SMAI Project Report: Applying Machine Learning in the NBA

Authors

Indraneil Paul, Akshay Pawale and Nayan Joshi

Abstract

The explicit aim of the semester project was to find a series of methods and metrics by which to better model the National Basketball League games. We abandon the traditional metrics that are conventionally thought to be a good measure of individual player's worth and go in search of a completely new method that we develop from scratch, to quantify a player's value in the context of what he brings to the team and how he skews the results in his presence. Armed with these unconventional and revolutionary views of a player's ability we then sought to model the positive and negative synergies between these players which gave us new insights into our team modelling process and permitted us to model a team as more than just a sum of its parts. Finally using this information and the respective histories of the two teams and the adjudged abilities of the respective coaches, we seek to bring all of our work together and factor in field and venue advantages to try and predict the result of a hypothetical encounter between any two teams in the Atlantic Division of the Eastern Conference of the National Basketball League.

Introduction

Arguably the most important aspect of our lives that machine learning has had the greatest impact on is the ability to predict the outcomes of seemingly unpredictable events. It is with this in mind that we sought to apply various data processing and learning methods to be able to answer as well as possible questions such as – Who would win tomorrows match between Golden State Warriors and Boston Celtics? In the presence of highly structured data the NBA and its affiliates maintain there have been several attempts at trying to model the world of basketball in such a way that allows us to predict games and makes the unforeseeable, foreseeable and hence opening up a whole new world of possibilities in the world of sports analytics, sports betting and sports coaching.

Arguably the greatest challenge that the people attempting to have this this problem face is to model the data in an unbiased and exhaustive way that takes into account all the contributions of each of the players towards the results that a team achieves. Most of the conventional statistics are highly offensively biased and favor players who score, shoot and assist rather than those who stymy the opponents game by harrying their players, intercepting passes and blocking their shots [Image2].

We also take a look at a revolutionary new metric called the Player Efficiency Rating(PER) that has taken over the sports analytics world over the last 8 years and try to combine that with a whole battery of methods that take into account the position a player plays in and what should be considered a good or bad performance for a player in that position.

Another important challenge we tackle is to model how the synergies between players affect team performance and how the addition of a third player can not only affect the abilities and effectiveness of the first two players but also drastically change the synergy between them. Armed with these completely new ways to look at player and team level statistics we attempt to gauge the chances a team has to win a game against any other team given their player abilities, the synergies between them, the aptitudes of their respective coaches, the form of the two teams coming into the match and the history these teams have against each other.

Related Work

The problem of predicting Major League games has been well studied and several quirky and unconventional solutions and approaches have been proposed that have at times been so successful that they have subsequently gained mainstream acceptance. Fearnhead and Taylor's (2010) method of adjusted plus minus scores has a pertinent point when it seeks to asses a player's defensive contribution for the time has plays on court. However, the model does not make any room for the existence of synergies between players and is hence not nuanced enough. Ryan Parker's (2010) method of applying logistic regression on points per possession statistic also normalizes the statistic plating time that inflates the difference between starters and reserves. However, there is no inclination to model defense and its effect on results, David Berri's (1999) attempt to find the League Most Valuable Player led him to regress to a formula that we used as an anchor point to verify our results from time to time. Alan Maymin (2012) introduced the multi agent system concept of synergy graphs into basketball and introduced the valuable concept of positive and negative synergies between players that could be modelled from both an offensive and defensive standpoint. We have adopted this technique in a modified way in our eventual implementation. Also we have used this directly in the final team level predictions as a valid proxy for team chemistry. Savic and Rodovjiic's (2014) method of applying data envelopment analysis shows a lot of promise but starts to fail when the two teams have equal strength and the prediction can be a tossup. Finally, Alex Block's (2014) thesis gave us useful markers when it came to modelling defense as a combination of individual statistics.

All the aforementioned methods have their own positives and negatives and we have tried to accommodate as many of their plus points as possible along with introducing0 changes of our own that range from minor tweaks to wholesale sweeping changes in approach. In the subsequent sections we outline the structure of our data and go on to show the details of our implementation pipeline explaining their advantages and drawbacks and eventually suggest ways to improve results further.

