Working
Paper
Series

doi.org/10.5287/ora-2ajz1epnk

2301

Machine Learning in the Prediction of Human Wellbeing

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

April 2023

cite this paper

Oparina, E., Kaiser, C., Gentile, N., et al. (2023). *Machine Learning in the Prediction of Human Wellbeing*. University of Oxford Wellbeing Research Centre Working Paper 2301. doi.org/10.5287/ora-2ajz1epnk

Machine Learning in the Prediction of Human Wellbeing

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

April 6, 2023

Abstract

Subjective wellbeing data are increasingly used across the social sciences. Yet, our ability to model wellbeing is severely limited. In response, we here use tree-based Machine Learning (ML) algorithms to provide a better understanding of respondents' self-reported wellbeing. We analyse representative samples of more than one million respondents from Germany, the UK, and the United States, using the data between 2010 and 2018. In terms of predictive power, our ML approaches perform better than traditional ordinary least squares (OLS) regressions. 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, personality traits, 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.

KEYWORDS: Subjective wellbeing, prediction methods, machine learning.

*These authors are joint first authors. The displayed order of these authors is random and determined by the AEA tool, confirmation code $oJsh\_ZMZJwhH$. Correspondence to casparkaiser@gmail.com. Author affiliations: Ekaterina Oparina: LSE; Niccolò Gentile, Conchita D'Ambrosio and Alexandre Tkatchenko: University of Luxembourg; Caspar Kaiser and Jan-Emmanuel De Neve: University of Oxford; Andrew E. Clark: PSE - CNRS.
We thank Filippo Volpin for excellent research assistance and Sid Bhushan for early discussions on the topic. We are grateful to Martin Huber, Daniel Kahneman, Christian Krekel, Andrew Oswald, Nattavudh Powdthavee, Sorawoot Srisuma and seminar participants at the LSE Wellbeing Seminar and Loughborough University, as well as the STATEC Wellbeing 2022 Conference, the 2022 IAAEU Workshop on Health and the Labour Market, and the Oxford Wellbeing & Policy Conference for comments and suggestions. This work was supported by the ERC [grant agreement n. 856455], the ESRC [grant number ES/T014431/1]; and the Institute for Advanced Studies, University of Luxembourg [grant DSEWELL]. We thank The Gallup Organization for providing access to their data for this research project.

Statement of Relevance

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 respondents' wellbeing, limiting our ability to predict it. We utilise tree-based Machine Learning (ML) algorithms to improve predictions. First, we find that these ML algorithms indeed yield better predictive performance than standard methods, and establish an upper bound on the predictability of evaluative wellbeing with survey data. Second, we use ML to identify the key drivers of evaluative wellbeing. We show that the variables emphasised in the earlier intuition- and theory-based literature also appear in ML analyses. Third, we illustrate how ML can be an impartial arbiter in questions about functional forms, including the existence of satiation points in the effects of income and the U-shaped relationship between age and wellbeing.

1 Introduction

A substantial and interdisciplinary literature on the correlates and determinants of subjective wellbeing has emerged over the past 50 years (Diener et al., 2018). In parallel, international organisations (OECD, 2020) and national governments (ONS, 2021) have turned to subjective wellbeing data as a key tool for policy analysis.

Subjective wellbeing data has been extensively validated, and has been shown to correlate well with objective outcomes (Diener et al., 2013). Generally, such data also do better in predicting future behaviour than many other standard social science variables (Benjamin et al., 2014; Charpentier et al., 2016; Kaiser and Oswald, 2022). On that basis, we might expect that the answers respondents provide to questions about their wellbeing to be well-predicted using standard regression equations. However, to the contrary, and despite the widespread use of these scores, our current ability to model wellbeing is surprisingly limited. Standard approaches, where variables are selected based on intuition or theory, explain relatively little of the variation. Individual-level models typically yield R-squared figures of no more than 15% (Layard et al. 2014 is one typical example). Yet, especially in economics and psychology, the prediction and explanation of individual wellbeing are one of the discipline's core tasks. Our limited ability to make predictions about wellbeing would thus seem to be a major shortcoming.

Notwithstanding the above, the existing literature has largely reached a consensus on the main sources of wellbeing. These include good health (Lucas, 2007; Oswald and Powdthavee, 2008), unemployment (Kassenboehmer and Haisken-DeNew, 2009; Lucas et al., 2004), social relations (Blanchflower and Oswald, 2004; Rohrer et al., 2018), as well as personality traits (Anglim et al., 2020; Boyce, 2010; Boyce and Wood, 2011). For wider overviews of the interdisciplinary literature, see Clark (2018), Dolan et al. (2008), Kong et al. (2019), Nikolova and Graham (2022), and Ryan (2001).

Despite this broad consensus, some questions are still actively debated. Two examples are whether wellbeing is U-shaped in age (e.g. Frijters and Beatton 2012 *versus* Cheng et al. 2017;

Wunder et al. 2013), and whether income beyond a 'satiation point' yields no further increase in wellbeing (e.g. Jebb et al. 2018; Kahneman and Deaton 2010 *versus* Killingsworth 2021; Stevenson and Wolfers 2013). The answers researchers provide to these questions can depend on their prior beliefs and their subsequent modelling choices. Machine learning algorithms, on the contrary, are indifferent about the conclusions they reach: they have no 'axe to grind'. The use of ML would therefore seem to be particularly apt for the resolution of controversial academic debates in a disinterested manner. We will below investigate the effects of income and age on subjective wellbeing to illustrate the more-general potential of ML as an impartial arbiter in this kind of scientific disputes.

We pose three research questions:

RQ1: Do ML algorithms predict wellbeing substantially better than conventional linear models, and what is the upper limit on our ability to predict wellbeing based on survey data?

RQ2: Are the variables that ML algorithms identify as important in the prediction of wellbeing the same as those emphasised in the existing literature?

RQ3: Can ML help to resolve ongoing debates about the form of the relationships between wellbeing, income, and age?

We will apply random forests (Breiman, 2001; Hastie et al., 2009), gradient boosting (Friedman, 2001; Natekin and Knoll, 2013), and penalised regressions (Tibshirani, 1996). Random forests and gradient boosting are tree-based algorithms that have been shown to perform well with standard social-science data that is organised in rows and columns (*i.e.* 'tabular data', as opposed to text or images: see Shwartz-Ziv and Armon 2022).[1] Penalised regressions are a convenient tool for analyses that involve a large number of covariates, as will be the case in some of our specifications (Tibshirani, 1996).

To the best of our knowledge, this paper represents the first systematic attempt to evaluate the (dis-) advantages of using ML for the analysis of self-reported wellbeing at a global scale with survey data. Earlier work focused on single drivers of wellbeing like age (Kaiser et al., 2022) and relatively limited country-, year- or age-specific samples (Margolis et al., 2021; Prati, 2022), or used objective biomarkers in small-N studies (Dukart et al., 2021; Just et al., 2017).

We focus on life satisfaction as a key cognitive-evaluative measure of wellbeing (OECD, 2013). In supplementary analyses, we also analyse more-affective measures of subjective wellbeing: both positive affect (as measured by e.g. the rate of smiling) and negative affect (as measured by e.g. the rate of feeling anger). Our main findings extend to these measures (see Online Appendix A.1). Our analyses are based on three of the largest currently-available datasets that include wellbeing information: the German Socio-Economic Panel (SOEP; yearly N=30,000 with around 400 variables), the UK Household Longitudinal Study (UKHLS; yearly N=40,000 with around

[1]Neural networks often perform poorly on tabular data (Borisov et al., 2022). In preliminary analyses, the performance of feed-forward neural networks was no better than that of OLS. This is why we did not consider them further.

500 variables) and the American Gallup Daily Poll (yearly N=200,000 with around 60 variables).

As is standard in the ML literature – but not so in the social sciences – we exclusively evaluate model performance using data that these models have not previously seen (the 'test' set). Model parameters are determined via a separate 'training' set. Model evaluation then refers to the quality of the out-of-sample prediction, as the test and training sets do not overlap. It is important to underline that there is no automatic improvement in model performance by the addition of more variables: any advantage of ML models over OLS cannot therefore be attributed to mechanical overfitting.[2]

The performance of black-box ML methods can be used to benchmark the highest possible predictive ability provided by a given set of characteristics (see (Fudenberg et al., 2022)). Regarding **RQ1**, we find that ML algorithms predict better than standard linear models. The size of this improvement is moderate in absolute terms, but substantial when compared to the predictive power of key variables such as health. Although the use of ML as a more-flexible modelling framework does provide some gains, much more substantial improvement comes from adding other relevant individual characteristics to the model. Model performance, as judged by the R-squared on the unseen 'test set' data, roughly doubles for both the OLS and ML approaches when the set of variables is expanded from a standard set (we call this the 'Restricted Set') to all of the available data (the 'Extended Set'). Independently of the type of algorithm, an R-squared of around 0.30 appears to be the feasible maximum for individual-level models given the available data. This is approximately half of all the predictable wellbeing variance, as determined by the test-retest correlations obtained in earlier work (Krueger and Schkade, 2008).

