**IST718 – Big Data Analytics**

**Lab 1**

**David Doman**

**Introduction**

This analysis attempts to determine how to best recommend the optimal salary for Syracuse’s next head football coach. Various data sets are combined to build a data frame that can be used to perform analysis. The goal is to determine the attributes that best can predict what a college football coach deserves to earn in his/her salary. The initial dataset used is the coach’s dataset, which contains basic information regarding coaches. Additional datasets including information regarding a football schools’ standings and record, the schools’ student athlete graduation rates and Stadium information will be combined and used for analysis.

The following questions will be the focus of this analysis:

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

o What would his salary be if we were still in the Big East? What if we went to the Big Ten?

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

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

o How good is our model?

o What is the single biggest impact on salary size?

**Datasets**

**Coaches Dataset**

The Coaches dataset contains 9 attributes with 130 records, each record representing the coach of a college football team. Attributes include School, Conference, School Pay, Total Pay, Bonus, Bonus Paid, Assistant Pay & Buyout.

Below is an example of the dataset:

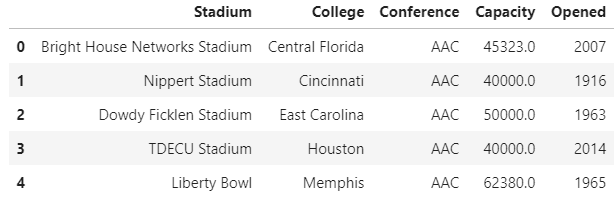


Certain attributes that were not numeric but needed to be were converted. School Pay, Total Pay & Bonus were converted, while Bonus Paid, Assistant Pay & Buyout were left alone since they will be removed eventually due to not being beneficial for this analysis. The ‘Bonus Paid’ attribute doesn’t make sense to use since contracts for each coach are different and when bonus’ are paid varies. ‘Assistant Pay’ has null values so this was not needed and was removed. ‘Buyout’ just simply provides no substantial insight into what a coach’s salary should be, so it was dropped.

**Stadium Dataset**

The Stadium dataset contains 6 attributes with again 130 records. Capacity is the only numeric attribute in this dataset and will be used in the analysis.

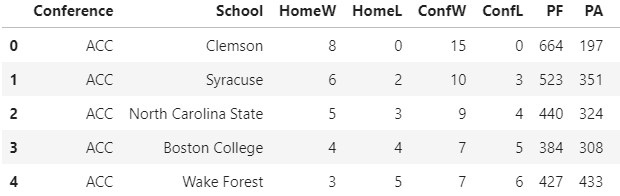
Below is an example of the Stadium dataset:



**College Records Dataset**

This dataset contains additional potentially useful attributes including Home Wins, Home Losses, Conference Wins, Conference Losses, Points For & Points Against.

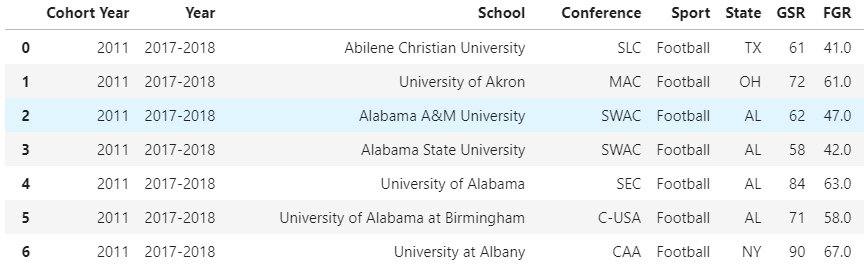
Below is an example of the dataset:



**Graduation Dataset**

This dataset contains potentially useful attributes such as Graduate Success Rate and Federal Graduation Rate. This information explains how each school’s student athletes perform in the classroom.

Below is an example of the dataset:

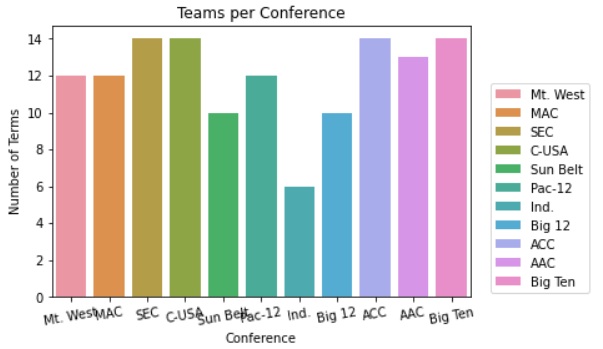


Due to differing naming conventions for conferences between the two datasets, for example, “Atlantic Coast Conference” vs. “ACC”, a single naming convention is needed so the two datasets can be merged. This conversion making naming conventions between the datasets uniform was also performed on team names. The next step was to merge the datasets into a single data frame so that it can be utilized for analysis. This was done by using the “FuzzyWuzzy” package, which allows to find certain characters in a string and match on them. Fuzzy string matching is the process of finding strings that match a given pattern. Basically, it uses Levenshtein Distance to calculate the differences between sequences. FuzzyWuzzy has been developed and open-sourced by SeatGeek, a service to find sport and concert tickets.