The Data

We have modelled the problem so as to attack the problem at two levels. Firstly, we evaluate individual player abilities and then armed with the player level information we then approach the problem at a team level.

For each player in question in a certain fixture we collect his overall game stats (both offensive and defensive) and his advanced stats along with his stats in the playoff separately. The playoff stats if applicable are collected separately so as to account for the fact that a player in a team that makes it through the league stages is likely to be better than a player who doesn't and his performance in these knockout games is indicative of his performance under pressure.

At the team level we have used the box score data of the matches that a team has played in the recent past. The box score sows team totals when playing against a certain team and are valuable in the team level modelling.

In the subsequent sections we go on to show how the data was processed to make it fit for learning and the various stages of the learning pipeline.

Gauging Similarity

To efficiently find out a player's value, we need to first determine what defines the value of a certain player. For that however, we first need to figure out up to 10 most similar players to the player in question, across the NBA. The role of similarity scores in the determination of a player's value, will be explained in the next section.

Thus to go about finding the set of most similar players, we first need to develop a measure of similarity, that can be applied between any two players. Here we take a leaf out of the incredibly popular Sabermetrics book. The Sabermetrics system used in baseball analysis in Major League Baseball aims to evaluate a player only by the skills that are retrospectively deemed most important for a player of his age across the NBA from all eras, moving on to the next year.

Hence, for example to assess the value of a 26-year-old point guard in a certain team, the evaluation of his abilities will only be done by the retrospective evaluation of the career paths of all 26-year-old point guards in the history of the NBA across all teams, through the next year [Image4].

Also important, is to make sure that the players deemed to be similar also have similar career trajectories. Hence to evaluate similarity we add up the win shares for each season and evaluate the age distribution of the player's best seasons up to the age of the player in question. Ensuring a similar distribution of performance up to a certain age and similar statistics, allows us to claim that the aforementioned techniques are a reliable substitute foe similarity.

Assessing Player Value

In order to avoid the mistake of naively judging a player by the means of conventional statistics that are heavily biased toward offensive players with low accuracy who make a lot of attempts at the basket and score some, we need to first answer the very important question – By what set of statistics and by what aspects of the game should a player who plays in a certain position be judged by in order to evaluate his effectiveness? In order to find out the set of pertinent statistics with respect to a player we use the career curves of the players most similar to him across the league as was shown in the last section.

Our next challenge was to select a suitable dependent variable for regression from the commonly available. After investigating several candidate attributes like offensive win shares, defensive win shares and adjusted plus minus ratings, we finally settled on Player Efficiency Rating (PER). This statistical measure developed by ESPN normalizes the player's ratings with respect to their positions and also accounts for a per-minute measure thus not being biased to fast paced teams who have more attempts on basket and have more turnovers [Image1].

We then proceed to regress using least squares regression with feed forward selection that greedily goes on to select the attribute that combines best with the already selected set of attributes to provide the best fit to a certain pre-selected dependent variable. However, certain attributes tend to have larger values compared to others and hence will dominate. To prevent this, we normalize the values by subtracting the average value of the same statistic for all players of the opposition. This also has the highly desirable side effect of only considering the abilities of arch player with respect to the opposition in question.

This set of most important attributes for a player is determined by running the regression technique on each of the aforementioned set of most similar players. Then we proceed to take the union of the sets obtained from each of the players. This union of attributes effectively forms the attributes by which we judge the player in question. The coefficients of the regression are used to obtain a value that can be considered a reliable proxy for the effectiveness of a player [Image3].

Synergy

With the player level modelling covered we move on to meet our next challenge – modelling the team level synergies between pairs of players. For this we referred to the skills plus-minus method that allows us to effectively simulate the events of the game and figure out the probable outcome if any two players were to play together as compared to if only one of them was to play [Image5].

In the skills plus-minus method we view the game as a series of successful shots or turnovers of possessions. The six events on which the simulation is run are Basket Score, Block, Steal, Out of Bounds, Offensive The probability of each of these events occurring is evaluated from the statistics of each of the 10 players on court at the time. We first run the simulation with the two players in question and replace the statistics of the remaining eight with the average of the team. Then we run the simulation with each of the two players individually and replace the remaining nine players with the average of the team, Now the points difference is calculated for each of the three simulations and the synergy is calculated by comparing the result of the first one with the sum of the second and third. If the first simulation has a larger value than the sum of the other two, then the synergy is positive else the synergy is negative. Again here the synergies are not found out between two players at one go but it is found skill by skill. This is done by maintaining the actual values for only the skill in question for the player or players involved and replacing all the others by replacements level values i.e. the lowest values for that skill in the team roster.

