Topic: College Football Coach Salary Analysis - how can we recommend the best salary (total compensation, minus bonus) for our next head football coach?

EXECUTIVE SUMMARY

Using a merged dataset consisting of 125 college coaches, their associated salaries, and features including team record, collected revenue and stadium capacity, researchers produced multiple predictive models using Ordinary-Least-Squares Regression (OLS). The models were evaluated through cross-validation, and the highest-performing model was used to estimate Syracuse's next football coach salary at approximately $3 million .

ABOUT THE DATA

Researchers compiled 4 sets of data from publicly available information around college football in the recent past. Only fields used in the models are listed below:

Dataset 1 – Coaches Dataset
Source: Provided by Syracuse University for this analysis

Fields	Description	Example
School	Institution in charge of program	Alabama
Conference	Governing Body for program's competition	SEC
Coach	Name of Head Coach	Nick Saban
TotalPay	Salary of Head Coach	885000

Dataset 2 – College Football Stadium Comparisons
Source: https://www.collegegridirons.com/comparisons-by-capacity/ - College Football Stadium Comparisons, cross-checked with Wikipedia

Fields	Description	Example
Capacity	Stadium Capacity where games are played	65000

Dataset 3 – Graduation Rates by School with NCAA program
Source: NCAA http://fs.ncaa.org/Docs/newmedia/public/rates/index.html

Fields	Description	Example
FGR	Federally defined Graduation Rate	0.7
GSR	Graduation Rate defined by NCAA	0.62

According to the NCAA, "The NCAA GSR differs from the federal calculation in two important ways. First, the GSR holds colleges accountable for those student-athletes who transfer into their school. Second, the GSR does not penalize colleges whose student-athletes transfer in good academic standing."

Dataset 4 – Total Revenue by College Football Program
Source: College Athletics Financial Information (CAFI) Database – Knight Commission on Intercollegiate Athletics

Fields	Description	Example
Total Revenue	Total Revenue obtained by team in Calendar Year	14,866,061.00

Dataset 5 – 2019 College Football Stats
Source: Kaggle Contributor Jeff Gallini, Original Source: https://www.ncaa.com/stats/football/fbs
https://www.kaggle.com/jeffgallini/college-football-team-stats-2019

Fields	Description	Example
Games	Number of Games played in last season	12
Off_rank	Rank of Offense Team Performance	51
Def_rank	Rank of Defense Team Performance	6

What schools were dropped from the data, and why?
Schools that failed to meet the following criteria were excluded from this analysis:
Publicly reported data on total revenue
'Air Force', 'Army', 'Charlotte', 'Georgia State', 'Maryland', 'Navy', 'Old Dominion', 'South Alabama', 'Texas-San Antonio'
Publicly available data on coach compensation
Baylor, BYU, Rice, SMU (all private universities)
Obtainable graduation rate data in the same time frame as the remaining universities
'Georgia State', 'North Carolina', 'South Alabama', 'Texas-San Antonio'

Noting the overlap in schools excluded, the final list includes:
Air Force, Army, Baylor, Brigham Young University, Georgia State, University of Maryland, Navy, Old Dominion, Rice University, Southern Methodist University, University of South Alabama, UNC-Charlotte, and University of Texas – San Antonio.

DATA PREPROCESSING

Data was obtained from the sources listed above and stored in CSVs now available on GitHub. These flat files were processed and merged using Python, where extraneous information was removed, and relevant information was transformed into their proper data types before being merged together into a single data frame for analysis.