For **RQ2**, our data-driven ML results largely confirm the findings in the conventional literature. Variables reflecting respondents' social connections, health, and material conditions are consistently the most predictive of their wellbeing. The extended-set analysis identifies the following additional variables as being particularly important in predicting self-assessed wellbeing scores: personality traits, relationship quality, additional measures of health, and perceptions of the local area. Many of these are not consistently included in wellbeing analyses. There is a substantial correlation in variable-importance rankings across algorithms ($\rho = 0.58$ to $\rho = 0.83$). Hence, ML approaches and OLS are largely in agreement regarding what matters for wellbeing.

Last, with respect to **RQ3**, we find support for a U-shaped relationship between age and wellbeing in all three datasets. For income, there is evidence of satiation for equivalent household incomes of over 40,000 GBP (50,000 EUR) in the UK (German) data. We do not find evidence of satiation in the US, which, we suggest, may reflect the way in which income is measured in the Gallup data.

[2]Overfitting occurs when a statistical model is too closely aligned to a particular set of data, and as a result it may fail to perform accurately against unseen data, defeating its purpose.

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 (annual N=115,192 to N=351,875 after removing incomplete information). Self-reported evaluative wellbeing is measured by the Cantril Ladder of Life (Cantril, 1965), which is recorded on a scale from 0 to 10.[3] The SOEP and UKHLS are respectively representative of the German and UK adult population, with interviews conducted in person (SOEP, 2021; UKHLS, 2021). To enable comparison with the Gallup data, we consider the survey period between 2010 and 2018 (SOEP annual N=26,089 to N=32,333; UKHLS annual N=29,605 to N=40,679). Life satisfaction is measured on a scale from 0 to 10 (SOEP) or 0 to 7 (UKHLS). The descriptive statistics and histograms of each wellbeing measure appear in Online 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 estimate **Ordinary Least Squares (OLS)** regressions. The OLS estimates are the solution to the problem $\arg\min_b \sum_{i=1}^{N}(x_i'b - s_i)^2$. 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. When using OLS, the researcher implicitly assumes that reported wellbeing is a linear combination of the chosen set of explanatory variables x.

OLS estimates can have large variances when the number of explanatory variables is large, leading to poor predictions. The second algorithm, the **Least Absolute Shrinkage and Selection Operator (LASSO)**, tackles this issue by adding a penalty for the sum of coefficient magnitudes. Specifically, LASSO estimates are the solution to $\arg\min_b \sum_{i=1}^{N}(x_i'b - s_i)^2 + \lambda \sum_{k=1}^{K}|b_k|$. Here, λ is a hyperparameter, the preferred value of which is found using a grid search. LASSO and OLS are equivalent for $\lambda = 0$. The LASSO tends to shrink coefficients on variables with little 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

[3]There is debate over whether these kinds of variables allow for inference about underlying wellbeing (Bond and Lang, 2019; Chen et al., 2022; Kaiser and Vendrik, 2022; Schröder and Yitzhaki, 2017). We remain agnostic about this, and ask which algorithms and models best predict the answers to wellbeing questions, without making claims about how survey responses relate to respondents' underlying feelings.

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 (Hastie et al., 2009): $D_{(L,k)} = \{x|x_j < \tau_{k,j}\}; D_{(R,k)} = \{x|x_j \geq \tau_{k,j}.\}$. 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 residual sum of squares at either side of the split. 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 mean squared error (MSE). If the size of the tree is not limited, the algorithm will overfit the data in the training set. This issue can be tackled by aggregating the predictions from multiple smaller trees. Random forests and gradient boosting are both examples of this strategy (Hastie et al., 2009).

Specifically, **Random Forests**, the third algorithm we consider, average across a large number of trees (set to 1,000 throughout). Each individual tree is grown on a separate bootstrap sample of the original data. At each split, only a random subset of all covariates is considered. Both operations reduce the correlation between trees, thereby reducing the variance of the resulting overall predictions. The size of the variable 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 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 overall predictions are then given by $\hat{s}_{T_1} + \hat{e}_{T_1} = \hat{s}_{T_2}$. This process is repeated *Ntrees* times, producing increasingly accurate predictions of s. Since gradient-boosted collections of trees overfit in the training set 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.

The algorithms are trained on the training set, which here contains 80% of the sample. Each algorithm's performance is on the contrary evaluated only with the unseen test set, which contains the remaining 20% of observations. The optimal hyperparameters are chosen via 4-fold cross validation on the training-set data. The chosen hyperparameters can be found in Online Appendix Table A1. These algorithms are implemented in the scikit-learn library in Python (Pedregosa et al., 2011). We evaluate the stability of our results over time, where feasible, by training each algorithm separately on each wave within each survey.

[4] We here use a standard implementation of gradient boosting. We also evaluated the performance of extreme gradient boosting (XGBoost; Chen and Guestrin (2016), which yielded only negligible improvements.

2.3 Explanatory variables

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

We first consider a restricted set of variables that are observed in all three datasets. This set includes: 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, body mass index, labour-force status, working hours, home ownership, area of residence, and interview month. A more-detailed description of these variables appears in Online Appendix Table A2. These variables are typical in the conventional literature on subjective wellbeing, and allow us to assess the performance of ML algorithms relative to OLS in a standard estimation setting.

We second consider 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).[5] Our extended sets include all of the available variables in each survey, apart from variables that are direct measures of subjective wellbeing (such as domain satisfaction, happiness and subjective health) and mental health, as well as extraneous meta-data (such as respondents' identification numbers). The resulting Gallup dataset contains 67 variables, and around 450 variables are retained in the SOEP and UKHLS. The variables concern the respondents' family relationships, social life, neighbourhood and residence, incomes and expenditures, attitudes, personality traits and other characteristics. The summaries of the variables in each dataset are presented in Online Appendix Table A3.[6]

Some variables have no predictive power. We therefore use LASSO as a device to select the explanatory variables (Ahrens et al., 2020; Tibshirani, 1996). We have carried out the estimations on both the full and post-LASSO extended sets, with both specifications having similar performance.[7] For simplicity, we only display the results for the approach that performed better in each individual analysis.

2.4 Assessing Model performance

We evaluate model performance using data (the 'test set') that was unseen by the model during the training stage. In this way, we can evaluate each model's out-of-sample prediction quality. We first compare the performance of the OLS model to those of several ML algorithms, using the restricted set of variables that are standard in the wellbeing literature; we then carry out an analogous comparison for the extended set of individual characteristics.

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

[6] The full list of variables is available here: https://data.mendeley.com/datasets/pgrvssrwy6. We exclude variables with more than 50% missing values. 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 sets of dummy variables, one for each category. Creating these dummies and removing perfectly collinear variables yields 210, 542, and 957 effective explanatory variables in the Gallup, SOEP and UKHLS datasets respectively.

[7] Applying LASSO to the restricted variable set produced a similar performance to OLS, with the optimal λ being close to 0.

As we evaluate model performance by out-of-sample prediction, any performance improvements from the more-flexible ML framework or the extended set of variables is not a mechanical result of overfitting. The improvements we observe, therefore, genuinely indicate that the literature's standard regression models do not utilise all of the relevant information contained in social surveys.

We should not of course expect all of the observed variance in reported wellbeing to be predictable. For example, responses to wellbeing questions can be influenced by random and extraneous factors that are not relevant for global evaluative wellbeing, such as passing moods or social desirability. With this in mind, we should interpret reported wellbeing levels as a combination of a potentially-predictable latent state and a measurement error (see Bertrand and Mullainathan (2001) and Oparina and Srisuma (2022) for a similar approach). Our aim should therefore be to successfully predict the share of variance in reported wellbeing that can be attributed to a respondent's latent state. Following Krueger and Schkade (2008), we approximate this share of the explainable variance by the test-retest correlation in reported wellbeing (see Silk (1977) for a formal derivation). In particular, Krueger and Schkade (2008) find a test-retest correlation of 0.59 for life satisfaction, which is in line with earlier findings on smaller samples (see *e.g.* Kammann and Flett (1983)). We take this as an upper bound for any model's ability to predict wellbeing.

