



Home Credit Scorecard Model

by Bernadetta Quinta

<https://github.com/Bernadettaquin/Homecredit-scorecardmodel>

The screenshot shows a GitHub repository page for 'Home Credit Scorecard Model'. The repository was created by Bernadettaquin and has one contributor. It has zero issues, stars, and forks. The repository aims to predict credit scores to ensure that customers with the ability to repay are not rejected during the loan application process. This also allows loans to be granted with a defined timeline.

Bernadettaquin/Homecredit-scorecardmodel

Home Credit aims to predict credit scores to ensure that customers with the ability to repay are not rejected during...

1 Contributor 0 Issues 0 Stars 0 Forks

Bernadettaquin/Homecredit-scorecardmodel: Home Credit aims to predict credit scores to ensure that...

Home Credit aims to predict credit scores to ensure that customers with the ability to repay are not rejected during the loan application process. This also allows loans to be granted with a define...

GitHub



Data Overview

Periode: -

Populasi: 47.074 data nasabah dengan 122 fitur

Testing: 20% → 9.415 data

Modeling: 80% → 37.659 data

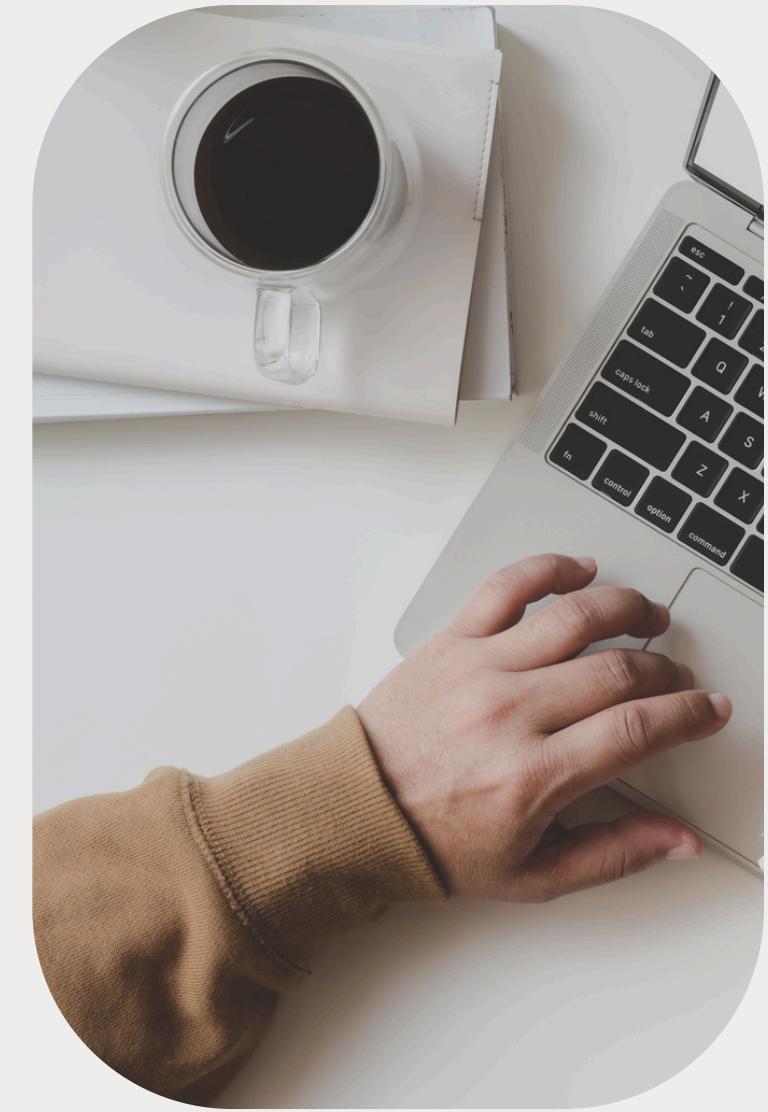
Proporsi data: 80:20

List of Variable: EXT_SOURCE_3,
EXT_SOURCE_2,EXT_SOURCE_1,EMPLOYED_YEARS,NAME_EDUCATION_TYPE,
CODE_GENDER, AMT_GOODS_PRICE, AMT_CREDIT, OWN_CAR_AGE, AMT_ANNUITY,
DAYS_LAST_PHONE_CHANGE, AGE_YEARS, AMT_REQ_CREDIT_BUREAU_YEAR,
DAYS_ID_PUBLISH, DAYS_BIRTH, NAME_CONTRACT_TYPE, DAYS_EMPLOYED,
AMT_INCOME_TOTAL, DEF_60_CNT_SOCIAL_CIRCLE, FLAG_DOCUMENT_3



