

Prediction Model

ID/X Partners - Data Scientist

Presented by
DISYA NURUL ARIZA

DISYA NURUL ARIZA

Data Scientist Intern

Seorang fresh graduate Teknologi Informasi Universitas Bina Sarana Informatika angkatan 2021 dengan minat besar pada analisis data, statistika, dan visualisasi data, serta memiliki keterampilan dalam Python, MySQL, RapidMiner, dan tool visualisasi seperti Tableau dan Power BI untuk analisis dan pengolahan data.



Serang Baru, Kabupaten Bekasi, Jawa Barat



Disyaaarizaaa24@gmail.com



www.linkedin.com/in/disya-nurul-ariza24

Courses and Certification

SERTIFIKASI

- Sertifikat BNSP Database Programmer – LSP Teknologi Digital | [Link Sertifikat](#) | Tanggal Ujian: 30 Mei 2025
- Sertifikat BNSP Analis Program – LSP Universitas Bina Sarana Informatika | [Link Sertifikat](#) | Tanggal Ujian: 14 Januari 2025
- Sertifikat TOEFL – Lembaga Bahasa | [Link Sertifikat](#) | Tanggal Ujian: 10 Februari 2025
- Surat Keterangan Kompeten Associate Data Scientist - Digital Talent Scholarship 2025 | [Link Sertifikat](#) | Tanggal Ujian: 11 September 2025

PELATIHAN

- Intermediate Data Science – Digital Talent Scholarship 2025 (Juli 2025) | [Link Sertifikasi](#)
- Fundamental Data Science – Digital Talent Scholarship 2025 (Juni 2025) | [Link Sertifikasi](#)
- AI for Beginners – HP LIFE (November 2024) | [Link Sertifikasi](#)
- Data Science & Analytics – HP LIFE (November 2024) | [Link Sertifikasi](#)
- Python Fundamental for Data Science – DQLab (November 2024) | [Link Sertifikasi](#)
- Fundamental SQL Using SELECT Statement – DQLab (November 2024) | [Link Sertifikasi](#)
- R Fundamental for Data Science – DQLab (November 2024) | [Link Sertifikasi](#)
- Data Science – Dicoding Indonesia (September 2024) | [Link Sertifikasi](#)
- Data Analysis: Fullstack Intensive Bootcamp – MySkill (Maret – Mei 2024) | [Link Sertifikasi](#)
- Mini Course Data Analytics – RevoU (September 2023) | [Link Sertifikasi](#)
- Short Class Data Analysis – MySkill (Agustus 2023) | [Link Sertifikasi](#)

About Company

ID/X Partners adalah perusahaan konsultan teknologi dan data yang membantu bisnis memaksimalkan pemanfaatan data, mulai dari pengumpulan, pengolahan, analisis, hingga pembuatan model prediksi untuk mendukung strategi bisnis. Dengan keahlian di bidang data engineering, data science, machine learning, dan business intelligence, ID/X Partners melayani berbagai industri seperti perbankan, e-commerce, dan telekomunikasi, guna membantu klien mengambil keputusan yang lebih akurat, efisien, dan berbasis data.



Project Portfolio

Deskripsi Project:

Project ini bertujuan untuk mengembangkan model machine learning yang dapat memprediksi risiko kredit (credit risk) berdasarkan dataset pinjaman (loan dataset). Latar belakang masalah adalah kebutuhan perusahaan multifinance untuk meningkatkan akurasi dalam menilai kelayakan kredit, sehingga keputusan persetujuan pinjaman lebih tepat dan risiko gagal bayar dapat diminimalkan.

Problem Statement:

Bagaimana membangun model machine learning yang mampu mengklasifikasikan peminjam menjadi kategori Good Loan dan Bad Loan dengan akurasi tinggi, sehingga perusahaan dapat mengoptimalkan keputusan bisnis dan mengurangi potensi kerugian akibat kredit macet.

Output yang dihasilkan:

Model machine learning menggunakan Logistic Regression, Decision Tree, KNN, dan Random Forest.
Evaluasi model dengan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC.
Visualisasi hasil berupa confusion matrix dan ROC curve.

