

Business Intelligence and Data Analytics

How can they be used in sports for gaining competitive advantage?

Szerző: Koncz Tamás

Utoljára módosítva: 2014.07.17

Table of Contents

I. Introduction.....	2
II. The framework for case studies.....	3
2.1 Why use analytics?.....	3
2.2 How are sports organizations going through the ‘analytical transition’?.....	7
2.3 Challenges during the transition.....	7
2.4 What skills are needed to be a good analyst?	8
III. Case Studies	9
3.1 Oakland A’s	9
3.2 Orlando Magic.....	12
The Business need.....	12
The solution.....	13
System used	15
Results and the future.....	15
IV. Building a showcase model.....	17
4.1 Data collection	17
4.2 Data Understanding	18
4.3 Models.....	20
4.4 Interpreting results	23
V. Summary and conclusions	23
Sources.....	25

I. Introduction

The use of data and interpreting numbers was always a big interest for me. I like how ‘real life’ events can be described and better understood with the use of statistics and the field of analytics. Also, I am a big fan of basketball. Having played it for ten years and watching games regularly made me wonder – what can be done with all that data that is presented during a game?

Sports fans were always into numbers, and TV channels and sports organisations made sure that they get their load of them. What was interesting to me, is that whether they using are this data. How the very best optimize their strategy given the huge data sets available to them?

To answer these questions I was looking at examples from North-America, home of the Major League Baseball, the National Basketball Association, the National Football League and many other leagues which are undoubtedly the strongest of their kind, and where competition is the fiercest. In this paper I will build a framework of how to use analytics in a sport organisation, look at case studies which proved to be extremely successful, and at the end build my own model for analysing winning factors in the NBA.

II. The framework for case studies

This chapter is based on the books of Davenport, Harris and Morison (2010) and Alamar (2013). As they gave a really well-grounded and thorough overview of the topic, this chapter mostly builds on their works and my thoughts derived from them.

Before going into details, a definition for the term ‘sports analytics’ should be made. I will use Alamar’s definition here, which says that sports analytics is *‘the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their organizations in gaining a competitive advantage in the field of play’* (Alamar, 2013, page 22).

2.1 Why use analytics?

In general, sports teams have two incentives when deciding to use analytics: one is saving time in the decision-making process, and the other one is gaining unique insight about their operation (Alamar, 2013).

There are multiple factors which can reduce the time needed for reaching a conclusive decision. The most important ones are the following:

- **Having ‘one version of truth’:** a common problem for sports organizations was (and in many cases, still is) that the people involved in solving an issue – like answering the question which player to draft – had different information. If they are looking at different factors, then this can be resolved by simply sharing their resources. But what if they can’t agree on the same thing? For example, imagine two coaches arguing about a player’s stamina – they can have completely different opinions if one of them is looking at athletic testing data and the other one is looking at gameplay data. With business intelligent solutions all the data and analysis can be shared, therefore everybody will see the whole picture.
- **Spending less time gathering and organizing, and more time analysing:** previously I may have suggested that sharing information is easy. But don’t let it

fool anyone. It is really, unbelievably hard. Information has to be organized, processed, and made available to make it a benefit, not a burden. The data coming from different sources has to go through the phases of standardization, centralization, and integration.

Once the data is made accessible the analysts can start to work on utilizing it for gaining unique insight. The goal is to help making the right decision – one that is rational and not affected by biases.

To better understand these two concepts (saving time and gaining unique insight) think about a very wide-spread, old-time iconic tool for coaches: the magnet board. It can be used to visualize and share information very easily – even though it has its advantages, this tool still might be outdated. First, it takes a considerable amount of time to set up a magnet board for a complex game situation, trade, etc. This means that using a magnet board is simply too slow nowadays. Second, a magnet board can give you a good overview of a topic, but it cannot be used for further investigation – common features, like drill-down in dashboards offer much better solutions for these kind of uses.

The different uses of analytics can be group by two dimension, time frame and innovation power (Davenport, Harris, & Morison, 2010):

	Past	Present	Future
Information	What happened? (Reporting)	What's happening now? (Alerts)	What will happen? (Extrapolation)
Insight	How, why ...? (Modeling)	What's the next best action? (Recommendation)	What can happen? (Prediction, Optimization, Simulation)

1. Table: Information vs. Insight in different time frames

Gaining insight is more than just generating information in the sense that insight is something that helps understanding and unrevealing the causes. Among others, it can be leveraged to:

- **Know what's really working:** relying on expertise is not bad, but for understanding complex problems modelling and analytical tools can be a considerably large advantage.
- **Cut costs, improve efficiency:** If you know what the key performance drivers are, you can save time, effort, and money by not wasting them on something else.
- **Manage risks:** by identifying and measuring potential risk factors.

A common point where teams first use analytics is picking the best players at the draft (almost every team in the big leagues use analytics for player selection). Figure 1 can help us understand this trend better: In the last twenty years, NBA salary caps¹ increased to more than tenfold of their level in the mid-eighties (teams usually pay almost the full possible amount to their players). The growth rate in player salaries highly exceeded inflation and also the growth in other sectors. Players cost more than ever, which makes 'experimenting' extremely expensive. From a business perspective, only teams that have a high success rate in selecting the players with the highest potential for future success can survive.

