Question 1

Introduction to Survey
Welcome, and thank you for taking the time to participate in this survey. This study compares how various explainable artificial intelligence (XAI) methodologies and techniques affect people's behaviour and cognition. This survey aims to learn more about how the actual user conceives explanations. By explanations, we mean techniques that convey information about how machine learning models arrive at their conclusions to humans.
Further information about machine learning models will be provided on the next page. We'll then present you with two distinct explanation methods and pose a task for each method, followed by some general questions to obtain your opinion.

The information we collect complies with Art. 13 of the European Union's General Data Protection Regulation (EU GDPR Reg. n. 2016/679). This study is meant for research purposes only and is based on non-personal or anonymous data which is provided during your voluntary participation. The management of your data will take place in complete accordance with the principles of confidentiality, allowing you to make an informed decision about sharing it.

It is expected to take you about 5-10 minutes to complete the survey. Should you have any questions regarding this study, please address them to Muhammad Suffian at suffiankhursheed(at)gmail(dot)com, the data manager of the study.

Giving consent to the conditions indicated below is mandatory to continue. You can proceed by clicking on the below text. By agreeing to take part in this study, you confirm that:

○ You have reached the age of maturity. You acknowledge that your participation is completely voluntary. You are aware that you may terminate taking the survey at any time. You acknowledge that your anonymous responses may be used for research purposes in accordance with General Data Protection Regulation.

The only thing we are interested in is your opinion; there are no right or wrong responses!

How familiar are you with Machine Learning models?

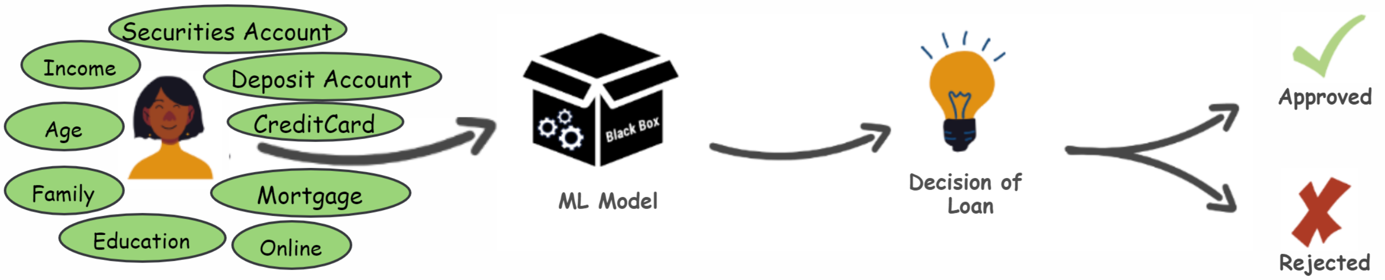
○ I'm not really familiar with machine learning models

○ I have knowledge of machine learning models from my studies and/or employment

○ I am expert of machine learning models

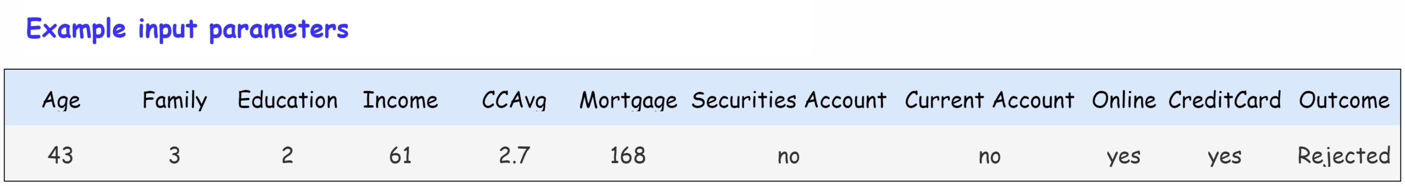
Machine Learning

Nowadays, decisions directly impacting people's lives are being made using machine learning models in various real-world applications. These models frequently function as a "black box", rendering the decision opaque to the people who may be impacted. However, people have a right to an explanation under the General Data Protection Regulation. Throughout this study, we will look at the following scenario: An ML model needs to decide if a loan applicant will likely get approved or denied based on the input values for different parameters required to fill the loan application. Note: The image is redesigned according to our case study. The theme of the survey and the image's original source is adopted from GitHub.



