Marriage and Identity

Computational Social Science

Master's Degree in Data Science Research Project

Patrick Montanari - 226515

Univeristy of Trento - Academical year 2021-2022

Abstract: The conception of marriage is a social construct which has permeated every human society ever since the first settlements in ancient times. Despite having different meaning and importance across history, it has always been considered a milestone of adult life and, as such, all the traditions and celebrations are passed on to each generation. For this research I wanted to see if one's social status had an impact on opinions regarding marriage, specifically depending on whether one belongs to an ethnic minority and/or is a religious person. Different models were tested to see which of them was performing better; how results differ between each country is also a key part of it.

Contents

1 Introduction

1.1 What is marriage?

The conception of marriage is a social construct which has permeated every human society ever since the first settlements in ancient times. Despite having had different meanings and importance tied to it across history, it has always been considered a milestone of adult life and, as such, deeply respected and recognized as a unequivocal rite of passage by the whole community. Moreover, it used to be the moment where young women left the father's household to start a family of their own; in the modern age, thanks to Feminist movements and the improvements achieved with the research of technologies, marriage has become more of a confirmation of an already formed bond, aimed at sanctioning the union of two mature individuals (extended beyond the common idea of meeting across genders of heterosexual relationships).[1] Other forms of union, like cohabitation, are widely embraced.

The power dynamics are also being shifted; with the man no longer being the primary agent of producing income as default, marriage now represents both a boundary and a desirable outcome for both partners involved. It often represents a trivial source of happiness and life satisfaction.[2] Among minorities, it represents an instrument to strengthen kinship and consolidate one's identity.[3] External pressure from the social group could play a role as well. On this note, marriages between different ethnic groups are more likely to lead to a divorce.[4]

1.2 Research Design

With all these factors taken in consideration, my research will try to demonstrate how many factors tied to social status play a role in defining our idea of marriage (more specifically, the ideal age to get married at). My main hypothesis is that both belonging to an ethnic minority group and belonging to an organized religion are associated with a lower ideal age of marriage reported; I also decided to include several socio-demographic parameters, such as age, gender and educational level. Do note that a similar study concerning immigrants and idea of marriage had been previously published; however, as I will explain in the next section, both the aspect observed, and the dimension of analysis diverge because of the methods implemented.

[1] Howe R. T., 2017.

[2] Lelkes O., 2008.

[3] Beck-Gernsheim, 2007.

[4] Milewski N., 2014.

2 Previous Studies

The primary inspiration for my research is the article "Ideal ages for family formation among immigrants in Europe".[5] Published in 2012, it mainly revolved around the correlation between ideal age of father/motherhood and marriage with the subject's background: education level, parents education level, average marriage age in the country of provenience. Those and many others control variables were fitted by the author in a regression model.

The main difference between this work and the said articles is in two main points: my focus is not on the background, but rather on an individual's characteristics as they define him as a whole. All the emphasis put by the author on years spent in the country, parents' education level and average marriage age for the state are replaced with life conditions aspects, to better describe one's actual and current conditions. The second point is on the methodology aspect: I will include different models (regression trees, generalized linear models, naïve-Bayes classifiers), and discuss which of those could be more appropriate for it. Moreover, the first study is based on 2006 data, while mine is taken from data traced back to 2018, twelve years later. Global and local changes could produce different results and, as a consequence, I am definitely sure that the cross-sectional aspect will prove this statement.

Another category of marriage related studies involves the comparison between civil and religious marriage, cohabitation, and other forms of mutual agreement between partners set after the second major demographic transition[6]: according to Lesthaeghe and many other demographers, procreation is no longer strongly connected to marriage. Concurrently, a decrease in fertility has affected all European countries in the last decades, currently below the replacement level (less people are born than those dying in the same time frame).[7]

A key publication for the conception and design of this research is "Explaining cross-national differences in marriage, cohabitation, and divorce in Europe, 1990–2000".[8] As said by the author: "While marriage rates have declined in many Western countries (albeit not simultaneously or to the same degree), the differences among countries at any point in time remain striking." This made me decide to include a multilevel aspect, comparing and filtering results by country of respondent to show how it impacted average ideal age of marriage's distribution.

[5] Holland J.A. et al., 2013.