The above plus-minus simulation is run several hundred times between various randomly selected pairs of players. Then out of the five skills the offensive ones, namely offensive rebounds and field basket are regressed against the sum of the offensive win shares of the two players in question Similarly the defensive skills, namely steals, blocks, and defensive rebounds are regressed against the sum of the defensive win shares for the two players in question. In both the cases ordinary least squares method is used. This leaves us the coefficients for the various contributing factors to synergy that can be used to evaluate the most plausible synergy values between any two randomly selected players.

However, the above method still leaves something to be desired. This is because it views a team of n players as having n(n-1)/2 pairwise synergies and completely discounts the ability of a third player to alter the synergy between the first two players along with the ability to affect the effectiveness of the first two players (which is already modelled by synergy).

To solve this challenge, we take a leaf from the field of multi agent systems by using the concept of synergy graphs. We apply an inverse transform to the synergy values so that all synergy values are positive and higher positive synergy values have the smallest values and the higher negative synergy values have the largest values. Then we proceed to construct a synergy graph with the players as the nodes and the synergy between any two of them being the weight of the edge connecting the nodes. Now, using Dijkstra's Algorithm we update the synergy values between players with the length of the shortest path between them. After updating all the synergy values we then reverse the previously applied inverse transform to obtain the updated synergy values. This allows for situations such as when two players may not gel well together but combine well with a third and hence the three of them play effectively when together.

Team Level Predictions

Armed with the individual player effectiveness values and the team wide synergy values we seek to solve the final problem of predicting the winner given the above mentioned data. Also thrown into this information mix are several pertinent team level statistics such as the box score tables of the previous five matches of each team (to take form into account), the box score records of the past five head-to-heads between the two teams and the coach statistics for both the teams, such as win percentages vs a certain team.

Using this large feature vector for each match (~600 attributes) we then use two classifier methods to arrive to the predicted result and subsequently verify them. After speculating on and testing out various parameter values we came to the conclusion that a multi-layered back propagation hyperbolic tangent neural network works best. After tinkering around with the parameters we came to the conclusion that a double hidden layered setup with each layer having roughly 1/10th the number of neurons as the original input feature vector. Also a learning rate of 0.1 and momentum of 0 were deemed suitable for our purposes [Image6].

We also used an SVM classifier to back up our results. Several types of kernels were tinkered with. Stated below are the recorded accuracies of prediction using the various kernels on a relatively limited subset of the available data.

Kernel Used	Recorded Accuracy
Linear	61%-68%
Sigmoid	52%-71%
Polynomial (Degree 2)	58%-74%
Polynomial (Degree 3)	60%-76%
RBF	55%-66%

Results and Conclusion

The neural network classifier's prediction accuracy varies from 64% to 72% for various runs on the dataset. However, due to the enormous bank of data present about, players, coaches and teams and the difficulty in collecting and processing it we are more or less constrained to this accuracy as we could only manage to train our pipeline on a sixth of the league (the Atlantic Division of the Eastern Conference, that is riddled with offensive minded teams). A more concerted and a larger scale effort could certainly push these boundaries further.

Also, in order to focus on the final classifier, important parts of the pipeline were handicapped in sophistication. For starters, the player value determination could be done using a more nuanced regression technique such as ridge regression and such methods could open up a whole new world of possibilities for improvement.

