Facebook Political Data analysis

Jayanti Prasad

Jan 24, 2020

Abstract

In the present analysis I show that political choice of the Indians is mainly determined by the factors which do not vary over a long period of time, such as religion, caste, language and geographic location. Using a small data set with around 300 rows and 6 columns (representing five features & one target) I show that the primary language in which people consume news has much more impact on the political choice of people as compared to other variables such as education & exposure to other cultures and economic status. The main reason for this strong dependency is the gap between the quality of content present in english and in other native Indian languages including Hindi.

1. Introduction

One of the main foundations of democractic systems is that the popularity of an elected government will vary according to the performance of the government on various counts, such as employment, development, law and order situation, economic growth etc. This means people will evaluate political parties on the basis of these factors and so will have a choice to change the government. In an ideal situation this type of system keeps the political parties under pressure for the performance. However, if the political parties can shift the evaluation criteria from the performance based measures to other measures which are much more stable (do not vary),

such as religion, caste, language or other measures which can be easily controlled & manipulated such as renaming places etc., then it can work as effective tactics to be in power without actually doing anything significant. Apart from these factors, if political parties can make use of the inherited loyalties (for example if a candidate is replaced by his family member without having any political loss) then the job can again become easier.

2. Modeling & Data

In the present work I have considered a small set of data based on my facebook friends and used that to predict whether a person will support right wing (hindu nationalist) political party on the basis of a set of binary features which are given as follows:

S. No	Feature Name	Explanation
1	Is Ph.D	The idea is that a Ph.D or similar degree which can isolate a person for more than 5 years may help him/her to broaden his/her horizon.
2	Is Hindu	This is again a fact that the right wing political party in India represents only one religion. This is based on the observation that 0 seats were allocated to a community which make 20 % of the population.
3	Have been Abroad	Again this factor is assumed to help people to broaden their horizon outside their immediate neighbourhood.
4	Primary Language	There is a huge gap in the quality of content in English & other Indian languages, including Hindi. People who follow Indian

		language newspapers and watch Indian language news channels are less likely to change their views, mainly because these sources reinforce the same views which people believe in.
5	Is from North India	There are huge north Indian states with social and economic backwardness that makes them to be easily manipulated.

Table1 : Features with explanation

The target variable is to predict whether a person is a supporter of the right wing hindu nationalist party.

All the features variables have binary values ‘0’ and ‘1’ and so all the columns will have binary values. Since these are numerical variables so we can direct feed them to our classifier, which is a decision tree classifier.

2.1 Decision Tree classifier

The main idea of a decision tree classifier is to identify features which can iteratively group a population in such a way that elements belonging to a group have the same class or target variables. If there were just one feature which could be used to group a population into a set of pure classes then we may not need to use other variables. For example, in our data set it is not always the case that having a Ph.D degree will avoid a person to become a right wing political supporter so we need other feature variables also.

There are two important measures which are used to quantify the homogeneity of a group - information gain and gini index.

In order to compute the information gain we split the data on the basis of a feature variable into two groups and compute the entropy of both the groups and consider the weighted sum of that.

Let us compute the entropy of the target class.

Number of positive cases = 121
Number of negative cases = 169
Total = 290

Root Entropy = - Σ Pi log2 (P_i) = **0.9800**
With :
P1 = 121/290 = 0.417
P2 = 169/290 = 0.582

Now if we split the data on the basis of some feature then we have two groups of size n1 and n2. If the entropy of the groups are E1 and E2 then we can compute the weighted entropy as :

E = (N1/n) * E1 + (N2/n) * E2
With N = N1 + N2

If we split the data on the basis of language we get the entropy:
0.4237, which is minimum as compared to the other cases so we can use the third feature ‘language’ for the first split.

We get similar results if we use gini index, which is defined below, for this purpose.