[6] Perelli-Harris B. et al., 2012.

[7] Lesthaeghe R., 2014.

[8] Kalmijn M., 2007.

3 Methodology

3.1 Data Description

Most of the analysis were conducted on R (version 4.1.3), with each method repeated on Python (version 3.10.4) as a proof of reliability. First and foremost, I gathered all Data needed from European's Social Survey;[9] I chose the 9th round for two main reasons: first, because it was the most recent one; second, due to the fact that it includes the rotating module "timing of life", which contains my response variable: *iagmr*, ideal age to get married for each subject.

Aside from this one, 9 others were selected: four I define as "socio-demographic parameters": age (*agea*, discrete), gender (*gndr*), country of origin (*cntry*) and years spent in the education system (*eduyrs*, discrete). The other five involved one's "social status", both as identity aspects and personal views: belonging to a particular organized religion (*rlgblg*), ever been married (*evmar*), ever given birth or fathered a child (*bthcld*), importance attributed to following traditions (*imptrad*) and, lastly, belonging to a minority ethnic group (*blgetmg*). All of these variables were examined during the data cleaning step in order to visualize outliers and exclude missing variables. I was initially planning to use the educational level variable (*edulvlb*); however, its high correlation with years spent in education (a quantitative variable, more suited for this project) led me to choose the latter.

Initially I included another variable: *dscrgrp*, member of a group discriminated. What led me to remove it from my analysis was the very low statistical relevancy that it had, bearing a p-value much higher than 0.05 for each of the 4 methods and iterations I tested on R (lm, naïve-Bayes, regression tree, boosting).[10] I then proceeded by filtering the database, excluding missing values of the response and predictors, and removing variable's labels for smoother processing using the R library haven's method "zip_labels".

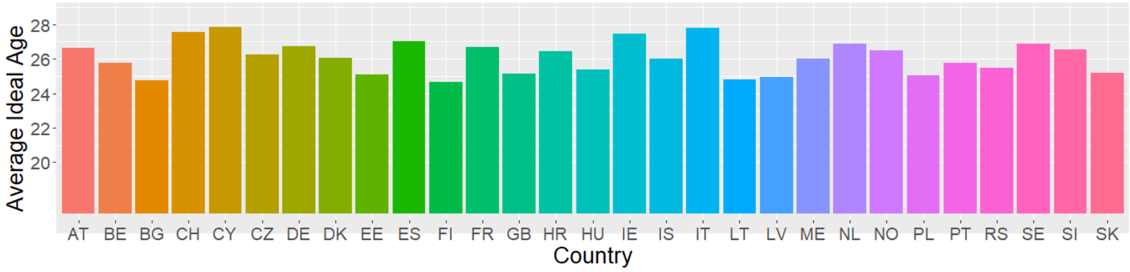
Out of those 10 variables, only three are proper quantitative discretely distributed: *agea*, *iagmr* and *eduyrs*. For the response variable I applied two selections: removed all the outliers and exclude those whose ideal age of marriage was below 18.[11] What was left was a sample of 39'271 observations, with *iagmr* having an average of 26 and 3.55 as standard deviation. Age reported was highly influenced by the country of origin, as seen in the image below.

[9] https://www.europeansocialsurvey.org/data

[10] Even using interaction with *blgetmg* the variable was not statistically significant.

[11] Being 18 the legal age of adulthood across all European Countries, I wanted my research to reflect this; also, less than 1% of the answers reported an age between 14 and 17, effectively minimizing the loss.

Figure 1: Barplot visualization of iagmr mean value by country.

