Michelle Mak
IST 718
Lab 1

Introduction

College athletics is a proven source of publicity and revenue. However, some are concerned that funding is being wrongly allocated towards athletic programs instead of academics. As the interest and fanbase of college athletics grow, the field has become competitive not only for the players, but for the coaches who lead the teams to championships.

One way to combat the issue of excessive spending in this department might be to evaluate the salary of college football coaches. This investigation will attempt to find an optimal model to predict a competitive salary for the next Syracuse football coach.

Analysis

Data Overview

Five datasets are used in this experiment. The first contains the salary information of coaches in the NCAA from 2016, which was provided by Syracuse University. The second details the stadium capacity of university football stadiums.[1] Graduation rate for the 2006 cohort[2], the NCAA Football standings for the most recent championship[3], and donation information[4] is also included amongst the datasets.

Data Preparation and Preprocessing

For uniformity, all datasets are converted to lowercase, and columns that contained numbers are converted into numeric columns. For the coaches data, non-numeric values are replaced with zeros. In the interest of organization, superfluous information such as assistant pay, buy out, years of contribution, etc. are removed from their respective dataframes.

For the stadium data, the column of "team" is renamed to "school" because school is the key value on which all datasets will merge. In addition, a number of schools in the stadium data need to be renamed to match the other dataframes.

```
# replace acronym with school name
stadiums['school'] = stadiums['school'].replace(['ucf'], 'central florida')
stadiums['school'] = stadiums['school'].replace(['usf'], 'south florida')
stadiums['school'] = stadiums['school'].replace(['utsa'], 'texas-san antonio')
stadiums['school'] = stadiums['school'].replace(['byu'], 'brigham young')
stadiums['school'] = stadiums['school'].replace(['utep'], 'texas-el paso')
stadiums['school'] = stadiums['school'].replace(['tcu'], 'texas christian')
stadiums['school'] = stadiums['school'].replace(['unlv'], 'nevada-las vegas')
stadiums['school'] = stadiums['school'].replace(['smu'], 'southern methodist')
stadiums['school'] = stadiums['school'].replace(['niu'], 'northern illinois')
stadiums['school'] = stadiums['school'].replace(['miami (oh)'], 'miami (ohio)')
stadiums['school'] = stadiums['school'].replace(['fiu'], 'florida international')
stadiums['school'] = stadiums['school'].replace(['umass'], 'massachusetts')
```

[1] Stadium Data: https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv

[2] Graduation Data: http://www.ncaa.org/about/resources/research/graduation-rates

[3] 2018 Standings: https://www.sports-reference.com/cfb/years/2018-standings.html

