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.

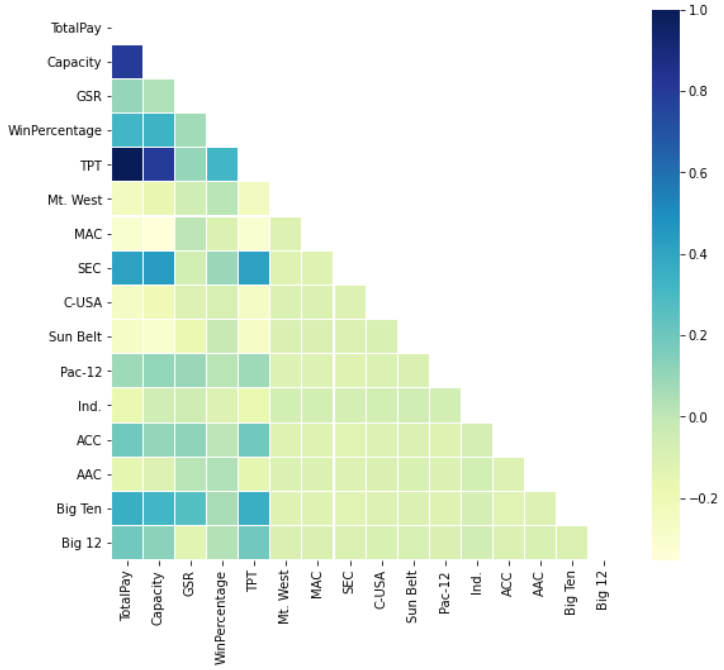
ATTRIBUTE	DESCRIPTION
School	Names of the colleges/universities
State	Name of the US state in which the college/university is.
Coach	Name of the football coach
Conference	Name of the conference in which the college/university participates
TotalPay	Salary of the football coach
Stadium	Name of the stadium 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] The other column was superfan which identified if the school had a big fanbase for their football team.[2]

EXPLORATORY DATA ANALYSIS



1

2

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.

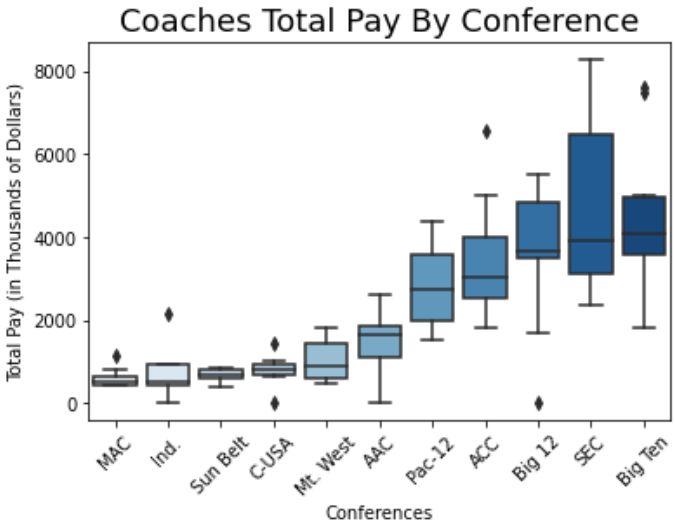


Figure 2. Coaches Total Pay by Conference Boxplot

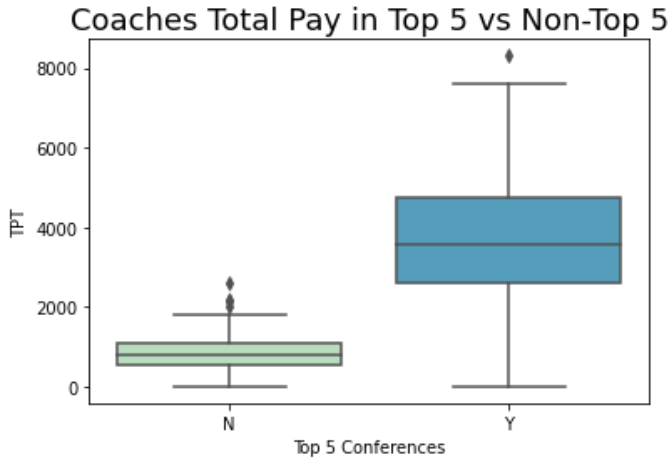


Figure 3. Coaches Total Pay in Top 5 Conferences Boxplot

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.