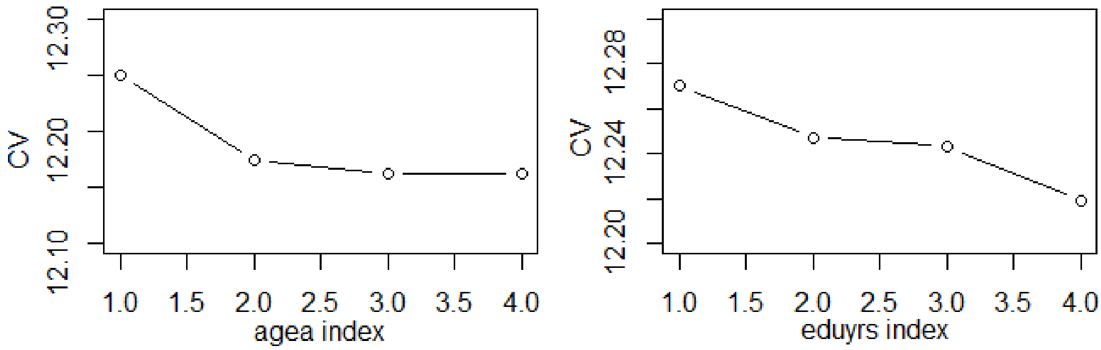


3.2 Statistical Models

The relationship between my predictor variables and the response was tested using different models; LDA and logistic regression had to be left out, as they would not be adequate for a discrete dependent variable. My choice was between a linear model, a Naïve-Bayes classifier. and regression trees. Random forest's output would be more precise than the regression tree, having significantly higher computations complexity as the only downside.

Using the LOOCV method "cv.glm", taken from the boot package, I tested whether the relationships between *agea* and *iagmr* or between *eduyrs* and *iagmr* was linear or polynomial. Due to the very low improvements (less than 0.5%) reached with higher degrees for both parameters (as shown in the next figure), I decided to opt for a linear model.

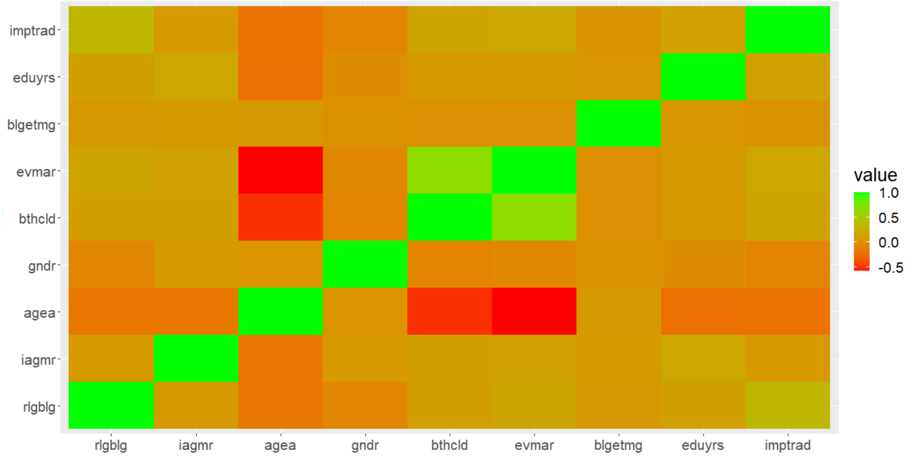
Figure 2: LOOCV performed on agea and eduyrs's influence on iagmr based on degree.



I divided my sample in testing and training sets, both chosen randomly and with 10 different splits for a 10-fold cross-validation. A linear regression was performed on both using the eight variables described, with all coefficients' significance corroborated by ANOVA. I also performed a dummied regression only including the four socio-demographic parameters, which proved to be less precise as it had much higher AIC and BIC.[12]

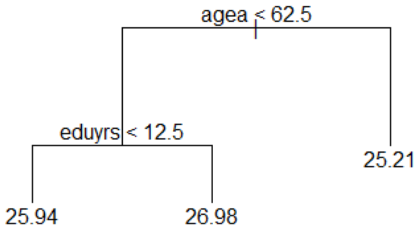
A Naïve-Bayes classifier algorithm was then applied, converting predicted values in categories by rounding the decimals to match with the age reported (considered a categorical variable instead of discrete one for this step). The main assumption underlying this process was the independence of the parameters (due to the low correlation between them, as shown in the table below created using reshape's "melt" function and ggplot2).[13]

Figure 3: Correlation matrix for all variables except cntry (incomparable).



Lastly, I performed a regression tree method using R's library tree to see the discriminating thresholds for the discrete variables. After pruning, this is what it looks like (described later).

