Approach : Doceree Machine Learning Hackathon

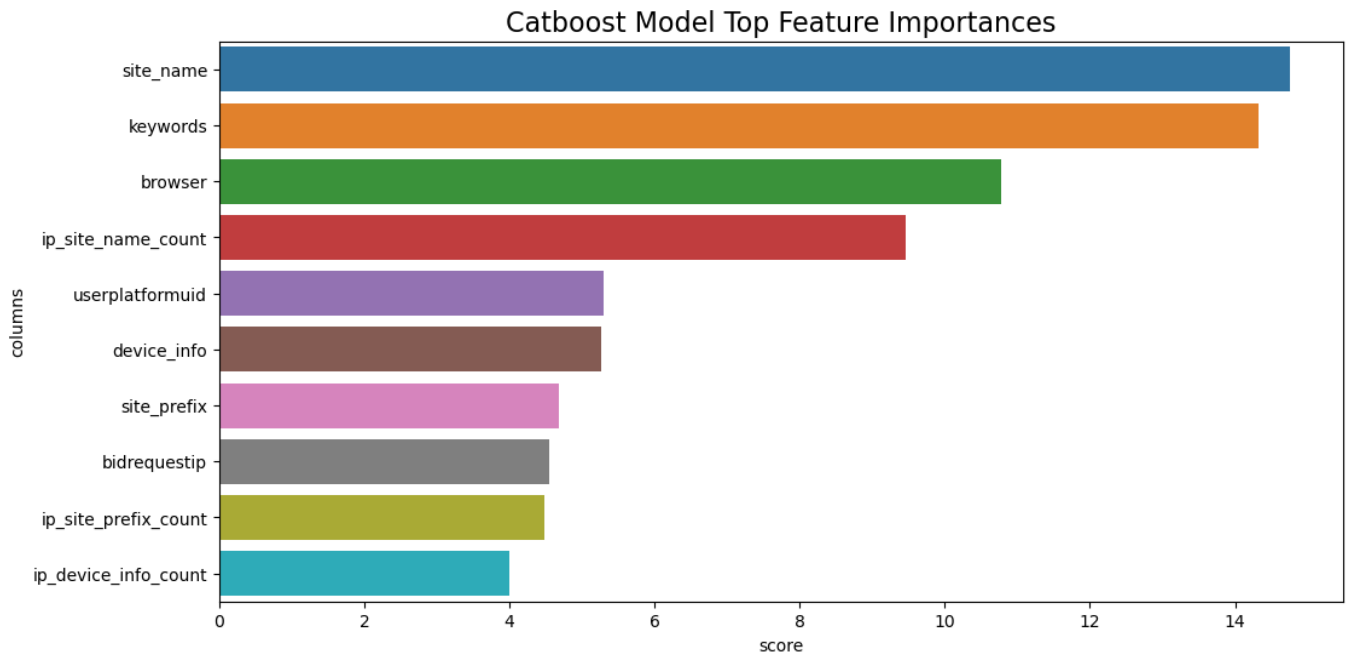
Create a model that can accurately predict whether a user belongs to the HCP(Healthcare Professional) category and its specialization id/taxonomy based on ad server logs information.

- Basic exploratory data analysis using pandas, matplotlib, seaborn.
- Data pre-processing
    - Change column names to lower case
    - Feature Engineering
        - Extract device, browser information from useragent column.
        - Extract the site name, domain name, site prefix information from URL column.
        - Create brand name from device info.
        - Categorical columns level count by IP group.
    - The final features for the model
        - 0_devicetype
        - 1_platform_id
        - 2_bidrequestip
        - 3_userplatformuid
        - 4_platformtype
        - 5_channeltype
        - 6_keywords
        - 7_device_info

