Piramal ML Hackathon - IIT Delhi

Mar 01, 2024, 10:00 AM IST (Asia/Kolkata) - Mar 10, 2024, 11:55 PM IST (Asia/Kolkata)

Predict Rate of Interest (ROI) from Bureau Data

Max. score: 100

This problem is no longer available for practice. Apology for any inconvenience!

As a leading finance firm, we frequently encounter data incompleteness that hampers our ability to make conclusive decisions or predictions. One such scenario arises when determining the Rate of Interest for an existing loan from a customer taken with an external Financial Institute. However, we do obtain other loan-related details for such tradelines by subscribing to bureaus such as CIBIL and Experian.

The challenge at hand involves using this data to accurately compute or be closest to the actual Rate of Interest (ROI) across various types of loans. Participants are encouraged to leverage mathematical equations in solving this problem while maximizing the available data. The final solution should be robust enough to handle various loan types, including *Housing Loan, Property Loan, Business Loan, and Personal Loan.*

Task

Participants are expected to ideate on a scalable approach to the problem & submit a document that briefly describes the solution. The document must include a flow diagram & a description of each component. Participants are also expected to build a Machine learning model, based on the data provided, that predicts the rate of interest(ROI) . The ML model must take in data from all sources provided. Unavailability of data points over a subset of the sample should not pose any restriction.

Dataset details

- **train_main_loan.csv:** 69958 x 24
- **train_all_loan.csv:** 1066009 x 19
- **test_main_loan.csv:** 29983 x 23
- **test_all_loan.csv:** 452990 x 19
- **sample_submission.csv:** 5 x 2

The data set consists of financial related data of a customer against a loan & also demographical data belonging to the same individual.

There are multiple files provided.

1. **train_main_loan.csv** : This will have the main loan on which the participant needs to build the model using the info pertaining to that loan along with Date Of Birth, Gender, Occupation type

2. **test_main_loan.csv:** Another file is provided with different set of IDs to validate your model. Once you are able to build model, then you are supposed to get predictions on this dataset using model fit from training dataset. This file would be split between PUBLIC and PRIVATE set on our side. Final evaluation would be done internally on PRIVATE dataset alone.

3. There are other related files: **train_all_loan.csv & test_all_loan.csv** of train & test IDs respectively. These files contain customer's complete credit history including other Live & Closed loans from the past. Participant can use this to extract more credit behavior of the customer and thereby in the model building part.

Name	Description
ACCOUNT_TYPE	Type of Loan taken by the individual
HIGH_CREDIT_OR_SANCTIONED_AMOUNT	This field contains the amount of loan sanctioned. This amount is a whole number and is assumed to be positive.
DATE_OPENED	This is the date of first disbursement of the account. Format is DDMMYYYY.
CURRENT_BALANCE	The entire amount of credit/loan, including the current and overdue portion, if any, outstanding as of the date in the Date Reported and Certified field. This amount must be a whole number and can be either positive (+) specifically with debit balance or negative (-) specifically with credit balance. If the balance is zero, report as 0. If a sign appears in this field, it must appear at the end of the field. If no sign appears, it is assumed to be positive. This field must contain the value zero in case of all closed accounts where balance is zero.
EMI_AMOUNT	Scheduled Payment Amount of each installment
REPAYMENT_TENURE	To be the count of number of months for repayment
AMOUNT_OVERDUE	The amount past due as of the date reported in the Date Reported and Certified field. This amount is a whole number and is assumed to be positive.
PAYMENT_HISTORY_1 & PAYMENT_HISTORY_2	The most recent 36 months of Payment History. Each month displays the Number of Days Past Due (NDPD)/Asset Classification (AC) for that month. The first value in the payment history string is the NDPD/AC, as of the date reported in the Date Reported and Certified field. For accounts that are closed and where the Date Closed field is provided, the first value in the payment history string is the NDPD/AC, as of the date reported in the Date Closed field. If the value for a month is reported as XXX, it means the Reporting Member did not report data for that account for that particular month. If, instead of NDPD, AC is reported, the following values will be displayed: STD = Standard SMA = Special Mention Account SUB = Substandard DBT = Doubtful LSS = Loss ?

