



ID/X PARTNERS — RAKAMIN ACADEMY
FINAL PROJECT DATA SCIENTIST INTERN

CREDIT RISK PREDICTION

BY MUHAMMAD GHAZA EKA PUTRA



BUSINESS UNDERSTANDING

OVERVIEW

Latar Belakang Masalah :

- Perusahaan multifinance menghadapi risiko gagal bayar dari nasabah.
- Penilaian kelayakan kredit secara manual tidak efisien dan rawan subjektivitas.

Tujuan :

- Mengembangkan sistem prediksi kelayakan kredit untuk mengklasifikasikan nasabah sebagai berisiko tinggi atau tidak.
- Mengurangi risiko kerugian akibat gagal bayar dan mendukung profitabilitas perusahaan.

Manfaat :

- Pengambilan keputusan kredit lebih cepat dan objektif.
- Mengurangi kerugian dari nasabah yang tidak mampu melunasi pinjaman.





DATASET OVERVIEW

UKURAN DATASET

- Total baris: 466.285
- Total kolom: 75

TIPE DATA

- Numerik: 53 kolom
- Kategorikal: 75 kolom

KUALITAS DATA

- Terdapat banyak Missing Value
- Tipe data yang tidak sesuai

HIGHLIGHT

- Kolom **loan_status** adalah target untuk prediksi risiko kredit.
- Banyak kolom terkait dengan riwayat kredit, penghasilan, dan detail pinjaman.

CONTOH KOLOM PENTING

- loan_amnt, term, int_rate, grade, emp_length, annual_inc, loan_status, dti, recoveries dll

DATA PREPARATION

DATA CLEANING

- Menghapus kolom dengan missing value > 50%
- Menghapus kolom tidak relevan (e.g., desc, url, emp_title, member_id)
- Mengisi missing value:
 - Numerik → median
(Presentase Sedang 20% -40%)
 - Tanggal → modus
(Presentase kecil < 10%)

ENCODING

Encoding Fitur Kategorikal

- term: 36 Month → 0 dan 60 Month → 1
- pymnt_plan: n → 0 dan y → 1
- verification_status: Not Verified → 0, lainnya → 1
- grade: A–G → 0–6 (ordinal)

Encoding Target

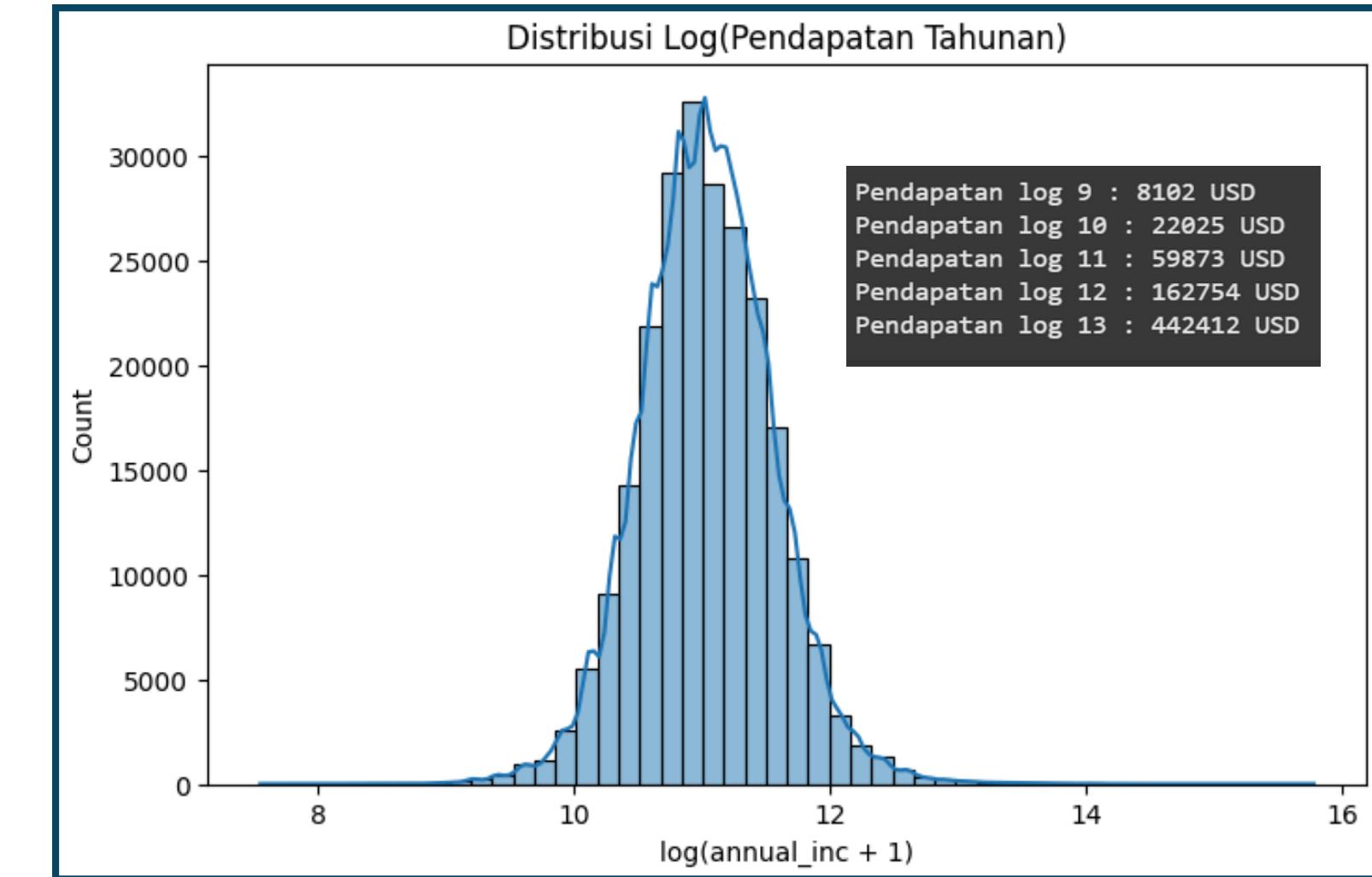
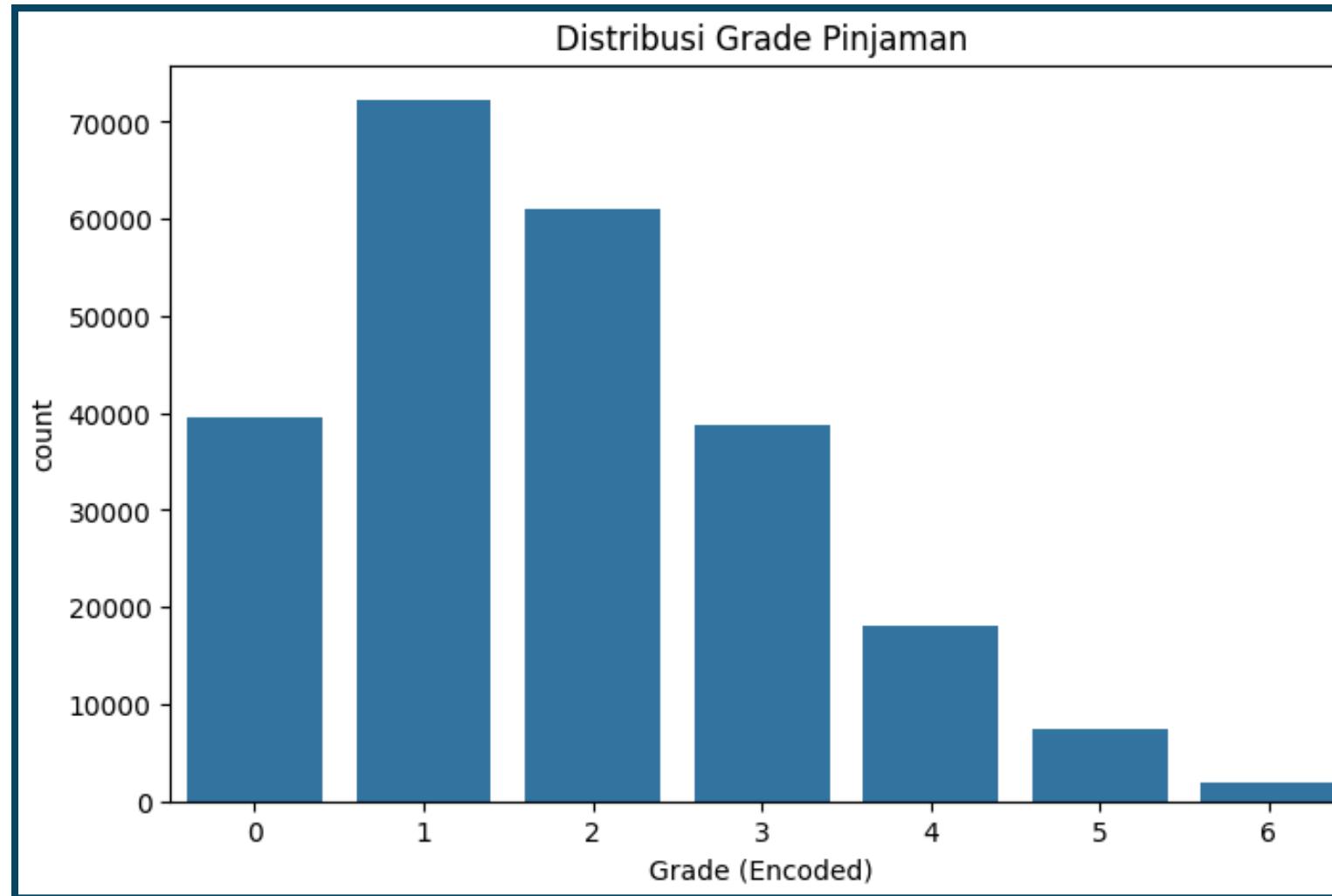
- loan_status diklasifikasikan menjadi 2 kelas:
 - 0 (Good): Fully Paid
 - 1 (Bad): Charged Off, Default, Late, dll

