Aurelia Arnett

IST 718 – Big Data Analytics

Coaches

**Introduction**

For many universities, football can serve as a large source of income for the school and in the case of Division I, coaches can make millions of dollars a year. As the current football season is coming to an end, the Syracuse athletic director would like to determine the best salary for head football coach Dino Babers for next season by considering the different factors that contribute to both athletic and academic success. This case study explores current salaries across all Division 1 universities, graduation rates, stadium size, rankings, and enrollment to recommend next year’s salary.

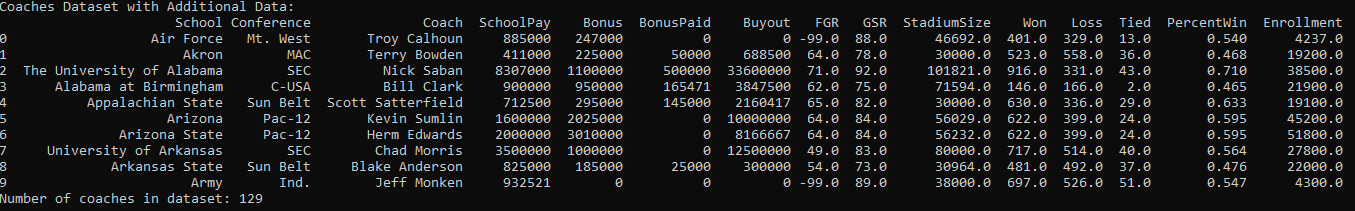
The caveat to this recommendation, or a weakness to the analysis, is that the data is collected from a variety of different years. In future research, we would obtain a larger amount of historical data and create a time series or collect all data within the same year. For the purpose of this analysis we assume temporal consistency has no impact on the strength of the models.

**About the Data**

A structured csv dataset containing Division 1 coaches, their school and conference, their salaries, and any potential external sources of income is first provided and read into the program. A for loop reads the csv file, formats the coaches data into a pandas data frame, and cleanses the salary and income data to remove non-numerical characters and convert these to integers for calculation.

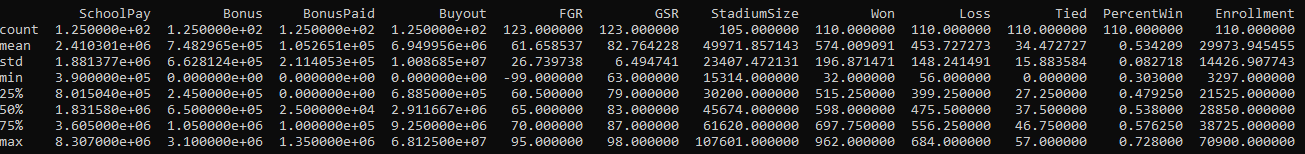
Graduation rates are read into the program from downloadable tsv files provided by the NCAA and also formatted into a pandas data frame with only the school name, FGR rate, and GSR rate pulled in as the column data. The school names in both the coaches and the graduation rates datasets are printed to compare the spelling of the schools and how much of the title is included. A series of .replace() functions replaces mismatched school names in the coaches dataset with an updated spelling of the school name from the graduation rates data frame, then the two are patched together into one using a lookup function that matches the graduation rates by school name. A new coaches data frame is created for analysis.

Three more datasets containing stadium size, win-loss records, and enrollment information are read into the program from Wikipedia tables using the Beautiful Soup method. Each table is formatted into its own pandas data frame and matched to the new coaches data frame using the same lookup function method. The final data frame reads the following:



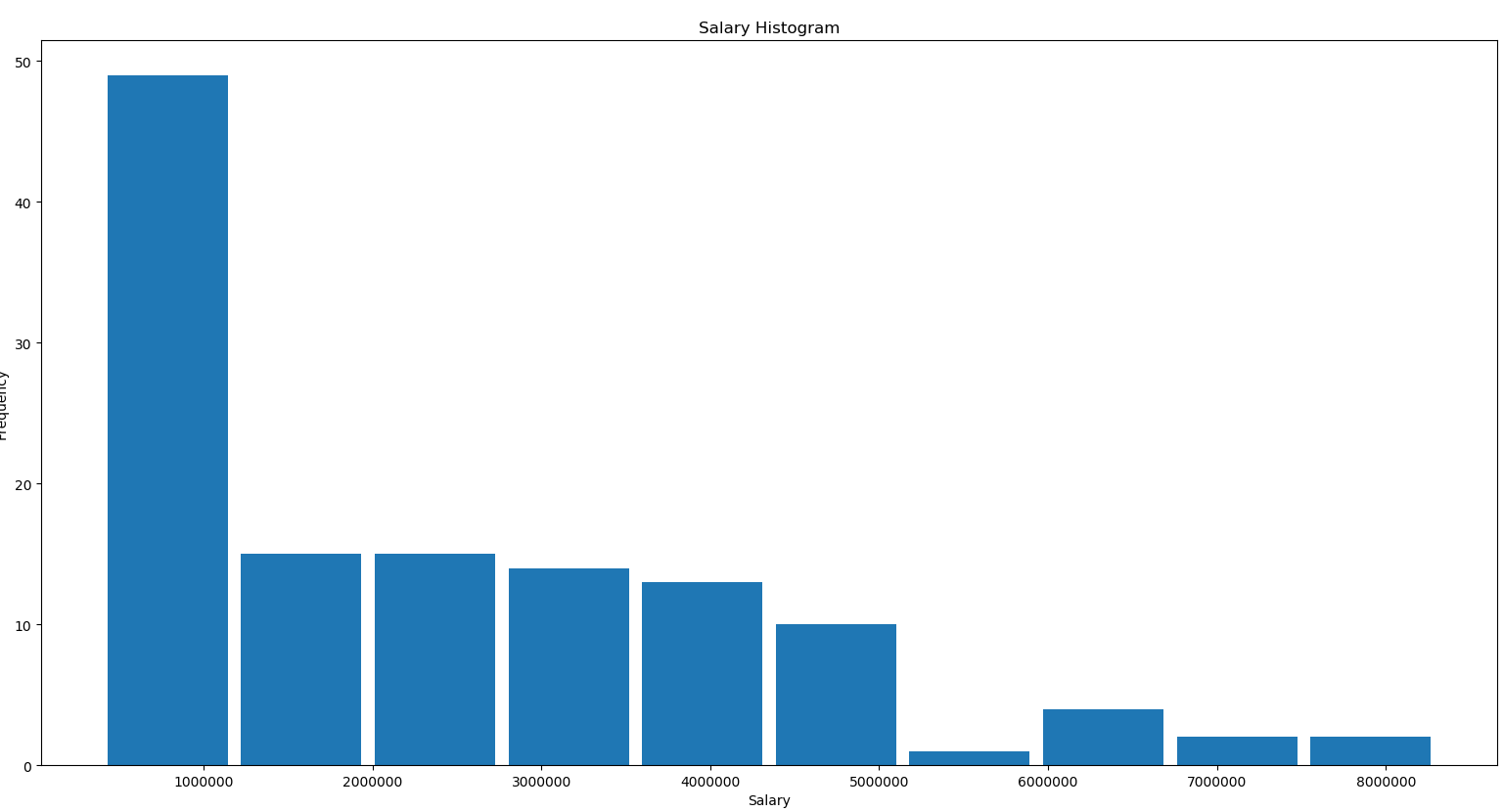
Coaches data frame with additional data included for analysis

The .describe() function is run on the data frame to explore the min, max, and mean values for the variables of all of the coaches in Division 1 schools. In the first round, it is found that some coaches do not have salary (notated as the SchoolPay variable) which could skew the data modeling. These schools (Baylor, Brigham Young, Rice, and Southern Methodist) are removed from the data frame and the descriptive statistics reads the following:



Descriptive statistics output for coaches data

The descriptive statistics output shows that the mean salary for coaches in Division 1 schools is $2,410,301 and we see a large range with the min salary being $390,000 and the max being $8,307,000. This indicates that the mean may not accurately represent the dataset as a whole and the other variables will be important to consider in the model, such as win-loss records data, graduation rates, and enrollments for example where we see smaller ranges. A histogram on the SchoolPay variable shows a consistent conclusion with the descriptive stats as the data is skewed right, meaning the mean salary is larger than what most coaches in the data set make and multiple factors aside from just conference must be considered:



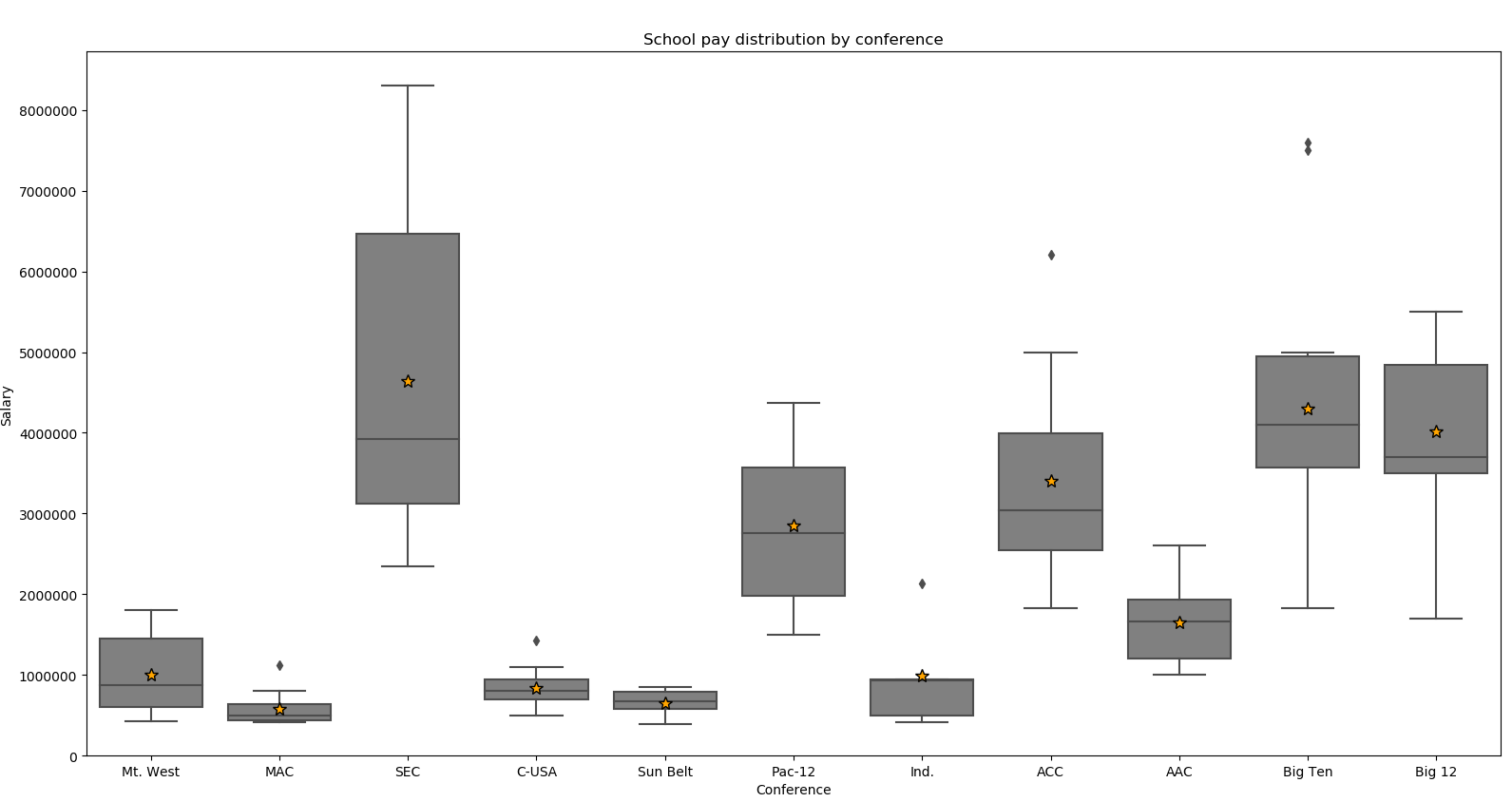
Representation of salary skewness among coaches in Division 1

