




DATABASE PROFILING

Marketing Analytics Term Project
Group 1

Inès ODDO, Binxiang XIANG
Emilie LEBLANC, Hippolyte JACOMET



What is database profiling ?

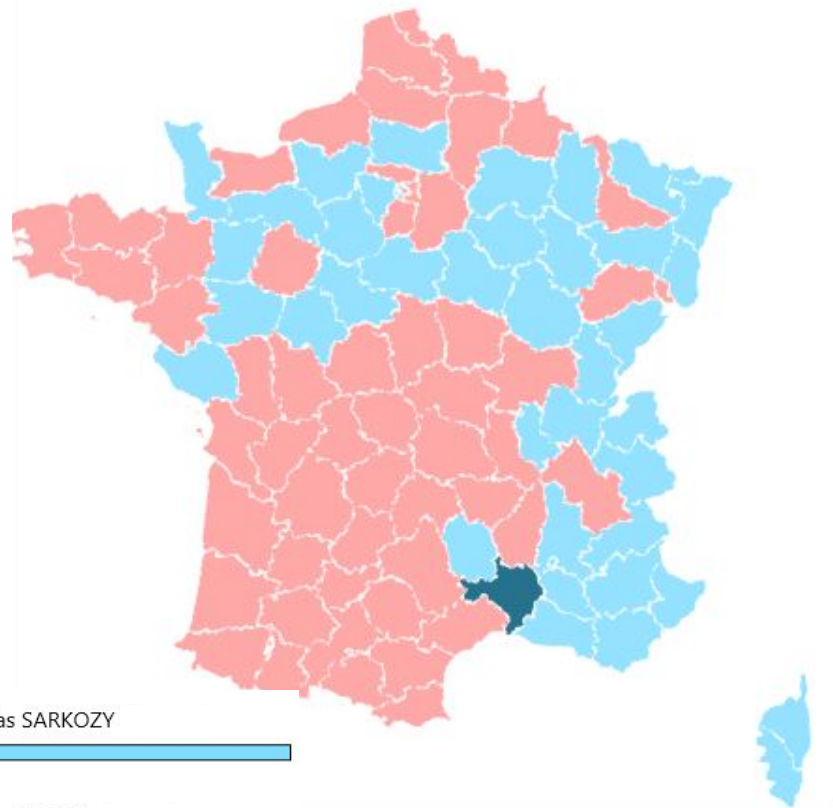
Can we find valuable insights on our donors' database from external sources ?

A way to:

- find key customer drivers
- personalise marketing campaigns

Our project:

Infer political partisanship based on contact's geographical codes.



Nicolas SARKOZY



Marine LE PEN



François HOLLANDE



Naive Approach: Method

Step 1: Pre-processing the data

- Correct geographic codes to fit external database best
- Fill missing values

Our method:

SQL script that duplicates our contacts table so as to safely perform changes

Load town name - geo code correspondence table from INSEE

Update rows accordingly

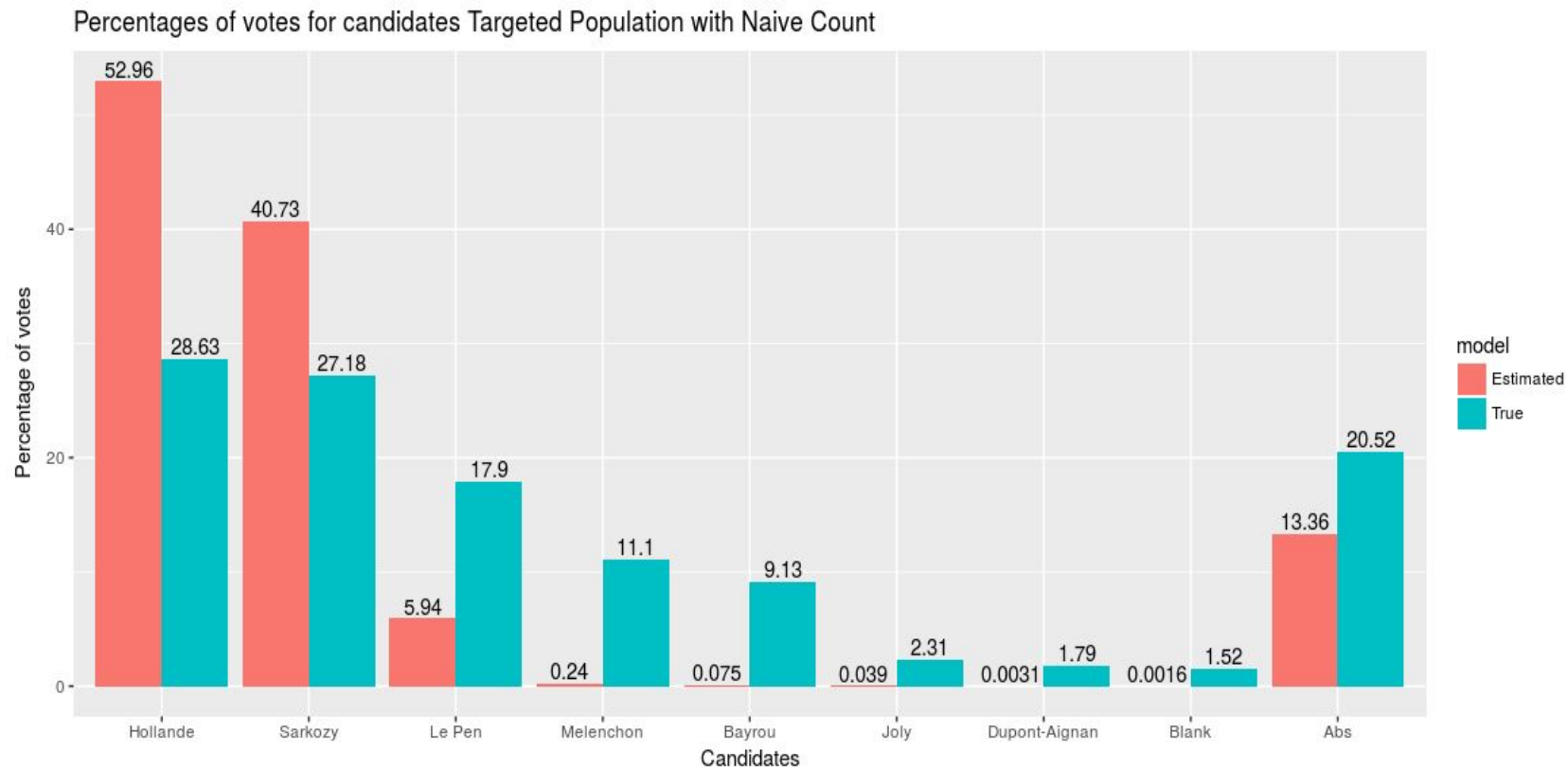
Step 2: Estimating probabilities with proportion

Here we simply infer political partisanship based on the election results by geographical code (INSEE):

The probability of a donor's partisanship to a given candidate is the candidate's proportion of votes in their respective towns.

