Sammy Pardes
IST 718
Week Lab 3
7/24/2021

Introduction

Football is a draw to many prospective college students. They look for the atmosphere and the comradery that come with attending a university with a winning NCAA football program. To construct an alluring team, it's imperative to employ a stellar football coach. Many coaches come at a serious premium and competitive programs continue to pay. So, what is an appropriate price for a competitive university such as Syracuse to pay their head football coach?

Analysis and Models

About the Data and Data Cleaning

The data used to determine NCAA football coach salary came from four different sources. The first was the coaches9.csv file that contained 129 observations of 9 variables pertaining to coaches and their yearly salaries.

The second data set was obtained from sports-reference.com. 2014 college football standing data was pulled from the site. There were 128 schools in this data set of 15 variables. The metrics grabbed from the Sports Reference table included overall wins (W), losses (L), win-loss percentage (Pct), points per game (Off), opponent points per game (Def), and strength of schedule (SOS) for the 2014 season. Another attribute used from this data set was called SRS (Simple Ratings System). The SRS is a score which combines the average point differential and SOS. The data was matched with the coaches data on the School column and sparse columns were dropped from the dataframe.

Information pertaining to the football stadiums was gathered from collegegridirons.com. 131 stadiums were included in this data set along with college, conference, stadium capacity, and the year the stadium opened. Just like the standings data, the stadium capacity data was joined with the merged data set by the name of the university.

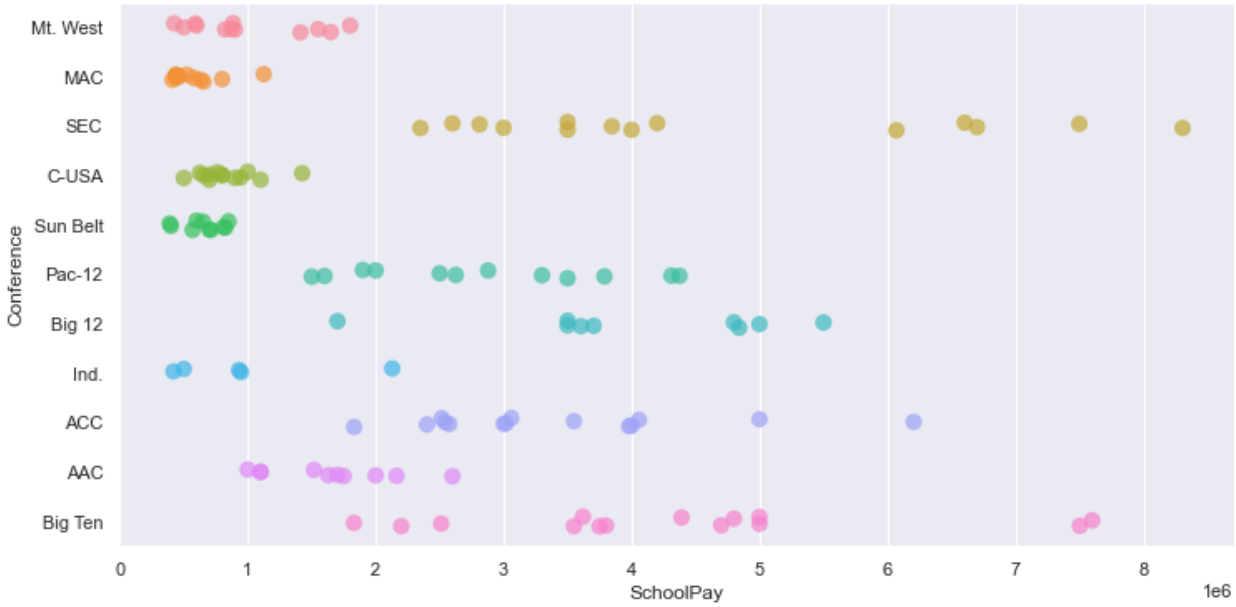
The final piece of the puzzle used for this analysis was graduation rate data. This information was pulled into a .csv file after filtering for 2014 football programs on ncaa.org. There were 248 rows in this file with 8 columns. All attributes aside from school, state, and graduation rates were dropped. Again, the data was fuzzy-matched on the School name and merged with the coach, stadium, and standing data sets.

