NTT Data 2022 AI Hackathon – Improve outcomes for detection and management of diseases

Atrij Talgery/Team ATR21

Started on 14 April, 2022

Contents

Chapter 1

Introduction

1.1 Sponsor's message

Following is the problem statement as provided by the Hackathon sponsor:

> Health and wellbeing are central to the human experience. Whoever you are, wherever you live, being and feeling healthy is a universal aspiration. Yet, nine million people die every year without proper healthcare services. Access to comprehensive, quality healthcare is critical. With more than 50 years of experience in healthcare services, NTT DATA understands the gravity of the situation, and our goal is to improve the healthcare journey for everyone. All while reducing costs and enabling new forms of value to flourish.
>
> We believe artificial intelligence (AI) can help improve outcomes for detection and management of diseases, whether physiological, genetic or psychological. For instance, someone with crippling social anxiety may ask a friend to make phone calls to doctors on their behalf because they're unable to. If, however, people in that position had access to tools like emotional support bots, phone-based cancer screenings or primary care virtual doctors (using natural language processing), it could help them reassert control of their lives.
>
> In this hackathon, contestants will focus on applications of AI for:
>
> - Primary care
> - Early disease detection
> - Patient support for disease management
> - Access to the right healthcare
> - Emotional support
>
> You should capture and label the data you use to train the AI models. Once you clear the round 1 MCQ, you'll be able to access more information on the problem statement.
>
> We welcome you to take this opportunity to test your limits and find out what you're capable of.

1.2 Methodology used

We use the agile planning and development methodology which includes:

- Understanding the theme/initiative.
- Breaking the initiative down into epics/projects.
- Deciding on tasks/user stories for each project.

Chapter 2

Agile structure description

We use the traditional agile structure for our project as described below:

2.1 Initiative/Theme

The initiative theme is taken directly from the sponsor's message.

2.2 Epic/Project(s)

The theme lends itself to the typical problems listed. For this hackathon, we shortlist `early stage disease detection` with a simple bot interface as our project. Such a solution would help patients seeking easy access to healthcare.

2.3 Tasks/User stories

This being a data science project is different from typical software development projects. There are no user stories other than those that dictate the user interface and deployment.

Rather, the iteration structure is repetitive and is as follows.

- Lit. review and dataset gathering.(LIT)
- Data Exploration and analysis.(EDA)
- Algorithm devpt., data modeling (ADDM)
- Result Analysis and Evaluation (RAE)
- Review (REV)
- Deployment (DEP)

Each iteration will progressively do all the steps above to give a minimum viable product (MVP) to start with, that improves over each iteration. Each iteration will also progressively move towards more and more advanced versions of each of the steps. For instance, we could keep on improving the EDA/ADDM/DEP steps across iterations.

Additional backlog items

- Create deployable models for serialization and storing on disk.
- Design user interface for the web app.
- Develop the Flask web app
- Deploy and test

We use week-long sprints to iterate over our tasks and backlog items.

2.3.1 Define project concept and explore feasibility 20220415

- Explore availability for use of datasets for training deep learning model on health conditions.

https://www.kaggle.com/datasets/aanya08/hypertension-data https://www.kaggle.com/datasets/mazharkarimi/ disease-and-stroke-prevention https://www.kaggle.com/datasets/ruckdent/hypertension-dataset-india

- Explore training neural network with variable length input sequences.

  Example: We might have hypertension data with missing age and/or weight, which should still be capable of training our model.

- Explore the possibility of implementing a simple rule-based bot to interact with and guide the user.

2.3.2 User Story: Deployment

The deployment should facilitate access over different computing devices including mobiles. The interface should be easy to learn.

2.3.3 Task and to-do lists

Sprint 1

1. (LIT)Finalise the cardio-vascular dataset for use.
2. (EDA)Do basic EDA on the dataset(incl. imputation, feat selection)
3. (ADDM)Explore the logistic reg. and random forest models.
4. (RAE)Analyse, evaluate the classifiers.
5. (REV)Review iteration and update backlog for next iteration.
6. (DEP)Basic deployment in JupyterLab.

Sprint review

```
The cardiovascular dataset was analysed(EDA) and a random forest model was
fitted around the data. Hyperparameter tuning was done using Grid Search CV.
Deployment of the model is still within JupyterLab.
```

Sprint 2

1. (LIT)Finalise the diabetes prediction dataset for use.
2. (EDA)Do basic EDA on the dataset(incl. imputation, feat selection)
3. (ADDM)Explore the logistic reg. and random forest models.
4. (RAE)Analyse, evaluate the classifiers.
5. (REV)Review iteration and update backlog for next iteration.
6. (DEP)Basic deployment in JupyterLab.

Sprint review

```
The diabetes dataset gives us very high accuracy with a random forest classifier.
But this is small consolation because this is a highly imbalanced dataset and
consequenty, precision and recall matter more than accuracy. We also found
about 5% of the data to be mislabeled. The next sprint should perhaps focus on
```

imputing labels and improving the classification quality of both this and the cardio dataset.

Appendix

Datasets

- https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
- https://www.kaggle.com/datasets/houcembenmansour/predict-diabetes-based-on-diagnostic-measures

---