Lab Exercise 1

Bing-Je Wu

IST 718 – big data analytics

Table of Contents

[Introduction 2](#_Toc30495939)

[Dataset 2](#_Toc30495940)

[Methods for Analysis 2](#_Toc30495941)

[Data and Data Processing 2](#_Toc30495942)

[Exploratory Analysis 3](#_Toc30495943)

[Data Modeling - Multiple Linear Regression 5](#_Toc30495944)

[Model 1 5](#_Toc30495945)

[Model 2 5](#_Toc30495946)

[Model 3 6](#_Toc30495947)

[Final Model 8](#_Toc30495948)

[Questions 10](#_Toc30495949)

[Conclusion 10](#_Toc30495950)

# Introduction

American football is the most popular sport in the United States. It has a huge market either in the National Football League or in the National Collegiate Athletic Association. The lab exercise is to predict the projected salary of the football coaches for each school and answer the following research 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?

# Dataset

This lab exercise contains 5 datasets from different sources: (1) The coaches9 dataset is provided by the instructor as the main dataset for the entire lab exercise ( <https://github.com/2SUBDA/IST_718/blob/master/Coaches9.csv> ). (2) The football\_stadium dataset is from the Wikipedia as supplement for providing the information of stadium capacity ( <https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums> ). The program for the web scrapping is ‘Web scrapping - wiki stadium tables.ipynb’. (3) The 2006\_corhort dataset is scrapped from the NCAA website for the GSR information ( <https://web3.ncaa.org/aprsearch/gsrsearch> ). The program for the web scrapping is ‘Web scrapping - 2006 Cohort 2012-2013 Year.ipynb. (4) The division\_I dataset is scrapped from the Wikipedia containing the common name and official school name information as a dictionary to join all the table together. The program name is ‘Web scrapping - wiki division I.ipynb’. ( <https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_institutions> ). (5) The 2019\_record was obtained from the NCAA website providing the information of the school performance in the year of 2019 ( <https://www.ncaa.com/standings/football/fbs> ).

# Methods for Analysis

The salary will be predicted by implementing multiple learning regression algorithm with OLS (Ordinary least squares) technique. The 5 datasets gathered from different sources will be used as training data to build the final model to predict the salary. Three candidate models will be proposed for comparison. The model with the highest performance will be chosen as the final model and to answer the research questions above.

# Data and Data Processing

After importing the five datasets, some cleaning steps need to be performed. For the coaches9 dataset, the data should be normalized, such as removing ‘$’, '--', and ',' from the strings and converting those strings into int datatype. Four school records have no information on 'SchoolPay', 'TotalPay', 'Bonus', 'BonusPaid', 'AssistantPay', 'Buyout'. Because it could not contribute on ‘SchoolPay’ prediction, those four records (Baylor, Brigham Young, Rice, and Southern Methodist) were dropped from the dataset. A new feature, ‘SuccessRate’, was added, derived from ‘BonusPaid’ and ‘Bonus’ (BounsPaid/Bonus).

Because there are many schools in the US and school names used from each source are different. The division\_I dataset was used as the bridge to connect the coach9 dataset, 2006\_corhor dataset, football\_stadium dataset and 2019\_record dataset. It contains the ‘School’ variable which is the longer school name and ‘Common Name’ variable which is the shorter name. Thus, it is used as the foreign keys to join others. Some discrepancies of the names were revised for smoothing the join process. The rules for join are as following:

1. Coaches9 dataset inner join with Division\_I dataset
2. The merged dataframe from the last step left join with ncaa\_2006\_cohort dataset
3. The merged dataframe from the last step left join with 2019\_record dataset
4. The merged dataframe from the last step left join with football\_stadium dataset

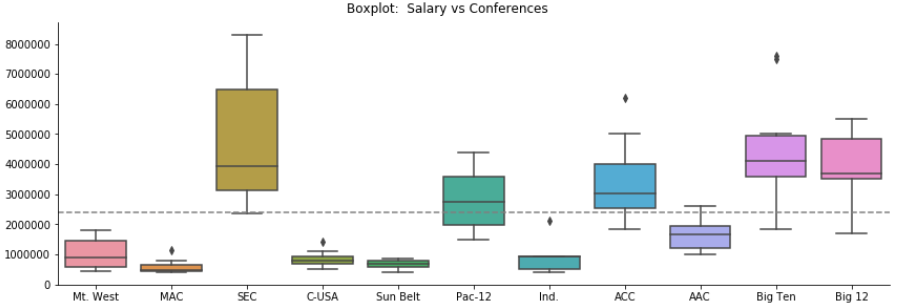
Some of the data still have missing values. To fill in the missing value, the median replacement technique was implemented, or other rules were taken in placed. For the missing values of GSR, graduation success rate, 4 schools cannot find the GSR for 2006 cohort. As an alternative, filled in the most recent GSR or use the average of the GSRs in the following cohort years.