As Bill Gates put it: *"The most meaningful way to differentiate your company from your competitors, the best way to put distance between you and the crowd is to do an outstanding job with information. How you gather, manage and use information will determine whether you win or lose."* (Gates, 1999)

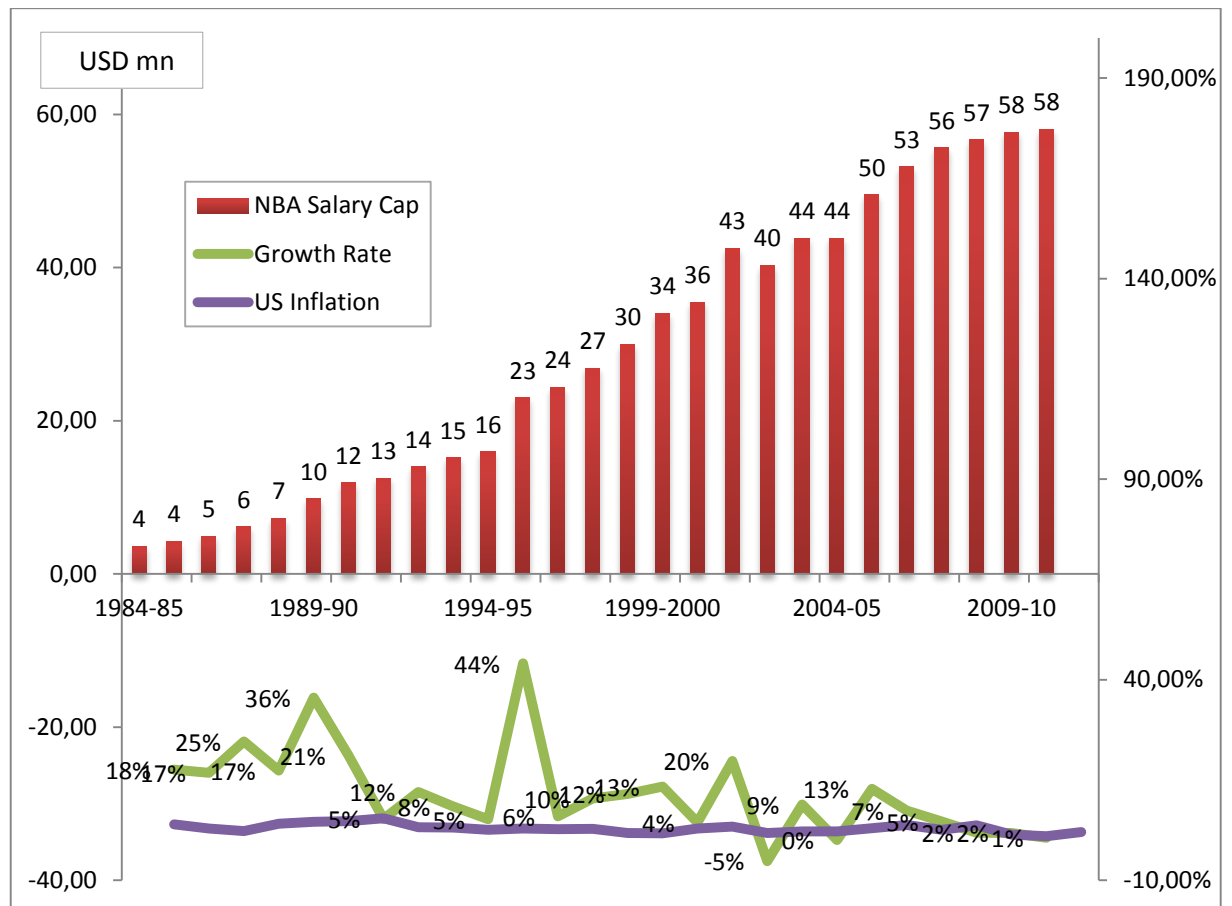
Analytics at Work (Davenport, Harris, & Morison, 2010) gives a good guidance about which business processes lend themselves to analytics:

- **Data-rich.** In modern sports everything is getting measured.
- **Resource-intensive.** Increasing salaries and short supply of talent is the new era for every major sport league in North-America.
- **Depend on spend and timing.** Playing a new team every other team definitely creates time pressure on those who have to come up with the winning strategy.

Low average success rate: Only one team wins the championship every year. Only a handful of player will be really as good as scouts predicted. Only... We can see the trend

¹ Salary caps are the highest amount of money that a team can pay in salary to its players in a given year (salaries are also regulated on the individual level in the NBA)

here (Analytics at Work actually uses batting-rate from baseball as a metaphor – I think this is really telling).



1. Figure: NBA Salary Cap Growth vs. US Inflation

2.2 How are sports organizations going through the ‘analytical transition’?

Davenport, Harris, & Morison (2010) identifies five key factors for success, grouped under the DELTA acronym:

- D: Accessible, high quality **data**
- E: **Enterprise** orientation
- L: **Leadership**
- T: Strategic **Targets**
- A: **Analysts**

Data’s importance is quite straightforward: ‘garbage-in garbage-out’ is true for every field of modelling and analysis. Without an enterprise-wide perspective, it is really likely that the new initiatives will stay local in the organization. Leadership is needed for good change management: analytics has to become a habit, deeply embedded in the culture of the organization to really have an effect. Specific targets are needed because analytical resources are in short supply, and teams shouldn’t waste them on minor aspects of their operations. Lastly (but not being less important than the factors before), analysts have two functions: building the models for decision making and making decision makers appreciate and apply them.

2.3 Challenges during the transition

Three main factors can be identified where the transition can go wrong:

- **Leadership:** even if you have all the analytical capabilities that one can imagine, it won’t help unless you **use them to make and execute better decisions** (Davenport, Harris, & Morison, 2010). Sports organizations are built having common assumptions and relying on conventional wisdom. Challenging the status quo can lead to resistance against the innovative solutions - especially in rather hierarchical organizations like professional teams. A very strong backer (a project champion) is needed to overcome these obstacles.
- **Understanding:** modern algorithms, methods and models can operate as black boxes to the key decision makers in these organizations, which can lead to strong

reluctance to use the results of analysis. People have to get comfortable with trusting computers with tasks they have been long making, and in fact believe that their decisions can be better and more consistent.