Let's build Understanding about ML and different Methods of Explainable AI (XAI)

Machine Learning (ML) models are frequently opaque to humans, it is impossible to determine how and why a certain decision was made. Since it is more challenging to trust systems, this is a severe flaw in employing ML models. By outlining the rationale behind their decisions, ML models can become more persuasive, transparent, intelligible, and capable of influencing human cognition and behaviour. In this survey, we will present you with two different strategies/methods which provide an explanation to reason the black-box decisions.



Example input parameters

Age	Family	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard	Outcome
43	3	2	61	2.7	168	no	no	yes	yes	Rejected

In our case, ML model concludes that the loan applicant is likely to be "**Rejected**". The loan applicant has the right to an explanation of why the model came to this outcome based on the given attributes.

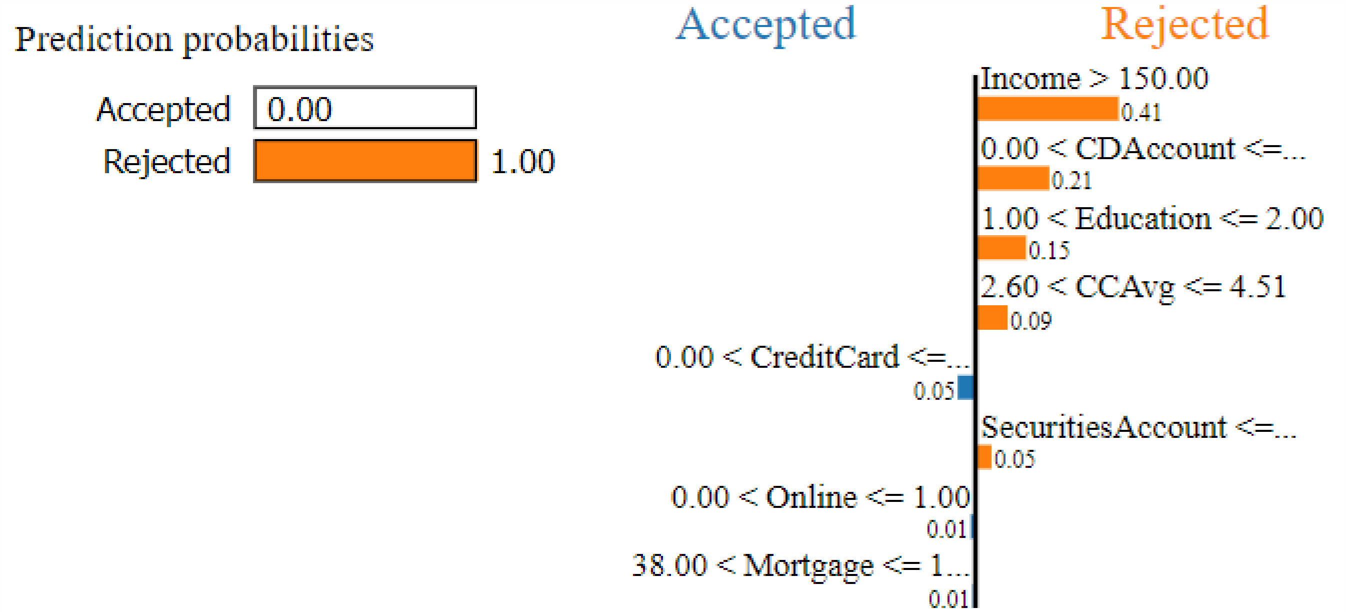
The following two methods are subjected to provide

explanation why the ML model concluded to Reject the loan.