Transformasi Lain

- emp_length dikonversi dari string ke numerik

EXPLORATORY DATA ANALYSIS

EDA

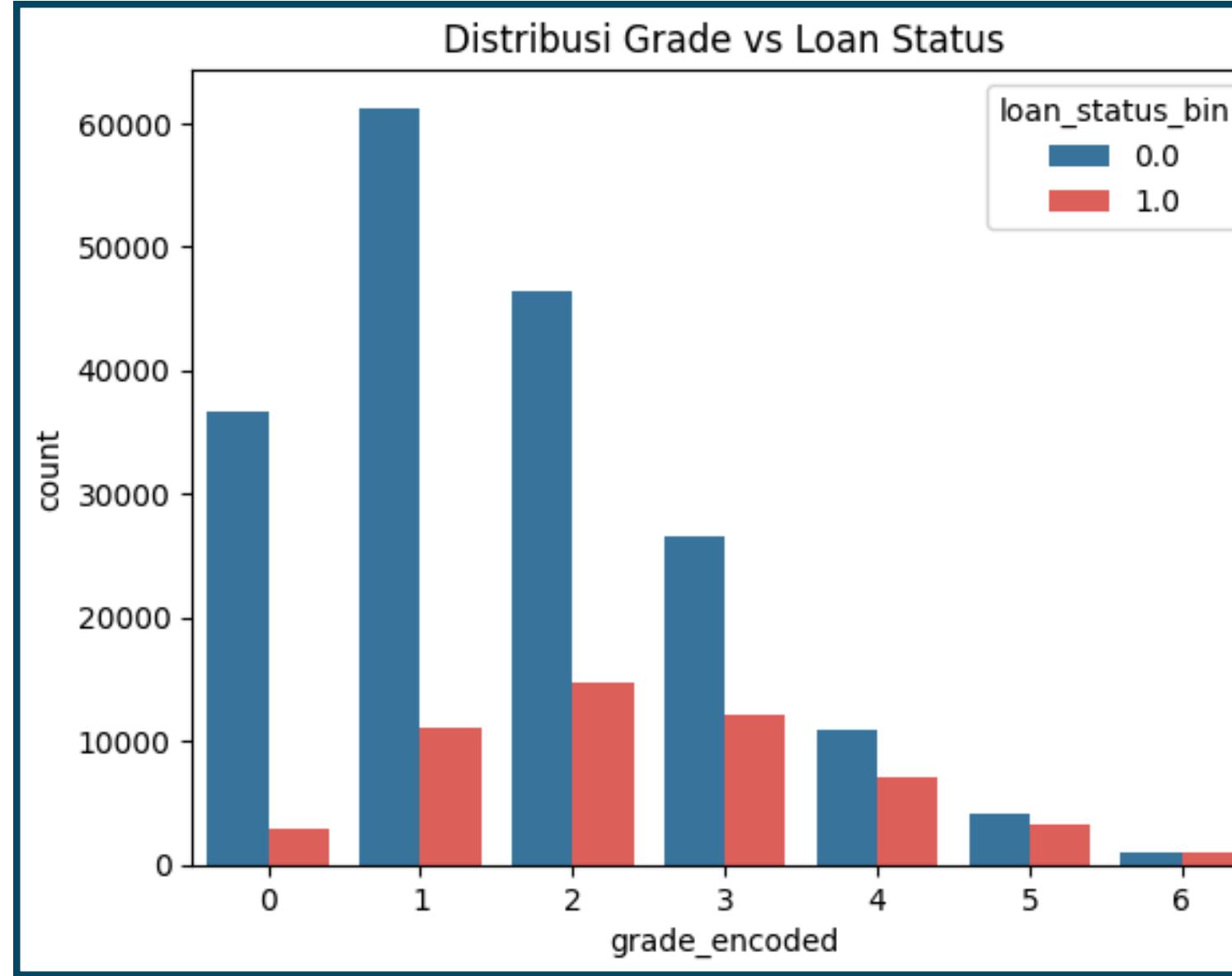


- Mayoritas peminjam berasal dari grade B dan C, menunjukkan risiko menengah dalam data.

