PORTFOLIO 2025

Disya Nurul Ariza
Virtual Internship
Experience

Home Credit Default Risk – *Predicting Loan Default Using Machine Learning*

PROBLEM STATEMENT

Problem
Perusahaan menghadapi risiko kredit yang tinggi akibat nasabah gagal membayar pinjaman (default), yang berdampak pada meningkatnya Non-Performing Loan (NPL) dan kerugian finansial perusahaan.

Objective
1. Memprediksi risiko gagal bayar (default) nasabah secara akurat
2. Mendukung proses persetujuan kredit berbasis data dan objektif
3. Membantu perusahaan dalam menekan tingkat Non-Performing Loan (NPL)

DATASET DESCRIPTION

Analisis ini menggunakan beberapa dataset historis dari Home Credit yang saling terhubung untuk memprediksi risiko gagal bayar (default) nasabah.

Dataset yang digunakan:

- **application_train.csv**
  Data utama nasabah yang berisi informasi demografis, finansial, serta label target (**TARGET**) yang menunjukkan status default.
- **bureau.csv**
  Berisi riwayat kredit eksternal nasabah dari lembaga lain.
- **previous_application.csv**
  Menyimpan informasi pengajuan pinjaman nasabah sebelumnya.
- **installments_payments.csv**
  Mencatat riwayat pembayaran cicilan dan keterlambatan pembayaran.

Key Identifier:
Seluruh dataset diintegrasikan menggunakan **SK_ID_CURR** sebagai primary key.

PROCESS ANALYSIS & MODELLING

Data Preprocessing

- Menggabungkan seluruh dataset menggunakan SK_ID_CURR sebagai primary key
- Menangani missing value
  - Data numerik diisi dengan nilai median
  - Data kategorikal diisi dengan kategori “Unknown”
- Melakukan encoding pada variabel kategorikal
- Melakukan feature scaling menggunakan StandardScaler
- Membagi data menjadi 80% data latih dan 20% data uji

Feature Engineering

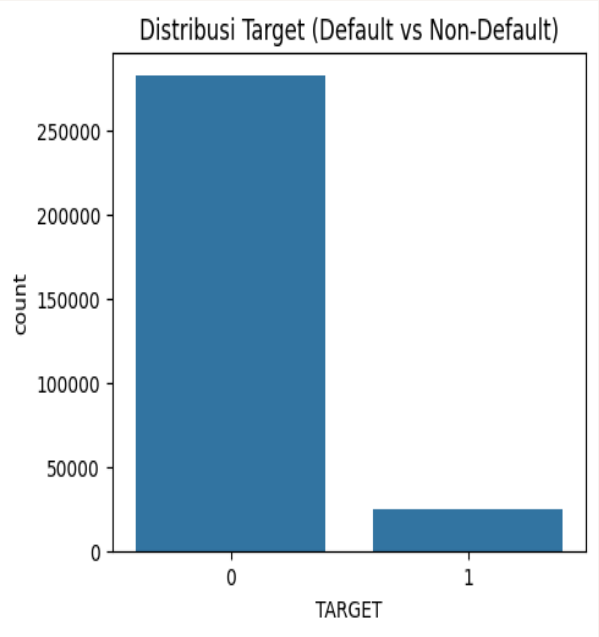
- Membuat fitur agregasi dari data riwayat kredit eksternal (**bureau**)
- Menghitung rata-rata nilai pengajuan pinjaman sebelumnya (**previous_application**)
- Menghitung selisih pembayaran cicilan sebagai indikator keterlambatan (**installments_payments**)

