Xerox Mortality Prediction Challenge

Problem Statement

Xerox Data Challenge: Round 1

UPDATES

October 12. 03:00 PM. File 'id_time_labs_train.csv' in the training dataset has been updated. Extra column from previous file removed. **Please use this new file from the same download link. Please do not use the file uploaded earlier.**

October 12. 03:00 PM. Description of vitals as given in **List 1: Vitals and Units** below has been corrected. There is no change in the data file.

The aim of this challenge is to predict risk of death (mortality) in patients admitted to intensive care units (ICU) within hospitals.

The data consists of 5990 (simulated) patient records where each **patient record** has the following **variables**:

1. ID: a unique identifier for each patient
2. Age
3. 6 Vitals: see list 1
4. 25 Labs: see list 2
5. Timestamps: measurement time relative to first measurement for patient
6. ICU flag: indicates whether a patient is in ICU or not at a given time
7. Mortality label: indicates whether a patient survived or died (the label or outcome variable)

A **timestamp** shows when a **vital or lab measurement** is made relative to the first measurement. There can be one or more measurements at each timestamp. The first timestamp for each patient is always 0 (zero). Subsequent timestamps show time elapsed from this first timestamp in seconds. Regular intervals between measurement timestamps *cannot* be assumed. The length-of-stay of each patient is different and so the number of measurements (hence timestamps) are different in each patient record.

NOTE Timestamps are not synchronized across different patients. Timestamps only show relative differences in time within each patient. For example, timestamp 0, does not indicate the same time across different patients. It only indicates the first measurement time of each patient (which may be different for different patients). All other timestamps are with respect to the first measurement for each patient. **Do not compare timestamps across different patients.**

Vitals and Labs are **time-series variables** associated with timestamps. Each vital or lab variable may be measured multiple times (at different timestamps) for a patient. Not all variables are measured at all timestamps. Some variables may not be measured at all in a patient record.

At each timestamp the **ICU flag** indicates whether the patient is inside an ICU or outside (in the hospital).

Every patient in this dataset visits the ICU exactly once during the hospital stay. Thus the period when the patient is in the ICU is indicated by a series of consecutive timestamps where the ICU flag is set to 1. Patient records have measurements both inside as well as outside the ICU.

Each patient has a single age value. Age values for all patients are given.

The **mortality indicator**, a single label per patient, indicates whether the patient survived or died at the end of the hospital stay. Time of death is not provided. For patients that die, the time of death is after the last measurement recorded.

List 1: Vitals and Units

1. Systolic Blood Pressure in mmHg
2. Diastolic Blood Pressure in mmHg
3. Heart Rate in bpm
4. Respiration Rate in bpm
5. Oxygen Saturation in %
6. Temperature in Fahrenheit

List 2: Lab investigations (Labs) and Units of measurement

1. Arterial blood PH in ph
2. Partial Pressure of Carbon dioxide (PaCO2) in mmHg
3. Partial Pressure of Oxygen (PaO2) in mmHg
4. Sodium in mmol/L
5. Potassium in mmol/L
6. Bicarbonate in mmol/L
7. Blood Urea Nitrogen in mg/dL
8. Serum Creatinine in mg/dL
9. WBC Count in x 10^3/µL
10. Hematocrit %
11. Platelet Count in x 10^3/µL
12. Bilirubin in mg/dL
13. Urine Output in ml
14. LDL Cholesterol in mg/dL
15. Lactic Acid in mmol/L
16. Troponin I in ng/ml
17. Troponin T in ng/ml
18. Random Blood Glucose in mg/dL
19. Fasting Blood Glucose in mg/dL
20. Fraction of Inspired Oxygen (FiO2) in %
21. Albumin in g/dl

22. Alkaline Phosphatase in IU/L
23. Alanine in IU/L
24. HDL Cholesterol in mg/dL
25. Magnesium in mg/dL

The hyperlinks point to websites giving more information about the investigation. These are for your information only.

Patient records are divided into 3 sets: train, validation and test sets. Train set contains 3594 patient records. Validation and test sets each contain 1198 patient records. Train, Validation and Test sets will be made available to participants at different times during competition. More details below.

Data format

The **train set** consists of 4 files:

1. id_time_vitals_train.csv
2. id_time_labs_train.csv
3. id_age_train.csv
4. id_label_train.csv

DOWNLOAD THE TRAINING DATASET HERE: https://s3.amazonaws.com/istreet-assets/PCZGuJ03isy6UZpXT4v7kw/Traning_Dataset.zip

File **id_time_vitals_train.csv** contains timestamped vitals measurements of all the train patients along with the ICU flag for each timestamp. The column headers in the file are:

ID, TIME, V1,...,V6, ICU

File **id_time_labs_train.csv** contains timestamped labs measurements of all train patients. The column headers in the file are:

ID, TIME, L1,..., L25

Each row contains measurements made for a patient at a given timestamp. Measurements that are not made at a timestamp has the value 'NA'. The first timestamp for a patient is always 0. Subsequent timestamps show time elapsed from this first timestamp in seconds. V1 – V6 indicate the six vital measurements and L1 – L25 indicate 25 lab measurements (given in lists 1 and 2 above).

