5-2022

Forecasting Razorback Baseball Game Outcomes

Austin Raabe
University of Arkansas, Fayetteville

Forecasting Razorback Baseball Game Outcomes

by

William "Austin" Raabe

Dr. Ron Freeze

An Honors Thesis in partial fulfillment of the requirements for the degree Bachelor of Science in Business Administration in Accounting and Information Systems.

Sam M. Walton College of Business
University of Arkansas
Fayetteville, Arkansas

May 11, 2022

Table of Contents

Why Razorback Baseball?

Despite the disappointing end to the 2021 Arkansas Razorback baseball year, the team's success provided hog fans something to look forward to next season. While they will be without the 2021 Golden Spikes Award winner, Kevin Kopps, and four All-SEC team selections, the 2022 roster has promising new and returning talent. With fifty percent of the players who played significant time last year coming back (minimum ten hits or ten innings pitched), the arrival of several impact transfers from major conferences, and a recruiting class ranked in the top five according to Perfect Game, there is reason to believe that last season can be used as a good predictor of what will happen in 2022.

As the sports analytics industry continues to grow, with increased use in college sports like baseball, there is a potential benefit for all parties involved. Being a fan of baseball, and a die-hard Arkansas Razorback fan, I decided to create a score predictor model that determined the runs scored for a particular Arkansas baseball game. The model would include different variables and statistics that deemed to be significant to the team's score output. To provide a complete score prediction and more insights to the project, I created another equation that determines the runs scored by the Razorback's upcoming opponent.

One goal of this project is to provide the Arkansas coaching staff and players with influential statistics to focus on as they scout their upcoming opponents. Also, it will allow them to key in on their offensive and defensive strengths, and make next-game adjustments. Another goal of this project is to give sports fans, and more specifically baseball fans, an accurate scoring model that they can use for potential sports betting. From a fan perspective, sports analytics has seen a significant growth in popularity. For example, just like when you play fantasy football and make roster decisions based on projected fantasy points for a certain player, this scoring model would give sports fans the opportunity to make more accurate bets on the number of runs scored, the winner of the game, or the margin of victory. The last goal of this project is to give Arkansas fans, like myself, a thorough projection of what to expect from the Razorbacks, and their next opponent, going into any given gameday or weekend series.

Box Score Stats and More…

With the goal of this project being to create the most accurate scoring models for the upcoming Razorbacks baseball games, I gathered all of the stats you would find in the post-game box score from each of the sixty-three games played in the 2021 season. As seen explained in more detail in the Data Dictionary, I collected over fifteen batting and fielding statistics from the Razorbacks and each of their opponents. Also, I included other variables I thought would be important such as the team's ranking, the day of the week the game was happening, the date and location, and if the game were against a Power 5 school or SEC opponent.

Making the Model Fully Predictive

To ensure the scoring models were completely predictive, there were key steps taken to manipulate the data in order to make that possible. First, I took the Razorback baseball team box score stats, and changed them to reflect the average of the previous ten games played rather than the actual stats the team incurred in the game. For example, Arkansas played Oklahoma last season, and going into the game their average at-bats were 32.2. So, this means that for the ten games played prior to the Oklahoma game, the Razorbacks had an average of 32.2 at-bats in those games. This process repeated for each of the box score stats, and the only numbers gathered that weren't previous ten game averages was the actual score of the games and the day of the week in which the game happened.

For the opponents stats, instead of gathering the box score stats incurred from the games that happened, I also replicated the same format as done for the Razorbacks. For each of the opponents that Arkansas faced in the 2021 season, I took their previous ten game averages for all of the same box score stats. For example, Arkansas played against Florida last season, and going into the game their average hits allowed were 9.8. So, this means that for the ten games played prior to our matchup, the Gators had an average of 9.8 hits allowed in those games.

Also, there were general assumptions that made when gathering the data for this project. Firstly, I chose the ten game sample size for the previous game averages to ensure the models would hold the most statistical relevance. When predicting the games for this 2022 season, I decided to only predict the scores and outcomes for the first twenty-two games of the season. Next, the entire dataset collected is from strictly the 2021 season, except for the first ten games of the 2021 season which includes previous ten game averages from the last ten games of the 2020 season. Lastly, I assumed with players graduating or leaving the program, and new to Arkansas players coming in, that the 2021 season won't be a perfect representation of what will happen the upcoming year. However, with a head coach that has been around for years and key players returning or transferring in, the 2021 data should still hold significance and be a starting point for what to expect in the 2022 season.

Comparing Each of the Models

With the goal of the project being to create the most accurate score predictor for Arkansas baseball games in the upcoming 2022 season, I developed six linear regression models for both Arkansas, and their opponents. Each model contained both categorical and numerical variables, and I primarily focused on the R-Squared and Root MSE when evaluating each of the equations. Firstly, I began by creating models for the Razorbacks, so when it came time to make the opponent models, I could use the same variables in both equations. For example, when I used the variable, "Doubles," in the Razorbacks linear regression model, I would use the variable, "Opponent Doubles," in the opponents model. Once I established my strongest linear regression models for both the Razorbacks and the opponents, I created two logistic regression models, one for each, that mirrored those linear regression variables. These models evaluated just the projected outcome of the game, a win or a loss, rather than the score.

Predicting Arkansas Runs

Due to the overhaul of variables and statistics gathered from the 2021 season, I figured it was best to begin the model development process by using three automated variable selection methods. These methods included: forward, backward, and stepwise. For each linear regression model run, regardless of team, the predictor variable was runs scored. Once I established those models, I also developed my own model by inputting all of the variables relevant to the Arkansas score prediction model, and then taking them out until they no longer improved the R-Squared. The variables chosen for the Razorback scoring model included Arkansas batting and opponent pitching stats. For example, statistics like home runs, stolen bases, opponent walks allowed, and opponent pitching strikeouts were used in some of the equations. Also, categorical variables like conference and opponent ranking were inputted into the model. Next, when I took an in-depth look into each of the equations, I noticed that there were variables that did not make as much statistical sense as other variables. For example, a variable such as home runs, which should have a positive coefficient, had a negative coefficient. Then, a variable such as opponent pitching strikeouts, which should have a negative coefficient, had a positive coefficient.

So, for the next model I created, I took only the variables that made statistical sense and inputted them into one equation. Also, I decided to leave the categorical variables in, as they could have a negative or positive coefficient. Lastly, I took the Explained Variables model that I created, and ran a backward regression to see if further condensing the model would improve its statistical evaluation. Below is each of the following models that I created, the amount of variables in the equation, the R-Squared, the Root MSE, and the average deviation from the prediction model and the actual runs scored in the game.

Scoring Model	Variables in Equation	R-Squared	Root MSE	Average Deviation from Actual Runs Scored
Model #1 (Forward)	6	0.3337	4.47	4.93
Model #2 (Backward)	6	0.4013	4.16	7.70
Model #3 (Stepwise)	5	0.3378	4.42	5.20
Model #4 (Self-Created)	12	0.4896	4.20	11.71
Model #5 (Explained Variables)	10	0.2270	4.86	4.02
Model #6 (Backward of #5)	3	0.1934	4.67	4.13

When creating these models I was looking for the highest R-Squared, which is the percentage of the model explained by the variables chosen. I was also looking for the lowest Root MSE, which is the expected deviation from the predicted value and the actual value. For example, in the fourth model, expect +/- 4.2 runs on the predicted scoring output. Even though the R-Squared and Root MSE were best in the first four scoring models, their average deviation from the actual runs scored were the furthest away. In the first and third model, they slightly underperformed as the average deviation was barely over the Root MSE. However, with the second and fourth model, the average deviation was way over the Root MSE. Then, despite the lower R-Squared, and slightly higher Root MSE, the fifth and sixth models proved to be more accurate. In each of the two models, the average deviation from the actual runs scored was less than the Root MSE, and instead did a better job than the other models at predicting Arkansas' runs scored. I decided to choose the fifth model for further evaluation.

Predicting Opponent Runs

After I ran each of the Arkansas scoring models, I used those same variables for each respective equation and inputted them in the equation for the opponents. As displayed previously in the Arkansas linear regressions table, below is each of the following models that I created. For each model I put the amount of variables in the equation, the R-Squared, the Root MSE, and the average deviation from the prediction model and the actual runs scored in the game.

