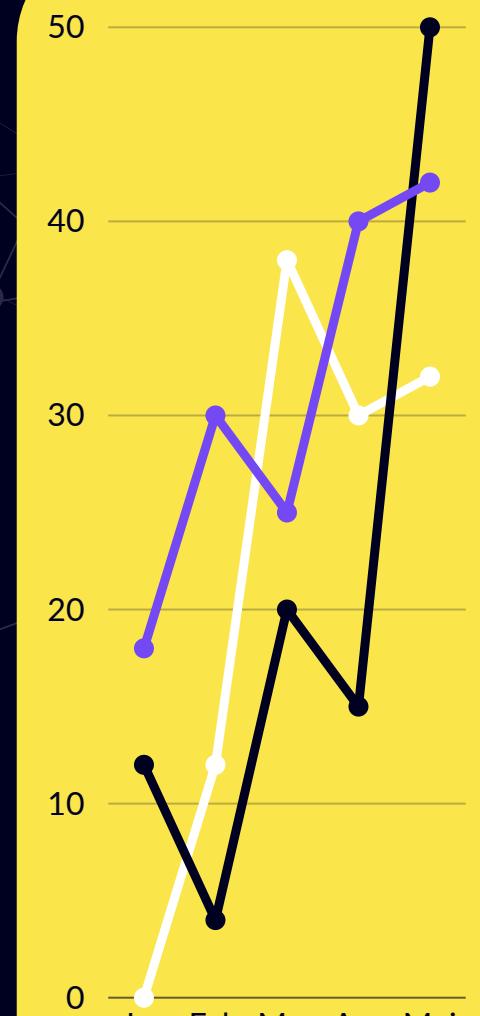


PREDICTING CUSTOMER CHURN IN A DIGITAL BANK

A Machine-Learning Case Study



Disusun Oleh : Elqory Ceasardy





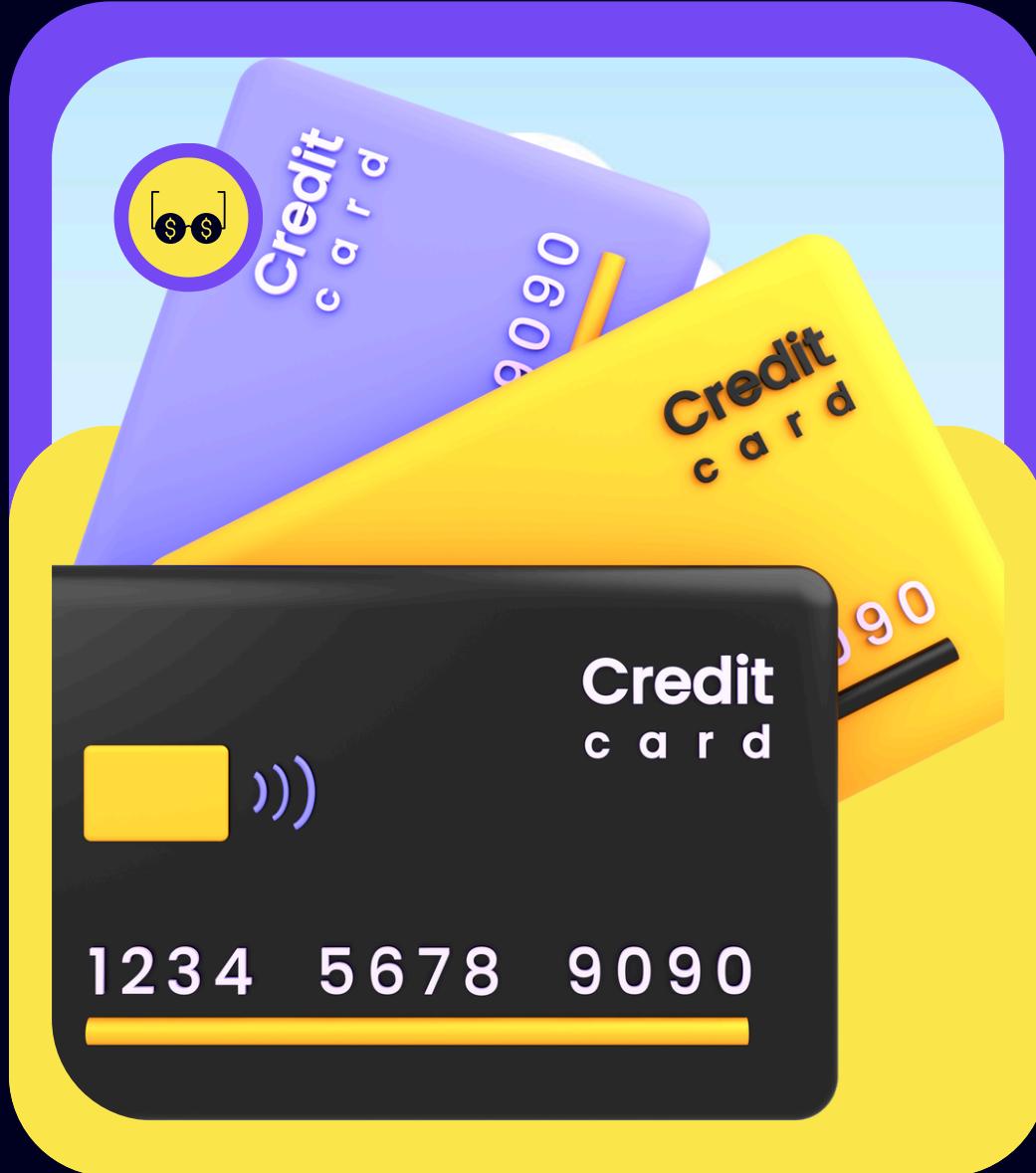
SELF-OVERVIEW

Perkenalkan, saya Elqory Ceasardy, peserta Data Science Bootcamp dengan fokus pada machine learning dan business analytics. Saya memiliki ketertarikan dalam mengolah data untuk menghasilkan insight yang dapat digunakan dalam pengambilan keputusan bisnis. Pada presentasi ini, saya akan membahas studi kasus Churn Analysis pada sektor perbankan digital sebagai bagian dari project akhir.



Content

- 01** Introduction
- 02** Data Cleaning & Preporcessing
- 03** Eda
- 04** Model
- 05** Insight
- 06** Closing



Background

“Churn atau hilangnya nasabah merupakan tantangan besar bagi industri perbankan digital. Tingginya tingkat churn dapat menyebabkan kerugian signifikan, karena biaya untuk mendapatkan nasabah baru biasanya jauh lebih tinggi dibandingkan mempertahankan nasabah lama. Oleh karena itu, analisis churn menjadi sangat penting untuk memahami pola perilaku nasabah, mengidentifikasi faktor-faktor yang memengaruhi keputusan mereka untuk meninggalkan layanan, serta menyusun strategi retensi yang lebih efektif. Dengan pendekatan ini, perusahaan dapat meningkatkan loyalitas nasabah sekaligus menjaga keberlanjutan bisnis.”

