NSM 2022
NORDIC STATISTICAL MEETING

Machine learning methods for estimating the Census population

Margherita Zuppardo, Statistics Iceland, margherita.zuppardo@hagstofa.is
Violeta Calian, Statistics Iceland, violeta.calian@hagstofa.is
Ómar Harðarson, Statistics Iceland, omar.hardarson@hagstofa.is

Abstract

The dramatic development of machine learning and A.I. in the last decade opens up many new possibilities of improvement for the 2021 register-based Icelandic Census. We focus here on one of the main purposes of the census, which is to accurately describe the population of residents of Iceland. However, identifying the individuals belonging to this population is not trivial since many people do not notify national registers about their change of residence when moving abroad. Ignoring this phenomenon may create biases in statistical estimates of demographic or social characteristics e.g. age distributions, fertility or mortality rates, migration flows, employment and education profiles.

This is a binary classification problem where the status of any individual may be either in or out of the country. In this paper, we propose a systematic solution based on rigorous statistical methods, implemented as machine learning algorithms by using open source R-packages. To our knowledge, such techniques were not previously applied to demographic problems of this type. The data set used for training and testing the algorithms was built by using information regarding presence/absence of individuals from surveys combined with register data regarding for instance employment status, income and taxes, education level, changes in civil and residency status, family composition, previous migration events („signs of life“).

We trained several classification models such as random forests, classification trees, neural networks as well as their stacked versions. We assessed their performance according to measures which include sensitivity, specificity, confusion matrices, accuracy and information rates and their confidence intervals. We discuss the results obtained by applying these methods to the Census data.

Keywords: Census 2021, machine learning, signs of life, open source code

1. Introduction

The population overestimation by register-based population statistics and/or census is a well-known problem of official statistics, and it is due to the fact that not all people

leaving the country would promptly de-register. This over-coverage effect can create unwanted consequences for the statistical reporting of demographic, social and economic characteristics.

Statistics Estonia has previously developed, for this purpose, an approach based on a residency index [2] built as a function of binary variables describing education, health care, social support, employment ("signs of life") measures and calibrated on training data of certain outcomes. Statistics Sweden used a scoring method [3], based on tracing changes in registers concerning characteristics related to education, income, migration, civil status or residency and analysing the impact of the over-coverage on mortality and fertility estimates. For the 2011 edition of the Census, Statistics Iceland solved this problem with a logistic regression type of model, applied to foreign citizens residing in the country.

We have now formulated this problem as a standard statistical classification one, that will be solved and employed in the Icelandic census 2021. We provided a solution by investigating a spectrum of machine learning algorithms [1], including ensemble/stacked classifiers. All these models were fitted on survey data, enriched with financial, social, household attributes observed at previous points in time. In this paper, we show some important steps to the process of solution based on classification algorithms:

(i) choosing the best performing model according to well defined metrics (e.g. accuracy, sensitivity, specificity), and adding domain-specific constraints, such as accepting higher numbers of false-presence than of false-absence due to the difference in likelihoods of these two states in the real population;

(ii) optimising the algorithm parameters. In particular, the value of the cut-off classification probability and of the under-sampling/stratification proportion needed. This is useful due to the very unbalanced sizes of the present/absent classes in our sample, and greatly improves the results according to our metrics;

(iii) providing some type of interpretability to the machine learning models, notoriously complex and not always transparent.

2. Training and testing data: sources and analysis

The data set used for training and testing the machine learning (ML) models has the Labour Force Survey (LFS) survey over the 2014-2018 period as a main source of information. The sample includes over 17000 individuals aged 16 and over. During the survey, participants are asked whether they still reside in the country. We use their answers, together with de-registrations dates to define the binary variable 'Presence'. In addition, register data concerning demography, income, employment, and real estate ownership were involved.

