**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.

**About Data:**

There are 6 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
* **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.

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' }

**Meta data**.

●`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.

**Tasks :**

1. Explore the data and engineer new features
2. Build a model to predict the `**outage**\_**duration**‘ for records given in `test\_data.csv` file. (use cross validation and hyperparameter tuning)
3. Answer questions from the operations team using EDA

**Questions from the operations team:**

The operations team at small basket 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.

**Free to use any model of choice and choose the appropriate evaluation metric**