FDS Project 19/20

Group members:

- Simone Marretta
- Luca Scofano
- Daniele Trappolini

Private score: **0.78115**

Description of the project:

We are given 7 different datasets with multiple features each. The main goal is to predict whether a new individual will be able to repay the loan or not. We tackled different steps to achieve our goal:

1. Importing all the datasets
2. Quick analysis of the main datasets
3. Feature Engineering as well as One hot encoding
4. Cross- validation model (K-fold)
5. Light GBM model and ROC-AUC score

Comments:

We tried with different types of models, but Light GBM + K-folds produces the highest score. **What is K- fold?** It is used to reduce variability, in most machine learning methods multiple rounds of cross-validation are performed using different random partitions, and the validation are averaged at the end. This method is known as k-fold cross validation. In essence, it splits it into K folds, trains on K-1 and then tests on the left-out.

We decided to use all datasets, as each one have them have lots of important features. After some feature engineering we merged all the datasets into one, that we called “df” and the classification is done on this last one. Most feature engineering is done to create ratios, these new variables are useful to summarize some meaningful aspects.

Another important aspect was to fix categorical variables since most statistical operations could not be done. Since most variables don’t have an order, we decided to use One Hot Encoding that converts variables into binary ones. (An example is provided in the Notebook)