City University of London
MSc in Data Science
Project Report
2019

Visualisation and Verbalisation for Explainable AI (XAI)

Author: Heiko Maerz
Supervisor: Dr. Cagatay Turkay
Submitted: 01.10.2019

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Heiko Maerz

Abstract

Artificial Intelligence (AI) is gaining importance, and the generated predictions have an impact in many fields such as healthcare, banking, justice, and education. But what determines these predictions? The user of such a system needs to know how credible and reliable these predictions are, and will want to understand why the particular prediction was generated. Explainable AI (XAI) is the field of research that works towards addressing this need.

This project has developed two different approaches to generate explanations for AI model predictions. Building on top of the design space proposed by Sevastjanova et al., 2018 visualisation, verbalisation and interaction techniques fed into two explainer applications. The explanations were evaluated by five domain experts in a qualitative user research.

Key findings are that the alignment of the explanations with the mental model are fundamental. In this case the explanation is understandable and valuable, it can inform the user and lead to new discovery and knowledge. The participants wished to understand the system, and the cognitive complexity of the system has to be balanced with the value of the prediction. The system must also be designed and implemented in a form that engages the user.

Keywords:
Explanation, Qualitative User Research, Verbalisation, Visualisation, XAI

Contents

Chapter 1

Introduction and Objectives

1.1 Background

1.1.1 Introduction

An increasing number of decisions in a variety of domains such as banking and finance, criminal justice, education, healthcare, insurance, job recruitment, law enforcement, military, retail, and transportation (autonomous vehicles) are based on predictions generated by artificial intelligence (AI). Complex machine learning (ML) models such as support vector machines (SVM), random forests, or deep neural networks (DNN) are used to generate these predictions in high-dimension scenarios. Many models behave like black boxes, and their predictions are hard to understand.

There is a growing demand for AI to be more explainable and accountable, and explainable AI (XAI) has seen a surge of interest for research in this field.

The demand for explanations exists since the emergence of expert systems in the 1970s. It is a multi-disciplinary task that includes the fields computer science, data science, machine mearning (ML) and artificial intelligence (AI), and human-computer interaction (HCI) (Adadi et al., 2018), robotics (Kirsch, 2017), as well as social sciences (T. Miller, 2018), cognitive sciences, law, and philosophy (Mittelstadt et al., 2019).

1.1.2 Explainable AI and Interpretable AI

Adadi et al., 2018 list the terms used in XAI, including 'Black-box', 'Explainable AI', 'Interpretable AI', 'Opaque AI', 'Transparent AI', 'Understandable AI', and 'White-box'. These terms are not used consistently throughout the literature.

The following definitions derived from papers by Doran et al., 2018 and Lipton, 2018 will be used in this report:

- **Interpretability** is defined as a description of how a model works. The aim of interpretability is to make the mathematical functions transparent that map the inputs to the outputs.
- **Explainability** is an answer for a human why an AI model predicted a specific output. The description needs to reflect the individual's expertise and the domain and task. The resulting explanation should allow the individual to understand the decisions and reasoning that lead to the specific prediction.

- **Interpretable models** or **White-box models** are transparent in regard to the mathematical functions, learned weights, and parameters as a result of model training.
- **Non-interpretable models** or **Black-box models** are not transparent. This is due to complexity, because the individual parameters can still be inspected (for example weights and biases in deep neural nets).
- **Opaque AI models** offer no insights for either interpretability of explainability [1].
- **Explainable AI** not only generates a prediction, but also output decisions and reasoning for that prediction that allow insights to the end-user.

Goebel et al., 2018 note that an interpretable system is not necessarily an explainable one, a neural network for example might be transparent in its learned weights and biases, but due to its complexity and number of layers and neurons the prediction might not be explainable to an individual.

1.1.3 Stakeholders

Based on Preece et al., 2018 and Tomsett et al., 2018 this report identifies a number of stakeholders with the following requirements:

- **Researchers, teachers, and students:** This group is concerned with the theory on which the systems and models are based.
- **Data scientists and developers:** They are interested in model understanding, debugging, quality control, identifying bias in the training data, and model improvement.
- **Business stakeholders and decision-makers:** They are the end-users and consumers of the generated predictions, they need to understand predictions, gain trust in the system, and have the necessary confidence in AI to base their decisions on the outputs. It is not expected that members of this group have a background or knowledge in AI or ML. This group is called **domain expert** in this project and is the target audience for explanations.
- **Regulators:** This group audits the models to ensure legal compliance.
- **Individuals:** The people in this group are affected by the predictions of ML or are individuals whose data was used during model training or both. They want to understand model results, detect mistakes, and challenge predictions. It is not expected that members of this group have a background or knowledge in AI or ML. GDPR, which was implemented in Europe in May 2018 gives these individuals the legal right to an explanation, and the proposed Algorithmic Accountability Act aims to implement this right in the US as well.

[1] An AI system might be closed to preserve trade secrets.

These diverse groups have different needs for XAI, among them acceptance of technology, accountability, causality, confidence, credibility, ethics and fairness, justifiability, legal compliance and adherence to moral standards, privacy, reliability, responsibility, robustness, safety, transparency, trust, and unbiased behaviour.

XAI has gained importance, and much research is done to explain the mathematical functions that map model inputs to model output. At the same time there is demand for what Adadi et al., 2018 term "human-like explanations, human-friendly explanations", and the authors suggest combining different approaches to achieve this goal.

Sevastjanova et al., 2018 have proposed an explainability framework which combines visualisation and verbalisation. Goebel et al., 2018 call for XAI to use both visual and verbal modalities.

1.2 Purpose, Product, and Beneficiaries

This project will focus on the last two groups of stakeholder, that is end-users of an AI system and affected individuals. The purpose if this research is to generate explanations that allow these individuals to understand the cause of the prediction rather than the inner workings of the model using the framework proposed by Sevastjanova et al., 2018. It is 'why' rather than 'how' a prediction was made.

Framed as a research question:

RQ1 Can the use of visualisation and verbalisation provide an explanation of an AI/ML model prediction to the user or affected individual of the system?

Supporting research questions:

RQ2 Which approaches can be used to generate these explanations?

RQ3 How effective are the approaches, considering different scenarios, tasks, and algorithms?

Products of this project will be:

- An interpretable and a non-interpretable ML model for an appropriate dataset and target group.
- An explainer application that will generate explanations for both.
- Participant research to assess the explanations.
- An evaluation on the effectiveness of the explanations based on the user research.

Beneficiaries of the project are both end-users and affected individuals without a background in AI and ML. The generated explanations should allow them understand the predictions, make them transparent, highlight capabilities and limitations of the deployed models, and allow to question and challenge them.

1.3 Objectives

The project will be broken down into a number of tasks, each of which will build on the previous. This results in a list of objectives:

1. Identify the requirements for an explanation in the human context. Compile evaluation measures to assess the comprehensibility of the explanations.

2. Select a domain, a target audience, and an appropriate data set. Generate domain specific explanations and tailor them for the chosen audience.

3. Build applications that explain the predictions of different ML models.

4. Conduct qualitative user research to evaluate the comprehensibility of the generated explanations.

5. Evaluate and critically reflect the results of the different approaches to generate the explanations. Report the findings.

1.4 Methods

A literature search was performed to understand explanations in a social and AI context and how to evaluate them, to get an overview of appropriate interpretable ML models and model-agnostic explanation methods, and both verbal and visual approaches, methods, techniques, and software libraries necessary to create an explainer application.

Interview participants with domain knowledge were recruited, and an appropriate dataset was found.

This project explores two approaches to explainability based on the framework of Sevastjanova et al., 2018. This requires training, evaluation, and selection of interpretable ML models, non-interpretable ML models, and model agnostic explanation methods.

The results of these steps are used to build two prototypes of an explanation application that employ both visualisation and verbalisation. One explainer uses an interpretable ML model, and the generated explanation includes both the transparent algorithm of the model as well as explanations for selected observations and their predictions. The second explainer uses a non-transparent ML model. The explanations are generated using a model-agnostic explanation method. These explanations will use insights from social sciences to generate a verbal explanation.

Both understandability and satisfaction of the generated explanations were evaluated with a qualitative user research.

The results of the user research were used to critically evaluate both explanation approaches and the use of visualisation and verbalisation. This evaluation will result in the detection of three themes that will be briefly discussed.

1.5 Metrics for Success

End users, called domain experts in this report, utilise AI systems to aid them in their decision-making tasks. They need to have the confidence and trust to use the

generated outputs of the system in this manner. The main focus of this research is the generation and evaluation of such explanations. A qualitative user research will assess the explanations. The subjective satisfaction with the explanation and the user's understanding of the model is evaluated, taking into account the combination of specific approaches to visualisation, verbalisation, and interaction.

1.6 Work Plan

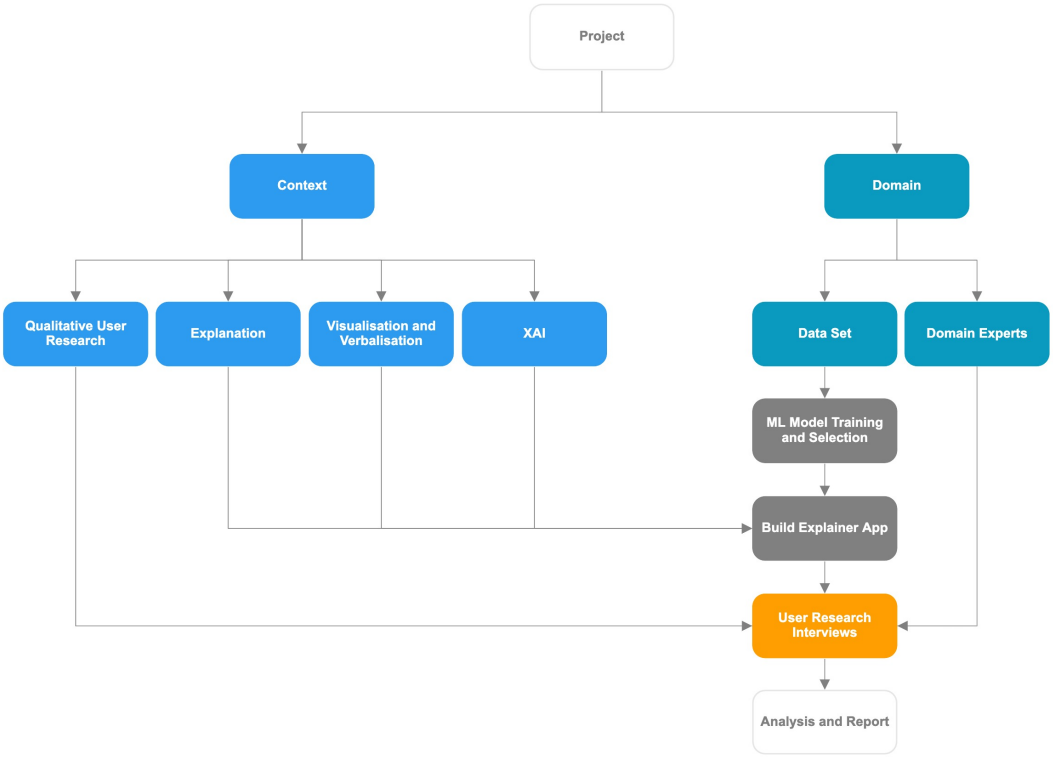


Figure 1.1: Work Plan

The work plan is shown in figure 1.1. Context research is shown in blue, domain identification in cyan, explainer development in grey, and qualitative user research in orange. It demonstrates how the work steps inform each other.

The initial stage included context research regarding explanations in a human context, aspects of XAI such as models, methods, visualisation, and verbalisation, as well as gaining on overview of qualitative user research.

The search for a domain was conducted concurrently. The identification of the domain results in both the data-set and the recruitment process for research participants. This process was iterative and proved harder than expected.

ML model training and selection was begun as soon as the data-set was defined.

Development of the explainer apps builds both on the ML models and the insights gained from the context research into the field of XAI.

The final steps comprised the qualitative user research using the explainer apps, and analysis and report of the findings.

1.7 Changes in Goals and Methods

The focus of this project was reduced to one task (binary classification), one scenario (predict which team would have the MotM), two approaches as in regards to ML model, explanation, and presentation.

