Centre for Economic Performance

Discussion Paper ISSN 2042-2695

No.1853
June 2022

What makes a satisfying life? Prediction and interpretation with machine-learning algorithms

Andrew E. Clark
Conchita D'Ambrosio
Niccoló Gentile
Alexandre Tkatchenko

LSE The London School of Economics and Political Science

UKRI Economic and Social Research Council

Abstract
Machine Learning (ML) methods are increasingly being used across a variety of fields and have led to the discovery of intricate relationships between variables. We here apply ML methods to predict and interpret life satisfaction using data from the UK British Cohort Study. We discuss the application of first Penalized Linear Models and then one non-linear method, Random Forests. We present two key model-agnostic interpretative tools for the latter method: Permutation Importance and Shapley Values. With a parsimonious set of explanatory variables, neither Penalized Linear Models nor Random Forests produce major improvements over the standard Non-penalized Linear Model. However, once we consider a richer set of controls these methods do produce a non-negligible improvement in predictive accuracy. Although marital status, and emotional health continue to be the most important predictors of life satisfaction, as in the existing literature, gender becomes insignificant in the non-linear analysis.

Keywords: life satisfaction, well-being, machine learning, British cohort study
JEL: I31; C63

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 Nick Powdthavee for helpful comments. We gratefully acknowledge financial support from the University of Luxembourg's Audacity project "IAS - DSEWELL". Andrew Clark acknowledges financial support from the EUR grant ANR- 17-EURE-0001.

Andrew E. Clark, Paris School of Economics and Centre for Economic Performance, London School of Economics. Conchita D'Ambrosio, University of Luxembourg. Niccoló Gentile, University of Luxembourg. 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

One of the major domains of Social Science is the understanding of individual well-being, with the aim of predicting what makes a successful life. This success in well-being terms can be defined either objectively or subjectively: the former focuses on measures such as income or consumption, where those with more economic resources are considered to be better-off, while the latter relies on individuals' own evaluations of how well their life is going. We here consider this second type of measure, commonly called subjective well-being.

The prediction of subjective well-being starts with the analysis of its associations with a set of key observable characteristics, which can be at either the individual or a more-aggregated level (see Clark, 2018, for a survey). We will here focus on individual-level characteristics. One of the central individual variables is income, both in absolute terms and expressed relative to others (Clark and Oswald, 1996, and Luttmer, 2005, are two analyses including relative income), and another (conditional on income) is unemployment (Winkelmann and Winkelmann, 1998, and Clark and Oswald, 1994, among many others). With respect to other non-pecuniary characteristics, the married are more satisfied than the non-married (see Stutzer and Frey, 2006, for a discussion of selection into marital status), and the correlations with both physical and mental health are typically positive (Dolan *et al*., 2008), with Layard *et al*. (2014) and Clark *et al*. (2018) finding the correlation with emotional health to be larger. The association between subjective well-being and education is on the contrary more ambiguous (see Chapter 3 of Clark *et al*., 2018). Women are often found to be more satisfied with their lives (Helliwell *et al*., 2016) but at the same time report more stress (Kahneman and Deaton, 2010). While there is a vibrant literature on subjective well-being and age, this will not be relevant in the analysis we carry out here, which is based on one wave of a birth-cohort dataset (in which all respondents are therefore the same age).

The vast majority of these findings regarding the individual correlates of well-being come from parametric models. These models are, however, more useful in terms of explaining rather than predicting the dependent variable, at a potential cost in terms of predictive accuracy. The related statistical and methodological arguments will be presented below. At the same time, the growing computing power of current machines (including computers) has recently made Machine

Learning (henceforth ML) widely available. Broadly, ML looks for a pattern (in general, non-linear) that maps a set of explanatory variables to the dependent variable of interest in a training set of data, and then focuses on generalizations, *i.e.* on obtaining good predictions of the dependent variable on data from outside of this training set.

Our aim here is to see whether two key ML algorithms – Penalized Linear Models and Random Forests – can provide more-accurate predictions of subjective well-being than does the more-traditional linear model (which we will henceforth call non-penalized linear regression). The model we analyze is that in Layard *et al.* (2014), the aim of which (as indicated in their article title) is the prediction of life satisfaction; this thus provides a natural starting point for our analysis.

The greater predictive accuracy of ML models comes at the cost of being less-easily interpretable than non-penalized linear regressions. Following Kim *et al*. (2016), interpretability refers to the degree to which a human can consistently predict the model's result. We will below apply model-agnostic methods to our results in order to render the predictions from Random Forests more interpretable.

The remainder of the paper is organized as follows. Section 2 describes the British Cohort Study data that we use in our empirical applications. The results are then presented in Section 3, and interpreted in Section 4. Last, Section 5 concludes.

2 Data

We use the same dataset as in Layard *et al.* (2014), the British Cohort Study (BCS). This is a birth-cohort study, covering all individuals in the UK who were born in the second week of March 1970 (cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/). Since the birth wave of the survey in 1970, there have been ten other waves ('sweeps') at ages 5, 10, 16, 26, 30, 34, 38, 42, 46 and 51. Layard *et al.* (2014) focus on the life satisfaction that respondents report at age 34. Of the 17 000 initial births recorded, 8 867 individuals provided information at age 34 on all of the variables that we will use in the analysis, as listed below.

We initially consider only the eight adult age-34 explanatory variables that appear in Layard *et al.* (2014): these are our explanatory variables, which we use to predict *Life Satisfaction*, our

dependent variable. The only variable that we treat differently from them is health. Our health measure comes from the BCS analysis in Clark and Lepinteur (2019), and is the number of conditions from which the individual suffers; that in Layard *et al.* (2014) is instead self-assessed health at age 26 measured on a scale of 1 to 4 (from 'Bad' to 'Excellent'). We prefer an objective health measure for common-method variance reasons (even if the subjective health measure in Layard *et al.*, 2014, is lagged by two waves).

Our eight initial explanatory variables are the following:

- *Ln*(*income*) at age 34. Household equivalent disposable income using the OECD equivalence scale, expressed in Pounds.
- *Educational Achievement* at age 34. This is a single variable with six distinct cardinal values, obtained from a regression of male log full-time earnings on having a family, childhood emotion and conduct, and five education dummies. The resulting values are 0.750 (PhD or Master), 0.486 (Degree), 0.237 (A-level), 0.188 (GCSE), 0.043 (CSE), and 0 (No qualifications; this was the omitted category in the regression).
- *Employment* at age 34. A dummy variable for not being unemployed at the time of the interview.
- *Has a Partner* at age 34. This is a single variable with four distinct cardinal values, obtained from a regression of life satisfaction on three family dummies and a number of life-success variables. The resulting estimated coefficients on the family dummies are 0.685 (Married and cohabiting with children), 0.530 (Married/cohabiting without children), -0.004 (Single with children), and 0 (the omitted category: Single without children).
- *Good Conduct* between ages 16-34: One unit of 'crime' here is being found guilty by a criminal court or formally cautioned at a police station. Good Conduct is the maximum observed number of crimes between ages 16 and 34 years in the BCS sample (25 crimes) minus the individual's own number of crimes.

- *Physical Health* at age 26. This is a cardinal variable for the number of health conditions from which the individual suffers, from a list of 15 (see Appendix B.[1] We multiply this figure by -1, so that higher values refer to better physical health.
- *Mental Health* at age 26. This is the sum of the respondent's replies at the age-26 BCS wave to 24 questions covering aspects such as worry and irritation, and physical symptoms like poor appetite and headache. The total number of conditions, multiplied by -1, is our index of mental health. 665 individuals had missing values for mental health at age 26; for these individuals we take their value at age 30 instead.
- *Gender*. 1 if female, 0 for male.

The dependent variable is *Life satisfaction* at age 34. This comes from the following question: "Here is a scale from 0-10. On it "0" means that you are completely dissatisfied and "10" means that you are completely satisfied. Please tick the box with the number above it which shows how dissatisfied or satisfied you are about the way your life has turned out so far."

Our expanded analysis of life satisfaction adds 16 additional explanatory variables reflecting life at age 34: *Number of people in the household, Number of natural children of the Cohort Member in the household, Number of non-natural children of the Cohort Member in the household, Number of rooms in the household, Type of accommodation, BMI, Alcohol units per week, Cohort Member's main activity, Highest academic qualification, Disability status, Whether the mother is alive, Whether the father is alive, Marital status, Weekly smoking habits, Tenure status,* and *Whether health limits everyday activities*. These new explanatory variables are likely highly correlated with some of the eight original explanatory variables: we will discuss this issue below when presenting the results. The descriptive statistics of all our variables appear in Appendix Table A, which also contains the coding details for all the variables, including Type of accommodation, Alcohol units per week, and Cohort Member's main activity.

The treatment of missing values depends on the nature of the variable. Missing values for categorical variables are not imputed. The rationale here is that the missing values are not at

[1] We retain the two-wave lag (i.e. using age-26 values) in order to be consistent with Layard *et al*. (2014). Information on some, but not all, of the conditions used to construct the Physical Health index are also available at age 34.

random, and potentially contain additional information about the individual. We instead consider the missing categories (there may be more than one for a given variable) as separate values to be used in the empirical analysis. Of the 16 new explanatory variables proposed above, the only categorical variable with significant missing information is Alcohol units per week (which is measured in categories), with 1683 missing values. These correspond to individuals who reported never drinking or only on special occasions (these individuals are assigned a missing value code of -1 in the BCS questionnaire). The next most-frequent occurrences of missing values are for BMI and Whether the father is alive, with much smaller numbers of 246 and 121.

In the linear regression models, we create dummies for each value of the following categorical variables: Type of accommodation, BMI, Alcohol units per week, Cohort Member's main activity, Highest academic qualification, Disability status, Whether the mother is alive and Whether the father is alive (both of these are categorical, as they distinguish between the living parent being in or outside of the household), Marital status, Weekly smoking habits, Tenure status, and Whether health limits everyday activities.

The numerical variables Number of people in the household, Number of natural children of the Cohort Member in the household, Number of non-natural children of the Cohort Member in the household, and Number of rooms in the household have, respectively, 25, 25, 25, and 53 missing values, also labelled via negative numbers. We impute the negative missing values for these variables by the mean of the observed value. Nonetheless, there are only few observations that have missing values for these numerical variables in the dataset (between 0.3% and 0.6% of the observations), and our findings are unaffected if we instead simply drop the observations with missing values for these numerical variables. In the Random Forest analysis, these missing negative values were left as they appear in the data, as here the different explanatory variables' values only serve to define the sample splits, with the actual numerical values not affecting the calculation of the value of the dependent variable.

3 Machine-Learning Algorithms: Presentation and Results

The choice of the ML technique to be used depends on the *interpretability–predictive accuracy* trade-off (see James *et al.,* 2013, for a discussion). In general, the most internally-

interpretable algorithms are the least flexible: these less-flexible algorithms provide straightforward intuitions about the relationship between each of the explanatory variables and the dependent variable. If we wish the model to be *interpretable*, we may then prioritize less-flexible models. If, on the contrary, we are more concerned about accurate *prediction*, we may sacrifice interpretation in favor of more-flexible complex models. Accurate prediction may be at a premium, for example, in contexts in which we already have strong theoretical arguments regarding the explanatory variables-dependent variable relationship, and want to establish the best-possible predictive map. Linear Regression and Deep Feedforward Neural Networks can be considered as two polar examples in this trade-off continuum.

Nonetheless, the interpretability-predictive accuracy trade-off is not a strict dichotomy. As we will see below, model-agnostic interpretative tools also allow for inference in more-flexible methods. Equally, inflexible methods can produce similar (or even better) performance than more-flexible ones (for example, if the joint distribution of the explanatory variables and the dependent variable is relatively simple to model).

We will start our analysis of subjective well-being in the BCS data in the following sub-section by considering linear models. Computations were performed using the *scikit-learn* library in Python (Pedregosa *et. al.*, 2011), and *glmnet* in R (Friedman *et al.,* 2010).

