Centre for Economic Performance

Discussion Paper

ISSN 2042-2695

No.1863
July 2022

Human wellbeing and machine learning

Ekaterina Oparina
Caspar Kaiser
Niccolò Gentile
Alexandre Tkatchenko
Andrew E. Clark
Jan-Emmanuel De Neve
Conchita D'Ambrosio

LSE
THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI
Economic and Social Research Council

Abstract
There is a vast literature on the determinants of subjective wellbeing. International organisations and statistical offices are now collecting such survey data at scale. However, standard regression models explain surprisingly little of the variation in wellbeing, limiting our ability to predict it. In response, we here assess the potential of Machine Learning (ML) to help us better understand wellbeing. We analyse wellbeing data on over a million respondents from Germany, the UK, and the United States. In terms of predictive power, our ML approaches perform better than traditional models. Although the size of the improvement is small in absolute terms, it is substantial when compared to that of key variables like health. We moreover find that drastically expanding the set of explanatory variables doubles the predictive power of both OLS and the ML approaches on unseen data. The variables identified as important by our ML algorithms – i.e. material conditions, health, and meaningful social relations – are similar to those that have already been identified in the literature. In that sense, our data-driven ML results validate the findings from conventional approaches.

Key words: subjective wellbeing, prediction methods, machine learning
JEL: C63; C53; I31

This paper was produced as part of the Centre's Community Wellbeing Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We thank Filippo Volpin for excellent research assistance and Sid Bhushan for early discussions on the topic. We are grateful to the participants of the LSE Wellbeing Seminar, the STATEC Wellbeing 2022 Conference, and the 2022 IAAEU Workshop on Health and the Labour Market for their comments and suggestions. Funding via the ERC Grant Agreement n. 856455, and the Institute for Advanced Studies, University of Luxembourg, Grant DSEWELL is gratefully acknowledged. We thank The Gallup Organization for providing access to their data for this research project.
Andrew E. Clark, Paris School of Economics and Centre for Economic Performance, London School of Economics. Conchita D'Ambrosio, University of Luxembourg. Jan-Emmanuel De Neve, University of Oxford and Centre for Economic Performance, London School of Economics. Niccolò Gentile, University of Luxembourg. Caspar Kaiser, University of Oxford. Ekaterina Oparina Centre for Economic Performance, London School of Economics. Alexandre Tkatchenko, University of Luxembourg.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Over the last 40 years, researchers from various fields have established an immense literature on the correlates and determinants of subjective wellbeing (Clark (2018), Diener et al. (2018), Nikolova and Graham, 2020). In parallel, international organisations (OECD (2020)) and national governments (ONS (2021)) have turned to subjective wellbeing data as a key tool for policy analysis. Such data are also increasingly used to proxy individual welfare in economic scholarship (e.g. Benjamin et al. (2014)).

However, despite the widespread use of wellbeing scores, our current ability to predict wellbeing is limited. Conventional linear models, where variables are selected based on intuition or theory, explain little individual-level variation. Typically, models of individual wellbeing produce an R-squared of no more than 15%.

In response, we here evaluate whether machine learning (ML) algorithms can improve our capacity to understand wellbeing. We answer two research questions:

> **RQ1:** Are ML algorithms significantly better at predicting wellbeing than conventional linear models? What is the upper limit on our ability to predict wellbeing based on survey data?
> **RQ2:** Are the variables that are identified by ML algorithms as important in predicting wellbeing the same as those in the conventional literature?

To answer these questions, we use random forests (Breiman (2001), Hastie et al. (2009)), gradient boosting (Friedman (2001), Natekin and Knoll (2013)), and penalised regressions (Tibshirani (1996)) as examples of ML algorithms. Random forests and gradient boosting are tree-based algorithms that have been shown to perform well with tabular data (Shwartz-Ziv and Armon (2022)).[1] Penalised regressions are a convenient tool for analyses that involve large number of covariates, like ours (Tibshirani (1996)). Generally, these techniques can identify more-complex models of wellbeing than traditional linear models, potentially improving predictive performance. Unlike standard regression techniques, these algorithms allow for the inclusion of an arbitrary number of variables, and, in the case of our tree-based methods, can identify nonlinearities and interactions between variables.

To the best of our knowledge, this paper is the first systematic attempt to evaluate the (dis) advantages of using ML for studying wellbeing at a global scale. Earlier work focused on relatively small country- and year-specific samples (Margolis et al. (2021)), or particular drivers of wellbeing, such as age (Kaiser et al. (2022)).

We carry out our empirical analysis using three of the largest currently-available datasets that include wellbeing information: the German Socio-Economic Panel (SOEP), the UK Household Lon-

[1]Other ML algorithms like neural networks tend to perform poorly on tabular data. We therefore do not consider them here (Borisov et al. (2022)). In preliminary analyses we indeed also found that feed-forward neural networks yielded performances that were no better than OLS.

gitudinal Study (UKHLS), and the American Gallup Daily Poll. The SOEP has data on about 30,000 unique respondents and 400 distinct variables; the UKHLS surveys on around 40,000 individuals in each wave and over 500 distinct variables; and each year of the Gallup data has information on around 200,000 respondents with approximately 60 distinct variables. We can thus study the extent to which utilising more information about individual respondents improves the predictive power of wellbeing models.

Regarding RQ1, we find that ML algorithms predict somewhat better than standard linear models. The size of this improvement is small in absolute terms, but substantial when compared to the predictive power of key variables, such as health. Increasing the number of variables in the model from a standard set (we call this the "Restricted Set") to all available data (the "Extended Set") has a far larger effect on predictive model performance. Predictive accuracy, judged by the R-squared on unseen data, roughly doubles for both OLS and ML methods. Independently of the type of algorithm, an R-squared of 0.30 appears to be the feasible maximum given the available data.

For RQ2, our data-driven ML results validate the findings of the conventional literature. We find that variables related to respondents' social connections, health and material conditions are consistently among the most important in predicting wellbeing. Variable importance is assessed using permutation importances (Breiman (2001), Kuh et al. (2002)) and by computing pseudo partial effects for all algorithms, including OLS. In general, there is substantial correlation in variable-importance rankings across algorithms ($\rho = 0.58$ to $\rho = 0.83$), so that ML approaches and OLS are largely in agreement in terms of what matters most for wellbeing.

2 Methods

2.1 Data

We analyse data from three nationally-representative surveys over the 2010 to 2018 period: the German Socio-Economic Panel (SOEP), the UK Longitudinal Household Survey (UKHLS), and the US Gallup Daily Poll (Gallup).

The Gallup data covers the US adult population, with daily cross-sectional telephone-based surveys of 500 (1000 until 2012) respondents. After removing incomplete data, this yields an annual sample ranging from N=115,192 to N=351,875. Wellbeing is measured by the Cantril Ladder of Life (Cantril (1965)), which is recorded on a scale from 0 to 10, with equal steps between response options.[2]

[2]There has been controversy about whether such data can support inferences about underlying wellbeing (Bond and Lang (2019), Chen et al. (2019), Kaiser and Vendrik (2020), Schröder and Yitzhaki (2017)). We remain agnostic about this issue, and merely ask which algorithms and models best predict answers to wellbeing questions, without making further claims about how these answers relate to respondents' underlying feelings.

The SOEP is representative of the German adult population, with interviews conducted in person (SOEP (2021), Goebel et al. (2019)). To allow for a direct comparison with the Gallup data, we here consider the survey period between 2010 and 2018. In each year, between N=26,089 and N=32,333 observations are available. Life satisfaction is measured on a scale from 0 to 10.

The UKHLS is representative of the UK adult population (UKHLS (2021)). Interviews are conducted in person. We confine our analysis to the same 2010-2018 period (corresponding to Waves 2 to 10). The number of available annual observations is between N=29,605 to N=40,679. Life satisfaction is measured on a 1 to 7 scale.

Descriptive statistics and histograms of each wellbeing measure are shown in Appendix Figure A1.

2.2 Algorithms

We model wellbeing using four kinds of algorithms. First, as our baseline and corresponding to the workhorse of a great deal of research on subjective wellbeing, we apply **Ordinary Least Squares (OLS)** regressions. OLS estimates are the solution to the problem:

$$\arg\min_b \sum_{i=1}^{N} (x_i'b - s_i)^2 \tag{1}$$

Here, x_i is a vector of explanatory variables and b the vector of coefficients. The wellbeing of respondent i is denoted by s_i. Let $\hat{b}$ be the solution of Equation 1. Then, the predicted wellbeing level on the respondent i is $\hat{s}_i = x_i'\hat{b}$. When using OLS, the researcher implicitly assumes that reported wellbeing is a linear combination of the chosen set of explanatory variables x. If these assumptions are an appropriate description of the true data-generating process, OLS will provide accurate predictions of individual wellbeing. However, in applications with a large number of covariates, the performance of OLS may degrade due to overfitting or multicollinearity between included explanatory variables.

The second algorithm we consider, the **Least Absolute Shrinkage and Selection Operator (LASSO)**, tackles this issue by adding a penalty for the sum of the magnitudes of the estimated coefficients. In particular, LASSO estimates are the solution to:

$$\arg\min_b \sum_{i=1}^{N} (x_i'b - s_i)^2 + \lambda \sum_{k=1}^{K} |b_k| \tag{2}$$

Here, λ is a hyperparameter, the preferred value of which is found using a grid search. LASSO and OLS are equivalent for $\lambda = 0$. Although LASSO may improve predictions by reducing the risk of overfitting, the algorithm continues to assume an additive functional form. Nevertheless, one helpful property of LASSO is that it shrinks coefficients on the variables with low explanatory power to zero. In some specifications, we thus use LASSO as a device for variable selection.

The third and fourth algorithms we consider – **Random Forests (RF)** and **Gradient Boosting (GB)** – are based on regression trees (Breiman (1984)). Regression trees are generated via a recursive binary splitting algorithm. The algorithm splits the sample along values of covariates and predicts the outcome in each subsample, or *node*, as the mean outcome within each node. More formally, at each step k, the data D is split into two nodes $D_{L,k}$ and $D_{R,k}$. The location of the split within the data is determined by some variable x_j and an associated threshold $\tau_{k,j}$. The nodes $D_{L,k}$ and $D_{R,k}$ are defined as (see Hastie et al. (2009)):

$$D_{(L,k)} = \{x|x_j < \tau_{k,j}\}; D_{(R,k)} = \{x|x_j \geq \tau_{k,j}\} \tag{3}$$

The predicted values are the mean value of s within each node, *i.e.* $\hat{s}_{D_{m,k}} = N^{-1}_{D_{m,k}} \sum_{i:X_i \in D_{m,k}} s_i$, for $m \in \{L, R\}$, where $N_{D_{m,k}}$ is the number of respondents in each node. At each step, the splitting variable x_j and the threshold $\tau_{k,j}$ are determined by minimising the following residual sum of squares:

$$\sum_{i:X_i \in D_{L,k}} (s_i - \hat{s}_{D_{L,k}})^2 + \sum_{i:X_i \in D_{R,k}} (s_i - \hat{s}_{D_{R,k}})^2 \tag{4}$$

Finally, the nodes $D_{L,k}$ and $D_{R,k}$ are in turn used as inputs for the next step. This procedure is repeated until some final number of *leaves* is found. By construction, every split reduces the in-sample mean squared error (MSE). Hence, if the size of the tree is not limited, the algorithm will overfit the data. Limiting the maximum tree size can ameliorate this issue. However, this comes at the cost of increasing the bias of the resulting estimates (Hastie et al. (2009)). Alternatively, the variance in the predictions can be reduced by aggregating the predictions from multiple trees. Random forests and gradient boosting are both examples of this strategy.

Specifically, **Random Forests**, the third algorithm we consider, rely on averaging across a large number of trees (which we set to 1,000 for all the three datasets).[3] Each individual tree has low bias but high variance. When the correlation between the trees is low, averaging across the predictions of multiple trees reduces the variance of the predictions without introducing additional bias. To carry out this procedure, each individual tree is grown on a nonparametrically-bootstrapped sample of the original data. The correlation between trees is further reduced by considering only a random subset of all covariates at each split. The size of this subset, *Nvars*, is a hyperparameter that we select based on a grid search.

The fourth algorithm, **Gradient Boosting**, proceeds by sequentially fitting regression trees on the residuals of the predictions of the previous collection of trees.[4] Intuitively, each subsequent tree

[3]The performance of random forests is non-decreasing in the number of trees. In our application, increasing the number of trees to 2,000 for UKHLS and Gallup and to 10,000 for SOEP yielded similar results. The final number of trees was chosen to render the optimisation less computationally-expensive.