For the ‘Buyout’ variable, there were 20 missing values. It was replaced by the median based on each of conference that the school belongs to.

|  |
| --- |
| Conference Median  AAC 1174999.75  ACC 5593750.00  Big 12 7150000.00  Big Ten 12065243.25  C-USA 2850000.00  Ind. 0.00  MAC 965150.00  Mt. West 1432868.50  Pac-12 7927125.25  SEC 12375000.00  Sun Belt 846145.50 |

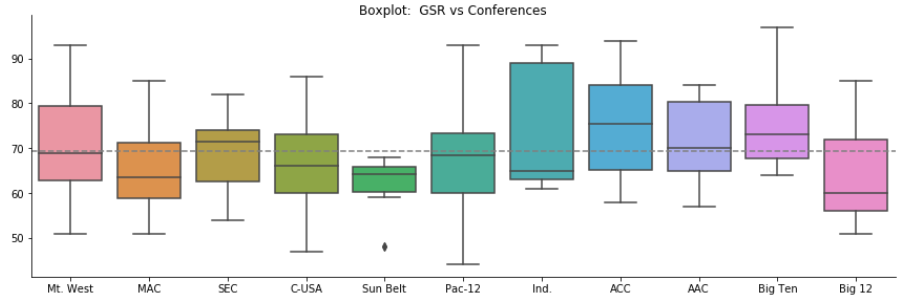
# Exploratory Analysis

The merged dataframe was used for statistics analysis to decide what variables are going to be used as predictors to predict the salary. Since we have known there is a pattern on the conference variable in practice. We are going to look deeper on this variable. Three boxplots were shown as below and the dotted line represents the mean of the y metric: (1) Salary vs Conferences (2) GSR vs Conferences (3) PF vs Conferences

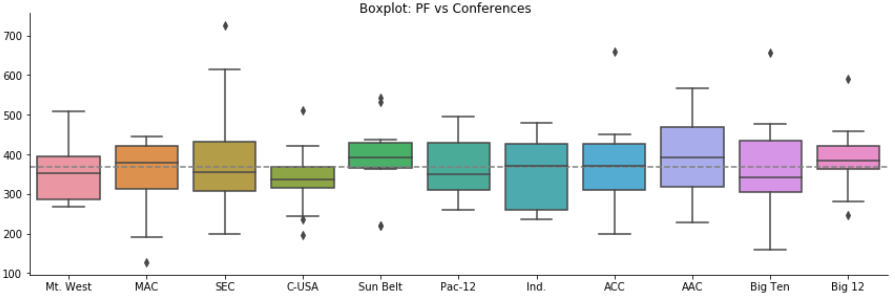


The Salary vs Conferences plot shows the difference of the salary range on each of conferences. Thus, we can infer that the conference will be a key factor to predict the salary.

The GSR vs Conferences plot shows that most of the range of GSR are overlapping with each other. It may explain why the GSR variable is not a significant predictor in our final model. Its p-value is bigger than the threshold, 0.05. And the confidence interval of the GSR coefficient from the final model overlaps with zero. That indicates the ‘GSR’ is not the effective predictor for salary prediction.

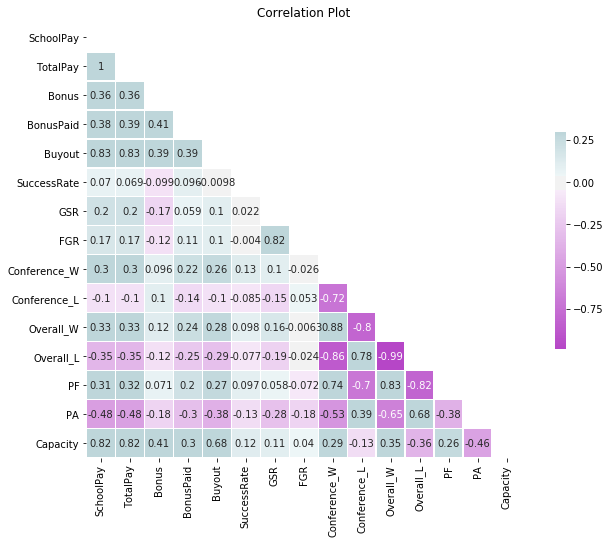


The PF vs Conferences plot shows that not all the conferences overlap with the mean, 369.192. It implies that somehow the PF could contribute a small degree of effect on the salary prediction. The final model also provides the evidence that the coefficient of the PF variable is 3069.9392 and its p-value is significant.