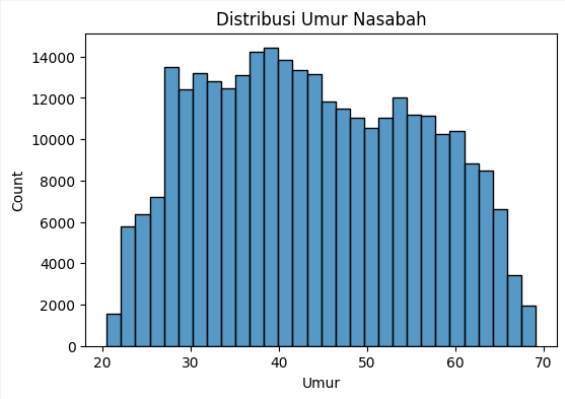
Machine Learning Model

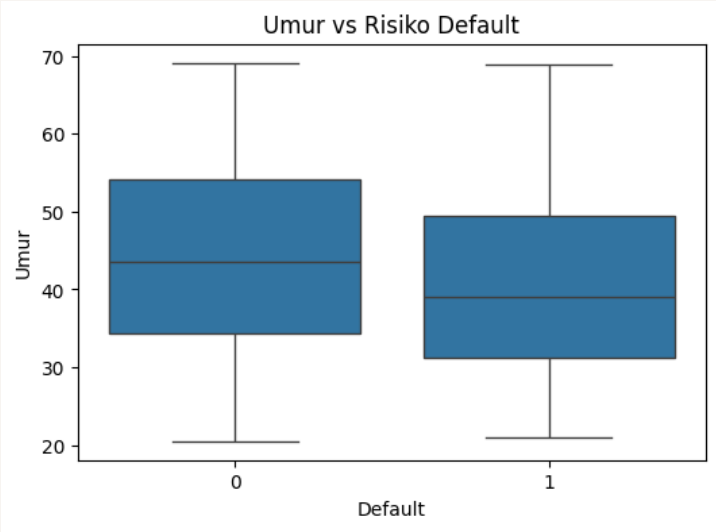
- Model yang digunakan: Logistic Regression
- Model digunakan untuk memprediksi probabilitas nasabah mengalami gagal bayar (default)

Model Evaluation

- Evaluasi performa model menggunakan AUC–ROC
- Analisis hasil prediksi menggunakan classification report



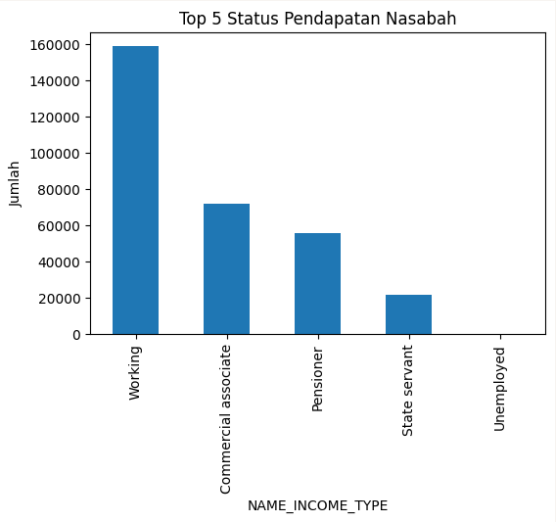


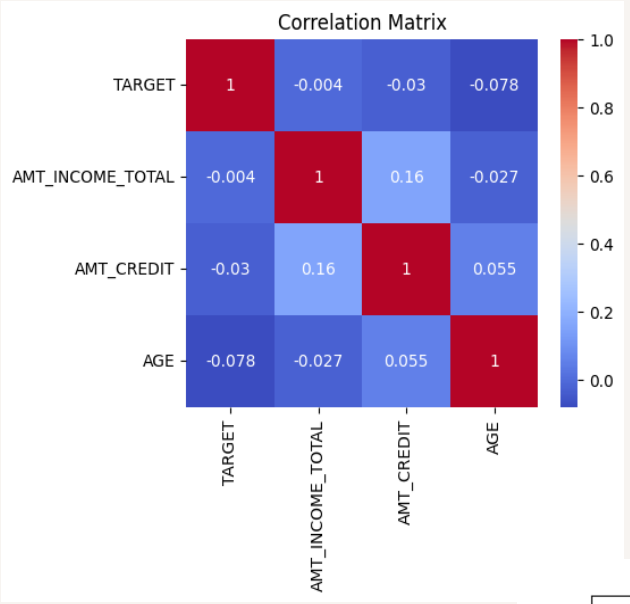


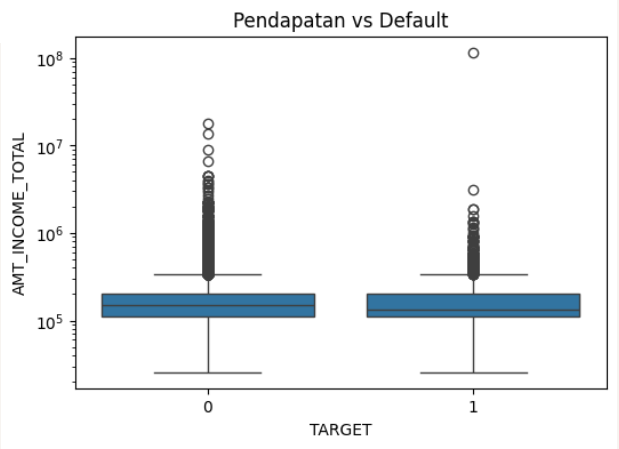
INSIGHT (EDA)

Hasil exploratory data analysis menunjukkan bahwa nasabah dengan pendapatan lebih rendah memiliki kecenderungan risiko gagal bayar (default) yang lebih tinggi.

Perusahaan dapat menerapkan pengetatan persetujuan kredit atau penyesuaian limit untuk segmen nasabah dengan pendapatan rendah.