Method-A: Feature Attribution

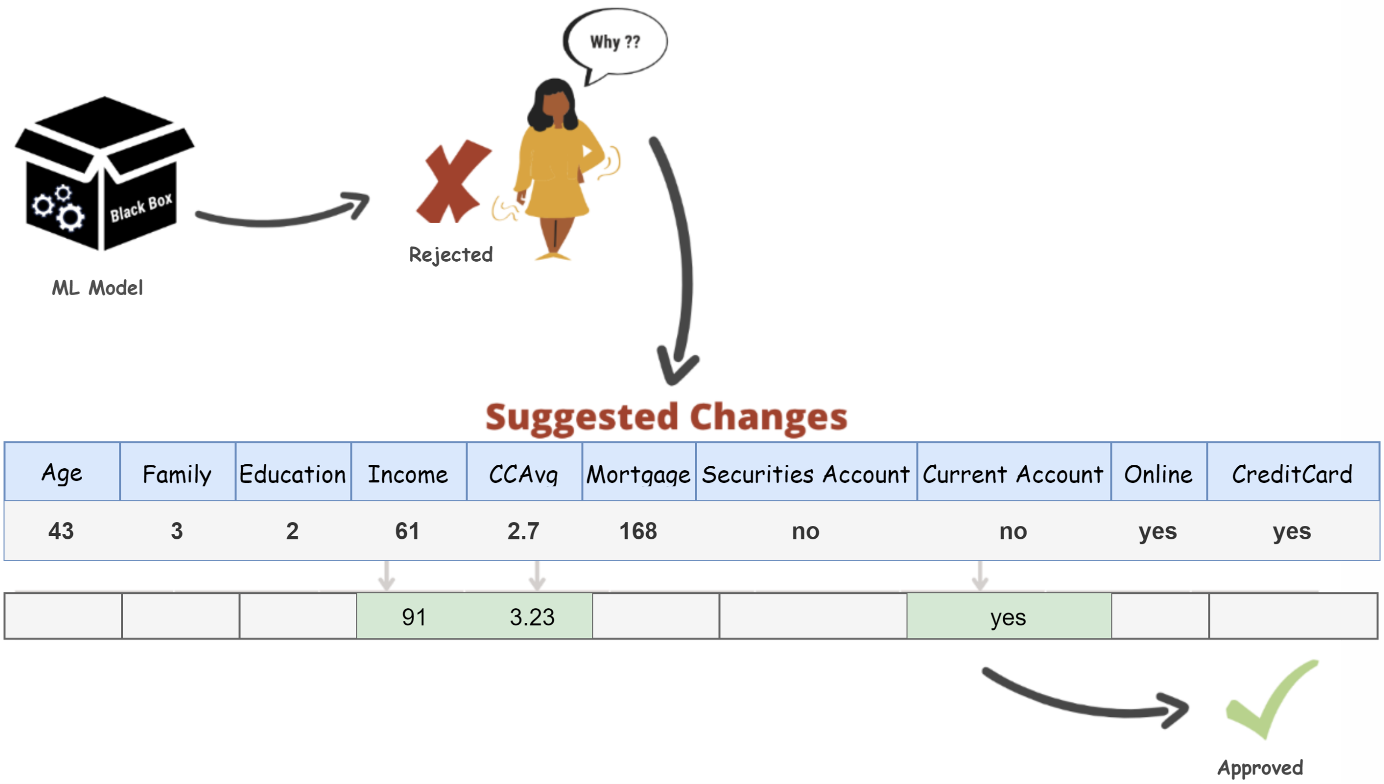
This method is one approach to provide such an explanation. It takes model inputs and produce an importance score for each feature (input parameter) based on how much that feature contributed to the model's output.

Local feature attribution is a feature attribution that emphasizes the significance of a feature and its impact on a trained model's output for a particular input.



Local feature attribution (feature scores are computed with LIME). The feature on the left (blue bars), represent their importance towards loan approval, and features on the right (yellow bars), represent their importance towards loan rejection. The prediction probabilities refer to the confidence of the underlying ML model of LIME on the given data. For example, Income value in user input was low, however, it was the most important feature (longer bar) which played a role in the prediction (Rejected). In other words, if the values of longer bar features were raised as per suggestion above the bars, it could alter the prediction. In general, Method-A informs about the features which are most relevant and play an important role in the prediction.

Method-B: Counterfactual Explanations
Counterfactual explanations (CE) are one another method for offering such an explanation. In order to provide an alternative (preferred) result, it explains by highlighting the features (which and to what extent) that must be altered. Consider the following example:
A loan applicant was "Rejected" and she is curious about how the ML model arrived at this conclusion. A counterfactual explanation outlines the features that are necessary to get the desired result, likely to be "Approved".



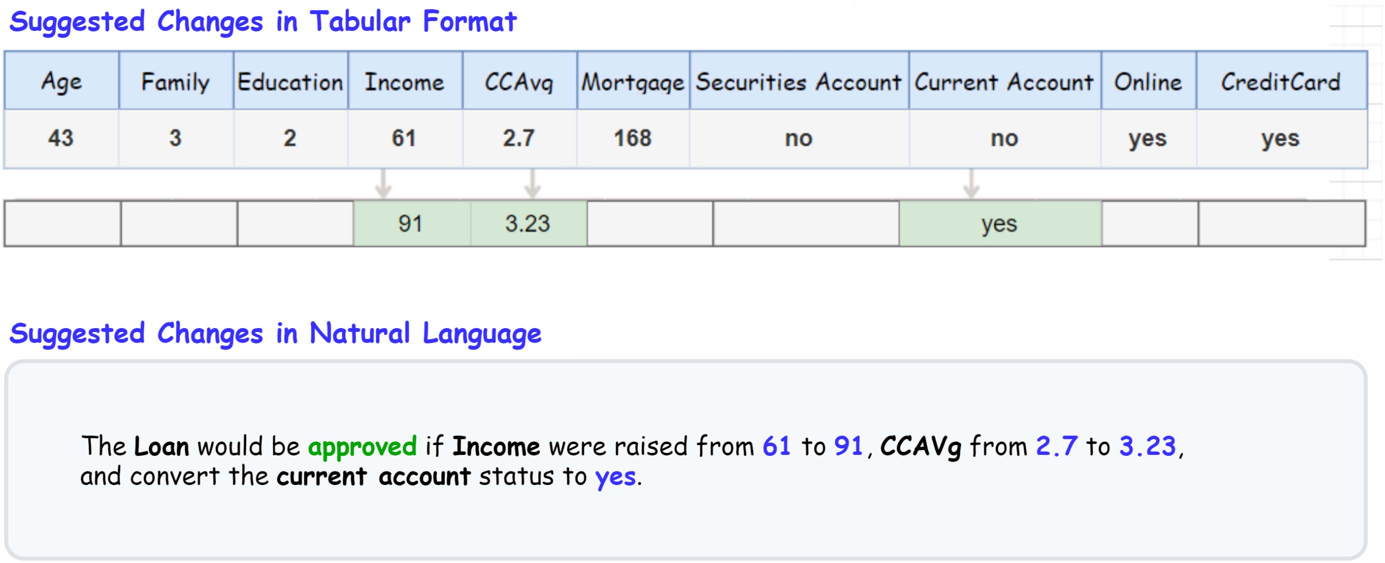
Suggested Changes

Age	Family	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard
43	3	2	61	2.7	168	no	no	yes	yes
			91	3.23			yes		

Method-B has highlighted the features which must be updated to suggested changes to get the desired outcome (Loan Approved).

Counterfactual Explanations (tabular + textual)

This is an addition to method-B, it presents textual explanation along with tabular explanation. It is human-friendly as it provides the natural language explanation as well.



Suggested Changes in Tabular Format

Age	Family	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard
43	3	2	61	2.7	168	no	no	yes	yes
			91	3.23			yes		

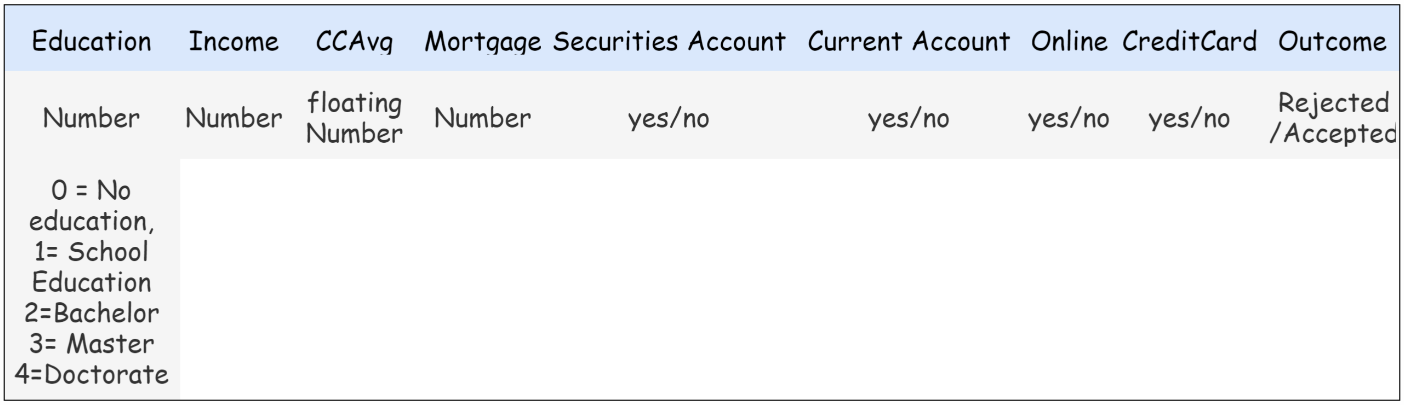
Suggested Changes in Natural Language

The **Loan** would be approved if **Income** were raised from 61 to 91, ***CCAVg*** from 2.7 to 3.23, and convert the ***current account*** status to yes.

The textual explanations in natural language are easy to understand for all types of users to take actions accordingly to obtain the desired outcome.

How the attributes (features) can change

We consider only the changeable features (Age and Family are not considered due to their protected nature in real-world applications).

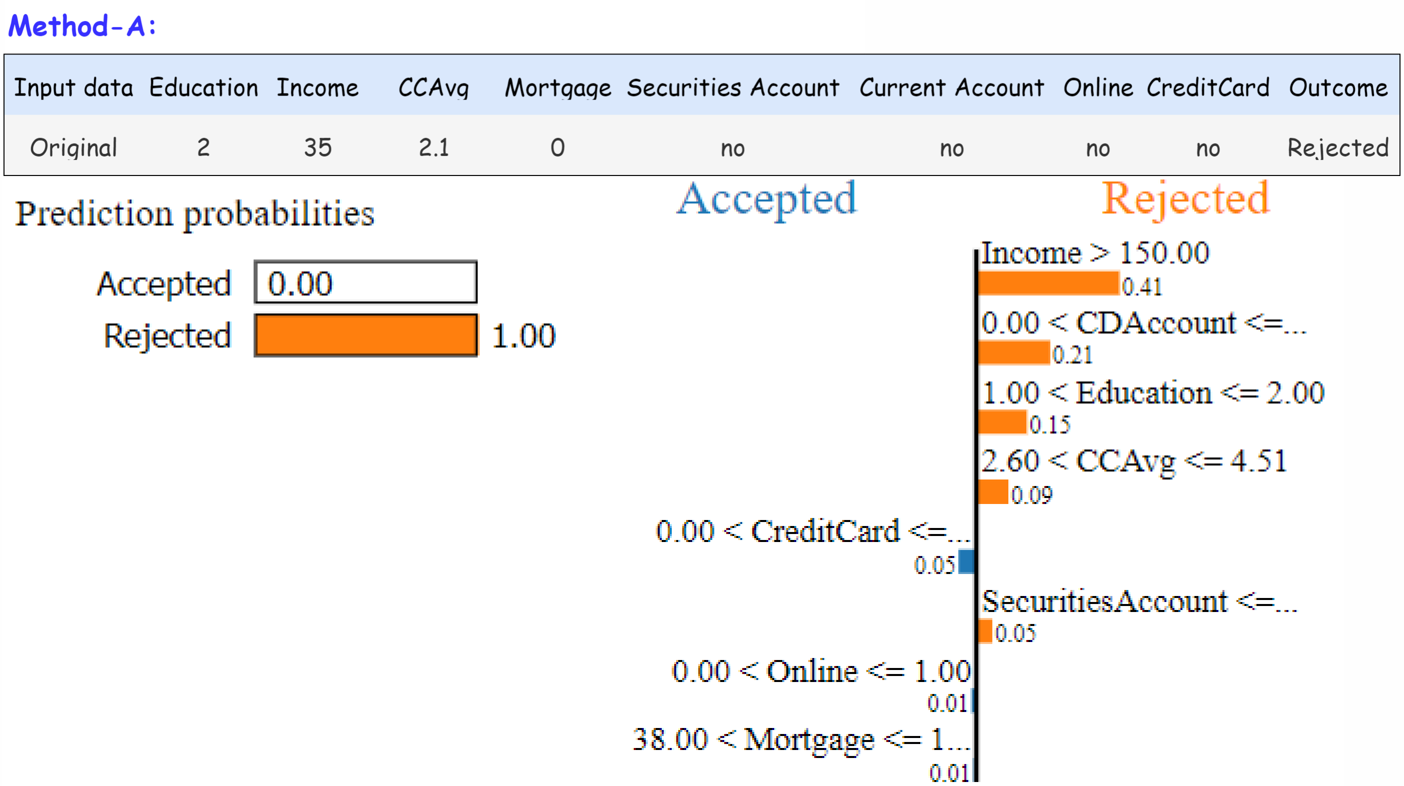


Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard	Outcome
Number	Number	floating Number	Number	yes/no	yes/no	yes/no	yes/no	Rejected /Accepted
0 = No education, 1= School Education 2=Bachelor 3= Master 4=Doctorate								

Decision-Making based on Method-A

Here, you have to utilize your knowledge and understanding based on Method-A to decide whether the presented explanation and your action accordingly could lead to the model outcome as "Accepted".
The output of Method-A (local feature attribution method i.e., *LIME*) is a list of explanations reflecting the contribution

of each feature to the prediction of a data sample. It also allows us to determine which feature changes will most impact the prediction. The prediction probabilities refer to the confidence of the underlying ML model on the given data.



Method-A:

Input data	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard	Outcome
Original	2	35	2.1	0	no	no	no	no	Rejected

What do you think about the Loan outcome? select the choice regarding the Loan decision in your opinion.

How well does this method explain what needs to be changed to get the "Accepted" model outcome?

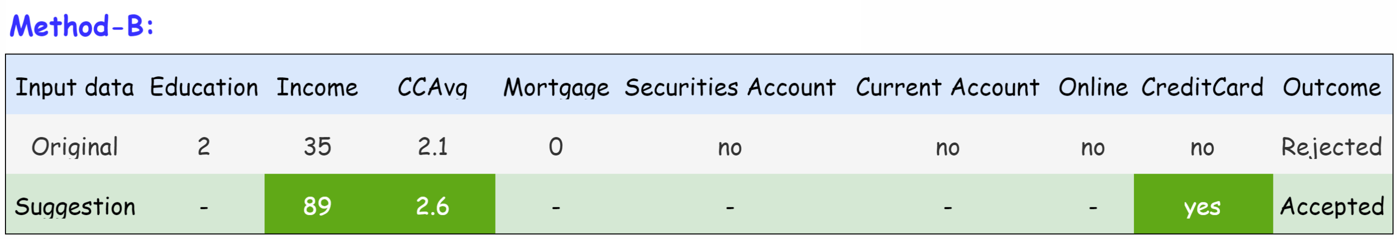
○ Not at all ○ Very Little ○ Somewhat ○ Very Well

Could you realistically act upon the suggestions to change the model outcome to "Accepted"?

Not at all	Very Little	Somewhat	Fully
○	○	○	○

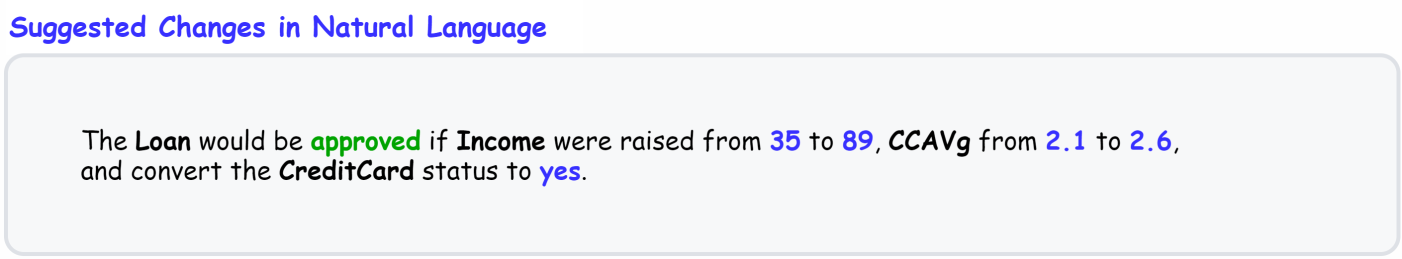
Decision-Making based on Method-B

Method B provides exact values for each input parameter, making the action more understandable. Also, the suggested values which only need to change in the actual input are highlighted along with a textual explanation. Again, you have to utilize your knowledge and understanding based on Method B to make your opinion regarding the suggested changes, which could result in a loan being Accepted.



