Jennifer Williams

IST718 Section 2022-706

Lab 1 (pseudonym Lab 3)

**Lab Objective:**

This lab seeks to exercise data understanding to produce meaningful analytics and predict the salary of the Syracuse Head Coach.

**Data:**

Data provided for this exercise included Coaches9 data which incorporated records from 129 schools and specifically each head football coach’s names, pay, bonus options, bonuses paid, and buyouts. This data set also included Assistant Pay but with no values and therefore provided no value to the analysis. For data purposes and consistency, selected variable included TotalPay in lieu of SchoolPay as in most cases, the two values were very close if not identical. Combining the TotalPay and BonusPaid variables created a Salary variable that more accurately reflected the amount paid out by the collage or university to the coach.

It became necessary to incorporate additional data to create a wider overall dataset with greater analytic value and predictive capabilities. Data incorporated included Stadium Size to address capacity, football program revenue generated, graduation rate for the football program athletes, NCAA records/rankings and win percentage, and Conference performance in the form of a win percentage derived from Wins and Losses in conference play. This produced a data set distilled down to key variables: Coach, School, Conference, Salary, Conference Win Percentage [ConfWins], Total Game Win Percentage [Win\_Percentage], Rank, Stadium Capacity [Capacity], Revenue Generated [Revenue], and Graduation Rate [GradRate]. This range of data enabled me to evaluate if success in conference play had a greater impact than overall winning percentage and rank, if schools reward coaches who bring in big revenue, and if fostering a sense of commitment to education and getting their players to graduate translates to more money for coaches.

Data cleaning focused on a consistent naming of columns to facilitate merging of the data sets. Considerable data cleaning efforts went into establishing a consistent naming convention of the schools. Some differences were rather straight forward such as the use of State versus the abbreviation of St. Others included wildly different variations of school names between all datasets. Additional data cleaning included modifying the data types, cleaning the values within the records to be both consistent and machine readable such as removing punctuation in dollar amounts, and replacing blank or placeholder records with NaN values. Once cleaned, all the disparate datasets were merged and then pared down to focus just on the key variables.

Once merged, it was possible to deal with records containing NaN values. As these NaN values spanned multiple datasets, I opted to exclude those records from my final dataset. The following is a list of the 28 schools excluded from the final dataset and the data gap which prompted the exclusion. In cases with multiple categories of unavailable data for a given school, I noted only a single category in the table below.

Upon removing the excluded schools, the dataset included a total of 101 records.



**Table 1 - Table of Excluded Schools with Reason[[1]](#footnote-1)[[2]](#footnote-2)**

**Visualization(s):**

Starting first with some visualizations to help understand the data and how it looks helped me understand the interactions between the variables. Being a big college football fan, I immediately understood and appreciated the role of the conference on salaries and wins. Interlaced with conference strength is the ever nebulous “strength of schedule” argument used in the CFP Championship ranking. Some schools are however, and always will be, football schools. Those schools hang their athletic department hat on their football programs and build stadiums according, veritable monuments to the greatness of the sport.

**Figure 1 – Split of Boxplot Total Pay by Conference and Scatterplot of Stadium Capacity by Conference relating to Revenue Generated**