- **Recruiting talent:** Analytics is a new thing in the life of a general manager or a coach. This is natural and does not cause a big problem if the right talent is present in the organization for developing the new culture. But who can recruit that talent? This task falls on people who never used analytics, therefore don't have the necessary experience for identifying the needed qualifications and people who have them.

2.4 What skills are needed to be a good analyst?

A good applicant definitely has to have the quantitative background and technical foundation, although this is far from enough. He or she also needs a satisfying knowledge of the given sport. Design skills for developing easily understandable charts, dashboards, etc. are also welcomed. Coaching and staff development will also be in an analysts' job description, so being competent at teaching is something that can give a great lift to the team which wants to speed up its transition. Last but not least, good communication and relationship skills are essential for getting the key organizational players start to use new measures and models.

III. Case Studies

3.1 Oakland A's

Everybody heard of 'Moneyball' in the field of sports – the experiment of Billy Beane and the team he was the General Manager of, Oakland A's completely redefined the modern thinking about the science of sports.

Their situation at the time can be described briefly as the following: given is a small market team with one of the smallest budgets (around \$40 million) of the MLB, which has to battle in the big leagues with teams that can overspend them without thinking (for example, the New York Yankees had a budget of approximately \$150 million, not including a \$100 million in reserves). A study on the economics of baseball (Levin, Mitchell, Volcker, & Will, 2000) showed that difference between the poorest and richest teams is the widest in baseball (4:1), and much smaller in basketball (1,75:1) and American football (1,5:1).

This posed an enormous challenge for teams like the A's: they couldn't follow what the big teams did - since they only had the fraction of their money, they would lose every time against them.

The problem lied in the traditional ways of operations of the baseball organisations and their very hierarchical structure.

Player scouts defined who can play the game – therefore defined how the game is played. It can be effortlessly seen that picking the right players to play is the most important factor for winning in baseball.

But this was not the case with scouts. For instance, they had the tendencies to prefer high school pitchers, who had almost no chance of becoming great players. Scouts' inability to choose wisely can be derived from common psychological traits of the human nature (Anonymus, 2014):

- people tend to overgeneralize their own experience
- recent happenings have more influence on them than earlier ones
- bias towards thing seen with one's own eyes

During a draft, scouts would go over hundreds of names, and on most of them they have no deep knowledge. Because of this they used a 2 step selection process:

1. Breakdown to 2 groups: no-gos and players worth consideration
2. Elimination, when they ranked players by post and only the best remain in their pool

Instead of relying on instinct, Billy Beane wanted an understanding of how the game works, using scientific investigation.

The question was: *what is the **efficient** way to spend money on players?*

To answer this, they went back to Bill James, the inventor of “sabermetrics”, the quantitative science of baseball. His work was long overlooked by managers, and only fans read his writings. The teams did not get that the point was **understanding**, not statistics. Bill James challenged conventional wisdom (and traditional metrics), arguing that things could be much better done than they were.

The Oakland A’s believed him. They started using his “on-base percentage” metric, which was the probability that a batter will not make an out. This became the most important one-dimensional statistic of baseball.

While the baseball salaries were skyrocketing, the radical advances in computer technology shrunk the costs of compiling and analysing vast amounts of baseball data. The A’s were taking advantage of these advances by calculating the most important metrics for every player in their scope and evaluating them continuously.

Since they know that scored runs mattered more than batting average (the metric used by almost everybody in the MLB to decide a player’s value) they realized that the market for players is inefficient. It cared too much about looks, so the A’s started looking for players who:

- don’t look good in uniform
- couldn’t play anything else but baseball (but were pretty good at playing baseball itself)
- gone to college (as it was mentioned, for some reason scouts preferred high school players, therefore the price of college players were lower than it should have been)

The other big inefficiency was that defensive metrics were overvalued.

The A's signed AVM systems which provided a solution for recording everything that happened on the baseball field. They came up with the model which allowed them to see that how anything that happens on the field contributes to the team's chances of scoring a run (Lundegaard, 2014). They could reduce the game to a social science.

The results were staggering: only two stats, offensive ones (on-base % and slugging %) seemed to be in significant correlation with winning percentage. Once they found out that the ability to get on base was under-priced, they started operating as a Wall-Street broker smelling a big opportunity for risk-free gains. The finding that the variance between the best and worst fielders is a lot smaller than between the best and worst hitters seemed to be showing that the market overpriced defence.

It would not have been this easy to exploit these results if the baseball teams had been operating rationally. But they did not – they were not looking for new ways to improve their operations. Other teams assumed that Oakland could not afford “normal” players, and this is the reason why they were after odd ones. The Oakland, on the other hand, was trying to recreate the “aggregate”, not the individual players.

The structure of the MLB season also played to Billy Beane's hand – he had a phenomenal ability to improve mid-season. Why? Because the market allowed him to do things he couldn't other times. As bad teams lost hope during the season, they had a desire to cut costs by selling players – and the Oakland went on a shopping spree to get the best of them. They put an exact \$ figure on every player, to know how much they worth to them. Just as it is with trading bonds or stock, the one who has better information could make a killing. Oakland had better information than everyone else.

