

# Designing an Interface for an Explainable Machine Learning Risk Model for Predicting Illegal Shipbreaking

*Sander Treur*

A dissertation submitted in partial fulfilment  
of the requirements for the degree of  
**Master of Science**

in

**Human Computer Interaction**

under the supervision of

Hanna Hauptmann, Utrecht University

Michael Behrisch, Utrecht University

Paul Merkx, IDLab at ILT

Stephanie Wassenburg, IDLab at ILT

Victor Ciulei, IDLab at ILT



**Utrecht  
University**

Department of Information and Computing Sciences

July 13th, 2022

# Abstract

In the battle against 'beaching', referring to the illegally dismantling of ocean-going vessels, the IDLab has developed machine learning models to provide predictive risk assessments of currently active ships. As the predictions that the models produce will support decision-making of the ship inspectors of the Dutch Human Environment and Transport Inspectorate, explainable artificial intelligence (XAI) techniques were used to find explanations for the predictions. As these inspectors are experts in their domain, but novices in the field of (X)AI and data science, challenges arise with regard to making the model results accessible for them. This exemplifies a larger question on how humans interact with (X)AI and concerns aspects such as visualisation and interaction, with the aim to make predictive machine learning models accessible, understandable and trustworthy for the decision-making end-users. As existing XAI visualisation studies mainly target data scientists, the current research contributes to getting a better understanding of how to effectively design XAI visualisations for end-users. In this research, a dashboard interface design is proposed, which was created following a systematic top-down approach, including a literature review, requirements analysis with stakeholders, brainstorm and sketching session, low-fidelity prototypes, focus group sessions, the implementation of a high-fidelity prototype and a final experiment with the target users. The resulting prototype was evaluated in terms of understandability, usability and reliance, and indicated promising results. The interface was received positively by the inspectors and findings from the evaluation show no reason to assume that there are major flaws in the design. Furthermore, the proposed design for the model explanations was found to be understandable in terms of visuals, while also opening the door to new challenges regarding trust in XAI models and interpretability of their explanations.

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	Motivation . . . . .	8
1.2	Problem Statement . . . . .	10
1.3	Research Question . . . . .	11
<b>2</b>	<b>Literature Review</b>	<b>12</b>
2.1	Explainable Artificial Intelligence . . . . .	12
2.1.1	Definition . . . . .	12
2.1.2	The Need for XAI . . . . .	12
2.1.3	Explainable Artificial Intelligence in Human Computer Interaction	13
2.2	Types of Explanation . . . . .	13
2.2.1	Global versus Local Explanations . . . . .	13
2.2.2	Model-Specific versus Model-Agnostic Methods . . . . .	13
2.2.3	Intrinsic Models versus Post-Hoc Methods . . . . .	14
2.2.4	Local Explainability: Feature Importance . . . . .	14
2.3	Visual Analytics . . . . .	14
2.3.1	Definitions . . . . .	14
2.3.2	Visualisation Process . . . . .	15
2.3.3	Types of Users . . . . .	16
2.4	Visualising Feature Importance . . . . .	17
2.4.1	Bar Charts . . . . .	17
2.4.2	Heatmaps . . . . .	17
2.4.3	Scatter Plots . . . . .	17
2.4.4	Tables . . . . .	18
2.4.5	Word Clouds & Word Highlighting . . . . .	18
2.4.6	Force Plots . . . . .	18
2.4.7	Dashboards . . . . .	19
2.5	Designing for AI Novices . . . . .	19
2.5.1	Design Guidelines . . . . .	19
2.5.2	Design Goals . . . . .	20
2.6	Related Design Studies . . . . .	21
2.6.1	EluciDebug . . . . .	21
2.6.2	XAI Interface for Customer Turnover Predictions . . . . .	22

---

2.6.3	WEET Interface . . . . .	23
2.6.4	Common Design Themes . . . . .	24
<b>3</b>	<b>Background</b>	<b>26</b>
3.1	Data Structure of the Models . . . . .	27
3.2	Model Output and Explanation . . . . .	27
<b>4</b>	<b>Methodology</b>	<b>29</b>
4.1	Precondition Phase . . . . .	29
4.1.1	Learn (Literature Review) . . . . .	29
4.1.2	Winnow . . . . .	30
4.1.3	Cast . . . . .	30
4.2	Core Phase . . . . .	31
4.2.1	Discover (Requirements Analysis) . . . . .	31
4.2.2	Design . . . . .	31
4.2.2.1	Brainstorm and Sketching Session . . . . .	32
4.2.2.2	Low-Fidelity Prototypes . . . . .	32
4.2.2.3	Focus Groups . . . . .	32
4.2.3	Implement . . . . .	32
4.2.4	Deploy . . . . .	32
4.3	Analysis Phase . . . . .	32
4.3.1	Reflect . . . . .	32
4.3.2	Write . . . . .	32
<b>5</b>	<b>Requirement Analysis</b>	<b>33</b>
5.1	Findings from Inspector Interview . . . . .	33
5.2	Final Set of Requirements . . . . .	33
5.2.1	Persona . . . . .	34
5.2.2	Requirements Sources . . . . .	35
5.2.3	Requirement Priorities . . . . .	35
5.2.4	Functional Requirements . . . . .	36
5.2.5	Non-Functional Requirements . . . . .	37
5.2.6	User Goals . . . . .	38
<b>6</b>	<b>Design</b>	<b>39</b>
6.1	Brainstorm and Sketching . . . . .	39
6.1.1	Participants . . . . .	39
6.1.2	Materials . . . . .	39
6.1.3	Procedure . . . . .	39
6.1.4	Analysis . . . . .	40
6.1.5	Results . . . . .	40
6.1.6	Final Sketches for Prototyping . . . . .	40
6.2	Low-Fidelity Prototypes . . . . .	42

---

6.2.1	Usage of Colours . . . . .	42
6.2.2	Prevention of Information Overload . . . . .	42
6.2.3	Flexibility for Users . . . . .	43
6.3	Focus Groups . . . . .	44
6.3.1	Participants . . . . .	44
6.3.2	Materials . . . . .	44
6.3.3	Procedure . . . . .	44
6.3.4	Analysis . . . . .	45
6.3.5	Results . . . . .	45
6.3.6	Final Dashboard Compositions . . . . .	47
6.3.7	Final Visual Explanation Design . . . . .	47
6.4	High-Fidelity Prototype . . . . .	50
6.4.1	Home Screen . . . . .	50
6.4.2	Detail Screen: Overview . . . . .	52
6.4.3	Detail Screen: Ship Characteristics . . . . .	53
6.4.4	Detail Screen: Explanation . . . . .	54
6.4.5	Detail Screen: Similar Ships . . . . .	55
6.4.6	Detail Screen: Company . . . . .	56
6.4.7	Detail Screen: Notes . . . . .	57
6.4.8	Detail Screen: Distribution Charts . . . . .	58
6.4.9	Help screen . . . . .	59
<b>7</b>	<b>Evaluation</b>	<b>60</b>
7.1	Methodology . . . . .	60
7.1.1	Participants . . . . .	60
7.1.2	Materials . . . . .	61
7.1.2.1	Usability . . . . .	61
7.1.2.2	Trust and Reliance . . . . .	61
7.1.2.3	Understandability . . . . .	61
7.1.2.4	Task-based Analysis . . . . .	62
7.1.3	Procedure . . . . .	62
7.1.4	Analysis . . . . .	63
7.2	Results . . . . .	64
7.2.1	General Trust in Technology Scale . . . . .	64
7.2.2	System Usability Scale . . . . .	64
7.2.3	Understandability . . . . .	65
7.2.4	Reliance . . . . .	65
7.2.5	Findings from Observations and Thinkaloud . . . . .	65
<b>8</b>	<b>Discussion</b>	<b>69</b>
8.1	Understandability . . . . .	69
8.2	Usability . . . . .	70
8.3	Trust in Technology . . . . .	70

---

8.4	Reliance . . . . .	71
8.5	Visualisation of Model Explanations . . . . .	71
8.6	Research Approach . . . . .	73
8.7	Limitations . . . . .	73
8.8	Recommendations for Future Research . . . . .	74
8.9	Implications for the IDLab . . . . .	74
<b>9</b>	<b>Conclusion</b>	<b>75</b>
9.1	Main findings . . . . .	75
<b>A</b>	<b>Inspector Interview Questions</b>	<b>82</b>
A.1	Mapping the general process steps . . . . .	82
A.1.1	Expectation of the process, based on prior meetings . . . . .	82
A.1.2	Questions . . . . .	82
A.2	Diving deeper into each process step . . . . .	82
A.2.1	Signaling . . . . .	82
A.2.2	Investigation . . . . .	83
A.2.3	Visiting . . . . .	83
A.2.4	Monitoring . . . . .	83
A.2.5	Closing . . . . .	83
<b>B</b>	<b>Brainstorm Results: Thoughts on User Goals</b>	<b>84</b>
<b>C</b>	<b>Sketching Results</b>	<b>86</b>
<b>D</b>	<b>Low-Fidelity Prototypes</b>	<b>93</b>
<b>E</b>	<b>Focus Group Results</b>	<b>100</b>
E.1	Focus Group 1 (ID-lab members) . . . . .	100
E.2	Focus Group 2 (HCI Master students) . . . . .	107
<b>F</b>	<b>System Usability Scale</b>	<b>111</b>
<b>G</b>	<b>Propensity to Trust Technology Scale</b>	<b>112</b>

# List of Abbreviations

---

<b>Abbreviation</b>	<b>Description</b>
AI	Artificial Intelligence
GDPR	General Data Protection Regulation of the European Union
HCI	Human-Computer Interaction
IDLab	Innovation and Data-lab department within the ILT
ILT	Human Environment and Transport Inspectorate (In Dutch: "Inspectie Leefomgeving en Transport")
IV	Information Visualisation
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
SHAP	Shapley Additive Explanations
VA	Visual Analytics
XAI	Explainable Artificial Intelligence

---

## Chapter 1

# Introduction

### 1.1 Motivation

Ocean-going vessels have an average lifespan of 30 to 40 years (Sarraf et al., 2010). When these ships reach end-of-life, they have to be dismantled. Preferably, this takes place in countries in Europe or the United States, where regulations on dismantling make sure that it is carried out in a safe and environment-friendly way. However, currently most ships end up in South Asia, where they are dismantled much cheaper, but under inhumane and hazardous conditions. This practice is known as 'beaching', and happens mostly on three beaches in India, Bangladesh and Pakistan (NGO Shipbreaking Platform, 2020). Here, ships are dismantled barefoot and without proper protective clothing, resulting in numerous (fatal) incidents (Claeys & Bisschop, 2018; NGO Shipbreaking Platform, 2020). Also, it has great ecological impact, as heavy metals such as cadmium, lead, mercury, arsenic, chromium and nickel leak into the ocean (Litehauz, 2015). Moreover, beaching constitutes unfair competition for shipyards that dismantle ships in an environmentally responsible way (i.e. more expensive), turning an ecological advantage into an economic disadvantage (Claeys & Bisschop, 2018).

The decision to send a ship for disassembly to a less developed country is purely economically motivated (Stopford, 2008). Ship owners evaluate expected future earning potential and the expected cost of keeping a ship in operation against the revenue obtained when the vessel is sold for scrap. As the unhealthy and unsafe dismantling methods in South Asian yards are a lot cheaper, these yards are able to offer much higher prices for ships (Demaria, 2010). The process of selling a ship to a shipbreaker depends mostly on scrap dealers known as 'cash buyers', who function as middle men. They operate from London, Dubai, Singapore and Hamburg, and in contrast to traditional ship brokers, become ship owners themselves (Demaria, 2010). As the flag of a ship determines under whose regulatory control it falls, the cash buyers change the flag of a ship to a country that has less restrictive regulations on ship recycling; a so-called 'flag of convenience'. The cash buyers then bring the vessel to their final destination, bypassing liabilities and regulations of the country where the ship owner is located. As the ships are often registered through post box companies, it is very difficult for authorities to trace cash buyers (NGO Shipbreaking Platform, 2019, 2020).

Currently, regulations aiming to restrain the destructive beaching business exist,



but it is still very easy for ship owners to circumvent these laws. For example, the EU have developed maritime regulations (Waste Shipment Regulation, 2006), stating that no transport of hazardous waste is allowed between European ports and countries who are not a member of the Organisation for Economic Co-operation and Development (OECD). However, in practice, since it only applies to ships carrying a European flag, simply reflagging to a non-European flag exempts them from this legislation (Claeys & Bisschop, 2018). Different regulations make it theoretically possible for enforcement authorities to detain a ship if it departs for a non-OECD country. Nonetheless, in most cases there is not enough evidence and cash buyers make this more difficult by stopping at a non-European port before sailing to its beaching destination (Argüello Moncayo, 2016).

Although difficult, it is possible to take enforcement action if proven that a shipping company intentionally sold a ship to a beaching company. However, as this does not happen in most of the cases, shipping owners are apparently willing to take the risk. Also, focusing on enforcing regulations does not directly solve the problem, because the beaching has already happened. To combat this, more effort should be put in the prevention phase, before a ship has the chance to depart to a beaching yard. As there were more than 50.000 active ocean-going vessels in 2021 already (UNCTAD, 2021), manually monitoring every ocean-going ship for its chances of violation is an impossibly time-consuming task.

Artificial Intelligence (AI) techniques have a great potential for solving these types of problems, as they have the power to process large sets of data and are able to exceed human performance for particular tasks. Also, it can be used as a means of justification in decision-making situations. It allows for decision-making backed by factual data; instead of subjective rationales only. In this context, machine learning (ML) methods are particularly valuable. They are generally used for automated detection of meaningful patterns in data (Osisanwo et al., 2017). By finding patterns in typical beached ships, such methods are expected to be supportive in the prevention of illegal shipbreaking. However, the impact of decisions based on such models is large and relying solely on the output of ML systems only would not be justifiable. Therefore, the ML system should be able to explain its behaviour, which opens the doors to the field of Explainable Artificial Intelligence (XAI): ML algorithms revealing the paths to the output they produce.

Along with the use of XAI, there is an increasing relevance for designing effective interfaces for visualising XAI model explanations, which is related to the field of Human-Computer Interaction (HCI). One of the key elements in HCI design is the focus on the target user of a system, as design is greatly affected by the audience it was designed for. Existing research on visualising XAI focuses mostly on supporting model developers in understanding their models. However, visualising for the actual end-users of XAI models, referring to the people who use XAI models to support their decision-making, is still under-represented in literature.

## 1.2 Problem Statement

The Human Environment and Transport Inspectorate (In Dutch: "Inspectie Leefomgeving en Transport", ILT) has the responsibility of improving safety, confidence and sustainability in regard to transport, infrastructure, environment and housing (Ministerie van Infrastructuur en Waterstaat, 2022). In The Netherlands, they are responsible for countering illegal shipbreaking and are dealing with similar problems regarding prevention of these practices. The Innovation and Data-lab (IDLab) is a department of the ILT, which aims to increase the use of data-driven AI solutions for supporting inspection tasks (Digitale Overheid, 2020). In the battle against illegal shipbreaking, the IDLab has developed the *shipbreaking* and *beaching* model, ML models aiming to predict whether a ship is likely to be dismantled and beached in the future. This allows inspectors to contact the owners of a ship to inform them about the regulations and required permits on ship recycling. Additionally, it alerts ship owners that they are being monitored by the inspectorate. This is expected to positively influence the behaviour of ship owners and to result in less ships being beached in South Asia.

As the models' predictions will be used and valued by human inspectors, it is important for the models to be transparent. Generally, machine learning models follow the "black box" analogy, referring to the idea that models take inputs and provide outputs without producing interpretable information about its decisions (McGovern et al., 2019). In contrast, explainability in AI focuses on making AI models understandable, trustworthy and manageable (Goebel et al., 2018). First of all, it enables developers to enhance the model's robustness and to prevent bias, unfairness and discrimination (Confalonieri et al., 2020). Secondly, model explanations assist the end-user who uses the model's output for decision-making by gaining their trust in the system (Confalonieri et al., 2020). Because the inspectorate using the model is a government agency, it is even more important that decisions based on the model can be properly substantiated (Confalonieri et al., 2020).

The *shipbreaking* and *beaching* models were trained using a large dataset including ships that have been dismantled and ships that have been beached, based on records from the NGO Shipbreaking Platform and supplemented with data scraped from the web. Some of the data features that were collected included a unique ship identifier (IMO-number), the type of ship, flagging history, build year, information about ship size, and some technical details about the engines. Through supervised learning, the models were trained to identify typical dismantled and beached ships.

The predictions of the models include a prediction score for each ship, together with an explanation of this score. These local explanations follow the SHapley Additive exPlanations (SHAP) approach, which is a framework for interpreting predictions of machine learning models. In each prediction, features are assigned a SHAP value, which are calculated by comparing the prediction score with and without a certain feature to a baseline prediction score. These values express to what degree and in which direction (increase or decrease in score) a feature influenced the prediction score (S. M. Lundberg & Lee, 2017).

The ILT shipbreaking project is currently in the pilot phase. The goal is to experiment on how the models perform in the field, and how it can be implemented in the workflow of the inspectors. A problem that remains unsolved, and becomes increasingly important as the project progresses, entails the interaction between the inspectors and the output of the models. This HCI related problem exemplifies a larger question on how humans interact with AI. It concerns aspects such as visualisation and interaction, with the aim to make the predictive machine learning model accessible, understandable and trustworthy for the decision-making users. The current research focuses on designing visualisations specifically for the inspectors, as they are the end-users in this case. This type of user can be seen as an expert in the domain, but a novice in the field of AI and data science. They have different motives for using an XAI system compared to other users, such as the developers of a model. As existing XAI visualisation studies have a strong focus on this last user type, the current research contributes to getting a better understanding of how to effectively design XAI visualisations for end-users.

### 1.3 Research Question

The primary goal of this thesis is to design an interface to be used by the ship inspectors of the ILT which helps them understand, interact and gather valuable insights from the data that is produced by the ML models. In an earlier study by Haas (2021), existing industry standard SHAP-explanations for visualising the *beaching model* were tested and compared. It was concluded that these types of visualisations are understandable and useful for data scientists, but not suitable for inspectors. As the inspectors often lack data science skills, it is expected that they have difficulties understanding the detailed and technical visualisations.

The current study takes a user centered approach and aims to design an interface that prioritises usability, trust, reliance and understandability for the intended end-user. For this, the following research question was formulated:

*"How can data visualisations of XAI models effectively support ship inspectors with preventing illegal shipbreaking?"*

## Chapter 2

# Literature Review

### 2.1 Explainable Artificial Intelligence

This section provides a background on XAI, which is one of the main topics of this study.

#### 2.1.1 Definition

The term *explainability* or *Explainable AI* is in most studies used to describe the general ability of understanding the behaviour of machine learning model (Bhatt et al., 2020; Samek et al., 2019). Explanations can have different goals, for example through explaining people why a system rejected an application for a loan, or by providing explanations to understand sales predictions in terms of customer up-sell (Gade et al., 2019). The term is sometimes used interchangeably with *interpretability* (Du et al., 2019). However, in a survey by Mohseni et al. (2021), a strong differentiation is made between *Explainable AI* and *Interpretable AI* as parts within the overarching concept of *Transparent AI*. According to this, interpretability refers to models that are inherently human-interpretable (or *intrinsic*, see 2.2.3) and explainability refers to the ability to explain underlying black-box models with accurate and comprehensible explanations (or *post-hoc*, see 2.2.3). Furthermore, Mohseni et al. (2021) define the Explainable AI system, serving as a connector between the end-user and a model's data, as a cooperation between a machine learning model and an explainable interface. The current study focuses on the design of the explainable interface, which is an essential and underexposed aspect in the research field of Explainable Artificial Intelligence.

#### 2.1.2 The Need for XAI

As the application of AI for decision-making increases, along grows the demand for transparent AI models. Ming (2017) defines three main aspects for the need of explainability: humans' curiosity about knowledge, limitations of current intelligent algorithms, and moral and legal issues.

- **Curiosity of the human.** Algorithms are applied for generating knowledge, and humans are generally curious about new knowledge. Especially when AI achieves great accuracy, humans want to know how this was achieved. At the same time insight into the behaviour of a model allows for improving the performance of the model.

- **Limitations of machines.** As machines and therefore algorithms are imperfect, human knowledge and context are required to complement them, especially when AI models are used for real-world decision-making. Providing explanations allows for establishing trust in a system. Additionally, for developers of a model, understanding its behaviour allows for detecting and understanding possible bias in the data or model.
- **Moral and legal issues.** According to the General Data Protection Regulation (GDPR) of the European Union, humans have a "right to explanation". This regulation has raised debates on to which extent automatic decision-making systems should provide explanations about its decisions. A concrete example of why this regulation is important, is that people who are turned down a loan because of AI decision-making should be able to get an explanation of those decisions. The regulation is mainly driven by the fear for discrimination of AI models, which may be easily neglected during the development phase and could potentially lead to unfair treatment of humans.

### 2.1.3 Explainable Artificial Intelligence in Human Computer Interaction

Research on XAI is multidisciplinary and priorities drastically differ per domain (Mohseni et al., 2021). In the domain of ML, research focuses on developing new interpretable models, while research in visual analytics design focuses more on designing tools and methods for visualising complex models for domain experts. In the field of HCI, research focuses on the needs of the end-user such as trustworthiness in a system and understandability of the explanations.

## 2.2 Types of Explanation

ML models can be explained in various ways. The most salient differences that are relevant for this study are listed below.

### 2.2.1 Global versus Local Explanations

Explanations of ML models can be broadly categorised in two main types: global and local. Global explanations describe the overall working of machine learning models and provide a global understanding of the knowledge that has been acquired by the trained models. They do not only help to understand a model's predictions, but are also important for ascertaining trust in the model. Especially during the development phase, as a certain confidence about a model's real-world performance is often needed before deciding to deploy it in the field (Ribeiro et al., 2016a). In contrast, local explanations describe the behaviour of a model for individual predictions. These types of explanations are thought to be more accessible for users that are less familiar with data science or artificial intelligence (Mohseni et al., 2021). Understanding individual model predictions are essential to ascertain trust in a specific prediction.

### 2.2.2 Model-Specific versus Model-Agnostic Methods

Model-specific explanation tools are limited to specific classes of ML models. For example, interpreting regression weights in a linear model is model-specific to regression

models, as regression scores only apply to such models. On the other hand, model-agnostic tools can be used on any machine learning model and are applied after a model has been trained, usually by analysing a model's feature input and output (Molnar, 2019).

### 2.2.3 Intrinsic Models versus Post-Hoc Methods

The degree to which ML models can be interpreted by humans are inversely proportional to the size and complexity of the model. Simpler models such as decision trees and linear regression models are easier to understand but have limited performance on high-dimensional data. Complex models such as the random forest model (collection of hundreds of decision trees) offer high performance but are difficult to interpret (Mohseni et al., 2021). Intrinsically interpretable models are model-specific ML models that are considered interpretable due to their simple structure. Post-hoc methods refer to the application of interpretation methods after model training. These model-agnostic methods separate the explanation from the model, meaning that they can be used on any ML model and are especially useful for complex models. Also, the decoupling allows for comparing different models through a single post-hoc explanation method (Molnar, 2019).

### 2.2.4 Local Explainability: Feature Importance

From the different types of explainability, local explainability can be relevant for end-users of a system, as it allows them for assessing individual predictions. Various different local explainability techniques exist, of which *feature importance* is studied and applied most (Bhatt et al., 2020). This is because most traditional ML models rely on feature engineering, in which raw data is transformed into features. Such models can be explained through feature importance, which indicates the contribution of each feature to the underlying model (Du et al., 2019). This helps to understand how each prediction was made and to what extent each feature affected the prediction. Different methods for calculating feature importance exist, but they have in common that the output consists of a set of importance scores for each feature. The importance is expressed as a number between -1 and 1, representing the strength of attribution (0 = no attribution, 1 or -1 = maximal attribution) and its direction (1 = positive, -1 = negative). The methods mainly differ from each other in the way they calculate these scores. For the scope of this visualisation research, mathematical details of the explainability algorithms will not be elaborated. Focus lies on understanding what the output represents.

## 2.3 Visual Analytics

This section aims to provide a background on the research field of *Visual Analytics* (VA), including a definition and important aspects that need to be considered when designing a visualisation tool.

### 2.3.1 Definitions

VA can be loosely defined as the use of interactive user interfaces with advanced visualisation modules, aiming to provide in-depth understanding of how an underlying

model works (Choo & Liu, 2018; Hohman et al., 2019; Spinner et al., 2019). Keim et al. (2008) describe VA as "more than just visualisation. It can rather be seen as an integral approach to decision-making, combining visualisation, human factors and data analysis." The application of VA is often seen as mainly for *developers* of a model, in order to support understanding and improvement of the models they build (Hohman et al., 2019; Mohseni et al., 2021). For end-users of a model, research refers to *Information Visualisation (IV)*, which is seen as a simpler form of visualising data. While IV focuses on how to visualise data, VA focuses on providing tweakable visualisations giving the user more power and responsibility for gaining intelligence from data (Keim et al., 2008). Opposed to this viewpoint, the current study aims to broaden the definition of VA by bridging the gap between VA and the end-user through discovering how VA tools can be designed for this target audience.

### 2.3.2 Visualisation Process

In order to systematically approach the development of a visualisation, Munzner (2009) proposes a framework to guide and validate the creation of visualisation systems. The nested model consists of four levels: domain, abstraction, idiom and algorithm.

- **Domain situation.** In the first level, the domain situation should be characterised by the visualisation designer. This involves getting a clear understanding of the target audience, their problems and their data. This can be achieved through requirements analysis techniques and involves actual engagement with any target user of the system. Understanding the domain is essential for designing relevant visualisations, it does not matter how clever a visualisation may be if it does not fit the domain it is used in. The output of this level is often a detailed set of questions that the target users have on the data (in the language of the target user). Section 2.3.3 goes more in-depth about different types of target users and their goals.
- **Abstraction.** In the abstraction level, the results of the domain level are translated to the vocabulary of visualisation. This involves rephrasing the requirements to tasks, and transforming the raw data into data types that can be addressed by visualisation techniques. The output of this level is a description of operations and data types, which are the input required for making visual encoding decisions in the next level.
- **Idiom.** The idiom level addresses the problem of how to visualise the data, in two ways: how to draw (visual encoding) and how to manipulate (interaction). This depends on the output of the earlier levels; the target audience, what the audience wants to achieve and what the data looks like. An overview of common visualisations is given in Section 2.4.
- **Algorithm.** Eventually, the system should be implemented; an algorithm should be developed to carry out the visual encoding and interaction designs automatically. As the current study only takes some first steps from IV to VA, this level is only slightly applicable in terms of algorithm development. Still, the principle of

this level is about implementing a system and programming the visualisations, which is also relevant in this study.

### 2.3.3 Types of Users

How explanations and visualisations should be designed, depends strongly on the person who will be using it. According to Preece et al. (2018), there are four types of target users when designing for XAI. Additionally, Mohseni et al. (2021) state that, depending on the type of user, there are different priorities and goals when using a visualisation tool. Also, different types of users have different expertises. In line with the outer layer of the nested model of Munzner (2009) described in section 2.3.2, these studies endorse the importance of defining the target user in order to be able to design relevant and useful XAI tools. The different types can be defined as follows.

- **Developers.** People that are responsible for building the XAI applications. They are the ones that develop ML models and are skilled in the domain of data science. Developers use explainability mainly for quality assurance of the model, i.e. to support model testing and comparison, debugging, evaluation and to improve robustness of their models. Industry-standard explanation and visualisation tools such as LIME (Ribeiro et al., 2016b) or SHAP (S. M. Lundberg & Lee, 2017) are often used to achieve this.
- **Theorists.** AI experts who are concerned with advancing AI theory, mostly active in academic or industrial research units. They are designers of ML algorithms and interpretability techniques for XAI systems. Explainability helps them to better understand the workings of ML models or neural networks. It is also used for model debugging, in order to improve the architecture and training process of models.
- **Ethicists.** People that are interested with fairness, accountability and transparency of AI systems. These can be both technically skilled people as well as AI novices. Developers can also have the role of ethicist; they will apply explainability techniques in order to evaluate and improve the model on the aforementioned matters. The other type of ethicist has motives for XAI that go beyond the quality of software. These can be lawyers, journalists, politicians, or anyone who is involved in ethical usage of machine learning.
- **Users.** The people who use (X)AI systems in daily life but have very little expertise on ML. This group is generally less interested in the technical background of a system, but mainly needs explanations to understand how a complex model works for other reasons. In order to make real-world decisions based on the output of a machine learning model, its explanations provide the necessary background to be able to judge a model's decision and assess the reliability of the system. Also, a better understanding of the model's output can improve the user experience of XAI systems.

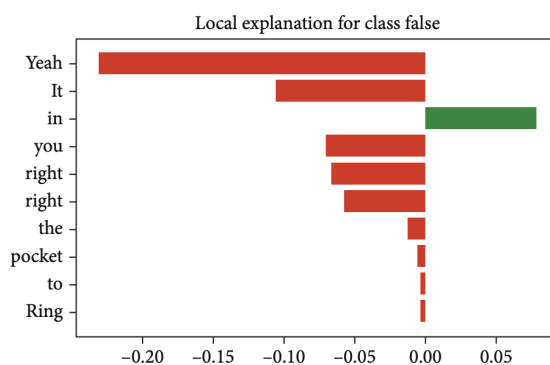


## 2.4 Visualising Feature Importance

There are multiple ways to visualise feature importance. Within these methods, the level of detail of the information is generally inversely proportional to the understandability. Some methods are easy to understand, but lack detailed information, others which may contain lots of detailed information are rather difficult to comprehend. The best methods should be chosen based on the skills of the target user, the type of data and the desired perspective on the explanation. The most commonly used or otherwise relevant types of visualisation are elaborated below. In practice, visualisation types can also be combined or enhanced for specific use cases.

### 2.4.1 Bar Charts

Bar charts are a common used and multi-purpose type of visualisation, expressing a numeric value through the size of a bar. Colour variations can be used to convey additional information, such as the direction (positive/negative) of a value (see Figure 2.1). Bar charts are often easy to comprehend, but are limited to the level of details they can express. Examples of feature importance bar charts can be found in S. Lundberg (2019) and Kumar et al. (2021).



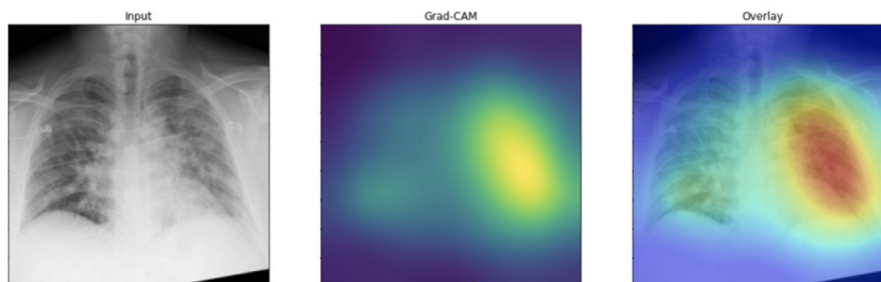
**Figure 2.1:** Example of a bar chart providing explanation for feature importance in a text classifier (Kumar et al., 2021)

### 2.4.2 Heatmaps

Heatmaps are a way of representing data values through a spectrum of colours. They are often used in explaining image classifications by highlighting the degree in which parts of an image attributed to a prediction or classification (see Figure 2.2). Also, heat maps appear in other forms, such as matrices, tables, or any form of colour spectrum usage for visualising data values. Examples of heatmaps can be found in Gohel et al. (2021), Ming et al. (2017), and Wehbe et al. (2021)

### 2.4.3 Scatter Plots

Scatter plots use dots to represent values across two dimensions. They can provide insight into the relation between two variables, and can be helpful seeing data points in context. Also, they can provide visual aid in finding clusters and outliers to these clusters. When visualising feature importance, a scatter plot could be used to find the relationship between a single feature and the model prediction, in order to understand



**Figure 2.2:** Example of a heat map that represents feature importance on an image classification model (Wehbe et al., 2021)

how a feature globally behaves. Additionally, the colour of the data points can be used to express a third variable. Depending on the selected axes, scatter plots can be moderately easy to comprehend and are able to provide a different perspective on the data compared to other types of visualisation. An example can be found in Dave et al. (2020).