Images

Image 1: Strong positive correlation between PER and Team Win Ratio

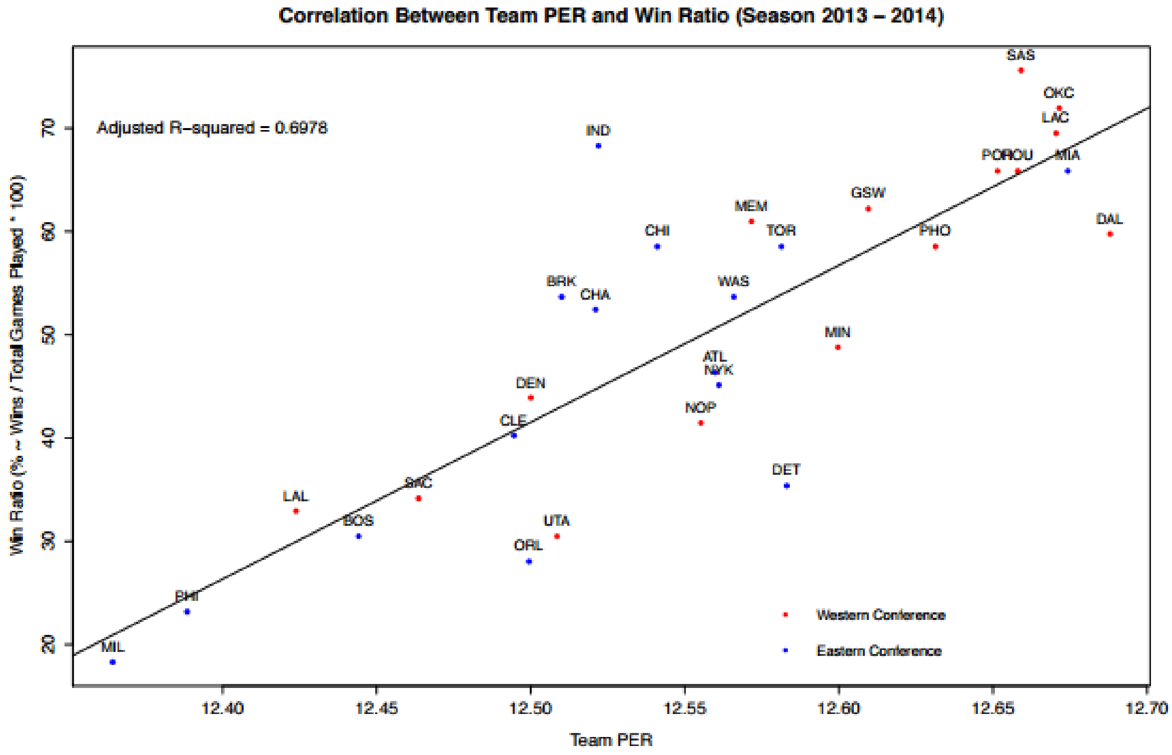


Image 2: Kernel Density Estimation for Offensive Win Shares of Atlanta Hawks 2011-1012

