Predicting NCAA Softball Bid - Report

Thomas Sheehy and George Hunt

December 2022

1 Introduction

Coaches and managers have always had what used to be known as an instinct for determining the immeasurables in sports. Over the past few decades, significant strides have been made in sports statistics and data analytics which help quantify those immeasurable intuitions. There has been an increasing desire to compile, utilize, and interpret large sources of information. That being said, NCAA Softball is one of the few sports which lacks the accessibility of compiled data due to the excessive time constraint of collecting it.

2 Abstract

Our group wanted to create an original project by taking the time to compile our own data. This data can be analyzed by a model which makes it presentable and understandable. This model would take inputs, and then output the probability of NCAA softball teams receiving a bid in a given year. We will display the relationship between variables through plots and correlation matrices to determine the importance of each of them in receiving a playoff bid. These representations can be used to advise sports professionals to make more informed decisions based on their already existing instincts.

3 Objective and Limitations

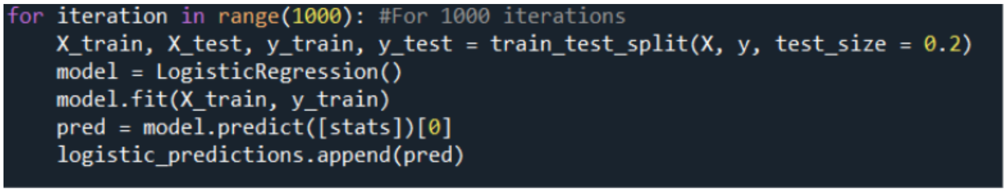
Our original goal had a different focus than our current goal. We originally intended to create a program that would take inputs, and then output the best possible schedule for an NCAA softball team to maximize the possibility of a bid. The primary variable in this model would be the schedule for each team. This would allow the model to associate many different schedules with whether or not that schedule received a bid. That schedule would include every team that each team has played, and then all of the box scores and statistics for each game. In theory, this idea sounded great, but in practice, we found that it would be extremely difficult to implement. In most popular professional sports, box scores and game-by-game data are widely available, but this wasn't the case for NCAA softball. The limitations of the proposed model include a lack of time and a lack of information. The scheduling data that we desired was either non-existent or found on hundreds of different websites which would take weeks of tedious work to compile. Due to these complexities in the "schedule" variable, we planned on finding a way to narrow the scope of the project. One way we thought of was to quantify a team's schedule. This could be done by assigning each schedule a value from 0.0 to 1.0. After brainstorming many versions of quantification, we decided that strength of schedule would be the simplest and most efficient. Only to find out in a few minutes of research, that there was absolutely not a single database that contained the strength of schedule of NCAA softball teams, which was beyond surprising. The complications of the lacking data led our project into a different direction. We wanted our project to still be used in the application of NCAA softball which could provide use to softball coaches and experts, so we created a new model that would take inputs, and then output the probability of NCAA softball teams receiving a bid in a given year. We would also display the correlation between variables to determine the importance of each of them in receiving a bid. Even though the path of deriving information is different, the intention of finding usefulness in data is still the same.

4 Determining Model

The first step in creating a machine learning model is determining the type to use. We narrowed down the type of machine learning to supervised learning since supervised learning is used when the outputs of our data are known, which are bids. More specifically, we found that classification would be best since our output variable (bid) is binary, either a 0 or 1. Based on the format of this program, we hypothesized that a logistic regression model would best fit the data. This is due to the nature of logistic models, which output a number between 0 and 1. To verify this claim, we used other types of models by training and testing them with the same data, and then compared their accuracy scores. This process consisted of using our existing code and simply replacing the model type for each test. We confirmed our hypothesis by discovering that logistic regression had by far the highest accuracy score out of any model we tested.

5 User Inputted Stats

A unique feature of our code is the ability to predict the probability of a specific (or arbitrary) team making the NCAA Softball Playoffs, after inputting the team's data. The input data for each variable was an average of the team's previous three years, in this case 2016-2018. Our program first creates a logistic regression model, which is trained and tested on randomly split samples from the data. With that, it will determine whether or not an inputted team's stats will result in a bid for a given year, which in this case was 2019. It then iterates through the code 1000 times and takes the average of the 1000 iterations to determine the percentage of iterations that this team received a bid.

