AI Modeling for Auburn University Baseball Team

Cycle 3 Report

Authors:
Isaiah Gallardo
Chase Garner
Sarah Holland

Comp 4710 Senior Design Team #10

Department of Computer Science and Software Engineering
Samuel Ginn College of Engineering, Auburn University

Table of Contents

1. System Metaphor – Sarah

Our system involves the creation of a specialized system tailored for the Auburn University Baseball Coaching staff, with the primary objective being the facilitation of easy access to specific statistics concerning pitcher-batter combinations. This system is envisioned to function as the "brain" of the coaches, serving as a central repository of essential data and insights. Our aspect of this system specifically revolves around pitching metrics and allowing the coaches to visualize each player on the roster, thus enabling them to use this information to make decisions.

Key features of our contribution to the system include providing comprehensive statistics related to pitcher-batter matchups, pitching performance metrics, historical outcomes, and pitch preferences. Ultimately, the system aims to empower the coaching staff with valuable insights and data-driven decision-making capabilities, thereby optimizing pitcher performance and enhancing the team's overall competitive edge on the field.

2. Previous Development – Isaiah

In the architectural spike, our team developed a table to determine correlation between variables, cleaned the data provided, and developed additional models to explore and make stronger. In addition, we developed statistics measuring the accuracy of each model. During this phase, we also set up communication channels such as discord and set up a Google Colab.

In Cycle 1, our team developed code for generating heatmaps of a pitcher's pitches. We experimented with different values for the pitch types and different amounts of spread to make the heatmaps less sparse. Additionally, we separated the code into usable well documented modules that will be easily incorporated into other parts of the project. In this cycle we also migrated our codebase to GitHub as it works better for collaborating and maintaining different versions of the project.

In Cycle 2, our team developed code to generate PDF reports for the pitchers. We created code that generated all the data for the reports including heatmaps using our previous development. We then created functionality to generate .tex files that can be compiled into a PDF report. Additionally we worked to refine the code from the previous cycle. Additional comments and documentation were added to the code along with optimizations to help with performance.

3. Cycle Intent – Sarah

Our goals for this cycle were to integrate our code with the database team and produce output on the UI for our client. This involved changing the aesthetics of the heat maps and making our code more legible. Additionally, we aimed to research more on clustering methods, learn how we could apply these methods to this problem, and begin the preliminary code to incorporate a similarity score. We aimed to present our final product to the customer after integrating each modeling team's code with the database team.

4. Future Work – Sarah

For this team, there is no more future work. Since this is our last cycle, we can leave only recommendations of uncompleted user stories for the next teams.

We believe it is critical that future teams focus on the similarity aspect of this project. There are a lot of insights that can be discovered by utilizing clustering methods, and we believe if this functionality was built into the UI, the k-means method would be highly effective. Utilizing this

scoring method would allow coaches to determine how similar a certain pitcher is to someone they have never seen. The purpose of this is, even if someone is not in the system, the client will still be able to search for someone with similar characteristics. compare these heatmaps with a pitcher they have never seen before, given characteristics.

5. User Stories – Chase

Name:	Calculate linear weights for every possible result
Summary:	As a user, I'd like to be able to calculate the linear weights for each possible result (single, double, walk, etc) based on run values.
Description:	The user will be able to see the value of each result such as a single or a triple, in terms of run values since the number of runs determines which team will win.
Planned Hours:	4
Planned Hours This Cycle:	0
Actual Hours:	2
Actual Hours This Cycle:	0
Coder:	Chase, Isaiah
Tester:	Sarah
Reviewer:	Chase, Isaiah, Sarah
Status:	Collaboration

Name:	Develop a model for the possibility of an outcome given ball-in-play characteristics
Summary:	As a user, I'd like to be able to access a model predicting the probability of future outcomes given the current ball-in-play characteristics
Description:	The user would be able to access a model informing them of the possibilities of several outcomes given what is currently happening on

	the baseball field.
Planned Hours:	2
Planned Hours This Cycle:	0
Actual Hours:	2.5
Actual Hours This Cycle:	0
Coder:	Sarah
Tester:	Chase, Isaiah
Reviewer:	Chase, Isaiah, Sarah
Status:	Completed

Name:	Run value based on count
Summary:	As a user, I'd like to calculate the potential run value based on the current count (balls - strikes).
Description:	For each count, the user would be able to predict the run value given the current number of balls and strikes.
Planned Hours:	4
Planned Hours This Cycle:	0
Actual Hours:	2
Actual Hours This Cycle:	0
Coder:	Sarah
Tester:	Isaiah
Reviewer:	Chase, Sarah, Isaiah
Status:	Collaboration