1.8 Structure of the Report

Chapter 2 gives the context for the work. It reviews explanations, explanation evaluation, the proposed framework on which this project is based, approaches to verbalisation, visualisation and interaction. Explainable AI, including interpretable ML models and model-agnostic explanation methods are researched. The chapter concludes with a quick overview over qualitative user research.

Chapter 3 describes the methods that were used for the model training, model selection, and development stages, shown in grey in the work plan. The chapter concludes with an outline of the qualitative user research.

Chapter 4 describes the results of model training and model and method experiments. Both chapter 3 and chapter 4 inform each other, the results obtained from developments lead to the selection of a component. This component then feeds into the next step of chapter 3. The analysis of the qualitative user research ends this chapter.

Chapter 5 discusses the results and compares them to the original objectives.

Chapter 6 evaluates the report and highlights its deficiencies. It looks at future work and includes a reflection of the author.

The supporting references and source are listed.

Appendix A includes the original project proposal.

Appendix B includes the Research Ethics Review, supporting documents, and approval.

Appendix C lists all additional files that were submittes as part of this project.

Chapter 2

Context

2.1 Explanation

The online edition of the Oxford Dictionary defines an **explanation** as an "action or process of explaining something (...) a statement that makes things intelligible." (OED, 2019). In the context of of XAI and the target audience for this project Molnar et al., 2019 refines the definition as follows: "An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way."

2.1.1 Properties and Qualities of an Explanation

Insights from Humanities

T. Miller, 2018 argues that people expect human-like behaviour from AI systems. An explanation generated by an AI system should therefore be modelled on human behaviour. He has conducted a survey on research in cognitive sciences, philosophy, and psychology to determine how people explain decisions or predictions to each other.

Explanation as a Social Interaction

Hilton, 1990 argues that an explanation consists of three parts, the individual who explains (the explainer), the subject of the explanation (the explanandum), and the individual to whom the subject is explained (the explainee). The explanation has to close a knowledge gap to be appropriate, meaningful, and relevant. Goebel et al., 2018 confirm that explanations are the foundation of communication and of understanding which results in learning. T. Miller, 2018 expands that everyday explanations are social interactions between individuals in which the explainer transfers knowledge to the explainee. The explainer will tailor the explanation to the situation, the domain, and to the explainee. This transfer of knowledge happens in form of an interaction or conversation.

Tailor to the Target Audience

What makes an explanation understandable to a human? Ribera et al., 2019 argue that to be understandable, the explanation has to satisfy the requirements of the

target audience. Defining the goals, the content, and the generation method of an explanation is only possible when the background, the expectations, and the knowledge of the stakeholder is taken into account. This is confirmed by Wachter et al., 2018, who also focus on the stakeholders.

What make an Explanation

The target audience of this project is an individual with domain knowledge. There is no presumption that the individual has prior knowledge in AI or ML.

According to Narayanan et al., 2018 the explanation should reflect the causal structure, describe the necessary conditions and be plausible. An explanation will describe a causal relation on why something occurred, specifically why the system returned a specific prediction.

Explanations are selective, the explainer will concentrate on important causes rather than the comprehensive scientific relationship (T. Miller, 2018). Zemla et al., 2017 confirm this, stating that an explanation should use the smallest number of causes necessary while still be consistent with the evidence. G. A. Miller, 1956 explores the cognitive capacity of individuals to process information, and sets the limit to "seven, plus minus two" cognitive chunks.

Coherence, generality, and simplicity are desired properties of the explanation.

Counterfactual Explanation

Explanations are answers to a 'why'-question ('why was P predicted?'), and T. Miller, 2018 found that they are often contrastive ('why was P predicted rather than Q?') where counterfactual causes will result in a different prediction (the **foil**), in a binary classification scenario the opposite prediction.

A counterfactual explanation helps the individual understand what should change (the **counterfactual**) to receive a different, possibly a desired outcome (Wachter et al., 2018). Many counterfactuals exist for a given prediction, the so called "Rashomon effect" (Fisher et al., 2018). The counterfactual that is closest to the observation usually provides the most insight (Molnar et al., 2019).

2.1.2 How to Evaluate an Explanation

The requirement for an explanation in the context of this project is to be understandable to a domain expert. The explanation has to be comprehensible, that is a human individual has to understand it (Molnar et al., 2019).

Doshi-Velez et al., 2017 insist that the target audience has to be taken into account when evaluating an explanation. They argue to use principles from human-computer interaction. Kirsch, 2017 call for the use of user testing to be able to determine the understandability or comprehensibility of the explanation.

A number of evaluation methods are suggested by Mohseni et al., 2018 for domain experts as target audience, that is stakeholders who will use the system for analysis and decision-making:

- **Human-machine task performance:** Measures the impact of an explanation on the effectiveness and efficiency in the analysis and decision-making process.

- **Mental model of the user:** Measures the degree to which the user understands the system.
- **Satisfaction with the explanation:** Measures subjective factors to determine the understandability of the generated explanation.
- **Trust in the system:** Measures the confidence the user has in the model to make the correct prediction.

A mental model is the internal representation of reality that individuals derive based on their observations. It allows them to understand the real world and to make predictions (Jones et al., 2011). Kulesza et al., 2013 use the mental model as a proxy of a user's understanding of a system based on generated explanations. Mohseni et al., 2018 suggest two tasks that can be used to test the understanding of a user:

- **"Prediction verification":** Can the user justify the prediction of a system based on a observation and the explanations?
- **"Counterfactual reasoning":** Based on a observation, a prediction, and the explanation, can the user change the observation data that would change the prediction?

2.2 Visualisation and Verbalisation for XAI

In their survey Mohseni et al., 2018 have found many examples of the use of visualisation (such as salience maps) and verbalisation, using natural language to describe a ML model.

Scenario specific analysis is required to determine if visualisation, verbalisation, or a combination of both yields the best approach to fulfil the user's requirement for a XAI system.

Sevastjanova et al., 2018 propose combining both approaches to leverage the strengths of each. Visualisation allows a good overview of large amounts of data, while verbalisation allows for detailed information. They suggest strategies for both explanation generation and explanation presentation to guide the decision of which aspect of the explanation to visualise and which aspect to verbalise.

2.2.1 Explanation Generation

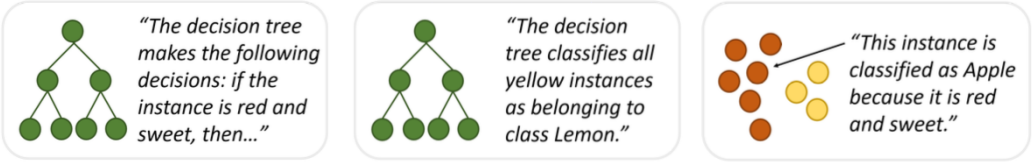


Figure 2.1: Explanation Generation, source: Sevastjanova et al., 2018

Visualisation and verbalisation should augment each other to generate useful explanations. Sevastjanova et al., 2018 suggest a few methods:

- Combining visualisation and verbalisation to present the same information, using both mediums to increase the understandability of the explanation at the expense of efficiency (figure 2.5, left, **"double encoding"**).

- They also suggest a number of ideas that use visualisation for an overview of the data and verbalisation for detailed descriptions of the mapping of inputs to outputs (figure 2.5, middle, **"overview and summary"**), of metadata of the ML model, or to explain the prediction for specific observations (figure 2.5, right, **"overview and detail"**).

2.2.2 Explanation Presentation

Figure 2.2: Explanation Presentation, source: Sevastjanova et al., 2018

The presentation depends on the domain, the task, and the background and preference of the user. Sevastjanova et al., 2018 suggest an interactive user interface to allow the user to explore observations and predictions. The user could select on observation on the visual overview and receive a detailed verbal explanation (figure 2.2, left, **"details on demand"**), or the system could point out some observations in combination with a verbal counter-factual explanation (figure 2.2, middle, **"data-driven guidance"**). An **"agent based dialog system"** (figure 2.2, right) can take the interaction even further.

2.2.3 Visualisation

S. Liu et al., 2017 list a number of visualisation approaches for explainable AI, yet many are geared towards ML experts to understand, debug, and improve the ML model. Many examples can be found for deep neural nets in the domain of computer vision, displaying the input image representation in the various layers of the net (see for example M. Liu et al., 2016).

The target audience for this project however is a different one. Figure 2.2 and 2.2 reveal two demands that the domain expert might have:

1. Enable the user to pick observations to gain a detailed, but local explanation.

2. Gain a global understanding of the system and an understanding of the algorithm that leads to an explanation. The prerequisite is an interpretable ML model, such as a decision tree.

For a classification task the first demand requires a visualisation that adequately represents the observations and clustering according to similarity of features or membership to a class. It also requires the integration of the "visual encoding and interaction idiom" (Munzner et al., 2015) to affect which data is is displayed (Kirk, 2016). Interaction should encourage discovery and lead to new knowledge.

The second demand requires an adequate visual representation of the internal workings of the model. If a decision tree algorithm was chosen for the interpretable ML model a tree graph is the obvious choice. A tree is non-cyclical network consisting parent nodes connected to child nodes. Each child node references exactly one parent node, and the path from the highest parent node, the root node, to a given child node is unique (Munzner et al., 2015). The tree therefore is a good visual representation of the decision path for a prediction.

2.2.4 Verbalisation

To display a verbal explanation non-linguistic input, numerical or categorical data, needs to be transformed into text (Gatt et al., 2018). Natural language generation (NLG) researches this domain. Ideally the generated text should "flow" in well-formed sentences, including pronouns and names, and avoid repetition.

Of note is that feature engineering might result in transformations that are less comprehensible than the underlying original features.

Reiter; Dale, 2000 propose an advanced method for natural language generation that consists of three sub-systems:

1. **"Document planner":** determines the structure and the content of the text, which depends on the data, the user, and the domain;

2. **"Microplanner"**, determines the syntactic structure, including lexicalisation (words and phrases), aggregation (text that can be combined to avoid repetition), and referring expression generation to identify objects;

3. **"Surface realiser"** that maps the abstract text specification into an actual text.

This system is complex. Deemter et al., 2005 argue that template based systems, comparable to "mail-merge" functionality of word processors, can be used instead. Texts generated by template systems usually are less fluent and not as readable (they lack textual variation) as text generated by advanced NLG methods.

2.3 Explainable AI (XAI)

This section does not aim to be a comprehensive survey of methods for explaining AI models, as that is beyond the scope of this report. Molnar et al., 2019 provides an overview that is currently updated, and some surveys can be found here: Biran et al., 2017, here: Došilović et al., 2018, here: Guidotti et al., 2019, and here: Nunes et al., 2017.

Instead, this section will explore key concepts and samples of models and methods that inform this project given the chosen data-set (supervised machine learning with a binary classification task) and target audience which comprises domain experts.

Explanations can be either global or local. A global explanation covers the entire model, either directly in case of interpretable models or via a surrogate in case of model-agnostic methods. A local explanation only covers a group of observations or a single observation. The local observation can focus on features specific to that

observation and ignore large parts of the model. This reduces complexity and may increase human understanding.

Two approaches are explored:

- Interpretable ML models;
- Non-interpretable models in combination with model-agnostic explanations;

2.3.1 Interpretable Models

Fundamentally interpretable ML models are the most straightforward way to generate explanations and justifications. Lipton, 2018 breaks this down into "simulatability", which describes the entire model, "decomposability", which describes the individual inputs, parameters, and outputs, and "algorithmic transparency", which describes the process in which the model is trained.

Rule based models are interpretable, such as decision trees and decision rules (Biran et al., 2017).

Decision rules are a combination of IF-THEN statements that are used to make a prediction.

