Data Mining and Visualization (83676)

Final Project

Part 2

Due: 12/5/22, 11:59 pm

General

In this second part of the project, you are requested to create a classification model that will be used by the marketing managers to decide which new potential customers to contact.

The submission will be in the same pairs as in the first submission and constitutes 60% of the project final grade. It should include three files:

- Python notebook file (.ipynb) with the full code.
- A report (word or PDF) with explanations of the process and results.
- A CSV file with a predictions for the attached test data set.

The grade evaluation takes into consideration:

- Techniques used: Did you select appropriate and diverse techniques and justify why?
- The process you followed: Is it correct (given the techniques you used), did you describe it well?
- Interpretation of results: Did you correctly understand and interpret the results you obtained?
- Quality of writeup: Did you present your work well, in an understandable and usable manner?
- The classification model quality: The test set predictions will be ranked based on accuracy relative to the class, and the top three will receive a 5 point bonus (with a maximum grade of 100 points).

Instructions

- Implement the pre-process on the attached test dataset and describe what adjustments you made (consider the notes given to you in the first part of the project to improve your data).
- Split the data to train and validation set and use cross-validation method, show the differences between iterations.
- Choose and explain what are the most appropriate evaluation metrics for this problem, show at list 4 evaluation metrics and explain their importance. Add visualizations.

- Train **at least** five different classifiers and present evaluation metrics to compare between them, describe the results and provide your opinion. Include at least one new classification model that wasn't covered in class, and explain it briefly. Add visualizations to show the performances of the different classification models.
- Perform hyperparameters tuning for each of the models you train. Explain the choices you made regarding the hyperparameters to calibrate and their range of values.
- Choose the best model, in terms of statistical significance, to apply to the attached test data set to generate predicted classification. Save the results in a CSV file. Make sure to explain why you chose a model to be the "best model".
- Incorporate visualizations whenever possible. Make sure to be creative and make graphs and visualizations as pretty as possible and readable!
- Note: You can change the pre-processing from part 1 if you think it will lead to better results, write and explain the changes in the report.

  Good Luck!