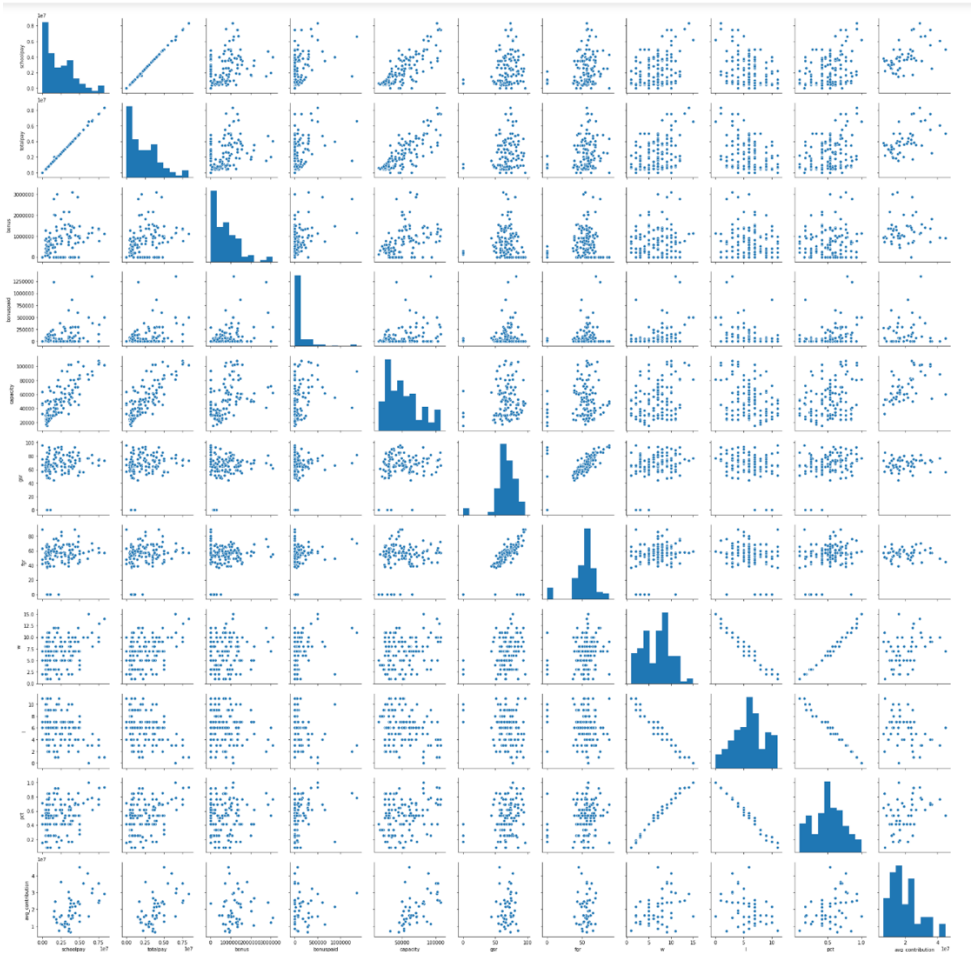
[4] Donation Info: https://virginiatech.sportswar.com/article/2017/07/26/power-5-donations/

Finally, all datasets are merged via “left join” on the “school” column. Some datasets, such as the donation data, are smaller the coaches data, which is the key dataset, so empty cells must be dropped when creating the regression model. However, **merging on “left join” will ensure that none of the schools present in the original coaches data are mistakenly dropped.**

Exploratory Analysis

An initial pair plot is created to get a quick overview of the relationships within the data. At first glance, it is clear that a number of variables have strong correlations with each other. For example, capacity and any monetary variable (school pay, total pay, bonus, average donor contribution) strong positive correlations. Graduation rates also seem to have some sort of positive correlation with average donor contribution.

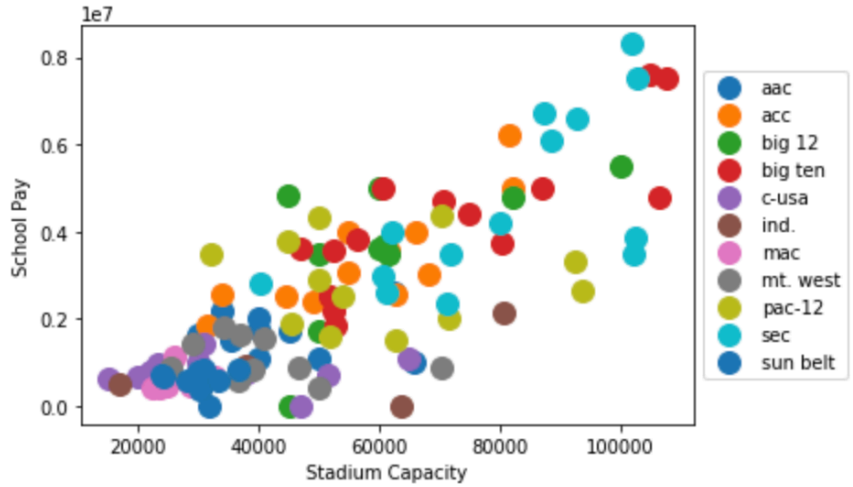
This histograms for pay-related variables also seem to skew heavily to the left. None of the variables seem to have a normal distribution.



A scatter plot of school pay and stadium capacity is created based on these initial findings. The plots are grouped by conference to illustrate any difference in tiers that might exist.