[4]This construction of the trees, when the residuals from the previous tree are used to build the following tree, is specific to a case when the partitioning of the tree is chosen to minimise the sum of squared residuals in each node. See Friedman (2001) and Hastie et al. (2009) for the general case. We here use a standard implementation of gradient boosting. In preliminary tests, we also evaluated the performance of extreme gradient boosting (XGBoost; Chen and Guestrin (2016)). Using XGBoost only yielded negligible improvements compared to standard gradient boosting, which is why we focused on the latter.

attempts to explain the variance that was not explained by the previous trees. We begin with the predictions $\hat{s}_{T_1}$ of a first tree T_1 and calculate the residual $\hat{s}_{T_1} - s_i = e_{T_1}$. A second tree T_2 is then fitted on these residuals to obtain predicted residuals, $\hat{e}_{T_1}$. The updated overall predictions are then given by $\hat{s}_{T_1} + \hat{e}_{T_1} = \hat{s}_{T_2}$. A third tree is subsequently trained on the residuals $\hat{s}_{T_2} - s_i = e_{T_2}$. This process is repeated *Ntrees* times, producing increasingly accurate predictions of s. Since gradient-boosted collections of trees overfit with large *Ntrees*, we select this hyperparameter via a grid search. To further reduce overfitting, the size of the update at each step is reduced by adding a penalty $0 < \gamma \leq 1$, and predictions are updated with the rule $\hat{s}_{T_k} + \gamma \hat{e}_{Tk} = \hat{s}_{Tk+1}$. The penalty γ is also selected via a grid search.[5]

As is customary, the algorithms are trained on the training set, which here contains 80% of the sample. Each algorithm's performance is then estimated on the test set, which contains the remaining 20% of observations. The optimal hyperparameters are chosen via 4-fold cross validation[6] on the training set using grid search. The optimal hyperparameters for all the datasets can be found in Appendix Table A1. Each of these algorithms are implemented using the scikit-learn library in Python (Pedregosa et al. (2011)). To evaluate the stability of our results across time, where feasible, we train each algorithm on each survey-wave combination separately.

2.3 Explanatory variables

We evaluate each algorithm's performance for two different sets of explanatory variables.

As noted above, we first consider a restricted set of variables that are observed in all three of the datasets, which cover basic demographics as well as economic and health variables. We specifically include: sex, age, age-squared, ethnicity, religiosity, number of household members, number of children in the household, marital status, log household income (equivalised used the modified OECD scale), general health status, disability status, body mass index, labour-force status, working hours, home ownership, area of residence, and interview month. A more detailed description of these variables is provided in Appendix Table A2. These variables are typical in the conventional literature on subjective wellbeing. This restricted set of variables will then allow us to assess the performance of ML algorithms relative to OLS in a standard estimation setting.

We also evaluate each algorithm on much larger extended sets of explanatory variables. Here, we only use the 2013 Wave of Gallup and SOEP, and Wave 3 of the UKHLS (which covers 2011-2012).[7] Our dataset includes all of the available variables, apart from direct measures of subjective wellbeing

[5] The maximum size of each tree in gradient-boosting is significantly smaller than in random forests. Consequently, the individual trees in such ensembles are called *weak learners* (Freund (1995), Freund and Schapire (1999)).

[6] Cross-validation is used to mimic the predictive performance of a machine learning model on unseen data. The training set is split into 4 sub-samples. At each step one of the sub-samples is held out while the algorithm is fitted on the remaining 3 sub-samples. Performance is then evaluated on the hold-out sample. Finally, the parameters that are associated with the best predictive performance are chosen.

[7] These waves/years were chosen as they include personality traits in the SOEP and UKHLS.

(such as domain satisfaction, happiness, or subjective health) or mental health. We also exclude variables with more than 50% missing values. The resulting Gallup dataset contains 67 variables, and around 450 variables are retained in the SOEP and UKHLS. Missing values for continuous variables are assigned the observed means, while missing values for categorical variables are assigned a new category. We convert categorical variables into a set of dummies, one for each category. [8]

The large number of variables in the extended set produces significant computational burden. At the same time, it is evident that some portion of these variables will have no predictive power for wellbeing. We therefore use LASSO as a device to select the explanatory variables (Tibshirani (1996), Ahrens et al. (2020)).[9] We have carried out the estimations on both the full extended set and the post-LASSO extended set. Typically, both approaches perform similarly. For simplicity, we only show results for the approach that performed better in each individual case.

2.4 Assessing Variable importance

To answer our second research question, we need to assess how important each explanatory variable is in enabling our algorithms to predict wellbeing. We do so in two ways.

We first use *permutation importances* (PIs) to measure the degree to which each algorithm relies on a given variable in making its predictions (Molnar (2022)).[10] PIs are calculated by randomly shuffling a given variable's observed values across individuals in the test data and evaluating the extent to which the predictive performance (in terms of R-squared) of a given algorithm falls when permuting the variable's values. This operation is carried out 10 times. The reported PI is the average change in the R-squared across these 10 iterations. The greater the average fall in the R-squared, the more important is the variable.

To understand the direction of our variables' effects we also report *pseudo partial effects* (PPEs). These are calculated by taking the difference in predicted wellbeing after setting each explanatory variable to a given set of values. Specifically, for continuous and ordinal variables we set the variable to the third and first quartile of their distributions and calculate the mean difference in predicted wellbeing. For binary variables (including dummies for all of the categorical variables), we predict wellbeing when setting each individual's value to either 0 or 1.

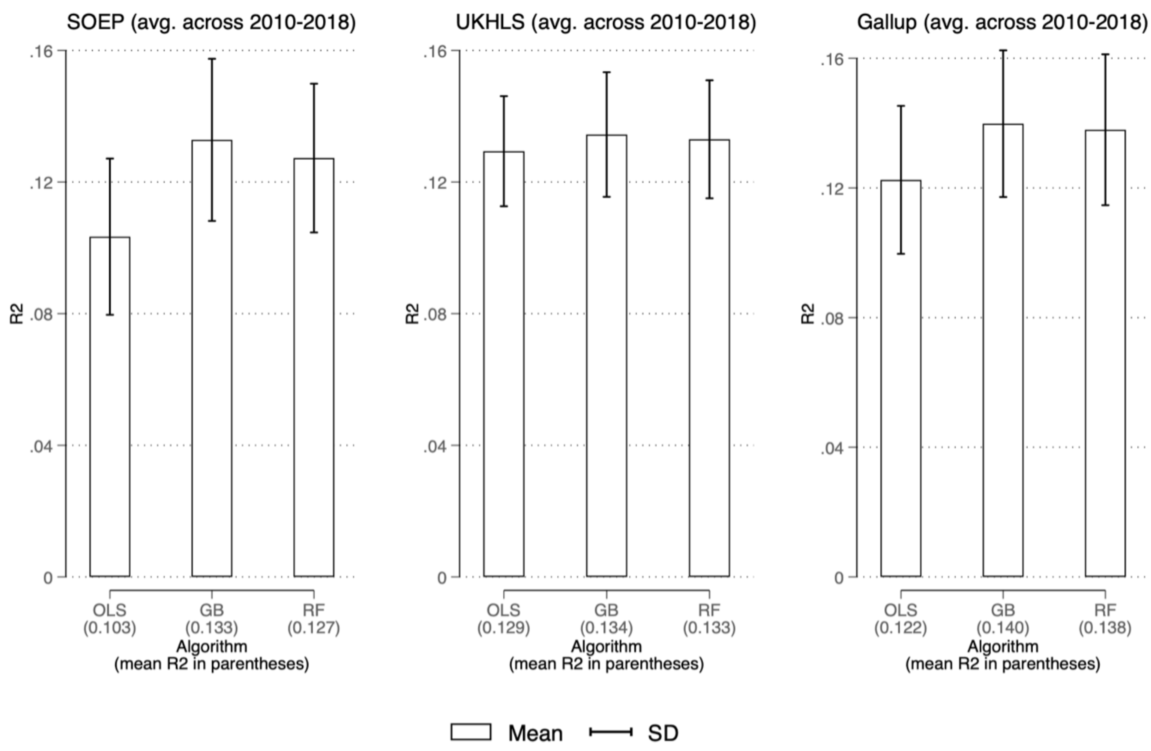
A key advantage of PIs and PPEs is that they can be used with any kind of algorithm, allowing us to compare the way in which each algorithm makes use of the available data.

[8] Processing categorical variables and removing perfectly collinear variables respectively yields 210, 542, and 957 effective explanatory variables in the Gallup, SOEP and UKHLS datasets. The full list of variables in this extended set is available from the following repository link.

[9] Using LASSO on the restricted set of variables produced a similar performance to OLS, with optimal λ close to 0.

[10] Shapley values are an alternative option to assess feature importances. We did not compute Shapley Values because of their substantial computational complexity (Lundberg et al. (2018);Yang (2021)) and since our pseudo marginal effects already allow us to identify the direction of variables' effects.

Figure 1: R-squared figures from OLS, GB and RF using the restricted set of variables. The R-squareds are computed using the unseen testing data.



3 Results

3.1 Model performance

We begin with RQ1, *i.e.* whether ML algorithms significantly outperform OLS in predicting wellbeing. As noted, OLS is the standard approach followed in the conventional literature.

3.1.1 The Restricted Set of explanatory variables

We start with the analysis based on the restricted set of covariates, which includes the variables that are typical in many conventional wellbeing estimations. Figure 1 depicts the performance of each algorithm on the test-set portion of each dataset. We use R-squared as our primary evaluation metric in order to facilitate the comparison with previous analyses.

In Figure 1 each algorithm is trained separately for each year between 2010 and 2018. The values refer to the average R-squared across these years and their standard deviations. The R-squareds are very similar across datasets, ranging from 0.10 (SOEP) to 0.14 (Gallup). Gradient boosting (GB) and random forests (RF) yield larger R-squared values than OLS in each case. Specifically, random forests yield absolute increases in R-squared of 0.024 (SOEP), 0.004 (UKHLS) and 0.016 (Gallup); the respective improvements from using gradient boosting are slightly larger, with respective R-

Table 1: An illustration of the size of the improvements from using ML.

	OLS, full	OLS, no health	GB	GB gain as % of loss from removing health
Panel A: Restricted set of variables				
SOEP	0.103	0.075 (Δ=0.028)	0.133 (Δ=0.030)	107%
UKHLS	0.129	0.095 (Δ=0.034)	0.134 (Δ=0.005)	15%
Gallup	0.122	0.093 (Δ=0.029)	0.140 (Δ=0.018)	62%
Panel B: Extended set of variables				
SOEP	0.284	0.240 (Δ=0.043)	0.318 (Δ=0.035)	81%
UKHLS	0.215	0.197 (Δ=0.018)	0.243 (Δ=0.028)	155%
Gallup	0.270	0.240 (Δ=0.031)	0.280 (Δ=0.018)	58%

Notes: The figures refer to the R-squared values from the test-set.

squared gains of 0.030, 0.005, and 0.018.[11] ML algorithms thus do outperform linear regressions, and gradient boosting always outperforms random forests.

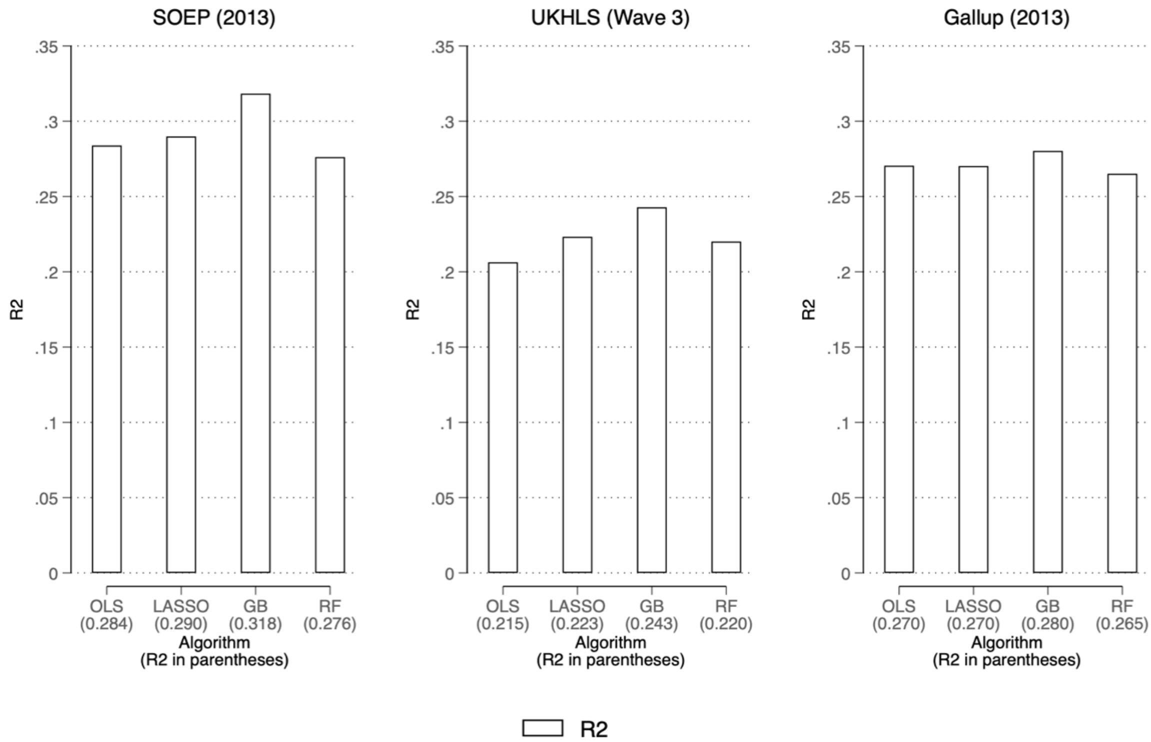
These gain figures considered on their own are hard to interpret. To illustrate the substantive size of these improvements, we compare them to the change in predictive performance when omitting information on respondent's health status – a key wellbeing predictor – in our baseline OLS regressions. Panel A of Table 1 lists the changes in the test-set R-squared of the OLS regression when omitting this information and compares this figure to the gain from using gradient boosting. As benchmarked against the gain from adding health information, the prediction-improvement figure from gradient boosting (as our best ML algorithm) lies between 15% and 107%. When evaluated in this way, the gains from using ML do look substantial.