3.1 Non-Penalized and Penalized Linear Regressions

The standard linear non-penalized regression is our benchmark. This is a special case of an *Elastic Net Regression*, the general form of which is (see Zou and Hastie, 2005):

$$\min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 \quad + \quad \lambda\left[\frac{1-\alpha}{2}\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right] \tag{1}$$

where λ and α are hyperparameters: parameters that are used to regulate the learning process, whose value has to be determined before the estimation of the β's. Penalizing by the sum of squares of the betas produces coefficient shrinkage, balancing the bias and variance of the estimates. It does not however yield a parsimonious model as all variables are retained – none of the coefficients are shrunk to 0. Automatic variable selection instead comes from penalizing the sum of the absolute values of the betas (Zou and Hastie, 2005). The values of λ and α can either be input manually (*ex ante*) or discovered via cross-validation (*tuning*: see Section 3.1.2 below).

We first consider five different values of α, (0, 1, 0.25, 0.50 and 0.75), and in each case use 5-fold cross-validation on the training set (which will cover 80% of the individuals) to find the optimal value of λ.

The linear non-penalized regression empirical loss function (*i.e.* that of OLS) is given by Equation (1) with $\lambda = 0$. When $\lambda \neq 0$, a value of $\alpha = 0$ corresponds to the *Ridge Regression* minimization problem, and $\lambda \neq 0$ and $\alpha = 1$ to the *Lasso Regression* minimization problem (where *Lasso* stands for *Least Absolute Shrinkage and Selection Operator*).

In linear regression, the goal is to estimate the unknown mapping under the assumption that the dependent variable is linear in the parameters, by minimizing the squared distance between the predicted and observed values.

We analyze these four cases ($\lambda = 0$, and $\lambda \neq 0$ with α either 0, 1, or in the interval) in turn, discussing the rationale for each case and the ensuing results.

3.1.1 Linear Regression - Non-Penalized

The standard linear regression model corresponds to $\lambda = 0$. Defining $X \in \mathbb{R}^{n \times k}$ as the matrix whose element $x_{i,j}$ is the value of the j^{th} explanatory variable for the i^{th} individual, the (non-penalized) linear regression minimization problem is usually presented as:

$$min_{\boldsymbol{\beta} \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2 \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of values of the continuous dependent variable for each of the n individuals in the sample. The underlying assumed mapping is linear in the parameters:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \sigma^2 I_n)\ . \tag{3}$$

Additional standard requirements are the conditional mean independence of the error term with respect to the explanatory variables (formally, $E(\boldsymbol{\varepsilon}|X) = \mathbf{0}$), no perfect multicollinearity, so that no one column in X can be expressed as a linear combination of the others (or more simply that $rank(X) = k < n$) and that the error term $\boldsymbol{\varepsilon}$ is distributed as in (3). The latter can be relaxed by allowing for *heteroscedasticity* (where the variance of the error term's distribution is individual-dependent), which often appears as a robustness check. Under these conditions, it is well-known (the Gauss−Markov Theorem) that the Least Squares estimator solving (2)

$$\widehat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y} \tag{4}$$

is the Best Linear Unbiased Estimator (*BLUE*), in that it has the lowest variance of all the unbiased linear estimators.

Given its additive structure, the linear regression is arguably the most-interpretable model, as $\hat{\beta}_{OLS,j}$ is the predicted change in y_i following a unit change in $x_{i,j}$, for all individuals i and keeping all other explanatory variables $\boldsymbol{x}_{i,-j}$ constant. If the variables are standardized, a similar interpretation holds in terms of the correlation between standard deviations, and the square of each estimated coefficient $\hat{\beta}_{OLS,Stand,j}$ shows how much the explanatory variable $\boldsymbol{x}_j$ contributes to the dependent variable's variance, ignoring its covariance with the other explanatory variables (Layard *et al.*, 2014). Nonetheless, linear regression is inflexible due to the stringent parametric linearity assumption and the other requirements noted above.

We will compare the performance of our models using the Test Mean Squared Error (MSE), considering a random split where 80% of the individuals appear in the training set (*S*) and the remaining 20% are in the test set (*T*). In general, *S* and *T* have no individuals in common and come from the same data-generating process. We train our algorithms on the set *S* to learn the mapping $\hat{f}: \mathbb{R}^{\mathrm{k}} \rightarrow \mathbb{R}$. We then assess the empirical quality of this mapping via the following statistic:

$$MSE_{test} = \frac{1}{n(T)} \sum_{i:(\boldsymbol{x_i}, y_i) \in T} \left(\hat{f}(\boldsymbol{x_i}) - y_i \right)^2 \tag{5}$$

where *n(T)* represents the cardinality of the test set T. For instance, in the case of linear regression:

$$\hat{f}(\boldsymbol{x_i}) = \boldsymbol{x_i}' \widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS,train}} \tag{6}$$

for all the individuals *i* in *T*, with $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS,train}}$ having been learned from the training set. We add the subscript 'train' to the estimated coefficients to stress that these come from the training set, but are evaluated in terms of their ability to map the explanatory variables onto the dependent variable using the data from the test set.

We now present the Test MSEs for predicting life satisfaction, as well as the Training MSEs, defined as in (5) but over the elements in the Training Set *S*. All non-dummy explanatory variables are standardized (standardization is a normalization and does not affect the quality of the fit). *Original* refers to the model including only the eight adult explanatory variables from Layard *et al.* (2014), and *Extended* to the 21-explanatory variable model (which become 96 once the

dummies are created from the categorical variables) corresponding to five of the eight original explanatory variables in Layard *et al.* (2014) and the 16 new explanatory variables. Three of the eight original explanatory variables are dropped (or rather expanded) in the Extended model. The Original explanatory variable 'Has a partner' is now redundant, as the newly-added variables include both respondent marital status and the number of natural and non-natural children. Equally, educational achievement is replaced by the highest academic qualification, and the original employed dummy is now one of the categories of the newly-added respondent main-activity variable. In order to avoid potential multicollinearity issues, we omit the most-populous category for each categorical explanatory variable, and drop entirely all categories covering fewer than 15 individuals: these dropped categories are listed in Appendix D.[2] As a result, the number of explanatory variables falls from 96 to 72.

All models were fitted 100 times with 100 different randomly-drawn train–test splits (in all of which 80% of observations were assigned to the training set). Table 1 lists the average Mean Squared Errors from these 100 different splits, with their associated standard deviations in parentheses.

Table 1. The Performance of the Linear Regression

	Training MSE	**Test MSE**
Original	2.78 (0.03)	2.79 (0.11)
Extended	2.57 (0.02)	2.65 (0.09)

Notes. These figures show the average performance of linear regressions in predicting life satisfaction in 100 different train-test splits, with 80% of the sample in the training set and the error calculated on the remaining 20% test-set individuals. Standard deviations are in parentheses.

Adding the 16 new explanatory variables – for a total of 72 plus the constant - in the Extended model improves the Test Set performance, with a reduction in the MSE of 5.3% (from a figure of 2.79 to 2.65). Moreover, while in the Original dataset the training and testing accuracy figures are

[2] Without this exclusion, there are 18 perfectly multicollinear cases (out of the 100). This occurs with sparse categorical dummy cells, when all of the 1's are randomly-allocated to the test set (producing a column of 0's in the training set).

almost identical, in the Extended case the Training MSE is 3% lower than the Test MSE (2.65 *vs*. 2.57).

The procedure to avoid multicollinearity does nonetheless involve a potentially substantial loss of information. Considering, for instance, Marital Status, we of course have to drop one category in order to estimate the coefficients on the other categories: here we drop the most-populous category ('Married', with 4817 observations); we in addition drop 'Widowed' (12 observations) and 'Other missing' (3 observations), for which we therefore do not estimate a coefficient. However, these small groups may still be of policy interest – especially the Widowed, whose life satisfaction (as we will see with Random Forests) is particularly low. As such, machine-learning techniques that are capable of dealing with multicollinear datasets can be of use, as they allow us to model the relationship with the dependent variable for individuals in these more sparsely-populated categories. To this end, in what follows we consider Penalized Linear Regressions that allow for the inclusion of all of the response categories, for a total of 96 explanatory variables.

3.1.2 Multicollinearity and Ridge Regressions

The Ridge Regression estimator (Hoerl and Kennard, 1970) corresponds to the minimization of (1) with $\alpha = 0$:

$$\min_{\boldsymbol{\beta} \in R^k} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'\beta})^2 + \frac{\lambda}{2} \sum_{j=1}^{k} \beta_j^2 \tag{7}$$

where λ is a tuning parameter. It can be shown that the Ridge Regression estimator from (10) is:

$$\widehat{\boldsymbol{\beta}}_{Ridge} = (X'X + \lambda I_k)^{-1} X'\boldsymbol{y}. \tag{8}$$

The Ridge estimator can be calculated even under perfect multicollinearity, as $\lambda > 0$. In the case of harmful, but not perfect, multicollinearity, it can be seen that the presence of λ reduces the absolute values of the estimates. The larger is the chosen λ (via hyperparameter tuning or ex-ante choice), the greater is the coefficient shrinkage - although the coefficients never become zero.

The variance of the Ridge estimator is:

$$Var(\widehat{\boldsymbol{\beta}}_{Ridge} \mid X) = \sigma^2 (X'X + \lambda I_k)^{-1} X'X (X'X + \lambda I_k)^{-1} < \sigma^2 (X'X)^{-1} = Var(\widehat{\boldsymbol{\beta}}_{OLS} \mid X). \tag{9}$$

This variance is smaller than that from OLS for every $\lambda > 0$. However, $E(\widehat{\boldsymbol{\beta}}_{Ridge} \mid X) \neq \beta$ due to shrinkage, so that the coefficients are *biased* under the linearity assumption, whereas $E(\widehat{\boldsymbol{\beta}}_{OLS} \mid X) = \beta$. The broad idea behind the use of the Ridge estimator is that by introducing some bias into the estimates, we can reduce the variance up to a point at which the associated MSE is lower than that from OLS.

The Ridge Regression results appear in Table 2. The optimal λ^* here is chosen from a grid of 100 values via *5-fold cross−validation*[3] on the training set solving (7). The λ^* producing the smallest average cross-validated MSE is then introduced into (7), producing the Ridge estimator in (8). Last, the fitted model is used to assess the quality of the fit on the data in the test set, measured via the Test Set MSE as in (5). The procedure is again applied with 100 different random train-test splits, and the results refer to the average performance and associated standard deviations. We also list the mean and standard deviation of λ^*. Note that standardization is required here for all explanatory variables, including the dummies, given the presence of the penalization term.

Table 2. The Performance of the Ridge Regression

	Training MSE	**Test MSE**	**$\lambda*$**
Original	2.78 (0.03)	2.79 (0.11)	0.06 (0.001)
Extended	2.57 (0.02)	2.65 (0.10)	0.35 (0.08)

Notes: This table lists the mean performance and optimal $\lambda*$ of the Ridge regression predicting life satisfaction over 100 different train-test splits, each with 80% of the sample in the training set and the error calculated on the remaining 20% of individuals in the test set. The λ^* obtained for each split comes from a 5-fold cross-validation on the training set. Standard deviations appear in parentheses.

Prediction in the test set using the Ridge estimator on the Extended dataset now always produces a reasonable Test Error, even without dropping any dummy variables.

In the Original dataset, the Ridge estimator's lower variance does not suffice to offset the loss in accuracy: the Test MSE of the Ridge estimator is 2.79. This reflects the absence of

[3] The training set is split into k equally-sized blocks for k-fold cross-validation. One of these k blocks is used for validation, and the model is fitted on the remaining k-1 blocks. This process is repeated k times until each of the k blocks has been used for validation. The cross-validated score for a given hyperparameter value is the average validation score (the MSE in our case) over the k folds (we here use 5 folds).

multicollinearity in the Original model. On, the contrary, the estimated average value of λ^* in the Extended model is almost six times that in the Original model: this reflects the multicollinearity discussed above. The standard deviation of λ^* is small as compared to its mean, so that the optimal values found across the 100 train-test splits were very similar to each other.

