Final Project - Machine learning

Introduction

This first part of the final project will analyze the `FICO_Dataset.csv` file. 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.

You can find the original source of the dataset in its webpage. However, **the dataset has been modified to ease the tasks you must complete in this project**.

Dataset

The FICO dataset has the following **input variables**:

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.

And the **output variable**:

10. `Risk Performance`: Paid as negotiated flag (12-36 months). Class variable: 0 (`'Good'`) or 1 (`'Bad'`).

Special characters

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 in Moodle before April 24th**:

1. A report in `.pdf` format that contains the **justified analysis 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 (**screen captures are not allowed**). The length of the report must be **8 pages or less** (cover, index and bibliography pages do not count).
2. A compressed folder, in `.zip` or `.7z` format, with all your code files and any additional file that you might want to attach (*e.g.* a saved model which takes too long to train). You must include a `README.md` file with the steps to reproduce the results of your report.
3. **The quality of the code will be assessed and may penalize the final grade of the assignment.**
4. **The format of the report will be assessed and may penalize the final grade of the assignment.**

Questions

Part 1: Exploratory Data Analysis (1.75 pts)

The objective of this part is to explore and obtain information from a real dataset. Load the dataset `FICO_Dataset.csv` and perform the following tasks:

1. **Exploratory Data Analysis** (EDA) (1.5 pts)
2. Divide data into **train** and **test sets** (0.25 pts)

Part 2: Classification (4.5 pts)

The objective of this part is to compare the performance of different classification algorithms with a real dataset. Based on the training and test sets from Part 1, do the following:

1. **Identification** and **fitting** process of **classification models**. (1.0 pts)
2. **Comparative analysis** of the fitted models. (2.5 pts)
3. **Creativity and innovation** (1.0 pts)

Part 3: Unsupervised Learning (2.25 pts)

The objective of this part is to compare different clustering algorithms with a real dataset. Based on the training and test sets from Part 1 performed:

1. **PCA** fitting process and analysis. (0.5 pts)
2. **Identification** and **fitting** process of **clustering algorithms**. (0.75 pts)
3. **Creativity and innovation** (1.0 pts)

Part 4: Conclusions (1.5 pts)

The objective of this part is to analyze the results from parts 1 to 3. Here you should answer questions such as, but not limited to:

1. Are the hypothesis about variable importance from Part 1 confirmed in Part 2?
2. Which is the best classification algorithm? Why?
3. If the clustering methods are used for the classification task, would the performance be similar to the algorithms in Part 2?

→ (other possible questions *you* may come up with...)

Hint #1:

For the creativity and innovation questions, you should look for resources on the internet and/or in the reference textbooks that apply to this assignment and were not taught in class. These resources can be concepts, techniques, packages, etc... that are applicable in this exercise. Extra effort on the other questions will also be accounted for in this section.