3.1.2 The Extended Set of explanatory variables

Adding further explanatory variables should increase our ability to predict wellbeing. Given the greater flexibility of the ML algorithms, we should expect these to benefit more from additional variables than OLS. To test this, we estimate all of our models on the extended sets of variables. As explained in Section 2.3, these extended sets include all of the variables available in the 2013 waves

[11] These gains are calculated from the test set, which was not used for training the algorithm. In the training set, *i.e.* the data that is observed by each algorithm, the improvement of the ML algorithms over OLS is larger (see Appendix Figure A2). Therefore, the predictive capacity of the ML algorithms does not seem constrained by underfitting. Of course, performance in the training set is not *per se* indicative of the quality of an algorithm: A decision tree with as many leaves as training individuals would yield an MSE of 0. However, this model would perform extremely poorly when used to assess unseen test data.

Figure 2: R-squared figures from OLS, LASSO, GB and RF using the extended set of variables. The R-squareds are computed using the unseen testing data.



of the SOEP and Gallup, and Wave 3 of the UKHLS.

Figure 2 depicts our main results.[12] The R-squared figure approximately doubles using the extended set for all algorithms, including OLS. The OLS R-squared is now 0.28 for the SOEP, 0.21 in the UKHLS and 0.27 for Gallup. As such, standard economic specifications do not fully exploit the predictive information available in typical large-scale survey data.[13]

Gradient boosting remains the best-performing algorithm and clearly predicts better than OLS. The absolute gain in the R-squared from gradient boosting over OLS is now 0.034, 0.028 and 0.010 for the SOEP, UKHLS and Gallup respectively. Random forests now tend to perform poorly, under-performing OLS for SOEP and Gallup. This has also been observed in other empirical applications where covariates were measured with error (Reis et al. (2018)).

We again interpret the size of the gains from gradient boosting by comparing them to those from the inclusion of respondents' health information when using OLS.[14] The results in Panel B of Table 1 illustrate that these gains are again substantial, being approximately equivalent to the role of health

[12]The results for the training set can be found in Appendix Figure A3.

[13]All of these R-squared estimates are obtained using the test set. Hence, these improvements cannot be attributed to a mechanical increase in the share of explained variance due to adding more variables to the model.

[14]In these extended specifications, there are multiple variables related to health in each dataset. We remove 21, 19 and 12 health-related variables in the Gallup, the SOEP and UKHLS respectively.

in predicting wellbeing.

We thus conclude that tree-based ML algorithms can provide improvements in predictive performance over conventional methods. These gains are moderate in absolute terms, but are meaningful when compared to the predictive power of health. However, we also note that these gains are obtained with algorithms that take up to 100 times longer to estimate.[15] The use of ML algorithms thus involves a trade-off between computational burden and predictive performance.

There are several reasons why nonlinear ML methods do not yield an even larger improvement in predicting human wellbeing. First, most of the independent variables we include are binary or categorical. Such data cannot exhibit non-linearities except via interactions between variables. Second, it is possible that non-linear relationships do exist but the variables concerned have a small contribution to the outcome. This is particularly likely if there are many variables contributing to wellbeing, as is the case in our extended set of independent variables. Finally, it may be that non-linearities are present only at the extremes of the distribution where only few observations exist.

That said, apart from improvements in predictive performance, ML may also indicate new, and potentially-overlooked, variables that are key in explaining subjective wellbeing. The next section explores this idea.

3.2 Variable importance

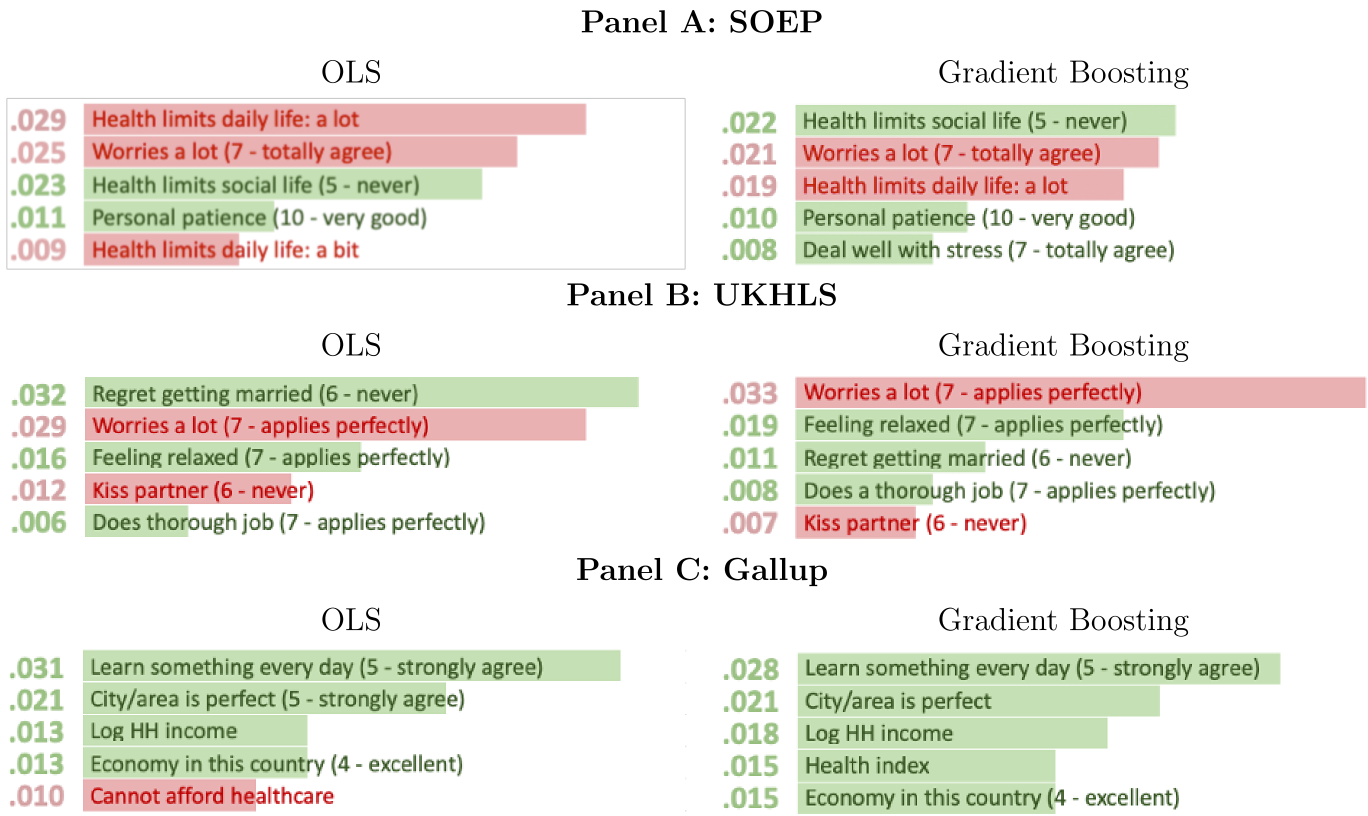
In this section we ask whether the variables that ML identifies as important in predicting life satisfaction correspond to those emphasised in the conventional literature. We do so by estimating variable importances, as discussed in Section 2.4. Our ML-based findings turn out to fit well with the results in previous analyses.

We start by estimating variable importances in the extended dataset, which provides more possibilities for the identification of important variables that do not appear in conventional wellbeing models. Figure 3 lists the five most-important variables identified in OLS and GB, which is the best-performing ML algorithm, in each dataset.[16] The bars and numerical values refer to permutation importance, *i.e.* the drop in the model's R-squared when the values of the variable are randomly permuted across respondents. The variables that are negatively associated with average wellbeing are in red, and those with a positive association in green. In all three countries, individual health and interpersonal relationships are among the most-important predictors. As expected, respondents whose health limits their activities are on average less satisfied, while people with fulfilling relationships are typically more satisfied with their lives. The directions of the estimated effects are in line with those in the previous conventional work. ML algorithms and OLS thus generally agree on the direction and approximate size of the most-important variables (see Appendix Table A3 for the

[15]This figure is based on a comparison between OLS and RF on the Gallup data with the extended dataset.

[16]We present the Top-10 most-important variables for OLS, RF and GB in the three datasets in Appendix Table A3.

Figure 3: Permutation importance and pseudo partial effects of OLS and GB on the extended set of variables, five most-important variables.



Notes: The bars and numerical values represent permutation importances and are coloured red for variables with negative pseudo partial effects and green otherwise. For Likert scale variables, the highest category is reported.

effect-size estimates).

As a more systematic measure of the degree of agreement between ML and OLS, we calculated the correlations of the ranks (in terms of their permutation importance) of each variable across algorithms and datasets. The detailed results appear in Appendix Table A4. Substantively, there is strong agreement between GB and RF in all three datasets, with the rank correlation figure never falling below 0.79. The correlations with the OLS ranking are somewhat lower, with a minimum value of 0.58 (OLS *vs.* RF in SOEP). Nevertheless, we can strongly reject ($p < 0.001$) the null hypothesis that the rankings are uncorrelated, supporting our conclusion that the OLS and ML algorithms are in broad agreement.

Apart from the conventional variables used in wellbeing analyses, such as health and interpersonal relationships, the algorithms also identify personality traits as important predictors in the UKHLS and SOEP. Personality traits, unfortunately, do not appear in the Gallup survey. In the UK data, measures associated with (the absence of) neuroticism (*i.e.* worrying, and feeling relaxed) appear in the Top-3. In German data, worrying a lot, being able to deal with stress, and patience are among the most-important variables in all empirical approaches. This is line with previous research underlining the potential advantages of including personality traits in wellbeing regressions (Ferrer-i-Carbonell and Frijters (2004), Proto and Zhang (2021)).

Beyond these similarities, there are some cross-country differences. The most striking concern financial factors. These are important in the US (e.g., household income and being able to pay for healthcare) but not in the other countries. To see whether this is a genuine finding or a consequence of differences in variable availability across countries, we carry out the same analysis using the restricted set (for which we have a common set of variables). When we do so, the cross-country differences in the importance of income largely disappear. More generally, the variables identified as most important in these harmonised datasets are very similar across the three countries (see Appendix Table A5 for detailed results). They include health, income, marital and employment status, as well as home-ownership – a proxy for wealth – and age. Sex and ethnicity are only important in the US. Education is among the most important factors in the US and Germany, but not in the UK.