The overall conclusion from the histogram is that this year’s current salary information is not valuable on its own when considering next year’s salary and outlier salaries exist which raises the mean. Other variables to consider can be represented visually, and the next series of graphics outputted are a series of trellis and lattice plots. These plots are developed to explore if patterns exist between the numerical, non-monetary, categorical variables individually with the coaches salary current salary, to begin considering which variables will be important to include in data modeling:

|  |  |  |
| --- | --- | --- |
| Salary trend as enrollment increases | Salary trend as stadium size increases | Salary trend as percent wins increases |
| Salary trend as GSR increases | Salary trend as FGR increases | Salary trend as FGR increases, outliers removed |

The trellis and lattice plots show that generally, as the school enrollment, stadium size, and coaches percent-wins increases, a coaches salary will be higher. For graduation rate, there seems to be no relationship between the FGR rate and a coaches salary though the majority of athletes graduate with an FGR score around 60-70. This seems to be consistent even when outlier schools (Air Force, Army, and Navy) are dropped from the data set. For GSR scores, there seems to be a slight relationship between the GSR score and coaches salary, though the majority of athletes have a GSR score around 80-85. When building a model, it will be important to consider how the enrollment, stadium size, and percent wins impacts the significance of the model. It may be interesting to explore how graduation rates may impact the significance of the model, as the graduation rate varies by conference.

A boxplot representing the relationship the relationship between the conference and the coaches salary is built to understand where additional outliers may exist:

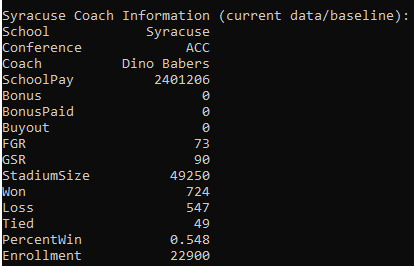


Boxplot representation of the relationship between conference and salary

The boxplots show that generally across each conference, the mean salary for coaches within each conference falls near or exactly inline with the median salary, accurately representing the data and showing a fairly small range. It is important to note that while this relationship exists, a range still exists indicating that other factors impact the salary.

This boxplot also shows that within a few conferences, a large range exists indicating the mean doesn’t accurately represent the conference salary, other factors become more crucial, and outlier coaches exist. For this analysis, we consider removing the outlier coaches in the SEC conference for data modeling and in a perfect world would also remove outlier coaches in the ACC, Big Ten, and Big 12 conferences as well. Due to a small dataset, we will not remove coaches in the ACC, Big Ten, or Big 12 conference as the mean isn’t as far from the median as seen in the SEC conference.

Finally, the Syracuse coaches data is printed to understand how present the data compares to the descriptive statistics on all coaches and to set a baseline of where coach Dino stands today:

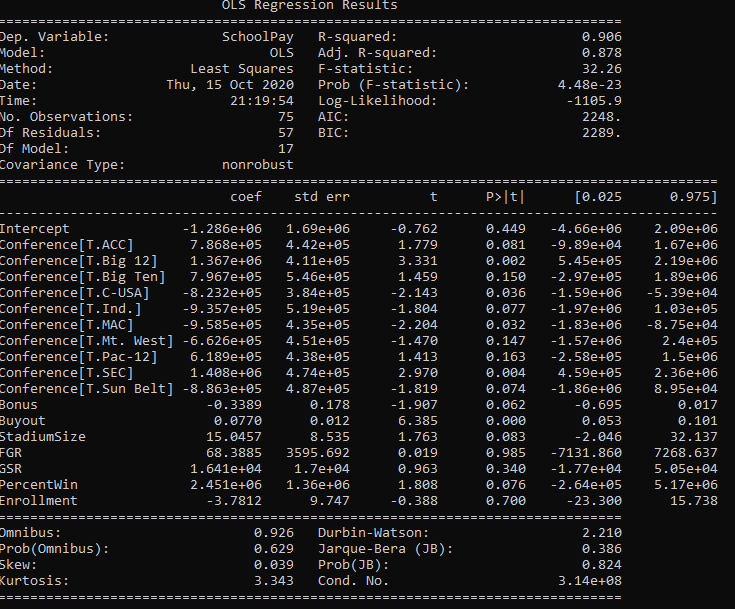


Syracuse coach information for this season

This output shows that this year, Dino makes almost exactly the mean of the salary for all coaches as a whole, which indicates that Dino may not be making the appropriate amount since the mean doesn’t strongly represent the dataset.

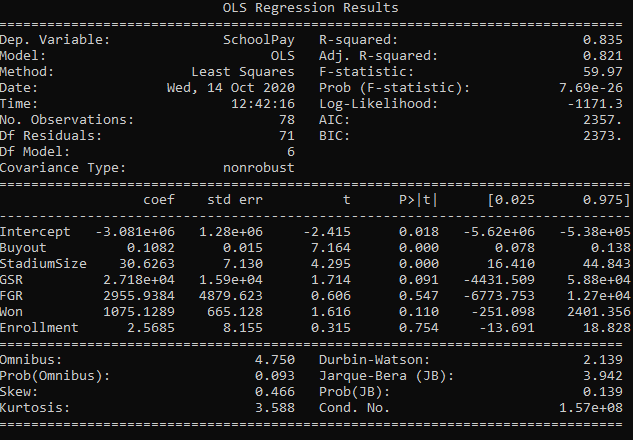
**Data Analysis**

To better understand factors that impact coaches salaries, a series of models are built, starting with a model that considers all of the variables (with collinear variables such as Won and PercentWin removed). The models are first applied to all coaches (no outliers removed) and then compared to models with the outlier coaches removed. The model considering all variables serves as a baseline only (no regression results are stored for later comparison) and reports the following:



Regression results for a model considering all variables

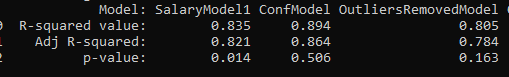
While this model returns a high R2 value, accounting for most of the error in the model, the p-value shows that this model is not significant and looking at the variables deeper, conference doesn’t seem to be a good variable to consider as the model compares each conference individually as its own categorical variable. This indicates that coaches within a conference could be considered amongst other coaches within that conference rather than within the Division 1 as a whole. This model also shows that graduation rates and enrollment seem to be insignificant variables as their p-values are much larger than 0.05. Model 1 is tuned by removing the conference variable permanently and considering the removal and addition of the other variables. Model 1 is tuned to print and store the following results:



Model 1 regression results

While model 1 shows to have a smaller R2 value than a model with all variables, indicating the potential of more error to exist, the p-value proves the model to be significant at 95%. The variables included in this model again show the FGR and enrollment variables to be insignificant, however the removal of these causes a slight decrease in R2 and slight increase in p-value so these variables will remain in this model to avoid error as much as possible. This model is applied to the coaches data (with no outliers removed) and the suggested output salary for Dino at Syracuse is $2,509,127.52 which is slightly higher than his current salary of $2.4M.

