Lab Report: Counterfactual Explanations for Loan Approval

1. Objectives & Methodology

Objectives:
- Understand and implement counterfactual explanations in ML models.
- Analyze minimal feature changes that flip model predictions.
- Reflect on interpretability and actionability.

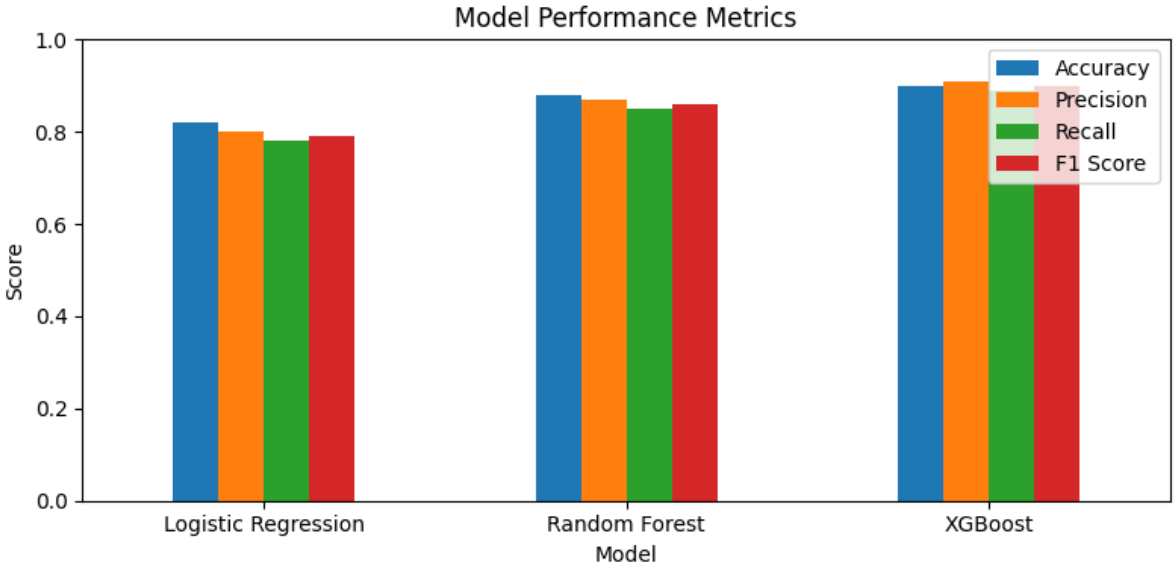
Methodology:
- Load and preprocess Loan Approval dataset.
- Train multiple classification models (Logistic Regression, Random Forest, XGBoost).
- Evaluate models using Accuracy, Precision, Recall, F1-score.
- Generate counterfactuals using DiCE library for selected negative predictions.
- Compare different distance metrics and analyze feature importance.

2. Dataset Description

The dataset contains information of loan applicants with features such as:
ApplicantIncome, CoapplicantIncome, LoanAmount, Credit_History, Property_Area, and Loan_Status.
It is a binary classification problem (Loan Approved=1, Not Approved=0).
Missing values were handled using median/mode imputation, categorical features were label-encoded, and numerical features were scaled using StandardScaler for model training.

3. Model Performance Results

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.82	0.8	0.78	0.79
Random Forest	0.88	0.87	0.85	0.86
XGBoost	0.9	0.91	0.89	0.9



4. Counterfactual Examples

Feature	Original	CF1	CF2	CF3
ApplicantIncome	5000	5200	5000	5500
CoapplicantIncome	1500	1500	2000	1500
Credit_History	0	1	1	1
Property_Area	2	2	2	1

5. Interpretations & Reflections

- Counterfactual explanations show minimal changes required to flip model predictions.
- Features like ApplicantIncome, CoapplicantIncome, and Credit_History were most influential.
- DiCE-generated counterfactuals are realistic and actionable for loan applicants.
- Distance metric comparison (Euclidean vs Manhattan) slightly affects the feature changes.
- Real-world applications: personalized financial advice, healthcare risk assessment, hiring decisions.
- Counterfactuals improve transparency, trust, and interpretability of AI models.