In terms of performance, the Ridge estimator produces a Test MSE that is 5.3% lower in the Extended (2.65) than that in the Original model (2.79). The new explanatory variables provide more-detailed information on the socioeconomic determinants of individual well-being, including marital status and wealth (approximated by housing-tenure status and the number of rooms in the household). In the Original model, these latter were limited to the explanatory variables of Has a Partner and Log Income. In order to estimate a coefficient for each of the categories of each categorical explanatory variable, we have however had to introduce bias into the estimates, in that the Ridge coefficients are biased estimates of the true β. A better way of describing the determinants of subjective well-being more thoroughly appears in the discussion of the Shapley Values in the Random Forest in Section 4 below.

3.1.3 Variable Selection and Lasso Regression

An alternative to the Ridge is the *Lasso* regression (Tibshirani, 1996). The empirical loss function here comes from setting $\alpha = 1$ in (1):

$$\min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 + \lambda \sum_{j=1}^{k} |\beta_j|. \tag{10}$$

The Lasso minimization problem in (10) may have multiple solutions, although they always produce the same predicted values, so that the Test MSE remains a valid measure of the quality of fit (Tibshirani, 2013). Outside of some particular cases, no closed-form expression for the Lasso estimator exists. There are a number of numerical methods of solving (10), including *Coordinate Descent*, the method used in the *glmnet* package of R. Additional details on Coordinate Descent and other solution techniques can be found in Friedman *et al.* (2010) and Van Wieringen (2020).

The key characteristic of the Lasso penalization is that it induces *variable selection*: even with not particularly large values of λ, one or more of the $\hat{\beta}_{Lasso,j}$ may be shrunk to 0; this is only the case for the Ridge estimator when the estimated coefficients were already zero in the OLS

estimation without penalization. The difference between the two approaches reflects the shapes of the constraints imposed on the estimates by the two penalizations. A more detailed explanation can be found in Hastie *et al.* (2009, Ch.3).

The optimal λ^* values were obtained using the same procedure as described above for the Ridge estimator. The results appear in Table 3, which also lists the number of *non-zero* coefficients associated with the optimal cross-validated λ^*. The figures refer to standardized values and show the means and standard deviations over 100 different random train–test splits.

Table 3. The Performance of the LASSO Regression

	Training MSE	**Test MSE**	**λ∗**	**Non-zero coefficients**
Original	2.78 (0.03)	2.79 (0.11)	0.002 (0.002)	9 [out of 9] (0.20)
Extended	2.58 (0.02)	2.64 (0.09)	0.02 (0.004)	51 [out of 97] (5.60)

Notes: These figures show the average performance, optimal λ^* and number of non-zero coefficient figures in a Lasso regression predicting life satisfaction over 100 different train-test splits, each with 80% of the sample in the training set and errors calculated over the remaining 20% of individuals in the test set. λ^* is obtained from 5-fold-cross-validations on the training set. Standard deviations appear in parentheses.

The predictive performance of the Lasso regression is comparable to that of the Ridge regression, and the same conclusions regarding bias and variance, overfitting and underfitting as in the OLS and Ridge case apply.

In the Original model, shrinkage to 0 was confined to one explanatory variable out of the 9 (8 plus the constant) in four cases out of the 100 train-test splits. On the contrary, in the Extended model an average of 46 coefficients (out of 97) were shrunk to 0.

As for the Ridge estimator, the Lasso estimator solves the numerical multicollinearity issues found for the OLS estimator in the Extended model when we did not drop the dummy associated with the most populous category and all categories with fewer than 15 individuals. The 16 new explanatory variables (with their 97 associated categories) yield a greater predictive accuracy of 5.7% in testing.

While the performances of the Ridge and Lasso estimators are then comparable, the latter has the advantage of automated explanatory-variable selection via the shrinkage to zero. This may help reduce model complexity, further reducing its variance and making it easier to interpret. We

nonetheless may still wish to obtain estimates for all of the coefficients, after explanatory-variable selection has been carried out *ex ante*. In general, Tibshirani (1996) concludes that with $n > k$ (*i.e.* more observations than independent variables) the Ridge estimator outperforms the Lasso estimator. Furthermore, if two explanatory variables are collinear, the Lasso estimator does not shrink both of the associated $\hat{\beta}_{Lasso,j}$ coefficients, but rather only one of them. As such, Lasso does not have the desirable *Grouping Effect*, where two highly-correlated explanatory variables should attract similar estimated coefficients (and identical coefficients in absolute value if the two are perfectly correlated: see Zou and Hastie, 2005).

The *Elastic Net*, first developed by Zou and Hastie (2005), is considered to overcome the weaknesses of the Lasso estimator, but retains its attractive explanatory variable-selection property.

3.1.4. Between Ridge and Lasso: The Elastic Net

The general Elastic Net minimization problem in Zou and Hastie (2005) was set out in Equation (1) above, of which OLS, Ridge and Lasso are special cases. In general, the estimator that solves this problem is

$$min_{\boldsymbol{\beta}\in R^k} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 + \frac{\lambda_2}{2}\sum_{j=1}^{k}\beta_j^2 + \lambda_1 \sum_{j=1}^{k}|\beta_j| \qquad (11)$$

where $\alpha \in (0,1)$ in Equation (1) is the ratio of λ_1 over $\lambda_1 + \lambda_2$, and thus shows the relative weights given to the two types of penalization.

Were we to optimize over pairs of (λ_1, λ_2), we may find the same cross-validated log-likelihood for two different pairs and thus not be able to distinguish between them: the same log-likelihood can come from a very sparse model in which more coefficients are shrunk to 0 ($\lambda_1 >> \lambda_2$) or one that is not sparse ($\lambda_2 >> \lambda_1$). We thus instead optimize over α, rephrasing the Elastic Net minimization problem in (11) as that in (1). The introduction of α allows us to tune the model over the pairs (λ, α). We here consider three possible values for α, 0.25, 0.50 and 0.75, hence either giving 3/4 of the weight to one of the two forms of penalization or weighting them equally. The results are listed in Table 4.

Table 4. The Performance of the Elastic Net Regression

	$\alpha = 0.25$				$\alpha = 0.50$				$\alpha = 0.75$			
	Train MSE	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$	**Train MSE**	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$	**Train MSE**	**Test MSE**	$\lambda*$	$\widehat{\beta_J} \neq 0$
Original	2.78 (0.03)	2.79 (0.11)	0.007 (0.003)	9 (0.14)	2.78 (0.03)	2.79 (0.11)	0.004 (0.003)	9 (0.20)	2.78 (0.03)	2.79 (0.11)	0.003 (0.002)	9 (0.17)
Extended	2.58 (0.02)	2.64 (0.09)	0.07 (0.01)	54 (5.35)	2.58 (0.02)	2.64 (0.09)	0.04 (0.01)	52 (5.37)	2.58 (0.02)	2.64 (0.09)	0.03 (0.01)	51 (6.01)

Notes: These figures show the average performance, optimal λ^* and number of non-zero coefficient figures in three elastic-net regressions predicting life satisfaction. 100 different train-test splits are carried out, each with 80% of the sample in the training set and the error calculated on the remaining 20% of individuals in the test set. The values of α are *ex-ante* fixed and reflect the relative weights on the two penalization terms. λ^* was obtained via 5-fold cross-validations on the training set. Standard deviations appear in parentheses.

As can be seen in Table 4, the three variants of the Elastic Net we consider do not yield much improvement in terms of predictive performance over the Ridge or Lasso regressions. From the $\widehat{\beta_J} \neq 0$ columns, there is shrinkage for over 40 explanatory variables in all three Elastic-Net estimations.

Our main conclusion from considering penalized and non-penalized linear regressions is then that there is no reason to believe that the linear non-penalized regression overfits the Original data and, given the reliability of the estimates in the training data with no evidence of harmful multicollinearity, it is probably preferable to avoid introducing bias. Conversely, in the Extended dataset, the 16 additional explanatory variables improve the Test Set performance with a reduction in the MSE of 5.3%. Moreover, while in the Original dataset the training and testing accuracy were almost identical, in the Extended model we observe a Training MSE that is 2.3% lower than the Test MSE.

We next introduced penalization, and retained all of the dummies in the analysis. We do not observe any additional improvement here: the Test MSE for the Ridge estimator is 5.3% lower in the Extended (2.65) than in the Original model (2.79), as was the case for the non-penalized regression. In general, fitting a multicollinear linear regression can be of interest in any case, as we may wish to assess the marginal effects of some explanatory variables while adding other (possibly correlated) controls. Moreover, the addition of (relevant) multicollinear explanatory variables can in theory still lead to improved test accuracy, and hence a fuller model to interpret (although this is not the case in the data that we analyze here).

In what follows, we move beyond linear estimation to the next algorithm in the interpretability-complexity trade-off: *Regression Trees* and their *ensemble*, the *Random Forest*. For the latter, we will explore two *Model-Agnostic Interpretable Algorithms – Permutation Importance* and *Shapley Values* – that will help us to interpret the results.

3.2 Regression Trees and Random Forest: Stratifying the Explanatory Variable Space

Classification and Regression Trees have a considerable history. The Regression Trees we now turn to were presented in Breiman *et al.* (1984). The overall idea is to divide the explanatory variable space into *J* distinct and disjoint sets, the *terminal nodes* or *leaves* of the tree. The dependent variable value for each individual in a leaf is the mean of the dependent variable of all the individuals who are in the same leaf. Individuals fall into a leaf by moving along one of the branches of the tree, depending on values of their explanatory variables.

The subsequent splits along the *branches* of the tree define the *internal nodes* obtained by *recursive binary splitting*. Starting from the top of the tree – at which point every individual belongs to the same set (so that this is a *top−down* approach) - a *greedy* procedure is implemented, where the preferred split is that which is the best at that specific point, independent of any subsequent steps.

These procedures tend to overfit the training set, producing deep trees with too-long branches, and so produce estimators with high variance and low bias. There will be only few training individuals in each of the final leaves, and a poorly-defined outcome variable, $\hat{y}_{t_{k,train}}$. For this reason, *Random Forests (ensembles* of trees) are preferred, along with regularization criteria for each tree.

Random Forests are constructed via *bootstrap aggregation*, which can be either *non-parametric* or *parametric*. In the former case, no assumptions are made regarding the data-generating process, and new observations are constructed by sampling with reintroduction from the training set. On the contrary, in the latter we assume a well-defined parametric model for the data-generating process.

As we are looking for evidence *against* the linearity (parametric) assumption, we consider non-parametric bootstrapping, which is the general practice in the applied Random Forest

literature. *Bootstrap Aggregation* or *bagging* consists in averaging the prediction of B fitted models, each labelled b, over the S^b different bootstrapped samples, with the aim of reducing the variance of the final estimator.

The entire Random Forest, and each Regression Tree in it, has the same expected value, and hence the same bias. As Tibshirani (2013, p.596) notes, "Increasing the number of trees does not cause the Random Forest sequence to overfit".

A key element in the lower variance is the number m of explanatory variables used for the split at each internal node of each tree, $m \leq k$, where k is the total number of explanatory variables. The correlation between two generic trees in the forest rises with m, although the bias falls with m.