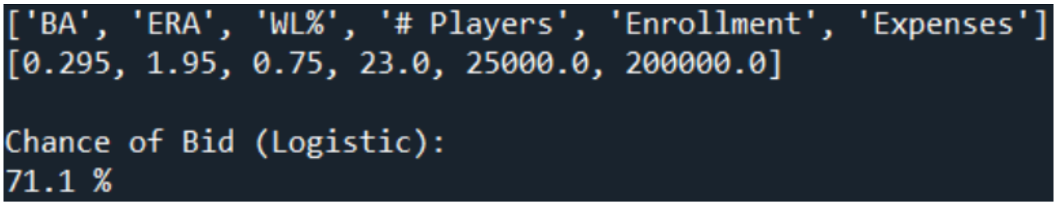


```
for iteration in range(1000): #For 1000 iterations
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)
    model = LogisticRegression()
    model.fit(X_train, y_train)
    pred = model.predict([stats])[0]
    logistic_predictions.append(pred)
```

Figure 1: User Inputted Stats - Code

An example of the statistics that could be inputted are the following: AVG = 0.295, ERA = 1.95, W/L% = 0.75, Number of Players = 23, Enrollment = 25,000, Expenses =

200,000. Based on these specific statistics that we put into the code, the program will output the probability of that team getting a bid. In this case, the program outputted that the chances of this team receiving a bid is 71.1% as shown in the code below.

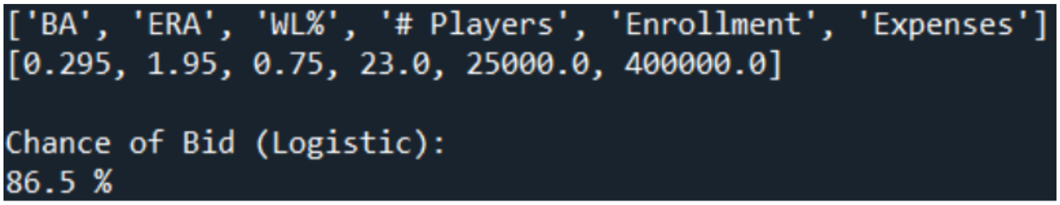


```
['BA', 'ERA', 'WL%', '# Players', 'Enrollment', 'Expenses']
[0.295, 1.95, 0.75, 23.0, 25000.0, 200000.0]

Chance of Bid (Logistic):
71.1 %
```

Figure 2: Bid Chance Before Expense Increase

This process could be repeated for any set of statistics. For example, a head coach of a softball team could see how much the chance of receiving a bid would increase if they doubled their team's operating expenses.



```
['BA', 'ERA', 'WL%', '# Players', 'Enrollment', 'Expenses']
[0.295, 1.95, 0.75, 23.0, 25000.0, 400000.0]

Chance of Bid (Logistic):
86.5 %
```

Figure 3: Bid Chance After Expense Increase

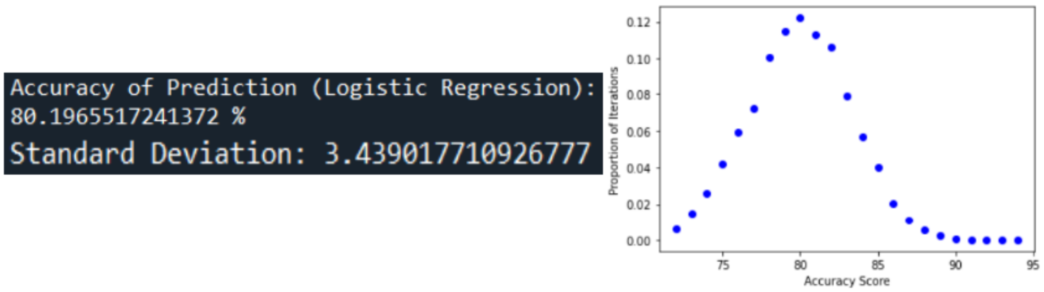
As shown above, every statistic was held constant except for the team's operating expenses. The coach now knows that doubling their operating expenses would increase their odds of a bid by roughly 15%.

6 Model Code and Accuracy

Our first prototype of the model used three key variables for success. These included batting average (AVG), earned run average (ERA), and winning percentage (WL%). The accuracy score of this model was roughly 80.2%, which is much better than just simply randomly guessing a bid or not at 50% accuracy; essentially a coin flip.

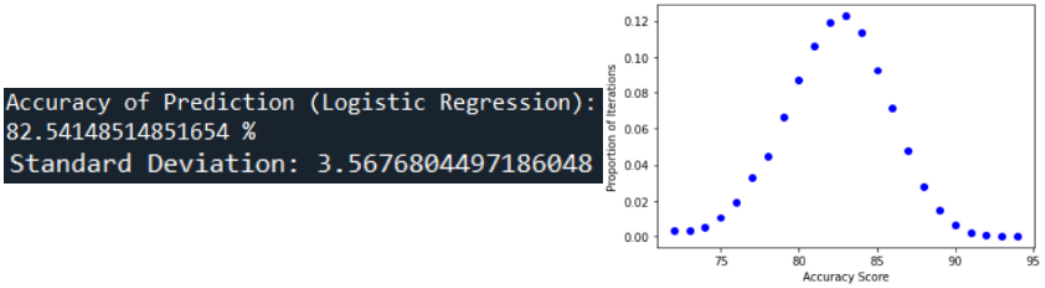
7 Additional Data - Adjusted Model Accuracy

In the hopes to raise our accuracy score, we discovered some more information that may aid the model in predictive ability. These new variables included team operating expenses, undergraduate enrollment, number of players on the team, and 4-year graduation rate. The new model has a predictive accuracy of approximately 82.5%, which was a 2.3 percent increase from the previous model.



```
Accuracy of Prediction (Logistic Regression):
80.1965517241372 %
Standard Deviation: 3.439017710926777
```

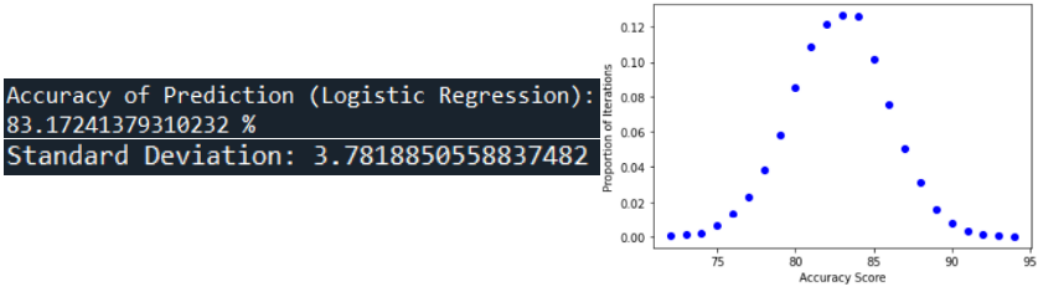
Figure 4: Model Accuracy with 3 Variables (AVG, ERA, WL)



```
Accuracy of Prediction (Logistic Regression):
82.54148514851654 %
 Standard Deviation: 3.5676804497186048
```

Figure 5: Model Accuracy with 7 Variables (AVG, ERA, WL, Expenses, # Players, # Undergraduates, 4-year Grad Rate)

8 Further Model Adjustments



```
Accuracy of Prediction (Logistic Regression):
83.17241379310232 %
Standard Deviation: 3.7818850558837482
```

Figure 6: Model Accuracy with 6 Variables (AVG, ERA, WL, Expenses, # Players, # Undergraduates)

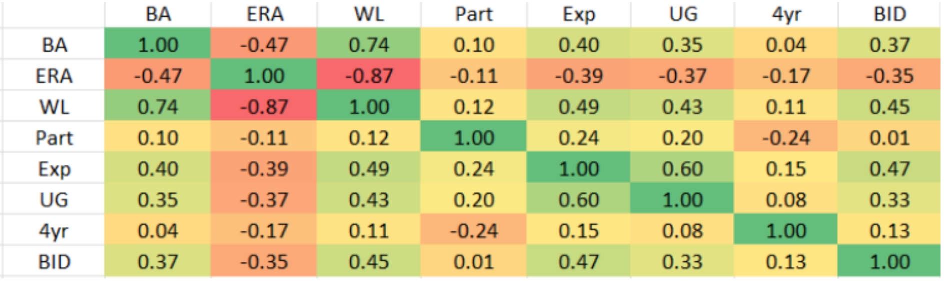
We spent a good chunk of time removing some variables and adding some in from the data we already have, and found that one variable in particular lowers our accuracy score. This was surprising to our group since we thought the logistic model would be “smart” enough to know when a variable was either irrelevant or hurting its accuracy level since it trains itself. However, that was not the case; the new accuracy score of the 6-variable model (4-year graduation rate removed) is roughly 83.2%, in comparison to the 82.5% with the graduation rate. Keep in mind that these accuracy scores are computed over 1000 iterations, so that any error of the average accuracy score is mostly removed.

9 Teams with a Bid

With the updated adjusted model accuracy, we were able to make a list of the teams with the highest chance of receiving a bid, and compare that list to the actual results. While comparing, we found that our top 64 teams closely resembled the 64 teams that received a bid from the NCAA committee in 2019. We were able to accurately predict 50 out of the 64 teams in the 2019 NCAA Softball playoffs, which equates to 78% of the teams. This value is not exactly the 83.2% accuracy due to slight variations in running the model.

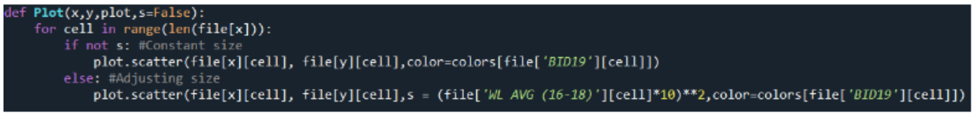
10 Visualizing with Plots and Correlation Matrices

Another way to demonstrate this data is with a correlation matrix. In the correlation matrix that we crafted, we display how correlated each variable is to all of the other variables. The correlation of the variable we care the most about is whether or not each team received a bid. As the data below shows, Win Loss Percentage and Team's Expenses are the two variables that correlated most strongly with receiving a bid. We also see a smaller, but still strong, correlation with batting average and earned run average in determining a bid.



	BA	ERA	WL	Part	Exp	UG	4yr	BID
BA	1.00	-0.47	0.74	0.10	0.40	0.35	0.04	0.37
ERA	-0.47	1.00	-0.87	-0.11	-0.39	-0.37	-0.17	-0.35
WL	0.74	-0.87	1.00	0.12	0.49	0.43	0.11	0.45
Part	0.10	-0.11	0.12	1.00	0.24	0.20	-0.24	0.01
Exp	0.40	-0.39	0.49	0.24	1.00	0.60	0.15	0.47
UG	0.35	-0.37	0.43	0.20	0.60	1.00	0.08	0.33
4yr	0.04	-0.17	0.11	-0.24	0.15	0.08	1.00	0.13
BID	0.37	-0.35	0.45	0.01	0.47	0.33	0.13	1.00

Figure 7: Variable Correlation Matrix



```
def Plot(x,y,plot,s=False):
    for cell in range(len(file[x])):
        if not s: #Constant size
            plot.scatter(file[x][cell], file[y][cell],color=colors[file['BID19'][cell]])
        else: #Adjusting size
            plot.scatter(file[x][cell], file[y][cell],s = (file['WL AVG (16-18)'][cell]*10)**2,color=colors[file['BID19'][cell]])
```

Figure 8: Adaptable Plotting Function

Another interesting portion of our code is the function that we created which allows the user to input any two variables and compare their relationship by creating a scatter plot of the two variables. We also color-coded the points on the graph: green dots = bid, red dots = no bid.

With this function, we were able to compare many sets of variables to help visualize and understand each variable. The four plots are fairly self-explanatory given that their variables are labeled on the axes. It is also apparent that many trends are forming between the two variables on each plot, along with the presence of green and red (bid and no bid). The relationships in these plots can be further explained by the previously shown correlation matrix.

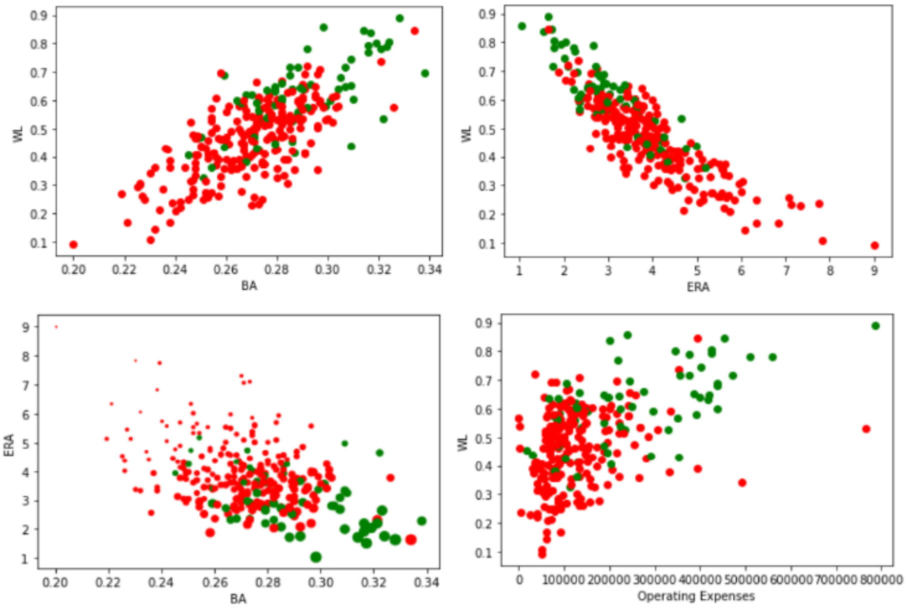


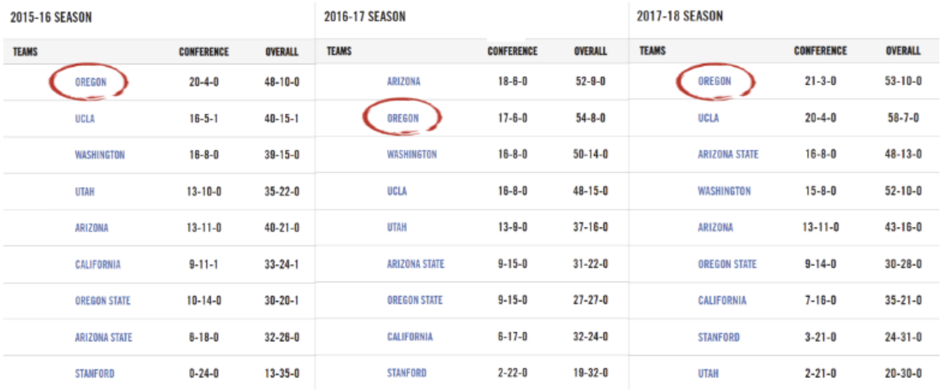
Figure 9: Variable Plots

11 Explaining Outliers In the Data and Plots

The lack of accessible offseason data moves led to a number of outliers. Just because you were good in the past, doesn't mean you will be good in the following years, however, our model doesn't understand that. To eliminate the outliers, we would need to take in other variables that would cover offseason moves, such as the number of incoming transfer players, number of incoming recruited players, number of leaving players due to transfers and graduation, and coaching changes, etc.

12 Oregon Debacle

The Oregon Ducks present a great example of a flaw in our code. As you can see in the image above, the Oregon Ducks were, arguably, the best team in the PAC 12 and in all of NCAA softball from the years 2016-2018, which is when we gathered the data. After putting their numbers into our program, we predicted that Oregon had an 86.7% chance of receiving a bid. What our program does not know is that Oregon's head coach, Mike White, left to go to Texas in the offseason following the 2018 season. After he left, Oregon lost 10 players to the transfer portal, and many incoming recruits decomitted, which left Oregon in a troubling spot for the 2019 season.