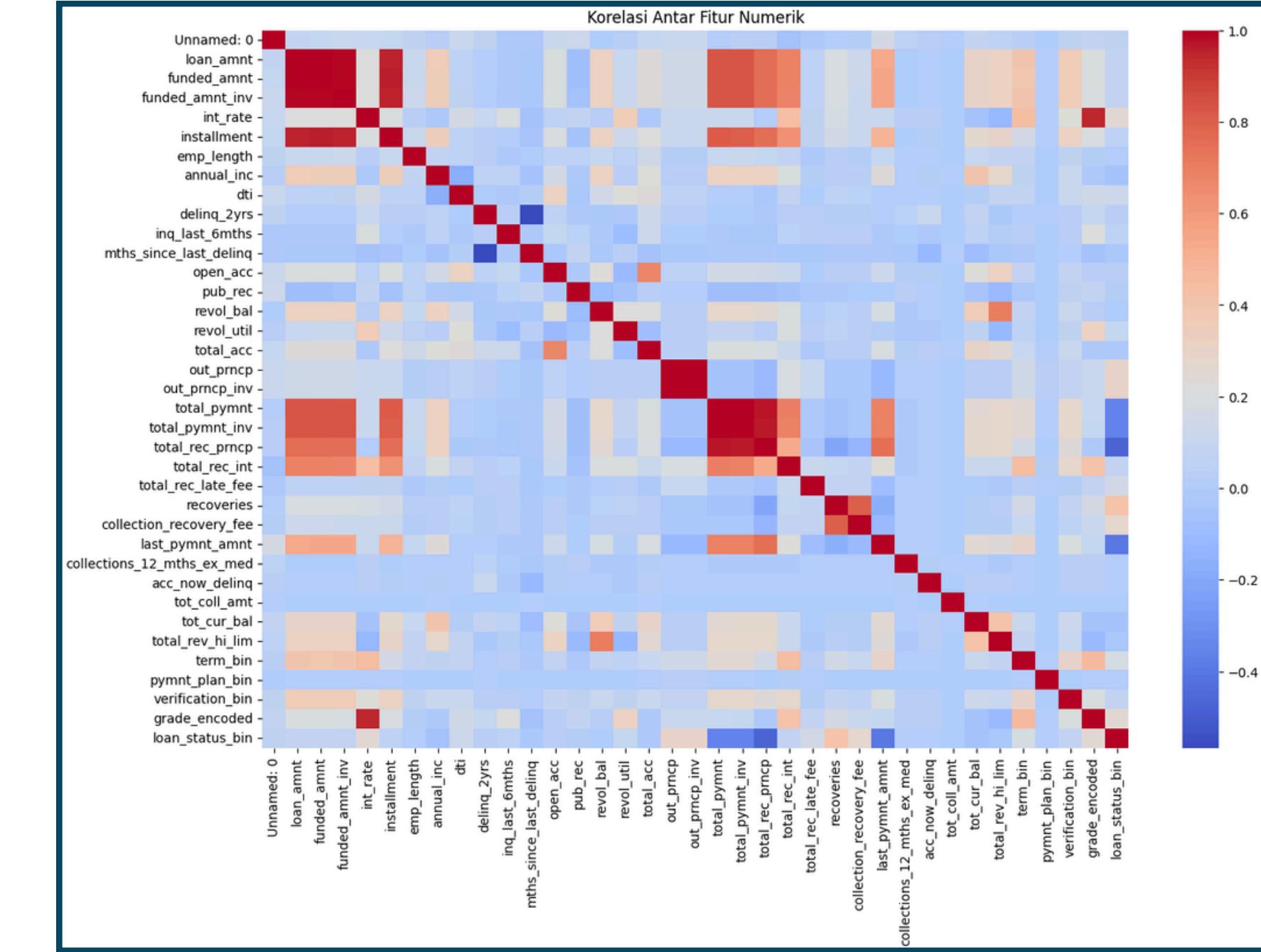
- Distribusi pendapatan (annual_inc) bersifat right-skewed; sebagian besar peminjam memiliki pendapatan tahunan < \$100K dan hanya sedikit yang memiliki pendapatan tinggi

EXPLORATORY DATA ANALYSIS

EDA



- Grade yang lebih rendah (seperti D, E, F) cenderung memiliki proporsi gagal bayar (bad loan) yang lebih tinggi dibandingkan grade A atau B.



- Pendapatan tahunan berkorelasi positif dengan jumlah pinjaman dan cicilan, namun tidak berpengaruh signifikan terhadap kelancaran pembayaran.

MODELING: PEMILIHAN FITUR & DATA SPLIT

FITUR PREDIKSI (X)

Total 24 fitur, terdiri dari:

- Fitur finansial: loan_amnt, installment, int_rate, annual_inc, dti, dll.
- Fitur riwayat kredit: delinq_2yrs, revol_bal, revol_util, total_acc, dll.
- Fitur pemulihan dan pembayaran: recoveries, collection_recovery_fee, last_pymnt_amnt, dll.
- Fitur kategorikal (encoded): term_bin, grade_encoded

TARGET PREDIKSI (Y)