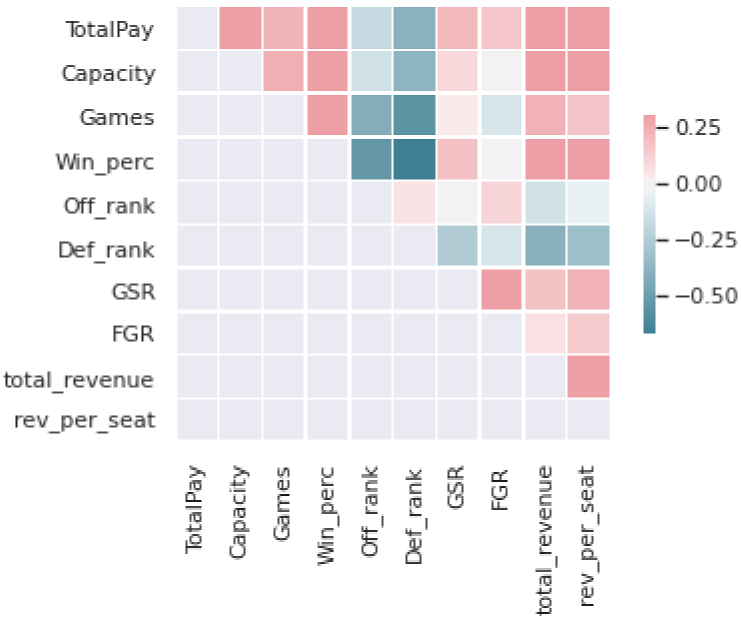
As part of this processing step, two fields were created to help account for potential multi-collinearity issues with the model

Fields	Description	Example
Win_perc	% of Last Season’s games resulting in victory	0.85
Rev_per_seat	Revenue generated per seat in stadium	301.85

DATA ANALYSIS

A high-level view of the data points and their relationship to the primary factor – Coach Compensation – proves insightful as to how these features will prove useful to the final model.

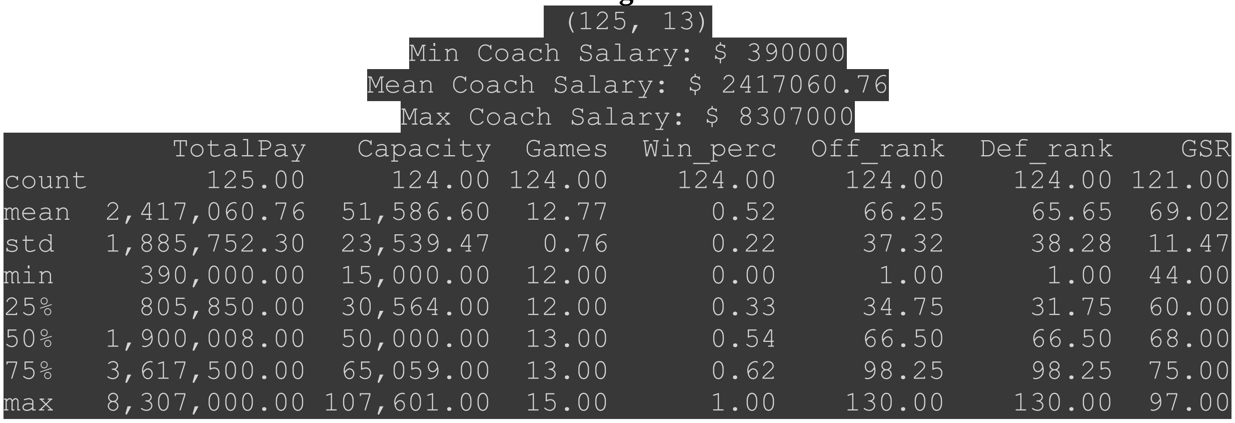
Fig. 1



Reading from the first row across (red = positive correlation, blue = negative), Compensation (TotalPay) is positively (albeit weakly <= .25) related to Stadium Capacity, win percentage, both styles of grad rate, and to the potential alternative measures, total revenue and revenue per seat. Compensation is also negatively correlated with offensive and defensive rank, which makes sense conceptually – a lower rank is indicative of higher performance.

