ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF ECONOMIC SCIENCES

DEPARTMENT OF ECONOMICS

THE ECONOMICS OF HAPPINESS: A MACHINE LEARNING APPROACH

GAVRIIL NIKOLAOS

Dissertation submitted

in partial fulfilment of the necessary prerequisites

for the acquisition of the MSc Degree

Athens

January, 2018

We approve the dissertation of Gavriil Nikolaos

Supervisor

Professor Kyriazidou Ekaterini ……

Examiner 1

Professor Vettas Nikolaos ……

Examiner 2

Professor Palivos Theodoros ……

Abstract

The emotional state of happiness has been for a long time studied mainly by the field of psychology. Economists lately acknowledge the fact that self-reported measurements of happiness should play a bigger role in policy and economic theory. In this study we use economic and development indicators to model the relationship between happiness and a selected subset of these indicators. To produce models of high accuracy and interpretability advanced methods from the fields of statistical learning and data mining are used. GDP per capita, GDP growth and unemployment are generally considered as indicators of high importance and this study confirms that statement. Moreover we provide evidence on the relationship between happiness and variables that indicate life expectancy, immigration and ethical values like gender equality, although more research should be conducted in order to validate the relationship of the these indicators and the aforementioned fundamental variables.

Journal of Economic Literature (JEL) Classification: *C50 D60 I31*

Keywords: *happiness economics, welfare, subjective well-being, statistical learning*

Acknowledgements

First, I would like to express my sincere gratitude to my dissertation advisor Ekaterini Kyriazidou. Her teaching excellence and guidance definitely provided me with the necessary tools to choose the right direction and successfully complete my dissertation. I would also like to thank my friends Panos, George and Giannis for the stimulating discussions, the sleepless nights we were working together before deadlines, and for all the fun we have had during the master's program.

Contents

1 Introduction 5
2 Literature Review 8
2.1 Importance of happiness 8
2.2 Measurement and validity 9
2.3 Important factors of happiness 9
2.3.1 Findings on income 9
2.3.2 Findings on unemployment 10
2.3.3 Findings on relationships 10
2.3.4 Findings on education 11
2.3.5 Findings on health 11
3 Data and methods 12
3.1 Data 12
3.2 Methods 13
3.2.1 The problem 14
3.2.2 The bias-variance tradeoff 15
3.2.3 Ridge regression 15
3.2.4 The lasso 17
3.2.5 Principal components analysis 17
3.2.6 Decision trees 18
3.2.7 Ensembles 20
4 Empirical Analysis 21
4.1 Model selection 21
4.1.1 Preprocessing data 21
4.1.2 Prediction performance 23
4.2 Feature selection 28
4.3 Modeling 32
4.3.1 Model 1: gradient boosting 32
4.3.2 Model 2: least squares 33

4.3.3 Model 3: decision tree 34
5 Discussion / Results 35
5.1 Performance results 35
5.2 Important indicators 36
5.2.1 GDP growth 36
5.2.2 Merchandise exports to high-income economies 37
5.2.3 GDP per capita 40
5.2.4 Unemployment 42
5.2.5 Population ages 65 and above 42
5.2.6 Women in national parliaments 44
6 Conclusions 46
List of tables 51
List of figures 52

1 Introduction

This paper seeks a deeper understanding of the happiness of nations. Happiness is studied by many fields, each one defining it in a different manner. In economics the study of happiness belongs to the field of happiness economics which is a subfield of behavioral economics. Behavioral economics is the subfield of economics that studies the bounds of rationality of economic agents. While happiness is an emotional state, so well-known and experienced by every human being, at the same time it is hard to understand what is actually happening under the hood.

There are different reasons that make happiness such a hard concept to understand. One of them is the fact that happiness is correlated with many other variables and these variables share similar connections between them. For instance, being healthy and rich increases your chances of being happy, but health and income are correlated with each other since rich individuals can invest in their health and healthy individuals are more productive and thus can increase their income. This situation poses several issues in statistical modeling. For example, taking this to an extreme, we could have income and health being perfectly correlated. Then by using both variables in a least squares model and estimating the parameters we would be unable to identify the marginal effect of each variable, since any (constrained) linear combination of the coefficients could minimize the error of the predictive model. Even if these variables are not fully correlated, still there are problems that emerge and have to be solved.

Another problem we encounter in the study of happiness is the plethora of externalities. Such an externality is the case of employment. Many individuals, consider working as a mundane process and feel that less hours invested in it would result in greater levels of happiness. While this could be true for extreme case, there are studies that suggested otherwise. Research conducted in elderly people found that those who retire later tend to live longer. Other studies find that if people are given exactly the same compensation for being unemployed with the salary of others that work, they would be less happy even if they enjoy the same income but have more free time. The previous examples indicate that there are other factors as well that distort happiness and are closer to emotional experiences rather that rational thinking. These externalities give rise to problems in science too, since one has to maintain an interdisciplinary approach in order to understand these results better.

In this study we take a step back and evaluate some of the empirical findings in the happiness economics literature. Subjective measurements of happiness are used as the target variable. A representative sample of citizens from different countries is asked how they would rate their life lately using as their evaluation

metric a ladder with 11 steps (from zero to ten). In order to construct predictive models of high accuracy we select the independent variables that will be used by following an exploratory approach. Starting with a big set of economic and development indicators, we work towards narrowing down the search to those that actually matter. Our goal is to find indicators that relate with happiness but can be broken down to more fundamental concepts. Proving causation is not the main target of this analysis and in some steps of the analysis we keep a more qualitative attitude towards the topic.

In the first chapter we start by presenting some important research findings in the field that relate to our analysis. We include some factors that influence happiness and the way they relate with it. More specifically, happiness is linearly related with the logarithm of income. Countries with lower levels of GDP per capita have considerable differences with respect to their reported happiness levels for different values of GDP per capita, while countries with higher levels of that indicator have differences in reported happiness levels that cannot be explained by income. Besides income, unemployment has been the subject of many studies and results have been found both on the way people are affected by unemployment on a personal and general level. Apart from economic indicators, a short reference is provided regarding the relationship between happiness and non-economic variables like relationships, education and health.

In the second chapter we provide information regarding the methodology of this paper. We present the data sources and the reasons for choosing the particular datasets. Later in the chapter we explain the problems we face when trying to model the relationship between happiness and the rest of the indicators and how these problems are solved using modern statistical and machine learning approaches. In particular, the data suffer from multicollinearity and classical statistical techniques are unable to overcome this obstacle. Biased techniques like ridge and lasso drop variance levels and achieve better predictive performance. In terms of prediction error gradient boosting provides outstanding results and like lasso, a subset of the initial features can be selected with respect to their importance. While ensemble methods are superior to single trees in terms of accuracy, the latter provide high interpretability and are generally a simple and intuitive way to model statistical relationships.

The third chapter contains the empirical analysis. We start with data preprocessing and then we select the most important features and most efficient ways to model their relationship with happiness. While prediction accuracy is certainly a highly important target, compromises are made for methods that provide better interpretation and thus can offer a more spherical view of the underlying relationship. While the statistical package R would be the ideal framework for the analysis, certain libraries from the Python 3.6 scripting language are more helpful when it comes to data cleaning and preprocessing. In addition, 75% of coding time was spent in the preprocessing stage and thus Python was the selected programming language for the empirical analysis.

Finally, we discuss our findings and qualitatively show how these results apply to some characteristic examples. We suggest that GDP growth plays an important role in happiness at least in short term and present some cases where for some countries many indicators that relate with happiness remain constant and GDP growth moves in a similar fashion with happiness. We also present some evidence of conditional importance. Simply put, some indicators seem to be useful only for certain parts of the sample. For instance GDP per capita is highly correlated with happiness in poor countries, while the same correlation in wealthy countries is very close to zero. Other highly important features are the percentage of population aged over 65 and the percentage of seats held by women in national parliaments. The former variable seems to indicate longevity at its lower levels and immigration at its higher. The latter may indicate the existence of moral values in different societies. Quite interesting is the fact that the benefit in terms of happiness from increasing the income of poor countries is more or less the same with the benefit from establishing moral values in rich societies.

2 Literature Review

Utility is one of the major concepts in economic theory. Neoclassical economists have inferred the utility that an agent derives from goods or services, by his revealed preferences. This methodology is consistent under the assumptions that the agents are rational, fully informed and seek to maximize their utility. In this context the agents will choose the action that maximizes expected utility.

The flaws of the assumptions above have given rise to the field known as behavioral economics. Behavioral economists accept subjective well-being (SWB) or happiness as a certain utility notion and consequently the effects of economic situations can be studied under this framework. In this section we provide the most important factors of happiness as they were presented in the papers of Dolan et al. (2007), Frey et al. (2002) and Wiese.

2.1 Importance of happiness

Happiness is a highly complex concept. In philosophy, the term is defined in terms of living a good life, rather than a feeling and it was used to translate the Greek word eudaimonia. According to Sonja Lyubomirsky, happiness is *"the experience of joy, contentment, or positive well-being, combined with a sense that one's life is good, meaningful, and worthwhile."* Even though happiness is one of the few goals in life shared by so many people, economists have left its study to other disciplines, especially psychology. There are other goals in life like economic prosperity but they are considered as secondary since they follow a different end.

One of the reasons why economists are interested in the concept of happiness is economic policy. An important research finding is the consistently strong influence of non-financial variables on well-being. Many governments, acknowledging the fact have changed their priorities by taking into account other measures of well-being instead of GDP per capita. One characteristic example is that of Bhutan. In 1971, the country introduced an alternative indicator known as gross national happiness (GNH). They have adopted policies that give higher priority to the happiness of their citizens and less to economic indicators like GDP. Studying happiness can also provide knowledge on the effects that happiness has on different variables like productivity according to Oswald et al. (2015). They found 12% increases in productivity due to high happiness levels that reached as high as 20% above the control group.

2.2 Measurement and validity

An important issue for economists that study happiness and life satisfaction is the validity of subjective well-being data in formal analysis. Kahneman and Krueger (2006) find that self-reports concerning subjective well-being correlate with smiling frequency or rating of one's happiness by friends and thus can be considered as a solid indicator of true happiness.

Another issue is the distinction between happiness and life satisfaction. These two concepts are not really synonyms and this is reflected also in the data. The World Value Survey poses the following question: *"Taking all things together, would you say you are (i) Very happy, (ii) Rather happy, (iii) Not very happy, (iv) Not at all happy, (v) Don't know".* The Gallup World Poll, on the other hand, uses the Cantril Ladder question and asks respondents to evaluate their life: *"Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?".* These two measurements are closely related but not identical. The most common way to analyze data is by taking national averages. It has been also found that there is understanding among humans about what it means to be happy and that linguistic differences are not important.

2.3 Important factors of happiness

2.3.1 Findings on income

According to the standard economic literature, utility is increasing on income while marginal utility is diminishing. Higher income leads to higher consumption and thus higher utility. Brickman et al. (1978) found empirical evidence that lottery winners where only slightly happier after a year. The average winner reported 4/5 life satisfaction which was only 0.2 points more than the control group which had substantially less wealth. Furthermore, Easterlin (1974) claimed that money doesn't buy happiness. Specifically, he showed that even though there were considerable differences between rich and poor people in the same country, the pattern would disappear in comparison among different countries (developed nations). In addition, he analyzed time-series data from 1946 to 1970 and found that while there was an increase in income in the United States, reported happiness would remain quite constant and even decline between 1960 and 1970, creating the well-known Easterlin paradox. Veenhoven (2003) concluded that there was no paradox, and countries did indeed get happier with increasing income. In 2008 Stevenson and Wolfers concluded that increases in absolute income were linked to increased self-reported happiness, for both individuals and whole countries. In 2010, Easterlin published data from a sample of 37 countries reaffirming the paradox and soon came a response from Wolfers. However, Wolfers and Stevenson admit in a New York Times (2008) interview that their *"time-series evidence is fragile".* Further, they just look at a short time frame and

disregard the adaptation effects which might play a role in a long-term comparison. Last but not least, their theory cannot clarify why some countries' life satisfaction data has remained unchanged even though GDP has risen over a long time. Interestingly, the approach of Easterlin and Stevenson/Wolfers differs in many ways. For example the first uses growth rates while the second GDP levels. Besides, they even work with different datasets and different data series. Hence, their contradictory results can easily be influenced by these differences and shows the limitation of econometric research in general. Gilovich and Kumar (2015) suggested that *"experiential purchases (such as vacations, concerts, and meals out) tend to bring more lasting happiness than material purchases."* Their explanation for the aforementioned result was that *"Compared to possessions, experiences are less prone to hedonic adaptation."* Kesebir and Oishi, argued that inequality hides the effect that increased GDP may have on national happiness and that could explain the paradox to some extent.

