university of groningen / faculty of science and engineering

Explanation methods for regression tasks

Bachelor's Project Thesis

Julia Boers, s3632644, j.j.a.boers@student.rug.nl,
Supervisor: Dr. Matias Valdenegro-Toro

Abstract: Explainable artificial intelligence (XAI) is the field that aims to make machine learning models explainable, e.g. by developing explanation methods. The majority of research on this topic is focused on classification problems, while real-world applications are often regression problems, and explanation methods developed for classification cannot thoughtlessly be applied to regression. In this Bachelor thesis three attribution-based explanation methods are compared when applied to regression models. Two gradient-based explanation methods, Guided Backpropagation (GBP) and Integrated Gradients (IG), and one model-agnostic method, Local Interpretable Model-agnostic Explanations (LIME), were applied to two different regression models, a wine quality prediction (WQP) model and an age prediction (AP) model. The explanations were evaluated using the Deletion Area Under the Curve (DAUC) and Insertion Area Under the Curve (IAUC) metrics and a user study was performed. The evaluations did not point to one best-performing explanation method. For the WQP model IG performed best according to the DAUC score, but not significantly. LIME performed best according to the IAUC score. For the AP model GBP performed best according to the DAUC, LIME performed best according to the IAUC score and IG received the most votes in the user study, although the differences in votes were not significant.

1 Introduction

1.1 Context

Explainable artificial intelligence (XAI) is the field concerned with making machine learning models interpretable. A model is interpretable if users can understand the reason why a model made a certain prediction [Miller, 2019]. XAI can be achieved by using intrinsically interpretable models or by applying post-hoc explanation methods to uninterpretable black box models [Ancona et al., 2019].

Within the literature the potential trade-off between interpretability and performance is discussed [Rudin, 2018]. In some situations complex models achieve a higher performance than interpretable models, due to their non-linearity. Tasks like image processing are more suitable to be solved by a black box model [Loyola-González, 2019]. On the other hand, Rudin [2018] argues that there is no significant difference between the performance of black box models and interpretable models, given that the training data is structured and meaningful. Post-hoc explanation methods aim to explain black box models and make it possible to get the performance without losing interpretability [Lipton, 2016]. Furthermore Loyola-González [2019] argues that it is not necessary for users to understand the inner workings of machine learning models if they are provided with an understandable explanation.

1.2 Motivation

XAI has several objectives: transparency, causality, privacy, fairness, trust, usability and reliability [Fiok et al., 2022]. Explanation methods can provide transparency, fairness and trust.

Transparency Transparent models are able to provide explanations for their decisions. Since 2018 it is required by the General Data Protection Regulation (GDPR) that companies that use machine learning models to aid with decision making provide algorithmic transparency [Wachter et al., 2017].

Fairness Models can be biased against certain populations if they are trained on biased data. For example, Amazon discovered a hiring tool they were developing was biased against women [Dastin, 2018]. Explanation methods can help detect biases in training data [Anders et al., 2022].

Trust Machine learning models are more likely to be deployed and users are more likely to to take action based on the model outcomes if they are trusted to behave reasonably [Ribeiro et al., 2016]. According to Jacovi et al. [2021] a model should reason in a way that is "agreeable" in order to establish trust. By explaining models users can decide if they trust the model based on how reasonable they think the model is.

1.3 Problem statement

As Letzgus et al. [2021] observed, most research on explanation methods is focused on explaining classification models, while many real world applications of machine learning are regression models (e.g. stock market prediction [Parmar et al., 2018], signal processing applications [Sarwate and Chaudhuri, 2013] and house price prediction [Madhuri et al., 2019]). Explanation methods may assume or are optimised for a categorical output of the underlying model, which is why they should not be applied to regression problems without consideration [Letzgus et al., 2021].

1.4 Objectives

In this Bachelor thesis the performance of three explanation methods applied to two different regression models is evaluated and compared. The regression models are a wine-quality prediction (WQP) model trained on tabular data and an age prediction (AP) model trained on image data. The explanation methods are Guided Backpropagation (GBP), Integrated Gradients (IG) and Local Interpretable Model-agnostic Explanations (LIME). Figure 1.1 shows a schematic overview of the comparisons.

The aim of this Bachelor thesis is formulated in the following research question: **How do different explanation methods compare when applied to regression models?**

The process of answering the research question is guided by the following sub questions:

1. How do different explanation methods of regression tasks on tabular data compare?
2. How do different explanation methods of regression tasks on image data compare?

1.5 Contributions

The contributions of this Bachelor thesis include an exploratory analysis of explanation methods applied to regression tasks. GBP, IG and LIME were evaluated and compared through two different quantitative evaluation methods and a user study.

2 State of the Art

2.1 Explanation method taxonomy

There are numerous explanation methods that can be categorised based on different properties.

Model-agnostic methods do not use any internal details of the model they explain. *Model specific* methods do use model internals, and can therefore only be applied to specific model classes [Molnar, 2022].

Global explanation methods explain the overall decision strategy of models, while *local* explanation methods only provide explanations for a single sample [Molnar, 2022].

Explanations can be presented in different forms, e.g. in the form of a counterfactual explanation [Wachter et al., 2017], a selection of features that had an effect on the model prediction, a full description of the feature attributions [Molnar, 2022], or in the form of an interpretable model that approximates the complex model [Molnar, 2022].

2.2 Attribution-based methods

This Bachelor thesis focuses on local attribution-based explanation methods. An attribution-based explanation consists of a vector of attributions of each input feature on the output.

$$R_i(x) = x_i \cdot \frac{\delta y_i}{\delta x_i}(x) \tag{2.1}$$

Equation 2.1 is a simplified description of an attribution-based explanation [Ancona et al., 2019]. R is a vector containing the attributions of all the

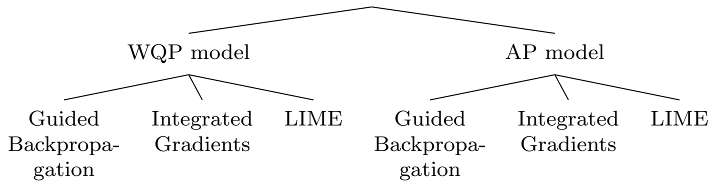


Figure 1.1: Overview of comparisons

input features x at instance i. Features with a high attribution, also called a high saliency, had a large effect on the model output. y_i is the output of the model at instance i. The explanation is equal to the model gradient at x_i, as the gradient is defined as the partial derivative of the output with respect to the independent variables.

In the case of image data the explanation can be visualised in a saliency map, where the vector is plotted as a heatmap over the original image, highlighting the pixels that had the greatest positive effect on the model prediction.

2.3 Explaining regression models

As mentioned above, most of the research on explanation methods is focused on classification problems. Bennetot et al. [2021] and Linardatos et al. [2021] provide flowcharts that help developers choose what explanation method is best suitable for their task. However, the distinction between classification and regression tasks is not made.

Explanation methods for regression models should receive more attention because the problems they are solving are inherently different from classification problems. Firstly, due of the difference in output type the the decision boundary for classification problems is more clearly defined than for regression problems, where ambiguities may occur. This notion is especially relevant when developing counterfactual explanation methods [Spooner et al., 2021].

Secondly, regression models can adhere to the conservation property, meaning that the output of the model can be expressed in the measurement unit of the data, which is often not the case for classification problems. It enhances interpretability if explanation models adhere to this property of conservation [Letzgus et al., 2021].

Finally, the explanation of a regression model needs to be compared to some reference value to contextualise the explanation. Letzgus et al. [2021] give an illustrative example involving an auction with two bidders. Bidder 1 bids €900, bidder 2 bids €1100. An explanation of the regression model predicting the final price based on these two bidders needs to be formulated around a reference value in order to make sense. The question "which of the two bidders had the greatest influence over the final price?" is difficult to answer. But formulate the question around a reference value: "which of the two bidders had the greatest influence over the final price being higher than €1000?", and the answer becomes clear: bidder 2. Current explanation methods often are implemented around implicitly defined reference values, and lack the possibility for the user to set them.

3 Methods

3.1 Datasets

The WQP is trained on a white wine quality dataset from Cortez et al. [2009], containing 11 independent variables describing properties of Portuguese wines. These features represent the physiochemical properties which can determine the quality of a wine: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

The AP model is trained on an image age prediction dataset from Frentescu [2020], containing 40.440 images of faces of people from the ages 20 to 50, divided into 33432 train images and 7008 test images.

