, steals, assists, blocks). By using these technologies, models are created in order to predict the result of the game.

According to [17], "Data mining is a process of discovering new sensible correlations, patterns and guidelines by observing a large amount of information stored, and by using pattern recognition technologies and statistical and mathematical techniques". There are also other definitions:

- "Data mining is the analysis of (mostly large) observed data set in order to find certain connections and to sum up data in new ways, being intelligible and useful to the data owner" [18].
- "Data mining is an interdisciplinary bough merging techniques from machine learning, pattern recognition, statistics, databases and visualization, in order to solve the question of obtaining information from large data bases" [19]

Some companies are trying to use data mining in their own way, depending on level of inertia in certain departments. Inter-industrial standard was obviously necessary, being

independent on industry branch, independent on tools and independent on application. For this purpose, special international industrial standard has been developed, independent on industry type, tool and application. Analysts from DaimlerChrysler, SPSS and NCR have developed the Cross-Industry Standard Process for Data Mining (CRISP-DM) in 1996. The CRISP is a non-profit standard freeware, intended for fitting of data mining to general problem-solving strategies for business and research purposes. Most often tasks given to data mining are: Description, Estimation, Prediction, Classification, Clustering and Association. We will touch every one with few lines. Regarding description, we must say that researchers and analysts are trying to find ways to describe patterns and tendencies present in given data. Data mining models must be as transparent as possible. This means that results of data mining model must explain clear patterns responsible for intuitive interpretation and explanation. Regarding estimation, it is similar to classification except variables sought are numerical and not categorical. Models are built using "complete" records, providing values for target variables, and predictors (defining variables). For new observation, estimated values of target variables are given regarding values of predictor. Prediction is similar to classification and estimation, except that result of prediction is located in the future. Examples of prediction in business and science areas may be predicting a winner of this year's championship in football, basketball or any other sport, based on comparing statistics of given teams. Any method or technique used for classification and estimation may also be used for prediction, under certain conditions. We also must emphasize that traditional statistical methods must be included, such as point estimation and estimation of confidence intervals, linear regression, correlation an multiple regression, as well as data mining methods and methods of retrieving knowledge such as neural networks, decision trees and methods of k-closed neighbor. Regarding classification, we must point that there is a target variable belonging to some category. Data mining model investigates a large set of records, where every record contains information about target variable and a set of predictor variables. Clustering goes for grouping records, and for observing and sampling classes of similar objects. Cluster is a collection of records similar to each other, and different from records in other clusters. Clustering is differing from classification since there is no target variable for clustering. Task of clustering is not to classify, estimate or predict value of target variable. Cluster algorithms tend to divide a data set to relatively homogenous subgroups (clusters), with maximum similarity between records, while similarities between records in different clusters are minimal. Aim of associations in data mining is to find out which attributes "belong together". This is prevailing technology in a business world, where it is called affinity analysis or market analysis. Its task is to find rules to describe connections between several attributes. Rules of association are in "*if hypothesis then consequence*" form, together with measurements of support and confidence associated to a rule.

3. Neural network

The modern discipline of neural networks was created as a combination of several quite different ways of research: signal processing, neurobiology and physics [20] and therefore is

a typical interdisciplinary branch of science [21]. It is basically an effort to comprehend intricacies of a human brain, as well as to apply new insightsto processing complex information [22]. There is a number of progressive, non-algorithmic systems, as learning algorithms, genetic algorithms, adaptive memory, associative memory, fuzzy logic. General opinion is that neural networks are presently the most mature and most applicable technology [23].

Conventional computers' work is based on logic: deterministic, sequential or with a very low level of parallelism. Software written for such computers must be literally perfect in order to work properly. This requires a long and therefore expensive process of perpetual design and testing.

Neural networks belong to the category of parallel asynchronous distributed processing. The network is damage-resilient or only a relatively low number of neurons falls out of function. It is also tolerant to noise in input signal. Every memory element is delocalized - situated in network as a whole - and it is impossible to identify in which part it is stored. Classic addressing is nonexistent, since memory is approached using contents, and not the address [24].

Basic component of neural network is a neuron, as shown in figure 4:

