Contents

Problem Description 2
About Data 2
Objectives for the Hackathon 2
Evaluation Metric 3

Submission Timeline 4

Broadband Outage Detection

Problem Description:

India is seeing an explosion of new competitors in the Broadband space. 'India Broadband' is a company that is now seeing a lot of customer churn due to customer dissatisfaction because of broadband outages.

The company has now curated a dataset, where it tracks several variables that it believes impact the `outage_duration`. They have tracked three different outage durations, `0` for no outage, `1` for short outages that last anywhere between a few minutes and a maximum of 2 hours, and `2` for long outages that can last from 2 hours to sometimes even a couple of days.

You will now have to use these metrics that the company has tracked to create a machine learning model that will be able to predict the `outage_duration` so that the company can better handle outages and improve customer satisfaction and therefore reduce customer churn.

In this hackathon, you will now have to use these metrics that the company has tracked to create a machine learning model that will be able to predict the `outage_duration` so that the company can better handle outages and improve customer satisfaction and therefore reduce customer churn.

About Data:

There are 7 CSV files provided to us, they are described below:

- `**train_data.csv**`: It has a unique event `id` for each observation of the `outage_duration` in a particular `area_code`

- `**test_data.csv**`: Similar to the train dataset, we are provided with an `id` and an `area_code`, we are expected to predict the `outage_duration` for each of the records. (This will be provided to you later on 4th July)

- `**broadband_data.csv**`: For each of the event `id`s mentioned in the `train_data.csv` and `test_data.csv` files and also some additional `id`s there is a record of the `broadband_type` that is stored in the dataset. There are `10 different types` of broadbands that are observed in the dataset

- `**outage_data.csv**`: For each of the event `id`s mentioned in the `train_data.csv` and `test_data.csv` files and also some additional `id`s there is a record of the `outage_type` that is stored in the dataset. There are `5` different `outage_type`'s recorded in the dataset.

- `**report_data.csv**`: For each event `id` there are `log_report_type` and `volume` columns are recorded. `log_report_type` is a type of the recorded report generated by a technical team member after evaluating the outage. `volume` is the volume of data handled in the area at the time of report in custom company specific units.

- `**server_data.csv**`: For each of the event `id`s mentioned in the `train_data.csv` and `test_data.csv` files and also some additional `id`s there is a record of the `transit_server_type` that is stored in the dataset. Transit Servers handle the requests and responses of the customers.

- `**sample_submission.csv**`: The format of CSV file required for submission to the evaluation backend. **(Please remember that the prediction file which you are going to upload to tool, to check out what is your score should be of the same format as this file)**

The different `broadband_type`'s are given below:

```
{
        broadband_type_8  : 'ADSL 1',
        broadband_type_2  : 'ADSL 2',
        broadband_type_6  : 'ADSL 2+',
        broadband_type_7  : 'Cable',
        broadband_type_4  : 'Fiber 1',
        broadband_type_9  : 'BPL',
        broadband_type_3  : 'Fiber 2',
        broadband_type_10 : 'Fiber High Speed',
        broadband_type_1  : 'Fiber Ultra',
        broadband_type_5  : 'Fiber Ultra Max'
}
```

Description of the columns present in the dataset.

- `id` is the instance where the event was recorded when there was an outage in the broadband connectivity in an area
- `area_code` is a categorical column, in which each unique value refers to an area where the `outage_duration` has been measured
- `broadband_type` is the technology that the ISP uses for delivering broadband internet connection, there can be multiple types of broadband connections in a single area

- `outage_type` signifies the `5` different types of outages as classified by the engineering experts who remotely diagnose the issue, once reported
- The `log_report_type` column signifies one of the `386` different types of reports generated by customer service representatives who record issues and classify them as one of the 386 different types of issues
- `transit_server_type` is the type of transit server that handles the traffic of data and route the incoming and outgoing web traffic
- `volume` is the recorded data, in masked units, for 10 minutes prior to the time of recording the observation as per custom company specific units.

Objectives:

In this hackathon, you are expected to:

1. Explore the data and engineer new features
2. Predict the `outage_durtion‘ for records given in `test_data.csv` file
3. Answer questions from the operations team using EDA

Answering questions from the operations team:

The operations team at ‘India Broadband’ has asked you the following questions

- Which areas are most prone to long outage durations?
- Which broadband types are suspect of long outage durations?
- Any other recommendations to improve the detection of outage durations.

Evaluation Metric:

The evaluation metric used for this hackathon would be the **F1 Macro Average**

Submission Timelines:

Submission No	File	Submission Format	Start Date	End Date
Submission - I	1. Exploratory Data Analysis 2. Use insight from EDA to answer the questions from the operations team	Zip format. Include your R-file or. ipynb file plus a converted html file of your R or Python notebook	28th June (9:00 AM)	3rd July (8:00 PM)
Submission - II	Predictions on test.csv (Target attribute: outage_type)	Prediction on test data (format will be same as sample_submission.csv)	4th July (9:00 AM)	5th July (8:00 PM)
Submission - III	Predictions based on improved model and feature engineering	Prediction on test data (format will be same as sample_submission.csv)	8th July (9:00 AM)	8th July (8:00 PM)
Submission - IV	Predictions based on improved model and feature engineering	Prediction on test data (format will be same as sample_submission.csv)	12th July (9:00 AM)	12th July (8:00 PM)
Submission - V	Predictions based on improved model and feature engineering	Prediction on test data (format will be same as sample_submission.csv)	15th July (9:00 AM)	15th July (8:00 PM)
Submission - VI	a final report including all tasks	Zip file format. It should contain all your code files, and if you have prepared any report/ppt/pdf document to present your analysis on the day of viva	16th July (9:00 AM)	17th July (8:00 PM)