



ACTUARIAL SOCIETY
OF SOUTH AFRICA

Advancing Actuarial Excellence

IDENTIFYING DECISION BOUNDARIES TO EXPLAIN BLACK-BOX MACHINE LEARNING PREDICTIONS

Adriaan Rowan, PhD

June 2026

Providing an Explanation Is Not Optional

Legal, ethical and commercial drivers



The Law Requires It

South Africa's POPI Act requires automated decision models to provide sufficient information about underlying logic to the individuals affected.



Clients Demand It

Insurance, credit and health decisions affect lives.

Customers want clear, actionable explanations, not just a result.



Profitability depends on it

Increase sales, by increasing the number of clients who qualify.

Reduce defaults, by finding and addressing each client's risk factors .

Explainability methods in South African Financial Services

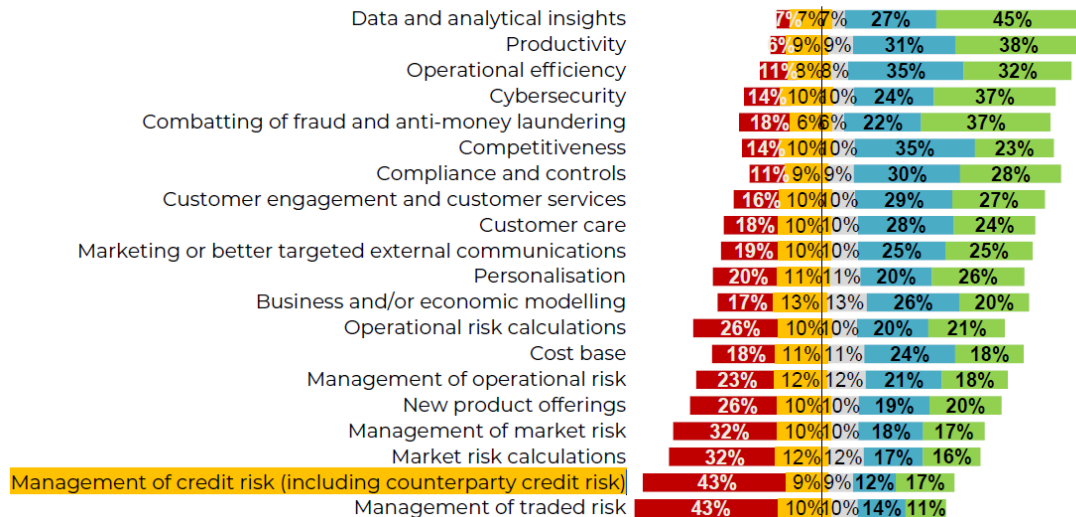


Counterfactual vs other methods: FSCA-PA Artificial Intelligence in the South African Financial Sector Survey

Benefits of using AI

Benefits of AI/ML Usage N=220

Low benefit Moderate benefit High benefit Very high benefit

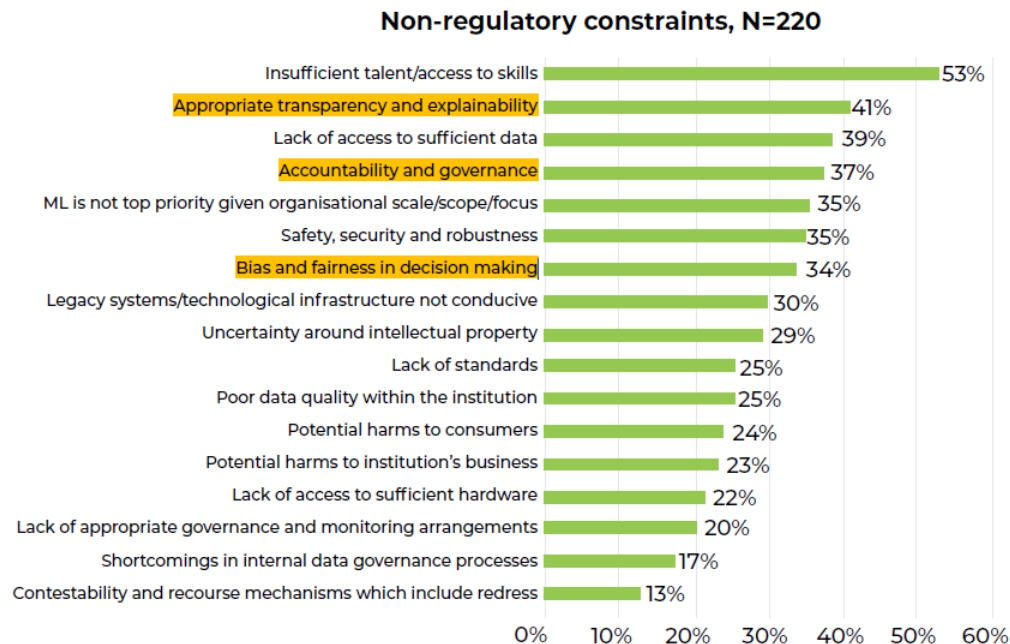


Explainability methods in South African Financial Services



Counterfactual vs other methods: FSCA-PA Artificial Intelligence in the South African Financial Sector Survey

Non-Regulatory Constraints to using AI

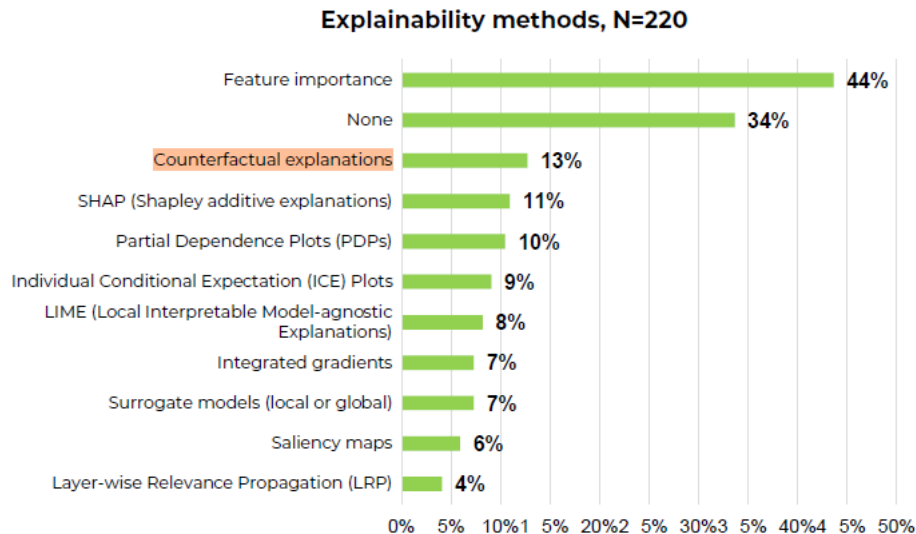


Explainability methods in South African Financial Services



Counterfactual vs other methods: FSCA-PA Artificial Intelligence in the South African Financial Sector Survey