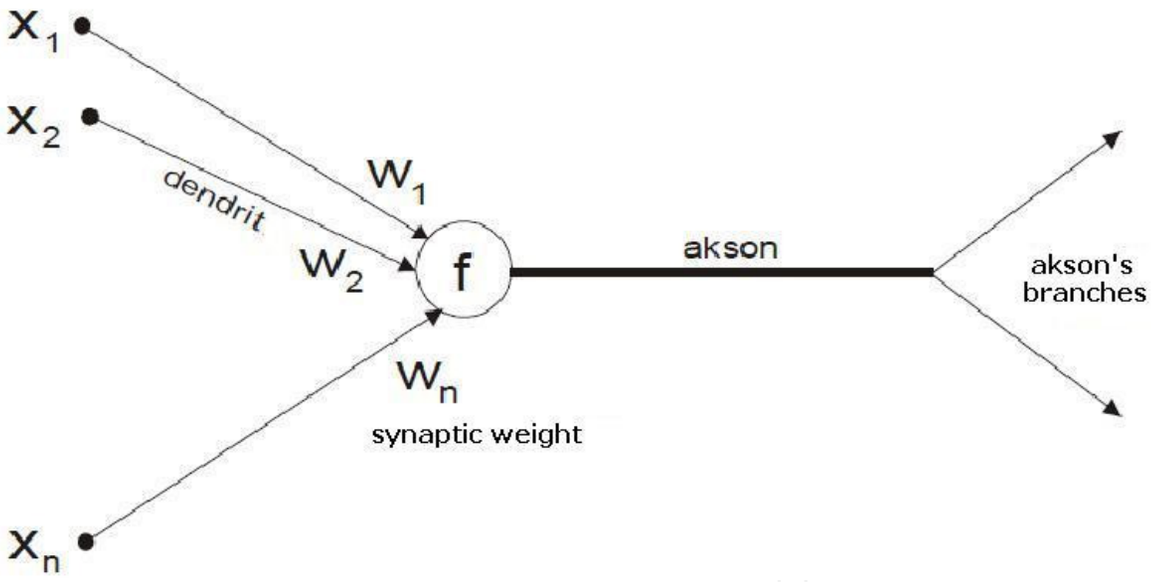


Figure 2. Basic components of neural network

3.1. Backward propagation

Neural network is a controlled learning method, demanding a large training set of complete records, including target variable. Since every observation from training set is conducted through the network, output value is obtained at the output node. This value is compared to the real value of target variable for given observation in training set, and the difference between the real and the predicted value is calculated. This prediction error is corresponding

to prediction error in regression model. In order to measure how much the output prediction is consistent to real target value, most neural networks uses sum of error squares (SSE):

$$SSE = \sum_{records} \sum_{output\ nodes} (actual - output)^2 \quad (1)$$

where prediction error square is summed over all output nodes and all records in a training set.

The main problem is therefore to create set of difficulty models, which would minimize the SSE. In this way, difficulties correspond to parameters of the regression model. "Real" values of difficulty that would minimize SSE are unknown, so it is our task to assume ones for the input data set. Due to non-linear nature of the sigmoid function extending through the network, there is no ready-made solution for minimizing SSE.

3.2. Sensitivity analysis

One of downsides of neural networks is their vagueness. The same exquisite flexibility that enables a neural network to model a wide range of nonlinear behaviors, at the same time limits our ability to interpret results using easily formulated rules. Unlike decision trees, there is no clear procedure for translating complexities of neural network to a compact set of decision rules.

There is still a procedure that may be used – sensitivity analysis, enabling us to measure relative effect of every attribute on output result. By using testing data, this analysis works as follows:

1. Generates a new observation x_{mean}, where every attribute value in x_{mean} is equal to mean value of different attributes for all records in a testing data set.
2. Finds network output for x_{mean} input. We will call it $output_{mean.}$
3. One attribute at a time, changes x_{mean} in order to represent attribute minimum and maximum. Then it finds network output for every variation and compares it to *output mean*

Sensitivity analysis will determine that change of certain attributes to minimum value has more effect on resulting output of a network than some other attributes.

3.3. C4.5 algorithm

C4.5 algorithm is a Quinlan extrapolation of its own ID3 algorithm for creating a decision tree [25]. It recursively visits every decision node, choosing optimal division while divisions are possible. The basic properties of the C4.5 algorithm are:

- C4.5 algorithm is not limited to binary divisions; it creates trees with triple, or multiple branching.
- For categorical attributes, C4.5 by default creates separate branch for every value of a categorical attribute. This may bring to excessive complexity, since some values may have very low frequency or be connected to other values.

C4.5 algorithm uses concept of information gain of entropy reduction in order to choose optimal division. Let us suppose that we have variable X with k possible values with probability $p_1, p_2, \ldots, p_k$. Which is the minimal number of bits needed, in average per symbol, to conduct a sequence of symbols representing values for X observed? Answer is called entropy for X and it is defined as:

$$H(X) = -\sum_j p_j \log_2(p_j)$$

Where this entropy formula comes from? For event with value p, average amount of information in bits needed to conduct result is $-log_2(p)$. For instance, result of coin toss-ups, with probability 0.5, can be conducted using $-log_2(0.5)$ =1 bit, i.e. zero or one, depending on toss-up result. For variables with several possible outcomes, we use difficulty sum for $log_2(p_j)$, with difficulties equal to probabilities of outcomes as seen in formula

$$H(X) = -\sum_j p_j \log_2(p_j)$$

C4.5 uses entropy concept in a following way. Let us suppose that we have potential division S, which divides training data set T to several subsets, $T_1, T_2, \ldots, T_k$.. Mean information demands may be calculated as a difficulty sum of entropy for all individual subset:

$$H_s(T) = -\sum_{i=1} P_j H_s(T_j)$$

where Pi is a data proportion in subset i. Then we may define our information gain as $gain(S) = H(T) – Hs(T)$, or as increase of information brought by division of training set T by potential division S. At every decision node, C4.5 chooses optimal division with highest information gain, $gain (S)$.

4. Data mining in sport

Huge amounts of data are present in all areas of sport. These data may show particular traits of any player, or events that happened during a game, and/or how a team is performing as a unit. It is important to determine which data to store and to comprise a way for their best usage [26]. By finding the best method to obtain new facts from these information and to transform it to a particular data, sports organizations provide themselves a leverage in comparison to other teams [27]. Such approach to knowledge seeking may be applied to a whole organization – from players who may improve their performance using techniques of video analysis, to scouts who use statistic analysis and projection techniques in order to identify which talented youth would develop the most and become a good player [28].

The first part of a problem is to determine performances metrics [29]. A lot of present sports metrics may be used in an inappropriate way (performances are not measured with the aim to score more points than opponent, which is an ultimate goal of every sport organization).

The second part is to find patterns of interest in data observed. These patterns may include tendencies and tactics of opposing players or teams, origin of player's injuries based on monitoring exercise performances, as well as predictions based on earlier data. Professional sports organizations are the multi-million companies and certain decisions are worth large amounts of money. With this kind of capital, a single wrong decision may potentially set them years back. Due to high risk and need to make correct decisions, sport industry is just the right environment to apply data mining technologies.

