Utrecht University

Usability Study of an Explainable Machine Learning Risk Model for Predicting Illegal Shipbreaking

Casper de Haas

Utrecht University

University of Utrecht

Usability Study of an Explainable Machine Learning Risk Model for Predicting Illegal Shipbreaking

Master's Thesis

To fulfill the requirements for the degree of
Master of Information Science in Human Computer Interaction
at University of Utrecht under the supervision of
Dr. H. Kaya (Department of Information and Computing Sciences, University of Utrecht)
and
Dr. A.A.A. (Hakim) Qahtan (Department of Information and Computing Sciences, University of Utrecht)
and at ILT under the supervision of
Dr. S.I. (Stephanie) Wassenburg (Department of IDLab, ILT)
P.P.A.B. (Paul) Merkx, MSc (Department of IDLab, ILT)

Casper de Haas (5679303)

November 1, 2021

Contents

Acknowledgments

I would like to thank the following people for helping me with this research project:
My first supervisor from the Utrecht University Heysem Kaya for all your help guiding me through this project. Second supervisor Hakim Qahtan for giving me interesting feedback during the last part of the thesis. My external supervisors from ILT Stephanie Wassenburg and Paul Merkx for their amazing support and feedback during the weekly meetings. The colleagues at IDLab for providing feedback during presentations and participating in the pilot study. The team of waste inspectors that were interviewed and provided me with helpful feedback. The team of port state controllers for participating in the online experiment. And finally, my family for their support during the project.

Abstract

The goal of this project conducted at the IDLab of the ILT was to implement information driven solutions to support inspection tasks with the use of Artificial Intelligence AI. The AI solutions should complement current work activities by providing predictions based on historic data. The algorithms and data processing add complexity to the AI models. The AI models therefore are required to be made understandable for future users, by creating insight into the workings of the model through transparency and trust, so that end users can adopt the model predictions into their work activities. For this reason, an information representation is required that presents useful information and model predictions to the user. The explanations of individual model predictions should provide insight into the workings of the model to support users with their decision making process and increase trust. This requires techniques to make model output interpretable and explainable.

In this study Explainable Artificial Intelligence (XAI) has been used to make a complex AI model interpretable for end users while bridging the gap between data science and the field of application. The SHapley Additive exPlanations (SHAP) method was used to explain individual predictions. By researching the effectiveness of additional visual- and textual model explanations on the trust and perceived usefulness of the end user, a decision support system that better matches expectations of end users could be developed. A predictive model for detecting illegal shipbreaking has been evaluated with end users. The model predicts the likelihood of the event of beaching, which means the illegal dismantling of ships at beaches in Asia. The beaching model uses features of ships to predict whether the ship is likely to be beached. Semi-structured interviews that used the think-aloud method were conducted with experts during the first part of the research to evaluate the implementation of SHAP and elicit detailed information on the effect of SHAP-based explanations. After analyzing the results from the interviews, the findings were used to create the final online user testing experiment. Qualtrics was used to create an online experiment to evaluate SHAP-based information representations. A group of Port State Controllers were invited to participate in the experiment. The provided SHAP-based explanations that were used during the experiment described individual predictions of the beaching model. The individual instances or ships were evaluated with and without additional visual- and textual explanations from predictions of the beaching model. Metrics on System Usability Score (SUS), cognitive load and response time were gathered during the experiment. Statistical tests were performed to create findings on the proposed hypotheses.

Results from the final experiment showed the effects of visual- and textual model output explanations and how to effectively implement interpretable machine learning solutions. These findings contributed to a set of guidelines for designing machine learning applications while focusing on usability and interpretability of the model output. It was found that the SHAP waterfall plot visualisation as an addition to the feature values and the prediction score does not necessarily contribute to the usability of the explanation. SHAP-based text explanations however did provide a significant positive contribution to the usability in terms of improving understandability of individual explanations. By integrating the perceptions of explainablity between inspectors and data scientists, the adoption and sustained use of machine learning systems within ILT could be facilitated.

List of Abbreviations

AI Artificial Intelligence. *Glossary:* AI

API Application Programming Interface. *Glossary:* API

GISIS Global Integrated Shipping Information System. *Glossary:* GISIS

HCI Human Computer Interaction. *Glossary:* HCI

IDLab Innovation- and Data Lab. *Glossary:* IDLab

ILT Inspectie Leefomgeving en Transport. *Glossary:* ILT

LIME Local Interpretable Model-agnostic Explanations. *Glossary:* LIME

NGO Shipbreaking Platform Non-Governmental Organisation Shipbreaking Platform. *Glossary:* NGO Shipbreaking Platform

NLG Natural Language Generation. *Glossary:* NLG

SBRI Ship Breaking and Recycling Industry. *Glossary:* SBRI

SHAP SHapley Additive exPlanations. *Glossary:* SHAP

SUS System Usability Scale. *Glossary:* SUS

XAI eXplainable Artificial Intelligence. *Glossary:* XAI

Glossary

AI Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.. 6, 13

GISIS GISIS (Global Integrated Shipping Information System) is developed, maintained and headed by the International Maritime Organisation (IMO).. 9

HCI Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers.. 13

IDLab The Innovation- and Data Lab (IDLab) uses state-of-the-art techniques to analyse big amounts of data to create innovative solutions for the ILT.. 6

ILT The ILT, which stands for Human Environment and Transport Inspectorate, works at improving safety, confidence and sustainability in regard to transport, infrastructure, environment and housing.. 6, 9

LIME Local surrogate models are interpretable models that are used to explain individual predictions of black box machine learning models.. 15

NGO Shipbreaking Platform The NGO Shipbreaking Platform is a global coalition of organisations working to reverse the environmental harm and human rights abuses caused by current shipbreaking practices and to ensure the safe and environmentally sound dismantling of end-of-life ships worldwide.. 9

NLG Natural language generation (NLG) is the use of artificial intelligence (AI) programming to produce written or spoken narratives from a data set.. 10, 20

SBRI The ship breaking and recycling industry (SBRI) converts end-of-life ships into steel and other recyclable items.. 9

SHAP SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. 6, 10, 18

SUS The System Usability Scale (SUS) is a reliable tool for measuring the usability.. 6, 24

XAI eXplainable Artificial Intelligence (XAI) is artificial intelligence in which the results of the solution can be understood by humans. It contrasts with the concept of the ”black box” in machine learning where even its designers cannot explain why an AI arrived at a specific decision.. 6, 10

1 Introduction

1.1 Problem Statement

The Inspectie Leefomgeving en Transport (ILT) has the task to ensure that companies, organisations, and government agencies comply with the laws and regulations concerning environmental sustainability, physical safety and the housing corporation sector. The purpose of the IDLab is to increase the use of data-driven AI solutions that support inspection tasks within the ILT. Predictive information systems can provide helpful insights which can be used to take timely proactive actions.

At the ILT there is a need to create a system for inspectors that could predict whether a ship has a high risk of committing a violation. This violation would comprise the illegal act of having their ship dismantled in developing countries, also known as 'beaching'. When a ship reaches its end-of-life, the owner can decide to follow European rules regarding shipbreaking or resort to illegal ways of recycling the ship. The shipbreaking method of 'beaching' describes the open-beach breaking of end-of-life ships. Barua et al. describes that beaching is performed by the Ship Breaking and Recycling Industry (SBRI) and creates the exposure of hazardous materials to employees and the environment [1]. Employees have to work in extremely dangerous and unhealthy conditions while dismantling the ships.
The shipbreaking risk model developed by the IDLab department of the ILT can be used to predict whether a ship is likely to be beached. This allows inspectors to get into contact with the owner of the ship to inform them about the needed permits for shipbreaking to prevent the ship from being beached. The predictions of the model are used to intervene and notify the shipowners with end-of-life ships that they are being monitored. By preventing the illegal beaching method of shipbreaking, the model can make a contribution to people and the environment. This requires the predictions from the shipbreaking model to be easily explainable to supervised entities. End users should also trust the model predictions through model accuracy and explanation fidelity [2]. This could for example be achieved with an interactive dashboard that provides the inspector with visualisations of the predictions of the model. Providing system feedback to the end user can also be supported by natural language processing using semantic classification to create standardized reports [3].

To create a predictive model capable of making accurate predictions on possible illegal behaviour, a large dataset was needed. This dataset included the current records of already dismantled ships on the NGO Shipbreaking Platform between 2016 and 2019. Additional specifications of the ships were gathered and added to the dataset through web scraping. These specifications were scraped from the GISIS information system [4] and matched to the specific ship. In the Python programming environment, the random forest ensemble machine learning method was used to train (or create) two random forest models based on the shipbreaking data: 1) The shipbreaking model looks at active ships and dismantled ships while mostly containing technical features to predict whether a ship is likely to be dismantled. 2) The model on beaching behaviour only includes ships that were either dismantled or beached in order to make a prediction on illegal beaching. Information on feature values of ships from historic data on past shipbreaking events were used to predict future shipbreaking events. The beaching model makes its predictions predominantly on flag behaviour in combination with other features. Both models share some features on individual ship entries.
The performance of the previously trained models on shipbreaking and beaching were evaluated with cross validation. An implementation set for testing the model has been created that can be used in the field. The implementation set includes currently active ships with some form of Dutch owner-

ship. To create an interpretable output containing the risk factors and feature importance that lead to a prediction for each individual ship both models use SHAP scores, which stand for *SHapley Additive exPlanations* [5]. SHAP compares the prediction of the model with and without a specific feature to calculate the feature importance. Currently, the outputs of the models are in the form of an excel sheet, where SHAP values are represented through conditional formatting. This representation format was chosen to conveniently display the different feature values and SHAP scores calculated by the models. However, the excel file is just a temporary solution for displaying the model output. SHAP offers different types of representations for visualising individual predictions of the model. Therefore, research is necessary on what type of representation would be the most effective in terms of explainability, transparency, usability, informativeness, etc.

Interpretable machine learning is essential in the development of explainable models in predictive modelling. A machine learning model can have explanations of its decisions that lead to an increased level of interpretability of the model, making it easier to be interpreted by users, therefore increasing their likeliness to use the model. When users understand the models' decisions and the decision making process, it becomes more intuitive to incorporate predictions from machine learning into their daily work activities. The field of work that focuses on making machine learning more explainable is called explainable artificial intelligence (XAI). In the field of XAI there is still much to be achieved on making models more explainable to end users [6]. It is however important to involve the user into the development process of XAI systems (e.g. to prevent reasoning failures due to cognitive biases) [7].
This project focused on the beaching model since the results from this model can be used to evaluate the shipbreaking model. Evaluating both models would be too time-consuming for this study. The beaching model contains information concerning ship features that are commonly used in the areas of illegal beaching activity of ships. This requires that the participants that are invited to the study had to have knowledge in the field of ships. The performance and evaluation results of the beaching model were used to evaluate the shipbreaking model. The focus of the project was on the end users of the explainable machine learning application. This raised questions such as: 'How do we make the predictions from the model understandable for the end users?', 'How do we present the prediction?', and finally 'Does the current presentation of the prediction raise enough trust within the users?'. It also encompassed a review of different methods in the field of explainable artificial intelligence. An in-depth user study using semi-structured qualitative interviews provided insight concerning end user requirements to make the transition into AI-supported decision making. Based on the results from the interviews, a user study was developed. The user study included SHAP visualisations and additional natural language generated (NLG) texts to explain the model predictions. The calculated SHAP score for each individual feature is used to create a textual explanation of the model prediction for a subset of features based on their individual contribution to the prediction. A total of 15 user tests were performed to measure the interpretability and informativeness of the different model explanations.

1.2 Research Questions

To summarize, this thesis focuses on the following problems:
The main research question:

RQ1. *"How does the inclusion of SHAP-based instance wise explanation in a random forest ensemble model affect the users' view in terms of system usability scale?"*

With the corresponding sub-questions:

SQ1. *"What is the effect of the newly proposed representations of SHAP-based explanations on the systems' usability score?"*

SQ2. *"What is the effect of the newly proposed representations of SHAP-based explanations on the perceived cognitive load?"*

The second research question focused on the evaluation of the natural language generation (NLG) method by implementing text explanations as an addition to the SHAP explanation. Therefore the second research question was defined as:

RQ2. *"Using natural language generation (NLG) as an addition to SHAP representations, how does the SUS improve w.r.t. the model interpretability and individual explanations provided to the end users?"*

With the corresponding sub-questions:

SQ1. *"How many features should be included in the text explanations with respect to the informativeness and interpretability of the explanation?"*

SQ2. *"How well do the newly proposed NLG explanations as an addition to the visualisations convince the end-users and what is their effect on the systems' usability score and perceived cognitive load?"*

The above RQs were be answered for a new set of SHAP visualisations and NLG text explanations to be evaluated looking at the interpretability and informativeness of the SHAP and NLG explanations.

1.3 Thesis Outline

Below holds the current outline of the proposed research and the corresponding methods. The methods are chronologically ordered to indicate different stages in the research process.

1. **Literature Review**
   An in depth literature review was performed to acquire an understanding of the current research within the field. The current gap in the literature was also pointed out during the literature study phase. This created the possibility to effectively formulate possible contributions that this study could potentially make to the current body of knowledge. The qualitative research methods for evaluating interactive systems have been reviewed in the literature study section. The applicability of the different methods will now be discussed below. To gather insights into the field of research it is necessary to characterize and reflect on qualitative research methods to decide which method(s) will be used for the future study.

2. **Expert Interviews**
   The next step was to gather information from participants on the usability of the system through a number of expert interviews. The individual expert interviews guided the participant through several steps while measuring their actions through the 'think-aloud' method. This method asked the participant to vocalize his or her thoughts and actions during the interviews. A short qualitative interview at the start of the experiment created an overview of the individual user experience in the field of ML. The next part of the experiment consists of a number of scenarios for the participant to solve with- or without the support of the system. At the end of the experiment there was room for questions.

3. **Pilot and Testing**
   To gather information on the effect of alternative representations on the system usability scale, a pilot was made and tested with participants from the IDLab. The pilot was created in the form of static instances that allow the users to explore the data and several alternative representations. The pilot asked the participant to solve a number of scenarios with- or without the availability of additional visual- and/or textual representations of the model output. The actions of the participants were recorded through the Qualtrics questionnaire environment. The participants answered several questions during- and after the experiment. After completing the questionnaire, there was some room for remarks from the participants.

4. **Online User Experiment**
   A new set of natural text explanations were created and added to the proposed SHAP visualisations. The effect of the natural language explainability and the individual text explanations of the additional information representations on the systems usability scale was tested. The previously designed online questionnaire was used to measure the effect of additional text explanations of the proposed natural language explanation representations based on the system usability score.

2 Literature Review

This section includes the literature review on the topics of explainable artificial intelligence and different approaches on interpretable machine learning.

2.1 Explainable Artificial Intelligence

2.1.1 AI System Transparency

Decision support through machine learning has become a part of everyday life. While current research focuses on transparent machine learning methods, there is a need for investigating the interaction between humans and Artificial Intelligence (AI). How does human-AI interaction benefit from current advancements in machine learning transparency? A study on the impact of transparency and the effect of risk and ambiguity on the users' trust in the AI system stated that transparency does not necessarily improve human-AI interaction [8]. The study also shows that the calibration of the level of transparency within an AI system, between the AI system and human users, is required to prevent algorithmic bias. Decreasing the algorithmic bias that is caused by the transparency of the AI system can help in more responsible usage of AI systems. For optimal calibration, it is necessary to have a comprehensive understanding for both sides of the interaction. This requires insight into the functionality of the system but also knowledge about the personality traits of human users. Figure 1 shows a framework for measuring the effectiveness of explanations which can help with analysing the usability of individual explanations. The design of intelligent systems is a Human Computer Interaction (HCI) problem as stated by Stumpf et al. [9]. Meaningful interaction includes approaches that are useful for the user and the system. This requires continuous evaluation with end users, as seen in Figure 2 However, the human part is currently underrepresented in the literature, creating a gap between transparent AI systems and users. Therefore, more research should be conducted on bridging the gap between AI system transparency and user expectations.

2.1.2 XAI Goal

The goal of explainable artificial intelligence (XAI) is to produce explainable models using new or modified ML techniques. These models enable the end-user to understand, trust and manage the system [10]. The concept of XAI is therefore to provide explanations to users so that they understand the system's decision-making model and are able to appropriately use it. An example of how an XAI system query would look like can be seen in Figure 6. The concept is user-centered proposing several questions:(1) how to produce more explainable models, (2) how to design explanation interfaces, (3) how to understand the psychologic requirements for effective explanations.

2.1.3 XAI Applications

Explanations of decisions are often essential to establish trust between people. Although these social norms can be less important for AI systems there are still many reasons for making artificial intelligence explainable. The most important reasons why XAI is needed are:

- *Verification of the system:* black box systems can not be trusted due to the lack of verifiability. Especially in health care the need for interpretable and verifiable models is high since there is less room for error as stated by Vellido [11].

- *Improvement of the system:* weaknesses in systems that are interpretable are easier to point out. Biases in the model or dataset are also detected with less effort. The interpretability of the model also aids the process of selecting the most appropriate model. Therefore an increased understanding of the model and its decision making makes it easier to make improvements to it.

- *Learning from the system:* AI systems are trained on massive datasets to observe patterns in the data. With explainable AI systems we can gather insights from this distilled knowledge in the previously observed patterns.

