The Million Dollar Question

Written By Laura C. Larregui For IST 718 Big Data Analytics

# Introduction

The goals of this laboratory are to demonstrate the ability to combine datasets and produce a meaningful analysis. In this case, the salaries of college football coaches are examined in order to give a recommendation for Syracuse’s next football coach.

## Questions

What is the recommended salary for Syracuse’s football coach?

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

What Syracuse if went to the Big Ten?

What schools were drop from the data, and why?

What effect does graduation rate have on the projected salary?

# Analysis & Models

This section includes information about the nature of the dataset, any changes made to it, and models made from such dataset.

## About the Data

The initial dataset was provided by the course professor. The dataset had 9 attributes and 129 records. Out of the 9 attributes, 6 were dropped since they did not provide enough information for the analysis. Thus, 4 additional datasets were needed to provide more insight. **Table 1** illustrates the essential variables compiled from all the imported datasets.

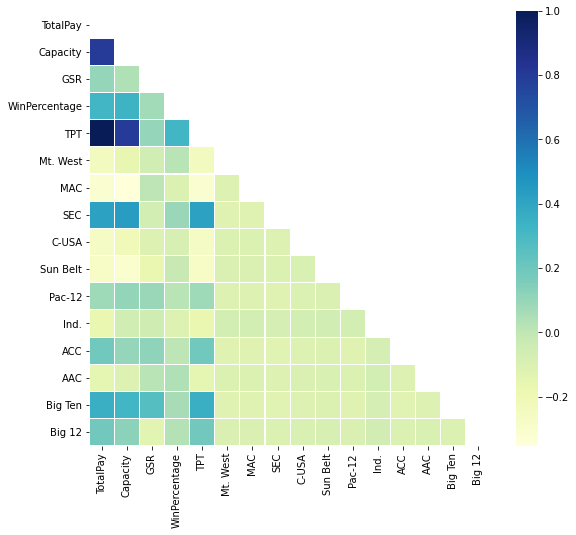
|  |  |
| --- | --- |
| **ATTRIBUTE** | **DESCRIPTION** |
| **School** | Names of the colleges/universities |
| **State** | Name of the US state in which the college/university is. |
| **Coach** | Name of the football coach |
| **Conference** | Name of the conference in which the college/university participates |
| **TotalPay** | Salary of the football coach |
| **Stadium** | Name of the stadium where the college/university plays |
| **Capacity** | Stadium capacity |
| **GSR** | Graduation Rate |
| **WinPercentage** | Winning Record Percentage |
| **TPT** | Total pay in thousands of dollars |
| **top5** | Identifies if the college/university is part of the top 5 conferences |
| **superfan** | Identifies if the college/university has an extraordinary fanbase |

Merging the datasets required extensive data cleaning. The most challenging task was matching school names. For this variable, the school names from the initial dataset were used as reference. For the rest of the data, datatypes were checked and changed, as necessary. There were a few records that had missing values for GSR, Capacity, and WinPercentage. Even though the missing data was minimal, records with missing values were dropped for the dataframe. Consequently, records which had ‘Charlotte’ ‘Texas-San Antonio’, ‘Southern Mississippi’, and ‘Liberty’ as School were dropped.

Lastly, two extra columns were created. The first one was top5 which identified if the school was part of the top 5 conferences of college football.[[1]](#footnote-1) The other column was superfan which identified if the school had a big fanbase for their football team.[[2]](#footnote-2)

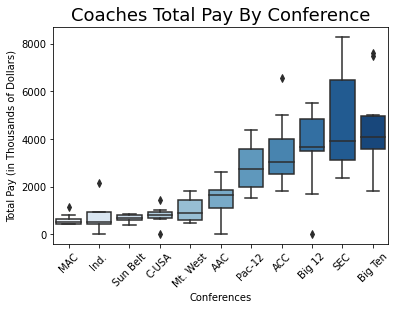


### Exploratory Data Analysis

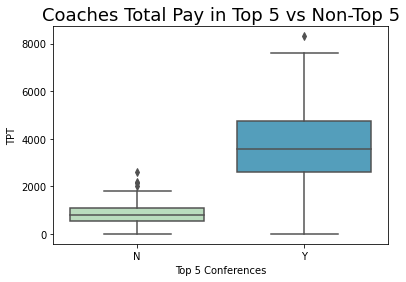


**Figure 1. Correlation Matrix**

A correlation matrix was made to see if there were any relationships between variables. From **Figure 1**, one can see that capacity has a strong effect on TotalPay. Moreover, most of the Top 5 conferences (SEC, Big Ten, ACC, Big 12, and Pac-12) have strong correlation with TotalPay and Capacity.