#### 2.4.4 Tables

Although mostly not as visually appealing as other visualisations, tables can be useful for displaying multiple records with lots of flexibility for the user. Records can be sorted and filtered by row values or column names according to the user's or system's preferences. Combining this with feature importance, records could be ranked in a table with model features as columns, giving a quick insight into the model's predictions and record data. However, for larger datasets, tables can become too big and therefore unclear and difficult to comprehend. In such cases, filtering would be necessary (e.g. only show records with highest prediction score, or only show features with highest attribution globally), but this reduces the informative value. An example of a table can be found in Kulesza et al. (2015).

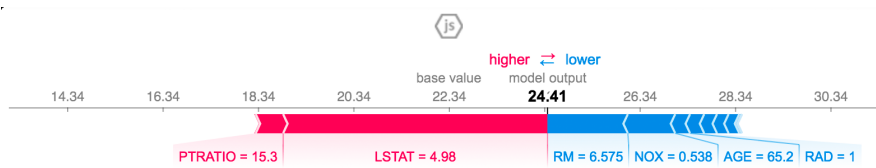
#### 2.4.5 Word Clouds & Word Highlighting

When a model involves the analysis of textual sentences, a word cloud can be used to explain which words attributed to a decision, in which the font size expresses the amount of attribution and the text color the direction of the attribution (positive or negative). A word cloud does not give detailed information, but is simple to understand thus can be used to give a quick insight. When it is useful to know the context of the words, designers may choose to display the original text (partially) and highlight the words with a high feature importance. Examples of this can be found in Kulesza et al. (2015) and Kumar et al. (2021).

#### 2.4.6 Force Plots

The force plot is one of the industry-standard visualisations that is provided by the SHAP package and was specifically designed for visualising feature importance values (S. Lundberg, 2019; S. M. Lundberg et al., 2018). It can be used to visualise feature importance for a single prediction (see Figure 2.3), but they can also be combined to globally express the importance of a single feature in all known predictions. The force plot

and most alternative SHAP visualisations (e.g. waterfall plot and decision plot) are generally hard to understand for non-AI experts (Haas, 2021), and are rarely used in XAI systems for end-users.



**Figure 2.3:** Example of a force plot from the Python SHAP package (S. Lundberg, 2019)

### 2.4.7 Dashboards

Generally, designers do not settle for a single visualisation but rather choose for a combination of multiple visuals to highlight different parts or perspectives of the data. Often these are combined in a *dashboard*. An example can be found in a text classification XAI design study by Kulesza et al. (2015), which resulted in a dashboard with at least eight different types of visualisation. The interface consisted of a combination of a *table view* for selecting individual records, a *word cloud* and *bar chart* to explain decisions locally, supplemented with *text highlighting* to inform which words were used for the decision. Also, there are various visualisations that explain different global model aspects.

## 2.5 Designing for AI Novices

The current study focuses on the last of the aforementioned user types: the end-user. As this group has generally little knowledge in the technical details of a ML model, they prefer simple explanations and representation interfaces. This contrasts with expert users, who seek for detailed information and are able to manage complex interactive visualisations (Mohseni et al., 2021). This illustrates how designing for a specific user type sets a different starting point in the design process. Hence, there are also specific guidelines, principles and goals when designing for an end-user. These are presented in the following section.

### 2.5.1 Design Guidelines

Although there are multiple studies suggesting guidelines for designing XAI interfaces, most of them are targeted at developers of a model. Other guidelines that are developed, focuses on specific use cases that do not match the scope of the current study (Kulesza et al., 2015). However, a survey study by Chromik and Butz (2021) presents a set of four user centered design principles that can be well applied to the user category of non-technical end-users. The principles will serve as design guidelines in the design phase of this study and are expected to improve the design in terms of user experience.

1. **Complementary Naturalness:** as visuals may be difficult to comprehend, especially for non-experts, they should be combined with rationales in natural language. Providing extra information in the language of the end-user could support their understanding of visual cues.

2. **Responsiveness Through Progressive Disclosure:** offer hierarchical or iterative functionalities that can be used in initial explanations. This builds upon the concept of *information overload*, referring to presenting too much information at once. However, too little explanation diminishes the goal of explainable AI. The difficult part of this principle is that user's individual need for information depends on their level of knowledge and skills, thus is different for every user. A good interface provides not too much information at once, but has functionalities to get more information for the more experienced user.
3. **Flexibility Through Multiple Ways to Explain:** offer multiple explanation methods and modalities to improve understanding for different types of people. As every human gains understanding differently, there is often no best way to explain.
4. **Sensitivity to the Mind and Context:** offer functionalities to adjust and personalise explanations, in order to improve the interaction between the user and the system. This principle is more complex, and includes the need for constructing a computer model of the user's mental model to be able to alter the interfaces timing of actions for an optimal interaction. However, for the scope of the current research, the principle is interpreted in a simpler manner: the user should be able to personalise the system to their preferences and context, in order to improve user experience and their understanding of the visualisations.

### 2.5.2 Design Goals

There are four main themes that are mentioned often in the XAI visualisation studies that were reviewed. The first theme is *trust*, which is also one of the main motivations for using explainability in general. Trust is a critical component in AI systems for them to be adopted by users. Meske and Bunde (2020) describe a possible relationship between human and an explanation interface as follows: the AI system is the trustee, the human user the trustor, while the explanation interface is the mediator role between those, aiming to inspire trust in the AI.

Next to this, there is *reliance*. Explaining a machine learning model's decisions allows for end-users to judge those decisions and actually enables them to act upon these decisions. Note that this also relies on the level of trust a user has in the system; without trust in a system, the user will not rely on a system. Otherwise, trusting a system does not mean that users will rely on it in decision making. The term can be split up into two subdimensions. *Under-reliance* refers to users not relying on a system, leading to ignorance to the model entirely if it's output would not fit their expectations. On the other end of the spectrum lies *over-reliance*, referring to users blindly following machine learning decisions. As stated before, machines are imperfect and therefore it is desired to have a *human-in-the-loop* when ML supports real-world decision making. Well-designed and comprehensible explanations help users to understand a model and therefore allows them to actually use them better.

The third theme is *understandability*, referring to the degree of success in which the visualisations convey the explanations to the user. This is influenced by various

factors, such as the choice of visualisation and its interactions, but also the perceived user experience of the XAI interface and how it deals with the problem of *information overload*.

Finally, there is *usability*, which is a common theme in HCI research in general. It refers to users being able to actually use a system and interact with it. Although not the main focus of this research, the system to be designed should be usable 'enough' for the target audience.

These four themes are considered as the main design goals for this research and will play an important role in the design and evaluation phases.

## 2.6 Related Design Studies

In the field of VA, most studies addressing the visualisation of explainable AI models focus on model developers or other expert users as target audience. Therefore, these are considered of limited relevance in the scope of the current study. Nonetheless, some examples of visualisation studies targeting end-users can be found. This chapter describes three studies from different domains aimed at designing interfaces of XAI models. These studies were used to define a set of design themes, representing categories of recurring design elements in XAI dashboards. The resulting six design themes are presented at the end of this chapter.

### 2.6.1 EluciDebug

The study by Kulesza et al., 2015 proposes an interface for the "Explanatory Debugging" approach. Herein, the system gives explanation for its reasoning to the user, who in turn explains corrections back to the system. The main goal of this research is to find out if this approach helps end-users to build better mental models of a system. This is not directly related to the goals of the current research, but the part of the interface that was designed to explain the reasoning of a systems predictions is especially relevant.

The prototype was designed to visualise a text classification model for predicting topics of email texts (Figure 2.4). Most of the interface components are commonly used visualisation techniques. Firstly, component B gives an overview of all instances in a tabular view, which also serves as an instance selector. Secondly, component C uses the 'word highlighting technique' to explain which words in the text are important for the prediction. Component D gives a more detailed explanation on the predicted topic, using word clouds, bar charts and pie charts. The last two components are both local explanations, meaning that their contents are dependent on the selected row in the table. An explanation on global level is given in component E, where the importance of all used words in all predictions are plotted.

The prototype was evaluated in two conditions: a control group using a version without any (visual) explanation versus a treatment group using the full-featured prototype as described above. The evaluation does not specifically discuss every individual component, but it is concluded that participants had no problems with understanding and using the visual explanations; instead, they stated that their visual explanations led to significant increase of understandability and an improved overall user experience.

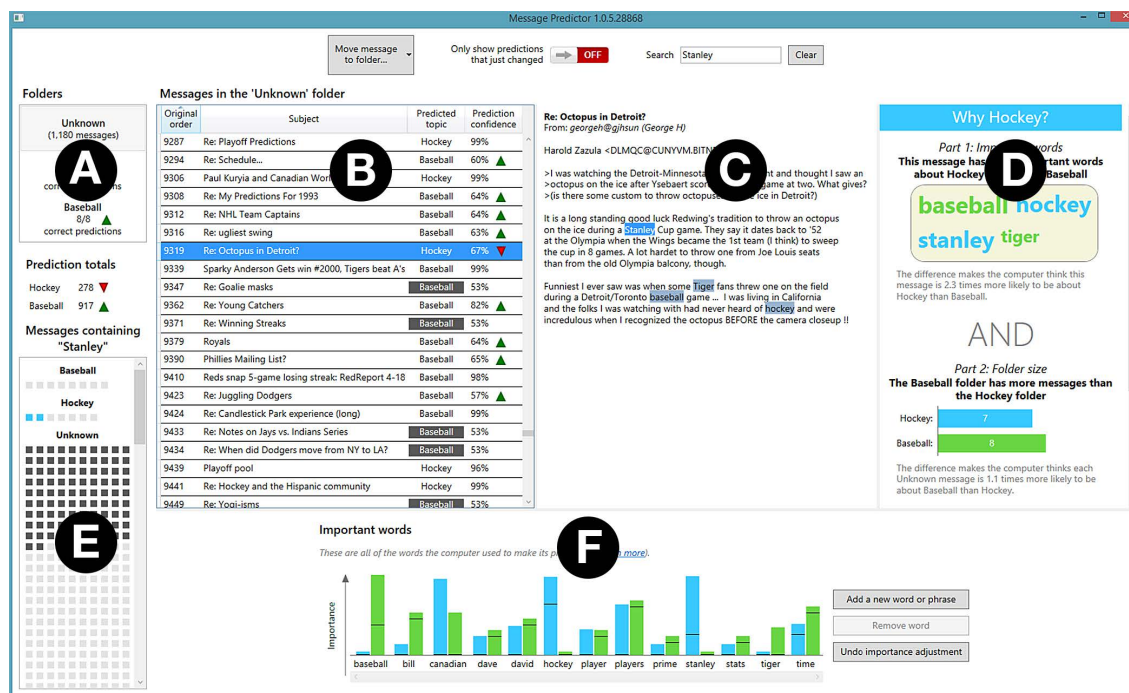


Figure 2.4: The EluciDebug prototype. (Kulesza et al., 2015)

## 2.6.2 XAI Interface for Customer Turnover Predictions

Souza and Leung (2021) researched a case study aiming to design a "user-friendly interface for non-expert users". The interface was designed for an explainable model predicting customer churn. Churn is the rate of customers who stopped using a service or product in a given time frame. The model and explanations are similar to the current research; the studies use the same algorithm for the predictions and they both use Shapley values for the explanation. Additionally, Souza and Leung (2021) use an algorithm to generate *contrastive explanations*. This refers to a set of techniques for explaining the prediction of an individual instance based on what should be done differently for changing the current prediction. The text that is presented is in the form of "The model predicted A instead of B because feature  $x$  is greater than  $\langle number \rangle$  and feature  $y$  is smaller than  $\langle number \rangle$ ." These contrastive explanations are presented as recommendations for the user.

The prototype consists of four screens (Figure 2.5): the home and expected loss (A), local feature importance (B+C), global feature importance (D) and model recommendation (E). On the home screen (A), a summary of all customer predictions is shown, categorised in three risk categories (high, medium and low) with corresponding colours. Selecting one of the categories brings the user to a tabular overview of all the customers within the category, including some information about that user (similar to B). Selecting an instance in the table leads the user to the local feature importance explanation, which is a modified visualisation of the SHAP force plot aimed to be more understandable for a non-expert user. The user can choose to view an alternative local explanation (the contrastive explanation, shown in E) as recommendation for further actions. In

this case, the explanation says that value of the "Growth rate ratio" was the determining feature for predicting a probably of churn for the customer. This information could be used for giving the customer a call to check if he is satisfied with the investment's growth level, aiming to prevent the customer from churning. Finally, the global explanation screen (D) uses a bar chart to give an overview of the importance of all features.

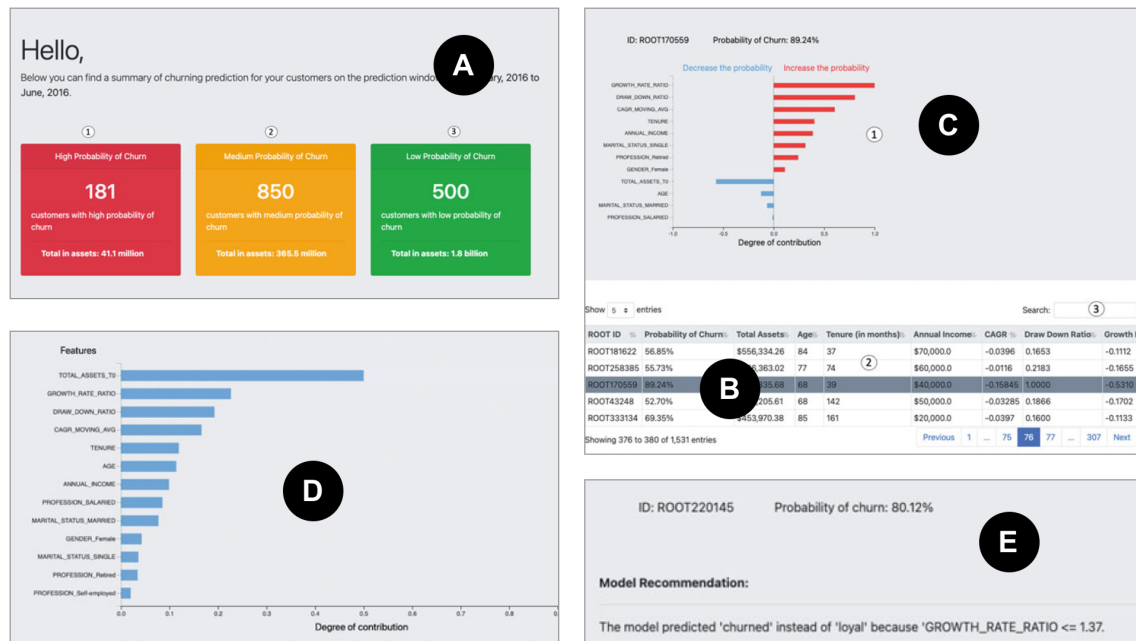


Figure 2.5: The Customer Turnover Interface prototype. (Souza & Leung, 2021)

The study was evaluated by comparing functionalities between their solution and existing ones. The researchers conclude that the solution they propose is the only one ticking all the important boxes: it has global explanation, local explanation, contrastive explanation, a search table, the ability to target specific instances and designed explicitly for a non-expert audience.

### 2.6.3 WEET Interface

A recent thesis study by Ekhart, 2022 proposes an interface design for the Website Evaluation Tool (WEET). This is a tool that uses XAI techniques for detecting bona fide and mala fide factors of webshops, providing support for the Dutch police in their fight against malicious webshops. The interface consists of three screens: the "homepage", "detail page" (Figure 2.6) and "additional information page". The homepage consists of an overview table with all websites that were analysed, including some basic information and the prediction results of the XAI model. The detail page consists of four main components. Component A shows a list of factors that were found, either bona fide or mala fide and serves as the local explanation of the model prediction. Component B and C give an overview of police reports for the selected webshop, using a bar chart and tabular visualisations. In component D, the prediction of the model is presented, after which the user has to draw their own conclusion based on the prediction and explanation. The interface operates strongly on the "human-in-the-loop" principle in which the XAI models give suggestions and arguments, but the user makes the final decision.

There is also an "additional information page" on which the user can find general details of a webshop, such as their Chamber of Commerce registration number.

The prototype was evaluated using a cognitive walkthrough, in which the three participants were asked to carry out small tasks. For each task it was registered if they were carried out following the correct method of completion. Overall results indicated that the interface is understandable and effective in terms of usability. Furthermore, the researcher concludes that existing user interface and user experience heuristics (e.g. Nielsen's heuristics (Nielsen, 1994)) can be effectively applied to XAI interface design.

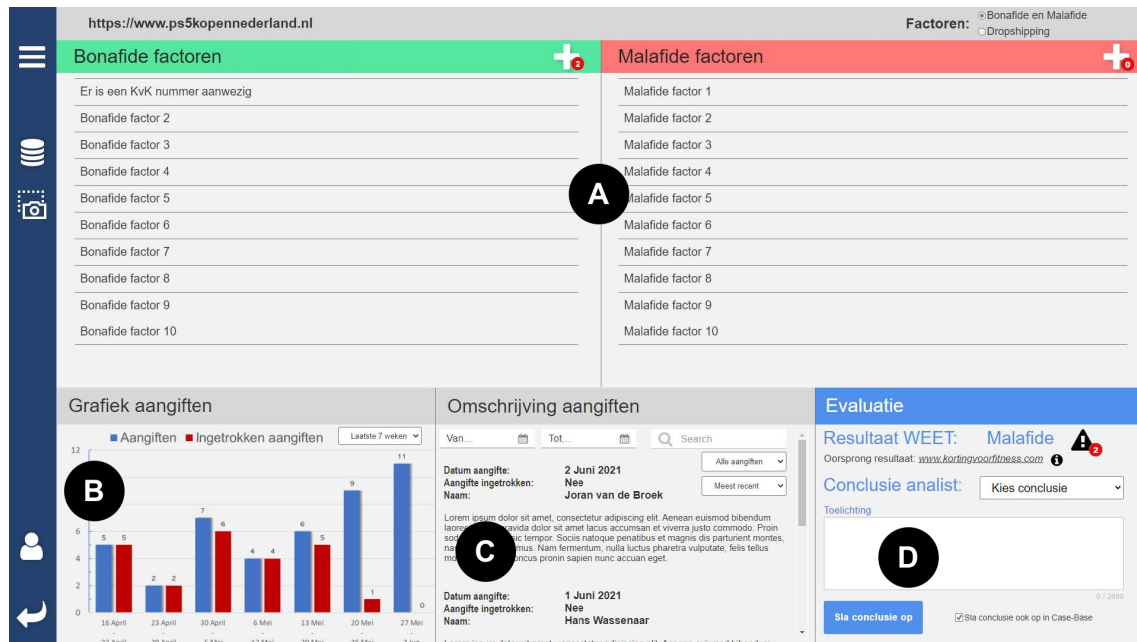


Figure 2.6: The WEET dashboard prototype. (Ekhart, 2022)

### 2.6.4 Common Design Themes

Based on the related design studies and aforementioned information visualisation techniques, six common design themes for XAI interfaces were discovered. Each theme has different goals and uses different visualisation techniques. The purpose of the theme definition is not to prescribe how dashboards should be built, but rather to describe categories of individual interface components.

1. **Dataset Overview:** provides an overview of all the instances, including their predicted outcomes and additional relevant information. By ordering or categorising the instances, the user can find the most important instances that may require their attention. This is in most cases visualised using a table, at the same time functioning as the start page or main instance selector in a dashboard. Additional filtering, sorting and searching functionalities help finding specific instances in the data. All three related studies have this included (figures 2.4 B and 2.5 B).
2. **Local Explanation:** provides an explanation of the model predictions on the level of individual instances. They should provide more information to the user for getting a better understanding of the model prediction. The three related studies



included local explanations in varying forms. They use text highlighting (Figure 2.4 C), word clouds (Figure 2.4 D), feature importance visualisation (Figure 2.5 C), contrastive explanations (Figure 2.5 E) or list views (Figure 2.6 A).

3. **Global Explanation:** provides insight in the the model's behaviour over all predictions. For end-users, knowing how a model globally works may help understand individual predictions better. Two of the related studies include this, both using bar charts for visualising the global feature importances (figures 2.4 F and 2.5 D).
4. **Providing Context:** another way of helping users understand individual predictions better, is by providing context to individual instances within the entire dataset. This theme overlaps with the "Dataset Overview" theme, but the main difference is that the latter has a more top-down approach (use the overview of all instances to find a specific instance), while the current theme is more bottom-up (start at a specific instance and widen the scope to the entire dataset). An example can be seen in the EluciDebug prototype: after searching for a word it presents a list of all messages containing that word including their predicted topic (Figure 2.4 E).
5. **Instance Data:** providing the (additional) data of an instance is often needed for the final decision making. This does not necessarily only include the data that was used in the models for the prediction. For example, in the WEET prototype, users can get an overview of all the data of a specific website, which gives them the necessary information about the website.
6. **Process Support:** this theme is not directly related to the predictions or explanations. It refers to the inclusion of additional components that support the user's work and decision-making. Some concrete examples are shown in the WEET prototype. It presents an overview of the number the police reports (Figure 2.6 B), the contents of these reports (Figure 2.6 C) and a small form for their final decision on taking down the webshop (Figure 2.6 D).

## Chapter 3

# Background

This chapter provides background knowledge on the two models that were used in the current research: the *shipbreaking* and *beaching* model. A brief explanation is given of how the model's data is structured and what the output of the models look like. The inner workings of the models are not extensively discussed, because that is less relevant for the design of the interface. This also implies that if at some point in the future a different algorithm is chosen to produce the prediction scores and its explanations expressed as feature importances, the interface will still be usable.

Name	Type	Description
Age	Numerical	Age of the ship (unit: months)
Type	Categorical	Type of ship
Propulsion	Categorical	Type of propulsion
Gross tonnage	Numerical	Total volume of the ship (unit: gt)
Deadweight	Numerical	Total deadweight (unit: tons)
TEU	Numerical	Cargo capacity for container ships (unit: TEU)
Insulated capacity	Numerical	Capacity of insulated cargo space (unit: m3)
Total length	Numerical	Total length of the ship (unit: meters)
Length between perpendiculars	Numerical	Length of the ship between the perpendiculars (unit: meters)
Service speed	Numerical	Average speed maintained by the ship (unit: knots)
Main engines: model max. age	Numerical	The maximum age of this engine type that can be found in the entire dataset (unit: months)
Main engines: designer max. age	Numerical	The maximum age of this engine designer that can be found in the entire dataset (unit: months)
Main engines: builder code max. age	Numerical	The maximum age of this engine builder code that can be found in the entire dataset (unit: months)
Main engines: amount	Numerical	Total number of main engines (unit: amount)
Main engines: power	Numerical	Maximum power of the main engines (unit: kW)

**Table 3.1:** Overview of predictive features used in the shipbreaking model

Name	Type	Description
Age	Numerical	Age of the ship (unit: months)
Type	Categorical	Type of ship
Gross tonnage	Numerical	Total volume of the ship (unit: gt)
Classification society	Categorical	Classification organisation that certified the ship
Port of registry	Categorical	Place where the ship is registered
Previous flag country	Categorical	Previous registered flag state of ship
Current flag country	Categorical	Current registered flag state of ship
Time since flag swap	Numerical	Amount of time since the final flag change (unit: years)
Population of previous flag country	Numerical	Total population of the previous flag's country (unit: amount)
GDP of previous flag country	Numerical	Total GDP of the previous flag's country (unit: euros)
Population of current flag country	Numerical	Total population of the current flag's country (unit: amount)
GDP of current flag country	Numerical	Total GDP of the current flag's country (unit: euros)

**Table 3.2:** Overview of predictive features used in the beaching model

### 3.1 Data Structure of the Models

The models were trained using a dataset consisting of around 2500 ships having at least one portcall in the Rotterdam harbour, of which the data was scraped from the GISIS information system and merged with open data of the NGO Shipbreaking Platform. Ships were labelled to be either active, (legally) dismantled or (illegally) beached. As the training set is fairly small, it could lead to the detection of patterns that are not representative for the entire population. The number of predictive features in the models was limited to prevent this. As a result, the shipbreaking model is more focused on physical features, such as the length, deadweight and engines. The beaching model mostly uses features regarding registration details, such as the port of registry and flag behaviour. An overview of all features is shown in Tables 3.1 and 3.2.

### 3.2 Model Output and Explanation

The output that is generated is two-fold: the model produces the prediction scores and then a post-hoc method produces an explanation for this prediction. The prediction score represents the likelihood for a ship to be dismantled (shipbreaking model) or beached (beaching model). The values lie between 0 and 1, with a higher score meaning a higher likelihood. Additionally, a final score is calculated by multiplying the score of both models, which indicates the overall risk of a ship. Ships scoring high in both models will have a high final score, as they are likely to be dismantled soon and to be sent to a beach.

The explanation is expressed in SHAP values. For each prediction and each feature in the model, a score is calculated that indicates how much that single feature affected the prediction. Importance values can be either negative or positive. A positive value means that the feature led to a higher prediction score, whereas a negative value means that the feature led to a lower prediction score. An example of a local explanation for a shipbreaking model prediction is shown in Table 3.3. In this example, the ship's *age* had the largest contribution to a higher prediction score, *type* the second one, the *length between perpendiculars* the third, and so on. At the bottom of the list, there are three features that affected the prediction score negatively, thus contributed to a lower prediction score. Generally, high-scoring predictions are expected to have more features with positive values and vice versa.

Feature	Feature importance
Age	0,203
Type	0,127
Length between perpendiculars	0,092
Length overall	0,068
Main engines: Builder code	0,042
Deadweight	0,025
Gross tonnage	0,021
Main engines: power	0,011
Service speed	0,003
Main engines: designer	0,003
Main engines: model	0,002
TEU	0
Insulated capacity	-0,001
Main engines: amount	-0,007
Propulsion	-0,014

**Table 3.3:** Example of local feature importance values in a shipbreaking model prediction.

## Chapter 4

# Methodology

This study follows the nine-stage design study framework from Sedlmair et al. (2012), which builds on the nested model of visualisation development (Munzner, 2009). The framework was developed based on a strong foundation of combining wide-ranging experiences in design studies and extensive literature reviews. This framework is chosen as it gives practical guidance on how to conduct design studies from beginning to end, while other work on methodologies mainly focuses on evaluation of designs only. The methodology consists of the following stages: *learn*, *winnow*, *cast*, *discover*, *design*, *implement*, *deploy*, *reflect* and *write* (Figure 4.1). The stages express a linear process, but can be considered dynamic; stages often overlap and jumping backwards to previous stages is also common in practice. Although not all stages are equally relevant for this research, they will all be covered shortly.

### 4.1 Precondition Phase

The precondition stages *learn*, *winnow* and *cast* focus on the preparation of the research, and establishing and defining useful collaborations with the project's stakeholders.

#### 4.1.1 Learn (Literature Review)

This stage describes the gathering of knowledge of visualisation literature, which serves as necessary background information for the subsequent stages. It includes knowledge

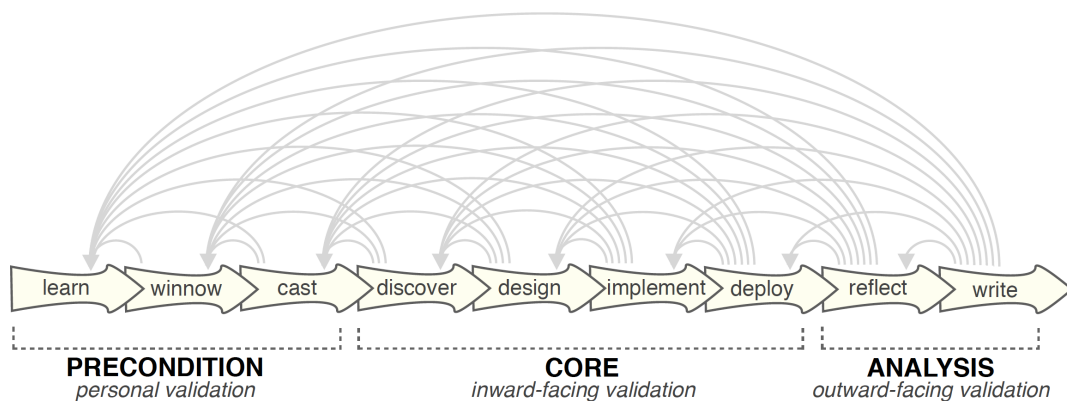


Figure 4.1: Design study framework from Sedlmair et al. (2012)

such as visual encoding and interaction techniques, design guidelines, and evaluation methods. The results of this background research were covered in chapter 2 (Literature Review).

#### 4.1.2 Winnow

The goal of this stage is to identify collaborations for a design study. The framework describes this as a lengthy process of talking to many people and making a careful selection. A set of questions are suggested to be asked in order to identify viability of a collaboration. These are questions regarding the availability of real data, the amount of time and the relevance of the research question. Although the collaboration with the IDLab was already set up prior to starting this research, it satisfies the potential problems that were mentioned in the framework paper.

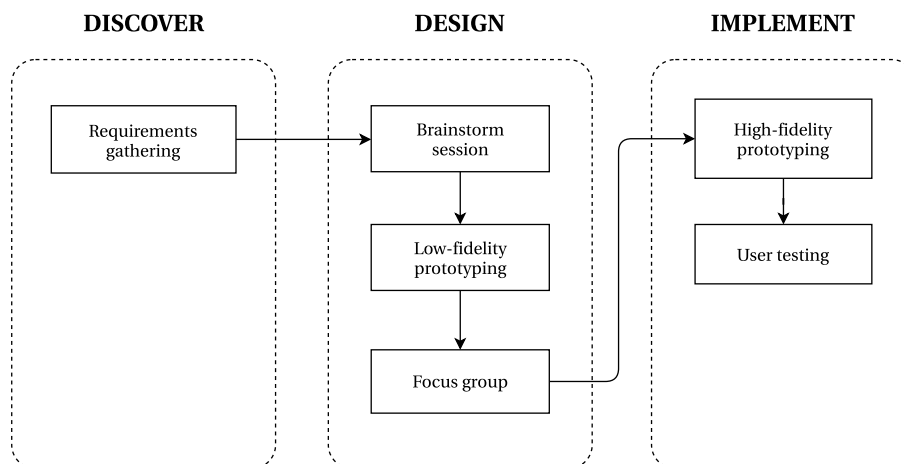
#### 4.1.3 Cast

After selecting collaborations, roles within the project need to be defined. The roles that are suggested in the framework do not fit the type of research of a Master's thesis. However, it is also mentioned that not all design studies have the same roles. Therefore, more general roles common in user centered design literature are used, which were also mentioned in the framework: user, stakeholder and researcher.

