# DataLife: Influencer Prediction Tool

August 8, 2021

## 1 Problem

After our analysis on both social media and surveys we identify the problem of blood banks and associations to attract the younger generation of donors. It is clear that nowadays young people are taking in high consideration the actions and opinions of internet influencers and blood banks and blood donation centers should start leveraging this opportunity. Our analysis underlined the importance of choosing the right influencers. Avoiding the wrong one is not always an easy task and identifying the perfect fit for the campaign is not trivial. Even with the list of tips we proposed on our website, the choice still remains quite complex and the decision is not always straightforward. For this reason we show with this work our solution to the problem involving Machine Learning techniques. We develop also a PoC that can be used to evaluate the effectiveness of the solution on a real customer

## 2 Data

The dataset was built manually by considering the set of features we first identify as can be seen in Table 2. Label can be done manually for a few hundreds of profiles and later can be done by the model itself with the possibility to, of course, change the label. The decision to follow this strategy was mostly due to the fact we knew, based on our experience, who has the best characteristics to generate a good engagement in social media. A qualitative approach founded in data analysis was our choice.

## 3 Model

We decided to deploy a decision tree model as a binary classifier to solve the problem. Our model was trained on 70% of the training data using a validation set to do hyperparameter tuning. The 30% of data left behind was used at the end to assess the accuracy of the decision tree and evaluate how to proceed. What was crucial in the decision of the model was its explainability. The model should be fully explainable and therefore it is required that the online tools provide a clear explanation of what are the reasons for a certain decision.

## 4 Results

The results are encouraging. Our simple model performed well enough on our very small dataset, giving the impression it could improve significantly once the data is augmented. The model learned that the politician feature is the most important to consider for the task as it is shown by Figure 1.

Table 1: Dataset's features

| **Features** | **Source** |
|---|---|
| N° Followers | NotJustAnalytics |
| N° Posts | NotJustAnalytics |
| Avg. Like | NotJustAnalytics |
| Avg. N° comments | NotJustAnalytics |
| Avg. Video Views | NotJustAnalytics |
| Politician | Manual |
| Engament Rate | NotJustAnalytics |

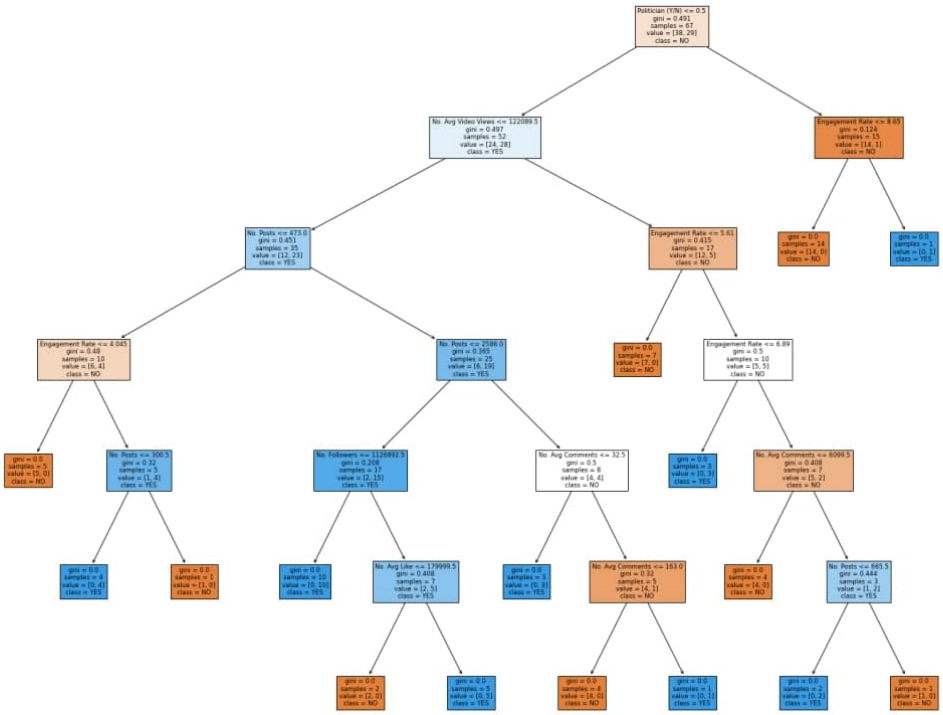


Figure 1: Decision Tree

# 5 Conclusions

This process is not engineered at the moment but it can be. Regarding the data, it is possible to build a dataset and update it in an automatic manner. Most of the features, seen in 2, can be extracted by accessing the NotJustAnalytics API [1]. For the ones who are left, such as politician status and location it is possible to extract this information by collecting the first bunch of posts associated with the profile. From this information we can extract the location very rapidly and extract the caption from the first bunch of posts the account has posted. In this way it is possible to extract the few other features that are needed by the model.

It is suggested to train with the new data every night and provide monitoring of performances through the model's lifetime.

Important would be to provide not only a global explanation of the model as shown in Figure 1, but also a local explanation for each new entry. This can be done through the LIME library [2] that can provide local explanation via input perturbation.

[1] https://www.ninjalitics.com
[2] https://github.com/marcotcr/lime