As always, the experiment was messier than the conducting scientist is usually willing to admit. The big problem with statistics is that one always relies on the sample size, and has to accept that past performance is an uncertain guide to the future.

Oakland could rule the regular season – it made history on 2002.09.04 when it won its 20th game in a row. But it did not help them in the playoffs, where teams played a lot less games, resulting in a smaller sample size for analysis, and higher volatility (Terbush, 2014).

Although the A's has never won a championship, they could change the way modern sports are played. After Moneyball was out (Michael, 2003) Billy Beane was a role model for many of the General Managers of sports teams in North-America – almost everybody cared about him, but baseball teams. But they had to follow, which eventually led to the overvaluation of once undervalued metrics – keeping the competition to come up with new and better metrics alive.

3.2 Orlando Magic

The second case study will be about the NBA team Orlando Magic. The Magic introduced its system years after the Oakland A's, and its uses are rather focused on the fan experience than on the sport itself (although the system has function which can be leveraged by the coaches).

The Business need

As any other small market NBA team (or for that matter, any sports team) the Magic was struggling to build up a big enough revenue stream which would allow it to compete with the teams from bigger markets.

The introduction and growth of a formal secondary game ticket market by the NBA hit every team at their weakest spot: season tickets were losing value, since it became easier to obtain a ticket any time somebody wanted to. Buying otherwise unused ticket from a trustworthy source is also possible now, which decreases ticket revenues for the teams as well.

For Orlando Magic, ticket sales are making up 28% of total revenues and 80% is coming from season ticket holders (Crawford, 2014). Since it is the single biggest revenue stream for the team, it had to come up with a solution.

The solution

Moving to a new arena in 2010 gave the Magic a chance it could not miss: they turned to Business Intelligence, and with SAS they implemented a system which is utilized in every portion of the business (Anonymus, Sas.com, 2013).

3 goals were set by the team:

- Improve **fan engagement**: get to know and focus on what fans want for their experiences in attending games
- Identify and exploit **new revenue streams**
- **Fill** the most **seats** possible (ideally, all of them for every home game – 41 in the regular season, and a couple more during the playoffs)

The Magic wanted to get **data driven insight**, so they decided to enter a partnership with SAS, which would provide the needed software platforms and help develop the hardware background. The below quote shows the team's satisfaction with its choice:

"Our biggest challenge is to customize the fan experience, and SAS helps us manage all that in a very robust way," says Alex Martins, CEO of the Orlando Magic (Anonymus, Sas.com, 2013).

A few years before the new system the team had to overcome the following challenges in its everyday operations (Crawford, 2011):

- Multiple data systems were present with no or very little communication, therefore the analysts had to spend a lot of time and effort with merging the segmented data into one
- Reports were prepared manually
- Lack of data inventory
- The above and the lack of essential tools resulted in very limited data mining capabilities

A new data warehouse was developed, which included internal (e.g. revenue streams) and external (e.g. secondary ticket market) data sources.

The main data collected are from (Anonymus, kdnuggets.com, 2010):

- ticket sales
- operations
- retail
- sponsorships
- fan demographics
- social networks
- website visits

Instead of the printed tickets an electronic card was introduced, making it possible to collect and connect data of fans (Kyler, 2014).

A real-time ‘heat map’ was developed, where managers can track ticket and suite sales during the games. It is also possible to drill-down from the aggregated data, therefore helping the understanding of customer behaviour – this is a huge improvement to the previously used every-24-hour Excel reports which lacked any such functionality (Anonymus, Sas.com, 2013).

To boost revenues the team needed to understand what people were doing once they were inside the building. Using the new electronic ticket system (matching bar codes) it now can link ticketing with merchandise, and food & beverages (Anonymus, Sas.com, 2013). With all these information the Magic now scores season ticket holders and provides up-to-the-minute standardized reports on them, which are accessible for all the relevant decision makers.

For example, based on the purchasing data of a customer and the renewal patterns, a decision tree was build which buckets subscribers to 3 categories:

- most likely to renew
- least likely to renew
- fence sitters, or in other words, potential customers who are likely to be undecided now, but has a considerable chance of buying

This way the team can predict which season ticket holders are at risk, and take action accordingly.

The Magic also uses the new business intelligence system for the following:

- Price tickets: with the availability of the secondary ticket market data, the team is in a better place to efficiently price tickets in the first place, therefore generating more revenue and value
- Analyse concession and product merchandise sales, and leverage this information to optimize inventory, decide what products to hold, and to deliver better product placement in the shops
- Line-up optimization: or simply put, moneyballing. Without covering it in detail, it has to be mentioned that the coaches also have access to the system, and they can built models for understanding opponents and the best strategies in the basketball court

The Magic's analyst team operates as an in-house consulting team, so everyone in the organization can ask for their help and use their resources to derive business value.

System used

The following SAS solutions are implemented (Anonymus, kdnuggets.com, 2010):

- **SAS Data Integration Environment**, for the data warehouse, data collection, consolidation, integration
- **SAS for Patron Value Optimization**: originally developed for gaming and hospitality companies, it helps to maximize customers' lifetime value and satisfaction
- **SAS Marketing Automation**, for planning, testing and executing marketing campaigns efficiently and with ease
- **SAS Enterprise BI Server**, for portals, preparing report and making dashboards
- **SAS Miner**, for data mining activities

Results and the future

With the solutions described above Orlando – although just the 20th biggest market – is among the top revenue earners in the NBA. Its ticket sales are the 7th biggest (Anonymus, Sas.com, 2013). The Magic together with SAS redefined BI as “Basketball Intelligence”.

One of the most important achievements is that the decision makers in the organization are now more confident about what strategies to follow and in making operational decisions (McCharty, 2014).

