Modeling Salaries for NCAA Football Coaches

By Leland Ball

2020-10-15

Introduction

Salary compensation for the next head football coach of any university is an important matter. The average coach's salary carries a $2.4 million dollar price tag, which may not be justified at some university's football programs.

This analysis brings data to bear on the question of just what a head coach is worth paying based on metrics found to be important in the performance of other teams in the NCAA. The result of this analysis is a model that predicts with reasonable accuracy the salary of a prospective coach, given their track record. This model can be a starting point for future salary negotiations.

Analysis and Models

About the Data

Five sources of data were pulled together for exploration, though only metrics from several sources proved to be worthwhile. The data-sources have been arranged below:

1. The coaches dataset was provided, giving 129 coaches, their salaries, schools, and additional financial breakdowns. This dataset was the foundation and starting point for merging in other data.
2. Graduation success rate data was acquired from the NCAA. It was from this site that the GSR and FGR scores for each university were taken for the 2006 cohort year. [1]
3. Win and loss data was taken from Wikipedia, and represents the current 2019 list of football coaches and their win/loss records. [2]
4. Bowl game wins were retrieved from sports-reference.com, which provides statistics for each coach for bowl wins over their entire careers. Only some 83 coaches were represented in this dataset, and only 25% of those could be found in the 129 coaches that this analysis took into account. [3]
5. Finally, school finance totals, as well as financial revenue breakdowns were obtained from the NCAA at a different URL. This data provided not only a school's total revenue from their football programs, but also more detailed information about ticket revenues. [4]

The astute reader will note that many of these data-sources do not contain data from the same years. This fact is noted, but not addressed in this paper. It could be argued that a coach's performance will remain valid over a number of years, and many coaches stay with the same university for a long period

of time. However, additional accuracy should be obtainable by synchronizing the data across the temporal dimension.

Data was scraped from the various websites listed above using Python from within Google's Colaboratory environment. The Python notebook accompanying this analysis is able to scrape the sites and save modified copies of the data. Upon subsequent runs, these .csv and .json files may simply be uploaded to the cloud environment. While the Python has been included with this paper, it may be more convenient to access the URL, available in the appendix. [5] The code to re-scrape (and thereby re-acquire) the data will be at the bottom of the notebook.

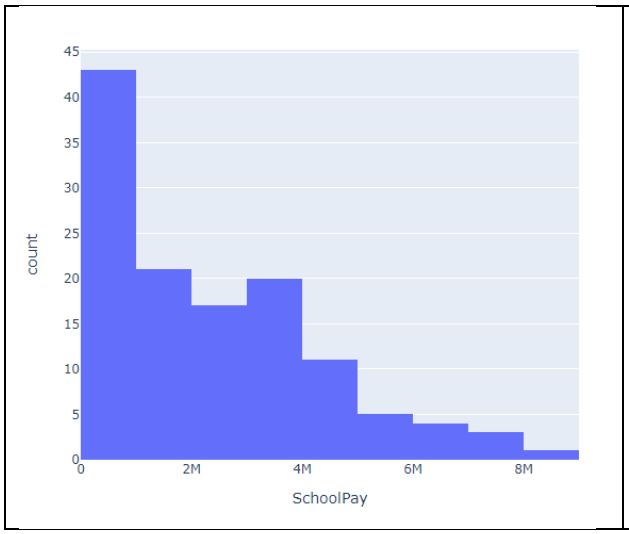
This data was then expediently joined together using a system of fuzzy matching between coach names and university names between the main dataset and each of the sources listed above. Not all coaches and universities were found to match these data, and the following percents show the amount of data successfully matched on