Scoring Model	Variables in Equation	R-Squared	Root MSE	Average Deviation from Actual Runs Scored
Model #1 (Forward)	6	0.0779	3.55	1.83
Model #2 (Backward)	6	0.0850	3.57	1.92
Model #3 (Stepwise)	5	0.0670	3.54	1.97
Model #4 (Self-Created)	12	0.2336	3.46	3.97
Model #5 (Explained Variables)	10	0.1738	3.49	2.25
Model #6 (Backward of #5)	3	0.0331	3.54	1.66

Unlike a couple of the Arkansas score prediction models, the opponent models average deviation from actual runs scored was less than the Root MSE. As seen in the chart above, Model #4 was the only model that had a greater average deviation than Root MSE. Next, another takeaway from the opponent models was that the R-Squared on each of the equations was less than that of the Razorback models. The lowest R-Squared for any of the Arkansas equations was .1934, while the highest of any of the opponent models was .2336. However, the Root MSE, which in this case is the difference between predicted and expected runs, for each of the models was less than any of the Razorback models that I created. Despite less variables than any of the other opponent equations evaluated, the last model I made proved to have the best average deviation. In order to give the most accurate score prediction from both teams, I decided that the variables for both equations should be the same. Due to the amount of variables in the equation, and the accuracy of the deviation from actual runs scored, I decided to stay consistent in my evaluation of both teams and further evaluate my fifth model.

Predicting Arkansas Victories

In order to keep consistency between models, I decided to use the same variables from each equation when creating my logistic regression models. Also, because Models #1-4 proved to be more inaccurate than the final two linear regression models, I only replicated Models #5-6 for the logistic regression equations. For each of the logistic regression models, regardless of team, the predictor variable was game outcome. Below is each of the following models that I created, the amount of variables in the equation, the Misclassification Rate, and the percentage of correctly predicted games.

Scoring Model	Variables in Equation	Misclassification Rate	% of Correctly Predicted Games
Model #1 (Explained Variables)	10	0.2222	77.27%
Model #2 (Backward of #1)	3	0.2063	81.82%

As seen in the table above, both models were around 80% accurate on predicting the Razorbacks game outcomes. This makes sense as they both had misclassification rates around 20% and had an expectation to be right about four out of five times. Also, because I had chosen the "Explained Variables" model for each of the two linear regression equations, I elected to stay with that model again for my logistic regression. This decision was to ensure that there was no discrepancies between the variables used for each of the models. Even though my first logistic regression had a greater misclassification rate and was slightly less accurate predicting games, I don't believe there was a significant difference between the two models. Overall, given its larger number of variables, I think the first of my logistic regression model provides a more accurate representation and gives variable consistency across the linear and logistic regression models that I chose to further evaluate.

Predicting Opponent Losses

Similarly to the logistic regression models for the Razorbacks, I elected to only run two logistic regression models for the opponents. These variables were reflective of the variables used in the Arkansas logistic regression equations. As displayed for the Arkansas linear regressions, below is each of the following models that I created, the amount of variables in the equation, the Misclassification Rate, and the percentage of correctly predicted games.

Scoring Model	Variables in Equation	Misclassification Rate	% of Correctly Predicted Games
Model #1 (Explained Variables)	10	0.1746	81.82%
Model #2 (Backward of #1)	3	0.2063	81.82%

For the opponents, both logistic regression models had misclassification rates the same or lower than the models created for the Razorbacks. Each equation had near 82% accuracy, which made statistical sense once again given each of their misclassification rates. As mentioned for my Arkansas logistic regression model, I decided to choose the first of my two models to keep consistency in the variables that I used for my models that I wanted to further evaluate.

Evaluating the Best Model for the 2022 Season

Now that I have chosen the "Explained Variables" model for each of the logistic and linear regression equations for both teams, it is now time to see how accurately they predicted the scores and outcomes of the first twenty-two games of the 2022 season. While each of the models had the same variables across the board, each one had its own unique predictions, and found itself closer or further away from the others in accurately predicting its desired outcome. Once again, the goal for the linear regression models was to predict the runs scored for Arkansas and its opponents, and the goal for the logistic regression models was to predict the result of the game for both teams.

Arkansas Runs Scored Model

Beginning with the Arkansas linear regression model, this model had a .2270 R-Squared and a 4.86 Root MSE. This means that only 22.7% of the predicted run score was represented by this model, and that the average margin of error between actual runs scored and its prediction was 4.86 runs. While these metrics aren't not very favorable for a model, this linear regression equation far out exceeded its expectations. Even though the Root MSE was 4.86 runs, which is not a desirable number given the small number of runs scored in a baseball game, the model only exceeded that threshold seven out of the twenty-two times. Below is the equation for the Arkansas linear regression, with each of the variables included and the rounded coefficients for each variable.

Arkansas Linear Regression Equation

$$\hat{Y}_{Score} = 9.02 + 2.33x_{Power\_5} + -2.38x_{Average\_Score} + -0.22x_{Batting\_Strikeouts} + 1.02x_{Home\_Runs} + 0.05x_{Opponent\_Earned\_Runs} + 0.34x_{Opponent\_Hits\_Allowed} + -0.26x_{Opponent\_Pitching\_Strikeouts} + 0.39x_{Opponent\_Walks\_Allowed} + 2.68x_{Walks} + -0.14x_{Weekday}$$

In the equation above, there are ten statistically significant variables making up the linear regression model. With runs scored as the desired predictor variable, each input has a positive or negative coefficient that contributes to the output. For example, for every batting strikeout that Arkansas has, their score expects to decrease by 0.22 runs. Another example, for every walk allowed by the opponent, Arkansas' score expects to increase by 0.39 runs.

Date	Location	Opponent	Actual Result	Actual Score	Model #5 Projected Score	Deviation
2/18/2022	Home	Illinois State	Loss	2	11.4	9.4
2/19/2022	Home	Illinois State	Win	5	10.7	5.7
2/20/2022	Home	Illinois State	Win	4	9.5	5.5
2/25/2022	Neutral	Indiana	Win	5	8.3	3.3
2/27/2022	Neutral	Stanford	Loss	0	7.9	7.9
2/27/2022	Neutral	Louisiana	Win	6	10.7	4.7
3/2/2022	Home	Omaha	Win	15	12.9	2.1
3/4/2022	Home	Southeastern	Loss	3	11.7	8.7
3/5/2022	Home	Southeastern	Win	4	14.1	10.1
3/5/2022	Home	Southeastern	Win	11	15.5	4.5
3/10/2022	Home	UIC	Win	12	12.7	0.7
3/10/2022	Home	UIC	Win	5	10.1	5.1
3/12/2022	Home	UIC	Win	10	11.8	1.8
3/13/2022	Home	UIC	Win	10	10.9	0.9
3/15/2022	Home	Grambling State	Win	14	10.3	3.7
3/16/2022	Home	Grambling State	Win	13	8.5	4.5
3/18/2022	Home	Kentucky	Win	6	4.6	1.4
3/19/2022	Home	Kentucky	Win	9	4.6	4.4
3/20/2022	Home	Kentucky	Win	3	3.8	0.8
3/25/2022	Away	Missouri	Win	7	4.4	2.6
3/26/2022	Away	Missouri	Loss	5	4.7	0.3
3/27/2022	Away	Missouri	Win	6	6.5	0.5

Opponents Runs Scored Model

Now looking at the opponents linear regression model, this model had a .1738 R-Squared and a 3.49 Root MSE. This means that only 17.4% of the predicted run score was represented by this model, and that the average margin of error between actual runs scored and its prediction was 3.49 runs. These metrics are once again still unfavorable for a typical model, but just like the last model, this linear regression continues to outperform the expectations. For the opponents equation, the Root MSE was 1.37 runs lower than the Arkansas model, which is closer to a more desirable three run error margin, and only saw the model exceed that number six out of the twenty-two times. Below is the equation for the Opponents linear regression, with each of the variables included and the rounded coefficients for each variable.

Opponent Linear Regression Equation

$$\hat{Y}_{Opponent\_Score} = 10.29 + 0.07x_{Power\_5} + 0.73x_{Opponent\_Average\_Score} + -0.26x_{Opponent\_Batting\_Strikeouts} + -2.18x_{Opponent\_Home\_Runs} + -1.99x_{Earned\_Runs} + 1.30x_{Hits\_Allowed} + -0.57x_{Pitching\_Strikeouts} + 0.06x_{Walks\_Allowed} + -0.92x_{Opponent\_Walks} + 0.24x_{Weekday}$$

In the equation above, there are ten statistically significant variables making up the linear regression model. With opponent runs scored as the desired predictor variable, each input has a positive or negative coefficient that contributes to the output. For example, for every batting strikeout that the opponent has, their score expects to decrease by 0.26 runs. Another example, for every walk allowed by the Razorbacks, the opponent's score expects to increase by 0.06 runs.

