Universidad Carlos III de Madrid

Survey Methodology II: Final Project

Support towards the transgender community: analysis of individual and country level variables.

Authors:

Clara Espinosa Acevedo

Daniel Pérez Gutiérrez

Carlos San Juan Baeza

Introduction.

A big part of the civil rights debate in Europe in recent decades has been centered on the situation of transgender individuals. Legal Gender Recognition is the one step towards an equal situation between cis-gender and transgender people. The situation in Europe regarding Legal Gender Recognition is very diverse, the difference usually resides on the process of changing the information and the grade of self-determination allowed. In this project we want to understand the relation between an individual being in favour or against Legal Gender Recognition and different variables.

To fulfil this objective, we will be using data from the 2019 Eurobarometer Survey, from which we will take individual-level variables (whether it is sociodemographic information or variables regarding opinion). Also, we will add variables from three external sources: we will extract the political sign of the government with three variables that indicate the percentage of politicians in government that are center, right or left-winged from the Comparative Political Data Set; we will create a Transgender Protection Index with information from the TGEU Trans Rights Map; and lastly, we will add socioeconomic indicators extracted from The World Bank.

We part from two research questions: (1) is there a difference between countries when it comes to a level of support towards Legal Gender Recognition measures? And (2) are we able to predict if a European Citizen would support a measure regarding Legal Gender Recognition?

Methodology.

We will take as a target variable the question QC19 from the Questionnaire that says: "Do you think that transgender persons should be able to change their civil documents to match their inner gender identity?" The possible answers to this question are Yes, No and Don't Know (Which we will interpret as NAs). The methodology will consist, first on a descriptive analysis of the interaction of this target variable with other features (on an individual and country level). Then we will carry out the imputation of missing data on certain variables. After this first part, we will use Unsupervised-Learning methods to understand the most important variables (PCA) and the distribution of the countries in different groups (Clustering). Finally, we will implement Supervised-Learning methods to carry out the prediction and see the results of the best method.

Descriptive analysis.

The target variable in our dataset is split almost at 50% (47.29% doesn't support Legal Gender Recognition, and 52.71% supports it) (see Figure 1). On a country level, the opinion of citizens is normally related to the steps their government takes towards equality (Kwiatkowski, 2023), this can be seen in the map shown in Figure 2 in the Appendix: those countries that show a higher proportion of people being in favour, also show a higher Protection of Trans People Index, the correlation between these variables is 0.3 (for more information on the correlation of the target variable with other variables we recommend looking to the code or to Figures 3, 4 and 5 in the Appendix).

Other country-level variables that are correlated with the proportion of support to Legal Gender Recognition are GDP per capita (Jones et al., 2018) and the political sign in the government, especially the percentage of right-wing politicians in the government, the visualization of the interaction of these variables can be seen in Figure 6.

On an individual level, we can see how gender, social class and age interact with the target variable on Figure 7, males and people from the lower classes seem to show less support than middle, high classes and female respondents. The evolution of Support can be seen in Figure 8, there is a clear drop of support in elder people.

Preparation of data: NAs imputation.

To continue with our analysis, we have to decide if we should impute NAs in any variable, our dataset didn't have many variables with high proportion of NAs. We imputate our variables with different methods and then using trans_discrimination, sexorient_discrimination (opinion about trans people and gay, lesbian and bisexual people being discriminated), as well as social class to compare the distributions and see how well our imputatitions fit our data.

We finally select random forest imputation based on the barplots (see Figure 9)

PCA and Clustering analysis.

We conduct a Principal Component Analysis (PCA) that is a technique that reduces the dimensionality of complex datasets by transforming correlated variables into a smaller set of uncorrelated variables. These principal components capture most of the variability present in the original data, facilitating visualization, interpretation, and analysis. Within the context of assessing countries' development in combating transphobia, PCA could uncover patterns and correlations among various indicators that might not be immediately apparent.

What is the most developed country at the time to fight to transfobia?

What are the main indicators contributing to the effective fight against transphobia?

Which countries are most important in explaining the variability of the data?

Can we rank the countries considering all the variables at the same time?

First Principal Component: The first principal component (PC1) accounts for 20.5% of the total variability in our data (see Figure 10) This component appears to be primarily linked to aspects related to the socioeconomic and demographic development of the countries included in our analysis. In other words, the variables that contribute most significantly to the direction and magnitude of PC1 are related to the economic, social, and demographic status of these countries.

Second *Principal Component:* The variables with the highest weight in PC2, related to age and discrimination, reflect important aspects of demographic structure and social dynamics (see Figure 11)

Additionally, clustering based on the scores of PC1 and PC2 provides an effective way to group countries according to their position in this reduced-dimensional space. By jointly analyzing these two principal components, we can identify clustering patterns and explore how common socioeconomic and demographic characteristics may influence the formation of these groups. Visualization in a two-dimensional plane facilitates the interpretation of results and helps understand the relationships between countries in terms of their development.

As we can see in Figure (12) , we can group countries into 4 clusters based on PC1 and PC2.

Group 1: These countries, including Cyprus, Croatia, and Portugal, among others, exhibit potential for growth in both economic and social aspects. While they may not currently boast the highest levels of development, they present opportunities for improvement in terms of policies and support systems for the LGBT community.

Group 2: Luxembourg emerges as a standout performer in socio-economic development and legal protection for transgender individuals. With robust laws safeguarding the rights of transgender people and a society characterized by high levels of acceptance, Luxembourg provides an exemplary model of inclusivity and support for the transgender community.

Group 3: This cluster comprises countries like the Netherlands, Finland, and Spain, which excel in both economic and social development. Notably, they demonstrate strong legal and social support for the transgender community, reflecting progressive attitudes and policies in this regard.

Group 4: Eastern European countries, such as Romania and Hungary, dominate this category, characterized by lower levels of socio-economic development and limited legal protections for transgender individuals. These nations face challenges in providing adequate support and legal frameworks for the LGBT community.

Group 5: Representing economically advanced countries with room for improvement in LGBT rights, this group includes the United Kingdom, France, and Ireland. While they exhibit significant economic advancement, they still require enhancements in legal and social measures to fully protect and support the LGBT community.

Prediction of values.

In this part of the research, we try to respond our second Research Question. For the prediction of our target variable, we have used different models, starting with a basic Logit Model. The variables chosen for these models are those with highest correlation with the target variable. The Logit Model reveals significant patterns in variables such as age or life_satisfaction (for more information we recommend seeing Table 1). Also, age and gender are relevant in the model: younger people, especially women are more likely to support the transgender collective (Figure 13 and 14).

After performing the Logit Model, we tried predicting with for more advanced Machine Learning models: Random Forest, Linear Discriminant Analysis (LDA), ElastcNet and Gradient Boosting. The first three show similar metric results (a comparison can be seen in Table 2): all around 70% for Accuracy and 68% for Sensitivity. Nevertheless, we decided to use Gradient Boosting, since is the one that has obtained better results. To support this decision, we have plotted the ROC curve for all models, and we've observed that Gradient Boosting shows the highest AUC (Figures 15 to 18).

