ROBERT GORDON UNIVERSITY ABERDEEN

Towards Feasible Counterfactual Explanations A Taxonomy Guided Template-based NLG Method

Pedram Salimi, Nirmalie Wiratunga, David Corsar, Anjana Wijekoon

Abstract Counterfactual Explanations outline minimal feature value changes to alter outcomes, but they typically overlook the feasibility of such changes. In response, this research introduces the n-XAI^{T} approach, using the Feature Actionability Taxonomy (FAT) to categorise features by mutability and guide the generation of natural explanations. The taxonomy, informed by user studies across six Fair AI datasets, distinguishes between mutable and immutable feature-based sentence-level templates to generate natural explanations, with special consideration for sensitive immutable features. The study evaluates n-XAI^{T} on 6 datasets across three domains: health, finance, and education. Our findings from a user study show that the taxonomy-guided n-XAI^{T} provides more articulate, feasible, and acceptable explanations compared to baseline methods. It also confirmed that while n-XAI^{T} is able to present actions related to sensitive features appropriately, XAI platforms should use personalised and interactive methods tailored to individual user circumstances.

Understanding How to Compose Counterfactuals

Objective: To understand how counterfactuals are authored and identify reusable linguistic constructs with non-domain-expert users.

Methodology:

- **Counterfactual scenarios:** Using DICE, counterfactuals with 4-5 actionable feature changes for seven rejected loan applications were chosen.
- **Presentation formats:** Counterfactuals were presented in 4 formats: Natural-XAI (n-XAI^{B1} & n-XAI^{B2}), zero-centred visual chart, and tabular format.
- **User Task:** Participants received a query and corresponding counterfactual in alternative formats depending on their cohort. They were asked to write an explanation based on given information and then rank the given explanations
- **Findings:**
  - Natural-XAI improved user created natural explanation quality.
  - Thematic analysis led to Natural-XAI templates groupings and concepts for FAT

Textual Explanation using SHAP Importance $n-XAI^{B1}$

Your loan application would be successful if you increase loan amount to 14500.0, increase average current balance to 20500.00, decrease recoveries to 0.0, and change loan grade to B in the exact order of priorities.

Textual Explanation using action grouping $n-XAI^{B2}$

Your loan application would be successful if you increase these features: loan amount and average current balance to the corresponding values of 14500.0 and 20500.00, decrease recoveries to 0.0, and change loan grade to B.

Graph Explanation

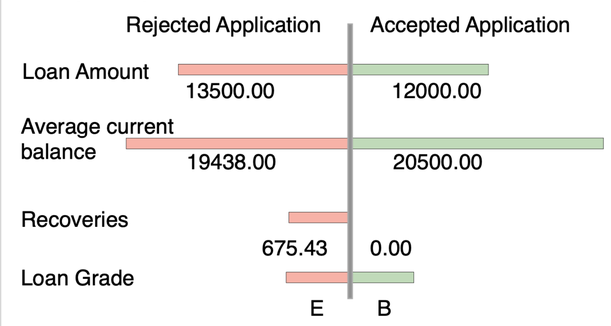
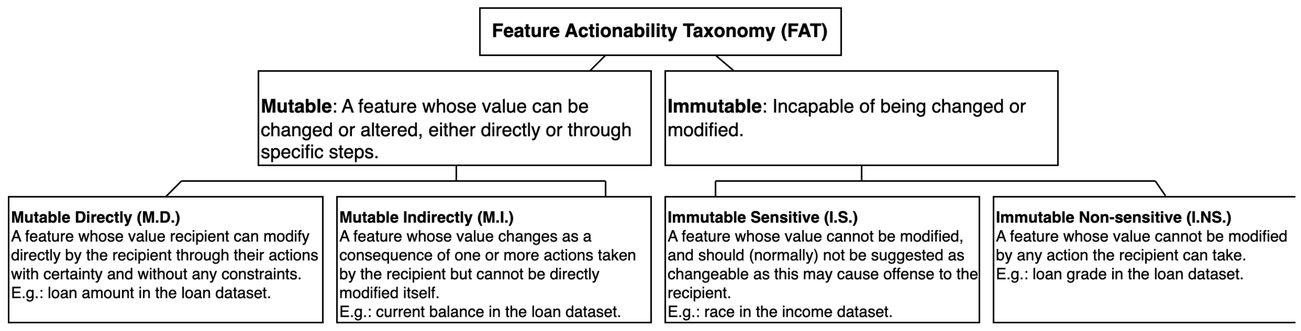


Table Explanation

	Query	Counterfactual
Loan Amount	13500.00	12000.00
Average Current Balance	19438.00	20500.00
Recoveries	675.43	0.00
Loan Grade	E	B

