IST718 Big Data Analytics - Lab 1

Martin Alonso - 2019-02-02

Introduction

The objective of this project is to estimate how much should the head coach of the Syracuse football program be paid. This project will use data collected through the NCAA website, Wikipedia, and other websites, to gauge how much NCAA FBS Division I coaches earn, their standings and team statistics for the 2017-2018 college football season, their respective schools student-athletes graduation rates, and their ranking among the top 25 football schools. These data will also allow to answer the following questions:

- What would the Syracuse head coach's salary be if he were in the Big East? If he were in the Big Ten?
- What schools did we drop from the data and why?
- What is the effect of graduation rate (GSR and FGR) on coach salary?
- How good is the model?
- What is the biggest impact on salary?

Analysis

The starting point of this analysis is a data set containing 129 observations, one for each school in the NCAA FBS program, with nine variables. These variables include School Conference, School Name, Coach Name, Salary, Total Salary, Bonus, Bonus Paid, Assistant Pay, and Buyout. The objective for this exercise is to build a model capable of predicting Salary for each coach, especially the Syracuse coach.

In addition to this initial data set, four more data sets were obtained with data pertaining to each school's football statistics and season standings, their graduation rate, and their rank within the NCAA FBS. These four data sets were imported into Python using the Pandas package, and merged using a combination of the Pandas and Fuzzywuzzy packages.

After merging, the data frame was checked for missing information and errors. Four schools had no information regarding their coach's salary, data that was completed by using the website Coaches Hot Seat. Additionally, two schools, Liberty and Coastal Carolina, were missing stadium capacity; the data was obtained from each school's Wikipedia page.

For schools in the Independent Conference, no home games were registered during the 2017-2018 season, for which their home record was listed as a missing value; these were subsequently changed to .000. Finally, there were six schools that did not feature a FGR number. Using Pandas groupby and fillna functions, these values were replaced with the mean graduation rate of each school's conference.

Having cleaned the data, the final data set featured 129 observations spread across 25 columns.

However, the current data was used to create an additional 12 columns that calculated the following variables for each school:

1. Away Wins: Games won as visiting team; calculated by subtracting Home Wins from Conference Wins.
2. Away Losses: Games won as losing team; calculated by subtracting Home Losses from Conference Losses.
3. Home Record: Home Wins divided by number of games played at home.
4. Away Record: Away Wins divided by number of games played away.
5. Conference Record: Wins divided by total number of games played.
6. Rank difference: Difference in rank between 2016-2017 season and 2017-2018 season.
7. Average Points Scored: Average Points scored per Conference.
8. Average Points Allowed: Average Points allowed per Conference.
9. Offensive Rating: Points scored per team minus Average Points Allowed divided by Average Points Scored.
10. Defensive Rating: Points allowed per team minus Average Points Scored divided by Average Points Allowed.
11. Points Per Game: Points scored per number of games played.
12. Points Allowed Per Game: Points allowed per number of games played.

Having created these columns, the objective is to identify what their relationship is to the salary each coach is paid.

The first question that should be answered is how are the salaries distributed. We can see in Fig. 1 that coach salaries are not normally distributed, with the majority of coaches earning a salary below US$2 million. Salaries like Alabama A&M's Nick Saban are more the exception than the rule. However, despite this skewness, there may be a case that salaries depend more on conference than anything else.

