FINAL TASK: HOME CREDIT SCORECARD MODEL

Muhammad Ali Umar

X

Rakamin Academy

LATAR BELAKANG

Home Credit saat ini sedang menggunakan berbagai macam metode statistik dan Machine Learning untuk membuat prediksi skor kredit. Manager memintaku untuk membuka potensi maksimal dari data Home Credit. Dengan melakukannya, kita dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman dan pinjaman dapat diberikan dengan *principal, maturity, dan repayment calendar* yang akan memotivsi pelanggan untuk sukses. Evaluasi akan dilakukan dengan mengecek seberapa dalam pemahaman analisa yang aku kerjakan. Sebagai catatan, aku perlu menggunakan setidaknya 2 model **Machine Learning** dimana salah satunya adalah **Logistic Regression.**

TUJUAN

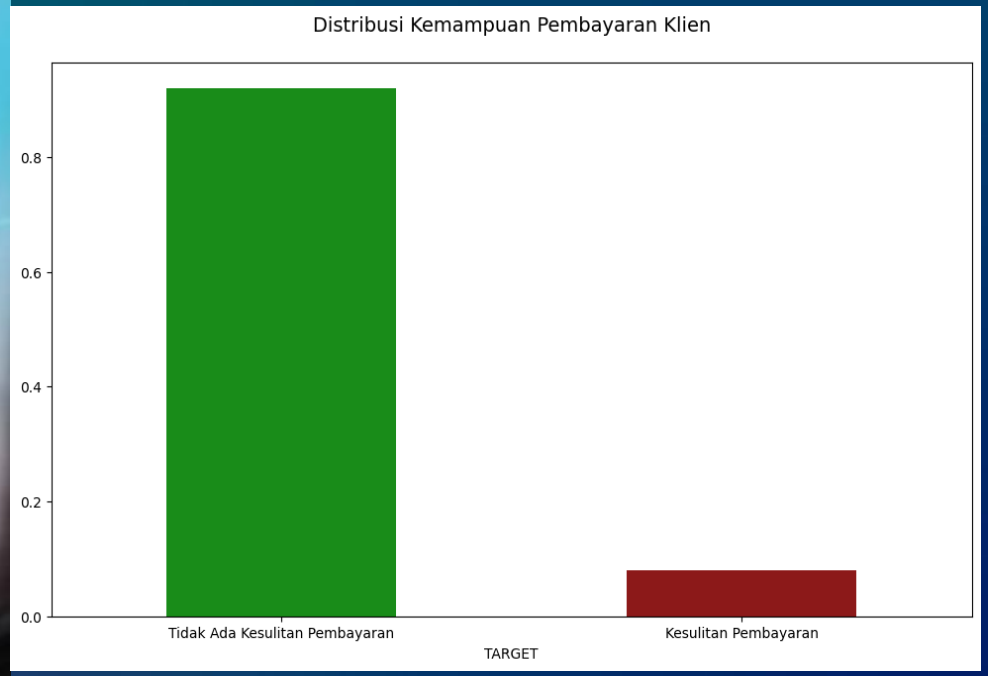
1. Mengidentifikasi karakteristik calon klien yang akan mengalami kesulitan dalam membayar pinjaman.

2. Memprediksi kemampuan pembayaran klien.

RINGKASAN DATASET

Dataset application_train yang digunakan untuk analisis ini memiliki:

- Jumlah Baris: 307.511
- Jumlah Kolom: 122



DISTRIBUSI VARIABEL TARGET ('TARGET')

Variabel target menunjukkan apakah klien mengalami kesulitan dalam membayar pinjaman (TARGET = 1) atau tidak (TARGET = 0). Distribusi kelas adalah sebagai berikut:

- Tidak Ada Kesulitan Pembayaran (TARGET = 0): Sekitar 92% (282.686 pemohon)
- Kesulitan Pembayaran (TARGET = 1): Sekitar 8% (24.825 pemohon)

Ini menunjukkan ketidakseimbangan kelas yang signifikan, dimana sebagian besar klien tidak mengalami masalah pembayaran.