Link code [here!](#)

Project explanation video [here!](#)

1. Data Understanding

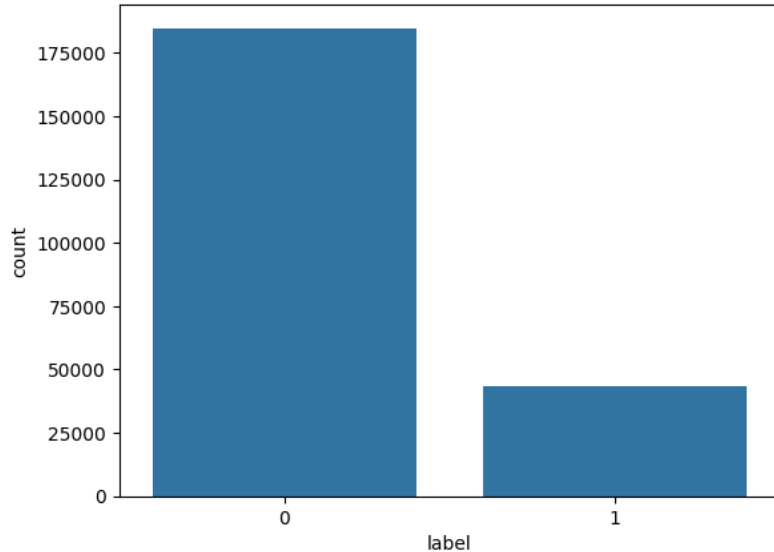
- Dataset memiliki 466.285 baris dan 75 kolom dengan informasi tentang pinjaman, data peminjam, dan status pembayaran.
- Dataset terdiri dari fitur numerik (loan_amnt, int_rate, annual_inc, dti) dan kategorikal (grade, sub_grade, home_ownership), dengan target loan_status yang ditransformasi menjadi biner. Good Loan (0 = Fully Paid) dan Bad Loan (1 = Charged Off/Default).
- **Jumlah Pinjaman (Rupiah), loan_amnt:** Mean = 14.800 | Median = 12.000 | Min = 500 | Max = 40.000
funded_amnt: Hampir sama dengan loan_amnt.
- **Cicilan (Rupiah per bulan), installment:** Mean = 430 | Median = 350 | Min = 15 | Max > 1.700
- **Suku Bunga (% per tahun), int_rate:** Mean = 13% | Median = 12,7% | Min = 5% | Max = 30%
- **Pendapatan Tahunan (Rupiah), annual_inc:** Mean = 75.000 | Median = 65.000 | Max > 1 juta (outlier)
- **Debt-to-Income Ratio (DTI) (%), dti:** Mean = 18 | Median = 16 | Max > 300 (anomali)
- **Riwayat Kredit (jumlah/kejadian):** open_acc: Rata-rata 11 akun | Max > 80, akundelinq_2yrs: Mayoritas 0 (tidak ada tunggakan), inq_last_6mths: Rata-rata < 2 kali, mayoritas rendah
- Data cenderung imbalanced (tidak seimbang) karena mayoritas Good Loan
- Terdapat missing value di beberapa kolom, misalnya: desc, mths_since_last_record, annual_inc
- Beberapa kolom redundan atau tidak relevan (misalnya url, member_id, id, desc, title) dan perlu dihapus.
- Mayoritas pinjaman diselesaikan dengan baik (distribusi tidak seimbang).
- Ada potensi outlier pada kolom annual_inc (income sangat tinggi pada sebagian kecil peminjam).
- Fitur seperti int_rate (bunga pinjaman) dan dti (debt-to-income ratio) diperkirakan punya pengaruh kuat terhadap status pinjaman.
- Mayoritas Good Loan punya int_rate rendah & dti kecil.
- Banyak Bad Loan muncul di kelompok dengan income lebih rendah.