Counterfactual based methods are increasing in prominence





Four Distinctions when Providing an Explanation

Getting the terminology right before proposing a solution and giving an explanation

Global explanation

How the model works overall — feature patterns, boundary shape

vs

Local interpretation

Why THIS specific prediction was made — and what to change

Model-specific method

Works only for one model type (e.g. gradient-based)

vs

Model-agnostic method

Works for any classifier — black box or transparent

Interpretable by design

GLM, decision tree: transparent structure, weaker predictions

vs

Post hoc explanation

Apply explanation layer AFTER fitting any complex model

Understanding WHY my loan application was rejected

Attribution: which features drove the prediction?

vs

Knowing WHAT TO CHANGE to get approved

Recourse: what actionable changes flip the outcome?

Applications if you know WHAT causes outcomes to change

Where in the loan application and management process can the information be used



Targeted Acquisitions

Convert near-declines into targets for alternative terms and specific actions to meet requirements. This supports targeted client messaging during the application process.

Industry approval rates are only ~70%, with approval at a second bank after an initial decline ~43%. Therefore, 14% uplift possible by finding credit worthy clients that need individual actions to be approved.

Risk based Credit Pricing

ML models don't naturally give PD rates, only binary outcomes. Applies risk-based pricing based on the distance from the decision boundary acts as an approximation of the probability of default (PD).

Can also provide information on profile changes required by the customer to move to a better pricing tier, supporting affordability discussions and targeted upsell.

Correct allowance for Impairments

Model the change in a client's position and distance relative to the decision boundary to approximate the change in probability of default.

Reduction in distance provides evidence of increased credit risk, even before the client misses a payment. Also provides variable attribution explanations for the factors driving provision movements, which supports reporting and audit of the change in provisions.

Targeted Collections

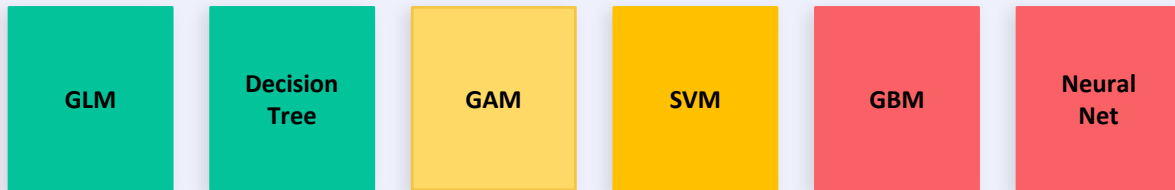
Focus the collection team's efforts and offer improved terms to clients near the decision boundary, while writing off clients far from the boundary. Effectively repricing, but determining what has gone wrong

Determine customers whose distress is driven by actionable factors, and provide solutions based on the credit model's internal logic i.e. term extension that reduces payments



The False Trade-Off

Accuracy vs interpretability — a dilemma that can be resolved



← *More Interpretable by design*

More Accurate black-box models →

Example: Smoking & Cancer Research

To conclude that smoking causes cancer, you need: (1) an accurate model predicting mortality, (2) a method to isolate each variable's contribution, and (3) a clear definition of 'significance'. Without interpretability, even a perfectly accurate model cannot prove causation.

The Resolution

- Interpretability results in accuracy, not the reverse
- Explaining a model reveals poor features that should be removed
- High-stakes contexts demand both accuracy AND an explanation
- Adding a model-agnostic explanation layer helps explain any type of model

Interpreting a GLM Model

GLM models gives an analytical solution for identifying the counterfactual and the change vector to reach it



Logistic Regression with two variables

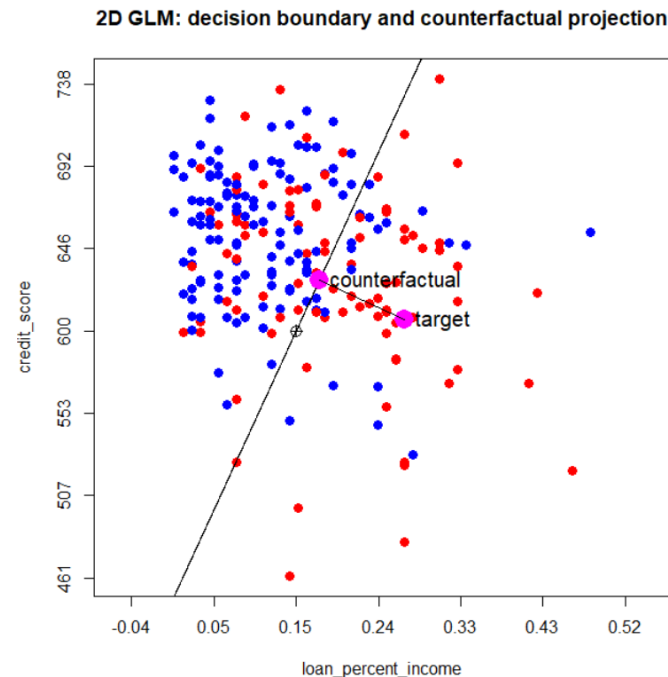
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.3662	0.1510	-2.426	0.0153 *
loan_percent_income	0.8512	0.1715	4.965	6.89e-07 ***
credit_score	-0.3920	0.1606	-2.440	0.0147 *

- +/- signs means moving to/from class 1 by increasing the variable value.
- Large coefficients indicate a bigger contribution towards the change, if standardized.
- The p-value provides a feature importance indication

For two input variables the boundary is a line in 2D space.
The projection of the data onto the boundary gives the feature values needed to change the prediction outcome

2D View — Change Vector from Target to Decision Boundary





How to interpreted a data point: Counterfactuals

The closest possible world where the outcome flips

A counterfactual is the nearest data point to an individual observation — but with the opposite predicted outcome.

