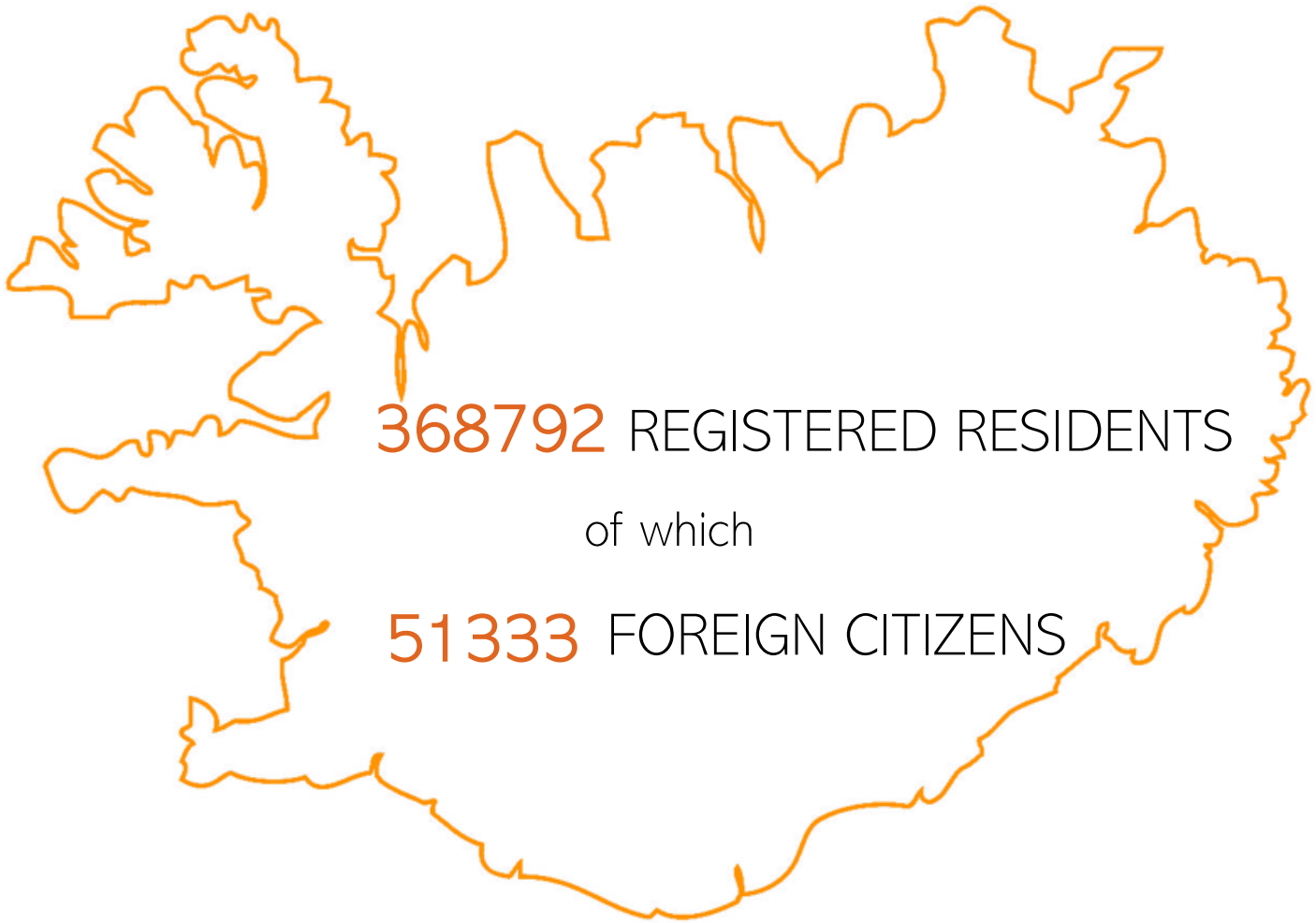
Margherita Zuppardo
Violeta Calian
Ómar Harðarson

MACHINE LEARNING METHODS FOR ESTIMATING THE CENSUS POPULATION

Statistics Iceland

NSM 2022

2021 DIGITAL CENSUS



HOW MANY ARE ACTUALLY HERE?

Registers Iceland

People Change of address → **Moving from Iceland**

Moving from Iceland

Notify a change of address

Attn.

