Darrell Nelson

Big Data Analytics (IST – 736)

7/20/2019

Lab 01

Background

Research question: Using predictive analytics, how can we recommend the best salary for our (Syracuse) next head football coach?

Data: An excel file containing relevant data of the coaches and other pertinent information in a [118 x 23] matrix.

Data Preparation

- Import data into Spyder and convert into dataframe using pandas.
- Select which features are appropriate:
    - School is a placeholder for a primary key (PK) which is needed for labelling, but will not be used in the analysis.
    - Conf, MedianConfSal, and StadSize are appropriate features.
    - FBREV16 will not be used as Syracuse does not contain that feature.
    - For the most part SchoolPay == TotalPay. There are 10 instances where TotalPay > SchoolPay by a small percentage. Therefore, TotalPay was used as the metric for how much coaches are paid excluding bonuses.
    - Seat Rank and StadSize rank the same metric (which team has the most seats). When doing modeling it will be preferable to have the actual number of seats as a metric as it is the fans sitting in these seats that best dictate revenue. The more butts in seats the higher the revenue, and the more the coach can be paid. The relative number of seats compared to the competition is inconsequential.
    - Graduation Rate (GSR) is preferred over GSR rank for the same reason. The graduation rate is a variable more under the control of the school/coach and not their relative ranking across other schools.
    - W/L Ratio = does the coach produce on-field results
    - Offense/Defense metrics
        - Which one is a better indicator of winning?
        - Do offensive or defensive coaches get paid more?
    - Unsure how the PointsPerGame metric is calculated. The numbers do not signify the actual points per game the team scored on average and it's not the average total number of points scored for and against the team per game. Due to this ambiguity this variable will be left in the model to see if it is important.

```
Shape of new matrix: (118, 10)
Columns of new matrix: ['School', 'Conf', 'MedianConfSal', 'TotalPay', 'StadSize', 'Graduation Rate
(GSR)', 'Ratio', 'OffenceScore', 'Defense Score', 'PointsPerGame']
```

With the variables set, all schools with NAs in any of the 10 columns were removed from the dataset. The rows were reduced from 118 to 103, a 12.71% reduction.

```
New shape: (103, 10)
Percentage of rows removed: 12.71%
```

Based off the correlation matrix it appears total pay is highly correlated with all the selected attributes.