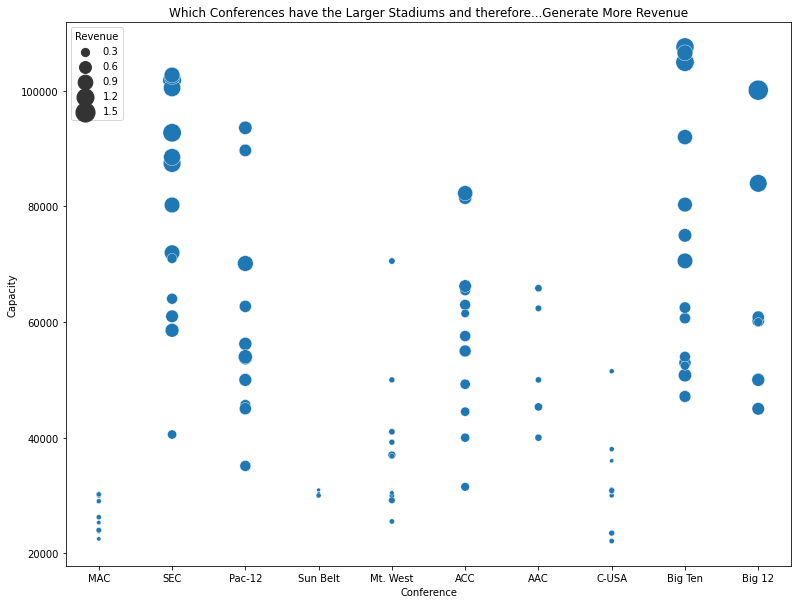
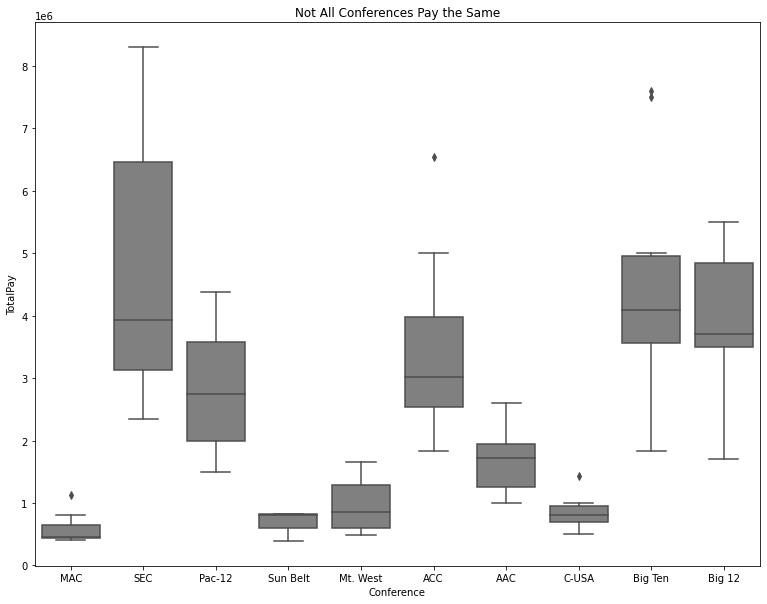
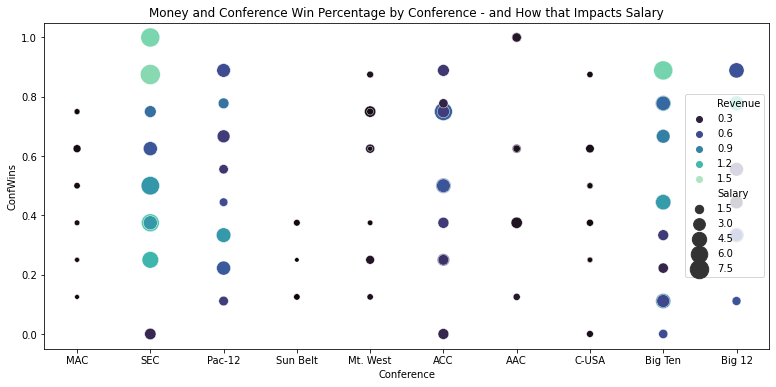