Different sports associations have varying approach to such data. This approach may be divided into five levels:

- There is no connection between sports data and their use
- Experts from a given field are working on predictions using their instinct and hunch
- Experts from a given field are working on predictions using data collected
- Use of statistics in decision-making process
- Use of data mining in decision-making process.

The first type of approach is when there is no connection between sports data and their use. These sport organizations often obtain certain information about players during their games and they ignore all of it. This is characteristic for amateur sports clubs, since their emphasis is on fun or on introducing the sports basics.

The next type of approach is based on an expert from a given field who is predicting based on his personal experience. It used to be widely accepted notion that these experts (coaches, managers, scouts) might efficiently use their insights and experience in order to reach correct decisions. Decisions made from this type of approach are usually based on predictions or instincts, and not on real data and information. These decisions may include playing certain types of actions or certain player changes since such decision "looks right".

The third type approach is the one when experts start using collected data. Decisions on this level include playing with certain players, for which it was proven that they cooperate well and using actions that score points more often.

The fourth type of approach includes use of statistics as a help in decision-making process. Such statistical measurements may be simple, for instance the frequency of certain events, or complex, dividing performance of a whole team and assigning merits to each player for every game or the competition. Statistics is used as a tool, thus helping experts in making right decisions.

The fifth type of connection between sports information and their use is using the data mining techniques, since they certainly might help predictions. Statistics techniques are still in the core of data mining, but statistics is being used to extract from the background noise a pattern or any other underlying system (inclinations of opponent players). Statistics or statisticians never elucidate relations between such data, since this is a task of data mining. This type of method may be used either in order to help other professionals to make appropriate decision or to make such decisions even without experts. Use of data mining techniques without human influence is often exempted from certain errors. For instance, a

scout may especially appreciate certain qualities in a player, neglecting some flaws. Most of sports organizations use the third or fourth type of approach, somewhere in between data and their use, and only a small number of them use data mining techniques. Although data mining was relatively recently introduced in sport, results of teams who apply these methods are outstanding [30]. Estimations are being done using strong analysis and scientific investigations. Rising number of sports organizations embrace the digital era, and it is possible that sport will soon became a battle of better algorithms or better metrics for performances measurement, so analysts will be equally important as players.

Applying statistics in decision-making process is certainly a step forward in comparison to decisions based on hunch, but statistics may also sway decisions to wrong direction, if there is no knowledge regarding base of a problem. This may happen because of imprecise measurements of performances of due to over-enhancement of certain quality by sports community [31]. For example, certain player may have extraordinary individual statistics, bit he or she still may have only a minor influence on the team as a whole. Sports statistics suffers from imprecision, since statistical metrics may not measure completely influence of all players. For instance, defensive rebound is a measure how many times certain player in defense caught a ball after unsuccessful shot by opponent players. In order to have a defensive rebound, other player from his team must block opponent players and therefore they are equally important in this action. Having in mind the way of noting rebounds, only the player who caught the ball is noted in statistics and rewarded a defensive rebound.

Besides imprecision and incorrect use, another problem in sports statistics is how to determine a risk value. Defense player may risk by sliding in order to intercept the ball. It may result in his fall out of the game so opponent team may have an extra man and score easily. However, if the player succeeds in intercepting the ball, this is a big plus for his team since they obtain a new attack.

4.1. Effect of shot percentage on winning

Basketball is a competitive game between two teams with the aim to win. A win is accomplished by scoring more points than the other team. Sometimes, coaches like to claim that the aim is to receive less points than the opponent does, so the game is won by defense actions. In both cases, the winner is decided by the number of points scored by shots.

Shots may be scored in several ways, and therefore are bringing different points. The hardest to achieve are long-distance shots, so they bring the most points. There is a line drawn on the floor at 6.25 meters from the basket, and shots from outside this line bring three points (in some leagues this border line is drawn even further from the basket in order to be more difficult to score, and therefore make the game more interesting to the spectators). If attempted from inside of this line, every succesful shot brings two points. During a game, sometimes a player is irregularly interrupted by rival players, and this is called a foul. If the foul is done during an attempt to score, or if the team committing the foul have already surpassed the limit (four fouls committed during one period, or quarter),

then fouled player gets a chance to score from the free-throw line. Every shot scored from this line brings one point. Depending on whether a foul was committed while a player was trying to score for two or three points, he or she will have opportunity to try two or three free throws, respectively. [32]

When shooting for two points, three points and when throwing free throws, a player may make the shot or miss the shot, i.e. score or miss, respectively.The ratio between shots and scores is called shooting percentage. In basketball statistics, there are separate percents for one-, two- and three-point shots.

The aim of this paper is to measure effect on shooting percentage from different position on outcome of the game, in order to establish at which position it is most important to be precise.

4.1.1. Data understanding phase

In keeping statistics for the Serbian First "B" basketball league for men, Basketball Supervisor (BSV) software is used. This program allows noting all data relevant for a basketball game. At the end of each quarter, statistics collected by this program is printed and distributed to home and visiting players, commissioner, TV crew (if it is covered live) and journalists covering the game. After the game, all collected data are being sent to the Basketball Federation of Serbia where, they are stored for further analysis.

In this paper we analyzed statistics collected at all the games of the Serbian First "B" Basketball league for men for 2006/07, 2007/08, 2008/09, 2009/10 and 2010/11 seasons. Databases used for storing data are named yubadata_0607, yubadata_0708, yubadata_0809, yubadata_0910 and yubadata_1011.

