Structured Data Assignment

Data Description -

The folder shared with you contains following four files

1) Train.parquet - Dataset to be used for training
2) Test.parquet - Dataset to be used for testing
3) Sample_submission.csv - a sample csv file showing how the output should be
4) Final_submission.csv - csv file to be submitted finally after generating the output

Brief Description of the Dataset -

The dataset in question contains a comprehensive collection of electronic health records belonging to patients who have been diagnosed with a specific disease. These health records comprise a detailed log of every aspect of the patients' medical history, including all diagnoses, symptoms, prescribed drug treatments, and medical tests that they have undergone. Each row represents a healthcare record/medical event for a patient and it includes a timestamp for each entry/event, thereby allowing for a chronological view of the patient's medical history

The Data has mainly three columns

1) Patient-Uid - Unique Alphanumeric Identifier for a patient
2) Date - Date when patient encountered the event.
3) Incident - This columns describes which event occurred on the day.

Patient-Uid	Date	Incident
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2015-09-22	DRUG_TYPE_7
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-04-13	SYMPTOM_TYPE_2
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-05-02	DRUG_TYPE_7
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-11-23	DRUG_TYPE_0
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-11-23	SYMPTOM_TYPE_0

Above Table represents a patient journey for patient a0db1e73-1c7c-11ec-ae39-16262ee38c7f This patient took drug of type 7 on 2015-09-2022, then the patient had a symptom of type 2 on 2018-04-13, and then again the patient took drug of type 7 on 2018-05-02 and so on

There are total 27K unique patients present in train.parquet

Similar to drug_type_7, there is also drug called 'Target Drug", this drug is of interest for the assignment, there are total 9K patients in train.parquet who have taken "Target Drug" atleast once.

Link to download data
https://drive.google.com/file/d/1oHnw-M9jOshB3WkbKrMBWepjIEHAdwA1/view

Problem Statement

Problem 1 - The development of drugs is critical in providing therapeutic options for patients suffering from chronic and terminal illnesses. "Target Drug", in particular, is designed to enhance the patient's health and well-being without causing dependence on other medications that could potentially lead to severe and life-threatening side effects. These drugs are specifically tailored to treat a particular disease or condition, offering a more focused and effective approach to treatment, while minimising the risk of harmful reactions.
The objective in this assignment is to develop a predictive model which will predict whether a patient will be eligible*** for "Target Drug" or not in next 30 days. Knowing if the patient is eligible or not will help physician treating the patient make informed decision on the which treatments to give.

*** - A patient is considered eligible for a particular drug when they have taken their first prescription for that drug. Below table gives an example

Patient-Uid	Date	Incident
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2015-09-22	DRUG_TYPE_7
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-04-13	SYMPTOM_TYPE_2
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-05-02	DRUG_TYPE_7
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-11-23	TARGET DRUG
a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2018-12-30	TARGET DRUG

In above example, we see that the patient took his first prescription of "Target Drug" on 2018-11-23, so it can be assumed that on this particular day the patient became eligible for "Target Drug".

Please follow below steps for developing the model -

A. Come up with a positive and negative set for developing the model, here the positive point is the patient who has taken 'Target Drug". Make sure you are also taking into account the time aspect while coming up with a positive & negative set because the aim is to predict 30 days in advance whether a patient is going to be eligible or not.
B. Come up with the right kind of feature engineering for developing your model. The features can be frequency-based, time-based etc. If possible can also leverage deep learning techniques
C. Evaluate the model on validation set & come up with the right strategy to reduce false positives & false negatives.
D. Once you have developed your predictive model, use your model to generate predictions for patients in test.parquet, each patient in test.parquet should be labelled as 1 or 0 using your predictive model and the final generated predictions should be submitted in final_submission.csv
E. The evaluation metric for the assignment is F1-Score(candidates with the highest F1 score would be prioritised)

   *** 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.

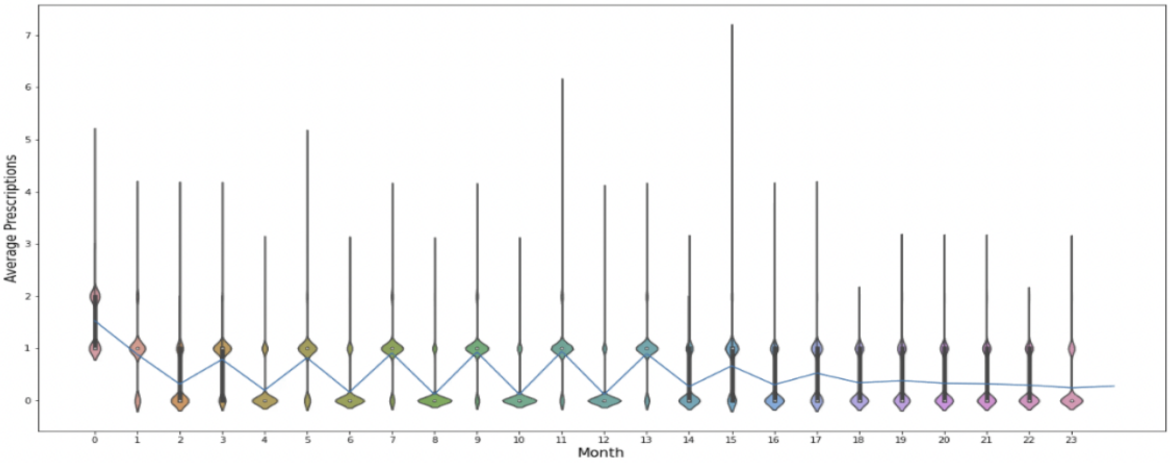
Problem 2 - Drugs are generally administered/prescribed by the physicians for a certain period of time or they are administered at regular intervals, but for various reasons patients might stop taking the treatment . Consider following example for better understanding

Let's say you get a throat infection, the physician prescribes you an antibiotic for 10 days, but you stop taking the treatment after 3 days because of some adverse events.

In the above example ideal treatment duration is 10 days but patients stopped taking treatment after 3 days due to adverse events. Patients stopping a treatment is called dropoff. We want to study dropoff for "Target Drug", the aim is to generate insights on what events lead to patients stopping on "Target Drug".
Assume ideal treatment duration for "Target Drug" is 1 year, come up with analysis showing how drop-off rate is, dropoff rate is defined as number of patients dropping off each month. Then come up with analysis to generate insights on what events are driving a patient to stop taking "Target Drug".

Problem 3 - A drug is generally administered to a patient in certain patterns or in regular intervals of time. For example Chemotherapy which is drug treatment in case of Cancer is generally given to patients in an interval 3-4 weeks, i.e. every 3-4 weeks patients are administered with the drug.

Similarly to Chemotherapy, “Target Drug” is also administered/prescribed in certain patterns, we want to analyse in what patterns “Target Drug“ is administered/prescribed to patients, there might be multiple patterns in which “Target Drug” is administered/prescribed, come up with an analysis which to extract the dominant patterns in the data using clustering or other unsupervised techniques.
Visualise the prescription patterns with time on X-axis (month) and prescriptions on Y-axis for each of the patterns you are able to extract(Below 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).



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

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“

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