3.2 Model architectures

3.2.1 Wine quality prediction model

The WQP model predicts the quality of wines, expressed as a value between 1 and 6, based on 11 descriptive properties called features. The architecture consists of an input layer of 11 neurons, representing the 11 features, followed by 2 hidden layers of 16 and 8 neurons respectively and an output layer. The source code of the model can be found in Appendix B. All layers are fully connected and modified by the the rectified linear unit (Relu) activation function. For the final layer the sigmoid activation function is used. The network is trained by minimizing the mean squared error (MSE) loss function, using the Adam optimiser.

The model was trained for 250 epochs and achieved a training error of 0.57 MSE and a validation error of 0.55 MSE.

3.2.2 Age prediction model

The AP model predicts the age of people based on an image of their face. The model is a residual neural network (Resnet) as originally proposed by He et al. [2015]. A Resnet is a convolutional neural network, but with so-called skip-connections, connecting layers of the neural network skipping some layers in between. This avoids the problem of vanishing gradients which occurs when the values of the gradients go to zero and the network cannot learn anymore. By adding skip-connections, the value of a few layers back is added to the weights of the current neuron, therefore the value is higher and does not "vanish" to zero.

The Resnet is based on an implementation from Singh [2019], with 20 residual layers, adapted to accommodate regression problems with a sigmoid activation after the residual layers. A final lambda layer was added to scale the model output. Appendix B contains the source code of the architecture. The parameters given were input shape $= (64, 64, 3)$ and depth $= 20$. The network is trained by minimizing the coefficient of determination (R^2) loss function, using the Adam optimiser.

The model was trained for 250 epochs, with a batch size of 32 and a learning rate of 1e-1 and achieved a training error of $0.54R^2$ and a validation error of $0.51R^2$.

3.3 Explanation methods

3.3.1 Guided Backpropagation

Guided Backpropagation [Simonyan et al., 2013] results in a saliency vector, describing the attribution of each input feature on the output. This is achieved by propagating backwards through the model, setting all the negative gradients to zero using the Relu activation function. Gradients represent the rate of change of the target value with respect to all the inputs, so by setting all the negative gradients to zero only the features with a positive effect on the model output will be left. This results in an explanation w as described in

$$w = \frac{\delta S_c}{\delta I} \mid I_0 \tag{3.1}$$

where S_c is a first-order Taylor expansion of the class score function $S_c(I)$, the output of a classification model for some class. I is the symbol for the input data and I_0 represents the specific sample. w then represents the derivative of the class score function with respect to I_0, meaning that it is a vector containing the attributions of the features of sample I_0.

3.3.2 Integrated Gradients

Integrated Gradients [Sundararajan et al., 2017] aims to satisfy the *Sensitivity* and *Implementation Invariance* axioms. The Sensitivity axiom requires that if two inputs to the model that differ in one feature and result in a different model prediction, then this feature should be given a non-zero attribution. It also requires that if the outcome of the model does not depend on some feature, this feature should be given a zero attribution. The Implementation Invariance axiom requires that attributions of two functionally equivalent networks are always identical.

The method starts by generating a baseline input x', and a straight line path through the feature space from the baseline input to the input x. Then the integral of the gradients along that path is calculated.

This process is described in the following formula:

$$IG_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha \times (x - x'))}{\delta x_i} d\alpha \tag{3.2}$$

where $F : \mathbb{R}^n[0,1]$ is the function of the to-be-explained model and

$$\frac{\delta F(x)}{\delta x_i} \tag{3.3}$$

represents the gradient of $F(x)$ with respect to x_i. The fraction inside the integral therefore represents the gradient of the variations of the inputs along the path, perturbed by interpolation constant α. Furthermore, $(x_i - x'_i)$ represents the distance of the difference from the currently considered perturbation to the baseline. Which means that the result of the integral is scaled to the distance of the input to the baseline, so gradients of inputs far from the baseline weigh heavier than those close to it.

In practice the integral is approximated using a Riemann sum (Equation 3.3.2), where m is the number of steps in the Riemann integral approximation, and the feature perturbation constant k is scaled.

$$IG_i(x) ::= (x_i - x'_i) \times \Sigma_{k=1}^{m} \frac{\delta F(x' + \frac{k}{m} \times (x - x'))}{\delta x_i} \times \frac{1}{m} \tag{3.4}$$

3.3.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016] is model-agnostic, meaning that it does not use any inner details of the underlying model, making it fundamentally different from the previously described methods. The aim of LIME is to generate an interpretable model (e.g. a linear model) that locally approximates the underlying complex model. LIME generates a custom dataset of inputs and outputs by perturbing the sample data and collecting the model outputs to train the interpretable model on. Formally the explanation ξ of sample x by LIME is defined by the following optimization formula:

$$\xi(x) = \underset{g \in G}{\arg\min}\, L(f, g, \pi_x) + \Omega(g) \tag{3.5}$$

where g is an interpretable model out of the class of potentially interpretable models G. As not every model of this class is actually readily interpretable, $\Omega(g)$ represents a measure of complexity that is added as a penalty for models that are too complex. L represents the loss function of the interpretable model, f represents the underlying model and π_x represents the neighbourhood of sample x. In simpler terms, LIME finds the optimal interpretable model in terms of the performance and complexity in the proximity of the complex model.

3.4 Evaluation

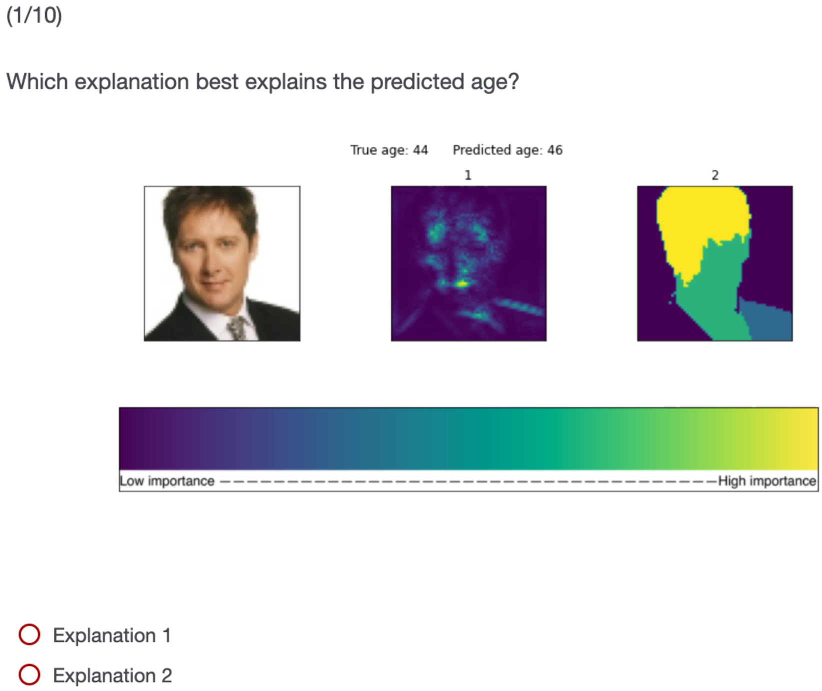
Evaluating and comparing different explainability methods is a difficult task because of the lack of a ground truth explanation [Kindermans et al., 2019], which is the true reason why a model gave a certain output. To find it is the goal of explanation methods. However, ideally the generated explanations are evaluated by comparing them to the ground truth explanation. As this is not possible, other methods have been developed to assess the quality of explanations. In this Bachelor thesis explanations were evaluated using two different evaluation metrics, and an additional qualitative evaluation was done on the explanations of the AP model in the form of a user study.

3.4.1 Quantitative evaluation

Deletion Area Under the Curve (DAUC) and Insertion Area Under the Curve (IAUC) are two quantitative evaluation metrics developed by Petsiuk et al. [2018].

DAUC step-wise removes the most salient features, according to the explanation, from the input and measures the change in model outcome. A big increase in error indicates that the deleted features were indeed important, and the explanation method was correct in assigning them a high saliency. Therefore a high area under the curve, in other words a high DAUC score, indicates a good explanation.

IAUC starts with a randomised version of the input and adds the most salient features. As features are added step-wise to the randomised input the change in model outcome is measured. A big decrease in error indicates that the added features were indeed important, meaning that a low area under the curve, in other words a low IAUC score, indicates a good explanation.


(1/10)

Which explanation best explains the predicted age?