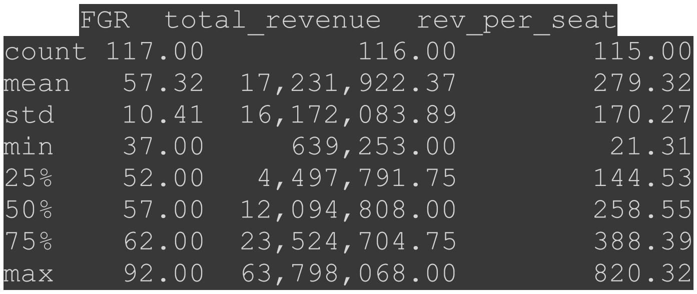
Also worthy of note is the strong negative correlation between win percentage and defensive rank – this could be a statistical sign-off on the concept that the best offense is a good defense!

Fig. 2



```
                         (125, 13)
                 Min Coach Salary: $ 390000
                Mean Coach Salary: $ 2417060.76
                 Max Coach Salary: $ 8307000
            TotalPay   Capacity  Games  Win_perc  Off_rank  Def_rank    GSR
count         125.00     124.00 124.00    124.00    124.00    124.00 121.00
mean  2,417,060.76  51,586.60  12.77      0.52     66.25     65.65  69.02
std   1,885,752.30  23,539.47   0.76      0.22     37.32     38.28  11.47
min     390,000.00  15,000.00  12.00      0.00      1.00      1.00  44.00
25%     805,850.00  30,564.00  12.00      0.33     34.75     31.75  60.00
50%   1,900,008.00  50,000.00  13.00      0.54     66.50     66.50  68.00
75%   3,617,500.00  65,059.00  13.00      0.62     98.25     98.25  75.00
max   8,307,000.00 107,601.00  15.00      1.00    130.00    130.00  97.00
```

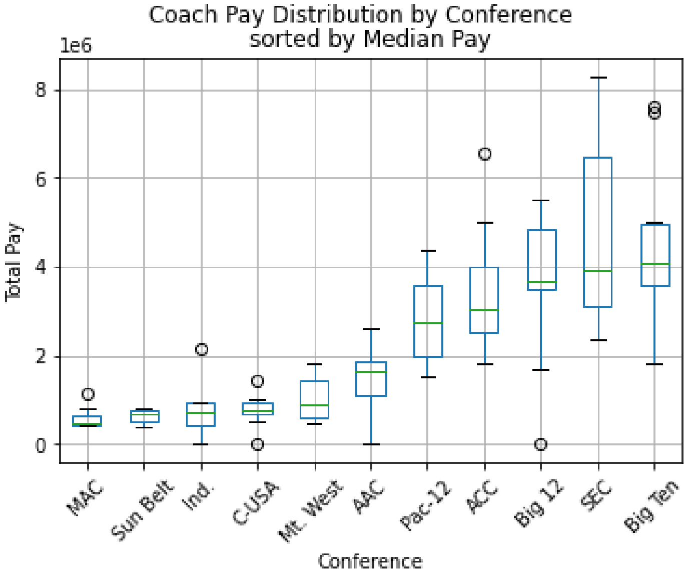
Fig. 2 cont.



```
        FGR  total_revenue  rev_per_seat
count 117.00         116.00        115.00
mean   57.32  17,231,922.37        279.32
std    10.41  16,172,083.89        170.27
min    37.00     639,253.00         21.31
25%    52.00   4,497,791.75        144.53
50%    57.00  12,094,808.00        258.55
75%    62.00  23,524,704.75        388.39
max    92.00  63,798,068.00        820.32
```

Looking closer at Salary and the makeup of the features that may influence it, it's immediately clear through an examination of the possible salary range (standard deviation, range between minimum and maximum values under TotalPay) that there is notable variation in coach pay across the nation.

Fig. 3



Noting the wide variation in coach salaries alluded to in Fig. 2, this boxplot helps to contextualize the relationship between a team's conference membership and their possible pay range. Syracuse's membership in the ACC places it in the higher half of conferences' pay distributions, though both the wide confidence interval in the SEC pay distribution and cluster of outliers in the Big Ten are also of note. It will be essential to include conference in any model estimating coach pay, as the variation here is stark.