- **Researchers** being myself, supported and guided by University supervisors and responsible for conducting the study, leading design sessions, developing prototypes, setting up and conducting experiments and reporting about the results.
- **IDLab team members** as important stakeholders of the project. Data scientists from this team were responsible for gathering the data and developing the machine learning model. They were asked to participate in some of the design sessions.
- **Inspectors** are the domain experts and the targeted end-user of the system. They are involved at the start of the project for gathering requirements and at the end for the evaluation of the interface.
- **Peer students** mimicking users in some of the research phases for which not necessarily domain experts were needed. They were asked to participate in the design phase to assess the design proposals from a HCI perspective.

## 4.2 Core Phase

The core of the design study contains the stages *discover*, *design*, *implement* and *deploy*. An overview of these stage and the particular methods used in the current study can be found in Figure 4.2.



**Figure 4.2:** Design cycle as described by Sedlmair et al. (2012), with particular methods used in the current research. As the current thesis research does not allow for deploying the system in the field, this phase is left out in the diagram.

### 4.2.1 Discover (Requirements Analysis)

The main goal of this stage is to gather and define requirements for the system to be developed. Well-defined requirements are considered an essential foundation for a successful design study. Requirements that fail to characterise the problem well, will most likely result in design ideas that do not offer a suitable solution.

A semi-structured interview was conducted with one of the main shipbreaking inspectors from the ILT, with the goal of gaining better understanding about the domain and the work processes of inspectors. Certain themes and some specific questions were prepared (see Appendix A) and supplemented by questions that arose during the interview. The results of the interview and the final set of requirements are described in chapter 5 (Requirements Analysis).

### 4.2.2 Design

In this stage, the researcher can start with designing a visualisation solution. It is described as the generation and validation of data abstractions, visual encodings and interaction mechanisms. The design cycle is an iterative process, starting with a broad consideration space during the generation of ideas, moving towards a more narrow proposal space after a selecting the best ideas, based on design guidelines and principles. The authors of the framework suggest to create low-level prototypes for the selected ideas, in order to get quick feedback before going into the prototype stage. In the current research, the design phase consists of three substages: a brainstorm and sketching session, the creation of low-fidelity prototypes, and focus groups.

#### **4.2.2.1 Brainstorm and Sketching Session**

In order to generate ideas for the visualisations, a brainstorm session was held. This is a widely used technique for the generation of design ideas, as it allows for a creative environment in which participants can inspire each other (Benyon, 2013). The session was focused on generating ideas on three levels, as described by (Sedlmair et al., 2012): the visual encoding, interaction with the visualisations, and data abstractions. Further details on the methodology and the results are presented in chapter 6.1 (Brainstorm and Sketching).

#### **4.2.2.2 Low-Fidelity Prototypes**

The sketches and ideas from the previous session were transformed into low-fidelity prototypes. They were created on an individual component level which are presented in chapter 6.2 (Low-Fidelity Prototypes).

#### **4.2.2.3 Focus Groups**

The low-fidelity prototypes were used in focus groups, in order to gather feedback on the components regarding understandability and usability, and to find out how the components should be arranged in a dashboard layout. This was done through a discussion and a practical dashboard creation task. Further details on the methodology and the results are presented in chapter 6.3 (Focus Groups).

### **4.2.3 Implement**

After gathering feedback on the low-fidelity prototypes, the next step was to develop a high-fidelity prototype to be evaluated in an experiment with the target users. The high-fidelity prototype is presented as part of the design creation in chapter 6.4 (High-Fidelity Prototype)

### **4.2.4 Deploy**

This stage involves deploying the tool and gathering feedback in the field. Due to the limited time available in this thesis research, this stage is not part of the research. Yet, the results of the study will be presented to the IDLab, who is responsible for the actual field deployment.

## **4.3 Analysis Phase**

The final stages of the framework are *reflect* and *write*.

### **4.3.1 Reflect**

Reflection is an important part of any research. After conducting the design study, a critical reflection should be done about the methods and findings. This includes reflecting on how the study relates to the larger research area, and how previously proposed design guidelines within the field can be improved. In the current research, the reflection is reported in chapters 8 (Discussion) and 9 (Conclusion).

### **4.3.2 Write**

The final stage involves reporting on the study. In most research, this is done through a design study paper. In the current research, this was done in the form of this report.



## Chapter 5

# Requirement Analysis

Requirements are necessary attributes in a system, as they provide direction in the design process and allow for evaluating design ideas. Good requirements are measurable, unambiguous, complete, concise and implementation-free (Lamsweerde, 2009).

### 5.1 Findings from Inspector Interview

The work of the inspectors is rather difficult to generalise, as each case of a (potentially) illegally dismantled ship could be triggered for different reasons and because decision-making often differs per use case. Nevertheless, on a high level the work process can be described in three consecutive phases: *signalling*, *investigating* and *visiting*. In the signalling phase, the inspectors get notified about a ship with a high risk of being beached in the future, or when it already has been beached. Signals can come from many directions, from partners in their network to media reports. The shipbreaking model provides another way of signalling the inspectors. After being signalled about an individual ship, it is investigated by looking at the ship's characteristics, location and history. Regarding the shipbreaking model, this is the phase in which the model explanation are relevant. Based on the signal and further investigation a decision to act follows, which in most cases leads to visiting the shipping company. During a visit, inspectors will mainly inform a shipping company about the existing regulations around ship dismantling, as it is difficult to prove a company's intention to beach. Additionally, it might be necessary for an inspector to give an explanation about the reason for their visit. When this is based on results of shipbreaking model, they want to be able to give a brief explanation about the decision of the model.

Structuring the work process in these three phases is an important aspect of gaining domain understanding. Also, it provides a starting point for the system's requirements, which are presented in the next section.

### 5.2 Final Set of Requirements

The requirements are based on three main sources: the interview with the inspector, multiple meetings with one of the model developers who had also spoken to various shipbreaking inspectors, and guidelines from literature. Firstly, the end-user is described in order to get a better understanding of their reasons and goals for using the system. Secondly, the requirements, categorised as functional and non-functional, are

presented. Functional requirements define specific functionalities that a tool must provide. Non-functional requirements define constraints describing behaviour of the overall system (Lamsweerde, 2009). The functional requirements are grouped by higher-level themes representing the stages that were elicited from the interview. Also, they are prioritised following the MoSCoW-rule: Must have (critical), Should have (important, but not necessary), Could have (desirable, but not necessary), Won't have (not important). The non-functional requirements are grouped by themes regarding system quality or constraints. As in the current research only a prototype will be developed, not all requirements can be satisfied. Therefore, focus lies on implementing the must-haves and should-haves in the most concrete form possible with the data and resources available.

### 5.2.1 Persona

The system will be used by the inspectors of the ILT. These inspectors are responsible for the prevention and punishment of illegal shipbreaking by owners of large ships in the Netherlands. In general, ships are monitored and site visits are made when there are signals that a ship will be illegally dismantled. These signals often come from their network, for example national partners such as the police, border control, or other inspectorates. Also foreign stakeholders or reports in the media may be a signal for the inspectors to further investigate a ship. These signals are officially reported in a system that is used by the ILT.

When an inspector gets signalled, further investigations into the ship are conducted to determine whether it is necessary to visit this shipping company. Currently, the Lloyds List Intelligence database is consulted, which is the most important source of information for the inspectors, with data about the ship such as the flag, owner, current destination, weight and other technical details. There is no clear answer to how a decision is made to proceed with a visit. It is always a combination of factors, depending on the signals they have received about the ship, how strong the evidence is and sometimes the reputation of the ship owner. All in all, if there is a strong suspicion that a ship is going to be beached, an inspector can decide to visit the ship owner. The main purpose of such visits are to inform shipowners about the applicable legislation. The system to be developed plays a key role in providing the information that the inspector needs to make a decision.

In addition to their current way of working, the ML models that were developed aim to provide a predictive risk assessment of ships. In order to understand the predictions of the models, inspectors want to know how these predictions were produced. In terms of skills, inspectors are domain experts in the field of (illegal) shipbreaking, while also being novices in the field of AI and data science.

### 5.2.2 Requirements Sources

Source description	
<b>S1</b>	Interviews with data scientist at IDLab who is mainly responsible for the machine learning models that were created. This person had already spoken to many inspectors and therefore had extensive knowledge of how they work and what their goals are.
<b>S2</b>	Interview with chief inspector who talked on behalf of the other inspectors. Focus lied on the current working process, their information needs and their future goals.
<b>S3</b>	A Revised Set of Usability Heuristics for the Evaluation of Interactive Systems (Pribeanu, 2017)
<b>S4</b>	Design Principles for AI Novices (Chromik & Butz, 2021)

### 5.2.3 Requirement Priorities

Must have	Critical requirements
Should have	Important, but not necessary for the prototype to be validated.
Could have	Desirable, but not necessary. Can be included if time and resources permit.
Won't have	Least-critical

## 5.2.4 Functional Requirements

### Theme 'Signalling'

ID	Requirement	Source	Priority
R1.1	The user should be able to see the model scores for each ship	S1/S2	Must
R1.2	The user should be able to see how the ship scores / compares in context (relative to the other ships)  The user should be able to see at a glance which ships require attention:	S1	Must
R1.3	- when a ship has a high model score	S1/S2	Must
R1.4	- when a ship recently changed its flag	S1/S2	Could
R1.5	- when a ship recently changed its owner	S1/S2	Could

### Theme 'Investigation'

ID	Requirement	Source	Priority
R2.1	The user should be able to browse through all ships that are in the model	S1	Must
R2.2	The user should be able to search for ships based on its name or ship code	S2	Must
R2.3	The user should be able to change ordering of the presented ships by one of their characteristics	S2	Should
R2.4	The user should be able to filter the presented ships by one or more of their characteristics	S2	Should
R2.5	The user should be able to choose which ship characteristics are presented in the overview	S2	Should
R2.6	The user should be able to see when the data of a ship were fetched most recently	S1	Could
R2.7	The user should be able to get more detailed information about the characteristics of a ship	S2	Must
R2.8	The user should be able to see the model (prediction and explanation) history for the ships	S1	Could

### Theme 'Explanation'

ID	Requirement	Source	Priority
R3.1	The user should be able to get an explanation on how a prediction was made	S1/S2	Must

### 5.2.5 Non-Functional Requirements

#### Usability

ID	Requirement	Source
R4.1	The system should guide users toward making specific actions	S3
R4.2	The system should provide appropriate feedback as a response to user's actions within reasonable time	S3
R4.3	The system should provide a clear structure of the application	S3
R4.4	The system should provide means to group similar objects and distinguish between different classes of objects	S3
R4.5	The system should provide similar meanings and design choices in similar contexts	S3
R4.6	The system should provide means to the users' perceptual and cognitive load	S3
R4.7	The system should minimise the number of actions needed to accomplish a task goal	S3
R4.8	The system should ensure that only actions requested by the users are processed and only when these are requested	S3
R4.9	The system should provide the means to initiate and control the system processing	S3
R4.10	The system should provide means to customise the interface and select the preferred way to accomplish a goal	S3/S4
R4.11	The system should provide means to match the users' characteristics with the characteristics of the user interface	S3
R4.12	The system should provide the user with the procedure and associated support (forms, documents, etc.) needed to perform specific tasks.	S3
R4.13	The system should provide means to prevent, diagnose, correct, and recover from errors.	S3
R4.14	The system should provide online help and documentation	S3
R4.15	Visual information should be supplemented with textual alternatives	S4
R4.16	The system should provide an appropriate amount of information, with functionalities to get more details	S4
R4.17	Explanations should be provided through multiple methods and modalities	S4

#### Reliability

ID	Requirement	Source
R5.1	The data should be updated at least every 3 months	S1

**Technical**

---

<b>ID</b>	<b>Requirement</b>	<b>Source</b>
R6.1	The system should preferably be developed in R Shiny	S1
R6.2	The model and ship data should be read from CSV files	S1

---

**5.2.6 User Goals**

As the list of requirements is rather detailed, it was not practical to present these to the participants in the design phase. Therefore, the requirements were mainly used to support decision making on the design proposals. For usage in the hands-on design sessions, the requirements were transformed to a list of main goals for the system. The resulting five goals set the boundaries for the ideas to be generated during the design phase.

- Inspectors want a system that they understand and is easy to use;
- Inspectors want to be able to find the ships that have the potential to being illegally dismantled in the future;
- Inspectors want to understand why the system thinks that a ship has a low or high risk of being illegally dismantled;
- Inspectors want themselves to be able to explain to others why a ship has a low or high risk of being illegally dismantled;
- Inspectors want to be able to see the features of a ship.

## Chapter 6

# Design

The design cycle in this study focuses on the generation of design solutions. First, a brainstorm session was held to generate ideas. The best ideas were used for the low-fidelity prototypes, which were then used to discuss in focus groups with non-domain experts. Details about the intended participants, materials, procedure and analysis are further discussed in this chapter.

### 6.1 Brainstorm and Sketching

The session was focused on generating ideas on three levels, as described by (Sedlmair et al., 2012).

- **Visual encoding:** what type of visualisations, or combinations, can be used?
- **Interaction mechanisms:** how could the user interact with the visualisations?
- **Data abstractions:** which data might be relevant for the user, and how could it be encoded?

#### 6.1.1 Participants

The session was held with the members of the IDLab project group who are responsible for supervising thesis projects around the shipbreaking model. This is a group of four people (N=4, 2 female, 2 male) with varying expertise in data science, psychology and behavioural sciences. Also, the participants had different levels of knowledge about the shipbreaking domain. The variety of the group members was expected to positively affect the generation of ideas, as the participants had different perspectives on the topic.

#### 6.1.2 Materials

During the session, participants were handed post-its and A4 papers for writing or sketching their ideas. The ideas were shared with each other by sticking the post-its on a publicly visible table. Additional comments about ideas were recorded through note-taking, which are discussed in the results section.

#### 6.1.3 Procedure

A few days prior to the session, the participants were provided the necessary background information on the research, including the requirements document and an explanation of what was expected from them during the brainstorm. When starting the session, this information was shortly summarised as a refresher, and the ground rules

of brainstorming session were explained (quantity over quality, all ideas matter, no negativity, wild ideas are encouraged, build on other's ideas).

At the start of the brainstorm session, the four user goals that were defined in the requirements analysis were used to get the discussion going. Each goal was presented, after which the participants were asked to think about what they thought the each goal could imply. This includes thoughts about how they could be achieved, how they could be visualised, or what challenges they pose. All thoughts were written down on post-its and presented to the group, in order to allow people to build on each other's ideas. The main goal of the discussion phase was to make sure that every participant had thoroughly reflected on the goals and had seen multiple perspectives, before starting the sketching session.

After the brainstorm, the sketching session was held. Participants got one minute for sketching one or multiple ideas that fit one or more of the four user goals. Then, everyone presented their sketches and they were shortly discussed. A few of these sketching rounds were held, until the participants ran out of new ideas.

#### **6.1.4 Analysis**

After the brainstorm session, all sketches were grouped by similarity and notes about comments on ideas were structured. After this, the ideas passed a first selection round based on the discussion feedback, additional comments that were made and existing design guidelines. The resulting sketch ideas were used for the low-fidelity prototyping phase.

#### **6.1.5 Results**

The brainstorm discussion on the user goals resulted in various thoughts about functionalities, visualisation ideas and challenges. As they served mostly for creating a shared background for the participants, they are not further discussed here. However, to give an idea about the session contents, the complete list of thoughts and ideas is provided in Appendix B.

After four sketching rounds, a total set of 13 sketches were created. Three sketches were highly similar, involving a scatterplot for giving an overview of all instances. Therefore, the two that were least clearly drawn were discarded, leaving one scatterplot and a total of 11 sketches. A full list of the sketches is provided in Appendix C.

#### **6.1.6 Final Sketches for Prototyping**

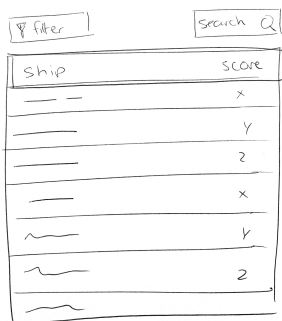
The initial set of sketches were evaluated using the feedback and thoughts that were given during the session. This revealed that some ideas were not feasible (e.g. due to lack of available data), not fitting the target audience (e.g. too much use of charts) or falling outside the scope of the requirements. This resulted in a set of final sketches which were used for creating the low-fidelity visualisations. The final sketches are provided in Figure 6.1.

On top of the sketches, a specific idea came up during the brainstorm session to provide the inspectors information about which ships are similar to other ships. The rationale behind the idea is that it gives the inspectors a different perspective on the

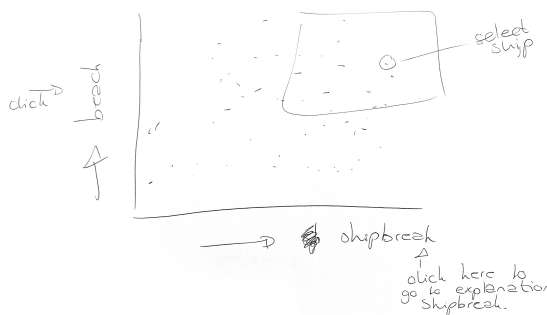


dataset, and it may help them compare prediction scores similar ships to get a better understanding of the underlying model.

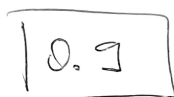
Finally, it was decided to not include global feature importance in the interface, as participants were afraid that inspectors could not comprehend the conceptual differences between global and local feature explanation. As this is in line with findings of Mohseni et al. (2021), it was decided to adopt this standpoint. Various visualisations such as the scatterplot, distribution charts and similar ships page were thought to still give the inspectors a global view on the data, without compromising understandability.



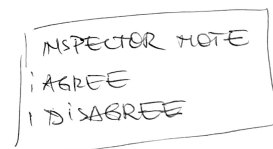
(a) Simple instance selection table, including at least ship name and prediction scores. Supplemented by filtering and searching functionalities.



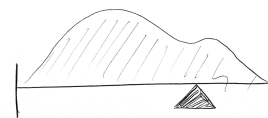
(b) Scatterplot to provide an overview of all ships, with the beaching model and shipbreaking model scores as axes.



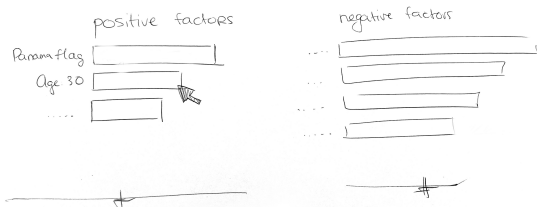
(c) The final score of the selected ship, presented as a score between 0 to 10 instead of 0 to 1.



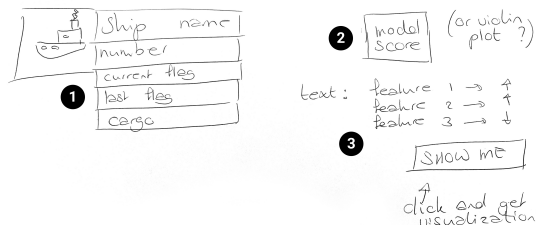
(d) Feature for adding feedback to the system, which could be given to the model developers, or eventually even the model itself.



(e) Chart for visualising the distribution of a ship's feature (e.g. age), providing context through marking the value of the currently selected ship.



(f) Idea for displaying the feature importances as lists with bars. The positive and negative factors are separated, and only the X most important features are initially shown. Clicking the "plus" button expands the list with all features.



(g) Idea for the detail page. (1) Summary of most important ship details, including an image of the ship. (2) Presentation of the model score. (3) A textual list of three most important features, supplemented by a button triggering a visual representation of all the feature importances.

**Figure 6.1:** Overview of the final sketches which are used for low-fidelity prototyping. High resolution versions are provided in Appendix C.

## 6.2 Low-Fidelity Prototypes

Based on the results of the brainstorm and sketching session, low-fidelity visualisations were created. The visualisations were created digitally following the visual identity guidelines of the Rijksoverheid and the IDLab. This was chosen over paper sketches or simple wireframes, as they were expected to lack the detail or clarity that was needed for evaluating the prototypes with non-experts. The visualisations were created on an individual component level without yet deciding on how the components should be arranged in the dashboard view, as this was to be decided with the results from the focus groups in the next phase. Figure 6.2 shows an overview of visualised components parallel to the previously presented sketches in Figure 6.1. The entire overview of components is provided in Appendix D. For the creation of these low-fidelity visualisations, some important design decisions were made, which are discussed below.

### 6.2.1 Usage of Colours

Firstly, the usage of colours is one that stands out. Provided by the IDLab was a colour palette with a variety of options. Although red is the main theme colour of the IDLab, the more neutral blue colour was chosen for general buttons, anchor links or charts. Green and red, or the spectrum in between fading from yellow and orange, was used for visualising numeric values. This makes distinguishing values from each other easier. An example are the prediction scores (Figure 6.2a and 6.2c). Especially in an overview scenario, it helps to easily distinguish instances from each other and attract attention to instances with the highest scores. Furthermore, in the feature importance visualisation (Figure 6.2f), the features leading to a higher prediction score are displayed in red, while the features leading to a lower score are green.

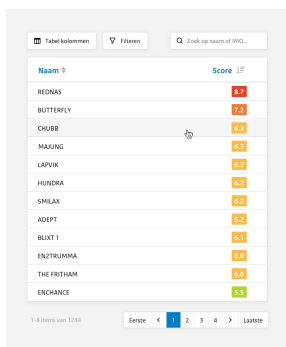
Although good colour usage is essential for enhancing the user experience, it is important that the transmission of information does not solely rely on colours, as this can lead to information loss for people with visual impairments. Therefore, for all information presented through colour, there is also a textual alternative. In the score visualisation (Figure 6.2c), this is done through the score value itself. In the feature importance visualisation (Figure 6.2f), there are titles above the explanations in the form of "Reasons for *[higher/lower]* chance of shipbreaking".

### 6.2.2 Prevention of Information Overload

Other types of design decisions were made regarding the prevention of information overload. A concrete example of this can be found in the number of explanations that are shown. The textual explanation, which is focused on reducing the amount of information, only shows the three most important features and users can click a button to show all details (Figure 6.2h). The visual explanation includes details such as the relationship between the features, but with a maximum of five. Lastly, having a summary of most important ship characteristics (Figure 6.2g) is also a form of information overload prevention; users can navigate to the entire list using the "all details" button.

### 6.2.3 Flexibility for Users

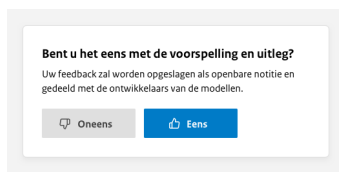
Finally, some design decisions have deliberately not been made, but were converted into flexibility for the user in the form of functionalities for choosing what the interface should look like. An example of this can be found in the overview table, in which the end-user can select any of the ship characteristics as table columns (Figure 6.2a). Another example can be found in the scatterplot, in which the user is able to select which values should be presented on the axes (Figure 6.2b).



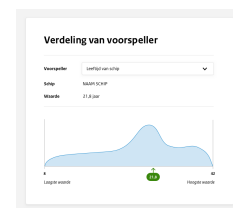
(a) Simple instance selection table, including at least ship name and prediction scores. Supplemented by filtering and searching functionalities. (b) Scatterplot to provide an overview of all ships, with the beaching model and shipbreaking model scores as axes.



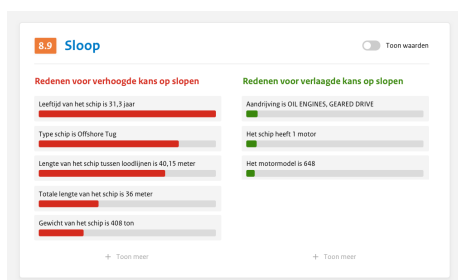
(c) The final score of the selected ship, presented as a score between 0 to 10 with a coloured background.



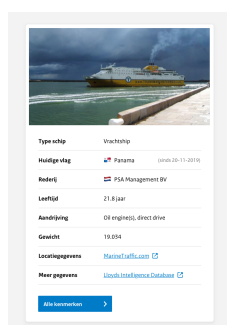
(d) Feature for adding feedback to the system, which could be given to the model developers, or eventually even the model itself.



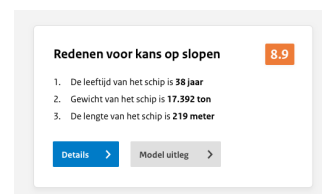
(e) Chart for visualising the distribution of a ship's feature (e.g. age), providing context through marking the value of the currently selected ship.



(f) Visualisation of the feature importance values, with separated positive and negative factors and only the 5 most important features initially shown.



(g) Summary of most important ship details, including an image of the ship.



(h) Textual list of three most important features, supplemented by a button triggering a visual representation of all the feature importances.

**Figure 6.2:** Overview of visualised components (parallel to the sketches in Figure 6.1), which were created for the focus group. A full overview of all components, including high resolution versions and alternative visualisations are provided in Appendix D.

## 6.3 Focus Groups

The aim of the focus groups was to answer to the following main questions, which was done through a discussion and a practical dashboard creation task.

- Are the proposed component designs understandable?
- Are the proposed component designs well-designed in terms of usability?
- How should the dashboard be composed?
  - Which components are most important?
  - Which components need to be combined in one view?
  - Which components can be hidden behind an extra ‘click’?

### 6.3.1 Participants

Two focus groups were planned. The first session was conducted on team members of the IDLab (N=6, 4 male, 2 female), having a background in data science. The second was done with peer students following the HCI Master at the Utrecht University (N=5, 3 male, 2 female). It was expected that the first group would gather useful feedback as they have a general idea of who inspectors are and how they work. The second group was chosen to mainly gather feedback from a usability perspective, but also to gather insights from non-experts in data science.

### 6.3.2 Materials

The focus groups were centered around the low-fidelity component designs that were created. These were all printed and separately cut out for the practical dashboard composition part of the session. Photos were taken of the dashboard compositions. Feedback and commentaries were recorded through note-taking.

### 6.3.3 Procedure

A few days prior to the focus group, participants were provided information regarding the background of the research, an explanation of the session’s goals and outline, and an introduction about the six design themes as defined in chapter 2.6.4. When starting the session, this information was shortly repeated, and the ground rules of focus group session were explained (quantity over quality, all ideas matter, no negativity, wild ideas are encouraged, build on other’s ideas).

Next, all individual components (as presented in Appendix D) were shortly explained. The individual components were presented within their corresponding design theme, in order to support the participant’s understanding of the components, and to give an overview of the available visualisation choices within each theme. Participants were asked for each component if they found it understandable, usable and if they had suggestions for improving them. In the case a component had different versions, it was asked which they preferred. Participants were further encouraged to share with the group if any of these thoughts would come up during the rest of the session. After the discussion rounds, the participants were as a group asked to compose dashboard

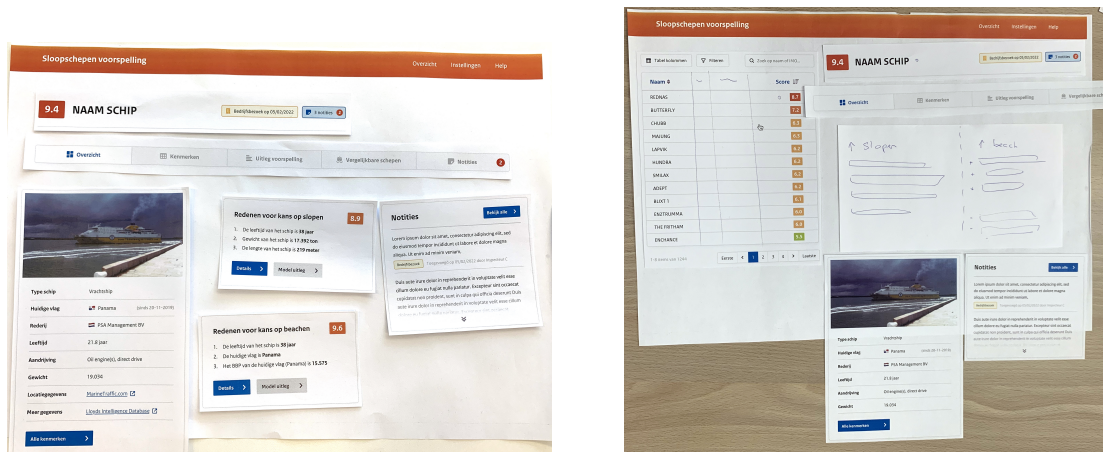
screens using the component cutouts. They were also allowed to draw or write annotations on the components, or sketch entirely new components on blank pieces of paper. Participants were asked to motivate their design choices and openly discuss these with each other. The dashboard layout compositions iterated until at least every selected component was used at least once. The total duration of the focus group session was two hours.

### 6.3.4 Analysis

The gathered feedback was structured to a list of general conclusions, a list of motivations and thoughts per dashboard composition, and a list of feedback that was considered not to be relevant enough for the scope of the current research, but can be interesting for future work. This information, together with the photos of the dashboard compositions, was used for making decisions on the final dashboard compositions.

### 6.3.5 Results

The IDLab group created a total of six dashboard compositions for different screens. The students decided to compose only one main screen, and mostly gave further feedback on individual components. Examples of dashboard compositions are shown in figures 6.3a and 6.3b. A full overview of the session results, including all dashboard compositions and individual component feedback is provided in Appendix E.



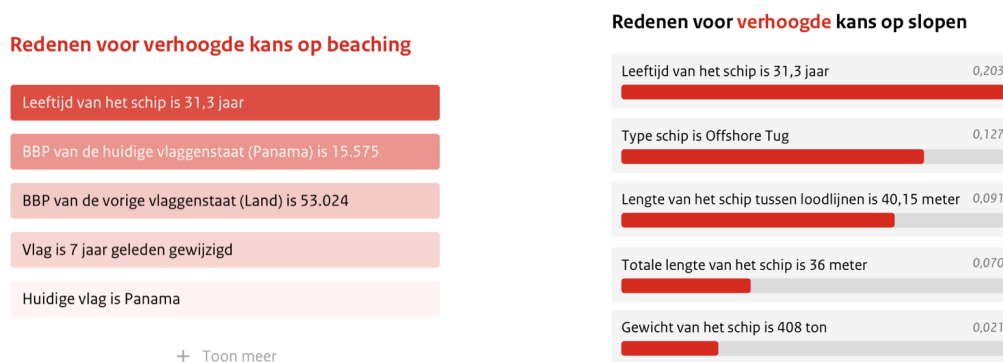
(a) Dashboard composition of the details overview screen, created in the first focus group.

(b) Dashboard composition as created during the second focus group with students. The overview table is placed at the left of the screen, in which instances can be selected. Upon selection, the right part of the screen will update to show all the details of the selected instance.

**Figure 6.3:** Resulting dashboard compositions

Overall, both groups were generally positive about the designed components, but there were some textual and visual improvement suggestions. These were on the level of changing specific words, colours or adding interface elements as buttons or icons. Furthermore, in both groups there were misinterpretations about the design of the "similar ships" visualisation. When discussing this with the participants, some suggestions for improvements were posed. Next, a new idea emerged in the IDLab focus group for creating a "company page", giving the inspectors an overview of all the ships that are registered to a company, which provides yet another way for browsing through the dataset.

