2021-0407 IST 718 Big Data Analytics

Lab 3 Assignment

Prepared For: Jon Fox

Prepared By: Will Holt

Due: Saturday, April 24, 2021

The Syracuse football team finished its 2017 season with an overall record of 4-8, which included a 2-6 record within the ACC. While the season did end with five consecutive losses and a missed opportunity to play in a Bowl game, it was highlighted by a 27-24 upset win over No.2 Clemson. Struggles could be attributed to the team playing the 19th most difficult schedule according to SportsReference (https://www.sports-reference.com/cfb/schools/syracuse/2017-schedule.html), but that did not reduce the concern many fans had about head coach Dino Babers. In two seasons, the former Bowling Green coach, has compiled a pair of 4-8 records, which has led Syracuse Athletic Director, John Wildhack, to review the contract for Babers.

Wildhack has asked the Syracuse iSchool to review the salary for Babers and other coaches throughout the country to address the following key questions:

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

In order to address these questions, it is important to provide the proper background and context so that Wildhack and other decision makers can understand how the questions were answered and to allow for anyone else to understand the approach so that they can identify where areas of improvement can be made.

The Data

Variables

An initial dataset of 129 coaches was provided to the iSchool and contained data for each coach's salary, which was broken down by School Pay, Total Pay, Bonus, Bonus Paid, and Buyout, as well as data related to Assistant Pay, and the conference in which he coaches. The iSchool determined that additional data may be helpful for this analysis and supplemented it with each coach's career wins (CareerW), career losses (CareerL), career ties (CareerT), career winning percentage (CareerWinPct), average simple rating system (AvgSRS), average strength of schedule faced (avgSOS), the number of seasons in which the coach's team finished the season ranked in the top 25 (SeasonsRanked), stadium capacity for home games (StadiumCapacity), the team's graduation success rate (GSR), the federal graduation rate (FGR), the school's total wins in 2017 (SchoolWins17), and the schools total losses in 2017 (SchoolLosses17).

CareerW, CareerL, CareerT, CareerWinPct, avgSRS, avgSOS, SeasonsRanked, SchoolWins17, and SchoolLosses17 data were obtained from SportsReference College Football (https://www.sports-reference.com/cfb/). GSR and FGR were obtained from the NCAA (https://web3.ncaa.org/aprsearch/gsrsearch). StadiumCapacity was obtained from College Gridirons (https://www.collegegridirons.com/comparisons-by-capacity/). Before proceeding further, it should be noted that avgSRS is a metric calculated by SportsReference and is "a rating system that takes into account average point differential and strength of schedule. The rating is denominated into points above/below average, where zero is average."

Removed Data

Certain schools lacked data or were not initially included. Baylor, BYU, Rice, and SMU are private schools and elected to not provide data related to salary. However, after some research, the salary for Rice and SMU was found through USA Today's football coach database (https://sports.usatoday.com/ncaa/salaries/football/coach), although the Rice data was later removed for a different concern. Baylor and BYU were ultimately dropped from the dataset due to a lack of salary data. If future research is done by someone else it is certainly acceptable to replace the missing values with the mean, median, or mode derived from other coach's salaries, but it would be advised that missing values are not replaced with $0. This could negatively impact any models that are built.

