# NCAA Coach Salaries

# Introduction

# In 2021 the NCAA had a revenue of $1.15 billion (NCAA earns). For successful teams football is a major source of revenue and media attention. The success of football team can be the result of many variables not least of which is the performance of the head coach. For this analysis the salary of a football head coach will be investigated and suggestions for future coaches will be made.

# Obtaining the Data

The first source of data was the coaches data set contained data about 129 universities including the NCAA conference it belongs to the head coach’s name, his total pay, his bonus, the bonus amount he was actually paid, the assistant’s pay and the Buyout amount.

The second piece of data used was NCAA gameplay stats for each university in the year 2019 from https://www.sports-reference.com/ . With this dataset the goal was to see if expectations of the quality of gameplay, as measured through gameplay statistics, impacted how much a school saw the value of their athletic program and ultimately the salary they are willing to pay the head coach. The year 2019 was chosen because it appeared to be the best approximation of the when the coaches dataset was collected based on salaries paid in 2019. This second dataset contained a ranking of each team based on a poll by the Associated Press, the number of wins, loses, OSRS (the offensive component of the simple rating system which creates a score for each team based on their points differential), DSRS, (the defensive component of the SRS), SRS, average points per game, average points per game for their opponents, average passing yards per attempt, opponents average passing yards per attempt, rushing yards per attempt, and opponent’s rushing yards per attempt.

The third data set obtained was hosted was from the US Department of Education and available on Kaggle <https://www.kaggle.com/datasets/kaggle/college-scorecard?select=MERGED2013_PP.csv>. This data set contained thousands of rows about universities across the US with each campus having its own entry. From this data set a subset was taken with attributes including the state, highest degree available at that university including (associates bachelors, graduate or certificate), the latitude, longitude, whether it was a historically black university, whether it is currently predominantly black university, if the university has a religious affiliation, the admission rate, the median SAT reading, median SAT math score, median SAT writing score, number of undergrads, the percent of undergraduates that are white, the percent of undergraduates that are black, the cost of attendandce, the amount of turition revenue per student, and the average faculty salary at the university. One final dataset was used from the NCAA site to pull in GSR or graduation success rate for each school (Graduation Rates, n.d.),

# Scrubbing the Data

Scrubbing the coaches data set required cleaning all dollar amounts to remove commas and dollar signs and format them as numbers. Dummy variables were created for each conference in the dataset so that they could later be used in modeling. Combining the coaches data set and the gameplay data set required minimal effort. Some universities needed to be renamed in order to be mapped onto the coaches dataset and many variables needed to be turned into numeric.

Most of the scrubbing came with using the US Department of education dataset. There were of thousands of schools listed and tens of thousands of attributes for each. The first challenge was to identify which variables could be of interest without expanding the work substantially. In the end the variables were chosen that could give a snapshot about the size of the university, difficulty of admissions, some student demographic information, and the financial standing of the university. The next step was to map these columns onto the previously combined dataframe. This took a tremendous amount of effort as each campus of each university was listed. Many NCAA team names such as ‘California’ or ‘Texas’ in the coaches data set could be grouped with over 20 entries from the US Department of education dataset. Time was taken to research which university belonged to each team and when there were more than campuses listed for that university more time was used to research which campus was either the main campus or which was historically the first campus of that university. The main / first campus was then renamed with the name of NCAA team name in the coaches dataset and finally this dataset was able to be mapped onto the other combined dataframe. The final list of columns in the dataframe is below.

|  |  |
| --- | --- |
| School | Conference\_x\_MAC |
| Coach | Conference\_x\_Mt. West |
| TotalPay | Conference\_x\_Pac-12 |
| Buyout | Conference\_x\_SEC |
| Rank | Conference\_x\_Sun Belt |
| Poll Rank | State |
| Wins | highestDegree |
| Losses | Latitude |
| OffenSRS | Longitude |
| DefSRS | HistBlack |
| SRS | PredomBlack |
| PointsPerGame | Religous |
| OppPointsPerGame | AdmisRate |
| Passing | midSATread |
| OppPassing | midSATmath |
| Rushing | midSATwriting |
| OppRushing | undergrads |
| TotalYards | percentWhite |
| OppYards | percentBlack |
| Conference\_x\_ACC | costOfAttend |
| Conference\_x\_Big 12 | turitionRevenueperStud |
| Conference\_x\_Big Ten | facultySalary |
| Conference\_x\_C-USA | GSR |
| Conference\_x\_Ind. |  |

