CREDIT RISK MODELING HOME CREDIT

Prediksi Gagal Bayar Nasabah

“Menggunakan dataset Home Credit untuk memprediksi nasabah yang berisiko mengalami keterlambatan bayar.”

Grace Gabriella Herald

PROBLEM YANG INGIN DISELESAIKAN

Problem bisnis:

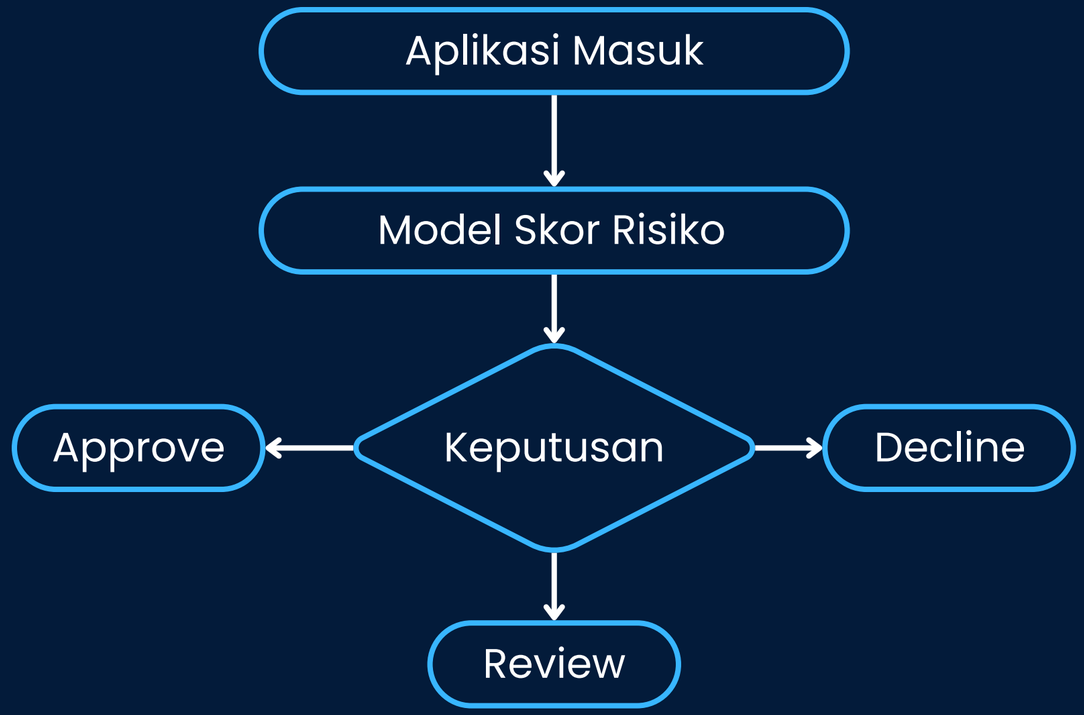
- Perusahaan ingin menurunkan risiko gagal bayar tanpa menolak terlalu banyak nasabah yang sebenarnya layak.

Pertanyaan utama:

- Nasabah seperti apa yang berisiko gagal bayar (TARGET=1)?
- Bagaimana model machine learning bisa membantu mem-filter aplikasi berisiko tinggi?

Tujuan:

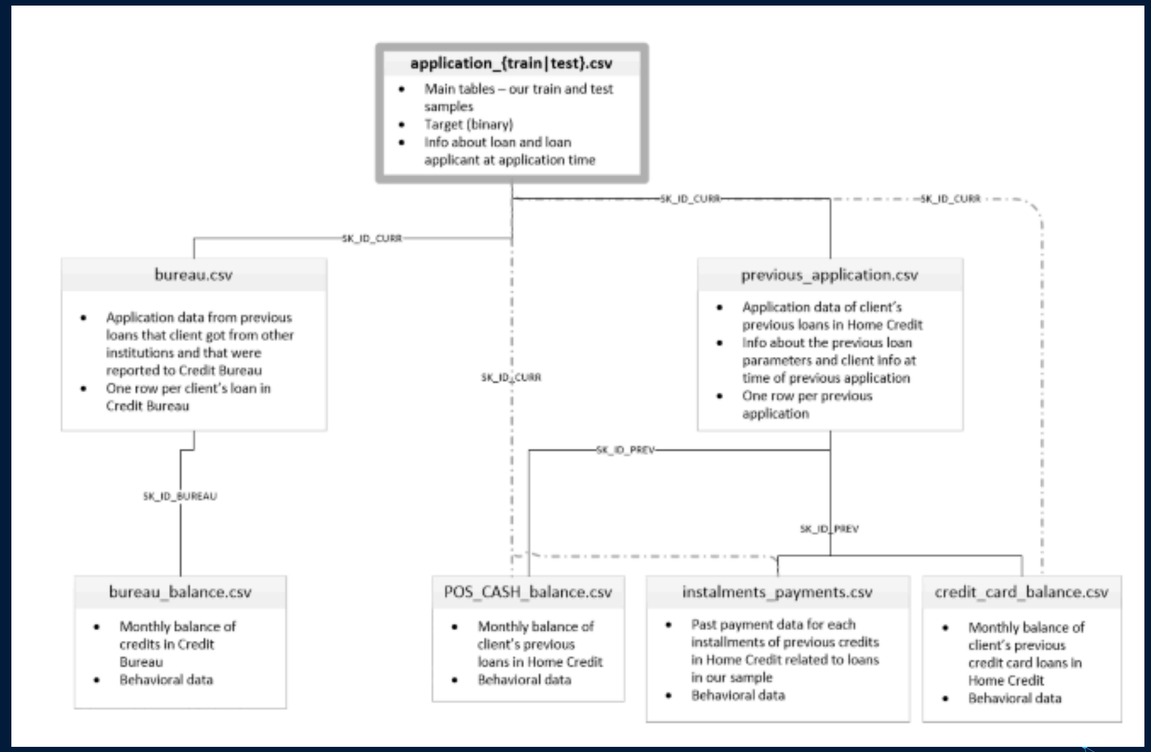
- Membangun model klasifikasi untuk memprediksi probabilitas gagal bayar.
- Memberikan rekomendasi kebijakan kredit & kampanye marketing berdasarkan insight data.



DATASET YANG DIGUNAKAN

Tabel Utama

- application_train.csv
→ data aplikasi + kolom TARGET.
- application_test.csv
→ data aplikasi tanpa TARGET.
- “1 baris = 1 pengajuan kredit nasabah.”



Tabel Pendukung

- Bureau & bureau_balance → riwayat kredit di lembaga lain.
- previous_application → semua aplikasi Home Credit sebelumnya.
- POS_CASH_balance, installments_payments, credit_card_balance → histori pembayaran & saldo bulanan di Home Credit.

DATA PRE-PROCESSING

Data Cleaning

- Menghapus kolom ID dari fitur (SK_ID_CURR dll, hanya dipakai untuk join).
- Mengisi missing value numerik dengan median; kategorik dengan label “Unknown”.

Feature engineering

- AGE_YEARS = -DAYS_BIRTH / 365.
- EMPLOY_YEARS = -DAYS_EMPLOYED / 365.
- CREDIT_INCOME_RATIO = AMT_CREDIT / AMT_INCOME_TOTAL.
- ANNUITY_INCOME_RATIO = AMT_ANNUITY / AMT_INCOME_TOTAL.

