Counterfactual Explainations for Loan Approval Dataset

Megha Akanksha G

2303A52058

Batch 39

**Objectives and Methodology**

The primary objective of this lab is to evaluate how small, actionable changes in applicant features can influence loan approval decisions using counterfactual explanations. The overall methodology involves loading and preprocessing a large real-world loan approval dataset, training multiple classifiers, and generating counterfactual instances using the DiCE library. The workflow explicitly covers data cleaning, encoding, model training (Logistic Regression and Random Forest), and interpretability analysis through counterfactuals.

**Dataset Description**

The loan approval dataset contains thousands of records with features relevant to credit risk, including:

Demographics: education status, number of dependents.

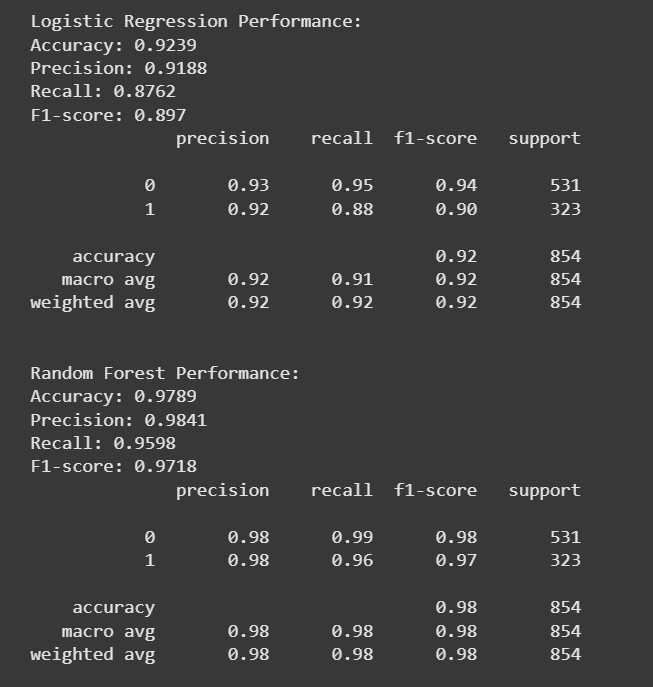
Financial attributes: income, loan amount, loan term, CIBIL score, residential/commercial/luxury/bank asset values.

Target variable: loanstatus (Approved/Rejected).

Irrelevant columns (e.g., “loanid”) are dropped during preprocessing, and categorical variables (e.g., Graduate/Not Graduate, Self Employed) are encoded for modeling. Missing values in numeric columns are imputed with medians, while categorical ones use the mode. The label distribution is non-trivial, allowing for meaningful model training and evaluation.

**Model Performance Results**

Both Logistic Regression and Random Forest classifiers are trained on the dataset with proper stratified train-test splitting. Metrics reported on the test set include:

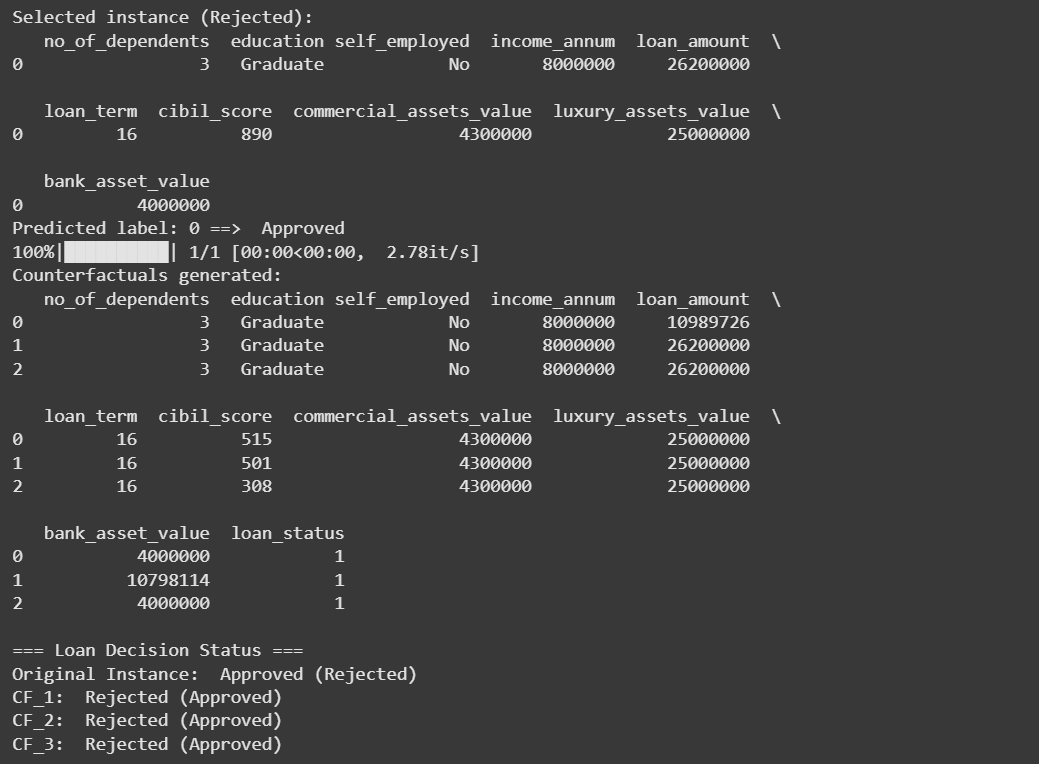


The Random Forest model demonstrates superior performance and is chosen for counterfactual analysis. Model results indicate solid predictive ability, but also highlight practical limitations in feature sensitivity.

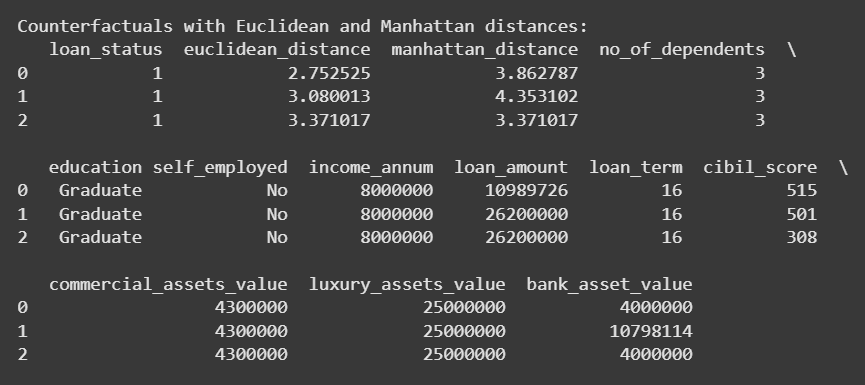
**Counterfactual Examples**

Counterfactual explanations are generated for randomly chosen “Rejected” instances using the DiCE library. By varying all features, the system creates examples that “flip” the model’s decision to “Approved” with minimal change(see the generated image above):

Each counterfactual is compared with the original rejected application, showing the smallest edits required for approval.



Distances (Euclidean and Manhattan) between the original and counterfactual instances are computed, illustrating how close the applicant is to approval in the feature space.



**Interpretations and Reflections**

The experiments prove that small, actionable changes (income, loan amount, credit history) can alter model outcomes for applicants.

Counterfactuals provide clear “what-if” scenarios for stakeholders, increasing trust in the decision process and showing tangible paths to approval.

Such transparency empowers end-users by clarifying how their profiles affect outcomes and what minimum improvements can change a rejection into approval.

Model interpretability through counterfactuals expands the practical utility of machine learning for financial services.