There were three major differences between the two focus groups. Firstly, the IDLab group unanimously concluded that the dashboard should present the textual explanation first, and the more detailed visual explanation only if the user asks for it. However, the students group strongly preferred to present the visual explanation only, as they thought the usage of visuals made it easier to quickly distinguish the most important factors and compare them with those from other instances. The second difference is that the IDLab group voiced a preference for the visual explanation using the bar's opacity for representing the feature importance values (Figure 6.4a) opposed to the progress-bar like visualisation (Figure 6.4b). In contrast, the students group mentioned that the opacity focused visualisation would not meet accessibility standards. It would be difficult for people with visual disabilities (e.g. colour blindness) to distinguish the bars. Also, the transmission of information would become dependent on the quality of the user's computer screen. Since not all screens can display the same detailed levels of contrast, information might get lost. The third major difference between the focus groups, was that the IDLab group composed their layouts with the idea of having a home screen for an overview of instances, and a separate details screen which can be accessed after selecting an instance. On the contrary, the students group was unanimously in favour of not having a separate 'start' and 'detail' screen, but to combine them in one (Figure 6.3b). This came from the idea that the end-users might want to quickly browse through ship instances. It would then lead to a bad user experience



(a) Visualisation of feature importance value through the opacity of the coloured bars. (b) Visualisation of feature importance value through the width of the colored bars.

**Figure 6.4:** Comparison of two different ideas for visualising feature importance.

if users have to go back and forth between the overview and details screen every time, with additional risk for users to get lost in navigation.

### 6.3.6 Final Dashboard Compositions

Based on the gathered insights, decisions were made for the layout of the dashboard and individual visualisations. First of all, the improvement suggestions that were given for individual components are processed in the next phase, during the creation of the high-fidelity prototypes. Furthermore, design choices that were common to both focus groups have been adopted.

Regarding the conflicting opinions between the groups, the following decisions were made. Firstly, both the textual and visual explanation will be included in the dashboard. The IDLab group was strongly in favour of this because of a possible unwanted information overload if all details would be shown. As this group has a bit more knowledge about the target audience, it was decided to follow their reasoning. Additionally, by including it in the dashboard prototype, there would be the opportunity to evaluate both explanation visualisations with the end-users themselves. Secondly, regarding the choice of visual explanation (as presented in Figure 6.4), it was decided to use the progress bar-like visualisation, in order to prevent any accessibility issues. Lastly, the decision was made to create a dashboard with a separate 'overview' and 'details' screen. Although the argument about navigation back and forth is sound, no use case was described of users wanting to skip through all the instances that quickly. Additionally, combining the two screens may lead to an unwanted information overload for the end-user.

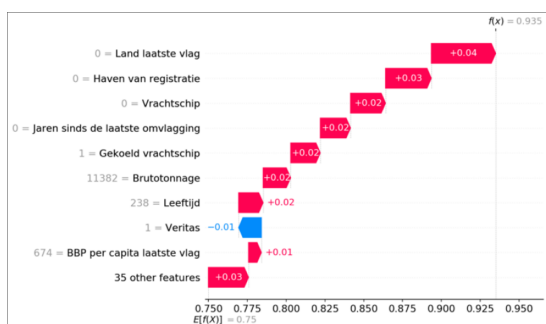
The resulting dashboard layouts and improved visualisations were processed during the creation of the high-fidelity prototype, of which the results are presented in the next chapter.

### 6.3.7 Final Visual Explanation Design

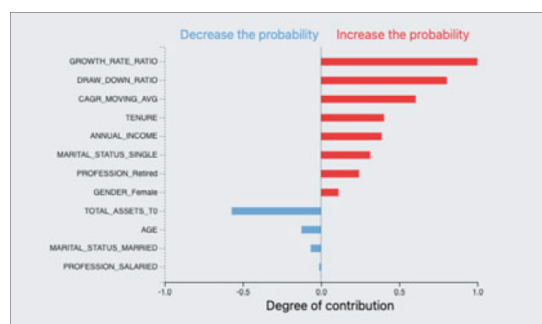
As the visual model explanation is one of the main components of the interface, it is worthwhile to elaborate on the proposed design in more detail. To explain the design decisions and how these were specifically at the target audience, it can be compared with an industry-standard SHAP visualisation (Figure 6.5a) and an enhanced version from one of the mentioned related studies (Figure 6.5b, Souza and Leung (2021)).

The first difference that stands out in the proposed design, is the reader-friendly textual description of each feature in the format of "*[feature] is [value]*". This is in line with the findings of Haas (2021) and Szymanski et al. (2021), stating that providing textual explanation with a visual explanation benefits the overall understandability. In the proposed design, these are not treated as separate but are seamlessly combined in one visualisation. The second key aspect in the proposed design is that it contains visual and textual hints about how to interpret the information. Positive and negative features are split and colouring and texts in the format of "*Reasons for a [higher/lower] risk of [shipbreaking/beaching]*" are assumed to support understandability and decrease the

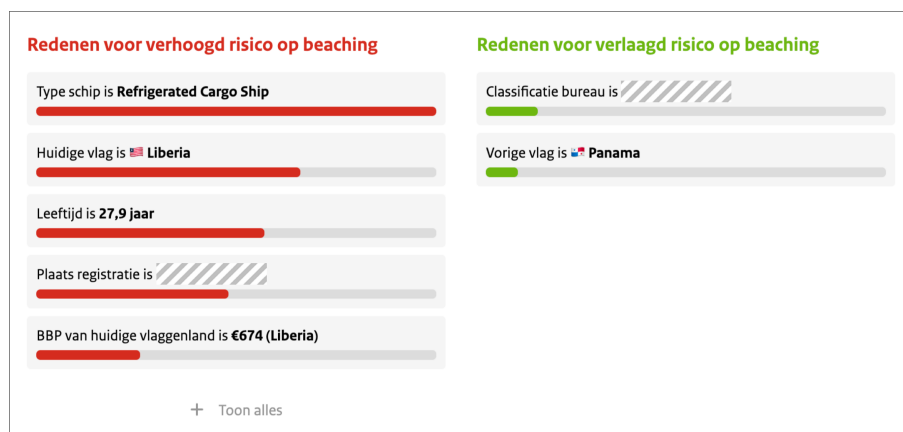
user's cognitive load. This was done similarly in the design presented by Souza and Leung (2021). The third major difference is that the raw feature importance values are not presented textually. As mentioned in one of the focus groups, it was expected that the end-users would not need the exact numbers and including them would only confuse them. Therefore, they are only presented through sizes of coloured bars. Finally, only the five most important features are presented to avoid information overload and guide users to the most important reasons. Manual assessment of the data showed that a number of five features in most cases provided a good balance between the presentation of relevant features and a manageable amount of information. Limiting the amount of presented features is similarly applied in industry-standard SHAP visualisations, but most of them provide no option to expand the list to show all features.



(a) Waterfall plot of the shipbreaking model as evaluated by Haas (2021)



(b) Custom enhanced SHAP visualization as evaluated in The Customer Turnover Interface prototype (Souza & Leung, 2021)



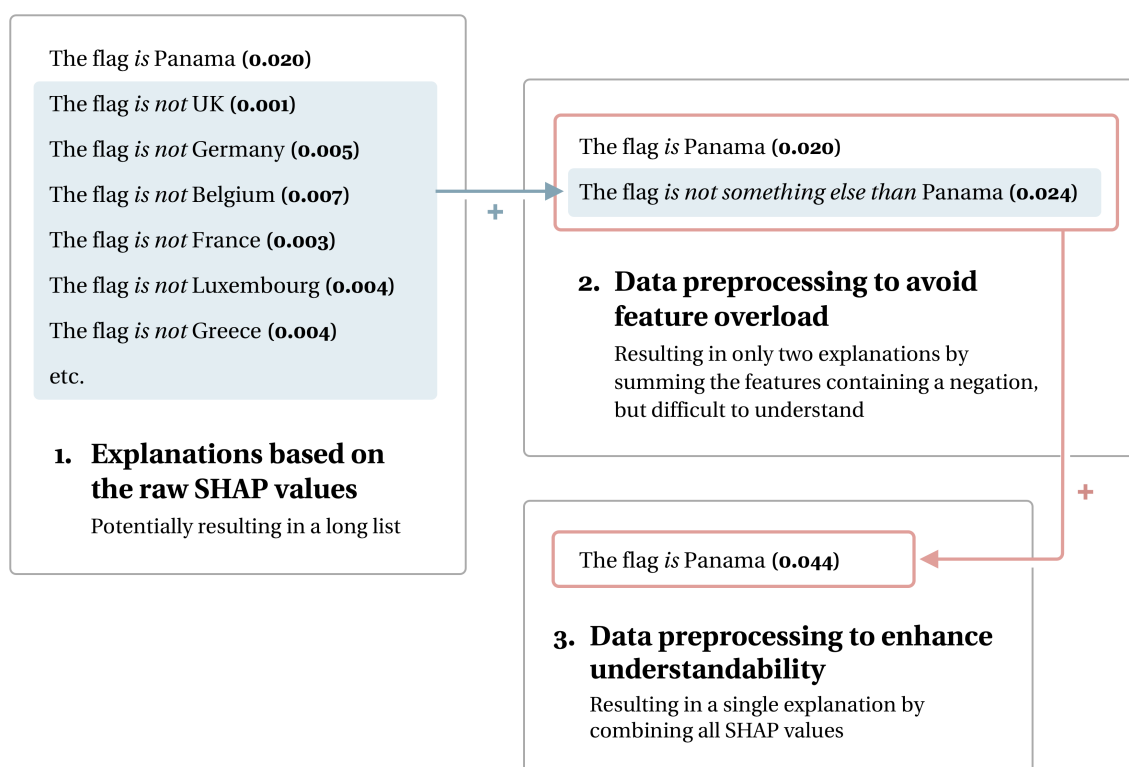
(c) Proposed design in the current study

**Figure 6.5:** Comparison of local feature importance visualisations.

Another decision concerns the widths of the coloured bars. As ships are mainly viewed individually, the bar widths have a local meaning. Hence, the highest SHAP value within each model prediction is visualised as a 100% width bar for that specific ship, from which the bar widths of the other features are calculated. Finally, some data preprocessing had to be carried out for the categorical variables in the model. This is due to the fact that categorical variables sometimes contained a feature importance value for negations. For example, a score for a ship carrying the Panama flag could be



explained by the fact that it did *not* carry the Dutch flag (and so on for every possible flag). To avoid an extensive list of these features, the first step was to combine all the feature values containing a negation into a single feature by summing their SHAP value. In a first trial, this led to explanations such as "The «feature» is not something else than «value»". As this was expected to undermine understandability, it was decided to combine all features to a single one for the categorical variables. This seemed to be the best option in terms of understandability, while still reflecting the underlying meaning of the feature's explanation. To illustrate the steps that were taken, an example is shown in Figure 6.6.

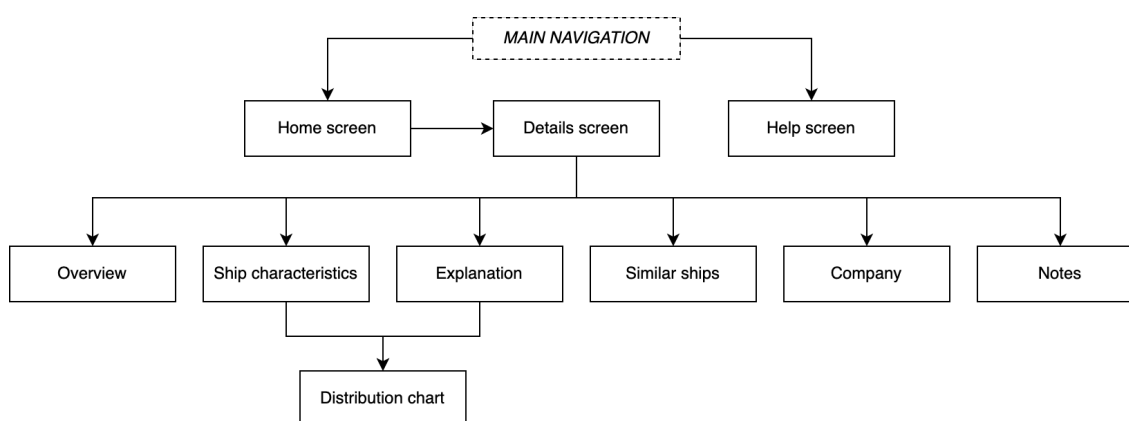


**Figure 6.6:** Visual description of the data preprocessing steps that were done for categorical variables. The features and SHAP values are fictitious and only aimed to illustrate the steps.

## 6.4 High-Fidelity Prototype

The high-fidelity prototype was created using the low-fidelity prototypes and proposed dashboard layouts in the focus groups. Since all the data of the ships, including the model predictions and explanations were available, there was the opportunity to create a product as realistic as possible. The ability to evaluate a real working product with the intended end-users was expected to result in higher quality and more detailed experiment results.

The dashboard's front-end was created in Vue.js, a Javascript framework allowing for creating highly interactive web-based tools. The data of all ships was loaded from a single JSON-formatted file. By decoupling the dashboard from the data, future deployment plans allow to easily update the prototypes with new data, without having to make alterations to the dashboard itself. The structure of the dashboard is shown in Figure 6.7 and all the individual screens are presented in the next sections. Note that privacy-sensitive information about the ships or companies are masked in the attached screenshots.



**Figure 6.7:** Structure of the dashboard prototype

### 6.4.1 Home Screen

Upon entering the dashboard, users will be presented the home screen as shown in Figure 6.8. Here, the user is presented an overview table of all instances (**A**), including the prediction scores for the shipbreaking model, beaching model and the aggregated final score. Initially, the instances are sorted in ascending order of their final score, but users can modify the ordering by clicking the column headers. Furthermore, the table columns can be changed as well (**B** opens **F**). The instances that are shown can be filtered by a large number of characteristics (**C** opens **G**), for example by ship type or prediction scores. Finally, users can quickly search for specific instances by typing in the search bar (**D**), which filters the results on their unique ship number, ship name or owner name.

Next to the table, a scatterplot is presented with an overview of all the instances (E). The user is able to select which ship characteristics should be shown on the axes, which allows them to explore the dataset and relations between characteristics. The scatterplot is linked to the table, meaning that the ship that is hovered is highlighted in the scatterplot. When clicking on one of the ships in the table, the user will be navigated to the detail screen.

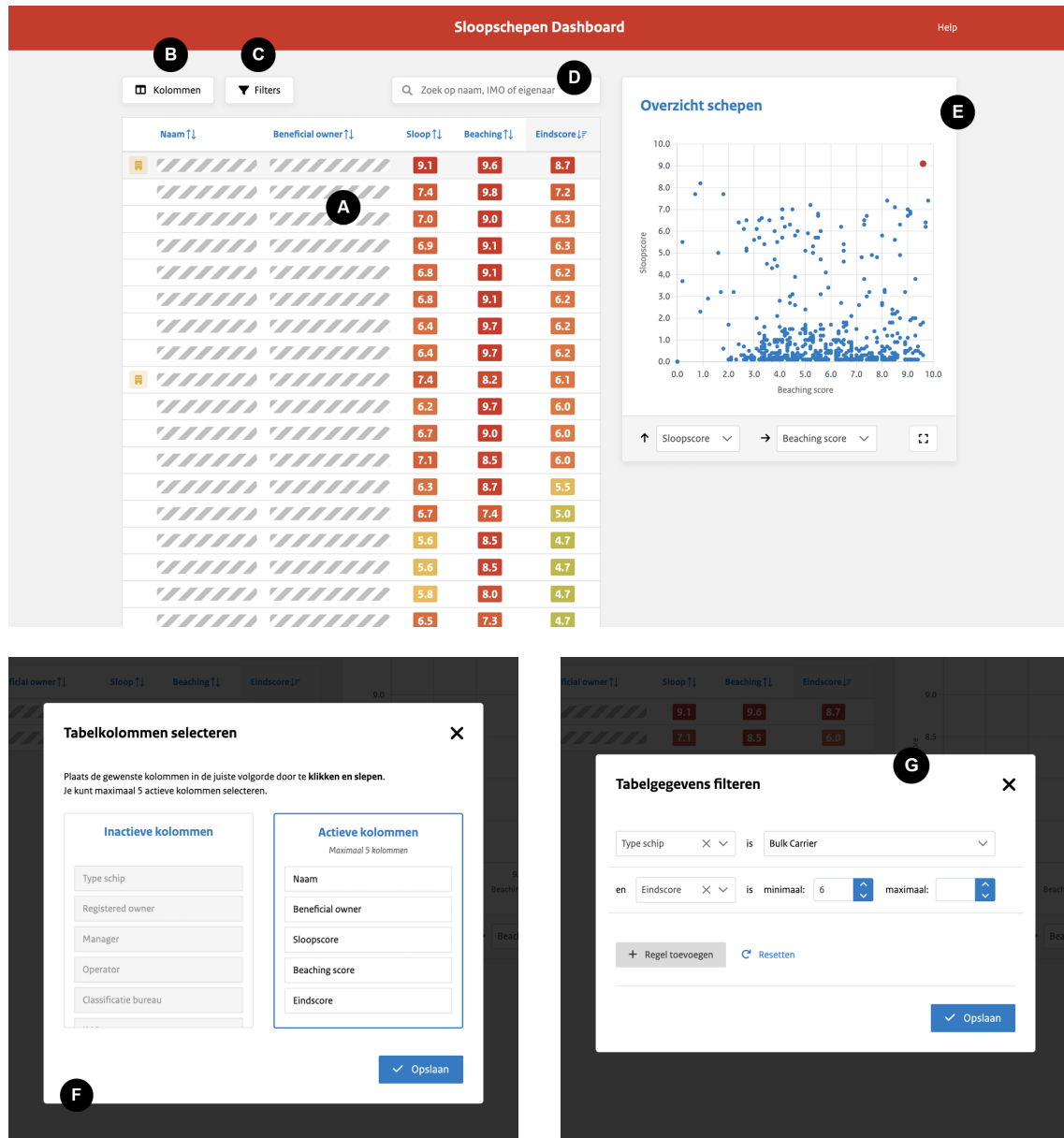


Figure 6.8: Screenshots of prototype home screen.

### 6.4.2 Detail Screen: Overview

The detail screen consists of six child screens, of which the overview screen (*Overzicht*, Figure 6.9) is the one that is initially shown. At the top of the detail screen page, the final score and title of the selected ship are shown (A). Below this, there is a tab bar for navigating between the other screens (B). On the overview screen itself includes three components. Firstly, a summary of the ship characteristics (C) that are expected to be the most important, based on the requirements interview and focus group results. Also, an image of the ship is presented here, to get a quick glance at the type and size of the ship. For more details, the user can click on a button to show "all characteristics", which leads to the "Ship Characteristics" screen as described in section 6.4.3.

Next to this, there is a textual overview of the prediction explanation (D). For both the shipbreaking and beaching model, the prediction score is shown, including a list of the three most important features with a positive importance value. Again, for a more detailed look, the user can click the "details" button which leads to the "Explanation" screen as described in section 6.4.4.

At the right of the page, an overview of all the notes that were added to the ship is displayed (E). Clicking a note opens the full text in an overlaying modal. Again, for a more detailed overview of all notes, the user can click the "see all" button which leads them to the "Notes" screen as described in section 6.4.7.

All in all, the overview page provides the user a quick glance at some of the most elementary information about a ship. It gives an overview of the ship characteristics itself (C), a quick understanding of the predictions (D), and some historic information such as recent company visits (E). As the main goal of the dashboard is to decide if a ship's company should be visited, this decision is also shown in the page header (F).

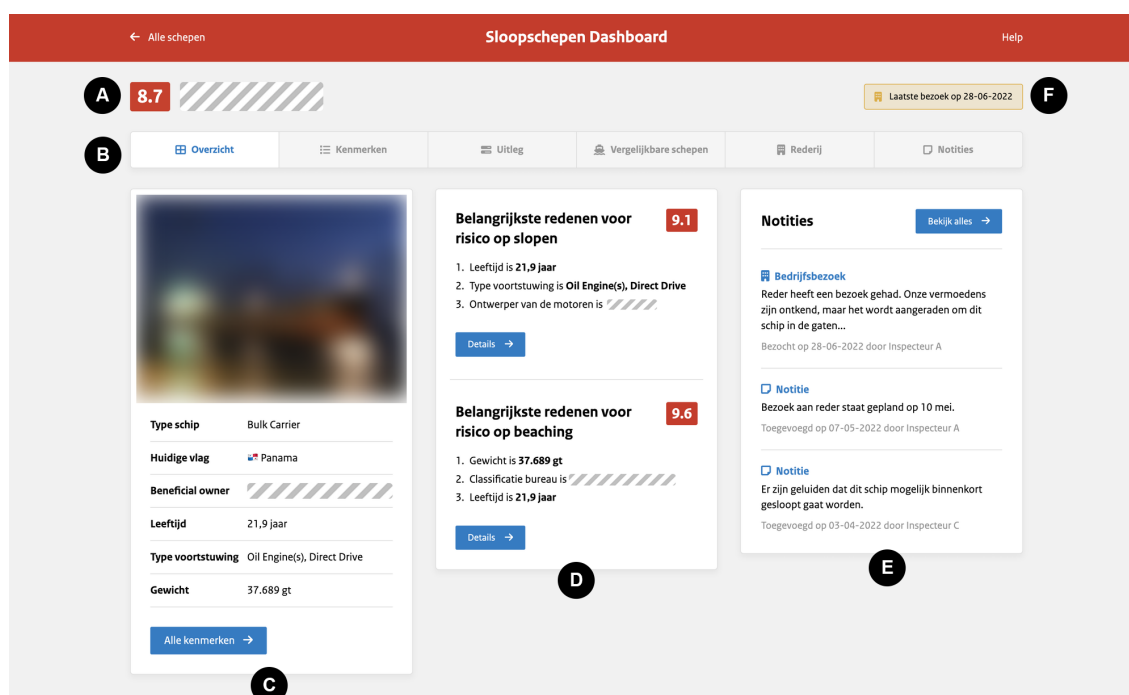


Figure 6.9: Screenshot of prototype detail overview screen.

### 6.4.3 Detail Screen: Ship Characteristics

On the ship characteristics tab (*Kenmerken*, Figure 6.10), an overview is provided of all the known data of a ship (A). This includes data which was used in the shipbreaking and beaching models, but also some extra information that is expected to be of value to the user, such as the unique ship number (IMO) and ownership data (registered owner, manager, beneficial owner, operator). Also, some links are provided to external websites for even more information. Most characteristics include a little chart icon (B), which is a button linking to the corresponding distribution chart that is further elaborated in section 6.4.8.

The screenshot displays the 'Sloopschepen Dashboard' for a specific ship. The interface is organized into several sections:

- Registratiegegevens (A):** A list of registration details including IMO, Registered owner, Manager, Beneficial owner, Operator, Huidige vlag (Panama), Vorige vlag (Panama), Tijd sinds vlaggenwissel (11 jaar), Classificatie bureau, and Plaats registratie.
- Technische kenmerken (B):** A list of technical specifications including Type schip (Bulk Carrier), Leeftijd (21,9 jaar), Type voortstuwing (Oil Engine(s), Direct Drive), Gewicht (37.689 gt), Draagvermogen (72.497 ton), Totale lengte (225 meter), Lengte tussen loodlijnen (217 meter), Gemiddelde vaarsnelheid (14,5 knopen), Capaciteit van de koelruimtes (0 m³), Totaal aantal motoren (1), Maximale vermogen van de motoren (12.181 kW), Model van de motoren (6RTA62), and Laadvermogen (0 TEU).
- Overig:** A section with links for 'Meer foto's' (MarineTraffic galerij), 'Locatiegegevens' (MarineTraffic.com), and 'Meer details' (Lloyds Intelligence Database).

Figure 6.10: Screenshot of prototype characteristics screen.

### 6.4.4 Detail Screen: Explanation

The explanation tab (*Uitleg*, Figure 6.11) provides a more detailed view on the prediction explanation. Separate explanations are shown for the shipbreaking (A) and beaching model (B). The width of the bars are defined by their importance value in which the importance value within each prediction is visualised as 100% bar width. From here, the width of the other bars are calculated. If there are more than five features in the explanations, a "show all" button appears (C) which triggers the expansion of the full list. Most feature explanation again include the chart icon (D), which triggers the distribution chart that is further elaborated in section 6.4.8.

Next to the explanations, a short help text is provided including a button for more information (E). This button leads to the "help" screen as described in section 6.4.9. Finally, the user is asked to give their feedback about the prediction and explanations (F).

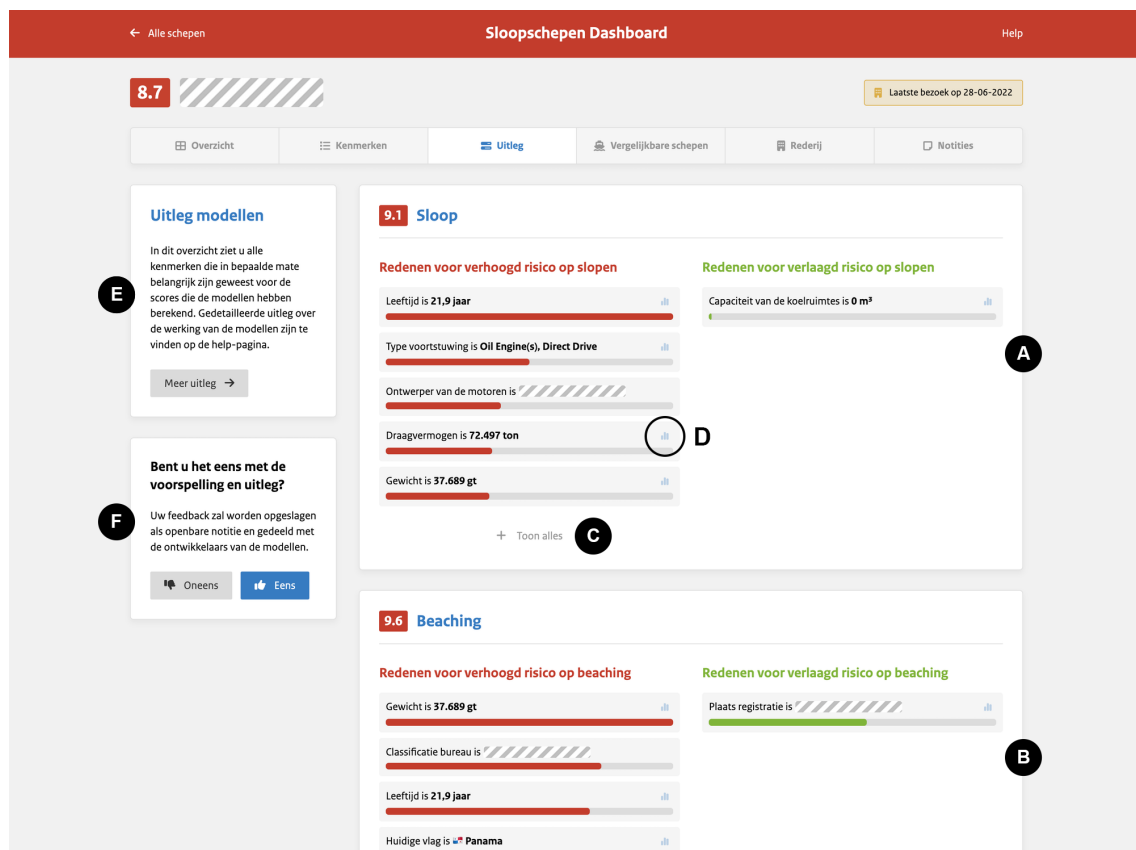


Figure 6.11: Screenshot of prototype explanation screen.

### 6.4.5 Detail Screen: Similar Ships

For the similar ships tab, comparison scores were calculated between each of the ships. This was done using mostly physical characteristics, such as the ship type, length, volume, speed and age. For calculating the similarity scores, the Gower distance metric was used (Gower, 1971), as this allowed for calculating a difference between records with mixed data types. After manual assessment of the calculated scores, it was decided to include only ships with a comparison score above 0.9 (in a scale of 0 to 1) and a maximum of 20 ships.

Using the feedback of the focus groups, this screen (*Vergelijkbare schepen*, Figure 6.12) was redesigned to show a card of the currently selected ship (**A**) with a scrollable list of all similar ships ordered by their similarity score (**B**). As this score might be too abstract or not relevant for the user, it was decided not to present this score in the interface but only use it for the ranking order. The cards in the list are clickable and lead to the detail page of the corresponding ship.

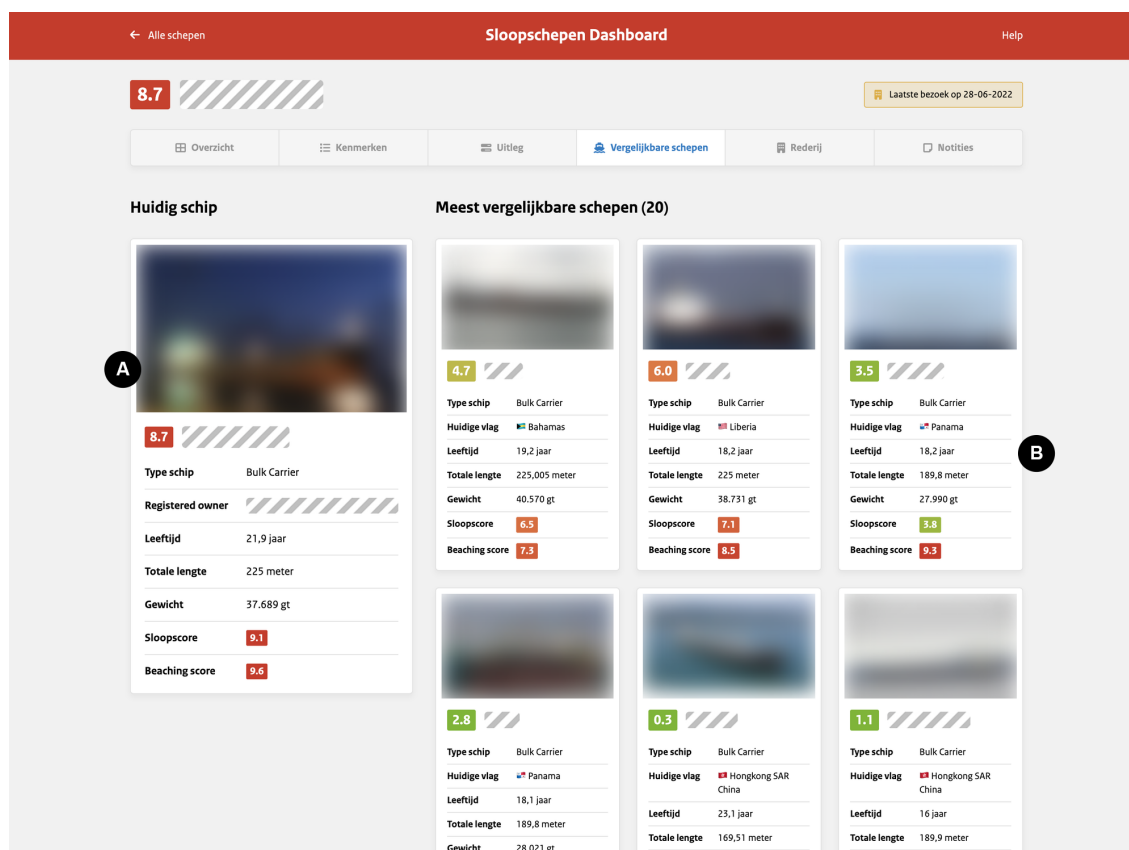


Figure 6.12: Screenshot of prototype similar ships screen.

