Basketball: Regression Model To Correlate Performance With Salary

Mihir Sathe
University of Southern California, Computer Science Dept.
Los Angeles, CA 90089
msathe@usc.edu

Abstract National Basketball Association (NBA) is the premium men’s professional basketball league in the world. NBA players are world’s best-paid sportsmen by average annual salary. Since the on-field performance is naturally the most important factor in order to decide the salary, a performance measure is necessary that spans over multiple aspects of a player’s game and can be scaled to the level of the actual salary. Through integration and analysis of a wide range of data available from multiple sources throughout the World Wide Web, this project intends to find a correlation between various performance measures and the salaries and use it to define a model to assign salaries based on performance.

Keywords. Information Integration, Basketball, Regression

1 Introduction

Over the time, many performance measures have been created to assess the performance of basketball players. Some of them are direct observations (such as Number of Goals) while others are derived statistics (such as Win Shares). Unfortunately, none of these measures is a ‘wide spectrum’ one that can inform about the overall performance of the player in one particular season (or over multiple seasons).

One obvious thought would be to consider the salary that a player gets as his overall performance measure. Up to certain extent this notion makes sense since in an ideal world, teams do pay for what performance they get from the player. Although in the practice, there are many more factors other than performance that quite majorly affect the salary. Popularity of the player, success (or failure) in previous seasons and need and resources of the offering team are some of the important factors.

Another major factor specific to NBA is the multi-season contracts. A contract signed over multiple seasons assures a certain sum of salary to the player largely irrespective of his performance during this period. This makes finding the correlation between season performance-season salaries almost meaningless. Also, it is really hard to find the exact and accurate information of the contracts over a large historical span. Together, all these factors make the problem of finding the performance-salary correlation quite complex.

This paper tries to solve two problems 1) Using the salary data of a player to predict the possible contracts using clustering 2) Use this pseudo-contracts to fit a sensible model between performance over a contract vs. salary of the next contract.
2.1 Data

All the data required for this project is not available from any single source. So multiple sources (web sites) are used. The data was collected by screen scraping web sites using C# and open source library HTML Agility pack[1]. The collected data was then interlinked with record linkage using Fine Grained Record Integration and Linkage (FRIL)[2]. The collected and connected data was then stored in RDBMS (PostgreSQL) database and views were created as and when required to make it easy to query the data. All the tools used were open source (except for .NET framework itself).

2.2 Contract Predictions

This is a very interesting problem considering the variety of complications and details involved in the NBA contracts. I will start with a simple case. Table 1 states the salaries of an NBA player Brendan Haywood[3].

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
<td>$4,000,000</td>
<td>$4,500,000</td>
<td>$5,000,000</td>
<td>$5,500,000</td>
<td>$6,000,000</td>
<td>$6,900,000</td>
<td>$7,624,000</td>
<td>$1,886,312</td>
</tr>
</tbody>
</table>

The pattern in this case is quite visible. Clearly 2005-2009 follows the same pattern of increments (Which happens to be with Wizards). Followed by that 2010-2011 are together (with Mavericks). 2012 is again an obviously different and significantly lower valued contract with Bobcats. This type of contracts can be easily predicted with a linear time algorithm. But things can get more complicated than this. Table 2 shows the partial salary information of Chris Webber.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>Kings</td>
<td>76ers</td>
<td>76ers</td>
<td>Pistons</td>
<td>76ers</td>
<td>Warriors</td>
</tr>
<tr>
<td>Salary</td>
<td>$17.53M</td>
<td>$19.12M</td>
<td>$17.6M</td>
<td>$.66M</td>
<td>$19M</td>
<td>$.6M</td>
</tr>
</tbody>
</table>

This looks very unusual and complicated as compared to the one before. As a matter of fact, 76ers waived him in 2006 (but of course they had to pay him till the official end of contract). In 2006, he played as a free agent for Pistons and in 2007, he played as a free agent for Warriors. The simple linear approach cannot help in finding
this type of ‘hidden’ information. So instead, we can use the approach of hierarchical clustering.

**Hierarchical Clustering**

Hierarchical clustering is K-Neutral in the sense it automatically finds as many number of clusters as it can in the given data. This property makes it a very powerful tool for processing the unknown patterns \(^6\). This approach is very useful for our problem since we cannot know the number of contracts beforehand. We initially start with all the points as their own clusters. Then we start decreasing the threshold of similarity to make fewer clusters. We use the average of all the containing elements to find the ‘salary’ value over an entire cluster.

Through empirical information about the contract increments, 88% similarity (or 12% difference) threshold seems to be working great for most of the cases. Although this technique can capture the nested deals as separate contracts, one big issue with this technique is that it only looks for similar numbers (irrespective of their chronological order).

**De-clustering**

This requires us to follow the clustering with another process, which I will refer to as ‘de-clustering’. This is a simple linear process of visiting every single cluster in a series of salaries and breaking it down into multiple smaller clusters, which are chronologically contiguous.