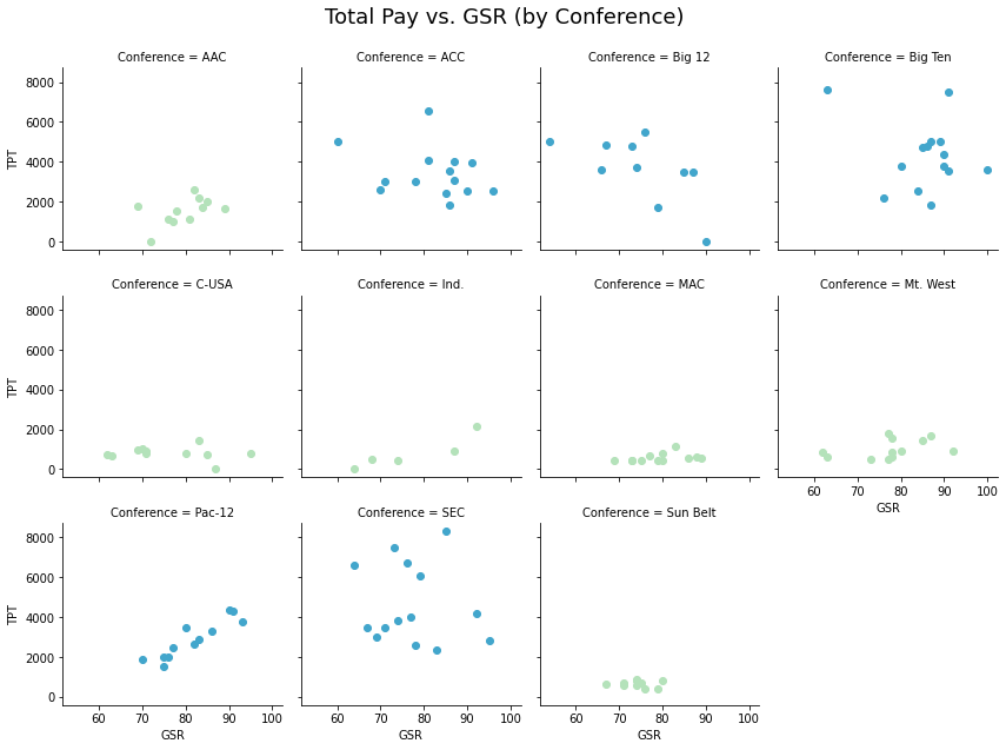


Figure 4. Coaches Total Pay VS GSR by Conference Scatterplot

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.