2.5 Assessing variable importance

To answer our second research question, we need to establish the importance of each explanatory variable in predicting 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).[8] 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 the variable's values are permuted in this way. 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 drop in the R-squared, the more important is the variable in predicting wellbeing.

To understand the direction of the 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 then calculate the mean difference in predicted wellbeing. For binary variables (including the dummies calculated from the categorical variables), we predict wellbeing when setting each individual's value to either 0 or 1.

[8]Shapley values are an alternative way of assessing variable importance. We did not calculate Shapley Values here due to their substantial computational complexity (Lundberg et al., 2018; Yang, 2021), and as the pseudo partial effects, discussed below, also identify the direction of the variables' effects.

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.

3 Results

3.1 Model performance

We begin with RQ1: whether ML algorithms substantially outperform OLS (the most-common approach in the existing literature) in predicting wellbeing.

3.1.1 The Restricted Set of explanatory variables

Figure 1 depicts the improvements in each algorithm's performance over OLS. The results here are based on the 'test-set', and can thus be interpreted as an assessment of the models' ability to make out-of-sample predictions. The results in Panel A are for the 'restricted' set of covariates, which only includes variables that are typically used in the literature. We use the R-squared as our primary evaluation metric, to facilitate comparisons with previous analyses.

In Panel A of Figure 1 each algorithm is trained separately for each year between 2010 and 2018. For each year we take the difference between the R-squared from each ML algorithm and the R-squared from the OLS estimation. The figure reports the average differences across these years and their standard deviations. In absolute terms, 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 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 at 0.030, 0.005, and 0.018.[9] ML algorithms thus do outperform linear regressions, and gradient boosting always outperforms random forests. We here focus on predicting wellbeing in the cross-section. Online Appendix A.2 shows substantially the same results when exploiting the panel dimension of the German and UK data. [10]

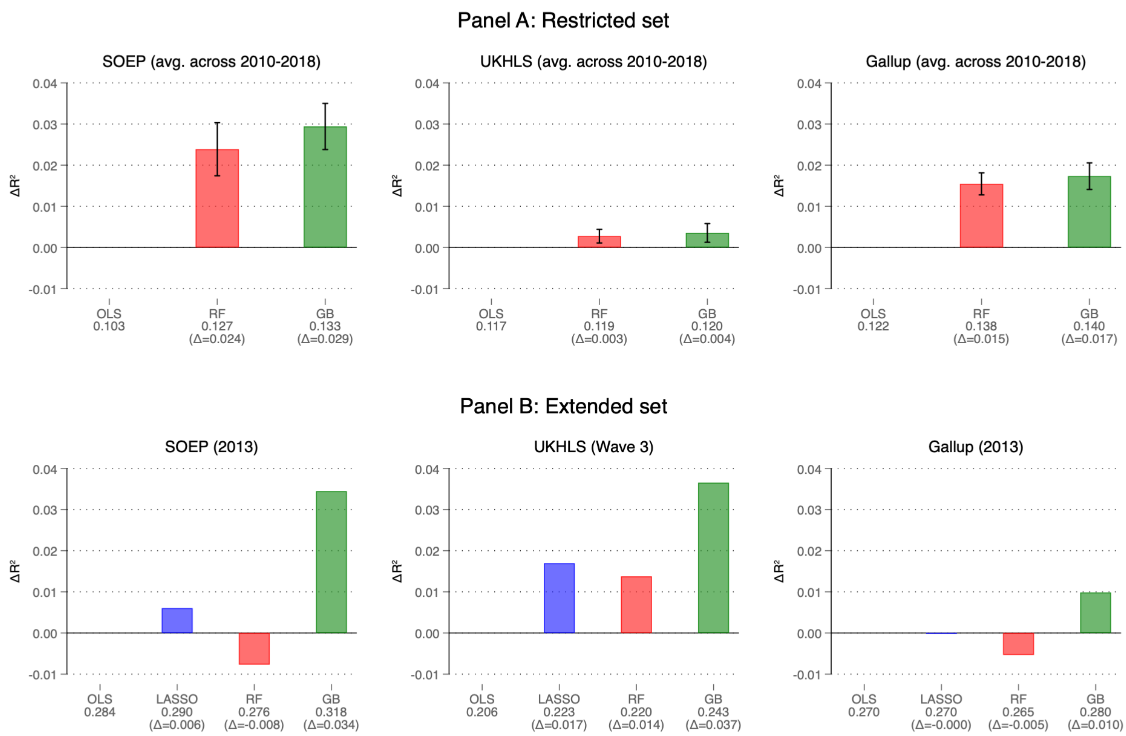
On their own, these gain figures are hard to interpret. We therefore illustrate their size by comparing them to the change in predictive performance when omitting the respondent's health status – a key wellbeing predictor – from the baseline OLS regressions. The first two columns in Panel A of Table 1 list the OLS test-set R-squared figures with and without health, and column 3 the R-squared from gradient boosting. Benchmarking one against the other in column 4, the

[9]These figures refer to performance in the test set, which was not used to train the algorithm. In the training set, the improvement of ML over OLS is larger (see Online Appendix Figure A2). The predictive capacity of the ML algorithms does not therefore seem to be constrained by underfitting. Performance in the training set does not *per se* indicate algorithm quality: a decision tree with as many leaves as individuals in the training set would produce an MSE of 0, but would perform poorly when assessing unseen test data.

[10]There, *levels* of self-reported wellbeing are analysed longitudinally. In further analyses, we have also predicted the individual-level changes in wellbeing from one wave to another. Again, random forests and (especially) gradient boosting outperform OLS by a moderate amount.

improvement in prediction from gradient boosting (our best ML algorithm) is between 15% and 107% of the role of health in predicting wellbeing. Evaluated in this way, the gains from using ML are not negligible.

Figure 1: Differences in out-of-sample performance between OLS and ML.



Notes: Panel A shows the differences in the mean R-squared figures between OLS and GB/RF, using the restricted set of variables across all years (2010-2018). Whiskers show standard deviations. Panel B shows the analogous figures for the extended set of variables, comparing OLS with LASSO, GB and RF. 2013 data are used here. All R-squareds are calculated from the unseen test data. The figures below each bar indicate the absolute out-of-sample R-squared and those in parentheses the differences in the out-of-sample R-squared compared to OLS. Throughout, gradient boosting (GB) yields the best predictive performance.

3.1.2 The Extended Set of explanatory variables

Adding further covariates will (weakly) increase our ability to explain wellbeing in the training set. It is also possible, although far from mechanical, that they will improve our ability to predict wellbeing in the test set. Given the greater flexibility of the ML algorithms, we may expect these latter to benefit more from additional variables than OLS. The extended sets of variables we consider here include all of the variables available in the 2013 waves of the SOEP and Gallup, and Wave 3 of the UKHLS.

Panel B of Figure 1 depicts the results with the extended set of variables, again as the

Table 1: An illustration of the performance improvement 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=0.029)	107%
UKHLS	0.117	0.095 (Δ=0.022)	0.120 (Δ=0.003)	14%
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=0.034)	81%
UKHLS	0.206	0.197 (Δ=0.009)	0.243 (Δ=0.037)	155%
Gallup	0.270	0.240 (Δ=0.031)	0.280 (Δ=0.010)	58%

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

improvement in R-squared over OLS. The R-squared figure for the extended set of variables is around double that for the restricted set in all of the algorithms. 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.[11]

Given the large number of variables in the extended set, we now also estimate LASSO regressions, which serve as a device for variable selection (see Section 2.2). LASSO regressions marginally outperform the corresponding OLS models. 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 perform poorly, and worse than 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 compare the performance gains from gradient boosting to those from the inclusion of health information in OLS estimation.[12] The results in Panel B of Table 1 show that these gains are again approximately equivalent to the role of health in predicting wellbeing.

We thus conclude that tree-based ML algorithms can predict wellbeing better than conventional methods. These gains are moderate in absolute terms, but are meaningful when compared to the predictive power of health. However, these gains do come from algorithms that take up to 100 times longer to estimate.[13] ML algorithms thus involve a trade-off between computational

[11] All of these R-squared figures come from the test set, and do not reflect the mechanical increase in the share of explained variance in the training set due to adding more variables to the model. The results for the training set can be found in Panel B of Online Appendix Figure A2.

[12] In these extended specifications, there are multiple health variables in each dataset: we drop 21, 19 and 12 health-related variables in the Gallup, the SOEP and UKHLS respectively.