2.3.2 Findings on unemployment

Empirical studies have provided results both on the effects of personal and general unemployment. Oswald et al. (1994) found that unemployment reduces well-being more than any other factor, including important negative ones such as divorce, while controlling for other indirect effects like the income loss. Additionally, it seems that being without a job is more devastating for those who are male, highly educated and middle-aged (Clark-Oswald 1994, Clark et al. 2006). As mentioned before (Oswald et al. 2015), there are also findings on the reverse causation, since lower happiness levels reduce productivity and lower productivity may result in unemployment. The main causation though seems to run from unemployment to unhappiness. Besides personal unemployment, people are also unhappy about unemployment even if they themselves are not put out of work according to Di Tella et al. (2001). People may feel bad about their fortune or fear that they may share the same fate. They may also consider the high taxes or increased criminality that often follow unemployment. There is also evidence of different effects of unemployment for different reference groups. For instance, someone that is already unemployed feels less discomfort if unemployment hits his close environment (Clark 2000).

2.3.3 Findings on relationships

Being single seems to be worse for SWB than being part of a relationship. Regular sex was also associated with higher levels of SWB, while the effects were strongest when this was with the same partner (Blanchflower & Oswald 2004). Wildman and Jones (2002) report that men and women appear to suffer equally following widowhood, divorce and separation, while single women may actually have higher well-being than married women. There is also evidence of adaptation after divorce or widowhood. While some people recover faster than others, finding someone new is often associated with a return to the original levels of well-being.

The effects of having children on happiness are not significant, while there exists a positive relationship to life satisfaction as Haller and Hadler (2006) report. The idea behind it is that even though children reduce positive emotions on a daily

basis, people consider them an important part of their overall well-being as they provide meaning to their lives. It appears also that seeing family and friends is positively associated with SWB (Pichler 2006).

2.3.4 Findings on education

There are mixed results concerning education. Some studies find that each additional level of education relates with higher levels of SWB (Blanchflower & Oswald 2004), while others find that middle level education is related to the highest life satisfaction. (Stutzer 2004). The coefficient on education is often responsive to the inclusion of other variables within the model. Education is likely to be positively related with income and health and if these are not controlled for we would expect the education coefficient to be more strongly positive. However, the inclusion of correlated variables raises a multicollinearity problem.

2.3.5 Findings on health

Both psychological and physical health have a strong impact on SWB. Of course, the relationship is strong due to the fact that SWB may as well cause better health and especially psychological health. Shields & Wheatley Price (2005) study the effect that several negative situations like heart attacks and strokes have on health. Oswald and Powdthavee (2006) present evidence that individuals partially adapt to disability status (although never fully recover), finding that as the length of time an individual experiences the disability increases, the negative impact of disability drops.

3 Data and methods

Economists and psychologists are trying to solve the mystery of what brings people joy and deeper meaning and recent advances in technology could definitely provide assistance towards that end. The cost of data collection has decreased, statistical and machine learning techniques have evolved and that gives us the opportunity to develop highly accurate models using the plethora of available data. The purpose of this paper is to take advantage of that by choosing a more exploratory attitude. Instead of forming a series of hypothesis, there has been followed a rather dynamic approach. A huge variety of indicators are collected and advanced machine learning techniques are used for recognizing patterns that may not be identifiable with standard statistical methods. The most important factors of human happiness are investigated, as well as the way they influence happiness. Of course, the field of happiness economics is quite young and hence there are certain limitations regarding the availability and quality of data.

3.1 Data

The underlying source of the happiness scores is the Gallup World Poll, a poll using sample of people from over 130 countries around the globe. The sample consists of people aged 15 and older and was drawn from urban areas. The dataset contains measurements that range from 2005 to 2016. The questionnaire covered several aspects of well-being, including an overall measure of life satisfaction, as well as health and economic indicators. The main poll question was: *"Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"*. The questionnaire remained the same in all countries and that provides an opportunity to make cross-country comparisons. No previous poll has provided national measurements for so many countries, but that is just the first reason why we will use this dataset for our study. The second reason is related to the fact that the answers follow an 11-point scale and thus we hope that the greater sensitivity will present patterns that may not be easily recognized in a 4-point scale used in other studies. This dataset provided only the life satisfaction question (or Life Ladder as we will refer to it from now on).

The features that are used as predictors of Life Ladder in this study were drawn from the World Bank database under the name *"World Development Indicators"*. More specifically, WDI is the primary collection of development indicators from officially recognized international sources. The data range from 1960 to 2016 and

contain over one thousand national indicators and estimates from GDP per capita, unemployment and mortality rates to percentages of population over 65 years and seats held by women in national parliaments in different countries. The huge diversity of indicators provides a valuable opportunity in exploring the concept of happiness and the forces that drive its movement. The dataset comes with some challenges with respect to its size, shape and number of missing values. For instance, some indictors are present at some years while absent in others.

3.2 Methods

Conventional statistical and econometric techniques have various issues when it comes to modeling relationships using big datasets. Besides the fact that most of them seek mostly linear relationships, there are more fundamental problems that are related to the fitting process. And even if we decide to use classical techniques we first have to make some decisions as to which features are more relevant to our needs, using feature selection techniques. So, one reason why machine learning techniques should be used by economists is for pattern recognition in the first stages of their analysis.

Machine learning techniques are mostly about prediction. Given a set of features a machine learning algorithm can be trained to predict the future values of a related variable taking its history under consideration. However, these algorithms are not so useful in predicting the causal impact that a variable may have on another. Varian (2014), presents how machine learning techniques can be used indirectly for causal inference. More specifically, he provides an example where a policy maker should choose the number of police officers that should be assigned to a precinct for criminality reduction purposes. In order to estimate the causal impact that the increased police force has in criminality, one should compare the criminality rates after the change in police force, with the ones that would exist if the change had not taken place. A machine learning model could be used then to predict with great accuracy what the outcome would be without the intervention and then estimate the causal effect.

In our problem the variable of interest is continuous and the dataset is high dimensional. Therefore, techniques like lasso, ridge and principal components regression are considered a good starting point. Additional attention will be given to non-parametric techniques like decision trees and gradient boosting. Bellow we provide a formal presentation of our problem, the theoretical background of the techniques that will be applied and the main reasons that make them more appropriate for the specific problem.

3.2.1 The problem

For the definition and analysis of the problem we use the same notation with Hastie et al. (*An Introduction to Statistical Learning*). Consider a dataset that contains different variables and n samples for each one. Suppose then, that we choose one of them as a target, in the sense that we would like to use the rest of the features to predict its behavior. Let's define as y the target and $X_1, X_2, \dots, X_K$ the rest of the variables that we will use as features. We can make the hypothesis that a linear function $h(\cdot)$ could be used to predict the target y pretty well:

$$h_\beta(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \quad (1)$$

where X is a $n \times (K+1)$ design matrix that contains the features as columns, plus a column of ones for the intercept, and each row represents a sample. We can then, compute a vector of parameters β that when inserted in (1) would make the hypothesized function as close as possible to the target variable y. To measure the efficiency of our model we can use some cost function. Consider the function $J(\cdot)$ that measures the squared differences between each observation of the target variable and the predicted value of the linear model:

$$J(\beta_0, \beta_1, \dots, \beta_K) = \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{K} \beta_j x_{ij})^2 \quad (2)$$

where x_{ij} is the element that corresponds to the ith sample and jth column of the X matrix and y_i the ith element of the target vector. The above cost function is the sum of squared residuals and in matrix notation it can be expressed as:

$$RSS(\beta) = (y - X\beta)'(y - X\beta) \quad (3)$$

By setting the first derivative of (3) equal to zero we find the optimum value of the parameter vector β to be:

$$\beta^{ols} = (X'X)^{-1}X'y \quad (4)$$

The above solution (least squares parameter) is unique and can be computed if the number of samples is larger than the number of features and the X matrix has full column rank. Under some assumptions regarding the distribution of y we can also compute the variance of the least squares solution to be:

$$Var(\beta^{ols}) = (X'X)^{-1}\sigma^2 \quad (5)$$

where σ^2 is the variance of the target variable and assumed to be constant. In addition, if the y's are standardized, $\sigma^2 = 1$. The solution is unbiased and if the variables are standardized, $X'X = R$, where R is the correlation matrix of the independent variables. The variance of each least squares parameter is:

$$Var(\beta_j^{ols}) = \frac{1}{1 - R_j^2} \quad (6)$$

where R_j^2 is the R-squared value obtained from regressing X_j on the other independent variables and it's the variance inflation factor (VIF). VIF quantifies

the severity of multicollinearity in the least squares regression. To reach the least squares solution we assumed that there is no linear relationship between the independent variables. In the case where the independent variables are not related in a pure linear manner but are strongly correlated, the results will be distorted. The value of R_j^2 in equation (6) will be close to 1 and hence the variance of the solution will be inflated, which makes the solution unstable and highly dependent on the dataset used for estimation. A rule of thumb is that when VIF values are bigger than 10, or equivalently when an independent variable has R-squared value bigger than 0.9 with the rest, we face multicollinearity.

3.2.2 The bias-variance tradeoff

A central idea in accomplishing high performance predictions is the tradeoff between bias and variance of the model. Generally, a model suffers from high bias when it is rather simple (underfitting), while it suffers from high variance when it's more complex (overfitting). To be more specific, let's see how these concepts relate to our problem. Having a dataset with many independent variables increases the danger of multicollinearity. The dataset that will be used for the empirical analysis in the next chapter is even more prone to multicollinearity since many of the features may be indicators of the same variable or even worse represent the same variable in different measurement units. As mentioned in section 3.2.1 when multicollinearity is present the least squares estimates have inflated variance. One can show that the expected prediction error can be decomposed into:

$$E\left[(y_0 - \hat{h}(x_0))^2\right] = Var\left(\hat{h}(x_0)\right) + [Bias(\hat{h}(x_0))]^2 + Var(\varepsilon) \quad (7)$$

for a given value of x_0. We cannot decrease the noise variance but we can do something for reducing the prediction error. By introducing some bias we could reduce the variance of the model. Of course, this would reduce the prediction error if the decrease in variance is greater than the increase in bias. To present an extreme example consider having K regressors and instead of using all the parameters of model (1), keep just the intercept and set the rest equal to zero. Then, by minimizing the sum of squared residuals we find the intercept to be equal to the mean of the target variable. We have solved the variance issue but this estimator is highly biased.

3.2.3 Ridge regression

Ridge regression is a technique for analyzing multiple regression datasets that suffer from multicollinearity. This method decreases the variance of the model by introducing some bias. To do this, we regularize the coefficients by choosing how large they grow. As in the initial least squares problem, we seek to minimize some cost function. This function is slightly differentiated by an additional a term that controls the shrinkage of the coefficients and it's called penalized sum of squared residuals.

$$PRSS = RSS + \lambda \sum_{j=1}^{K} \beta_j^2 \quad (8)$$

The extra term in the $PRSS$ is called shrinkage penalty. Observe that when the tuning parameter λ is equal to zero the solution will be identical with that of least squares. As λ increases, instead of just minimizing the sum of squared residuals, some weight is also given to the values of the coefficients. As $\lambda \to \infty$, the impact of the shrinkage penalty increases and the coefficients approach zero. Of course, we need to shrink just the coefficients that are associated with the independent variables and not the intercept, which measures the mean value of the target variable when all X's are equal to zero. Unlike least squares, ridge produces a different solution for each value of the tuning parameter. To choose the optimum value for the tuning parameter we can split the dataset in k parts. Then, we can train the model using $k-1$ parts and test its performance on the remaining part. Repeating the process until all parts have been used as test sets and averaging over the prediction errors, offers a good estimate of the true prediction error. The minimization problem of $PRSS$ can be solved analytically:

$$\beta_{\lambda}^{ridge} = (X'X + \lambda I_K)^{-1} X'y \quad (9)$$

It can be easily shown that the above solution is biased. The geometric interpretation of ridge provides a better understanding of this process. The

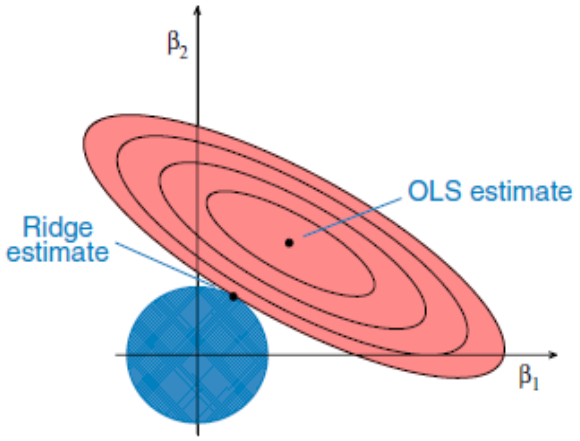

Figure 1 Solution of the error minimization problem in ridge regression. Source: PennState STAT 897D Applied Data Mining and Statistical Learning

ellipses correspond to the contours of sum of squared residuals. As we move towards the center of the ellipse the RSS decreases. The circle centered on zero corresponds to the shrinkage constraint. In the case of two regressors the constraint can be written as $\beta_1^2 + \beta_2^2 < constant$. The solution therefore will be found as the point where the ellipse and circle touch. As already mentioned ridge is biased and its relationship with the least squares solution is:

$$\beta_{\lambda}^{ridge} = \frac{n}{n+\lambda}\beta^{ols} \quad (10)$$

3.2.4 The lasso

The lasso method is quite similar with ridge. They are both built on top of least squares by adding some constraint to the original cost function in order to prevent overfitting. The difference lies in the choice of the constraint. While ridge uses the L2 norm (sum of squared coefficients), the lasso uses the L1 norm (the sum of absolute values of the coefficients). More formally, we can write the cost function of the lasso method:

$$PRSS = RSS + \lambda \sum_{j=1}^{K} |\beta_j| \quad (11)$$

This small change has quite an impact in the resulting coefficients. In the lasso case the coefficients can be shrunk to zero. This method is better than ridge in terms of interpretation since it performs feature selection as well. In geometric terms the new constraint represents a diamond and thus, the larger the number of dimensions the easier it is for a coefficient to be set to zero. Even if the lasso constraint is not differentiable, a variety of techniques can be used to spot the solution that minimizes the objective function, including subgradient methods, least-angle regression and proximal gradient methods. When the dataset includes similar indicators, lasso should be preferred as it keeps just one variable from those of the same correlation levels. In the rest of the cases, usually ridge performs better in terms of prediction error.

