CoachRank

Akilesh Potti, Dominick Twitty, Wenhai Yang

February 10, 2014

Contents

1 Introduction

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. 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.

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. That data was backed by a NCAA PDF 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 required 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 she thinks is better than herself. We assume the coaches to be rational: the coach doesn't take into account the opinions of amateurs and public media, and she only gives out votes to the coaches who have beaten her. 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.
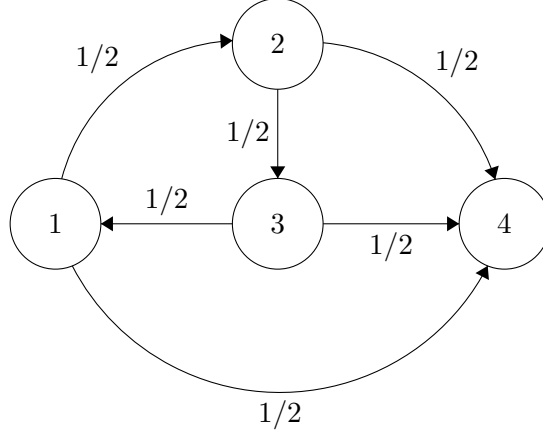
The Markov Chain above shows how the vote will be given out at each timestep: coach 1 will give out half of his vote to coach 2 and the other half to 3, coach 2 will give out half of her vote to coach 3 and the other half to coach 4, coach 3 will give out half of her vote to coach 1 and the other half to coach 4. Coach 4, since she never lost to anyone, will simply keep her vote to herself. 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.

Irreducible & Ergodic

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 algorithm used by Google:

1. For dead-ends (coach 4 in the example), in each step she will split her vote into equal pieces and give one piece to every coach, (go to a random state in the Markov Chain)

2. For non-dead-ends (coach 1, 2, 3 in the example), in each step with he will first split her vote into two, with proportion α and $1 - \alpha$. With the α proportion she did what she did before, spliting it and give the pieces to coaches who have beaten her, the remaining $1 - \alpha$ she will split into equal pieces and give one piece to every coach.

Therefore, after adjusting for deadends 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 can 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 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, she is going to put less weight than the game she lost later in her 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 vote to someone she lost to later in her 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 she will be to give her vote to her 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.

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

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

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

$$\mathbf{P}(i,j) = \begin{cases} \dfrac{w(v_i, v_j)}{\sum_{w(v_i,v_k)>0} w(v_i, v_k)}, & \text{if } d(v_i) \geq 1 \\[2ex] \dfrac{1}{n}, & \text{otherwise} \end{cases}$$

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 $h = (\text{number of games of a given importance})^{-1}$. The reason for this approximation is clear, the less frequent a type of sub-game is, the more it is valued. Say for example that a team playes 30 regular season games, 5 tournament games, and 1 championship game, then

$h(\text{season}) = 1/30$, $h(\text{tournament}) = 1/5$, $h(\text{championship}) = 1$. Similarly, in college Football, $h(\text{season}) = 1/12$, $h(\text{playoff}) = 1/12$, $h(\text{championship}) = 1$

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:

(a) Distribution of Basketball Score Difference
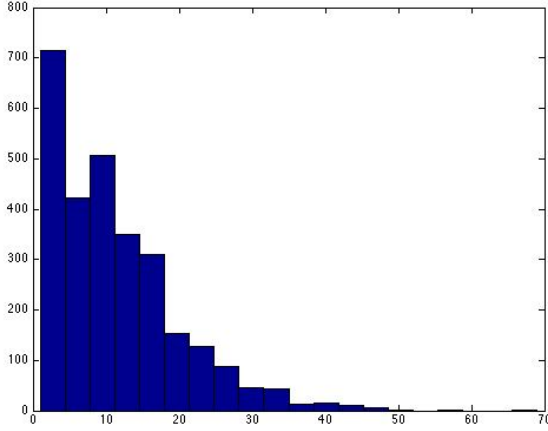
(b) Distribution of Football Score Difference
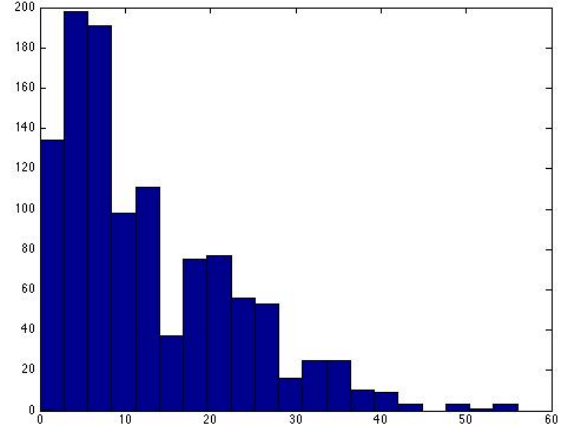
There are a lot more possible variations with edge weights in the graphical model. However, as more detailed information is added into the model, the less significant it is, especially among the top results returned. Therefore due to time constraints the above model and method of parameter estimation 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 refered 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 basketabll features