[13] This figure comes from the comparison between OLS and RF on the Gallup data with the extended dataset.

burden and predictive performance.

3.2 Variable importance

Based on estimating permutation importances across the extended set of variables, we now assess whether the variables that ML identifies as important in predicting life satisfaction are different from those in the conventional literature. Figure 2 lists the five most-important variables identified in OLS and GB (the best-performing ML algorithm) in each dataset.[14] The bars and numerical values refer to permutation importance, *i.e.* the drop in the 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 each country, individual health and interpersonal relationships are among the most-important predictors. As expected, respondents whose health limits their activities are on average less satisfied with their lives, while people with fulfilling relationships are typically more satisfied. The directions of the estimated effects are in line with those in previous conventional work. ML algorithms and OLS thus generally agree on the direction and approximate size of the most-important variables (see Online Appendix Table A4 for the effect-size estimates).

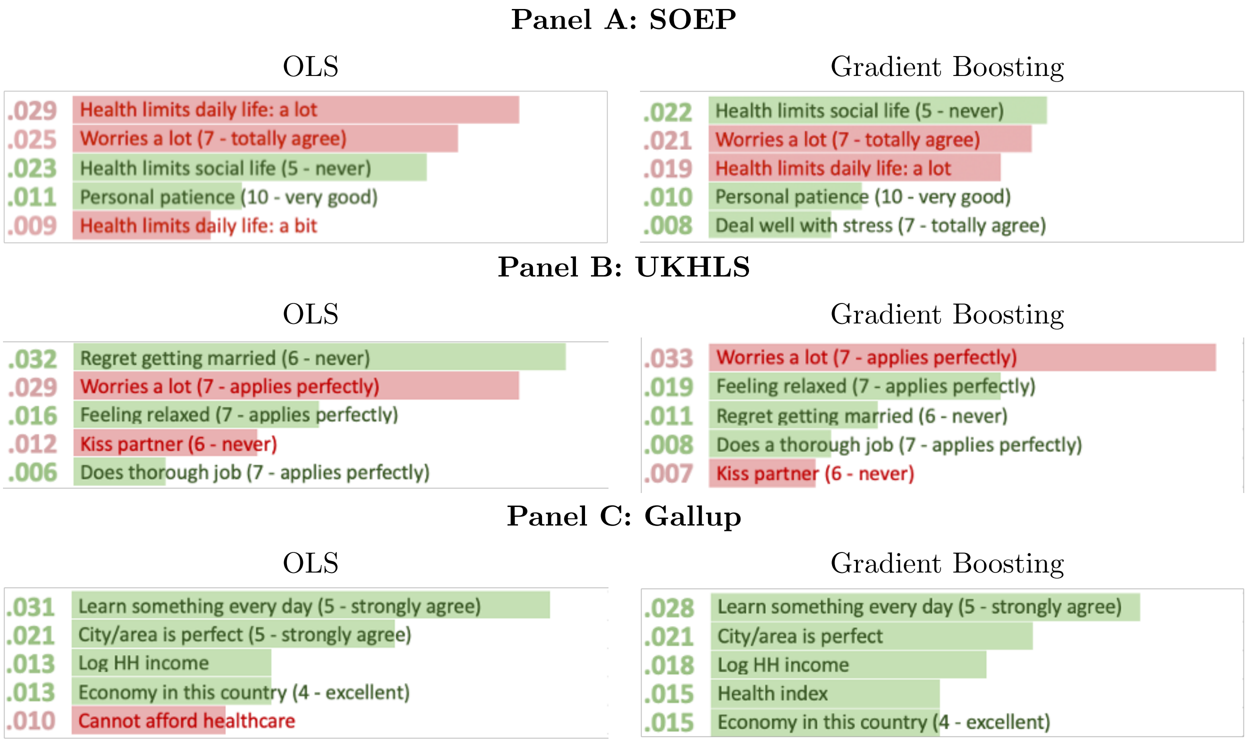
A more-systematic measure of agreement between ML and OLS is given by the rank correlations (in terms of their permutation importance) of each variable across algorithms and datasets. The results in Online Appendix Table A5 reveal strong agreement between GB and RF, with the rank correlation figure never being 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, personality traits are also identified as being important in the UKHLS and SOEP.[15] This is line with some previous research that has underlined the relationships between personality and wellbeing (Ferrer-i-Carbonell and Frijters, 2004; Proto and Zhang, 2021). In the UK, worrying a great deal and feeling very relaxed appear in the top-three wellbeing predictors. In the German data, worrying and patience are in the top-four predictors in both OLS and Gradient Boosting.

Beyond these similarities, there are some cross-country differences. The most striking concerns 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 due to the extended-set variables being different across countries, we reproduce the analysis for the restricted set, which has a common set of variables. There, 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 Online

[14] See Online Appendix Table A4 for the Top-10 most-important variables for OLS, RF and GB in each dataset.
[15] Personality traits are unfortunately not measured in the Gallup survey.

Figure 2: Permutation importance and pseudo partial effects of OLS and GB on the extended set of variables: the five most-important variables.



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

Appendix Table A6). 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 Wellbeing by age and income

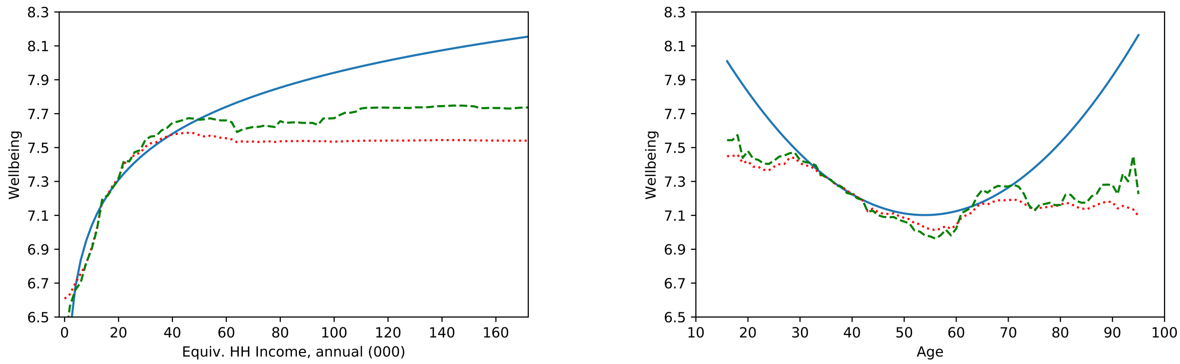
Whether the relationship between age and wellbeing is U-shaped, and whether there is a satiation point beyond which income no longer yields wellbeing are two open and hotly-debated questions. Tree-based algorithms freely estimate the most-appropriate functional forms. They are thus particularly well-suited to act as agnostic judges in these debates.

Figure 3 and Online Appendix Figure A3 depict average predicted wellbeing for different levels of age and income, holding the other covariates constant. In the OLS estimation, illustrated in blue, we assume a quadratic functional form for age and a log-linear functional form for income: these are both extremely common in the existing literature. The relationships estimated using RF are plotted in red, and those using GB in green.

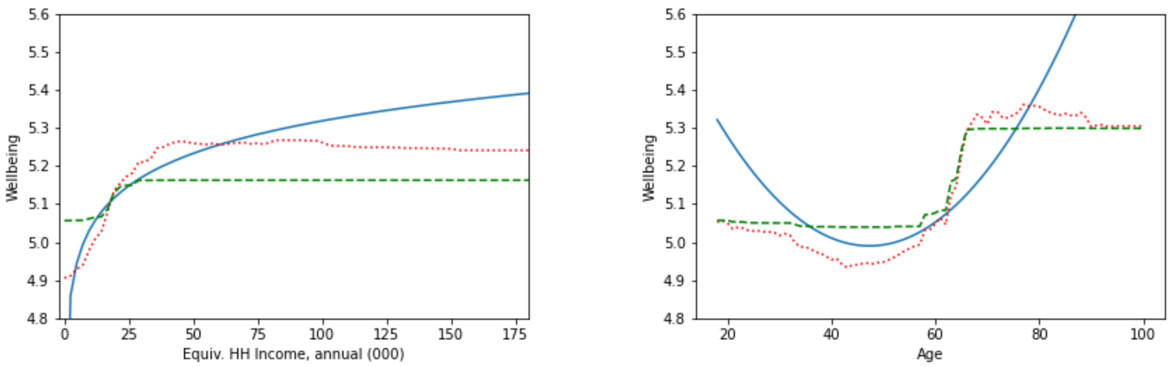
Both ML algorithms track the log-linear functional form for income over much of the income distribution. However, once we reach the relatively-high equivalised annual income figures of 50,000 EUR in the SOEP and 40,000 GBP in the UKHLS, ML suggests that wellbeing no longer

Figure 3: Wellbeing, age and household income: restricted set of variables.

