IST-718 Homework/Lab-1 (Week3)

Prof. Jonathan M Fox

Student: Sharat Sripada (vssripad)

## Introduction

This case study provides an opportunity to demonstrate the ability to combine datasets and produce meaningful analysis learned through the first three-weeks of the course.

Specifically, we are looking to answer the following questions:

* What would his salary be if we were still in the Big East? What if we went to the Big Ten?
* What schools did we drop from our data, and why?
* What effect does graduation rate have on the projected salary?
* How good is our model?
* What is the single biggest impact on salary size?

## Dataset

The dataset initially related to Coaches data (an excerpt is seen below) was loaded into a data-frame **df\_coaches**. This would be the master data-frame:

Graphical user interface, text

Description automatically generated with medium confidence

**NOTE:**

Two additional datasets related to Graduation success Rate (GSR)/Federal Graduation Rate (FGR) and Stadium capacities were merged into the master data-frame.

First, data related to GSR/FGR was retrieved from *<https://web3.ncaa.org/aprsearch/gsrsearch>* and loaded into a data-frame **df\_gsr\_fgr**:

A picture containing table

Description automatically generated

Finally, data related to stadium sizing was retrieved from source *<https://www.collegegridirons.com/comparisons-by-capacity/>* and loaded into data-frame **df\_stadiums:**

Graphical user interface, text, application, email

Description automatically generated

## Data cleaning/munging

Data was cleaned and munged in steps.

**STEP-1:**

On data-frame **df\_gsr\_fgr** missing values were replaced with statistical mean for the respective GSR/FGR data. This may not always be the best approach and could potentially skew data. Also, we convert the datatypes for the GSR and FGR columns to type int:

Graphical user interface, text, application, email

Description automatically generated

1. Data-frame **df\_coaches** were munged for the ‘Capacity’ column replacing characters to make them numbers for data-analysis:

A picture containing chart

Description automatically generated

Data from the two data-frames were then merged into the master data-frame **df\_coaches** before preparing the data for visualization and modelling:

* replace missing values/zeros with mean/avg where applicable
* check data types & modify appropriately using *astype(*)
* remove '$' symbol using Python regex and convert currency columns - *SchoolPay, TotalPay, Bonus, BonusPaid, AssistantPay, Buyout* to float.

A sample of the merged data-frame and columns is seen below:

Table

Description automatically generated

Data types on all columns within data-frame **df\_coaches**:

School object

Conference object

Coach object

SchoolPay float64

TotalPay float64

Bonus float64

BonusPaid float64

AssistantPay float64

Buyout float64

GSR int64

FGR int64

State object

Capacity float64

### Removal of data-records

During the process of data-retrieval there were instances where few records had to be removed. For instance:

* when retrieving data related to GSR/FGR certain schools from the Coaches dataset could not be matched or found. This resulted in *EIGHTEEN* records being removed or dropped
* similarly, *FIFTEEN* schools could not be matched or found in stadium capacity data. These were removed too.

**NOTE:**

The original data-frame was preserved, and all the trimming was saved to df\_coaches\_trim

After all, the size of Coaches data is shown below – comprising up to *SEVENTEEN* fewer records:

df\_coaches shape: (129, 13) **VS** df\_coaches\_trim shape: (112, 13)

Missing values in df\_coaches dataframe: False

## Exploratory Data Analysis

Since the problem we are trying to solve is related to predicting the Salary of coaches some of the exploratory data analysis or descriptive statistics and visualization are centered around the column TotalPay from the dataset.

Seen below are the min, max, mean and median values related to *TotalPay* (in USD) paid to Coaches and another column that seemed to have a strong correlation to it, *Capacity* of stadiums:

Graphical user interface, table

Description automatically generated

This distribution and variance can also be seen in a swarm-plot. This particularly indicates the larger set of salaries distributed *<4mi USD* and just one data-point *>8mi USD*:

Chart, scatter chart

Description automatically generated

The min and max *TotalPay* values were used to retrieve data of Coaches who were paid the least and most salaries:

Text

Description automatically generated

## Results

### Model-1

The first model will attempt to predict the salary of Coaches (here *TotalPay*) using Linear Regression/commonly known Ordinary Least squares method with variables (GSR, FGR, Capacity) from the dataset.

A summary of the model-fit and results as follows:

Text

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generated

Interpreting the results:

* The equation of linear regression can be written as:

*TotalPay = 72.2476 \* Capacity + 2.686e+04 \* FGR - 8909.0952 \* GSR - 2.041e+06*

* **R-square value**

R-square = 0.756 is an indication of a good fit and is reflective of how close the data is to line of

best-fit/regression.

* **P-value**
  + P-value for GSR/FGR >0.05 and may therefore not be statistically significant in this dataset.
  + Stadium Capacity with P-value = 0 is statistically significant.

Further, a plot of the line of least squares for each of the variables was made:

Chart, scatter chart

Description automatically generated

Based on the linear regression model/Model-1, the prediction of Coach Dino Babers' salary is approximated to **2,460,942USD** about 2.5% more than his current pay.

A picture containing table

Description automatically generated

### Model-2

This model will make a similar prediction as the first, by including an additional response variable *Conference*. Since this variable is categorical and non-numeric the values in the dataset will be mapped and converted to numeric data using a key-value map or Python dictionary:

{'Mt. West': 0, 'MAC': 1, 'SEC': 2, 'Sun Belt': 3, 'Pac-12': 4, 'Ind.': 5, 'ACC': 6, 'AAC': 7, 'C-USA': 8, 'Big Ten': 9, 'Big 12': 10}

A summary of the model-fit and results as follows:

A picture containing text

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generated

Interpreting the results:

* **R-square value**

A marginally better R-square = 0.776

* **P-value**
  + P-value conference\_num <0.05 indicates being statistically significant.

Pair plots of the line of least squares updated with Conference:

Chart, scatter chart

Description automatically generated

Based on the linear regression model/Model-2, the prediction of Coach Dino Babers' salary is approximated to **2,769,900USD** if Syracuse was in Big-Ten conference.

Graphical user interface, application

Description automatically generated

**NOTE:**

*Conference\_num* is changed to 9 which maps to the ‘Big Ten’ conference per the dictionary.

## Conclusions

The conclusions are mostly presented by answering the questions we originally sought out to answer and covered to some depth throughout this document.

**Questions:**

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

The dataset across all the data-frames did not comprise any data from Big East.

However, taking column *Conference* into consideration when training the model, prediction has it that Coach Dino Babers’ salary could be ~15% higher and approximately be **2,769, 900USD**.

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

As stated earlier in the *Data cleaning/munging section*, had to drop 17-rows or schools owing to the following:

* + when retrieving data related to GSR/FGR certain schools from the Coaches dataset could not be matched or found. This resulted in *EIGHTEEN* records being removed or dropped
  + similarly, *FIFTEEN* schools could not be matched or found in stadium capacity data. These were removed too.
* *What effect does graduation rate have on the projected salary?*

Graduation rate or GSR does not seem to have a significant role. This is manifested by high P-values and the line of least square plots in the *Results section*.

* *How good is our model?*

One of the measures of a good linear regression model is an R-square value tending towards 1.0 and we have both models yield >0.75. making a moderately good model.

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

*Capacity* of stadiums has the largest impact on the Salary of coaches. This is corroborated with low P-values and low variance when plotted on the line of least squares.

Speaking in more general terms, bigger stadiums or more seating capacity would certainly put more revenue into the sport as it promotes bolstering the local economy (sale of merchandise, food/beverages, car-parking etc.) and this likely drives the stakes up for sustaining good Team performances, better Coaches and therefore higher Coach salaries.