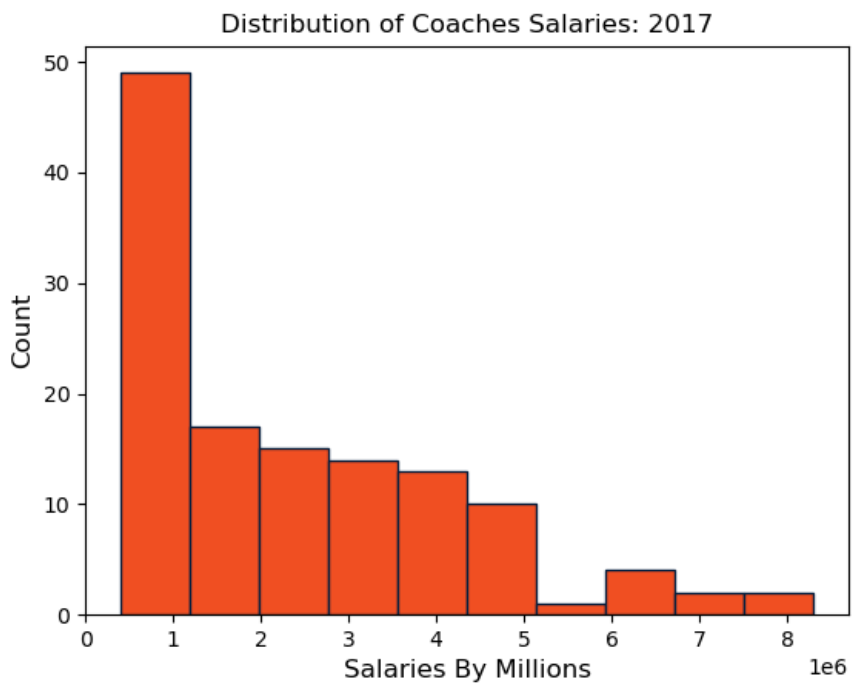
After further investigation, the following schools were removed from the dataset because 2017 marked the end of the current coach and 2018 saw the respective schools hire a first-time head coach. The iSchool did not feel it was fair to include data for a first-year head coach against Babers. However, if Syracuse was looking to construct a contract for a first-year coach then the schools would have been included. The schools that were dropped for this reason included:

- Arizona State
- Central Florida
- Coastal Carolina
- Kent State
- Louisiana-Lafayette
- Mississippi State
- Oregon State
- Rice
- South Alabama
- Tennessee

Prior to any models being built there were several columns that were removed:

- TotalPay
- Bonus
- BonusPaid
- AssistantPay
- Buyout

Total Pay was very similar to SchoolPay, which means that if SchoolPay and Total pay were used in the same model the output could be negatively impacted and it may lead us to not properly identifying if a variable is significant in explaining the outcome we are seeking, which in this case is the coach's salary. Assistant Pay was removed due to it not containing any values. Bonus, BonusPaid and Buyout contained too many missing values for the iSchool to feel comfortable moving forward with the mean, median or mode as replacements.



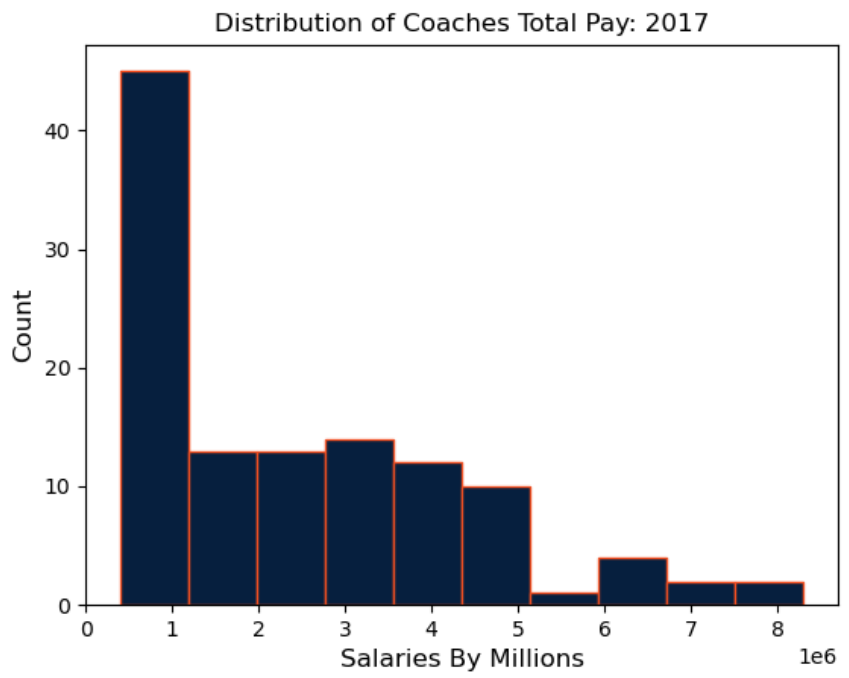


Figure1: Comparison of SchoolPay on the left and Total Pay on the right through a histogram. This shows how similar the variables are, which is why Total Pay was dropped