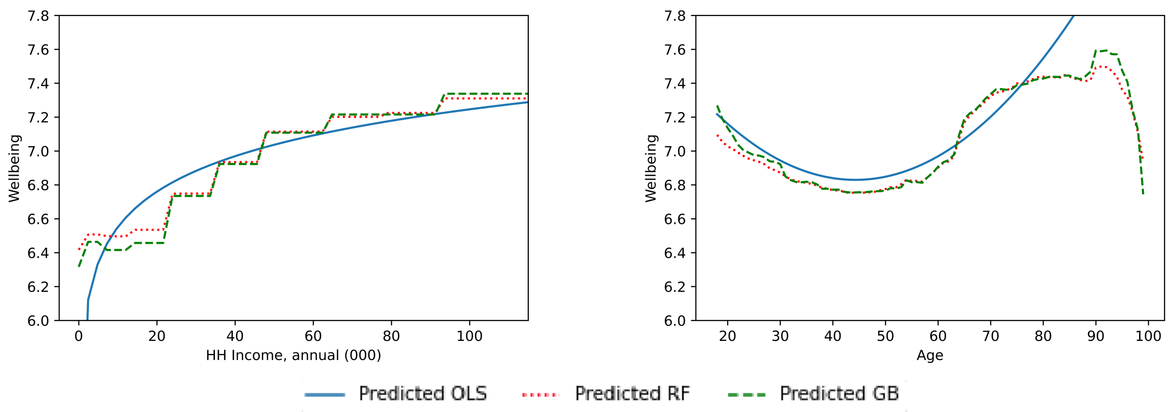
Panel A: SOEP



Panel B: UKHLS



Panel C: Gallup



Notes: Income is continuous in the SOEP and the UKHLS, and we use equivalence-scale adjusted household income in the analysis. For ease of presentation, we only depict the relationship up to equivalent household income figures of 180 000 in the local currency for these two datasets. Income is collected in income bands in Gallup, and there is no information on household size in 2013. The Gallup analysis thus refers to unadjusted household income.

rises with income. We cannot confirm this finding in the Gallup data, where income is measured in discrete bands with a highest value of 100,000 USD or above (equivalent to 70,000 GBP or 78,000 EUR in 2013) .[16] As we also cannot correct for household size, income in Gallup is not comparable to the adjusted equivalent incomes in the SOEP and UKHLS. Given these caveats, and in line with previous work on wellbeing from the US (Kahneman and Deaton, 2010; Killingsworth, 2021), we find no evidence of satiation in the relationship between income and evaluative wellbeing in the Gallup data.

With respect to the relationship between age and wellbeing, both ML estimations replicate the well-known U-shape up to age 70. This pattern appears in all countries, and is the least-pronounced in Germany and the most-pronounced in the US. However, unlike the smooth quadratic U-shape in the OLS results, we find a much more pronounced 'kink' at around age 65 in ML, which we suspect reflects higher wellbeing around the age of retirement (Gorry et al., 2018; Wetzel et al., 2016). Moreover, and in particular for the US, there is a steep drop in predicted wellbeing above age 90. Our results are then in line with the parametric findings in Cheng et al. (2017) of a clear U-shape in wellbeing during working age. They are also in line with the neural network-based results of Kaiser et al. (2022), which focused on Germany only.

4 Discussion

We draw four main conclusions.

First, tree-based ML approaches do indeed perform better at predicting wellbeing than conventional linear models, and gradient boosting consistently outperforms random forests. Although the gains in R-squared are modest in absolute terms, they are comparable with – and sometimes exceed – the extent to which information on respondents' health adds to wellbeing predictions. This finding is not mechanical: Performance is evaluated out of sample so that there is no guarantee that an unconstrained functional form will perform better (Mehta et al., 2019; Wolpert and Macready, 1995). The improved performance rather implies that there are genuine non-linearities in the drivers of wellbeing.

Second, when we use all of the non-wellbeing variables available as predictors, we more than double the explained variation in wellbeing for all estimation methods. The R-squared figure with this extended set of variables is around 0.3, which looks to be the maximum achievable with the current survey data. This is approximately half of the predictable wellbeing variance, as defined by test-retest correlations found in earlier work (Krueger and Schkade, 2008). Hence it seems that even if we use all the information available in standard social surveys, we still fail to explain about half of the in-principle explainable variance in individual wellbeing.

Third, almost all of the variables that turn out to be important in the extended-data specifications 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 (Diener et al., 2018). In that sense, machine-learning

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

approaches validate the previous human-guided search for the determinants of wellbeing. This looks to be good news for the field.

Unlike OLS, where some functional forms are imposed between the covariates and the outcome, random forests and gradient boosting involve no such *a priori* assumptions. Machine learning can thus be used as an 'impartial arbiter' regarding functional forms. We have here considered the relationships between wellbeing, income and age. Our last finding is that this arbiter provides support for the U-shape in age and, where comparable data is available, a satiation point beyond which higher incomes are unrelated to wellbeing.

We see two directions for future research. The first is to further explore the capabilities of ML models, *e.g.*, by using a combination of theory-based modelling and algorithmic approaches. Another potential approach is to combine unsupervised and supervised learning. For example, datasets can be split into overlapping clusters of individuals based on subsets of independent variables, and the predictive advantage of non-linear ML models may be higher when applied to clusters, as compared to the whole dataset at once. Moreover, the analysis here has been correlational, identifying key variables for the prediction of wellbeing. A natural next step is to apply machine learning to the variables that matter most for wellbeing in a causal sense (Wager and Athey, 2018).

The second direction is to extend this analysis beyond rich Western countries. Our findings may well not be reproduced in a more global setting, for example in countries where material needs are more pressing. Insofar as the scope for improving wellbeing is greater in low- and middle-income countries (Helliwell et al., 2022; McGuire et al., 2022; van Agteren et al., 2021) 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. *Stata Journal*, *20*(1), 176–235.

Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting Psychological and Subjective Well-Being from Personality: A Meta-Analysis. *Psychological Bulletin*, *146*, 279–323.

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.

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? implications for subjective survey data. *American Economic Review*, *91*, 67–72.

Blanchflower, D. G., & Oswald, A. J. (2004). Well-Being over Time in Britain and the USA. *Journal of Public Economics*, *88*(7-8), 1359–1386.

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]*.

Boyce, C. J. (2010). Understanding Fixed Effects in Human Well-Being. *Journal of Economic Psychology*, *31*(1), 1–16.

Boyce, C. J., & Wood, A. M. (2011). Personality Prior to Disability Determines Adaptation: Agreeable Individuals Recover Lost Life Satisfaction Faster and More Completely. *Psychological Science*, *22*(11), 1397–1402.

Breiman, L. (1984). *Classification and Regression Trees*. Routledge.

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

Cantril, H. (1965). *The Pattern of Human Concerns*. New Brunswick: Rutgers University Press.

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of Affective Decision Making: How Do Feelings Predict Choice? *Psychological Science*, *27*(6), 763–775.

Chen, L.-Y., Oparina, E., Powdthavee, N., & Srisuma, S. (2022). Robust Ranking of Happiness Outcomes: A Median Regression Perspective. *Journal of Economic Behavior & Organization*, *200*, 672–686.

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. *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., Inglehart, R., & Tay, L. (2013). Theory and Validity of Life Satisfaction Scales. *Social Indicators Research*, *112*(3), 497–527.

Diener, E., Oishi, S., & Tay, L. (2018). Advances in Subjective Well-Being Research. *Nature Human Behaviour*, *2*(4), 253–260.

Dolan, P., Peasgood, T., & White, M. (2008). Do We Really Know What Makes Us Happy? A Review of the Rconomic Literature on the Factors Associated with Subjective Well-Being. *Journal of Economic Psychology*, *29*(1), 94–122.

Dukart, J., Weis, S., Genon, S., & Eickhoff, S. B. (2021). Towards increasing the Clinical Applicability of Machine Learning Biomarkers in Psychiatry. *Nature Human Behaviour*, *5*(4), 431–432.

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

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

Frijters, P., & Beatton, T. (2012). The Mystery of the U-shaped Relationship between Happiness and Age. *Journal of Economic Behavior & Organization*, *82*(2-3), 525–542.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, *130*(4), 956–990.

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.

