Rakamin Academy X id/x partners

Prediction Model

ID/X Partners - Data Scientist

Presented by
Willy Agcely Heza

Jakarta Timur, Indonesia

willyagcely08@gmail.com

Willy agcely heza

Willy Agcely Heza

Data Science & Data Anlayst Entuthiast

Lulusan S1 Statistika Universitas Muhammadiyah Semarang (IPK 3,67) dengan pengalaman di Badan Pusat Statistik dan Eduwork. Terampil dalam pengolahan, analisis, dan visualisasi data menggunakan Excel, SQL, Python, Tableau, dan Looker Studio. Memiliki kemampuan komunikasi, teamwork, problem solving, serta inisiatif tinggi dalam mendukung pengolahan data dan pengambilan keputusan berbasis data.

About Company

ID/X Partners (PT IDX Consulting) didirikan pada tahun 2002 dan telah melayani perusahaan di seluruh wilayah Asia dan Australia dan di berbagai industri, khususnya layanan keuangan, telekomunikasi, manufaktur, dan ritel. ID/X Partners menyediakan layanan konsultasi yang mengkhususkan diri dalam memanfaatkan solusi data analytic and decisioning (DAD) yang dipadukan dengan manajemen risiko dan disiplin pemasaran terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis. Layanan konsultasi dan solusi teknologi yang komprehensif yang ditawarkan oleh mitra id/x menjadikannya sebagai one-stop service provider

Project Portfolio

Project ini, dalam konteks peran Data Scientist di ID/X Partners, berfokus pada pengembangan model machine learning untuk memprediksi risiko kredit pada perusahaan multifinance. Model dibangun menggunakan data pinjaman (baik atau buruk dalam meminjam) melalui tahapan Data Understanding, EDA, Data Preparation, Modelling, dan Evaluation guna meningkatkan akurasi penilaian risiko serta mengurangi potensi kerugian.

Link code here!

Link Drive here!

Project explanation video here!

Link Youtube here

1. Business Understanding

Business Objective

- Meningkatkan akurasi penilaian risiko kredit.
- Mengurangi potensi kerugian akibat kredit macet.
- Mendukung pengambilan keputusan approve/decline pinjaman secara lebih tepat.

Business Metrics

- **Default Rate** → menurunkan persentase kredit macet.
- **Approval Accuracy** → meningkatkan ketepatan persetujuan pinjaman.
- **Model Performance** → diukur dengan metrik machine learning seperti AUC, Precision, Recall, dan F1-Score untuk klasifikasi risiko kredit.

2. Data Understanding

Dataset Overview:

- Data pinjaman 2007–2014.
- Target: `loan_status` (baik / buruk).
- Variabel: informasi pinjaman, profil peminjam, histori kredit.

Data Exploration Fokus:

- Identifikasi variabel relevan (numerik & kategorikal).
- Cek kualitas data (missing values, distribusi, outlier).
- Analisis awal hubungan fitur dengan `loan_status`.

Output

- Dataset terpilih & bersih, siap dipakai untuk Feature Engineering dan Modeling.

```
print(df.head())
```

```
   loan_amnt        term  int_rate  installment grade emp_length  \
0       5000   36 months     10.65       162.87     B  10+ years
1       2500   60 months     15.27        59.83     C   < 1 year
2       2400   36 months     15.96        84.33     C  10+ years
3      10000   36 months     13.49       339.31     C  10+ years
4       3000   60 months     12.69        67.79     B     1 year

  home_ownership  annual_inc verification_status  loan_status         purpose  \
0           RENT     24000.0            Verified   Fully Paid     credit_card
1           RENT     30000.0     Source Verified  Charged Off             car
2           RENT     12252.0        Not Verified   Fully Paid  small_business
3           RENT     49200.0     Source Verified   Fully Paid           other
4           RENT     80000.0     Source Verified      Current           other

     dti  delinq_2yrs  inq_last_6mths  open_acc  pub_rec  revol_bal  \
0  27.65          0.0             1.0       3.0      0.0       13648
1   1.00          0.0             5.0       3.0      0.0        1687
2   8.72          0.0             2.0       2.0      0.0        2956
3  20.00          0.0             1.0      10.0      0.0        5598
4  17.94          0.0             0.0      15.0      0.0       27783

   revol_util  total_acc
0        83.7        9.0
1         9.4        4.0
2        98.5       10.0
3        21.0       37.0
4        53.9       38.0
```

3. Feature Engineering

loan_status
Fully Paid
Charged Off
Fully Paid
Fully Paid
Current
...
Current
Charged Off
Current
Fully Paid
Current

loan_status
1
0
1
1
1
...
1
0
1
1
1

Proses Label Encoding “loan_status”:
1 = Good status (Current, Fully Paid)
0 = Bad status (Charged Off, Default, Late)

Proses Label Encoding pada kolom “grade”

grade
B
C
C
C
B
...
C
D
D
A
D

grade
2
3
3
3
2
...
3
4
4
1
4

```python
# ========================================
# 3. Feature Groups
# ========================================
num_features = [
    "loan_amnt", "term", "int_rate", "installment", "emp_length", "annual_inc",
    "dti", "delinq_2yrs", "inq_last_6mths", "open_acc", "revol_bal",
    "revol_util", "total_acc", "pub_rec", "income_to_loan_ratio", "installment_to_income", "grade"
]

cat_features = ["home_ownership", "purpose", "verification_status"]

# ========================================
# 4. Preprocessing Pipelines
# ========================================
# Numerik → imputasi median + scaling
num_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])

# Kategorikal → imputasi modus + OneHotEncoding
cat_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])

# Gabungkan
preprocessor = ColumnTransformer(
    transformers=[
        ("num", num_transformer, num_features),
        ("cat", cat_transformer, cat_features)
    ]
)
```

Proses One Hot Encoder “loan_status”:
`home_ownership", "purpose",`
`"verification_status`

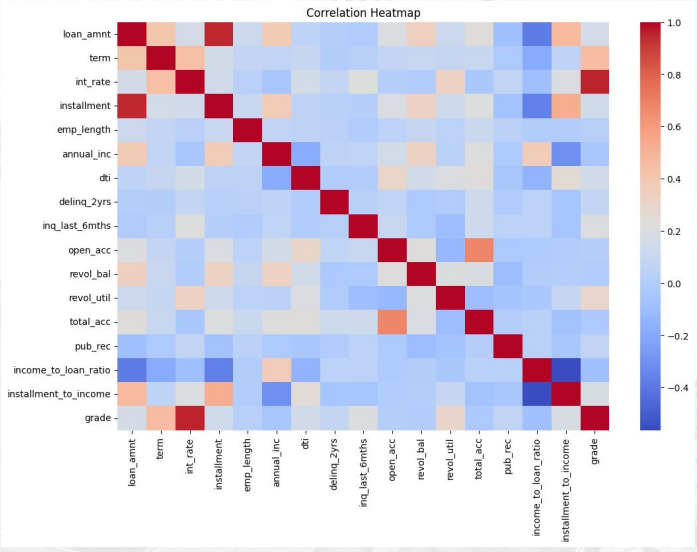
Penambahan Variabel (Feature)

- **`income_to_loan_ratio` = `annual_inc` / `loan_amnt`**
  - Mengukur seberapa besar pendapatan tahunan dibanding total pinjaman.
  - **Logika bisnisnya**:
    - Jika rasio tinggi → pendapatan jauh lebih besar dari pinjaman → kemungkinan lebih mampu melunasi (risiko rendah).
    - Jika rasio rendah → pinjaman terlalu besar dibanding pendapatan → risiko gagal bayar lebih tinggi.

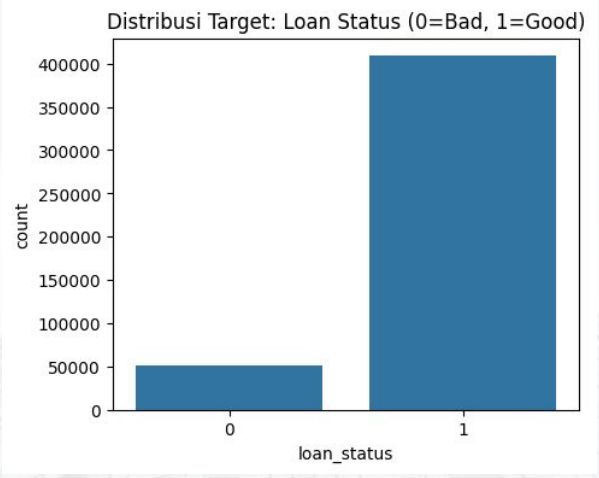
- **`installment_to_income` = `installment` / (`annual_inc` / 12)**
  - Mengukur seberapa besar cicilan bulanan dibanding gaji bulanan.
  - **Logika bisnisnya**:
    - Jika cicilan makan porsi besar dari gaji → rawan gagal bayar.
    - Jika cicilan kecil relatif terhadap gaji → lebih aman.