Figure 4: Pruned Regression Tree



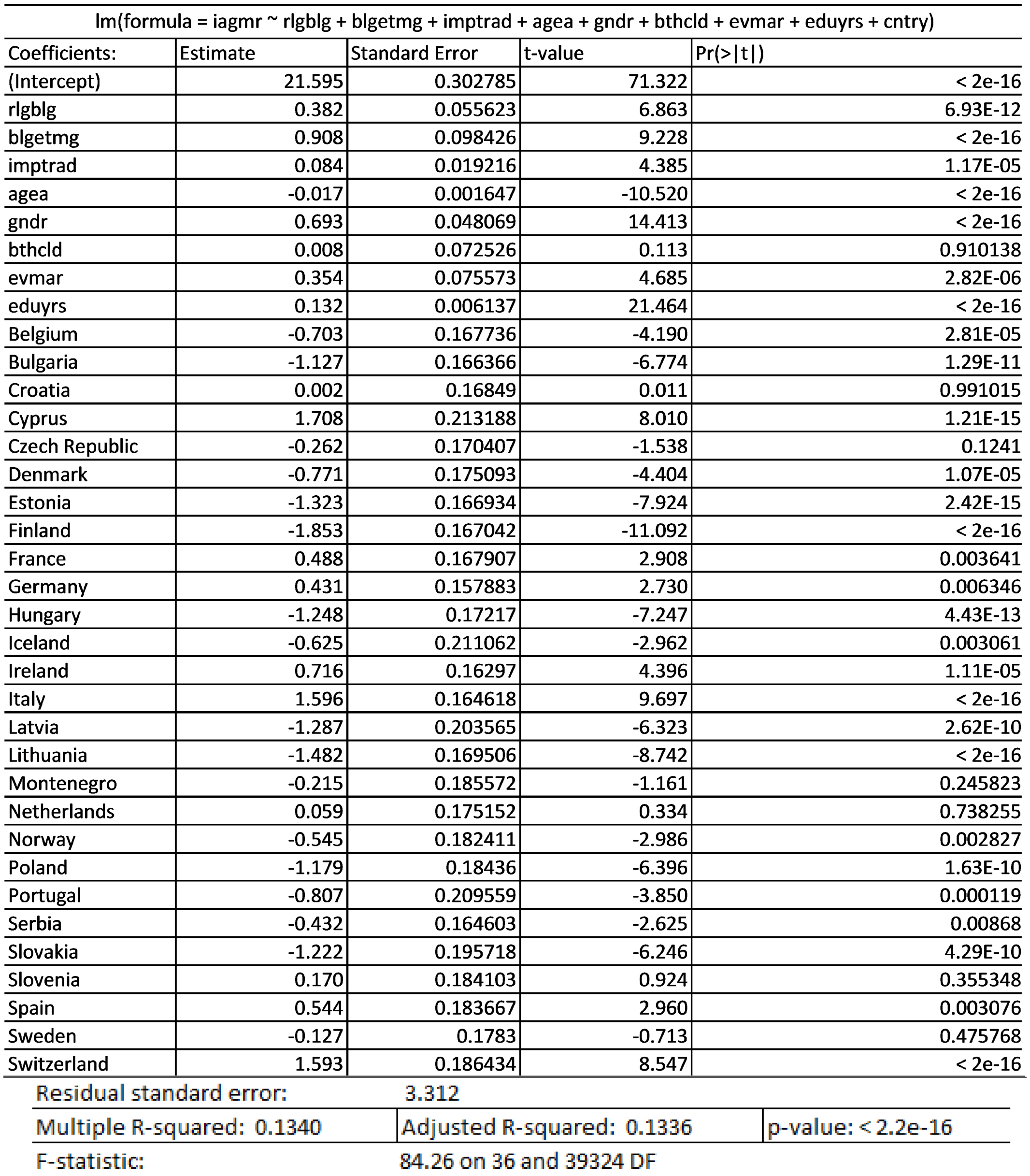
[12] Included due to the bigger number of parameters, which could cause overfitting.

[13] Do note that *iagmr* and *evmar* have a negative correlation, which however is expected and not justifying their exclusion: as one grows older, it is more likely that he/she will get married; same thing applies to *bthcld*, becoming more prone to having already had offspring.

4 Results

First and foremost, the regression model results show the presence of a relation between the independent variable and the response, showcased in detail in the following table.

