# Final Team Project

## Start Assignment

* **Due** Jun 23 by 11:59pm
* **Points** 200
* **Submitting** a text entry box, a website url, a media recording, or a file upload
* **Attempts** 0
* **Allowed Attempts** 3

## Introduction

This hands-on teamwork project in AAI-510 is one of the most important components of the course and learning outcome. It will consolidate the machine learning methods you have learned in this course by applying them to a real-world business case. You will be assigned to a group by your instructor. To understand the importance of teamwork in the AI space, read the [You should always think of AI and ML projects as a Team Sport.](https://www.forbes.com/sites/tableau/2022/03/10/why-you-should-think-of-ai-as-a-team-sport) article. You and your teammate(s) will work closely to solve a business problem using machine learning. **Carefully read all the instructions below.**

## Project Milestones / Timeline

* Module 2 (by the end of Week 2): The course instructor grouped students into teams of two to three members.
* Module 4 (by the end of Week 4): Each team selected and introduced a dataset. The team representative submitted the "Team Project Status Update Form."
* Module 7 (by the end of Week 7): Each team should submit the deliverables specified below for the course project in the final week.

\*\*NO extensions will be given for any of the project milestone due dates for any reason. Projects submitted after the final due date will not be graded.

## Project Task

Find an exciting business problem, find data, and solve the problem with machine learning in Python using the algorithms covered in the course. You do not need to follow the CRISP-DM methodology directly, but you should articulate the following in your final products produced:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have/need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How to get the model in production and ensure it works?

*Note: It is required that your team use GitHub to host your code, collaborate, and manage versions.* GitHub helps ensure traceability, allow rollback, and avoid unintended overwrites and loss of code. You can use the integration between Google Colab and GitHub to achieve the goals of the project if you are using Google Colab.

An important note about deployment - it is not necessary to fully deploy your model to a standalone application to satisfy the deployment component of the requirements. It is only mandatory to describe how you would perform the deployment, including any important considerations like the type (is this batch or a real-time system?) latency, cost, and how and where you might host the deployment. If you have time remaining, I’d encourage everyone to fully deploy the model to gain valuable experience, but it is not explicitly necessary in this assignment.

## Project Datasets

You will be responsible for working with your team to find a business-related dataset for this final team project. Final projects and datasets from prior courses cannot be used. However, you may use the [Home Credit Default.](https://www.kaggle.com/c/home-credit-default-risk/data) dataset that is used for other assignments from this course. There are also several [free data resource links provided in this list](https://sandiego.instructure.com/courses/19573/files/3074734?wrap=1)

### Freely Available Datasets

|  |  |
| --- | --- |
| [UC Irvine Machine Learning Repository](http://archive.ics.uci.edu/ml/index.php) | Currently maintaining 557 data sets as a service to the machine learning community |
| [Biomedical & Life Sciences Datasets](https://libguides.brown.edu/c.php?g=545426&p=3741199) | Library resources in Biology by Brown University |
| [Awesome Public Datasets](https://github.com/awesomedata/awesome-public-datasets) | A list of a topic-centric public data sources in high quality, collected from blogs, answers, and user responses |
| [Kaggle Datasets](https://www.kaggle.com/datasets) | Containing over 19,000 public datasets and 200,000 public notebooks to conquer any analysis in no time |
| [The home of the U.S. Government’s open data](https://www.data.gov/) | Data, tools, and resources to conduct research, develop web, mobile applications and design data visualizations |
| [Bureau of Labor Statistics](https://www.bls.gov/data/) | BLS Public Data gives the public access to raw economic data from all BLS programs |
| [Amazon Web Services (AWS) datasets](https://registry.opendata.aws/) | Help people discover and share datasets that are available via AWS resources |
| [BigQuery Public Datasets](https://cloud.google.com/bigquery/public-data/) | Public datasets are available for you to analyze using either legacy SQL or standard SQL queries |
| [YouTube-8M Segments Dataset](https://research.google.com/youtube8m/) | A large-scale labeled video dataset that consists of millions of YouTube video IDs, with high-quality machine-generated annotations from a diverse vocabulary of 3,800+ visual entities |
| [Harvard Dataverse](https://dataverse.harvard.edu/) | A repository for research data by Harvard University |
| [Large Health Data Sets](https://www.ehdp.com/links/datasets.htm) | A freely accessible repository for healthcare data |
| [World Bank Open Data](https://data.worldbank.org/) | Free and open access to global development data |
| [Popular Open-Source Datasets](https://enoumen.com/2021/04/23/data-sciences-datasets-data-visualization-data-analytics-big-data-data-lakes/#Energy_Dataset) | Popular open source and public data sets, data visualization, data analytics and data lakes |

General guidance for finding a dataset is it should strike a balance between a dataset that your team is interested in and excited about, but not be so complex to put together such that the data collection process takes up too much of the final project process.