	conf_reg_wins	conf_tourn_wins	ff_appearances	games	ncaa_appearances	start_year	wins	years_active
years_active	0.58	0.26	0.37	0.98	0.59	0.3	0.96	1
wins	0.64	0.34	0.41	0.98	0.71	0.2	1	0.96
start_year	0.16	0.39	0.33	0.18	0.04	1	0.2	0.3
ncaa_appearances	0.61	0.41	0.44	0.64	1	0.04	0.71	0.59
games	0.59	0.35	0.35	1	0.64	0.18	0.98	0.98
ff_appearances	0.38	0.01	1	0.35	0.44	0.33	0.41	0.37
conf_tourn_wins	0.37	1	0.01	0.35	0.41	0.39	0.34	0.26
conf_reg_wins	1	0.37	0.38	0.59	0.61	0.16	0.64	0.58

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 regerssion. 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

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 intializes 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: LARS of Basketball Features
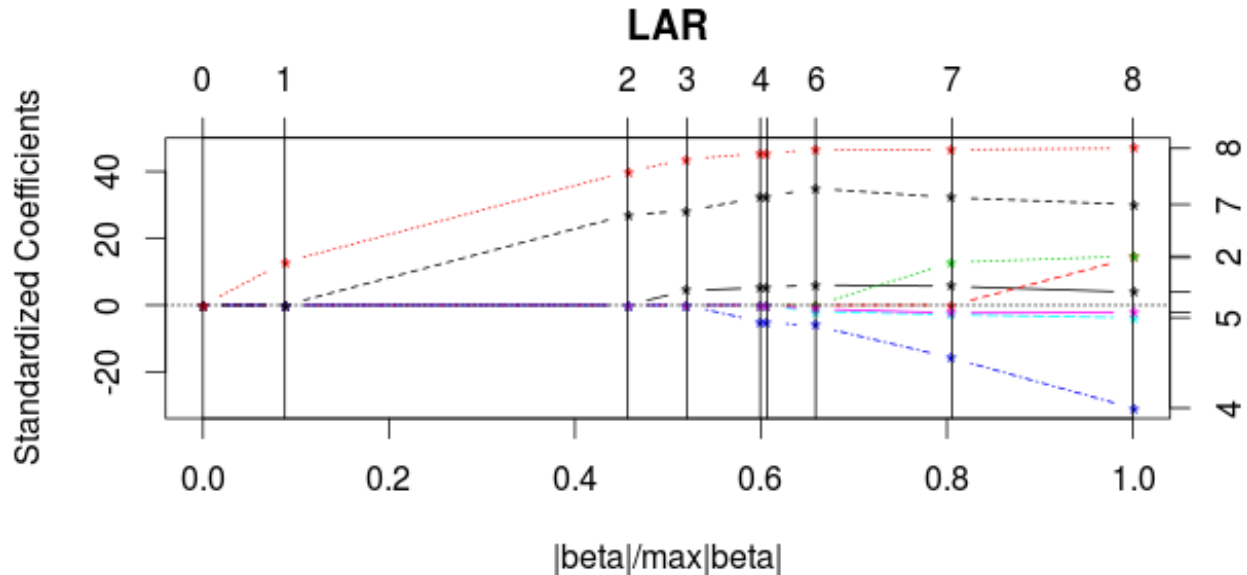
As expected, the $ncaa - appearances$ and $ff - appearances$ enter first and are thus deemed to be the most "important". Now it remains to figure out how to make the coefficients of the features more interpretable in a way that sports fans can understand it. Fortunately, intereptable models in supervised 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, which one of the group members is currently working on with Professor Cynthia Rudin at MIT, so are able to use the code although it's not publicly released yet.

Supersparse Linear Integer Model

Supersparse Linear Integer Models (SLIM) create predictive scoring systems that are both practical and interpretable. They are widely refered to as intrepretable 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 first 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 specifically 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, not where in either model do we construct the model on any 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. As far as different time periods go for the machine learning method, 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 we noticed in 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 in determining what constitutes as the "best" coach as the game and society has evolved: an interesting observation.

4 Results & Validation

4.1 Result of Graphical Model (CoachRank)

Using the Markov Chain model with the edge weight function that takes game's importance, and score difference into account, we computed the following result and showed the top five coaches for Male College Football and Male College Basketball:

Male College Football

The graph $G(V, E)$ for Male College Football have 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 have 763 nodes, 2582 edges, and 1 weakly connected component. The fact that G is weakly connected is really useful in that it allow us to compare every two coaches in the graph, even though they are playing in 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 connectivity data between baseball coaches, we resort to sorting by scaled win ratio.

Coach Name	$w(w/w + l)$
Ed Cheff	1361.60
Gene Stephenson	1343.38
Mike Martin	1316.73
Augie Garrido	1279.37
Gordie Gillespie	1259.56

If we plot the Vote % of all the College Basketball coaches from high to low, we can get the following histogram. Similar graphs were computed for Baseball and Football.