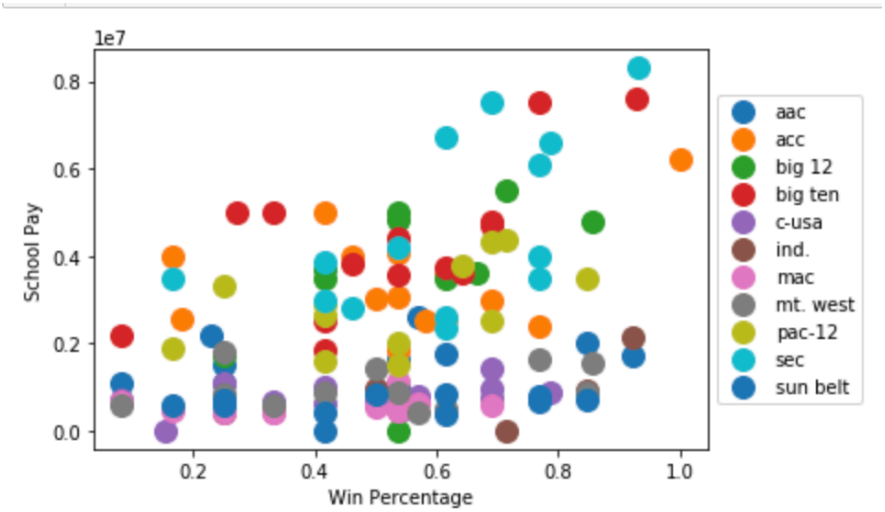
Michelle Mak
IST 718
Lab 1

The Big Ten conference and SEC both seem to be on the higher end of the salary spectrum, while schools in the Sun Belt, MAC, and AAC are crowded on the lower end. This plot shows a high correlation between school pay and stadium capacity.

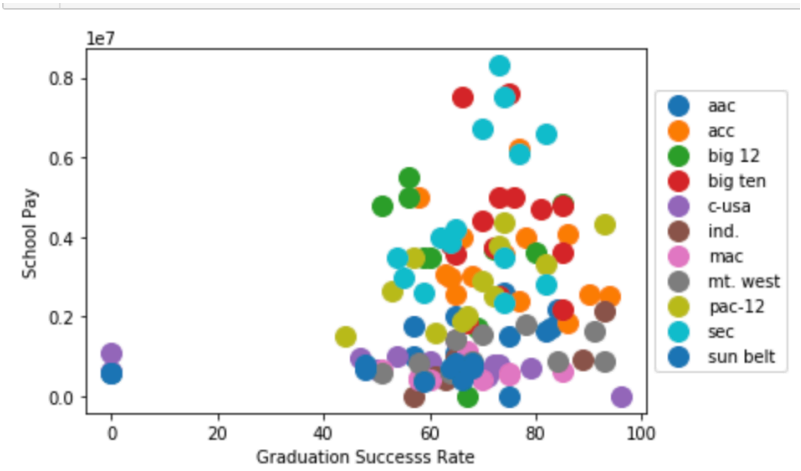


To further investigate, the win percentage of each school is also plotted with school pay and grouped by conference to see if this evens the playing field a bit across conferences.

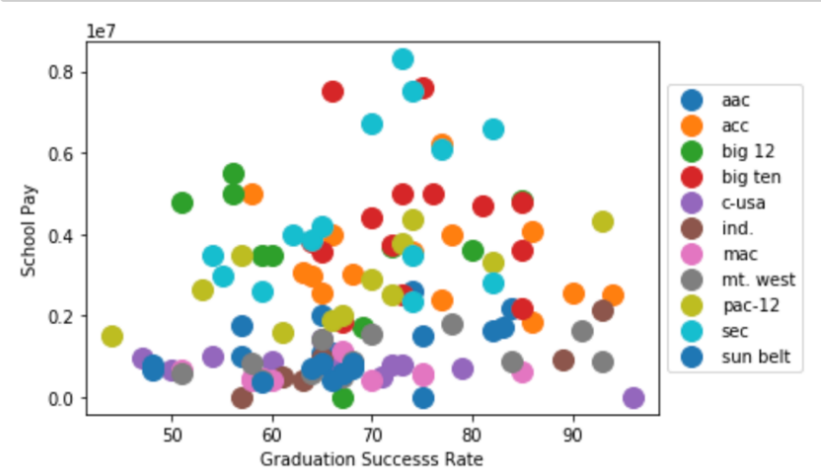
While there is less correlation between these two variables, the plot does show a little more variation in the winner's circle, with a handful of Sun Belt universities in the higher winning percentages.