Altering the Data

Building a model requires certain assumptions to be made, and one of them is that the data is normally distributed, which can be thought of as a bell curve. Figure 1 above clearly shows that SchoolPay does not fit a bell curve. The iSchool attempted to transform this data by taking the square root and the log of the data. The output in Figure 2 represents more of a uniform distribution and less of a normal distribution. Because of this transformations were not made to SchoolPay. No other columns were altered.

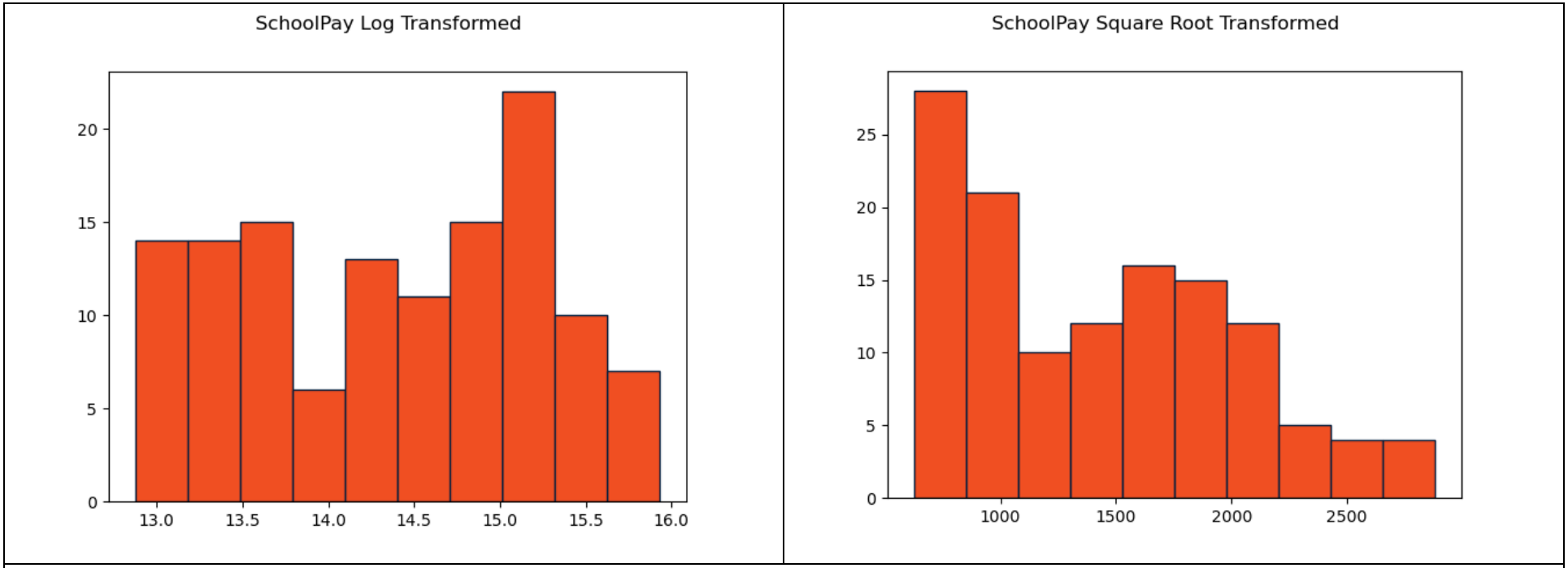


Figure2: SchoolPay was logged and the square root was applied. Neither produced the desired normal distribution

