Team Control Number

For office use only
T1 _____
T2 _____
T3 _____
T4 _____

31406

Problem Chosen

B

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2014 Mathematical Contest in Modeling (MCM) Summary Sheet
(Attach a copy of this page to your solution paper.)

Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page.

In this paper we employ mathematical models, specifically a combination of graphical and machine learning models, to identify the "all-time greatest male college coach" of the last century for basketball, football, and baseball. Since "best" is ambiguous, we employ a more qualitative approach to assessing our models' ranked coach lists through a combination of popular polls and general consensus.

We first consider the available data. For football and basketball, we collected career records and postseason appearances for thousands of coaches. We also collected the outcomes of postseason games (NCAA tournament for basketball and bowl games for football). Unfortunately, the amount of data readily available for baseball pales in comparison, and we had to develop a simpler model to choose the best baseball coaches.

We develop several models of varying complexity. The simplest models rank coaches based on the most easily-collected data, namely career win and loss totals. We discuss the inherent strengths and weaknesses of this approach, and develop a graphical model (which we call CoachRank) that makes more complete use of the data available and better models interactions between the coaches. CoachRank compares pairs of coaches based on games played between them to build a directed graph. We find that CoachRank's "best" coaches list is remarkably in line with top polls and intuitively make sense, and the supervised machine learning models extract championship appearances and win ratio as statistically significant attributes after removing heavily correlated attributes.

Here we give a preview of CoachRank results for football and basketball, and the results of our simpler model for baseball:

Rank	Football	Basketball	Baseball
1	Joe Paterno	Mike Krzyzewski	Ed Cheff
2	Mack Brown	Dean Smith	Gene Stephenson
3	Bear Bryant	Roy Williams	Mike Martin
4	Lloyd Carr	John Wooden	Augie Garrido
5	Pete Carroll	Rick Pitino	Gordie Gillespie

Furthermore, to provide intuition to the sports fans about how CoachRank comes up with these ranked lists, we use supervised machine learning methods and variable selection algorithms to extract out the most important coach attributes and comment about extending this model to other genders and sports.

College Coaching Legends: CoachRank

Control Number 31406

February 10, 2014

Abstract

In this paper we employ mathematical models, specifically a combination of graphical and machine learning models, to find the "best all time male college coach" in the last century for basketball, football, and baseball. Since "best" is ambiguous, we employ a more qualitative approach to assessing our models' ranked coach lists through a combination of popular polls and Wikipedia. We start with using linear and nonlinear transformations of simple metrics such as wins, losses, games, but move on to create a graphical approach that is better able to model the interactions between coaches which we call CoachRank. Due to the complicated nature of the graphical model, we seek to understand what the most important attributes are that CoachRank uses to find the "best coach". We do so through a supervised learning approach to construct interpretable models that sports fans can better understand, and generalize a model that works for across time periods and genders. We find that CoachRank's "best" coaches list is remarkably in line with top polls and intuitively make sense, and the supervised machine learning models extract championship appearances and win ratio as statistically significant attributes after removing heavily correlated attributes. Lastly, we partition the last century into different buckets and test both CoachRank and the machine learning model to find that CoachRank performs just as well while the machine learning model extracts out a smaller set of statistically significant attributes which may mean as the game evolved over time, people's notions of "best coach" has also evolved leading to a larger set of features now than a hundred years ago.

Contents

1 Problem Statement

Sports Illustrated, a magazine for sports enthusiasts, asks, "Who are the greatest college coaches, male or female, of the last century?" To answer this, we constructed a mathematical model to choose the top coaches in men's Basketball, men's Football, and men's Baseball. We also use these models to consider the difference between coaching in the past and in the present. Finally, we present an article for Sports Illustrated detailing our metrics and results at a non-technical level.

2 Model Assumptions and Data Collection

We make several assumptions in our modeling process:

- The idea of "best all time college coach" is subjective in nature, and to assess our model we have to focus more on qualitative assessment and justification, instead of a quantitative measure.
- A coach's rank is determined and only determined by their achievement in the sport they coach, and is unrelated to other factors in their personal life.
- The data we collect is considered true in all circumstances.

2.1 Data Collection

The accuracy and relevance of our results depended heavily on the quality and volume of data we could collect. That said, data collection was perhaps the most difficult part of our modeling process. We found that data on college sports is much harder to come by than data on professional sports. Within the scope of our search, less popular sports like Hockey simply did not have easily accessible data, especially from before 2001. We often found that when a site stated they had statistics on file, this meant that they had scanned copies of the original documents, which we could not parse.

For what was available, we had to write web scrapers to collect the data in its human-readable format. This made collecting data on the largest scales time-prohibitive, even with generous caching and multithreading. Finally, we had to be mindful of the data use policies of our sources, and ensure that our collection methods used as little bandwidth as possible.

We managed to collect a fair amount of data for Football, Basketball, and Baseball, the most popular college sports. All the data for these three sports were taken from Sports-Reference. [2] For Basketball, we collected the career statistics for all coaches listed, back to 1895, for a total of over 3500 coaches. These statistics include total wins and losses, conference championships and appearances, and NCAA tournament statistics. We also collected per-school statistics for each coach. Finally, we collected the outcomes and point data for every NCAA tournament game. Aforementioned limitations prevented us from collecting outcome data for every regular season game.

For College Basketball there are 8 career statistics features that we will use in our model:

- **years_active**: number of years the coach has been active
- **wins**: number of games the coach has won

- **start_year**: the year the coach started coaching
- **ncaa_appearances**: number of seasons the coach participated in the conference
- **games**: number of all games played by the coach
- **ff_appearances**: number of final four appearances
- **conf_tourn_wins**: number of conference tournament wins
- **conf_reg_wins**: number of conference regular season wins

We collected a similar scope of data for Football. We acquired career wins, losses, and bowl game statistics for every Football coach listed, back to 1877, for a total of over 2000 coaches. We also collected per-school statistics for each coach. Finally, we collected the outcomes and point data for every bowl game listed. Once again, time and data policies prevented us from collecting data on every game.