2015-16 SEASON

TEAMS	CONFERENCE	OVERALL
OREGON	20-4-0	48-10-0
UCLA	16-5-1	40-15-1
WASHINGTON	16-8-0	39-15-0
UTAH	13-10-0	35-22-0
ARIZONA	13-11-0	40-21-0
CALIFORNIA	9-11-1	33-24-1
OREGON STATE	10-14-0	30-20-1
ARIZONA STATE	6-18-0	32-26-0
STANFORD	0-24-0	13-35-0

2016-17 SEASON

TEAMS	CONFERENCE	OVERALL
ARIZONA	18-6-0	52-9-0
OREGON	17-6-0	54-8-0
WASHINGTON	16-8-0	50-14-0
UCLA	16-8-0	48-15-0
UTAH	13-9-0	37-16-0
ARIZONA STATE	9-15-0	31-22-0
OREGON STATE	9-15-0	27-27-0
CALIFORNIA	6-17-0	32-24-0
STANFORD	2-22-0	19-32-0

2017-18 SEASON

TEAMS	CONFERENCE	OVERALL
OREGON	21-3-0	53-10-0
UCLA	20-4-0	58-7-0
ARIZONA STATE	16-8-0	48-13-0
WASHINGTON	15-8-0	52-10-0
ARIZONA	13-11-0	43-16-0
OREGON STATE	9-14-0	30-28-0
CALIFORNIA	7-16-0	35-21-0
STANFORD	3-21-0	24-31-0
UTAH	2-21-0	20-30-0

Figure 10: Oregon Standings 2016-2018

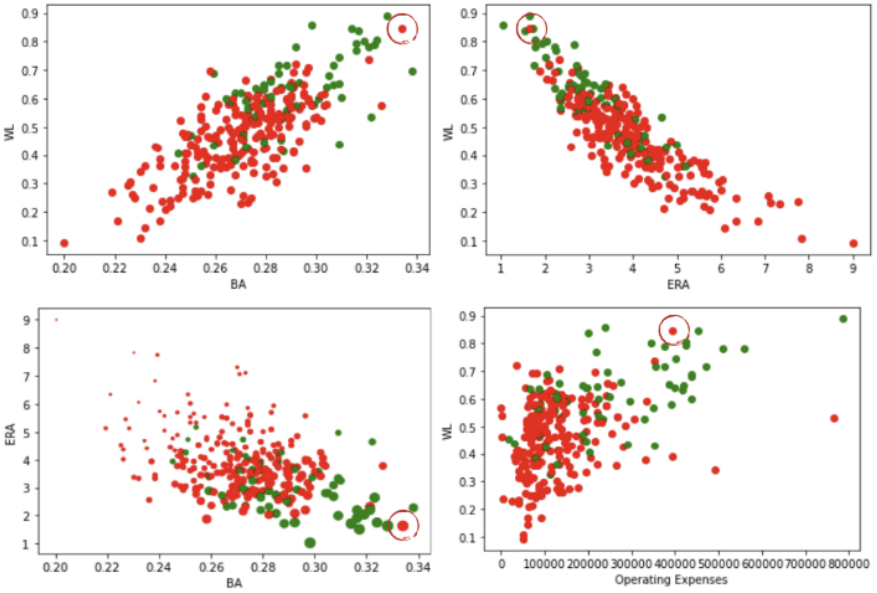
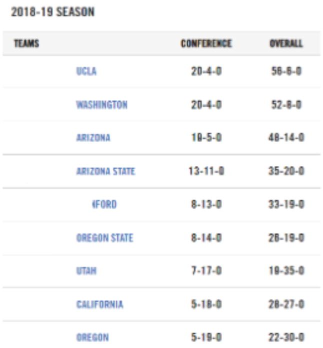


Figure 11: Variable Plots with Oregon Circled

Oregon finished last in their conference after being so good for 3 consecutive years. This led to disparities in our program.

In the graphs, which you have already seen, we have circled out Oregon. They are one of the only teams to not receive a bid given their high statistics. They were on pace to once again be one of the best teams in NCAA softball in 2019 according to our program, but it failed to predict that due to the flaws we discussed.



2018-19 SEASON

TEAMS	CONFERENCE	OVERALL
UCLA	20-4-0	56-6-0
WASHINGTON	20-4-0	52-8-0
ARIZONA	18-5-0	48-14-0
ARIZONA STATE	13-11-0	35-20-0
FORD	8-13-0	33-19-0
OREGON STATE	8-14-0	28-19-0
UTAH	7-17-0	18-35-0
CALIFORNIA	5-18-0	28-27-0
OREGON	5-19-0	22-30-0

Figure 12: Oregon Standings 2019

13 Improved Model - Current Season Data

To account for outliers like Oregon, we implemented current season statistics into the model. This would improve the accuracy of bid predictions as the season progresses by adding the current season record in as a variable. The model accuracy has now improved from 83.2% to 90%. An example of how Oregon's bid probabilities would have changed over the 2019 season is shown.