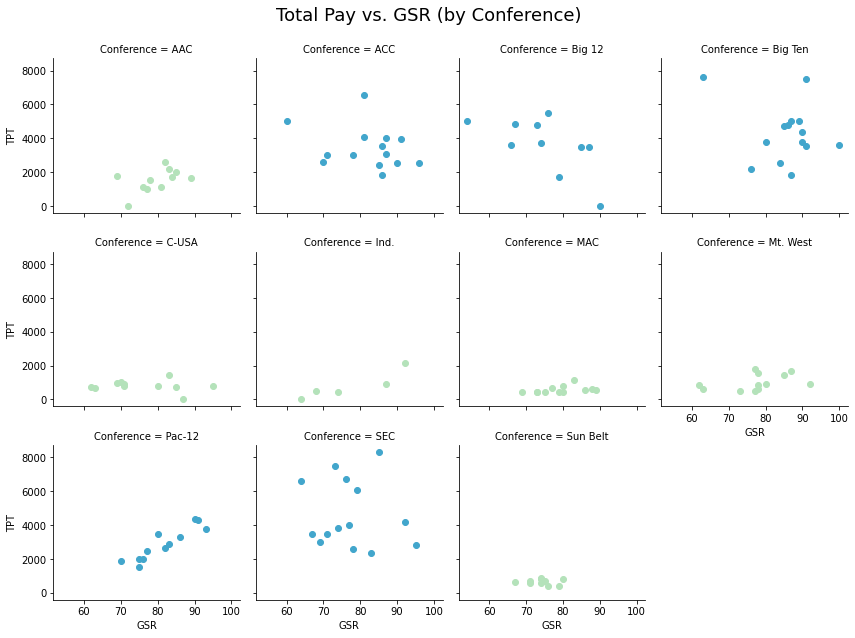


**Figure 2. Coaches Total Pay by Conference Boxplot**



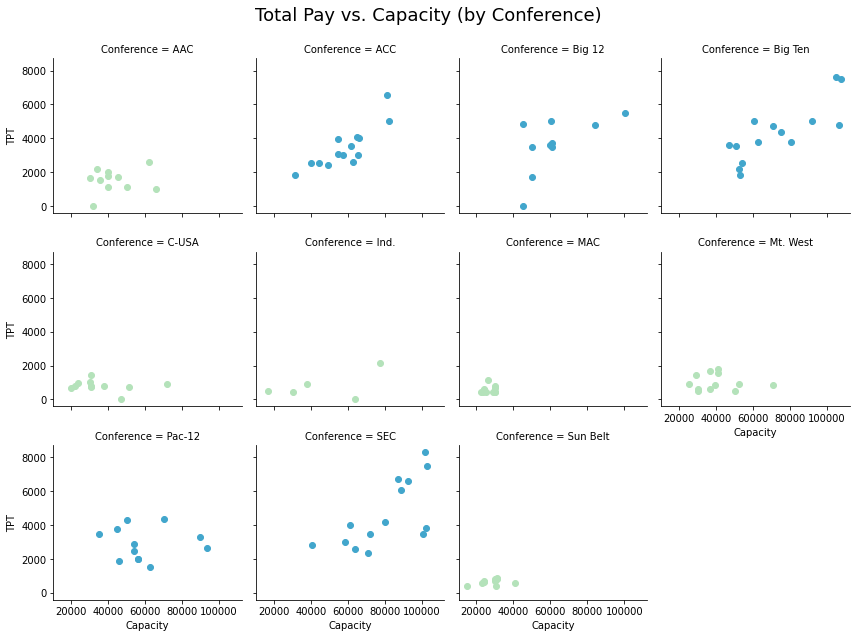
**Figure 3. Coaches Total Pay in Top 5 Conferences Boxplot**

The next visualizations made were boxplots. **Figure 2** shows that last five conferences (Pac-12, ACC, Big12, SEC, and Big Ten) have a greater median value of the TotalPay. Coincidentally, these 5 conferences are also the top 5. To emphasize this observation, **Figure 3** was made.

