Model for Determining Entertainment Value Index

Hack-a-Shaqers
Nick Ceneviva, Ishaan Rao, Neil Sharma, Arturo Torres

September 24, 2017

Table of Contents

1 Objective

Currently, the NBA All-Star Weekend 2017 is taking place. After the All-Star break there remains 460 games in the regular season, and the league is trying to determine the optimal method of allocating their promotional spending budget. To help with this issue, the league is looking to understand the entertainment value of each regular season game.

The objective of our project is to build a model to predict the entertainment value of a regular season game after the All-Star break. We have defined "entertainment value" to be the engagement of the fanbase over a certain regular season game. This metric includes three main components:

1. Game Attendance vs. Stadium Capacity
2. National and Regional TV Viewership of the Game
3. Web Traffic of the NBA Team's Website

These components will generate an Entertainment Value Index from 0 to 20, with 0 being a game without any entertainment value, and 20 being a game with the highest possible entertainment value. The league will be able to use this index to determine which games they should or should not allocate their marketing budget towards. Additionally, our model will include a measure of how confident we are in our prediction of the entertainment value.

2 Data

We used the following datasets to develop our model:

- Attendance and Capacity Data
- DMA Households Data
- Game Data
- Jersey Sales Data
- National and Regional TV Ratings Data
- Player Data
- Web Metrics Data
- NBA Team Possessions per Game [3]
- ELO Ratings [1][2]
- NBA Rivalries [4]

The Attendance and Capacity dataset provides us with training data to predict the attendance at future games for each team. We used the DMA Households dataset to determine the size of regional markets, so the regional TV viewership could be scaled as a percentage of households.

The Jersey Sales Data provided us insight into the most popular players. This is a key component of TV Viewership, because fans tend to watch games that feature popular players. We used the National and Regional TV Ratings to provide the training data to predict the viewership for games after the All-Star break. The Web Metrics dataset was also used to train the model to predict the web traffic for games after the All-Star break. We used the UNQ value on the day of a game to the teams' websites to determine web traffic for a matchup.

The NBA Team Possessions per Game was used to determine the pace of play for a specific game. We used this, because faster paced games tend to be more entertaining.

We used the Game Data set to determine the dates of every NBA game, allowing us to determine whether games occurred on weekdays, weekends, or holidays. We used this set to find if teams had met in the playoffs in a previous year as well. The Game Data set also provided the results of every NBA game, which is essential to calculating ELO.

Lastly, we used the values from Five Thirty Eight's NBA ELO Ratings to generate our own ELO values for every team after every game. Our process for calculating this is outlined in the methods section below. ELO is used as a measure of the strength of a team, and it originally was a measure used for chess ranking.

For comprehensibility, we have outlined the features used in our model and what dataset we fetched them from below.

Feature	Dataset	Definition
All-Stars	Player Data	Represents the number of all-stars involved in the game
ELO	ELO Ratings [1]	Represents the relative strength of an NBA team
Holiday	Game Data	Indicates whether the game took place on a holiday
Past Playoff Teams	Game Data	Indicates whether the teams met in the previous year's playoffs
Possessions per Game	Team Rankings [3]	Represents each team's average possessions per game over the season
Rivalry	Wikipedia [4]	Indicates whether the two teams have a historic rivalry
Top Jersey Sellers	Jersey Sales Ranking Data	Represents how many high jersey-selling players were involved in the game
Weekday	Game Data	Indicates whether the game took place on a weekday
Weekend	Game Data	Indicates whether the game took place on a weekend

Table 1: Features of the Model

3 Methods

Since we must predict the entertainment value of future games, we can only use data that would be available to us at the All-Star Break, which could be 1 week to 3 months before a prospective game. The features in the above table are all data points that would be available many weeks before a game, and therefore we will use these factors to develop our model.

3.1 ELO Ratings

Nate Silver of FiveThirtyEight.com developed a metric to evaluate the relative strength of an NBA team at any point in the season. [2]. This ELO rating is based off of chess ratings. We used this rating to evaluate the strength of matchups, because when two teams with higher ELO's play each other, the entertainment value tends to be higher. Since the dataset online only included ELO values until the 2014-2015 season, we calculated the values for the other seasons ourselves. We did this using the following formula:

$$ {}_nR_i =_o R_i + k \cdot MoV \cdot (S_{ij} - \mu_{ij}) \tag{1} $$

where:

${}_nR_i$ is the new ELO rating of some team, i

${}_oR_i$ is the old ELO rating of some team, i

k is a scaling coefficient, determined to be 20 for the NBA by Nate Silver [2]

