Team Control Number

29696

Problem Chosen

B

2014 Mathematical Contest in Modeling (MCM) Summary Sheet

College Coaches' Mount Rushmore

Summary

College coaches' exposure is growing these days, as a result of the increase on attention to college sports. In this paper, we build a mathematical model to rank college coaches among either gender across generations in varied sports.

Firstly, we build a factors pool to collect evaluation factors of all sports. By principle component analysis (PCA), we obtain three components used to evaluate the coach's performance. Then we take American college basketball as an example to build the Coaches Ranking Model (CRM). The first step of CRM is to classify the ranking coaches by Cluster Analysis. Next we compare all coaches with the "idealistic coach" whose factors are the best of factors pool by means of adjusted Cosine Similarity. We hereby obtain the top 5 college basketball coaches in the previous century.

Secondly, we optimize CRM across generations by modified coefficient of coaching difficulty and social influence. The coefficients are calculated by Regression Analysis. Based on the optimized model, we take the impacts of genders into consideration. Modified coefficient in genders is calculated by Analysis of covariance (AVOVA) and Data Whitening**.** We hereby obtain the optimized CRM.

Thirdly, the optimized model is applied to other sports in the worldwide. As for sports, we figure out an advanced ranking model related to different sports with the update of factors pool. Hence, we obtain the coaches ranking list of football, baseball and softball. As for countries, we obtain a list of top 5 college basketball coaches in China.

Finally, we test the sensitivity and robustness of the coaches ranking model. The test result illustrates the stability and adaptability of the model. Besides, based on the ranking results, we discuss the significance of the relation between ranking state distribution and economy, society, culture development in America. Generally, the prosperity of sports stimulates the development of the region.

Above all, our model is accessible to both genders across generations in varied sports. It can be applied further to other countries and organizations. Therefore, our model has a high degree of adaptability and stability.

Key Words: Adjusted Cosine Similarity, Principle Component Analysis, Data Whitening, Cluster Analysis, CRM

Model Execution Flow		
College Coaches Ranking Model		
STEP1: Collect evaluation factors of college coaches		
STEP2: Principle Component Analysis (PCA) of data for basketball coaches	Results and Conclusion	
	Classify 12 factors into 3 principle components Z_1 : coach-ability, Z_2 : coach-influence-school, Z_3 :coach-influence-society.	Principle components ranking by importance $Z_1 > Z_2 > Z_3$
STEP3: Cluster analysis of data	Results and Conclusion	
	Pick out the better groups in 9 for ranking with simplification the latter calculation.	
STEP4: Best coaches ranking by Cosine Similarity	Results and Conclusion	
	Ranking result： 1st:John Wooden 2nd:Geno Auriemma 3rd:Pat Summitt 4th:Dean Smith 5th:Mike Krzyzewski	
Model Optimization		
STEP1: Find the relation between generations andcoaching difficulty across time line horizons.	Process and Approach	Conclusions and Results
	Calculate modification coefficient of time line horizon by linear regression analysis	**1st**:John Wooden **2nd**:Geno Auriemma **3rd**:Pat Summitt **4th**:Dean Smith **5th**:Mike Krzyzewski
STEP2: Analysis of variance (ANOVA) of data and discussion of gender differences	Process and Approach	Conclusions and Results
	Analyze the significant difference in factors across the gender and modify data by Data Whitening.	1st:John Wooden 2nd:Geno Auriemma 3rd:Pat Summitt 4th:Dean Smith 5th:Mike Krzyzewski
Model Extension		
● Factors pool extension ● Sports extension: baseball, football, softball. ● Countries and organizations extension: China		
Further Discussion Upon Model		
Discussion upon sports impact on **economy, culture, social developments**.		

Sensitivity and Robustness Analysis

Looking for the "best all time college coaches"?   See here!

Over the previous century, outstanding college coaches have pushed college sports a lot further. Skill improvement, strategies update and tactics developments, such changes contributed by coaches stimulate college sports to sparkle. It's time to take stock of the coaching landscape in the past 10 decades.

In this article, a ranking system is applied to pick out the "best all time college coaches". This is not simply a ranking of who are the best coaches in the previous century. It's a scientific ranking of who would be the best to lead a team going forward. Through the system, we find out the top 5 best coaches in basketball, football and baseball as the following figure 1.

Figure 1

The rankings proved to be a tough job. The best coaches in college should be outstanding in all aspects: recruiting, teaching, tactics design and players motivation.

We are eager to find out the well-rounded coaches, who can wisely handle all. The following in figure 2 evaluation factors are weighed in our system.

Figure 2

So how are these factors weighed in the ranking system?

Firstly, we create an "idealistic coach" whose factors are the best among all ranked coaches in basketball, virtually, of course. Secondly, compare each coach with the "idealistic coach" by means of Cosine Similarity. In addition, we also keep an eye on gender differences and generation gap by means of Data Whitening. And eventually, we promote the basketball system to football and baseball across countries.

Why is this system accessible?

Genders, generations, countries and sports, such considerations complete the system. In advance, the ranking result makes our system convincing.

And why is this system reliable?

We tested the system's stability and sensitivity. The results were stable with slight factor fluctuation. Besides, a large amount of data is used to establish and test the system.

Therefore, we have strong confidence to carry out the ranking system.

What does the ranking result show?

We have to admit that coaching is one of the driving forces in building a national championship team. The job of a coach is multi-faceted and challenging. Coaches picked out in our system are not only excellent among their peers, but leave impression on the career for decades as well. The ranking system is capable of evaluating coaches among organizations across sports.

Table of Content

1. Introduction

Restatement of the problem

College Coaching Legends

Sports Illustrated, a magazine for sports enthusiasts, is looking for the "best all time college coach" male or female for the previous century. Build a mathematical model to choose the best college coach or coaches (past or present) from among either male or female coaches in such sports as college hockey or field hockey, football, baseball or softball, basketball, or soccer. Does it make a difference which time line horizon that you use in your analysis, i.e. , does coaching in 1913 differ from coaching in 2013? Clearly articulate your metrics for assessment. Discuss how your model can be applied in general across both genders and all possible sports. Present your model's top 5 coaches in each of 3 different sports.

In addition to the MCM format and requirements, prepare a 1-2 page article for Sports Illustrated that explains your results and includes a non-technical explanation of your mathematical model that sports fans will understand.

Background

The "Sports Illustrated" is a magazine for sports enthusiasts, as the problem above demonstrates. Therefore, after a brief survey of the formal rankings in this magazine, we sort out the following basic idea related to college sports coaches ranking, combined with data resources and other researches.

The National Collegiate Athletic Association (NCAA) is a nonprofit association of 1,281 institutions, conferences, organizations, and individuals that organizes the athletic programs of many colleges and universities in the United States and Canada. Based on the game statistics from this association and its similar conferences, we found that hockey, football, baseball or softball, basketball, and soccer teams are supposed to attend national conferences or games annually. The performance of a team, which is bond with its head coach, is the core element determining the result in a game, win or loss. Therefore, a coach's quality can be evaluated by the team's performance. In advance, the assessed coach is possible to have coached more than one team. Hence, the overall team-directing experience of a coach is considered, including the number of the games the coach has directed, wins and losses all counted, the winning rate, significant awards the coach has won and the his or her career length. Through this process, the fact of gender, time line horizon and varied sports can make a difference to the ranking result.

In order to solve this problem, we collected a large amount of statistics. Considered the purpose is to build a coach judging model as reasonable as possible and present our model's top 5 coaches in each 3 different sports, we applied all the statistics to test the model and ranking the coaches whose statistics have been pre-processed. Hereby, there is guarantee in accuracy and efficiency for the model.

2. Assumptions

Table 1.

General assumption

Assumptions	Justification of assumptions
The factors used in this model are qualified to represent a comprehensive ability of college sports coaches.	There are many ways of valuing a coach. For simplifying the mathematical model, we picked some of the quantifiable and non-quantifiable factors. Factors we haven't picked out can be explained by the factors we used in this model.
The data used in this paper are real and effective and the coach ability can be well-judged by the data.	The data source in this paper is authorized statistics website. College coach's ability can be judged by the data.
The relation between difficulty coefficient and time ling horizon is positive linear.	We adjust the difficulty coefficient according to the difficulty caused by different generations. After analysis, we find that the degree of difficulty varies with generation shows an increasing trend.
The coaching difficulty and social influence are the only factors affected by time line horizon.	Other factors caused by generation are considered to be explained by the two modification coefficient.

3. Factors Analysis Model Based on Principal Component Analysis

3.1. Model overview

In this topic, we found that there are 12 factors affecting the ability evaluation of coaches after data collecting [1]. In order to simplify the data processing, we used Principal Component Analysis (PCA) to find the core components.

The results provide a three-component factor assemblage. The three components each can represent an inner ability of the coach.

3.2. Justification of our approach

PCA, is a statistical procedure concerned with elucidating the covariance structure of a set of variables. In particular it allows us to identify the principal directions in which the data varies.

In this project, firstly, we make a survey on 60 coaches(30 males 30 females), who are from NCCAB (National Committee Association America Basketball). The game of basketball in American college is used as an example. Then we collected the data including tenure, the number of the games has been directed, experiences in NCCA, the number of the games wins and losses all counted respectively, the number of the teams have been directed, the winning rate, the times of final four in NCCA and the times of championships in NCCA, the number of championships in other games such as tournaments and regular seasons, social influence which mainly accounted by the amount of searches on Google and the number of significant awards have been won. In this process, the fact of gender, time line horizon is out of consideration.

Therefore, the factors mentioned above are listed as follows.

1. **Tenure:** the term during which the coach conducts a team
2. **Games:** directed: the number of the games have been directed
3. **Tenure in:** NCAA: tenure when the coach attended NCAA
4. **Games won:** the number of the games have been won
5. **Games lost:** the number of the games have been lost
6. **Teams directed:** the number of the teams have been directed
7. **Winning rate:** the percentage of the games has been won in al games directed
8. **Final four in NCAA:** the number of the times when the coach's team was in final 4 in NCAA
9. **Championship in NCCA:** the number of the times when the coach's team won the championship
10. **Championship in other games:** the number of the times when the coach's team won the championship in other games
11. **Social influence:** the amount of searches on Google
12. **Significant Awards:** the number of the significant awards have been won.