Gini index = ΣPi^2

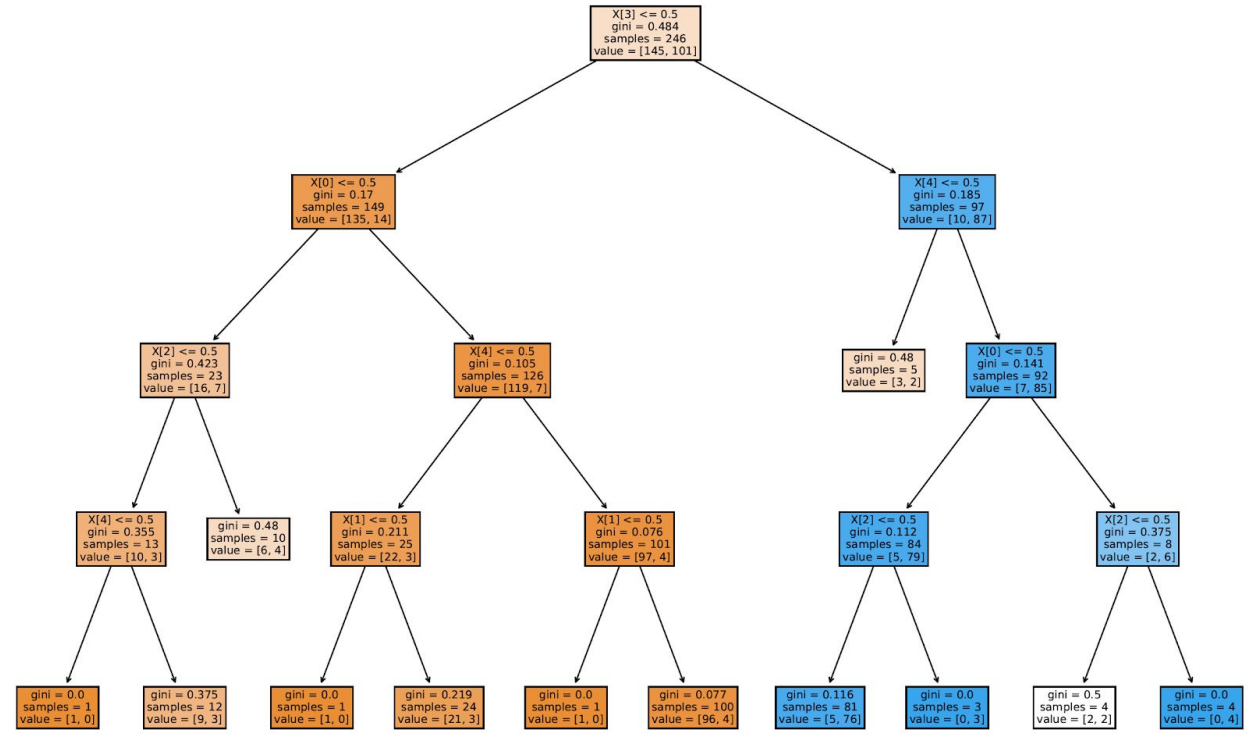
The following table gives the entropy and gini index for all the feature and target classes.

Name	Entropy	Gini Index
Root	0.9800	0.512613
Feature 1	0.6030	0.7465
Feature 2	0.9720	0.5173
Feature 3	.6206	0.7385
Feature 4	**0.4206**	**0.8424**
Feature 5	0.9477	0.5330

Table 2: Inhomogeneity measures (entropy and gini index) for the data.

3. Results

We split out data into training and testing parts where 247 data points are taken for training and 47 for the testing purpose. Out of 47 cases the classifier was able to correctly predict the class of 43 data points. The score we have received is 90 %. The complete decision tree split of the data is given on the next page.


X[3] <= 0.5
gini = 0.484
samples = 246
value = [145, 101]
X[0] <= 0.5
gini = 0.17
samples = 149
value = [135, 14]
X[4] <= 0.5
gini = 0.185
samples = 97
value = [10, 87]
X[2] <= 0.5
gini = 0.423
samples = 23
value = [16, 7]
X[4] <= 0.5
gini = 0.105
samples = 126
value = [119, 7]
gini = 0.48
samples = 5
value = [3, 2]
X[0] <= 0.5
gini = 0.141
samples = 92
value = [7, 85]
X[4] <= 0.5
gini = 0.355
samples = 13
value = [10, 3]
gini = 0.48
samples = 10
value = [6, 4]
X[1] <= 0.5
gini = 0.211
samples = 25
value = [22, 3]
X[1] <= 0.5
gini = 0.076
samples = 101
value = [97, 4]
X[2] <= 0.5
gini = 0.112
samples = 84
value = [5, 79]
X[2] <= 0.5
gini = 0.375
samples = 8
value = [2, 6]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.375
samples = 12
value = [9, 3]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.219
samples = 24
value = [21, 3]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.077
samples = 100
value = [96, 4]
gini = 0.116
samples = 81
value = [5, 76]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.5
samples = 4
value = [2, 2]
gini = 0.0
samples = 4
value = [0, 4]