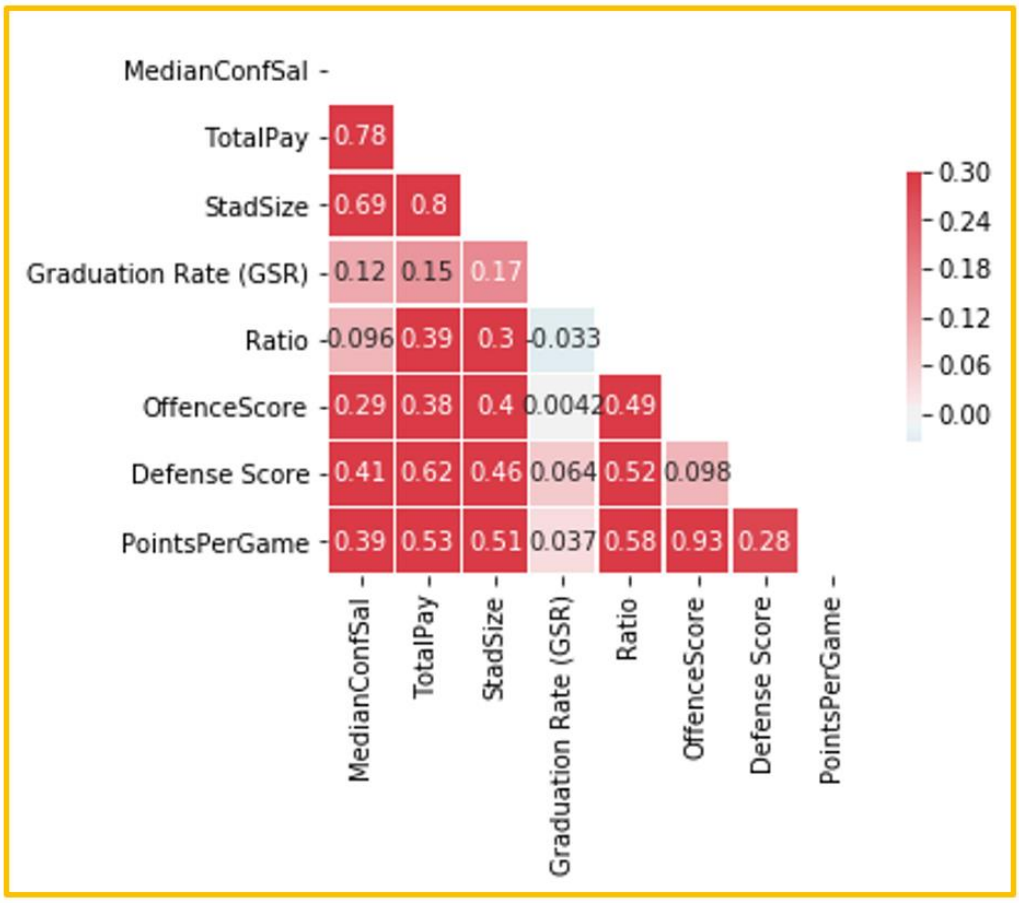

Figure 1. Correlation Matrix

It's of note to point out that GSR seems to be the least correlated with winning metrics (e.g. Ratio, Offence, and Defense) and getting the coach paid (TotalPay). Therefore, there appears to be a built-in bias for coaches to focus all their attention on football and neglect the academic side of their duties.

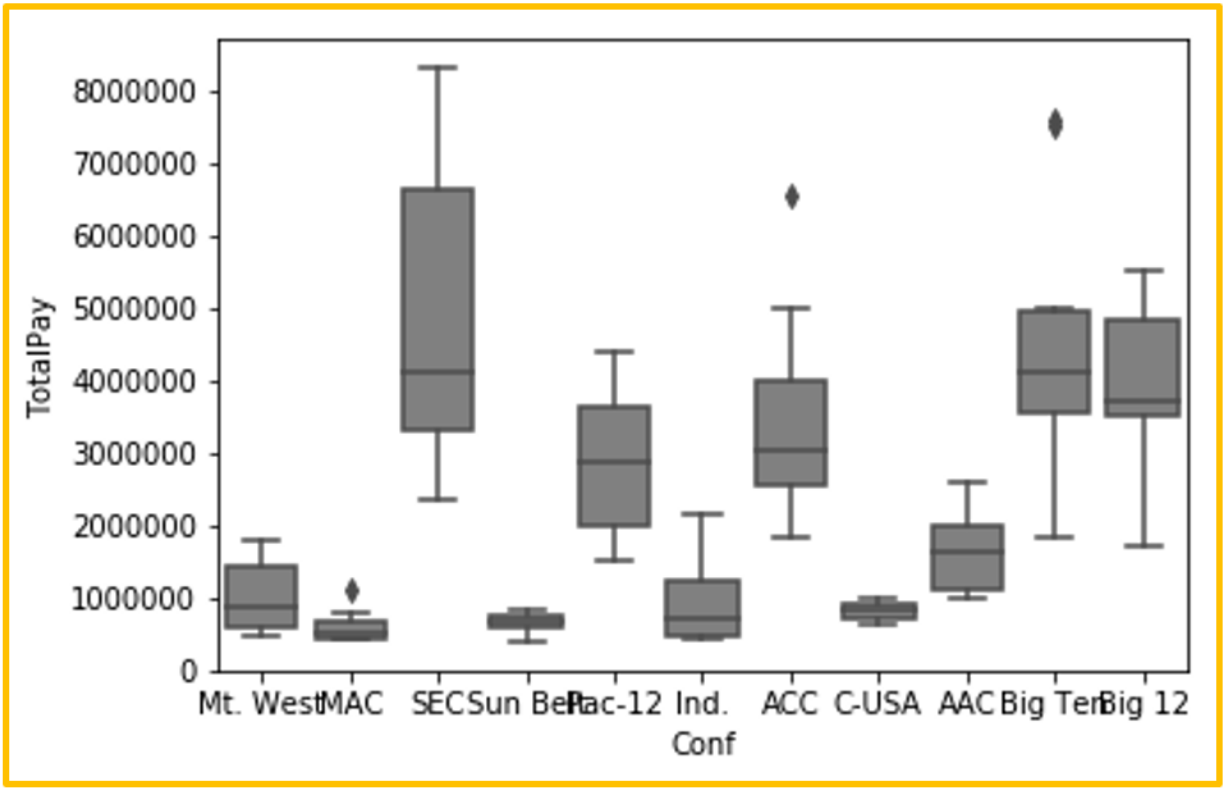

Figure 2. TotalPay vs Conference

It appears that the pay scale for coaches is conference dependent. With the SEC dominating the upper half and the non-power 5 conferences (Mt. West, MAC, Sun Belt, and C-USA) dominating the lower end. It suggests that a coach's pay is highly tied to the conference he is in. Is this because the upper end conferences have more revenue to pay for better coaching, or is it solely based on on-field production?