Some lines in the file may contain only 'NA's after the ID and timestamp. This could happen in the vitals file if there are only lab measurements at that timestamp and no vitals are measured. Similarly it could happen in the labs file if there are only vitals measured at that timestamp. Also note that the values may not look like real world values due to the noise added intentionally. E.g. Systolic Blood Pressure values are integers in real world, but may be float values in the dataset.

The ICU variable can take two values: 0 indicates the patient is not in ICU at the timestamp and 1 indicates the patient is in ICU at the timestamp.

File **id_age_train.csv** contains one line per patient, each line containing ID and age. Column headers are: ID, AGE

File **id_label_train.csv** contains one line per patient, each line containing ID and mortality label. Label 0

indicates survival and label 1 indicates death. Column headers are: ID, LABEL

Validation and test sets

The **validation set** consists of three files: id_time_vitals_val.csv, id_time_labs_val.csv, id_age_val.csv. The format is the same as that of the corresponding train files (with different patient IDs and data, of course!).

The labels for the validation set will **not** be provided. Once your model script/code (having training and testing part) generates output.csv on validation dataset, you can upload output.csv in HackerRank to compute the performance metrics and get scores on the leaderboard. More details below.

The test set will not be provided until the last 3 days of the challenge. It consists of three files again: id_time_vitals_test.csv, id_time_labs_test.csv and id_age_test.csv following the same formats as those of train and validation sets. The finalists of round 1 will be selected based on the performance of their models on this test dataset.

Prediction

Prediction must be made for each patient **only when the patient is in ICU** (i.e. when the ICU flag is set to 1). The prediction must be done in an **online manner**, that is, at a given timestamp the model can use any of the past data and data for the current timestamp (for that patient) to make a prediction for that timestamp. Predictions are to be made at **every** measurement timestamp while the patient is in ICU.

Thus, **each patient has a sequence of predictions**. Each prediction is a label 0 or 1 and the number of predictions for each patient is not more than the number of rows in the data, for the given patient, where ICU flag == 1.

For example, assume a patient has measurements at timestamps 100, 300, 500 and 1000 during ICU stay. There should be 4 predictions for this patient, one at each timestamp. At timestamp i, measurements from any timestamp **before or at i** can be used for prediction. But no data after timestamp i should be used for predicting at timestamp i. In the above example, at timestamp 500, data from measurements at 100, 300 and 500 may be used for prediction but the data at timestamp 800 cannot be used at timestamp 500.

Entries that do not perform online prediction in the manner described above will be disqualified.

Evaluation

We obtain a **final prediction** per patient as follows: If the sequence of predictions for the patients contains only zeros, then the final prediction is 0, otherwise 1.

Examples:

1. Prediction sequence: 0 0 0 1 0 0, final prediction: 1
2. Prediction sequence: 0 1 0 1 0 0, final prediction: 1
3. Prediction sequence: 0 0 0 0 0 0, final prediction: 0

Prediction time is only defined for patients whose final prediction is 1. It is the difference between the last timestamp (i.e. the last measurement made for the patient) and first timestamp with a prediction of 1.

We obtain patient-wise performance metrics. Each of TP, TN, FP and FN denotes number of patients for whom the corresponding condition, given below, holds.

1. True Positives (TP): Mortality Label == 1 and Final Prediction == 1

2. True Negatives (TN): Mortality Label == 0 and Final Prediction == 0
3. False Positives (FP): Mortality Label == 0 and Final Prediction == 1
4. False Negatives (FN): Mortality Label == 1 and Final Prediction == 0

Sensitivity = TP/(TP+FN), Specificity = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN)

Leaderboard score computation

Teams will be ranked based on the **Final score** that is in the range 0 to 100 (where, 0 is minimum and 100 is maximum score).

Failure cases: The Final Score is set to 0 if any one of the following is true.

1. Specificity is less than 0.99
2. Sensitivity is 0
3. Median prediction time is less than 5 hours

If there are no failure cases, then the final score is computed as the weighted sum of Sensitivity score, Specificity score, and Median prediction time score defined below.

Final_score = 75*Sensitivity_score + 20*Median_pred_time_score + 5*Specificity_score

Sensitivity_score = Sensitivity = TP/(TP+FN).

Specificity_score = (Specificity - 0.99)*100, where Specificity = TN/(TN+FP); Note that Specificity_score is assigned for specificity in the range 0.99 to 1.

Median Prediction Time score is assigned for median prediction time in range 5 hrs to 72 hrs. Median prediction time above 72 hrs is considered as 72 hrs only for computing the final score.

Median_pred_time_score = median_pred_time_clipped_at_72 / 72; where

if median_pred_time < 72 hrs: median_pred_time_clipped_at_72 = median_pred_time
else: median_pred_time_clipped_at_72 = 72 hrs.