One of the most interesting features of Random Forests is the possibility of *leaving categorical and ordinal explanatory variables as they are, without creating dummies.* We now present the Random Forest results for both the Original and Extended models. The average predictive performance continues to be calculated over 100 Random Forests with 100 different random train–test splits. In each of these, 400 trees were constructed with non-parametrically bootstrapped data. The procedure differs from that in the Penalized Linear Regressions, where we looked for the optimal λ^* *in each train-test split*. Conversely, the optimal structure of the trees in the forest was established, via 5–fold cross–validation, on a single train–test split (the first) using 4000 trees. The penalizations used were the number of explanatory variables at each split, the maximum depth of the branches and the minimum number of training individuals per leaf. The Shapley Values, describing the marginal effects of the different explanatory variables at an individual level, are instead calculated considering only the Random Forest in train-test split 1. The results are presented in Table 5.

3.2.2 Random Forest: Results

Cross-validation was used as the optimizing strategy, so as to be consistent with the linear regressions. Table 5 (fourth column) shows that the algorithm always prefers a random subset of the explanatory variables over including them all - in order to avoid overfitting - and over considering one variable only - which would have been too restrictive. More precisely, the

algorithm considers a subset composed of only the (rounded) square root of the number of all of the variables, which latter is a rule-of-thumb value to trade-off between overfitting and underfitting. Regarding the maximum depth of each branch of each tree, longer trees are unsurprisingly required in the Extended dataset of 21 explanatory variables, given the potential for more-complex relationships.

Table 5. The Performance and the Optimal Hyperparameters of the Random Forest

	Average Training MSE	Average Test MSE	Number of trees	Number of considered explanatory variables per split	Maximum depth of branches	Minimum individuals per leaf
Original	2.67 (0.03)	2.79 (0.10)	400	round(√8) = 3	8	15
Extended	2.19 (0.02)	2.66 (0.10)	400	round(√21) = 5	13	8

Notes: These figures show the average performance of 100 Random Forests over 100 different train-test splits in predicting life satisfaction. The optimal number of explanatory variables to be considered at each split of each tree, the maximum depth of each branch of each tree, and the minimum number of training individuals to be left in each leaf of each tree were *ex-ante* obtained via 5-fold-cross-validation on the first train-test split 1.

Table 6. The Performance of the Random Forest Compared to Linear Regression

	Lin. Reg. MSE Train	Lin. Reg. MSE Test	R.F. MSE Train	R.F. MSE Test	R.F. Improvement in Training Set	R.F. Improvement in Test Set
Original	2.78 (0.03)	2.79 (0.11)	2.67 (0.03)	2.79 (0.10)	4.12%	0%
Extended	2.57 (0.02)	2.65 (0.09)	2.19 (0.02)	2.66 (0.10)	17.35%	-0.38%

Table 6 compares the performance of linear regressions and Random Forests, in both training and testing. We first note a considerable improvement in training set accuracy over the linear regressions of 4.1% and 17.4% in the Original and Extended specifications respectively, while accuracy does not change much in testing. Comparing across both algorithms and specifications, the Extended-model Test MSE in the Random Forest (2.66) is a 4.9% improvement over the Original-model MSE in Unpenalized Linear Regression (2.79).

We now present *Permutation Importance* and the *Shapley Values* calculated for the Random Forest, and a comparison of the latter to the Linear Regression results. As well as discussing the

Random Forest's predictive accuracy, these will allow us to understand how the different explanatory variables affect life satisfaction.

4 Interpreting the Findings: Opening the Black Box

The interpretation of the ML results requires additional calculations beyond fitting, as opposed, for instance, to the interpretation of the explanatory variable coefficients in linear regressions. Model-agnostic tools are used to this end.

The choice of the best model-agnostic interpretability approach depends on a number of factors, including the complexity cost of the algorithm, and whether we are interested in *sparse* or *full* interpretations, or extracting new, derived predictive algorithms from the fitted model (see Molnar, 2019, for details). We will here consider *Permutation Importance* and *Shapley Values*, applied to the results from the Random Forest. We first focus on the *Shapley Values*, as they are interpretable in terms of both their importance - defined via their absolute mean for each explanatory variable - and their marginal effects, and provide a clearer image of the fitted model. *Permutation Importance* instead tells us which explanatory variables, once randomized, most increase the MSE. Last, *Learning Curves* allow us to understand the overall complexity of the underlying data-generating process.

4.1 Shapley Values and TreeSHAP

The Shapley Value is a solution concept from co-operative game theory introduced by Shapley (1951) and formalized in Shapley (1953). The underlying idea is that the way in which a certain sum obtained by a group of players is split depends on how much each member contributes to the outcome.

Applied to Machine Learning, the *game* is the predictive task and the *players* are the different explanatory variables that work together to produce the *gain*, namely the difference between the prediction for a given individual and the "average prediction in the sample" (Molnar, 2019, Chapter 5.9). The *Shapley Value* of an explanatory variable is "the average of all marginal contributions across all possible coalitions of explanatory variables" (Molnar, 2019, Chapter 5.9). Shapley Values are calculated at the individual level. If we have k explanatory variables and we

are interested in calculating the Shapley Value for one of them, say variable j, we will consider all of the possible 2^{k-1} coalitions of the remaining $k - 1$ explanatory variables.

In each of these 2^{k-1} coalitions, we calculate the difference between the predicted value *with* and *without* the value of the j^{th} explanatory variable for individual i, $x_{i,j}$. This reveals the *marginal contribution* of the explanatory variable j in predicting the dependent variable. The values of the explanatory variables that do not appear in a coalition are *eliminated,* by randomly replacing individual *i's* value of that explanatory variable with that of another individual. The Shapley Value for explanatory variable j for individual i is then the weighted average of its marginal contributions across all of the 2^{k-1} coalitions, with the weights depending (in a U-shaped way) on the number of explanatory variables included in the coalitions.

Formally, define $\boldsymbol{x}_i$ as the vector of explanatory variables for individual i, and $\{x_{i,1}, \ldots, x_{i,k}\}$ as the set of all of the values of the k explanatory variables considered for i. Let S be the coalitions of players considered in a given step - that is, the coalition of explanatory variables used in the model - and $f: 2^{k-1} \rightarrow \mathbb{R}$ a *value function*. The Shapley Value of the explanatory variable j for individual i is formally defined as:

$$\phi\ \left(x_{i,j}\right) = \sum_{S \subseteq \{x_{i,1},\ldots,x_{i,k}\} \setminus \{x_{i,j}\}} \frac{n(S)!(k-n(S)-1)!}{k!} [f_{\boldsymbol{x}_i}\left(S \cup \{x_{i,j}\}\right) - f_{\boldsymbol{x}_i}(S)]. \qquad (12)$$

The value taken by explanatory variable j for individual i then contributes $\phi\ \left(x_{i,j}\right)$ "to the prediction of this particular instance compared to the average prediction for the dataset" (Molnar, 2019, Chapter 5.9).

It is immediate to see that the calculation of Shapley Values is costly, as we calculate values for 2^{k-1} coalitions *for every individual* in the sample and *for every explanatory variable.* A number of ways of addressing this issue have been proposed, including Monte Carlo sampling by Štrumbelj *et al.* (2014).

We here consider the *TreeSHAP* algorithm of Lundberg *et al.* (2018), where the value function is the expected value of the prediction conditional on the explanatory variables in the coalition S: $f_{x_i}(S) = E\ [f(x_i) \mid S]$. The direct estimation of $f_{x_i}(S)$ would have computational complexity of $O(BL2^k)$, where B is the number of trees in the forest, L the maximum number of final leaves in any tree, and k the number of explanatory variables. The *TreeSHAP* algorithm greatly reduces the computational complexity to $O(BLD^2)$, where D is the maximum depth of any tree.

The key measure that can be derived from Shapley Values is the *Shapley Feature Importance,* that is, the mean absolute value of the Shapley Values for variable *j* calculated over all of the *i* individuals in the training set:

$$I_{Shap}(X_j) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} |\phi(x_{i,j})|. \quad (13)$$

We calculate $I_{Shap}(X_j)$ for each explanatory variable in each of the 100 train-test splits, and average these to produce *Average Mean Absolute Shapley Values* with their associated standard deviations. Formally, labelling the different train-test splits as *train(1), train(2),…,train(100)*, this average value is given by:

$$Avg[I_{Shap}(X_j)] = \frac{1}{100} \sum_{t=1}^{100} \frac{1}{n(train(t))} \sum_{i=1}^{n(train(t))} |\phi_t(x_{i,j})|, \quad (14)$$

where $n(train(t)) = 7093$ (*i.e.* 80% of the sample size of 8867) in all the 100 splits, and $\phi_t(x_{i,j})$ represents the Shapley Value of explanatory variable *j* for training individual *i* in the t^{th} training set. The results appear in Figure 1 and Table 7 for the Original model, and Figure 3 and Table 8 for the Extended model.

4.1.1 Average Mean Absolute Shapley Values: Original Model

The Average Mean Absolute Shapley Values are depicted in Figure 1: the most important explanatory variable is the composite variable "Has a Partner". This changes the absolute predicted value of life satisfaction by on average 0.36 over the 100 train–test splits; the second most important explanatory variable is Emotional Health, with an average effect of 0.19.

Figure 1: Average Mean Absolute Shapley Values in the Original Model.

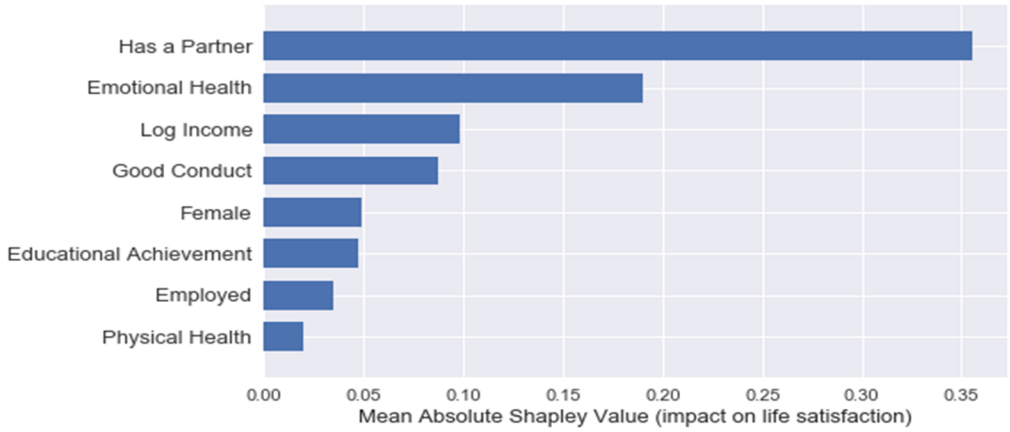


Table 7. Average Mean Absolute Shapley Values in the Original Model

Explanatory Variables	Average MASV	SD MASV
Has a Partner	0.36	0.01
Emotional Health	0.19	0.01
Log Income	0.10	0.01
Good Conduct	0.09	0.01
Female	0.05	0.01
Educational Achievement	0.05	0.01
Employed	0.03	0.00
Physical Health	0.02	0.00

Notes: This table shows the Average Mean Absolute Shapley Value (MASV) for each explanatory variable calculated over the same 100 different train–test splits considered in the Random Forests. Original model. Standard deviations are in parentheses.

The Shapley Values can also tell us in which direction the explanatory variables affect the findings. The values presented below refer *to one Random Forest only (that calculated on train-test split 1)*. Nonetheless, given that the performance of this Random Forest and the average over all 100 forests are similar, the results there are generalizable. The rankings of the Average MASVs in Table 7 (calculated over the 100 Random Forests) and those in Figure 2 below are also similar.

