

Implementasi Machine Learning untuk Prediksi Risiko Kredit (Credit Risk Scoring)

Final Task Data Analyst Id/x Partners
– Prediction Model–

Ghozi Alvin Karim

Latar Belakang

Perusahaan pendanaan perlu membedakan antara calon peminjam yang berisiko tinggi (gagal bayar) dan peminjam yang aman untuk menjaga stabilitas portofolio keuangan.



Proyek ini bertujuan untuk Membangun model klasifikasi biner yang mampu memprediksi apakah seorang nasabah akan masuk kategori "**Pinjaman Baik**" atau "**Pinjaman Tidak Baik**".

Menggunakan metodologi End-to-End Data Science, mulai dari pembersihan data (466rb+ baris), rekayasa fitur, hingga implementasi pipeline machine learning otomatis.



Dataset

Dataset pinjaman historis periode 2007 - 2014 terdiri dari 466.285 baris dan 75 fitur awal (sebelum pembersihan). Campuran antara data numerik seperti jumlah pinjaman dan bunga serta data kategorikal (seperti grade, tujuan pinjaman, dan status kepemilikan rumah). Target Variabel **loan_status** yang disederhanakan menjadi **1 (Baik)**: Status Current, Fully Paid dan **0 (Tidak Baik)**: Status Charged Off, Late.

Dimensi Dataset : (466285, 75)

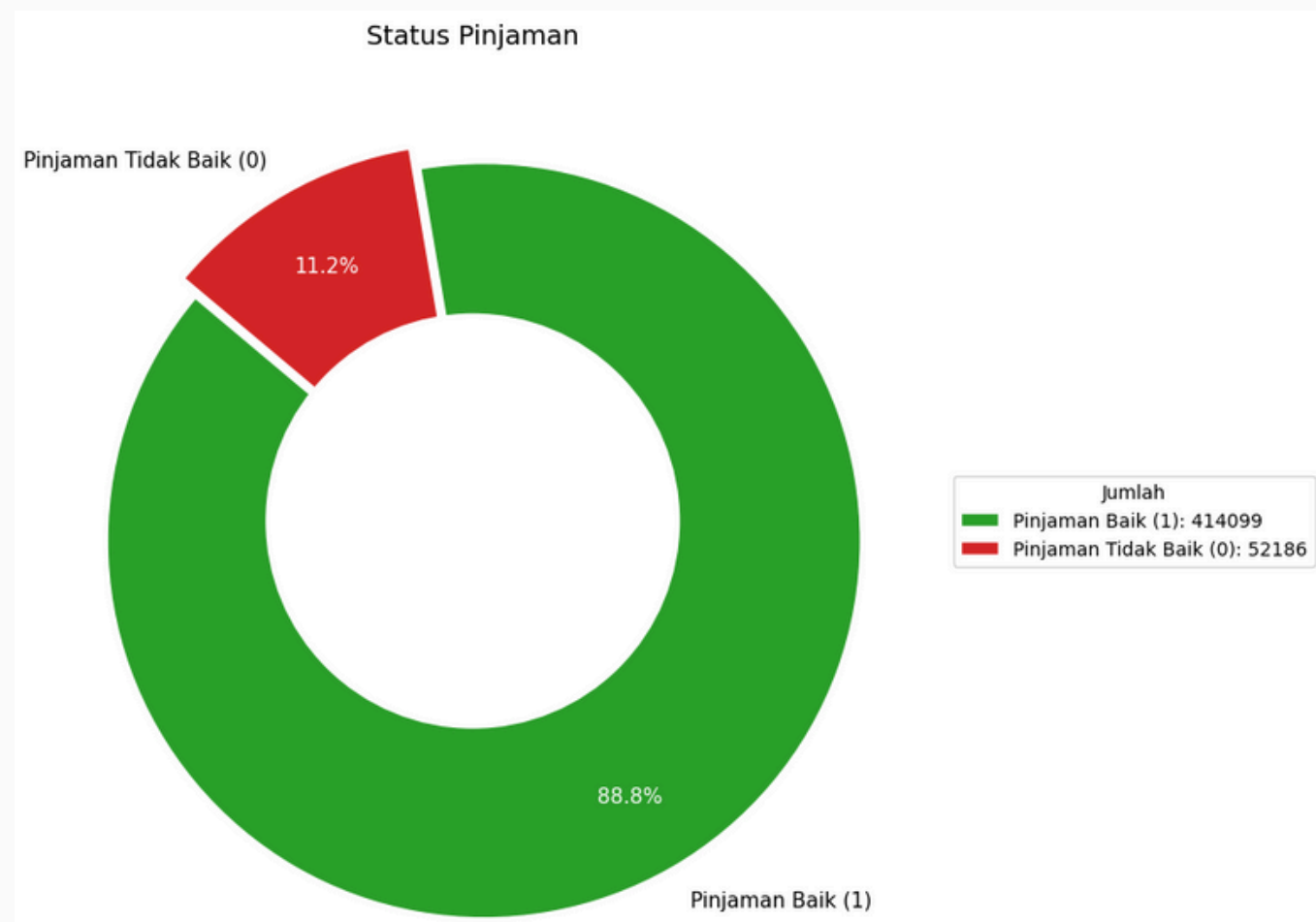
Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	...	total_bal_il	il_util	open_rv_12m	open_rv_24m	max_bal_bc	all_util
0	0	1077501	1296599	5000	5000	4975.0 36 months	10.65	162.87	B	...	NaN	NaN	NaN	NaN	NaN	NaN
1	1	1077430	1314167	2500	2500	2500.0 60 months	15.27	59.83	C	...	NaN	NaN	NaN	NaN	NaN	NaN
2	2	1077175	1313524	2400	2400	2400.0 36 months	15.96	84.33	C	...	NaN	NaN	NaN	NaN	NaN	NaN
3	3	1076863	1277178	10000	10000	10000.0 36 months	13.49	339.31	C	...	NaN	NaN	NaN	NaN	NaN	NaN
4	4	1075358	1311748	3000	3000	3000.0 60 months	12.69	67.79	B	...	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 75 columns



