Machine Learning Challenge #1

LIVE

Mar 17, 2017, 02:00 AM IST - Mar 27, 2017, 02:00 AM IST

Bank Fears Loanliness

Max. Marks: 100

Problem Statement

The Bank Indessa has not done well in last 3 quarters. Their NPAs (Non Performing Assets) have reached all time high. It is starting to lose confidence of its investors. As a result, it's stock has fallen by 20% in the previous quarter alone.

After careful analysis, it was found that the majority of NPA was contributed by loan defaulters. With the messy data collected over all the years, this bank has decided to use machine learning to figure out a way to find these defaulters and devise a plan to reduce them.

This bank uses a pool of investors to sanction their loans. For example: If any customer has applied for a loan of $20000, along with bank, the investors perform a due diligence on the requested loan application. Keep this in mind while understanding data.

In this challenge, you will help this bank by predicting the probability that a member will default.

18 LIVE EVENTS

Download Dataset

Data Information

There are files given: train, test and submission. Your submission file must adhere to format specified in the given submission file. This data set comprises of information captured in December 2016. Following is the description of variables given:

Variable	Description
member_id	unique ID assigned to each member
loan_amnt	loan amount ($) applied by the member
funded_amnt	loan amount ($) sanctioned by the bank
funded_amnt_inv	loan amount ($) sanctioned by the investors
term	term of loan (in months)

Variable	Description
batch_enrolled	batch numbers allotted to members
int_rate	interest rate (%) on loan
grade	grade assigned by the bank
sub_grade	grade assigned by the bank
emp_title	job / Employer title of member
emp_length	employment length, where 0 means less than one year and 10 means ten or more years
home_ownership	status of home ownership
annual_inc	annual income ($) reported by the member
verification_status	status of income verified by the bank
pymnt_plan	indicates if any payment plan has started against loan
desc	loan description provided by member
purpose	purpose of loan
title	loan title provided by member
zip_code	first three digits of area zipcode of member
addr_state	living state of member
dti	ratio of member's total monthly debt repayment excluding mortgage divided by self reported monthly income
delinq_2yrs	number of 30+ days delinquency in past 2 years
inq_last_6mnths	number of inquiries in last 6 months
mths_since_last_delinq	number of months since last delinq
mths_since_last_record	number of months since last public record
open_acc	number of open credit line in member's credit line
pub_rec	number of derogatory public records
revol_bal	total credit revolving balance
revol_util	amount of credit a member is using relative to revol_bal
total_acc	total number of credit lines available in members credit line
initial_list_status	unique listing status of the loan - W(Waiting), F(Forwarded)
total_rec_int	interest received till date
total_rec_late_fee	Late fee received till date

Variable	Description
recoveries	post charge off gross recovery
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	number of collections in last 12 months excluding medical collections
mths_since_last_major_derog	months since most recent 90 day or worse rating
application_type	indicates when the member is an individual or joint
verification_status_joint	indicates if the joint members income was verified by the bank
last_week_pay	indicates how long (in weeks) a member has paid EMI after batch enrolled
acc_now_delinq	number of accounts on which the member is delinquent
tot_coll_amt	total collection amount ever owed
tot_cur_bal	total current balance of all accounts
total_rev_hi_lim	total revolving credit limit
loan_status	status of loan amount, 1 = Defaulter, 0 = Non Defaulters

Evaluation Metric

Submissions will be evaluated based on AUC-ROC score. For more information about this metric, read here.

Update (17th March, 3pm) - If you've made any earlier submission based on logloss, kindly optimize it for AUC-ROC score. All the existing submissions will be re-evaluated using the AUC-ROC scoring criteria

LIVE EVENTS

Upload Prediction File

Please upload the prediction file in the format as stated in the problem.

Choose file No file chosen

Submit & Evaluate

Upload Source Files

You need to submit a zip or tar archive consisting of a text file explaining your approach, details about feature engineering, tools you used and the relevant source files.