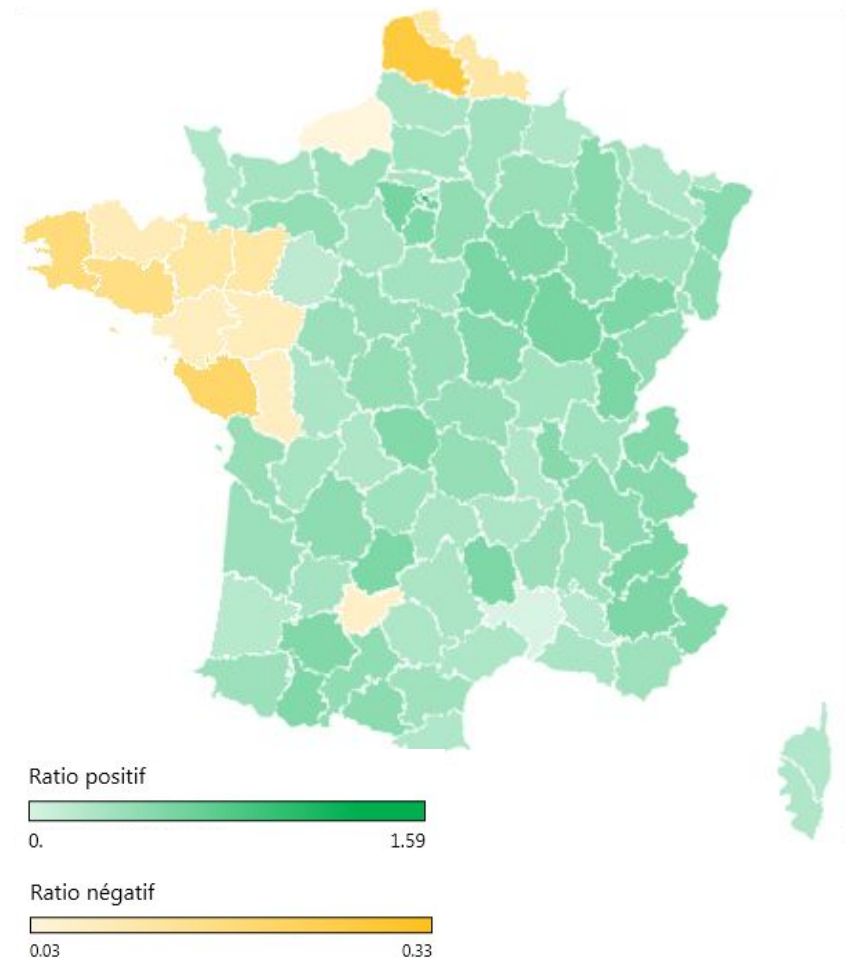
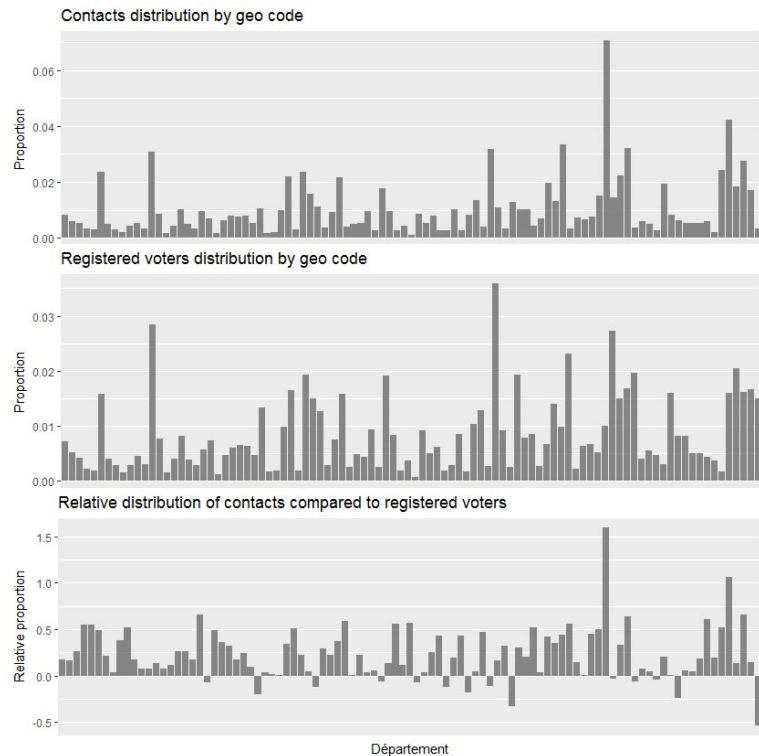
Then, for a given donor, we attribute the candidate with the biggest probability considering its living town. we estimate the candidate for which

Naive Approach: Results



Assessing the bias

Do donors constitute a representative sample of France's voting population ?



Maximum Likelihood Estimator: Method

Key idea: weight the election's results to take into account distribution bias.

Step 1: Same pre-processing step as naive method

Step 2: Weighting probabilities

Assign a relative weight to each candidate (the weight of abstention is set to 1 for reference).

Optimize the weight to maximize the likelihood of the geographical distribution of our donors.

Compute the probability of partisanship to a given candidate as the weighted proportion of the election's results.

Maximum Likelihood Estimator: Results - Weights

Weights obtained per candidate :

We note that the weights obtained by candidate are highly in favor of Eva Joly. Sarkozy and Mélenchon are second and third place.