Objektive

Identifikasi faktor utama penyebab churn nasabah

Bangun model machine learning prediksi churn

Berikan rekomendasi strategi retensi customer



Dataset Overview



Sumber : Dataset Bank Digital Dieropa

Jumlah Data : 10.000+ Data Pelanggan

Terdapat 13 Kolom (12 Fitur 1 Target)

Customer ID	Pengenal unik untuk setiap pelanggan.
Surname	Nama keluarga atau nama belakang pelanggan
Credit Score	Nilai numerik yang mewakili skor kredit pelanggan.
Geography	Negara tempat pelanggan tinggal (Prancis, Spanyol, atau Jerman).
Gender	Jenis kelamin pelanggan (Pria atau Wanita).
Age	Usia pelanggan
Tenure	Jumlah tahun pelanggan telah menjadi nasabah bank.
Balance	Saldo rekening pelanggan.
NumOfProducts	Jumlah produk bank yang digunakan pelanggan (misalnya, rekening tabungan, kartu kredit).
HasCrCard	Apakah pelanggan memiliki kartu kredit (1 = ya, 0 = tidak).
IsActiveMember	Apakah pelanggan adalah anggota aktif (1 = ya, 0 = tidak).
EstimatedSalary	Perkiraan gaji pelanggan.
Exited	Apakah pelanggan telah berhenti berlangganan (1 = ya, 0 = tidak).

Data Cleaning & Preprocessing

Data Cleansing

- Tidak terdapat Missing value
- Tidak Terdapat Duplikatad

Split Data

Split data train dan data test

Feature Creation

'balance_to_salary', 'products_ge_2',
'age_tenure_ratio', 'inactive_card',
'credit_score_bucket', 'log_balance',
'log_estimated_salary'

Feature Selection

Drop kolom yang tidak relevan
('CustomerId', 'Surname')
Multicollinearity Check

Feature Transformation

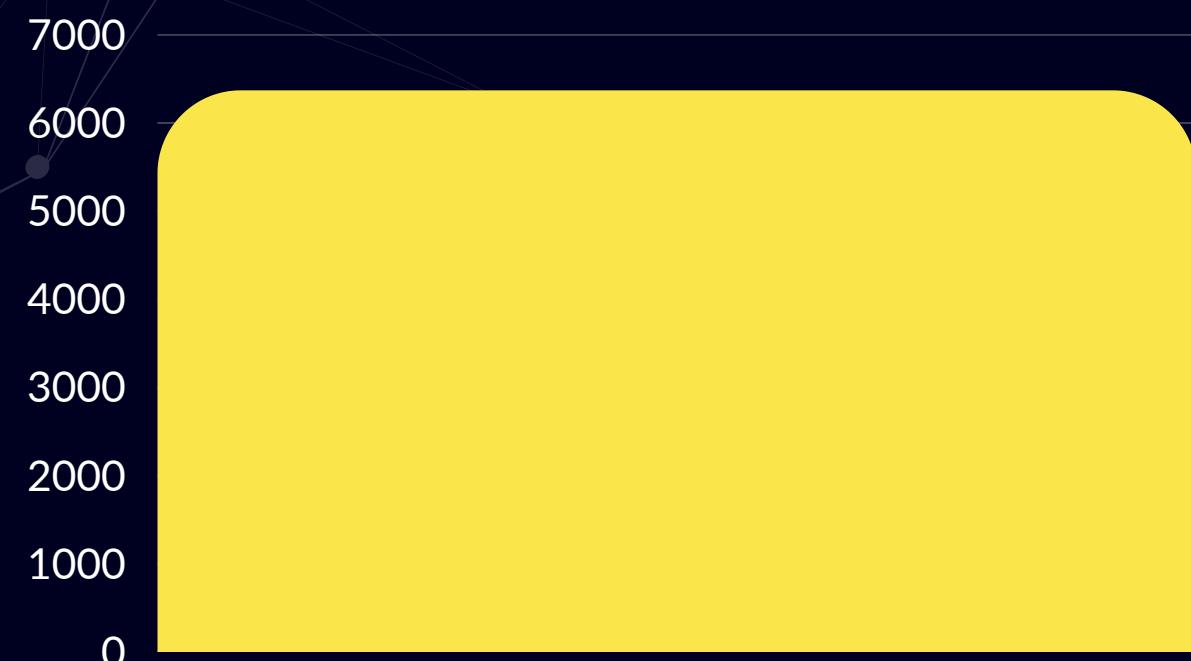
Encode Categorical Feature
StandardScaler

Data set already clean, next step eda

Mayoritas pelanggan bertahan, sekitar 80% pelanggan tetap menggunakan layanan, sedangkan 20% berhenti.

Karena target imbalanced, evaluasi bukan pakai accuracy aja. Next: PR-AUC, Recall@Top-10%, dan threshold berbasis profit.