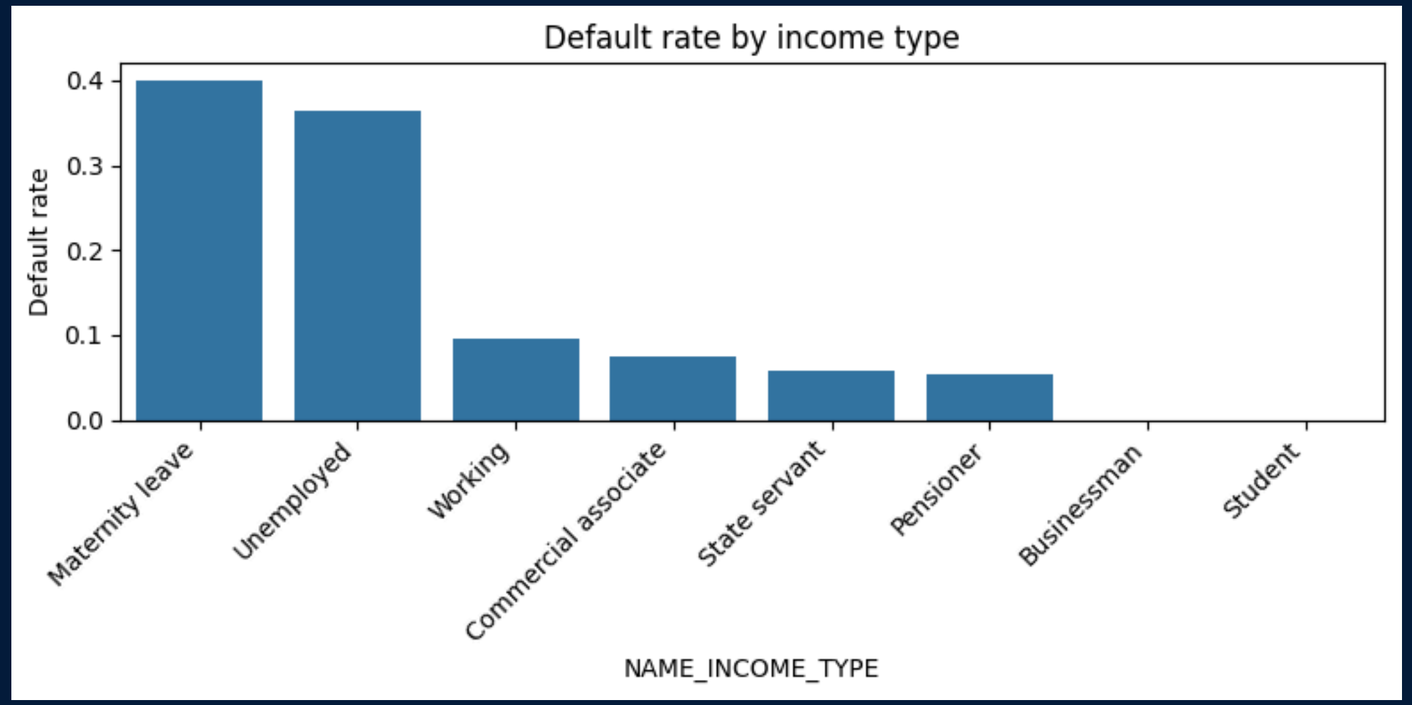
Agregasi tabel lain

- Dari bureau: jumlah kredit sebelumnya, total debt, maksimum DPD.
- Dari previous_application: jumlah aplikasi sebelumnya, approval rate.
- Dari installments_payments: rate telat bayar (LATE_PAYMENT_RATE).

Encoding & scaling

- One-hot encoding untuk fitur kategorik (gender, income type, housing type, dll).
- StandardScaler untuk fitur numerik.