Two additional models are built for analysis and the following regression results are outputted with suggested salaries for Dino, based on different datasets:



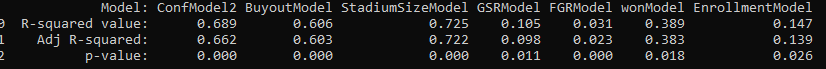
Regression results summary table, based on 3 models



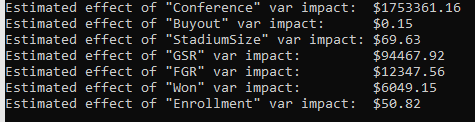
Suggested salaries for Dino, based on different models and datasets

The model ‘SalaryModel1’ model represents the first model discussed above. The ‘ConfModel’ model represents a model with the same variables included as model 1 and the addition of the Conference variable. This model was built to explore how Dino’s salary would look if the Syracuse athletic director decides to evaluate Dino’s salary compared to other coaches within the conference, rather than against all Division 1 coaches. While the R2 value increases, a very insignificant p-value is returned, suggesting this is not a good model to evaluate Dino’s salary. It is, however, applied to two different datasets: the first being only coaches in the ACC conference and schools in the Big Ten conference (with Syracuse modeled as a Big Ten school in this data frame). The salaries table shows that the ‘ConfModel’ suggests that Dino’s salary should be about $2.7M if the athletic director compares salaries to other coaches within the ACC conference and about $2.8M if Syracuse fell within the Big Ten conference. The final model built ‘OutliersRemovedModel’ represents a model with the same variables as model 1, but trained on a coaches data frame with outlier schools (such as those with an FGR score of 0 and extreme cases of high salaries in the SEC data frame) removed. We see an R2 value close to model 1, however this model showed to be insignificant as the p-value is larger than 0.05. It is not recommended to use this model as the final, though it is applied to the coaches data with outliers removed and suggests Dino should make $2.3M. In the real world, we would want the outlier coaches removed, however in this case that reduced our already small dataset even further, which could have been an impact on the model. In this analysis, model 1 would be the recommended model when considering Dino’s salary.

To understand the impact of individual variables, an additional 7 models are built to represent models based on one individual variable and the following regression results and salary impact is printed:



Regression results for individual variables considered in data modeling



Estimated effect of individual categorical variables on salary

All variables seem to prove significant on their own in data modeling, though in the real world and as demonstrated in model 1, do not serve powerful on their own.

**Recommendations**

The recommended salary for Dino at Syracuse is $2,509,127.52 based on the most significant model which had an R2 value of 0.835 (accounting for error within 83.5% of the data) and a significant p-value of 0.014. If Syracuse was in the Big Ten, Dino’s suggested salary is $2,828,073.80, though the model did not prove to be significant. This analysis does not provide a recommendation for Dino’s salary if Syracuse was still in the Big East, though research from 2011 suggests that the average head coaches salary for Syracuse was $261,016 which suggests that any data modeling would recommend a much lower salary suggestion. Further analysis would require data from the Big East conference that aligns with the variables collected for the final coaches data frame. For this analysis, that data could not be obtained.

During data exploration, 4 schools were removed from the coaches dataset because the coaches did not have salary information included. These schools were Baylor, Brigham Young, Rice, and Southern Methodist and this data became the primary coaches data frame. In model 1, no additional schools were dropped from the data because the dataset overall was too small and dropping schools resulted in weaker regression results. In the outliers model, schools with FGR scores of 0 were dropped from the dataset as these schools visually showed an impact on the relationship between graduation rates and salaries. These schools were the military schools Air Force, Army, and Navy. Other schools dropped in the outliers model were schools in the SEC conference that had coaches salaries well over the mean and median, resulting in an extreme range. These schools were Alabama, Auburn, University of Georgia, and Texas A&M and were removed to better distribute salaries within the SEC conference, impacting the mean salary for the coaches data as a whole.

Model 1 suggest that graduation rate can be a both significant and insignificant variable when considering coaches salaries. In model 1, the graduation rate variable GSR is significant with a p-value of 0.091 while FGR is insignificant at 0.547. The models considering only the FGR and GSR variables as independent variables show significant p-values however extremely low R2 values, indicating high error and are not recommended when considering a salary. These models, however, indicate that GSR has an impact of $94,467.92 and FGR has an impact of $12,347.56, and the single biggest impact on salary size seems to be conference with an estimated impact of $1,753,361.16 which is nearly half the mean of all coaches salaries.