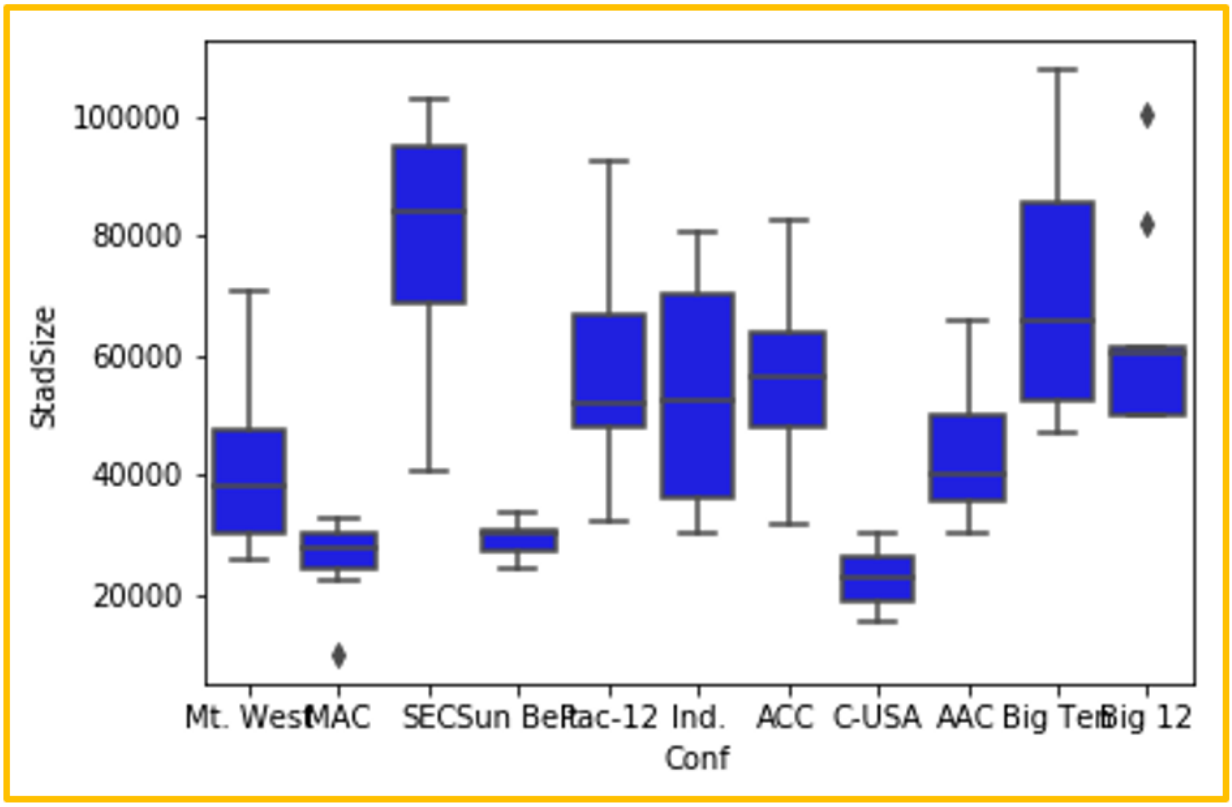

Figure 3. StadiumSize vs Conference

Looking at the total number of seats available in each stadium by conference (*Fig. 3*) reveals that the trends are the same between the two graphs. So which comes first, the higher the coach's salary the higher the revenue generated and thus increased stadium size? Or, is it bigger stadiums tend to be built based off how large the loyal fanbase is (those who purchase tickets no matter how the team performs) and that creates more revenue to spend on hiring a really good coach.

