Project 1

Deadline: November 4, 2021, 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 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 will also 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. (10 points)
2. Explore your features further in their distributions and plot their box plots. 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)
3. 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 handle missing data and why. Provide supporting visualizations with their analysis. (10 points)
4. What data pre-processing do you apply? E.g., encoding features, missing values, scaling, etc. Explain each process and why you use it. (10 points)
5. Analyze the balance or distribution of your target variable. Do you think any of these will present a problem and why? Provide supporting visualizations with their analysis. (10 points)
6. What kind of ML approaches and algorithms do you take and why? E.g., supervised, regression, classification, binary, multi-class, split rate of data, logistic regression, SVM, decision trees etc. Provide supporting visualizations with their analysis. (10 points)
7. 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? (10 points)
8. What is the solution you propose and implement to solve the problem? Explain your thought process as to what kind of stages and processes you have gone through to make decisions in each step. For instance, what led you to choose the evaluation metric you use? what motivated your selection of ML algorithms for prediction? why did you choose the preprocessing techniques you used? Provide supporting visualizations with their analysis. (20 points)

Guidelines:

Project 1 will be graded according to the following guidelines:

- A score between 0 and 100 will be assigned.
- If it is not submitted before the specified deadline, then a grade of 0 will be assigned.
- The team leader will submit a separate report that briefly describes the contribution of each team member to the joint effort. Also, the leader will provide a score between 0 and 5, where 0 indicates no effort and 5 indicates equal contribution.
- At the beginning of the project 0 report, please include the names of the team members specifying the team leader and GSU email addresses. Also, include the following text:
    - "Statement of Academic Honesty: The following code represents our own work. We have neither received nor given inappropriate assistance. We have not copied or modified code from any source other than the course webpage or the course textbook. We recognize that any unauthorized assistance or plagiarism will be handled in accordance with Georgia State University's Academic Honesty Policy and the policies of this course. We recognize that our work is based on an assignment created by the Institute for Insight at Georgia State University. Any publishing or posting of source code for this project is strictly prohibited unless you have written consent from the Institute for Insight at Georgia State University."

You need to submit to iCollege:

- Jupyter notebook. –comment your code as needed.
- Dataset
- Report (up to two pages). You don't have to repeat the same information from project 0. You can just refer to project 0, e.g., (see Project0).
- Contribution report (by the team leader)