Jay Mahapatra

Lab 3-How can we recommend the best salary for our next head football coach?

IST 718- Big Data Analysis

1/26/2019

**Goal:**

The main goal of this report is to most accurately predict what the salary of a Syracuse coach should be. Through the analysis the report should also show what factors play the biggest roles in predicting a coaches’ salary including win percentage, graduation rate of the football students, rank, capacity of stadium, and the conference in which the coach plays in.

**Hypothesis:**

In an ideal world, as Bill Thomas insinuated in case study two, the coaches’ salary should be somewhat tied to academic measures, however, prior to reading this case study I believed win percentage and points per game played the biggest factor in paying a coach. After reading the case study though I know conclude that capacity and win percentage play the biggest factor in a coaches’ salary.

**Data Sources:**

I ended up pulling in 4 data sources:

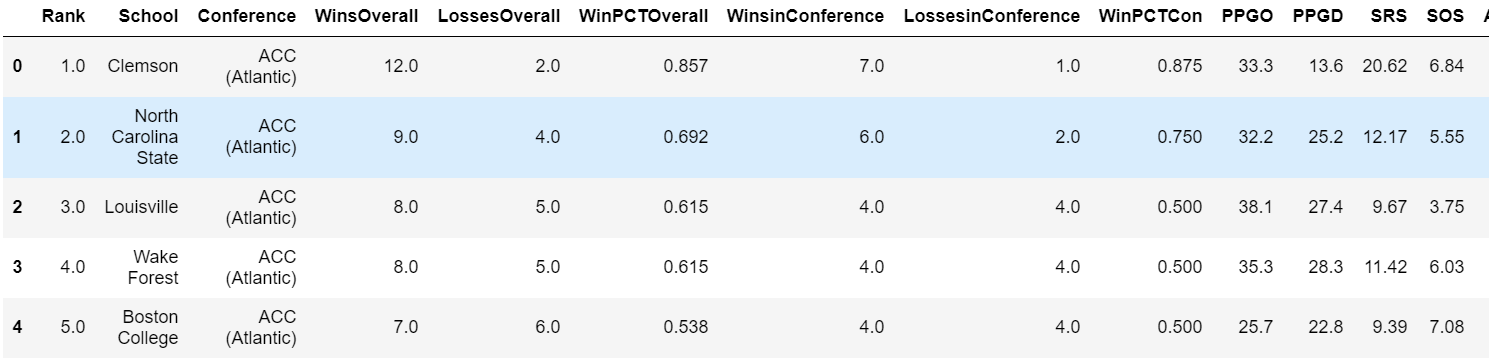
The Coaches9 data set which included the total salary of the coaches:



A table from collegegridirons.com that had information about the capacity of stadiums in relation to the colleges that play there:

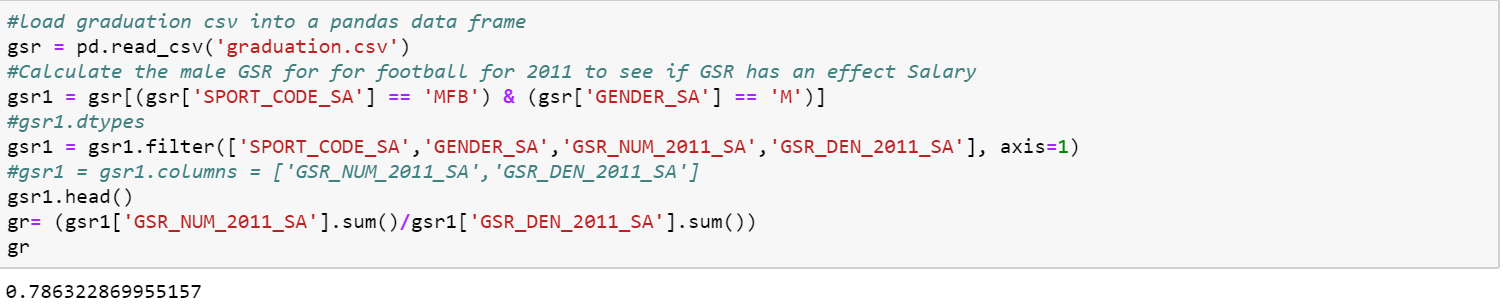


A table from sports reference.com that provided information around 2017 win/loss record, ranks, win percentage, and a couple other metrics:

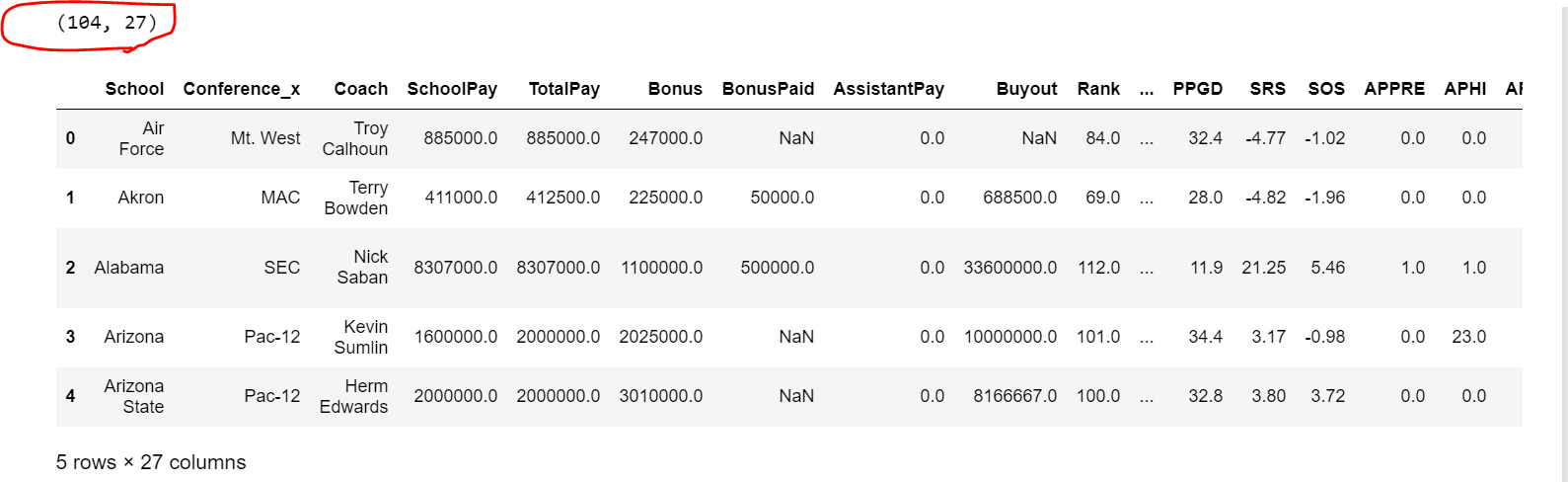




I also calculated a graduation rate from data provided by the NCAA:

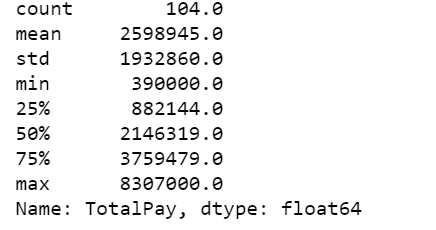


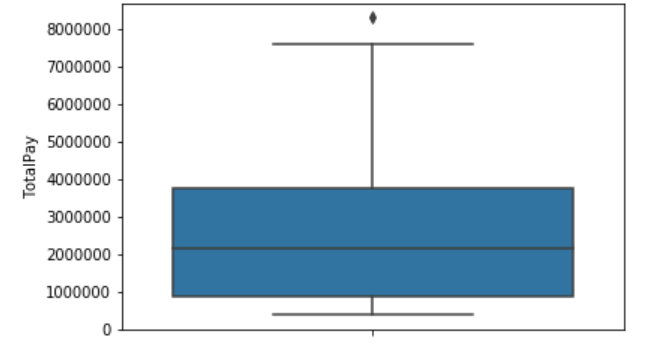
I then merged all the data together into one “Final” dataset consisting of 104 schools from the original 129.



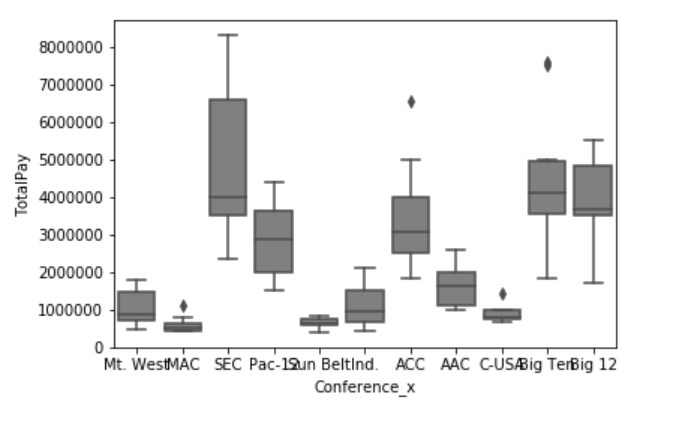
**Data Exploration:**

The initial data shows that there is a wide range in the pay of all the coaches ranging from aprox. $390K to $8.3M dollars but most of the coaches get paid between $882K-$3.75M.