Once we selected the model, we checked which variables affected the probability of supporting or not Legal Gender Recognition, and we identified that people who support same-sex marriages, who have positive attitudes towards the LGBT community and protection policies for transgender people are significant predictors of our target variable. In Figure 19 we can see that those variables increase the likelihood of supporting Legal Gender Recognition, as well as how they affect the target variable, in case the respondent is Spanish.

Conclusions.

The analysis of support towards the transgender community in Europe reveals a nuanced landscape shaped by both individual and country-level variables. Across Europe, attitudes towards Legal Gender Recognition (LGR) vary significantly, with most developed european countries like Netherlands, Finland or Spain exhibiting more advanced measures to support the transgender community compared to Eastern European countries such as Romania and Hungary, where there is a widespread issue with the protection and inclusion of the LGBT+ community, particularly for transgender individuals. These countries exhibit significantly lower levels of legal and social protection compared to other regions.

At the individual level, gender, social class, and age emerge as significant determinants of support for LGR. Women, younger individuals, and those from higher social classes are more likely to endorse Legal Gender Recognition measures.

In predictive modeling, machine learning techniques, particularly Gradient Boosting, demonstrate high accuracy and sensitivity in predicting support for LGR, outperforming traditional logistic regression. Key variables such as age, gender, attitudes towards the LGBT community, and support for same-sex marriages play pivotal roles in these predictions.

These findings have implications for policymakers and advocates aiming to promote greater support for Legal Gender Recognition across Europe. Tailored interventions informed by an understanding of the complex interplay of individual and country-level factors are crucial for advancing civil rights and fostering inclusivity for transgender individuals. Continued research and monitoring of attitudes and policies towards the transgender community are essential for creating a more equitable and supportive society in Europe.

Bibliography

Armingeon, K., Engler, S., Leemann, L., & Weisstanner, D. (2023). *Comparative Political Data Set 1960-2021*. University of Zurich, Leuphana University Lueneburg, and University of Lucerne: Zurich/Lueneburg/Lucerne.

Jones, P., Brewer, P. R., Hoffman, L. H., Lambe, J., & Young, D. (2018). Explaining Public Opinion toward Transgender People, Rights, and Candidates. Public Opinion Quarterly, 82(2), 252-278.

Kwiatkowski, P. (2023). The European Standard of Legal Gender Recognition. Teka Komisji Prawniczej PAN Oddział w Lublinie. https://doi.org/10.32084/tkp.5463

Transgender Europe (TGEU). (2024). *Trans Rights Map Dataset*. Retrieved from https://transrightsmap.tgeu.org/index

World Bank. (2024). *World Development Indicators* [Data set]. Retrieved from https://databank.worldbank.org/source/world-development-indicators

