Tree based methods to predict

Raffaele Sinani
raffaele.sinani@studenti.unitn.it

September 7, 2024

Abstract

With immigration on the rise, along with anti-immigrant feelings, the perception of immigration has been a sensitive and discussed topic for the past few years in Europe, as well as in the US. The aim of this work is going to be an attempt at modeling some subjective indicators of immigration using some social and economic factors. This should allow us to identify the more relevant factors that influence such perception. This work is going to employ tree-based methods, ranging from basic trees to boosting.

1 Introduction

The perception of immigrants and their danger has been a controversial topic. The main issue is the aspect of perception, as is is an entirely subjective matter, so it can be unrelated to factual evidence. However, in this work we are going to attempt to do so, utilizing two indicators that have been collected by European Union surveys:

- The number of immigrants relative to the country population. We are going to compare that number to the actual number.
- The perception of immigration as an opportunity or a danger. We are going to use the proportion of people who view immigration as more of a danger than an opportunity.

To perform the modelization some relevant data is going to be collected:

- GDP at parity of purchasing power. GDP values and GDP variation over the years are going to be utilized as indicators of current wealth and economic growth, respectively. In particular, we considered GDP increase over the past five years as an indicator of long-term economic growth and the GDP increase from last year for short-term growth.
- Poverty risk has been used to verify for possible societal imbalances in the distribution of wealth in the country. This indicator should show if the number of people in dire economical situation impacts immigration perception.
- The percentage of the population that attained a level of education of 3 or higher in the ISCED (International Standard Classification of Education). This is equivalent to Upper secondary education or higher. This statistic has been used as a stand-in for education level.
- The proportion of people that declared themselves "Very well informed" or "Fairly well informed" on immigration matter in the survey has been used to try to verify the level of information on the subject. This indicator might not be reliable as it is also dependent on the respondent's perception.
- Religion might be a factor influencing a country's perception. I considered religion as a categorical variable. Countries have been classified according to whether there is a singular dominant religion, being practiced by over 45% of the population. Atheism and agnosticism were also considered together. In every other case the country has been classified as *mixed*.

This can be useful because it can allow the legislator to have a glimpse at the most important factors that can allow a more seamless integration of foreign immigrants.

2 Literature Review

It has been widely recognized that the attitude towards immigrants is greatly affected by whether said immigrants are highly or lowly skilled (Hinmueller and Hiscox 2007 [HH07]). However, other studies highlighted that such perception also depends on the respondent's own skill level. Hinmueller and Hiscox, in particular, identified in economic literacy as one of the most influential factors in their study.

Those studies give ample importance to education, and the irrational beliefs that can stem from a lack of it. They argue that occupational worries, employment status and economical self-interest are, surprisingly, not as influential.

Chandler, Charles and Tsai (2001) found that some of the aspects that are typically associated with the fear of immigrants to be not statistically significant, such as race, income and fear of crime.

However, successive studies argue that "when considering respondents' overall beliefs regarding the net effect of immigration on household economic welfare, economic self-interest is a substantively important predictor of attitudes toward immigration across individuals and different types of immigrants" (Gerber et al. 2017 [Ger+17]).

Overall, there are many problematic aspects when addressing beliefs. Firstly, some respondents might utilize economic reasons to justify a broader negative attitude towards foreigners. Furthermore, beliefs are hard to measure properly. And finally, it is hard to determine causal relationships between variables, as most subjective views often depend on one another.

3 Methods

The main methodology is going to be the tree, as well as some of its more sophisticated variations. They have some very notable advantages, when compared to other methods. First of all, they don't require as much data to train as other machine learning methods, such as neural networks. This is the fundamental requirements, considering that the sample size is only going to be a survey on European Union countries. Another advantage is explainability, at least as far as basic trees are concerned, as they are easy to represent and understand for an outsider. Finally, this method should also handle the high correlation between subjective variables, as multiple splits on correlated variables wouldn't lead to high decreases of the splitting criterion. To determine some hyper-parameter values we are going to employ 10-fold cross-validation. It is a method that consists in dividing the dataset in ten parts and, for each of them, perform the training on the remaining nine, while testing on the one left out. This way we can estimate the best possible values for the necessary hyperparameters.