2. Feature Engineering

- Dari total 75 kolom pada dataset, hanya dipilih 17 fitur yang relevan dengan performa kredit, yaitu:
Numerik: loan_amnt, funded_amnt, int_rate, installment, annual_inc, dti, delinq_2yrs, inq_last_6mths, open_acc, pub_rec.
Kategorikal: term, grade, sub_grade, emp_length, home_ownership, verification_status, purpose.
Target/label: loan_status.
- Fitur kategorikal (term, grade, sub_grade, emp_length, home_ownership, verification_status, purpose) diubah ke bentuk dummy variables (One-Hot Encoding) (pd.get_dummies, drop_first=True).
- Outputnya Dataset dengan fitur numerik & dummy variables siap diproses.

3. Exploratory Data Analysis

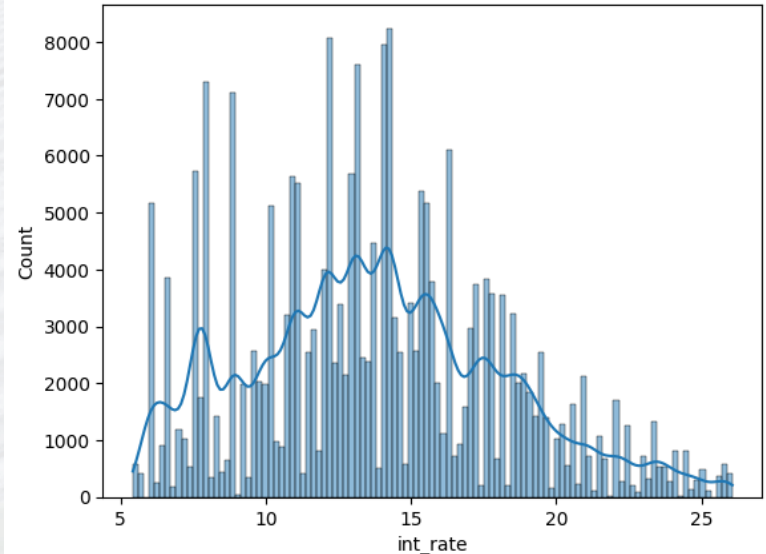
Distribusi Good vs Bad Loan



Dataset Imbalanced

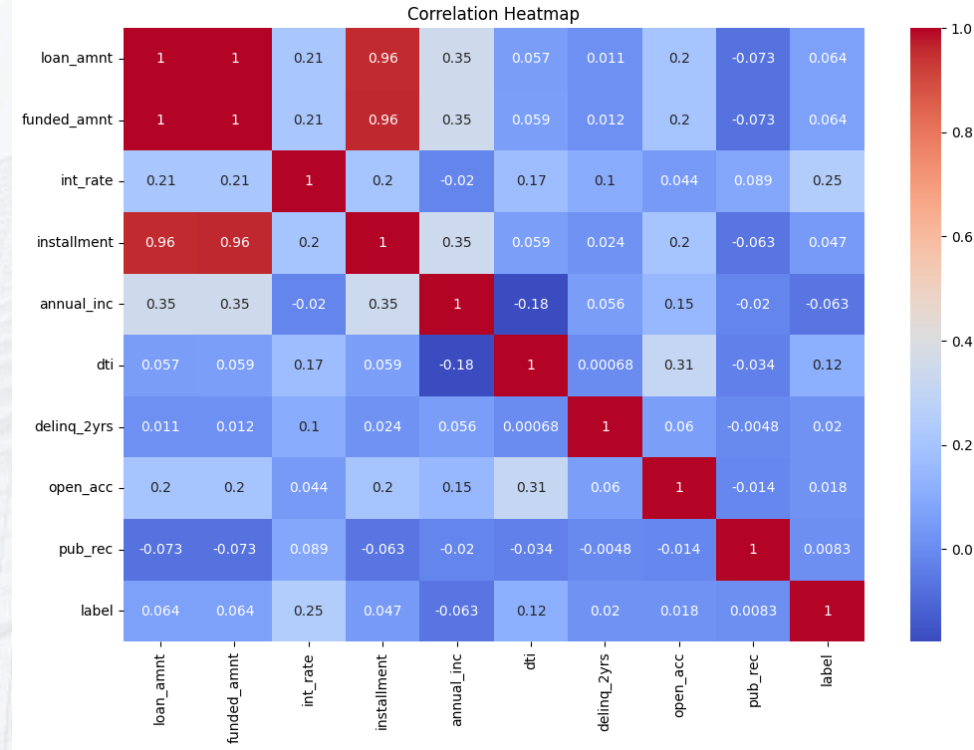
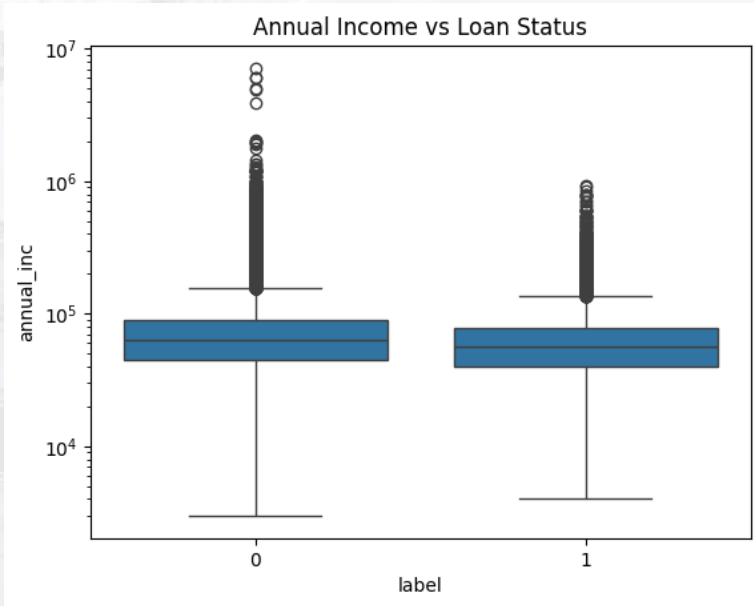
Analisis Univariat (satu variabel)