The factors are justified because:
- With the development of sports, coaches can be evaluated in increasing factors. But elder coaches are not included in some of the new factors. In order to be fair, we chose factors that are shared by coaches in different generations.
- There are also many non-quantifiable factors that can reflect the coach's ability. For a more comprehensive measure of evaluating the coaches, we quantify some non-quantifiable factors and take them into consideration. For example, as for the factor of social influence, we use searches on Google to represent it.

- Other minor factors out of the consideration can be covered by the existing factors, such as the factor of salary. In this way, we avoid the factor repetition and hereby simplify the calculation.

3.3. Model formulation

$X_1 \sim X_{12}$ are used to represent the 12 factors above. Hence, we get an 60×12 matrix. Considering the large sorts of the factors, we will apply dimension reduction with the PCA process.

Through PCA process, we analyze the standardized factors data and build a correlation matrix. Through the correlation matrix's eigenvalue and eigenvector, we figure out inner relation across several factors in the origin data sample. The factors whose inner relations with each other are strong are classified as a new component, named principle component. $Z_i \quad (i=1,2,...,p) \quad (p=12)$

$Z_1, Z_2, ..., Z_m \quad (m \le p)$ are used to represent the inner connection among some of the factors above. It can be demonstrated as the following equation,

$$\begin{cases} Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ Z_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p \end{cases}$$

where, $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \quad (1 \le i \le m)$ and $Z_1, Z_2, ..., Z_m \quad (m \le p)$ is the first, second,...principle component.

In order to eliminate the downturn impact caused by the dimension and the order of magnitude according to the statistics collected. We process the data with the approach of z-score standardization, based on the formula as follows,

$$X_i' = \frac{X_i - E_i}{S_i} \quad (1 \le i \le p)$$

where E_i is the mean value of the factors of group i,

S_i is the variance value of the factors of group i.

Then we can get the standardized data matrix., and we can work on the correlation coefficient through the standardized data. Accordingly, we get the correlation matrix R as follows,

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & \ddots & & \\ \vdots & & \ddots & \\ r_{p1} & & & r_{pp} \end{bmatrix}$$

where,

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{i=1}^{n}(x_{kj} - \bar{x}_j)^2}} \quad (i, j = 1, 2, 3..., p)$$

Each eigenvalue $\lambda_1, \lambda_2 \ldots \lambda_p (\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0)$, is with an eigenvector $e_i \, (1 \leq i \leq p)$, where λ_i refers to the eigenvalue while $e_i \, (1 \leq i \leq p)$ refers to the eigenvector. Each eigenvalue λ_i of the correlation matrix and its eigenvector e_i is with a possible principle component Z_i.

The eigenvalue of the correlation matrix satisfies the equation,

$$\det(\lambda_i I - R) = 0$$

where I is an identity matrix.

Then we can get the contribution rate V_i of the eigenvalues, based on the formula as follows,

$$V_i = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} \left(i = 1, 2, 3 \ldots, p \right)$$

In advance, the cumulative contribution rate V_m is,

$$V_m = \frac{\sum_{i=1}^{i} \lambda_i}{\sum_{k=1}^{p} \lambda_k} \left(i = 1, 2, 3 \ldots, p \right)$$

In general, the variable correspondence to the eigenvalues whose cumulative contribution rate V_m is from 95% to 75% is regarded as the first, second,.. principle component respectively. Based on the cumulative contribution rate of correlation matrix's eigenvalue, through SPSS17.0, we get the principle components Z_1 Z_2 and Z_3.The results is shown in **Table 2.**

Table 2.

Result of PCA process

Component	% of Variance	Cumulative %
1	**52.395**	**52.395**
2	**16.050**	**68.445**
3	**8.521**	**76.966**
4	6.323	83.289
5	4.898	88.187
6	4.416	92.602
7	3.788	96.390
8	1.742	98.132
9	1.098	99.230
10	0.474	99.705
11	0.295	100.000
12	0.13	100.000

The Component Matrix of Z_1, Z_2 and Z_3 is shown as follow.

Table3.

Component Matrix

Factors	Z_1	Z_2	Z_3
Tenure	.928	.271	.032
NCAA Tenure	.842	.100	-.151
Wins	.971	.142	-.003
Losses	.711	.598	.055
Games	.948	.276	.013
Winning Rate	.631	-.324	-.001
Teams	.255	.716	-.017
NCAA Final Fours	.791	-.473	.012
NCAA Champions	.574	-.582	-.031
Conference Champion	.757	-.271	.105
Googles	.042	-.037	.987
awards	.613	-.366	-.095

3.4. Result analysis

It is indicated that,

- The first Principal Component Z_1 is relatively bond with the factors of experiences in career, experiences in NCCA, the number of the games wins and losses, the winning rate, the times of final four in NCCA and the times of championships in NCCA, and the number of significant awards have been won. Consequently, Z_1 can be described as a coach-ability variable.
- The second Principal Component Z_2 is relatively bond with the factors of the number of the teams have been directed, and the times of championships in NCCA. Hence, Z_2 can be described as a coach-influence-college variable.
- The third Principal Component Z_3 is relatively bond with the factors of social influence which mainly accounted by the amount of searches on Google. Therefore, Z_3 can be described as a coach-influence-society variable.

The degree of influence count down is Z_1, Z_2 and Z_3. Based on the analysis of the eigenvalue, the eigenvalue of Z_1 is the biggest, with Z_2 and Z_3 following it. Therefore, we believe that the factors belonged to Z_1, which is the "coach-ability" variable contributes most to the "coach evaluation system". It is shown as follows in **Fig. 1**
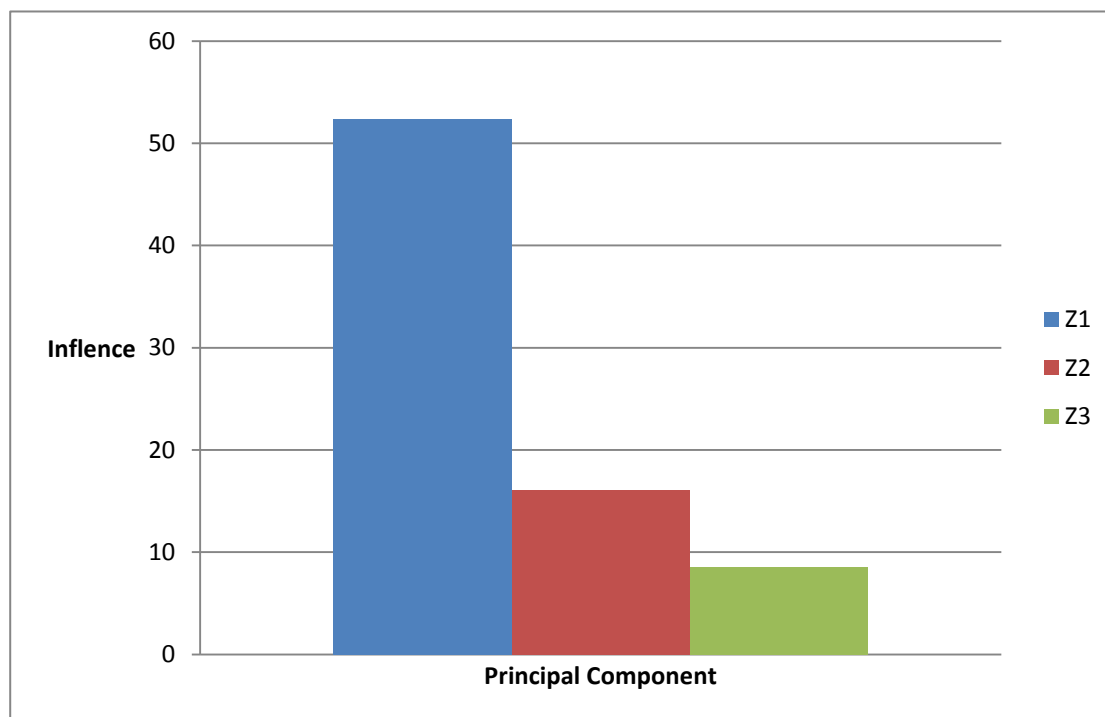
Fig 1. The Influence of Principal Components

4. Coaches Ranking Model Based on Cosine Similarity

4.1. Model overview

In this topic,   in order to select the best college coach or coaches (past or present) from among either male or female coaches in such sports as college hockey or field hockey, football, baseball or softball, basketball, or soccer, we take the basketball game is used as example. We establish a standard——"idealistic coach", which is a virtual coach whose factors data are the best of all. Then, based on the idea of vector space mapping, we apply Cosine Similarity and its adjustment to compare all coaches with the "idealistic coach". Coaches with higher similarity to the   "idealistic coach" are expected to take a better seat in the ranking list. When it comes to the computation of the model, we use Euclidean distance-based clustering analysis to process the data. In this way, we obtain the college coaches ranking result in the field of basketball.

4.2. Model justification

In this model, Cosine Similarity is used to compare the assessed coach's every factors to the all twelve factors set as the best.

The measure of similarity is to calculate the degree of similarity between individuals. Cosine Similarity is an approach to measure the gap between two identities. Firstly, we map the data of each individual to the vector space. Then we

measure the similarity between them by measuring the cosine value of the angle between the two individuals with dot product in vectors space [6]. The more it is close to 1, the more the vector angle is close to 0 degree, which means more similar the two vectors are. The more it is close to 0, the more the vector angle is close to 90 degree, which means less similar the two vectors are.
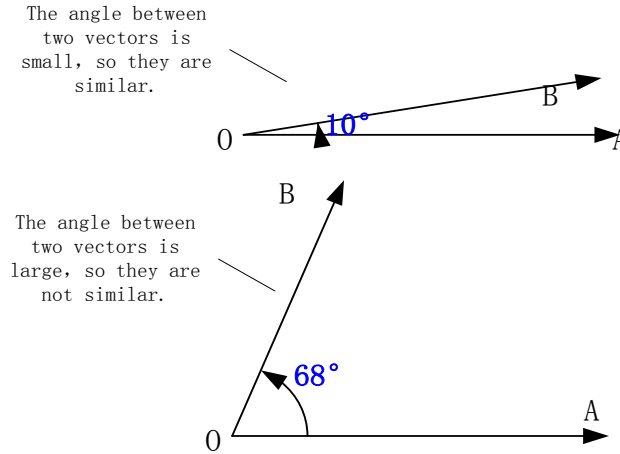
Fig 2. Explanation of Cosine Similarity

In this project, we rank all the coaches by calculating the similarity between the vector of each coach and the set vector of the virtual "idealistic coach". The higher the similarity is, the closer to the top seat in the ranking.