The Correlation plot shows the relationship between variables. Because we set 'SchoolPay' as the response variable, we need to look at the relationship between 'SchoolPay' and others. Both variable, 'Capacity' and 'Buyout', have high positive correlation with 'SchoolPay'. ‘Conference\_W', 'Overall\_W' and 'PF' have positive correlations which also make sense. 'PA' and 'Overall\_L' have negative correlations which make sense as well. 'TotalPay' is excluded since it is driven from 'SchoolPay'. 'Bonus' and 'BonusPaid' have many missing values. Thus, a re-engineered variable, 'SuccessRate' is added to predict Salary. Since 'FGR' is not fully provided, we will use 'GSR' instead.

To sum up, there are 10 numerical variables, 'Capacity', 'Buyout', 'GSR', 'SuccessRate' ,'Conference\_W', 'Overall\_W', 'PF', 'PA', 'Conference\_W' and 'Overall\_L' that we are using to predict Salary.



# Data Modeling - Multiple Linear Regression

Models were built by implementing the train-test split technique. Approximately 34% of the dataset were randomly selected as testing set with uniform distribution.

## Model 1

**SchoolPay ~ Conference + Buyout + SuccessRate + GSR + Conference\_W + Overall\_W + PF + Capacity**

The Adjusted R-squared for the model built in training set is 0.899. The Adjusted R-squared for the model built in the full data set is 0.875. The plot of prediction in testing set for the Model 1 is shown as below

## Model 2

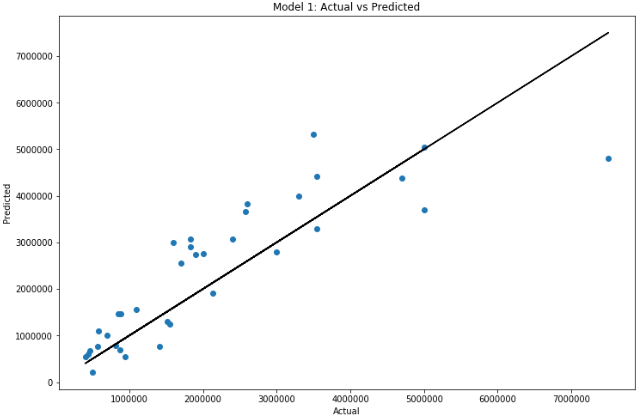
**SchoolPay ~ Buyout + SuccessRate + GSR + Conference\_W + Overall\_W + PF + Capacity**

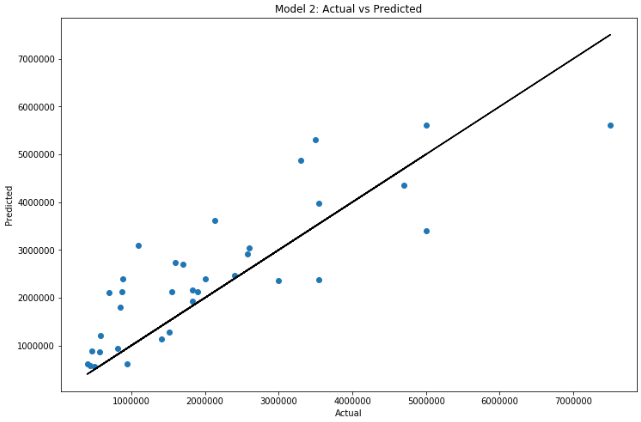
The Adjusted R-squared for the model built in training set is 0.835. The Adjusted R-squared for the model built in the full data set is 0.811. The plot of prediction in testing set for the Model 2 is shown as below

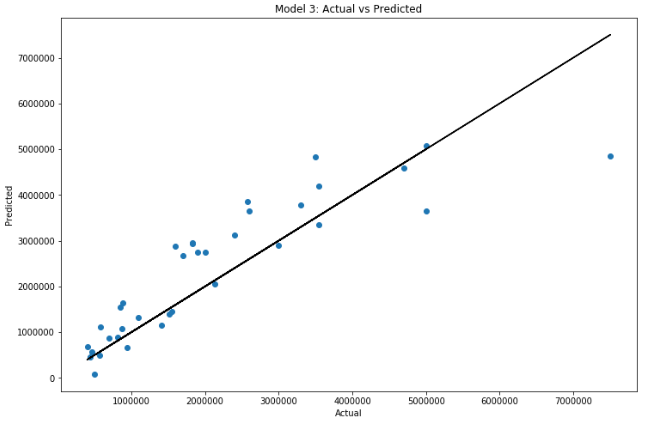
## Model 3

**SchoolPay ~ Conference + Buyout + SuccessRate + GSR + Capacity + Conference\_W + Conference\_L + Overall\_W + Overall\_L + PF + PA**