	XXX,STD & missing values could be treated as 000, while SMA as 060 & LSS,DBT,SUB to be equivalent to 090
PAYMENT_HISTORY_START_DATE	This is the date of the beginning of the payment history. Format is DDMMYYYY.
PAYMENT_HISTORY_END_DATE	This is the date of the end of the payment history. Format is DDMMYYYY.
OWNERSHIP_TYPE	Joint/Individual/Authorised User/Guarantor
COLLATERALVALUE	Value of Collateral in Rupees
TU_SCORE	Score range is between 300 and 900 for consumers with more than 6 months of credit history. Two zeroes (00) prefix the score when the range is between 300 to 900. A score of -1 is returned for subjects that are new to credit or do not have enough information to score*. A score of 0 is returned for subjects when there is an error
DOB	Customer's Date of Birth
GENDER	Represents the gender of customer
OCCUPATION_TYPE	Customer occupation profile will be provided in this
ACTUAL_ROI	Represents the ROI

Train Set: You are provided with 70,000 Loan level information and customer related original data, along with other tradeline data of other loans for same customer

Test Set: You are provided with 30,000 Loan level information and customer related original data

Submission File Format: Final results should be in .csv file containing only two fields as shown in below sample:

ID	ACTUAL_ROI
A001306408	9.70
A000650025	16.07
B3000514492	11.57
B000157664	14.75
B000101814	20.99

Evaluation metric

The metric to evaluate the performance of the solution is Root mean squared error

```
score = 100*(1-np.sqrt(metrics.mean_squared_error(actual, predicted)))
```

Test data is further divided into Public and Private data equally. Your initial responses will be checked and scored on the public data. The final rankings would be based on your private score which will be published once the competition is over.

Submission guidelines

?

- Please ensure that your final submission includes the following:
    - Solution file (.zip) containing code, computed ROI against each ID of the test set in an csv file (refer submission file format) mentioned above
    - A zipped file containing code & approach (Note that both code and approach document are mandatory for shortlisting)
        - Code: Clean code with comments on each part
        - Approach: Please share your approach to solve the problem (doc/ppt/pdf format). It should cover the following topics:
            - A brief on the approach used to solve the problem.
            - Which Data-preprocessing / Machine Learning ideas really worked? How did you discover them?
            - What does your final model look like? Mention all the steps of the model's journey?
- The index is *ID* and the *ACTUAL_ROI* is the *target* column.
- The size of this submission file must be *29983 x 2.*
- Correct names of columns as provided in the *sample_submission.csv* file

Dear Participants,

After reviewing comments on the hackathon portal and closely monitoring all submissions, we've observed that many entries are receiving a score of 0. This is primarily due to the RMSE value exceeding 1, resulting in a final score of 0.

To address this, on the private dataset at the end of the contest we intend to normalize the RMSE and then compute scores.

As per current method,

Score = 100*(1-RMSE);

Eg:

1) if RMSE > 1 then Score is floored to **0**;

2) if RMSE is between 0 & 1 say 0.75 then Score would be 100*(1-0.75) which is **25**

We shall evaluate private dataset at the end of competition with:

Score = 100*(1-Norm.RMSE)

Norm.RMSE = Square.Root (Avg[(original ROI -predicted ROI)/(original ROI)]^2)

We encourage you to keep submitting your models, however request that you strive to develop more sophisticated models that yield smaller RMSE values. Also make sure you submit a final upload that gives the best result from your iterations.

NOTE: We have also extended our deadline from 5th Mar to **7th Mar**

New Timeline

1st Mar 2024	Open the Hackathon at IIT D (Virtual or Physical)
1st – 7th Mar 2024	Registration and Submission

9th Mar 2024	Private Leaderboard calibration by Piramal team
11th Mar 2024	Top 3 names in the Leaderboard made public
14th Mar 2024	Top 3 Candidates travel to Kurla and meet the team and leaders
15th Mar 2024	In-person Presentation to Piramal Panel

Regarding submission file we highlight again to follow the format given in sample_submission.csv file i.e,:

1. File should contain your prediction Interest Rate against the IDs with these exact column names: **"ID","ACTUAL_ROI"**
2. File should be submitted only in **.csv** format

Thank you for your continued participation and efforts.

Download dataset