Conclusion :

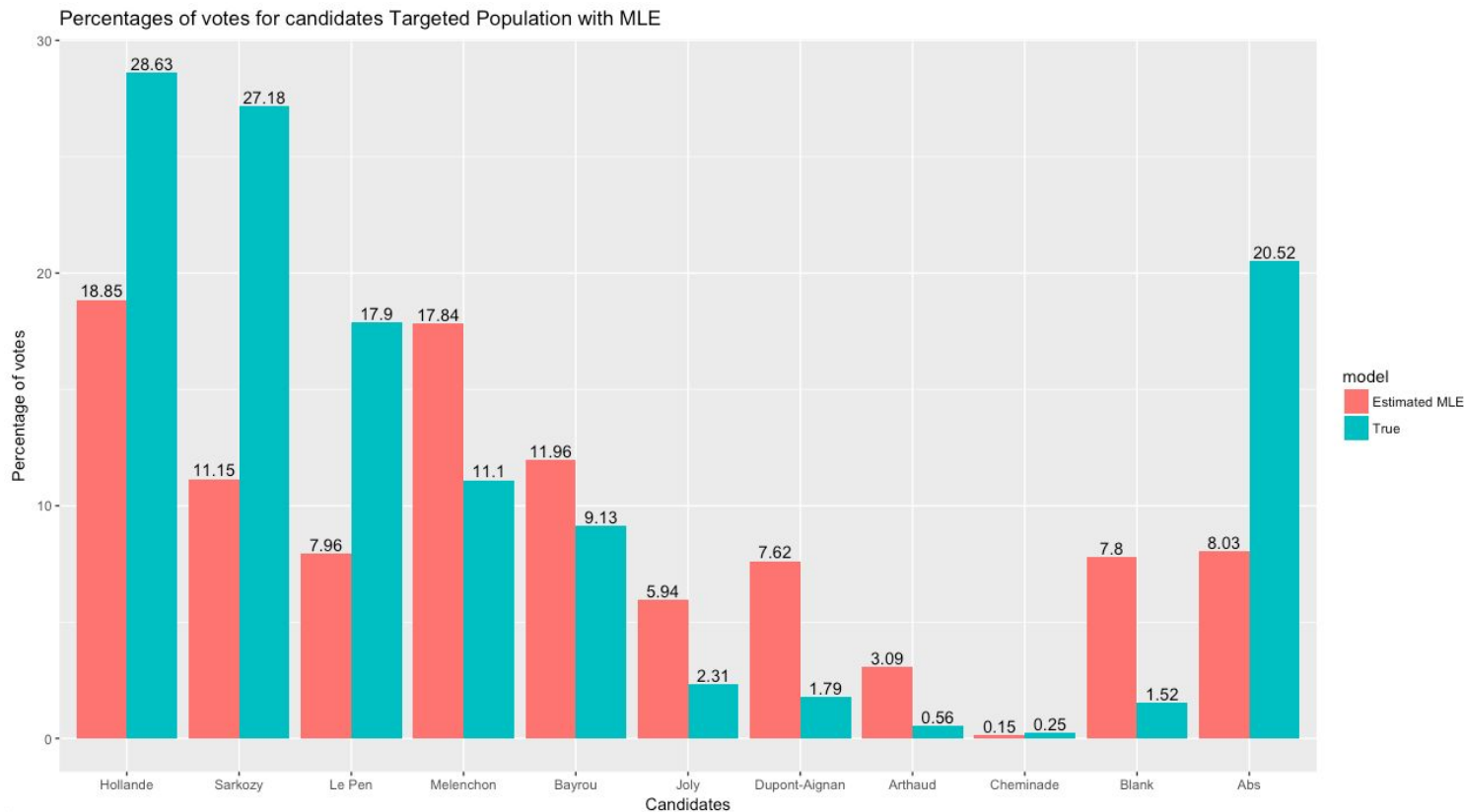
Our database appears to be highly biased in favor of Eva Joly.
Could it be an environmental-friendly charity ?