Further Exploration

Prior to building a model it was important to understand the averages of the variables and where coach Babers stands in relation to each:

Variable	Average	Coach Babers
School Pay	2464574	2401206
Career Wins	51	26
Career Losses	35	25
Career Win Percentage	0.536	0.510
AvgSRS	0.978	-0.28
AvgSOS	-0.524	1.04
Seasons Ranked	2.01	0
Stadium Capacity	51,540	49,250
GSR	77.38	85
FGR	62.59	70
School Wins 2017	6.91	4
School Losses 2017	5.81	8

Figure 3: Comparison of the data's average to coach Babers

The iSchool also wanted to examine how certain variables may impact the prediction of a coach's salary. The first variable reviewed was Conference. The image below shows that there is not much variance for each conference except for the SEC, where we see a lot of variance. We also noted that the highest paid coaches come from the traditional conferences, more commonly known as the Power 5 conferences (SEC, Pac12, ACC, Big Ten, and Big 12). This suggests there is a correlation between salary and the conference in which one coaches.

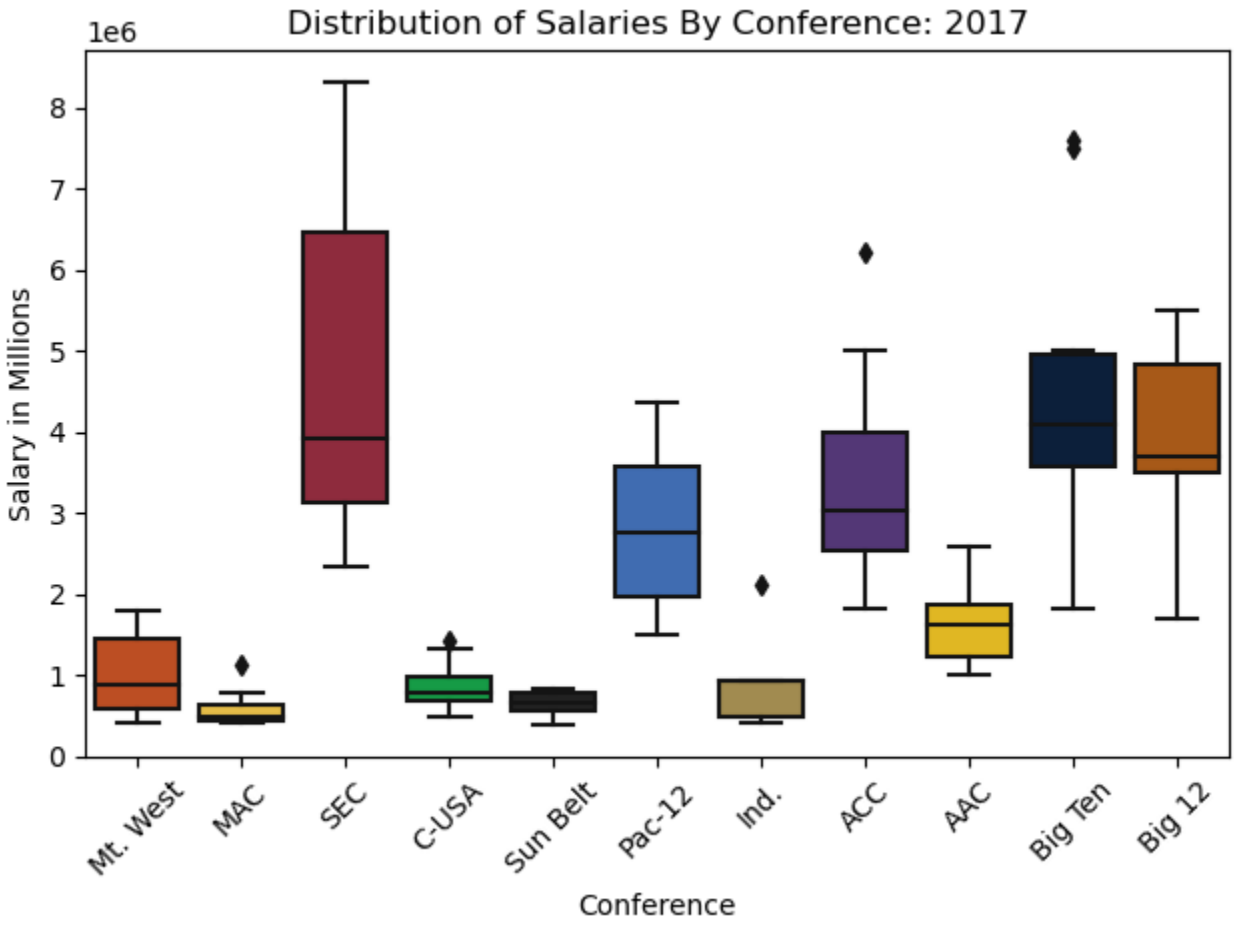


Figure 4: Boxplot of coach's salary base don the conferences they are in.

A relationship also began to emerge when looking at the number of times a coach finished the season with his team ranked in the Top 25 but the relationship does not become clear until a coach achieves about 5 such seasons. It does appear that schools may be overpaying for coaches that do not finish the season ranked in the Top 25.