Problem Research

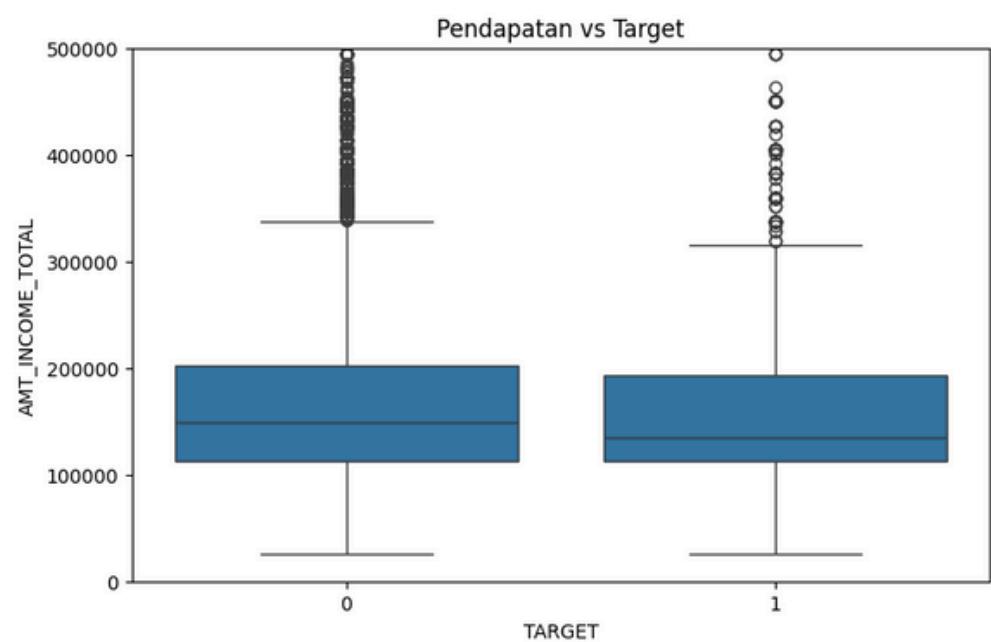
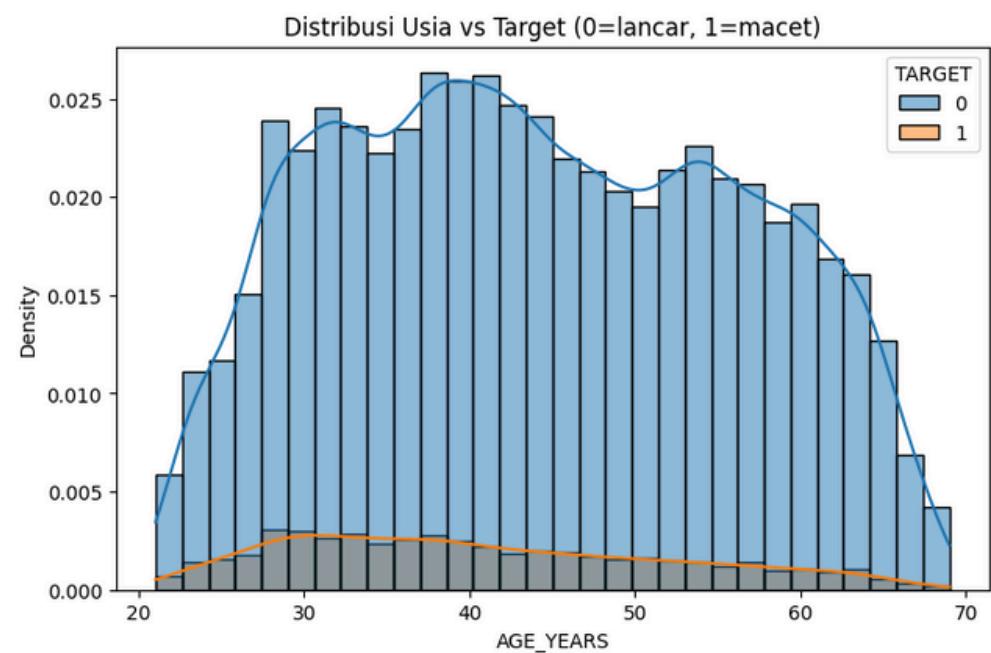
Perusahaan Home Credit ingin prediksi skor kredit. Dengan itu dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman, dan pinjaman data diberikan dengan principal, maturity, dan repayment calendar.





