Bracketology Report

Chris Campell, Chris Chalfant

Monday, March 14, 2016

This is an R Markdown document in. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

# Abstract

This paper will detail the methodology employed by our group when attempting to predict the winner of the ESPN NCAA 2016 basketball tournament via machine learning. Our approach is contrived from the Massey Method of Statistical Modeling [1]; which aims to accurately depict the correlation between a team’s rating and their obtained box scores. Through permutation testing we have found the statistical method proposed by Massey to have a 66% percentage rate at predicting the outcome of games correctly in the 2016 bracket. Our goal is to increase the efficiency in prediction of the linear model in order to surpass the accuracy of the Massey Method in determining match results correctly.

# Methodology

## Systems of Linear Equations and Matrix Vector Form

A Linear Equation takes the following form:

Where represents a feature and represents a weight applied to said feature.

The weight is unknown in the linear system, and it is our intent (during Linear Regression) to find all values of that minimize the generic linear equation:

We will define a system of linear equations that represent basketball games. We will then convert the linear system of equations into Matrix Vector Form. This conversion into matrices will allow us to employ Linear Regression to solve the complex system of linear equations for .

The Generic Matrix Vector form of a Linear System of Equations is defined by the following matrices:

Where matrix is comprised of the features of the linear system:

Matrix is comprised of the weights of the linear system:

And matrix is comprised of the targets of the linear system:

Combined, the matrix vector form of the system of linear equations appears as such:

The Massey Method of Statistical Modeling applies the approach detailed above to represent the ESPN NCAA tournament in Matrix Vector form as a system of linear equations.

## Applying the Massey Method

### Defining the Linear Model

The linear equation of the Massey Model is defined as such: . Where are the ratings of teams and respectively; and are the scores of team and . The general problem is as such: given a series of games with pre-recorded outcomes; can we assign an accurate rating to each team? The contrived rating should accurately represent the strength of the designated team; in that any following matchup between teams should be predictable by a comparison of the team's ratings alone.

As previously discussed; the generic equation for a linear system can be defined as: . The goal is to fit the equation of the Massey model to the generic linear form. The features of the linear system through in this case are the teams themselves. The weights of the linear system through are the ratings of the teams. The varaible is the sum of the ratings of the teams. For any given match; we know the teams (A and B), and we know the score differential , what we don't know is the ranking differential . We employed the method of Least Squares Linear Regression to solve for for each team. In doing so the computer minimized the equation giving us an accurate rating for team A. In order to do this; we must represent the Massey method’s linear system of equations in matrix vector form. The following sections describe the translation of the linear equation into matrix vector form.

### Defining the Feature Matrix

The Feature matrix in the context of the Massey Method represents the games played during a given year of the NCAA tournament. Each row in the "feature" matrix represents a single game (team A vs B). Each column in the matrix represents an NCAA team. Hence, matrix can be visualized as such:

In the Massey Method, each is represented by a linear equation. The default Massey Method expresses a row/game in the feature matrix using only the constants . This approach A in a given cell indicates that the game was played as an away game for that team/column. Similarly, a in a given cell indicates that the game was played as a home game for that team/column. A in a cell for a particular row indicates that the associated team (column) did not play in that game. Using this notation we can expect the matrix to resemble something like this:

The "features"" of the linear system are now encoded in matrix vector form in matrix . In the following section we will encode the "weights" of the linear system in matrix , and encode the "target" of the linear system in matrix .

### Defining the Weights Matrix

The "weights" matrix represent the weights in the generic linear equation . The weights matrix in the context of the Massey Method represents the vectorized lefthand side of the linear equation .

In the context of the Massey linear equation:

### Defining the Target Matrix

The "target" matrix represents the targets of the generic linear equation: .The target matrix in the context of the Massey Method represents the vectorized righthand side of the linear equation .

In the context of the Massey Linear equation:

### The Final Linear Model

Having now defined all the necessary matrices, we can now combine them in order to represent the Massey Model's linear equation in Matrix Vector Form. Combined, the generalized matrix vector form of the system of linear equations appears as such:

We can now see that the Massey Method linear equation in matrix vector form is as follows:

Each row in the Feature matrix constitutes a game played in the NCAA tournament. Each column in matrix F represents a unique team in the NCAA tournament. Each entry in the Weight vector represents the difference in rating between the two competing teams. Each entry in the Target vector Y represents the point difference between the two competing teams.

## Matrix De-Singularization

The final matrix representation of the linear system is a square matrix but it is not transposable [2]. This becomes a problem when utilizing the Least Squares method of Linear Analysis. In order to solve this, we must de-singularize the matrices so that they are transposable. By adding a row of all 1’s to the Feature matrix , as well as adding 0 to the Weight matrix and Target matrix ; we ensure that the matrices’ determinants are not zero. This allows us to utilize least squares to solve the linear system.

## Training the Linear Model

Having constructed a linear model of the NCAA tournament, it is now necessary to apply the process of Logistic Regression to solve the linear system for the values of that minimize the Massey linear equation. In the first step of this process, we train the model using the Feature and Target matrices and . The Weight matrix remains "unknown" and we rely on the Least Squares method of Linear Regression to populate the matrix accordingly.

# Variations to the Linear Model

Using the methodology described above in conjunction with the Massey model equation: as a template; we began to apply alterations to increase the validity of our predictions. The first of which was employing the observable effect of home court advantage on a given teams score.

## Adding Home Court Advantage

Given the default Massey equation: we can make the small modification (detailed in the Lets Go! PowerPoint [3]) to add home court advantage to the linear system: . The variable represents another unknown in the linear system of equations; whose value is approximated through Linear Regression.

Recall the generic equation for a linear system: . Adding the variable to the linear system is equavelent to the modification: where the functor represents the weight .

In terms of the Matrix Vector Form of the linear system, the addition of variable simply requires the addition of another column vector to the Feature matrix . This is due to the fact that the matrix already represents ; the left-hand side of the linear model equation. The addition of the term requires an addition to every row of the matrix in the form of the column vector:

We impose the restriction that to ensure that home court advantage does not add to zero.

## A Possessive Efficiency Approach

A somewhat different approach detailed by Statistician Ken Pomeroy emphasizes possessive efficiency over score differential. This approach attempts to acknowledge the fact that some teams play more aggressively than others (and as a natural result will incur more possessions). The Ken Pomeroy method attempts to boost a defensive-based-team’s rating by incorporating scores per possession for each game.

The utilization of the Ken Pomeroy method requires a re-writing of the Massey method's linear model: . Most basketball games do not keep track of possessions as part of the box scores. As a result, it is necessary to statistically estimate the possessions of a team during a given match. Pomeroy defines an equation that accomplishes this as such: where is the number of total field goals (shots) attempted, is the number of offensive rebounds, is the number of free throws (foul shots) attempted, and is the number of turn overs; for a given team: [4]. Similarly, the equation holds true for the opposing team of a given match in that:

Pomeroy suggests taking the average of the possessions of each match in that:

This modification when applied to the original Massey Equation yields:

Where is defined as stated above. Just as before, in relation to the generic linear equation the equation described above is just another feature in the linear system, so that:

for any given game in the Linear System.

# Assessing Model Validity

### The Laugh Test:

### Cross Validation:

Involves constructing and then training a linear model. Mathematically Cross Validation is defined as:

We employed the Test Set method described by Moore [3] with a 1:5 split; 80% modeling data and 20% test data. For the 2016 NCAA season [6] (with the data available up until 3/25/2016), we found the following results.

# Results

We narrowed down the methods we thought would be best for the group to three main formulae.  The three chosen methods were: the Massey Method, the Ken Pomeroy method, and a homebrewed method using point-per-possession efficiency.  We then trained the methods individually using cross validation, and began estimating the outcomes of the last 1/4th of games in a season.

While estimating the outcomes of the games, we tallied up the total number of correct predictions out of the total number of predictions made.  This gave us a percentage accuracy that we used to represent how effective each method was estimating the outcomes of games.  As a group, we decided that this was the best way to determine which method was ‘better’, since when it comes down to it, the most important guess to make is which team will win the match, rather than the point spread itself.

The results of our tests were as follows:

* Homebrewed Method: 72%
* Massey Method: 68%
* Ken Pomeroy Method: 68%

All three of our methods were tested on the same data sets, using the same games to train and test them.  The raw percentage of games predicted correctly may not have varied much between Massey and Ken Pomeroy, but the way the teams were ranked varied greatly between each of the three methods.

# Conclusion

After analyzing the data received from our tests, we came to the conclusion that the best method for predicting the outcome of sports games was our Homebrewed Method.  It ranked the highest in correct predictions.  Additionally, the teams it ranked highest this season appear to hold up against the brackets for the NCAA March Madness.

However, one of the biggest problems with estimating a basketball games, is that anything can happen.  Both before and during a basketball game, there are seemingly unlimited amounts of uncatalogued variables that should be accounted for; in order to obtain a truly confident prediction. Given the limitations of both current basketball statistics and technology, it is nearly impossible to be able to take all of these factors into account.

This is further reinforced by the fact that the NCAA teams do not always perform as the stats dictated they should. If our statistics were more accurate, machine learning would have a more realistic chance of predicting the outcome of any given game.  But at the end of the day, the second strongest team in the bracket can still be beaten in their first game by the 13th strongest, despite the statistics saying the 13th strongest should have had almost 0% chance of actually winning.

# Bibliography

|  |  |
| --- | --- |
| [1] | K. Massey, "Statistical Models Applied to the Rating of Sports Teams," Bluefield College, Bluefield, Virginia, 1997. |
| [2] | Sal Khan, *Determining Invertible Matrices,* Khan Academy, 2008. |
| [3] | D. M. Parry, "Let's go!," Appalachian State University, Boone, 2016. |
| [4] | K. Pomeroy, "Advanced Analysis of College Basketball," The Forecast Factory LLC, 1999. [Online]. Available: kenpom.com. [Accessed 29 March 2016]. |
| [5] | A. W. Moore, "Cross-Validation for Detecting and Preventing Overfitting," Carnegie Mellon University, Pittsburgh. |
| [6] | inversion, *Root Mean Squared Error (RMSE),* Kaggle, 2016. |