☰

**h**

13

LIVE EVENTS

18 : 20 : 03 : 46
DAY   HRS   MIN   SEC

# HackerEarth Machine Learning ...w effective is the STD drug?

LIVE

Mar 31, 2020, 07:30 AM IST - Apr 30, 2020, 07:30 AM IST

INSTRUCTIONS    PROBLEMS    SUBMISSIONS    LEADERBOARD    ANALYTICS    JUDGE

← Problems / Effectiveness of STD drugs

## Effectiveness of STD drugs

Max. Marks: 100

## Problem statement

A new pharmaceutical startup is recently acquired by one of the world's largest MNCs. For the acquisition process, the startup is required to tabulate all drugs that they have sold and account for each drug's effectiveness. A dedicated team has been assigned the task to analyze all the data. This data has been collected over the years and it contains data points such as the drug's name, reviews by customers, popularity and use cases of the drug, and so on. Members of this team are by the noise present in the data.

Your task is to make a sophisticated NLP-based Machine Learning model that has the mentioned features as the input. Also, use the input to predict the base score of a certain drug in a provided case.

## Data

The dataset has the following columns:

| Variable Name | Description |
|---|---|
| patient_id | ID of patients |
| name_of_drug | Name of the drug prescribed |
| use_case_for_drug | Purpose of the drug |
| review_by_patient | Review by patient |
| drug_approved_by_UIC | Date of approval of the drug by UIC |
| number_of_times_prescribed | Number of times the drug is prescribed |
| effectiveness_rating | Effectiveness of drug |
| base_score | Generated score (Target Variable) |

?

# Data description

The data folder consists of the following two **.csv** files:

- **train.csv** - (32165x 7)
- **test.csv** - (10760x6)

The **sample_submission** is described as follows:

```
patient_id,base_score
206461,9.05
95260,8.85
92703,5.26
138000,8.03
```

# Evaluation metric

$$score = 100 * max(0, 1 - RMSE(actual\_values, predicted\_values))$$

**Note**: To avoid any discrepancies in the scoring, you must ensure all the **patient_id** column values in the submitted file must match the values in **test.csv** provided.

Download dataset

## 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 (22) ⟳                                                SORT BY: Relevance ▾

Join Discussion...

Cancel        Po **13**

LIVE EVENTS

---

**Shruti Sarkar** ⚡ Admin  ✎ Edited 9 days ago

Announcement:
Unfortunately, since datasets for this challenge were readily available online, there were several instances of malpractice. Thus, the dataset has been updated and your test has been reset. Please retake the test by downloading the updated dataset again and submiting your predictions for the same.
We apologize for the inconvenience caused.

▲ 7 votes ● Reply ● Permalink

**Vasudev Awatramani** ✎ Edited 10 days ago

Hi, there appears to be a data leak in this competition (from UCI ML Drug Review dataset). I am not discouraging anyone to participate, but just informing to manage expectations and trying to keep the community spirit in an ethical direction.
Nevertheless, we can still discuss solutions or approaches in the forum for a better learning experience.

▲ 1 vote ● Reply ● Message ● Permalink

> **L.Hari Narayanan** 10 days ago
>
> then why have you lifted the rating values from there to get 100 percent accuracy
>
> ▲ 3 votes ● Reply ● Message ● Permalink

**Mustafiz Kaifee Mumtaz** 5 days ago

can someone help me how to solve this problem through any tutorial videos?

▲ 0 votes ● Reply ● Message ● Permalink

> **Subbhashit Mukherjee** 4 days ago
>
> please don't ask such questions here ,this is a competition,not any lecture or course ,after the competition is over just learn by observing the best submission,how are they solving the problem.
>
> ▲ 4 votes ● Reply ● Message ● Permalink

**Vipul Sharma** 9 days ago

There is blatant cheating and data leakage happening in this challenge. Please ban below comment's users and remove them from scoreboard immediately.
They are using UCI and kaggle repositories to get the test data beforehand.