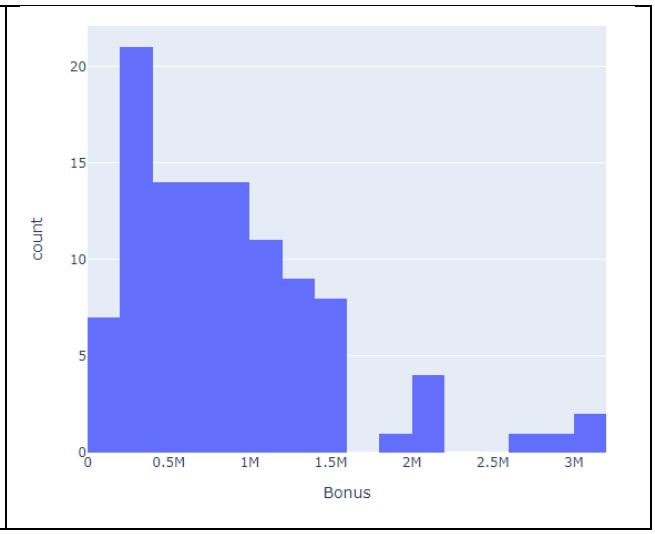
Data Source	Percent Joinable to Main Dataset
Bowl Wins	25%
School Finances	93%
Graduation Rates	85%
Wins and Losses	96%
Program Revenues	100%

Finally, missing values for these data sources were addressed. It was assumed that any coach not listed in the Bowl Wins dataset was likely a coach with no wins at all. Missing values from this dataset were zeroed out. The other datasets were first split into a 25/75 testing and training group split, before their own missing values were filled in with the median value in each relevant column. The medians were independently calculated so as to avoid contaminating the training set with incidental knowledge of the entire test/train dataset as a whole, thus avoiding a confounding factor in the assessment of model fitness. Furthermore, these medians were calculated on a per-conference basis. The operating assumption (later verified by boxplots) was that conference membership had a large effect upon salary. Taking the median for each sub-population provides a more accurate replacement for a missing-value.

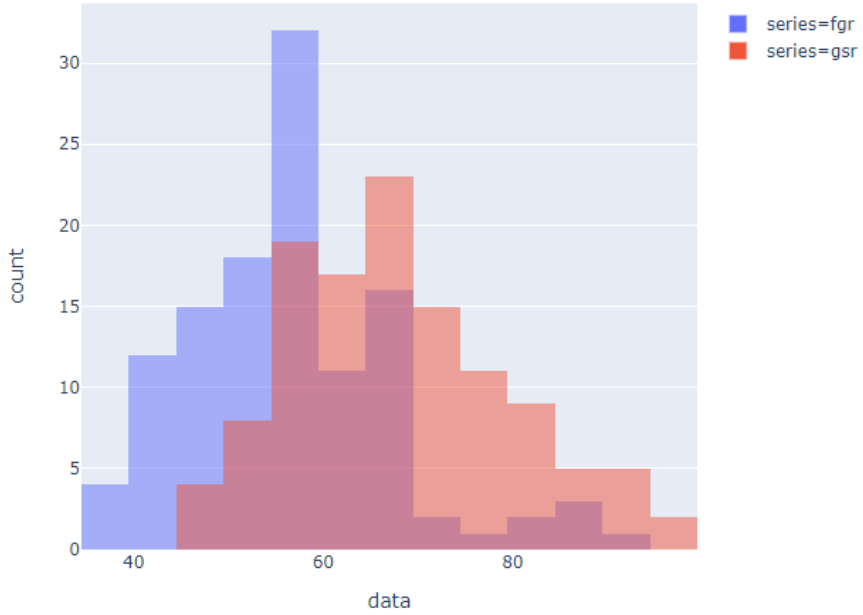
Data Exploration

Data exploration began with an investigation into the distribution of the predictor/response variable, "SchoolPay". This variable is the sum total of the money that a coach earns directly from a school. This is the response variable that the model was created to predict, and shows a gracefully decreasing distribution of salaries from the sub-$1M to above $7M. Bonuses given to each coach was also plotted and shows a more varied distribution.