The variables in the final data set are the following:

- *binary variables*:

gender, region (register address in/not in the capital region), Icelandic citizenship;

individual has ever had foreign residence, has dependent children, is a home owner, has been studying abroad in the past year

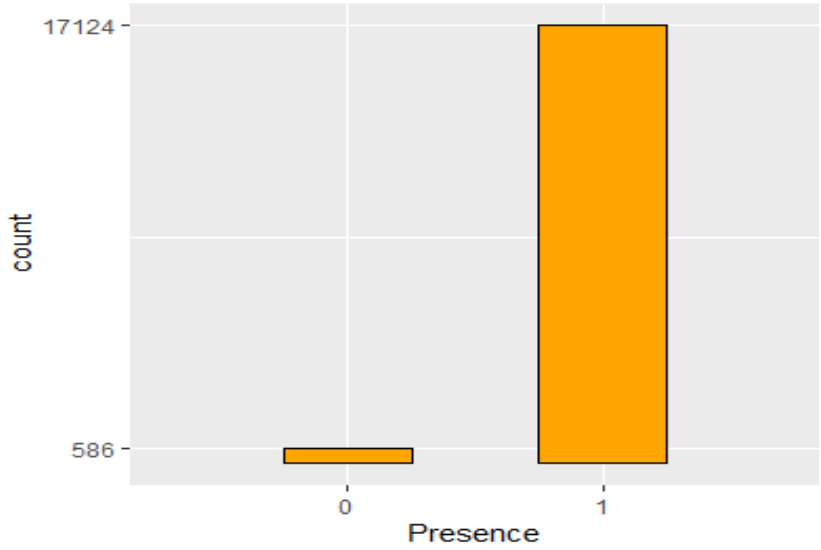
- *numeric variables, measuring:*

  age, the difference between present income and the highest past income (scaled by average income), the income increase in the past two years relative to the previous two years (also scaled), the number of months the person worked during the past year, the number of changes registered in public registers in the past 12 months, the number of adults in the household who have been in school in the country during the past 12 months, the number of children in the household who have been in school in the country during the past 12 months, the number of recorded changes of address in the past 3 years, the ratio between the time spent in the country and age (it is one for Icelandic citizens and between zero and one for foreign citizens), the number of years since earning the highest income.

All the individuals contacted in the LFS survey were in the country according to the population register. For this reason, only a small fraction was found to be residing outside of Iceland. Therefore our data sample has a disproportionately large number of individuals with present versus absent confirmed status (Figure 1). This indicates that

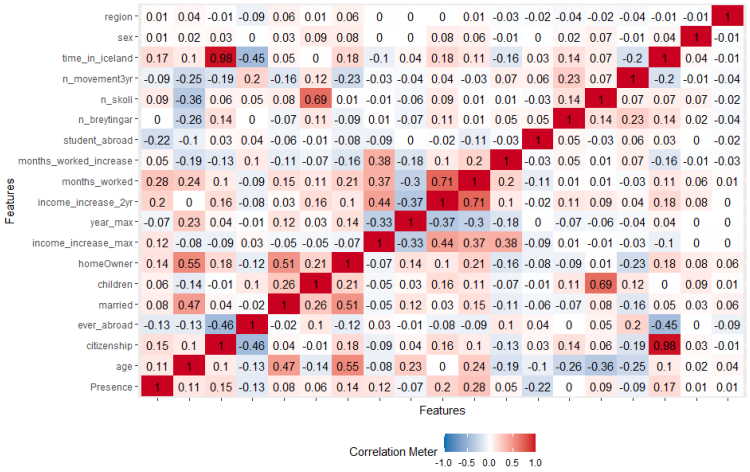
we are solving an imbalanced classification problem and need to apply a stratified random under-sampling for the training data, as will be discussed later.

Figure 1. The survey results on presence versus absence status



As part of the exploratory data analysis, checking the pairwise correlations between variables is a useful step (see Figure 2). This investigation shows that the present/absent status is correlated (anti-correlated) with the variables reflecting changes of location, changes of income, owning a house, several demographic characteristics and length of time lived in Iceland, thus these variables may be tested as predictors in the set of proposed models.

Figure 2. Correlations between the variables used for modelling



The preliminary data analysis confirmed the fact that the proposed predictors have a clear influence on the outcome class. We include here a few plots illustrating the contrast between distributions of some variables of interest on the two groups defined by presence/absence status.