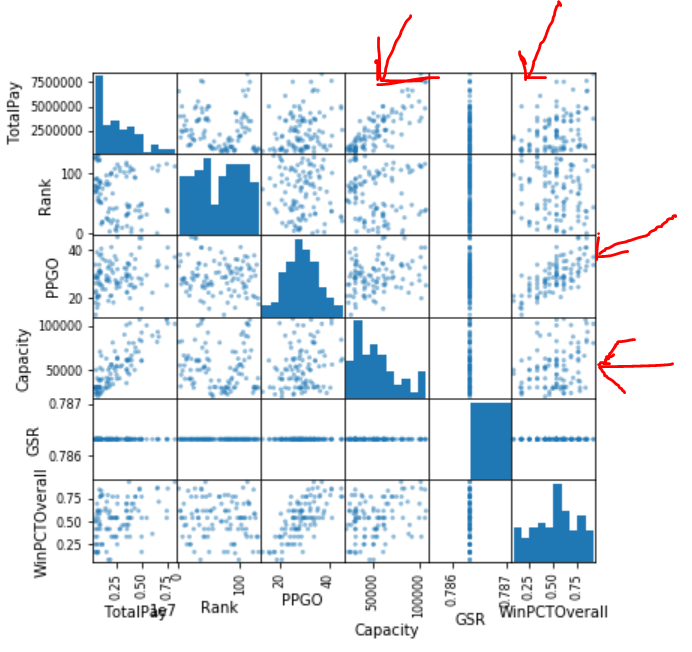
Looking at the range of pay by conference, it shows that there are big differences in paygrade between conferences. Initially suggesting that conference may play a significant factor in total pay, however as I find out later in the report, it does not play as big of a role as this graph suggests.



This matrix plot shows a lot of strong positive relationships. Some to note include:

* Total Pay and Capacity
* Total Pay and Overall Win Percentage
* Offensive Points per Game and Win Percentage
* Capacity and Win Percentage Overall

Increased Capacity and Win Percentage both seem to have strong correlations with Total Pay of a Coach.



**Questions/Results:**

***What schools did we drop from our data, and why?***

In total, I ended up cutting 25 schools from my final data set. 4 schools were initially cut off the coaches’ data set as there was no salary data for them.

* Baylor
* BYU
* Rice
* SMU

I could have put in the mean salary for them but decided to completely omit them as 3 of the schools would have had salaries well below the mean while Baylor would have had a salary well above the mean possible an outlier.

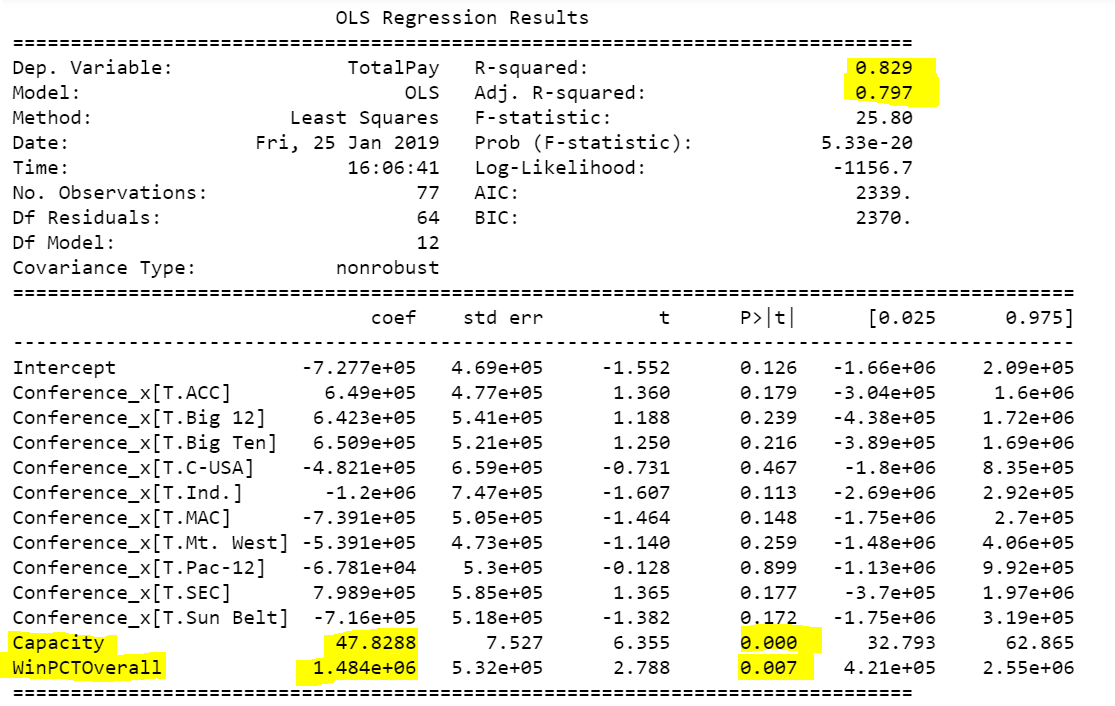
Another 21 schools were lost in the merge process of all the data sets but in looking at the schools that were dropped none of them were not a significant loss.

