Analisis dan Prediksi Skor Kredit untuk Home Credit Indonesia

Oleh
Anita Margareth D Silalahi

Tujuan

Proyek ini menggunakan machine learning untuk memprediksi risiko kredit pelanggan dengan menganalisis data dari aplikasi pinjaman dan riwayat kredit sebelumnya. Model Logistic Regression dan Random Forest digunakan untuk klasifikasi, dengan teknik preprocessing seperti imputasi, normalisasi, seleksi fitur, dan encoding kategori.

1

Menganalisis performa dua model machine learning dalam memprediksi kelayakan kredit pelanggan.

2

Mengevaluasi model menggunakan matrix accuracy, precision, recall, F1-score, dan AUC-ROC Score.

3

Memberikan rekomendasi bisnis berdasarkan hasil model.

4

Meminimalkan risiko kredit macet dan memastikan pinjaman diberikan kepada pelanggan yang tepat.

Dataset

Sumber Data

- Dataset utama: application_train.csv dan application_test.csv.
- Dataset tambahan: bureau.csv, bureau_balance.csv, previous_application.csv, installments_payments.csv, POS_CASH_balance.csv, credit_card_balance.csv.

Karakteristik Data:

- 307.511 baris data pelatihan dengan 122 kolom.
- Data mencakup informasi demografis, riwayat kredit, dan perilaku pembayaran pelanggan.

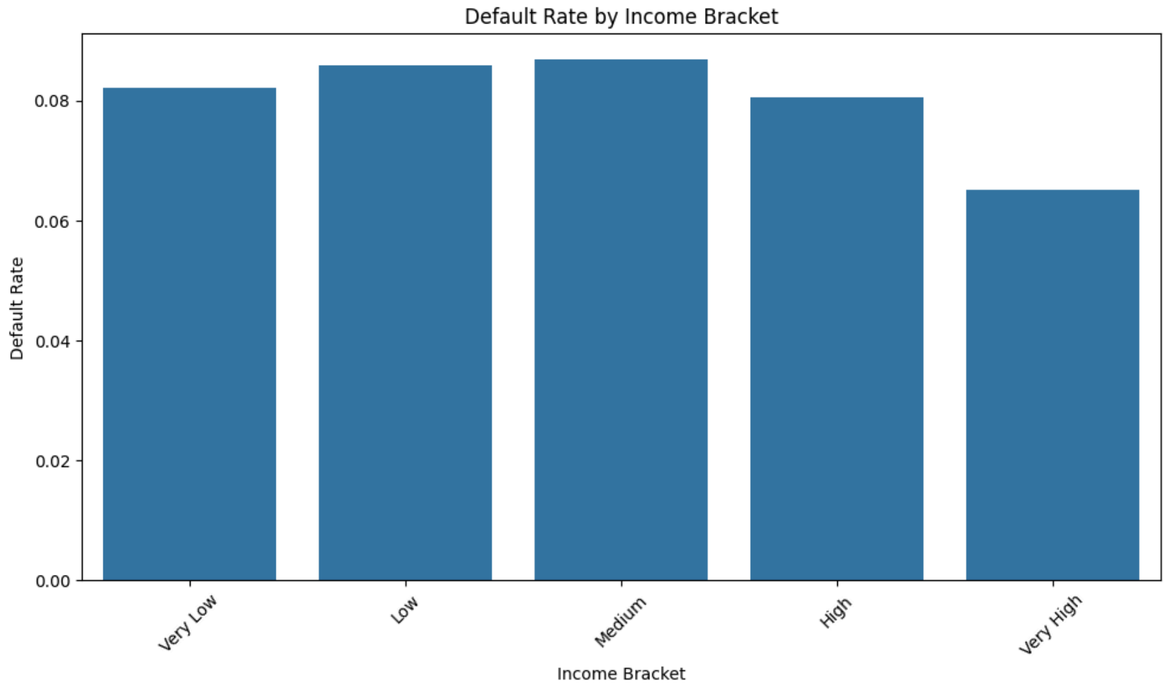
Tantangan:

- Data yang tidak seimbang (hanya 8% pelanggan dengan risiko kredit buruk).
- Banyak missing values dan data kategorikal yang perlu diproses.

