reunion

Data Science Assignment - REUNION

👉 **Note**: *Please go through the assignment and instructions carefully. Submissions will be accepted via the Submission Form only and not on email.*

Background

A person's creditworthiness is often associated (conversely) with the likelihood they may default on loans.

We're giving you anonymized data on about 1000 loan applications, along with a certain set of attributes about the applicant itself, and whether they were considered high risk.

0 = Low credit risk i.e high chance of paying back the loan amount

1 = High credit risk i.e low chance of paying back the loan amount

▼ Dataset

data.zip 41.4KB

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:
  - Primary_applicant_age_in_years (numeric)
  - Gender (string)
  - Marital_status (string)
  - Number_of_dependents (numeric)
  - Housing (string)
  - Years_at_current_residence (numeric)
  - Employment_status (string)
  - Has_been_employed_for_at_least (string)
  - Has_been_employed_for_at_most (string)
  - Telephone (string)
  - Foreign_worker (numeric)
  - Savings_account_balance (string)
  - Balance_in_existing_bank_account_(lower_limit_of_bucket) (string)
  - Balance_in_existing_bank_account_(upper_limit_of_bucket) (string)

2. `loan.csv`: This file contains data more specific to the loan application

- Target: `high_risk_application` (numeric)
- Other fields:
  - applicant_id (string)
  - Months_loan_taken_for (numeric)
  - Purpose (string)
  - Principal_loan_amount (numeric)
  - EMI_rate_in_percentage_of_disposable_income (numeric)
  - Property (string)

- Has_coapplicant (numeric)
- Has_guarantor (numeric)
- Other_EMI_plans (string)
- Number_of_existing_loans_at_this_bank (numeric)
- Loan_history (string)

TASK-1

1. Do the Exploratory Data Analysis & share the insights.
2. How would you segment customers based on their risk (of default).
3. Which of these segments / sub-segments would you propose be approved?
    - For e.g. Would a person with critical credit history be more creditworthy? Are young people more creditworthy? Would a person with more credit accounts be more creditworthy?
4. Tell us what your observations were on the data itself (completeness, skews).

TASK-2

Develop the ML model(s) to predict the credit risk(low or high) for a given applicant.

Business Constraint: Note that it is worse to state an applicant as a low credit risk when they are actually a high risk, than it is to state an applicant to be a high credit risk when they aren't.

Provide the answers for the below points:

1. Explain your intuition behind the features used for modeling.
2. Are you creating new derived features? If yes explain the intuition behind them.
3. Are there missing values? If yes how you plan to handle it.
4. How categorical features are handled for modeling.
5. Describe the features correlation using correlation matrix. Tell us about few correlated feature & share your understanding on why they are correlated.
6. Do you plan to drop the correlated feature? If yes then how.
7. Which ML algorithm you plan to use for modeling.

8. **Train two (at least) ML models** to predict the credit risk & provide the confusion matrix for each model.
9. How you will select the hyperparameters for models trained in above step.
10. Which metric(s) you will choose to select between the set of models.
11. Explain how you will export the trained models & deploy it for prediction in production.

Optional but good to have:

1. Build a dashboard using python/tableau/any other platform of choice to visualize the above
2. Tell us about other real-world data features that you would have liked to see in the dataset and why you think these features might have enriched your analysis

---

TECH STACK TO BE USED

Python, Jupyter Notebook or Collab, Pandas & other helping libraries like Scikit-Learn, Matplotlib, Pytorch etc.

SUBMISSION GUIDELINES

1. Complete TASK-1 & TASK-2 in two separate notebook or colab. Please add the proper heading & comments for each sub-task in the notebook before sharing.
2. Either share the Jupyter Notebook through Kaggle or share the working Colab
3. Also push the Jupyter Notebook or Colab to a ***private*** Github repo & share the access with Github users `kshitij-g` and `neekneeraj`
    a. Please note committed notebook should have all the cells output required for explanation.
4. Please document and explain your analysis [Note: Your write up should be preferably in pdf or doc format.]
    - Include visualizations (other than the dashboard) that you may have created
    - Include interesting insights that you may have found in the data
    - Include any other information that you feel is relevant
5. Share any other document, charts etc with us which you think will help you to explain things in a better way.

Submission Form

Please host all your documents in Google Drive or Dropbox or any other cloud service. Please make sure that all the documents are publicly accessible so that we can view them. Please submit your work using the form below.

Name *

Email ID *

Role Type *

A Internship

B Full-time

You solution

Please host your submission files in Google Drive, Dropbox, etc. Please make sure that all the shared folders are accessible (we will not evaluate the assignment if the access is denied). You can include your descriptions and explanations here.

Submit