The Magic cannot lay back and celebrate the result though – others are catching up quickly, most notably the Phoenix Suns, which has a similar program. The league also set up a marketing and business operations team to share information and best practices throughout the teams (Kyler, 2014).

The Orlando Magic can't rest reassured that its competitive advantage will last long – even though it now has a system which many 'real' business would envy, competition is fierce every other team can see how far utilizing data can get you in the NBA.

IV. Building a showcase model

In this part I will introduce an artificial case study, and show how analytics could be used for helping the decision maker's work. As basketball is the sport closest to my heart (and the one I have the most expertise in) I choose it for conducting the analysis. The question I'm trying to answer is the following: "how can a NBA team leverage game play data to gain advantage in its strategy?" - putting it in other words, what factors lead to winning in the NBA.

4.1 Data collection

To collect data I was using Excel and VBA to automatize the collection process. The data I collected is from <http://www.basketball-reference.com/>, where every game's Box Score and other advanced statistic measures can be found. With the help of VBA I wrote a macro which collected data for all of the games of the 2012-2013 regular season. The sample was of 1229 games, played by the 82 teams in the NBA. My target variable is whether the team won or not (represented by the binary variable 'HOME_WIN'), and I had 64 predictors (although in the end I only used 54 of them because of trivial relations between some predictors and the target, like it was the case with scored field goals).

A brief summary of the used variables can be seen below:

<i>Variable Summary</i>	Class	Levels	Min.Level.Size	Mean	SD
<i>HOME_WIN.</i>	factor	2	477	0,39	0,49
<i>MAX_PTS_1</i>	numeric	NA	NA	0,24	0,06
<i>MAX_PTS_2</i>	numeric	NA	NA	0,24	0,05
<i>MAX_FGA_1</i>	numeric	NA	NA	0,22	0,05
<i>MAX_FGA_2</i>	numeric	NA	NA	0,21	0,04
<i>FGA_1</i>	integer	NA	NA	81,96	7,08
<i>FG_pct_1</i>	numeric	NA	NA	0,45	0,06
<i>FTA_1</i>	integer	NA	NA	21,68	7,03
<i>FT_pct_1</i>	numeric	NA	NA	0,76	0,10
<i>ORB_1</i>	integer	NA	NA	11,04	3,87
<i>DRB_1</i>	integer	NA	NA	30,53	5,09
<i>TRB_1</i>	integer	NA	NA	41,57	6,26
<i>AST_1</i>	integer	NA	NA	21,43	4,94

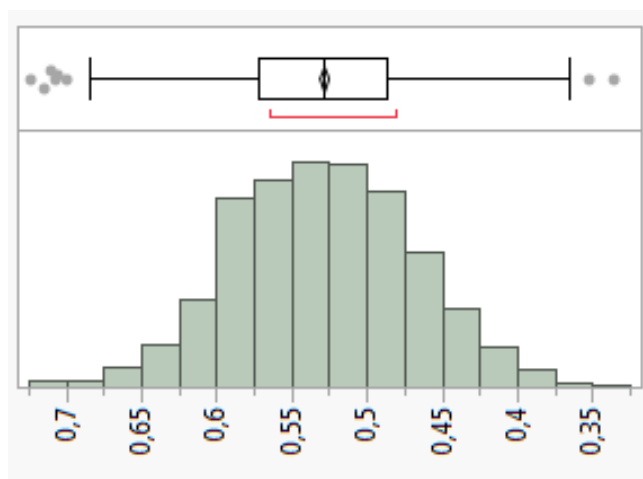
<i>STL_1</i>	integer	NA	NA	7,70	2,71
<i>BLK_1</i>	integer	NA	NA	4,90	2,71
<i>TOV_1</i>	integer	NA	NA	14,20	3,75
<i>PF_1</i>	integer	NA	NA	20,15	4,36
<i>FGA_2</i>	integer	NA	NA	82,08	7,07
<i>FG_pct_2</i>	numeric	NA	NA	0,46	0,06
<i>FTA_2</i>	integer	NA	NA	22,70	7,37
<i>FT_pct_2</i>	numeric	NA	NA	0,75	0,11
<i>ORB_2</i>	integer	NA	NA	11,30	4,05
<i>DRB_2</i>	integer	NA	NA	31,41	5,22
<i>TRB_2</i>	integer	NA	NA	42,71	6,58
<i>AST_2</i>	integer	NA	NA	22,84	4,80
<i>STL_2</i>	integer	NA	NA	7,91	2,84
<i>BLK_2</i>	integer	NA	NA	5,38	2,77
<i>TOV_2</i>	integer	NA	NA	13,70	3,59
<i>PF_2</i>	integer	NA	NA	19,54	4,17
<i>TS_pct_1</i>	numeric	NA	NA	0,53	0,06
<i>eFG_pct_1.2</i>	numeric	NA	NA	0,49	0,06
<i>ORB_pct_1.2</i>	numeric	NA	NA	25,82	7,57
<i>DRB_pct_1</i>	numeric	NA	NA	73,27	7,96
<i>TRB_pct_1</i>	numeric	NA	NA	49,34	5,25
<i>AST_pct_1</i>	numeric	NA	NA	58,55	10,60
<i>STL_pct_1</i>	numeric	NA	NA	8,34	2,85
<i>BLK_pct_1</i>	numeric	NA	NA	8,05	4,59
<i>TOV_pct_1.2</i>	numeric	NA	NA	13,43	3,41
<i>TS_pct_2</i>	numeric	NA	NA	0,54	0,06
<i>eFG_pct_2.2</i>	numeric	NA	NA	0,51	0,07
<i>ORB_pct_2.2</i>	numeric	NA	NA	26,73	7,96
<i>DRB_pct_2</i>	numeric	NA	NA	74,18	7,57
<i>TRB_pct_2</i>	numeric	NA	NA	50,66	5,25
<i>AST_pct_2</i>	numeric	NA	NA	60,47	9,41
<i>STL_pct_2</i>	numeric	NA	NA	8,56	2,96
<i>BLK_pct_2</i>	numeric	NA	NA	8,70	4,50
<i>TOV_pct_2.2</i>	numeric	NA	NA	12,96	3,30