Distribusi Interest Rate



3. Exploratory Data Analysis

Analisis Bivariat (hubungan dengan target)



Korelasi Antar Variabel Numerik

4. Data Preparation

- Input dari Feature Engineering
- Untuk fitur numerik, nilai hilang diganti dengan median (agar tidak dipengaruhi outlier).
- Fitur kategorikal, nilai hilang diganti dengan mode (nilai yang paling sering muncul).
- Digunakan StandardScaler agar semua fitur numerik berada dalam skala yang sama (mean = 0, std = 1).
- Hal ini penting terutama untuk algoritma yang sensitif terhadap skala data seperti KNN dan Logistic Regression.
- Dataset dibagi menjadi train set (80%) dan test set (20%) dengan stratifikasi agar distribusi target tetap seimbang.

5. Data Modeling

1. Pada tahap ini dipilih beberapa algoritma supervised learning (classification) untuk dibandingkan kinerjanya.

Model yang digunakan:

Logistic Regression, model linear untuk klasifikasi biner,

Decision Tree, mudah dipahami, tapi rentan overfitting,

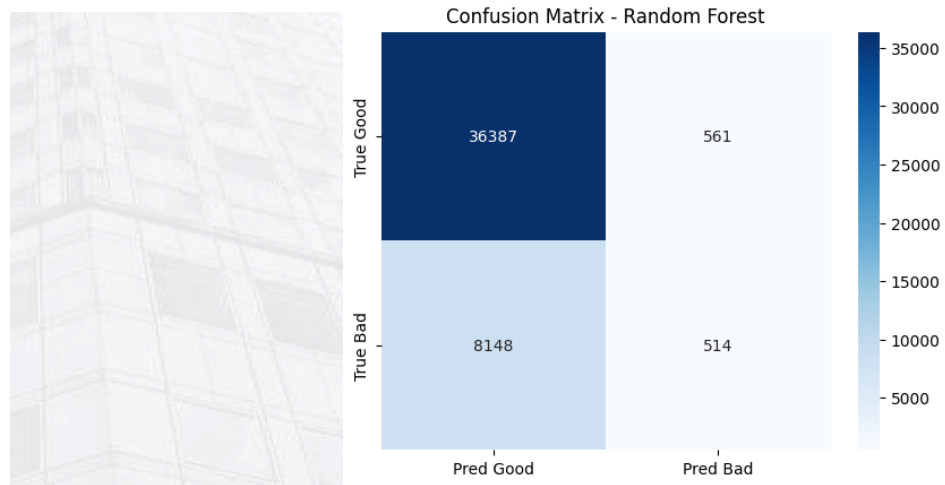
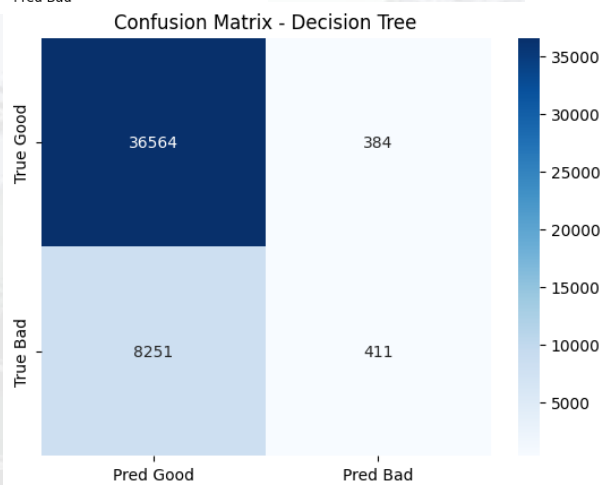
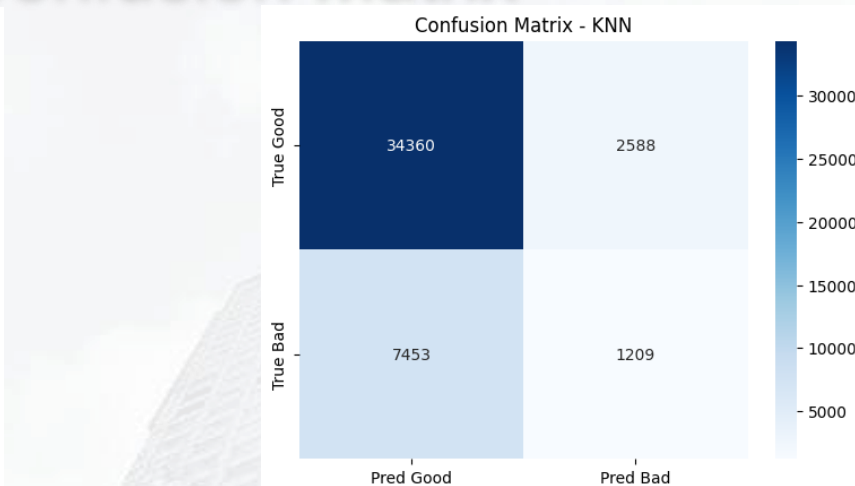
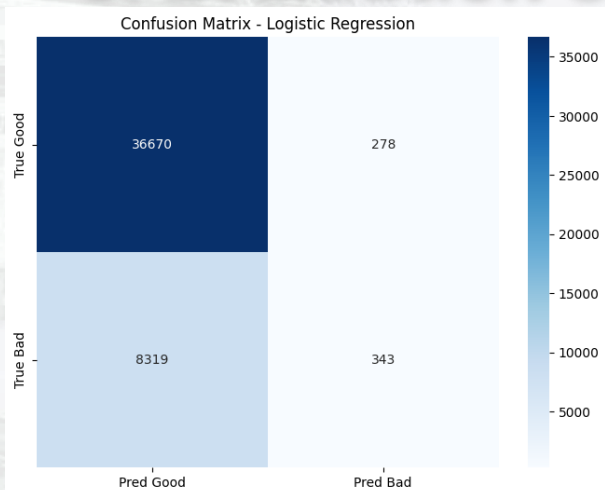
K-Nearest Neighbors (KNN), prediksi berdasar kedekatan data,

Random Forest, ensemble dari Decision Tree, lebih stabil & akurat,

Tujuan menggunakan banyak model adalah agar dapat memilih model terbaik berdasarkan evaluasi.