Next, we investigate whether graduation rate of student athletes has any relevance to the coaches salary. There seems to be a very slight positive correlation between these two variables. However, outliers with a graduation rate of 0 might be creating an issues in evaluating the data.



Therefore, a second plot is created with the 0 values replaced with the mean. The positive correlation is no longer as obvious. Here are the results:



Methods

With these insights in mind, regression models are created using the "statsmodels.apio". First, the data is split into a training and testing set. Then, several models are created in an attempt to achieve the highest R-squared possible.

The first model investigates the relevance of win percentage, along with graduation rates and salary. Since academics is the core part of attending college, the academic success of the players should theoretically be considered by the coaches.

However, this first model, which used the variables of win percentage, graduation success rate, and federal graduation rate, came back with a very low R-squared of 0.161. In addition, **both**

graduation measure have a high p-value, which means that they are definitely not significant in regard to projected salary.

```
    OLS Regression Results
==============================================================================
Dep. Variable:              schoolpay   R-squared:                       0.161
Model:                            OLS   Adj. R-squared:                  0.132
Method:                 Least Squares   F-statistic:                     5.575
Date:                Sat, 02 Feb 2019   Prob (F-statistic):            0.00152
Time:                        03:11:59   Log-Likelihood:                -1438.4
No. Observations:                  91   AIC:                             2885.
Df Residuals:                      87   BIC:                             2895.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -8.793e+04    8.4e+05     -0.105      0.917   -1.76e+06    1.58e+06
pct         3.065e+06   8.56e+05      3.581      0.001    1.36e+06    4.77e+06
gsr         2926.3899   1.94e+04      0.151      0.881   -3.57e+04    4.15e+04
fgr         1.173e+04   1.99e+04      0.590      0.557   -2.78e+04    5.12e+04
==============================================================================
Omnibus:                        5.920   Durbin-Watson:                   2.342
Prob(Omnibus):                  0.052   Jarque-Bera (JB):                6.065
Skew:                           0.610   Prob(JB):                       0.0482
Kurtosis:                       2.663   Cond. No.                         454.
==============================================================================
```

Several other models are attempted, but the best results came about when predicting for school pay using the variables of capacity, conference, and win percentage. This **rendered an R-squared of 0.795** in addition to having almost all significant variables (save for a handful of conferences), which means this model accounts for almost 80% of variance.

```
OLS Regression Results
==============================================================================
Dep. Variable:              schoolpay   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     24.30
Date:                Sat, 02 Feb 2019   Prob (F-statistic):           4.28e-21
Time:                        03:12:15   Log-Likelihood:                -1329.5
No. Observations:                  88   AIC:                             2685.
Df Residuals:                      75   BIC:                             2717.
Df Model:                          12
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                -1.208e+06   4.71e+05     -2.566      0.012   -2.15e+06    -2.7e+05
conference[T.acc]         1.412e+06   4.55e+05      3.107      0.003    5.07e+05    2.32e+06
conference[T.big 12]      1.251e+06   5.37e+05      2.332      0.022    1.83e+05    2.32e+06
conference[T.big ten]     1.549e+06   5.18e+05      2.988      0.004    5.16e+05    2.58e+06
conference[T.c-usa]       -1.26e+05   4.83e+05     -0.261      0.795   -1.09e+06    8.37e+05
conference[T.ind.]       -1.033e+06   5.96e+05     -1.731      0.087   -2.22e+06    1.56e+05
conference[T.mac]         -312.9618   4.93e+05     -0.001      0.999   -9.82e+05    9.81e+05
conference[T.mt. west] -1.124e+05   4.96e+05     -0.226      0.821    -1.1e+06    8.76e+05
conference[T.pac-12]      1.291e+06   4.82e+05      2.676      0.009     3.3e+05    2.25e+06
conference[T.sec]          1.51e+06   4.96e+05      3.044      0.003    5.22e+05     2.5e+06
conference[T.sun belt]      -1e+05   4.93e+05     -0.203      0.840   -1.08e+06    8.82e+05
capacity                   41.0005      7.017      5.843      0.000      27.022      54.979
pct                       1.543e+06   4.79e+05      3.222      0.002    5.89e+05     2.5e+06
==============================================================================
```

```
Omnibus:                        3.118   Durbin-Watson:                   1.903
Prob(Omnibus):                  0.210   Jarque-Bera (JB):                2.774
Skew:                          -0.183   Prob(JB):                        0.250
Kurtosis:                       3.789   Cond. No.                     6.63e+05
================================================================================
```

