Kerwin

Internship

01

02

03

Machine Learning

Credit Risk Prediction Model for Lending Company

End to End Data Science Solution

Final Project

25 Dec 2025

ID/X Partners x Rakamin Academy

- Lending companies face financial losses due to loan defaults
- Manual credit assessment is inconsistent and subjective
- Early prediction of credit risk enables better decision making

Objective

Build a machine learning model to predict the probability of loan default using historical loan data.

01

02

03

Business Problem

Applicant → Model → Approve / Reject

Dataset Overview

- Source: Lending Company Historical Loan Data (2007–2014)
- Total records: ~460,000 loans
- Data types:
- Borrower profile
- Loan characteristics
- Credit history
- Target variable: loan_status

Target Definition

- Bad Loan (1):
  - Charged Off
  - Default
  - Late (16–30 days, 31–120 days)
- Good Loan (0):
  - Fully Paid

Data Preparation & Cleaning

- Filtered only relevant loan statuses
- Converted loan outcome into binary target variable
- Removed data leakage features such as:
  - Total payment
  - Recovery amount
  - Outstanding principal
- Handled missing values:
  - Numerical: median imputation
  - Categorical: most frequent value

```python
bad_status = [
    "Charged Off",
    "Default",
    "Late (31-120 days)",
    "Late (16-30 days)"
]

good_status = ["Fully Paid"]

df = df[df["loan_status"].isin(bad_status + good_status)].copy()
df["target"] = (df["loan_status"].isin(bad_status)).astype(int)

df["target"].value_counts(normalize=True)
```

target	proportion
0	0.782249
1	0.217751

dtype: float64

01

02

03

Why this matter?

To Prevents unrealistic model performance and ensures real-world usability.

Modeling Approach

1

Logistic Regression

- Interpretable baseline model
- Suitable for binary classification

2

Random Forest

- Captures non linear patterns
- Handles feature interactions well

3

Data Split

- 80% Training
- 20% Testing
- Stratified sampling to preserve class distribution

01

02

03

Metric Used

- ROC AUC
  - Measures model's ability to distinguish bad vs good loans
  - Robust to class imbalance

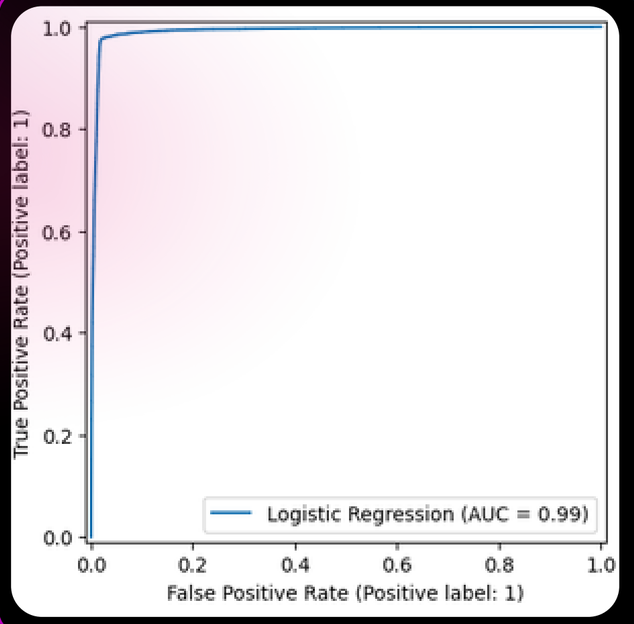
Results

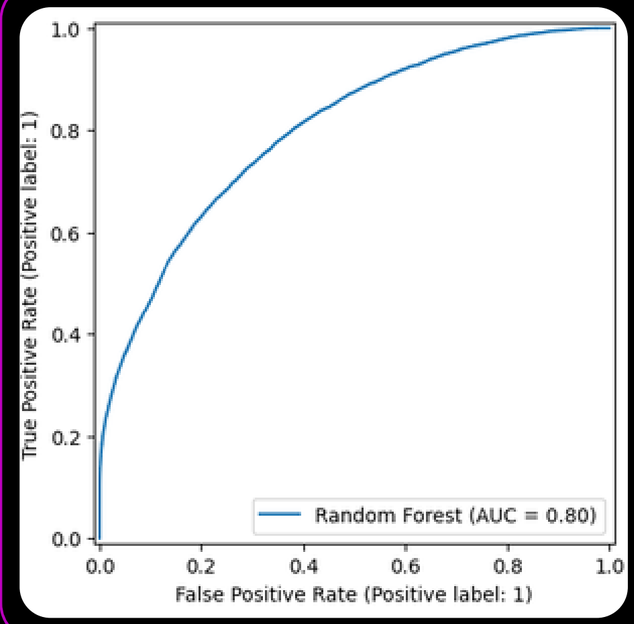
- Logistic Regression ROC AUC: ~0.7x
- Random Forest ROC AUC: ~0.8x (better)

Model Evaluation

roc_auc_score(y_test, y_pred_prob)

ROC Curve comparison





Kerwin

Key Risk Driver

Feature Importance

```
rf_feature_names = (
    rf_model.named_steps["preprocess"]
    .get_feature_names_out()
)

importances = rf_model.named_steps["model"].feature_importances_

feature_importance = pd.DataFrame({
    "feature": rf_feature_names,
    "importance": importances
}).sort_values(by="importance", ascending=False)

feature_importance.head(15)
```

Top contributing features identified by Random Forest:

- Interest rate
- Loan grade
- Debt to income ratio
- Annual income
- Loan term

Insight

Borrower credit profile and loan pricing strongly influence default risk.

Kerwin

Business Recomendation

- Random Forest selected as final model
- Use predicted default probability to:
  - Adjust loan approval thresholds
  - Support credit analyst decisions
  - Reduce default losses
- Model can be integrated into loan approval workflow

- Hyperparameter tuning
- Model monitoring
- Periodic retraining with new data

This solution helps balance loan growth and credit risk.

01

02

03

Thank You

FOR YOUR ATTENTION