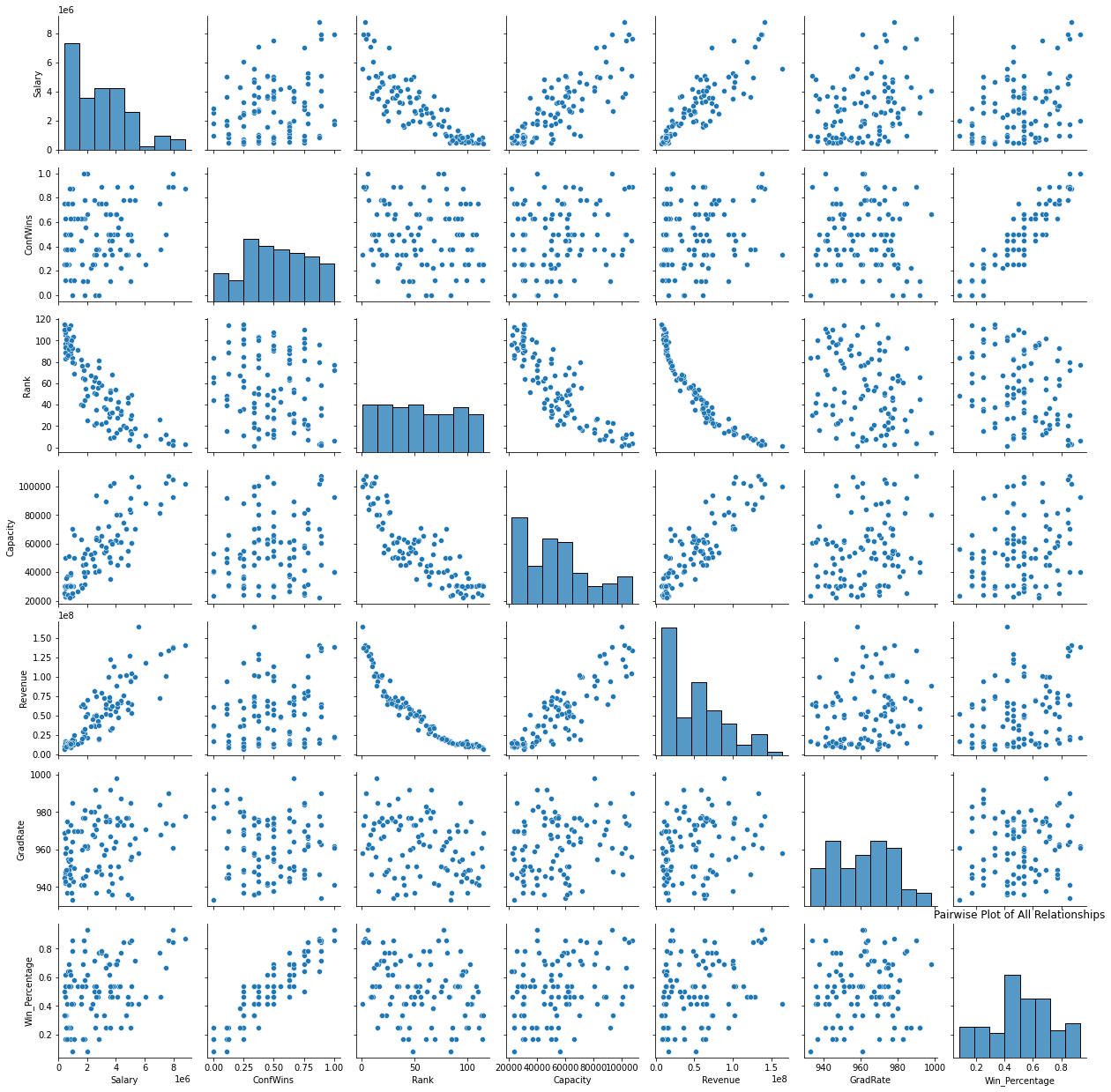
Figure 1 shows similar patterns in coach’s total pay by conference and stadium capacity by conference indicating that while pay stratifies strictly by conference, stadium capacity and revenue generated follow a very similar pattern. The conferences that build grand stadiums and generate revenue gravitate to the power conferences and pay their coaches handsomely.



**Figure 2 - Revenue and Conference Wins Impact on Salary - By Conference**

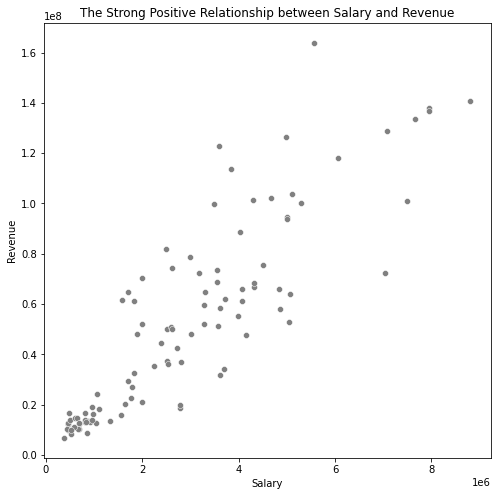
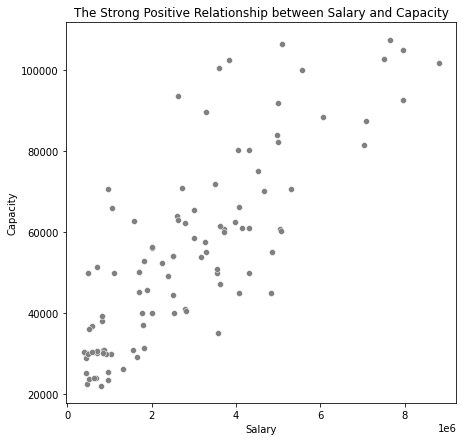
Figure 2 seeks to expound upon this understanding, specifically the role of the conference and revenue on a program, and evaluate it specifically under the auspice of coach salary and conference win percentage. While the SEC is still king, for the moment, some interesting observations start to emerge. In the middle, the ACC shows lower revenues than many of the other major conferences but salary that is commensurate with those larger more historically powerful conferences. The question then becomes, why might the ACC expend so much in salary when not generating nearly the same revenue as the SEC, Big Ten, Pac-12 or Big-12? I would argue, the ACC is trying to generate some momentum to convert a conference traditionally known for its basketball into a more diverse sports powerhouse. Additionally, football makes a school more money than any other sport.