For Baseball, we were only able to collect career statistics for about 100 coaches, each with over 1000 wins.[3] That data was backed by a NCAA PDF [4] that was presumably human-entered. Despite the huge gap in data volume, we chose Baseball because we could not find historical data on any level for the other sports listed.

2.2 Coaches vs Teams

One of the key assumptions we make in our approach is that in general, "best" coaches are largely independent of the team that they coach. This is a bold claim, but an important one as if this were true it allows us to simply analyze the coaches rather than the specific teams they coached and the players another way. Said another way, the mathematical models would be far more complex if the coaches and teams were tied together in terms of determining who is the "best coach".

We discuss the methodology used for testing this assumption for Basketball and for conserving space we don't discuss it for Football and Baseball. We test this assumption by observing the per-school statistics for each coach. For each (Coach, School) pair we have the percentage of games that the team won with that coach. From this dataset we know that if the coach has achieved approximately the same win percentage then we have strong reason to believe that the coaches and teams are independent enough to only construct the mathematical model around coach data without delving into player and team specific data. To test this formally, we use the Pearson chi squared statistical test to test if there is a discrete uniform distribution across the win percentage across all the teams coached. Out of the 3513 basketball coaches, 755 of them had coached more than one team in their career. Of these coaches, 87.6% of them had a discrete uniform distribution across the number of win percentages, which largely says that we are fine in constructing mathematical models with just the coach data and not diving into the player data. We repeated similar statistical tests for football and found similar results.

3 Models

3.1 Simple Win & Loss Models

We began with the simplest models and continually improved them. Naively, the most indicative measure of a coach's capability is the number of games won and the number of games lost, and this information is also easy to obtain. Therefore we started by considering different algorithms to rank coaches based only on their number of wins and losses.

3.1.1 Sort by Win Percentage

The simplest model involves computing the win percentage. That is, we score each coach by the ratio of wins to total games. This method has obvious flaws: it doesn't take into account the number of games played by the coaches: a coach with a 4-1 win-loss record is intuitively not better than a coach with a 100-70 record, though this model would say so.

3.1.2 Sort by Net Wins

Another simple model involves computing net wins. That is, we score each coach by wins minus losses. Intuitively, this takes into account the magnitude of the number of games won and lost. However, this can be shown to simply favor coaches who have played more games.

3.1.3 Sort by Scaled Win Ratio

A third simple model involves multiplying the win ratio by the number of wins. This model is more robust than the previous two models because it takes into account both ratio of wins and losses and the magnitude.

We subjectively considered the results produced by each of these model and decided that sorting by squared win and loss ratio produces the best results with no filtering of the data. This is the model we resort to given only career win and loss data.

3.2 Graphical Model

We had access to more data than career win and loss totals, and we incorporated connectivity data between coaches into a graph-based model. One can consider the nodes of the graph to be the coaches and the edges to represent relationships between pairs of coaches. Naturally, the edges between coaches in our graphs are based on games played between them. Depending on the situation, the edges could be weighted and/or directed either to or from the better coach.

The main advantage of a graph-based approach is that it can observe direct interactions between pairs of coaches rather than considering coaches in isolation. It also allows us to apply transitivity to our data. That is, the relationship between two coaches can now be influenced by the pair's interactions with outside coaches. The approach even allows inferences about coach pairs who never played against each other by looking at paths in the graph. Therefore we can draw conclusions about coaches even though they came from different time periods.

3.2.1 Topological Ordering

The first idea that came to mind was to apply a topological ordering to the graph. That is, if we consider edges between pairs to point to the better coach, we want to remove coaches from the graph such that at time of removal, that coach has no incoming edges (and is therefore the worst coach). The idea is that the coach removed later would be the better.

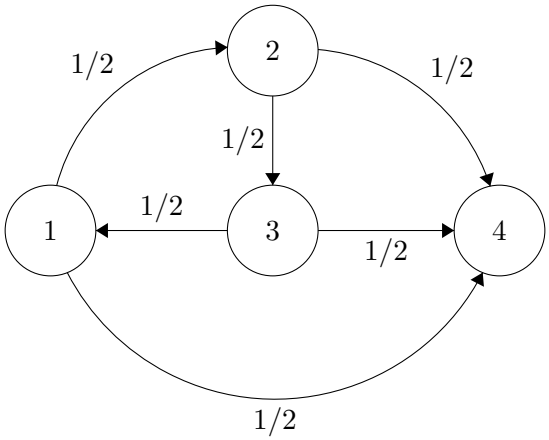
The major problem with this approach has to do with cycles. In order to construct a topological sort, we would either need a graph with no cycles (implying an already-existing partial order on coaches), or a mechanism for removing cycles from the graph. We could not create a satisfactory heuristic for removing cycles (that is, one that didn't seem too arbitrary).

3.2.2 Markov Chain (CoachRank)

Rather than a model that requires we remove cycles from our graph, we decided it would be best to create a model that relied entirely on the structure of the graph created from the data.

Imagine instead of having us, the spectators doing the voting and the ordering, each coach will have a **vote** of a certain **size**, which the coach can divide up into pieces. The coach can then give the chunks of the vote to different coaches which he thinks is better than himself. We assume the coaches to be rational: the coach doesn't take into account the opinions of amateurs and public media, and he only gives out votes to the coaches who have beaten him. Then we will rank the coaches based on how many votes each coach eventually gets. Since the coaches form a graph in which there might be cycles, at each timestep a coach can be both receiving and giving votes. What we are interested in is the distribution of the votes in the long term.