2. Data dibagi menjadi 80% data latih (train set) dan 20% data uji (test set). Stratifikasi dilakukan pada target agar distribusi Good Loan dan Bad Loan tetap seimbang di train-test split.
3. Masing-masing model dilatih (fit) menggunakan X_{train} dan y_{train} . Model mempelajari pola dari fitur numerik dan kategorikal (setelah di-*encoding* dan *scaling*).
4. Evaluasi dilakukan dengan beberapa metrik: Accuracy (%), Precision (%), Recall (%), F1-score (%), ROC-AUC
5. Cross-Validation, menggunakan Stratified K-Fold (5-fold). Data dibagi ke dalam 5 bagian, tiap bagian bergantian menjadi test, sisanya train
6. Confusion Matrix, Classification Report, ROC Curve.

6. Evaluation Confusion Matrix



6. Evaluation Classification Report

=== Logistic Regression ===

	precision	recall	f1-score	support
Good Loan	0.82	0.99	0.90	36948
Bad Loan	0.55	0.04	0.07	8662
accuracy			0.81	45610
macro avg	0.68	0.52	0.48	45610
weighted avg	0.77	0.81	0.74	45610

=== KNN ===

	precision	recall	f1-score	support
Good Loan	0.82	0.93	0.87	36948
Bad Loan	0.32	0.14	0.19	8662
accuracy			0.78	45610
macro avg	0.57	0.53	0.53	45610
weighted avg	0.73	0.78	0.74	45610

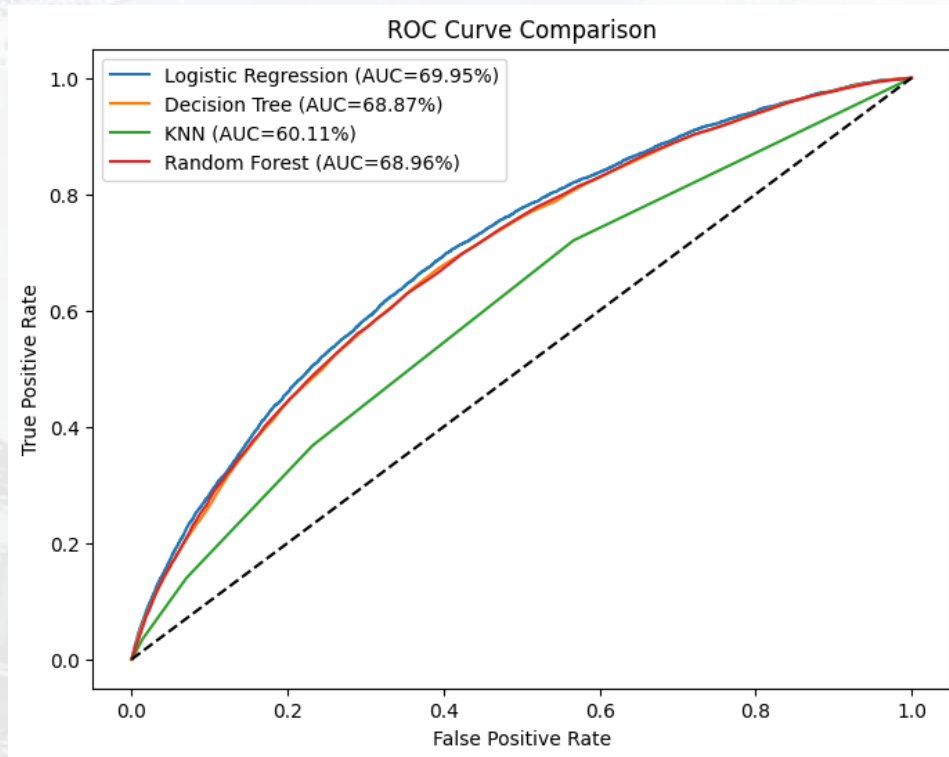
=== Decision Tree ===

	precision	recall	f1-score	support
Good Loan	0.82	0.99	0.89	36948
Bad Loan	0.52	0.05	0.09	8662
accuracy			0.81	45610
macro avg	0.67	0.52	0.49	45610
weighted avg	0.76	0.81	0.74	45610

