Big Data Analytics Homework One

Allison R. Hollingsworth Deming

IST 718: Big Data Analytics

Homework One Report

***Introduction***

Many things change in the Fall, especially in a college town. First, the population swells from the students returning to town, who are new to their alma mater and excited about everything the college experience brings. Other students have been at the school for weeks already. These are the athletes, particularly the football players, trainers, coaches, and significant support staff. They have spent the last few scorching weeks of Summer honing their craft before the football season itself starts. This is generally around the second week after school begins. This is far from being just a sport for student-athletes; for many, it is religion, and for most Division I schools, it is also big business. Depending on the school, the salary for the head coach can range from just under $100,000 to well over ten million per year. This depends on many things, but how does a school appropriately pick a salary for this vital position? With football bringing in large amounts of cash not just for its own operations but for the entire school, it is essential to get the right person at the right price. Here, multiple models will be produced to estimate what coaches in each conference should be paid, and more specifically, what Syracuse’s Coach, Dino Babers, should make at Syracuse and what he could make elsewhere.

***Obtaining the Data***

More is needed than just the dollars and cents of a wage to develop a model to predict the best salary for a head football coach. Certain schools may have caps to which they can pay their coach. Others with more competitive and extensive programs might negotiate contracts over multiple years with additional bonuses based on performance. One recently published example of this complexity is that of Dabo Swinney at Clemson University. He joined the school as the head football coach in 2008 with a team that had a win percentage of about 70%. Over the next ten-plus years, that percentage went up to 85%. In 2022, Coach Swinney signed a ten-year, $115 million contract to keep him at Clemson. That comes to about $15 million per year plus a bonus structure. The bonus structure includes $50,000 for playing in the ACC Championship Game (their conference), $200,000 for winning this game, and $75,000 each year for winning at least eight games. Playing in a New Year’s non-CFP (College Football Playoff) semifinal bowl: $150,000; for a CFP appearance, another $150,000; for a CFP semifinal appearance: $250,000; for a CFP championship appearance: $250,000; for a CFP championship win: $350,000 Single-year academic progress rate (APR) ≥ 950: $75,000; or Single-year APR ≥ 975: $100,000; if awarded ACC Coach of the Year Award: $25,000, and lastly if he receives the National Coach of the Year Award: $50,000. This shows how complex the bonus structure can be for a single coach. If Clemson decides to part ways with Dabo Swinney during this contract, toward the beginning of the ten years, Clemson will pay $64 million, which decreases to a low of $57 million toward the end of the contract. Coaches may also earn money through endorsements, licensing, appearances, and many things not factored into this analysis. In addition to the salary data, our study incorporates three different data sets: college graduation rate, football stadium, and the football teams’ wins and losses since 2010. It should be noted that the most recent years are not included in the data, but where that data is to be obtained, it could be filtered into the model for more appropriate (up-to-date) results.

The recent data about Coach Swinney’s contract is on the higher end of the pay scale, but that is not why it is included. It is included because it gives a basis for the data sets chosen to put into our model. Each of these data sets can be seen reflected in the 2022 Swinney contract. It shows what is essential to at least one school within the group we are analyzing. Additionally, it gives justification for win percentage and graduation rate being included along with the other monetary variables and how those are weighted within the overall contract.

*Data Set One*: The professor posted the Coaches’ data on the following website: https://github.com/2SUBDA/IST\_718. This has 129 entries and eight columns. The columns have variables related to the school’s football teams, such as School (The identity of the school), Conference (with which group they compete during the regular season), Coach (designated head coach), SchoolPay (the school pay), TotalPay (total pay of the coach), Bonus (any bonus pay contracted), BonusPaid (actual bonus paid out, generally based on performance that year), AssistantPay (what the assistants are paid), and Buyout (the amount the coach gets if the school decides to buy out their contract).

*Data Set Two*: The colleges’ graduation dataset is available at the following website:

https://www.ncaa.org/sports/2016/12/14/shared-ncaa-research-data.aspx. The data set

contains columns detailing the colleges’ football division, which depends on size and contracts. Additional data contained therein is if the school is private or public. The next piece designates historically black colleges and universities (HBCUs) and the graduation rate of their different sports programs. The model produced will only include that of the football teams.