▲ 2 votes ● Reply ● Message ● Permalink

**Kawaljeet Singh Maras** a day ago

I need a definition of the "base_score"
Is it any of the following:
1. Is the base_score specific for a particular drug(which is unlikely looking at the training data)
2. Is the base score according to the drug and disease combination
3. Is it according to the patient who is reviewing
Also what is the use of base_score

▲ 2 votes ● Reply ● Message ● Permalink

**UTSAV PORWAL** 7 days ago

Hey, I am getting 0 score can anyone explain to me?

▲ 0 votes ● Reply ● Message ● Permalink

> **Nitin kshatriya** 6 days ago

?

Since the evaluation is done as 100*max(0, 1- RMSE), if you get RMSE greater then 1, your LB score will be 0.

▲ 1 vote ● Reply ● Message ● Permalink

**L.Hari Narayanan** 9 days ago

Is it possible to change the data set and the problem statement for the challenge. @organisers please look into the same

▲ 0 votes ● Reply ● Message ● Permalink

**Nishant Prabhu** 9 days ago

Is there an issue with the leaderboard? When I submit my predictions, it gives me a score of about 27 but when I open the leaderboard, it says my score is 0. Anyone else facing this issue?

▲ 0 votes ● Reply ● Message ● Permalink

**Kalaiyarasan S** 8 days ago

how to generate a base_score.can anyone tell me plz.

▲ 0 votes ● Reply ● Message ● Permalink

**Baljeet Singh** 7 days ago

hi...does it take time to evaluate score or it's an instant thing?
Also, my base score is in 6 decimal places. Does it have to be 2 decimal places as in the format above?

▲ 0 votes ● Reply ● Message ● Permalink

> **Subhajeet** ⚡ Admin ✎ Edited 6 days ago
>
> Hi, evaluation of score is an instant thing.
> For this question the formula to calculate the score is 100*max(0,1-RMSE(actual_values,predicted_values)) and is independent of number of decimal places in the prediction.
>
> ▲ 0 votes ● Reply ● Permalink

> **Subbhashit Mukherjee** 4 days ago
>
> Yes it should be in 2 decimal places
>
> ▲ 0 votes ● Reply ● Message ● Permalink

**Sammed Shantinath Kagi** 5 days ago

Dear Admin,
Does the public leaderboard has the scores on the updated test set? If not could you completely reset the leaderboard so that the participants can run tests on the new test set and update their submission.

▲ 0 votes ● Reply ● Message ● Permalink

**D Moti Raj** 4 days ago

Are Indians eligible for the prize?

▲ 0 votes ● Reply ● Message ● Permalink

**Manimaran Paneerselvam** 3 days ago

I thought I could directly use wget download the dataset via command line, but it is not working, does anyone know any other alternative?

▲ 0 votes ● Reply ● Message ● Permalink

**Utkarsh Sharma** 2 days ago

What is the range of base_score values ? Is it same as effectiveness_rating ? (ie 1-10)

▲ 0 votes ● Reply ● Message ● Permalink

**Manish Prajapati** 16 hours ago

I can do it

▲ 0 votes ● Reply ● Message ● Permalink

**Pavan Kumar** 11 hours ago

Hello Shruti, My best score is higher than the score showing on LeaderBorad. How can i see my best highest score on Leaderborad? It's been quite a while i have uploaded my best scored file.

⌃ 0 votes ● Reply ● Message ● Permalink

**Ashish Tiwari** 10 hours ago

It there some kind of huge difference in test and train data? I am getting a RMSE of ~.30 on train but for test my score is 0(i.e RMSE > 1. Is anyone else facing this huge gap?

⌃ 0 votes ● Reply ● Message ● Permalink

**13**

LIVE EVENTS

## Resources

Tech Recruitment Blog

Product Guides

Developer hiring guide

Engineering Blog

Developers Blog

Developers Wiki

Competitive Programming

Start a Programming Club

Practice Machine Learning

## Solutions

Assess Developers

Conduct Remote Interviews

Assess University Talent

Organize Hackathons

## Company Service & Support

About Us

Press

Careers

Technical Support

Contact Us

+1-650-461-4192

contact@hackerearth.com

f 🐦 in

▶

?