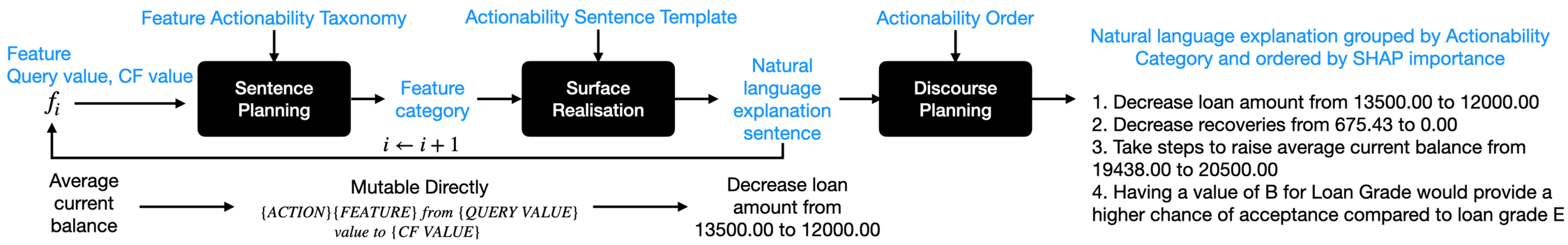
Counterfactual Natural-XAI Method



Template Variables with Synonym Examples

```
VERB={Take|Initiate|Undertake|Pursue|Negotiate}
OBJECT={steps|measures|actions|}
ACTION={Pos: (increase|improve|raise)| Neg:(decrease|reduce)}
COMPARATIVE={Pos: (increase|higher|better) | Neg: (decrease|lower|worse)}
OUTCOME={undesired:(rejected|fail) | desired:(accepted|pass)}
FEATURE= feature name in dataset, QUERY_VALUE= feature value from query, CF_VALUE=
 feature value from counterfactual, POSSESSIVE={Your}
```

Actionability Category	*Feature Sentence Template*
Mutable Directly	1. {ACTION} {FEATURE} from {QUERY_VALUE} value to {CF_VALUE} 2. {ACTION} {FEATURE} to {CF_VALUE}
Mutable Indirectly	1.{VERB} {OBJECT} to {ACTION} {FEATURE} from {QUERY_VALUE} to {CF_VALUE} 2.{VERB} {OBJECT} to {ACTION} {FEATURE} to {CF_VALUE}
Immutable Non-sensitive	Having a value of {CF_VALUE} for {FEATURE} would provide a {COMPARATIVE} chance of {DESIRED_OUTCOME} compared to a value of {QUERY_VALUE}
Immutable Sensitive	{POSSESSIVE} {FEATURE} has contributed to {OUTCOME}



Evaluating Actionability in Natural-XAI

Objective: To understand and evaluate the effectiveness of n-XAI^{T} in generating actionable counterfactual explanations across different domains.

Methodology:

- **Domains:** User study was conducted on the Prolific platform across three domains: health, education, finance.
- **User Task:** Participants received either n-XAI^{T} or baseline counterfactual explanations, not both. For each type, they evaluated two scenarios per domain based on Articulation, Acceptability, Feasibility, and Sensitivity. Participants rated responses on a 5-point Likert scale and provided their rationale.
- **Quantitative Findings:** n-XAI^{T} outperformed baseline with significant improvements in articulation (health), acceptability (health and education).
- **Qualitative Findings:** Participants particularly valued use of factual style explanations for sensitive features specifically in Health and Education domains. Users desired more detailed strategies for proposed changes in the Finance.
- Feedback suggests need for personalised and interactive explanation experiences.

Conclusion

The n-XAI^{T} approach, centred on actionability knowledge, significantly enhanced metrics such as articulation, acceptability, feasibility, and sensitivity for counterfactual explanations.

- The four-category Feature Actionability Taxonomy (FAT) guides explanation presentation, and is compatible with existing counterfactual explainers.
- Our study distinguished content-based and structure-based themes in human-composed counterfactual explanations, leading to the creation of a sentence-level template-based NLG method.
- FAT combined with feature attribution weights, informs the selection of NLG templates compatible with explainers like DICE, NICE, and DisCERN.
- FAT and sentence-level templates are open-sourced, for community input and enhancement (refer to QR code).
- Future research will prioritise the need:
  - for personalised and interactive explanation experiences
  - to identify what additional knowledge is necessary to enhance understanding of actionability for mutable features.

Supplementary Materials

Supported by chist-era EPSRC iSee

Sponsor Website
https://isee4xai.com/

Contacts
p.salimi@rgu.ac.uk
n.wiratunga@rgu.ac.uk
d.corsar1@rgu.ac.uk
a.wijekoon1@rgu.ac.uk