ACTUAL APPLICATION

Credit Score: 600

Loan to Debt: 28%

Age: 30 Years old

✗ DECLINED



COUNTERFACTUAL

Credit Score: 625 ← increase by 25

Loan to Debt: 15 % ← reduce by 13%

Age: 25 Years old ← reduce by 5 years

✓ APPROVED



However, Not Any Change Will Do

Counterfactuals must meet four key constraints to be useful

A

Actionable

Only change features the person can actually influence. Age is immutable. Debt level is actionable. The counterfactual must respect which variables can realistically change.

Do NOT suggest: 'be 5 years younger'

F

Feasible

The proposed values must fall within realistic bounds for each feature. An income of R 500 000/month may be mathematically optimal but practically impossible.

Stay within the training data range per variable

P

Plausible

The combination of feature values must make sense together — not just individually. Higher income & low age is ideas but = implausible. Changing one variable while ignoring correlated variables can produce unrealistic profiles.

Produce counterfactuals that considers the correlation

S

Sparse

Change as few features as possible. A counterfactual that requires changing 10 variables is not actionable. One or two key changes is better than ten small ones

Identify the smallest set of changes that flips the prediction.

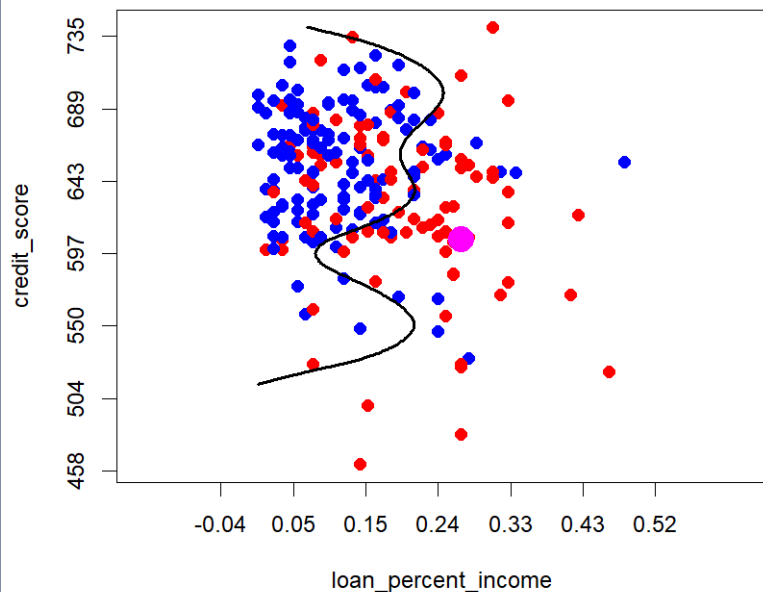
Why Machine Learning models are not interpretable

They are explainable using feature importance plot etc, but individual points are not interpretable



2D View of a GAM model with decision boundaries

2D GAM: decision boundary without projection



Three ways of finding an individual solution

Analytical Solutions: Possible for gradient based models, but limits the type of models that can be used

Grid search. Evaluate the model's predicted probability on a dense grid of input combinations: Required G^p point evaluations.

Heuristic optimisation (e.g. gradient-based search, genetic algorithms). Scale better than grid search, but may results in different solutions each time and provide no guarantee of finding the globally nearest boundary point.



The Problem is the Boundary Is Hidden

The model gives a score — but not a map of where the boundary lies or an equation that specifies the boundary value

What a machine learning model gives you

A score between 0 and 1

$f(x) = 0.73 \rightarrow \text{predicted class 1}$

No information about distance to boundary

Is this point just over 0.5, or solidly at 0.9?

No variable-level directions to the boundary

Which variable, changed by how much, would flip the prediction?

No map of the boundary shape

Where exactly does the model change its mind — and why?

If we COULD find the boundary...



Find counterfactuals

The nearest point on the boundary IS the nearest counterfactual — the minimal change that flips the prediction.



Explain variable contributions

The path from a data point to the boundary shows exactly which variables need to change — and by how much.



Measure model robustness

Distance to the boundary is a direct measure of prediction stability. Close = fragile. Far = robust.



Cluster and understand structure

Knowing boundary regions reveals how the model divides feature space — giving true algorithmic transparency.



The SHAP Problem: A Closer Look

It is widely adopted, but it answers a different question than practitioners and customers need

SHAP answers:

"Why is this prediction different from the average prediction?"

≠

What we need to know:

"What must change for this individual to get a different outcome?"

Wrong reference point

SHAP explains deviation from the global average — but a client doesn't care about the average. They want to know about THEIR situation.

Computationally expensive

Exact Shapley computation is exponential in the number of features. Approximations introduce error and instability in practice.

No actionable path

SHAP tells you what contributed — not what to change. A variable with high SHAP importance may be immutable (e.g. age). No recourse is provided.

Part 3

The Biplot Counterfactual Framework

*How biplots make decision boundaries visible,
counterfactuals findable, and predictions explainable*