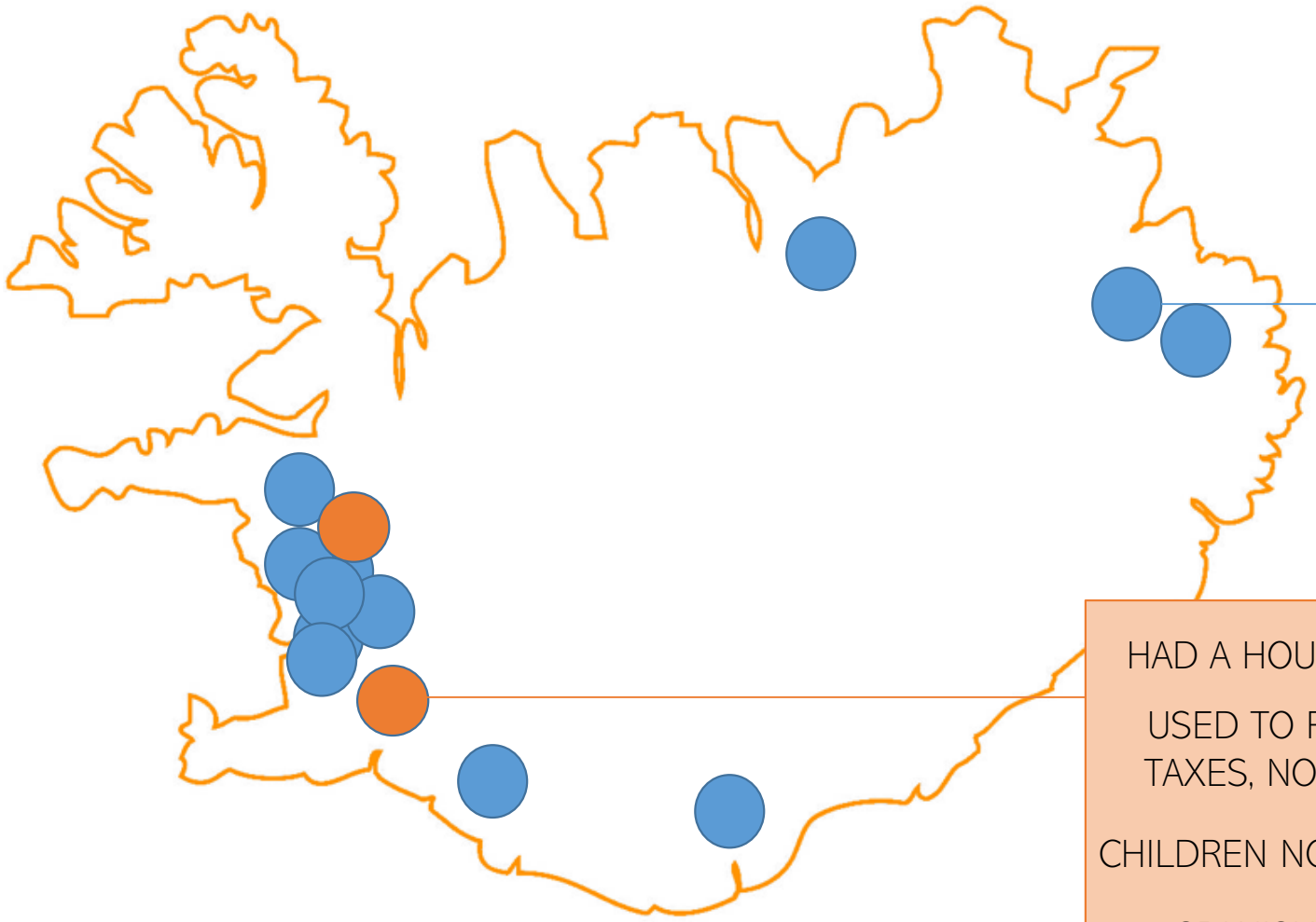
Transfers of legal domicile from Iceland **must be registered within 7 days**.

IN 2020:

7994 DE-REGISTRATIONS

5833 FROM NON-CITIZENS

SIGNS OF LIFE



OWNS A HOUSE

WORKS IN ICELAND

HAS CHILDREN IN SCHOOL

ETC

IN

HAD A HOUSE, SOLD IT

USED TO PAY INCOME TAXES, NOT ANYMORE

CHILDREN NOT IN SCHOOL

SPANISH SPOUSE

ETC

OUT

BINARY CLASSIFICATION!

SOLUTION

- TRAIN MACHINE LEARNING CLASSIFICATION ALGORITHMS
- CHOOSE THE BEST ALGORITHM BASED ON PERFROMANCE MEASURES
- FIND THE OPTIMUM REGIME OF THIS ALGORITHM

GITHUB: https://github.com/MargheritaZ/ML-Census2021
https://github.com/violetacln/SLOPA

STEP 1: BUILDING A DATAFRAME

SOURCES:

LFS survey 2014-2018

NATIONAL REGISTERS

TRAINING TABLE:

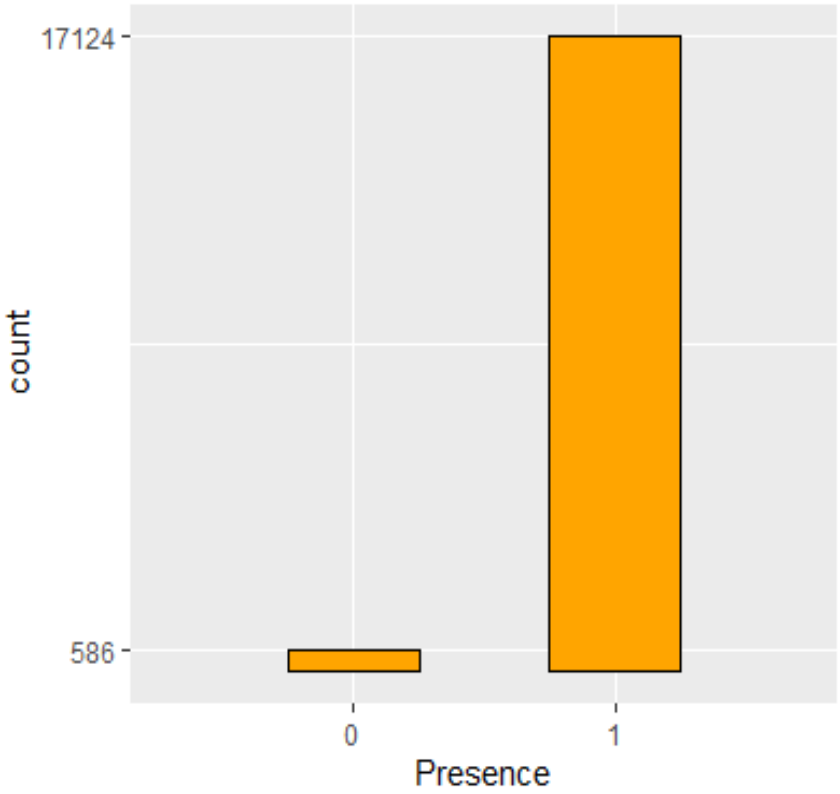
17710 ROWS

21 COLUMNS

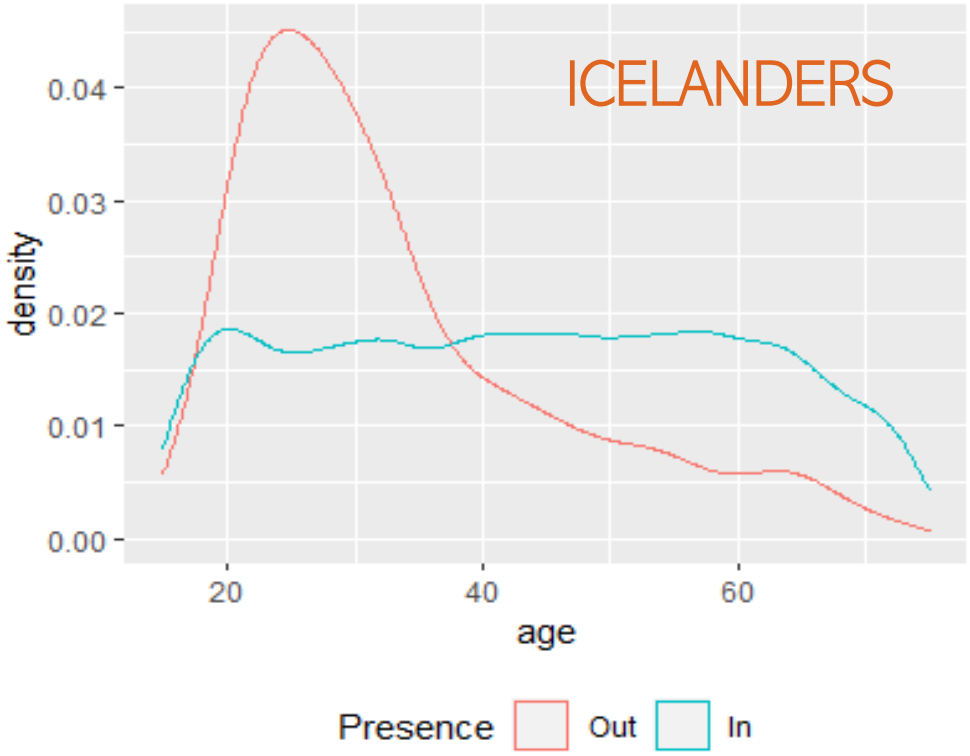
Presence

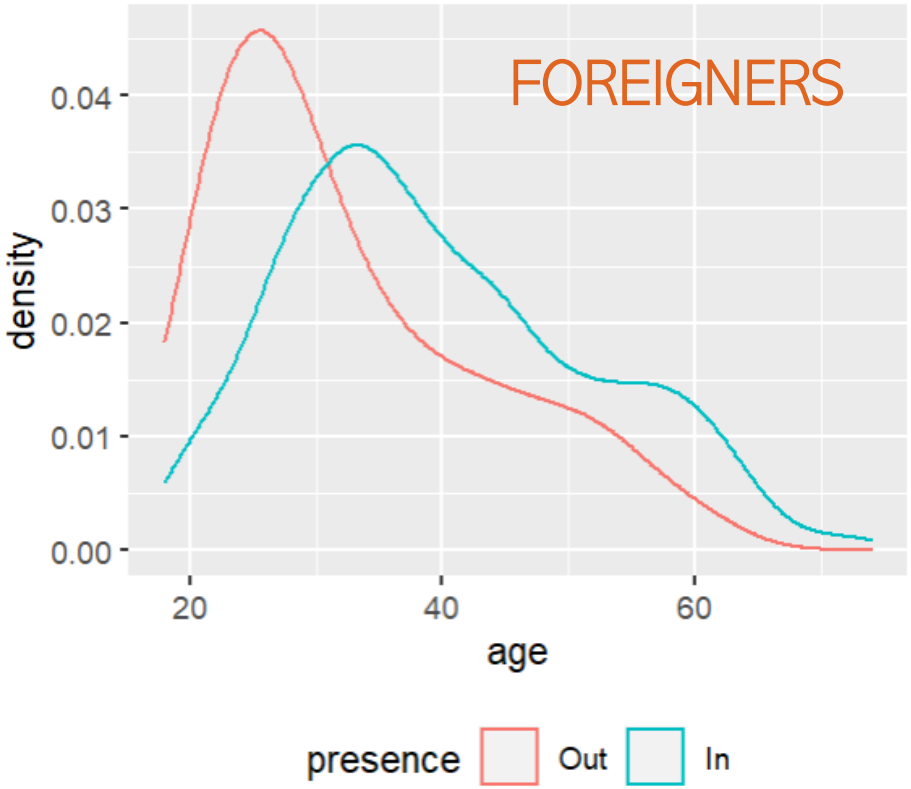
sex
country_of_birth
citizenship
region
ever_abroad
married
children
home_owner
student_abroad

age
income_1yr
income_2yr
months_worked_1yr
months_worked_2yr
year_max_income
year_max_brt
n_skoli
time_in_Iceland
n_movement3yr
n_changes



PREDICTORS


ICELANDERS
density
0.04
0.03
0.02
0.01
0.00
20
40
60
age
Presence
Out
In


FOREIGNERS
density
0.04
0.03
0.02
0.01
0.00
20
40
60
age
presence
Out
In

CHOOSING A MODEL

6 Available Models

The models below are available in `train`. The code behind these protocols can be obtained using the function `getModelInfo` or by going to the github repository.

Show 23 entries

Search:

Model	*method* Value	Type	Libraries	Tuning Parameters
AdaBoost Classification Trees	adaboost	Classification	fastAdaboost	nIter, method
AdaBoost.M1	AdaBoost.M1	Classification	adabag, plyr	mfinal, maxdepth, coeflearn
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptDA	model
Adjacent Categories Probability Model for Ordinal Data	vglmAdjCat	Classification	VGAM	parallel, link
Bagged AdaBoost	AdaBag	Classification	adabag, plyr	mfinal, maxdepth
Bagged FDA using gCV Pruning	bagFDAGCV	Classification	earth	degree
Bagged Flexible Discriminant Analysis	bagFDA	Classification	earth, mda	degree, nprune
Binary Discriminant Analysis	binda	Classification	binda	lambda.freqs
Boosted Classification Trees	ada	Classification	ada, plyr	iter, maxdepth, nu
Boosted Logistic Regression	LogitBoost	Classification	caTools	nIter
C4.5-like Trees	J48	Classification	RWeka	C, M
C5.0	C5.0	Classification	C50, plyr	trials, model, winnow
Cumulative Probability Model for Ordinal Data	vglmCumulative	Classification	VGAM	parallel, link
DeepBoost	deepboost	Classification	deepboost	num_iter, tree_depth, beta, lambda, loss_type
Diagonal Discriminant Analysis	dda	Classification	sparsediscrim	model, shrinkage
Distance Weighted Discrimination with Polynomial Kernel	dwdPoly	Classification	kernlab, kernlab	lambda, qval, degree, scale
Distance Weighted Discrimination with Radial Basis Function Kernel	dwdRadial	Classification	kernlab, kernlab	lambda, qval, sigma
Factor-Based Linear Discriminant Analysis	RFlda	Classification	HiDimDA	q
Flexible Discriminant Analysis	fda	Classification	earth, mda	degree, nprune
Fuzzy Rules Using Chi's Method	FRBCS.CHI	Classification	frbs	num.labels, type.mf
Fuzzy Rules Using Genetic Cooperative-Competitive Learning and Pittsburgh	FH.GBML	Classification	frbs	max.num.rule, popu.size, max.gen
Fuzzy Rules Using the Structural Learning Algorithm on	SLAVE	Classification	frbs	num.labels, max.iter, max.gen

https://topepo.github.io/caret/

CHOOSING A MODEL

	TRUE IN	TRUE OUT
PREDICTED IN	TP	FN
PREDICTED OUT	FP	TN

$$\text{Accuracy} = \frac{TN + TP}{N\_tot}$$

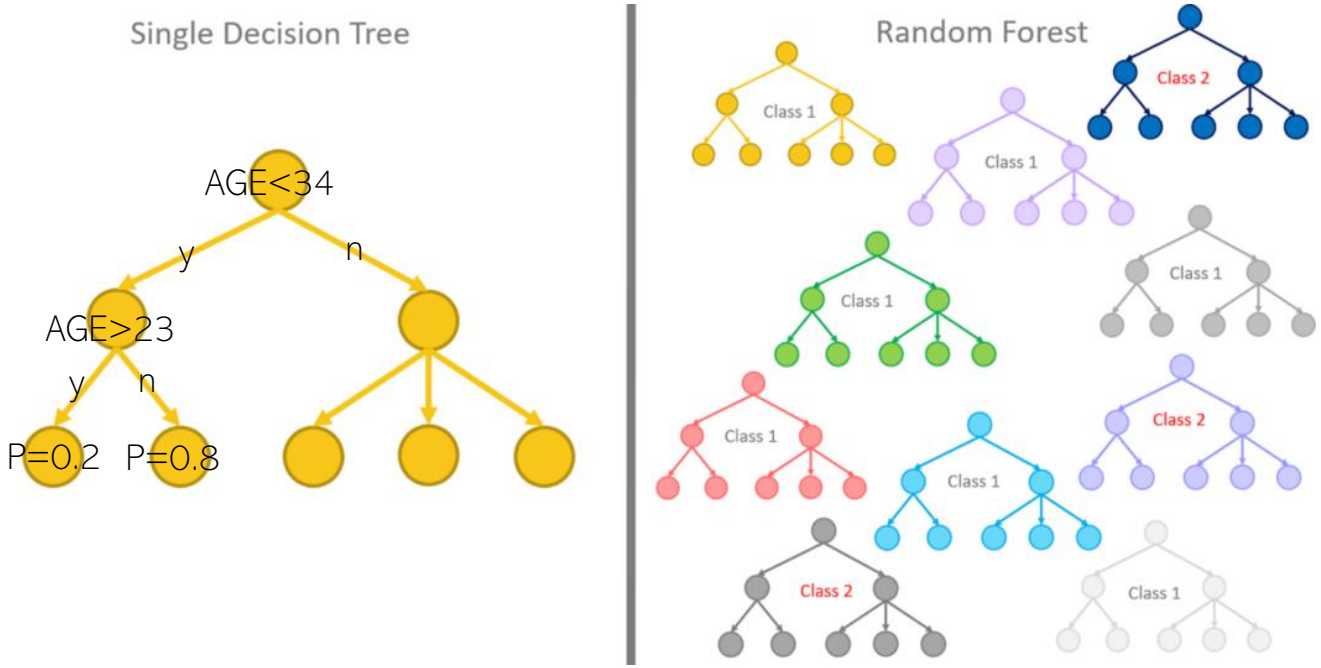
$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Population error} = \frac{|FP - FN|}{N_{tot}}$$

Method	Accuracy (%)	Specificity (%)	Sensitivity (%)	Population error (%)
Register data	<96.7	0.0	100.0	>3.3
Logistic regression	96.8	14.2	99.6	2.6
Decision tree	97.0	16.4	99.8	2.3
Neural network	96.9	17.5	99.6	1.6
AdaBoost	95.1	26.3	97.5	0.0
Random forest (untuned)	97.0	25.1	99.5	2.1
Optimized RF (final model)	96.0	48.0	98.0	0.04
Latest results (revised data)	96.7	56.0	98.2	0.2