The database is organized in such way that shooting data are entered to a table *utstat*. Appearance of this table is given at the Figure 3. It comprises of a large number of various parameters, and for us the following are of interest:

- ID_GAME – identification of the game being observed
- ID_CLUB – identification of the club
- ID_PLAYER – identification of the player for which statistics is entered
- P1OK – successfully realized one-point shots for the observed player
- P1SUM – total number of one-point shots for the observed player
- P21OK – successfully realized two-point shots from position one for the observed player
- P21SUM – total number of two-point shots from position one for the observed player
- P22OK – successfully realized two-point shots from position two for the observed player
- P22 SUM - total number of two-point shots from position two for the observed player
- P23OK - successfully realized two-point shots from position three for the observed player

- P23 SUM - total number of two-point shots from position three for the observed player
- P24OK - successfully realized two-point shots from position four for the observed player
- P24 SUM - total number of two-point shots from position four for the observed player
- P25OK - successfully realized two-point shots from position five for the observed player
- P25 SUM - total number of two-point shots from position five for the observed player
- P26OK - successfully realized two-point shots from position six for the observed player
- P26 SUM - total number of two-point shots from position six for the observed player
- P31OK - successfully realized three-point shots from position one for the observed player
- P31 SUM - total number of three-point shots from position one for the observed player
- P32OK - successfully realized three-point shots from position two for the observed player
- P32 SUM - total number of three-point shots from position two for the observed player
- P33OK - successfully realized three-point shots from position three for the observed player
- P33 SUM - total number of three-point shots from position three for the observed player
- P34OK - successfully realized three-point shots from position four for the observed player
- P34 SUM - total number of three-point shots from position four for the observed player
- P36OK - successfully realized three-point shots from position six for the observed player
- P36 SUM - total number of three-point shots from position six for the observed player

This table comprises all data regarding a shot. It does not comprise a final result of the game, i.e. who won. These data are given in table *game*. Parameters of interest in this table are:

- ID_GAME – identification of the game observed
- ID_CLUB1 – identification of the host club
- ID_CLUB2 – identification of the guest club
- SCORE_HOME – number of points scored by the host club
- SCORE_AWAY– number of points scored by the guest club

4.1.2. Data preparation phase

Data in a base are connected to particular players. Within analysis in this paper, we intend to compare the effect of shot precision for one, two and three points on win of observed team. Therefore, it is necessary to sum data regarding players, and to obtain data for a team as a whole.

Before summing up, we will merge tables *GAMESTAT* and *GAME*. This will be done using attribute *ID_GAME* so in every observed line we will have not only existing data, but also data regarding a result.

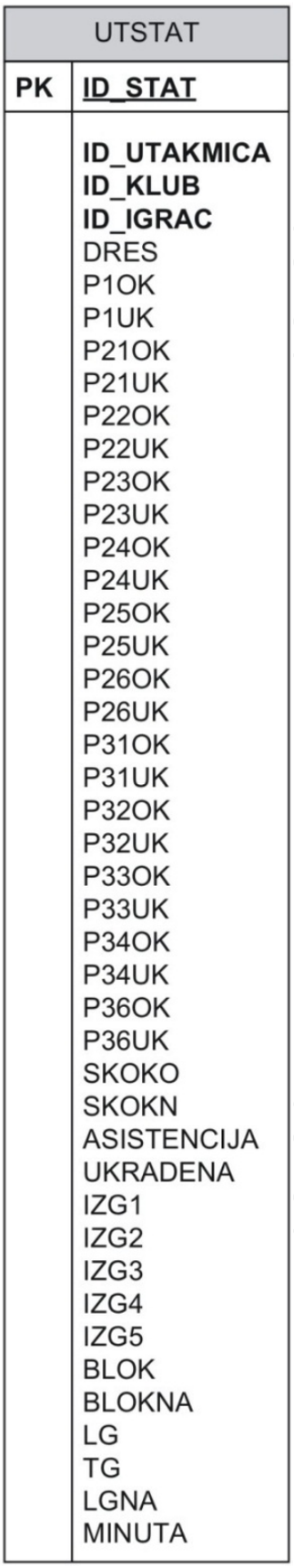


Figure 3. Appearance of the *utstat table*

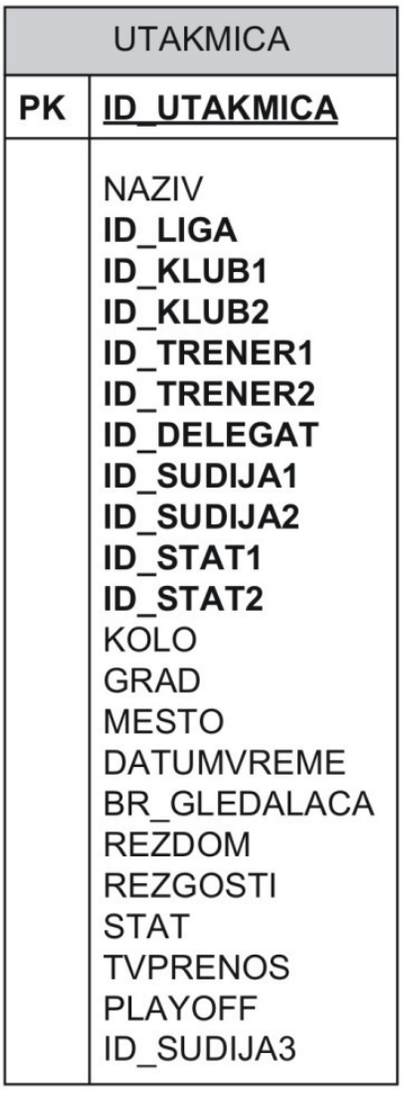


Figure 4. Appearance of the *game* table

Since we are interested in shot percent from certain positions, after summing up data for those positions we will divide values for successful shot from the position to values for total number of shots.