In this section

Introducing biplots

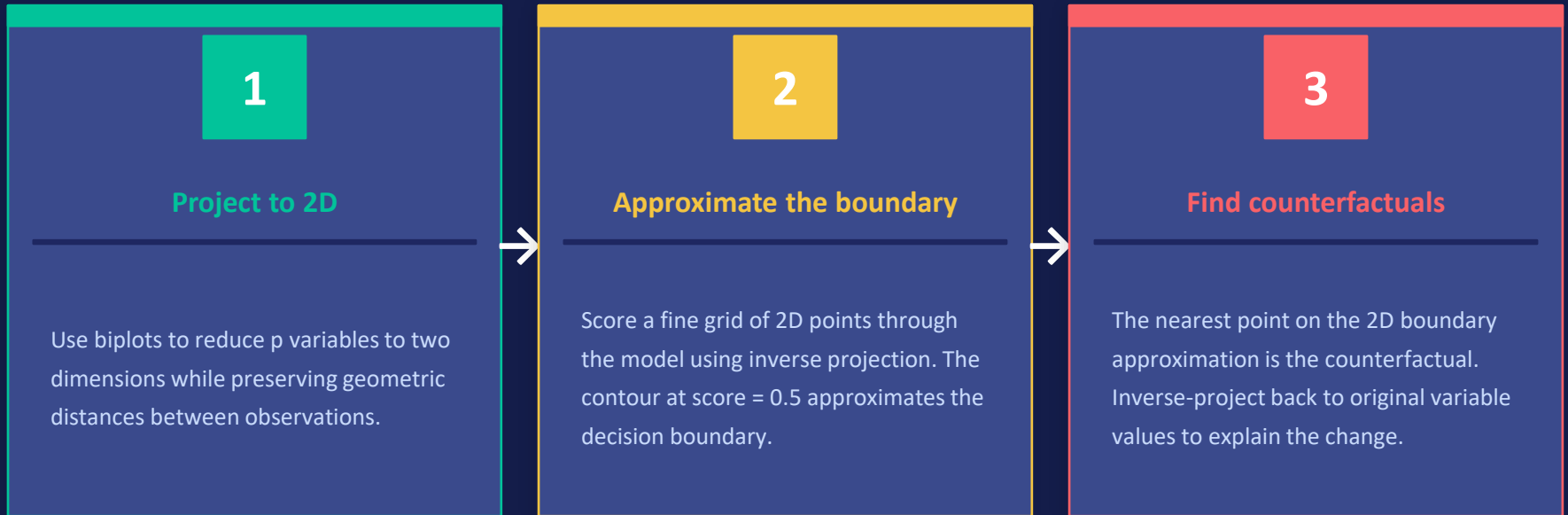
Visualising the decision boundary

Finding the counterfactual



How to find the boundary...

If we can project high-dimensional data into two dimensions — and preserve the geometric structure — we can approximate the decision boundary visually, find the nearest counterfactual, and recover the required variable changes.

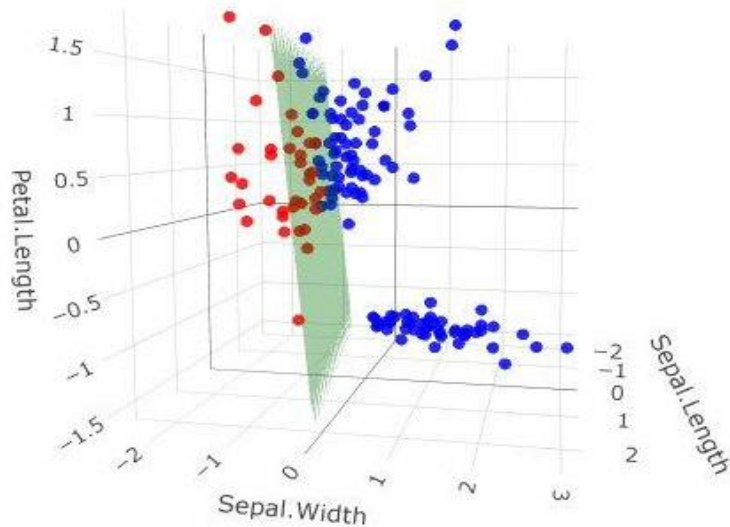


Introducing Biplots

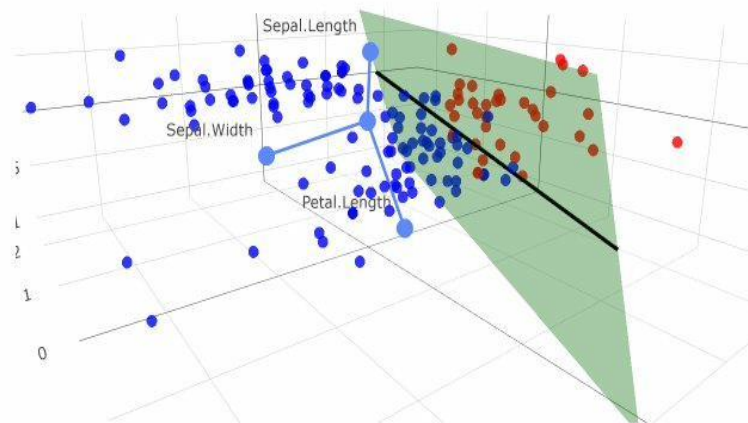
Create uncorrelated data points with the variable axes added as separate orthogonal components to the plots



