Holden Herrell
27 January 2020

IST 718: Lab 1

Introduction

In the widely competitive world of NCAA Division 1 (FBS) college football, schools are always trying to hire the best coach. A caveat to luring a highly esteemed or promising young coach away from their current school is offering them the right contract. There are many contributing factors when arriving at the right salary terms to offer a coach, but the true impact of these factors is relatively unknown. In order to provide insight on these factors, a study of 2019 coach salaries was sponsored by Syracuse University. Syracuse would like to use the findings of this study to determine the best salary for its head football coach.

About the Data

Coach Data

The primary source of data for this study was a list of coaches' salaries from a USA Today source (https://sports.usatoday.com/ncaa/salaries/). This data includes data point such as a coach's school, athletic conference, salary, bonus, and total pay. It has additional data points including bonus paid, assistant pay, and contract buyout; however, there are not values for all coaches within the data set. For the purpose of this study, these attributes were not used for analysis. Among the remaining data, there is no salary, bonus, or total pay listed for the coaches at Baylor, BYU, Rice, or SMU—thus they were also excluded from this study. Some coach bonus values were missing, so in lieu of removing the bonus attribute, it was recalculated as the difference between total pay and salary. In order to integrate this data with supplemental datasets, more commonly used school names were applied (i.e. Ole Miss instead of Mississippi).

Stadium Data

To determine the impact of stadiums on coaches' salaries, stadium data was sourced from CollegeGridirons.com (https://www.collegegridirons.com/comparisons-by-capacity/). This data includes a school's stadium name, stadium capacity, and stadium year of opening. UMASS (Massachusetts) had two stadiums assigned to it within the data, so Gillette Stadium (the temporary stadium during the 2012 season) was removed from the data (https://en.wikipedia.org/wiki/Gillette_Stadium). Liberty University's stadium was not listed in the initial stadium data, so it was added from (https://en.wikipedia.org/wiki/Williams_Stadium). In order to improve analysis, the stadium year opened attribute was used to derive the attribute stadium age (*2019 – YearOpened*). Again, in support of integration, school names were aligned with those in the coach data.

Graduation Success Data

To evaluate the effect of school graduation success rate data on coaches' salaries, data was gathered from the NCAA (https://web3.ncaa.org/aprsearch/gsrsearch). The most recent data for all schools within the coach salary data is from the 2012 NCAA football player graduation cohort. This data included both the Graduation Success Rate (GSR) and the Federal Graduation Rate (FGR); however, the FGR data was incomplete for 19 of the schools within the coach data. For this reason, FGR was omitted and GSR was used as the primary metric for football player graduation rate performance. School names were normalized to match those within the coach data.

Career Coaching Statistics Data

To measure the contributions of career coaching statistics on coaches' salaries, data was collected from Sports Reference (https://www.sports-reference.com/cfb/coaches/). This data included attributes such as years coached, games coached, wins, losses, ties, win percentage, bowl games coached, bowl wins, bowl losses, bowl

ties, and bowl win percentage. Among the coaches in the coach data there were 15 who had not coached in a bowl game. As a result, these did not have a calculated bowl win percentage in the data. For the purpose of this study, bowl record (wins, losses, ties) was deemed a sufficient metric for bowl game success and bowl win percentage was excluded. Coach names were aligned with those in the coach salary data to support further analysis.

2019 Team Statistics Data

To gauge the influence of 2019 team performance on coaches' salaries, team offensive and defensive statistic data was gathered from the NCAA (https://www.ncaa.com/stats/football/fbs/). Among this data are attributes including: games played, offensive touchdowns, defensive touchdowns allowed, offensive extra points made, defensive extra points allowed, offensive field goals made, defensive field goals allowed, offensive 2-point conversions, defensive 2-point conversions allowed, offensive safeties, defensive safeties, offensive points per game, defensive point allowed per game, total offensive points scored, and total defensive points allowed. Again, school names were normalized to match the coach data.