The cosine value between two vectors can be easily obtained through Euclidean dot product equation,

$$\vec{a} \bullet \vec{b} = |a| \times |b| \cos\theta$$

where $\cos\theta$ is the similarity between the two vectors.

Then we obtain the formula of individual similarity as,

$$\cos\theta = \frac{X_i \cdot X_s}{|X_i||X_s|} = \frac{(x_{i1}, x_{i2}...x_{in}) \cdot (x_{s1}, x_{s2}...x_{sn})}{\sqrt{\sum_{j=1}^{n}(x_{ij})^2} \times \sqrt{\sum_{j=1}^{n}(x_{sj})^2}}$$

4.3. Model formulation

4.3.1. Vector definition of the "idealistic coach" and ranking

In this topic, there are 12 factors that make impacts on the ranking of coaches, after data analysis. Accordingly, standardized vector's dimension is eleven.

Considering the large amount of the data, we classified all the coaches firstly. Then we ranked the coaches from the first group, who are the best coaches of all groups. In this case, we used Euclidean distance-based clustering analysis to classification.

We mapped the collected factors of 60 male and female coaches to the vector space. Then we obtained the coach's ability factor vector V_i $(1 \le i \le 60)$. V_i is a 1×12 row vector. And each vector component is the value of the factors.

$$V_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, x_{i10}, x_{i11}, x_{i12})$$

Firstly, we defined a standard vector to combine the factors with cosine similarity so as to ranking the coaches. We defined the vector in reality as a coach whose achievement is beyond all ever, which is the "idealistic coach". The data of his factors are picked from the highest ones. **Table 4.**

Table 4.
Data of 'idealistic coach'

T	NC T	W	L	G	W R	T	NFF	NCC	CC	G	A
23.1	16.3	1204	133	1337	0.9	2	13	10	41	18.1k	11

We compared each individual vector of all coaches to the standard vector, and then we calculate the similarity between them by cosine similarity. The bigger the cosine value is, the higher the similarity is, and hereby the better the coach will be. In the ranking list, the coach is closer to the top seat. The example is as follows.
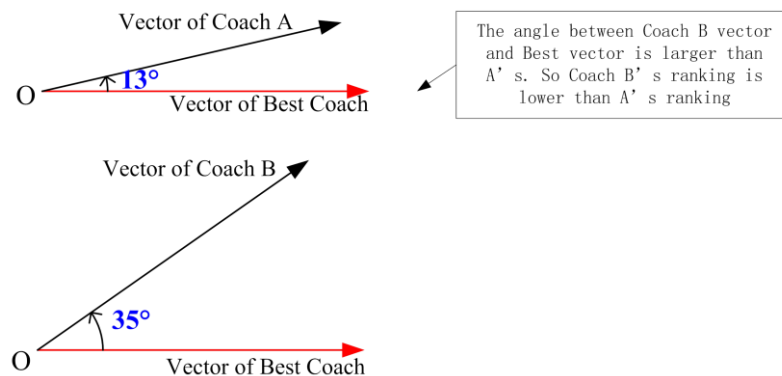
Fig 3. Comparison between coaches and the "idealistic coach"

4.3.2. Adjusted Cosine Similarity

Cosine Similarity is sensitive to direction instead of absolute value.[8] Value's differences in each dimension, therefore, are not accessible to measure. Cosine Similarity's insensitivity leads the result to be less accurate. [10]For example, we suppose two of a coach's factors are the highest in all, $x_1 = (10,5)$. While the two of another coach's factors are the lowest, $x_2 = (2,1)$. It is obvious that there is a gap between the two coaches' ability. But the angle between the two vectors is 0 degree so $\cos\theta$ is 1. It is illustrated in **Fig 4.**
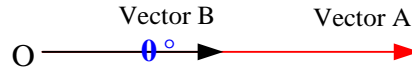
Fig 4. Situation when two vectors overlap

We hereby normalize all of the values to amend the unreasonable results.

The amendment equation is as follows,

$$x'_{ij} = \frac{x_{ij} - E_j}{X_{j\max} - X_{j\min}} \quad (1 \le i \le 60, 1 \le j \le 12)$$

where x_{ij} represents the factor j of the coach i, x'_{ij} is the standardized factor, $X_{j\min}$ represents the minimum of factor j, $X_{j\max}$ represents the maximum of factor j, E_j is the mean value of the data sample j.

We apply the adjustment in the example. Then we find that the angle between the vector $x'_1 = (0.5, 0.5)$ and the vector $x'_2 = (-0.5, -0.5)$ is 180 degree. In this case, the difference between the two coaches' ability is well presented. In fact, the example we used here is a special situation. Generally, the angle ranges from 0 degree to 180 degree and the cosine value ranges from -1 to 1, in which, 1 represents that the ability of the two coaches' are practically same, while -1 represents that the gap between the ability of the two coaches' are biggest.

4.3.3. Model computing by Euclidean distance-based clustering analysis

Cluster analysis is mainly used to classify factors or variables. Generally speaking, there are two procedures, data standardization and clustering.

Data standardization: Data standardization is used to normalize the statistics, in case of the statistics' dimension gap and value differences making unreasonable impacts. We suppose that there are n specimens, and each specimen has m factors, then each variable can be expressed as x_{ij}. The mean value is,

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

The standard deviation is,

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_{ij}\right)^2}$$

Finally, the standardization form is,

$$x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{s_j} \quad \left(s_j \ne 0\right)$$

Clustering: Based on the distance between specimens and sorts, we classify a larger sort of given sorts. At the beginning, each specimen is regarded as a single sort and the distances between sorts are equal. We choose two specimens whose distances are the smallest of all. Next we calculate the distance between the new sort and other

sorts. Then we merge the nearest two together. Continue this process until there is only one sort left. During this process, distance refers to the measurement used to estimate close degree among specimens. It is because each specimen and factor can form a matrix in the vector space. We suppose x_{ij} as the factor j, d_{ij} as the distance between specimen i and other specimen. There are several ways of calculating the distance between sorts, such as, single linkage, the longest distance method, group average method, and centroid method. In this problem, we apply single linkage to clustering analysis.

Single linkage can be described as follows.

First of all, we need to know about the distance called Minkowski. The most common method to calculate the distance is Minkowski

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^q \right]^{1/q}$$

When $q = 2$

d_{ij} is called Euclidean distance.

$$d_{ij}(2) = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2}$$

We suppose G_p、G_q、G_r as three respective sort, then we get the formula as follows,

$D_k(p,q) = \min\{d_{jl} \mid j \in G_p, l \in G_q\}$

Then we merge the sort G_p with sort G_q as sort G_r. So the distance formula is,

$$D_{kr}^2 = \min_{X_i \in G_k, X_j \in G_r} d_{ij} = \min\{ \min_{X_i \in G_k, X_j \in G_p} d_{ij}, \min_{X_i \in G_k, X_j \in G_q} d_{ij} \} = \min\{D_{kp}, D_{kq}\}$$

Fig 5 shows the demonstration of Cluster Analysis.
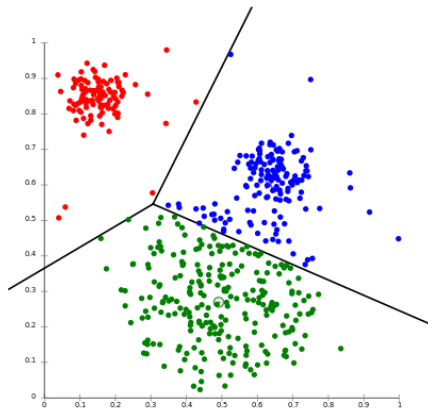
Fig 5. Demonstration of Cluster Analysis.[21]

Continuing the process, we classify the 60 coaches into nine levels, of which, there are 24 coaches in the first five levels. Finally, we get the results of the clustering analysis is shown in **Table 5.**

Table 5.

Result of clustering analysis

Name	Cluster	Name	Cluster	Name	Cluster
1:John Wooden	1	6:John Calipari	3	8:Tom Izzo	4
2:Dean Smith	2	7:Denny Crum	3	9:Billy Donovan	4
4:Mike Krzyzewski	2	12:Bill Self	3	19:Jack Gardner	4
5:Adolph Rupp	2	13:Rick Pitino	3	22:Thad Matta	4
11:Jim Calhoun	2	15:Jerry Tarkanian	3	23:Sean Miller	4
17:Jim Boeheim	2	16:Lute Olson	3	25:Jay Wright	4
3:Roy Williams	3	32:Tara VanDerveer	3	27:Anthony Grant	4

Name	Cluster	Name	Cluster	Name	Cluster
28:Bo Ryan	4	41:Brenda Fres	4	50:Sharon Versyp	4
29:Jim Valvano	4	42:Tina Martin	4	51:Pam Borton	4
30:Jamie Dixon	4	43:Jamelle Renee Elliott	4	52:Joanne Boyle	4
33:Kimberly Duane Mulkey	4	44:Deb Patterson	4	53:Nikki Caldwell	4
35:JoanneMcCallie	4	46:Sue Semrau	4	55:Kellie Harper	4
36:Sherri Coale	4	47:Suzy Merchant	4	56:Kim Barnes Arico	4
39:Jennifer Rizzotti	4	48:MaChelle Joseph	4	57:Melissa McFerrin	4

4.4. Results analysis

Table 6.