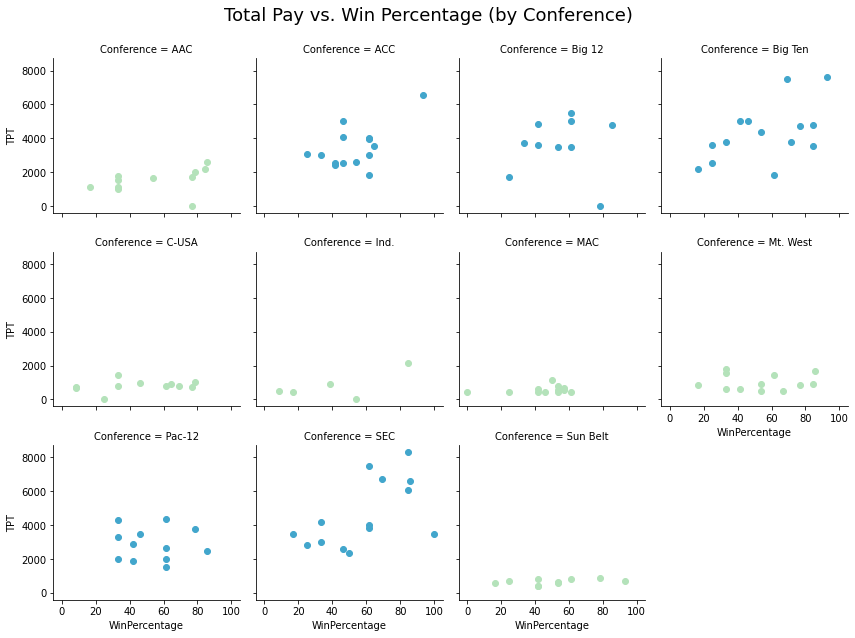


**Figure 4. Coaches Total Pay VS GSR by Conference Scatterplot**

Following the boxplots, scatterplots were made focusing in GSR, Capacity and Win Percentage. As one can see in **Figure 5** and **6**, there is a strong relationship between TotalPay and Capacity and Win Percentage. These relationships are strongly highlighted in the top 5 conferences. For GSR, the relationship cannot be discerned as easily. For most conferences, there is a minor increase in TotalPay when GSR increases.



**Figure 5. Coaches Total Pay VS Capacity by Conference Scatterplot**



**Figure 6. Coaches Total Pay VS WinPercentage by Conference Scatterplot**

## Models

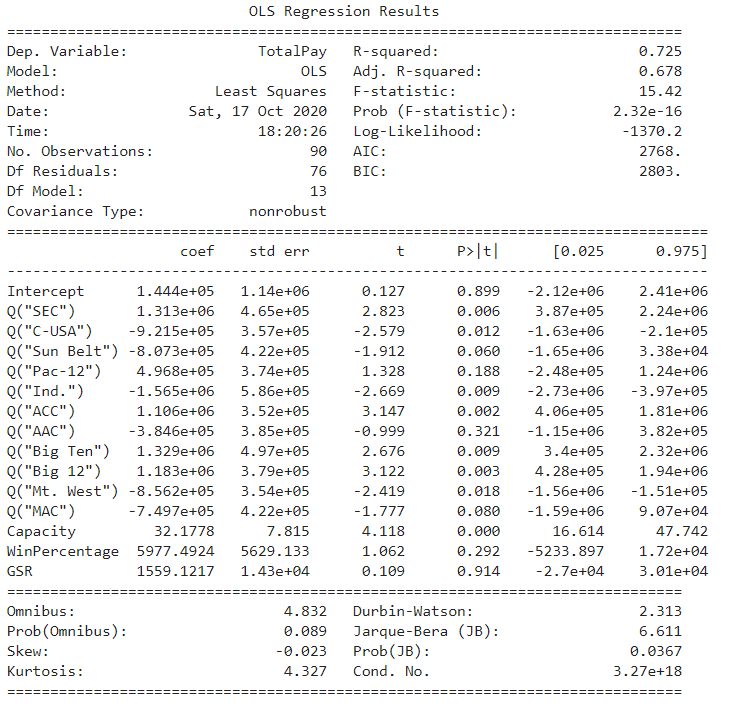
For this laboratory, linear regression was used to conduct multiple analysis. A linear regression model shows a relationship between a dependent variable and one or more independent variables.