## Project Scenario

In an ML Engineering team, you are responsible for convincing executives to launch ML solutions into production. The executive decision-making body often consists of executives from software development, QA, reliability, IT, cloud operations, and the CTO. Thus, establishing credibility while ensuring brevity is vital.

Before presenting to the executive decision-making body, there is a code review round and a technical solution review by a technical committee. The committee consists of fellow ML Engineers, Data Engineers, Architects, ML managers, and the Head of Machine Learning.

There are three key deliverables for this final team project:

1. Jupyter notebook (ipynb)
2. Business presentation slides (pptx/pdf)
3. Recorded video presentation (mp4)

The Juypter notebook is for the technical committee, whereas the business brief is targeted to a non-technical executive committee. Both committees will view the slides and video. Please see the following explanation for the requirements on each file. The Juypter notebook can be in a direct Jupyter instance, or a Google Colab notebook.

The Colab configuration may be easier for sharing across team members.

## Submission Requirements

This project requires that your team submit a Jupyter notebook (ipynb), business presentation slides (pptx/pdf), and a recorded video presentation (mp4).

### Jupyter Notebook with Python code

The Jupyter Notebook must be organized like a report where the code blocks are interspersed with text blocks. The text block that appears before the code block must cover the explanations of the approach. The text blocks that follow the output graphs and tables must contain inference, actionable insight, and recommendations. The code blocks themselves must be annotated with comments so they are readable. *Here is an entire textbook as Google Colab Notebooks:*[*Python Data Science HandbookLinks to an external site.*](https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks_v1/Index.ipynb)*. Here is another textbook for Jupyter Notebooks:*[*Computational and Inferential Thinking: The Foundations of Data ScienceLinks to an external site.*](https://inferentialthinking.com/chapters/intro.html)*.*

The notebook must contain the following sections:

* Problem statement and justification for the proposed approach.
* Data understanding (EDA) – a graphical and non-graphical representation of relationships between the response variable and predictor variables.
* Data preparation & Feature engineering – data pre-processing – missing values, outliers, etc.
* Feature Selection – how were the features selected based on the data analysis?
* Modeling – selection, comparison, tuning, and analysis – consider ensembles.
* Evaluation – performance measures, results, and conclusions.
* Deployment - A discussion of either the hypothetical deployment of the model or the actual deployment of the model if it has been deployed.
* Discussion and conclusions – address the problem statement and recommendation.

Please use the naming convention Final Project Section# Team Number (e.g., **Final Project SectionX-Team 1.ipynb**) and include a link to your GitHub code and provide access to the instructor(s).

### Business Presentation Slides

The business presentation slides should be no more than 12 slides prepared for a nontechnical executive audience. Curate slides for an Executive-level leader who is not familiar with statistical terminology, ML, and predictive modeling methods, i.e., a non-technical audience. Present your project as a **story** instead of a technical report. Articulate your final model results for everyone to understand with clear recommendations and actions. To improve your skill set, please read and practice [Top Ten Slide Tips.](https://www.bobpikegroup.com/trainer-blog/top-ten-slide-tips-by-garr-reynolds)  by Garr Reynolds.

* Introduce the problem – your understanding and empathy.
* Discuss the solutions that were explored and how the recommendation was chosen.
* Detail any relevant business, ethical, and regulatory risks; and recommend mitigations.

Please use the naming convention Final Project Section# Team Number (e.g., **Final Project SectionX-Team 1.pptx/pdf**)

### Video Presentation

Record a video presentation that is 15–20 minutes, the first part covering your business presentation (with slides) and the second the Jupyter notebook (screencast). Your audience this time is both the executive team as well as the technical peers and leaders that cover both aspects well. Watch this [Ted Talk by Hans RoslingLinks to an external site.](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen) for inspiration.