RANDOM FOREST

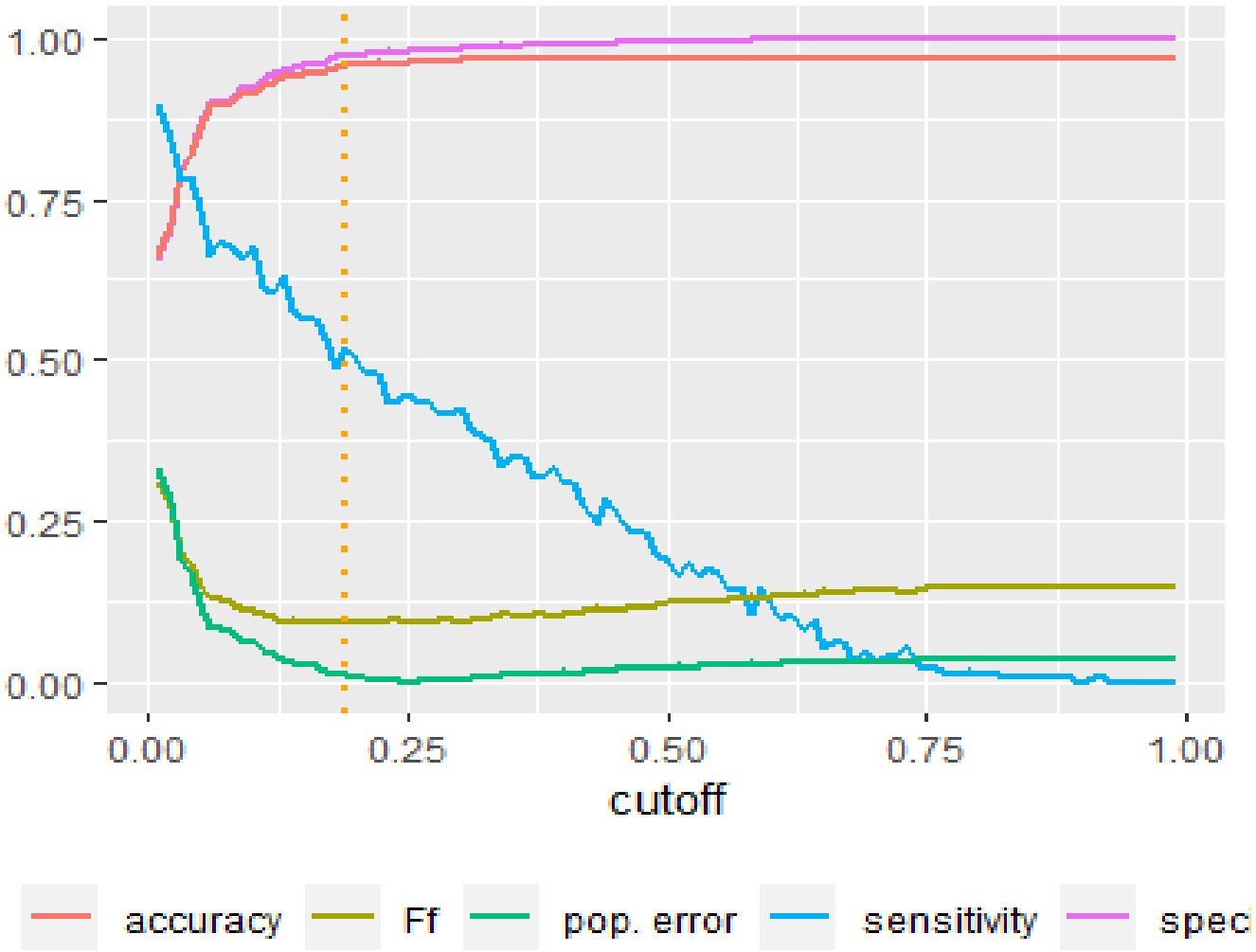


https://rosaria-silipo.medium.com/

OUTPUT: PROBABILITY TO BE INSIDE THE COUNTRY

https://cran.r-project.org/web/packages/randomForest/index.html

MODEL TUNING



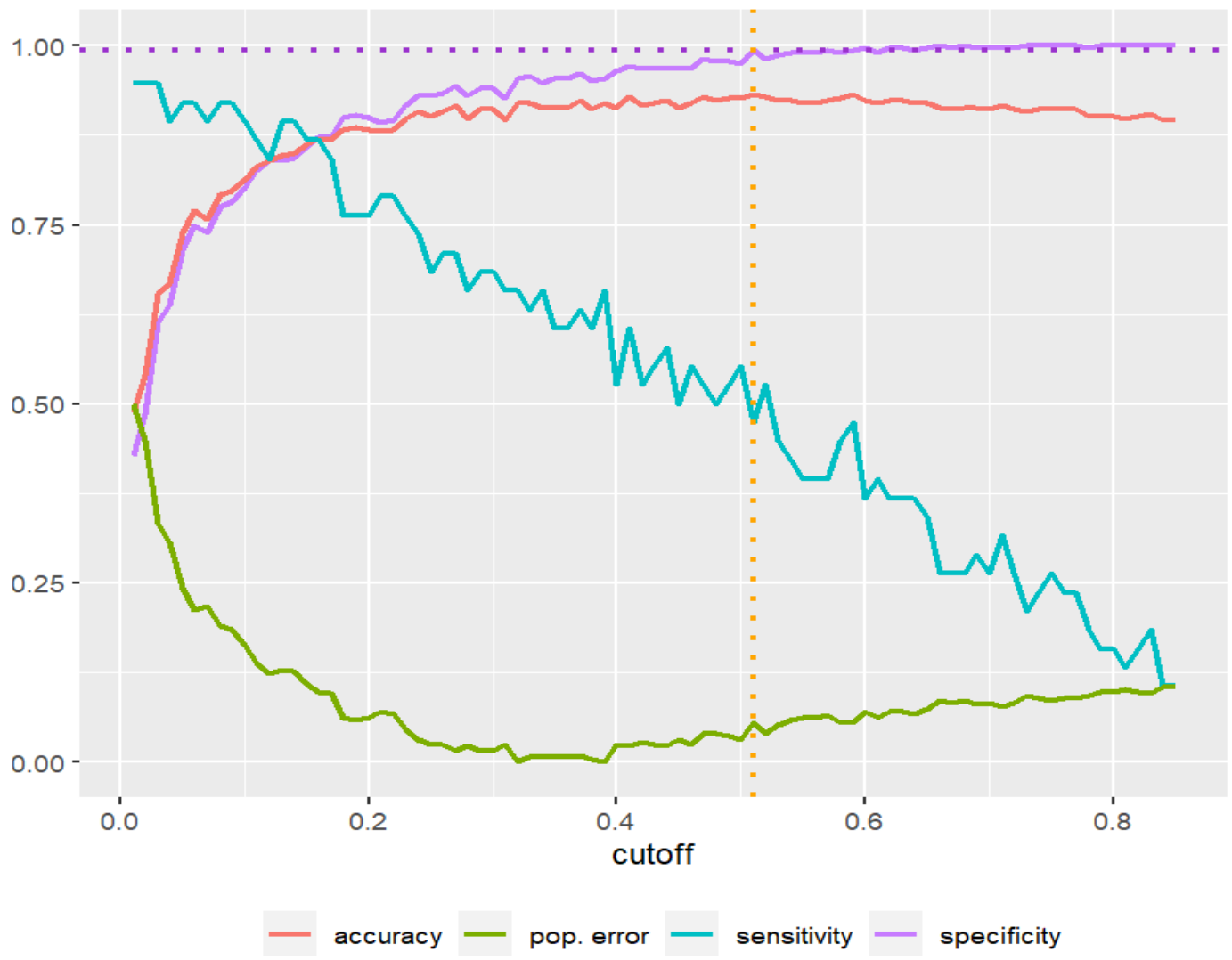
WE TUNE THE PROBABILITY CUTOFF

Specificity > 98%

WE OPTIMIZE EVERYTHING ELSE

96.0	48.0	98.0	0.04