They show that: most people absent from the country (although registered) are young, in the range 19 to 38 years of age (Figure 3). Home owners are more likely to be present in the country (Figure 4), while people not owning a house are more likely to be abroad. People with a recent increase in income are more likely to be present than absent (Figure 5). Also, the more time a foreign individual has spent in Iceland, the more likely he/she is to be currently present in the country (Figure 6). This confirms that short term migrants are de-registering less frequently than migrants who live in Iceland for a longer time or than Icelandic citizens.

Figure 3. The age distributions of present and absent individuals. The latter shows a clear peak around the age of 25.

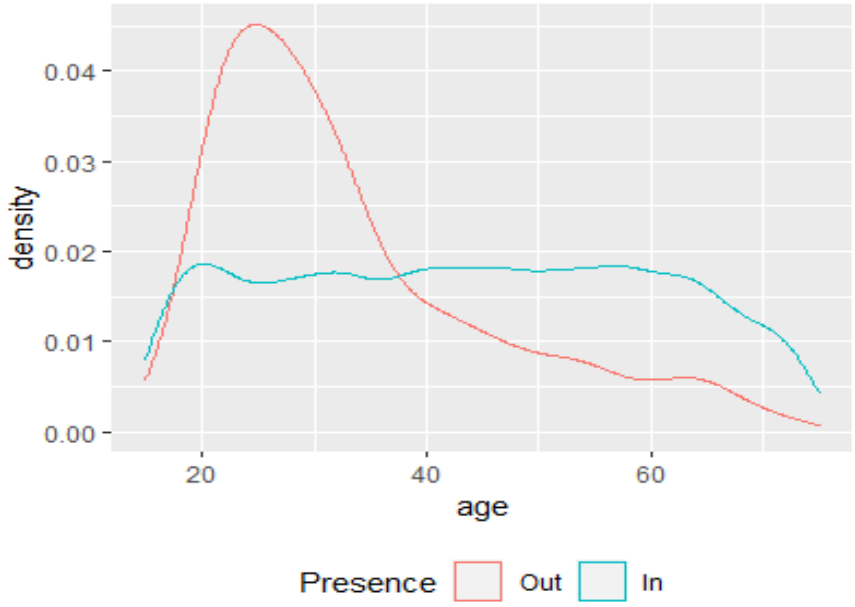


Figure 4. The number of people who are/are not home owners while present/absent from the country

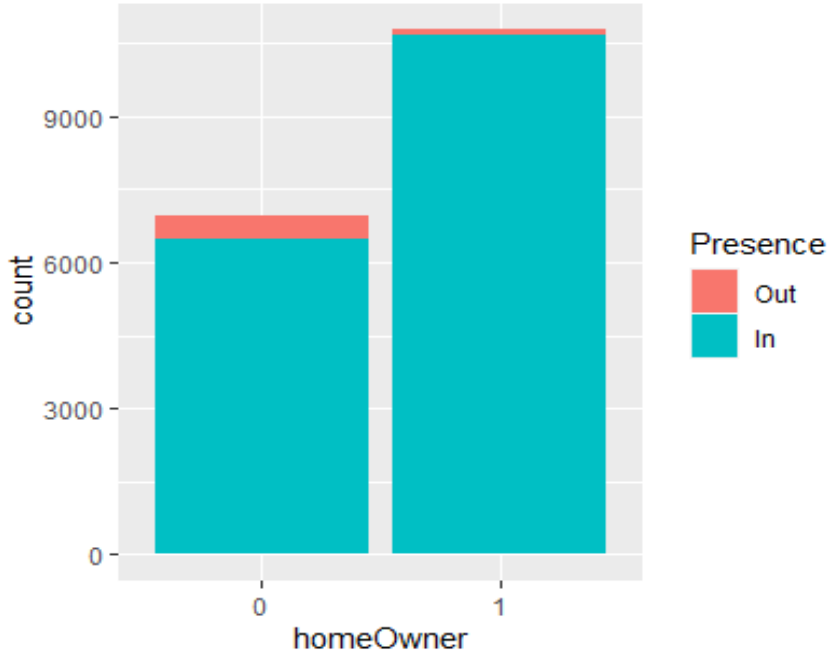


Figure 5. Comparing the distributions of the relative increase in income for present/absent individuals