EXPLORATORY DATA ANALYSIS (EDA)

Kemampuan Pembayaran Klien Berdasarkan Jenis Kelamin

	CODE_GENDER	TARGET	SK_ID_CURR
0	Laki-laki	Kesulitan Pembayaran	10655
1	Laki-laki	Tidak Ada Kesulitan Pembayaran	94404
2	Perempuan	Kesulitan Pembayaran	14170
3	Perempuan	Tidak Ada Kesulitan Pembayaran	188278
4	XNA	Tidak Ada Kesulitan Pembayaran	4

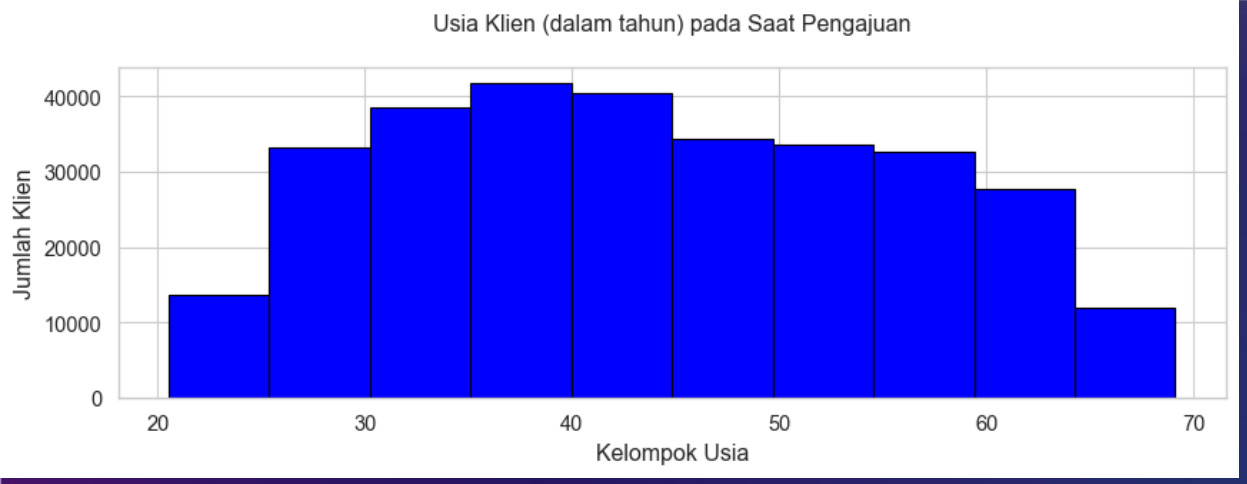
- Klien perempuan lebih banyak mengajukan pinjaman, namun klien laki-laki memiliki persentase kesulitan pembayaran yang sedikit lebih tinggi.
- Perempuan: 188.278 (Tidak Ada Kesulitan Pembayaran), 14.170 (Kesulitan Pembayaran). Tingkat default sekitar 7.0%.
- Laki-laki: 94.404 (Tidak Ada Kesulitan Pembayaran), 10.655 (Kesulitan Pembayaran). Tingkat default sekitar 10.1%.
- Meskipun perempuan lebih sering mengajukan pinjaman, laki-laki memiliki risiko default yang lebih tinggi. Informasi ini dapat digunakan untuk segmentasi risiko yang lebih baik.

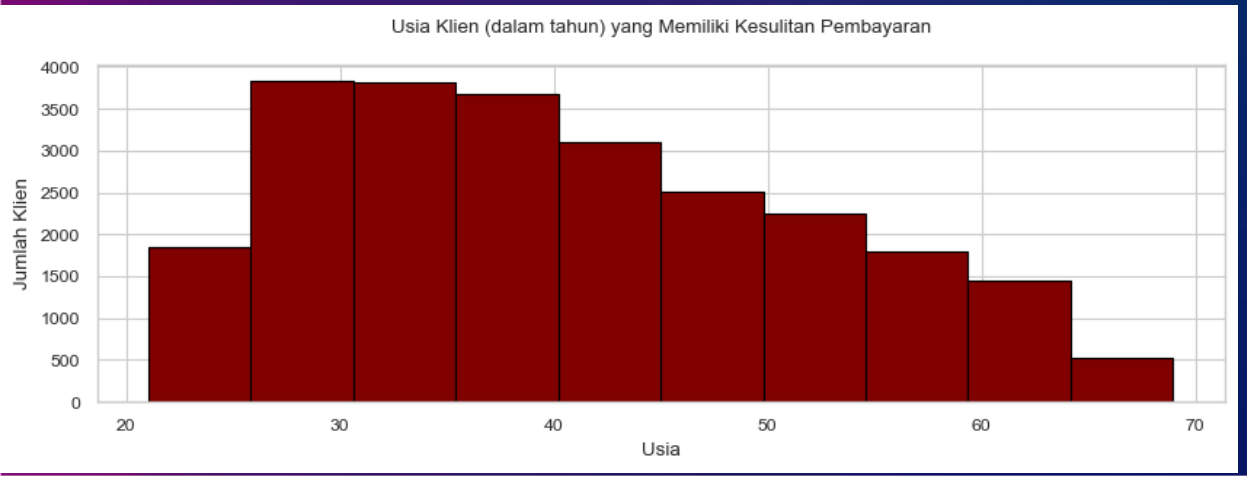
Kemampuan Pembayaran Klien Berdasarkan Status Keluarga