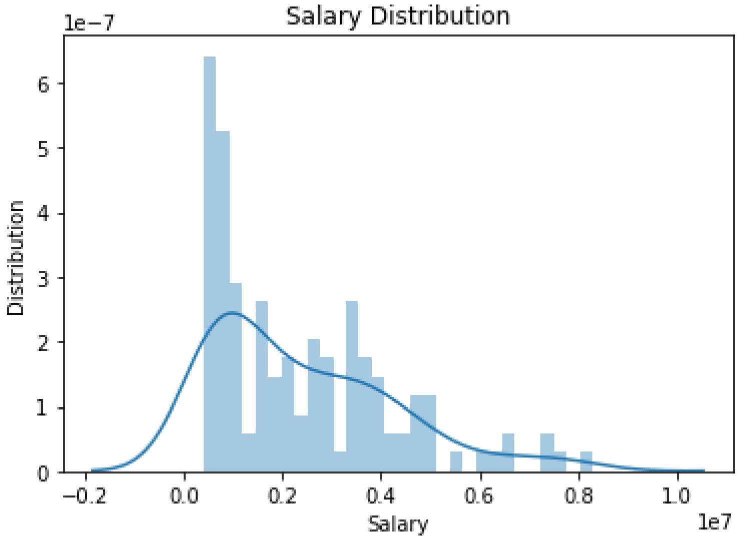

Fig. 1

Fig.2 proves that this is true: coach salaries are highly influenced by the conference in which the team plays in, with the SEC being notorious for having the highest spread between coach salaries, while the MAC and Sun Belt conferences feature the lowest salaries.

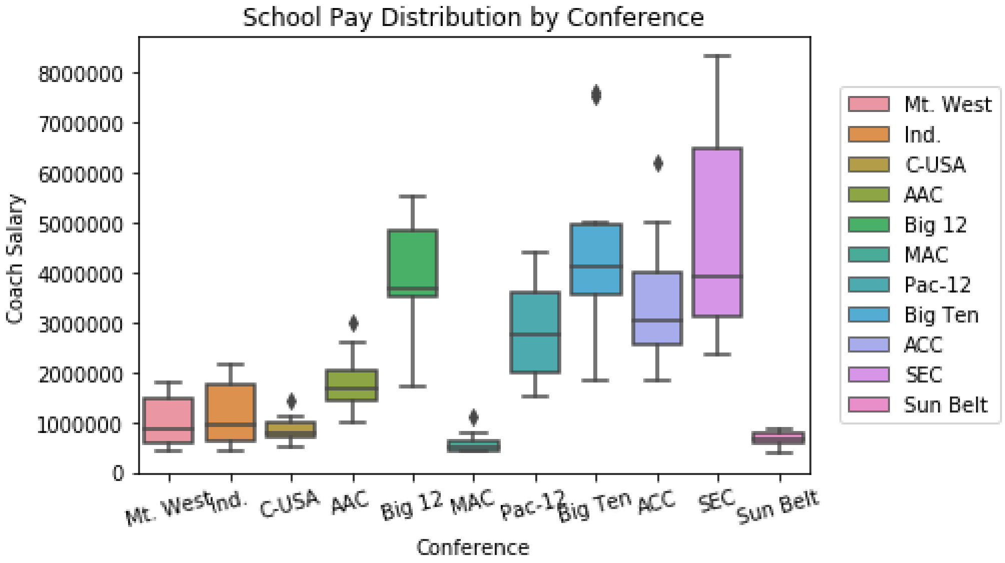

Fig.2

Playing in the AAC, Syracuse’s head coach should expect to be paid very close to the median salary for all coaches.

Nevertheless, salary does not fully depend on the conference in which a school plays. Coaches are expected have their teams perform consistently, with winning coaches being offered higher salaries, while coaches that constantly lose are either paid poorly or churned. Figures 3 and 4 show a clear relationship that salary has to a team’s number of wins and their winning percentage.

