Problem Statement

A person's creditworthiness is often associated (conversely) with the likelihood they may default on loans. Here, we ask you to look at data on loan applicants and build a model to predict whether or not an application can be deemed creditworthy.

We're giving you anonymised data about 1000 loan applications, along with a certain set of attributes about the applicant itself, and whether they were considered high risk. We want you to work your magic✨ and predict whether or not a future loan application is a high credit risk given this data.

Do note that it is worse to predict an applicant as a low credit risk when they are actually a high risk (cost=4 below), than it is to predict an applicant to be a high credit risk when they aren't (cost=1 below).

		Prediction	
		0	1
Actual	0	0	1
	1	4	0

This table contains a possible cost matrix where rows represent the actual classification and the columns the predicted classification with classes: 0 = Low credit risk, 1 = High Credit risk

What we will be looking for:

1. EDA (exploratory data analysis), any data preprocessing you performed on the data, and feature engineering to create a dataset for modelling. We want to see how your feature engineering evolves from your EDA!
2. Using your EDA, try and answer these questions (and use the write-up to explain how you arrived at the answers):
    a. Would a person with critical credit history be more creditworthy?
    b. Are young people more creditworthy?
    c. Would a person with more credit accounts be more creditworthy?
3. Choose an evaluation metric you will use to compare & evaluate model performance on a hold-out test set.
4. Train model(s) to predict the creditworthiness of a customer. Describe your strategy when choosing the model(s) you train and provide an explanation for why you feel a subsequent model would improve upon a prior one (and your insights when it does/doesn't).

5. Compare the model(s) performance with the evaluation metric(s) and choose a final model you feel is most optimal along with your reasoning.

Optional, but good to have:

- A non-ML predictor that you can use as a baseline to compare your models' performance.
- Some degree of model optimisation/fine-tuning. (As much as you're able to!)
- Modular, functional/object-oriented code is always appreciated!
- Using your real-world knowledge, tell us which other data features might have helped.

Expected Submission

We're flexible in terms of how you'd like to present your output though do remember to share your code/scripts (even simple scripts used to explore the data) and a brief write up about your solution (can even be a Jupyter - or any other - notebook). We would appreciate a README file with an overview of the files included.

Do use a standard archival format (.zip, .tar, .tar.gz, etc) vs proprietary archival formats please.

Code

- Please submit all the code/scripts/notebooks you've written, even if it is just a simple script to explore the data. You can even do it as an Jupyter (or any other) notebook.
- Please do not include any pyc files, ipynb checkpoints, or other libraries/output generated by the interpreter/ from the build process.

Write-up

This gives us a sense of your approach and thinking. Here are some ideas on what to include:

- A description of your approach to the solution.
- Your model evaluation metric and why you chose it.
- Relevant details on your model(s), along with the evaluation metric value for each on a held-out test set, and reason for choosing the final model.
- Any visualizations you may have created (along with corresponding observations).
- Any interesting insights that you may have found in the data.
- Any other information that you feel is relevant.
- Your write up should be preferably in pdf, odf, markdown or html.

Dataset Description

The dataset has two files:

1. **`applicant.csv`**: This file contains personal data about the (primary) applicant
   - Unique ID: `applicant_id` (string)
   - Other fields:
     i. Primary_applicant_age_in_years (numeric)
     ii. Gender (string)
     iii. Marital_status (string)
     iv. Number_of_dependents (numeric)
     v. Housing (string)
     vi. Years_at_current_residence (numeric)
     vii. Employment_status (string)
     viii. Has_been_employed_for_at_least (string)
     ix. Has_been_employed_for_at_most (string)
     x. Telephone (string)
     xi. Foreign_worker (numeric)
     xii. Savings_account_balance (string)
     xiii. Balance_in_existing_bank_account_(lower_limit_of_bucket) (string)
     xiv. Balance_in_existing_bank_account_(upper_limit_of_bucket) (string)

2. **`loan.csv`**: This file contains data more specific to the loan application
   - Unique ID: `loan_application_id` (string)
   - Target: `high_risk_application` (numeric)
   - Other fields:
     i. applicant_id (string)
     ii. Months_loan_taken_for (numeric)
     iii. Purpose (string)
     iv. Principal_loan_amount (numeric)
     v. EMI_rate_in_percentage_of_disposable_income (numeric)
     vi. Property (string)
     vii. Has_coapplicant (numeric)
     viii. Has_guarantor (numeric)
     ix. Other_EMI_plans (string)
     x. Number_of_existing_loans_at_this_bank (numeric)
     xi. Loan_history (string)