LAB 1

Salary of College Football Coaches

Alex Hyman

10/24/2018

Alex Hyman
10/18/2018
Lab 1

Table of Contents

TABLE OF CONTENTS **0**

INTRODUCTION **2**

OBTAIN **2**

SCRUB **4**

- Coaches Data Frame 4
- Record Data Frame 4
- Graduation Data Frame 5
- Rankings 6
- Cities Data Frame 6
- Merging Data Frames 6

EXPLORE **8**

MODEL **14**

- Pooled 14
- Non-Pooled Data 16
- Mixed Effects Model 17

INTERPRET **18**

QUESTIONS **18**

Alex Hyman
10/18/2018
Lab 1

Introduction

Fifty percent of Division I College Football coaches for the 2018-19 season are making a base salary of $2 million dollars or more. While this may not come as a surprise to the millions of spectators that watch college football every Saturday during the fall, students at these academic institutions may believe that the excessive spending on college football is a waste of their increasing tuition and fees. Nonetheless, college football is a potentially huge revenue source for these schools, and many see a successful team as an asset.

However, when a college athletics coach is the highest paid employee of the state, it is important that they are getting paid at a market value similar to other college coaches[1]. A model to estimate the salary of a college football coach can be created using public data and fitting the variables to a linear regression model of some type. The goal of this paper is to determine the best way to model the salary of a college football coach through the OSEMiN process, and to answer the following questions:

1. What is the recommended salary for the Syracuse football coach?
2. What would his salary be if we were still in the Big East? What if we went to the Big Ten?
3. What schools did we drop from our data, and why?
4. What effect does graduation rate have on the projected salary?
5. How good is our model?
6. What is the single biggest impact on salary size?

Obtain

The base data for this analysis was the Coaches8.csv file provided in this assignment. This data included 130 rows for the data objects, and had columns that included the School Name, Conference Name, Coach's Name, School Pay, Total Pay, Bonus, Bonus Paid, Assistant Pay, and Buyout. For this analysis we are interested in the salary of the coach, which will be calculated as the Total Pay subtracted by the Bonus Paid. This will be the target variable for our analysis. The conference of the school will be utilized in some of the models as either a categorical value or as a grouping for multi-level modeling.

Additional data was collected to act as predictor variables for this model. These datasets included variables related to the team's record, grades, stadium capacity, game attendance, and ranking as well as the coach's historical record. Additionally, geographic locations for each of the schools was obtained for some geographic analysis.

[1] College athletic coaches are the highest Paid state employee in 39 states. http://www.espn.com/espn/feature/story/_/id/22454170/highest-paid-state-employees-include-ncaa-coaches-nick-saban-john-calipari-dabo-swinney-bill-self-bob-huggins

Alex Hyman
10/18/2018
Lab 1

The team's record was obtained from the ESPN website and included variables of win-loss, if the team won the conference, and the generalized metric Football Power Index (FPI), which uses historical team performance to estimate the average point margin versus an average team on a neutral field. This website was saved as an html file on my local machine to allow for offline scraping of this data.

The AP ranking of a school was also scraped from a link on the ESPN website. However, because only 25 are ranked at a given time, this was the smallest dataset. The webpage was downloaded as an html file on my local machine to allow for offline scraping of the data.

The grades of the team were interpreted by via metrics provided by the NCAA's Academic Progress Rate Research Data. These metrics were the Graduation Success Rate and the Federal Graduation Rate. The Federal Graduation Rate is the percentage of college football players that begin at a given school as a Freshman and graduates within a six-year period. The Graduation Success Rate is the percentage of all college football players that participate at a university, regardless of whether they had transferred or began as an early-enrollee at the school. The webpage providing this data was downloaded as an html file on my local machine to allow for offline scraping.

The stadium capacity for a school was obtained from the website college gridirons[2], which included the school, stadium name, and stadium capacity. Stadium capacity data was scraped directly from the website. The football attendance data was obtained from a .pdf file provided by the NCAA, and it included the school, number of home games, average attendance, and total attendance. The school name and total attendance was manually copied from the pdf file and placed in a new csv file. Only total attendance was utilized in this analysis because I was more interested in the volume of tickets sold, as this metric would be more telling to the revenue brought into the school from spectators making purchases in and around the campus.

The historical coaching data was provided by the Statistics section on the NCAA website[3]. The search engine would provide the coach's career winning percentage, team winning percentage, and years spent at a given program as the head coach. The biggest downside of this data source was that it did not provide any information of coaches that had moved from a coordinator position to a head coach position. This resulted in a good bit of missing or incomplete data. The data provided by the coaching search engine was manually copied into a csv file that contained the coach's name, career winning percentage, school winning percentage, and years as a head coach at that school.

Finally, the geographic information of the schools was provided by a Wikipedia page[4] that had a table of the city and state for each Division I college football program. This web page was scraped for the table, that contained this geographic information.

---

[2] http://www.collegegridirons.com/comparisonscap.htm
[3] http://web1.ncaa.org/stats/StatsSrv/careersearch?searchSport=MFB
[4] https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs

Alex Hyman
10/18/2018
Lab 1

Scrub

While all of the data needed for this model has been obtained, it is not all in a useable format. Schools are referred to as different names (abbreviations, acronyms, etc.), numbers are formatted as strings, and some values are combined together in a single column (W-L). The scrubbing process for each of the datasets is described below.

Note: Because all the different data sources utilized different names for the schools (and there is no Primary Key), a dictionary was maintained to convert common names of the schools into the names used in the Coaches8.csv file.

Coaches Data Frame

The columns of interest in the Coaches8.csv file were the “School”, “Coach”, “Conference”, “TotalPay”, and “BonusPaid” columns. The school, coach, and conference columns were already formatted in an acceptable string format, as these are not continuous variables. The TotalPay and BonusPaid columns were formatted as a string with a “$” at the start of the string and commas separating every three digits of the number. Because we need to subtract the Bonus Paid from the Total Pay of the coach and use this as the response variable theses strings need to be converted into continuous numbers. Additionally, there were multiple columns that had “--” representing an unknown or missing value.

The TotalPay and BonusPaid columns were converted into integer values by mapping a function the removed all “$” and “,” from the cells and replaced all “-“ with a “0”. After the values were all numerical string of some sort, both columns were converted to int-type columns. The Bonus Paid was then subtracted from the Total Pay, and saved in a new column named “salary”. Finally, because there were some rows that had missing data for the salary, the data was indexed to have only rows with a salary column greater than 0. This resulted in removing SMU, BYU, Temple, Rice, and Baylor from the data because they did not have public salary data. The columns remaining after this scrubbing were “School”, “Conference”, “Coach”, and “salary”

Record Data Frame

The html file that contained the records for each team was read into a pandas data frame and contained columns of “Team”, “W-L”, “CONF WIN%”, and “FPI”. The Team column included the name of the school, along with the conference the school belonged to. The W-L column contained the overall record of the team, with the number of wins being separated by the number of losses by a “-“. The CONF WIN% column deviates between 0 and 100% during the season as an estimate of the probability a team wins their conference, but because this web page was updated after the season, teams that won their conference had a “100.0” as the CONF WIN% and the remaining teams had 0.0. Because pandas read in the table from a web page, all columns were formatted as a string. Additionally, because this was a large table online,

Alex Hyman
10/18/2018
Lab 1
the header was repeated multiple times within the data frame. These rows were filtered out by indexing all rows in the data frame that did not have a value of “W-L” in the W-L column.

To make this table usable for the modeling the salary of the head coach, the table was scrubbed to contain the team name, win percentage, FPI, and a binary conference winner column. The first step in making the Team column usable was applying a function that looked for the regular expression “, [A-Za-z0-9 -]+” and replaced it with an empty string and then stripped the team name of any punctuation. The team column was then renamed to “School” and an initial right merge was conducted onto the coaches data frame and the names of schools that were missing the salary column was indexed. The newly created data frame provided the nominal values for the schools in the Record data frame, and informed me of which schools I need to find in the coaches data frame. The values that were missing data were manually stored as a key in a python dictionary, and the corresponding nominal value in the coaches data frame was stored as the value. Finally a function was created to find the school name in the dictionary, and if it existed as a key, the new name replaced the name that existed in the data frame.

The win percentage was calculated in a few steps. First, the number of wins was extracted from the W-L column by splitting the string at the “-“ and storing 0-index of the list as an integer value in a column named “totalWins”. Next, a new function was created to split the W-L column into a list, convert the list into a list of integers, and return the sum of the list. This value was stored in a new column called “numGames”. The reason this was split into two different functions was to see if the number of wins or the number of games had any effect on the model. Finally the totalWins column was divided by the numGames column and stored in a column named “winPercent”.

A binary column was created for whether a team had won their conference by changing the value of all rows that had a value of “100.0” in the CONF WIN% column, and changing the value to an integer with the value of 1, the remaining rows were converted to an integer with value 0. This value was stored in a column named “confWin”.

Finally, the FPI of a school was converted from a string value to a float value. The columns that remained in the data frame for the analysis were “School”, “FPI”, “totalWins”, “numGames”, “confWin”, and “winPercent”.

Graduation Data Frame

The html file that contained the graduation data had an html table with the school, GSR, and FGR for the schools. To make this data usable in the analysis, all punctuation and instances of the words “at”, “of”, “University”, and punctuation were removed, and a merge on the school name, similar to the merge performed in the record scrubbing, was performed and new key names were added and mapped to the values of the school name in the coaches dictionary.

Alex Hyman
10/18/2018
Lab 1

Because the FGR and GSR are continuous data, both columns were converted to integer values, and the columns were assessed for missing values. The GSR did not have any missing values, but the FGR was missing some data. All rows that were missing the FGR were replaced with the median FGR as I felt this was the best estimate for the unknown values. The final columns for the graduation data frame were “School”, “GSR”, and “FGR”.

Rankings

The rankings html file contained an html table with the AP’s top 25 for the 2017-18 season. The two columns of interest were the school name and the ranking, even though the ranking would be removed later in the processing. The only processing that needed to occur was making the school names match the school names in the coaches data frame. Once again a function was mapped to the school column which removed all non-alphabetical characters and looked for any school names in the dictionary that needed to be converted to the nominal value for school in the coach’s data frame.

Cities Data Frame

The Wikipedia web page containing the city and state for all Division I schools contained an html table. The html table was converted into a pandas data frame and the columns School, City, and State were kept. Once again, all punctuation from the school name was stripped from the cell and the school name was looked up in a dictionary to see if the name of the school needed to be changed to the same value in the coaches data frame. A new column “cityState” was also created to combine the city name with the state name, separated by a comma.

For this to be a useful data frame, the latitude and longitude of the schools is required for any geographical analysis. The geopy package can accomplish this by passing a city, state pairing into a geocoder and a latitude, longitude tuple will be the output. The geocoder used for this conversion is the Open Street Maps *Nominatim*. A function using this geocoder will be applied to the cityState column and the output tuple will be stored in a new columns called “latlng”. Finally, the latitude and longitude for the schools will be extracted from this column and stored in a separate “lat” and “lng” column. The final columns in the cities data frame will be school, lat and lng.

Merging Data Frames

After all the processing of the original data sources, the following data frames have been formatted into a usable manner: Coaches, Record, Stadium, Attendance, Ranking, Cities, and Coaches Record. These data frames were then merged into one large data frame that contained all the information needed for exploring the data and modeling the variables that determine the salary of the coaches. The series of inner and left merges were conducted in a manner to retain the most information while ensuring we had complete cases of the data.

Alex Hyman
10/18/2018
Lab 1

First the Coaches data frame and the Record data Frame were merged via an inner join on the School column into a new data frame called data. The coaches data frame was missing the salary of five schools (SMU, BYU, Temple, Baylor, and Rice), and therefore were removed from the analysis. The Record data frame was missing the Liberty University row because they were competing as an FCS school until 2018, which could mean that their data is not representative of the other schools. An inner join between the Coaches and Records data frames would result in dropping Liberty and the five schools missing salary information from the final data set.

The next join conducted was an inner join between the previously joined data and the graduation data on the School column. The graduation data frame was missing the UNC Charlotte Football Program. This is likely due to the school being relatively new to College Football, as their first season began in 2013. Because the metrics of this dataset are based on graduation within six-years, the Charlotte program from the analysis.

An inner join on the School column between the previously joined data and the stadium capacity data was the next merge conducted. The stadium capacity was missing the schools Coastal Carolina and Alabama at Birmingham (UAB). This is most likely due to the data being collected before 2016. If this was the case, UAB had been disbanded in 2015 and Coastal Carolina was yet to become an FBS team. Both Coastal Carolina and UAB were excluded from the final analysis due to this merge.

An inner merge between the previously merged data and the attendance was then conducted on the School column, which resulted in no additional missing data; and a left join of the rankings onto the previously merged data followed. The left join ensured that schools that were previously in the combined data frame remained and had an NA value in the RK column, and the 25 schools that were in the ranking data frame had an integer value in the RK column. The RK values were changed for the unranked teams from NA to a 0, and the ranked teams had their ranking values changed to a 1. These operations resulted in the RK column being converted into a binary column indicating whether a school was ranked at the end of the football season.

The final two merges were inner merges of the cities and coaching record data frames on the combined data on the school column. Neither of these data frames were missing any schools, and therefore no more schools were excluded from the analysis. The final data frame for the analysis ended up missing the following schools:

- SMU – Private School, no salary data
- Rice – Private School, no salary data
- Baylor – Private School, no salary data
- BYU – Private School, no salary data
- Temple – No salary data
- Liberty – FCS Team in 2017-18, no FBS record
- Charlotte – New Program, No graduation data

Alex Hyman
10/18/2018
Lab 1

- UAB – Previously disbanded, No stadium data
- Coastal Carolina – FCS Team at time of data gathering, no stadium data

Explore

Salary is the response variable for the model using the data frame, therefore most of the EDA of this analysis will focus on the coach's salary; but first I wanted to get an idea of how the all the continuous variables are distributed. A tabular view of the of the count, mean, standard deviations, and the quartiles for each variable is provided below:

```
              salary         FPI     confWin  winPercent         GSR  \
count   1.210000e+02  121.000000  121.000000  121.000000  121.000000
mean    2.366055e+06    0.361983    0.090909    0.530254   75.710744
std     1.819090e+06   12.378074    0.288675    0.210925   10.017017
min     3.100000e+05  -27.200000    0.000000    0.000000   33.000000
25%     7.607500e+05   -8.100000    0.000000    0.416667   70.000000
50%     2.000000e+06    1.100000    0.000000    0.538462   76.000000
75%     3.508750e+06    8.900000    0.000000    0.692308   82.000000
max     7.807000e+06   28.400000    1.000000    1.000000   99.000000

              FGR       capacity     attendance  SeasonsAtSchool  \
count  121.000000     121.000000     121.000000       121.000000
mean    62.330579   52473.363636  271185.793388         4.322314
std     10.584098   23152.947043  179967.787882         4.399271
min     41.000000   20118.000000   49640.000000         0.000000
25%     56.000000   30964.000000  120650.000000         2.000000
50%     62.000000   50000.000000  223875.000000         3.000000
75%     68.000000   65236.000000  359660.000000         6.000000
max    100.000000  107601.000000  752464.000000        26.000000

       SchoolWinPerc  CareerWinPercent
count     106.000000        114.000000
mean        0.534198          0.570456
std         0.174147          0.155144
min         0.083000          0.083000
25%         0.414000          0.481750
50%         0.538000          0.592000
75%         0.670250          0.664250
max         0.901000          0.868000
```

The tabular data shows that there is a huge range in the response variable of salary, with the minimum salary being $310,000 and the maximum salary being $7,807,000 and median salary being $2,000,000. This would suggest that there is a bit of a positive skew, with a longer tail on the positive end.

FPI, however, appears to be a normally distributed variable distributed around 0. The table shows that the variable is normally distributed by the symmetry in the quartiles and the median being about 1. Other notable characteristics about the data discovered from the data is that

Alex Hyman
10/18/2018
Lab 1

GSR appears to be higher than FGR, at least in terms of measures of centrality; there is one school that had an attendance value less than the median stadium capacities of all schools; and most college football coaches have had a tenure of less than their the expected career of their first recruited class (expected length of recruits college career is 4 years, median tenure is 3 years).

A more interesting analysis would be into the relationship between variables, which can be accomplished with a pair plot. This plot will show a scatter plot between the two continuous variables. The top row will show the relationship of other variables with salary, and cells that show any sort of pattern will be of more interest. The pair plot is provided in Figure 1.

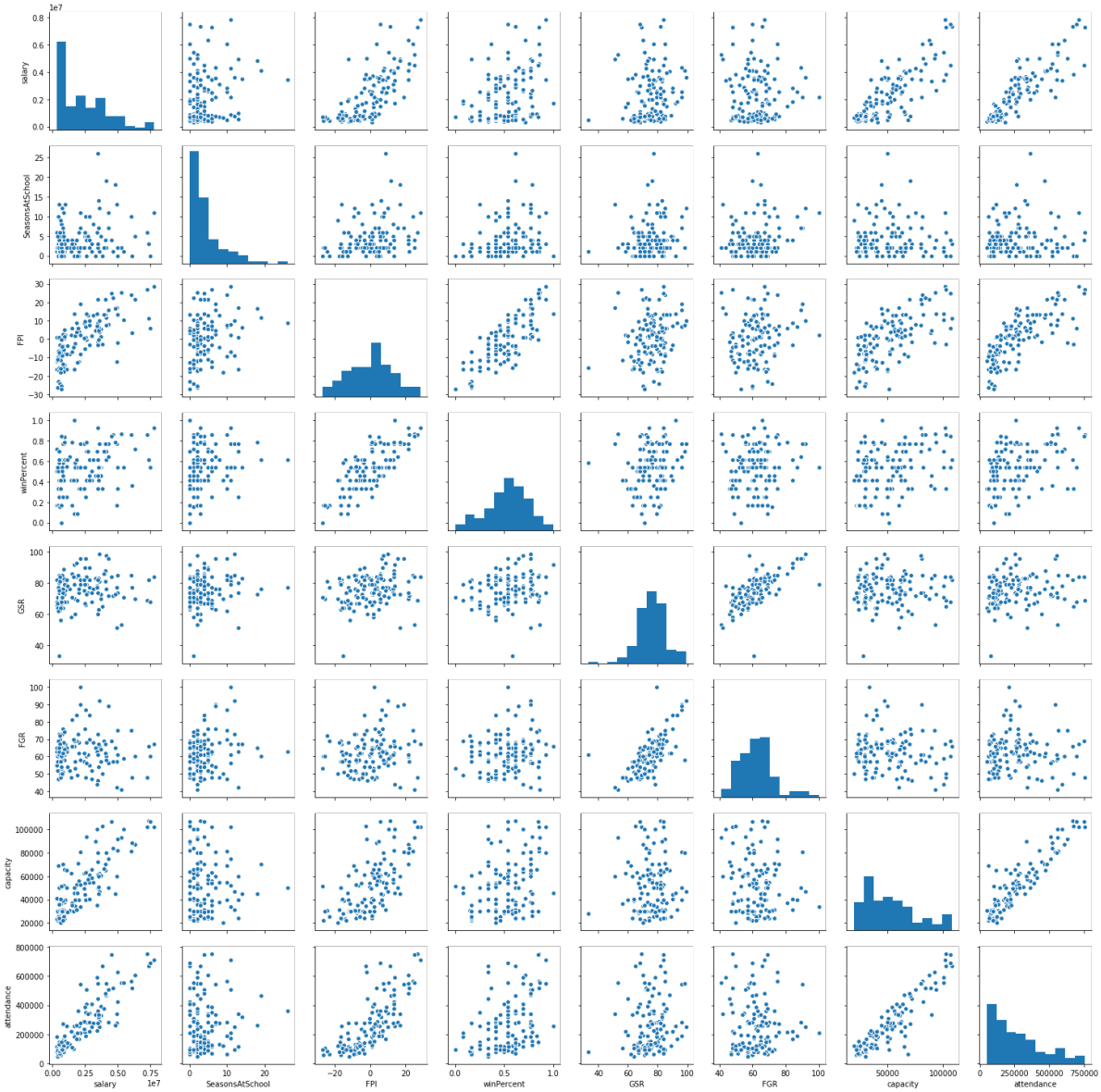


Figure 1 Pair Plot of Continuous Variables

Alex Hyman
10/18/2018
Lab 1

The pair plot shows that salary has some sort of relationship with both attendance, capacity, and FPI. There is also a minor positive relationship with win percentage. Salary's relationship with attendance and capacity appears to be a positive relationship, with salary increasing as both attendance and capacity increase. The relationship between salary and attendance appears to have a bit less variance than the relationship between salary and capacity. This is likely because the attendance figures are likely to be responsive to the success of the team. The FPI metric also appears to have a positive relationship with salary, but not as linear as the attendance and capacity relationships.

A way to get a numeric value representing the linear relationship between variables would be by a creating a correlation matrix. This can be accomplished with the correlation method on a the data frame of the variables. This correlation matrix is provided in Figure 2.

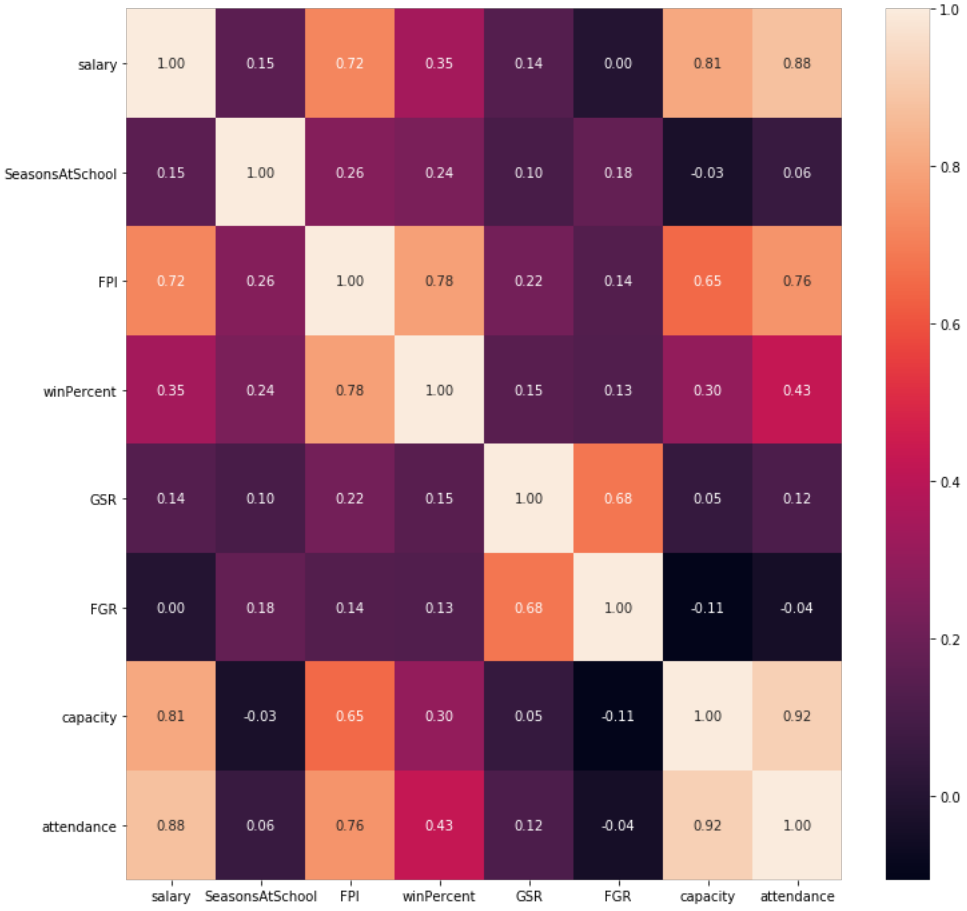


Figure 2 Heatmap of Correlations

As noted earlier, the capacity and attendance variables have the strongest linear relationship with salary, with attendance being a slightly stronger relationship than capacity. FPI is the next strongest linear relationship, followed by the team's winning percentage. It is interesting to

Alex Hyman
10/18/2018
Lab 1
note that the Federal Graduation Rate of the college has a correlation of zero with salary, and Graduation Success Rate barely appears to have any sort of correlation with salary.

A histogram of the coaches salary can be seen in the first diagonal of the pair plot. This histogram shows a heavy positive skew with a long tail on the positive side. There even appears to be an increase in the highest salary bin, indicating that there is a group of elite coaches that are most likely outliers. The boxplot of all the coaches in the analysis (Figure 3) shows that the Alabama is actually the only outlier with a salary more than 1.5 x IQR greater than the median of the salary.

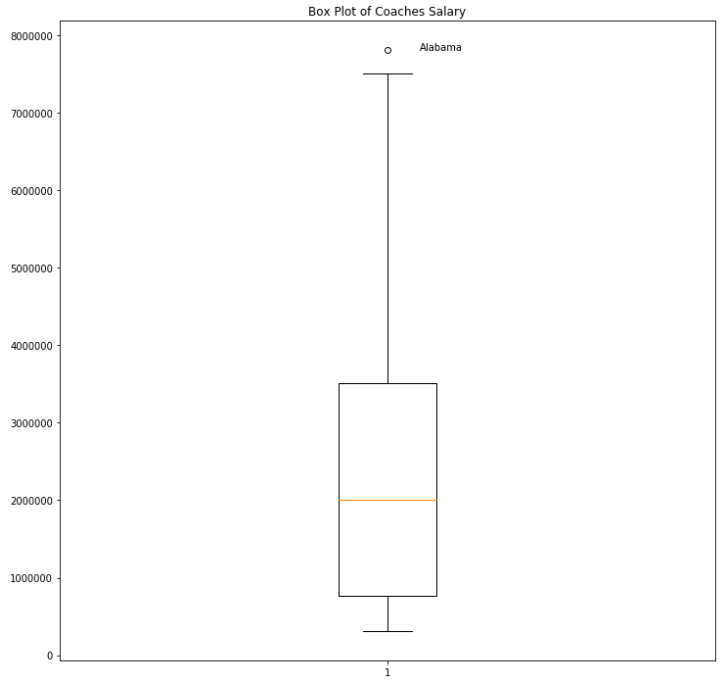
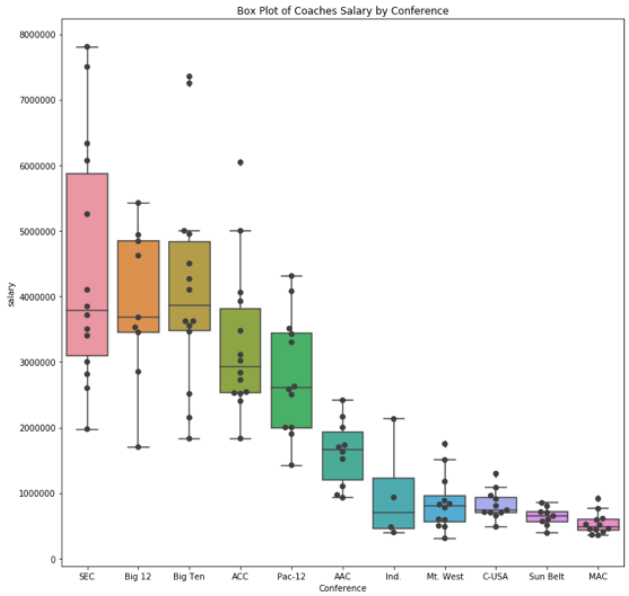


Figure 3 Boxplot of Head Coaches Salaries

The boxplot suggests that there is a large positive skew in the salary of coaches and that half of the college football coaches make a salary of $2,000,000 or more. A closer look at the distribution of coaches salary in the boxplot by conference shows that there is a discrepancy in salary distribution by conference. Within the context of the SEC, the salary of the Alabama coach is actually not an outlier. However, the SEC also has the greatest range compared to the other conferences. It is also important to note that there are only four schools that are Independent of other conferences. An aggregation by conference also shows that coaches in the SEC have the highest average salary, followed by the Big Ten and Big 12. A table with aggregations of average salary and counts by conference as well as a boxplot of salaries by conference are provided in Figure 4.

Alex Hyman
10/18/2018
Lab 1



Conference	Avg. Salary	Num. Schools
SEC	4.422027e+06	14
Big Ten	4.154371e+06	14
Big 12	3.892311e+06	9
ACC	3.285946e+06	14
Pac-12	2.804606e+06	12
AAC	1.615868e+06	10
Ind.	9.854498e+05	4
Mt. West	8.533545e+05	12
C-USA	8.216772e+05	11
Sun Belt	6.418333e+05	9
MAC	5.315078e+05	12

Figure 4 Boxplot and Average Salary By Conference

The most noteworthy item from these graphics and tables is that the salary does have an effect by conference; therefore, the conference of the team should be considered when modeling the salaries of coaches.

Additionally (as a bonus), the salaries of college football coaches were analyzed by geography to see where the college football coaches that were paid the most money were likely to be located. A base 10 log of the salary was taken of the coaches salary to make the distribution look a bit more normally distributed. The graphic is provided in Figure 5.

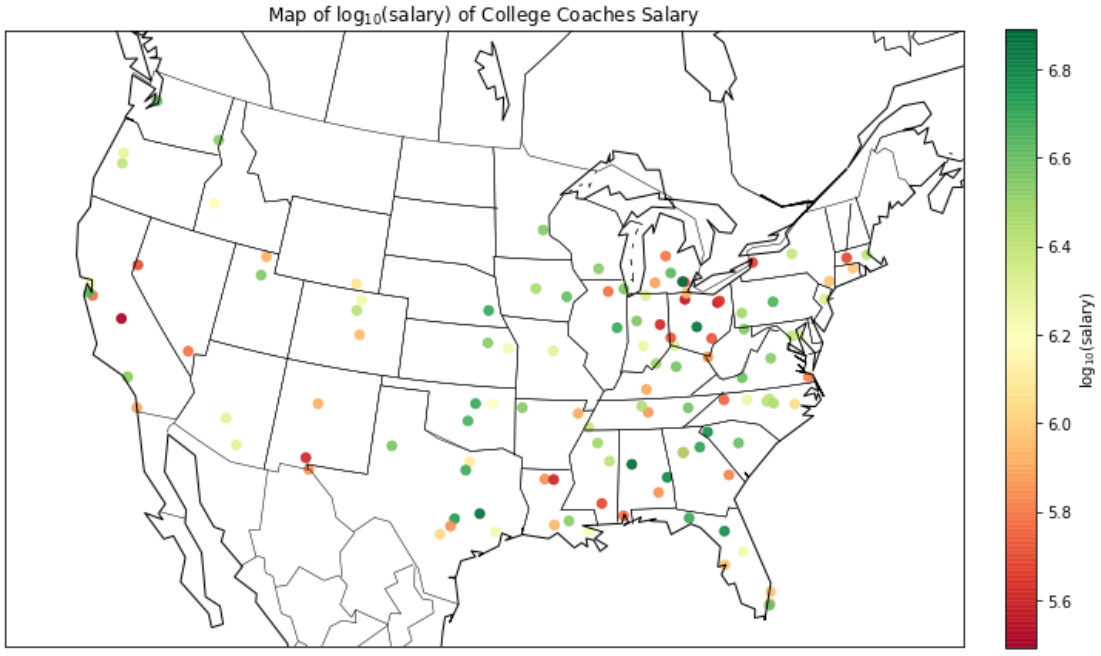


Figure 5 Map of College Football Coaches

Alex Hyman
10/18/2018
Lab 1

The graphic shows that a majority of college football teams are east of the Mississippi and there is a high concentration in the southeast, specifically in Alabama, Georgia, Florida, and South Carolina. There is also a few highly paid college coaches in the lower Midwest, specifically Texas, Oklahoma, and Nebraska. There is also a high concentration of college football teams in the upper Midwest, specifically in Michigan, Ohio, and Indiana, but a high percentage of these jobs are on the lower end of the pay scale. There is also a noticeable lack of schools out west and in the Northeast.

A more interesting graphic would generalize the salaries and locations of the schools by conference. The Figure 6 is a map of the conferences estimated by average latitude and longitude by school. The size of the point on the graphic represents the median salary of coaches in the conference.

The most noteworthy item from this analysis is that there appears to be five conferences (SEC, ACC, Big 12, Big 10, and Pac-12) that seem to operate at a higher salary than the other six conferences. It is also interesting that 8 of the 11 conferences are on average located east of the Mississippi and the conferences are clustered around the Southeast and upper Midwest (Ohio and Indiana).

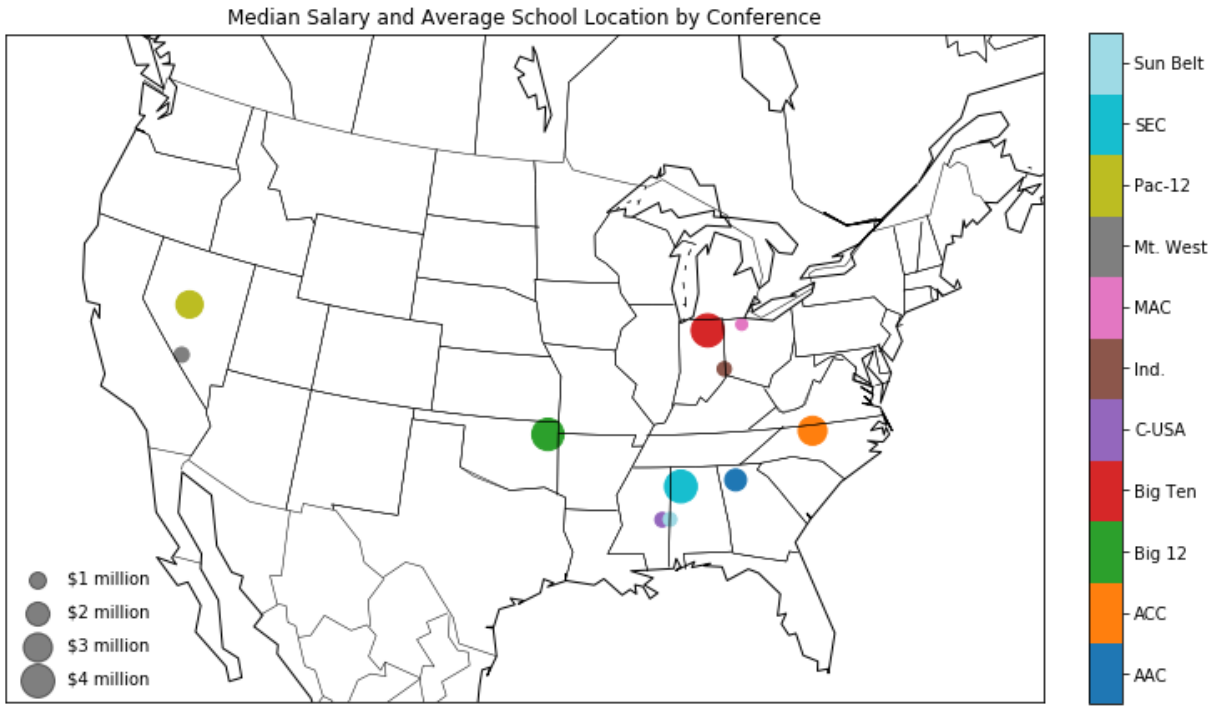


Figure 6 Average Location and Median Salary of Coaches Salary

Alex Hyman
10/18/2018
Lab 1

Model

The coaches data set was split into two separate datasets: a training set and a test set. The training set was used to fit the linear model and the test set was used to evaluate the model. The training set was set to be about 2/3 of the remaining data, and the test set was the remaining 1/3 of the dataset. The metric used to evaluate the model's performance on the test set was Mean Absolute Error (MAE). MAE was chosen as the evaluation metric in the performance of the model on the test set because this analysis will focus more on the median prediction of salary from the model than worry about outliers.

Three different models will be evaluated for their performance: Pooled, Non-Pooled, and Mixed Effects. The pooled model will not consider the school's conference in the model, the non-pooled model will use conference as a factor, and the mixed effects model will consider how the variance between the conferences affect the variance of the predictor variables. Each of these models will be evaluated once using the predictor variables of Ranking, Attendance, GSR, Conference Champions, team win percentage, FPI, and seasons the coach has spent at the school; the model will then be pruned for variables that are significant in the initial model, and evaluated again with just the remaining variables. GSR was chosen as the graduation attribute to model because it had a higher linear correlation coefficient with salary than FGR had.

Additionally, after the models have been fit and evaluated in initial dataset, the data will be indexed to only include coaches that are not considered outlier in the overall dataset, or outliers within their conference. Conferences that had fewer than five teams remaining were also excluded in the modeling due to the small sample size. Each of the models will then be fit and evaluated with the newly selected datasets. These models should provide a better understanding of the variance in the coach's salary at an average school. The performance of each of the models evaluated is provided in Table 1.

Table 1 Mean Absolute Error of Models Evaluated

	Pooled	**Non-Pooled**	**Mixed-Effects**
All Variables, All Data	$605,432.80	$694,926.39	$605,114.28
Sig Variables, All Data	$561,548.42	$673,599.41	$570,189.55
All Variables, No Outliers	$587,179.54	$691,935.14	$580,544.56
Sig Variables, No Outliers	$563,793.82	$666,538.93	$556,173.51

Pooled

The initial pooled data with all variables and outliers was evaluated to have an MAE of $605,432.80. This regression found that the variables win percentage, FPI, and attendance were

significant. An interesting find from this analysis was that win percentage had a negative coefficient. This would be counterintuitive, as common sense would expect the higher a winning percentage, the higher a coach's salary would be. The adjusted R-squared for this model was 0.785.

The pooled model with only win percentage, attendance, and FPI was evaluated to have an MAE of $561,548.42. Once again, win percentage was determined to be a negative coefficient and all variables had a p-value less than 0.05. The Adjusted R-squared for the model was 0.791, which means that nearly 80% of the variance in coach's salary is captured by the coefficients of FPI, attendance, and win percentage. A graph showing estimated salary vs. the actual salary of a coach is provided in Figure 7. The red line running diagonal through the figure represents what the model would look like if each school was predicted correctly. Dots that appear above the red line are coaches that are getting paid less than the model estimated, and dots below the red line are coaches getting paid more than the model estimated. This model was determined to be the best predictor of a coach's salary in a Pooled Model.

After removing the outliers from the dataset, the performance of the model improved to have a MAE of $587,179.54. Again, win percentage, attendance, and FPI were the variables the model determined to be significant, and winning percentage was determined to have a negative coefficient. The model using only significant variables improved upon the model using all variable to have an MAE of $563,793.82; however, the MAE for this model was not an improvement upon the model that had outliers in the dataset as well. Both models used the data with the outliers removed improved upon the adjusted R-squared (0.804 and 0.812), however, this model was most likely improved due reduced variation in the data, not necessarily because of the better model.

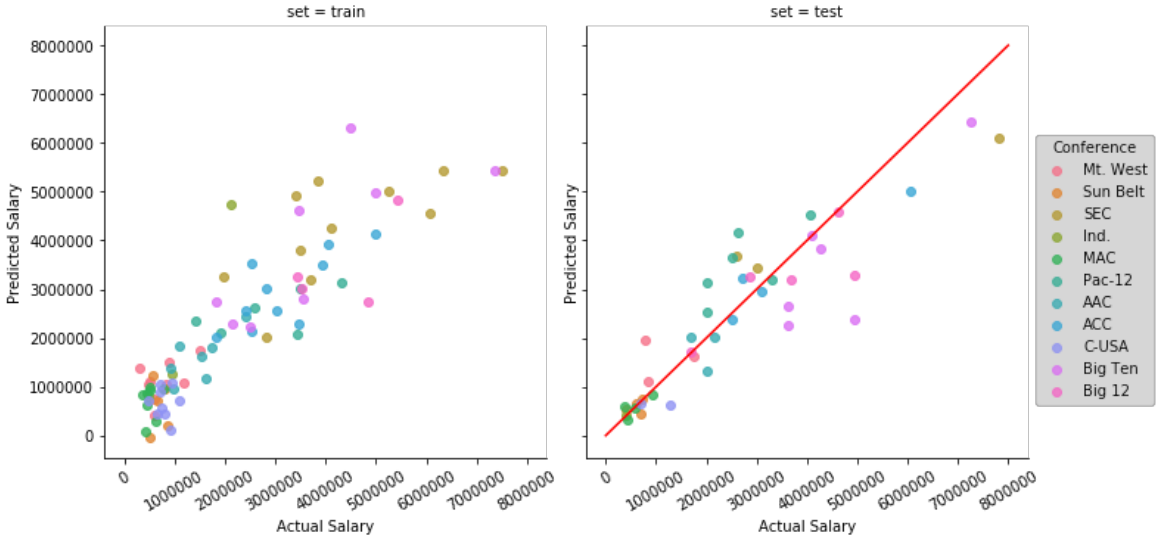


Figure 7 Pooled Model Actual Salary vs. Predicted Salary

Alex Hyman
10/18/2018
Lab 1

Non-Pooled Data

The Non-Pooled model included the conference as a factor, which effectively altered the y-intercept for each of the conferences while maintaining the same slope for all other variables. The initial model created using all the schools (including outliers) was evaluated to have a MAE of $694,926.39 on the test set. While this model performed worse on the test set than the Pooled model, the adjusted R-squared was calculated to be 0.810. This metric means that the model accounted for more variation in the training set than the Pooled model, but because of the significantly reduced performance in MAE, it is likely that the Non-Pooled model was overfit to the training set and would not perform well as an estimator.