	NAME_FAMILY_STATUS	TARGET	SK_ID_CURR
0	Civil marriage	Kesulitan Pembayaran	2961
1	Civil marriage	Tidak Ada Kesulitan Pembayaran	26814
2	Married	Kesulitan Pembayaran	14850
3	Married	Tidak Ada Kesulitan Pembayaran	181582
4	Separated	Kesulitan Pembayaran	1620
5	Separated	Tidak Ada Kesulitan Pembayaran	18150
6	Single / not married	Kesulitan Pembayaran	4457
7	Single / not married	Tidak Ada Kesulitan Pembayaran	40987
8	Unknown	Tidak Ada Kesulitan Pembayaran	2
9	Widow	Kesulitan Pembayaran	937
10	Widow	Tidak Ada Kesulitan Pembayaran	15151

- Status keluarga tertentu menunjukkan kecenderungan risiko yang berbeda terhadap kemampuan pembayaran.
- Klien dengan status keluarga **Civil marriage** memiliki persentase kesulitan pembayaran tertinggi (9.94%), diikuti oleh **Single/not married** (9.81%).
- Struktur keluarga dapat mempengaruhi beban finansial dan sumber daya yang tersedia untuk melunasi pinjaman. Status **Civil marriage** mungkin menunjukkan komitmen finansial yang berbeda dibandingkan status lainnya.

EXPLORATORY DATA ANALYSIS (EDA)





- Sebagian besar klien yang mengajukan pinjaman berada dalam rentang usia 35-40 tahun, diikuti oleh rentang 40-45 tahun.

- Hubungan dengan Target:
  - Klien yang tidak memiliki kesulitan pembayaran umumnya berada dalam rentang usia 35-45 tahun.
  - Klien yang memiliki kesulitan pembayaran lebih sering ditemukan dalam rentang usia yang lebih muda, yaitu 25-35 tahun.
  - Ini menunjukkan bahwa usia yang lebih matang (35-45 tahun) berkorelasi dengan kemampuan pembayaran yang lebih baik, sedangkan kelompok usia yang lebih muda (25-35 tahun) memiliki risiko kesulitan pembayaran yang lebih tinggi.

STRATEGI PEMBERSIHAN DAN PREPROCESSING DATA

PENANGANAN NILAI YANG HILANG (MISSING VALUES)

- Fitur-fitur dengan persentase nilai hilang lebih dari 50% (seperti COMMONAREA_AVG, OWN_CAR_AGE, EXT_SOURCE_1 dan banyak fitur terkait apartemen) dihapus. Ini bertujuan untuk mengurangi dimensi data dan menghindari imputasi yang berlebihan pada kolom yang minim informasi.
- Untuk nilai hilang yang tersisa pada kolom-kolom lain, strategi imputasi diterapkan:
  - Fitur Numerik: Nilai hilang diisi dengan median dari masing-masing kolom. Median dipilih karena lebih robust terhadap outlier dibandingkan mean, yang sering ditemukan pada distribusi data finansial dan demografi.
  - Fitur Kategorikal: Nilai hilang diisi dengan modus (nilai yang paling sering muncul) dari masing-masing kolom. Ini memastikan bahwa kolom kategorikal tetap dapat digunakan oleh model dan tidak ada kategori baru yang diperkenalkan secara acak.