Date	Location	Opponent	Actual Result	Actual Score	Model #5 Projected Score	Deviation
2/18/2022	Home	Illinois State	Win	3	5.8	2.8
2/19/2022	Home	Illinois State	Loss	1	6.3	5.3
2/20/2022	Home	Illinois State	Loss	2	6.7	4.7
2/25/2022	Neutral	Indiana	Loss	2	4.5	2.5
2/27/2022	Neutral	Stanford	Win	5	6.6	1.6
2/27/2022	Neutral	Louisiana	Loss	4	8.3	4.3
3/2/2022	Home	Omaha	Loss	3	4.0	1.0
3/4/2022	Home	Southeastern	Win	7	5.6	1.4
3/5/2022	Home	Southeastern	Loss	2	5.5	3.5
3/5/2022	Home	Southeastern	Loss	1	5.5	4.5
3/10/2022	Home	UIC	Loss	4	4.9	0.9
3/10/2022	Home	UIC	Loss	4	3.8	0.2
3/12/2022	Home	UIC	Loss	1	3.1	2.1
3/13/2022	Home	UIC	Loss	8	2.4	5.6
3/15/2022	Home	Grambling State	Loss	1	2.5	1.5
3/16/2022	Home	Grambling State	Loss	3	1.7	1.3
3/18/2022	Home	Kentucky	Loss	2	2.8	0.8
3/19/2022	Home	Kentucky	Loss	3	2.8	0.2
3/20/2022	Home	Kentucky	Loss	1	2.5	1.5
3/25/2022	Away	Missouri	Loss	5	5.1	0.1
3/26/2022	Away	Missouri	Win	7	5.5	1.5
3/27/2022	Away	Missouri	Loss	4	6.0	2.0

Arkansas Game Outcome Model

Now as we transition to evaluating the Arkansas logistic regression model, this model had a .2222 Misclassification Rate and 77.27% accuracy on predicted outcome. This means that when determining the outcome of the game, meaning who wins or loses, this logistic regression model had an error rate of 22%. When putting this model to the test, it is able to accurately predict four in every five games played by the Razorbacks. Given that any team has a fifty-fifty shot in every game played, a correct prediction percentage of 77% well exceeds that desired rate. Also, when you take one minus the misclassification rate, which should be the hypothetical percentage of correct predictions, it nearly equals the actual percentage of correct predictions displayed by the model. Overall, the model was fairly accurate in predicting the outcome of the game, and was only incorrect five out of the twenty-two times. Below is the equation for the Arkansas logistic regression, with each of the variables included and the rounded coefficients for each variable.

Arkansas Logistic Regression Equation

$$\hat{Y}_{Arkansas\_Win} = -1.33 + 1.66x_{Power\_5} + 0.23x_{Average\_Score} + 0.33x_{Batting\_Strikeouts} + -2.38x_{Home\_Runs} + -0.35x_{Opponent\_Earned\_Runs} + 0.50x_{Opponent\_Hits\_Allowed} + 0.21x_{Opponent\_Pitching\_Strikeouts} + -0.45x_{Opponent\_Walks\_Allowed} + 0.23x_{Walks} + -0.47x_{Weekday}$$

In the equation above, there are ten statistically significant variables making up the logistic regression model. With game outcome as the predictor variable, meaning a positive output would result in a win and a negative output a loss, each input has a positive or negative coefficient that contributes to the likelihood of the output. For example, because the walks and average score variables have positive coefficients, an increase in either one of those variables would result in a stronger possibility of an Arkansas win. However, on the other hand, because the weekday and opponent earned runs variables have negative coefficients, an increase in either one of those variables would result in a stronger possibility of an Arkansas loss.

Date	Location	Opponent	Actual Score	Actual Result	Model #1 Projected Outcome
2/18/2022	Home	Illinois State	2	Loss	Win
2/19/2022	Home	Illinois State	5	Win	Win
2/20/2022	Home	Illinois State	4	Win	Win
2/25/2022	Neutral	Indiana	5	Win	Loss
2/27/2022	Neutral	Stanford	0	Loss	Win
2/27/2022	Neutral	Louisiana	6	Win	Win
3/2/2022	Home	Omaha	15	Win	Win
3/4/2022	Home	Southeastern	3	Loss	Win
3/5/2022	Home	Southeastern	4	Win	Win
3/5/2022	Home	Southeastern	11	Win	Win
3/10/2022	Home	UIC	12	Win	Win
3/10/2022	Home	UIC	5	Win	Win
3/12/2022	Home	UIC	10	Win	Win
3/13/2022	Home	UIC	10	Win	Win
3/15/2022	Home	Grambling State	14	Win	Win
3/16/2022	Home	Grambling State	13	Win	Win
3/18/2022	Home	Kentucky	6	Win	Win
3/19/2022	Home	Kentucky	9	Win	Win
3/20/2022	Home	Kentucky	3	Win	Win
3/25/2022	Away	Missouri	7	Win	Win
3/26/2022	Away	Missouri	5	Loss	Win
3/27/2022	Away	Missouri	6	Win	Win

Opponent Game Outcome Model

Lastly, as we take a look at the opponents logistic regression model, this model had a .1746 Misclassification Rate and 81.82% accuracy on predicted outcome. This means that when determining the outcome of the game, this logistic regression model had an error rate of 17%. When putting this model to the test, it is able to accurately predict every four in five games played by Arkansas opponents. Similarly to the Razorback logistic regression model, being able to exceed the desired fifty-fifty winning opportunity that is available to either team is a nice statistical output to see, especially given the correct prediction percentage is near 82%. Also, just like the Arkansas regression model, when you take one minus the misclassification rate, it nearly equals the actual percentage of correct predictions displayed by the model. Overall, the model was fairly accurate in predicting the outcome of the game, and was only incorrect four out of the twenty-two times.

Opponent Logistic Regression Equation

$$\hat{Y}_{Opponent\_Win} = -6.74 + -2.61x_{Power\_5} + -0.13x_{Opponent\_Average\_Score} + 0.08x_{Opponent\_Batting\_Strikeouts} + 1.19x_{Opponent\_Home\_Runs} + -2.58x_{Earned\_Runs} + 1.23x_{Hits\_Allowed} + -0.29x_{Pitching\_Strikeouts} + 2.12x_{Walks\_Allowed} + -0.13x_{Opponent\_Walks} + 0.23x_{Weekday}$$

In the equation above, there are ten statistically significant variables making up the logistic regression model. With opponent game outcome as the predictor variable, each input has a positive or negative coefficient that contributes to the likelihood of the output. For example, because the opponent home runs and hits allowed variables have positive coefficients, an

increase in either one of those variables would result in a stronger possibility of an opponent win. However, on the other hand, because the pitching strikeouts and earned runs variables have negative coefficients, an increase in either one of those variables would result in a stronger possibility of an opponent loss.

Date	Location	Opponent	Actual Score	Actual Result	Model #1 Projected Outcome
2/18/2022	Home	Illinois State	3	Win	Loss
2/19/2022	Home	Illinois State	1	Loss	Loss
2/20/2022	Home	Illinois State	2	Loss	Loss
2/25/2022	Neutral	Indiana	2	Loss	Win
2/27/2022	Neutral	Stanford	5	Win	Win
2/27/2022	Neutral	Louisiana	4	Loss	Loss
3/2/2022	Home	Omaha	3	Loss	Loss
3/4/2022	Home	Southeastern	7	Win	Loss
3/5/2022	Home	Southeastern	2	Loss	Loss
3/5/2022	Home	Southeastern	1	Loss	Loss
3/10/2022	Home	UIC	4	Loss	Loss
3/10/2022	Home	UIC	4	Loss	Loss
3/12/2022	Home	UIC	1	Loss	Loss
3/13/2022	Home	UIC	8	Loss	Loss
3/15/2022	Home	Grambling State	1	Loss	Loss
3/16/2022	Home	Grambling State	3	Loss	Loss
3/18/2022	Home	Kentucky	2	Loss	Loss
3/19/2022	Home	Kentucky	3	Loss	Loss
3/20/2022	Home	Kentucky	1	Loss	Loss
3/25/2022	Away	Missouri	5	Loss	Loss
3/26/2022	Away	Missouri	7	Win	Loss
3/27/2022	Away	Missouri	4	Loss	Loss

Key Insights from Assessing the Models

After evaluating each of the linear and logistic regression models for both the Arkansas Razorbacks and their opponents, it is safe to say that the typical statistical measures for these models did not tell the entire story. If we were to go specifically off of R-Squared or Root MSE for the Arkansas linear regression equations, it would have been the original models produced by the automated variable selection methods, like backward or stepwise, that would have been the best to use for predicting runs scored. However, those models were significantly more inaccurate than the last two models created, for example the “Explained Variables,” which was Model #5 for the Arkansas Razorbacks linear regression models. This model was the most accurate out of that group, and I decided to use those variables throughout the rest of the equation selection process. Furthermore, the linear regression used for the opponents was an even better predictor than the Arkansas model, and was within two runs on thirteen of its twenty-two games predicted.