2. Table: Variables used in during the model building

4.2 Data Understanding

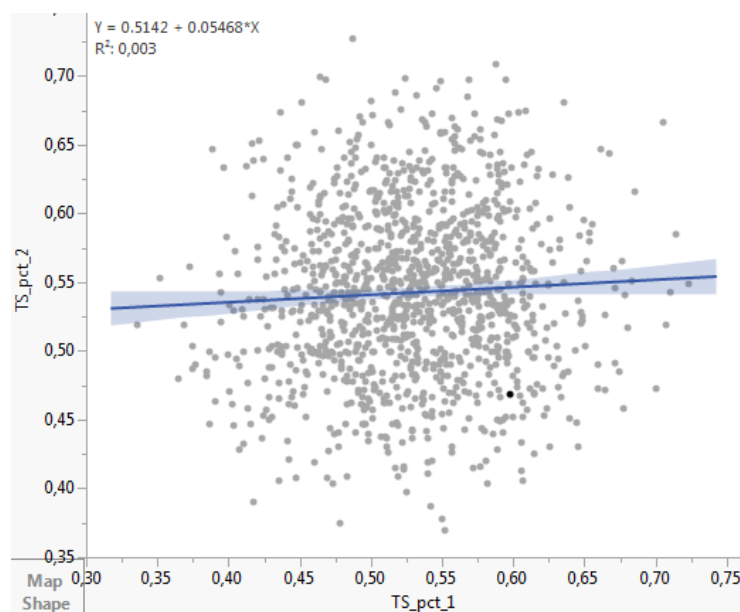
Before I started building my models I wanted to have a deeper understanding of the main variables I was going to use. I used JMP and its visual capabilities to their distributions and how they connect to each other.

In this paper I only have space to briefly cover the methodology and technics I was using. For this I chose two metrics. The first one is the true shooting percentage, which is the number of field goals per field goal attempts adjusted for the value for the 3 pointers and free throws. It has a mean of 53% with a standard deviation of 6%, and as it can be seen below, the distribution is really close to normal.



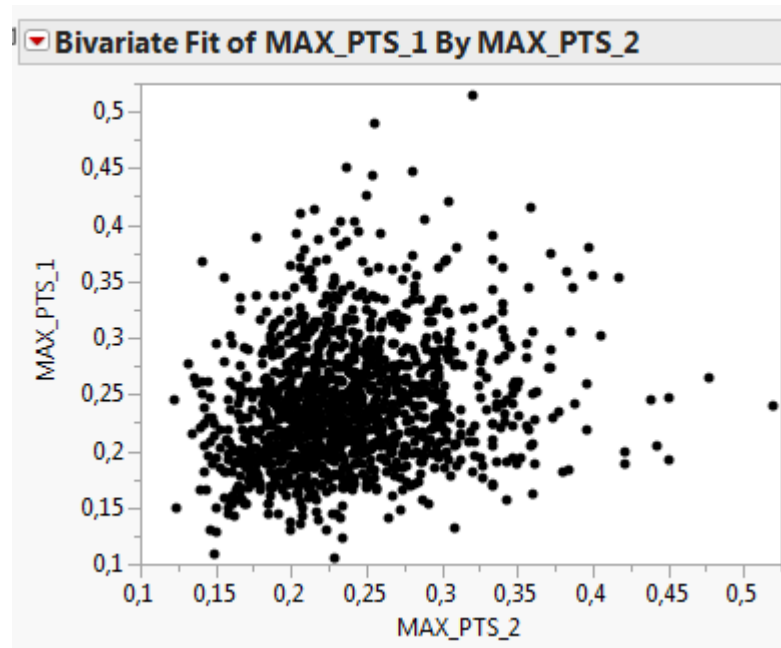
2. Figure: Distribution of true shooting %

I was interested in how the two teams TS% is connected to each other – but, as it can be seen on the below graph the relationship is rather random (the regression line is almost flat while the R squared is low).



3. Figure: True shooting percentages

The second metric is the ratio of the points scored by the best scorer of the team against the total points scored by the team. The interesting thing here is that there is a correlation between the ratios for the two teams, as it can be seen in figure 4. It is hard to interpret without further context, but one reason can be that the two best scorers are defending each other, and they put more emphasis on scoring than on defence.



