Text

Description automatically generated

**Course:** 2023-1004:16645 IST-718 Big Data Analytics

**Name:** Jacob N. Conard

**NetID:** jconard

**Assignment:** Laboratory Exercise - 1

**Date due:** 22 October 2023

**Date submitted:** 18 October 2023

**Number of Pages:** 14

**Introduction**

In the ever-evolving landscape of data science, the OSEMN model has emerged as a foundational framework that plays a pivotal role in transforming raw data into actionable insights. Acronymized from the stages of Obtaining, Scrubbing, Exploring, Modeling, and Interpreting data, the OSEMN model provides a structured and comprehensive approach for data scientists to navigate the complex process of extracting knowledge from diverse and often messy datasets. As the field of data science continues to expand and impact various industries, the OSEMN model stands as a crucial methodology, guiding practitioners through the intricate journey of data analysis, helping them uncover valuable patterns, make informed decisions, and derive meaningful conclusions from the vast sea of information.

The main research question for this paper is how can we recommend the best salary for the next head football coach at Syracuse University. We will do this by using a dataset that contains salary information for current head football coaches and the University where they work. Additionally, we will enrich the dataset to include stadium size information, graduation rate information, and the most recent school’s win / loss records from 2022. We are going to try to answer the following key questions in the report:

1. What is the recommended salary for the Syracuse football coach?
2. What would his salary be if Syracuse moved to the Big Ten? What would his salary be if Syracuse moved to the SEC?
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?
7. Develop a geographic visualization that in your view best depicts the conference’s median salary.

**Loading and Cleaning the Data**

In order to answer the main research question of how can we recommend the best salary for the head football coach at Syracuse we used four different datasets. The main coaches salary dataset that was given to us, and the other datasets and their locations:

1. Football stadium data 🡪 [*https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv*](https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv)
2. Graduation data 🡪 <https://data.world/databeats/college-completion>
3. Wins/Losses 🡪 Provided in the coaches .xls at the following github link <https://github.com/2SUBDA>

A screenshot of a computer program

Description automatically generated

Once the data was read in, the next step was to ensure that all of the column names matched so I decided to make all of the columns lowercase and all of the string data in the rows lowercase as well:

A screenshot of a computer code

Description automatically generated

The next data cleaning task that we had to accomplish was to normalize some of the school names since we were using data from four different places. The following code will show how those school names were normalized:

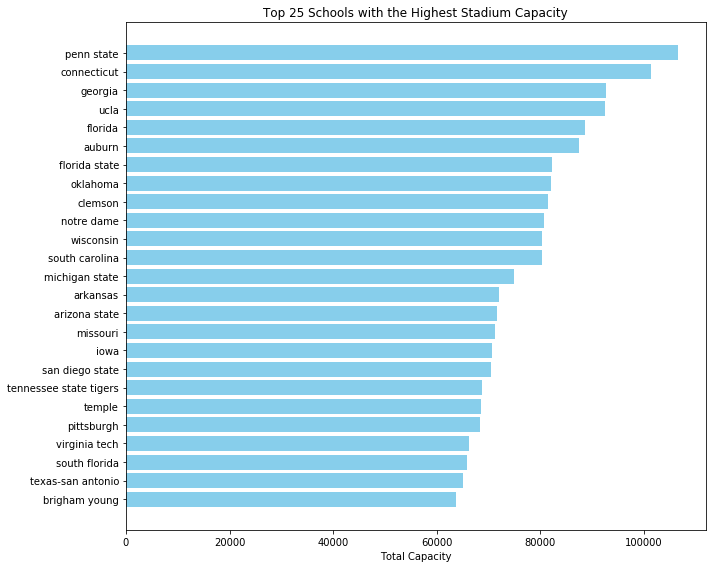
A screenshot of a computer program

Description automatically generated

After those small tweaks we are ready to start exploratory data analysis of the data.

**Exploratory Data Analysis [EDA]**

Now that the data cleaning and data normalization portion is complete, we can start doing the exploratory data analysis. We need to keep in mind the main research question that we are trying to answer, and that is how much should we pay the next Syracuse head football coach. Since we have four datasets that we are going to combine to help answer this question there are some interesting things that we would like to analyze during the EDA phase.

