

Project 2

Deadline: December 01, 2024, 11:59pm ET.

Note: You need to name your team and state the name in each document you submit in this project, including the report, presentation, and the source code.

Description:

In Project 1, you have analyzed your dataset and extracted insights following a thought process to acquire a good understanding of the data and the problem, as it was focused on the exploration and analysis of a dataset and a problem. In this project, you will use the insights and information that you gathered in project 1 to build a machine learning (ML) model that would potentially provide an effective solution to your problem. You are expected to focus on experimenting with multiple ML models to compare their performance using evaluation metrics. The choices you make in modeling using ML are expected to be motivated by the insights you extracted in project 1. So, you need to justify your choices for using a particular technique or algorithm in your approach through the outcomes from your project 1 or new insights you discovered in this project. Your justification can also be from past work documented in the form of academic literature, white paper, or a blog article, as the techniques or algorithms might have been used before, and in this case, you will need to provide references to these sources. Your evaluation of the models needs to be based on the chosen metrics as to which model is performing best. Based on your problem, you need to justify the evaluation metrics that you chose to use in your approach. Further, you need argue why this model will perform well and improve the business metrics (i.e., KPIs) when deployed in the workflow of the organization. Then, you need to provide a discussion as to why the model you pick is performing better and why it would perform well in a real-world application – in this case, your application. How your evaluation metric is informing the outcomes of your analysis and the predictions your model is producing. Please provide supporting visualizations with their analysis wherever needed.

Note: You may choose to change your problem and dataset for this project. If you do, you need to provide the exploratory analysis as in project 1 and provide justifications in this project as needed.

This project will have multiple deliverables that are a report, Jupyter notebook and in-person presentation. Their fraction in grades will be as follows: project report %40, source code in Jupyter notebook %30, project presentation %30.

1) Project Report:

You need to write your report emphasizing four aspects of your problem and approach as explained below. You are expected to provide visualizations and tables wherever it is needed. This report will be in PDF format.

- **Motivation and Problem:** First, you need to describe why this problem is important and why people should care enough to pay you for solving it. Then provide your problem statement in very clear terms as there should not be any ambiguity.
- **Solution:** Second, you need to provide a high-level description as to what is your solution to solve this problem, and why your approach is the right solution.
- **How it works:** Third, you need to explain how your approach works. Start with providing an architecture of your approach to give a big picture. Then, you need to provide ample details of the components of your approach that will explain their inner workings. Also, you need to provide justifications for why you chose those specific techniques or algorithms. Explain what steps you took to improve the model, e.g., tuning models, manipulating data.
- **Evaluation, Outcomes and Discussion:** Fourth, you need to provide results of the models based on the evaluation metrics you chose. Explain which models are performing better and why this might be the case. Provide a discussion as to why the model you pick will perform well, and state if there may be situations it may not perform well. What are key takeaways from your analysis and models?

2) Source code (Jupyter Notebook):

You need to provide your source code. You are expected to comment your code as it should be readable and understandable by any person with a background in python. If not commented properly, 5% will be deducted.

3) Project Presentation:

As your target audience is a partner organization, assume that you already have your data and model built and you're trying to convince the partner company to move forward with integrating your model into their systems. As a reminder, this presentation is not only for other data scientists, but also for non-technical people, such as CEO or other admins. You need to communicate effectively what you accomplished in this project and your outcomes to a semi or non-technical group of individuals.

Critical components:

- Clear introduction to the problem and motivation as to why this problem is important to solve.
- Visualizations showing from the data emphasizing your motivation and the problem.
- Describe the dataset, provide stats. with visualizations.
- Provide a high-level architecture of your proposed solution.
- Explain important insights from project 1 that informs your ML modeling approach.
- Describe your approach in building and experimenting with ML models.

- Present visualizations for your evaluation of the models and explain better performing models.
- Provide clear outcomes and key takeaways that will argue why your model will solve the problem of the organization and what would be the expected impact and outcomes from the model you built.

Extra Credit (5%):

- During the presentations, each student will vote for the best presentation. The teams cannot vote for their own. The team with the most votes will receive a 5-point bonus towards the final grade of project 2.

You need to submit to iCollege:

- Presentation (e.g., pptx, pdf)
- Jupyter notebook. –comment your code.
- Dataset
- Report (up to five pages) in pdf format. You can briefly recall and repeat the insights from projects 0 and 1, but you still need to refer to them, e.g., (see Project0).