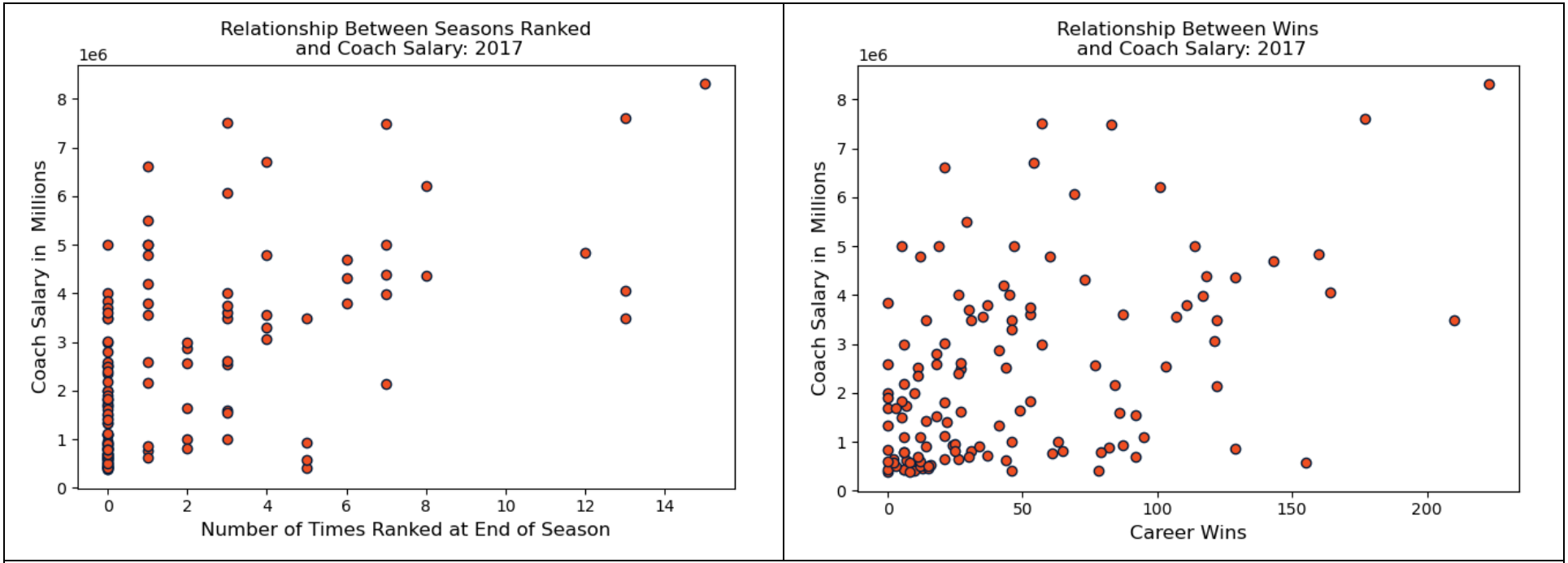


Figure5: The relationship between the number of times a coach finishes the season ranked inside the Top 25 appears to be weaker when compared to the relationship between Career Wins and salary.

The Big East

Before we examine the model, it should be noted that the iSchool did not feel it was appropriate to make a prediction of coach Baber's salary based on the what-if scenario of Syracuse being back in the Big East. The last time Syracuse was in the Big East was 2012 and it was comprised of Louisville, South Florida, Rutgers, Cincinnati, Pittsburgh, Connecticut, and Temple. The teams split into various conferences such as the ACC, AAC, and Big Ten. If we realigned the data to create the most-recent version of the Big East, we would also have to consider the impact that conference realignment had on salaries. In other words, if we used the salary data provided, we would need to figure out how much each team's salary was positively or negatively impacted by leaving the Big East.

The Models and Predictions

There was a total of eight different models that were built. Determining the best can mean different things depending on the person who is viewing and interpreting it. For the sake of this report the iSchool will call out some definitions of a good model. Some may only look at the R-squared number, which helps us understand how much of the variability is explained by the variables used in the model. The closer the score is to absolute value of 1 then the better the model. Others may look deeper into the model to identify variables that are not statistically significant even if the model's R-squared value is close to absolute value 1. The iSchool decided to select a model that had an R-squared score of 0.947 and also contained statistically significant variables with the exception of specific conferences. The iSchool wanted to maintain all conferences in the data because removing a conference may inflate or degrade the values of other conferences, one of which could