Appearance of a SQL command for selecting appropriate data is as follows:

```
select
        sum(p1ok)/sum(p1uk) p1_procenat,
        (sum(p21ok)+sum(p22ok)+sum(p23ok)+sum(p24ok)+sum(p25ok)+sum(p26ok))/
        (sum(p21uk)+sum(p22uk)+sum(p23uk)+sum(p24uk)+sum(p25uk)+sum(p26uk))
p2_procenat,
        (sum(p31ok)+sum(p32ok)+sum(p33ok)+sum(p34ok)+sum(p36ok))/
                (sum(p31uk)+sum(p32uk)+sum(p33uk)+sum(p34uk)+sum(p36uk))
p3_procenat,
        if(id_klub1=id_klub,
```

```
                    if(rezdom > rezgosti, 'pobeda', 'poraz'),
                    if(rezdom < rezgosti, 'pobeda', 'poraz')) rezultat
from yubadata_0506.utakmica ut, yubadata_0506.utstat st
where ut.id_utakmica = st.id_utakmica
group by ut.id_utakmica, st.id_klub
union
select
        sum(p1ok)/sum(p1uk) p1_procenat,
        (sum(p21ok)+sum(p22ok)+sum(p23ok)+sum(p24ok)+sum(p25ok)+sum(p26ok))/

        (sum(p21uk)+sum(p22uk)+sum(p23uk)+sum(p24uk)+sum(p25uk)+sum(p26uk))
p2_procenat,
        (sum(p31ok)+sum(p32ok)+sum(p33ok)+sum(p34ok)+sum(p36ok))/
                (sum(p31uk)+sum(p32uk)+sum(p33uk)+sum(p34uk)+sum(p36uk))
p3_procenat,
        if(id_klub1=id_klub,
                    if(rezdom > rezgosti, 'pobeda', 'poraz'),
                    if(rezdom < rezgosti, 'pobeda', 'poraz')) rezultat
from yubadata_0607.utakmica ut, yubadata_0607.utstat st
where ut.id_utakmica = st.id_utakmica
group by ut.id_utakmica, st.id_klub
union
select
        sum(p1ok)/sum(p1uk) p1_procenat,
        (sum(p21ok)+sum(p22ok)+sum(p23ok)+sum(p24ok)+sum(p25ok)+sum(p26ok))/

        (sum(p21uk)+sum(p22uk)+sum(p23uk)+sum(p24uk)+sum(p25uk)+sum(p26uk))
p2_procenat,
        (sum(p31ok)+sum(p32ok)+sum(p33ok)+sum(p34ok)+sum(p36ok))/
                (sum(p31uk)+sum(p32uk)+sum(p33uk)+sum(p34uk)+sum(p36uk))
p3_procenat,
        if(id_klub1=id_klub,
                    if(rezdom > rezgosti, 'pobeda', 'poraz'),
                    if(rezdom < rezgosti, 'pobeda', 'poraz')) rezultat
from yubadata_0708.utakmica ut, yubadata_0708.utstat st
where ut.id_utakmica = st.id_utakmica
group by ut.id_utakmica, st.id_klub
union
select
        sum(p1ok)/sum(p1uk) p1_procenat,
        (sum(p21ok)+sum(p22ok)+sum(p23ok)+sum(p24ok)+sum(p25ok)+sum(p26ok))/
        (sum(p21uk)+sum(p22uk)+sum(p23uk)+sum(p24uk)+sum(p25uk)+sum(p26uk))
p2_procenat,
```

```
        (sum(p31ok)+sum(p32ok)+sum(p33ok)+sum(p34ok)+sum(p36ok))/
                (sum(p31uk)+sum(p32uk)+sum(p33uk)+sum(p34uk)+sum(p36uk))
p3_procenat,
        if(id_klub1=id_klub,
                if(rezdom > rezgosti, 'pobeda', 'poraz'),
                if(rezdom < rezgosti, 'pobeda', 'poraz')) rezultat
from yubadata_0809.utakmica ut, yubadata_0809.utstat st
where ut.id_utakmica = st.id_utakmica
group by ut.id_utakmica, st.id_klub
union
select
        sum(p1ok)/sum(p1uk) p1_procenat,
        (sum(p21ok)+sum(p22ok)+sum(p23ok)+sum(p24ok)+sum(p25ok)+sum(p26ok))/

        (sum(p21uk)+sum(p22uk)+sum(p23uk)+sum(p24uk)+sum(p25uk)+sum(p26uk))
p2_procenat,
        (sum(p31ok)+sum(p32ok)+sum(p33ok)+sum(p34ok)+sum(p36ok))/
                (sum(p31uk)+sum(p32uk)+sum(p33uk)+sum(p34uk)+sum(p36uk))
p3_procenat,
        if(id_klub1=id_klub,
                if(rezdom > rezgosti, 'pobeda', 'poraz'),
                if(rezdom < rezgosti, 'pobeda', 'poraz')) rezultat
from yubadata_0910.utakmica ut, yubadata_0910.utstat st
where ut.id_utakmica = st.id_utakmica
group by ut.id_utakmica, st.id_klub;
```

By execution of inquiry, we obtain following data: one-point shot percentage, two-point shot percentage, three-point shot percentage and information whether the team has won or lost the game, in all games in five competition seasons. A part of data obtained is shown in Table 1:

p1_percent	p2_percent	p3_percent	result
0.5652	0.5789	0.3636	‘win’
0.7273	0.3556	0.4211	'loss'
0.6500	0.5517	0.2083	'win'
0.6200	0.4722	0.3333	'loss'
0.7368	0.5641	0.5333	'loss'
0.7368	0.7632	0.5882	'win'
...	...	...	...