3.2.5 Principal components analysis

The techniques mentioned so far belong to the category of supervised learning. In the context of supervised learning there is a target variable and a model that combines the features in order to predict the target. Since we have a target we can estimate how close or far the model stands from predicting the target accurately and thus have a metric of its performance. In contrary, unsupervised learning contains techniques that do not have a simple goal like the prediction of a response. These techniques are used in exploratory data analysis as a means of finding useful patterns in the data or visualizing large datasets for better understanding. Principal components analysis is an unsupervised learning technique used for linear dimension reduction. Big datasets usually contain many correlated variables and before the modeling phase it would be useful to drop the number of dimensions, either for memory or summarization purposes. In addition, if this dataset will be used only for predictive modeling, there is no reason in using correlated variables as they cause instability to the estimated model. So PCA reduces a set of, possibly correlated, high dimensional variables to a lower dimensional set of linearly uncorrelated synthetic variables called principal components. It does so while trying to maintain the largest part of the variability of the dataset.

To illustrate how this can be done consider a set of variables organized in a design matrix X. First step would be to subtract the mean of each variable from each element of that variable. Having done so the sample covariance matrix of the design matrix X is:

$$S = X'X/n \quad (12)$$

Our target is to find the eigenvectors of $X'X$. It can be done either by doing an Eigen decomposition of $X'X$ or by doing singular value decomposition from X. The eigenvectors are called principal components directions of X and if we project X onto the principal components directions we get the principal components. We can perform Eigen decomposition of $X'X$:

$$X'X = (UDV')'(UDV')$$

$$= VD^2V' \quad (13)$$

where V contains the eigenvectors and D contains the square roots of the eigenvalues of $X'X$ as its diagonal elements. We can then take the eigenvector (v_1) that is associated with the largest eigenvalue (d_1^2) and we have the first principal component direction. We can project X on v_1 and get the first principal component $z_1 = Xv_1$, which explains the largest part of the dataset's variance. Following this method, we can compute the rest of the principal components as well.

Having performed PCA, we can choose some of the first principal components and perform least squares regression to predict some target variable. This method is called principal components regression and will be used in the empirical analysis of the next chapter.

3.2.6 Decision trees

A decision tree is a machine learning technique, quite different from those already analyzed. This method performs non-linear modeling of the target variable and its strength lies in its interpretability and ease of explanation and visualization. It can be used for both regression and classification.

Consider the following example. Assume we have some data on height and weight for some people and we would like to predict the gender of the person given his

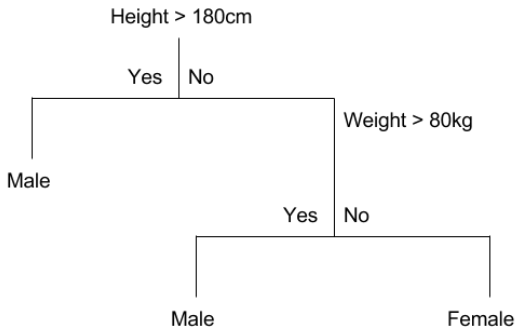

Figure 2 Decision tree representation. Source machinelearningmastery.com

characteristics. We could predict that if the person has height over 180 cm he is a probably a man, but for height less than 180 cm, we could use the information that the weight variable provides. So, if the person is under 180 cm with weight over

80 kg, then he is probably a man, while with weight under 80 kg probably a woman. This is quite an intuitive model. So how can we build one?

Roughly speaking, there are two steps in building a decision tree. First, we have to split the predictor space in j distinct and non-overlapping regions, $R_1, R_2, \dots, R_j$. Then we have to make a prediction for every region. The prediction can be the mean of the response variable in the region if the response is a continuous variable (regression) and the most frequent value of the response variable if not (classification). When we seek to estimate the response for a new set of input variables, we find the region that the sample belongs to and we use the prediction associated with that region. In our example we split the $weight \times height$ space in 3 regions. These can be written as:

$$R_1 = \{X | height > 180\}$$

$$R_2 = \{X | height < 180, weight > 80\}$$

$$R_3 = \{X | height < 180, weight < 80\} \quad (14)$$

The 3 regions are known as leaves of the tree. So, a new observation, according to the region it belongs, will be classified as male or female.

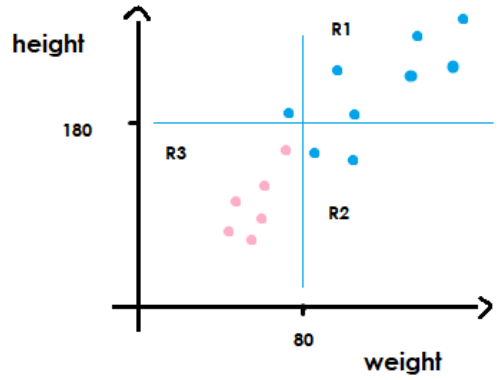

Figure 3 Plot of the tree regions

Specifically, there are some criteria that are used to choose the optimal size of the tree, the value of the split and if a variable will be used in the higher or lower levels of the tree. The process begins by choosing one variable X_j and splitting the X space in two regions according to that variable. The split should be done so that the sum of squared loss is minimized. We can then compute the prediction loss and repeat the process with all the other variables. The variable that provides the smaller prediction error is used as the first criterion in the top of the tree. In more detail, we seek the value s and define the two regions $R_1 = \{X | X_j \leq s\}$ and $R_2 = \{X | X_j > s\}$ so that equation (15) is minimized:

$$\sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2 \quad (15)$$

where $\hat{y}_{R_1}$ is the mean response in the first region and $\hat{y}_{R_2}$ in the second. We repeat the process for the rest of the variables and choose the variable X_j that minimizes the prediction error. We have two regions from the first split and we can repeat the process for each region, to grow the tree even more. As it becomes bigger it has the tendency to overfit. To avoid it, we can test its prediction accuracy using cross-validation or validation set and prune it so that it provides more general predictions.

3.2.7 Ensembles

Decision trees usually suffer from high variance, since they are sensitive in the subset of the data they are trained. However, if we grow several trees and average their predictions (in the continuous response case) we can have a dramatic decrease in prediction error. There are several ensemble methods like bagging, random forests and boosting. Bagging, bootstraps the dataset, trains a single tree in each bootstrap and averages the predictions. Random forests, provide an improvement over bagging, by decorrelating the trees. Each tree is trained on a bootstrapped training sample. The difference is that when a split is considered, only a subset of predictors is taken into account. Boosting, unlike bagging grows the trees sequentially. Each tree uses information from the previous tree. The process starts by estimating an initial model and each time a new decision tree is used to fit the residuals of that model. Then we add the tree to the fitted function to update the residuals. Generally, boosting outperforms random forests, which outperform bagging.

4 Empirical Analysis

In the first section of this paper, some findings from research in the field of happiness economics were presented. In the next section, we explained the issues that emerge from the existence of multicollinearity and the methods used to solve those problems. This section, contains our empirical analysis based on the data described in previous section. This chapter starts with a detailed demonstration of the data preprocessing. Afterwards, we establish some benchmarks based on the predictive performance and interpretability of the methods described earlier on a subset of the data and then we present in detail the results of the analysis. We also append the Python (3.6 version) code in the grey boxes.

4.1 Model selection

4.1.1 Preprocessing data

We start by dealing with some issues that emerge from the fact that the data we use come from different sources. There are countries that exist in the SWB dataset while missing from the development indicators (DI) dataset. Another issue is the fact that the DI dataset contains measurements in a bigger timeframe than the measurements contained in the SWB dataset. Finally the DI dataset has to be reshaped to follow the standards of a design matrix (each column represents a feature and each row represents an observation).

Importing the Python modules that are needed for the task:

```
import pandas as pd
import numpy as np
import pickle

happy = pd.read_csv('C:\\Users\\happinessdata.csv')
bigdata = pd.read_csv("C:\\Users\\indicators.csv")
```

For the first part of the analysis we will use the measurements that were drawn in 2014. Reshaping the two datasets:

```
#happy dataset
happy2014 = happy[happy.year==2014].copy()
happy2014.rename(columns={'country':'Country'},inplace = True)
happy2014.drop(["year"],axis=1,inplace=True)
happy2014 = happy2014[['Country','Life Ladder']]


#bigdata dataset
bigdata.rename(columns={'CountryName': 'Country'},
inplace=True)
bigdata.drop(["CountryCode","IndicatorCode"],axis=1,inplace=Tr
ue)
devdata2014 = bigdata[bigdata.Year==2014].copy()
devdata2014 =
devdata2014.pivot(index='Country',columns='IndicatorName',valu
es='Value')
devdata2014.insert(loc=0,column='Country',value=devdata2014.in
dex[:])
```

The DI data frame is full of non-available (NA) values. One way to get rid of that issue is by deleting the samples that contain NA values and then proceed to the analysis with fewer samples. In datasets with a small number of features this is rarely an issue. Since the specific data frame has 921 columns and given that one missing feature in some row is enough to cause the removal of the entire sample, this strategy is not efficient. Another strategy would be to delete the features that contain a sufficiently large number of NAs and the drop the observations that contain NAs to clean the data frame. This process gives us the freedom to choose the number of missing elements that one feature is allowed to contain in order to remain in the data frame. So we can iterate over different values of that parameter and choose its value so that we keep the maximum amount of information:

```
for i in range(15):
    df1 = pd.merge(happy2014,devdata2014,on='Country')
    blacklist =[]
    for col in range(df1.shape[1]):
        if df1[df1.columns[col]].isnull().sum()>i:
            blacklist.append(df1.columns[col])
    df1.drop(blacklist,axis=1,inplace=True)
    df1 = df1.dropna()
```

We keep a threshold of maximum 6 NAs per feature and then clean the remaining observations. The resulting dataframe that will be used for model selection has the following form (just the first 5 observations and 6 features along with the dependent variable):

Country	Life Ladder	Adolescent fertility rate (births per 1,000 women ages 15-19)	Age dependency ratio (% of working-age population)	Age dependency ratio, old (% of working-age population)	Age dependency ratio, young (% of working-age population)	Average precipitation in depth (mm per year)
Afghanistan	3.130896	76.7336	89.773777	4.620393	85.153384	327.0
Algeria	6.354898	10.7914	51.536631	8.794346	42.742285	89.0
Armenia	4.453083	23.5084	41.329330	14.952075	26.377256	562.0
Australia	7.288550	14.4050	50.231115	22.120440	28.110675	534.0
Austria	6.950000	7.3790	48.978845	27.705794	21.273051	1110.0

Table 1 Part of the dataset that will be analyzed

4.1.2 Prediction performance

For this part we will use the following Python modules:

```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import scale
from sklearn.model_selection import train_test_split, KFold,
cross_val_score
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge, RidgeCV, Lasso,
LassoCV
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble.partial_dependence import
plot_partial_dependence
from sklearn.ensemble.partial_dependence import
partial_dependence
from sklearn.metrics import r2_score,mean_squared_error
```

The dataset contains 103 rows and 138 columns. As we mentioned in the previous chapter, the problem we deal with is that of multicollinearity. The least squares method is useless here. If we perform linear regression using the entire set for training and then repeat the procedure splitting the dataset in train and test set, we can observe the effect of multicollinearity:

```
y = df.iloc[:,0]
X = df.drop(["Life Ladder"],axis=1)
Xtrain,Xtest,ytrain,ytest = train_test_split(X,y,random_state=2)
linreg = LinearRegression(normalize=True)
print("r2 before split :
"+str(r2_score(y,linreg.fit(X,y).predict(X)))+
     "\nr2 after split :
"+str(r2_score(ytest,linreg.fit(Xtrain,ytrain).predict(Xtest))))
```

The R^2 in the first case is exactly 1. This is clearly a sign of overfitting. When we split the dataset in train and test sets the R^2 drops to -9.01, which confirms our previous statement. To solve the multicollinearity problem we can use coefficient penalization. Ridge and Lasso perform that kind of process. In the two plots we can spot the effect that regularization with L1 and L2 norms have on the coefficients. As the value of lambda from equation (8) increases, the regression coefficients shrink. Observe that in the ridge case (left) the coefficients converge to zero smoothly, whereas in the lasso case the coefficients actually hit zero.