* Prepare a recorded video presentation of your project.
* You may use any recording software but export the video file to an mp4 format.
* Use your USD licenses to Zoom to record your screen and provide a voice narration. Ensure that the sound quality of your video is good and that each member presents an equal portion of the presentation. Export the video file to an mp4 format. Visuals are required in your presentation.
* Ensure that the sound quality of your video is good and background noise is minimal.
* Emphasis is on how you present your findings and recommendations vs. just the content of your slides and the Jupyter notebook.
* **ALL team members** must speak in the presentation with time divided evenly.

Please use the naming convention Final Project Section# Team Number (e.g., **Final Project SectionX-Team 1.mp4**).

Upload all three files to **Final Team Project** assignment page in Canvas before the deadline. Only one member of your team will need to submit these deliverables. Projects submitted after the final due date will not be graded.

You will submit the Peer Evaluation form individually using the separate assignment link in Module 7. Consult the syllabus for grading weights of the team project and peer evaluations.

**NOTE:** Team members may not get the same grade on the Final Team Project, depending on each team member's level of contribution.

This assignment has [Turnitin.](https://help.turnitin.com/integrity/student/canvas/assignments/submitting-an-assignment.htm) enabled for submissions which means that your instructor will obtain a Similarity Report that identifies specific parts of your writing that may indicate a high level of matching to external content. You are strongly encouraged to review your work without penalty by activating the [Draft Coach extension in your Google Docs.](https://help.turnitin.com/integrity/student/draft-coach/using-draft-coach.htm) prior to submitting your work for final grading.

To understand how your work will be assessed, view the scoring rubric below.

**Final Team Project Scoring Rubric**

| Final Team Project Scoring Rubric | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| GitHub and Jupyter Notebook Submission  50% | |  |  |  |  |  | | --- | --- | --- | --- | --- | | **100 pts**  **Meets or Exceeds Expectations**  All required sections are included in the repository, and adequately support the desired final outcome of the project. | **90 pts**  **Approaches Expectations**  All but 1-2 required sections are included in the repository, OR all but 1-2 sections adequately support the desired final outcome of the project | **82 pts**  **Below Expectations**  More than 2 required sections are NOT included in the repository,or more than 2 sections are not adequately support the desired final outcome of the project | **70 pts**  **Inadequate Attempt**  Submission does not resemble the requested outcome. | **0 pts**  **Non-Performance**  Non-performance. | | 100 pts |
| Business Presentation  25% | |  |  |  |  |  | | --- | --- | --- | --- | --- | | **50 pts**  **Meets or Exceeds Expectations**  The presentation clearly motivates the business problem, walks through the final solution and any alternative solutions considered, and gives clear actionable next-step recommendations. Any relevant business, ethical, or regulatory risks are highlighted. | **45 pts**  **Approaches Expectations**  The presentation does not clearly motivate the business problem OR does not walk through the final solution and any alternative solutions considered, and gives clear actionable next-step recommendations, OR relevant business, ethical or regulatory risks are NOT highlighted. | **41 pts**  **Below Expectations**  Two sections out of (1) motivating the business problem, (2) walking through the solution with recommendations, or (3) including relevant risks are not included or are well below standard. | **35 pts**  **Inadequate Attempt**  All three sections out of (1) motivating the business problem, (2) walking through the solution with recommendations, or (3) including relevant risks are not included or are well below standard. | **0 pts**  **Non-Performance**  Non-performance. | | 50 pts |
| Video Presentation  25% | |  |  |  |  |  | | --- | --- | --- | --- | --- | | **50 pts**  **Meets or Exceeds Expectations**  The video presentation would be acceptable to use in a practical business setting, all team members adequately participate, and the video is of reasonable length based on the requirements. | **45 pts**  **Approaches Expectations**  One of three is not included or is substandard (1) The video presentation would be acceptable to use in a practical business setting, (2) all team members adequately participate, and (3) the video is of reasonable length based on the requirements | **41 pts**  **Below Expectations**  Two of three are not included or are substandard (1) The video presentation would be acceptable to use in a practical business setting, (2) all team members adequately participate, and (3) the video is of reasonable length based on the requirements | **35 pts**  **Inadequate Attempt**  All three are not included or are substandard (1) The video presentation would be acceptable to use in a practical business setting, (2) all team members adequately participate, and (3) the video is of reasonable length based on the requirements | **0 pts**  **Non-Performance**  Non-performance. | | 50 pts |
| Total Points: 200 | | |