To clean the data, all dollar signs and commas were removed from the data frame using the replace function. Double dashes (--) were replaced with NaN values to later be excluded from the linear models. All numeric fields were converted to float data types.

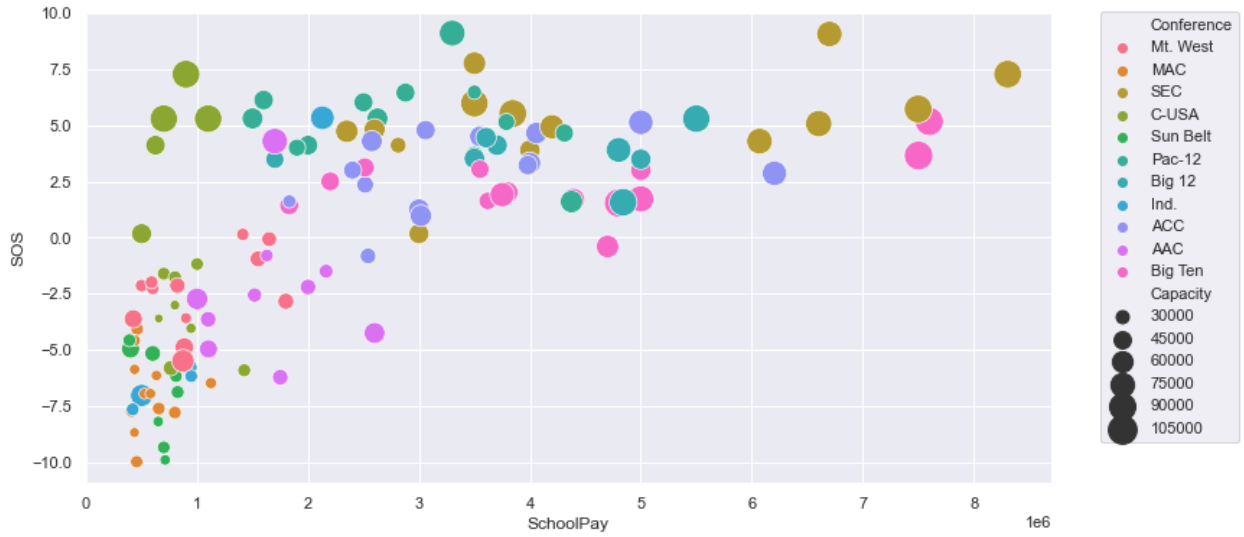
After the four data sets were cleaned and combined, 129 observations of 24 variables were ready for analysis. The SchoolPay variable was selected as the dependent variable of interest.

Visualizations

The first visualization created took a look at SchoolPay vs. Conference. It seemed that each school in a given conference tended to compensate their coaches a similar amount. One exception was the SEC which appeared to have two distinct salary groups. About half earned 2-4 million dollars per year, and the other half made 6-8 million. Every coach in the Mt. West, MAC, C-USA, and Sun Belt conferences earned under $2 million. There was a fairly large disparity in coach salary within the Pac-12, ACC, and Big Ten conferences.

