Lab assignment 1: Classification

Statement

Datasets:

This assignment will analyze the **FICO_Dataset.csv** dataset which contains the following information:

- The data contains information about Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years.
- Source:
  https://community.fico.com/s/explainable-machine-learning-challenge?tabset-158d9=3
- Attribute Information:
  1. **ExternalRiskEstimate**: A measure of borrower's riskiness based on consolidated external data sources.
  2. **NetFractionRevolvingBurden**: The proportion of an individual's current credit usage compared to their maximum allowed credit.
  3. **AverageMInFile**: The average duration, in months, of the trades in a borrower's credit file.
  4. **MSinceOldestTradeOpen**: The age, in months, of a borrower's oldest credit account.
  5. **PercentInstallTrades**: The percentage of a borrower's credit accounts that have fixed payment terms over a specified period.
  6. **NumSatisfactoryTrades**: Count of trades where a borrower has met obligations satisfactorily.
  7. **NumTotalTrades**: Number of Total Trades (total number of credit accounts).
  8. **MSinceMostRecentInqexcl7days**: Months since the last credit inquiry, ignoring the most recent week.
  9. **PercentTradesNeverDelq:** The percentage of a borrower's trades with no history of delinquency.
- Output:
  10. **Risk Performance**: Paid as negotiated flag (12-36 months). Class variable (0 or 1)

NOTE:

The dataset contains some special characters which correspond to the following situations:

-9	No Bureau Record or No Investigation
-8	No Usable/Valid Trades or Inquiries
-7	Condition not Met (e.g. No Inquiries, No Delinquencies)

Deliverables:

You will have to submit two files through **Moodlerooms** before October 24th:

- **A report in PDF format** that contains the developed code --screen captures of the code are not allowed--, justified answers to the proposed questions, and analyses of the results. The report does not need to be long, but should demonstrate that you worked through the whole statement. Do not include figures or code without a comment about it. Remember to include a conclusion section.
- **A compressed folder**, in .zip or .7z format, with all your code files and any additional file[1] that you might want to attach (for example, a model which takes too long to train).
- **Quality of the code will be assessed and may penalize the final grade of the assignment.**
- **Format of the report will be assessed and may penalize the final grade of the assignment.**

[1] https://pythonbasics.org/pickle/

Questions:

The objective of this practice is to compare different classification algorithms with a real dataset.

Load the dataset **FICO_Dataset.csv** and:

1. Exploratory Data Analysis (EDA)
2. Identification and fitting process of classification models
3. Comparative analysis of the fitted models
4. Creativity and innovation

   In this section you should look for resources on the internet that are applicable to this assignment and were not taught in class. These resources can be concepts, techniques, packages, etc... that are applicable in this exercise. For example, you could use packages that aid in dataset exploration or classification model analysis. Extra effort on the other sections would also be taken into account in this section.
5. Conclusions