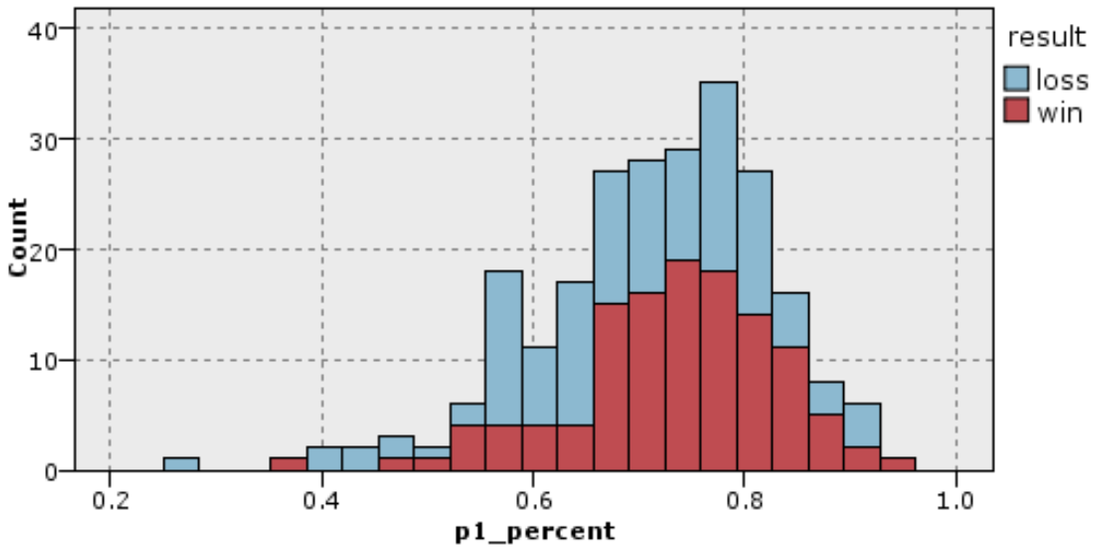
Table 1. Example of modeling data

Total of 1920 lines is obtained. This means that during five seasons, 960 games were played (1920 / 2 = 960), since for every game table contains data for home and guest team.

Following graphs represent analysis of statistic parameters in the First "B" Basketball league of Serbia for men. In this league, 130 games are played per year, including play off and play out. At every histogram, color blue is a number of losses by teams from competition observed. Red color shows number of wins by those teams.

Graph 1. shows effect of one-point shot on final outcome of the game. It is visible that number of wins abruptly rises when one-point shot percent exceeds 65% limit. Since number of wins and losses is approximately linear, regardless of the percent, it may be supposed that one-point shot is not crucial for outcome of the game. If one-point shot percent is below 65%, teams in most cases lose that game.

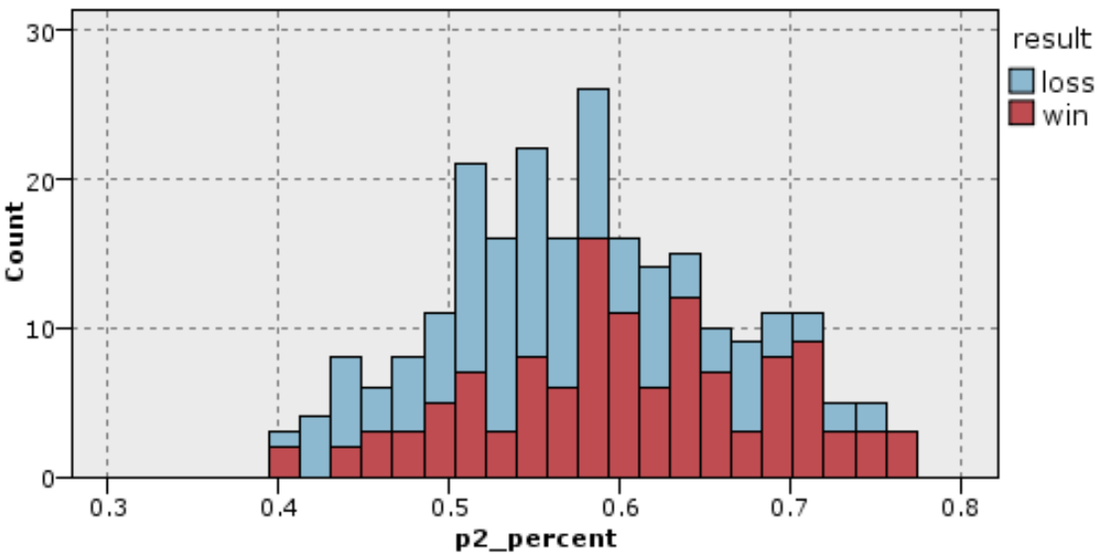
Statistical minimum for one-point shot percent is 25%, and statistical maximum is 96.2%. Average value for this type of shot, regarding all games in this league, is 71.4%. Standard deviation is 0.11.