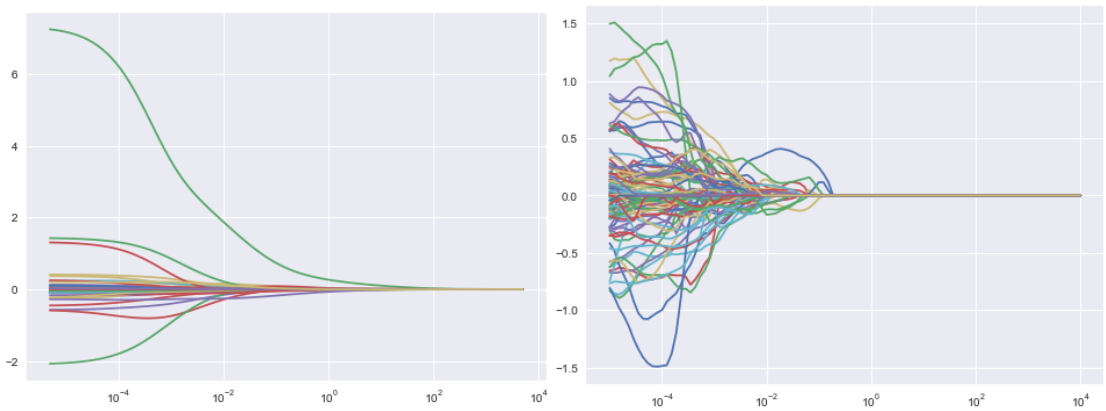

Figure 4 Size of the regression coefficients as a function of lambda for the ridge (left) and lasso (right) cases

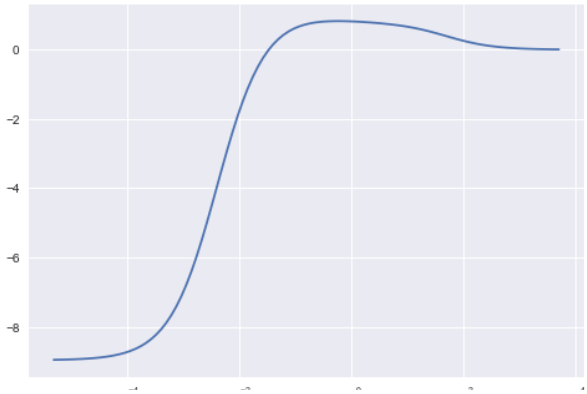

Figure 5 The mean squared error as a function of lambda for the ridge regression

We can also see the effect that penalization has on the prediction accuracy. When lambda is zero the ridge solution is the same with that of least squares. As lambda increases the accuracy improves until it reaches the optimal point where it drops again. The R^2 is negative before the shrinkage due to the fact that the prediction accuracy using all the coefficients is worse than using just the intercept parameter. As the coefficients converge to zero the only parameter left is the intercept and that leads to zero R^2. The task here, is to find the value of lamda that maximizes the prediction accuracy. To compute it we perform (10-fold) cross-validation. In the ridge case we find the optimum value of lambda to be around 2.16. The same steps are repeated for the lasso case where the optimal value of lambda is around 0.01. Since lasso performs feature selection as well, it's more sensitive to lambda. It corrects multicollinearity with shrinkage but some coefficients have reached zero so the effect of multicollinearity is eliminated faster. To compute the optimum lambda and estimate the prediction accuracy we can use the following code:

```
#find optimal alpha for Ridge
ridgecv = RidgeCV(alphas = alphas, scoring =
'mean_squared_error', normalize = True)
ridgecv.fit(Xtrain, ytrain)
#find optimal alpha for Lasso
lassocv = LassoCV(alphas = None, cv = 10, max_iter = 100000,
normalize = True)
lassocv.fit(Xtrain, ytrain)
print("optimal alpha for Ridge : "+str(ridgecv.alpha_))
print("optimal alpha for Lasso : "+str(lassocv.alpha_))

#find prediction accuracy
ridge = Ridge(alpha = ridgecv.alpha_, normalize = True)
ridge.fit(Xtrain, ytrain)
lasso.set_params(alpha=lassocv.alpha_)
lasso.fit(Xtrain, ytrain)
print("Ridge: "+ str(r2_score(ytest, ridge.predict(Xtest))))
print("Lasso: "+ str(r2_score(ytest, lasso.predict(Xtest))))
```

Ridge performs better with $R^2 = 0.77$ while in the lasso case we have $R^2 = 0.73$. For the two-model-comparison the R^2 coefficient should not be the only criterion. Lasso kept just 16 features for the initial set of 138. Thus interpretability should also be taken under consideration in model selection process. The features that survived along with their estimated coefficients are:

Average precipitation in depth (mm per year)	0.000023
Employment to population ratio, ages 15-24, male (%) (modeled ILO estimate)	0.002055
GDP per capita (current US$)	0.000004

GDP per capita, PPP (constant 2011 international $)	0.000003
Improved sanitation facilities, urban (% of urban population with access)	0.017542
Merchandise exports to high-income economies (% of total merchandise exports)	0.002088
Merchandise imports from developing economies in Middle East & North Africa (% of total merchandise imports)	-0.015379
Mortality rate, under-5 (per 1,000)	-0.003324
Price level ratio of PPP conversion factor (GDP) to market exchange rate	0.407129
Private credit bureau coverage (% of adults)	0.001299
Proportion of seats held by women in national parliaments (%)	0.006336
Rural population (% of total population)	-0.005004
Theoretical duration of secondary education (years)	-0.034140
Time required to register property (days)	-0.000469
Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	-0.010064
Urban population (% of total)	0.001141

Table 2 Selected features from lasso with their coefficients

The above results confirm the findings of other empirical researchers. Income, unemployment, education, health and exports seem to be important for SWB. Another way to reduce the multicollinearity is by using dimension reduction techniques. We described PCA in the previous chapter. To use this method for prediction purposes we have to reduce the dimensions of the original dataset and then use the remaining features to perform least squares regression. PCA has some issues when it comes to modeling. The first issue is that the only criterion for the dimension reduction is the reduction of the data size and not the performance of a predictive model. Another issue is the interpretability of the resulting model. The features that will be used for the regression model are linear combination of the initial features and thus we cannot interpret the marginal effects of the initial features. Nevertheless, this technique perform well under conditions. We start by computing the first principal components (using the SVD method). Then we perform cross validation to choose the optimal number of principal components:

```
pca = PCA()
Xreducedtrain = pca.fit_transform(scale(Xtrain))
n = len(Xreducedtrain)
CV = KFold( n_splits=10, shuffle=True, random_state=1)
mse = []
linreg = LinearRegression()
score = -1*cross_val_score(linreg, np.ones((n,1)),
ytrain.ravel(), cv=CV,
scoring='neg_mean_squared_error').mean()
mse.append(score)
for i in np.arange(1, 21):
    score = -1*cross_val_score(linreg, Xreducedtrain[:,:i],
ytrain.ravel(), cv=CV,
scoring='neg_mean_squared_error').mean()
    mse.append(score)
```

If we plot the relationship of mean squared error and the number of principal components we see that using the first 5 components offers a good predictive model. Using more than 15 predictors reduces the accuracy dramatically. Even if a model that used many predictors could achieve slightly better results than the one using five predictors it would be still reasonable to keep the latter since it performs well using only a small subset of principal components. The fact that the performance of the model that uses the first five principal components is poorer than that of ridge and lasso ($R^2 = 0.7$ in the PCA model), plus the low interpretability make this technique unsuitable for our purposes.

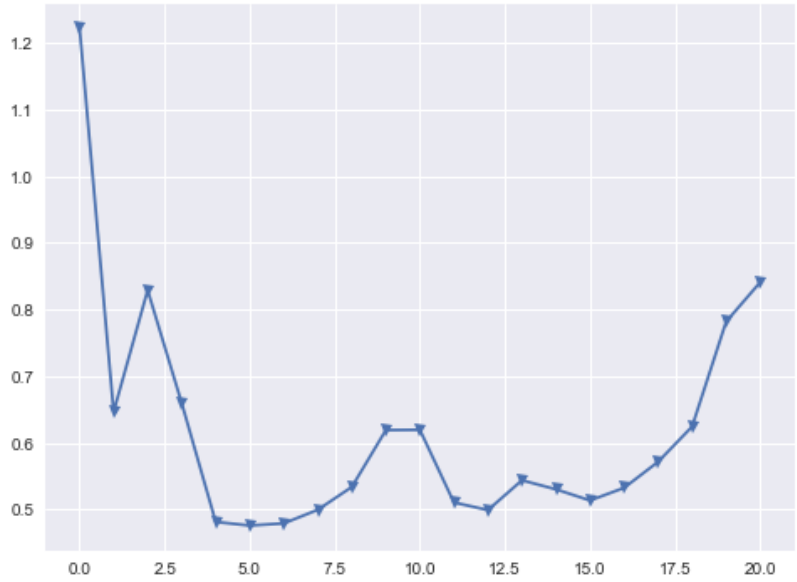

Figure 6 Mean squared error as a function of the number of principal components used in the model

Linear methods should be preferred for their simplicity and interpretability. In our case, using a linear model has devastating results in prediction accuracy. Using non-parametric statistical methods can be a solution. As we mentioned in the previous chapter ensemble methods combine many decision trees to decrease the variance and thus increase the predictive accuracy. While a single tree achieves $R^2 = 0.73$ using gradient boosting results in $R^2 = 0.85$.

```
params = {'n_estimators': 700, 'max_depth': 2,
'min_samples_split': 2,
          'learning_rate': 0.01, 'loss': 'ls','max_features'
:7,'random_state':0}
reg = GradientBoostingRegressor(**params)
reg.fit(Xtrain, ytrain)
print("Gradient Boosting : "+str(r2_score(ytest,
reg.predict(Xtest))))
```

We can avoid overfitting by using many trees with small depth. After some point the performance of the algorithm doesn't change by increasing the number of trees as we see in the following plot.

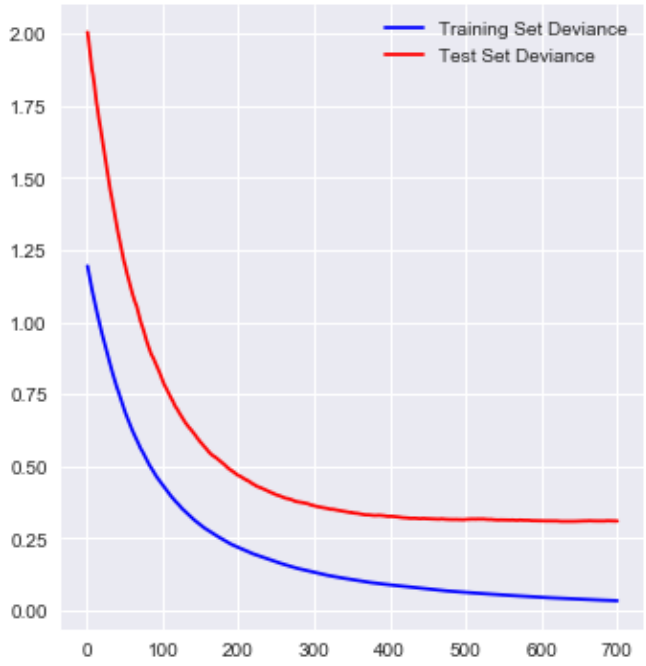

Figure 7 Training and test set deviance as a function of iterations

4.2 Feature selection

Even if these statistical techniques provide predictive models of high accuracy, to get better insights in the relationship between SWB and the indicators we have to narrow down the list of predictors. We start by removing the features that measure the same thing with minor modifications. For example we can keep *'GDP per capita, PPP (current international $)'* and drop *'GDP per capita (current US$)'*

or other similar indicators. Then we filter the remaining features using two feature selection techniques, randomized lasso and recursive feature elimination with cross-validation (RFECV). The final set of features are filtered one last time to remove features that indicate the same idea. For instance, we keep *'Documents to import (number)'* and drop *'Time required to register property (days)'*, since both features indicate the efficiency of the public sector.