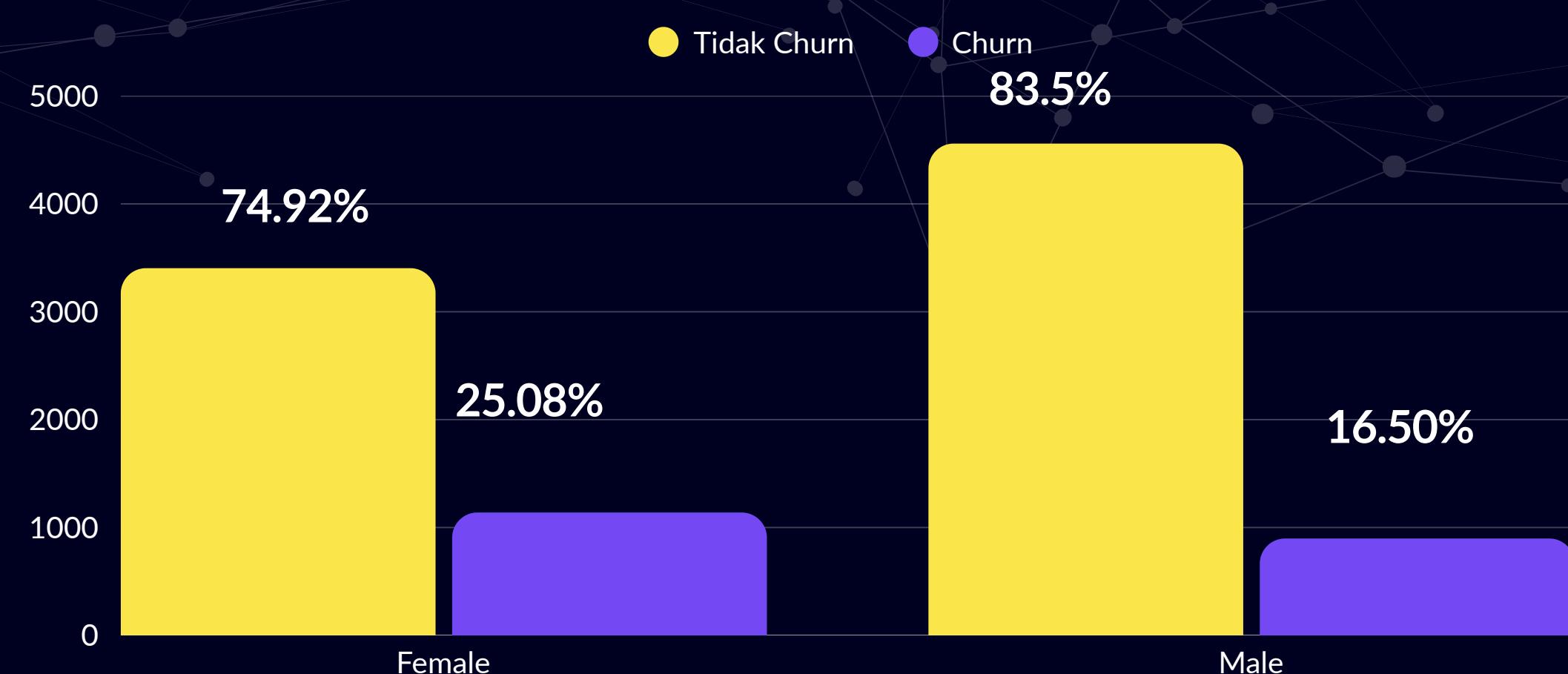
Bagaimana distribusi pelanggan yang churn?



Apakah gender pelanggan mempengaruhi churn?

Di train set, female churn rate = 25.1% sedangkan male = 16.5% (+8.6 pp). Female menyumbang 55.6% dari seluruh kasus churn meski hanya 45.2% dari populasi. Artinya, gender berpengaruh moderat—namun tetap perlu dilihat bareng faktor lain (balance, keaktifan, jumlah produk, credit score).

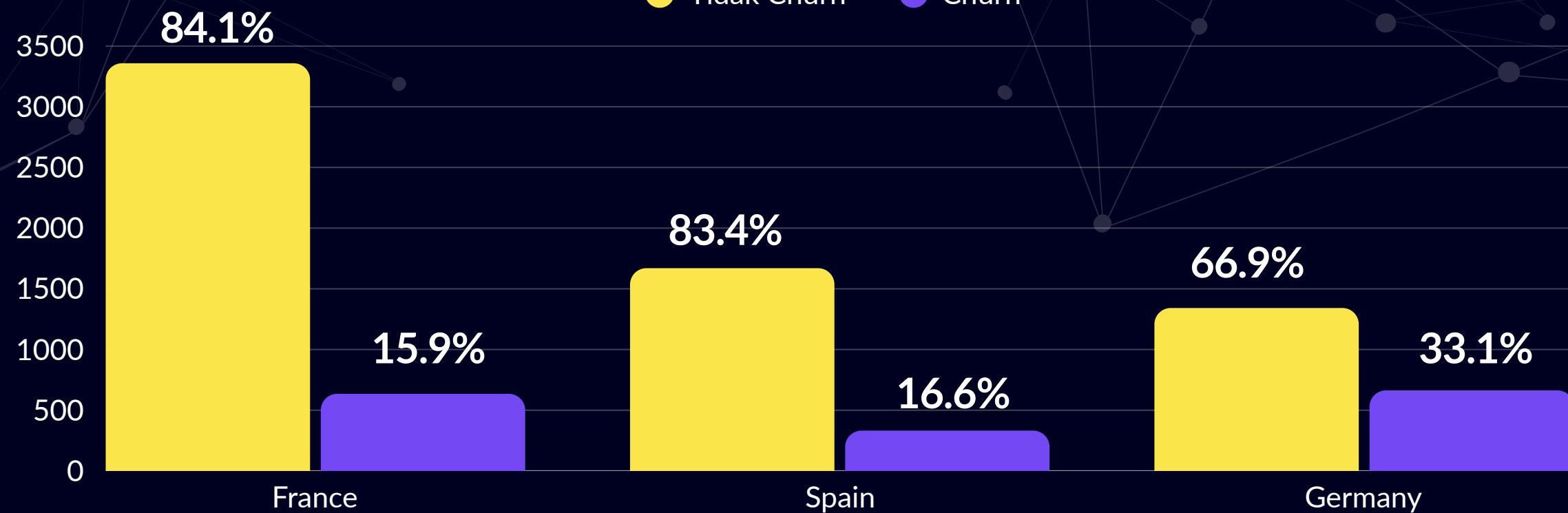
Prioritaskan segmen female untuk outreach, terutama yang inactive dan produk <2.



- Germany paling risk: churn rate 33.1% ($\approx 2 \times$ France/Spain).
- France & Spain relatif stabil (~16%).
- Artinya, nasabah Germany butuh prioritas retensi (biaya kebocoran terbesar ada di sini).
- Untuk fairness, selalu lihat rate (%) + jumlah N per negara (biar nggak bias volume).

- Prioritaskan outreach ke Germany terlebih dulu (top-10% skor tertinggi).
- Target spesifik: inactive member, produk <2, credit score rendah di Germany.
- Uji A/B insentif (fee waiver / cross-sell) khusus region ini.

Apakah geografis pelanggan juga menjadi faktor yang mempengaruhi churn?