- *Compliance to legislation:* AI systems are used in all kinds of situations. Some situations require the assignment of responsibility when the decision from the system is wrong. The legal aspect of being able to explain a decision is required to adapt to new European regulations on the "right to explanation" [12]. This describes the right to get an explanation for a specific output of an algorithm. An example could be the explanation of the decision making of a model used by a financial organisation on distributing loans based on health and income. The client has the right to receive an explanation on the decision of the algorithm.

2.2 Machine Learning

Machine learning is a field of artificial intelligence that focuses on building applications that automatically learn from data and improve from experience. Machine learning algorithms are used to create models that can make predictions based on historic data. Supervised learning is the process of creating a predictive model by training with an appropriate learning algorithm (e.g. random forests).

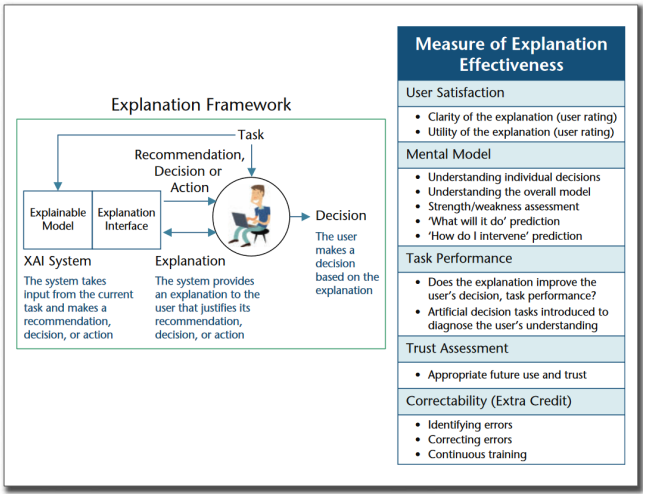


Figure 1: Measure of Explanation Effectiveness. Adopted from [10].

2.2.1 Decision Trees

A decision tree is a tree-like model that consists of multiple branches where each branch represents a possible decision. It is used to map the possible outcomes of a series of related choices by classifying the examples also known as a training set. The examples are sorted in a tree format. The decision tree typically starts with a single node which splits into different outcomes. Every split is a leaf/node and provides a classification for the example. It can be used to mathematically predict the best possible choice. For every node in the tree a specific attribute is considered. All edges that descend from the node indicate the outcome/classification for that example. For every subtree the process of sorting the examples by splitting at every leaf/node is repeated.

2.2.2 Random Forest Classifier

A random forest classifier is an ensemble learning method which can be used with machine learning to create models for classification. It consists of a large number of decision trees that can be combined

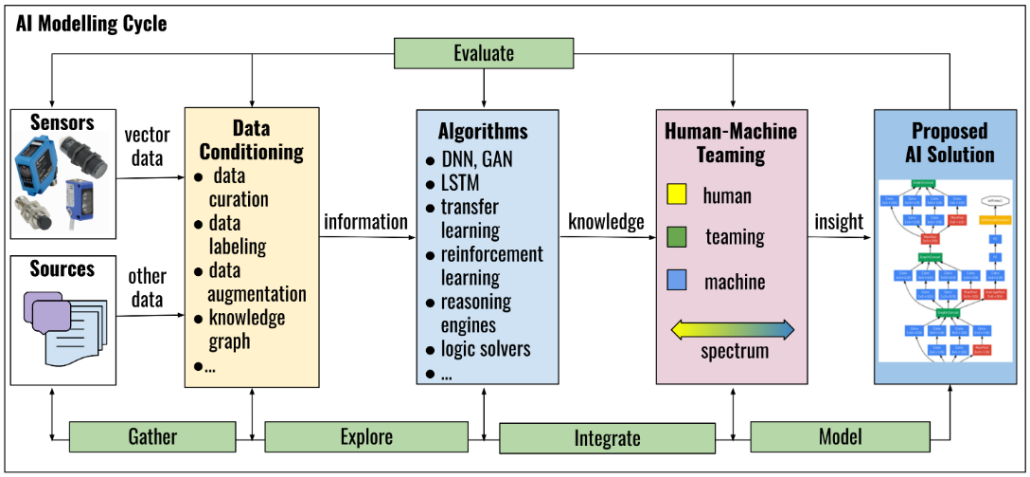


Figure 2: Steps of Developing AI Models. Adopted from [13]

to outperform any individual decision tree. The biggest difference between the random forest and the decision tree is that the decision tree creates a set of rules whereas the random forest builds new decision trees by randomly selecting features and observations. The results of the generated decision trees are averaged.

2.3 Interpretability in Machine Learning: Post Hoc Explanation Techniques

There are different methods for visualising the output of AI systems into interpretable explanations. A selection of these saliency methods used for highlighting importance of input components will be used in this study and are listed below.

2.3.1 LIME

Local Interpretable Model-Agnostic Explanations (LIME) are described by Ribeiro et al. as a technique for explaining the predictions of any machine learning classifier [14]. Local surrogate models are models that can be used to explain individual predictions of black box machine learning models due to the interpretability of the surrogate model. The surrogate model is trained to make predictions that resemble the original black box model. The big difference is that LIME does not train a global surrogate model but instead trains local surrogate models for explaining individual predictions. By providing the machine learning model with variations of the data, LIME tests what happens to the outcome of the individual predictions. LIME then creates a new dataset that holds the perturbed samples and prediction outcomes based on the original black box model. LIME uses this new dataset to train an interpretable model that it weighted by measuring the distance of the sampled instances to the instance of interest. The interpretable model should be an accurate estimation of the local predictions of the original machine learning model. The *local fidelity* describes the measurement of accuracy in local predictions. The amount of fit/loyalty of the explainable model to the original model is designated by the *general fidelity* of the model.
General fidelity means how much the explainable model or the explainer fits / is loyal to the original model.
Local surrogate models can be expressed mathematically with the following formula:

$$explanation(x) = argmin_g \varepsilon_G L(f, g, \pi x) + \Omega(g) \quad (1)$$

Instance x has the explanation model g where model g is a linear regression model. Model g is used to minimize loss L. Loss L stands for the similarity of the explanation to the original black box model prediction f. The aim is to keep the model complexity Ω at a minimum (e.g. for better explainability of the model). All possible explanations are represented as G in the equation. The field around instance x that is considered part of the explanation is referred to as pix which is also called the *proximity measure*. During application LIME only minimizes the loss of the prediction similarity which means that the user has to decide on the complexity of the explanation. This could for example be done by adapting the linear regression model outcome through changing the number of features it can use.
There are five steps in creating a local surrogate model using LIME:

- Pick a specific instance for which you want to create an explanation of the original prediction outcome.
- Make perturbations to your dataset and use the original black box model to create new predictions for the new input variable values.
- Applying weights to the newly created samples based on their similarity to the specific instance that was picked in step 1.
- Use the data with the variations from the new dataset to train an interpretable model that includes the new weights.
- Interpret the new local model to explain the prediction.

Since the user can effectively change the complexity of the local surrogate model by changing the number of features it is important to look at the feature importance of the model. A lower number of features makes the model more interpretable whereas a higher number of features increases the fidelity of the model. A good opportunity for reducing the number of features is by using Lasso. Lasso, which stands for *Least Absolute Shrinkage and Selection Operator*, uses shrinkage to create simpler and more sparse models by decreasing the number of model parameters through L1 regularization. The less important feature coefficients of the model are shrunken to zero which works well for feature selection. By removing unimportant features the model becomes less complex while also minimizing the accuracy loss of the predictions.

2.3.2 Shapley Values

The Shapley value formulated by Shapley in his study on cooperative game theory describes a method for calculating the individual player contribution (e.g. feature importance) on the total payout in order to assign payouts to the different players [15]. The payout that each player receives is based on their individual contribution to the end result. Players also cooperate and receive profit from the cooperation. To translate the Shapley cooperative game theory to the field of machine learning: the *game* describes the prediction task for an individual instance within the dataset, the *payout* is the prediction value of the instance minus the average prediction for all instances, and the *players* represent the different feature values that collaborate to achieve a specific prediction. The Shapley value of a feature value is determined by calculating the the average marginal contribution of the feature value over all possible coalitions. By increasing the number of features the computation time for calculating the possible coalitions increases exponentially.
The Shapley value of a feature value to determine its contribution to the payout can be described as the function *val* and players in S:

$$\phi j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\}\{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S)) \tag{2}$$

The S represents a subset of the features used in the model, x describes the feature values for a specific instance, and p indicates the number of features. $val_x(S)$ located in the function below describes, for all feature values in subset S, the predictions which are marginalized over features not included in subset S:

$$val_x(S) = \int \hat{f}(x_1,\ldots,x_p) d\mathbb{P}_{x \notin S} - E_x(\hat{f}(X)) \tag{3}$$

Multiple combinations for each feature that is not in subset S are calulated.
To fulfill the task of calculating a fair payout, the Shapley value ensures the properties Efficiency, Symmetry, Dummy and Additivity:

1. **Efficiency** describes that the difference of the prediction for x and the average must be the total contribution of the features.

2. **Symmetry** states that if two features have an equal contribution to all possible coalitions, their contribution (Shapley value) should be the same.

3. **Dummy** says that if a feature does not change the predicted value it should have a Shapley value of 0.

4. **Additivity** guarantees that you can calculate the Shapley value for each feature value decision and have the ability to average the feature value predictions to create a decision model.

In order to calculate the exact Shapley value of a feature x, all possible combinations of feature values with- and without feature x have to be calculated. Since the number of combinations increases exponentially with the addition of more features it can become difficult to find the exact solution to the problem. The study by Strumbelj et al. proposed Monte-Carlo sampling as a method of approximation [16]:

$$\hat{\phi_j} = \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(x_m^{+j}) - \hat{f}(x_m^{-j})) \tag{4}$$

The $\hat{\phi}_j$ part of the equation is the prediction for x with a random amount of feature values from a random point z in the data, excluding the value of feature j which is not replaced. The vector of x_m^{-j} is almost similar to x_m^{+j} however the value x_m^j comes from the randomly sampled z. Therefore, the new instances of M are assembled from two instances making them some sort of unique values.
Advantages of the Shapley value include that it allows for *contrastive explanations* for explaining predictions by comparing it to a subset or single data point instead of the average prediction for the complete dataset. Another advantage is that the *efficiency* property of the Shapley values lead to *fairly distributed* predictions that deliver a full explanation. The solid theory and fair distribution of the effects is important in giving the explanation a solid foundation. This could lead to more successful implementations of the technique in practice by increasing the trust and sense of fairness through complete explanation of all the effects.

2.3.3 SHAP

The study by Lundberg proposes SHapley Additive exPlanations (SHAP) that break down a prediction by measuring the impact of single variable scores on the final prediction [5]. SHAP is a game theoretic approach for explaining machine learning model output.
For the purpose of this study, SHAP was picked for explaining machine learning model output. This was motivated due to the extensive use of SHAP within the IDLab and the developed information products. The IDLab determined that SHAP was a suitable technique for providing explanations of machine learning model predictions. A comparison between the post hoc explanation techniques LIME and SHAP was not performed since this was not the purpose of the study. Comparing both explanation techniques in terms of usability could however be considered as future work. The paper by Slack et al. found that LIME and SHAP post hoc explanations can be fooled using input perturbations, making them not reliable [17]. This can make biased predictions unnoticeable for end users by creating explanations that do not reflect underlying bias. The study concludes that LIME is more vulnerable for this effect than SHAP. It is therefore suggested that existing post hoc explanation techniques are not suitable for detecting discriminatory behaviour of classifiers in sensitive applications.
By calculating the feature contribution for each feature to the final prediction SHAP tries to explain the prediction. The main idea of the shapley value based explanations is to use fair allocation results from cooperative game theory for allocating credit for a machine learning model's output between all its input features. To facilitate the implementation of game theory with machine learning models the model's input features should be matched with players in a game, and the model functionality should be matched with the rules of the game. SHAP implements the Shapley value explanation as an additive feature attribution method, therefore connecting LIME and Shapley values. The SHAP explanation is specified as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

In the formula, g is the explanation model, the coalition vector is represented as $z' \in \{0,1\}^M$, the maximum coalition size (number of iterations) is M, and the Shapley values feature attribution for feature j is $\phi_j \in \mathbb{R}$. For the instance x, the coalition vector x' indicates which features are taken into the equation when calculating the feature contribution:

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{6}$$

If a feature has the value '1' it is contributing to the prediction whereas features in the linear model of coalitions with the value or label '0' are not contributing to the prediction.
Besides satisfying the previously described properties Efficiency, Symmetry, Dummy and Additivity of Shapley values, SHAP also includes three more properties:

1.**Local accuracy**

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j \tag{7}$$

The Shapley efficiency property can be formulated as $\phi_0 = E_x(\hat{f}(x))$ with all the x'_j set to 1 indicating that all features are present. The difference is that SHAP uses a different name and implements a coalition vector for selecting specific features:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_X(\hat{f}(X)) + \sum_{j=1}^{M} \phi_j \tag{8}$$

2.**Missingness**

$$x'_j = 0 \Rightarrow \phi_j = 0 \tag{9}$$

The property of *missingness* states that when a feature is missing, it should have the attribution of zero. In the representation of present features using the coalition x'_j, the features that need to be explained have the value '1' and all absent features the value '0'. The Shapley values do not have the missingness property. For SHAP, the missingness property makes sure that features that do not contribute to the prediction get a value of 0.

3.**Consistency**

The *consistency* property describes that the Shapley value adapts to the marginal contribution of the feature value. For example, when the model changes while also increasing or not changing the feature value, the Shapley value of that feature would also increase or stay the same. This could be formulated as:

$$f'_x(z') - f'_x(z'_{\setminus j}) \geq f_x(z') - f_x(z'_{\setminus j}) \tag{10}$$

Where f and f' are two models and $f_x(z') = f(h_x(z'))$ and $z'_{\setminus j}$ show that $z'_j = 0$.

KernelSHAP estimates the contributions of each feature value for an instance x to the prediction. It uses 5 steps for calculating the Shapley values:

1.*Sample the coalitions* $z'_k \in 0, 1^M$, $k \in 1, ..., K$

2.*Get the prediction* for each feature z'_k

3.*Calculate the weight* for each feature using the SHAP kernel

4.*Fit the weighted linear model*

5.*Return the Shapley values* ϕ_k

By isolating the effect of a feature we can learn the most about the features effect on the prediction. If however many coalitions consist of half the features it is difficult to learn about an individual feature contribution. For that reason the SHAP kernel was proposed by Lundberg et al.:

$$\pi_x(Z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)} \tag{11}$$

Where the maximum coalition size is indicated as M and the amount of present features in instance z' as —z'—. Applying linear regression with the SHAP kernel weight produces Shapley values. The weighted linear regression model can be formulated as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{12}$$

The linear model g can be trained with the optimization of the loss function L:

$$L(f,g,\pi_x) = \sum_{z' \in Z} [f(h_x(z'))]^2 \pi_x(z') \tag{13}$$

Z represents the training data and the estimated coefficients π_j are the Shapley values of the model.

TreeSHAP is another variant for SHAP proposed by Lundberg which is suitable for tree-based machine learning models (e.g. decision trees, random forests, gradient boosted trees) [18]. TreeSHAP uses the conditional expectation $E_{X_S|X_C}(f(x)|x_S)$ instead of the marginal expectation used by KernelSHAP. Features that have no influence on the final predictions can however get a TreeSHAP value different from zero. This estimate can happen if the feature is correlated with another feature that has an influence on the prediction. The tree-specific algorithm of TreeSHAP reduces the computational complexity from $O(TL2^M)$ to $O(TLD^2)$. T represents the number of trees, L the maximum number of leaves of each tree, and D the maximum depth of the trees. TreeSHAP calculates the Shapley values in polynomial time instead of exponential. By decreasing the computational time TreeSHAP provides fast implementation for tree-based models. Finally, due to the *additivity* property, the Shapley values of a tree ensemble can be calculated by taking the average Shapley values of the individual trees. This also contributed to the popularity of the implementation of SHAP in tree-based models.

SHAP industry-standard visualisations include 3 approaches of visualising the model output in an explainable way: the force plot in Figure 3, the decision plot in Figure 4, and the waterfall plot in Figure 5. Each visualisation method tries to describes the feature contributions for a single, local prediction.