3.1 Trees

The main idea behind tree methodologies is to divide the feature space into a number of simple regions, that are as homogeneous as possible. The predicted outcome is then a function of all observations in the region the observation falls into. In our case, we are going to utilize regression trees, as both target variables are going to be numerical. Thus, the predicted outcomes are going to be the averages of all observations that fall in the same region. As a splitting criterion we opted for the natural choice of Mean Standard Error (MSE)

The main problem with this method is the tendency to overfit, which is slightly curbed by setting a maximum tree size. To tackle this issue, the cost-complexity pruning method was employed. It consists of fitting a very large tree and choose the subtree that minimizes the following

$$C_\alpha(T) = R(T) + \alpha|T|$$

where:

- $C_\alpha(T)$ is the cost-complexity of the tree T,
- $R(T)$ is the empirical risk, in our case the MSE, of the tree T,
- α is a regularization parameter that controls the trade-off between the complexity of the tree and its performance and is determined via cross validation,
- $|T|$ is the number of terminal nodes in the tree T.

3.2 Random Forest

Another approach to reduce the variance of tree models is the random forest. The first concept behind this method is bootstrap aggregation (*bagging*). This technique relies on the idea that fitting B trees on B independent training sets and then aggregating the results leads to a reduction in variance of a factor of $1/B$, at no bias cost. Naturally, the B datasets are not actually available and need to be built using the bootstrapping method. It consists of sampling with replacement from the original dataset. This new data should have the same exact distribution as the original one. Naturally, the datasets are not actually independent, so the actual reduction in variance is going to be lower than the expected $1/B$, but it still maintains a significant effect, making large values of B preferable, if computational costs are ignored.

Random forests build upon bagging by sampling m amongst all possible p predictors for every bagging iteration. This is done in order to decorrelate the trees, as the presence of strong predictors is going to create very similar and, therefore, highly correlated trees when employing regular bagging. The number m should be determined via cross-validation and is the tuning parameter that regulates the bias-variance trade-off for the forest. In fact, selecting a high value of m leads to high variance, as it is very likely that all strong predictors are still going to be selected, while a low value leads to high bias, as the trees are going to lose a lot of their accuracy.

3.3 Boosting

Another alternative to forests is boosting. This method employs *weak learners*, simple models with a single split, and subsequently improves them by fitting another tree on the residuals of the previous model. At each step, the tree is able to better explain the phenomenon, reducing bias and increasing variance. The number of maximum iterations M should be determined via cross-validation, as a high number leads to overfitting, and a low number leads to high bias. Another parameter called learning

rate (or *shrinkage*) is often employed to slow down the update of the model.

4 Results

4.1 Immigrant proportion perception

4.1.1 Trees

The first aspect that is worth noting is that almost every country amply overestimates the number of immigrant that is present in their country, with the few notable exceptions of Estonia, Croatia and Sweden. Therefore, the first model was an attempt at determining whether the actual proportion of immigrants is as significant of a factor as it should be. To verify that we fit a tree over all variables available and then we prune it. The number of terminal nodes of the tree, determined via cross-validation, is 4 and the output is shown in Figure 1.

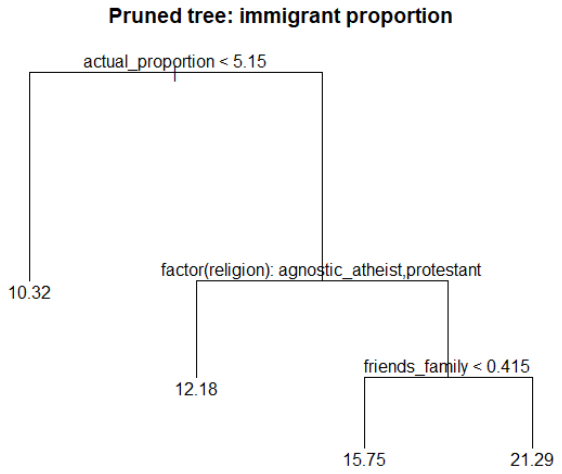

Figure 1: Unpruned tree for the proportion of immigrant people fit on all variables

As it could be expected, the actual proportion of immigrants is the most important variable at play. After that, it seems that countries that are predominantly atheistic/agnostic or protestant seem to have a lower perception of the number of immigrants. Finally, the number of people that declared to have immigrants as friends or family also seems to play a part, as a higher proportion also reflects into a higher perception of the proportion of immigrants in their country.