3.3 Additional analyses and robustness tests

3.3.1 Wellbeing by age and income

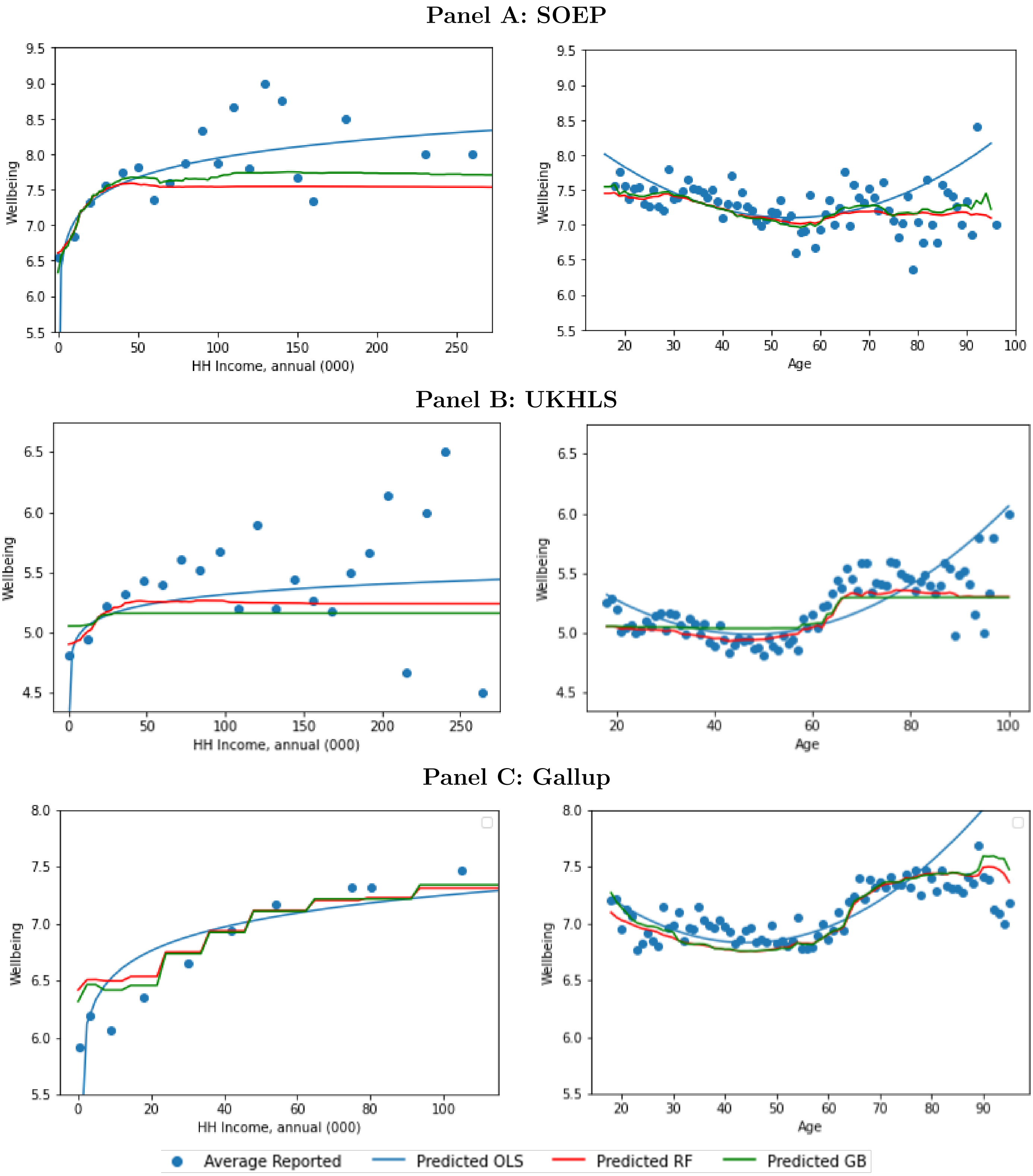
The preceding section concluded that the kinds of variables that machine learning finds to be important - and the estimated direction of their association with wellbeing - are largely in line with the results in the conventional literature. We here present a detailed analysis of two variables that have attracted a great deal of interest in the conventional literature: age and income. In OLS estimation, the functional forms associated with these two variables are imposed by the analyst, while they are instead freely estimated in our tree-based ML algorithms.

The results appear in Figure 4 and Appendix Figure A4. In the OLS estimation, illustrated in blue, we assume a quadratic form for age, and a log-linear functional form for income, which are very common functional forms in this literature. The relationships for RF are in red, and those for GB in green.

For low to medium incomes, both ML algorithms track the assumed log-normal functional form remarkably closely, in line with the conventional literature. However, once we reach relatively-high equivalised annual income figures, above 50,000 EUR in the SOEP or 40,000 GBP in the UKHLS, the ML algorithms suggest that wellbeing no longer increases with income. We cannot confirm this finding in the US, as income in Gallup appears in bands with the highest band being 100,000 USD or above. In 2013, 100,000 USD were approximately equivalent to 70,000 GBP or 78,000 EUR.[17] In addition, the Gallup 2013 wave did not collect data on household size. As a result, household income in Gallup is not directly comparable to the adjusted equivalent incomes in SOEP and UKHLS. Given these caveats, we do not find evidence of satiation in the US data. Our ML findings are therefore in

[17] https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm

Figure 4: The mean effects of age and household income on wellbeing, restricted set of variables.



Notes: For the UKHLS and the SOEP annual income is constrained to be less than or equal to a figure of 250 000 in the local currency. This covers over 99.9% of the income distribution in both countries. In SOEP and UKHLS, incomes are recorded as a continuous variable and equivalence-scale adjusted household income is used for the analysis. Income data in Gallup is collected in income bands, and household size data was not collected in 2013. Thus, we here use non-adjusted HH income data.

line with previous work on wellbeing using US data (Kahneman and Deaton (2010), Killingsworth (2021)).

With respect to the relationship between age and wellbeing, our ML estimations replicate the well-known approximate U-shape up to age 70 (e.g. Cheng et al. (2017)), which is more pronounced in the US. However, unlike the smooth U-shape assumed in the OLS approach, we find a much more pronounced "kink" at around age 65 for each dataset and ML-algorithm. We suspect that this kink reflects the gains in wellbeing following on from retirement (Gorry et al. (2018), Wetzel et al. (2016)).

3.3.2 Positive and negative affect

We also evaluate the performance of gradient boosting and random forests on measures of positive and negative affect. The results show that our findings are not specific to the use of life evaluations as the measure of subjective wellbeing, but generalise to affect (or mood). In the 2013 Gallup data, positive affect is measured by the average figure from dummy variables indicating whether the respondent felt happiness or joy, or smiled during the previous day. Negative affect is calculated analogously from dummies indicating pain, worry, sadness, and anger. In the German SOEP, positive affect is the self-reported frequency of being happy in the last 4 weeks (on a 1 to 5 scale), and negative affect as the average of the self-reported frequency from three questions about being angry, sad, or worried in the last four weeks (all measured on a 1 to 5 scale). The UKHLS dataset does not contain comparable affect data and is not used in this part of analysis.

The detailed results for the Gallup data appear in the top panels of Figure A5 and Table A6. It is notable that negative affect is easier to predict than positive affect. This finding holds across algorithms, with R-squared figures ranging from 0.423 and 0.464 for negative affect, and between 0.261 and 0.296 for positive affect. Random forests and gradient boosting outperform both OLS and LASSO. As was the case for life evaluations, gradient boosting again performs the best, with gains in R-squared over OLS of 0.041 for negative affect and 0.036 for positive affect. Regarding variable importances, Table A6 shows that good health is even more important for predicting positive and negative affect in the Gallup data than it was for life evaluation. Moreover, in line with previous work (e.g. Kahneman and Deaton (2010)), variables relating to material conditions – like income – do not feature in the list of the set of most-important variables when modelling affect.

Our results are qualitatively similar in the German data: gradient boosting again performs best, and positive affect is harder to predict than negative affect (see Appendix Table A7 and the bottom panels of Figure A5).

3.3.3 Panel data

Our main findings regarding the ML estimation of wellbeing are also robust to exploiting the panel dimension of the German SOEP and the UKHLS. As there is no standard procedure for the introduction of individual fixed effects in the ML algorithms that we use, we implement an approach similar to the Mundlak correction for linear models (Mundlak (1978), Wooldridge (2010)): we pool all years of the UKHLS and SOEP data, demean all covariates at the individual level and include both an individual's average value over time of each covariate as well as their year-specific deviations from their individual mean. The level of wellbeing is the dependent variable, as was the case in the analysis above.

The relative predictive performance of the OLS and ML in the pooled dataset is similar to the findings for individual years. In the UKHLS, the OLS R-squared is 0.140. The use of RF produces a small improvement, with the R-squared increasing to 0.143. Gradient boosting provides a further improvement, yielding an R-squared of 0.150. In the German SOEP, the OLS R-Squared is 0.122, with once again both the random forest and gradient boosting leading to better R-Squared figures of, respectively, 0.150 and 0.156. As shown in Tables A8 and A9, the most important variables predicting the level of wellbeing are almost exclusively the average values of the individual covariates. One exception in both the UKHLS and SOEP is the *Health limits activities* variable. As such, deviations in individual health status (from their average value) seem to be important for the level of individual wellbeing.

4 Discussion

We draw three main conclusions from our analysis above.

First, tree-based ML approaches do indeed perform better at predicting wellbeing than more-conventional linear models. Although the gains in R-squared we obtain are modest in absolute terms, they are comparable with – and sometimes exceed – the extent to which information on respondents' health can improve wellbeing predictions. Comparing the algorithms we consider, gradient boosting consistently outperforms random forests.

Second, when we use all of the non-wellbeing variables that are available in each dataset as predictors, we more than double the explained variation in wellbeing for all of the estimation procedures that we analyse. This extended set of variables produces R-squared figures of around 0.3. These values look to be the maximum achievable with the current survey data.

Third, almost all of the variables that turn out to be important in the specifications using of all the available data relate to health, economic conditions, personality traits, and personal relationships. This purely data-driven process thus picks out the same core determinants of wellbeing as have been identified in the conventional literature. In that sense, machine-learning approaches validate the previous human-guided search for the determinants of wellbeing. This looks to be good news for the field.

We see two directions for future research.

The first is to further explore the capabilities of ML models. We have focused our analysis here on tree-based methods, which are powerful algorithms that perform well in multiple contexts. However, given the specificities of wellbeing data, we might find further improvements by using other algorithms (e.g. Kernel Ridge, Vovk (2013)), or by using a combination of theory-based modelling and algorithmic approaches. Another potential approach is using a combination of unsupervised and supervised learning. For example, it might be possible to separate the dataset into overlapping clusters of individuals chosen based on subsets of independent variables. Then, the predictive performance of non-linear ML models could be substantially higher when applied to such clusters, as compared to using one global model as we have done in our work. Moreover, we have currently only focused on identifying variables that are key for the successful prediction of wellbeing. A natural next step is to extend the use of ML-based algorithms to investigate the variables that are most important for wellbeing in a causal sense (Wager and Athey (2018)).

Second, our analyses focused on rich Western countries. As such, it remains an open question whether our findings would also hold in a more global setting, e.g. in countries where material needs are much more acute. Insofar as there may be greater scope for improving wellbeing in low- and middle-income countries (Helliwell et al. (2022)), applying ML approaches in this setting may be particularly valuable going forward.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). Lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, *20*(1), 176–235.

Benjamin, D. J., Heffetz, O., Kimball, M. S., & Rees-Jones, A. (2014). Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices. *American Economic Review*, *104*(11), 3498–3528.

Bond, T. N., & Lang, K. (2019). The Sad Truth about Happiness Scales. *Journal of Political Economy*, *127*(4), 1629–1640.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep Neural Networks and Tabular Data: A Survey. *arXiv:2110.01889 [cs]*.

Breiman, L. (1984). Classification and regression trees.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cantril, H. (1965). *The pattern of human concerns*. New Brunswick: Rutgers University Press.

Chen, L.-Y., Oparina, E., Powdthavee, N., & Srisuma, S. (2019). Robust ranking of happiness outcomes: A median regression perspective. *arXiv preprint arXiv:1902.07696*.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cheng, T. C., Powdthavee, N., & Oswald, A. J. (2017). Longitudinal Evidence for a Midlife Nadir in Human Well-Being: Results from Four Data Sets. *The Economic Journal*, *127*(599), 126–142.

Clark, A. E. (2018). Four Decades of the Economics of Happiness: Where Next? *Review of Income and Wealth*, *64*(2), 245–269.

Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*, *2*(4), 253–260.

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How Important is Methodology for the Estimates of the Determinants of Happiness? *The Economic Journal*, *114*(497), 641–659.

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, *121*(2), 256–285.

Freund, Y., & Schapire, R. E. (1999). A Short Introduction to Boosting.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360.

Gorry, A., Gorry, D., & Slavov, S. N. (2018). Does retirement improve health and life satisfaction? *Health Economics*, *27*(12), 2067–2086.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Helliwell, J. F., Wang, S., Huang, H., & Norton, M. (2022). Happiness, Benevolence, and Trust During COVID-19 and Beyond (World Happiness Report), 40.

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *PNAS*, *107*(38), 16489–93.

Kaiser, C., & Vendrik, M. C. (2020). How threatening are transformations of happiness scales to subjective wellbeing research? *INET Working Paper*.

Kaiser, M., Otterbach, S., & Sousa-Poza, A. (2022). Using machine learning to uncover the relation between age and life satisfaction. *Scientific Reports*, *12*(1), 5263.

Killingsworth, M. A. (2021). Experienced well-being rises with income, even above $75,000 per year. *Proceedings of the National Academy of Sciences*, *118*(4).