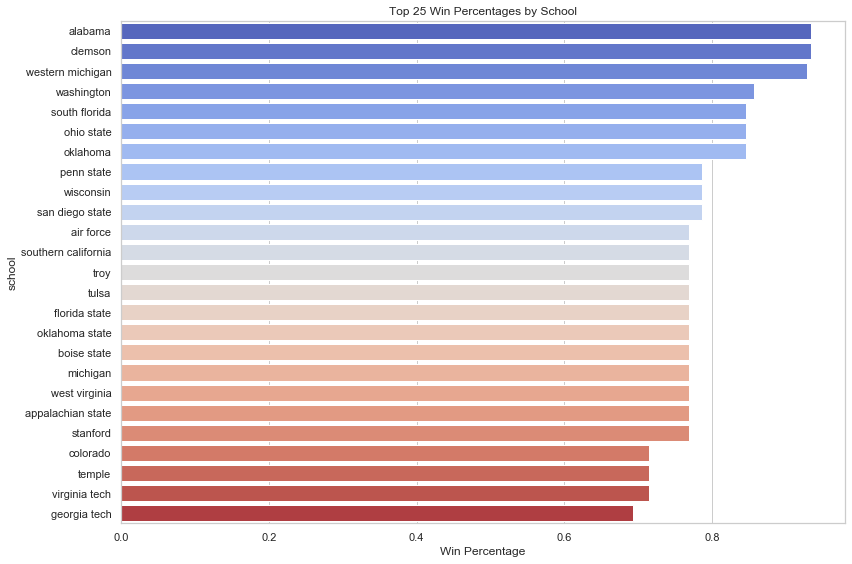
**Top 25 Schools Based on Stadium Capacity**

**Top 25 Schools Based on Coaches Salary**

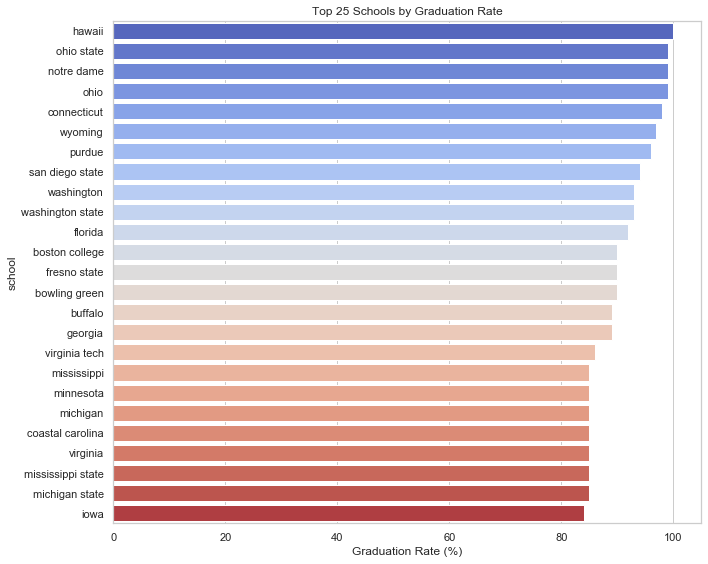
A red and white striped pattern

Description automatically generated

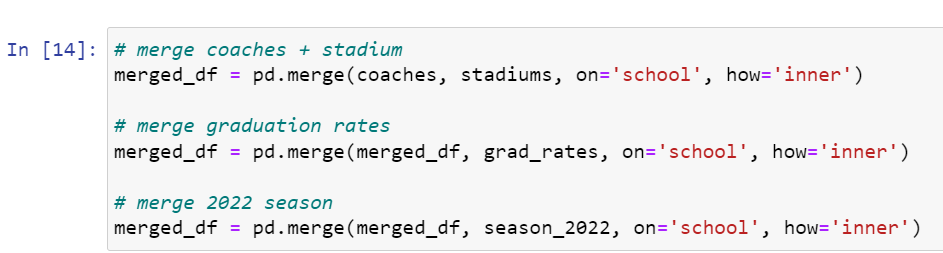
**Top 25 Schools Based on Win Percentage**



**Top 25 Schools Based on Graduation Rate**

****

After looking at basic EDA for all of the different datasets we then merged the datasets in order to utilize linear regression to answer the main research question.



Below is an example of the merged dataframe:

A screenshot of a computer

Description automatically generated

Now that the data is merged we can visualize a correlation matrix to try to get an idea of what variables may be related to each other:

A screenshot of a computer

Description automatically generated

Based on the correlation matrix we can visually see that the winning percentage is correlated to the total pay of the coach’s salary, we will ensure that we keep this in mind later when we utilize the data to answer the main research question as well as the follow-on questions that were mentioned in the introduction.