Figure 5: Linear regression output



lm(formula = iagmr ~ rlgblg + blgetmg + imptrad + agea + gndr + bthcld + evmar + eduyrs + cntry)

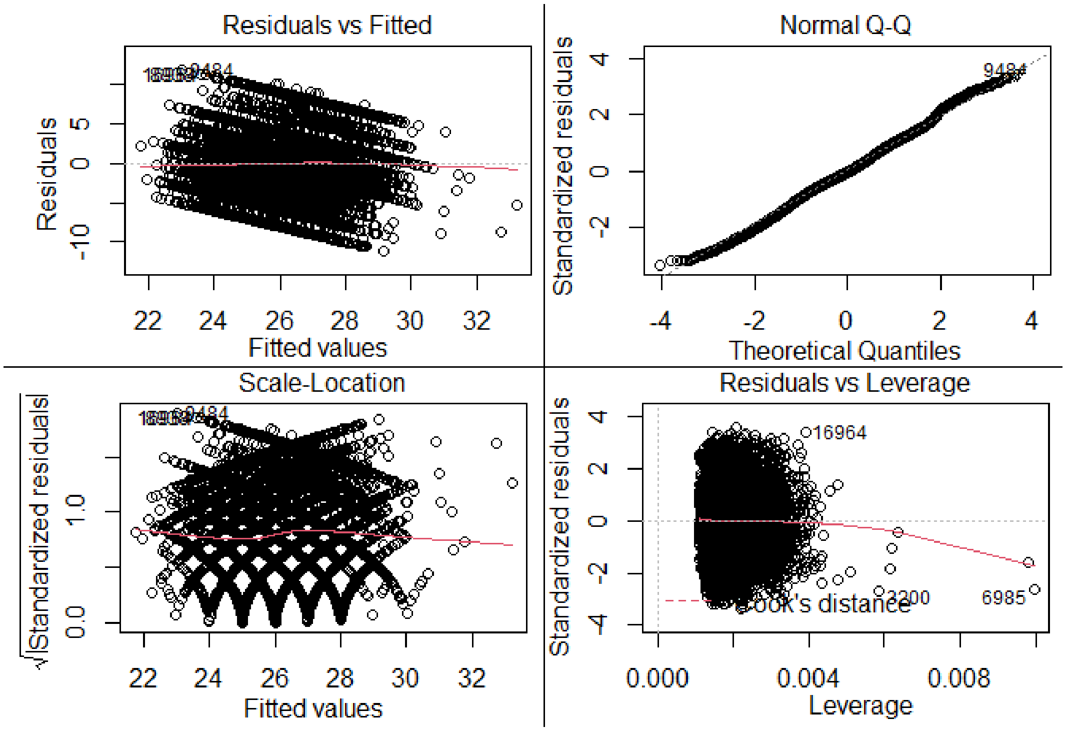
Coefficients:	Estimate	Standard Error	t-value	Pr(>|t|)
(Intercept)	21.595	0.302785	71.322	< 2e-16
rlgblg	0.382	0.055623	6.863	6.93E-12
blgetmg	0.908	0.098426	9.228	< 2e-16
imptrad	0.084	0.019216	4.385	1.17E-05
agea	-0.017	0.001647	-10.520	< 2e-16
gndr	0.693	0.048069	14.413	< 2e-16
bthcld	0.008	0.072526	0.113	0.910138
evmar	0.354	0.075573	4.685	2.82E-06
eduyrs	0.132	0.006137	21.464	< 2e-16
Belgium	-0.703	0.167736	-4.190	2.81E-05
Bulgaria	-1.127	0.166366	-6.774	1.29E-11
Croatia	0.002	0.16849	0.011	0.991015
Cyprus	1.708	0.213188	8.010	1.21E-15
Czech Republic	-0.262	0.170407	-1.538	0.1241
Denmark	-0.771	0.175093	-4.404	1.07E-05
Estonia	-1.323	0.166934	-7.924	2.42E-15
Finland	-1.853	0.167042	-11.092	< 2e-16
France	0.488	0.167907	2.908	0.003641
Germany	0.431	0.157883	2.730	0.006346
Hungary	-1.248	0.17217	-7.247	4.43E-13
Iceland	-0.625	0.211062	-2.962	0.003061
Ireland	0.716	0.16297	4.396	1.11E-05
Italy	1.596	0.164618	9.697	< 2e-16
Latvia	-1.287	0.203565	-6.323	2.62E-10
Lithuania	-1.482	0.169506	-8.742	< 2e-16
Montenegro	-0.215	0.185572	-1.161	0.245823
Netherlands	0.059	0.175152	0.334	0.738255
Norway	-0.545	0.182411	-2.986	0.002827
Poland	-1.179	0.18436	-6.396	1.63E-10
Portugal	-0.807	0.209559	-3.850	0.000119
Serbia	-0.432	0.164603	-2.625	0.00868
Slovakia	-1.222	0.195718	-6.246	4.29E-10
Slovenia	0.170	0.184103	0.924	0.355348
Spain	0.544	0.183667	2.960	0.003076
Sweden	-0.127	0.1783	-0.713	0.475768
Switzerland	1.593	0.186434	8.547	< 2e-16