Scheme 1. Effect of one-point shot percent on final outcome of the game

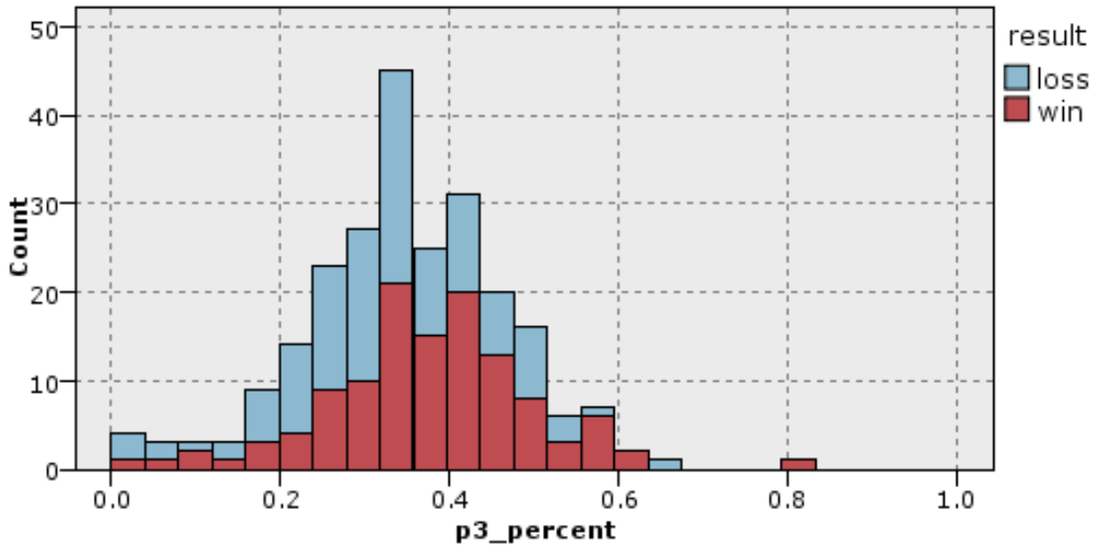
Graph 2 shows effect of two-point shots on final outcome of the game. Blue color is a number of losses by teams from competition observed. Red color shows number of wins by those teams. Graph shows that number of wins rises when a two-point shot percent is above 58%. Therefore, we may conclude that a two-point shot gas significant effect on outcome of a game. If two-point shot percent is below 58 %, number of losses is higher.

Statistical minimum for two-point shot percent is 39.5%, and statistical maximum is 83%. Average value for this type of shot, regarding all games in this league, is 58.2%. Standard deviation is 0.083. Two-point shot percent has higher minimum and lower maximum than one-point shot percent, because during a game shots for two points prevail.



Scheme 2. Effect of two-point shot percent on final outcome of the game

Graph 3 shows effect of three-point shots on final outcome of the game. Blue color in histogram is a number of losses by teams from competition observed. Red color shows number of wins by those teams. Graph shows that number of wins rises when a three-point shot percent is above 38%. If a team has three-point shot percent close to 80%, it will win all games. If this percent is above 59%, the team rarely loses a game. Three-point shot percent between 30 and 59% has less effect on the final outcome of a game, since number of wins and losses is approximately equal. If a three-point shot is below 30%, team loses a game more often. Therefore, we may conclude that a three-point shot percent has significant effect on outcome of a game.



Scheme 3. Effect of three-point shot percent on final outcome of the game

Statistical minimum for three-point shot percent is 0%, when a team does not score three-point shots at all and statistical maximum is 83%. Average value for this type of shot,

regarding all games in this league, is 34.9%. Standard deviation is 0.349 Three-point shot has more extreme values than two-point shot due to lower number of shots during a game.

When data about opponent team are collected, they will pass through three filters and get color and clarity during analyzing, selection, presentation and practicing game plan against this opponent. Every activity done in training should be incorporated into the overall game plan and in purpose of the game. Scouting system must also be incorporated in the demands of the game, and must be highly organized. When scouting, individual characteristics and habits of players from the database collected during the previous season must be taken into account.

1p season 2006/07				**2p season 2006/07**				**3p season 2006/07**			
Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV
24.00	95.20	0.814	0.121	38.50	74.40	0.516	0.071	0.00	85.30	0.311	0.125
1p season 2007/08				**2p season 2007/08**				**3p season 2007/08**			
Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV
0.00	94.10	0.561	0.127	27.10	81.2	0.581	0.101	0.00	74.00	0.290	0.156
1p season 2008/09				**2p season 2008/09**				**3p season 2008/09**			
Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV
24.99	95.80	0.712	0.119	38.50	76.40	0.581	0.081	0.00	81.10	0.399	0.124
1p season 2009/10				**2p season 2009/10**				**3p season 2009/10**			
Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV
23.01	91.21	0.735	0.123	39.50	73.40	0.216	0.091	0.00	82.10	0.411	0.129
1p season 20010/11				**2p season 20010/11**				**3p season 20010/11**			
Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV	Min %	Max %	MEAN	S.DEV
0.00	94.10	0.561	0.127	27.10	81.2	0.581	0.101	0.00	74.00	0.290	0.156

Table 2. Comparative review of one-, two- and three-point shots

Table 2. shows comparative review of one-, two- and three-point shots for First "B" Basketball league of Serbia for men in all 5 seasons observed. For the coach, these parameters are of utmost importance, especially one-point shots. At the every pause during a training, coaches ask players to practice one-point shots in order to improve this segment of the game. Table also shows that in a number of games there were no successful three-point shots, so this segment of a game calls for further improvements. It is also visible that teams were best at free throws in 2008-2009 seasons, at two-point shots in 2007-2008 season, at three-point shots in 2006-2007 season.

5. Conclusion

Basketball game is progressing rapidly. Number of quality players and teams is quickly growing. At high levels of competition, there are no teams that can count on a safe win for every game. Good preparation for the game may mean the difference between the average and best results. Scouting opponents is an important and indispensable element in these preparations. Scouting targets are not only players or team game, but also coaches, who usually have consistent approach to the game (coach philosophy). Use of data mining

techniques provides knowledge about individual qualities of players and their habits, in both offence and defense, so it is easier to predict where advantages of problems will occur in individual offence or defense situations. Having this in mind, a coach may pay more or less attention to certain segment of a game, reducing the number of information and allowing all players of the team the clear and identical idea how to play that game. When number of misunderstandings decrease, power of team play rises.