Figure 4: Male College Basketball Vote %
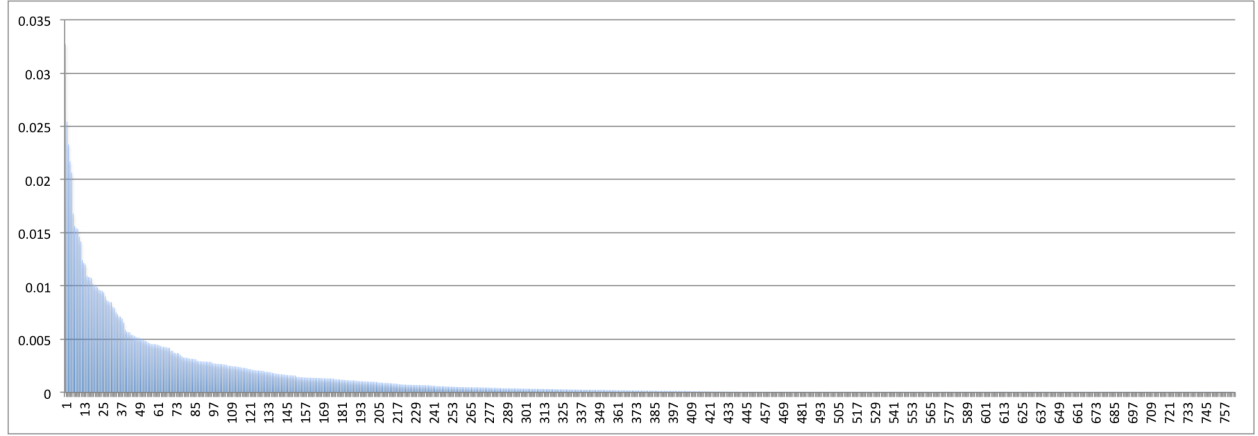
We find some very interesting facts from this graph:

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:

Figure 5: Vote % Distribution over Career Start Year
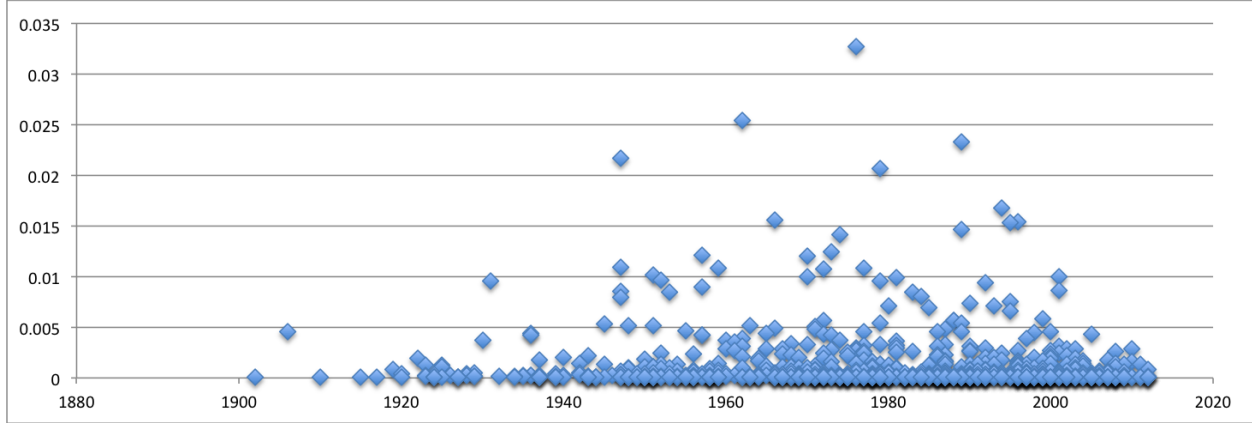
We can see from the graph that the scores of the coaches have a positive correlation with the start-year of their career. This is reasonbale because first techniques of sports tend to improve overtime and second due to the shortage of data for earlier periods we have more connectivity for coaches in 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 baed on information from public polls and professional sports media.

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 result is 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 game result, game score, and the game importance of the games they play against each other.

- We have an efficient implementation of using the power method to calculate the stationary 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 and Future Work

We believe our metrics for ranking coaches are justifiable and that our results are believable. The coaches returned by our model are all acclaimed in their sport, and we believe that our objective metrics match well with subjective polls. Finally, we consider our model to make the most complete use of the data we were able to obtain.

Given more time, we would have liked to come up with a more objective way of assessing the strengths and weaknesses of our models. We tailored our models to make the most use of the data available, and given more information on each coach we could keep improving our models. Likewise, having more data from each sport would increase the accuracy of our results.

7 Bibliography

References

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

[2] A stochastic optimization model for real-time ambulance redeployment,
`http://www.sciencedirect.com/science/article/pii/S0305054813000385`

[3] Ithaca Demographics, `http://www.ci.ithaca.ny.us/maps/index.cfm`

[4] Cornell Demographics, `http://www.cornell.edu/about/facts/stats.cfm`

[5] Ithaca College, `http://www.ithaca.edu/admission/facts/`