**Salary vs. Game’s Won**

A graph with red and blue dots

Description automatically generated

**Fit a Regression Model**

The next step in the assignment is to fit a regression model to the data. The steps of fitting a regression model can be seen below:

1. Data Collection: The first step is to collect the dataset that contains the variables we want to analyze. We should have at least two variables: one dependent variable (total coaches’ salary) and one or more independent variables (graduation rate, win percentage, etc.).
2. Data Preprocessing:

* Data Cleaning
* Feature Selection
* Data Splitting: Split the dataset into a training set and a test set. The training set is used to train the model, and the test set is used to evaluate its performance.

1. Model Selection: Choose linear regression as the modeling technique if we believe there is a linear relationship between the independent and dependent variables. There are different types of linear regression models, such as simple linear regression (one independent variable) and multiple linear regression (more than one independent variable).
2. Model Training: Fit the linear regression model to the training data. The model will estimate the coefficients that define the relationship between the independent and dependent variables.
3. Model Evaluation:

* Model Performance Metrics: Use appropriate metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), or R-squared (R2) to assess the model's performance on the test data.

1. Model Interpretation: Interpret the coefficients of the linear regression model. The coefficients indicate the strength and direction of the relationship between the independent variables and the dependent variable. A positive coefficient means that as the independent variable increases, the dependent variable also increases, and vice versa.
2. Model Validation: Ensure that the model's assumptions are met.

Fit the linear model from the merged dataset of coaches’ salary, graduation rate, win/loss record, and stadium size:

A screenshot of a computer

Description automatically generated

Now that we have the linear regression model fit how well is it performing?

A screenshot of a computer error

Description automatically generated

Based on these scores we can make the following assumptions:

* Mean Squared Error (MSE): The MSE is a measure of the average squared difference between the actual and predicted values. In this case, the MSE is approximately 1,265,310,163.05. A higher MSE value suggests that, on average, the model's predictions have relatively large errors when compared to the actual values.
* Root Mean Squared Error (RMSE): The RMSE is the square root of the MSE and provides a more interpretable measure of error. In this case, the RMSE is approximately 35,571.20. The RMSE gives us an estimate of the average absolute error in the same units as the target variable. In this context, the RMSE value indicates that, on average, the model's predictions are off by approximately 35,571 units (units in this case is related to coach’s salary).
* Mean Absolute Error (MAE): The MAE measures the average absolute difference between actual and predicted values. With a value of approximately 25,574.47, the MAE tells us that, on average, the model's predictions are off by around 25,574 units. The MAE is less sensitive to outliers compared to the MSE and RMSE.
* R-squared (R2): The R-squared score, which is very close to 1 (0.9996), is a measure of how well the independent variables explain the variance in the dependent variable. In this case, the R2 score is extremely high, suggesting that the model explains nearly all of the variance in the total pay. An R2 value close to 1 indicates a very good fit.

Now that we have a linear model fit, we can now attempt to answer the main research question as well as the other research questions.

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

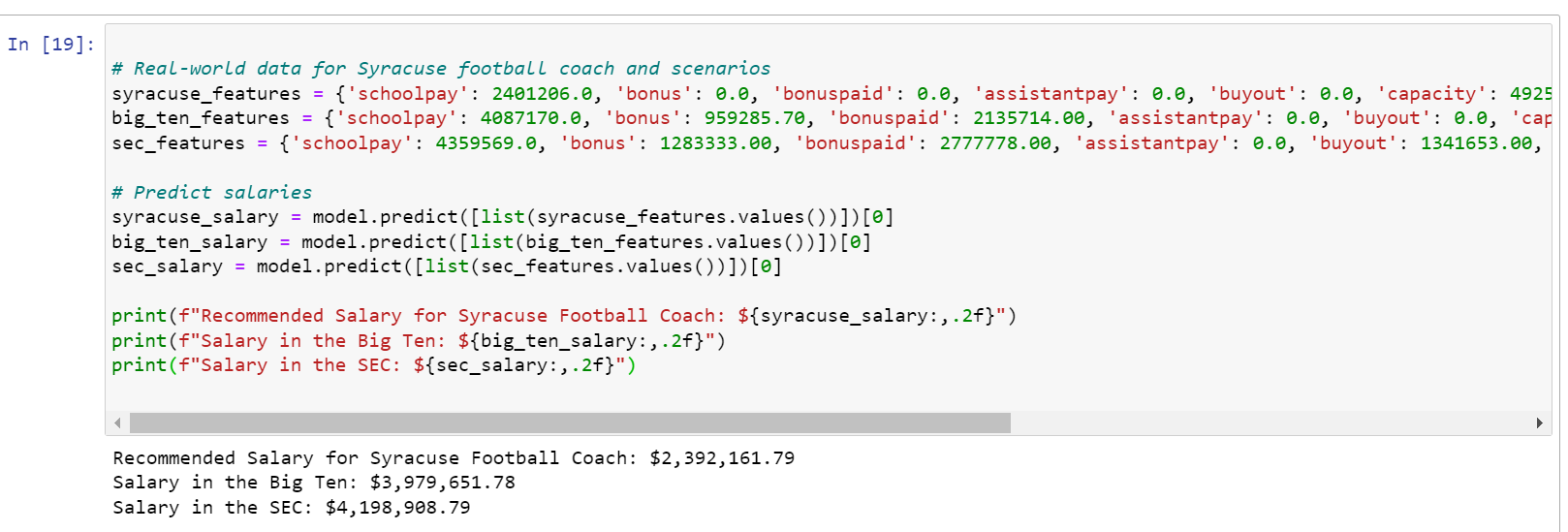
The recommended salary for the next Syracuse head football coach is $2,392,161.79.

