Nicholas Lichtsinn

IST 718

Lab 3

**Introduction**

Finding an acceptable salary range or minimum number is particularly important to most applicants to any job posting. Knowing what that number should be allows an applicant more control over the process and usually, better pay. For college football coaches and other positions with salaries that are open to the public, this data can be used to determine the market rate for a specific position.

This project will use linear regression models created with Python to predict the market rate for the Syracuse football coach based on a variety of factors.

**Analysis**

Data Preparation:

A csv file with data for school, coach name, salary, conference, bonus amounts and buyout clauses was provided and a tsv file of additional data on school’s academic progress rates, private vs public schools, historically black universities, and squad size was pulled from the ICPSR database.

Both files were loaded into python, an average APR was calculated from 2004-2014 for each school and the datasets were merged after cleansing to match school names.

The coaches9 csv contained 129 rows and 9 columns, after merging to include the additional variables the cleaned dataset had 68 rows and 7 columns.

Some initial plots were created to show the APR average by conference, as well as the distribution of the total pay variable and total pay by conference.

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A train dataset was created using 80% of the dataset and the remaining 20% was used to create a test dataset to test the accuracy of the models.

**Results**

Once the dataset was cleaned and train and test sets were created the linear regression models could be created. The first model build was a simple linear regression model to predict Total Pay as a function of Conference as Conference had the highest variability.

Table

Description automatically generatedThis model performed alright, with an R-Squared of 63.7% and Adjusted R-squared of 55.6%. For a simple model this is alright, a lot of the conferences are not statistically significant but describe about 64% of the variation in the Total Pay.

The second model created used APR average, Private, Squad Size and HBCU in addition to Conference to predict of Total Pay and was tested on the training data. This model had a higher R-squared at 75.9% and Adjusted R-squared at 68.5%. the only additional variable that was not significant was the Squad Size with a p value of 0.144.

Table

Description automatically generated

Chart, box and whisker chart

Description automatically generatedA third model was created taking out Squad Size and returned an R-squared of 74.7% and Adjusted R-squared of 67.6% performing slightly worse than the second model. Looking at Private vs Public to see if the data looks as it should based on the p value.

This boxplot looks as it should so the prediction will use this model moving forward.

The third model was tested on the test set of data and this model had an R-squared of 70.4% and an Adjusted R-squared of 63.9%. However, the APR average and HBCU variables were both not significantly significant which indicated overfitting of these variables on the train data and these variables were removed.

Table

Description automatically generated

A model of Total Pay as a function of Conference and Private vs Public achieved a slighter lower R-squared of 70.3% and a slightly higher Adjusted R-squared of 64.4% on the test data.

**Conclusion**

Using the final model, the predicted salary for the Syracuse football coach was $2,920,894 which is about $500,000 higher than what the current coach is being paid. The salaries of the 3 ex-Big East schools were used to determine what the salary should be if Syracuse was still in the Big East conference and the expected salary was $2,720,858, so higher than the current coach and lower than expected if Syracuse stays in the ACC. If Syracuse had joined the Big Ten however, this model expects the coach to earn $3,597,142 as the Big Ten has higher wages.

To create these models sixty-one schools were dropped from the data because they did not match up with schools in the dataset from ICPSR with an accompanying APR score to show their program’s focus on academics. In this model Graduation Rate or APR did not have a statistically significant impact on the model and was ultimately dropped as most conferences had a similar average APR apart from the Sun Belt Conference.

This model performed decently, accounting for 64.4% of the change in Salary is not terrible but not confident enough to rely on this model alone. The largest impact on Salary or Total Pay was Conference. Although the main difference being within a power five conference (ACC, Big Ten, Big 12, Pac 12, SEC) and being in a smaller conference as the variance within the two groups is significantly less than it is between them.