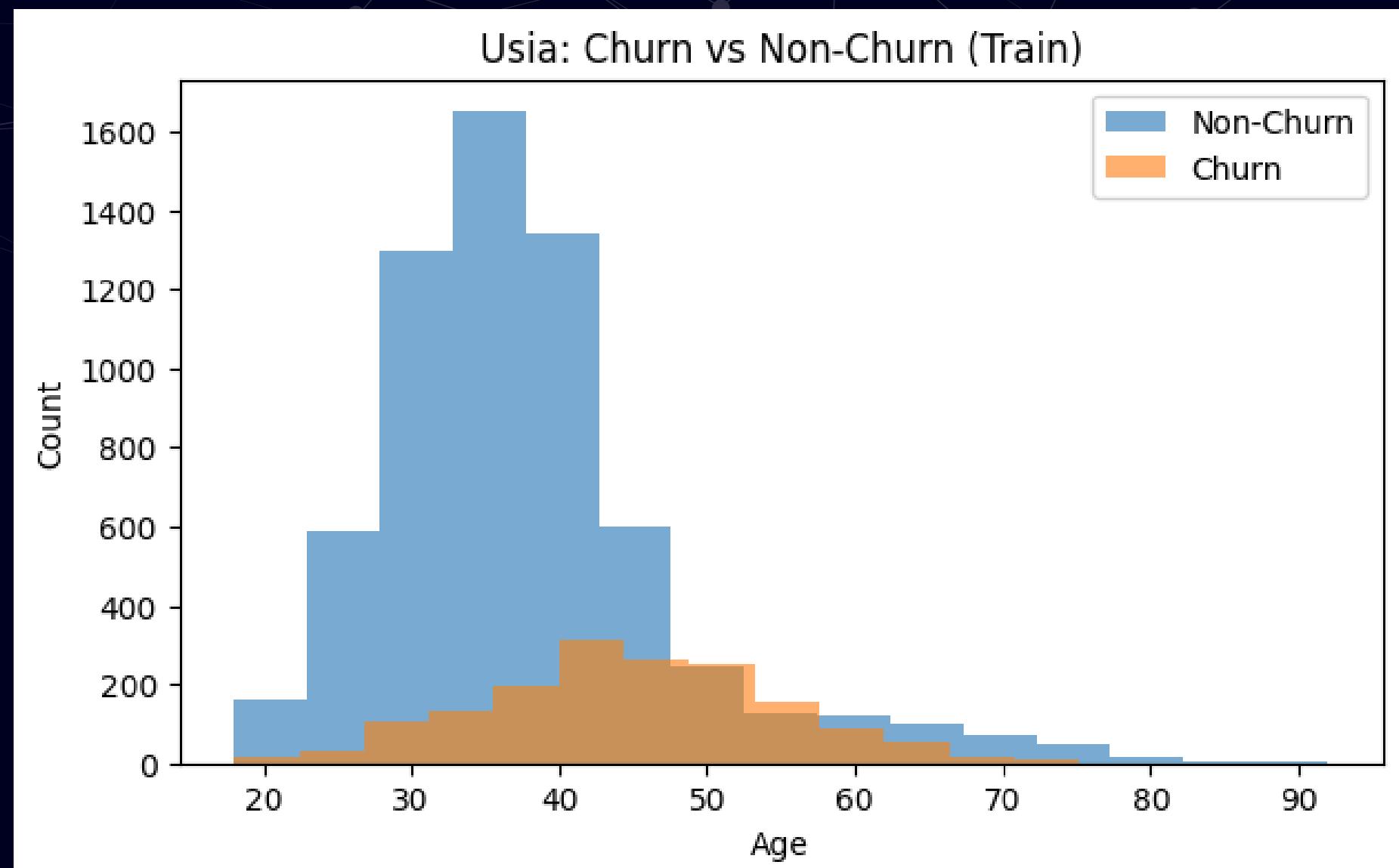


Bagaimana dengan usia pelanggan?

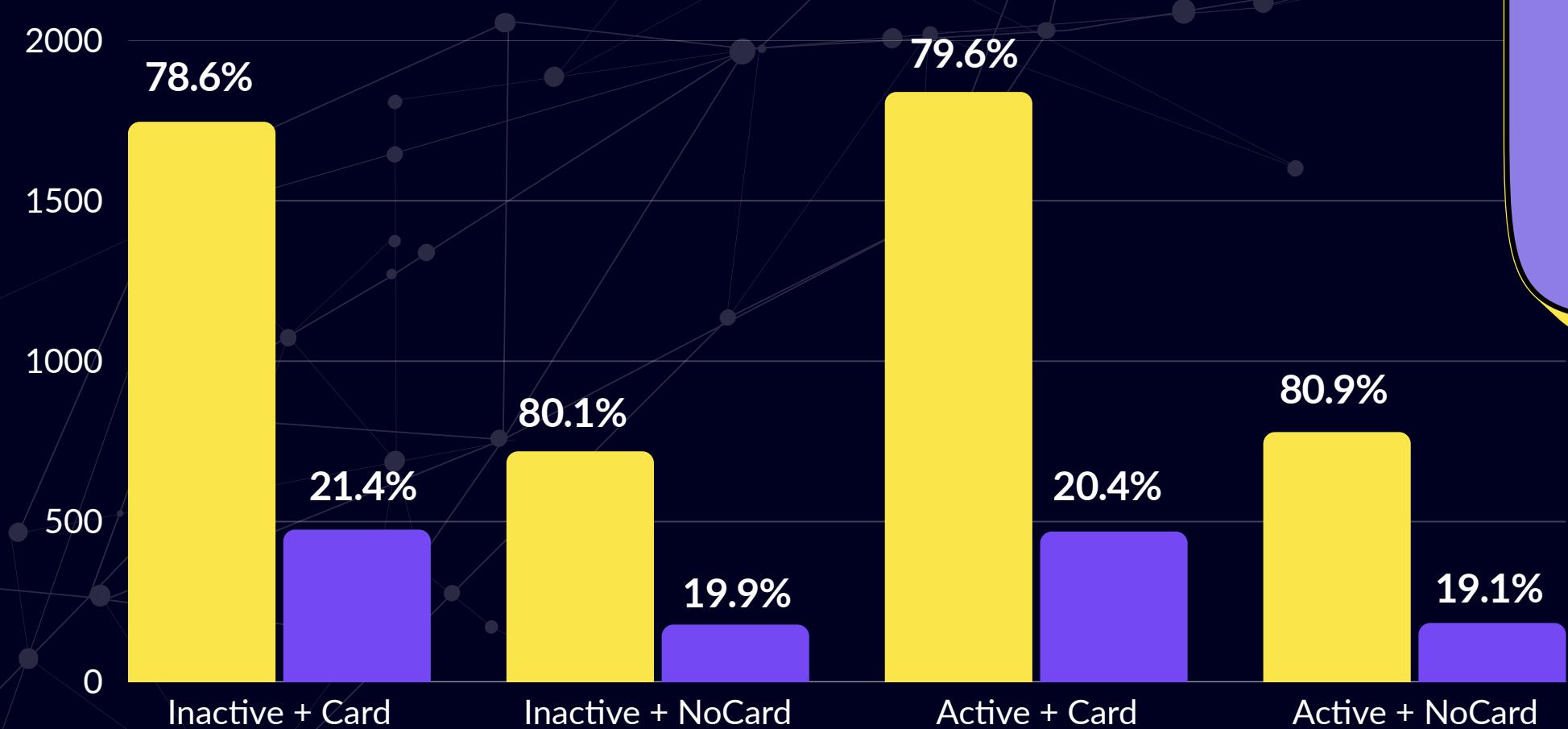
- Sebaran umur normal: mayoritas nasabah kumpul di 30-45 (peak).
- Bar oranye (churn) lebih ke kanan → yang churn cenderung lebih tua.
- Kalau lihat churn rate per rentang umur (decile/bucket), risikonya naik seiring usia dan biasanya paling tinggi di 45+.
- Artinya: usia itu sinyal, tapi bukan satu-satunya. Gabung dengan keaktifan, #produk, credit score biar targeting pas.



