Milestone 3 — Baseline Model Development (CreditPathAI)

Objective:
The goal of Milestone 3 was to **build a baseline machine learning model** for predicting loan default using logistic regression. This step establishes a reference performance level that future advanced models can improve upon. The milestone also includes applying preprocessing, generating key performance metrics, and producing visualization graphs to better understand model behaviour.

1. Objectives of Milestone 3

- Build the **first baseline model** using Logistic Regression.
- Implement a **feature engineering pipeline** to automate preprocessing.
- Train and evaluate the model using the cleaned dataset.
- Compute essential classification metrics such as:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - **AUC-ROC Score**
- Generate graphical visualizations of the model's performance.
- Save the trained model (logistic_baseline.pkl) for future milestones.

2. Feature Engineering Pipeline

To ensure consistent preprocessing, a **Column Transformer + Pipeline** structure was built that automatically applies:

Numerical Column Processing

- Standardization using **Standard Scaler**
- Ensures numerical features have consistent scale, improving model stability.

Categorical Column Processing

- One-hot encoding using **OneHotEncoder**
- Converts non-numeric features into model-friendly binary vectors.

This pipeline ensures:

- No data leakage
- Fully automated preprocessing during both training & inference
- Clean, reproducible machine learning workflow

3. Baseline Model: Logistic Regression

A Logistic Regression classifier was selected as the baseline model because:

- It is simple and interpretable
- Fast to train
- Provides a good starting point for comparing future models
- Works well with standardized and encoded data

Parameters used:

LogisticRegression(max_iter=1000)

The model was wrapped inside the pipeline to ensure preprocessing happens automatically.

4. Model Training & Evaluation

The dataset was split using an **80:20 train-test split** with a fixed random seed for reproducibility.

Metrics calculated include:

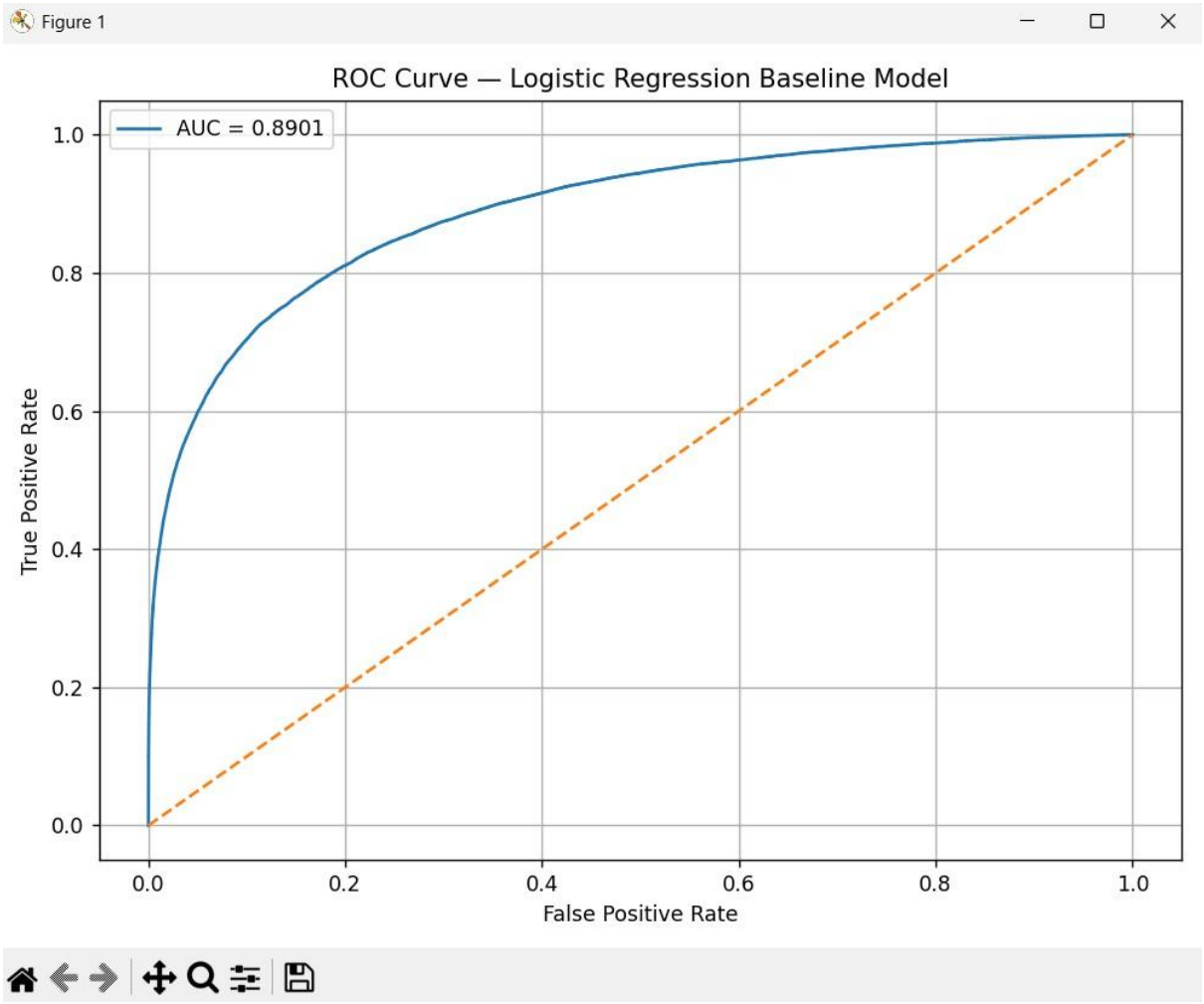
- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **AUC-ROC Score (key milestone requirement)**