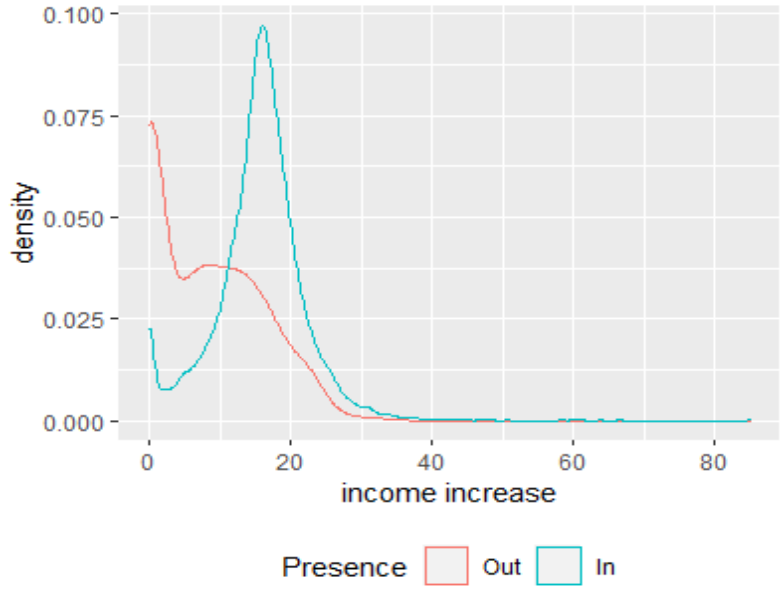
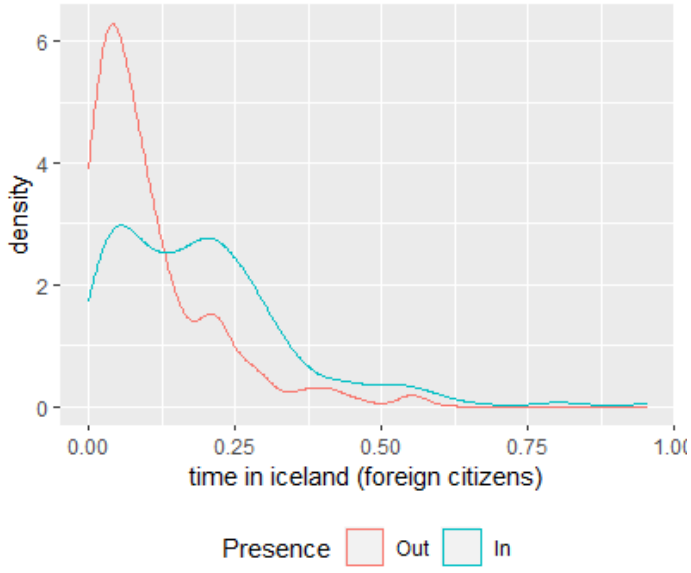


Figure 6. Density distribution of the (normalized by age) time spent in Iceland by foreign citizens for present/absent individuals



3. Results: choosing and optimizing the classification algorithm

In this section we describe the main results which concern the method of choosing the best performing ML algorithm and how to find its optimum cutoff (i.e. the value of classification probability) and stratification values.

The evaluation of classification algorithms is usually based on several performance measures:

(i) The confusion matrix (CM), with four components defined by: the number of cases correctly classified (in or out of the country), on the diagonal of the matrix but also by the number of cases incorrectly classified (as out while in or the other way around) on the anti-diagonal.

	actually out	actually in
predicted out	a	b
predicted in	c	d

(ii) Accuracy (X): the proportion of cases correctly classified, out of the total number of cases

(iii) Specificity (Y): the proportion of cases classified as “out of the country”, from the total number of cases which are indeed out of the country

(iv) Sensitivity (Z): the proportion of cases classified as “out the country”, from the total number of cases which are in fact *in* the country

The search for the best performing algorithm should thus find an optimum over the (X,Y,Z)-space. However, the purpose of classification may define priorities. For instance, in the case of census estimation, we prefer high sensitivity over specificity since it is better to overestimate the population than to underestimate it.