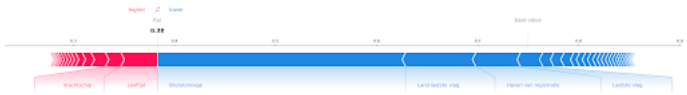


Figure 3: *This figure shows an example of the SHAP force plot.*

2.3.4 NLG

Natural Language Generation (NLG) is a subfield of AI and computational linguistics that describes how computer systems can be used to produce understandable text or speech from non-linguistic information [19]. Another study by Reiter discusses the challenges from the NLG perspective [20]. More specifically in the context of using NLG in XAI systems to explain AI reasoning to users. The main idea of NLG is that the generated texts serve a *communicative* purpose while helping users with decision making, changing their behaviour and entertaining them. Current real-world explanations of AI systems help with: developers debugging their AI system, helping users detect mistakes in AI reasoning *(scrutability)* and building trust in AI recommendations. More goals including Transparency, Effectiveness, Persuasiveness, Efficiency, and Satisfaction have been proposed by Tintarev and Masthoff for measuring effectiveness [21]. A study by Liu et al. showed that it can be used to help explain the decision made by a classification model [22]. They proposed the Generative Explanation Framework (GEF) to improve the quality of explanations. To test the effectiveness of the proposed

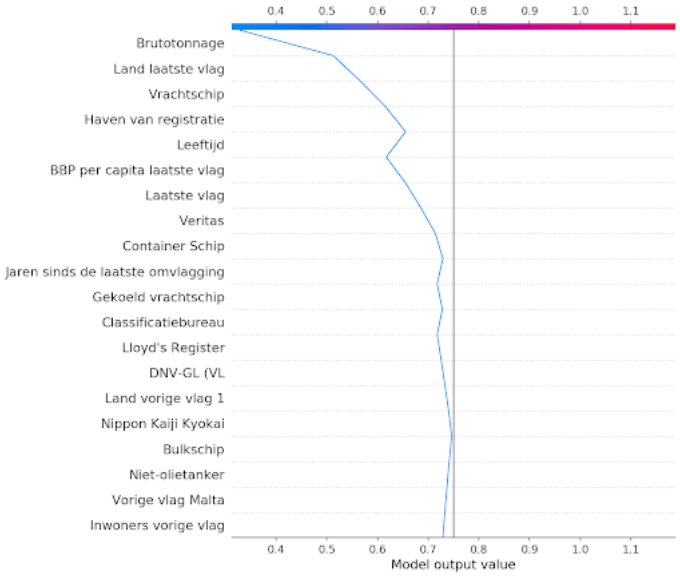


Figure 4: *This figure shows an example of the SHAP decision plot.*

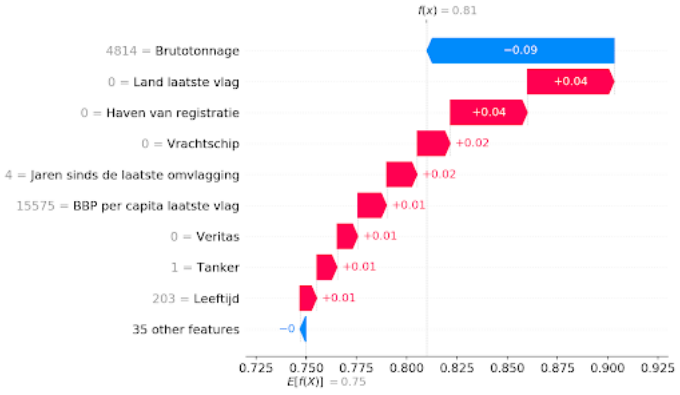


Figure 5: *This figure shows an example of the SHAP waterfall plot.*

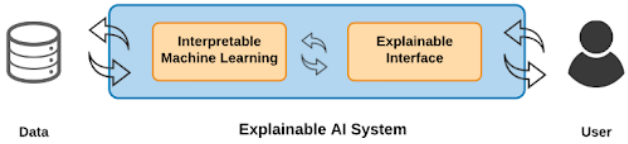


Figure 6: User interaction with explainable interface. The interpretable model interacts with the data to generate an explanation or a new prediction based for the user query. Adopted from [23].

explanation framework, experiments were conducted on two datasets where texts and numerical ratings were used. The findings include that GEF can enhance the performance of the base model while also improving the quality of the explanations generated by the model.

2.4 Research Methods

2.4.1 Qualitative Measurements and Methods

A widely used qualitative research method is a *Contextual Inquiry*. This is a field-gathering technique that collects data from a few specifically selected people in order to get a better understanding of their work practice [24]. Contextual Inquiry comprises detailed observations with interview questions that are focused on the participant's work and their interaction with technology and systems within that specific work environment [25]. The *Think-aloud* protocol is research method for collecting insights from participants by having them speak out their thoughts and feelings of the system whilst interacting with it. Participants are requested to articulate their thoughts loudly to facilitate the evaluation of the system's interface. The researcher fades into the background during the session and is only available for questions by the participant if necessary. The think-aloud method is mostly used in laboratory usability studies [25].
There are three variants to the think-aloud protocol. Firstly, the *co-discovery* method where two participants are being recorded while they are discovering and discussing about a system's interface together [26]. One advantage of the co-discovery method compared to the traditional think-aloud method is that discussing something feels more natural to the participant than articulating thoughts by yourself. The second variant to the think-aloud method is *Retrospective Testing*. During retrospective testing the participant is not verbalizing the thoughts during the interaction with the system. Instead participants are discovering the system silently while being recorded, and make comments on their thoughts, feelings and actions of the system afterwards [27]. The final variant is the *coaching protocol*. Here the researcher has a more active role during the evaluating of the system together with the participant. The researcher is allowed to help and guide participants when they are struggling with a task and point to certain areas of interest to discover [28].

2.4.2 Quantitative Research Methods

A study by Paas et al. describes the measurement of cognitive load which can be used to measure the effect of performing a specific task on the cognitive function of the participant [29]. The results from measuring cognitive load can describe whether a task is too simple or too complex for a participant. This task complexity can for example be caused by presenting a large number of items which decreases the task performance. The decline in task performance can be assigned to the cognitive overload that the participant might be experiencing [30].

Another quantitative measurement is measuring the task execution (completion) times. This describes the total time that the participant needed to finish the task. Comparing the results from different tasks can provide an estimation of the complexity of the task or usefullness of the provided system that is needed for completing the task. A specific system could for example be evaluated based on the task completion times by the participants. Task completion times between different systems can be compared to create insights regarding the usability of the systems. A study by Koch et al. described the effect of information integration on situation awareness and task completion time [31]. They found that the integration of information in an ICU display increased situation awareness but decreased task completion time. The authors stated that information integration can decrease errors, increase productivity, and increase reaction rate.
In the study of Iverson et al. statistical benchmarking is proposed as a method for estimating task execution times [32]. It is used for correcting inconsistensies between different estimates. The method of statistical benchmarking is useful for smaller experiments where the target variables have different frequencies. Information is used to adjust the sampling weights used in the experiment.
Since recommender systems and XAI systems use a relatively similar approach in producing explanations it makes sense to include the research done in evaluating recommender systems. The study by Tintarev and Masthoff describes the evaluation of the effectiveness of explanations by recommender systems [21]. To measure the effectiveness of the explanations a number of metrics have been proposed:

- *Acceptance of items known to the user:* are the features shown accepted by the user?
- *Use of the explanations:* are the explanations correctly used according to the user?
- *Perceived effectiveness before consumption:* how effective is the explanation according to the user before knowing the result/prediction score?
- *Perceived effectiveness after consumption:* how effective is the explanation according to the user after knowing the result/prediction score?
- *Similarity between liking items before and after consumption:* does the user change in opinion about the system after consumption?
- *Success rate in finding the best item:* how successful is the system in presenting an acceptable explanation to the user? The last proposed metric is difficult to test since the system in this particularly study has a decision-making support task. This means that the system does not need to satisfy the user by providing an acceptable prediction. The prediction stands on its own and acts as a supporting mechanism for the user to make its own decision. It is therefore not necessary to study whether the predictions provided by the predictive system are the expected perdictions by the user. It is however important that the system presents an acceptable explanation to the user.

The study by Schmidt and Biessmann looked at ways to measure trust and quality of explanations of machine learning predictions [33]. To measure the quality of the interpreting method they proposed a quantitative measure.
Discrete-choice experiments (DCEs) were first described in the paper of Louviere et al. [34]. The DCE method uses the generation and analysis of choice data by constructing a hypothetical market and applying a survey. DCEs are made up from several choice sets where each set contain mutually exclusive hypothetical alternatives. Respondents are then asked to choose an alternative based on their preference.

2.4.3 Measuring Trust

The study by Yang et al. looked at the effects of machine learning explanations on end users' trust [35]. During the research they measured the trust of the participant, calculated the effects of different visual representations and spatial layouts, and looked at the change of trust over time. They found that visual representation and performance feedback have a high effect on users' trust. The spatial layout showed an average effect on the users' trust. Furthermore, the study provides guidelines in the design and appropriate use of automated systems. Trust was measured as the willingness to follow the recommendation made by the explanation and participants self-confidence in the decision. They used the following questions to quantitatively measure trust:

1. "Will you follow this recommendation?" (scale: Follow, not follow)
2. "How do you feel about your decision above?" (7-point Likert scale)
3. "Was the explanation helpful in making the decision above?" (7-point Likert scale)
4. A linear "Trust Meter" ranged from completely distrust (-100) to completely trust (+100)

Another study by Nourani et al. investigated the effects of meaningful and meaningless explanations on user trust. They performed a controlled experiment with local explanations and found that explanations that are deemed human-meaningful can significantly affect the perception of the systems' accuracy [36].

2.5 Human-AI Interaction

The field of Human-AI Interaction (HAI) describes the investigation of the interaction of AI systems and humans and its effect on the human experience. The goal of HAI is harnessing the power of AI so that it is beneficial and useful to people. With the increasing work in the field of Human-AI interaction several AI design guidelines on usability have been created [37]. The guidelines describe how to effectively design AI systems in terms of utility to the user. They are also suitable for conducting a heuristic evaluation of the system.
An example of HAI are recommender systems which use explanations to provide the user with help in order to make a decision or take an action. Explanations support the user with performing a specific task. In algorithmic decision making the objective is to provide a transparent mechanism since there are no explicit choices presented to the user [38]. The effectiveness of AI systems should not only be evaluated in terms of the accuracy of the predictions. Other dimensions that relate to the acceptance of the AI system and its predictions should be measured. A user-centric evaluation approach can be used to measure the success of an AI system from the users' point of view [39].

2.5.1 System Usability Scale

The system usability scale (SUS) has been developed by Brooke as a survey scale that allows the usability researcher to measure the usability of a product or application [40]. The original SUS uses 10 statements which are scored on a 5-point Likert scale. The final SUS score ranges between 0 and 100, where higher scores indicate a greater level of usability. The study by Bangor et al. performed an empirical evaluation of the SUS [41]. They found that during their analysis approximately 90 percent of the collected SUS data used used a modified form of the original SUS. The original SUS statements and the modified statements can be found in table 8 in the appendix. The study by James

R. Lewis performed a review of the SUS and found that it has become the most widely used measure of assessing usability [42].
The study by Holzinger et al. proposes the System Causability Scale (SCS) for measuring the quality of explanations of an explainable AI system as an alternative for the system usability scale [43]. The SCS uses the Likert scale method which is a widely used psychometric scale for measuring human responses [44]. The SCS can be used to quickly evaluate whether an explanation is suitable for the intended purpose. The questions of the SCS for evaluating an explainable user interface in a fast manner include:

1. *I found that the data included all relevant known causal factors with sufficient precision and granularity.*
2. *I understood the explanations within the context of my work.*
3. *I could change the level of detail on demand.*
4. *I did not need support to understand the explanations.*
5. *I found the explanations helped me to understand causality.*
6. *I was able to use the explanations with my knowledge base.*
7. *I did not find inconsistencies between explanations.*
8. *I think that most people would learn to understand the explanations very quickly.*
9. *I did not need more references in the explanations: e.g., medical guidelines, regulations.*
10. *I received the explanations in a timely and efficient manner.*

The SCS evaluation tool has been applied to the Framingham Risk Tool (FRT) which is an example of a risk prediction model [45]. The results of this test can be found in the table below 1.

Table 1: Results of using SCS with Farmingham Model. Likert scale rating range from 1 = strongly disagree, to 5 = strongly agree

Question	**Rating**
01. Factors in data	3
02. Understood	5
03. Change detail level	5
04. Need teacher/support	5
05. Understanding causality	5
06. Use with knowledge	3
07. No inconsistencies	5
08. Learn to understand	3
09. Needs references	4
10. Efficient	5
SCS = sum of Ratings / 50	0.86

It is however important to look at the legitimacy of the Likert scale ordinal level of measurement. The authors indicate that picking the correct statistical technique is necessary to come to the right conclusions because the descriptive and inferential statistics differ for ordinal and interval variables.

2.5.2 Information Visualisation

Visualisation techniques have the ability to effectively pass on information in a way that is naturally comprehensible without explanation (source). By visualising information it can become easier and faster for users to understand the information. This is particularly useful in situations where users have to examine several instances in a relatively short time. The use of visualisation techniques provides the opportunity to present more information compared to traditional text representations. This creates possibilities in explaining the decision-making process of systems that use large amounts of information through the method of information visualisation.
Visual search: Several variables influence visual searches:

1. *The isolated-feature/combined-feature effect* shows that if the target differs from the irrelevant items in the display with respect to a simple feature such as color, observers can detect the target faster. Serial processing is needed when searching for a target that is a combination of two features. You have to pay attention to one item at the same time, which is more complex than when the target is isolated. (source)
2. *The feature-present/feature-absent affect* describes that our cognitive processes handle positive information better than negative information. Our visual search is faster when we are looking for a particular feature that is present. The visual search time increased dramatically when we are searching for a feature that is absent. (source)

Research on distributed attention: if you processed isolated features, then you should be able to rapidly locate a target among its neighboring, irrelevant items. That target should 'pop out' the screen. (source) *Research on focused attention:* if your target was an object (= a conjunction of features), you were forced to focus your attention on item at a time, using serial processing. This task is more complex. People need also more time to find the target when there are a large number of distracters. (source)
We can hold only a limited number of items in short-term memory. The study by Atkinson and Shiffrin states that the we can hold an average of seven items between five and nine items, with an average of 7 items [46]. (source) A *Chunk* is a memory unit that consists of several components that are strongly associated with one another. The study by Miller suggests that our short-term memory holds approximately seven chunks.

2.6 Key Challenges of AI System Lifecycle

After a machine learning problem has been solved in terms of using an appropriate amount of trainable data and a suitable learning model, the requirements outside the machine learning performance phase have to be addressed to create a system that is in line with the target operational environment [13]. The three important aspects that arise during this phase of the development process are: 1) Deployment challenge and computational resource constraints, 2) Data and software quality, and 3) Model validation and system verification including testing, debugging and documentation. The last point on system verification is important in the development of XAI systems especially in for example highly regulated target domains like the government. The success of an AI system for a specific domain depends, apart from the systems performance, highly on taking into account the following human-computer interaction (HCI) challenges:

- *Interpretability challenge:* Explainability and interpretability are crucial for delivering useful AI systems to end-users. By capitalizing on the effects of interpretability, AI applications can

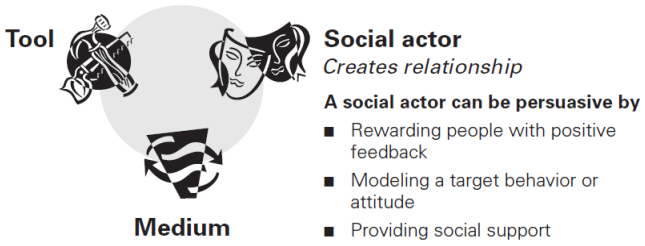


Figure 7: *Social actor according to the study by B.J. Fogg on persuasive technologies [48].*

become easier to use for end users. This is especially important when designing AI applications that require a high level of understanding of the decision making process of the algorithm. This is relevant since this research looks at the interpretability to study the effect of adding visualisations and text explanations of a machine learning based predictive model.

- *Trust challenge:* The increasing interaction between humans and AI through the fast growth of AI systems will make human-AI interaction the most popular form of HCI [37]. Research on human centered AI has become more relevant with the current attention to ethical topics on preserving human autonomy and preventing harm from AI systems. Recently released ethical guidelines for trustworthy AI reveal the general concerns with the rapid transition into AI [47].

2.7 Persuasive Technology and Behaviour

The study by B.J. Fogg on persuasive technology describes how computing applications can play the role of a social actor that tries to persuade human users [48]. Figure 7 shows computers as social actors. It is important to consider the research on persuasion of computer systems while researching or designing applications for end-users because persuasion can help with the successful commissioning of the system. If a system can persuade users that they are a team, users might be more accepting towards working together with the computer system. The computer can use persuasion in order to be identified by the human user as a teammate. This could for example increase the effectiveness of a specific task since the human user now also works together with the 'computer colleague' making the task easier to accomplish.
So how can we use computers to change people's attitudes and behaviour? Another study by B.J. Fogg et al. goes deeper into the subject of persuasion of users [49]. To create successful human-computer interactions (HCIs) requires the ability to motivate and persuade people. Interaction designers develop applications that try to change people. More specifically in the way of making people feel, what they think and finally changes in their behaviour. Common challenges include: How can users be motivated to use the application? How can designers make people persist in using and learning the (online) application? A successful interactive application often depends on people's attitudes or behaviors. HCI elements like creating a feeling of confidence or trust in the actions or results of computer systems can contribute to a more accepting attitude of users towards change. This change could for example be a change in their current workflow because they have to cooperate with a computer system.
Captology describes the study and design of computer systems as persuasive technology [50]. The definition of persuasion is described as *"a noncoercive attempt to change attitudes or behaviors"*. No coercion or force is used while attempting to persuade the user.

3 Background

This section includes the previous investigation of the training data and evaluation of the model performance. The work that is described below has already been performed before starting with this research.

3.1 Data used for Shipbreaking and Beaching Model

Two random forest models were trained, tested and applied: the shipbreaking-model and the beaching model. The variables that were used as predictors in the shipbreaking-model can be found in Table 2. The predictors used in the shipbreaking-model relate to the age, ship type and technical properties of the ship. The variables that were used as predictors in the beaching-model are located in Table 3. The predictors used in the beaching-model relate to (mutations in) the flagstate, classification society and registration harbour. The age, ship type and gross tonnage were also used in the model because the exploratory data analysis showed that these variables have a clear coherence with beaching.
Feature simplification
The relatively small amount of data can lead to features that detect patterns that are not representative for the population. Reducing the amount of features for training the model can help in preventing this. Therefore, only features that relate to ship characteristics are included in the shipbreaking model. The features that describe the flag behaviour of ships were selected for the beaching model.