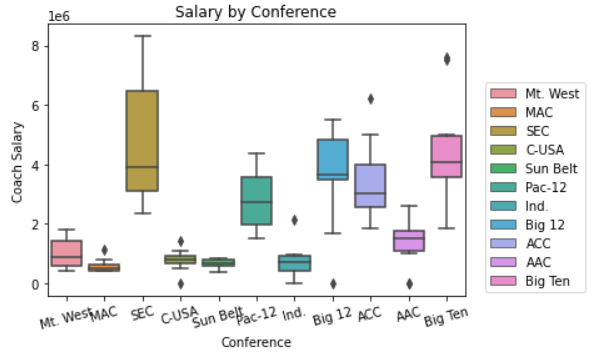
The FuzzyWuzzy package was utilized to combine the various datasets using the School attribute as the primary key. During the merging of datasets duplicate or unnecessary attributes appear in the final dataset. These were removed resulting in a dataset which is ready to use for analysis.

**Analysis**

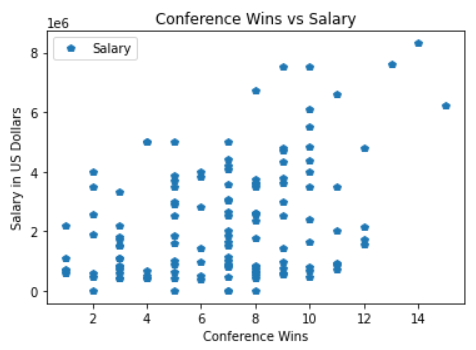
The first step is a simple bar graph produced to display the breakdown of teams by conference. This is just a way to get a general idea of how the data is distributed.



Another simple plot was produced to get a general idea of how salary by conference is distributed using a box plot.

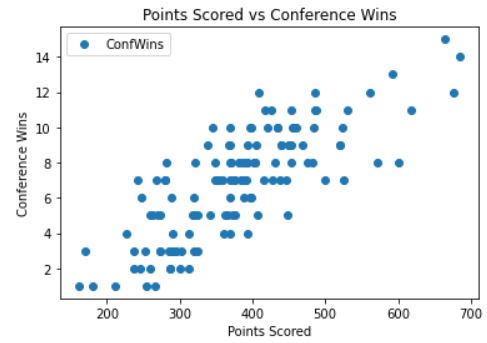


The initial exploratory analysis involved taking attributes that one would assume to be a contributing factor to a coach’s salary. These attributes were plotted against each other to see if any visible relationship exists.



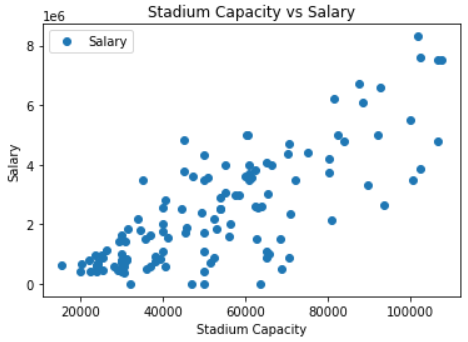
The first analysis plotted conference wins vs. Salary. There seemed to show a vague but somewhat visible relationship between conference wins resulting in a higher salary paid to a coach. Although, it also showed ambiguity with a 4 conference wins coach having an equal salary to a 12 conference wins coach. Overall, the data points are dispersed so much that no clear trend can be determined.

The next plot looked at the relationship between Points Scored and Conference Wins. There is a much clearer linear relationship between these two attributes shown in the graph below.



More points scored resulting in more conference wins is what would be expected. This makes sense and is not a surprise.

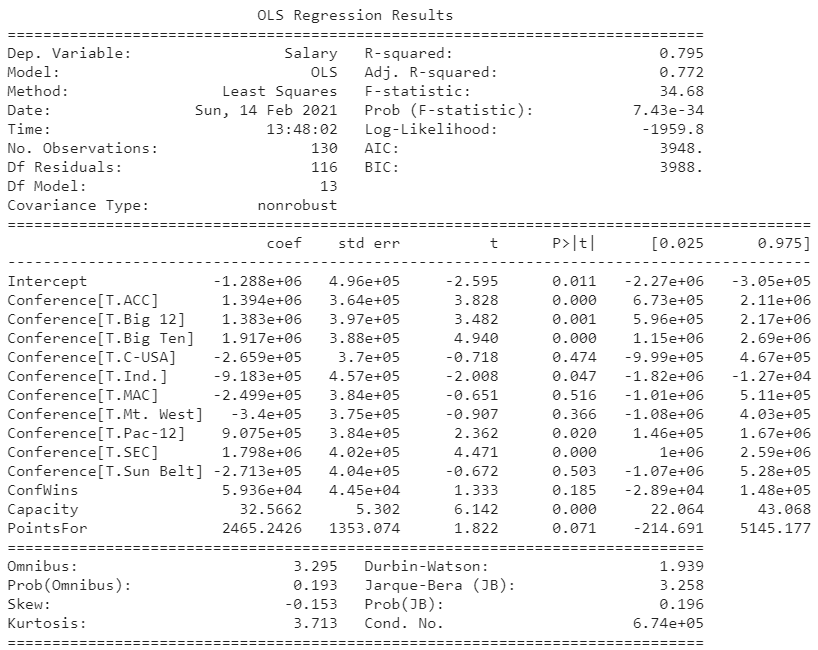
The next attributes compared was Stadium Capacity vs. Salary. Shown in the plot below there is a relationship between Stadium Capacity and Salary.



