Project Scope and Deliverables | Fall 2023 AI Studio

This document is designed to help your team understand, internalize, and align on the scope, goals, and technical aspects of your AI Studio Project Challenge.

Complete all 4 sections as a team based on information you have gathered through:

* The project overview doc(s) provided in your team’s Project Folder in Google Drive (e.g., company video or slides);
* Insights gained during your first Challenge Advisor meeting during Bridge to Studio;
* Referring back to your Machine Learning Foundations summer course modules;
* Additional research done by your team related to the project/industry

Once you’re done, one team member should submit it through the assignment page in your AI Studio course in Canvas (“Business Understanding” module) by **September 3rd**. Your team’s AI Studio TA will review your submission and provide some initial feedback.

During your team’s first “Full Group” meeting during the week of September 4th with your Challenge Advisor (and AI Studio TA if they’re available), review your completed Project Scope and Deliverables document together and make updates / fill in any gaps as needed.

|  |  |
| --- | --- |
| **Project Title:** | The Bat-Detection Sonar |
| **Team Members:** | 1. Linda Dominguez 2. Anders Freeman 3. Iris Yang (she/her) 4. Anusha Bandaru 5. Emily Lu (she/her) 6. Yuhan Wang (she/her) |
| **Challenge Advisor(s):** | 1. Noah Snyder, CEO Biointerphase, [noahsnyder@biointerphase.com](mailto:noahsnyder@biointerphase.com) |
| **AI Studio TA:**  *(aka Tutor or Course Support)* | Leandra Marie Tejedor |

**PART 1: PROJECT OVERVIEW**

**Project Description**

In your own words, what are you trying to accomplish? What type of ML problem is this? (e.g., “Supervised Learning: Classification”, “Unsupervised Learning: Clustering”, etc.)

|  |
| --- |
| The goal of our project is to develop a ML model that allows us to predict bat population decline in North America using sound data. The causation of bat population decline will be based on White Nose Syndrome. This will be a supervised learning problem utilizing classification and regression.  To use acoustic modeling and other data sources (air quality, spread of human traffic, and other sources) to predict if a bat colony is about to contract white nose syndrome. This model could output A) how much the population is going to decline (regression) B) if the population could go into decline (classification) or C) the odds that the population will go into decline at some point in the future. |

**Purpose of Project**

Why is this project important or relevant to your AI Studio host company/org?

|  |
| --- |
| The purpose of this project is to allow us to correctly address and track white nose syndrome and its effects on bat populations. The company can then use our results to identify possible factors resulting in a population’s decline and partner with organizations to work on solutions to counter that decline and improve an ecosystem’s biodiversity. |

**Ethics Considerations**

Are there any potential ethics-related considerations to take into account for your project?

|  |
| --- |
| We need to be cognizant of possible biases within the dataset, such as unequal distributions of bat species, as well as being transparent about the way we present our findings and the assumptions we have made in the model-building process. |
| We will be using Microsoft Teams to meet with the challenge advisor. |

**PART 2: PROJECT SCOPE**

**Project Requirements**

What is your Challenge Advisor expecting your team to deliver by December? Are there specific algorithms that you might use as part of model training/testing? (e.g, Linear Regression, KNN) How might you evaluate your model(s)? (e.g., F1 Score, RSME)

|  |
| --- |
| He is expecting a ML model that predicts bat population decline and identifies factors relating to White Nose Syndrome. Some algorithms we might use are Logistic Regression, Linear Regression, and Random Forests. We can evaluate our models using ROC and AUC performance metrics. |

**Python Libraries**

What Python libraries do you expect to use? (e.g., Pandas, NumPy, Scikit-learn, NLTK)

|  |
| --- |
| * Numpy * Pandas * Scikit-learn * Matplot * Seaborn * (Matlab - matlab toolboxes for audio analysis) * XGBoost |

**Other Resources**

What resources (e.g., online forums, recommended research papers, example code) does your team plan to consult while working on the project? Be specific where possible (e.g., listing a specific research paper relevant to your project)

|  |
| --- |
| * Medium Articles * BTTAI canvas page * Academic paper that was discussed during meeting with challenge advisor (but have not received a copy of) |