Ranking result of cosine similarity model

Ranking	Name.	Gender
1	John Wooden	M
2	Geno Auriemma	F(M)
3	Pat Summitt	F
4	Dean Smith	M
5	Mike Krzyzewski	M
6	Roy Williams	M
7	John Calipari	M
8	Bill Self	M
9	Adolph Rupp	M
10	John Thompson	M

From the results above, we obtain result analysis as follows,

- Analyzing the top few of the factors in the ranking list, we find that each basic factor has played a role in the ranking. Therefore, the pick of these rankings is reasonable.
- There are differences in some factors across different time line horizon, such as the searches on Google, winning rate, the games won by the coach and so on. Therefore, it should be a modification considering the impact of timeline horizon.
- There are differences in some factors across both genders.

5. Model Optimization of Time Line Horizon Based on Regression Analysis

5.1. Justification of our generation-based model optimization

With the gradual development of NCAA, the number of teams attending the NCAA basketball game is changing. Accordingly, the difficulty and pressure the coach faces is changing. Besides, with the development of science and technology, the social influence of basketball is affected by internet transmission. Therefore, model optimization is supposed to consist of the coaching difficulty optimization and the social influence optimization.

## 5.2. Generation-based model optimization

5.2.1. Giant effect

When considering the difficulty coefficient brought by time line horizon, we supposed that the difficulty is positively related to time line horizon. So we set the

coefficient as a value between 1 and 1.5. However, younger coaches can learn from the elder ones. With such endowed advantage, younger coaches are expected a relatively better performance. Still, we set the difficulty coefficient below 1.5. So we consider it as a minor influence instead of paying too much attention.

5.2.2. Coaching difficulty optimization

After analysis, we believe that there is a significant impact on the factor of winning rate caused by different coaching time line horizon. In early times, the teams that participated NCAA games are less, and hereby, they got a greater chance of winning, compared with those today. However, with the progressive development of NCAA games, teams participate in are much more. As a result, the pressure is higher to not only players but also coaches and the opportunity to win is harder to earn.

We carried on the statistics of the active coaches in NCAA from 1942 to 2013. [3].Here is the approximate linear relationship between the number of active coaches and the time line horizon in **Fig 6.** based on regression analysis.
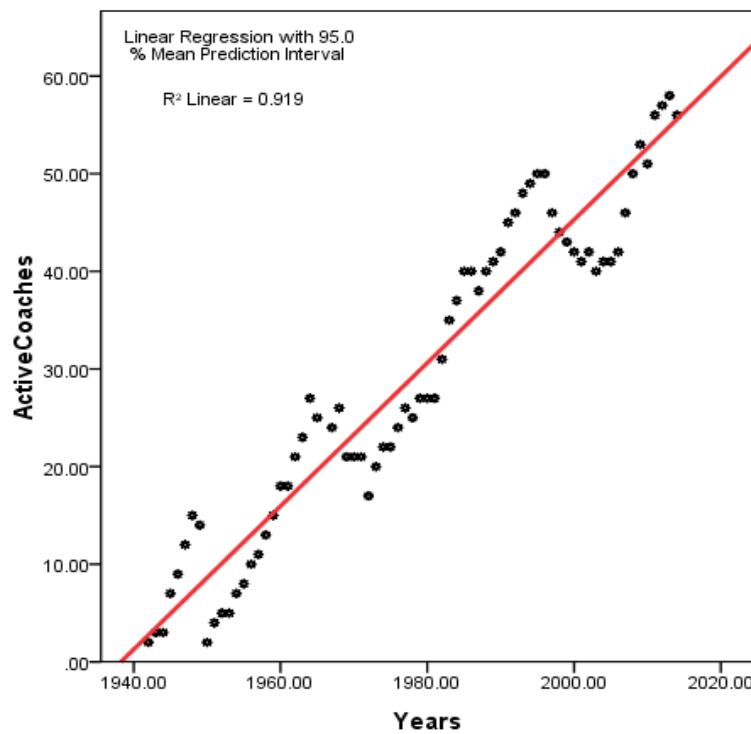
Fig 6. Approximate linear relationship between the number of active coaches and the time line horizon

Based on the result of regression analysis, we found that the number of active coaches and the time line horizon satisfy the equation,

$$A(t) = 0.734t - 1421.8$$

We believe that the difficulty coefficient of a match is directly proportional to the number of active coaches each year. Hence, it can be expressed as follows,

$$D(t) = k \times A(t)$$

Based on the formal equations, we get the relation between the difficulty coefficient of a match and time line horizon as follows,

$$D(t) = \frac{0.734t - 1421.8}{k}$$

where we suppose $k = 1$, therefore,

$$D(t) = 0.734t - 1421.8$$

It is shown in Fig 7.
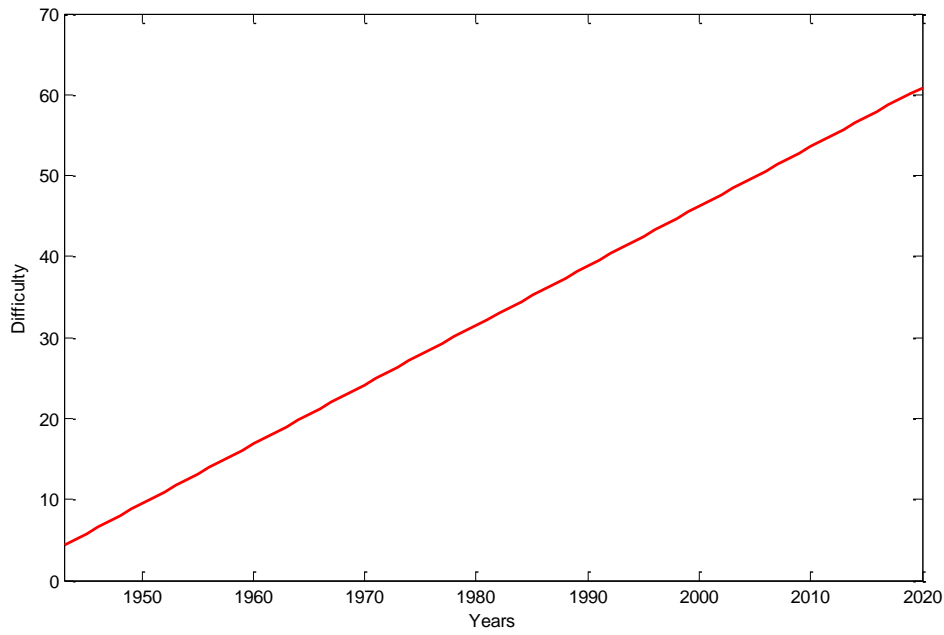
Fig 7. Relation between Coaching difficulty and Generation

Considering the fact that there are coaches whose career life is over a relatively long time period, we looked up the year when the coach started coaching and the year when the coach stopped coaching, instead of measuring the comprehensive difficulty coefficient by a single year. [5]We define C_i $(1 \le i \le 60)$ as the comprehensive difficulty coefficient in coach i's whole career life. The formula is as follows,

$$C_i = \frac{1}{T_i^{end} - T_i^{start}} \int_{T_i^{start}}^{T_i^{end}} D(t)\, dt \quad (1 \le i \le 60)$$

Standardizing C_i with Max-min method ,we can obtain C_{si}

$$C_{si} = \frac{C_i - C_{min}}{C_{max} - C_{min}} \quad (1 \le i \le 60)$$

Because of the 'Effect of Giants', we let all the C_{si} values of which is above 0.5 be 0.5. Hence, C' is as follows,

$$C_i' = \begin{cases} 1.5 & C_{si} \geq 0.5 \\ C+1 & C_{si} < 0.5 \end{cases} \quad (1 \leq i \leq 60)$$

Fig 8. Adjustment of Difficulty Coefficient

5.2.3. Social influence optimization

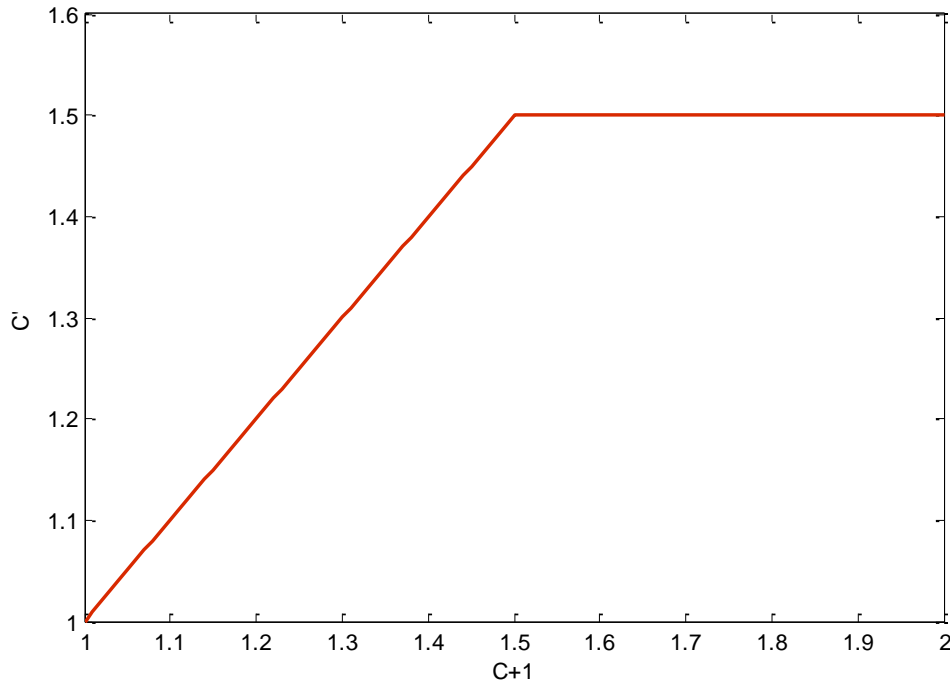With the popularization of sports events and development of internet technology, the social influence of basketball is rising. Hereby, college coaches' visibility is also rising. It is illustrated by the amount of searches on Google. Therefore, there is a significant impact on the amount of searches results on Google in different time line horizon. A reasonable optimization coefficient is needed to minimize this impact.