Table 2: Predictors Shipbreaking model

Predictor	**Definition**	**Type**
age_in_months	Age of ship in months	Numerical
GSS_Type	Type of ship	Categorical
GSS_Propulsion	Type of propulsion	Categorical
GSS_Main.engines..Model	Model of main engines	Nominal
GSS_Main.engines..Designer	Designer of main engines	Nominal
GSS_Main.engines..Builder.code	Builder code of main engines	Numerical
GSS_Gross.tonnage	Ship's overall internal volume	Numerical
GSS_Deadweight	Weight that the ship can carry	Numerical
GSS_TEU	Cargo capacity for container ships	Numerical
GSS_Insulated.capacity	Insulated cargo capacity	Numerical
GSS_Length.overall	Total length of the ship	Numerical
GSS_Length.between.perpendiculars	Length along summer load line	Numerical
GSS_Service.speed	Average speed maintained by ship	Numerical
GSS_Main.engines..Number.of.main.engines	Total amount of main engines	Numerical
GSS_Main.engines..Max..power	Maximum power of the main engines	Numerical

Table 3: Predictors Beaching model

Predictor	**Definition**	**Type**
age_in_months	Age of ship in months	Numerical
GSS_Type	Type of ship	Categorical
GSS_Gross.tonnage	Ship's overall internal volume	Numerical
GSS_Classification.society	Non-gov organisation for certifying ships	Categorical
GSS_Port.of.registry	Place where the ship is registered	Categorical
GSS_Country_Last_flag	Last registered flag state of ship	Categorical
GSS_Country_Previous_flag	Second-last registered flag state of ship	Categorical
GSS_years_since_final_flag_swap	Amount of time since final flag swap in years	Numerical
POP_Last_flag	Total population of last flag's country	Numerical
GDP_CAP_Last_flag	Total GDP of last flag's country	Numerical
POP_Previous_flag	Total population of second-last flag's country	Numerical
GDP_CAP_Previous_flag	Total GDP of second-last flag's country	Numerical

3.2 Pre-processing

The pre-processing of the data consisted of several steps:

- *Replacing categorical variables*
  Three engine variables have been replaced: "*GSSMain.engines..Model*", "*GSSMain.engines..Designer*" en "*GSSMain.engines..Builder.code*". The feature values have been replaced by the maximum of the 'age in months' feature with the *maxage'* prefix. This has been done to prevent leakage. Leakage describes the use of information during the training of the model that is not available at the time of prediction. This can cause overestimation of the model's utility which can result in a suboptimal model. By replacing the engine variables with the maximum age to prevent leakage, a decrease of variable importance of these variables has been created.

- *Selection of ships relevant for beaching*
  A number of ships were removed from the dataset based on the specific ship type. Ships in the category 'Other' were not relevant for the beaching problem and therefore removed from the dataset. This brought the size of the dataset from 9007 ships back to a total of 6078 ships. From the set of ships, 1381 were used to create an validation set. The remaining 4697 ships can be used to train and test the model.

- *Selection of ships with portcalls from Rotterdam*
  During the selection process only ships that have portcalls in Rotterdam are selected. This was done to even out the sample populations of the active ships and demolished ships. After the selection the dataset consists of 3627 ships where 2246 ships can be used for developing the model and 1381 ships for the validation set. The train/test set of 2246 ships contains a total of 2063 active ships, 45 dismantled ships and 138 beached ships. The set of ships has increased in

scarcity due to the selection in order to equalize the sample populations. This would make the model better at distinguishing between active and demolished ships.

- *Even out population of active and demolished ships*
  A threshold was used to increase the overall percentage of old ships, either demolished or active, in the dataset. The age of the youngest, most recent demolished ship is used as a threshold filter to sample the set of active ships. All active ships that were younger than the threshold, namely the youngest demolished ship, in the previous 5 years were removed from the dataset. This is an implicit rule that was put in the model as a way to make the ages of the active and demolished ships more evenly distributed.

3.3 Model

During the development of both random forest models, alternative modelling techniques (like single trees) were investigated. These simpler models however did not lead to the required performance. The use of simpler models would also not have required the implementation of SHAP to provide explainability of the model results. To provide the required model performance while still being able to explain the results, the ensemble random forest modelling technique was picked. Due to the complexity of the ensemble tree models, it was required to use SHAP to provide explainability and transparency of the model.
The Area Under the ROC curve (AUC) score is used to measure how well the classification model can distinguish two classes. It calculates the probability that two randomly selected samples are correctly ranked by classification algorithm. The AUC score for both models can be found in table 4.

Table 4: AUC scores

Predictor	**AUC - Shipbreaking**	**AUC - Beaching**
All predictors	0.995	0.918
Top 4 predictors removed	0.985	0.918
All possible leakage variables removed	0.98	0.858
Only ‘Age in Months’	0.847	0.631

Figure 8 shows a plot of the Z scores of the SHAP values showing the average SHAP scores. The correlation matrix of the predictors is shown in figure 9.

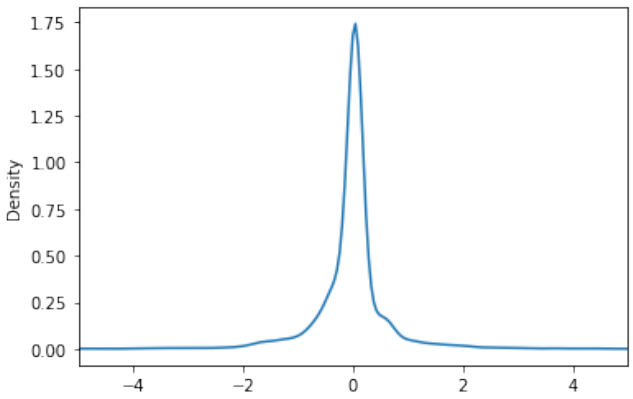


Figure 8: Plot of SHAP Z scores

Figure 9: Correlation matrix of predictors

4 Methodology

In this section the study design and methods will be explained along with the hypotheses.

4.1 Research Questions and Hypotheses

Based on the main research questions, two hypotheses were proposed:

$H1_1$: *"The use of newly proposed representations of SHAP visualisation leads to a higher level of perceived system usability, when compared to the absence of SHAP visualisations."*

$H2_1$: *"The use of feature-based natural language explanations as an addition to the SHAP visualisations significantly increases the SUS w.r.t. the model interpretability and individual explanations provided to the end users, when compared to the absence of additional feature-based natural language explanations."*

The first hypothesis stated that the use of newly proposed representations of SHAP to substantiate the model output would increase the usability of the system. The newly proposed representations of SHAP explanations should increase the users' understanding of the model concerning feature importance allowing for better model interpretability and informativeness.
The second hypothesis declared that the implementation of feature-based natural language explanations should have a positive or negative effect on the model interpretability and informativeness. The second hypothesis on the effect of text explanations as an addition to the SHAP visualisation can be tested with a two-tailed T-test. The Wilcoxon signed rank test was used as a statistical analysis method to test the effect of differences in feature quantities of the text explanations.

Two sub questions were proposed for each main research question. For the first research question, the hypotheses of the corresponding sub questions are listed below:

$H1_1$ *"The newly proposed representations of SHAP-based explanations positively affect the systems' usability score."*

$H2_1$ *"The newly proposed representations of SHAP-based explanations increase perceived cognitive load."*

For the second main research question, the hypotheses for the sub questions are also provided:

$H1_1$ *"The number of features included in the text explanations have an effect on the informativeness and interpretability of the explanation."*

$H2_1$ *"The newly proposed NLG explanations convince the end users and have a positive effect on the systems' usability score and decrease perceived cognitive load."*

4.2 Research Design

The research used a mixed methods approach to collect the results for answering the proposed research questions. The mixed methods approach is a combination of quantitative and qualitative design and was chosen because the research questions depends on both approaches. The research aimed

to create insights on the potential contribution of SHAP- and NLG-based explanations on the interpretability and informativeness of a prediction model output. The qualitative part of the research includes conducting and analyzing semi-structured interviews. For the quantitative parts a cross-sectional design had been chosen. This means that several questionnaires and statistical analyses were included.
User studies were performed to gather information regarding the perceived interpretability and informativeness of the explanations by the participants.

4.3 Qualitative Study

4.3.1 Feature Selection

As described in the method section, a total of 8 features from the beaching model were selected for the visualisation. During the first interview, only the force plot was used since the other 2 industry-standard plots were not explored yet. The focus of the interview was on comparing the additional visualisation- and textual explanations. The next two interviews consisted of a mix of all of the 3 industry standard SHAP plots: force plot, decision plot, and waterfall plot. Each plot was generated using the specified 8 features and the Dutch feature names. There effectiveness was measured and compared to textual explanations in terms of usability.
The experiment consisted of several instances of ships that included different information representations. As previously described, the qualitative results from the think-aloud interviews were used to make design decisions while creating the quantitative user experiment.

4.3.2 Semi-structured Interviews

The first phase of the experiment consisted of 3 semi-structured interviews. A total of 3 experts in the field of waste inspection were invited to join the semi-structured interviews. The participant group was made up of inspectors in the field of waste that have knowledge on ships and shipbreaking. During the first part of the interviews, questions were asked on the actions performed during their daily work activities. This part focused on the routine and information need of the inspectors. The interviews followed a protocol which can be found in the appendix. During the expert interviews the interaction of the experts with the presented information was recorded. The interviews also included several questions that focused on the usability of the representations that were shown. The results of each corresponding interview were analysed to create preliminary findings on the experienced level of usability. These findings helped with the design of the SHAP representations that were necessary for the quantitative part of the research.

4.3.3 Think-aloud Protocol

A field-gathering method seemed to be optimal for this study since the users needed to be observed during their work activities to effectively gather insights concerning their interaction with the AI system. Specifically their interaction with technology should be observed during the study. Therefore, a full contextual inquiry have been conducted to gather observations of the participant interacting with the AI system.

Since the think-aloud method is mostly used in a laboratory environment and to evaluate the interface of a system, the think-aloud technique was used during the qualitative study.

The respondents were selected based on the following criteria: function, prior knowledge, willingness to participate. The participants should not have much prior knowledge of data science and machine learning and should be closely connected to the shipping sector. This was required to test whether the explanations are intuitive enough to be used without giving instructions on them beforehand.
The second part of the interviews, which contained the think-aloud protocol, used the same group of participants as previously described. During this part of the interview, the experimenter provided different SHAP representations (features, SHAP visualisations, text explanations) to the participants. The participants used the Think-aloud protocol to give feedback on the presented representations. Results were gathered on the initial interpretability and informativeness of the representations. These results have been analysed to create preliminary findings on the usability of the system that had been taken into account while designing the final quantitative experiment.

The second phase of the research consisted of the final experiment. This was a more in-depth approach to the first phase. Participants were asked to evaluate different representations of the systems' output. The different instances were randomly presented to the participants. After each instance, the participant filled out a response about that specific instance. The response consisted of a small set of instance specific questions that need to be scored to asses the SUS. A subset of the SUS questions is presented below. The total set of SUS questions can be found in Appendix 8.

- "How informative was the presentation for understanding model behaviour?"
- "How accurate do you think this model prediction is?"
- "How understandable/clear is the presentation?"
- "The system/representation was clear."

4.3.4 Implications for Quantitative Study

Findings of the expert interviews led to the iterative development of the proposed SHAP-based visualisation and text explanations. The responses of the participants led to the identification of interpretation errors concerning the different types of information representations that were displayed during the interviews. Together with the participants, the visualisations and text explanations have been discussed. During the interviews the researcher recorded the experiences of the participants with the system while also asking questions regarding the usability of the system.
Results from the interviews were analyzed after each session and implemented within the experimental setup. Therefore, each interview was unique. All interviews focused on the experienced level of usability. Findings from the interviews include:

- Rename features and feature values to Dutch
- Provide accurate information for each feature and rename *'Other'* with actual feature value
- Sum the SHAP score of features that describe the same categorical feature (e.g. type of the ship)
- Positive responses on the use of colour within the visualisation therefore keep this unchanged within the SHAP visualisations
- Force plot was interpreted from left to right instead of interpreting it from the center to the outer edges

Handcrafted feature selection
During the interviews, the importance of specific features to the explanation were discussed. Several categorical features that indicated whether a condition was fulfilled or not were identified and required additional thinking steps from the participant. This included the ship type and specific flag features. These categorical features were analysed regarding their importance to the total explanation. This led to the selection of 8 features that contributed to the SHAP score.
Features that had a feature value 'other' indicated that they belonged to a residual category within the beaching model. This indicated that there were too few instances with that specific feature value. Therefore, they were given an all-encompassing label within the model. The information on the feature values was however perceived as important for the user and should therefore be specified, even though the model deems them inaccurate or not representative for the prediction.

4.4 Quantitative Experiment

4.4.1 Feature Selection

For the experiments, several subsets of the data were created based on the risk scores. This selection included instances with similar risk scores. These were used during the experiments for increased consistency of the results. The feature names were translated to Dutch. Categorical feature values had to be renamed to match the specific category. Feature values were added to the data to provide more complete visualisations.
Categorical variables were transformed to binary variables during the training of the model using one-hot encoding. A ship can for example have the type *"Tanker = 1"* and at the same time it is not a *"Cargoship = 0"*. To make the model output more practical and simpler to interpret, the binary features were merged into a single variable *"Ship type = tanker"*. The corresponding SHAP-values of the binary features were summed and added to the single variable. The binary classification values were replaced with the actual values for the specific ship. Indirect feature contributions were summed and added to the corresponding feature category that they represent. If for example *"Vrachtschip = 0"*, then the feature is removed and its feature contribution is added to the ship type categorical feature, like *"Tanker = 1"*. Some features were removed from the model since they were uninformative and not relevant for the specific use case of the experiment. These features include: POP previous flag, GDP previous flag, POP last flag, and GDP last flag. They may be individually not so meaningful for the user experiment but as mentioned earlier, a difference or ratio of GDP per capita of the current/former flag countries would make it meaningful. Their SHAP contribution was summed and added to the corresponding previous flag and last flag features of the ship.

4.4.2 Pilot

Qualtrics software was used to create an online experiment that could be shared among participants. The experiment was designed based on the previous qualitative findings from the interviews. The final version of the online user experiment was tested within the IDLab. From the IDLab, a total of 19 colleagues participated during the pilot of the experiment. The participants for the pilot were data scientists. This created interesting feedback on the experiment and helped in detecting specific errors. The pilot consisted of 20 instances of ships. The ships were picked based on their risk scores. Each ship was a real-world ship with the correct corresponding features and feature values. Participants were asked whether they would inspect a single ship or not based on the information that they receive for that specific ship: *"Probeer aan de hand van de informatie de keuze te maken of u wel of niet in gesprek wil gaan met de eigenaar vanwege het risico op beaching."* This can be seen in figure 10.

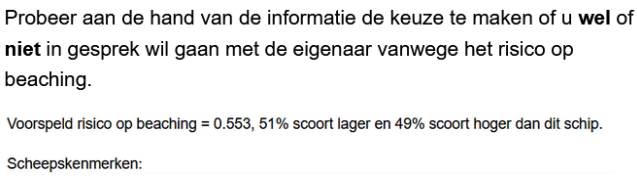


Figure 10: Example instance of pilot.

In order to create usable new representations that use SHAP feature importance, the above mentioned interviews were conducted with experts. After each consecutive interview, the results were analysed and the experiment design was adapted. These adaptations were created based on findings from the interviews and think-aloud experiment. The design of the final experiment went through several rounds of iterations. Results from the interviews, think-aloud procedure, and the pilot within the IDLab provided new insights for developing the final SHAP visual- and text explanations.

4.4.3 Participants

A total of 30 participants were invited to test the usability of the proposed treatments. The purposive sampling method was used to select participants suitable for participating in the research [51]. The participants were not closely related to the field of data science but had sufficient knowledge about ships. Participants were recruited from a group performing tasks as port state control (PSC) personnel. Each participant tested 20 instances in an online questionnaire application. Each treatment was presented to the participant through 5 different instances. Measurements concerning the cognitive load and usability of the system were taken for each instance. Participants were be asked to score each instance using a Likert scale from 1 to 5, where 1 indicates *"Strongly disagree"* and 5 indicates *"Strongly agree"*. Finally, 15 complete responses were gathered from the PSC group.
Each usability issue has a probability to be found during an evaluation. A minimal sample size of 10 should uncover more than 80 percent of usability issues [52]. A larger sample size including 20 participants is recommended since the minimal amount of the total usability issues found is increased to 95 percent.

4.4.4 Materials

In order to gather results, an online experiment was build using Qualtrics, an online tool for conducting experiments. A set of explanations without SHAP visualisations, with SHAP visualisations, with NLG explanations, and with SHAP visualisations and NLG explanations were generated.
A total of four conditions were tested:

1. *No explanation (baseline)*
2. *SHAP based visualisation (8 features)*
3. *SHAP text explanation (5 features)*
4. *SHAP visualisation (8 features) + text explanation (5 features)*

Different information representations that described individual ships were created for testing the explanations. These information representations included feature values, SHAP risk score, SHAP visualisations of feature values, and NLG text explanations. The SHAP visualisations consisted of waterfall plots that used a total of 8 features that the model deemed relevant for the prediction. A total

of 5 textual feature values were shown. These features were chosen from the top 5 highest SHAP feature contributions and are therefore important features that contribute to the prediction of the model. The focus of the research was on evaluating the usability of the proposed model explanations.
To research the use of SHAP and additional NLG explanations on the interpretability and informativeness of AI systems, cognitive load, usability and user trust was measured. The questions of the SUS and SCS were adapted for the experiment. This adaptation was required since the original SUS questions focused less on interpretability. This was however necessary for evaluating a predictive system. Therefore, the original SUS questions were changed so that they focused more on aspects that are important for evaluating AI solutions (*e.g. interpretability, informativeness, trust, etc.*.). They can be found in Appendix 9. Participants answered several questions that discuss usability that they experienced during their interaction with the system. Different explanation variants were shown to the participant. After each instance, the participant was asked to fill in a response for the specific explanation. Measurements were taken in the form of Likert-scale responses. The online questionnaire tool Qualtrics was used for the experiment.