INSIGHT 1 : DEFAULT RATE PER INCOME TYPE



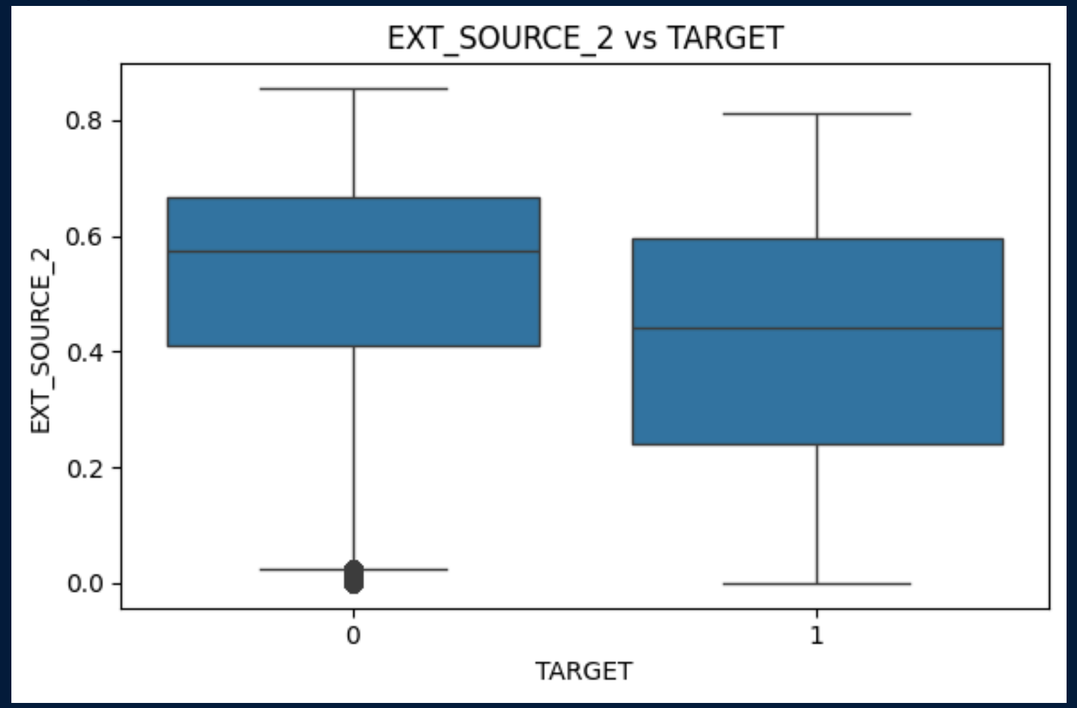
Keterangan

- Segmen dengan risiko gagal bayar tertinggi adalah nasabah dengan pendapatan ‘Maternity leave’ (default rate 40%) dan ‘Unemployed’ (36%), jauh di atas rata-rata segmen lain yang umumnya di bawah 10%
- Nasabah dengan pendapatan stabil seperti ‘Working’, ‘Commercial associate’, ‘State servant’, dan ‘Pensioner’ punya default rate jauh lebih rendah (sekitar 5–10%).
- Kategori ‘Businessman’ dan ‘Student’ pada data training tidak menunjukkan kasus gagal bayar, kemungkinan karena jumlah sampelnya sangat sedikit sehingga perlu dianalisis hati-hati.

Action yang disarankan:

- Perketat policy kredit untuk segmen Maternity leave & Unemployed (limit lebih kecil, butuh jaminan, atau proses review lebih ketat).
- Fokus kampanye akuisisi ke segmen berpenghasilan stabil seperti Working, Commercial associate, State servant, dan Pensioner untuk menjaga kualitas portofolio.

INSIGHT 2 : SKOR EKSTERNAL & RISIKO



Keterangan

- Nasabah yang tidak bermasalah bayar (TARGET=0) cenderung memiliki nilai EXT_SOURCE_2 lebih tinggi (median sekitar 0,55–0,60), sedangkan nasabah bermasalah (TARGET=1) punya skor lebih rendah (median sekitar 0,40–0,45).
- Sebaran skor eksternal untuk TARGET=1 juga lebih ‘turun ke bawah’, artinya banyak nasabah default berada di rentang skor rendah.

Action yang disarankan:

- Gunakan EXT_SOURCE_2 sebagai salah satu fitur utama dalam credit scoring: skor rendah → kategori risiko tinggi.
- Tetapkan ambang batas skor eksternal: di bawah threshold tertentu, aplikasi perlu manual review atau syarat tambahan (jaminan, down payment lebih besar).

PENDEKATAN MACHINE LEARNING

Tipe masalah:

- Binary classification: memprediksi TARGET (0 = good, 1 = default).

Data split:

- Train–test split 80:20 dengan stratified sampling agar distribusi TARGET seimbang.

Algoritma yang dicoba:

- Logistic Regression (baseline, interpretable).
- Random Forest (model non-linear dengan performa lebih baik).

Penanganan class imbalance:

- Gunakan class_weight='balanced' pada model untuk memberi bobot lebih besar ke kelas default.

```
log_reg = LogisticRegression(
max_iter=1000,
class_weight='balanced',
n_jobs=-1
)

log_reg.fit(X_train_scaled, y_train)

y_pred_lr = log_reg.predict(X_test_scaled)
y_prob_lr = log_reg.predict_proba(X_test_scaled)[:,1]
```

```
rf = RandomForestClassifier(
n_estimators=200,
max_depth=10,
n_jobs=-1,
random_state=42,
class_weight='balanced'
)

rf.fit(X_train, y_train) # RF tidak wajib di-scale

y_pred_rf = rf.predict(X_test)
y_prob_rf = rf.predict_proba(X_test)[:,1]
```