For example, consider following hypothetical input:

<table>
<thead>
<tr>
<th>Year</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>$1.3M</td>
</tr>
<tr>
<td>2003</td>
<td>$1.4M</td>
</tr>
<tr>
<td>2004</td>
<td>$5.2M</td>
</tr>
<tr>
<td>2005</td>
<td>$5.4M</td>
</tr>
<tr>
<td>2006</td>
<td>$1.45M</td>
</tr>
<tr>
<td>2007</td>
<td>$1.5M</td>
</tr>
</tbody>
</table>

Since clustering is unaware of the temporal parameters, the clusters that we will get are:

2. 2004, 2005

So the de-clustering algorithm will go through the first cluster and break it down into two chronologically contiguous clusters as follows:

2. 2006, 2007
3. 2004, 2005
Once both clustering and de-clustering is done, we are ready to go to the next step of finding a regression model that fits with this data. With this process, the data from Table 1 can be represented as:

Where each dot being a salary point and the ellipses indicating the detected clusters using the algorithm specified before.

### 2.2 Fitting a model

Now that our data is processed and we have the pseudo contract information ready, we will try to fit a regression model over the data. In the normal situation, upon the end of a contract term, a player gets his contract renewed (with same or different team). So we will compare the contract amount of a contract with the performance in the previous contract. It has been observed that the most important performance for the consideration for the contract renewal is the last season played before the previous contract ended.

So if we have contracts as in Table 4, we will use performance in 2006 and average salary from 2007, 2008 and 2009 as parameters for our training data. Clearly, the associated layer did not perform well during earlier contract; so he got a worse contract.

<table>
<thead>
<tr>
<th>Year</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>$10M</td>
<td>$5M</td>
<td>$5.2M</td>
<td>$5.3M</td>
</tr>
<tr>
<td>2006</td>
<td>$10.3M</td>
<td>$5.2M</td>
<td>$5.3M</td>
<td></td>
</tr>
</tbody>
</table>

To measure the performance, we will use the Advanced Statistics from BasketballReference.com (See appendix 1), which are essentially the derived statistics that assess particular aspects of a player’s performance over a season such as Rebounds, Win Sahres, Steals, Offensive rating, Usage etc.
For the experiment, I used two different time spans
1. 2001-2005
2. 2006-2010

We are only considering the contracts that ‘completely’ lie inside the said timeframe. Reason for using such short time windows is that we are trying to project performance measures over the salary. Equivalent performance in 1990 and 2013 would produce equivalent performance measures but it won’t produce the same salary (because of the inflation and other environmental factors like change in the cap amount). Keeping the experiment constrained within a small timeframe reduces this problem.

2.4 Results

<table>
<thead>
<tr>
<th>Year Range</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2005</td>
<td>.6886</td>
</tr>
<tr>
<td>2006-2010</td>
<td>.6574</td>
</tr>
</tbody>
</table>

Model

\[
\text{SALARY} = -544772.0397 \times \text{DEFENSIVE\_WINSHARES} + \\
-530268.0069 \times \text{OFFENSIVE\_WINSHARES} + \\
853711.618 \times \text{WINSHARES} + \\
12734909.4147 \times \text{WINSHARES\_48} + \\
62064.0623 \times \text{AGE} + \\
351713.8469 \times \text{BLOCK\_PERCENTAGE} + \\
-242227.9657 \times \text{DEFENSIVE\_REBOUND\_PERCENTAGE} + \\
-68354.5166 \times \text{GAMES\_PLAYED} + \\
2920.4088 \times \text{MINUTES\_PLAYED} + \\
-147009.7588 \times \text{OFFENSIVE\_REBOUND} + \\
-57933.3654 \times \text{OFFENSIVE\_RATING} + \\
-313272.5444 \times \text{STEAL} + \\
426503.3052 \times \text{TOTAL\_REBOUND\_PERCENTAGE} + \\
134676.2495 \times \text{USAGE} + \\
3197526.4921
\]

(See appendix 1 for meaning of these terms)
Conclusion

The model correlates with the actual salaries with a correlation coefficient of .6886 in 2001-05 and with .6574 in 2006-07. As stated before, in practice, there are many more factors other than performance those affects the salary such as popularity of player, previous performance, need of the team etc. If those factors can be accounted for in this model, the correlation would be better.

But as our goal is to find a performance measure that can scale to the salary, this model serves the purpose very well. This model will answer the query “What is the best evaluation of this player based on last contract performance?” which is indeed a very valuable input for deciding a contract amount.

This project can be expanded by creating models for other factors such as popularity vs. salary, need vs. salary etc. Although data collection, preparation and determination of the correct parameters for this would be quite challenging and interesting. Once that is done, we can combine these models and tell the exact expected salary for the player.

Appendix 1

Explanation of some parameters for the training data

1. Win Shares
   An estimate of wins contributed by the player.

2. Defensive Win Shares
   An estimate of wins contributed by the player due to his defense.

3. Offensive Win Shares
   An estimate of wins contributed by the player due to his offense.

4. Block Percentage
   Block percentage is an estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor.

5. Defensive Rebound Percent
   Defensive rebound percentage is an estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor.

6. Offensive Rebound Percentage
   Offensive rebound percentage is an estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor.

7. Offensive Rating
For players, it is points produced per 100 possessions, while for teams it is points scored per 100 possessions.

More information about these and other parameters can be found on the glossary page \(^5\) of basketball reference.

References

1. HTML Agility Pack http://htmlagilitypack.codeplex.com/
2. FRIL. http://fril.sourceforge.net/