**STANDARDFORM OUTPUT OR EXPORT[[1]](#footnote-1)**

|  |  |
| --- | --- |
| Name *[name user]* | Hanzhang Ren |
| Telephone number *[telephone number where the user can be reached when there are questions about this output]* | +1 203 285 7184 |
| Institution *[name institution]* | Rijksuniversiteit Groningen |
| Date *[date on which the output is produced]* | 22-09-2024 |
| Project *[contractnumber + name research]* | 9469 - Predicting Fertility Data Challenge (PreFer) |
| This output consists entirely of documentation *[syntax or text file without research results] If Yes, then you don’t need to fill in the rest of the form* | No |
| This output is an adjusted version of a stopped output. If Yes, then please fill in the date of the stopped output. | No |
| Data *[used datasets]* | We used the catboost package to write a machine learning algorithm that predicts whether Dutch individuals between 18 and 45 would have a child between 2021 and 2023, using data up until 2020. This outcome that we predict is a variable constructed by Lisa Sivak using data from GBAPERSOONTAB and KINDOUDERTAB.  We examined all features in GBAPERSOONTAB and GBAHUISHOUDENSBUS as potential predictors.  In addition, we examined a set of potential predictors curated by Lisa Sivak. These predictors come from GBAPERSOONTAB, KINDOUDERTAB, HOOGSTEOPLTAB, SECMBUS, SPOLISBUS, GBABURGERLIJKESTAATBUS, GBAVERBINTENISPARTNERBUS, GBABURGERLIJKESTAATBUS, INPATAB, INHATAB, VEHTAB, GBAADRESOBJECTBUS,  NABIJHEIDKINDOPVTAB, VBOWONINGTYPETAB, and  GBAHUISHOUDENSBUS |
| Purpose *[description of the purpose of the analysis in this output]* | We want to answer the following two sets of questions:   1. If someone trains a default catboost model on increasingly larger random samples of the training data, to what extent do the predictions become more accurate? How does this relationship change depending on the number of features used to train the model? 2. If one uses a package like catboost to train a reasonable machine learning model to predict fertility, would the accuracy of the predictions differ across social groups? Do these group differences depend the number of features used to train the model? |
| Relation with previous output *[description of the relation with previously submitted output]* | We have previously only submitted code that helped produce the output. |
| Content of the output  *[description of the output, for example: there will be frequencies or crosstabs or regression analysis[[2]](#footnote-2) in this output]* | The csv file sample\_size\_investigation.csv seeks to answer the first question mentioned above.  Each row of the csv file corresponds to a default catboost model trained with a particular bundle of features, a particular training set size, as well as a particular evaluation set.  There are 4 options for the bundle of features, as indicated by the “feature\_set” column  1. Just the person’s sex and birthyear  2. All predictors from GBAPERSOONTAB  3. All predictors from GBAPERSOONTAB and GBAHUISHOUDENSBUS. We call this bundle “key\_records”  4. All of the above plus any features curated by Lisa Sivak. We call this bundle “augmented\_records.”  There are 8 options for the size of the training set, as indicatd by the columns training\_set and n\_training\_set:   1. 100 2. 1000 3. 5000 4. 10000 5. 50000 6. 100000 7. 1000000 8. 3696041. Maximum number of people in the training data prepared by Lisa Sivak. It’s a random sample of households for whom the fertility outcome is available, and the analysis deals directly with household members who are Dutch residents and within the target age range of 18-45   There are 2 options for the evaluation\_set.  1. evaluation\_selection\_50\_percent\_split  These 206549 people are half of the evaluation set prepared by Lisa Sivak. This larger evaluation set is a random sample similar to the training set. When evaluating the model in this half of the evaluation\_set, we try different predicted probability thresholds that separate people who had kids and those who did not (e.g. we might classify a person as having kids if the predicted probability of having kids is above 30 percent). We choose the threshold leading to the best predictions. We record performance metrics for that best threshold. The threshold is also saved as one column in the results table.   1. evaluation\_test\_50\_percent\_split   These 206549 people are the other half of the evaluation data. We use the same threshold that we determined using evaluation\_selection\_50\_percent\_split to make sure that we are not artificially inflating model performance by trying many thresholds.  For each row in the csv file, we estimate prediction metrics including Log Loss, Mean Squared Error (MSE), the R2 Score, and the F1 Score. We calculate 95 percent confidence intervals for these measures (this corresponds to columns starting with ci\_lower and ci\_upper). These all represent the model’s average performance among a large group of people and reveal no microdata.  The csv file subgroup\_investigation.csv seeks to answer the second question mentioned above.  Each row of the csv file corresponds to a default catboost model trained with a particular bundle of features and evaluated on a particular evaluation set. The models always use the maximum training set size of 3696041. It’s the same training set as mentioned earlier.  There are two options for the bundle of features:   1. key\_records: It’s the same set of features as indicated by “key\_records” in the other csv file. It’s GBAPERSOONTAB plus GBAHUISHOUDENSBUS. 2. augmented\_records: “augmented\_records” is also defined identically across the two csv files.   There are 16 options for evaluation\_set:   1. Two of the evaluation sets are defined by the sex variable GBAGESLACHT from GBAPERSOONTAB. One set contains all men in the evaluation set curated by Lisa Sivak, the other contains all women. 2. Seven of the evaluation sets are defined by age group using GBAGEBOORTEJAAR from GBAPERSOONTAB. There are seven groups each covering a span of four years, and the total time range is between 1975 and 2002. 3. Six of the evaluation sets are defined by the highest level of education one has achieved in 2020 (with diploma) according to ISCED classification. Lisa Sivak constructed this variable using information from HOOGSTEOPLTAB. The levels include pre-primary, primary, lower-secondary, upper-secondary, short-cycle tertiary, bachelor’s degree, master’s degree, and doctor’s degree.   As indicated by the n\_evaluation\_set column, each of the evaluation sets have at least 1501 individuals.  We calculate the following prediction metrics: LogLoss, MSE, R2\_Score, F1\_Score (with threshold), as well as the area under the receiver operating characteristics curve (AUC). Again, these all represent the model’s average performance among a large group of people and reveal no microdata. |
| 1 – Does this output contain zeros? | No |
| 2 – Does this output contain values < 10 or is it possible to recalculate values < 10? | No |
| 3 – Is there any group disclosure or dominance in this output? | No |
| 4 – Does this output contain information on 1 institution/company/household/etc? | No |
| 5 – Are there any other points where this output does not meet the [output guidelines](file:///K:\Utilities\Regels_Richtlijnen\Guidelines%20for%20RA%20Output.pdf) ? | No |
| If you have answered “Yes” to at least one of the above 5 questions, please give a short but clear explanation on why there is no risk of disclosure in your opinion. |  |

1. If this form is not fully or correctly filled in, the output will not be checked. [↑](#footnote-ref-1)
2. In case the output contains results from models, please clearly mention the underlying number of observations. [↑](#footnote-ref-2)