For the Final Project, you will work with your group to collaboratively solve or analyze a problem using advanced ML methodologies. In your solution, you will incorporate transformer models, natural language processing (NLP) techniques, and other tools acquired throughout the course, in addition to at least one new technology that we haven’t covered together.

Here are the specific requirements:

1. Identify a problem worth solving or analyzing.
2. Find a dataset or datasets that are sufficiently large enough to effectively train a ML model or neural network with a high degree of accuracy to ensure that your results are reliable.
3. Evaluate the trained model(s) using testing data. Include any calculations, metrics, or visualizations needed to evaluate the performance.
4. You must use at least two of the following:

* scikit-learn
* Keras
* TensorFlow
* Hugging Face
* spaCy or Natural Language Toolkit (NLTK)
* LangChain
* OpenAI

1. You must use one additional library or technology NOT covered in class, such as:

* Valence Aware Dictionary for Sentiment Reasoning (VADER)
* Whisper (OpenAI’s automatic speech recognition system)
* DALL·E (OpenAI’s text-to-image model)
* Other OpenAI capabilities, including:
  + Text-to-speech
  + GPT-4 with vision (GPT-4V)
* PyTorch

For this project, you can focus your efforts within a specific industry, as detailed in the following examples.

**Finance**

* Build a customer service chatbot for a financial firm that analyzes a user’s request and makes customized recommendations in one or more languages.
* Develop a deep learning model that forecasts and predicts stock prices for at least three publicly traded companies.
* Use NLP, transformers, or OpenAI to summarize key takeaways from a company’s earnings call.

**Healthcare**

* Build a transformer model that captions medical images in one or more languages.
* Develop a deep learning model to distinguish between malignant and benign moles.
* Use NLP to de-identify medical data such as name, birthdate, and ID number.

**Custom**

We’ve only specified healthcare and finance, but any industry can benefit from applying NLP, transformers, or OpenAI technologies. Consider preparing a data deep dive or infrastructure review that shows ML in the context of what we’ve already learned.

* Develop an integrated AI model that accurately detects and filters out spam messages.
* Use NLP to analyze social media data or customer reviews to understand user sentiment about a product, service, or issue.
* Develop a transformer model that translates between two or more languages of your choice.

**Working with Your Group**

When working on an online group project, it’s crucial to meet with your group and communicate regularly. Plan for significant collaboration time outside of class. The following tips can help you make the most of your time:

* Decide how you’re going to communicate with your group members when you begin. Create a Slack channel, exchange phone numbers, and ensure that the group knows each group member’s available working hours.
* Set up an agile project by using [GitHub ProjectsLinks to an external site.](https://docs.github.com/en/free-pro-team@latest/github/managing-your-work-on-github/managing-project-boards) so that your group can track tasks.
* Create internal milestones to ensure that your group is on track. Set due dates for these milestones so that you have a timeline for completing the project. Some of these milestones might include:
  + Project ideation;
  + Data fetching;
  + Data exploration;
  + Data transformation;
  + Data analysis;
  + Data cleaning and preprocessing
  + Testing ML models;
  + Integrate AI tools into the project for deployment
  + Creating documentation; and
  + Creating the presentation.

Since this is a two-week project, make sure that you have completed at least half of your project by the end of the first week in order to stay on track.

Although you will divide the work among the group members, it’s essential to collaborate and communicate while working on different parts of the project. Be sure to check in with your teammates regularly and offer support.

**Support and Resources**

Your instructional team will provide support during classes and office hours. You will also have access to learning assistants and tutors to help you with topics as needed. Make sure to take advantage of these resources as you collaborate with your group on this project.

**Requirements**

**Model Implementation (25 points)**

* There is a Jupyter notebook that thoroughly describes the data extraction, cleaning, preprocessing, and transformation process, and the cleaned data is exported as CSV files for a machine or deep learning model, or NLP application. (10 points)
* A Python script initializes, trains, and evaluates a model or loads a pre-trained model. (10 points)
* At least one additional library or technology NOT covered in class is used. (5 points)

**Model Optimization (25 points)**

* The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself. (15 points)
* Overall model performance is printed or displayed at the end of the script. (10 points)

**GitHub Documentation (25 points)**

* GitHub repository is free of unnecessary files and folders and has an appropriate .gitignore in use. (10 points)
* The README is customized as a polished presentation of the content of the project. (15 points)

**Presentation Requirements (25 points)**

Your presentation should cover the following:

* An executive summary or overview of the project and project goals. (5 points)
* An overview of the data collection, cleanup, and exploration processes. Include a description of how you evaluated the trained model(s) using testing data. (5 points)
* The approach that your group took to achieve the project goals. (5 points)
* Any additional questions that surfaced, what your group might research next if more time was available, or a plan for future development. (3 points)
* The results and conclusions of the application or analysis. (3 points)
* Slides that effectively demonstrate the project. (2 points)
* Slides that are visually clean and professional. (2 points)

This project will be evaluated against the requirements and assigned a grade according to the following table:

| **Grade** | **Points** |
| --- | --- |
| A (+/-) | 90+ |
| B (+/-) | 80–89 |
| C (+/-) | 70–79 |
| D (+/-) | 60–69 |
| F (+/-) | < 60 |

**Project Guidelines**

The following project guidelines focus on teamwork, your project proposal, data sources, and data cleanup and analysis.