A screenshot of a computer program

Description automatically generated

**What would his salary be if Syracuse moved to the Big Ten? What would his salary be if Syracuse moved to the SEC?**

The recommended salary for the next Syracuse head football coach if Syracuse moved to the Big Ten is $3,979,651.78. The recommended salary for the next Syracuse head football coach if Syracuse moved to the SEC is $4,198,908.79.



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

We dropped the following schools from out data and that is because those schools had no values for total pay for their coaches, and that means that it is impossible to utilize this data in the linear regression model for prediction:

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer code

Description automatically generated

This is an interesting statistic that as the graduation rate increases the projected salary is expected to decrease meaning that there is a negative correlation between the projected salary of the head football coach and the graduation rate. This is likely because schools that are successful in college football tend to send more players to the National Football League and those players do not graduate. This makes sense that the schools that pay their football coaches more money would have a lower graduation rate and explains the negative correlation between graduation rate and projected salary.

**How good is our model?**

This was already explained and answered in the linear regression section; however, the details are explained and reiterated again below:

A screenshot of a computer error

Description automatically generated

Mean Squared Error (MSE): The MSE is a measure of the average squared difference between the actual and predicted values. In this case, the MSE is approximately 1,265,310,163.05. A higher MSE value suggests that, on average, the model's predictions have relatively large errors when compared to the actual values.

R-squared (R2): The R-squared score, which is very close to 1 (0.9996), is a measure of how well the independent variables explain the variance in the dependent variable. In this case, the R2 score is extremely high, suggesting that the model explains nearly all the variance in the total pay. An R2 value close to 1 indicates a very good fit.

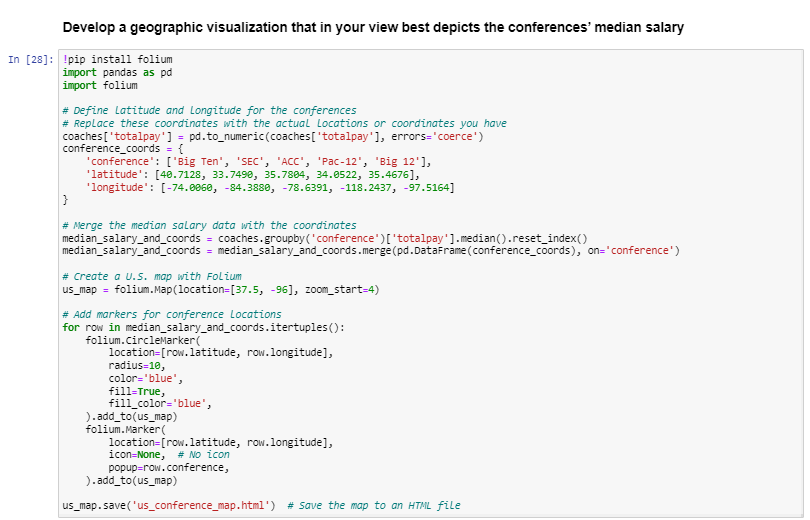
**What is the single biggest impact on salary size?**

The single biggest impact on salary size is winning percentage with a coefficient of 2335135.23. The coefficient of 2,335,135.23 signifies the impact or contribution of the winning percentage feature on the dependent variable (total salary). Specifically, if all other variables are kept constant, a one-unit increase in the winning percentage feature is associated with an increase in salary size by 2,335,135.23 units (the same unit of measurement as the dependent variable which in the case of this model is the total coach’s salary).

A screenshot of a computer program

Description automatically generated

**Develop a geographic visualization that in your view best depicts the conference’s median salary.**



A map with blue pins

Description automatically generated