Choose file No file chosen

Upload

COMMENTS (33) SORT BY: Relevance

Join Discussion...

Cancel Post

Kashyap Thacker a day ago
Are there going to be more problems?
5 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin a day ago
Are you asking about upcoming machine learning challenges ?
0 votes ● Reply ● Permalink

Kashyap Thacker a day ago
No I am talking about the current challenge. Will there be more problems? I am asking because we have a limit on total no of submissions
6 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 12 hours ago
For this challenge, there is just one problem (stated above) to solve.
1 vote ● Reply ● Permalink

João Paulo Vasques Camargo da Silva 11 hours ago
in my understanding there is a BIG problem with the score evaluation methodology.
By using the probability to measure the score (instead of right/wrong classification as 0 or 1) there is NO WAY to know the right answer. Not even those who proposed the challenge can know the right probability.
Philosophically, as the problem is stated (predicting the probability that a member will default.), the best possible answer is the simplest one. And there is no need to apply any machine learning technique to discover it.
1 vote ● Reply ● Message ● Permalink

Manish Saraswat Admin 10 hours ago
The predicted probability is scored against true labels. Considering real life used case, to detect defaulters, banks are more interested in determining the probability of a new customer rather than true labels. Another reason being, such situations (defaults) result in imbalanced data where accuracy metric isn't suitable. Hence, metrics like AUC-ROC is preferred.
2 votes ● Reply ● Permalink

João Paulo Vasques Camargo da Silva 10 hours ago
So, just to clarify, you are using, from our submissions against the test data via:
probabilities greater than 0.5 are labelled as 1 and less than 0.5 as 0?
1 vote ● Reply ● Message ● Permalink

Manish Saraswat Admin 7 hours ago
Here you should understand how AUC-ROC scoring works. Let me explain in a simple way: for a true prediction 1, let's say if your predicted probability is 0.14, this metric will punish the low probability and result in low score. On the other hand, considering the same case, if predicted probability comes out to be 0.83, the AUC score will be higher. Therefore, also stated above, for a true label, you have to help the bank to understand the chances of default of a member. Later on, bank can also do members profiling as well.

18 LIVE EVENTS

▲ 1 vote ● Reply ● Permalink

Vivek Kumar a day ago

Hello, here I have some question on understanding of data.
1. What is investor have to do with loan. Is investor is funding some or whole part of loan and bank is just mediator
2. What is batch allotment. On what basis bank if giving same batch to two members
3. Grade and Sub Grade, given by whom and on what basis
4. Whats difference in open_acc and total_acc
5. What is pymnt_plan
6. mths_since_last_record: What is public record

▲ 2 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 11 hours ago

1. Investors assist the bank in funding the loan. Along with bank, the due diligence is also done by a team set up by investors. Then, investor may decide if they want to fund the entire loan or some part of it.
2. Batch allotment code is generated by time. For a set of member joined in a particular time period have been assigned to one batch.
3. These grades are given by the bank. More information is not available. Data exploration might help you.
4. Both variables indicates the available credit lines for a member. Bank uses this information to decide funded amount.
5. It indicates if a member has specified his/her interest payment plan (term).
6. This bank uses public data to measure the activeness of a member. Again, data exploration will help you understand its nature better.

▲ 1 vote ● Reply ● Permalink

Smriti Sheel 2 days ago

In the submission file, is the loan status depicting the probability of that particular member to be defaulter?

▲ 1 vote ● Reply ● Message ● Permalink

Mathusuthan Kannan a day ago

Yes, you need to predict the probability of default.

▲ 0 votes ● Reply ● Message ● Permalink

Himanshu Jaju a day ago

Do we still need to do that? Or its 0 and 1 now ?

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin a day ago

Everything else is as is. You need to predict probability.

▲ 0 votes ● Reply ● Permalink

Himanshu Jaju a day ago

Is there a hidden test data? Also, why do we have to tick against the submissions we want to choose?