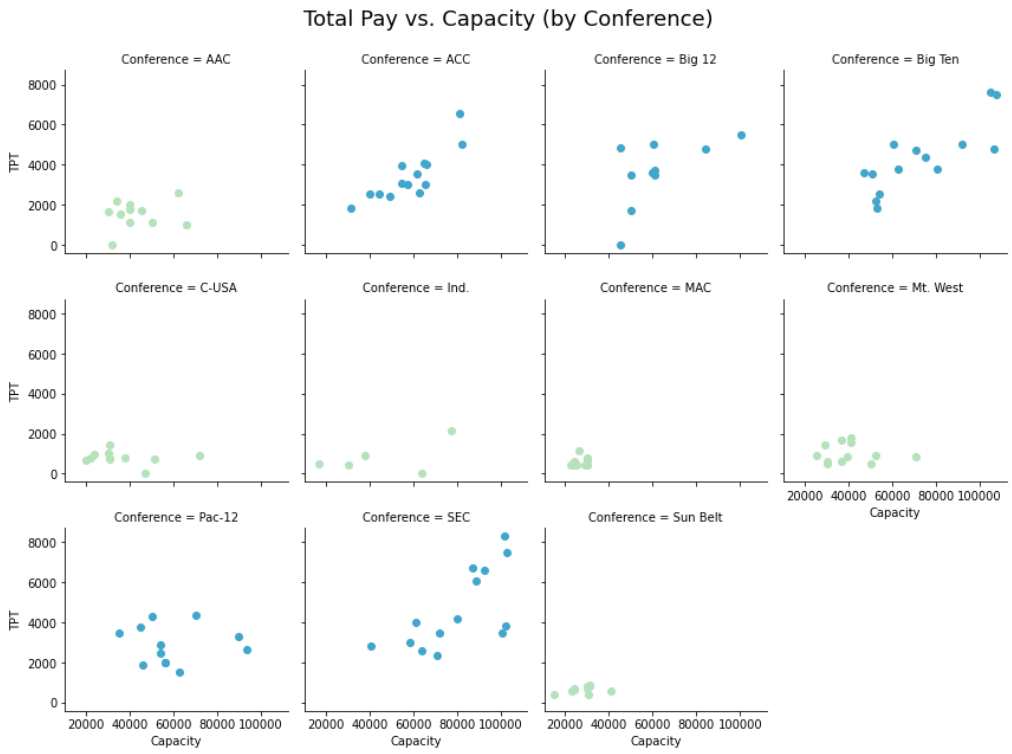


Figure 5. Coaches Total Pay VS Capacity by Conference Scatterplot

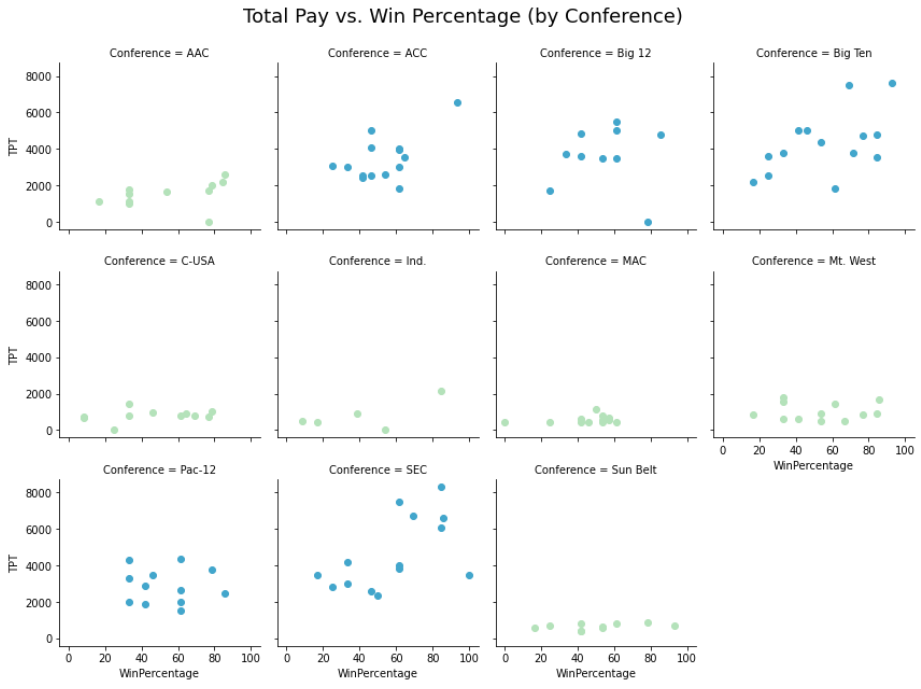


Figure 6. Coaches Total Pay VS WinPercentage by Conference Scatterplot

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.