Randomized lasso (or stability selection), performs the lasso method in different subsets of the original dataset. Then, the algorithm evaluates the importance of the features by ranking them with respect to their frequency of appearance in multiple iterations. This process works like the ensemble methods. By using different partitions of the dataset the variance of the estimated importance of the features drops. The second filter is RFECV. This feature selection algorithm starts by modeling the relationship of all available features (that survive the first round) and drop one feature at a time. The criterion behind the feature removal is the decrease in predictive power. If some feature can be removed without sacrifices in accuracy then that feature should be removed. The additional Python modules needed are:

```
from sklearn.linear_model import RandomizedLasso
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import RFECV
```

In the previous section a subset of the initial dataset (year 2014) was used in order to evaluate the performance of some statistical techniques. Since there are more data in our disposal we can use them as well. While many features are available for the 2014 case, we are limited to a smaller subset in case we use data from 2004 to 2014, since each data frame (that contains yearly data) should contain exactly the same features. So we start by preprocessing the initial dataset of indicators:

```
datalist = []
for year in range(2005,2016):
    data = bigdata[bigdata.Year==year].copy()
    data =
data.pivot(index='Country',columns='IndicatorName',values='Valu
e')
    data.insert(loc=0,column='Year',value=year)
    data=data.loc[:,finalkeep]

data.drop(list(data.isnull().sum()[data.isnull().sum()==247].in
dex.values),axis=1,inplace=True)
    data = data.dropna()
    datalist.append(data)
```

```
happylist = []
for yr in range(2005,2016):
    data = happy[happy.year==yr].copy()
    data.rename(columns={'country':'Country'},inplace = True)
    data = data[['Country','Life Ladder']]
    data = data.set_index("Country")
    happylist.append(data)
```

After collecting the SWB data as well, we have created two lists of data frames, one for the SWB data and one for the indicators. After that, only the common features are kept for each year in the first data list (indicators):

```
allcolumns = []
for frame in datalist:
    mycolumns = list(frame.columns.values)
    allcolumns.append(mycolumns)

result = set(allcolumns[0])
for s in allcolumns[1:]:
    result.intersection_update(s)
finalcols = list(result)

finaldata = []
for frame in datalist:
    fr = frame.loc[:,finalcols]
    finaldata.append(fr)
```

Finally, we can create a new list that contains as elements the data frames that are created by merging the SWB with the indicators for the same year and country:

```
combolist = []

for yr in range(0,10):

    d = pd.merge(happylist[yr],finaldata[yr],left_index=True,
right_index=True)

    combolist.append(d)

dfremake = pd.concat(combolist)
```

Since the dataset is ready for analysis we perform feature selection using randomized lasso:

```
X=dfremake.drop("Life Ladder",axis=1)

X = StandardScaler().fit_transform(X)

y= dfremake["Life Ladder"]

model =
RandomizedLasso(alpha=0.002,random_state=1,sample_fraction=.5)

model.fit(X,y)

filterdcols = model.get_support()

filterdcols =
pd.Series(True).append(pd.Series(filterdcols),ignore_index=Tru
e)

frame = dfremake[dfremake.columns[filterdcols]]
```

We are left with 16 features and using the RFECV method we can get rid of one more:

```
y = frame.iloc[:,0]

X = frame.drop(["Life Ladder"],axis=1)

reg = LinearRegression()

rfecv = RFECV(estimator=reg)

rfecv.fit(X, y)

newdf = frame.drop(X.columns[rfecv.ranking_>1],axis=1)
```

After performing manual feature selection as well we are left with the following indicators:

Improved sanitation facilities (% of population with access)
Population ages 65 and above (% of total)
Proportion of seats held by women in national parliaments (%)
Year
Time required to register property (days)
Unemployment, total (% of total labor force)
GDP growth (annual %)
Documents to import (number)

Merchandise exports to high-income economies (% of total merchandise exports)
Rural population (% of total population)
GDP per capita, PPP (current international $)

Table 3 Set of selected features through the 3-layer feature selection process

This three-layer filtering process has provided the most important features for modeling purposes. To get a good understanding of SWB one model is not enough. We provide a gradient boosting regression model and partial dependence plots to depict the nonlinearities in the relationship between SWB and the selected features. Then a linear model is presented, to compute the marginal effects of each feature and finally a decision tree, which will provide a rule-based approach to the analysis.

4.3 Modeling

4.3.1 Model 1: gradient boosting

Since we have selected the most important features we can use the data from 2004 to 2014 to train a gradient boosting regression model. The model is highly accurate with $R^2 = 0.90$. The importance of each feature in prediction is depicted in the next plot:

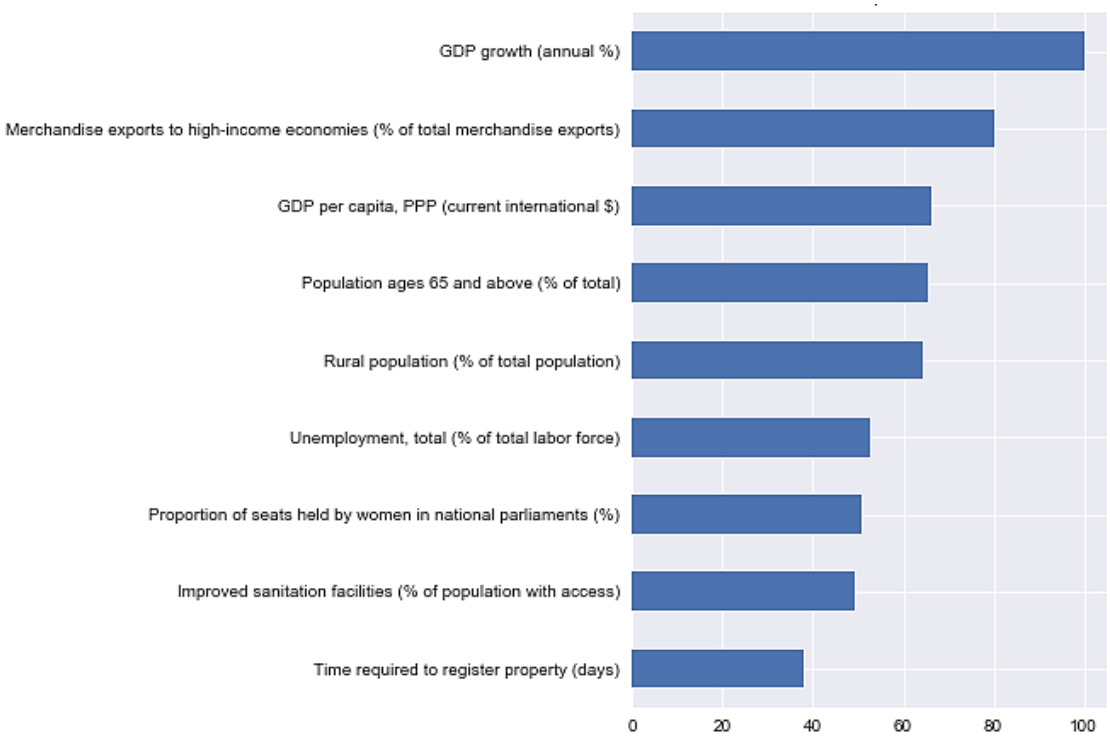

Figure 8 Feature importance plot for the gradient boosting model

We can gain considerable insight by using partial dependence plots. They were introduced by Friedman (2001), to address the lack of interpretability for his gradient boosting machine, but can also be applied for any complex predictive

model. The idea behind them is the following. Assume that we have a dataset with N observations and p features. The model we use to predict the kth element of our target variable, given some features, is:

$$\widehat{y_k} = F(x_{k,1}, x_{k,2}, \dots, x_{k,p}) \quad (16)$$

For some feature x_j we can compute the following quantity and plot it for different values x of that feature:

$$\varphi_j(x) = \frac{1}{N}\sum_{k=1}^{N} F(x_{k,1}, \dots, x, \dots, x_{k,p}) \quad (17)$$

In the next plot we see how unemployment influences the model predictions.

Figure 9 Partial dependence plot

Low levels of unemployment do not influence the predicted SWB, while after 12%, unemployment has a negative effect on SWB. In the next chapter, we discuss the effect that these indicators have on SWB.

4.3.2 Model 2: least squares

A linear model is inferior compared with gradient boosting when it comes to accuracy, since it achieves $R^2 = 0.75$, but still better than the simple lasso case, even with fewer features. Nevertheless, as a parametric method, it provides regression coefficients which we can use to estimate the mean marginal effect of each selected indicator. The estimated linear model is:

$$\widehat{SWB} = 5.462 + 0.364X_1 - 0.266X_2 + 0.231X_3 - 0.255X_4 + 0.175X_5 - 0.164X_6 + 0.118X_7 \quad (18)$$

and the features used to estimate it are presented in the following table:

X_1	Improved sanitation facilities
X_2	Rural population
X_3	GDP per capita
X_4	Unemployment
X_5	Proportion of women in national parliaments
X_6	Documents to import
X_7	Merchandise exports to high-income economies

Table 4 Features of the least squares model

4.3.3 Model 3: decision tree

The final model in this section provides a more intuitive approach to the understanding of happiness. As we mentioned in the previous chapter, ensemble methods achieve better accuracy than simple decision trees. The decision tree presented bellow achieves $R^2 = 0.77$, slightly better than the linear model, but offers better representation of the relationship:

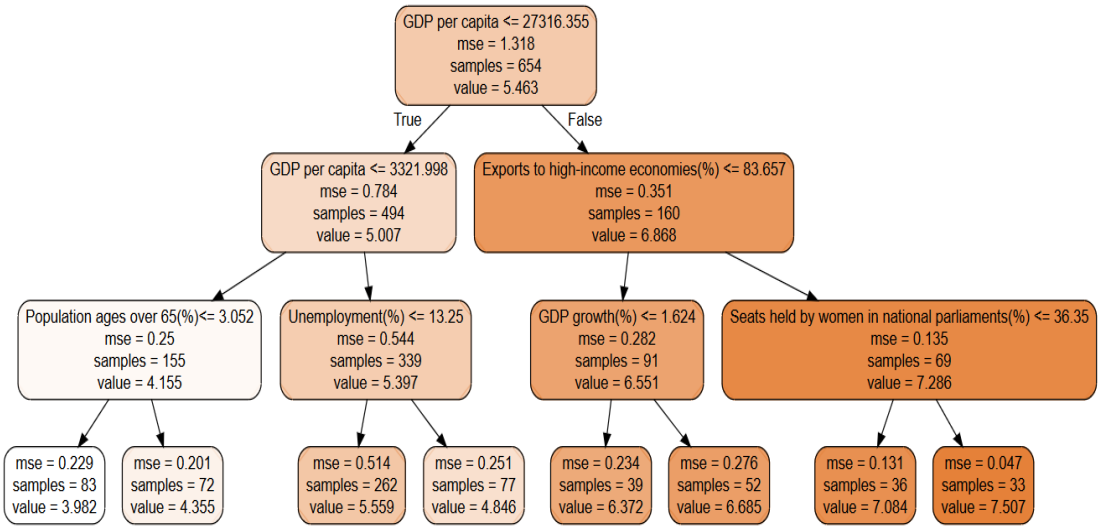

Figure 10 Decision tree

5 Discussion / Results

In this chapter, we discuss the results of our empirical study and how they relate with previous studies. Although the goal of this paper, like every other paper in the field, is to gain some insights regarding the mechanisms that promote happiness, our approach is differentiated. Starting with an initial set of over 1000 economic and development indicators, we narrow our search down to the most relevant features as they were chosen from different feature selection techniques. Then, using various statistical techniques, we model the relationship between subjective well-being and the selected features.

5.1 Performance results

The major problem of our analysis was the existence of multicollinearity and to tackle that, we used the appropriate tools. These tools were lasso, ridge, principal components, decision trees and gradient boosting regression. Even if our ultimate goal is the interpretation of the results, the results cannot be trusted if they come from inaccurate methods. In the table below we present the techniques used along with their goodness-of-fit (R^2) estimate, as they were computed in the previous chapter.

Technique	$\boldsymbol{R^2}$
Gradient boosting	0.85
Ridge	0.77
Lasso	0.73
Decision tree	0.73
Principal components	0.70

Table 5 Methods ranked according to their prediction accuracy

PCA neither performs well nor has interpretable results and thus was not used in the second phase of the analysis. Ridge performed better than lasso, but the latter offered feature selection as well. Since interpretation of a model that contains over 50 coefficients is not an option, lasso is considered as the best choice for the particular problem. A single decision tree had pretty much the same accuracy with that of linear models. It should be noted though that since we wanted a general and interpretable model, the tree was kept short with the maximum depth being three layers. Usually there is no point on using trees with more than three or four layers of nodes, since ensemble methods perform considerably better. In terms of accuracy, gradient boosting had the best performance. Given these results, the methods that had the best overall performance and were used in the second phase were gradient boosting, lasso and decision trees.