Based on the scatter plots created in the exploratory analysis, it was pretty clear that there are discrepancies between conferences. Therefore, further investigation leads us to using a mixed linear effects model, also called through the "statsmodel.api."

This model predicts school pay using the variables of stadium capacity and win percentage while grouping by conference. The results are very interesting – the significance of both capacity and win percentage grow, despite the standard error of this model being high.

```
  Mixed Linear Model Regression Results
============================================================================
Model:                MixedLM     Dependent Variable:     schoolpay
No. Observations:     92          Method:                 REML
No. Groups:           11          Scale:                  918347973340.5237
Min. group size:      4           Likelihood:             -1376.2371
Max. group size:      12          Converged:              Yes
Mean group size:      8.4
----------------------------------------------------------------------------
               Coef.        Std.Err.    z    P>|z|    [0.025       0.975]
----------------------------------------------------------------------------
Intercept      -880186.760 454194.954 -1.938 0.053 -1770392.511   10018.992
capacity            45.891      7.144  6.423 0.000       31.888      59.894
pct            1395981.251 471529.062  2.961 0.003   471801.272 2320161.230
Group Var 604280083095.954 412758.706
============================================================================
```

In order to answer the question of what the salary would be if Syracuse moved to the Big Ten conference, the conference code of Syracuse is changed to "big ten."

Results

Here are the results of predicted the predicted head coach salary for Syracuse when applying the best linear regression and the mixed linear regression:

Conference	Model	Prediction
ACC	Linear Regression	3.409961e+06
ACC	Mixed Linear Model	2.453455e+06
Big Ten	Linear Regression	3.546605e+06
Big Ten	Mixed Linear Model	2.453455e+06

The difference between the predicted results of the linear regression model and mixed linear regression model is about 1.1 million dollars. In addition, the predicted salary of the mixed linear regression model did not change at all Syracuse's conference was changed to Big Ten. However, the prediction of the mixed linear model is closer to the actual salary than that of the linear regression model.

Recommendations

To make the best decision, the Syracuse's stats are compared to the average of all the variables within the respective conferences.

ACC Salary

The current salary for the head football coach at Syracuse is $2.4 million, which is about $1 million less than the conference average. However, the Syracuse Carrier Dome's stadium capacity is also lower than the average. Syracuse's graduation rate is slightly above that of the conference average. Syracuse also has a relatively high win percentage at 77%.

Therefore, **the salary recommendation within the ACC conference, is $2,453,455**, based on the mixed effects linear regression model. This is largely because the stadium capacity of the Carrier dome is 7 percent less than that of the conference average and **stadium capacity is the most significant variable within the model**. However, since Syracuse is performing above average in both graduation rate and win percentage, the recommendation is to offer the possibility of a bonus, should certain benchmarks be met. This will allow the opportunity for the head coach's salary to be more aligned with the ACC conference average.