A confusion matrix was also generated to analyse prediction errors.

5. Graphical Representations

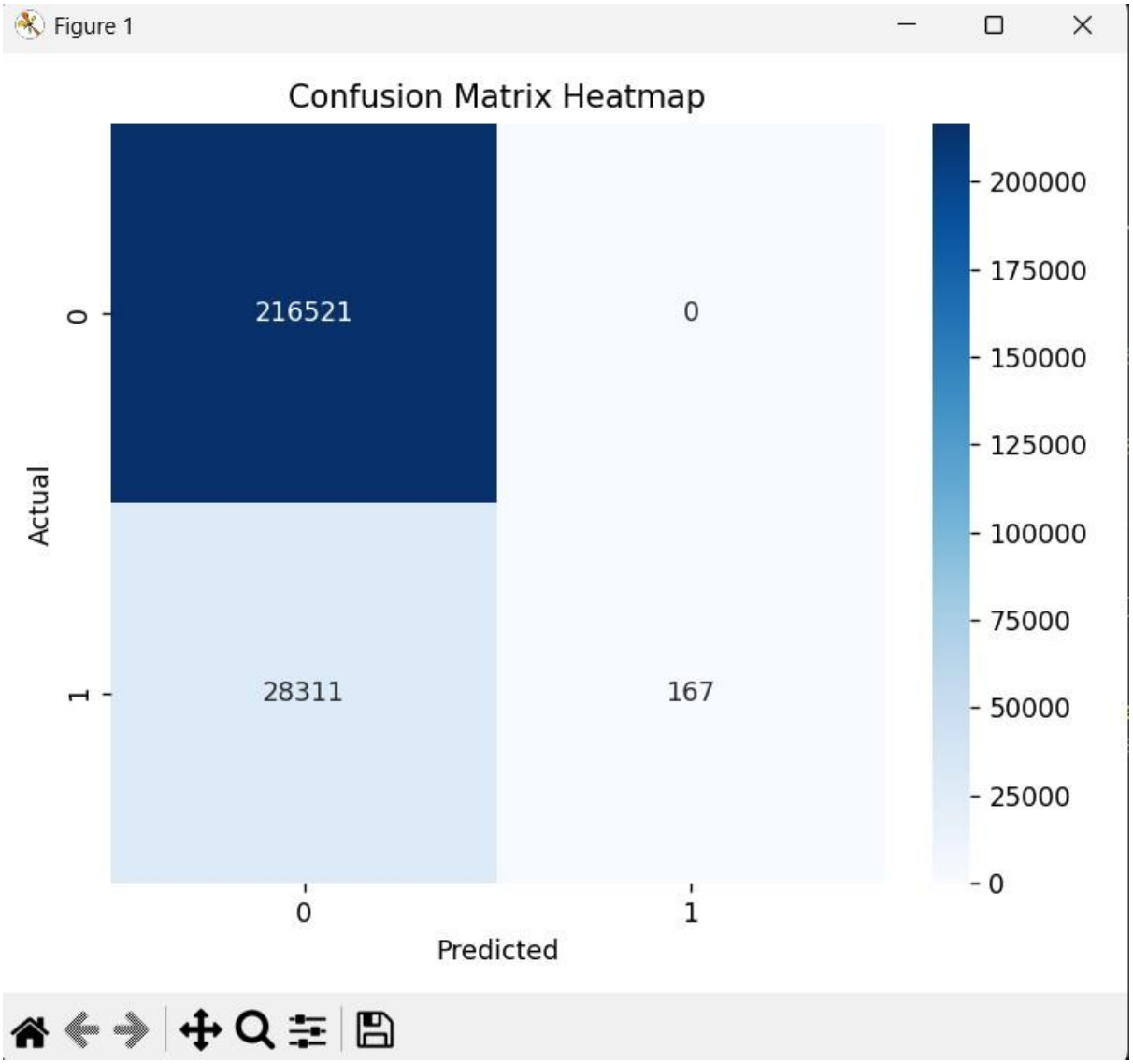
To make the evaluation more visual, four essential graphs were generated.