Appendix

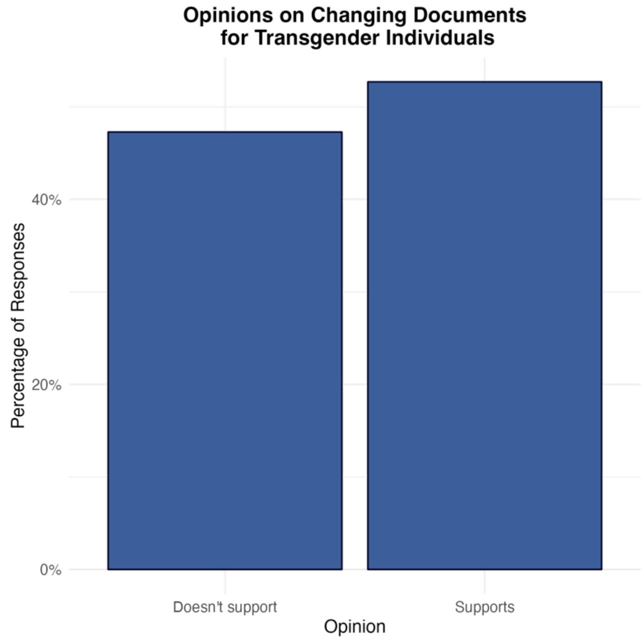

Figure 1. Distribution of the target variable

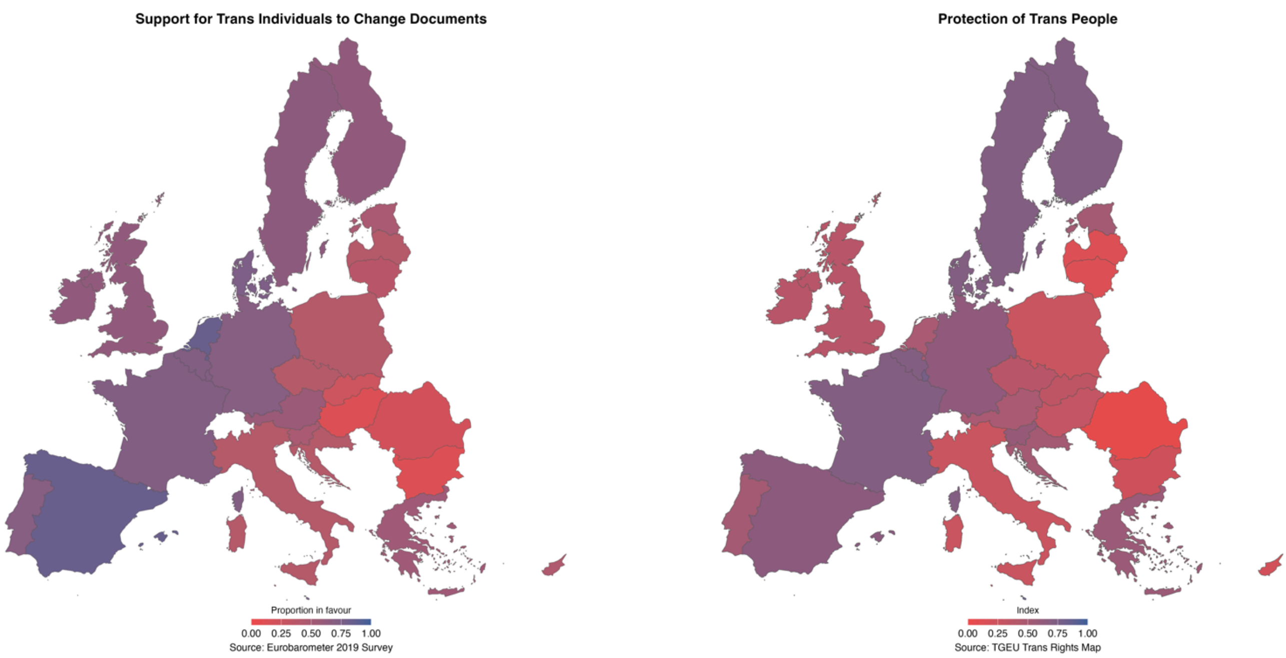

Figure 2. Comparison of Proportion Support on Survey and Protection Index

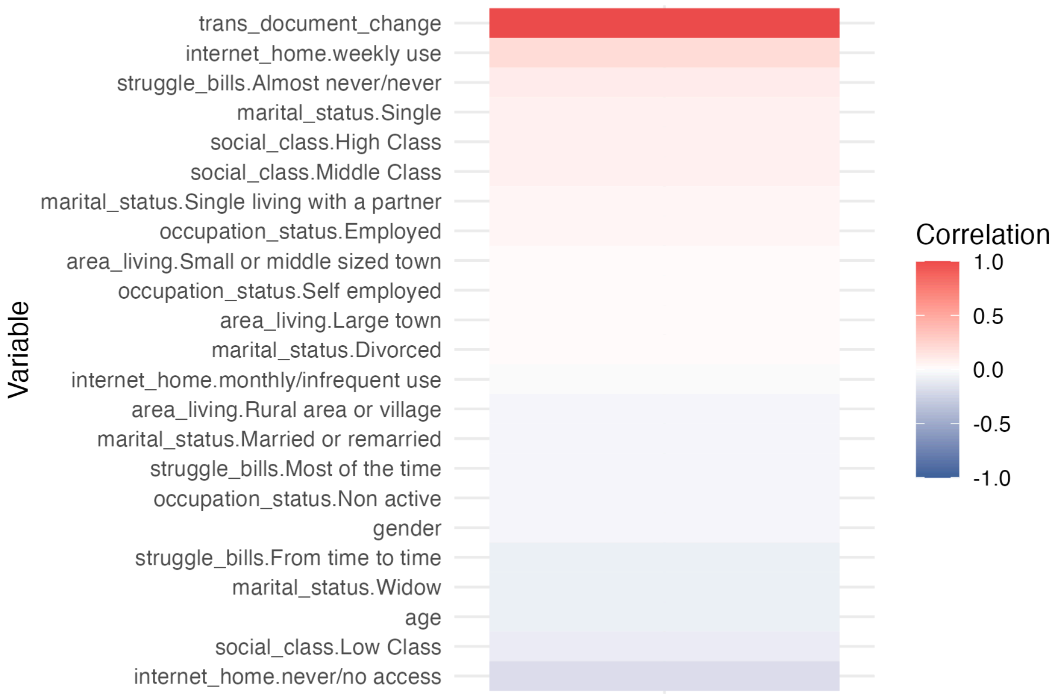

Figure 3. Correlation of Target Variable with Sociodemographic Variables

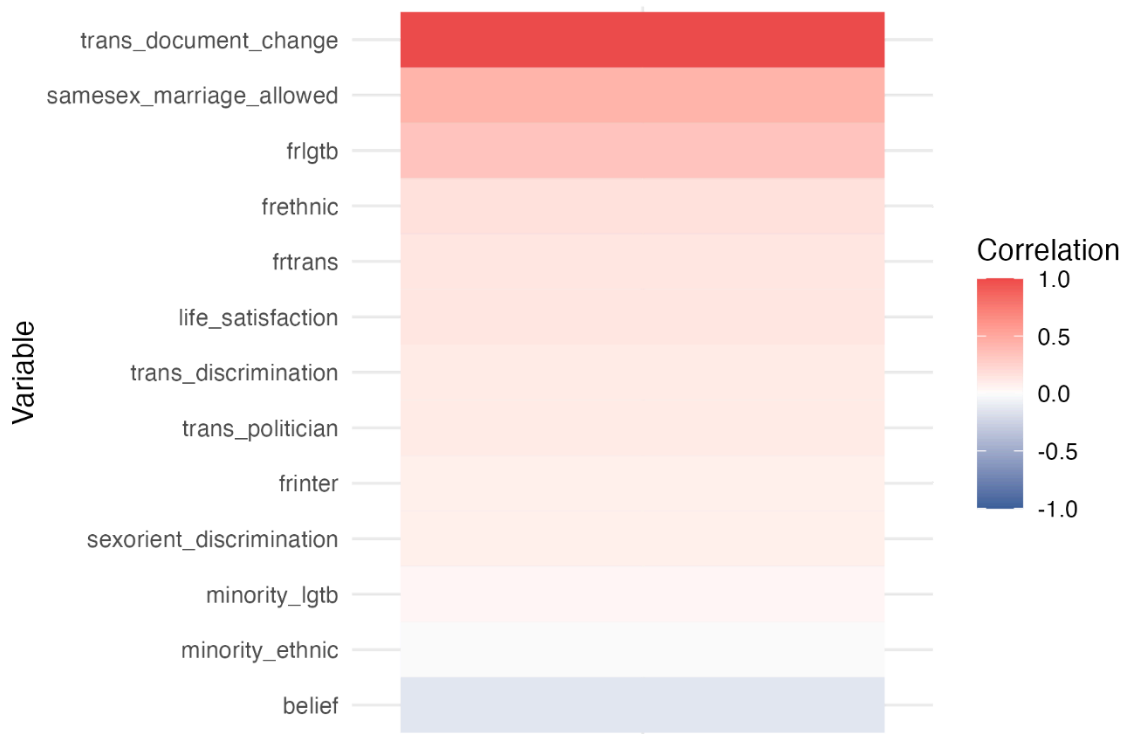

Figure 4. Correlation of Target Variable with Opinion Variables