Kuh, D., Hardy, R., Langenberg, C., Richards, M., & Wadsworth, M. E. J. (2002). Mortality in adults aged 26-54 years related to socioeconomic conditions in childhood and adulthood: Post war birth cohort study. *BMJ*, *325*(7372), 1076–1080.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint*.

Margolis, S., Elder, J., Hughes, B., & Lyubomirsky, S. (2021). *What Are the Most Important Predictors of Subjective Well-Being? Insights From Machine Learning and Linear Regression Approaches on the MIDUS Datasets* (tech. rep.). PsyArXiv.

Molnar, C. (2022). *Interpretable Machine Learning*.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 69–85.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*.

Nikolova, M., & Graham, C. (2020). The Economics of Happiness. *Unpublished Working Paper*.

OECD. (2020). *How's Life? 2020: Measuring Well-being*. OECD.

ONS. (2021). Well-being - Office for National Statistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830.

Proto, E., & Zhang, A. (2021). COVID-19 and mental health of individuals with different personalities. *Proceedings of the National Academy of Sciences*, *118*(37).

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, *157*(1), 16.

Schröder, C., & Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, *92*, 337–358.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

SOEP. (2021). Socio-economic panel (SOEP), data for years 1984-2019: SOEP-Core v36.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

UKHLS. (2021). United Kingdom Household Longitudinal Study Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009.

Vovk, V. (2013). Kernel ridge regression. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical inference: Festschrift in honor of vladimir n. vapnik* (pp. 105–116). Springer Berlin Heidelberg.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wetzel, M., Huxhold, O., & Tesch-Römer, C. (2016). Transition into Retirement Affects Life Satisfaction: Short- and Long-Term Development Depends on Last Labor Market Status and Education. *Social Indicators Research*, *125*(3), 991–1009.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Yang, J. (2021). Fast treeshap: Accelerating shap value computation for trees. *arXiv preprint*.

Author contributions: E.O., C.K., and N.G. equally contributed to the study design, data analysis, and data visualisation. A.E.C., J.D.N, A.T., and C.D. provided further input and discussion on the study design, data analysis, and data visualisation. N.G. wrote a significant portion of the initial set of codes for the analysis. C.K. undertook the initial manuscript writing with significant contributions by E.O. Input and revisions to the manuscript writing were provided by N.G. and A.E.C., with further contributions by J.D.N, A.T., and C.D.

Appendix

Table A1: Optimal hyperparameters used in the extended specifications (post-LASSO extended specification in parentheses)

	Panel A: Random Forest		
	SOEP	Gallup	UKHLS
MaxDepth	96 (70)	70 (70)	30 (20)
Nvars	225 (65)	80 (80)	400 (130)
Ntrees	1000 (1000)	1000 (1000)	1000 (1000)
MinLeaf	1 (1)	5 (5)	15 (5)
	Panel B: Gradient Boosting		
	SOEP	Gallup	UKHLS
MaxDepth	8 (8)	3 (3)	5 (7)
Nvars	75 (30)	40 (40)	100 (30)
Ntrees	6000 (2000)	16000 (16000)	2000 (2000)
MinLeaf	1 (1)	1 (1)	1 (1)
Learning rate (γ)	0.005 (0.01)	0.0063 (0.0063)	0.01 (0.01)

Notes: Hyperparameters are identified via a grid search by minimizing the average MSE across 4 folds of cross-validation. *MaxDepth* is the maximum depth of each branch of each tree. *Nvars* is the maximum number of randomly-picked variables used to perform splits within each tree. *MinLeaf* is the minimum number of training individuals that must be in each leaf of a given tree (fixed to 1 for gradient boosting). *Ntrees* is the number of trees fitted (fixed to 1,000 for random forests). The learning rate (γ) is the rate at which predictions are updated (only applicable to gradient boosting).

Table A2: List of variables in the restricted set.

Variable	SOEP	UKHLS	Gallup
Age	16 - 105	18 - 103	18 - 99
Area of residence	16 distinct values	12 regions	51 distinct values
BMI	11.10 - 84.50	11.80 - 74.20	7.19 - 152.56
Disability status	Binary	Binary	n.a.
Education	18 - 7 (years of education)	6 distinct values	6 distinct values
Labour-force status	Binary	12 distinct values	4 distinct values
Equivalised Log HH income	0 - 13.88	-0.80 - 12.52	3.40 - 9.90
Ethnicity/Migration background	3 distinct values (migration background)	18 distinct values (ethnicity)	5 distinct values (ethnicity)
Health	0 – 396 (doctor visits in prev. year)	Health limits activities (3 distinct values)	Binary (self-assessed health problems)
Housing status	4 distinct values	6 distinct values	n.a.
Marital status	5 distinct values	10 distinct values	6 distinct values
Month of interview	12 distinct values	24 distinct values	12 distinct values
Number of children in HH	0 - 11	0 - 9	0 - 15
Number of people in HH	1 - 16	1 - 16	1 - 99
Religion	10 distinct values	Binary	8 distinct values
Sex	Binary	Binary	Binary
Working hours	0 - 6669	0 - 180	4 distinct values

Notes: For continuous variables, the range is reported. For SOEP, possible values for the categorical variables are: *Area of residence*: Each of the 16 Bundesländer. *Ethnicity/Migration background:* No migration background, Direct migration background, Indirect migration background. *Housing status:* Main Tenant, Sub-Tenant, Owner, Nursing Home/ Retirement Community. *Marital status:* Married, Single, Widowed, Separated, Divorced. *Religion:* Catholic, Protestant, Christian Orthodox, Other Christian, Muslim, Muslim (Shiite), Muslim (Sunnite), Muslim (Alevite), Other, No religion. For UKHLS, possible values for the categorical variables are: *Area of residence:* North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland, Northern Ireland. *Education:* Degree, Other higher degree, A-level etc, GCSE etc., Other qualification, No qualification. *Labour-force status:* Self-employed, Paid employment(ft/pt), Unemployed, Retired, On maternity leave, Family care or home, Full-time student, LT sick or disabled, Govt training scheme, Unpaid, family business, On apprenticeship, Doing something else. *Ethnicity:* British/English/Scottish/Welsh/Northern irish, Irish, Gypsy or Irish traveller, Any other white background, White and black caribbean, White and black african, White and asian, Any other mixed background, Indian, Pakistani, Bangladeshi, Chinese, Any other asian background, Caribbean, African, Any other black background, Arab, Any other ethnic group. *Health limits moderate activities:* Yes, a lot; Yes, a little; No, not at all. *Housing status:* Owned outright, Owned/being bought on mortgage, Shared ownership (part-owned part-rented), Rented, Rent free, Other. *Marital status:* Single and never married/in civil partnership, Married, In a registered same-sex civil partnership, Separated but legally married, Divorced, Widowed, Separated from civil partner, A former civil partner, A surviving civil partner, Living as couple. For Gallup, possible values for the categorical variables are: *Area of residence:* 51 States. *Education:* Less than high school, High school, Technical/Vocational school, Some college, College graduate, Post-graduate. *Labour-force status:* Employed, Self-employed, Employed and self-employed, not employed. *Ethnicity:* White, Other, Black, Asian, Hispanic. *Marital status:* Single, Married, Separated, Divorced, Widowed, Living with partner (not married). *Religion:* Protestant, Catholic, Jewish, Muslim, Mormon, Other Christian, Other, No religion. *Working hours:* 30 or more hours per week, 15 to 29 hours per week, 5 to 14 hours per week, less than 5 hours per week.

Table A3: Permutation Importance (PI) and Pseudo Partial Effects (PPE) in OLS, RF and GB on the Extended Set of variables: the 10 most-important variables.

	OLS			Random forest			Gradient boosting		
	Variable name	PI	PPE	Variable name	PI	PPE	Variable name	PI	PPE
				Panel A: SOEP					
1	Health limits daily life: a lot	.029	-.780	Health limits social life	.032	.154	Health limits social life	.022	.172
2	Worry a lot	.025	-.146	Health limits daily life: a lot	.028	-.742	Worry a lot	.021	-.100
3	Health limits social life	.023	.187	Worry a lot	.020	-.113	Health limits daily life: a lot	.019	-.628
4	Personal patience	.011	.129	HH income	.018	.202	Personal patience	.010	.174
5	Health limits daily life: a bit	.009	-.266	Deal well with stress	.015	.160	Deal well with stress	.008	.128
6	Partner in HH	.008	.222	Personal patience	.008	.106	Health limits daily life: a bit	.006	-.220
7	No monthly savings	.008	-.186	No annual holiday trip	.007	-.114	Partner in HH	.006	.152
8	Deal well with stress	.006	.080	No monthly savings	.007	-.110	Risk tolerance	.006	.036
9	House needs repair	.005	-.126	Not unemployed	.006	.303	HH income	.006	.152
10	Hours of sleep on workday	.004	.077	Unemployment benefit	.005	-000	Number of doctor visits	.006	-.086
				Panel B: UKHLS					
1	Regret getting married	.032	.418	Worries a lot (Big 5)	.030	-.146	Worries a lot (Big 5)	.033	-.188
2	Worries a lot (Big 5)	.029	-.274	Feeling relaxed (Big 5)	.027	.238	Feeling relaxed (Big 5)	.019	.212
3	Feeling relaxed (Big 5)	.016	.240	Health limits kind of work	.009	.040	Regret getting married	.011	.209
4	Kiss partner	.012	-.218	Belong to neighbourhood	.009	-.179	Does a thorough job (Big5)	.008	.069
5	Does thorough job (Big 5)	.006	.112	Age squared	.009	.007	Kiss partner	.007	-.110
6	Share interests w. partner	.006	-.161	Regret getting married	.009	.137	Age squared	.007	.002
7	Belong to neighbourhood	.005	-.107	Health limits work amount	.008	.032	Health limits kind of work	.007	.053
8	Sociable (Big 5)	.005	.094	Does thorough job (Big 5)	.007	.053	Health limits work amount	.006	.049
9	Health limits work amount	.005	.070	Consider divorce (never)	.006	.106	Belong to neighbourhood	.006	-.162
10	Long term sick or disabled	.005	-.420	Sociable (Big 5)	.006	.081	Sociable (Big 5)	.006	.126
				Panel C: Gallup					
1	Learn something every day	.031	.43	Learn something every day	.033	.34	Learn something every day	.028	.35
2	City/area is perfect	.021	.32	City/area is perfect	.026	.42	City/area is perfect	.021	.39
3	Log HH income	.013	.15	Log HH income	.021	.30	Log HH income	.018	.26
4	Economy in this country	.013	.21	Cannot afford healthcare	.021	-.54	Health index	.015	.16
5	Cannot afford healthcare	.010	-.38	Economy in this country	.015	.21	Economy in this country	.015	.22
6	Health limits activities	.010	-.04	Physical health index	.013	.15	Cannot afford healthcare	.013	-.40
7	Health encouragement	.010	.12	Health limits activities	.010	-.03	Health encouragement	.008	.17
8	Physical health index	.010	.14	Health encouragement	.010	.17	Health limits activities	.008	-.01
9	Female	.008	.24	Female	.005	.13	Age and age-squared	.005	.03
10	Ever diag. w depression	.008	-.28	Ever diag. w. depression	.005	-.16	Female	.005	.25

Notes: The following variables are shown. SOEP: Dummies: *Health limits daily life a lot*, *Health limits daily life a bit*, *Partner in HH*, *No monthly savings*, *Not unemployed*, *No emergency reserves*, and *No annual holiday trip*. Likert scales: *Limited socially due to health* (1 – always to 5 – never), *Worries a lot* and *Deals well with stress* (1 – not at all to 7 – totally agree), *Personal patience* (0 – very bad to 10 – very good), *House needs repair* (1 – in good condition, 3 – needs major renovation). Continuous: *Log HH income*, *Hours of sleep*, *Number of Doctor visits*, *Risk Tolerance and Unemployment Benefit*. UKHLS: Dummies: *Health not limiting activities*. Likert scales: *Pain interferes with work* (1 – not at all to 5 – extremely), *Regret getting married*, *Share interests w. partner*, *Consider divorce* and *Kiss partner* (1 – all the time, 6 – never), *Health limits work amount* and *Health limits kind of work* (1 – all of the time, 5 – none of the time); Big 5 traits, including *Worries a lot*, *Feeling relaxed*, *Does thorough job*, *Is sociable* (1 – does not apply to 7 – applies perfectly), *Belong to neighbourhood* (1 – strongly agree – 5 strongly disagree). Continuous: *Age squared*. Gallup: Dummies: *Cannot afford healthcare*, *Female*, *Ever diagnosed with depression*. Likert scales: *Learn something every day*, *City/area is perfect* and *Receives Health encouragement* (1 – strongly disagree, 5 – strongly agree), *Economy in this country* (1 – poor to 4 – Excellent), *Health limits activities in the last month* (0 to 30 days). Continuous: *Age*, *age squared*, *Log HH income*, *Physical health index*.