The Adjusted R-squared for the model built in training set is 0.902. The Adjusted R-squared for the model built in the full data set is 0.882. The plot of prediction in testing set for the Model 3 is shown as below



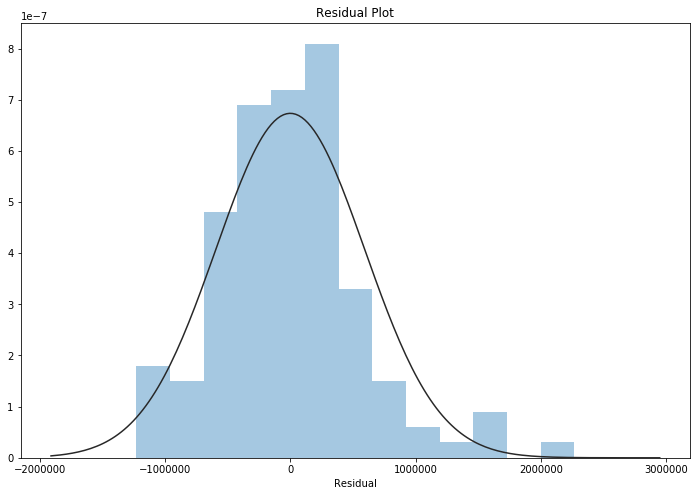




## Final Model

The model 3 was picked as the best model among three. It has highest adjusted R-squared value and the initial omnibus test of the R-squared is significant, suggesting that the R-squared is not equal to zero. The overall performance of the final model is pretty good. The confidence intervals for the coefficient of ‘SuccessRate’, ‘GSR’, ‘Conference\_W’, ‘Overall\_W’ and ‘Overall\_L’ overlap with zero. It implies that those variables might not be good predictors to the model since it has the chance that the coefficient of those variable might equal to zero.

|  |
| --- |
|  |



|  |
| --- |
| The median of the residual: 6794.519332478521  The Q1 of the residual: -392632.1469536894  The Q3 of the residual: 309400.1770760593 |
| Omnibus: 17.963534579371476,  p = 0.0001256805371990896 |

The residuals median is a bit far from zero and the absolute value of Q1 and Q3 are not similar indicating that there is a skewness in the residuals. The residual distribution is a bit skewed to the right (not symmetrically distributed). It indicates that the result of the regression analysis could be inaccurate.

The null hypothesis of the Omnibus test is that the distribution of the residual is normal. The p-value of the test is statistically significant and shows the evidence that the residuals are not normally distributed about zero.

The initial Omnibus test for the regression equation has F-statistic=49.28 and p-value < 0.001. The null hypothesis is that the R-squared is equal to zero. We reject the null hypothesis since the p-value for the F-test is extremely small. Thus, the overall R-squared is significant. Because we are using multiple linear regression model, we need to use adjusted R-squared to have unbiased estimate instead of overall R-squared. The adjusted R-squared shows independent variables are accounted for about 88.2% of variability in dependent variable.

# Questions

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

The recommended salary for the Syracuse football coach is $2,743,824

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

Since we don’t have School in the Big East conference in our dataset, the coach salary in the Big East conference cannot be predicted correctly. The salary for the Syracuse football coach in the Big East conference is approximately about $1,924,624. The salary for the Syracuse football coach in the Big Ten conference is $2,865,830

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

There are four records having missing values for 'SchoolPay', 'TotalPay', 'Bonus', 'BonusPaid', 'AssistantPay', 'Buyout'. Because we use the 'SchoolPay' as out response variable and those four records could not contribute on 'SchoolPay' prediction, the 4 schools (Baylor, Brigham Young, Rice, Southern Methodist) were dropped from the dataset.

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

Each unit of the graduation rate can contribute the increase of $7,833.8737 on projected salary.

1. **How good is our model?**

The model can explain about 88.2% of variability in dependent variable.

1. **What is the single biggest impact on salary size?**

Being in the Big 12 conference will be the biggest impact on the salary prediction. If a school is in the Big 12 conference, it can increase the projected salary by $1,196,000.

# Conclusion

The salary can be projected through several datasets, such as coaches' payment information, school graduation rate, stadium size, and team performance…etc. The model was built with an overall good performance and can explain 88.2% variability in the projected salary. More data or accurate data could increase the performance of the model. Since the model over-fitting issue was not discussed during the exercise, the k-fold cross validation technique could have been implemented to prevent over-fitting. The donation information was not used in this lab exercise. It could be combined into the dataset and build a better model in the future.