Residual standard error:	3.312	
Multiple R-squared: 0.1340	Adjusted R-squared: 0.1336	p-value: < 2.2e-16
F-statistic:	84.26 on 36 and 39324 DF	

Do note that the country chosen as default is Austria, meaning that all coefficients are to be compared with Austria's one (set as 0). All computations were executed on R and compared with output produced using the same formula on Python.

Females tend to have a higher reported average ideal age of marriage; same thing applies to those who never got married or had a child.[14] Coherent with my original hypothesis, members of ethnic minority tend to perceive marriage as something more impending, seen as a mean through which strengthen ties with the social group. Regarding the religious aspect, the higher reported value for non-believers might be caused by the fact that several faiths emphasize traditions' role in society, with marriage being one of the most important as seen in previous paragraphs. As people grow older, age reported lowers as everyone's life view change. The accuracy of predicted values is of 0.135 and the MSE is 10.994. The R^2 is 0.134, meaning that 13.4% of the response variable's distribution can be explained using the chosen independent variables.

Figure 6: Regression model residual plots



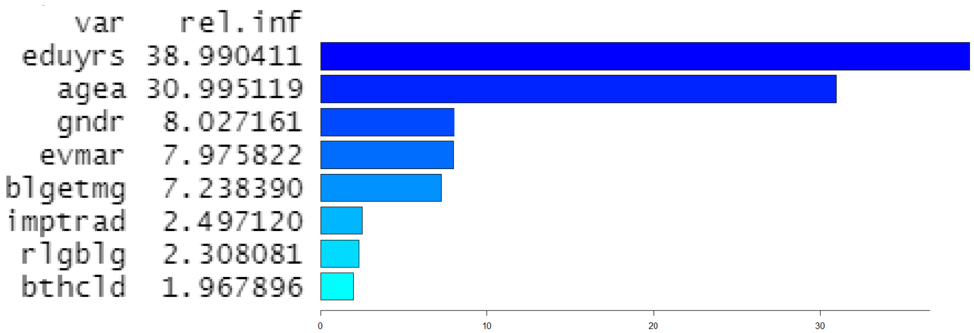
[14] Who could be attributing less importance to marriage due to their own life experiences.

Residual analysis showed that both the hypothesis of linearity and homoscedasticity hold with very few outliers. Based on Q-Q plot the residuals are normally distributed; according to the scale-location plot, residuals spread slightly further for higher values beyond 30. Lastly, considering Residual vs Leverage plot there seem to be some outliers, none of which being influential as the dots are still within Cook's Distance.

For Naïve-Bayes' computations I used the R package e1071; the output shows an accuracy of 0.300, which represents how precise the model's classification of the discrete variable is comparing it to the actual distribution.

For the regression tree I decided to exclude the variable *cntry*, as I wanted to focus on the impact of the other factors. As shown in Figure 4, individuals of an older age (above 62) think that marriage should happen before the age of 25, and those who studied more than 12 years[15] believe that marriage could happen later in life. The MSE for the pruned tree is 12.060. I then executed a boosting algorithm again excluding the country of origin (which had more than 50% of the relevance), choosing 300 as the number of trees through cross validation and setting interaction depth as default.