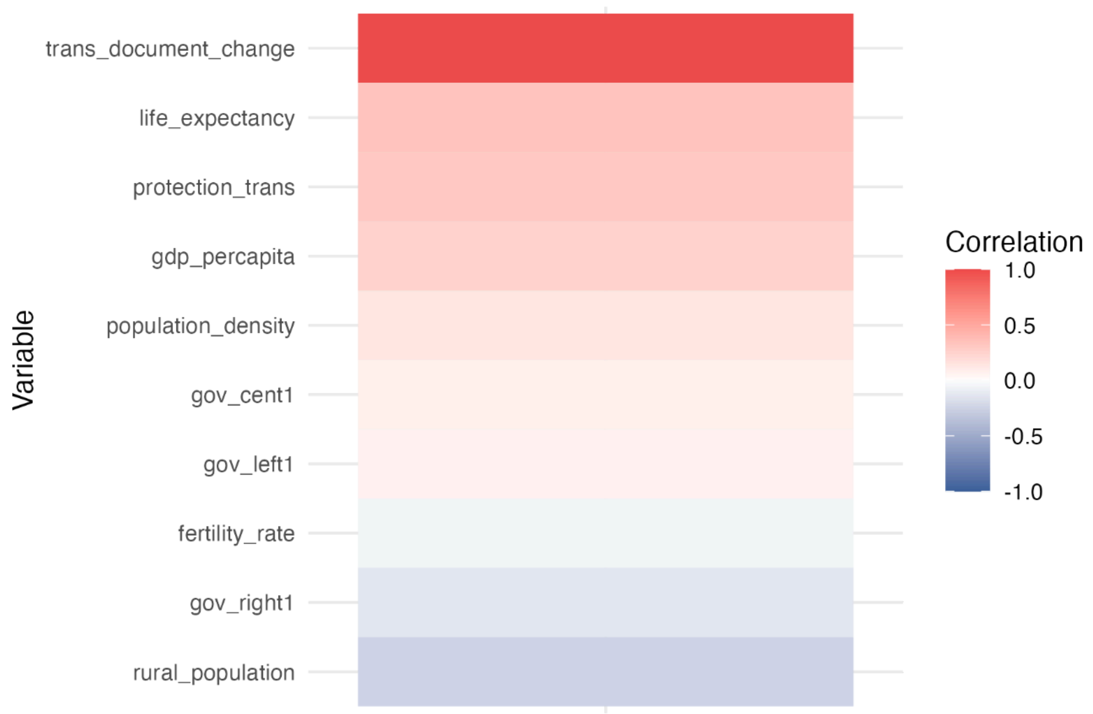

Figure 5. Correlation of Target Variable with Country-Level Variables

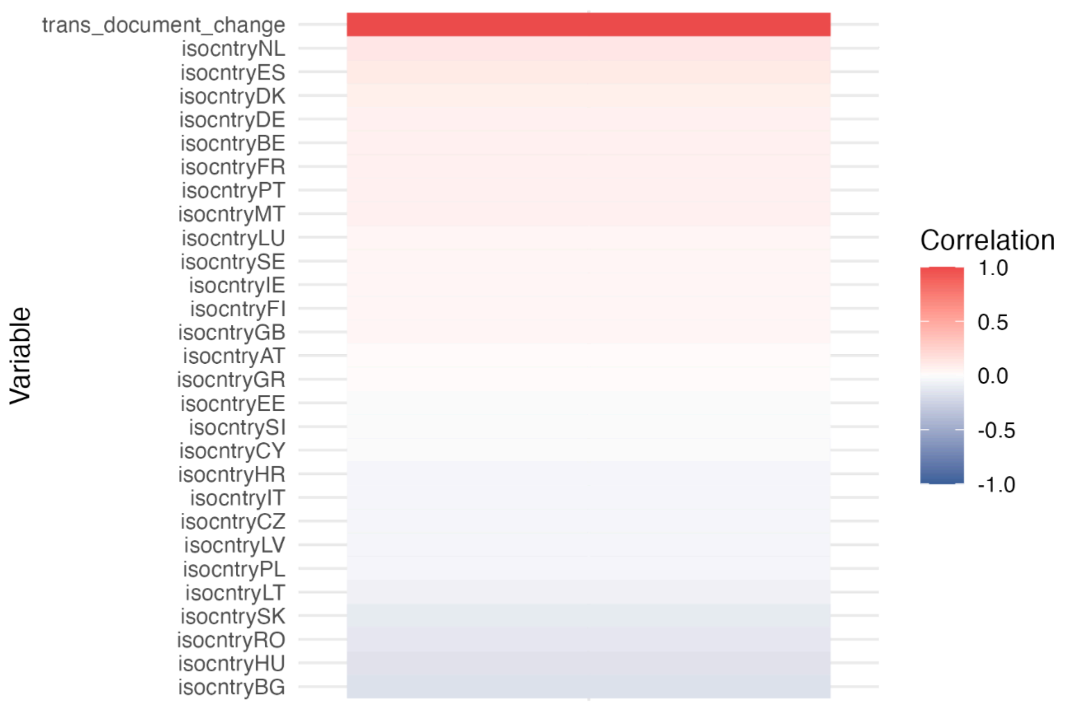

Figure 6. Correlation of Target Variable with Country Variables

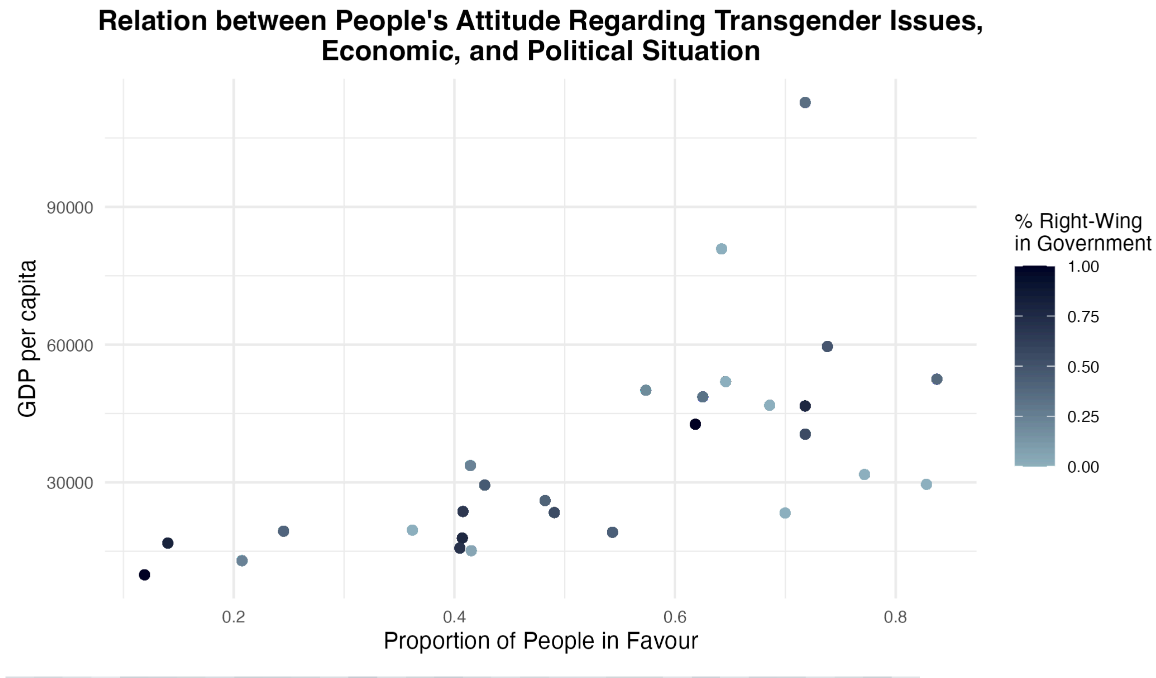

Figure 7. Interaction between Target Variable, GDP Per Capita, and Political Situation

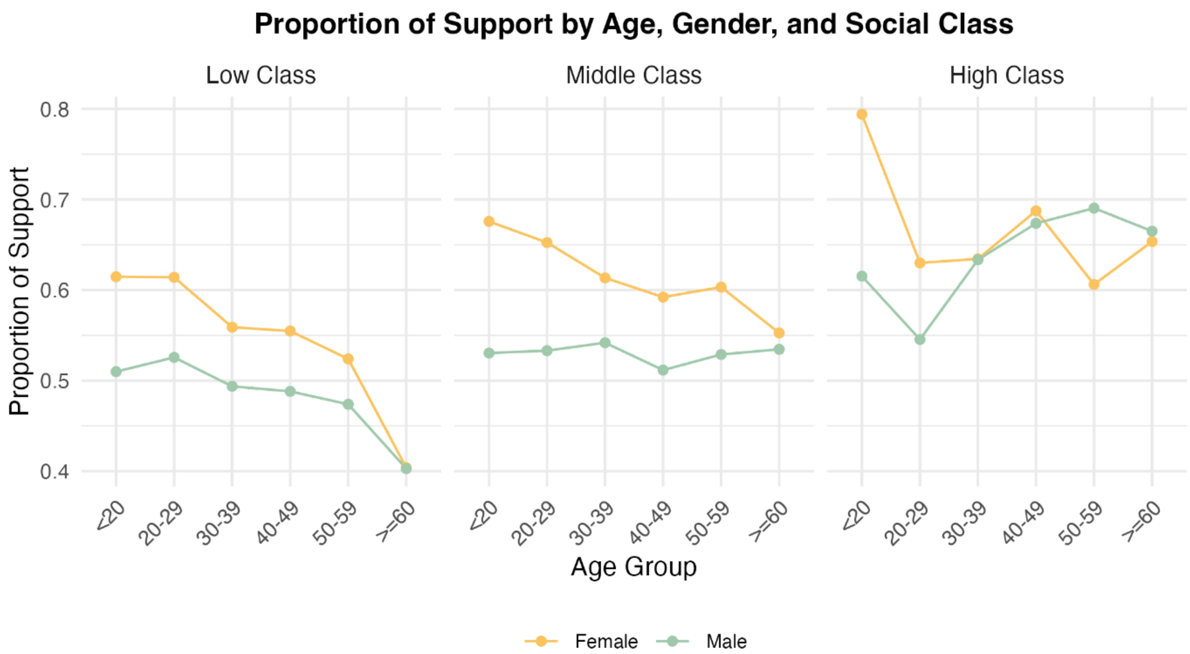

Figure 8. Relation between the Target Variable, Age, Gender and Social Class.