Data Preprocessing

- Cek duplikasi: ditemukan 0 duplikat, jadi aman.
- Handling missing value: beberapa variabel seperti DAYS_EMPLOYED yang bernilai anomali (365243) sudah diganti dengan NaN.
- Feature engineering dibuat variabel baru, yaitu:
 1. AGE_YEARS = usia nasabah dari DAYS_BIRTH.
 2. EMPLOYED_YEARS = lama bekerja dari DAYS_EMPLOYED.
- Encoding fitur kategorikal dilakukan Label Encoding untuk variabel bertipe object (contoh: jenis kelamin, tipe kontrak).
- Split dataset dipisahkan menjadi training dan validation/testing.



1 Distribusi Usia (AGE_YEARS) vs Target

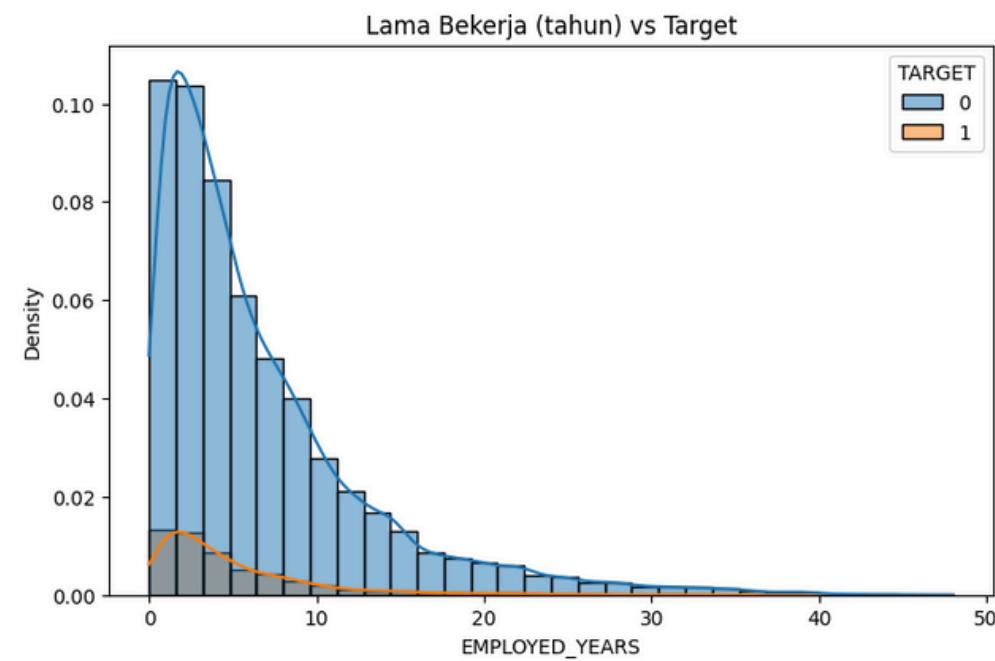
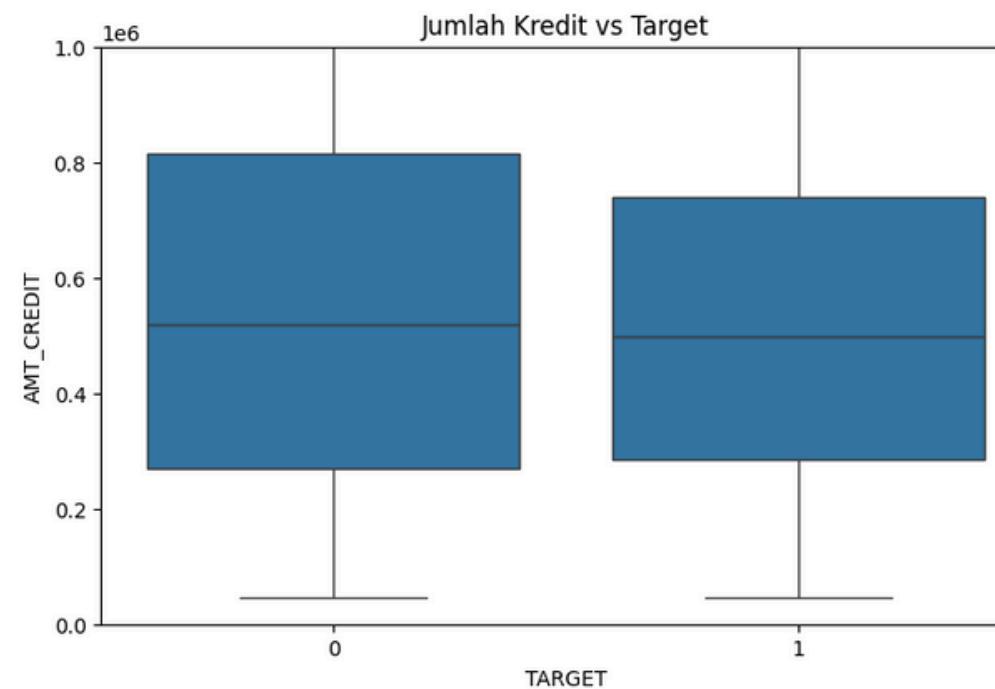
Mayoritas peminjam berada pada rentang 30–50 tahun. Nasabah yang lebih muda (<30 tahun) dan lebih tua (>60 tahun) cenderung memiliki risiko gagal bayar (TARGET=1) lebih tinggi.