When moving on to the logistic regression models, the same variables again held up in predicting the outcome of the game for both teams. Unlike the statistical measures used when evaluating the linear regression models, the misclassification rate for the logistic regression models was a way better indicator of the success of the equations. In each logistic regression equation, for both Arkansas and its opponents, there was over a 77% accuracy in predicting wins and losses. This was the exact same as the expected percentage of correct predictions displayed by the model, which you can calculate by taking one minus the misclassification rate. Overall, for the linear regression models, the story played out exactly as written. There was not as many surprises given the statistical measures evaluated for the logistic regression models, contrary to when looking at the linear regression models. In fact, the biggest surprise was not how accurate

the predictions were, but instead, how accurate the model was in predicting the accuracy of the game outcome predictions.

Lastly, after evaluating all of the models created throughout the process, and narrowing it down to four, more accurate and in-depth equations, there are a few takeaways that can be stated given my findings. While the 2022 Arkansas baseball season sample size is not large enough to fully evaluate the accuracy of these models, the early success of the models could be an indicator of how the rest of the season's predictions could play out. One takeaway I gathered is that there is definitely a practical use for these models within the Arkansas coaching staff. Whether it be the linear or logistic regression model, for either Arkansas or even the opponent, being able to see the coefficients of the variables used in the equations allows for the baseball team to see which statistic(s) they should focus on more than the others. Another takeaway that can I made is that each of the models holds up for potential sports betting. While the R-Squared and Root MSE did not look promising for potential run predictions, the accuracy of the linear regression models was a surprising outcome. Also, both of the logistic regression models proved to be the most accurate of the models, and each of them could be beneficial to potential bettors betting on Arkansas Razorback baseball games. Finally, as an Arkansas Razorback fan, this model provides great promise for the 2022 season, as we expect to have another strong regular season, and hopes for a deeper run in the NCAA Tournament.

Appendix

Data Dictionary

Name	*Type*	*Example*	*Description*
At-Bats	Decimal	33.1	Previous ten game average of at-bats by Arkansas
Average Score	Decimal	6.8	Previous ten game average of runs scored by Arkansas
Batting Strikeouts	Decimal	5.9	Previous ten game average of strikeouts by Arkansas batters
Conference	Boolean	Yes	Are they an SEC opponent? "Yes" or "No"
Date	Date	03/26/2022	Day on which the game is happening
Doubles	Decimal	1.8	Previous ten game average of doubles by Arkansas batters
Earned Runs	Decimal	4.6	Previous ten game average of earned runs by Arkansas pitchers
Errors	Decimal	0.8	Previous ten game average of errors by Arkansas fielders
Hits	Decimal	7.8	Previous ten game average of hits by Arkansas batters
Hits Allowed	Decimal	6.3	Previous ten game average of hits allowed by Arkansas pitchers
Home Runs	Decimal	1.5	Previous ten game average of home runs by Arkansas batters
Opponent	String	Kentucky	Team facing the Arkansas Razorbacks
Opponent At-Bats	Decimal	30.5	Previous ten game average of at-bats by Opponents
Opponent Average Score	Decimal	4.7	Previous ten game average of runs scored by Opponents
Opponent Batting Strikeouts	Decimal	8.8	Previous ten game average of strikeouts by Opponent batters
Opponent Doubles	Decimal	1.7	Previous ten game average of doubles by Opponent batters
Opponent Earned Runs	Decimal	4.9	Previous ten game average of earned runs by Opponent pitchers
Opponent Errors	Decimal	1.2	Previous ten game average of errors by Opponent fielders
Opponent Hits	Decimal	5.9	Previous ten game average of hits by Opponent batters

Opponent Hits Allowed	Decimal	7.1	Previous ten game average of hits allowed by Opponent batters
Opponent Home Runs	Decimal	1.5	Previous ten game average of home runs by Opponent batters
Opponent Pitching Strikeouts	Decimal	8.2	Previous ten game average of pitching strikeouts by Opponent pitchers
Opponent Ranking	String	Top Ten	Ranking of the Opponent at the time of the game
Opponent RBI's	Decimal	4.1	Previous ten game average of RBI's by Opponent batters
Opponent Score	Number	8	Score posted by the Opponent in the actual game
Opponent Stolen Bases	Decimal	0.7	Previous ten game average of stolen bases by Opponent runners
Opponent Triples	Decimal	0.4	Previous ten game average of triples by Opponent batters
Opponent Walks	Decimal	2.8	Previous ten game average of walks by Opponent batters
Opponent Walks Allowed	Decimal	4.4	Previous ten game average of walks allowed by Opponent batters
Outcome	String	Win	The result of the game for the Arkansas Razorbacks
Pitching Strikeouts	Decimal	7.9	Previous ten game average of pitching strikeouts by Arkansas pitchers
Postseason	Boolean	No	Is the game a postseason game? "Yes" or "No"
Power 5	Boolean	Yes	Is the opponent in a Power 5 conference? "Yes" or "No"
RBI's	Decimal	5.7	Previous ten game average of RBI's by Arkansas batters
Score	Number	12	Score posted by the Arkansas Razorbacks in the actual game
Stolen Bases	Decimal	0.8	Previous ten game average of stolen bases by Arkansas runners
Triples	Decimal	0.5	Previous ten game average of triples by Arkansas batters
U of A Ranking	String	Top Five	Ranking of the Arkansas Razorbacks at the time of the game
Walks	Decimal	3.6	Previous ten game average of walks by Arkansas batters
Walks Allowed	Decimal	2.3	Previous ten game average of walks allowed by Arkansas pitchers
Weekday	Number	6	Day of the week on which the game is happening (Sunday = 7, Monday = 1)

Visualizations

Chart 1 – Condensed Evaluation of Arkansas Linear Regression Equations

Scoring Model	Variables in Equation	R-Squared	Root MSE	Average Deviation from Actual Runs Scored
Model #1 (Forward)	6	0.3337	4.47	4.93
Model #2 (Backward)	6	0.4013	4.16	7.70
Model #3 (Stepwise)	5	0.3378	4.42	5.20
Model #4 (Self-Created)	12	0.4896	4.20	11.71
Model #5 (Explained Variables)	10	0.2270	4.86	4.02
Model #6 (Backward of #5)	3	0.1934	4.67	4.13

Chart 2 – Condensed Evaluation of Opponents Linear Regression Equations

Scoring Model	Variables in Equation	R-Squared	Root MSE	Average Deviation from Actual Runs Scored
Model #1 (Forward)	6	0.0779	3.55	1.83
Model #2 (Backward)	6	0.0850	3.57	1.92
Model #3 (Stepwise)	5	0.0670	3.54	1.97
Model #4 (Self-Created)	12	0.2336	3.46	3.97
Model #5 (Explained Variables)	10	0.1738	3.49	2.25
Model #6 (Backward of #5)	3	0.0331	3.54	1.66

Chart 3 – SAS Viya Output (Linear Regression Model #1 – Arkansas)

