Lab Exercise 1 - NCAA Football Coach Salary Analysis
IST 718 - Big Data Analytics | Kelly Hwang

Introduction

The highest paid college football coaches in the nation can earn millions of dollars a year. In return, schools generate even more millions of dollars in revenue off the games themselves. This exercise is tasked with tackling the question: how can we recommend the best salary (base compensation) for Syracuse's next head football coach, with the data points at hand? The focus of this lab will be using recent public salary, game, stadium, and graduation data to predict compensation values.

Goals:

1. Create a model for explaining the relationship between collected variables and salaries for college football coaches. Identify biggest impact.

2. Predict an appropriate salary for a football coach at Syracuse University.

3. Estimate salary change if conference for Syracuse changed to Big Ten.

4. Evaluate effects of graduation rate on projected salary.

Analysis and Models

About the Data

The following datasets were used in this analysis:

- School/Coach salary data (base)
- School graduation data
- School donation data (added but not used in modeling)
- School stadium data
- 2017 Game data

Key attribute definitions and use cases:

- School Pay: This is the base salary provided by the university and does not include any bonus or incentives. This is used as the response variable for salary.
- Total Pay: Sum of School Pay and additional compensation from non university sources. This is not be used as salary.
- Bonuses: not be used in salary calculations.

- GSR: Graduation Success Rate. This is the proportion of student-athletes on any given team who earn a college degree. This metric was developed to more accurately reflect college transfers among student athletes, as the original graduation rate metric (FGR) did not take this into account.
- FGR: Federal Graduation Rate which is compiled by the US Department of Education. Does not take into account transfers.
- Capacity: Max person capacity for the stadium
- PCT: win ratio for the season

After cleaning and merging the data into a final dataframe:

	schoolpay	totalpay	bonus	bonuspaid	buyout	gsr	fgr	avg donations	capacity	latitude	longitude	win	loss	pct
count	125.00	125.00	125.00	125.00	125.00	125.00	125.00	51.00	125.00	125.00	125.00	125.00	125.00	125.00
mean	2410300.71	2417060.76	748296.50	105265.10	6949955.77	0.68	0.57	19795658.98	51283.66	36.79	-92.00	6.69	5.86	0.52
std	1881377.04	1885752.30	662812.43	211405.28	10086846.94	0.12	0.11	9108547.41	23690.55	4.86	14.66	3.05	2.50	0.22
min	390000.00	390000.00	0.00	0.00	0.00	0.31	0.31	6572169.00	15000.00	21.37	-157.93	0.00	0.00	0.00
25%	801504.00	805850.00	245000.00	0.00	688500.00	0.60	0.52	12488029.50	30600.00	33.46	-97.37	5.00	4.00	0.42
50%	1831580.00	1900008.00	650000.00	25000.00	2911667.00	0.67	0.57	16575714.00	49250.00	36.89	-86.81	7.00	6.00	0.54
75%	3605000.00	3617500.00	1050000.00	100000.00	9250000.00	0.75	0.63	24609826.50	65326.00	40.76	-81.78	9.00	7.00	0.69
max	8307000.00	8307000.00	3100000.00	1350000.00	68125000.00	0.94	0.90	45024857.00	107601.00	47.65	-71.17	13.00	12.00	1.00

There are a total of 125 schools in the analysis. The original dataset contained 129 schools, but 4 were excluded in this analysis upon examination that there was missing salary data.