- loan_status_bin → 0 = Good, 1 = Bad

PREPROCESSING

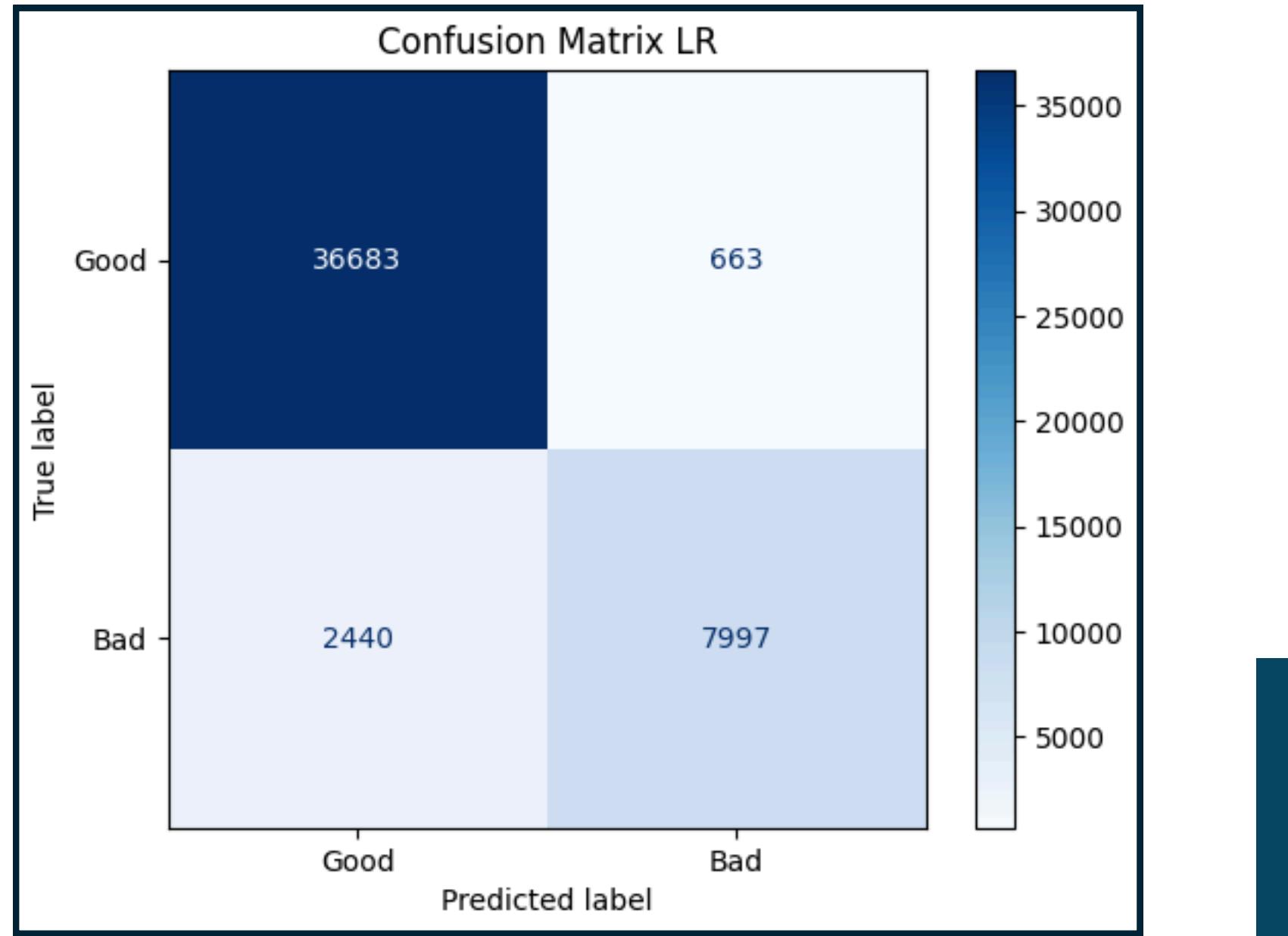
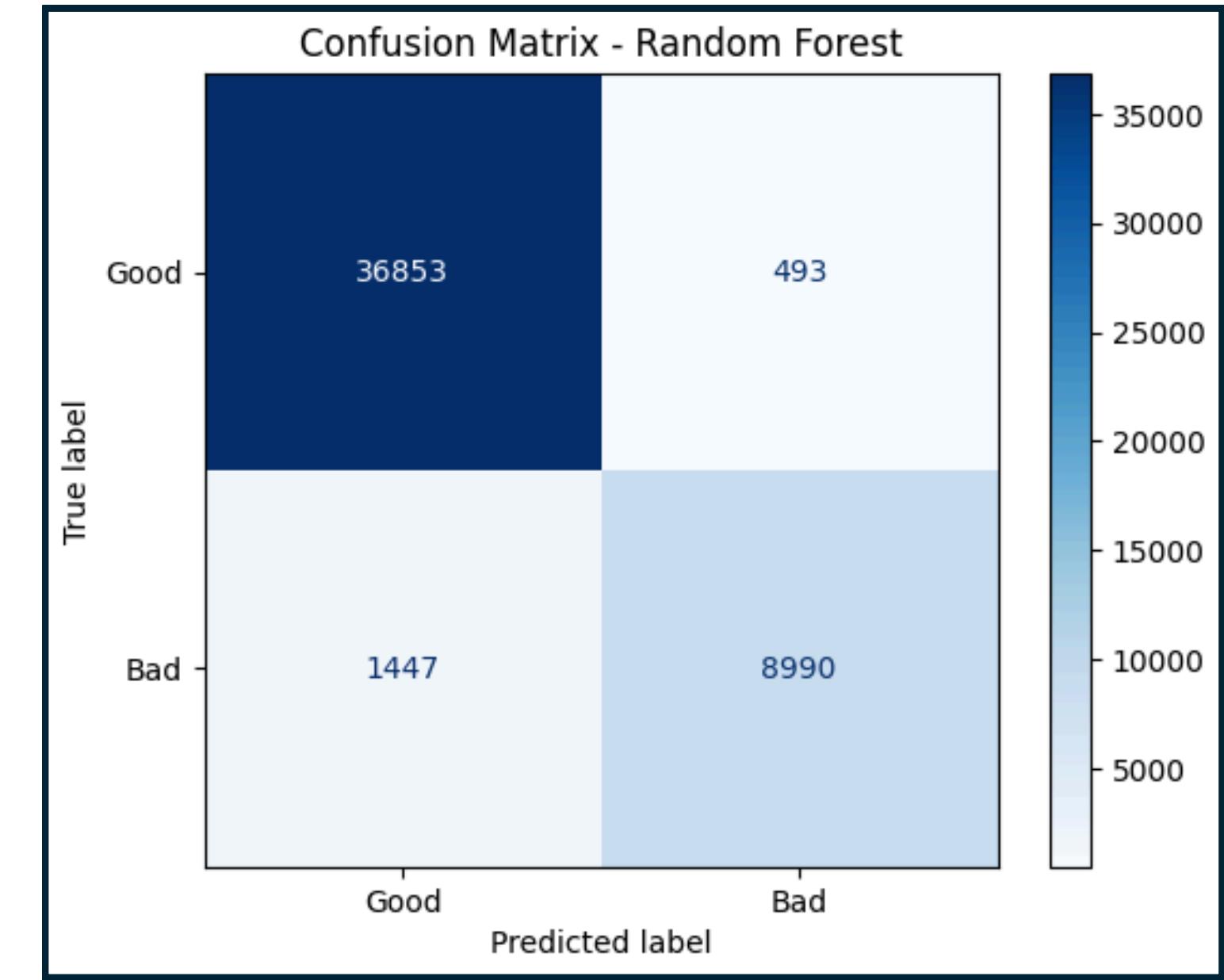
- Scaling: StandardScaler digunakan untuk menstandarkan semua fitur numerik

TRAIN-TEST SPLIT

- Proporsi: 80% data latih, 20% data uji
- Stratify: menjaga proporsi seimbang antara kelas 0 dan 1 pada train dan test