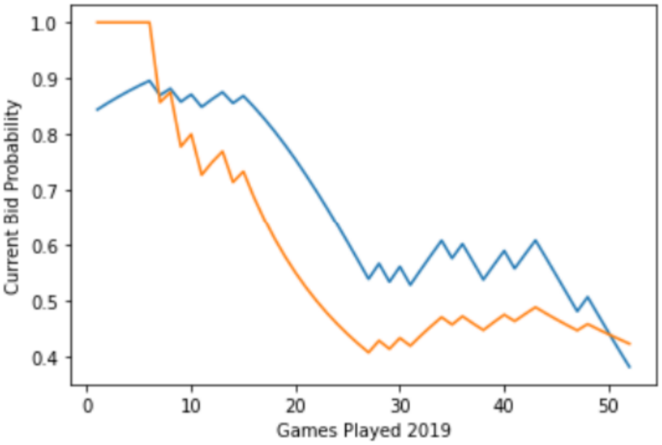


Figure 13: Oregon 2019 Model-Predicted Bid Probability (Blue) — Oregon 2019 Win Percentage (Orange)

14 Conclusion

Even with the flaws, the program we created analyzed the compiled data and was able to efficiently present the data in an organized fashion. We were able to create a model that determined a team receiving a bid at 83% accuracy. Furthermore, we increased that to 90%

when taking into account current season data. Our program showed the importance of each inputted statistic through a graph and correlation matrix, and it was able to take user input to predict their odds of receiving a bid.

15 Applications

We envision coaches, general managers, athletic directors, and even gamblers to be able to use this code to their benefit. Coaches and general managers have the opportunity to view their strengths and weaknesses through the program so they can adapt and be ready for the season. Athletic directors could now look at the projection of the upcoming season to see how their team will do, and even see the effects of spending more money on their program. Sports gamblers could look at the data to put money on the teams most likely to make the NCAA Softball playoffs. Slight adjustments to the code can allow for this program to be applied to most other sports, which displays just how universal our program is. The largest application of our program is the ability to operate user-inputted data. Coaches can input their situations and alter variables to determine the effect of potential adjustments to be made to the team.

16 Future Topics

A team like Oregon shows the perfect example of why the lack of offseason statistics led to flaws in our code. To address these flaws we would need to find easily accessible data on the following variables: team's number of outgoing players, team's number of incoming transfers, team's number of incoming recruits, team's coaching changes, team's number of conference titles won, team's strength of schedule. These statistics were not available for NCAA softball specifically, so once they are, we should be able to utilize them to improve our program. We would hope to increase our accuracy score and maybe find new ways to use the program such as determining an optimal schedule, which was the original goal of our project. We hope to develop a similar model for many other sports that have more data widely available, which should make for a more accurate model.

17 Acknowledgements

We would like to take the time to show our appreciation to former head softball coach Jacqueline Joseph for her interest in this project and the support of us. Additionally, we would like to thank Emilio Esposito for his guidance throughout this project. We would like to thank Nathan Mitchell for his assistance with the original state of the project. Finally, we would like to thank Dr. Albert Cohen for his great mentorship throughout the directed study of the whole project.

18 Sources Cited

"297 NCAA D1 Softball Colleges." Do It Yourself College Rankings, 5 Jan. 2022, https://www.diycollegerankings.com/ ncaa-d1-softball-colleges/6740/ #: :text=NCAA%20D1%20Softball %20Colleges%20%20%20%20Name, %20%206%2C031%20%20%2020%20more%20rows%20.

"NCAA Statistics." NCAA Statistics, http://stats.ncaa.org/rankings/change_sport_year_div.

NCAA.com. "2019 Division I Softball Official Bracket." 2019 Division I Softball Official Bracket — NCAA.com, NCAA.com, 25 Apr. 2022, https://www.ncaa.com/brackets/softball/d1/2019.

Nelson, Bruce. "The Death of a Program? Softball Meltdown at Oregon ..." FishDuck, 24 Jan. 2019, https://fishduck.com/2019/01/the-death-of-a-program-softball-meltdown-at-oregon/.