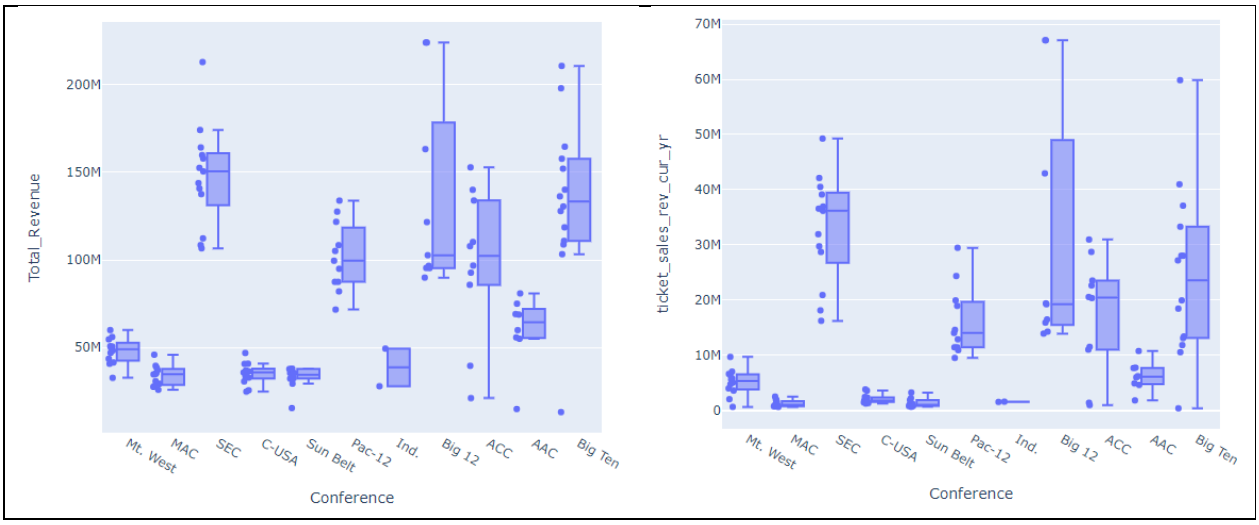




GSR and FSR (the variables representing graduation rates for the 2006 cohorts) were plotted against each other to ascertain the difference between them. A clearly shifted distribution can be seen, with several outliers in the FGR variable, while the GSR is more normally distributed.



While outliers were seen in both the Total Revenue and Ticket Sales data, no schools were removed from the test or training dataset. The outliers for these two variables can be seen in the boxplots below.



Additional Vectors

"Last Year's" records were investigated for inclusion into the model. Revenue from ticket sales was available for both 2019 and 2018, and were included in an exploratory linear model. These attempts never met with a statistically significant effect at p>a=0.05. Only a singular variable indicating revenue from ticket sales was significant. When including the prior year's along with the current year's data, the effect was diminished, harming the model. Additionally a column containing the difference between the prior year and the current year was created. This variable also had no significant effect upon the model.