- Prioritaskan outreach untuk nasabah 45+ yang inactive / produk <2 / score rendah.
- Segmen muda: fokus onboarding & cross-sell (biar nambah produk ⇒ risiko turun).



Korelasi diantara active member dan memiliki Card untuk churn ?



- Semua segmen ~19–21% churn.
- Tertinggi: Inactive+Card ~21.4%; terendah: Active+NoCard ~19.1% → gap ≈ 2.3 pp (kecil).
- Aktivitas > Kartu. Status active konsisten lebih aman dikit; punya kartu sendiri efeknya tipis.
- Jadi, korelasi ada tapi lemah. Driver yang bener-bener ngaruh di dataset kamu itu keaktifan (sedikit) dan # produk (jauh lebih kuat), bukan “punya kartu” doang.

- Prioritaskan reaktivasi semua inactive (apa pun status kartu).
- Dorong 1 → 2 produk (ini yang paling nurunin churn di EDA kamu).
- Buat flag operasional: inactive_card cukup buat targeting outreach (nggak wajib masuk model).



Machine Learning



- 01 HistGradientBoosting Classifier
- 02 RandomForest Classifier
- 03 LightGBMClassifier
- 04 XGBoostClassifier
- 05 LogisticRegression

Modeling Evaluation

	roc_auc		pr_auc		recall@10	
	Train	Test	Train	Test	Train	Test
HGB	0.8639	0.8610	0.6911	0.7080	0.8187	0.8250
RF	0.8612	0.8550	0.6777	0.6870	0.8125	0.8050
LGBM	0.8480	0.8410	0.6684	0.6770	0.8063	0.7850
XGBoost	0.8385	0.8430	0.6545	0.6760	0.7875	0.8050
LR	0.8312	0.8370	0.6120	0.6280	0.7250	0.7250

Kami pilih PR-AUC + ROC-AUC + Recall@10% karena data imbalance dan eksekusi berbasis top-N; metrik ini paling mencerminkan dampak bisnis

Model Terbaik

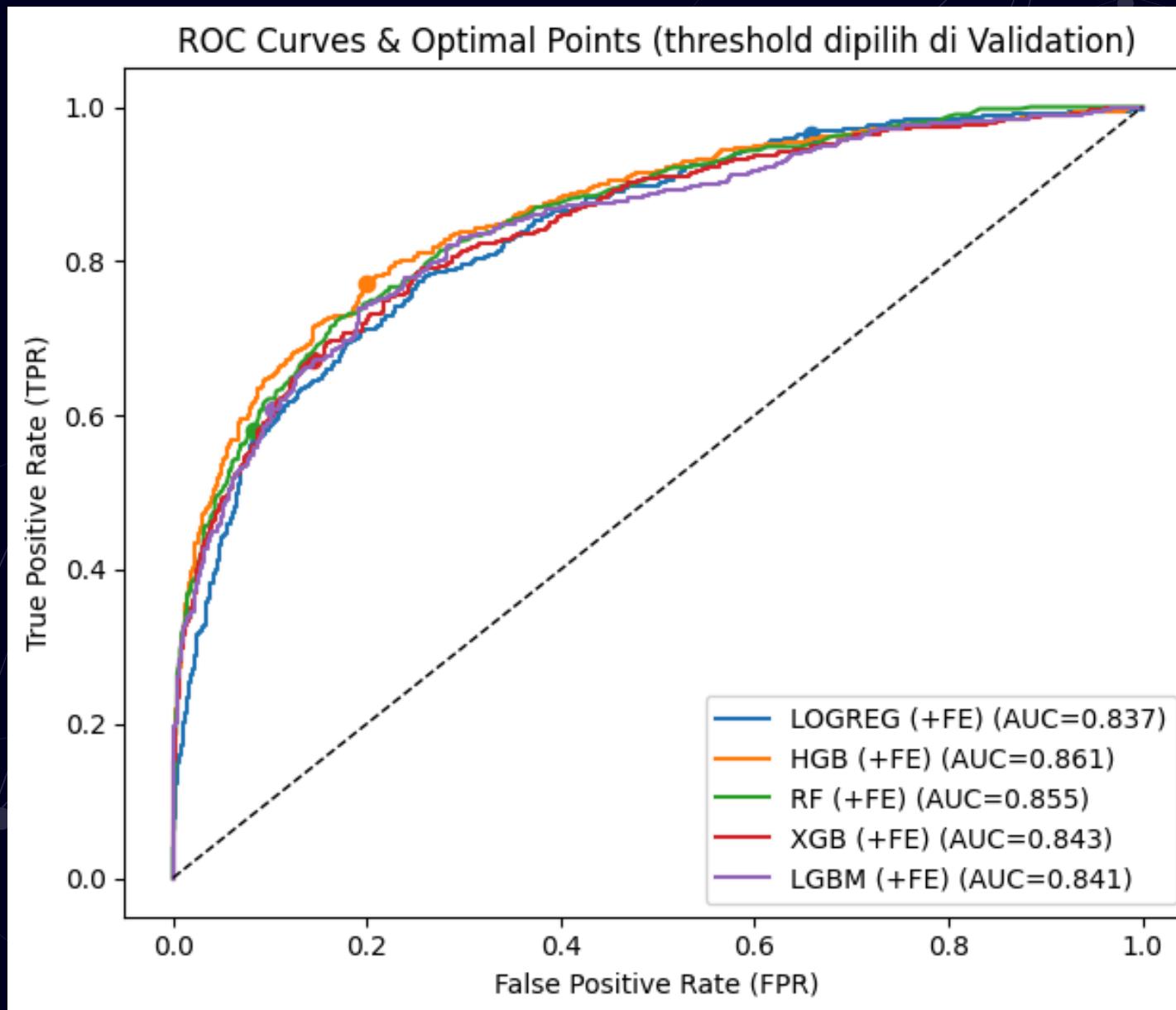


HistGradientBoosting

Kenapa HistGradientBoosting Terbaik ?