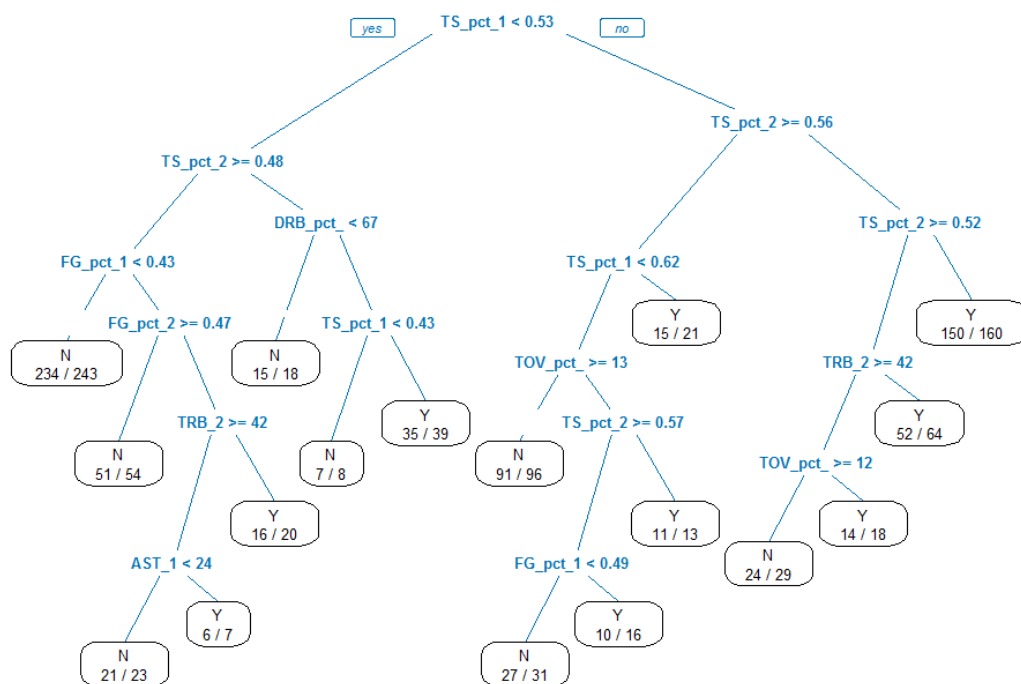
4. Figure: Scoring ratio of each team's best scorer

4.3 Models

With the help of R and its RcmdrPlugin.BCA GUI package (Putler & Krider, 2012) I was building three models:

- Logistic Regression
- Decision Tree
- Artificial Neural Network

The goal was to optimize the logistic regression as much as possible, since it gives the best understanding of each variable's effect on the target. I used the decision tree and the ANN for capturing non-linearities.



5. Figure: Decision Tree

Based on the decision tree the following variable seemed to be having the biggest effect on winning: True Shooting Percentage, the higher the better, but it is equally important to keep the other team shooting efficiency low.

The next step was the logistic regression, where I used the built-in stepwise variable selection method for picking the best predictor variables.

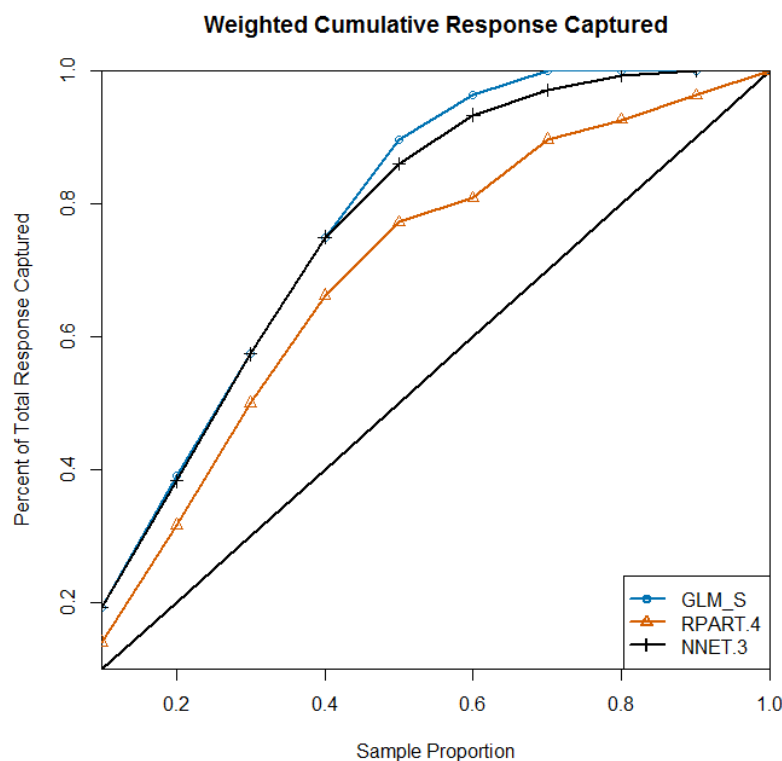
	<i>Estimate</i>	<i>STDE</i>	<i>z</i>	<i>value</i>	<i>Pr(> z)</i>
<i>AST_1</i>	0,19	0,03	6,34	0,0000	***
<i>AST_2</i>	-0,14	0,03	-4,42	0,0000	***
<i>DRB_1</i>	0,47	0,04	10,88	0,0000	***
<i>DRB_2</i>	-0,42	0,04	-10,97	0,0000	***
<i>FTA_2</i>	-0,11	0,02	-5,09	0,0000	***
<i>MAX_FGA_1</i>	6,58	2,69	2,45	0,0143	*
<i>ORB_2</i>	0,12	0,05	2,43	0,0150	*
<i>PF_2</i>	0,10	0,05	2,21	0,0272	*
<i>TOV_1</i>	-0,47	0,05	-10,13	0,0000	***
<i>TOV_2</i>	0,39	0,06	6,40	0,0000	***

6. Figure: Results of the logistic regression

As per the table above, the following variables deserve attention:

- Number of assists: assist is the pass which leads to scoring – therefore it is not surprising that having more is positively correlated with winning. Also, if the other team can give more assists it is bad for the team's chances of winning.
- Defensive rebounds: after the opponent misses a shot, the defending team needs to capture the rebound in order to not give a second chance for scoring in the same play. Usually offensive rebounds lead to easy points (as they are captured close to the basket) so it is extremely important for the defence to minimize these opportunities for the other team.
- Turnovers: Turnovers are not just wasted offensive plays, but gives the chance of easy scoring against a not structured defence

After I ran the ANN model I used a lift chart for comparing model performance:



7. Figure: Lift chart

The optimized General Linear Model was performing really similarly as the ANN, while the decision tree had the worst performance. It can be a bit surprising, but seems like that the decision tree was most prone to be overfitted, so it gave worse performance on the

validation sample. Based on the lift chart I concluded that there is not much left to optimize on the logistic regression, which was already performing well.