Decision Tree

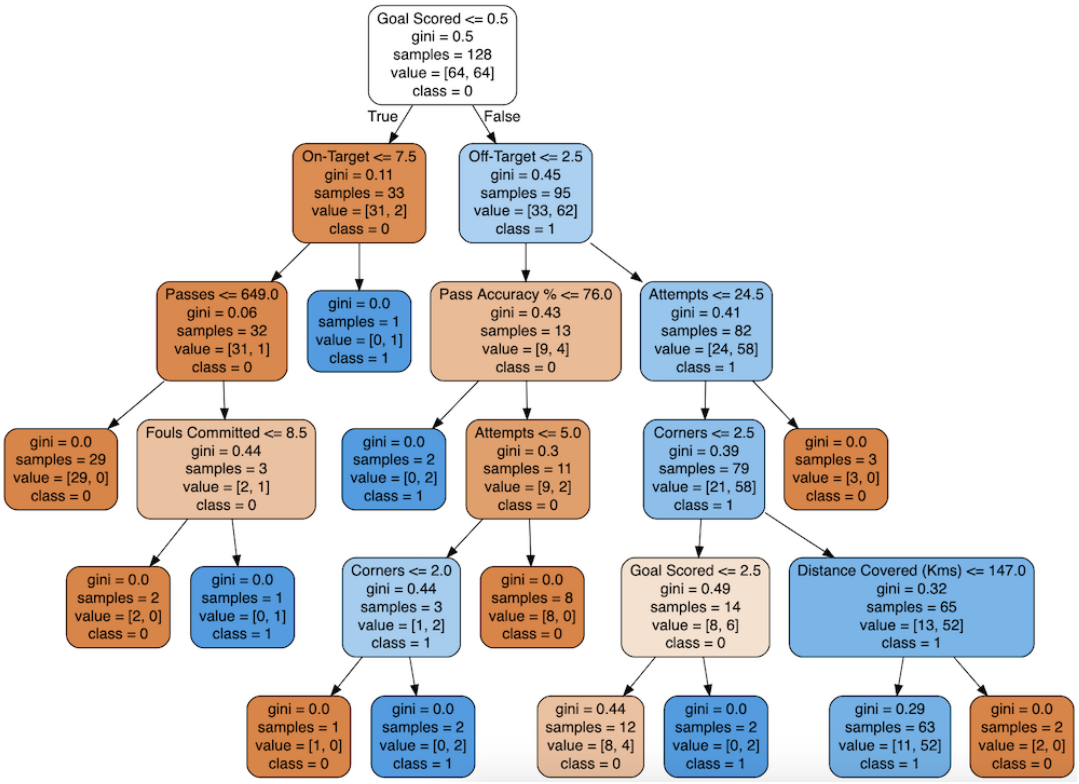


Figure 2.3: Decision Tree

A decision tree consists of a root node, a number of internal decision nodes that create a decision path based on features and feature value thresholds, and leaf or

end nodes that represent the regression or classification prediction. Decision trees can be visualised with their nodes and edges (figure 2.3). As long as they are not too deep they are very interpretable, but the number of leaf nodes grows quickly with increasing depth. It is easy to verbalise a decision path, counterfactuals can be constructed by changing the direction on a decision node. Feature importance corresponds to the distance of the decision node to the root node, the root node holds the most important feature and threshold.

Linear Models

Another family of interpretable models are linear models and their extensions. Linear regression is such linear model, predictions are derived by calculating the sum of weighted observation data. The weights for each feature explain the magnitude and the direction of influence of the feature. Extending linear regression with a logistic function allows the use for classification problems, but the interpretations is not as intuitive for a ML layperson.

An interpretable model provides valuable insights if its accuracy for the scenario in question is acceptable and its degree of complexity (such as tree depth) is such that an individual can comprehend it.

2.3.2 Model-Agnostic Methods

Model-agnostic methods allow to extract explanations from non-interpretable models. These explanations are extracted from the ML model post-hoc, either by training an interpretable model using inputs and outputs of the black-box model or by perturbing the inputs to the black-box model and observing the resulting predictions. Some model-agnostic methods use surrogate models. These surrogates are interpretable models trained on the predictions of the black-box model. Fidelity measures how well the surrogate approximates the underlying black box model.

Ribeiro et al., 2016a list a number of advantages to this approach:

- **"Model flexibility":** The machine learning model most appropriate for the given task can be chosen. The machine learning model can also be changed if necessary.
- **"Explanation flexibility":** The explanation method most appropriate for the given task, target group, or presentation can be chosen.
- **"Representation flexibility":** The feature representation most appropriate for the given task can be chosen.

Mittelstadt et al., 2019 argue that approximations and local explanations mimic human explanations. They point out that it is important to understand where the approximation is reliable, where it is uncertain, and where it is not accurate. If the explanation does not reflect the model it is not understandable and can be misleading.

One disadvantage shown by Alvarez-Melis et al., 2018 is the robustness of the model-agnostic method. They compared two methods, LIME and SHAP, and conclude that the explanations can vary locally for non-linear models.

LIME and SHAP are two commonly used model-agnostic explainers in tutorials (Kaggle, 2019a, PyDataNYC, 2018) They will be explored briefly:

LIME

LIME (Local Interpretable Model-agnostic Explanations) was proposed by Ribeiro et al., 2016b.

LIME uses local surrogates for specific predictions, see figure 2.4:

1. Perturb the data in the proximity of the selected observation (bold red +), and compute the black-box predictions;
2. The perturbed data and predictions are weighted by their distance to the selected observation;
3. Train an interpretable model on the resulting data-set, this example uses a linear decision boundary;
4. The selected observation is explained using the local surrogate model.

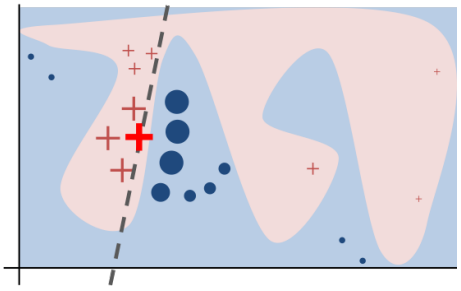


Figure 2.4: LIME, source: Ribeiro et al., 2016b

The example demonstrates how a complex global decision boundary results in a local linear one.

LIME is implemented for tabular data, text, and images.

SHAP

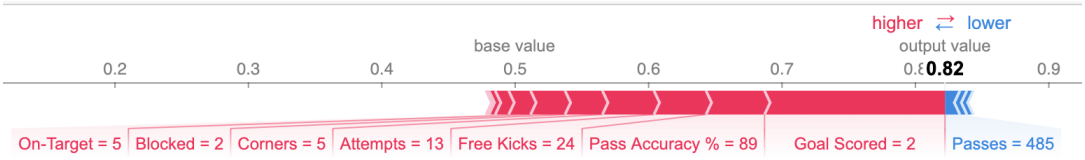


Figure 2.5: SHAP Explanation

SHAP (SHapley Additive exPlanations) was proposed by (Lundberg et al., 2017. It is based on coalition game theory and computes the average marginal contribution of a feature to the prediction of an observation. The marginal contribution corresponds to feature importance. The feature importance is generated for specific observations using the predictions of the black-box model.

Figure 2.5 shows an example of an explanation for one specific observation. The explanation starts with the base value, which is the prediction average for all observations (in this case 0.5). Each feature of the observation in question increases or decreases the probability of the prediction. In this example the features in red ('Goals Scored, 'Pass Accuracy', etc.) increase the probability, with 'Goals Scored' the feature with the highest marginal contribution. The feature 'Passes' lowers the prediction. Adding the marginal contributions of all features to the base value results in the model prediction of 0.82.

2.4 Qualitative User Research

To evaluate an explanation one has to question if it conveys valuable insights to an individual. Any evaluation will therefore require user testing (Kirsch, 2017).

2.4.1 Interview

Qualitative research is a method to gather scientific data. Interviews can be used to gather feedback. Interviews can be conducted with single individuals or with groups, and they can be structured, unstructured, or semi-structured (Oates, 2006).

The semi-structured interview is used in qualitative user research. It is based on a discussion guide which contains topics and open-ended questions (Given, 2008). The interview can follow the flow of the conversation, which allows the researcher to explore new topics, ask additional questions, and discover thoughts and opinions of the user that are not part of the discussion guide. This form of interview engages the participant (Galletta et al., 2013).

One form to conduct a semi-structured interview is to use the **think aloud technique**. Rather than asking an abstract question the participant is given a task and asked to think aloud while performing that task. The researcher thus gains data about the thought processes, knowledge, methods, and insights of the participant (Van Someren et al., 1994).

2.4.2 Analysis

Insights from qualitative data can be derived using **thematic analysis**. Thematic analysis "is a systematic approach to the analysis of qualitative data that involves identifying themes or patterns of cultural meaning; coding and classifying data, usually textual, according to themes; and interpreting the resulting thematic structures by seeking commonalities, relationships, overarching patterns, theoretical constructs, or explanatory principles." (Mills et al., 2010, p. 926).

After a good overview of the data initial codes are extracted and organised into clusters, these clusters then form themes. The themes are evaluated and organised into a "thematic map". The last step is to define the themes and describe their content and scope (King et al., 2018).

Chapter 3

Methods

3.1 Introduction

The aim of this project is to generate and evaluate explanations for the target group of domain experts. They are the end-users of an AI application and use the generated predictions in their decision-making process. One example is the demonstration of pacmed at PyDataAmsterdam, 2019.

These users possess domain expertise, but knowledge and understanding of AI and ML is neither required nor assumed.

The project will compare two applications, a Non-interpretable Model Explainer and an Interpretable Model Explainer. The explainers take the domain, the data-set, and the target group into account. This chapter describes the methods used to build and evaluate the applications in order to answer the research question.

For the development phase of the project this chapter and the next (Results) inform each other and build on each other. Each step described in this chapter is evaluated, this evaluation is described in Chapter 4. The evaluation leads to the selection of a particular model, method, or technique, and the selected component is the base for the step.

3.2 Data-set

A initial project to work with a NHS data-set unfortunately fell through, and an alternative had to be found fast. In order to keep the premise of domain knowledge the FIFA 2018 Man of the Match (MotM) [1] data-set from Kaggle was chosen, and the participants were recruited from football fans among friends.

In the 2018 world cup the Man of the Match was voted for by fans [2] via website, app, or twitter (FIFA.com, 2019) rather than being picked by a commentator or sponsor. This makes it a suitable task for ML. The data-set can be found on the Kaggle website (Kaggle, 2019b).

The FIFA 2018 datasets has game statistics for the 2018 World Cup matches hosted by Russia. Each observation corresponds to a team/match combination, it has 128 instances, two teams for each of the 64 World Cup matches. It is a

[1] The Man of the Match is awarded to an outstanding player in that particular match. The player usually belongs to the winning team, although the player could be from either team. Source: Wikipedia, 2019

[2] One of the research participants actually voted for a MotM

multivariate data-set with 27 columns. There are 20 continuous features (such as 'Goals Scored') and one binary target feature (the MotM belongs to the team, the MotM does not belong to the team). The remaining columns are descriptive, such as team, opponent, and date of the match.

3.3 Visualisation and Verbalisation Designspace

	Visualisation	**Verbalisation**	**Interaction**
Interpretable Model Explainer	decision tree, scatter plot	decision path	select instance for detail
Non-interpretable Model Explainer	scatter plot	explanation, counterfactual	select instance for detail

Table 3.1: Design Space

This project will implement two applications to explain the prediction if, for a given team and match, a team had the MotM as a member. Both will be evaluated and compared with a user research.

The applications use the visualisation and verbalisation design space proposed by Sevastjanova et al., 2018 and described in section 2.2. Two different approaches are used (see table 3.1), an interpretable model and a non-interpretable model in conjunction with a model-agnostic explainer. In both the user is able to pick an observation to explain from an interactive scatter plot, but visualisation and verbalisation will be specific to the model.

The explanation applications are implemented using the Python library Bokeh (Bokeh, 2019).

3.4 ML Model Training

This project compares explanations generated from an interpretable model with explanations derived from a model-agnostic method.

All models are trained using a grid search with 3-fold cross validation. The hyperparameters are listed in table 3.2. Due to the small size of the data-set, it only contains 128 entries, the data is not split into test, train, and validation. All models are trained on the full data-set and the training accuracy is used as evaluation criteria. This approach contravenes common ML training processes, because it does not guard against overfitting. It should be noted, however, that the aim of this project is not to train a model that generalises well against future unseen data, but to evaluate explanations. The risk is therefore acceptable. Once trained only accurate observations are retained, that is the prediction equals the ground truth. This is done to avoid confusing the user. Furthermore, only those observations are used in the explanation applications that are accurate for both models. This allows the user to pick the same teams when comparing both models.

