Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Identifying Predictors for Energy Poverty in Europe Using Machine Learning

Willem Eugène van Hove

Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Identifying Predictors for Energy Poverty in Europe Using Machine Learning

Author:	Willem Eugène van Hove
1st examiner:	Prof. Hans-Joachim Bungartz
2nd examiner:	Prof. Bob van der Zwaan
Assistant advisor:	Dr. Francesco Dalla Longa
Submission Date:	October 21, 2020

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

October 21, 2020 Willem Eugène van Hove

Acknowledgments

I would like to thank Bob van der Zwaan and Francesco Dalla Longa for their supervision, support, and guidance during this research. Their expertise helped me forward in my Master's thesis. I would also like to thank TNO for providing me with the internship and Professor Hans-Joachim Bungartz for his supervision and feedback.

"If a machine is expected to be infallible, it cannot also be intelligent."

-Alan Turing

Abstract

This thesis attempts to identify drivers for energy poverty in Europe using machine learning. The establishment of predictors for energy poverty valid across countries is a call made by many researchers active in the field of energy poverty. A previously defined framework to classify households as being at risk of energy poverty, based on income and energy expenditure, is applied to a data set from a survey conducted at the household level in 11 European countries with vastly different climates, cultures, and economies. A gradient boosting classifier to predict energy poverty risk is successfully trained on a set of socio-economic features hypothesized as predictors for energy poverty in a diverse set of countries in Europe. The classifier's internal model is analyzed, providing novel insights into the intricacies that underlie energy poverty. We find that besides the main driver - income - floor area and household size are confirmed as predictors. These features significantly assist the model in classifying a household as being at risk. The results suggest that house age and respondent age can be discarded as predictors. With regards to heating strategy and house detachment as predictors, the outcomes are inconclusive. We argue that this is possibly due to a severely limited data set in terms of both quality and quantity. In order to build a model that has the potential to capture the full complexity of the mechanisms that govern energy poverty, consistent high-quality data sets are needed. Currently, these were not found to be available for most European countries. To facilitate more advanced research into energy poverty in Europe, we recommend to increase household data collection efforts, both at the country- and EU-level. The establishment of energy poverty predictors valid across Europe could provide a basis to effectively target energy-poor households with adequate policy measures.

The thesis is structured in three parts. Part I is intended as a self-contained piece that is accessible for the general reader. Here the main aspects of the research, and the results are discussed. Part II provides a comprehensive description of the data, preprocessing steps, and the framework used to label the data. In Part III, a more technical, in-depth explanation of machine learning techniques and methods to interpret the resulting models, is presented.

Contents

Part I.

Main Analysis

1. Introduction

Traditionally, energy poverty is defined as the lack of access to modern energy services and is primarily studied in developing countries where access is not assured for a substantial share of the population. The transition to a more sustainable energy supply is expected to radically transform energy infrastructure, both in developing and developed countries. The cost of this transformation needs to be distributed in such a way that all households continue to be able to afford their energy bills. This has led to the expansion of the scope of energy poverty research to also include developed countries that are on the forefront of the energy transition. In those countries, energy poverty is defined as the inability of a household to afford its energy bills.

Humankind has always evolved around acquiring sufficient energy [1]; from the discovery and subsequent control of fire to keep warm and provide light, to later inventions, such as electric heating and LED lighting. The concept of energy poverty has been studied since the oil crises in the 1970s. In the United Kingdom, the term fuel poverty was coined; a person affected by fuel poverty was defined as "a person [who] is a member of a household living on a lower income in a home which cannot be kept warm at reasonable cost" [2]. In this research, we consider energy poverty a similar, but broader concept than fuel poverty. The field of energy poverty has become truly widespread on the European continent since 2008, as in that year EU institutions and consultative committees began calling for a Europe-wide definition of energy poverty [3]. In 2013, the European Economic and Social Committee called for "European energy poverty indicators to be established and for statistics to be harmonised in order to identify, prevent and tackle the problem more effectively at the European level and to generate solidarity in this area" [4]. In 2015, Sovacool studied the Warm Front (WF) scheme, a program in England that ran between 2000 and 2013 with the aim of combating fuel poverty. He concluded that the WF "had difficulty identifying fuel poor homes" [5]. Although fuel poverty is a field of active research, it is still poorly understood [6].

A large-scale study in 2019 found that an indicator for energy poverty valid across countries is still absent. This has resulted in a lack of research comparing countries, since no metric exists that can be used to contrast them [7]. Concurrently, Castaño-Rosa et al. observed a lack of standards to assess energy poverty across Europe and, instead, argued for a multiple indicator approach as starting point for policy decisions to reduce energy poverty in Europe [8]. A recent study assumed that constructing all-encompassing predictors to assess energy poverty is unfeasible [9]. These three papers have in common that they illustrate the call from researchers active in this field for the establishment of predictors for energy poverty across countries in order to enable cross-country comparisons and assist European legislation.

Since the 1950s, the field of Artificial Intelligence (AI) has dedicated efforts to en-

abling machines to act intelligently; this is in contrast to animals and humans that possess natural intelligence [10]. Recently, a sub field of AI, machine learning (ML), has shown incredible growth and has gained prominence in academic publications and news headlines. In ML, the approach starts with a generic model with a large number of parameters and a wide range of potential applications, depending on how these parameters are set. The goal is to make a machine "learn", through experience, how these parameters can be set in an optimal manner for the purposes at hand. The experience is passed to the model in the form of data [11]. ML models have demonstrated a potential to reach exceptional performance on diverse problems. Typically, only the ML model's input and output are visible, and the increasingly convoluted inner workings of the model are unknown. These unknown processes in ML models are referred to as the *black box* [12].

ML models play a role in many aspects of our lives: from talking to a virtual assistant on your smartphone to receiving a credit score from a bank. Simultaneously, the academic debate on our understanding of the black box approach is flourishing. Since 2018, EU legislation ensures that consumers have a basic right to an explanation of how an algorithm produced an output [13]. While this is a strong incentive for companies to understand their models, the extent of this regulation is heavily debated [14]. The field concerned with opening the black box is "interpretable or explainable artificial intelligence" (XAI). This is achieved with various different methods designed to provide users an easier, more comprehensible explanation of the output [15]. If successful, ML models are able to find complex statistical relations in data that would require excessive amounts of manual labor using statistical tests or running and evaluating standard regressions. As a result, XAI could reduce the time needed for researchers to understand the sophisticated systems they are working with. This was recently demonstrated, for example, when XAI enabled researchers to identify crucial predictive biomarkers of disease mortality briefly after the outbreak of the covid-19 pandemic [16].

ML has successfully been used to identify energy poverty predictors in the Netherlands [17]. The present research aims to use a similar approach to investigate energy poverty predictors within a larger geographic area, encompassing several countries in Europe. Finding predictors valid across Europe could aid the assessment of the prevalence of energy poverty in European countries, and subsequently assist adequate policy design. We attempt to identify pan-European predictors through the use of XAI, a method that has yet to be applied in the field of energy poverty. In Section 2 the ML technique of gradient boosting is described, as well as the energy poverty framework used to categorize households. In Section 3 the results of applying the classification framework to the data set are analyzed, and the predictors found with a gradient boosting model are discussed. In Section 4 the findings are interpreted and recommendations regarding the collection and accessibility of data are given, in order to better address energy poverty in Europe, and provide guidance for future research endeavors in this field.

2. Methodology

For this research, a survey on energy use in Europe conducted in 2018 by Enable-EU was used [18]. Enable-EU is an ongoing endeavor funded by the European Union's Horizon 2020 research and innovation program with the mission statement: "[Enable-EU] seeks to understand what determines people's choices in three key consumption areas: transportation, heating & cooling, and electricity" [19]. The survey was targeted at a group of 11 diverse countries in Europe: Bulgaria (BG), France (FR), Germany (DE), Hungary (HU), Italy (IT), Norway (NO), Poland (PL), Serbia (RS), Spain (ES), Ukraine (UA), and the United Kingdom (UK). A report outlining and comparing the outcomes of participating countries was published along with the data set [20]. While some questions, for example on prosumers, were country-specific, all respondents were asked to complete those sections of the survey with generic and socioeconomic questions, which provides us with a complete data set on these topics. Most notably, participants were asked to report their income and energy expenditure, two variables crucial in the labelling of data points. Thus making the data set eligible for (so-called supervised) machine learning. Most questions were multiple-choice, i.e. one answer was to be selected from a list of possible options (categories). The diversity in the assessed countries makes this data set particularly interesting for the purpose of investigating energy poverty predictors at the European level.

The energy poverty classification framework proposed by Dalla Longa et al. is used to categorize each household in our data set into one of four energy poverty risk categories [17]. The framework operates on an income-energy expenditure grid that is divided into four quadrants using two thresholds, one for each axis. This is illustrated in Figure 2.1. A household with income above and energy expenditure below the respective thresholds is categorized as "No risk" (green). If one of the variables crosses a threshold, the label "Income risk" (yellow) or "Expenditure risk" (orange) is assigned. If both thresholds are crossed, the household is categorized as being at highest risk of energy poverty, labelled "Double risk" (red).

In order to apply this classification framework to the Enable-EU data, the energy expenditure first had to be derived using two questions in which respondents were asked to report their latest heating and their latest electricity costs. During preprocessing, all currencies were converted to Euros and all costs to annual costs, after which the yearly energy expenditures were defined for all households. Combined with a question on income, all households in the data could now be labelled as one of the risk categories.

The Enable-EU survey required respondents to classify their income into the corresponding decile in their respective country, leaving the results to be categorized into 10 brackets. This enabled a normalized measure of income across the participating countries. The income threshold is set for all countries between income deciles three and

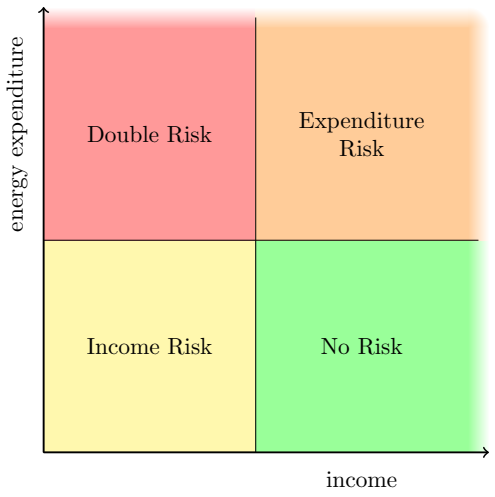

Figure 2.1.: Visual representation of the quadrants that the classification framework uses to assign households to a risk category.

four. This particular threshold choice was made for three main reasons. First, analysis showed that for most countries the respective minimum wage corresponds to bracket 3. Second, the Low Income High Costs (LIHC) indicator applied in the UK, results in the vast majority of households categorized as energy poor being in the first three income deciles [21]. Third, this threshold ensures that the risk classes contain enough data points to produce reliable classifiers [17]. The energy expenditure threshold is set at the 80th quantile of the absolute energy expenditure in the respective country, resulting in a different value for every country.

The thresholds used are statistically determined thresholds: the 30th quantile of income and the 80th quantile of energy expenditure. Consequently, every country has the same share of households that cross the thresholds and are thus at risk, irrespective of mean income or energy expenditure. This allows the framework to be applied to a heterogeneous set of countries such as ours. Due to the thresholds' statistical nature, for example, if energy becomes cheaper overall, the energy expenditure threshold moves down with it. The same holds for income. As a result, the only class that can be "manipulated" by policy or legislation is the intersection between these two groups: the double risk category. These are the people that earn least money but are still amongst the households with highest energy expenditure. Therefore, the size of this risk group can be used as an indication of the prevalence of energy poverty in a set of households. Policy can attempt to minimize energy poverty by aiming to move the households labelled as being in double risk, out of the red quadrant, towards the orange or the yellow quadrants.

Using ML, we try to identify predictors to classify households at risk of being in energy poverty. Gradient boosting is an ML technique that incrementally adds weak

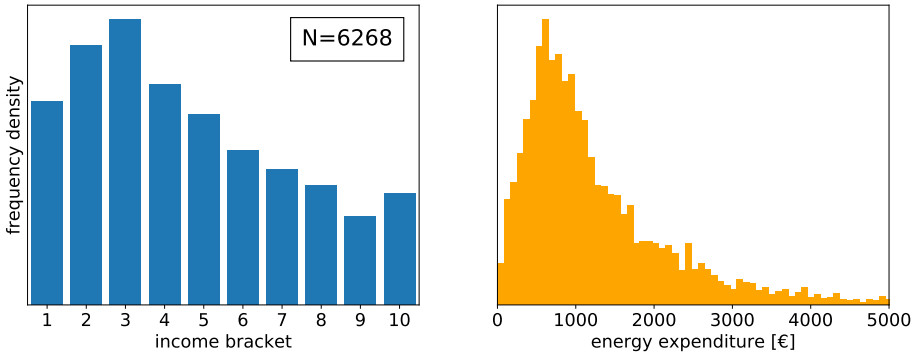

Figure 2.2.: Income distribution of the whole data set (left), and distribution of energy expenditure per household per year in the whole data set (right).

prediction models, in our case decision trees, to obtain a model with a better performance [22], the resulting model is known as an ensemble. Gradient boosting achieves state-of-the-art performance on many modelling tasks [23]. Especially on tabular data, decision tree methods perform best and provide a vast range of tools to analyze the internal model [15, 24], which we can exploit to gain insight into the complex origins of energy poverty in Europe. As the data set contains many categorical features, CatBoost was chosen as gradient boosting library. CatBoost is a relatively new library by the Russian tech company Yandex. As opposed to other popular gradient boosting libraries, it can deal with categorical features without any preprocessing steps required [23, 25].

We attempt to assess possible predictors by estimating the influence a certain feature has on the model outcome, i.e. the feature importance, according to different metrics. A feature's assigned importance can be an indicator of its predictive power. Therefore, if a feature is assigned high importance for a good model, that would suggest that it is a true predictor of energy poverty. The preprocessing steps and classification framework are further detailed in Part II. Gradient boosting, CatBoost, and the some of the methods used to obtain our results are explained in Part III.

The distribution over the income deciles in the entire data set can be found in Figure 2.2. Ideally, a representative sample of a population has the same number of respondents in each category. However, in the current data set a bias towards the lower-income brackets can be observed, this is known as *selection bias*. On the country level, this discrepancy is more prominent, as some income brackets are completely un- or severely underrepresented (Figure A). This skewness of the income distribution on the country level is not deemed a fundamental problem, as we are seeking predictors that are valid across Europe, and over the entire data set the selection bias is less pronounced.

