April 2020

Jennifer Mead

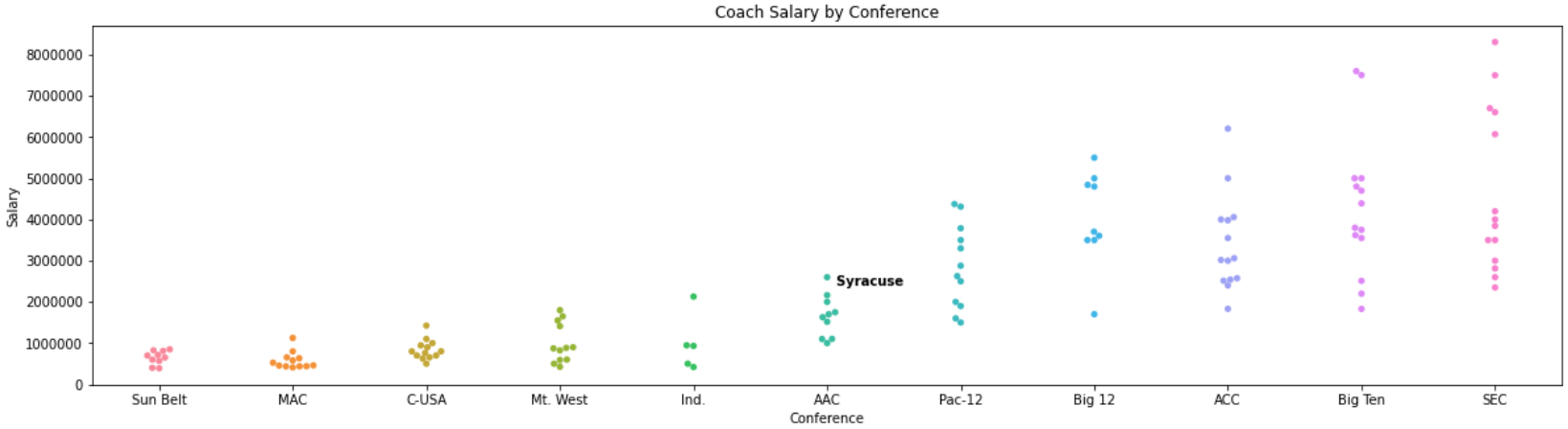
Report for the Athletic Director regarding salary negotiations with the Syracuse football coach

Problem statement

Salary negotiations are coming up, and we want to offer the head coach of the football team a salary that is good enough for him to stay at Syracuse happily, but not more than he is worth. But what is a coach worth? Going into this, we knew that conference plays a big role in coach salaries, but ideally we'd like a salary model that is based on more than conference. Football is a good money-maker for the university—ticket sales, parking, concessions, merchandise, etc.—and we'd like to incentivize the coach to make the team as successful as possible while also reinforcing Syracuse's values (such as a high graduation rate for its players).

The data

I was given a dataset of coach salaries with some additional information like school name, conference, bonus, total pay (salary plus media bonuses). Looking at the data, conference did differentiate salaries as we expected. Syracuse is at the top of its conference.



I then collected additional data such as:

- Win percentage—a winning team could be worth more
- Graduation rate—GSR and FGR give a score based on how many players graduate
- Stadium size—a bigger stadium means more tickets sold and more money made. It also indicates that alumni and local residents are engaged in the sports programs
- College sports fan rank—ranking for cities in the US based on how much their residents support university sports (including but not limited to football)
- State median income—if the people in the state had lower income, would they have less to spend supporting the football program?

There were other data that I looked for but couldn't find, or didn't find anything useful, such as:

- Coach tenure—data only available on some coaches, not the 100+ we are using
- Team rank—this is largely based on win percentage and conference, so adding this data didn't give us anything we didn't already have
- School donations to football program—could not find a reliable, complete source

Taking all of those data points, I combined them and cleaned them up in several ways to make them ready for analysis. For example:

- Four schools were deleted because there was no pay information available in the original dataset (Baylor, Brigham Young, Rice, Southern Methodist).
- FGR was removed because it was missing data on a third of the schools and was very similar to the GSR data we already had

- I separated the media bonus out of total pay, then removed the remainder of the pay columns because they didn't give additional information past SchoolPay.
- When pay information was missing, I replaced the values with the median value for the conference. There were other variables that I replaced missing values by looking up the most current values on the internet. For example, the size of the University of Alabama of Birmingham's stadium was missing, so I looked it out. A perfect dataset has no blanks, though that isn't always practical.

For more details on data handling, see the technical document.

The models

I used ordinary least squares regression (also known as linear regression) to model suggested salaries.

Results

I ran nine variations with many combinations of variables to find which worked the best, where "best" is defined by a high R-squared. R-squared represents the percentage of the data that is explained by the model. A perfect R-Squared would be 100%. Each model gives a range of salaries around its recommendation.

Of the nine models I made, their R-squared values ranged from 0.26 to 0.80. Just using conference gives an R-squared of 71%, so we want to do better than that. Each recommendation is given as a range; no model or dataset is perfect, and ranges help us remember that.