With understanding of the data basics, it is now possible to begin looking closer at relationships.



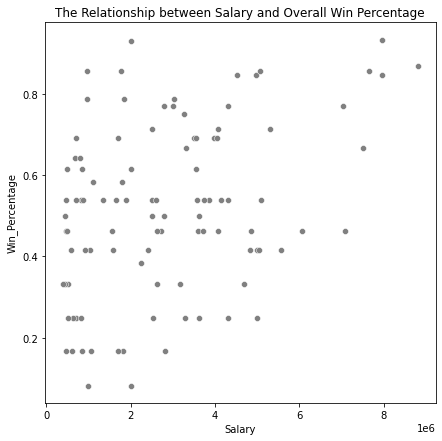
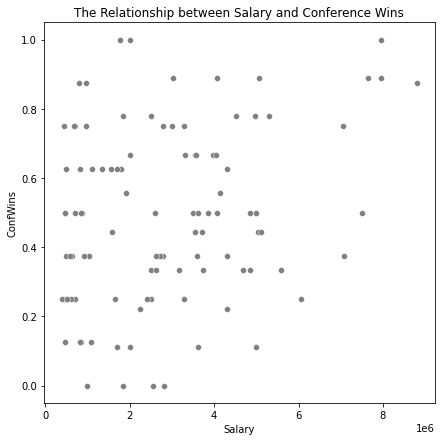
**Figure 3 - Pairwise Plot of Key Variable Relationships**

The pairwise plot shows a degree of multicollinearity between revenue and capacity as well as win percentage, conference wins and rank. As increased capacity tends to result in greater revenue, the linkage between these two variables is understandable. Figure 4 shows the extent of this by depicting the relationship between Salary and Revenue and Salary and Capacity side by side with a near identical pattern.

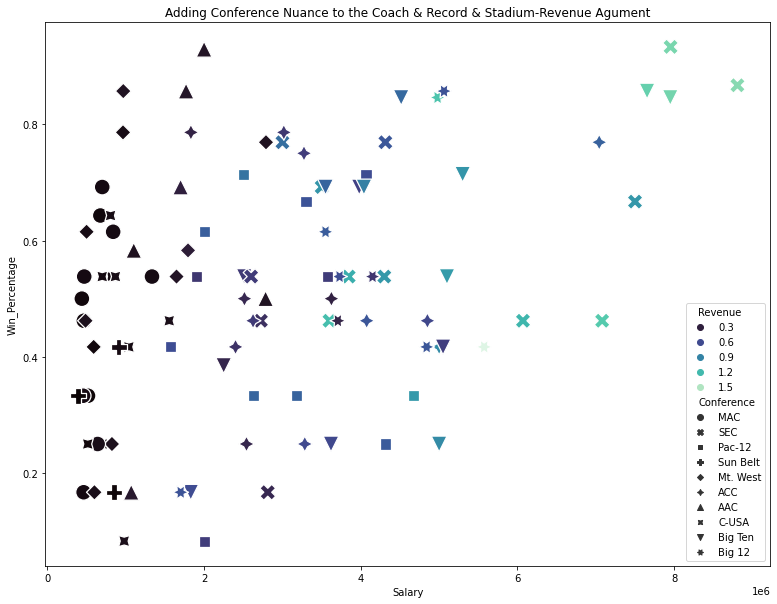


**Figure 4 - Side-by-side Scatterplots of Salary and Revenue:Capacity**

The relationship between Salary and overall Win Percentage and Salary and Conference Wins highlights another such example of this showing a strikingly similar pattern (Figure 5). Because of the multicollinearity between these variables, revenue:capacity and win percentage:conference wins:rank, I opted to create models with various combinations to assess individual impacts.



**Figure 5 - Side-by-side Scatterplots of Salary and Overall Wins and Conference Wins**



**Figure 6 - Adding Conference Distinctions to Salary/Winning/Revenue Evaluation**

By isolating which if these key variables have the greatest impact on Salary, it will then be possible to incorporate conference back more effectively into the equation and quantify the value of a conference (Figure 6).