To remove this influence from now on we are going to consider, as the outcome variable, the difference between the perceived proportion and the actual value.

Repeating the same procedure we obtain the tree shown in Figure 2

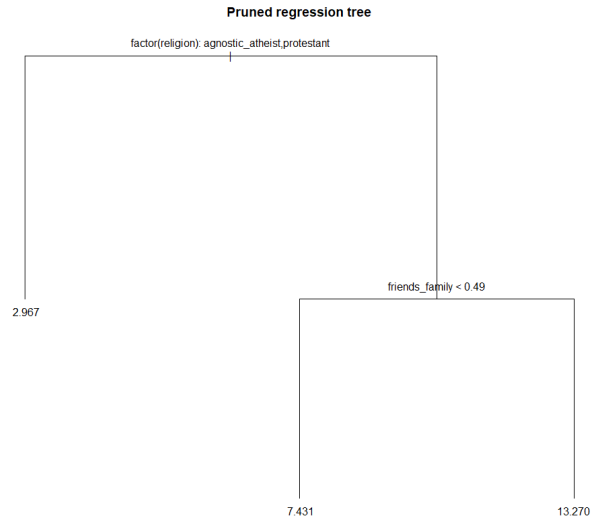

Figure 2: Pruned tree for the difference between the proportion of immigrant people and the perceived proportion

The tree has 3 terminal nodes and is very similar to the previous one, which is expected. Similarly, atheistic/agnostic or protestant countries have more accurate estimates of the actual value, and having immigrants as friends or family tends to increase the overestimation.

4.1.2 Random Forest

We can now fit a random forest with 5000 trees over all our variables. With this methodology we lose the ability to show a simple figure as we could with trees. The resulting forest only explains 22.33% of the training data variability. This can be due to a high number of unimportant variables being sampled. To verify which variables hold the most value we can look at the importance table (Table 1).

Variable	**%IncMSE**
GDP 2017	-0.8993
Long-term GDP change	-3.0638
2017 GDP change	0.6049
Percentage at poverty risk	1.0658
Percentage educated (Level 3+)	10.0396
Percentage informed	2.2929
Percentage with immigrant friends/family	9.4688
Religion	2.4956

Table 1: Variable importance based on % MSE increase

In this case the economic variables still seem to hold the

least weight, with 2017 GDP and long term GDP change being a negative inclusion in the model.

To improve the model we can gradually remove the least important variables, alongside those that appear to be detrimental. The best model only includes the amount of people with immigrants as friends or family, percentage of educated people, religion and the proportion that declared themselves informed in order of decreasing importance. This model is able to explain 38.21% of the variability. This seems to confirm that economic condition might not be the most relevant factor when it comes to the perception of the amount of immigrants.

Variable	**%IncMSE**
Percentage educated (Level 3+)	9.2964
Percentage with immigrant friends/family	11.4049
Religion	6.8283
Percentage informed	2.2201

Table 2: Variable importance based on % MSE increase in the final model

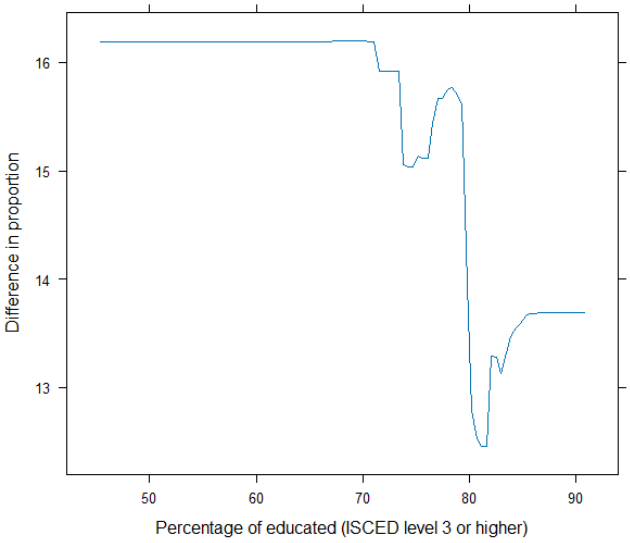

Figure 3: Partial dependence of the difference in proportion over the percentage of people with level 3 or higher education

4.1.3 Boosting

Finally, we can fit trees using the boosted model. Once again, the simple explainability of trees is lost, having to rely once again on relative importance (Table 3).