KONVERSI NILAI NEGATIF KE POSITIF

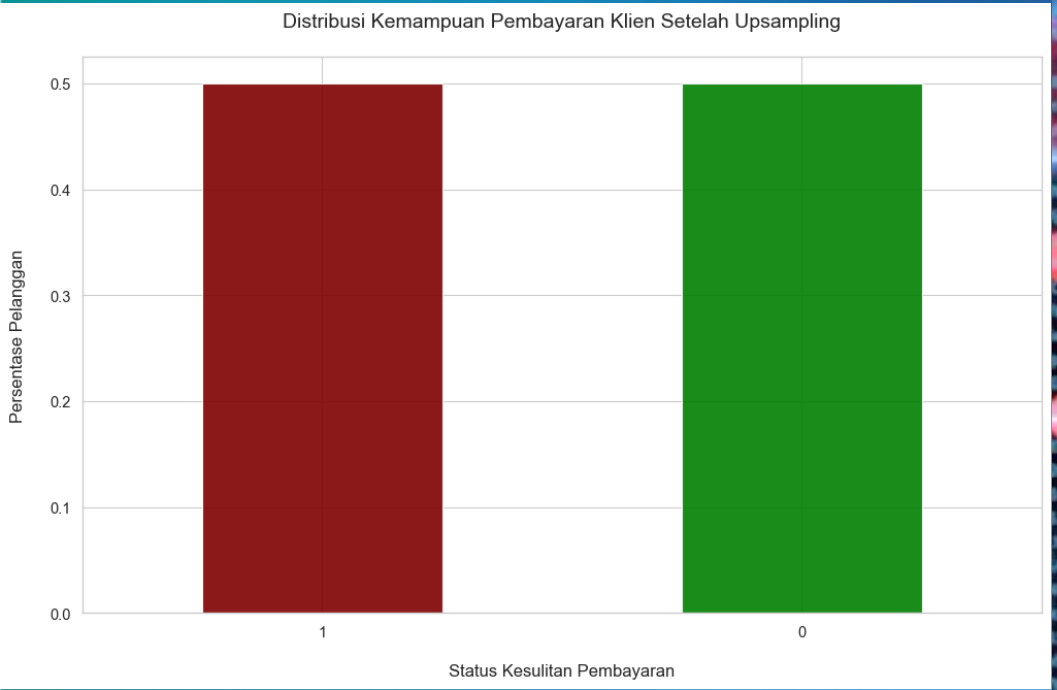
- Beberapa fitur temporal seperti DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, dan DAYS_LAST_PHONE_CHANGE awalnya direpresentasikan dalam nilai negatif (misalnya, jumlah hari sejak tanggal aplikasi). Untuk kemudahan interpretasi dan konsistensi, nilai-nilai ini dikonversi menjadi absolut (positif), merepresentasikan durasi dalam hari.

LABEL ENCODING UNTUK FITUR KATEGORIKAL

- Fitur-fitur kategorikal nominal (misalnya, NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, OCCUPATION_TYPE, WEEKDAY_APPR_PROCESS_START, ORGANIZATION_TYPE) dikonversi menjadi representasi numerik menggunakan Label Encoding. Setiap kategori unik dalam kolom tersebut diberi nilai integer yang berbeda. Ini diperlukan karena sebagian besar algoritma machine learning memerlukan input data dalam format numerik.

```
Sebelum upsampling:
Kelas mayoritas (0): 282686
Kelas minoritas (1): 24825
```

```
Setelah upsampling:
Kelas mayoritas (0): 282686
Kelas minoritas (1): 282686
```



STRATEGI PENANGANAN KETIDAKSEIMBANGAN DATA

STRATEGI PENANGANAN KETIDAKSEIMBANGAN DATA

Untuk mengatasi ketidakseimbangan kelas pada variabel target, digunakan metode **Upsampling** pada kelas minoritas. Kelas **Kesulitan Pembayaran** (TARGET = 1) di-*upsample* agar jumlahnya sama dengan kelas mayoritas **Tidak Ada Kesulitan Pembayaran** (TARGET = 0).

- Sebelum *Upsampling*:
  - Kelas Mayoritas (0): 282.686
  - Kelas Minoritas (1): 24.825

- Setelah *Upsampling*:
  - Kelas Mayoritas (0): 282.686
  - Kelas Minoritas (1): 282.686

Hal ini memastikan bahwa model tidak bias terhadap kelas mayoritas dan dapat belajar secara efektif dari kedua kelas.

