HOME CREDIT SCORECARD MODEL

Project-Based Internship Rakamin x Home Credit Indonesia

Muhammad Tri Wahyudi

Project Objective

Latar Belakang : Latar belakang project ini adalah perlunya sistem yang dapat membantu perusahaan dalam menilai calon peminjam.

Pinjaman yang gagal bayar bisa menyebabkan kerugian finansial. Metode penilaian tradisional sering subjektif dan memakan waktu. Maka dari itu, diperlukan solusi berbasis data dan machine learning untuk mengukur risiko borrower secara lebih objektif dan cepat.

Problem:

- Bagaimana memprediksi borrower yang berisiko gagal bayar?
- Data memiliki ratusan fitur dan multi-table (kompleks)
- Target default sangat kecil (imbalanced)
- Model harus mampu generalisasi dengan baik

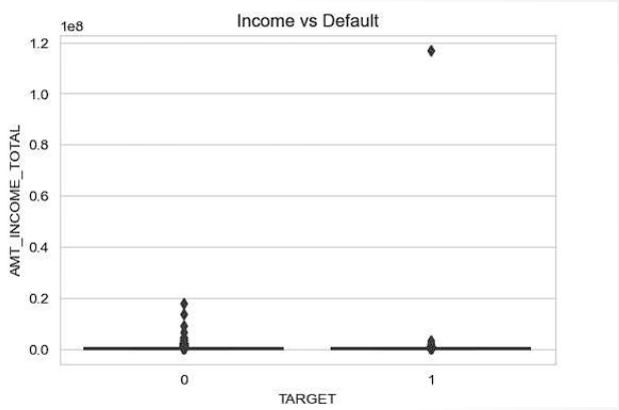
Dataset Summary

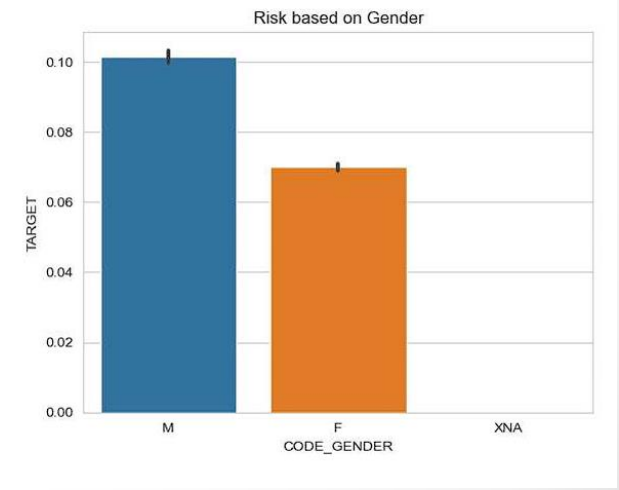
Dataset: application_train.csv (307.511 baris, 122 kolom), application_test.csv (48.744 baris, 121 kolom), plus dataset pendukung seperti bureau.csv (1.716.428 baris, 17 kolom), previous_application.csv (1.670.214 baris, 37 kolom), dll.

Masalah bisnis: Data imbalanced (non-default ~91.92%, default ~8.08%), perlu model akurat untuk approval pinjaman.

EXPLORATORY DATA ANALYSIS

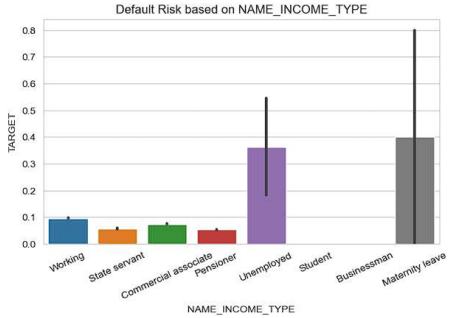
- Banyak fitur skewed (e.g., AMT_INCOME_TOTAL tinggi pada non-default).
- Hubungan dengan TARGET: Income default lebih rendah (boxplot menunjukkan median ~150.000 vs non-default ~168.000).
- Kategorik vs TARGET: Laki-laki risiko (~0.10), perempuan ~0.07 (dari 202.448 F, 105.059 M).