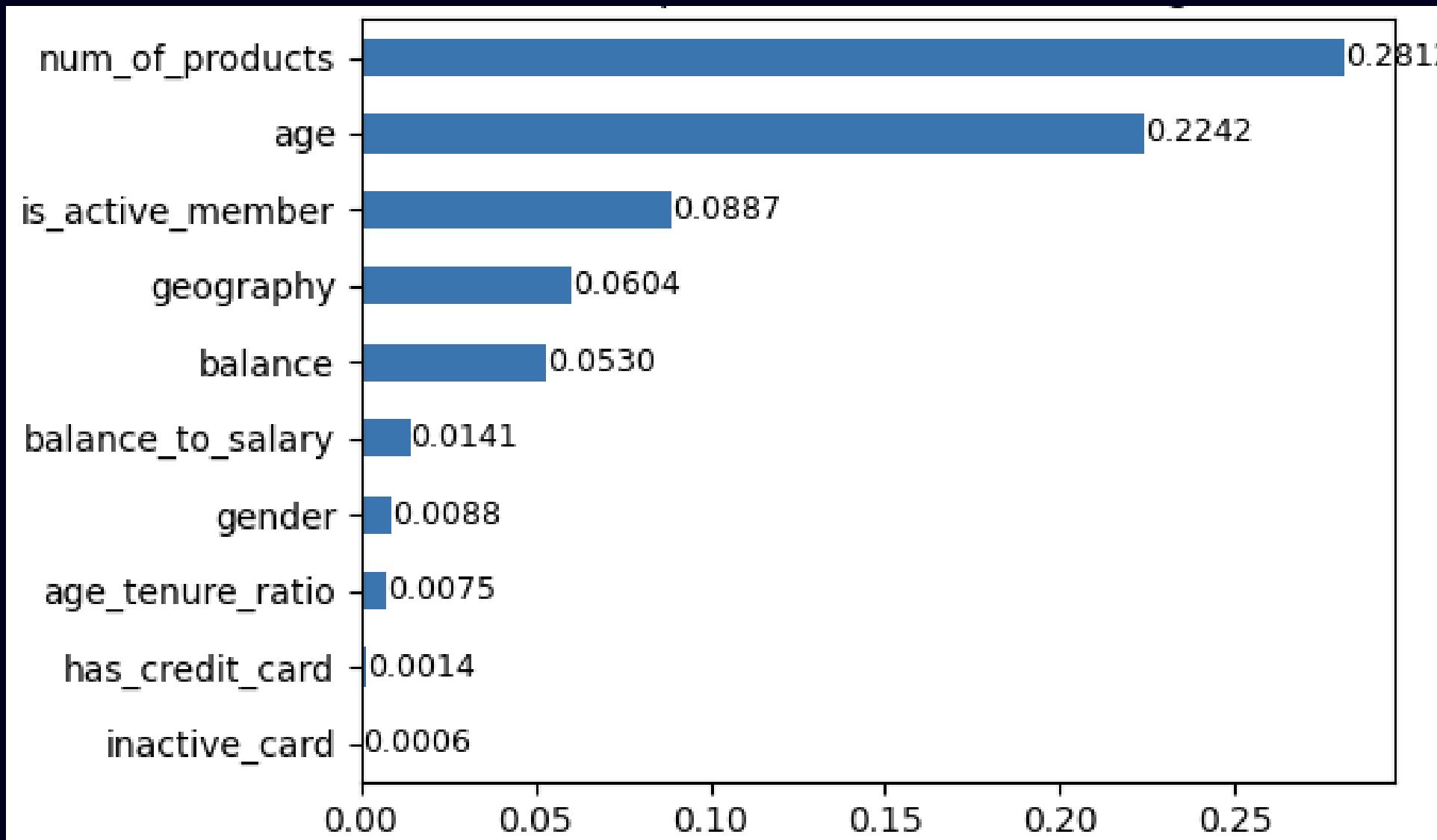
- Metrik utama unggul. HGB meraih PR-AUC & Recall@10% tertinggi (Test: PR-AUC ~0.708; Recall@10% ~0.825) → paling banyak menangkap calon churn di top-10% outreach.
- Generalize bagus. Gap Train–Test kecil (ROC-AUC ~0.864 vs ~0.861) → ranking stabil di data baru, risiko overfitting rendah.
- Tangkap non-linear & interaksi. Gradient boosting + binning histogram efektif membaca fitur turunan (ratio, bucket, log transform) dan interaksi (active × card) yang linear model sulit tangkap.
- Tahan outlier & skala. Binning histogram membuat model robust terhadap skala/ekstrem pada balance & salary; tidak terlalu sensitif ke scaling.
- Efisien & mudah dituning. Latihan cepat dan cukup stabil tanpa tuning berat; good baseline untuk produksi.
- Dukungan feature importance/SHAP memudahkan alasan bisnis (faktor terbesar pendorong churn).

ROC-AUC Metrics



- Slide ini menampilkan kurva ROC pada data Test untuk semua model.
- AUC mengukur kemampuan model memisahkan churn vs non-churn; makin tinggi makin baik.
- Model terbaik: HistGradientBoosting dengan AUC = 0,861.
- Titik pada kurva menandai threshold optimal dari validation berbasis profit ($\approx 0,20$).
- Dengan HGB @0,20, bank dapat menargetkan ±10% nasabah paling berisiko secara efisien.

Feature Importance (HistGradientBoosting)



- Top driver churn: #1 num_of_products, #2 age, #3 is_active_member, #4 geography, #5 balance.
- num_of_products → risiko melonjak di nasabah 3-4 produk.
- age → risiko naik di usia lebih tua (~45+).
- is_active_member → inactive jauh lebih berisiko.
- geography → Germany paling risk.
- balance → saldo tinggi + pasif = kombinasi rawan.
- Fitur minor (pengaruh langsung kecil): has_credit_card, gender, age_tenure_ratio, balance_to_salary – sering penting saat berinteraksi (mis. inactive × card), yang ditangkap model boosting.
- Aksi prioritas: fokus outreach ke (1) 3-4 produk, (2) usia 45+, (3) inactive, (4) Germany, (5) saldo tinggi, lewat program re-aktivasi, edukasi/right-product, dan early warning untuk kombinasi faktor di atas.