MoV is the margin of victory multiplier

S_{ij} is the actual result of the match

μ_{ij} is the expected result of the match

S_{ij} evaluates to 1 if team, i, wins the game and evaluates to 0 if they lose. The formulas for MoV and μ_{ij} are found as follows:

$$MoV = \frac{(Pd + 3)^{0.8}}{7.5 + 0.006 \cdot ({}_oR_i -_o R_j + HCA)} \tag{2}$$

$$\mu_{ij} = \frac{1}{1 + 10^{\frac{{}_oR_i - {}_oR_j}{400}}} \tag{3}$$

where:

${}_oR_i$ is the old ELO rating of some team, i

${}_oR_j$ is the old ELO rating of some opponent, j

Pd is the point differential in the final score of the game

HCA evaluates to 100 if the team, i, has home court advantage, and 0 if they do not

3.2 Attendance and Capacity

We decided that the percentage of a stadium's capacity that is filled is a better predictor of a game's entertainment value, rather than the raw number of people attending. Additionally, we assumed that half of a team's tickets were sold to non-individual plans. Therefore we calculated the percentage of individual tickets sold like so:

$$\%ofIndividualTicketsSold = \frac{attendance - \frac{capacity}{2}}{\frac{capacity}{2}} \cdot 100 \tag{4}$$

3.3 Linear Regressions

In order to predict Game Attendance, TV Viewership, and Web Traffic, we had to determine which features would be most impactful in our regression model. For game attendance, we found that information regarding the home team, such as their number of All Stars and the number of top jersey sellers on their team, were the largest determinants in predicting game attendance. For TV viewership, however, we found that the information regarding the overall level of play between the two teams, such as the total number of All Stars, the total number of top jersey sellers, and the total pace of the play, were crucial in our prediction model. We were unable to find any features which correlated heavily with web traffic.

4 Results

4.1 Game Attendance

We used the following features as predictors for game attendance:

- Home All-Stars
- Weekend
- Home Top Jersey Sellers
- Away Top Jersey Sellers
- Rivalry
- Past Playoff Teams
- Home Team's Possessions per Game
- Away Team's Possessions per Game
- Home Team's ELO
- Away Team's ELO

We modeled the distribution of the percentage of individual tickets sold and came up with this figure.