Aggregate Coach Data

Each of the aforementioned data sets were aggregated into the final data set used for the analysis of this study. The merging was done by either joining on the school name (stadium, graduation success, team statistics) or the coach name (career coaching statistics). To allow modeling based on athletic conference, additional binary attributes were added to the aggregate for each of the conferences (i.e. coaches in the ACC have a value of 1 for the attribute 'ACC').

Initial Analysis

As part of the initial exploratory analysis of the aggregate coach data, basic descriptive statistics were calculated on all 123 records. Results of selected attributes are shown in Figure 1 (below).

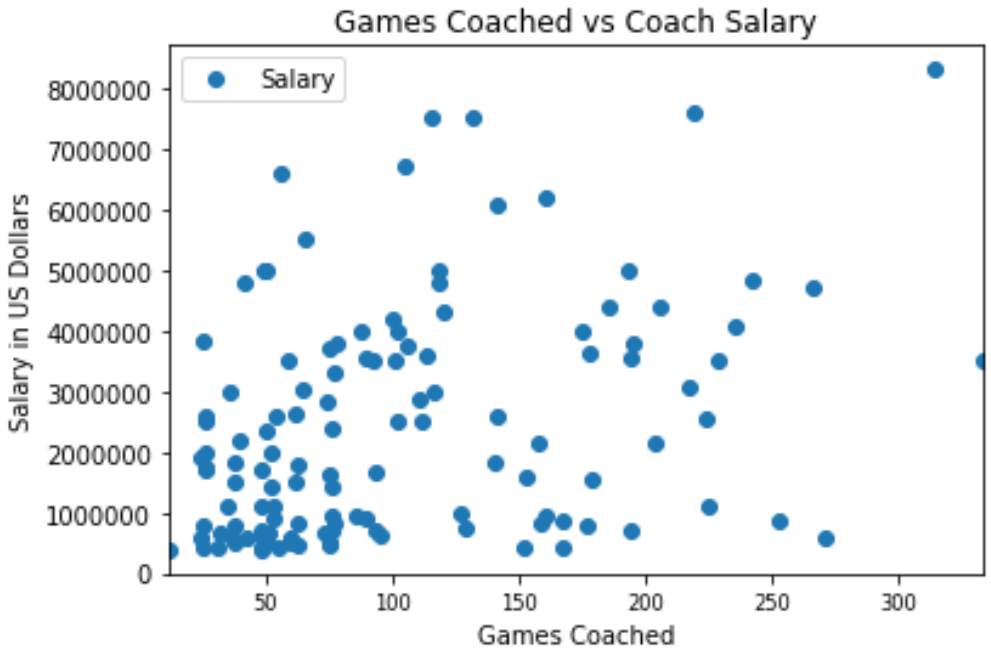
Figure 1:

	Salary	TotalPay	Bonus	Stadium Capacity	Stadium Age	2012 Grad Success Rate	Years Coached	Games
mean	$ 2,434,452	$ 2,441,322	$ 6,869.97	51,713.32	62.878	79.049	8.268	103.211
std	$ 1,886,997	$ 1,891,342	$ 47,293.82	23,603.66	29.646	8.765	5.500	69.964
min	$ 390,000	$ 390,000	$ -	15,000.00	2.000	54.000	1.000	12.000
25%	$ 800,752	$ 804,427	$ -	30,528.00	42.500	73.000	4.000	50.000
50%	$ 1,900,008	$ 2,000,000	$ -	50,000.00	59.000	79.000	6.000	77.000
75%	$ 3,612,388	$ 3,618,638	$ -	65,118.00	90.500	86.000	11.500	146.500
max	$ 8,307,000	$ 8,307,000.00	$ 400,000.00	107,601.00	115.000	100.000	27.000	333.000