○ Explanation 1

○ Explanation 2

Figure 3.1: Example of a survey question

3.4.2 Qualitative evaluation

In order to determine how agreeable the explanations are according to users, a user study in the form of a survey was conducted on the explanations of the AP model. The goal of the survey was to determine which saliency map users most agree with, which would indicate if the most salient pixels according to the explanation method match with the area on the image that people would look at to determine the persons' age.

The survey consisted of 10 questions, randomly picked from a question pool of 474 questions. Each question showed a reference image and two different saliency maps, as can be seen in Figure 3.1. The saliency maps displayed the results of two of the explanation methods randomly. The question also showed the actual age and the predicted age of the person on the image. Participants were asked to select the saliency map that they thought best explained the predicted age.

Before the survey started, participants were introduced to the basic idea behind black box models and explanation methods, and an example was given on how to reason about saliency maps. This introductory text can be found in Appendix C.

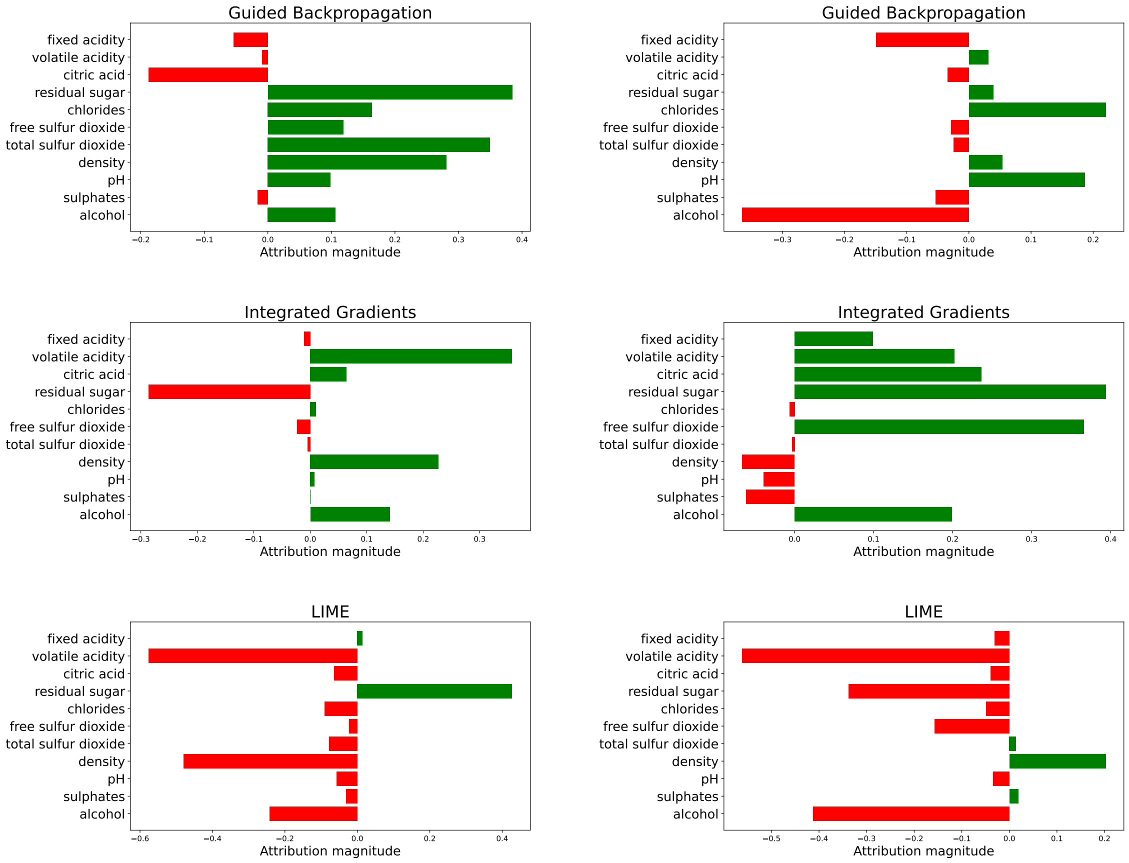
Finally, participants were asked to indicate if they had experience working or studying in the field of AI because the survey topic was domain-specific. If the answers of the two groups would be different, this might indicate that not all participants understood what was asked of them due to a flaw in the introductory text.

4 Results

In this section the results of the explanation methods are shown and evaluated. Section 4.1 displays the results of the explanation methods. Sections Section 4.2 and 4.3 evaluate the explanations of the wine quality prediction (WQP) model and age prediction (AP) model respectively.

4.1 Explanations

Figure 4.1 shows examples of WQP model explanations. The three figures on the left display the different explanations of a low-error sample. The true



True quality: 2 Predicted quality: 2.31 **True quality: 3 Predicted quality: 1.48**

Figure 4.1: Explanations generated by GBP, IG and LIME of the WQP model

quality of this sample was 2, the WQP model predicted it to be 2.3. The three figures on the right display the different explanations of a high-error sample. The true quality of this sample was 3, the WQP model predicted it to be 1.4.

The size of the bars indicate the magnitude of the effect of that feature on the model prediction if all other features would remain constant. The color and direction indicate whether this effect was positive or negative. The colors indicate an increase (green) or decrease (red) in model error.

Figure 4.2 shows examples of AP model explanations. The figure on the left displays the different explanations of a low-error sample. The true age of the person on the image was 23, the AP model predicted it to be 18. The figure on the right displays the different explanations of a high-error sample. The true age of this person was 35, the AP model predicted it to be 96.

The colors of the saliency maps indicate the magnitude of the positive effect of that pixel on the model prediction if all other pixels would remain constant. A yellow color indicates a high importance. In case of the low error example it can be seen that GBP, IG and LIME highlight the shoulder, face and part of the background, and hair and jawline respectively. For the high error example GBP and IG highlight the face, while LIME focuses on the forehead and ear.

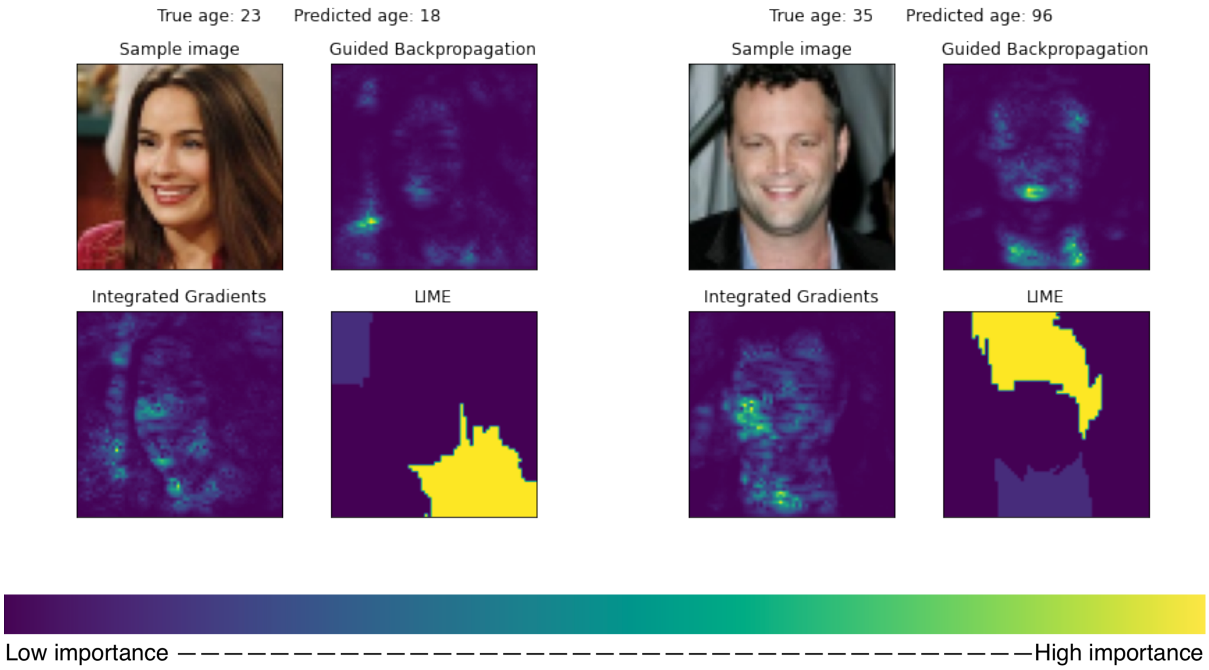


Figure 4.2: Explanations generated by GBP, IG and LIME of the AP model

4.2 Evaluation of WQP model explanations

4.2.1 Evaluation individual sample

Figure A.1 shows error curves of DAUC and IAUC applied to WQP model explanations of an individual data sample with true quality 3 and predicted quality 1.98.