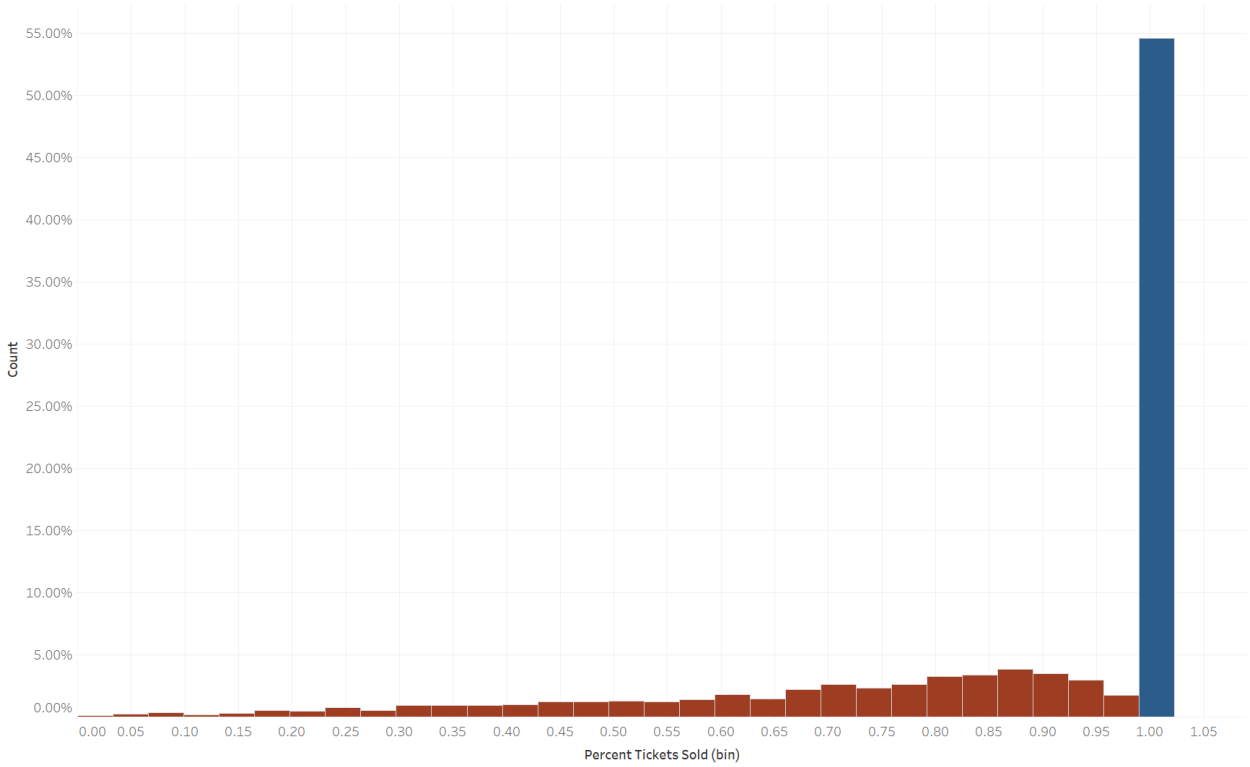

Figure 1: Percent Individual Tickets Sold

As seen above, the large majority of NBA games sold out. So, changing the inputs of our model did not have much impact on the percentage of tickets sold. We were more interested in what situations would result in the distribution highlighted in red in the above figure. Therefore, we changed our metric to predict the probability that a game would sell out, rather than predicting what percent of individual tickets would be sold. Our model had a 81 % accuracy in predicting whether or not a game would sell out.

4.2 National and Regional TV Viewership

We used the following features as predictors for national and regional TV viewership:

- Total All-Stars
- ELO
- Rivalry
- Top Jersey Sellers
- Possessions per Game
- Past Playoff Teams
- TV Network
- Total Households in the Region
- Holiday

We used these as predictors, because fans tend to watch games on TV that have something at stake, feature good teams, feature popular players, and are fast paced. Using this model, we developed a linear regression with the above features as independent variables, and the National and Regional TV Viewership as the dependent variable. We fitted a linear regression with an R^2 value of 0.79. The figure below displays a visualization of our regression.

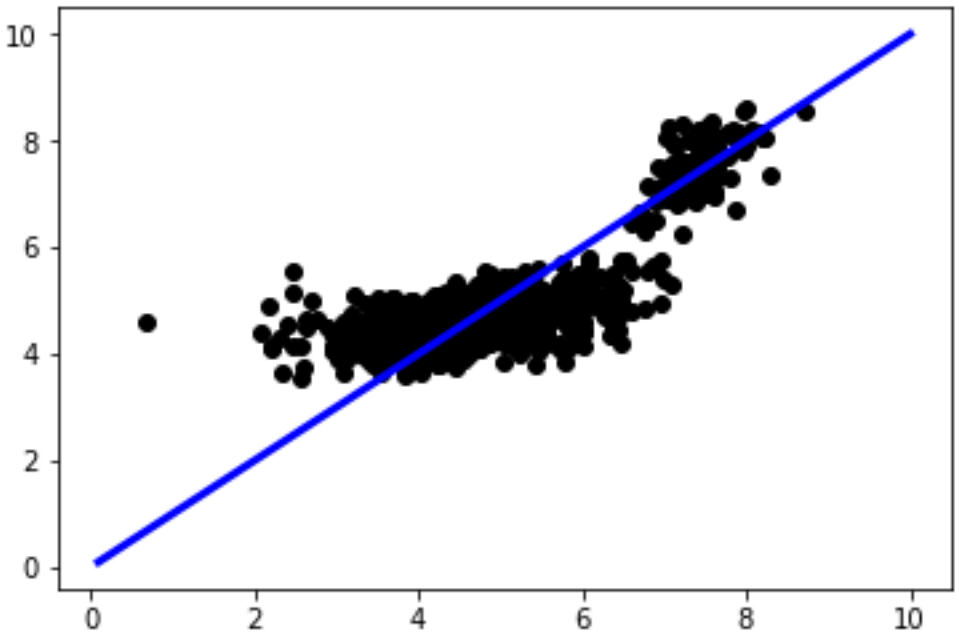

Figure 2: National and Regional TV Viewership

4.3 Web Traffic

Originally, we aimed to predict web traffic using certain features, such as ELO Ratings, Total All-Stars, and Weekday/Weekend. However, we found that no combination of inputs resulted in any correlation with web traffic. We still believe that web traffic is a good metric to predict entertainment value, but we did not have the correct features to accurately predict web traffic. If we had more access to social media data or other web traffic, then we may have been able to implement this into our model. Because of this, we did not include web traffic in our final model.