Variable	**Relative Importance (%)**
Percentage educated (Level 3+)	28.3677
Religion	24.9543
Percentage with immigrant friends/family	13.3034
2017 GDP	9.9224
Percentage at poverty risk	7.3467
Percentage informed	5.9349
2017 GDP change	5.2157
Long-term GDP change	4.9548

Table 3: Relative importance of variables in the boosted model

The most relevant variables in the boosted model are the same three as before: percentage of educated people, religion and proportion of people with immigrants as friends or family. We can verify the marginal effect of the *education* varaible by producing a partial dependence plot (Figure 3). It shows that as the number of educated people grows, lower the difference between perception and reality gets, which is the effect we expected.

Similarly for the proportion of respondents that declared having immigrants as friends or family in Figure 4. This plot seems to confirm that having direct contact with immigrants seems to increase their perceived number.

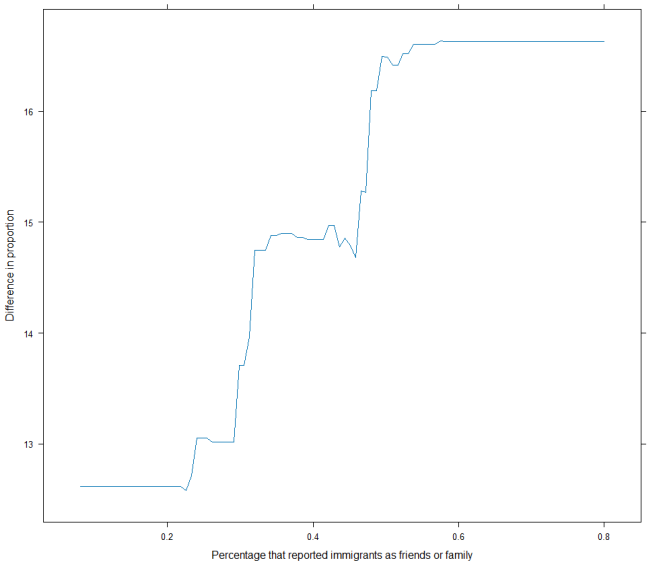

Figure 4: Partial dependence of the difference in proportion over the percentage of people who reported immigrants as friends or family

4.1.4 Brief conclusions

All models that tried to explain the difference in the perception of the proportion of immigrants and the real value seem to confirm the importance of cultural and social aspect. The experience of having immigrants in their close circle, the religion and the level of education seem to be the dominant predictors in all the analysis.

4.2 Perception of immigration as a problem

4.2.1 Trees

The second analysis regards the proportion of respondents that view immigration as more of a problem than a resource.

The first method, using every variable to fit a tree (Figure 5), doesn't show promising results. Cross-validation indicates 2 to be the best number of terminal nodes, which indicates that the model is not really capable of making reasonably accurate predictions. This is almost certainly due to the very low sample size. In any case the only variable that seems to explain part of the variability is the economic factor of the country's GDP in the current year.

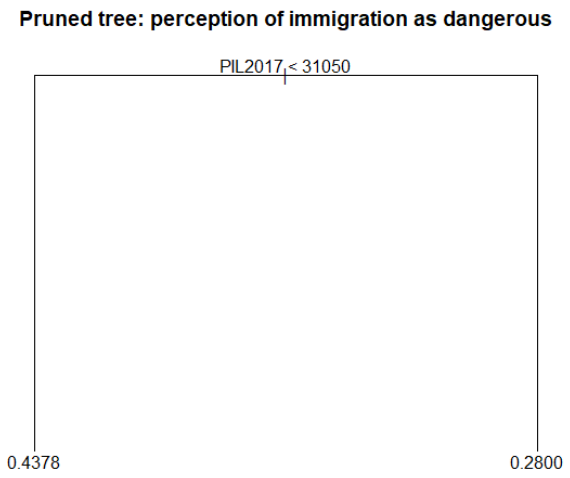

Figure 5: Pruned tree of the proportion of individuals considering immigration as a danger over all variables

4.2.2 Random Forest

Fitting a random forest to all the data, doesn't show any more promising results. The full model actually explains negative variance, performing worse than a base model that takes the sample average. To try and improve the model we look at the importance table (4).