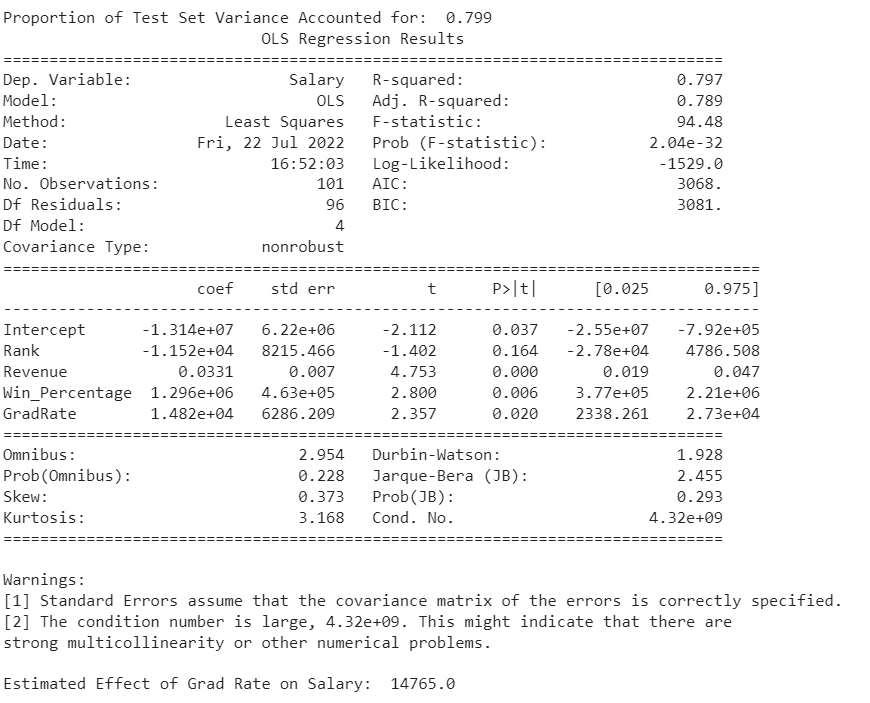
**Model(s):**

As mentioned, I utilized the key variables of Rank, Overall Win Percentage, Conference Win Percentage, Graduation Rate, Revenue Generated and Stadium Capacity to predict Salary which is the combination of Total Pay and Bonus Paid. I began modeling with all variables and then, based on the multicollinearity mentioned above, began generating select models with combinations of the variables. I opted to first strip out capacity and use revenue instead with the presumption that I would run all variation models and then test the best performing model with capacity instead of revenue to isolate the precise impact of those two variables. I created test and train subsets of the data to initially run each model. After running each model and because the dataset is rather small, I ran the model using my full data set in lieu of my test set to produce an out of sample result. Table 2 below includes a snapshot of each of the models including the adjusted R squared value from the out of sample summary.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Variables** | **Adjusted R Square** | |
| All\_Model | Salary ~ Rank/ Revenue/ Capacity/ ConfWins/ GradRate/ Win\_Percentage | | 0.789 |
| Select\_model1 | Salary ~ Rank/ Revenue/ ConfWins/ GradRate | | 0.787 |
| **Select\_model2** | **Salary ~ Rank/ Revenue/ GradRate/ Win\_Percentage** | | **0.789** |
| Select\_model3 | Salary ~ Revenue/ ConfWins/ GradRate | | 0.784 |
| Select\_model4 | Salary ~ Revenue/ GradRate/ Win\_Percentage | | 0.787 |
| Select\_model5 | Salary ~ Revenue/ ConfWins/ GradRate/ Win\_Percentage | | 0.785 |
| Select\_model6 | Salary ~ Rank/ Revenue/ Win\_Percentage | | 0.779 |
| Select\_model7 | Salary ~ Rank/ Capacity/ GradRate/ Win\_Percentage | | 0.761 |
| Select\_model8 | Salary ~ Rank/ Revenue/ GradRate/ Win\_Percentage/ ConfWins | | 0.787 |

**Table 2 - Model Description and Results**

While the model which contains all variables and Select\_model2 have the same adjusted R2, Select\_model2 has a far superior F-statistic indicating that the effects of the variables working together present a more statistically significant result. Below is the summary output for the Select\_model2 (Figure 7). I did swap capacity for revenue in Select\_model7, which otherwise mirrored model 2, to see if capacity or revenue were really the key. The output lends itself towards revenue in lieu of stadium capacity. The final model integrated conference into the optimal model to evaluate the impact of each conference along with the other critical identified variables (Rank, Revenue, Graduation Rate and Overall Win Percentage) on coaches’ salary.