4.4.5 Procedure

The online survey tool Qualtrics was used to create and distribute the questionnaire to the participants of the study. An email had been composed that included the instructions and the link to the experiment. A password was also provided since the data consists of real ships which could be identified through feature values. By including a password, the privacy of the shipowners was ensured. The email was send by the experimenter to the team leader of the PST (Port State Controller) participant group and forwarded to each participant. The participants were invited to the study through an online invitation link. They were given two weeks to perform the experiment.
The total experiment took around 20 minutes to complete. Participants could take a break during the experiment since their progress is saved. During the time that participants could complete the experiment, several reminders were send to the participants. This was necessary since the experiment took place during the holidays.

4.4.6 Data Analysis Plan

The total SUS score including all of the questions was calculated using the following formula:

$$(\sum PV_i - (1 * PT)) + ((5 * NT) - \sum NV_i) * 2.5) \tag{14}$$

Where PV represents the values of all positively-phrased questions, PT is the total number of positively-phrased questions, NV are the values of the negatively-phrased questions, and NT is the total amount of negatively-phrased questions [41].
For answering the first research question, the results of treatment 1 on the SUS were analysed with a focus on interpretability and informativeness. A statistical analysis was performed that looks at the: 1) possible significant improvement for each participant, and 2) an item-wise analysis for each explanation.
The second research question was answered by validating the use of text as an explanation method. The results of the three alternative treatments to the baseline, by using static dashboards that hold text explanations, were analyzed. The statistical analysis was performed through the method of statistical benchmarking which is described in the literature section.
To analyze the results of the Likert-scale responses several statistical tests will be performed:

- *ANOVA and ANCOVA*

  The first hypothesis describes the expected results from the experiments on the first research question regarding the effect of SHAP visualisations on increasing the interpretability and informativeness of the system output compared to feature values. An ANOVA was used on the results of all 4 conditions. The ANCOVA has also been performed in order to include the response time measurements within the analysis.

- *Two-Tailed T-Test*

  A two-tailed T-test was executed for the results considering the effect of providing additional visual SHAP-based explanations on the interpretability and informativeness of the system output compared to no additional text explanations. The t-test provided insight into the effectiveness of visualisations on increasing the usability of information representations. It was used to provide insight into the first hypothesis on the effect of visualisations on the SUS. The t-test was a suitable statistiscal analysis method since the data from the conditions was normally distributed.

- *Wilcoxon signed rank test*

  The effect of using a different types of information representation (visualisation vs. text) has been evaluated using the Wilcoxon signed rank test. Wilcoxon signed rank test has been used to analyse whether the conditions have different distributions. This is necessary when analysing variables that are represented through an ordinal scale. The Wilcoxon test was chosen for answering the second hypothesis on the effect of text explanations on explainability. By grouping the SUS questions based on their corresponding category, findings could be created.

4.5 Experimental Setup

4.5.1 Dataset and Tools

Jupyter Notebook with python version 3.7 was used to load the data and the beaching model. Python was used throughout the whole research to make changes to the data and model results that were required for the experiment. The SHAP in Python was used to create the visualisations and feature importance for individual predictions of the beaching model.

4.5.2 Dataset Beaching Model

The CRISP DM (CRoss Industry Standard Process for Data Mining) process was used during the research to organize the data science project. The first two steps of the process were already described in the previous sections. In the introduction, the business understanding was described. The data understanding part can be found in the background section.
The next steps of the CRISP DM process that are discussed in this section include the preparation of the data together with the modeling. The evaluation and deployment are the final steps of the process and described in the experimental setup and results section

4.5.3 Data Preparation

The data was split into subsets to extract instances of ships with similar SHAP scores. The averages were calculated by taking the average over the absolute SHAP values of features for the implementation set. A total of five subsets were created from the implementation set based on their average SHAP score:

1. feature values high (80 to 100)
2. feature values low (0 to 20)
3. feature values medium high (60 to 70)
4. feature values medium low (40 to 30)
5. feature values middle (50 to 60)

4.5.4 Alternative Visualizations of SHAP Scores

The shap package was used to calculate the SHAP scores for each ship. The visualisations of the SHAP scores were generated using the calculated SHAP scores. The three industry-standard SHAP visualisations were created to give instance-wise visual information of the SHAP scores. The visualisations include:

- *The force plot*, can be found in figure 11
- *The waterfall plot* can be found in figure 12
- *The decision plot* can be found in figure 13

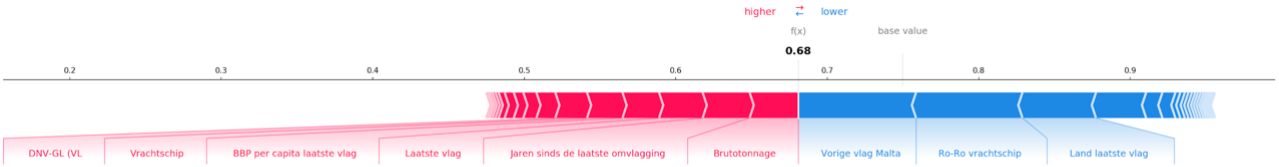


Figure 11: *SHAP force plot*

The visualisations that were created from the data of the implementation set functioned as a test design for the interviews. The goal was to test which one of the three visualisation types was the most

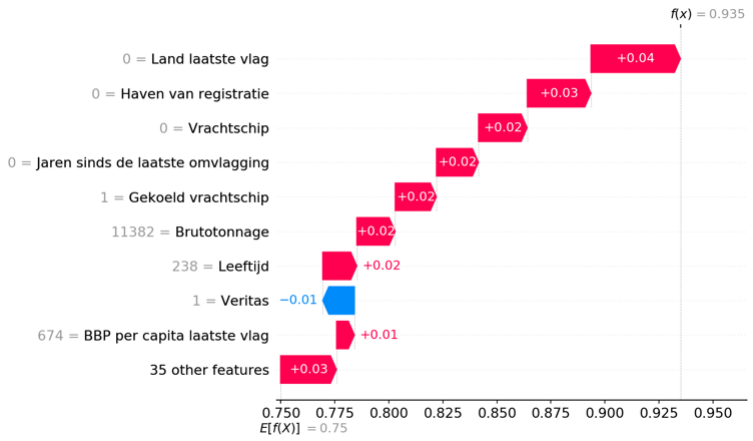


Figure 12: *SHAP waterfall plot*

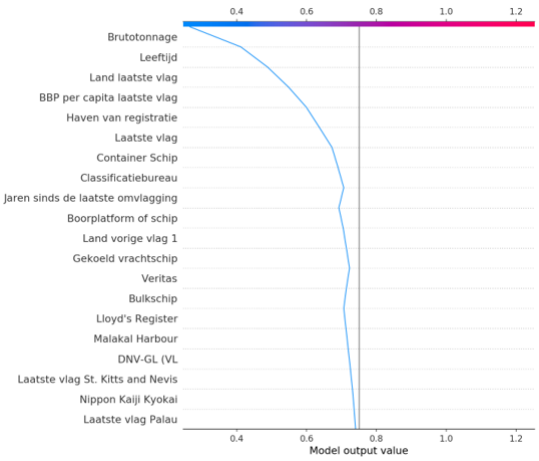


Figure 13: *SHAP decision plot*

intuitive. This would be helpful for designing the final evaluation of the model representations. This resulted in a test design that was iteratively changed according to the feedback from each consecutive participant.
Some of the changes included different parameter names and settings. The grouped feature 'Overig' was removed from the model since it was difficult to interpret by participants and added relatively few informational value to the model explanation. The model description was also changed according to the feedback.
The model explanations was qualitatively assessed through the input of the project team IDLab and the interviews with shipbreaking experts. The final assessment of the model explanations had been performed using quantitative research methods with experts in the field of ships.

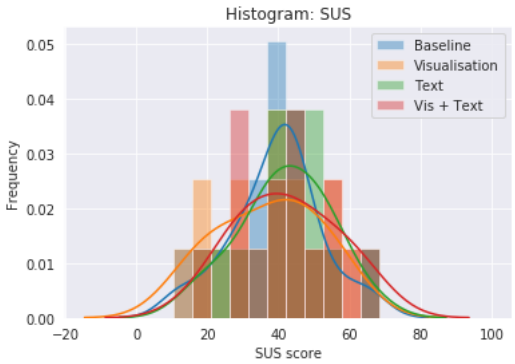


Figure 14: *This figure shows the normal distribution of the SUS scores for each condition.*

5 Results

The results section contains the experimental results for explainable model output using SHAP representations that were gathered during the user experiment.

5.1 Quantitative Results: User Experiment

In this subsection, the results of the quantitative user experiment using Qualtrics are shown.

5.1.1 Descriptive Statistics

In the tables 5 and 6, the System Usability Score (SUS) and cognitive load measurement per condition are shown.

Since the experiment used 7 positively phrased questions and 2 negatively phrased questions, the SUS was calculated as *((sum positive - 7) + (10 - sum negative)) x 2.5*. Since the maximum score of the SUS is 40, the multiplication by 2.5 is made to provide a scale from 0 to 100. Note that the SUS does not provide a percentage. It is simply describing the usability of the application through a score.

5.1.2 Data Distribution

In Figure 14, the corresponding SUS scores for each condition are displayed. The results are normally distributed among the participant responses. It can be seen that condition 4 (visualisation + text) has the highest mean and maximum score (mean = 42,333, max = 67,5).

Figure 15 provides a distribution of the cognitive load scores for each condition. It can be spotted that conditions that included visualisations had higher average cognitive load scores.

Finally, Figure 16 shows the distribution of the response time. Note that the included response times were filtered to 100 seconds or less to be included in the distribution. Responses that took longer than 100 seconds could indicate that people were distracted or doing other things instead of looking at the specific instance. The total cutoff is based on the maximum standard deviation *(SD = 48,9)* times two. This was done to remove outliers that could influence the data.

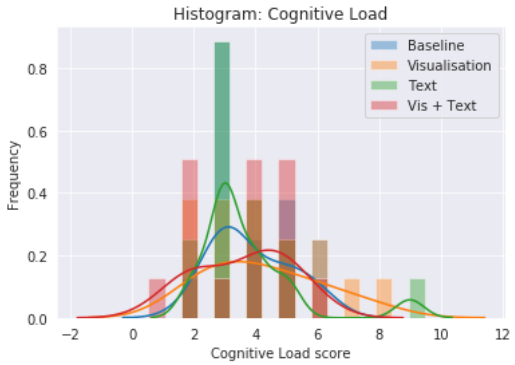


Figure 15: *This figure shows the normal distribution of the cognitive load scores for each condition.*

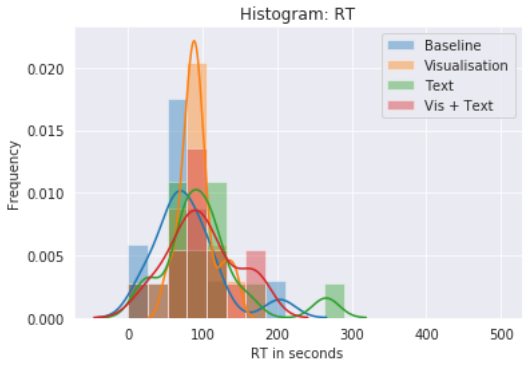


Figure 16: *This figure shows the normal distribution of the response time for each condition.*

Table 5: Descriptive Statistics System Usability Score

	SUS_Base	SUS_Vis	SUS_Text	SUS_Vis+Tekst
Valid	15	15	15	15
Mean	39.3	37.1	42.0	42.3
Median	40.0	40.0	42.5	45.0
Std. Deviation	12.8	14.6	12.7	14.1
Minimum	12.5	12.5	17.5	17.5
Maximum	65.0	60.0	65.0	67.5

Table 6: Descriptive Statistics Cognitive Load

	CogLoad_Base	CogLoad_Vis	CogLoad_Text	CogLoad_Vis+Tekst
Valid	15	15	15	15
Mean	3.8	4.2	3.7	3.6
Median	3.0	4.0	3.0	4.0
Std. Deviation	1.2	1.8	1.7	1.5
Minimum	2.0	2.0	2.0	1.0
Maximum	6.0	8.0	9.0	6.0

5.1.3 Comparative Analysis

The comparative analysis looks at the comparison between the results of the different conditions that were tested. The results on the total SUS per condition where analysed using the Analysis of Variance (ANOVA). A one-way ANOVA was performed for each of the metrics. The metrics were the dependent variables and included the System Usability Score (SUS), cognitive load, and response time. The independent variables included the 4 different conditions. The different p-values for each of the conditions were compared to the significance level of 0.05. This was done to assess whether the null-hypothesis could be rejected.

- A p-value $\leq$ the significance level α indicates that the mean difference are statistically significant
- A p-value $>$ the significance level α indicates that the mean difference are not statistically significant

The results of the ANOVA showed that there was no overall significant effect of the representation type on the SUS, *F(3, 56) = 0.479, MSE = 88.715,* $p > 0.05$, $\eta^2 = 0.025$. There was also no overall significant effect when looking at the cognitive load measurements for each representation, *F(3, 56) = 0.487, MSE 1.244,* $p > 0.05$,$\eta^2 = 0.025$. The last ANOVA on the response time provided a significant effect of the type of the representation on the measured response time, *F(3, 34) = 4.329, MSE = 5078.738,* $p < 0.05$ *(p = 0.011),* $\eta^2 = 0.276$.

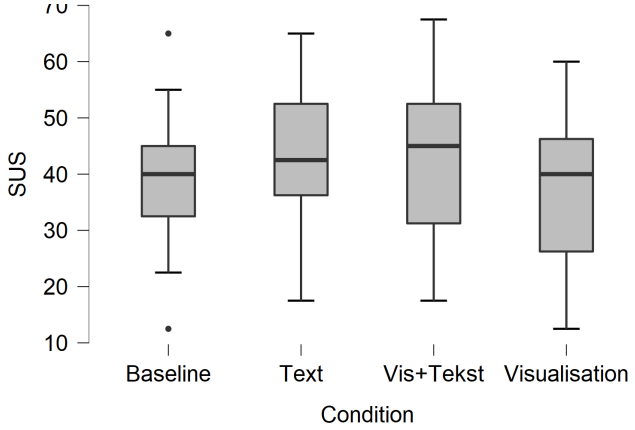


Figure 17: Boxplot of SUS results from online user experiment.

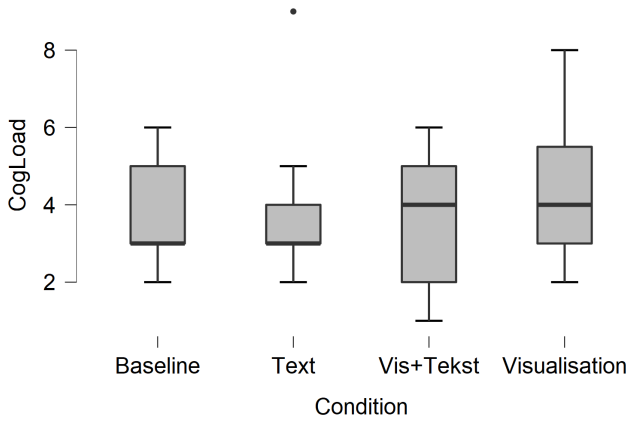


Figure 18: Boxplot of cognitive load results from online user experiment.

Since the overall effect was not significant, the posthoc comparisons of the ANOVA's were also taken into consideration. No significant effect was found for the conditions 2 Visualisation *(M = 37.167, SD = 14.604)* and 4 Visualisation + Text *(M = 42.333, SD = 14.157)* with visualisations versus the baseline *(M = 39.333, SD = 12.834)*.

The next step was to compare the group means. This can be visualised through a boxplot. In Figures 17, and 18 you can see the different median values for each of the 4 conditions. This provides a clear view of the results from the participants. The two outliers within the baseline condition can be spotted.
By using the ANCOVA we can compare the effect of different metrics on the dependent variable, in our case the SUS. The results on perceived cognitive load were analysed using the ANCOVA and can be found in table 7. This showed a significant effect for SUS versus Cognitive Load, $F(1, 32) = 5.627$, $p = 0.024$, $\eta^2 = 0.141$.

In the descriptive results it can be found that cognitive load has a significant impact on the SUS *(p = .024)*. The text-based explanations as well as the baseline have a lower cognitive load compared to the explanations that contain visualisations which had a high perceived cognitive load.
The Wilcoxon signed rank test was used to compare the subset of SUS responses on understandability. Here, condition 2 visualisation is compared to condition 4 on visualisation and text regarding the SUS

Table 7: ANCOVA - SUS

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Condition	401.591	3	133.864	0.709	0.554	0.053
CogLoad	1062.895	1	1062.895	5.627	0.024	0.141
RT	4.076	1	4.076	0.022	0.884	5.425e-4
Residuals	6044.301	32	188.884			

scores on understandability. This lead to a significant effect (p = 0.02).

