PROJECT ASSIGNMENT

- Work individually.
- Submit a code and a short explanation along the code. The names of the files should be YOURNAME.py (or YOURNAME.ipynb) and YOURNAME.docx (e.g. OrsanOzener.py, OrsanOzener.docx)
- Any type of plagiarism will not be tolerated and will lead to disciplinary actions.
- **Due Date: TBD. Please submit your report through LMS. Only ONE submission per person.**

1 Introduction

In this assignment, you are expected to work on a relatively large loan data from a Kaggle competition (https://www.kaggle.com/wendykan/lending-club-loan-data). Based on their description: "These files contain complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables. A data dictionary is provided in a separate file."

On this data you are suppose to performs three tasks (use any method we have discussed in class):

- Predict time until loan is paid in full, charged off, or defaults (this is in fact the question on the competition, so please do not spend too much time on this one as my real questions are the ones below).
- Classify the loans with respect to "loan_status" column. Hence, obviously you need a train/test split (use 80-20 split).
- Compare the difference between "loan_amnt" column and "funded_amnt_inv" column. Basically, what I am asking is predict which applications will be underapproved by the investors and by how much?

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