Yearly energy expenditure varies per country: Ukraine has the lowest median per household of €[1]267, and Norway has the highest median of €2311. The majority of

[1] All euros reported in this research are in €(2018)

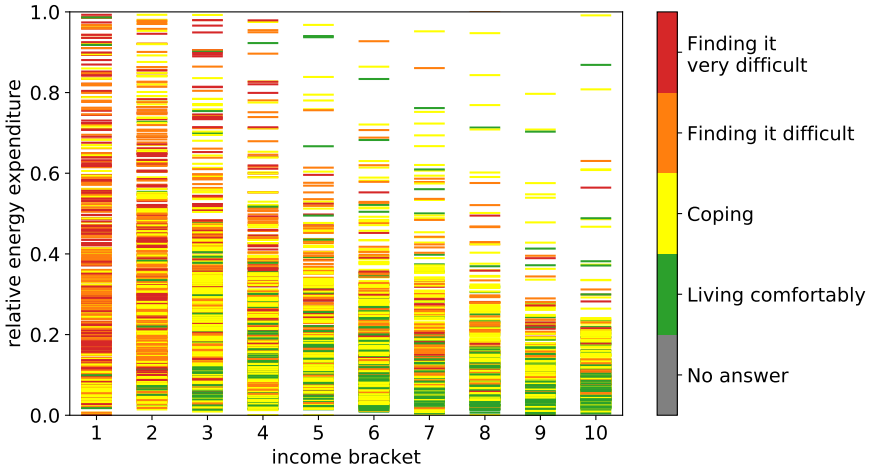

Figure 2.3.: Poverty as reported by all households in the data set.

households in the data set spend up to €2000 per year on energy, with those spending more than that predominantly being from Norway and, to a lesser extent, from other Western European countries. The median over the entire data set is €997 per year.

The employed classification framework is supported by Figure 2.3, which depicts the answers to a self-reported poverty question for the entire data set. Each household is a colored line; the color depicts the response to the question "which of the descriptions bellow [sic] comes closest to how you feel about your household's income nowadays?". All households are plotted on a grid with income bracket on the x-axis and approximated relative energy expenditure on the y-axis. The relative energy expenditure is approximated and used to provide a uniform measure of energy expenditure, thus allowing all households to be plotted on the same axis. The approximation is detailed in Section 5.2.

A gradient from red ("finding it very difficult on current household") in the top left, to green ("living comfortably on current income") in the bottom right can be identified. It shows similarity to the energy poverty classification framework depicted in Figure 2.1, applied to Ukrainian households in Figure 2.4. The resemblance corresponds to the close connection between energy poverty and poverty. These plots, with a wide range of households from different countries and socioeconomic backgrounds, together with the fact that the framework makes intuitive sense, allows us to confidently claim that the framework is robust in identifying energy poverty in a diverse set of countries.

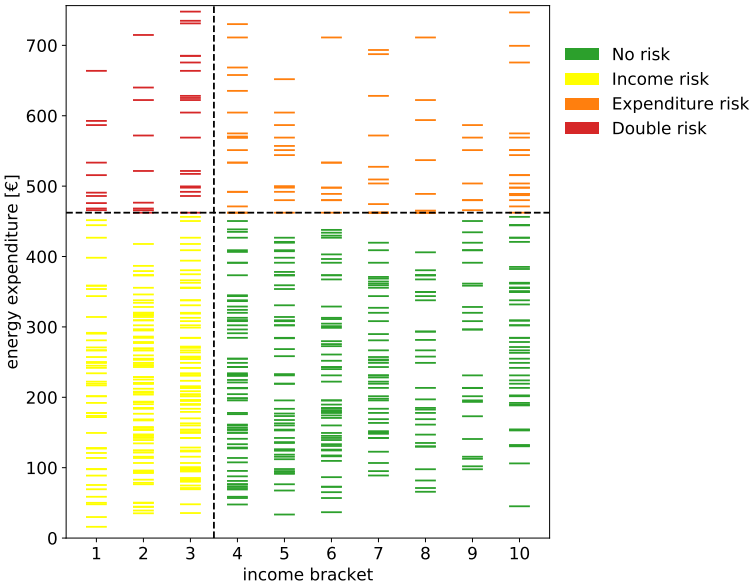

Figure 2.4.: Energy poverty classification for Ukraine.

3. Results

The overall findings of applying the framework to the data and the trained ML model are first presented. In the next section, results obtained by applying XAI to assist in interpreting the model are detailed.

3.1. Overall findings

The energy poverty class composition that resulted from applying the risk framework on the data set, for each of the 11 countries in the Enable-EU data set, is plotted in Figure 3.1. The countries are sorted by the share of households classified in the double risk category. The proportion of households classified as being at risk of energy poverty is readily apparent from this figure. For comparison, data from previous research conducted by Dalla Longa et al. in the Netherlands was added, in which a similar energy poverty classification framework was applied [17, 26]. The Dutch data set is different from the Enable data set used in this research; it contains more data points, continuous numeric answers, and concerns neighborhood averages instead of single households. As the expenditure threshold was set at the 80th quantile, a horizontal line denoting this threshold between income risk and expenditure risk can be observed. An exact 80th quantile split cannot be defined for all discrete numbers, therefore, some countries in the data set show a minor deviation from this boundary.

In a perfectly representative sample of a country, the income risk and double risk group combined should add up to 30% as the income threshold is set between the third and the fourth income decile of a country. However, this is not the case in our data set, where the sum of the two groups ranges between 15% for Germany, to 82% for Poland. This is because the distributions of sampled households are skewed to the higher and lower income deciles, for Germany and Poland respectively. As previously discussed, this is not deemed problematic to the research goal. It does, however, limit the conclusions that can be drawn regarding the prevalence of energy poverty in each of the countries, as this can also be caused by selection bias.

We observe a distinction between Western European countries, characterized by lower double risk shares, and Central and Eastern European (CEE) countries, where double risk shares are more pronounced. The UK is found to be an outlier in this respect: while widely considered a Western European country, its significant share of households categorized in the double risk category position it closer to the CEE countries. Energy poverty is researched extensively in the UK, and several government policies aimed at reducing it have been implemented. This could be because disproportionately more households experience energy poverty in the UK than in other Western European coun-

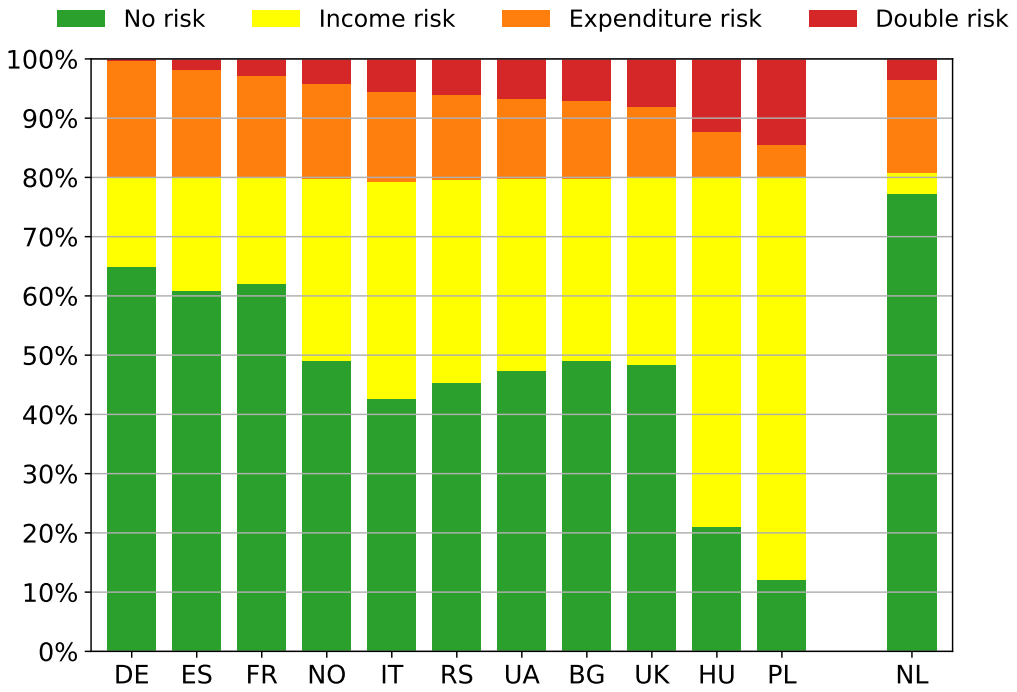

Figure 3.1.: Energy poverty classification distribution for countries in the Enable-EU data set (the columns for the Netherlands is derived from previous research [17]).

Table 3.1.: Features used in the model, type, and number of unique values.

name	type	# unique values
income bracket	integer	10
floor area	integer	7
household size	integer	17
house detachment	integer	5
house age	integer	9
birth year respondent	integer	79
heating strategy	categorical	5

tries. Energy poverty is shown to affect public health [27]. Its prevalence might coincide with the UK having the second most (after Ireland) long-term excess winter mortalities among 30 European countries [28]. We speculate that this difference could be explained by the fact that the UK is the birthplace of capitalism. To this day it is being considered closer to the free-state market than mainland European countries, generally regarded to have larger welfare states.

The observed double risk shares are a consequence of two phenomena at play. On the one hand, we have genuine energy poverty prominence in a country, resulting in a large number of points in the double risk category. On the other hand, we have the non-representative sampling of income deciles in countries, resulting in misplaced thresholds and misrepresented risk groups. To what extent each factor plays a role, differs per country and falls beyond the scope of this research. Consequently, although providing an intriguing indication for the prevalence of energy poverty in these countries. Further research with improved data quality and quantity is necessary to confirm these observations.

A CatBoost model was trained to classify households into one of the four energy poverty risk categories. A household is represented by seven selected features that can be found in Table 3.1. These features are hypothesized as potential predictors and selected as a result of extensive data analysis and domain expertise. Some features, such as floor area, are categorical but have a distinct ordering to them, and are represented by an integer. These features are therefore depicted as type "integer" and do not require any treatment before being used in a decision tree model. *House detachment* corresponds to the survey question "Which best describes your home?" and has four possible answers. The answers range from "single-family house detached from any other house" to "apartment in a building with 6 or more flats," and are interpreted as a scale that determines how well insulated the home is by surrounding homes. Multiple country level studies have hypothesized that the level of house detachment plays an important role in energy poverty [29, 30]. *Heating strategy* corresponds to a question on what heating methods households employ. The feature value can be one of 5 different heating strategies described. This is a categorical feature with no clear ordering, and thus cannot be used directly. It has type "categorical" and is handled by the CatBoost

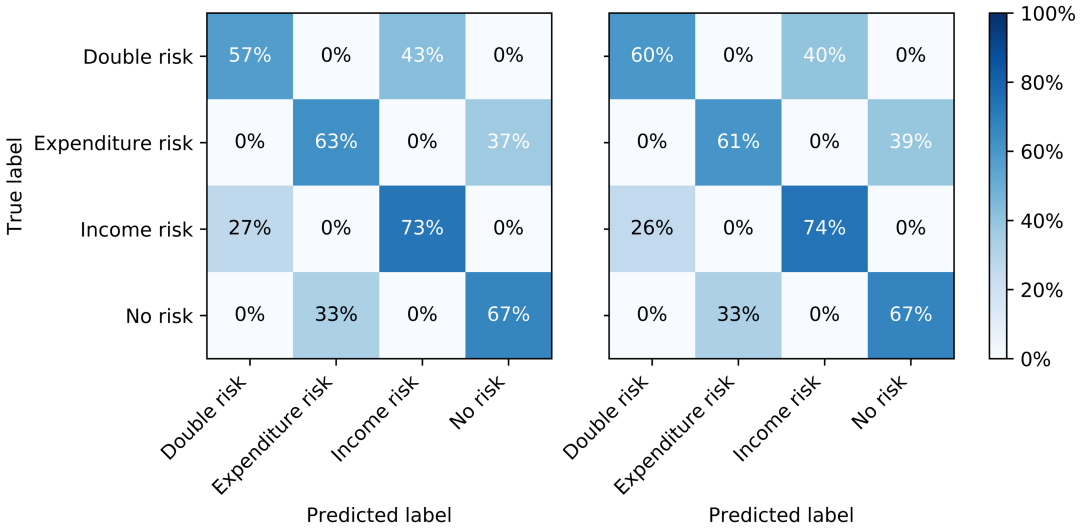

Figure 3.2.: Confusion matrix of the final model, on the validation (left) and test (right) set.

library. Further details are explored in Part III.

In Figure 3.2, a confusion matrix is used to visualize the performance of our model. A confusion matrix tabulates the labels as predicted by the model versus the true labels. In this figure, we have normalized all rows to sum to 100%. Values on the diagonal of a confusion matrix are known as the true positive rates: a model able to correctly identify all instances would result in a 100% true positive rates for all categories. A model randomly assigning labels, i.e. random guessing, would result in a score of 25% in all cells of the confusion matrix. Given our classification framework (Figure 2.1), using income as the only feature would allow a model to learn the income threshold and split the task into two binary classification tasks. Subsequently, the performance would deteriorate to almost random guessing (50% diagonal), this has been plotted in Figure A.4. The performance of such a one-feature model can be improved on by including the other features introduced in Table 3.1. The resulting model performs much better than a 50% diagonal in the confusion matrix, yielding true positive rates between 60% and 74% on the test set. Similar scores were achieved on the validation set, indicating that the model is not just memorizing the training data - known as overfitting - but that the performance generalizes well to unseen data.

The true positive rates vary per category. Compared to a similar study of the Netherlands, where diagonals scores ranging between 73% and 82% were achieved [17], the results achieved in this study are worse. The performance difference can be explained by the small amount of data available for this research, compared to the Dutch research. While the modelling tasks is extended from one country to multiple countries. This likely resulted in the model being too simple to capture the full complexity of the con-

cept; this is known as underfitting.

3.2. Interpretable Artificial Intelligence

In this section, the model we have trained on the data set is analyzed. We attempt to open the black box to gain insight into the inner workings of the model.

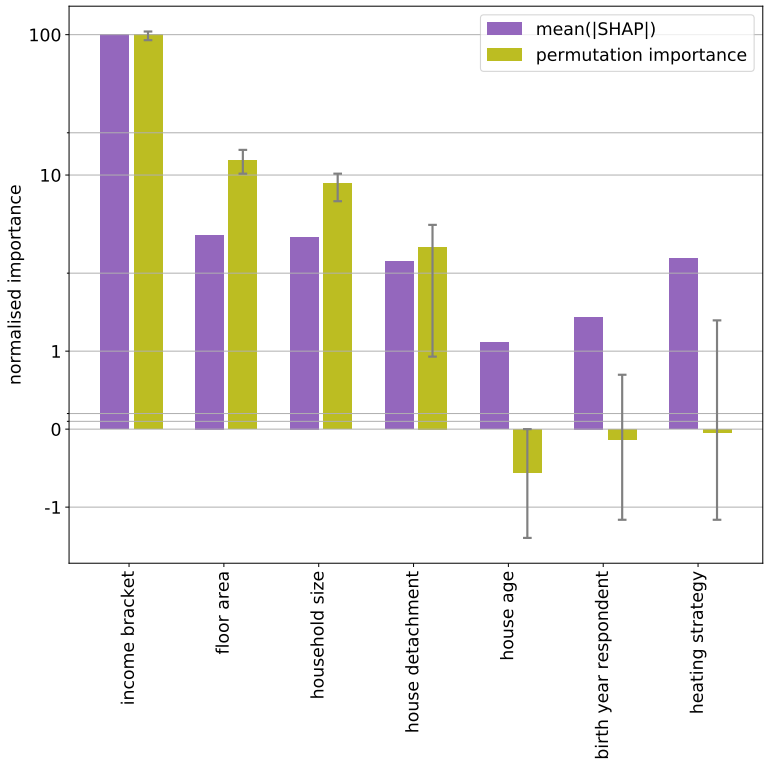

Figure 3.3.: Permutation and mean absolute SHAP value feature importance of the seven features in our model plotted with a symmetrical logscale on the y-axis.