Warning:

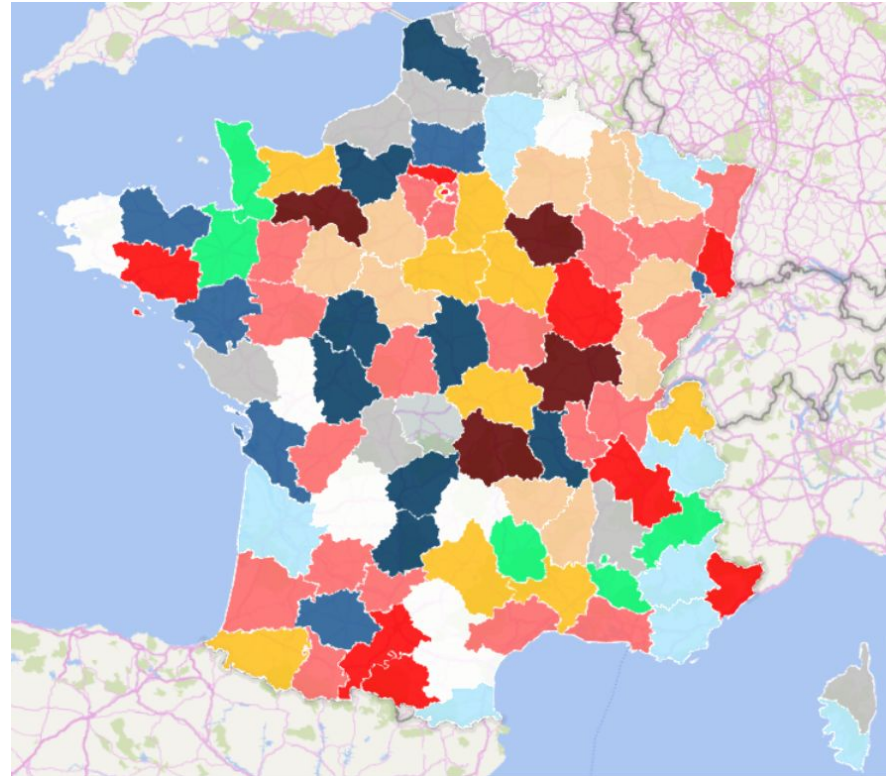
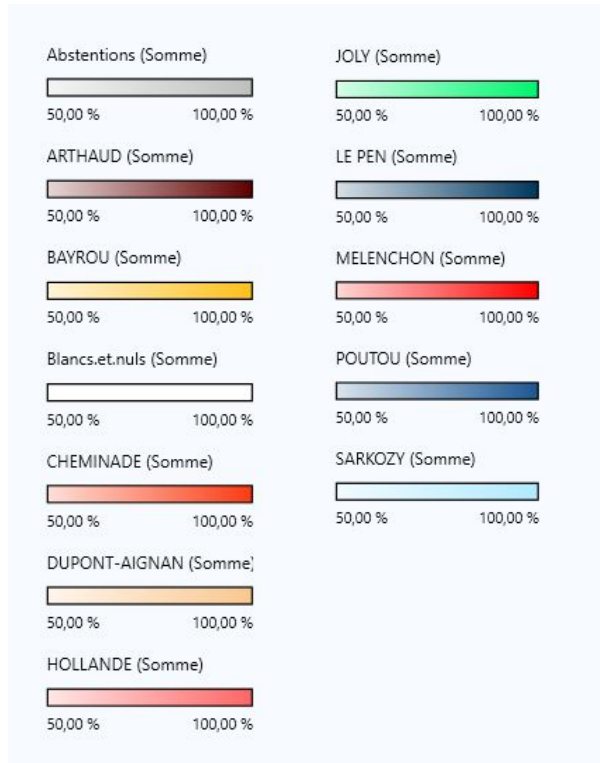
In this methods, the weights are set relatively to each-other. So we fixed the weight of abstention to 1 at the initialisation.

	Weights
Blancs et nuls	1,71699254
Abstentions	0
JOLY	11,388862
LE PEN	-38,915561
SARKOZY	8,58325025
MÉLENCHON	7,96326506
POUTOU	4,84774052
CHEMINADE	1,99430041
DUPONT-AIGNAN	2,57241261
BAYROU	0,93068011
HOLLANDE	0,09033638
ARTHAUD	1,72388605

Maximum Likelihood Estimator: Results - Nationwide



Maximum Likelihood Estimator: Results - Per department



Conclusion

The naive approach can suffer from a biased distribution.

We find that the MLE methods yield significantly different results.

Limitations of the model:

There are no means to evaluate the veracity of our inferred information.

Database profiling should be used with caution.