ROC Curve Graph: -



The ROC curve illustrates the model's ability to distinguish between default and non-default classes across different thresholds. AUC shows overall model quality.

Confusion Matrix Heatmap: -

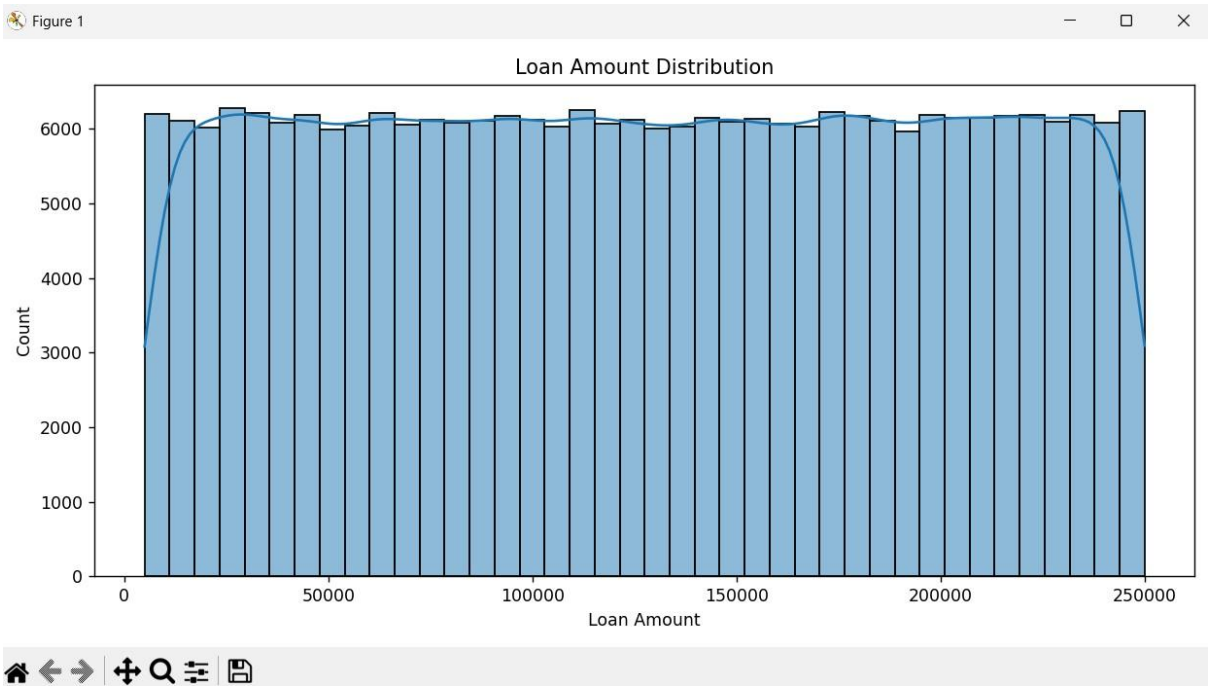


This heatmap visualizes:

- True Positives
- True Negatives
- False Positives
- False Negatives