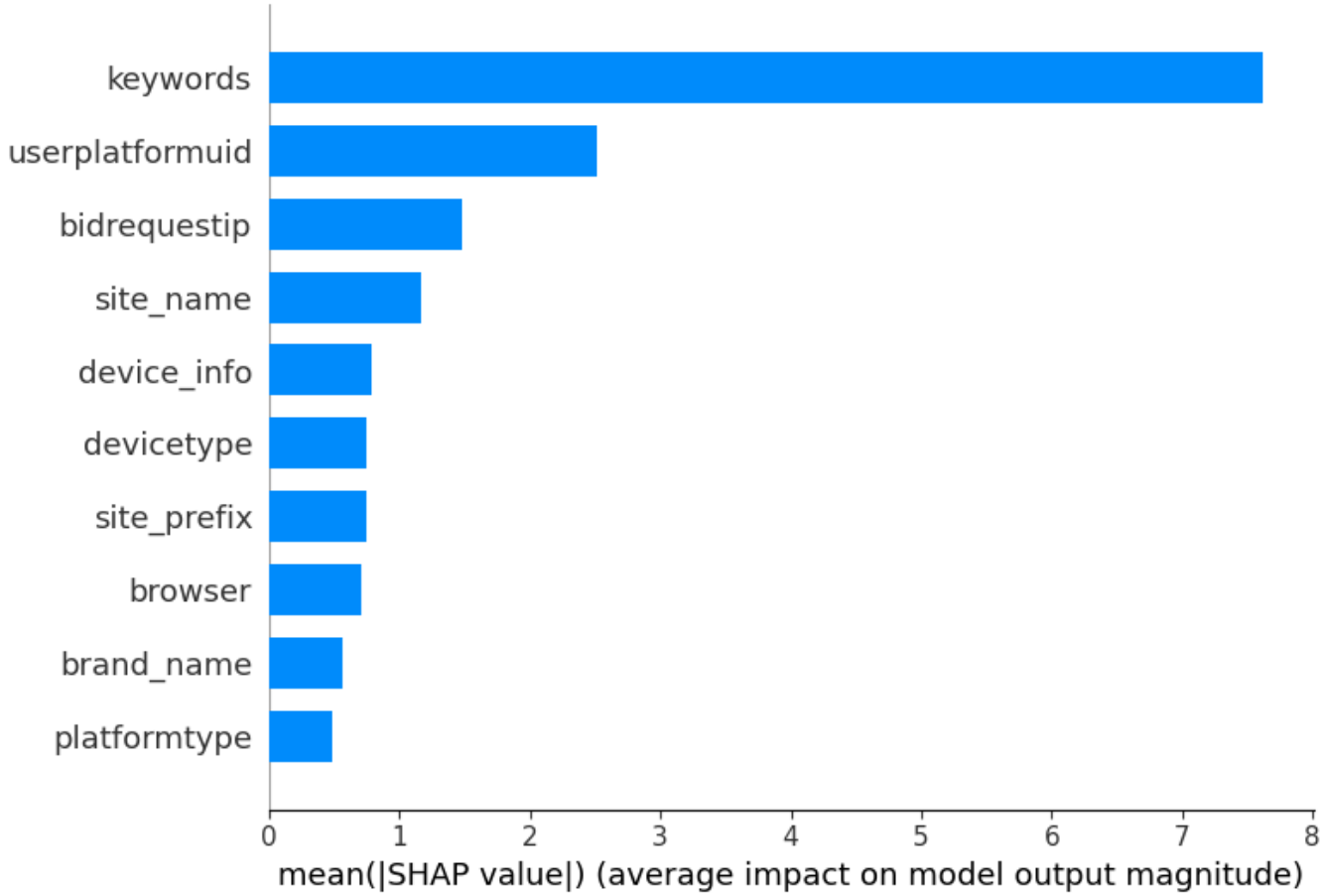
        - 8_brand_name
        - 9_browser
        - 10_site_name
        - 11_domain_name
        - 12_site_prefix
        - 13_userplatformuid_len
        - 14_ip_user_count
        - 15_ip_devicetype_count
        - 16_ip_userplatformuid_count
        - 17_ip_platformtype_count
        - 18_ip_channeltype_count
        - 19_ip_keywords_count
        - 20_ip_device_info_count
        - 21_ip_brand_name_count
        - 22_ip_browser_count
        - 23_ip_site_name_count
        - 24_ip_domain_name_count
        - 25_ip_site_prefix_count
- Created catboost classifier model and tuned hyper parameters by using optuna framework. Model evaluated by Accuracy. After 100 trials,
    - The best score is 0.9972

- The best hyper parameters are,
  - reg_lambda: 0.00144054576027747 71
  - learning_rate: 0.074077168614526 32
  - n_estimators: 337
  - max_depth: 11
  - random_state: 500
  - colsample_bylevel: 0.079092366429850 43
  - boosting_type: Ordered
  - min_data_in_leaf: 65
  - random_strength: 0.61654658585398 36
  - od_type: IncToDec
  - od_wait: 126
  - bootstrap_type: Bernoulli
  - subsample: 0.29632877167360 31