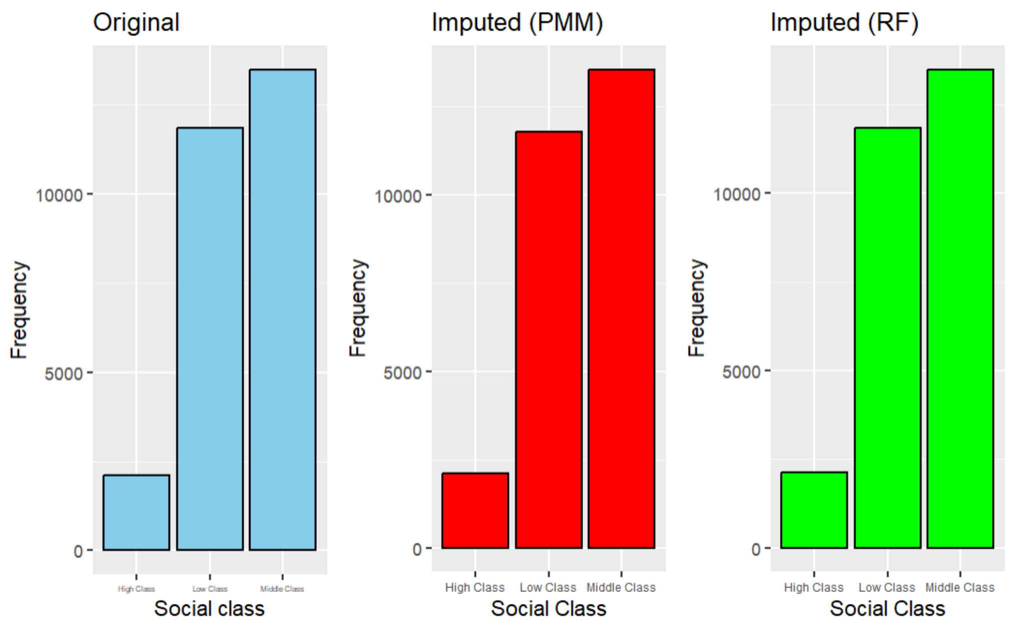

Figure 9. Comparation of imputation methods.

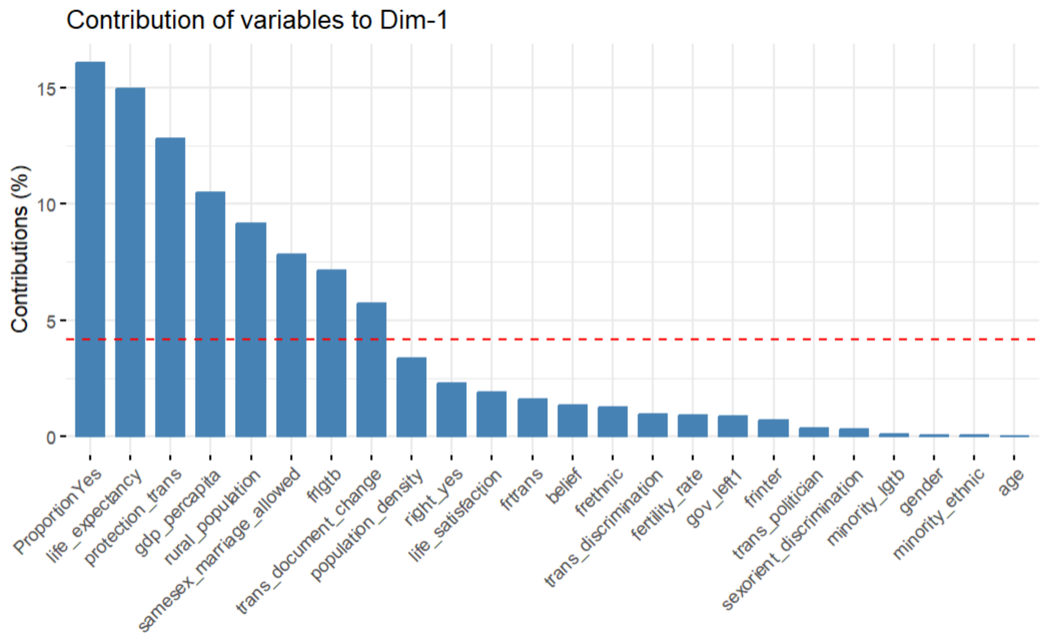

Figure 10. First Component Analysis (PC1)

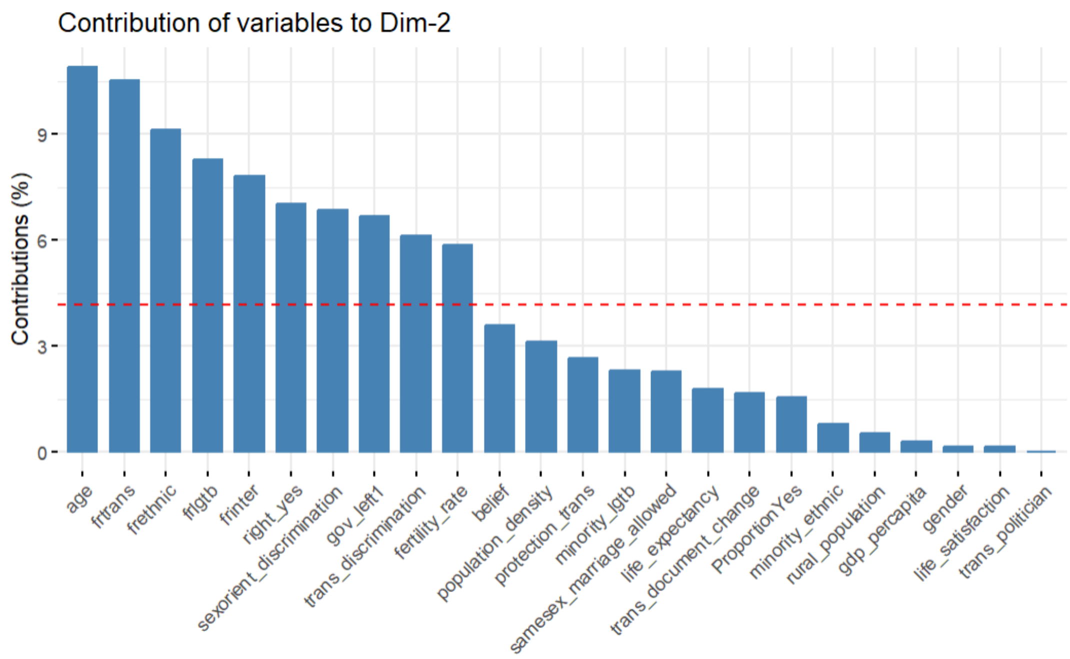

Figure 11. Second Component Analysis (PC2)