6 Discussion and Conclusions

The primary objective of this research was to perform an assessment of the effectiveness concerning the intuitiveness, interpretability, and explainability of SHAP-based explainable AI and to allow other researchers and practitioners in the field who are using explainable AI to get a better understanding of the possibilities through a range of different information representation techniques. More specifically, this research focused on how the inclusion of instance wise explanation using SHAP has an effect on the usability of AI models in information driven solutions designed for end users. Even more specifically, on how the beaching model from ILT can be improved w.r.t. the SUS so that inspectors can make better use of it in practice. The central part of the research was to convince end users in terms of interpretability and informativeness by comparing visual- and text explanations as an addition to the traditional approach. To further understand the usability of SHAP-based explainable AI, it is necessary to collect information concerning the success rates of implementing the technique into relevant use cases. Through qualitative and quantitative performance measures, a better understanding of the effectiveness of SHAP-based explainable AI representations can be gained.

6.1 Findings

This research tried to provide insight into the effect of visual- and textual information representations on the usability of machine learning model output. The results from the conducted expert interviews and online user experiment have been analysed to provide findings related to answering the proposed hypotheses.

The research questions and their corresponding sub questions are stated below:

RQ1. *"How does the inclusion of SHAP-based instance wise explanation in a random forest ensemble model affect the users' view in terms of system usability scale?"*

$H1_1$ *"The use of newly proposed representations of SHAP visualisation leads to a higher level of perceived system usability, when compared to the absence of SHAP visualisations."*

RQ2. *"Using natural language generation (NLG) as an addition to SHAP representations, how does the SUS improve w.r.t. the model interpretability and individual explanations provided to the end users?"*

$H2_1$ *"The use of feature-based natural language explanations as an addition to the SHAP visualisations significantly increases the SUS w.r.t. the model interpretability and individual explanations provided to the end users, when compared to the absence of additional feature-based natural language explanations."*

The first research question focused on the effect of additional visualisations on the perceived usability of the explanation. No significant results were found regarding the effect of visualisations on the SUS score compared to the baseline. This means that there is no evidence for supporting the first main hypothesis. The experiment shows that visualisations decrease the perceived usability. This is in contrast to the first sub question of RQ1 and the proposed hypothesis on visualisations contributing to the SUS score. This can be explained by looking at the response time- and cognitive load measurements of the visualisation condition. Both measurements increase when visualisations are added. A higher response time indicates that participants need more time to interpret the visualisation. This

was expected since participants need more time when interpreting more information. The lower response time for the visualisation and text is interesting and shows that adding more information does not necessarily increase processing time. This shows that participants find it difficult to interpret the visualisation. The increased cognitive load showed that participants required more effort while interpreting the information. This provides evidence for accepting the hypothesis of the second sub question stating that the newly proposed visualisations increase perceived cognitive load.

Results supporting the second research question on the effect of text on the usability of explanations can be found when looking at the specific SUS questions regarding understandability. Results from the literature and the interviews showed that the amount of features included in the explanations had an effect on the informativeness and interpretability of the explanation. The first hypothesis of the second main research questions can therefore be accepted. The additional NLG text explanations were added to increase the interpretability and informativeness of the model explanation. Here, a significant effect can be found between adding a visualisation and alternatively adding text, when looking at the perceived understandability. These results were sufficient for rejecting the null hypothesis and accepting the alternative hypothesis of the second subquestion on the effect of NLG explanations. The effect of adding a visualisation compared to adding a visualisation and text also yielded a significant result. These results indicate that the addition of text positively influences the level of understandability regarding the explanation.

Both hypotheses contributed to understanding the effect of visualisations and text on the usability of explanations, more specific, of machine learning model output. No significant effects were found on the effect of the addition of the proposed SHAP-based visualisation of the model output compared to the baseline condition. This could have been caused by the complexity of the visualisation or the lack of explanation on interpreting the visualisation. The measured increase in response time and cognitive load confirm the increased effort that participants had during the interpretation of the visualisation. This indicates that the proposed SHAP-based visualisations might not be intuitive enough to be used without prior instructions on how to interpret them. Visualisations should therefore be accompanied by SHAP-based text explanations of the model output or instructions on how to use the visualisations. Additional SHAP-based text explanations on the other hand do significantly increase the usability of the model output concerning the understandability when comparing condition 2, where visualisations are added with condition 3, where text is added. The research showed that using text to represent the output of the model has a positive effect on the usability of the system. The online user experiments provided results on the perceived usability of the explanations. The results indicate that the use of additional text explanations increases the usability of the information system while also decreasing the cognitive load. These findings showed that inspectors have a preference for textual explanations when interpreting information representations that represent the output of a predictive model. This is relevant for the ILT since they are developing machine learning models that need to be used by inspectors. The results of the beaching model can be generalized to determine the usability of the shipbreaking model. The findings from this study could contribute to the development of future applications that use explainable AI techniques such as SHAP. By developing a set of design guidelines for implementing XAI, future applications could benefit from better usability.

Another interesting finding is the effect of additional visual explanations on the systems' usability score and cognitive load. Adding visualisations without text lead to a decrease in the SUS score compared to the baseline while increasing the cognitive load. This could indicate that for our case, text is a more intuitive method of transferring information than visualisations. Specifically, while representing information using 8 variables for visualisations and 5 for textual explanations. Finally it

was found that the explanations that used a combination of visualisation and text scored highest on the SUS but only slightly higher than the use of text explanations.

6.2 Limitations

The representative sample used in the experiment included participants that had no prior experience with graph interpretation. This could have influenced the results on the presented SHAP-based visual explanations. As before mentioned, these visualisation might have been intuitive enough for the participants to interpret them without instructions. This could explain why there was no significant result on the effect of adding visualisations to the usability of the model output.
Because the experiment required participants with a background on ships, the participants were recruited from a specific group. The group of port state controllers allowed for a limited amount of participants due to the group size. The relatively small amount of participants that participated during the study could have had an effect on the accuracy of the results.

6.3 Future research

Accordingly, future research could be performed with data that resembles the reality more closely. By examining real-life instances, participants might have different experiences concerning perceived usability, model trust, cognitive load, etc.
Another prospect for future research could be studying the effect of visualisations with the inclusion of additional explanation on interpreting the visualisation. Since the end-users are no data scientists, the experiment should also take this into consideration by offering instructions on how to interpret and use the visual explanation.
In future experiments, researchers could change the number of features of the explanations. Instead of using a fixed amount of features, future studies could experiment with visualisation- and text explanations that use more or less features to study the effect of using different amount of features on usability, interpretability, trust, and response time. This could lead to interesting quantitative results on improving information representations of model output. This could also be done for the comparison between SHAP and LIME post hoc explanations.
Furthermore, several other analyses on the data from the user study could still be performed. This includes:

- *Improvement for each participant*

  Here we look at the improvement for each of the individuals. Looking at the results from single participants, we might find different insights concerning the effect of the different conditions on the usability. This is especially important when looking at usability testing since the topic of usability can be very subjective. For some participants, a clear usability improvement or decrease can be found whereas other participants for example might experience each information representation as confusing leading to incomprehensible results.

- *Item-wise analysis*

  The item-wise analysis looks at individual metrics in order to find significant differences. This can be useful to create insights into for example an improvement that is not visible when looking at the average of all the questions combined. Some questions of the SUS can be ambiguous to the participant rendering them less valuable for calculating the final level of usability. By looking at the scores of individual questions, the researcher might be able to create more precise insights concerning changes of the perceived usability.

6.4 Conclusions and Design Guidelines

A clear preference for the SHAP-based adapted visualisation can be found within the results. As well as the need for more information and explanation of the model output. During the analysis of the experiment results, several design guidelines were composed:

- Adapted SHAP-based waterfall visualisations without extensive instructions do not increase the usability of the information product. No significant evidence was found that visualisations increase usability.
- Significantly longer processing times have been identified when visualisations are added. This was expected since the visualisations increase the total amount of information that is presented.
- Text as an additional explanation provides benefits to the overall explainability of the information. The results show that text significantly improves the usability of the model output, compared to the proposed SHAP-based visualisation.
- Text explanations can be combined with visualisations to maximize the usability of the information product used in this study.
- It is recommended to present SHAP-based waterfall visualisation with an additional instruction to improve understandability since the visualisation in itself is not sufficiently intuitive.
- Finally, it is recommended for SHAP experts to research other ways of intuitively presenting local SHAP-based results of individual model predictions.

The research provided insight into the design of explanations that focus on usability in the field of explainable machine learning. Implementing the proposed design guidelines while developing explainable machine learning applications could lead to applications that are more suitable for the end-user making the final product more usable. Although this study focused on explaining machine learning results, the results of this study can be generalized to other subjects concerning end-user design.

Bibliography

[1] S. Barua, I. M. Rahman, M. M. Hossain, Z. A. Begum, I. Alam, H. Sawai, T. Maki, and H. Hasegawa, "Environmental hazards associated with open-beach breaking of end-of-life ships: a review," *Environmental Science and Pollution Research*, vol. 25, no. 31, pp. 30880–30893, 2018.

[2] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust," *arXiv preprint arXiv:1907.12652*, 2019.

[3] C. V. Guimaraes, R. Grzeszczuk, G. S. Bisset III, and L. F. Donnelly, "Comparison between manual auditing and a natural language process with machine learning algorithm to evaluate faculty use of standardized reports in radiology," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 550–553, 2018.

[4] I. Secretariat, "Global integrated shipping information system."

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[6] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020.

[7] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15, 2019.

[8] P. Schmidt and F. Biessmann, "Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 431–449, Springer, 2020.

[9] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker, "Interacting meaningfully with machine learning systems: Three experiments," *International journal of human-computer studies*, vol. 67, no. 8, pp. 639–662, 2009.

[10] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[11] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural computing and applications*, pp. 1–15, 2019.

[12] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[13] L. Fischer, L. Ehrlinger, V. Geist, R. Ramler, F. Sobiezky, W. Zellinger, D. Brunner, M. Kumar, and B. Moser, "Ai system engineering—key challenges and lessons learned," *Machine Learning and Knowledge Extraction*, vol. 3, no. 1, pp. 56–83, 2021.

[14] M. T. Ribeiro, S. Singh, and C. 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*, pp. 1135–1144, 2016.

[15] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[16] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[17] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," 2020.

[18] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[19] E. REITER and R. DALE, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, p. 57–87, 1997.

[20] E. Reiter, "Natural language generation challenges for explainable ai," *arXiv preprint arXiv:1911.08794*, 2019.

[21] N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems," in *Recommender systems handbook*, pp. 479–510, Springer, 2011.

[22] H. Liu, Q. Yin, and W. Y. Wang, "Towards explainable NLP: A generative explanation framework for text classification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5570–5581, Association for Computational Linguistics, July 2019.

[23] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2018.

[24] K. Holtzblatt and H. Beyer, *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[25] A. Blandford, D. Furniss, and S. Makri, "Qualitative hci research: Going behind the scenes," *Synthesis lectures on human-centered informatics*, vol. 9, no. 1, pp. 1–115, 2016.

[26] K. H. Lim, L. M. Ward, and I. Benbasat, "An empirical study of computer system learning: Comparison of co-discovery and self-discovery methods," *Information Systems Research*, vol. 8, no. 3, pp. 254–272, 1997.

[27] M. Van Den Haak, M. De Jong, and P. Jan Schellens, "Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue," *Behaviour & information technology*, vol. 22, no. 5, pp. 339–351, 2003.

[28] E. L. Olmsted-Hawala, E. D. Murphy, S. Hawala, and K. T. Ashenfelter, "Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2381–2390, 2010.

[29] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[30] D. Kirsh, "A few thoughts on cognitive overload," 2000.

[31] S. H. Koch, C. Weir, D. Westenskow, M. Gondan, J. Agutter, M. Haar, D. Liu, M. Görges, and N. Staggers, "Evaluation of the effect of information integration in displays for icu nurses on situation awareness and task completion time: a prospective randomized controlled study," *International journal of medical informatics*, vol. 82, no. 8, pp. 665–675, 2013.

[32] M. A. Iverson, F. Ozguner, and L. C. Potter, "Statistical prediction of task execution times through analytic benchmarking for scheduling in a heterogeneous environment," in *Proceedings. Eighth Heterogeneous Computing Workshop (HCW'99)*, pp. 99–111, IEEE, 1999.

[33] P. Schmidt and F. Biessmann, "Quantifying interpretability and trust in machine learning systems," *arXiv preprint arXiv:1901.08558*, 2019.

[34] J. J. Louviere and G. Woodworth, "Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data," *Journal of marketing research*, vol. 20, no. 4, pp. 350–367, 1983.

[35] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

[36] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan, "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 97–105, 2019.

[37] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

[38] E. Rader, K. Cotter, and J. Cho, "Explanations as mechanisms for supporting algorithmic transparency," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018.

[39] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*, pp. 157–164, 2011.

[40] J. Brooke, "Sus: a "quick and dirty'usability," *Usability evaluation in industry*, vol. 189, 1996.

[41] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[42] J. R. Lewis, "The system usability scale: past, present, and future," *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, 2018.

[43] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (scs)," *KI-Künstliche Intelligenz*, pp. 1–6, 2020.

[44] R. Likert, "A technique for the measurement of attitudes.," *Archives of psychology*, 1932.

[45] J. Genest, J. Frohlich, G. Fodor, and R. McPherson, "Recommendations for the management of dyslipidemia and the prevention of cardiovascular disease: summary of the 2003 update," *Cmaj*, vol. 169, no. 9, pp. 921–924, 2003.

[46] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of learning and motivation*, vol. 2, pp. 89–195, Elsevier, 1968.

[47] Anonymous, "Ethics guidelines for trustworthy ai," Nov 2020.

[48] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 2, 2002.

[49] B. Fogg, G. Cueller, and D. Danielson, "Motivating, influencing, and persuading users: An introduction to captology," in *The human-computer interaction handbook*, pp. 159–172, CRC press, 2007.

[50] B. J. Fogg, "Persuasive computers: perspectives and research directions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 225–232, 1998.

[51] M. D. C. Tongco, "Purposive sampling as a tool for informant selection," *Ethnobotany Research and applications*, vol. 5, pp. 147–158, 2007.

[52] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 3, pp. 379–383, 2003.

A System Usability Scale

Table 8: The original SUS statements and modified statements

Original SUS Statements	**Modified SUS Statements**
I think that I would like to use this system frequently	I think that I would like to use this product frequently
I found the system unnecessarily complex	I found the product unnecessarily complex
I thought the system was easy to use	I thought the product was easy to use
I think that I would need the support of a technical person to be able to use this system	I think that I would need the support of a technical person to be able to use this product
I found that the various functions in this system were well integrated	I found that the various functions in this product were well integrated
I thought that there was too much inconsistency in this system	I thought that there was too much inconsistency in this product
I would imaging that most people would learn to use this system very quickly	I would imaging that most people would learn to use this product very quickly
I found the system very cumbersome to use	I found the product very awkward to use
I felt very confident using the system	I felt very confident using the product
I needed to learn a lot of things before I could get going with this system	I needed to learn a lot of things before I could get going with this product

Table 9: The final SUS statements used in the user experiment

Final SUS Questions Dutch
1) De informatie helpt mij te begrijpen hoe het model werkt.
2) De informatie is voldoende gedetailleerd.
3) De informatie laat mij weten hoe precies het model is voor individuele schepen.
4) De informatie laat mij weten hoe betrouwbaar het model is.
5) Ik vond de informatie onnodig ingewikkeld.
6) Ik denk dat ik extra uitleg nodig heb van een expert.
7) De informatie van het model is erg voorspelbaar.
8) Ik verwacht dat ik betere keuzes kan maken met hulp van de informatie uit het model.
9) Ik heb vertrouwen in het model. Ik denk dat het goed werkt.

B Consent Form

Experiment evaluation of explainable AI interpretability and informativeness. Hi! Thank you for participating. In this study you will provide demographic information, complete an online evaluation task and answer closed questions about the task. Please read the consent form shown below.

- I confirm that the goal of the research project has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily.
- I am aware of and consent to any potential risks of the research project. And if I don't feel comfortable doing a task, I can quit at any moment.
- I consent to the material I contribute being used to generate insights for the research project.
- I understand that my participation in this research is voluntary, that it is not a requirement of my work activities, and that I may withdraw from the study at any time.
- I consent to allow the fully anonymised data to be used for future publications and other scholarly means of disseminating the findings from the research project.
- I confirm that I am 18 years of age or over.
- I understand that I can request any of the data collected from/by me to be deleted.
- I agree to take part in the above study.

My name is _____ and I accept the statements mentioned above.

C Protocol Qualitative Interviews with Think-Aloud method

C.1 Protocol of the semi-structured interview

Before the experiment starts (2 min)

1. Greet the participant and explain that there will be room for questions and comments after the experiment.
2. Ask if screen and audio can be recorded during the experiment.
3. Room for questions, informative and on the subject of work (activities, current situation, etc).
4. Begin screen sharing and show the presentation with the introduction and experiment.
5. Introduce the topic of the research.
6. Explain the Think-aloud method and the experiment.
7. Ask if the participant has any question. If no questions, begin the experiment:

During the experiment (10 min)

1. Show the first instance

2. Write down comments concerning the behaviour of the participant
3. Notice whether the participant understands the think-aloud protocol
4. When the participant is finished with the instance: check whether every part of the instance has been discussed by the participant. If a part was missed, point it out and ask what the participant thinks about it.
5. If everything has been discussed, move to the next instance.
6. Repeat steps 2 to 5 until all instances have been discussed

After the experiment (5 min)
Discussing SHAP values and NLG explanations (5 min)
Intervention point: For the next questions I'd like to explain about SHAP and natural language generated explanations and how they can be used to create interpretable explanations. SHAP uses individual feature importance to create visualisations of the model output that allow for explainability of the decision making. NLG also uses feature importance to create a text-based summary of the most important features of the model output (while also providing a percentage of their total contribution). In the context of improving explanations made by information systems I would like to review the different SHAP and NLG representations with you.