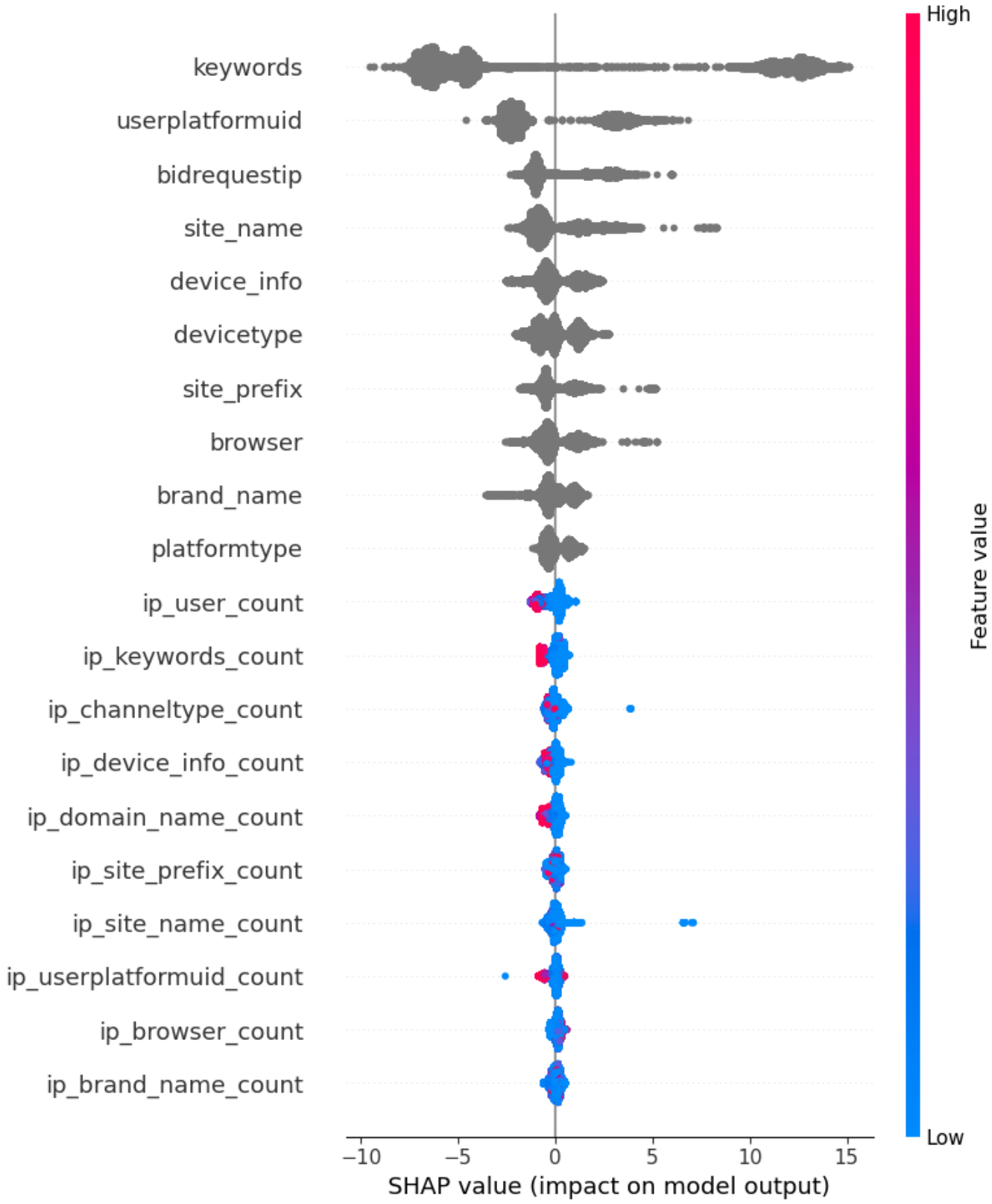
- Catboost model default feature importance’s