=== Random Forest ===

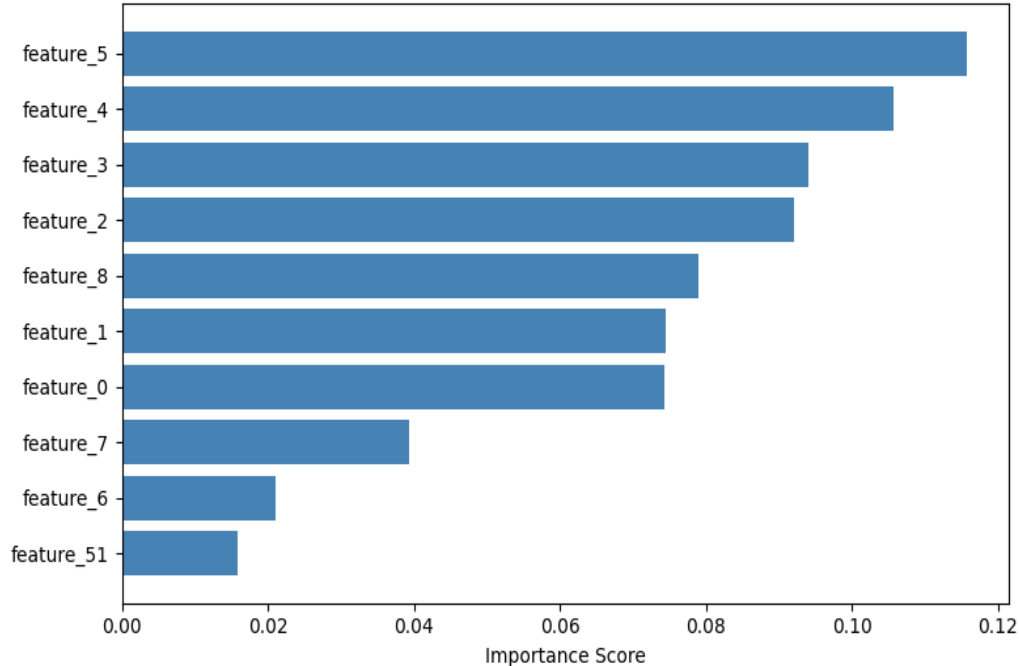
	precision	recall	f1-score	support
Good Loan	0.82	0.98	0.89	36948
Bad Loan	0.48	0.06	0.11	8662
accuracy			0.81	45610
macro avg	0.65	0.52	0.50	45610
weighted avg	0.75	0.81	0.74	45610

6. Evaluation ROC Curve



7. Feature Importance – Random Forest

Top 10 Feature Importance - Random Forest



Berdasarkan hasil *Feature Importance* model Random Forest, fitur yang paling berpengaruh terhadap status pinjaman adalah interest rate (11,6%), DTI (10,5%), dan annual income (9,4%). Semakin tinggi suku bunga dan rasio utang terhadap pendapatan, semakin besar risiko gagal bayar (*Bad Loan*), sedangkan pendapatan yang tinggi cenderung menurunkan risiko tersebut.

8. Conclusion

1. Random Forest terbaik dalam menangkap bad loan dengan recall dan f1-score tertinggi, sehingga paling efektif meminimalisir risiko gagal bayar.
2. Logistic Regression unggul di ROC-AUC dan precision bad loan, tetapi recall lebih rendah sehingga lebih banyak bad loan terlewat.
3. Decision Tree mudah diinterpretasi namun rentan overfitting.
4. KNN akurasi terendah karena sensitif terhadap skala dan distribusi data.
5. Hasil evaluasi menegaskan bahwa Random Forest paling efektif membedakan Good Loan (pinjaman yang disetujui/lunas) dan Bad Loan (pinjaman yang ditolak atau gagal bayar) pada dataset yang disediakan.