Consider the following example. There are four coaches: coach 1, coach 2, coach 3, coach 4. Coach 1 has been beaten by coaches 2 and 4, coach 2 has been beaten by coach 3 and coach 4, coach 3 has been beaten by coach 4 and coach 1, coach 4 has never been beaten.



The Markov Chain above shows how the votes will be given out at each timestep: coach 1 will give out half of his vote to coach 2 and the other half to 3, coach 2 will give out half of his vote to coach 3 and the other half to coach 4, coach 3 will give out half of his vote to coach 1 and the other half to coach 4. Coach 4, since he never lost to anyone, will simply keep his vote to himself. This can

be modeled as a Markov Chain, and the distribution of votes at some large time is the stationary probability distribution of this Markov Chain:

$$\pi = \lim_{n\to\infty} \phi_0 P^n \text{ for some initial distribution } \phi_0$$

where $\mathbf{P}$ is the probability transition matrix, with $\mathbf{P_{i,j}}$ equal to the proportion of coach i will give to coach j (can also be interpreted as the probability of going from i to j in the Markov Chain):

$$\mathbf{P} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

We know that if the Markov Chain is irreducible and ergodic, then a stationary distribution exists and is unique. This is the left eigenvector of $\mathbf{P}$ for the eigenvalue 1.

Irreducibility & Ergodicity

The example we show doesn't satisfy the property because it's possible for coaches to be undefeated. To make sure the matrix we encounter will be irreducible and ergodic, we use a similar technique to the PageRank[1] algorithm used by Google:

1. Dead-end coaches (coach 4 in the example), will split his vote into equal pieces at each timestep and give one piece to every coach, (go to a random state in the Markov Chain)
2. For non-dead-end coaches (coaches 1, 2, 3 in the example), in each step, will first split their vote into two, with proportion α and $1-\alpha$. With the α proportion he did what he did before, splitting it and give the pieces to coaches who have beaten him, the remaining $1-\alpha$ he will split into equal pieces and give one piece to every coach.

Therefore, after adjusting for dead-ends in $\mathbf{P}$, the new probability transition matrix can be represented as:

$$\mathbf{D} = \alpha\mathbf{P} + (1-\alpha)\mathbf{U}$$

where $\mathbf{U}$ with all entries equal to $\frac{1}{n}$, n being the total number of coaches (or number of states in the Markov Chain).

With the new probability transition matrix, since all entries in $\mathbf{D}$ are positive, we know by Perron-Frobenius theorem that $\mathbf{D}$ is both irreducible and ergodic, and that the stationary distribution π exists and is unique. With π we can find the coaches with the top 5 probability and they will be the "best" college coaches of all time since they get the most proportion of votes from all coaches in the long-run.

Formal Representation

$$G(V, E) \text{ represents the graph of coaches}$$

$$V = \{v_1, v_2 ... v_n\} \text{ where } v_i \text{ is coach i}$$

$$(v_i, v_j) \in E: \text{ coach j beats coach i more often than coach i beats coach j}$$

$$\mathbf{P}(i,j) = \begin{cases} \frac{I_{\{(v_i,v_j)\in E\}}}{d(v_i)}, & \text{if } d(v_i) \geq 1 \\ \frac{1}{n}, & \text{otherwise} \end{cases}$$

$$\mathbf{D}(i,j) = \alpha \mathbf{P}(i,j) + (1-\alpha)\frac{1}{n}$$

Efficient Implementation

Since the matrix $\mathbf{D}$ is huge and all entries are non zero, calculating its eigenvalues and eigenvectors directly could be very inefficient computationally. However $\mathbf{P}$ is a sparse matrix and most of its entries are zero. To make sure we will be able to produce the eigenvector that represents the stationary distribution efficiently, we use the Power Method:

Algorithm 1 The Power Method

$v = (1/n, ..., 1/n)$
while $\|\alpha v\mathbf{P} + (1-\alpha)(1, ..., 1) - v\|_2 < e^{-8}$ **do**
    $v = \alpha v\mathbf{P} + (1-\alpha)(1, ..., 1)$
return v

Since $v\mathbf{P}$ can be calculated using sparse matrix $\mathbf{P}$, this algorithm is much more computationally efficient than simply calculating the eigenvectors of $\mathbf{D}$. As α increases, the time it takes for v to converge is shorter, but the difference in the ranking might be less significant, therefore this is a tradeoff and we pick $\alpha = 0.95$.

Edge Weights

In the model we introduced above, in situations where $d(v_i) \geq 1$, $\mathbf{P}(i,j) = \frac{I_{\{(v_i,v_j)\in E\}}}{d(v_i)}$. However there is a lot more information we can capture:

1. Importance of the game: The more important the game is, the game result will be more informative, and more likely coach i will be to give the vote to coach j.

2. Career: If it is early in the career of coach i when he lost to coach j, he is going to put less weight than the game he lost later in his career, since it is typical for a coach to learn from failure early and become a better coach later. Therefore coach i will be more convinced to give votes to someone he lost to later in him career.

3. Game Score: A close game is less convincing than a game entirely dominated by an opponent. Therefore the more points coach i lost, the more convinced he will be to give votes to his opponent.

