Bachelor's Thesis Computing Science

Radboud University Nijmegen

Subjective Well-Being and Temperature

Insights from Advanced Analytical Techniques

Author:
Marco Sousa-Poza
s1059057

First supervisor/assessor:
Dr. Ir. Tom Claassen

Second assessor:
Dr. Yuliya Shapovalova

August 26, 2023

Abstract

In the context of rising temperatures due to climate change, this study embarked on a bottom-up investigation of the association between temperature and subjective well-being. Utilizing the Socio-Economic Panel (SOEP), an extensive German longitudinal survey dataset, and open-source weather data, various machine learning models were trained without initial assumptions. The analysis applied techniques such as permutation importance and partial dependence to identify variables correlating with subjective well-being.

The findings reveal a weak yet robust connection between temperature and well-being in both linear and non-linear models. Predominantly, health and economic factors were identified as strong associates with well-being, aligning with current knowledge in the field.

Despite its insights, the study acknowledges limitations, including the lack of focus on specific demographic subgroups that might suffer more from hot weather, and the absence of causal analysis. These areas are suggested for future research.

This research, among the first in Europe to analyze this subject, indicates that there is not yet a significant negative correlation between warm weather and well-being, highlighting the importance of continued investigation in the light of global climate change.

Contents

1 Introduction 3

2 Related Work 5
2.1 Subjective Well-Being 5
2.2 Temperature and Well-Being 6

3 Methodology 8
3.1 Data Collection & Preprocessing 8
3.1.1 Source of Data 8
3.1.2 Data Cleaning 9
3.1.3 Target Variable 11
3.1.4 Feature Engineering 11
3.2 Analytical Methods: Overview 13
3.2.1 Introduction to Methods 14
3.2.2 Linear Regression 14
3.2.3 Random Forest 16
3.2.4 Histogram-based Gradient Boosting 17
3.2.5 Multilayer Perceptron (MLP) 19
3.2.6 Model Refinement and Hyperparameter Tuning 19
3.2.7 Feature Importance 20
3.2.8 Partial Dependence 21
3.2.9 Resampling 23

4 Experimental Results 24
4.1 Software 24
4.2 Feature Importance 24
4.3 Feature Importance - Findings 26
4.4 Effect of Temperature 27
4.5 Simplification 29

5 Conclusions 34
5.1 Research Findings 34
5.2 Future Prospects 34

A Appendix **39**

A.1 Methods Definitions 39

A.1.1 Decision Trees 39

A.1.2 Multilayer Perceptron (MLP) 39

A.2 Tables 41

A.2.1 Variables 41

Chapter 1

Introduction

The study of well-being presents an extensive and complex field, with numerous variables contributing to our understanding of what constitutes a satisfactory life. This thesis investigates a relatively unexplored variable in this domain - the influence of high temperatures on our well-being, a topic gaining increased attention in the face of mounting climate change.

High temperature can cast a wide net of influence, potentially affecting mental and physical health[19], sleep quality[18], and thereby, our overall well-being. As we stand on the cusp of an era marked by climate change, with rising global temperatures and frequent heatwaves becoming an increasingly common reality, this investigation takes on an added urgency. Surprisingly, to our knowledge, the link between weather conditions and subjective well-being (SWB) has not yet been thoroughly investigated in the context of Europe. This gap in the research landscape underlines the need for and relevance of this investigation.

Before embarking on the main exploration, it is necessary to clarify what is meant by 'well-being' in the context of this thesis. There are many different philosophical opinions on what constitutes well-being, each emphasizing different aspects. This thesis adopts the concept of subjective well-being (SWB), a more holistic variable. SWB encompasses an individual's overall assessment of life satisfaction and emotional well-being. It is often measured through self-reporting methods, such as asking individuals to rate their satisfaction with life on a scale from 1 to 10. This approach allows for a comprehensive understanding of well-being that takes into account both the positive and negative aspects of life. Given its multi-dimensional nature and substantial exploration in previous research, SWB provides a stable foundation for comparative analysis.

This investigation is unique in several ways. We harness the power of a vast and unique dataset, the German Socio-Economic Panel, which contains data spanning back to 1984.[9] This extensive temporal range allows us to look for patterns and impacts that might only reveal themselves over time.

Furthermore, unlike conventional well-being research that leans towards hypothesis testing using 'top-down' models, our analysis adopts a 'bottom-up' approach. Here, we refrain from making pre-emptive assumptions about the relationships between variables, choosing instead to use the richness of the available data to unearth potential effects.

In this context, our 'bottom-up' approach is manifested through the use of two specific machine learning methodologies: variable selection, for identifying important variables for our well-being and given those, if there exists a relation between high temperatures and well-being. And secondly, dimensionality reduction to provide a high-level overview of the life aspects impacting well-being. Both methods pointed towards a range of factors influencing well-being, and hinted at a consistent, albeit weak, link between well-being and temperature.

In concluding the introduction, it is essential to clarify a specific aspect of this thesis concerning SWB. Although the relationship between SWB and other variables is complex and multifaceted, the goal of this research is not to identify causal relationships. The multifaceted nature of SWB presents significant challenges in defining a clear sequence of events and eliminating alternative explanations. This complexity is a recognized issue within this field. Instead, this thesis focuses on investigating whether there is a statistical association between the variables, without attempting to establish causality.

The subsequent chapter provides a brief overview of well-being research, particularly in relation to temperature. Thereafter, a chapter dedicated to data exploration delves into the German Socio-Economic Panel dataset and presents the theoretical underpinnings of our chosen machine learning methodologies, explaining our methodological steps. Chapter 3 details our research procedure and chapter 4 the experimental results. In conclusion, the thesis summarises our findings in chapter 5 and proposes potential pathways for future research, picking up from where this study leaves off.

Chapter 2

Related Work

2.1 Subjective Well-Being

Subjective well-being (SWB) is a multifaceted construct that has garnered significant attention across various fields of study. It encompasses numerous factors that shape an individual's perceived sense of well-being, reflecting a broad spectrum of influences that vary across personal and societal dimensions.

Health stands as an essential contributor to SWB. Research underscores the intricate relationship between both physical and mental health and an individual's perceived well-being, emphasizing the mutual influence these aspects exert on one another [19].

Income's role in shaping SWB has been widely acknowledged. Empirical studies have found a positive correlation between income and well-being. However, this relationship exhibits diminishing returns beyond an income threshold of 74,000 U.S. dollars, indicating a complex interaction between financial resources and perceived contentment [14].

Demographically, both cohort and age serve as vital components that influence SWB. Well-being appears to vary across different generations and age brackets, reflecting the intricate interconnection between generational effects and the aging process [5, 23].

Parenthood presents a complex relationship with SWB, characterized by conflicting findings. A comprehensive model exploring the link between parenthood and well-being identifies various mediators and moderators, such as financial strain, emotional factors, and social roles that can either enhance or diminish happiness in parents [17].

The phenomenon known as scarring, or the lingering negative effect of unemployment on SWB, has been substantiated through robust empirical research. This enduring impact persists even after reemployment, with the scars affecting overall life satisfaction for a minimum of five years, and is particularly exacerbated by repeated unemployment periods [7].

Urban housing pressure has emerged as a noteworthy determinant of SWB. A large-scale study in China found both subjective and objective housing pressures to be negatively correlated with SWB, with nuanced variations observed across different city sizes and housing tenure groups [25].

The nexus between social relationships and SWB has been thoroughly examined, highlighting marriage's positive correlation with SWB, although the strength of this relationship has been found to be weaker than initially anticipated [11]. Moreover, a broader meta-analysis showed that strong social relationships increase survival likelihood by 50%, a factor not directly tied to SWB but potentially related as an indirect aspect, extending the influence of social relationships beyond mere well-being to overall health and survival [12].

This extensive collection of studies illustrates the multifaceted and complex associations with well-being. Understanding these preexisting relationships is crucial when evaluating causal links. It should be noted that this collection of academic papers is illustrative rather than exhaustive and primarily represents the most well-established variables associated with well-being. Furthermore, cultural differences and other contextual factors may lead to variations across different datasets, adding further complexity to the understanding of SWB.

2.2 Temperature and Well-Being

There is to our knowledge only very little literature seeking the association between temperature and subjective well-being in Europe. However, as hinted at in the introduction, well-being is a term that can be interpreted in many ways. So there is still valid literature motivating a association between temperature and well-being.

One paper[13], estimate the effects of local temperature on depression in China using two waves of the China Family Panel Studies (CFPS). They merge data from 699 weather stations in China with the CFPS and then estimate a series of fixed-effects models with the two waves of the CFPS. Their analysis shows that high temperatures negatively affect (among others) elderly individuals and females. Especially temperatures above 30 degrees increases the probability of depression.

In a recent and much-publicized report, Gallup analyzed the impact of rising temperatures on people's lives by using geospatial information on respondent locations together with Gallup survey data from 1.75 million people in 160 countries [8]. Their results show that rising temperatures have a significant effect on well-being, especially among older generations and people in poorer countries.

There is are of course also other effects that temperature that whose results could potentially be mirrored in well-being. For instance the effect weather has on physical health as indicated by an increase in hospitalisations in hot weather[22]. There are also indications that there is a negative correlation between ambient tmeperature and sleep-quality[18] or productivity[21].

In summary there is a solid foundation in literature that motivates that there could be an association between well-being and temperature.

Chapter 3

Methodology

3.1 Data Collection & Preprocessing

Within the confines of this section, an exhaustive and detailed overview of the data at hand is laid out. This encompasses a comprehensive examination of the data's assembly, merging, and purification processes. Furthermore, an exposition of the feature engineering steps is explained, showing the methodologies employed to structure each variable in a manner that is compatible with the machine learning algorithms employed in this research.

3.1.1 Source of Data

The final dataset is an amalgamation of two distinct datasets. The primary and more extensive one is the Socio-Economic Panel (SOEP) dataset, furnished by the "Deutsches Institut für Wirtschaftsforschung" (DIW) or in English, the German Institute for Economic Research. Established in 1984, the SOEP is a multi-modular longitudinal study that annually gathers data on private households in Germany, offering a comprehensive look into the fluctuating living conditions and beliefs of the German populace. It spans a broad range of topics, from income, employment, and education to health, life satisfaction, and household composition.

To streamline the analysis, an initial step was taken to refine the variables under consideration. From the various modules in the SOEP study, only those that seemed to contain relevant variables were included. Four modules were singled out, each encompassing a broad set of variables[1]:

- pl: This contains personal data that encapsulates individual information.
- hl: A module dedicated to household data.

[1] https://paneldata.org/soep-core/datasets/

- regionl: Focusing on region-specific information, this module provides details on household locations, proving pivotal for the subsequent merging with meteorological data from Meteostat.
- health: A dedicated health module offering more in-depth health data than what's present in pl. A notable component of this module is the 12-Item Short Form Survey (SF-12), a widely-used tool in research to provide a health summary statistic[24].

The SOEP dataset, while extensive, does not inherently encompass weather data. However, investigating the potential relationships between weather conditions, particularly temperature, and well-being is an area of interest. To facilitate this analysis, weather data was sourced from an external, publicly available resource known as Meteostat.[16] Meteostat provides historical weather data from different weather stations, including information about temperature, precipitation, humidity, and other meteorological variables across Germany.

The raw Meteostat data underwent a transformation using a rolling average across a 7-day span. This transformation enabled the variables to approximate long-term exposure to specific climatic conditions. A 7-day window was chosen because it provided a time frame that was sufficiently extended to potentially influence well-being, but not so lengthy that it caused all variation to converge to the mean. Although inspecting slightly larger timeframes could be performed to test for robustness, considering the results obtained, further exploration along this avenue was not pursued in this thesis.

The SOEP data was spatially joined with Meteostat's meteorological dataset at the NUTS 1 level, which in Germany corresponds to the Bundesländer or federal states.[16]. It is an inherent assumption that the average temperature within a Bundesland is sufficiently precise for this study's objectives. Thus, while a more detailed level might have offered richer insights, the combined dataset, blending socio-economic and meteorological data, provides a unique interdisciplinary view.

3.1.2 Data Cleaning

The current state of the data still has some slight issues. The SOEP study is ever evolving and new variables are added and removed ever year. This results in a big set of variables that have many missing values. This is clearly visible in figure 3.1 where only in some years there are some non-missing values.

In order to avoid having a very sparse dataset with many missing values some preliminary variable screening techniques were employed. As a primary screening technique only data points after the year 2000 were considered and variables that had maximal missing rates of 80%.

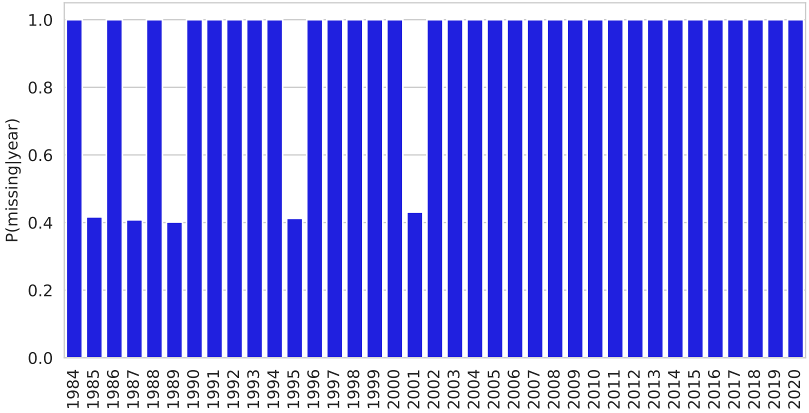

Figure 3.1: Example: Missing Values

The reason for choosing the cutoff year 2000 is because the SOEP study introduced a new health module that gives more detailed information on an interviewees health than in other modules.[3] Given that health is a strong contributor to well-being and more importantly a possible confounder between weather and well-being this module was included.

The missing rate cutoff of 80% had two reasons: • With this prescreening step still around 80 variables remain making it a sizeable dataset. • Most of the variables are actually very sparsely represented in the data. Many survey years do not include certain variables because the variable was only temporarily important during the SOEP study. Figure 3.1 depicts one of these sparsely represented variables labelled *Job Is High Stress* as an example. This step therefore removes all these sparse variables.

The data provided by Meteostat does not require cleaning as there are no missing values. A complete description of all variables can be found in the table A.1.

An issue with the current data is that most interviews are conducted in winter by the DIW. This generates a strong bias towards low temperature observations. To counter this bias the observations lying between the months May and September were sampled from the overall population. Figure 3.2 depicts the change in distribution when using this sampling step. The subsample made it easier to focus the investigation on high temperature observations than when taking the whole population.

Subsequent to all preprocessing steps, the dataset was comprised of $44'457$ observations.

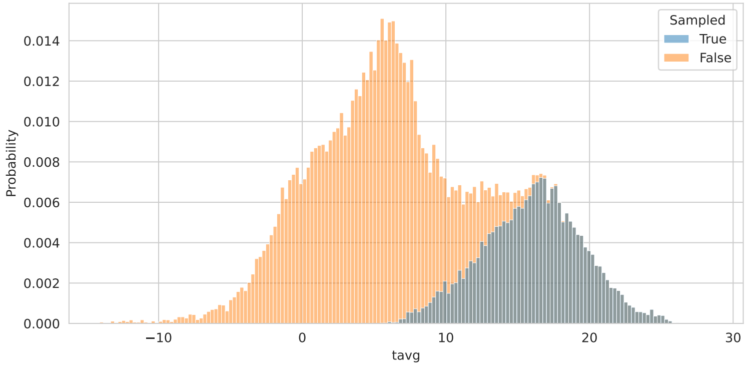

Figure 3.2: Distribution of Temperature

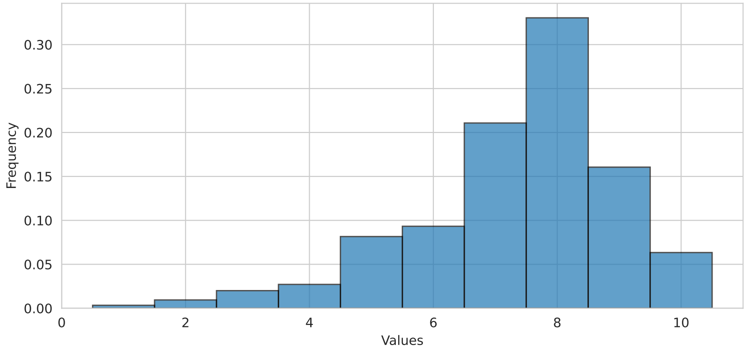

Figure 3.3: Distribution of Subjective Well-Being

3.1.3 Target Variable

The target variable, i.e. the variable we want to be able to predict, is an index called subjective well-being, or life satisfaction. The way it was collected by the DIW is by asking the question "How satisfied are you with your life from 1 to 10 all things considered?". The figure 3.3 displays the distribution of this variable.

3.1.4 Feature Engineering

In this section it is discussed which post merging steps were taken to make new variables and transform existing ones to make them suitable for our machine leaning models.

The former concerns only a small number of variables. One of which was the variable *age* which was calculated from the variables *birth year* and *survey year*. The age variable is included for one because research has shown that there exists a so-called "ageing effect"[23]. Furthermore, this allows the models to find potentially different effect sizes across different demographic groups.

Given these variables the next step included the preparation of the dataset to our machine learning models. The preprocessing was intended to be homogeneous such that each machine learning algorithm could use the transformed data. The steps included:

1. **Variable Removal:** We eliminated certain columns from the dataset that were deemed unnecessary for the analysis. These included variables identified as irrelevant, those sharing a prefix and potentially carrying redundant information, and identifying variables such as pid. Most of these variable were key columns uniquely identifying either individuals our households.

2. **Isolating Target Variable:** The target variable is then spitted from the main data. All subsequent processing steps are not performed to the target to avoid inducing biases.

3. **Missing Data Treatment:** Missing values were interpolated using backfilling with a limit of one year within each respondent. The assumption here is that most variables do not change drastically over the period of one year. The remaining missing values were dropped.

4. **Data Segregation:** Initially, the variables were systematically classified into distinct categories: Nominal, Ordinal, Discrete, and Continuous. Depending on their classification, each category underwent a specific treatment to optimize data representation and analysis.

    - **Nominal:** Nominal variables underwent a transformation process called dummy encoding. In dummy encoding, a categorical variable is transformed into a set of binary variables that capture all the information of the original variable. One challenge of dummy encoding is the potential for dataset expansion, especially when dealing with variables of high cardinality. However, in this dataset, we were fortunate as none of the nominal variables exhibited high cardinality, making dummy encoding a practical choice.

    - **Ordinal:** All the ordinal variables were adjusted using the min-max scaler, which transformed their values to lie within the interval $[0, 1]$. The benefits of this method of rescaling are twofold:
        - It retains the inherent order and relative significance of the original values.
        - It eradicates any biases that could be introduced due to the original scale of the variable.

As an example, consider an ordinal variable representing the education level, ranging from 'Primary' to 'PhD'. Using the min-max scaler, 'Primary' might be scaled to 0, and 'PhD' to 1, with all other levels appropriately scaled in between based on their order.

- **Discrete:** Discrete variables also benefitted from the min-max scaler to standardize their range. The rationale behind this approach is the same as that for ordinal variables: to ensure consistent data interpretation while preserving the relative differences among the discrete data points.
- **Continuous (Numerical):** For continuous variables, we aimed to center and normalize their distribution. Thus, every continuous variable was adjusted to have a mean of 0 and a standard deviation of 1. Such scaling is essential when dealing with data that follows a specific distribution, like a normal distribution. As an example, consider a variable like income. While the exact monetary value might not always be crucial, understanding where an individual's income stands in relation to the broader population can be informative. By normalizing the income data, we can assess relative income levels more effectively.

5. **Data Consolidation:** The processed feature classes were then integrated to generate a comprehensive dataset.

Note 3.1.1. Rescaling variables is a foundational step in many statistical analyses and machine learning models because it ensures all variables operate on a consistent scale. When variables are on disparate scales, it can be challenging to directly compare the magnitude of their effects. For instance, a change in a variable measured in thousands (e.g., yearly income) might seem minuscule when compared to a change in a variable measured in single units (e.g., number of children). By rescaling, both variables could range between 0 and 1, which means that a unit change in either variable represents a proportionate change in its range, facilitating a clearer and more direct comparison. For example, after rescaling, a change from 0.2 to 0.3 in both the normalized income and the number of children would reflect a consistent relative change, making it easier to inspect and compare the magnitudes of their effects in a model.

3.2 Analytical Methods: Overview

In this section, we delve into the primary analytical methods employed, offering a comprehensive overview of their respective strengths and limitations. Additionally, we clarify the rationale behind selecting these particular methodologies. Care is also taken to guide readers on appropriate interpretation of the results, highlighting potential pitfalls in over-analysis or misinterpretation.

3.2.1 Introduction to Methods

The idea of this particular analysis is to create different models that can accurately predict the well-being of a person given its living situation (economic, social and other statuses). Due to the nature of the well-being target variable, as described in section 3.1.3, this will be a regression problem.

These trained models are then inspected on, firstly, how well they actually work, and secondly, what the concrete effects of each variable are on the predictions. This allows us to form a deeper understanding of how the model is making predictions.

Concerning the accuracy of the predictions we will employ two different measures. Namely $R2$ and means-squarred error (MSE). We use the two because they measure two things that are both important for the analysis. The MSE value is a good indicator of how accurate the individual predictions are. The $R2$ value on the other hand is a good suggestor on how well the variance of the target variable is explained. Given the distribution of the target variable a dummie predictor such as a mean predictor would actually work quite well with an average MSE score of 2.8882. However, such a predictor would not explain the variance very well and therefore receive a very low $R2$ value, or in this case and $R2$ score of 0. The two metrics are displayed in equations 3.1 and 3.2. Another reason to use $R2$ is to make this work comparable to existing research as most models there achieve an $R2$-value between 5% and 30% depending of the covariates included in the analysis using the same dataset.[6, 10, 4]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.2}$$

In the next sections a brief explanation of the machine learning models used is given with their respective strengths and weaknesses.

3.2.2 Linear Regression

Linear Regression is arguably among the most prevalent machine learning models, revered for its straightforwardness and interpretability. At its core, this model seeks to deduce the optimal linear relationship between predictors and the outcome. Mathematically, this endeavor can be depicted as $y = \beta_0 + \beta X + \epsilon$, where y represents the dependent variable's vector, β denotes the coefficients under estimation, X is the matrix of predictors, and ϵ signifies the residual error.

The primary optimization goal in linear regression is to determine the coefficients (β) that minimize the mean squared error between the predicted outcomes ($\hat{y}$) and the actual values of the dependent variable (y). We can formalize this as an objective function J that needs to be minimized:

$$J(\beta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 \tag{3.3}$$

While its simplicity is an advantage, linear regression also brings with it several assumptions. Key among these is the presumption of a linear relationship between predictors and the dependent variable, and the expectation that predictors are not highly correlated with one another. Deviations from these assumptions can compromise the model's validity, potentially yielding biased estimations.

When dealing with the risk of overfitting to the training data in linear regression, especially with datasets populated by a large number of predictors, regularization techniques such as Ridge and Lasso regression come into play. These methods modify the linear regression loss function by adding a penalty term, constraining the magnitude of coefficients.

1. **Ridge Regression (L2 Regularization)** The objective function (loss function) for Ridge Regression is:

   $$J(\beta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.4}$$

   The first part of this equation is just the Mean Squared Error (MSE) of a linear regression. The second part is the $l2$ penalty, with λ being the regularization strength. A larger λ means a stronger penalty, pushing coefficients toward zero but not exactly to zero.

2. **Lasso Regression (L1 Regularization)** The objective function for Lasso Regression is:

   $$J(\beta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.5}$$

   The first segment again represents the MSE of a linear regression. The second segment introduces the $l1$ penalty. The strength of this penalty is controlled by λ. A notable feature of Lasso is its ability to reduce some coefficients to an exact zero, effectively performing variable selection.

3. **Elasticnet Penalty:** This is a combination of the two penalties. The objective function can be defined as:

$$J(\beta) = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \beta_j^2 \tag{3.6}$$

To clarify, Ridge regression corresponds to the $l2$ regularization, adding a penalty equivalent to the square of the magnitude of coefficients. On the other hand, Lasso regression corresponds to $l1$ regularization, penalizing the absolute value of the coefficients. The choice between Ridge and Lasso typically hinges on the specific problem and the nature of the dataset at hand. Note that the assumption of linearity and Independence equally hold for these regression models.

Note 3.2.1. It is essential to recognize that utilizing regression may not optimally address the analysis of our target variable, given its ordinal character. The underlying assumption of consistent differences in regression might be challenged when considering our subjective well-being index: the difference between values of 8 and 9 might not inherently convey the same significance as that between 3 and 4. Nonetheless, some of the models delineated below possess the capability to accommodate these non-linear relationships within the target variable indices (see sections 3.2.3, 3.2.4, 3.2.5).

3.2.3 Random Forest

A random forest is a member of the ensemble method family. Ensemble techniques harness the collective power of multiple weak learners to craft a more robust and high-performing model. While applicable to both regression and classification tasks, our discourse will predominantly revolve around the regression scenario. Nonetheless, it is crucial to understand that the foundational principles remain largely congruent across both realms.

The nomenclature "random forest" originates from the method's reliance on "decision trees" as its constituent weak learners. At its core, a decision tree seeks to recursively pinpoint a split based on a single feature. This split aims to minimize the variance of the resulting subgroups. A more comprehensive and formal exposition on this can be found in appendix A.1.1.

However, a caveat with single decision trees is their susceptibility to overfitting. This vulnerability arises because:

1. High Complexity: Trees, especially deep ones, can capture noise in the data, mistaking it for a pattern.

2. Sensitivity to Small Variations: Minor changes in the data can result in dramatically different tree structures.

3. Bias Toward Features with More Levels: Trees can show a preference towards features with numerous levels, as they offer more opportunities to split and thus can seem more informative than they truly are.

In contrast, random forests mitigate these issues through ensemble learning, bringing in the diversity of multiple trees and ensuring a more generalized model. The way it works is as follows:

1. **Bootstrap Sampling:** Given a dataset D, we randomly select $N \leq |D|$ samples with replacement to form a new dataset D_i. This process is repeated M times to create M different bootstrap datasets. Note that N and M can be tuned as you like. But generally we choose $N = |D|$ if the dataset is not too large, and M can be as large as you wish as long as it stays computationally feasible.

2. **Train Decision Trees:** For each bootstrap dataset D_i, train a decision tree T_i. During the training of each tree, at each node, a random subset of features is chosen to determine the best split (this introduces more randomness and decorrelation between trees).

3. **Prediction:** Given a new data point $x \in D_i$, each tree T_i in the forest gives a prediction $\hat{y}$. The final prediction of the random forest for regression is the average of all these predictions:

$$y(x) = \frac{1}{M} \sum_{i=1}^{M} T_i(x) \tag{3.7}$$

Random Forests are renowned for their robustness against overfitting, a common ailment of standalone decision trees. By combining multiple models, they often achieve high accuracy, tapping into the collective ”wisdom of the crowd.” Furthermore, they offer valuable insights into feature importance and can be parallelized for faster training, making them suitable for both regression and classification tasks, even in the presence of missing data. However, these benefits come with trade-offs. The complexity of a Random Forest, consisting of multiple trees, can make it a challenge to interpret. Training can be computationally intensive, especially with voluminous datasets. Also, while predictions can be accurate, the need to traverse all trees can introduce latency in real-time applications. Lastly, in scenarios with abundant noisy features, they may inadvertently introduce bias, overshadowing the genuinely informative features.

3.2.4 Histogram-based Gradient Boosting

The Histogram-based Gradient Boosting algorithm, belongs to the ensemble and boosting umbrella of methods. Boosting methods operate on the philosophy of training weak learners sequentially, with each subsequent model attempting to correct the errors of its predecessor. These techniques are flexible and can be adapted to both regression and classification problems. However, for the sake of our present discussion, the emphasis will predominantly be on the regression context.

Gradient Boosting's name is derived from its procedure of employing gradient descent to minimize loss, combined with the boosting technique. The fundamental idea is to add new models to the ensemble sequentially. Each new model fits to the negative gradient (or the residual errors) of the cumulative ensemble of preceding models. Over iterations, this leads to a collective model that can capture intricate patterns in data, often outperforming standalone models.

Contrary to Random Forests, which build trees in parallel and merge their predictions, Gradient Boosting constructs trees sequentially. This key difference can be broken down as follows:

1. **Initialize:** Begin with a simple model, often just a constant value.
2. **Compute Residuals:** Calculate the residuals (or the negative gradient of loss) between the predictions of the current ensemble and the actual values.
3. **Fit a Tree:** Train a shallow decision tree to these residuals.
4. **Update Ensemble:** Add this new tree to the ensemble, often with a learning rate to prevent overfitting.
5. **Iterate:** Repeat the process, each time fitting trees to the residuals of the current ensemble's predictions.

The ”Histogram” in the name signifies a key optimization in the algorithm. Instead of using original continuous features, Gradient Boosting algorithms discretize them into discrete bins, thereby speeding up the training process without a significant compromise on accuracy.

Histogram-based Gradient Boosting combines the strengths of decision trees with the prowess of gradient descent. Its sequential nature aids in addressing the biases and errors of prior trees, often leading to remarkably accurate models. The algorithm can also inherently handle missing values, thereby obviating the need for imputation. However, this comes at a price. Gradient Boosting models, especially when not tuned adequately, can be susceptible to overfitting, especially on noisy datasets. Training is inherently sequential, which can be time-consuming for large datasets and contrasts with the parallelizability of Random Forests. Lastly, the model's additive complexity can challenge interpretability, similar to the intricacies posed by Random Forests.

3.2.5 Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP), commonly known as a neural network, is a type of feedforward artificial neural network. It has at least three distinct layers: an input layer, one or more hidden layers, and an output layer. Every node within a layer connects to every node in its adjacent layers through weighted pathways. For a deeper dive into its mechanics, see appendix A.1.2.

Central to the MLP is its objective function, usually tied to a loss function. This measures the difference between the predicted output and the actual data. For regression tasks, the Mean Squared Error (MSE) is a frequent pick, described in equation 3.1.

MLPs offer several advantages. They can model complex, non-linear relationships, making them more versatile than some linear models. Additionally, with enough neurons, an MLP can theoretically approximate any continuous function. They're also scalable and, when designed right, can handle large datasets and find subtle patterns within.

However, they have their challenges. MLPs can overfit, particularly if they aren't regularized or if they have more parameters than training samples. Their intricate structure can make them hard to interpret, often labeled as a "black box" model. They perform best with a lot of data; limited or noisy data can diminish their effectiveness. Furthermore, training an MLP can be resource-intensive and may require careful tuning of various settings.

3.2.6 Model Refinement and Hyperparameter Tuning

Most of the models discussed sofar have parameters that are required to be specified upfront when constructing the model. These parameters are not derived from the data but are set prior to training, thus the term "hyperparameters". For instance, in Lasso or Ridge regression, a pivotal hyperparameter is the regularization strength, denoted by the λ parameter (see equations 3.5, 3.4 and 3.6). For Random Forests, hyperparameters might include the maximum depth of trees, the criterion for splitting nodes, the number of trees in the forest, and the percentage of features considered at each split, among others.

Finding the optimal set of hyperparameters for a specific problem is an ongoing field of research, with various techniques proposed over the years. One popular method is the Tree-structured Parzen Estimator (TPE) sampler. The TPE sampler operates based on a Bayesian optimization framework. At a high level, the TPE builds a probabilistic model that tries to estimate the likelihood that a given set of hyperparameters will yield a performance improvement over previously evaluated sets. Instead of performing a grid or random search over the entire hyperparameter space, TPE selectively samples the regions of the space that are likely to offer better results, thus typically leading to faster convergence to optimal values.

3.2.7 Feature Importance

Up to this point, we have discussed how to develop predictors based on specific criteria, the construction of the model, and techniques to refine it. If the sole goal was to have a reliable predictor, this would be the end of our discussion. However, the objective of this analysis extends beyond mere prediction: it is to understand and interpret the data. After training the model, it's pivotal to examine it to discern which variables significantly influence its predictions. While there's a plethora of ways to measure variable importance, we'll delve into one of the simpler yet powerful methods.

For this research, we have adopted *Permutation importance* as our feature importance measure. At its core, permutation importance involves randomly shuffling a single predictor or feature and measuring how much the model's performance deteriorates. If the model heavily relies on this predictor for its predictions, we'd expect its performance to drop significantly; otherwise, the decline would be minimal. The pseudo-code for this method can be found in pseudocode 1.

Algorithm 1 Permutation Importance

```
function PermutationImportance($X, y$, model)
    $n \leftarrow$ length($X$)
    $m \leftarrow$ length($X[1]$)
    importance $\leftarrow$ array of zeros($m$)
    for $i \leftarrow 1$ to $m$ do
        original_scores $\leftarrow$ model_score($X, y$, model)
        permutation $\leftarrow$ shuffle($X[i]$)
        permuted_scores $\leftarrow$ model_score($X, y$, model)
        importance[$i$] $\leftarrow$ mean(original_scores $-$ permuted_scores)
    end for
    return importance
end function
```

The allure of permutation importance lies in its model-agnostic nature. Unlike methods specific to certain models, like in linear regression where coefficients denote feature importance, permutation importance can be applied across different models. However, this approach is not without its challenges:

1. The accuracy of the results hinges on the model being well-trained and validated. Before conducting permutation importance, it is crucial to ensure that the model has been rigorously evaluated. To enhance reliability, the permutation importance procedure should be repeated several times. Another strategy to augment robustness is to retrain the model with different bootstraps, then execute the permutation importance algorithm on each variant. Consistent results across these models would corroborate the significance of the variable.

2. Permutation importance can be computationally taxing. This is mainly because each feature must be individually shuffled and the model's performance recalculated, which can be time-consuming, especially with large datasets or complex models.
3. A high permutation importance score does not automatically equate to a feature's meaningfulness or real-world relevance. It merely underscores the feature's significance in making predictions. For instance, a binary variable might greatly influence predictions, but if it's heavily imbalanced, shuffling its values might affect only a fraction of the predictions. Thus, its score change might be negligible, deeming it less significant than it actually is in the context of the model.
4. Correlated features can distort permutation importance scores. If two or more variables are highly correlated, shuffling one might not impact the model's performance much since the other correlated variable(s) can still provide similar information. This can lead to underestimating the importance of correlated variables.

Another use case of feature importance in general is to select a subset of variables from the data that will still result in optimal predictions. In the case of this analysis this will be very usefull as most of the variables included in the dataset will be irrelevant for the predictions and therefore only inject more noise into our model, effectively enhancing the chances of overfitting.

Note 3.2.2. One thing that should be kept in mind is the magnitude of the permutation importance, as it depends on the scoring method used. The correct way of thinking about the permutation importance is to view it as the change in the respective score when the variable has been permuted. For instance, if using mean absolute error as the scoring method, a permutation importance of 0.05 for a variable means that shuffling its values leads to the model making less accurate predictions by a margin of 0.05 with respect to the unshuffled data. This underscores the variable's influence on the prediction.

3.2.8 Partial Dependence

In assessing the impact of high temperatures on well-being, merely relying on permutation importance might fall short. For instance, given the temperature variable distribution, as illustrated in figure 3.2, high temperature values are underrepresented. To unravel the effects of notably high values, we turn our attention to an analytical tool named partial dependence.

Partial Dependence (PD) offers a window into the model by showcasing how a feature affects predictions, holding other features constant. In essence, it varies a feature of interest across its range and computes the average prediction. PD plots, thus, help us discern how the model's output varies across different feature values. For a more detailed explanation see pseudocode 2.

Furthermore, PD plots can elucidate interactions between two features. For instance, one might investigate the joint effect of temperature and humidity on well-being, keeping all other features static. This capability can be invaluable, especially when interactions are suspected or when they're of particular interest.

A key point to be aware of is that while PD plots portray an average effect, the relationship might not be strictly monotonic for all instances. In other words, a plot displaying a mostly increasing relationship doesn't necessarily dictate that every individual prediction follows this pattern. It's a representation of the aggregate effect.

Lastly, while tools like feature importance highlight influential features, PD plots go a step further, illuminating the nature of said influence. They might reveal insights such as: "As temperatures rise, well-being tends to decline, but only up to a certain threshold, after which there's minimal effect."

However, a note of caution: as with permutation importance, the reliability of PD is tethered to the model it springs from. A flawed model will yield misleading PD plots. Hence, ensuring model robustness is paramount. One effective strategy to reinforce trust in these plots is to compute the partial dependence over various sets of resampled data. This not only validates the observed trends but also gives a sense of their stability.

Algorithm 2 Partial Dependence Computation

```
Require: Trained Model $M$, Dataset $D$, Feature of interest $F$
Ensure: Partial Dependence values for feature $F$
  Initialize an empty list $PD$
  for each bin $v$ in feature $F$ do
      Set feature $F$ in all instances of $D$ to value $v$
      Predict using model $M$ on modified $D$ to get predictions $P$
      Compute the average prediction $avgP = \frac{1}{|D|} \sum_{i=1}^{|D|} P_i$
      Append $avgP$ to list $PD$
  end for
  return $PD$
```

The method of studying Partial Dependence (PD), especially when combined with resampling techniques, can serve as a potent strategy for revealing subtle connections in the data. This is particularly valuable when dealing with features that have infrequent values, such as elevated temperatures.

3.2.9 Resampling

Resampling techniques, such as bootstrapping, are used to estimate the distribution of a statistic (like the mean or variance) by drawing with replacement from the data. Bootstrapping, in particular, provides a measure of robustness to the estimates derived from a model.

Why is this important? When building any statistical or machine learning model, the robustness of the model's outcomes is essential. If slight changes in the data lead to significantly different results, the model may not be reliable. Resampling methods like bootstrapping give a way to simulate these slight changes by repeatedly resampling the data and recalculating the metric or model.

Furthermore, in techniques like cross-validation, it is critical to ensure that the distribution of the target variable is consistent across different folds, especially if the data is imbalanced. Stratified KFold addresses this by ensuring that each fold retains the same distribution of the target variable as in the entire dataset. In other words, if 20% of your entire dataset belongs to class 1 and 80% to class 2, Stratified KFold will maintain this distribution in each of its K folds. This is vital for producing reliable and consistent model evaluations, as each fold will be a representative sample of the entire dataset.

Chapter 4

Experimental Results

4.1 Software

In the following sections it is discussed how the above mentioned methodologies were applied to this problem. To begin, specific python packages were used to make this analysis possible. Most of the analysis is done using the $scikit-learn$ package.[20] This package offers a versatile set of tools allowing to create robust machine learning pipelines. another advantage of $scikit-learn$ is that many other machine learning packages offer wrapper functions that allow for easy integration of custom models into the $scikit-learn$ pipeline. One such package used in this project is the $tenserflow$ API, which is a framework for building custom neural networks.[1]

The last module required for this project was $optuna$ which is quite a new hyper-parameter optimization framework. It already provides the aforementioned TPE sampler to find an optimal set of hyper-parameters.[2]

The complete code is posted on GitHub[1] Unfortunately, due to contractual limitations, I am prohibited from sharing the data. It would need to be requested separately from the German Institute of Economics (DIW).

4.2 Feature Importance

The initial phase of our analysis began by employing all models on the preprocessed dataset as outlined in section 3.1. This preliminary exercise provided insights into the key variables that influence subjective well-being. The series of actions executed during this phase include:

1. The dataset was segregated into training and testing subsets.

2. Every model detailed in section 3.2 underwent fine-tuning via a TPE sampler. Each proposed set of hyperparameters underwent validation

[1] `https://github.com/marcosousapoza/bachelor_thesis`

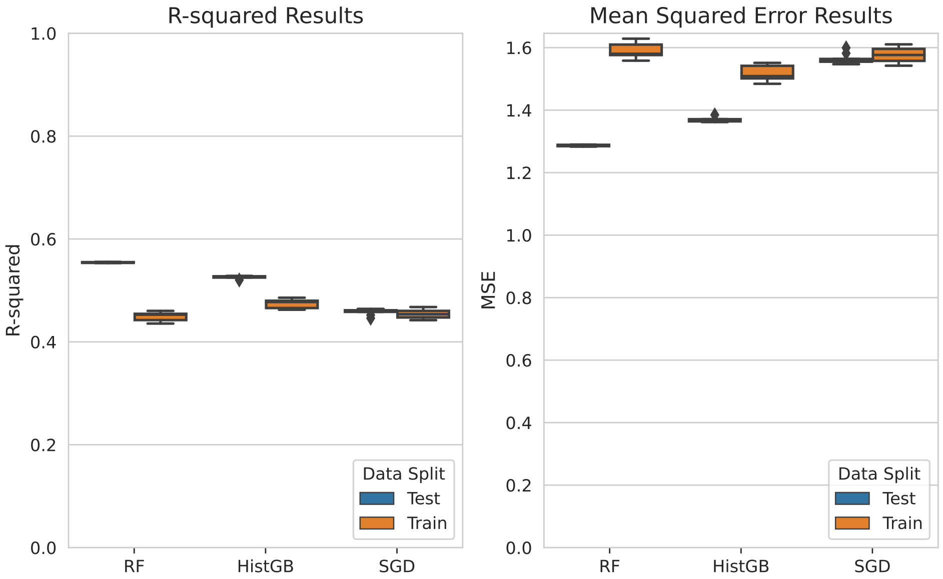

Figure 4.1: Model Performance

across 5 stratified folds, ensuring the robustness of the hyperparameters. The final score was determined by averaging the R^2 and MSE across the five test folds. A total of 150 different variable combinations were proposed, and the outcomes indicated a convergence in performance. The superior performance of the best models is illustrated in figure 4.1.

The graph reveals that the showcased models have an impressive R^2 test score exceeding 40%. These findings therefore outperform the exisintg models assessing subjective well-being by explaining at least 10% more of the variance in existing research. However, a discernible degree of overfitting is apparent in the ensemble methods. The models tend to yield better outcomes on training data than testing data, possibly due to the multitude of inconsequential variables in the dataset, causing noise and model confusion. Notably, the linear regressor, labeled as *Stochastic Gradient Descent*, is immune to this as it deploys an elastic net penalty, effectively sidelining irrelevant variables. The neural network model is absent in the plots due to its subpar performance in these tests, attributed to the excessive unrelated variables. However, subsequent tests revealed enhanced performance.

3. The third phase exclusively focused on pinpointing pivotal features. For this endeavor, permutation importance was used. As emphasized in section 3.2.7, robustness was assured by re-evaluating and retraining the model across multiple data segments. Here, 10 stratified folds were

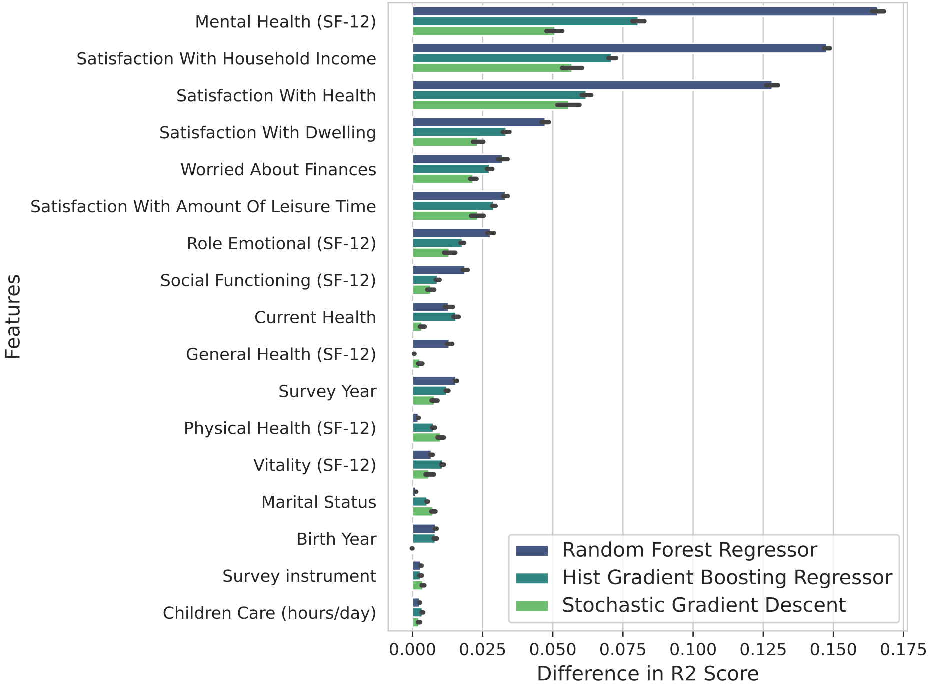

Figure 4.2: Permutation Importance

used to retrain the model discovered by the TPE sampler. Following this, the permutation importance for each variable was determined based on the trained models and the reserved validation subset. The most influential variables, based on permutation importance scores, are depicted in figure 4.2. The y-axis portrays variables, each prefixed with its respective category. For readers keen on the variable definitions, a detailed table is accessible in appendix A.1. The paramount variables are elaborated upon in the ensuing section.

4.3 Feature Importance - Findings

Figure 4.2 showcases all variables that have a permutation importance higher than 0.0025. These findings are consistent with previously conducted research. Here is a closer examination of some key variables:

1. **Mental Health**: Sourced from the SF-12 questionnaire and z-transformed, its prominence in affecting well-being is not unexpected, given the extensive literature supporting such an association.[19]

2. **Satisfaction With Household Income**: Ranking second in importance, its connection to life satisfaction has been well-documented.

The correlation between life satisfaction and income is widely accepted in literature.[14]

3. **Satisfaction With Health**: This raises concerns about causality, especially considering its potential strong correlation with mental health. The question of whether life satisfaction affects mental health or vice-versa is valid and merits further exploration. Nonetheless, these inquiries exceed this thesis's scope and primary focus. However, it is crucial to approach these findings judiciously and recognize their potential limitations.

4. **Satisfaction With Amount Of Leisure Time**: Though its predictive power is waning compared to earlier variables, its significance in well-being literature remains firm.[15]

5. **Satisfaction With Dwelling**: This too is acknowledged as a pivotal well-being factor in related studies.[25]

6. **Worried About Finances**: The cause-and-effect relation here, especially concerning **Satisfaction With Household Income**, invites scrutiny.

7. **Role Emotional**: The "role emotional" domain, or RE, tends to have questions that evaluate how much emotional problems might have hindered someone's ability to perform their daily tasks or roles.

8. **Current Health**: Similar to earlier variables, the causality of its relation with the target remains under deliberation.

9. **Survey Year Identifier**: This can be perceived as reflecting the cohort effect, a phenomenon also substantiated in literature.[23]

It is crucial to emphasize that the omission of a variable from this figure does not negate its significance. For example, although weather variables might seem less important in this context, their outliers could still have notable implications for the results.

4.4 Effect of Temperature

Now that all the most important variables have been identified the whole pipeline is reiterated with a subset of all the variables. With most important variables is meant all the variables with a higher permutation importance score than 0.0025. The decision to utilize a permutation importance cutoff of 0.001 was not arbitrary, but a product of meticulous analysis and considerations of model optimization. The cutoff value represents a balance between ensuring model simplicity and maintaining predictive performance.

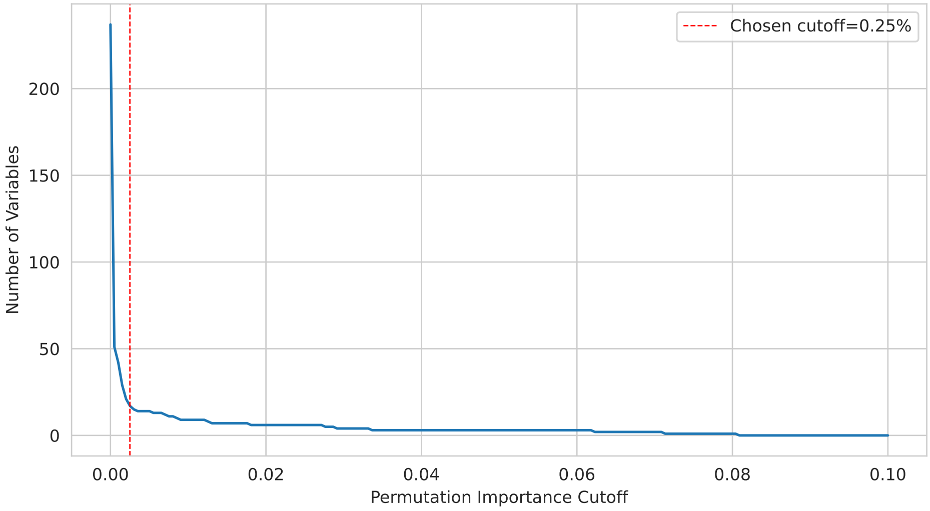

Figure 4.3: Permutation Importance Cutoff

By examining Figure 4.3, it's evident that beyond this threshold, the incremental gain in prediction accuracy, for the addition of more variables, begins to plateau. Including too many variables might introduce unnecessary complexity, making the model harder to interpret and potentially prone to overfitting. On the other hand, setting the cutoff too high might exclude potentially important variables, thereby reducing the model's efficacy. The selected threshold of 0.0025 offers a sweet spot, ensuring that the model remains both robust and interpretable, while capturing the most salient features that drive predictions.

Since the primary goal of this inviestigation was to see the effects of temperature on well-being the weather related variables were exempted from the selection criterion.

Other than the selection of the variables all the steps as performed in section 4.3 are identical. The performance on the test and validation sets can be found in figure 4.4. There is no notable difference in performance when comparing Figures 4.1 and 4.4 which further justifies the selection of the variables.

After evaluating the models, the partial dependence between temperature and subjective well-being (SWB) was analyzed. This analysis highlights varying effect sizes of temperature across different quantiles of its distribution. For enhanced reliability, the models were retrained using 10 stratified folds, and partial dependence was computed on the validation set. The findings are illustrated in Figure 4.5. The x-axis displays the temperature feature's value range, and the y-axis shows the average predicted SWB outcome. A flat line indicates minimal impact of temperature on SWB. The

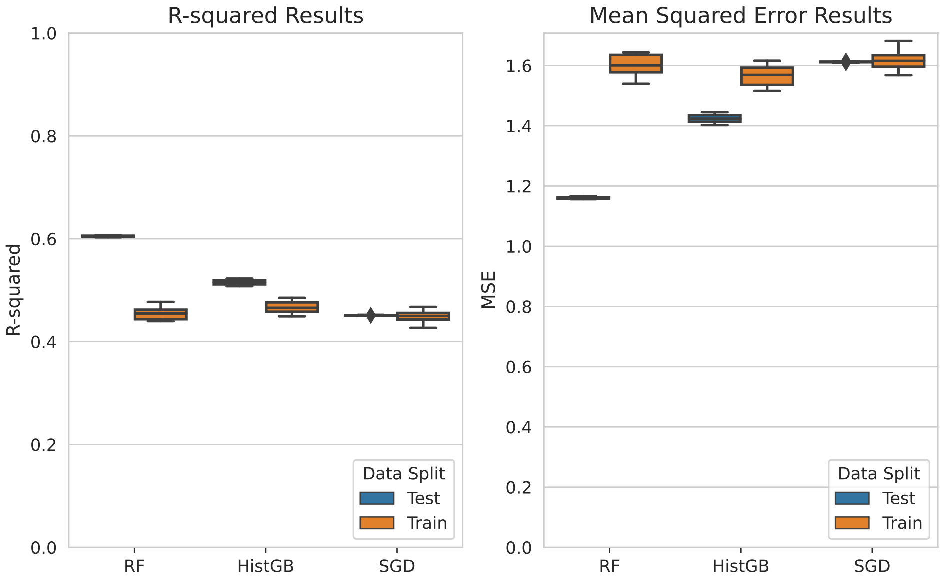

Figure 4.4: Model Performance

95% confidence interval, represented by whiskers, confirms the robustness of these results, revealing minimal variation upon repeated testing.

4.5 Simplification

One of the prevailing challenges in our analysis has been the models' diminished capacity to generalize, seemingly attributed to an undue emphasis on certain variables. To rectify this, one might consider a transformation of the data and the creation of summary statistics through dimensionality reduction. Dimensionality reduction aims to represent voluminous data more compactly, without sacrificing its intrinsic variability or complexity.

Principal Component Analysis (PCA) is a popular choice for such endeavors. PCA seeks to determine a set of new orthogonal axes, termed as principal components, such that most of the variance in the data can be captured by the first few components. It's akin to compressing data, but with the pivotal goal of maximizing the retention of its inherent information.

Yet, PCA is not without its challenges:

- **Assumption of Linearity:** PCA operates under the assumption that the data structure is linear. This means it expects the data points to lie in a hyperplane or a linear manifold. If the underlying data structure is nonlinear, PCA may not capture the main features efficiently.

- **Sensitivity to Outliers**: PCA is notably sensitive to outliers. Anomalous data points can significantly skew the primary components de-

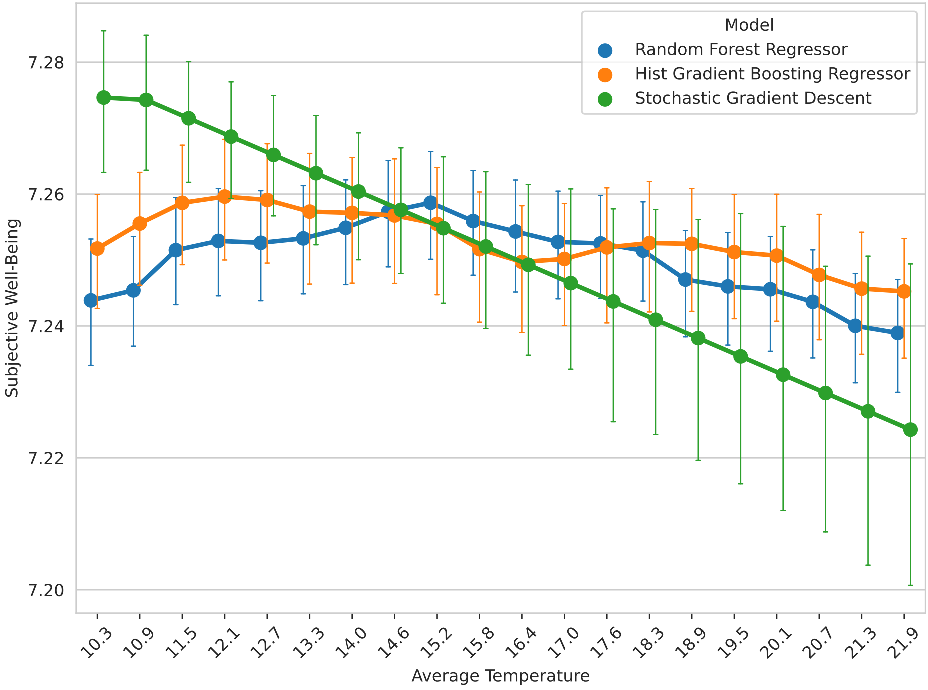

Figure 4.5: Partial Dependence of Subjective Well-Being on Temperature Across Distribution Quantiles

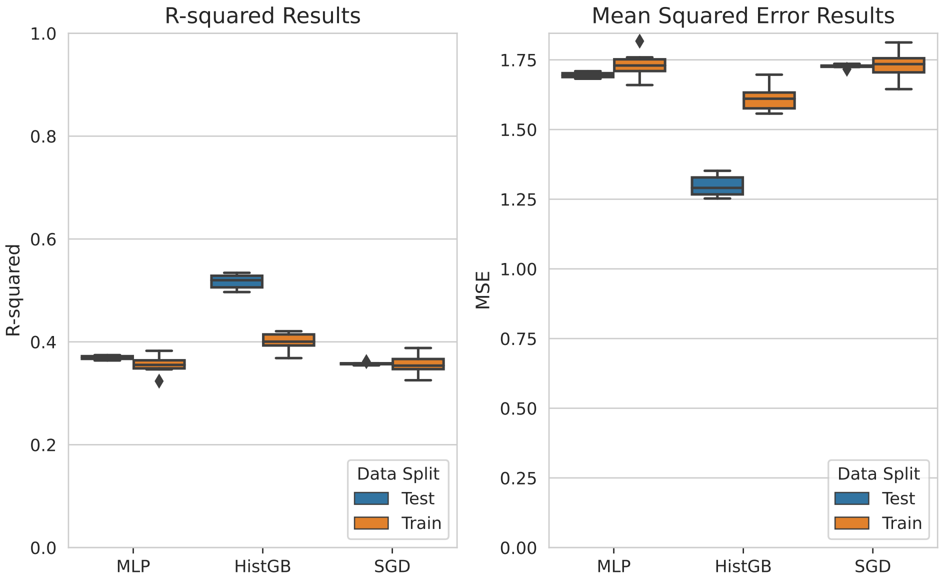

Figure 4.6: Model Performance

rived, leading to potentially misleading representations of data.

- **Variance vs. Importance**: PCA prioritizes variables with high variance, which may not always align with the most 'important' or 'meaningful' variables from a domain-specific perspective.

- **Loss of Interpretability**: As PCA generates synthetic variables (or principal components) by blending original variables, the interpretability often diminishes. These components don't have a direct real-world connotation, making it challenging to infer their significance intuitively.

One pivotal advantage of PCA is its flexibility in dimension specification. Practitioners can decide the number of dimensions (or components) they want to retain post-reduction. This is commonly done based on the cumulative variance captured by the components, retaining as many components as needed to explain, for example, 95% of the original variance.

To address the interpretability challenge posed by PCA, we took an additional step. Before the PCA transformation, variables were categorized into distinct socio-economically relevant categories, as shown in table A.1.

Our analytical approach, therefore, consists of initial data preprocessing (as discussed in previous sections), followed by the categorization of variables into their socio-economic classes. Subsequently, the PCA transformation is applied. The PCA was applied by using 4 principle components of each class which retained more than 90% of the variance from each class.

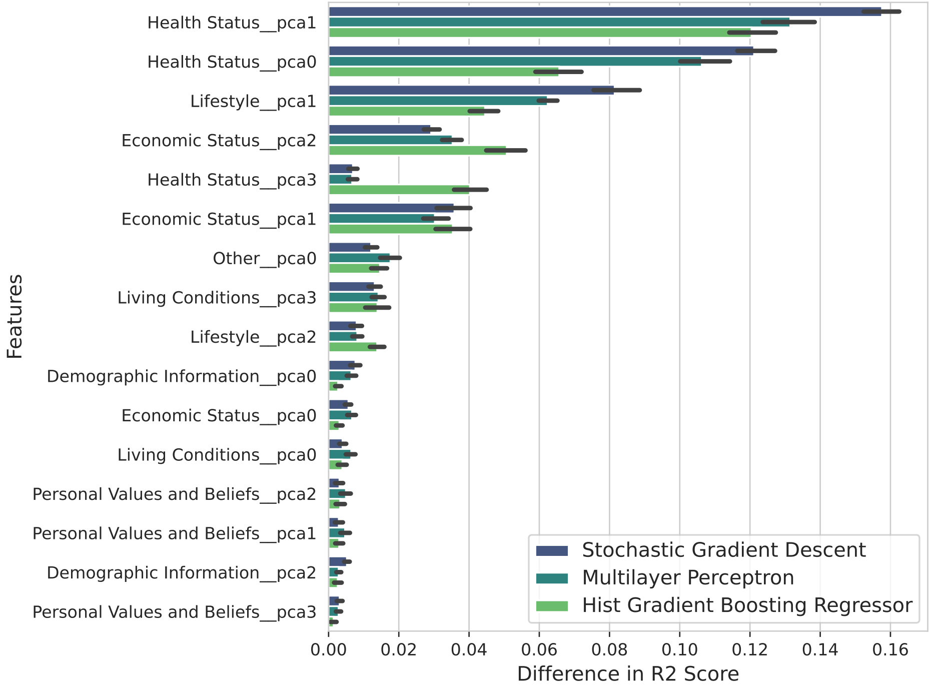

Figure 4.7: Permutation Importance

Again, the subsequent analysis is identical to the ones already performed in the prior two sections. The best performing model scores are displayed in Figure 4.6. It is clearly visible that the accuracy of the models has suffered somewhat with a loss of an R^2 score of around 0.02. This is probably due to the issue described above as *Variance vs. Importance*. There is also still some overfitting present in the hist gradient booster. In this analysis, however, the nerual network seems to perform quite well.

The feature importance was also performed equally as in section 4.2 and the results are displayed in figure 4.6. Similar to the permutation importance discussed in section 4.2 the most important denominators of well-being are health related variables reaching the highest permutation importance scores.

In this follow-up analysis, partial dependence was again assessed on the trained models, as in the previous section. The results are depicted in Figure 4.8. The plot averages predictions over different temperature quantiles. A slight positive effect is evident, particularly in the Neural Network and Stochastic Gradient Descent models. However, these results are less reliable than those of the earlier analysis, as indicated by the wider confidence intervals represented by the whiskers. The presence of such inconsistencies between the two analyses, especially given the minimal magnitude of the effect, suggests that the observed differences could be attributable to noise or

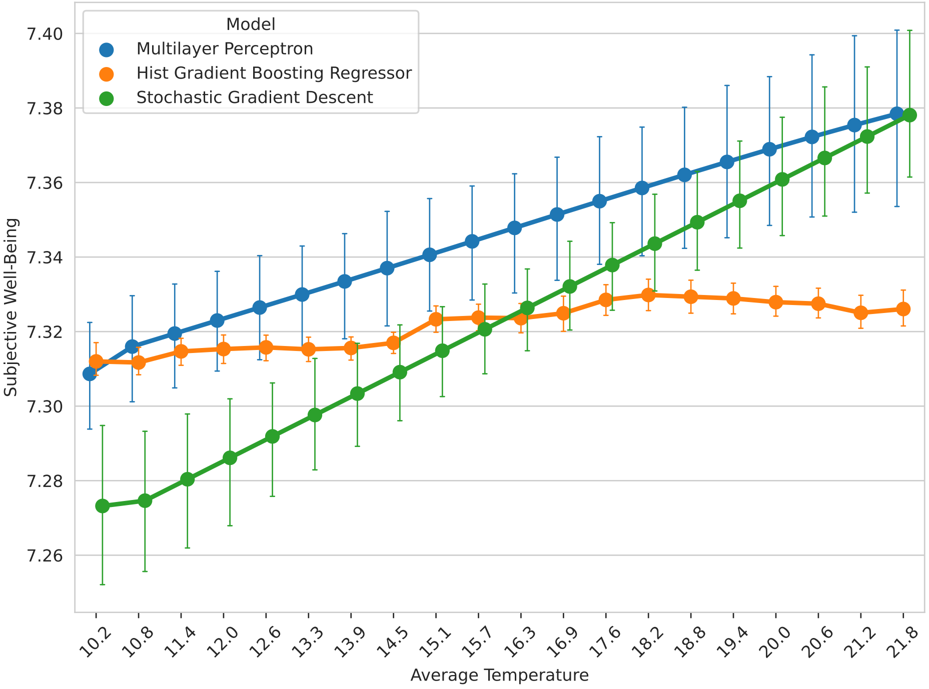

Figure 4.8: Partial Dependence Plot

other confounding factors. While these findings may initially appear to contradict those in Figure 4.5, the small effect size actually supports the claim that the influence of temperature on well-being is not clearly observable.

Chapter 5

Conclusions

5.1 Research Findings

The research integrated a dual analysis that showed among others associations such as health[19], income[14] and dwelling[25] as in existing studies, as noted in section 4.3. As shown in figures 4.2 and 1, variables of significant importance in predicting well-being in Germany are mainly health and economic related. This wording is chosen carefully to avoid assuming a direct cause-and-effect relationship. Higher economic status may not directly lead to better well-being but may enable it through what it can afford.

In terms of the relationship between higher temperatures and well-being, the evidence in this dataset is limited but robust, as evidenced in figures 4.5 and 4.8, where both results suggest a almost non-existent relationship between SWB and temperature. Although the impact of temperature on well-being is less pronounced, this does not mean the influence is non-existent. The minor observed impact might be due to Germany's current climate, or certain demographic factors not captured in the data. Thorough conclusions would need diverse datasets, an endeavor that exceeds this thesis and should be pursued in future research.

5.2 Future Prospects

The suggested future research path by this study can be outlined as follows:

- **Analyzing Data at a Finer Level:** Leading to higher variability in weather conditions and requiring secure data centers to protect individual data.

- **Incorporating Time-Series Analysis:** This might unearth vital information on how well-being changes over time. Investigating if temperature changes align with individual well-being could reveal stronger connections.

- **Examination of Specific Demographic Groups:** An essential aspect of future research should focus on specific demographic groups that might be more sensitive to temperature variations, contributing to a more comprehensive understanding of the relationship between temperature and well-being.

- **Potential Reduction in Personal Bias and Focus on Intra-Individual Changes:** Focusing on changes within individuals, rather than their absolute values, could yield more valuable and objective insights. Studying how well-being shifts in relation to weather, income, and health may reveal the true factors influencing these changes. This focus offers a more nuanced understanding of well-being and has broader implications.

In conclusion, this thesis serves as a robust foundation for future research to explore the complexities of well-being. It paves the way for more sophisticated, longitudinal investigations that can enhance our understanding and handling of both individual and societal well-being.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[3] Hanfried H. Andersen, Axel Mühlbacher, Matthias Nübling, Jürgen Schupp, and Gert G. Wagner. Computation of standard values for physical and mental health scale scores using the SOEP version of SF-12v2. *Journal of Contextual Economics – Schmollers Jahrbuch*, 127(1):171–182, jan 2007.

[4] Jan Michael Bauer, Victoria Levin, Ana Maria Munoz Boudet, Peng Nie, and Alfonso Sousa-Poza. Subjective well-being across the lifespan in europe and central asia. *Journal of Population Ageing*, 10(2):125–158, may 2016.

[5] David G. Blanchflower and Andrew J. Oswald. Is well-being u-shaped over the life cycle? *Social Science - Medicine*, 66(8):1733–1749, apr 2008.

[6] David Dorn, Justina A.V. Fischer, Gebhard Kirchgässner, and Alfonso Sousa-Poza. Is it culture or democracy? the impact of democracy and culture on happiness. *Social Indicators Research*, 82(3):505–526, sep 2006.

[7] Andreas Eberl, Matthias Collischon, and Tobias Wolbring. Subjective well-being scarring through unemployment: New evidence from a long-running panel. *Social Forces*, 101(3):1485–1518, mar 2022.

[8] Gallup. Climate Change and Wellbeing Around the World. 2022.

[9] Jan Goebel, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp. The german socio-economic panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239(2):345–360, aug 2018.

[10] Wencke Gwozdz and Alfonso Sousa-Poza. Ageing, health and life satisfaction of the oldest old: An analysis for germany. *Social Indicators Research*, 97(3):397–417, aug 2009.

[11] Marilyn Haring-Hidore, William A. Stock, Morris A. Okun, and Robert A. Witter. Marital status and subjective well-being: A research synthesis. *Journal of Marriage and the Family*, 47(4):947, nov 1985.

[12] Julianne Holt-Lunstad, Timothy B. Smith, and J. Bradley Layton. Social relationships and mortality risk: A meta-analytic review. *PLoS Medicine*, 7(7):e1000316, jul 2010.

[13] Junyi Hua, Yuan Shi, Chao Ren, Kevin Ka-Lun Lau, and Edward Yan Yung Ng. Impact of urban overheating and heat-related mortality in hong kong. In *Advances in Sustainability Science and Technology*, pages 275–292. Springer Nature Singapore, 2022.

[14] Daniel Kahneman and Angus Deaton. High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences*, 107(38):16489–16493, sep 2010.

[15] Lauren Kuykendall, Louis Tay, and Vincent Ng. Leisure engagement and subjective well-being: A meta-analysis. *Psychological Bulletin*, 141(2):364–403, mar 2015.

[16] Christian Sebastian Lamprecht. Meteostat python, 2022.

[17] S. Katherine Nelson, Kostadin Kushlev, and Sonja Lyubomirsky. The pains and pleasures of parenting: When, why, and how is parenthood associated with more or less well-being? *Psychological Bulletin*, 140(3):846–895, may 2014.

[18] Nick Obradovich, Robyn Migliorini, Sara C. Mednick, and James H. Fowler. Nighttime temperature and human sleep loss in a changing climate. *Science Advances*, 3(5), may 2017.

[19] Morris A. Okun and Linda K. George. Physician- and self-ratings of health, neuroticism and subjective well-being among men and women. *Personality and Individual Differences*, 5(5):533–539, jan 1984.

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

[21] O Seppanen, W J Fisk, and Q H Lei. Room temperature and productivity in office work. 7 2006.

[22] Toki Sherbakov, Brian Malig, Kristen Guirguis, Alexander Gershunov, and Rupa Basu. Ambient temperature and added heat wave effects on hospitalizations in california from 1999 to 2009. *Environmental Research*, 160:83–90, jan 2018.

[23] Beatriz Fabiola López Ulloa, Valerie Møller, and Alfonso Sousa-Poza. How does subjective well-being evolve with age? a literature review. *Journal of Population Ageing*, 6(3):227–246, apr 2013.

[24] JOHN E. WARE, MARK KOSINSKI, and SUSAN D. KELLER. A 12-item short-form health survey. *Medical Care*, 34(3):220–233, mar 1996.

[25] Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang, Li Chen, and Yunxiao Dang. The impact of housing pressure on subjective well-being in urban china. *Habitat International*, 127:102639, sep 2022.

Appendix A

Appendix

A.1 Methods Definitions

A.1.1 Decision Trees

Below is given a concise mathematical formalization of decision trees. First, define the variance for a set of samples, $S \subseteq X$, at a given node:

$$\text{Variance}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y_S})^2 \tag{A.1}$$

where S represents the samples at the current node, y_i denotes the target value of the i-th sample, and $\bar{y_S}$ signifies the average target value of the samples in S.

Considering a split of S into two partitions, S_{left} and S_{right}, the associated cost function becomes:

$$C(S, S_{left}, S_{right}) = \frac{|S_{left}|}{|S|}\text{Variance}(S_{left}) + \frac{|S_{right}|}{|S|}\text{Variance}(S_{right}) \tag{A.2}$$

The recursive objective function for the decision tree regressor can then be written as minimizeing the cost of every partition:

$$\begin{aligned} S_{left}, S_{right} &= \underset{S_l, S_r}{\text{argmin}} \ C(S, S_l, S_r) \\ J(S) &= J(S_{left}) + J(S_{right}) \end{aligned} \tag{A.3}$$

Keep in mind that S_{left} and S_{right} is a partition of S.

A.1.2 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) extends the concepts from linear regression (Section 3.2.2) into a nonlinear context through the process of forward propagation. It consists of multiple layers of nodes, where each layer transforms its inputs into a more abstract representation.

For each hidden layer l, the transformation is:

$$h_{l,j} = \sigma\left(\sum_{i=1}^{m_{l-1}} w_{l,j,i} \cdot h_{l-1,i} + b_{l,j}\right) \tag{A.4}$$

where $h_{l,j}$ is the output of the j-th node in the l-th layer, $w_{l,j,i}$ and $b_{l,j}$ are the weights and bias, and σ is an activation function.

Forward propagation refers to the process of computing the output by applying the transformations through each layer sequentially from the input to the output layer.

The output layer in regression is:

$$\hat{y} = \sum_{i=1}^{m_L} w_{L+1,1,i} \cdot h_{L,i} + b_{L+1,1} \tag{A.5}$$

The loss function is:

$$\text{Loss}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{A.6}$$

The objective function includes optional regularization techniques (L1 and L2) applied to each layer's weights:

$$J(\theta) = \text{Loss}(y, \hat{y}) + \sum_{l=1}^{L}\left(\lambda_1 \cdot ||w_l||_1 + \lambda_2 \cdot ||w_l||_2^2\right) \tag{A.7}$$

where w_l represents the weights at layer l, and λ_1 and λ_2 control the strengths of the L1 and L2 penalties, respectively.

Note A.1.1. In addition to L1 and L2 regularizations, other techniques such as dropout may also be used to prevent overfitting, although they are not elaborated here. The choice of regularization and other hyperparameters depends on the specific problem and dataset, as outlined in section 3.2.2. Activation functions add an additional layer of complexity, allowing for the modeling of non-linearities, a feature not present in traditional linear regression.

A.2 Tables

A.2.1 Variables

Table A.1: Summary of Variables

variable	variable_label	type	category
pid	Unveraenderliche Personennummer	Nominal	key
hid	Aktuelle Haushaltsnummer	Nominal	key
cid	Case ID, Ursprungshaushaltsnummer	Nominal	key
pnr	Lfd. Personennummer	Nominal	key
syear	Erhebungsjahr (SurveyYear)	Discrete	Other
pla0009_v2	Geschlecht [1984-2019]	Nominal	Demographic Information
plb0018	Bezahlte Arbeit letzte 7 Tage	Nominal	Economic Status
plb0019_v2	Mutterschutz/Elternzeit [2001-2020]	Nominal	Demographic Information
plb0021	Arbeitslos gemeldet	Nominal	Economic Status
plb0022_h	Erwerbsstatus [harmonisiert]	Nominal	Economic Status
plb0024_v3	Laenger als 6 Wochen krank gemeldet [1999-2020]	Nominal	Health Status
plb0282_h	Seit Anfang Vorjahr aus Beruf ausgeschieden [ha...	Nominal	Economic Status
plb0282_v2	Seit Anfang Vorjahr aus Beruf ausgeschieden [20...	Nominal	Economic Status
pld0131_v1	Familienstand [1984-2018]	Nominal	Demographic Information
ple0008	Gesundheitszustand gegenwaertig	Ordinal	Health Status
ple0010_h	Geburtsjahr [harmonisiert]	Discrete	Demographic Information
ple0010_v2	Geburtsjahr (Viersteller) [1984, 1999-2020]	Discrete	Demographic Information
ple0040	Erwerbs-, Schwerbehinderung	Nominal	Health Status
ple0053	Krankenhausaufenthalt Vorjahr	Nominal	Health Status
ple0097	Art der Krankenversicherung	Nominal	Health Status
ple0160	Kassenwechsel in Vorjahr	Nominal	Health Status
plg0012_v1	Derzeit in Ausbildung [1984-2020]	Nominal	Lifestyle
plg0072	Seit Vorjahr Ausbildung abgeschlossen	Nominal	Other
plh0007	Interesse fuer Politik	Ordinal	Personal Values and Beliefs
plh0011_h	Allgemeine Parteienpraeferenz [harmonisiert]	Nominal	Personal Values and Beliefs
plh0011_v2	Allgemeine Parteienpraeferenz [1984-2020]	Nominal	Personal Values and Beliefs
plh0032	Sorgen allgemeine wirtschaftliche Entwicklung	Ordinal	Personal Values and Beliefs
plh0033	Sorgen eigene wirtschaftliche Situation	Ordinal	Economic Status
plh0035	Sorgen eigene Gesundheit	Ordinal	Health Status
plh0036	Sorgen Umweltschutz	Ordinal	Personal Values and Beliefs
plh0038	Sorgen Friedenserhaltung	Ordinal	Personal Values and Beliefs
plh0040	Sorgen Kriminalitaetsentwicklung in Deutschland	Ordinal	Personal Values and Beliefs
plh0171	Zufriedenheit Gesundheit	Ordinal	Health Status
plh0175	Zufriedenheit HH-Einkommen	Ordinal	Economic Status
plh0177	Zufriedenheit Wohnung	Ordinal	Living Conditions
plh0178	Zufriedenheit Freizeit	Ordinal	Lifestyle
plh0182	Lebenszufriedenheit gegenwaertig	Ordinal	Target
pli0038_h	Beruf, Lehre, Nebenerw. Std., Werktg. [harmonis...	Continuous	Economic Status
pli0038_v4	Beruf/Lehre/Nebenerwerb Std./Werktag [1992-2020]	Continuous	Economic Status
pli0040	Besorgungen Std., Werktg.	Continuous	Economic Status
pli0043_h	Hausarbeit Std., Werktg. [harmonisiert]	Continuous	Lifestyle
pli0043_v3	Hausarbeit Std./Werktag [1992-2020]	Continuous	Lifestyle
pli0044_h	Kinderbetreuung Std., Werktg. [harmonisiert]	Continuous	Lifestyle
pli0044_v3	Kinderbetreuung, Mo.-Fr., Stunden [1992-2020]	Continuous	Lifestyle
pli0046	Versorgung Pflegebeduerftiger, Werktg.	Continuous	Lifestyle
pli0047_v1	Aus- u. Weiterb., Lernen Std., Werktg. (erwerbs...	Continuous	Lifestyle
pli0049_h	Reparaturen etc. Std., Werktag [harmonisiert]	Continuous	Lifestyle
pli0049_v3	Reparaturen/Gartenarbeit Std./Werktag [1992-2020]	Continuous	Lifestyle
pli0051	Hobbies, Freizeit Std., Werktg.	Continuous	Lifestyle

variable	variable_label	type	category
plj0014_v3	Deutsche Staatsangehoerigkeit [1996-2020]	Nominal	Demographic Information
plj0022	2. Staatsangehoerigkeit vorhanden	Nominal	Demographic Information
plj0046	Sorgen Zuwanderung	Ordinal	Personal Values and Beliefs
plj0047	Sorgen Auslaenderfeindlichkeit	Ordinal	Personal Values and Beliefs
plj0151	Keine Zahlung	Nominal	Other
pinta_v2	Befragungsform [1985-2020]	Nominal	Other
pmonin	Monat des Interviews	Nominal	Other
ptagin	Tag des Interviews	Nominal	Other
hlc0005_h	Monatliches HH-Netto-Einkommen [harmonisiert]	Continuous	Economic Status
hlc0005_v2	Monatliches HH-Netto-Einkommen (Euro) [2002-2020]	Continuous	Economic Status
hlc0007	Miet- u.Pachteinnahmen Vorjahr	Nominal	Economic Status
hlc0039_h	Kindergeldbezug letztes Jahr [harmonisiert]	Nominal	Economic Status
hlc0039_v3	Kindergeldbezug letztes Jahr [1996-2020]	Nominal	Economic Status
hlc0044_h	Kindergeldbezug heute [harmonisiert]	Nominal	Economic Status
hlc0055_h	Hilfe Lebensunterhalt Vorjahr [harmonisiert]	Nominal	Economic Status
hlc0067_h	Hilfe Lebensunterhalt heute [harmonisiert]	Nominal	Economic Status
hlc0077	Leistungen der Pflegeversicherung Vorjahr	Nominal	Health Status
hlc0080_h	Wohngeld,Lastenzuschuss Vorjahr [harmonisiert]	Nominal	Economic Status
hlc0080_v1	Wohngeld 2016 [1984, 1991-2020]	Nominal	Economic Status
hlc0083_h	Wohngeld heute [harmonisiert]	Nominal	Economic Status
hlc0085_h	Pflegevers. Leistungen [harmonisiert]	Nominal	Economic Status
hlc0113_h	Abzahlung Kredite [harmonisiert]	Nominal	Economic Status
hlc0119_h	Sparbetrag monatlich [harmonisiert]	Nominal	Economic Status
hlf0001_h	Haupt-, Untermieter, Eigentuemer [harmonisiert]	Nominal	Living Conditions
hlf0001_v3	Miete oder Eigentum (auch Altersheim) [1999-2020]	Nominal	Living Conditions
hlf0006	Eigentuemerwechsel Vorjahr	Nominal	Living Conditions
hlf0018	Groesse der Wohnung veraendert	Nominal	Living Conditions
hlf0019_h	Wohnflaeche insgesamt in qm [harmonisiert]	Continuous	Living Conditions
hlf0019_v1	Qm Wohnflaeche [1984, 1998-2020]	Continuous	Living Conditions
hlf0021_h	Anzahl der Wohnraeume [harmonisiert]	Continuous	Living Conditions
hlf0021_v1	Anzahl der Wohnraeume [1984-1990, 1998-2020]	Continuous	Living Conditions
hlf0071_h	Beurteilung der Wohnungsgroesse [harmonisiert]	Ordinal	Living Conditions
hlf0071_v1	Beurteilung der Wohnungsgroesse [1984, 1998-2020]	Ordinal	Living Conditions
hlf0261	Putz-,Haushaltshilfe beschaeftigt	Nominal	Living Conditions
hlf0291	Hilfe-,Pflegebeduerft. Person im HH	Nominal	Living Conditions
hlk0044_v1	Kinder im HH, in oder nach 2004 geboren [1984-2...	Nominal	Demographic Information
hlk0056	Durchfuehrung der Befragung	Nominal	Other
hlk0059	Tag des Interviews	Nominal	Other
hlk0060	Monat des Interviews	Nominal	Other
valid	Vollstaendigkeit der Generierung des SOEPvSF12	Nominal	Health Status
mcs	MCS: Summary scale Mental (NBS)	Continuous	Health Status
pcs	PCS: Summary scale Physical (NBS)	Continuous	Health Status
pf_nbs	Physical functioning (NBS)	Continuous	Health Status
rp_nbs	Role physical (NBS)	Continuous	Health Status
bp_nbs	Bodily pain (NBS)	Continuous	Health Status
gh_nbs	General health (NBS)	Continuous	Health Status
vt_nbs	Vitality (NBS)	Continuous	Health Status
sf_nbs	Social functioning (NBS)	Continuous	Health Status
re_nbs	Role emotional (NBS)	Continuous	Health Status
mh_nbs	Mental health (NBS)	Continuous	Health Status
bmi	Body-Mass-Index	Continuous	Health Status
height	Body Height in cm	Continuous	Health Status
fheight	Imputation Flag for Height$$	Nominal	Other
weight	Weight in kg	Continuous	Health Status
fweight	Imputation Flag for Weight$$	Nominal	Other
bula	Bundesland nach AGS (1-2 Stelle)	Nominal	Demographic Information
nuts1	NUTS-Systematik Level 1 (Bundesland)	Nominal	key

variable	variable_label	type	category
tavg	Temperature	Continuous	Other
prcp	Precipitation	Continuous	Other
wspd	Wind Speed	Continuous	Other
pres	Level Air Pressure	Continuous	Other
tsun	Sunshine Duration	Continuous	Other
age	Age	Continuous	Demographic Information
intid	Interviewer ID	Nominal	Other
time	Time of Interview	Nominal	key
h_pnr	Houshold Interviewee	Nominal	Other