	Wins	Win Percentage	Bowl Games	Bowl Wins	Wins Per Year	2019 Games	2019 Points PerGame Allowed	2019 Points PerGame
mean	59.919	53.484%	5.447	2.846	6.677	12.756	27.320	28.511
std	49.647	15.326%	5.074	3.183	2.143	0.750	6.873	6.860
min	5.000	12.500%	0.000	0.000	1.500	12.000	12.600	10.500
25%	21.000	42.800%	2.000	0.000	5.056	12.000	22.100	24.150
50%	44.000	55.600%	4.000	2.000	6.833	13.000	28.100	29.000
75%	84.000	63.650%	8.000	4.500	7.981	13.000	31.900	32.500
max	248.000	85.700%	25.000	15.000	12.000	15.000	52.700	48.400

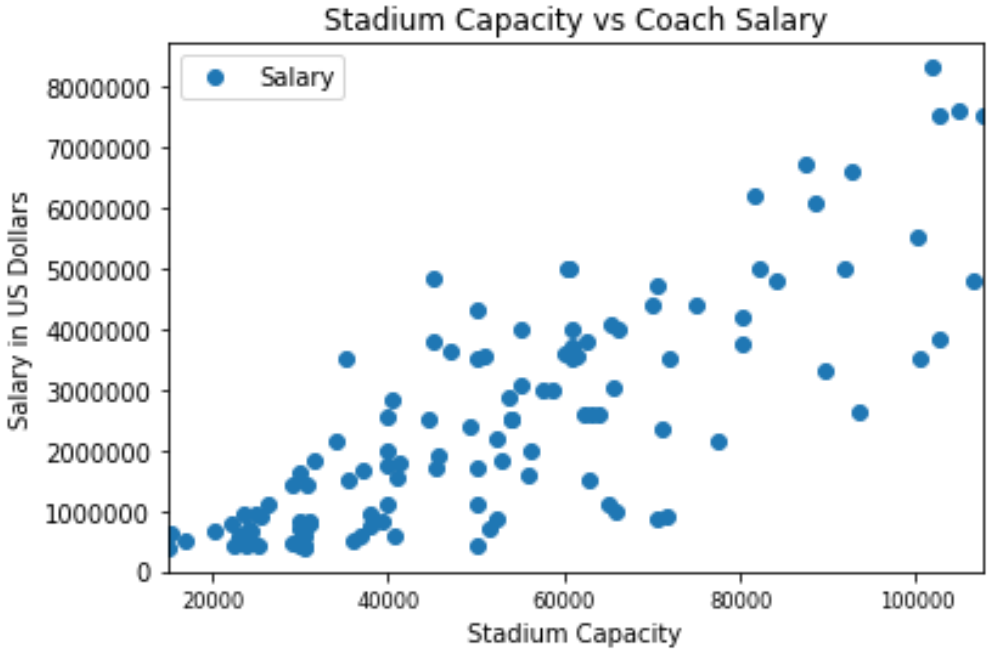
Coach salaries were found to be rather widely spread with a standard deviation of ~$1.9 million from a mean salary of ~$2.4 million. There also appears to be a fair amount of disparity in coaching experience with standard deviations in years coached and games coached both equating to roughly 5 seasons of coaching experience. In order to explore the potential relationship between salary and game coached, a scatter plot was created (see Figure 2 below). From this analysis, there is no easily visible linkage between games coached and salary—there are instances of both highly paid inexperienced coaches and relatively underpaid experienced coaches.

Figure 2:



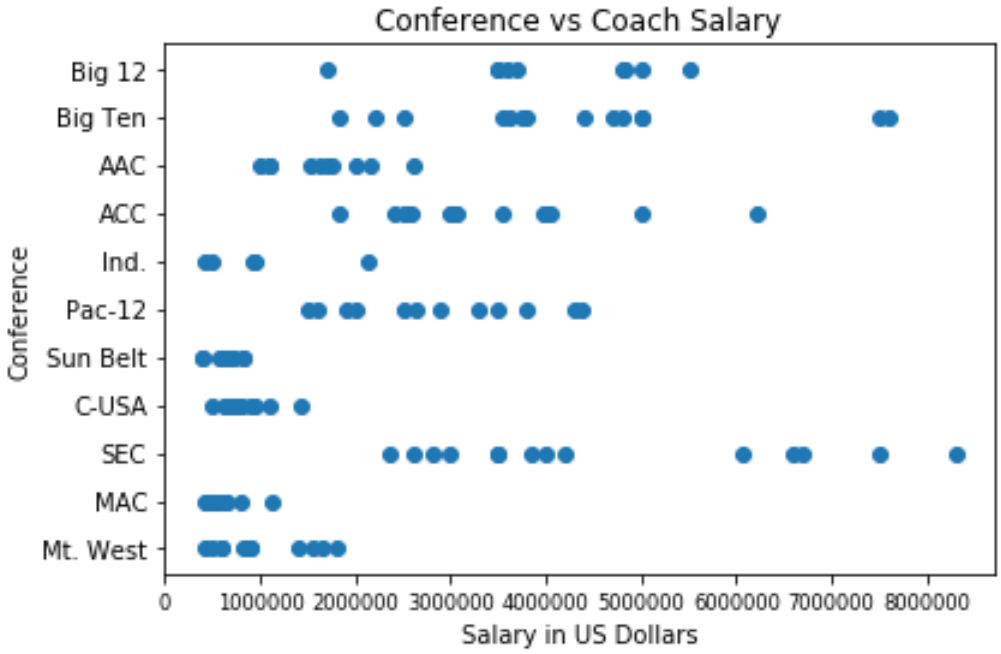
Another explored potential linkage that was the relationship between stadium capacity and coach salary. Figure 3 shows that there is a somewhat positively correlated relationship.

Figure 3:



The final exploratory analysis was conducted on the relationship between athletic conference and coach salary (Figure 4). There appear to be some conferences that are paid higher than others, however, there also appears to be clustering within the conferences.

Figure 4:



Models and Results

Linear Regression Model

Using the aggregate coach data, a linear regression model was developed using the following attributes as predictors of coach salary:

Stadium Capacity, Stadium Age, 2012 GSR, Games, Wins, Bowl Games, Bowl Wins, 2019 Points/game, 2019 Points Allowed/game, 2019 Games, AAC, ACC, Big12, Big Ten, CUSA, Independent, MAC, Mountain West, Pac12, SEC, and Sun Belt.

Model Output

Salary = *$22.39[StadiumCapacity] - $519.30[StadiumAge] - $6,234.52[GradSuccessRate2012] - $16,656.84[GamesCoached] + $36,128.55[CoachWins] - $47,445.50[BowlGamesCoached] + $8,550.57[BowlWins] + $20,785.27 [PointsPerGame2019] - $1,772.01[PointsPerGameAllowed2019] - $75,882.13[Games2019] - $390,857.515[AAC] + $774,113.20[ACC] + $1,082,530.76[Big12] + $1,378,965.49[BigTen] - $929,707.68[CUSA] - $740,432.45[Independent] - $948,774.99[MAC] - $924,055.66[MtnWest] + $199,494.82[Pac12] + $1,533,583.50[SEC] - $1,034,859.48[SunBelt]*

r^2 = 0.8594563054400856

This model was able to account for nearly 86% of the variation in coach salary (r^2 value) which should make it fairly accurate at predicting a coach's salary. Based on this model, the Syracuse head coach should have a salary of $1,087,432. Per the original coach data, Syracuse pays its current coach $2,401,206; therefore, according to the model, they are overpaying for their head coach by $1.3 million. According to the model, coach wins have the greatest impact on coach salary—if Syracuse were to relocate from the ACC to the BigTen, it expects the head coach salary to change to $1,692,284. If the model were correct, it would cost Syracuse ~$605K in additional coach salary if it were to change conferences and expect to keep its current coach. Also noteworthy is the fact that GSR negatively impacts coach salary, although it is relatively less impactful.