### 6.4.6 Detail Screen: Company

The company page (*Rederij*, Figure 6.13) shows a list of all ships from the same company as the currently selected ship (A), including some ship details. The cards in the list are clickable and lead to the detail page of the corresponding ship. The ordering of the cards can be changed to alphabetically or by either shipbreaking, beaching or final score (B).

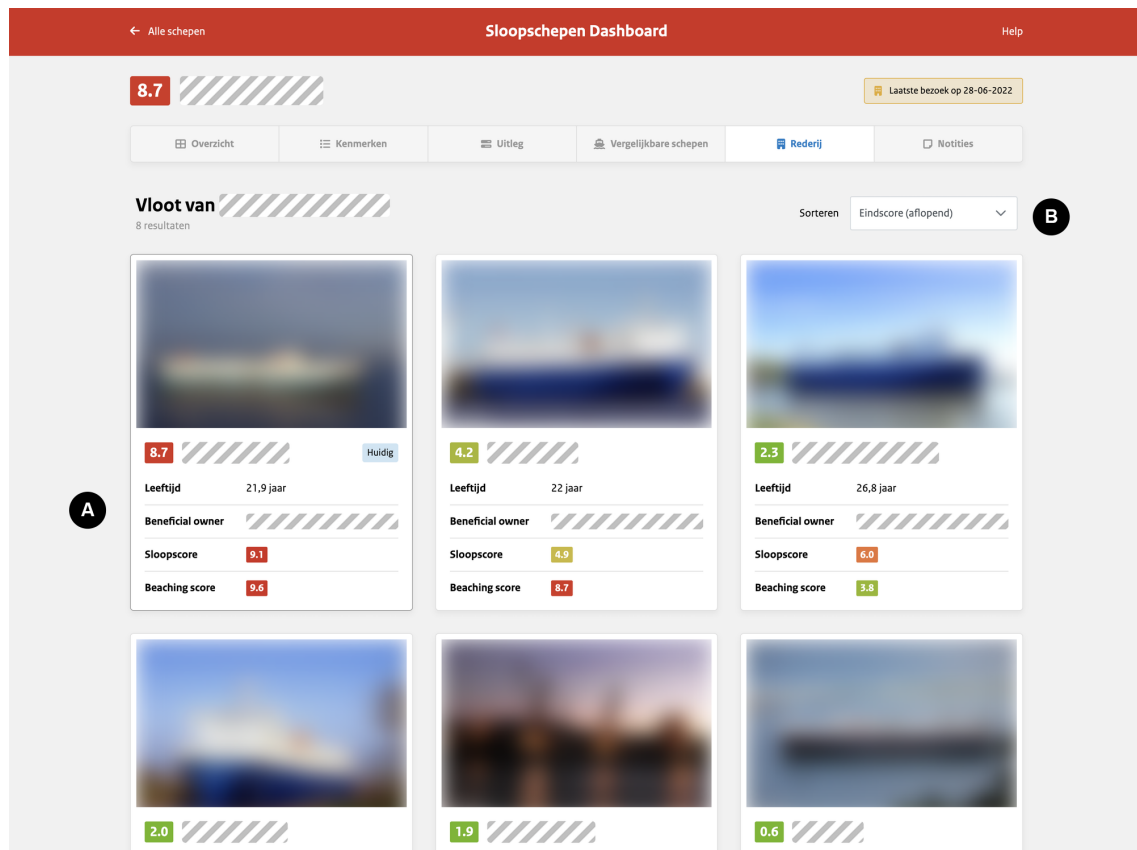


Figure 6.13: Screenshot of prototype company screen.



### 6.4.7 Detail Screen: Notes

The notes page (*Notities*, Figure 6.14) shows a full overview of memos that were added by the inspectors. There are three types of notes: general notes, company visits and feedback. The general notes (**A**) can be used to document any event about a ship, such as alerts from external parties. Company visits (**B**) can be used to document the event of when any inspector has visited a company for an interview or inspection. If a company visit was added, this is also shown in the page header (as described in section 6.4.2) and in the overview table on the home screen (as described in section 6.4.1). Lastly, users can add feedback about a prediction or explanation, which can be of use for the model developers and to colleague inspectors (**C**).

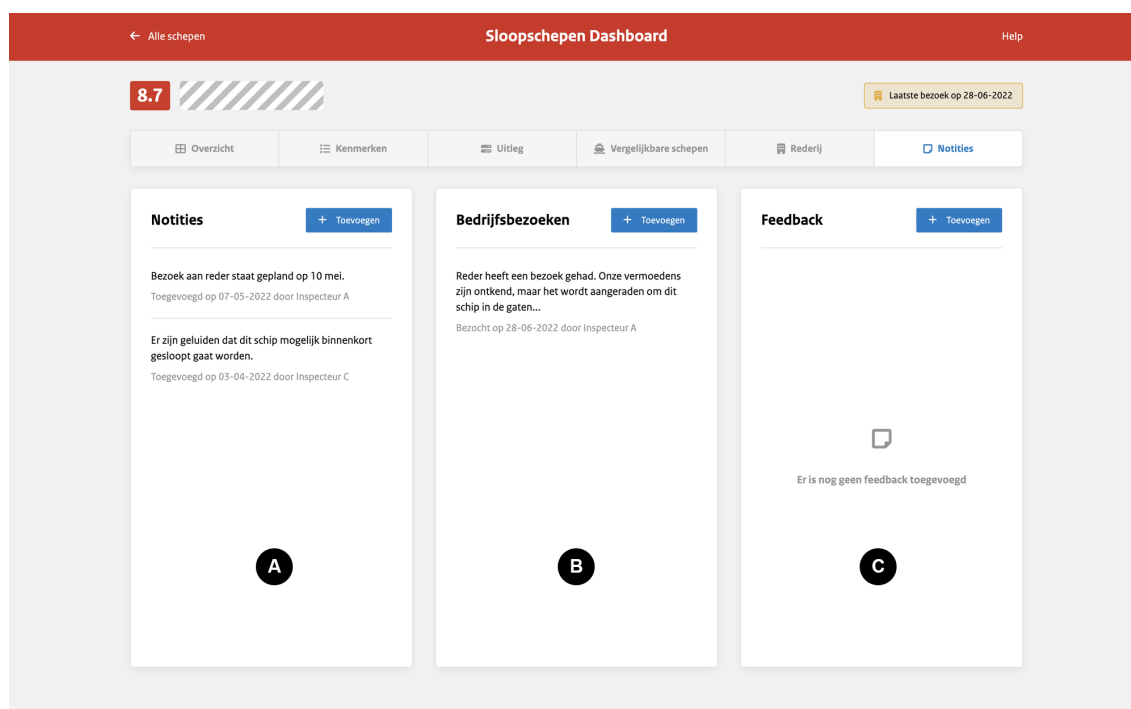
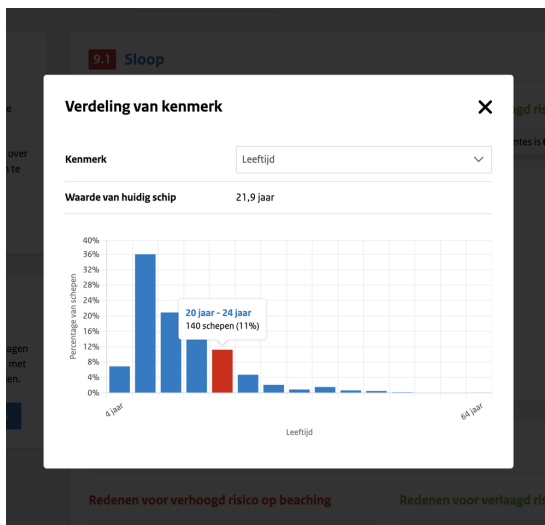


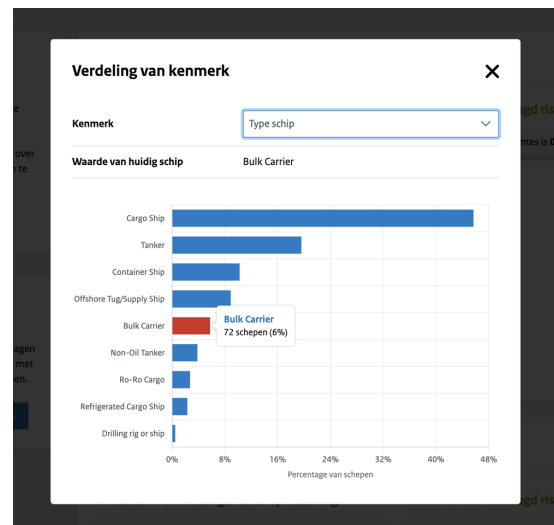
Figure 6.14: Screenshot of prototype notes screen.

### 6.4.8 Detail Screen: Distribution Charts

The distribution charts (Figure 6.15) are triggered by the chart icon on either the characteristics page (section 6.4.3) or explanation page (6.4.4). They open as an overlay on the active page and show a distribution of the selected feature for all ships in the dataset. For features with numerical values such as the age, volume and deadweight of a ship, a histogram is presented (Figure 6.15a). For categorical features, such as the ship type or the flag it carries, a horizontal bar chart is used to visualise the frequencies (Figure 6.15b). The value of the currently selected ship is highlighted and by hovering the bars, more information is shown in a tooltip.



(a) Numerical feature values: histogram



(b) Categorical feature values: frequency bar chart

**Figure 6.15:** Screenshots of prototype distribution charts screen.

### 6.4.9 Help screen

Lastly, the help screen (Figure 6.16) aims to provide users a background on the models, how they were created and how they should be interpreted. The page can be accessed at all times from the main navigation bar.

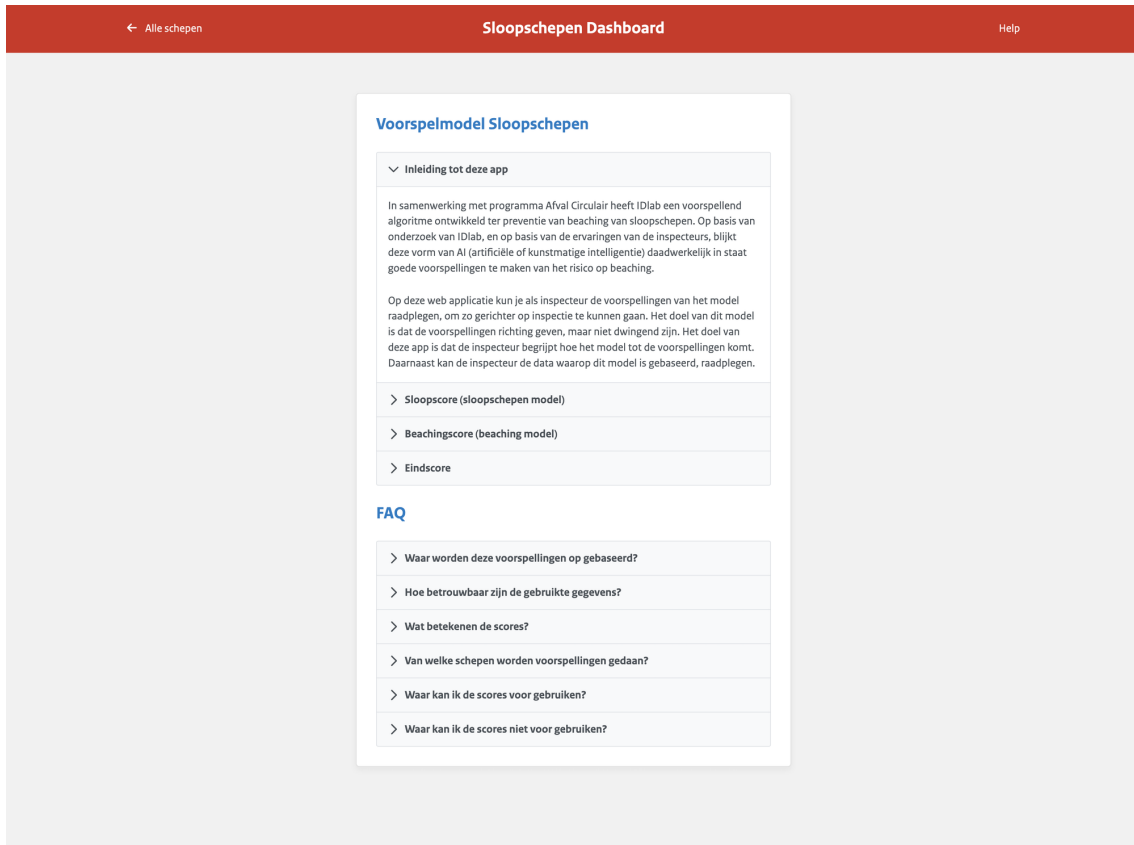


Figure 6.16: Screenshot of prototype help screen.

## Chapter 7

# Evaluation

After the creation of a high-fidelity prototype, the next step was to evaluate the design solutions that have been created in this research. The proposed interface design was evaluated by measuring participants' perceived usability, understandability, trust and reliance, and by gathering feedback from the intended users of the system.

### 7.1 Methodology

The main method used in the experiment is a task-based analysis, which is a common approach in HCI research. In this method, participants are asked to carry out specific tasks, while their actions are being observed. Also, participants are asked to share their thoughts, which is known as *thinkaloud* and is seen as a valuable method for analysing cognition of test users (Nielsen, 1994). A combination of these methods can be used to analyse how people use and interact with an interface. The entire evaluation was conducted in Dutch, as the interface is also in Dutch and it could not be assumed that all participants would speak English proficiently.

Before carrying out the experiment with real inspectors, one pilot session was held to validate the experiment setup. Based on the feedback from the pilot participant, a few minor bugs in the interface were corrected and the instructions at the start of the experiment were revised.

#### 7.1.1 Participants

The evaluation was conducted with the intended end-users of the system: shipbreaking inspectors. Invitations were sent to a number of inspectors working on the topic of illegal shipbreaking. As there are only around six people dedicated to this subject, only a small number of participants would be available.

Four participants (N=4, all male, age 45-65) were recruited to participate in separate one-to-one experiment sessions. Knowledge about ships and illegal shipbreaking varied; two participants were senior inspectors, who worked in the field for a long time and therefore knew a lot about the shipping business. The other two participants only had general knowledge, of which one mentioned to be mainly concerned with the legal aspects of illegal shipbreaking. Knowledge regarding the models also varied; one participant (P1) has been involved in the development of the models and was therefore somewhat familiar with the output of the models. The other three participants had no prior knowledge about the models.

### **7.1.2 Materials**

The experiment was centered around the high-fidelity prototype that was created. This prototype was loaded onto a company laptop of the same brand and type, in order to avoid bias caused by unfamiliarity of the hardware. During the entire session, after consent, the audio was recorded as well as the laptop screen during the task-based session.

#### **7.1.2.1 Usability**

Usability was measured using the System Usability Scale (SUS). This is a widely-used scale consisting of ten questions, each with five response options ranging from "Strongly agree" to "Strongly disagree". The questionnaire results in a SUS score between 0 and 100, of which the rule-of-thumb standard defines systems with a score above 70 passable (Bangor et al., 2008). Usability is not the main theme of the current study, but should be validated to ensure the tool is usable enough for the target audience. The questions of the SUS can be found in Appendix F.

#### **7.1.2.2 Trust and Reliance**

The user's trust in the system was researched in two ways. First, through analysing reliance on the visualisations in the inspector's decision-making. Although ideally trust should be evaluated in more detail, it would take an entire research project to study this effectively. Also, it could be highly affected by the duration of the experiment and the technological readiness of the prototype. Therefore, trust was measured in terms of reliance on the visualisations of the data. More specifically, the aim was to gain understanding about how individual visualisations support the participants in achieving their goals. Thus, participants were asked to rank each individual interface component by their perceived importance, and to motivate their ranking. For this, all individual components were printed on an A4 paper, so that the participant could easily select and rearrange the ranking. The second way for evaluating trust was by using a scale to identify the participant's general trust in technology. This was expected to support interpretation of the participants' responses, for example to avoid overinterpreting responses from general sceptics. The Propensity to Trust Technology (PTT) scale was used, containing six questions about the general trust in technology (see Appendix G).

#### **7.1.2.3 Understandability**

Understandability was examined qualitatively through a set of questions aiming to learn if the participant had problems understanding certain visualisations, and to validate their understandings. Each visualisation was shortly discussed, after which the following questions were asked:

- Do you think this visualisation is clear and understandable?
- After hearing my explanation, are there things that you yourself thought would work differently?
- After seeing the component, and hearing the explanation: do you agree with the choices that were made? Would you visualise things differently?

#### 7.1.2.4 Task-based Analysis

For the task-based analysis, four tasks were created based on expected real world scenarios and related to the user goals that were defined in the requirements analysis (chapter 5). The tasks were as follows.

- What would be the next ship you would visit?

*Reasoning for task: aiming to find out how the participant approaches the presented data, what information is used in decision-making and how they reason about the decision.*

- A report is coming in about the ship «name of a ship with a fairly high prediction score» that it may soon be dismantled. Assess whether the owner of this ship should be visited.

*Reasoning for task: aiming to find out if the participant is able to use the searching functionalities, to understand how the participant assesses ships with a medium high score and if he is able to recognise this by seeing the ship in context of other ships.*

- A report is coming in about the company «name of a company with 10+ ships with varying prediction scores, but all less than 6.5», assess whether any of their ships are potentially at risk.

*Reasoning for task: aiming to find out if the participant understands how to get an overview of company ships (e.g. through table filtering, or the company page), and how what decision is made for a company owning multiple ships that are not very high scoring.*

- A report is coming in about the ship «name of a ship with a shipbreaking score lower than 3.0 and beaching score higher than 9.0» that it certainly will be dismantled. Assess whether the owner of this ship needs to be visited. Imagine I am a colleague of yours, how would you explain it to me?

*Reasoning for task: aiming to find out if the participant understands that predictions can be wrong, and how is dealt with this fact. Furthermore, it aims to get an understanding of how a participant passes on the information he has obtained. This concerns the choice of words, but also the information that apparently is important for him.*

#### 7.1.3 Procedure

A few days prior to each experiment, the participant received an email explaining the research, the goals of the experiment and a brief outline.

At the start of the session, the participants were asked to introduce themselves and tell something about their domain expertise, and knowledge about (X)AI and experience with such systems. Next, the researcher introduced himself, in which the participants were provided a more elaborate background information on the research, including an explanation of what was expected from them.

First, participants were asked to fill in the PTT scale. Second, participants were given an introduction about the predictive models that were developed. This included an explanation about why they were created, how they were trained, which data was used, what output is produced by the model and how this output can be interpreted. Participants were allowed to ask questions, and during the explanation they were frequently asked if they could still follow. Then, the participant was given a tour through the prototype. After each explanation of an individual component, the participant was asked the three questions regarding understandability. After this first half of the session, which mainly focused on providing the participant with the necessary background information, there was time for a short break. After the break, the second half started, which focused more on the task-based thinkaloud analysis.

Participants were then invited to take place behind the laptop running the prototype. They were instructed that they would be asked to perform a number of tasks in the dashboard. Furthermore, they were told that they could click on anything and that the most important thing is to think out loud; so please say everything that comes to mind: why you are clicking on something, what you think of the information that you see, what you like or what you don't like. Also important: there is no right or wrong, there is no right order of actions which are being checked in the background, there is no time limit and participants were allowed to ask questions about the interface, for example if they could not find specific information or buttons. After this instruction, a task was given to the participant (as described in 7.1.2.4). When a task was fulfilled, or it was decided that enough information was gathered, the next task was given. If necessary, the researcher could ask questions to the participant for clarification of their actions or thoughts.

After the task-based session, participants were asked to fill in the SUS. Next, for gathering information on reliance, all printed components were spread around the table. Participants were asked to select the components they thought were important for achieving their goals, and to rank these in order of importance. They were allowed to leave out components in their ranking if they found them not useful at all. When the participant finished their ranking, they were asked to motivate their choices. Finally, to close off the experiment session, a few more questions were asked to the participant about how they thought about the interface, if they thought it could help the inspectors in their work and if they had anything else to share which was not discussed yet. Then, the recording was ended and participants were thanked for their time.

#### **7.1.4 Analysis**

The audio and video recordings were played back by the researcher, while all findings, interesting thoughts, opinions or observations were noted down. These were then categorised, after which overlapping annotations were merged or filtered out. The remaining set of findings are discussed in the next section, along with the results from the SUS and PTT scales and reliance rankings.

## 7.2 Results

### 7.2.1 General Trust in Technology Scale

For interpreting and comparing the participants' general trust in technology, the Likert-scale answers were transformed to scores ("Strongly disagree" = -2, "Disagree" = -1, "Neutral" = 0, "Agree" = 1, "Strongly Agree" = 2) and summed to a total score. Hence, the maximum score is 12 and the minimum -12. It could be said that a positive score means that the participant has a general trust in technology, while negative scores imply a general distrust in technology. The resulting scores are presented in Table 7.1.

All PTT scores were positive. Two participants (P2 & P3) had the highest scores of 6, one participant (P1) a score of 4 and one participant (P4) the lowest score of 2.

Participant	PTT1	PTT2	PTT3	PTT4*	PTT5	PTT6	Total score
P1	Agree	Agree	Agree	Disagree	Neutral	Neutral	4
P2	Agree	Agree	Agree	Disagree	Agree	Agree	6
P3	Agree	Agree	Agree	Disagree	Agree	Agree	6
P4	Neutral	Agree	Agree	Neutral	Neutral	Neutral	2

**Table 7.1:** Overview of scores in the Propensity to Trust Technology Scale. The item with an asterisk is the inverse question. An overview of the questions are provided in Appendix G

### 7.2.2 System Usability Scale

The resulting scores of the SUS are presented in Table 7.2. The SUS score was acceptable for two participants (P2 & P4), marginally acceptable for one other (P3) and borderline acceptable for the other one (P1).

Question	P1	P2	P3	P4
SUS1	Agree	Agree	Agree	Agree
SUS2*	Disagree	Strongly Disagree	Disagree	Disagree
SUS3	Neutral	Agree	Agree	Agree
SUS4*	Agree	Disagree	Neutral	Disagree
SUS5	Neutral	Agree	Agree	Agree
SUS6*	Disagree	Strongly Disagree	Neutral	Disagree
SUS7	Agree	Agree	Agree	Agree
SUS8*	Disagree	Disagree	Disagree	Disagree
SUS9	Neutral	Agree	Neutral	Agree
SUS10*	Neutral	Agree	Disagree	Disagree
<b>Score</b>	<b>60</b>	<b>75</b>	<b>67,5</b>	<b>75</b>

**Table 7.2:** Overview of results in the System Usability Scale. Items with an asterisk are inverse questions. An overview of the questions are provided in Appendix F



### 7.2.3 Understandability

Regarding understandability from a visual perspective, most components were found to be clear and understandable. However, some issues were found regarding understandability of the scatterplot on the home screen. Three participants either did not understand what was presented, or did not see any added value of the plot. Therefore, they stated to strongly prefer the instance table for getting an overview of the ships.

From a more informational perspective, participants stated to have some troubles understanding the contents of the presented model explanations. As this issue also came up during the task-based session, it will be further described in section 7.2.5.

### 7.2.4 Reliance

The individual reliance rankings were averaged for determining a final ranking of all 11 components, which are presented in Table 7.3. Components that were left out of rankings by the participants were assigned the highest possible ranking score of 11. Furthermore, participant 3 found multiple components equally important, hence the duplicate ranking scores.

The overview table on the home screen was ranked first by all participants. Both explanation components (visual and textual) were ranked fairly high, too; resulting in a second and third position in the final ranking. In between comes the notes page, ship characteristics, and company page. The distribution chart and scatterplot were by most participants either not included in their ranking, or ranked low. The similar ships page was not included in the rankings of three participants and ranked low by the other, leading to a last position in the average rankings.

Final Rank	Component Name	P1	P2	P3	P4	Average Ranking
1	Home Screen: Overview Table	1	1	1	1	1,00
2	Detail Screen: Visual Explanation	3	2	2	4	2,75
3	Detail Screen: Textual Explanation	2	3	2	11	4,50
4	Detail Screen: Notes Page	6	5	2	6	4,75
5	Detail Screen: Ship Characteristics Page	5	11	2	3	5,25
6	Detail Screen: Company Page	7	4	8	5	6,00
7	Detail Screen: Ship Characteristics Summary	4	11	2	11	7,00
8	Detail Screen: Distribution Chart	11	11	2	8	8,00
9	Home Screen: Scatterplot	11	11	11	2	8,75
10	Help Screen	11	11	10	7	9,75
11	Detail Screen: Similar Ships Page	11	11	9	11	10,50

**Table 7.3:** Overview of component rankings by the participants.

### 7.2.5 Findings from Observations and Thinkaloud

Several findings have emerged from the observation during the task-based session, thinkaloud and responses from the participant during the experiment. The list of findings were labelled and grouped and will be described in this section. The participants

perceived the task scenarios they had to perform to be representative of what they encounter in their actual work. Furthermore, participants mentioned to only recognise a few of the ships that were in the interface, but they could not remember the history of these ships. Therefore, it was not expected that having the real ship data in the interface influenced the decisions of the participants during the tasks.

### **Next Company to Visit**

In the first task, participants were asked which shipping company would be the next to visit. All participants started at the top of the overview list, but their responses varied. Only participant 1 noticed correctly that the ship with the highest score was already visited and chose for the second ship after a short assessment. Participant 2 changed the sorting of the table to the beaching score specifically and noticed that the top 5 ships were all from the same owner. After taking a look at some of these ships and their prediction explanations, there was a clear choice for visiting this company. Participant 3 had a similar reason for his choice, as he noticed in the initial overview table (still sorted by final score) that a specific shipping company owned five ships that were in the top 10. Visiting this company could have a large impact, as it could prevent illegal shipbreaking for a large number of high-risk ships. Without further assessing the individual ships, the decision was made to visit this company. Participant 4 explained a more traditional approach and stated to be triggered by external events or alerts, which could make him want to investigate a ship's predictions. He also mentioned that a high shipbreaking score alone can be a reason to visit a company, because even when ships are dismantled legally, the companies often do not have the right permits. According to him, the inspectors could try to prevent this by also visiting companies that have ships with a high shipbreaking score, regardless of their risk of being beached.

### **Interpreting Prediction Explanations**

During the tasks it became clear that most participants were interested in the explanations for the predictions. Only participant 4 did not seem to be curious for the prediction explanations. At one moment, he came across a ship of which he could not understand why it had a fairly high beaching score. He then started to assess the ship himself on the basis of the photo and its features as he would traditionally do, whilst mostly ignoring the presented explanations. In order to still get his thoughts on the model explanation, the last task was slightly modified to force the participant into looking at the explanations.

Overall, the participants seemed to grasp the visual aspect of the model explanations well. They were able to reproduce the presented explanations in their own words and some participants specifically mentioned that they found the visual explanation page well-organised. However, all participants experienced some serious difficulties with interpreting the content of the model explanations. It was mentioned that some features were difficult to use for assessing the prediction score. An example is the *classification society*, which was presented for a ship as a reason for beaching. On this, a participant (P1) said: *"It does not make sense to me that this could be a reason for a higher risk of beaching."* On the same feature, another participant (P2) said: *"I understand that*

*the classification society is apparently part of a pattern that was found by the model, but it still does not help me to understand the prediction score".* About the explanations in general, participant 1 sometimes expected to find different reasons: *"I understand why age and the flag are reasons for beaching, but I would expect the length to be an important feature here."* Furthermore, participant 3 asked for more information: *"The system says that these are important reasons, but it should also give me information on why this is a reason. So if the system tells me that a specific type of ship is an important reason for shipbreaking, it should also explain me for example that this is because these ship types are being phased out."* The last quote indicates that the participant did not have a clear understanding of how the prediction models work, and what they are and are not capable of.

Although these issues regarding the interpretability of the explanations mainly address the internals of the models and not the visualisation, they negatively impact factors such as understandability and trust. This was confirmed by participant 1, who mentioned that the predicted scores actually seemed to be accurate for the ships he had seen, but that he began doubting his trust in the model after seeing the unexpected prediction explanations. Therefore, these are relevant concerns.

### **Navigational Issues**

All participants experienced in some way issues with navigating through the system. Three participants had trouble finding the company page, because they had to select a ship first before being able to access the company page. One participant tried to click on the company name in the overview table and was surprised that he was not navigated to the page of this company. Furthermore, participants were sometimes lost after navigating to another ship's detail page through the company page. After investigating the other ship, they wanted to go back to the previous ship which turned out to be rather difficult as the back-button in the browser had to be clicked multiple times for each "detail page tab" they had visited.

### **Charts**

Similar to the findings in the understandability section, participants had difficulties understanding the charts that were in the interface. The scatterplot turned out to be the most troublesome. Participants did attempt to use the plot, but in most cases they did not seem to find it very useful and returned to the more intuitive overview table. Also, the scatterplot included functionalities to change both the axes, allowing for exploration of relations between ship features and prediction scores. This functionality remained untouched during all sessions. Despite these remarks, one participant mentioned it to be useful that the hovered instance in the table was highlighted in the scatterplot, as it gave him an idea of the score in relation to the other ships. Yet, the three other participants preferred having a larger table with more columns and to omit the scatterplot entirely or hide it behind a button.

The distribution charts also generated some comments by the participant. In contrast to the scatterplot, participants seemed to be able to understand and interpret the distribution charts. Three participants thought the charts could be useful, but they also

doubted if they would really use it in practice. It was seen as a more advanced feature that could provide some background on the data in the beginning, until the inspector has some knowledge about certain characteristics. Participant 1 thought the charts could be useful, but as some types of ships have a much longer life span than others, an overview of the age of all ships such as implemented in the prototype hardly provides valuable information on the relative age of a ship. According to him, it would only be useful if the charts only include similar ships to the one selected.

### **Similar Ships Page**

Although participants mentioned at the beginning of the experiment session to find the similar ships page useful, they seemed not to be interested in the page for the tasks. While three participants ignored the page entirely, one other did try to interpret the information. However, he mentioned to not be sure how the information could be useful for him. One of the other participants (P4) argued that it was just too much information for him.

### **General Opinion**

At the end of the session, the participants were asked what they thought of the interface, and whether they thought it would be potentially useful in the field. All participants felt that the system provides a strong basis for something they could use. To be more specific, participant 2 said to be pleased with the result and that it is definitely a step in the right direction. The way the system provides searching and filtering options to explore the data and predictions was mentioned to be very useful. Participant 3 said that he expected the dashboard to be more difficult, but that he actually liked it and that it was generally understandable and nice to use.

## Chapter 8

# Discussion

The aim of this research was to gain insights on how data visualisations of XAI models could effectively support ship inspectors with preventing illegal shipbreaking. This contributes to the knowledge gap that was identified with regard to making XAI accessible to end-users, being the ones responsible for decision-making based on outcomes of such models.