Hyperparameter	Value
Decision Tree	
criterion	entropy, gini
splitter	best, random
Random Forest	
number of estimatiors	5, 10, 20, 50, 100, 200
maximum features	1∼10
maximum depth	None, 5∼20
bootstrap	True, False
criterion	entropy, gini
Support Vector Machine	
C	0.001, 0.005, 0.01, 0.05,,0.1, 1, 10
gamma	0.001, 0.01, 0.05, 0.1, 0.5, 1, 'scale'
kernel	'linear', 'poly', 'rbf'
degree	2,3,4

Table 3.2: Hyperparameters

All models are trained with SKLearn libraries found at SKLearn, 2019. The Jupyter notebook
'Thesis_Heiko_Maerz_Vis_and_Verb_XAi.ipynb'
contains the code to train the models and is part of the thesis submission.

Once trained the models are saved to be used in the explanation application.

3.4.1 Train Interpretable Models

Decision trees are easy to interpret because one can follow the decision path from the root node through the decision nodes to the model prediction, see section 2.3.1. For this reason decision trees are used for interpretable model explanations while controlling tree depth. G. A. Miller, 1956 argues that the number of of objects that an individual can keep in short-term memory is "seven, plus minus two". Three trees are therefore trained, with maximum depth set to 5, 7, and 9 respectively. The tree selected for the user research will be based on subjective complexity, taking training accuracy into account.

3.4.2 Train Non-Interpretable Models

Random Forest and Support Vector Machines are used as black-box ML models. The particular ML models were chosen based on the author's past experience with the models during the programme. This ML model will be used in combination with a model-agnostic method, and the actual model is interchangeable (**model flexibility**). For that reason the choice of the model is arbitrary and of no major consequence to this research. It should be noted that the use of a neural net is negated by the small size of the training data-set. The model with the best accuracy is chosen as base for the model-agnostic method.

3.4.3 Model-agnostic Explanation Method

The model-agnostic method provides the input for the explanation text. Both LIME (Ribeiro et al., 2016b) and SHAP (Lundberg et al., 2017) are evaluated in a prototype in the Jupyter notebook. The most suitable one is used in the application.

3.5 Non-interpretable Model Explainer

3.5.1 Non-interpretable Model Explainer Application

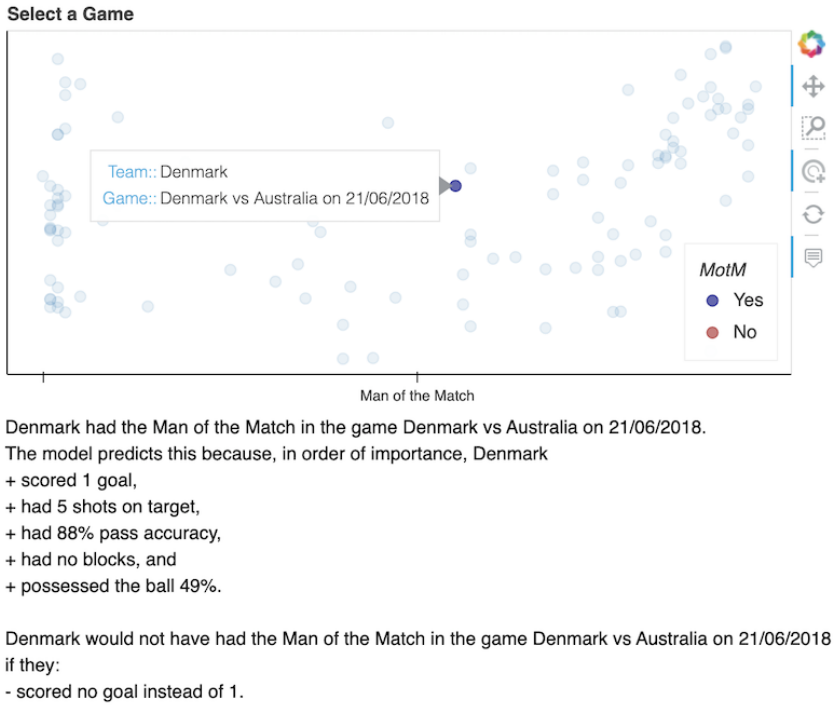


Figure 3.1: Application: Model-agnostic Explainer

This explainer application will generate explanations for a black-box ML model. It is not possible to understand exactly how the ML model arrives at the prediction. A model-agnostic method can be used, however, to generate explanations by example (post-hoc, Lipton, 2018).

This application uses the random forest model based on the evaluation in section 4.2.
The application consists of two parts:

- An interactive scatter-plot of all teams and games (see section 3.5.2). The user is encouraged to browse and select a game.
- The explanation text generated for the selected team and game (see section 3.5.4). The text describes the features and their values that have led to the prediction, as well as a counter-factual for the opposite prediction. This text should allow the user to form an understanding of the model (Ribera et al., 2019).

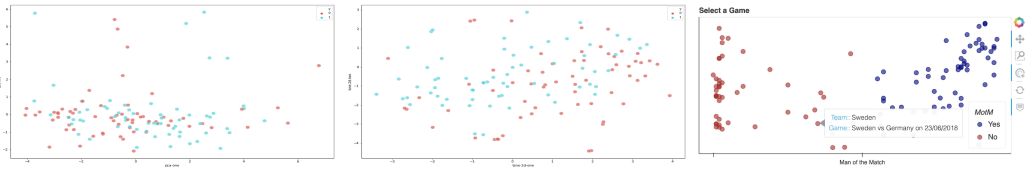
The explainer is developed as an interactive Python Bokeh application (Bokeh, 2019).

A screen-shot is shown in figure 3.1. The user has selected team Denmark in the game Denmark vs. Australia on the scatter plot. The explainer has returned the important features for this prediction. The number of features is a parameter of the explainer, for this project it equals the tree depth of the Interpretable Model Explainer to facilitate comparison. The explanation text lists them in order of importance.

The closest counterfactual for class 'the MotM is not in team Denmark' has a different value only for the features, 'Goals Scored'. The counterfactual text is displayed with the feature name, the counterfactual value, and the original value.

The explanation texts for all games are uploaded to Moodle in the additional files section.

3.5.2 Visualisation: Interactive Team Selection



(a) PCA (b) t-SNE (c) Predicted Probability

Figure 3.2: Scatter plots

This app uses an interactive graph to allow the user to pick a game for explanation. It follows "overview and detail" explanation presentation and "details on demand" explanation generation (Sevastjanova et al., 2018).

The initial premise is to visualise the games in a scatter-plot. The classes are colour-coded (red for a team that did not have the MotM, blue for a team that did have the MotM). Three different experiments are conducted (see figure 3.2) and evaluated (see section 4.2.1):

- Dimensionality reduction using 2-component PCA (see figure 3.2a).
- Dimensionality reduction using 2-component t-SNE (see figure 3.2b).
- ML model prediction probability on the x-axis, 1-component dimensionality reduction on the y-axis (see figure 3.2c).

The evaluation and selection of the scatter-plot is discussed in section 4.2.1.

The scatter-plot is interactive (see figure 3.2c). A hover-tool displays the team and the game on mouse-over. The graph offers tools to zoom, pan, and re-set the image. A click on the glyph selects the team and game.

3.5.3 Verbalisation: Explanation

Both model-agnostic methods evaluated in section 4.1.3 include functions that return a list of features and feature importance values (the code is available in the submitted Jupyter notebook).

The explanation text is based on this list of features, sorted by importance. It is passed to a template based NLG. Each feature is mapped to a specific template, e.g. 'Goal Scored' maps to 'scored {n} goals' while 'Ball Possession %' maps to 'possessed the ball n%'. The mapping takes the feature value (0, 1, >1) into account, in the example of 'Goal Scored' the mapping returns 'scored no goal', 'scored 1 goal', and 'scored {n} goals' respectively, as suggested by Reiter, 1995. This step returns a list of text tokens. This list of tokens is joined according to its length, one token returns just the text, two tokens are joined by an 'and'. In case of more than two tokens all are separated by a comma except the penultimate and the last, which are separated by an oxford comma and an 'and'. The text is terminated by a period.

The number of features to be used is determined by a parameter. For this experiment that parameter is set to equal the tree depth of the decison tree for the Interpretable Model Explainer, i.e. 5.

3.5.4 Verbalisation: Counterfactual

Computation

As suggested by T. Miller, 2018 verbal counterfactuals are used in this application. Each observation and prediction will have many counterfactuals, the so called Rashomon effect (Molnar et al., 2019). A method is required to pick the right one. To be useful for an explanation the chosen counterfactual should be the 'closest' to the observation. This section describes the approach used for this project. A major downside of the approach was discovered during user research which is described in section 4.3.2.

This project uses a post-hoc, naïve, brute force approach to pre-compute the counterfactuals. These are saved in a file and retrieved by the explanation app. This is done to speed up the computational process when interacting with the explainer application.

The logic roughly follows Waa et al., 2018 and Wachter et al., 2018 (see section 2.1.1):

1. For each observation i predict the class (the **fact**), determine the opposite class (the **foil**), and list the n most important features.

2. Using the data for this observation i, create a local data set by perturbing the values of the most important features returned by the model-agnostic method (see section 3.5.3). The new feature values are based on the distribution of the training data set (see table 3.3. To keep the permutations manageable the count of permuted values for each feature are limited.

3. Predict all local observations, and only consider instances of class foil (that is the opposite class).

4. Use a distance function to find the 'closest' counter-factual to this observation. This project uses euclidean distance with normalised features.

Text generation

The explanation application reads the data for both the selected game and the corresponding counter-factual. The feature values are compared. If the feature

Feature	Perturbation Values
Attempts:	3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25
Ball Possession %:	25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75
Blocked:	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Corners:	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Distance Covered (Kms):	80, 87, 94, 101, 108, 115, 122, 129, 136, 143, 150
Fouls Committed:	5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25
Free Kicks:	5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27
Goal Scored:	0, 1, 2, 3, 4, 5, 6
Goals in PSO:	0, 1, 2, 3, 4
Off-Target:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Offsides:	0, 1, 2, 3, 4, 5
On-Target:	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Own goals:	0, 1
Pass Accuracy %:	67, 70, 73, 76, 79, 82, 85, 88, 91, 94
Passes:	180, 240, 300, 360, 420, 480, 540, 600, 660, 720, 780, 840, 900, 960, 1020, 1080, 1140
Red:	0, 1
Saves:	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Yellow & Red:	0, 1
Yellow Card:	0, 1, 2, 3, 4, 5, 6

Table 3.3: Counterfactuals: Perturbation Values

values differ the feature name, its value, and its counter-factual are passed to a template based NLG. This returns a list of text-tokens that are then aggregated in the same manner as described in section 3.5.3.

3.6 Interpretable Model Explainer

3.6.1 Interpretable Model Explainer Application

The application will generate explanations for the interpretable ML model. A decision tree of depth 5 was chosen because it is a good compromise between model complexity and accuracy (see section 4.1.1). The specific tree for this app was chosen based on its complexity.

The explanation consists of two parts:

- Global explanation: The user is given the opportunity to study the algorithm and understand the logic that leads to a prediction. The inner workings of the model are shown, and the user can re-trace the prediction.
- Local explanation: The user can select a team and game and receives a verbal description of the decision path with the feature names, the thresholds, and observation values.