Having successfully built an ML model to categorize energy poverty in a diverse group of European countries, we can now assess to what extent the various features included in the model influence its outcome. In order to accomplish this we introduce two measures of feature importance: The permutation importance and the mean absolute SHapley Additive exPlanation (SHAP) value. The values of these two metrics for all seven features in our model are plotted in Figure 3.3. For both methods, the feature importance values are normalized, such that the highest value equals 100 and the rest of the values are given relative to that one. As income bracket is clearly the most important feature for both methods, to accommodate easy comparison, the bars are plotted on a symmetric logarithmic y-axis. To avoid having the plot go to infinity around zero,

the plot is linear in range $[-2, 2]$. For permutation importance, the features are permuted several times, each time yielding a (slightly) different feature importance. The resulting deviation between runs is depicted by the error bars in the plot. Permutation importance assigns an importance score to a feature based on the effect of shuffling its values on the model performance [31]. SHAP values have a strong mathematical foundation in cooperative game theory [32]. For a single input, each feature's contribution to the final prediction of the model is computed as if it were a coalition game in which each feature would get a "payout". This is done for each individual input in the unseen test set. The mean absolute value of all contributions is used to attribute each feature an importance.

Both feature importance measures identify income bracket, floor area, and household size as the main drivers of our model. The features house age and birth year respondent are assigned significantly lower importance values in both metrics. For permutation importance, the error bars span from negative to positive feature importances, and are centered close to zero for these features. This implies that these are not good predictors for the model and thus can be disregarded as such. This is confirmed by the findings of the mean absolute SHAP values, that also indicate they are of little importance. Although not as overtly as with permutation importance. All features identified as being important to our model, except income, are directly related to the heat demand of the household. Household size additionally affects electricity demand, while this is relatively unaffected by floor area and house detachment.

With respect to heating strategy and house detachment, the results of the two methods diverge. Permutation importance results rank heating strategy as having no impact on the model performance with error bars running from negative to positive values, centered around zero. However, mean absolute SHAP value assigns it equal importance to house detachment. SHAP values approximate different feature coalitions to determine the feature importance. This can be exactly determined by training a new model with a different set of features, this was done several times. The results from these models led us to confirm the results of the permutation feature importance. Models with the feature house detachment included perform slightly better than those with the feature heating strategy. However, in order to assess these features as true predictors, further research is necessary.

A way to assess the effect of a feature value model output is by using Partial Dependence Plots (PDPs). Friedmann proposed PDPs as a comprehensive summary of the model dependence in his paper introducing gradient boosting [22]. The PDPs for each of the risk classes and each of the features are plotted in Figure 3.4. It shows the partial dependence between the model's output for every energy poverty risk group and every single feature of the model, averaging over all other features. The y-axis depicts the partial probability of the category value being predicted as the line's corresponding class. A straight line suggests that the feature has no influence on the model's prediction. A very volatile line, for example for income, suggests that the feature has a big impact on the model. The shared y-axis enables a simple comparison of the volatilities of the partial dependencies. PDPs do not provide a complete explanation of how the

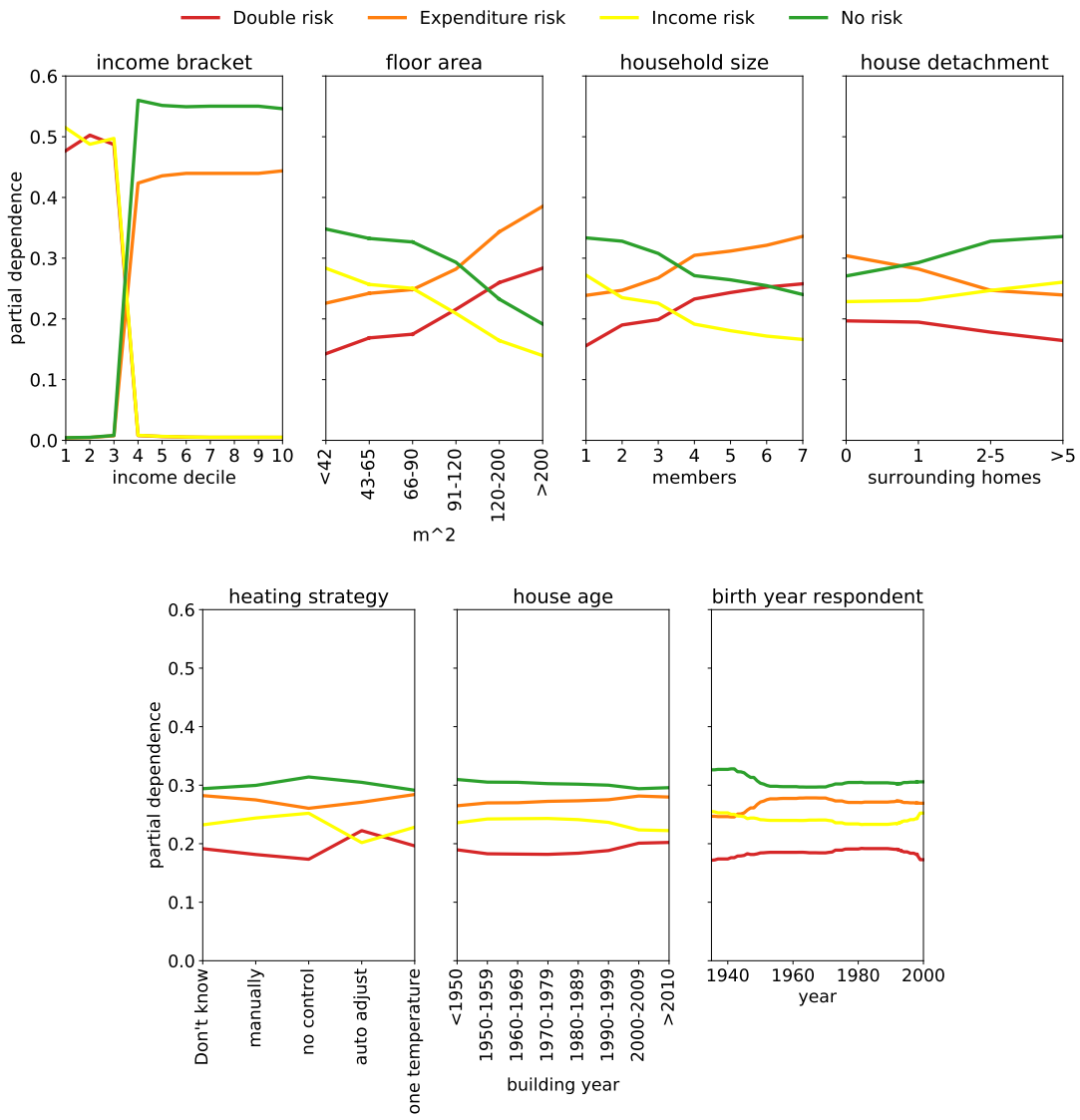

Figure 3.4.: Partial dependence plots for the seven features of our model.

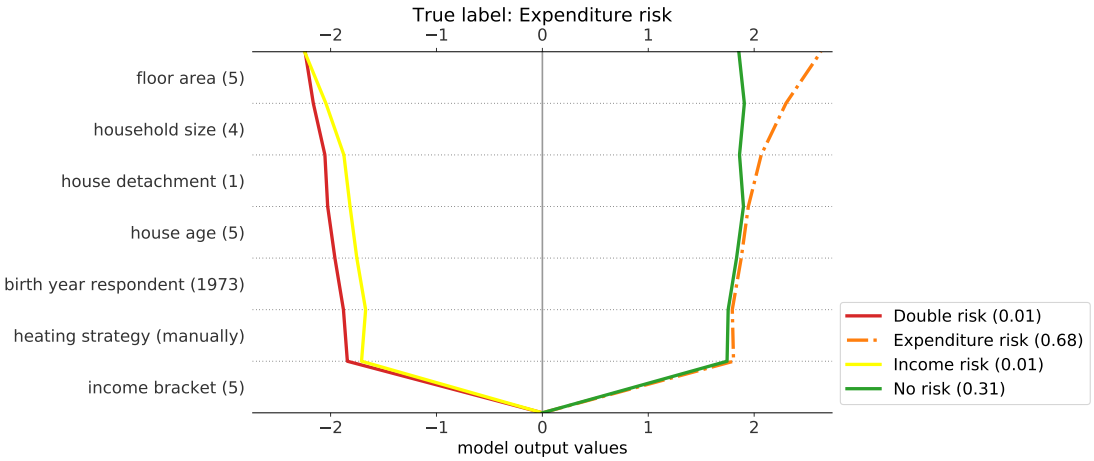

Figure 3.5.: Decision plot of a single sample in the test data set. The feature values are in brackets.

features affect the model's output but can provide valuable insight into it. Interactions between features could be averaged out to get this 2-dimensional visualization, thus are missed. The PDP for income confirms what can also be observed in the confusion matrices. Based on income bracket, the model reduces the multiclass classification task to two binary ones: if the income is below the income threshold, probabilities for the no risk and expenditure risk categories drop to zero. If the income is above the threshold the inverse is true, and the probabilities of income risk and double risk drop to zero.

In the other plots we can distinguish two pairs of classes with similar profiles. The no risk and income risk, and the expenditure risk and double risk classes show the same trends for floor area, household size, and house detachment. Whereas in the features with a low permutation feature importance, the partial dependence does not deviate much for each of the answers. The pairs could be interpreted as being the less- and more at risk classes, respectively, for both binary regimes, based on whether the household's energy expenditure spending exceeds the threshold. The plots indicate that a larger house floor area increases the likelihood of a household being classified in the higher risk category. The PDP for the feature household size shows a similar profile, where a larger household shows increased probabilities of being at risk. For house detachment an inverse relation can be observed, where increasing along the x-axis indicates a home more insulated by others, results in a decreased probability of being in the higher risk category as the house detachment category increases.

In order to provide additional insight in the internal workings of our ML model, we introduce the notion of a decision plot. A decision plot [33] exploits the additive nature of the feature contribution assigned by SHAP values by providing an effective visual summary. A visualization for a single household from the test set can be seen in Figure 3.5. The y-axis has the features followed by the feature value in brackets. The effect of the feature is depicted between the two horizontal lines. The values on the x-axis corre-

spond to the model output, these values are transformed by the model to probabilities (Equation 7.7), which are depicted in brackets in the legend. The dashed line is the eventual classification done by the model. This plot depicts how a large household living in a large home, detached from other houses, living of an income in the fifth decile, is correctly classified as being in the expenditure risk group. The split of the task into two binary classification tasks can clearly be observed (red and yellow versus green and orange lines). The little impact the features heating strategy, birth year respondent, and house age have on the prediction is also apparent. By analyzing these plots, one could in principle differentiate several characteristic decision paths corresponding to certain types of households that are more prone to be in energy poverty. We believe this could be beneficial for policy design in the future. In order to properly carry out this type of analysis, however, one would need a larger and more consistent data set than the one our current model is based upon.

4. Conclusion and Discussion

We have successfully applied ML to find features that have demonstrated their predictive power in a heterogeneous data set comprised of 11 countries, representative of the European continent. The resulting model is better at classifying households categorized as being at risk of energy poverty than a theoretical model with only income as an input. However, it does not attain the same level of accuracy as a gradient boosting model achieved in a similar study performed in the Netherlands. This leads us to hypothesize about the presence of two types of predictors for energy poverty in Europe. *Universal* predictors are indicators valid across a varied set of countries, e.g. for the whole of Europe. Besides these, there are also predictors concerned with the local specificities of a country, which we call *contextual* predictors. The universal predictors can serve as a starting point for European countries to establish an overarching framework to assess energy poverty. These can then be complemented in individual countries (or regions) by contextual predictors, thus adequately assessing the prominence of energy poverty at (sub)national level. This concept of having a national definition of energy poverty that is complementary to a common European definition has been proposed by other studies in the literature [3], our research supplies the first empirical support.

The features income, floor area, and household size were all found to be of significant importance to our energy poverty risk classifier. The results suggest these three features as potential universal predictors on the European continent. While our results were inconclusive with regards to the house detachment and heating strategy, the features house age and birth year of respondent have been found to be insignificant on a supranational scale. House age has previously been found to be an indicator of energy poverty [34]. We assign this apparent discrepancy to the heterogeneous nature of the data set used. As was established in [20], house age cannot be considered a proxy for insulation on a supranational level. Nonetheless, it could potentially be a contextual predictor. Further research, based on higher quality data sets, is necessary to confidently claim or disregard any of these features as predictors.

The results reported in this thesis should be considered in the light of certain limitations. Absolute energy expenditure was used to categorize households into an energy poverty risk group, as the data set did not allow for easy conversion to relative energy expenditure. As a result, some households might have been miscategorized as energy poor. Furthermore, households that severely under-consume are also missed by this framework; this atypical form of energy poverty is an open research question and beyond the scope of this research.

The lack of data might have caused our model to underfit, resulting in it only solving part of the complex modelling task. Not enough high quality data was available to effectively train a more sophisticated model. Therefore, we recommend better data

collection. The lack of data regarding energy poverty has long been recognized in the field [35, 36, 37, 30, 7, 38]. This is seemingly in the process of being resolved. The Energy Poverty Observatory (EPOV) was founded with this goal in mind: to improve and harmonize data collection [39].

The Enable data set suffered selection bias with respect to the income deciles, we suggest improving data collection by taking more representative samples. Homogeneity across income is an important aspect due to the close relationship between energy poverty and poverty. Furthermore, to explore the contextual predictors in a country, more data needs to be collected on the national level. Many national bureaus of statistics were consulted for this research, and data regarding energy poverty at the household level was requested, but no comprehensive and complete data sets were available for our purposes.

To stimulate the energy transition to a more sustainable energy supply, some sort of carbon taxation is imposed in many states in the US and EU. Spending the revenues of this on energy R&D and technology innovation is a commonly heard suggestion. Another suggestion is to use those revenues to assist people in meeting energy poverty challenges, which may be exacerbated due to stringent climate change mitigation measures. One question raised is how to determine which people will deserve and receive assistance, and which do not. Our analysis can help in determining who may be in dire straits when it comes to affording basic energy services, and ultimately help determine appropriate types of assistance for different groups of energy poor households. XAI methods could be instrumental in efforts to distribute greenhouse gas taxes through policy, if the targets include the alleviation of poverty and establishment of equity among consumers of energy services.