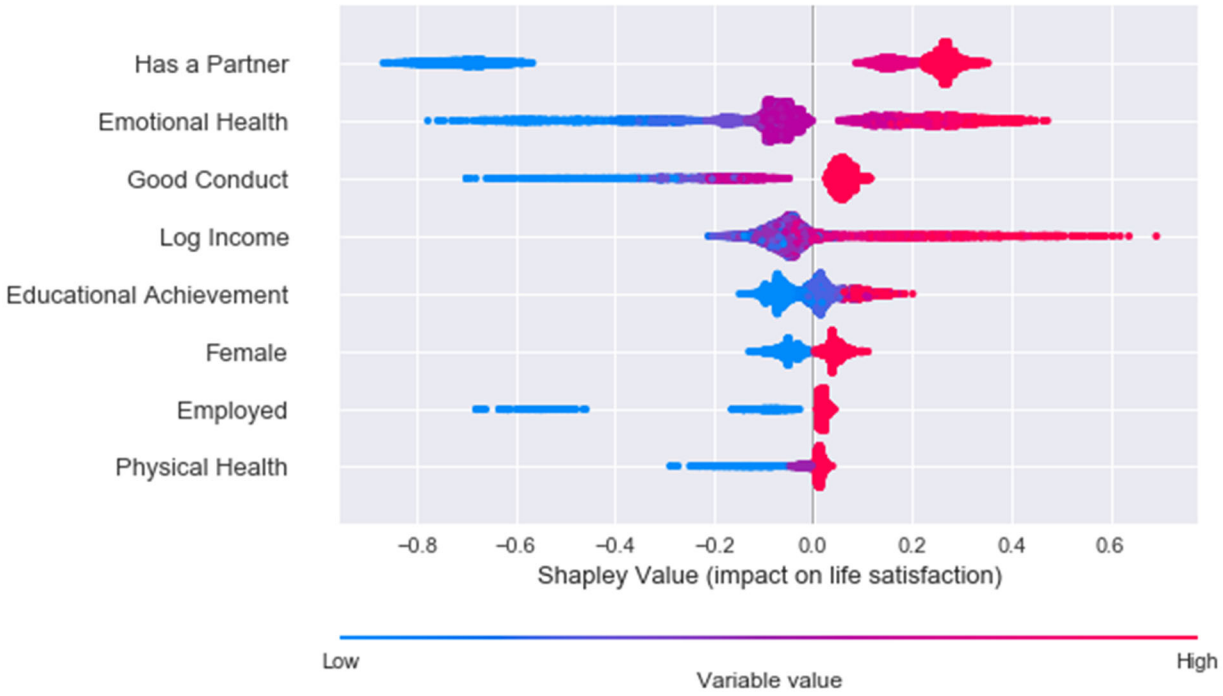
The dots depicted in Figure 2 are the Shapley Values by individual by explanatory variable, the $\phi \ \ (x_{i,j})$ in Equation (12), with the explanatory variables on the vertical axis and the Shapley Values on the horizontal axis. The explanatory variables are ranked from the most Shapley Important (Has a Partner) to the least (Physical Health), as shown in Figure 1.

The colors of the dots reveal whether the explanatory variable for that individual has a high or low value, ranked by color intensity ranging from red (high) to blue (low). Overlapping dots create 'clouds' that help to illustrate the distribution of the Shapley Values.

The patterns in Figure 2 allow a more-detailed understanding of the average absolute values plotted in Figure 1. Consider, for instance, Has a Partner, which is the most important explanatory variable: the two highest values of this variable, from Section 2, are 0.685 and 0.530, for being married with and without children respectively. The associated Shapley Values for these two highest values of Has a Partner in the first row of Figure 2 are represented by the red and purple dots, respectively. As can be seen, the Shapley Values of the Has a Partner variable are mostly

clustered in the [0.1, 0.35] or [–0.9, –0.5] intervals: being Married with or without children increases life satisfaction, on average, by 0.1 to 0.35 points relative to the "average prediction for the dataset" (Molnar, 2019). Conversely, the two lowest values that the Has a Partner variable takes, 0 and -0.004 (for being single with and without children respectively), correspond to the blue Shapley Values and are associated with lower life satisfaction of 0.5 to 0.9 points.

Figure 2: Shapley Values by individual by explanatory variable – Original Model



Notes: The dots in each line represent the Shapley Values (as shown on the horizontal axis) for each individual for the variable indicated. The redder dots refer to higher values of the explanatory variable in question, and the bluer dots to lower values. Shapley Values at the individual level are calculated from the Random Forest fitted on training-test split 1.

The results are even more interesting for Emotional Health. As this explanatory variable is more continuous, the Shapley Values are distributed more uniformly. Having a high value of Emotional Health increases life satisfaction by 0.1 to 0.45 points. There is also a long left tail: predicted life satisfaction can be up to 0.8 points lower for the individuals with the lowest values of emotional health.

Criminality (Good Conduct) is the third-most important variable. The highest value here is for those who reported no crimes. As is evident from the figure, having no criminal record has only a small impact on predicted life satisfaction; instead, having committed crimes can sharply reduce satisfaction by up to 0.7 points. The logic here is that while no criminal record is normal (and so does not make the individual much more satisfied with life), having reported crimes is associated with sharply lower satisfaction. The same pattern is found for being employed and good physical health: being employed and not having health problems do not have positive effects on life satisfaction, but the lack of them (being unemployed or having health problems) has a sizeable negative effect. Health problems having such a large effect may reflect the relatively young age (34) of our sample.

Last, low income does not strongly negatively affect life satisfaction (the majority of the blue-dot Shapley Values are close to zero), but there is a large positive impact of higher income, of up to 0.7 points.

This ranking of explanatory variables is important for policy. Population life satisfaction can then be improved by focusing on the individuals in the left tails of the Shapley Values. Here Emotional Health, Family situation, Unemployment and Criminality appear central, as the explanatory variables associated with the largest drops in life satisfaction.

4.1.2. Average Mean Absolute Shapley Values: Extended Model

Figure 3 and Table 8 show the results for the Extended model. Marital Status and Emotional Health behave similarly to Has a Partner and Emotional Health in the Original model. The individual Shapley Values for this extended set of variables, analogous to those for the Original model in Figure 2, appear in Figure 4. Many of these variables seem to have a systematic relationship with life satisfaction, as revealed by the separate clusters of dots according to the variable's different values (and the color of the individual dots).

Figure 3: Average Mean Absolute Shapley Values in the Extended Model.

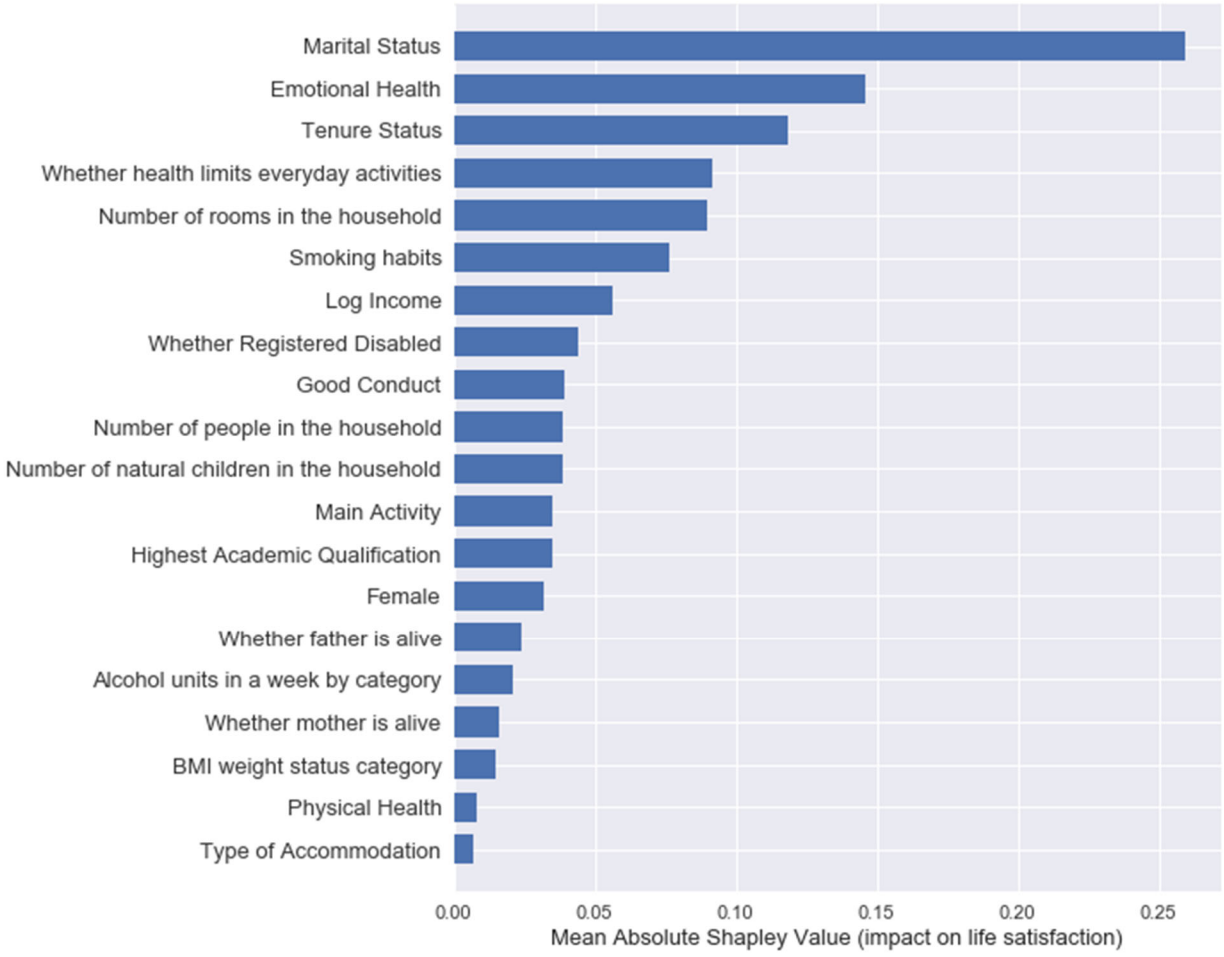
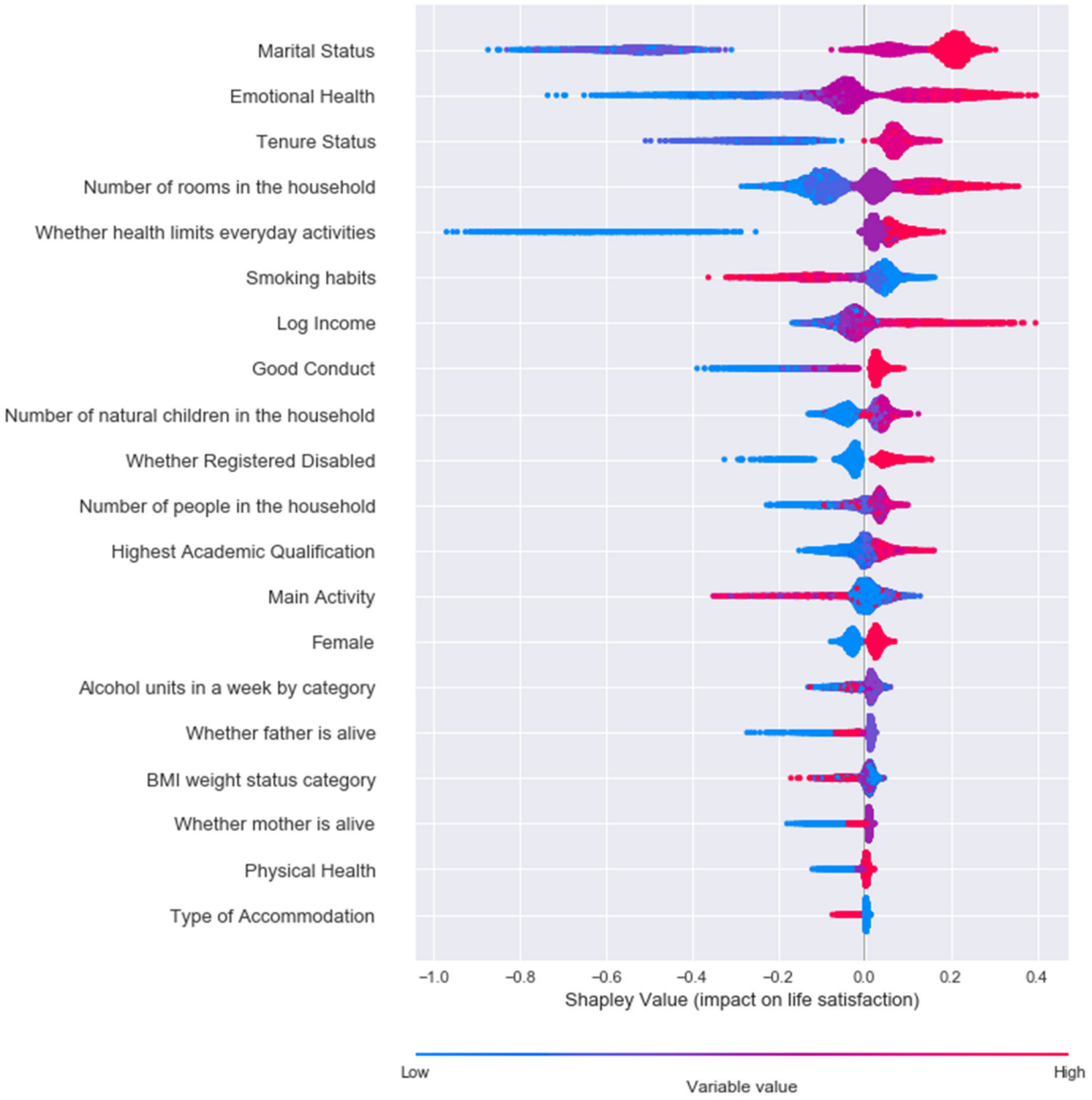