be the ACC, and the iSchool felt it was important to understand how much conference affiliation impacts salaries, especially the conference in which Syracuse particpates. This is especially important as there is always a chance that additional conference realignment could happen and may help Syracuse understand the impact a move could have on the coach's salary demands.

The output below provides us with some key insights and can help us address some of the key questions mentioned above:

OLS Regression Results

Dep. Variable:	SchoolPay	R-squared (uncentered):	0.947
Model:	OLS	Adj. R-squared (uncentered):	0.940
Method:	Least Squares	F-statistic:	140.4
Date:	Sat, 24 Apr 2021	Prob (F-statistic):	2.82e-59
Time:	21:52:08	Log-Likelihood:	-1729.2
No. Observations:	116	AIC:	3484.
Df Residuals:	103	BIC:	3520.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CareerWinPct	1.36e+06	4.99e+05	2.722	0.008	3.69e+05	2.35e+06
SeasonsRanked	1.357e+05	2.69e+04	5.038	0.000	8.23e+04	1.89e+05
StadiumCapacity	22.6167	5.009	4.515	0.000	12.682	32.551
ACC	8.45e+05	3.07e+05	2.749	0.007	2.35e+05	1.45e+06
Big 12	1.247e+06	3.46e+05	3.603	0.000	5.6e+05	1.93e+06
Big Ten	1.488e+06	3.3e+05	4.513	0.000	8.34e+05	2.14e+06
C-USA	-7.592e+05	2.8e+05	-2.712	0.008	-1.31e+06	-2.04e+05
Ind.	-6.187e+05	3.73e+05	-1.659	0.100	-1.36e+06	1.21e+05
MAC	-7.822e+05	2.82e+05	-2.776	0.007	-1.34e+06	-2.23e+05
Mt. West	-6.325e+05	2.78e+05	-2.277	0.025	-1.18e+06	-8.16e+04
Pac-12	2.992e+05	3.35e+05	0.894	0.373	-3.64e+05	9.63e+05
SEC	1.924e+06	3.6e+05	5.346	0.000	1.21e+06	2.64e+06
Sun Belt	-6.182e+05	3.4e+05	-1.818	0.072	-1.29e+06	5.61e+04