**Figure 7 - Summary Output Select\_Model2**

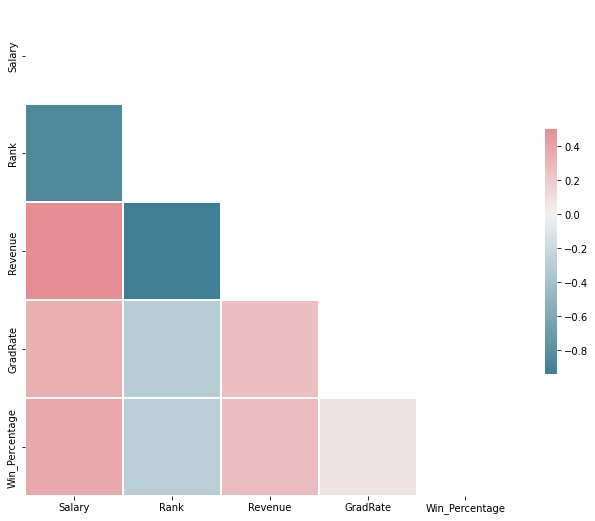
In evaluating how “good” my models performed, I compared the in and out of sample adjusted R squared values as well as the proportion of the test set variance accounted for by the variables considered. Table 3 below shows this information and shows that not only does Select\_model2 consistently outperform the other models but its proportion of variance accounting for the residual variance between salary and predicted salary appears sound, making it a good model. Select\_model2 wins.



**Table 3 - Model "Goodness" Evaluation**

**Analysis:**

What is the single biggest impact on salary size? The correlation matrix presented in Figure 8 below shows that Revenue and Rank have the greatest correlation to salary. As these two variables represent very different numeric intervals, comparing the coefficients and p values in the models may be misleading.



**Figure 8 - Correlation Matrix**

Therefore, it may be possible to distinguish between the importance of rank on salary versus revenue on salary by looking at model performance when they were or were not present. Models 3 through 5 excluded rank and had an average adjusted R2 of 0.785 which is very close to the winning model Select\_Model2 (0.789). This is juxtaposed when revenue is swapped out for capacity and the adjusted R2 dropped all the way from 0.789 to 0.761. For that reason, I believe **revenue generated is the single biggest impact on salary size**.

What is the effect of graduation rate on salary? Looking again at the correlation matrix, graduation rate is not strongly correlated. Based on the model with all variables as well as the “winning” Select\_model2, **the estimated effect of the graduation rate on Salary is $14,765**. With each point increase in graduation, a coach can expect a healthy bump in his salary.

What is the recommended salary (salary range) for the Syracuse Football Coach? What if Syracuse was in the AAC or the Big Ten? By adding the ordered conference variable into my Select\_model2, I isolated the “value” of each conference. By taking the coefficient and comparing it to the base conference rate it yields a target salary for each of the three conferences. Then by looking at the 0.025 and 0.925 lower and upper bound, it is possible to ascertain a salary range for our stalwart Syracuse coach in varied conference options. As postulated earlier, the ACC is throwing more money at their coaches than a similar coach with similar statistics would yield in a different conference, presumably to build the football brand within the ACC.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Salary |  |  | Salary Range | |
|  | Coefficient | $3,050,714.00 (Base Conf Rate) | 0.025 | 0.975 | Lower Bound | Upper Bound |
| ACC | $719,594.00 | $3,770,308.00 | -8.33E+05 | 2.21E+06 | $2,937,308.00 | $5,980,308.00 |
| Big Ten | $489,944.00 | $3,540,658.00 | -1.16E+06 | 2.14E+06 | $2,380,658.00 | $5,680,658.00 |
| AAC | $4,941.00 | $3,055,655.00 | -1.49E+06 | 1.50E+06 | $1,565,655.00 | $4,555,655.00 |

**Table 4 - Salary Predictions for Syracuse Coach**

**Conclusion:** Life is good for a football coach and life is good in the ACC. While it serves coaches financially to promote a sense of commitment and encourage graduation for their students, making money for their College/University plays a much bigger role. The impacts of these variables are not equal in all the conferences and that is likely to change next year with the coming conference shakeup. Viva la ACC and GO HEELS!!!!!

1. Conference exclusions fell into two situations, either schools that did not participate in conference play or won zero of their conference games and therefore a win percentage could not be calculated. [↑](#footnote-ref-1)
2. These two schools, for some reason, would not properly merge and the naming convention could not be made consistent across the multiple datasets despite repeated efforts. [↑](#footnote-ref-2)