Rekomendasi Bisnis (1)

High-Risk Customer ($p \geq 0.20$)

Aksi Utama (48 jam pertama)

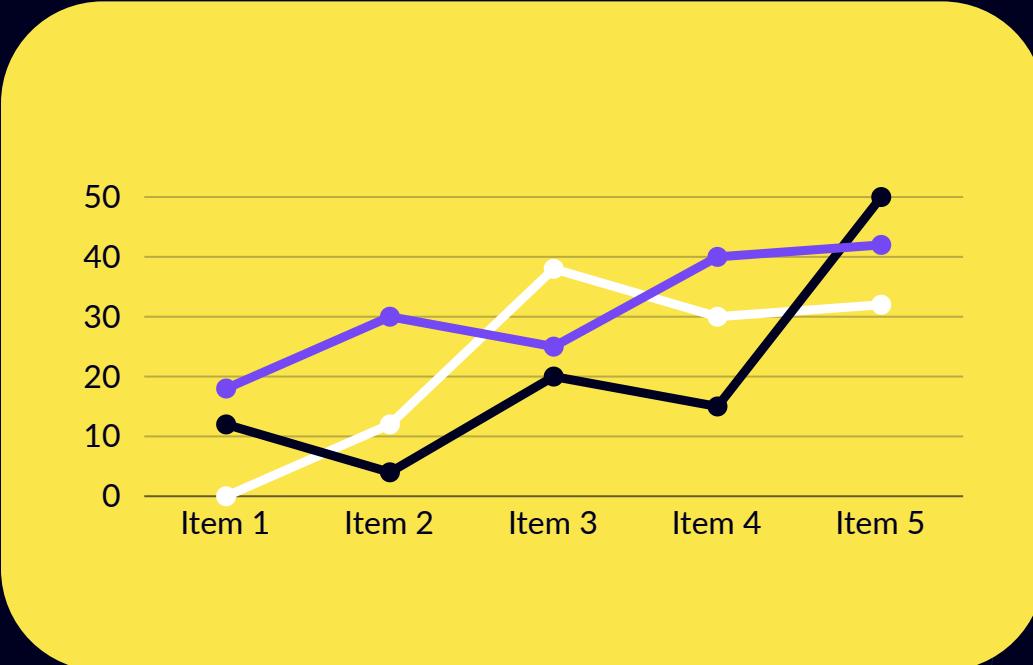
- RM Call Prioritas: Hubungi ≤ 48 jam. Tawarkan solusi berbasis masalah:
- Relief pembayaran: fee waiver 1x/late fee reversal, penjadwalan ulang/tenor ringan.
- Re-activation untuk yang inactive: pandu aktivasi aplikasi + bonus transaksi pertama.
- Simplifikasi produk (jika ≥ 3 produk): audit produk “nganggur”, tawarkan bundle simplification / downgrade sukarela untuk kurangi friksi & biaya.
- Autopay & reminder tagihan: bantu set autodebit + kalender tagihan.
- Penyesuaian lokal (Germany): jam kontak sesuai zona, materi lokal, perbaiki fee policy yang sensitif.

Channel & Insentif

- Utama: RM call + follow-up WA verified.
- Insentif: cashback kecil targetted / fee-rebate 1x bila aktif ≥ 3 transaksi/bulan.

Target KPI (60–90 hari)

- Penurunan churn High-Risk $\geq 5-8$ pp vs. kontrol.
- $\geq 60\%$ terhubung; $\geq 35\%$ aktif kembali (≥ 3 transaksi/bulan).
- ROI+ (benefit retensi > biaya insentif & tenaga).



Rekomendasi Bisnis (2)

Medium-Risk Customer ($0.10 \leq p < 0.20$)

Aksi Utama

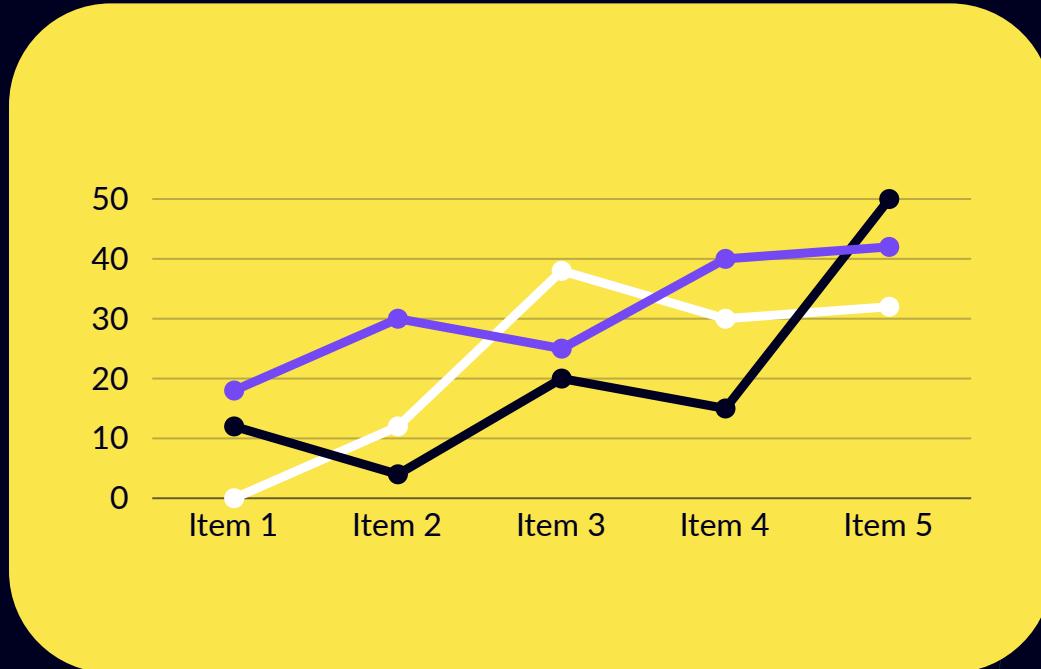
- Digital nudges: in-app/WA/email—autopay, due date, fitur yang belum dipakai.
- Right-product offer: produk relevan/lebih sederhana untuk pemilik 2-3 produk; light incentive (poin/fee-rebate kecil).
- Edukasi & kebiasaan: tips keamanan, budgeting, dan “3 transaksi/bulan” sebagai activation habit.