Top Insight (Exploratory Data Analysis)

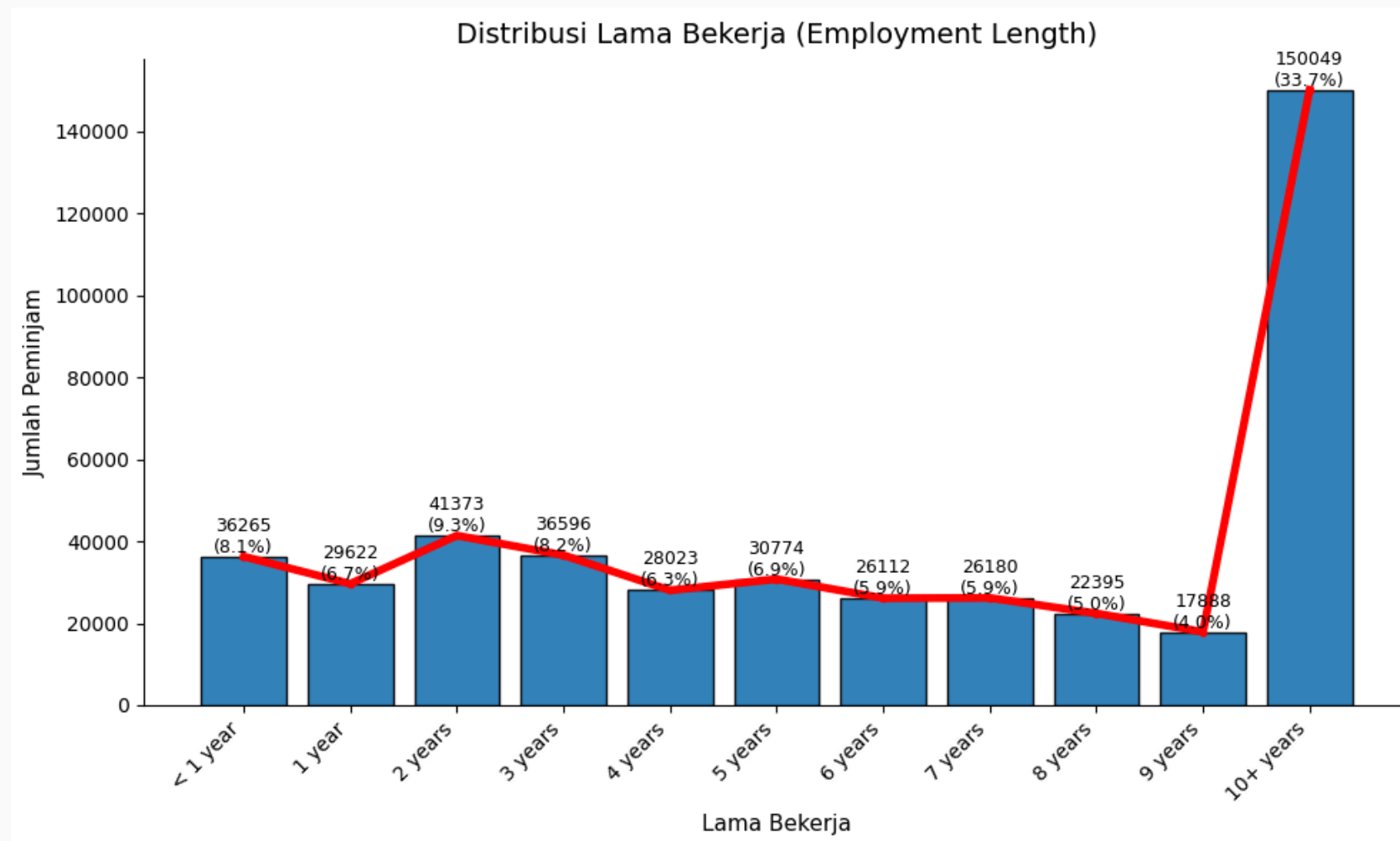
1. Distribusi Status Pinjaman



Mayoritas portofolio berada dalam kategori Pinjaman Baik (**88,8%**) dengan total 414.099 akun. Terdapat 11,2% (52.186 akun) yang masuk kategori Pinjaman Tidak Baik, yang menjadi fokus utama prediksi model.

Top Insight (Exploratory Data Analysis)