Table 8. Average Mean Absolute Shapley Values in the Extended dataset

Explanatory variable	Average MASV	SD MASV
Marital Status	0.26	0.008
Emotional Health	0.15	0.009
Tenure status	0.12	0.009
Number of rooms in the household	0.09	0.008
Whether health limits everyday activities	0.09	0.006
Smoking habits	0.08	0.007
Log Income	0.06	0.004
Number of natural children in the household	0.04	0.005
Number of people in the household	0.04	0.003
Good Conduct	0.04	0.004
Whether registered disabled	0.04	0.006
Main Activity	0.04	0.003
Highest academic qualification	0.03	0.005
Female	0.03	0.004
Whether father is alive	0.02	0.003
Alcohol units in a week by category	0.02	0.003
Whether mother is alive	0.02	0.003
BMI category	0.01	0.002
Type of accommodation	0.01	0.002
Physical Health	0.01	0.001
Number of non-natural children in the household	0.01	0.001

Notes: This table shows the Average Mean Absolute Shapley Values in the Extended dataset. Standard deviations appear in the right-hand column.

Figure 4: Shapley Values by Individual by Explanatory Variable – Extended Dataset



Notes: The dots in each line represent the Shapley Values (as shown on the horizontal axis) for each individual for the variable indicated. The redder dots refer to higher values of the explanatory variable in question, and the bluer dots to lower values. Shapley Values at the individual level are calculated from the Random Forest fitted on training-test split 1.

Marital Status in the Extended model is a different variable from Has a Partner in the Original model, as it now does not include the presence of children (children appear in a separate variable),

and takes on more values than simply Single or Married, now including Separated, Divorced, and Widowed (which are assigned the values of 3, 2 and 1 respectively, the lowest values for this variable). There is wide variation in the marginal effects for marital status, where the highest values (representing Married and Cohabiting, with values of 6 and 5) have a positive impact of up to 0.3 life-satisfaction points, but Single, Separated, Divorced or Widowed have large negative effects of 0.3 to 0.9 points.

Health limiting everyday activity has the largest negative impact on predicted life satisfaction, of up to 1 point, and behaves in the same way as Disability and Criminality (Good Conduct). Physical health, which is towards the bottom of Figure 4, has almost no effect on life satisfaction. We might wonder whether this reflects the inclusion of both disability and health limitations in the Extended Model. However, dropping these latter two continues to produce only very small Shapley Values (as illustrated in Figures 1 and 2, where this is the only physical-health variable). Our age-34 respondents report only few of the 15 health conditions in Appendix B: over-three quarters have none, and only 5% report two or more.

The impact of Emotional Health is again more-continuously distributed, with a large effect as illustrated in Figure 3. Some of the other explanatory variables are of more marginal importance, including gender, education, number of children, number of people in the household, and the type of accommodation. The first two of these were equally relatively unimportant in the Original Model.

4.2 Comparing Mean Absolute Shapley Values to the Linear Regression Coefficients

The MASV associated with an explanatory variable is its average absolute marginal effect on the predicted dependent variable. This measure is intuitively comparable to the coefficients from linear regression, which also reflect the marginal effect of a unitary change in the explanatory variable on the dependent variable. We here compare the two, taking only the Random Forest with 4000 trees fitted on training set 1. Insignificant coefficients (p-values > 0.05) are reported as 0. We start with the Original model.

4.2.1 Shapley Values and Regression Coefficients: Original model

It is intuitive to compare the MASVs, which reflect the mean absolute marginal impact of each explanatory variable, to the absolute linear regression coefficients. The results appear in Table 9, where the variables are ranked by MASV. The ranking in the two columns is identical for the continuous variables (which are all standardized). The comparison between the two columns is more difficult to carry out for Employed and Female, as these two coefficients are not standardized. The estimated coefficients are therefore larger than they would have been had the variables been standardized. On the other hand, standardization has no impact in Random Forests.

Table 9. Random Forest Mean Absolute Shapley Values and Absolute Linear Regression Coefficients – Original Model

Explanatory variable	MASV	|Coefficients|
Has a Partner	0.355	0.470
Emotional Health	0.177	0.293
Good Conduct	0.096	0.134
Log Income	0.092	0.117
Ed. Achievement	0.051	0.078
Female	0.047	0.216
Employed	0.034	0.988
Physical Health	0.021	0.000

Notes: This table compares Mean Absolute Shapley Values calculated from the optimized Random Forest to the Absolute Linear Regression Coefficients. All variables are standardized but the Employed and Female dummies in the Original model.

4.2.2 Shapley Values and Regression Coefficients: Extended Model

The comparison in the Extended Model is less straightforward. While the Shapley Values in this case can be interpreted in the same way as for the Original Model, this is not the case for the Ridge Regression Coefficients, as in the Extended Model we have added multiple (ordinal) multiclass categorical explanatory variables that are divided into dummies. We thus require a unique measure for these explanatory variables that is comparable to the MASVs from all of the coefficients on the associated dummies. We here choose the absolute weighted mean coefficient over all of the associated dummies, with the weights being the fraction of individuals in each of

the explanatory-variable categories. In this case, since the coefficients are from a Ridge regression, they also are standardized.

Formally, suppose that the explanatory variable X_j is a multiclass categorical variable with k categories, split into k dummies for the Ridge Regression. Let $\chi_{j,l}$ be the proportion of individuals in the training set in the l^{th} category of explanatory variable j:

$$\chi_{j,l} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} I\left(x_{i,j} = l\right) \tag{15}$$

where $I(x_{i,j} = l)$ is the indicator function with value 1 if individual i belongs to the l^{th} category of the j^{th} explanatory variable, and 0 otherwise. Then, given the $\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,k}$ estimated Ridge Regression coefficients, the *Derived Coefficient* is:

$$\hat{\beta}_j = \sum_{l=1}^{k} \chi_{j,l} \left|\hat{\beta}_{j,l}\right| \ . \tag{16}$$

We only carry out this calculation for the multiclass categorical explanatory variables (which are indicated by underlined coefficients in the final column below). Numerical discrete variables (whether binary, such as Female, or with multiple values, like Number of People in the Household), and the variables that are treated as numerical continuous (Log Income and Emotional Health) enter the Ridge Regression as they are, and the absolute coefficient in the table below is entered directly from the regression output.

Table 10. Random Forest Mean Absolute Shapley Values and Absolute Ridge Regression Coefficients – Extended Model

Explanatory variable	**MASV**	**|Coefficients|**
Marital Status	0.260	0.292
Emotional Health	0.138	0.190
Tenure Status	0.113	0.105
Number of rooms in the household	0.100	0.110
Whether health limits everyday activities	0.089	0.104
Smoking Habits	0.066	0.076
Log Income	0.055	0.067
Good Conduct	0.047	0.060
Number of natural children in the household	0.043	0.001
Registered disabled	0.041	0.050
Number of people in the household	0.039	0.022
Main Activity	0.031	0.090

Highest academic qualification	0.031	0.048
Female	0.029	0.136
Alcohol units in a week by category	0.024	0.054
Father is alive	0.022	0.090
BMI category	0.018	0.034
Mother is alive	0.017	0.047
Physical Health	0.008	0.009
Type of Accommodation	0.007	0.009
Number of non-natural children in the household	0.006	0.029

Notes: This table compares Mean Absolute Shapley Values calculated from the optimized Random Forest to the Absolute Ridge Regression Coefficients. The results are from the Extended model. The underlined coefficients in the final column are calculated using Equation (16).

In Table 10, the values of the multiclass categorical explanatory variables (calculated via Equation (16)) are underlined. The two most important variables in both columns are marital status and emotional health. The most-notable difference between the two columns of Table 10 is the estimated effect of Female (which is standardized in Ridge Regression): here the MASV is more than four times smaller than the associated Ridge coefficient. In the Ridge Regression, Female is the third most-important explanatory variable. But in terms of MASVs it is only the 14th most-important explanatory variable However, the estimated Ridge Regression coefficients should perhaps be taken with a grain of salt, as there is some risk that they overestimate the expected marginal impact of the explanatory variables on the dependent variable, given the assumed linearity of the dependent variable in the parameters and, under the linearity assumption, their bias. We conclude this section by discussing Permutation Importance, to assess the impact of each explanatory variable in determining the model's predictive accuracy.

4.3 Permutation Importance

The idea of Permutation Importance is simple. Once we have randomized, via shuffling, one of the explanatory variables in the test set, say the j*th*, its Permutation Importance is defined as the difference between the scoring metric that we consider (in our case, the MSE) calculated from the actual X_j and its shuffled version, X_{j^*} , keeping all of the other variables unshuffled at their original values. This operation is performed multiple times, and Permutation Importance is then

calculated as the average difference in the scoring metric across the multiple repetitions. While this operation can be carried out for both the test and training sets (see Breiman, 2001), we here consider only the Test Set, as this represents a diagnostic measure of predictive accuracy. The results refer to the Random Forest on train-test split 1.

Table 11. Random Forest Permutation Importance – Original Model

Explanatory variable	Weight (Standard Deviation)
Has a Partner	0.113 (0.011)
Emotional Health	0.043 (0.006)
Log Income	0.024 (0.005)
Good Conduct	0.013 (0.004)
Employed	0.009 (0.002)
Female	0.004 (0.002)
Physical Health	0.004 (0.002)
Educational Achievement	0.002 (0.001)

Notes: This table shows Permutation Importance calculated on the Test Set of the Original Model considering the best-performing Random Forest, measuring the fall in predictive accuracy across 100 shuffles of each explanatory variable. The figures in parentheses are standard deviations.

Table 12. Random Forest Permutation Importance –Extended Model

Explanatory variable	Weight (Standard Deviation)
Marital Status	0.064 (0.007)
Whether health limits everyday activities	0.028 (0.005)
Emotional Health	0.027 (0.004)
Log Income	0.011 (0.003)
Main activity	0.010 (0.002)
Tenure Status	0.010 (0.003)
Smoking habits	0.007 (0.002)
Number of rooms in the household	0.006 (0.003)
Good Conduct	0.004 (0.002)
Number of natural children in the household	0.003 (0.001)
Whether Registered Disabled	0.003 (0.002)
Number of people in the household	0.003 (0.002)
Whether father is alive	0.002 (0.001)
Female	0.002 (0.001)
Whether mother is alive	0.001 (0.001)
Highest Academic Qualification	0.001 (0.001)
Alcohol units in a week by category	0.001 (0.001)

Type of Accommodation	0.000 (0.000)
Number of non-natural children in the household	0.000 (0.000)
Physical Health	0.000 (0.000)
BMI weight status category	0.000 (0.001)

Notes: This table shows Permutation Importance calculated on the Test Set of the Extended Model considering the best-performing Random Forest, measuring the fall in predictive accuracy across 100 shuffles of each explanatory variable. The figures in parentheses are standard deviations.