According to the DAUC curves all three methods do not perform ideally. The error decreases rather than increases, as the most salient features are deleted. According to the DAUC score LIME provides the best explanation.

The IAUC curves show that the error decreases as the first features are inserted. However, especially for LIME the error increases after that, meaning that the quality prediction becomes worse as more features are added to the sample. According to the IAUC score IG provides the best explanation.

4.2.2 Average evaluation

Figure A.2 shows the error curves for DAUC and IAUC applied to the WQP model explanations, averaged over multiple samples. The graphs show the change in error as features are deleted/added (depending on the evaluation method), averaged over 500 samples.

The DAUC curves show that the error increases fast, as is expected to happen for good explanations. According to the DAUC score IG best explains the underlying model. A one-way ANOVA revealed that there is no statistically significant difference in DAUC scores between at least two methods ($F(2, 1497) = 0.35, p = 0.70$).

The IAUC curves show that the error increases as more features are added to the sample. LIME is the best method according to the IAUC score. A one-way ANOVA revealed that there is a statistically significant difference in IAUC scores between at least two methods ($F(2, 1497) = 219.80, p =< 2 \cdot 10^{-16}$). Tukey's HSD Test for multiple comparisons found that the mean value of IAUC score is significantly different between IG and GBP ($p = 0.50 \cdot$

$10^{-3}, 95\%$ C.I. $= 1.05, 4.52$), between LIME and GBP ($p =< 0.01 \cdot 10^{-3}, 95\%$ C.I. $= -13.55, -10.08$) and between LIME and IG ($p =< 0.01 \cdot 10^-3, 95\%$ C.I. $= -16.33, -12.86$).

4.3 Evaluation of AP model explanations

4.3.1 Evaluation individual sample

Figure A.3 shows error curves of DAUC and IAUC applied to the AP model explanations of an individual data sample.

The DAUC curves initially increase for all three explanation methods, indicating good explanations, but after that the error fluctuates. According to the DAUC score LIME provides the best explanation.

The IAUC curves show that only for GBP the error decreases as the most salient pixels are inserted, indicating a good explanation. However, according to the IAUC score LIME provides the best explanation.

4.3.2 Average evaluation

Figure A.4 shows the average curves of DAUC and IAUC applied to the AP model explanations, averaged over multiple samples. The graphs show the change in error as features are deleted/added (depending on the evaluation method), averaged over 500 samples.

The DAUC curves do not look as expected. Instead of the error increasing as pixels are deleted, the error decreases instead. According to the DAUC score GBP performs best. A one-way ANOVA revealed that there is a statistically significant difference in DAUC scores between at least two methods ($F(2, 1497) = 10.19, p = 4.04 \cdot 10{-}5$). Tukey's HSD Test for multiple comparisons found that the mean value of DAUC scores is significantly different between LIME and GBP ($p = 0.17 \cdot 10^{-3}, 95\%$ C.I. $= -130.10, -34.33$), and between LIME and IG ($p = 0.48 \cdot 10^{-3}, 95\%$ C.I. $= -124.98, -29.22$). There is no statistically significant difference between IG and GBP ($p = 0.97$).

The IAUC curves of GBP and IG also initially do not show expected results as the error increases at first, but then the error does go down when about 40% of the pixels are added. According to the IAUC score LIME performs best. A one-way ANOVA revealed that there is a statistically significant difference in IAUC scores between at least two methods ($F(2, 1497) = 9.55, p = 7.58 \cdot 10^{-5}$). Tukey's HSD Test for multiple comparisons found that the mean value of IAUC scores is significantly different between LIME and GBP ($p = 0.25 \cdot 10^{-3}, 95\%$ C.I. $= -107.65, -27.22$) and between LIME and IG ($p = 0.91 \cdot 10^{-3}, 95\%$ C.I. $= -102.19, -21.75$. There is no statistically significant difference between IG and GBP ($p = 0.95$).

4.4 Results user study

57 respondents filled in the survey, of which 32 had no experience working or studying in the field of AI and 21 had a little or a lot of experience.

Figure 4.3 shows the vote counts of respondents without AI experience. IG received the most votes (203), followed by LIME (186) and GBP (171). A chi-square test of independence was performed and showed no significant differences between the vote counts: $\chi^2(2, N = 1) = 0.8, p = 0.66$.

Figure 4.4 shows the vote counts of respondents with AI experience. IG received the most votes (141), followed by GBP (133) and LIME (126). A chi-square test of independence was performed and showed no significant differences between the vote counts: $\chi^2(2, N = 1) = 2.7, p = 0.25$.

Figure 4.5 shows the results of the two groups together.

Overall IG received the most votes (344), followed by LIME (312) and GBP (304). However a chi-square test of independence showed no significant difference between the vote counts: $\chi^2(2, N = 1) = 2.8, p = 0.25$.

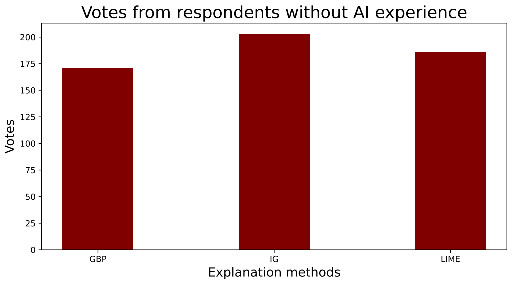


Figure 4.3: Vote counts from respondents without AI experience

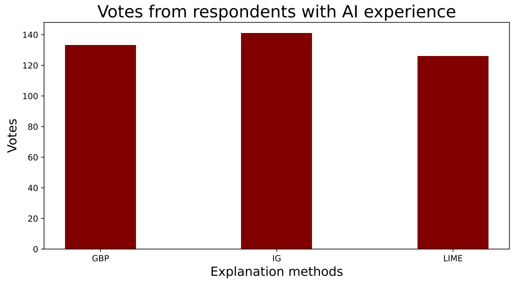


Figure 4.4: Vote counts from respondents with little or a lot of AI experience

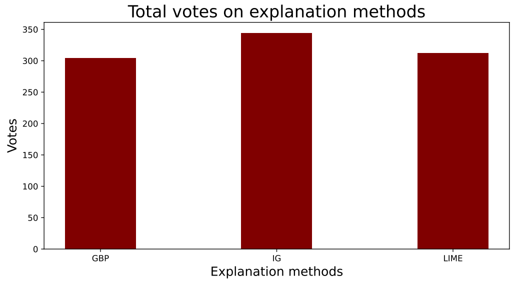


Figure 4.5: Total vote counts

5 Discussion

5.1 Summary of results

Table 5.1 shows a summary of the results of the average evaluation. The first two columns specify the model and the evaluation method. The third column is split into three and displays the highest scoring explanation methods from left to right, left being the highest scoring. The last column indicates whether the difference between the best explanation methods and the others was significant.

For the WQP model IG performed best according to the DAUC score, but not significantly. LIME performed best according to the IAUC score.

For the AP model GBP performed based according to the DAUC, LIME performed best according to the IAUC score and IG received the most votes in the user study, although the differences in votes were not significant.

The analyses done in this Bachelor thesis did not point to one favourite explanation model, but LIME performed best in two comparisons with significant differences. The fact that this method is model-agnostic and can therefore be applied to any model [Molnar, 2022] might have had an influence on this outcome. It would be interesting to compare the results of LIME with other model-agnostic methods like SHAP [Lundberg and Lee, 2017].

A notable result that can be pointed out is that the error on the averaged DAUC graphs (Figure A.4) of the AP model explanations decreases to the same low error when 100% of the pixels are deleted from the sample. This would mean that the model predicts the age of an all-black image very well, which is questionable.

5.2 Influencing factors

As discussed in the introduction, most research on explanation methods is done on classification problems, which is why most methods are designed or optimised for classification models. This research can be seen as an exploratory analysis of GBP, IG and LIME applied to regression models. It is difficult to say if the application of these methods was successful, as there are a variety of factors that could have influenced the results.

5.2.1 Data type

Figure A.4 shows that all DAUC curves and two out of three IAUC curves display unexpected results, indicating low-quality explanations. This may suggest that all three explanation methods perform poorly when applied to a regression model trained on image data.