3-Variable Plot with Decision Boundary



3-Variable data represented by 3D-biplot & Decision boundary

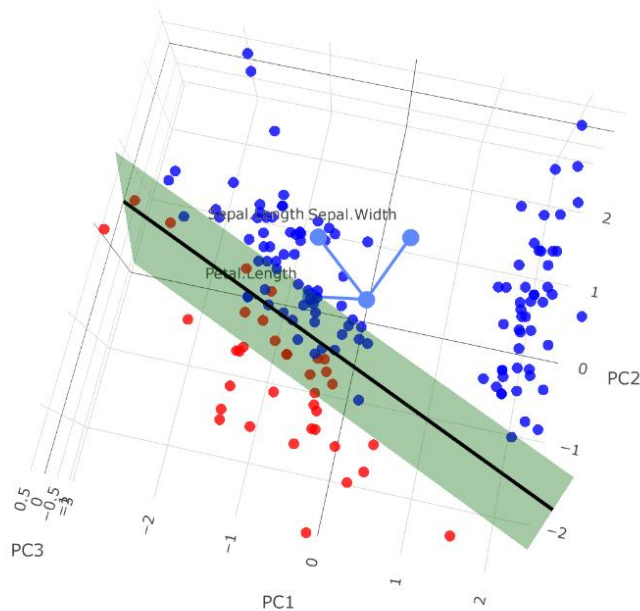


Reducing the number of dimensions

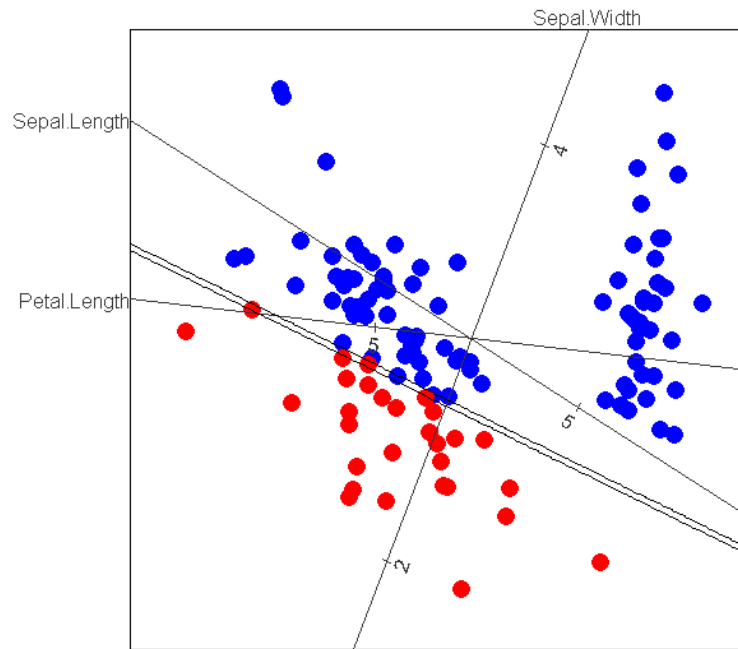
Distance between the data points indicate their similarity, axes direction indicate the correlation between variables



3-Variable data represented by 3D-biplot & Decision boundary



Projecting 3D Biplot Onto a 2D Biplot Plane

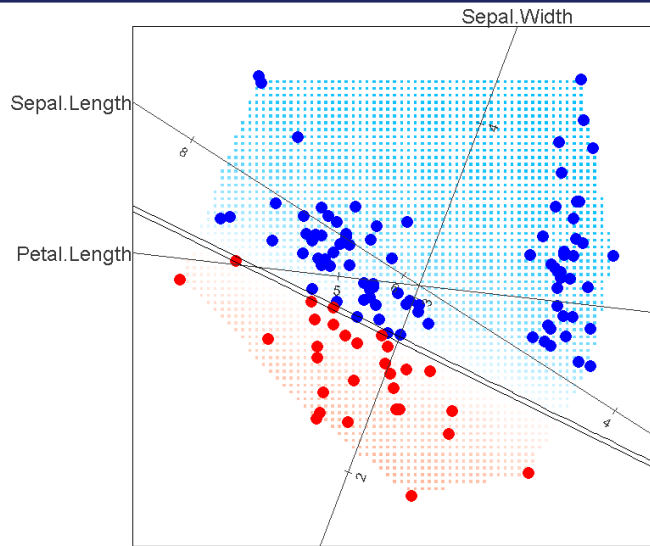


Finding Local Decision Boundaries and Counterfactuals

Each observation has its own nearest boundary segment — its local decision boundary

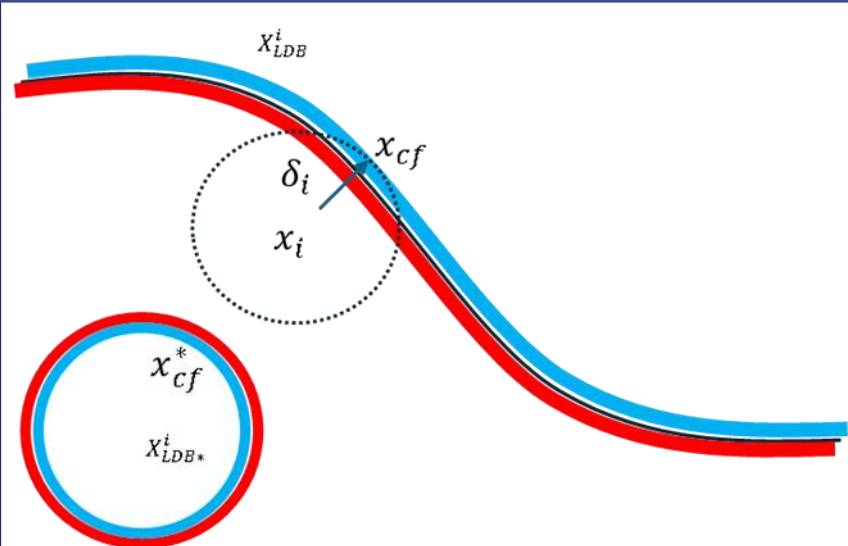


Create a grid of datapoints on the biplot plane



The grid points are predicted back to n-dimensional data as $\hat{\mathbf{X}} = \mathbf{ZV}'_2$.
These are then predicted by the model as a score between 0 and 1.
Decision boundary $\mathbf{Z}_{LDB}$ represented by contour lines along 0.5 ± 0.01 .

Multiple observations — shared boundaries



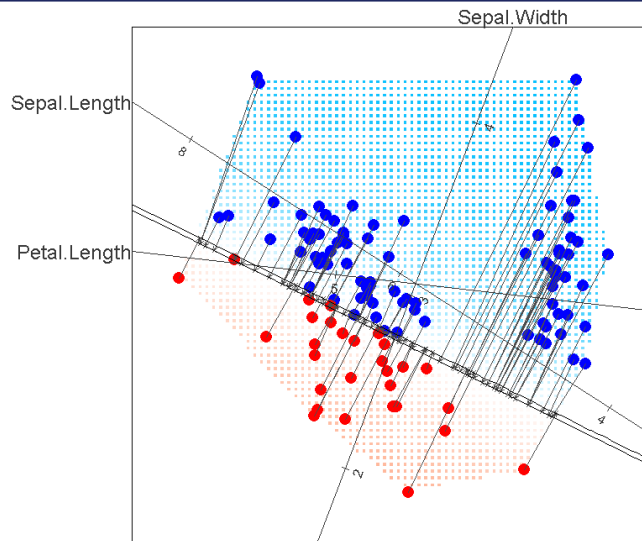
The dashed circle shows the search radius of each data point.
The nearest point on $\mathbf{Z}_{LDB}$ represents the counterfactual $\mathbf{x}_{cf}$.
Choose the nearest point between different sets of local boundaries.