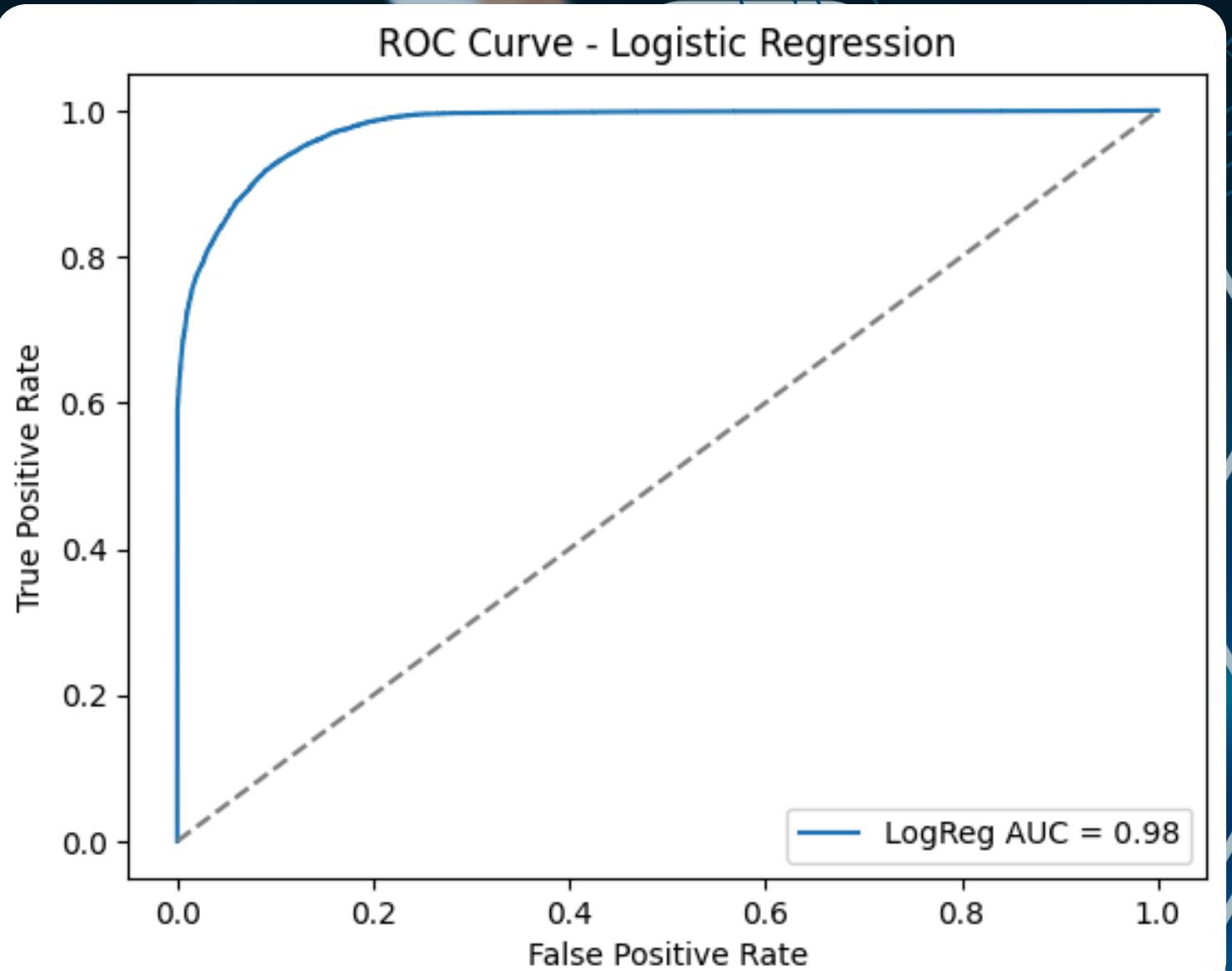
MODEL TRAINING & EVALUATION

**LOGISTIC REGRESSION****RANDOM FOREST**

LOGISTIC REGRESSION

Hasil Model :

- Akurasi: 94%
Model memprediksi status pinjaman dengan ketepatan yang sangat tinggi.
- Precision (Default): 92%
Dari prediksi gagal bayar, 92% memang benar gagal bayar.
- Recall (Default): 77%
Model berhasil mengidentifikasi 77% dari nasabah yang benar-benar gagal bayar.
- F1-score (Default): 84%
Keseimbangan antara kemampuan mendeteksi dan ketepatan klasifikasi default.
- ROC AUC Score: 0.98
Model sangat baik dalam membedakan nasabah berisiko dan tidak berisiko gagal bayar.

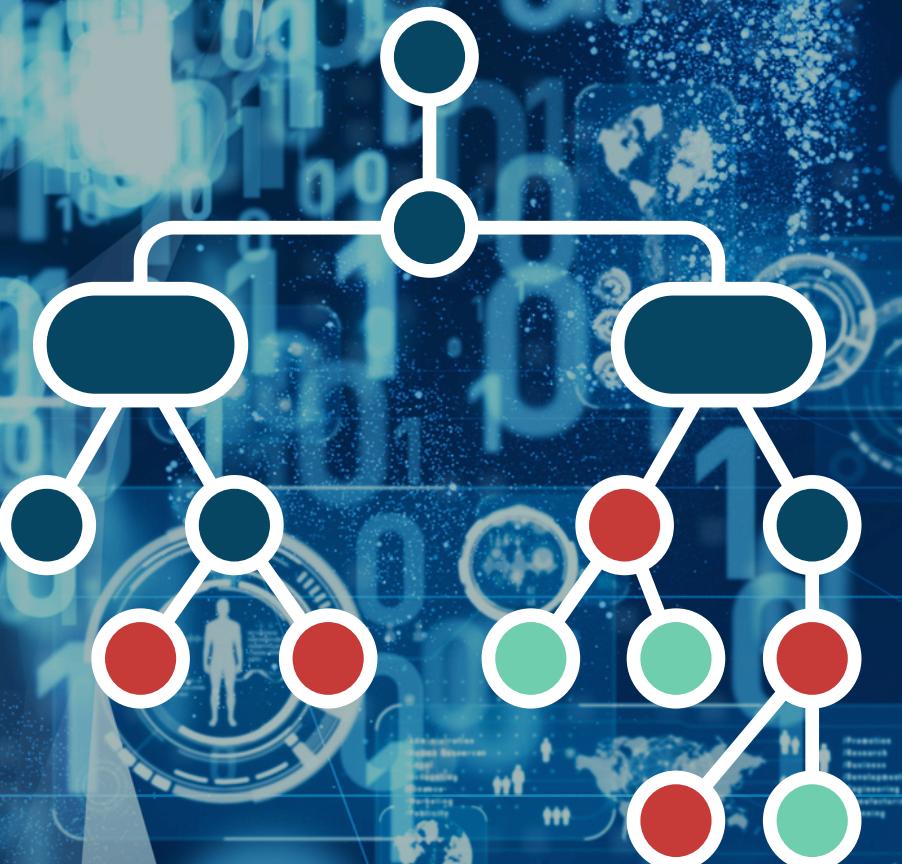


	precision	recall	f1-score	support
Good	0.94	0.98	0.96	37346
Bad	0.92	0.77	0.84	10437
accuracy			0.94	47783
macro avg	0.93	0.87	0.90	47783
weighted avg	0.93	0.94	0.93	47783