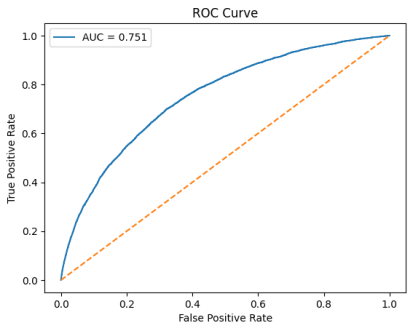


INSIGHT (EDA)

Riwayat keterlambatan pembayaran cicilan memiliki hubungan yang kuat dengan peningkatan risiko gagal bayar nasabah.

Penerapan early warning system dan evaluasi berkala terhadap nasabah dengan riwayat pembayaran buruk.

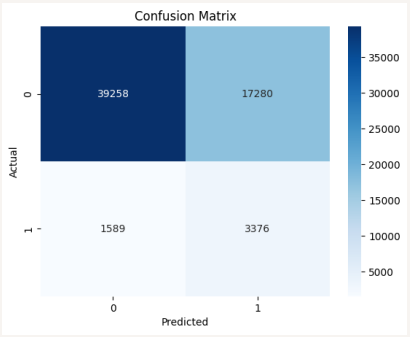
MODEL EVALUATION & INTERPRETATION



Model dievaluasi menggunakan AUC–ROC dan Confusion Matrix untuk mengukur kemampuan model dalam membedakan nasabah berisiko dan tidak berisiko.

Hasil evaluasi menunjukkan bahwa model Logistic Regression memiliki performa yang cukup baik dan stabil dalam memprediksi risiko gagal bayar, sehingga dapat digunakan sebagai dasar pengambilan keputusan kredit.

	Feature	Coefficient
21	FLAG_EMP_PHONE	7.648409
16	DAYS_EMPLOYED	7.596906
6	AMT_CREDIT	0.857427
71	BASEMENTAREA_MEDI	0.411817
127	PREV_CREDIT_MEAN	0.391584
45	YEARS_BUILD_AVG	0.320416
80	LIVINGAPARTMENTS_MEDI	0.317032
53	LIVINGAREA_AVG	0.243245
42	APARTMENTS_AVG	0.227562
4	CNT_CHILDREN	0.216138



Feature Coefficient Analysis
Hasil analisis koefisien pada model Logistic Regression menunjukkan bahwa beberapa variabel memiliki pengaruh signifikan terhadap risiko gagal bayar nasabah.

Variabel seperti riwayat keterlambatan pembayaran, jumlah kredit, dan pendapatan nasabah memiliki kontribusi terbesar dalam meningkatkan maupun menurunkan probabilitas default.

BUSINESS RECOMMENDATION

Berdasarkan hasil exploratory data analysis dan pemodelan machine learning, perusahaan pembiayaan dapat memanfaatkan output probabilitas default sebagai dasar pengambilan keputusan kredit yang lebih objektif dan berbasis data. Model dapat digunakan untuk mengelompokkan nasabah ke dalam beberapa tingkat risiko guna mendukung proses persetujuan kredit.

Nasabah dengan probabilitas default rendah dapat diberikan persetujuan kredit secara otomatis dengan proses yang lebih cepat. Untuk nasabah dengan risiko menengah, perusahaan disarankan melakukan evaluasi tambahan seperti verifikasi data atau penyesuaian limit kredit. Sementara itu, nasabah dengan risiko tinggi dapat dikenakan penolakan pengajuan atau pembatasan kredit untuk meminimalkan potensi kerugian.

Selain itu, perusahaan dapat menerapkan sistem early warning berbasis riwayat pembayaran cicilan dan perilaku kredit nasabah untuk memantau potensi gagal bayar lebih awal. Implementasi strategi ini diharapkan dapat membantu perusahaan dalam menurunkan tingkat Non-Performing Loan (NPL) serta meningkatkan kualitas portofolio kredit secara keseluruhan.

CONCLUSION

Project ini berhasil membangun solusi prediktif untuk mengidentifikasi risiko gagal bayar nasabah menggunakan data historis Home Credit dalam skema Project Based Learning Internship. Proses analisis dimulai dari exploratory data analysis untuk memahami karakteristik nasabah, dilanjutkan dengan feature engineering, preprocessing data, serta pemodelan machine learning.

Model Logistic Regression yang digunakan menunjukkan performa yang stabil dalam membedakan nasabah berisiko dan tidak berisiko, serta didukung oleh hasil evaluasi menggunakan ROC–AUC dan confusion matrix. Analisis feature coefficient juga memberikan insight penting terkait faktor-faktor utama yang memengaruhi risiko default, sehingga model dapat diinterpretasikan secara bisnis.

Secara keseluruhan, hasil project ini menunjukkan bahwa pendekatan data-driven dapat membantu perusahaan pembiayaan dalam mendukung pengambilan keputusan kredit, meningkatkan efisiensi proses persetujuan, serta mengurangi risiko kredit di masa mendatang.

Disyaaarizaaa24@gmail.com

Link Code Here!

DISYA NURUL ARIZA

Data Scientist Intern

Seorang fresh graduate Teknologi Informasi Universitas Bina Sarana Informatika 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

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