Variable	%IncMSE
GDP 2017	7.6985
Actual proportion	-0.0820
Total GDP change	2.4239
2017 GDP change	1.3035
Percentage at poverty risk	-1.3094
Percentage educated	-3.8819
Percentage informed	0.9004
Percentage with immigrant friends/family	3.7693
Religion	-0.5207

Table 4: Variable importance based on % MSE increase

By keeping only the three variables of 2017 GDP, Long-term GDP change and proportion of respondents with immigrants as friends and family we obtain a model that is at least better than the base one, being able to explain 10% of the sample variance. These results are, however, quite unreliable as everything is very dependent on the random seed selected.

4.2.3 Boosting

Finally, we can fit a boosted model to the data. Looking at the importance table (Table 5

Variable	Relative Importance (%)
Actual proportion	19.434
Friends with family	16.414
2017 GDP	12.890
2017 GDP change	12.374
Percentage at poverty risk	10.478
Percentage informed	10.270
Total GDP change	10.207
Religion	7.933

Table 5: Relative importance of variables in the boosted model

There is very little information to be taken from this table, as almost all variables have similar relative influence.

4.2.4 Brief conclusions

The low amount of data, as well as the limited amount of resources do not allow any significant conclusion from this analysis. The only aspects that seem important are the presence of the economic aspect of the latest GDP and the proportion with immigrant friends or family as important variables in all models.

5 Discussion and conclusion

Overall, the results of both analysis should not at all be considered definitive. This is due to the overarching

problem of the lack of data, as 28 data points from a single survey are most definitely not enough to develop a proper model. However, they should be still able to offer at least some information, as trees are less sensitive to the small sample and the methods employed (cross-validation, pruning, bootstrapping and boosting) should somewhat alleviate the overfitting problem.

In particular, as far as the first analysis is concerned, being able to explain almost 40% of the variance is not a meaningless result at all. It should be safe to say that the three variables individuated (percentage of educated people, percentage with immigrant friends or family and religion) do, in some way, influence the perception on the number of immigrants in a country.

On the other hand, the second analysis brought much less clear results. The methods employed were able to explain very little of the sample variability, making any conclusion derived from them quite unreliable. None the less, the very few results shown do not completely eliminate the possibility of obtaining much more accurate models given the availability of more data and considering other variables that might not have been considered in this work.

Theory References

[CT01] Charles R. Chandler and Yung-mei Tsai. "Social factors influencing immigration attitudes: an analysis of data from the General Social Survey". eng. In: *The Social science journal (Fort Collins)* 38.2 (2001), pp. 177–188. ISSN: 0362-3319.

[Ger+17] Alan S. Gerber, Gregory A. Huber, Daniel R. Biggers, and David J. Hendry. "Self-Interest, Beliefs, and Policy Opinions: Understanding How Economic Beliefs Affect Immigration Policy Preferences". In: *Political research quarterly* 70.1 (2017), pp. 155–171. ISSN: 1065-9129.

[HH07] Jens Hainmueller and Michael J. Hiscox. "Educated Preferences: Explaining Attitudes Toward Immigration in Europe". In: *International Organization* 61.2 (2007), pp. 399–442. DOI: 10.1017/S0020818307070142.

Data Sources

[ETF19] ETF. *KEY INDICATORS ON EDUCATION, SKILLS AND EMPLOYMENT*. 2019. URL: https://www.etf.europa.eu/sites/default/files/2019-12/kiese_2019.pdf.

[Eur15] European Commission. *Special Eurobarometer 437: Discrimination in the European Union*. Accessed: 2024-09-06. 2015. URL: https://europa.eu/eurobarometer/surveys/detail/2077.

[Eur18] European Commission. *Special Eurobarometer 469: Integration of immigrants in the European Union"*. Accessed: 2024-09-07. 2018. URL: https://europa.eu/eurobarometer/surveys/detail/2169.

Code repository

[] *Code Repository*. Accessed: 2024-09-07. URL: https://github.com/SanzioMeh/CSS_project.