RANDOM FOREST

Classification Report:

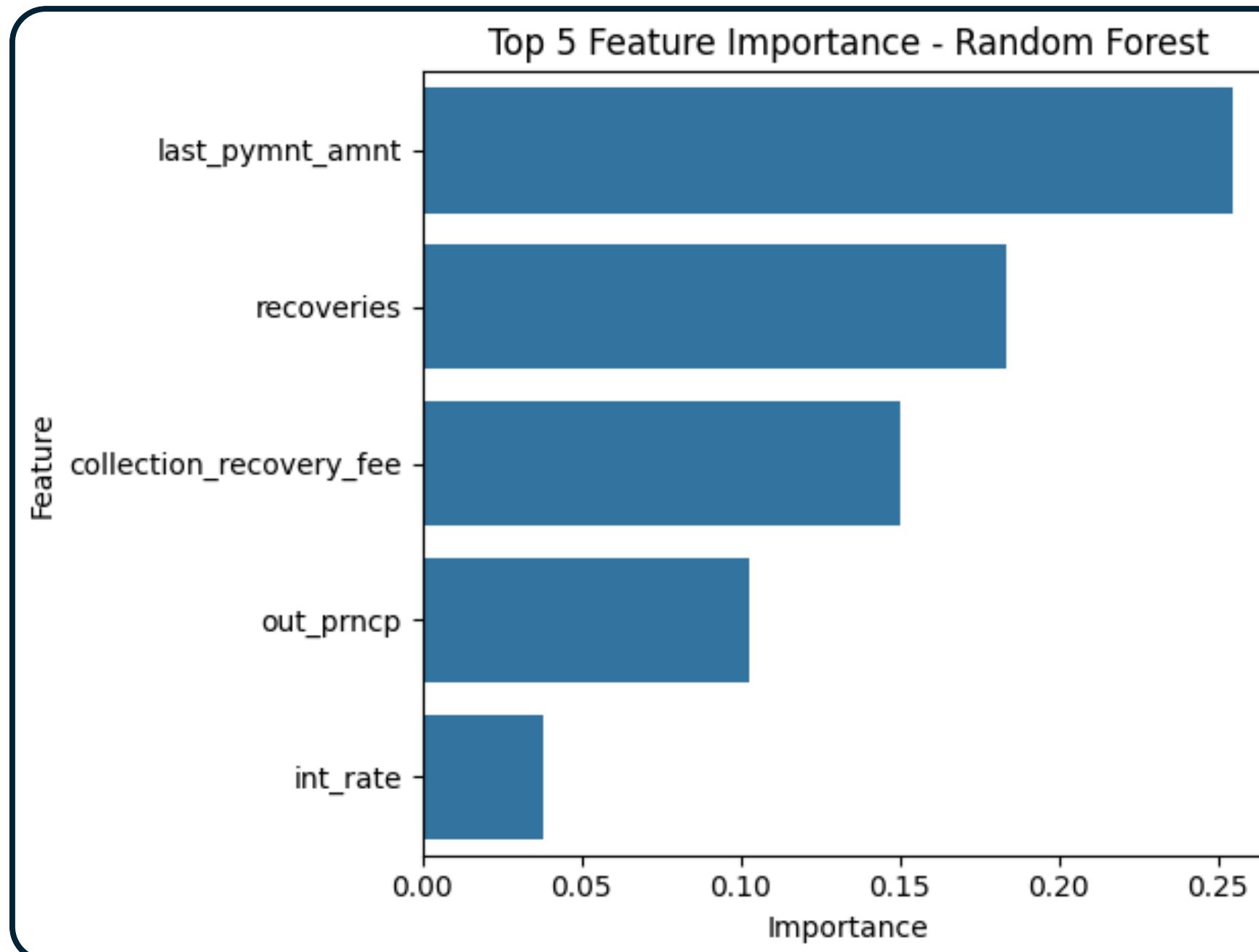
	precision	recall	f1-score	support
Good	0.96	0.99	0.97	37346
Bad	0.95	0.86	0.90	10437
accuracy			0.96	47783
macro avg	0.96	0.92	0.94	47783
weighted avg	0.96	0.96	0.96	47783



Hasil Model :