- Segmen usia produktif (30–50 tahun) lebih layak untuk prioritas penawaran kredit, sementara kelompok risiko tinggi perlu diberi monitoring lebih ketat (misalnya dengan limit kredit lebih kecil atau syarat tambahan).

2 Pendapatan (AMT_INCOME_TOTAL) vs Target

Distribusi pendapatan nasabah lancar (TARGET=0) dan macet (TARGET=1) cukup mirip, tetapi terlihat outlier pada nasabah berpendapatan sangat tinggi. Tidak selalu pendapatan tinggi menjamin pelunasan lancar.

- Perlu menambahkan variabel stabilitas pekerjaan (lama bekerja, jenis pekerjaan) selain hanya income. Kredit sebaiknya tidak hanya berbasis pendapatan, tetapi juga pada riwayat pekerjaan dan faktor eksternal.



3

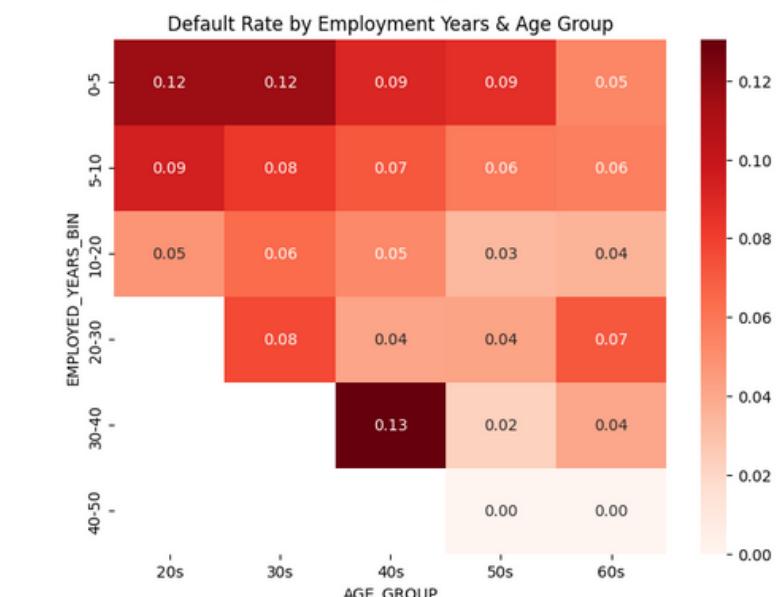
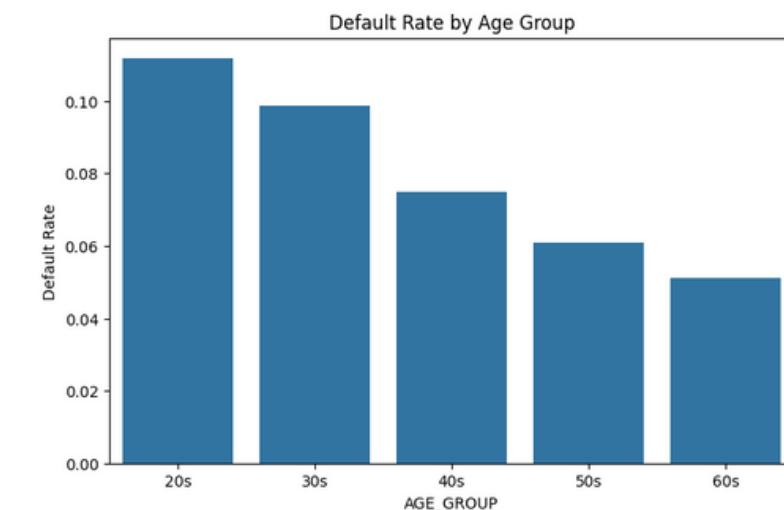
Jumlah Kredit vs Target