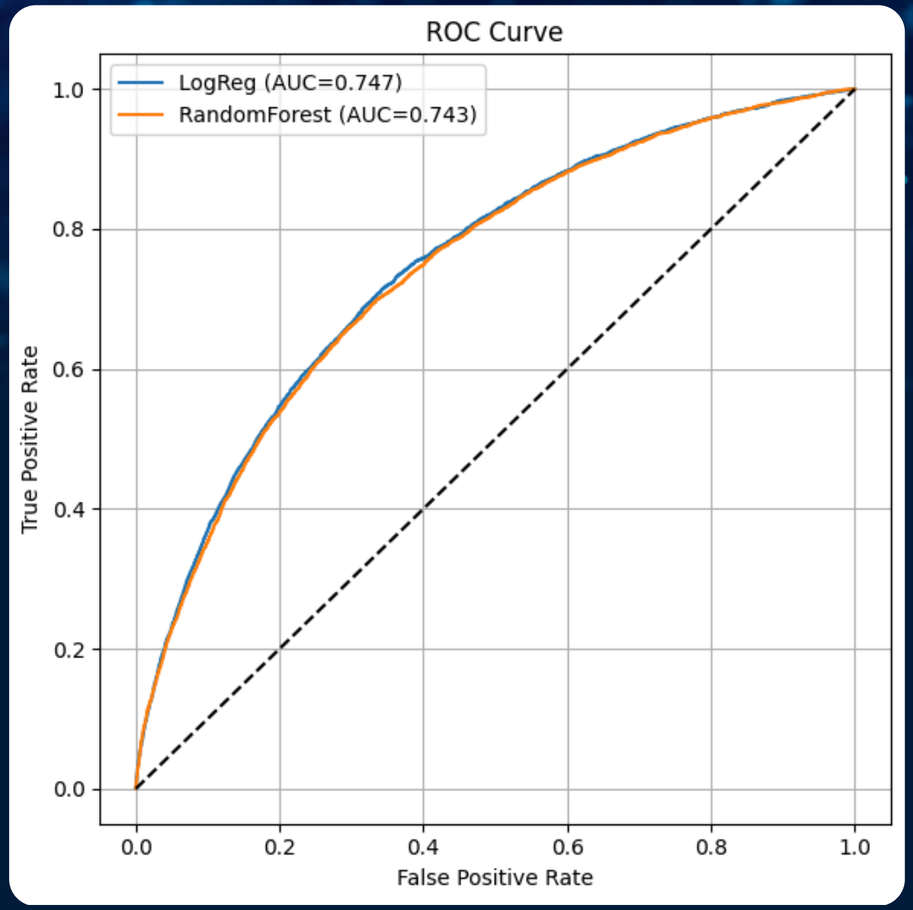
PERFORMANCE MODEL

Tabel Perbandingan Performance Model yang digunakan

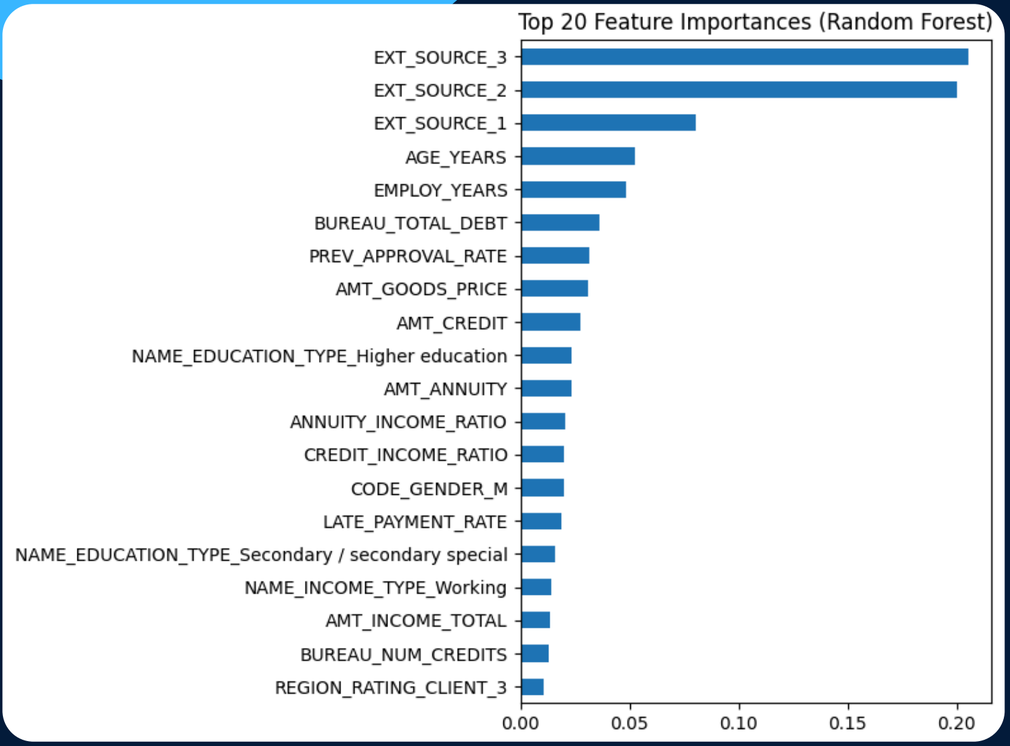
Model	AUC-ROC	Recall (TARGET=1)	Precision (TARGET=1)
Logistic Reg	0.7474	0.68 (68%)	0.16 (16%)
Random Forest	0.7429	0.63 (63%)	0.17 (17%)