Table A4: Correlations between the Permutation Importance ranks in different algorithms.

	OLS *vs.* GB	OLS *vs.* RF	GB *vs.* RF
SOEP	0.70	0.58	0.79
UKHLS	0.75	0.67	0.86
Gallup	0.86	0.69	0.82

Notes: The correlation figures refer to the Top-100 variables (using the OLS ranking). These are Spearman rank correlations.

Table A5: Permutation Importance (PI) and Pseudo Partial Effect (PPE) in OLS, RF and GB on the Restricted Set of variables: the 10 most-important variables.

	OLS			Random forest			Gradient boosting		
	Variable name	PI	PPE	Variable name	PI	PPE	Variable name	PI	PPE
				Panel A: SOEP					
1	Age and age-squared	.10	-1.70	Adjusted Income	.13	.27	Adjusted Income	.14	.46
2	Adjusted Income	.10	.26	Age and age-squared	.12	-.14	Age and age-squared	.13	-.18
3	Number of doctor visits	.08	-.14	Number of doctor visits	.11	-.28	Number of doctor visits	.12	-.63
4	Marital Status - Single	.07	-.40	Disability Status	.04	-.40	Disability Status	.03	-.45
5	N of children in HH	.06	.30	N of children in HH	.03	.07	Working hours	.02	-.29
6	Disability Status	.04	-.52	N of people in HH	.03	.02	N of years of education	.02	.17
7	N of people in HH	.03	-.17	N of years of education	.02	.07	N of children in the HH	.02	.08
8	N of years of education	.03	.11	House Ownership: Owner	.02	.12	N of people in HH	.02	-.16
9	Marital Status – Divorced	.02	-.38	Working hours	.01	.04	Marital Status – Single	.02	-.19
10	Marital Status - Separated	.02	-.74	BMI	.01	-.02	Marital Status - Separated	.01	-.53
				Panel B: UKHLS					
1	Health limits activities: a lot	.024	-.670	Age	.040	.052	LT sick or disabled (empl.)	.018	-.587
2	Single	.020	-.336	HH income	.015	.161	Age	.015	.052
3	LT sick or disabled (empl.)	.017	-.797	Health limits activities: a lot	.014	-.377	Health limits activities: a lot	.012	-.377
4	Age	.018	.015	Not disabled (health)	.014	.215	Not disabled (health)	.010	.215
5	Health limits activities: a bit	.014	-.327	Health limits activities: a bit	.012	-.226	Renting house	.007	-.106
6	Not disabled (health)	.011	.240	LT sick or disabled (empl.)	.011	-.587	Health limits activities: a bit	.007	-.226
7	Retired	.010	.235	Unemployed	.006	-.193	HH income	.006	.161
8	Renting house	.008	-.208	Renting house	.005	-.106	Unemployed	.006	-.193
9	Unemployed	.008	-.343	Single	.005	-.136	Retired	.005	.099
10	HH income	.008	.083	Retired	.003	.099	Single	.003	-.136
				Panel C: Gallup					
1	Health limits activities	.064	.84	HH income	.062	.48	HH income	.067	.48
2	HH income	.049	.30	Health limits activities	.057	.69	Health limits activities	.054	.71
3	Post-graduate education	.026	.58	Age and age-squared	.046	.43	Age and age-squared	.041	.44
4	Married	.013	.33	Married	.013	.26	Married	.013	.27
5	College Graduate	.010	.37	Female	.010	.23	Female	.013	.29
6	Female	.010	.29	Post-graduate education	.008	.43	Post-graduate education	.008	.34
7	Age and age-squared	.008	.24	Body Mass Index	.005	.29	Body Mass Index	.005	-.12
8	Hispanic	.003	.28	Working Hours Missing	.005	-.12	Hispanic	.003	.15
9	Atheist	.003	-.19	Hispanic	.003	.06	Black	.003	.10
10	High school graduate	.003	.17	Asian	.003	.02	Working Hours Missing	.003	-.06

Notes: The total set of variables available in the restricted set appears in Table A1.

Table A6: Permutation Importance (PI) and Pseudo Partial Effect (PPE) in OLS, RF and GB for positive and negative affect: the top 10 most-important variables (using 2013 Gallup data with the Extended Set of variables).

	OLS			Random forest			Gradient boosting		
	Variable name	PI	PPE	Variable name	PI	PPE	Variable name	PI	PPE
				Panel A: Positive affect					
1	Age	.14	-.26	Physical health index	.07	.42	Physical health index	.16	.62
2	Age squared	.09	-.26	Learn something every day	.06	.43	Learn something every day	.05	.49
3	Physical health index	.09	.66	Not treated with respect	.03	-1.39	Not treated with respect	.03	-1.13
4	Learn something every day	.05	.82	Health encouragement	.02	.13	Health encouragement	.02	.14
5	Not treated with respect	.03	-1.52	Diagnosed w. depression	.01	.27	BMI	.01	.02
6	Health encouragement	.02	.23	City/area is perfect	.00	.17	Diagnosed w. depression	.01	.34
7	In workforce	.01	.44	Health limits activities	.00	-.01	Has any health problems	.01	-.26
8	Diagnosed w. depression	.01	.52	BMI	.00	.09	City/area is perfect	.00	.17
9	Not working	.00	-.32	Age squared	.00	-.11	Health limits activities	.00	.21
10	Tuesday	.00	-.33	Age	.00	-.11	Female	.00	.20
				Panel B: Negative affect					
1	Physical health index	.26	-.11	Physical health index	.31	-.15	Physical health index	.50	-.18
2	Not treated with respect	.03	.16	Not treated with respect	.04	.17	BMI	.04	-.02
3	Diagnosed w. depression	.02	-.09	BMI	.03	-.01	Not treated with respect	.03	.15
4	Age squared	.01	-.03	Diagnosed w. depression	.02	-.07	Has any health problems	.02	.06
5	BMI	.01	-.03	Health limits activities	.01	-.02	Diagnosed w. depression	.02	-.07
6	Has any health problems	.01	.04	Has any health problems	.01	.02	Health limits activities	.02	-.06
7	Cannot afford healthcare	.01	-.05	Cannot afford healthcare	.01	-.04	Had a cold yesterday	.01	.07
8	Wednesday	.00	.05	City/area is perfect	.00	-.02	Cannot afford healthcare	.01	-.04
9	Neck or backpain	.00	-.03	Neck or backpain	.00	-.02	Headache yesterday	.00	.02
10	Time Zone E	.00	.03	Age	.00	-.04	City/area is perfect	.00	-.02

Notes: The following variables are shown. Dummies: *Cannot afford healthcare*, *Female*, *Ever diagnosed with depression*, *Not treated with respect*, *In workforce*, *Has any health problems*, *Tuesday*, *Wednesday*, *Neck or backpain*, *Time Zone E*. Likert scales: *Learn something every day*, *City/area is perfect*, *Receives Health encouragement* (1 – strongly disagree, 5 – strongly agree), *Economy in this country* (1 – poor to 4 – Excellent), *Health limits activities in the last month* (0 to 30 days). Continuous: *Age*, *age squared*, *Log HH income*, *Physical health index*.

Table A7: Permutation Importance (PI) and Pseudo Partial Effect (PPE) of OLS, RF and GB for positive and negative affect of the top 10 most-important variables (using 2013 SOEP data with the Extended Set of variables).

	OLS			Random forest			Gradient boosting		
	Variable name	PI	PPE	Variable name	PI	PPE	Variable name	PI	PPE
				Panel A: Positive affect					
1	Partner in HH	.03	.21	Partner in HH	.03	.17	Partner in HH	.03	.17
2	Worry a lot	.02	-.05	Health limits social life	.02	.03	Worry a lot	.02	-.03
3	Health limits social life	.02	.08	Number of close friends	.01	.07	Health limits social life	.01	.05
4	Deal well with stress	.01	.04	Worry a lot	.01	-.02	Number of close friends	.01	.08
5	Excursions/short trips	.01	-.07	Deal well with stress	.01	.04	Deal well with stress	.01	.04
6	Number of close friends	.01	.04	Excursions/short trips	.01	-.03	Excursions/short trips	.01	-.06
7	Last Word Fin. Decisions-NA	.01	-.08	HH income	.00	.04	Hours of childcare per day	.00	.01
8	Importance: to help others	.01	-.07	Attend cinema/concerts	.00	-.04	Use of social networks	.00	-.06
9	Health limits daily life: a lot	.00	-.12	Am sociable	.00	.01	Importance to help others	.00	-.05
10	Psychiatric problems	.00	-.16	Visit neighbours/friends	.00	-.01	Personal patience	.00	.04
				Panel B: Negative affect					
1	Worry a lot	.11	.13	Worry a lot	.13	.03	Worry a lot	.12	.05
2	Health limits social life	.04	-.12	Health limits social life	.04	-.08	Health limits social life	.04	-.12
3	Female	.03	.20	Deal well with stress	.03	-.03	Female	.02	.16
4	Deal well with stress	.02	-.06	Female	.02	.15	Deal well with stress	.02	-.03
5	Hours of sleep	.01	-.10	Psychiatric problems	.01	.18	Number of doctor visits	.01	.08
6	Health limits daily life: a lot	.01	.18	Number of doctor visits	.01	.05	Hours of sleep	.01	-.07
7	Psychiatric problems	.01	.25	Hours of sleep	.01	-.04	Psychiatric problems	.01	.22
8	Personal patience	.01	-.06	Annual pension	.01	.00	Personal Patience	.01	-.05
9	Health affects tiring tasks	.01	.15	Personal Patience	.01	-.03	Annual pension	.01	.00
10	Number of doctor visits	.01	.03	Physical pain last 4 weeks	.00	-.03	Health limits daily life: a lot	.00	.12

Notes: The following variables are shown.: Dummies: *Health limits daily life a lot*, *Health limits daily life a bit*, *Partner in HH*, *No monthly savings*, *Not unemployed*, *No emergency reserves*, *Last word in financial decisions-NA*, *Psychiatric problems*, *Female*, and *No annual holiday trip*. Likert scales: *Limited socially due to health* (1 – always to 5 – never), *Worries a lot*, *Importance: To help others* (1 – Very Important to 4 – Not important), *Deals well with stress* (1 – not at all to 7 – totally agree), *Personal patience* (0 – very bad to 10 – very good), *House needs repair* (1 – in good condition, 3 – needs major renovation), *Attend cinema/concerts* (1 – Daily to 4 - Infrequent), *Am Sociable* (1 to 7), *Visit neighbours/friends* (1 – Daily to 5 - Never), *Use of social networks* (1 – Daily to 5 - Never), *Health affects tiring tasks* (1 – A lot to 3 - Not at all), and *Physical pain last 4 weeks* (1 – Always to 5 - Never). Continuous: *Log HH income*, *Hours of sleep*, *Number of doctor visits*, *Risk tolerance*, *Unemployment benefit*, *Excursions/short trips*, *Number of close friends*, *Hours of childcare per day*, *Annual pension*.

Table A8: Permutation Importance (PI) of OLS, RF and GB for levels of wellbeing of the 10 most-important variables (using pooled UKHLS data with the Restricted Set of variables). For each covariate, the models include the average value and the annual deviation from that average.

	OLS		Random forest		Gradient boosting	
	Variable name	PI	Variable name	PI	Variable name	PI
1	Health limits activities: a lot (avg.)	.041	Age (avg.)	.025	Age (avg.)	.026
2	Not disabled (health) (avg.)	.020	Not disabled (health) (avg.)	.020	Not disabled (health) (avg.)	.022
3	Married (avg.)	.019	Health limits activities: a lot (avg.)	.018	Health limits activities: a lot (avg.)	.021
4	Health limits activities: a bit (avg.)	.017	Health limits activities: a bit (avg.)	.014	Health limits activities: a bit (avg.)	.014
5	LT sick or disabled (empl.) (avg.)	.015	LT sick or disabled (empl.) (avg.)	.011	HH income (avg.)	.012
6	Age (avg.)	.013	HH income (avg.)	.009	LT sick or disabled (empl.) (avg.)	.012
7	Retired (avg.)	.012	Married (avg.)	.006	Married (avg.)	.009
8	HH income (avg.)	.010	Retired (avg.)	.005	Retired (avg.)	.006
9	Unemployed (avg.)	.007	Unemployed (avg.)	.004	Unemployed (avg.)	.005
10	Rents the house/flat	.005	Health limits activities: a bit	.003	Health limits activities: a lot	.004

Notes: All covariates apart from month, ethnicity and sex are split into individual means and deviation from the mean. Individual averages are denoted by *(avg.)*; variables without additional notes are the deviations from the individual means.

Table A9: Permutation Importance (PI) of OLS, RF and GB for deviations from the average wellbeing and individual level of wellbeing of the 10 most-important variables (using pooled SOEP data with the Restricted Set of variables). For each covariate, the models include the average value and the annual deviation from that average.