Jebb, A. T., Tay, L., Diener, E., & Oishi, S. (2018). Happiness, Income Satiation and Turning Points Around the World. *Nature Human Behaviour*, *2*(1), 33.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine Learning of Neural Representations of Suicide and Emotion Concepts Identifies Suicidal youth. *Nature Human Behaviour*, *1*(12), 911–919.

Kahneman, D., & Deaton, A. (2010). High Income Improves Evaluation of Life but not Emotional Well-Being. *PNAS*, *107*(38), 16489–93.

Kaiser, C., & Oswald, A. J. (2022). The Scientific Value of Numerical Measures of Human Feelings. *PNAS*, *119*(42), e2210412119.

Kaiser, C., & Vendrik, M. (2022). How Much Can We Learn from Happiness Data? *Discussion 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.

Kammann, R., & Flett, R. (1983). Affectometer 2: A scale to measure current level of general happiness. *Australian Journal of Psychology*, *35*(2), 259–265.

Kassenboehmer, S. C., & Haisken-DeNew, J. P. (2009). You're Fired! the Causal Negative Effect of Entry Unemployment on Life Satisfaction. *Economic Journal*, *119*(536), 448–462.

Killingsworth, M. A. (2021). Experienced Well-Being Rises with Income, even above $75,000 per Year. *PNAS*, *118*(4), e2016976118.

Kong, F., Ding, K., Yang, Z., Dang, X., Hu, S., Song, Y., & Liu, J. (2019). Examining gray matter structures associated with individual differences in global life satisfaction in a large sample of young adults. *Social, Cognitive and Affective Neuroscience*, *10*, 952–960.

Krueger, A. B., & Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, *92*(8-9), 1833–1845.

Layard, R., Clark, A. E., Cornaglia, F., Powdthavee, N., & Vernoit, J. (2014). What Predicts a Successful Life? A Life-Course Model of Well-Being. *Economic Journal*, *124*(580), F720–F738.

Lucas, R. E. (2007). Long-term Disability is Associated with Lasting Changes in Subjective Well-Being: Evidence from Two Nationally Representative Longitudinal Studies. *Journal of Personality and Social Psychology*, *92*, 717–730.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2004). Unemployment Alters the Set Point for Life Satisfaction. *Psychological Science*, *15*(1), 8–13.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888*.

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.

McGuire, J., Kaiser, C., & Bach-Mortensen, A. M. (2022). A Systematic Review and Meta-Analysis of the Impact of Cash Transfers on Subjective Well-Being and Mental Health in Low- and Middle-Income Countries. *Nature Human Behaviour*, *6*(3), 359–370.

Mehta, P., Bukov, M., Wang, C.-H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A High-Bias, Low-Variance Introduction to Machine Learning for Physicists. *Physics Reports*, *810*, 1–124.

Molnar, C. (2022). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, *46*(1), 69–85.

Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, a Tutorial. *Frontiers in Neurorobotics*, *7*.

Nikolova, M., & Graham, C. (2022). The Economics of Happiness. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics* (pp. 1–33). Springer International Publishing.

OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being*.

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

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

Oparina, E., & Srisuma, S. (2022). Analyzing subjective well-being data with misclassification. *Journal of Business & Economic Statistics*, *40*(2), 730–743.

Oswald, A. J., & Powdthavee, N. (2008). Does Happiness Adapt? A Longitudinal Study of Disability with Implications for Economists and Judges. *Journal of Public Economics*, *92*(5), 1061–1077.

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. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prati, G. (2022). Correlates of quality of life, happiness and life satisfaction among european adults older than 50 years: A machine-learning approach. *Archives of Gerontology and Geriatrics*, *103*, 104791.

Proto, E., & Zhang, A. (2021). COVID-19 and Mental health of Individuals with Different Personalities. *PNAS*, *118*(37), e2109282118.

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *Astronomical Journal*, *157*(1), 16.

Rohrer, J. M., Richter, D., Brümmer, M., Wagner, G. G., & Schmukle, S. C. (2018). Successfully Striving for Happiness: Socially Engaged Pursuits Predict Increases in Life Satisfaction. *Psychological Science*, *29*(8), 1291–1298.

Ryan, E., R.M. Deci. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, *52*, 141–166.

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.

Silk, A. J. (1977). Test-Retest Correlations and the Reliability of Copy Testing. *Journal of Marketing Research*, *14*(4), 476.

SOEP. (2021). *SOEP-Core v36* (tech. rep.). SOEP Survey Papers.

Stevenson, B., & Wolfers, J. (2013). Subjective Well-Being and Income: Is there any Evidence of Satiation? *American Economic Review*, *103*(3), 598–604.

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.

van Agteren, J., Iasiello, M., Lo, L., Bartholomaeus, J., Kopsaftis, Z., Carey, M., & Kyrios, M. (2021). A Systematic Review and Meta-Analysis of Psychological Interventions to Improve Mental Wellbeing. *Nature Human Behaviour*, *5*(5), 631–652.

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.

Wolpert, D. H., & Macready, W. G. (1995). *No Free Lunch Theorems for Search* (tech. rep.). Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT press.

Wunder, C., Wiencierz, A., Schwarze, J., & Küchenhoff, H. (2013). Well-Being over the Life Span: Semiparametric Evidence from British and German Longitudinal Data. *Review of Economics and Statistics*, *95*(1), 154–167.

Yang, J. (2021). Fast TreeSHAP: Accelerating SHAP Value Computation for Trees. *arXiv preprint arXiv:2109.09847*.

Author contributions: E.O., C.K. and N.G. contributed equally 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 regarding 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.

Online Appendix

A Additional analyses and robustness tests

A.1 Positive and negative affect

We have also evaluated the performance of gradient boosting and random forests on measures of positive and negative affect. The findings for evaluative wellbeing discussed in the main text generalise to these measures. 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 over the past 4 weeks (on a 1 to 5 scale), and negative affect as the analogous average of being angry, sad or worried. The UKHLS dataset does not contain comparable affect data and is not used in this part of analysis.

The detailed Gallup results appear in the top panels of Figure A4 and Table A7. It is striking 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 performs the best, with gains in R-squared over OLS of 0.041 for negative affect and 0.036 for positive affect. Regarding variable importance, Table A7 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 set of the most-important variables when modelling affect.

Our results are qualitatively similar in the German data: gradient boosting performs best, and positive affect is harder to predict than negative affect (see Online Appendix Table A8 and the bottom panels of Figure A4.

A.2 Panel data

Our main findings regarding the ML estimation of evaluative wellbeing are 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 demeaned level of wellbeing is the dependent variable.

The relative predictive performance of OLS and ML in the panel specification is similar to that in the cross-section analysis for the 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 A9 and A10, the most-important variables predicting 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 in predicting wellbeing.

B Tables and Figures

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

	SOEP	Gallup	UKHLS
Panel A: Random Forest			
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: The hyperparameters are identified via a grid search by minimising 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: *min - max, mean (sd)*.

Variable	SOEP	UKHLS	Gallup
Age	16 - 105	18 - 103	18 - 99
	47.08 (17.27)	49.32 (17.74)	52.74 (18.08)
Area of residence	16 distinct values	12 regions	51 distinct values
BMI	11.10 - 84.50	11.80 - 74.20	10.62 - 114.17
	26.33 (4.53)	26.33 (3.19)	27.40 (5.66)
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
Log HH income	0 - 13.88	-0.80 - 12.52	3.40 - 9.90
equiv. in UKHLS and SOEP	9.94 (0.67)	7.40 (0.70)	8.39 (1.02)
Ethnicity/Migration background	3 distinct values (migration background)	18 distinct values (ethnicity)	5 distinct values (ethnicity)
Health	0 – 396 (doctor visits in prev. year) 0.12 (0.33)	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
	0.85 (1.19)	0.56 (0.95)	0.55 (1.05)
Number of people in HH	1 - 16	1 - 16	n.a.
	2.99 (1.52)	2.83 (1.46)	n.a.
Religion	10 distinct values	Binary	8 distinct values
Sex	Binary	Binary	Binary
Working hours	0 - 6669	0 - 180	4 distinct values
	1033.73 (1074.71)	21.13 (20.81)	

Notes: For continuous variables, the range is reported. For SOEP, the values for the categorical variables are as follows. *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, the values for the categorical variables are as follows. *Area of residence:* North East, North West, Yorkshire and Humberside, 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 qualifications. *Labour-force status:* Self-employed, Paid employment (FT/LT), 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, the values for the categorical variables are as follows. *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: List of variables in the extended set.