2. Profil Lama Bekerja Peminjam

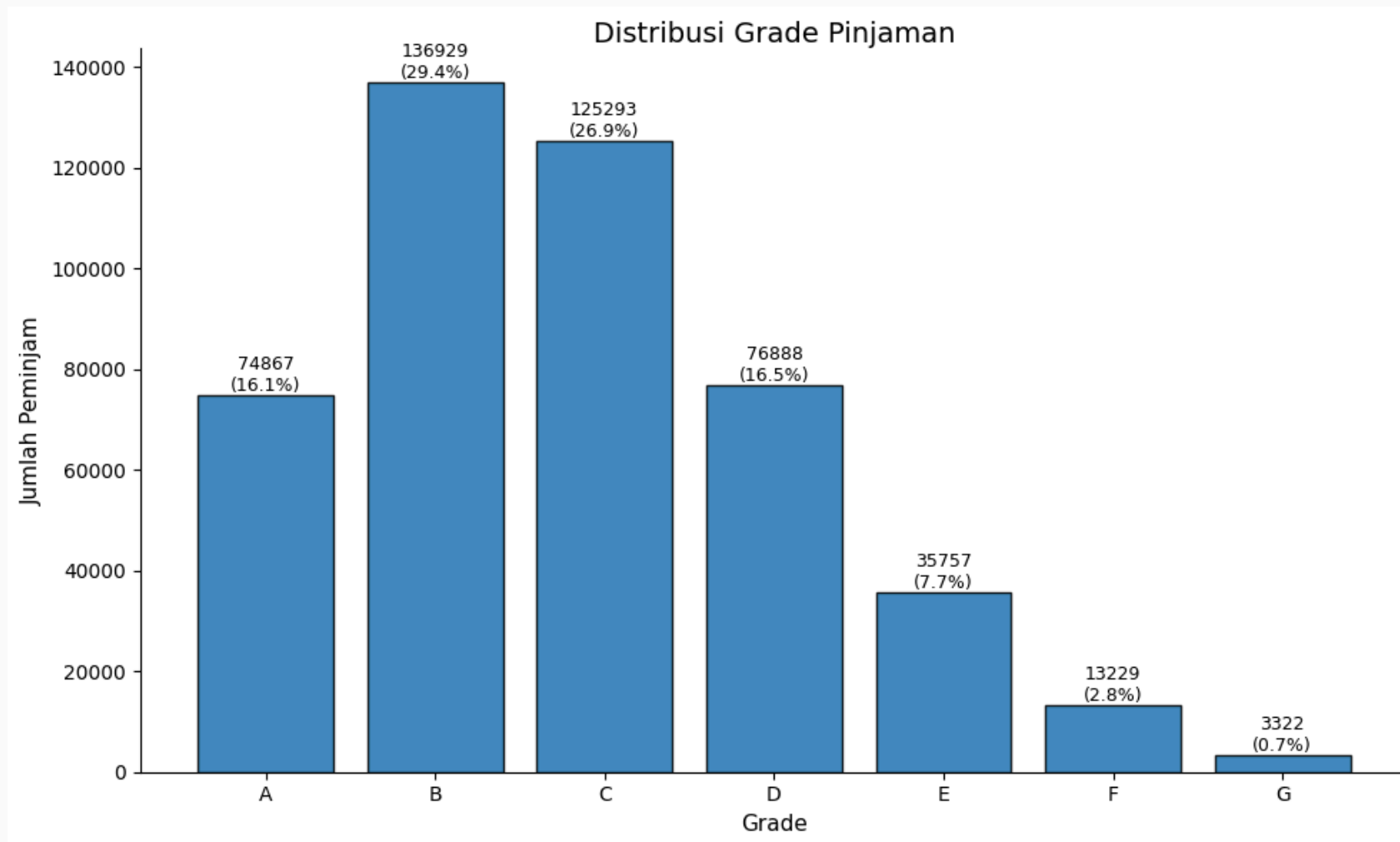


Peminjam dengan masa kerja 10+ tahun merupakan kelompok terbesar (**33,7%**), menunjukkan bahwa nasabah senior memiliki kecenderungan meminjam lebih tinggi.

Kelompok masa kerja 2 tahun dan 3 tahun berada di urutan berikutnya dengan proporsi masing-masing sekitar **9%** dan **8%**.

Top Insight (Exploratory Data Analysis)