It is obvious that coaches in earlier years are less possible to obtain an idealistic search result. So the start coaching year is positively relation to the search results. In order to weaken the positive relation, we let the search results divided by a coefficient which is above 1. That is to say, we add 1 to the standardized result. We hereby get statistics in the section of [1,2]. The formula is as follows,

$$G = \frac{s - s_{min}}{s_{max} - s_{min}} + 1$$

Where s is the start coaching year for each coach,

G is the social influence optimization coefficient.

5.3. Generation-based model Optimization results

Based on the formal analysis, we multiply the optimization coefficient C', and divide the amount of searches results on Google by G. Then we continue to model calculate through cosine similarity as illustrated before.

5.4. Generation-based results analysis

The final result is in **Table 7**.

Table 7.

Final ranking result

Ranking	Name.	Gender
1	John Wooden	M
2	Geno Auriemma	F(M)
3	Pat Summitt	F
4	Dean Smith	M
5	Mike Krzyzewski	M
6	Roy Williams	M
7	John Calipari	M
8	Bill Self	M
9	Adolph Rupp	M
10	Billy Donovan	M

Table 8.

Camparson of the 2 ranking result

	before time optimized	After time optimized
Ranking	Name.	Gender
1	John Wooden	John Wooden
2	Geno Auriemma	Geno Auriemma
3	Pat Summitt	Pat Summitt
4	Dean Smith	Dean Smith
5	Mike Krzyzewski	Mike Krzyzewski
6	Roy Williams	Roy Williams
7	John Calipari	John Calipari
8	Bill Self	Bill Self
9	Adolph Rupp	Adolph Rupp
10	**John Thompson**	**Billy Donovan**

Therefore, we analyzed the result as follows,

● When considering the difficulty coefficient, we put the difficulty of winning with increasing numbers of teams year by year as a core indicator. However, Young coaches can usually learn from the reference according to previous experience. The foundation laid by the old coach cannot be ignored. Therefore, the difficulty coefficient should not be too high. Eventually, we set its maximum as 1.5.

- It is obvious that, after comparison the ranking of before and after the optimization, the overall ranking hasn't changes tremendously. Only those whose tenure has some obvious features are adjusted slimly in the ranking.
- After comprehensive analysis of the factors data adjusted of the ranked coaches, we can see that the modified ranking result is more reasonable. Therefore, the optimization model is relatively correct.

6. Model Optimization of Gender Based on Data Whitening

6.1. Justification of our gender-based model optimization

By the optimization model as above, we considered the time line horizon and achieved the modifying of the origin model. Still, the gender difference has a significant impact on the coach ranking result. Therefore, we believe that there is significant difference among some factors, affected by the intense degree and competition pressure in matches across both gender.

In order to figure out which factor will be affected by the factor of gender, we apply analysis of variance (ANOVA) to test whether or not there is a significant difference of factors modified by time line horizon across both genders. Because the gender factor is the single consideration, we apply one way analysis of variance (ANOVA) in this problem.[23]

x_i is supposed as one single factor of coaches, and $\overline{x_i}$ is supposed as the mean value of the factor. N refers to the number of the coach kinds, which is 2 considering the gender factor. Therefore,

Factor：

$$SSA = \sum_i n_i \overline{y_i}^2 - N \overline{y_i}^2$$

Error：

$$SSE = \sum_i \sum_k y_{i,k}^2 - \sum_i n_i \overline{y_i}^2$$

Mean squares：

$$MSSA = \frac{SSA}{I-1}, \quad MSSE = \frac{SSE}{N-I}$$

F value:

$$F = \frac{MSSA}{MSSE}$$

Probability $p = P\left(F_{I-1,N-I} > c\right)$

In this problem, we use hypothesis tests in the analysis of variance, that is , to propose H_0 assuming that all the average observed measures are the average observed measures are the same. If the value of the probability $p < \alpha$, with $\alpha = 0.05$

the selected confidence level, the hypothesis H_0 should be rejected. Otherwise, the hypothesis cannot be rejected.

We input the two groups of data belong to male and female coaches respectively in SPSS, and the result is illustrated in **Table 9.**

Table 9.

Result of ANOVA

Factors	NC T	W	L	G	W R	T	NFF	NCC	CC	G
Sig. （p）	0.002	0.381	0.004	0.013	0.006	0.002	0.008	0.301	0.000	0.108

From the given table, for the factor of Los, $p > \alpha$, we hereby accept the hypothesis. That is to say, this factor makes less difference across both genders. However, for the factors of, NCAAExp(NCT), Win（W）, WinRate,(WR) $p < \alpha$, so we rejected the hypothesis ,which is, the factors mentioned have significant difference across both genders.

Hence,

- The factor of gender makes difference on majority of the factors. A modifying approach is needed.
- Not all of factors across both genders are affected. Consequently, our purpose of optimizing the model can be completed by modifying the factors influenced.

6.2. Gender-based model optimization

In this topic, we apply the approach of Data Whitening [23]to eliminate the impact caused by the gender difference. The factor of games won is used to demonstrate how the approach is applied.

According to the results from ANOVA as above, we know that the factors of games won across both genders are different. The variance of the factors of games won is demonstrated in **Table 10.**

Table 10.

Var. comparison between genders on 'win games'

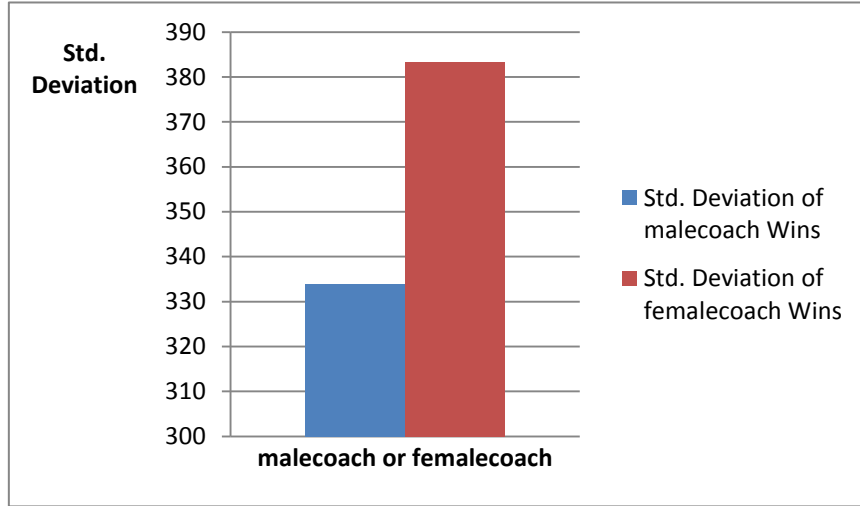| Factors	N	Std. Deviation	Variance
Games Won of male	30	333.96398	111531.941
Games Won of female	30	383.33084	146942.534

Fig 9 Variance Comparison of Wins across gender

Then we use the approach of Data Whitening to make the factor of games won consistent, in order to eliminate the impact of gender difference.

The mean value of the factor of games won by male coaches is supposed as E_{mw}, ($n = 30$)

$$E_{mw} = \frac{x_{mw1} + x_{mw2} ... + x_{mwn}}{n}$$

The mean value of the factor of games won by female coaches is supposed as E_{fw},

$$E_{fw} = \frac{x_{fw1} + x_{fw2} ... + x_{fwn}}{n}$$

The variance of the factor of games won by male coaches is supposed as $S_{mw}{}^2$

$$S_{mw}{}^2 = \frac{\sum_{i=1}^{n}\left(x_{mwi} - E_{mw}\right)^2}{n-1}$$

The variance of the factor of games won by female coaches is supposed as $S_{fw}{}^2$

$$S_{fw}{}^2 = \frac{\sum_{i=1}^{n}\left(x_{fwi} - E_{fw}\right)^2}{n-1}$$

Because the variance value of female coaches is higher, we can make the variance value of female and the variance value of male be consistent by data whitening formula. Hence, the optimization formula based on variance is as follows,

$$x'_{fwi} = E_{fw} + \frac{\left(x_{fwi} - E_{fw}\right)}{\sqrt{\dfrac{S_{fw}{}^2}{S_{mw}{}^2}}}$$

6.3. Gender-based optimization results

The optimized variance result is demonstrated in **Table 11.**

Table 11.

Var. of 'win games' after adjusted

Factors	N	Std. Deviation	Variance
Games Won of male	30	333.96398	111531.941
Games Won of female	30	333.89086	111483.103
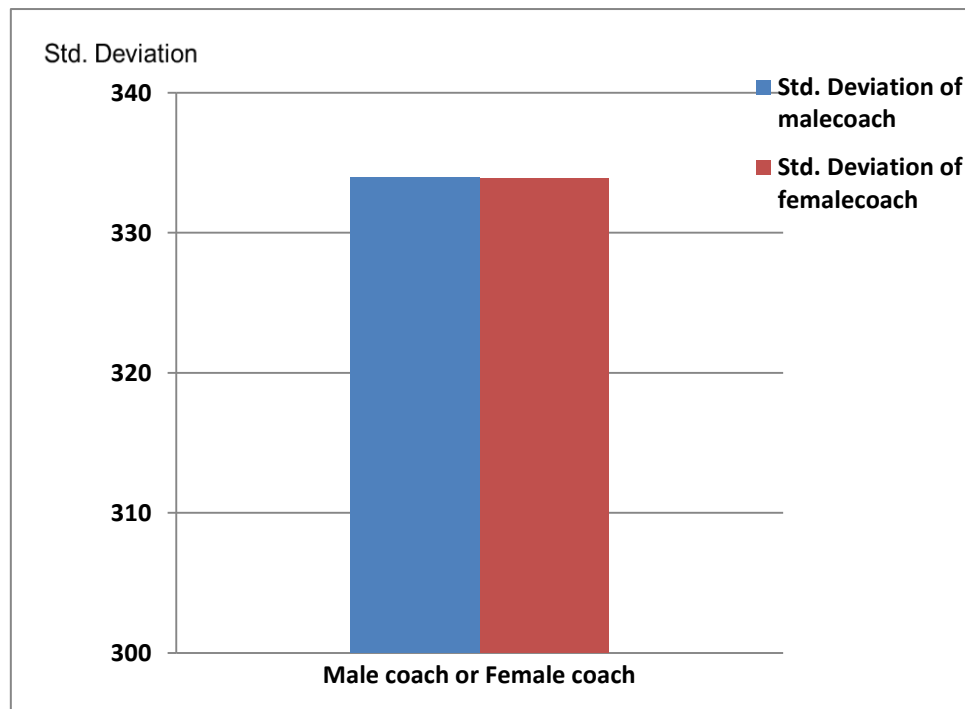
Fig 10.　Modified Variance of Wins

In this way, the influence made by the factor of gender can be eliminated. So other factors affected by the gender can be processed by data whitening in the same way. The result is in appendix.