Overall, results of the evaluation with four shipbreaking inspectors indicate that the proposed interface design is promising. It was received positively by the target users and results regarding usability, understandability and reliance show no reason to assume that there are major flaws in the design.

### 8.1 Understandability

In terms of understandability, it can be concluded that most interface components were well designed for the target user. This was according to expectation, as the inspectors' expertise level regarding information visualisation was strongly taken into account during the entire process. Only the scatterplot, which can be seen as a more complex visualisation, was found to be difficult to understand. This was also anticipated to a certain degree, given the non-statistical background of the inspectors. Yet, the plot was sketched by three participants and rated as an important component in one of the focus groups. Another argument for adding the plot to the interface, was because it provided an alternative route for getting an overview of the data, next to the tabular overview. This is in line with the design guideline that users should be provided information in multiple forms (Chromik & Butz, 2021). On the one hand, the fact that the inspectors found the scatterplot difficult to understand may have been influenced by the short duration of the experiment, as they may not have had the time to discover its value. On the other hand, findings of the observations indicate a pattern in which the inspectors often tended to focus on a local scope of the data. In the beginning of design phase, it was already decided to give the inspectors a local prediction explanation only, as the global explanation was expected to be too abstract for them. Alternative approaches, such as the scatterplot, distributions charts and similar ships page, aimed to give the inspectors some contextual understanding of the data. Still, they seemed to be mainly interested in local views. Both the scatterplot and similar ships page were not considered very relevant. About the distribution charts, which gave an overview of a specific feature in

the dataset, a participant mentioned that it would only be relevant if they contain ships of the same type only, which also comes down to narrowing his scope. These findings are confirmed by Liao et al. (2020), suggesting that users with an AI or analytic background are indeed more likely to seek global explanations. The inspectors' behaviour is thought to be caused by their traditional way of working, in which there is generally a strong focus towards assessing individual ships or companies. For them, taking a more global perspective is new and it might take more time and practice to find out if and why that information could be relevant to them. As global explanations of XAI systems are thought to support users' understandability and therefore usability of those systems (Alam, 2020; Liao et al., 2020), future work should explore how global information could support end-users effectively.

## 8.2 Usability

In terms of usability, the scores of two participants indicate an acceptable level, while those of the other two are just under the threshold of what is considered acceptable. Examination of their responses in the questionnaire may have revealed an explanation for these lower scores. Participant 1 in particular seems to have answered the questions from a broader view than usability only. For example, he answered to "Agree" on the statement "I think I need technical support for using this product". This is not consistent with the observations during the experiment regarding usability, as the participant seemed to be able to use the interface without issues. Yet, he did encounter problems in his expectations of certain prediction explanations, which may have influenced his responses. The participant might have meant to need technical support for *interpreting* the product's presented information, but not for *using* it. It is thought that it was not stated clear enough that the statements in the SUS were specifically meant to be assessed on the level of usability, or that the phrasing of the statements might have been ambiguous. Overall, looking at the usability scores and the observations from the task-session together, it is fair to conclude that no concerning issues were found. Although participants experienced some navigational issues, it is expected that these will subside in long term usage after spending more time with the interface. Still, there seems to be some room for improvement, especially in the placement of the company page and easier switching between ship pages.

## 8.3 Trust in Technology

In terms of trust, results indicate that all participants have a low to medium level of propensity to trust technology. One participant stands out with a score of only 2, which could be related to his older age and him mentioning to be more "old-fashioned" than his average colleague. The outcome suggests that the participant is in general more sceptical towards the use of technology, which could mean that he is less keen on incorporating the models and dashboard in his work. This was reflected in the task session, in which he took a rather traditional viewpoint regarding the ship assessments and seemed not to be able to really implement the prediction scores and explanations in his decision making process. Over the longer term, it is expected that this problem will

subside, as inspectors will become younger and predictive models will be used more and more in their daily work. For now, it poses challenges to reform their current way of working, which starts by developing a usable and understandable product.

## 8.4 Reliance

In terms of reliance, four things stand out. Firstly, the overview table was ranked as the most essential component by all participants, which was in line with expectations. The table provides the inspectors a quick overview of the ships that need most attention and the searching, sorting and filtering functionalities provide ways for browsing through the dataset. Therefore, it makes up the starting point of their work process. Also noteworthy are both the explanation components, which were ranked second and third. This suggests that inspectors value the explanations and need the information for their decision-making. As the model explanations deserve a more extensive discussion, it will be addressed in the next paragraph. The third notable remark concerns the two charts, being the distribution chart and scatterplot, which were ranked fairly low. Regarding the scatterplot, this follows the earlier comments regarding understandability. However, the distribution charts were found to be understandable and its contents relevant while participants still ranked the component as non-essential. On top of the aforementioned tendency towards a local scope, this is thought to be due to a general hesitance towards using charts, especially since all other information in the interface is presented in a different way that is more intuitive for the inspectors. Perhaps the addition of a textual alternative, in line with the aforementioned design principles (Chromik & Butz, 2021), would make the information more accessible. The final notable remark concerns the similar ships page, which was ranked the lowest and was not even included in the ranking of three participants. Again, the local scope tendency can play a role here. However, during the design phase it was thought that adding this page would be very useful for the inspectors, even though it did not directly fulfil one of the user goals. As with the scatterplot, long-term use may make inspectors realise the potential and value of the similar ships page. For now, however, it can be concluded that it would have been better to stick more closely to the user goals, as these evidently have accurately reflected what the inspectors want.

## 8.5 Visualisation of Model Explanations

For the model explanations, as main component of the interface, a custom design was proposed that specifically targeted novice users. The design differs from traditional visualisations in four ways: it provides reader-friendly textual descriptions of the features, it includes visual and textual hints about how to interpret the information, the raw feature values are represented by sizes of coloured bars, rather than numbers, and it aims to avoid information overload by presenting only the most important features.

Whereas the study by Haas (2021) found that inspectors experienced difficulties with interpreting the visual meanings of the industry-standard SHAP plots, the current study shows that inspectors mainly commented on the contents of the explanations, instead of the visual aspects. The absence of remarks on the visualisation could be a

sign that the proposed visualisation of the model explanations was effective in terms of visual understandability and that it has taken the problem of explaining XAI predictions to novices towards the next phase.

This next phase refers to the challenges regarding interpretability of the explanations and trust in the model. As observed during the tasks and resulting from the reliance questions, the inspectors seemed intrinsically motivated to gain an understanding of the reasons behind the predicted scores. Also, inspectors slightly preferred these visual explanations over the text-only alternative, which indicates that they are interested in the details of the explanation. These are positive observations, as one of the main motives for making AI explainable is because we do not want users to blindly rely on ML predictions (Ming, 2017). However, the findings also show that inspectors struggled to grasp the meanings of the presented prediction reasons. This mainly concerns the internals of the model and is therefore, for the most part, beyond the scope of this study. However, as it has strong implications on both the perceived understandability, trust and reliance, it is an important point of discussion. The observations in the user experiment indicate that the problem can be categorised in two groups, either (1) not being able to *interpret* the information or (2) *expecting* a different explanation. The first category seemed to apply to the three participants who were introduced to the models for the first time at the beginning of the experiment session. It is not surprising that they lacked the skills to fully comprehend what the model explanations represent, which can be described as a mismatch between the inspectors' *mental model* and the system. This concept refers to the mental representation of a system and its functioning, which "allow humans to mentally simulate aspects of a system, for instance, in order to understand the causes of its decision-making" (Langer et al., 2021). Good mental models are expected to lead to the development of appropriate trust and a better performance in using AI (Hoffman et al., 2018). There is no straightforward solution for matching the user's mental model to a system. On the one hand, mental models evolve as users interact with the model (Eiband et al., 2018) and users can be educated to get a better understanding of the system, e.g. by explaining underlying ML concepts and teaching about the strengths and limitations of the models. On the other hand, it could mean that the XAI model itself needs to be changed to align it to the mental models of the ship inspectors. A relevant approach to this is described by Eiband et al. (2018), which starts by mapping the mental model of the end-users, defining the desired mental model and through iterative prototyping and evaluation tailor the explanations until the desired mental model is reached. Applying this to the shipbreaking and beaching models, it could mean that features that fail to become recognised or interpretable by inspectors, could better be removed or explained differently.

In contrast to the problem of *interpretability*, there was one inspector who was already familiar with the models, had seen some explanations before and had already discussed these with the model developers. He seemed to have a better mental model and was able to interpret the results, but in some cases expected a different outcome. As he pointed out himself, this led to a distrust in the models. A similar observation is



described by Langer et al. (2021), stating that users with a lower degree of understanding of the AI system will be more likely to trust it (although inadequately), while a higher degree of understanding could expose awareness of the system's problematic features or limitations, which eventually may decrease trust in the system. Hence, people with less understanding are likely to have problems *interpreting* the information, whereas with better educated users *trust* may be the main obstacle for acceptance of the system. Nevertheless, distrust may not be necessarily problematic, as the actual goal of the explanations is to provide the handles for assessing the predictions manually, and not to ensure that predictions are always consistent with the user's expectations. After all, the model's predictions can also be inaccurate. The challenge regarding trust is captured in the concept of 'appropriate trust', stating that for proper use of automation, users should neither *overrely* nor *underrely* on automated systems (Arrieta et al., 2019; Langer et al., 2021). Further research on long-term system usage is needed to find out how the explanations can be optimised for an appropriate level of trust.

## 8.6 Research Approach

In terms of research methodology, it can be concluded that the approach was effective. The research followed the stages in the design study framework by Sedlmair et al. (2012). It started with getting to know the topic, the target audience and their needs, which was then used for the generation of design solutions to be implemented and then evaluated with the target users. This structured approach provided an organised top-down pipeline, in which the results of one stage were used as input for the next while iterating to a more scoped result in every step. Also, the approach pushed the focus of this study towards the characteristics and needs of the target audience in each stage, which ensured the end result to be well tailored to its target users.

## 8.7 Limitations

The main limitation of this study is the short time of deployment. As described in the design study framework (Sedlmair et al., 2012), the implemented prototype should preferably be deployed for several months in order to get the highest quality of feedback on a design proposal. As this was not feasible within the total time span of this thesis study, the deployment stage was reduced to a single experiment session. A drawback of this, is that it could not be fully investigated how users interpret the explanation results and how the presented information will actually be used in their work. However, for the relatively short amount of time that the users worked with the interface, it is already encouraging that it was mostly perceived as an intuitive and understandable interface, which is expected to further improve when users will be using the interface more regularly.

The second limitation was the small pool of target users. As illegal shipbreaking is not something that occurs on a daily basis with ships of Dutch companies, only few inspectors are working on the topic of illegal shipbreaking. This was clear at the beginning of this study and it was therefore decided to involve the inspectors only in the two main stages of the project. First, during the requirements analysis, which laid the

groundwork for the designs and for which it was therefore essential to involve the target users to understand their work and their goals. Second, during the final evaluation, to discover whether the designs effectively matched the users' level of knowledge and their needs. Retrospectively, looking at the highly overlapping results between the participants, it can be concluded that these decisions worked out well.

## 8.8 Recommendations for Future Research

The study results and limitations leave a number of opportunities for future work.

Firstly, more research should be done to the problem that was raised regarding interpretability of the explanation content. The assumption was made that users will get better at grasping the meaning of the model explanations over time (aligning their mental model), after applying them in their field of work and be taught about the underlying concepts of ML. On top of that, the models themselves might need to be altered to make them more interpretable. A similar and overlapping research opportunity lies in the second problem that was discussed, regarding users' expectations of models and gaining a level of appropriate trust. Since the current study could not examine these problems in depth, further research should focus on the obstacles for this type of users and how ML models can be adapted to make them more interpretable and trustworthy.

Furthermore, although the current study showed that the proposed designs were effective in terms of visuals, future research should compare them with alternative designs for further optimisation. Examples are the visualisation of the prediction scores, the local feature importance and more advanced explorational functionalities such as the scatterplot or distribution charts for providing a more global scope.

Also, the visualisation of feature importance was specifically designed for the use case of illegal shipbreaking, while other use cases using XAI models with explanations in the form of SHAP values exist. It is expected that the visualisation would also be effective in other domains with the same type of target users, but further research should be done to prove this.

Finally, the next step is to go more in the direction of *machine teaching*; which refers to giving input back to the model (teaching the model how to behave). This would bring the research a step closer to the field of Visual Analytics, and it is expected to positively impact the creation of mental models and improve building a level of appropriate trust. In the current research, there was already a suggestion for this in the form of giving feedback on predictions, but future research should investigate how to design it effectively.

## 8.9 Implications for the IDLab

Because of the positive results, the prototype is going to be used in the pilot phase of ILT's project; meaning that it will be tested in the field. Additionally, it will be investigated how certain components from the interface can be reused in other domains of the ILT. These exciting developments are expected to yield new interesting findings.

## Chapter 9

# Conclusion

In this research, an Explainable Artificial Intelligence web interface was designed aiming to produce more understandable and usable explanations of ML systems for end-users. The study was conducted through the use case of visualising predictive models for the prevention of illegal shipbreaking. For this, the following research question was formulated: *"How can data visualisations of XAI models effectively support ship inspectors with preventing illegal shipbreaking?"* The methods used for answering the question followed a user centered top-down approach, including a literature review, requirements analysis with stakeholders, brainstorm and sketching session, low-fidelity prototypes, focus group sessions, the implementation of a high-fidelity prototype and a final experiment with the target users. The resulting prototype was evaluated in terms of understandability, usability, trust and reliance.

### 9.1 Main findings

Overall, the results from this study are promising and provide guidance on how XAI interfaces for end-users can be designed effectively. From the evaluation, the following main conclusions can be drawn:

- *The structured approach towards a final interface prototype proved to be effective.* The top-down research approach, with a strong focus on the characteristics and needs of the target audience, proved to be effective, as the evaluation results indicate that the final prototype was well tailored to its target users.
- *Ship inspectors perceive an acceptable understandability when working with the proposed interface.* Although the inspectors found it hard to understand the scatterplot, most other interface components were found to be understandable. Interestingly, the inspectors seemed to be especially focused on a local scope of the available information, suggesting a more global scope to be less intuitive and more difficult to understand.
- *Ship inspectors perceive an acceptable usability when working with the proposed interface.* Even though there were some minor navigational issues, the inspectors liked the interface and were able to use it intuitively.

- *Ship inspectors showed a consistent reliance on the individual components of the proposed interface.* Similar reliance rankings were found for all participants, indicating that the overview table of all ship predictions was most important. Second, the model explanations, both visual and textual. The only two charts in the interface were considered hardly relevant and the similar ships page was found to be least relevant by almost all inspectors.
- *The proposed design for local feature importance was effective in terms of visuals.* A custom design was proposed for visualising the local feature importance values, making it more accessible for novice users compared to existing visualisations. The evaluation results indicate that the inspectors were able to well understand the visual aspects of the design.
- *Ship inspectors experienced difficulties regarding the interpretability of the local explanation contents.* Although the ship inspectors seemed to be intrinsically motivated to gain an understanding of the model's predictions, they found it difficult to interpret the explanations. It is thought to be caused by a mismatch between the inspectors' mental model and the system. From the inspectors' perspective, this is expected to improve after using the system more and through education about the underlying ML concepts. However, the ML models might need to be changed to align them to the mental models of the inspectors.

# Bibliography

- Alam, L. (2020). Investigating the impact of explanation on repairing trust in ai diagnostic systems for re-diagnosis.
- Argüello Moncayo, G. . (2016). International law on ship recycling and its interface with eu law. *Marine Pollution Bulletin*, 109(1), 301–309. <https://doi.org/10.1016/j.marpolbul.2016.05.065>
- Arrieta, A. B., Rodriguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *CoRR*, *abs/1910.10045*. <http://arxiv.org/abs/1910.10045>
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Benyon, D. . (2013). *Designing interactive systems: A comprehensive guide to hci, ux interaction design, 3rd ed.* (Comprehensive). Trans-Atlantic Publications, Inc.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3351095.3375624>
- Choo, J., & Liu, S. (2018). Visual analytics for explainable deep learning. *CoRR*, *abs/1804.02527*. <http://arxiv.org/abs/1804.02527>
- Chromik, M., & Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. *Human-computer interaction – INTERACT 2021* (pp. 619–640). Springer International Publishing. [https://doi.org/10.1007/978-3-030-85616-8\\_36](https://doi.org/10.1007/978-3-030-85616-8_36)
- Claeys, J., & Bisschop, L. (2018). Een schip op het strand is een baken in zee. *Tijdschrift voor Criminologie*, 60(1), 3–33. <https://doi.org/10.5553/TvC/0165182X2018060001001>
- Confalonieri, R. ., Coba, L. ., Wagner, B. ., & Besold, T. R. (2020). A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1). <https://doi.org/10.1002/widm.1391>
- Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable AI meets healthcare: A study on heart disease dataset. *CoRR*, *abs/2011.03195*. <https://arxiv.org/abs/2011.03195>

- Demaria, F. (2010). Shipbreaking at alang–sosiya (india): An ecological distribution conflict [Special Section: Ecological Distribution Conflicts]. *Ecological Economics*, 70(2), 250–260. <https://doi.org/https://doi.org/10.1016/j.ecolecon.2010.09.006>
- Digitale Overheid. (2020, October 9). *Overzicht data-initiatieven*. Retrieved January 13, 2022, from <https://www.digitaleoverheid.nl/initiatief/id-lab/>
- Du, M. ., Liu, N. ., & Hu, X. . (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
- Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. *23rd International Conference on Intelligent User Interfaces*. <https://doi.org/10.1145/3172944.3172961>
- Ekhart, N. (2022). *Taking down malicious webshops: designing Explainable AI against growing e-commerce fraud* (tech. rep.).
- Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). Explainable AI in industry. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. <https://doi.org/10.1145/3292500.3332281>
- Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable ai: The new 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction* (pp. 295–303). Springer International Publishing.
- Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *CoRR*, *abs/2107.07045*. <https://arxiv.org/abs/2107.07045>
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.
- Haas, C. d. (2021). Usability study of an explainable machine learning risk model for predicting illegal shipbreaking.
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *CoRR*, *abs/1812.04608*. <http://arxiv.org/abs/1812.04608>
- Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8), 2674–2693. <https://doi.org/10.1109/tvcg.2018.2843369>
- Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. . (2019). The measurement of the propensity to trust automation. *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, 476–489. [https://doi.org/10.1007/978-3-030-21565-1\\_32](https://doi.org/10.1007/978-3-030-21565-1_32)
- Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (1996). *Usability evaluation in industry*. CRC Press.
- Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. *Lecture notes in computer science* (pp. 154–175). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-70956-5\\_7](https://doi.org/10.1007/978-3-540-70956-5_7)

- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*. <https://doi.org/10.1145/2678025.2701399>
- Kumar, A., Dikshit, S., & Albuquerque, V. H. C. (2021). Explainable artificial intelligence for sarcasm detection in dialogues (M. R. Khosravi, Ed.). *Wireless Communications and Mobile Computing, 2021*, 1–13. <https://doi.org/10.1155/2021/2939334>
- Lamsweerde, V. . A. . (2009). Fundamentals of requirements engineering. *Requirements engineering: From system goals to uml models to software specifications* (1st ed., pp. 3–60). Wiley.
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? – a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Liao, Q. V., Gruen, D. M., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. *CoRR, abs/2001.02478*. <http://arxiv.org/abs/2001.02478>
- Litehauz. (2015, October). *Intertidal zone study*. Copenhagen: Litehauz Maritime Environmental Consultancy.
- Lundberg, S. (2019, November 6). *Github - slundberg/shap: A game theoretic approach to explain the output of any machine learning model*. Retrieved February 10, 2022, from <https://github.com/slundberg/shap>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749.
- McGovern, A. ., Lagerquist, R. ., John Gagne, D. ., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. . (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society, 100*(11), 2175–2199. <https://doi.org/10.1175/bams-d-18-0195.1>
- Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. *Artificial intelligence in HCI* (pp. 54–69). Springer International Publishing. [https://doi.org/10.1007/978-3-030-50334-5\\_4](https://doi.org/10.1007/978-3-030-50334-5_4)
- Ming, Y. (2017). A survey on visualization for explainable classifiers. *Hong Kong*.

- Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., & Qu, H. (2017). Understanding hidden memories of recurrent neural networks.
- Ministerie van Infrastructuur en Waterstaat. (2022, January 3). *About the ilt*. <https://english.ilent.nl/about-the-ilt>
- Mohseni, S. ., Zarei, N. ., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4), 1–45. <https://doi.org/10.1145/3387166>
- Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*.
- Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 921–928. <https://doi.org/10.1109/tvcg.2009.111>
- NGO Shipbreaking Platform. (2019, February 25). *Flags of convenience*. Retrieved January 13, 2022, from <https://shipbreakingplatform.org/issues-of-interest/focs/>
- NGO Shipbreaking Platform. (2020, February 10). *The problem*. Retrieved January 11, 2022, from <https://shipbreakingplatform.org/our-work/the-problem/>
- Nielsen, J. . (1994). *Usability engineering (interactive technologies)* (1st ed.). Morgan Kaufmann.
- Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128–138.
- Preece, A. D., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. *CoRR*, *abs/1810.00184*. <http://arxiv.org/abs/1810.00184>
- Pribeanu, C. (2017). A revised set of usability heuristics for the evaluation of interactive systems.
- Ribeiro, M. T., Singh, S. ., & Guestrin, C. . (2016a). "why should i trust you?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2939672.2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Model-agnostic interpretability of machine learning.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
- Sarraf, M., Stuer-Lauridsen, F., Dyoulgerov, M., Bloch, R., Wingfield, S., & Watkinson, R. (2010). The ship breaking and recycling industry in bangladesh and pakistan. <https://doi.org/10.11588/XAREP.00003749>
- Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12), 2431–2440.
- Souza, J., & Leung, C. K. (2021). Explainable artificial intelligence for predictive analytics on customer turnover: A user-friendly interface for non-expert users. *Explainable AI within the digital transformation and cyber physical systems* (pp. 47–67).



- Springer International Publishing. [https://doi.org/10.1007/978-3-030-76409-8\\_4](https://doi.org/10.1007/978-3-030-76409-8_4)
- Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. <https://doi.org/10.1109/tvcg.2019.2934629>
- Stopford, M. (2008). *Maritime economics 3e*. Routledge.
- Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. *26th International Conference on Intelligent User Interfaces*.
- UNCTAD. (2021). *Merchant fleet by country of beneficial ownership, annual*. Retrieved January 14, 2022, from <https://unctadstat.unctad.org/wds/tableView/tableView.aspx?ReportId=80100>
- Waste Shipment Regulation. (2006, June 14). *The european union waste shipment regulation. no 1013/2006*. [https://ec.europa.eu/environment/topics/waste-and-recycling/waste-shipments\\_en](https://ec.europa.eu/environment/topics/waste-and-recycling/waste-shipments_en)
- Wehbe, R. M., Sheng, J. ., Dutta, S. ., Chai, S. ., Dravid, A. ., Barutcu, S. ., Wu, Y. ., Cantrell, D. R., Xiao, N. ., Allen, B. D., MacNealy, G. A., Savas, H. ., Agrawal, R. ., Parekh, N. ., & Katsaggelos, A. K. (2021). Deepcovid-xr: An artificial intelligence algorithm to detect covid-19 on chest radiographs trained and tested on a large u.s. clinical data set. *Radiology*, 299(1), E167–E176. <https://doi.org/10.1148/radiol.2020203511>

## **Appendix A**

# **Inspector Interview Questions**

## **A.1 Mapping the general process steps**

### **A.1.1 Expectation of the process, based on prior meetings**

- Inspector receives a signal about a ship that has potential for being illegally dismantled
- Inspector researches the ship for making a decision for visiting the shipping company
- Inspector visits the shipping company, with main goal to prevent the ship of being illegally dismantled
- Inspector monitors the ship or shipping company, to intervene if necessary

### **A.1.2 Questions**

1. Can you explain what the work of a shipbreaking inspector involves?
2. Present the 4 steps as described above (possibly adapted in response to previous question); ask if this is correct and if there are any additions.

## **A.2 Diving deeper into each process step**

### **A.2.1 Signaling**

Currently, inspectors are mainly driven by signals from their network.

3. What could these 'network' sources be?
4. What does the information they provide look like? What do or can they say?
5. Is the information they provide focused on a shipping company, a specific ship or a group of ships?
6. How specific is the information given? Is it based on rumours, or is there already evidence?
7. How do these sources get their information?
8. Are there ways to be signalled outside the network? Do you monitor parameters yourself to find out which ships need further attention?

9. How do you as inspectors communicate this information to each other?

### **A.2.2 Investigation**

10. After there has been a signal to further investigate a particular ship, what does an inspector do?
11. Which sources are consulted for this?
12. Which information is important for this / which ship features are being investigated?
13. How is the decision made to visit a shipping company? What is ultimately decisive for this decision?

### **A.2.3 Visiting**

14. What does a visit to a shipping company look like?
15. What is the main purpose of such a visit?
16. What influence can you exert as an inspector and how does that determine which information you want to provide and in what form?
17. What information do you give to a shipping company about the reason for the visit? / How do you explain the reason for the visit to the shipping company?
18. What if an inspector decides NOT to visit a shipping company? What does the situation look like then? What are the next steps?

### **A.2.4 Monitoring**

19. What do inspectors do after visiting a shipping company?
20. Is there a structural / generic approach, or is it different for every shipping company/ship?

### **A.2.5 Closing**

21. Are there things we haven't discussed yet, but are important to understand what the process looks like?
22. Is there anything else you would like to share with us?

## Appendix B

# Brainstorm Results: Thoughts on User Goals

**User goal 1: Inspectors want to be able to find ships that have the potential of being illegally beached in the future (signaling)**

### Components / functionalities

- Prediction score
- A “pop-up” notification (where to receive it? what info to be included in the notification already?)
- Open search bar to find specific ships
- Ranking feature (e.g. rank ships by age)
- Filtering feature (e.g. filter view on specific ship types)

### Visualization ideas

- Geographic filtering (is there a high risk ship near me?)
- Show me on a map literally where these ships currently are
- Signaling out of a context: table/list, map, thumbnails

**User goal 2: Inspectors want to understand why the system thinks a ship has a low or high risk of being beached (investigation / explanation)**

### Components / functionalities

- Interactions between features? How to explain? (SHAP interactions?)
- SHAP values
- Clear connection of SHAP values to ship features
- Most important features for an individual decision. Then: how many features to show?
- Prediction history
- Being able to find outliers by placing ships into context

### Visualization ideas

- Bar chart visualization
- Visually highlight ship parts / opaqueness

- Textual explanation vs visual
- Explain by comparison to other ships

### **Challenges**

- How to cope with categorical features? Group them?
- How heavily should an inspector rely on this score? Should this be presented?
- How sure is the system? → Confusion matrix for model performance
- Why is the score different/similar to their 'expectation'?

## **User goal 3: Inspectors want to be able to explain to others why a ship has a low or high risk of being beached (explanation / re-production)**

### **Components / functionalities**

- Making sure visualizations are understandable enough so they can describe it in their own words
- Purpose of explaining to others: accountability, proving it is real, etc.
- Combination of own expertise which is supported by the data

### **Visualization ideas**

- By providing explanation in natural language, or at least both text and visuals

### **Challenges**

- Should we explain (not illegal) shipbreaking or beaching?
- To whom? Colleagues, other governmental shipbreaking agencies.
- On a local or global level? Or both?

## **User goal 4: Inspectors want to be able to see additional general characteristics of a ship (investigation)**

### **Components / functionalities**

- Use it to provide context (show relation to other ships)
- An easy way to show/hide features

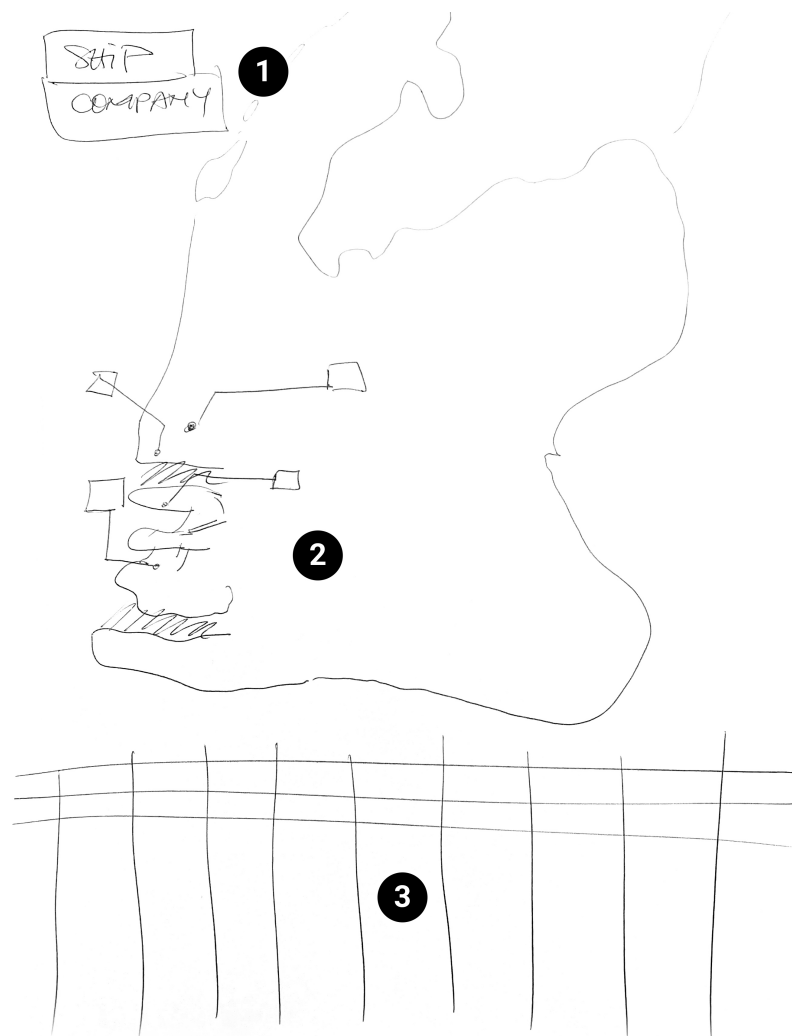
### **Visualization ideas**

- Links to external websites (e.g. marinetraffic)
- Company profile (e.g. reputation, or log of all ships from the company)
- Satellite images
- Remote sensing (drones)
- Societal / market trends?
- Trajectories of ships on a map

## Appendix C

# Sketching Results

### Category: Instance Overview / Selection



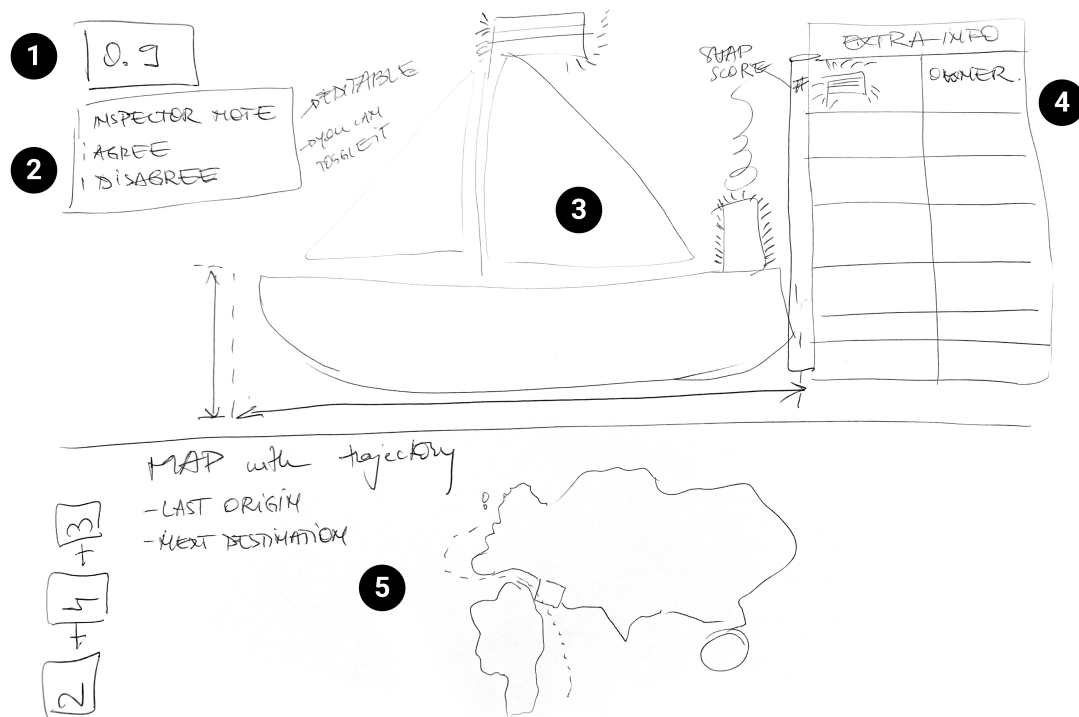
**Figure C.1:** Idea for the overview page. (1) Buttons for changing the overview of either ships or companies. (2) Map with an overview of ship/company locations. (3) Tabular view with the same instances as shown in the map, with general information about the instances and the prediction details.