$$y = a + bx$$

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

OLS Regression Results

Dep. Variable:	TotalPay	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.678
Method:	Least Squares	F-statistic:	15.42
Date:	Sat, 17 Oct 2020	Prob (F-statistic):	2.32e-16
Time:	18:20:26	Log-Likelihood:	-1370.2
No. Observations:	90	AIC:	2768.
Df Residuals:	76	BIC:	2803.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.444e+05	1.14e+06	0.127	0.899	-2.12e+06	2.41e+06
Q("SEC")	1.313e+06	4.65e+05	2.823	0.006	3.87e+05	2.24e+06
Q("C-USA")	-9.215e+05	3.57e+05	-2.579	0.012	-1.63e+06	-2.1e+05
Q("Sun Belt")	-8.073e+05	4.22e+05	-1.912	0.060	-1.65e+06	3.38e+04
Q("Pac-12")	4.968e+05	3.74e+05	1.328	0.188	-2.48e+05	1.24e+06
Q("Ind.")	-1.565e+06	5.86e+05	-2.669	0.009	-2.73e+06	-3.97e+05
Q("ACC")	1.106e+06	3.52e+05	3.147	0.002	4.06e+05	1.81e+06
Q("AAC")	-3.846e+05	3.85e+05	-0.999	0.321	-1.15e+06	3.82e+05
Q("Big Ten")	1.329e+06	4.97e+05	2.676	0.009	3.4e+05	2.32e+06
Q("Big 12")	1.183e+06	3.79e+05	3.122	0.003	4.28e+05	1.94e+06
Q("Mt. West")	-8.562e+05	3.54e+05	-2.419	0.018	-1.56e+06	-1.51e+05
Q("MAC")	-7.497e+05	4.22e+05	-1.777	0.080	-1.59e+06	9.07e+04
Capacity	32.1778	7.815	4.118	0.000	16.614	47.742
WinPercentage	5977.4924	5629.133	1.062	0.292	-5233.897	1.72e+04
GSR	1559.1217	1.43e+04	0.109	0.914	-2.7e+04	3.01e+04

Omnibus:	4.832	Durbin-Watson:	2.313
Prob(Omnibus):	0.089	Jarque-Bera (JB):	6.611
Skew:	-0.023	Prob(JB):	0.0367
Kurtosis:	4.327	Cond. No.	3.27e+18

MODEL 2: PAYTOTAL ~ WINPERCENTAGE + CAPACITY

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.560
Model:                           OLS   Adj. R-squared:                  0.550
Method:                Least Squares   F-statistic:                     55.46
Date:               Sat, 17 Oct 2020   Prob (F-statistic):           2.96e-16
Time:                       17:33:50   Log-Likelihood:                -1391.4
No. Observations:                 90   AIC:                             2789.
Df Residuals:                     87   BIC:                             2796.
Df Model:                          2
Covariance Type:            nonrobust
==============================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -9.521e+05   4.06e+05     -2.347      0.021   -1.76e+06   -1.46e+05
Capacity         58.7366      6.062      9.689      0.000      46.687      70.786
WinPercentage  5628.7086   6467.100      0.870      0.386   -7225.354    1.85e+04
==============================================================================
Omnibus:                        0.179   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.914   Jarque-Bera (JB):                0.059
Skew:                          -0.063   Prob(JB):                        0.971
Kurtosis:                       2.999   Cond. No.                     1.75e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.75e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

