Final Project: End-to-End Machine Learning

Project Overview

In this final project, you will apply your skills in an end-to-end data science workflow to solve a real-world problem. You will start from scratch by collecting and cleaning your own data, and proceed through exploration, modeling, and interpretation. The goal is to bring together the techniques we've covered in this course, spanning both **unsupervised learning** (e.g., clustering, dimensionality reduction) and **supervised learning** (e.g., classification, regression), to tell a compelling, data-driven story.

This project reflects the structure of a typical **data science pipeline**, which includes:

- **Data Collection & Cleaning:** Gathering data from reliable sources and preparing it for analysis is often the most time-consuming part of a data science project. Your ability to identify, understand, and clean data is foundational for all subsequent work.
- **Exploratory Data Analysis (EDA):** EDA helps you understand the shape, patterns, and anomalies in your dataset. Through visualizations and summary statistics, you'll detect relationships and guide your modeling strategy. Incorporating **unsupervised learning** at this stage can reveal hidden structure in your data and provide valuable intuition.
- **Feature Engineering & Preprocessing:** A key step that often determines the success of a machine learning model. Creating new features, transforming existing ones, and applying dimensionality reduction or normalization can significantly improve performance.
- **Model Building & Evaluation:** You will build and compare **supervised models** for predictive tasks. This includes training models, tuning hyperparameters, and choosing appropriate evaluation metrics. More importantly, you'll justify your choices and interpret your results in context.
- **Model Selection & Communication:** Ultimately, data science is not just about building accurate models, but about making informed, transparent decisions and communicating them effectively. Your final deliverables should clearly explain what you did, why you did it, and what you learned.

This project is designed to simulate what data scientists do in practice - deal with messy data, make modeling decisions under uncertainty, and communicate results to both technical and non-technical audiences. The technical rigor and clarity of your thought process will be just as important as the performance of your final model.

Through this project, you will demonstrate not only your technical competency, but also your ability to synthesize diverse tools into a coherent and reproducible analysis pipeline.

Project Objectives

This project is designed to give you hands-on experience with the full data science lifecycle, from raw data to final model and interpretation. It emphasizes both technical rigor and thoughtful communication and will help you develop practical skills that are essential for any data science role.

By completing this project, you will:

- Practice acquiring, cleaning, and preparing data for analysis
- Explore complex datasets to uncover patterns, trends, and anomalies
- Apply unsupervised learning techniques to support exploration or preprocessing
- Engineer and transform features to improve downstream modeling
- Build and evaluate supervised learning models using sound methodology
- Compare modeling approaches and articulate the strengths and limitations of each
- Select a final model and justify your choice based on performance and context
- Communicate your workflow, insights, and decisions in a clear, structured, and reproducible report

Project Tasks

This project can be approached as a unified analysis that incorporates both **unsupervised** and **supervised** learning. The workflow should reflect a thoughtful and iterative approach to understanding and modeling data. Below is a high-level outline of the tasks you are expected to complete.

1. **Data Acquisition & Preparation**
    - Identify and obtain a dataset from a credible source (or combine multiple sources).
    - Perform necessary data cleaning, including handling missing values, inconsistencies, formatting issues etc.
    - Provide a brief summary of the raw data and the steps taken to prepare it for analysis.

2. **Exploratory Data Analysis (EDA)**
    - Conduct an initial investigation to understand the structure, trends, and potential issues in the data.
    - Use visualizations and descriptive statistics to generate insights and guide your next steps.
    - Incorporate **unsupervised learning** methods (e.g., clustering or dimensionality reduction) to discover structure, reduce complexity, or engineering features.
3. **Feature Engineering & Preprocessing**
    - Create, transform, or encode variables to support effective modeling.
    - Apply appropriate preprocessing techniques (e.g., normalization, scaling, dimensionality reduction) where needed.
    - Justify your choices based on your understanding of the data and modeling goals.
4. **Model Development**
    - Frame a predictive question and identify a target variable.
    - Build at least three distinct **supervised learning models**, applying cross-validation or other performance validation techniques (you might also want to consider a stacked model).
    - Tune model parameters as appropriate and assess performance using relevant metrics.
5. **Model Comparison & Selection**
    - Compare the performance, interpretability, and suitability of your models.
    - Choose a final model based on both quantitative evaluation and qualitative considerations, that is, justify your model selection not only based on performance metrics, but also interpretability and robustness.
    - Re-train the selected model using appropriate data and document your rationale.
6. **Communication & Reporting**
    - Organize your entire workflow into a well-structured report (bonus: and interactive product, for example, web application).
    - Use visuals and narrative to clearly explain what you did and what you found.
    - Ensure your report is reproducible and understandable to an informed but non-specialist audience.

Project Deliverables [Due on May 2nd at 11:59PM]

Your submission should include the following components. Together, these materials should clearly communicate your analytical process, support reproducibility, and demonstrate your ability to apply the full data science pipeline in practice.

1. **Final Report**
   - A clear, well-organized report that walks the reader through your entire workflow, from data collection to model selection and conclusions (in addition to other story-telling aspects, your report must describe project tasks 1-5).
   - NOTE: You project report must also include each member's contribution to the project.
2. **Code Files**
   - A well-commented Python and/or R script(s) containing the full workflow (these files should be submitted in a GitHub repository with proper documentation; include a README file with instructions on how to run the code).
3. **Datasets**
   - The raw and final clean, processed datasets (these files can be submitted in a GitHub repository).
4. **Project Enhancements (optional)**

   You are welcome to include optional elements such as:

   - A Shiny app or interactive dashboard.
   - Additional exploratory work or model experimentation beyond the main pipeline.
5. **Oral Presentation [April 30th, class time]**
   - A brief (10 minutes) in-class presentation summarizing your project goals, methods, results, and insights.
   - It should highlight your main findings and modeling choices; visual aid (e.g., slides, dashboards, or live demo) are encouraged.

