Project II

- Work individually.
- Submit a code and a detailed project report. The names of the files should be YOURNAME.py (or YOURNAME.ipynb) and YOURNAME.docx (e.g. OrsanOzener.py, OrsanOzener.docx)
- Either submit the data file you have used in the project or provide a link to the file.
- Any type of plagiarism will not be tolerated and will lead to disciplinary actions.
- **Due Date: 15th of June 2020, 13:00. Please submit your report through LMS. Only ONE submission per person.**

1 Introduction

This project will be a follow-up task of Project I. Remember in the first project assignment, you were supposed to work on a problem and data of your choice and assess the benefits of feature elimination. In this assignment, you are supposed to take this to next level and perform one of the task below (you may choose to do both):

- Option 1: Consider all the relevant methods suitable for your problem and choose the best model, with the best features and with the best hyper-parameter setting. You are supposed to have three data partitions, training, validation and test, perform cross-validation steps when necessary, and report the results with respect to an appropriate criteria based on the problem setting you choose. Note that for this option, I am expecting a quite detailed submission, so basically performing all the relevant methods, discussed in class.
- Option 2: I rather prefer you to choose this option as this is something I wished to cover in class but I could not due to limited time. Thus, I think this project will be an venue for you to learn this concept. Recall that I have mentioned that there exist several methods to assess feature importance. One such methods borrows concepts from Lloyd Shapley, a Nobel Laureate, which is called the Shapley Value. The concept, Shap Value, is based on assessing the feature importance on a micro level, based on each individual observation, therefore providing both a global (aggregate) and local (individual) feature importance vector. Based on this one, you may cluster your observations and come up with conclusions/insights such as “in this clusters these

features are important and in that clusters some others are important", in fact you may even use different methods for those clusters for prediction and inference. Hence, you are supposed to assess the feature importance based on Shap Value and using these results improve the prediction performance by feature elimination, decomposing the problem into clusters, etc. whatever method you think is relevant.

If you have any questions, please send an email to: orsan.ozener@ozyegin.edu.tr