Accuracy	Sensitivity	Specificity	Pop. error

FOREIGN DATA

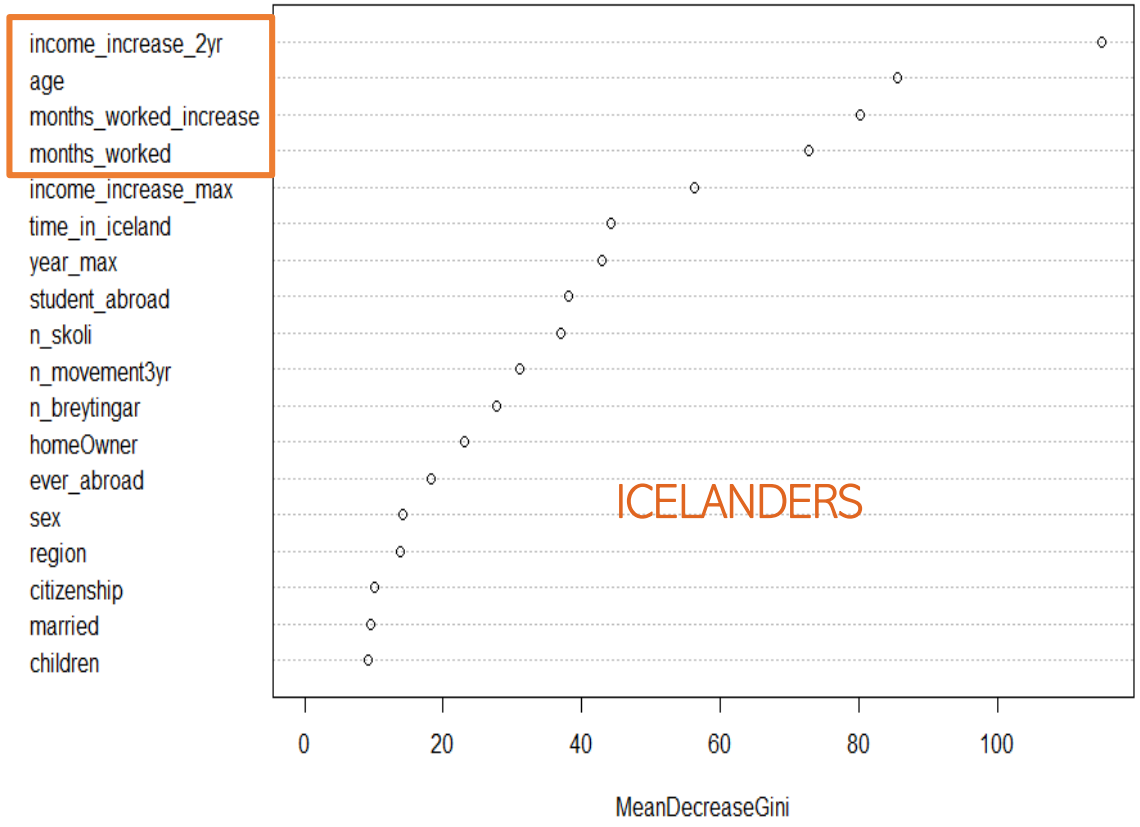


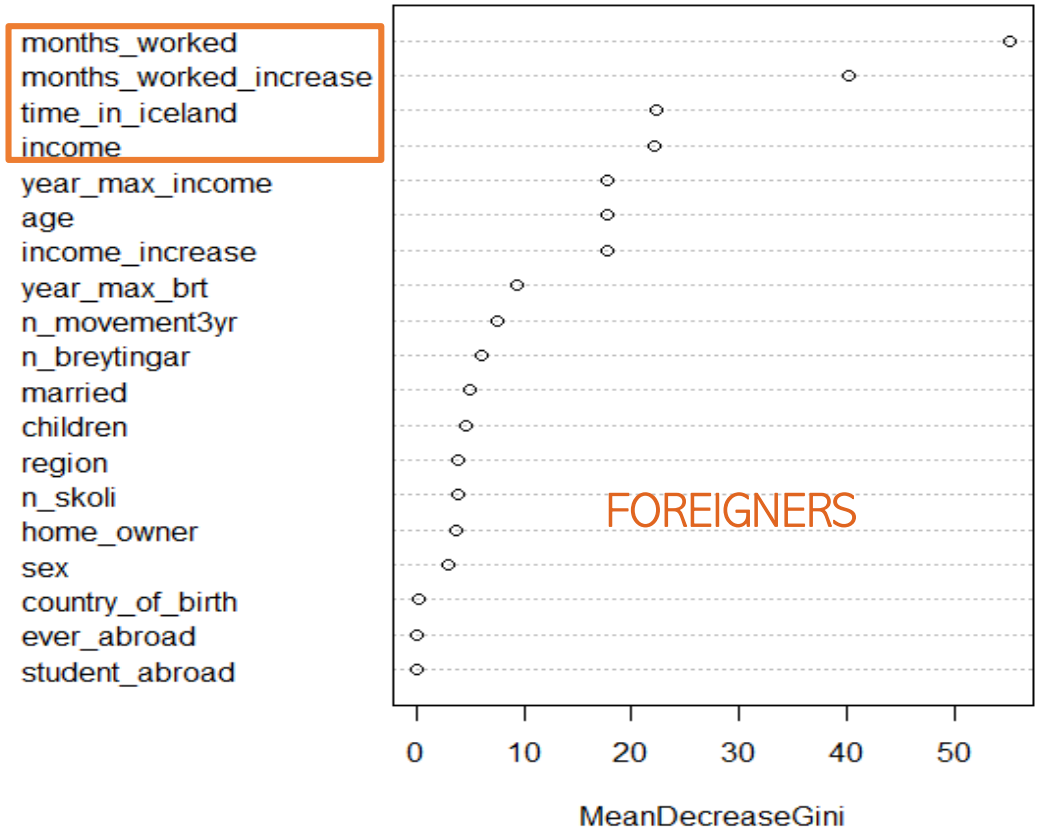
Specificity > 98%

MUCH NOISIER!

92.0	44.6	98.2	0.05
Accuracy	Sensitivity	Specificity	Pop. error

IMPORTANCE OF PREDICTORS





CENSUS RESULTS

368 792 REGISTERED RESIDENTS

7100 PEOPLE OUT

OUT OF WHICH

3770 FOREGNEIRS