In the linear regression formula from above,

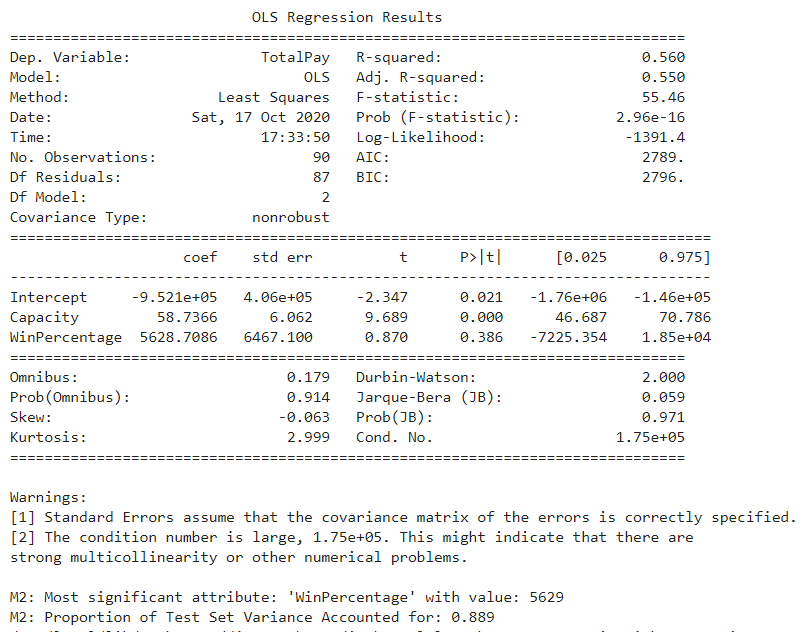
* **y** represents what is being predicted
* **a** is the constant
* **b** is the coefficient of x (slope)
* **x** is what is predicting the value of x

For the analysis conducted, cross validation was made where two thirds of the dataset (90 records) were used for training and one third (35 records) was used for testing.

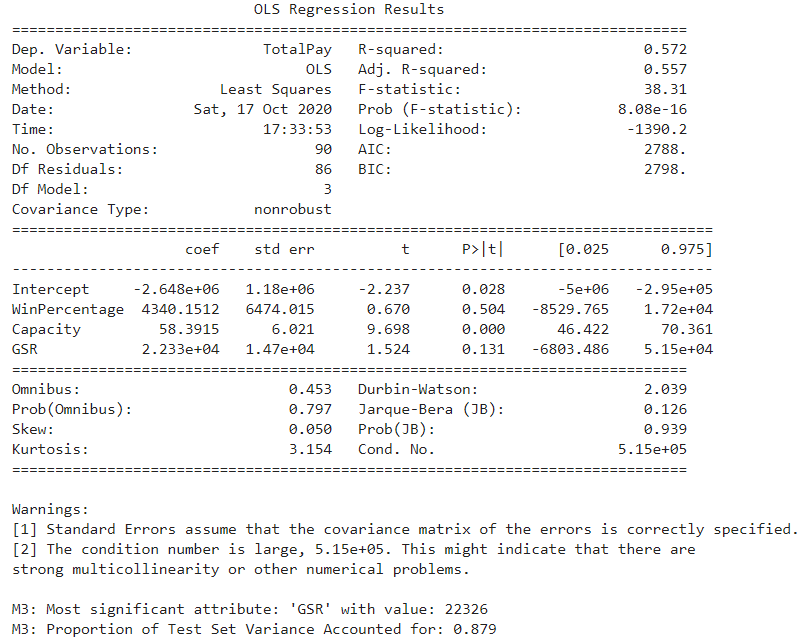
### Model 1: PayTotal ~ WinPercentage + Capacity + GSR + CONFERENCES



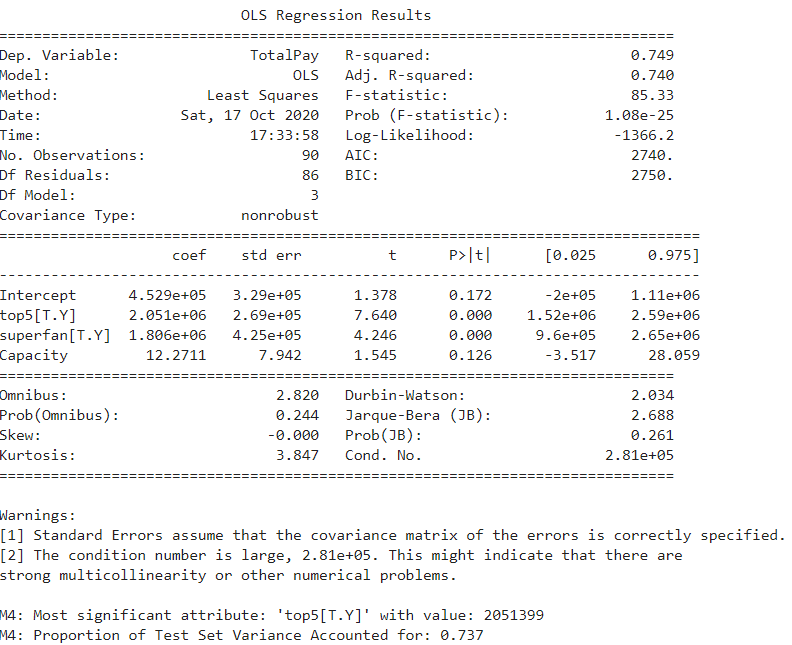
### Model 2: PayTotal ~ WinPercentage + Capacity