The application consists of three parts:

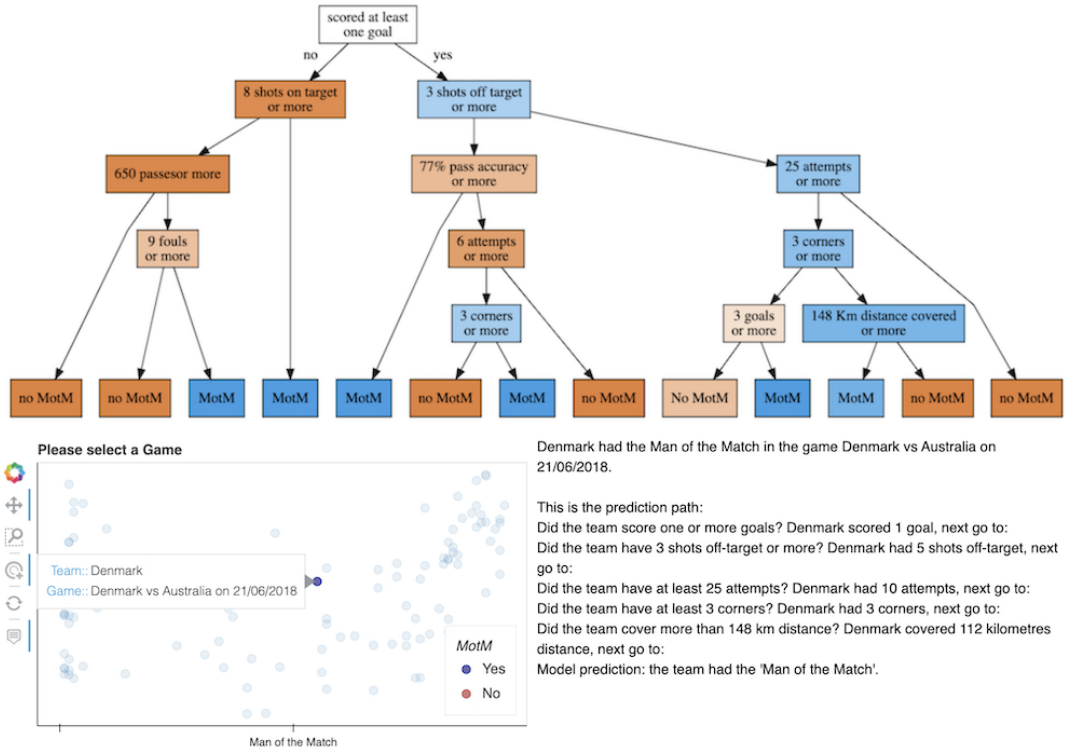


Figure 3.3: Application: Interpretable Model

- A static visualisation of the decision tree, including the decision nodes with their features and thresholds, and the prediction depicted as leaf nodes.
- An interactive scatter-plot to select teams and games. This part is common to both explainer applications.
- The explanation text that verbalises the decision path of the selected team and game.

Both tree visualisation and decision path verbalisation inform each, the information is "double-encoded".

The explainer is developed as an interactive Python Bokeh application (Bokeh, 2019).

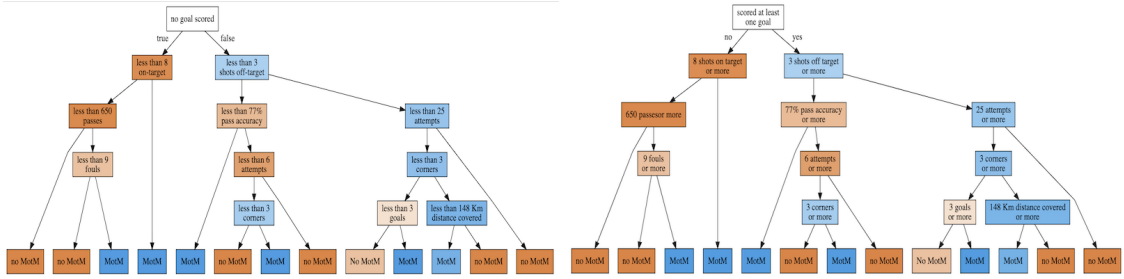
Figure 3.3 shows a screen-shot of the explanation application for the interpretable model.

The user has selected a team on the scatter-plot, for comparison it is the same as in section 3.5.1, the team Denmark in the game Denmark vs. Australia. The text describes the decision path using the feature values of the observation and the decision nodes. The text and the decision tree graph correspond to each other.

The explanation texts for all games are uploaded to Moodle in the additional files section.

3.6.2 Visualisation: Decision Tree

The decision tree is well suited for visualisation. It enables the user to understand the model on a global level (see section 2.3.1). Each prediction can be traced from the



(a) Model Mapping (b) Intuitive Text

Figure 3.4: Explanation App: Decision Trees

root node along the decision nodes to the model prediction at the leaf nodes. Counterfactuals can be generated visually by changing direction at the decision nodes. Depending on their complexity (tree depth) these models are highly interpretable.

The tree visualisation was created using the SKLearn tree.export_graphviz method, which exports the tree in DOT format (see figure 2.3). The text in the nodes is manually changed for use in the explanation app (figure 3.4). A first direct translation of the rules at each node is on the left, and a more intuitive version on the right. The colour of the nodes corresponds to the majority class, and the saturation to the proportion of classes in the node.

3.6.3 Visualisation: Scatterplot

For consistency this explainer application uses the same interactive scatter plot (figure 3.2c) for game selection as the Non-interpretable Model Explainer (section 3.5.2). This allows the users to easily compare both applications by selecting the same games.

3.6.4 Verbalisation

The explanation text describes the decision path along the tree for a selected observation. The SKLearn method sklearn.tree.DecisionTreeClassifier.decision_path() returns the traversed nodes. Each node is mapped to a text that reflects the decision or the final prediction. The features and their values generate text tokens using the same templates as described in section 3.5.3. The list of tokens is aggregated, taking into account whether the text token represents a decision node or the model prediction.

3.7 Qualitative User Research

3.7.1 User Research Goal

The user research is conducted to collect data about how the two different models (decision tree and SHAP) are understood by domain experts, in this case football fans that were exposed to data from the FIFA World Cup 2018.

The research question for the user research is in line with the research question for this project: How do domain experts with no knowledge of ML or AI understand

the explanations delivered by the two different models? The key research questions are listed in table 3.4.

Non-interpretable Model Explainer	
Explanation Evaluation	
1)	Does the explanation explain the prediction?
	Verbalisation: do explanation and counterfactual inform each other?
Model Understanding	
2)	Does the participant understand how the system works?
	Does the participant pick up feature importance?
	Can the participant construct a counterfactual?
Visualisation and Verbalisation	
3)	Does the interactive scatter-plot encourage the user to explore the sys- tem?
Interpretable Model Explainer	
Explanation Evaluation	
1)	Does the explanation explain the prediction?
Model Understanding	
2)	Does the participant understand how the system works?
	Do the global visualisation of the decision tree and the individual prediction inform each other?
	Does the participant pick up feature importance?
	Can the participant construct a counterfactual?
Visualisation and Verbalisation	
3)	Does the double-encoding of visualisation of the decision tree and verbalisation of local explanations inform each other?
	Does the interactive scatter-plot encourage the user to explore the system?
Comparison	
1)	Which model does the participant prefer and why?
2)	How does accuracy influence the preference?

Table 3.4: Interview Questions

3.7.2 Recruitment and Supporting Documentation

Ethics approval was requested and granted before the start of the interviews (see Appendix **??**). All interviews are anonymous and notes collected will be anonymised.

Recruitment

The research participants were football fans recruited through friends and family. Most participants were men (4), as it seemed difficult to recruit female football fans. The participants were between 32-49 years old. They all received a £20 Amazon gift voucher as compensation for their participation.

Supporting Documentation

The informed consent form and the participant information sheet were based on the guidelines from City University of London. Both were sent to the participants prior to the interview (see appendix **??**).

3.7.3 Semi-structured Interview

For the semi-structured interview a written discussion guide with open ended questions was developed (see appendix **??**). During the task based part of the interview participants were requested to use the think out loud method so that the researcher could capture the thinking behind their actions.

Data Collection

The interviews were scheduled to last up to 45 minutes and consisted of four parts:

1. **Introduction to the interview:** Explanation and review of the informed consent form together with the participant, signing of the consent form and answer any questions the participant might have.

2. **Warm-up question:** Give the participant an easy start into the interview by briefly inquiring about their interest in football. This will not be evaluated.

3. **Introduction to the project:** Short overview over AI and XAI and information about the project.

4. **Task based evaluation of the models:** Participants are asked to explore why a team has the man of the match.

   - Show one of the models first (Interpretable or Non-interpretable Model Explainer) and ask participants to evaluate their understanding of the explanation.
   - Show the other model next (Non-interpretable or Interpretable Model Explainer) and ask participants to evaluate their understanding of the explanation.
   - Ask participants to compare both models.

   The model sequence (Interpretable or Non-interpretable Model Explainer) was alternated between the interviews. This was done to prevent any bias in the collected data (Salkind, 2010).

The full discussion guide can be found in section **??**. The interviews were audio-recorded. A note-taker assisted the researcher [3], the subsequent analysis was based on these notes and augmented by the audio files. Notes and captured audio is submitted to Moodle in the 'Additional Files' section.

3.7.4 Thematic Analysis

This project follows the six steps of thematic analysis as defined by Braun et al., 2013.

1. **"Becoming familiar with the data"**: the researcher studied the notes and listened to the recordings multiple times.

[3] She is the author's wife.

2. **"Generating and grouping codes":** creation of tags (colour codes) from the data, these were attributed to interesting findings relevant to the interview research questions.

3. **"Searching for themes":** the findings were collated by themes into a spreadsheet. The findings were organised by research participants for attribution.

4. **"Reviewing themes":** a revision of the spreadsheet allowed the identification of key themes relevant to the research questions of this interview.

5. **"Defining and naming themes":** the analysis resulted in the identification of three key themes that can be clearly identified and named.

6. **"Production of a report":** the results are reported in section 4.4

Chapter 4

Results

4.1 Model Training

The evaluation of models is found in the Jupyter notebook submitted in the 'Additional Files' section on Moodle.

4.1.1 Interpretable ML Model

Decision Tree 5

Accuracy: 88.28%
F1-score: 88.89%
Confusion matrix:

		Predicted	
		yes	no
Truth	yes	**53**	11
	no	4	**60**

Decision Tree 7

Accuracy: 92.97%
F1-score: 92.13%
Confusion matrix:

		Predicted	
		yes	no
Truth	yes	**58**	6
	no	3	**61**

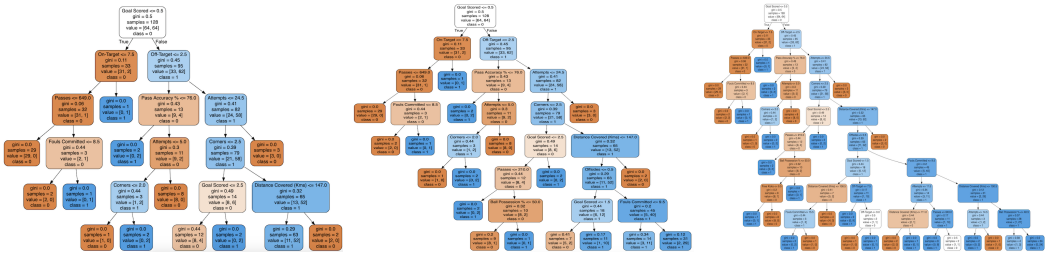
Decision Tree 9

Accuracy: 98.44%
F1-score: 98.44%
Confusion matrix:

		Predicted	
		yes	no
Truth	yes	**63**	1
	no	1	**63**

Table 4.1: ML Model Evaluation: Decision Tree

The ML model selected for the Interpretable Model Explainer is the decision tree. Three trees are trained, with maximum depth set at 5, 7, and 9 respectively. The choice of decision tree will be based both on its accuracy and f1-scores (see table 4.1) subject to its perceived complexity (see figure 4.1) .



(a) Tree Depth 5 (b) Tree Depth 7 (c) Tree Depth 9

Figure 4.1: Decision tree complexity

The hyperparameters (see table 3.2) determined by using grid-search returned 'gini' for 'criterion' and 'best' for 'splitter' for all three trees.

Trading subjective perceived complexity against accuracy the shortest tree will be used in the Interpretable Model Explainer. The qualitative user research will yield further insights if the choice was correct.