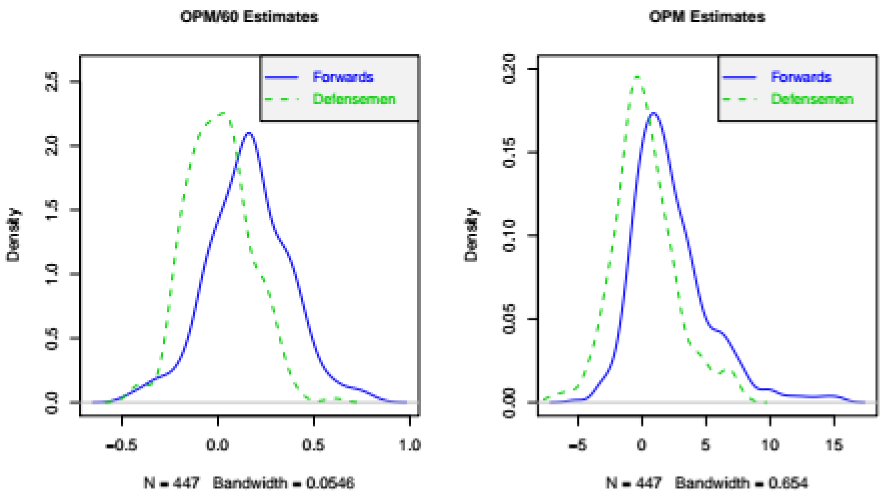
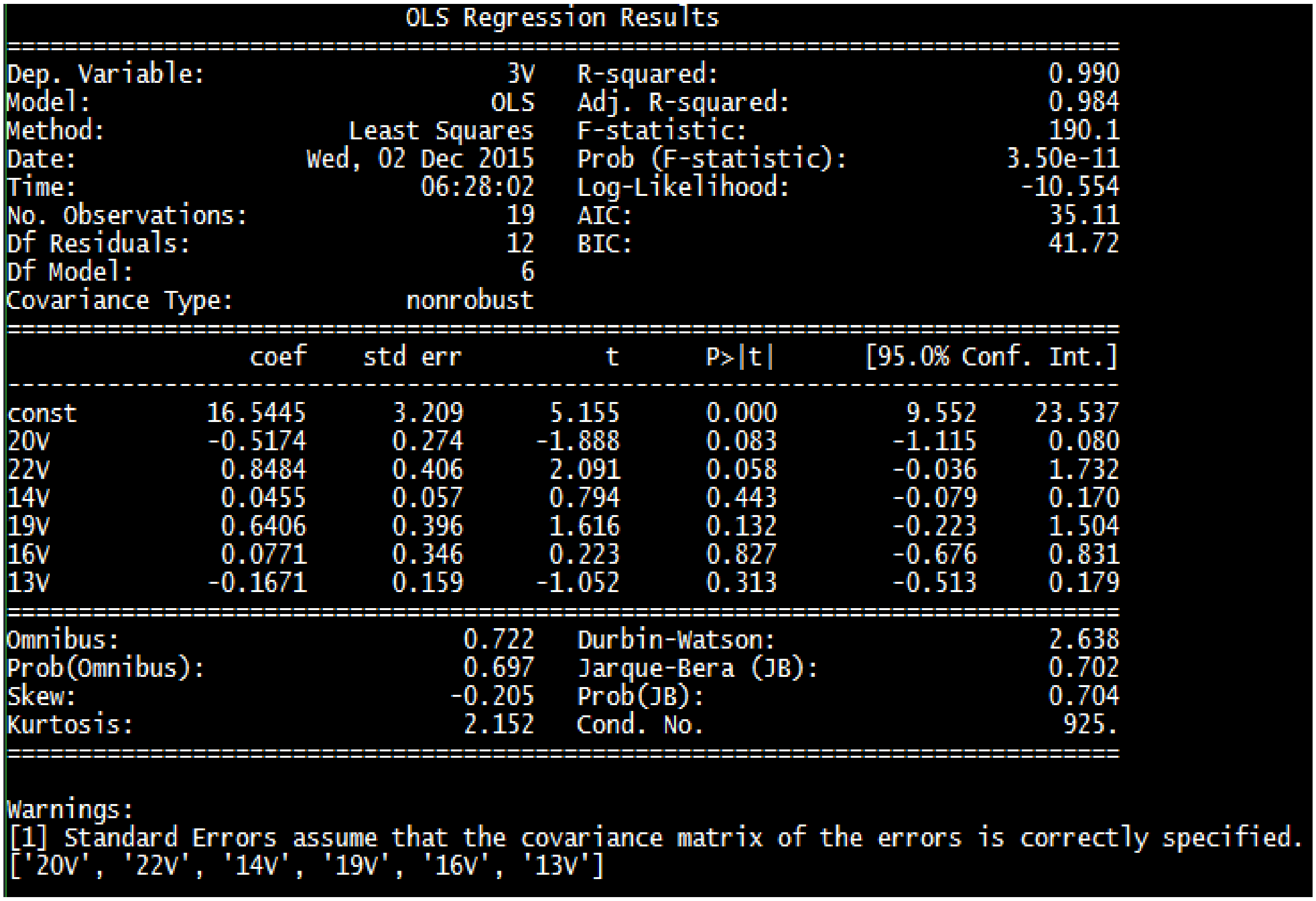


Image 3: Results of regression to find the set of important attributes for LeBron James