*Data Set Three*: The football stadium dataset is also from GitHub:

https://raw.githubusercontent.com/gboeing/data-visualization/main/ncaa-footballstadiums/

data/stadiums-geocoded.csv. This dataset contains variables that describe stadium

features such as the city, state, stadium name, capacity, build, and location.

*Data Set Four*: The football teams’ win and loss percentages are available at: https://www.teamrankings.com/ncf/trends/win\_trends/?range=yearly\_since\_2010. The dataset includes the teams and their win-loss record overall.

***Exploratory Data Analyses***

The four datasets were imported into Anaconda with Python using the PANDAS package. When each of the individual datasets was processed and explored, changing the strings’ columns to integer values. Overall, the data frames contained integers, objects, and floats. Lastly, NaN values were changed to dashes for easier processing. Each data frame requires a lot of filtering of columns to those that are needed for the analysis and those that are not. The exception is the wins and losses data frame, which contains the historical data of wins and losses for each team; all columns were kept and imported. The stadium dataset tells the size and upgrades to the stadiums and multiple columns that indicate the school's investment in the football team facilities. Lastly, the graduation set gives the graduation rate of student-athletes, but the only part that will be used in this analysis is that of the football players.

Graphical user interface, application

Description automatically generated

Figure 1: Stadium Data Column Names and Types

Table

Description automatically generated

Figure 1.1: First Four Rows of Stadium Data

Table

Description automatically generated

Figure 1.2: Columns and Rows of the Coach Data

Table

Description automatically generated with medium confidence

Figure 1.3: First Five Rows of the Coach Data Frame

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1.4: Wins and Losses Data Frame Information

Table

Description automatically generated

Figure 1.5: Graduation Date Columns and Type

Graphical user interface

Description automatically generated

Figure 1.6: Graduation Rate Data Frame Information

***Building a Uniquely Combined and Clean Data Frame***

Building this data frame was complex. It turned out that merging all four original datasets into a new data frame was the best idea. Several columns needed to be renamed to merge with the other data frames' columns properly. For example, the “School” column needed to be changed to “team” in the Coaches dataset to match the other data frames. Each had a column labeled “team” with the same information. Next, the SPORT column was filtered to the MFB column to capture male football players. The column SCL\_NAME needed to be scrubbed of common phrases such as “University” and “of” along with punctuation, then renamed as “team” like the other data frames. Everything was merged, and the data frame needed for this analysis was completed. Some schools ended up dropped from the data frame following the merge. The reason for this is unknown, but the number was small and likely due to inconsistencies between the different data frames. In the end, there were 50 rows and 25 columns with the following named columns: TEAM, DESCRIPTION OF TEAM, STADIUM, CITY, STATE, WIN-LOSS RECORD, and WIN PERCENTAGE.

Table

Description automatically generated

Figure 2: Final Data Frame Columns and Type (Partial)

***Scrubbing the Data Frame***

Additional scrubbing needs to be performed before the final data frame is ready to use. For example, the columns which should be numbers such as ‘Totalpay’, ‘Schoolpay’, “Bonus’, ‘BonusPaid’, ‘AssistantPay’, ‘Buyout’, and “Win %’all need to be numeric columns, Also ‘Win %’ needs to be divided by ten to get what is needed for the analysis. Three functions were made for scrubbing. The first will call items column names that contain NaN or empty row values.

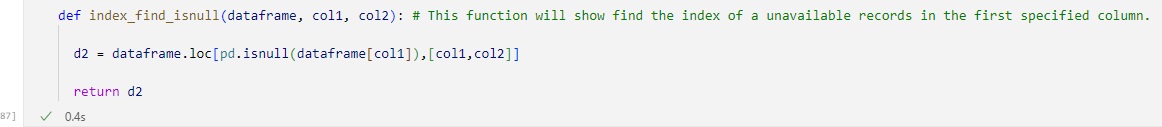


Figure 3: Function 1 for Scrubbing

**Graphical user interface, application

Description automatically generated**

Figure 3.1: Columns with NaN Rows

The second function that was made takes those NaN values from the numeric data frames and replaces them with the mean of the conference. This is true except for the Independent Conference, which is alone in its category. In this situation, it was determined that a combined mean of all teams would be best for NaN values.

Graphical user interface, application, Word

Description automatically generated

Figure 4: Function 3 Fill s NaN Rows

The third and final function that was made for data scrubbing out new and improved combined data frame uses our second function and does that final filling of NaN values in the columns with the appropriate mean discussed above. As can be seen from the code, ‘mean2’is calculated from the TotalPay and Conference variables. Once complete, all three functions have been run through those money variables that need to be numeric and cleaned, such as TotalPay and Bonus.

Graphical user interface, application

Description automatically generated

Figure 5: Function That Cleans and Replaces the Bonus Using ‘mean2’ and Conference

Graphical user interface, text, application

Description automatically generated

Figure 6: Using Cleanmerge to Take a Look at our Columns, Type, and NaNs in the Data Frame

Table

Description automatically generated

**Visual and Numeric Exploratory Data Analysis**

Sometimes the best way to see how your data is distributed is to plot the data quickly and eloquently. Early in this project, variables were looked at in one or two dimensions only to see the distributions and variability of a coach’s salary within or between the different conferences. By looking at the information, those conferences, such as the SEC with large franchises, spend tons of money and value their football program. These programs include but are not limited to, Alabama, Ole Miss’, Florida State, and Florida. While other powerhouse franchises (such as the example discussed above at Clemson above) exist in other conferences, there seems to be a culture of more significance being better in all areas of the SEC. The ACC has a lower salary average than the SEC. That is interesting to know, but what is more interesting is if the full coach’s salary affects the winning percentage and vice versa. Below is a stacked bar chart that is colored by conference, has an x-axis of the coach’s total pay (divided by $100,000), and a y-axis of schools count.

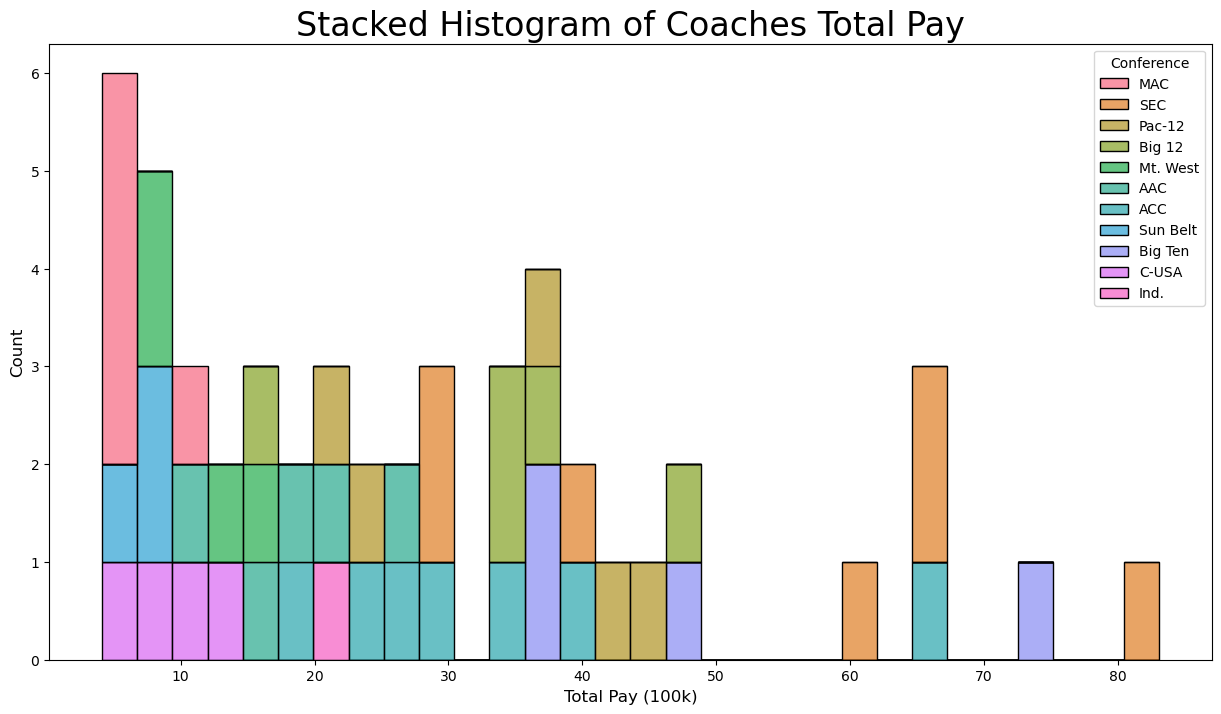


Figure 7: Stacked Histogram of Coach’s Salaries by Conference

The first and most obvious observation is that there is a significant right skew of the data, with several outliers on the right. The four most right-hand columns are the highest-paid coaching jobs, and the burned orange color filling makes up most of these high salary bars.

Chart, box and whisker chart

Description automatically generated

Figure 7.1: Boxplot of Coach’s Salary By Conference

Above, in Figure 7.1 is a Boxplot of each conference on the x-axis and salary divided by 100K on the left-hand or y-axis. The thick lines in the middle of each rectangular box represent the median of the data rather than the mean; This makes for easy comparisons across the conferences for the middle salary. This is less susceptible to outliers when compared to the mean. The ACC, Big Ten, and Pac-12 have the highest three medians in this representation; It is simple to visualize what each conference is paying with this type of graph.

Chart, scatter chart

Description automatically generated

Figure 7.2: Coach Salary Versus Win Percentage by Conference

This plot is a scatterplot of the win percentage on the x-axis and the coach’s salary on the y-axis. Encoded within the color of the data is the conference the school belongs to, and the size of the dot is the Total Pay variable for the coaches. The first thing seen is a linear relationship that looks moderately positive, with a low win percentage predicting a lower salary. The higher the win percentage, the more one sees a higher salary. This plot provides a good start for a linear model down the road.

Chart, scatter chart

Description automatically generated

Figure 8: Graduation Rate Versus Salaries By Conference

This plot is like the one above, except the win percentage is removed, and the Graduation Rate is substituted. In the Clemson contract reviewed above, Swinney would receive bonuses for improving academic performance amongst his players. This is relatively rare regarding the importance of most big-name coaches and schools. It should be noted that the program at Clemson works with many NFL-bound students graduating in three years and another large group that completes a master’s degree in addition to their bachelor’s all in four years before eligibility runs out. Our plot does not show a linear relationship, but with many more variables, such as Public or Private, something could still be found hidden in this data. Further explanation is needed.

***Correlation and Regression Models***

As mentioned above, some things may be hidden in our dataset. It is known that there is a linear relationship between the primary variable we are trying to predict (the coach’s salary) and at least one of the other variables (win percentage). Now, a correlation matrix can be run to show how all variables relate to each other in terms of correlation strength and direction.

Chart, bar chart

Description automatically generated

Figure 9: Correlation Matrix of the Data Frame

***Modeling the Data***

Three models were used to investigate the compiled data frame of coaches’ salaries in more depth. This was done with multivariate regression, with one output variable and multiple input variables. The first one used TotalPay as the main variable we are trying to predict. The predictor variables were Win Percentage, SCL\_PRIVATE (whether a school is public or private), GSR (graduation rate of football players), and Bonus. The results can be seen in Figure 10.0. Overall, the Adjusted R Squared, or variance accounted for in TotalPay by the predictor variables, was 58.4%. When fit to the testing data compiled earlier in the analysis, the accounted-for variance was shown to be .329.

A screenshot of a computer

Description automatically generated with low confidence

Figure 10: First Model Using Multivariate Regression

The second model, which output can be seen entirely in Figure 10.1 below, was similar in the result but was an entirely different model. TotalPay was still the primary variable we were trying to predict, but this time using Win Percentage, Capacity, and Graduation Rate. The Adjusted R Squared for the model was 93%, with the testing data accounting for 48% of the variance. Unfortunately, there was an error in the additional output that the model had high multicollinearity, which means that at least two of the variables are far too much alike numerically via correlation to both be in the analysis. Some technical errors can cause this, but it is likely because of variable correlation.