There is more information we can obtain; however, due to the difficulty of data collection, we come up with the following updated formula to calculate the weight for edge weight:

Let T be the set of all games
$f : T \to \mathbb{R}$ maps a game to the score difference of the game (positive winning score, 0 if draw)
$h : T \to \mathbb{I}$ returns the importance multiplier of a game, it is sport specific since different sport have different game structure.
β : the score difference adjustment, it is specific to a type of sport.

$$T_{i,j} = \{t \in T : \text{ coach i lost to coach j in } t\}$$

$$w(v_i, v_j) = \max(0, \sum_{t \in T_{i,j}} h(t) \log(1 + \beta f(t)) - \sum_{t \in T_{j,i}} h(t) \log(1 + \beta f(t)))$$

$$(v_i, v_j) \in E \text{ if } w(v_i, v_j) > 0$$

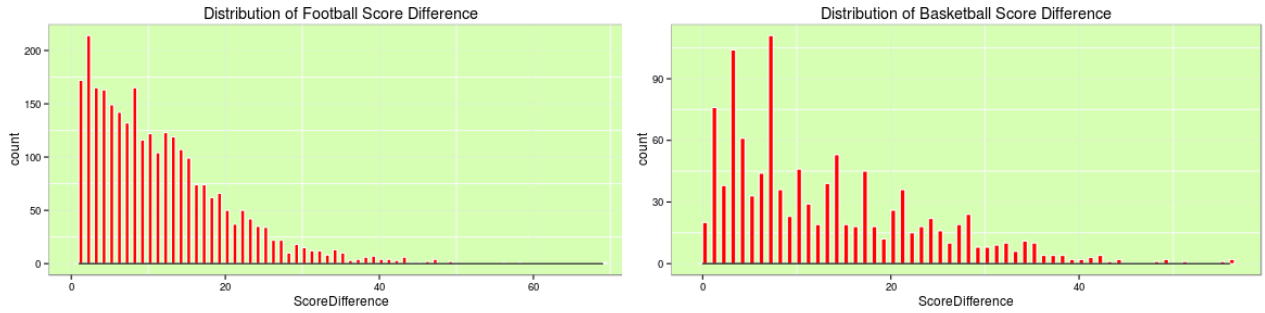
$$\mathbf{P}(i,j) = \begin{cases} \dfrac{w(v_i, v_j)}{\sum_{w(v_i,v_k)>0} w(v_i, v_k)}, & \text{if } d(v_i) \geq 1 \\ \frac{1}{n}, & \text{otherwise} \end{cases}$$

The intuition for this formula is really straightforward: for example, if a coach lost 30 points in championship game, it definitely is going matter much more than a loss of 4 points in an early-season game. The weight for an edge is a combination of the two factors: the importance of the game and the score difference.

Parameter Estimation

h : we let $h =$ (number of games of a given importance)$^{-1}$. The reason for this approximation is clear, the less frequent a type of game is, the more it is valued. Say for example that a team plays 30 regular season games, 5 tournament games, and 1 championship game, then h(season) = 1/30, h(tournament) = 1/5, h(championship) = 1. Similarly, in College Football, h(season) = 1/12, h(playoff) = 1/12, h(championship) = 1

Let β : be the inverse of the median of score differences of a sport in the data. From the data, we calculated the median and set $\beta_{basketball} = \frac{1}{9}$, $\beta_{football} = \frac{1}{10}$. We do not use the mean because there are outliers in the score difference distribution, as we can see in the following graph:



There are a lot more possible variations with edge weights in the graphical model. However, as more detailed information is added into the model, the less significant it is, especially among the top results returned. Therefore due to time constraints the above model and method of parameter estimation is the final version we decided to use for our graphical model. As we can see in the results and validation section, the results are promising.

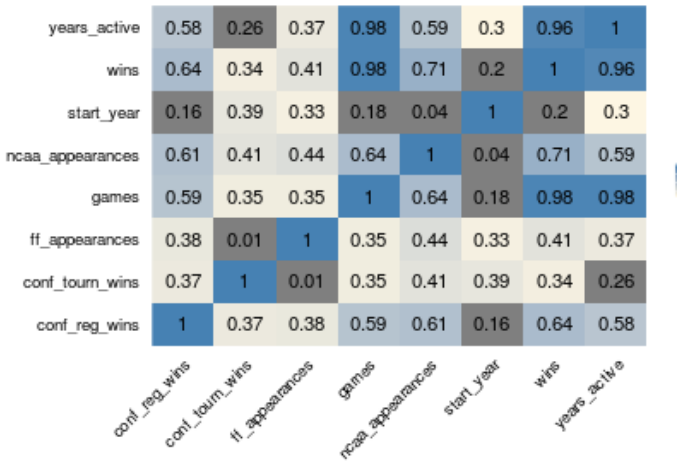
3.3 Machine Learning Model

So far we have algorithms that produce the top 5 "best coaches" according to historical data, but we could also interpret "best" as who has the best chance of winning in the future, or is overall "best" looking forward. Additionally, it is difficult to assess the CoachRank graphical model or more importantly figure out what attributes of a coach are the most important for defining "best coach". There are a lot of sports fans who may not understand the whole graphical model, but can understand certain key attributes of coaches. This leads us to the use of machine learning methods for tackling this unsupervised learning problem- we found significant difficulty finding structure within the massive amounts of sports data, but we realized that the CoachRank graphical model posed above can bring *structure* to formulate a supervised learning problem. Furthermore, we can use variable selection methods to extract out which are the most important attributes for deciding who the best coach is.

Feature Correlation

Before we start to use machine learning methods, it's important to get a feel for the data to make sure the models we produce are intuitively sound. After the initial data munging phase, we realized that a lot of the attributes (now on referred to as features) were correlated, which made sense because our data sets had wins, losses, total games played, etc. To make it the most interpretable and simplest model, we removed features that had a high degree of correlation. Below is the Pearson correlation matrix for basketball:

Figure 2: Pearson Correlation matrix for basketball features



From the matrix above, we can clearly see that there are some attributes (wins, losses, games, etc.) that are colored dark blue, which we removed moving forward to create a compact set of features that explained what constituted as the "best coach" without redundancy. There are a wealth of supervised machine learning methods that we can use, but we have two defined goals: figure out what features make up the best coach and assess how well the graphical model works. To best achieve these goals, we decided to use multiple linear regression to figure out how each feature correlates with the coach's CoachRank value. This is also easy to explain to sports fans.