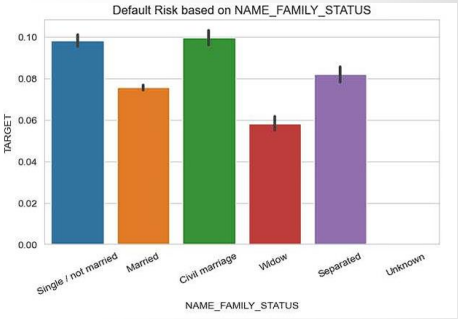


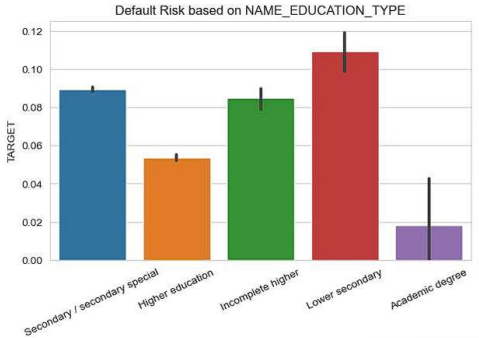


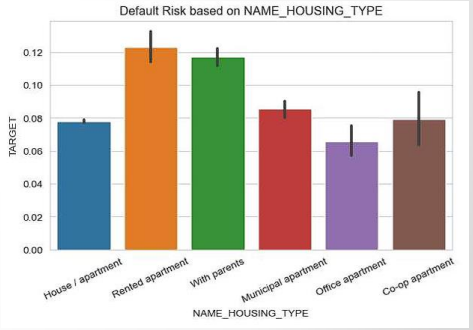
EXPLORATORY DATA ANALYSIS

- Orang yang sedang cuti melahirkan paling berisiko (~40% gagal bayar), diikuti pengangguran (~36%). Yang pekerja biasa jauh lebih aman (~8%).
- Pendidikan rendah (lower secondary) risiko tertinggi (~11%), sedangkan pendidikan tinggi (higher education) paling rendah (~5%).
- civil marriage (~10% risiko), lebih tinggi daripada yang sudah menikah resmi (~7%).
- Yang sewa apartemen (rented) paling risky (~11%), tinggal sama orang tua (~9%), sedangkan punya rumah/apartemen sendiri lebih aman (~8%).









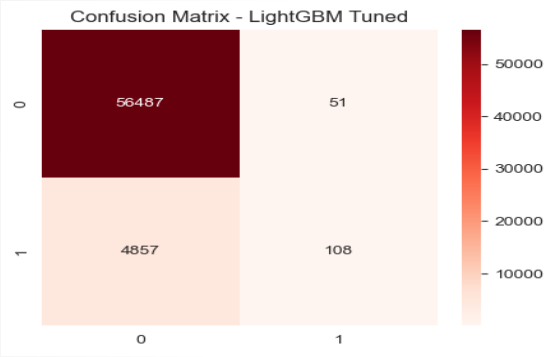
Model Performance

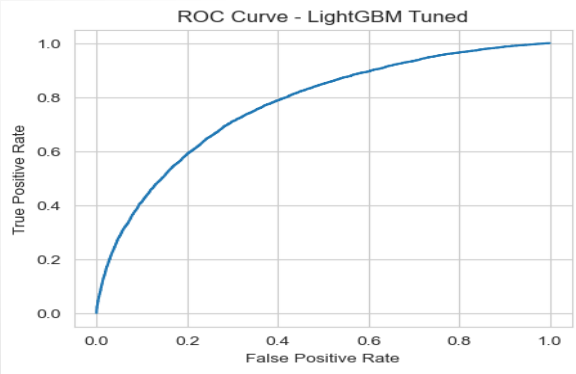
- LightGBM memberikan AUC tertinggi (0.773) → performa keseluruhan terbaik
- Gradient Boosting unggul pada Recall Default → lebih sensitif terhadap borrower default
- Logistic Regression baseline sederhana namun kurang menangkap pola kompleks
- Random Forest cukup baik namun belum optimal dibanding boosting model
- LightGBM dipilih sebagai model final lalu dilakukan Hyperparameter Tuning

Model	AUC	Recall Default
Logistic Regression	0.64	0.06
Random Forest	0.74	0.55
Gradient Boosting	0.76	0.80
LightGBM Tuned (Final)	**0.773**	0.03

Alasan LightGBM dipilih:
✓ Cepat pada dataset besar
✓ Efektif pada data dengan banyak fitur
✓ Mampu menangani missing value & non-linear pattern