Group	Description		
	SOEP	UKHLS	Gallup
Area	State of residence.	Country of residence, government office region, urban or rural area.	State of residence.
Cognitive skills		Numeric ability, verbal fluency and word recall scores; self-rated memory, interviewer rated language ability and anxiety.	
Education	Education level, currently in education, need training.	Educational qualifications, age of leaving school and further education, additional training.	Learn something every day, education.
Employment	Employment status, hours worked, current job characteristics.	Current and past employment characteristics, including full- or part-time employment, number of jobs, hours worked, industry, socio-economic classification, unemployment spells, location and commute.	Employment status, working hours, work environment, occupation, company and supervisor characteristics.
Finances	Income from different sources, entitlement to other forms of allowances, dept and assets.	Incomes from various sources, spendings on energy, food and alcohol, problems paying for housing, bills or council tax.	Income, basic access index, not enough money for food or shelter.
Friends and socialising	Number of close friends, use of social website, going out socially.	Number of close friends and their characteristics, belonging to social website, going out socially.	Treated with respect.
Health	Physical and mental health conditions, hospital stays, eventual limitations, and behaviours, including sleep, smoking and diet.	BMI, health conditions, health limits activities.	BMI, health conditions, doctor visits, pain, health limits activities, health insurance and health behaviours, including smoking and diet.
Household composition and family relationships	Household composition, marital status, relationship with other generations in the family, leisure and housework.	Household composition, marital status, family members outside of the household, providing or receiving help, caring responsibilities, quality of the relationships with the partner.	Number of children, marital status.
Interview	Month of interview.	Interview characteristics, including month, year, language of the interview, respondent's cooperation and understating, other people present.	Month, day of the week, time zone.
Membership in organisations		Being a member and being active in organisations, including political party, professional organisations, community groups, social or sports clubs, trade unions or others.	Member of labour union, served in the U.S. military.
Neighbourhood	Noise and pollution, quality of area, relationships with neighbours, local amenities.	Tenure in the neighbourhood and the neighbourhood characteristics, including safety, standards of local services, social cohesion, trust, interactions with neighbours.	City and area characteristics, including safety, affordability of food, medicine and places for exercise, appreciation of the city/area.
News sources		Reported news sources, including TV, internet, newspapers and others; most frequent TV channel, hours of TV per week.	
Personality traits	BIG 5 personality traits, risk preferences.	BIG 5 personality traits.	
Residence	Home ownership and characteristics of the dwelling, including the items in the accommodation.	Characteristics of the residence, including the items in the accommodation (e.g. television, washing machine, etc.), number of bedrooms, value of property; past and current tenancy status.	
Socio-demographic characteristics	Age, ethnicity, gender, nationality, religion.	Age, sex, ethnic group, migration status, religion, parents' education, ethnic group and county of origin.	Age, race, gender, religion.
Views and beliefs	Political preferences, importance of being able to afford something, having children, helping others, being socially and politically active.	Sense of civic duty, attitudes towards voting, perceived political influence, level of interest in politics, partisan support.	State of current economy, national economy is getting better.

Table A4: 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	Equiv. 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	Equiv. 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	HH income	.013	.15	HH income	.021	.30	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: *Equiv. 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 with 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*, *HH income*, *Physical health index.*

Table A5: 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 A6: 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	Equiv. HH Income	.13	.27	Equiv. HH Income	.14	.46
2	Equiv. HH 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	No. of children in HH	.06	.30	No. of children in HH	.03	.07	Working hours	.02	-.29
6	Disability Status	.04	-.52	No. of people in HH	.03	.02	No. of years of education	.02	.17
7	No. of people in HH	.03	-.17	No. of years of education	.02	.07	No. of children in the HH	.02	.08
8	No. of years of education	.03	.11	House Ownership: Owner	.02	.12	No. 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	Equiv. 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	Equiv. 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	Equiv. 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

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

Table A7: Permutation Importance (PI) and Pseudo Partial Effect (PPE) in OLS, RF and GB for positive and negative affect: the 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 A8: Permutation Importance (PI) and Pseudo Partial Effect (PPE) of OLS, RF and GB for positive and negative affect of the 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 A9: 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	Equiv. HH income (avg.)	.012
6	Age (avg.)	.013	Equiv. HH income (avg.)	.009	LT sick or disabled (empl.) (avg.)	.012
7	Retired (avg.)	.012	Married (avg.)	.006	Married (avg.)	.009
8	Equiv. 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 A10: 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	Equiv. HH Income (avg.)	.059	Equiv. HH Income (avg.)	.049
3	Equiv. HH 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.

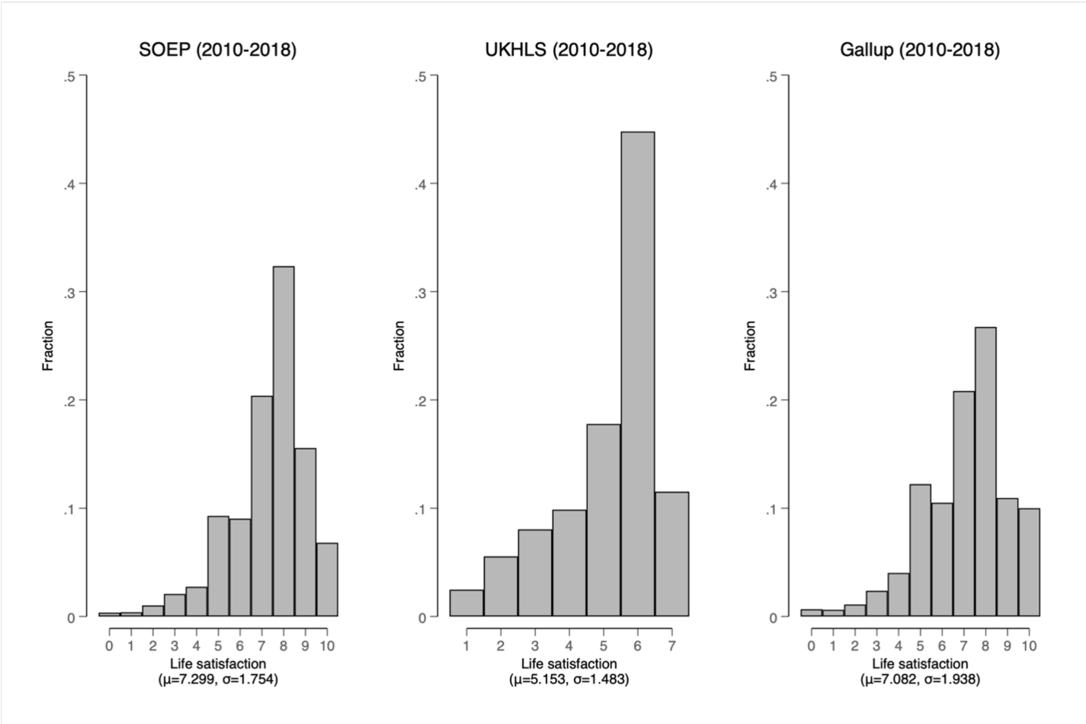


Figure A2: Differences in within-sample performance between OLS and ML. The R-squareds are calculated from the training data and are not representative of out-of-sample performance.