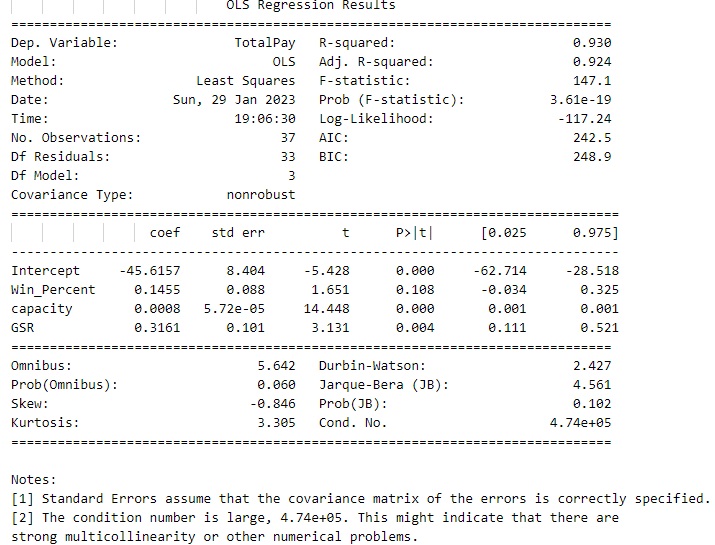


Figure 10.1: Second Model Using Multivariate Regression

A third model was created; the full results can be seen in Figure 10.2 below. This model tried to predict TotalPay for a third time, with Win Percentage, capacity, Graduation Rate, and built. The results from this analysis can be seen below in Figure 10.2. The Adjusted R Squared Value was on par with the other two analyses, at 92.2%, and the amount of variance accounted for in the testing set was 47.2%. Unfortunately, there was an error in the additional output that the model had high multicollinearity, which means that at least two of the variables are far too much alike numerically via correlation to both be in the analysis. Some technical errors can cause this, but it is likely because of variable correlation.

A picture containing text, receipt

Description automatically generated

Figure 10.2: Third Model Using Multivariate Regression

***Conclusion and Final Thoughts***

A picture containing text, receipt

Description automatically generated

Figure 11: ACC Multiple Regression Prediction

Based on the ACC data only. The recommended Coaches salary for DINO BABERS at Syracuse, based on his Win Percent of 43.5%, the school is Private, the school Graduation Rate of 85%, and a Bonus of $1587639.0, their salary will be $1804855 +/- 1558906.

A picture containing text, receipt

Description automatically generated

Figure 12: SEC Multiple Regression Prediction

If an opening in the SEC has significantly higher salaries, as we have seen throughout this analysis, his predicted salary with the same win percentage and parameters above his salary range would be from $3,233,348+/-$1,795,332.

This modeling and regression in a new programming language have been challenging, but something that was luckily just doable. Most of the time was spent on data preparation and scrubbing, getting the database into a form from the original four separate files down to a single file that could answer the established questions. This took many different steps, outlined in this report and the Jupyter Notebook containing the code. One should find notations about decisions and problems at the coding time. Many different limitations should be noted. First, the final database was less extensive and representative than possible for a multivariate analysis like this and any testing and training scenario. Many more schools might tell a different story than the one we found in the data. Aside from the sample size, there were problems with multicollinearity in our database when the regression models were run. This means that one should go back and analyze the variables until it is figured out which two or more are correlated enough to cause the problem. The variables could be eliminated down to one or combined in some way where both are still represented in the data set. Lastly, upon pondering the variables in the analysis, it was realized that all of the numeric and dollar-based variables come from one place: the school’s overall budget for the football program each year. That may be an exorbitant amount of money in many situations, but overall, it limits our variables and how much they can change. The ability to change represents the ability to have relationships and is the basis of the statistics we use to examine the data.