In addition, we impose the condition that the difference between wrongly classified cases should be very small when compared to the total number of cases, (b-c)|/N, since this is a measure of the *relative error of census population estimate*.

We split our data set into training and testing subsets, and fitted and evaluated several models, comparing their performance as shown in Table 1. The table includes the classification according to the register data as a baseline.

We chose the random forest classifier [4] for its best overall performance, potential for improvement and convenience. We tuned it further, by searching for the best combination of internal parameters (probability cutoff, under-sampling stratification) values which give the “optimised RF” classifier. Most recently, data has been slightly revised by improving the quality of the income-variation variables and by re-running the analysis we obtained an even better performance as indicated by the last line in Table 1.

Table 1. Comparing performances of main tested algorithms

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Total population error (%)
Register data	96.7	0.0	100.0	16.2
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

The search in the (X,Y,Z) – space and the tuning of the best algorithm are illustrated in what follows.

Figure 7 shows how the algorithm performance changes with the cutoff probability value. This value determines how cases are classified, depending on the probability of belonging to absent/present class. As it can be seen in the figure, the specificity and accuracy of the prediction stay relatively high for a wide range of the cutoff parameter, while the sensitivity and population error are more responsive to the tuning. For this reason, we chose the cutoff corresponding to the highest sensitivity value

which allows us to keep the specificity above 98% level, as indicated by the dotted orange line. This, in addition, corresponds to a low relative population error as well.

To further improve the specificity, we tuned the ‘sampsize’ parameter [4], which allows resampling of the training data with different proportion of people ‘in’ and ‘out’ of the country. This resulted in a small but significant performance improvement.

Figures 8 and 9, illustrate the tuning of the algorithm along the sensitivity/specificity dimensions, as a function of the cutoff and stratification parameters. As the 1-dimensional tuning shows in Figure 7, the sensitivity is maximized while keeping a 98% level for specificity. This is illustrated in Figure 10, where the uniformly colored region corresponds to parameter values that are not allowed due to low specificity values.

Based on these results, we chose the range 0.3-0.4 for the cutoff and 17-19 for the strata proportion. Further tuning can be done within these bounds.

Figure 7. Main performance metrics and a linear combination of them (Ff) as functions of the probability cutoff value of the classifier.