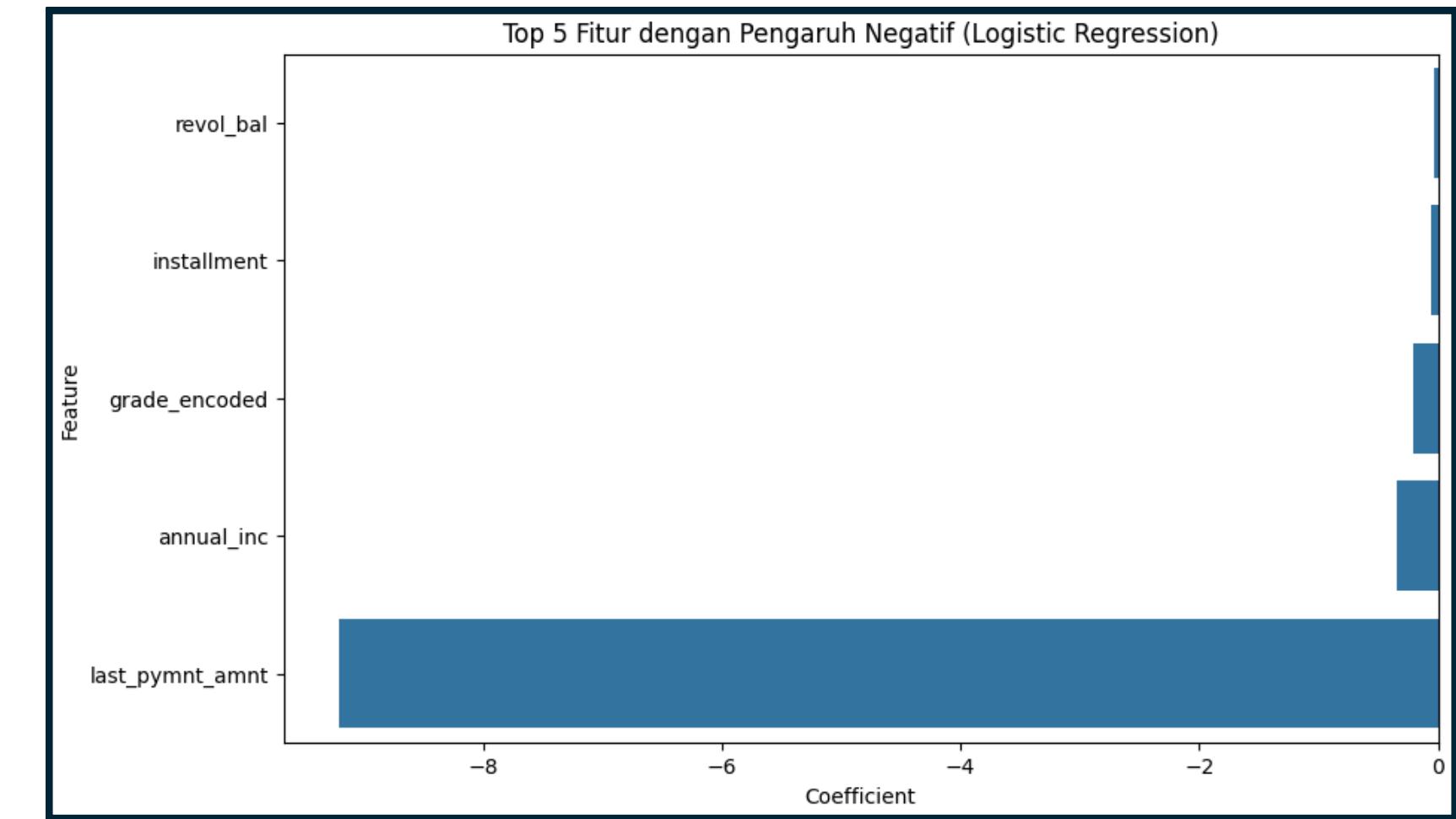
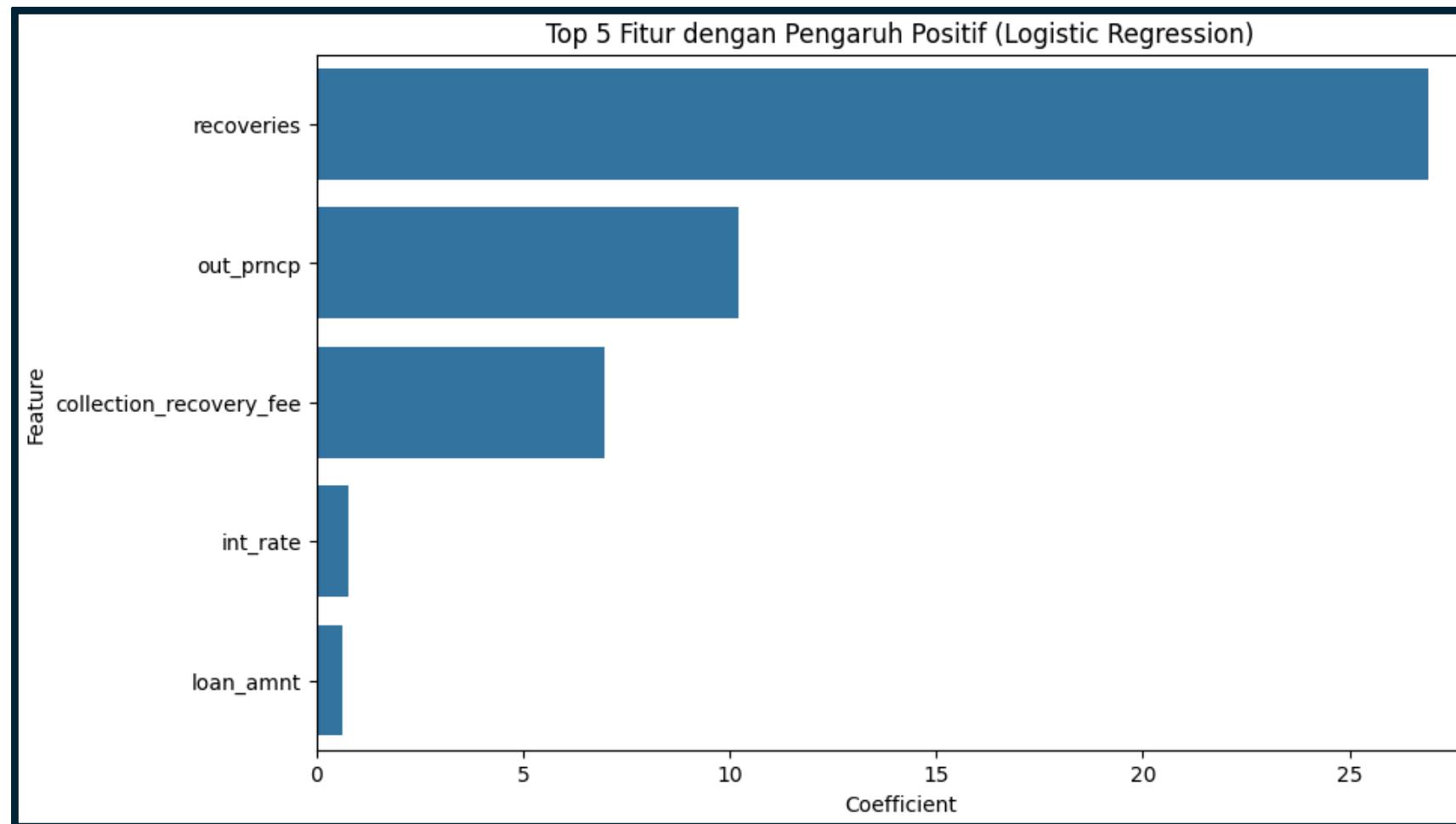
- Akurasi: 96%
Model sangat akurat dalam memprediksi status pinjaman.
- Precision (Default/Bad): 95%
Dari seluruh prediksi gagal bayar, 95% benar-benar gagal bayar.
- Recall (Default/Bad): 86%
Model berhasil mendeteksi 86% dari nasabah yang benar-benar gagal bayar.
- F1-score (Default/Bad): 90%
Performa model seimbang antara presisi dan kemampuan deteksi gagal bayar.

FEATURE IMPORTANCE (RANDOM FOREST)