In the second phase of the analysis, we used a three-layer feature selection process and supplied the chosen features to the selected methods of the first phase. The number of features was considerably lower and the available samples also increased since we used all the available data. Under these circumstances, the accuracy improved, especially for the gradient boosting method. On the other hand the accuracy of the decision tree didn't have any considerable increase, but this was expected since it was kept short on purpose.

5.2 Important indicators

5.2.1 GDP growth

The results from least squares and gradient boosting are contradictory for the specific indicator. In the least squares case its coefficient was close to zero and had the smallest absolute value from the set of selected indicators. In gradient boosting, even if it was ranked as the most important feature, in the partial dependence plot there is no visible relationship between GDP growth and SWB.

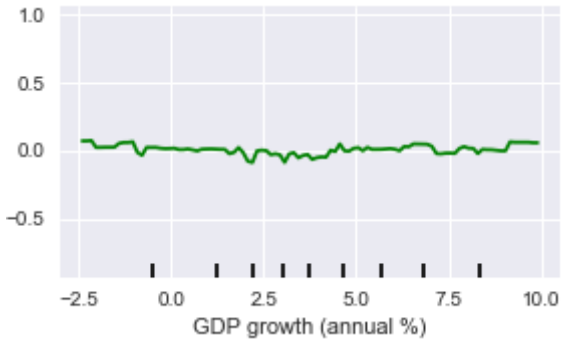

Figure 11 Partial dependence plot

This plot suggests that if a country has a 5% growth while another country has 0% growth and they reach the same GDP per capita level (and keeping all other things constant), then these countries will have the same SWB levels. The problem with this proposition lies in the fact that each sample contains just the growth with respect to the previous year and thus each sample does not contain evidence of consistent growth. It seems reasonable that signs of consistent growth would increase positive feelings regarding the situation in the future.

There are cases where GDP growth seems to correlate with SWB. Take Bahrain for example. In the next plot we have time series for SWB, merchandise exports to high-income economies, percentage of population over 65 years, unemployment, GDP per capita and GDP growth. Observe that SWB has intense fluctuations that seem to correlate with GDP growth. All other variables apart from GDP per capita have small changes. Even if the growth rate is positive for the 10-year timeframe, the SWB levels seem to be influenced by GDP growth. Observe that the GDP growth increase from 2011 to 2012 is the same, but the year before 2011 the growth rate

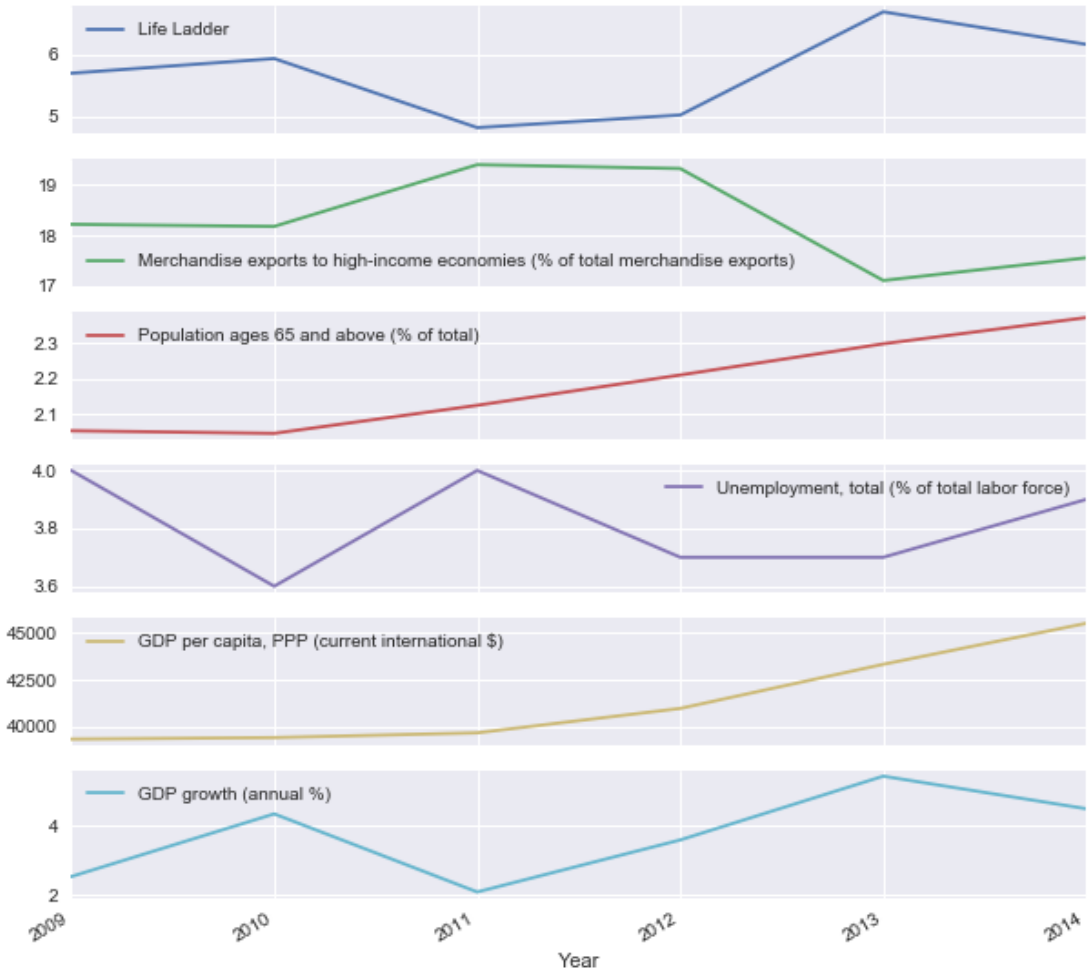

Figure 12 Time series for Bahrain

was negative and that could explain the difference in the SWB growth in periods 11-12 and 12-13. An interpretation of the plot is that even though one year growth may not cause happiness, consistent growth may do it. Another observation that seems to empower the relationship of growth and happiness is the fact that GDP per capita was increasing the whole period and that didn't prevent the drop in happiness level in the period 13-14.

We also can't ignore the strong psychological findings regarding adaptation. There are some matters that should be investigated further. For instance, what is the connection between growth rate and adaptation? A consistent growth rate may promote higher happiness levels but what happens when do people get used to it?

5.2.2 Merchandise exports to high-income economies

This is the second most important feature according to the gradient boosting model. Like GDP growth the direct relationship is quite poor, both in the partial dependence plot and in the least squares model (small value coefficient). Like GDP growth this indicator seems to promote happiness. A reasonable question at this point is how come these two aforementioned indicators seem to be so important without having any direct relationship. To answer this question we have to revisit the mechanism behind the decision making in the feature importance part of

gradient boosting. Some feature can be characterized as important with respect to the improvement it offers in the prediction error. The indicator *merchandise*

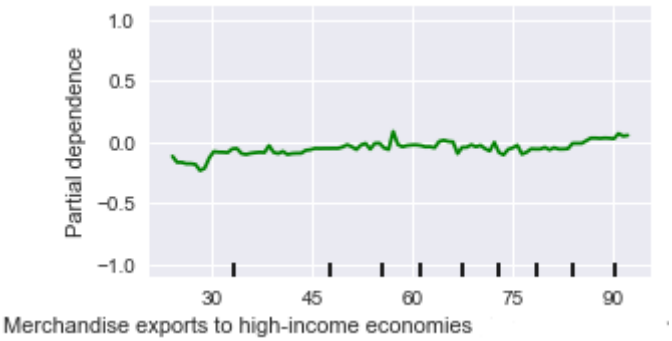

Figure 13 Partial dependence plot

exports to high-income economies offers improvement in accuracy but in a conditional way. It seems to reduce the prediction error when it is used in lower levels of the decision trees that are aggregated in order to develop the gradient boosting model. Consider the decision tree we presented in the previous chapter.

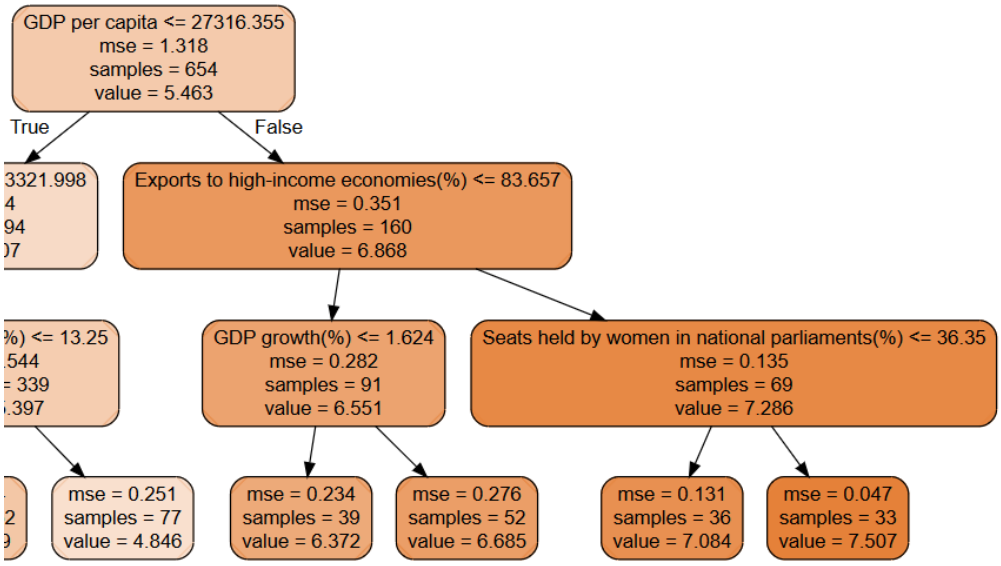

Figure 14 Part of decision tree

In the first split, coutries are separated in two nodes with respect to their GDP per capita levels (since this is the feature that provides the biggest decrease in prediction error for the first split). For the set of countries that the proposition $GDP\ per\ capita \leq 27316$ is true, the best choice for the next split is merchandized exports to high-income economies.

To show that let's split the dataset according to the first criterion, pick the subset that contains the countries that correspond to the false proposition of the first split and then resplit according to the second criterion. In the following matrix we

present the mean values of some important features in the two nodes that resulted from the split that used exports to high-income economies as a criterion. Observe that the class containing countries of higher SWB status (by 0.7 of a unit) attains pretty much the same values in all indicators apart from exports to high-income economies, where the difference is dramatic.

Indicator	High class mean	Low class mean
Life Ladder	7.236390	6.546076
Merchandise exports to high-income economies (% of total merchandise exports)	89.021567	65.366382
Unemployment, total (% of total labor force)	6.745098	7.069298
GDP per capita, PPP (current international $)	44319.490695	45770.796142
Improved sanitation facilities (% of population with access)	98.321569	99.216667
Rural population (% of total population)	19.261137	19.466140

Table 6 Mean values of some important features for the two classes of countries of the second split

The question that emerges from the specific pattern is the following. Since merchandise exports to high-income economies is the only difference from the most important indicators in the specific subset of data, does this feature cause happiness?

To answer this we need to specify what the aforementioned feature represents. It doesn't make sense, countries where people have similar income levels to have different happiness levels due to exports and especially to high-income economies. In the next matrix we present the countries that belong in these two classes:

High class countries	Low class countries
Austria	Australia
Belgium	Bahrain
Canada	Cyprus
Czech Republic	Greece
Denmark	Israel
Finland	Italy
Iceland	Japan
Ireland	Kuwait
Luxembourg	Malta
Netherlands	Oman
Norway	United Arab Emirates
Sweden	Qatar

Switzerland	Saudi Arabia
	Trinidad and Tobago
	Portugal
	Singapore

Table 7 The countries that belong to the two classes as they were separated by the second split

So what is the difference between the exports of these two sets of countries? The countries that have high percentage of exports to high-income economies have as their most exported good: machinery equipment, motors and vehicles, computers, chemicals and electrical equipment. On the other hand, countries that rank lower in the specific feature have as the most exported good: coal, petroleum, food and beverages, oil and agricultural products. So the actual difference between them is that high class countries export products that need highly skilled workers while lower class countries export mainly products that do not share the same constraint. It seems that education plays an important role here but further research should be conducted in order to support this result.

5.2.3 GDP per capita

Perhaps the most important indicator of happiness since it's ranked third in the gradient boosting and in the linear model. In the three-layer decision tree we presented, it is used twice as a splitting criterion. What makes it even more important is the fact that many development indicators are highly correlated with it since income is used as a means to advanced education, health and consumption. In the partial dependence plot of gradient boosting, unlike the previous indicators,

Figure 15 Partial dependence plot

there exists a significant relationship with SWB. After $40K the relationship seems to disappear. The fact that income is important in countries that have less of it can be seen in many ways. If we split the original dataset in the countries with GDP per capita over $30K and under $10K and measure the correlation of GDP per capita with SWB in the two cases we find that in the former it reaches 0.61 while in the latter -0.05. In the multiple linear regression model the value of the GDP per capita coefficient is 0.231. Since the features were standardized before the fitting process to interpret the coefficient we have to rescale it dividing by the standard

deviation of GDP per capita. We find that an additional amount of 10000 dollars increases happiness by 0.121 units on average.