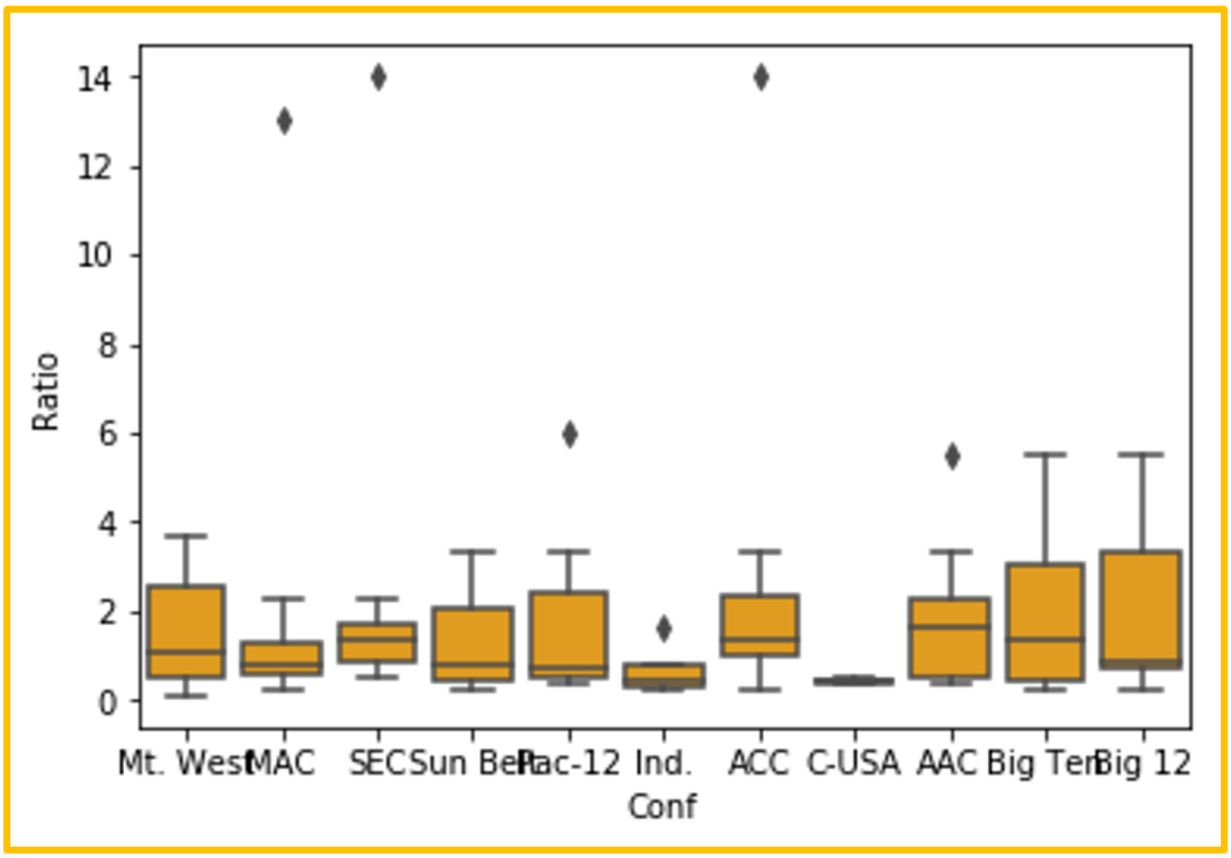

Figure 4. Ratio vs Conference

Figure 4. shows a pareto distribution in terms of less than 20% of all the football programs do all the winning and the rest are buried in mediocrity (at least for this year). There appear to be very little “really good” teams (teams that win 10 out of 12 games or have a win ratio of 5), some over achievers (Ratio > 10), and the rest are all bottom dwellers (3 < Ratio < 0).

Ordinary Least Squares

An Ordinary Least Squares (OLS) model was built to predict coaches' pay based off the metrics mentioned above.