4.4 Interpreting results

What can we learn from the above metrics?

- Giving assists matters: Assists are a good measure of team play – it means that the team can build scoring opportunities not just from individual plays, but from connecting teammates. This leads to a more balanced offensive front, which is harder to defend than one-on-one plays.
- A team has to appreciate the ball: When getting a rebound you are giving yourself a chance to lead the offensive instead of your opponent. Also, by minimizing turnovers, a team can have more scoring attempts (and avoid suffering fast breaks). Both rebounds and turnovers have a multiplier of around 0.4 – this suggests that there is no real difference between how you capture the ball, an interesting result for building defensive tactics.
- The below table shows five players ranked by the logistic regression's equation – all of them are believed to be among the top players in the league. The two very best, Kevin Durant and LeBron James might have ranks a bit lower than expected. This can be caused by the fact that direct scoring measures had to be excluded from the model because of their high correlation with the target variable. Overall, it seems like that the model gives reasonable results even on the individuals' level.

<i>RANK</i>	<i>Player</i>	<i>VALUE</i>
1	Kevin Love (MIN)	4,78
2	DeAndre Jordan (LAC)	4,63
3	LaMarcus Aldridge (POR)	4,13
19	Kevin Durant (OKC)	3,03
25	LeBron James (MIA)	2,87

V. Summary and conclusions

As we've seen in this paper the most innovative teams already started leveraging analytics in their operations. The very best not only use it to optimize their playing strategy, but to

optimize their business as well – this is a very important use for them as the post-crisis world challenges their revenue streams.

During the modelling exercise a very basic understanding of basketball strategy could be built – to go deeper more context is needed. In the future I plan to use a bigger sample (using data from more seasons, and also look at the playoffs) with different metrics to revisit my model.

The major leagues in North America are already big consumers of analytics (sometimes they are even leaders of this field for other industries, as the Oakland A's showed us) – I expect this trend to strengthen in the future and that more and more sports will use analytics for better performance.

1. Figure: NBA Salary Cap Growth vs. US Inflation.....	6
2. Figure: Distribution of true shooting %	19
3. Figure: True shooting percentages	19
4. Figure: Scoring ratio of each team's best scorer	20
5. Figure: Decision Tree.....	21
6. Figure: Results of the logistic regression	21
7. Figure: Lift chart	22

Sources

- Alamar, B. C. (2013). *Sports Analytics: A Guide for Coaches, Managers, and other Decision Makers*. New York, USA: Columbia Press University.
- Anonymus. (downloaded at 2014. 03 23). <https://www.boundless.com/management/decision-making/barriers-to-decision-making/personal-biases/>.
- Anonymus. (downloaded at 2014.01.10).
<http://www.businesswire.com/news/home/20101206005116/en/ADDING-REPLACING-NBA%E2%80%99s-Orlando-Magic-Bring-SAS%C2%AE#.UuWF3RDH9dh>
- Anonymus. (2014. 01 21). <http://www.kdnuggets.com/2010/12/nba-orlando-magic-sas-analytics.html>
- Anonymus. (2014. 01. 20). http://www.sas.com/en_us/customers/orlando-magic.html
- Crawford, M. (2014. 01. 22). <http://blogs.sas.com/content/sascom/2011/04/25/orlando-magic-and-nba-shaped-by-sas-analytics/>
- Davenport, H. T., Harris, G. J., & Morison, R. (2010). *Analytics at Work: Smarter Decision, Better Results*. USA: Harvard Business School Publishing Corporation.
- Gates, B. (1999), <http://www.luchzanirato.com/business-strategy-advice-from-worlds-billionaires/>
- <http://www.usinflationcalculator.com/inflation/historical-inflation-rates/>. (2013. 12.04).
- insidehoops.com. (2013. 12. 04). <http://www.insidehoops.com/nba-salary-cap.shtml>.
- Kyler, S. (2014. 01. 21). <http://www.usatoday.com/story/sports/nba/2013/03/06/nba-marketing-fans-business-orlando-magic-phoenix-suns/1968975/>
- Levin, R. J., Mitchell, J. G., Volcker, P. A., & Will, F. G. (2000). *The Report of The Independent Members of the Commissioner's Blue Ribbon Panel on Baseball Economics*. Major League Baseball.
- Lundegaard, E. (2014. 0 23). <http://eriklundegaard.com/item/imoneyballi-revisited>.

McCharty, V. (2014.01.20). <http://www.idevnews.com/stories/4467/NBAs-Orlando-Magic-Bring-SAS-Analytics-Center-Court>

Michael, L. (2003). *Moneyball: The art of winning an unfair game*. New York, USA: W. W. Norton & Company, Inc.

Putler, D. S., & Krider, R. E. (2012). *Customer and Business Analytics*. USA: Chapman & Hall / CRC.

Terbush, J. (2014. 03 23). <http://theweek.com/article/index/251077/why-doesnt-moneyball-work-in-the-playoffs>.