M2: Most significant attribute: 'WinPercentage' with value: 5629
M2: Proportion of Test Set Variance Accounted for: 0.889
```

MODEL 3: PAYTOTAL ~ WINPERCENTAGE + CAPACITY + GSR

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.572
Model:                           OLS   Adj. R-squared:                  0.557
Method:                Least Squares   F-statistic:                     38.31
Date:               Sat, 17 Oct 2020   Prob (F-statistic):           8.08e-16
Time:                       17:33:53   Log-Likelihood:                -1390.2
No. Observations:                 90   AIC:                             2788.
Df Residuals:                     86   BIC:                             2798.
Df Model:                          3
Covariance Type:            nonrobust
==============================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -2.648e+06   1.18e+06     -2.237      0.028      -5e+06   -2.95e+05
WinPercentage  4340.1512   6474.015      0.670      0.504   -8529.765    1.72e+04
Capacity         58.3915      6.021      9.698      0.000      46.422      70.361
GSR            2.233e+04   1.47e+04      1.524      0.131   -6803.486    5.15e+04
==============================================================================
Omnibus:                        0.453   Durbin-Watson:                   2.039
Prob(Omnibus):                  0.797   Jarque-Bera (JB):                0.126
Skew:                           0.050   Prob(JB):                        0.939
Kurtosis:                       3.154   Cond. No.                     5.15e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.15e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

M3: Most significant attribute: 'GSR' with value: 22326
M3: Proportion of Test Set Variance Accounted for: 0.879
```

MODEL 4: PAYTOTAL ~ CAPACITY + TOP5 + SUPERFAN

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.749
Model:                            OLS   Adj. R-squared:                  0.740
Method:                 Least Squares   F-statistic:                     85.33
Date:                Sat, 17 Oct 2020   Prob (F-statistic):           1.08e-25
Time:                        17:33:58   Log-Likelihood:                -1366.2
No. Observations:                  90   AIC:                             2740.
Df Residuals:                      86   BIC:                             2750.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      4.529e+05   3.29e+05      1.378      0.172      -2e+05    1.11e+06
top5[T.Y]      2.051e+06   2.69e+05      7.640      0.000    1.52e+06    2.59e+06
superfan[T.Y]  1.806e+06   4.25e+05      4.246      0.000     9.6e+05    2.65e+06
Capacity         12.2711      7.942      1.545      0.126      -3.517      28.059
==============================================================================
Omnibus:                        2.820   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.244   Jarque-Bera (JB):                2.688
Skew:                          -0.000   Prob(JB):                        0.261
Kurtosis:                       3.847   Cond. No.                     2.81e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.81e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

M4: Most significant attribute: 'top5[T.Y]' with value: 2051399
M4: Proportion of Test Set Variance Accounted for: 0.737
```

MODEL 5: PAYTOTAL ~ WINPERCENTAGE + CAPACITY + TOP5 + SUPERFAN

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.750
Model:                            OLS   Adj. R-squared:                  0.738
Method:                 Least Squares   F-statistic:                     63.66
Date:                Sat, 17 Oct 2020   Prob (F-statistic):           8.90e-25
Time:                        17:34:01   Log-Likelihood:                -1366.0
No. Observations:                  90   AIC:                             2742.
Df Residuals:                      85   BIC:                             2755.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      3.18e+05   3.92e+05      0.811      0.420   -4.62e+05     1.1e+06
top5[T.Y]      2.05e+06   2.69e+05      7.609      0.000    1.51e+06    2.59e+06
superfan[T.Y]  1.759e+06   4.33e+05      4.062      0.000    8.98e+05    2.62e+06
Capacity         11.8803      7.993      1.486      0.141      -4.013      27.773
WinPercentage  3186.6462   5011.991      0.636      0.527   -6778.535    1.32e+04
==============================================================================
Omnibus:                        3.712   Durbin-Watson:                   2.069
Prob(Omnibus):                  0.156   Jarque-Bera (JB):                4.088
Skew:                          -0.092   Prob(JB):                        0.130
Kurtosis:                       4.028   Cond. No.                     3.00e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large,  3e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

M5: Most significant attribute: 'top5[T.Y]' with value: 2050229
M5: Proportion of Test Set Variance Accounted for: 0.747
```

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.