Multiple Linear Regression

Multiple Linear Regression is a supervised machine learning method that predicts a real valued output given a set of features. More formally, the multiple linear regression function is:

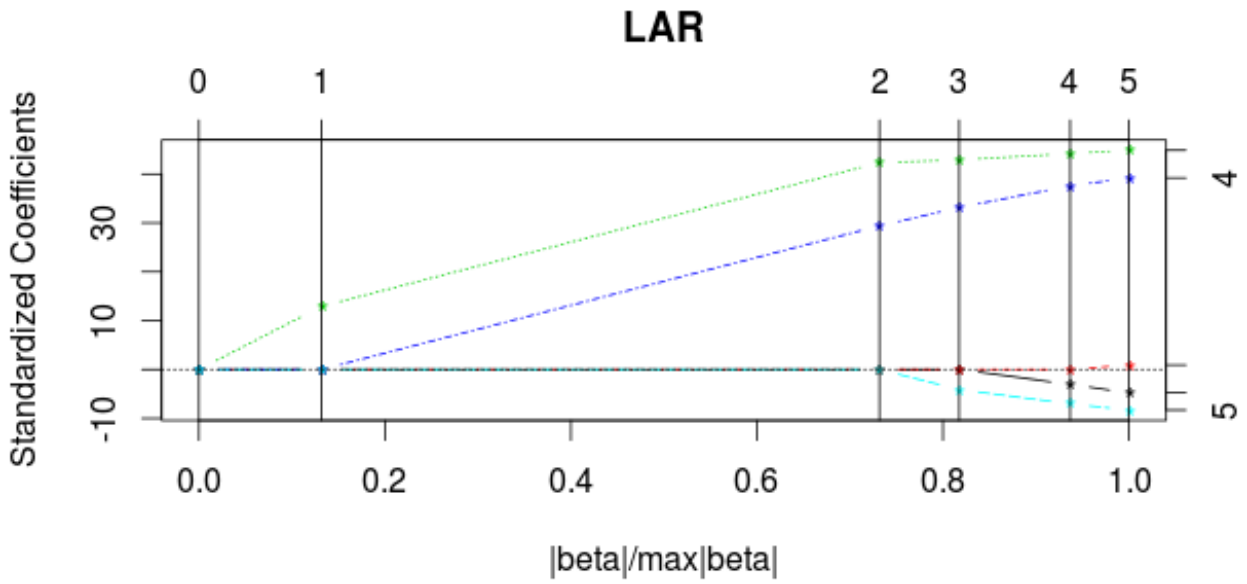
$$E[CoachRankValue|X] = \gamma + \sum_{i=0}^{n} X_i$$

That is, we are finding the conditional expectation of the CoachRank value of a coach given a set of features X. Since we would not only like to predict CoachRank values for coaches but also individually analyze the features, we handled multicollinearity by using the Pearson correlation matrix above to ensure none of the features were correlated above a minimum threshold of 0.9. We also looked at the distribution of the features to make sure the homoscedasticity, weak exogeneity, and independence assumptions of linear regression held true. Once we performed the analysis, we observed the plots of the residuals to ensure that they were in fact normally distributed as that's another major assumption of regression. Below are the results of the multiple linear regression model:

Features	Coefficients
conf_reg_wins	-0.02572
conf_tourn_wins	0.24730
ff_appearances	1.47954
ncaa_appearances	0.24730
games	0.00210

Although not as important, this model resulted in an adjusted R^2 of 0.892, which implies that the set of features fit the CoachRank values quite well. More importantly, the features that were statistically significant at the highest threshold were **ncaa-appearances** and **ff-appearances**, which intuitively make sense and in fact are used in the construction of the CoachRank graphical model. Additionally, the sign of the features make sense as the more final four and NCAA appearances a coach has the more popular or "good" they are likely to be. What's not intuitive is the real-valued coefficients of the features, and the number of features that we have: does one feature "predict" the CoachRank value of the coach better than other features? That is commonly referred to as variable selection, where there currently exist a wealth of methods. We use least-angle regression (LARS) for computing which features best correlate with the CoachRank. In a nutshell, LARS initializes all the coefficients of the features at zero and takes the largest step in the direction of the most correlated variable with CoachRank, until some other feature is more correlated, at which point LARS proceeds in a equiangular direction to both features, hence the name "least-angle". At the end of the day, LARS provides us which features that are the most "important" in determining the CoachRank values. The plot is shown below:

Figure 3: LAR of Basketball Features



Features	Label
conf_reg_wins	1
conf_tourn_wins	2
ff_appearances	3
ncaa_appearances	4
games	5

As expected, the **ncaa_appearances** and **ff_appearances** enter first and are thus deemed to be the most "important". Now it remains to figure out how to make the coefficients of the features more interpretable in a way that sports fans can understand it. Fortunately, interpretable models in

supervised learning are a growing area of research within predictive modeling and machine learning. In the next section we discuss a natural extension to the multiple linear regression model discussed above known as Supersparse Linear Integer Models [5].