According to Easterlin, richer people inside a country are happier than poorer. He argues that the same relationship does not exist between countries and also in time series between GDP per capita and happiness. Our findings indicate that there is certainly a relationship in cross section data between GDP per capita and SWB but that doesn't hold in time series. We analyzed the contribution of GDP growth in the Bahrain case and how SWB correlates more with growth than with GDP per capita. Bahrain is not the only example. Strong relationship can be found

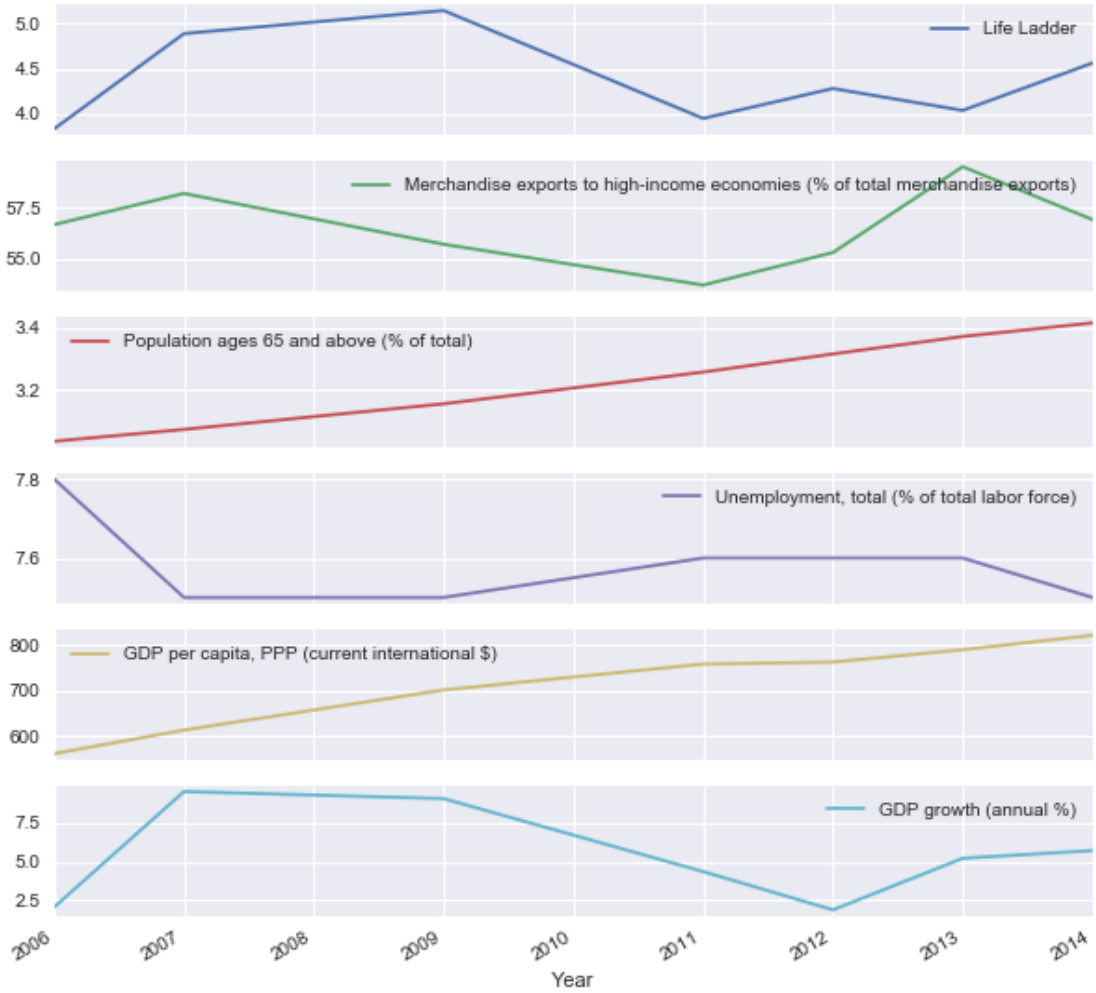

Figure 16 Time series for Malawi

in the Malawi case too. In the relevant plot we see that SWB is highly correlated with growth and even though growth is positive in that timeframe and thus GDP per capita is strictly increasing, SWB doesn't follow the same pattern. Observe that in this case, as in Bahrain, all other variables remain pretty constant. The Malawi case is stronger that Bahrain, since Bahrain has high levels of GDP per capita and thus we expected no relationship between GDP per capita and SWB. In the Malawi case, with GDP per capita under $1K, we expected a stronger relationship between GDP per capita and SWB. We should note that these time series cover only one

decade and thus more research should be done for solid results between the relationship of income and SWB.

5.2.4 Unemployment

In the partial dependence plot of the gradient boosting model, we see that unemployment influences SWB when it reaches some threshold. In particular, until unemployment reaches 13% there is neither negative nor positive effect, while after that threshold SWB drops. Simply put, unemployment doesn't influence happiness in lower levels but has a negative effect when it reaches some threshold. In the first chapter we presented the double effect of unemployment on happiness. Unemployment drops happiness for the unemployed but on the same

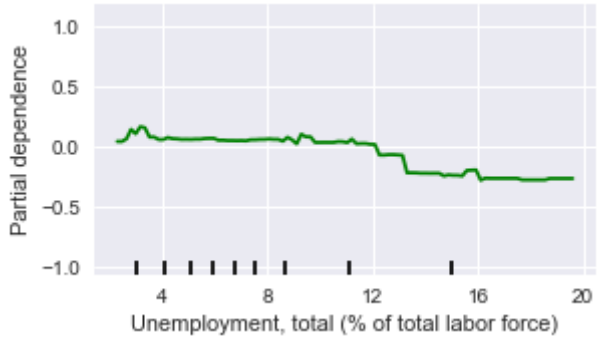

Figure 17 Partial dependence plot

time causes some concern for those who currently work but fear that this might change. Of course this effect is observed in higher unemployment levels and thus our results confirm the established findings.

5.2.5 Population ages 65 and above

In the partial dependence plot below we can see that SWB is concave with respect to that variable. This feature is important but rather complex. The reason is that it

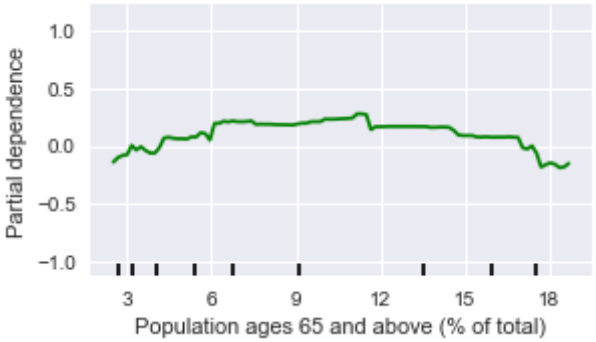

Figure 18 Partial dependence plot

can be used as an indicator of more than one variables. For instance, in the decision tree of the previous chapter it was the splitting criterion for the poorest

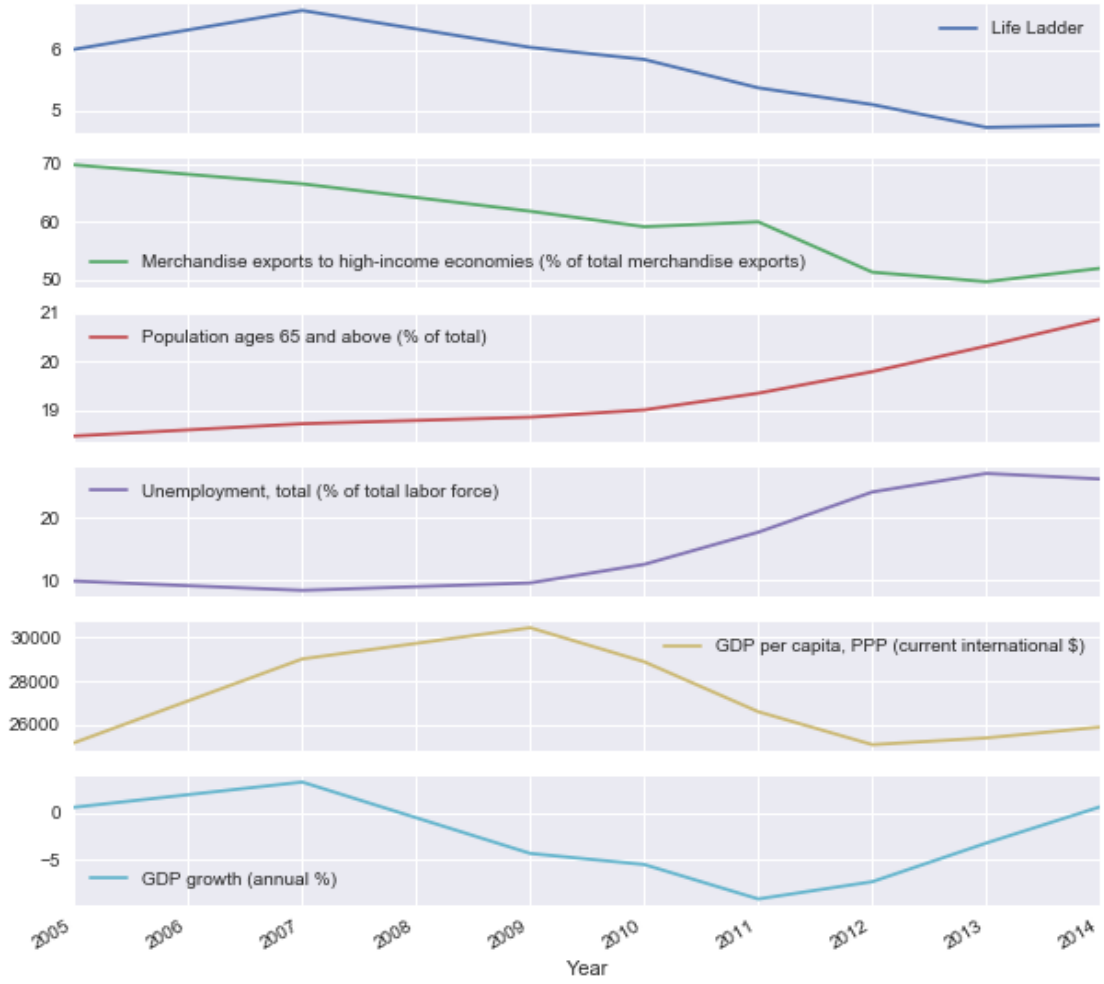

Figure 19 Time series for Greece

countries. In that sense, it can be used to indicate the longevity of people living in a country and thus countries with lower percentages will have lower SWB levels due to bad health and generally bad living conditions. So countries with higher levels of that feature will usually have better living conditions. Longevity is not the only thing indicated by this feature. There is another way for this variable to be inflated. In the short term, increased population percentages of senior citizens indicate that young citizens leave the country. Taking Greece as an example. As observed, after the 2009 crisis, GDP per capita took a dive and unemployment went from 10% to 20% in just a couple of years. As unemployment started increasing so did the percentage of senior citizens. In four years Greece had an increase in the percentage of senior citizens of around 2%. 2% was also the percentage of population that left the country during these four years. This explains the relationship of this feature with SWB. It indicates longevity in the lower levels and thus with its increase SWB is also increased, while in higher levels it indicates that young people are leaving the country and that drops the happiness levels for many reasons, like the splitting of families.

5.2.6 Women in national parliaments

In the partial dependence plot, this features seems to boost SWB levels when it reaches some threshold (around 32%). In the single decision tree it is used as a

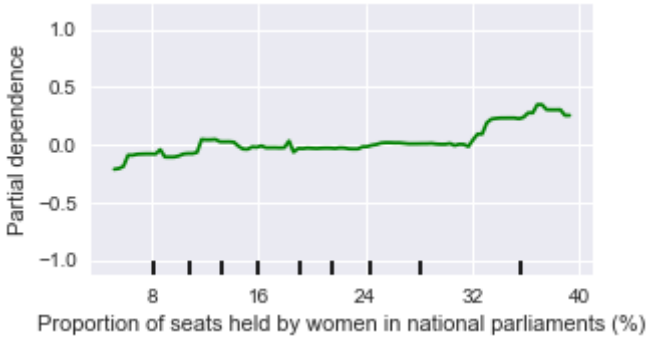

Figure 20 Partial dependence plot

splitting criterion for the top countries in the SWB rankings. In particular, it splits the countries that have already high GDP per capita and high percentage of exports in high-income economies in two new groups. The group that has higher levels of that indicator are happier by ½ of a point in the 11-point-scale. In the next matrix we present the mean values of some important indicators for the countries that where split according to that feature. Observe that these countries have similar statistics in all indicators, except for proportions of seats held by women in national parliaments.