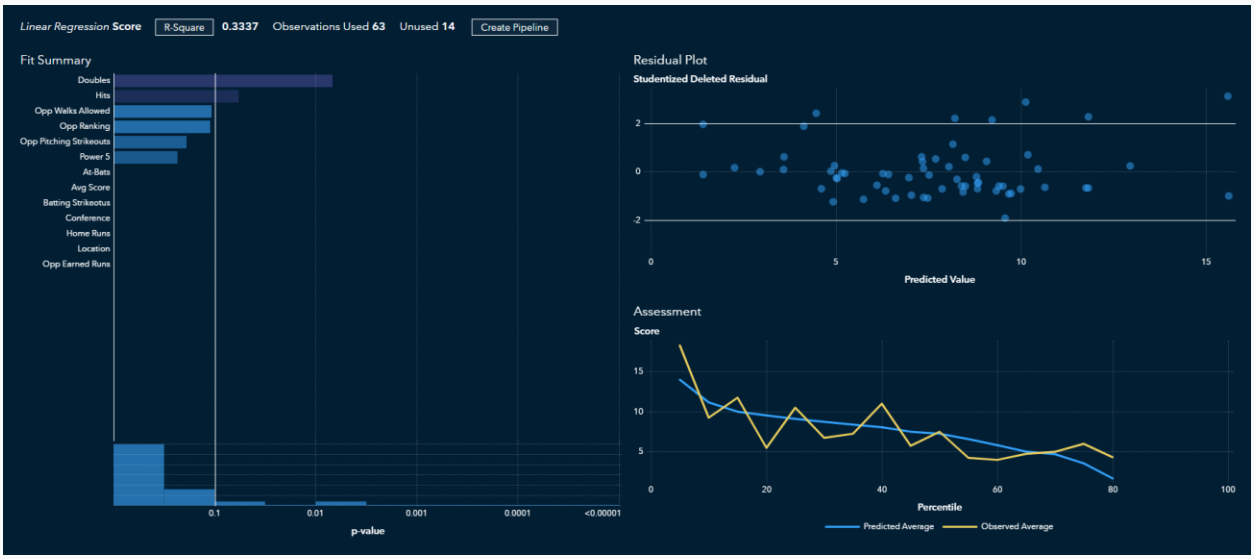

Chart 4 – SAS Viya Output (Linear Regression Model #2 – Arkansas)

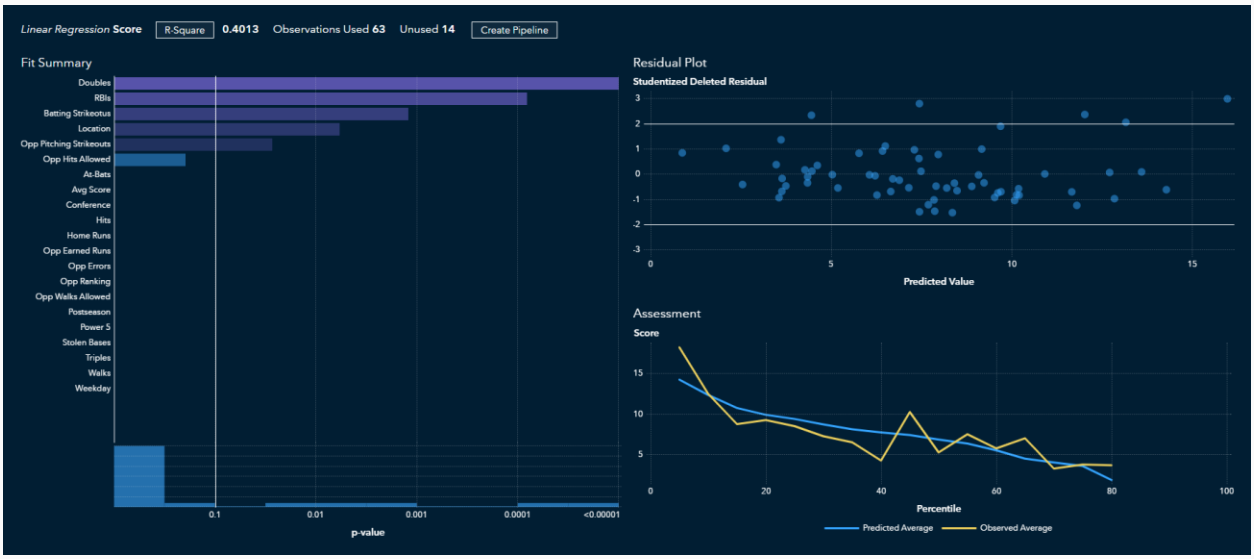

Chart 5 – SAS Viya Output (Linear Regression Model #3 – Arkansas)

Chart 6 – SAS Viya Output (Linear Regression Model #4 – Arkansas)

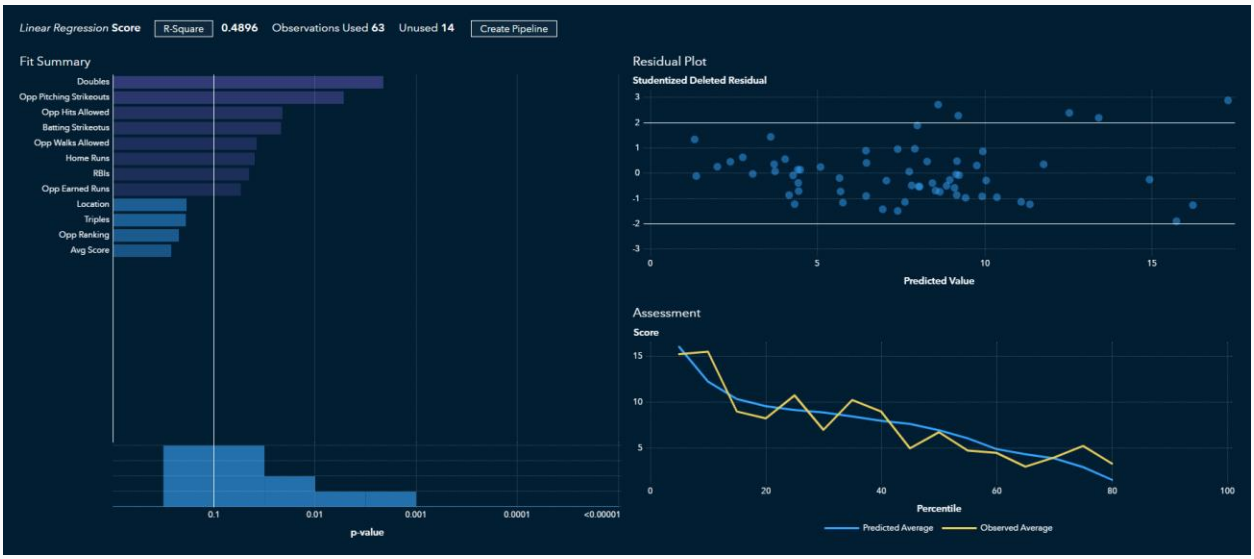

Chart 7 – SAS Viya Output (Linear Regression Model #5 – Arkansas)

Chart 8 – SAS Viya Output (Linear Regression Model #6 – Arkansas)

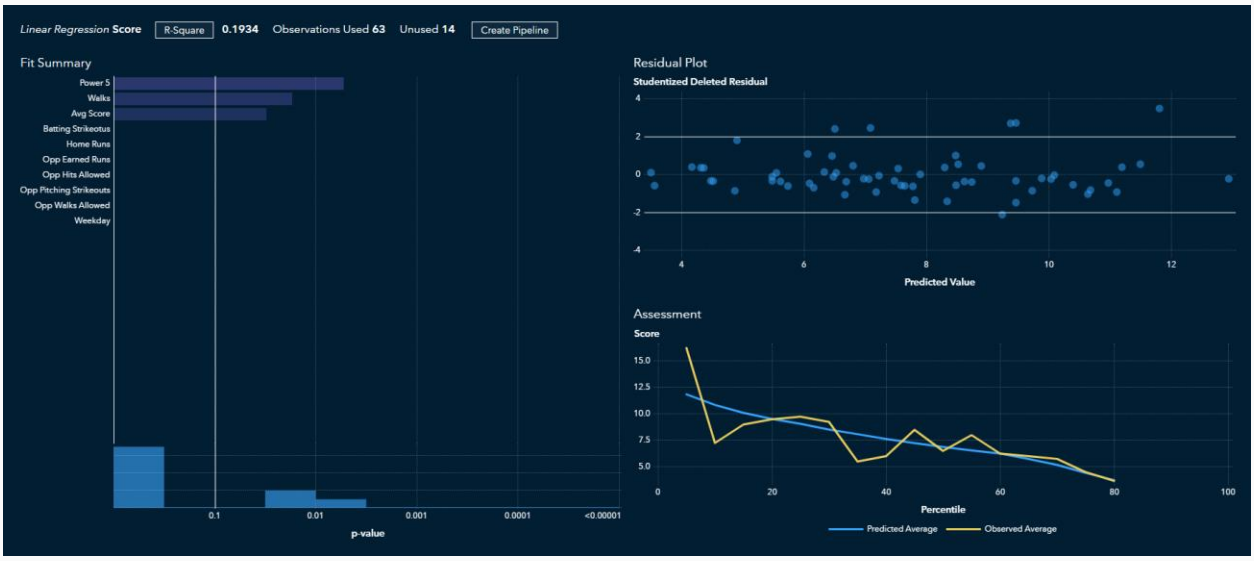

Chart 9 – SAS Viya Output (Linear Regression Model #1 – Opponent)