Supersparse Linear Integer Model

Supersparse Linear Integer Models (SLIM) create predictive scoring systems that are both practical and interpretable. They are widely referred to as interpretable models because they require users to perform only a few operations to make a prediction. They further restrict the coefficients to a certain set of integral values so that we can better understand the effect of the features on the predicted value. SLIM is formulated as a mixed integer program with an objective function that minimizes 0-1 training loss and L_0 and L_1 norm to ensure both accuracy, sparsity, and interpretability respectively. Although this is a computationally hard problem (NP-Hard), due the size of our data set and CPLEX solver from IBM we are able to obtain results within a reasonable amount of time. Additionally, as this is a classification model, we had to have a surjective mapping of our real valued CoachRank values to discrete buckets (< 0.01, > 0.02, etc). Since we had a imbalanced data set that resulted, that is the number of coaches falling into each bucket varied widely, we used the imbalanced formulation of SLIM. After doing so, we obtained results that, as expected, are close to the results of our multiple linear regression model but now have integral values for the features.

Can we abstract this?

Okay, so we have these neat supervised machine learning methods that complement the complicated graphical CoachRank algorithm that we discussed earlier- so what? Well, the purpose of the machine learning methods was two fold: one to extract features from the CoachRank graphical model to actually determine how it's figuring out which coaches are the "best" and two to understand how these features interact with each other and affect the CoachRank value in a practical and interpretable way, both of which were done through the multiple linear regression model and supersparse linear integer model discussed above.

A natural question that one may now ask is how can we extend this trained machine learning models to other time periods and other genders. Well, a nice feature about both the graphical model and the supervised learning model is that they are constructed for specific sports: sure, they might have inputs as "final four appearances" or "bowl games", but almost every sport has major competitions and team games that can be substituted for this. Additionally, nowhere in either model do we use gender specific criteria, so our assumption is that these models above will hold for both genders *assuming* both genders have similar abstract characteristics, which we believe at a first order approximation to be true but would test rigorously if we had more time. As far as different time periods go for the machine learning methods, we partitioned the coaches into different data sets depending on their start year and found that the most significant extracted features were in fact largely the same, which is no surprise because there should not be a significant difference between the most important features that determine the best coach between different time periods. However, it is important to note for earlier time periods (1913-1940), there were less features that were found as statistically significant which might mean that the notion of "best" coach was more simplistic way back when but now has grown more complicated as the game and society have evolved: an interesting observation.

4 Results & Validation

Below we present the results of our graphical model. Using the graphical model with edge weights that take game importance and score difference into account, we computed the following results and showed the top five coaches for Male College Football, Male College Basketball, and Male College Baseball:

Male College Football

The graph $G(V, E)$ for Male College Football has 529 nodes, 1032 edges, and 27 weakly connected components. The fact that there is no one single connected component is possibly due to the small size of our dataset. The top five coaches are:

Coach Name	Vote %
Joe Paterno	0.02421
Mack Brown	0.01728
Bear Bryant	0.01663
Lloyd Carr	0.01491
Pete Carroll	0.01338

Male College Basketball

The graph $G(V, E)$ for Male College Basketball has 763 nodes, 2582 edges, and 1 weakly connected component. The fact that G is weakly connected is really useful because it allows us to compare every two coaches in the graph, even though they came from different time horizon. The top five coaches are:

Coach Name	Vote %
Mike Krzyzewski	0.03272
Dean Smith	0.02544
Roy Williams	0.02329
John Wooden	0.02170
Rick Pitino	0.02066

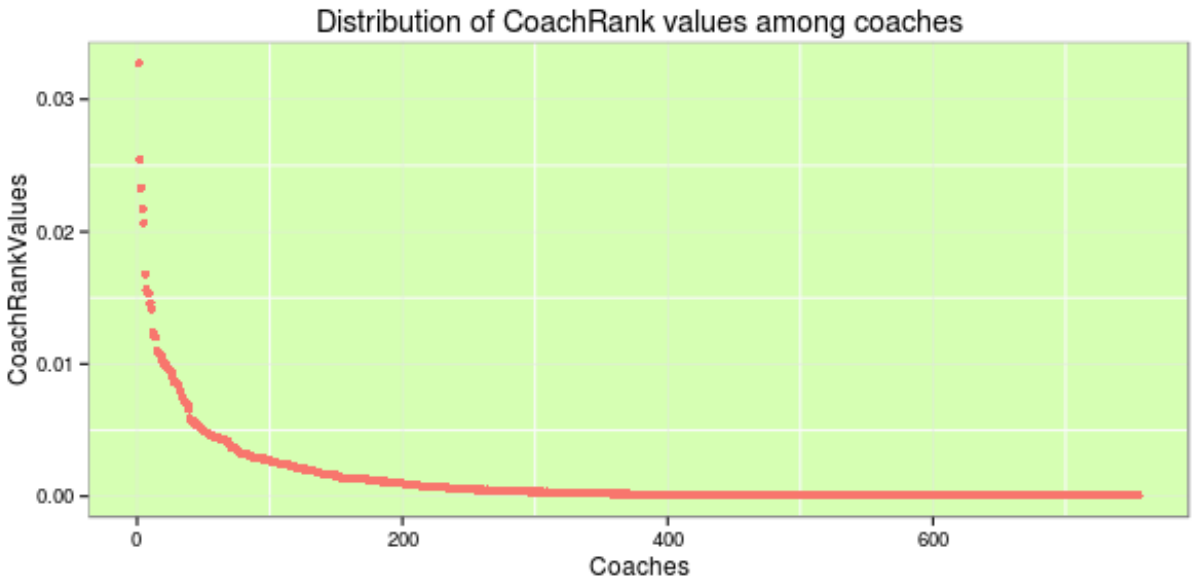
Male College Baseball

Due to the lack of game data between baseball coaches, we resort to sorting by scaled win ratio.

Coach Name	Scaled Ratio Score
Ed Cheff	1361.60
Gene Stephenson	1343.38
Mike Martin	1316.73
Augie Garrido	1279.37
Gordie Gillespie	1259.56

Analysis of Results

If we plot the sorted vote % of all the College Basketball coaches, we can get the following histogram. Similar graphs were computed for Baseball and Football.