3. Karakteristik Risiko Berdasarkan Grade



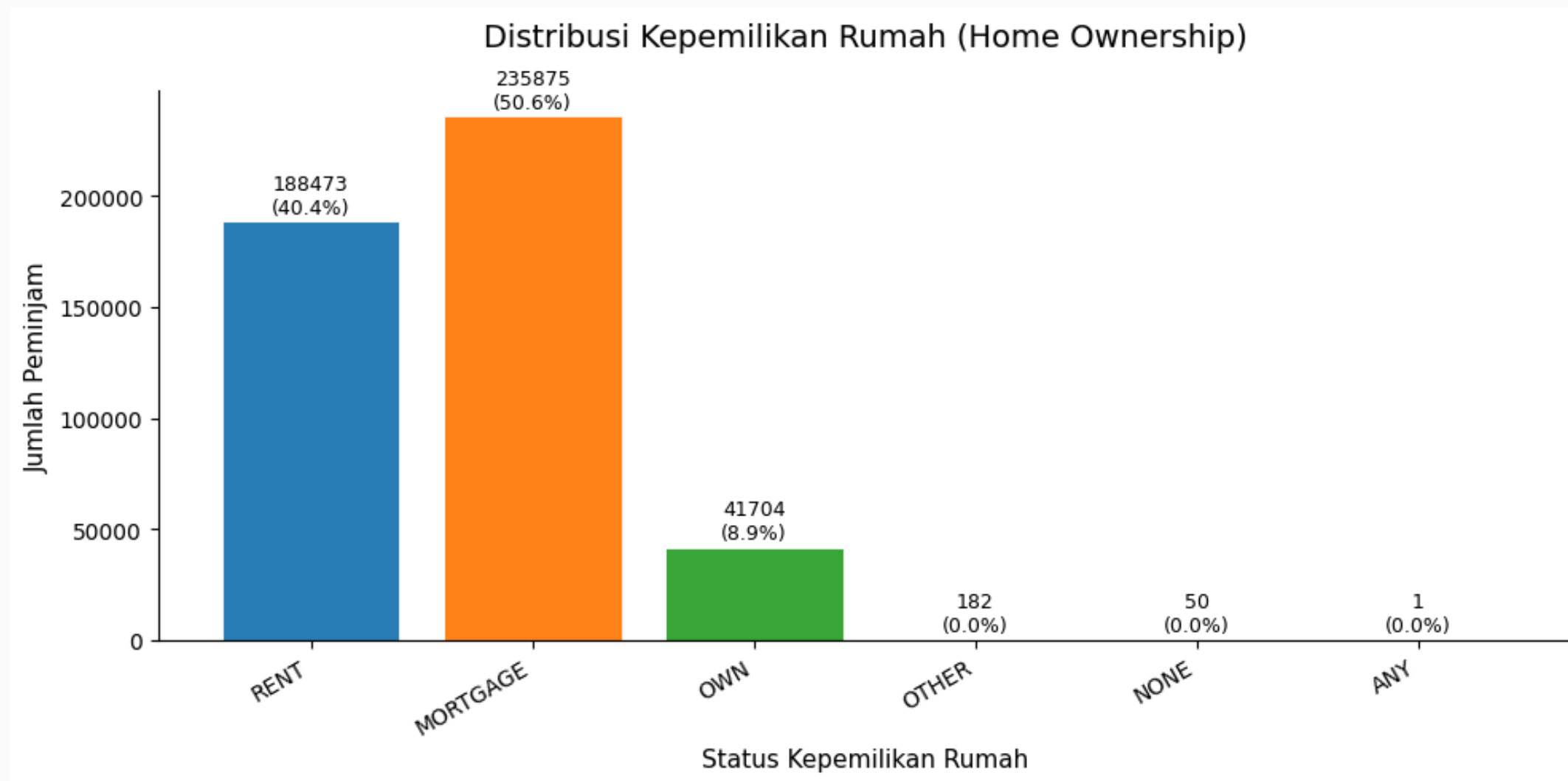
Grade B (29,4%) dan **Grade C (26,9%)** mendominasi dataset, mencerminkan nasabah dengan risiko menengah.



Nasabah Grade A (Risiko paling rendah) hanya mencakup 16,1% dari total populasi.

Top Insight (Exploratory Data Analysis)