Bennetot et al. [2021] provide a schematic overview of what explanation method to use for different data types. For explaining a model trained on tabular data they suggest to use SHAP while for image data they suggest to use Grad-CAM. SHAP is model-agnostic, just like LIME, and can be seen as the improved version of LIME. Perhaps SHAP would be able to explain regression models better than LIME. Grad-CAM is a gradient-based method, just like GBP and IG. However Grad-CAM is developed for classification models [Selvaraju et al., 2016], and is therefore not applicable to regression problems. These examples suggest ideas for future research on explanation methods for regression models that keep the underlying datatype in mind.

Model	Evaluation method	Explanation method			Significant difference?
WQP	DAUC	IG	LIME	GBP	No
	IAUC	LIME	GBP	IG	Yes
AP	DAUC	GBP	IG	LIME	Yes
	IAUC	LIME	IG	GBP	Yes
	User study	IG	LIME	GBP	No

Table 5.1: Overview results of the average evaluation, displaying the model type, the evaluation method, the explanation methods ordered by evaluation score (best scoring on the left) and an indication of significance

5.2.2 Evaluation metrics

The quality of the DAUC and IAUC evaluation metrics is a topic of discussion in the literature. Gomez et al. [2022] argue that DAUC and IAUC are not the best options for evaluating explanations of CNN models. DAUC and IAUC only look at the ranking of saliency values, and not at the actual values themselves. This is problematic because it is possible to change the saliency values while keeping the same rank. The saliency map will look vastly different, while the DAUC and IAUC scores remain the same. Gomez et al. [2022] suggest two other methods, that use two different properties of the saliency methods ignored by DAUC and IAUC: sparsity and calibration. Sparsity is a term describing the focus of the saliency maps. Saliency maps that highlight specific parts of the image are easier to interpret for humans, and could therefore be considered better explanations. Calibration would be visible in the saliency map by the luminosity of that pixel, representing the importance of that pixel on the class score. This is not taken into account in the DAUC and IAUC scores. The evaluation methods that Gomez et al. [2022] suggest are the Sparcity Metric, Deletion Correlation (DC) and Insertion Correlation (IC). It would be useful to investigate the results of these evaluations in future research.

5.2.3 Model quality

Both regression models considered in this Bachelor thesis did not achieve a high performance since the validation error was 0.55 MSE for the WQP model and $0.51R^2$ for the AP model. The model quality might have had an effect on the explanations, as according to Letzgus et al. [2021] post-hoc explanation methods assume the underlying model is the best performing model for that task. Future research could apply GP, IG and LIME to high-performing models to see if model quality had an effect of the results found here.

5.3 Future research

As discussed above, interesting factors to consider in future research are other model-agnostic explanation methods, data type, evaluation metrics and regression models with better performance. Furthermore, as was discussed in the introduction there is a debate in the literature between interpretable models and black box models. Another interesting path of future research is to extend the comparison to interpretable models as well.

This Bachelor thesis focused on local attribution-based explanation methods. It can be argued that these methods therefore produce somewhat limited explanations [Letzgus et al., 2021]. Global explanation methods might more closely approach the ground truth explanation, which is why one can argue that instead more focus should go towards global explanation methods. Guidotti et al. [2019] describe a global explanation method consisting of several local approximations (similar to the explanations generated by LIME) combined to form a global approximation of the global decision function. Guidotti et al. [2019] developed a two-step approach: the *local step* that will generate all the local explanations, and the *local-to-global* step, which will combine different local explanations based on similarity to form the global explanation. However, a situation in which a black box model is explained using a black box explanation method should be avoided. The notion of similarity used in the *local-to-global* step is a mathematical description of similarity, which is quite abstract.

Khakzar et al. [2022] observed that different attribution-based explanation methods applied to the same model can point to different salient features, which is what happened in this Bachelor thesis as well. They suggest to do additional experiments to test if provided explanations are in agreement with certain axioms. These axioms are formalisations of desirable properties that an attribution-based explanation method should have. As an example, one of the axioms is the Null-player axiom, similar to the Sensitivity Axiom as discussed in Section 3.3.2. Future research could perform the experiments suggested by Khakzar et al. [2022] on the results of the explanation methods discussed here.

Finally, other interesting directions of future research could be to compare post-hoc explanations of black box models to the results of inherently interpretable models, for example a decision tree, and to develop new explanation methods tailored to regression problems.

6 Conclusions

The question that this Bachelor thesis aimed to answer was: How do different explanation methods compare when applied to regression models?

The process of answering the research question is guided by the following subquestions:

1. How do different explanation methods of regression tasks on tabular data compare?

2. How do different explanation methods of regression tasks on image data compare?

Two regression models were implemented, a wine quality prediction (WQP) model trained on tabular data and an age prediction (AP) model trained on image data.

To these models three different local attribution-based explanation methods were applied and the results of those were evaluated and compared. The explanation methods used were Guided Backpropagation (GBP), Integrated Gradients (IG) and Local Interpretable Model-agnostic Explanations (LIME), all falling under the category of local attribution-based methods. The results of the explanation methods were evaluated using the Deletion Area Under the Curve (DAUC) and Insertion Area Under the Curve (IAUC) evaluation metrics, and an additional qualitative evaluation was done on the AP model explanations in the form of a user study.

The results of the quantitative evaluation methods were displayed for explanations of individual data samples, and averaged over the explanations of 500 data samples.

These analyses allowed the research question to be answered. In the case of the WQP model IG performed best according to the DAUC score, although not significantly, and LIME performed best according to the IAUC score. In the case of the AP model, GBP performed best according to the DAUC score and LIME performed best according to the IAUC score. Additionally, according to the user study IG performed best, however the difference between the number of votes was not significant.

Future research could go into applying and evaluating different existing explanation methods to regression problems, using different evaluation metrics, testing explanation methods on models with higher performance, global explanation methods or developing experiments to scientifically evaluate explanations. Furthermore, post-hoc explanations could be compared to inherently interpretable models and new explanation methods could be developed especially for regression models.

The results of this Bachelor thesis could help reduce the focus on classification problems and give some attention to regression problems in XAI.

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. *Gradient-Based Attribution Methods*, pages 169–191. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. 10.1007/978-3-030-28954-6_9. URL https://doi.org/10.1007/978-3-030-28954-6_9.

Christopher J. Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022. ISSN 1566-2535. https://doi.org/10.1016/j.inffus.2021.07

.015. URL https://www.sciencedirect.com/science/article/pii/S1566253521001573.

Adrien Bennetot, Ivan Donadello, Ayoub El Qadi, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Saranti, Silvia Tulli, Maria Trocan, Raja Chatila, Andreas Holzinger, Artur d'Avila Garcez, and Natalia Díaz-Rodríguez. A practical tutorial on explainable ai techniques, 2021. URL https://arxiv.org/abs/2111.14260.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. ISSN 0167-9236. https://doi.org/10.1016/j.dss.2009.05.016. URL https://www.sciencedirect.com/science/article/pii/S0167923609001377. Smart Business Networks: Concepts and Empirical Evidence.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. visited on 2022-05-03.

Krzysztof Fiok, Farzad V Farahani, Waldemar Karwowski, and Tareq Ahram. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, 19(2):133–144, 2022. 10.1177/15485129211028651. URL https://doi.org/10.1177/15485129211028651.

Maria Frentescu. Age prediction, 2020. URL https://www.kaggle.com/datasets/mariafrenti/age-prediction.

Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. *CoRR*, abs/2201.13291, 2022. URL https://arxiv.org/abs/2201.13291.

Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. The ai black box explanation problem. *Ercim News*, 116:12–13, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. 2021.

Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best, 2022. URL https://arxiv.org/abs/2203.02269.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019.

Simon Letzgus, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. Toward explainable AI for regression models. *CoRR*, abs/2112.11407, 2021. URL https://arxiv.org/abs/2112.11407.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021.

Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. URL http://arxiv.org/abs/1606.03490.

Octavio Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 2019. 10.1109/ACCESS.2019.2949286.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

CH. Raga Madhuri, G. Anuradha, and M. Vani Pujitha. House price prediction using regression techniques: A comparative study. pages 1–5, 2019.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. https://doi.org/10.1016/j.artint.2018.07.007. URL `https://www.sciencedirect.com/science/article/pii/S0004370218305988`.

Christoph Molnar. *Interpretable Machine Learning*. Creative Commons, Mountain View, California, 2022.

Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, and Lokesh Chouhan. Stock market prediction using machine learning. pages 574–576, 2018.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. *CoRR*, abs/1806.07421, 2018. URL `http://arxiv.org/abs/1806.07421`.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL `http://arxiv.org/abs/1602.04938`.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2018. URL `https://arxiv.org/abs/1811.10154`.

Anand Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *Signal Processing Magazine, IEEE*, 30: 86–94, 09 2013. 10.1109/MSP.2013.2259911.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL `http://arxiv.org/abs/1610.02391`.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013. URL `https://arxiv.org/abs/1312.6034`.

Y. Singh. Resnet20. `https://github.com/yasharvindsingh/ResNet20`, 2019.

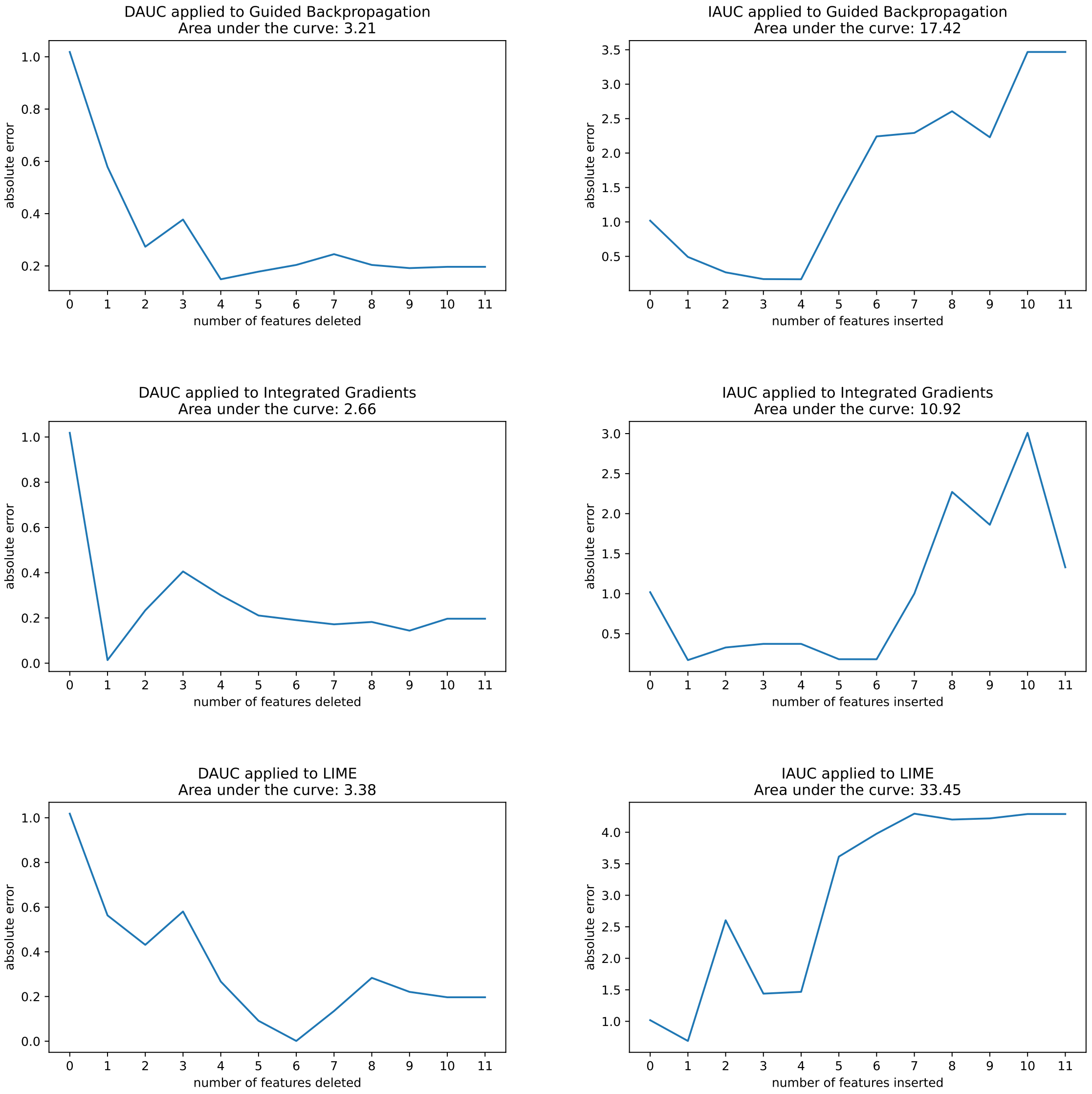
Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. Counterfactual explanations for arbitrary regression models. *CoRR*, abs/2106.15212, 2021. URL `https://arxiv.org/abs/2106.15212`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL `http://arxiv.org/abs/1703.01365`.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. URL `http://arxiv.org/abs/1711.00399`.

A Results quantitative evaluations

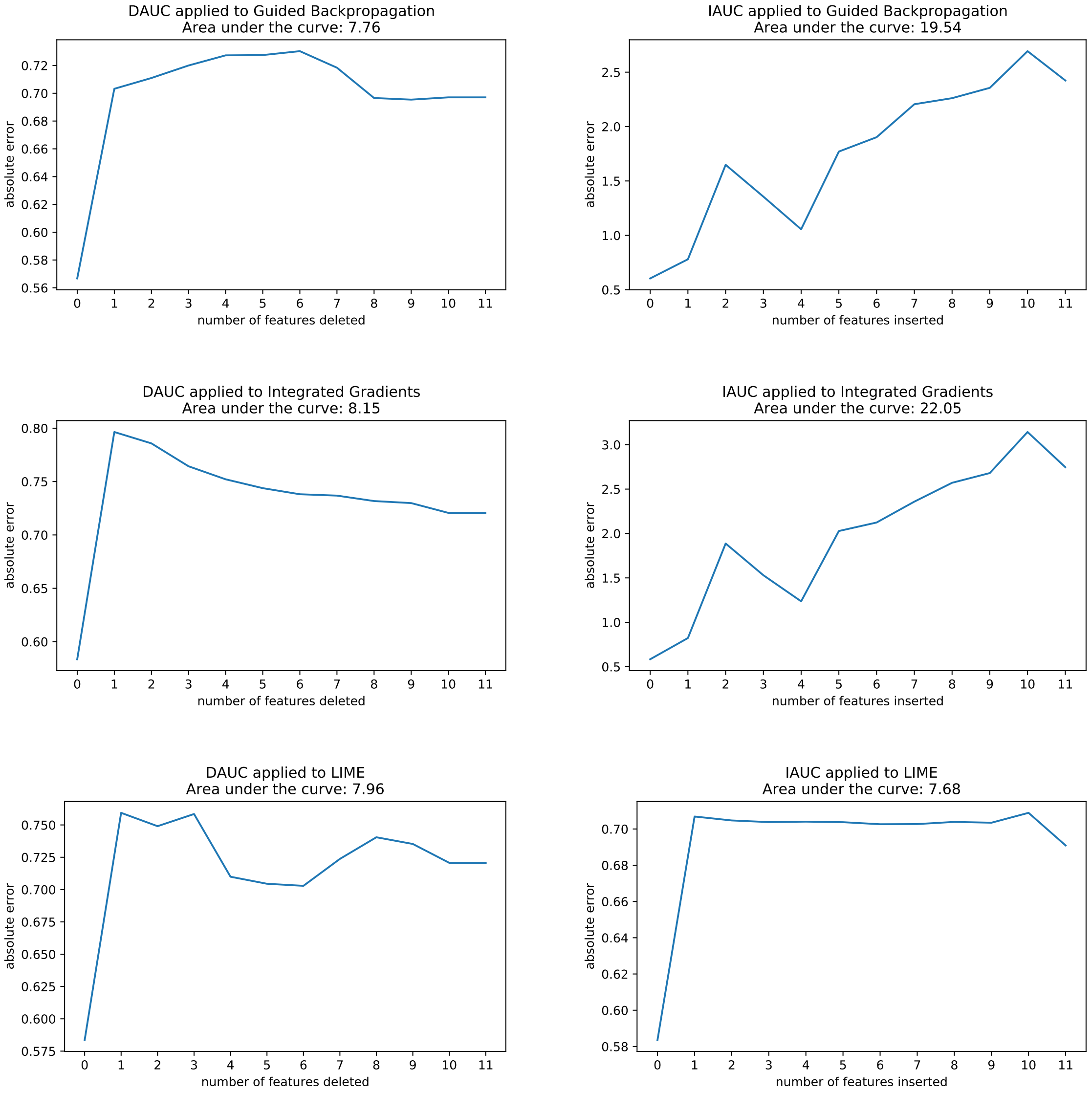
A.1 Evaluation of WQP model explanations