	OLS		Random forest		Gradient boosting	
	Variable name	PI	Variable name	PI	Variable name	PI
1	Age (avg.)	.082	Age (avg.)	.126	Age (avg.)	.124
2	Number of doctor visits (avg.)	.039	Adjusted Income (avg.)	.059	Adjusted Income (avg.)	.049
3	Adjusted Income (avg.)	.039	Number of doctor visits (avg.)	.041	Number of doctor visits (avg.)	.042
4	No. of children in the hh (avg.)	.025	Not disabled (health) (avg.)	.021	Not disabled (health) (avg.)	.016
5	Not disabled (health) (avg.)	.016	No. of people in hh (avg)	.014	Age	.010
6	Single (avg.)	.016	No. of children in hh (avg.)	.011	No. of people in hh (avg.)	.009
7	Divorced (avg.)	.007	House Owner	.009	No. of children in hh (avg.)	.008
8	No. of people in hh (avg.)	.006	Age	.008	Number of doctor visits	.007
9	Number of doctor visits	.005	Number of doctor visits	.005	Single	.006
10	House Owner	.005	Number of years of education	.005	House Owner	.006

Notes: All covariates apart from month, ethnicity and sex are split into individual means and deviation from the mean. Individual averages are denoted by *(avg.)*; variables without additional notes are the deviations from the individual means.

Figure A1: Histograms of life satisfaction for SOEP, UKHLS and Gallup data.

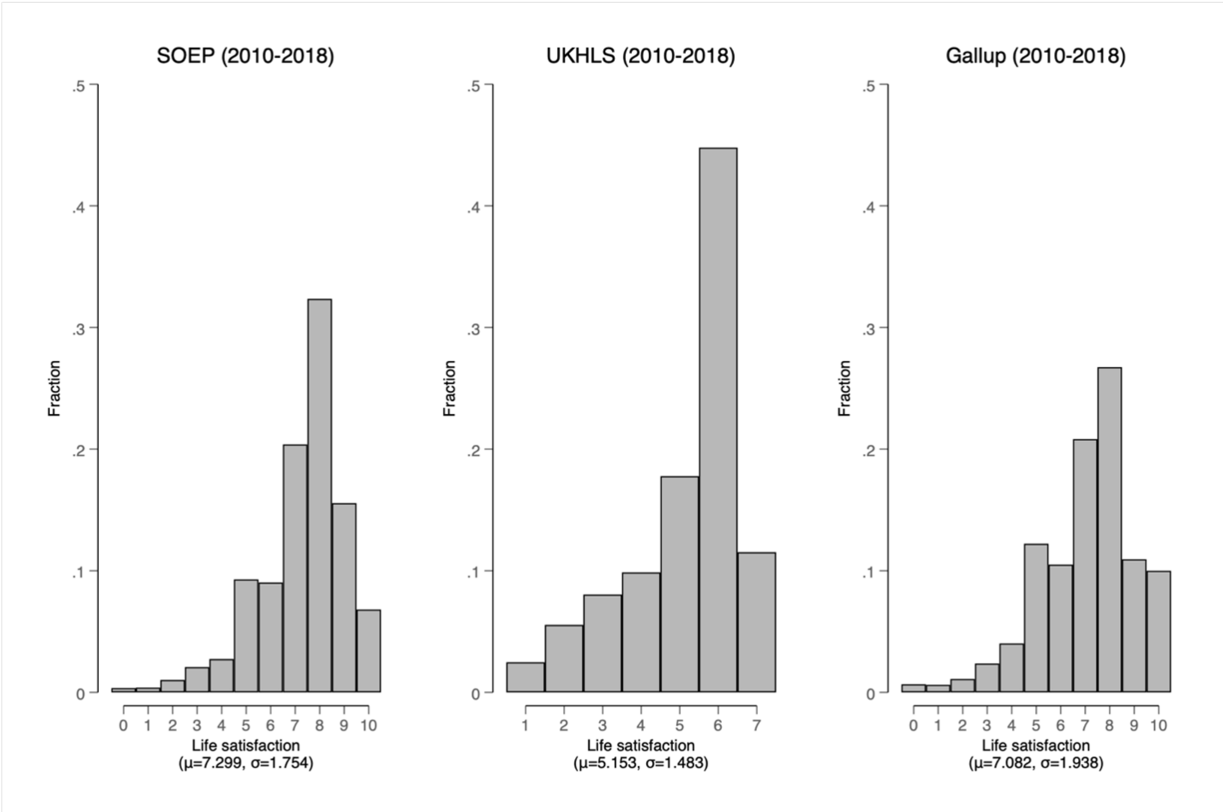


Figure A2: The R-squared from OLS, GB and RF on the Restricted Set of variables. The R-squareds are calculated from the training data and are not representative of out-of-sample performance.

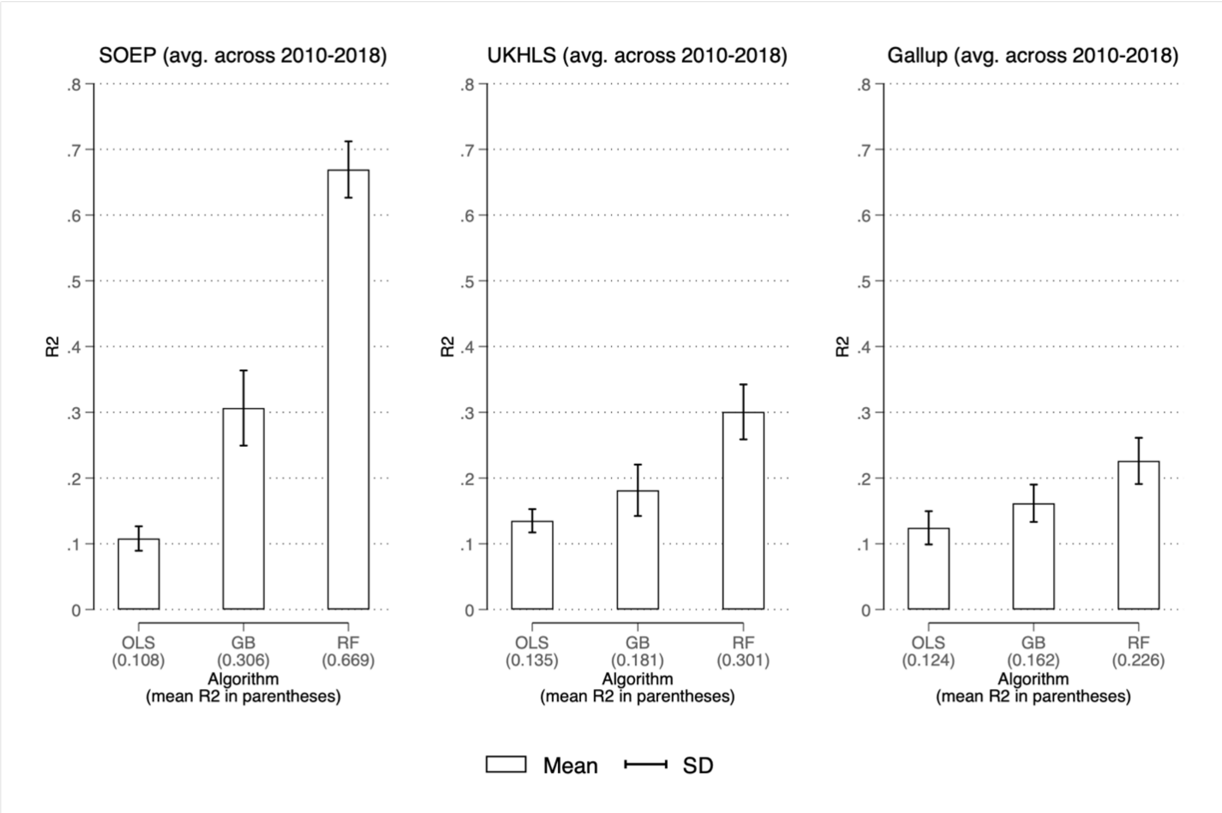


Figure A3: The R-squareds from OLS, LASSO, GB, RF, and mean on the Extended Set of variables. The R-squareds are calculated from the training data and are not representative of out-of-sample performance.

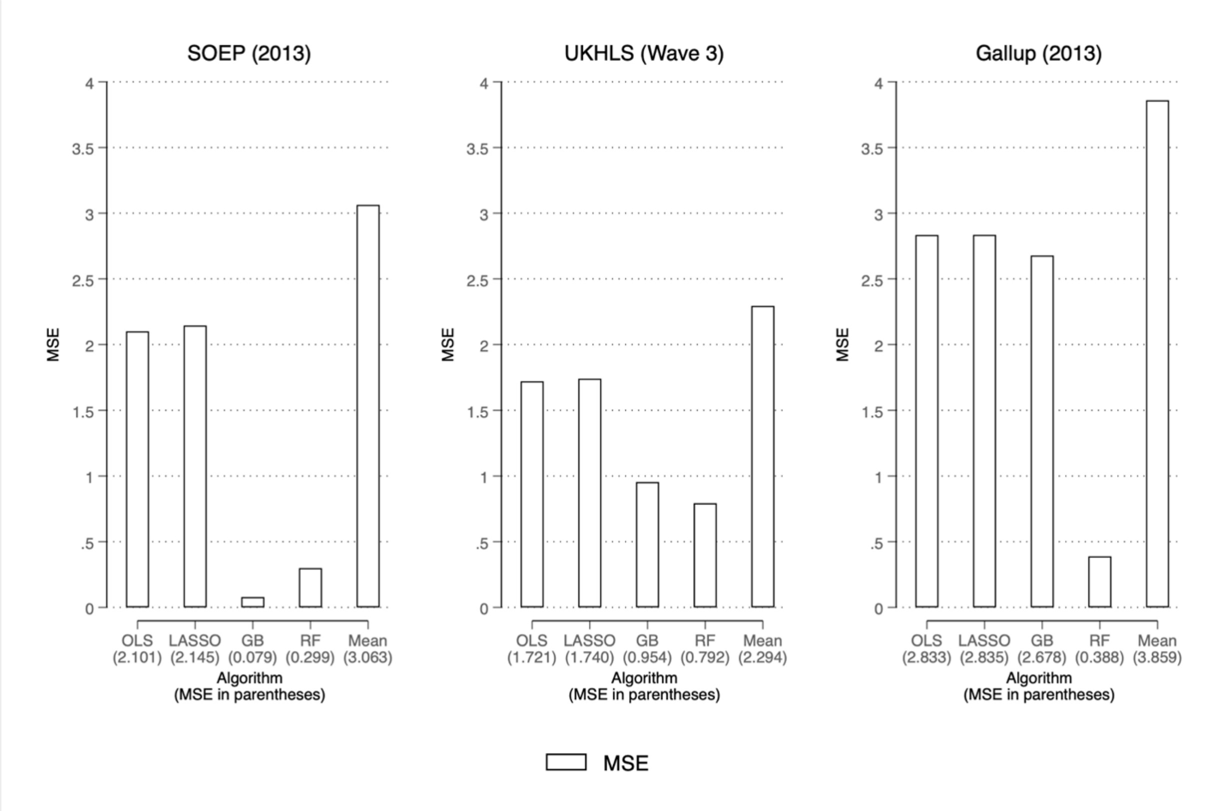
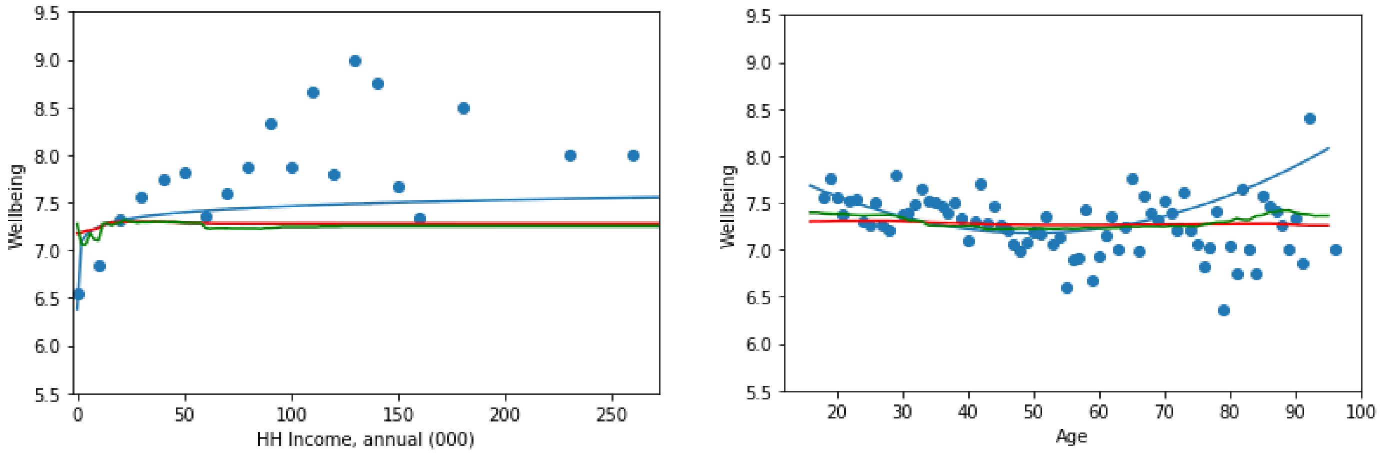
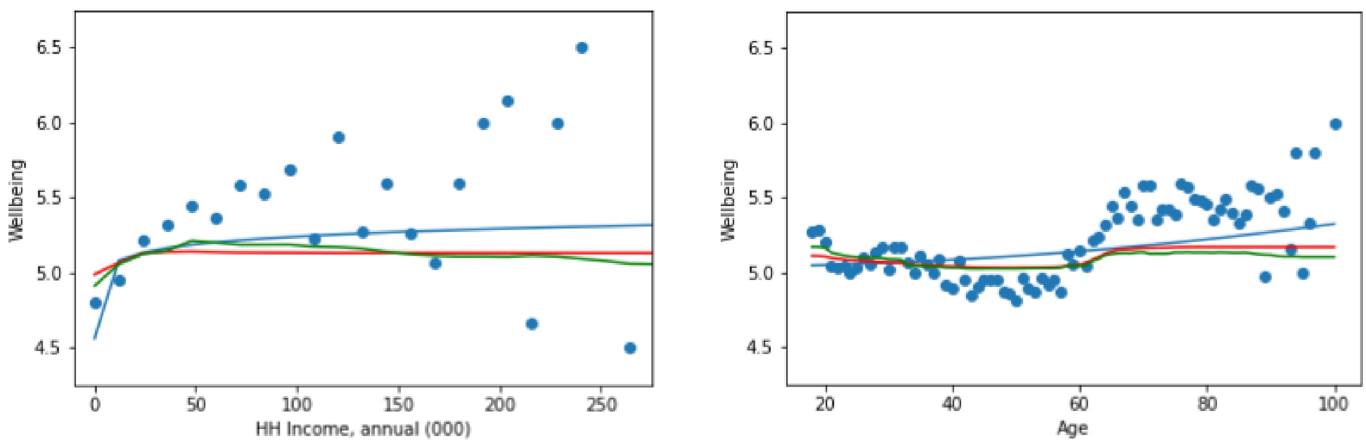


