Effective Strategies For

CREDIT CARD APPROVAL PREDICTION

Otomatisasi dengan Machine Learning

Student Batch 35 : Fuad Hasyim

AGENDA

01 Executive Summary

02 Background

03 Data Introduction

04 Exploratory Data Analysis (EDA)

05 Modelling

06 Recommendation

Executive Summary

Background

Perusahaan menghadapi tantangan dalam menilai kelayakan nasabah baru untuk produk kredit. Kesalahan dalam menerima nasabah berisiko tinggi (Bad) dapat menyebabkan **kerugian finansial**, naiknya *Non-Performing Loan*, serta tekanan operasional. Oleh karena itu, proses penilaian risiko perlu ditingkatkan menggunakan pendekatan berbasis Machine Learning.

Objectives

1. **Bisnis** – Meminimalkan risiko gagal bayar dengan mendeteksi sebanyak mungkin calon nasabah yang berisiko tinggi.
2. Machine Learning – Mengembangkan model prediksi yang **mengutamakan Recall untuk kelas Bad (high risk)**, sehingga meminimalkan False Negative (nasabah berisiko tetapi lolos).
3. Memastikan model tetap **interpretable** agar dapat digunakan oleh tim risk management.

Results Summary

Hasil analisis menunjukkan bahwa dataset memiliki ketidakseimbangan kelas yang sangat tinggi Bad (high risk) hanya 1.71%.

Model-model yang diuji menunjukkan performa beragam, namun **Gradient Boosting dengan hyperparameter tuning memberikan recall test tertinggi (0.85)** pada kelas Bad.

Fitur seperti **Total Income, Age, dan Family member** menjadi faktor utama dalam menentukan risiko gagal bayar.

Model-model linear dan ensemble cenderung kurang efektif untuk pola risiko non-linear pada dataset ini.

Secara keseluruhan, model yang direkomendasikan adalah **Gradient Boosting**, konsisten dengan prioritas bisnis untuk **memaksimalkan deteksi nasabah berisiko tinggi (recall).**

Background

Latar Belakang

Perusahaan kartu kredit harus menentukan apakah calon nasabah layak diberikan kredit. Keputusan ini sangat penting karena:

- **Jika menyetujui nasabah berisiko tinggi** → potensi kerugian finansial (default loss).
- **Jika menolak nasabah yang sebenarnya baik** → hilangnya potensi pendapatan (opportunity cost).

Saat ini perusahaan memprioritaskan **keamanan finansial**, artinya:

- ✔ Lebih baik menolak sebagian calon nasabah yang sebenarnya baik
- ✖ Daripada menyetujui calon nasabah yang berpotensi gagal bayar.

Company Context

Kondisi Perusahaan Saat Ini

- Riwayat NPL stabil, tetapi tetap dalam pengawasan.
- Regulasi internal & kondisi ekonomi mendorong perusahaan lebih konservatif.
- Strategi terbaru:
  - Fokus menjaga keamanan finansial
  - Menggunakan filter risiko yang lebih ketat melalui model ML

Business Problem

Permasalahan Utama

Perusahaan ingin **mengidentifikasi calon nasabah berisiko tinggi gagal bayar sedini mungkin.**

Prioritas

- **Minimalkan risiko gagal bayar (default risk).**

Konsekuensinya:

- ✔ Perusahaan rela menerima lebih banyak False Positive
  - (menolak nasabah baik)
- ✖ Tidak mau False Negative
  - (meloloskan nasabah buruk → risiko kerugian)

Trade-Off

- Recall ↑ (lebih banyak Bad terdeteksi)
- Precision ↓ (beberapa Good ikut ditolak)

Goal

Membangun model yang memaksimalkan recall, sehingga sebanyak mungkin nasabah berisiko tinggi (Bad) dapat terdeteksi, demi meminimalkan potensi kerugian akibat gagal bayar.

Data Introduction

Sumber dataset

- Dataset ini berasal dari Kaggle

Dataset terdiri dari dua tabel utama:

- **Application Record**
  - 438,557 rows | 18 columns
  - Berisi data demografi dan kondisi sosial-ekonomi pemohon.
- **Credit Record**
  - 1,048,575 rows | 3 columns
  - Berisi histori pembayaran bulanan setiap nasabah.

Merged and cleaned data: **29,165 rows**

- Bad: 499 **nasabah (1.71%)**
- Good: 28.666 **nasabah (98,29%)**

Column : 20 Column (**Original**)

Credit Record

Feature name	Explanation	Remarks
ID	Client number	
MONTHS_BALANCE	Record month	Hitungan mundur: 0 = bulan ini, -1 = bulan lalu, dst.
STATUS	Credit status	0 = 1-29 hari telat bayar; 1 = 30-59 hari; … 5 = >150 hari, C = lunas bulan itu, X = tidak ada pinjaman

Application Record

Feature name	Explanation	Remarks
ID	Client number	Unique ID untuk setiap pemohon
CODE_GENDER	Gender	Laki-laki / Perempuan
FLAG_OWN_CAR	Is there a car	1 = punya mobil, 0 = tidak
FLAG_OWN_REALTY	Is there a property	1 = punya properti, 0 = tidak
CNT_CHILDREN	Number of children	Jumlah anak
AMT_INCOME_TOTAL	Annual income	Total pendapatan tahunan pemohon
NAME_INCOME_TYPE	Income category	Contoh: Working, State servant, etc.
NAME_EDUCATION_TYPE	Education level	Basic, Secondary, Higher, Academic
NAME_FAMILY_STATUS	Marital status	Menikah, Single, Separated, dll.
NAME_HOUSING_TYPE	Way of living	Contoh: Rented apartment, House, etc.
DAYS_BIRTH	Birthday	Negatif: hitungan mundur dari hari ini; -1 = kemarin
DAYS_EMPLOYED	Start date of employment	Negatif = lama bekerja; positif = tidak bekerja
FLAG_MOBIL	Has mobile phone	1 = ya
FLAG_WORK_PHONE	Has work phone	1 = ya
FLAG_PHONE	Has phone	1 = ya
FLAG_EMAIL	Has email	1 = ya
OCCUPATION_TYPE	Occupation	Tipe pekerjaan
CNT_FAM_MEMBERS	Family size	Jumlah anggota keluarga

Data Cleaning

Handling Missing Values

Missing value kolom **OCCUPATION_TYPE** : 11.323 Rows Nan (30% Nan)
Handling → Rows **Nan** → Drop
Penggunanaan modus tidak tepat karena fitur tidak cukup berbeda.

Handling Duplicate

No duplicate rows

Handling Outliers

- Jumlah anggota keluarga (Family member count)
- Pendapatan (Income)
- Lama bekerja (Employment length)

Menggunakan Interquartile Range (IQR) untuk mendeteksi outlier.
Data yang berada di luar 3 × IQR dari kuartil 25% dan 75% akan dihapus.

Data Manipulation

Merging Dataset

Menggabungkan **application_record.csv** dan **credit_record.csv** menggunakan kolom ID.

Label Creation (is high risk? Yes or No)

Mengonversi kolom STATUS menjadi label biner:

- **Yes (1)** → Status mengandung **2, 3, 4, 5** (tunggakan **≥ 60 hari**).
- **No (0)** → Status **0,1, C, X** (lancar/tunggakan tidak melebihi batas toleransi perusahaan)

Transformasi umur & employment length

Encoding variabel kategorikal

Feature scaling untuk model tertentu

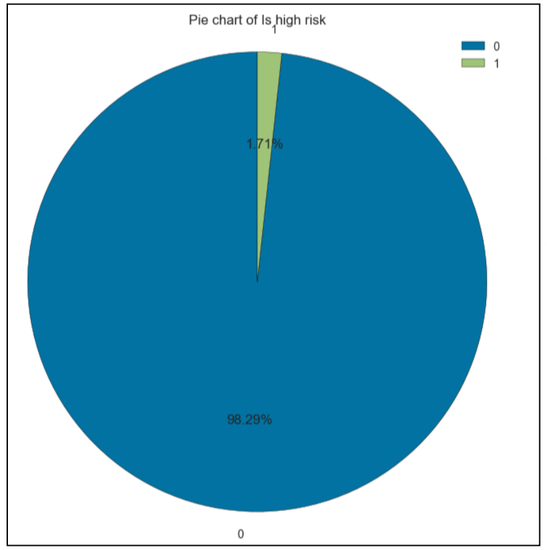
Penyesuaian struktur data untuk modeling

Exploratory Data Analysis

“Data Demografi Nasabah”

Demografi nasabah : Gender dan Education level

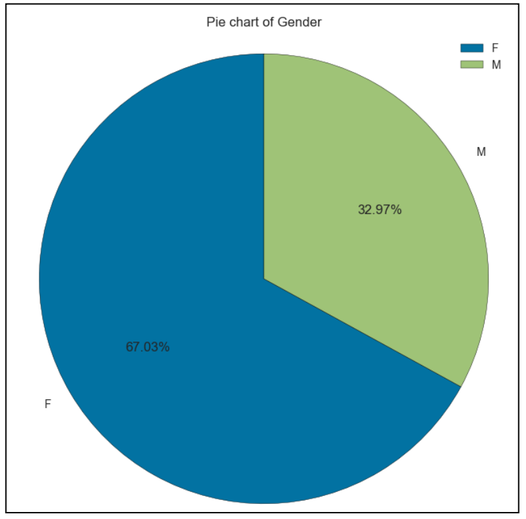
Label / Target



Distribusi Label (Target)

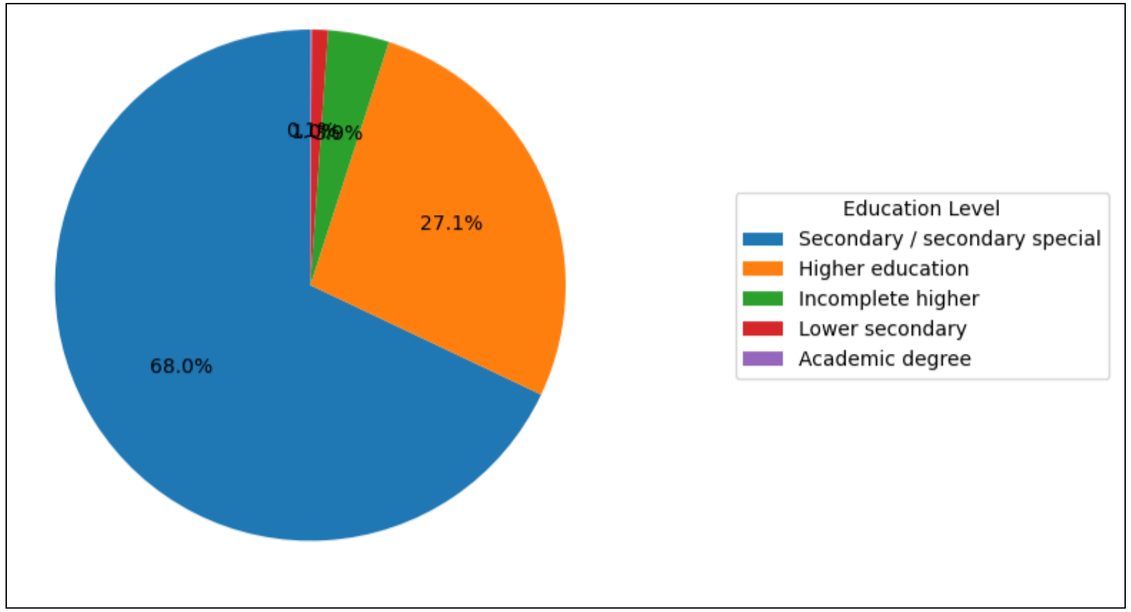
- Mayoritas nasabah memiliki status **Good** (**98,29%**).
- Hanya **1,71%** nasabah yang tercatat **Bad**, menunjukkan data sangat tidak seimbang.

Distribusi Gender



Distribusi Gender Nasabah
Laki-laki : 33%
Perempuan : 67%
Ini menunjukkan bahwa perempuan lebih sering mengajukan kartu kredit dalam dataset.

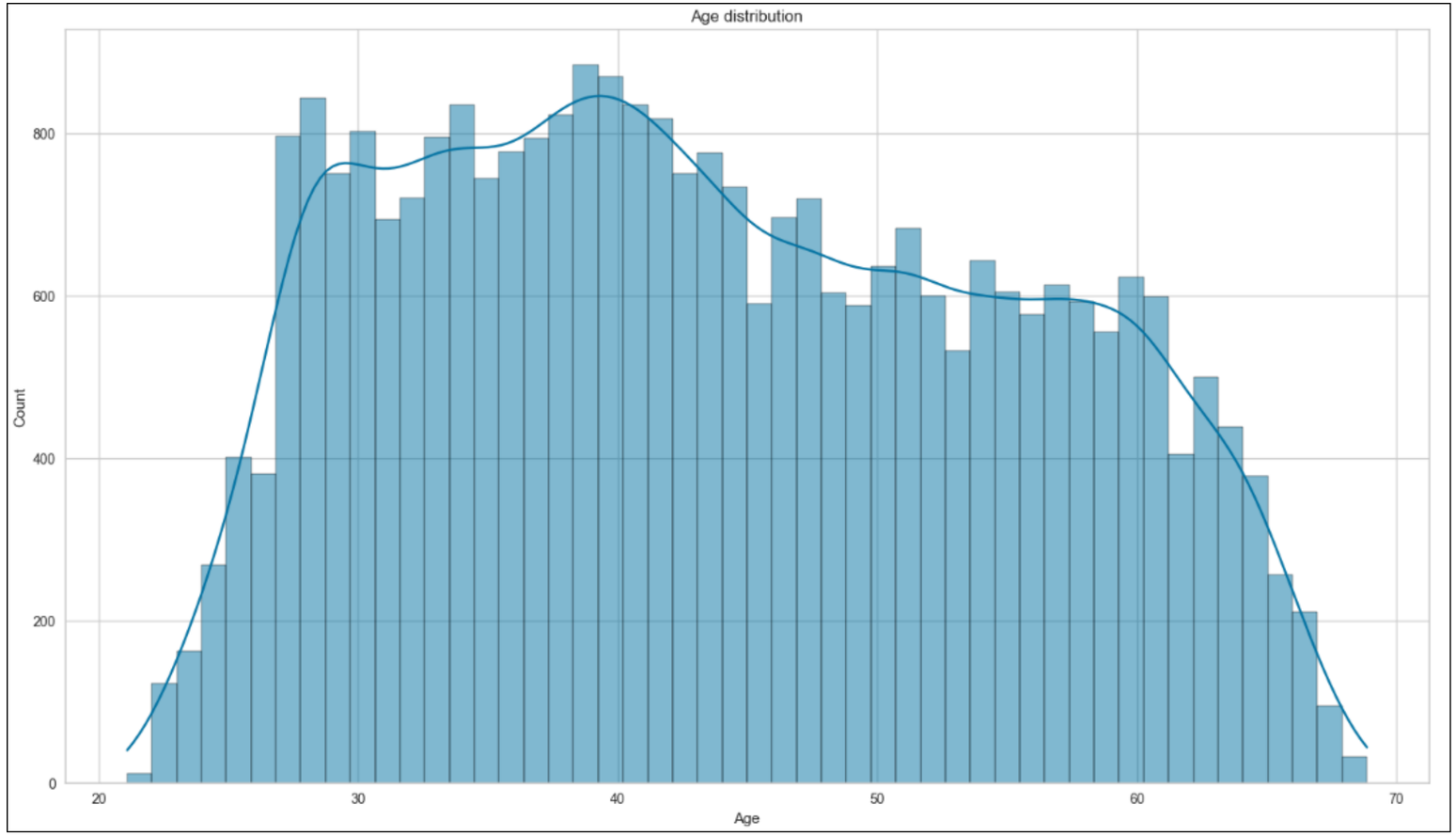
Distribusi Education level

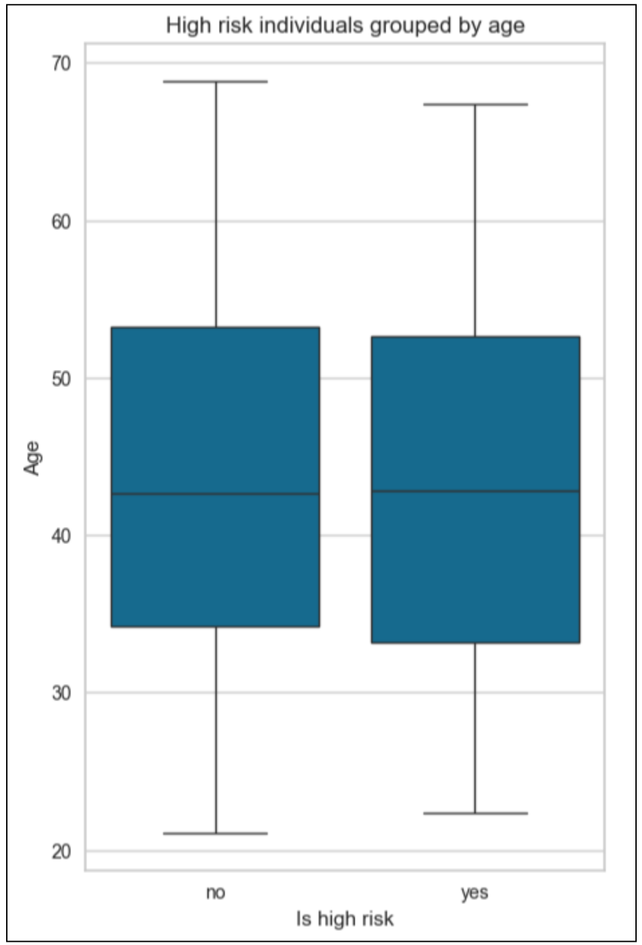


lebih banyak dari nasabah yang lulusan setaran SMA (68%), diikuti oleh yang berpendidikan setara sarjana (27,1%)

Demografi Nasabah : Age

Age nasabah



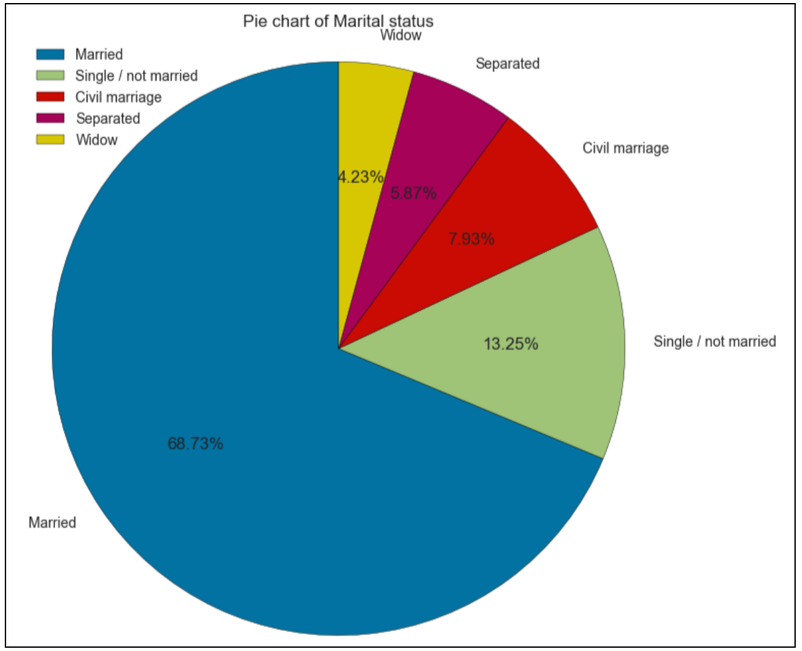


Distribusi Usia Nasabah

- Nasabah termuda berusia 21 tahun, sedangkan yang tertua berusia 68 tahun, dengan rata-rata usia 43,7 tahun dan median 42,6 tahun (median tahan terhadap outlier).
- Fitur usia tidak terdistribusi normal, cenderung sedikit miring ke kanan (positively skewed).
- Tidak terdapat perbedaan yang signifikan antara rata-rata usia nasabah berisiko tinggi dan rendah.

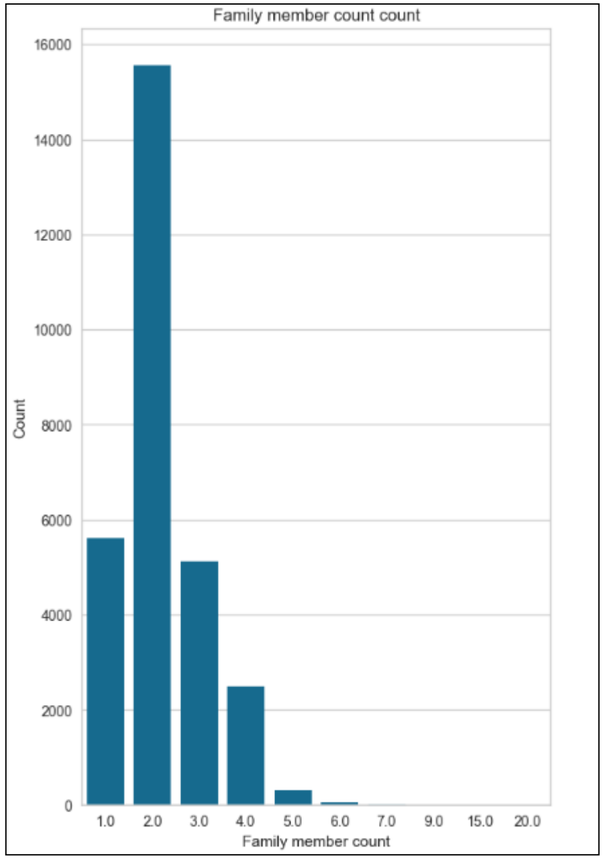
Demografi Nasabah : marital status, Family member and Children Count

Marital Status

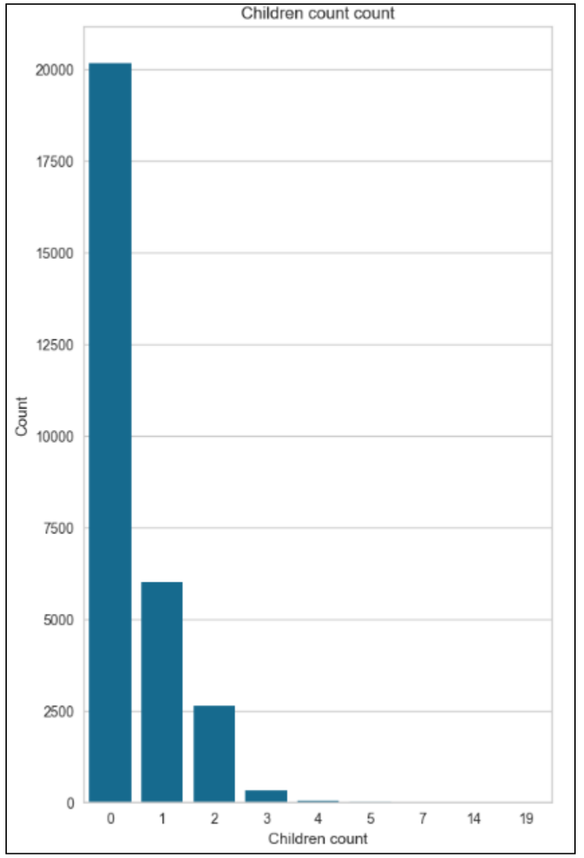


Secara garis besar nasabah pada data kali ini **sudah menikah**

Family Member



Children Count



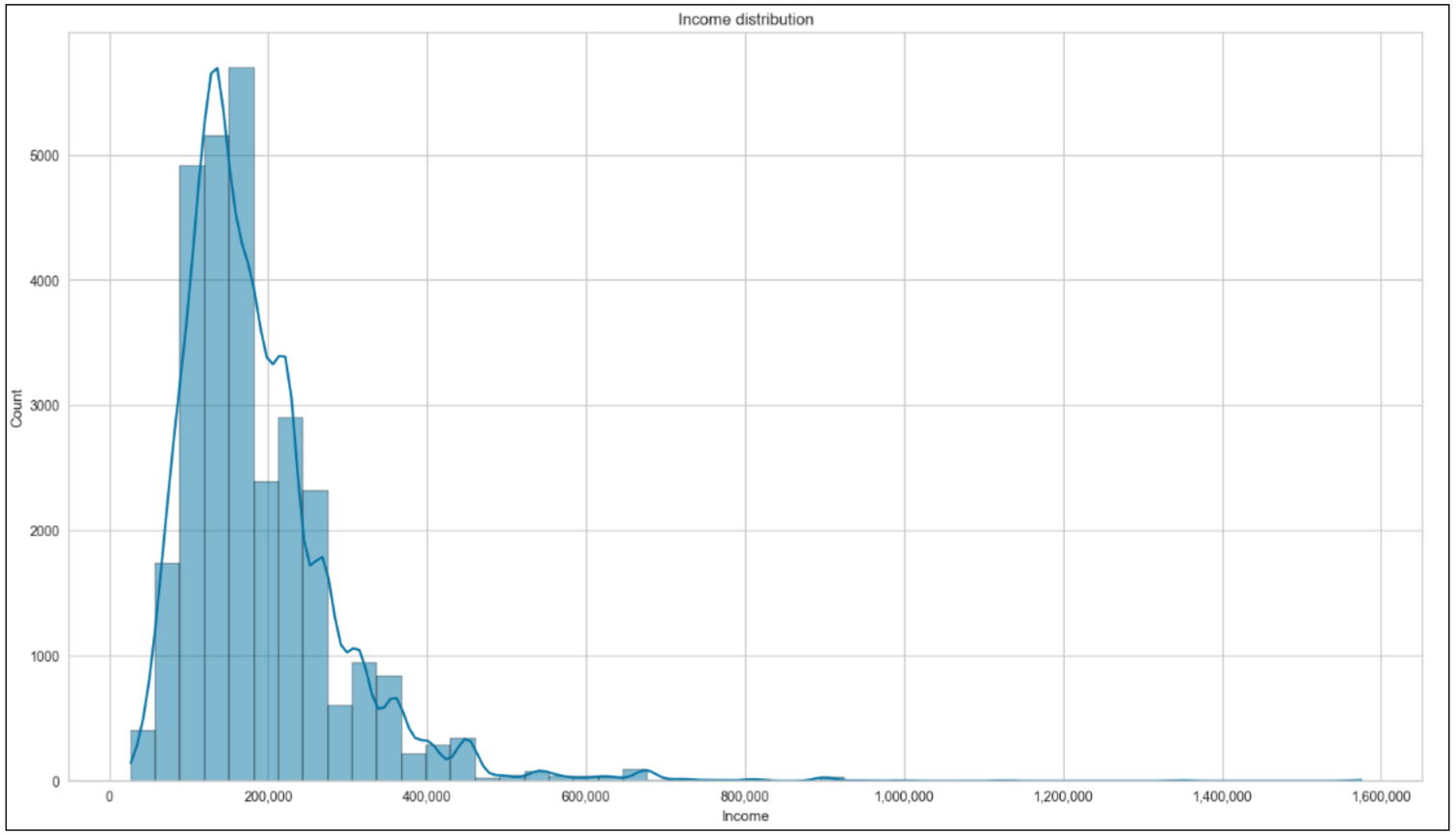
Sebagian besar nasabah **memiliki dua anggota** dalam rumah tangga, yang juga sejalan dengan fakta bahwa **kebanyakan tidak memiliki anak.**

Exploratory Data Analysis

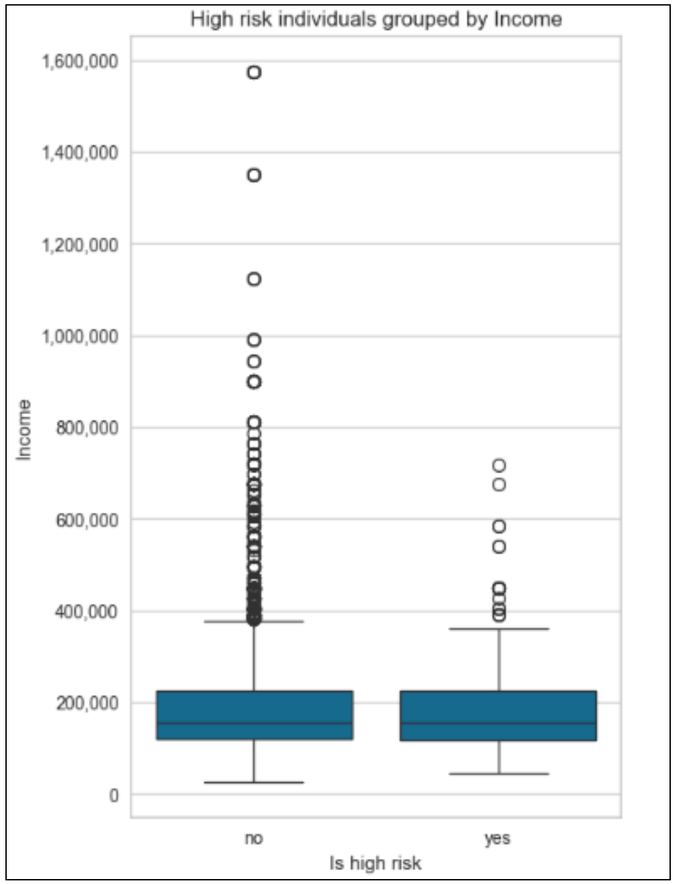
“Kondisi Ekonomi & Pendapatan”

Kondisi Ekonomi & Pendapatan : Total Income

Distribusi Pendapatan



Distribusi Pendapatan dan Risiko

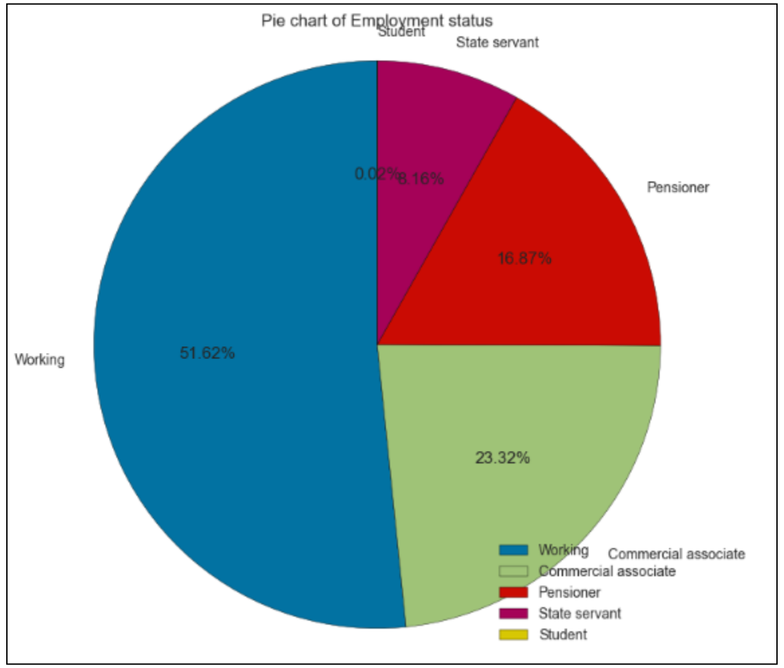


Distribusi Pendapatan dan Risiko

- Rata-rata pendapatan nasabah adalah 186.890, namun angka ini dipengaruhi oleh outlier. Jika outlier diabaikan, sebagian besar nasabah memiliki pendapatan sekitar 157.500.
- Terdapat 3 nasabah dengan pendapatan lebih dari 1.000.000.
- Fitur pendapatan cenderung miring ke kanan (positively skewed).
- Nasabah berisiko tinggi dan rendah memiliki pendapatan yang kurang lebih sama.

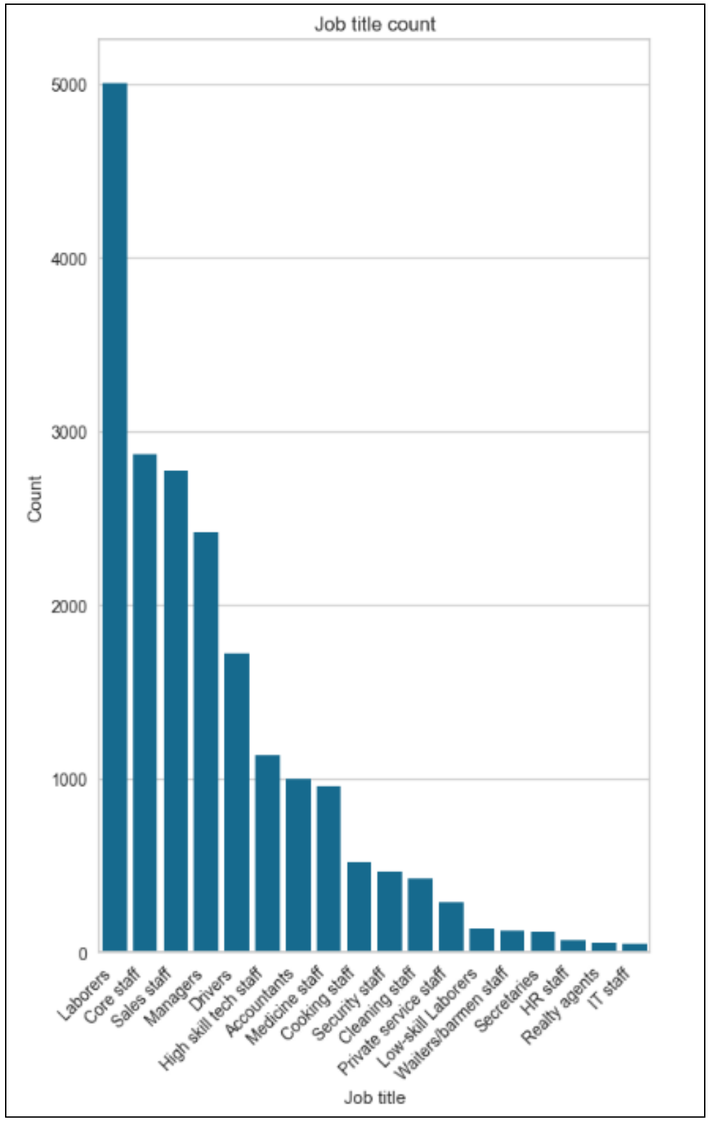
Kondisi Ekonomi & Pendapatan : Employment status & Tipe pekerjaan

Employment status



- *Sebagian besar nasabah sedang bekerja atau memiliki pekerjaan.*
- *Pekerjaan nasabah paling banyak adalah **Laborers** atau pekerja fisik*

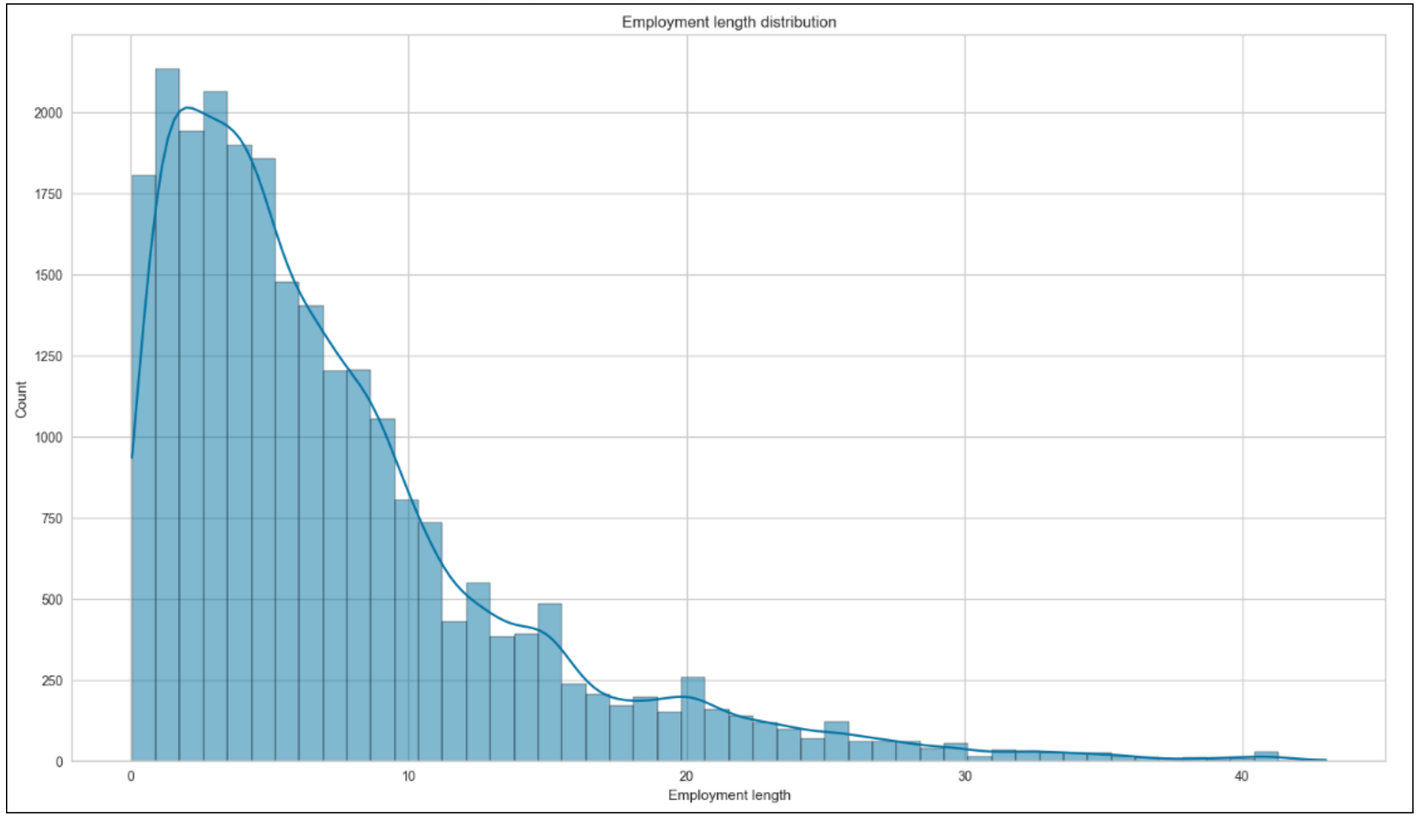
Distribusi Pekerjaan



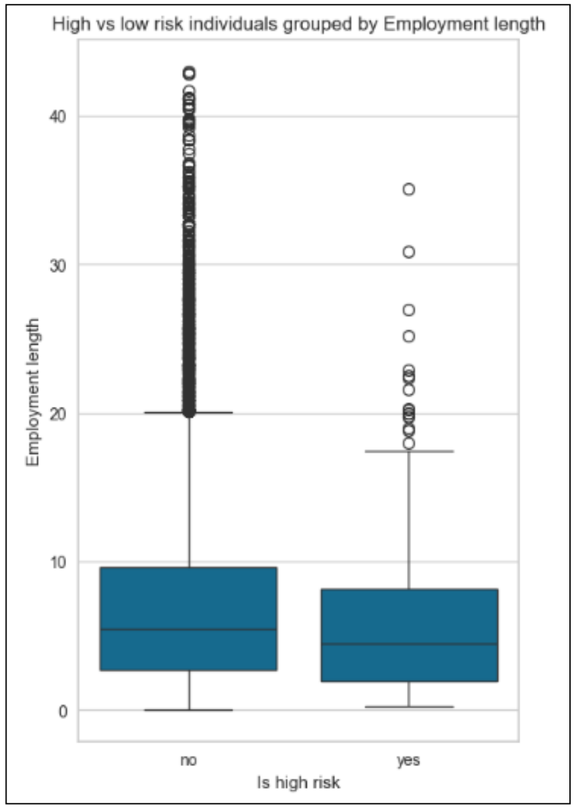
Job Title	Count	Frequency (%)
Laborers	5,004	24.85
Core staff	2,866	14.23
Sales staff	2,773	13.77
Managers	2,422	12.03
Drivers	1,722	8.55
High skill tech staff	1,133	5.63
Accountants	998	4.96
Medicine staff	956	4.75
Cooking staff	521	2.59
Security staff	464	2.30
Cleaning staff	425	2.11
Private service staff	287	1.43
Low-skill Laborers	138	0.69
Waiters/barmen staff	127	0.63
Secretaries	122	0.61
HR staff	72	0.36
Realty agents	60	0.30
IT staff	48	0.24

Kondisi Ekonomi & Pendapatan : Lama masa pekerjaan

Employment Length / lama masa bekerja



Employment Length by target



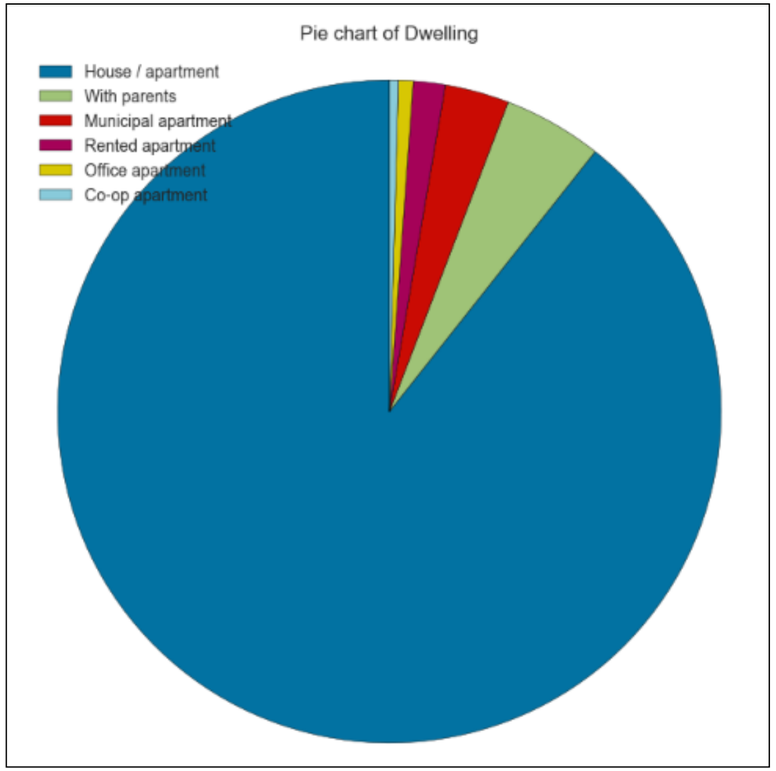
- Sebagian besar nasabah rata-rata telah bekerja selama **5 hingga 7 tahun.**
- Terdapat sejumlah **outlier yang telah bekerja lebih dari 20 tahun.**
- Lama bekerja cenderung miring ke kanan (positively skewed).
- Nasabah yang **berisiko tinggi** memiliki **lama bekerja** lebih rendah, yaitu **sekitar 5 tahun**, dibandingkan (7 tahun) pada nasabah berisiko rendah.

Exploratory Data Analysis

“Kepemilikan aset & Tempat tinggal”

Kepemilikan aset & Tempat tinggal : Properti dan Tipe tempat tinggal

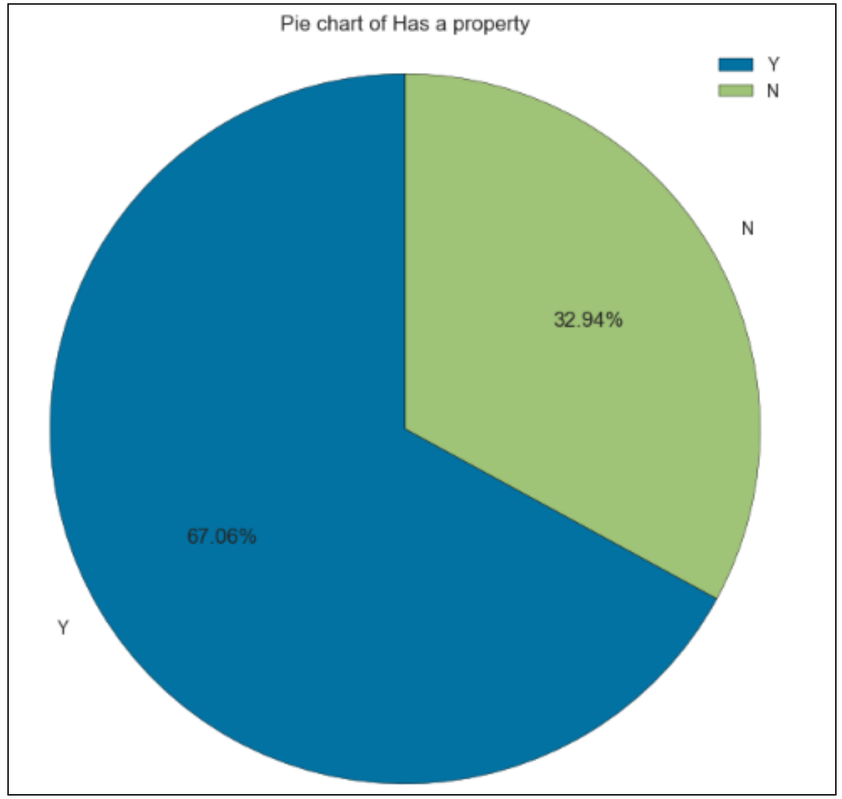
Dwelling type



Dwelling Type	Count	(%)
House / apartment	26,059	89.35
With parents	1,406	4.82
Municipal apartment	912	3.13
Rented apartment	453	1.55
Office apartment	208	0.71
Co-op apartment	127	0.44

- 89,35% nasabah tinggal di house / apartment, menunjukkan stabilitas tempat tinggal yang tinggi
- Nasabah yang tinggal dengan orang tua atau menyewa memiliki proporsi kecil dan merepresentasikan stabilitas ekonomi yang lebih rendah

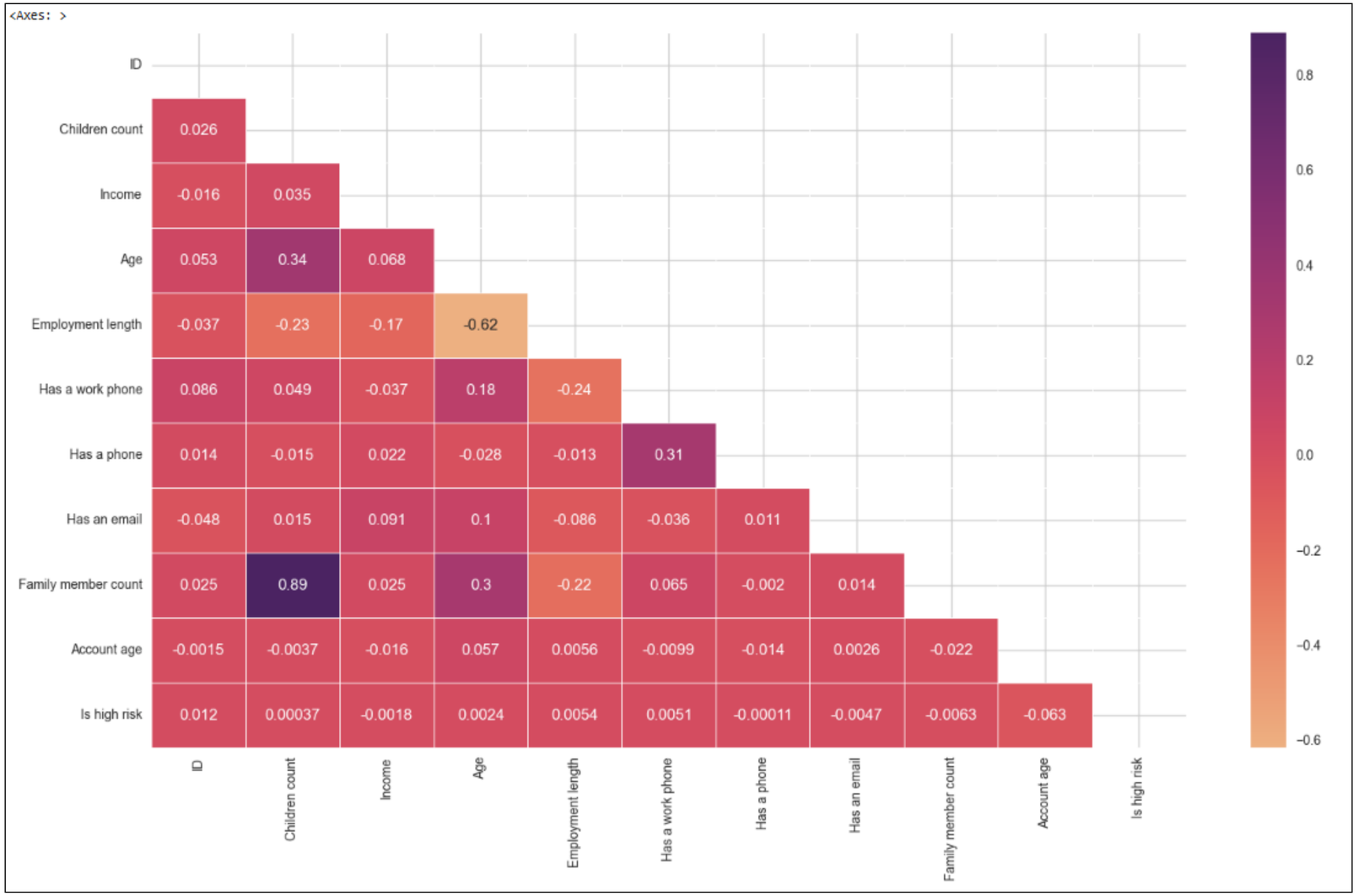
Has a property



Sebagian besar nasabah (**67,2%**) memiliki properti.

Correlation Heatmap

Correlation Heatmap



- Tidak ada fitur yang berkorelasi langsung dengan fitur target.
- Jumlah anggota keluarga sangat berkorelasi dengan jumlah anak
- Usia memiliki korelasi positif dengan jumlah anggota keluarga dan jumlah anak; semakin tua seseorang, semakin besar kemungkinan memiliki keluarga yang lebih besar.
- Korelasi positif lainnya terlihat antara memiliki telepon dan memiliki telepon kerja.
- Korelasi positif juga terlihat antara usia dan kepemilikan telepon kerja; semakin muda seseorang, semakin kecil kemungkinan memiliki telepon kerja.
- Terdapat juga korelasi negatif antara lama bekerja dan usia, seperti yang telah diamati sebelumnya.

“Feature Engineering”

Feature Engineering

Time Conversion

- Variabel **Age dan Employment Length** awalnya disimpan dalam satuan hari **(negatif**)
- Dilakukan **konversi menjadi nilai absolut** agar lebih mudah diinterpretasikan
- Tujuan: meningkatkan keterbacaan fitur dan konsistensi data numerik

Employment Length

- Nilai **365243** pada Employment Length merepresentasikan **nasabah pensiunan**
- Nilai tersebut diubah menjadi **0 tahun masa kerja**
- Tujuan: menghindari nilai ekstrem yang dapat mengganggu proses modeling

Skewness Handling

- Variabel **Income dan Age** memiliki distribusi miring (skewed)
- Dilakukan **cubic root transformation** untuk menormalkan distribusi
- Tujuan: mengurangi pengaruh outlier dan meningkatkan performa model

Binning (Numerical to Categorical)

- Variabel biner numerik:
  - Has a work phone
  - Has a phone
  - Has an email
- Diubah dari **0/1 → N/Y**
- Tujuan: meningkatkan interpretabilitas dan konsistensi encoding kategorikal

One-Hot Encoding

- Diterapkan pada fitur kategorikal nominal:
  - **Gender**
  - **Marital status**
  - **Dwelling**
  - **Employment status**
  - **Has a car**
  - **Has a property**
  - **Has a work phone**
  - **Has a phone**
  - **Has an email**
- Menggunakan **OneHotEncoder** dengan feature names
- Tujuan: mencegah asumsi urutan dan menjaga interpretabilitas fitur

Feature Engineering

Ordinal Encoding

- Diterapkan pada **Education level**
- Menggunakan **Ordinal Encoder** sesuai dengan urutan tingkat pendidikan
- Tujuan: mempertahankan makna hierarki pendidikan

Feature Scaling (Min–Max Scaling)

- Diterapkan pada fitur numerik utama:
  - **Age**
  - **Income**
  - **Employment Length**
- Menggunakan **Min–Max Scaling (0–1)**
- Tujuan: menyamakan skala fitur dan mendukung performa model

Target Variable Transformation

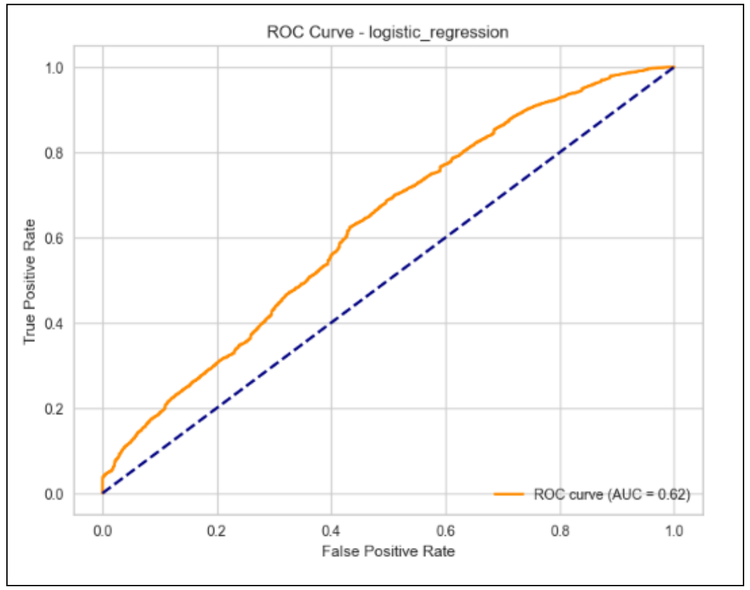
- Target **Is high risk** dikonversi menjadi **numerik**
- Tujuan: memastikan kompatibilitas dengan algoritma machine learning

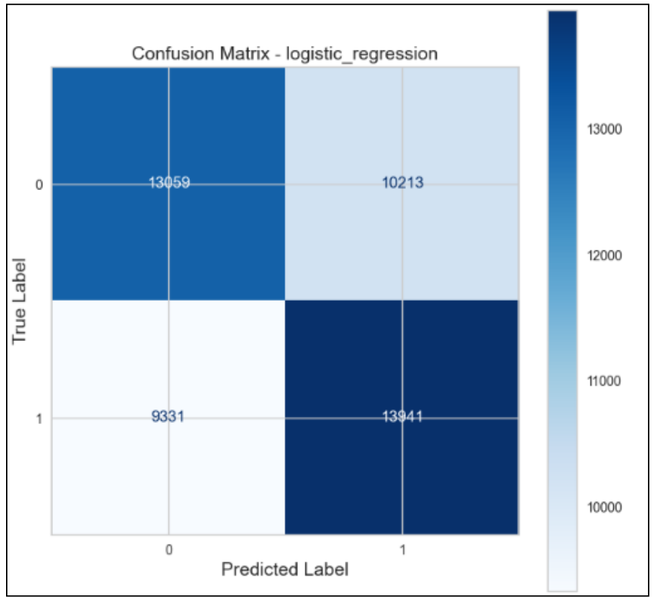
Oversampling (Handling Imbalanced Data)

- Data target mengalami **class imbalance**
- Diterapkan **SMOTE** untuk menyeimbangkan kelas minoritas
- Tujuan utama: meningkatkan **recall kelas high risk** dan meminimalkan false negative

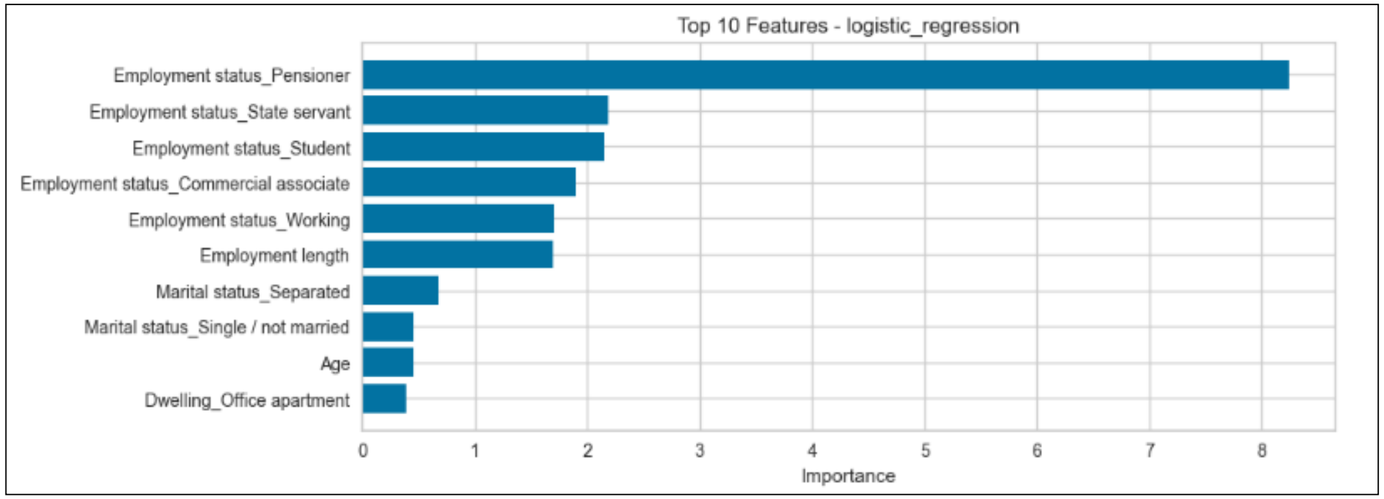
“Modeling”

Models Training : Logistic Regression





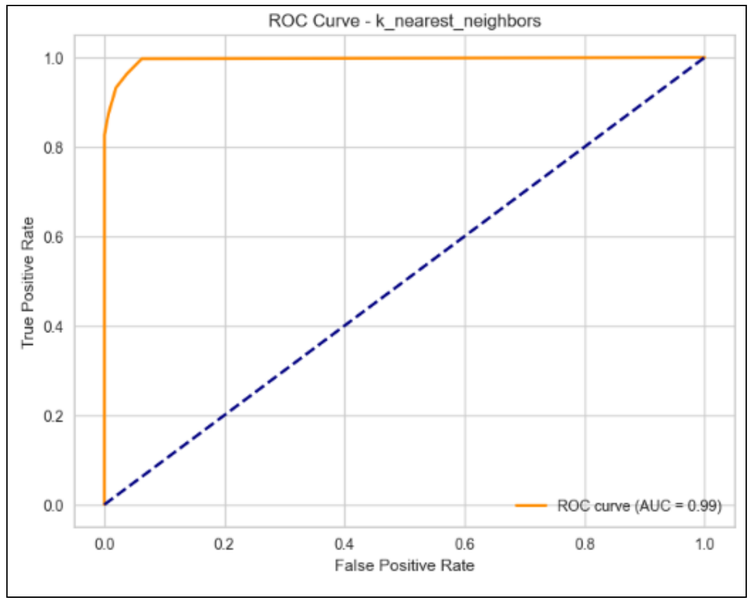
Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.58	0.56	0.57	23,272
1 (High Risk)	0.58	0.60	0.59	23,272
Accuracy			**0.58**	**46,544**
Macro Avg	0.58	0.58	0.58	46,544
Weighted Avg	0.58	0.58	0.58	46,544

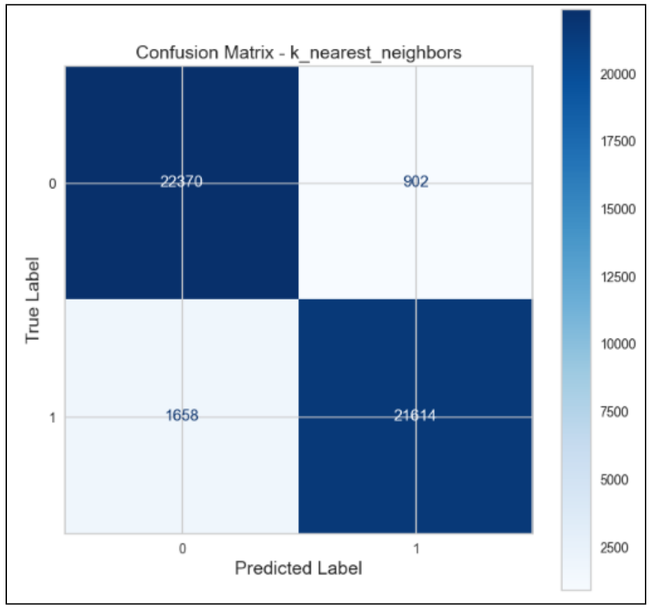


Kurva ROC menunjukkan **AUC sekitar 0,62**, yang berarti kemampuan model membedakan nasabah berisiko dan tidak berisiko masih dalam kategori rendah–menengah

Recall untuk model ini rendah untuk kategori **high risk (0.60) rendah** dengan akurasi (0,58)

Models Training : K-Nearest Neighbors



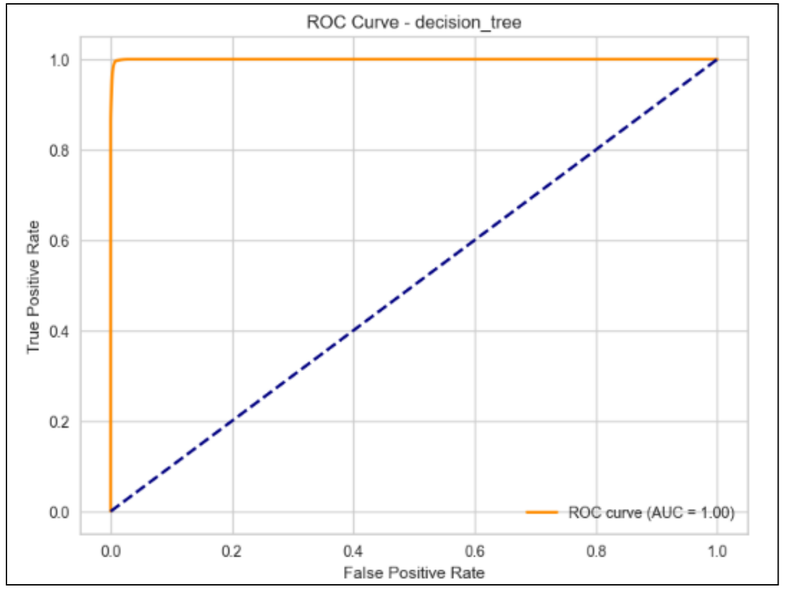


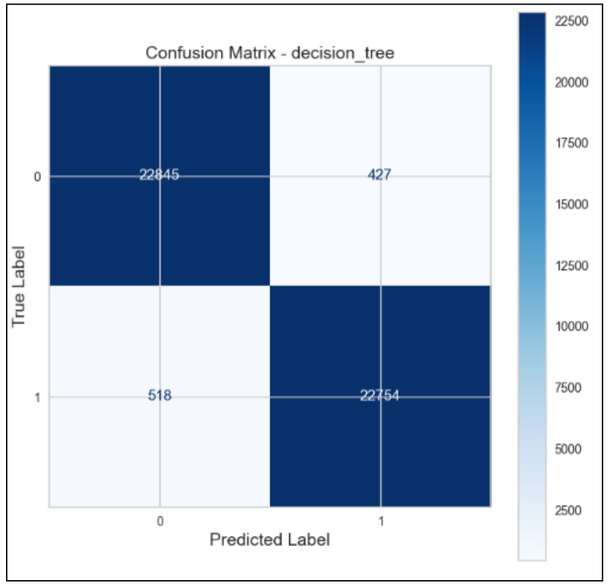
Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.93	0.96	0.95	23,272
1 (High Risk)	0.96	0.93	0.94	23,272
Accuracy			**0.94**	**46,544**
Macro Avg	0.95	0.94	0.94	46,544
Weighted Avg	0.95	0.94	0.94	46,544

Kurva ROC menunjukkan **AUC sekitar 0,99**, yang berarti kemampuan model membedakan nasabah berisiko dan tidak berisiko sangat sempurna.

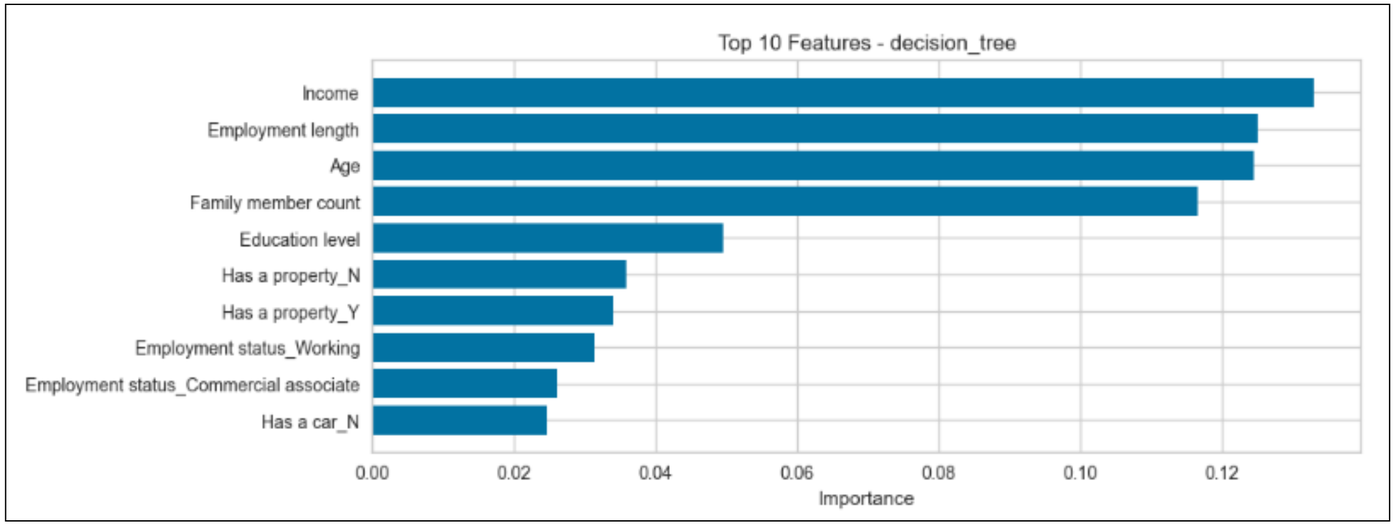
Recall untuk model ini rendah untuk kategori **high risk (0.93) rendah** dengan akurasi (0,94)

Models Training : Decission Tree





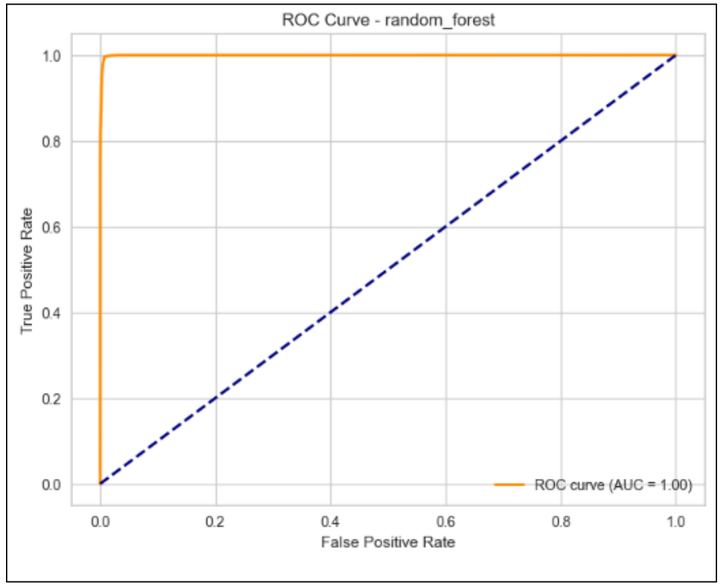
Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.98	0.98	0.98	23,272
1 (High Risk)	0.98	0.98	0.98	23,272
Accuracy			**0.98**	**46,544**
Macro Avg	0.98	0.98	0.98	46,544
Weighted Avg	0.98	0.98	0.98	46,544

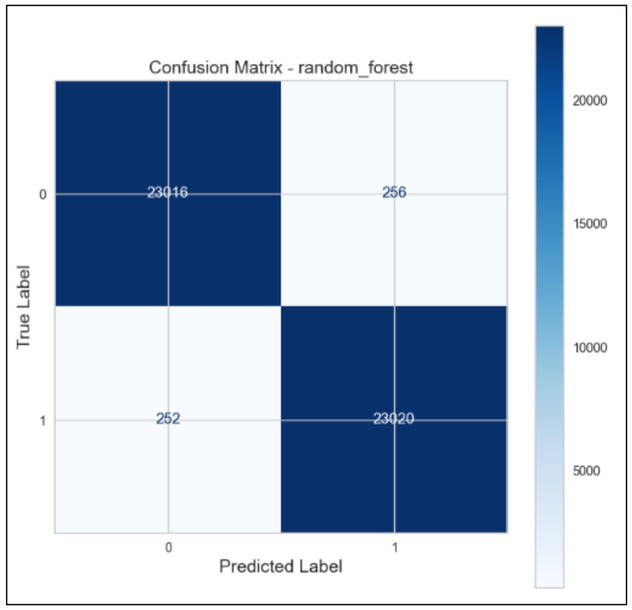


Kurva ROC menunjukkan **AUC sekitar 1.00**, yang berarti kemampuan model membedakan nasabah berisiko dan tidak berisiko sangat sempurna, namun sangat mungkin bisa overfiting.

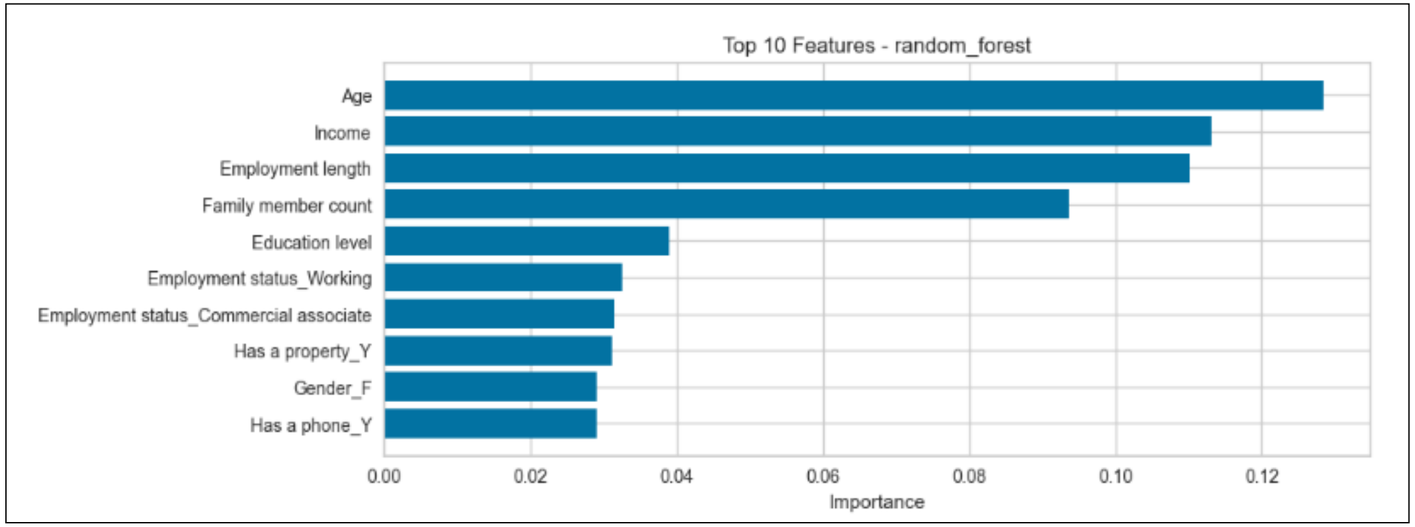
Recall untuk model ini rendah untuk kategori **high risk (0.98) tinggi** dengan akurasi (0,98)

Models Training : Random Forest





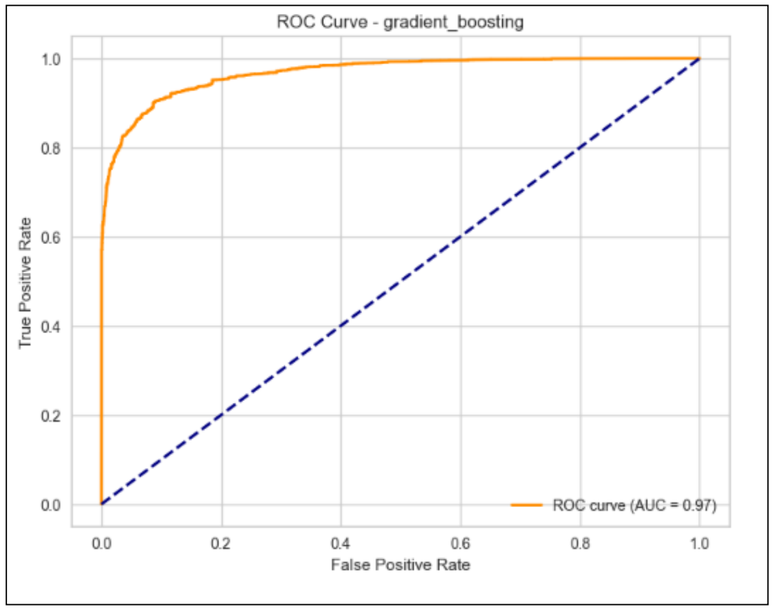
Class	Precision	Recall	F1-Score	Support
Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.99	0.99	0.99	23,272
1 (High Risk)	0.99	0.99	0.99	23,272
Accuracy			**0.99**	**46,544**
Macro Avg	0.99	0.99	0.99	46,544

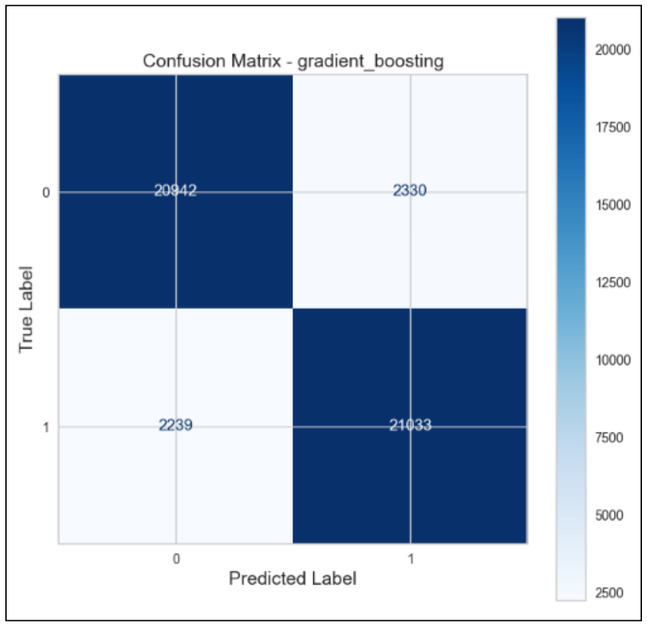


Kurva ROC menunjukkan **AUC sekitar 1.00**, yang berarti kemampuan model membedakan nasabah berisiko dan tidak berisiko sangat sempurna, namun sangat mungkin bisa overfiting.

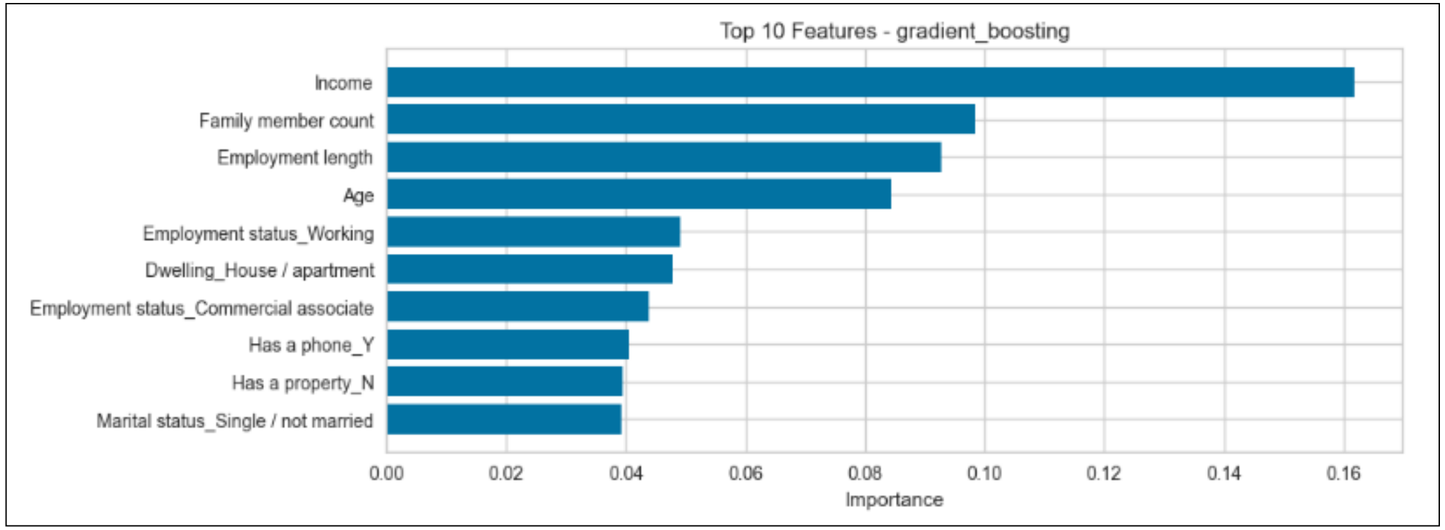
Recall untuk model ini rendah untuk kategori **high risk (0.99) tinggi** dengan akurasi (0,99).

Models Training : Gradient Boosting



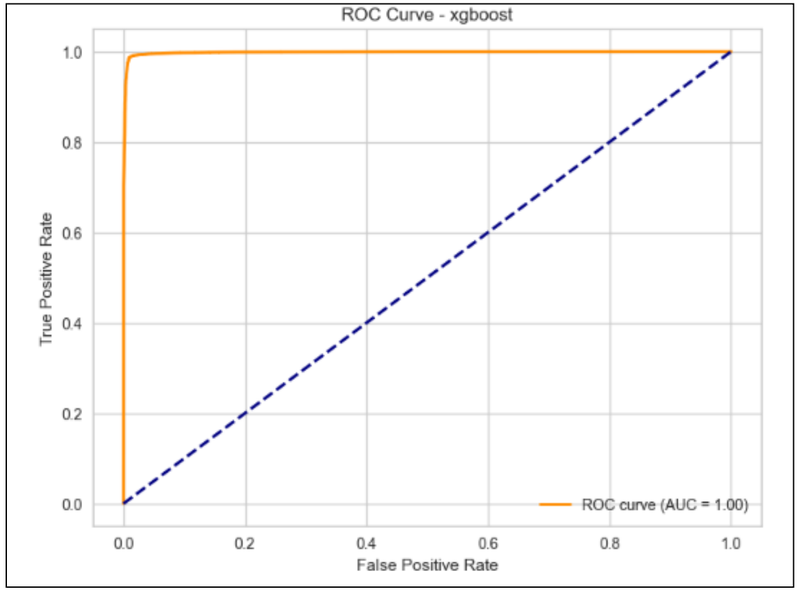


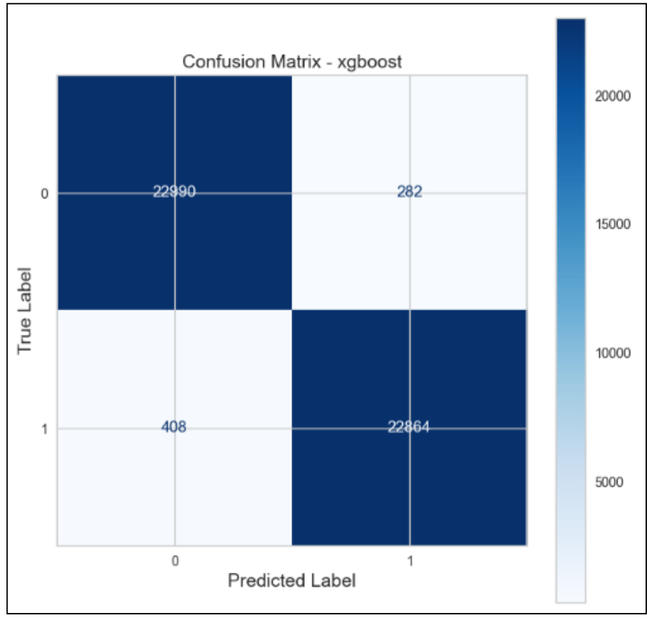
Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.90	0.90	0.90	23,272
1 (High Risk)	0.90	0.90	0.90	23,272
Accuracy			**0.90**	**46,544**
Macro Avg	0.90	0.90	0.90	46,544
Weighted Avg	0.90	0.90	0.90	46,544



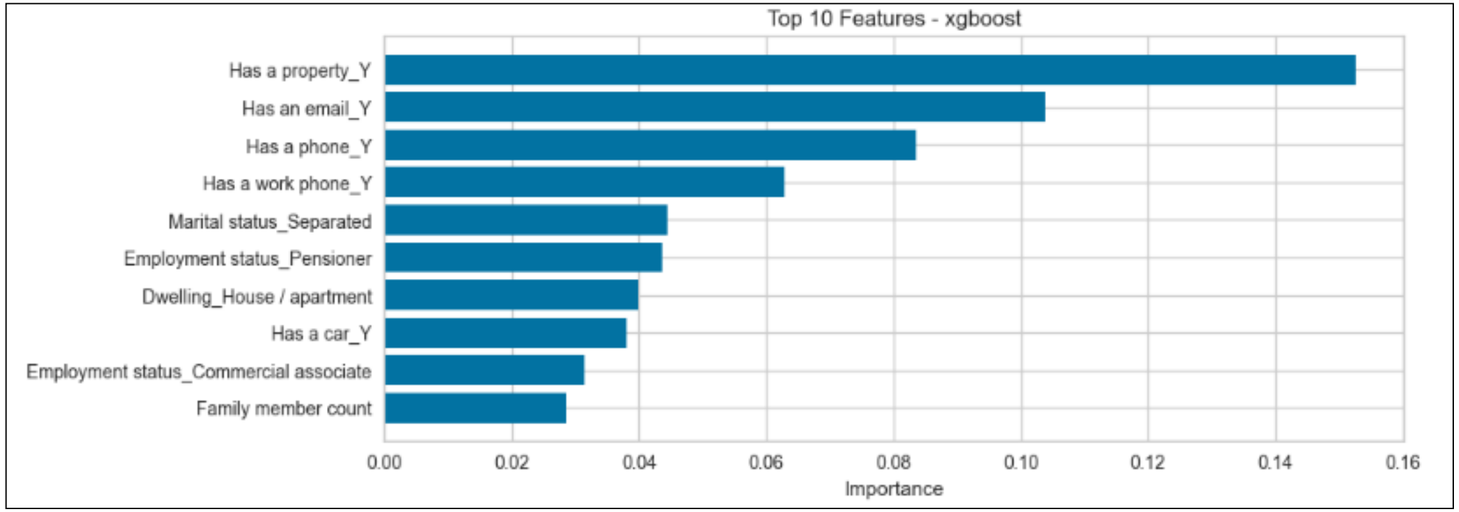
- Gradient Boosting memberikan **keseimbangan yang baik** antara precision dan recall
- Performa **lebih realistis dan stabil** dibandingkan model dengan skor mendekati sempurna
- Recall kelas **high risk (0.90)** sudah cukup tinggi untuk kebutuhan mitigasi risiko

Models Training : XGBoost





Class	Precision	Recall	F1-Score	Support
0 (Not High Risk)	0.98	0.99	0.99	23,272
1 (High Risk)	0.99	0.98	0.99	23,272
Accuracy			**0.99**	**46,544**
Macro Avg	0.99	0.99	0.99	46,544
Weighted Avg	0.99	0.99	0.99	46,544



Kurva ROC menunjukkan **AUC sekitar 1.00**, yang berarti kemampuan model membedakan nasabah berisiko dan tidak berisiko sangat sempurna. Indikasi overfiting.

Recall untuk model ini rendah untuk kategori **high risk (0.98) rendah** dengan akurasi (0,99)

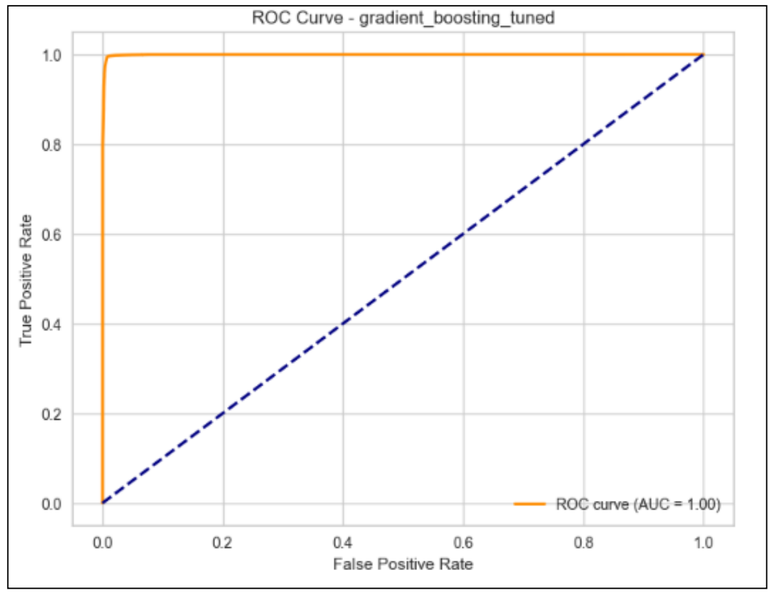
Pemilihan Top model

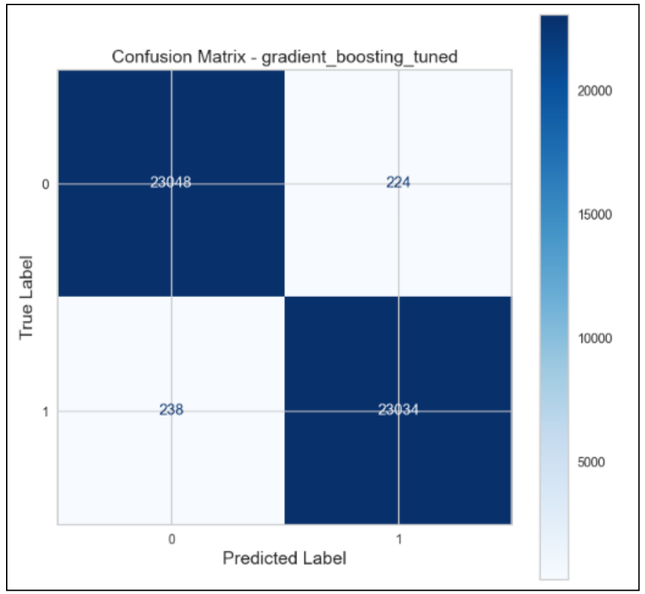
Model	Accuracy	Recall (Class 1)	ROC–AUC	Keterangan
Logistic Regression	0.58	0.60	0.62	Model awal tanpa optimasi
K-Nearest Neighbors	0.94	0.93	0.99	Sensitif skala, performa baik
Decision Tree	0.98	0.98	1.00	Interpretatif, rawan overfitting
Random Forest	0.99	0.99	0.99	Akurat, kompleks
XGBoost	0.99	0.98	1.00	Performa sangat tinggi, indikasi overfitting
Gradient Boosting	**0.90**	**0.90**	**0.97**	**Seimbang & model final**

- Model Gradient Boosting Alasan Utama: **Model Gradient Boosting** memiliki metrik performa yang sangat baik (ROC 0.97) dan Feature Importance yang lebih masuk akal secara bisnis/statistik. Fitur-fitur seperti Pendapatan dan Lama Bekerja adalah prediktor yang solid dan dapat diinterpretasikan.
- Model XGBoost, KNN, Decission tree & Ranomforest Alasan Utama: Meskipun metriknya hampir sempurna (ROC 1.00), ini adalah red flag.
- Overfitting: **ROC AUC 1.00 hampir pasti menunjukkan model overfit pada data latih** dan kemungkinan akan berkinerja buruk pada data baru di dunia nyata.
- contoh pada model xgboost Data Leakage: Ada kemungkinan kebocoran data di mana fitur Has a property_Y secara langsung atau tidak langsung terkait dengan variabel target.
- **Kesimpulan Akhir:** Pilih Gradient Boosting. **Metriknya kuat dan Feature Importance-nya dapat dipercaya**, menjadikannya model yang lebih andal dan dapat diinterpretasikan untuk produksi.

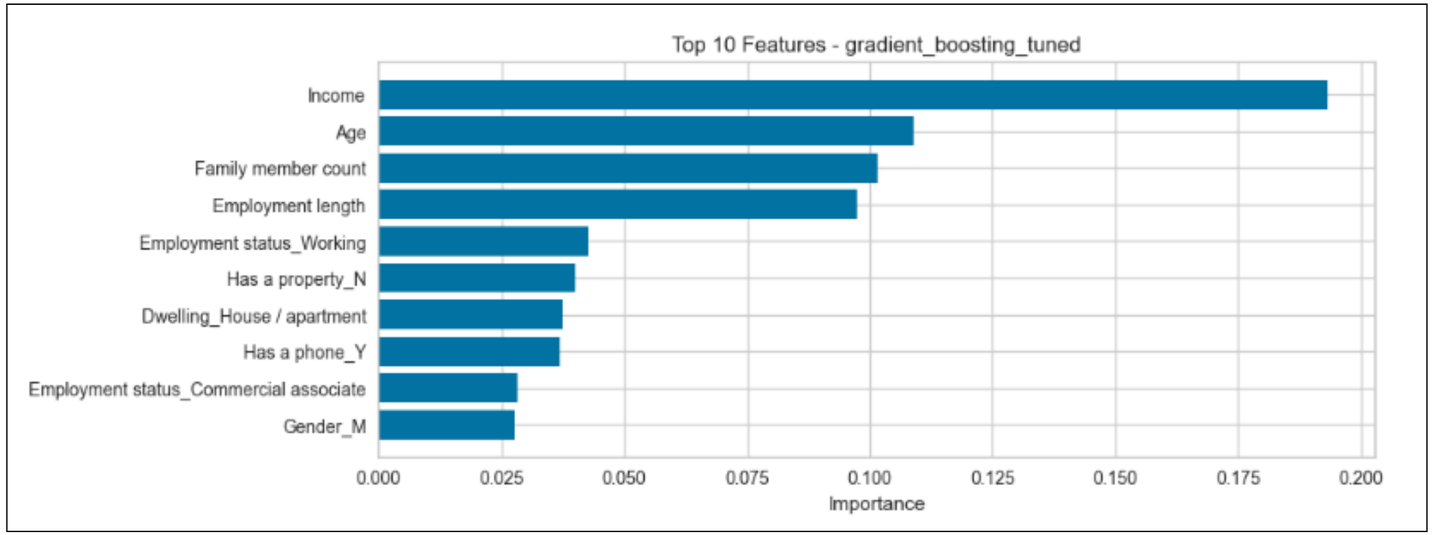
“Hyperparameter Tuning untuk Best model”

Models Training Gradient Boosting tuned & Evaluation recall Test





Class	Precision	Recall	F1-Score	Support
0 (Low Risk)	0.99	0.99	0.99	23.272
1 (High Risk)	0.99	0.99	0.99	23.272
Accuracy			**0.99**	46.544
Macro Avg	0.99	0.99	0.99	46.544
Weighted Avg	0.99	0.99	0.99	46.544



- **Recall test (kelas 1): 0.98**
- Hasil evaluasi menunjukkan bahwa model Gradient Boosting menghasilkan recall test sebesar 0.98 pada kelas high risk, yang menandakan kemampuan generalisasi yang sangat baik dalam mengidentifikasi nasabah berisiko tinggi. Performa ini konsisten dengan hasil training dan mengindikasikan tingkat overfitting yang rendah.

Kesimpulan dan Rekomendasi

- **Kesimpulan**

- Proses persetujuan kartu kredit memiliki risiko finansial tinggi jika nasabah **berisiko gagal bayar (high risk)** tidak terdeteksi dengan baik
- Model **Gradient Boosting** yang dikembangkan mampu mengidentifikasi **98% nasabah high risk pada data test**, menunjukkan kemampuan yang sangat baik dalam mengurangi risiko kredit macet
- Pendekatan end-to-end mulai dari **data cleaning, feature engineering, hingga model tuning** menghasilkan model yang stabil, akurat, dan dapat digeneralisasi
- Variabel seperti **usia, pendapatan, lama bekerja, status pekerjaan, dan kepemilikan aset** terbukti berperan penting dalam pengambilan keputusan kredit
- Dengan penerapan model ini, perusahaan dapat melakukan **keputusan persetujuan kredit yang lebih objektif, konsisten, dan berbasis data**

Rekomendasi

1. **Implementasikan model sebagai decision support system**, bukan pengganti penuh keputusan manusia, khususnya untuk kasus borderline

2. **Early Warning System (Post-Approval Monitoring)**
   - Gunakan fitur yang sama untuk:
     1. **monitor nasabah aktif**
     2. deteksi dini potensi default
   - Contoh trigger:
     1. penurunan pendapatan
     2. perubahan status pekerjaan

   Model jadi **preventive**, bukan hanya selektif

3. Gunakan **feature importance Gradient Boosting** untuk menyusun **credit scoring rules** meningkatkan transparansi keputusan kredit kepada stakeholder

4. **Prioritaskan recall kelas high risk** sebagai KPI utama untuk meminimalkan potensi kerugian finansial

Terima Kasih Banyak