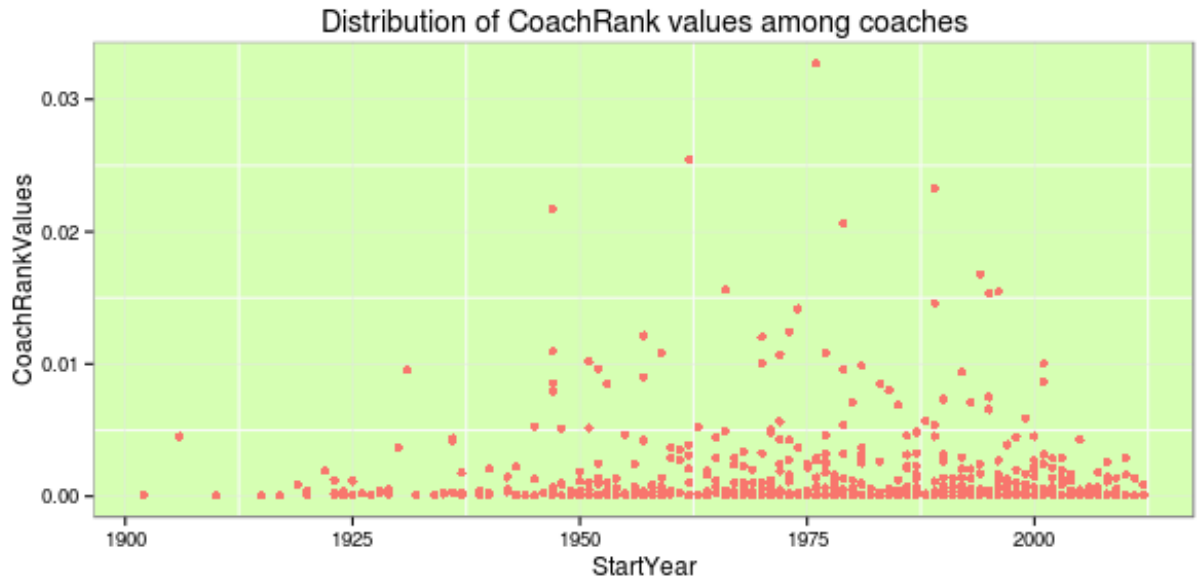


We find some very interesting facts from this graph:

1. The curves drop off very quickly, which means there are large differences between coaches with the top votes.
2. A small proportion of coaches hold the majority of the votes. In Male College Basketball, 5.6% of the coaches hold 50% of the votes. In Male College Football, 12.4% of the coaches hold 50% of the votes.

Comparison Across Time Horizon

If we plot the vote % of all the College Basketball coaches over the start year of their careers, we can get the following histogram:



We can see from the graph that the scores of the coaches have a positive correlation with the start-year of their career. This is reasonable because first techniques of sports tend to improve overtime and second due to the shortage of data for earlier periods we have more connectivity for coaches in later periods in the graphical model.

We also run the graphical model in separate time periods, and the results for the different periods (1900-1930, 1930-1970, 1970-2010) are reasonable based on information from public polls and professional sports media. Here are the results of segmenting our dataset into three time periods.

1900-1930	Vote %	1930-1970	Vote %	1970-2010	Vote %
Elmer Ripley	0.15947	John Wooden	0.07941	Mike Krzyzewski	0.05984
Vadal Peterson	0.12466	Dean Smith	0.04433	Roy Williams	0.03931
Howard Cann	0.12262	Fred Taylor	0.04429	Rick Pitino	0.03429
Chick Davies	0.08383	Pete Newell	0.04315	Bill Self	0.03383
Harol Oolsen	0.08383	Phil Woolpert	0.03836	Billy Donovan	0.02777

Applying the Model to Other Sports

Applying the graphical model to other sports are very straightforward. First obtain enough data of games played between coaches, the scores of the games, and the importance of the games. Then adjust β and h based on the median score difference and league structure. With this information we can run the graphical model and calculate vote % for each coach and select the top 5 results as the "best college coaches of all time".

Assessment

Due to the subjective nature of this problem, assessment of the results can be tricky. By comparing both of our results with public polls and professional sports media such as ESPN and Fox News,

etc., and looking at the achievements of the top coaches, we conclude that the results are consistent with public opinions.

Robustness

Varying parameters α from 0.75 to 0.95, and scaling β upwards and downwards by 10%, the top 5 coaches returned doesn't have much difference, just with minor rank movement among the top 10 coaches. This is in part due to the quick drop-off rate of the curves above, and the large score differences between top coaches. Therefore, we conclude that the result of the graphical model is valid and it is robust to small perturbations in parameters.

5 Strengths and Weaknesses

5.1 Strengths

- The graphical model allows us to be more objective in our ranking algorithm since we can understand it as the coaches voting among themselves based on their game history, instead of arbitrary tweaking of heuristics.
- We take into account not only the career data of a coach, but also the results, score differences, and importance of the games they play against each other.
- We have an efficient implementation using the Power Method to calculate the stationary probability distribution of the Markov Chain, and the results show clear differentiation between coaches, especially high-ranking ones.

5.2 Weaknesses

- Due to the limited amount of data we could collect, we could not consider coaches not in our dataset. For example, John Gagliardi, the coach with the most wins in college Football history, was not in our data set because he competed in the NAIA and NCAA Division III leagues.
- Our graphical model used only postseason games as input. We justify this heuristically by saying that only skilled coaches will be play in the postseason. However, this produces somewhat sparse graphs, especially compared to those we could have generated if we were able to collect data on every game.
- Assessment is also difficult since ranking is in itself subjective. Our methods of assessment are limited to public polls, professional sport media, coaches' achievements, and cross-validation between our different models.