	IN	OUT
In school 6-15 of age	45886	808
Not in school 6-15	537	8710
Working in November	181904	2325
Having a car	153608	1915

IN CONCLUSION

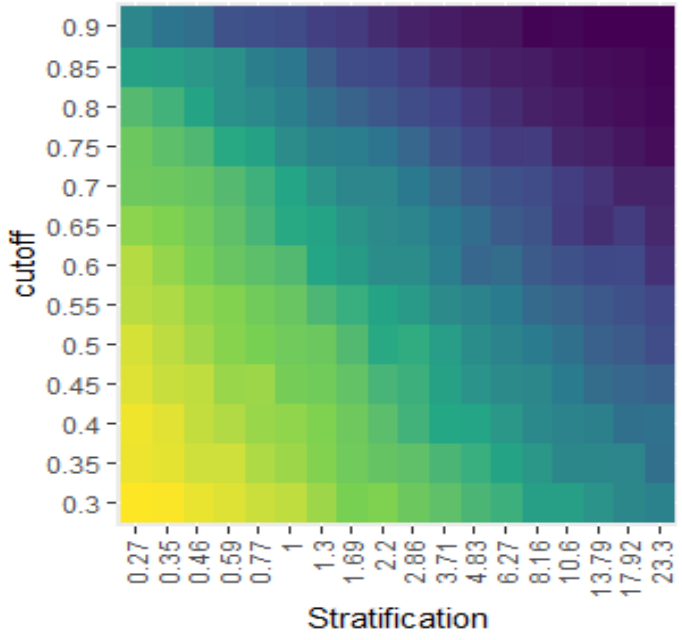
WE TRAINED A RANDOM FOREST MODEL TO PREDICT THE ACTUAL 2021 POPULATION IN ICELAND

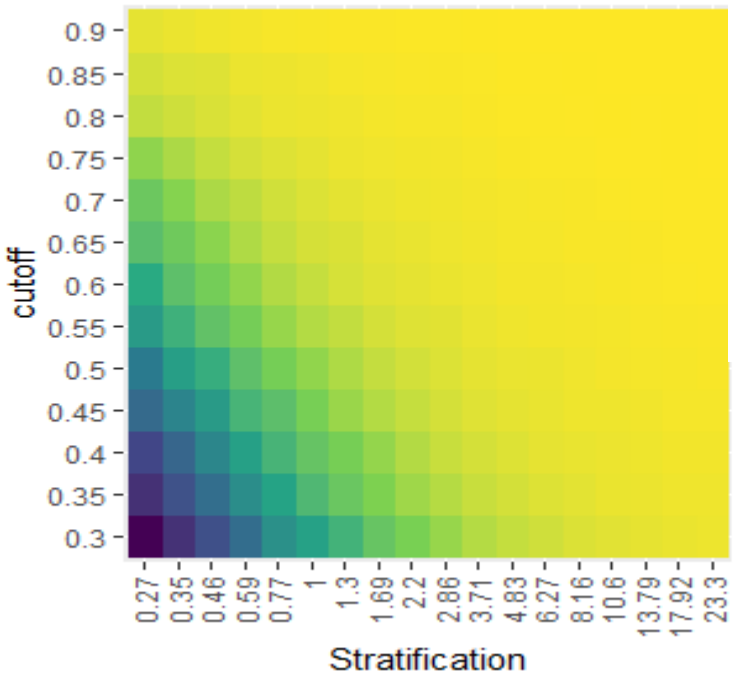
THE MODEL IS ACCURATE BUT 'SAFE' (HIGH SENSITIVITY)