Fig. 4

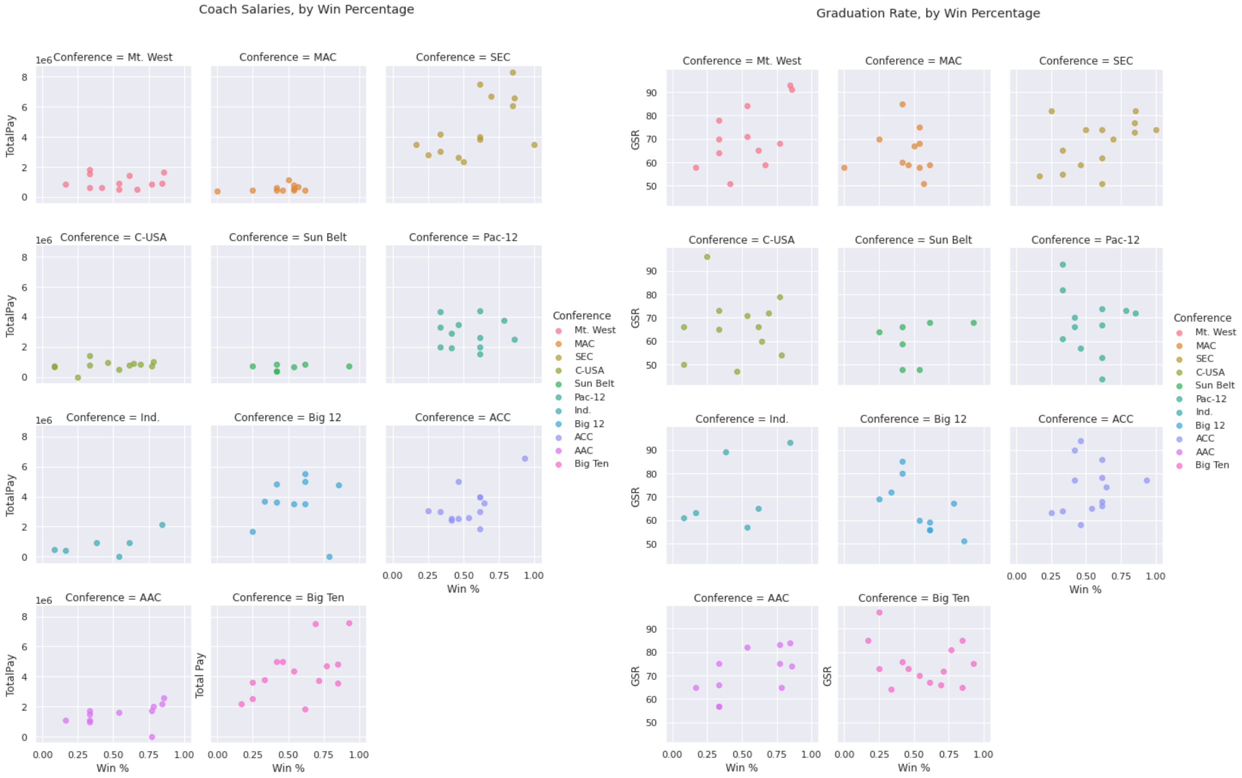


Figure 4 allows for a comparison of Coach salary and their school's associated graduation rate respectively, both in terms of their relationship to the team's win percentage in the previous year. On the left, there appears to be a positive relationship between coach salary and win percentage in the SEC (top right corner) and the Big Ten (bottom center). This relationship is less defined in conferences like the Mt. West (top left corner) and Pac-12 (second row, right). Worthy of note on the right is that winning teams have more graduates in the SEC, though the opposite seems true in the Big 12 (Third Row, center).

Given the wide variation of multiple features on a conference-by-conference basis, it is vital to run multiple models and compare their ability to predict out-of-sample before estimating an optimal coach salary for Syracuse.