Decomposing the change vector to the counterfactual

The Local Decision Boundary gives the counterfactual, which can be decomposed into the contribution by each variable

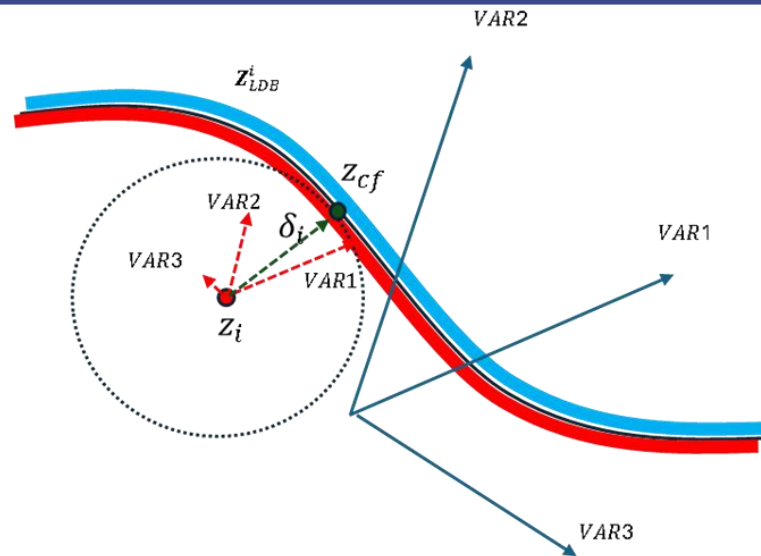


Find the change vector to the counterfactual in 2D-space



Change from the data point to its nearest counterfactual is represented by the vector $\delta_i = (z_{cf} - z_i)$. This can be converted to the change in the original dimension space $d_{ij} = (z_{cf} - z_i)V'_2$

Determine the direction and magnitude of contributing variables



This gives the direction and magnitude of the required change to reach the local decision boundary along each variable axis. A larger vector component implies a greater contribution to the prediction and change required.



Why Biplots Give Better Counterfactuals

Five advantages over existing optimisation and gradient-based methods



Visual & intuitive

The decision boundary and counterfactual are directly visible on the biplot. The impact of each variable axis on the required change is immediately apparent.



Consistent — not heuristic

The method always produces the same counterfactual for the same input. No randomness, no sensitivity to initialisation, no local minima problems.



Model-agnostic & always finds a solution

Works for any binary classifier with tabular data. Does not require gradients or model internals. Unlike analytic methods, always returns a counterfactual.



Covers the entire dataset

Counterfactuals are found simultaneously for all observations — not just one at a time. This enables both local and global interpretation in a single pass.



Local and global interpretations

Individual counterfactuals explain single predictions. Aggregated distances across all observations provide global feature importance — a novel contribution.

Part 4

Global and Local Interpretations

Using Biplots to give Global and Local interpretations

In this section

Global interpretation

Local interpretation

SHAP Comparison

Limitations

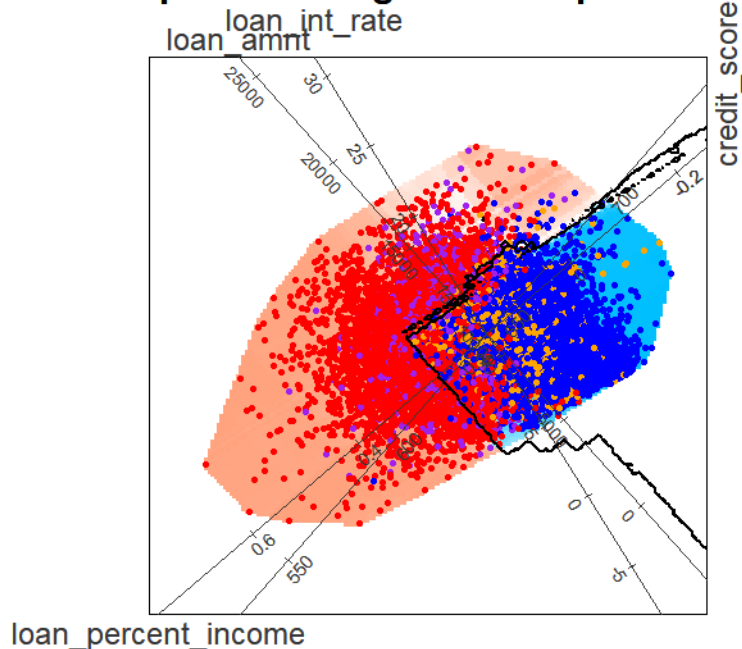
Global Explanations with the Distance to Boundary

The global biplot shows the observations and counterfactuals, allowing the change vector to be decomposed per variable