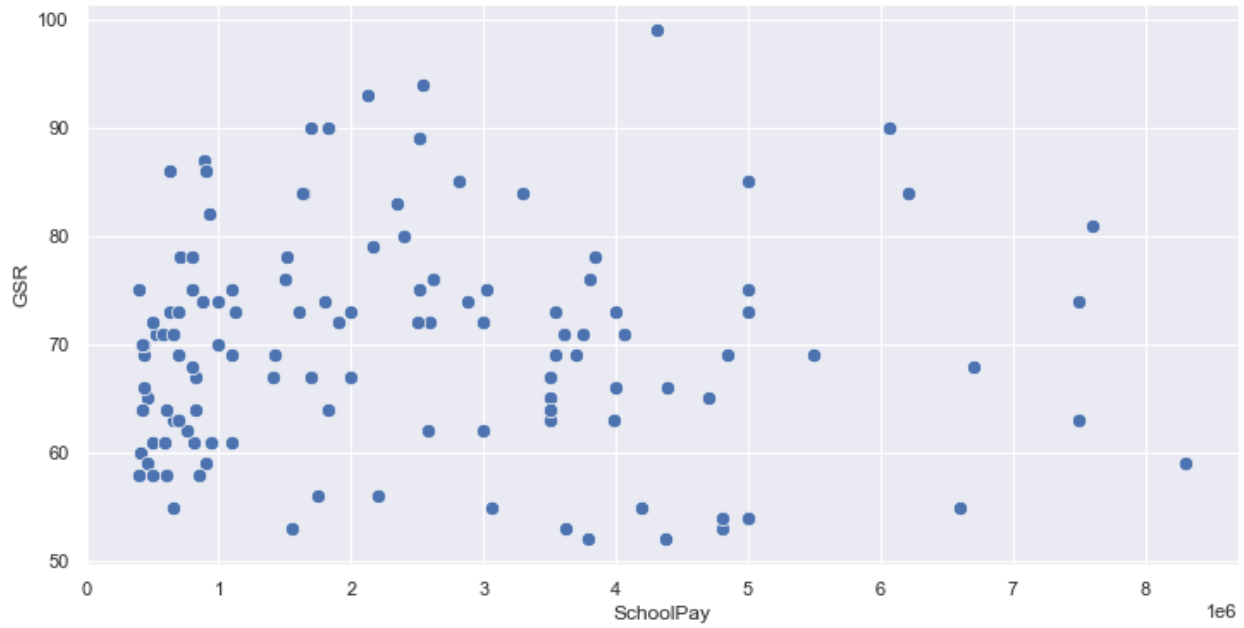


Looking at the Sports Reference data, strength-of-schedule seemed to positively correlate with school pay. The stronger the schedule, the higher the coaches salary. Schools with similar schedule strengths seemed to land near their conference peers. Again, the Mt West, MAC, and Sun Belt conferences were clustered around one another, appearing at the lower end of both the SOS and SchoolPay spectrums. Additionally, stadium capacity seemed to be related to salary. Schools with larger stadiums appeared to pay their coaches higher salaries. With more money to be made from ticket sales, these big venue schools could afford to pay their football coaches more money than their smaller counterparts.



After plotting SchoolPay to GSR (graduation rate), there seemed to be no obvious relationship between the two variables.

The only school dropped from this data set was Georgia, which had a GSR value of 0%. This left all remaining schools with GSRs between 52 and 99%.



Models

The combined data set was split into a testing and a training set using the uniform.rvs function. Two thirds of the data (84 rows) was put into the training set and the remaining third (44 rows) was reserved for testing.

Since conference and capacity seemed to be related to the SchoolPay value, these metrics were included in all data models. Each model was built with the ols function from the statsmodels.formula.api package. Null values were dropped from the models with the missing = “drop” parameter.

In addition to Conference and Capacity, the first model was created using the SRS value in lieu of the attributes it encompassed (SOS, Off, and Def). Graduation rate was also included in the first linear model. This yielded a r-squared value of 0.772, meaning this model accounted for over 75% of variability in the data.

SchoolPay ~ SRS + Conference + Capacity + GSR:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.772
Model:	OLS	Adj. R-squared:	0.728
Method:	Least Squares	F-statistic:	17.49
Date:	Sat, 24 Jul 2021	Prob (F-statistic):	1.05e-16
Time:	16:00:38	Log-Likelihood:	-1215.4
No. Observations:	81	AIC:	2459.
Df Residuals:	67	BIC:	2492.
Df Model:	13		
Covariance Type:	nonrobust		

Using the same parameters, but replacing SOS with its components (SOS, Off, and Def), the r-squared value increased slightly to 0.783. The p-value for SOS was a bit high at 0.618, but was kept in the model as it seemed to strongly align with SchoolPay based on the visualizations above.