A hand-drawn sketch of a table interface. At the top left is a box containing a funnel icon and the word "filter". At the top right is a box containing the text "Search Q". Below these is a table with two columns: "Ship" and "score". The table contains eight rows. The "Ship" column has various wavy lines representing ship names. The "score" column has the values "x", "y", "2", "x", "y", "2".

Ship	score
_____	x
_____	y
_____	2
_____	x
~~~~~	y
~~~~~	2
~~~~~	

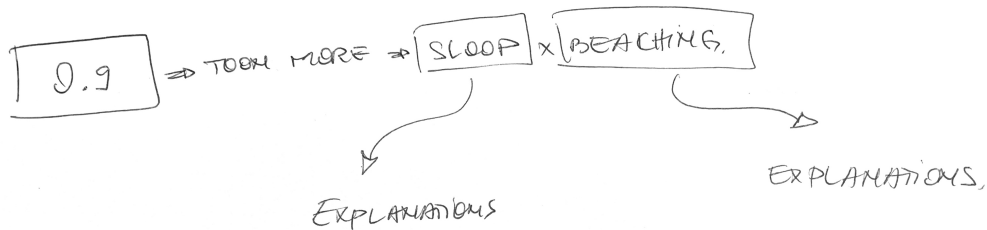
**Figure C.2:** Simple instance selection table, including at least ship name and prediction scores. Supplemented by filtering and searching functionalities.

## Category: Detail Page / Explanations

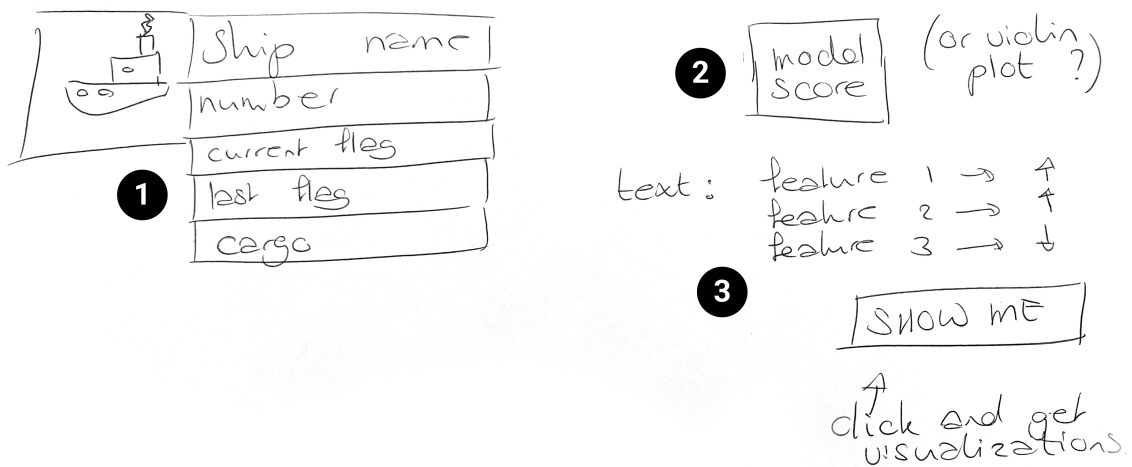


**Figure C.3:** Idea for the detail page, after selecting a specific instance. (1) The final score of the selected ship, presented as a score between 0 to 10 instead of 0 to 1. (2) Feature for adding feedback to the system, which could be given to the model developers, or eventually even the model itself. (3) A visual illustration of the ship, with all the values and ship details on it, e.g. the ship's height, length, engine types, flag. (4) A list of ship features, sorted by their feature importance score. (5) A map with a time-line of trajectories.

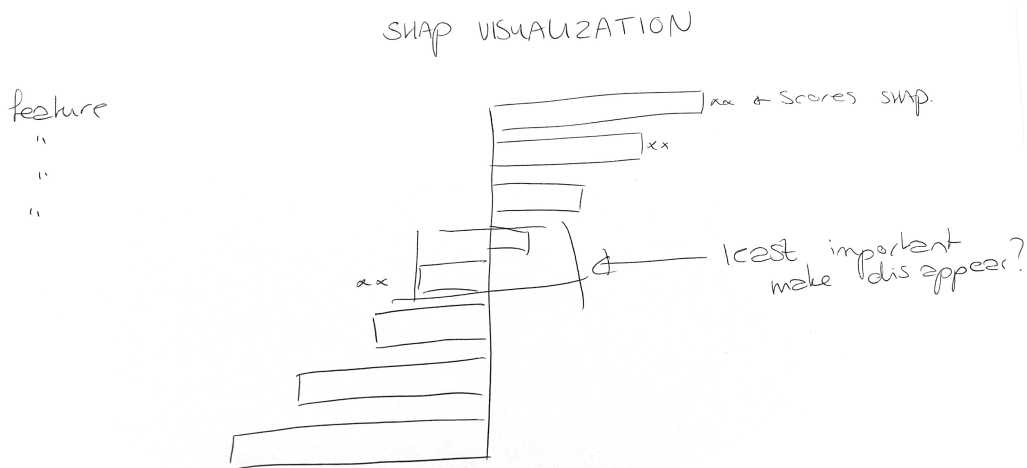




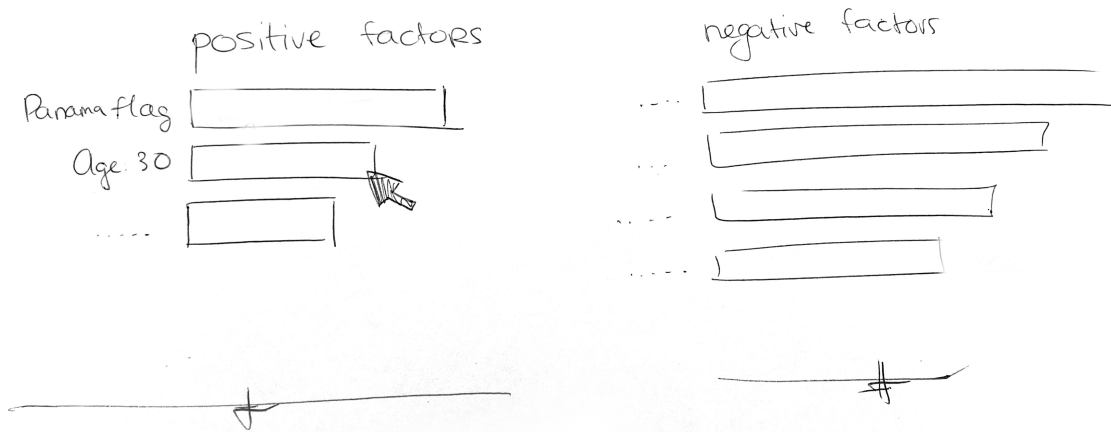
**Figure C.4:** Score visualisation: final score with a hover/click action to show how the scores of the two underlying models.



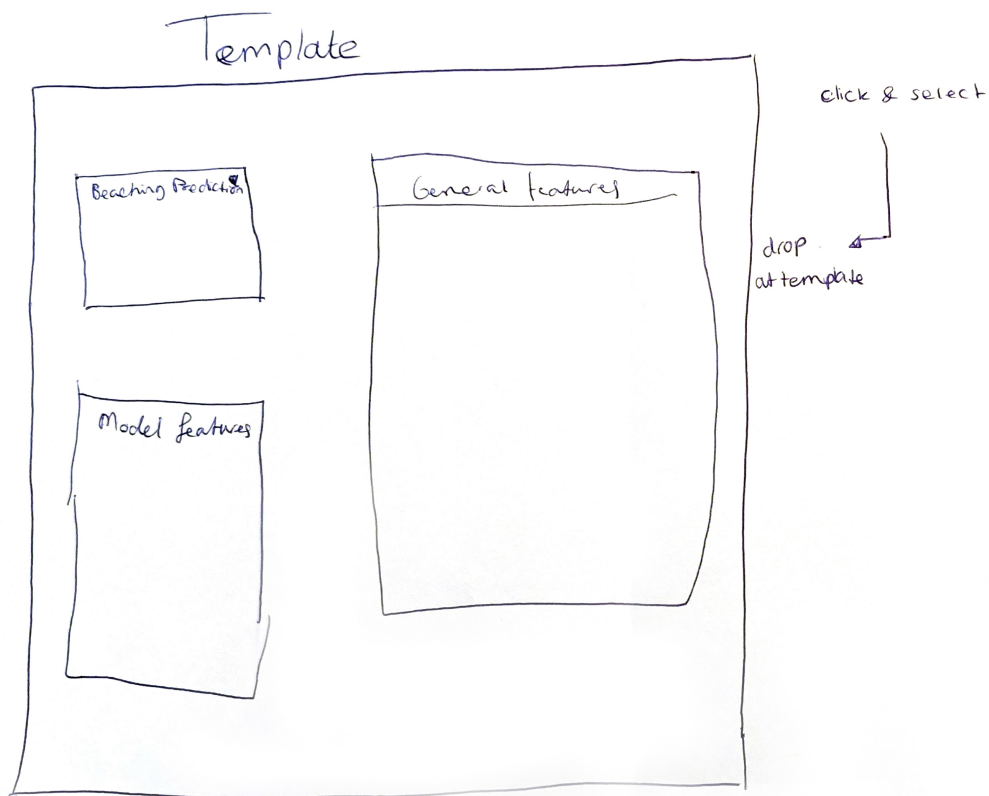
**Figure C.5:** An idea for the detail page. (1) Summary of most important ship details, including an image of the ship. (2) Presentation of the model score. (3) A textual list of three most important features, supplemented by a button triggering a visual representation of all the feature importances.



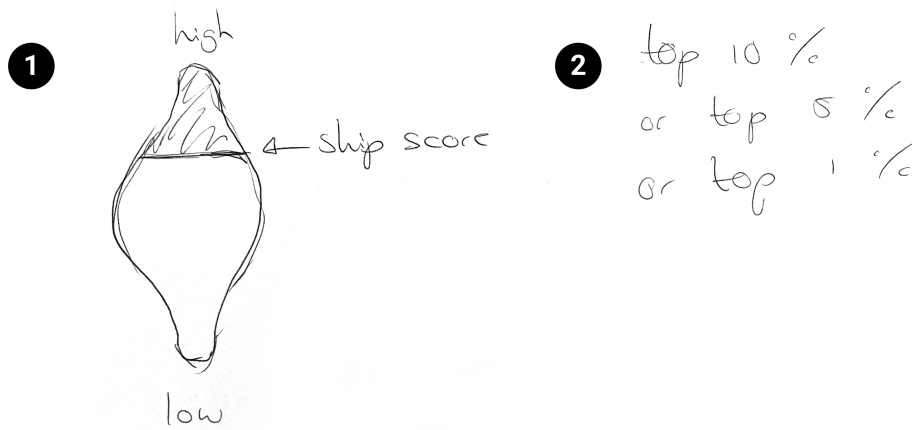
**Figure C.6:** A visualisation idea for the feature importances, using a bar chart.



**Figure C.7:** Alternative idea for displaying the feature importances as lists with bars. The positive and negative factors are separated, and only the X most important features are initially shown. Clicking the "plus" button will expand the list with all features.



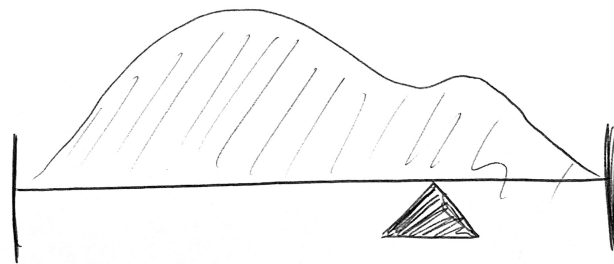
**Figure C.8:** Personalisation idea for the detail page: giving users the ability to arrange their own dashboard views with the information they need.

**Category: Context**

**Figure C.9:** Providing context to scores, (1) using a violin plot to visualise the distribution of scores. Additionally, provide (2) a textual explanation to indicate how high or low a score is compared to others.



**Figure C.10:** Scatterplot to provide an overview of all ships, with the beaching model and shipbreaking model scores as axes.



**Figure C.11:** Chart for visualising the distribution of a ship's feature (e.g. age), providing context through marking the value of the currently selected ship.

## Appendix D

# Low-Fidelity Prototypes

All the data that is included in the prototype screens, such as images, ship names and owner names, is fictitious.

Overzicht schepen

Tabel kolommen Filteren Zoek op naam of IMO...

Naam	Type	Leeftijd	Huidige vlag	Rederij	Sloopscore	Beachingscore	Eindscore
PRINCESS II	Vrachtschip	9,9 jaar	Nederland	Luckybank Tankers	9,1	9,6	8,7
STORSLAGEN	Vrachtschip	8,6 jaar	Nederland	Janssen Shipmanagement BV	7,4	9,8	7,2
ADEPT	Containerschip	8,3 jaar	Nederland	Shipping Skyr BV	7,0	9,0	6,3
ENZTRUMMA	Tanker	12,3 jaar	Nederland	ED Ship Management BV	6,9	9,1	6,3
BLIXT 1	Vrachtschip	11,4 jaar	Nederland	De Jong Shipping BV	6,4	9,7	6,2
SMILAX	Vrachtschip	24,5 jaar	Nederland	Mermaid Carriers AS	6,4	9,7	6,2
HUNDRA	Vrachtschip	7,5 jaar	Singapore	Norwegian Navigation BV	6,8	9,1	6,2
REDNAS	Vrachtschip	7,2 jaar	Nederland	Schiffahrts Markus Zeus	6,8	9,1	6,2
SEPTEMBRE	Vrachtschip	10 jaar	Nederland	Cargoship BV	7,4	8,2	6,1
CHUBB	Vrachtschip	11,3 jaar	Nederland	SPAX Shipping BV	7,1	8,5	6,0
PEARL PRIZE	Vrachtschip	11,8 jaar	Noorwegen	Overfind Ships BV	6,7	9,0	6,0
BUTTERFLY	Vrachtschip	8,1 jaar	Turkije	JTB Maritime Pte Ltd	6,2	9,7	6,0
ENCHANCE	Tanker	15,6 jaar	Singapore	SPAX Shipping BV	6,3	8,7	5,5

Eerste < 1 2 3 4 > Laatste

1-8 items van 1244

Figure D.1: Large overview table with functionalities for selecting the table columns, filtering and searching in the data and pagination.

Overzicht schepen

Tabel kolommen Filteren Zoek op naam of IMO...

Naam	Type	Leeftijd	Huidige vlag	Rederij	Sloopscore	Beachingscore	Eindscore
PRINCESS II	Vrachtschip	9,9 jaar	Nederland	Luckybank Tankers	9,1	9,6	8,7
STORSLAGEN	Vrachtschip	8,6 jaar	Nederland	Janssen Shipmanagement BV	7,4	9,8	7,2
ADEPT	Containerschip	8,3 jaar	Nederland	Shipping Skyr BV	7,0	9,0	6,3
ENZTRUMMA	Tanker	12,3 jaar	Nederland	ED Ship Management BV	6,9	9,1	6,3
BLIXT 1	Vrachtschip	11,4 jaar	Nederland	De Jong Shipping BV	6,4	9,7	6,2
SMILAX	Vrachtschip	24,5 jaar	Nederland	Mermaid Carriers AS	6,4	9,7	6,2
HUNDRA	Vrachtschip	7,5 jaar	Singapore	Norwegian Navigation BV	6,8	9,1	6,2
REDNAS	Vrachtschip	7,2 jaar	Nederland	Schiffahrts Markus Zeus	6,8	9,1	6,2
SEPTEMBRE	Vrachtschip	10 jaar	Nederland	Cargoship BV	7,4	8,2	6,1
CHUBB	Vrachtschip	11,3 jaar	Nederland	SPAX Shipping BV	7,1	8,5	6,0
PEARL PRIZE	Vrachtschip	11,8 jaar	Noorwegen	Overfind Ships BV	6,7	9,0	6,0
BUTTERFLY	Vrachtschip	8,1 jaar	Turkije	JTB Maritime Pte Ltd	6,2	9,7	6,0
ENCHANCE	Tanker	15,6 jaar	Singapore	SPAX Shipping BV	6,3	8,7	5,5

Eerste < 1 2 3 4 > Laatste

1-8 items van 1244

Figure D.2: Similar visualisation to figure D.1, but with coloured final scores.

**Overzicht schepen** Tabel kolommen Filteren

Naam	Type	Leeftijd	Huidige vlag	Rederij	Sloopscore	Beachingscore	Eindscore
PRINCESS II	Vrachtschip	9,9 jaar	Nederland	Luckybank Tankers	9,1	9,6	8,7
STORSLAGEN	Vrachtschip	8,6 jaar	Nederland	Janssen Shipmanagement BV	7,4	9,8	7,2
ADEPT	Containerschip	8,3 jaar	Nederland	Shipping Skyr BV	7,0	9,0	6,3
EN2TRUMMA	Tanker	12,3 jaar	Nederland	ED Ship Management BV	6,9	9,1	6,3
BLIXT 1	Vrachtschip	11,4 jaar	Nederland	De Jong Shipping BV	6,4	9,7	6,2
SMILAX	Vrachtschip	24,5 jaar	Nederland	Mermaid Carriers AS	6,4	9,7	6,2
HUNDRA	Vrachtschip	7,5 jaar	Singapore	Norwegian Navigation BV	6,8	9,1	6,2
REDNAS	Vrachtschip	7,2 jaar	Nederland	Schiffahrts Markus Zeus	6,8	9,1	6,2
SEPTEMBRE	Vrachtschip	10 jaar	Nederland	Cargoship BV	7,4	8,2	6,1
CHUBB	Vrachtschip	11,3 jaar	Nederland	SPAX Shipping BV	7,1	8,5	6,0
PEARL PRIZE	Vrachtschip	11,8 jaar	Noorwegen	Overfind Ships BV	6,7	9,0	6,0
BUTTERFLY	Vrachtschip	8,1 jaar	Turkije	JTB Maritime Pte Ltd	6,2	9,7	6,0
ENCHANCE	Tanker	15,6 jaar	Singapore	SPAX Shipping BV	6,3	8,7	5,5

Eerste < 1 2 3 4 > Laatste  
1-8 items van 1244

**Figure D.3:** Similar visualisation to figure D.1, but with coloured rows based on the final score.

Tabel kolommen Filteren

Naam	Score
REDNAS	8.7
BUTTERFLY	7.2
CHUBB	6.3
MAJUNG	6.3
LAPVIK	6.2
HUNDRA	6.2
SMILAX	6.2
ADEPT	6.2
BLIXT 1	6.1
EN2TRUMMA	6.0
THE FRITHAM	6.0
ENCHANCE	5.5

1-8 items van 1244 Eerste < 1 2 3 4 > Laatste

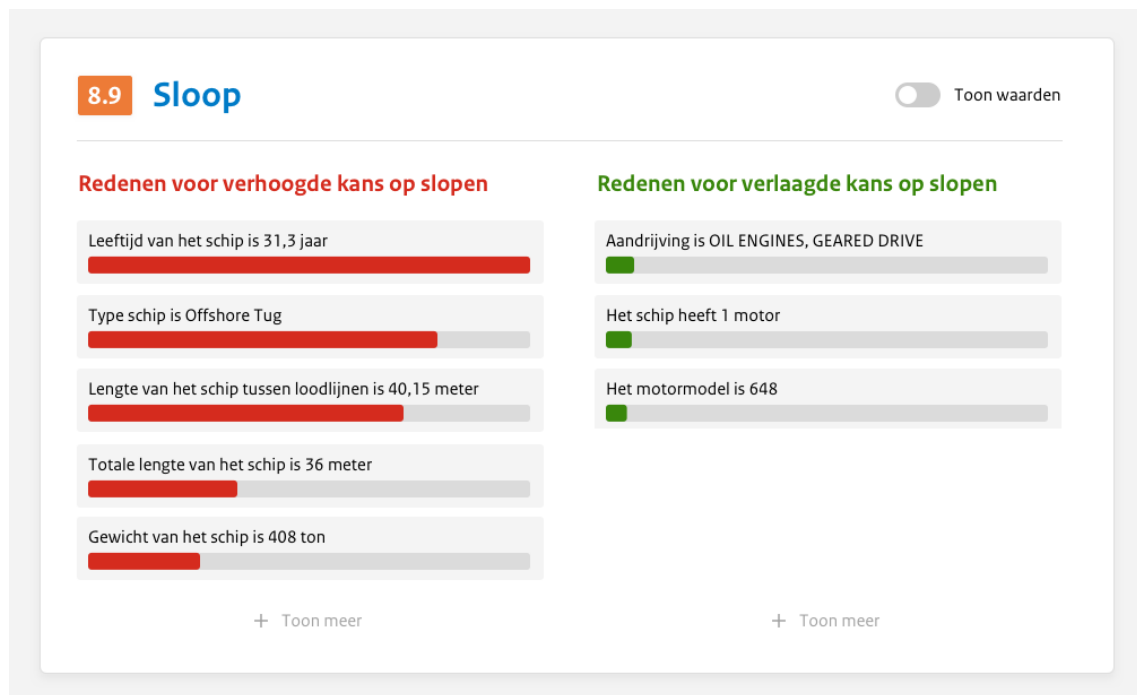
**Figure D.4:** Similar visualisation to figure D.2, but a smaller table with less columns in order to make room for other components on the same page.

**Redenen voor kans op slopen** 8.9

- De leeftijd van het schip is **38 jaar**
- Gewicht van het schip is **17.392 ton**
- De lengte van het schip is **219 meter**

[Details](#) [Model uitleg](#)

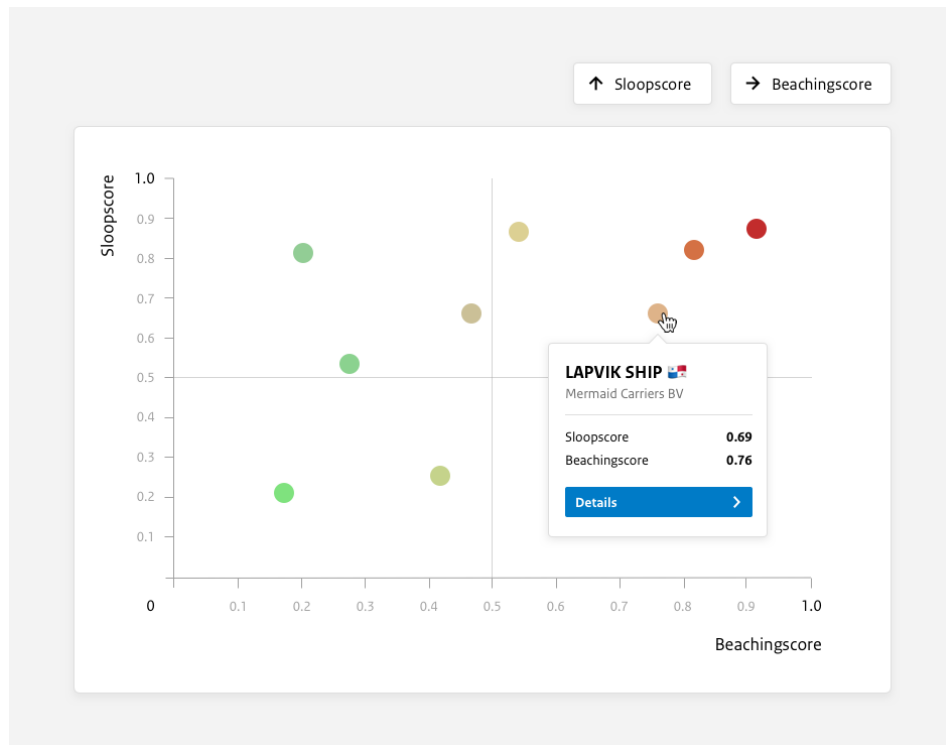
**Figure D.5:** Textual explanation of a prediction score, showing the three most important features



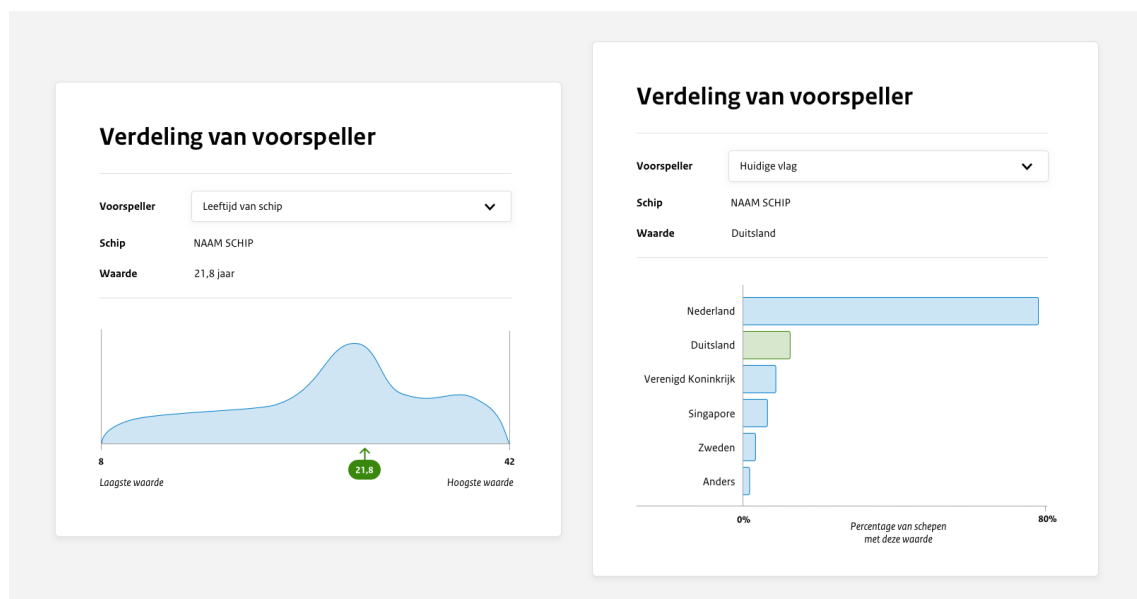
**Figure D.6:** Visual explanation of a prediction score, initially showing a maximum of five most important features (both positive and negative). The feature importance values are visualised through the width of each bar. Optionally, the user can choose to show the raw values for each feature (checkbox "Toon waarden").



**Figure D.7:** Similar to figure D.6, but here the feature importance values are visualised by the opacity of the bar colors.

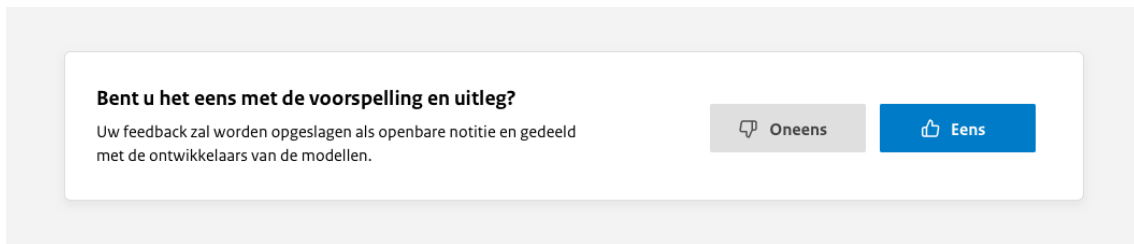


**Figure D.8:** Scatterplot with an overview of all instances. Default axes are the ship-breaking and beaching scores, but the user is able to choose any of the ship characteristics.

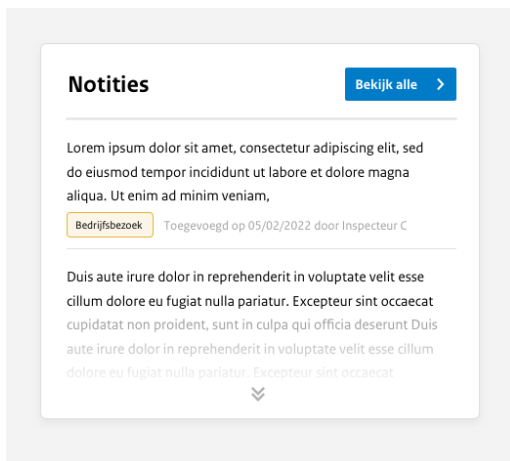


**Figure D.9:** Distribution charts of predictors. Users can choose for which predictor (e.g. age or flag) they want to get an overview. The charts give a distribution of values for that predictor in the entire dataset, and the value of the currently selected ship is highlighted.

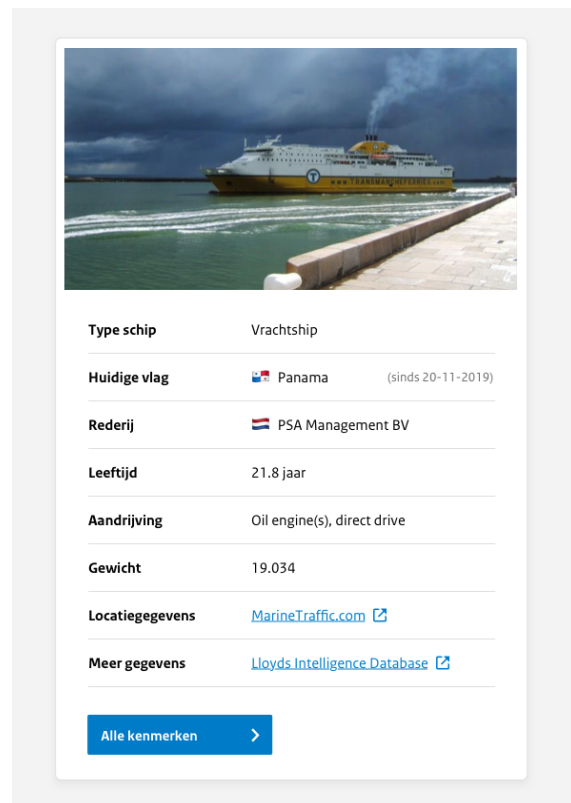




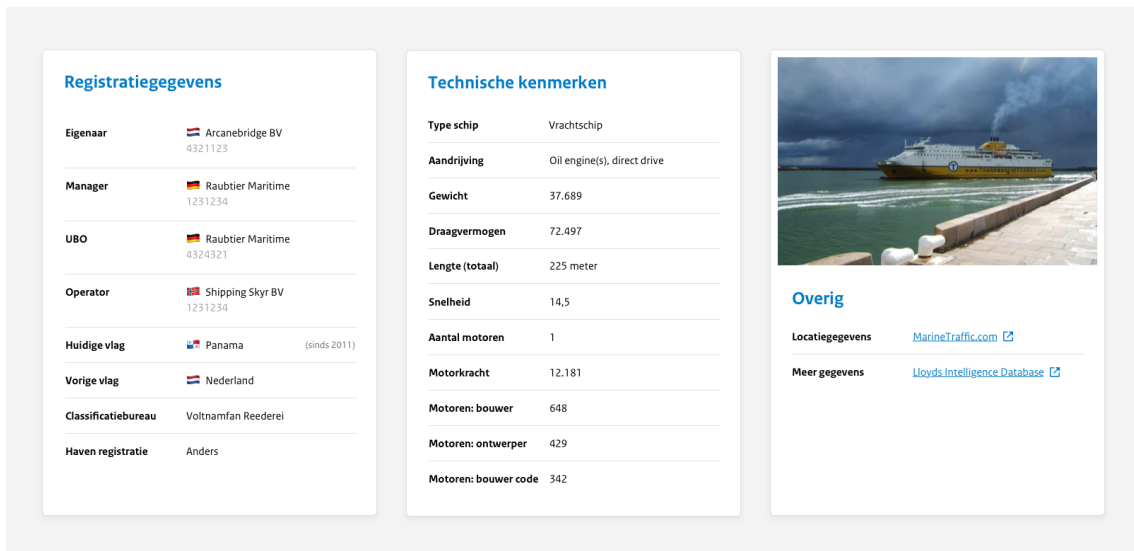
**Figure D.10:** Functionality for giving feedback about the prediction and explanations.



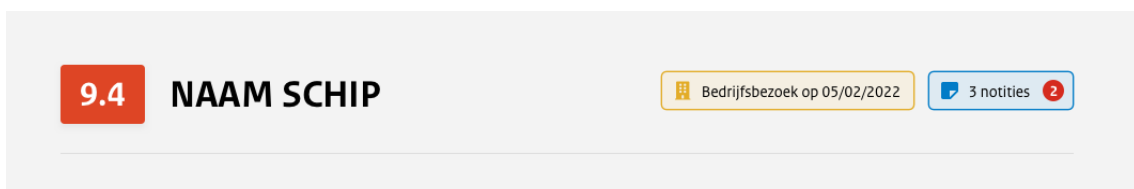
**Figure D.11:** Overview of notes added to an instance, in order to improve work efficiency between the end-users.



**Figure D.12:** Summary of ship characteristics



**Figure D.13:** Overview of all ship characteristics. This includes data that was used for the models, but also other information that might be relevant for making decisions (such as the ownership).



**Figure D.14:** Header component including the score (in colored box visualising a spectrum from green to red), the name of the ship and information about added notes and company visits.

**Vergelijkbare schepen**









Naam	Vergelijkbaarheid	Eindscore
A2B FUTURE	90%	6,3
CHEM STREAM	83%	6,3
Estime	56%	7,2
SYLVIA	48%	8,7

[Bekijk alle >](#)

**Figure D.15:** Small table with similar ships to the ship that was selected, sorted by their similarity.

**Vergelijkbare schepen**

Totaal **9.4** Sloop **8.9** Beaching **9.6**

Huidig schip	100% - 90% vergelijkbaar	89% - 80% vergelijkbaar	79% - 70% vergelijkbaar	69% - 60% vergelijkbaar
 <b>9.4</b> HUNDRA	 <b>9.4</b> STORSLAGEN	 <b>8.6</b> SEPTEMBRE	 <b>5.4</b> THE ELEPHANT	 <b>9.4</b> ENHANCE
	 <b>5.4</b> SMILAX		 <b>9.4</b> PEARL PRIZE	
	 <b>8.6</b> LAPVIK			

**Figure D.16:** Similar content to that of figure D.15, but a larger and more visual overview of similar ships.

## **Appendix E**

# **Focus Group Results**

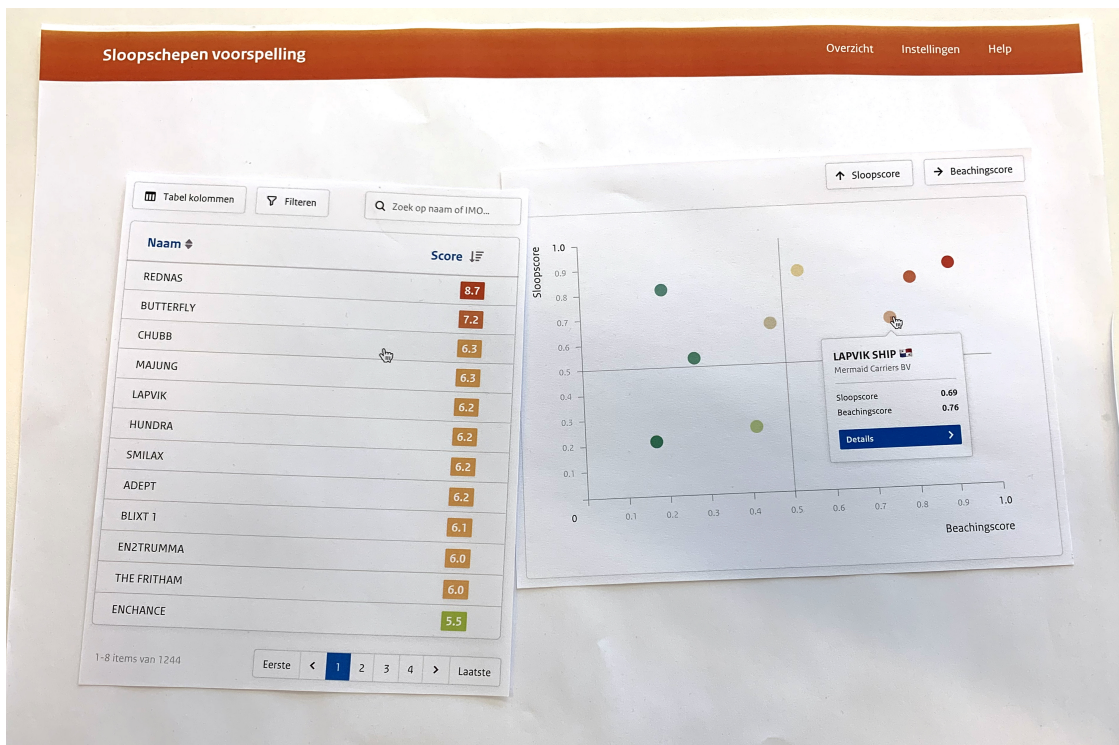
### **E.1 Focus Group 1 (ID-lab members)**

#### **General Conclusions**

- Participants were overall quite happy with the proposed designs, although there were some textual and visual suggestions
- There was a large preference for presenting textual explanation first (vs. visual)
- The 'similar ships' feature was deemed an important aspect of the dashboard
- There was an idea for a new page: the shipping company page with information about a shipping company and a grid/list with all ships and their scores.

### Composition 1: Start screen

The best start screen was described as a table view with a scatterplot next to it (figure E.1).



**Figure E.1:** Dashboard composition of the start screen

- This dashboard was chosen over the large table view, as the scatterplot was deemed to add an easy-to-understand overview of the instances.