income_to_loan_ratio	installment_to_income
4.800000	0.081435
12.000000	0.023932
5.105000	0.082595
4.920000	0.082759
26.666667	0.010169
...	...
5.978261	0.047197
3.545455	0.089615
2.222222	0.134176
41.500000	0.009049
4.600000	0.095890

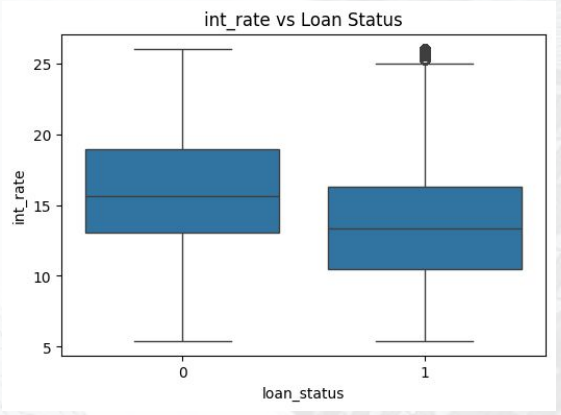
4. Exploratory Data Analysis



Terdapat korelasi kuat antara beberapa variabel utama seperti `loan_amnt`–`installment` dan `grade`–`int_rate`, sementara sebagian besar fitur lain relatif independen sehingga tetap relevan digunakan dalam prediksi risiko kredit.



Terjadi Imbalanced class



Pinjaman dengan suku bunga lebih tinggi cenderung memiliki risiko gagal bayar lebih besar.

5. Data Preparation

	loan_amnt	term	int_rate	installment	grade	emp_length	home_ownership	annual_inc	verification_status	loan_status	purpose	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_
0	5000	36 months	10.65	162.87	B	10+ years	RENT	24000.0	Verified	Fully Paid	credit_card	27.65	0.0	1.0	3.0	0.0	13648	
1	2500	60 months	15.27	59.83	C	< 1 year	RENT	30000.0	Source Verified	Charged Off	car	1.00	0.0	5.0	3.0	0.0	1687	
2	2400	36 months	15.96	84.33	C	10+ years	RENT	12252.0	Not Verified	Fully Paid	small_business	8.72	0.0	2.0	2.0	0.0	2956	
3	10000	36 months	13.49	339.31	C	10+ years	RENT	49200.0	Source Verified	Fully Paid	other	20.00	0.0	1.0	10.0	0.0	5598	
4	3000	60 months	12.69	67.79	B	1 year	RENT	80000.0	Source Verified	Current	other	17.94	0.0	0.0	15.0	0.0	27783	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
466280	18400	60 months	14.47	432.64	C	4 years	MORTGAGE	110000.0	Source Verified	Current	debt_consolidation	19.85	0.0	2.0	18.0	0.0	23208	
466281	22000	60 months	19.97	582.50	D	10+ years	MORTGAGE	78000.0	Verified	Charged Off	debt_consolidation	18.45	0.0	5.0	18.0	1.0	18238	
466282	20700	60 months	16.99	514.34	D	7 years	MORTGAGE	46000.0	Verified	Current	debt_consolidation	25.65	0.0	2.0	18.0	0.0	6688	
466283	2000	36 months	7.90	62.59	A	3 years	OWN	83000.0	Verified	Fully Paid	credit_card	5.39	3.0	1.0	21.0	0.0	11404	
466284	10000	36 months	19.20	367.58	D	10+ years	MORTGAGE	46000.0	Verified	Current	other	22.78	1.0	0.0	6.0	0.0	11325	