SchoolPay ~ SOS + Off + Def + Conference + Capacity + GSR:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	15.60
Date:	Sat, 24 Jul 2021	Prob (F-statistic):	4.49e-16
Time:	16:00:38	Log-Likelihood:	-1213.5
No. Observations:	81	AIC:	2459.
Df Residuals:	65	BIC:	2497.
Df Model:	15		
Covariance Type:	nonrobust		

Removing the graduation rate dropped the r-squared value slightly to 0.780.

SchoolPay ~ SOS + Off + Def + Conference + Capacity:

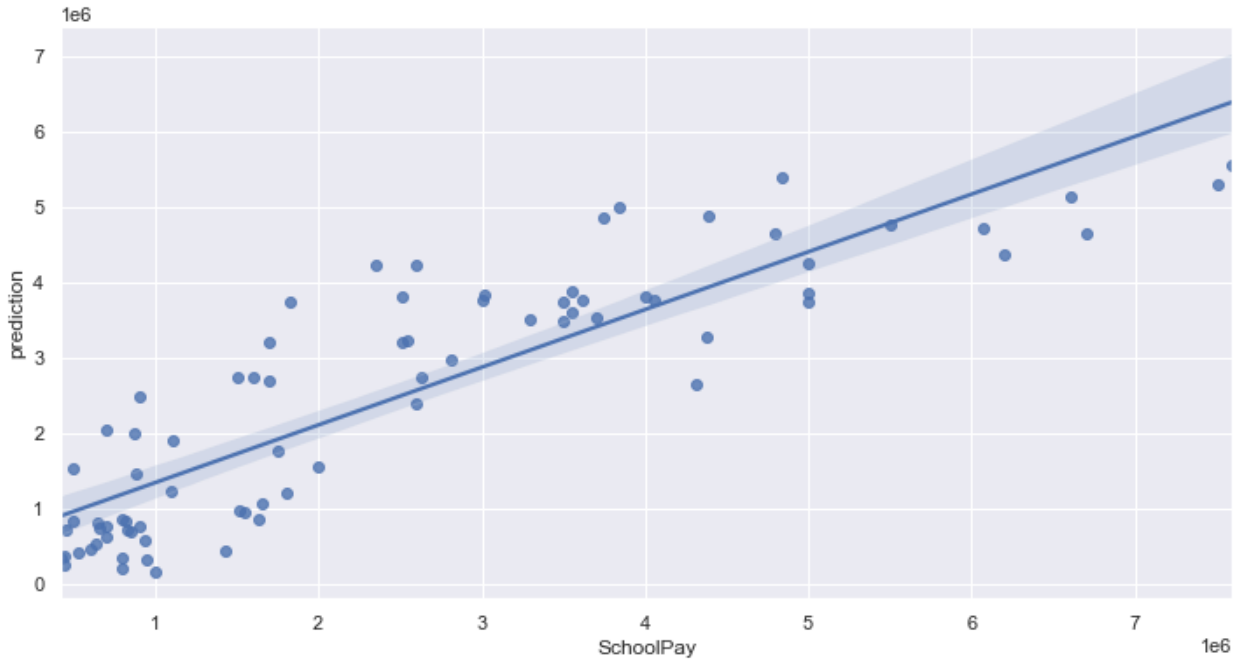
OLS Regression Results

Dep. Variable:	SchoolPay	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	16.74
Date:	Sat, 24 Jul 2021	Prob (F-statistic):	1.48e-16
Time:	16:00:38	Log-Likelihood:	-1214.0
No. Observations:	81	AIC:	2458.
Df Residuals:	66	BIC:	2494.
Df Model:	14		
Covariance Type:	nonrobust		