Figure6: Output of best model

Starting with R-Squared in the top right we see that 94.7% of the variability in the data is accounted for by the variables in the model. This was not the highest R-Squared score achieved but it was the highest score when also considering if variables, outside of the conferences, were significant. The significance of each variable can be found in the P>|t| column, which is the same as a p-value. Any value less than 0.05 is considered significant and can be verified by looking at the

two columns immediately to the right. If the range for those two numbers does not cross zero, then we are given another indication that the variable in the model is significant.

The coef column allows us to make predictions by leveraging the data points we have available. We can do this by multiplying the value of each variable for coach Babers and the coefficient value, and then adding the totals. In the case of each conference, we only need to add the coefficient for ACC and set all others to zero. This will then allow us to make a prediction for his salary if Syracuse was in the Big Ten.

Based on the output below we can see that the recommended salary for coach Babers is $2,652,205.81. His success at Bowling Green contributed to this recommendation as his winning percentage at Syracuse has fallen over the last two years. It would have been interesting to conduct this study prior to coach Babers first season to understand if schools overpay for an up-and-coming coach, and if so, by how much? We are also able to see that a move to the Big Ten would lead to a salary over $3 million.

Prediction for Syracuse Coach			
CareerWinPct	1360000.0	0.509804	693333.3333
SeasonsRanked	135700.0	0	0
StadiumCapacity	22.6	49250	1113872.475
ACC	845000.0	1	845000
Big 12	1247000.0	0	0
Big Ten	1488000.0	0	0
C-USA	-759200.0	0	0
Ind.	-618700.0	0	0
MAC	-782200.0	0	0
Mt. West	-632500.0	0	0
Pac-12	299200.0	0	0
SEC	1924000.0	0	0
Sun Belt	-618200.0	0	0
Predicted Salary			$ 2,652,205.81
Actual			$ 2,401,206.00
Difference			$ 250,999.81

Prediction if Syracuse played in the Big Ten			
CareerWinPct	1360000.0	0.509803922	693333.3333
SeasonsRanked	135700.0	0	0
StadiumCapacity	22.6	49250	1113872.475
ACC	845000.0	0	0
Big 12	1247000.0	0	0
Big Ten	1488000.0	1	1488000
C-USA	-759200.0	0	0
Ind.	-618700.0	0	0
MAC	-782200.0	0	0
Mt. West	-632500.0	0	0
Pac-12	299200.0	0	0
SEC	1924000.0	0	0
Sun Belt	-618200.0	0	0
Predicted Salary			$ 3,295,205.81
			$ 2,401,206.00
			$ 893,999.81

Figure 7: Predictions versus actuals

One variable that is particularly interesting was Graduation Success Rate. While the importance of graduation is often talked about, the iSchool was able to determine that it had very little importance in any model. In fact, when the iSchool added it into the model above it produced p-value of 0.408, which is considered high. It also produced a coefficient of -3468.9523, which means the coach's salary drops by about $3,500 for each one point increase in GSR.

An argument can be made for which variable had the biggest impact on the model and additional models going forward. From a pure numbers perspective it is career win percentage, but there is little adjustment that a coach will see from one year to the next considering there are about 12 games per season. As an example if coach Babers went 12-0 next year his predicted salary would increase by $109,289.62. But if he were to finish the season ranked inside the Top 25 his predicted salary would go up by $135,700. Conference clearly plays a major factor as well but this is something the coach has very little control over. Based on those considerations the iSchool would say the biggest factor in the current model is career winning percentage but going forward, coach Babers can see the biggest increase from a prediction occur as a result of finishing the season ranked in the Top 25.