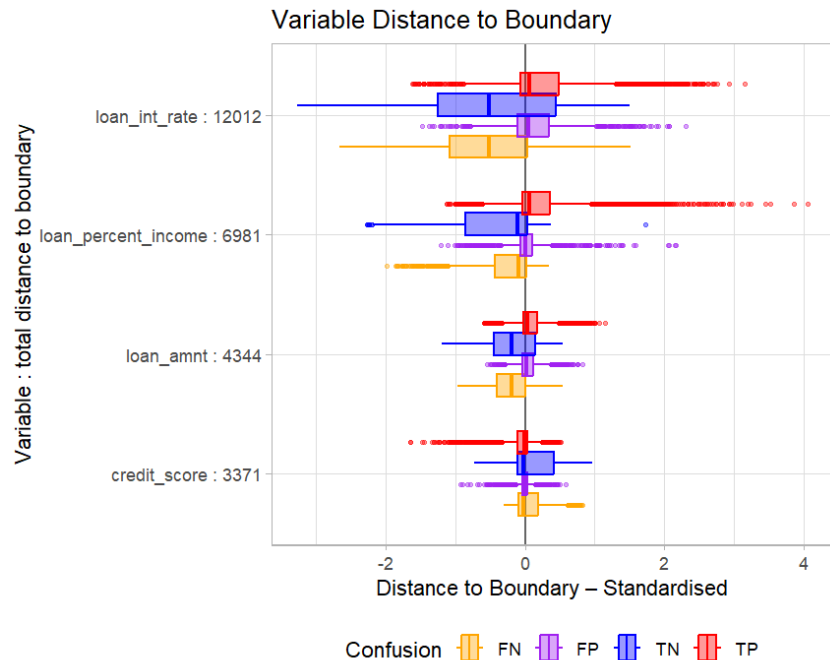


Individual Counterfactual and Decision Boundary

Biplot indicating Global Importance



Shapley Contribution Plot decomposes the contribution of each variable

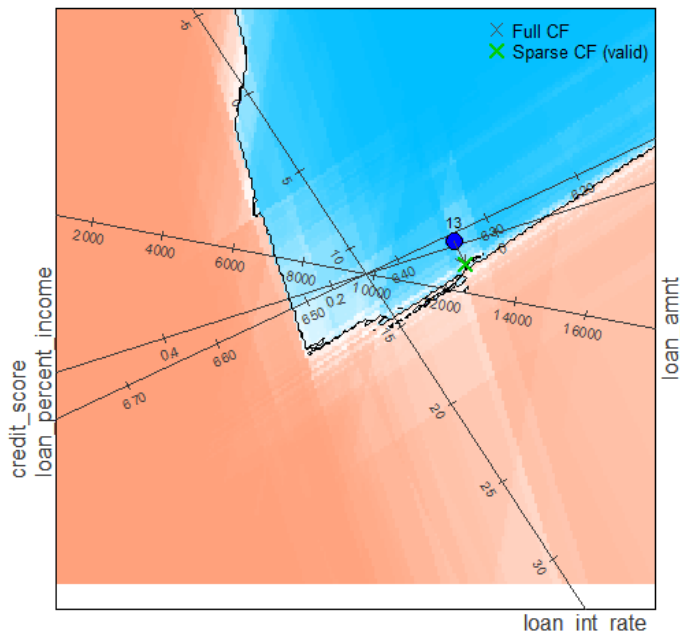


Local Interpretation using Shapley Values

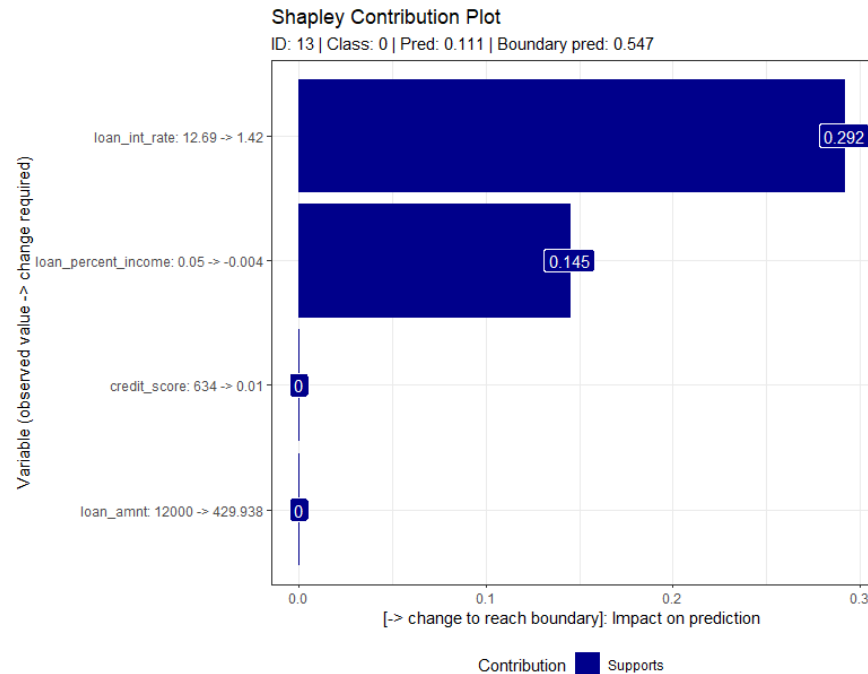
The local biplot shows the observation, its counterfactual, and the boundary. The Shapley value decomposes the contributions



