Data Mining and Visualization (83676)

Final Project

Part 2

Due: 6/6/21, 11:59 pm

<u>General</u>

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

The submission will be in the same pairs as in the first submission and 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:

- The process you followed: Is it correct (given the techniques you used), did you describe it well?
- Techniques used: Did you select appropriate techniques and justify why?
- Interpretation of results: Did you correctly understand and interpret the raw results you obtained?
- Quality of writeup: Did you present your work well, in an understandable and usable manner?
- The classification model quality: Your predictions on the test set will be ranked relative to the class and 20% of your score will be according to this ranking.

<u>Instructions</u>

- Split the data to train and validation set and use cross-validation method.
- Choose and explain what are the most appropriate evaluation metrics for this problem.
- Train **at least** three different classifiers and present evaluation metrics to compare between them, describe the results and provide your opinion.
- Perform hyperparameters tuning for each of the models. Explain the parameters you have chosen to calibrate and why.

- 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.
- 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. Furthermore, don't forget to apply the pre-process to the test set as well.

Good Luck!