Insight 1

Pengaruh Tingkat Pendapatan Terhadap Risiko Gagal Bayar

- Kelompok pendapatan Sangat Tinggi menunjukkan tingkat gagal bayar terendah (6,52%)
- Kelompok pendapatan Menengah memiliki tingkat gagal bayar tertinggi (8,68%)
- Di luar dugaan, pendapatan rendah tidak secara langsung berkorelasi dengan gagal bayar yang lebih tinggi



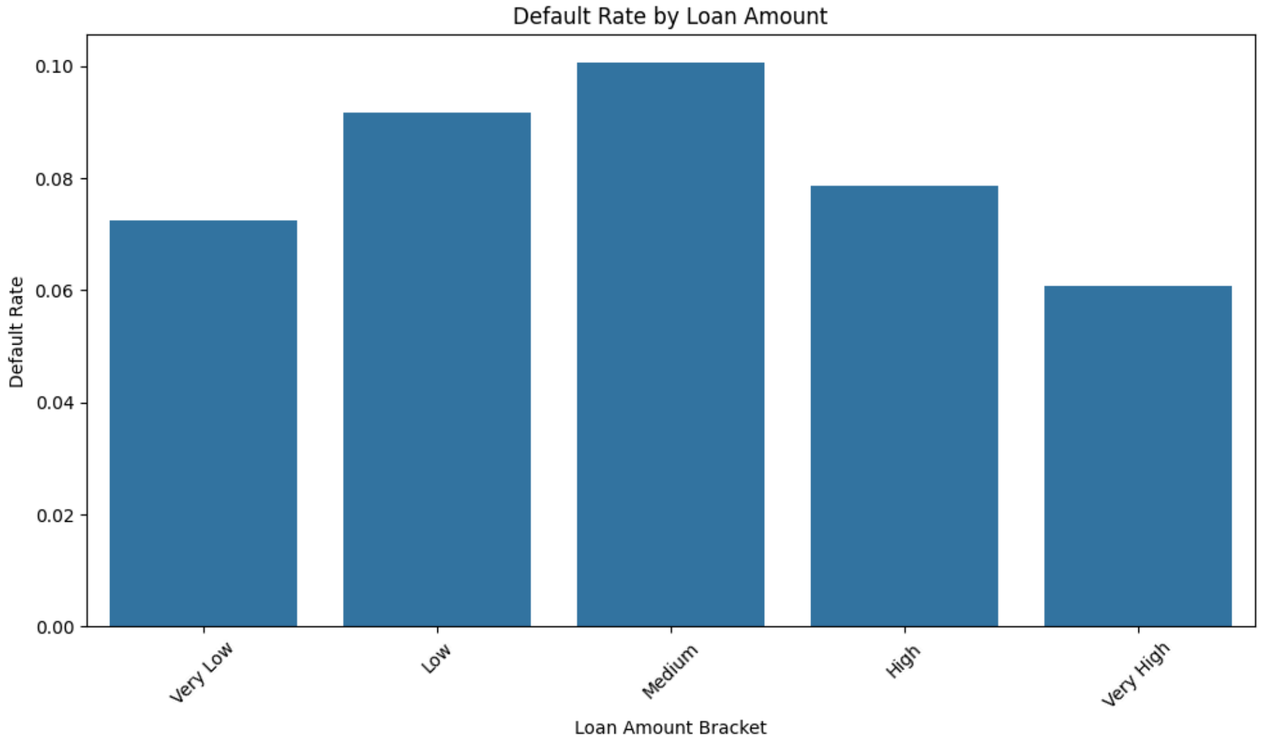
Action :

- Mengembangkan produk yang disesuaikan untuk segmen pendapatan tinggi
- Meningkatkan penilaian risiko untuk kelompok pendapatan menengah
- Membuat strategi pemasaran yang terarah

Insight 2

Pola Risiko Berdasarkan Jumlah Pinjaman

- Pinjaman ukuran menengah memiliki tingkat gagal bayar tertinggi (10,05%)
- Jumlah pinjaman sangat tinggi menunjukkan tingkat gagal bayar terendah (6,08%)
- Terdapat pola risiko berbentuk U yang jelas di seluruh jumlah pinjaman