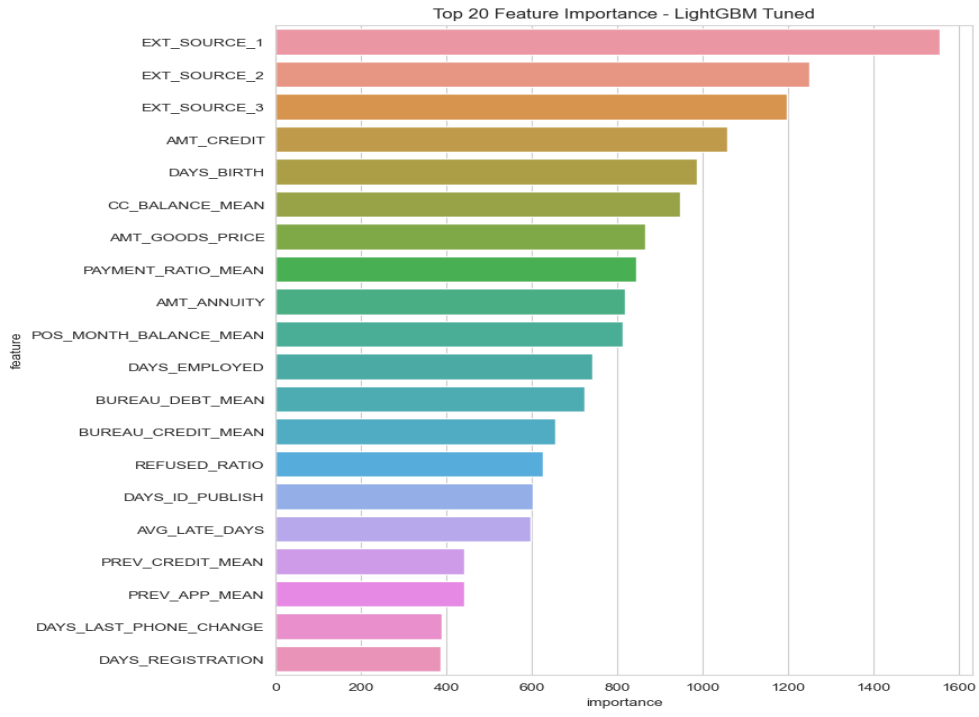
Evaluation Result – LightGBM





- LightGBM menghasilkan performa terbaik secara keseluruhan dengan AUC 0.773
- ROC Curve naik mendekati sudut kiri atas → menandakan model cukup kuat
- Namun confusion matrix menunjukkan banyak default tidak terdeteksi (4857 predicted non-default padahal default)
- Artinya model masih lebih konservatif dan cenderung aman dalam memprediksi
- Model cocok sebagai baseline credit scoring → next step bisa naik recall dengan threshold tuning

Feature Importance



- EXT_SOURCE_1/2/3 → faktor paling kuat menentukan risiko default
- Semakin rendah nilai EXT_SOURCE → semakin tinggi kemungkinan gagal bayar
- Fitur finansial seperti AMT_CREDIT dan AMT_ANNUITY juga berkontribusi signifikan
- DAYS_BIRTH → nasabah lebih muda cenderung memiliki risiko lebih tinggi
- Riwayat pembayaran & penggunaan kartu kredit (CC_BALANCE, PAYMENT_RATIO) berpengaruh pada stabilitas finansial
- Previous loan refused → indikasi risiko gagal bayar sebelumnya

BUSINESS INSIGHT

Insight 1 – External Score (EXT_SOURCE_1/2/3) : Skor eksternal jadi faktor paling kuat dalam prediksi risiko. Semakin tinggi skor → risiko default makin rendah.

Rekomendasi: Prioritaskan approval & beri limit lebih besar pada peminjam dengan skor tinggi untuk memperbesar portofolio aman.

Insight 2 – Besaran Kredit & Anuitas (AMT_CREDIT / AMT_ANNUITY) : Pinjaman dan cicilan besar meningkatkan potensi gagal bayar.

Rekomendasi: Batasi plafon kredit pada calon berisiko tinggi & cek rasio cicilan terhadap income untuk menekan keterlambatan.

Insight 3 – Riwayat Pengajuan Ditolak (REFUSED_RATIO & PREV_APP_MEAN) : Sering ditolak sebelumnya mengindikasikan risiko lebih tinggi.

Rekomendasi: Lakukan verifikasi tambahan dan monitoring ketat pada kategori high-risk agar default bisa dicegah lebih awal.

Insight 4 – Usia Nasabah (DAYS_BIRTH) : Nasabah lebih muda cenderung memiliki disiplin finansial lebih rendah.

Rekomendasi: Terapkan batas income minimum pada borrower muda & dorong penggunaan auto-debit untuk menekan keterlambatan pembayaran.

THANK YOU

⚜ **Full Project & Notebook :**

Click Here to open GitHub Repository