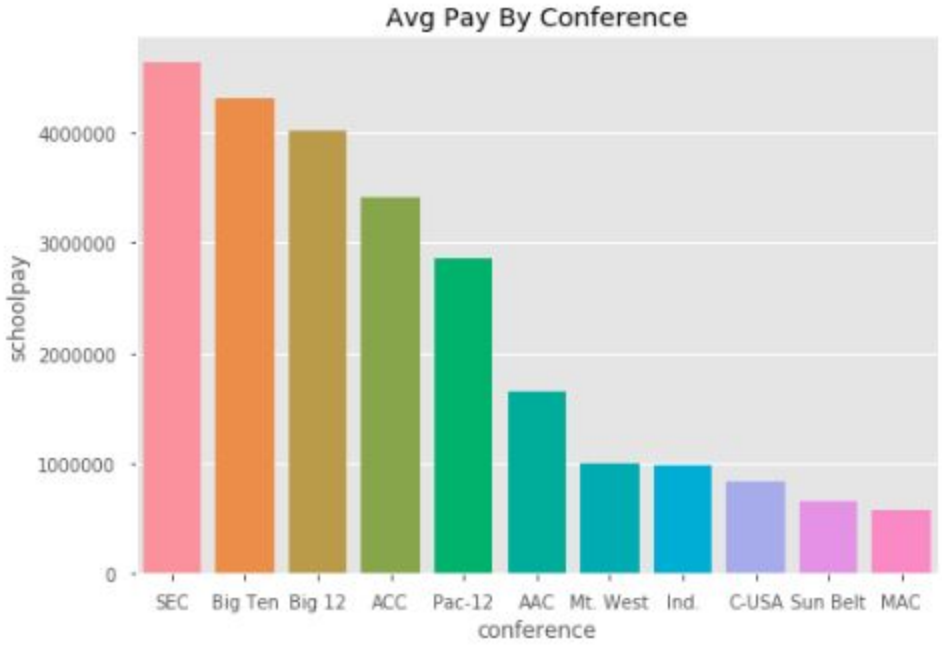
The 4 schools that were excluded:

	school	conference	coach	schoolpay	totalpay	bonus	bonuspaid	assistantpay	buyout
12	Baylor	Big 12	Matt Rhule	--	--	--	--	$0	--
16	Brigham Young	Ind.	Kalani Sitake	--	--	--	--	$0	--
91	Rice	C-USA	Mike Bloomgren	--	--	--	--	$0	--
99	Southern Methodist	AAC	Sonny Dykes	--	--	--	--	$0	--

Donation data only contained matching data for 51/125 schools, so donation data was excluded from analysis as well.

Exploring the Data - Initial Analysis

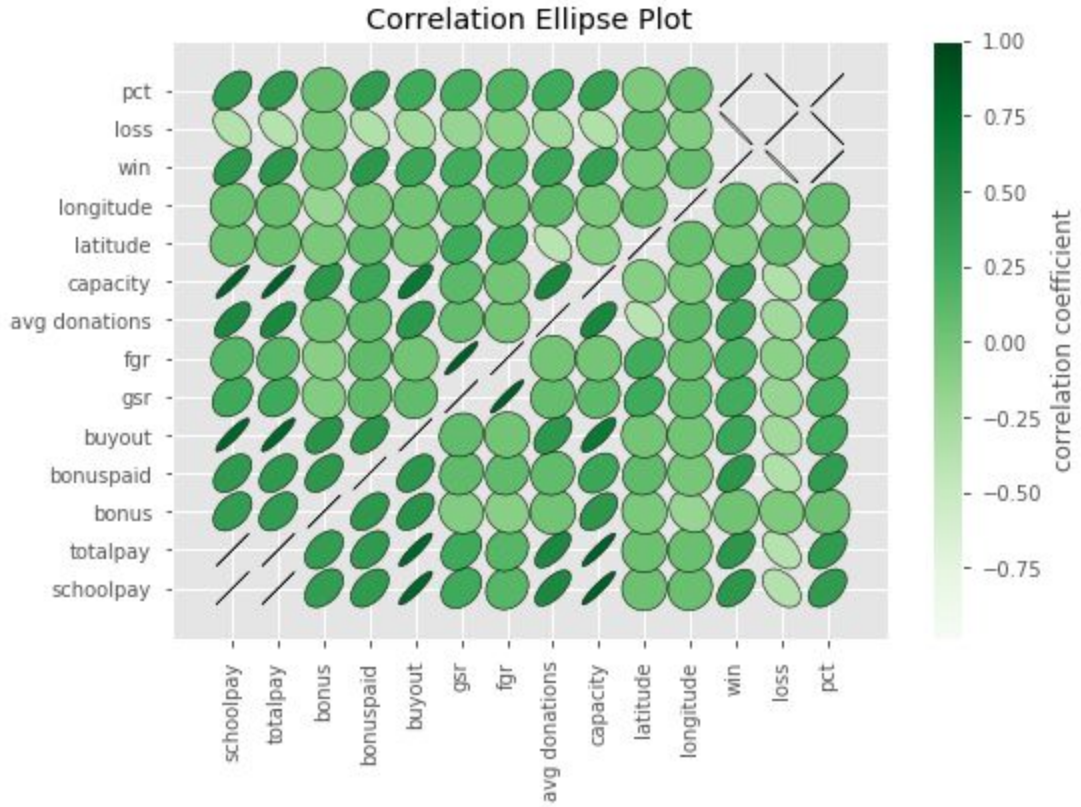
Sorted in order of most to least paid, the following are the average top paying conference sectors:



Syracuse is in the ACC conference. ACC is about the middle of the spectrum of average pay. In general, Big Ten colleges are paid on average $2,647,746.01 more than ACC colleges. Should Syracuse move from the ACC into Big Ten, there is a likelihood of an increase in salary. This is the breakdown of number of schools per conference in the dataset:

conference	mean	count
SEC	4642080.21	14
Big Ten	4304013.71	14
Big 12	4016755.67	9
ACC	3409628.86	14
Pac-12	2856577.75	12
AAC	1656267.70	10
Mt. West	1000642.00	12
Ind.	985816.00	5
C-USA	839518.54	13
Sun Belt	650650.00	10
MAC	579836.17	12

Plotting correlation between all variables:



There appears to be large positive correlations between pay and stadium capacity. There is also small positive correlations between pay and win rate. Most likely these will be variables to be used in regression modeling.

Models / Results

3 OLS regression models were fitted and tested. Schoolpay was used as the response variable. Different explanatory variables were tested in each model.

Model 1
Variables used: Capacity, PCT, GSR, FGR

OLS Regression Results

Dep. Variable:	schoolpay	R-squared:	0.737
Model:	OLS	Adj. R-squared:	0.724
Method:	Least Squares	F-statistic:	54.72
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	6.83e-22
Time:	22:06:12	Log-Likelihood:	-1258.2
No. Observations:	83	AIC:	2526.
Df Residuals:	78	BIC:	2539.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.67e+06	6.68e+05	-3.998	0.000	-4e+06	-1.34e+06
capacity	58.2621	4.431	13.150	0.000	49.441	67.083
pct	3.127e+05	5.56e+05	0.563	0.575	-7.93e+05	1.42e+06
gsr	1.991e+06	1.51e+06	1.321	0.190	-1.01e+06	4.99e+06
fgr	6.72e+05	1.71e+06	0.393	0.696	-2.74e+06	4.08e+06

Omnibus:	1.030	Durbin-Watson:	1.812
Prob(Omnibus):	0.597	Jarque-Bera (JB):	0.587
Skew:	-0.181	Prob(JB):	0.746
Kurtosis:	3.195	Cond. No.	1.20e+06

With these 3 variables, R-squared implies that ~74% of variation in salary can be explained by these variables. However, P value for Model 1 implies that PCT, GSR, and FGR do not have a stastistically significant effect on salary and as such can be dropped from the model.

Model 2
Variables used: Capacity

OLS Regression Results

Dep. Variable:	schoolpay	R-squared:	0.706
Model:	OLS	Adj. R-squared:	0.702
Method:	Least Squares	F-statistic:	194.1
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	3.28e-23
Time:	22:07:02	Log-Likelihood:	-1263.0
No. Observations:	83	AIC:	2530.
Df Residuals:	81	BIC:	2535.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.255e+05	2.5e+05	-3.297	0.001	-1.32e+06	-3.27e+05
capacity	59.6089	4.279	13.931	0.000	51.095	68.122