PERBANDINGAN PERFORMA MODEL MACHINE LEARNING

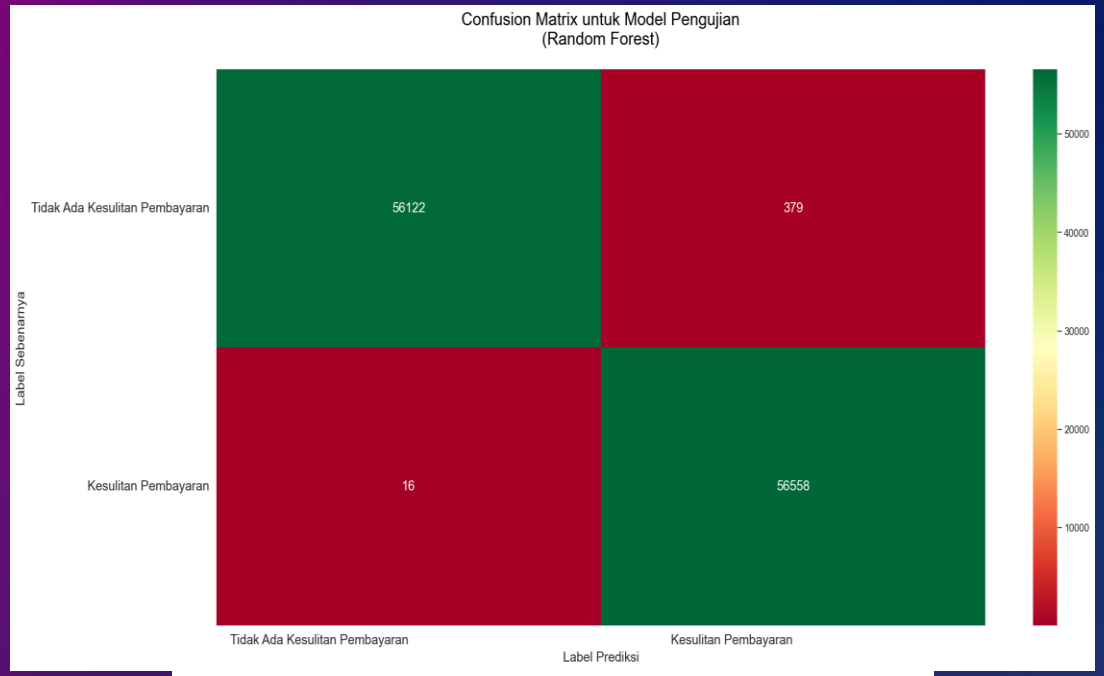
- Setelah melatih dan mengevaluasi beberapa model Machine Learning, berikut adalah ringkasan perbandingan performa masing-masing model berdasarkan akurasi pelatihan, akurasi pengujian, dan skor ROC AUC.

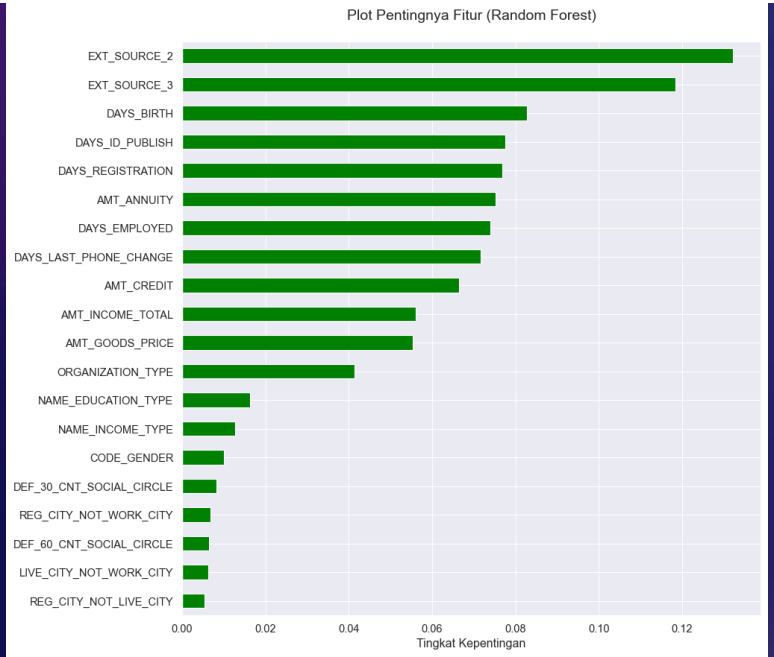
Ringkasan Performa Model

Model	Akurasi Training (%)	Akurasi Testing (%)	Skor ROC AUC
Random Forest	100.0	99.62	0.9962
Pohon Keputusan	100.0	88.35	0.8835
Regresi Logistik	67.16	67.30	0.6730
Naive Bayes Gaussian	60.24	60.39	0.6040

- Berdasarkan metrik akurasi pengujian dan skor ROC AUC, **Random Forest** adalah model terbaik dengan akurasi pengujian sebesar **99.62%** dan skor ROC AUC sebesar **0.9962**. Model ini menunjukkan kemampuan prediksi yang sangat tinggi dalam mengidentifikasi klien yang memiliki kesulitan pembayaran.