DAUC applied to explanations individual sample

IAUC applied to explanations individual sample

Figure A.1: DAUC and IAUC applied to three different WQP model explanations of an individual sample with true quality 3 and predicted quality 1.98

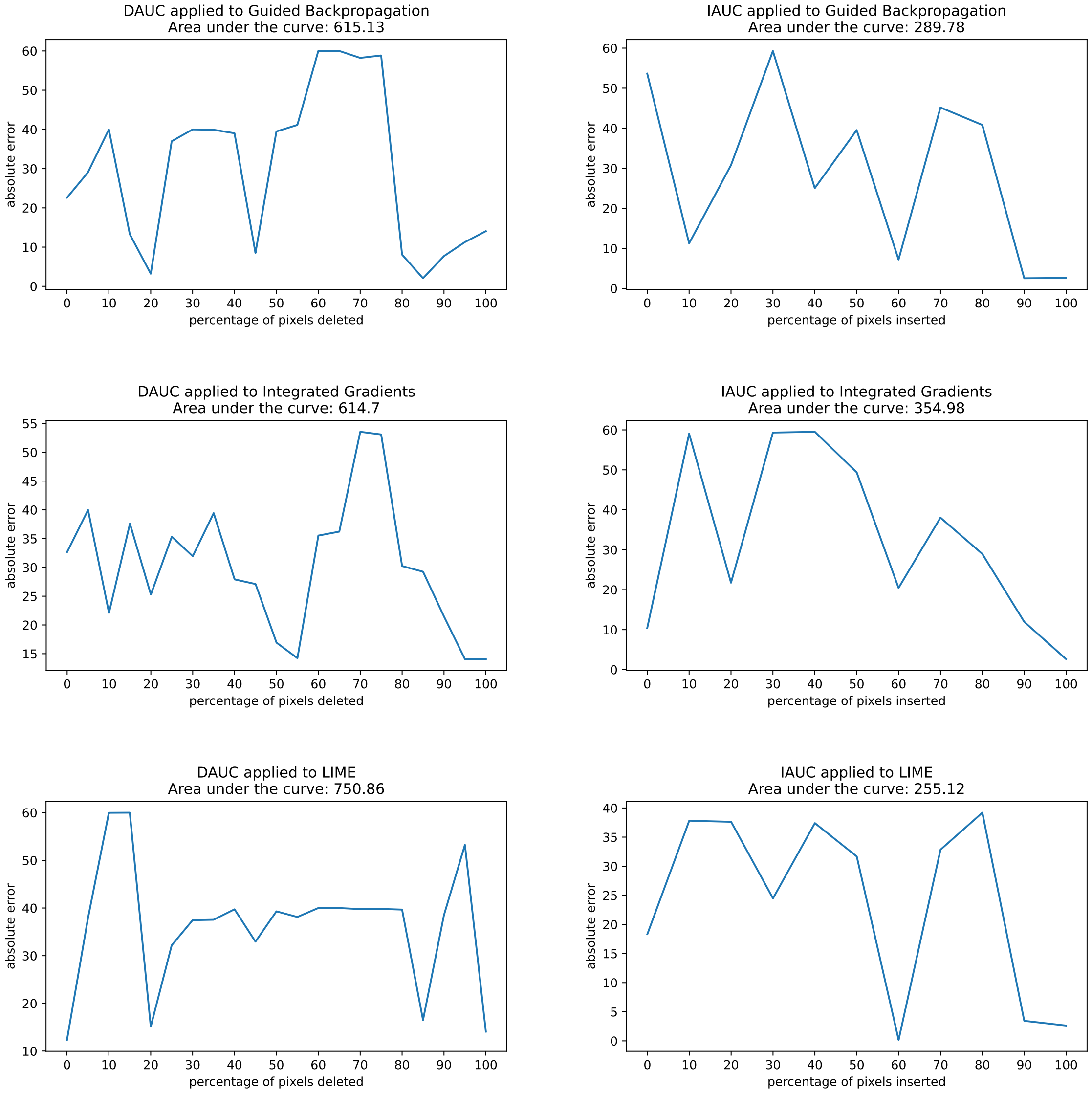


Average DAUC over 500 samples

Average IAUC over 500 samples

Figure A.2: Average curves of the DAUC and IAUC evaluation methods of GBP, IG and LIME explanation methods applied to the WQP model

A.2 Evaluation of AP model explanations


DAUC applied to Guided Backpropagation
Area under the curve: 615.13

IAUC applied to Guided Backpropagation
Area under the curve: 289.78

DAUC applied to Integrated Gradients
Area under the curve: 614.7

IAUC applied to Integrated Gradients
Area under the curve: 354.98

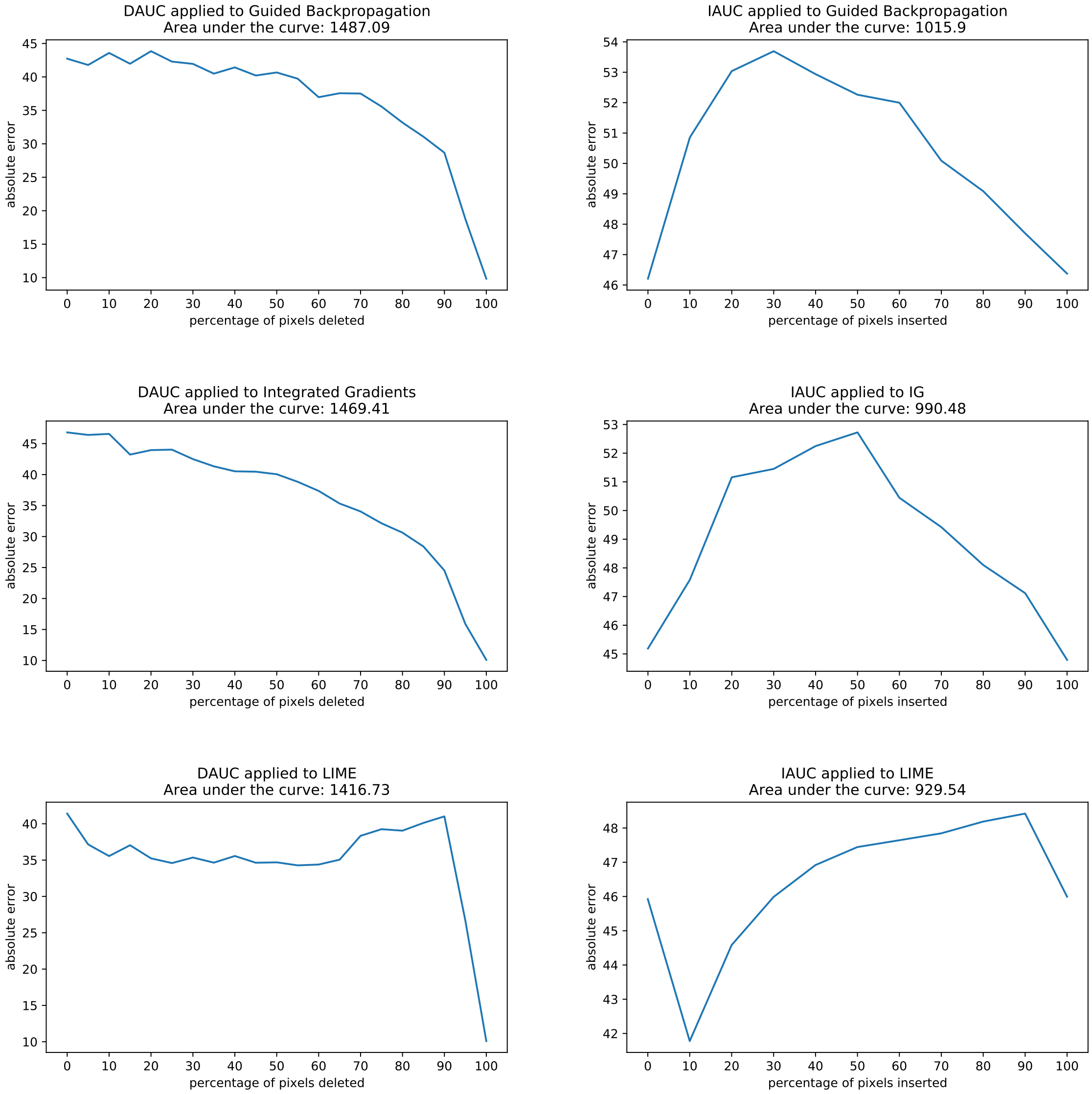
DAUC applied to LIME
Area under the curve: 750.86