MODELING COACH SALARY

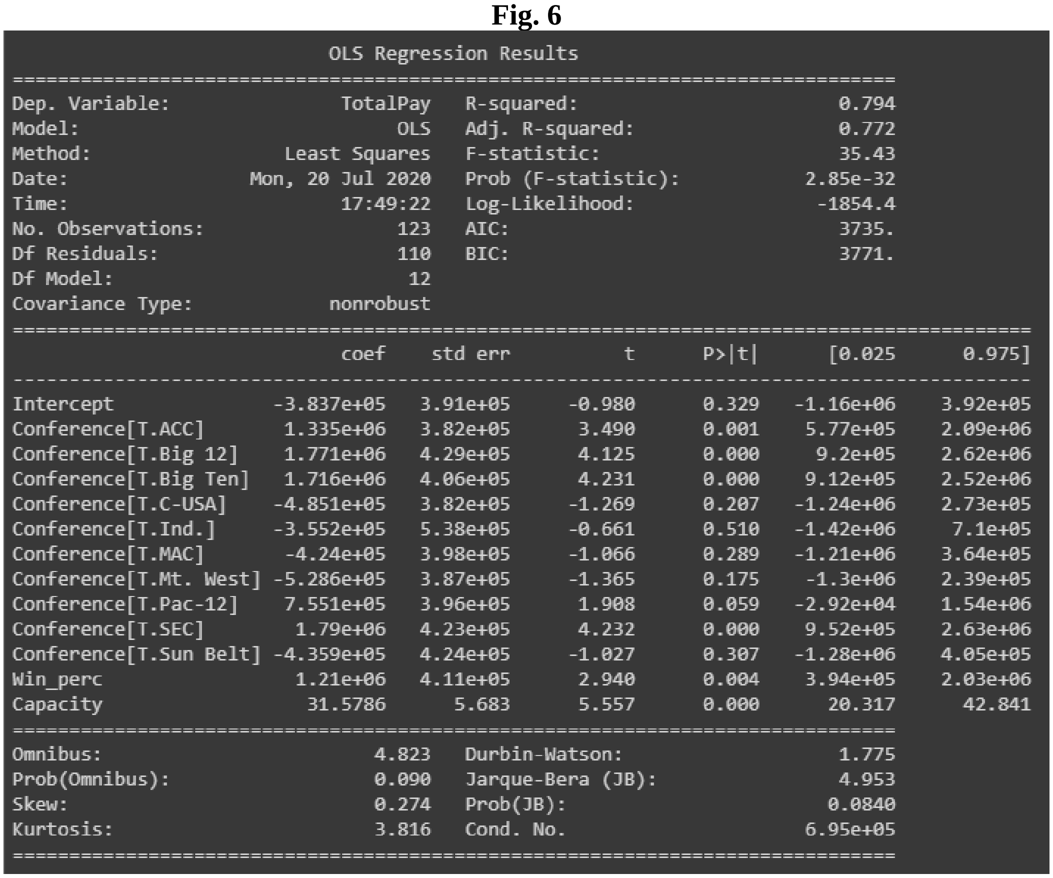
With the goal of recommending the best representative salary for Syracuse's new Head coach, four models were produced from this data. After all models were developed, they were evaluated based on both the ability of each model to accurately predict salary on a training and test set of the provided data. yielding the following results:

Fig. 5

Method	Training Set Result	Test Set Result
Conference Only	0.71	0.59
Stadium Size (Capacity), Conference, Win Percentage	0.82	0.69
Revenue Per Seat, Conference, Win Percentage	0.83	0.66
Stadium Size, Conference, Win Percentage, GSR, Rev Per Seat	0.84	0.63

Though the final model (with the most included variables) performed the best on the training set (0.84), its predictive power on the test dataset was the second lowest (0.63). Model 2 was ultimately selected to generate a coach salary estimate for Syracuse, as the slight drop in training prediction (.84 to .82, or -.2) was offset by the increase in predictive accuracy on the test data set (0.63 to 0.69, or +.6). Fig. 6 is a summary of Model 2 ran against the entire dataset.