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     3V   R-squared:                       0.990
Model:                            OLS   Adj. R-squared:                  0.984
Method:                 Least Squares   F-statistic:                     190.1
Date:                Wed, 02 Dec 2015   Prob (F-statistic):           3.50e-11
Time:                        06:28:02   Log-Likelihood:                -10.554
No. Observations:                  19   AIC:                             35.11
Df Residuals:                      12   BIC:                             41.72
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         16.5445      3.209      5.155      0.000         9.552    23.537
20V           -0.5174      0.274     -1.888      0.083        -1.115     0.080
22V            0.8484      0.406      2.091      0.058        -0.036     1.732
14V            0.0455      0.057      0.794      0.443        -0.079     0.170
19V            0.6406      0.396      1.616      0.132        -0.223     1.504
16V            0.0771      0.346      0.223      0.827        -0.676     0.831
13V           -0.1671      0.159     -1.052      0.313        -0.513     0.179
==============================================================================
Omnibus:                        0.722   Durbin-Watson:                   2.638
Prob(Omnibus):                  0.697   Jarque-Bera (JB):                0.702
Skew:                          -0.205   Prob(JB):                        0.704
Kurtosis:                       2.152   Cond. No.                         925.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
['20V', '22V', '14V', '19V', '16V', '13V']
```

Image 4: Similarity scores for players most similar to Kobe Bryant

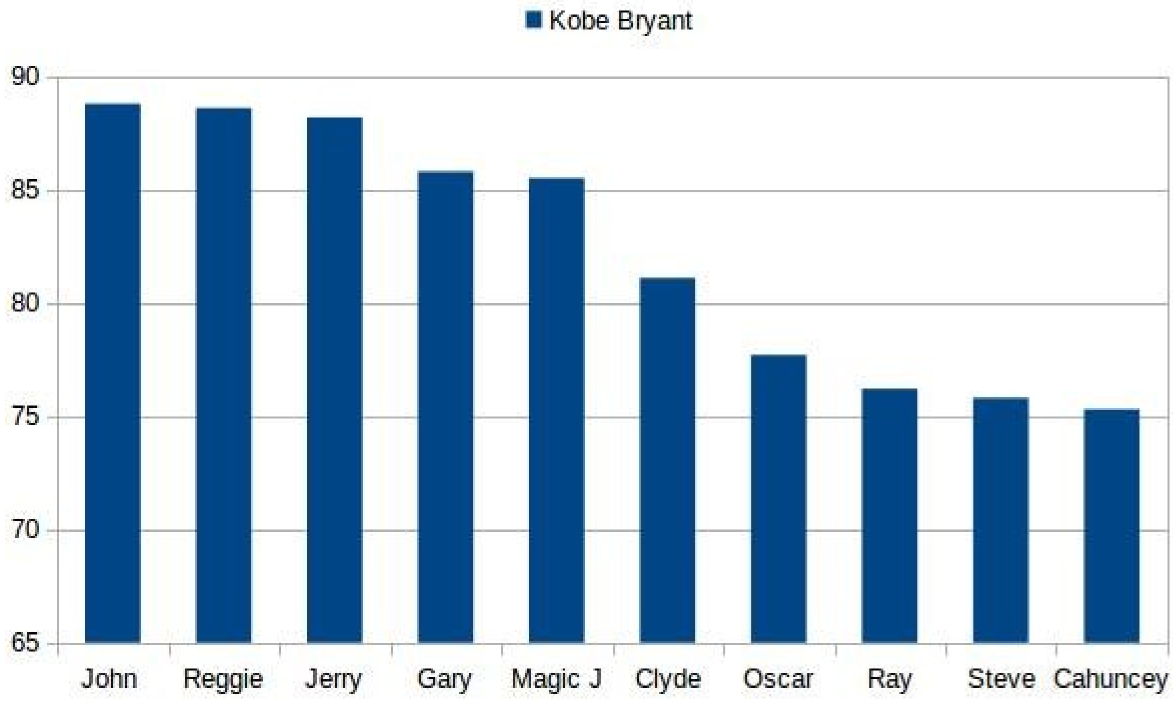


Image 5: Kobe Bryant Offensive Plus-Minus 2010-2011

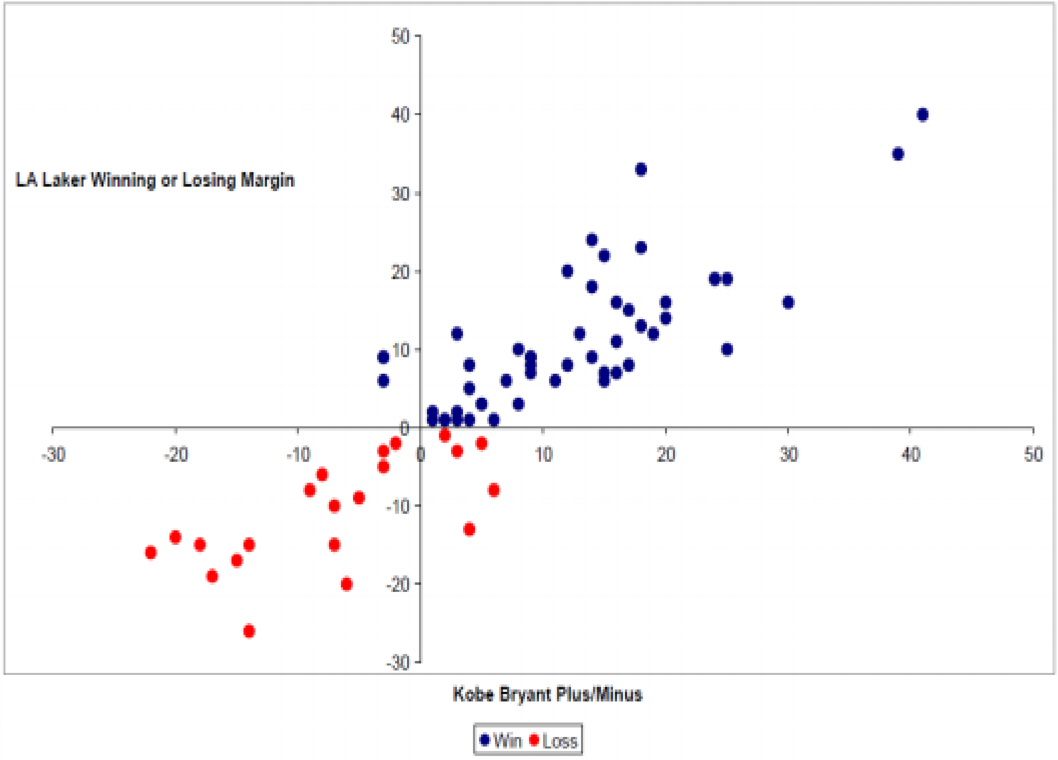
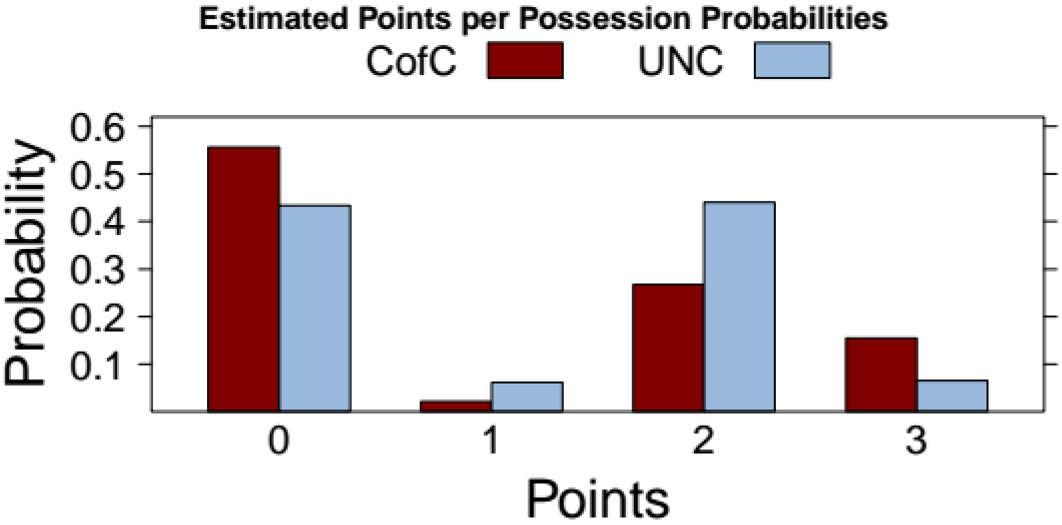


Image 6: Estimated probabilities of the home team scoring 0,1,2 or 3 points in an amateur game


Estimated Points per Possession Probabilities
CofC
UNC
Probability
Points