Chart 10 – SAS Viya Output (Linear Regression Model #2 – Opponent)

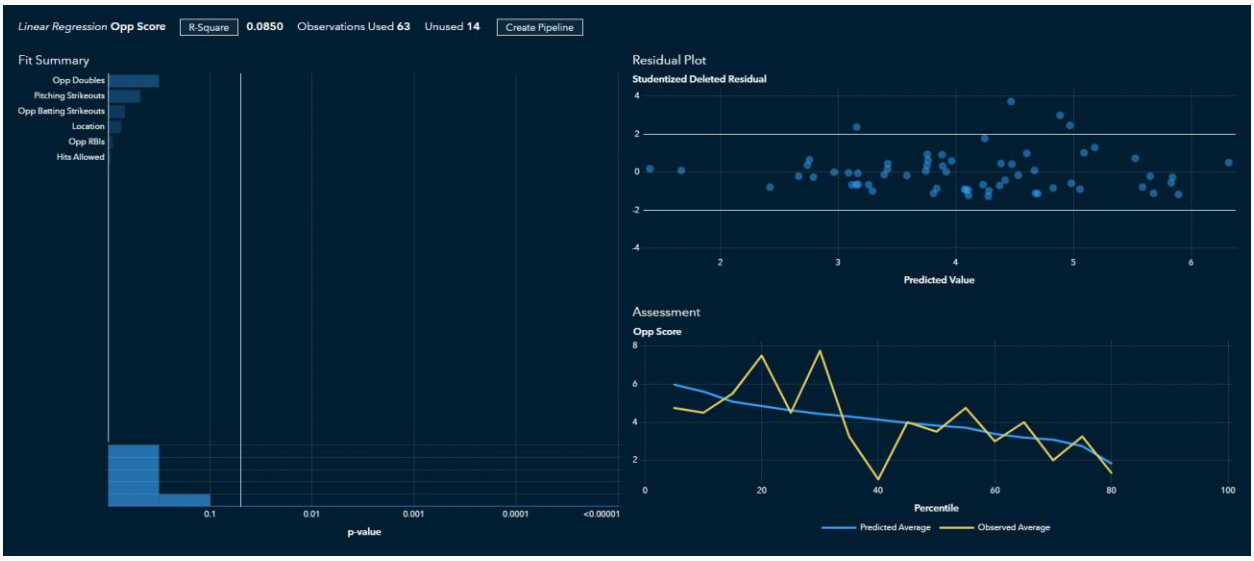

Chart 11 – SAS Viya Output (Linear Regression Model #3 – Opponent)

Chart 12 – SAS Viya Output (Linear Regression Model #4 – Opponent)

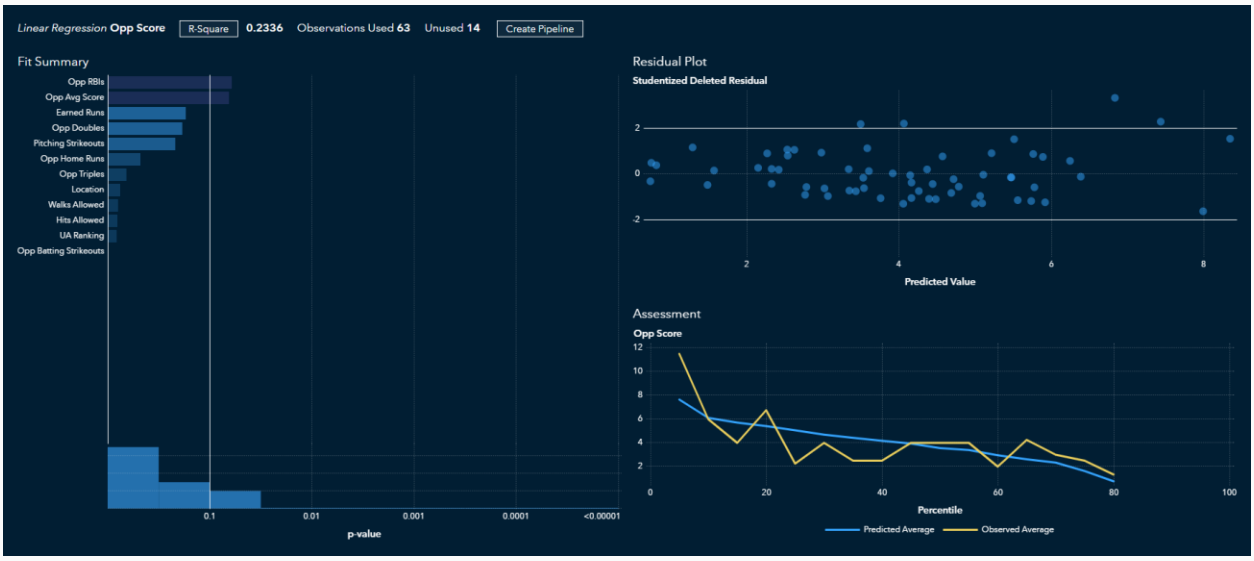

Chart 13 – SAS Viya Output (Linear Regression Model #5 – Opponent)

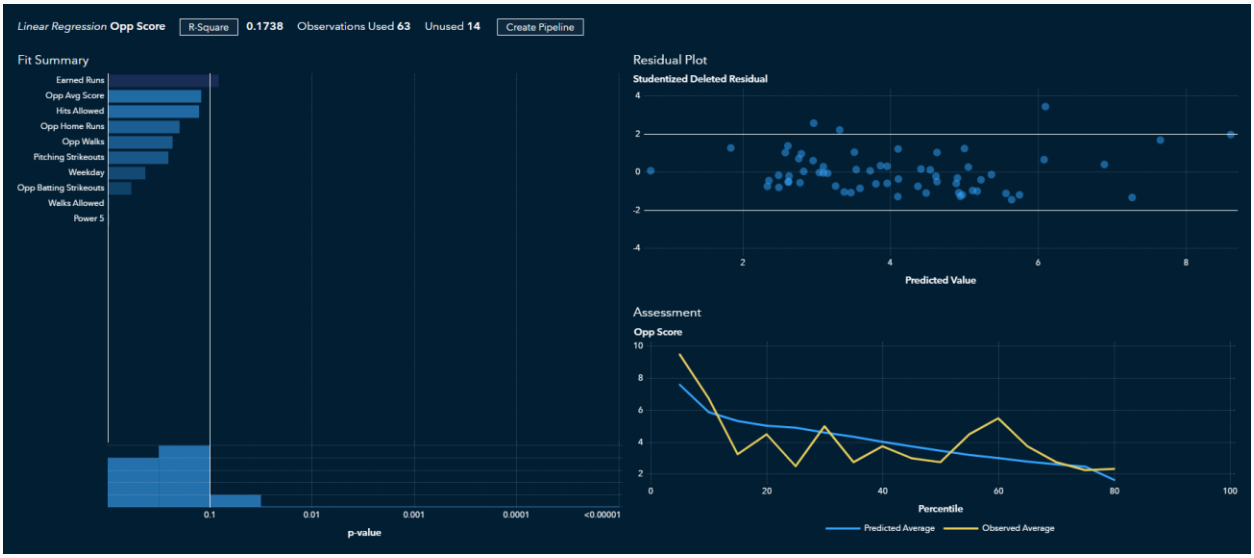

Chart 14 – SAS Viya Output (Linear Regression Model #6 – Opponent)

Chart 15 – Condensed Evaluation of Arkansas Logistic Regression Equations

Scoring Model	Variables in Equation	Misclassification Rate	% of Correctly Predicted Games
Model #1 (Explained Variables)	10	0.2222	77.27%
Model #2 (Backward of #1)	3	0.2063	81.82%

Chart 16 – Condensed Evaluation of Opponents Logistic Regression Equations

Scoring Model	Variables in Equation	Misclassification Rate	% of Correctly Predicted Games
Model #1 (Explained Variables)	10	0.1746	81.82%
Model #2 (Backward of #1)	3	0.2063	81.82%

Chart 17 – SAS Viya Output (Logistic Regression Model #1 – Arkansas)

Chart 18 – SAS Viya Output (Logistic Regression Model #2 – Arkansas)