IAUC applied to LIME
Area under the curve: 255.12

DAUC applied to explanations individual sample

IAUC applied to explanations individual sample

Figure A.3: DAUC and IAUC applied to three different explanations of the AP model of an individual sample with true age 40 and predicted age 42.63



Average DAUC over 500 samples **Average IAUC over 500 samples**

Figure A.4: Average curves of the DAUC and IAUC evaluation methods of GBP, IG and LIME explanation methods applied to the AP model

B Model architectures

B.1 MLP - Wine quality prediction model

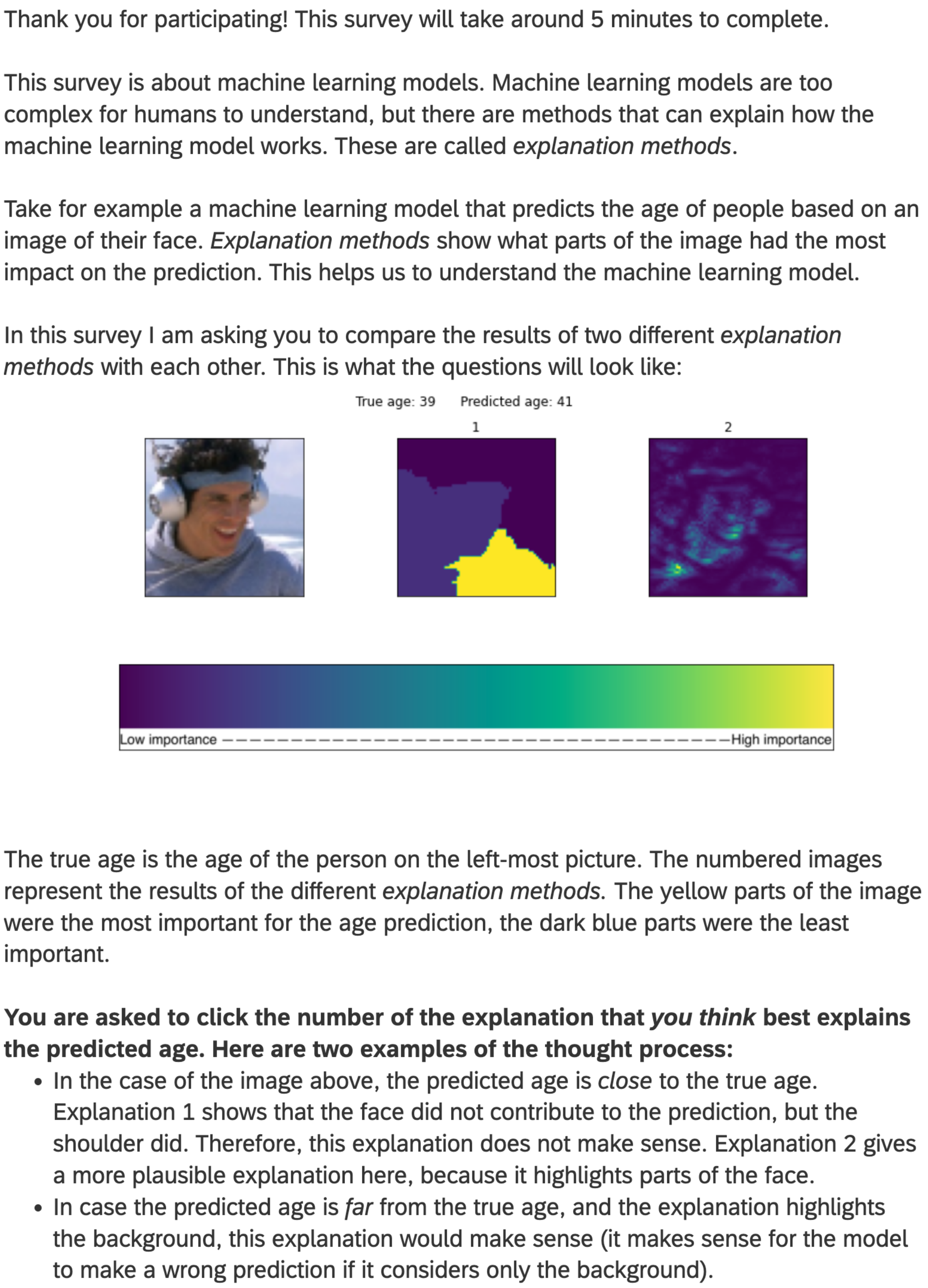
```
def custom_layer(tensor):
    return tensor * 6

inputs = tf.keras.Input(shape=(num_features,))
x = tf.keras.layers.Dense(16, activation='relu')(inputs)
x = tf.keras.layers.Dense(8, activation='relu')(x)
outputs = tf.keras.layers.Dense(1, activation='sigmoid')(x)
outputs = tf.keras.layers.Lambda(custom_layer)(outputs)
```

B.2 Resnet - Age prediction model

```
def resnet_layer(inputs, num_filters=16, kernel_size=3, strides=1, activation='relu',
    batch_normalization=True, conv_first=True):

  conv = Conv2D(num_filters, kernel_size=kernel_size, strides=strides, padding='same')

  x = inputs
  if conv_first:
      x = conv(x)
      if batch_normalization:
          x = BatchNormalization()(x)
      if activation is not None:
          x = Activation(activation)(x)
  else:
      if batch_normalization:
          x = BatchNormalization()(x)
      if activation is not None:
          x = Activation(activation)(x)
      x = conv(x)
  return x

def custom_layer(tensor):
    return tensor * 100

def resnet_v1(input_shape, depth):

    if (depth - 2) % 6 != 0:
        raise ValueError('depth should be 6n+2 (eg 20, 32, 44 in [a])')
    # Start model definition.
    num_filters = 16
    num_res_blocks = int((depth - 2) / 6) # 3

    inputs = Input(shape=input_shape)
    x = resnet_layer(inputs=inputs)

    # Instantiate the stack of residual units
    for stack in range(3):
        for res_block in range(num_res_blocks):
            strides = 1
            if stack > 0 and res_block == 0:  # first layer but not first stack
                strides = 2  # downsample
            y = resnet_layer(inputs=x,num_filters=num_filters,strides=strides)
            y = resnet_layer(inputs=y,num_filters=num_filters,activation=None)
            if stack > 0 and res_block == 0:  # first layer but not first stack
                # linear projection residual shortcut connection to match
                # changed dims
```

```python
                x = resnet_layer(inputs=x,num_filters=num_filters,kernel_size=1,strides=strides,activation=None,batch_normalization=False)
            x = keras.layers.add([x, y])
            x = Activation('relu')(x)
            x = Dropout(rate=0.25)(x)
        num_filters *= 2

    # v1 does not use BN after last shortcut connection-ReLU
    x = AveragePooling2D(pool_size=8)(x)
    y = Flatten()(x)
    outputs = Dense(1, activation='sigmoid')(y)
    outputs = keras.layers.Lambda(custom_layer, name="lambda_layer")(outputs)

    # Instantiate model.
    model = Model(inputs=inputs, outputs=outputs)
    return model
```

C Introductory text user study



Thank you for participating! This survey will take around 5 minutes to complete.

This survey is about machine learning models. Machine learning models are too complex for humans to understand, but there are methods that can explain how the machine learning model works. These are called *explanation methods*.

Take for example a machine learning model that predicts the age of people based on an image of their face. *Explanation methods* show what parts of the image had the most impact on the prediction. This helps us to understand the machine learning model.

In this survey I am asking you to compare the results of two different *explanation methods* with each other. This is what the questions will look like:

The true age is the age of the person on the left-most picture. The numbered images represent the results of the different *explanation methods*. The yellow parts of the image were the most important for the age prediction, the dark blue parts were the least important.

You are asked to click the number of the explanation that *you think* best explains the predicted age. Here are two examples of the thought process:

- In the case of the image above, the predicted age is *close* to the true age. Explanation 1 shows that the face did not contribute to the prediction, but the shoulder did. Therefore, this explanation does not make sense. Explanation 2 gives a more plausible explanation here, because it highlights parts of the face.
- In case the predicted age is *far* from the true age, and the explanation highlights the background, this explanation would make sense (it makes sense for the model to make a wrong prediction if it considers only the background).