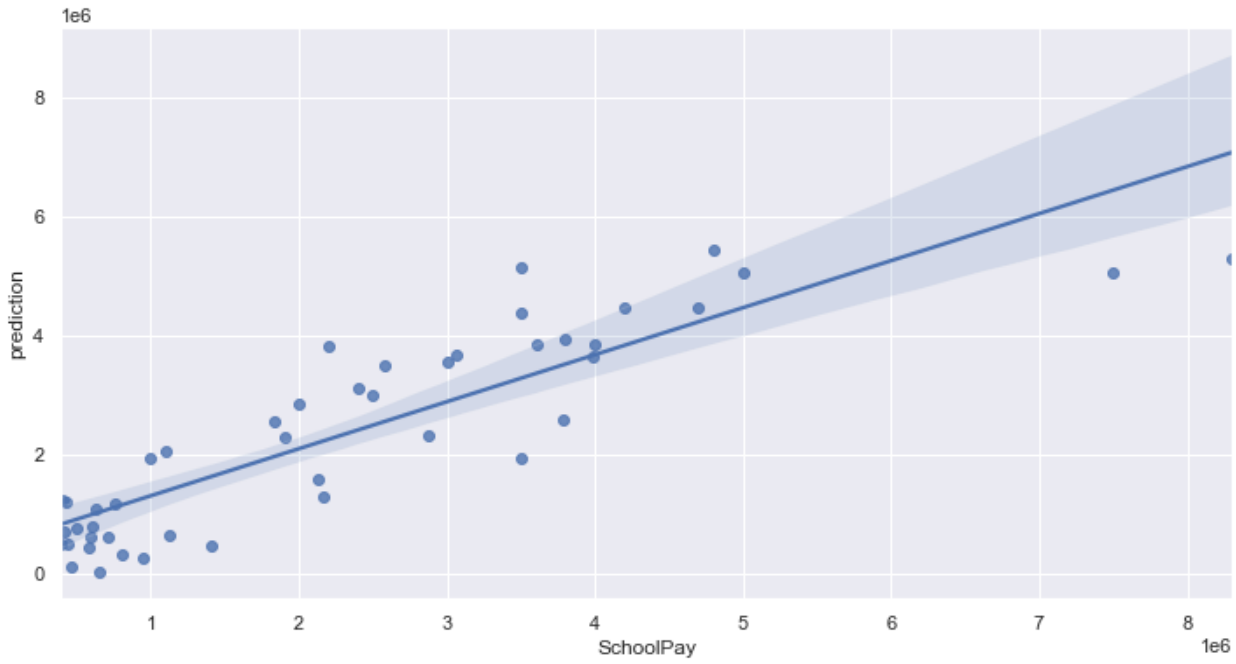
Dropping Conference from the model dropped the r-squared value about 20%. This indicated that Conference was the most important variable in determining football coach salary.

The second model which included SOS, Off, Def, Conference, Capacity, and GSR was chosen to predict the salary for the future Syracuse football coach because it had the highest r-squared value. The regression plots for this model, run on the training and testing sets, are below.

Training Data:



Testing Data:



Results and Recommendations

Running the second linear model on the Syracuse row of data, the model predicted that the next Syracuse football coach should be paid an annual salary of **$2,368,859.00**. This is slightly below coach Dino Babers's salary of $2,401,206.00 from the coaches9 file.

After updating the conference to Big Ten, the recommended salary increased to **$2,882,778.00**. This could be because there are a couple of coaches in the Big Ten that make significantly more than those in the ACC, at about $7.5 million, moving up the average Big Ten salary.

Without any other data from the Big East conference, it's difficult to make a salary recommendation using the second linear model. To make an accurate prediction, it would be beneficial to gather data from other schools in the Big East and re-run the model with the additional data.

Conclusion

It's important to recognize that a couple of the variables in this model are based on past coaching performance. Points scored and opponent points scored are likely impacted by the coach during a given season. However, it's unlikely an established program like Syracuse would pursue a coach with no NCAA experience.

Based on the variables that are attributed to the university such as graduation rate, strength of schedule, conference, and capacity, it's possible to provide a reasonable estimate of a given university's football coaches salary.

To create a better model with a higher r-squared value, more analysis should be conducted to see if there are additional variables that may impact a NCAA coaches salary, such as program donations or university ranking.