Part II.

Data Preparation

5. Data

Enable-EU provides the data set used for this research [18]. In this section the data set is described, and the preprocessing steps performed are discussed.

5.1. Data description

The data set is available on the website of Enable-EU and was published in March 2018. It is derived from a survey conducted in eleven countries in Europe. Income and energy costs, comprising of heating and electricity, are reported for households, making the data set eligible to apply the framework to. The survey, accompanied by instructions for the conducting parties, was published along with it. The data set was published in Excel format. It contains 473 columns and 11.265 rows. The diverse group of eleven countries can be considered representative of Europe, and thus serves as an excellent sample to conduct our research of finding energy poverty predictors valid across Europe.

The survey is divided into several different sections: general questions, mobility, shift to prosuming, heating and cooling, use of electricity, and governance framework. Only the general questions were surveyed in each country, namely a section "home / building characteristics and household possessions" and "social and economic characteristics". As the research aims to find European predictors, the data set was limited to only questions in these two sections.

The general section comprised of 24 questions, with sub questions, resulting in 86 columns being available. Most questions required respondents to answer categorically, the answers were encoded in the data set by integer values representing the categories. For each of the columns there was a description of the question it represented available in the Excel file, and the correspondence between the answers and the integer values was also provided. Missing values were handled inconsistently: for many columns there was no data, represented by a Not a Number (NaN) value, for some there was a special integer value indicating that no value was available, and for some there was an additional indicator column to indicate that the respondent did not answer the question.

5.2. Preprocessing

The data set had to go through several preprocessing steps before it was ready to be used. Data preprocessing steps and data handling were performed in the Python programming language [40], in combination with the pandas library [41].

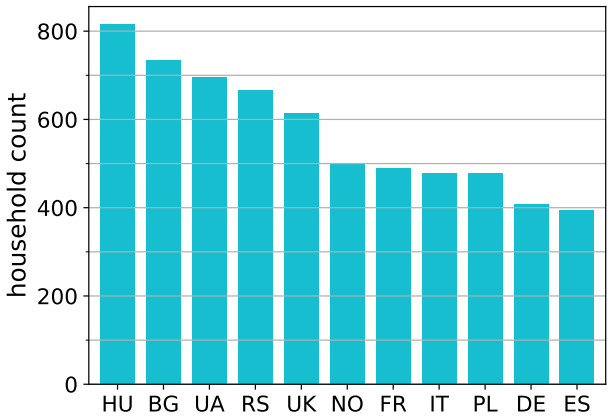

Figure 5.1.: Number of respondents from each respective country in the data set after cleaning.

In order to label the households in our data set, two variables were necessary: income and energy expenditure. The former could either be reported monthly or annually, therefore a simple check whether one of these two held a valid value was sufficient to guarantee income was filled. Energy expenditure was derived from 7 columns regarding heating and electricity cost. Two of these were columns indicating whether the question was answered by a respondent, four concerned yearly and monthly, heating and electricity costs. The last column corresponded to a question on the number of months the households paid for heating in the last heating season. After this data cleaning, the number of respondents per country ranged between 711 and 1500 respondents and is plotted in Figure 5.1.

Income was reported in two columns depicting yearly and monthly income, corresponding to how the people in the respective country generally calculate their income. The survey required respondents to classify their income into the corresponding decile in their respective country, leaving the results to be categorized into 10 brackets, represented by integer values 1 to 10. For some rows, the entry contained a NaN value; however, a value of 98 and 99 corresponds to "refused to answer" and "do not know" respectively. The data was cleaned accordingly and a column *income bracket* was engineered representing a uniform income measure.

The feature energy expenditure is constructed by summing the heating and electricity costs of each household, and - if necessary - convert them from monthly to yearly costs. The conversion of electricity monthly costs to annual costs is simply done by multiplying it by 12. Heating costs reported as monthly costs, had an accompanying column with a question on the number of months in which heating was required. If this column was filled, the costs were multiplied by it in order to obtain annual costs. If it was not filled, the costs were multiplied by the median number reported in the coun-

country	currency	exchange rate
Bulgaria	Bulgarian lev	0.511292
France	Euro	1
Germany	Euro	1
Hungary	Hungarian forint	0.003226
Italy	Euro	1
Norway	Norwegian krone	0.102712
Poland	Polish złoty	0.239406
Serbia	Serbian dinar	0.008446
Spain	Euro	1
Ukraine	Ukrainian hryvnia	0.029633
United Kingdom	Pound sterling	1.124859

Table 5.1.: Table of countries in the data set with their currency and the first available exchange rate to euros in 2018.

try. Costs were reported in the currency used in the country in question. Therefore, the yearly energy expenditures were all converted to euros using the first available exchange rate to euros in 2018. This was retrieved from online currency conversion tool xe [42]. The currency in each of the countries, and the exchange rate are tabulated in Table 5.1.

Household size was was derived from 6 columns making a distinction between age and gender of members of the household. Summing up the household composition gives us a household size, irrespective of the gender and age.

As absolute income was not reported in the survey, relative energy expenditure was not directly available. The thresholds used to delimit the income brackets of each country in the survey were not supplied in the data set, and a request for this data to the authors did not yield a response. In the instructions, it is stated that the income deciles as given by the national statistics for each country. To assign an absolute income to every income decile, the disposable income of the European Union Statistics on Income and Living Conditions survey was used [43], such that all approximations came from the same source. Even though it concerns disposable income instead of gross income, this was the preferred method to ensure consistency. For the deciles, the 9 cutoffs points were reported. Linear interpolation and extrapolation was used to approximate an income corresponding to the reported decile, depicted in Figure 5.2. Unfortunately, there was no data available for Ukraine, therefore an estimation was performed based on the relative GDP per capita compared to Serbia using data from the World Bank.

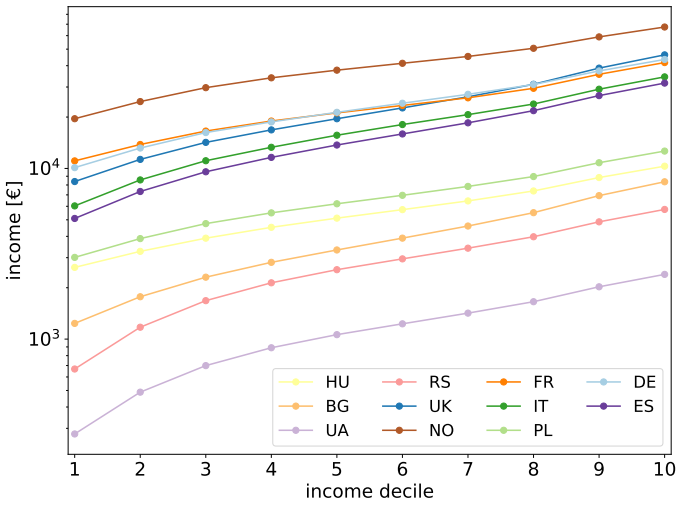

Figure 5.2.: Approximated yearly incomes corresponding to each income deciles plotted for all countries.

After all preprocessing was complete, the data could be labelled. The class distribution of the data set can be found in Table 5.2, and is visualized per country in Figure 3.1. The data set is imbalanced, indicating that the data points are not evenly distributed per class. This has consequences for the training of an ML model, handling the class imbalance is discussed in Section 7.5. The data was split into a training set and a test set, with the test set containing 20% of the samples. This was done in a stratified fashion which keeps the class distribution similar for both sets.

Risk group	Number of households	
No risk	2774	44.3%
Income risk	2228	35.5%
Expenditure risk	841	13.4%
Double risk	425	6.8%

Table 5.2.: The energy poverty risk label distribution of the data set

5.3. Categories

feature	category	value	count
floor area (H3)	Up to 42 m2	1	501
	43 – 65 m2	2	1527
	66 – 90 m2	3	1774
	91 – 120 m2	4	1502
	120 – 200 m2	5	717
	More than 200 m2	6	200
	Missing value		47
heating strategy (H9)	Manually adjust the temperature (e.g. at night or when no one is at home)		2261
	Our household does not have control over the equipment		1734
	Set one temperature and leave it there most of the time		1277
	Program the thermostat to automatically adjust the temperature during the day and night at certain times		960
	Don't know / No answer		36
house age (H2)	Before 1950	1	867
	1950 to 1959	2	624
	1960 to 1969	3	970
	1970 to 1979	4	1182
	1980 to 1989	5	973
	1990 to 1999	6	577
	2000 to 2009	7	466
	2010 to 2016	8	237
	Missing value		372
house detachment (H1)	Single-family house detached from any other house	1	2738
	Single-family house attached to one or more other houses (for example: duplex, row or terraced house, or townhome)	2	897
	Apartment in a building with 2 to 5 flats	3	662
	Apartment in a building with 6 or more flats	4	1955
	Missing value		16

Table 5.3.: The categories and corresponding values for each feature, with the corresponding question that can be found in Appendix B, in brackets.

6. Energy poverty classification framework

In this section, frameworks for classifying energy poverty in Europe are discussed. Indicators for energy poverty can be divided into two categories: self-reported/qualitative indicators, and measurable/quantitative indicators.

6.1. Existing frameworks

Fuel poverty has been studied for a longer time, resulting in more literature on identifiers being available for fuel poverty. A fundamental academic work for fuel poverty was published by Brenda Boardman [44] in which she argues for a clear distinction between poverty and fuel poverty. It is claimed that energy poverty is not exclusively a consequence of financial hardship. Consequently, eliminating energy poverty requires different households to be targeted than those targeted to eliminate poverty. It was proposed that in energy poverty, three important factors are at play: income, household energy requirements, and fuel prices.

When in the UK fuel poverty was first defined, a 10% relative energy expenditure threshold was used to classify households as being fuel poor. Later, this definition changed to the Low Income High Costs (LIHC) indicator. This classifies households as being at risk when they have required fuel costs that are above average (the national median level) and when this amount is deducted from their income, the household would be left with a residual income below the official poverty line [37]. This new indicator received a lot of criticism, and was even argued unlikely to have positive impacts for most fuel poor households [45].

The EPOV was founded by the European Commission to improve the knowledge on energy poverty and thereby help move forward in eliminating energy poverty [39]. The EPOV proposes four primary indicators: arrears on utility bills, low absolute energy expenditure (M/2), high share of energy expenditure in income (2M), and inability to keep the home adequately warm. The first three are quantitative, and the last one is a qualitative, self-reported indicator.

The 2M indicator is a statistically determined indicator that uses twice the median relative energy expenditure of a country as a threshold. A household is labelled as energy poor if its energy expenditure, as share of their gross income, exceeds this threshold. Another statistical indicator is the M/2 indicator, that uses half the median of the absolute energy expenditure in a country. A household energy expenditure that is less than this threshold is believed to be abnormally low. This could be caused by severe under-consumption to reduce costs, and consequently induce energy poverty [39].

Additionally, a pool of secondary indicators believed to be relevant to energy poverty

is mentioned by the EPOV. However, the indicators show contradictory results when applied to European countries. No coherent framework on how these indicators should be combined is provided by the EPOV. It is believed that a combination of indicators is necessary to adequately identify all households suffering from energy poverty [8].

In Table 6.1, the 2M and M/2 indicators are tabulated for all countries in the data set using an approximated income (derived as described in Section 5.2). Western European countries have a 2M indicator that is close to the 10% indicator previously used in the UK. However, Eastern European countries spend a substantially higher share of their income on energy. Therefore, a uniformly set threshold, such as the 10% one, appears unfit as a European indicator. Moreover, the table suggests that the heterogeneity of the data causes a large deviation in both relative and absolute energy expenditure. This suggests that this threshold should be determined for every country individually.

6.2. Employed framework

The energy poverty classification framework proposed by Dalla Longa et al. requires an income and an energy expenditure threshold to be set [17]. The framework is visualized in Figure 2.1. In that paper, the energy expenditure threshold is set at the 80th quantile. This approach was adapted for our research, such that every participating country has a distinct energy expenditure threshold corresponding to the 80th quantile.

The data presents a uniform income metric with the deciles it was reported in. This allows for one threshold to be set for all countries. The minimum wages in 2018 for each of the countries are retrieved from Eurostat and the corresponding deciles determined. The third income decile is the mode. Norway and Italy do not have a nationwide minimum wage [46, 47] and minimum wages for Ukraine were not available on Eurostat. A study conducted in Spain found that for three different energy poverty indicators, 99% of the households classified as being energy poor, are in the first three income deciles [21]. This lead us to set the income threshold between the third and fourth deciles in the current research.

Country	Households	Median	2M	M/2	Minimum wage bracket
Norway	499	€ 2311	12.6%	€ 1156	-
United Kingdom	613	€ 1215	13.3%	€ 607	3
Germany	407	€ 1900	13.7%	€ 950	3
Spain	395	€ 960	15.8%	€ 480	3
France	489	€ 1890	16.4%	€ 945	3
Italy	478	€ 1890	28.9%	€ 945	-
Poland	477	€ 862	37.5%	€ 431	4
Hungary	815	€ 890	48.2%	€ 445	5
Bulgaria	734	€ 798	53.1%	€ 399	4
Ukraine	696	€ 267	54.4%	€ 133	-
Serbia	665	€ 760	64.8%	€ 380	7
Total	6268	€ 997	31.1%	€ 499	-

Table 6.1.: The countries in the Enable data set with median annual energy expenditure, two conventional indicator values, and the income decile the minimum wage corresponds to.

Part III.

Interpretable Artificial Intelligence

7. Model building

This section involves the process of obtaining an ML black box model. First, the concept of gradient boosting is explained, followed by the library of choice in this research, CatBoost. Sections 7.3, 7.4, and 7.5 consider the steps taken to employ CatBoost for the model creation.