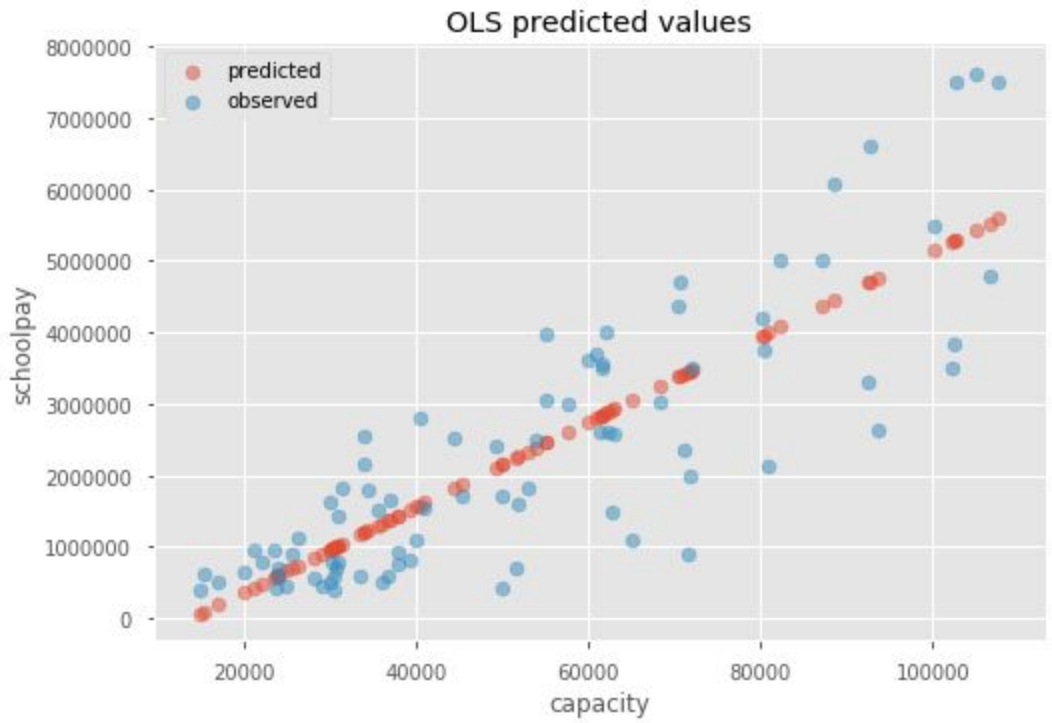
Omnibus:	0.583	Durbin-Watson:	1.732
Prob(Omnibus):	0.747	Jarque-Bera (JB):	0.375
Skew:	-0.164	Prob(JB):	0.829
Kurtosis:	3.024	Cond. No.	1.34e+05

Model 2 R-squared only dropped to 70% after dropping all the other variables. However, although R-squared is relatively high, the model doesn't appear very precise as the std error for capacity is relatively high, at 4.279.

Using Model 2's prediction formula, the following Syracuse salary was calculated:

```
actual salary for Syracuse: $ 2,401,206
recommended salary for Syracuse: $ 2,111,770.0
difference: $ 289,436.0
```

Running the model on test data, the average error between actual and predicted salary outcomes is an estimated $532,272.53. Below is a chart of OLS predicted vs observed values:



Model 3

Variables used: Capacity, PCT

OLS Regression Results

Dep. Variable:	schoolpay	R-squared:	0.713
Model:	OLS	Adj. R-squared:	0.706
Method:	Least Squares	F-statistic:	99.38
Date:	Sat, 02 Feb 2019	Prob (F-statistic):	2.06e-22
Time:	22:13:46	Log-Likelihood:	-1261.9
No. Observations:	83	AIC:	2530.
Df Residuals:	80	BIC:	2537.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.122e+06	3.22e+05	-3.479	0.001	-1.76e+06	-4.8e+05
capacity	57.6760	4.456	12.943	0.000	48.808	66.544
pct	7.85e+05	5.44e+05	1.444	0.153	-2.97e+05	1.87e+06