Individual Counterfactual and Decision Boundary



Shapley Contribution Plot decomposes the contribution of each variable

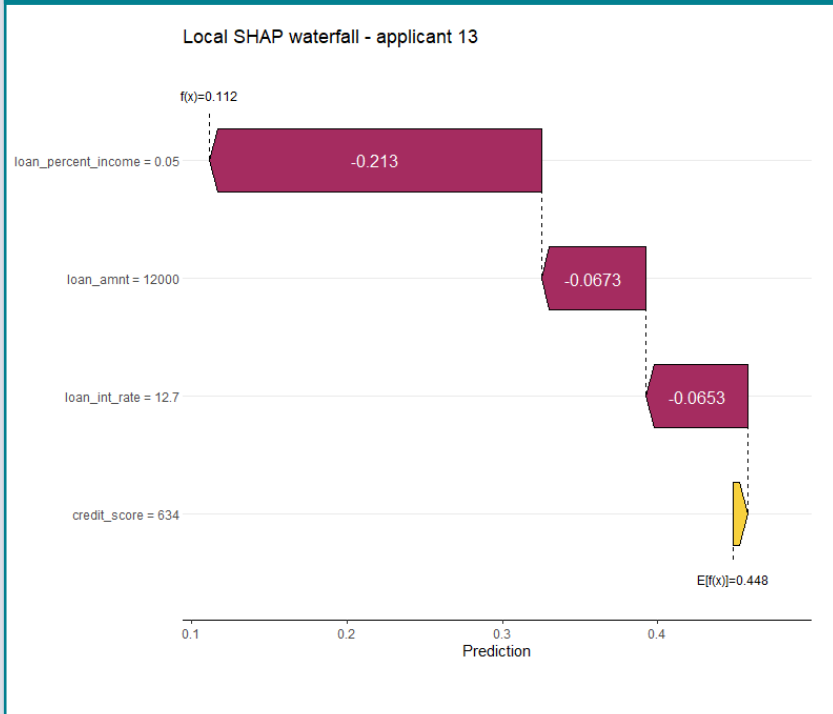




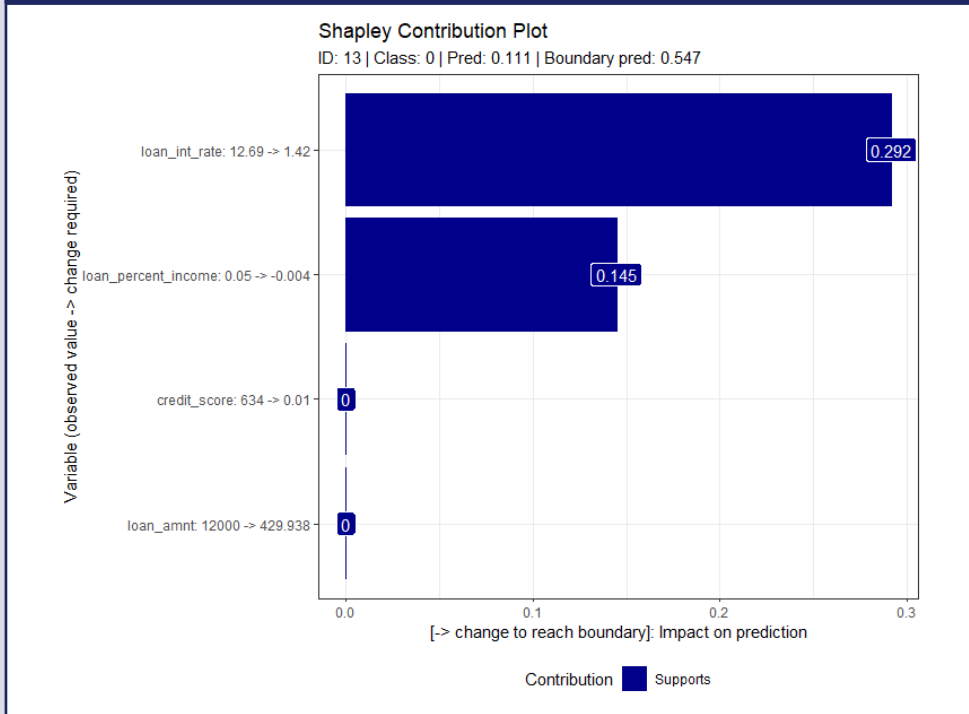
Local Interpretation compared to SHAP

Comparing the outcome of SHAP vs Distance-to-Boundary for an individual data point

SHAP equivalent values



Shapley Contribution Plot decomposes the contribution of each variable

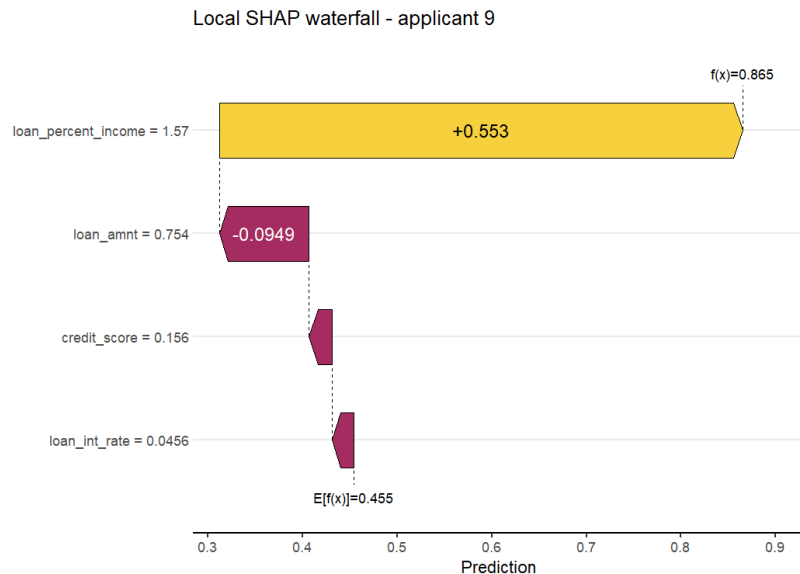


Local Interpretation compared to SHAP

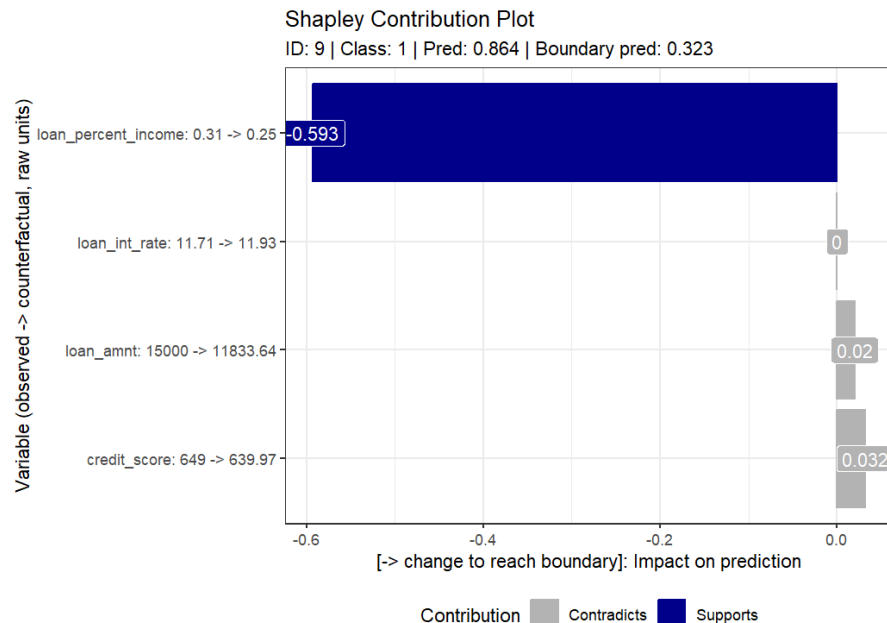
Comparing the outcome of SHAP vs Distance-to-Boundary for an individual data point



SHAP equivalent values



Shapley Contribution Plot decomposes the contribution of each variable





Limitations of Biplot Estimation of Counterfactuals

Every approximation has a cost — here is what this method trades off



2D approximation of the boundary

The counterfactual is found in a 2D projection — closer counterfactuals may exist in the full p -dimensional space. The method finds the best approximation, not the guaranteed global minimum.



Shapley computation scales with variables

Exact Shapley calculation is exponential in the number of variables. For large feature sets, permutation approximation is used — introducing some estimation error. However, because it is local (single point), computation time remains manageable.



Distance-to-boundary as importance proxy

Using boundary distance as a feature importance metric is a fast alternative to Shapley values across the full dataset — but it is not as precise. It is a useful ranking tool, not an exact attribution.



Current scope: binary, continuous, tabular

The method currently handles binary outcomes with continuous variables and tabular data. Extensions to multi-class problems, ordinal variables, and regression outcomes are directions for future work.

boundary_logic Github Account

Download the R package



Web address

Github:

https://github.com/AI-Rowan/boundary_logic

In R:

```
remotes::install_github(  
  "AI-Rowan/boundary_logic",  
  ref = "method_developments",  
  dependencies = TRUE)
```

Email: Adriaan.Rowan@yala.co.za

QR Code