- Target 0 (lancar bayar) dan Target 1 (macet) sama-sama memiliki distribusi jumlah kredit (AMT_CREDIT) yang relatif mirip. Sedangkan Median jumlah kredit peminjam yang lancar sedikit lebih tinggi dibanding peminjam yang macet.
- Jumlah kredit yang diajukan tidak terlalu menentukan apakah peminjam akan macet atau tidak. Faktor lain (misalnya penghasilan, lama bekerja, atau sumber eksternal skor) lebih berpengaruh.
 - Home Credit sebaiknya tidak menolak otomatis berdasarkan besarnya pinjaman saja. Perlu mempertimbangkan faktor pendukung lain seperti penghasilan tetap, stabilitas pekerjaan, dan skor eksternal.

4

Lama Bekerja (Years Employed) vs Target

- Mayoritas peminjam, baik lancar maupun macet, punya lama bekerja < 10 tahun. Namun, distribusi peminjam macet (Target 1) cenderung lebih tinggi pada kelompok dengan lama bekerja sangat singkat (0–2 tahun). Sementara, peminjam dengan lama bekerja panjang (10+ tahun) cenderung lebih banyak masuk ke kelompok lancar.
 - Stabilitas pekerjaan terbukti penting. Pelanggan dengan pekerjaan jangka panjang memiliki kemungkinan lebih besar untuk lancar membayar.
 - Home Credit bisa memberi skor tambahan atau bunga lebih rendah bagi mereka yang punya riwayat kerja panjang, dan berhati-hati memberi pinjaman besar ke peminjam dengan lama kerja singkat.



5

Income vs Credit Amount by Target

- Grafik scatter menunjukkan hubungan antara pendapatan total (AMT_INCOME_TOTAL) dan jumlah kredit (AMT_CREDIT).
- Sebagian besar data terkonsentrasi pada pendapatan rendah–menengah (< 200.000) dengan kredit sekitar < 500.000.
- Baik peminjam dengan status default (TARGET = 1) maupun yang tidak (TARGET = 0) tersebar mirip, artinya income dan credit amount tidak cukup kuat sebagai pembeda langsung.
- Namun, peminjam dengan pendapatan rendah lebih rentan default dibanding yang berpendapatan tinggi (meski secara visual masih bercampur).

6

Default Rate by Age Group

- Semakin tua usia peminjam, semakin rendah default rate.
- 20s punya default rate tertinggi (~11%), diikuti 30s (~10%) dan usia 40s–60s lebih stabil dan rendah (5–7%).
- Usia muda lebih berisiko gagal bayar karena mungkin stabilitas pekerjaan dan finansial mereka masih lemah.

7

Default Rate by Employment Years & Age Group

- Pola risiko default dipengaruhi kombinasi lama bekerja dan usia.
- Risiko tinggi (default > 10%) banyak terjadi pada usia 20–30 tahun dengan pengalaman kerja < 5 tahun dan Usia 30–40 tahun dengan pengalaman 30–40 tahun
- Risiko menurun signifikan pada kelompok dengan pengalaman kerja > 10 tahun, terutama untuk usia 40+.
- Semakin lama pengalaman kerja, semakin kecil kemungkinan default, kecuali ada outlier tertentu.

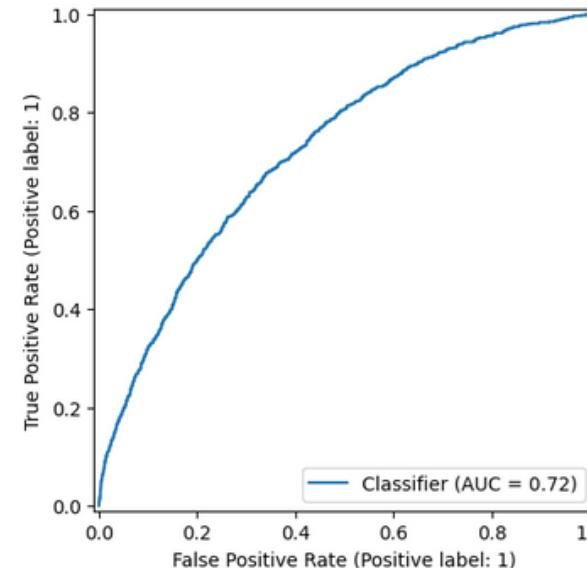
Data Modeling

Data Modeling: setelah pembersihan & transformasi, dataset dibagi dengan proporsi 80% train (± 37.659 data) dan 20% validasi/test (± 9.415 data).