Logistic Regression memberikan trade-off yang sedikit lebih menguntungkan untuk mendeteksi nasabah gagal bayar (recall 68% dengan AUC tertinggi), sedangkan Random Forest menawarkan peningkatan kecil pada precision namun dengan recall yang sedikit lebih rendah. Pada konteks mitigasi risiko kredit, model dengan recall lebih tinggi (Logistic Regression) lebih menarik karena mengurangi jumlah nasabah bermasalah yang terlewat.

ROC curve dua model (LogReg & RandomForest)



FITUR PALING BERPENGARUH



Pemodelan menunjukkan bahwa skor eksternal, faktor usia, riwayat pinjaman, dan perilaku pembayaran adalah kunci dalam membedakan nasabah berisiko tinggi dan rendah. Fitur-fitur ini layak dijadikan dasar utama kebijakan credit scoring dan monitoring risiko.

Fitur paling berpengaruh dalam prediksi gagal bayar adalah skor eksternal:

- EXT_SOURCE_3, EXT_SOURCE_2, EXT_SOURCE_1

(skor kredit dari pihak ketiga sangat menentukan risiko nasabah)

Fitur demografi dan keuangan juga penting:

- AGE_YEARS, EMPLOY_YEARS (umur & lama bekerja—nasabah lebih tua/berpengalaman cenderung lebih aman)
- BUREAU_TOTAL_DEBT (total utang di lembaga lain)
- PREV_APPROVAL_RATE (persentase aplikasi sebelumnya yang disetujui)
- AMT_GOODS_PRICE, AMT_CREDIT (besar pinjaman & harga barang)

Fitur perilaku & riwayat pembayaran:

- LATE_PAYMENT_RATE (proporsi telat bayar di pembayaran cicilan sebelumnya

Fitur pendidikan & jenis income mendukung pemisahan risiko:

- NAME_EDUCATION_TYPE_Higher education, NAME_INCOME_TYPE_Working
- Biasanya nasabah berpendidikan/income tetap lebih aman.

BUSINESS RECOMMENDATION & REFERENSI

1.Business Recommendation

- Gunakan Logistic Regression sebagai model utama credit scoring karena memberikan AUC tertinggi (0,7474) dan recall lebih tinggi untuk nasabah default (68%) dibanding Random Forest.
- Terapkan segmen risiko berdasarkan probabilitas output model:
  - Low risk: probabilitas default rendah → auto-approve dengan limit normal/lebih tinggi.
  - Medium risk: probabilitas sedang → butuh manual review & dokumen tambahan.
  - High risk: probabilitas tinggi → limit diturunkan signifikan atau aplikasi direject.
- Jadikan EXT_SOURCE_1/2/3, umur, lama bekerja, rasio kredit terhadap pendapatan, total utang di bureau, dan LATE_PAYMENT_RATE sebagai fitur utama dalam aturan bisnis (misal: batasi plafon jika rasio kredit/income tinggi atau skor eksternal terlalu rendah).
- Fokus akuisisi dan promo ke segmen yang menurut model relatif aman: income type 'Working / Commercial associate / State servant / Pensioner' dan skor eksternal tinggi, untuk menjaga kualitas portofolio.
- Lakukan monitoring berkala (misalnya tiap kuartal) terhadap AUC, recall default, dan default rate per segmen; lakukan retraining model ketika terdapat data baru atau pola risiko berubah.

2.Referensi & Implementasi

- Dataset: Home Credit Default Risk
- Konsep evaluasi model: dokumentasi Precision–Recall, ROC–AUC, dan metrik klasifikasi dari Google ML Crash Course / scikit-learn.
- Implementasi: notebook Python & script preprocessing + modeling disimpan di GitHub pribadi

3. Kesimpulan

- Model ini dapat diintegrasikan ke dalam pipeline onboarding kredit sebagai modul scoring otomatis, dengan tetap memberikan ruang bagi tim risk/credit analyst untuk override berdasarkan kebijakan perusahaan

Dokumentasi di github > https://github.com/EadLim/Rakamin_HomeCredit