Figure A4: The mean effects of age and household income on wellbeing in the Extended Set of variables.

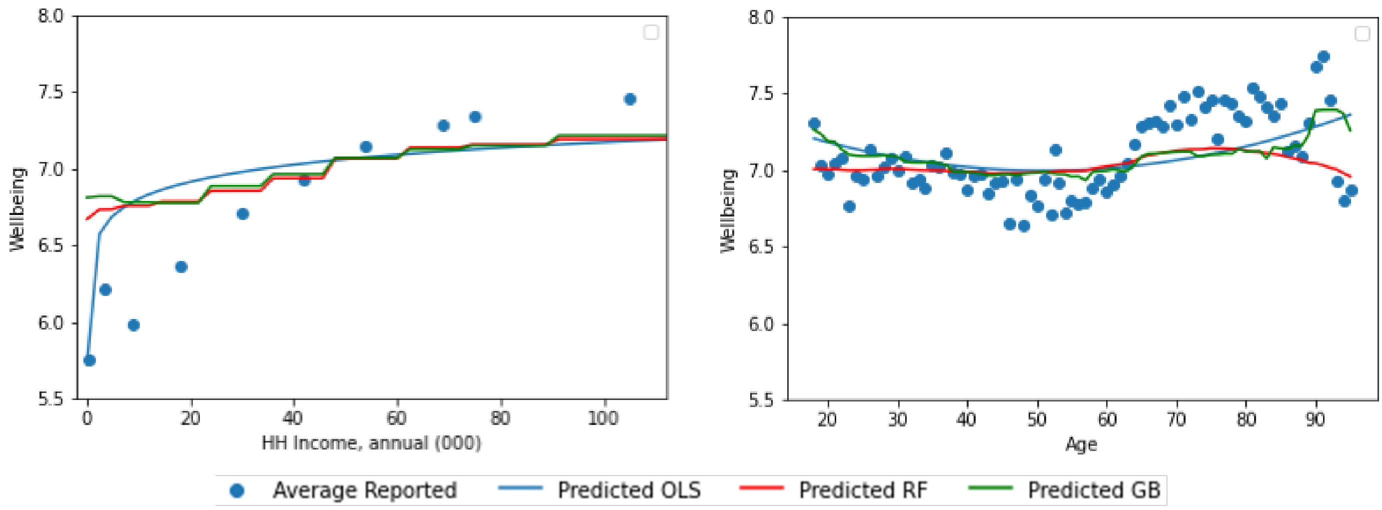
Panel A: SOEP



Panel B: UKHLS



Panel C: Gallup



Notes: For the UKHLS and the SOEP annual income is constrained to be less than or equal to a figure of 250 000 in the local currency. This covers over 99.9% of the income distribution in both countries. In SOEP and UKHLS, incomes are recorded as a continuous variable and equivalence-scale adjusted HH income is used for the analysis. Income data in Gallup is collected in income bands, and household size data was not collected in 2013. We here thus use non-adjusted HH income data.

Figure A5: The R-squared from OLS, LASSO, GB and RF when modelling positive and negative affect using 2013 Gallup and 2013 SOEP data with the Extended Set of variables. The R-squareds are calculated from unseen 'testing data'.

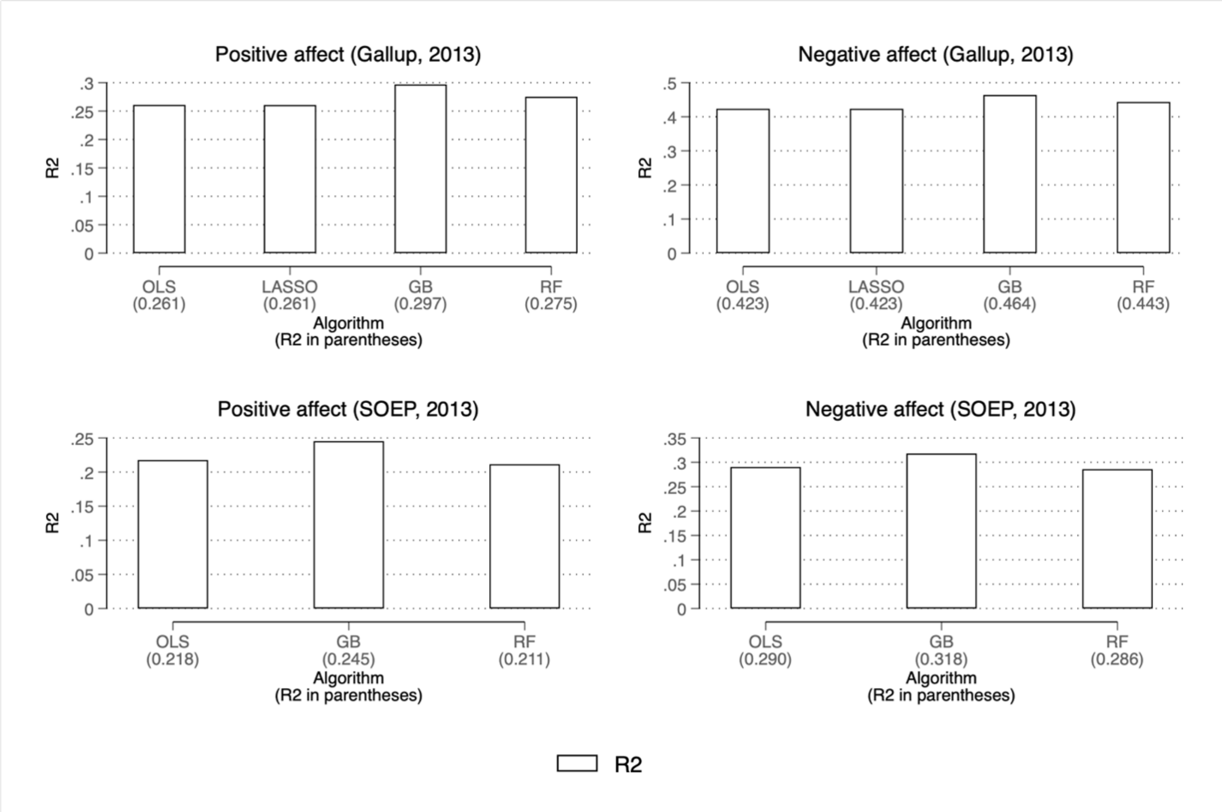
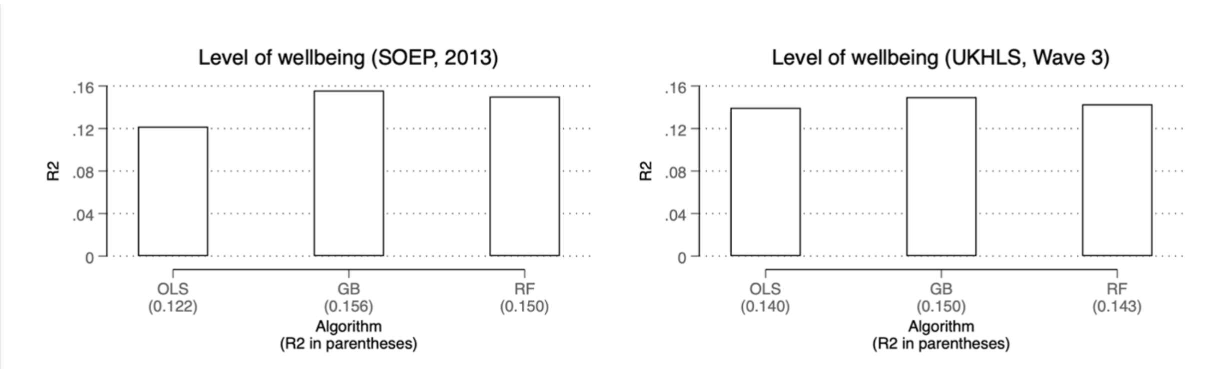


Figure A6: The R-squared from OLS, GB and RF when modelling the level of wellbeing with Mundlak terms using 2013 SOEP and Wave 3 UKHLS data with the Restricted Set of variables. The R-squareds are calculated from unseen 'testing data'.



CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1862	Fabrizio Leone Rocco Macchiavello Tristan Reed	Market size, markups and international price dispersion in the cement industry
1861	Elias Einio Josh Feng Xavier Jaravel	Social push and the direction of innovation
1860	Xiao Chen Hanwei Huang Jiandong Ju Ruoyan Sun Jialiang Zhang	Endogenous cross-region human mobility and pandemics
1859	Xavier Jaravel Danial Lashkari	Nonparametric measurement of long-run growth in consumer welfare
1858	Leonardo Bursztyn Jonathan Kolstad Aakaash Rao Pietro Tebaldi Noam Yuchtman	Political adverse selection
1857	Attila Lindner Balázs Muraközy Balázs Reizer Ragnhild Schreiner	Firm-level technological change and skill demand
1856	Jeremiah Dittmar Ralf R. Meisenzahl	The university, invention and industry: evidence from German history
1855	Donna Brown Jonathan Wadsworth	Accidents will happen: (de)regulation of health and safety legislation, workplace accidents and self employment
1854	Fabrizio Leone	Foreign ownership and robot adoption

1853	Andrew E. Clark Conchita D'Ambrosio Niccoló Gentile Alexandre Tkatchenko	What makes a satisfying life? Prediction and interpretation with machine-learning algorithms
1852	Xiang Ding Teresa C. Fort Stephen J. Redding Peter K. Schott	Structural change within versus across firms: evidence from the United States
1851	Christos Genakos Eleni Kyrkopoulou	Social policy gone bad educationally: unintended peer effects from transferred students
1850	Antonin Bergeaud Cyril Verluise	A new dataset to study a century of innovation in Europe and in the US
1849	Jo Blanden Mattias Doepke Jan Stuhler	Education inequality
1848	Martina Manara Tanner Regan	Ask a local: improving the public pricing of land titles in urban Tanzania
1847	Rebecca Freeman Kalina Manova Thomas Prayer Thomas Sampson	Unravelling deep integration: UK trade in the wake of Brexit
1846	Nicholas Bloom Scott W. Ohlmacher Cristina J. Tello-Trillo Melanie Wallskog	Pay, productivity and management
1845	Martin Gaynor Adam Sacarny Raffaella Sadun Chad Syverson Shruthi Venkatesh	The anatomy of a hospital system merger: the patient did not respond well to treatment

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE