A   PRELIMINARY STUDY - TASK DESIGN

A.1   Description of the task

A bank has implemented a new loan application system where potential customers apply for a loan online and then the company assesses the eligibility of the customer for the loan.

Given the latest technological advances and in an effort to make loan decisions in a timely manner, the loan application process is now fully automated. An artificial intelligence system receives the online requests and evaluates each case. An email is sent to the applicants with the final verdict.

Kim, a potential customer, is looking for funding opportunities to buy a house and has thus decided to apply for a home loan through the bank's online platform. As part of the home loan application process, the bank has requested the following information:

- Credit score
- Date of birth
- Employment status
- Education
- Gender
- Loan amount requested
- Loan amount term (months)
- Loan purpose
- Marital status
- Number of dependents
- Total annual income

The artificial intelligence system uses this information for making the loan decision.

A few hours after sending the requested information, Kim has received an email with the final decision: **the loan has been rejected**. In the email received by Kim, an explanation of how the decision-making system has reached the conclusion is included.

Figure 1: Description of the presented scenario.

A.2 Tested explanations

Factor importance-based explanation:

The email includes the importance that each piece of information provided by Kim had in the final decision. Factors are listed from the most important to the least important factor based on the bank's criteria. This means that, for example, credit score is a factor that the bank considers more important than the loan amount requested.

Credit Score> Loan amount requested> Total annual income> Loan purpose> Employment status> Loan amount term (months)> Date of birth> Marital status> Number of dependents> Education> Gender

Figure 2: Factor importance-based explanation.

Input influence-based explanation:

The email includes the importance that each piece of information provided by Kim had in the final decision. Factors are listed from the most important to the least important factor based on the bank's criteria. The magnitude of the contribution of each piece of information (negative (-) means that it contributed to the rejection decision) is added between brackets:

Credit Score (-0.15)> Loan amount requested (-0.12)> Total annual income (-0.09)> Loan purpose (+0.02)> Employment status (+0.02)> Loan amount term (months) (-0.03)> Date of birth (+0.03)> Marital status (+0.01)> Number of dependents (-0.07)> Education (+0.02)> Gender (+0.04)

Figure 3: Input influence-based explanation.

Case-based explanation:

"This decision was based on thousands of similar cases from the past. For example, a similar case to yours is a previous customer, Taylor. Taylor also applied for a home loan with an annual income and credit score similar to yours. Taylor was trying to get a house as expensive as yours and ended up not being able to pay it back."

Figure 4: Case-based explanation.

Counterfactual explanation:

The individual would have been granted the loan if one of the following scenarios had been true:

- The loan amount requested had been 5% lower
- The total annual income of the individual had been 10% higher
- The credit score of the individual had been "Very Good"

Figure 5: Counterfactual explanation.

Input influence-based AND counterfactual explanation:

The email includes the importance that each piece of information provided by Kim had in the final decision. Factors are listed from the most important to the least important factor based on the bank's criteria. The magnitude of the contribution of each piece of information (negative (-) means that it contributed to the rejection decision) is added between brackets:

Credit Score (-0.15)> Loan amount requested (-0.12)> Total annual income (-0.09)> Loan purpose (+0.02)> Employment status (+0.02)> Loan amount term (months) (-0.03)> Date of birth (+0.03)> Marital status (+0.01)> Number of dependents (-0.07)> Education (+0.02)> Gender (+0.04)
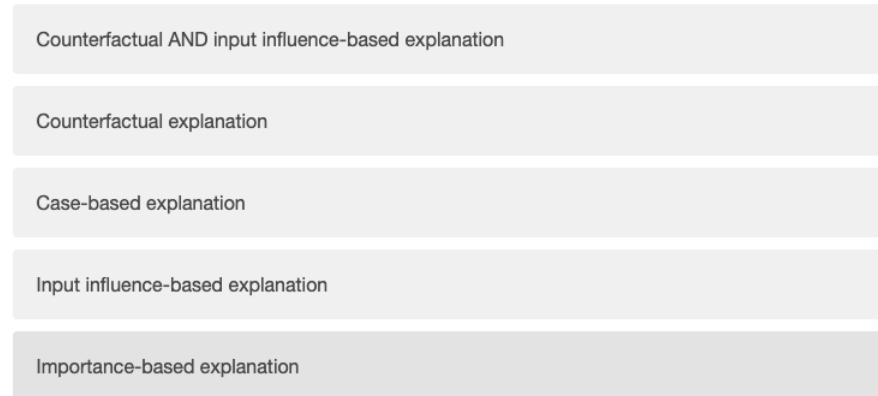
The email also includes information about scenarios where the individual would have been granted the loan. Kim would have been granted a loan if one of the following scenarios had been true:

- The loan amount requested had been 5% lower
- The total annual income of the individual had been 10% higher
- The credit score of the individual had been "Very Good"

Figure 6: Combination of input influence-based and counterfactual explanation.

A.3   Measurements

Select **two explanations that would make Kim best understand how the decision was made.**

Counterfactual AND input influence-based explanation

Counterfactual explanation

Case-based explanation

Input influence-based explanation

Importance-based explanation

Figure 7: Selecting understandable explanations.

Select **two explanations that give Kim the most explicitly actionable information.**

Case-based explanation

Importance-based explanation

Counterfactual AND input influence-based explanation

Input influence-based explanation

Counterfactual explanation

Figure 8: Selecting actionable explanations.

Select **two explanations that thanks to which Kim would best know what information to use to contest the decision** in case of disagreement.

Importance-based explanation

Input influence-based explanation

Case-based explanation

Counterfactual explanation

Counterfactual AND input influence-based explanation

Figure 9: Selecting explanations that support contestability.

Overall select the **best explanation.**

Case-based explanation

Counterfactual AND input influence-based explanation

Counterfactual explanation

Input influence-based explanation

Importance-based explanation

Figure 10: Selecting the preferred explanation.

If Kim was able to start an appeal process, what aspect of the decision-making should Kim contest? Why?

How does the best explanation help identify the aspect that Kim should contest?

Figure 11: Open-ended questions.