1. Did you realize that SHAP and NLG explanation methods were used, and can you explain how?
2. How did you interpret the model explanations?

   Use notes from the observation if needed.
3. Can you think of interaction techniques (like buttons or visualisations) that can help you with interpreting the model outcome?
4. How could visualisations and text be used to increase (model interpretation) the effectiveness of the explanations?

Summary and conclusion (3-5 min)

1. Summarize the answers and conclude.
2. Ask if there are any questions or final remarks.
3. Thank the participant for participating in the study.
4. Stop and save the audio recording.

Phase 3: After the interview

1. Write up a short conclusion instantly after the interview.
2. Save and gather all important collected data (consent forms, notes, recordings and conclusion) on your laptop under one folder for each focus group.

1. Ask if the participant has any questions or comments.
2. Thank the participant for participating in the research.

D Protocol Pilot Quantitative Questionnaire

D.1 Testing the Questionnaire

Before sending out the experiment

1. Prepare the questions and decide on what kind of feedback is preferred.
2. Go through the complete pilot questionnaire and check for errors
3. Send the pilot to thesis supervisors for final check
4. Process final feedback from thesis supervisors

Sending the experiment to the IDLAB

1. Prepare introductory mail with information on the research topic and goal of the pilot
2. Include examples of feedback interesting for the final questionnaire (length, duration, etc)
3. Include the link to the questionnaire and the required password
4. Mention the deadline for completing the questionnaire
5. Send a reminder one day before the deadline

E Protocol Quantitative Questionnaire with System Usability Scale

E.1 Protocol of the prototype evaluation

Part 1: Before the experiment

1. Determine the minimum number of evaluations necessary for sufficient statistical power.
2. Determine the participant selection criteria.
3. Generate suitable questions by adapting System Usability Scale.
4. Generate a standard consent form for the participants to store their responses.
5. Decide upon a suitable online questionnaire environment.

Part 2: Conduct the evaluations of the newly proposed SHAP visualisations with Quantitative Questionnaire

1. Send an email with the introduction of the research
2. Send the link with the questionnaire and the password.

Introduction (5min)

1. Greet the participant, and (if needed) familiarize them with the evaluation.

2. Explain the purpose of the study (evaluating explainability approaches).

3. Explain research methods of the study including think-aloud protocol.

4. Ask for consent (refer to consent form) to gather answers from the participant during the session via note-taking, audio recording and screen recording.

5. Send participants the consent form to sign.

Perception of interaction with SHAP-based representations of model output based on observations (30 min)

1. How could these techniques be implemented?

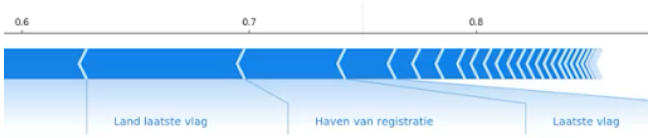
F Qualitative results think-aloud interviews:

F.1 Interview 1: 07-04-2021

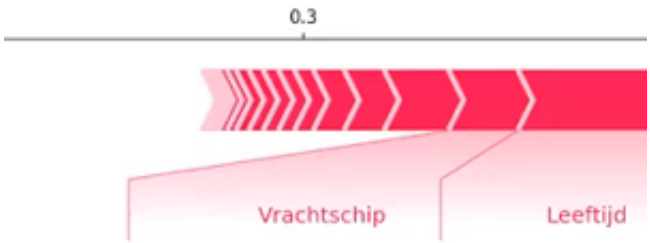
Interpretation of each instance

Interpretatie feature bijdrage SHAP platen (eerste gepresenteerde SHAP plaat)

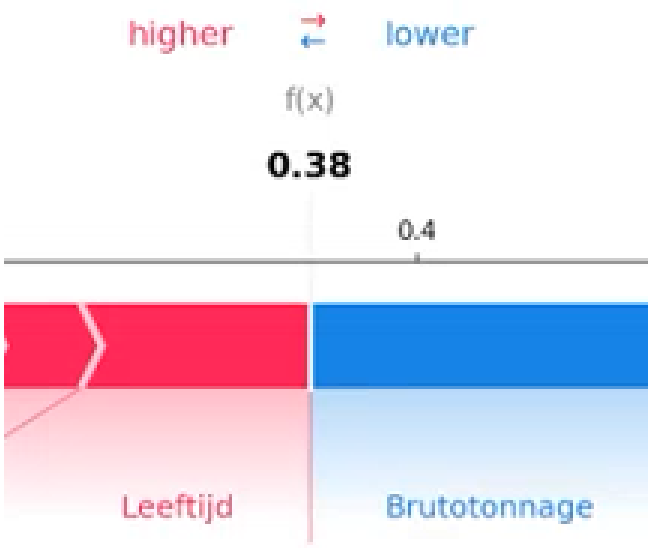
1. Verbaast zich over het plaatje waarbij ‘land laatste vlag’,’haven van registratie’, en ‘laatste vlag’ als risico verlagend zijn weergegeven



Hij zou de laatste vlag juist als een hoog risico verwachten Mogelijk denkfout: laatste vlag heeft hier juist een verlagend effect op de voorspelling omdat het land geen geschiedenis heeft van beaching Misschien dacht hij dat de feature ‘laatste vlag’ altijd voor het model als verlagend wordt meegenomen. Dit verschilt natuurlijk per instance/schip. 2. Kan moeilijk geloven dat het ‘type’ van het schip zoveel bijdraagt aan het verhogen van de voorspelling



Type kan heel erg afhankelijk van de markt zijn (marktontwikkeling, bepaalde vraag naar bepaalde schepen) of is de vraag misschien helemaal weggevallen 3. “Leeftijd kan een rol spelen.” 4. “Als het gaat om beachen dan zal ook het brutotonnage een rol spelen.”



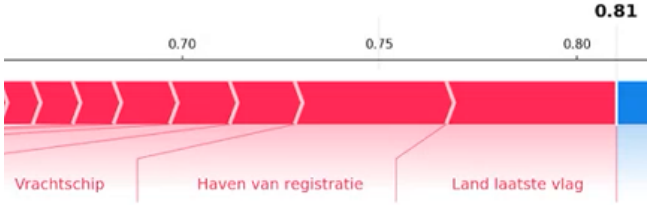
“Omdat er voor grotere schepen minder mogelijkheden zijn” (minder slooplocaties aanwezig)

Interpretatie feature bijdrage SHAP platen (tweede gepresenteerde SHAP plaat)

1. Vraagt zich af of ‘jaren sinds laatste omvlagging’ echt een toegevoegde waarde heeft op de voorspelde kans 2. “Verbaast mij heel erg dat Veritas hoog scoort als risico”

Dit is een interpretatiefout aangezien. Hij leest van links naar rechts en interpreteert dat de features helemaal links het meeste bijdragen aan de verhoging van de voorspelling. Het is lastig om de feature ‘Veritas’ te koppelen aan een stukje van de SHAP force plot balk. Hierdoor is het moeilijker om de feature importance af te lezen. 3. “Hetzelfde als het BBP per capita. Die link is niet helemaal duidelijk.”

Hij weet niet of de specifieke vlaggen die gebruikt worden voor beachen gelieerd zijn met deze feature. 4 “Jaren liever in maanden of misschien wel weken.” 5. “Bij de anderen kan ik mij wel iets voorstellen. Land laatste vlag en haven van registratie eigenlijk gekoppeld aan elkaar.”



Is ook wel te zien in hun bijdrage aan de voorspelling. 6. Vraag of hij zich kan vinden in de relatief hoge voorspelling (0.81): “Maar het verbaast mij het meest dat ie Veritas en dat BBP dat ie daar het meest op scoort.” “En die 4 jaar...” (duidt op ‘jaren sinds laatste omvlagging’) “Als je dat uitdrukt in weken of maanden dan zou ik me iets voor kunnen stellen op een verhoogde kans op overtreding. Eerder dan 4 jaar. Als hij nog 4 jaar in de vaart gehouden moet worden dan had je (als eigenaar) de intentie om er iets anders mee te doen.” Na uitleg over de SHAP plaat en hoe je deze moet interpreteren

(balkje laat zien hoeveel invloed een feature heeft, vanuit het midden lezen): “Goed dat je zegt dat in het midden de grootste risico’s liggen en niet aan de zijkant.” “Land laatste vlag en haven van registratie zou je bijna als een enkele feature kunnen zien. Ik weet niet waarom haven en land uit elkaar getrokken is. Haven is Panama en land is Panama. Wat is nou de toegevoegde waarde? Het is mij niet bekend dat voor sommige landen bepaalde registratie haven een nog groter risico vormen. Het gaat vaker om de vlaggen die een hoger risico zijn dan de havens voor zover ik weet.”

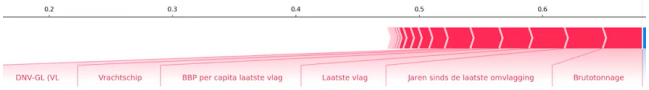
Interpretatie feature bijdrage SHAP platen (derde gepresenteerde SHAP plaat) Bij de tekst explanation was hij het niet eens met dat de feature Veritas de kans op overtreding verlaagde.

- De voorspelde kans op overtreding:
  - Wordt verhoogd door: leeftijd (7 jaar en 8 maanden)
  - Neemt af vanwege: bruto tonnage (2989), classificatie bureau (Veritas)

Dit kan ik verkeerd toegevoegd hebben dus zal ik het verwijderen voor het volgende experiment

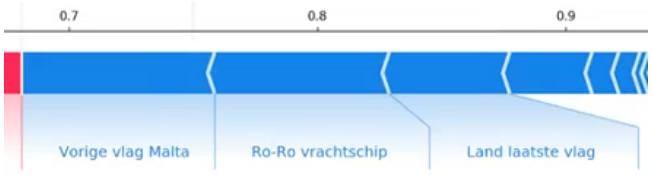
Interpretatie feature bijdrage SHAP platen (vierde gepresenteerde SHAP plaat)

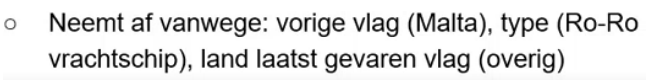
Leeftijd	27 jaar (+ 8 maanden)
Type	Ro-Ro vrachtschip
Bruto tonnage	11866
Jaren sinds laatste omvlagging	1



- De voorspelde kans op overtreding:
  - Wordt verhoogd door: bruto tonnage (11866), aantal jaren sinds laatste omvlagging (1 jaar)
  - Neemt af vanwege: vorige vlag (Malta), type (Ro-Ro vrachtschip), land laatst gevaren vlag (overig)

“Ik kan me wel iets voorstellen bij leeftijd en tonnage. En ook laatste jaren sinds omvlagging dat dat een bepaalde rol speelt.”





“Alleen niet qua vlag zie ik niet dat het risico afneemt. En Ro-Ro, het type, dat weet ik ook niet. Want we weten wel dat Ro-Ro schepen ook gewoon gesloopt worden op stranden in Azië. Dus waarom het (de voorspelling) daar vermindert is me nog niet duidelijk. En ook Malta, waarom dat de kans zou verminderen. Een schip varend onder Maltese vlag wat verkocht wordt naar een van die andere exotische staten zegmaar. Volgens mij kan dat gewoon voorkomen. Ik zou dat niet zelf inschatten dat daardoor het risico afneemt. Als je zegt ‘Er zijn weinig schepen varend onder Maltese vlag die op de stranden belanden’ dat zou dan misschien kunnen. Omdat ze misschien toezicht uitoefenen. Malta is ook een EU vlag. Dus als het recent zou zijn (deze beaching/omvlagging) dan zou het heel erg slim zijn. Maar ik denk ook dat ze om inspecties te omzeilen bij de verkoop omvlaggen naar een exotische vlag. Dus dat zou wel een verklaring kunnen zijn. Niet dat je zegt ‘Schepen onder Maltese vlag worden niet omgevlagd om daarna gesloopt te worden op stranden’ dat zou ik niet durven zeggen.” Vraag: Zou je bij dit schip de keuze maken om wel de eigenaar te inspecteren of niet? “Ja ik zou hier wel een inspectie op uit kunnen voeren zegmaar want er zijn wel risico’s. De leeftijd en het tonnage, het type gaat ook die kant op om gesloopt te worden.”

Laatst gevaren vlag	Overig
Vorige vlag	Malta
Classificatie bureau	DNV-GL (VL)
Haven van registratie	Overig

“En van belang is dan nog wel de laatst gevaren vlag (nu ‘overig’).” “Ik weet niet hoe je dat jaar nu moet zien (‘jaren sinds laatste omvlagging’ is 1). Is dit nu een schip wat gebeached is 1 jaar na de laatste omvlagging? Of komt die in het model van de kans op beachen?” Huib geeft aan dat het qua uitlegbaarheid een beetje dubbel is dat: De voorspelling afneemt naarmate de ‘jaren sinds laatste omvlagging’ toeneemt De voorspelling toeneemt als de ‘leeftijd’ toeneemt “Als ze (schepen) worden omgevlagd naar een bepaalde vlaggenstaat die risicovol is (bijv. Saint Kitts and Nevis), dan heb je dat risico van een verlaagde kans (bij jaren die verstrijken) wel. “ “Maar als het schip wordt omgevlagd van een Maltese vlag naar de Nederlandse vlag bijvoorbeeld, dan is dat risico van ‘jaren sinds laatste omvlagging’ bijna nihil. Want vanuit NL vlag bijna niks gebeached.” “Dat (feature ‘jaren sinds laatste omvlagging’) moet je echt koppelen aan de laatste vlag.” Wat zijn je ideeën over de tekst uitleg?:

- De voorspelde kans op overtreding:
  - Wordt verhoogd door: bruto tonnage (11866), aantal jaren sinds laatste omvlagging (1 jaar)
  - Neemt af vanwege: vorige vlag (Malta), type (Ro-Ro vrachtschip), land laatst gevaren vlag (overig)

“Je moet dan wel toevoegen waar we het net over hadden. Het is goed dat je zegt waarom het risico verhogend of verlagend werkt. ” “Maar in deze zou dus het risico neemt af vanwege die vlaggen dat zie ik dus niet en ook niet het type en ook niet het land laatst gevaren vlag.” “Wel goed dat het uitgelegd wordt. Los van het feit dat wat er uitgelegd wordt juist is of dat je het er mee eens bent.”

Interpretatie feature bijdrage SHAP platen (vijfde en laatste gepresenteerde SHAP plaat)

Leeftijd	19 jaar (+ 10 maanden)	**Laatst gevaren vlag**	Liberië
Type	Gekoeld vrachtschip	**Vorige vlag**	Duitsland
Bruto tonnage	11382	**Classificatie bureau**	Veritas
Jaren sinds laatste omvlagging	Minder dan 1 jaar	**Haven van registratie**	Monrovia

1. “Neemt af vanwege classificatie bureau Veritas. Daar heb ik eerder opmerkingen over gegeven.” 2. “De anderen (features) zijn wel verklaarbaar.” 3. “Maar bij deze moet je ook wel naar de eigenaren kijken. Want ik weet nog uit het model... En dan weten we dat een bepaald bedrijf of eigenaar vaak naar voren komt die in het verleden gebeached is en die ook daarvoor veroordeeld is. Dan kun je je afvragen: dit was zijn werkwijze in het verleden. Het zou enorm stom zijn als hij dat nu nog zou doen.” “Als het dat bedrijf is dat ik denk dat het is dan moet je er wel aan koppelen in het model dat je bepaalde eigenaren ondanks hun overtredingen in het verleden vanwege hun veroordelingen het risico misschien wel is afgenomen (vanwege bekendmaking werkwijze, extra oplettendheid).” “Ik weet dat dit model gebruik maakt van lijsten van schepen die in het verleden gebeached zijn. Bepaalde koelschepen kwamen naar voren en die werden vaak ook gekoppeld aan eigenaren. Als een eigenaar dan veroordeeld is dan zou je je af kunnen vragen of het schip/type (koelschepen) inderdaad nog zo hoog moeten scoren en dat de kans op overtreding nog zo hoog is? Misschien moet hij nog wel gecontroleerd worden. Maar je kan je wel afvragen of de voorspelde kans op overtreding nog zo heel hoog is omdat hij al een keer veroordeeld is. Neemt dan de kans toe of juist af (na een veroordeling)?” “Iemand die een overtreding heeft gepleegd betekent niet dat hij altijd in overtreding zal blijven. Sommige bedrijven leren ook van hun fouten en overtredingen en zien misschien dat het niet loont. Als dat eraan gekoppeld is, dat haal ik hier niet uit, maar dan heb ik daar nog wel een

vraagteken bij.” “Dus dit (hogere kans van koelschepen) komt uit het model maar dan zal je niet als inspecteur in een keer 1 op 1 dat model over nemen, maar als slechts 1 van een x aantal middelen moeten gebruiken om te gaan selecteren en om inspecties te gaan doen. Dus niet als het enige middel maar als een aanvulling op andere zaken.” 4. “Haven van registratie Monrovia maar volgens mij heeft Liberia maar 1 haven. Dus als je dat koppelt. Volgens Mij gaat het meer om het land dan de haven.”

