universität wien

Masterarbeit / Master's Thesis

Titel der Masterarbeit / Title of the Master's Thesis

Pairwise Evaluation of Explanation Methods for Black-Box Models on Tabular Data

verfasst von / submitted by

B.Sc. Tobias Wetzel

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (M.Sc.)

Wien, 2025 / Vienna, 2025

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 926
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Wirtschaftsinformatik
Betreut von / Supervisor:	Assoz. Prof. Dipl.-Ing. Dr.techn. Sebastian Tschiatschek, BSc
Mitbetreut von / Co-Supervisor:	Timothée Schmude, M.A.

Abstract

Advances in black-box model performance have increased interest in understanding their internal mechanisms, particularly as these models demonstrate exceptional predictive capabilities across critical domains. While numerous explanation methods exist, research has not systematically measured their combined effects, creating a knowledge gap regarding the potential benefits of integrating multiple techniques. This thesis addresses this gap through a systematic evaluation of established explanation methods—including SHAP, LIME, and Partial Dependence Plots—and three novel approaches, with a specific focus on models processing tabular data.

The research methodology combines synthetic data experiments with a real-world user study to examine whether different explanation methods vary in subjective and objective understanding and if their combination enhances the overall understanding of the model. The user study, involving 129 participants, was designed to quantitatively and qualitatively assess the understanding of the explanation method and the underlying model. The assumptions of the synthetic data evaluation were specifically designed to test the characteristics of the underlying data that each explanation method was intended to uncover.

The synthetic data evaluation validated key assumptions of the novel approaches. At various breakpoints in the training process, SHAP values are embedded into a lower-dimensional space using UMAP (Uniform Manifold Approximation and Projection), a non-linear dimensionality reduction technique for high-dimensional data visualization. Increased model complexity correlates with larger shifts in embedding positions across these breakpoints, and distinct clusters that highlight differences among subgroups are observed. Frequent Pattern Mining indicated that features with strong interactions tend to have high Lift values—a metric that quantifies how much more frequently two features occur together compared to what would be expected by chance—and that Feature Context Embeddings, which generate vector representations of features based on the contexts in which they are used, effectively group related features. The findings of the synthetic data evaluation demonstrate that the novel explanation approaches are capable of capturing properties of the underlying data, which establishes a practical basis for real-world applications.

Results from the user study demonstrate that extending existing explanation techniques, most notably by using UMAP to enhance SHAP to better expose model learning patterns, produces statistically significant improvements in subjective understanding and feature importance interpretation. The findings indicate that the UMAP-based SHAP extension was well understood subjectively, although objective understanding scores were lower than for PDP and standard SHAP, likely due to varying abstraction depths. Education significantly impacted both types of understanding, with higher education correlating with better subjective understanding but not with improved objective performance. Notably, the enhanced understanding of the model, generated by combining two explanations, could be reliably predicted by summing their individual contributions, and the results suggest a potential diminishing effect when integrating different approaches, although this effect was not statistically significant at the 5% level. The results of the user study demonstrate that combining two explanation method does improve overall understanding of the model.

Kurzfassung

Fortschritte in der Leistung von Black-Box-Modellen haben das Interesse an einem tieferen Verständnis ihrer internen Mechanismen gesteigert, insbesondere da diese Modelle in kritischen Bereichen außergewöhnliche Vorhersagefähigkeiten demonstrieren. Obwohl zahlreiche Erklärungsverfahren existieren, wurde deren kombinierte Wirkung bisher nicht systematisch untersucht, was eine Wissenslücke hinsichtlich der potenziellen Vorteile der Integration mehrerer Techniken schafft. Diese Arbeit schließt diese Lücke durch eine systematische Evaluierung etablierter Erklärungsverfahren – darunter SHAP, LIME und Partial Dependence Plots – sowie drei neuartige Ansätze, wobei ein besonderer Schwerpunkt auf Modellen liegt, die tabellarische Daten verarbeiten.
Die Arbeit kombiniert eine Analyse anhand synthetischer Daten mit einer realen Nutzerstudie, um zu untersuchen, ob verschiedene Erklärmethoden in Bezug auf das subjektive und objektive Verständnis variieren und ob ihre Kombination das Gesamtverständnis des Modells verbessert. Die Nutzerstudie mit 129 Teilnehmern wurde so konzipiert, dass sie das Verständnis der Erklärmethode sowie des zugrunde liegenden Modells sowohl quantitativ als auch qualitativ bewertet. Die Annahmen für die Evaluierung synthetischer Daten wurden speziell entwickelt, um die Eigenschaften der zugrunde liegenden Daten zu testen, die jedes Erklärungsverfahren aufdecken soll.
Die Evaluierung mit synthetischen Daten bestätigte zentrale Annahmen der neuartigen Ansätze. An verschiedenen Breakpoints im Trainingsprozess wurden die SHAP-Werte in einen Embedding-Raum mittels UMAP (Uniform Manifold Approximation and Projection) projiziert – einer nichtlinearen Dimensionsreduktionsmethode zur Visualisierung hochdimensionaler Daten. Eine zunehmende Modellkomplexität korrelierte mit stärkeren Verschiebungen in den Embedding-Positionen an diesen Breakpoints, und es bildeten sich Cluster, die Unterschiede zwischen den Subgruppen hervorhoben. Frequent Pattern Mining ergab, dass Merkmale mit starken Interaktionseffekten tendenziell hohe Lift-Werte aufweisen – ein Maß dafür, wie viel häufiger zwei Merkmale gemeinsam auftreten als zufällig erwartet – und dass Feature Context Embeddings, die Vektorrepräsentationen der Merkmale basierend auf dem Kontext ihrer Verwendung erzeugen, verwandte Merkmale effektiv gruppierten. Die Ergebnisse der Evaluierung mit synthetischen Daten zeigen, dass die neuartigen Erklärungsansätze in der Lage sind, Eigenschaften der zugrunde liegenden Daten zu erfassen und somit eine Grundlage für ihre Anwendbarkeit in realen Szenarien zu schaffen.
Die Ergebnisse der Nutzerstudie zeigen, dass die Erweiterung bestehender Erklärungsverfahren – insbesondere durch den Einsatz von UMAP zur Verbesserung von SHAP, um die Lernmuster des Modells besser sichtbar zu machen – statistisch signifikante Verbesserungen im subjektiven Verständnis und in der Interpretation der Merkmalswichtigkeit bewirkt. Die Ergebnisse deuten darauf hin, dass die auf UMAP basierende SHAP-Erweiterung subjektiv gut verstanden wurde, wohingegen das objektive Verständnis niedriger ausfiel als bei PDP und SHAP, was vermutlich auf unterschiedliche Abstraktionstiefen zurückzuführen ist. Die Bildung hatte einen signifikanten Einfluss auf beide Verständnistypen, wobei ein höherer Bildungsgrad mit einem besseren subjektiven Verständnis korrelierte, jedoch nicht mit einer verbesserten objektiven Leistung. Das durch die Kombination zweier Erklärungen entstandene Modellverständnis konnte zuverlässig durch die Summe ihrer Einzelbeiträge vorhergesagt werden, und die Ergebnisse deuten auf einen möglichen abnehmenden Effekt bei der Integration unterschiedlicher Ansätze hin, obwohl dieser Effekt auf dem 5%-Niveau statistisch nicht signifikant war. Insgesamt zeigen die Ergebnisse der Nutzerstudie, dass die Kombination von zwei Erklärungsverfahren das Gesamtverständnis des Modells verbessert.

Contents

Chapter 1

Introduction

The annual increase in data is enormous. Although the exact figures can vary depending on the source and measurement method, it is generally assumed that the volume of data generated and stored increases by around 25 to 30% each year. Along with this continuous growth in data, there is also a growing trend towards ever larger and more complex machine learning models in order to make the best possible use of this data. The noteworthy progressions in machine learning algorithms, specifically with the introduction of Random Forests and Deep Neural Networks, have led to a new era of outstanding predictive performance.

Well-known platforms such as OpenAI, Google and Amazon operate models with billions of parameters and the areas of application in which machine learning models can be used are constantly becoming more extensive. Prominent areas of application include medicine, where models can be used to diagnose certain diseases or develop personalized medicine, biology, where machine learning models can be used to analyze genomes and their function, or chemistry, where they can be applied to develop lighter, more conductive or more elastic materials. The list of areas of application is long. Chui et al. [48] list 19 industries in which machine learning models are already being used. In dermatology, for example, convolutional neural networks have been used to accurately detect skin cancer with high accuracy[57]. Saliency Maps have been used to highlight areas, which were effected by the cancer [23]. Mertes et al.[49] found that applying Counterfactuals to explain diagnoses can significantly increase trust, satisfaction with the explanation and prediction accuracy when asked to predict the network's prediction.

In many of these applications, it is beneficial or even essential to understand models, either in part or exhaustively, to draw meaningful conclusions from their use and ensure they perform their assigned tasks effectively. However, measuring explainability remains a complex challenge. Additionally, the depth of understanding required varies depending on the context. High-stakes decisions, such as whether to proceed with surgery, demand a far deeper understanding than models with no direct impact on individuals' well-being.

In certain cases, understanding the decision-making process is not just desirable but legally required. Recent regulatory changes, particularly under the General Data Protection Regulation (GDPR) [73], explicitly state that "data subjects should have the right to [...] obtain an explanation of the decision, and to challenge the decision" and "shall have the right to obtain meaningful information about the logic involved" [24].

Although evaluating explanation methods that clarify model behavior is an active area of research, little attention has been given to the combined effects of multiple explanation methods on model interpretability. Moreover, widely used explanation methods such as SHAP, LIME, and Partial Dependence Plots each provide insights into different aspects of a model but only capture a single perspective and often to comprehensively explain increasingly complex models.

This thesis aims to develop and identify explanation methods—and their combinations—that are particularly helpful in understanding machine learning models for tabular data. It addresses four main research questions through synthetic data analysis and a user study:

- **RQ1:** Can the new explanation approaches depict specific model properties effectively? In particular, do UMAP-based SHAP embeddings accurately reflect variability in feature importance projections and form distinct clusters? Do features with strong interaction effects, as revealed by Frequent Pattern Mining, exhibit higher Lift values? And do Feature Context Embeddings, which create vector representations of features based on the context in which they are used, group related features effectively?

- **RQ2:** Are there differences in the subjective and objective understanding of the explanation methods when various influencing factors are taken into account?
- **RQ3:** To what extent does the perceived usefulness of individual explanation methods affect the overall added value when multiple explanation methods are combined? Does the relative usefulness of one explanation method influence this effect?
- **RQ4:** Does a higher subjective and objective understanding of the explanation methods result in more accurate inferences based on the provided explanations?

The remainder of this thesis is structured as follows. Chapter 2 discusses the desirable properties and various approaches for evaluating and classifying explanation methods. Chapter 3 presents the explanation methods, including three novel approaches. Chapter 4 outlines two evaluation strategies—one based on synthetic data and another on a real-world user study—to assess these methods. Chapter 5 demonstrates which model properties can be successfully recovered using the novel explanation methods on synthetic data and presents the findings from the user study. Finally, Chapter 6 concludes the thesis with a summary of the work, key contributions and findings, limitations of the work and future research directions.

Chapter 2

Related Work

Numerous methods have been developed to explain the decisions of machine learning models, both holistically and in terms of the contributions of individual features or specific decisions. However, not all models require such explanations due to their inherent simplicity. A key distinction can be made between models that are inherently interpretable due to their straightforward structure or built-in transparency, and those whose structure and operation are complex and therefore difficult to interpret.
Murdoch et al. [56] argue that a model should have several properties in order to be inherently interpretable. According to the authors, an inherently interpretable model should have the smallest possible number of parameters (sparsity), the ability to reproduce the decision process itself (simulatability), the ability to interpret individual aspects of the model separately (modularity), and the ability to consider the interaction of individual features in combination, either through meaningful indices such as BMI or price-earnings ratio (domain-based feature engineering) or through dimensionality reduction methods (model-based feature engineering). Other work [41] does not emphasize sparsity but instead introduces algorithmic transparency, which demands clarity in the procedures and training processes behind the model.
Inherently interpretable models include linear regression, in which the individual characteristics are assigned coefficients that directly measure the influence of the characteristics or decision trees, which represent a series of yes-no decisions that are easy to understand and visualize. Each node in the tree represents a feature, and the branches to the child nodes represent the possible answers to the question posed by the node.
However, many modern machine learning models, such as deep neural networks, random forests, and gradient-boosted trees, are not inherently interpretable due to their complexity and the interactions between numerous parameters and therefore require explanation methods to provide insights into their decision-making processes. These methods are often referred to as black-box models.
Depending on the use-case either black-box or inherently interpretable model might be more applicable.
Some researchers argue that "for any given task, the set of almost-equally performant models typically includes at least one simple and explainable model" [83]. They suggest that practitioners should prioritize interpretable models [69] and consider additional factors that may influence a model's success beyond accuracy alone, such as resource availability and risk [19]. Others challenge the notion of a strict trade-off between accuracy and explainability, and find no significant difference in explainability between black-box and interpretable models and claim that black-box models can often be both the most accurate and the most explainable models to end users [9]. Holliday et al. [30] on the other hand came to the conclusion that explanations did lead to an increase in user trust in machine learning algorithms. In cases where researchers opt for a black-box model and subsequently employ explanation methods for deeper understanding, those methods should yield clear and actionable insights.

2.1 Desiderata for Explanation Methods

To evaluate whether an explanation method is genuinely effective, clear objectives have to be established that a useful explanation should meet. This section outlines several desiderata that will serve as criteria for assessing the quality of the selected explanation methods.

2.1.1 Explanation Relevance

Nauta et al. [58] define twelve explanation properties (the so-called Co-12 properties). The properties are grouped into three categories. The content category assesses the intrinsic quality of the explanation, and properties belonging to this group include correctness (how accurately the explanation reflects the behavior of the underlying system), completeness (the extent to which the explanation covers all aspects of the system's behavior), consistency (the degree to which the explanation method yields the same result for identical inputs), continuity (how smoothly the explanation changes with small variations in input), contrastivity (the capability of the explanation to distinguish between different outcomes), and covariate complexity (the level of intricacy in feature interactions within the explanation). The second category, the presentation category, focuses on how the explanation is delivered and consists of compactness (the succinctness of the explanation), composition (the organization and structure of the explanation), and confidence (the clarity and accuracy of the probability or certainty information provided). The third category, the user category, evaluates how well the explanation meets the needs of its audience. Properties in this group include context (how relevant the explanation is to the user's specific situation), coherence (the extent to which the explanation aligns with the user's existing knowledge and beliefs), and controllability (the degree to which the user can interact with or influence the explanation).
Another categorization of the desiderata is provided by Zhou et al. [85], who divided the desiderata of explainability into interpretability and fidelity. Interpretability is defined as the capacity to provide a comprehensible explanation to a human audience. The authors identified three key aspects of interpretability: clarity, meaning that the explanation is unambiguous; parsimony, indicating that the explanation is simple and concise; and broadness, describing the general applicability of the explanation methods. Fidelity is further subdivided into completeness and soundness. Completeness refers to the extent to which the explanation accounts for the entire model, while soundness refers to the correctness of the explanation.
According to Murdoch et al. [56], three aspects should be considered when choosing a suitable explanation method. First, the explanation method should have a high degree of predictive accuracy in order to be able to make predictions by approximating the relationships in the data. If the underlying model is unable to express these relationships, the explanation method will hardly allow insightful conclusions about the actual relationships in the data. Second, the explanation method should maximize descriptive accuracy by effectively capturing the relationships learned from complex models and deciphering non-linear relationships between variables. Third, the explanation method must be relevant to the target audience [63]. Perhaps the information is useful to statisticians while providing little useful information to the people affected by an algorithm's decisions. who are affected by the decisions of an algorithm. Buschek et al. [15] argue that relevance is based on mindset, engagement and knowledge outcomes. Murdoch et al. further argue that increased relevance might be introduced by using a novel form of output and show the model and data from different perspectives.

2.1.2 Explanation User Groups

Langer et al. [40] as well as Arrieta et al. [8] determine relevance based on five different stakeholder groups. The authors distinguish between users, developers, affected parties, deployers, and regulators—each of whom has distinct expectations for explainability. Developers, for instance, might focus on explainability to enhance debugging and verification processes, while regulators are more concerned with fairness, accountability, and legal compliance. Mohseni et al. [52] broadly distinguish between AI Novices, Data Experts, and AI Experts. AI Novices are defined as users who interact with AI products daily without possessing a deep understanding of machine learning systems. The authors identify four desiderata from the perspective of AI Novices: (1) algorithmic transparency, which enables them to construct a mental model of the system; (2) trust and reliance, which are critical for applications such as recommendation systems and autonomous systems; (3) bias mitigation, with examples including criminal risk assessment and insurance rate prediction; and (4) data privacy, which is particularly relevant in personalized advertising contexts. In contrast, Data Experts primarily focus on model visualization and inspection as well as model tuning and selection whereas AI Experts are concerned with model interpretability and model debugging.
Langer et al. [40] further highlight that the satisfaction of desiderata involves two key dimensions: epistemic and substantial satisfaction. Epistemic satisfaction is achieved when stakeholders can evaluate and understand whether an AI system meets their needs, such as assessing fairness or transparency. Substantial satisfaction, on the other hand, is achieved when the system inherently exhibits the desired attributes, such as genuine fairness, transparency, or usability. The key dis-

tinction, therefore, is that epistemic satisfaction is tied to perception, while substantial satisfaction concerns the system's actual characteristics.

2.2 Classification of Explanation Methods

The explanation methods used in this work can be categorized based on different criteria, depending on their scope, applicability, and how they interact with the underlying model. Two fundamental classification schemes distinguish between model-specific and model-agnostic methods, as well as local and global approaches.

2.2.1 Model-specific and Model-agnostic explanation methods

Explanation methods can broadly be categorized into model-specific and model-agnostic approaches, depending on whether they are tailored to a particular model architecture or can be applied universally across different models [40].
Model-specific methods are adapted to the structure and function of the respective model, which may be due to the peculiarities of the model or to the different areas of application. In image generation, methods such as Grad-CAM are used to emphasize image areas that were relevant in the decision-making process of a neural network, whereas other methods identify abstract features that are most relevant for a specific outcome. In text generation, attention maps can be used to visualize which parts of the input text were given particular attention when generating the output text. Model-specific approaches have also been developed for other types of models, such as those handling structured or sequential data.

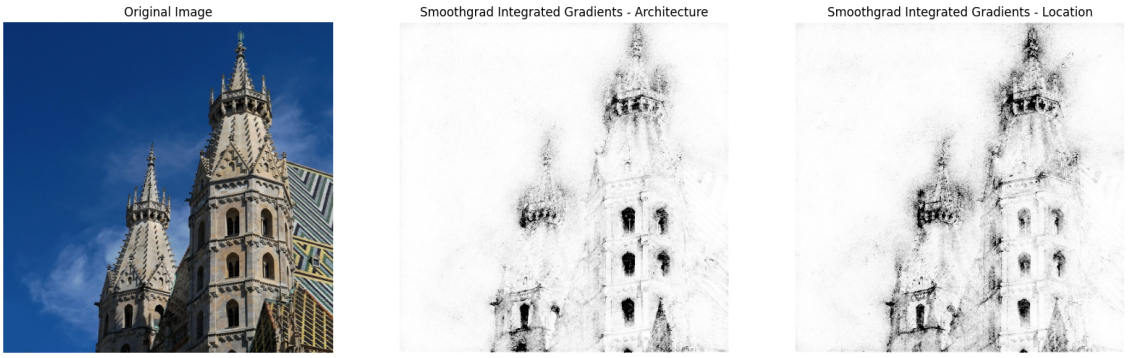


Figure 2.1: Left: Original image of the St. Stephen's Cathedral in Vienna. Center: Saliency map produced by computing the gradient with respect to the architectural style. Right: Saliency map computed with respect to the gradient discerning indoor versus outdoor characteristics of the image. The underlying model is an adapted version of the VGG architecture, using a multi-task output, trained on a synthetic dataset.

Model-agnostic methods are not tied to a specific model and can therefore be used in various areas of application. For the classic application area of regression and classification-based evaluation of tabular data, methods such as SHAP or LIME are used to visualize the influence of individual characteristics on the prediction[11]. Other techniques like Permutation Feature Importance[4] and Partial Dependence Plots[25] offer complementary insights by quantifying the overall effect of features on model outputs.

2.2.2 Local and global explanation methods

Another fundamental way to categorize explanation methods is by distinguishing between *local* and *global* approaches [40].

Global explanation methods aim to provide insights into the overall behavior of a model. Notable examples include Feature Importance, which quantifies the contribution of each feature to model predictions, Partial Dependence Plots, and Global Surrogate Models.

In contrast, *local* explanation methods focus on explaining the model's output for specific instances, either real or representative. Examples of *local* approaches include LIME (Local Interpretable

Model-Agnostic Explanations) [66], which constructs a linear model around a point of interest by considering similar instances, Counterfactuals [82], which illustrate how the model's output would change if certain input features were modified, and Anchors [67], which define rule-based conditions under which other input variations do not affect the prediction.

2.2.3 Other classification approaches

Apart from the applicability and scope of explanation methods, other dimensions exist by which explanations can be classified. One such dimension is the functioning-based approach [8][71][79], which categorizes explanation methods according to how they extract information from the model. Speith [79] identifies five categories within the functioning-based approach. The first category, local perturbations, involves slightly altering input values to determine the importance of features—LIME being an example of this approach. The second category, leveraging structure, makes use of the model's inherent architecture and closely relates to the model-specific explanation methods discussed in subsection 2.2.1. For instance, two of the three novel explanation methods introduced in section 3.5 and section 3.6 exploit the disentangled nature of decision trees to extract information. The third category, meta-explanation, does not operate directly on the machine learning model but instead works with explanations that have already been generated for that model; the novel approach presented in section 3.2 falls into this category. The fourth category, architecture modification, involves altering the model's structure to improve interpretability, as seen in Self-Explaining Neural Networks [6]. Finally, the fifth category is the example-based approach, which creates examples that meet certain criteria, with Anchors being a notable example by generating if-then rules.
Another classification is based on the result produced by an explanation [79][44], which groups explanations into feature importance, surrogate models, and example-based explanations. Additionally, [41] identifies text explanations, where a language model is used to generate a natural language description of the model's behavior.
Mohseni et al. [52] group the explanations into six categories, "How Explanations", which holistically show how the model works, "Why Explanations", which demonstrate why a prediction was made, "Why-Not Explanations", which describe why a model did not arrive at a specific conclusion, "What-If Explanations", which investigate how changes would affect the model output, "How-to Explanations", which describe how to arrive at a different conclusion, and "What-Else Explanations", which show other examples that yield similar model outputs.

2.3 Evaluating Post-Hoc Explanation Methods

Explanation methods that are not intrinsic are typically referred to as post-hoc explanation methods [53]. The following section reviews various evaluation techniques for these post-hoc methods, in order to assess the chosen explanation approaches and verify that they meet the desiderata outlined in section 2.1.

According to Gilpin et al. [28], '[a]n explanation can be evaluated in two ways: according to its interpretability and according to its completeness'. The authors define completeness as the degree to which an explanation allows conclusions to be drawn about other situations, while interpretability is defined as a set of 'descriptions that are simple enough to be understood by a person using a vocabulary that makes sense to the user', using understandable terms[22], either in visual representations or in a vocabulary in the sense of words.

2.3.1 Evaluation Metrics and Nested Frameworks

Mohseni et al. [52] identify five different measurable aspects: the mental model of the system, explanation usefulness and satisfaction, user trust and reliance, human-AI performance, and computational measures. The first aspect, the mental model, can be evaluated in three ways: by assessing participants' understanding of the model using subjective instruments (free-text questions and Likert scales), by posing questions about the model's output, and by investigating model errors. User satisfaction and usefulness can be measured similarly—through subjective feedback methods (again using free-text questions, Likert scales, and self-reports) and by case studies involving domain and AI experts—while also considering engagement with the explanation, task duration, and cognitive load. Trust and reliance can be evaluated using these subjective methods as well. Although there are no direct measures of trust, sustained use of explanations is taken

as an indication. Performance can be quantified both in terms of the model's accuracy and the users' ability to perform tasks or detect errors. Finally, computational methods can be applied, including simulation studies (where synthetic data is generated to test whether the explanation method captures key data aspects), sanity checks (such as input variance tests to see how slight image transformations affect saliency maps), and comparative evaluations—as done in the study by Samek et al. [70], where saliency maps are compared by flipping the most salient pixels to observe the resulting performance drop.
In addition, the authors develop a design and evaluation framework intended to help determine which evaluation methods are appropriate at each stage of building an explainable system. The framework is organized as a nested model with three layers that guide the design and evaluation of intelligible machine learning systems: At the outer layer, the overall system goals are defined, determining what should be explained and why, and these goals are linked to evaluation criteria that capture user needs, regulatory requirements, and desired outcomes. These goals are based on the points addressed in section 2.1. The middle layer focuses on the explanation interface. At this stage, an explanation interface is created and evaluated to determine whether the explanation is understood, whether users are satisfied with it, and whether they can develop a mental model of the system. The previously mentioned evaluation methods are employed here. At the innermost layer, the focus is on interpretable algorithms, which requires evaluating whether inherently interpretable models or post-hoc explanation techniques should be chosen based on assessing the model's trustworthiness and the fidelity of the ad-hoc explainer.

2.3.2 Alternative Categorization and Qualitative Evaluations

Nauta et al. [58] list a wide range of different evaluation methods that are mapped to the Co-12 properties listed in section 2.1. For instance, correctness can be assessed using synthetic data. Here, data is created in a way which would force the model to have a particular structure, the explanation method should then successfully recover this information. Completeness can be assessed by giving the input of the explanation method to the model. If the explanation is complete, the model should arrive at the same conclusion as if it had had the original input. Similar to the evaluation done by Samek et al. [70], completeness can also be assessed by removing the whole explanation rather than individual pixels. If all features are deleted at once, the accuracy should drop significantly. A third way of assessing the completeness can be achieved by comparing the output of the explanation to the output of the model being explained. Surrogate Models (described in chapter 3) are particularly straightforward to evaluate in this regard, as the coefficient of determination directly measures how much the Surrogate Model reveals about the underlying model. The authors additionally list the Kullback-Leibler divergence or correlation as assessment options for completeness. Continuity can be measured using perturbations. Here, small adjustments are made to the input data, which should then be reflected by minor changes in the output of the explanation. Similarly, Alvarez-Melis et al.[5] propose Local Lipschitz estimates which measures the stability of neighboring points around a point of interest by adding local perturbations. A high Lipschitz constant indicates that small changes in input can lead to large variations in the explanation, which is an indication of instability. Other properties among the Co-12, namely Consistency, Compactness, Composition, Confidence can be assessed qualitatively, as they are based on design choices of the explanation method.
[85] and [22] group the evaluation of explanation methods in three different categories. On one side, there are human-centered evaluations, which are either based on a real-world application, focusing on end-users or human-grounded, which involve experiments with lay people. On the other side, there are functionality-grounded evaluations, which evaluate an explanation method based on a formal definition. According to the authors, the depth of a decision tree would be one such metric. Doshi-Velez et al.[22] more generally list sparsity as one metric which functionality-grounded explanations can be evaluated with. The paper furthermore introduces the idea of constructing a matrix (as seen in Figure 2.2) consisting of the domain as one dimension and the methods as the other. The authors hypothesize that due to the high dimensionality of the different domains where machine learning models might be employed there could be a lower dimensional latent domain space where different explanation methods are needed. The authors list global/local models, time constraints and severity of incompleteness and user expertise as potential dimensions in this latent space and additionally argue that the cognitive work associated with understanding and applying each explanation method can be split into what the authors refer to as "chunks". These cognitive chunks can be differentiated with respect to the object being explained (e.g. feature values, individual instances), the number of cognitive chunks a method consists of and the compositionality of the cognitive chunks (if one cognitive chunk depends on another). The authors argue that this

matrix can be used to find suitable explanation methods if the characteristics of a new domain are similar to that of a domain where explanation methods have already been evaluated.

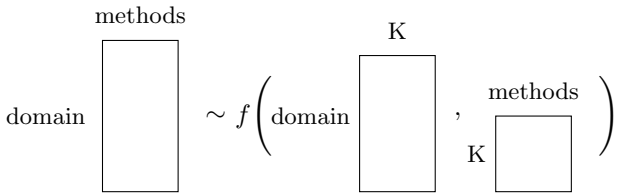


Figure 2.2: Domain-Method Matrix as proposed by Doshi-Velez et al.

Argawal et al.[1] list several metrics which can specifically be applied to feature attribution methods which measure the agreement between post-hoc explanations and ground truth explanations by comparing feature importance, rank and direction (coefficient of the effect).
Munoz et al.[55] use the spread among the importance of features as a metric for complexity. Few features with high importance are less complex to understand than many features with relatively balanced importance. They further introduce the α-Feature-Importance, which is the proportion of the features that is required to capture at least $\alpha \times 100\%$ of the total feature importance of the model and the Fluctuation Ratio, which measures the stability of Partial Dependence Plots of numeric data as well as metrics to evaluate Surrogate Models, such as the Performance Degradation, which measures how much of the original model's performance is lost.

2.3.3 Joint Evaluation of Explanation Methods

Krishna et al. [35] as well as Neely et al. [59] address the *disagreement problem* that arises when multiple explanation methods (e.g. LIME, SHAP, Integrated Gradients, DeepLIFT) produce conflicting outputs for the same prediction. Their empirical study, conducted across various datasets and models, shows that different explainers can yield widely varying feature importance rankings. Furthermore, a user study with data scientists revealed that practitioners frequently encounter these disagreements, often lacking a principled method for reconciling them. Min et al. [51] address this by proposing an *ensemble interpretation* framework that integrates multiple explanation methods capable of producing feature importance rankings (e.g., LIME, SHAP, PDP) to generate model explanations that are both more stable and comprehensive. They evaluated the consistency of the ensemble by using ranking correlation indices to determine if it consistently identified the same key features as human experts, and then leveraged the ensemble as a tool for feature selection. In comparing the ensemble-based feature selection to a method based on correlation analysis, they found that the ensemble approach achieved higher accuracy.
Brdnik et al. [13] evaluate eight different explanation techniques within an educational analytics system that predicts student grades. Their user study with college students indicates that local, feature-based explanations (such as bar charts displaying feature importance) tend to improve user understanding and satisfaction compared to other forms. However, they also find that certain explanation types, particularly those based on confidence measures, may not be as effective.
Labarta et al. [37] conducted a user study with six state-of-the-art XAI techniques (LRP, GradCAM, LIME, SHAP, Integrated Gradients, Confidence Scores) to see how well users could judge, trust, and question AI decisions with each. They found that each individual method excelled at different user goals – e.g. Confidence Scores best helped users judge decision correctness, certain visual attributions (GradCAM, LRP) best helped build trust, and SHAP best enabled users to question a decision. No single explainer scored highest on all metrics. The authors conclude that "using individual explanation methods is not sufficient" for effective user understanding and they see the need for an interactive framework to focus on the user's needs. Jeyakumar et al. [32] also find that the preferred explanation method depends on the task at hand.

Chapter 3

Background

In the following chapter, six explanation methods will be introduced. The chapter will cover three established explanation methods, namely SHAP, Partial Dependence Plots, Surrogate Models and three custom explanation methods which are based on preexisting techniques. All methods are either entirely global or a hybrid approach (SHAP). Among the six methods presented, four are model-agnostic, while the remaining two novel approaches are model-specific.

3.1 SHAP

SHAP (Shapley Additive Explanations)[42] values quantify the contribution of each feature to the prediction for individual instances locally and can also be aggregated globally across all instances. SHAP values are defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} w(S)[f(S \cup \{i\}) - f(S)] \tag{3.1}$$

where:

- ϕ_i is the SHAP value for feature i,
- $f(S)$ is the prediction for subset S,
- $f(S \cup \{i\})$ is the prediction for subset S including feature i.
- $w(S)$ are weights that determine the significance of each feature subset.

This formula calculates the SHAP value ϕ_i by summing over all subsets S of features that exclude i. It uses the weight $w(S)$, which accounts for all permutations of S and $N \setminus S$, and the difference between the predictions with and without feature i to determine its impact on the model's output. An example calculation for a Shapley value is given in Appendix A.

In particular, the weights $w(S)$ are defined as:

$$w(S) = \frac{|S|!(|N| - |S| - 1)!}{|N|!} \tag{3.2}$$

where:

- S is a subset of all features excluding i,
- N is the set of all features,

The weight $w(S)$ determines the significance of adding the feature i to the subset S within the context of all possible feature subsets. As $|S|$ approaches $|N|$, the weight increases. Likewise, if $|S|$ approaches $\emptyset$, the weight also increases. The function assigns more weight to the prediction difference, if there are few features in S, as the resulting prediction difference is more closely corresponding to the pure effect of i as well as sets where S is nearly the complete set of N, as significant changes after the inclusion of a feature inside an almost complete set of features indicate a strong effect of this feature on the prediction.

SHAP values are "the Shapley values of a conditional expectation function of the original model" [42]. Unlike traditional Shapley values, which can be applied to a wide range of allocation and distribution problems, SHAP values are specifically designed to explain how different features contribute to a model's predictions.
SHAP values are grounded in three fundamental properties that make them particularly desirable for machine learning explanation methods:

- **Local Accuracy:** The sum of all SHAP values plus the base value equals the model's prediction for each individual instance.
- **Missingness:** If a feature is absent in the input, its contribution to the prediction is exactly zero.
- **Consistency:** If the contribution of a feature to the model's output increases or remains unchanged, regardless of other features, its SHAP value will not decrease.

3.2 Embedded SHAP

For large datasets, having a large list of feature importances is not interpretable according to the sparsity constraint outlined in Chapter 2. One possibility that can be used in order to comply to this constraint is to utilize embeddings as a method to reduce the dimensionality of the feature space while retaining the most significant and interpretable aspects of the data.
Figure 3.1 shows three common methods for dimensionality reduction consisting of Principal Component Analysis (PCA) [60], Variational Autoencoders (VAE) [34], and Uniform Manifold Approximation and Projection (UMAP) [47].

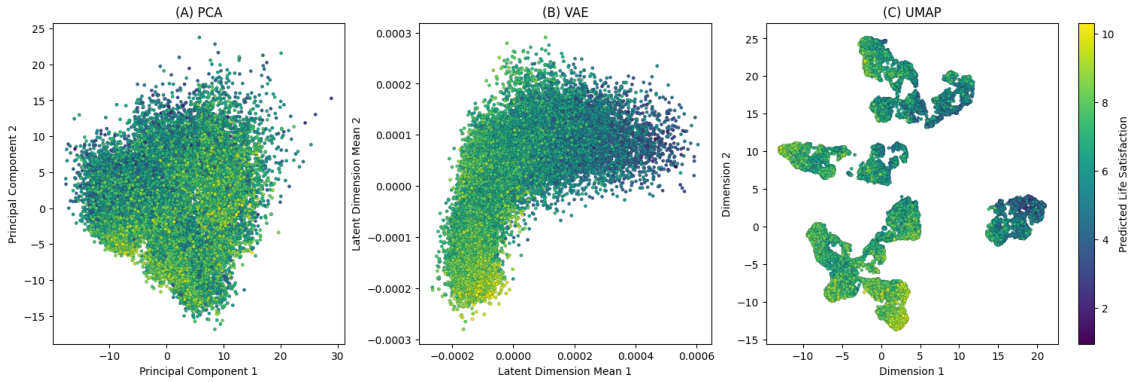


Figure 3.1: Two-dimensional embeddings along with the associated predictions obtained by fitting a tree-based gradient boosting model (LightGBM) to a dataset consisting of 4185 features. (A) Principal Component Analysis (PCA): The data is projected onto a plane with the highest variance, represented by the first two principal components. (B) Variational Autoencoder (VAE): The data is embedded into a two-dimensional latent space. (C) Uniform Manifold Approximation and Projection (UMAP): The data is visualized in two dimensions using Supervised UMAP, a minimal distance of 0.2, 50 neighbors and a negative sample rate of 20.

An essential consideration in dimensionality reduction, especially when explaining a machine learning model, is how effectively the method maintains the distances between data points in the reduced-dimensional space. The three outlined methods handle this challenge in distinct ways:

- **PCA** does not explicitly aim to preserve pairwise distances between data points. Instead, it seeks to maximize the variance along the principal components, which can often result in a rough approximation of the distance relationships in the original space, particularly for linear relationships.
- The **VAE**'s primary goal is to encode the data into a latent space in a way that allows for accurate reconstruction. While the encoder-decoder framework does not explicitly preserve distances, the latent space often captures the underlying data structure.
- Uniform Manifold Approximation and Projection (**UMAP**) aims to preserve both the local and global structure of the data, by pulling similar points together and pushing dissimilar points apart through optimization processes.

Aligned UMAP extends UMAP by using Procrustes analysis [29] to align embeddings from related datasets. In machine learning model training, SHAP values across iterations can be viewed as such related datasets. Each training iteration adjusts feature impacts on the target, leading to corresponding adjustments in SHAP values for each observation. Similarly, in forest-based explanation methods, each new tree can be considered a related dataset, as it refines and builds upon the results of previous trees.

Procrustes analysis seeks to find an orthogonal matrix Ω that best aligns two sets of point representations, given by matrices A and B. The primary objective here is to minimize the Frobenius norm of the difference between ΩA and B, where Ω is constrained to be orthogonal, thus ensuring that it represents a pure rotation (or rotation combined with reflection) which preserves the geometric properties of the points in matrix A, such as distances and angles, while aligning them as closely as possible to the points in matrix B:

$$\begin{aligned} &\min_{\Omega} \|\Omega A - B\|_F \\ &\text{subject to } \Omega^T \Omega = I \end{aligned} \tag{3.3}$$

where:

- $\|\cdot\|_F$ denotes the Frobenius norm
- A and B are matrices representing SHAP embeddings,
- Ω is an orthogonal matrix,
- I is the identity matrix, ensuring $\Omega^T \Omega = I$ confirms the orthogonal property of Ω.

It can be shown that the optimal rotation matrix Ω can be derived by:

$$\Omega = UV^T.$$

where U and V are orthogonal matrices obtained from the singular value decomposition (SVD) of the matrix $M = BA^T$. U contains the left singular vectors and V contains the right singular vectors of M. An example calculation for the rotation matrix can be found in Appendix B.

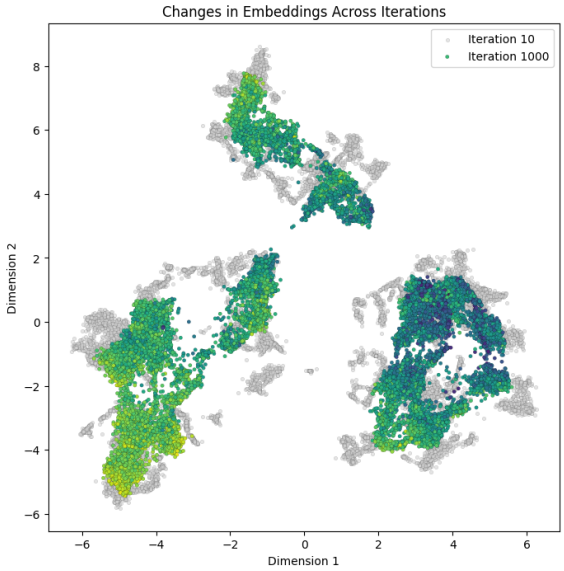


Figure 3.2: The gray points correspond to the UMAP embeddings generated after 10 iterations. At iteration 1000, the points have shifted. There is a clear correspondence between the clusters at both iterations.

Figure 3.2 demonstrates that points are well-aligned between iterations. In addition to preserving local and global relationships between points, UMAP has the valuable property of forming clusters. These clusters can be further analyzed by identifying the features that were most influential in the predictions at specific stages of the model training process for each cluster. A natural choice

for clustering clear, distinct clusters is density-based clustering, particularly Hierarchical Density-Based Clustering (HDBSCAN) [46], which is effective for both sparse and dense clusters. Figure 3.3 illustrates that HDBSCAN successfully separated the different clusters. The methodology of calculating SHAP values, embedding the iterations individually, aligning them, and analyzing the resulting clusters will hereafter be referred to as Embedded SHAP.

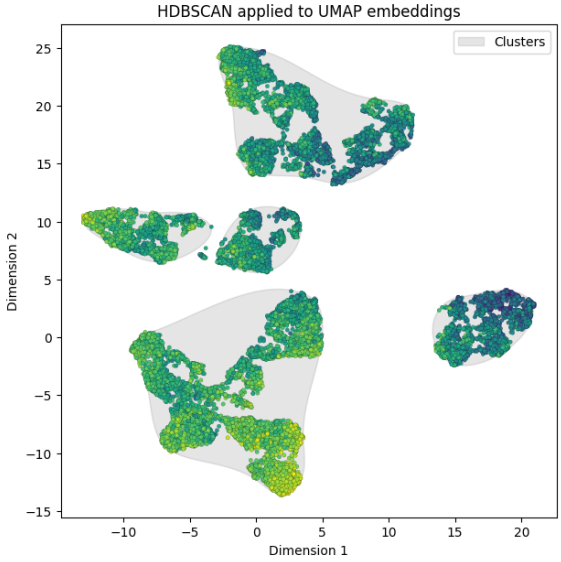


Figure 3.3: Cluster Analysis Using HDBSCAN: B-Spline Smoothed Concave Hulls Indicated in Gray

3.3 Partial Dependence Plots

Partial Dependence Plots [25] show the marginal effect of a variable on the prediction of the model by holding all other variables constant and only varying the values of the variables under consideration. The partial dependence plot is calculated using the following formula:

$$f_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)}) \tag{3.4}$$

where $f_{x_S}(x_S)$ represents the partial dependence function, f is the original model, x_S are the features of interest, $x_C^{(i)}$ are the values of other features for the i-th instance, and n is the total number of instances.
A major point of criticism is that Partial Dependence Plots ignore interactions between variables and can therefore lead to incorrect interpretations. If, for example, the variable 'income' is considered in isolation with regard to life satisfaction, PDP could lead to the conclusion that a higher income always leads to higher life satisfaction, although life satisfaction possibly consists of a combination of income and working hours and a higher income, while working hours remain the same, would mean that a higher hourly rate was earned.
Accumulated Local Effects (ALE) [7] were developed to avoid such erroneous interpretations. ALE differs from PDP in that it considers the effects of a variable on the prediction of the model in the vicinity of other variable values. The value range of a variable is divided into several intervals and the average effects of a small change within these intervals are calculated, which is intended to avoid unrealistic feature combinations. Despite these advantages, Partial Dependence Plots were chosen for the following analysis due to their simplicity.

3.4 Surrogate Models

In addition, there are explanation approaches that summarize the core elements of a complex model, so-called surrogate models. These models serve as substitutes for the original model and attempt to approximate its functionality in a simplified, easily understandable way.

$$\hat{f}(x) \approx f(x) \tag{3.5}$$

Where:

- $\hat{f}(x)$ is the surrogate model's approximation.
- $f(x)$ is the true, complex model output being approximated.
- x is the vector of input variables.

Which simplified model is used depends primarily on the structure of the model to be described. Regression and classification models are often represented by linear models or decision trees. LIME [66] can be classified as a surrogate model which only approximates the original model in the proximity of a specific point of interest.
In addition to local models, there are also rule-based models that work on a global level such as RuleFit [26]. Here, rules are created that show the most important decision paths within the underlying model. In contrast to linear regression, the interactions of the individual characteristics are also taken into account. Unlike the original model, which tries to predict the target feature as well as possible, a surrogate model instead tries to predict the prediction from the original model.

$$x^* = \arg\min_x L(\hat{f}(x), f(x)) \tag{3.6}$$

Where:

- x^* represents the input values that minimize the loss function L.
- $L(\hat{f}(x), f(x))$ is the loss function that measures the discrepancy or error between the surrogate model's predictions $\hat{f}(x)$ and the true objective function $f(x)$.
- $\hat{f}(x)$ is the surrogate model that approximates the true objective function $f(x)$.
- $f(x)$ is the true objective function.
- $\arg\min_x$ is the argument that minimizes the loss

3.5 Frequent Pattern Mining (Lift)

Ensemble methods such as Random Forests and Gradient Boosting inherently use features in the sequence in which splits occur, which allows a direct analysis of the decision process. Previous work has investigated how random forests could leverage bivariable splits to allow direct interpretation [31] or Random Intersection Trees to construct feature interactions [36]
Another approach to directly leverage the untangled nature of decision trees is to treat the paths from the root to each leaf as a set and identify recurring patterns in the decision-making process, which can be achieved through *frequent itemset* or *sequential pattern* mining on the set of decisions. *Sequential pattern* mining is particularly useful when pairs of adjacent nodes are believed to have differing conditional probabilities. For instance, if feature A frequently precedes feature B, but feature B does not equally frequently precede feature A, sequential pattern mining would be the preferred method. If the pairs of adjacent nodes possess identical conditional probabilities, making the order of feature occurrences irrelevant, *frequent itemset* mining is preferred.[3]
A commonly used method for frequent itemset mining is the Apriori Algorithm [2]. The Apriori Algorithm is a breadth-first search algorithm that first scans the dataset for single items that meet a specified threshold. Items that do not meet this threshold are discarded. The remaining items are then combined into pairs, and pairs that do not meet the threshold are again discarded. This process continues until itemsets of the desired maximum length are formed. The algorithm relies on the principle that a superset cannot be frequent if any of its subsets are infrequent. Thus, if A is a subset of $A \cap B \cap C$, then $A \cap B \cap C$ cannot meet a frequency threshold if A does not.

In this context, the itemsets are derived from all the splits made by the model across all trees. Consecutive features are treated as items, using a sliding window that includes several features at a time.
To identify frequently co-occurring itemsets, *Lift* can be used as a metric to measure the strength of associations. It evaluates how much more often two items A and B occur together than would

be expected if they were independent. The Lift for the features A and B in the model is defined as:

$$\text{Lift}_{\text{model}}(A, B) = \frac{P_{\text{model}}(A \cap B)}{P_{\text{model}}(A) \cdot P_{\text{model}}(B)} \tag{3.7}$$

where:

- $P_{\text{model}}(A \cap B)$: The joint probability of A and B occurring together in consecutive splits.
- $P_{\text{model}}(A)$: The probability of feature A occurring in a split.
- $P_{\text{model}}(B)$: The probability of feature B occurring in a split.

A Lift value greater than one indicates that the co-occurrence of features A and B in the model is more frequent than would be expected if they were independent.

However, this calculation does not account for dependencies between features, where the presence or absence of one feature is influenced by another. For example, in a dataset capturing people's backgrounds, if an individual has lived in a third country, it inherently means they have also lived in a first and second country. Likewise, the year someone ceased experiencing financial difficulties implies there was a specific year when those difficulties began. To address such dependencies, an alternative perspective on Lift can incorporate the features from the training instances as a secondary population. In this approach, the second Lift value represents the joint probability of a specific feature combination occurring *in the data the model was trained on*, relative to the probabilities of the individual features being present independently.

$$\text{Lift}_{\text{training}}(A, B) = \frac{P_{\text{train}}(A \cap B)}{P_{\text{train}}(A) \cdot P_{\text{train}}(B)} \tag{3.8}$$

where:

- $P_{\text{train}}(A \cap B)$: The joint probability of features A and B being set together in the training data.
- $P_{\text{train}}(A)$: The probability of feature A being set in the training data.
- $P_{\text{train}}(B)$: The probability of feature B being set in the training data.

The ratio Equation 3.9 measures the strength of dependencies (or lack thereof) under two distributions

$$\text{Ratio}_{\text{lift}}(A, B) = \frac{\text{Lift}_{\text{splits}}(A, B)}{\text{Lift}_{\text{training}}(A, B)} \tag{3.9}$$

A ratio greater than 1 indicates a stronger dependency between the features in the model than between the same features in the dataset.

3.6 Feature Context Embedding

Another method for uncovering relationships between features is word embedding [50], a widely used technique in natural language processing that evaluates the closeness of words based on their contextual usage. Similarly, in feature analysis, a model can infer a given feature—such as health, language, age, wealth, or origin—by analyzing its contextual neighbors. These neighbors are derived from the paths within a decision tree or a forest-based explanation method. The context of each node (and consequently each feature) is defined by incorporating the preceding split (parent) and the subsequent splits (children).

Mikolov et al. introduced two primary approaches for embedding words based on their context. The Continuous Bag of Words (CBOW) approach predicts the probability of a word based on the words surrounding it, whereas the Skip-gram model uses the central word to predict the individual words around it. In both methods, embeddings are constructed through an embedding layer within a neural network. This embedding layer maps each word to a continuous vector space of fixed dimensionality, which are adjusted so that words used in similar contexts are mapped to nearby points in the vector space. Within each context window, proximity to the word does not influence the embedding of the word.

Text	**Skip-Grams**	**CBOW**
[health language age] wealth origin	language →health language → age	(health, age) → language
health [language age wealth] origin	age → language age → wealth	(language, wealth) → age
health language [age wealth origin]	wealth → age wealth → origin	(age, origin) → wealth

Table 3.1: Skip-Gram and CBOW in Splits with Context Windows Specified as Brackets

The key parameters of this method are the context window size and the dimensionality of the vector embeddings. The example in Table 3.1 illustrates a context window size of one. In the CBOW model, only the immediately preceding and following words are considered when predicting the central word, whereas the Skip-Gram model uses the central word to predict the surrounding words. There are other models that generate static word embeddings. Most notably GloVe[62] and BERT[21].

Chapter 4

Methodology

The six explanation methods introduced in chapter 3 will be evaluated using two distinct approaches. First, the three novel techniques will be validated through synthetic data analysis (RQ1) to confirm they accurately depict inherent relationships between predictors and the target variable. In the second phase, a user study is created which investigates participants' subjective and objective understanding (RQ2), explores how the perceived usefulness of individual methods influences the perceived added value of their combinations (RQ3), and determines whether understanding the methods leads participants to draw correct inferences from the explanations provided (RQ4).

4.1 Synthetic Data Evaluation Methodology

The novel explanation approaches will be evaluated using synthetic data. Using synthetic data is a common approach to test for correctness [58], test if the explanation covers key aspects [52] and to test if the explanation leads to a simulatable result [56]. To evaluate the hypotheses regarding the expected outcomes of these explanations, tests will be conducted using a LightGBM model with 100 estimators and default parameters.

4.1.1 Embedded SHAP

To evaluate the Embedded SHAP values, two key hypotheses are proposed:

- **RQ1H1: Variability in Feature Importance Projections.** The first hypothesis is that complex dependencies between a feature and the target variable lead to greater fluctuations in the assigned importance of that feature, and consequently, in the corresponding embeddings. This is because no simple approximation can adequately capture the complexity of such dependencies. As a result, the SHAP values are likely to adjust more frequently, resulting in visible changes within the lower-dimensional embedding space. Simpler dependencies, however, should result in embeddings that remain mostly static.
- **RQ1H2: Cluster Formation.** The second hypothesis is that instances significantly different from the rest of the population will form a separate cluster. If this subgroup is small, its separation should only be clearly observed in later iterations. This is particularly true for forest-based machine learning models, as the initial splits aim to minimize the overall error, rather than finding nuances for small subgroups.

RQ1H1: Variability in Feature Importance Projections

The explanation method should be sensitive to underlying functions' complexity and update the embeddings more frequently for instances associated with more challenging functions. To examine the differential reactions of groups to functions of varying complexity, two underlying functions, A and B, are defined:
The impact of A on the target variable is defined as:

$$\text{Impact}_A(x) = \frac{1}{\pi} \cdot x + 1 \tag{4.1}$$

The impact of B on the target variable is defined as:

$$\text{Impact}_B(x) = \sin(10x) \tag{4.2}$$

The scaling factor of 10 amplifies the oscillation, making the function $\text{Impact}_B(x)$ more difficult for a machine learning model to learn.
The target variable y is calculated by summing the impacts of A and B:

$$y_i = \text{Impact}_A(A_i) + \text{Impact}_B(B_i) \tag{4.3}$$

Each instance in the population is assigned to either underlying function A or underlying function B.

$$(A_i, B_i) = \begin{cases} (\text{Uniform}(0, 2\pi),\ 0) & \text{with probability 0.5,} \\ (0,\ \text{Uniform}(0, 2\pi)) & \text{with probability 0.5.} \end{cases}$$

The expected result is that instances belonging to group B will exhibit more frequent shifts in their embedding space representation compared to those in group A, given that underlying function A is a straightforward linear function. Within the range $[0, 2\pi]$, $\text{Impact}_A(x)$ spans from 1 at $x = 0$ to 3 at $x = 2\pi$, matching the value range of 2 seen in the function $\text{Impact}_B(x)$, which varies from -1 to 1. For the simulation, 1000 samples were generated for both functions within the range $x = 0$ to $\frac{2}{\pi} - 1$ and combined into a single dataframe. Both features were then used to predict the target variable in Equation 4.3.

RQ1H2: Cluster Formation

To simulate the formation of clusters based on underlying feature interactions and the influence of a small subgroup, three independent features, A, B, and C, are generated. These features are then subjected to quadratic functions. Additionally, a small subgroup, comprising 5% of the population, is assigned a fixed impact on the target variable through a dummy variable.
The independent features are defined as:

$$A_i, B_i, C_i \in [0, 10], \quad i = 1, \ldots, N$$

The target variable y is calculated by summing the squared impacts of each feature, with an additional large impact of 100 applied only to the instances belonging to the subgroup identified by the dummy variable:

$$y_i = A_i^2 + B_i^2 + C_i^2 + \text{Dummy}_{0.05} \times 100 \tag{4.4}$$

Where $\text{Dummy}_{0.05}$ is a binary variable that equals 1 for the 5% subgroup and 0 otherwise.

Within the value range $[0, 10]$, the impact of $\text{Dummy}_{0.05}$ is as large as each of the features A, B, and C can get at most. However, unlike $\text{Dummy}_{0.05}$, features A, B, and C impact the entire population. Therefore, the cluster formation should only become visible in a later iteration.

4.1.2 Frequent Pattern Mining (Lift)

The frequent patterns that emerge should reveal interaction effects between the features in the feature space. Consequently, two hypotheses are proposed to address the general magnitude and order of the interaction values:

- **RQ1H3: Co-occurrence Drives Lift.** Features with strong interaction effects are expected to frequently appear together along the paths of a decision tree, and their co-occurrence should be more common than their individual appearances. This will result in a higher Lift value.
- **RQ1H4: Interaction Strength Correlates with Lift.** The greater the strength of the interaction, the higher the expected Lift value. If an interaction is strong, the features involved should seldom occur without the corresponding interaction feature being present.

The synthetic dataset, developed to validate these hypotheses, consists of three primary features and four interaction features, each sampled from a uniform distribution over the interval $[0, 10]$. The primary features are designed to have a direct impact on the target variable, while the interaction features contribute to y through interaction effects.
The primary features are denoted as PrimA, PrimB, and PrimC, and are defined as follows:

$$\text{PrimA}_i,\ \text{PrimB}_i,\ \text{PrimC}_i \sim \text{Uniform}(0, 10), \quad i = 1, \ldots, N$$

The interaction features, denoted as InterA, InterB, InterC, and InterD, are similarly defined but with the majority of their values randomly set to zero. Specifically, 95% of the values for each interaction feature are set to zero to simulate sparse interaction effects:

$$\text{Inter}_i \in \{\text{InterA}_i, \text{InterB}_i, \text{InterC}_i, \text{InterD}_i\}, \quad \text{Inter}_i = \begin{cases} \text{Uniform}(0, 10), & \text{with probability } 0.05 \\ 0, & \text{with probability } 0.95 \end{cases}$$

Two interaction effects were calculated: one between InterA and InterB, and another between InterC and InterD:

$$\text{InterAB}_i = \text{InterA}_i \times \text{InterB}_i$$

$$\text{InterCD}_i = \text{InterC}_i \times \text{InterD}_i$$

The target variable y is calculated as the sum of the squared values of the primary features and the scaled squared interaction effects:

$$y_i = \text{PrimA}_i + \text{PrimB}_i + \text{PrimC}_i + 2 \times \text{InterAB}_i^2 + 4 \times \text{InterCD}_i^2 \tag{4.5}$$

Given Equation 4.5, InterCD should result in a larger Lift value than InterAB. Both features should have a larger Lift value than any other feature combination.

4.1.3 Feature Context Embeddings

For testing the context embeddings, the following two hypotheses are proposed:

- **RQ1H5: Contextual Clustering.** Feature context embeddings should be arranged in a 2D space such that features appearing in similar contexts are clustered.
- **RQ1H6: Global Similarity Optimization.** The overall arrangement should maximize inter-cluster similarity.

A dataset consistent with this arrangement consists of two distinct feature groups, each influencing the target variable y independently. Additionally, a shared feature contributes to y regardless of which group is active.
Each group contains five features that independently contribute to the target variable y when the group is active.

$$\text{Group}_i = \begin{cases} 1, & \text{with probability } 0.5 \\ 2, & \text{with probability } 0.5 \end{cases}$$

Here, 1 and 2 represent the two distinct feature groups, with the active group being chosen independently for each instance.
The features in each group were generated uniformly over the interval $[0, 10]$ and are defined as:

$$\text{Group\_1\_Feature}_i \sim \text{Uniform}(0, 10), \quad i = 1, \ldots, 5$$

$$\text{Group\_2\_Feature}_i \sim \text{Uniform}(0, 10), \quad i = 1, \ldots, 5$$

The shared feature is defined as:

$$\text{Shared\_Feature} \sim \text{Uniform}(0, 10)$$

The target variable y is determined by summing the values of the active features within the selected group and adding a shared feature that influences y across both groups.

$$y_i = \sum_{j=1}^{5} \text{Group\_1\_Feature}_{ij} + \sum_{j=1}^{5} \text{Group\_2\_Feature}_{ij} + \text{Shared\_Feature}_i \tag{4.6}$$

Two window configurations are tested. One window configuration consists of the parent of the node, the node itself as well as one of its children. Two windows are generated for each child - one for each child. (Parent-Node-Child 1, Parent-Node-Child 2). The other configuration only considers the children and does not consider the parent (Child 1-Node-Child 2). This corresponds to a window size of 1.
The resulting embedding should display two distinct clusters, each representing one of the groups in Group_i. The *Shared_Feature* is expected to be positioned between these clusters, as it is not associated with either group.

4.2 Real-World Data Evaluation Methodology

Additionally, all explanation methods will be evaluated in a user study involving laypeople recruited from MTurk. The study aims to assess the comprehensibility of the methods using both subjective and objective measures, determine whether understanding the model leads to correct conclusions, and investigate how combining explanation methods influences model understanding.

4.2.1 Dataset

The dataset originates from the Survey of Health, Ageing and Retirement in Europe (SHARE) [77], a longitudinal study conducted across the European Union and Israel since 2004. The dataset comprises more than 4000 features from 160,000 participants. The survey covers many categories such as health, economic conditions, social and family networks, labor market participation, and demographic information.
The categories included in the studies remained largely constant throughout the surveys. Categories that were covered in all waves can be seen in Table C.1. Other categories, such as Computer Use and Saving Regrets, were included in some waves but not consistently across all of them. In order to be eligible for the survey, participants had to

- be at least 50 years old at the time of the survey.
- reside in one of the countries involved in the survey.
- have a partner in the same household that is eligible, regardless of the persons age.
- be capable of completing the interview, either independently or with assistance.

Target Feature Definition

Life satisfaction has been chosen as the feature of choice which was included in the survey across all waves with the exception of wave 1 and 3. Given its universal importance, it serves as a meaningful target variable that is relevant across different demographic and socio-economic backgrounds. The effect of various features on life satisfaction has been studied extensively[39][78][17]. Personality traits such as Neuroticism and Openness[43], Socioeconomic Status[84] and other community related features[18] have all shown to impact life satisfaction.

Data Preprocessing

Missing values were categorized as either "Missing Not at Random" (MNAR) or "Missing at Random" (MAR) based on whether the individual in question had responded to any of the survey questions in a particular category. If a respondent did not answer any questions within a category, it was assumed that they had not participated in the survey wave where that category was covered, and their missing values were classified as MAR. In contrast, if a respondent participated but left certain questions unanswered, these omissions were classified as MNAR. This classification assumes that specific questions were inapplicable, such as a question about children for a respondent who has indicated they have none.
The datasets were subsequently merged and imputed in both a forward and backward direction for each individual participant. For instance, if a person has provided the number of children in one survey but not the other, the values were transferred across waves.
The data displayed a skew towards higher life satisfaction values. To achieve a more balanced distribution and ensure each individual is represented only once in the final dataset, the wave which included the minimum life satisfaction value recorded for each person across all survey participations was kept, while other waves were excluded.
Moreover, all features from the activities survey were excluded from the analysis. The survey contained several questions that, while highly correlated with the target variable of life satisfaction, likely do not exert a causal influence on it. Instead, these items may act as proxies for life satisfaction itself, complicating the interpretation of any predictive modeling efforts aimed at identifying independent predictors. Examples of such features include:

- **AC021_LifeMean**: Respondents feel their lives have meaning.
- **AC025_FutuGood**: Respondents express optimism about their future.
- **AC030_Happy**: Respondents felt happy most of the time during the previous week.

- **AC032_EnjLife**: Respondents enjoyed life over the past week.

Furthermore, numeric outlier values that fell outside 1.5 times the interquartile range were dropped and replaced by MAR, and categorical features that had more than 99% [1] missing data or had more than 600 unique categorical values were also removed.

4.2.2 Model Training

Several options were evaluated as viable regression models, all of which could handle missing data. All considered methods are based on gradient boosting techniques (LightGBM[33], XGBoost[16], and CatBoost[64]). Gradient boosting algorithms are frequently observed to outperform other models, such as linear models or neural networks, particularly when employed for tabular regression [45]. Although on large datasets, neural networks can exhibit superior performance compared to traditional tree-based learning algorithms[12].
Table D.1 shows that CatBoost had the highest performance on the test set with the fewest features and the smallest gap between training and test scores, while also using the most trees. Although LightGBM is the fastest running model, it had a tendency to overfit. XGBoost, on the other hand, uses the least number of trees and leaves. With the default regression parameters, LightGBM was ultimately selected, as it delivered performance nearly equal to CatBoost but with significantly faster execution. To further optimize the hyperparameters, a tree-structured Parzen estimator[10] was used. The final hyperparameters can be seen in Table D.2.

Features that appeared only once in all trees were omitted for simplicity. The complete list of all features of the model can be found in Appendix G. The obtained result from the fine-tuned LightGBM model is comparable to other studies. Shen et al.[78] achieved a R^2 value of 0.436 using support vector regression by selecting features from the RAND Health and Retirement Study[81] dataset that had a Pearson correlation coefficient with the target greater than 0.2. Further feature refinement was performed using LASSO regression, resulting in a final set of 18 features. Notably, their model primarily included features that were explicitly excluded from the models depicted in Table D.1, which focused primarily on subjective factors such as emotional well-being, social support, and personality traits. In comparison, the study "Understanding Key Predictors of Life Satisfaction in a Nationally Representative Sample of Koreans"[17] used data from the Gallup World Poll[27]. They used multiple linear regression and obtained an R-squared value of 0.307. They used 27 characteristics, including demographic and psychological variables such as satisfaction with standard of living, household income, positive affect, social support, and education level. Malvaso et al. reported an R^2 of 0.514. The most influential factors were found to be satisfaction with spouse, social life, and six other satisfaction measures, as well as personality traits such as neuroticism or openness.
The chosen LightGBM model treated all levels of life satisfaction equally, without assigning different weights to them. Consequently, less frequently occurring values, particularly those at the lower end of the scale, were not predicted as accurately. As shown in Figure 4.1, there is no significant discrepancy between the training and testing sets in terms of the direction and magnitude of the error.
The hyperparameter range was chosen to reduce overfitting and handle features with high cardinality. In particular, the number of leaves was set to a relatively low range (considering the large dataset size), as shown in Table D.2 (30–40; default is 31). Similarly, the minimum number of child samples required per leaf was adjusted to a higher range compared to the default (80–150; default is 20). Consequently, the model prioritizes generalization over capturing influential factors that affect the life satisfaction of only a small fraction of individuals in the dataset. The training process was performed using early stopping. If adding another tree did not improve the performance of the validation set, training was stopped.

4.2.3 Model Explanation

This section applies the six previously introduced explanation methods to the LightGBM model. Additionally, two metrics specific to tree-based methods — Gain and Split — will be examined. Gain measures the total reduction in loss achieved by adding a specific node, with the overall Gain for a feature calculated as the sum of error reductions across all nodes where the feature is used. Split represent the count of times a feature is selected for decision-making splits.

[1] 1% or less corresponds to less than 2096 instances with a value for a given categorical feature as there were 209.606 instances in the dataset

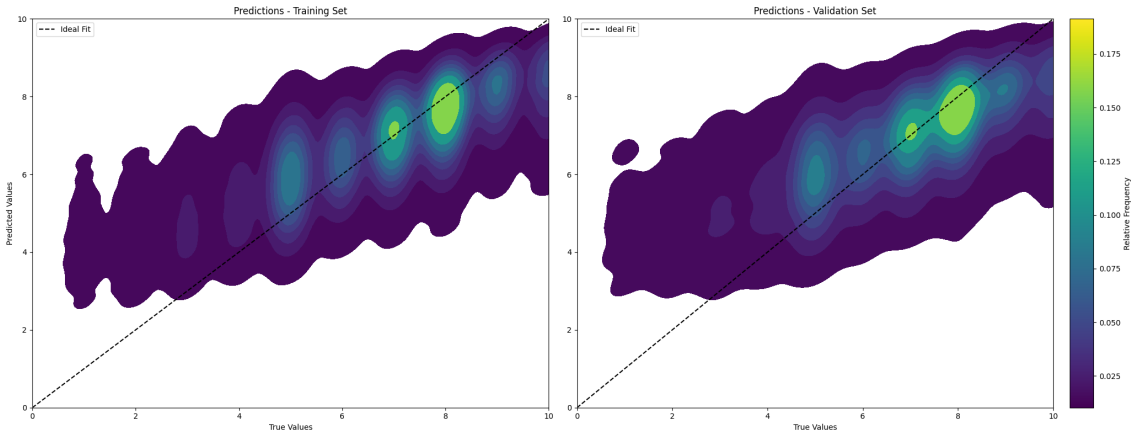


Figure 4.1: Kernel Density Estimation (KDE) plots of the true vs. predicted values for the training set (left) and validation set (right). The dashed line represents the ideal fit where predicted values match true values.

Table E.1 lists the features grouped by the categories assigned in the survey. Mental and physical health stand out as the most important factors affecting life satisfaction, accounting for 44% of the model's cumulative gain, even though they are not frequently used in splits. They appeared in only 7% and 8% of splits respectively. The category of retrospective accommodation appeared in the most splits (26%), followed by demographics and networks, and language.
The demographics and networks category includes characteristics such as the country of birth of the individual's parents. Retrospective accommodation includes the regions where individuals have lived for more than one year since birth, covered by characteristics `ra015c_1` through `ra015c_30`. Consumption focuses on financial management, specifically whether a person can cover daily expenses. Expectations include various features such as the expected lifespan, financial outlook, possibilities of inheritance, and factors like personal trust and political views. General life covers financial troubles, happiness, stress, and health issues throughout life. Interviewer Observations note how willing and able the respondent was to answer the questions.
Table E.2 lists the top ten features ranked by gain, which together contribute 62% of the total gain. Self-rated physical health alone accounts for nearly a quarter of the model's cumulative gain. Features `ph003_` and `co007_` dominate their respective groups (physical health and consumption), accounting for almost all of their impact. In comparison, the impact of mental health is more spread across several characteristics. The top three characteristics in the mental health category (`mh002_`, `mh037_`, and `mh003_`) together account for only 51% of the total mental health influence.
Table E.3 shows the features ranked by split occurrences. Features with a high split count also tend to have higher cardinality than most other features. Features belonging to the category of retrospective accommodation have the second highest cardinality in the data set. The features had between 257 and 273 unique values. Country of birth of father and mother also have high cardinality as the values of these features also refer to smaller regions within each country.
Six of the top ten features relate to origin, accounting for 25.6% of all splits. Notably, the sequence of accommodation nodes often follows a pattern, with the first and second accommodations frequently succeeding each other in the splits. It is, in fact, the most common pattern observed in the model. The feature ranking also matches the sequence of places a person has lived. The first place a person lived was more important than the second place that person lived, followed by the second and third places that person lived. One possible explanation is that early childhood accommodations have a more significant impact than those later in life. Alternatively, it may reflect the certainty that everyone has lived in at least one place, while fewer people may have lived in a second or third place. Notably, the feature `ra015c_1` was missing in 57% of the data (both at random and not at random), compared to `ra015c_2` with 0.59% missing and `ra015c_3` with 69% missing.
Social Network Satisfaction, trust in other people, and a household's ability to make a living each had fewer than 10 unique values but appeared frequently in splits.

SHAP

The global SHAP values ranking is largely identical to the features ranked by gain. The retrospective region of residence has a lower SHAP ranking as it appears only once. Instead, trust in other

people and the self assessed likelihood of still living in ten years is ranked higher.

	Feature	Description	Abs. SHAP	Avg. SHAP
0	ph003_	Self assessed health status	8102.342089	0.316041
1	co007_	Is household able to make a living	5731.110799	0.223548
2	language	Language of questionnaire	4722.081054	0.184190
3	mh002_	Sad or depressed last month	3460.723439	0.134989
4	sn012_	Social Network satisfaction	2451.080341	0.095607
5	ex026_	Trust in other people	2098.157471	0.081841
6	ra015c_1	Region of residence (not current) - coded	1989.911767	0.077619
7	mh037_	Feels lonely	1967.004700	0.076725
8	ex009_	Self assessed likelihood of still living in ten years	1911.557966	0.074562
9	mh003_	Hopes for the future	1739.364384	0.067846

Table 4.1: Absolute Summed SHAP values and Average per Person

Embedded SHAP

Individual SHAP values were clustered using UMAP with a three-dimensional output and clustered using HDBSCAN and various different values for the min samples hyperparameter. Figure 4.2 shows the clusters with min samples set to 60. The clusters did not differ significantly with respect to the features that were highly important in each cluster. Self-assessed physical and mental health, household consumption, or network satisfaction were the most important SHAP values in each group. None of the other features were the most important feature in any of the clusters. The clusters form different combinations of the most significant features. Cluster 1 in Figure 4.2 includes people with good health, sufficient consumption but a lack of a strong social network, and depression. Cluster 2, which had the lowest average life satisfaction, included people with poor physical health. This is in line with the surrogate tree from Figure 4.5 which also assigned the lowest predicted value of life satisfaction when physical health was poor (Table E.3 Split 6). Instances in and around cluster 3 were given the highest predicted values. Like in the case of the cluster with the lowest predicted value, the most important feature was again the self-assigned health status.
The SHAP values were clustered at different breakpoints. Figure 4.3 shows that the general location of the clusters hardly moves between each breakpoint, which indicates that the initial trees approximate the similarity between the instances in terms of SHAP values well. However, trees later built into the model still have a substantial impact on the predictions. The predictions after ten trees were built were still close to the average prediction of 6.9.

Partial Dependence Plots

Figure 4.4 shows the partial dependence of the features `ph003_` and `co007_`, representing physical health and financial situation, respectively. The values were not averaged as in Equation 3.4 but instead show the distribution of the predicted values of $f(x_S, x_C)$. The results indicate diminishing returns on life satisfaction for both attributes. While initial improvements in health and financial status have a substantial positive impact, the effect diminishes as these characteristics reach higher levels. This assumes that the differences between individual ordinal categories are approximately equally spaced.
Furthermore, language, which ranks third in terms of gain, exhibits a notable variation in its effect on life satisfaction, the gap between the language associated with the highest predicted average life satisfaction (Hebrew) and the lowest (Estonian) is 0.42.
The country in which the survey was carried out is strongly correlated with the partial dependence of language. The Nordic countries and Switzerland have the highest average predictions. Likewise, Estonia and Hungary are predicted to have the lowest average life satisfaction.

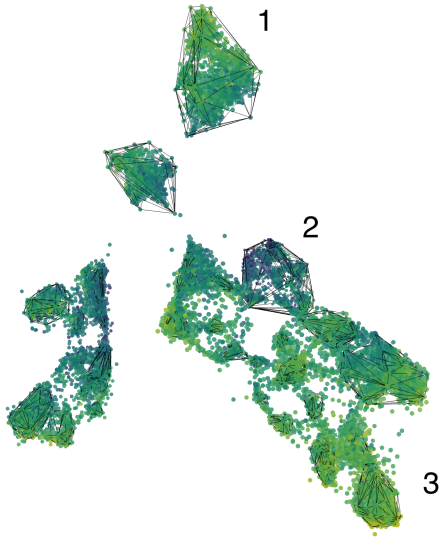
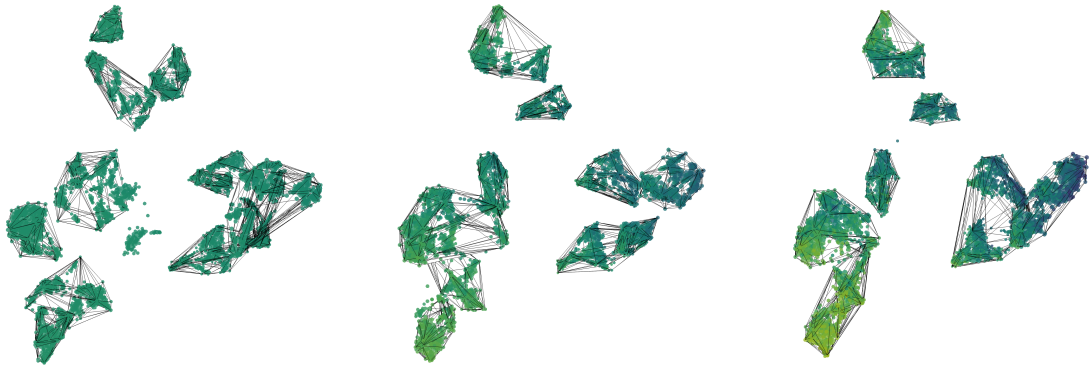


Figure 4.2: UMAP Embedded SHAP values in three-dimensional Space. Lines show the Convex Hull around the Clusters generated using HDBScan. Lighter points correspond to a higher predicted life satisfaction



(a) Tree 10 (b) Tree 100 (c) Tree 300

Figure 4.3: Aligned UMAP Embeddings after 10, 100 and 200 Trees have been built. The colors change as the model is adjusting the predictions.

Surrogate Model

The Surrogate Tree seen in Figure 4.5 is a LightGBM model with ten leaves. The features that appear in the simplified tree are also among the most important features of the original model when ranking by gain. The simplified model achieves an R^2 value of 0.58 with respect to the predictions of the larger original model. Physical health is used for both the first split and one of its children. Poor physical health leads to the lowest value of predicted life satisfaction (Split 1-6-7) paired with mental health. The highest predicted life satisfaction score was achieved by people who had good physical health, had no difficulties to make a living, did not speak certain languages, and were fully satisfied with their social network (Split 1-2-4-5).

Frequent Pattern Mining (Lift)

Table 4.3 highlights the pairs of adjacent nodes with the highest Lift values across all nodes. The itemsets were generated using all three consecutive nodes across all trees, with a minimum support threshold of 0.1%. Given the limited number of feature pairs and the fact that no more than two features were evaluated in a single feature set, this low threshold was considered appropriate. The resulting combinations often involve features within the same category. For instance, features related to periods of financial hardship frequently co-occur, as illustrated in rules 1 and 2. In particular, recent financial hardships had the most substantial joint impact on life satisfaction, while prolonged periods of hardship also had a strong negative joint effect. In contrast, periods of

Figure 4.4: Partial Dependence Plots for Physical Health, Financial Situation and Language

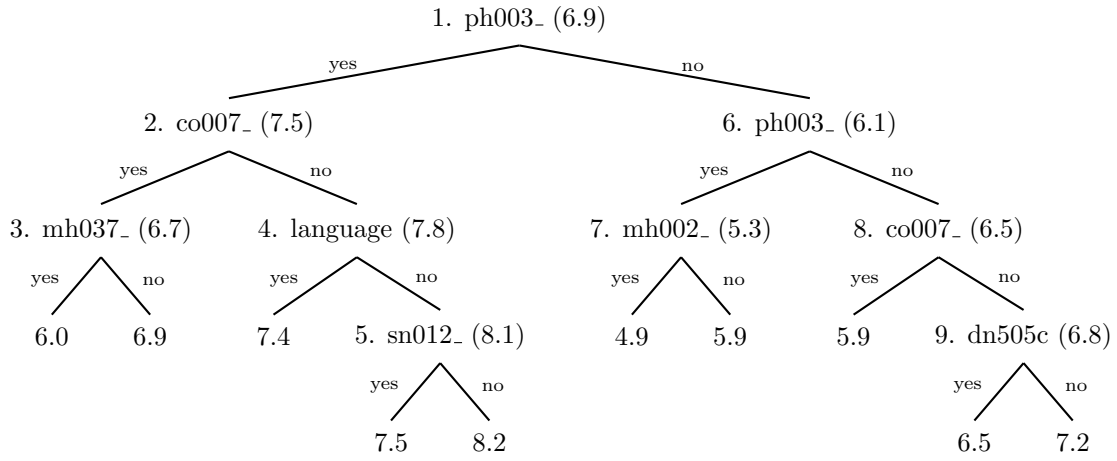


Figure 4.5: Surrogate Decision Tree. Predicted Values are given in Brackets

financial hardship that occurred long ago had minimal joint influence on current life satisfaction. The features `gl013_` and `gl011_` co-occurred 94 times more often than would be expected if they were statistically independent and appeared together in 24 itemsets. Furthermore, features from the mental health category commonly co-occurred, as shown in rules 6, 7, and 10. Retrospective living conditions demonstrated some positive associations with other retrospective accommodation features, as indicated by rules 8 and 9.

When comparing these Lift values with the results obtained from the simulated dataset (see subsection 5.1.2), there appear to be some similarities. Features measured jointly tended to have higher Lift values (`gl013_` was defined for 15% of the data, `gl012_` for 20%, with both being simultaneously defined in 15% of the data). This suggests that high interaction scores may arise not only from actual interactions within the population but also from features that are jointly missing or jointly set. section E shows $\text{Ratio}_{\text{lift}}(A, B)$ as defined in Equation 3.9, which accounts for cases where features are jointly set.

Feature Context Embedding

Both settings (Child 1- Node - Child 2, Parent - Node - Child) introduced in section 3.6 were considered for the embeddings. The Child 1- Node - Child 2 configuration showed better convergence and more stable loss and was chosen over the Parent - Node - Child configuration.
Figure 4.6 shows the two dimensional embeddings of the features that were used in the model. The model was constructed using CBOW and a window size of 1, only considering the left and right

Split	Feature		Values
1	Self assessed health status (ph003_)	=	Excellent, Good, Very good (6.9)
2	Is household able to make a living (co007_)	=	With some difficulty and 3 more (7.5)
3	Feels lonely (mh037_)	=	Often, Some of the time (6.7)
4	Language	=	French, Spanish and 16 more (7.8)
5	Social Network satisfaction (sn012_)	$\leq$	8.500 (8.1)
6	Self assessed health status (ph003_)	=	Poor (6.1)
7	Sad or depressed last month (mh002_)	=	Yes (5.3)
8	Is household able to make a living (co007_)	=	With great/some difficulty (6.5)
9	Country of birth coded: father (dn505c)	=	Austria and 24 more (6.8)

Table 4.2: Decision Rules for each Split

children and the parent respectively. The embeddings align with the rules generated by frequent itemset mining. Pairs of features that have high lift values in Table 4.3 also tend to be close in the embedding space.

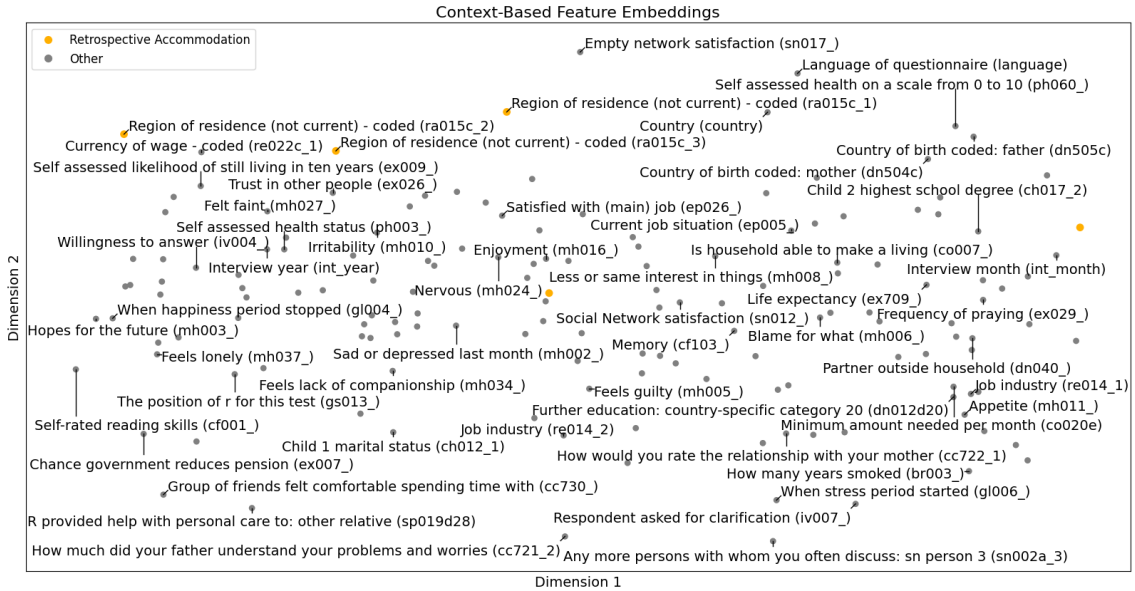


Figure 4.6: Feature Context Similarity. Features positioned closely together are used within similar contexts of the decision-making process of the model

Features related to retrospective accommodation cluster towards the upper left of the embedding space. Not only do they frequently co-occur as seen in Table 4.3, they are also used in similar contexts.

Summary

The key takeaways of the explanations are:

1. The model performs best around the average of the real target values, as gradient boosting-based methods base their initial prediction on the average but fails to properly capture low satisfaction levels. (Figure 4.1)

2. Physical and mental health, household consumption as well as features related to origin such as language and retrospective region of residence dominate the impact on life satisfaction. (Table E.2, Table 4.1)

	Feature 1	Feature 2	Lift	Support
1	When financial hardship period stopped (gl013_)	Period of financial hardship (gl011_)	94	24
2	When financial hardship period started (gl012_)	When financial hardship period stopped (gl013_)	29	30
3	Suicidal feelings or wish to be dead (mh004_)	Afford to pay an unexpected expense without borrowing money (co206_)	9	26
4	Current job situation (ep005_)	Country (country)	8	25
5	Irritability (mh010_)	Partner outside household (dn040_)	8	34
6	Less or same interest in things (mh008_)	Hopes for the future (mh003_)	7	29
7	Sad or depressed last month (mh002_)	Hopes for the future (mh003_)	7	80
8	Region of residence (not current) - coded (ra015c_5)	Region of residence (not current) - coded (ra015c_4)	6	80
9	Region of residence (not current) - coded (ra015c_4)	Region of residence (not current) - coded (ra015c_6)	6	25
10	Suicidal feelings or wish to be dead (mh004_)	Feels left out (mh035_)	6	25

Table 4.3: Top 10 Feature Combinations Ranked by Lift

3. Features related to financial hardship frequently occur together. Features related to mental health or the region of residence also frequently occur in sequential order. (Table E.1)

4. Features such as mental health or depression are used in contexts that are unique to these features. (Figure 4.6)

5. Health, household consumption, as well as other ordinal-scaled features, show a diminishing effect for higher satisfaction levels, assuming they are approximately equally spaced. (Figure 4.4)

6. Living in a nordic country or speaking a nordic language has a positive impact on predicted life satisfaction. Living in the baltics or speaking a baltic language has a negative impact on the predicted value. (Figure 4.4)

7. Poor physical health leads to the lowest predicted values. Good physical health leads to the highest predicted values. (Figure 4.5, Figure 4.2)

8. Instances that exhibit similarity in their SHAP values during early iterations tend to remain similar in later iterations. Therefore, the similarity among instances based on the features influencing life satisfaction can be reasonably approximated using only the early iterations of the model (Figure 4.3).

4.2.4 Study Procedure

To evaluate the applicability of three novel explanation methods along with Partial Dependence Plots, a Surrogate Model, and the SHAP values, as well as the helpfulness of explanation combinations in understanding the model and factors influencing understanding, a survey was constructed. Initially, the survey was planned to include all explanation methods. However, early trials revealed that the study took a significant amount of time, which could have hindered participant recruitment. To address this issue, participants were split into groups and participants in each group evaluated two explanation methods. Different groups were exposed to distinct pairs of explanation methods.
Of the 15 possible pairs that can be formed from the six explanation methods, only a subset was selected for use. These include combinations involving SHAP, a Surrogate Model, and Partial

Dependence Plots (SHAP - Surrogate Model, SHAP - Partial Dependence Plots, Surrogate Model - Partial Dependence Plots), as well as the combinations Context Embeddings - Lift and SHAP - Embedded SHAP. There are three reasons for making this choice:

1. Including a larger number of combinations would require a larger overall sample size to ensure meaningful testing within the analysis. Limiting the combinations helps maintain feasibility.

2. Embedded SHAP requires an additional introduction to SHAP when paired with other explanation methods other than SHAP itself, therefore, other combinations were excluded to minimize cognitive overhead.

3. Because Embedded SHAP, Lift, and Context Embeddings are newly developed explanation methods, their use was limited to a single subset to mitigate the potential impact of negative reception of either method.

The study primarily employed a close-ended survey format, where participants selected responses from multiple-choice lists or rated items on a Likert scale, along with two open-ended questions.

Survey Structure

The survey was divided into three parts. The first section contained general questions about the participants' background and interest in the topic. The second section focused on explanation-specific questions, which evaluated the participants' understanding and perceived helpfulness of each explanation. The third and final section addressed the evaluation of combined methods. It explored the combined helpfulness of the methods, their relative contributions to the participants' understanding of the model, and included text boxes for respondents to provide feedback on aspects of the model they felt were insufficiently covered and to indicate which explanation they found more useful (Appendix F; item 8). An additional attention-check question was included to assess participants' focus, and it was used solely to determine which submissions were accepted. All questions and available answers are shown in Appendix F.

1. **General Questions**

   Prior to responding to questions targeting the specific explanation, participants were asked to indicate their general familiarity with mathematical concepts on a Likert scale. In addition, two questions were included to collect demographic information about the participants: their age range and educational background. The age range was categorized into five-year intervals, and the educational background included commonly recognized degrees such as high school, bachelor's, master's, and doctoral degrees, as well as their equivalents. Participants were also asked to rate their interest in machine learning explanations and features influencing life satisfaction.

2. **Explanation-Specific Questions**

   Each explanation method was first described, focusing on an abstract, high-level overview of the explanation method. Participants were then asked to rate their subjective understanding of the explanation method on a Likert scale. They were subsequently asked to select one of the four available answers related to the conceptual approach of the explanation method to measure their objective understanding. Participants then had to answer two questions related to the application of the explanation method on the LightGBM machine learning model from subsection 4.2.2 and had to rank the helpfulness of the explanation method on a Likert scale.

3. **Combined Method Evaluation**

   In the final section of the survey, participants were asked to evaluate the additional benefit provided by the combination of methods and to indicate which of the two methods contributed more to their understanding of the model. They had to indicate which aspects of the model were not sufficiently explained by the two explanations and had to explain why they found one explanation more useful than the other. A graphical representation of each section and their respective questions is shown in Figure 4.7.

Participation Constraints

The pool of participants was recruited on MTurk[20], with the goal of recruiting 15 workers per combination of explanation methods. This number was determined based on a simulation study designed using data modeled to align with the hypotheses. Participants had to meet three criteria:

they were required to be MTurk Masters, workers recognized for consistently high performance across a wide range of tasks over an extended period[74]; they needed an approval rate of 90% or higher, which is associated with improved data quality[61]; and they were allowed to participate in only one of the surveys. The final criterion was implemented after some workers had already participated in multiple surveys. Thus, the dataset includes three submissions from workers who had previously participated. This low number was deemed negligible, and these submissions were retained.
An additional criterion considered was restricting participation to workers with a specific educational background, due to the high complexity of some explanation methods. However, this was ultimately excluded because it would have reduced the variance in the education variable, thus limiting its predictive power. Workers were compensated $3 per submission, with an additional $0.75 per worker charged by MTurk for selecting Masters status participants. The survey was carried out in batches, each containing nine participants, and batches were published until at least 15 accepted submissions were obtained for each pair of explanations.
In total, 129 submissions were collected, of which 94 were approved. Submission acceptance was based on four criteria: the two text-based questions, the attention check question, and the Likert scale responses. Submissions were rejected if workers failed the attention check, provided responses in the open-text fields that did not align with the questions asked, submitted responses that appeared to be AI-generated, or selected the same response for all Likert scale items. Table 4.4 lists the first five rejected submissions along with the question that led to their rejection and the feedback given to the participants. Participants were given 30 minutes to complete the survey, and among the approved submissions, workers spent 18.64 minutes on average on the task.

Explanation method preference	**Requester feedback**
It was an important term used for representing words for text analysis in the form of real values...	The answers given in the open text do not match the questions
Shallow neural network models with an input layer, hidden layer, word embeddings, and output layer...	Answers do not match questions
The explanation method is a teaching technique where the instructor breaks down complex concepts into manageable parts...	The answers given in the open text do not match the questions
Some of the best techniques for learning include retrieval practice, spaced practice, and collaborative learning...	The answers given in the open text do not match the questions
CBOW uses a shallow neural network with a single hidden layer. The input layer consists of the aggregated embeddings...	The answers given in the open text do not match the questions

Table 4.4: Reasons for rejection of submissions, including truncated responses to the question "Which of the two explanation methods did you find more useful and why" alongside the corresponding rejection reasons. The responses were overly verbose and ultimately did not address the question properly.

Multiple choice questions were graded using the formula in Equation 4.7. Scores below zero were adjusted to zero.

$$\text{Score} = \frac{\text{Correct Selections}}{\text{Total Correct Answers}} - \frac{\text{Incorrect Selections}}{\text{Total Possible Incorrect Answers}} \tag{4.7}$$

Limitations in the Study

This subsection addresses two limitations that arose during the execution of the study and outlines the measures taken to address them.

1. **Missing Questions**

   In the combination of the Surrogate Model and Partial Dependence Plots (PDP) explanation methods, three questions were inadvertently omitted. These questions pertained to participants'

interest in machine learning explanation methods, factors influencing life satisfaction, and confidence in their mathematical abilities. This omission was identified after the data collection phase, and submissions with missing data in these columns were excluded from RQ2.

In the combination of SHAP and Partial Dependence Plots (PDP), the second question regarding the applied explanation method and the question about subjective helpfulness were omitted. To address this, submissions with missing data in these columns were excluded from RQ3 and RQ4, and additional batches were published to ensure at least 15 valid samples for this combination. As a result, the SHAP–Partial Dependence Plots (PDP) combination had 31 observations, while the remaining combinations had either 15 or 16 observations.

It is unlikely that the presence or absence of these questions influenced participants' decisions to take part in the study. All survey batches had identical titles and descriptions, and participants were unaware of any differences in content between batches. Consequently, the survey results of the affected submissions were included in analyses for research questions unaffected by the missing data.

2. **Duplicate Participation Across Batches**

   In three cases, submissions from participants who had already completed a survey from a different combination of explanation methods were accepted. Although this overlap was not originally intended, the inclusion of these submissions is justified due to the distinct nature of the questionnaires for each combination. Furthermore, since all batches were presented with identical titles and descriptions, it can reasonably be assumed that participation in one batch is independent of participation in another. Re-collecting the data was deemed unnecessary due to the minimal likelihood of bias affecting the results and the limited scale of the issue.

4.2.5 Research Questions and Hypotheses

The survey seeks to evaluate how effectively the explanation methods were understood, identify factors that affected their comprehension, determine whether the explanation methods improved objective or subjective understanding of the model, explore how a preference for one model over another impacted the perceived overall helpfulness of the explanation methods, and examine whether specific combinations of explanation methods were favored over others.
Although Likert scale measurements represent ordinal data, they were treated as having cardinal properties. Therefore, linear regressions were used instead of ordinal logistic regressions to reduce the number of predictors and for its ease of interpretation, both in terms of the coefficients of the predictors and the coefficient of determination. Each Likert scale question was included as a single predictor, rather than modeling each available response separately.

Research Question 2 (RQ2)

Research Question 2 is aimed to determine whether there were significant differences in the subjective understanding of different explanation methods.

Hypothesis RQ2H1 The subjective understanding of each explanation method depends significantly on a participant's background and the specific method presented to them. Higher education is expected to lead to greater subjective understanding. Additionally, participants who spend more time on the survey are anticipated to demonstrate a better subjective understanding. Interest in the topic—whether related to life satisfaction or machine learning models—is expected to enhance subjective understanding by increasing engagement with the content. Confidence in one's mathematical abilities is also expected to positively impact subjective understanding.

A linear regression model was used to model this relationship . The notation `C(...)` represents dummy variables for categorical data, where one category was automatically dropped to avoid perfect multicollinearity.

$$\begin{aligned} \text{subjective\_understanding} \sim\ & \text{work\_time\_in\_minutes} + C(\text{method}) \\ & + C(\text{age\_range}) + C(\text{education}) \\ & + \text{interest\_ml\_explanations} \\ & + \text{interest\_life\_satisfaction} + \text{math\_confidence} \end{aligned} \tag{4.8}$$

Hypothesis RQ2H2 There exist significant differences in the objective understanding of each explanation method when considering a participant's background. The magnitude and direction of each predictor are anticipated to conform to the influence of features outlined in RQ2H1. Although the objective understanding across explanation methods is expected to follow the same direction, the magnitude is expected to differ.

Objective understanding was measured using a single correct answer, therefore, a logistic regression was used to test this hypothesis.

$$\begin{aligned} \text{objective\_understanding} \sim\ & \text{work\_time\_in\_minutes} + C(\text{method}) \\ & + C(\text{age\_range}) + C(\text{education}) \\ & + \text{interest\_ml\_explanations} \\ & + \text{interest\_life\_satisfaction} + \text{math\_confidence} \end{aligned} \tag{4.9}$$

Hypothesis RQ2H3 There exist significant differences in the subjective understanding of paired explanation methods.

This hypothesis was tested using paired t-tests for all five pairs of methods. Comparisons between methods not used together in a subset were not considered.

$$\begin{aligned} \text{statistic, p\_value} &= \text{ttest\_rel}(\text{data1}, \text{data2}) \\ \text{where } (\text{method1}, \text{method2}) &\in \{(\text{SHAP, PDP}), (\text{SHAP, Surrogate}), \dots\} \end{aligned} \tag{4.10}$$

Research Question 3 (RQ3)

Research Question 3 is aimed to investigate how subjective helpfulness of individual explanation methods influenced overall subjective helpfulness when methods were combined, and whether individual helpfulness scores interacted.

Hypothesis RQ3H1 The subjective helpfulness of each explanation method positively affects the perception of the helpfulness of the explanation combination. Additionally, as the combined usefulness of the explanation methods increases, the incremental benefit to the helpfulness of the combination decreases. .

$$\text{combined\_helpfulness} \sim \text{helpfulness\_sum} + \text{helpfulness\_sum}^2 \tag{4.11}$$

Hypothesis RQ3H2 The more equally both explanation methods contribute to understanding the model, the higher the subjective helpfulness of the combined explanation.

The factor `abs_method_contribution_centered` represents the absolute distance from the midpoint, where both methods contribute equally to the overall helpfulness.

$$\text{abs\_method\_contribution\_centered} = |\text{method\_contribution} - 3| \tag{4.12}$$

In addition, the sum of the helpfulness of each explanation method was included.

$$\text{combined\_helpfulness} \sim \text{helpfulness\_sum} + \text{abs\_method\_contribution\_centered} \tag{4.13}$$

Research Question 4 (RQ4)

Research Question 4 is intended to explore the extent to which an individual's understanding of an explanation method influenced their ability to draw accurate conclusions about the underlying model.

Hypothesis RQ4H1 Subjective and objective understanding of explanation methods has a positive impact on the ability to correctly interpret the explanation method in relation to the model.

A linear regression was used to test this hypothesis, as the `applied_score` was calculated as the sum of the two scores (see Equation 4.7) from the questions where participants applied the two explanation method to the given model (section *Applying the explanation* in Figure 4.7).

$$\begin{aligned} \text{applied\_score} \sim \text{subjective\_understanding} + \text{objective\_understanding} \\ + C(\text{method}) \end{aligned} \tag{4.14}$$

Figure 4.7 illustrates all the questions included in each survey. The explanation-specific section was included twice in each survey, once for each model in the group. Overall, the survey comprises five introductory questions, ten explanation-specific questions (five for each explanation), and four summary questions, totaling 19 questions. Two features used in the hypotheses are not listed here: the dummy variable for each explanation method and the time spent completing the survey before submission.

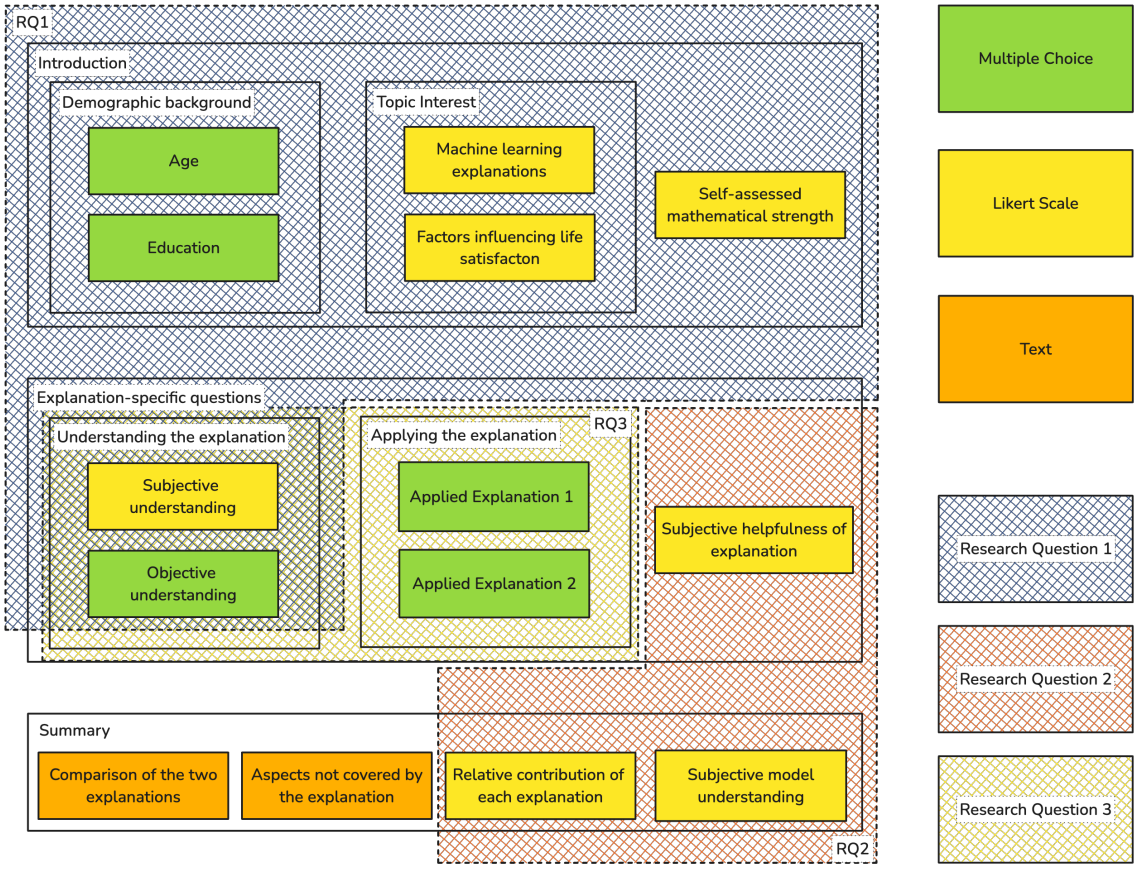


Figure 4.7: Survey Structure

Chapter 5

Findings

This chapter shows the findings from the synthetic data for the three novel approaches as introduced in section 4.1 as well as the user study using all six explanation methods as shown in section 4.2.

5.1 Synthetic Data Evaluation (RQ1)

5.1.1 Embedded SHAP

RQ1H1: Variability in Feature Importance Projections

The function introduced in Equation 4.3 should lead to frequent changes in the embedding space for instances belonging to group B, the embeddings of instances belonging to group A should remain mostly static.

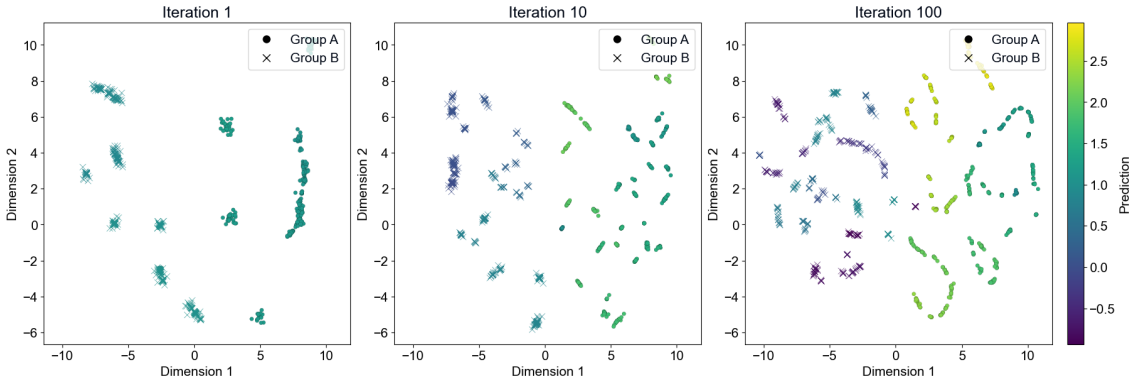


Figure 5.1: Instances belonging to group A correspond to instances of impact function A. Instances belonging to group B correspond to instances of impact function B

After the 10th iteration, the points assigned to the underlying function B have shifted significantly whereas the points assigned to the underlying function A have only shifted slightly. In the final iteration, the instances affected by the underlying function A have aligned in several line segments as seen in Figure 5.1. Some line segments can also be seen in group B.

The positional change in the embeddings is assessed by computing the Euclidean distance between consecutive iterations for each instance, which evaluates the extent of movement in the embeddings from one iteration to the next:

$$\text{Average Euclidean Distances} = \frac{1}{I-1} \sum_{i=1}^{I-1} \sqrt{\sum_{j=1}^{n} (D_{i+1,j} - D_{i,j})^2} \tag{5.1}$$

Where:

- I is the total number of iterations.
- n is the number of features for each instance.

- $D_{i,j}$ represents the j-th feature of the data at iteration i.

Applying Equation 5.1 to the synthetic datasets shows that instances with the impact function defined in Equation 4.1 changed their position by 0.46, whereas instances impacted by Equation 4.2 changed their position by 1.06. This represents a notable difference of approximately 130%. Another metric that highlights the differences in complexity is the number of tree nodes traversed by each instance. With 100 trees and default settings, instances influenced by Equation 4.1 traversed approximately 735 nodes in total, whereas those affected by Equation 4.2 traversed around 931 nodes, which is a percentage difference of approximately 27%. This confirms RQ1H1 for the given function and dataset. The more complex function resulted in greater movement within the embedding space.

RQ1H2: Cluster Formation

The function introduced in Equation 4.4 should lead to a cluster formation of all instances part of the group $\text{Dummy}_{0.05}$. The cluster should only form in later iterations, as it affects only a small fraction of the population.

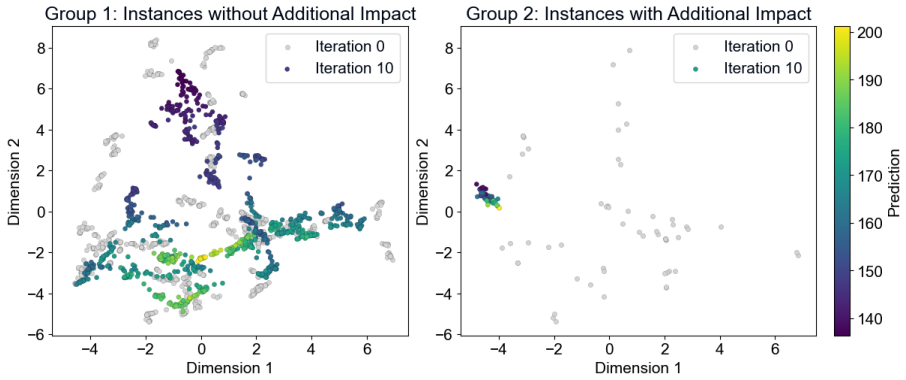


Figure 5.2: Group 1: Instances with $\text{Dummy}_{0.05}$ set to zero. Group 2: Instances with $\text{Dummy}_{0.05}$ set to one. The two images show instances from the same embedding space. For visual clarity, each of the subsets are displayed separately.

As seen in Figure 5.2, the 5% subgroup, influenced by the dummy variable, forms a cluster due to the additional impact. Initially, in the first iteration, the dummy variable was not present in the constructed tree. However, after 10 iterations, all instances with non-zero coefficients for the dummy variable have clustered towards the left side of the image, which confirms RQ1H2. Instances belonging to a subgroup will form a separate cluster once their separating factor is considered and sufficiently large in comparison to the remaining features.

5.1.2 Frequent Pattern Mining (Lift)

The two hypotheses that will be examined are:

- **RQ1H3: Co-occurrence Drives Lift.** Features with strong interaction effects should have larger Lift values.
- **RQ1H4: Interaction Strength Correlates with Lift.** The greater the strength of the interaction, the higher the expected Lift value.

Based on the function introduced in Equation 4.5, the features $InterC$ and $InterD$ should have the largest Lift value, as the interaction $InterCD$ has the largest impact on the target variable y. The interaction $InterAB$ should have the second highest lift value. The lift values of the features $PrimA$, $PrimB$ and $PrimC$ should be close to one, as they have no interaction effect with any other feature.

Features $InterC$ and $InterD$ indeed result in the largest Lift value in the model, appearing in 121 splits. The primary features exhibit practically no positive association with each other, which is expected, as they are uncorrelated with other features and have no joint effects on the target. The second interaction effect also showed the second highest Lift value. Accordingly, the two

	Antecedents	Consequents	Lift	Support
1	interD	interC	6.498389	121
2	interB	interA	5.141388	195
3	interA	interC	4.010025	104
4	interA	interD	3.924647	75
5	interD	interB	3.782593	102
6	interC	interB	3.761331	139
7	primC	primA	1.044772	1267
8	primB	primC	1.024657	1167
9	primA	primB	1.003067	1060

Table 5.1: Frequent 2-itemsets for interaction and primary features with a lift value greater than one.

hypotheses RQ1H3 and RQ1H4 have been confirmed for the given dataset. Features with a strong interaction effect often formed sequences in the tree splits, leading to high Lift values. In this dataset, the strength of the interaction determined the magnitude of the Lift value.

5.1.3 Feature Context Embeddings

The two hypotheses that will be examined are:

- **RQ1H5: Contextual Clustering.** Feature context embeddings should be arranged in a 2D space such that features appearing in similar contexts are clustered.
- **RQ1H6: Global Similarity Optimization.** The overall arrangement should maximize inter-cluster similarity.

Based on the function defined in Equation 4.6, features belonging to Group_1 and features belonging to Group_2 should be clustered separately. The feature *Shared_Feature* should not be clustered with either of the groups, as it affects the target y independently.

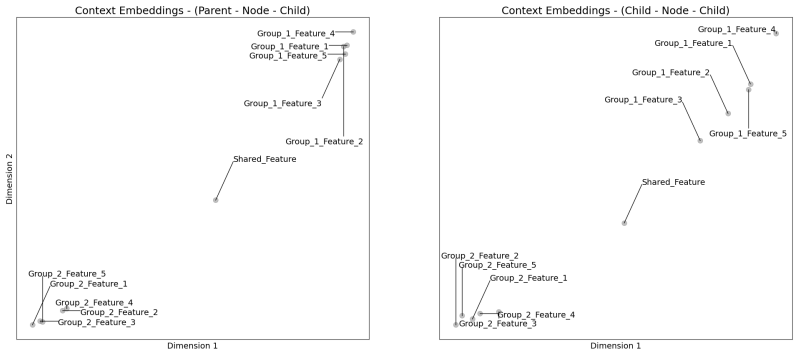


Figure 5.3: Scatter plots illustrating the contextual embeddings of features from each configuration.

As seen in Figure 5.3, the two groups are clearly clustered separately from one another which validates RQ1H5. The configuration including the parent clusters the groups more tightly. Moreover, the shared feature is located in the middle between the two groups, which confirms RQ1H6. The arrangement maximizes inter-cluster similarity.

5.2 Real-World Evaluation (RQ2 - RQ4)

5.2.1 Descriptive Analysis

The custom explanation methods consistently demonstrated notable performance differences both in understanding and in helpfulness metrics (see Figure 5.4). These methods occupy both the top

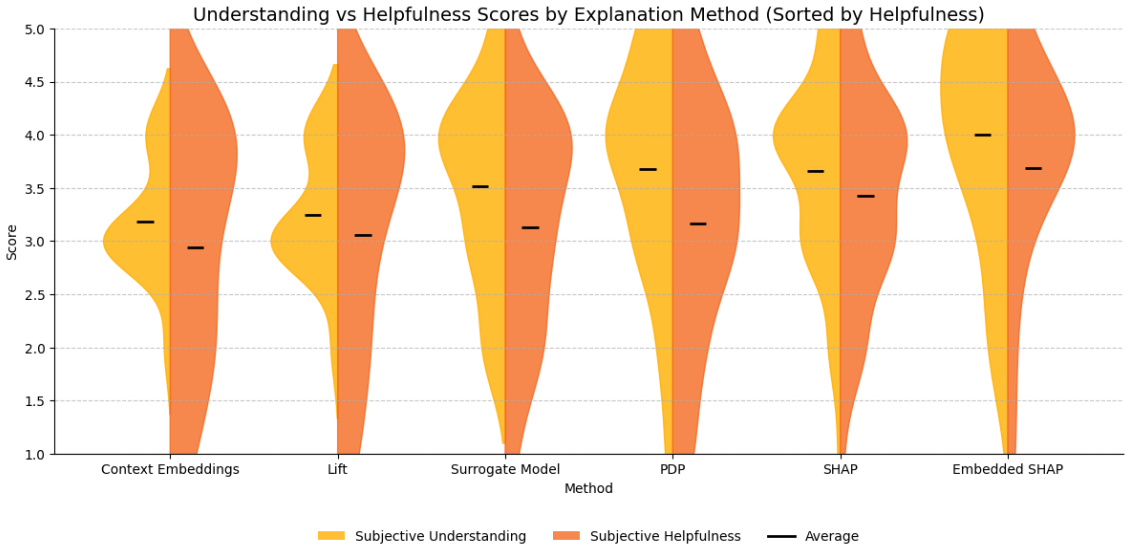


Figure 5.4: Explanation methods with higher understanding scores were generally rated as more useful. Note that the y-axis starts at 1, which corresponds to the lowest point on the Likert scale. Helpfulness Score: 1 = Did not help in understanding the model, 5 = Completely clarified the model. Understanding Score: 1 = Did not understand the explanation, 5 = Completely understood the explanation.

and bottom positions in the hierarchy of understanding and helpfulness. Specifically, Embedded SHAP was found to be the most understood (Ø 4.00) and helpful (Ø 3.69) method, while Context Embeddings (CBOW) scored the lowest in both understanding (Ø 3.19) and helpfulness (Ø 2.94). For each individual explanation method, individuals with higher education levels demonstrated a higher average subjective understanding. Participants with a high school diploma had the lowest average understanding scores, followed by those with a bachelor's degree, and then by participants with a master's degree or higher. This aligns with the findings from [13] where Master students reported higher self-reported understanding. This also holds true for the subjective helpfulness scores that participants assigned to each individual explanation method. The higher the education level, the higher the subjective helpfulness of the explanation method. The only exceptions are SHAP and Context Embeddings, where participants with a bachelor's degree had slightly higher averages (Embedded SHAP: Bachelor = Ø 4.14, Master = Ø 4.00; Context Embeddings: Bachelor = Ø 3.42, Master = Ø 3.25).

Enhanced mathematical confidence correlates with higher subjective understanding for several explanation methods. SHAP understanding scores increase from an average of 2 at the lowest confidence level to 4.25 at the highest, and Partial Dependence Plots also exhibit improved understanding scores as confidence rises. The Surrogate Model, evaluated only among intermediate groups (since no participants reported a confidence of 1 or 5), shows an increase from an average of 2.5 to 3.33 at the highest measured level, as does Embedded SHAP, which rises from 3.0 at a confidence level of 3 to 4.23 at a level of 4, with no data available for other levels. Understanding scores for the Context Embeddings and Lift remain relatively consistent, with subjective understanding scores fluctuating between 3 and 4 regardless of math confidence.

A clear correlation exists between the perceived understandability of a model and its perceived helpfulness. However, across all methods, models scored consistently higher in understanding than in helpfulness. This suggests that while participants could grasp the underlying explanatory mechanisms, they found the methods less directly applicable in explaining model behavior. An additional explanation for the discrepancy between understanding and helpfulness is that some explanation methods contain more information than others. In the case of Partial Dependence Plots, only three plots were shown, despite the model containing numerous features. The most pronounced disparity between understanding and helpfulness was observed in the Partial Dependence Plots. Despite achieving a relatively high average understanding score of 3.68, these plots were rated significantly lower in terms of helpfulness at an average of 3.17.

The two methods that demonstrated the greatest individual helpfulness also exhibited the highest level of enhanced understanding of the model. (SHAP and Embedded SHAP). Additionally, the combination of Lift and Context Embeddings, which both relate to the context in which a feature is utilized, is perceived as more beneficial. A potential explanation for similar explanations that

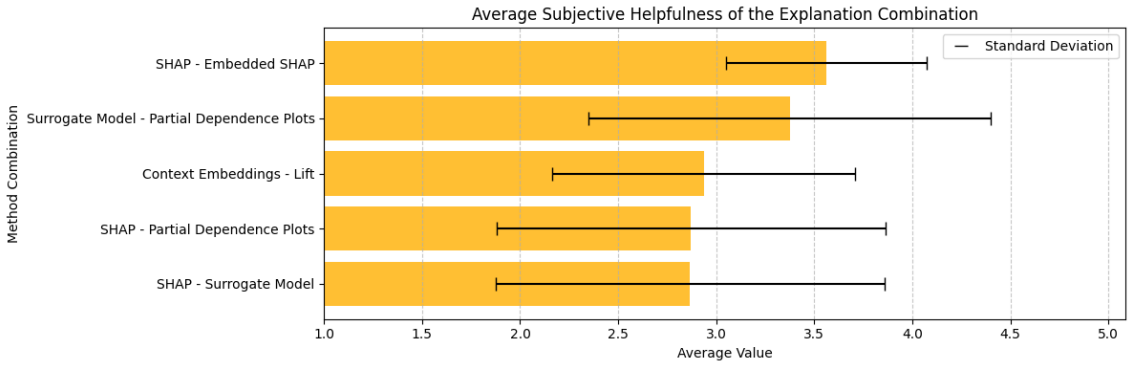


Figure 5.5: Comparison of average subjective helpfulness ratings for different combinations of explanation methods. The chart shows how participants perceived the combined effectiveness of two methods in enhancing their understanding of the model compared to the potential contribution of a single method. 1 means the combination did not enhance model understanding at all, 5 means the combination led to a complete or near-complete understanding of the model

receive high scores in terms of enhanced understanding of the model is that individuals prefer further clarification of a specific aspect of the model over having two distinct explanations that cover different perspectives on the model.

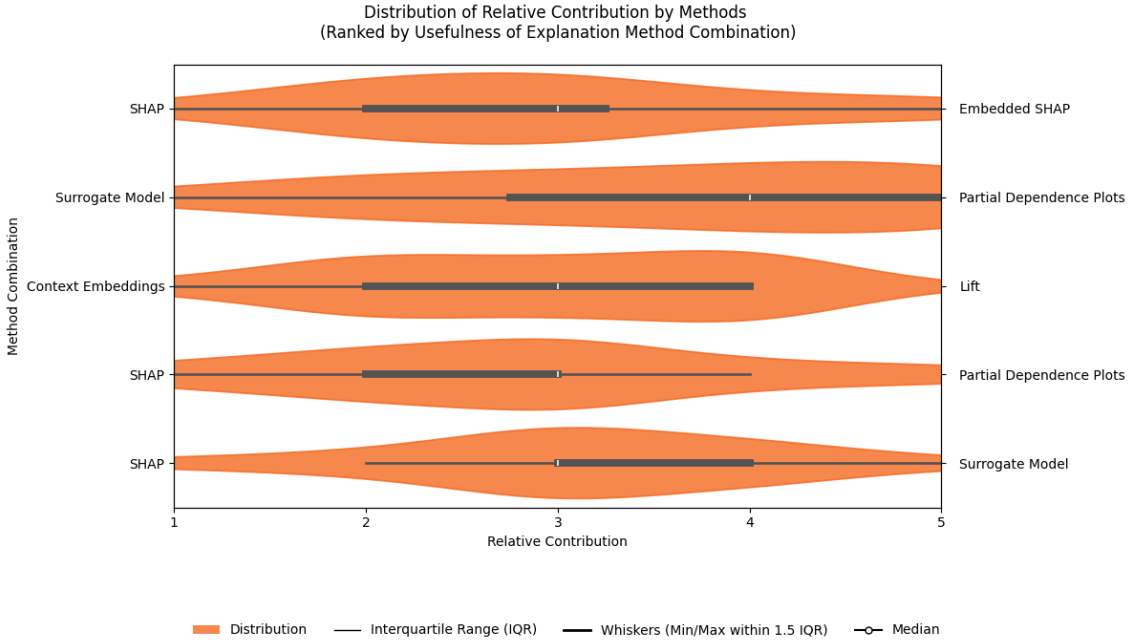


Figure 5.6: Distribution of relative contribution ratings grouped by method combination. Each violin shows the kernel density estimate of the ratings. 1 means the first method contributes significantly more than the second, 5 means the second method contributes significantly more than the first

Figure 5.6 shows that the contribution of each method follows a distribution similar to a normal curve. In particular, all explanation combinations, except for the pair Context Embeddings - Lift, appear to have a normal distribution. Although Embedded SHAP had a higher average understanding score as seen in Figure 5.4, it contributed less to the overall understanding of the model, as the distribution is shifted closer to SHAP.
Figure 5.7 indicates that most participants held a bachelor's degree or equivalent, with only one participant holding a doctorate. As a result, the categories "Master's degree or equivalent" and "Doctorate degree or equivalent" were merged into a single category, "Master's degree or higher." The median age of the participants was in the 36 to 40 range. Since only one participant fell into the 21 to 25 age group, the age brackets 21 to 25 and 26 to 30 were combined into a single bracket, 21 to 30.

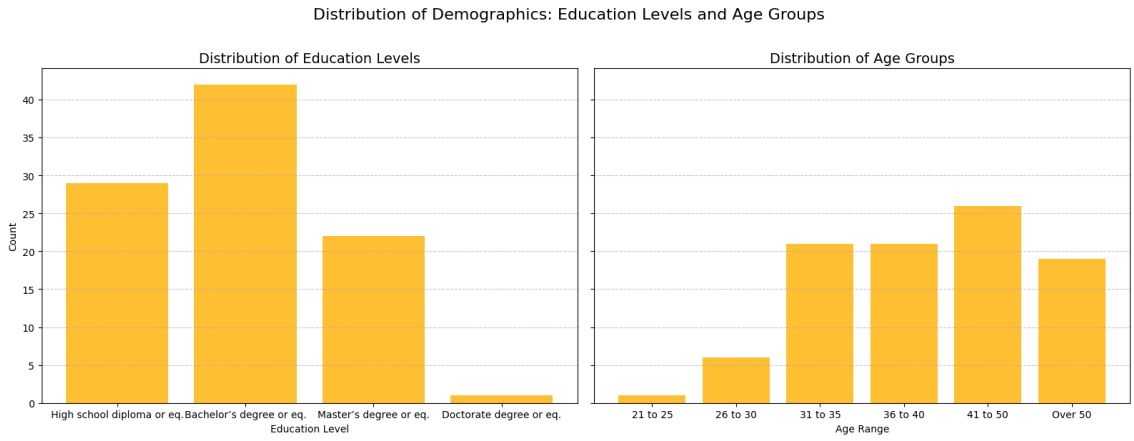


Figure 5.7: Education level and age range.

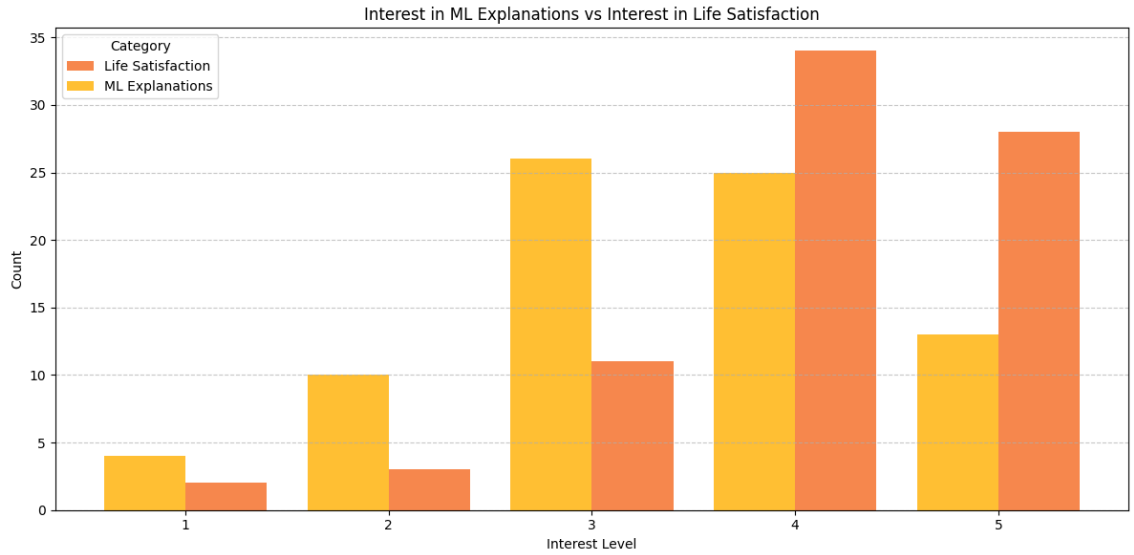


Figure 5.8: Interest in machine learning and life satisfaction.

Figure 5.8 indicates that the participants were generally more interested in life satisfaction than in machine learning explanations. On average, participants rated their interest in machine learning explanations at 3.42, while their interest in life satisfaction was higher at 4.06. Additionally, there is a moderate positive correlation ($r = 0.45$) between the participants' interest in machine learning explanations and their interest in life satisfaction.

5.2.2 Empirical Analysis

To test the hypotheses presented in subsection 4.2.5, logistic and linear regression models were employed.
Linear regression models were evaluated to ensure that they satisfy five key assumptions of linear regression models [72], which are outlined in Table 5.2:

- **Linearity:** To test whether the linear relationship assumption held, RESET (Regression Specification Error Test) [65] tests were performed to assess whether any higher-order predictors were significant. The test was performed using quadratic terms, while cubic and higher-order terms were not considered. If no predictor was significant at the 5% significance level, linearity was assumed.
- **Normality:** Normality was tested using a Shapiro-Wilk test [76], which compares the ordered residuals with a normal distribution.
- **Multicollinearity:** Multicollinearity was assessed using the Variance Inflation Factor (VIF) [54], which was calculated by regressing each predictor on all the other predictors and measuring how much the variance of its coefficient was inflated due to multicollinearity.

- **Independence:** Independence of the residuals was assumed for regressions with a single observation per participant, as there was no reason to assume that one participant's responses could depend on another participant's responses. Although three workers participated twice, their impact on overall independence was considered insignificant. When there was more than one observation per worker, because each worker received two explanation methods, the independence of the pairs was tested using a Pearson correlation test.
- **Homoscedasticity:** Homoscedasticity was tested using the Breusch-Pagan test [14]. This test regresses the squared residuals of the model on the predictors (or functions of them). If the predictors explained a significant amount of variance in the squared residuals, this suggested heteroskedasticity.

Identifier	Assumption	Test
Linearity	Predictors relate linearly to the target.	Ramsey's RESET Test
Normality	Residuals are normally distributed.	Shapiro-Wilk Test
Multicollinearity	Predictors lack strong intercorrelation.	Variance Inflation Factor
Independence	Residuals are independent.	Pearson Correlation Test
Homoscedasticity	Residual variance is constant across predictors.	Breusch-Pagan Test

Table 5.2: Tests for Linear Regression

Logistic regression relies on the following assumptions [72]:

- **Binary Target:** The target variable has to be binary. This assumption was not explicitly tested because the analysis was restricted to binary outcomes.
- **Independence:** Residuals are expected to be independent. As in linear regression, Pearson Correlation tests were conducted when there were two observations per participant, as in the case of having two explanation methods per participant.
- **Multicollinearity:** Multicollinearity was assessed using the Variance Inflation Factor to ensure that the predictors were not highly correlated.
- **Linearity:** The predictors are required to have a linear relationship with the log-odds of the outcome. This was tested by examining interactions with log-transformed predictors.
- **Sample Size:** The assumption of adequate sample size was considered satisfied. The smallest group (participants aged 21 to 30) had 7 observations, which was considered sufficient.

Identifier	Assumption	Test
Independence	Residuals are independent.	Pearson Correlation Test
Multicollinearity	Predictors lack strong intercorrelation.	Variance Inflation Factor
Linearity	Predictors relate linearly to log-odds.	Log-transformed interactions

Table 5.3: Tests for Logistic Regression

RQ2H1

Table 5.4 shows that the combinations of Lift / Context Embeddings and the Surrogate Model differ significantly from the reference group, Embedded SHAP. Lift and Context Embeddings were combined into a single dummy variable because they only appeared in a single combination, resulting in perfect collinearity. All explanation methods had negative coefficients, consistent with the observation that Embedded SHAP was subjectively better understood than any other explanation method. Given that the target variable is a Likert scale ranging from 1 to 5, the coefficients indicate that the understanding of the Surrogate Model is approximately 0.76 points lower than that of Embedded SHAP.

Variable	Coeff.	Std. Error	z-value	P-value
Intercept	**1.6962**	**0.6019**	**2.8180**	**0.0056**
C(method)[T.Lift/Context_Embeddings]	**-0.8002**	**0.2389**	**-3.3491**	**0.0011**
C(method)[T.PDP]	-0.2969	0.2398	-1.2382	0.2179
C(method)[T.SHAP]	-0.1099	0.2240	-0.4905	0.6246
C(method)[T.Surrogate]	**-0.7561**	**0.2806**	**-2.6947**	**0.0080**
C(age_range)[T.31 to 35]	0.3141	0.4224	0.7437	0.4584
C(age_range)[T.36 to 40]	0.6292	0.4186	1.5031	0.1353
C(age_range)[T.41 to 50]	0.6278	0.4089	1.5356	0.1271
C(age_range)[T.Over 50]	0.2684	0.4149	0.6469	0.5189
C(education)[T.High school diploma or equivalent]	-0.0769	0.1637	-0.4699	0.6393
C(education)[T.Master's degree or higher]	**0.6564**	**0.1719**	**3.8185**	**0.0002**
work_time_in_minutes	0.0225	0.0116	1.9466	0.0538
interest_ml_explanations	**0.1480**	**0.0743**	**1.9927**	**0.0485**
interest_life_satisfaction	0.0265	0.0821	0.3225	0.7476
math_confidence	0.1610	0.0879	1.8311	0.0695

Table 5.4: Results from the Linear Regression Model. Significant predictors ($p < 0.05$) are highlighted in bold. The target is subjective understanding of the explanation method. The model explains 35.3% of the variance in the dependent variable ($R^2 = 0.353$).

In contrast, none of the dummy variables for age were significant, with the age bracket of 21 to 30 serving as the reference group. While there is some indication that subjective understanding is higher in higher age brackets, the effect is not significant.
Education, however, has a significant influence on subjective understanding. A master's or higher degree leads to a significant increase in perceived understanding.
There appears to be a small but positive effect of work time on subjective understanding, an additional 10 minutes of work time increases subjective understanding by 0.23 points. Interest in the topic, both in terms of interest in machine learning explanations and interest in life satisfaction, had a positive effect on the target. However, only interest in machine learning explanations had a statistically significant effect.
Confidence in one's mathematical abilities falls just outside the range of statistical significance, suggesting some evidence that higher confidence may lead to higher subjective understanding of the machine learning explanation methods.

Assumption	Test	Condition	Value
Linearity in Model Form	Ramsey's RESET Test	P-value > 0.05	0.2021 (Pass)
Homoscedasticity of Residuals	Breusch-Pagan Test	P-value > 0.05	0.0861 (Pass)
Normality of Residuals	Shapiro-Wilk Test	P-value > 0.05	0.2631 (Pass)
No Predictor Multicollinearity	Variance Inflation Factor	Max VIF < 10	8.7074 (Pass)
Independence of Residuals	Pearson Correlation	P-value > 0.05	0.1126 (Pass)

Table 5.5: Diagnostic checks for RQ2H1

The diagnostic checks summarized in Table 5.5 confirm that the assumptions of linear regression are met. All tests satisfy the predefined thresholds. There is no indication of non-linearity, the residuals are independent and approximately normally distributed, though there is mild evidence of heteroscedasticity ($p = 0.0861$). Furthermore, no significant multicollinearity is detected among the predictors. There appears to be a slight positive correlation between observations for the same

participants, but it is not statistically significant.

RQ2H2

Variable	Coeff.	Std. Error	z-value	P-value
Intercept	-0.8625	1.8533	-0.4654	0.6417
C(method)[T.Lift/Context_Embeddings]	1.0268	0.7082	1.4499	0.1471
C(method)[T.PDP]	**1.5501**	**0.7231**	**2.1439**	**0.0320**
C(method)[T.SHAP]	**2.3439**	**0.7191**	**3.2597**	**0.0011**
C(method)[T.Surrogate]	1.1415	0.8314	1.3730	0.1698
C(age_range)[T.31 to 35]	1.1485	1.2808	0.8967	0.3699
C(age_range)[T.36 to 40]	1.6914	1.2894	1.3118	0.1896
C(age_range)[T.41 to 50]	0.5093	1.2261	0.4153	0.6779
C(age_range)[T.Over 50]	1.2324	1.2573	0.9802	0.3270
C(education)[T.High school diploma or equivalent]	-0.4324	0.5472	-0.7902	0.4294
C(education)[T.Master's degree or higher]	**-1.5547**	**0.5423**	**-2.8670**	**0.0041**
work_time_in_minutes	-0.0232	0.0367	-0.6328	0.5269
interest_ml_explanations	-0.1277	0.2562	-0.4984	0.6182
interest_life_satisfaction	0.1278	0.2727	0.4686	0.6393
math_confidence	0.0656	0.2853	0.2300	0.8181

Table 5.6: Results from the Logistic Regression Model. Significant predictors ($p < 0.05$) are highlighted in bold. The target is objective understanding of the explanation method. The model's pseudo R^2 is 0.1726.

The results presented in Table 5.6 show that among the explanatory methods, Partial Dependence Plots (PDP) and SHAP show significant differences from the reference group, Embedded SHAP. Both PDP ($p = 0.0320$) and SHAP ($p = 0.0011$) have positive coefficients, in contrast to the results in Table 5.4.
The combinations of Lift / Context Embeddings and Surrogate Model have positive coefficients, but are not significantly different from the reference group. None of the age groups (31 to 35, 36 to 40, 41 to 50, and over 50) differ significantly from participants aged 21 to 30 in predicting objective understanding. Education appears to be a significant factor, with a Master's degree or higher associated with a decrease in subjective rating ($p = 0.0041$, coefficient $= -1.5547$). This finding contradicts RQ2H1, where a Master's degree or higher was associated with an increase in subjective understanding compared to a Bachelor's degree or equivalent. In contrast, participants with a high school diploma or equivalent do not show significant differences in objective understanding compared to those with a bachelor's degree.
Other variables, including work time ($p = 0.5269$), interest in machine learning explanations ($p = 0.6182$), interest in life satisfaction ($p = 0.6393$), and confidence in mathematical abilities ($p = 0.8181$), do not significantly affect the result. However, both work time and interest in machine learning explanations had a significant effect on subjective understanding in RQ2H1.
.
The logistic regression assumptions were tested (Table 5.7) and all were met as multicollinearity showed no significant concerns with a maximum VIF of 8.7074, while linearity in log-odds passed with a minimum p-value of 0.4785, and residual independence was confirmed with a correlation of -0.1543 and a p-value of 0.2391.

RQ2H3

Table 5.8 indicates that among the five combinations of explanation methods, only SHAP and the Surrogate Model showed a significant difference in subjective understanding. Within this pair, the

Assumption	Test	Condition	Value
No Predictor Multicollinearity	Variance Inflation Factor	Max VIF < 10	8.7074 (Pass)
Linearity in Log Odds	Log-transformed interactions	P-value > 0.05	0.4785 (Pass)
Residuals are independent	Pearson Correlation	P-value > 0.05	0.2391 (Pass)

Table 5.7: Diagnostic checks for RQ2H2.

Method 1	Method 2	Statistic	p-value
SHAP	PDP	0.764	0.457
SHAP	**Surrogate**	**2.553**	**0.023**
PDP	Surrogate	1.168	0.261
Context Embeddings	Lift	-0.436	0.669
SHAP	Embedded SHAP	0.000	1.000

Table 5.8: Paired t-test results for statistical equality of subjective understanding across explanation method combinations. The table shows the t-statistic and corresponding p-value for each method pair. Significant predictors ($p < 0.05$) are highlighted in bold.

SHAP values were on average considerably better understood than the Surrogate Model (3.73 vs. 3.13). SHAP and Embedded SHAP, on the other hand, had identical average values (4) in their group, resulting in a p-value of 1. Therefore, although Table 5.5 suggests a statistically significant difference in subjective understanding, the difference is only explained by the combinations of the explanation method SHAP - PDP and SHAP - Surrogate but is not present in the SHAP - Embedded SHAP combination.

RQ3H1

Variable	Coefficient	Std. Error	z-value	P-value
Intercept	0.1128	0.6720	0.1680	0.8670
helpfulness_sum	**0.6289**	**0.2230**	**2.8240**	**0.0060**
I(helpfulness_sum ** 2)	-0.0239	0.0180	-1.3390	0.1850

Table 5.9: Results from the Linear Regression Model. The dependent variable is the combined helpfulness of the two explanation methods. Significant predictors ($p < 0.05$) are highlighted in bold. The model explains 47.1% of the variance in the dependent variable ($R^2 = 0.471$).

Table 5.9 shows that increases in the sum of perceived helpfulness scores (`helpfulness_sum`) significantly correspond to an increase in combined helpfulness ($p = 0.006$). However, the squared term for `helpfulness_sum`, $I(\texttt{helpfulness\_sum}^2)$, is not statistically significant ($p = 0.185$), suggesting no evidence of a nonlinear relationship, despite showing the anticipated negative coefficient. The intercept is also not significant, indicating no baseline level of understanding unrelated to the predictors.

According to Table 5.10, the Shapiro–Wilk test indicates that the residuals do not strictly follow a normal distribution. As shown in the kernel density plot (Figure 5.9), the residuals remain slightly skewed with two local maxima but appear broadly normal overall despite the test result. The variance inflation factor (28.9248) exceeds the usual threshold of 10, which is unsurprising given that both predictors stem from the helpfulness measures of individual explanation methods. The remaining assumptions—linearity and no endogeneity—are satisfied.

Assumption	Test	Condition	Value
Linearity in Model Form	Ramsey's RESET Test	P-value > 0.05	0.5608 (Pass)
Homoscedasticity of Residuals	Breusch-Pagan Test	P-value > 0.05	0.8444 (Pass)
Normality of Residuals	Shapiro-Wilk Test	P-value > 0.05	0.0001 (Fail)
No Predictor Multicollinearity	Variance Inflation Factor	Max VIF < 10	28.924 (Fail)

Table 5.10: Diagnostic checks for RQ3H1.

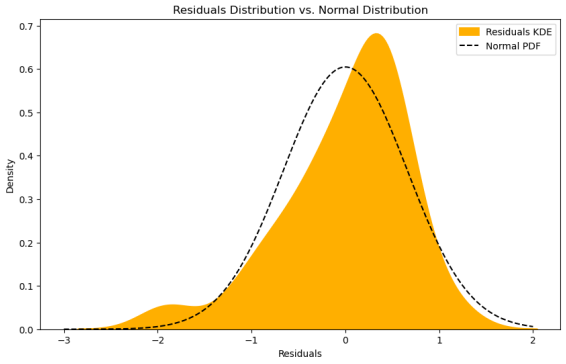


Figure 5.9: Distributon of the Residuals from RQ3H1.

RQ3H2

Variable	Coefficient	Std. Error	z-value	P-value
Intercept	**1.0645**	**0.2988**	**3.5632**	**0.0006**
helpfulness_sum	**0.3345**	**0.0413**	**8.0933**	**0.0000**
abs_method_contribution_centered	-0.1499	0.1027	-1.4590	0.1487

Table 5.11: Results from the Linear Regression Model. The dependent variable is the combined helpfulness of the two explanation methods. Significant predictors ($p < 0.05$) are highlighted in bold. The model explains 47.3% of the variance in the dependent variable ($R^2 = 0.473$).

Table 5.11 shows that while `helpfulness_sum` remains a significant predictor ($p \sim 0$), the deviation from the midpoint where the two explanation methods contribute equally is not significant, although it has the expected coefficient, indicating that there is some positive effect of the two explanation methods contributing equally to overall understanding.

Assumption	Test	Condition	Value
Linearity in Model Form	Ramsey's RESET Test	P-value > 0.05	0.1752 (Pass)
Homoscedasticity of Residuals	Breusch-Pagan Test	P-value > 0.05	0.9549 (Pass)
Normality of Residuals	Shapiro-Wilk Test	P-value > 0.05	0.0002 (Fail)
No Predictor Multicollinearity	Variance Inflation Factor	Max VIF < 10	1.0006 (Pass)

Table 5.12: Diagnostic checks for RQ3H2.

Table 5.12 is consistent with the results of Table 5.10. The normality assumption for the residuals still does not hold. The distribution appears to be approximately identical to the distribution seen in Figure 5.9 and still appears to be generally normal.

Variable	Coefficient	Std. Error	z-value	P-value
Intercept	**1.9913**	**0.2647**	**7.5238**	**0.0000**
C(method)[T.Lift/Context_Embedding	**-0.7660**	**0.1893**	**-4.0473**	**0.0001**
C(method)[T.PDP]	**-0.8070**	**0.1767**	**-4.5680**	**0.0000**
C(method)[T.SHAP]	**-0.5810**	**0.1801**	**-3.2256**	**0.0015**
C(method)[T.Surrogate]	**-0.8922**	**0.1874**	**-4.7609**	**0.0000**
subjective_understanding	**-0.1253**	**0.0531**	**-2.3611**	**0.0194**
objective_understanding	0.1100	0.1018	1.0805	0.2815

Table 5.13: OLS regression results. The dependent variable is the applied score from Equation 4.14. Significant predictors ($p < 0.05$) are highlighted in bold. The model explains 15.6% of the variance in the dependent variable ($R^2 = 0.156$).

RQ4H1

Table 5.13 indicates that the ability to draw correct conclusions from individual explanation methods strongly depends on the specific explanation method used. All explanation methods differ significantly from the reference group, Embedded SHAP, consistent with the findings in Equation 4.8 and Figure 5.4. Subjective understanding has a significant negative effect on the correct application of explanation methods. Although objective understanding appears to have a positive influence on making correct conclusions, this effect is not statistically significant at conventional thresholds ($p = 0.2815$).

5.2.3 Qualitative Analysis

The following section summarizes the two open text questions at the end of the survey, which were not included in the three research questions (see Figure 4.7).

SHAP - Embedded SHAP

Among the participants who received both SHAP and Embedded SHAP, half of them (8 out of 16) felt that the explanations were adequate. Three participants expressed a desire for more examples, while the remaining participants either wished for a more in-depth explanation or found the explanations too technical. In terms of preference, SHAP values were favored by 9 participants, while 6 participants preferred Embedded SHAP. The primary reason for preferring SHAP was the ease of understanding its general concept. One participant had no preference. Those who favored Embedded SHAP appreciated the ability to visually interpret the proximity between points rather than relying on bar charts. Those findings are in line with the relative contribution to the understanding of the model as seen in Figure 5.6 where participants attributed a higher contribution to SHAP.

Lift - Context Embeddings

For the participants who received Lift values and Context Embeddings, five individuals found the explanations to be adequate. However, two participants expressed confusion about how the embeddings were constructed based on decision trees. The remaining participants either found the explanations insufficient, overly complex, or lacking in certain aspects, such as higher-order interactions. Lift and Context Embeddings were therefore the least understood combination out of all six combinations. This matches the findings in Figure 5.4. In terms of preference, 7 participants favored Lift, while 6 preferred Context Embeddings. Three participants expressed no preference. Those who preferred Lift valued its simplicity in providing a single interpretable number, whereas those who favored Context Embeddings appreciated its ability to visually illustrate the proximity between features.

SHAP - Partial Dependence Plots

Among the participants who were introduced to both SHAP values and Partial Dependence Plots, those with a clear preference favored SHAP twice as often as PDP, with 6 participants preferring

SHAP compared to 3 for PDP. However, 6 participants expressed uncertainty or did not explicitly indicate a preference. Regarding the adequacy of explanations, 8 participants felt that the models were sufficiently explained. One participant found the explanation somewhat adequate but required a simpler presentation. Five participants believed that important aspects of the models were missing or unclear. Common concerns included a lack of examples, difficulty understanding feature interactions, insufficient clarity in graph-based explanations, and challenges in predicting model outputs. Additionally, one participant stated that they lacked enough information to properly assess the explanations.

Partial Dependence Plots - Surrogate Model

For participants who received explanations of Partial Dependence Plots and the Surrogate Model, 7 individuals preferred PDP due to its visual clarity and intuitive interpretation, while 4 preferred the Surrogate Model for its logical structure. Two participants found both methods equally useful, and three had unclear preferences. In terms of explanation adequacy, 8 participants felt that the models were well explained, whereas 1 participant found the explanation somewhat adequate but expressed confusion regarding the Surrogate Model. Five participants believed that aspects of the models remained unclear, due to difficulties in understanding numerical values in trees, feature correlations, and linking different graphs and features. Additionally, 2 participants provided unclear responses, making general statements about model behavior without indicating clarity.

SHAP - Surrogate Model

Among the participants who received explanations of SHAP values and the Surrogate Model, 7 participants preferred the Surrogate Model for its logical structure and clarity, while 5 preferred SHAP due to its ability to simplify individual feature importance. Three participants did not express a clear preference. Regarding explanation adequacy, 9 participants felt that the model's behavior was well explained, while 3 participants found the explanations somewhat adequate but struggled with understanding feature interactions. Two participants indicated that aspects of the model were not sufficiently explained, stating that more step-by-step guidance or better insights into complex nodes might improve the explanation. One participant provided an unclear response that did not indicate whether they fully understood the explanations.

5.3 Summary

5.3.1 General Applicability of the Novel Explanation Methods

The findings from Research Question 1 show that the novel explanation approaches could successfully uncover relationships in the synthetic data with respect to the target.
In particular, the evaluation of the Embedded SHAP approach demonstrated that the method is sensitive to function complexity, instances influenced by more complex functions exhibited significantly greater movement within the embedding space compared to those affected by simpler functions. Additionally, clusters of small subgroups formed as anticipated.
The observed Lift values aligned with the predicted outcomes: features involved in stronger interactions yielded higher Lift values and primary features without interaction effects maintained Lift values near one.
Finally, the analysis of the Context Embeddings showed that the embeddings are organized based on context similarity. The clear separation of features into distinct clusters demonstrated that the method accurately captured the underlying relationships among feature groups.

5.3.2 Impact of Education and Stakeholder Background on Understanding

The findings from Research Question 2 demonstrate a significant relationship between participants' educational backgrounds and their subjective and objective understanding of explanation methods. Higher education consistently corresponds to increased subjective understanding and perceived helpfulness of explanation methods (subsection 5.2.1). However, objective understanding does not follow this trend: participants with a master's degree exhibited lower objective understanding scores despite higher subjective assessments. This discrepancy may result from an illusion of explanatory depth, where individuals with higher education overestimate their abilities, as supported by the statistically significant shorter survey completion times for this group ($p = 0.0284$) [68].

Additionally, increased mathematical confidence is generally correlated with improved subjective understanding.

5.3.3 Stakeholder-Specific Relevance

With respect to the relevance of the novel approaches, Embedded SHAP is not considered relevant for most stakeholders. The movement between iterations in Embedded SHAP was mentioned only once as a reason for preference. Additionally, it primarily reveals model construction details, which are mainly relevant to developers rather than users or affected parties (Langer et al. [40]).
Lift and Context Embeddings, with their focus on inter-feature relationships, appear primarily relevant for regulators concerned with model transparency and developers prioritizing model performance. Neither of the two explanation methods provides a detailed explanation for the model overall. Using both explanation methods along with other explanations in a multidimensional framework such as the one proposed by Mohseni et al. [52] and Langer et al. [40] could enhance overall helpfulness scores. Both papers point out the importance of tailoring explanations to different stakeholder groups, such as AI novices and experts. While profession itself was not explicitly measured in this survey, education and self-assessed mathematical confidence likely correlate with professional expertise, although no single explanation method was exclusively understood by a particular educational group.

5.3.4 Subjectively Best and Worst Understood Methods

Participants consistently demonstrated strong subjective understanding of SHAP values and their embeddings, and rated them highly for helpfulness regardless of educational background. SHAP values were notably better understood and perceived as more helpful than the Surrogate Model and Partial Dependence Plots (PDP). The differences observed in subjective understanding between SHAP and the Surrogate Model were statistically significant (Table 5.6, Table 5.8).
In contrast, Lift and Context Embeddings were poorly understood and considered less helpful and received the lowest subjective scores overall. The technical complexity of Lift, presented with its mathematical formula, likely intimidated participants and negatively impacted their subjective perception of both Lift and Context Embeddings, which were presented together. Aligning explanation complexity with the target audience's expertise, as suggested by Lage et al. [38], who argue for simplicity in explanations for laypeople, might help in this regard.

5.3.5 Consistency and Abstraction Level Across Explanation Methods

The results from Research Question 2 show strong differences in subjective and objective understanding across methods. This difference can likely be attributed to varying abstraction depths, which refers to the level of detail and complexity provided when describing and presenting an explanation method to participants. Previous research on cognitive load theory suggests that the abstraction level significantly impacts learning outcomes [80]. Embedded SHAP, for instance, received high subjective understanding scores but notably lower objective understanding scores compared to traditional SHAP. Furthermore, the questions assessing objective understanding might also differ with respect to their difficulty. Evaluating other intrinsic properties of explanation methods, such as consistency and continuity through controlled perturbations, might provide a better picture of the overall quality of the explanation methods [5]. However, assessing these properties might be challenging for explanations based solely on embeddings, as their interpretability is less direct and measurable.
The findings from Research Question 4 showed that there is a weak relationship between understanding an explanation and the ability to make correct assessments based on it. Similar to the questions regarding objective understanding, these questions may vary in difficulty.

5.3.6 Relevance and Combination of Explanation Methods

In Research Question 3, the overall helpfulness of the explanation methods was effectively predicted by their individual ratings, although the observed diminishing effect did not reach statistical significance (Table 5.9).
Notably, combinations such as SHAP with Embedded SHAP resulted in the highest perceived combined helpfulness. Overall, participants indicated that explanation combinations ranged from "slightly" to "significantly" enhancing their understanding of the model.

Furthermore, combining PDP with the Surrogate Model significantly increased overall model understanding, despite neither method individually receiving high subjective understanding scores, which suggests a complementary effect.

Chapter 6

Conclusion

6.1 Summary of the Thesis

This work examined the factors that determine the understanding and successful application of machine learning explanation methods. It aimed to investigate whether combinations of explanation methods offer added value for users and to compare these combinations. Moreover, the study examined whether novel explanation methods can also enhance the understanding of models by shedding light on facets that are not covered by established explanation methods.
In chapter 2, stakeholder groups and their desiderata concerning model understanding were presented. Additionally, the individual explanation methods were classified, and various classification approaches were introduced. The chapter presented several frameworks designed to help apply explanation methods effectively, along with metrics intended to evaluate these methods based on different criteria, such as Likert scales and synthetic data evaluation.
In chapter 3, six different explanation methods were introduced. These include three established methods (SHAP, Partial Dependence Plots, and Surrogate Models) and three novel approaches (Lift, Feature Context Embeddings, and Embedded SHAP) that capture facets not addressed by the established methods.
These explanation methods were examined using the methodology described in chapter 4, by using both an analysis of synthetic data and a user study. The evaluation using synthetic data was limited to the novel approaches, whereas the user study examined all six explanation methods by generating pairs to measure the added value of the combinations. The findings were presented in chapter 5.

6.2 Key Contributions and Findings

The findings from section 5.1 show that all three novel approaches could successfully uncover properties of the underlying function on which the data creation depended. The movement in the embedding space using Embedded SHAP showed that instances with a complex underlying target function were harder for the model to predict, and thus their similarities with other instances fluctuated more. Moreover, instances that were similar with respect to the impact of the features on the target clustered together. Feature dependencies were correctly uncovered by using Lift values, with features that had no joint effect on the target receiving a Lift value of one accordingly. Embedding the features of a synthetic dataset revealed that features used in similar contexts were clustered together. A third feature, which belonged to neither of the two groups, was positioned between them, as expected, since it was not used particularly often in contexts similar to those of the two groups.
The user study revealed that the explanation methods differ significantly regarding both subjective (Table 5.4, Table 5.8) and objective understanding (Table 5.6). Increased subjective helpfulness of individual methods positively influences their combined helpfulness, although this effect appears to diminish with greater individual helpfulness, as indicated by the results in Table 5.9. Additionally, there is weak evidence suggesting that overall helpfulness increases when both explanation methods contribute equally (Table 5.11). Finally, there exists only a weak relationship ($R^2 = 0.156$) between understanding an explanation and correctly interpreting its implications, as demonstrated by Table 5.13.
The novel approaches, especially Embedded SHAP, helped users to gain a better subjective understanding of the model. All three novel explanation methods were able to reveal aspects of the

underlying model that were not exposed by the other established explanation methods.

6.3 Limitations of the Work

The use of multiple-choice questions to assess the objective understanding of both the explanation method and the model did not yield the expected results. One potential improvement could be to replace the objective assessment with alternative approaches, such as those described by Nauta et al. (see subsection 2.3.2). Additionally, evaluating the explanation methods via an online survey may not sufficiently incentivize participants to fully engage with and correctly apply the methods. Although participants had the option to indicate what was unclear about the explanations, there was little motivation for them to invest significant time in their responses. Conducting in-person questionnaires might reveal more insights into the shortcomings of the explanations and potential misunderstandings. Furthermore, issues such as missing questions in some survey responses and duplicate participation in three cases, as noted in section 4.2.4, were also problematic. However, the impact was seen as small and was addressed in the case of the missing question.

6.4 Future Research Directions

In future studies, it could be valuable to examine a pure between-subjects design, which might reveal statistically significant differences in understanding. To better account for different abstraction depths, incorporating the number of characters in the explanation may also help. However, relying solely on character count may be insufficient, given that technical complexity can vary widely between texts. Comparing explanations that address different aspects of a model remains challenging. One potential solution is to have participants reconstruct a decision tree and then assess whether a particular explanation method yields predictions more closely aligned with actual outcomes. This approach would enable the evaluation of methods that emphasize various aspects of the model, such as interaction between features. Participants could build a new decision tree after each explanation is presented, allowing for the quantification of how each additional method enhances prediction accuracy. A similar approach has also been suggested by Doshi-Velez et al. [22], where users approximate model decisions. Alternatively, as described in section 2.3, an application-based evaluation could focus on task performance rather than solely on model understanding.

With respect to novel explanation methods, slightly modified approaches could be explored. For instance, deriving Embedded SHAP values using variational autoencoders might yield softer cluster boundaries. Although Procrustes analysis was applied with UMAP in this study, it could also be used in conjunction with other dimensionality reduction techniques.

Regarding the Lift-based explanation method, it may be beneficial to extend the analysis beyond adjacent pairs of decision nodes. Including the entire path from the root node to the target node and weighting distances based on path length could offer further insights. Moreover, when accounting for co-definedness in the dataset (as described in Equation 3.9), a more nuanced measure such as mutual information [75] could be employed instead of a binary variable indicating whether a value is defined, as outlined in Equation 3.8.

Bibliography

[1] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: towards a transparent evaluation of post hoc model explanations. In *Proceedings of the NIPS '22*, NIPS '22, page 1148, Red Hook, NY, USA, nov 2022. Curran Associates Inc.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.

[4] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 04 2010.

[5] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, Sweden, 2018.

[6] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7786–7795, Red Hook, NY, USA, 2018. Curran Associates Inc.

[7] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B*, 82(4):1059–1086, 2020.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. Accessed: 2024-12-03.

[9] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022.

[10] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, 2011.

[11] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, jun 2023.

[12] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, jun 2024.

[13] Saša Brdnik, Vili Podgorelec, and Boštjan Šumak. Assessing perceived trust and satisfaction with multiple explanation techniques in xai-enhanced learning analytics. *Electronics*, 12(12):2594, 2023.

[14] Trevor S. Breusch and Arthur R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, 1979.

[15] Daniel Buschek, Malin Eiband, and Heinrich Hussmann. How to support users in understanding intelligent systems? an analysis and conceptual framework of user questions considering user mindsets, involvement, and knowledge outcomes. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–27, nov 2022.

[16] 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*, pages 785–794. KDD, 2016.

[17] Yun-Kyeung Choi, Mohsen Joshanloo, Jae-Ho Lee, Hong-Seock Lee, Heung-Pyo Lee, and Jonghwan Song. Understanding key predictors of life satisfaction in a nationally representative sample of koreans. *International Journal of Environmental Research and Public Health*, 20(18):6745, sep 2023.

[18] James Chowhan, Hossein Samavatyan, and Farimah HakemZadeh. Life satisfaction and the roles of work, family, and social factors in a social production function framework. *Journal of Happiness Studies*, 25, feb 2024.

[19] Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 316–324, 09 2023.

[20] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald, editors, *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[22] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.

[23] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[24] European Commission. 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) (text with eea relevance), 2016.

[25] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), oct 2001.

[26] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

[27] Gallup. Global research: See the world in data. Online, 2024. Accessed: 2024-06-07.

[28] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, oct 2018.

[29] John C. Gower and Garmt B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

[30] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168. ACM, 2016.

[31] Roman Hornung and Anne-Laure Boulesteix. Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects. *Computational Statistics & Data Analysis*, 171:107460, mar 2022.

[32] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4211–4222. Curran Associates, Inc., 2020.

[33] 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 *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. NeurIPS, 2017.

[34] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[35] Satyapriya Krishna, Tessa Han, Alex Gu, Zhiwei Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Trans. Machine Learning Research (TMLR)*, 2024.

[36] Karl Kumbier, Sumanta Basu, James B. Brown, Susan Celniker, and Bin Yu. Refining interaction search through signed iterative random forests. *CoRR*, abs/1810.07287, 2018.

[37] Tobias Labarta, Elizaveta Kulicheva, Ronja Froelian, Christian Geiß ler, Xenia Melman, and Julian von Klitzing. Study on the helpfulness of explainable artificial intelligence (xai). In *Proceedings of the World Conference on Explainable Artificial Intelligence (XAI)*, pages 294–312, 2024.

[38] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019.

[39] Louise Lambert, Tatiana Karabchuk, and Mohsen Joshanloo. Predictors of life satisfaction in the united arab emirates: Results based on gallup data. *Current Psychology*, 41(6):3827–3841, jun 2022.

[40] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? — a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, jul 2021.

[41] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018.

[42] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[43] Antonio Malvaso and Weixi Kang. The relationship between areas of life satisfaction, personality, and overall life satisfaction: An integrated account. *Frontiers in Psychology*, 13, 2022.

[44] John A. McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. Artificial intelligence explainability: The technical and ethical dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2207):20200363, 2021.

[45] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, page 3337, Red Hook, NY, USA, nov 2023. Curran Associates Inc.

[46] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[47] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[48] McKinsey and Company. Notes from the ai frontier: Insights from hundreds of use cases. Technical report, McKinsey and Company, 2018. Accessed: 2023-05-29.

[49] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in Artificial Intelligence*, 5, 2022.

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[51] Chao Min, Guoyong Liao, Guoquan Wen, Yingjun Li, and Xing Guo. Ensemble interpretation: A unified method for interpretable machine learning. *arXiv preprint arXiv:2312.06255*, 2023.

[52] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):Article 24, 45 pages, Aug 2021.

[53] Christoph Molnar. *Interpretable Machine Learning*. Self-published, 2 edition, 2022. Online book, accessed on 2025-02-14.

[54] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 6th edition, 2021.

[55] Cristian Muñoz, Kleyton da Costa, Bernardo Modenesi, and Adriano Soares Koshiyama. Evaluating explainability in machine learning predictions through explainer-agnostic metrics. 2023. Withdrawn.

[56] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, oct 2019.

[57] Maryam Naqvi, Syed Qasim Gilani, Tehreem Syed, Oge Marques, and Hee-Cheol Kim. Skin cancer detection using deep learning—a review. *Diagnostics*, 13(11):1911, 2023.

[58] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s), July 2023.

[59] Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the court: Explainable ai methods prone to disagreement. arXiv preprint arXiv:2105.03287, 2021.

[60] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[61] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, 46(4):1023–1031, 2014.

[62] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[63] Alun D. Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable AI. *CoRR*, abs/1810.00184, 2018.

[64] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648. NeurIPS, 2018.

[65] James B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):350–371, 1969.

[66] 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, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[68] Leonid Rozenblit and Frank C. Keil. The illusion of explanatory depth. *Cognitive Science*, 26(5):521–562, 2002.

[69] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[70] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[71] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, chapter 1, pages 5–22. Springer International Publishing, Cham, CH, 2019.

[72] Deanna Schreiber-Gregory and Karlen Bader. Logistic and linear regression assumptions: Violation recognition and control. In *Proceedings of the Joint SESUG, WUSS, SCSUG, PharmaSUG, MWSUG Conference*, jan 2018.

[73] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, nov 2017.

[74] Amazon Web Services. Mechanical turk documentation: Qualification requirement data structure, 2024. Accessed: 2024-12-23.

[75] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[76] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[77] SHARE-ERIC. Share - survey of health, ageing and retirement in europe. Online, 2024. Accessed: 2024-06-01.

[78] Xiaofang Shen, Fei Yin, and Can Jiao. Predictive models of life satisfaction in older people: A machine learning approach. *International Journal of Environmental Research and Public Health*, 20(3):2445, 2023.

[79] Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 2239–2250, New York, NY, USA, 2022. ACM.

[80] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[81] University of Michigan. Health and Retirement Study (HRS) Public Use Dataset. Produced and distributed by the University of Michigan, 2023. Funded by the National Institute on Aging (grant number NIA U01AG009740).

[82] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[83] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, page 1023, Red Hook, NY, USA, nov 2022. Curran Associates Inc.

[84] Guoan Yue, Weilong Xiao, and Qinghui Fan. The influence of subjective socioeconomic status on life satisfaction: The chain mediating role of social equity and social trust. *International Journal of Environmental Research and Public Health*, 19(23):15652, 2022.

[85] Jianlong Zhou, Amir Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10:593, mar 2021.

Appendix A

Shapley Value Calculation

Assume the prediction function f is defined as follows:

$$f(A, B, C) = A + 2B + 3C$$

where A, B, and C are binary features (either 0 or 1).
To calculate the Shapley value for A, we consider all subsets S of the set $\{B, C\}$ that do not include A. These subsets are: $\emptyset$, $\{B\}$, $\{C\}$, $\{B, C\}$.

Calculation Steps

1. **Subset $\emptyset$ (No features)**
   - $f(\emptyset) = 0$
   - $f(\{A\}) = 1$
   - Contribution when A is added $= 1 - 0 = 1$
   - Weight $w(\emptyset)$: $\frac{0!\cdot(3-0-1)!}{3!} = \frac{2}{6} = \frac{1}{3}$
2. **Subset $\{B\}$**
   - $f(\{B\}) = 2$
   - $f(\{A, B\}) = 3$
   - Contribution when A is added $= 3 - 2 = 1$
   - Weight $w(\{B\})$: $\frac{1!\cdot(3-1-1)!}{3!} = \frac{1}{6}$
3. **Subset $\{C\}$**
   - $f(\{C\}) = 3$
   - $f(\{A, C\}) = 4$
   - Contribution when A is added $= 4 - 3 = 1$
   - Weight $w(\{C\})$: $\frac{1!\cdot(3-1-1)!}{3!} = \frac{1}{6}$
4. **Subset $\{B, C\}$**
   - $f(\{B, C\}) = 5$
   - $f(\{A, B, C\}) = 6$
   - Contribution when A is added $= 6 - 5 = 1$
   - Weight $w(\{B, C\})$: $\frac{2!\cdot(3-2-1)!}{3!} = \frac{2}{6} = \frac{1}{3}$

Summing up the contributions weighted by their respective weights:

$$\phi_A = \frac{1}{3} \times 1 + \frac{1}{6} \times 1 + \frac{1}{6} \times 1 + \frac{1}{3} \times 1 = 1$$

The Shapley value of 1 for feature A indicates that A contributes 1 unit to the model output, based on averaging A's impact across all combinations of other features.

Appendix B

Rotation Matrix Calculation

Given two matrices A and B, where:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 2 \\ -1 & 0 \end{bmatrix}$$

The goal is to find an orthogonal matrix Ω that minimizes the Frobenius norm of the difference between ΩA and B, subject to Ω being orthogonal.

Since A is the identity matrix:

$$A^T = A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Multiplying B by A^T:

$$M = BA = \begin{bmatrix} 0 & 2 \\ -1 & 0 \end{bmatrix}$$

The SVD of M results in:

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad V^T = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Given the matrices from the SVD:

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad V^T = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

The calculation of Ω becomes:

$$\Omega = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

This matrix Ω represents a 90-degree rotation (counterclockwise).

Appendix C

Dataset Categories

Feature	Examples
Demographics and Networks	Gender, year of birth, country of birth, information about parents
Children	Number, age, gender, education, current living situation of children
Physical Health	Chronic illnesses, symptoms, physical functioning, overall health
Behavioral Risks	Smoking, alcohol consumption, physical activity, diet
Cognitive Function	Memory, mental speed, executive function through standardized tests
Mental Health	Depression, anxiety, life satisfaction, stress levels
Health Care	Visits to doctors, hospital stays, medication usage
Employment and Pensions	Employment history, current employment status, job characteristics
Grip Strength	Physical strength measured through handgrip tests
Social Support	Availability and quality of support from family, friends, social networks
Financial Transfers	Transfers between the respondent and others, such as children and parents
Housing	Type of housing, ownership status, housing quality
Household Income	Income sources and amounts, wages, pensions, other benefits
Consumption	Expenditure on various goods and services
Assets	Financial and non-financial assets owned
Activities	Daily activities, leisure activities, social participation
Expectations	Future expectations regarding health, financial situation, and other aspects

Table C.1: Regular Questionnaire Modules in all Waves with the exception of Wave 3

Appendix D

Model Training

Metric	LightGBM	LightGBM Fine-Tuned	XGBoost	CatBoost
RMSE - Test	1.34	1.34	1.37	1.33
R2 - Test	0.48	0.47	0.45	0.48
R2 - Train	0.60	0.55	0.55	0.54
Training Time (s)	42.20	15.62	99.51	121.72
Num Trees	318	666	47	590
Num Features	1698	471	1129	1127
Total Leaves	9858	25308	3414	37760

Table D.1: Model Performance Metrics for Training and Testing Sets

Hyperparameter	Min Value	Max Value	Final Value
lambda_l1	0.1	10.0	1.22
lambda_l2	0.1	10.0	1.58
num_leaves	30	40	38
bagging_fraction	0.3	0.7	0.58
bagging_freq	1	7	2
min_child_samples	80	150	128
learning_rate	0.001	0.1	0.0147

Table D.2: Hyperparameter Tuning Ranges and Final Values

Appendix E

Model Explanations

The tables Table E.1, Table E.2, and Table E.3 show the percentage of total gain, cumulative gain, the percentage of splits in which a feature appeared, and the rank a feature or category had in terms of splits.

Gain and Split

27% of the total gain was achieved by features related to physical health. Physical and mental health together accounted for 44% of the total gain. Physical health appeared in 7% of the splits, making it the 6^{th} most important category in terms of splits. The category that appeared in the most splits was retrospective accommodation.

Category	**Prefix**	**% Gain / Cumulative**	**% Split / Rank**
Physical Health	ph	0.27 / 0.27	0.07 / 6
Mental Health	mh	0.17 / 0.44	0.08 / 4
Retrospective Accommodation	ra	0.12 / 0.56	0.26 / 1
Consumption	co	0.11 / 0.67	0.04 / 8
Language	la	0.08 / 0.75	0.08 / 3
Demographics and Networks	dn	0.07 / 0.82	0.09 / 2
Expectations	ex	0.05 / 0.87	0.08 / 5
Social Networks	sn	0.03 / 0.89	0.04 / 9
Employment and Pensions	ep	0.02 / 0.91	0.04 / 7
General Life	gl	0.01 / 0.92	0.02 / 12
Retrospective Employment	re	0.01 / 0.94	0.03 / 10
Interviewer Observations	iv	0.01 / 0.94	0.01 / 18
Children	ch	0.01 / 0.95	0.02 / 11

Table E.1: 95% of Total Gain Grouped By Category and Ranked By Gain

Feature	Description	% Gain / Cumulative	% Split / Rank
ph003_	Self assessed health status	0.2288 / 0.2288	0.0159 / 11
co007_	Is household able to make a living	0.0906 / 0.3193	0.0183 / 9
language	Language of questionnaire	0.0761 / 0.3954	0.0767 / 3
ra015c_1	Region of residence (not current) - coded	0.0637 / 0.4591	0.1228 / 1
mh002_	Sad or depressed last month	0.0382 / 0.4973	0.0080 / 17
ra015c_2	Region of residence (not current) - coded	0.0366 / 0.5339	0.0784 / 2
mh037_	Feels lonely	0.0278 / 0.5617	0.0065 / 21
sn012_	Social Network satisfaction	0.0215 / 0.5833	0.0192 / 7
mh003_	Hopes for the future	0.0206 / 0.6038	0.0047 / 37
dn505c	Country of birth coded: father	0.0195 / 0.6234	0.0243 / 5

Table E.2: Top 10 Most Important Features Ranked by Gain

The first region of residence `ra015_c1` appeared in most splits (12.28%). It ranked 4$^{\text{th}}$ in terms of gain.

Feature	Description	% Gain / Rank	% Split / Cumulative
ra015c_1	Region of residence (not current) - coded	0.0637 / 4	0.1228 / 0.1228
ra015c_2	Region of residence (not current) - coded	0.0366 / 6	0.0784 / 0.2012
language	Language of questionnaire	0.0761 / 3	0.0767 / 0.2779
ra015c_3	Region of residence (not current) - coded	0.0137 / 13	0.0335 / 0.3114
dn505c	Country of birth coded: father	0.0195 / 10	0.0243 / 0.3357
ph060_	Self assessed health on a scale from 0 to 10	0.0164 / 11	0.0228 / 0.3585
sn012_	Social Network satisfaction	0.0215 / 8	0.0192 / 0.3777
ex026_	Trust in other people	0.0134 / 15	0.0188 / 0.3965
co007_	Is household able to make a living	0.0906 / 2	0.0183 / 0.4149
dn504c	Country of birth coded: mother	0.0109 / 18	0.0166 / 0.4315

Table E.3: Top 10 Most Important Features Ranked by Split Importance

Lift Ratio of Model Features

Table E.4 presents the ten highest-ranking feature combinations based on the Lift ratio derived from Equation 3.9. Feature combinations that also ranked among the top 10 based on their lift value from the model are highlighted in bold.

	Feature 1	Feature 2	R. Lift	D. Lift	L. Ratio
1	When poor health period stopped (gl010_)	Region of residence (not current) - coded (ra015c_2)	0.91	0.0003	3257
2	Region of residence (not current) - coded (ra015c_2)	Satisfied with job achievements (wq032_)	1.19	0.0005	2206
3	Born a citizen of country of interview (dn503_)	Self assessed health on a scale from 0 to 10 (ph060_)	2.36	0.028	84
4	Further education: country-specific category 20 (dn012d20)	Self assessed health on a scale from 0 to 10 (ph060_)	2.42	0.031	77
5	**Period of financial hardship (gl011_)**	**When financial hardship period stopped (gl013_)**	**94.00**	**1.45**	**65**
6	Further education: country-specific category 15 (dn012d15)	Self assessed health on a scale from 0 to 10 (ph060_)	1.19	0.027	44
7	**Country (country)**	**Current job situation (ep005_)**	**8.08**	**1.00**	**8**
8	**Afford to pay an unexpected expense without borrowing money (co206_)**	**Suicidal feelings or wish to be dead (mh004_)**	**8.68**	**1.08**	**8**
9	**Partner outside household (dn040_)**	**Irritability (mh010_)**	**7.69**	**1.00**	**8**
10	Month depression for the last time (mh031_)	Region of residence (not current) - coded (ra015c_1)	0.88	0.12	7

Table E.4: Lift Ratio = L. Ratio, R. Lift = Rules Lift, D.Lift = Dataframe Lift.

Appendix F

Survey Questions and Answers

All participants received the questions from item 1 (topic interest, confidence in mathematical abilities and background information) as well as item 8 (combined effect, open questions). Furthermore, participants received two out of the available six different explanation methods (item 2 to item 7).

1. Topic interest, confidence in mathematical abilities and background information
    - What is your age range? (Age brackets spanning five years)
    - What is the highest level of education you have completed? (No formal education, high school diploma, bachelor's degree, master's degree, doctorate)
    - How interested are you in explanation methods for machine learning models? (1 means no interest, 5 means high interest)
    - How interested are you in understanding the factors that influence life satisfaction? (1 means no interest, 5 means high interest)
    - How confident are you in your abilities to comprehend various types of graphs and mathematical concepts? (1 means no confidence, 5 means high confidence)
2. SHAP Values
    - Do you think you understood the concept of SHAP values? (1 means did not understand the concept, 5 means did fully understand concept)
    - What do SHAP values indicate?
        - SHAP values show how accurate the model predictions are compared to actual outcomes.
          Explanation: False. SHAP values do not measure the accuracy of model predictions; instead, they explain the contribution of each feature to the model's prediction.
        - SHAP values indicate the importance of each feature in the model's prediction.
          Explanation: True. SHAP values are designed to indicate the contribution of each feature to a specific prediction.
        - SHAP values assess the quality and cleanliness of the data used in the model.
          Explanation: False. SHAP values do not provide information about data quality or cleanliness.
        - SHAP values measure the error of the model's prediction.
          Explanation: False. SHAP values do not measure prediction error, they only explain how features contribute to individual predictions.
    - Based on the feature importance plot seen above the following can be said about the feature "self assessed health status":
        - Each instance is more strongly influenced by health than any other feature.
          Explanation: False. The plot shows the absolute sum of SHAP values for each feature, which provides average feature importance over the entire dataset. While health may have a strong average impact, this does not mean that each individual instance is more influenced by health than other features.
        - Self assessed health status had the largest average impact on a person's life satisfaction
          Explanation: True. Self-assessed health status has the highest value, this feature has the largest average impact on life satisfaction.

- Self assessed health status had a strong positive impact on life satisfaction (roughly 8000 life satisfaction points).
  Explanation: False. SHAP values reflect both positive and negative contributions. The absolute value indicates the strength of the impact, regardless of its direction.
- Self-assessed health status cannot directly be compared to the other features, as SHAP values cannot be compared.
  Explanation: False. SHAP values can be compared across features in terms of their absolute values, as shown in the feature importance plot.

- According to the ten most important features in absolute terms "Social Network Satisfaction" is the fifth most important feature yet it does not appear in the force plot shown above. Mark all reasons why this could be the case:
  - "Social Network Satisfaction" is not that important for this person
    Explanation: True. While Social Network Satisfaction may be important on average across the dataset, for this particular individual, it might not have a significant impact on the model's prediction, so it does not appear in the force plot.
  - This person has an average value for "Social Network Satisfaction", which does not have a strong effect
    Explanation: True. If the person's Social Network Satisfaction is close to the average, its SHAP value might be near zero, which means it has little to no effect on the prediction and thus is not highlighted in the force plot.
  - "Social Network Satisfaction" is influenced by other features, so the effect does not appear in the plot
    Explanation: False. SHAP values are additive and meant to explain individual feature contributions in a way that accounts for interaction effects. If Social Network Satisfaction was influential, it would still appear in the force plot, regardless of interactions with other features.
  - "Social Network Satisfaction" is cancelled out by other features, which reduces its effect on life satisfaction
    Explanation: False. SHAP values explain individual feature contributions even when interactions exist between features.
- How much did SHAP help you gain a better understanding of the model? (1 means did not help, 5 means completely clarified the model)

3. Embedded SHAP Values

- To what extent have you understood the concept of the plotted SHAP values? (1 means did not understand concept, 5 means did fully understand concept)
- If there are two clusters (Cluster A and Cluster B) formed by using embedded SHAP and the instances in Cluster A move around whereas the instances in Cluster B remain static this indicates that:
  - The impact of the feature values in Cluster A impact the instances differently than the feature values in Cluster B. The model corrected the relative similarity between instances within Cluster A
    Explanation: True. The movement of instances in Cluster A suggests that the model is adapting its understanding of how feature values influence predictions differently for these instances compared to those in Cluster B, which remain static. This is backed by the findings in section 4.1.1.
  - The impact of the feature values in Cluster A is larger than the impact of the feature values in Cluster B. The model learned more about instances in Cluster A than about instances in Cluster B
    Explanation: False. While instances in Cluster A may be moving more, it does not necessarily mean their feature values have a larger impact. The movement could be due to the model refining its predictions based on the variability in the feature values, rather than indicating overall larger impacts compared to Cluster B.
  - The features in Cluster A are more important than the features in Cluster B. The model therefore moved instances in Cluster A more than in Cluster B.
    Explanation: False. The movement of instances in Cluster A does not directly imply that features are more important, it may simply reflect the model's ongoing adjustment to different feature complexities in Cluster A compared to the static nature of Cluster B.

  - The model focuses on Cluster A as there are more instances in Cluster A.
    Explanation: False. The presence of more instances in Cluster A does not inherently mean the model will focus on them more, the static nature of Cluster B might indicate a different structural relationship or feature importance that leads to fewer adjustments, regardless of instance count.

- The instances did not move a lot from their initial positions, this indicates that:
  - The model did not learn a lot.
    Explanation: False. Limited movement of instances does not necessarily indicate a lack of learning, it may imply that the model has effectively captured the relationships among instances.
  - The first trees summarize the model well with respect to the similarity between instances.
    Explanation: True. If instances remain largely in their initial positions, it suggests that the initial trees are providing an adequate representation of the relationships among the instances.
  - The relative similarity stayed the same between instances throughout the iterations.
    Explanation: True. Minimal movement may imply that the model's understanding of instance similarities has not changed significantly, therefore the relative positions of instances remain stable through iterations.
  - The model cannot determine how similar the instances are to one another.
    Explanation: False. The lack of movement does not mean the model cannot determine similarity, rather, it suggests that the model has accurately identified and maintained the relationships between instances without needing to adjust their positions significantly.
- There are only a couple features that are the most important features (health, language, financial situation) across clusters. This indicates that:
  - The model did not learn a lot.
    Explanation: False. The presence of only a few important features does not imply that the model lacks learning, it may indicate that these features are consistently influential in predicting outcomes across clusters.
  - These features have many values, so the model has to make many splits to distinguish the feature values effectively.
    Explanation: False. While features with many values can lead to more splits, this statement does not explain the importance of only a few features across clusters, rather, it suggests complexity in handling those features.
  - The model ignores other features that might influence life satisfaction.
    Explanation: False. The focus on a few features does not necessarily mean that the model ignores others, it could be that those features are simply the most relevant in explaining life satisfaction for the given clusters.
  - Health, language, and financial situation are important for all groups of instances.
    Explanation: True. The consistent ranking of these features across clusters suggests that they have a significant and universal impact on life satisfaction. They are crucial for all groups of instances.
- How much did the plotted context similarity values help you to get a better understanding of the model? (1 means did not help, 5 means completely clarified the model)

4. Surrogate Model

- To what extent have you understood the concept of a surrogate model? (1 means did not understand, 5 means completely understood)
- Which explanation summarizes the functionality of a surrogate model?
  - A surrogate model removes features from a more complex model to create a simplified version.
    Explanation: False. While a surrogate model may involve simplification, it does not necessarily remove features, rather, it approximates the original model's predictions without focusing on feature elimination.
  - A surrogate model uses the predictions from a more complex model to create a simplified version.

Explanation: True. A surrogate model approximates the predictions of a complex model to provide interpretability and insights without the full complexity.

- A surrogate model predicts the real values of a target feature, using only a subset of the original feature set.
  Explanation: False. This statement implies that a surrogate model limits itself to a subset of features to predict a specific target feature, which is not the primary purpose of a surrogate model. Instead, it mimics the overall behavior of the original model.
- A surrogate model estimates how many features are needed to correctly predict the real values of a target feature.
  Explanation: False. A surrogate model is not designed to assess the number of features required for accurate predictions. It focuses on providing a simplified representation of the predictions made by a more complex model.

- The difference between the two paths indicates that:
  - Feeling depressed is more important to people with bad physical health than to people with good physical health.
    Explanation: True. The feature of feeling depressed plays a more significant role in influencing the outcomes for individuals with poor physical health compared to those with good physical health.
  - People are more likely to be depressed when their physical health is bad, which is why it's not as relevant for people with good physical health.
    Explanation: True. Poor physical health may lead to increased depression, making it less relevant for those who are in good health, as their overall experience may differ significantly.
  - Feeling depressed has the same effect on life satisfaction as feeling lonely or feeling a lack of companionship.
    Explanation: False. While both feelings may influence life satisfaction, the paths indicate that feeling depressed is more directly associated with poor physical health, while loneliness is addressed for those with good health.
  - People with bad physical health are not affected by a lack of companionship.
    Explanation: False. Companionship may still be relevant to people with bad physical health, however, other features may be more important than companionship in determining their life satisfaction.
- Physical Health can take the values Excellent, Very good, Good, Fair and Poor. Do you think that it is more important for a person to distinguish between poor and fair physical health (Node A in the image) or between excellent, very good or good physical health (Node B in the image)?
  - This question cannot be answered because the order in which nodes appear does not determine the importance of the feature.
    Explanation: False. While node order does not always indicate importance, in this case, the limited number of splits suggests that the feature's relevance and the splits can provide insight into its importance.
  - Distinguishing between poor and fair physical health is more important, as it appears higher up in the tree.
    Explanation: True. Given that the feature only appears in two splits (excluding the root), a higher position in the tree generally indicates a more significant impact on the model's predictions for distinguishing between these categories.
  - This question cannot be answered as the feature occurs multiple times, which indicates that there is an error in the model.
    Explanation: False. The occurrence of the feature in only two splits does not indicate an error. A feature may appear multiple times.
  - Distinguishing between excellent, very good and good physical health is more important, as it appears further down the tree.
    Explanation: False. If the feature occurs in only two splits, being further down the tree does not imply higher importance, rather, it may suggest that these distinctions are less critical for the model's predictive capability.
- The surrogate decision tree helped me to get a better understanding of the model. (1 means do not agree, 5 means do fully agree)

5. Partial Dependence Plots (PDPs)

- To what extent have you understood the concept of a partial dependence plot? (1 means did not understand, 5 means completely understood)
- What does a flat line in a partial dependence plot indicate about the relationship between the feature and the target variable?
  - The feature strongly influences the target variable.
    Explanation: False. A flat line suggests that changes in the feature do not significantly affect the target variable.
  - The target variable is highly variable.
    Explanation: False. A flat line does not imply high variability in the target variable, rather, it shows a consistent response across different values of the feature.
  - The feature is correlated with other features.
    Explanation: False. A flat line does not directly indicate correlation with other features, it simply reflects that the feature does not influence the target variable in the current model context.
  - The feature influence on the target variable does not change across its values.
    Explanation: True. A flat line indicates that regardless of the feature's value, the target variable remains constant.
- What can you infer based on the two partial dependence plots shown above?
  - Self assessed health and household consumption both roughly have the same impact magnitude on life satisfaction as the range spans from 6.4 to 7.2.
    Explanation: True. This indicates that the effects of both features on life satisfaction are comparable, as their scores fall within a similar range (6.4 to 7.2). This suggests that improvements in either self-assessed health or household consumption contribute equally to life satisfaction.
  - There are fewer large feature values, so the average effect on the target decreases for each feature value decreases.
    Explanation: False. Partial Dependence Plots replace all existing feature values for a given feature. Therefore, the real values of a feature do not affect the values displayed in the plot.
  - Health has a positive but diminishing effect on the target as health improves from poor to excellent.
    Explanation: True. The plots illustrate that while improvements in health status lead to higher life satisfaction, the increase in satisfaction diminishes as one moves from poor to excellent health.
  - Both features are irrelevant to the model as the increase from "poor" to "excellent" and from "with great difficulty" to "easily" is linear.
    Explanation: False. The linearity of an increase does not imply irrelevance, it suggests a consistent effect on life satisfaction, although the effect is diminishing here and not linear.
- There are two questions regarding health in the survey, one which asks participants to rank their health into the categories Poor, Fair, Good, Very Good, and Excellent and another question which asks participants to rank their health on a scale from 0 to 10. What effect might this have on the outcome of the Partial Dependence Plot?
  - Including both features could show that either or both of the two has a large significant effect on the target, even though their impact is much smaller.
    Explanation: False. While including both features may lead to a situation where their combined influence appears significant, this does not accurately reflect their individual contributions, which could be diminished when both are included.
  - Including both features has no effect on the Partial Dependence Plots, because the features are likely independent of one another.
    Explanation: False. The inclusion of both features can interact in ways that affect the interpretation of the Partial Dependence Plots.
  - Including both features could show that either or both of the two has no significant effect on the target, even though their impact is significant if only one of them had been used as a feature.
    Explanation: True. The presence of both features may lead to confounding effects.
  - Including both features will show that both of the features have a significant effect, because only the feature values of one feature will be changed.

Explanation: False. Including both features does not guarantee that their effects will be significant, as the interaction between them can obscure their individual contributions.

- How much did the partial dependence plots help you to get a better understanding of the model? (1 means did not help, 5 means completely clarified the model)

6. Lift

   - To what extent have you understood the concept of Lift? (1 means did not understand, 5 means completely understood)
   - What does it signify if a pair of features, such as physical health and mental health, has a lift value of 2?
     - The two features appear twice less frequently together than if they were statistically independent.
       Explanation: False. A lift value of 2 indicates a positive association between the features, meaning they appear together more often than expected, not less.
     - The lift value of 2 implies that the occurrence of one feature has no influence on the occurrence of the other.
       Explanation: False. A lift value greater than 1 suggests that the occurrence of one feature positively influences the occurrence of the other.
     - The two features appear twice as often together as they would were they statistically independent from one another.
       Explanation: True. A lift value of 2 means that the features are twice as likely to occur together compared to what would be expected if they were independent.
     - Physical health occurs twice as often as mental health.
       Explanation: False. The lift value does not provide information about the individual frequencies of each feature, it only indicates the degree of association between them.
   - Select all true statements:
     - Questions related to being able to pay for an unexpected expense and suicidal feelings are likely to appear in sequential order in the model
       Explanation: True. The lift value of 9 indicates a significant joint effect, suggesting that these features are positively associated with one another in the context of predicting life satisfaction.
     - Features related to periods of financial hardship often occur together in the model.
       Explanation: True. The lift value of 94 between "When period of financial hardship stopped" and "Period of financial hardship" shows a strong association.
     - The question regarding a person's country appears 8 times.
       Explanation: False. The value 8 does not refer to the number of times an individual feature occurs.
     - The questions regarding current job situation and country appeared together 8 times.
       Explanation: False. The value 8 does not refer to the number of times features occur jointly.
   - The question regarding the current job situation of a person and the country the person is residing in are always likely to occur together. This statement is:
     - False, because the question about the current job situation may only be relevant for specific countries.
       Explanation: True. This statement correctly identifies that the relevance of job situations may vary by country.
     - True, because the question about the current job situation is independent of the country that a person is residing in.
       Explanation: False. This statement is incorrect because the occurrence of job situation and country may be related.
     - False, because they only appeared together 8 times, which is insignificant if the model is large.
       Explanation: False. A lift value of 8 does not reveal any information about how often the two features appeared together.
     - True, because according to the lift value, no other feature appears as often with country as the feature "Current job situation."

Explanation: False. While the lift value indicates some association, it does not guarantee that the two features are always likely to occur together regardless of their feature values.

- How much did the lift values help you to get a better understanding of the model? (1 means did not help, 5 means completely clarified the model)

7. Context Embeddings and CBOW

- To what extent have you understood the concept of context embeddings and CBOW? (1 means did not understand, 5 means completely understood)
- If two features are located at the same point in a two-dimensional space, they:
  - frequently occur together in the contexts.
    Explanation: False. Similar vector representations indicate that the features are used in similar contexts, not that they co-occur.
  - are indistinguishable with respect to the context they are used in.
    Explanation: True. If two features lie on the same spot in the embedding space, it suggests that they are used in indistinguishable ways within the same contexts according to the model's learned representations.
  - have the same effect on the target feature.
    Explanation: False. Even if two features share a similar position in the embedding space, they may still differ in their effect on the target.
  - have no effect on the target feature.
    Explanation: False. Being similar in the embedded space doesn't imply they have no effect on the target feature, they could still significantly impact the model's predictions.
- The plot above shows the context embeddings for all the features in the model. The feature "Hopes for the future (mh003_)" can be seen towards the left of the plot. It is rather isolated from the other features. This means that:
  - mental health does not frequently co-occur with the same sets of features as other features do.
    Explanation: True. The feature's relationship with other features is distinct, which is why it is positioned separately in the embedding space.
  - there is no other feature that can act as a direct replacement for mental health.
    Explanation: True. The isolation in the embedding space indicates that no other feature can serve as a direct substitute or express similar information in the model.
  - it is used in more contexts than other features.
    Explanation: False. Isolation in the embedding space doesn't imply that the feature is used in more contexts. It only shows that its context is different from the others.
  - it has a strong effect on predicted life satisfaction.
    Explanation: False. The embedding plot shows relationships between features based on their contextual similarity, not their strength of impact on the target variable. The strength of the effect would need to be assessed using other feature importance methods.
- Features regarding mental health (MH) are predominantly found towards the left of the plot, whereas features related to the country a person was born in and lived earlier in life are found towards the top of the plot. This indicates that:
  - Features regarding mental health are similar to features regarding a person's origin with respect to the context they are used in.
    Explanation: False. The spatial separation of the features on the plot shows they are used in different contexts.
  - Features regarding mental health and features regarding a person's origin are used in different contexts.
    Explanation: True. The fact that these feature groups are located in different parts of the plot indicates that they do not co-occur in the same contexts.
  - Features regarding mental health have a higher similarity score than features regarding a person's origin.
    Explanation: False. The plot shows the context embedding, not the direct similarity score comparison.

  - The two features are used in different models.
    Explanation: False. The separation indicates different contexts within the same model.
- How much did the plotted context similarity values help you to get a better understanding of the model? (1 means did not help, 5 means completely clarified the model)

8. Combined Effect

- How much did the combination of both methods enhance your understanding compared to what you think a single method would have provided? (1 means the combination did not enhance understanding at all, 5 means the combination led to a complete or near-complete understanding)
- How do the two explanation methods contribute to your understanding of the model? (1 means the first method contributes significantly more than the second, 5 means the second method contributes significantly more than the first)
- Which of the two explanation methods did you find more useful and why?
- Are there any aspects of the model's behavior that you feel are still not adequately explained by the combined approach?

Appendix G

Model Features

	Feature	Gain / Cumulative	Split / Rank	Description
1	ph003_	0.2288 / 0.2288	0.0159 / 11	Self assessed health status
2	co007_	0.0906 / 0.3193	0.0183 / 9	Is household able to make a living
3	language	0.0761 / 0.3954	0.0767 / 3	Language of questionnaire
4	ra015c_1	0.0637 / 0.4591	0.1228 / 1	Region of residence (not current) - coded
5	mh002_	0.0382 / 0.4973	0.0080 / 17	Sad or depressed last month
6	ra015c_2	0.0366 / 0.5339	0.0784 / 2	Region of residence (not current) - coded
7	mh037_	0.0278 / 0.5617	0.0065 / 21	Feels lonely
8	sn012_	0.0215 / 0.5833	0.0192 / 7	Social Network satisfaction
9	mh003_	0.0206 / 0.6038	0.0047 / 37	Hopes for the future
10	dn505c	0.0195 / 0.6234	0.0243 / 5	Country of birth coded: father
11	ph060_	0.0164 / 0.6397	0.0228 / 6	Self assessed health on a scale from 0 to 10
12	mh004_	0.0142 / 0.6539	0.0058 / 27	Suicidal feelings or wish to be dead
13	ra015c_3	0.0137 / 0.6676	0.0335 / 4	Region of residence (not current) - coded
14	ex709_	0.0136 / 0.6812	0.0149 / 13	Life expectancy
15	ex026_	0.0134 / 0.6945	0.0188 / 8	Trust in other people
16	ex009_	0.0132 / 0.7077	0.0152 / 12	Self assessed likelihood of still living in ten years
17	dn012d15	0.0129 / 0.7206	0.0062 / 24	Further education: country-specific category 15
18	dn504c	0.0109 / 0.7315	0.0166 / 10	Country of birth coded: mother
19	mh035_	0.0105 / 0.7420	0.0037 / 45	Feels left out
20	mh036_	0.0096 / 0.7516	0.0034 / 48	Feels isolated from others
21	mh034_	0.0091 / 0.7607	0.0027 / 61	Feels lack of companionship
22	mh016_	0.0080 / 0.7687	0.0053 / 31	Enjoyment
23	iv004_	0.0073 / 0.7760	0.0066 / 20	Willingness to answer
24	gl004_	0.0066 / 0.7825	0.0061 / 25	When happiness period stopped

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
25	mh008_	0.0063 / 0.7888	0.0030 / 57	Less or same interest in things
26	co020e	0.0061 / 0.7949	0.0018 / 98	Minimum amount needed per month
27	co206_	0.0057 / 0.8006	0.0021 / 79	Afford to pay an unexpected expense without borrowing money
28	ra015c_4	0.0055 / 0.8061	0.0146 / 14	Region of residence (not current) - coded
29	re022c_1	0.0051 / 0.8112	0.0089 / 15	Currency of wage - coded
30	mh010_	0.0050 / 0.8162	0.0035 / 47	Irritability
31	mh024_	0.0047 / 0.8209	0.0040 / 41	Nervous
32	dn040_	0.0042 / 0.8250	0.0033 / 52	Partner outside household
33	co202_	0.0040 / 0.8291	0.0023 / 70	Afford to go on holiday at least once a year (a week long)
34	dn044_	0.0038 / 0.8329	0.0029 / 58	Marital status changed
35	country	0.0036 / 0.8365	0.0071 / 19	Country
36	dn012d20	0.0034 / 0.8399	0.0040 / 41	Further education: country-specific category 20
37	dn503_	0.0031 / 0.8430	0.0026 / 63	Born a citizen of country of interview
38	ep026_	0.0031 / 0.8461	0.0060 / 26	Satisfied with (main) job
39	ep005_	0.0031 / 0.8492	0.0039 / 43	Current job situation
40	mh013_	0.0028 / 0.8520	0.0004 / 299	Fatigue
41	gl012_	0.0027 / 0.8547	0.0047 / 37	When financial hardship period started
42	mh020_	0.0027 / 0.8574	0.0037 / 44	Ever treated for depression by doctor or psychiatrist
43	mh006_	0.0024 / 0.8598	0.0034 / 49	Blame for what
44	ex802_	0.0023 / 0.8621	0.0025 / 68	Financial situation today compared to expectations at age 45
45	ph089dno	0.0022 / 0.8643	0.0004 / 283	Bothered by frailty: none
46	hc125_	0.0022 / 0.8666	0.0043 / 39	Satisfaction with own coverage in basic health insurance/national health system
47	ex029_	0.0022 / 0.8688	0.0055 / 28	Frequency of praying
48	ph061_	0.0022 / 0.8710	0.0032 / 54	Health problem that limits paid work
49	dn005c	0.0022 / 0.8732	0.0084 / 16	Foreign country of birth coding
50	mh027_	0.0022 / 0.8753	0.0026 / 67	Felt faint
51	br033_	0.0021 / 0.8775	0.0021 / 79	Not eating meat, fish or chicken more often because ...
52	language_x...	0.0021 / 0.8795	0.0080 / 17	Language: End of life interview
53	mh023_	0.0021 / 0.8816	0.0037 / 45	Fear of the worst happening
54	ph006d18	0.0020 / 0.8836	0.0019 / 92	Other affective/emotional disorders: ever diagnosed/currently having
55	wq009_	0.0017 / 0.8853	0.0027 / 61	Work gave recognition
56	ex001_	0.0016 / 0.8868	0.0048 / 36	Introduction and example

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
57	cc722_1	0.0016 / 0.8884	0.0032 / 54	How would you rate the relationship with your mother
58	ep054_	0.0016 / 0.8900	0.0063 / 22	Kind of industry working in last job
59	ch017_1	0.0015 / 0.8915	0.0063 / 23	Child 1 highest school degree
60	cf002_	0.0015 / 0.8929	0.0019 / 94	Self-rated writing skills
61	re042c	0.0015 / 0.8944	0.0054 / 30	Currency of wage at end of main job - coded
62	ra015c_5	0.0014 / 0.8959	0.0041 / 40	Region of residence (not current) - coded
63	ph005_	0.0013 / 0.8972	0.0015 / 112	Limited in activities because of health
64	mh007_	0.0013 / 0.8985	0.0013 / 129	Trouble sleeping
65	ep337_	0.0013 / 0.8998	0.0026 / 65	Currently looking for job
66	re022c_2	0.0013 / 0.9011	0.0052 / 32	Currency of wage - coded
67	mobirthp	0.0013 / 0.9024	0.0055 / 28	Month of birth spouse/partner
68	ph072_3	0.0013 / 0.9037	0.0011 / 149	Been diagnosed with cancer since last interview
69	ph050_	0.0013 / 0.9050	0.0006 / 235	Help activities
70	gs010d1	0.0012 / 0.9062	0.0013 / 129	Why not completed gs test: r felt it would not be safe
71	ch017_2	0.0012 / 0.9074	0.0049 / 35	Child 2 highest school degree
72	wq727_	0.0011 / 0.9086	0.0022 / 73	Satisfaction with job
73	ph049d8	0.0011 / 0.9097	0.0018 / 96	Difficulties: preparing a hot meal
74	sn017_	0.0011 / 0.9108	0.0025 / 68	Empty network satisfaction
75	ph049d2	0.0011 / 0.9119	0.0016 / 105	Difficulties: walking across a room
76	ph011d10	0.0011 / 0.9130	0.0014 / 121	Drugs for: anxiety or depression
77	cf103_	0.0011 / 0.9141	0.0028 / 60	Memory
78	int_month	0.0011 / 0.9151	0.0051 / 33	Interview month
79	dn010_	0.0011 / 0.9162	0.0050 / 34	Highest school degree obtained
80	mh021_	0.0010 / 0.9172	0.0012 / 133	Ever admitted to mental hospital or psychiatric ward
81	gl013_	0.0010 / 0.9183	0.0013 / 125	When financial hardship period stopped
82	ep067_	0.0010 / 0.9193	0.0021 / 76	How became unemployed
83	ph072_4	0.0010 / 0.9203	0.0007 / 217	Suffered a hip fracture since last interview
84	re014_1	0.0009 / 0.9212	0.0030 / 56	Job industry
85	mh022_	0.0009 / 0.9222	0.0021 / 79	Ever told affective or emotional disorders
86	dn014_	0.0009 / 0.9231	0.0020 / 86	Marital status
87	sn005_1	0.0009 / 0.9240	0.0032 / 54	Network relationship: sn person 1
88	ep110d1	0.0009 / 0.9249	0.0011 / 144	Received public benefits: old age pension

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
89	cf001_	0.0009 / 0.9258	0.0013 / 125	Self-rated reading skills
90	yrbirthp	0.0009 / 0.9267	0.0020 / 82	Year of birth spouse/partner
91	sp003_1	0.0008 / 0.9275	0.0034 / 50	Who gave help: person 1
92	sn005_2	0.0008 / 0.9283	0.0028 / 59	Network relationship: sn person 2
93	ex011_	0.0008 / 0.9291	0.0026 / 63	Self assessed likelihood of still living in ten years
94	hh025_	0.0008 / 0.9299	0.0020 / 86	If I were in trouble, there are people in this area who would help me
95	co211_	0.0007 / 0.9306	0.0010 / 163	To help keeping living costs down: postponed visits to the dentist
96	gl011_	0.0007 / 0.9314	0.0005 / 256	Period of financial hardship
97	hh017e	0.0007 / 0.9321	0.0016 / 105	Total income received by all hh members an average month last year
98	dn020_	0.0007 / 0.9328	0.0022 / 75	Year of birth of former partner
99	br002_	0.0007 / 0.9335	0.0015 / 112	Smoke at the present time
100	dn021_	0.0007 / 0.9342	0.0033 / 51	Highest educational degree of former partner
101	ph049d9	0.0007 / 0.9349	0.0011 / 149	Difficulties: shopping for groceries
102	as003e	0.0007 / 0.9356	0.0015 / 115	Amount bank account
103	cc722_2	0.0007 / 0.9363	0.0016 / 109	How would you rate the relationship with your father
104	br029_	0.0007 / 0.9369	0.0020 / 82	How often serving of fruits or vegetables
105	as051e	0.0007 / 0.9376	0.0019 / 89	Amount selling cars
106	re037c_1	0.0007 / 0.9383	0.0026 / 65	Currency of pension benefit - coded
107	dn012d17	0.0007 / 0.9389	0.0005 / 256	Further education: country-specific category 17
108	ep678v1	0.0006 / 0.9396	0.0012 / 138	Bracket value 1
109	sn005_3	0.0006 / 0.9402	0.0022 / 73	Network relationship: sn person 3
110	int_year	0.0006 / 0.9408	0.0017 / 100	Interview year
111	as054d6	0.0006 / 0.9414	0.0015 / 118	Owe money: student loans
112	mh005_	0.0006 / 0.9421	0.0018 / 98	Feels guilty
113	co003e	0.0006 / 0.9427	0.0018 / 98	Amount spent on food outside the home
114	hs003_	0.0006 / 0.9433	0.0009 / 177	Childhood health status
115	ex028_	0.0006 / 0.9438	0.0022 / 73	Left or right in politics
116	hh022_	0.0006 / 0.9444	0.0017 / 102	Feeling part of this area
117	ph064_	0.0006 / 0.9450	0.0019 / 92	Health worse last wave
118	ra015c_6	0.0006 / 0.9455	0.0016 / 105	Region of residence (not current) - coded
119	sn009_1	0.0005 / 0.9461	0.0015 / 118	Network closeness: sn person 1
120	br026_	0.0005 / 0.9466	0.0021 / 76	How often serving of dairy products
121	ph072_1	0.0005 / 0.9471	0.0007 / 211	Had a heart attack since last interview

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
122	hs060_1	0.0005 / 0.9476	0.0020 / 86	When did illness period 1 stop
123	ch006_1	0.0005 / 0.9481	0.0019 / 89	Child 1 year of birth
124	ch016_1	0.0005 / 0.9486	0.0020 / 82	Child 1 employment status
125	ep326_	0.0005 / 0.9491	0.0005 / 256	Received severance payment since last interview
126	rh792_	0.0005 / 0.9496	0.0012 / 138	Postponed taking medication because of cost
127	as061_	0.0005 / 0.9501	0.0016 / 105	Reason for not having a bank account
128	ex010_	0.0005 / 0.9506	0.0015 / 118	Chance standard of living will be better
129	sn002a_3	0.0005 / 0.9510	0.0009 / 184	Any more persons with whom you often discuss: sn person 3
130	wq032_	0.0005 / 0.9515	0.0015 / 112	No description
131	mobirth	0.0005 / 0.9519	0.0022 / 71	Month of birth
132	dn003_	0.0004 / 0.9524	0.0015 / 115	Year of birth
133	hc114_	0.0004 / 0.9528	0.0009 / 169	Could not see a doctor because of cost
134	rc028_1	0.0004 / 0.9533	0.0015 / 115	Year of death other child
135	wq013_	0.0004 / 0.9537	0.0009 / 169	Work employees treated fair
136	cf014_	0.0004 / 0.9541	0.0020 / 86	Numeracy: 6000 is two-thirds what is total price
137	xt008_	0.0004 / 0.9546	0.0019 / 92	Month of decease
138	br003_	0.0004 / 0.9550	0.0013 / 129	How many years smoked
139	ch017_3	0.0004 / 0.9554	0.0019 / 94	Child 3 highest school degree
140	age2011	0.0004 / 0.9559	0.0008 / 194	Age in 2011
141	ep328_	0.0004 / 0.9563	0.0020 / 82	Retirement month
142	sr004dno	0.0004 / 0.9567	0.0009 / 177	Negative Shock: None of these
143	re014_2	0.0004 / 0.9571	0.0016 / 109	Job industry
144	ep032_	0.0004 / 0.9575	0.0012 / 138	Receive recognition for work in (main) job
145	ph049d5	0.0004 / 0.9579	0.0009 / 169	Difficulties: getting in or out of bed
146	iv007_	0.0004 / 0.9583	0.0013 / 129	Respondent asked for clarification
147	ep052_	0.0004 / 0.9587	0.0017 / 100	Name or title of last job
148	ph049d1	0.0004 / 0.9591	0.0007 / 211	Difficulties: dressing, including shoes and socks
149	ex104_	0.0004 / 0.9595	0.0009 / 169	Partner ever done paid work
150	iv008_	0.0004 / 0.9599	0.0011 / 144	Respondent understood questions
151	ep097_	0.0004 / 0.9603	0.0011 / 154	Pension claims
152	co002e	0.0004 / 0.9607	0.0014 / 123	Amount spent on food at home
153	mh011_	0.0004 / 0.9610	0.0011 / 149	Appetite
154	ex025_	0.0004 / 0.9614	0.0011 / 149	Chance to work after age of 63
155	gl010_	0.0004 / 0.9618	0.0016 / 105	When poor health period stopped

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
156	re022c_3	0.0004 / 0.9622	0.0016 / 105	Currency of wage - coded
157	cc730_	0.0004 / 0.9625	0.0011 / 149	Group of friends felt comfortable spending time with
158	ex007_	0.0004 / 0.9629	0.0011 / 154	Chance government reduces pension
159	mh025_	0.0004 / 0.9632	0.0010 / 163	Hands trembling
160	co207_	0.0004 / 0.9636	0.0008 / 203	To help keeping living costs down: continued wearing clothing that was worn out
161	ep111_1_1	0.0004 / 0.9639	0.0014 / 121	Receive old age pension period 1 from month
162	ep127_21	0.0003 / 0.9643	0.0013 / 125	Period from month (unemployed)
163	re014_3	0.0003 / 0.9646	0.0008 / 194	Job industry
164	gl006_	0.0003 / 0.9650	0.0013 / 125	When stress period started
165	hc760_	0.0003 / 0.9653	0.0009 / 184	Needed medication but could not afford last 12 months
166	mh031_	0.0003 / 0.9656	0.0012 / 138	Month depression for the last time
167	gs010d5	0.0003 / 0.9660	0.0005 / 256	Why not completed gs test: r did not understand the instructions
168	ph049d14	0.0003 / 0.9663	0.0007 / 217	Difficulties: leaving the house independently/accessing transportation
169	gs013_	0.0003 / 0.9666	0.0012 / 138	The position of r for this test
170	ph745_	0.0003 / 0.9669	0.0007 / 217	Have hearing aid
171	ch012_1	0.0003 / 0.9672	0.0011 / 144	Child 1 marital status
172	ex601_	0.0003 / 0.9675	0.0011 / 144	Start of non proxy section
173	ho060_	0.0003 / 0.9678	0.0009 / 169	Partner years in accomodation
174	cf005_	0.0003 / 0.9681	0.0012 / 133	Date: year
175	ep213_1	0.0003 / 0.9684	0.0012 / 138	First year received income source c1
176	sp009_1	0.0003 / 0.9687	0.0014 / 121	To whom did you give help: person 1
177	wq030_	0.0003 / 0.9690	0.0009 / 177	No description
178	ph089d4	0.0003 / 0.9693	0.0004 / 299	Bothered by frailty: fatigue
179	sp019d28	0.0003 / 0.9696	0.0008 / 194	R provided help with personal care to: other relative
180	ft003_1	0.0003 / 0.9698	0.0012 / 138	To whom given gift, person 1
181	ep649_	0.0003 / 0.9701	0.0006 / 225	Years worked in last job
182	as070e	0.0003 / 0.9704	0.0009 / 184	Interest or dividend income
183	pf003_	0.0003 / 0.9707	0.0011 / 149	Value first measurement
184	ph009_1	0.0003 / 0.9710	0.0011 / 144	Age heart attack or other heart problems
185	cf012_	0.0003 / 0.9712	0.0013 / 132	Numeracy: chance disease 10% of 1000
186	hc066_	0.0003 / 0.9715	0.0011 / 154	Total nights stayed in other institutions
187	it003_	0.0003 / 0.9718	0.0002 / 346	Computer skills

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
188	ph043_	0.0003 / 0.9720	0.0010 / 163	Eyesight distance
189	ph048d1	0.0003 / 0.9723	0.0006 / 235	Difficulties: walking 100 metres
190	ep018_	0.0003 / 0.9726	0.0011 / 154	Kind of industry working in
191	ph051_	0.0003 / 0.9728	0.0009 / 169	Help meets needs
192	ex111_	0.0003 / 0.9731	0.0010 / 158	Planning horizon of saving and spending
193	ho067e	0.0003 / 0.9733	0.0009 / 177	Amount similar dwelling todays market
194	sp014_	0.0003 / 0.9736	0.0009 / 184	Looked after grandchildren
195	mh030_	0.0002 / 0.9738	0.0006 / 235	Year depression for the last time
196	dn041_	0.0002 / 0.9741	0.0009 / 169	Years education
197	xt010_	0.0002 / 0.9743	0.0009 / 184	Age at the moment of decease
198	sn007_3	0.0002 / 0.9746	0.0009 / 177	Network contact: sn person 3
199	ex009age	0.0002 / 0.9748	0.0008 / 194	Life expectancy target age
200	age2020	0.0002 / 0.9750	0.0010 / 163	Age in 2020
201	re702_	0.0002 / 0.9753	0.0008 / 194	First time computer at work
202	ep078e_1	0.0002 / 0.9755	0.0009 / 177	Average payment income source c1 last year
203	gs009_	0.0002 / 0.9757	0.0008 / 194	2nd measurement: right hand
204	rp004c_1	0.0002 / 0.9760	0.0010 / 158	When relationship start
205	gs007_	0.0002 / 0.9762	0.0008 / 203	2nd measurement: left hand
206	wq011_	0.0002 / 0.9764	0.0007 / 217	Work had adequate support
207	cc721_1	0.0002 / 0.9766	0.0006 / 244	How much did your mother understand your problems and worries
208	gl007_	0.0002 / 0.9769	0.0009 / 177	When stress period stopped
209	cc721_2	0.0002 / 0.9771	0.0007 / 211	How much did your father understand your problems and worries
210	ch007_1	0.0002 / 0.9773	0.0010 / 158	Child 1 where does child live
211	gs010d4	0.0002 / 0.9775	0.0005 / 256	Why not completed gs test: r tried but was unable to complete test
212	cf019_	0.0002 / 0.9777	0.0005 / 270	Instruction for CF
213	ep152isco	0.0002 / 0.9780	0.0005 / 256	ISCO code: respondent's last job
214	sn006_3	0.0002 / 0.9782	0.0007 / 217	Network proximity: sn person 3
215	br010_	0.0002 / 0.9784	0.0010 / 158	Days a week consumed alcohol last 6 months
216	ep110d2	0.0002 / 0.9786	0.0002 / 387	Received public benefits: early retirement pension
217	mh019_	0.0002 / 0.9788	0.0006 / 244	Age depression symptoms first time
218	age_int	0.0002 / 0.9790	0.0006 / 225	Age of respondent at the time of interview
219	ep110d7	0.0002 / 0.9792	0.0003 / 323	Received public benefits: public-longterm care insurance

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
220	ch020_1	0.0002 / 0.9794	0.0009 / 184	Child 1 year of birth youngest child
221	dn009_	0.0002 / 0.9796	0.0005 / 256	Country-specific question
222	ho024ub	0.0002 / 0.9798	0.0009 / 184	Value of property ub
223	dn002_	0.0002 / 0.9800	0.0010 / 163	Month of birth
224	wq003_	0.0002 / 0.9802	0.0007 / 211	Work was uncomfortable
225	ph006dno	0.0002 / 0.9804	0.0008 / 194	None: ever diagnosed
226	cc010a_	0.0002 / 0.9806	0.0008 / 203	Relative position to others when ten: language
227	ch014_2	0.0002 / 0.9808	0.0009 / 184	Child 2 contact with child
228	sn006_1	0.0002 / 0.9810	0.0008 / 194	Network proximity: sn person 1
229	sp019d1	0.0002 / 0.9812	0.0006 / 225	R provided help with personal care to: spouse/partner
230	cf010_	0.0002 / 0.9814	0.0008 / 203	Verbal fluency score: number of animals
231	ph049d13	0.0002 / 0.9816	0.0003 / 333	Difficulties: managing money
232	xt002_	0.0002 / 0.9818	0.0008 / 203	Relationship to the deceased
233	ph049dno	0.0002 / 0.9820	0.0003 / 333	Difficulties: none of these
234	ph049d6	0.0002 / 0.9821	0.0004 / 299	Difficulties: using the toilet, incl getting up or down
235	rc024_1	0.0002 / 0.9823	0.0008 / 194	Year of birth other child
236	co209_	0.0002 / 0.9825	0.0004 / 299	To help keeping living costs down: put up with feeling cold
237	sn005_7	0.0002 / 0.9827	0.0009 / 169	Network relationship: sn person 7
238	ch021_	0.0002 / 0.9829	0.0008 / 203	Number of grandchildren
239	ep033_	0.0002 / 0.9831	0.0006 / 244	Salary or earnings are adequate in (main) job
240	dn027_1	0.0002 / 0.9833	0.0008 / 194	Age of death of parent: mother
241	ep129_21	0.0002 / 0.9834	0.0004 / 299	Period to month (unemployed)
242	sn009_3	0.0002 / 0.9836	0.0006 / 244	Network closeness: sn person 3
243	ep678v3	0.0002 / 0.9838	0.0004 / 299	Bracket value 3
244	yrbirth	0.0002 / 0.9840	0.0006 / 235	Year of birth
245	gs006_	0.0002 / 0.9842	0.0006 / 235	1st measurement: left hand
246	cc010_	0.0002 / 0.9844	0.0008 / 194	Relative position to others when ten: mathematically
247	re035_1	0.0002 / 0.9845	0.0006 / 225	Situation in after last job
248	ho044_	0.0002 / 0.9847	0.0004 / 299	Changed place of residence
249	gs010d2	0.0002 / 0.9849	0.0004 / 283	Why not completed gs test: iwer felt it would not be safe
250	rc032_1	0.0002 / 0.9851	0.0006 / 225	Maternity benefit amount
251	as649_	0.0002 / 0.9852	0.0004 / 313	Number of cars
252	agep2020	0.0002 / 0.9854	0.0007 / 211	Age of partner in 2020

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
253	mh012_	0.0002 / 0.9856	0.0004 / 283	Eating more or less
254	partnerinh...	0.0002 / 0.9857	0.0004 / 299	Partner in household - after CA update
255	wq010_	0.0002 / 0.9859	0.0006 / 244	Work had adequate salary
256	wq014_	0.0002 / 0.9861	0.0007 / 211	Work health risk reduced
257	ex008_	0.0002 / 0.9862	0.0004 / 283	Chance government raises retirement age
258	cc733_	0.0002 / 0.9864	0.0006 / 235	Family was pretty well off financially, about average, or poor
259	gs008_	0.0002 / 0.9865	0.0006 / 244	1st measurement: right hand
260	relrpers	0.0002 / 0.9867	0.0008 / 203	Relation to coverscreen respondent
261	mh026_	0.0002 / 0.9868	0.0005 / 256	Fear of dying
262	br028_	0.0002 / 0.9870	0.0007 / 211	How often serving of meat, fish or chicken
263	ex003_	0.0001 / 0.9871	0.0006 / 235	Chance inheritance more than 50000
264	ch007_SHL_...	0.0001 / 0.9873	0.0008 / 203	Child 1 (SHARELIFE: biological) where does child live
265	pf004_	0.0001 / 0.9874	0.0006 / 244	Value second measurement
266	cs002_	0.0001 / 0.9876	0.0006 / 235	Safe to do cs
267	hc602_	0.0001 / 0.9877	0.0007 / 217	Times talked to medical doctor/nurse about your health last 12 months
268	rp004b_1	0.0001 / 0.9879	0.0006 / 225	Year started living with married partner
269	ep016_	0.0001 / 0.9880	0.0006 / 235	Name or title of job
270	ch016_2	0.0001 / 0.9881	0.0006 / 225	Child 2 employment status
271	dn027_2	0.0001 / 0.9883	0.0006 / 225	Age of death of parent: father
272	hc005_	0.0001 / 0.9884	0.0006 / 244	Most recent consulted specialist
273	fs004_	0.0001 / 0.9885	0.0001 / 407	Ever had any mutual funds
274	hc841dno	0.0001 / 0.9887	0.0005 / 270	Forgo care due to cost: none of these
275	cf116tot	0.0001 / 0.9888	0.0006 / 225	Ten words list learning delayed recall total
276	xt038e_3	0.0001 / 0.9889	0.0005 / 270	Value of assets: cars
277	re002_	0.0001 / 0.9890	0.0005 / 256	Year finished fulltime education
278	yrbirth_xt	0.0001 / 0.9892	0.0005 / 256	Year of birth of the deceased
279	ph009_15	0.0001 / 0.9893	0.0004 / 299	Age other fractures
280	wq008_	0.0001 / 0.9894	0.0004 / 283	Work allowed development of skills
281	hc002_	0.0001 / 0.9896	0.0006 / 225	How often seen or talked to medical doctor last 12 months
282	cc729_	0.0001 / 0.9897	0.0004 / 299	Lonely for friends in childhood
283	as003ub	0.0001 / 0.9898	0.0006 / 244	Amount bank account ub
284	ex004_	0.0001 / 0.9899	0.0004 / 299	Chance of leaving inheritance more than 50000

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
285	ho024e	0.0001 / 0.9900	0.0004 / 283	Value of property
286	xt009_	0.0001 / 0.9902	0.0004 / 283	Year of decease
287	rh026dot	0.0001 / 0.9903	0.0005 / 270	Why no regular dental care: other reasons
288	re011_1	0.0001 / 0.9904	0.0004 / 283	Year started job
289	br030_	0.0001 / 0.9905	0.0005 / 270	How many cups a day drinks of tea, coffee, water, milk, fruit, soft drinks
290	as054dno	0.0001 / 0.9906	0.0003 / 323	Owe money: none of these
291	hh011e	0.0001 / 0.9907	0.0003 / 323	Additional income received by all hh-members last year
292	wave	0.0001 / 0.9908	0.0004 / 313	wave
293	iv010_	0.0001 / 0.9910	0.0005 / 256	Type of building
294	rh025_	0.0001 / 0.9911	0.0005 / 256	Frequency regular dentist
295	ra006_2	0.0001 / 0.9912	0.0005 / 256	Start living at residence
296	hc068_8	0.0001 / 0.9913	0.0004 / 283	Current health insurance coverage: private hospitals
297	dn019_	0.0001 / 0.9914	0.0005 / 256	Since when widowed
298	ph044_	0.0001 / 0.9915	0.0005 / 270	Eyesight reading
299	ra006_1	0.0001 / 0.9916	0.0004 / 299	Start living at residence
300	hc014_	0.0001 / 0.9917	0.0005 / 256	Total nights stayed in hospital
301	ep329_	0.0001 / 0.9918	0.0004 / 283	Retirement year
302	dn127_1	0.0001 / 0.9919	0.0005 / 256	Year of death of parent: mother
303	br027_	0.0001 / 0.9921	0.0005 / 270	How often serving of legumes or eggs
304	gl003_	0.0001 / 0.9922	0.0004 / 313	When happiness period started
305	gender	0.0001 / 0.9923	0.0004 / 299	Male or female
306	ph009_18	0.0001 / 0.9924	0.0004 / 313	Age affective or emotional disorders
307	ph013_	0.0001 / 0.9925	0.0005 / 270	How tall are you?
308	sn006_2	0.0001 / 0.9926	0.0005 / 270	Network proximity: sn person 2
309	ch007_2	0.0001 / 0.9927	0.0005 / 270	Child 2 where does child live
310	ep678v2	0.0001 / 0.9928	0.0001 / 407	Bracket value 2
311	re026_1	0.0001 / 0.9929	0.0004 / 313	Year stopped in this job
312	hh017ub	0.0001 / 0.9930	0.0005 / 270	Total income received by all hh-members an average month last year ub
313	gl740d5	0.0001 / 0.9931	0.0001 / 407	Discriminated against father: engaged in combat operations/fighting
314	br016_	0.0001 / 0.9932	0.0003 / 323	Activities requiring a moderate level of energy
315	dn127_2	0.0001 / 0.9933	0.0005 / 270	Year of death of parent: father
316	rp011_1	0.0001 / 0.9934	0.0004 / 283	Year of death partner
317	ch006_2	0.0001 / 0.9935	0.0004 / 299	Child 2 year of birth
318	ph012_	0.0001 / 0.9936	0.0004 / 283	Weight of respondent

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
319	st006_	0.0001 / 0.9937	0.0004 / 283	Month of birth of respondent
320	hc068_10	0.0001 / 0.9938	0.0004 / 313	Current health insurance coverage: nursing care at home
321	ph009_13	0.0001 / 0.9939	0.0004 / 313	Age cataracts
322	rel_relati...	0.0001 / 0.9939	0.0004 / 313	Relation to household member (wave 8)
323	re022c_4	0.0001 / 0.9940	0.0004 / 299	Currency of wage - coded
324	rc024_2	0.0001 / 0.9941	0.0004 / 299	Year of birth other child
325	cf107tot	0.0001 / 0.9942	0.0004 / 299	Ten words list learning first trail
326	pf002_	0.0001 / 0.9943	0.0003 / 333	Feels safe to do the test
327	ep078e_4	0.0001 / 0.9944	0.0002 / 346	Average payment income source c4 last year
328	dn051_2	0.0001 / 0.9945	0.0004 / 283	Highest school certificate/degree: father
329	sp019d22	0.0001 / 0.9945	0.0002 / 346	R provided help with personal care to: grandchild
330	age2007	0.0001 / 0.9946	0.0002 / 346	Age in 2007
331	rh050dot	0.0001 / 0.9947	0.0003 / 323	Why no regular blood pressure checks: other reasons
332	ph089d2	0.0001 / 0.9948	0.0002 / 346	Bothered by frailty: fear of falling down
333	age2004	0.0001 / 0.9949	0.0003 / 323	Age in 2004
334	ep050_	0.0001 / 0.9949	0.0003 / 323	Year last job ended
335	cf015_	0.0001 / 0.9950	0.0004 / 313	Numeracy: amount in the savings account
336	rp008_1	0.0001 / 0.9951	0.0003 / 323	Year married
337	ep051_	0.0001 / 0.9952	0.0004 / 313	Employee or a self employed in last job
338	mc010_	0.0001 / 0.9952	0.0003 / 323	Childhood health status
339	ra015_2	0.0001 / 0.9953	0.0003 / 323	Region of residence (not current)
340	dn051_1	0.0001 / 0.9954	0.0004 / 313	Highest school certificate/degree: mother
341	mh015_	0.0001 / 0.9955	0.0002 / 366	Concentration on reading
342	br025_	0.0001 / 0.9955	0.0002 / 366	How many meals a day
343	ep213_4	0.0001 / 0.9956	0.0003 / 333	First year received income source c4
344	ft002_	0.0001 / 0.9957	0.0002 / 346	Given financial gift 250 or more
345	ph048d10	0.0001 / 0.9958	0.0002 / 366	Difficulties: picking up a small coin from a table
346	hc049v1	0.0001 / 0.9958	0.0002 / 346	Bracket value 1
347	sp019d12	0.0001 / 0.9959	0.0003 / 333	R provided help with personal care to: child 3
348	dn012d18	0.0001 / 0.9960	0.0002 / 387	Further education: country-specific category 18
349	dn023dno	0.0001 / 0.9960	0.0002 / 366	Further education former partner: none

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
350	sp016_1	0.0001 / 0.9961	0.0002 / 346	How often did you look after child of child 1
351	ep128_13	0.0001 / 0.9962	0.0002 / 366	Period from year (working)
352	ph087d1	0.0001 / 0.9963	0.0001 / 407	Pain location: back
353	gs004_	0.0001 / 0.9963	0.0002 / 346	Dominant hand
354	age2015	0.0001 / 0.9964	0.0002 / 366	Age in 2015
355	cf016tot	0.0001 / 0.9965	0.0002 / 346	Ten words list learning delayed recall total
356	co004e	0.0001 / 0.9965	0.0002 / 366	Amount spent on telephones in last month
357	ph009_8	0.0001 / 0.9966	0.0003 / 333	Age arthritis or rheumatism
358	hc068_5	0.0001 / 0.9966	0.0002 / 366	Current health insurance coverage: dental care
359	dn006_	0.0001 / 0.9967	0.0003 / 333	Year came to live in country
360	ph048d8	0.0001 / 0.9968	0.0002 / 366	Difficulties: pulling or pushing large objects
361	as068_	0.0001 / 0.9968	0.0003 / 333	Risk aversion
362	agep2017	0.0001 / 0.9969	0.0002 / 387	Age of partner in 2017
363	it004_	0.0001 / 0.9970	0.0001 / 407	Use of internet in past 7 days
364	re013_3	0.0001 / 0.9970	0.0003 / 333	Job description
365	sp011_1	0.0001 / 0.9971	0.0003 / 333	How often given help to person 1
366	as055e	0.0001 / 0.9971	0.0002 / 366	Amount owing money in total
367	age2013	0.0001 / 0.9972	0.0002 / 366	Age in 2013
368	ra015_3	0.0001 / 0.9972	0.0002 / 366	Region of residence (not current)
369	ep030_	0.0001 / 0.9973	0.0002 / 366	Opportunity to develop new skills in (main) job
370	cc008_	0.0001 / 0.9973	0.0002 / 366	Number of books when ten
371	cc725_2	0.0001 / 0.9974	0.0002 / 346	Father physical harm
372	ra015_1	0.0001 / 0.9974	0.0002 / 346	Region of residence (not current)
373	dn015_	0.0001 / 0.9975	0.0002 / 366	Year of marriage, if living together
374	ph009_5	0.0001 / 0.9976	0.0002 / 346	Age diabetes
375	hh024_	0.0001 / 0.9976	0.0002 / 346	Area is kept very clean
376	ra006_3	0.0001 / 0.9977	0.0002 / 366	Start living at residence
377	hs063d4	0.0001 / 0.9977	0.0002 / 387	Consequences of illness period: made social life more difficult
378	hc034_	0.0000 / 0.9978	0.0002 / 366	Hours received professional nursing care
379	ep098dno	0.0000 / 0.9978	0.0002 / 387	Type of pension entitled to: none of these
380	ep010_	0.0000 / 0.9979	0.0002 / 346	Start of current job (year)
381	ho067v1	0.0000 / 0.9979	0.0002 / 366	Bracket value 1
382	ho035_	0.0000 / 0.9980	0.0002 / 366	Years in community

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
383	rh789_	0.0000 / 0.9980	0.0002 / 387	Number of periods postponed dentist visit
384	ho024v2	0.0000 / 0.9980	0.0002 / 387	Bracket value 2
385	hc035_	0.0000 / 0.9981	0.0002 / 366	Weeks received paid domestic help
386	co011e	0.0000 / 0.9981	0.0002 / 387	Value of home produced food
387	ph095_	0.0000 / 0.9982	0.0002 / 366	How much loss of weight (in kg)
388	hc033_	0.0000 / 0.9982	0.0002 / 387	Weeks received professional nursing care
389	hh017v2	0.0000 / 0.9983	0.0002 / 387	Bracket value 2
390	ph062_	0.0000 / 0.9983	0.0002 / 387	Compare health last wave
391	xt019e_1	0.0000 / 0.9984	0.0002 / 387	Costs: care from a general practitioner
392	ex006_	0.0000 / 0.9984	0.0002 / 346	Chance of leaving inheritance more than 150000
393	ph009_6	0.0000 / 0.9985	0.0002 / 366	Age chronic lung disease
394	hh017v1	0.0000 / 0.9985	0.0002 / 387	Bracket value 1
395	xt019e_6	0.0000 / 0.9985	0.0001 / 407	Costs: medication
396	age2017	0.0000 / 0.9986	0.0002 / 387	Age in 2017
397	partnerinh...	0.0000 / 0.9986	0.0001 / 407	Partner in household
398	rc024_3	0.0000 / 0.9987	0.0002 / 387	Year of birth other child
399	ho005e	0.0000 / 0.9987	0.0002 / 387	Amount last rent payment
400	ch020_3	0.0000 / 0.9987	0.0002 / 387	Child 3 year of birth youngest child
401	rh793_	0.0000 / 0.9988	0.0001 / 407	Number of periods postponed taking medication
402	hhsize_upd...	0.0000 / 0.9988	0.0001 / 407	Household size - after CA update
403	ch007_3	0.0000 / 0.9988	0.0002 / 366	Child 3 where does child live
404	st007_	0.0000 / 0.9989	0.0001 / 429	Year of birth of respondent
405	ep201e	0.0000 / 0.9989	0.0001 / 407	Taken home from work after tax, (main) job
406	iv014_	0.0000 / 0.9990	0.0002 / 387	Age of interviewer
407	rh041_	0.0000 / 0.9990	0.0002 / 387	Year regular blood pressure checks started
408	sp019d10	0.0000 / 0.9990	0.0001 / 407	R provided help with personal care to: child 1
409	re704_	0.0000 / 0.9990	0.0001 / 429	Computer training for job
410	dn018_	0.0000 / 0.9991	0.0001 / 407	Since when divorced
411	ho024v3	0.0000 / 0.9991	0.0001 / 407	Bracket value 3
412	hc049e	0.0000 / 0.9991	0.0001 / 407	Paid out-of-pocket for prescribed drugs
413	gs010d6	0.0000 / 0.9992	0.0001 / 407	Why not completed gs test: r had surgery, injury, swelling, etc.
414	rh790_3	0.0000 / 0.9992	0.0001 / 429	Years postponed dentist visit 1 to 5
415	sn023_1	0.0000 / 0.9992	0.0001 / 407	Reason did not mention sn person 1

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
416	agep2015	0.0000 / 0.9992	0.0001 / 407	Age of partner in 2015
417	gl740dno	0.0000 / 0.9993	0.0001 / 429	Discriminated against father: none of these
418	ch020_2	0.0000 / 0.9993	0.0001 / 407	Child 2 year of birth youngest child
419	ep106_1	0.0000 / 0.9993	0.0001 / 407	Expected age to collect pension c1
420	co005e	0.0000 / 0.9993	0.0001 / 407	Amount spent on all goods and services in last month
421	rh794_2	0.0000 / 0.9994	0.0001 / 429	Years postponed taking medication 1 to 5
422	ep049_	0.0000 / 0.9994	0.0001 / 407	Years working in last job
423	xt025_	0.0000 / 0.9994	0.0001 / 407	Hours of help necessary during typical day
424	xt007_	0.0000 / 0.9994	0.0001 / 429	Year of birth proxy
425	ch019_2	0.0000 / 0.9995	0.0001 / 429	Child 2 number of children
426	rc029_1	0.0000 / 0.9995	0.0001 / 429	Left job because of child
427	cf105tot	0.0000 / 0.9995	0.0001 / 429	Ten words list learning first trail
428	sn027_1	0.0000 / 0.9995	0.0001 / 429	Year of birth sn person 1
429	ep078e_12	0.0000 / 0.9995	0.0000 / 456	Typical payment of pension in c12 last year (ep324d2)
430	ch015_2	0.0000 / 0.9996	0.0001 / 429	Child 2 year moved out
431	ph046_	0.0000 / 0.9996	0.0001 / 429	Hearing
432	dn029isco_...	0.0000 / 0.9996	0.0001 / 429	ISCO code of father when respondent was 10
433	cs003_	0.0000 / 0.9996	0.0001 / 429	Number of people living in household when ten
434	chselch2	0.0000 / 0.9996	0.0001 / 429	Child number 2 selected child
435	as003v1	0.0000 / 0.9996	0.0001 / 429	Bracket value 1
436	as049_	0.0000 / 0.9997	0.0001 / 429	Number of cars
437	re007_	0.0000 / 0.9997	0.0001 / 429	Situation in gap after education
438	ph009_4	0.0000 / 0.9997	0.0001 / 429	Age stroke or cerebral vascular disease
439	sr004d1	0.0000 / 0.9997	0.0001 / 429	Negative Shock: Bad health affected work
440	ep078e_5	0.0000 / 0.9997	0.0000 / 456	Average payment income source c5 last year
441	sp009_2	0.0000 / 0.9997	0.0001 / 429	To whom did you give help: person 2
442	rc024_4	0.0000 / 0.9997	0.0001 / 429	Year of birth other child
443	rh782_1	0.0000 / 0.9998	0.0000 / 456	Years could not afford doctor 1 to 5
444	ph009_7	0.0000 / 0.9998	0.0000 / 456	Age asthma
445	rh790_2	0.0000 / 0.9998	0.0000 / 456	Years postponed dentist visit 1 to 5
446	xt038e_4	0.0000 / 0.9998	0.0000 / 456	Value of assets: financial assets, e.g. cash, bonds or stocks
447	co002v1	0.0000 / 0.9998	0.0000 / 456	Bracket value 1

Continued on next page

	Feature	Gain / Cumulative	Split / Rank	Description
448	ph010d12	0.0000 / 0.9998	0.0000 / 456	Bothered by: fatigue
449	ch001_	0.0000 / 0.9998	0.0000 / 456	Number of children
450	ph009_3	0.0000 / 0.9998	0.0000 / 456	Age high blood cholesterol
451	rc028_2	0.0000 / 0.9998	0.0000 / 456	Year of death other child
452	mc007_	0.0000 / 0.9999	0.0000 / 456	Relative position to others when 10: language
453	ra021_3	0.0000 / 0.9999	0.0000 / 456	Stopped living at residence
454	ph054_	0.0000 / 0.9999	0.0000 / 456	Who answered the questions in ph
455	sp011_2	0.0000 / 0.9999	0.0000 / 456	How often given help to person 2
456	as003v2	0.0000 / 0.9999	0.0000 / 456	Bracket value 2
457	rh781_	0.0000 / 0.9999	0.0000 / 456	Number of periods could not afford doctor
458	ra733_3	0.0000 / 0.9999	0.0000 / 456	Start year living first time: Mother-in-law
459	ho008e	0.0000 / 0.9999	0.0000 / 456	Amount charges and services
460	ph048d4	0.0000 / 0.9999	0.0000 / 456	Difficulties: climbing several flights of stairs
461	ho012_	0.0000 / 0.9999	0.0000 / 456	Year acquired property
462	xt024_	0.0000 / 0.9999	0.0000 / 456	Time the deceased received help
463	ph048d6	0.0000 / 0.9999	0.0000 / 456	Difficulties: stooping, kneeling, crouching
464	ph004_	0.0000 / 1.0000	0.0000 / 456	Long-term illness
465	xt019e_3	0.0000 / 1.0000	0.0000 / 456	Costs: hospital stays
466	ep103_1	0.0000 / 1.0000	0.0000 / 456	Years contributing to plan, pension c1
467	sp020_	0.0000 / 1.0000	0.0000 / 456	Someone in this household helped you regularly with personal care
468	ws013_	0.0000 / 1.0000	0.0000 / 456	Time of second walking speed test
469	br019_	0.0000 / 1.0000	0.0000 / 456	How many drinks in a day
470	ep082e_1	0.0000 / 1.0000	0.0000 / 456	Total amount of lump sum payment income source c1
471	ch015_3	0.0000 / 1.0000	0.0000 / 456	Child 3 year moved out