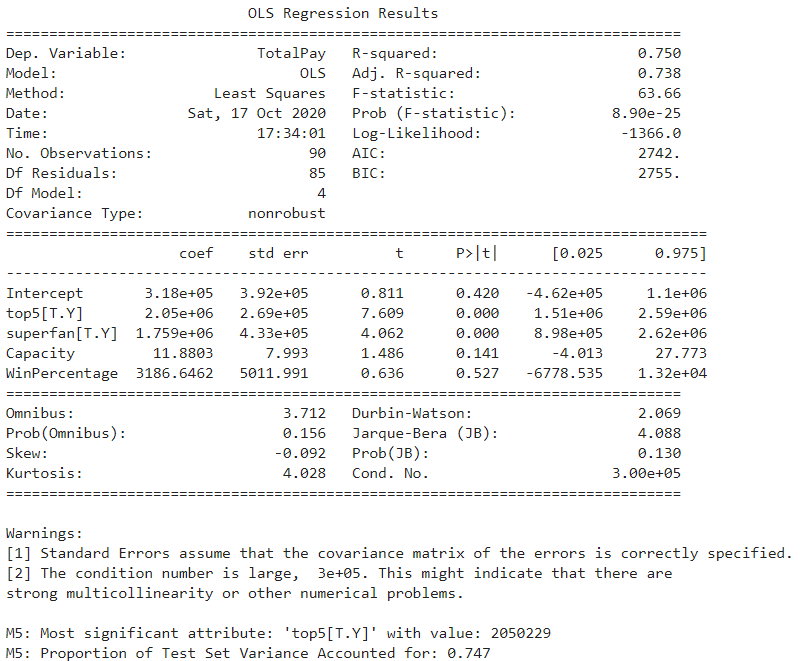
### Model 3: PayTotal ~ WinPercentage + Capacity + GSR



### Model 4: PayTotal ~ Capacity + top5 + superfan



### Model 5: PayTotal ~ WinPercentage + Capacity + top5 + superfan



# Results

The best model overall was Model 5. It scored an R-squared of 0.750 (75% Accuracy), which was the highest among all models. Model 4 was also a good model, with a 0.749 for R-squared. The only downside of Model 5 was that WinPercentage scored a p-value of 0.5, which is not ideal. Preferably, one would like to have lower p-values. Nonetheless, Top 5 was the variable with greater significance for TotalPay for both models. To make salary predictions in relation of conferences was Model 1 because it had the necessary coefficients and had a R-square value of 0.725.

# Conclusion

Currently Syracuse’s football coach has a salary of $2,401,206.00. Using Model 5, the recommended salary is $2,921,754.10. Between the two salaries, there is an 18% difference. Predicting the coach’s salary for the Big East was a small obstacle in the analysis. In the dataset used for this laboratory, Big East was not included. The reason for this is because Big East does not longer exist since 2013. It was succeeded by the AAC. Thankfully, the AAC was included in the dataset.

In order to get the salary prediction in the Big East, the coefficient for ACC (as shown in Model 1) was extracted from the recommended salary since Syracuse is part of this conference. Then, the coefficient for AAC was added. The result was $684,477.81, which is a huge difference.

To predict the salary in relation to the Big Ten conference, the same process was done with the only difference being that the coefficient for the Big Ten was added. The result was 2,398,525.62.

# References

Athlon Sports. (2020) College Football 2020 Conference Power Rankings. Retrieved October 16, 2020 from athlonsports.com/college-football/college-football-2020-conference-power-rankings.

College Grid Irons. (n.d). College Football Stadium Comparisons. Retrieved October 13, 2020 from https://www.collegegridirons.com/comparisons-by-capacity/

NCAA. (2019). Graduation Success Rate Retrieved October 13, 2020 from https://web3.ncaa.org/aprsearch/gsrsearch

TeamRanking. (2019). College Football Team Win Trends - All Games, 2019. Retrieved October 13, 2020 from https://betiq.teamrankings.com/college-football/betting-trends/win-loss-records/?season=2019

The Spun. (2019). The 10 Most Powerful Fanbases in College Football. Retrieved October 16, 2020 from https://thespun.com/college-football/the-10-most-powerful-fan-bases-in-college-football.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)