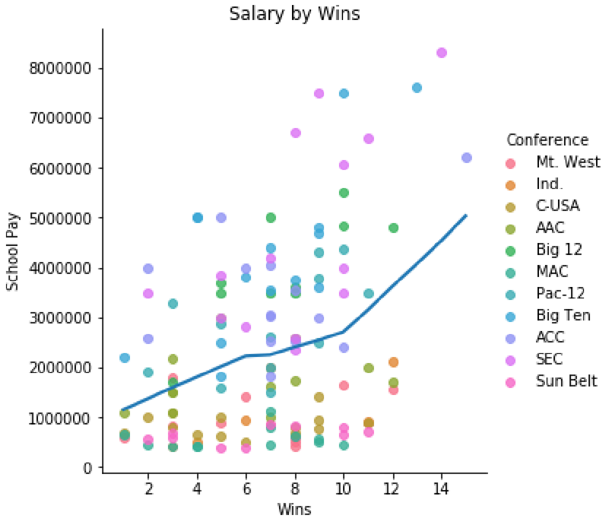

Fig. 3

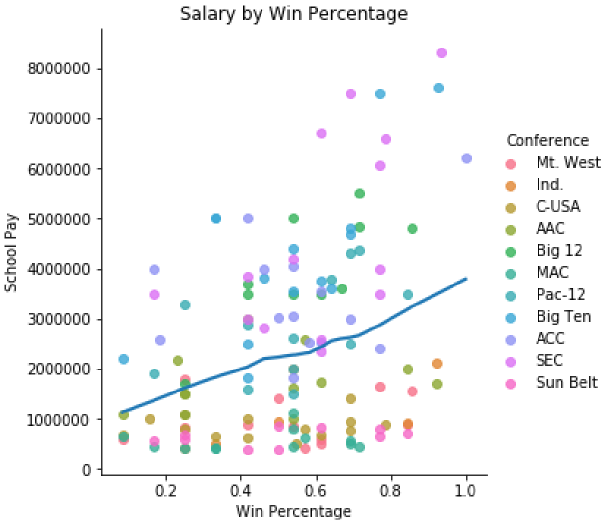

Fig. 4

Nevertheless, despite salary being results driven, their is a preference that these results are offense related rather than defense related. Teams with higher points per game and higher offensive ratings (Figure 5) tend to have coaches with higher salaries, while teams with higher points allowed per game and higher defensive ratings (Figure 6) tend to have lower salaries. Though this may seem counterintuitive, it makes sense as teams that allow too many points to be scored are teams that usually lose and, therefore, are teams which won't reward coaches as handsomely as teams that constantly win.

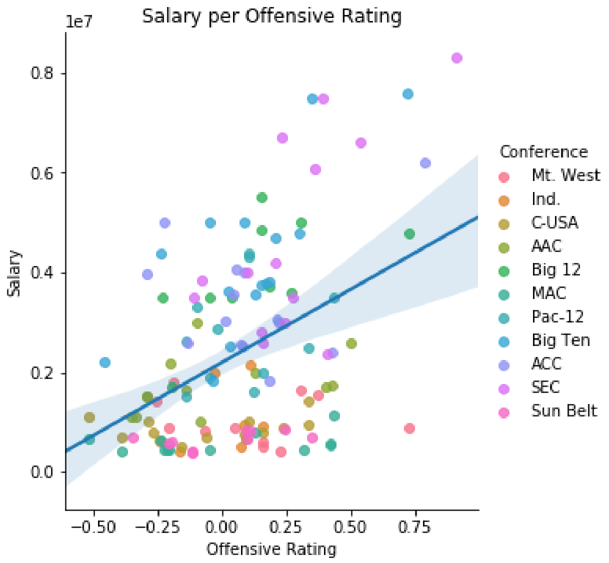

Fig. 5

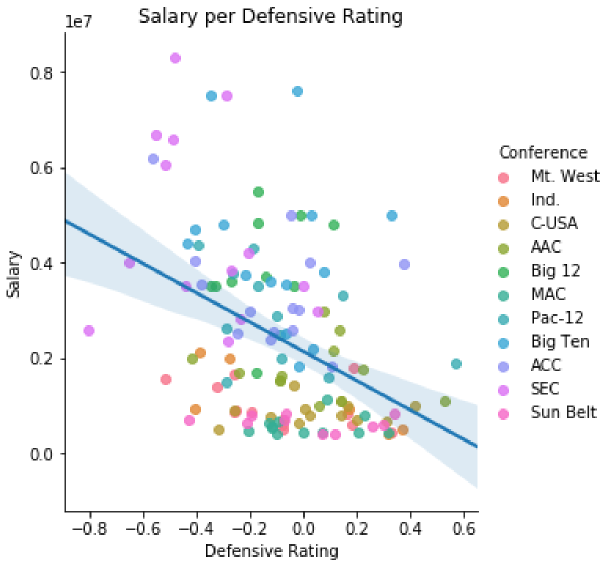

Fig. 6

Where there is no clear relationship is regarding the effect of both GSR (Figure 7) and FGR (Figure 8) on a coach's salary. However, a coach's main objective is to win football games with the team he is provided, not to ensure that his student-athletes graduate. That responsibility is in hands of the educators of the institution and depends on the student-athlete's priorities. From these two graphs, we can conclude that both GSR and FGR have no impact on a coach's salary.