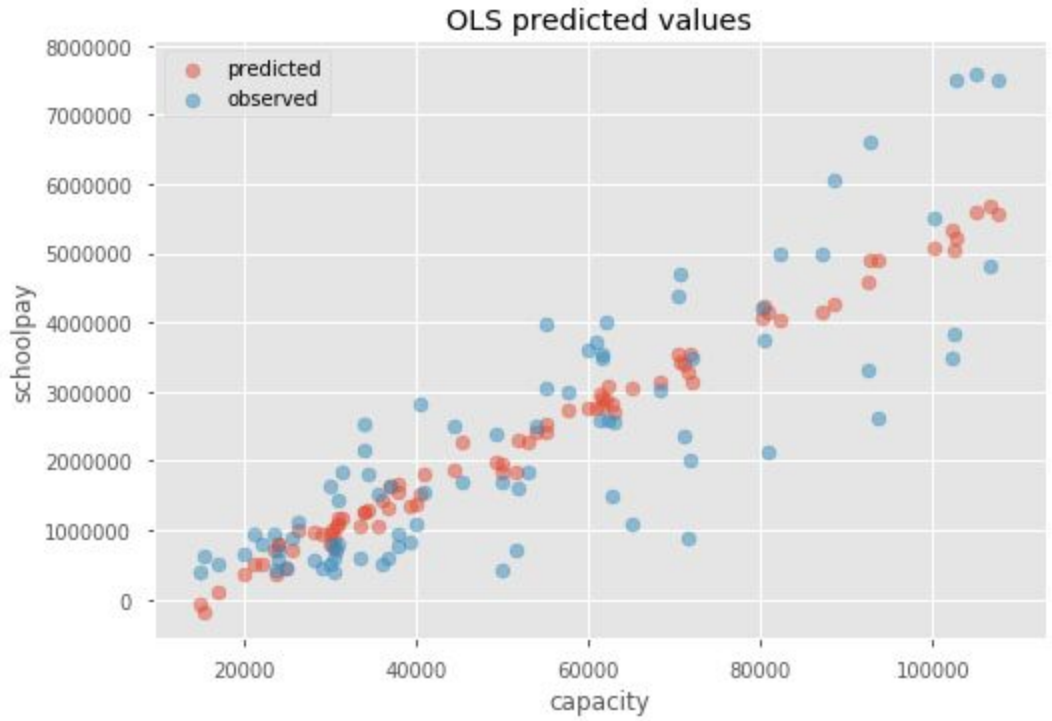
Omnibus:	0.535	Durbin-Watson:	1.733
Prob(Omnibus):	0.765	Jarque-Bera (JB):	0.340
Skew:	-0.156	Prob(JB):	0.844
Kurtosis:	3.019	Cond. No.	3.18e+05

Although R-squared slightly increased, error rate is still high. The P value for win rate, although still relatively insignificant, has become slightly more significant in this model without graduation rates.

Using model 3's formula, the following Syracuse salary was calculated:

```
actual salary for Syracuse: $ 2,401,206
recommended salary for Syracuse: $ 1,977,790.0
difference: $ 423,416.0
```

The precision of the model is worse than Model 2 for this particular school. But, when running Model 3 on test data, the avg error between actual and predicted salary outcomes overall improved slightly, at $485,451. Plot for OLS predicted vs observed values:



Conclusions

- With the data sets on hand, there are not enough variables to predict coach salary with high precision. One observation is clear: that stadium size and sell-out potential are highly correlated to a larger salary.

- Model 2 produced the best estimate for Syracuse salary: $2,111,770.

- It is possible that if Syracuse moved to the Big Ten, salary could increase on average upwards of $2,647,746.01. There is no data for Big East in this data set.

- The following schools were removed from analysis due to missing salary information: Baylor, Brigham Young, Rice, and Southern Methodist.

- Graduation rate does not have a statistically significant effect on projected salary.

- All 3 models had a high R-Squared value of over 70%. However, accuracy of the predictions were off by ~$500,000 on average.

- The single biggest impact on salary (in the available data for this analysis) is stadium capacity.

Given more time and data, a better model could be fitted to improve accuracy of salary predictions.

Resources

Code:
https://stackoverflow.com/questions/34556180/how-can-i-plot-a-correlation-matrix-as-a-set-of-ellipses-similar-to-the-r-open

Data:
http://sports.usatoday.com/ncaa/salaries
http://grfx.cstv.com/photos/schools/sdsu/genrel/auto_pdf/what-is-grad-success-rate.pd
http://www.ncaa.org