4. Status Kepemilikan Rumah



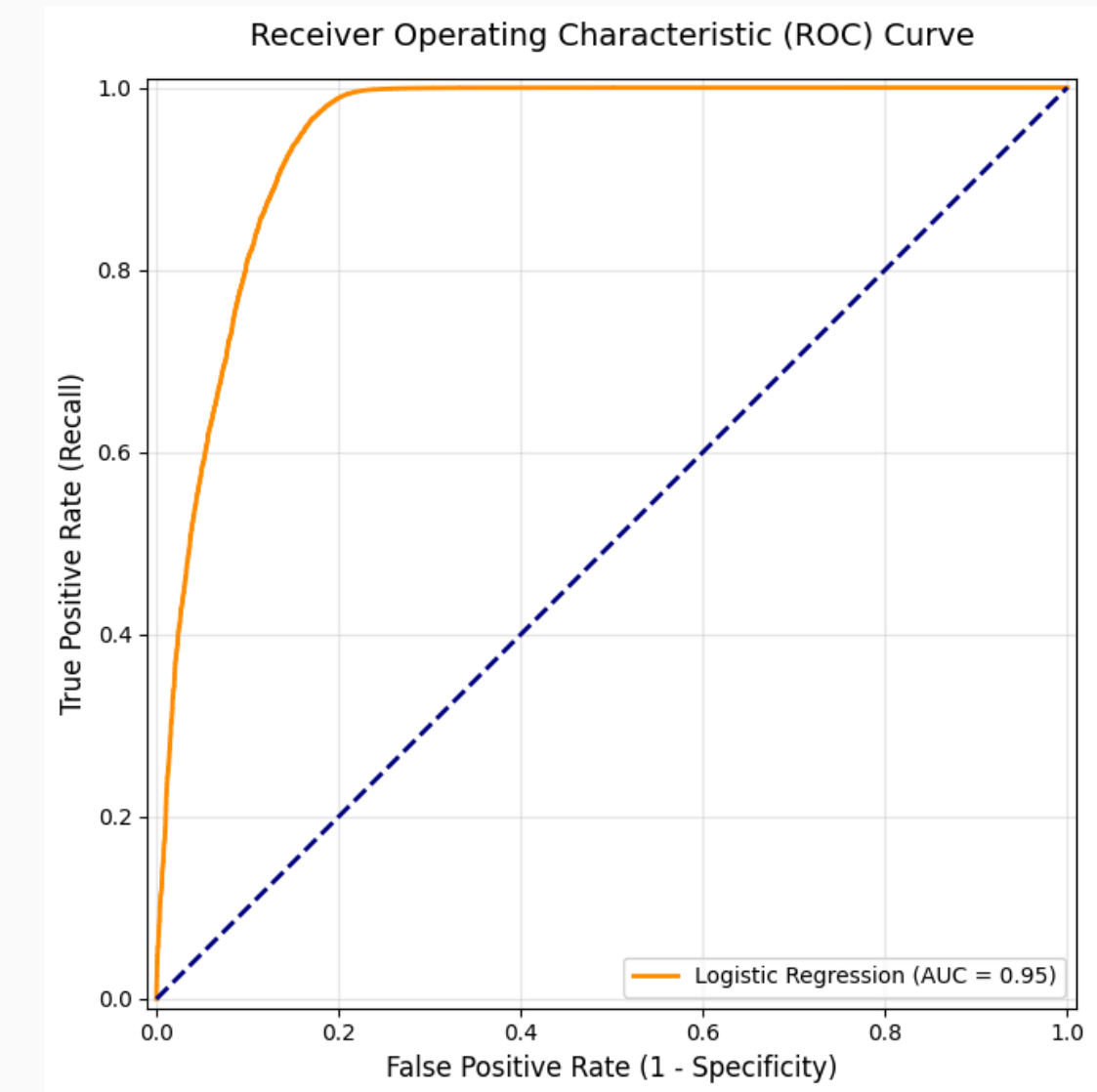
Sebanyak **50,6%** nasabah memiliki status MORTGAGE (KPR), diikuti oleh nasabah yang menyewa rumah (RENT) sebesar **40,4%**.

Our Final Model

Logistic Regression dipilih karena memberikan performa tertinggi dibandingkan Multinomial NB dan KNN, serta memberikan kemudahan interpretasi koefisien risiko. Berdasarkan hasil pengujian pada data test (20% dari dataset):

	precision	recall	f1-score	support
0	0.99	0.74	0.85	10437
1	0.97	1.00	0.98	82820
accuracy			0.97	93257
macro avg	0.98	0.87	0.92	93257
weighted avg	0.97	0.97	0.97	93257

Model memiliki Precision sebesar 0.99 untuk kelas 0, artinya model sangat jarang salah saat menandai seseorang sebagai risiko tinggi (minim False Positive).



Rekomendasi dan Kesimpulan

Kesimpulan

Model *Logistic Regression* yang dikembangkan sangat handal dengan skor **ROC AUC 0.95**, yang berarti model memiliki kemampuan **95%** untuk membedakan nasabah yang gagal bayar dan yang tidak.

Rekomendasi Strategis

- **Fokus pada Grade B & C** : Mengingat kelompok ini paling dominan, tim marketing dapat menyesuaikan suku bunga atau promosi khusus untuk menjaga loyalitas nasabah di segmen ini.
- **Monitoring Nasabah Non-MORTGAGE** : Perlu perhatian lebih bagi peminjam dengan status RENT, karena profil mereka mencakup hampir setengah dari populasi peminjam.
- **Pemeliharaan Model** : Disarankan melakukan retraining model setiap 6-12 bulan untuk menyesuaikan dengan kondisi ekonomi terbaru dan pergeseran pola perilaku peminjam.



Thank You Very Much

Referensi & Repositori GitHub

Detail lebih lanjut dan notebook Python lengkap tersedia di
repositori GitHub:

Dokumentasi
Github