Then we apply the cosine similarity to the results.

6.4. Gender-based results analysis

The ranking list after gender optimization and time line horizon optimization is shown in **Table 12.**

Table 12.

Ranking result after gender factor adjusted

Ranking	Name.	Gender
1	John Wooden	M
2	Geno Auriemma	F(M)
3	Pat Summitt	F
4	Dean Smith	M
5	Mike Krzyzewski	M
6	Roy Williams	M
7	Adolph Rupp	M
8	John Calipari	M
9	Tara VanDerveer	F
10	Tom Izzo	M

Based on the results, we have analysis as follows,

- It is obvious that, after comparison the ranking of before and after the optimization, the overall ranking hasn't changed tremendously. Part of the ranking of female coaches declined slightly overall.
- The influence caused by the gender factor can be eliminated at some point by data whitening.
- Although the ranking of female coaches declined slightly overall, the ranking of female coaches whose seat is in front of others haven't changed a bit. Therefore, the optimization model is relatively correct.
- We obtain the final ranking list with the consideration of gender and time line horizon.

7. Model Extension

7.1. Establishment of factors pool

The mathematical model above has provided a relatively reasonable ranking result. Hence, we can classify sports coaches according to the factors by pushing this model further to other sports. Considering that the factors in each sport can be varied, but the principle and basic model are the same, we hereby establish a factors pool, which includes all the factors across sports when ranking the college coaches.   In this way, a coach in some sport can be ranked by the factors belong to the sport. After screening out the factors, we apply the formal mathematical model to ranking the coaches.

The factors pool is demonstrated in Fig 11. The sports of basketball, football and baseball are considered.
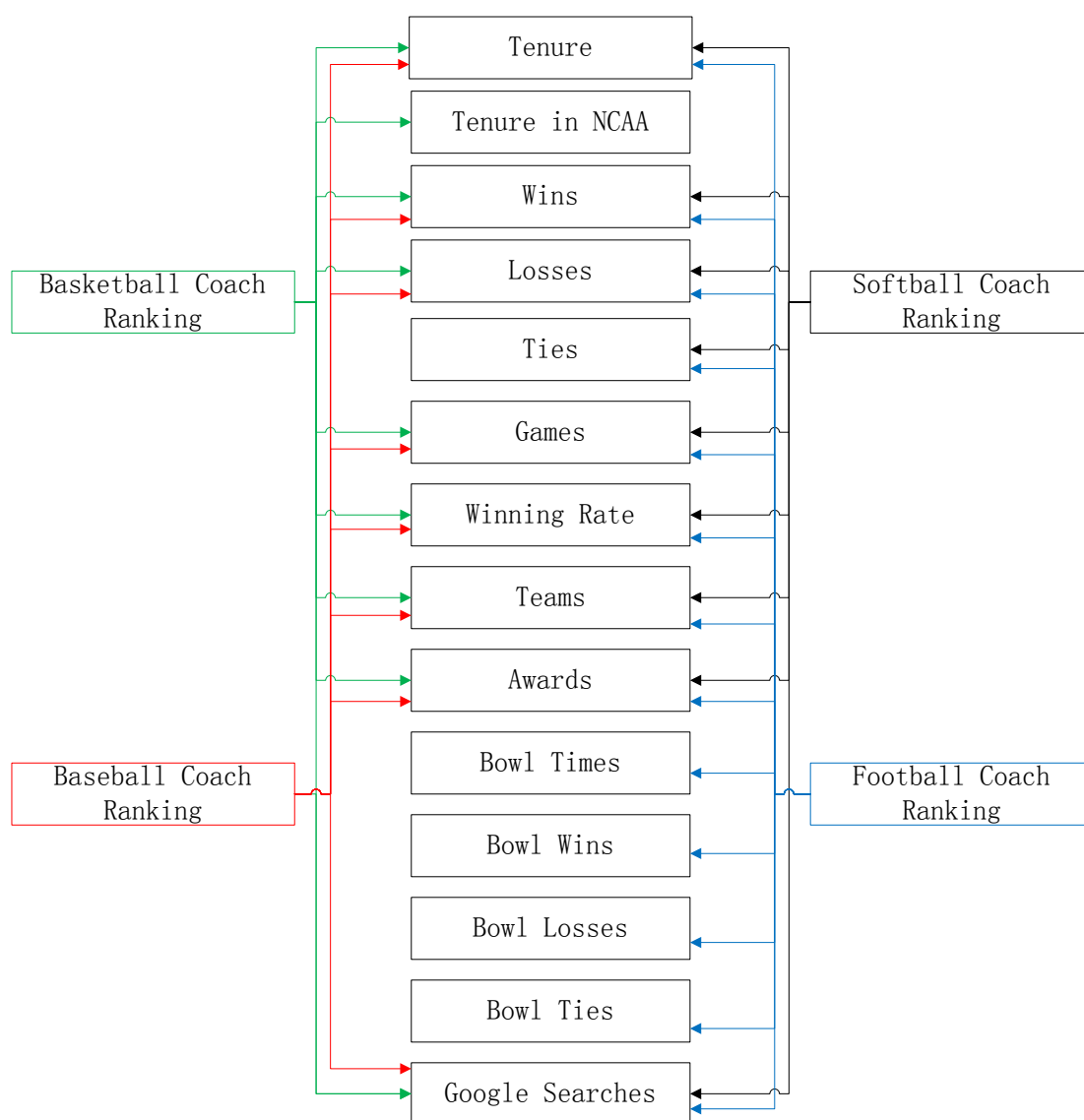
Fig 11. The demonstration of Factors Pool

Merits of establishing the factors pool:
● The mathematical model can be pushed further to other sports in this way. So the application range of the model is expanded.
● Referring to one kind of sports, we can pick out the factors belong to the sport in the pool and complete the later ranking process. The operation is simplified.
● The factors pool is a universal data base including factors of varying sports. Such feature makes the factors pool easy to be up-dated. Other sports factors can be added to the pool at any time. plus, the factors existed in the pool can be refreshed immediately.

7.2. Sports extension

From the factors pool, we notice that rankings for coaches across different sports are based on the factors. Therefore, we take the ranking for football as an example to illustrate the ranking model extension.

It is illustrated in the factors pool that there are factors belong to not only basketball but also football, such as the winning rate value, games lost, tenure and so on. There are also factors that belong to football instead of basketball such as games end up with tie, winning rate of bowls and wins, losses, ties in bowls. Likewise, there are factors that belong to basketball instead of football, such as championships in NCAA and tenure in NCAA.

Hence, we use the factors needed and their data to rank the college football coaches. Then the optimized model can be applied to the coaches ranking.

Based on all of above, the ranking result of college football coaches is illustrated in **Table 13.**

Table 13.
Football Coach Ranking

Ranking	Names
1	Bear Bryant
2	Vince Dooley
3	Joe Paterno
4	Lou Holtz
5	Darrell Royal

Likewise, the ranking result of college softball and baseball coaches is shown in **Table 14. and Table 15.**

Table 14.
Softball Coach Ranking

Ranking	Names
1	Sue Enquist
2	Sharon Backus
3	Patty Gasso
4	Sandy Jerstad
5	Judi Garman

Table 15.
Softball Coach Ranking

Ranking	Names
1	Ron Fraser
2	Ray Tanner
3	Bill Holowaty
4	Mike Martin
5	Ron Fraser

7.3. Country extension

It is demonstrated from the ranking across sports that the ranking model is universal. We hereby try to apply the model to other countries. However, there may be some difference referring to some sports in varying countries. So an update and adjustment to the formal factors pool is needed. After that, the optimized mathematical model can be used to rank college coaches in the country across sports. Through our model, we figure out the top five college basketball coaches in China is shown in **Table 16.**

Table 16.
Chinese Coach Ranking

Ranking	Names
1	Baoqiang Sun
2	Changshan Li
3	Daohong Chen
4	Guangbi Xiao
5	Huaiyu Wang

7.4. Factors extension

7.4.1. Salary

A college coach's salary demonstrates his status in his expertise at some point. Salary levels of college sports coaches can therefore explain the coaches' personal abilities from outside. So salary could be a factor when considering the ranking of coaches. Still, salary level of some coach is always determined by the coach's directing performances, which include factors of tenure, wins and losses, winning rate and so on. That is to say, in the formal model, the influence caused by salary can be covered with factors considered.

7.4.2. Popularity of college

The popularity of the college the coach is in plays a role in social influence and reflection of audience, with varied degree of social concern. Hence, the investment of the college team and the environment players faced in the middle of a game may be affected. So the team's performance is inevitably affected. As a result, the factors such as wins and losses, winning rate may have a difference. Still, the impact brought by popularity of college is overweighed compared with the 12 prior factors. So we consider it as a minor influence instead of paying too much attention.
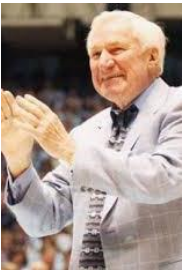
7.4.3. Formal performance of the college team

Factors used to judge coaches are based on the coaches' teams. In fact, the formal performance of a team takes a critical part in the team's later performance, which, hereby, influences the coach's data. Still, the impact brought by formal performance of the college team is overweighed compared with the 12 prior factors. So we consider it as a minor influence instead of paying too much attention.