The first intuitive finding from Tables 11 and 12 is that, in the Original Model with 8 explanatory variables, the average marginal impact of randomizing explanatory variables on predictive accuracy is greater than in the richer Extended Model with 21 explanatory variables. It is also clear that Permutation Importance is not monotonic with respect to the cardinality of the explanatory variable. Take, for example, Has a Partner and Log Income in the Original model. The former takes on only 4 different values, while the latter is continuous. Hence, when randomizing (shuffling) the former, the probability that an individual's shuffled value is the same as their original value is higher, which in turn should mechanically reduce its Permutation Importance. Nonetheless, the Permutation Importance of Has a Partner is almost five times higher than that of Log Income: Permutation Importance then does capture the actual importance of an explanatory variable in predicting life satisfaction, rather than simply modeling the noisy characteristics of the explanatory variable itself, such as its cardinality.

5 Conclusions

In this paper we have constructed a predictive model for life satisfaction using data from the British Cohort Study (BCS). We evaluate the predictive performance of our models relative to the benchmark OLS regression in Layard *et al.* (2014). We first use only the eight original adult variables that appeared there (with a different version of self-assessed physical health, as updated in Clark and Lepinteur, 2019), and then turn to an Extended model that has 21 explanatory variables: 5 of the original 8, plus 16 new variables (some of which are more-detailed versions of the other 3 of the original 8). Splitting these categorical variables up into their separate values produces 96 dummy variables.

We found no evidence of improvement in model fit using more-advanced ML methods. In the Extended model, we first have to penalize the linear models due to numerical problems including

multicollinearity, or exclude from the analysis some of the least-populated categories. The Extended Model with the 16 new explanatory variables allows us to improve the predictive accuracy, in testing, by 5.3% in terms of a lower Average Test MSE figure.

The best-optimized Random Forest produced no improvement over the Penalized Linear Regressions on the test set in the Extended Model.

Last, to help interpret the importance of the different explanatory variables in the prediction of life satisfaction, we considered two model-agnostic interpretability tools applied to the Random Forest: Permutation Importance and Shapley Values. The latter allows the comparison of the machine-learning results to the estimated coefficients from Penalized Linear Regressions.

Shapley Values assess the marginal impact of the (significant) different explanatory variables at the individual level. In other words, Shapley Values do not pick up the *average* effect of a one-unit change in the explanatory variable (as for the coefficients of a linear regression model) but the marginal impact of *every single value of that explanatory variable*. Another advantage of using a Machine Learning algorithm like Random Forest, where the explanatory variables do not need to be split in dummies (as long as they are ordinal), is that we can take into account the categories that we dropped in the Linear Unpenalized Regression. The comparison of the Random Forest Shapley Values to the estimated Ridge Regression coefficients suggests that some caution should be exercised regarding coefficient size in the latter. This in particular applies to gender: in the Extended dataset, there is a significant difference between the Female MASV and the linear regression coefficient, with the latter being nine times larger than the former. This is in line with Oparina and Srisuma (2022) who, in non-parametric estimation of the measurement error in reported life satisfaction, find a negative relation between female and *latent* life satisfaction (i.e. the true value of the variable), but a positive coefficient for *reported* life satisfaction.

Our work here has considered the subjective judgment of life satisfaction, but we believe that the prediction of objective variables will also benefit from non-linear machine-learning analyses.

Regarding the most important predictors of life satisfaction, our comprehensive analysis confirms that Marital Status as well as Emotional and Physical Health (in terms of limitations to everyday activities) are always the most important explanatory variables, in line with the findings from the existing literature.

References

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, New York: Chapman and Hall, Wadsworth.

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

Clark, A.E., and Lepinteur, A. (2019), "The Causes and Consequences of Early-adult Unemployment: Evidence from Cohort Data", *Journal of Economic Behavior & Organization*, 166, 107–124.

Clark, A.E., Flèche, S., Layard, R., Powdthavee, N., and Ward, G. (2018), *The Origins of Happiness: The Science of Well-being over the Life Course*, Princeton University Press, Princeton NJ.

Clark, A.E., and Oswald, A.J. (1994), "Unhappiness and Unemployment", *Economic Journal*, 104, 648–59.

Clark, A.E., and Oswald, A.J. (1996), "Satisfaction and Comparison Income," *Journal of Public Economics,* 61, 359–381.

Dolan, P., Peasgood, T., and White, M. (2008), "Do We Really Know What makes us happy? A Review of the Economic Literature on the Factors Associated with Subjective Well-being" *Journal of Economic Psychology*, 29, 94–122.

Friedman, J., Hastie, T., and Tibshirani, R., (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent", *Journal of Statistical Software,* 33, 1–22.

Hastie, T., Tibshirani, R., and Friedman, J., (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Helliwell, J.F., Huang, H., and Wang, S. (2016), "The Distribution of World Happiness", *World Happiness Report*.

Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12, 55–67.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, New York: Springer.

Kahneman, D., and Deaton, A. (2010), "High Income Improves Evaluation of Life but Not Emotional Well-being." *Proceedings of the National Academy of Sciences*, 107, 16489–16493.

Kim, B., Khanna, R., and Koyejo, O.O. (2016), "Examples are Not Enough, Learn to Criticize! Criticism for Interpretability". *Advances in Neural Information Processing Systems, 29, 2280–2288.*

Layard, R., Clark, A.E., Cornaglia, F., Powdthavee, N., and Vernoit, J. (2014), "What Predicts a Successful Life? A Life-Course Model of Well-Being", *Economic Journal*, 24, 720–738.

Lundberg, S.M., and Lee, S.I., (2017), "A Unified Approach to Interpreting Model Predictions", Proceedings of the *31st Conference on Neural Information Processing Systems,* 4768–4777.

Lundberg, S.M., Erion, G.G., and Lee, S. (2019), "Consistent Individualized Explanatory Variable Attribution for Tree Ensembles", *Preprint at arXiv:1802.03888.*

Luttmer, E. (2005), "Neighbors as Negatives: Relative Earnings and Well-Being," *Quarterly Journal of Economics*, 120, 963–1002.

Molnar, C. (2019), *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

Oparina, E., and Srisuma, S. (2022), "Analyzing Subjective Well-Being Data with Misclassification", *Journal of Business & Economic Statistics*, 40, 730–743.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011), "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, 12, 2825–2830.

Shapley, L.S. (1953), "A Value for n-person Games", *Contributions to the Theory of Games, Princeton University Press, Princeton,* 2, 307–317.

Štrumbelj, E., and Kononenko, I. (2014), "Explaining Prediction Models and Individual Predictions with Explanatory Variable Contributions", *Knowledge and Information Systems*, 41, 647–665.

Stutzer, A., and Frey, B.S. (2006), "Does Marriage Make People Happy, Or Do Happy People Get Married?", *Journal of Socio-Economics*, 35, 326–347.

Tibshirani, R.J. (2013), "The Lasso Problem and Uniqueness", *Electronic Journal of Statistics*, 7, 1456–1490.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society B*, 58, 267–288.

Winkelmann, L., and Winkelmann, R. (1998), "Why are the Unemployed so Unhappy? Evidence from Panel Data," *Economica*, 65, 1–15.

Zou, H., and Hastie, T. (2005) "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society B*, 67, 301–320.

Appendix A: Descriptive Statistics

Explanatory Variables	Mean	SD	Min	Max
Log Income	9.28	0.598	6.23	12.4
Educational Achievement	0.20	0.251	0	0.75
Employed	0.98	0.130	0	1
Has a Partner	0.48	0.285	0.00	0.66
Good Conduct	24.50	1.699	0	25
Female	0.52	0.500	0	1
Marital Status - Other missing	0.00	0.018	0	1
Marital Status - Married	0.54	0.498	0	1
Marital Status - Cohabiting	0.21	0.404	0	1
Marital Status - Single (never married)	0.19	0.394	0	1
Marital Status - Separated	0.02	0.149	0	1
Marital Status - Divorced	0.03	0.182	0	1
Marital Status - Widowed	0.00	0.037	0	1
Type of Accommodation - Not Applicable	0.01	0.076	0	1
Type of Accommodation - A house or bungalow	0.88	0.326	0	1
Type of Accommodation - Flat or Maisonette	0.11	0.310	0	1
Type of Accommodation - Studio flat	0.00	0.044	0	1
Type of Accommodation - A room / rooms	0.00	0.041	0	1
Type of Accommodation - Something else	0.00	0.057	0	1
Tenure Status - Refusal	0.00	0.065	0	1
Tenure Status - Do not Know	0.00	0.015	0	1
Tenure Status - Own (outright)	0.05	0.221	0	1
Tenure Status - Own - buying with help of a mortgage/loan	0.69	0.462	0	1
Tenure Status - Pay part rent and part mortgage (shared/equity ownership)	0.01	0.067	0	1
Tenure Status - Rent it	0.19	0.393	0	1
Tenure Status - Live here rent-free	0.04	0.185	0	1
Tenure Status - Squatting	0.00	0.015	0	1
Tenure Status - Other	0.02	0.147	0	1
Main Activity - Do not know	0.00	0.015	0	1
Main Activity - Full-time paid employee	0.58	0.494	0	1
Main Activity - Part-time paid employee (under 30 hours a week)	0.16	0.365	0	1
Main Activity - Full-time self-employed	0.08	0.273	0	1
Main Activity - Part-time self-employed	0.02	0.127	0	1
Main Activity - Unemployed and seeking work	0.02	0.137	0	1
Main Activity - Full-time education	0.01	0.092	0	1
Main Activity - On a government scheme for employment training	0.00	0.028	0	1
Main Activity - Temporarily sick/disabled	0.00	0.042	0	1

Main Activity - Permanently sick/disabled	0.02	0.153	0	1
Main Activity - Looking after home/family	0.10	0.302	0	1
Main Activity - Other	0.01	0.108	0	1
Highest Academic Qualification - Do not know	0.00	0.034	0	1
Highest Academic Qualification - None	0.09	0.286	0	1
Highest Academic Qualification - CSE	0.15	0.359	0	1
Highest Academic Qualification - GCSE	0.09	0.289	0	1
Highest Academic Qualification - GCE O Level	0.24	0.428	0	1
Highest Academic Qualification - A/S Level	0.02	0.128	0	1
Highest Academic Qualification - Scottish School Certificate, Higher School Certificate	0.02	0.145	0	1
Highest Academic Qualification - GCE A Level (or S Level)	0.05	0.225	0	1
Highest Academic Qualification - Nursing or other para-medical qualification	0.02	0.128	0	1
Highest Academic Qualification - Other teaching qualification	0.01	0.086	0	1
Highest Academic Qualification - Diploma of Higher Education	0.08	0.267	0	1
Highest Academic Qualification - Other degree level qualification such as graduate membership	0.05	0.217	0	1
Highest Academic Qualification - Degree (e.g. BA, BSc)	0.12	0.325	0	1
Highest Academic Qualification - PGCE-Post-graduate Certificate of Education	0.02	0.135	0	1
Highest Academic Qualification - Higher degree (e.g. PhD, MSc)	0.04	0.205	0	1
Whether Registered Disabled - Do not know	0.00	0.030	0	1
Whether Registered Disabled - Yes	0.02	0.132	0	1
Whether Registered Disabled - No but long-term disability	0.63	0.482	0	1
Whether Registered Disabled - No and no long-term disability	0.35	0.477	0	1
Whether health limits everyday activities - Yes	0.07	0.258	0	1
Whether health limits everyday activities - No but health problems since last interview	0.51	0.500	0	1
Whether health limits everyday activities - No and no health problems since last interview	0.42	0.494	0	1
BMI weight status category - Insufficient data	0.03	0.164	0	1
BMI weight status category - Underweight (< 18.5)	0.01	0.119	0	1
BMI weight status category - Normal (18.5-24.9)	0.47	0.499	0	1
BMI weight status category - Overweight (25-29.9)	0.33	0.470	0	1
BMI weight status category - Obese (30 and above)	0.16	0.368	0	1
Smoking habits - Other missing	0.00	0.011	0	1
Smoking habits - Never smoked	0.45	0.498	0	1
Smoking habits - Ex smoker	0.24	0.425	0	1
Smoking habits - Occasional smoker	0.07	0.246	0	1
Smoking habits - Up to 10 a day	0.09	0.290	0	1
Smoking habits - 11 to 20 a day	0.13	0.337	0	1
Smoking habits - More than 20 a day	0.02	0.144	0	1
Smoking habits - Daily but frequency not stated	0.00	0.026	0	1
Alcohol units in a week by category - Never drinks or only on special occasions	0.19	0.392	0	1
Alcohol units in a week by category - None reported	0.08	0.266	0	1
Alcohol units in a week by category - 1 to 14	0.48	0.500	0	1