Name:	Model to update upon receiving new data

Summary:	As a user, I'd like the model to update as it receives new data.
Description:	Whenever something new is added to the database, the model will update its output accordingly.
Planned Hours:	4
Planned Hours This Cycle:	4
Actual Hours:	4
Actual Hours This Cycle:	2
Coder:	Isaiah, Chase, Sarah
Tester:	Isaiah, Chase, Sarah
Reviewer:	Isaiah, Chase, Sarah
Status:	Collaboration

Name:	Visualization for run values
Summary:	As a user, I'd like the model to provide clear visualizations of the run values.
Description:	The model will provide visual tools to make analyzing the data for the run values easier.
Planned Hours:	3
Planned Hours This Cycle:	2
Actual Hours:	4
Actual Hours This Cycle:	2
Coder:	Sarah
Tester:	Chase, Isaiah
Reviewer:	Chase, Isaiah
Status:	Collaboration

Name:	Heat map for pitches
Summary:	As a user, I'd like to be able to generate a heat map to show the strengths and weaknesses of my pitchers.
Description:	The model will generate a heat map to help the coach help their pitchers improve on their weak points.
Planned Hours:	4
Planned Hours This Cycle:	5
Actual Hours:	14.5
Actual Hours This Cycle:	4
Coder:	Sarah, Isaiah
Tester:	Chase, Isaiah
Reviewer:	Chase, Sarah
Status:	Collaboration

Name:	Player Strengths and Weaknesses
Summary:	As a user, I'd like to analyze the run values of different players to identify any strengths and weaknesses in their performance.
Description:	The model will help the coach identify any strengths and weaknesses of each player to allow the coach to make more informed decisions.
Planned Hours:	6
Planned Hours This Cycle:	5
Actual Hours:	9

Actual Hours This Cycle:	2.5
Coder:	Chase, Isaiah, Sarah
Tester:	Sarah
Reviewer:	Chase
Status:	Collaboration

Name:	Analyze player performance
Summary:	As a user, I'd like to use the model to analyze player performance to make decisions regarding the lineup.
Description:	The user will be able to compare and analyze player performance to make an informed decision regarding the lineup.
Planned Hours:	4
Planned Hours This Cycle:	3
Actual Hours:	8
Actual Hours This Cycle:	2
Coder:	Chase
Tester:	Sarah, Isaiah
Reviewer:	Sarah
Status:	Collaboration

Name:	Pitch Strategy
Summary:	As a pitcher, I want the model to provide information on how my pitch selection influences the team's run values, allowing me to improve my pitching strategy.
Description:	The model will use the data given to inform the

	pitcher on what the optimal pitch strategy might be.
Planned Hours:	3
Planned Hours This Cycle:	5
Actual Hours:	8.5
Actual Hours This Cycle:	3
Coder:	Sarah
Tester:	Chase, Isaiah
Reviewer:	Sarah, Chase, Isaiah
Status:	Collaboration

Name:	Training programs
Summary:	As a coach, I want the model to help address specific performance areas for our training programs that were identified through run value analysis.
Description:	The run value analysis given by the model will help the coach figure out what areas to focus on during training.
Planned Hours:	3
Planned Hours This Cycle:	0
Actual Hours:	0
Actual Hours This Cycle:	0
Coder:	N/A
Tester:	N/A
Reviewer:	N/A
Status:	Unstarted

6. Design Documentation – Isaiah
- Architecture

Our application is split into two parts, .py files with utility functions for the different features we are implementing and then the updateheatmaps.py file for integrating our project into the database. So far for our utility modules we have developed heatmap.py and reportutils.py. All code for our project can be accessed on our GitHub repository.

The first module, heatmap.py, implements the heatmaps and related functions. The heat maps are constructed with the different combinations of batters, pitchers, and pitch types. The heatmap function takes the data along with the ranges of both axes and the desired resolution of the heatmap. From this it generates a 2D array representing the heatmap. This heat map array can be visualized using libraries like pyplot or any other graphing library.

The second module, reportutils.py, deals with generating data and the .tex file to be able to generate a pdf report for all the pitchers. The gen_report_data() function first generates and saves all the data for the report; heatmaps, pitcher statistics, etc. The report_to_latex() function generates the .tex file that uses the data generated from the gen_report_data() function. This .tex file can then be compiled to a pdf with any latex compiler.