This makes sense since bigger schools with a large fan base tend to be the schools that have more success and have larger budgets to provide higher salaries to football coaches. Not all schools with large fan bases are successful but these institutions tend to have the money to provide for football coaches, resulting in higher salaries being provided.

**Models**

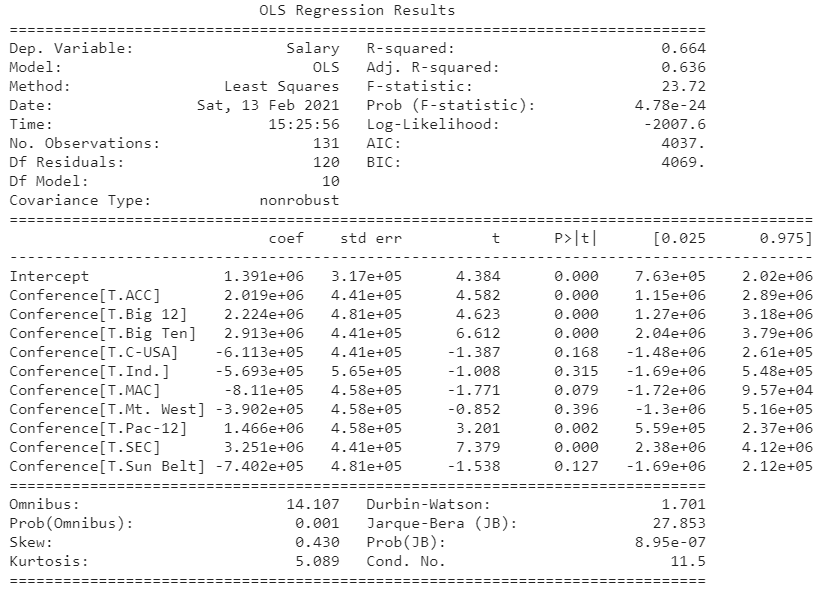
Now that the final dataset has been created and exploratory analysis has been performed it is time to build Linear Regression Models. The first model created uses the attributes Conference, Conference Wins, Conference Losses, Capacity, Points For, Points Against and Bonus. It was found that Bonus, Conference Losses and Points Against were very insignificant so they were removed, and the model was run again. Below shows the output:



As shown above, this model achieved an adjusted R-Squared of .772 and provided the insight that Capacity is a significant attribute in determining a coach’s salary. The following is the equation from this model:

*Salary = 5.936e+04\*Conference Wins + 32.5662\*Capacity + 2,465.2426\*Points For*

A second model was created to look solely at Conference. These results are shown below:

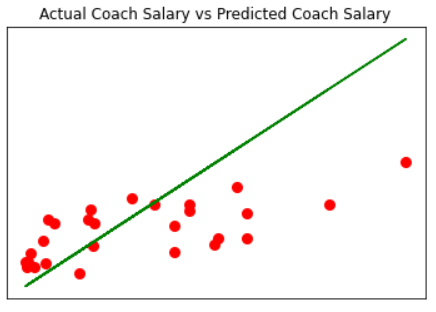


As shown above, being a coach in the ACC, Big 12, Big Ten, Pac 12 & SEC are all significant when determining a coach’s salary. This makes sense since the conferences listed previously are in the bigger conferences and they tend to have higher salaries regardless of if they are successful or not. This model resulted in an R-Squared of .636.

The final model utilized was the sklearn package creating a training set and a testing set using a 75% split for the training dataset. The attributes used in this model were Conference Wins, Conference Losses, Points For, Points Against and Bonus. The resulting model produced a .337 R-Squared and following formula:

*Salary = 4.14\*Conference Wins + 3.38\*Conference Losses + 1.75\*Points For + 4.09\*Points Against + 9.54\*Bonus*

This model did not produce a good R-Squared. The plot for this model is shown below:



**Conclusions**

From the sklearn model, it was shown the Syracuse head football coach is overpaid by about $1.1Million. If he were in the Big Ten, it seems Syracuse’s coach would be compensated accordingly at the current salary. From the analysis, Stadium Capacity and Conference best determined what a coach’s salary should be. Further analysis would provide a clearer insight into the appropriate salary’s coaches should earn.