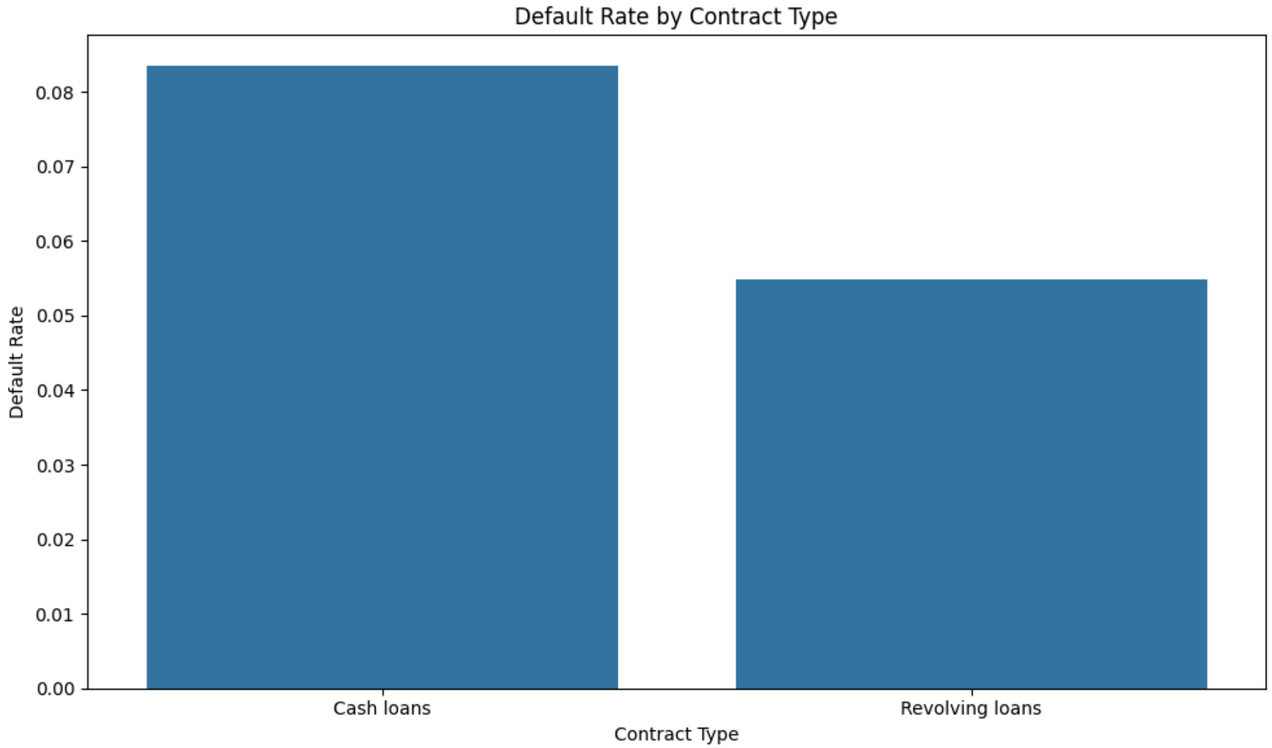
Action :

- Meningkatkan kehati-hatian dan screening untuk pinjaman kategori Medium
- Menawarkan suku bunga lebih baik untuk pinjaman sangat tinggi untuk mendorong pengambilan
- Menyesuaikan kriteria approval berdasarkan jumlah pinjaman

Insight 3

Dampak Jenis Kontrak

- Pinjaman tunai (90,48% dari portofolio) menunjukkan tingkat gagal bayar 8,35%
- Pinjaman bergulir (9,52% dari portofolio) menunjukkan tingkat gagal bayar 5,48%
- Pinjaman bergulir berkinerja jauh lebih baik (-35% tingkat gagal bayar)



Action :

- Meningkatkan pangsa portofolio pinjaman bergulir
- Menerapkan kebijakan kredit berbeda berdasarkan jenis kontrak
- Mengembangkan kampanye pemasaran yang menyoroti manfaat pinjaman bergulir