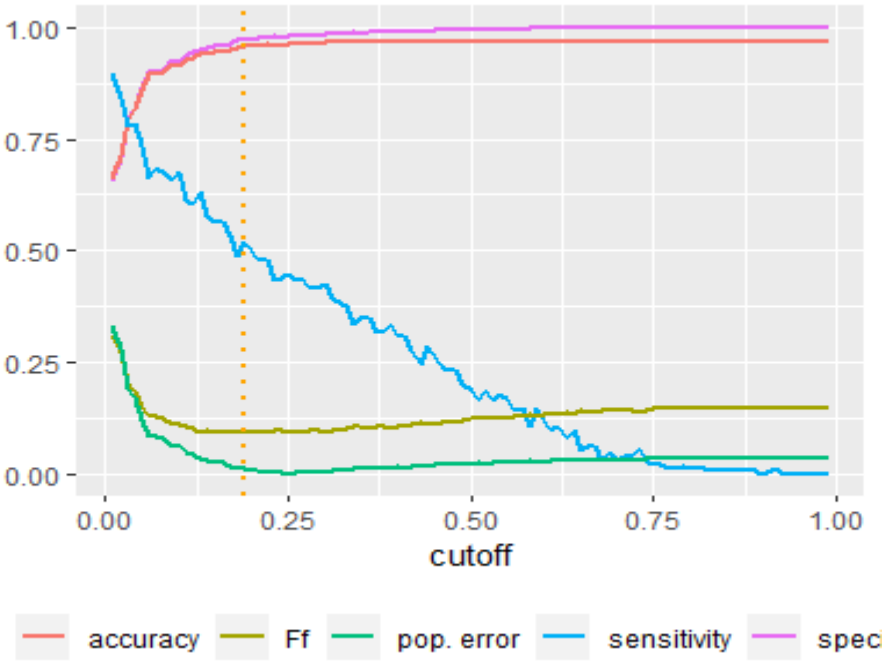


Figure 8. Sensitivity as a function of the probability cutoff value and stratification proportion

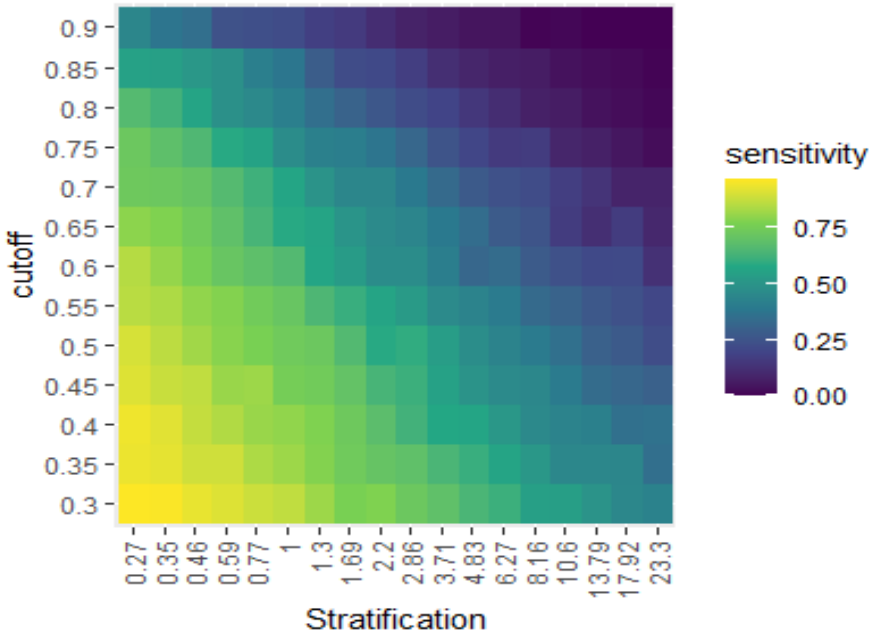


Figure 9. Specificity as a function of the probability cutoff value and stratification proportion