Linear Model

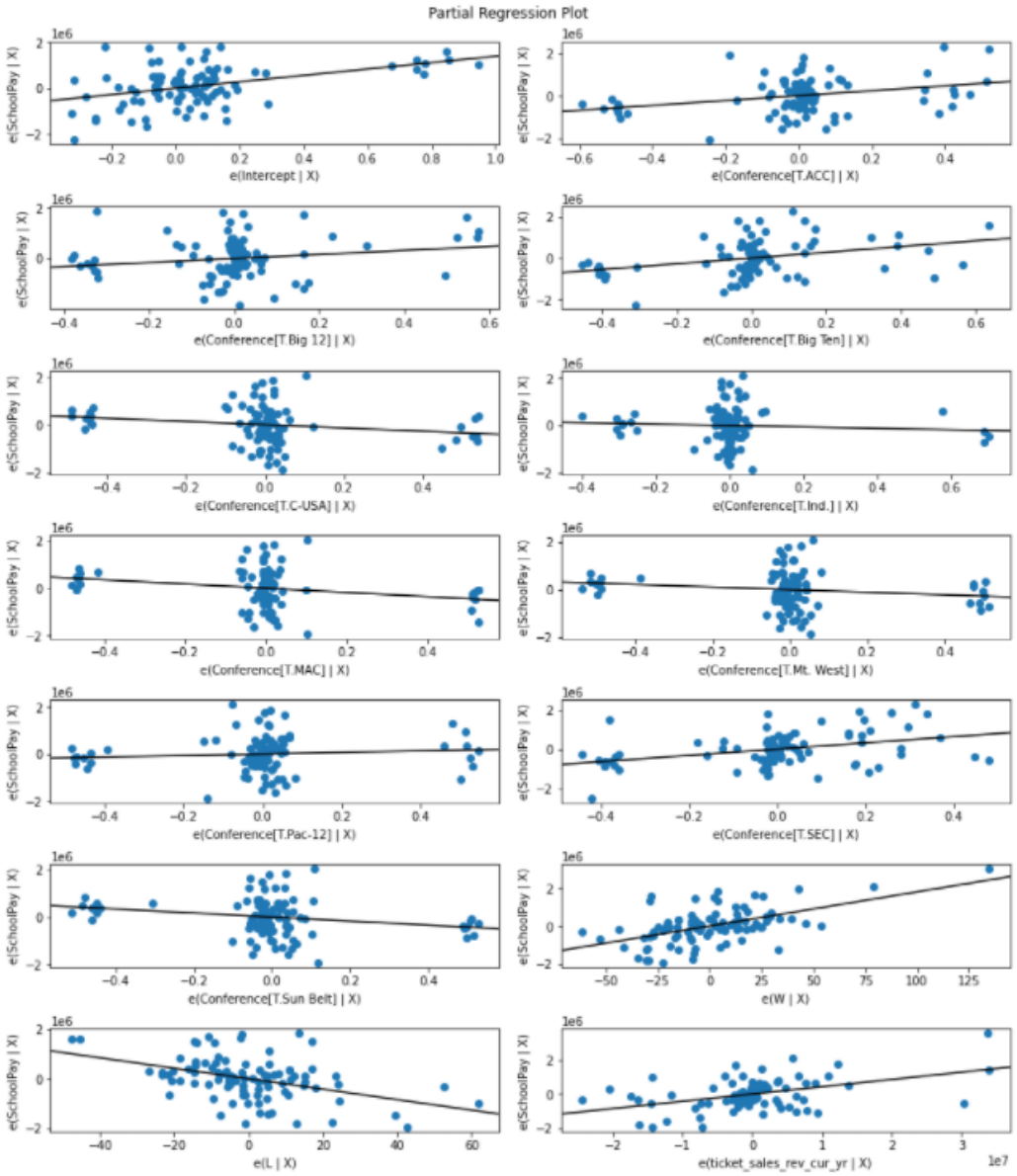
A linear regression model was prepared to predict coach salary from the four variables of: conference, wins, losses, and ticket revenue. The conference variable was categorical, spanning eleven different groupings of football teams. This variable was dummy-encoded. The others were floats representing integer or float values. This linear model was fit to the training dataset using ordinary least squares regression.

This linear model proved to be significant at the p<a=0.05 level. The F-statistic associated with this model was large at 32, with 13 degrees of freedom on 84 observations. These findings lead to the rejection of the Null Hypothesis, which states that this model is no better than the intercept-only model.

Indeed, after finding the model significant, it was found to have an Adjusted R-squared value of 0.830, implying that around 83 percent of the variation in a coach's salary can be explained by the variables in the linear model.

The model was then further assessed for fitness by using the model to perform predictions upon the training data. The simple accuracy of the model was found to be 71%. This training data was fully 25% the size of the original dataset.

A plot of the partial regression coefficients (included below) show that revenue from ticket sales, wins, losses, and membership in certain conferences show expected distributions. Certain conferences show some abnormalities associated with dummy-coding to some extent, and insignificant effects on the model in some cases.



The single biggest impact on salary size

As can be seen in the table below, the coefficient with the single largest effect upon salary size is **membership in the SEC conference**. This effect must be evaluated in the same unit that the variable is in. While membership in a different conference is a binary status- wins, losses, and especially ticket revenue can increase or decrease in larger numbers. Smaller magnitudes of effect will be magnified by large swings in a variable's range. An interesting thought-experiment could be: how many wins would a coach need to provide in order to secure a salary increase large enough to overshadow a coach moving

from the Big Ten to the SEC? The difference between the Big Ten and the SEC is 202,000. This is roughly equivalent (in the model) of staying in the Big Ten and leading their team to win about 11 more times.

Coefficient	Magnitude of Effect
Conference[T.SEC]	1588000.00
Intercept	1392000.00
Conference[T.Big Ten]	1386000.00
Conference[T.ACC]	1159000.00
Conference[T.MAC]	885600.00
Conference[T.Sun Belt]	852700.00
Conference[T.Big 12]	793000.00
Conference[T.C-USA]	677100.00
Conference[T.Mt. West]	546000.00
Conference[T.Pac-12]	332400.00
Conference[T.Ind.]	295900.00
L	21220.00
W	18090.00
ticket_sales_rev_cur_yr	0.04

Linear Model Without 'Conference'

An additional experiment was carried out to ascertain the effect of removing the 'Conference' variable from the Linear Model. The model was trained as before on a 25/75 test/train split. It was then found to have all statistically significant coefficients at the p>a=0.05 level: Wins, Losses, ticket revenue. These variables were used to predict the salary of a coach. The model was still found to be statistically significant at the same p>a=0.05 level, with a high F-statistic of 74, df=3 with 84 observations.