466285 rows × 19 columns

Proses Preprocessing Data & Handling Missing Value

	loan_amnt	term	int_rate	installment	emp_length	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	...	purpose_medical	purpose_moving	purpose_other	purpose_renewable_energy	purpose_small_
0	-1.127043	-0.616710	-0.725243	-1.108095	1.137658	-0.897908	1.328142	-0.356759	0.207709	-1.644664	...	0.0	0.0	0.0	0.0	
1	-1.428830	1.621509	0.333851	-1.531520	-1.593664	-0.788592	-2.067429	-0.356759	4.064459	-1.644664	...	0.0	0.0	0.0	0.0	
2	-1.440901	-0.616710	0.492027	-1.430842	1.137658	-1.111948	-1.083797	-0.356759	1.171896	-1.845477	...	0.0	0.0	0.0	0.0	
3	-0.523469	-0.616710	-0.074198	-0.383044	1.137658	-0.438781	0.353428	-0.356759	0.207709	-0.238969	...	0.0	0.0	1.0	0.0	
4	-1.368472	1.621509	-0.257591	-1.498810	-1.449910	0.122374	0.090956	-0.356759	-0.756478	0.765099	...	0.0	0.0	1.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
460385	0.490535	1.621509	0.150458	0.000480	-0.587387	0.668953	0.334316	-0.356759	1.171896	1.367540	...	0.0	0.0	0.0	0.0	
460386	0.925108	1.621509	1.411283	0.616305	1.137658	0.085935	0.155937	-0.356759	4.064459	1.367540	...	0.0	0.0	0.0	0.0	
460387	0.768179	1.621509	0.728145	0.336213	0.275136	-0.497083	1.073315	-0.356759	1.171896	1.367540	...	0.0	0.0	0.0	0.0	
460388	-1.489187	-0.616710	-1.355656	-1.520178	-0.874895	0.177032	-1.508084	3.411063	0.207709	1.969980	...	0.0	0.0	0.0	0.0	
460389	-0.523469	-0.616710	1.234768	-0.266873	1.137658	-0.497083	0.707638	0.899182	-0.756478	-1.042223	...	0.0	0.0	1.0	0.0	

460390 rows × 40 columns

Setelah dilakukan feature engineering

6. Data Modeling

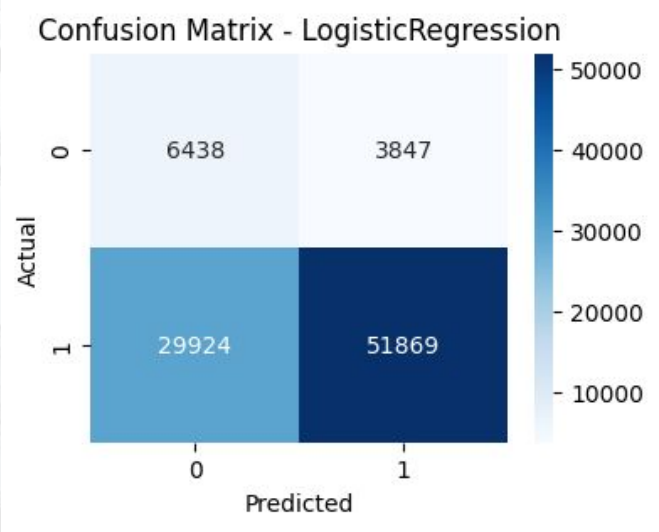
Logistic Regression → `max_iter=1000` agar model konvergen, `class_weight="balanced"` untuk mengatasi imbalance, `random_state=42` biar hasil konsisten.

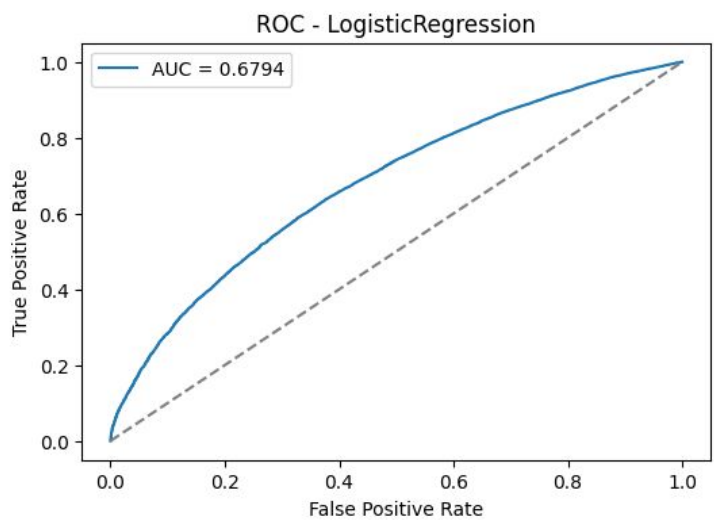
Random Forest → `n_estimators=200` jumlah pohon, `class_weight="balanced"` untuk imbalance, `random_state=42` untuk reprodusibilitas.