Model Final

Model yang Digunakan

- Logistic Regression
- Random Forest

Data dan Features

- Dataset terdiri dari 61.503 sampel
- Label terdiri dari dua kelas (0 dan 1)
- Fitur yang digunakan belum dijelaskan, tapi biasanya untuk model seperti ini melibatkan fitur numerik dan kategorikal yang telah diolah

Preprocessing Data

- Data Cleaning
- Feature Engineering
- Splitting Data
- Oversampling/Undersampling

Model Final

Evaluasi Model

Logistic Regression

- Akurasi: 91.92%
- Precision, Recall, F1-score untuk kelas 1 masih rendah (indikasi masalah imbalance data)
- Confusion Matrix menunjukkan model cenderung lebih baik dalam mengenali kelas mayoritas (0) dibanding kelas minoritas (1)

Random Forest

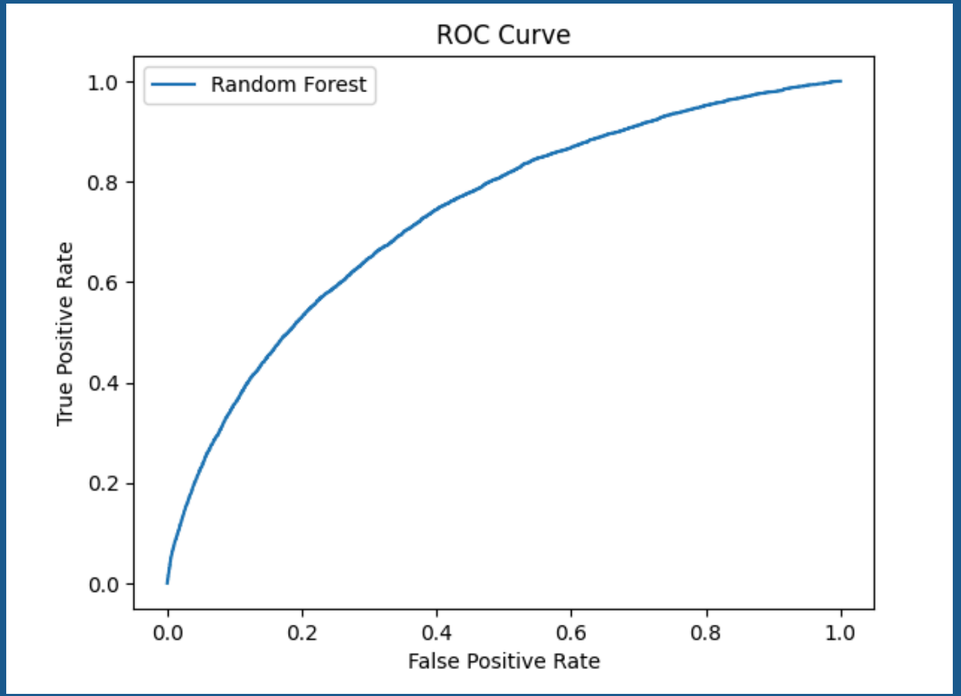
- Akurasi: 91.95% (sedikit lebih baik dari Logistic Regression)
- Tidak ada prediksi yang benar untuk kelas 1 → Precision = 0, Recall = 0
- Warning dari sklearn menunjukkan bahwa model tidak bisa memprediksi kelas 1 sama sekali
- Masih ada masalah imbalance data yang cukup signifikan

Model Final

Evaluasi Model

AUC-ROC Score

- 0.735, menunjukkan bahwa model masih memiliki kemampuan klasifikasi yang terbatas
- Perlu perbaikan sebelum diterapkan dalam bisnis



Referensi

S. Sathyanarayanan, “Confusion Matrix-Based Performance Evaluation Metrics,” African Journal of Biomedical Research, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

A. C. . Rencher and G. Bruce. Schaalje, Linear models in statistics. Wiley-Interscience, 2008.

A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” in Ensemble Machine Learning, New York, NY: Springer New York, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7_5.

Link Git Repository

Link Repositori

Terima Kasih

anitasilalahi18@gmail.com

Universitas Brawijaya

Linked In | Anita Silalahi