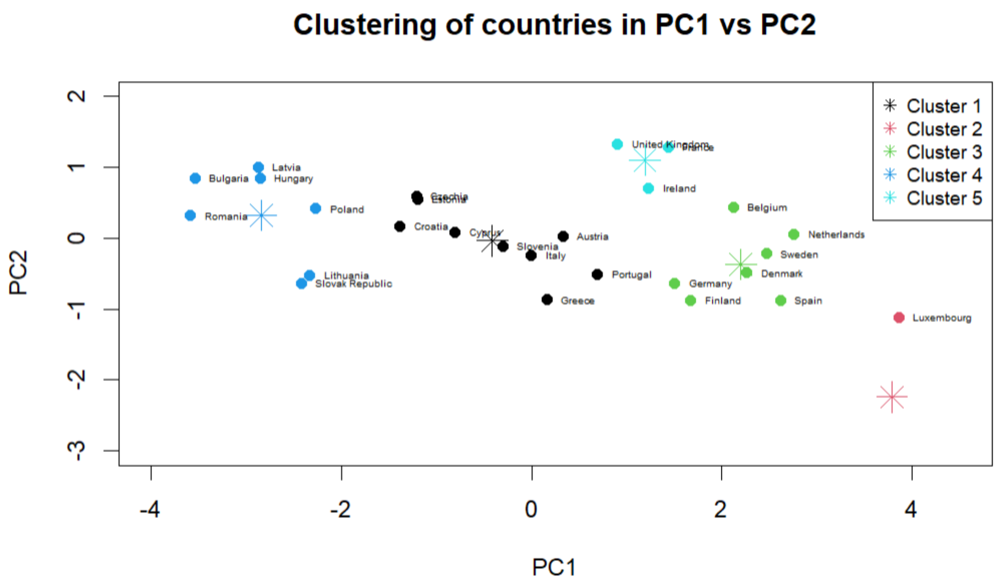

Figure 12. Clustering of countries in PC1 vs PC2

Table 1. Logit Model.

```
Call:
glm(formula = trans_document_change ~ age + life_satisfaction +
    frethnic + frlgtb + frtrans + minority_ethnic + sexorient_discrimination +
    trans_politician + gender + protection_trans + life_expectancy +
    gdp_percapita + population_density + rural_population + fertility_rate +
    gov_right1 + samesex_marriage_allowed + belief + countrynameBulgaria +
    countrynameHungary + countrynameRomania + countrynameNetherlands +
    countrynameSpain, family = binomial(link = "logit"), data = training)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.872e+00  1.015e+00   2.830 0.004651 **
age                      -4.042e-03  9.510e-04  -4.251 2.13e-05 ***
life_satisfaction         2.497e-01  4.527e-02   5.515 3.49e-08 ***
frethnic                  3.254e-01  3.523e-02   9.237  < 2e-16 ***
frlgtb                    4.731e-01  3.967e-02  11.924  < 2e-16 ***
frtrans                   3.850e-01  6.161e-02   6.249 4.13e-10 ***
minority_ethnic          -3.074e-01  9.153e-02  -3.359 0.000783 ***
sexorient_discrimination  2.151e-01  3.289e-02   6.538 6.23e-11 ***
trans_politician          5.397e-01  6.570e-02   8.214  < 2e-16 ***
gender                   -2.031e-01  3.248e-02  -6.254 4.01e-10 ***
protection_trans          9.340e-01  1.278e-01   7.310 2.67e-13 ***
life_expectancy          -4.986e-02  1.218e-02  -4.092 4.27e-05 ***
gdp_percapita            -1.041e-06  1.229e-06  -0.847 0.396867
population_density        4.374e-04  9.721e-05   4.500 6.80e-06 ***
rural_population         -8.755e-03  2.048e-03  -4.274 1.92e-05 ***
fertility_rate           -3.328e-01  1.323e-01  -2.515 0.011899 *
gov_right1               -1.314e-03  6.746e-04  -1.947 0.051508 .
samesex_marriage_allowed  1.389e+00  3.671e-02  37.847  < 2e-16 ***
belief                   -1.388e-01  4.129e-02  -3.361 0.000777 ***
countrynameBulgaria      -1.536e+00  1.278e-01 -12.024  < 2e-16 ***
countrynameHungary       -1.821e+00  1.208e-01 -15.077  < 2e-16 ***
countrynameRomania       -6.028e-01  1.105e-01  -5.454 4.94e-08 ***
countrynameNetherlands    4.635e-01  1.084e-01   4.278 1.89e-05 ***
countrynameSpain          7.494e-01  1.170e-01   6.407 1.49e-10 ***
---
Signif. codes:  0 ´***´ 0.001 ´**´ 0.01 ´*´ 0.05 ´.´ 0.1 ´ ´ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30371  on 21950  degrees of freedom
Residual deviance: 23674  on 21927  degrees of freedom
AIC: 23722

Number of Fisher Scoring iterations: 4
```

