TECHGIG CODE GLADIATORS 2023

docereee™

Machine Learning Hackathon

Team Name- ***ArtMachine***

Team Leader Name- ***Shalaka Thorat***

Team Leader Email Address- ***shalaka.thorat.432@gmail.com***

Brief Description of the Problem at hand:

- We are provided with ad server logs of users that contain information like IP addresses, geographic locations, site URLs and related data.

- The task is to build a machine learning model using the ad server logs data and predict whether a user belongs to HCP (Healthcare Professionals) category or not.

- And if the user belongs to HCP category, we must predict the taxonomy of the HCP as well.

Solution proposed and description:

- We start with data transformation such as, Data pre-processing and Label Encoding on the Test Data we have, so as to make the data suitable for our ML Model.
- Later on, we feed the data to our ML model, which in turn outputs the predictions with respective ID and IS_HCP, whether user is HCP or not (1: HCP, 0: Non-HCP).
- Furthermore, if the user is HCP, the model also provides the Taxonomy of HCP. The prediction files and their contents are as follows.

- File Name: Doceree_HCP_Submission.csv
- Columns:
  - ID: Key
  - IS_HCP: Indicates whether user is Healthcare Professional (HCP) or not

- File Name: Doceree_Taxonomy_Submission.csv
- Columns:
  - ID: Key
  - IS_HCP: Indicates whether user is Healthcare Professional (HCP) or not
  - Taxonomy: Specialization code for HCP

Approach:

- Firstly, we import the train data and perform data analysis and pre-processing steps. We remove the data where IS_HCP value is null, Target column being null will not be useful for prediction. Later, we study each column and steps namely, null value identification and replacement, importance of the column, removing redundant columns.

- Depending on our study, we perform *Feature Engineering* such as, we create CIDRBLOCK and URLDOMAIN columns by leveraging data from BIDREQUESTIP and URL columns resp. We create DEVICE column and try to find device type using USERAGENT column, where DEVICETYPE is Unknown. Furthermore, we perform *Label Encoding, Building Train and Validation Sets*.

- Next, we enter *Model Building* phase, we choose 2 algorithms depending on data size, and speed: Decision Tree Algorithm and Random Forest Algorithm, and compare their accuracy performance. *Decision Tree* model outperforms, so we choose it as our Final Model. Finally, we import our test data, perform data transformation to make the data suitable for the model and get the predictions of whether the person is HCP or not.

- For Taxonomy prediction, we take subset of Training data where IS_HCP is 1, perform steps such as Data Analysis, Pre-processing, Label Encoding, and train our ML model. For Test Data Prediction, if IS_HCP is predicted as 1 by our HCP Prediction Model, it is further passed on for Taxonomy Prediction. We store all these predictions in a csv file as ID, IS_HCP and Taxonomy columns.

- Detailed Approach: https://github.com/shalaka-thorat/IdentifyingHealthcareProfessionalsAndTaxonomy/blob/main/Approach.pdf

Execution Demo (Video/ Screenshots) of the solution:

- **Project Screenshots URL:**

https://github.com/shalaka-thorat/IdentifyingHealthcareProfessionalsAndTaxonomy/tree/main/Screenshots/

Source code in ZIP file / GitHub URL:

- **GitHub URL:**

https://github.com/shalaka-thorat/IdentifyingHealthcareProfessionalsAndTaxonomy/

Additional comments (optional):

- A separate jupyter notebook “Doceree_Taxonomy.ipynb” is created which includes the code for predicting Taxonomy of HCP.
- Output file named “Doceree_Taxonomy_Submission.csv” is created to store the taxonomy predictions.

- Both these files are in the zip folder submitted through `Upload Source Files` option. These files can be accessed from GitHub URL as follows:
- https://github.com/shalaka-thorat/IdentifyingHealthcareProfessionalsAndTaxonomy/

THANK YOU!