1

Logistic Regression

Metric	Train Set	Validation Set
ROC-AUC	0.762	0.725
Accuracy (valid)	-	0.69
Precision (class 1)	-	0.15
Recall (class 1)	-	0.63
F1-score (class 1)	-	0.25



2

LightGBM Classifier

Metric	Non-Default (0)	Default (1)	Overall
Precision	0.94	0.23	-
Recall	0.91	0.32	-
F1-Score	0.92	0.27	-
Accuracy	-	-	0.86
Macro Avg (P/R/F1)	-	-	0.58 / 0.61 / 0.59
Weighted Avg (P/R/F1)	-	-	0.88 / 0.86 / 0.87

Confusion Matrix (LightGBM, Threshold=0.60)		
True Label	Pred: Non-Default	Pred: Default
Actual: Default	7659	1110
Actual: Non-Default	472	300

Logistic Regression cukup baik untuk baseline, tapi kurang optimal untuk prediksi kredit karena:

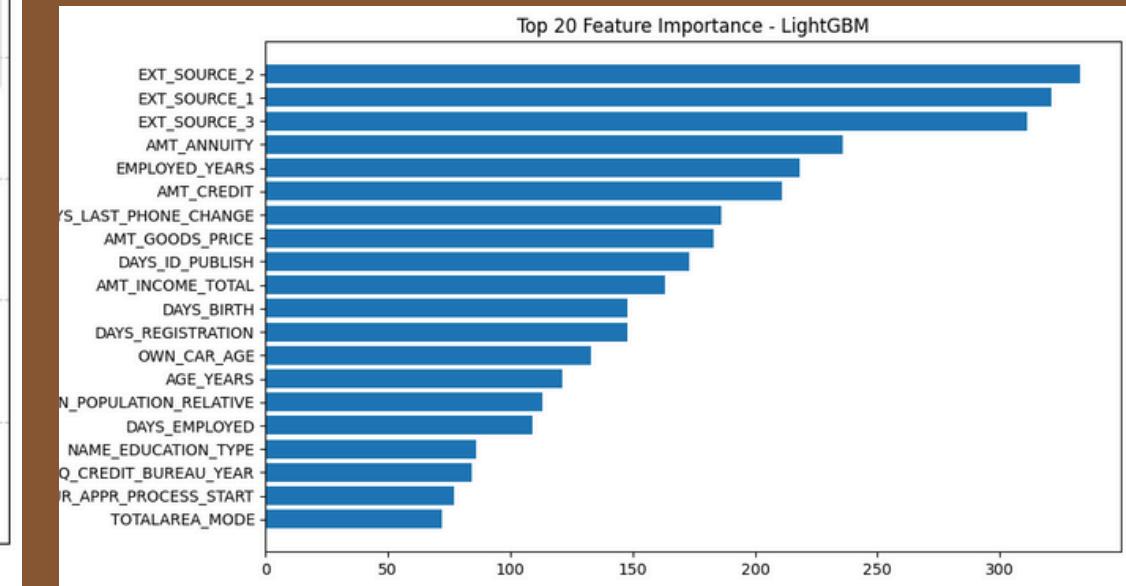
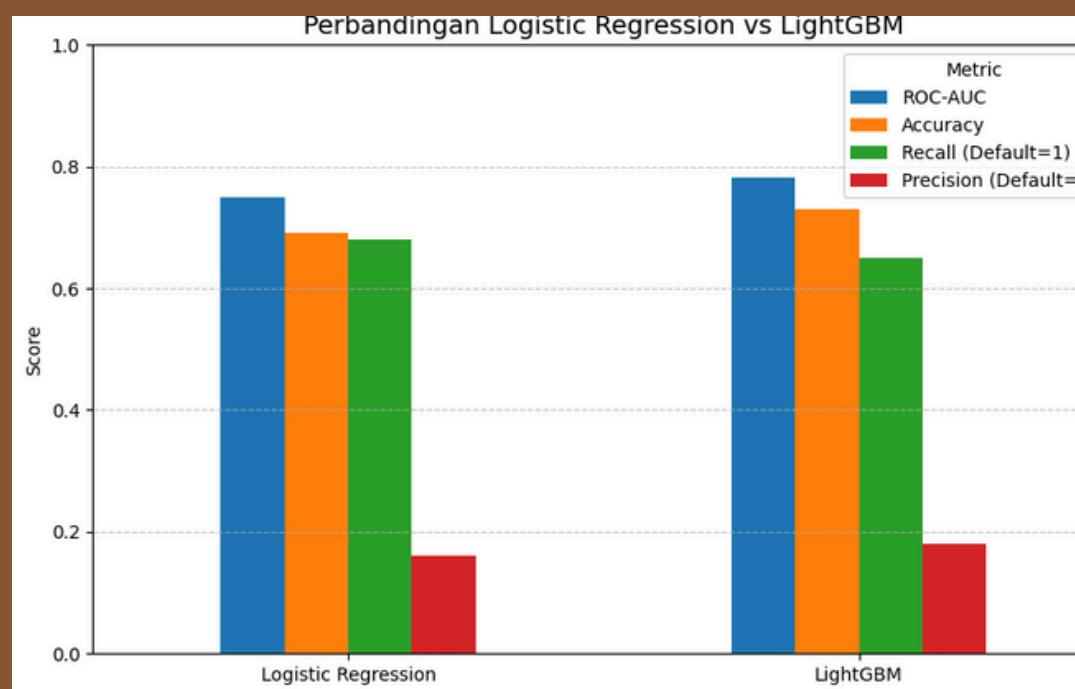
- Sensitif pada imbalance data dan Tingkat kesalahan prediksi nasabah gagal bayar masih tinggi (precision rendah).

