Project 1

Deadline: December 05, 2025, 11:59pm Eastern Time

Description:

In Project 0, you have brainstormed ideas and identified a problem, dataset, and a probable solution. In this project, you are expected to explore and analyze the dataset, and extract insights following a thought process that will allow you to acquire good understanding of the data and the problem. It will also enable you to come up with an effective solution. This project is focused on the exploration and analysis of a dataset and problem, while you do ***not*** have to create a ML model at the end. Provide supporting visualizations with their analysis wherever needed.

1. Show overall descriptive statistics of your dataset; number of data points, number of descriptive features, type of features, your target feature and its type, descriptive features for different target feature values. (10 points)
2. Determine if any features have missing data and what should be done with the missing data. Explain why the decision was made for each feature. If there is no missing data, explain how you would handle missing data and why. Provide supporting visualizations with their analysis. (10 points)
3. Explore your features further in their distributions and plot their bar and box plots. How are individual features distributed? Show outliers for each feature. Do you think any of the outliers may impact your analysis? Why? Provide supporting visualizations with their analysis. (20 points)
4. What data pre-processing techniques do you apply? E.g., encoding features, missing values, scaling, etc. Explain each process and why you use it. (10 points)
5. Analyze distribution of your target feature. Is it balanced or imbalanced? Do you think any of these may cause a problem and why? Provide supporting visualizations with their analysis. (10 points)
6. Analyze distribution of descriptive features. Do you think any of these may cause a problem and why? Do you see any correlation among them or against target feature? If so, what problem it may create and how would you handle this situation? Provide supporting visualizations with their analysis. (10 points)
7. What kind of new features would you engineer to represent your data points (i.e., rows) better and why? Explain what information it represents. How new features distribution look like through visualization? (15 points)
8. [Bonus] What kind of ML approaches and algorithms do you choose to use and why? E.g., supervised, regression, classification, binary, multi-class, split rate of data, logistic regression, SVM, decision trees etc. What evaluation metrics you used to evaluate the performance of your model. Discuss the results of your model as to which model performs better and why this would be the case. How would your model perform based

on the results? What shortcomings your model has and possible implications? What would you do to improve the results? (10 points)

9. Reflect on your thought process, steps and explain what kind of stages and processes you have gone through to make decisions in each step. For instance, why did you choose the preprocessing techniques you used? what led you to choose the evaluation metric you use? what motivated your selection of ML algorithms for prediction? Provide supporting visualizations with their analysis. (15 points)

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 0 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 extract insights, e.g., manipulating data, etc..
- **Outcomes and Discussion:** Fourth, provide a discussion as to what are key takeaways and insights from your analysis? How it will impact the business objective and outcomes? How it will improve KPIs your company tracks periodically?

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 0 that informs your analytics approach.
- Describe your approach in building analytics approach and experimenting with data.
- Present visualizations for your analysis of data and explain key insights.
- Provide clear outcomes and key takeaways that will argue why your appraoch will address the problem of the organization and what would be the expected impact and outcomes from the approach 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 1.

You need to submit to iCollege:

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