Indicator	High class mean	Low class mean
Life Ladder	7.461254	7.072507
Merchandise exports to high-income economies (% of total merchandise exports)	88.726158	89.236865
Unemployment, total (% of total labor force)	6.462791	6.950847
GDP per capita, PPP (current international $)	43064.576351	45234.089284
Proportion of seats held by women in national parliaments (%)	40.655814	24.608475

Table 8 Mean values of some important features for the two classes of countries of the final split

Scandinavian countries (Finland, Denmark and Sweden) and Netherlands belong in the high class according to that split and also have top ranks in SWB measurements. This feature indicates the equality between sexes and the values that citizens of some country have in general. As a matter of fact, these countries have also low Gini coefficients.

Using the previous results and especially the single decision tree of the previous chapter we can present the general idea behind happiness. Countries that face serious problems like hunger and bad health can increase their happiness levels by dealing with them. Education and moral ethics may help in the long run but are not going to increase their happiness in the present. Countries that do not deal with such issues, may face unemployment and low income and thus they should focus towards solving these issues. In the top of the SWB ranking, we find countries that do not face the previous problems. Increasing their income will not influence their happiness and thus they should focus on their education and quality of values inside their society.

6 Conclusions

The purpose of this study was to get a better understanding of the mechanics behind the emotional state of happiness. We carefully selected from a set of more than 1000 development indicators, those that suit our purpose and used them to model their relationship with subjective well-being. Most of them were not directly connected with the factors that are already found to promote happiness e.g. life expectancy, but can be used to indicate these factors. One example is the population percentage with ages over 65 years. We suggest that when in low levels, that variable indicates longevity, while in higher levels it might be connected with immigration. We used different statistical methods that varied from fully parametric to non-parametric to model the aforementioned relationship. Furthermore, we applied these techniques to rank the initial set of indicators and filter the most important. We also find that while GDP per capita is related with SWB in cross section data, this didn't seem to be the case in time series and thus the latter finding agrees with the Easterlin paradox. GDP growth on the other side was found highly important according to the gradient boosting method. In addition, some visualization patterns emerge as well in time series, but this matter should be investigated further. Finally, we presented the findings and showed how they apply to different countries. One characteristic example is that of Bahrain, where the SWB would fluctuate in a similar way with that of GDP growth, while the other important features were almost constant during the ten year period.

A major issue in this study emerges from the data quality and quantity. Many of the indicators that may be useful in our analysis had many missing values and thus could not be used for analysis. Even if we could somehow recover the partially missing indicators, the SWB observations were limited to a smaller timeframe. In particular the indicators ranged from 1960 to 2015 while the happiness data from 2005 to 2015. This limitation prevents us from having a clear view of the correlation between some variables through time. On the other hand this was a tradeoff we intentionally chose in order to increase the number of countries and the precision of measurement, since the particular dataset contained self-reported measurements of happiness according to the Cantril Ladder system (11-scale) for more than 120 countries.

Further research should be conducted regarding the nature of the indicators that were important predictors of SWB. Specifically, it seems that these features can be used to indicate more fundamental variables, and investigating this relationship would help in breaking them down in their fundamental components and then search for their connection with happiness. The field of happiness economics is a

relatively new and promising one that can potentially change the way we make policy and understand the human motives in general.

References

Easterlin, Richard A. (1974) "Does Economic Growth Improve the Human Lot? Some Empirical Evidence". In: Paul A. David and Melvin W. Reder (eds), *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz*. New York: Academic Press, pp. 89–125.

Easterlin, R. A.; McVey, L. A.; Switek, M.; Sawangfa, O.; Zweig, J. S. (2010). *"The happiness-income paradox revisited".* Proceedings of the National Academy of Sciences. 107 (52): 22463–22468. doi:10.1073/pnas.1015962107. PMC 3012515 . PMID 21149705.

Justin Wolfers (13 December 2010). *"Debunking the Easterlin Paradox, Again".* Freakonomics.com.

Diener, Ed, Ed Sandvik, Larry Seidlitz and Marissa Diener (1993) "The Relationship Between Income and Subjective Well-Being: Relative or Absolute?" *Social Indicators Research*, 28(3), pp. 195–223.

Frey, Bruno S. and Alois Stutzer (2002) *The Economics of Happiness*. World Economics, vol. 3, No. 1

Thomas Gilovich; Amit Kumar (2015). "*We'll Always Have Paris: The Hedonic Payoff from Experiential and Material Investments*". Chapter Four – Advances in Experimental Social Psychology (PDF). 51. Elsevier Inc. pp. 147–187. doi:10.1016/bs.aesp.2014.10.002. ISSN 0065-2601.

Kahneman, Daniel, Ed Diener and Norbert Schwarz (eds, 1999) *"Well-Being: The Foundations of Hedonic Psychology".* New York: Russell Sage Foundation.

Kahneman, D., & Krueger, A. B. (2006). "*Developments in the measurement of subjective well-being*". The journal of economic perspectives, 20(1), 3-24.

Stevenson, Betsey; Wolfers, Justin. *"Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox".* Brookings Papers on Economic Activity. 2008 (Spring): 1–87. JSTOR 27561613.

Veenhoven, Ruut (1993*) "Happiness in Nations: Subjective Appreciation of Life in 56 Nations 1946–1992".* Rotterdam: Erasmus University Press.

Veenhoven, Ruut (2000) "*Freedom and Happiness: A Comparative Study in Fortyfour Nations in the Early 1990s*". In: Ed Diener and Eunkook M. Suh (eds) *Culture and Subjective Well-Being*, pp. 257–288. Cambridge, MA: MIT Press.

Hagerty, M. R.; Veenhoven, R. (2003). *"Wealth and Happiness Revisited – Growing National Income Does Go with Greater Happiness".* Social Indicators Research. 64: 1–27. doi:10.1023/A:1024790530822.

Wildman, J. & Jones, A. (2002). *"Is it absolute income or relative deprivation that leads to poor psychological well-being? A test based on individual-level longitudinal data*". University of York: YSHE.

Alesina, Alberto, Rafael Di Tella and Robert MacCulloch (2001) "*Inequality and Happiness: Are Europeans and Americans Different?" NBER Working Paper No. 8198*. Cambridge, MA: National Bureau of Economic Research.

Becker, Gary S. (1976) *The Economic Approach to Human Behavior*. Chicago: Chicago University Press.

Blanchflower, David G. and Andrew J. Oswald (2000) "*Well-Being Over Time in Britain and the USA"*. *NBER Working Paper No. 7487*. Cambridge, MA: National Bureau of Economic Research.

Blanchflower, D. G., & Oswald, A. J. (2004a). "*Money, sex and happiness: An empirical study*". Scandavian Journal of Economics, 106(3), 393–415.

Blanchflower, D. G., & Oswald, A. J. (2004b). "*Well-being over time in Britain and the USA*". Journal of Public Economics, 88, 1359–1386.

Bowles, Samuel (1998) "*Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions*". *Journal of Economic Literature*, 36(1), pp. 75–111.

Brickman, Philip, Dan Coates and Ronnie Janoff-Bulman (1978) "*Lottery Winners and Accident Victims: Is Happiness Relative?" Journal of Personality and Social Psychology*, 36(8), pp. 917–927.

Clark, Andrew E. (2000) "*Unemployment as a Social Norm: Psychological Evidence from Panel Data"*. Mimeo. University of Orléans, France.

Clark, Andrew E. and Andrew J. Oswald (1994) "*Unhappiness and Unemployment*". *Economic Journal*, 104(424), pp. 648–659.

Clark, Andrew E. and Andrew J. Oswald (1996) *"Satisfaction and Comparison Income"*. *Journal of Public Economics*, 61(3), pp. 359–381.

Clark, A - Diener, E. - Georgellis, Y. – Lucas, R. 2006. *Lags and Leads in Life Satisfaction: A Test of the Baseline Hypothesis*. Working paper, CNRS and DELTAFederation Jourdan.

Di Tella, Rafael, Robert J. MacCulloch and Andrew J. Oswald (2001) *"Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness"*. *American Economic Review*, 91(1), pp. 335–341.

Di Tella, R., MacCulloch, R., Oswald, A. (2003) "*The Macroeconomics of Happiness"*. Review of Economics and Statistics 85, no. 4: 809–827.

Diener, E., Diener, M., Diener, C. (1995) "*Factors predicting the subjective well-being of nations"*. Journal of personality and social psychology, 69(5), 851.

Dolan, P., Peasgood, T., White, M. (2007) "*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 (2008) 94–122

Haller, M., & Hadler, M. (2006). "*How social relations and structures can produce happiness and unhappiness: An international comparative analysis*". Social Indicators Research, 75, 169–216.

Pichler, F. (2006). "*Subjective quality of life of young Europeans. Feeling happy but who knows why?*" Social Indicators Research, 75, 419–444.

Oswald, A., Proto, E. and Sgroi, D. (2015), *"Happiness and Productivity",* Journal of Labor Economics 33(4).

Oswald, A. & Powdthavee, N. (2006)."*Does happiness adapt? A longitudinal study of disability with implications for economists and judges*". Institute for the Study of Labour, IZA DP No. 2208.

Stutzer, A. (2004). "*The role of income aspirations in individual happiness*". Journal of Economic Behaviour and Organisation, 54, 89–109.

Shields, M., & Wheatley Price, S. (2005). "*Exploring the economic and social determinants of psychological wellbeing and perceived social support in England*". Journal Royal Statistical Society(Part 3), 513–537.

Wiese, T., "*A literature review of Happiness and Economics and guide to needed research".* Retrieved from: http://competitio.unideb.hu/wp-content/uploads/2016/03/XIII-1/8_Wiese%20Thomas.pdf.

Frederick, S. – Loewenstein R. 1999. "*Hedonic Adaptation"*. In Well-Being: The Foundations of Hedonic Psychology, ed. D. Kahneman, E. Diener, and N. Schwarz. Russell Sage Foundation.

[dataset] World Happiness Report 2017, Country averages of self-reported life satisfaction (Question: Cantril Ladder), source: Gallup World Poll http://worldhappiness.report.

[dataset] The World Bank, Data Bank, World Development Indicators, http://databank.worldbank.org/data/home.aspx.

"*When Economic Growth Doesn't Make Countries Happier*". Harvard Business Review. Retrieved 2016-12-09.

Esteban Ortiz-Ospina and Max Roser(2017), "*Happiness and Life Satisfaction",* Our World in Data, https://ourworldindata.org/happiness-and-life-satisfaction.

Ron Pearson (2017), "*Interpreting Predictive Models Using Partial Dependence Plots*", https://cran.r-project.org/web/packages/datarobot/vignettes/PartialDependence.html.

James,G., Witten, D., Hastie, T., Tibshirani, R. (2017), "*An Introduction to Statistical Learning: with Applications in R".*

Varian, H. (2013), *"Big Data: New Tricks for Econometrics"*.

List of tables

Table 1 Part of the dataset that will be analyzed 23
Table 2 Selected features from lasso with their coefficients 26
Table 3 Set of selected features through the 3-layer feature selection process 32
Table 4 Features of the least squares model 34
Table 5 Methods ranked according to their prediction accuracy 35
Table 6 Mean values of some important features for the two classes of countries of the second split 39
Table 7 The countries that belong to the two classes as they were separated by the second split 40
Table 8 Mean values of some important features for the two classes of countries of the final split 44

List of figures

Figure 1 Solution of the error minimization problem in ridge regression. Source: PennState STAT 897D Applied Data Mining and Statistical Learning 16
Figure 2 Decision tree representation. Source machinelearningmastery.com..... 18
Figure 3 Plot of the tree regions..... 19
Figure 4 Size of the regression coefficients as a function of lambda for the ridge (left) and lasso (right) cases 24
Figure 5 The mean squared error as a function of lambda for the ridge regression 24
Figure 6 Mean squared error as a function of the number of principal components used in the model 27
Figure 7 Training and test set deviance as a function of iterations 28
Figure 8 Feature importance plot for the gradient boosting model..... 32
Figure 9 Partial dependence plot 33
Figure 10 Decision tree..... 34
Figure 11 Partial dependence plot 36
Figure 12 Time series for Bahrain..... 37
Figure 13 Partial dependence plot 38
Figure 14 Part of decision tree 38
Figure 15 Partial dependence plot 40
Figure 16 Time series for Malawi 41
Figure 17 Partial dependence plot 42
Figure 18 Partial dependence plot 42
Figure 19 Time series for Greece..... 43
Figure 20 Partial dependence plot 44