6 Conclusions

We believe our metrics for ranking coaches are justifiable and that our results are believable. The coaches returned by our CoachRank graphical model are all acclaimed in their sport, and we believe that our objective metrics match well with subjective polls. Additionally, we confirmed

the complicated nature of CoachRank through supervised learning to extract a set of features that intuitively made sense since they were included in the construction of the graphical model and edge weight schemes. On a larger level, we used a graphical model to provide structure to make an unsupervised learning problem into a supervised one. Finally, we consider our model to make the most complete use of the data we were able to obtain.

7 Future Work

There is a lot of work that can be done to expand on our mathematical models discussed in the paper. To start with, we found it very difficult to come up with a concrete definition of "best coach", so we would spend more time doing due diligence on what constitutes "best". If we are able to quantify "best" we could remove the dependency of the supervised machine learning model on the graphical model. Furthermore, the simplistic models that generated lists of ranked coaches using an absolute metric were rather subjective; we would like to test these subjective metrics on a training set and see how much predictive ability they have on a test set.

Furthermore, we would have liked to look at the underlying distribution differences between gender to assess whether our mathematical models can in fact be generalized to other sports and genders. We have done initial analysis regarding this in our paper but certainly not enough to guarantee the model working on other genders and sports. Lastly, we would have liked to spend more time on data munging as the data we got is widely varied and in increasing volumes across sports.

8 Bibliography

References

[1] Pagerank Algorithm, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[2] Football and Basketball Data, http://www.sports-reference.com/

[3] Baseball Data, http://en.wikipedia.org/wiki/List_of_college_baseball_coaches_with_1,000_wins

[4] NCAA Coaching PDF, http://fs.ncaa.org/Docs/stats/baseball_RB/2011/Coaching.pdf

[5] Supersparse Linear Integer Models, http://web.mit.edu/rudin/www/UstunTrRuAAAI13.pdf

College Coaching Legends

Any sports ranking is going to be contentious. Players, coaches, and teams are often judged subjectively, not so much by how many games they win so much as how inspiring their story is. When we were given the challenge of deciding the greatest college coaches of all time, we knew that we'd have to be as objective as possible. Rather than choose any one way of picking the greatest coaches, we started with simple models and then extended to models that captured win-loss records between coaches, which we then compared to existing opinion polls and awards.

1 Simple Models

We considered how well the most common metrics for rating coaches work in this situation. Win ratio (wins over total games) is often used. We found that this doesn't work because it will put a coach with a 1-0 record above one with a 200-20 record. The reason why win ratio works when ranking good coaches is usually because every coach surveyed has played a similar number of games. We could have filtered our data (only consider coaches with a bowl win, etc.), but we wanted to remove arbitrary factors from our process. Another metric we tested was the win-loss difference. We found that this simply favored coaches with long careers.

To make up for the flaws of win ratio and net wins, we gave more weight to coaches that had a high win ratio but also have won more games by multiply win ratio by wins, which when we tested produced great results without any filtering. This effectively dismisses coaches who have played too few games, but doesn't allow one coach to dominate simply because of a high number of games.

2 Connecting the Dots

One problem with this model is that it doesn't consider the connections between coaches. Surely career stats don't matter when one coach consistently beats another. We explored the interactions between coaches with what is known as a graph model. The idea is like this: I get points from every coach that I beat, and I evenly distribute those points to coaches who beat me. We give every coach the same amount of points to start, and see where they stabilize. This adds transitivity to our model: if I beat another coach and he beats 100 other coaches, then I get a good share of all those points. This also has the effect of eliminating small outliers. A coach with a 15-0 record gets a lower score than one with a 200-30 record simply because the better coach has a better inflow of points.

For Basketball, we collected score data from every NCAA tournament game, and for Football we collected the results of every postseason bowl game. The original plan was to collect data on every regular season game ever as well, but that would have violated the data use policies of our sources. We were also unable to find any more data than tournament final scores for Baseball (the data exists but not in a usable format), and only considering two coaches every year would not have been enough data.

With the data that we could get, we did some math relating the score differences and game importance. For instance, a first-round game is not as important as a Final Four game, and an easy victory weighs more heavily than a close win. We used this data to say whether one coach beats another over the course of their games together. With this, we used a variant of Google's Pagerank algorithm (originally designed to rank website results) to rank coaches. We call this system CoachRank.

3 Results

Based on our models, our top five coaches in football, basketball, and baseball are:

Rank	Football	Basketball	Baseball
1	Joe Paterno	Mike Krzyzewski	Ed Cheff
2	Mack Brown	Dean Smith	Gene Stephenson
3	Bear Bryant	Roy Williams	Mike Martin
4	Lloyd Carr	John Wooden	Augie Garrido
5	Pete Carroll	Rick Pitino	Gordie Gillespie

4 Best Coach Attributes

Okay cool, so we have this model that spits out the top 5 coaches for a sport by looking at interactions between the coaches. Wouldn't it be interesting to find out what attributes of a coach made them show up in the top 5? It's really difficult to tell this from the graphical model because there's a lot of thousands of coach interactions going on. So to combat this, we took the CoachRank values that the graph model spit out for each coach and tried to put a line through the points based on attributes we thought might be important, such as when they started coaching, how many games they won, how many times they ended up winning the championship, etc. Once we ran our new model, we found that the attributes that gave the coach a high rank were number of times they appeared in the final round of a sport and how many games they had won. This also confirmed our graph model because although we didn't explicitly use those attributes in making the model, they still showed up as indicators of a good coach.

We believe these models have a minimum amount of subjectivity, which sets them apart from many other rankings and opinion polls. These models make the best use of the data available to us, but that doesn't mean they are infallible. The final say on who's the greatest coach can only truly be tested on the field, and we think we do well in capturing that.