# **Approach to the problem statement**

# **JOB-A-THON**

Submitted by-

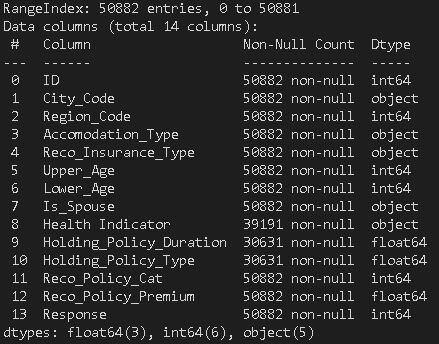
Rohit Pant

Github link - [*https://github.com/r-pant*](https://github.com/r-pant)

**Problem statement:**

To build a model to predict whether the person will be interested in their proposed Health plan/policy given the data.

**Features:**



Link to Data hack – [click here](https://datahack.analyticsvidhya.com/contest/job-a-thon/%23ProblemStatement)

**Score:**

|  |  |
| --- | --- |
| **Data** | **Score** |
| Validation set | 0.7906 |
| Leader score | 0.8077 |

**Approach:**

I used Excel and Tableau for initial exploration of data. This helped me to find out some common patterns that later helped in creating features while feature engineering.

After that, used python for further processing and building the model.

I broke down my approach into below steps to get a stable model.

**Steps:**

Data validation

* Exploring the data
* Finding out the missing values
* Imputing missing values –
  + Used new user column to impute corresponding values
  + Imputed rest values with max. to feed into Catboost
* Checking datatypes

EDA

* Univariate EDA
* Using plots
* Using groupby

Feature engineering

* Used insights from previous steps to create features as:
  + Top\_tier\_cities – Cities having max number of users
  + New\_user – used missing historical data as a way to differentiate between new and old user
  + Single\_user
  + Av\_age
  + Age\_Bracket
  + AgeBin

Removing Outliers

* Used quantile method to remove extreme values from train data

Preparing data for modelling

* Splitting the data
* Encoding (based on, which ml algorithm is used)

ML Models used

* Logistic Regression
* Decision tree
* LGBM
* Catboost

Catboost – Hyper-parameter tuning

* Applied Randomsearch CV for tuning the model for best combination of parameters

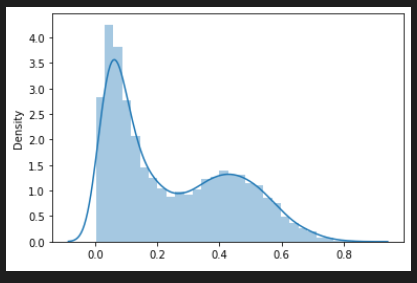
**Catboost** came out as best performing model on the dataset.

Hyper-parameter tuning helped to optimize the model and get better results.

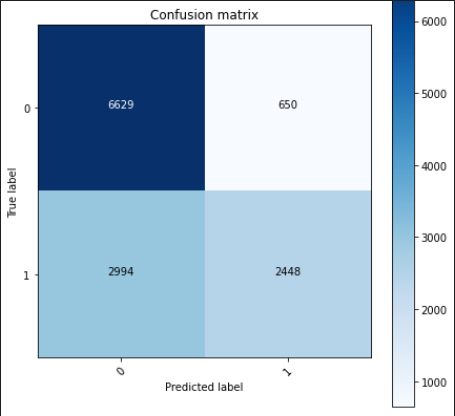
**Final model outcome:**

Model outcome on the validation data is as:

Probability distribution –



Confusion matrix (maximizing True positives by changing the threshold):



**Model Analysis**

* Score improved when missing values are imputed based on new user.
* Removing outliers helped in improving the prediction probability.

Features that were helpful

* Age\_Bracket, New\_user, Top\_tier\_cities

Features that were not helpful

* Single\_user, Av\_age

**Further steps to improve the model:**

* Capturing the region wise impact to improve the model output.
* Using feature selection to remove redundancy
* Using ensemble of different models to improve the model

--------\*\*\*\*\*\*-------