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:** | “Detecting the Next Data Breach” (Accenture Team 3B ) |
| --- | --- |
| **Team Members:** | 1. Ashley Garcia Cervantes 2. Nusaiba Mahmud 3. Zaira Garcia 4. Aida Bozulan 5. Chloe Nguyen 6. Sakshi Gondkar |
| **Challenge Advisor(s):** | 1. Sanjay Ananthanarayan, Custom Software Engineering Sr Analyst, s.ananthanarayan@accenture.com |
| **AI Studio TA:**  *(aka Tutor or Course Support)* | Wenjie Wang |

**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.)

| This can be either a regression or NLP type problem depending on what we wish to accomplish.  Regression if we would like to predict where, which type of organization (e.g. Health Care, Financial, etc.), the next data breach would happen  NLP if we want to predict the method in which the next breach will come (e.g. spam email, password, etc.) |
| --- |

**Purpose of Project**

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

| The healthcare company we are working with hopes to build an ML algorithm that can help them better prepare for the next attack. Knowing which type of organization and which method is predicted to be in the next attack can help them better prepare their systems. |
| --- |

**Ethics Considerations**

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

| Since we are working with some information gathered by a specific company, we must be prudent with this information and keep it private. |
| --- |

**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)

| Our advisor recommended to aim for a 65-90% accuracy |
| --- |

**Python Libraries**

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

| Pandas  Numpy  Matplotlib  Scikit-learn |
| --- |

**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)

| CA’s recommended platforms to work with:  Google CodeLabs and Github-trello  CA’s recommended resources:  <https://www.deeplearningbook.org/>  Google  Github |
| --- |

**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** |
| --- | --- | --- | --- | --- | --- |
| **Decide the label, what we want to predict** | | **Label specification, project goal/direction** | | **Sept 8th** | **Sept 9th** |
| **Research our project info** | | **Understanding of data and circumstances of the project** | | **Sept 11th** | **Sept 17th** |
| **Decide on the best platform to code our model** | | **Platform ready to start coding** | | **Sept 18th** | **Sept 24th** |
| **Data Clean up** | | **Data ready for model** | | **Sept 25th** | **Sept 29th** |
| **Start the AI model and set task to each member to start working on** | | **Have a skeleton of what we will work on** | | **Sept 30th** | **Oct 6th** |

**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)

| Source: Company Data  Type: Numbers and Words  Format: Tabular data  Data Storage: .csv file |
| --- |

**Data Understanding**

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

| Entity  Year  Records  Organization Type  Method  Sources (to be removed) |
| --- |

**Data Preparation**

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

| We will first remove the ‘sources’ column as suggested by our advisor.  For the column ‘method’, we will replace the values that are labeled as ‘hacked’, with something more concise such as “weak password’. Lastly, we must clean some of the values in the data. |
| --- |

**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 meeting day and time with Challenge Advisor: Fridays at 3pm |
| --- |

**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?

| We will be communicating via email. |
| --- |

**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?

| For this project we will be working with Sanjay and Timo. |
| --- |

###