However, the model's explanatory power decreased. The new Adjusted R-squared value was 0.728, and the testing dataset was found to have resulted in an accuracy score of 64%.

Recommended Salary for Syracuse University Coach

The current salary for Dino Babers is $2,401,000. According to the model, his team's performance, current conference membership, and the revenue generated in ticket sales should give him a projected salary of $3,382,053. This is a marked increase.

Dino Baber's salary assuming he was still in the Big East conference is something of a mystery, as the Big East conference no longer exists. Operating under the assumption that the Big East went defunct because it was making enough money (and would therefore rank among the conferences with the lowest revenues currently accounted for by the model) then the Sun Belt would do as a passable stand-in conference for the purpose of estimating a coach's salary. In this conference, with all other variables being the same, Dino Baber is projected to make only $1,370,735.

If the conference was changed to the Big Ten (Syracuse currently exists in the ACC), then Dino Baber's salary is projected to increase to $3,609,522.

Results

The linear model presented in the analysis above was the end result of a series of explorations of the data, with many dead-ends. It was expected that certain variables would have more of an effect upon the data than they actually had. Much effort was spent to obtain some of this data, only to find that certain variables were not statistically significant coefficients in the model.

Effect of Graduation Rates

The GSR and FGR variables representing grades had little effect on the model. While the beginnings of distinct populations can be seen in the data drawn from the 2006 cohort, neither of these variables was a statistically significant coefficient in the linear model. Thus it can be said that graduation rate had no noticeable effect on the model

Effect of Bowl Game Performance

It was expected that a coach's participation in a bowl game, and especially a coach's ability to win a bowl game, would have a significant impact upon their salary. However, neither their participation in such bowl games, nor their wins within that category were found to be statistically significant. This may be related to the lower number of coaches having bowl game wins at all. This effect is suspected to be seen in the many missing values present when this dataset was joined to the original list of coaches. Only 25% of the coaches present in the original dataset were found to exist in the bowl game data.

Effect of Conference

The conference category proved to be a significant variable within the model. However, this variable was dummy-encoded, and each category was able to be scrutinized for statistical significance. It was found that membership in ACC, Big Ten, and SEC were significant, while other categories did not have a significant impact upon salary. These coefficients were left in the original model, as they are tied in with the significant coefficients. Future work could involve grouping the conferences into several groups, which may maintain the overall model fitness while removing insignificant coefficients.

Effect of Prior Year's Data

Data for the prior year's data for ticket sales revenue was included in the initial linear model. This was found not to have a statistically significant effect upon the model, when included with the current year's data. In addition, a feature column of data was generated with the calculated difference between the current year's ticket sales data and the prior year's. There was no statistically significant effect seen with any of these variables, and only the singular ticket sales data was a significant predictor of the coach's salary.

Conclusions

In conclusion, a linear model that uses conference membership, ticket sales revenue, coach wins, and coach losses has the ability to serve as a starting point for salary negotiations based upon data. This

model can account for 75%-80% of the variation in a given coach's salary. Conversely, graduation rates, prior year's ticket sales data, and performance in bowl games were not seen to be significant predictors of how well a coach was paid.

Appendix

URLs

[1] University Football Teams Graduation Rate: NCAA: https://web3.ncaa.org/aprsearch/gsrsearch

[2] Football Team Win Loss Data: Wikipedia: https://en.wikipedia.org/wiki/List_of_current_NCAA_Division_I_FBS_football_coaches

[3] Coach Bowl Wins and Losses: Sports Reference: https://www.sports-reference.com/cfb/leaders/win-loss-pct-post-coach-career.html

[4] NCAA School Finance Totals: USA Today: https://sports.usatoday.com/ncaa/finances/

[5] Football Coach Analysis Python Notebook: Google Colab: https://colab.research.google.com/drive/1yTUkZpJyN0hE1UiJy-Md92qPNFFXW1Qu?usp=sharing