8. Results and results analysis

8.1. Results

Table 17.
Basketball Coach Ranking

Ranking	Name	Tenure	Wins	Games	Winning Rate
1	John Wooden	29	644	806	0.8
2	Geno Auriemma	27	862	995	0.866
3	Pat Summitt	37	1098	1306	0.84
4	Dean Smith	36	879	1133	0.78

| **5** | | **39** | **975** | **1277** | **0.76** |

Mike Krzyzewski

Table 18.
Football Coach Ranking

Ranking	Name	Tenure	Wins	Games	Winning Rate
1	 Bear Bryant	**38**	**323**	**425**	**0.76**
2	 Vince Dooley	**25**	**201**	**288**	**0.697917**
3	 Joe Paterno	**46**	**409**	**548**	**0.74635**

| 4 | | 33 | 249 | 388 | 0.641753 |

Lou Holtz
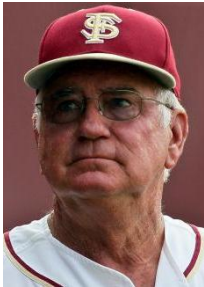
| 5 | | 23 | 184 | 249 | 0.738956 |

Darrell Royal

Table 19.
Baseball Coach Ranking

Ranking	Name	Tenure	Wins	Games	Winning Rate
1	 Don Schaly	40	1442	1784	0.76
2	 Ron Fraser	30	1271	1718	0.70

3	Ray Tanner	25	**1133**	**1625**	**0.75**
4	Bill Holowaty	45	**1404**	**1936**	**0.64**
5	Mike Martin	34	**1771**	**2386**	**0.74**

Images come from Google Image Search

8.2. What the results illustrate

We draw a map of the United States based on the performance of the basketball regionally.

Fig.12 Map of U.S.A. [22]

Indeed, sports have an influence on every aspects of society and all walks of life.

Sports effect on society: Sports, with their impact and significance, have always been a critical role in society. As for players and coaches, sports bring competition, self-confidence integrity and ambition. As for audience, sports bring the sense of excitement, amusement and challenge. The most of all, however, is the spiritual power they bring us.

Sports effect on economy: The sports industry as a whole brings roughly $14.3 billion in earnings a year — and that's not even counting the Niagara of indirect economic activity generated by Super Bowl Sunday (well-known for being the second foodiest day in the country, behind Thanksgiving). The industry also contributes 456,000 jobs with an average salary of $39,000 per job", quoted from American website "Find the best". It is obviously that profits brought by sports are hard to be ignored. The basketball distribution difference hereby leads to a gap among states in American.

Sports effect on culture: Sports players are icons for some people today. Sports are also about pride and integrity. When people share a sports idol, they become a community, and when the number of those people grows, there would be an increase on sense of unity even patriotism. Unity is never a downturn of progress of either economy or others, which is also illustrated in the map above.

9. Model Stability Analysis

9.1. Sensitivity analysis

In this paper, the basic model is established on the approach of cosine similarity. Then we optimized the model from the aspect of time line horizon and gender. The ranking list before and after the optimization is as follows (basketball as an example). It is illustrated in the result that the relatively better coaches' ranking hasn't changed tremendously. Obviously, the model is stable. The slight adjustment of the parameters change hasn't led to a dramatic change of the result. It is shown in the **Table 20.**

Table 20.

Ranking list before and after the optimization

Ranking before adjustment	Ranking before adjustment
John Wooden	John Wooden
Geno Auriemma	Geno Auriemma
Pat Summitt	Pat Summitt
Dean Smith	Dean Smith
Mike Krzyzewski	Mike Krzyzewski
Roy Williams	Roy Williams
Adolph Rupp	John Calipari
John Calipari	Bill Self
Tara VanDerveer	Adolph Rupp

Tom Izzo	John Thompson

9.2. Robustness analysis

Robustness testing is a quality assurance methodology focused on testing the robustness of software and mathematical model. Robustness testing has also been used to describe the process of verifying the correctness of the system.

By adding input noise to the first 15 coaches' factors, we obtain the changed data of the first 15 coaches. The data comparison of a coach is shown in **Table 21**.

Table 21.

Data comparison of John Wooden

T	NC T	W	L	G	W R	T	NFF	NCC	CC	G	A
29	16	759	162	921	0.8241	2	12	10	16	5752963	12
29	16	755	162	917	0.8231	2	12	10	16	5743239	11
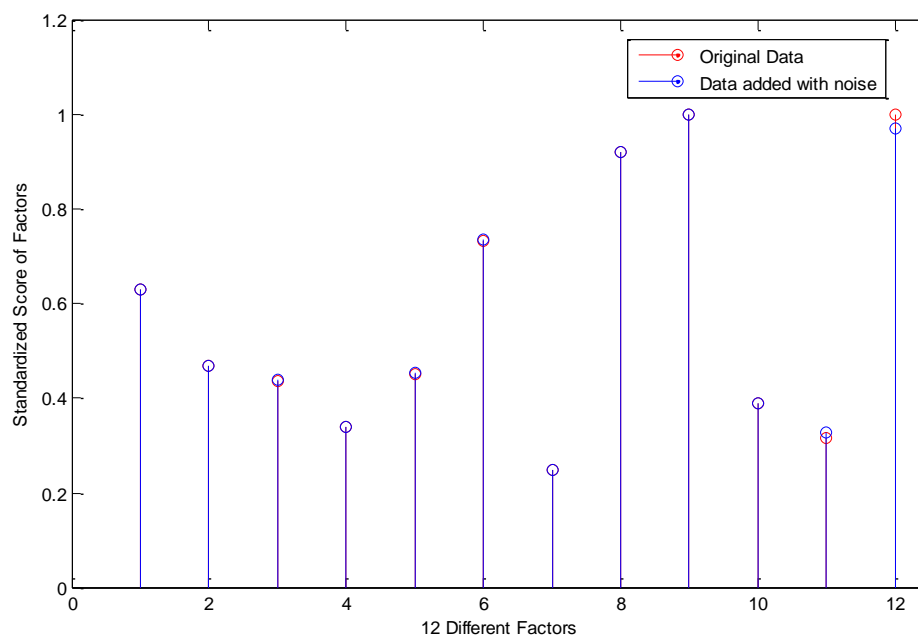
Fig. 13 Data Comparison of John Wooden before and after noise input

The new ranking is shown in **Table 22.**

Table 22.

Comparison of adding noise

Ranking list after noise input	Ranking list before noise input
John Wooden	John Wooden
Geno Auriemma	Geno Auriemma
Pat Summitt	Pat Summitt

Dean Smith	Dean Smith
John Calipari	Mike Krzyzewski
Rick Pitino	Roy Williams
Roy Williams	John Calipari
Tom Izzo	Bill Self
John Thompson	Adolph Rupp
Billy Donovan	Tom Izzo

Analyzing the ranking list of coaches after adding input noise, we find that the ranking of the first 5 coaches hasn't changed. It proves that the robustness of our model is relatively well. The noise influence is little.

9.3. Strengths and Weaknesses

9.3.1. Strengths

Objectively analyzed, the core strengths of our model are,

- **Credibility**: Coaches ranking result in this model is convincing compared to college coaches evaluation from famous advisers and media.
- **Accessibility**: Our model can be applied to genders, varied time line horizons and different countries.
- **Self-improvement:** The approaches of optimization in our model simplify the overall ranking system and push it further to other data sources in different countries among sports.
- **Reliability**: Our model is based on a large amount of data. The stability and robustness of our model are guaranteed. The accuracy and sensitivity is also presented in the results analysis section.
- **Adaptability**: The basic idea of our model can be applied to other ranking systems due to its scientific approach and easy-handled model.

9.3.2. Weaknesses

With merits above, there is still limitation in our model.

- The model can be less flexible when the evaluation factors are narrowed down to one or two. So it can be less reliable when ranking coaches on s specific factor.
- The model can be inaccurate referring to coaches whose ability is less impressing.

10. Conclusions

- The approach of cosine similarity is used for optimization in aspects of time line horizon and gender. We obtained a best coaches ranking list in basketball as an example. The best coach of basketball will be John Wooden. The model was then applied to other sports. We ranked the coaches of baseball, softball and football.

- The factor of time line horizon makes an important impact. So we analyzed the factors contributing to the difficulty gap, which is the number of active coaches in NCAA. We supposed that the difficulty in coaching is positively linear with it. The difficulty in coaching is also positively linear with time line horizon.

- The factor of gender contributes to the pressure difference across genders. Therefore, there are significant differences in factors between male and female coaches. It is also proven by the analysis of variance. In order to fix the problem, we apply the approach of data whitening to modify the data with a significant difference. After the elimination of the difference, we rank the coaches again.

- We applied the ranking model further to other sports. To achieve this purpose, we established a factors pool. When considering one specific sport, we picked factors in the factors pool and then applied the same model to rank the coaches.

- At the end, we changed several parameters and add some input noise. The final result is consistent with the formal one. Therefore, the stability and robustness is considerably well.