7.1. Gradient Boosting

Gradient boosting is a popular machine learning method, demonstrating state-of-the-art performance on many modelling tasks, especially on tabular data with meaningful features [24, 15]. It is highly suitable for interpretation, making it a useful ML method for this research. Gradient boosting was developed as a combination of steepest descent optimization (gradient) and additive modelling (boosting) to solve a function estimation problem. It was first proposed in 1999 by Jerome Friedman [22]. In a function estimation problem, as in Equation 7.1, one approximates an unknown function that maps an input vector $\mathbf{x} = \{x_1, \cdots, x_n\}$ to an output value y. In practise, a natural process is being estimated, thus the true function is a natural process. To estimate the true function, the expected value of some loss function $\Psi(y, F(x))$ should be minimized for a set of observations of this true function. Those observations make up the training data $\{y_i, \mathbf{x}_i\}_1^N$. Rather than the more abstract function estimation, gradient boosting using decision trees is described here.

$$F^*(\mathbf{x}) = \underset{F(x)}{\operatorname{argmin}} E_{y,\mathbf{x}} \Psi(y, F(\mathbf{x})) \tag{7.1}$$

The true function is approximated using a method derived from additive modelling, which is a technique for the creation of nonparametric regression models by summing simple functions to approximate a more complicated one [48]. An initial guess $F^{(0)}$ ($F^{(j)}$ denotes the jth iteration of our approximate function) is incrementally combined with a simple function $f_m(\mathbf{x})$ that helps to minimize the loss, this is known as *boosting*.

$$F^{(M)}(\mathbf{x}) = F^{(0)}(\mathbf{x}) + \sum_{m=1}^{M} f_m(\mathbf{x}) \tag{7.2}$$

Each addition to the model consists of a *base learner*[1] that attempts to minimize the loss function value. In our case a base learner consists of a simple decision tree with L

[1] Some authors call these *boosts* or *steps*

leaf nodes. The decision trees can be parameterized to get $f_m(\mathbf{x}) = h_m(\mathbf{x}; \mathbf{a}_m)$ where $\mathbf{a}_m$ denotes the splits of the decision tree in the m-th iteration.

$$\mathbf{a}_m = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(y_i, F^{(m-1)}(\mathbf{x}_i) + h(\mathbf{x}; \mathbf{a})) \tag{7.3}$$

We minimize the objective in Equation 7.1, using steepest descent. The base learner is fit by parameters optimization to the negative gradient of the loss function, known as *pseudo-residuals*.

$$\widetilde{r}_{im} = -\left[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F^{(m-1)}(\mathbf{x})} \tag{7.4}$$

The feature space is partitioned into L disjoint regions $\{R_{lm}\}_{l=1}^{L}$ each corresponding to one leaf node. The decision tree base learner assigns the value corresponding to the leaf node where the input $\mathbf{x}$ ends up at.

$$h(\mathbf{x}; \{R_{lm}\}_{l=1}^{L}) = \sum_{l=1}^{L} \widetilde{y}_{lm} 1(x \in R_{lm}) \tag{7.5}$$

The value assigned to each leaf node is the mean of the pseudo residuals of all training samples that are in the corresponding feature region: $\widetilde{y}_{lm} = mean_{x_i \in R_{lm}} \widetilde{r}_{im}$. A regularization term ν is added to allow the model to depend on more base learners and produce a more robust model, less prone to overfitting. Typically, a value ≤ 0.1 is used as shrinkage factor. Putting this all together results in the iterative procedure known as gradient boosting.

$$F^{(m)}(\mathbf{x}) = F^{(m-1)}(\mathbf{x}_i) + \nu \left(\widetilde{y}_{lm} 1(x \in R_{lm})\right) \tag{7.6}$$

For classification with K classes, K ensembles are created. Each with an output value that is transformed using a *softmax* function that normalizes the output to a probability distribution. The function F_k depicts the ensemble that gives an output value for class k. Each class gets a probability, such that all sum to 1.

$$p_k = \sigma(\mathbf{x})_k = \frac{e^{F_k(\mathbf{x})}}{\sum_{l=0}^{K-1} e^{F_l(\mathbf{x})}} \tag{7.7}$$

For multiclass logistic regression with cross-entropy loss function. The vector y_i consists of all zeros, except for the correct class, which has a value of one. The loss reduces to the negative log of the probability assigned by the model to the correct class. In practise, a regularization term is added that penalizes complicated tree structures, guiding the algorithm to build simpler trees [49]. This is done in order to prevent overfitting.

$$\Psi(\{y_k, F_k(\mathbf{x})\}_1^K) = -\sum_{l=0}^{K-1} y_k log(p_k(\mathbf{x})) = -log(p_{truelabel}(\mathbf{x})) \tag{7.8}$$

The technique of gradient boosting was improved on, by introducing stochasticity to it. Instead of using the full data set to calculate the gradient and fit a decision tree to, bootstrap aggregating (*bagging*) is used [50]. Here, in each iteration, a random subsample of the data set is selected to fit a decision tree on. This procedure prevents the model from overfitting on the training data. Every gradient boosting library makes its own minor tweaks to the algorithm, through which it attempts to maximize performance.

7.2. CatBoost

CatBoost is an innovative new gradient boosting library developed by the Russian tech company Yandex [23, 25]. As opposed to many other popular gradient boosting libraries - such as scikit-learn [51], XGBoost [24], and LightGBM [52] - it can deal with categorical features without any preprocessing steps required. It also drastically improves prediction time and has many model analysis tools available.

Most gradient boosting libraries only support numerical features, such that two values can be compared using a $>$ or $<$ operation, necessary to perform splits. Categorical data can have ordered categories, allowing them to be split using these operations. For example, income deciles are categories where each one represents a range of incomes, and therefore, can easily be compared. This type of categorical data is known as *ordinal* data. Decision trees are invariant to monotonic transformations [53]. As a result, it is easy to assign values to an ordinal feature, as only the ordering is relevant. There are also cases where there is no clear ordering among the categories, for example, a sentence describing the heating strategy of a household. Data corresponding to this type of category, is known as *nominal* data. Nominal data has to be processed before it can be used in most gradient boosting algorithms. CatBoost performs the preprocessing steps required for nominal data internally.

There are several methods that process nominal data. A frequently used method is the one-hot-encoding (OHE) method, in which the feature is transformed into as many columns as it has unique values, the number of unique values is also known as its *cardinality*. Each of the columns is a binary column; the column corresponding to the answer is assigned a value of one, and all others are set to zero. This introduces a lot of sparsity into the data. The nominal data can be transformed to ordinal data. This is done by assigning each category a useful numeric representation, thus resulting in the data having a new artificial ordering. This is usually done by using the mean target value for a category over all samples. CatBoost uses a combination of the two methods described above. If the cardinality of a categorical feature is equal to 2, OHE is used to encode the feature. Otherwise, meaningful numeric values are assigned, calculated using the training data.

$$\frac{\sum_{j=1}^{n}[x_{j,k} = x_{i,k}] * Y_j}{\sum_{j=1}^{n}[x_{j,k} = x_{i,k}]} \tag{7.9}$$

Along all n samples, each with m features , $(x_{p,q}) \in \mathbb{R}^{n \times m}$, every category of feature

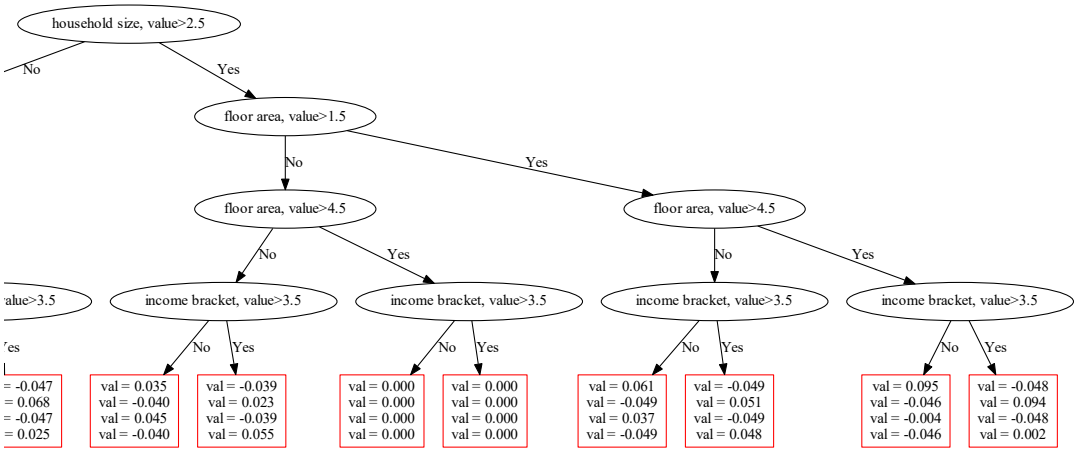

Figure 7.1.: One side of an oblivious tree applied in CatBoost.

k is assigned a numeric value using this equation. For all samples logical proposition $x_{j,k} = x_{i,k}$ is evaluated in Iverson brackets, represented by $[\cdot]$. It equals 1 if the proposition is true, i.e. the feature values are the same, and 0 otherwise. A modified version of Equation 7.9, including a novel scheme involving bagging to prevent overfitting, is used in CatBoost to deal with categorical data [25].

The first tree of the final CatBoost model is plotted in Figure 7.1. CatBoost uses oblivious trees as base learners, those are trees that use the same criterion for every split on the same level in the tree. Consequently, all trees have the same number of leaf nodes. This is claimed to prevent overfitting of the model. Instead of inducing K trees for a classification task with K classes, a contribution is assigned to every class at each leaf node. The observation done in Section 3 that the model splits the multiclass classification task into two binary ones, can also be observed in this figure. The split on income bracket is made at 3.5, the income threshold is set at exactly this value. The model has learned this correctly. The oblivious trees lead to more, increasingly redundant splits being performed. This is illustrated by certain leaf nodes assigning zeros to all classes as no input can reach it. As a result, training can take longer, but it enables an implementation for prediction that is significantly faster than other libraries.

7.3. Feature selection

The most influential feature to our model is income. Poverty and energy poverty are closely connected, income is one of two features used to label the data. It was decided to include income as a feature as it is an easy metric to get hold of for policy makers. The additional features were chosen using domain expertise. From this point, a lot of manual labor with trial and error resulted in the final selection of features. These

resulted in the best model, and also included the features previously hypothesized to be at play in predicting energy poverty. The final features selected can be found in Table 3.1.

The features: *income bracket*, *house age*, *house detachment*, and *floor area* have a distinct ordering of the corresponding answers, thus are ordinal features. *Household size* and *birth year respondent* are numeric answers, represented by positive integer numbers. The aforementioned features can be handled in a gradient boosting model with no preprocessing necessary. Lastly, *heating strategy* corresponds to a question on what heating strategy the households employ. This is one of 5 different heating strategies described which can be found in Table 5.3. This is a nominal feature and is treated by CatBoost with no preprocessing required.

Feature selection by feature elimination with cross validation [54] has been attempted. With this technique, a model is trained on an initial set of features. The feature with lowest assigned permutation feature importance is dropped and a new model trained. This is iteratively done until the new model no longer performs better than the previous. The resulting model indeed showed better results than our final model on the validation set. However, it did not generalize well and performed poorly on the test set. It barely performed better than random guessing after a split was performed on income, resulting in diagonals around 50%.

7.4. Hyperparameter optimization

CatBoost has shown that it performs better with default parameter settings, than three other popular gradient boosting libraries with tuned hyperparameters on several publicly available data sets [25]. As a result, less time need to be spent on optimizing the hyperparameters. However, to maximize performance, hyperparameters are still to be tuned. CatBoost provides several optimization tools such as grid search and random search. Grid search is a simple form of hyperparameter optimization, in which a grid of hyperparameters specified by the user, is exhaustively searched. For each point a model is trained, and its performance compared to models trained with other hyperparameters on the grid. In random search, random hyperparameter settings are tested, this is often used to find a region in which to apply grid search. The three main hyperparameters in gradient boosting that can be tuned are the shrinkage factor ν, often called the *learning rate*; the number of iterations or trees M; and the number of leaf nodes L of the base learners. Besides these, many other hyperparameters can be tuned regarding, for example, regularization and bagging.

In his original paper, Friedman used L to denote the number of leaf nodes. As in the CatBoost library oblivious trees are used, only the depth of the tree is a hyperparameter. The depth of the trees determines what level interactions between features the model can capture. This can be seen by observing the *ANOVA* expansion of a function.

$$F(\mathbf{x}) = \sum_{j} f(x_j) + \sum_{j,k} f(x_j, x_k) + \sum_{j,k,l} f(x_j, x_k, x_l) + \ldots \tag{7.10}$$

For a depth of 1 we only get the effects of the first sum, these are known as *main effects*. Each function in it only depends on one input variable. Every summation we go further, an interaction between input variables is added to the expansion. Every extra level of depth we allow our base learners to go to, adds another summation of the expansion that can be approximated by the model. Generally, the first few sums explain most of the variance in the original function [22]. Grid search resulted in an optimal tree depth of 4. This is less than the default value of 6.

Grid search determined the optimal learning rate to be at 0.05. The number of iterations, trees, was set to 500. However, as CatBoost saves the best model it encountered during training, the resulting model did not consist of 500 trees. The model started overfitting long before this point. The training process is described in Section 7.5 and plotted in Figure 7.2.

The data set available was imbalanced, as can be observed in Table 5.2. This can be circumvented by taking a subsample of the training data with the same number of points from each class, known as undersampling. Undersampling is a robust and effective approach to counter class imbalance [55]. The data set used in this research does not have enough data points available to perform undersampling, as this would result in too few samples to effectively train a model. Another method is to subsample with replacement. This would include the same data point several times in the data set used by the model for training. This is known as oversampling. There are also advanced methods available to synthetically create additional data points of the minority class. Most methods use some variation of taking a random point on a line in feature space, between two data points from the same class [56, 57].

As we are trying to understand the mechanism that drive energy poverty in Europe, artificially generated data points might lead to results not reflected by real data. Therefore, *class weights* were passed to the classifier. CatBoost multiplies the gradient of a sample with the weight corresponding to its true label. Weight were calculated on the data set relative to the reciprocal of a class count in the data set. This has a similar effect on training as oversampling. With oversampling, if from one class with two data points, one is doubly sampled. The gradient of this data point is used twice. Whereas, by using class weights, both data point's gradient is multiplied by 1.5. As a result, we get a "smoother" form of oversampling.

7.5. Training

The model is trained using weighted multiclass logistic regression loss function, depicted in Equation 7.11. The weights for each class are determined over the training set. Splitting the data that is stratified by their labels ensures that the training, validation, and test set have a similar class distribution.

$$-\frac{\sum_{i=1}^{N} w_i log\left(\frac{e^{s_{i,y_i}}}{\sum_{j=0}^{M-1} e^{s_{i,j}}}\right)}{\sum_{i=1}^{N} w_i} \tag{7.11}$$

Catboost has many tools to monitor training, all metrics specified by the user are plotted in real time. The model that performed best on the optimization metric, loss function, is saved. The library also enables the user to use a different metric to evaluate the performance of the model. This can be a non-differentiable function, that is evaluated at the end of each iteration. The evaluation metric is used to determine the best model, and the differentiable loss function is used for optimization, to calculate the gradient to construct the next tree for our ensemble.

For model evaluation, some important measures are the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). For classification with more than one or two classes, all classes other than the one evaluated are assumed to be of the negative class. In a confusion matrix with absolute numbers, the TP representing the samples correctly classified as the class being evaluated is the value on the diagonal. TN is the number of samples correctly not classified as the class: all entries except the entries in the row and column corresponding to the class. FP are all samples that were incorrectly classified as the class: the sum of all entries in the column, except the diagonal. FN are all samples of our class that were not classified as being of the class: the sum of the entries in the row except the diagonal.

Precision $= \frac{TP}{TP+FP}$, and recall $= \frac{TP}{TP+FN}$ use the above described metrics to give one number that encapsulates the relevance of the found positives by the model. Both metrics can effortlessly be maximized by either classifying all (recall), or only a minor fraction (precision) of the samples as positive. The F1 score is the harmonic mean of the precision and recall and evaluates a model on both metrics simultaneously. It is used as the evaluation metric in this study. It has long been known as the Sørensen–Dice coefficient to measure the degree of similarity between two sets [58, 59] given in Equation 7.12.

$$F1 = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FN + FP} = \frac{2(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{7.12}$$

However, since we have four classes, we calculate the F1 score for each of the classes and weigh them to get a total F1 score.

$$TotalF1 = \frac{\sum_{k=0}^{M-1} w_k F1_k}{\sum_{k=0}^{M-1} w_k} \tag{7.13}$$

The training and validation learning curves are plotted in Figure 7.2. The loss function starts around 1.3, which makes intuitive sense as then it would do little more than random guessing, which would result in a loss of $-log(\frac{1}{4}) \approx 1.39$ for all classes. Overfitting of the model can be observed in both plots where the training and validation curves start to diverge. The iteration with the best score on the validation set is indicated with a dashed black line. This is at iteration 126 with a TotalF1 score of 0.65, and at iteration 264 with a loss function value of 0.64. The model results generalized poorly at the point of minimum loss function value. We hypothesize this could be caused by some of the risk classes being so small that the model starts overfitting on the few samples that are in the training and validation set. A decision boundary tighter around these samples

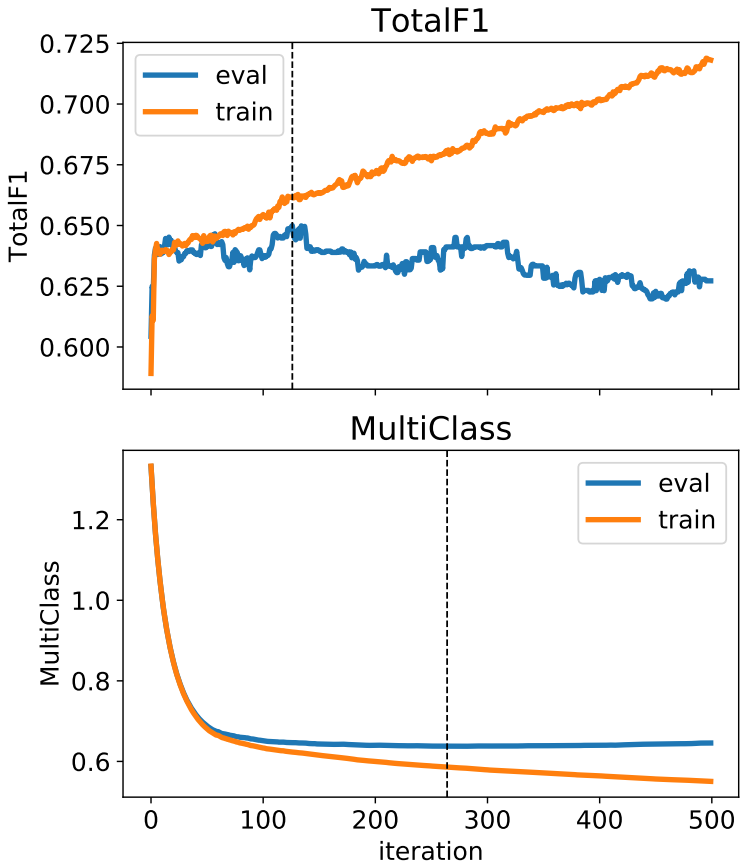

Figure 7.2.: Scores on validation and training set during training time. The dashed black line indicates the iteration where the best validation score occurred.

would decrease the loss but hinders the generalization of the model. Five fold cross validation shows very similar learning curves with the best average TotalF1 score of 0.63 at iteration 118, iteration 254 minimizes the average loss function at a value of 0.65. These results confirm our findings and show that the results are robust.

8. Model analysis

The methods used in this research are detailed in this section. Two methods that are used in this research to analyze the internal model of the classifier, are discussed. A relatively simple, but robust method, namely permutation feature importance is discussed. Another more advanced method that uses ideas from the Shapley value from game theory is also explained. Model analyses can be performed for a single sample, known as a *local* explanation. The model can also be analyzed as a whole, which is known as a *global* explanation.

Many gradient boosting packages provide a default method of measuring feature importance using an impurity metric. This is determined by the distribution of labels on each side of the split. A perfect split, has only samples of the positive class on the one side, and only negatives on the other. Evaluating the impurity on each side of the splits made by a variable results in a measure of its importance. This is a fast method and can provide some insights. However, this has a bias towards features used for overfitting. A model that is highly overfitted will assign a high feature importance to the features used to overfit. Features with few unique answers, low cardinality, are hard to overfit on. Contrarily, features with high cardinality tend to get an overestimated feature importance assigned. This holds especially true for categorical features [60]. Therefore, more advanced feature importance measures will be discussed: permutation importance and mean absolute SHAP values.

8.1. Permutation importance

Permutation importance is a global measure of feature importance for machine learning models. The feature values are randomly permuted and assigned an importance based on the resulting performance change of the model. Therefore, the assigned values can be positive when the performance of the model drops, but also negative when the model performs better after the feature values are scrambled. A negative or near-negative importance suggests that the feature does not aid the model in prediction and might be an indication of the model using the feature solely to overfit. Permutation feature importance was first described by Breiman [31]. An implementation by Python library Scikit-learn [51] was used in our analysis. The feature permutation feature importances are plotted in Figure 3.3.

The simplicity of the method is appealing and have lead people to try and improve on this method [61]. Permutation importance has been proposed as an unbiased feature importance measure [60]. It is computationally very cheap with a complexity of only $\mathcal{O}(n)$, and is statistically robust. moreover, the method makes intuitive sense. However,

the method has also received criticism in recent years as it tends to overestimate the feature importance of heavily correlated features [62].

8.2. SHAP values

In 2020, Lundberg et al. published set of tools for interpreting the results of tree-based machine learning models [15]. Notably, they presented TreeExplainer, a simple model to explain the prediction of a complex tree-based model. This is known as an explanation model. TreeExplainer augments conventional Python tree-based models with local and global explanations of predictions. The best explanation model for any model, is the model itself. However, as models get more complex, they rapidly become too complicated for a human to comprehend. An easier model to interpret the outputs and provide the user with more insight into the workings of a complex one become valuable.

TreeExplainer uses Shapley Additive exPlanation (SHAP) values, based on Shapley values. First proposed in 1953 by Lloyd Shapley, they have a solid theoretical foundation from cooperative game theory [63]. Shapley values are a method to fairly distribute the payout of a game among the participating parties, known as coalitions. The Shapley values measure the "importance" of each member of the coalition to determine the payout. For the development of SHAP, Lundberg et al. borrowed this idea and adapted it to determine a feature importance for an ML model.

SHAP values works with additive feature attributions: all feature attributions sum up to the model output. The explanation model $g(z)$ is a linear combination of the binary coalition variable $z \in \{0,1\}^M$ representing the input features, where M is the number of input features. The function $m_x(z) = x$ maps the coalition vector to the real input values. Explanation models often use simplified inputs x', especially in image processing to bundle pixels into meta or super pixels. However, as we are dealing with tabular data, in this section we will not use simplified features and just work with input vector x.

$$g(z) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i = f(m_x(z)) \tag{8.1}$$

During prediction all feature are "present", $z = \vec{1}$, this simplifies to

$$g(x) = \phi_0 + \sum_{i=1}^{M} \phi_i = f(x) \tag{8.2}$$

Now it become apparent that ϕ_i represents the contribution of the corresponding feature, x_i, to the final score of the model we are trying to interpret. This linearization of the model only holds locally, for this input vector x. The bias ϕ_0 is defined as the output of the model when none of the features are present, a coalition vector of $\vec{0}$. It is

determined by the expected value over all samples in the data set [64].

$$\phi_0 = f(m_x(\vec{0})) = \mathop{\mathbb{E}}_{X}\big(f(x)\big) \tag{8.3}$$

Lundberg et al. formulate three desirable properties that an additive feature attribution method should adhere to. Local accuracy states that the explanation model prediction should equal the output of the model f (Equation 8.2). The property of missingness says that missing features should have zero contribution assigned. This is to ensure that we have a unique solution, and in practice only used when a feature is constant for all samples. Consistency states that for two models with the same input features, if

$$f'(m_x(z)) - f'(m_x(z \setminus i)) \leq f(m_x(z)) - f(m_x(z \setminus i)) \;\; \forall z \tag{8.4}$$

$$\text{then} \;\; \phi_i(f', x) \leq \phi_i(f, x) \tag{8.5}$$

The classic properties described in the original paper by Shapley: linearity, dummy, and symmetry, follow from this property [32]. Additive feature attributions, adhering to the three desired properties leads to a unique solution. SHAP is the method proposed that adheres to these properties. Computing exact Shapley values is NP-hard [32]. By focusing on tree-based machine learning models, TreeExplainer can compute the SHAP values based on exact Shapley values in polynomial time [15].

Another additive feature attribution method is LIME [12]. It linearizes the input of a model in a similar fashion and was an inspiration for the SHAP values. LIME values are calculated by minimizing a function. Imposing certain constraints on that function enables LIME to be used to calculate the SHAP values. This is known as the Shapley kernel [32].

A limitation of SHAP values is the assumed additive attributions of variables inputted to a model. If the model is not additive, then the SHAP values may be misleading. The missingness property might be violated for our model. The permutation feature importance results suggest that three features are of no importance at all to our model. However, the missingness property states that a feature's contribution can only be zero is the feature is constant over the entire data set.

The mean absolute SHAP values attributed to each of the features in our model are plotted in Figure 3.3. Although SHAP values were developed to provide local explanations, taking the mean of the absolute SHAP values provides valuable insights into a feature's importance. The mean absolute value indicates how much the feature affects the prediction on average. For our research, each feature has different SHAP values for every class as the model assigns each class a probability. The feature importances were averaged over the classes. SHAP values have theoretical assurances of accuracy on consistency, it is claimed to give a better global importance score than other feature importances [15].

The SHAP values allow for both interesting and attractive visualizations. Decision plots can be used to visualize how a model reached a prediction. These are plotted

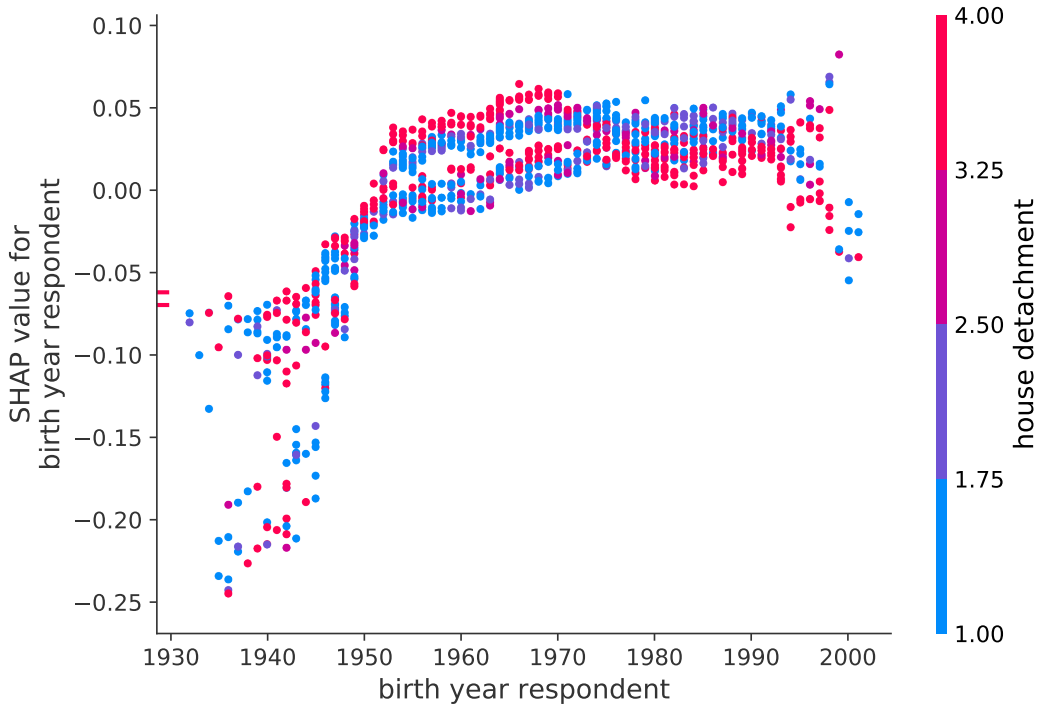

Figure 8.1.: A dependence plot for the feature birth year respondent, colored by the value for house detachment for the class expenditure risk.

in Figure 3.5, and Figure A.3. Dependence plots can be used to identify how pairs of features interact. This is plotted for our model in Figure 8.1. The SHAP value for one feature is plotted, and the data points are coloured using another feature. Before interpreting a model, a minimum requirement is that the model should be performing well. However, as stated in Section 3, as we do not consider this requirement to be met, we do not make any claims regarding these plots. The plots are there to demonstrate the capabilities of the method. SHAP values provide an extensive set of powerful tools to analyze tree based models.

Appendix

A. Additional plots

In Figure A.1, the absolute and approximated relative energy expenditure of households from each country are plotted. The y-axis depicts the frequency density: normalized frequency. The vertical red line indicates the median. The number of respondents is given after each country name.

In Figure A.2, the income distributions are plotted. The y-axis depicts the frequency density that is shared in absolute values by all plots in the row. The dashed black line is located at, ideally representative, frequency density of 10 %.

Four decision plots are shown in Figure A.3. We see two correctly classified households with very different SHAP values assigned to the features. The third decision path shows a household that the model confidently misclassifies as being in the expenditure risk group. The last plot shows a household where the probabilities for both classes are very close. There are two classes that are moving together, these are the classes corresponding to the two binary classification tasks that remain when the model splits based on income. What really matters for the classification is how the two relevant classes move relative to each other.

In Figure A.4, the confusion matrix to a theoretical classifier only splitting on income would result in. By the framework (Figure 2.1), a classifier would be able to split on the multi-class classification task in two binary ones. This greatly reduces the complexity of the modelling problem at hand.

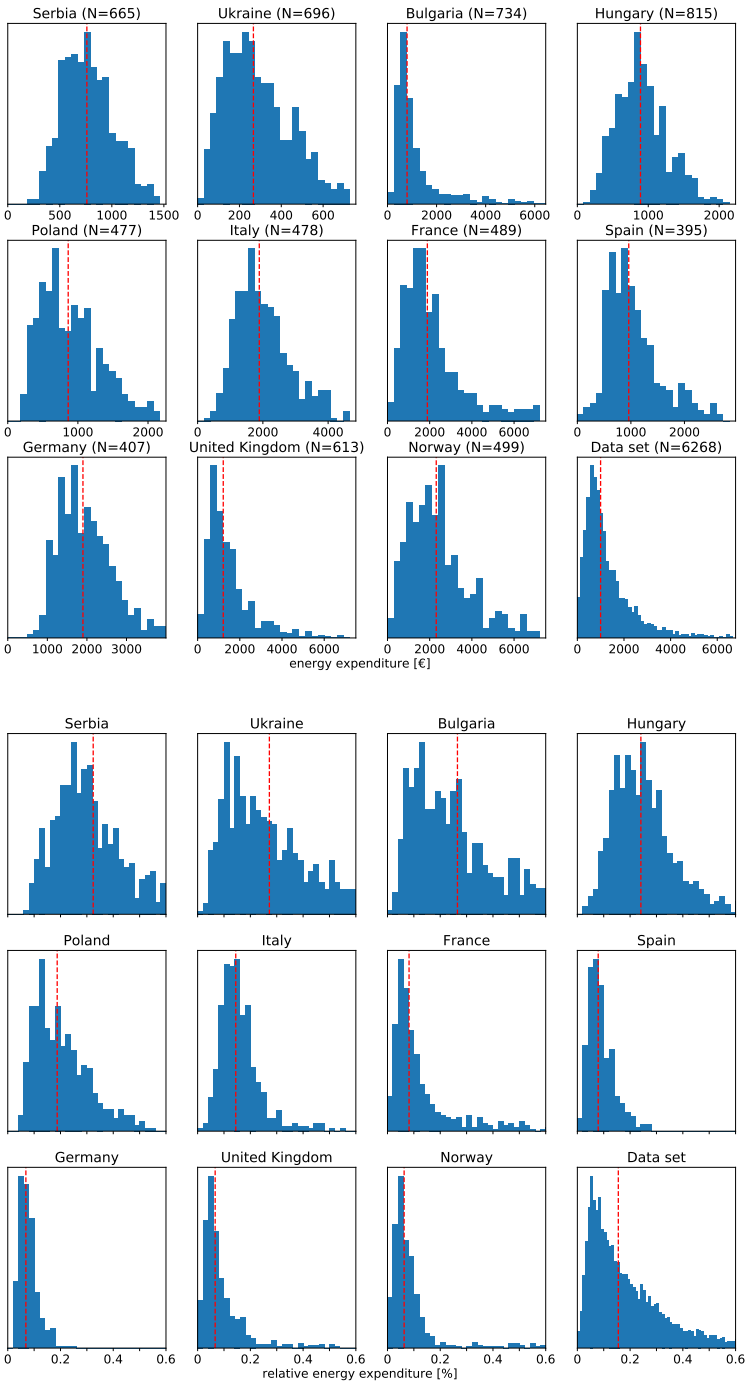

Figure A.1.: Absolute (top), and approximate relative energy expenditure (bottom) for all countries in the data set.

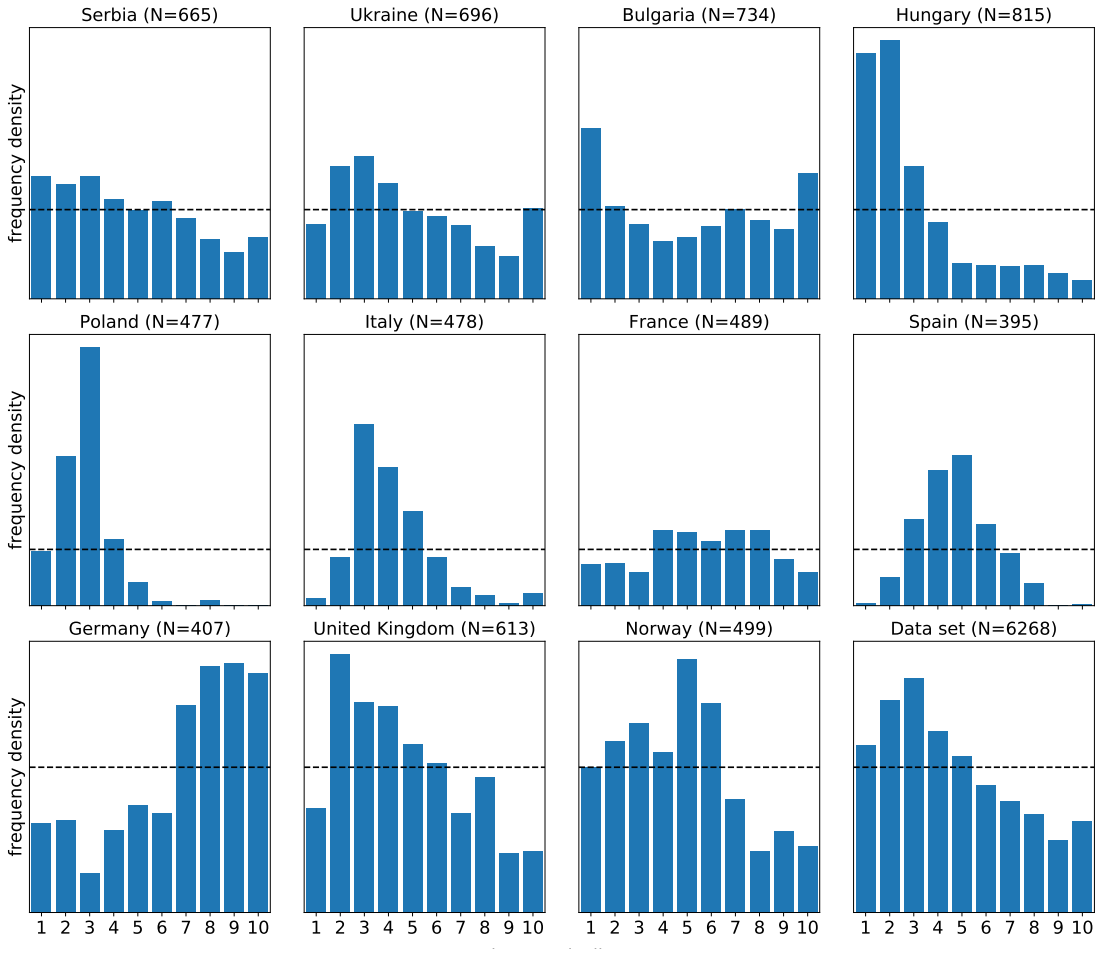

Figure A.2.: Income distributions of respondents per country.

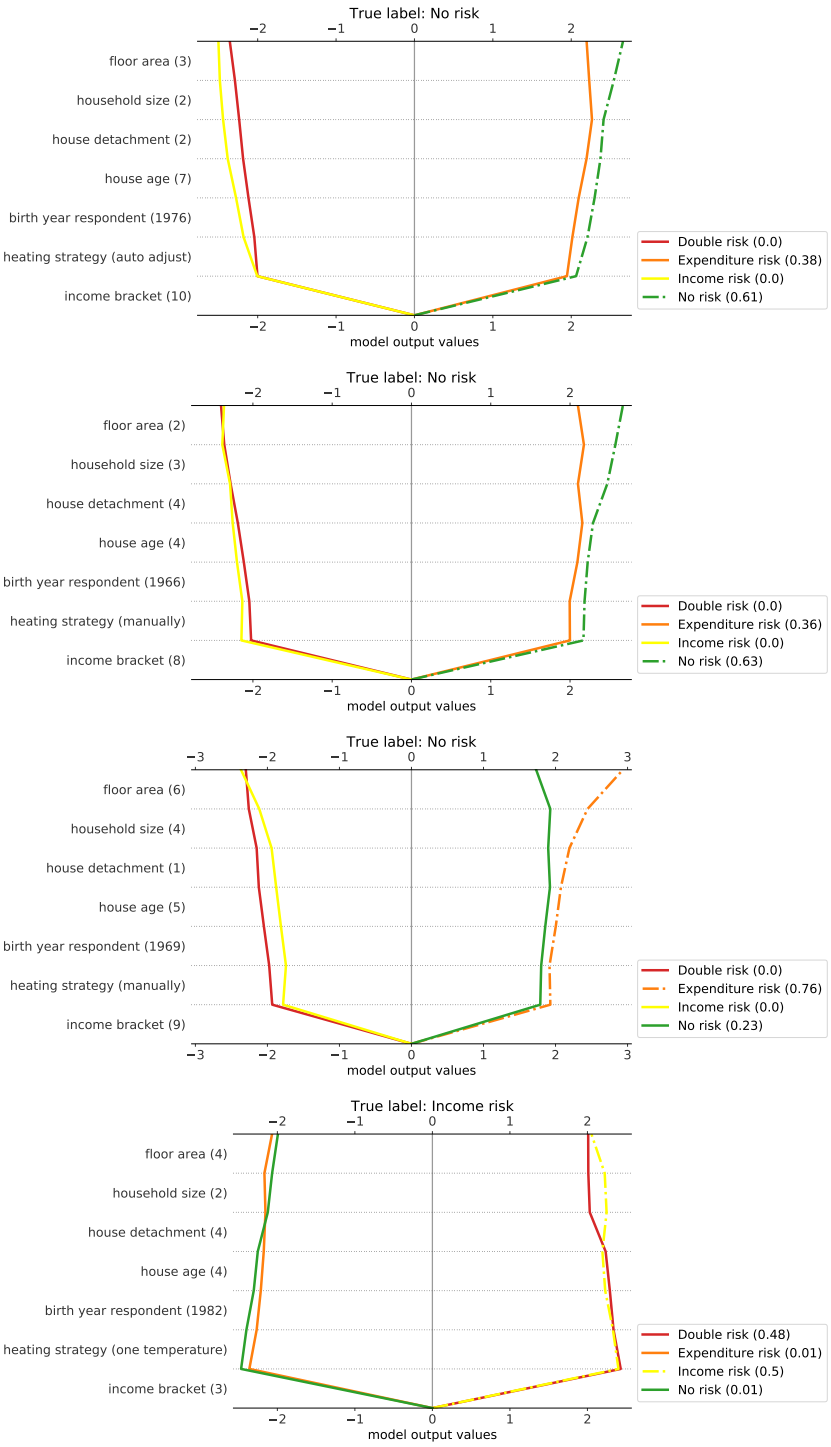

Figure A.3.: The decision plots of four additional samples from the test data set.

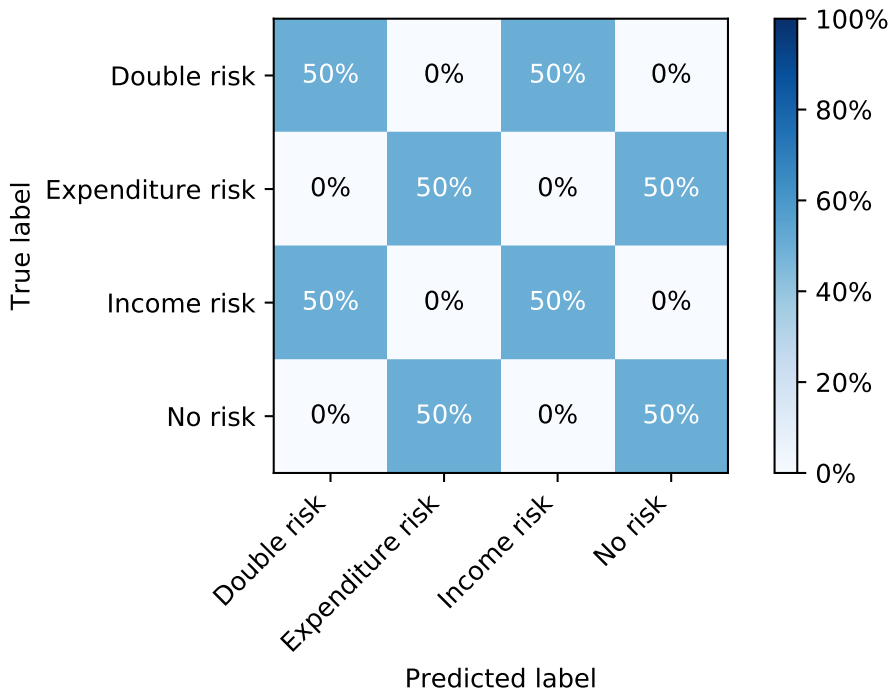

Figure A.4.: A theoretical confusion matrix that a model that only splits on income would achieve

B. Survey questionnaire [20]

GENERAL questions: to be asked in ALL countries

(Bulgaria, France, Germany, Hungary, Italy, Norway, Poland, Serbia, Spain, Ukraine, United Kingdom)

Section H - HOME / BUILDING CHARACTERISTICS AND HOUSEHOLD POSSESSIONS

H1. Which best describes your home?

Only ONE answer.

1. Single-family house detached from any other house
2. Single-family house attached to one or more other houses (for example: duplex, row or terraced house, or townhome)
3. Apartment in a building with 2 to 5 flats
4. Apartment in a building with 6 or more flats

H2. As far as you know, when was your home built?

Only ONE answer.

1. Before 1950
2. 1950 to 1959
3. 1960 to 1969
4. 1970 to 1979
5. 1980 to 1989
6. 1990 to 1999
7. 2000 to 2009
8. 2010 to 2016

99. (Don't know)

Instruction to the survey company: Please, use the answers with the relevant measurement system. Delete the unnecessary column.

H3. In which group does your home belong?

Only ONE answer.

1	Up to 42 m^2	1	Up to 455 ft^2
2	43 – 65 m^2	2	456 – 700 ft^2
3	66 – 90 m^2	3	701 – 970 ft^2
4	91 – 120 m^2	4	971 – 1295 ft^2
5	120 – 200 m^2	5	1296 – 2160 ft^2
6	More than 200 m^2	6	More than 2160 ft^2
7	Doesn't know/ didn't answer	7	Doesn't know/ didn't answer

H4. How many of the following vehicles your household owns?

One answer per row

		Don't have	Number of vehicles			(Don't know)
			1	2	3+	
A	Petrol car	1	2	3	4	99
B	Diesel car	1	2	3	4	99
C	Alternative fuelled car (methan, LPG)	1	2	3	4	99
D	Electric car	1	2	3	4	99
E	Hybrid car	1	2	3	4	99
F	Motorcycle (or Scooters)	1	2	3	4	99
G	Electric Motorcycle (or Scooter)	1	2	3	4	99
H	Van, truck, caravan	1	2	3	4	99
I	Bicycle	1	2	3	4	99
J	Electric bicycle	1	2	3	4	99

H5. Does your home have any of the following types of insulation?

Tick all that apply

1. Attic and/or roof insulation
2. Cavity wall insulation
3. External wall insulation
4. My home does not have any additional insulation.

99. (Don't know)

H6. What is the approximate percentage share of the energy sources you use for heating?

Indicate the approximate percentage share, based on the bills you paid

1. Electricity (including under floor heating)	……..%
2. District heating, different than using natural gas from a central source?	……..%
3. Natural gas from a central source / propane or bottled gas	……..%
4. Wood	……..%
5. Coal or coke	……..%
6. Pellets	……..%
7. Fuel oil	……..%
8. Waste/garbage	……..%
9. Biomass	……..%
10. Geothermal or air-source heat pump	……..%
11. Other source, please specify…….	……..%
99. Don't know	

H7. What was the cost of heating for your home for the last heating season? Indicate the cost per month or for the whole heating season, depending on how you pay your bills.

Fill only ONE of the answers, most suitable for you:

1. About …… [national currency] average per month	**Continue with the NEXT question**
2. About …… [national currency] for the whole heating season	**Skip the NEXT question**
99. Don't know	

H7A. Number of months, you pay for heating in the last heating season?

1. Number of months ……

99. (Don't know)

Instruction to the survey company: Use only one of the following two questions. If there is a country, where the two options are presented, ask both questions

H8A. What was the average monthly bill for electricity of your household over the last 12 months?

……. [National currency]

H8B. What was the last annual bill for electricity of your household?

……. [National currency]

H9. Which of the following best describes how your household controls your main heating equipment most of the time?

Only ONE answer.

1. Set one temperature and leave it there most of the time
2. Manually adjust the temperature (e.g. at night or when no one is at home)
3. Program the thermostat to automatically adjust the temperature during the day and night at certain times
4. Our household does not have control over the equipment

H10. Does your household use electricity or heating, generated by any of the following technologies, which are owned by you or by you and your neighbours/community?

Tick all that apply

1. Solar photovoltaic panels (PV) for generation of electricity and/or heat
2. Using biomass for generation of electricity and/or heat
3. Solar collectors for water heating
4. Geothermal or air-source heat pumps
5. None of the previous

H11. About how old are the most used electrical appliances in your home?

One answer per row. If you have more than one appliance of a given age, please answer for the most often used ones.

		Up to 3 years old	4-10 years old	Older than 10 years	Don't have	Don't know
A	Cooker (stove, oven, cooktops)	1	2	4	5	99
B	Dishwasher	1	2	4	5	99
C	Clothes washer / Washing machine (*Do not include community clothes washers that are located in the basement or laundry room of your apartment building*)	1	2	4	5	99
D	Refrigerator / freezer	1	2	4	5	99
E	Air conditioning units at your home	1	2	4	5	99
F	Portable electric heater(s)	1	2	4	5	99
G	Standalone electric water heater (boiler)	1	2	4	5	99
h	TV set / Home theater system	1	2	4	5	99

H12. What portion of the light bulbs inside your home are:

One answer per row

		All	Most	About half	Some	None	Don't know
A	Incandescent bulbs ("old" classic bulbs)	1	2	3	4	5	99
B	Energy efficient bulbs (e.g. LED, compact fluorescent bulbs or halogen bulbs)	1	2	3	4	5	99

H13. Does your home have any of the following "smart meters", which records energy consumption in real time and sends this information to your utility company and in some cases includes also a monitor to see (and control) your energy usage?

One answer per row.

	Yes	No	Don't know
Electricity smart meter	1	2	99
Gas smart meter	1	2	99
Heating smart meter	1	2	99
	Skip the NEXT question	**Continue with the**	

		NEXT question	

H14. What are the main reasons not to have a "smart meter" at you home?[20]

Tick all that apply.

1. Smart meters are still not adopted by the utility companies
2. Smart meters are adopted by the utility companies but they are not compulsory
3. The cost of smart meters is too high
4. Smart meters violate my privacy, sharing information about my consumption habits
5. The utility company could misuse the data from the smart meters
6. I don't know whether I can use smart meters at home
7. I heard that smart meters can be harmful to health
8. Other, please specify …….

H15. How much do you agree with the following statements?[21]

ONE answer per row

	Strongly disagree	Disagree	Agree	Strongly agree	Don't know
I am not willing to do anything about the environment if others don't do the same	1	2	3	4	99
Environmental impacts are frequently overstated	1	2	3	4	99
Environmental issues should be dealt with primarily by future generations	1	2	3	4	99
I am willing to make compromises in my current lifestyle for the benefit of the environment	1	2	3	4	99
Policies introduced by the government to address environmental issues should not cost me extra money	1	2	3	4	99
Environmental issues will be resolved in any case through technological progress	1	2	3	4	99
Protecting the environment is a means of stimulating economic growth	1	2	3	4	99

[20] Removed from the survey questionnaire in Norway as not relevant due to factual reasons – the government started a campaign for installing smart meters to all households by 2019.

[21] Even the question is in the General section, it is mandatory to be asked only in the countries covered by the "Mobility" and "Heating and cooling" sections. In the rest of the countries (Bulgaria, Serbia and the UK) it should be included, if possible.

B. Reducing the CO2 emissions from the industry and the building sector						
C. Increasing the share of energy, generated by RES						
D. Improving the energy efficiency of the residential sector						
E. Mitigate the effects of the climate change						
F. Lowering the energy intensity of the industry						

Section S - SOCIAL AND ECONOMIC CHARACTERISTICS

S1. How many women and men at the following ages, live in this household for at least 6 months of the year?

Indicate the number of people in each cell. If there are no people at the given age, write "0".

		Up to 18 year old	18-65 year old	Above 65 year old
A.	Women	_ _	_ _	_ _
B.	Men	_ _	_ _	_ _

S2. What is the highest level of studies, you have completed?

Only ONE answer.

1	No formal education or below primary
2	Primary education
3	Secondary and post-secondary non-tertiary education
4	Tertiary education first stage, i.e. bachelor or master
5	Tertiary education second stage (PhD)
9	(Don't know)

S3. What best describes your current employment status?

Only ONE answer.

1	Employed full-time
2	Employed part-time
3	Long time not employed (more than 3 months)
4	Retired / pensioner
5	Student
6	Other economically inactive person
99	(Don't know)

S4. What year were you born?

1. ……

99. (Don't know / refuse to answer)

S5. What is your gender?

Only ONE answer.

1. Male
2. Female

S6. Which phrase describes best the area where you live?

Only ONE answer.

1. A big city (more than 0,5 mln people)
2. The suburbs or outskirts of a big city
3. A town or a small city
4. A country village
5. A farm or home in the countryside
6. (Don't know)

S7. Has your household or any member of it received any financial aid from a public institution, which has helped you to pay your energy bills in the last 12 months (incl. so called social tariffs)?

Only ONE answer.

1. Yes -> for Ukraine ONLY: continue with the NEXT question
2. No -> for Ukraine ONLY: Skip the next question

Question to be asked ONLY in Ukraine

S7UA. What type of energy supplies are covered by the financial aid, received by you?

Tick all that apply

1. Gas supply
2. Electricity supply
3. Heat supply
4. Water supply
5. Other (please specify) ……

S8. Which of the descriptions bellow comes closest to how you feel about your household's income nowadays?

Only ONE answer.

1. Living comfortably on present income
2. Coping on present income
3. Finding it difficult on present income
4. Finding it very difficult on present income

99. (Don't know)

Instruction to the survey company: You can remain only one of the columns below ("per month" or "per year") if the people in the country calculate their income correspondingly.

S9. What was the average total monthly income of your household, after tax and compulsory deductions, from all sources, over the last 12 months? If you don't know the exact figure, please give an estimate.

Please, tick only ONE answer.

	Per month	Per year

1	Up to …. [national currency][42]	Up to …. [national currency]
2	…	…
3	…	…
4	…	…
5	…	…
6	…	…
7	…	…
8	…	…
9	…	…
10	Over … [national currency]	Over … [national currency]
98	Refused to answer	
99	(Don't know)	

Conclusion

[42] Deciles of the income as given by the national statistics

Bibliography

[1] Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. Harper, 2014.

[2] UK Parliament. Warm homes and energy conservation act, 2000. https://www.legislation.gov.uk/ukpga/2000/31/contents.

[3] Harriet Thomson, Carolyn Jane Snell, and Christine Liddell. Fuel poverty in the European Union: a concept in need of definition? *People, Place & Policy online*, pages 5–24, 2016.

[4] Pierre-Jean Coulon and Hernández Bataller. Opinion of the European Economic and Social Committee 'for coordinated European measures to prevent and combat energy poverty. *Official Journal of the European Union*, C 341:21–27, 2013.

[5] Benjamin K. Sovacool. Fuel poverty, affordability, and energy justice in England: Policy insights from the Warm Front Program. *Energy*, 93:361–371, 2015.

[6] Ryan Walker, Paul McKenzie, Christine Liddell, and Chris Morris. Area-based targeting of fuel poverty in Northern Ireland: An evidenced-based approach. *Applied Geography*, 34:639–649, 2012.

[7] Hossein Hassani, Mohammad Reza Yeganegi, Christina Beneki, Stephan Unger, and Mohammad Moradghaffari. Big Data and Energy Poverty Alleviation. *Big Data and Cognitive Computing*, 3(4):50, 2019.

[8] Raúl Castaño-Rosa, Jaime Solís-Guzmán, Carlos Rubio-Bellido, and Madelyn Marrero. Towards a multiple-indicator approach to energy poverty in the European Union: A review. *Energy and Buildings*, 193:36–48, 2019.

[9] Siddharth Sareen, Harriet Thomson, Sergio Tirado Herrero, João Pedro Gouveia, Ingmar Lippert, and Aleksandra Lis. European energy poverty metrics: Scales, prospects and limits. *Global Transitions*, 2:26–36, 2020.

[10] Nils J. Nilsson. *Principles of Artificial Intelligence*. Elsevier Inc, Morgan Kaufmann, 1982.

[11] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2014.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" : Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[13] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L 119:1–88, 2016.

[14] Lilian Edwards and Michael Veale. Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review*, 16(1):1–65, 2017.

[15] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[16] Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, Xiang Huang, Ying Xiao, Haosen Cao, Yanyan Chen, Tongxin Ren, Fang Wang, Yaru Xiao, Sufang Huang, Xi Tan, Niannian Huang, Bo Jiao, Cheng Cheng, Yong Zhang, Ailin Luo, Laurent Mombaerts, Junyang Jin, Zhiguo Cao, Shusheng Li, Hui Xu, and Ye Yuan. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2(5):283–288, 2020.

[17] Francesco Dalla Longa, Bart Sweerts, and Bob van der Zwaan. Exploring the Complex Origins of Energy Poverty in The Netherlands with Machine Learning. 2020. Manuscript submitted for publication.

[18] Enable-EU. D4.1 dataset for the comparative sociological analysis of the household survey results. http://www.enable-eu.com/wp-content/uploads/2019/10/enable_eu_dataset_households.zip. Accessed 01-04-2020.

[19] Enable-EU. The context. http://www.enable-eu.com/about-us/. Accessed 24-06-2020.

[20] Todor Galev, Alexander Gerganov Csd, and Thomas Pellerin-carlin. D4.1 Final report on comparative sociological analysis of the household survey results. Technical report, Enable-EU, 2018.

[21] José Carlos Romero, Pedro Linares, and Xiral López. The policy implications of energy poverty indicators. *Energy Policy*, 115:98–108, 2018.

[22] Jerome Harold Friedman. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, 29(2):1189–1232, 1999.

[23] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, pages 6638–6648, 2018.

[24] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[25] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: gradient boosting with categorical features support. In *NIPS 2017*, pages 1–7, 2018.

[26] CBS Statistics Netherlands. StatLine publicaties Kerncijfers wijken en buurten. https://www.cbs.nl/nl--nl/dossier/nederland--regiona, 2018. Accessed 21-07-2020.

[27] Christine Liddell and Chris Morris. Fuel poverty and human health: A review of recent evidence. *Energy Policy*, 38(6):2987–2997, 2010.

[28] Christine Liddell, Chris Morris, Harriet Thomson, and Ciara Guiney. Excess winter deaths in 30 European countries 1980–2013: a critical review of methods. *Journal of Public Health*, 38(4):806–814, 2015.

[29] Lefkothea Papada and Dimitris Kaliampakos. Measuring energy poverty in Greece. *Energy Policy*, 94:157–165, 2016.

[30] Carlo Andrea Bollino and Fabrizio Botti. Energy Poverty in Europe: A Multidimensional Approach. *PSL Quarterly Review*, 70(283):473–507, 2017.

[31] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[32] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[33] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018.

[34] John D Healy and J.Peter Clinch. Quantifying the severity of fuel poverty, its relationship with poor housing and reasons for non-investment in energy-saving measures in Ireland. *Energy Policy*, 32(2):207–220, 2004.

[35] Shonali Pachauri and Daniel Spreng. Measuring and monitoring energy poverty. *Energy Policy*, 39(12):7497–7504, 2011.

[36] Sabina Scarpellini, Pilar Rivera-Torres, Inés Suárez-Perales, and Alfonso Aranda-Usón. Analysis of energy poverty intensity from the perspective of the regional administration: Empirical evidence from households in southern Europe. *Energy Policy*, 86:729–738, 2015.

[37] Harriet Thomson, Stefan Bouzarovski, and Carolyn Snell. Rethinking the measurement of energy poverty in Europe: A critical analysis of indicators and data. *Indoor and Built Environment*, 26(7):879–901, 2017.

[38] Lilia Karpinska and Sławomir Śmiech. Invisible energy poverty? analysing housing costs in central and eastern Europe. *Energy Research & Social Science*, 70:101670, 2020.

[39] European Union. Energy Poverty Observatory, 2020. www.energypoverty.eu/.

[40] Guido van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[41] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[42] Xe.com Inc. Online foreign exchange tools and services company, 2020. www.xe.com. Accessed April, 2020.

[43] Eurostat. Distribution of income by quantiles - eu-silc and echp surveys (ilc_di01), 2018. data retrieved from Eurostat 06-2020, https://ec.europa.eu/eurostat/web/products-datasets/-/ILC_DI01.

[44] Brenda Boardman. *Fuel Poverty: From Cold Homes to Affordable Warmth Hardcover*. John Wiley & Sons Ltd, 1991.

[45] Lucie Middlemiss. A critical analysis of the new politics of fuel poverty in England. *Critical Social Policy*, 37(3):425–443, 2016.

[46] Asbjørn Rødseth and Steinar Holden. *Wage formation in Norway*. IIES, 1989.

[47] Marco Tufo. The minimum wage in italy during the eurozone crisis age and beyond. *IUSLabor. Revista d'anàlisi de Dret del Treball*, pages 205–231, 2018.

[48] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[49] Tianqi Chen. Introduction to Boosted Trees. *Data Mining with Decision Trees*, pages 187–213, 2014.

[50] Leo Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

[51] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, 2011.

[52] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

[53] Glenn De'ath and Katharina E. Fabricius. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000.

[54] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[55] X. Liu, J. Wu, and Z. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[56] Lawrence O. Hall, Kevin W. Bowyer, Philip W. Kegelmeyer, and Nitesh V. Chawla. SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Journal of Artificial Intelligence Research*, 2009(Sept. 28):321–357, 2006.

[57] Haibo He, Yang Bai, Edwardo Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 1322 – 1328, 2008.

[58] Thorvald Julius Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34, 1948.

[59] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[60] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[61] Carolin Strobl, Anne Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11, 2008.

[62] Giles Hooker and Lucas Mentch. Please Stop Permuting Features: An Explanation and Alternatives. *arXiv e-prints*, page arXiv:1905.03151, 2019.

[63] Lloyd Stowell Shapley. A value for n-person games. In Harold William Kuhn and Albert William Tucker, editors, *Contributions to the Theory of Games (AM-28), Volume II*, chapter 17, pages 307–318. Princeton University Press, 1953.

[64] Christoph Molnar. *Interpretable machine learning: a guide for making Black Box Models interpretable*. Lulu, Morisville, North Carolina, 2020.