4.1.2 Non-interpretable ML Models

Random Forest

Accuracy: 92.19%
F1-score: 92.75%
Confusion matrix:

		Predicted	
		yes	no
Truth	yes	**54**	10
	no	0	**64**

SVM

Accuracy: 79.69%
F1-score: 79.03%
Confusion matrix:

		Predicted	
		yes	no
Truth	yes	**53**	11
	no	15	**49**

Table 4.2: ML Model Evaluation: Non-interpretable Model

A complex ML model is used in conjunction with a model-agnostic method for the Non-interpretable Model Explainer. The ML model is considered a black-box. For that reason acceptable accuracy and f1-score (see table 4.2) are the only selection criteria. A random forest and a SVM are trained in the first iteration, if neither model is a candidate for the Explainer additional ML models will be included in further iterations.

The random forest returned acceptable accuracy and f1-score, in line with the trained decision trees, and is the candidate for the Non-interpretable Model Explainer. The hyperparameters determined by the grid-search (see table 3.2) returned 'False' for 'bootstrap', 'entropy' for 'criterion', 5 for 'max_depth', 10 for 'max_features, and 10 for 'n_estimators'.

The SVM did not perform as well and is discarded. The hyperparameters determined by the grid-search are linear kernel, gamma = 0.001, and C = 1.

4.1.3 Model-agnostic Method

Both model-agnostic methods used to explain a specific observation include functions that return a list of features and their importance for the respective model prediction (lime.LimeTabularExplainer.explain_instance() and shap.TreeExplainer..shap_values()) where not restricted in the number of features used for the list.

The generated explanation text (as well as the computation of the counterfactuals) will be based on this list. This section compares both and proposes one to be used in the applications.

Table 4.3 shows an example of the feature list compiled for team Mexico in the game Mexico vs Korea on 23/06/2018. The model predicts that Mexico has the MotM with 82% probability.

LIME	
Feature	**LIME value**
Goal Scored	0.174371
Corners	0.039549
On-Target	0.032702
Free Kicks	0.032427
Pass Accuracy %	0.027028
Distance Covered (Kms)	0.018725
Offsides	0.008791
Yellow Card	-0.007615
Ball Possession %	0.006971
Saves	0.00358

SHAP	
Feature	**SHAP value**
Goal Scored	0.13525
Pass Accuracy %	0.043887
Free Kicks	0.039015
Attempts	0.038653
Corners	0.028014
Blocked	0.025821
On-Target	0.016404
Passes	-0.011185
Yellow Card	0.010077
Offsides	-0.005236
Distance Covered (Kms)	0.004727
Saves	-0.004725
Off-Target	0.002166
Fouls Committed	-0.000606
Ball Possession %	0.000421
Yellow & Red	0
Red	0
Goals in PSO	0
Own goals	0

Table 4.3: Model Agnostic Method

Both LIME and SHAP values have a magnitude and a direction, that is their relative importance as well a direction if they contribute to the prediction probability or push the prediction in the opposite directions (negative value in the 'vale' column).

The list returned by the SHAP explainer however allows a direct interpretation of the prediction probability: starting from a baseline (which is the class probability of the training dataset, in this case 0.5), each feature contribution (the SHAP value in the list) pushes the prediction towards the model probability or away from it. The SHAP values are local, they are specific to the selected observation (see section 2.3.2). Adding the sum of each feature contributions to the baseline results in the prediction probability returned by the model. To allow future expansions of the verbal explanation generation the SHAP implementation is chosen for this project.

4.2 Visualisation

4.2.1 Interactive Scatter Plot

Neither PCA (see figure 3.2a) nor t-SNE (see figure 3.2b) returned clusters that correspond to the class labels of the observations. The resulting graph should encourage the user to interactively select games, explore the explanations, and allow them build a mental model of the AI system. Instead, both graphs appear confusing rather than assisting in this task.

The third option was chosen instead. In case of the Non-interpretable Model Explainer the x-axis does convey information as it represents the prediction probability. The y-axis is based on a 1-component t-SNE dimensionality reduction and

unfortunately does not convey any intuitive information to the user.

For consistency both applications will use the same interactive scatter-plot. This is done to ensure consistency and to allow the user to pick the same team and game for the two explainers to enable comparison.

The user research showed that the plot was very confusing for the participants.

4.2.2 Decision Tree

The decision trees are based on the .DOT files generated by the function from sklearn.tree.export_graphviz() (see figure 2.3). The .DOT files were manually changed so that the text in the decision nodes reflects the target audience. Figure 3.4a is a direct translation of the thresholds, the flow direction is to the left if the decision is true, to the right otherwise. This proved confusing in early interviews and was changed to the format seen in 3.4b.

4.3 Verbalisation

4.3.1 Explanation text

While not satisfying the requirements set by Reiter; Dale, 2000 for a flowing, natural text, the explanation text was deemed as adequate by the researcher, and user research confirmed this.

4.3.2 Counterfactual

Computation

The computation had one major flaw which only became apparent during user testing. The values used for the perturbations were derived from the data distribution in the training data-set. Some attributes however had large ranges, for example 'Ball Possession' (min=25, max=75), 'Pass Accuracy' (min=67%, max=94%), 'Distance Covered' (min=80km, max=148km), and 'Passes' (min=189, max=1137). The number of permuted values per feature was limited to the data shown in table 3.3. This approach was chosen to keep the total number of local permutations for each observation manageable and computational run-times acceptable. The result however were counterfactuals which, for certain observations and explanations, were either confusing or, worse, counter-intuitive.

A simple adaptation of the the code would include the original values for changed features into the local dataset. Other more sophisticated approaches are discussed by Wachter et al., 2018 and Rathi, 2019.

The effect will be demonstrated using three examples:

Example 1: Egypt is predicted not to have the MotM. The counterfactual will list the features that need to change and the new values so that the system predicts Egypt does have the MotM:

> **Egypt did not have the Man of the Match in the game Egypt vs Uruguay on 15/06/2018.**

The model predicts this because, in order of importance, Egypt
+ scored no goal,
+ had 8 attempts,
+ had 3 shots on target,
+ had no corners, and
+ had 78% pass accuracy.
Egypt would have had the Man of the Match in the game Egypt vs Uruguay on 15/06/2018 if they:
- scored 1 goal instead of 0,
- had 11 attempts instead of 8,
- had 4 shots on target instead of 3, and
- had 88% pass accuracy instead of 78.

This counterfactual is intuitive and corresponds to the prior mental model of the participants. The values of the features for the counterfactual change in the direction expected by the user.

Example 2: Serbia is predicted to have the MotM. The counterfactual will list the features that need to change and the new values so that the system predicts that Serbia does not have the MotM:

Serbia had the Man of the Match in the game Serbia vs Costa Rica on 17/06/2018.
The model predicts this because, in order of importance, Serbia
+ scored 1 goal,
+ had 19 free kicks,
+ covered 109 kilometres distance,
+ had 2 blocks, and
+ had 3 offsides.
Serbia would not have had the Man of the Match in the game Serbia vs Costa Rica on 17/06/2018 if they:
- scored no goal instead of 1 and
- *covered 108 kilometres distance instead of 109.*

This counterfactual is confusing: the user expects to see a change from 1 goal to no goal for the counterfactual, however the change from 109km to 108km distance covered is unexpected, and it is not clear why such a small change for this particular feature is necessary to change the prediction.

The underlying reason is that the local data-set never included a value of 109km for 'Distance Covered', but rather 108km and 115km. The observation with 108km had a shorter Euclidian distance to the original observation and was selected as counterfactual.

Example 3: Switzerland is predicted to have the MotM. The counterfactual will list the features that need to change and the new values so that the system predicts that Switzerland does not have the MotM.

Switzerland had the Man of the Match in the game Switzerland vs Costa Rica on 27/06/2018.

The model predicts this because, in order of importance, Switzerland
+ scored 2 goals,
+ had 87% pass accuracy,
+ had 594 passes,
+ had 12 attempts, and
+ had 6 corners.
Switzerland would not have had the Man of the Match in the game Switzerland vs Costa Rica on 27/06/2018 if they:
- scored no goal instead of 2,
- *had 88% pass accuracy instead of 87,*
- *had 600 passes instead of 594, and*
- had 11 attempts instead of 12.

This counterfactual in counter-intuitive: the user expects to see a change for feature 'Goals Scored' from 2 to 0, but increasing both pass accuracy and passes is impossible to integrate into the prior mental model. If the user came across a counterfactual similar to this example the explanations would not be accepted.
The underlying reason is again the range of values selected for perturbation, the local dataset tested permutations with pass accuracy of 85% and 88% but not 87% and 540 and 600 passes but not 594. The counterfactual dataset with pass accuracy 88% and 600 passes had the smallest Euclidian distance to the original observation.

4.4 Qualitative User Research

The qualitative user research set out to answer how the use of visualisation and verbalisation provides an explanation of an AI/ML model prediction to the user or affected individual of the system. The two different type of explainers were tested with five domain experts. The following sections show the key findings.

4.4.1 Non-interpretable Model Explainer Explanation Evaluation

Does the explanation explain the prediction?

All participants had a prior mental model of which features and which data range for these features should predict the MotM. Whether participants felt that an explanation explained a prediction was dependent on their prior mental model and / or their willingness to amend their prior mental model. If it did not fit the explanation was rejected. During the interview the participants tried to imagine the match (or remember it, had they seen it) using the explanations. They then tried to consolidate the model's prediction, their prior mental model, and the explanation. This resulted in the following scenarios:

- The explanation and the prior mental model correspond: the explanation is accepted. The participants thought that the explanation is indeed explaining the prediction.

- The explanation and the prior mental model do not correspond, either because the feature itself, the feature rank, or the value was not part of the prior mental

model.The participants then tried to fit the explanation into their mental model.

  - The explanation is intuitive and could be included into the prior mental model, this peaked their interest, and the explanation is accepted.
  - The explanation could not be included into the prior mental model, and the explanation is rejected.

Verbalisation: do explanation and counterfactual inform each other?

Similar to the result for the previous question the participants tried to consolidate the counterfactual and their prior mental model.

The following scenarios were observed:

- If the counterfactual fits the prior mental model it is accepted and informs the participant. One participant ignored the counterfactual if it was 'obvious'.
- The counterfactual does not fit the prior mental model, specially if caused due to the computational issue of counterfactuals describe in section 4.3.2:
  - The counterfactual is rejected.
  - The counterfactual confuses the participant.

  Both cases result in the rejection of the explanation and a lack of trust in the system.

If accepted, the participants considered the counterfactual very useful both to understand the explanation and to build a mental model of the AI, as described by T. Miller, 2018. The only rejected counterfactuals encountered during the experiment were caused by the computational issue of counterfactuals described above.

NOTE: In the first experiment the counterfactuals listed all features, including the ones which had not changed. The word 'instead' was used to highlight a change in data. This proved to be confusing and hard to understand. It was therefore changed for the subsequent interviews. Only the changed features, the old value for class fact, and the new value for class foil were displayed. This was done so that the following participants could immediately grasp the changes rather than having to look for them and could therefore evaluate the counterfactual on its own merit.

Model Understanding

Does the participant understand how the system works?

Due to time constraints the participants explored between 3 and 5 games, and these were not enough for the participants to fully understand how the system works.

It was therefore not feasible for the participants to consolidate all experiments for the Non-interpretable Model Explainer to rank the other features by importance, specially if a feature used by the model did not correspond to the prior mental model.

If the explanation and the counterfactual corresponded to the prior mental model or could be consolidated into the prior mental model the explanation was generally accepted and deemed satisfactory.

Does the participant pick up feature importance?

All participants picked up 'Goals Scored' as the most important feature in most of the predictions. This, however, corresponds to their prior mental model, and it is difficult to say if they derived this by observing and evaluating the explanations or if it was a confirmation of their existing beliefs. They observed, understood, and valued that the predictions used different features in different order of importance.

Can the participant construct a counterfactual?

All participants were able to create a counterfactual changing the number of goals scored, but failed to add other features. The most important feature in the prior mental model of the participant corresponds to the most important feature in the counterfactual generated by the system.

Visualisation and Verbalisation

Does the interactive scatter plot encourage the user to explore the system?

None of the participants thought the scatter plots with clusters of teams were useful or intuitive. In fact only two out of five participants were actively exploring the games. The glyphs deliberately only showed the game on mouse hover to encourage random selections of games rather than search for specific games, however the two participants did actively look for games they remembered. Both were not interested in teams that did not have the MotM.