Method-B:

Input data	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard	Outcome
Original	2	35	2.1	0	no	no	no	no	Rejected
Suggestion	-	89	2.6	-	-	-	-	yes	Accepted



Suggested Changes in Natural Language

The **Loan** would be approved if **Income** were raised from 35 to 89, **CCAVg** from 2.1 to 2.6, and convert the **CreditCard** status to yes.

Select the choice regarding the loan decision in your opinion.

How well does this method explain what needs to be changed to get the model outcome as loan "Accepted"?

Not at all	Very Little	Somewhat	Very Well
○	○	○	○

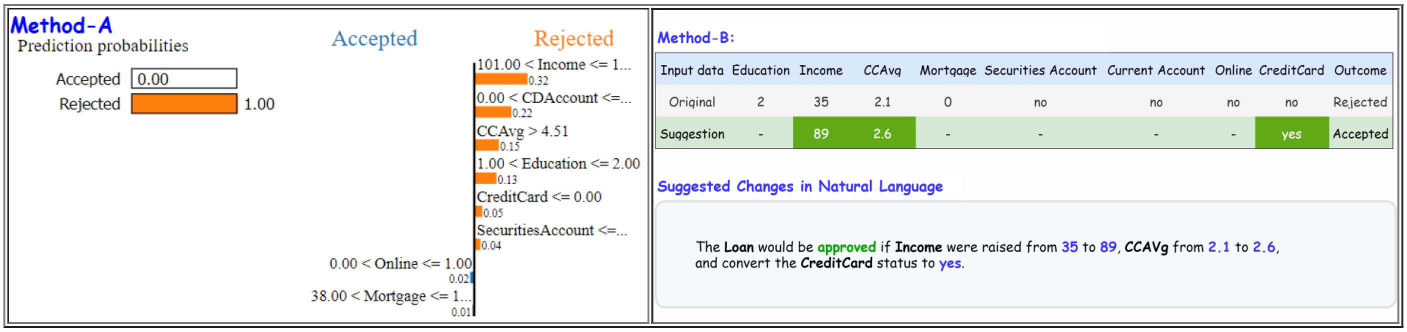
Could you realistically act upon the suggestions to change the model outcome to "Accepted"?

Not at all	Very Little	Somewhat	Fully
○	○	○	○

comparative-questions

Your Opinion

You are asked some questions below regarding different evaluation criteria for both methods. Mark the checkbox which you consider more appropriate by recalling the previously presented Methods of explanation.



Method-A

Method-B:

Input data	Education	Income	CCAvg	Mortgage	Securities Account	Current Account	Online	CreditCard	Outcome
Original	2	35	2.1	0	no	no	no	no	Rejected
Suggestion	-	89	2.6	-	-	-	-	yes	Accepted

Suggested Changes in Natural Language

The **Loan** would be approved if **Income** were raised from 35 to 89, **CCAVg** from 2.1 to 2.6, and convert the **CreditCard** status to yes.

Checkmark the Method you think fulfill the question requirements.

	Method-A	Method-B	Both Method-A and Method-B
From different explanation methods, which method you feel was better understandable?	○	○	○
Which method presented sufficient details?	○	○	○
Which method you consider as useful to decision-making goals?	○	○	○
Which method you feel as trustworthy?	○	○	○
Which method has changed your behaviour to take actions appropriately?	○	○	○

demographic

Demographic Information

Would you tell us a bit about yourself?

Privacy statement: The demographic information collected at this stage is entirely anonymous. No other personal data were collected at any moment during the survey. The user cannot be identified using the data from this screen. It is solely used to describe the sample of survey participants.
* - mandatory question.

How old are you?

What is your gender?

What is your education level?

What is your region of residence?

What is your English language proficiency level?

Powered by Qualtrics