***How good is our model?***

***What is the single biggest impact on salary size?***

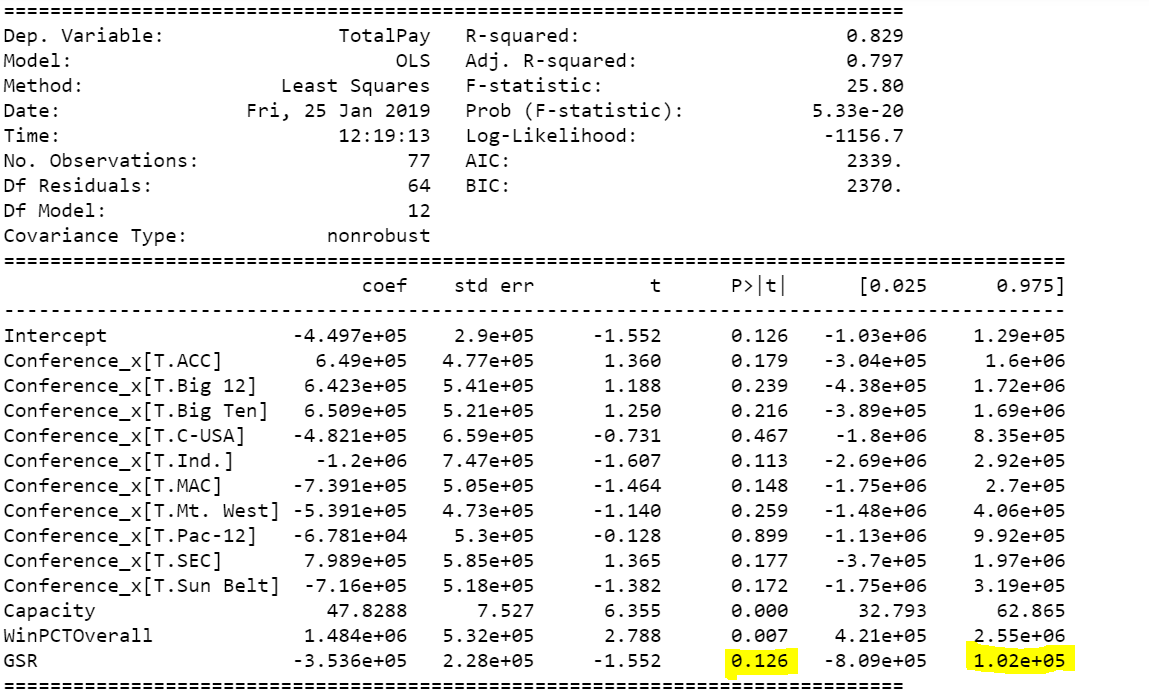
The model is reasonably good. Based on R-Squared values 80% of Coaches pay can be accounted for in the model.

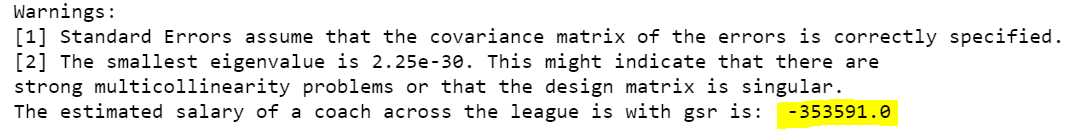
In addition, the model states that win percentage accounts for the biggest impact on a coaches’ revenue. Based on the coefficient, win percentage accounts for $1.4 Million dollars of a coaches’ total pay.