F.2 Results categorized

Content
Requirements: Moet een sterk verhaal zijn (goed onderbouwd)
Wat ze terug horen Waarom de keuzes gemaakt zijn? Zonder de details prijs te geven Daarom zijn de keuzes gemaakt Als het fout gaat moet je bewust zijn dat je dingen moet gaan uitleggen Waarom bepaalde zaken gedaan of niet gedaan In dit werkveld bepaalde media die erg gericht zijn hoe de overheid werkt en dit werkveld Nu heel relevant transparantie en openheid van de overheid
Vaak inspecteren op basis van een gevoel Risico’s management, analysis prioritering Flexibel en wijzigen Iemand een gevoel geven van een grote pakkans
hoe leg je uit dat het een steekproef is? Geeft niet aan waarom die in een steekproef valt Dan zou je het profiel uitleggen Je valt in een profiel om gecontroleerd te worden Je moet wel kunnen duiden waarom je die ene partij wel controleert en de ander niet
Features: Vlag allesbepalend voor de wetgeving Zelfde als de haven Brutotonnage Schepen minder dan 500 ton vallen niet onder
Usability (navigatie, informativiteit, zichtbaarheid)
Profielen → gebaseerd op bepaalde criteria Nu met machine learning is nieuws → moet uitlegbaar zijn voor ondertoezichtstaanden Inspecteurs werken al 25 jaar in bepaald patroon Keuze maken door een bepaald patroon Machine learning is helemaal nieuw Ander aspect is de privacy gevoeligheid Manier vinden om als inspecteur uit te leggen waarom het zo is Zonder de techniek in te gaan Soort generatiekloof/leeftijdskloof Team, meesten zijn 45+ Ook een aantal 60 ers Gat in kennis en gewenning
Visualisation (grootte, informativiteit)
Jaren sinds laatste omvlagging Niet specifiek genoeg Jaren is te groot Als een schip gesloopt is Vlak ervoor omgevlagd Gebeurd vaak om regelgeving en moeilijke inspecties te vermijden Beter maanden of misschien wel weken
Type Liever wat specifieker Miste IMO nummer Erg kenmerkend voor het schip blijft uniek en identificatienummer van het schip Erg relevant Bepaalde vlaggen worden vaak gebruikt bij beaching Vlaggenstaten die erg bekend staan als makkelijk Geen inspecties
Laatst gevaren vlag erg belangrijk maar komt niet zo voor in het model/visualisatie
Veritas scoort hoog Niet heel duidelijk BBP per capita laatste vlag Ook niet veel toegevoegde Weet niet of dat gelinkt is aan de laatste vlag Jaren sinds laatste omvlagging Liever in maanden of aantal weken
Type anders beschrijven Zie niet in waarom de kans afneemt als het classificatiebureau Veritas is Geen toegevoegde waarde
Ziet niet waarom Malta en Ro-Ro type de kans vermindert Schip varend onder Maltese vlag
Ook naar de eigenaar kijken Bepaalde eigenaar die vaak voorkomt Die verleden heeft met beachen Die is daarvoor veroordeeld Neemt dan de kans toe of af Als hij al in overtreding is geweest Blijft hij dan een hoger risico Gekoelde vrachtschepen Bias hier vanwege eerdere veroordeling Monrovia niet veel toegevoegde waarde Gaat meer om het land dan echt de haven

Tekst generation: Tekst uitleg wordt gewaardeerd De inhoud is wat minder Maar fijn dat het er is Belangrijkste features als tekst 3 tot 5

F.3 Interview 2: 10-5-2021

Findings

Naja dat is dan vaak toch in combinatie met bepaalde dingen die niet zo 123 voor de hand liggen. Dat maakt het wel weer lastiger hé hoe je dat vervolgens moet uitleggen. Want dat zal een bepaalde combinatie zijn die een hogere score oplevert.

Vrachtschip ja zegt ook niet zoveel er zijn natuurlijk veel vrachtschepen.

Nou kijk die kleurtjes maken het wel helder

Want dat weet ik dus niet wat dat dan voor invloed heeft. Dus die kans op overtreding en de laatst gevaren vlag 'overig' ja dat ehh ja goed 'overig' ja dus niet bekend als ik kijk naar het plaatje boe dan wordt het toch wel weer... bulkschip, classificatiebureau, containerschip.,, dat ja dat denk ik dat ja daar kan ik eigenlijk geen chocola van maken wat dat nou eigenlijk ook betekent.

dat zou dan vreemd zijn want hij staat wel bij dat uitgeschreven stukje (text explanation) laatst gevaren vlag 'overig'. Dus is wel kans verhogend terwijl als je daar geen uitspraken over kan doen omdat hij te weinig voorkomd.

Dat zou dan kunnen omdat die zo weinig voorkomt dat het niet een gebruikelijke vlaggenstaat is en dus wel een risico. Dan moet je het op die manier uitleggen. Dan zit wel die uitleg van Paul die zit er wel bij, is er wel bij nodig om te weten dat het dan weinig voorkomende vlag ehh land is.

Als je kijkt naar de balk ehh na goed dat is rood jaren sinds de laatste omvlagging, ja dat klopt dan denk ik wel een dingetje. Ehh bruto tonnage ja dat zal wel het is niet zo'n super groot schip maar het kan meespelen, dat weet ik niet. Vorige vlag Malta nou dat is inderdaad Malta Europees dus dat is weer blauw. Roll on Roll of vrachtschip

Maar de combinatie met andere factoren maken het denk ik weer wel dan hoog scorend.

het wat lastig zeker ook dat moet je in combinatie zien met het type schip en dat is ook niet al te eenduidig dat een schip van 1000 ton direct hoog scoort. Het is niet iets waar je dan direct op triggert.

Nou de leeftijd, 40 jaar, is wel een hoge leeftijd maar misschien in combinatie met het type schip juist weer minder hé dat die boorschepen toch over het algemeen langer meegaan dat kan natuurlijk. Hé dus die combinatie zou je dan moeten leggen.

Type gekoeld vrachtschip is wel een bepaald type schip. Dat kan wel een bepaalde reden zijn dat ehh het minder vracht zou kunnen hebben (weer een combinatie van features).

Vorige vlag Duitsland, ja dat is van Duitsland een Europees land naar Liberie buiten Europa ja dat is wel een belangrijk ding denk ik (ook een combinatie).

Brutotonnage beetje gemiddeld schip. Leeftijd ook wel iets, niet heel jong schip maar ook niet heel oud. Meer de combinatie denk ik ook met die andere factoren.

Sowieso jaren sinds laatste omvlagging dat is ook wel kenmerkend altijd van 'oeps' zeker omdat ie voorheen in Duitsland was. Dat staat er niet bij, dat mis ik dan misschien, dat zou ik misschien verwacht hebben. Want de vorige vlag was Duitsland en nu ineens Liberie, dat is wel een bijzonder kenmerk.

Ik mis misschien nog wel dat land van Duitsland naar Liberie dat is wel een dingetje. Zowel het termijn als het type land waarnaartoe het omgevlagd is dat mis ik dan nog wel.

Maar dat is wel, die combinatie dus niet die factor alleen, maar die combinatie van die twee factoren maakt het, of eigenlijk die drie factoren, maakt het bijzonder. Hé dus: land vorige vlag, land laatste vlag, en het termijn daartussen (jaren sinds laatste omvlagging) dan denk je van 'oeps', waarom is die omgevlagt van Duitsland naar Liberie? Waarom is dat gebeurt? He plus die 19 jaar, is niet super

hoog maar goed gezien ook weer dat koelschip dan denk je van dat zou wel kunnen. Dus dat is wat je ook zou verwachten dan op het moment dat je zo'n hoge score ziet (0.93).

Kijk die wordt verhoogd door de laatst gevaren vlag, Liberie, dat geloof ik. Maar als je die combinatie met Duitsland en dat minder dan 1 jaar (omvlagging) dat is denk ik de echte trigger.

F.4 Interview 3: 27-5-2021

Findings

ik weet niet hoe ze dat hebben ingeschat.

Dus ik weet niet hoe je op die getallen komt natuurlijk.

alleen ja die lijst daar links (features van de decision plot) dat denk ik dat je daar nog wat andere dingen bij moet.

Nou ik vind het vorige plaatje (decision) vond ik beter. Ja die vond ik echt beter want dan kan je in één keer zien met al die strepen van hoe is die afweging weet je wel. Dat is van alle die ik nu gezien heb het beste

Ja wat de afweging is snap je en dan kan je er zelf ook nog even naar kijken van 'oke' of dat dan wel klopt als je het gaat controleren. Want dat is het mooiste van dat plaatje wat je van de vorige keer had. Alleen dan moet je hem nog wel met een aantal dingen uitbreiden

Maar het kan een signaal zijn, want soms krijgen wij ja ff denken ja soms doet port state controles aan boord van schepen. Vroeger was dat meer dan nu omdat we nu niet meer zo dicht bij elkaar zitten maar dan kregen we nog wel eens een seintje.

Ja ja en dan hadden ze misschien nog wat aan boord gehoord van bemanningsleden en dat soort dingen. Maar dat krijg je nooit in dit systeem natuurlijk verwerkt maar dan zou je echt wat ik eerder al gezegd heb, iets van internet moeten afhalen, bepaalde berichten en ehh ja hoe je dat in het systeem kan krijgen ja dat weet ik natuurlijk niet

Want je gaat niet zo'n klein scheepje, of je kan ze stapelen, wat ik eerder vertelde. Dat kan ook nog hé dat zou in principe kunnen. Maar dit plaatje opzich is wel mooi. Alleen die afweging die verbaast mij weer maarja oke.

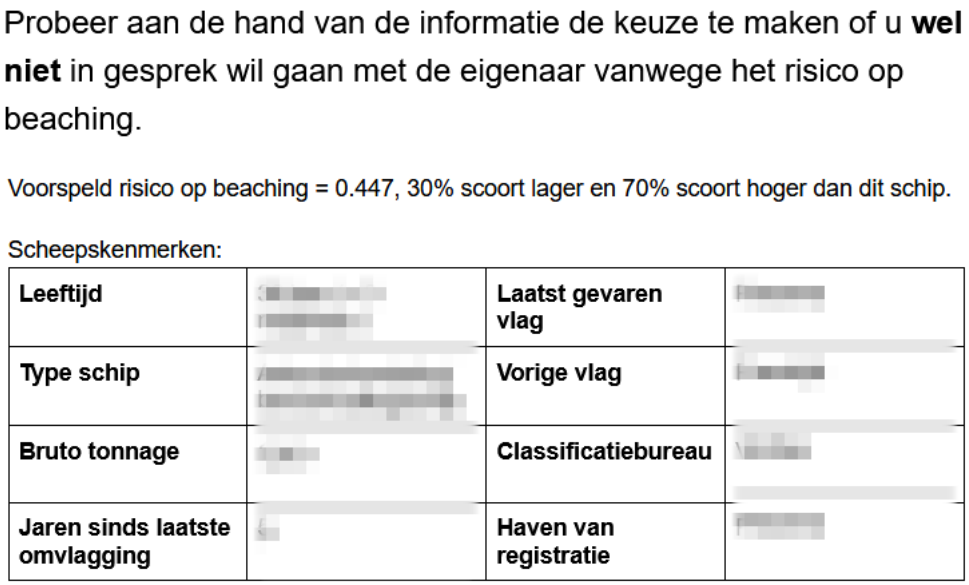
En even kijken, de schrootprijs staat die hier ook bij? Nee die staat er ook niet bij.

Als je voorspelmodellen wil maken met de afvalbranche, is dat ehh ja, moet dat ook kunnen. En dan moet je het eigenlijk per afvalstroom doen: koperafval, aluminium afval, kunststof afval.

en de vorige vlag Duitsland!, zie je het? Heb ik net verteld, zie je dus, vorige vlag is Duitsland en nou wordt het ineens Liberia, zie je het, en minder dan een jaar, en hij is behoorlijk oud, zie je. Dus dit is inderdaad een groot risico. Maar dan kan je beter dat andere plaatje (decision plot) doen, dat vind ik dan duidelijker.

eigenlijk zou je ook onderhoud van het schip of iets dat je dat ergens weghaalt,

G Final Experiment Design



Probeer aan de hand van de informatie de keuze te maken of u **wel** of **niet** in gesprek wil gaan met de eigenaar vanwege het risico op beaching.

Voorspeld risico op beaching = 0.447, 30% scoort lager en 70% scoort hoger dan dit schip.

Scheepskenmerken:

Leeftijd		**Laatst gevaren vlag**	
Type schip		**Vorige vlag**	
Bruto tonnage		**Classificatiebureau**	
Jaren sinds laatste omvlagging		**Haven van registratie**	

Figure 19: *This figure example instance of condition 1 of the online experiment.*

Probeer aan de hand van de informatie de keuze te maken of u **wel** of **niet** in gesprek wil gaan met de eigenaar vanwege het risico op beaching.

Voorspeld risico op beaching = 0.74, 75% scoort lager en 25% scoort hoger dan dit schip.

Scheepskenmerken:

Leeftijd		**Laatst gevaren vlag**	
Type schip		**Vorige vlag**	
Bruto tonnage		**Classificatiebureau**	
Jaren sinds laatste omvlagging		**Haven van registratie**	

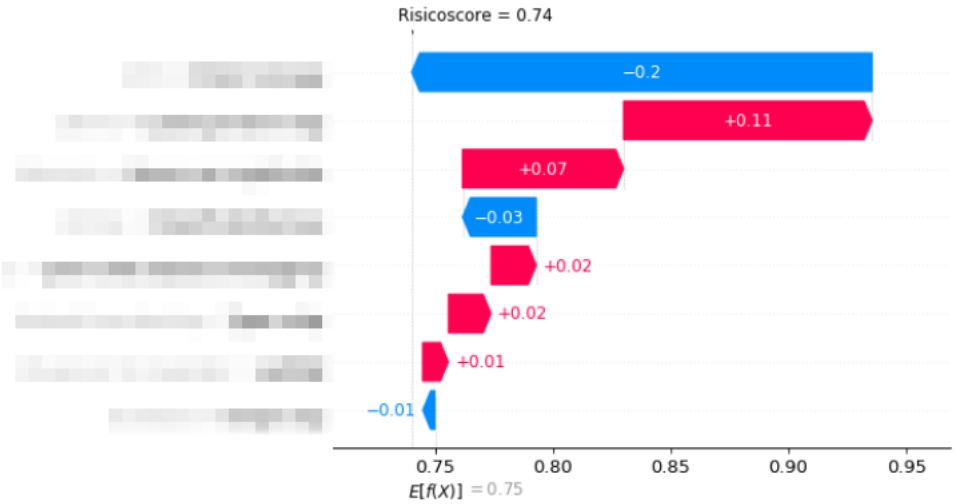


Figure 20: *This figure example instance of condition 2 of the online experiment.*

Probeer aan de hand van de informatie de keuze te maken of u **wel** of **niet** in gesprek wil gaan met de eigenaar vanwege het risico op beaching.

Voorspeld risico op beaching = 0.411, 25% scoort lager en 75% scoort hoger dan dit schip.

Scheepskenmerken:

Leeftijd		**Laatst gevaren vlag**	
Type schip		**Vorige vlag**	
Bruto tonnage		**Classificatiebureau**	
Jaren sinds laatste omvlagging		**Haven van registratie**	

Hoger risico door:	
Lager risico door:	

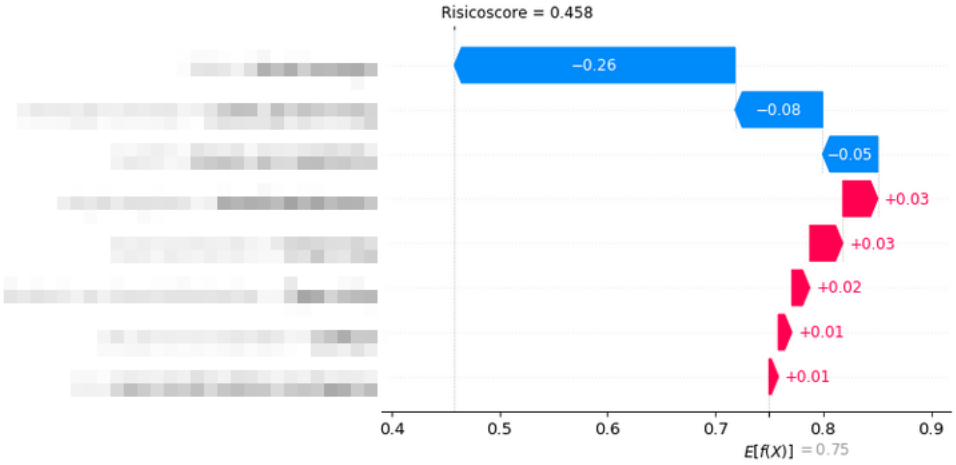
Figure 21: *This figure example instance of condition 3 of the online experiment.*

Probeer aan de hand van de informatie de keuze te maken of u **wel** of **niet** in gesprek wil gaan met de eigenaar vanwege het risico op beaching.

Voorspeld risico op beaching = 0.458, 32% scoort lager en 68% scoort hoger dan dit schip.

Scheepskenmerken:

Leeftijd		**Laatst gevaren vlag**	
Type schip		**Vorige vlag**	
Bruto tonnage		**Classificatiebureau**	
Jaren sinds laatste omvlagging		**Haven van registratie**	



Hoger risico door:	
Lager risico door:	

Figure 22: *This figure example instance of condition 4 of the online experiment.*