Evaluation Rubrics

1. **Data Collection & Preparation [0 – 10pt]:**
   - **Basic [3pt]:** Chooses a dataset without considering its complexity or data quality. Dataset has minimal/no data quality issues, making it relatively straightforward to work with.
   - **Intermediate [6pt]:** Selects a dataset with moderate complexity and some data quality issues.
   - **Advanced [10pt]:** Selects a dataset with high complexity and significant data quality challenges.

   NOTE: Data complexity encompasses factors such as data quality, data structures (e.g., structured or unstructured data), and data sources (e.g., public repositories, web scraping), among others.

2. **Exploratory Data Analysis (EDA) [0 – 10pt]:**
   - **Basic [3pt]:** Performs basic data exploration with limited visualizations and insights.
   - **Intermediate [6pt]:** Conducts thorough data exploration with appropriate visualizations, identifying key patterns and trends.
   - **Advanced [10pt]:** Conducts extensive data exploration, utilizing advanced visualizations and statistical techniques to uncover nuanced insights and relationships; incorporates unsupervised learning to uncover structure or improve features.

3. **Data Pre-processing [0 – 10pt]:**
   - **Basic [3pt]:** Implements basic data cleaning techniques but overlooks some issues.
   - **Intermediate [6pt]:** Applies standard data preprocessing techniques effectively, handling most issues and optimizing data for modeling.
   - **Advanced [10pt]:** Implements advanced data preprocessing techniques with meticulous attention to detail, effectively addressing all data quality issues and optimizing data for optimal model performance.

4. **Feature Engineering [0 – 10pt]:**
   - **Basic [3pt]:** Implements basic/no feature engineering techniques without considering the full potential of feature manipulation.
   - **Intermediate [6pt]:** Applies standard feature engineering techniques effectively, including normalization and standardization, and handling missing values appropriately.
   - **Advanced [10pt]:** Implements advanced feature engineering techniques with a high level of creativity and sophistication. Demonstrates a deep understanding of feature

engineering principles and applies them effectively to enhance the predictive power of the dataset.

5. **Supervised Modeling [0 – 10pt]:**
   - **Basic [3pt]:** Builds a single model without tuning or validation; limited explanation of choice.
   - **Intermediate [6pt]:** Trains multiple models with reasonable hyperparameter tuning and evaluation.
   - **Advanced [10pt]:** Builds and compares multiple models using strong validation strategies; clearly articulates modeling decisions.

6. **Model Evaluation & Selection [0 – 10pt]:**
   - **Basic [3pt]:** Uses a limited set of performance metrics; model selection based mainly on one metric.
   - **Intermediate [6pt]:** Compares models using appropriate metrics and explains trade-offs.
   - **Advanced [10pt]:** Applies a wide range of metrics; justifies final model choice based on both quantitative and contextual considerations.

7. **Communication & Interpretation [0 – 15pt; + 10pt]:**
   - **Basic [5pt]:** Provides a simple written report that lacks detail and coherence; lacks clarity in conveying key points; fails to provide sufficient explanation or analysis of the project tasks, methodologies used, and findings.
   - **Intermediate [10pt]:** Prepares a clear and organized written report that effectively summarizes the project's approach, methodologies, and findings. Presents information in a logical manner, provides a moderate level of detail and analysis, offering some insight into project tasks and outcomes.
   - **Advanced [15pt]:** Produces a comprehensive and well-articulated written report that demonstrates a deep understanding of the project tasks and methodologies. Presents detailed explanations and insightful analysis of each project task, including challenges faced and solutions implemented. Communicates findings effectively, providing clear conclusions; tells a compelling, data-driven story supported by visuals and clear interpretation.
   - **Bonus [10pt]:** Bonus points may be awarded for creating a web application or interactive dashboard that effectively the data science workflow, key insights, or predictive model in a user-friendly and dynamic format.

8. **Creativity & Depth of Analysis [0 – 15pt]:**
    - **Basic [5pt]:** Follows a formulaic approach; limited insight or creativity.
    - **Intermediate [10pt]:** Demonstrates some initiative through thoughtful model choices or analysis paths.
    - **Advanced [15pt]:** Shows originality in problem framing, modeling, or interpretation; explores interesting extensions or goes beyond basic requirements.

9. **Oral Presentation [0 – 25pt]:**
    - **Basic [5pt]:** The presentation lacks clarity and structure. Key components of the project (such as data exploration, modeling, or conclusions) are missing or underdeveloped. Visuals are minimal, poorly integrated, or absent. Delivery is hard to follow due to poor pacing, low engagement, or unclear explanations. The overall communication does not effectively convey the project's main contributions.
    - **Intermediate [15pt]:** The presentation covers all essential elements of the project in a mostly clear and organized manner. There is a logical flow, and visuals support the narrative. Delivery is generally effective, though there may be some awkward transitions, rushed sections, or overreliance on notes. The audience gains a solid understanding of the work, though some parts could be more polished or insightful.
    - **Advanced [25pt]:** The presentation is clear, engaging, and professionally delivered. It tells a compelling story of the data science process from beginning to end, highlighting key decisions and findings with strong visuals and confident communication. The pacing is smooth, the content is thoughtfully curated, and the delivery demonstrates deep understanding and enthusiasm. The audience is left with a strong grasp of the project's goals, process, and impact.