In case of the Non-interpretable Model Explainer the x-axis represents the probability of the prediction, and one of the active participants used this deliberately to search for games.

4.4.2 Interpretable Model Explainer

Explanation Evaluation

Does the explanation explain the prediction?

All the participants evaluated the predictions based on the decision tree that is displayed in the Interpretable Model Explainer (see the next section) rather than on the merit of the explanation itself.

An explanation is accepted if the decision path uses features and thresholds in its decision nodes that do correspond to the prior mental model. An explanation has lower acceptance if its decision path passes through nodes that do not correspond to the prior mental model of the participant.

Model Understanding

Does the participants understand how the system works?

The participants understand how the model derives a prediction, it is they understand the inner workings of the model, the algorithm, and the decision paths. They

studied the visualisation of the decision tree in detail. Three thought the logic was easy to follow. All participants questioned some of the features and thresholds. Two participants used the colour coding of the nodes representing the majority class in the node to further their understanding of the model.

All participants wanted to confirm that the decision tree was trained by an ML model rather than a human, and one thought it would be more interesting to understand the training algorithm rather than the resulting tree.

Do the global visualisation of the decision tree and the individual prediction inform each other?

Two participants commented that understanding the classification algorithm made it more trustworthy. The algorithm allowed them to compare it directly to their prior mental model. Four of the participants found the rules restrictive, limiting the ability to predict complex scenarios "it's a bit too simplistic ... it kind of weakens its authority" (participant 5).

Does the participant pick up feature importance?

All participants studied the visualisation of the decision tree and understood the features used and their order. All agreed with the root node (goal scored or not), but all questioned some of the features and thresholds used (e.g. distance covered), and these features had a lower acceptance.

Can the participant construct a counterfactual?

All participants were able to create a counterfactual using the decision tree and constructing an alternative decision path, even though they contested some of the features used (specially distance covered, but also fouls).

Visualisation and Verbalisation

Does the double-encoding of visualisation of the decision tree and verbalisation of local explanations inform each other?

The participants did not think the double encoding was useful. Four thought the double encoding of visualisation and verbalisation was redundant. Of these four, one preferred the visualisation and suggested highlighting the path and integrating the text. The other three preferred the verbalisation.

The remaining participant had an aversion to flow charts and preferred to have just the text, rejecting the visualisation outright.

Does the interactive scatter plot encourage the user to explore the system?

Both explanation applications used the scatter plot of the random forest for consistency and to allow the users to compare the two systems using the same game. The downside however was that the x-axis was meaningless in when used in combination with the decision tree. This confused the participants, and they did not interact with the scatter plot.

However, once the participants understood the algorithm of the decision tree they lost interested in exploring specific games and explanations. The scatter-plot had already lost the participant's interest.

Three participants found the number of nodes was too much, while one thought the tree depth of 5 was too short.

4.4.3 Model Comparison

Which model does the participant prefer and why?

Three participants prefer the Non-interpretable Model Explainer no matter which model they saw first. All three found that the verbal explanation and counterfactual provided them with a better understanding of the prediction.

- Research participant 2 likened the generated text to a human-like verbal explanation that leads to understanding. The Interpretable Model Explainer is not as convincing because its explanation is based on the algorithm, and to accept the algorithm the participant wishes to understand the logic that derived it. The participant also complained about information overload caused by the complexity of the decision tree. This participant saw the Interpretable Model Explainer first, the Non-interpretable Model Explainer second.

- Research participant 3 compared the verbalisation to a human-like conversation as well: the AI presents its reasons in form of explanation and counterfactual, the human responds, and a debate ensues. The participant found this interaction appropriate for a subjective and emotional matter such as choosing the MotM. This participant saw the Non-interpretable Model Explainer first, the Interpretable Model Explainer second.

- Research participant 4 found the verbalisation well presented, sufficiently descriptive, and easy to follow. The decision tree looks dated in comparison, and the decision path is not beneficial. This participant saw the Interpretable Model Explainer first, the Non-interpretable Model Explainer second.

Only one participant preferred the Interpretable Model Explainer due to the transparency of the algorithm. This allowed the participant to directly compare it to their prior mental model. The participant however did find the counterfactuals of the Non-interpretable Model Explainer very informative. This participant saw the Non-interpretable Model Explainer first, the Interpretable Model Explainer second.

The last research participant saw the Non-interpretable Model Explainer first. The participant had no preference, but weighted both against each other:

- The participant valued the transparency of the algorithm of the decision tree, but found it too rigid and thought not enough features were taken into account into deriving the decision. The decision tree simplifies too much, which weakens its authority.

- The Non-interpretable Model Explainer was intriguing to the participant because the explanation and counterfactual shows the smallest change for a prediction to flip to the opposite class. The participant called it interesting, "in a blowing your mind kind of way". The participant disapproves the lack of transparency with which the explanations are generated.

How does accuracy influence the preference?

After asking their preference the participants were informed that the non-interpretable model was more accurate in its prediction than the interpretable model (88% vs. 92%). Due to overrunning times research participant 2 was not asked this question.

Research participant 4 and research participant 5 were not surprised, as they thought that the Non-interpretable Model Explainer used a wider variety of features tailored to the specific prediction while the decision tree felt more limited and simplistic.

Research participant 3 felt that accuracy in this subjective and emotional matter is of no consequence. The participant appreciated the human-like nature of explanation and counterfactual of the Non-interpretable Model Explainer.

Research participant 1 was surprised. The participant linked transparency to accuracy. The decision tree had gained trust because the root node (Goal Scored) corresponded to their prior mental model. Reflecting on this information the participant reached the same conclusion as the other three, that the Non-interpretable Model Explainer took a wider range of features into account and varied them for individual predictions.

Chapter 5

Discussion

5.1 Evaluation of Objectives

5.1.1 Identify Requirements and Evaluation Measures for an Explanation

Objective: Identify the requirements for an explanation in the human context. Compile evaluation measures to assess the comprehensibility of the explanations.

The targeted stakeholder of this project is the domain expert, but layperson in regards to ML and AI. The literature review resulted in detailed information about explanations in a human context, which is the research focus of this project. Insights are gathered from the fields of artificial intelligence, human-computer interaction, social sciences, cognitive sciences, and psychology and provide a rich source of information.

The literature also revealed comprehensive insights into evaluating explanations, going back to evaluation of expert systems. This includes measuring the effectiveness of an explanation, the subjective satisfaction with the explanation, and the comprehensibility and understandability of explanations.

The objective was met and laid the foundation for the explanation generation and the qualitative user research.

5.1.2 Domain, Target Audience, and Data Set

Objective: Select a domain, a target audience, and an appropriate data set.

This objective was a prerequisite, because the explanations has to be targeted to the domain and the intended stakeholder. The challenge was access to both domain data and domain experts.

Meeting this objective proved surprisingly difficult. A number of possibilities for domains and stakeholders did not come through, and football was chosen as a fall-back. One downside was the size of the data set. Training ML models with a total of 128 observations limited the selection of algorithms to select.

The project proceeded, keeping these limitations in mind.

5.1.3 Explainer Application

Objective: Build applications that explain the predictions of different ML models.

The design space proposed by Sevastjanova et al., 2018 helped to narrow down the choice to the two implemented approaches, a Non-interpretable Model Explainer generating verbal explanation and counterfactual for selected observations, and an Interpretable Model Explainer with a visual representation of the algorithm and also the option for verbal explanations of selected observations.

The framework yielded a solid foundation for the design of both explainer applications. Common to both was the interactive feature to select individual explanations. Both applications provided counterfactuals, the Non-interpretable Model Explainer explicit as part of the verbal explanation, the Interpretable Model Explainer implicit in the visualisation of decision tree.

Both explainer applications had downsides that became apparent during the qualitative user research. A common downside was the scatter plot that failed to engage the participants. This was a case of research bias, the author was so focused in showing clusters that he neglected to take the subjective aspect of the domain into account. One participant suggested using country flags or football kits as glyphs. It is worth investigating if this approach would have encouraged a more active discovery of individual predictions and observations, and what effect this would have on the understanding of the system as a whole.

A drawback in the implementation of the Non-interpretable Model Explainer was the computation of the counterfactuals. The lack of domain knowledge of the researcher meant that counter-intuitive counterfactuals were not identified, and the issue was not caught before the interviews.

The use of the Python Bokeh library proved to be straight forward. The development of visualisations follows the 'grammar of graphics' and enabled the researcher to follow known patterns. Deploying the applications in a browser facilitated the qualitative user research.

5.1.4 Qualitative User Research

Objective: Conduct qualitative user research to evaluate the comprehensibility of the generated explanations.

The research participants represented an unexpected opportunity. Although the domain was subjective and emotional, all had a wealth of domain knowledge, a well founded prior mental model on how to pick the MotM, and were very vocal during the interviews. It should be noted that the researcher has limited knowledge of football and approached the domain as an outsider.

The interviews yielded many interesting insights that flowed into the evaluation and will be discussed in the next section

5.1.5 Evaluation of Approaches to Generate Explanations

Objective: Evaluate and critically reflect the findings gained in the qualitative user research of the different approaches to generate the explanations. Report the findings

Presenting counterfactuals that were either confusing or contradictory led to the rejection of the explanation of selected observations. This issue was not part of the proposed research, but rather an interference that detracts. Intuitive counterfactuals, even if they did not correspond to the prior mental model of the participant, were considered valuable, interesting, and thought provoking. /newline
Despite this limitation the following key themes can be identified:

- **The prior mental model is key in the acceptance of the explanation:** Explanations and counterfactuals are perceived as convincing, satisfactory, and useful as long as they fit or can be included into the participant's prior mental model. Once either explanation or counterfactual does not fit into the prior mental mode they are rejected. This was particularly the case when counterfactuals were perceived as counter-intuitive.

- **Cognitive complexity is at odds with perceived value of the prediction:** The decision tree was considered too complex to be grasped at once and kept in short term memory. It was, however, deemed too simplistic and rigid to generate good decisions. A major criticism was the apparent lack of features used in the prediction.

- **Appropriate visualisation is fundamental to user engagement:** The type of visualisation used in the interactive scatter plot was not engaging enough and participants therefore preferred engagement with the verbalisation.

5.2 Research Questions

5.2.1 Can Visualisation and Verbalisation Provide an Explanation?

Main research question: Can the use of visualisation and verbalisation provide an explanation of an AI/ML model prediction to the user or affected individual of the system?

This research cannot answer the question. The proposed approaches generate explanations that were not satisfactory in specific scenarios. The reasons for rejection however can be identified. The prior mental model of the user is key in the evaluation and acceptance of the explanation. The verbal counterfactuals generated in the proposed application for specific observations contradicts the prior mental model of the user, which results in their rejection. However the issue of the counterfactual is of algorithmic rather than fundamental nature and can be addressed.

If explanation and counterfactual align with the prior mental model, even if they do not correspond, the research participants considered them understandable, comprehensible, and useful. Further research in this area is therefore necessary.

The same criticism applies to the proposed visualisation. They did not encourage the users to explore the system and were not appropriate for the subjective domain of the MotM in a world cup football match. Visualisations more suited to the domain should be explored.

5.2.2 Which Approaches Can Be Used?

Supporting research question: Which approaches can be used to generate these explanations?

This project focused on two different approaches:

- A model-agnostic based verbal explanation in combination of any domain-appropriate ML model. The explanations for this approach consist of a verbal explanation in combination with a verbal counterfactual. This approach yielded promising results if implemented correctly. The participants responded favourably to the human aspect of the generated texts.

- An interpretable model with a visualisation of the underlying algorithm and a verbalisation of user-selected model outputs. Complexity (along with accuracy) was a selection criteria, and the chosen decision tree had a maximum depth of 5. The trained tree had 11 decision nodes and 13 leaf or prediction nodes, but was considered too complex to grasp on a global level by the research participants. When evaluating individual predictions however the algorithm was considered too rigid without taking an appropriate number of features into account. This contradiction puts the use of a similar interpretable model in question in a scenario which uses 21 features.

5.2.3 How Effective Are These Approaches?

Supporting research question: How effective are the approaches, considering different scenarios, tasks, and algorithms?