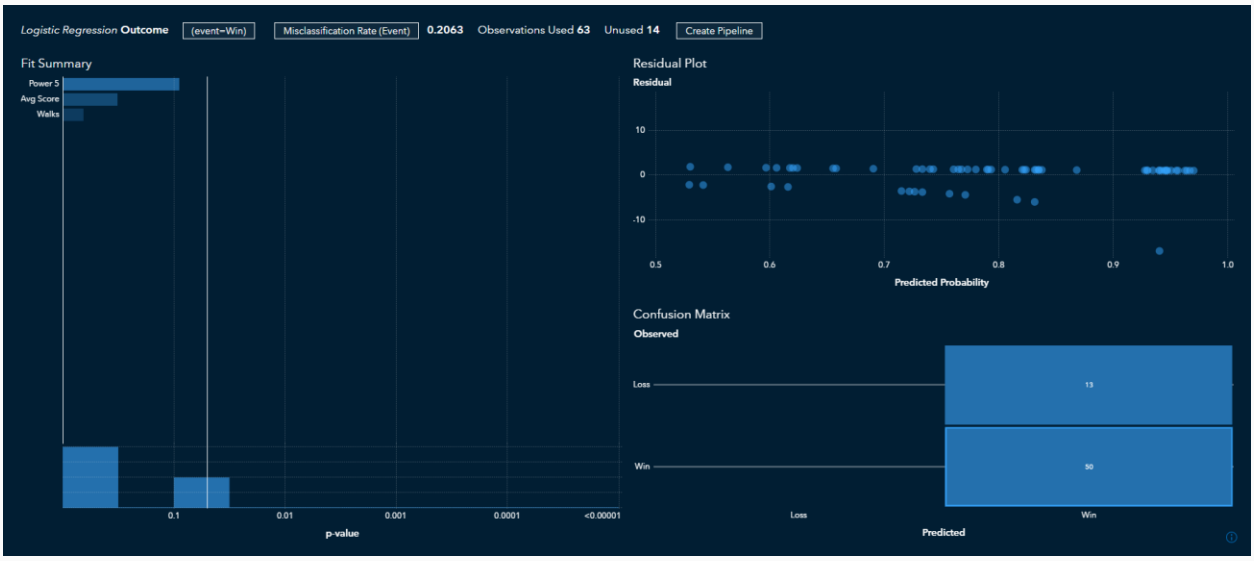

Chart 19 – SAS Viya Output (Logistic Regression Model #1 – Opponent)

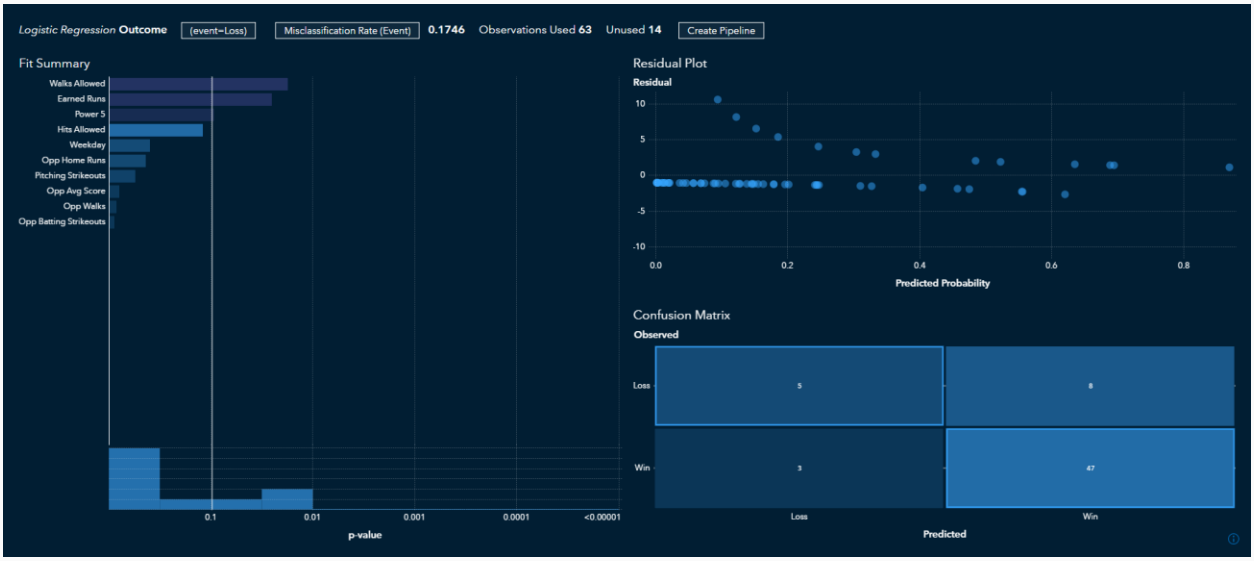

Chart 20 – SAS Viya Output (Logistic Regression Model #1 – Opponent)

Chart 21 – Projected Score Outputs (Best Linear Regression Model – Arkansas)

Date	Location	Opponent	Actual Result	Actual Score	Model #5 Projected Score	Deviation
2/18/2022	Home	Illinois State	Loss	2	11.4	9.4
2/19/2022	Home	Illinois State	Win	5	10.7	5.7
2/20/2022	Home	Illinois State	Win	4	9.5	5.5
2/25/2022	Neutral	Indiana	Win	5	8.3	3.3
2/27/2022	Neutral	Stanford	Loss	0	7.9	7.9
2/27/2022	Neutral	Louisiana	Win	6	10.7	4.7
3/2/2022	Home	Omaha	Win	15	12.9	2.1
3/4/2022	Home	Southeastern	Loss	3	11.7	8.7
3/5/2022	Home	Southeastern	Win	4	14.1	10.1
3/5/2022	Home	Southeastern	Win	11	15.5	4.5
3/10/2022	Home	UIC	Win	12	12.7	0.7
3/10/2022	Home	UIC	Win	5	10.1	5.1
3/12/2022	Home	UIC	Win	10	11.8	1.8
3/13/2022	Home	UIC	Win	10	10.9	0.9
3/15/2022	Home	Grambling State	Win	14	10.3	3.7
3/16/2022	Home	Grambling State	Win	13	8.5	4.5
3/18/2022	Home	Kentucky	Win	6	4.6	1.4
3/19/2022	Home	Kentucky	Win	9	4.6	4.4
3/20/2022	Home	Kentucky	Win	3	3.8	0.8
3/25/2022	Away	Missouri	Win	7	4.4	2.6
3/26/2022	Away	Missouri	Loss	5	4.7	0.3
3/27/2022	Away	Missouri	Win	6	6.5	0.5

Chart 22 – Projected Score Outputs (Best Linear Regression Model – Opponents)

Date	Location	Opponent	Actual Result	Actual Score	Model #5 Projected Score	Deviation
2/18/2022	Home	Illinois State	Win	3	5.8	2.8
2/19/2022	Home	Illinois State	Loss	1	6.3	5.3
2/20/2022	Home	Illinois State	Loss	2	6.7	4.7
2/25/2022	Neutral	Indiana	Loss	2	4.5	2.5
2/27/2022	Neutral	Stanford	Win	5	6.6	1.6
2/27/2022	Neutral	Louisiana	Loss	4	8.3	4.3
3/2/2022	Home	Omaha	Loss	3	4.0	1.0
3/4/2022	Home	Southeastern	Win	7	5.6	1.4
3/5/2022	Home	Southeastern	Loss	2	5.5	3.5
3/5/2022	Home	Southeastern	Loss	1	5.5	4.5
3/10/2022	Home	UIC	Loss	4	4.9	0.9
3/10/2022	Home	UIC	Loss	4	3.8	0.2
3/12/2022	Home	UIC	Loss	1	3.1	2.1
3/13/2022	Home	UIC	Loss	8	2.4	5.6
3/15/2022	Home	Grambling State	Loss	1	2.5	1.5
3/16/2022	Home	Grambling State	Loss	3	1.7	1.3
3/18/2022	Home	Kentucky	Loss	2	2.8	0.8
3/19/2022	Home	Kentucky	Loss	3	2.8	0.2
3/20/2022	Home	Kentucky	Loss	1	2.5	1.5
3/25/2022	Away	Missouri	Loss	5	5.1	0.1
3/26/2022	Away	Missouri	Win	7	5.5	1.5
3/27/2022	Away	Missouri	Loss	4	6.0	2.0

Chart 23 – Projected Game Outputs (Best Logistic Regression Model – Arkansas)