4.4 Final Model

Our final model will be a linear combination of the Game Attendance (GA) and TV Viewership (TVV). We gave each of these aspects equal consideration in determining the Entertainment Value Index, so in the equation below, the value for $c1$ equals that of $c2$.

$$EVI = c1 \cdot GA + c2 \cdot TVV \tag{5}$$

We arbitrarily chose the weights to be 10, so $c1 = c2 = 10$, and therefore the final EVI is on a scale from 0 to 20, with 0 being a game without any entertainment value, and 20 being a game with the highest possible entertainment value.

5 Conclusions

Using our model for Entertainment Value Index (EVI), we calculated the 10 most entertaining and 10 least entertaining regular season games of 2017 after the All-Star Break. The following graphic illustrates these results.

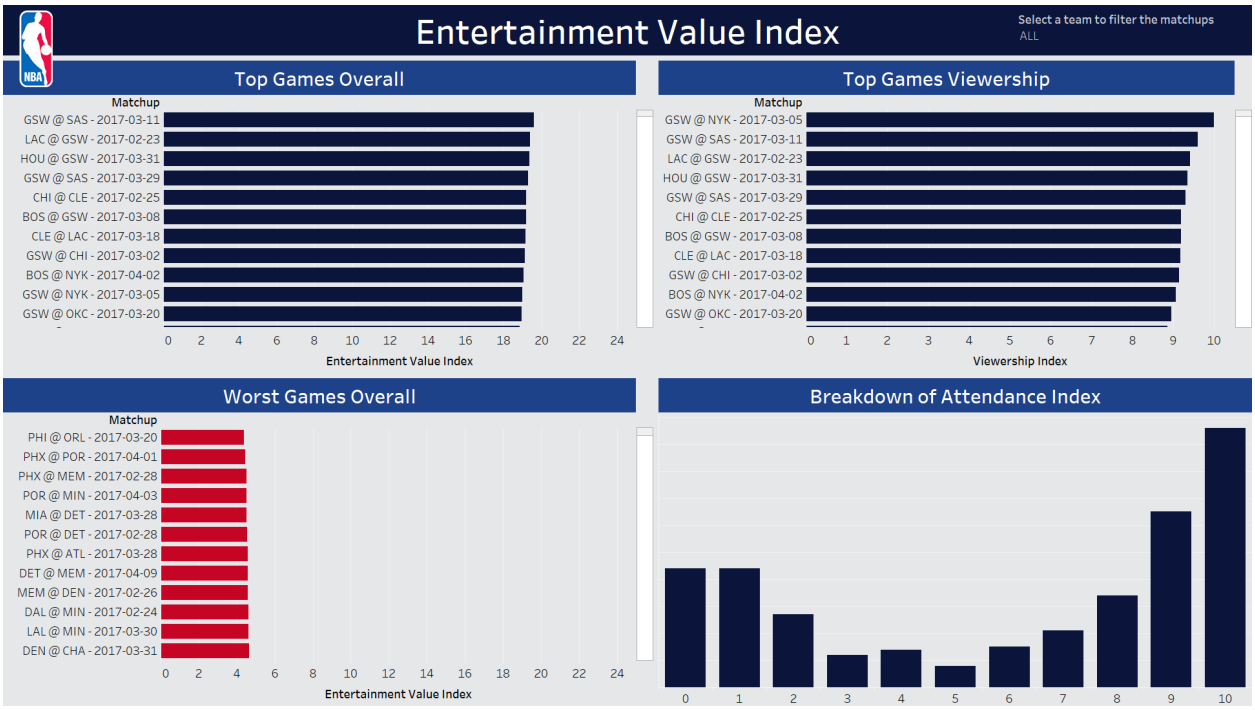

Figure 3: Model Results

The left side of the graphic above displays what we predicted to be the best and worst games overall, as measured by our Entertainment Value Index. The top right of the graphic displays what we predict to be the top games in just TV viewership. The bottom right of the graphic illustrates the distribution of the likelihood of a game selling out.

5.1 Confidence

We are fairly confident in our findings because all of the features incorporated in our model are known at the time of the All-Star Break, with the exception of the ELO rating of a given team before a game. We found that the standard error of ELO ratings over the whole course of the season was only 27 points, which is why we took ELO ratings to be constant at the point of the All-Star break. Therefore, our confidence in our model is in line with the results. Our R^2 value of 0.79 for the TV Viewership along with our 81% accuracy in predicting whether or not a game will sell out indicates a high level of confidence that our metric is an accurate representation of the Entertainment Value of a game.

References

[1] https://github.com/fivethirtyeight/data/tree/master/nba-elo

[2] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

[3] https://www.teamrankings.com/nba/stat/possessions-per-game

[4] https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_rivalries