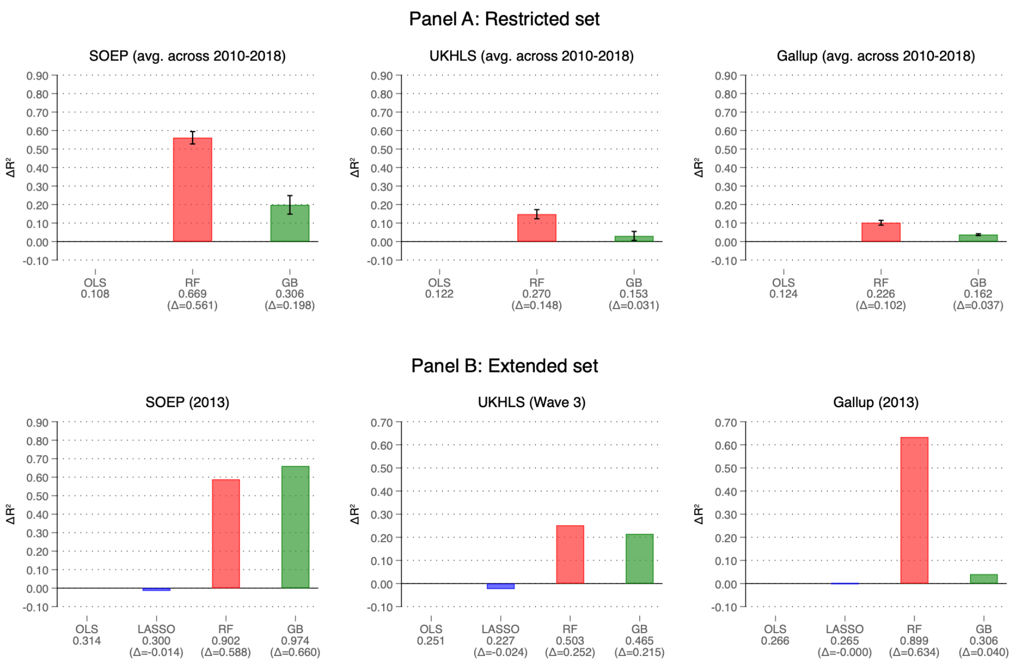
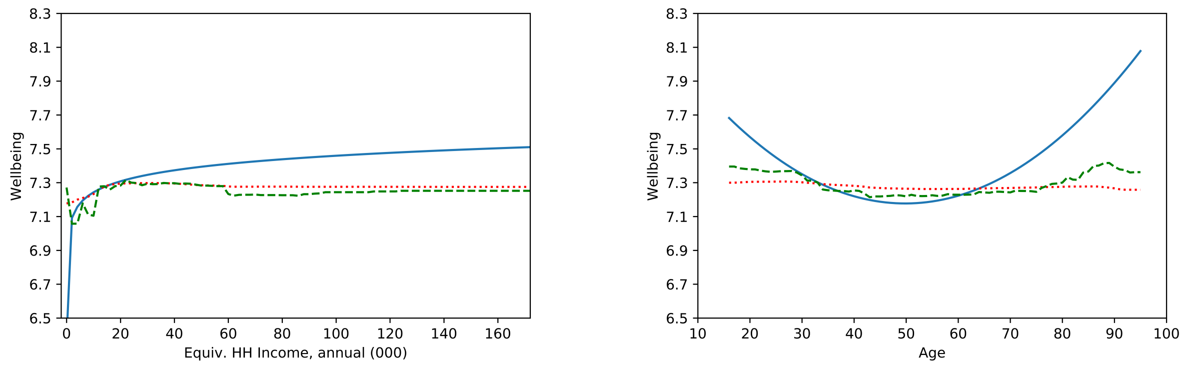
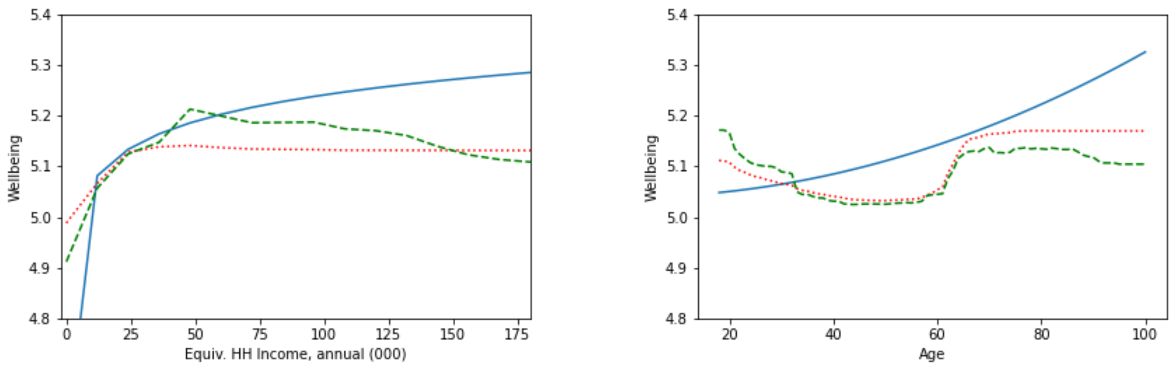


Figure A3: The mean effects of age and household income on wellbeing in the Extended Set of variables.

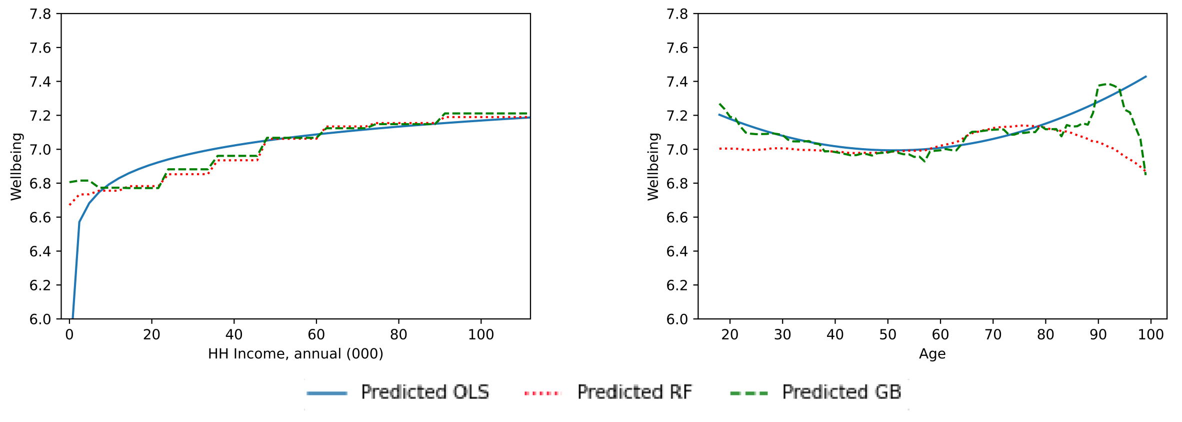
Panel A: SOEP



Panel B: UKHLS



Panel C: Gallup



Notes: Income is continuous in the SOEP and the UKHLS, and we use equivalence-scale adjusted household income in the analysis. For ease of presentation, we only depict the relationship up to equivalent household income figures up to 180 000 in the local currency for these two datasets. Income is collected in income bands in Gallup, and there is no information on household size in 2013. The Gallup analysis thus refers to non-adjusted household income.

Figure A4: Differences in out-of-sample performance between OLS and ML 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'.

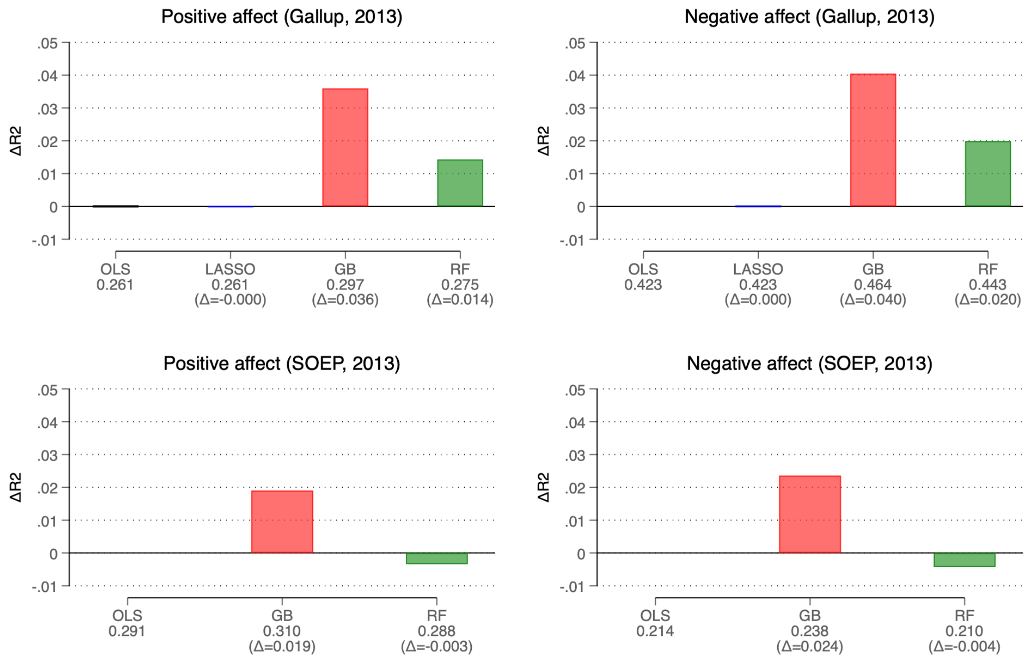


Figure A5: Differences in R-squared between OLS and when modelling the level of wellbeing with Mundlak terms using 2013 SOEP and Wave 3 UKHLS data with the Restricted Set of variables. The R-squareds are calculated from unseen 'testing data'.