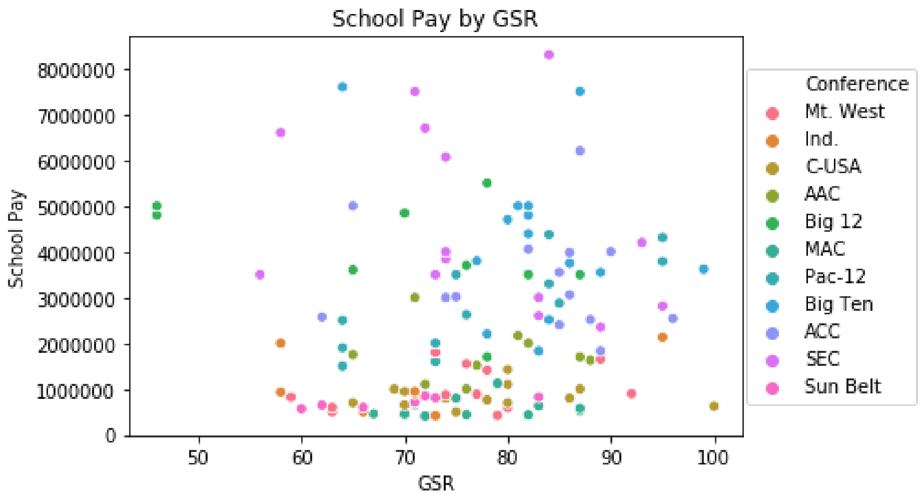

Fig. 7

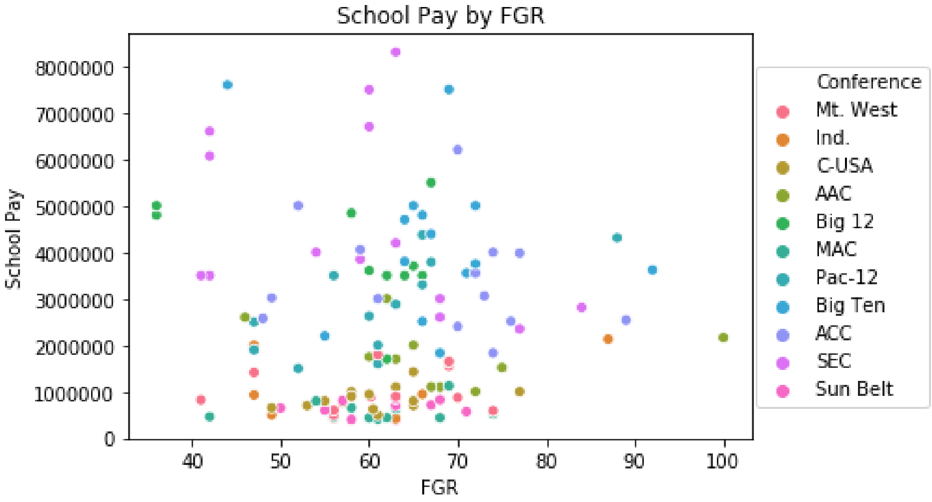

Fig. 8

Out of all the variables that are featured in the data set, the one that most closely correlates positively with a coach's salary is stadium capacity (Figure 9). The larger a stadium, the higher the salary. This can be due to the fact that stadiums with more seating capacity generate more revenue for the team, allowing the school to use part of the profits from tickets, and sales from other merchandise, to pay for the coach's salary.

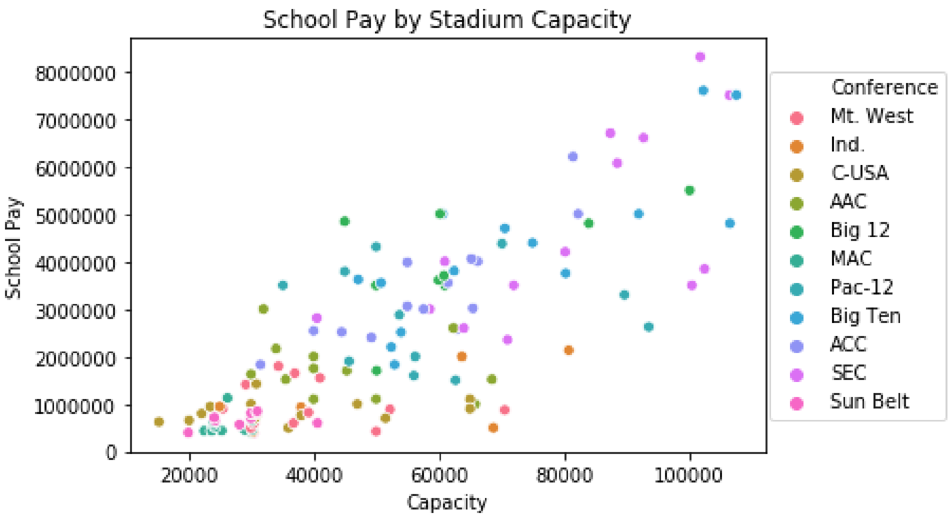

Fig. 9

Yet, this relationship may be a product of Simpson's paradox, as the relationship is clear when looking at the data set as a whole, but disappears, or diminishes, when looking at the relationship for each conference.

Modeling

Using Pythons sklearn and statsmodel.api packages, three models were built to test the data, the aim being predicting a coach's salary. In order to do this, the data were divided into three groups:

- Training set: Two-thirds of the data set were used for the training set.
- Test set: One-third of the data set was used for the test set.
- Validation set: Single observation featuring the Syracuse coach. This observation was removed from the training and test sets.

For the first model, the following input variables were used: Conference, Conference Wins, Conference Losses, Points Scored, Points Allowed, GSR, FGR, Stadium Capacity, Away Record, Points Per Game, Points Allowed per Game, Offensive Rating, and Defensive Rating.

```
OLS Regression Results
==============================================================================
Dep. Variable:          SchoolPay           R-squared:          0.916
Model:                  OLS                 Adj. R-squared:      0.901
Method:                 Least Squares       F-statistic:        60.88
Date:                   Wed, 30 Jan 2019    Prob (F-statistic): 8.95e-34
Time:                   16:10:24            Log-Likelihood:     -1304.2
No. Observations:       86                  AIC:                2634.
Df Residuals:            73                 BIC:                2666.
Df Model:               13
Covariance Type:        nonrobust
==============================================================================
                    coef           std err           t      P>|t|          [0.025          0.975]
-----------------------------------------------------------------------------------------------------
Conference     -1.697e+05       4.43e+04      -3.833    0.000       -2.58e+05      -8.15e+04
ConfW          1.519e+05        2.19e+05      0.694     0.490       -2.84e+05      5.88e+05
ConfL          2.177e+05        2.17e+05      1.002     0.319       -2.15e+05      6.51e+05
PF             3453.8087        1.29e+04      0.267     0.790       -2.23e+04      2.92e+04
PA             -9568.9121       1.39e+04      -0.688    0.494       -3.73e+04      1.82e+04
GSR            1823.9657        1.68e+04      0.109     0.914       -3.17e+04      3.53e+04
FGR            5956.8502        1.71e+04      0.349     0.728       -2.8e+04       4e+04
Capacity       35.8789          6.414         5.594     0.000       23.095         48.663
AwayRecord     4.591e+04        6.38e+05      0.072     0.943       -1.23e+06      1.32e+06
PPG             -3.22e+05       1.41e+05      -2.282    0.025       -6.03e+05      -4.07e+04
PAG            3.201e+05        1.71e+05      1.874     0.065       -2.02e+04      6.6e+05
OffRat         9.52e+06         2.33e+06      4.086     0.000       4.88e+06       1.42e+07
DefRat         -5.931e+06       1.67e+06      -3.545    0.001       -9.27e+06      -2.6e+06
==============================================================================
Omnibus:                 0.162          Durbin-Watson:          1.739
Prob(Omnibus):            0.922         Jarque-Bera (JB):       0.028
Skew:                    -0.044         Prob(JB):               0.986
Kurtosis:                3.010          Cond. No.               1.35e+06
==============================================================================
```

This first model is very good at predicting the total salary for a coach, having an adjusted R-squared of 0.901. However, GSR, FGR, ConfW, ConfL, PF, PA, and AwayRecord aren't significant predictors for determining a coach's salary. As stated previously, a coach’s job description doesn’t include a clause requiring high student-athlete graduation rates, which means that both GSR and FSR should not impact a coach’s salary.

ConfW, ConfL, PF, PA, and AwayRecord, on the other hand, are more confusing as any person would think that these are results that a coach can be held responsible for and, as such, should be reviewed as a part of their salary negotiations.

With this input, a second model was built, only removing GSR, FGR, which weren’t significant; but the other non-significant variables were retained to gauge their impact.

```
OLS Regression Results
==============================================================================
Dep. Variable:           SchoolPay              R-squared:              0.915
Model:                   OLS                    Adj. R-squared:          0.903
Method:                  Least Squares          F-statistic:            73.40
Date:                    Wed, 30 Jan 2019       Prob (F-statistic):     1.53e-35
Time:                    16:11:28               Log-Likelihood:         -1304.5
No. Observations:        86                     AIC:                    2631.
Df Residuals:             75                    BIC:                    2658.
Df Model:                11
Covariance Type:         nonrobust
===============================================================================
                       coef          std err           t      P>|t|         [0.025          0.975]
-----------------------------------------------------------------------------------------------------
```

```
Conference    -1.772e+05    4.22e+04     -4.195   0.000        -2.61e+05     -9.3e+04
ConfW         2.205e+05     1.83e+05     1.206    0.231        -1.44e+05     5.85e+05
ConfL         2.805e+05     1.84e+05     1.527    0.131        -8.54e+04     6.46e+05
PF            1910.6927     1.24e+04     0.154    0.878        -2.28e+04     2.67e+04
PA            -1.074e+04    1.37e+04     -0.786   0.434        -3.79e+04     1.65e+04
Capacity      35.5398       6.323        5.621    0.000        22.945        48.135
AwayRecord    2.2e+04       6.3e+05      0.035    0.972        -1.23e+06     1.28e+06
PPG           -3.111e+05    1.38e+05     -2.252   0.027        -5.86e+05     -3.58e+04
PAG           3.34e+05      1.68e+05     1.992    0.050        -79.215       6.68e+05
OffRat        9.695e+06     2.25e+06     4.303    0.000        5.21e+06      1.42e+07
DefRat        -5.93e+06     1.66e+06     -3.583   0.001        -9.23e+06     -2.63e+06
==============================================================================
Omnibus:                0.171       Durbin-Watson:          1.727
Prob(Omnibus):          0.918       Jarque-Bera (JB):       0.070
Skew:                   -0.069      Prob(JB):               0.966
Kurtosis:               2.982       Cond. No.               1.32e+06
==============================================================================
```

This second model shows slight improvement, increasing the adjusted R-squared to 0.903. Nevertheless, ConfW, ConfL, PF, PA, and AwayRecord are still not significant at any p-value below 0.10. Because of this, a third model was built, removing these five variables.

```
OLS Regression Results
==============================================================================
Dep. Variable:          SchoolPay           R-squared:              0.911
Model:                  OLS                 Adj. R-squared:          0.904
Method:                 Least Squares       F-statistic:            136.5
Date:                   Wed, 30 Jan 2019    Prob (F-statistic):     6.76e-40
Time:                   16:12:16            Log-Likelihood:        -1306.4
No. Observations:       86                  AIC:                    2625.
Df Residuals:            80                 BIC:                    2640.
Df Model:               6
Covariance Type:        nonrobust
==============================================================================
                  coef          std err        t               P>|t|    [0.025        0.975]
--------------------------------------------------------------------------------------------
Conference    -1.471e+05    3.57e+04     -4.125          0.000   -2.18e+05     -7.61e+04
Capacity      38.5104       5.969        6.452           0.000    26.632       50.389
PPG           -2.222e+05    4.66e+04     -4.770          0.000   -3.15e+05     -1.29e+05
PAG           2.366e+05     5.06e+04     4.672           0.000   1.36e+05      3.37e+05
OffRat        7.728e+06     1.35e+06     5.733           0.000   5.05e+06      1.04e+07
DefRat        -6.799e+06    1.42e+06     -4.791          0.000   -9.62e+06     -3.97e+06
==============================================================================Omni
bus:                    0.357       Durbin-Watson:          1.652
Prob(Omnibus):           0.837      Jarque-Bera (JB):       0.310
```

```
Skew:                    -0.141    Prob(JB):                 0.857
Kurtosis:                 2.920    Cond. No.              1.05e+06
==============================================================================
```

This final model features a small improvement, increasing the adjusted R-squared to 0.904. But, contrary to the other two models, all variables were significant at a p-value of 0.01.

Results

With the three models built, they were ran against the test sets with the results shown in Table 1.

Table 1

Metric	Model 1	Model 2	Model 3
R-squared	0.708	0.712	0.726
RMSE	879,807	873,740	852,476
MAE	713,092	697,238	656,450

We can gauge that all three models suffer from an overfitting problem, given that their training error is larger than the testing error. However, the third model shows to be the superior model as the difference between the error rates is the smallest. Similarly, the models Root Mean Squared Error and Mean Absolute Error are the smallest, meaning that there is less deviation between the actual salaries and the predicted salaries.

Given these results, the salary for the Syracuse coach was predicted as follows:

Table 2

School	Actual	Model 1	Model 2	Model 3
Syracuse	2,401,206	3,355,886	3,291,394	3,385,218

Despite being the best model, the third model is the more bullish of the three, predicting that the coach salary is less than US$ 1 million than it actually should be. This gives the head of the athletic department ample breathing room to negotiate a fair salary with the coach. Nevertheless, this salary does not include any bonus whatsoever, meaning that any remaining room between the agreed salary and the recommended salary can be used towards this bonus.

Conclusion and Answers

Given the results, the initial questions can be answered. Regarding the Syracuse coach's salary, we should expect him to earn around US$3,385,218. If the objective is to give advice on how much he should be paid, the model can be interpreted as giving a hard cap on what he should be earning and consider every dollar above that value to be overpaying.

If the Syracuse coach were coaching in the Big 12 Conference, then his expected salary according to the model would be US$3,238,164. If he were coaching in the Big Ten Conference, then he would receive an expected salary of US$3,091,111. Given his track record, the third model feels that the Syracuse head coach has a better chance of earning a higher salary if he stays in the ACC.

As to the data itself, no schools were dropped. At one point, however, the observations for Baylor, Brigham Young, Rice, and Texas Christian were dropped as they didn't have any salary information for their coaches. Fortunately, data was found for these coaches, so the observations didn't need to be removed.

Regarding the effects of GSR and FGR on a coach's salary, the model shows that for every additional percentage point gained in each variable, a coach's salary is expected to increase by US$1,823 and US$5,957. Nevertheless, the model also shows that these two variables are not significant to the model, as they have p-values of 0.914 and 0.728. Given these results, both variables can be removed from the model without affecting the model's results, but improving it, as can be seen by the results of the second model.

As to the model, all three models had very good results, all have adjusted R-squares above 0.900. But, if we compare the R-square of the training sets to those of the testing sets, the superior model turns out to be the third model, which has an R-squared of 0.904 on the training set and a 0.726 on the test set. Though all three models suffer from slight overfitting, the third model has the lowest margin of error between both sets. From this third model, as well, we can gauge that the variable that has the most impact on a coach's salary is Offensive Rating which increases the coach's salary by US$ 7.7 million dollars for every additional point; however, Defensive Rating reduces a coaches salary by US$6.8 million dollars for every additional point.

Bonus questions

If we were to fit a hierarchical model (also known as a Linear Mixed Effects Model), the model shows that the Convergence of the model has a p-value below 0.05, meaning that the model is significant at the 5 percent level.

If we ran this model against the test set, we get the following results: R-squared of 0.726, RMSE of 852,797, and an MAE of 672,848. When compared to the third model, this model produces almost equal results regarding the R-squared and RMSE. But, the MAE performance worsens to that of the third linear regression model.

Double Bonus questions

If we ran the third model once again but this time removing the Conference variable, the model performs worse overall, but not by much, as the adjusted R-squared goes down to 0.885. As to the testing set, the R-squared becomes 0.715, with an RMSE of 869,946 and an MAE of 659,337; all three values performing worse than the third model but still better than the second model.