Lastly, the updateheatmaps.py file handles connecting to the database and updating the heatmaps stored there. It first queries the pitch data for each pitcher on each team, then it calculates the heatmaps, and finally inserts them into the database for later use.

7. Management Plan – Chase
- Task Assignments

At the beginning of the cycle, we met as a group and allocated certain tasks and user stories to each group member. The assignments for each member were as follows:
- Sarah Holland:
  - Change the color scheme of the heat maps to make them more legible.
  - Integrate our project into the database.
- Chase Garner:
  - Work on getting text statistics working.
  - Integrate our project into the database.
- Isaiah Gallardo
  - Work on getting text statistics working.
  - Commenting and cleaning up code to make it easier for the group that takes over next semester.
  - Integrate our project into the database.

Many of the tasks involved a collaboration between 2 or more group members.

- Development Schedule
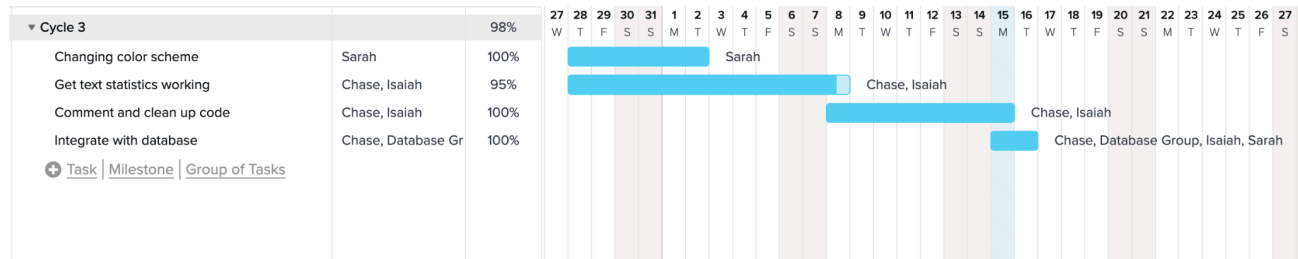
8. Risk Mitigation – Isaiah

For risk management and mitigation, our team has identified very few risks. Our biggest risk is the possibility of providing bad/misleading data to the client. We plan to mitigate this by evaluating our models for accuracy and precision. We will only release to the client when we have confidence in the performance of models that meet a desired level of performance. Another way to mitigate risk is to involve our client in the testing phase and use their domain-specific expertise to "sanity check" results so we can make corrections. Another small risk we have identified is all our code needs to be efficient. If it takes too long to run it will not be easily usable once incorporated into the database. To mitigate this we are going back through the code we have written and making any optimizations possible to speed up the run time. One last risk is in improperly accessing or updating the database. To mitigate this we have had meetings with the database team to review our code and ensure that it will not work properly with their system.

9. Testing Plan – Isaiah

ID	Test #1
Title	Model Evaluation Metrics
Pre-Conditions	The models are trained using a pre-divided portion of the data (the training set).
Test Steps	1. Select the appropriate model accuracy scoring metric. 2. Train the model with the training set. 3. Test the accuracy of the model's results against the actual values from the testing set.
Expected Results	Some of the models we create will have a higher accuracy/better performance on the data.
Actual Results	Same as expected results. Test Passed.

ID	Test #2
Title	Heat Map Evaluation
Pre-Conditions	The list of data to create the heatmap from is created as well as the range of each axis
Test Steps	1. Pass the data and the ranges to the heatmap function. 2. Get the array representing the heatmap 3. Use pyplot library to visualize the heatmap array
Expected Results	A heat map representative of the data is created and displayed
Actual Results	Same as expected results. Test Passed.

ID	Test #3
Title	Report Generation Testing
Pre-Conditions	The data needed to generate the report has been loaded into a dataframe
Test Steps	1. Pass the dataframe into the gen_report_data() function 2. Call the report_to_latex function() function to generate the .tex file 3. Use a latex compiler to generate the report pdf 4. Verify that the report pdf contains the heatmaps for all the players
Expected Results	A PDF report showing info on each pitcher is created
Actual Results	Same as expected results. Test Passed.

ID	Test #4
Title	Update Database Testing
Pre-Conditions	The data needs to be accessible on the database and the connection info saved to db_info.txt
Test Steps	1. Run the updateheatmaps.py file 2. Query the database for the heatmaps that should have been updated 3. Ensure that the queried data matches hat was supposed to be written to the database

Expected Results	The heatmaps should have been successfully updated and written to the database.
Actual Results	Same as expected results. Test Passed.