Fig. 6



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.794
Model:                            OLS   Adj. R-squared:                  0.772
Method:                 Least Squares   F-statistic:                     35.43
Date:                Mon, 20 Jul 2020   Prob (F-statistic):           2.85e-32
Time:                        17:49:22   Log-Likelihood:                -1854.4
No. Observations:                 123   AIC:                             3735.
Df Residuals:                     110   BIC:                             3771.
Df Model:                          12
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept              -3.837e+05   3.91e+05     -0.980      0.329   -1.16e+06    3.92e+05
Conference[T.ACC]       1.335e+06   3.82e+05      3.490      0.001    5.77e+05    2.09e+06
Conference[T.Big 12]    1.771e+06   4.29e+05      4.125      0.000     9.2e+05    2.62e+06
Conference[T.Big Ten]   1.716e+06   4.06e+05      4.231      0.000    9.12e+05    2.52e+06
Conference[T.C-USA]    -4.851e+05   3.82e+05     -1.269      0.207   -1.24e+06    2.73e+05
Conference[T.Ind.]     -3.552e+05   5.38e+05     -0.661      0.510   -1.42e+06     7.1e+05
Conference[T.MAC]       -4.24e+05   3.98e+05     -1.066      0.289   -1.21e+06    3.64e+05
Conference[T.Mt. West] -5.286e+05   3.87e+05     -1.365      0.175    -1.3e+06    2.39e+05
Conference[T.Pac-12]    7.551e+05   3.96e+05      1.908      0.059   -2.92e+04    1.54e+06
Conference[T.SEC]        1.79e+06   4.23e+05      4.232      0.000    9.52e+05    2.63e+06
Conference[T.Sun Belt] -4.359e+05   4.24e+05     -1.027      0.307   -1.28e+06    4.05e+05
Win_perc                 1.21e+06   4.11e+05      2.940      0.004    3.94e+05    2.03e+06
Capacity                  31.5786      5.683      5.557      0.000      20.317      42.841
==============================================================================
Omnibus:                        4.823   Durbin-Watson:                   1.775
Prob(Omnibus):                  0.090   Jarque-Bera (JB):                4.953
Skew:                           0.274   Prob(JB):                       0.0840
Kurtosis:                       3.816   Cond. No.                     6.95e+05
==============================================================================
```

INTERPRETING THE MODEL

How good is the model?

As previously noted, four models were compared by each's ability to account for the variance in both the train and test datasets created. The selected model was able to account for 82.4% of the variance in the test dataset, and 68.9% of the variance in Total Pay in the test dataset, the best overall performance of the models considered. This would fall under good but not great, an acceptable result knowing the limits of prediction using OLS Regression.

What is the recommended salary for the Syracuse football coach?

When Syracuse's features are entered into this model, the resulting predicted salary is **$3,010,342**. Given that the previous coach pay was closer to $2.4 million, the model-recommended salary would mean a new head coach should be offered approximately *25% more* than their predecessor.

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

This scenario is calculated by removing the modeled "ACC effect" and estimating from the baseline value in the model, (AAC), a reduction of approximately $1,334,772, or approximately **$1,675,570**.

Big Ten?

This scenario is calculated by removing the modeled "ACC effect" from the estimated salary, a reduction of $1,334,772, followed by an addition of the associated "Big Ten effect", $1,716,093.33, or approximately **$3,391,663.33**.

What effect does graduation rate have on projected salary?

In short, it doesn't--the effect isn't statistically significant when included in the selected models, and its inclusion in said models only resulted in a decreased predictive accuracy on the test dataset.

What is the single biggest impact on salary size?

The single biggest impact is the conference in which a team plays, demonstrated both in the descriptive analysis above and made clear in the modeling exercise. Over 60% of the variance in this dataset could be explained by conference placement alone, indicating it plays a tremendous role in a coach's compensation.

FINAL CONCLUSION

With the use publicly available datasets, data blending/preparation via python and the power of Ordinary-Least-Squares Regression (OLS), multiple predictive models were created and evaluated through cross-validation to estimate the recommended salary for Syracuse's next head coach. This recommended compensation estimated at approximately **$3 million**, and graduation rate was determined to be a statistically insignificant factor in this estimation.