In preparation for building a model, variables that will be used in the models with missing values including percent of students that are white, percent of students that black, and average faculty salary were filled with the means of those columns. The following columns were normalized by taking their score:

['SRS' , "OppPointsPerGame" , "Losses" , "Rank", "PointsPerGame", "turitionRevenueperStud" , "GSR" , "percentWhite" , "percentBlack" , "facultySalary" , "TotalYards" , "OppYards" , "DefSRS" , "Wins"]

# Exploration

Chart

Description automatically generated The first step to exploring the data set was to look at summary statistics of coaches’ total pay across the NCAA. The table shows the mean salary is 2417060.76 and the standard deviation 1885752.30. In order to get a better idea of the distribution of salaries the histogram below was also created. The histogram shows that the distribution is skewed right with the vast majority of coaches earning less than $1,000,000 a year but the distribution has a long tail with some earning more $8,000,000 a year.

|  |
| --- |
| Total Pay |
| count | 125.00 |
| mean | 2417060.76 |
| std | 1885752.30 |
| min | 390000.00 |
| 25% | 805850.00 |
| 50% | 1900008.00 |
| 75% | 3617500.00 |
| max | 8307000.00 |

The school conference was the first variable explored to see if it influenced total pay. One histogram was created for total pay of coaches in each conference. Below is the plot of these histograms.

Graphical user interface

Description automatically generated

From the distributions above its apparent that conference likely has a major impact on coach salary. Within each conference there is still some variation To determine which other variables may be having an impact on coaches’ salaries a correlation matrix was created. The total pay column of the correlation matrix was then reordered and turned into heatmap. Variables on either end of the heatmap have the highest correlation with coaches’ total pay. The fill color for positive correlation is green and the fill color for negative correlation is red.

Table

Description automatically generated

## From this plot it appears defensive coach performance has a high correlation with coach salary with opponents points per game, opponents yards, opponent passing, and opponent rushing yards having strong negative correlation with total pay. Offensive stats have a positive correlation with coach total pay. Also SRS score and team rank have a high degree of correlation with coach pay. Interestingly the median SAT math and reading scores have a negative correlation with coach total pay. This may imply more academically competitive schools pay their coaches less. Cost of attendance for the school is also negatively correlated with coaches’ total pay. A scatter plot was created for total pay vs each of the top variables with the strongest correlation with total pay to better visualize their relationships as well as one to explore the relationship between graduation success rate at a university and the coaches’ total pay a scatter plot was created. From this scatter plot it appears that there a weak positive correlation between graduation success rate and the coach’s total pay.

A picture containing text, window, display

Description automatically generated

## Analysis: Models

Three multiple linear regression models were created to predict coach total pay based on a mix of independent variables. The original data set was randomly split into a training data set which contained 67% of entries and a test data frame containing 33% of the original data. All of the models were trained using only the training data set. Model 1was a multiple regression model that only considered the conferences as independent variables to predict total pay.

'TotalPay ~ Conference\_x\_ACC+ Conference\_x\_Big\_12 + Conference\_x\_Big\_Ten + Conference\_x\_C\_USA + Conference\_x\_MAC + Conference\_x\_MtWest + Conference\_x\_Pac12 + Conference\_x\_SEC + Conference\_x\_Sun\_Belt'

A picture containing text, receipt

Description automatically generated

Model 1 had an adjusted R2 of 0.6737. From the p values of the individual conferences of this model all conferences except Conference USA, MAC, Mountain West, and the Sunbelt conference had a statistically significant impact on predicting total pay but none of these will be excluded because they are all aspects of the one attribute which is conference.

Model 2 used gameplay statistics including the team’s SRS score, Defensive SRS, the opponent’s points per game, the team’s total yards, and the opponent’s yards. Some additional school demographic and financial variables such as percent of students that are white, percent that are black, the average faculty salary and the tuition revenue per student were also included.