DETAIL MODEL TERBAIK DAN FEATURE IMPORTANCE





Confusion Matrix (Data Pengujian)

Visualisasi *Confusion Matrix* dari model **Random Forest** pada data pengujian menunjukkan:

- **True Negatives (TN):** 56122 klien diprediksi tidak memiliki kesulitan pembayaran, dan memang tidak memiliki kesulitan pembayaran.
- **False Positives (FP):** 379 klien diprediksi tidak memiliki kesulitan pembayaran, namun sebenarnya memiliki kesulitan pembayaran.
- **False Negatives (FN):** 16 klien diprediksi memiliki kesulitan pembayaran, namun sebenarnya tidak memiliki kesulitan pembayaran.
- **True Positives (TP):** 56558 klien diprediksi memiliki kesulitan pembayaran, dan memang memiliki kesulitan pembayaran.

Matriks ini menunjukkan bahwa model **Random Forest** sangat baik dalam mengidentifikasi baik klien yang bermasalah maupun yang tidak, dengan jumlah kesalahan prediksi yang sangat minimal.

Feature Importance

Plot Feature Importance mengidentifikasi fitur-fitur yang paling berpengaruh dalam prediksi model Random Forest. Fitur-fitur teratas meliputi:

- **DAYS_EMPLOYED:** Jumlah hari klien telah bekerja. Ini menunjukkan bahwa durasi pekerjaan adalah prediktor yang sangat kuat terhadap kemampuan pembayaran.
- **DAYS_BIRTH:** Usia klien (dalam hari). Usia merupakan faktor demografi penting yang memengaruhi risiko kredit.
- **AMT_CREDIT:** Jumlah kredit pinjaman. Besarnya pinjaman memiliki dampak langsung pada kemampuan pengembalian.
- **AMT_GOODS_PRICE:** Harga barang yang dibeli dengan pinjaman. Serupa dengan jumlah kredit, nilai ini juga krusial.
- **AMT_INCOME_TOTAL:** Total pendapatan klien. Tingkat pendapatan secara langsung memengaruhi kapasitas pembayaran.

Fitur-fitur ini sangat relevan karena mencerminkan stabilitas finansial (pekerjaan, pendapatan) dan demografi (usia) klien, serta beban finansial pinjaman itu sendiri. Memahami fitur-fitur ini membantu dalam mengambil keputusan kredit yang lebih tepat dan merancang strategi mitigasi risiko.

KESIMPULAN DAN REKOMENDASI BISNIS

Proyek ini berhasil mengembangkan model prediksi credit scoring yang efektif untuk Home Credit. Analisis data yang mendalam mengungkap karakteristik klien yang berisiko tinggi dan rendah, serta faktor-faktor kunci yang memengaruhi kemampuan pembayaran. Model Random Forest menunjukkan kinerja terbaik dengan akurasi dan skor ROC AUC yang sangat tinggi pada data pengujian, menjadikannya alat yang kuat untuk pengambilan keputusan kredit.

1. REKOMENDASI PEMASARAN:

- Targetkan kampanye pada klien berusia 35-45 tahun yang memiliki kemampuan bayar lebih baik
- Fokus pada profesional seperti akuntan, manajer, dan staf teknologi terampil tinggi
- Klien dengan kepemilikan properti menunjukkan risiko yang lebih rendah

2. MANAJEMEN RISIKO:

- Berhati-hati dengan klien berstatus cuti hamil dan pengangguran
- Tinjau ulang aplikasi dari klien dengan pendidikan menengah pertama
- Monitor ketat klien dengan tipe pekerjaan 'Low-Skilled Worker’

3. PENGEMBANGAN PRODUK:

- Pertimbangkan produk khusus untuk klien dengan pinjaman berputar
- Kembangkan program edukasi keuangan untuk klien usia muda (25-35 tahun)
- Buat skema pembayaran fleksibel untuk klien dengan pendapatan tidak tetap

4. OPTIMISASI OPERASIONAL:

- Gunakan model Random Forest untuk *screening* aplikasi kredit
- Prioritaskan fitur-fitur penting seperti DAYS_EMPLOYED, AMT_GOODS_PRICE, AMT_CREDIT
- Implementasikan sistem monitoring berkelanjutan untuk performa model

TERIMA KASIH

AL.

+6287785588499

al.emailkerja@gmail.com

GitHub

Youtube