- The table view with filter/search options adds the ability to search for a specific ship, or filter on ships from a certain company, and these are expected to be the two most important ways of searching through the data (next to the default 'sorted by prediction score' functionality).
- The table should include at least 4 configurable columns. That means that the table will probably become a little larger and the scatterplot smaller.
- The scatterplot will show the items that are shown in the table, and they should immediately show if an item was already visited or planned to visit.

## Composition 2: Details Overview

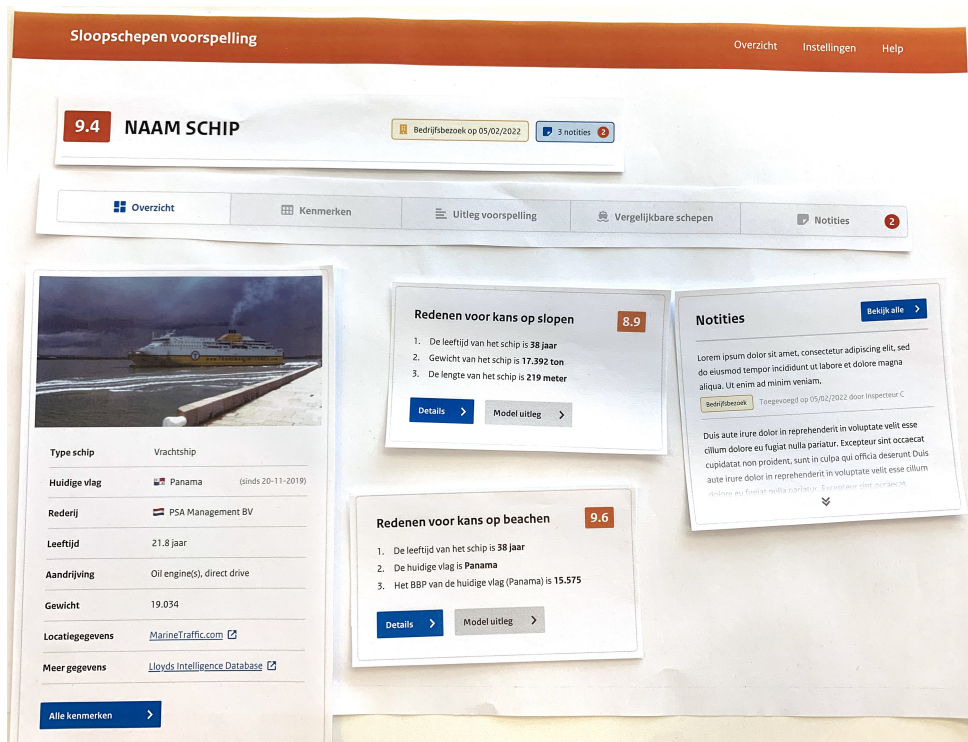
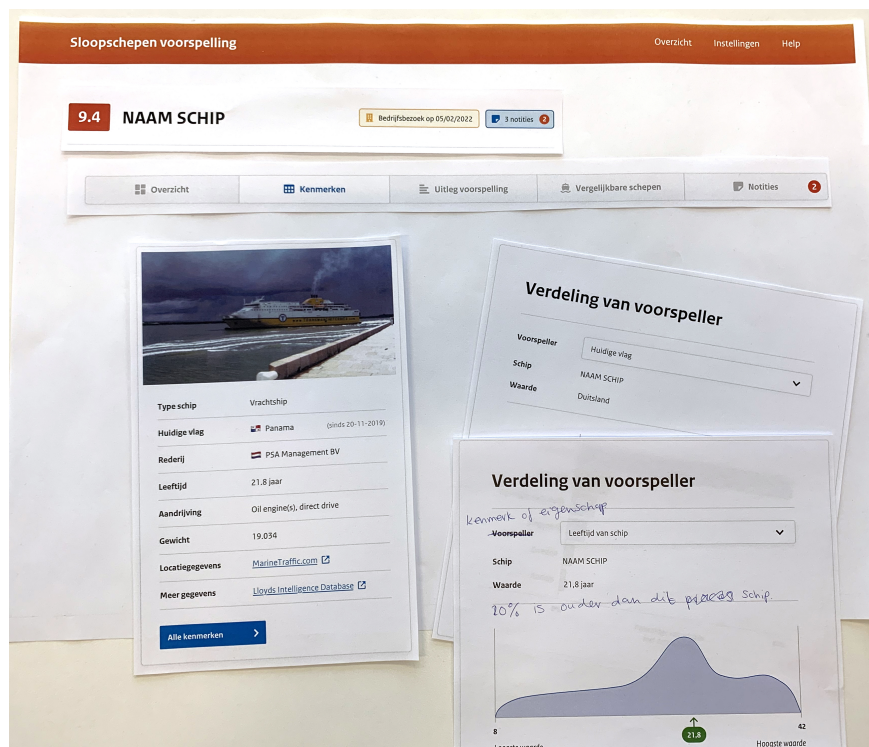


Figure E.2: Dashboard composition of the details overview screen

This was thought the best overview layout of a ship detail page. Most important the ship features table, then the explanation and notes. Other components that were opted to be placed on this were deliberately left out. Some comments about the choices:

- The overview page should not contain too much information, as it might be too overwhelming.
- A textual explanation of all the features with a positive SHAP score should be shown. When clicking 'Details' or the navigation bar item 'Uitleg voorspelling' you will get the visual explanations.
- The notes section could also be just a list of notes/feedback/company visits, their dates and if they are unread. The contents itself could then be hidden behind a click-action.

### Composition 3: Ship Characteristics

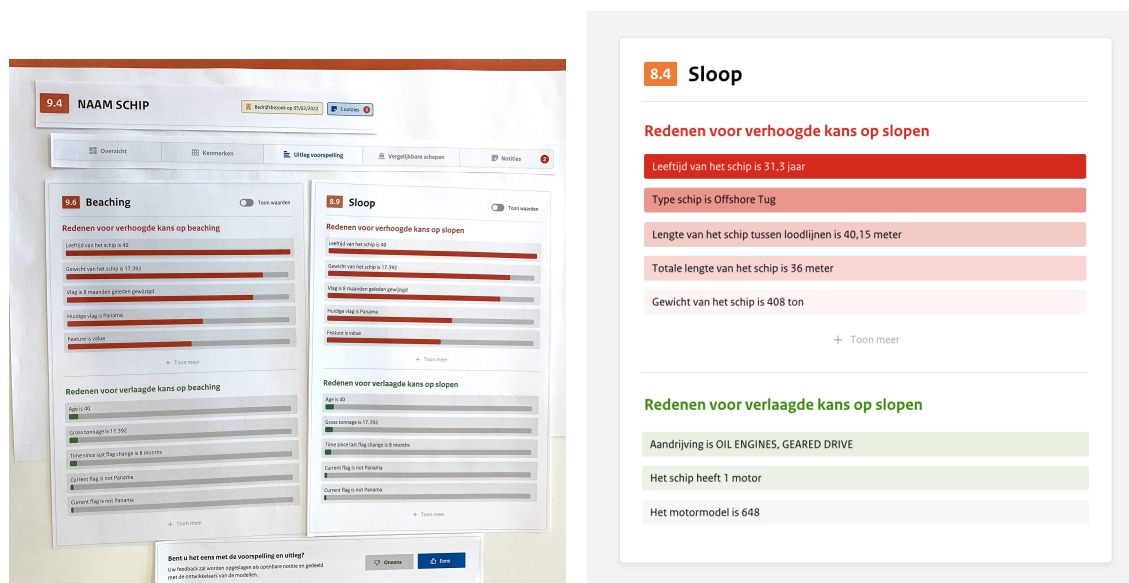


**Figure E.3:** Dashboard composition of the ship characteristics screen

This page should contain all the information (registration details, physical features, etc). This was not a printed component, but an example is shown below. When clicking one of the features, a popup could appear with the context reports. Two comments on those screens (which are also written on the print):

- Add a textual explanation of the context chart, for example: “10% of ships are older”.
- Change the word “voorspellers” to “kenmerken” or “eigenschappen”

## Composition 4: Detailed Explanation



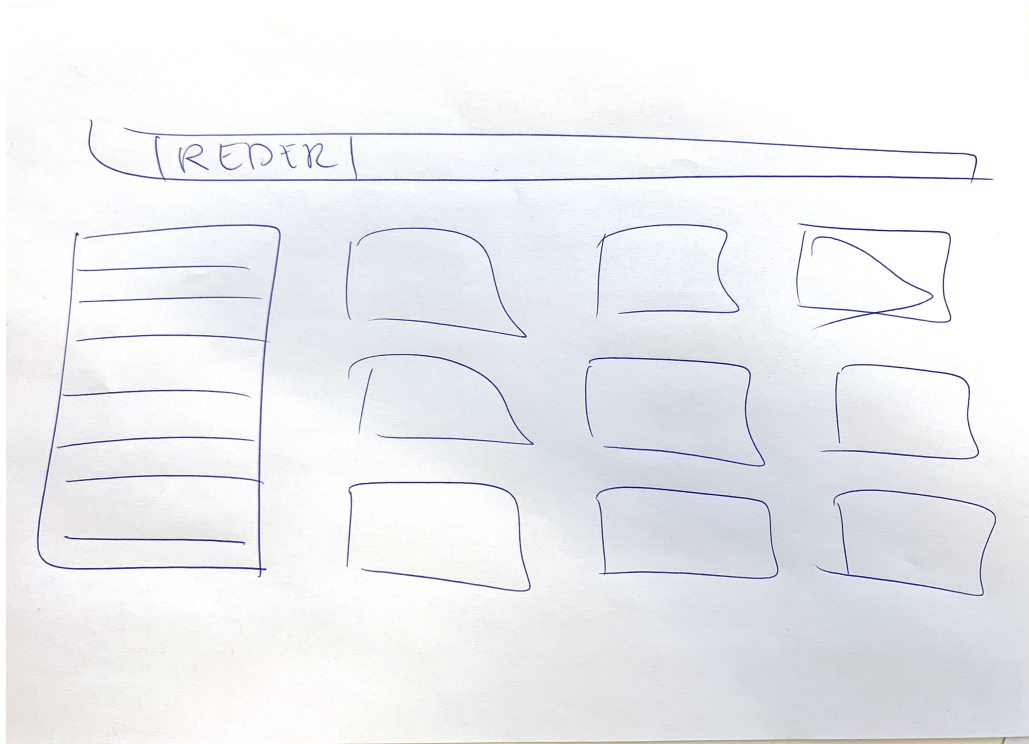
(a) Dashboard composition of detailed explanation screen (b) Visualisation of model explanation using opacity for representing SHAP values

**Figure E.4:** Composition and visualisation of model explanations

There was consensus about placing the visual explanations on the “Uitleg voorspelling” page, instead of the overview (“Overzicht”) page. On this page, the full explanation could be shown of both models. Participants thought that any functionality for showing the raw SHAP values should not be included, as these values were assumed not to be understandable for the inspectors. It could also lead to an undesirable scenario where inspectors will compare values between the two models, whereas SHAP values should be interpreted only relatively within a model.

Participants were not enthusiastic about the visualisations itself. The bars feel too much like ‘progress bars’. When showed some other earlier designs from the prototyping phase, there was a preference for the opacity-visualisations (see figure E.4b), in which the opacity of a feature’s background colour represents the SHAP score. Because of this feedback, it was decided to add this component to the next focus group.



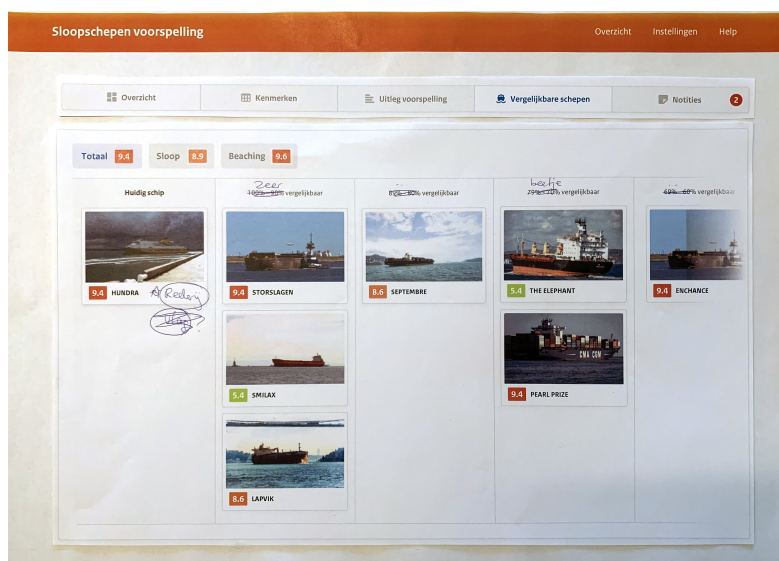
**Composition 5: Company's Ships**

**Figure E.5:** Dashboard composition of the company's ships screen

This is a new component / dashboard page. Shipping company page with information about a shipping company and a grid/list with all ships and their scores. This was deemed important, as inspectors would need this in order to understand more about a company's reputation.

The cards to show can be similar to the ones shown on the 'similar ships' page (see Composition 6).

## Composition 6: Similar Ships



**Figure E.6:** Dashboard composition of the similar ships screen

There was an unanimous preference to include this in the dashboard as a separate page, only with a few alterations:

- Don't mention percentages in the row headings, but use words ('Very similar', 'Rather similar', ..., 'Not similar').
- Add the name of the shipping company to the ship cards.

### Other Feedback and Future Suggestions

- There should be an option to set the colour spectrum minimal and maximal value. Default can be overall minimum score = green and highest score = red. But inspectors might manually decide that all ships under a certain score are not interesting, so they should be able to manually set all ships with a specific score or lower to green.
- Users should be able to configure the table features shown in the table on the overview page.
- When the inspector's feedback is saved, it should also be saved to what model score and explanation this was given to. Otherwise, the score/explanation could change and the inspectors feedback becomes outdated.
- Future work: Add a link to DocGen for generating templated reports
- Future work: the notes section should ideally be connected to ILT's Holmes system.

## **E.2 Focus Group 2 (HCI Master students)**

### **General Conclusions**

- Participants were overall quite happy with the proposed designs, although there were some textual and visual suggestions
- There was a large preference for presenting visual explanation first, as it was assumed easier to quickly see the most important factors, and compare these with other ships.
- The 'similar ships' feature led to varying responses. One participant did not see the added value of the screen. One other could see potential advantages. The main use case that was mentioned: when an inspector gets to know about a ship being beached, (s)he could look up the specific ship in the dashboard and see how similar ships score.
- The design of the similar ships feature was not clear enough for some participants. They suggested that it would be better to choose a cutoff point regarding similarity percentage and only show the ships with the minimal percentage or higher. Also, they should just be shown in a grid instead of in columns based on similarity percentage.

### Composition 1: Start screen

The participants unanimously said that there should not be a separate ‘start’ and ‘detail’ screen, but they should be combined in one. This came from the idea that inspectors might want to quickly browse through ship instances. It would then be a bad experience if users have to go back to the previous screen every time.

At the left, there should be the table with instances. While at the right the ship details (Figure E.7).

- The table should have at least 4 columns, defaulting to “Name, Ship type, Shipping company, Score”
- There was a large preference for presenting visual explanation in the overview tab, as it was assumed easier to quickly see the most important factors, and compare these with other ships.
- Participants overall agreed with the tabbed navigation for when inspectors want to have more details. The “Uitleg voorspelling” can be skipped when the visual local explanations are used in the “Overzicht” page already.
- The scatterplot for providing context was not deemed important enough, and relatively hard to interpret. Therefore, this was left out from the dashboard by the participants.

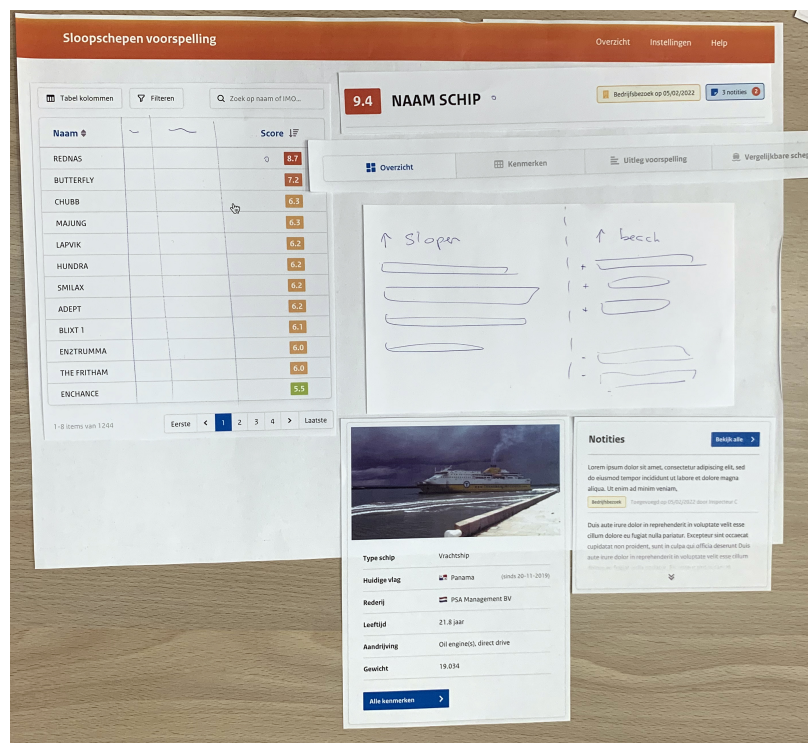


Figure E.7: Dashboard composition of the start screen

## Distribution Chart

Some minor alterations to this visualisation were proposed: Add a dot on the line, with a dashed line and the value below on the x-axis (Figure E.8).



Figure E.8: Sketched feedback on the distribution chart

## Ship Characteristics

Clickable Info-buttons should be added to the characteristics table, which should open the distributions chart (previous page) in a pop-up (Figure E.9).

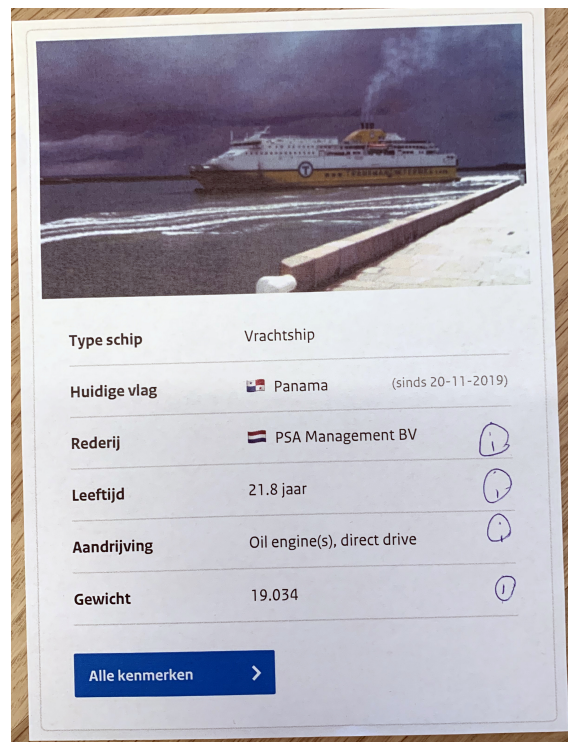


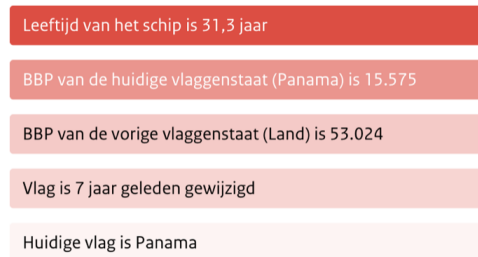
Figure E.9: Sketched feedback on the ship characteristics component

## Visual Explanation Type

The progress-bar like design (right) was preferred over the opacity-differing visualisation (left), as the latter would not be readable for people with visual disabilities (Figure E.10).

### Does not comply accessibility standards

#### Redenen voor verhoogde kans op beaching



+ Toon meer

### Better option

#### Redenen voor verhoogde kans op slopen

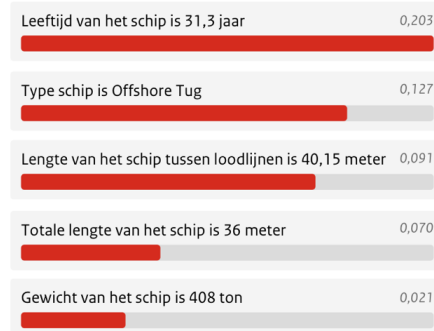


Figure E.10: Comparison of visual explanations

## Explanation Layout

The explanations should be combined in one box (Figure E.11), with positive and negative factors aligned vertically. This gives the opportunity to show both model explanations in one view, instead of hiding either one of them behind a tab. See an example (sketch and design draft) of this idea.

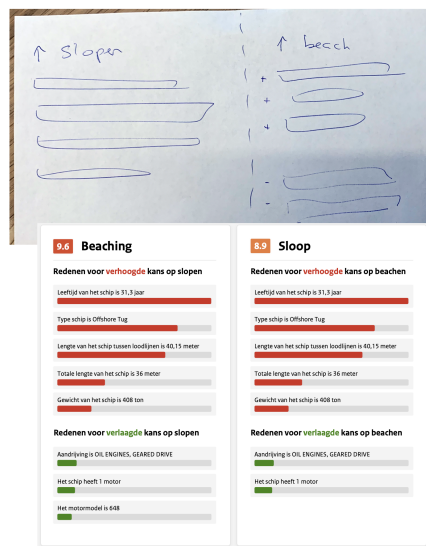


Figure E.11: Comparison of visual explanations

## Appendix F

# System Usability Scale

These are the 10 questions that should each be answered choosing from one of the following options: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree (Jordan et al., 1996).

1. I think that I would like to use this system frequently (Ik denk dat ik dit product frequent zou willen gebruiken)
2. I found the system unnecessarily complex (Ik vond het onnodig ingewikkeld)
3. I thought the system was easy to use (Ik vond het product makkelijk te gebruiken)
4. I think that I would need the support of a technical person to be able to use this system (Ik denk dat ik technische support nodig heb om het product te gebruiken)
5. I found the various functions in this system were well integrated (Ik vond de verschillende functies van het product goed met elkaar geïntegreerd)
6. I thought there was too much inconsistency in this system (Ik vond dat er te veel tegenstrijdigheden in het product zaten)
7. I would imagine that most people would learn to use this system very quickly (Ik kan me voorstellen dat de meeste mensen snel met het product overweg kunnen)
8. I found the system very cumbersome to use (Ik vond het product omslachtig in gebruik)
9. I felt very confident using the system (Ik voelde me zelfverzekerd tijdens het gebruik van het product)
10. I needed to learn a lot of things before I could get going with this system (Ik moest veel over het product leren voordat ik het goed kon gebruiken)

## Appendix G

# Propensity to Trust Technology Scale

These are the 6 questions that should each be answered choosing from one of the following options: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree (Jessup et al., 2019).

1. Generally, I trust technology (Over het algemeen vertrouw ik technologie)
2. Technology helps me solve many problems (Technologie helpt me om veel problemen op te lossen)
3. I think it's a good idea to rely on technology for help (Ik vind het goed om te vertrouwen op technologie voor hulp)
4. I don't trust the information I get from technology (*Reverse question*) (Ik vertrouw de informatie die technologie oplevert niet)
5. Technology is reliable (Technologie is betrouwbaar)
6. I rely on technology (Ik vertrouw op technologie)