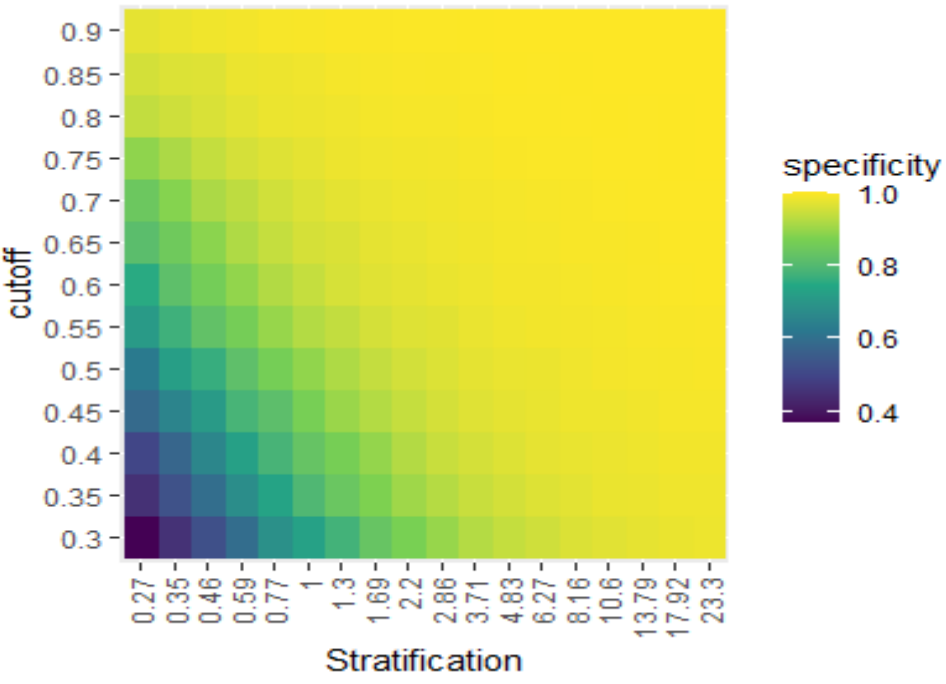
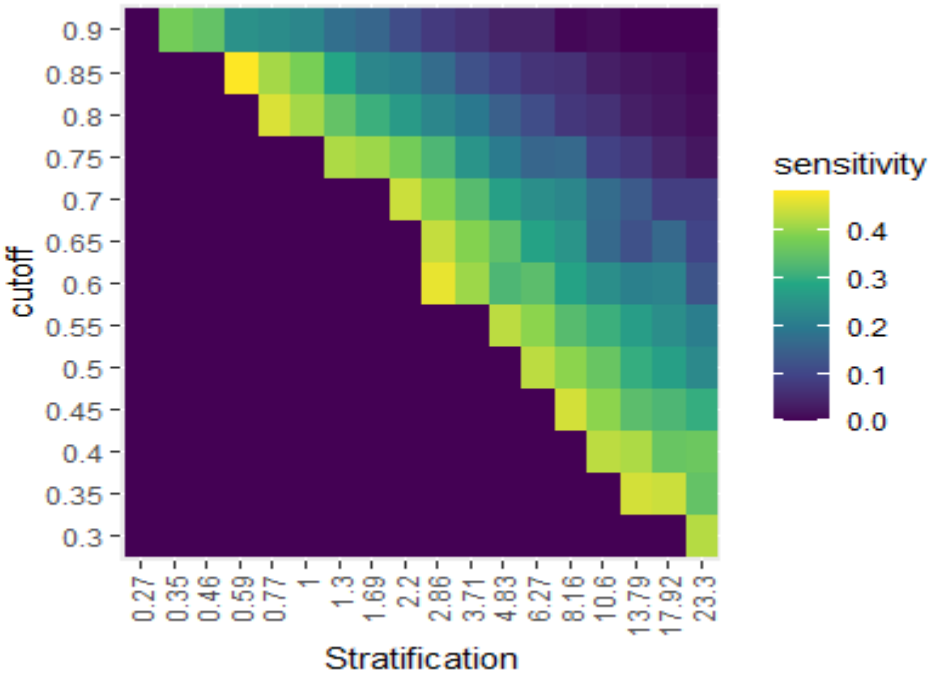


Figure 10. Sensitivity as a function of the cutoff and stratification values, *given that specificity is above 98%.* This shows for instance that the optimal values are around cutoff=0.35 and stratification=13.79.