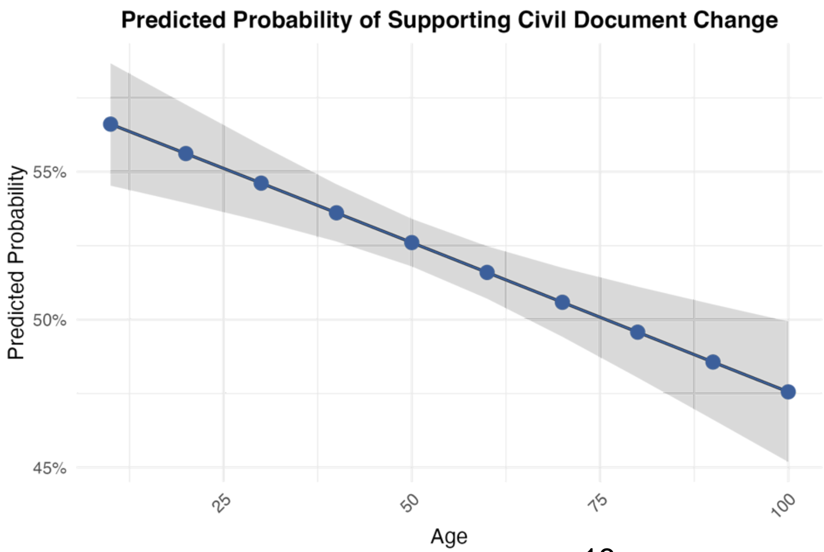

Figure 13. Predicted probabilities for Age in Logit Model

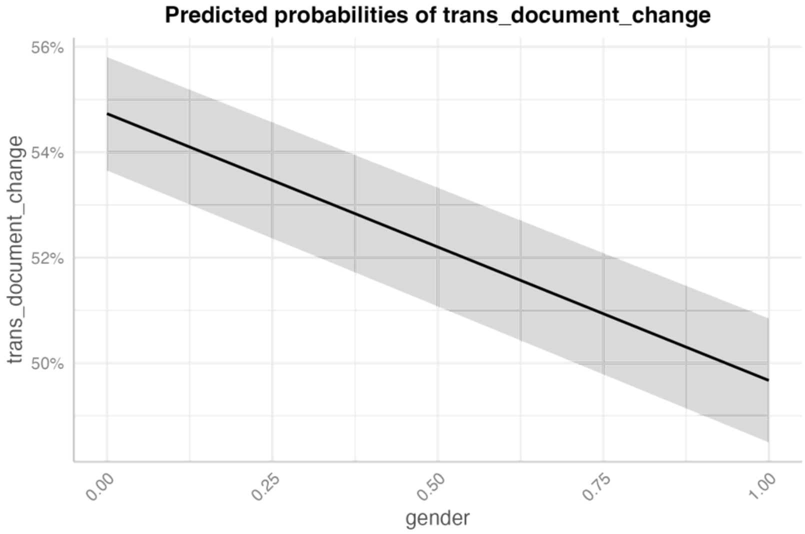

Figure 14. Predicted probabilities for Gender in Logit Model

Table 2. Comparison of Metrics for Different Models.

	Logit Model	Random Forest	LDA	ElasticNet	Gradient Boosting
Accuracy	74.07%	73.1%	74.3%	74.3%	74.18%
Sensitivity	67.54%	70.22%	66.91%	66.95%	70.07%
AUC		0.7971	0.7957	0.7956	0.8045

Table 3. Output of LDA model.

```
Call:
lda(trans_document_change ~ age + life_satisfaction + frethnic +
    frlgtb + frtrans + minority_ethnic + sexorient_discrimination +
    trans_politician + gender + protection_trans + life_expectancy +
    gdp_percapita + population_density + rural_population + fertility_rate +
    gov_right1 + samesex_marriage_allowed + belief + countrynameBulgaria +
    countrynameHungary + countrynameRomania + countrynameNetherlands +
    countrynameSpain, data = training)

Prior probabilities of groups:
        0         1
0.4740559 0.5259441

Group means:
       age life_satisfaction  frethnic    frlgtb    frtrans minority_ethnic
0 53.36767         0.7610994 0.5227753 0.2291947 0.04987507      0.03815107
1 50.14353         0.8855782 0.6948463 0.5652663 0.14118666      0.02572542
  sexorient_discrimination trans_politician    gender protection_trans
0                0.4567557        0.9039977 0.4750144        0.3902307
1                0.5474231        0.9570377 0.4371589        0.5197921
  life_expectancy gdp_percapita population_density rural_population
0        79.29388      29034.51           128.3285         29.90013
1        81.12461      39578.48           192.9368         23.75169
  fertility_rate gov_right1 samesex_marriage_allowed    belief
0       1.554299   43.39672                0.3743994 0.8362483
1       1.533704   33.86283                0.8092681 0.7136423
```

```
  countrynameBulgaria countrynameHungary countrynameRomania
0          0.07168941        0.068902556         0.06505862
1          0.00822867        0.008575141         0.01472499
  countrynameNetherlands countrynameSpain
0             0.01316548       0.01278109
1             0.05881334       0.05872672

Coefficients of linear discriminants:
                                   LD1
age                      -3.383118e-03
life_satisfaction         1.981010e-01
frethnic                  2.756476e-01
frlgtb                    4.462126e-01
frtrans                   2.674750e-01
minority_ethnic          -2.582760e-01
sexorient_discrimination  1.740619e-01
trans_politician          4.418286e-01
gender                   -1.623776e-01
protection_trans          8.256169e-01
life_expectancy          -3.529943e-02
gdp_percapita            -7.301750e-07
population_density        3.388543e-04
rural_population         -8.052836e-03
fertility_rate           -2.398755e-01
gov_right1               -1.171671e-03
samesex_marriage_allowed  1.353490e+00
belief                   -1.124902e-01
countrynameBulgaria      -1.014657e+00
countrynameHungary       -1.276988e+00
countrynameRomania       -4.447913e-01
countrynameNetherlands    3.368942e-01
countrynameSpain          4.963962e-01
```

