Instructions

1. This is an individual-participant challenge.
2. After you have built your application, you must do the following:
    - Submit the solution in .zip format
    - At the end of the page, you must complete the following tasks:
        - Upload the following as a zip, tar, or tar.zip archive:
            - Prediction file
            - Source file
        - Submit your solution
3. The submissions are evaluated automatically. The evaluation parameters include the following:
    - Functionality of the code
    - Design aesthetics
    - Usability of the application
4. The evaluation of submissions can take up to a week's time.
5. When the challenge is live, your output will be evaluated only for 50% of the test data. After the challenge is over, your output for the remaining 50% of the test data will be evaluated and the final rank will be awarded.
6. If you do not select a submission file for the offline evaluation, your best submission will be automatically considered.
7. You will have to upload your submissions on the **Problems** page in the format specified in the problem statement.
8. In addition to your final submission, you will also have to submit your source file and other files as a .zip or .tar compressed archive.
9. The total number of submissions allowed for a participant throughout a challenge is 600. The maximum number of submissions that a participant can make in a day is 10.
10. You can use any tools or libraries to build your solution. There is no restriction on the tools that you can use.
11. The Intellectual Property (IP) of the product/code of the winners will belong to HackerEarth (only when they accept the prize). Other participants will retain the IP over their product/code. They can choose to put it in an open source domain under any license.
12. You will receive your prize after the announcement of results on the contest page in the last week of the following month. But note that if your nation does not accept PayPal payments, we will not be able to send you any cash prizes.
13. Prizes above are mentioned per individual.
14. In order to claim the prize, your leaderboard score must be reproducible from your code files.
15. Use of external dataset is prohibited for this challenge. Participants found using external dataset will be disqualified.

By participating in this challenge, you are agreeing to HackerEarth's terms and conditions.

Predictive Lending - Piramal Finance

Piramal Finance caters to the diverse lending needs of the people of Bharat with a portfolio of products ranging from Secured loans to Unsecured loans. Acquisition models are used to judge the credit worthiness of a customer before giving out loan. Post giving out a loan, various internal metrics are used to track the performance of the loan to take necessary steps whenever required.

Task
You are given a set of loans and various parameters on it. Your objective is to train a binary classification model which outputs the probability of a loan going bad.

Dataset description
The train dataset is provided with ground truth for model development. The test dataset will be used for evaluation, and you need to make submission for these. Each of the train and test datasets are further divided into three parts following the same format as
following file

- â€˜xyz_1.csvâ€™ (where â€˜xyzâ€™ can be â€˜trainâ€™ or â€˜testâ€™) contain the primary data with both the identifier columns, â€˜loan_id and â€˜idâ€™, and â€˜labelâ€™ in case of train data.
- The files â€˜xyz_2_1.csvâ€™ and â€˜xyz_2_2.csvâ€™ contains additional secondary data with â€˜idâ€™ as identifier.

The â€˜train_1.csvâ€™ contains the â€˜labelâ€™ column which takes value 1 if there was a default in the loan. These two datasets contain same columns but the values belong from two different time periods. For example, if development time period is of September 2023, the data in â€˜xyz_2_1.csvâ€™ and â€˜xyz_2_2.csvâ€™ will be of August 2023 and April 2023 respectively.

SNo	File Name	Description	Columns
1	train_1.csv	Primary data to be used for modelling. This contains the loan performance over a period of time.	loan_id, label, id, prod, col 1â€¦.col n
2	train_2_1.csv	Secondary data which contains bureau data for time period 1	id, add 1â€¦.add n
3	train_2_2.csv	Secondary data which contains bureau data for time period 2	id, add 1â€¦.add n
4	test_1.csv	Primary data to make predictions (analogous to train_1.csv)	loan_id, id, prod, col 1â€¦.col n
5	test_2_1.csv	Secondary data for test (analogous to train_2_1.csv)	id, add 1â€¦.add n
6	test_2_2.csv	Secondary data for test (analogous to train_2_2.csv)	id, add 1â€¦.add n
7	sample_submission.csv	Sample file to illustrate final submission format	loan_id, prob

The columns provided in the dataset are as follows:

Column name	Description
loan_id	Represents primary key (loan ID) in the train and test dataset to be used for final prediction
id	Represents the secondary key (customer ID) to be used to combine additional bureau data
prod	Represents the product categorisation (masked)
col_1 to col_164	Represents the features in train and test data
add_1 to add_677	Represents the additional features in bureau data
label	Represents the ground truth in the train data
prob	Represents the predicted probability for final submission

Evaluation

```
score = 100*metrics.roc_auc_score(actual, predicted)
```

Public Leaderboard

- A public leader board will be available for participants to test their performance and compare it with others.
- The public leader board will be based on a subset of datapoints from test dataset
- The public leader board is provided for the sole purpose of helping participants compare their performance with others.

Private Leaderboard

- A separate private leader board will be used for final evaluation on remaining datapoints from the test dataset.
- This will not be visible to participants.
- Final evaluation will be based on model performance on private leader board.

Result submission guidelines

- The index is *loan_id* and the target is the *prob* column.
- The submission file must be submitted in *.csv* format only.
- The size of this submission file must be *100000 x 2*.

Notes

Ensure that your submission file contains the following:

- Correct index values as per the test_1.csv file
- Correct names of columns as provided in the *sample_submission.csv* file

Tips and Tricks

1. The dataset contains outliers, missing values and other data related issues. How one handles these will impact the model performance.
2. Apart from the given features, participants are encouraged to make more features by combining the existing ones.
3. Following process can be followed for model development â€“ data cleaning, feature engineering, feature selection, model training, internal evaluation.
4. You are free to use any ML/DL technique for model development.

Download dataset