Date	Location	Opponent	Actual Score	Actual Result	Model #1 Projected Outcome
2/18/2022	Home	Illinois State	2	Loss	Win
2/19/2022	Home	Illinois State	5	Win	Win
2/20/2022	Home	Illinois State	4	Win	Win
2/25/2022	Neutral	Indiana	5	Win	Loss
2/27/2022	Neutral	Stanford	0	Loss	Win
2/27/2022	Neutral	Louisiana	6	Win	Win
3/2/2022	Home	Omaha	15	Win	Win
3/4/2022	Home	Southeastern	3	Loss	Win
3/5/2022	Home	Southeastern	4	Win	Win
3/5/2022	Home	Southeastern	11	Win	Win
3/10/2022	Home	UIC	12	Win	Win
3/10/2022	Home	UIC	5	Win	Win
3/12/2022	Home	UIC	10	Win	Win
3/13/2022	Home	UIC	10	Win	Win
3/15/2022	Home	Grambling State	14	Win	Win
3/16/2022	Home	Grambling State	13	Win	Win
3/18/2022	Home	Kentucky	6	Win	Win
3/19/2022	Home	Kentucky	9	Win	Win
3/20/2022	Home	Kentucky	3	Win	Win
3/25/2022	Away	Missouri	7	Win	Win
3/26/2022	Away	Missouri	5	Loss	Win
3/27/2022	Away	Missouri	6	Win	Win

Chart 24 – Projected Game Outputs (Best Logistic Regression Model – Opponents)

Date	Location	Opponent	Actual Score	Actual Result	Model #1 Projected Outcome
2/18/2022	Home	Illinois State	3	Win	Loss
2/19/2022	Home	Illinois State	1	Loss	Loss
2/20/2022	Home	Illinois State	2	Loss	Loss
2/25/2022	Neutral	Indiana	2	Loss	Win
2/27/2022	Neutral	Stanford	5	Win	Win
2/27/2022	Neutral	Louisiana	4	Loss	Loss
3/2/2022	Home	Omaha	3	Loss	Loss
3/4/2022	Home	Southeastern	7	Win	Loss
3/5/2022	Home	Southeastern	2	Loss	Loss
3/5/2022	Home	Southeastern	1	Loss	Loss
3/10/2022	Home	UIC	4	Loss	Loss
3/10/2022	Home	UIC	4	Loss	Loss
3/12/2022	Home	UIC	1	Loss	Loss
3/13/2022	Home	UIC	8	Loss	Loss
3/15/2022	Home	Grambling State	1	Loss	Loss
3/16/2022	Home	Grambling State	3	Loss	Loss
3/18/2022	Home	Kentucky	2	Loss	Loss
3/19/2022	Home	Kentucky	3	Loss	Loss
3/20/2022	Home	Kentucky	1	Loss	Loss
3/25/2022	Away	Missouri	5	Loss	Loss
3/26/2022	Away	Missouri	7	Win	Loss
3/27/2022	Away	Missouri	4	Loss	Loss

References

"College Recruit Rankings Class of 2022: Perfect Game USA." Perfect Game. Accessed April 21, 2022. https://www.perfectgame.org/rankings/recruiting/Rankings.aspx.

Team Stat Pages

"2021 Baseball Cumulative Statistics." Auburn University at Montgomery athletics. Accessed May 1, 2022. https://aumathletics.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." Florida Gators. Accessed May 1, 2022. https://floridagators.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." Grambling State University Athletics. Accessed May 1, 2022. https://gsutigers.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." La Tech athletics. Accessed May 1, 2022. https://latechsports.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." Little Rock Athletics. Accessed May 1, 2022. https://lrtrojans.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." Mississippi State. Accessed May 1, 2022. https://hailstate.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." Missouri State. Accessed May 1, 2022. https://missouristatebears.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." New Jersey Institute of Technology Athletics. Accessed May 1, 2022. https://njithighlanders.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." Ole miss athletics. Accessed May 1, 2022. https://olemisssports.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." TCU Athletics. Accessed May 1, 2022. https://gofrogs.com/sports/baseball/stats/2021/.

"2021 Baseball Cumulative Statistics." Texas A&M Athletics - 12thman.com. Accessed May 1, 2022. https://12thman.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." Texas Tech Red Raiders. Accessed May 1, 2022. https://texastech.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." UA Pine Bluff Athletics. Accessed May 1, 2022. https://uapblionsroar.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." University of Alabama Athletics. Accessed May 1, 2022. https://rolltide.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." University of Central Arkansas Athletics. Accessed May 1, 2022. https://ucasports.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." University of Memphis Athletics. Accessed May 1, 2022. https://gotigersgo.com/sports/baseball/stats/2021#game.

"2021 Baseball Cumulative Statistics." University of Tennessee Athletics. Accessed May 1, 2022. https://utsports.com/sports/baseball/stats/2021.

"2021 Baseball Cumulative Statistics." University of Texas Athletics. Accessed May 1, 2022. https://texassports.com/sports/baseball/stats/2021.

"2022 Baseball Cumulative Statistics." Grambling State University Athletics. Accessed May 1, 2022. https://gsutigers.com/sports/baseball/stats.

"2022 Baseball Cumulative Statistics." Illinois State University Athletics. Accessed May 1, 2022. https://goredbirds.com/sports/baseball/stats/2022.

"2022 Baseball Cumulative Statistics." Southeastern Louisiana University Athletics. Accessed May 1, 2022. https://lionsports.net/sports/baseball/stats/.

"2022 Baseball Cumulative Statistics." UIC Athletics. Accessed May 1, 2022. https://uicflames.com/sports/baseball/stats/2022.

"2022 Baseball Cumulative Statistics." University of Kentucky Athletics. Accessed May 1, 2022. https://ukathletics.com/sports/baseball/stats.

"2022 Baseball Cumulative Statistics." University of Missouri Athletics. Accessed May 1, 2022. https://mutigers.com/sports/baseball/stats/.

"2022 Baseball Cumulative Statistics." University of Nebraska Omaha Athletics. Accessed May 1, 2022. https://omavs.com/sports/baseball/stats/2022?path=baseball.

"Arkansas State." Arkansas State University. Accessed May 1, 2022. https://astateredwolves.com/sports/baseball/stats/2021?path=baseball.

"The Automated Scorebook Overall Statistics for Vanderbilt (as of Jun 30 ..." Accessed May 1, 2022. https://vucommodores.com/wp-content/uploads/2021/06/Vanderbilt-Baseball-Season-Stats-6.30.21.pdf.

"Game-by-Game Results." Arkansas. Accessed April 21, 2022. https://arkansasrazorbacks.com/stats/baseball/2021/teamgbg.htm.

"Game-by-Game Results." Arkansas. Accessed May 1, 2022. https://arkansasrazorbacks.com/stats/baseball/2022/teamgbg.htm.

"Game-by-Game Results." LSU. Accessed May 1, 2022. https://static.lsusports.net/assets/docs/bb/21stats/teamgbg.htm.

"Game-by-Game Results." Murray State - game-by-game results. Accessed May 1, 2022. https://static.goracers.com/custompages/Stats/Baseball/2021/teamgbg.htm.

"Game-by-Game Results." Southeast Missouri - game-by-game results. Accessed May 1, 2022. https://static.gosoutheast.com/custompages/stats/baseball/2021/teamgbg.htm.

"Indiana University." Indiana University Athletics. Accessed May 1, 2022. https://iuhoosiers.com/sports/baseball/stats/2022.

"North Carolina State University Athletics." NC State University Athletics. Accessed May 1, 2022. https://gopack.com/sports/baseball/stats/2021.

"Ragin' Cajuns - University of Louisiana at Lafayette Athletics." Louisiana Ragin' Cajuns. Accessed May 1, 2022. https://ragincajuns.com/sports/baseball/stats/2022.

"Stanford Athletics." Stanford University Athletics. Accessed May 1, 2022. https://gostanford.com/sports/baseball/stats/2022.

"University of Georgia Bulldogs." University of Georgia Athletics. Accessed May 1, 2022. https://georgiadogs.com/sports/baseball/stats/2021.

"University of Nebraska Athletics." University of Nebraska. Accessed May 1, 2022. https://huskers.com/sports/baseball/stats/2021.

"University of Oklahoma Athletics." University of Oklahoma. Accessed May 1, 2022. https://soonersports.com/sports/baseball/stats/2021#game.

"University of South Carolina." University of South Carolina Athletics. Accessed May 1, 2022. https://gamecocksonline.com/sports/baseball/stats/2021#game.