LOAN DEFAULT
PROBABILITY PREDICTION

Presented By: Nisarga Kadam

01

PROBLEM STATEMENT

- In 2025, rising interest rates and record consumer debt have intensified the need for accurate credit risk assessment across global lending markets.
- As the financial sector accelerates adoption of AI-driven decision systems, institutions are turning to machine learning models to quantify borrower risk and promote transparent, responsible lending.
- This project develops a logistic regression baseline to predict the probability of loan repayment, establishing a foundation for more advanced AI models such as LightGBM or XGBoost.
- Evaluation Metric: Area Under the ROC Curve (AUC), measuring the model's ability to rank borrowers by repayment likelihood.

DATA OVERVIEW 02

- 593994 Rows, 13 features (including target variable)
- Features describe the demographics of borrowers, such as income, credit history, loan attributes, and risk grades
- Mix of numeric and categorical variables:
  - Numeric: annual_income, debt_to_income_ratio, credit_score, loan_amount, interest_rate
  - Categorical: gender, marital_status, education_level, employment_status, loan_purpose, grade_subgrade
- Target variable: loan_paid_back (0 = not repaid, 1 = repaid)

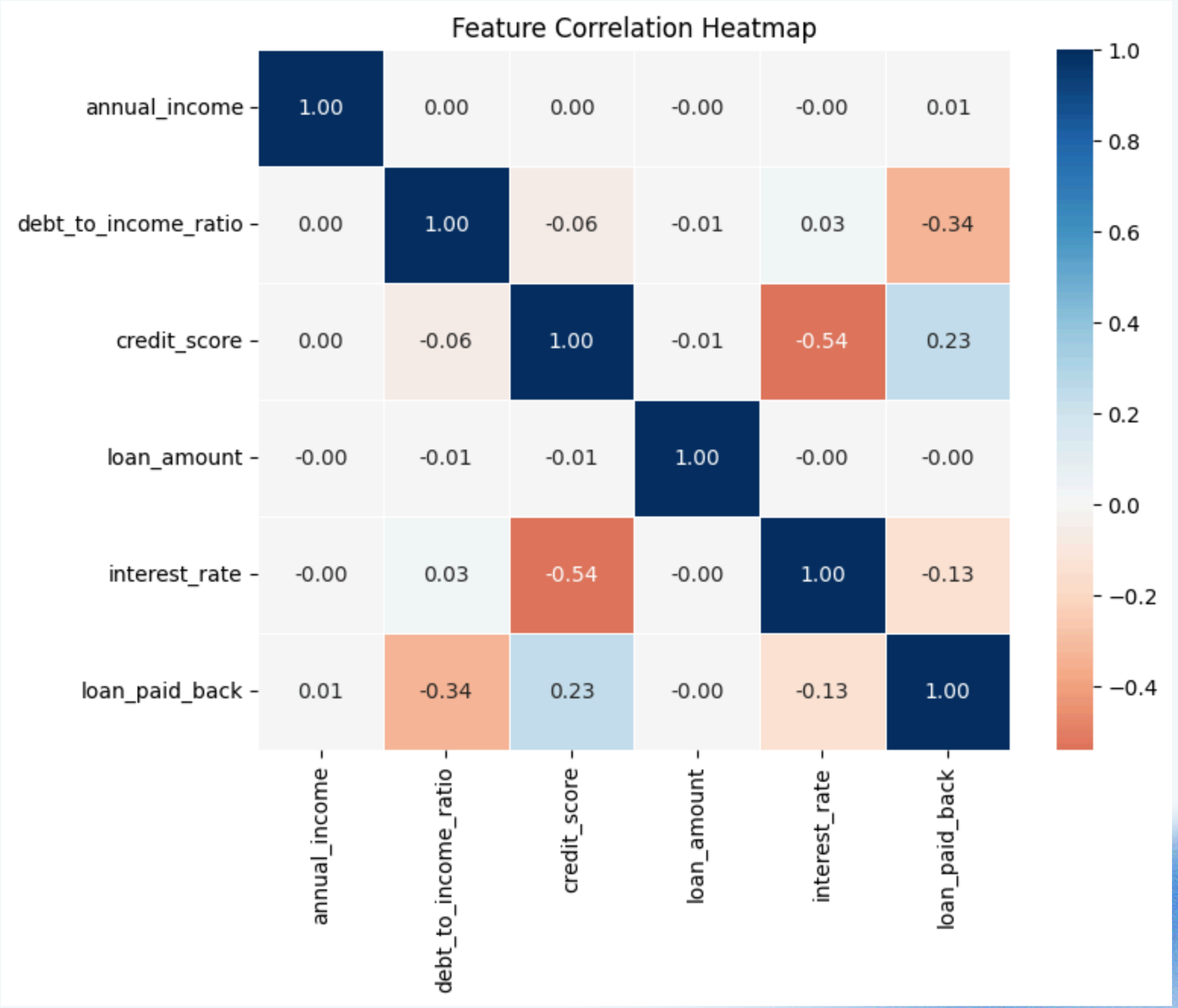
METHODOLOGY

- Selected key financial and demographic variables: annual_income, debt_to_income_ratio, credit_score, loan_amount, interest_rate, and borrower profile features
- Applied preprocessing pipeline with one-hot encoding for categorical data and standard scaling for numeric variables
- Trained a Logistic Regression model for binary prediction of loan_paid_back (0 = default, 1 = repaid)
- Configuration:
  - Solver: lbfgs
  - Max iterations: 1000
  - Parallelization: n_jobs = -1
- Why Logistic Regression?
  - Provides interpretable coefficients for each feature's impact on repayment likelihood
  - Efficient and stable for linearly separable data
  - Produces probabilistic outputs required for AUC evaluation

03

04

The correlation matrix shows that higher credit scores and lower debt-to-income ratios are associated with a higher likelihood of repayment. Most features are weakly correlated, suggesting a diverse and non-redundant feature set.



RESULTS

- The model effectively distinguishes between high- and low-risk borrowers, achieving a 0.90 AUC on validation data.
- Captures intuitive financial relationships: borrowers with higher credit scores and lower debt-to-income ratios show higher repayment likelihood, while higher interest rates slightly increase default risk.
- The correlation matrix confirms these trends, reinforcing the model's interpretability and alignment with real-world credit behavior.
- Provides a strong, explainable baseline for future experimentation with ensemble and gradient-boosting models.

05

IMPLICATIONS

- Demonstrates how AI-driven credit scoring can enhance transparency, speed, and fairness in global lending decisions.
- As financial institutions modernize risk assessment pipelines, interpretable ML models like logistic regression remain vital for regulatory compliance and model auditability.
- Long-term, such probabilistic models can integrate with automated credit platforms, improving credit access and reducing systemic default risk worldwide.
- Forms a foundation for scaling toward ensemble learning, real-time credit analytics, and AI-governed financial risk systems.

06