As a general conclusion of all analyses, it could be said that game under the hoop is a key to winning a game. In defense, it is very important to catch a ball after opponents shot and prevent them from another attack, while in offence it is important to maintain a high level of two-point shots and not to miss "safe shots".

Data collected are applicable for Basketball league of Serbia for men, and such a model may be applied to other leagues of similar quality. It is to be expected that higher-quality leagues (NBA) or those for younger players (juniors or cadets) would create somewhat different models.

Author details

Branko Markoski, Zdravko Ivankovic and Miodrag Ivkovic
University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia

Acknowledgement

This work was partially supported by the Serbian Ministry of Education and Sciences (Grant No: 171039).

6. References

[1] FIBA - Official Basketball Rules. (2011)
[2] Markoski, B., Radosav, D., Vasiljević, P. Milošević, Z.: Košarkaška statistika,. International Conference Dependability and Quality Management ICDQM 2009, Belgrade, Serbia, 776—781 (2009)
[3] Markoski, B., Adžić, Đ.: Razlike u vođenju statistike. Trener X, broj 42, Serbia, 24-27 (2006)
[4] Markoski, B., Adžić, Đ.: Statistika - neki problemi u praksi. Trener XI, broj 46-47, Serbia, 24-28 (2007)
[5] Oliver, D.: Basketball on paper - Rules and tools for performance analysis. Washington DC (2004)
[6] Vasiljević P., Markoski B., Ivanković Z., Ivković M., Šetrajčić J., Milošević Z. "Basket Supervisor - collecting statistical data in basketball and net casting", Technics Technologies education management, Vol.6-2011, 169-178, TTEM, ISSN 1840-1503
[7] Trninić, S.: *Analiza i učenje košarkaške igre*. (1996)
[8] B. Markoski, M. Ivković, P. Vasiljević, P. Pecev, Z. Milošević, "Collecting statistical data in basketball", Information and Communication Technologies for Small and Medium Enterprises, ICT-SME's 2011, Aranđelovac, Republic of Serbia, 2011

[9] José Manuel Sánchez Santos, Ana Belen Porto Pazos, Alejandro Pazos Sierra, "Team Performance in Professional Basketball: an Approach Based on Neural Networks and Genetic Programming", XIII IASE and III ESEA Conference of Sports, Prague, May 2011.
[10] B. Markoski, P. Pecev, L. Ratberg, M. Ivković, Z. Ivanković, "Appliance of Neural Networks in Basketball - Basketball board for Basketball Referees, CINTI 2011 - 12th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, 2011, pp 133 – 137.
[11] Bernard Loeffelholz, Earl Bednar, Kenneth W. Bauer "Predicting NBA Games Using Neural Networks", Journal of Quantitative Analysis in Sports: Vol. 5, Issue 1, Article 7, 2009.
[12] Michael E. Young, "Nonlinear Judgment Analysis: Comparing Policy Use by Those Who Draft and Those Who Coach", Psychology of Sport and Exercise, Volume 9, Number 6, November 2008.
[13] Ratgeber, L. Play from a Game: (Head Coach). Mizo Pecs 2010. 2007/2008. Mizo Pecs 2010 vs. Euroleasing Sopron.
[14] FIBA - Official Basketball Rules. (2012)
[15] Markoski B., Ivetić D., Šetrajčić J., Mirjanić D., Ivanković Z., **"**Košarkaški scauting**",** Infoteh Jahorina BIH, Volume 8, ref. E - III - 24, P. 628-630 mart 2009
[16] Markoski B., Ivanković Z., Ivković M., Ivković F., Nikolić M., Berković I. "Application of on line transmission in basketball", International symposium ZEMAK 2010, FYROM Macedonia,
[17] Bhandari, I., E. Colet, et al. 1997. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery* 1(1): 121-125
[18] David Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001
[19] Peter Cabena, Pablo Hadjinian, Rolf Stadler, JaapVerhees, and Alessandro Zanasi, *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ, 1998
[20] Reed, R.D., and Marks, R.J, II (*1999*), Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, Cambridge, MA: The MIT Press.
[21] Quinlan, J. Ross (1992) *C4.5 Programs for Machine Learning* Morgan Kaufman Publishers Inc.
[22] Fitzgerald, R., Blackall, L., Lyons, K., & Neill, J. (2011). eResearch in **....** Data. *Data Mining* and Knowledge Discovery (*2005*). Ausport, September, 20.
[23] O'Reilly, N. & P. Knight 2007. Knowledge Management Best Practices in National Sport Organizations. *International Journal of Sport Management and Marketing* 2(3): 264-280
[24] Markoski B., Ivanković Z., Šetrajčić J., Vasiljević P., Milićević V., "Razmena on-line podataka upotrebom Web Servisa", Zbornik radova konferencije Informacione tehologije i razvoj tehničkog informatičkog obrazovanja", Zrenjanin 2009, Serbia, 158-162
[25] Peter Cabena, Pablo Hadjinian, Rolf Stadler, JaapVerhees, and Alessandro Zanasi, *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ, 1998
[26] Stefani, R. 1999. "A Taxonomy of Sports Rating Systems", *IEEE Transactions on Systems, Man, and Cybernetics - Part A* 29(1): 116-120.
[27] Branko Markoski, Predrag Pecev, Laszlo Ratgeber, Miodrag Ivković, Zdravko Ivanković, "A new approach to decision making in basketball – BBFBR program", Acta Polytechnica Hungarica Vol. 8, No. 6, 2011