"""TotalPay ~ SRS + OppPointsPerGame + Losses + Rank+ PointsPerGame+ turitionRevenueperStud + GSR + percentWhite + percentBlack + facultySalary + TotalYards + OppYards + DefSRS + Wins""""

Table

Description automatically generated

The adjust R2 for Model 2 was also 0.609 so this model performed worse. Variables with a p-value below 0.1 were SRS, Rank, and points per game and will be include in the next model along with a few other variables that showed high correlation with Total Pay in the heatmap. These variables will be used in the next model. Graduation success rate will still be included because to answer the question on how graduation rate effects salary.

The input variables for the third model are below.

model3 = str("""TotalPay ~ SRS + Rank+ PointsPerGame+ GSR +

Conference\_x\_ACC+ Conference\_x\_Big\_12 + Conference\_x\_Big\_Ten + Conference\_x\_Pac12 + OppYards + Wins +Conference\_x\_SEC""")

Table

Description automatically generated

This model has an adjusted R2 of 0.73 which is the highest R2 yet. This indicates that the model fits the data better than previous models but analysis of the test dataset is still needed.

To determine how well each model performs on unseen data the actual coaches’ salaries were plotted against the predicted coach salaries. If a perfect model was plotted the line of best fit would have a slope of 1 and would pass through the origin with a 45 degree angle. This would mean as the salary increases the predicted salary increases by the exact same amount. The closer the plot is this ideal plot the better the model.

Chart, scatter chart

Description automatically generated

From these plots its clear that the first and third models performed better than the second model on predicting the test data set. A correlation matrix was also created for the total salary vs each of the predicted salaries below.

Table

Description automatically generated

This correlation matrix shows that predictions from the first and third model have the highest degree of correlation with the actual salaries in the testing dataset. Based on this and the fact that Model 3 performed best on the training data set with a higher adjusted R2 model 3 will be used for all predictions in this report moving forward.

# Discussion Questions

1. The recommended salary for the Syracuse head football coach was calculated based on model 3 and is
   1. $ 3,283,734.73
      1. The 95 Confidence interval of Syracuse head coach would be: $1,225,415.90 to $4,703,136.23
2. What would his salary be if we were still in the Big East? What if we went to the Big Ten?
   1. Big East salary recommendation -> $ 1438886.62.
      1. The Big East is not part of this data set. This recommendation is based on a team with 0 inputted for all dummy variable columns that does not belong to any of dummy variables listed. This method most closely puts Syracuse in the AAC which became the home of 5 AAC member universities after leaving the Big East. To answer this question more fully a dataset from the time when the Big East was still a conference would have to be analyzed and a new model would need to be created. The current method would most
      2. The 95 Confidence interval of Syracuse the head coach in the Big East: $277,415.90 to $1,963,136.23
   2. Big Ten Salary recommendation -> $ 3277350.63.
      1. The 95% Confidence interval of Syracuse head coach: $1,085,415.90 to $4,833,136.23
3. No schools were dropped from this data set. Extensive work was done to map all schools.
4. What effect does graduation rate have on the projected salary?
   1. The coefficient for GSR in the linear model is $143,600. This means as the graduation success rate increase by one z score the coach’s salary is predicted to increase by $143,600. The p value for GSR was somewhat high so the actually effect that the graduation success rate has on a coach’s salary many not be statisticaly significant.
5. How good is our model?
   1. The adjusted R squared of the model is 0.74 meaning this model accounts for 74% in the variation of coaches’ salaries. In the testing of this model, the predicted salaries had a correlation of 0.86 with the actual salaries.

# 

# Citations

AP NEWS. 2022. NCAA earns $1.15B in 2021 as revenue returns to normal. [online] Available at: <https://apnews.com/article/coronavirus-pandemic-sports-business-health-indiana-c0f4a0a468c22ae071ad84562b03c827> [Accessed 17 April 2022].

NCAA.org. n.d. Graduation Rates. [online] Available at: <https://www.ncaa.org/sports/2013/11/19/graduation-rates.aspx> [Accessed 19 April 2022].

Kaggle.com. 2018. US Dept of Education: College Scorecard. [online] Available at: <https://www.kaggle.com/datasets/kaggle/college-scorecard?select=CollegeScorecardDataDictionary-09-12-2015.pdf> [Accessed 17 April 2022].

Page, C. and Ratings, 2., 2022. 2019 School Ratings | College Football at Sports-Reference.com. [online] College Football at Sports-Reference.com. Available at: <https://www.sports-reference.com/cfb/years/2019-ratings.html> [Accessed 17 April 2022].