The non-pooled model with only FPI, attendance, and win percentage was evaluated to have an MAE of $673,599.41, and an adjusted R-squared of 0.820. This was a slight improvement upon the Non-Pooled model with all the variables, but the model was once again likely overfit to the training data. The sample sizes of the conferences likely are too small to find a proper adjustment based on conferences alone.

When the outliers were removed from the analysis, the models once again improved slightly. The Non-Pooled model using all variables found the MAE to be $691,935.14 on the test set, and had an adjusted R-squared of 0.824. The best estimator of the Non-Pooled models was determined to be the one with the outlier removed and only significant variables used. The MAE on the test set was determined to be $666,538.93 and the adjusted R-squared on the training set was found to be 0.835. A plot of the estimated and actual salary of the coaches is provided in Figure 8. Points that fall along the red line are coaches that have an actual salary equal to the predicted salary. Ponts above the red line are considered to be underpaid coaches according to the model, and points below the line are coaches that are considered overpaid.

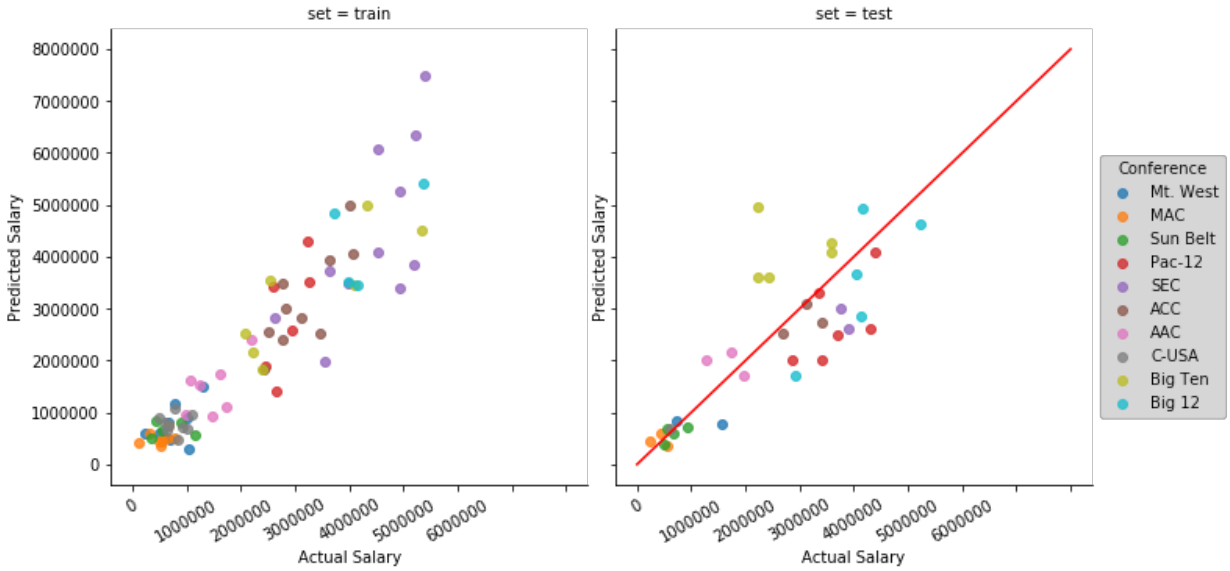


Figure 8 Actual Salary vs. Predicted Salary, Non-Pooled

Alex Hyman
10/18/2018
Lab 1

Mixed Effects Model

The last model evaluated was the Mixed Effects Model, which takes in account the random effects of the conference on the other variables predicting salary. The initial evaluation of the Mixed Effects model resulted in a MAE on the test set of $605,114.28. This was the best performing model that included all of the variables. Once again, team win percentage, FPI, and attendance were determined to be the significant variables and the win percentage was a negative coefficient.

Using only the significant variables found from the initial model, a new Mixed Effects model was created with only win percentage, attendance, and FPI. This model was evaluated to have a MAE on the test set of $570,189.55. This model's performance was not as good as the pooled model, but performed significantly better than any of the Non-Pooled models.

After removing the outliers, the Mixed Effects models were fit and evaluated again. The model that included all of the variables was evaluated to have an MAE of $580,544.56, and the Mixed effects Model with only the significant variables had an MAE of $556,173.51. The Mixed Effects model using only win percentage, FPI, and attendance turned out to be the best performing model out of all models evaluated. A graph of the actual salary vs. the predicted salary is provided in Figure 9. The red line in the plot would be a perfect prediction of salary, points above the red line are coached being paid less than expected and coaches below the line are being paid more than expected.