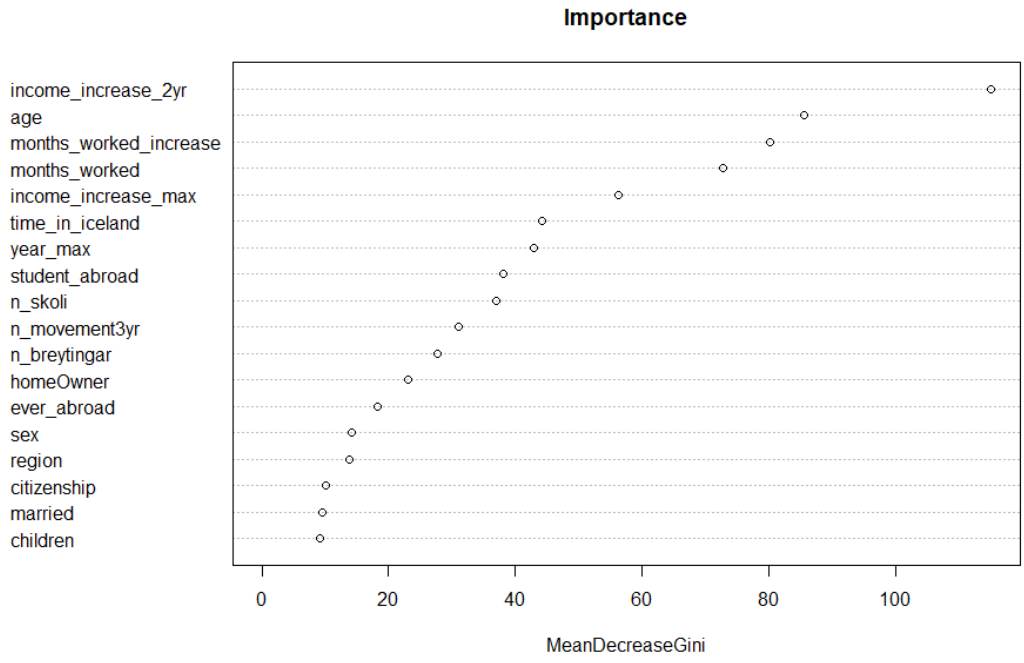
4. Interpretability of the machine learning algorithm

A frequent criticism of ML methods for official statistics is the lack of transparency of the ML models. In recent years however, new methods have been developed for conveying easy interpretations to the users of ML generated results.

We enumerate here only a few:

(i) the feature importance. This is a score which measures the increase in model's prediction error after permuting a given feature. We illustrate this for our random forest algorithm in Figure 11.

Figure 11. Feature importance calculated by using the *iml* R-package and showing how much each variable affects the predictions.



(ii) feature effects and interactions. They measure how the predictions change locally when a feature is varied and how much of the variance in outcome is explained by interactions between features.

(iii) surrogate decision trees can be built, connecting the input data and predictions, as an attempt to identify simple rules that classify cases and using an adjustable number of splitting levels.

5. Discussion

We showed an example of how machine learning can be used for population statistics. By using survey and register data, we fitted a class of Random Forest models and used them for the 2021 Census population data. These models allow to predict the presence/absence status of any individual in the country, improving the population estimate given by registries. In particular, as shown in Table 1, by misclassifying only about 2% of the people that are in the country (see the 98% specificity of the final model), we were able to identify more than 50% of the individuals that left the country

without de-registering. This process is accompanied by a very small error in estimating the total population (0.2%).

The methods and models[1] proposed in this paper may be applied to any data set with similar type of variables for the goal of training classifiers and predicting presence/absence (or any binary) status. The optimisation procedure described here gives good results for our problem, but, as well as the estimation of the model uncertainty, can be further improved. The latter is part of a wider topic still under development and the object of a new research project, mainly driven by simulations, but interesting for both theoretical and applied purposes.

6. References

[1] V. Calian, M. Zuppardo, Corecting for population overestimates by using statistical classification methods, *NTTS (2021),* https://coms.events/NTTS2021/data/abstracts/en/abstract_0088.html

[2] E. Maasing, E.-M. Tiit, M. Vahi, Residency index – a tool for measuring the population size, Acta et Commentationes Universitatis Tartuensis De Mathematica (2017), 129-139.

[3] A. Monti, S. Drefahl, E. Mussino, J. Härkönen, Over-coverage in population registers leads to bias in demographic estimates, Population Studies (2019), 1-19.

[4] Leo Breiman, Adele Cutler, Andy Liaw and Matthew Wiener. Breiman and Cutler's Random Forests for Classification and Regression. URL https://cran.r-project.org/web/packages/randomForest/index.html

[5] R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[6] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software (2008) 1-26.

[1] See the shared R-code at https://github.com/MargheritaZ/ML-Census2021

[7] B. Venables, B. Ripley, VR: Bundle of MASS, class, nnet, spatial. R package version 7.2-42 (1999) http://CRAN.R-project.org/package=VR