XGBoost → `n_estimators=200` jumlah boosting round, `scale_pos_weight=neg/pos` untuk atasi imbalance, `eval_metric="logloss"` metrik evaluasi, `random_state=42` untuk konsistensi, `use_label_encoder=False` menonaktifkan warning versi lama.

```
train pos/neg: 327172 41140 scale_pos_weight: 0.12574425684349516

Training LogisticRegression ...
LogisticRegression -- Acc:0.6332 Prec:0.9310 Rec:0.6341 F1:0.7544 AUC:0.6794

Training RandomForest ...
RandomForest -- Acc:0.8883 Prec:0.8883 Rec:1.0000 F1:0.9408 AUC:0.6690

Training XGBoost ...
XGBoost -- Acc:0.6575 Prec:0.9284 Rec:0.6658 F1:0.7755 AUC:0.6798
```

7. Evaluation



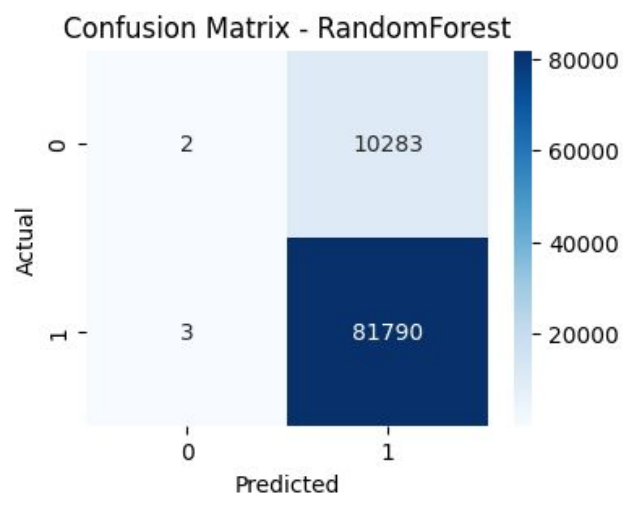


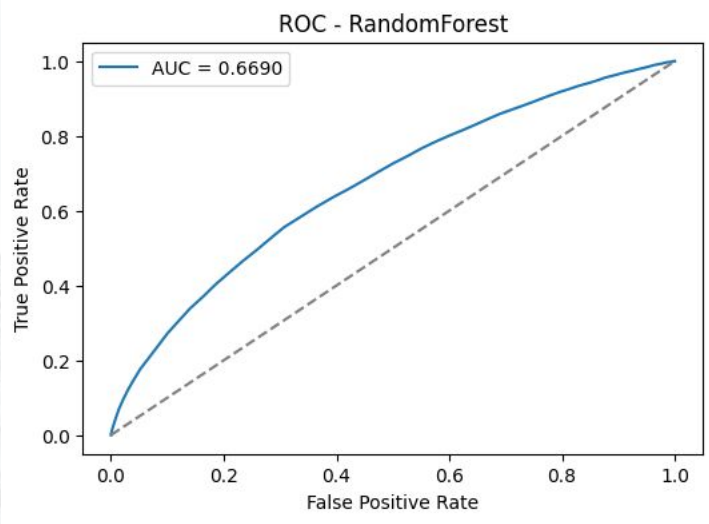
Model: LogisticRegression

	precision	recall	f1-score	support
0	0.1771	0.6260	0.2760	10285
1	0.9310	0.6341	0.7544	81793
accuracy			0.6332	92078
macro avg	0.5540	0.6301	0.5152	92078
weighted avg	0.8467	0.6332	0.7010	92078

Regression cukup baik sebagai baseline model, **tetapi performa masih bisa ditingkatkan dengan model lain atau optimasi lebih lanjut.**