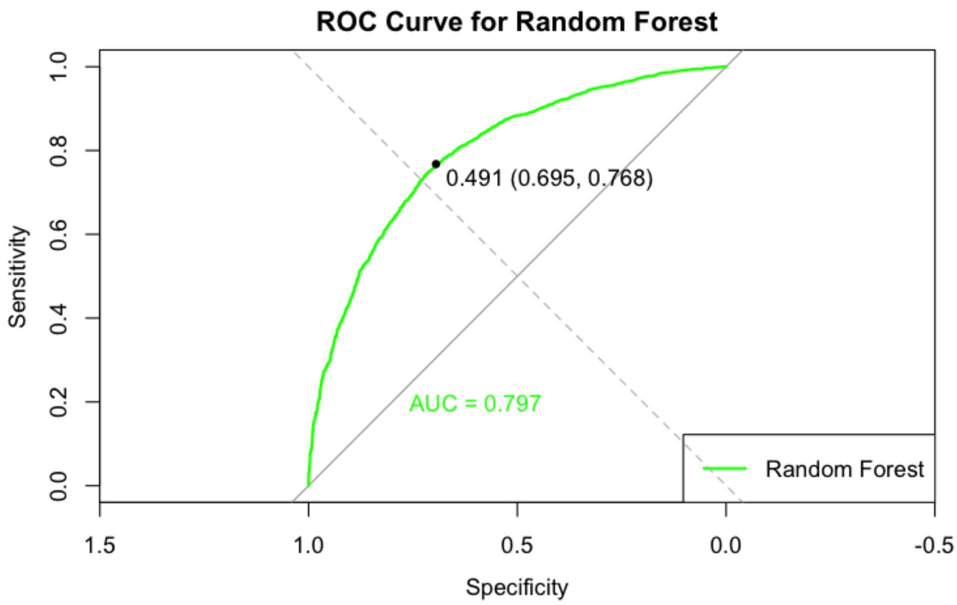

Figure 15. ROC Curve for Random Forest

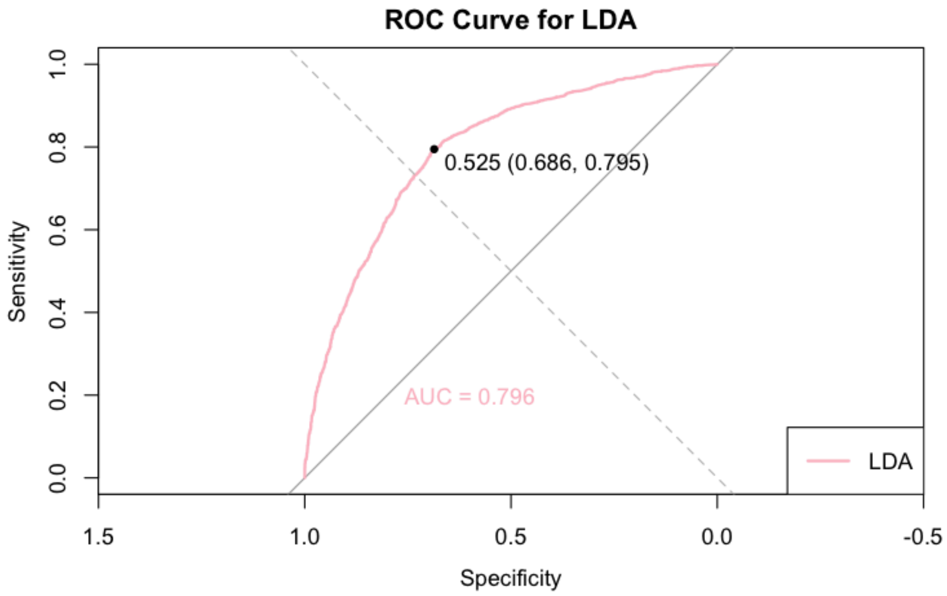

Figure 16. ROC Curve for LDA

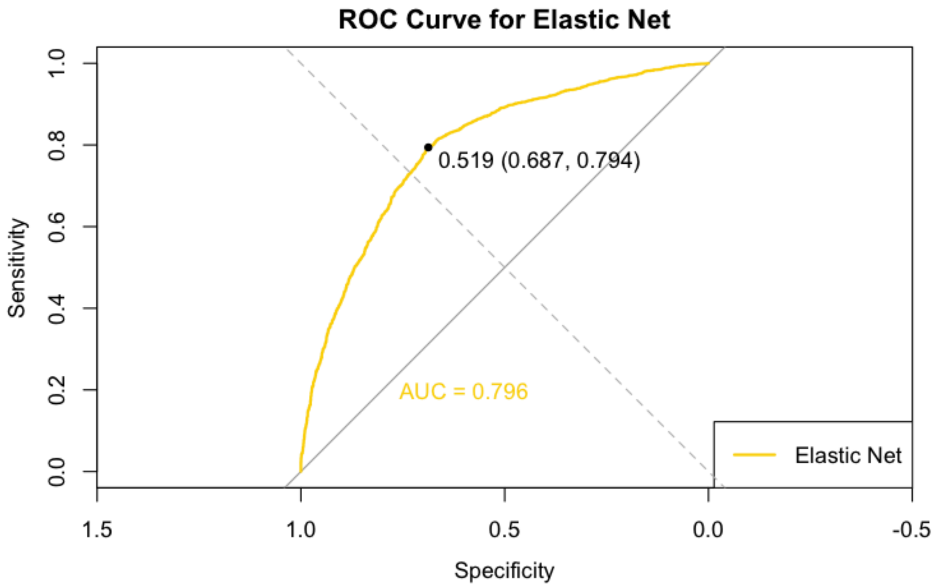

Figure 17. ROC Curve for ElasticNet

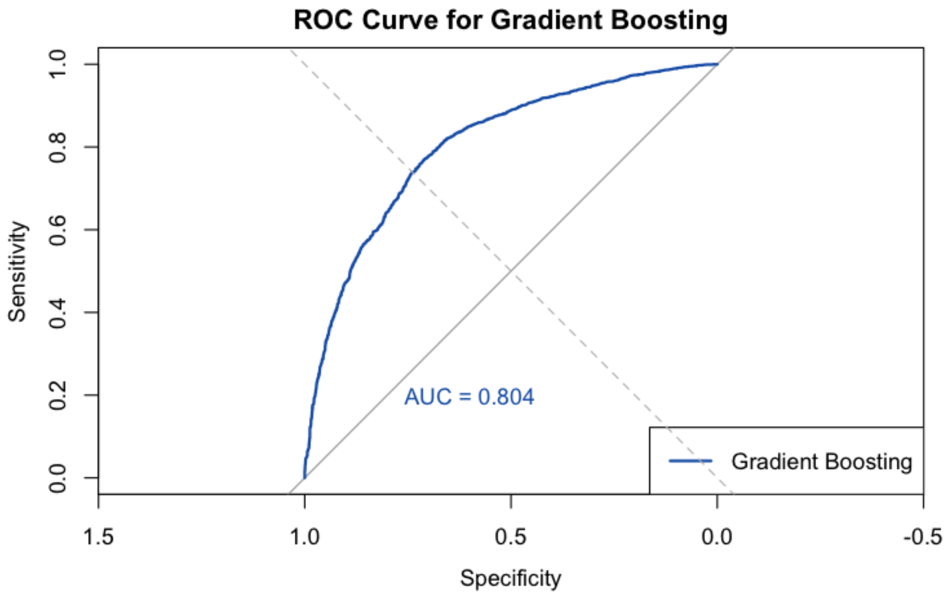

Figure 18. ROC Curve for Gradient Boosting

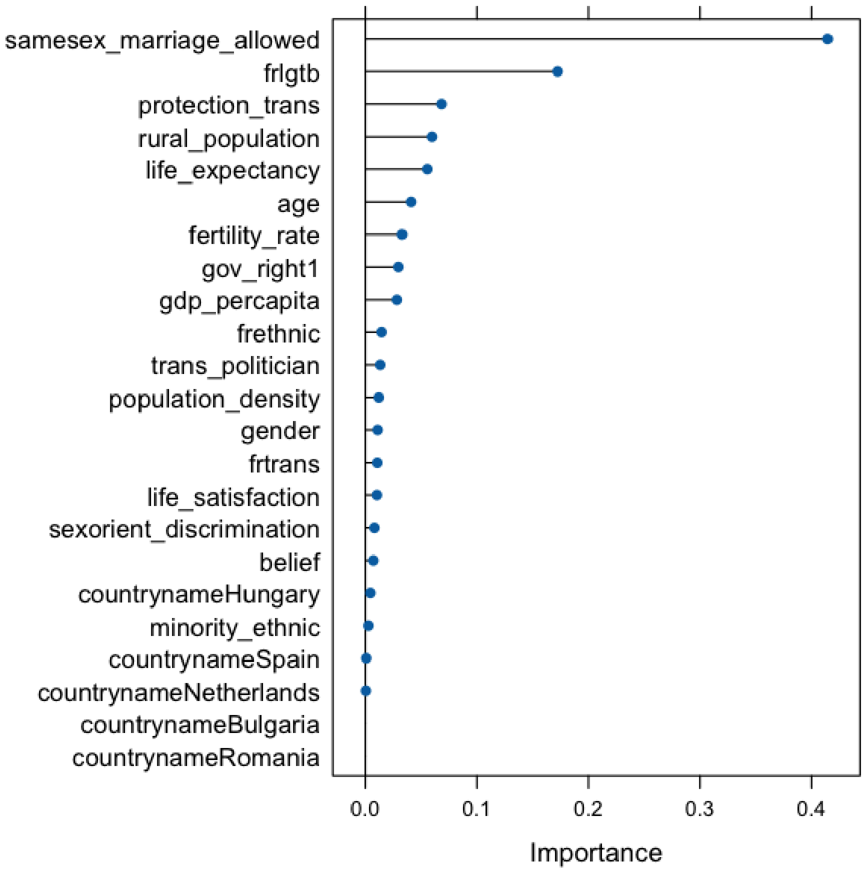

Figure 19. Importance of variables in Gradient Boosting Model