DS ASSIGNMENT

Data Description:
file format:

train.parquet
test.parquet
sample_submission.csv
final_submission.csv

Health records of patients diagnosed with a certain disease are present in both train and test files, various categories of health records like diseases diagnosis, tests, symptoms, and drug treatments are captured in the records along with the date of occurrence for each patient

Patient-Uid: Unique Identifier for each patient
Date: Date on which the patient encountered the event
Incident: Actual event encountered

Total Number of unique patients present in train.parquet 27K

An event called “Target Drug ” is present in the incident column of the training set for around 9K patients at least once in their journey.

Datalink: https://drive.google.com/file/d/1oHnw-M9jOshB3WkbKrMBWepjlEHAdwA1/view

Problem Statment:

1. Drugs are developed in therapeutic areas to boost the patients’ condition against chronic and terminally ill diseases, the “Target Drug” is one such and it can boost the patients’ health without making them dependent on the other drugs that can lead to life-threatening side effects. The objective is to build a predictive model to estimate if a patient is eligible for the first prescription of the “Target Drug” in the next 30 days so that the physician who is treating the patient could be informed on the better treatment choices.

    A. Come up with the right strategies to create positive and negative data samples from the data(avoid any biases while sampling the data to build a good predictive model)
    B. Build a predictive model by doing the appropriate feature engineering(eg: frequency-based, time-based features, e.t.c). The predictive model could also leverage Deep Learning based techniques.
    C. Evaluate the model on your own validation set and come up with the right threshold to minimize false positives and false negatives

DS ASSIGNMENT

D. Some of the patients present in the test file are eligible for the drug prescription within a month and some of them are not, using each patient's historical data predict if he/she is eligible for the "Target Drug"
E. Each patient-uid should be labeled with a binary value of 1 or 0 using the built model, 1 is considered as eligible for the "Target Drug" in the next 30 days and 0 considered as un-eligible

F. Submit the final_submission.csv file by filling the predicted label for each patient in the "label" column, make sure you don't change the format, file name, and the order of patients to be consistent with evaluating algorithm(refer sample_submission.csv).
G. The evaluation metric for the assignment is F1-Score(candidates with the highest F1 score would be prioritized)

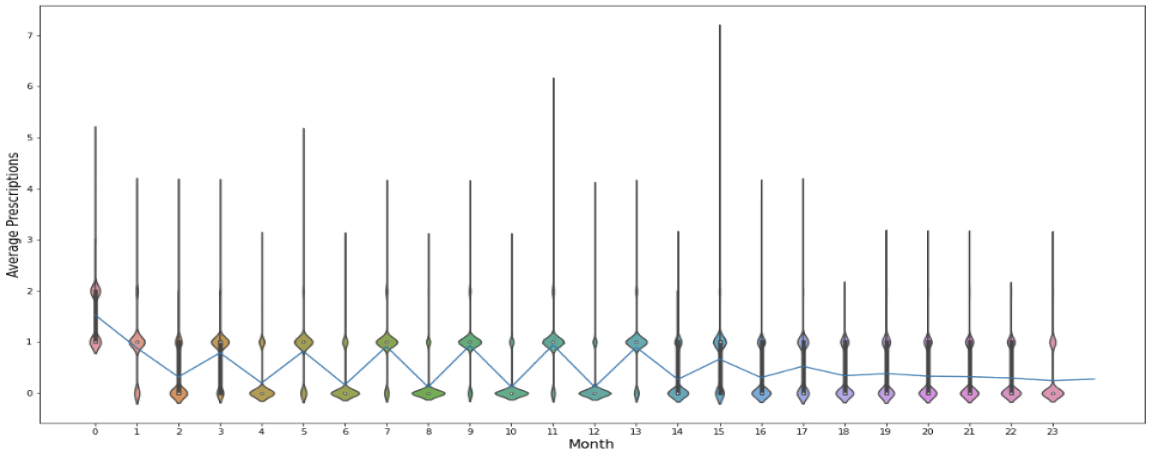
*** Note ipython notebooks should be submitted along with the results by documenting the assumptions/process along the way(better documentation is also given weightage in the final score) and make sure the results are reproducible within an acceptable range of standard deviation

2. Drugs are administered to patients in certain patterns to get the best outcome, but for various reasons, patients drop off early without finishing the regimen. Assume that the patients are required to stay on the "Target drug" for as long as one year, come up with your own analysis using different visualization techniques on how the drop off rate is present each month after the first prescription is made and what are the events contributing to the drop-off

   Should submit the ipython notebooks along with visualizations with some text describing your insights

3. A drug could be administered to patients in different patterns. From the data, extract dominant patterns in which the "Target Drug" is being administered across the patients using clustering techniques.
   Visualize the prescription patterns with time on X-axis and prescriptions on Y-axis for each of the patterns you are able to extract(following is an example of a prescription pattern, where a prescription is made at least once in the first two months followed by one prescription for every two months)

DS ASSIGNMENT



Should submit the ipython notebooks along with code and visualizations of prescription patterns

Submission guidelines:

1. Submissions for questions 2 and 3 are considered only if the results for question 1 are uploaded.
2. Results should be reproducible and the code should be re-runnable.
3. The assignment will be evaluated for 15 points, 10 points for problem statement 1, 3 points for problem statement 3, and 2 points for problem statement 2

4. Uploading code

Maintain separate Jupyter notebooks for each of the problem statements, and naming convection for the Jupyter notebooks should be as below

a. 001.ipynb ==>problem statement 1
b. 002.ipynb ==>problem statement 2
c. 003.ipynb ==>problem statement 3

Note: You can create multiple notebooks for a single problem statement, example 001.1.ipynb

Use the following convection for documenting the process, steps and results.

a. 001.pdf ==>problem statement 1
b. 002.pdf ==>problem statement 2
c. 002.pdf ==>problem statement 3

Note: You can also describe what more could be done if you have more time

Package all the files in a zip format and name it in the following structure and upload it in the allocated field in the google form
“yourname_structureddata_solution.zip“

DS ASSIGNMENT

5. **Uploading results**
   Problem statement 1 -upload the final_submission.csv to the google form in the allocated field and we will consider 001.ipynb and 001.pdf
   Problem statement 2- we will consider 002.ipynb and 002.pdf
   Problem statement 3- we will consider 003.ipynb and 003.pdf

**Note: submissions are expected to be in the order of hundreds, please comply with guidelines for automating the validation.

WE WISH YOU GOOD LUCK