9. Insight

1. Berdasarkan hasil *feature importance* dari model Random Forest, tiga fitur paling berpengaruh terhadap status pinjaman adalah: Interest Rate (*int_rate*), Debt-to-Income Ratio (*dti*), Annual Income (*annual_inc*).
2. Fitur *int_rate* memiliki pengaruh paling tinggi terhadap prediksi risiko kredit, diikuti oleh *dti* dan *annual_inc*, menunjukkan bahwa suku bunga dan rasio utang terhadap pendapatan merupakan indikator utama dalam menilai kelayakan pinjaman.
3. Nasabah dengan suku bunga rendah, rasio utang (DTI) kecil, dan pendapatan tahunan tinggi cenderung masuk kategori Good Loan. Sebaliknya, peminjam dengan pendapatan rendah dan DTI tinggi memiliki peluang lebih besar menjadi Bad Loan.
4. Distribusi target yang tidak seimbang (mayoritas Good Loan) menegaskan pentingnya teknik penanganan imbalanced data agar model tidak bias.

10. Weakness

1. Data tidak seimbang (imbalanced), mayoritas data termasuk kategori *Good Loan*, sehingga model cenderung bias ke kelas tersebut dan sulit menangkap *Bad Loan*.
2. Outlier dan anomali (contoh: pendapatan tahunan yang sangat tinggi, atau debt-to-income ratio yang tinggi dapat memengaruhi stabilitas model.
3. Fitur masih terbatas: hanya menggunakan data internal pinjaman, tanpa integrasi data eksternal (misalnya riwayat kredit dari lembaga keuangan lain atau perilaku transaksi nasabah).
4. Logistic Regression memiliki recall rendah untuk kelas *Bad Loan* (4%), artinya sekitar 96% peminjam berisiko tidak terdeteksi (false negative), yang dapat meningkatkan risiko gagal bayar. Sementara itu, model **Random Forest** menunjukkan performa yang lebih stabil dengan kemampuan untuk mengidentifikasi faktor risiko utama melalui *Feature Importance Analysis*.
5. Decision Tree: mudah dipahami tapi rentan *overfitting* pada data training, sehingga performa bisa menurun ketika diuji pada data baru.
6. KNN: memberikan akurasi terendah karena sangat sensitif terhadap perbedaan skala data dan distribusi yang tidak merata.

11. Recommendation

1. Gunakan Random Forest sebagai baseline model: model ini sudah terbukti memiliki recall dan f1-score tertinggi dalam mendeteksi Bad Loan, sehingga paling sesuai untuk implementasi awal.
2. Tangani data imbalance dengan menerapkan metode seperti SMOTE, oversampling, undersampling, atau penyesuaian class weight untuk meningkatkan kemampuan model menangkap Bad Loan.
3. Eksperimen dengan algoritma lanjutan dengan mencoba XGBoost, LightGBM, atau CatBoost, yang secara umum lebih unggul dibanding Random Forest terutama pada data besar dan kompleks.
4. Perbaiki feature engineering dengan melakukan pembersihan outlier, normalisasi distribusi, atau pembuatan variabel turunan (misalnya rasio pinjaman terhadap pendapatan) agar model lebih robust.
5. Integrasi data eksternal dengan menambahkan sumber data lain seperti skor kredit pihak ketiga, data perilaku transaksi, atau histori pembayaran untuk memperkuat prediksi risiko.
6. Gunakan interpretability tools dengan memanfaatkan SHAP atau feature importance agar hasil model lebih transparan, sehingga pihak manajemen bisa memahami alasan di balik keputusan prediksi.

12. Business Implication

1. Mengurangi risiko kredit macet, dengan kemampuan mendeteksi peminjam berisiko lebih baik, perusahaan bisa menekan jumlah pinjaman yang gagal bayar.
2. Proses persetujuan pinjaman menjadi lebih cepat karena model dapat melakukan screening otomatis, sehingga beban tim analis kredit berkurang.
3. Calon peminjam yang layak lebih cepat disetujui, sementara yang berisiko bisa ditolak atau dipantau lebih ketat.
4. Kerugian akibat kredit macet berkurang, profit perusahaan bisa meningkat.
5. Implementasi machine learning dalam credit scoring memperkuat daya saing perusahaan dalam industri multifinance yang semakin kompetitif.
6. Hasil *feature importance* dapat dimanfaatkan oleh tim bisnis untuk fokus pada calon peminjam dengan bunga tinggi, DTI besar, dan pendapatan rendah sebagai segmen yang berisiko tinggi dan memerlukan perhatian khusus.

Thank You



Rakamin
Academy



id/x partners