	school	conference	coach	schoolpay	totalpay	bonus	bonuspaid	stadium	capacity	gsr	fgr	w	l	pct	avg_contribution
102	syracuse	acc	dino babers	2401206.0	2401206.0	0.0	0.0	carrier dome	49250.0	77.0	64.0	10.0	3.0	0.769	NaN

	schoolpay	totalpay	bonus	bonuspaid	capacity	gsr	fgr	w	l	pct	avg_contribution
count	1.400000e+01	1.400000e+01	1.400000e+01	14.000000	12.000000	14.000000	14.000000	14.000000	14.000000	14.000000	8.000000e+00
mean	3.409629e+06	3.433797e+06	7.323512e+05	147851.214286	57675.333333	74.714286	62.857143	7.142857	5.714286	0.544000	1.783211e+07
std	1.165548e+06	1.229711e+06	7.534458e+05	254303.852926	16271.463500	11.193601	12.126621	3.207135	2.462912	0.213062	5.632102e+06
min	1.831580e+06	1.831580e+06	0.000000e+00	0.000000	31500.000000	58.000000	47.000000	2.000000	0.000000	0.167000	9.605338e+06
25%	2.549446e+06	2.549446e+06	0.000000e+00	0.000000	48062.500000	65.250000	53.500000	6.250000	5.000000	0.471500	1.513880e+07
50%	3.038868e+06	3.038868e+06	6.675000e+05	25000.000000	58250.000000	75.500000	60.500000	7.000000	6.000000	0.538000	1.649964e+07
75%	3.995108e+06	3.995108e+06	1.345000e+06	187500.000000	66774.750000	84.000000	69.000000	8.000000	7.000000	0.615000	2.197844e+07
max	6.205000e+06	6.543350e+06	2.165000e+06	869917.000000	82300.000000	94.000000	86.000000	15.000000	10.000000	1.000000	2.664437e+07

Big Ten Salary

For the hypothetical scenario of having Syracuse move to the Big Ten conference, Syracuse stats are one again compared to the conference averages.

The current salary for the head football coach about 25% lower than the average Big Ten conference salary. The Syracuse Carrier Dome's stadium capacity is lower than the average Big Ten stadium by 17%.

The salary recommendation for the Syracuse head football coach within the Big Ten conference, is $3,546,605, based on the linear regression model. Should Syracuse actually move into the Big Ten conference, a considerable salary increase should be justified in order to maintain a competitive salary amongst peers. A larger salary will also incentivize a maintenance

of the above average stats that Syracuse has in terms of graduation rate and win percentage when compared to Big Ten competitors. The recommended salary is still 8% less than the conference average since the Carrier Dome's stadium capacity is so much smaller and, therefore, will probably generate proportionally less revenue.

	school	conference	coach	schoolpay	totalpay	bonus	bonuspaid	stadium	capacity	gsr	fgr	w	l	pct	avg_contribution
102	syracuse	big ten	dino babers	2401206.0	2401206.0	0.0	0.0	carrier dome	49250.0	77.0	64.0	10.0	3.0	0.769	NaN

	schoolpay	totalpay	bonus	bonuspaid	capacity	gsr	fgr	w	l	pct	avg_contribution
count	1.500000e+01	1.500000e+01	1.500000e+01	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	1.200000e+01
mean	4.177160e+06	4.177160e+06	9.194445e+05	139666.666667	70353.933333	74.266667	60.266667	7.066667	5.666667	0.544667	1.700710e+07
std	1.715859e+06	1.715859e+06	7.136893e+05	177184.998828	22127.837973	7.274875	5.861090	3.127451	2.497618	0.220009	6.502862e+06
min	1.830000e+06	1.830000e+06	0.000000e+00	0.000000	47130.000000	64.000000	50.000000	1.000000	1.000000	0.083000	7.451606e+06
25%	3.031000e+06	3.031000e+06	6.375000e+05	0.000000	52489.500000	68.500000	57.000000	5.000000	4.000000	0.417000	1.248272e+07
50%	3.800000e+06	3.800000e+06	9.500000e+05	50000.000000	60670.000000	73.000000	58.000000	7.000000	6.000000	0.538000	1.548437e+07
75%	4.900000e+06	4.900000e+06	1.145000e+06	235000.000000	83706.000000	79.000000	65.000000	9.000000	7.000000	0.692000	2.372238e+07
max	7.600000e+06	7.600000e+06	2.875000e+06	600000.000000	107601.000000	85.000000	72.000000	13.000000	11.000000	0.929000	2.526268e+07

Appendix

The requested hypothetical scenario of Syracuse returning to the Big East conference is not possible because there is not Big East data in the coaches data.