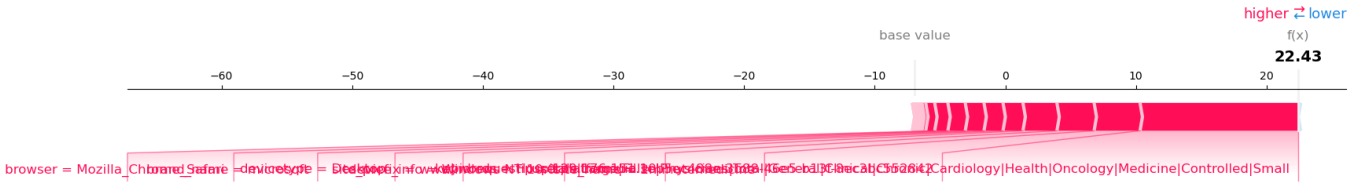
- Catboost SHAP feature importance’s



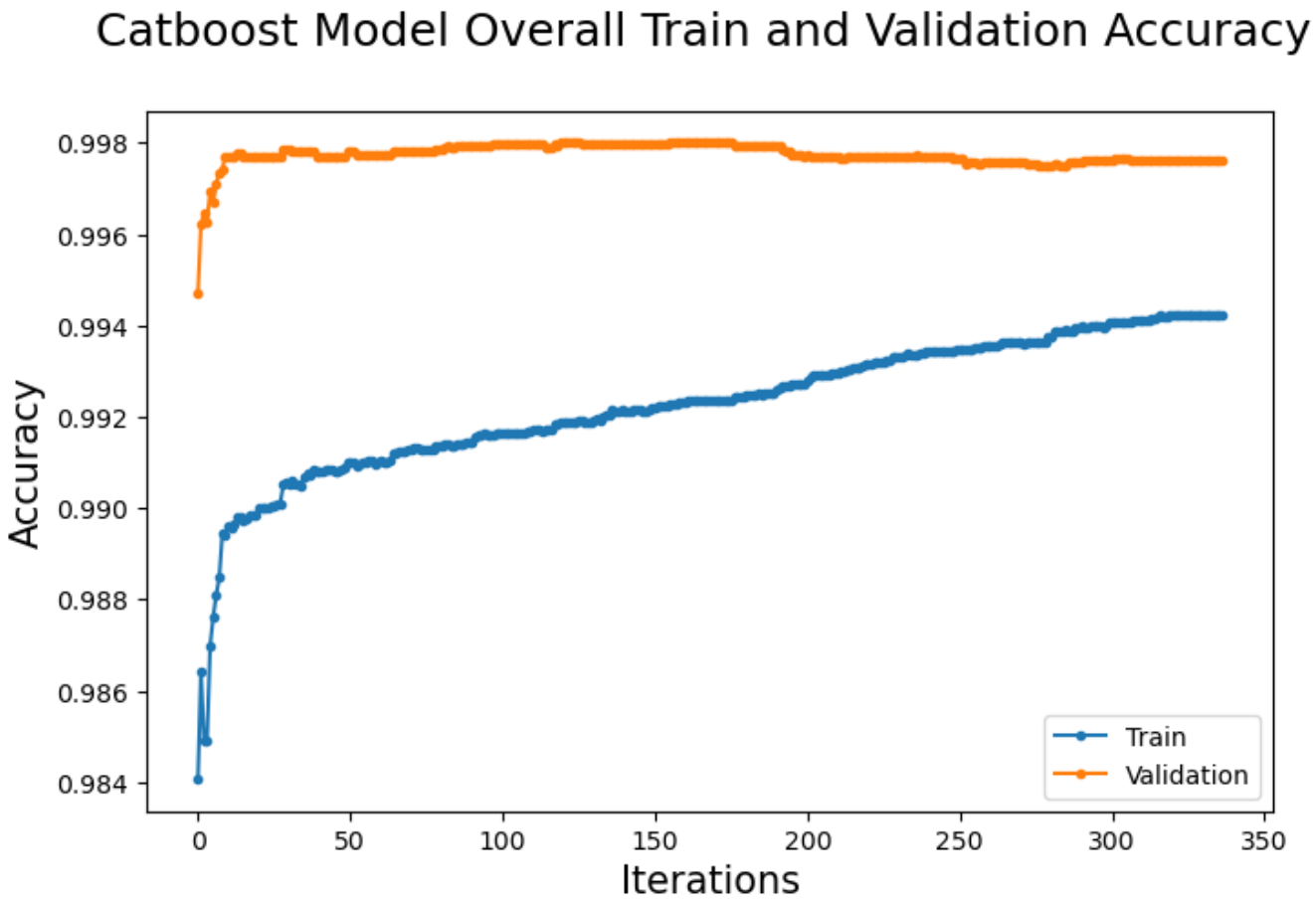
- Catboost SHAP top features impact the model



- Top feature influences for class 1



Overall Train and Validation Accuracy



- Validation data Confusion matrix