▲ 0 votes ● Reply ● Message ● Permalink

Sudip Maji Admin a day ago

Yes Himanshu, while challenge is live we assign score for half of the testcases and after challenge is over, the submission you choose (optional) will run against all the testcases and new score will be assigned. If you do not choose any submission, the best submission will be picked automatically.

▲ 0 votes ● Reply ● Permalink

Sharathkumar Anbu a day ago

How the score in leaderboard is calculated ? I could see some kind of junk numbers for top 2 in leaderboard.

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin a day ago

Current leaderboard ranks aren't stable. We are fixing the leaderboard issue. Within 12 hours, you should see the leaderboard with updated rankings. Till then, validate your models locally. Apologies for the inconvenience.

▲ 0 votes ● Reply ● Permalink

deepak gupta a day ago

Is there any limit on number of submissions?

▲ 0 votes ● Reply ● Message ● Permalink

deepak gupta a day ago

Read it just now in instructions. Its 40.

▲ 0 votes ● Reply ● Message ● Permalink

Gopinath Venkataraman a day ago

Is there any limits in the number of times of submissions. ?

▲ 0 votes ● Reply ● Message ● Permalink

Anthony Gracias a day ago

40 submission limit.

▲ 0 votes ● Reply ● Message ● Permalink

Aditya Guruprasad Rao a day ago

Should the "loan_status" in the output 1 or 0 .. or it is taken as a float value
The sample submission value is set to 0.5 ?

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 11 hours ago

loan_status output should be probability (float value).

▲ 0 votes ● Reply ● Permalink

Mayank Jain 21 hours ago

Will there be more questions for this contest or we need to solve just one problem?

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 11 hours ago

For this challenge, there is just one problem (stated above) to solve.

▲ 0 votes ● Reply ● Permalink

Smriti Sheel 21 hours ago

Why is leaderboard showing my rank as 172 when there are 51 people in the list

▲ 0 votes ● Reply ● Message ● Permalink

Smriti Sheel 21 hours ago

Also, how is the accuracy of the submission being calculated.

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 11 hours ago

Submissions are evaluated based on AUC-ROC score.

▲ 0 votes ● Reply ● Permalink

Manish Saraswat Admin 11 hours ago

Higher your score, your rank will improve. Kindly make your submission again. This anomalous behaviour might be due to update in leaderboard earlier. Rest assured, it's fixed now.

▲ 0 votes ● Reply ● Permalink

Baala Srinivas 10 hours ago

What is test_indessa and train_indessa, should we evaluate the result for anyone of those files or both of the files?

▲ 0 votes ● Reply ● Message ● Permalink

Manish Saraswat Admin 7 hours ago

train_indessa data file should be used for model training. test_indessa file should be used for generating future predictions and making submission.

0 votes ● Reply ● Permalink

Abhijit Annaldas 6 hours ago

any comments on training time? I'm using svm.SVC linear kernel and it's taking really long time.

0 votes ● Reply ● Message ● Permalink

ABOUT US

Blog

Engineering Blog

Updates & Releases

Team

Careers

In the Press

HACKEREARTH

API

Chrome Extension

CodeTable

Developer Profile

Resume

Get Badges

Campus Ambassadors

Get Me Hired

Privacy

Terms of Service

DEVELOPERS

AMA

Code Monk

Judge Environment

Solution Guide

Problem Setter Guide

Practice Problems

HackerEarth Challenges

College Challenges

College Ranking

Organise Hackathon

Hackathon Handbook

Competitive Programming

Open Source

18

LIVE EVENTS

EMPLOYERS

Developer Sourcing

Lateral Hiring

Campus Hiring

Hackathons

FAQs

Customers

REACH US

Ground Floor, Salarpuria Business Center,
4th B Cross Road, 5th A Block,
Koramangala Industrial Layout,
Bangalore, Karnataka 560095, India.

contact@hackerearth.com

© 2017 HackerEarth