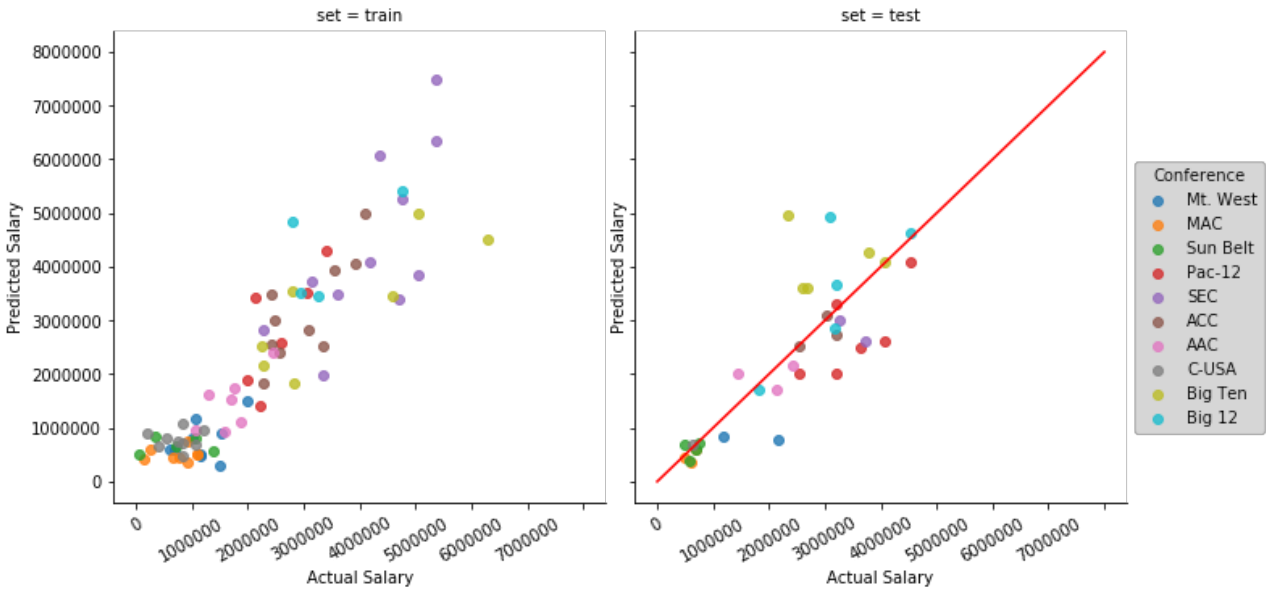


Figure 9 Actual Salary vs. Predicted Salary, Mixed Effects

Alex Hyman
10/18/2018
Lab 1

Interpret

The models that have been created can provide some insight into how much money a college football coach should be paid. Firstly, the conference a football team is in does have an effect on the salary of a college football coach, but the differences between and within the conferences are too vast to generalize solely with a changing y-intercept.

Attendance and FPI are also the two most significant positive coefficients in the models. The difference between the two variables is that attendance can only be greater than zero, which means every additional person that attends a can only increase the estimated salary of a coach, while the domain of FPI is negative infinity to positive infinity. Only coaches that have a positive FPI will increase their salary while coaches with a negative FPI will lose money, according to the models.

The win percentage variable tells a different story. The domain of the win percentage variable is from 0 to 1, and according to the model, a coach that wins 100% of their games will lose between $800,000 and $2,000,000 depending on the model. This feature of the model is likely meant to bring back outlier in attendance towards the mean. This likely has nothing to do with causation.

FGR and GSR were also found to be insignificant in the models when determining a coach's salary. This would suggest that the academic performance of a college football coach's team has little to do with how much they are valued.

Finally, if we wanted to predict a college football coach's salary, the Mixed Effects model with the outliers removed would be the best model, provided that the college needing a salary recommendation is not extraordinary in terms of football. This is especially the case in estimating the salary of the coach at Syracuse University, as the school would not be considered an outlier. The pooled model would also appear to be a decent model for estimating the salary of coaches. This would signify that it is potentially possible to capture the variance caused by the conferences by the predictor variables we have chosen.

Questions

1. **What is the recommended salary for the Syracuse football coach?**

The Mixed Effects level model with the outliers removed was used to recommend a salary for the Syracuse coach. With the variables associated with the current Syracuse head coach, the model recommends a salary of $2,587,532.36. This is compared to the actual salary of the head coach which is $2,401,206.00.

Alex Hyman
10/18/2018
Lab 1

2. What would his salary be if we were still in the Big East? What if we went to the Big Ten?

If we wanted to use see how a change in conference alters the salary prediction, a pooled model would be required. The pooled model does not consider the conference, and the mixed effects model only considers it when trying to limit the variation within the conferences.

It would not be possible to predict the salary of the Syracuse University head coach in the scenario we are in the Big East; this is because there are not any schools still competing as a member of the Big East, which means there is no data we could base our prediction on. However, if Syracuse were to have joined the Big Ten instead of the ACC, the Non-Pooled Model predicted that the head coach would have a salary of $2,240,892.76. This is compared to the Non-Pooled salary prediction of Syracuse in the ACC of $2,778,111.11. This is expected as the coefficient for the ACC was greater than the coefficient for the Big Ten.

3. What schools did we drop from our data, and why?

SMU, Rice, Baylor, BYU, Temple, Liberty, Charlotte, UAB, and Coastal Carolina were dropped from the analysis all together. SMU, Rice, Baylor, BYU, and Temple did not have any salary data listed. SMU, Rice, Baylor, and BYU are private schools and therefore exempt from being required to report salaries of employees; however, Temple is a Public school, but did not have any salary data listed for their coach.

Liberty was excluded from the analysis because they did not compete as an FBS school in the 2017-18 football season, and would have been the only non-FBS school in the analysis. Because they were not in the same league in the year the analysis is being conducted, they were determined to be an outlier for this analysis.

Charlotte was excluded from the analysis because there was not any graduation data. The likely reason graduation data was missing for Charlotte is because the graduation rate is based on a six-year timeline, and Charlotte did not have a program until 2013.

UAB and Coastal Carolina were excluded from the analysis because there was not any stadium capacity data. UAB was likely excluded from the stadium capacity data because their program had been disbanded in the mid-2010s, which is likely when this data was collected. Coastal Carolina was likely excluded from the data because 2017 was their first year as an FBS football team.

The data was also analyzed a second time through in which outliers in the from all schools, and outlier within the conference were removed. The school that was removed because they were an outlier of the population was Alabama. The schools that were removed due to being outliers within their conference were Clemson, Colorado State, Michigan, North Texas, Ohio State, and Toledo.

Alex Hyman
10/18/2018
Lab 1

4. What effect does graduation rate have on the projected salary?

It was determined that graduation rate did not have any real effect on the projected salary. In each of the models that included either GSR, the p-value was always greater than 0.6, and the FGR confidence intervals crossed zero in every model. Additionally, the FGR was found to have 0.00 correlation with the salary and GSR was found to have 0.14 correlation. This lack of correlation indicates that there is not a linear relationship between any graduation metric and a college coach's salary.

5. How good is our model?

When evaluated on a withheld test set, the best model (Mixed Effects, outlier removed) had an mean average error of $556,173.51. This means that when a prediction is made on a coach's salary, the average error $556,173.51 either greater or less than the actual salary. I would not consider this to be a great model, as the median salary is only $2 million. If the mean average error is more than $500k, that would mean the prediction is off by more than a quarter of the median salary.

6. What is the single biggest impact on salary size?

Each of the models determined that attendance, FPI, and team win percentage were the only significant predictive variables when linear modeling a head coach's salary. Using the Mixed Effects model, it was determined that attendance had a coefficient of 6.22, FPI had a coefficient of 63,426 and win percentage had a coefficient of -2,121,961.

To determine the overall impact of the significant variables of the model, a new Mixed Effects Model was created using the same training set previously used, but the variables of attendance, FPI, and win percentage were standardized to have an average of 0 and a standard deviation of 1. This allowed the variables to be within the same domain and comparable.
Because the variables were comparable in terms of scale, the absolute value of the coefficients were now comparable in terms of impact on the model. The new model determined that the scaled attendance variable had the largest absolute impact with a coefficient of 1,094,483.54; FPI had the second largest absolute impact on the model with a coefficient of 709,630.83; and win percentage had the smallest absolute impact on the model with a coefficient of -407,085.44.