Data Scientist - Challenge (No Time Limit)

Objective:

You are asked to build the most accurate model you can to predict target column for `data_test.csv`. The metric to reflect accuracy can be defined by yourself.
The column details are below:

- `id`: id column for data_train, data_test, respectively
- `num*`: numerical features
- `der*`: derived features from other features
- `cat*`: categorical features
- `target`: target column, only exists in data_train. it is binary.

There are potentially missing values in each column. The goal is to predict `target` column for `data_test.csv`.
The solution should have a result csv file with two columns:
1. 'id': the id column from data_test.csv
2. 'target': the predicted probability of target being 1
The corresponding code to reproduce the result csv file should be included as well.

Discussion Questions

In addition to a code submission and solution, please provide short answers to the following discussion questions in a README (plain-text or Markdown format):

- Briefly describe the conceptual approach you chose! What are the trade-offs?
- What's the model performance? What is the complexity? Where are the bottlenecks?
- If you had more time, what improvements would you make, and in what order of priority?

Preparation

Download the zipfile containing training and testing data from from this link.

Assessment Criteria:

In no specific order:

- If your solution satisfies the requirements
- How the code and functionality is tested
- The understandability and maintainability of your code
- The cleanliness of design and implementation
- Time performance on a standard laptop
- Answers to the discussion questions.