Evaluation



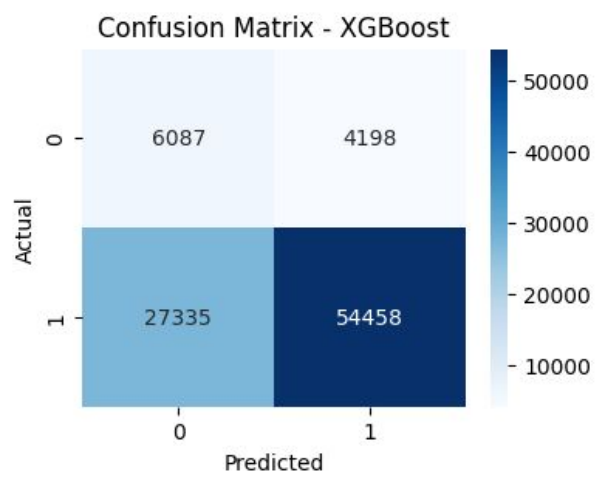


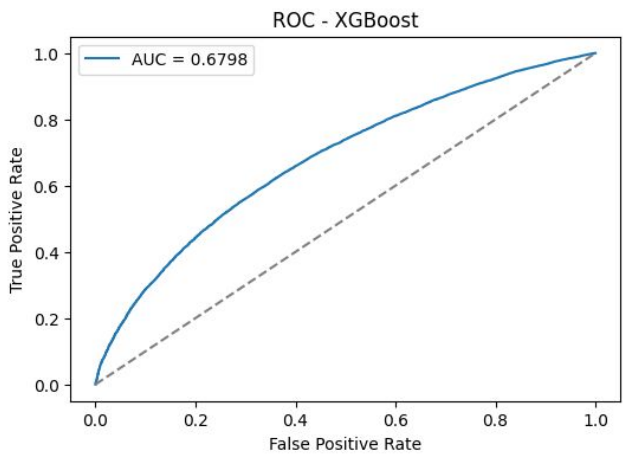
Model: RandomForest

	precision	recall	f1-score	support
0	0.4000	0.0002	0.0004	10285
1	0.8883	1.0000	0.9408	81793
accuracy			0.8883	92078
macro avg	0.6442	0.5001	0.4706	92078
weighted avg	0.8338	0.8883	0.8358	92078

Random Forest menghasilkan akurasi tinggi, namun gagal mengidentifikasi nasabah berisiko (bad loan), sehingga kurang cocok untuk tujuan bisnis yang ingin **meminimalkan risiko gagal bayar**

Evaluation





Model: XGBoost

	precision	recall	f1-score	support
0	0.1821	0.5918	0.2785	10285
1	0.9284	0.6658	0.7755	81793
accuracy			0.6575	92078
macro avg	0.5553	0.6288	0.5270	92078
weighted avg	0.8451	0.6575	0.7200	92078

XGBoost memberi trade-off terbaik, meski akurasi tidak tinggi, tapi **lebih adil menangani imbalance** dan lebih efektif dalam mendeteksi nasabah berisiko.

8. Conclusion

Hasil evaluasi menunjukkan:

- **XGBoost** menjadi model terbaik berdasarkan **AUC (0.6798)** dan memiliki **recall 0.6658** serta **precision 0.9284**. Artinya, model ini cukup seimbang dalam mendeteksi nasabah yang gagal bayar (tidak terlalu banyak yang lolos) sekaligus menjaga ketepatan prediksi bagi yang lancar bayar.

- **Logistic Regression** punya performa mirip XGBoost (AUC 0.6793) tapi recall lebih rendah, sehingga lebih berisiko meloloskan nasabah bermasalah.

- **Random Forest** mencatat akurasi tinggi (0.8883) dan recall hampir sempurna, tapi precision rendah, artinya model cenderung terlalu “longgar” dan bisa banyak salah dalam memprediksi nasabah gagal bayar.

👉 **Rekomendasi**: gunakan **XGBoost** sebagai model utama karena memberikan keseimbangan terbaik antara **minim risiko gagal bayar (recall cukup tinggi)** dan **efisiensi bisnis (precision tinggi)**, sesuai dengan tujuan bisnis untuk **mengurangi potensi kerugian akibat kredit macet**.

Thank You

X

Logo Company