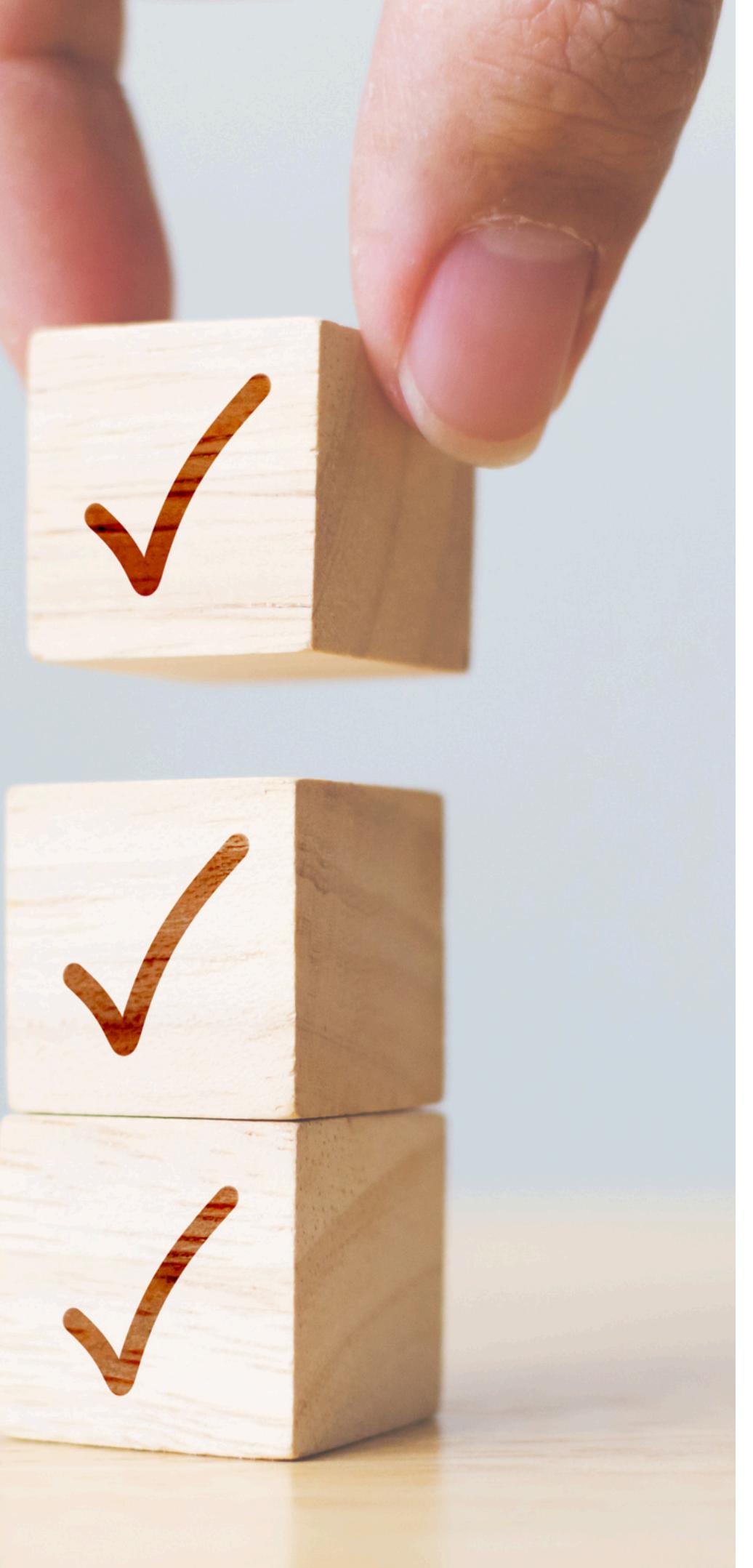


- Grafik di samping menunjukkan 5 fitur paling berpengaruh dalam prediksi risiko kredit menggunakan model Random Forest.
- Fitur `last_pymnt_amnt` (jumlah pembayaran terakhir) memiliki pengaruh terbesar, diikuti oleh `recoveries` (jumlah dana yang berhasil dipulihkan) dan `collection_recovery_fee` (biaya penagihan).
- Ketiga fitur ini berkaitan erat dengan perilaku pembayaran dan efisiensi pemulihan pinjaman, yang menjadi indikator kuat terhadap kemungkinan gagal bayar.

COEFFICIENT INTERPRETATION (LOGISTIC REGRESSION)

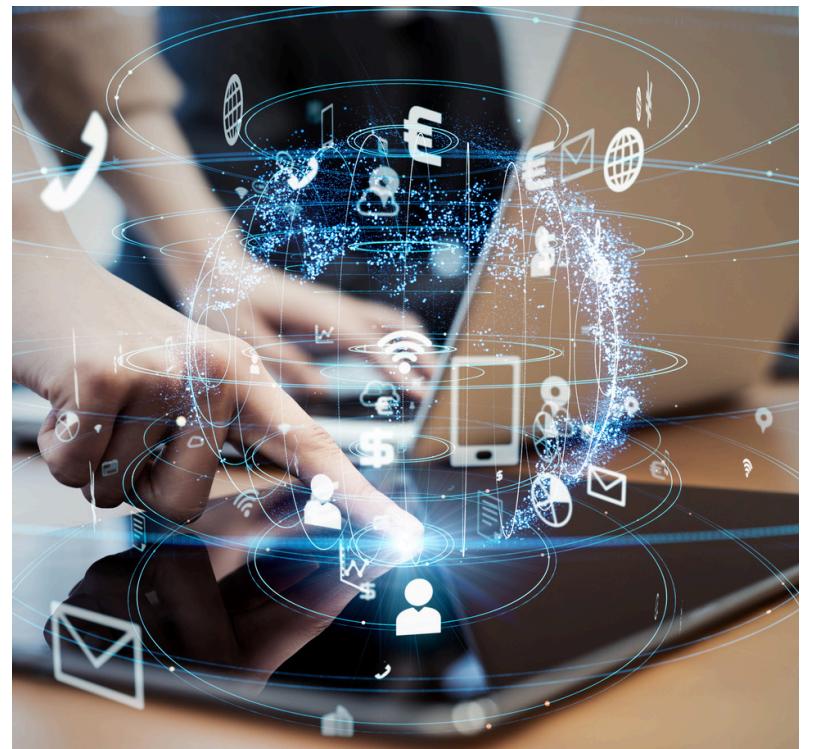


Adanya koefisien positif dan negatif menunjukkan bahwa setiap fitur memiliki arah pengaruh yang berbeda terhadap risiko gagal bayar. Fitur dengan koefisien positif meningkatkan kemungkinan peminjam masuk kategori Bad, sedangkan fitur dengan koefisien negatif cenderung terkait dengan peminjam yang tetap Good. Hal ini membantu memahami faktor-faktor yang memperbesar atau memperkecil risiko kredit secara lebih interpretatif.



KESIMPULAN PROJECT

- Logistic Regression memiliki akurasi 94% dan AUC 0.98 → cocok untuk baseline karena seimbang dan interpretable.
- Random Forest memiliki akurasi 96% → unggul mendekripsi nasabah gagal bayar, cocok untuk deployment.
- Fitur paling berpengaruh: last_pymnt_amnt, recoveries, dan out_prncp, berhubungan dengan riwayat pembayaran & pemulihan.
- Pada Logistic Regression, koefisien positif menaikkan risiko gagal bayar (Bad), negatif menandakan nasabah aman (Good).
- Kedua model menunjukkan performa tinggi dalam klasifikasi risiko kredit, dan dapat diandalkan untuk mendukung keputusan bisnis.



THANK YOU

✉️ ghazaputra99@gmail.com

LinkedIn <https://www.linkedin.com/in/muhammadghazaekaputra>

Github <https://github.com/ghaza-putra>