OLS Regression Results

Dep. Variable:	TotalPay	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.810
Method:	Least Squares	F-statistic:	28.13
Date:	Sun, 21 Jul 2019	Prob (F-statistic):	2.04e-27
Time:	20:11:32	Log-Likelihood:	-1543.0
No. Observations:	103	AIC:	3120.
Df Residuals:	86	BIC:	3165.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.059e+06	2.36e+06	-2.573	0.012	-1.07e+07	-1.38e+06
Conf[T.ACC]	1.024e+06	3.95e+05	2.592	0.011	2.39e+05	1.81e+06
Conf[T.Big 12]	2.034e+06	6.42e+05	3.169	0.002	7.58e+05	3.31e+06
Conf[T.Big Ten]	1.908e+06	6.65e+05	2.869	0.005	5.86e+05	3.23e+06
Conf[T.C-USA]	8.175e+04	7.3e+05	0.112	0.911	-1.37e+06	1.53e+06
Conf[T.Ind.]	-1.283e+06	6.7e+05	-1.916	0.059	-2.61e+06	4.79e+04
Conf[T.MAC]	-8.86e+05	5.71e+05	-1.552	0.124	-2.02e+06	2.49e+05
Conf[T.Mt. West]	-7.584e+05	4.69e+05	-1.618	0.109	-1.69e+06	1.73e+05
Conf[T.Pac-12]	9.402e+05	4.77e+05	1.971	0.052	-7956.969	1.89e+06
Conf[T.SEC]	2.095e+06	6.94e+05	3.019	0.003	7.15e+05	3.48e+06
Conf[T.Sun Belt]	-7.321e+05	5.13e+05	-1.428	0.157	-1.75e+06	2.87e+05
MedianConfSal	-0.3043	0.314	-0.970	0.335	-0.928	0.320
StadSize	28.8890	6.127	4.715	0.000	16.710	41.068
GraduationRate	4724.3073	6309.420	0.749	0.456	-7818.405	1.73e+04
Ratio	5.851e+04	5.35e+04	1.093	0.277	-4.79e+04	1.65e+05
OffenceScore	-8.021e+04	4.77e+04	-1.682	0.096	-1.75e+05	1.46e+04
DefenseScore	5.013e+04	2.03e+04	2.467	0.016	9740.451	9.05e+04
PointsPerGame	8.922e+04	4.09e+04	2.184	0.032	8008.042	1.7e+05

Omnibus:	3.693	Durbin-Watson:	1.749
Prob(Omnibus):	0.158	Jarque-Bera (JB):	3.071
Skew:	-0.328	Prob(JB):	0.215
Kurtosis:	3.534	Cond. No.	1.16e+20

Table 1. OLS Results

All of the high revenue earning conferences have low P-values and therefore are deemed significant, the bottom-feeder conferences (C-USA, Independent, Mt. West, Sun-Belt) are not statistically significant in this analysis. MedianConfSal also has little effect on coach's pay. Ratio of wins and losses is also deemed as statistically insignificant which is very interesting. Primarily, because a coach's performance should be the main indicator of how much they are paid, but that's not the case from a birds-eye-view. Digging down into the granularities of winning, DefenseScore and PointsPerGame are two metrics that pay

dividends. So, as long as a coach produces a team that is great on defense and has a high PointsPerGame metric they can expect the cash to start rolling in. Graduation rate does have a positive effect on salary, however the P-value is too high to deem reliable. This model accounts for 81.0% of the total variation in salaries.

Removing all features with a P-value > 0.05 yields an adjusted R^2 of 80.6%, and removing the Conf attribute from the 2nd model results in an adjusted R^2 of 73.1%.

Questions

1. What is the recommended salary for the Syracuse football coach?
    a. Predicted salary for Syracuse head coach: 2,548,076.0
2. What would his salary be if we were still in the Big East? What if we went to the Big Ten?
    a. Predicted salary for Syracuse head coach if still in AAC (Big East): 1,524,076.0
    b. Predicted salary for Syracuse head coach was in Big Ten: 3,432,076.0
3. What schools did we drop from our data, and why?
    a. Schools that had null values in the below attributes were removed from the dataset.

```
Shape of new matrix: (118, 10)
Columns of new matrix: ['School', 'Conf', 'MedianConfSal', 'TotalPay', 'StadSize', 'Graduation Rate
(GSR)', 'Ratio', 'OffenceScore', 'Defense Score', 'PointsPerGame']
```

4. What effect does graduation rate have on the projected salary?
    a. Graduation rate does have a positive 6-figure effect (coefficient*GSR ~ 1e6) on salary, however it is not statistically significant. In all, I'd say there is not enough evidence to support that graduation rate has any effect on salary.
5. How good is our model?
    a. Accounts for 81% of the variation in pay between coaches.
6. What is the single biggest impact on salary size?
    a. The feature with the highest correlation to total salary was stadium size.