# Bot or Not?

**Hassan Sherwani** **23.09.2019**

**Problem Statement**

*Build a better model that can help us identify bot traffic coming from search engines (NHT-search), other bot traffic (NHT-other) and human traffic (HT)*

For this problem, I have made a binary classification problem i.e Human traffic vs non-human traffic. I have also assumed for this exercise that non-human traffic aka bots are of great interest. Hence, I will keep them as class of my interest. Once, I knew that given problem is a classification problem and not a regression problem, it is easier to think of key machine learning algorithms. I have applied linear, non-linear, tree and neural networks on given data. With an optimal model, I came up with good results that were later tuned to provide even better outcome. I will explain very briefly steps and progress in two of notebooks.

**EDA Notebook:**

First of all, I gave a look to my given dataset. Idea is to check if there are any missing values, any duplicate entities, any outliers, or any sort of interesting results that might help me for initial assumptions.
Dataset was very simple with most of its features as “object” type meaning that they belong to categorical variable. I clean very few missing values with a blank space. These missing values were not in our feature of interest so, nothing to worry. How do I know? Those features that contain epochms , session id . url information will not make much difference in our machine learning analysis. So, I chose visitor_recognition_type','country_by_ip_address','region_by_ip_address', and ua_agent as my feature of interest. Next step is to encode all categorical features which are in feature of interest. Then I created a new variable called “labels” that will contain information related to our target feature. I could see that there were more classes than I need. I dropped “special” class as I didn’t know if it belongs to human traffic or non-human traffic. Then I divided Robot and Browser robot as Non-human traffic whereas; Browser, Browser webview, Cloud application, mobile app, hacker are taken as human traffic. I was not fully sure and hence, I might need to develop my domain knowledge. Anyway, this is classification that I came up. My

assumption can be wrong too. This exercise is to practice machine learning models so, we might ignore this issue.
Finally, I checked patterns and unusual trends in my data. One thing that I suspected that all feature" visitor_recognition_type" contains ANONYMOUS type and it might belong to traffic generated from bot. I checked it and found that it was not the case however; all NHT is in ANONYMOUS category. This insight might be help in model results.

**ML Model Notebook**

As I have done all preprocessing, it is time to apply model. One important information is that classes are not in balance. I have 64.5% belong to HT therefore, I assume that my model will be biased. If that is the case then I might come back and balance my classes. For now, I ll continue as it.

I did apply all models one by one and it took time. To save extra time and coding space, I applied k-fold method to fit and validate all given models effectively. Choice of model is in line with baseline model i.e KNN. I used logistic classification, SVM, Naïve Bayes, Decision Tree, Random Forest, Neural network model. Idea is to check if other models perform better than baseline model. I found that logistic classification, random forest, decision tree and neural network model performed better than KNN model. I have got accuracy of 98 % in all four model however; neural network gave accuracy of 98.1% whereas decision tree gave accuracy of 98.07% and random forest showed 98.08%. These two models perform better than decision tree but, I ll go for decision tree. Reason is that neural network and random forest are somewhat complex and time consuming due to computation. I tried to tune decision tree using RandomizedSearchCV technique and after finding best parameters, I could get same 98.1% accuracy from decision tree. So, decision tree is the most optimal model. Note I didn't say the best model. It is simple, efficient, easy to tune, and if we have bigger data (data scaling) then this model will still be functional.

One key issue with classification problems is to find proper evaluation matric for model. Accuracy is not a matric to evaluate our model. As I assume that we are more interested in bot traffic i.e encoded as 1 and hence accuracy does not tell how close we are in predicting bot traffic. It is important to know;

- True positives (TP): These are cases in which we predicted yes (bot traffic), and actually they do have bot traffic i.e non-human.
- True negatives (TN): We predicted non-bot search traffic, actually they were.
- False positives (FP): We predicted bot traffic, but actually they were human generated traffic. (Also known as a "Type I error.")
- False negatives (FN): We predicted human traffic i.e non-bot, but they actually were bots. (Also known as a "Type II error.")

We can see if we want to predict NHT (encoded as 1) then recall is the most feasible evaluation metric. On other side, if we want to see human traffic (encoded as 0) then precision is better evaluation metric. In either case, accuracy only provides a sense how well model is doing in overall prediction. But, if we need to know what classification category is of our interest then we better go deeper.

**My two cents**

- In this problem, I tried to focus on problem statement and solve this in a simpler way. There could be a solution with multi-class problem as was given in example. I simplified this problem to detect if traffic is human or not. I provided explanation of most optimal model, how is it trained, how is it evaluated, how is it tuned and finally how to evaluate given model. Furthermore, I have come up with solution for improving recall score. Even though classes were not balanced, I still managed to get good results.
- I have applied concept of ROC curve and AUC. Through this approach, we can change basic threshold and by decreasing it to 0.90, we can get better prediction of bot traffic. If we decrease it 0.1 then we may get better human generated traffic detection. Through this approach, we may get even better and more precise outcomes. As I was not asked for any specific value hence, I assume that recall score of 0.993 is good enough.