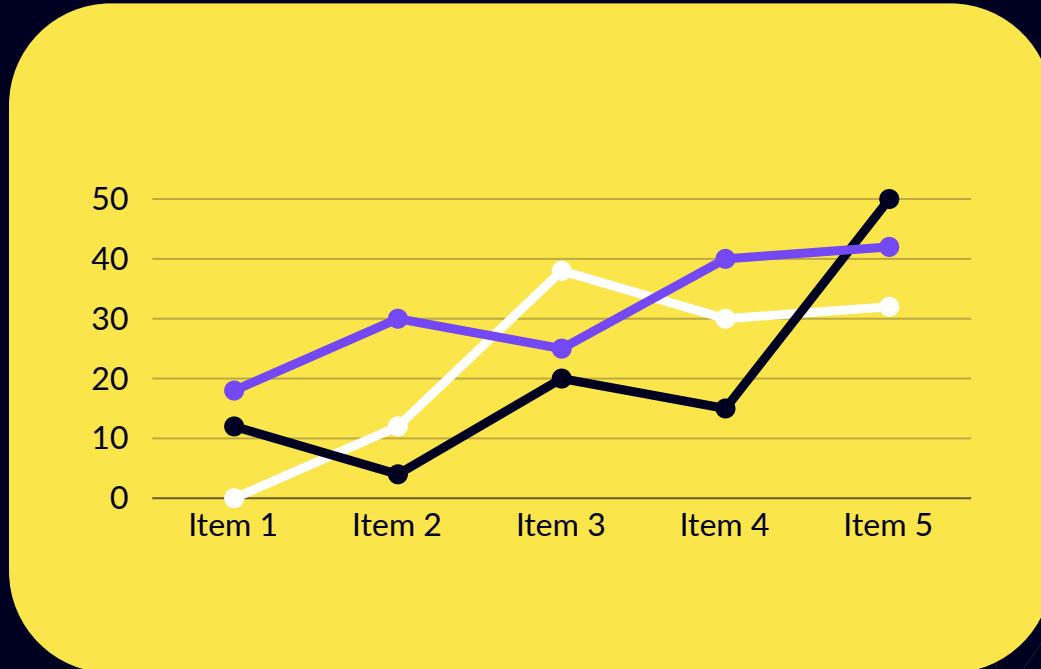
Channel & Insentif

- Utama: in-app + email + WA broadcast resmi (tanpa telefon).
- Insentif: poin kecil jika capai 3-5 transaksi/bulan selama 2 bulan

Target KPI

- Penurunan churn Medium-Risk $\geq 2-3$ pp vs. kontrol.
- Kenaikan feature adoption (autopay, aktivasi app) $\geq +15\%$.





Rekomendasi Bisnis (3)

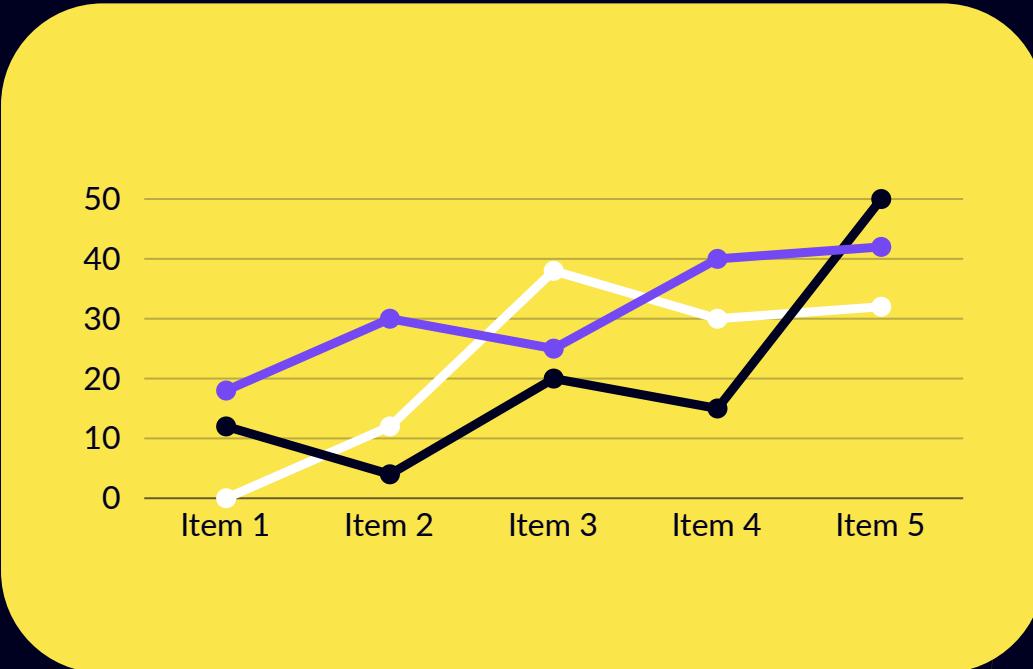
Low-Risk Customer ($p < 0.10$)

Aksi Utama

- Loyalty & advocacy: referral bonus, early-access fitur/premium ringan.
- Cross-sell ringan: produk bernilai tambah, tanpa diskon besar.
- Monitoring: pantau tren skor mingguan; do-not-disturb untuk yang tak responsif.

Target KPI

Stabil churn Low-Risk; NPS↑, up/cross-sell moderat.



Kesimpulan

- Model terbaik: HistGradientBoosting (HGB) – paling konsisten di PR-AUC dan Recall@Top-10%, serta gap train-test kecil (generalizable).
- Performa uji (Test): ROC-AUC ≈ 0.86 , PR-AUC ≈ 0.70 , Recall@Top-10% $\approx 0.82-0.83 \rightarrow$ mampu menangkap mayoritas calon churn di 10% nasabah teratas.
- Ambang operasional: Threshold ≈ 0.20 (Top-10%) sebagai titik terbaik antara jangkauan dan efisiensi biaya outreach.
- Driver utama churn: # Produk, Usia (\uparrow), Status aktif (inactive), Geografi (Germany), Balance \rightarrow dasar personalisasi penawaran & pesan.
- Strategi aksi:
- High-Risk ($p \geq 0.20$): RM call + relief biaya/penyederhanaan produk + aktivasi autopay.
- Medium-Risk (0.10–0.20): nudge digital, edukasi, light-incentive.
- Low-Risk ($p < 0.10$): loyalty & referral, monitor tanpa ganggu.
- Dampak bisnis yang ditargetkan (60–90 hari): penurunan churn High-Risk 5–8 pp, Medium-Risk 2–3 pp, ROI positif (benefit retensi > biaya insentif & tenaga).
- Rencana implementasi: scoring mingguan, routing ke CRM, A/B test insentif & pesan, monitoring metrik lift vs control dan kesehatan model (PR-AUC, calibration).

INFORMASI KONTAK



www.vyruz.id



elqoryc@gmail.com