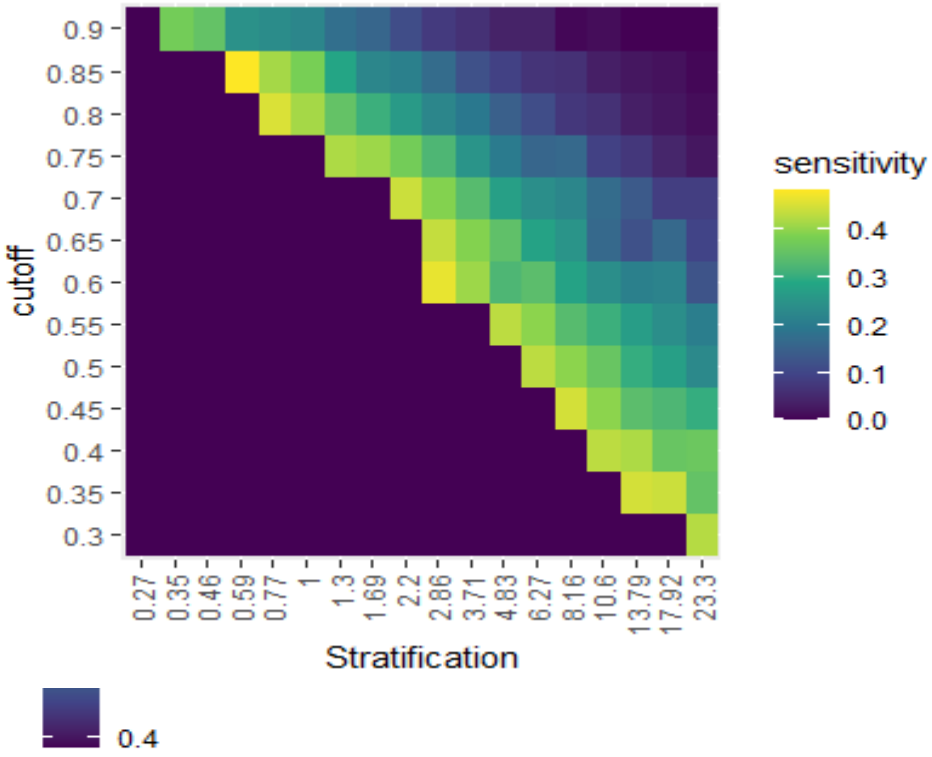
IT CAN BE IMPROVED BY INCREASING THE QUALITY OF THE DATA, AND MAY BE USED REGULARLY IN THE FUTURE TO ESTIMATE THE ICELANDIC POPULATION

Statistics Iceland

APPENDIX: DATA STRATIFICATION







CORRELATIONS

Features	region	sex	time_in_iceland	n_movement3yr	n_skoli	n_breytingar	student_abroad	months_worked_increase	months_worked	income_increase_2yr	year_max	income_increase_max	homeOwner	children	married	ever_abroad	citizenship	age	Presence
region	0.01	0.04	-0.01	-0.09	0.06	0.01	0.06	0	0	0	0.01	-0.03	-0.02	-0.04	-0.02	-0.04	-0.01	-0.01	1
sex	0.01	0.02	0.03	0	0.03	0.09	0.08	0	0	0.08	0.06	-0.01	0	0.02	0.07	-0.01	0.04	1	-0.01
time_in_iceland	0.17	0.1	0.98	-0.45	0.05	0	0.18	-0.1	0.04	0.18	0.11	-0.16	0.03	0.14	0.07	-0.2	1	0.04	-0.01
n_movement3yr	-0.09	-0.25	-0.19	0.2	-0.16	0.12	-0.23	-0.03	-0.04	0.04	-0.03	0.07	0.06	0.23	0.07	1	-0.2	-0.01	-0.04
n_skoli	0.09	-0.36	0.06	0.05	0.08	0.69	0.01	-0.01	-0.06	0.09	0.01	0.01	-0.03	0.14	1	0.07	0.07	0.07	-0.02
n_breytingar	0	-0.26	0.14	0	-0.07	0.11	-0.09	0.01	-0.07	0.11	0.01	0.05	0.05	1	0.14	0.23	0.14	0.02	-0.04
student_abroad	-0.22	-0.1	0.03	0.04	-0.06	-0.01	-0.08	-0.09	0	-0.02	-0.11	-0.03	1	0.05	-0.03	0.06	0.03	0	-0.02
months_worked_increase	0.05	-0.19	-0.13	0.1	-0.11	-0.07	-0.16	0.38	-0.18	0.1	0.2	1	-0.03	0.05	0.01	0.07	-0.16	-0.01	-0.03
months_worked	0.28	0.24	0.1	-0.09	0.15	0.11	0.21	0.37	-0.3	0.71	1	0.2	-0.11	0.01	0.01	-0.03	0.11	0.06	0.01
income_increase_2yr	0.2	0	0.16	-0.08	0.03	0.16	0.1	0.44	-0.37	1	0.71	0.1	-0.02	0.11	0.09	0.04	0.18	0.08	0
year_max	-0.07	0.23	0.04	-0.01	0.12	0.03	0.14	-0.33	1	-0.37	-0.3	-0.18	0	-0.07	-0.06	-0.04	0.04	0	0
income_increase_max	0.12	-0.08	-0.09	0.03	-0.05	-0.05	-0.07	1	-0.33	0.44	0.37	0.38	-0.09	0.01	-0.01	-0.03	-0.1	0	0
homeOwner	0.14	0.55	0.18	-0.12	0.51	0.21	1	-0.07	0.14	0.1	0.21	-0.16	-0.08	-0.09	0.01	-0.23	0.18	0.08	0.06
children	0.06	-0.14	-0.01	0.1	0.26	1	0.21	-0.05	0.03	0.16	0.11	-0.07	-0.01	0.11	0.69	0.12	0	0.09	0.01
married	0.08	0.47	0.04	-0.02	1	0.26	0.51	-0.05	0.12	0.03	0.15	-0.11	-0.06	-0.07	0.08	-0.16	0.05	0.03	0.06
ever_abroad	-0.13	-0.13	-0.46	1	-0.02	0.1	-0.12	0.03	-0.01	-0.08	-0.09	0.1	0.04	0	0.05	0.2	-0.45	0	-0.09
citizenship	0.15	0.1	1	-0.46	0.04	-0.01	0.18	-0.09	0.04	0.16	0.1	-0.13	0.03	0.14	0.06	-0.19	0.98	0.03	-0.01
age	0.11	1	0.1	-0.13	0.47	-0.14	0.55	-0.08	0.23	0	0.24	-0.19	-0.1	-0.26	-0.36	-0.25	0.1	0.02	0.04
Presence	1	0.11	0.15	-0.13	0.08	0.06	0.14	0.12	-0.07	0.2	0.28	0.05	-0.22	0	0.09	-0.09	0.17	0.01	0.01

Features

Correlation Meter

-1.0 -0.5 0.0 0.5 1.0

GINI INDEX

$$\text{Gini Index} = 1 - \sum (P(x=k))^2$$