Final Score will also be 0 if any of the following conditions is not satisfied.

1. All the prediction labels are either 0 or 1.
2. Predictions for all patients in validation/test are provided.
3. For each patient, predictions are present for all timestamps where ICU flag == 1.
4. For each patient, predictions are present for only those timestamps where ICU flag == 1.

Our evaluation metrics are chosen to measure the ability of the model to identify high risk patients at 1% false positive rate (99% specificity) and to identify high risk patients as early as possible. Participants should use suitable strategies (like setting sample weights/misclassification costs based on the class during training) to achieve 0.99 or higher specificity. Note that an increase in sensitivity by 0.01 increases the final score by 0.75, increase in specificity (above .99) by 0.001 increases final score by 0.5 and increase in median prediction time by 1 hour increases the final score by 0.27.

The leaderboard will show **final score, sensitivity, specificity, accuracy** and **median prediction time** measured over the validation set.In the last 3 days of the challenge, another ‘final leaderboard’ will be set up that will show final score, sensitivity, specificity, accuracy and prediction time measured over the test set. The top 10 finalists will be chosen based on Final score on test data. See timeline below for more details.

Deliverable

Every team must upload code and a file named **output.csv** to get the final score.

The submitted code should run from the command line taking three arguments which are test/validation filenames containing vitals, labs and age of patients (following the formats of the validation set files) respectively. Input filenames may be assumed to be present in the same directory. These filenames will be passed as command line arguments to the code.

All the code (for training and prediction) must be present in a single file. The train filenames can be hardcoded within the program (assuming them to be present in the same directory).

There are no timing constraints on the execution of the code. The code must generate an output file named **output.csv**.

Example: if the submitted code is run_model.py, it should run on the command line as follows:

$python run_model.py id_time_vitals_test.csv id_time_labs_test.csv id_age_test.csv

Participants have to run their code locally and create output.csv. Note that you have to submit **both** output.csv and the code. We will compute the performance using the uploaded file output.csv.

Each line of output.csv must have the following format.

Patient ID,Prediction timestamp,Prediction at that timestamp\n

Example (the first few lines):

3595,0,0

3595,115,0

3595,2118,0

3595,4197,0

...

The timestamps must be present in id_time_vitals_test.csv (or id_time_vitals_val.csv in the case of validation) and the icu label must be 1 at these timestamps. The number of predictions per patient must be exactly equal to the number of timestamps at which the patient is in ICU (with ICU flag set to 1). Predictions must be present for all patient IDs (all patients visit the ICU during hospital stay). **Output files that do not follow these rules will be given a score of zero.**

Additional documentation describing the entire approach taken will be required from the top 10 finalists. If any mistakes/bugs are found in the code then the entry will be discarded and the next team with highest performance in the leaderboard will be considered for the top 10.

Timeline

Oct 12: Training dataset available on Hackerrank (can be downloaded here: https://s3.amazonaws.com/istreet-assets/PCZGuJ03isy6UZpXT4v7kw/Traning_Dataset.zip)

Oct 19: Validation dataset (without labels) available on Hackerrank (can be downloaded). Participants can run their models on this dataset and check performance using script at Hackerrank **up to a maximum of 100 times per team**. Leaderboard updates with every performance check.

Oct 22: Test dataset (without labels) available on Hackerrank (can be downloaded). Participants can run their models on this dataset and check performance using script at Hackerrank **up to a maximum of 3 times per team**. Final Leaderboard updates with every performance check.

Oct 25: Challenge ends. Final leaderboard shows top 10 teams.

Programming Language

Only open-source languages are allowed, that can run on Linux (Ubuntu) operating systems. Python is the preferred choice. Proprietary software, like MATLAB, are not allowed.

Eligibility

Participants must be registered students **within India** in

- pre-final and final year of an undergraduate program or
- a Master's/PhD program

Participant teams can have up to 3 members.

Ineligible entries will be disqualified and not considered for the top 10 finalists.

The dataset has been simulated from real patients' data. No measurement in the dataset belongs to any real patient. Any match with a real patient's measurement is purely coincidental.

The objective of this competition is similar to a competition held by Physionet in 2012: http://physionet.org/challenge/2012/. The data is not the same although many of the measurement variables used are same. Participants are encouraged to see the approaches used by winners of the challenge that are described in papers here: http://physionet.org/challenge/2012/papers/.

IMPORTANT DETAILS REGARDING SUPPORT:

For any queries relating to the problem statement or the HackerRank platform, drop a note in the "discussions" forum in HackerRank.

Your query will be answered between 9 am to 5 pm IST.

Other alternative means for reaching out to us are:

1. Drop a message on the Xerox Facebook page: https://www.facebook.com/XRCIOpen2016
2. Drop an email to XRCI.Open@xerox.com

Explanation

Please upload all code (for training and prediction) in a single file and the output file with name output.csv.