**Collaborating with Your Team**

Remember that these projects are a group effort. The experience of close collaboration will create better project outcomes and help you in your future careers. Specifically, you’ll learn collaborative workflows that will enable you to approach and solve complex problems. Working in groups allows you to work smart and dream big. Take advantage!

**Project Proposal**

Before you start writing any code, your group should outline the scope and purpose of your project. This will help provide direction and safeguard against **scope creep** (the tendency for projects to become more complex after work begins).

The proposal is essentially a brief summary of your interests and intent. Be sure to include the following details:

* The kind of data you’d like to work with and the field you’re interested in (finance, healthcare, etc.)
* The questions you’ll ask of the data
* A possible source for the data

Use the following example for guidance:

The aim of our project is to uncover patterns in credit card fraud. We’ll examine relationships between transaction types and location, purchase prices and times of day, purchase trends over the course of a year, and other related relationships derived from the data.

**Finding Data**

Once your group has written a proposal, it’s time to start searching for data. We recommend the following curated sources of high-quality data:

* [UC Irvine Machine Learning RepositoryLinks to an external site.](https://archive.ics.uci.edu/)
* [data.worldLinks to an external site.](https://www.data.world/)
* [KaggleLinks to an external site.](https://www.kaggle.com/)
* [Data.govLinks to an external site.](https://www.data.gov/)
* [Awesome Public DatasetsLinks to an external site.](https://github.com/awesomedata/awesome-public-datasets)

**IMPORTANT**

Whenever you use a dataset or create a new dataset based on other sources (such as existing datasets or information scraped from websites), make sure to use the following guidelines:

1. Check for copyright protections, and make sure that the way you plan to use this dataset is within the bounds of fair use.
2. Document how you intend to use this dataset now and in the future. Find any licenses or terms of use associated with the dataset, and review them to confirm that your intended use is in compliance.
3. Investigate how the dataset was collected. Identify any indicators that the data was obtained from a source that the compilers were not authorized to access.

You’ll likely have to adjust your project plan as you explore the available data. That’s okay! This is all part of the process. Just make sure that everyone in the group is aligned on the project’s goals as you make changes.

Make sure that your datasets are not too large for your personal computer. Big datasets are difficult to manage locally, so consider using data subsets or different datasets altogether.

**Data Cleanup and Analysis**

Now that you’ve picked your data, it’s time to tackle development and analysis. This is where the fun starts!

The analysis process can be broken into two broad phases: (1) Exploration and cleanup, and (2) analysis.

As you’ve learned, you’ll need to explore, clean, and reformat your data before you can begin answering your research questions. We recommend keeping track of these exploration and cleanup steps in a dedicated Jupyter notebook to stay organized and make it easier to present your work later.

After you’ve cleaned your data and are ready to start crunching numbers, you should track your work in a Jupyter notebook dedicated specifically to analysis. We recommend focusing your analysis on multiple techniques, such as aggregation, correlation, comparison, summary statistics, sentiment analysis, and time-series analysis. Don’t forget to include plots during both the exploration and analysis phases. Creating plots along the way can reveal insights and interesting trends in the data that you might not notice if you wait until you’re preparing for your presentation. Presentation requirements will be further explained in the next module.

**Presentation Guidelines**

This section lists the Final Project presentation guidelines. Each group will prepare a formal, 10-minute presentation (7 minutes for the presentation followed by a 3-minute Q&A session) that covers the following points.

* An executive summary or overview of the project and project goals:
  + Explain how the project relates to the industry you selected.
* An overview of the data collection, cleanup, and exploration processes:
  + Describe the source of your data and why you chose it for your project.
  + Describe the collection, cleanup, and exploration processes.
* The approach that your group took to achieve the project goals:
  + Include any relevant code or demonstrations of the application or analysis.
  + Discuss any unanticipated insights or problems that arose and how you resolved them.
* The results or conclusions of the application or analysis:
  + Include relevant images or examples to support your work.
  + If the project goal was not achieved, discuss the issues and how you attempted to resolve them.
* Next steps:
  + Briefly discuss potential next steps for the project.

It’s crucial that you find time to rehearse before presentation day.

On the day of your presentation, each member of your group is required to submit the URL of your GitHub repository for grading.

**Presentation Day**

Your group will have a total of 10minutes—7 minutes for the presentation followed by a 3-minute Q&A session. It’s crucial that you find time to rehearse before presentation day.

On the day of your presentation, each member of your group is required to submit the URL of your GitHub repository for grading.