LightGBM lebih baik dibanding Logistic Regression karena ROC-AUC & akurasi lebih tinggi.

- Threshold 0.65 membuat model lebih akurat.
- Cocok kalau tujuan bisnis adalah mengurangi salah tolak nasabah yang layak (reduce false positives)



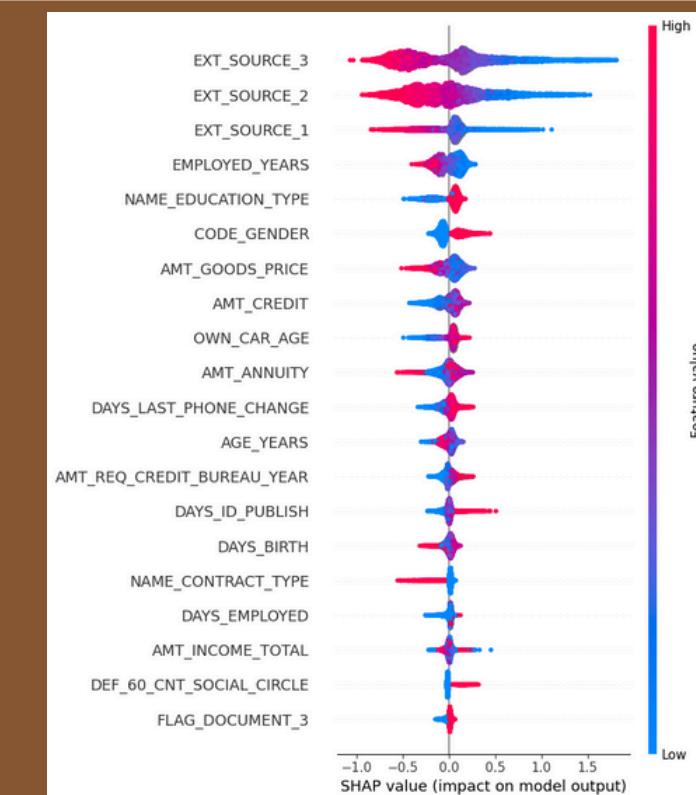
Komparasi 2 Model



- Logistic Regression lebih baik dalam Recall, sehingga cocok jika perusahaan ingin meminimalkan jumlah gagal bayar yang terlewat (false negative) meskipun banyak false positive.
- LightGBM unggul pada ROC-AUC, Accuracy, dan Precision, sehingga lebih baik secara keseluruhan untuk keseimbangan deteksi risiko dan akurasi prediksi.