Figure 7: Boosting results and visualization



```
    var   rel.inf
 eduyrs 38.990411
   agea 30.995119
   gndr  8.027161
  evmar  7.975822
blgetmg  7.238390
imptrad  2.497120
 rlgblg  2.308081
 bthcld  1.967896
```

Boosting proved that age and years spent in education are the two most important factors, with the others still contributing (with the parenting condition as the last). This boosted algorithm had an MSE of 11.777, slightly lower than the pruned tree and with similar results.

My initial plan was to use R's library "modEvA" to perform a comparison between the models I have used for this paper, but I was unable to carry out this process for two main reason: first, because it is optimized for binary distributions or values between 0 and 1. Secondly, because the actual comparison can be made

[15] which means that they reached tertiary education, university or other institutions with corresponding level of formation.

only between *glm*, *gbm* or *RandomForest* types of models. In the final section of this paper, I will try to elaborate on what presented so far, describing how it correlates with the hypothesis of my research and what else could be added to improve both performance and reliability.

5 Conclusion

Both religious affiliation and foreign ethnic background seem to determine a different conception of marriage, to which a more urgent necessity is represented by the fact that the ideal age of marriage reported is lower when either one of these conditions are satisfied across all states of Europe. These two elements of one's identity do not have a significant degree of correlation and, therefore, should be considered two distinct features not interacting with each other.[16]

What surprised me the most was the impact of both one's age and education level, which are the two biggest contributors in defining one's opinion on the matter among those chosen for the assignment.

The output of all these models constitutes empirical proofs of what I intended to demonstrate; however, the way in which this occurs differs according to the method, which places more emphasis on certain aspects rather than others. All things considered, I consider the linear model the most appropriate, not only because of the linear relationship between discrete variables proved by the LOOCV procedure but also because unlike the others it shows country of origin's impact, whose exclusion is detrimental for the analysis.[17]

For this study I decided not to include random forest methods, as the intention was mostly to provide an explanatory research and size constraint would not have given justice to models with such depth. If this project was to be expanded, random forest would provide stable and clear results requiring heavier computations due to the sample size (which could also be expanded, for example including more observation of European Social Survey's round 6[18] in this study).

I do not claim to provide a definitive and categorical answer to this association; nonetheless, I believe that showcasing how the two elements concerning one's affiliations and identity interact with the conception of marriage can be an adequate starting point for future analysis on the matter.

[16] Trying an interaction factor in the regression proved this statement.

[17] excluding country's relative influence during boosting was intentional but does not imply that it is negligible.

[18] which also contains the parameter *iagmr* and would introduce the aspect of passing of time and how society's opinions diverge on different time frames.

6 Bibliography

˜ Atkinson J.. “Gender Roles in Marriage and the Family: A Critique and Some Proposals”. Journal of Family Issues. Vol. 8. No. 1. 1987. pp. 5-41.

˜ Beck-Gresnheim E.. “Transnational lives. transnational marriages: a review of the evidence from migrant communities in Europe”. Global networks. Vol. 7. No. 3. 2007. pp. 271-288.

˜ Berelli-Harris P. et al.. “How Similar Are Cohabitation and Marriage? Legal Approaches to Cohabitation across Western Europe”. Population and development review. Vol. 38. No. 3. 2012. pp.435-467.

˜ Holland J. A. et al.. “Ideal ages for family formation among immigrants in Europe”. Advances in Life course Research. Vol. 18. No. 4. 2013. pp. 257-269.

˜ Howe R. T. “Marriages and Families in the 21st Century: a Bioecological Approach”. 2017.

˜ Kalmijn L. “Explaining cross-national differences in marriage. cohabitation. and divorce in Europe. 1990–2000”. Population Studies. Vol. 61. No. 3. 2007. pp. 243-263.

˜ Lesthaeghe R. “The second demographic transition: A concise overview of its development.” Proceedings of the National Academy of Sciences. Vol. 111. No. 51. 2014. pp. 18112-18115.

˜ Lelkes O.. “Happiness Across the Life Cycle: Exploring Age-Specific Preferences”. Policy Brief 3.2. 2008.

˜ Milewski N.. “Mixed Marriages in Germany: A High Risk of Divorce for Immigrant-Native Couples”. European Journal of population. Vol. 30. No. 1. 2014. pp. 89-113.