Alcohol units in a week by category - 15 to 21	0.10	0.305	0	1
Alcohol units in a week by category - 22 to 39	0.10	0.294	0	1
Alcohol units in a week by category - More than 39	0.05	0.226	0	1
Whether mother is alive - Do not know	0.00	0.032	0	1
Whether mother is alive - Missing	0.00	0.055	0	1
Whether mother is alive - Yes in household	0.07	0.254	0	1
Whether mother is alive - Yes	0.86	0.346	0	1
Whether mother is alive - No	0.03	0.158	0	1
Whether mother is alive - No reported dead last sweep	0.04	0.197	0	1
Whether father is alive - Do not know	0.01	0.105	0	1
Whether father is alive - Missing	0.00	0.051	0	1
Whether father is alive - Yes in household	0.05	0.220	0	1
Whether father is alive - Yes	0.79	0.410	0	1
Whether father is alive - No	0.05	0.218	0	1
Whether father is alive - No reported dead last sweep	0.10	0.300	0	1
Number of people in the household	3.11	1.274	1	10
Number of natural children in the household	1.09	1.090	0	8
Number of non-natural children in the household	0.07	0.357	0	4
Number of rooms in the household	4.70	1.531	1	12
Physical Health	0.30	0.610	0	4
Emotional Health	0.83	0.119	0	1

Appendix B: Physical Health

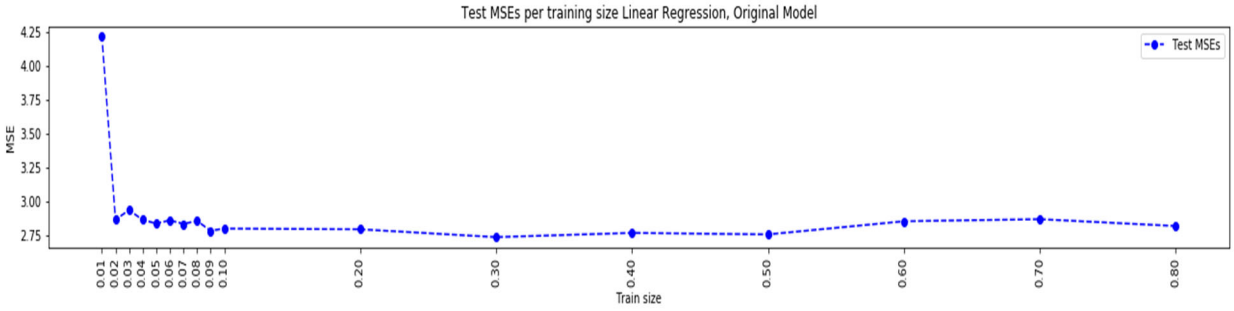
Physical Health
Please tick all that apply. Have you suffered from any of these…
Hay Fever
Asthma
Bronchitis
Wheezing when you have a cold flu
Skin problems
Fit, convulsions, epilepsy
Persistent joint of back pain
Diabetes
Persistent trouble with teeth, gums or mouth
Cancer
Stomach or other digestive problems
Bladder or kidney problems
Hearing difficulties

Frequent problems with periods or other gynecological problems
Other health problem

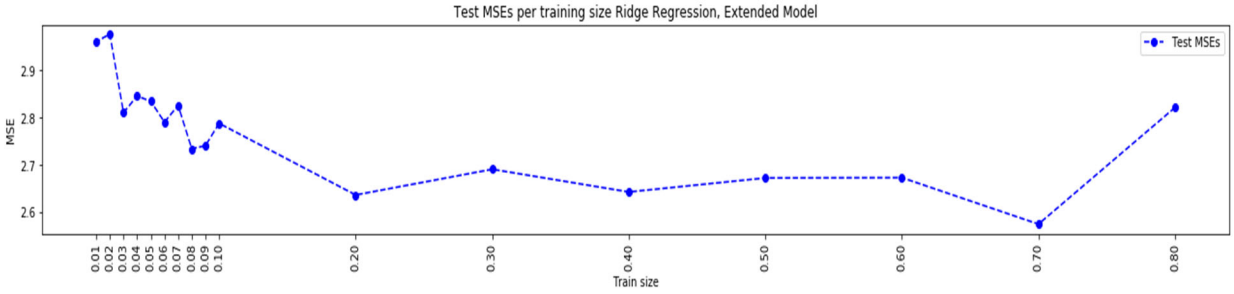
Appendix C: Learning Curves

Learning Curves refer to the behavior of the MSE calculated on the Test Set as function of the size of the training set, based on the idea that more-complex data generating processes (DGP) may require larger training sets. The understanding of the necessary size of the training set required to correctly learn the DGP is useful for a number of reasons. First, should we be interested in carrying out new analyses on the same data, we can save time by fitting the new algorithms only to the required amount of training data. Second, this can help us to better understand the complexity of the underlying DGP. And last, it can provide guidance for the training set size required for the analysis of similar, but not identical, data. In the Extended model, we limit our discussion to the Ridge Regression, and for the Original model we present the Unpenalized Learning Curves. In both the Original and Extended models, all of the five different Penalized Linear Regressions considered have similar learning behavior. We also plot the curves from Random Forests for both models, trained on non-standardized values.

C.1 Learning Curves: Linear Regressions

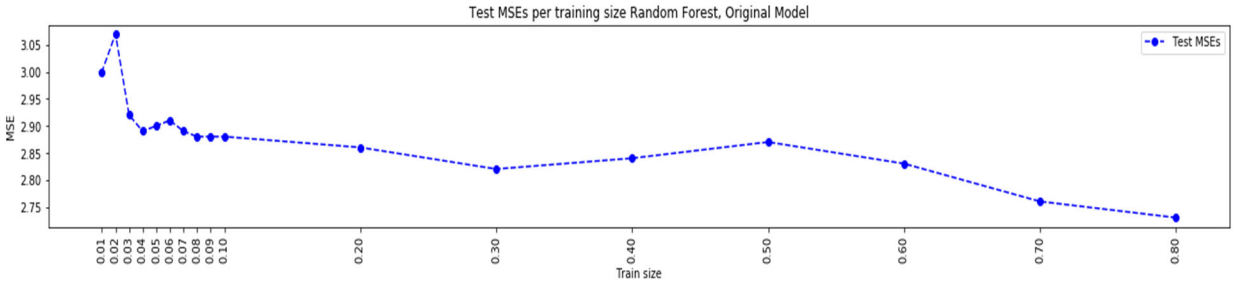


In the Original model, we start with the Unpenalized Linear Regression. Here the DGP is already fully learned with only 2% of individuals in the training set. This is consistent with our finding that an Unpenalized Linear Regression is the best choice for these data, and that the linearity assumption holds: the correct DGP is learned very quickly. Here, the MSEs converge to the bias only.

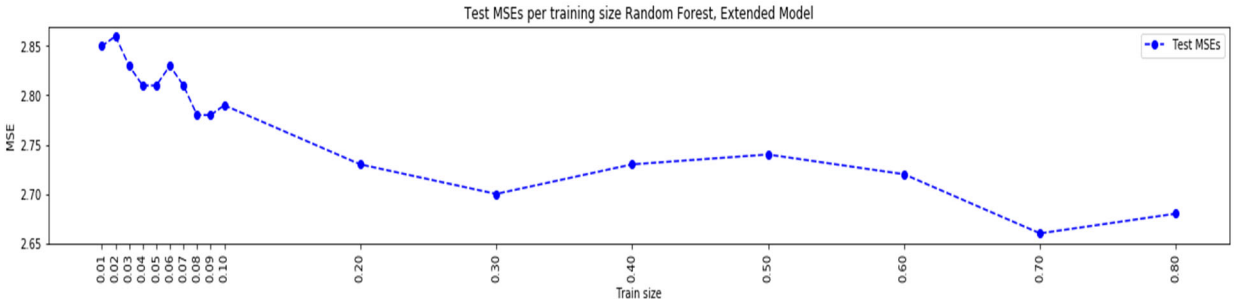


The behavior in the Ridge Regression on the Extended dataset is similar to that in the Unpenalized Linear Regression in the Original dataset. In this case, the Test MSE also stabilizes for training sets including more than 20% of individuals.

C.2 Learning Curves: Random Forest



For the Learning Curves in the Random Forest, in the Original model the DGP is learned confidently with 3% of observations in the training set.



The DGP is also confidently learned with 10% of individuals in the training set in the Extended model, with the Test MSE thereafter remaining constant.

Appendix D: Categories with at most 15 individuals:

Categories with at most 15 individuals	Number of individuals
Type of Accommodation - A room / rooms	15
Main Activity - Don't Know	2
Main Activity - On a government scheme for employment training	7
Main Activity - Wholly Retired	1
Whether registered disabled - Don't Know	8
Highest academic qualification - Don't Know	10
Marital Status – Widowed	12
Marital Status - Other missing	3
Whether mother is alive - Don't Know	9
Smoking habits - Daily but frequency not stated	6
Smoking habits - Other missing	1
Tenure Status – Squatting	2
Tenure Status - Don't Know	2

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

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
1844	Tomaz Teodorovicz Raffaella Sadun Andrew L. Kun Orit Shaer	How does working from home during Covid-19 affect what managers do? Evidence from time-use studies

1843	Giuseppe Berlingieri Frank Pisch	Managing export complexity: the role of service outsourcing
1842	Hites Ahir Nicholas Bloom Davide Furceri	The world uncertainty index
1841	Tomaz Teodorovicz Andrew L. Kun Raffaella Sadun Orit Shaer	Multitasking while driving: a time use study of commuting knowledge workers to access current and future uses
1840	Jonathan Gruber Grace Lordan Stephen Pilling Carol Propper Rob Saunders	The impact of mental health support for the chronically ill on hospital utilisation: evidence from the UK
1839	Jan Bietenbeck Andreas Leibing Jan Marcus Felix Weinhardt	Tuition fees and educational attainment
1838	Jan De Loecker Tim Obermeier John Van Reenen	Firms and inequality
1837	Ralph De Haas Ralf Martin Mirabelle Muûls Helena Schweiger	Managerial and financial barriers during the green transition
1836	Lindsay E. Relihan	Is online retail killing coffee shops? Estimating the winners and losers of online retail using customer transaction microdata
1835	Anna D'Ambrosio Vincenzo Scrutinio	A few Euro more: benefit generosity and the optimal path of unemployment benefits