The focus of this project was reduced to one task (binary classification), one scenario (predict which team would have the MotM), two approaches as described in the previous question. The effectiveness is therefore discussed in the primary research question.

Chapter 6

Evaluation, Reflections, and Conclusions

6.1 Choice of Objectives

The choice of objectives seems adequate for this project, specially after reducing the scope to one task, one scenario, and two explanation approaches with a binary classification.

Explainable AI is a very active field of research, exemplified by the DARPA XAi programme (Gunning, 2017) and the number of new papers published on platforms such as arxiv. Researching literature for the context was therefore very rewarding

The identification of a suitable domain, domain experts, and dataset almost proved to be a stumbling block. Two fall-back scenarios were considered, one using a generic data-set (such as the Titanic or the UCI Adult one) and changing the target audience from from domain-experts to individuals, or changing from a professional to a leisure field and use a sports theme and sport aficionados. The original intention was to evaluate explanations given domain knowledge and prior mental models with the research participants, and the latter option was chosen.

As discussed in the chapters dealing with results and discussion the proposed application failed in the task of generating counterfactuals. The root cause is identified and reflects more on the inexperience of the author rather than a fundamental issue. A renewed literature search returned a paper uploaded to arxiv in June of this year by Rathi, 2019 which promises valuable insights into this issue.

The qualitative user research generated useful data nevertheless, and the evaluation of explanation and counterfactuals as suggested by T. Miller, 2018 shows promising results, which have to be confirmed by future research.

6.2 Limitations of the Project

The original research question could not be answered. This project however uncovered the obstacles that led to that outcome and offers suggestions on how to overcome them.

6.3 Future Work

Addressing the issue of counter-intuitive counterfactuals is an immediate follow-up. The root cause and possible solutions are identified, but too late to be included into this research.

The computation of effective counterfactuals is an interesting field. A counterfactual that enhances the comprehensibility of an explanation and provides the user with new insights needs to be as 'close' to the explained observation, and the definition of the distacne function is not trivial. The data used in this project included only continuous numerical featues, but including categorical values increases the level of complexity.

Varying the number of features used in an explanation is another direction for further research. This project fixed the number to five, variations with both more or less features and conducting further qualitative research could identify an optimum.

The explanations generated for this project only included SHAP values that supported the prediction. Including features that push the prediction in the opposite direction (see table 4.3) would provide richer, more balanced and nuanced explanations, and research in the understandability and perceived value could prove interesting.

6.4 Reflection on the Project

The failure to provide an answer to the research question is a major disappointment. In hindsight this could have been prevented by using an iterative, agile approach, where each prototype would have been tested with users and would inform the next. Both counterfactuals and lack of engagement could have been identified and addressed. It is worth speculating that even a second iteration would have produced two applications that, once evaluated in qualitative user research, would have generated more valuable data. This approach was not feasible due to the time constraints of the project and access to research participants.

The research interview itself would benefit from a second attempt, listening back to the original audio recordings demonstrated the in-experience of the author in this field, and the experience gained in the first round would improved the second one considerably.

The project produced interesting insights to the author nevertheless, and conducting the interview with engaged and passionate participants was a rewarding experience.

Bibliography

ADADI, Amina; BERRADA, Mohammed, 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. Vol. 6, pp. 52138–52160.

ALVAREZ-MELIS, David; JAAKKOLA, Tommi S, 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

BIRAN, Or; COTTON, Courtenay, 2017. Explanation and justification in machine learning: A survey. In: *Explanation and justification in machine learning: A survey. IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8, p. 1.

BOKEH, 2019. *Bokeh 1.3.4 documentation* [online] [visited on 2019-07-14]. Available from: `https://bokeh.pydata.org/en/latest/index.html`.

BRAUN, Virginia; CLARKE, Victoria, 2013. *Successful qualitative research: A practical guide for beginners*. Sage.

DEEMTER, Kees Van; THEUNE, Mariët; KRAHMER, Emiel, 2005. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*. Vol. 31, no. 1, pp. 15–24.

DORAN, D; SCHULZ, SC; BESOLD, TR, 2018. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In: *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. CEUR Workshop Proceedings*. Vol. 2071.

DOSHI-VELEZ, Finale; KIM, Been, 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

DOŠILOVIĆ, Filip Karlo; BRČIĆ, Mario; HLUPIĆ, Nikica, 2018. Explainable artificial intelligence: A survey. In: *Explainable artificial intelligence: A survey. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215.

FIFA.COM, 2019. *2018 FIFA World Cup Russia™ - Man of the Match Rules - FIFA.com* [online] [visited on 2019-07-18]. Available from: `https://www.fifa.com/worldcup/awards/man-of-the-match/rules`.

FISHER, Aaron; RUDIN, Cynthia; DOMINICI, Francesca, 2018. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*.

GALLETTA, Anne; CROSS, William E., 2013. *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication*. NYU Press.

GATT, Albert; KRAHMER, Emiel, 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*. Vol. 61, pp. 65–170.

GIVEN, Lisa M., 2008. *The Sage encyclopedia of qualitative research methods*. Sage Publications.

GOEBEL, Randy; CHANDER, Ajay; HOLZINGER, Katharina; LECUE, Freddy; AKATA, Zeynep; STUMPF, Simone; KIESEBERG, Peter; HOLZINGER, Andreas, 2018. Explainable AI: the new 42? In: *Explainable AI: the new 42? International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 295–303.

GUIDOTTI, Riccardo; MONREALE, Anna; RUGGIERI, Salvatore; TURINI, Franco; GIANNOTTI, Fosca; PEDRESCHI, Dino, 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*. Vol. 51, no. 5, pp. 93.

GUNNING, David, 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*. Vol. 2.

HILTON, Denis J, 1990. Conversational processes and causal explanation. *Psychological Bulletin*. Vol. 107, no. 1, pp. 65.

JONES, Natalie; ROSS, Helen; LYNAM, Timothy; PEREZ, Pascal; LEITCH, Anne, 2011. Mental models: An interdisciplinary synthesis of theory methods. *Ecology and society*.

KAGGLE, 2019a. *Machine Learning Explainability — Kaggle* [online] [visited on 2019-07-18]. Available from: `https://www.kaggle.com/learn/machine-learning-explainability`.

KAGGLE, 2019b. *Predict FIFA 2018 Man of the Match — Kaggle* [online] [visited on 2019-07-18]. Available from: `https://www.kaggle.com/mathan/fifa-2018-match-statistics`.

KING, Nigel; BROOKS, Joanna, 2018. *The sage handbook of qualitative business and management research methods: Thematic analysis in organisational research*. SAGE Publications Ltd.

KIRK, Andy, 2016. *Data visualisation: a handbook for data driven design*. SAGE.

KIRSCH, Alexandra, 2017. Explain to whom? putting the user in the center of explainable AI. In: *Explain to whom? putting the user in the center of explainable AI*.

KULESZA, Todd; STUMPF, Simone; BURNETT, Margaret; YANG, Sherry; KWAN, Irwin; WONG, Weng-Keen, 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In: *Too much, too little, or just right? Ways explanations impact end users' mental models. 2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pp. 3–10.

LIPTON, Zachary C, 2018. The mythos of model interpretability. *Communications of the ACM*. Vol. 61, no. 10, pp. 36–43.

LIU, Mengchen; SHI, Jiaxin; LI, Zhen; LI, Chongxuan; ZHU, Jun; LIU, Shixia, 2016. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*. Vol. 23, no. 1, pp. 91–100.

LIU, Shixia; WANG, Xiting; LIU, Mengchen; ZHU, Jun, 2017. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*. Vol. 1, no. 1, pp. 48–56.

LUNDBERG, Scott M; LEE, Su-In, 2017. A unified approach to interpreting model predictions. In: *A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems*, pp. 4765–4774.

MILLER, George A, 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*. Vol. 63, no. 2, pp. 81.

MILLER, Tim, 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

MILLS, Albert J; DUREPOS, Gabrielle; WIEBE, Eiden, 2010. Thematic analysis. *Encyclopedia of case study research*, pp. 926–928.

MITTELSTADT, Brent; RUSSELL, Chris; WACHTER, Sandra, 2019. Explaining explanations in AI. In: *Explaining explanations in AI. Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288.

MOHSENI, Sina; ZAREI, Niloofar; RAGAN, Eric D, 2018. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*.

MOLNAR, Christoph et al., 2019. Interpretable machine learning: A guide for making black box models explainable. *URL: https://christophm. github. io/interpretable-ml-book/, visited on 12/08/2019*.

MUNZNER, Tamara; MAGUIRE, Éamonn, 2015. *Visualization analysis & design*. CRC Press.

NARAYANAN, Menaka; CHEN, Emily; HE, Jeffrey; KIM, Been; GERSHMAN, Sam; DOSHI-VELEZ, Finale, 2018. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

NUNES, Ingrid; JANNACH, Dietmar, 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*. Vol. 27, no. 3-5, pp. 393–444.

OATES, Briony J., 2006. *Researching information systems and computing*. SAGE Publications.

OED, Oxford English Dictionary, 2019. *explanation, n. : Oxford English Dictionary* [online] [visited on 2019-04-20]. Available from: `https://0-www-oed-com.wam.city.ac.uk/view/Entry/66604?redirectedFrom=explanation#eid`.

PREECE, Alun; HARBORNE, Dan; BRAINES, Dave; TOMSETT, Richard; CHAKRABORTY, Supriyo, 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.

PYDATAAMSTERDAM, 2019. *Hidde Hovenkamp: SHAP and Beyond — PyData Amsterdam 2019 - YouTube* [online] [visited on 2019-08-21]. Available from: `https://www.youtube.com/watch?v=xwl8WhtJNs0`.

PYDATANYC, 2018. *Introduction to Model Interpretability at PyData, NYC* [online] [visited on 2019-08-12]. Available from: `https://github.com/klemag/pydata_nyc2018-intro-to-model-interpretability`.

RATHI, Shubham, 2019. Generating Counterfactual and Contrastive Explanations using SHAP. *arXiv preprint arXiv:1906.09293*.

REITER, Ehud, 1995. NLG vs. templates. *arXiv preprint cmp-lg/9504013*.

REITER, Ehud; DALE, Robert, 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos, 2016a. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos, 2016b. Why should I trust you?: Explaining the predictions of any classifier. In: *Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

RIBERA, Mireia; LAPEDRIZA, Àgata, 2019. Can we do better explanations? A proposal of user-centered explainable AI. In: *Can we do better explanations? A proposal of user-centered explainable AI. IUI Workshops*.

SALKIND, Neil J., 2010. *Encyclopedia of research design*. Sage Publications, Inc.

SEVASTJANOVA, Rita; BECK, Fabian; ELL, Basil; TURKAY, Cagatay; HENKIN, Rafael; BUTT, Miriam; KEIM, Daniel A; EL-ASSADY, Mennatallah, 2018. Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models. In: *Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models. Workshop on Visualization for AI Explainability at IEEE VIS*.

SKLEARN, 2019. *scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation* [online] [visited on 2019-07-30]. Available from: `https://scikit-learn.org/stable/index.html`.

TOMSETT, Richard; BRAINES, Dave; HARBORNE, Dan; PREECE, Alun; CHAKRABORTY, Supriyo, 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

VAN SOMEREN, MW; BARNARD, YF; SANDBERG, JAC, 1994. *The think aloud method: a practical approach to modelling cognitive*. Citeseer.

WAA, Jasper van der; ROBEER, Marcel; DIGGELEN, Jurriaan van; BRINKHUIS, Matthieu; NEERINCX, Mark, 2018. Contrastive explanations with local foil trees. *arXiv preprint arXiv:1806.07470*.

WACHTER, Sandra; MITTELSTADT, Brent; RUSSELL, Chris, 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. Vol. 31, no. 2.

WIKIPEDIA, 2019. *Player of the match - Wikipedia* [online] [visited on 2019-07-18]. Available from: `https://en.wikipedia.org/wiki/Player_of_the_match`.

ZEMLA, Jeffrey C; SLOMAN, Steven; BECHLIVANIDIS, Christos; LAGNADO, David A, 2017. Evaluating everyday explanations. *Psychonomic bulletin & review*. Vol. 24, no. 5, pp. 1488–1500.