10. Lessons / Challenges – Sarah

This cycle, some of the challenges we encountered include the size of the dataset/time it takes to run things, as well as breaking things down into manageable pieces. First, for the size of the dataset, we thought that this would not be a problem, since our preliminary models did not take too long to run. However, this cycle proved differently. Even sectioning the data into each pitcher causes our heat map generation to take time, likely due to the pyplot function. This was a problem whenever we began to integrate with the database team. We knew our client wanted things to be updated after each game at least, so we brainstormed different ways to produce this output. Eventually, with feedback from the database team, we decided to create svg images that are stored in the cache. This helps to speed up the amount of time it takes to generate the images.

Additionally, we also faced challenges in the vision of the project. Our team started to have many ideas that we wanted to implement, since we understood the definition of the project better. However, we had to prioritize what was most important to the success of this cycle specifically, and that process was slightly difficult. We want to produce the best product possible for our client, and that means managing our time more efficiently. Therefore, we focused on integration this cycle and held off on things that were not as important.

11. Meeting Minutes / Memoranda – Sarah

Some of our meeting minutes have been highlighted below.

Date & Time	3/25/2024 at 8:00AM
Attendees	Sarah, Chase, Isaiah
Team Updates	Sarah– Gave database team our list of inputs so they could begin making us an API to connect to the database Chase– Worked on text statistics section of the report function Isaiah– Worked on speeding up heatmap function
Questions	What's the best way to cluster the heatmaps? What method should we use? How will we build this into the UI
Challenges	None identified

Date & Time	4/3/2024 at 8:00AM
Attendees	Sarah, Chase, Isaiah
Team Updates	Sarah– Changed heat map colors to make more aesthetically pleasing to client

	Chase– Collaborated with database UI team to determine what each player's page looked like and how our heatmaps would be integrated Isaiah– Commented code to ensure legibility and clearness to future teams
Questions	How will we get user inputs for the similarity score function? What is the easiest way to test our code before integrating with the database team?
Challenges	Not enough time to complete entirety of clustering/similarity score idea

Date & Time	4/8/2024 at 8:00AM
Attendees	Sarah, Chase, Isaiah
Team Updates	Sarah– Worked on sections of Cycle Report Chase– Commented code to ensure legibility and clearness to future teams Isaiah– Shared GitHub with Teaford, Adam, and other baseball groups; worked on sections of the Cycle Report
Questions	None identified
Challenges	None identified

12. Source Code – All

The link to our Github Repository is: https://github.com/IsaiahGallardo/SDProjectAIOrange

13. Presentation Slides – All

The link to our Presentation slides is: ☐ Cycle 3 Presentation

14. Previous Cycles / Grade Sheets – All

The link to our previous cycle submissions and grade sheets is:
Architectural Spike:
Grading:
https://auburn.instructure.com/courses/1550430/assignments/15029774/submissions/3734088
Report:
https://docs.google.com/document/d/1sse1iJIx3h0ze82KHbymZskL2AB0b14BeBycvujpRXs/edit?usp=sharing
Presentation:
https://docs.google.com/presentation/d/12kYKADzI35PH2-04PAUhCtCJw155HDDLywxbCqX02bM/edit?usp=sharing
Cycle 1:
Grading:
https://auburn.instructure.com/courses/1550430/assignments/15029749/submissions/3734088
Report:

https://docs.google.com/document/d/1txrs91IbXbjILU6IXkCrFZrEZKbviuV9aXCPskH-3RA/edit?usp=sharing

Presentation:
https://docs.google.com/presentation/d/1w6CxO_oHtq9HAuuollLCWK9_c7_dePkUdv-Dgjb0i-U/edit?usp=sharing

Cycle 2:

Grading:
https://auburn.instructure.com/courses/1550430/assignments/15029758/submissions/3734088

Report:
https://docs.google.com/document/d/1RM166SqKJQdYd-PM2e4peUT5GBDqSd4h3jxZaDSQpD8/edit?usp=sharing

Presentation:
https://docs.google.com/presentation/d/15X15A49OdHF8K8bE5qDzW4Yl5DRSZD1WRU6uqdQSFDk/edit?usp=sharing

15. Sponsor Approval – Everett Teaford

Coach Teaford is pleased with our progress thus far, and he frequently mentions this to Dr. Dozier and other faculty. Additionally, the team meets with him in-person weekly to check in, show demos, and make any adjustments as necessary. Also, we have reached out to Coach Teaford for approval on work done since our last in-person meeting via text and his response is below.