1. Fitur Utama yang Berpengaruh
- EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 : merupakan skor eksternal (misalnya credit bureau score) yang menjadi prediktor paling kuat untuk risiko gagal bayar. Nilainya tinggi cenderung menurunkan kemungkinan default.
 - EMLOYED_YEARS & DAYS_EMPLOYED: lama bekerja berpengaruh besar. Semakin lama seseorang bekerja, semakin rendah risiko gagal bayar karena ada kestabilan penghasilan.
 - AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE: besarnya pinjaman, cicilan, dan harga barang menentukan kemampuan pelunasan. Pinjaman besar relatif meningkatkan risiko jika tidak sebanding dengan pendapatan.
 - AMT_INCOME_TOTAL: total pendapatan rumah tangga, semakin tinggi semakin menurunkan risiko default.
 - AGE_YEARS & DAYS_BIRTH: usia berhubungan dengan profil risiko. Nasabah usia lebih tua biasanya lebih stabil secara finansial.
 - NAME_EDUCATION_TYPE: tingkat pendidikan juga berpengaruh, pendidikan lebih tinggi cenderung berhubungan dengan kemampuan melunasi.
 - OWN_CAR_AGE : umur mobil bisa mencerminkan stabilitas ekonomi (aset).
 - DAYS_LAST_PHONE_CHANGE & DAYS_ID_PUBLISH: indikator kestabilan data (misalnya pergantian dokumen atau nomor HP terlalu sering bisa menandakan risiko lebih tinggi).

Model LightGBM yang dipilih



Business Recommendation



Menjadikan Skor Eksternal sebagai Faktor Utama Screening

- Dari hasil model, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 adalah prediktor paling kuat.
- Rekomendasi:
 - Jadikan skor eksternal sebagai prioritas utama dalam menilai kelayakan kredit.
 - Tetapkan ambang batas skor minimum agar nasabah dengan risiko sangat tinggi bisa difilter sejak awal.

Sesuaikan Jumlah Pinjaman dengan Pendapatan & Cicilan (Risk-based Lending)

- Fitur AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE dibandingkan dengan AMT_INCOME_TOTAL menentukan risiko gagal bayar.
- Rekomendasi:
 - Terapkan aturan Debt-to-Income Ratio (DTI) untuk menentukan besarnya pinjaman maksimal.
 - Cicilan bulanan (annuity) jangan melebihi persentase tertentu dari pendapatan nasabah.

Mempertimbangkan Stabilitas Pekerjaan dan Lama Bekerja

- Fitur EMPLOYED_YEARS, DAYS_EMPLOYED menunjukkan kestabilan penghasilan.
- Rekomendasi:
 - Nasabah dengan pekerjaan stabil bisa diberi tenor lebih panjang (maturity lebih fleksibel).
 - Untuk nasabah dengan masa kerja pendek atau sering berganti, berikan pinjaman lebih kecil dengan tenor pendek.

Segmentasi Berdasarkan Usia & Pendidikan

- AGE_YEARS dan NAME_EDUCATION_TYPE berhubungan dengan risiko default.
- Rekomendasi:
 - Usia produktif dengan pendidikan tinggi bisa diberi penawaran pinjaman lebih besar.
 - Usia terlalu muda atau pendidikan rendah perlu evaluasi lebih ketat.

Gunakan Indikator Stabilitas Data untuk Validasi

- Fitur DAYS_LAST_PHONE_CHANGE dan DAYS_ID_PUBLISH bisa mendeteksi ketidakstabilan identitas.
- Rekomendasi:
 - Jika terlalu sering ganti nomor atau dokumen, lakukan verifikasi tambahan.
 - Hal ini membantu mengurangi risiko fraud atau gagal bayar.

Implementasikan Repayment Calendar yang Fleksibel

- Dengan memahami faktor risiko, perusahaan bisa menyesuaikan repayment calendar.
- Rekomendasi:
 - Untuk nasabah dengan risiko rendah → tawarkan cicilan fleksibel (tenor panjang, bunga lebih rendah).
 - Untuk nasabah dengan risiko menengah → cicilan lebih ketat, bunga sedikit lebih tinggi.
 - Untuk nasabah dengan risiko tinggi → tawarkan pinjaman kecil dengan tenor pendek, atau lakukan reject bila terlalu berisiko.

Kesimpulan:

Penerapan risk-based lending dapat memanfaatkan skor eksternal dan variabel stabilitas (income, pekerjaan, usia, pendidikan, data personal) untuk menyaring nasabah, menentukan besaran pinjaman (principal), serta menyesuaikan tenor & repayment calendar. Dengan begitu, perusahaan bisa meminimalkan risiko gagal bayar, namun tetap memberikan kesempatan pinjaman bagi nasabah yang mampu melunasinya.