Scikit-Learn Linear Regression Model

Once again, using the aggregate coach data, a machine learning linear regression model was developed using the following attributes as predictors of coach salary: Stadium Capacity, Stadium Age, 2012 GSR, Games, Wins, Bowl Games, Bowl Wins, 2019 Points/game, 2019 Points Allowed/game, 2019 Games. This model was derived using the Scikit-learn machine learning package (Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011). 80% of the data was used to train the model, and 20% of the data was used to test the model. For this model, conference was not included as a predictor of coach salary.

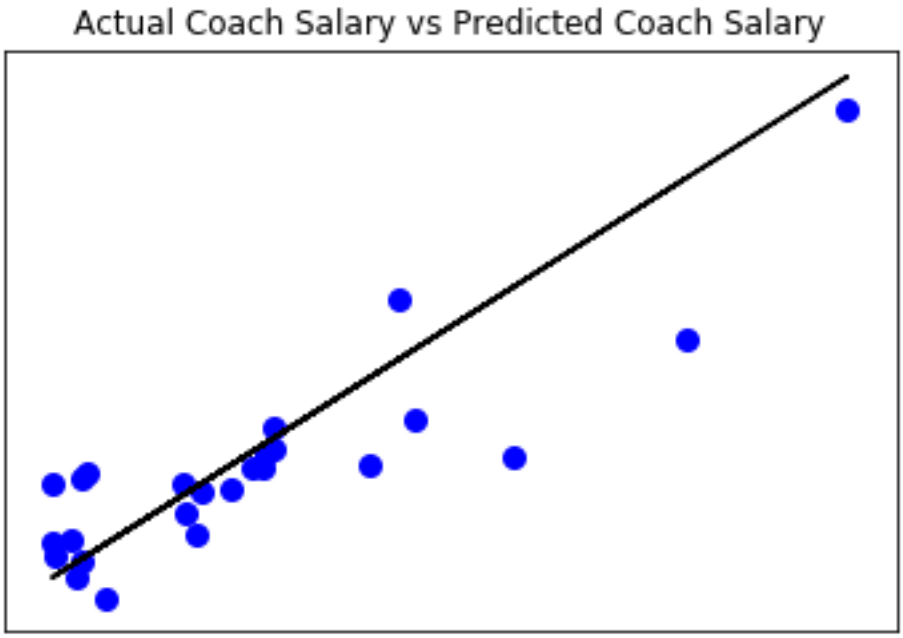
Model Output

Salary = *$51.26 [StadiumCapacity] - $3,615.57[StadiumAge] + $69.31[GradSuccessRate2012] - $958.04[GamesCoached] + $10,171.82[CoachWins] + $27,486.40[BowlGamesCoached] + $190,733.70[BowlWins] + $4,792.92 [PointsPerGame2019] - $7,991.38[PointsPerGameAllowed2019] - $143,365.62[Games2019]*

r^2 = 0.7320086368302925

This model was not able to account for as much variability in coach salary, but that can be expected with the retraction of conference as a predictor. Figure 5 (below) shows how well the predicted values of this model (in blue) match the actual salary values (solid line).

Figure 5:



Contrary to the previous model, the SciKit model indicates the most influential predictor of coach salary is bowl wins. The GSR seems to have very little impact on coach salary relative to the other predictors in this model. The predicted salary for Syracuse's head coach using the SciKit model is $1,340,030 which is still well below the current ~$2.4million.

Conclusions

In summation, Syracuse sought to determine the impact of different contributing factors to football coach salaries. When considering stadiums, graduation rates, career coaching statistics, and 2019 team statistics, it appears that the two most prominent predictors of coach salary are athletic conference and bowl wins. Both the linear regression model and machine learning linear regression model indicate that the current Syracuse head football coach is overpaid by ~$1.3 million. If Syracuse were to relocate to the BigTen, they would still be overpaying their coach, but by ~$605K less (the cost of relocating in expected salary). While stadium size was somewhat impactful, it seems that GSR has little to no impact on coach salary, so it could likely be excluded as a predictor in the future.