Counterfactual Explainations for Loan Approval Dataset

Humera Nuzhat
2303A52083
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:

```
Logistic Regression Performance:
Accuracy: 0.9239
Precision: 0.9188
Recall: 0.8762
F1-score: 0.897
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       531
           1       0.92      0.88      0.90       323

    accuracy                           0.92       854
   macro avg       0.92      0.91      0.92       854
weighted avg       0.92      0.92      0.92       854


Random Forest Performance:
Accuracy: 0.9789
Precision: 0.9841
Recall: 0.9598
F1-score: 0.9718
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       531
           1       0.98      0.96      0.97       323

    accuracy                           0.98       854
   macro avg       0.98      0.98      0.98       854
weighted avg       0.98      0.98      0.98       854
```

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.

```
Selected instance (Rejected):
   no_of_dependents  education self_employed  income_annum  loan_amount  \
0                 3   Graduate            No       8000000     26200000

   loan_term  cibil_score  commercial_assets_value  luxury_assets_value  \
0         16          890                  4300000             25000000

   bank_asset_value
0           4000000
Predicted label: 0 ==>  Approved
100%|██████████| 1/1 [00:00<00:00,  2.78it/s]
Counterfactuals generated:
   no_of_dependents  education self_employed  income_annum  loan_amount  \
0                 3   Graduate            No       8000000     10989726
1                 3   Graduate            No       8000000     26200000
2                 3   Graduate            No       8000000     26200000

   loan_term  cibil_score  commercial_assets_value  luxury_assets_value  \
0         16          515                  4300000             25000000
1         16          501                  4300000             25000000
2         16          308                  4300000             25000000

   bank_asset_value  loan_status
0           4000000            1
1          10798114            1
2           4000000            1

=== Loan Decision Status ===
Original Instance:  Approved (Rejected)
CF_1:  Rejected (Approved)
CF_2:  Rejected (Approved)
CF_3:  Rejected (Approved)
```

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

```
Counterfactuals with Euclidean and Manhattan distances:
   loan_status  euclidean_distance  manhattan_distance  no_of_dependents  \
0            1            2.752525            3.862787                 3   
1            1            3.080013            4.353102                 3   
2            1            3.371017            3.371017                 3   

   education self_employed  income_annum  loan_amount  loan_term  cibil_score  \
0   Graduate            No       8000000     10989726         16          515   
1   Graduate            No       8000000     26200000         16          501   
2   Graduate            No       8000000     26200000         16          308   

   commercial_assets_value  luxury_assets_value  bank_asset_value  
0                  4300000             25000000           4000000  
1                  4300000             25000000          10798114  
2                  4300000             25000000           4000000  
```

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.