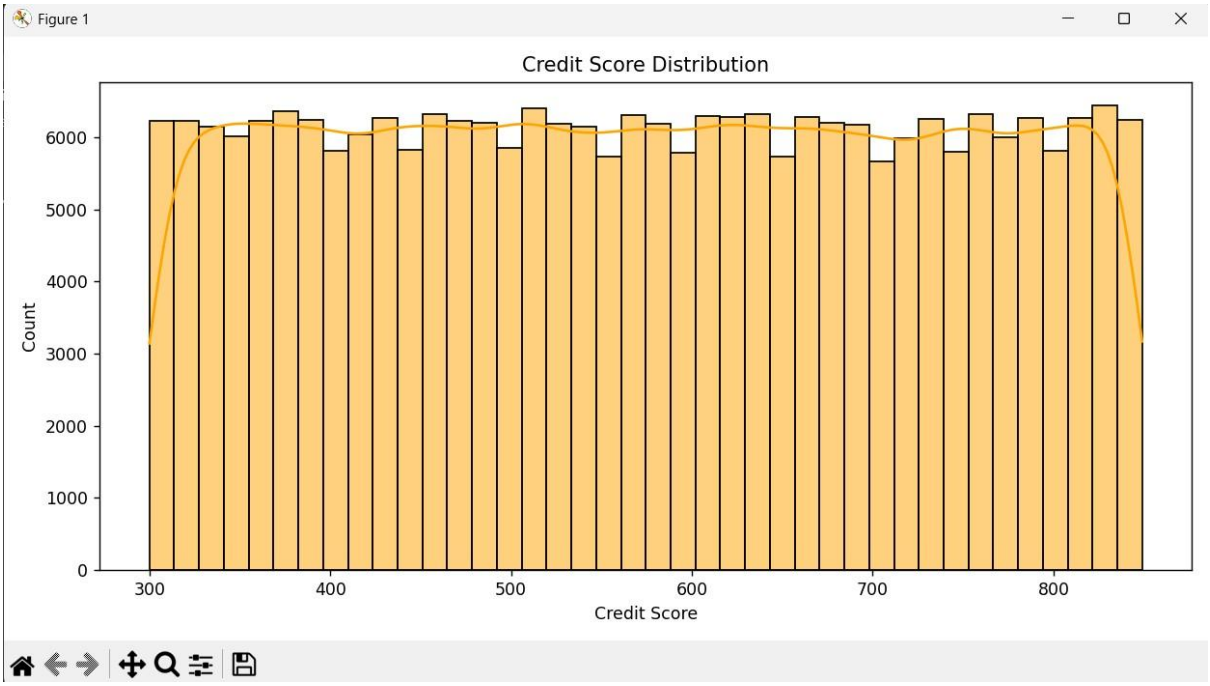
It helps identify biases, misclassifications, and class imbalance effects.

Loan Amount Distribution Graph: -



Shows how LoanAmount is distributed across borrowers and helps understand feature behaviour.

Credit Score Distribution Graph: -



Shows the density distribution of Credit Score and helps identify creditworthiness patterns.