Stat 486 Semester Project

The aim of this project to understand the end-to-end process of machine learning, from conceptualizing a meaningful question and sourcing the data, to analyzing the data using various machine learning techniques, and finally, deploying or presenting their findings in an effective manner. Through this project, you will get hands-on experience with the practical aspects of machine learning.

For the semester project, you should perform a thorough analysis of a dataset, motivated by a general topic and corresponding question of interest.

There are four levels or tiers available for you complete the semester project. Each tier offers varying levels of complexity and deliverables. Your chosen tier will guide the expectations for your project, allowing you to decide the level of challenge you wish to undertake. A semester project of at least Tier 4 is required in order to pass the class.

Project Description:

1. **Question Formulation:** Start with developing an intriguing question that can be answered using data. This question should be of interest to a larger audience and not be overly trivial.

2. **Data Acquisition:** Source an appropriate dataset that can be used to answer the formulated question. This dataset can be obtained from public repositories, APIs, or other resources. Ensure the dataset is legal to use and gives due credit if needed.

3. **Data Analysis and Model Building:** Depending on the data and the question, use appropriate machine learning models to derive insights or predictions. This should involve (*some steps may not be applicable to every problem*):

    - Preprocessing and exploring the data.
    - Engineering features.
    - Applying supervised learning methods.
    - Applying unsupervised learning methods.
    - Implementing interpretable AI methods to understand predictions and ascertain variable importance.
    - Evaluating and selecting models.

4. **Deliverables:** Create a public GitHub repository with the following:

    - Readme.md file that explains the purpose of your repository and the contents of the repo
    - Neat, documented, and organized code used for the project
        - *does not need to be one file*
    - File with a brief description of the feedback and iteration process you used
        - *I'm thinking that this will be a markdown file, but you can use any format that can be viewed directly on GitHub*

        - Not required for "Tier 4"
    - Final report that summarize your project
        - *See the "Final Report" section for more details

Project Tiers:

Tier 4 (Lowest level of engagement)

- **Data**: Simple
    - ~50 or fewer instances
    - ~6 or fewer features of similar type
    - Very little processing and cleaning
- **EDA:**
    - Basic descriptive statistics
    - Exploration of within-variable distributions
- **Analysis:** The project should include at least two of the following:
    - Minor feature engineering
    - Implementation of at least two standard supervised machine learning models without extensive parameter tuning
    - Implementation of at least one advanced supervised machine learning with some degree of hyperparameter tuning
    - An exploration of variable importance
- **Feedback & Iteration:**
    - None
- **Deliverables:**
    - GitHub Repo

Tier 3

- **Data:** Moderately simple
    - ~200 or fewer instances
    - ~10 or fewer features of similar type
    - Very little processing and cleaning
- **EDA:**
    - Basic descriptive statistics
    - Exploration of within-variable distributions
    - Analysis of bivariate relationships
    - Identification and treatment of outliers
- **Analysis:** The project should include at least three of the following:
    - Minor feature engineering

    - Implementation of at least one standard supervised machine learning models without extensive parameter tuning
    - Implementation of at least two advanced supervised machine learning with some degree of hyperparameter tuning
    - Cluster analysis or anomaly detection
    - An exploration of variable importance
- **Feedback & Iteration:**
    - On final report, specifically ask peer reviewers:
        - Is anything unclear, too wordy/detailing, too vague?
        - Are there any obvious typos or formatting issues?
- **Deliverables:**
    - GitHub Repo

Tier 2

- **Data**: Moderately complex (satisfies several of the following)
    - ~1000 or more instance
    - Many features of varying types
    - Non-tabular data
    - Some processing and cleaning
- **EDA**
    - Basic descriptive statistics
    - Exploration of within-variable distributions
    - Analysis of bivariate relationships
    - Identification and treatment of outliers
    - Analysis of multivariate relationships using pair plots, parallel coordinate plots, 3D scatterplots, and/or tSNE plots
- **Analysis:** The project should include at least four of the following:
    - Significant or innovative feature engineering
    - Implementation of at least two standard supervised machine learning models without extensive hyperparameter tuning
    - Implementation of 2-4 advanced supervised machine learning methods with detailed hyperparameter tuning
    - Cluster analysis and/or anomaly detection
    - Use of another unsupervised learning method, such as meaningful dimension reduction, apriori or recommender systems
    - An exploration of variable importance using an explainable AI method (e.g., SHAP, LIME)
    - Use of a significant ML method we didn't learn in class
- **Feedback & Iteration:**
    - After analysis step, specifically ask peer reviewers:
        - Is there anything that is incorrect?
        - Are there other models or methods that could be considered?
    - On final report, specifically ask peer reviewers:

        - Is anything unclear, too wordy/detailing, too vague?
        - Are there any obvious typos or formatting issues?
- **Deliverables:**
    - GitHub Repo:
    - A web application, dashboard or blog post

Tier 1 (highest level of engagement)

- **Data:** Moderately complex to complex (satisfies several of the following)
    - ~1000 or more instance
    - Many features of varying types
    - Non-tabular data
    - Some processing and cleaning
- **Analysis:** The project should include at least FIVE of the following:
    - Significant or innovative feature engineering
    - Implementation of at least two standard supervised machine learning models without extensive hyperparameter tuning
    - Implementation of 2-4 advanced supervised machine learning methods with detailed hyperparameter tuning
    - Cluster analysis and/or anomaly detection
    - Use of another unsupervised learning method, such as meaningful dimension reduction, apriori or recommender systems
    - An exploration of variable importance using an explainable AI method (e.g., SHAP, LIME)
    - Use of a significant ML method we didn't learn in class
- **Feedback & Iteration:**
    - After EDA, specifically ask peer reviewers:
        - Are there any aspects that you didn't explore?
    - After analysis step, specifically ask peer reviewers:
        - Is there anything that is incorrect?
        - Are there other models or methods that could be considered?
    - On final report, specifically ask peer reviewers:
        - Is anything unclear, too wordy/detailing, too vague?
        - Are there any obvious typos or formatting issues?
- **Deliverables:**
    - GitHub Repo:
    - A web application, dashboard or blog post

Final Report

Writing up results in machine learning, especially in situations where multiple models are explored, requires a balance between thoroughness and brevity. Here's a guideline on how to approach this:

1. **Introduction:**

- Start with the problem statement and what you aimed to achieve. This gives context to any reader unfamiliar with the project.
- Briefly touch on why multiple models were deemed necessary for exploration.

2. **Exploratory Data Analysis (EDA):**
    - Detail any key findings from the EDA, as this can justify certain modeling choices or feature engineering efforts.

3. **Overview of Models Tried:**
    - Present a concise list or table of all models attempted. For each model, provide:
        - A very brief description (especially if it's a less common model).
        - Key hyperparameters explored.
        - Highest-level results (e.g., a range of accuracy scores, or best and worst scores).

4. **Discussion on Model Selection:**
    - Without going into deep details, mention any patterns observed across models, such as:
        - Were tree-based models consistently outperforming linear models?
        - Did ensemble models show significant promise over single models?
    - Talk about major pitfalls or challenges faced with certain models, e.g., overfitting with a deep neural network, or convergence issues with a certain algorithm.
    - Briefly touch upon why certain models didn't make the cut. This doesn't have to be detailed but can include reasons like:
        - Poor performance on validation data.
        - Overfitting issues.
        - Computationally too intensive for the marginal gain in accuracy.
        - Difficulty in hyperparameter tuning.
        - If/how any models provided value in the exploration phase, even though they were not deemed optimal for the problem at hand.

5. **Detailed Discussion on Best Model:**
    - Go in-depth into the model that performed the best.
    - Discuss hyperparameter tuning, feature importance (if applicable), and any post-processing steps (like threshold adjustments for classification).
    - Visualize results, confusion matrices, ROC curves, or other relevant metrics to make the case for this model's superiority.
    - Talk about the interpretability of this model, if possible. Can you explain why it's making certain predictions or variable contribute most to the predictions?

6. **Conclusion and Next Steps:**
    - Conclude by reinforcing the choice of the best model and its implications.
    - Discuss any potential future steps or improvements that could be made, perhaps leveraging models or techniques not yet explored.

General Tips:

- **Visuals Over Text:** Graphs, charts, and tables can sometimes communicate model comparisons more efficiently than prose. For instance, a table comparing the accuracy, recall, precision, and F1 score of all models tried can be a quick way for readers to understand the performance landscape.
- **Use Appendices:** If there's a concern that some details about discarded models might be of interest to some readers, consider placing those details in an appendix. This way, the main report remains concise, but those who wish to dive deeper have the option.
- **Stay Objective:** While it's tempting to defend choices passionately, it's crucial to remain objective. Let the data and results drive the narrative, not personal preferences or biases.