***What effect does graduation rate have on the projected salary?***

Very little. Looking at the P-value it is not a significant variable. In addition, another indicator is if all the variables are accounted for including GSR the coaches would be paying the school due to graduation rate.



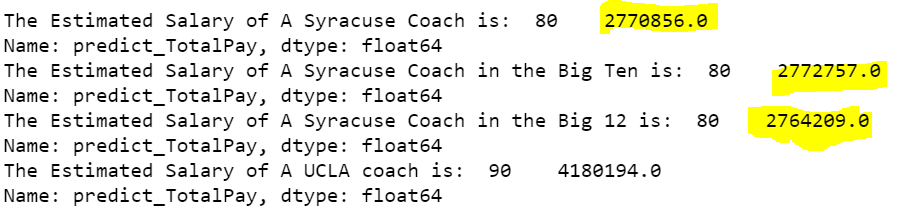


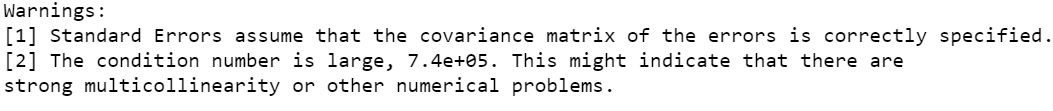
***What is the recommended salary for the Syracuse football coach?***

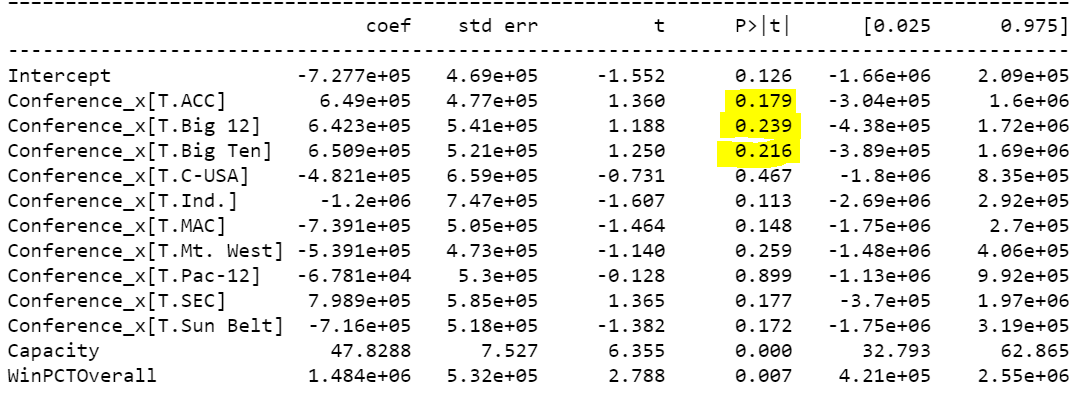
***What would his salary be if we were still in the Big East?***

***What if we went to the Big Ten?***

The recommended salary of a Syracuse coach based on the model is $2,770,856 with a standard error of $700,000. The Syracuse coach should get paid in the range of $2.0M-$3.4 based on win percentage and stadium capacity. In addition, as the p-values show below conference is not a significant indicator of total pay and as such the predicted salary for being in the Big 10 and Big 12 are similar to the original prediction. (I also did UCLA as that was my under grad and I was just curious about it. Guess Chip is getting under paid).







**Conclusion/Next Steps:**

The results of my model align closely with my hypothesis in that win percentage is a big factor in a college football coaches’ salary. I did not expect capacity to play a factor but in thinking about the results it does logically make sense that the larger the stadium the more tickets would get sold and factor into a coach’s salary.

Although my model did account for 80% of the reason why a coach gets paid a certain salary there is still another 20% of variability that is unaccounted for in the model. As next steps, if I had more time, I would like to explore adding even more variables such as rank, start of the year rank, and win/loss percentage from the year before to reduce the unaccounted-for variability. In addition, I would also like to explore how a mixed linear model would do in predicting coach’s salary.

**References:**

<http://www.collegegridirons.com/comparisons.htm>

<https://www.sports-reference.com/cfb/years/2017-standings.html>

<http://www.ncaa.org/about/resources/research/shared-ncaa-research-data>

Also thanks to everyone in the class who pitched in and helped with python code and links to data!