Current Syracuse coach salary: $2,401,206

Model name	Variables	R-sq	Low salary	Recommended mid-point salary	High salary
Just conference	Conference	0.71	2,296,267	3,375,859	4,455,451
Best linear model	Conference, graduation rate, win percentage, median income of the state, rank of town for college sports fandom, rank of media region	0.80	2,611,851	3,625,620	4,639,389

The process of making models largely involves looking at R-squared and then looking at whether each variable helps the model be more specific. Those that don't help are removed. And "help" doesn't necessarily mean "increase salary". For example, graduation rate is inversely correlated with salary, meaning that higher graduation rates tend to be associated with lower salaries. Graduation rate (GSR) and college fan rank had some influence; the biggest influence was conference (as we expected).

Evaluating the model

The model is good, but not *great*. It is a useful tool, but not the only tool to guide negotiation decisions.

The ideal model would match predictions exactly to actual salaries, like the solid blue line in the middle of this graph.

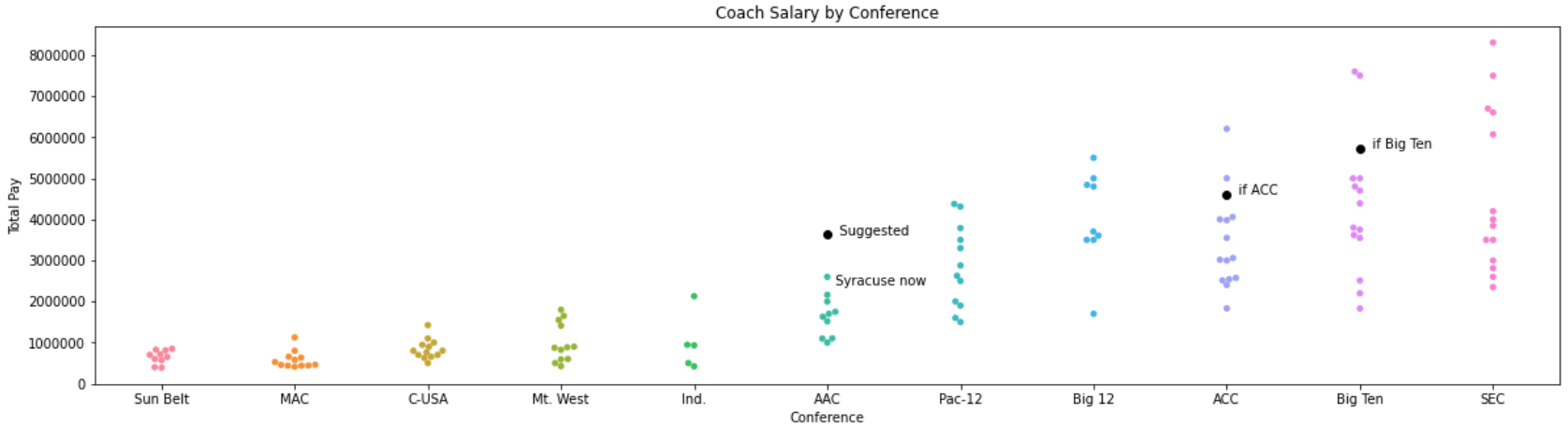
A great model would have most of the points inside the shaded regions. Our model has points further outside. This is like looking at True Catch Rate for wide receivers: you'd like them to make 100% of catches during games, but most don't. Most don't even catch 80%. If all of your wide receivers caught that well, you'd be thrilled…and you'd still have the receivers coach work with them to improve.

Recommendations

What should we offer?

Revisiting the salary chart from earlier, the suggested salary range from the model is between $2.6 million and $4.6 million. The chart shows the midpoint at $3.6 million. It's quite a bit higher than the rest of the salaries in the conference; the low point of the suggested range would be just slightly above the current salary.

As a comparison, I also used the model to estimate what the Syracuse coach's salary should be if Syracuse were in the ACC or Big Ten (like some of our former Big East cohorts). The suggested mid-point salaries for "alternate reality Syracuse" are higher for those conferences, which both have higher minimum and maximum salaries.



With any model, it is important to take a look at the results in context. **Aiming to negotiate a salary between the low point ($2.6 million) and the midpoint ($3.6 million) would make sense in comparison to the other salaries in the AAC.**

How can we improve the analysis for next year?

Just like football teams, data models can always be improved. Ideally, we'd want a model that was more than 90% accurate and had a salary range within $500,000 instead of a range of $2 million. For next year's analysis, consider:

Getting more observations. Since we don't control the number of schools in the NCAA, that likely means collecting data for several years back. This would give us more datapoints, and also allow us to analyze trends in salary over time. More observations mean error bars that are narrower.

Trying more kinds of models and data transformations. I used linear regression here, but there are other kinds of models, such as Random Forest, Decision Trees, and Naïve Bayes. It's worth trying these to see if any are well-suited to the problem. Other kinds of data transformations can also help, such as taking the log of the salary.

Getting more quantitative data. Some potentially useful datasets: coach tenure, assistant coach salaries, and the rate at which players go pro. More data helps the model be more specific in its recommendations.

Getting more qualitative data. The current dataset has little information about the effectiveness of the coach based on opinions. For example, if we knew how every player and assistant coach in the NCAA rated their head coach on factors like "knowledge of football", "teaching ability", and "team building", we might learn more about why the most successful coaches are winning. Can we develop those skills in our coach? Would our coach be motivated to improve if it meant a higher salary? Which of Syracuse's values can be represented in data?