**Timeline and Deliverables**

What tasks and outcomes do you plan to accomplish in the first few weeks? The first couple of months? List out specific steps for achieving your objectives.

|  |  |  |  |
| --- | --- | --- | --- |
| **Task**  (what will be done) | **Outcome**  (expected result of task) | **Start Date** | **Target Completion Date** |
| **Read paper + background information on WNS**  [**https://cnhp.colostate.edu/wp-content/uploads/download/documents/misc/ColoradoResultsReport-BAOApproved.pdf**](https://cnhp.colostate.edu/wp-content/uploads/download/documents/misc/ColoradoResultsReport-BAOApproved.pdf)  [**https://www.nabatmonitoring.org/methods-and-sample-design**](https://www.nabatmonitoring.org/methods-and-sample-design) | **Understand what previous work has been done** | **9/9/2023** | **9/17/2023** |
| **Get situated with Matlab** | **Understand how to use Matlab’s audio analysis tools  Note: we may not be using MatLab, put on hold** | **9/9/2023** | **9/24/2023** |
| **Start data familiarization** | **Understand what kind of data is in the dataset, and what and how we will need to clean the data** | **9/17/2023** | **9/24/2023** |
| **Start data cleaning** | **Fill in missing values, account for biases, etc.** | **9/24/2023** | **10/1/2023** |
| **Project understanding** | **Understand what it is we are going to be doing** | **9/11/2023** | **9/19/2023** |

**PART 3: DATA UNDERSTANDING**

**Data Structure and Source**

What is the source of the data? What is the data type? (e.g., numerical, time series, text, images, etc.) What is the data format? (e.g. tabular, nested, array, etc.) How much data has been or will be provided? Where will it be stored and in what format? (e.g., csv files)

|  |
| --- |
| The data is sourced from publicly available data. The data type includes: bat population data (numerical data), acoustic recordings, and White Nose Syndrome data. The data is nested and the amount of data we use isn’t limited. They are in csv files and are stored in our shared Google drive folder. |

**Data Understanding**

What are some of the variables/features of the dataset(s)?

|  |
| --- |
| Bat population numbers, white noise syndrome occurrence, and other external factors such as temperature. Some other numerical features include species, GRTS ([Generalized Random Tessellation Stratified](https://archive.epa.gov/nheerl/arm/web/pdf/grts_ss.pdf) algorithm type), year, mean population, lower bound, upper bound, regional sampling data. Categorical data: region, quantity, region type, admin 1(states and provinces). |

**Data Preparation**

What data preprocessing steps will be required? (e.g. cleaning, missing value imputation, feature engineering, etc.)

|  |
| --- |
| We need to handle missing values and account for biases within the data. In addition, we will need to homogenize and normalize data types and data values. We also need to one hot encode categorical features. Bat populations tracked across various different datasets and studies will need to be identified and grouped together for modeling. |

**PART 4: WAYS OF WORKING**

**Biweekly Meeting Details**

What will be the recurring meeting day and time for your 2 monthly virtual check-in meetings with your Challenge Advisor (“Full Group” meeting in week 1 and Challenge Advisor meeting in week 3)? Please note if these meetings will not happen in weeks 1 and 3 because of scheduling difficulties or preferences.

|  |
| --- |
| Recurring meetings will be on the first and third Monday, from 1:15-2:00 EST. |

**Challenge Advisor Communications**

How will you communicate with your Challenge Advisor outside of your biweekly virtual check-in meetings - do they prefer Slack and/or email? How will you share your meeting agendas with them 48 hours prior to each meeting - Slack, email, or a Google Drive link?

|  |
| --- |
| Meeting agendas will be shared over microsoft teams going forwards. For this meeting, agenda could not be shared because we had not received the microsoft teams invite. |

**Additional Project Stakeholders**

Are there any other stakeholders from your host company/org that your Challenge Advisor mentioned, and who your team might want to connect with to discuss the project?

|  |
| --- |
| The employees that were present during our first challenge advisor meeting may help us. |

###