11. Reference

[1] http://www.usatoday.com/sports/college/salaries/

[2]2012: NCAA Men's Basketball Tournament Records of Active Coacheshttp://www.dbwoerner.com/basketball/coaches/coach112.html

[3] http://www.masseyratings.com/theory/massey97.pdf

[4]10 Greatest Coaches in NCAA Basketball History http://bleacherreport.com/articles/1341064-10-greatest-coaches-in-ncaa-basketball-history

[5] Sports-Reference.com - Sports Statistics and History http://www.sports-reference.com/

[6]Cosine similarity - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Cosine_similarity

[7]Division I (NCAA) - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Division_I_(NCAA)

[8] Adjusted Cosine Similarity   http://www10.org/cdrom/papers/519/node14.html

[9] Recommender Systems > PART I: Introduction to basic concepts > 4.3.1 Defaults - Pg. 89e

http://my.safaribooksonline.com/book/-/9780521493369/2dot2-item-based-nearest-neighbor-recommendation/221_the_cosine_similarity_meas#X2ludGVybmFsX0J2ZGVwRmxhc2hSZWFkZXI/eG1saWQ9OTc4MDUyMTQ5MzM2OS84OQ

[10] Xin J. and Bamshad M. School of Computer Science, Telecommunication and Information Systems DePaul University USING SEMANTIC SIMILARITY TO ENHANCE ITEM-BASED COLLABORATIVE FILTERING

www.csee.umbc.edu/~kolari1/Mining/papers/JM03.pdf

[11] http://www.sports.yahoo.com/news/ncca-college-basketball/

[12]http://www.articles.sun-sentinel.com/sports/

[13]http://www.Find the best.com/

[14]http://www.wiseGEEK.com/how-are-college-football-teams-ranked.htm/

[15]http://www.economicmodeling.com/sports/

[16] Matrix-based Methods for College Football Rankings Vladimir Boginski1, Sergiy Butenko and Panos M. Pardalos1 University of Florida, USA Texas A&M University, USA

[17] College Football Rankings: Do the Computers Know Best? By Joseph Martinich College of Business Administration University of Missouri - St. Louis Final Version: May 7, 2002

[18] An overview of some methods for ranking sports teams Soren P. Sorensen University of Tennessee Knoxville

[19] http://college-basketball-coaches.findthebest.com/

[20]http://bbs.hupu.com/91418.html

[21]http://www.slate.com/articles/sports/slate_labs/2013/10/united_sports_of_america_map_if_each_state_could_have_only_one_sport_what.html

[22] http://en.wikipedia.org/wiki/Cluster_analysis

[23]Dingyu Xue 2009 *Solving Applied Mathematical Problems with MATLAB* Bejing: Tsinghua

12. Appendix

Original Data

Name.	G	Exp	NExp	W	L	G	WR	T
John Wooden	M	29	16	644	162	806	0.8	2
Geno Auriemma	F(M)	27	27	862	133	995	0.866	1
Pat Summitt	F	37	31	1098	208	1306	0.84	1
Dean Smith	M	36	27	879	254	1133	0.78	1
Mike Krzyzewski	M	39	29	975	302	1277	0.76	2
Roy Williams	M	26	23	715	187	902	0.79	2
John Calipari	M	22	14	585	171	756	0.77	3
Bill Self	M	21	15	524	169	693	0.76	4
Adolph Rupp	M	41	20	876	190	1066	0.82	1
John Thompson	M	27	20	596	239	835	0.71	1
Billy Donovan	M	20	13	470	188	658	0.71	2
Tom Izzo	M	19	16	458	181	639	0.72	1
Tara VanDerveer	F	36	33	891	203	1094	0.81	3
Denny Crum	M	30	23	675	295	970	0.7	1
Bob Knight	M	42	28	899	374	1273	0.71	3
Rick Pitino	M	28	18	681	239	920	0.74	4
Lon Kruger	M	28	14	531	338	869	0.611	5
Jim Calhoun	M	40	23	877	382	1259	0.7	2
Jerry Tarkanian	M	30	18	761	202	963	0.79	3
Lute Olson	M	34	28	776	285	1061	0.73	3
Jim Boeheim	M	38	30	942	314	1256	0.75	1
Jack Gardner	M	28	8	486	235	721	0.67	2
Kimberly Mulkey	F	14	14	376	81	457	0.85	1
Gene Keady	M	27	18	550	289	839	0.66	2
Thad Matta	M	14	11	370	109	479	0.77	3
Anthony Grant	M	8	3	171	90	261	0.66	2
Hank Iba	M	40	8	752	333	1085	0.69	3
Sean Miller	M	10	6	237	91	328	0.72	2
Mike Montgomery	M	32	16	670	312	982	0.68	3
Jay Wright	M	20	10	399	231	630	0.63	2
Tonya Cardoza	F	20	20	107	57	164	0.65	2
Joanne McCallie	F	21	19	457	180	637	0.73	3
Bo Ryan	M	15	12	339	145	484	0.7	2
Sherri Coale	F	17	14	381	179	560	0.68	1
Eddie Sutton	M	37	26	806	329	1135	0.71	5
Jim Valvano	M	18	9	337	200	537	0.63	3
Jamie Dixon	M	11	9	281	90	371	0.757	1
Muffet McGraw	F	31	23	714	258	972	0.73	5

Lisa Stockton	F	26	20	464	221	685	0.56	2
Brenda Fres	F	14	12	335	123	458	0.7	3
Charlaine Vivian	F	43	32	901	332	1233	0.73	3
Deb Patterson	F	17	17	339	207	546	0.62	1
Tina Martin	F	18	17	341	181	522	0.65	2
June Daugherty	F	24	20	369	335	704	0.52	3
JenniferRizzotti	F	14	8	276	146	422	0.45	1
Suzy Merchant	F	18	17	341	172	513	0.66	3
Lisa Bluder	F	29	19	617	297	914	0.69	4
Sue Semrau	F	18	13	309	203	512	0.6	1
Sharon Versyp	F	14	10	283	144	427	0.66	3
MaChelle Joseph	F	11	7	204	126	330	0.44	1
Nikki Caldwell	F	6	5	133	52	185	0.72	2
Pam Borton	F	10	10	283	185	468	0.45	2
Joanne Boyle	F	11	9	244	118	362	0.51	3
Melissa McFerrin	F	6	1	155	147	302	0.51	2
Connie Yori	F	21	12	410	275	685	0.47	4
Kellie Harper	F	10	10	171	135	306	0.52	3
Beth Couture	F	25	20	436	275	711	0.44	4
Regina Miller	F	22	19	235	226	461	0.51	4
Kim Barnes Arico	F	10	8	292	216	508	0.42	2
Jamelle Elliott	F	5	3	72	49	121	0.41	1

Modified Data

Name.	G	Exp	NExp	W	L	G	WR	T
John Wooden	M	29	16	755	162	917	0.823337	2
Geno Auriemma	F(M)	27	27	1204	133	1337	0.9005236	1
Pat Summitt	F	37	31	1512	208	1720	0.8790698	1
Dean Smith	M	36	27	1300	254	1554	0.8365508	1
Mike Krzyzewski	M	39	29	1462	302	1764	0.8287982	2
Roy Williams	M	26	23	1072	187	1259	0.8514694	2
Adolph Rupp	M	41	20	897	190	1087	0.825207	1
John Calipari	M	22	14	877	171	1048	0.8368321	3
Tara VanDerveer	F	36	33	1241	203	1444	0.8594183	3
Tom Izzo	M	19	16	687	181	868	0.7914747	1
Billy Donovan	M	20	13	705	188	893	0.7894737	2
John Thompson	M	27	20	894	239	1133	0.7890556	1
Bill Self	M	21	15	786	169	955	0.8230366	4
Denny Crum	M	30	23	1012	295	1307	0.7742923	1
Rick Pitino	M	28	18	1021	239	1260	0.8103175	4
Bob Knight	M	42	28	1348	374	1722	0.7828107	3
Jerry Tarkanian	M	30	18	1141	202	1343	0.8495905	3
Jim Calhoun	M	40	23	1315	382	1697	0.7748969	2

Lute Olson	M	34	28	1164	285	1449	0.8033126	3
Jim Boeheim	M	38	30	1413	314	1727	0.8181818	1
Kimberly Duane Mulkey	F	14	14	569	81	650	0.8753846	1
Jack Gardner	M	28	8	542	235	777	0.6975547	2
Thad Matta	M	14	11	555	109	664	0.8358434	3
Gene Keady	M	27	18	825	289	1114	0.7405745	2
Anthony Grant	M	8	3	256	90	346	0.7398844	2
Mike Montgomery	M	32	16	1005	312	1317	0.7630979	3
Sean Miller	M	10	6	355	91	446	0.7959641	2
Tonya Cardoza	F	20	20	217	57	274	0.7919708	2
Jay Wright	M	20	10	598	231	829	0.721351	2
Bo Ryan	M	15	12	508	145	653	0.7779479	2
Joanne P. McCallie	F	21	19	674	180	854	0.7892272	3
Hank Iba	M	40	8	752	333	1085	0.6930876	3
Sherri Coale	F	17	14	575	179	754	0.7625995	1
Eddie Sutton	M	37	26	1209	329	1538	0.7860858	5
Lisa Stockton	F	26	20	684	221	905	0.7558011	2
Jamie Dixon	M	11	9	421	90	511	0.8238748	1
Jennifer Rizzotti	F	14	8	438	146	584	0.75	1
Muffet McGraw	F	31	23	1010	258	1268	0.79653	5
Jim Valvano	M	18	9	505	200	705	0.7163121	3
Brenda Fres	F	14	12	515	123	638	0.80721	3
MaChelle Joseph	F	11	7	344	126	470	0.7319149	1
Charlaine Vivian Stringer	F	43	32	1254	332	1586	0.7906683	3
Tina Martin	F	18	17	523	181	704	0.7428977	2
Deb Patterson	F	17	17	520	207	727	0.7152682	1
Pam Borton	F	10	10	447	185	632	0.7072785	2
Suzy Merchant	F	18	17	523	172	695	0.752518	3
Sharon Versyp	F	14	10	447	144	591	0.7563452	3
Lisa Bluder	F	29	19	883	297	1180	0.7483051	4
Sue Semrau	F	18	13	481	203	684	0.7032164	1
Joanne Boyle	F	11	9	396	118	514	0.770428	3
June Daugherty	F	24	20	559	335	894	0.6252796	3
Kim Barnes Arico	F	10	8	459	216	675	0.68	2
Connie Yori	F	21	12	613	275	888	0.6903153	4
Nikki Caldwell	F	6	5	251	52	303	0.8283828	2
Beth Couture	F	25	20	647	275	922	0.7017354	4
Kellie Harper	F	10	10	300	135	435	0.6896552	3
Lon Kruger	M	30	14	531	338	869	0.611	5
Melissa McFerrin	F	6	1	279	147	426	0.6549296	2
Jamelle Renee	F	5	3	171	49	220	0.7772727	1

Elliott								
Regina Miller	F	22	19	384	226	610	0.6295082	4