Rakamin Academy × id/x partners

PREDICTION MODELING CREDIT RISK ID/X PARTNERS - DATA SCIENTIST

Presented by :
Aldrian Syafril Lubis

Jakarta Barat, DKI Jakarta

aldriansyafril@gmail.com

linkedin.com/in/aldriansyafrillubis

Halo! Perkenalkan nama saya...

Aldrian Syafril Lubis

Saya adalah seorang Data Enthusiast dan mahasiswa tingkat akhir di Universitas Negeri Jakarta dengan latar belakang Pendidikan Administrasi Perkantoran.

Saya senang berkolaborasi dalam tim untuk memecahkan masalah berbasis data (data-driven) dan berharap dapat berkontribusi secara positif bagi pertumbuhan organisasi, sekaligus terus mengasah kemampuan saya di bidang data-driven insights dan business intelligence.

ABOUT COMPANY

Rakamin Academy

ID/X Partners didirikan pada tahun 2002 oleh mantan bankir dan konsultan manajemen yang memiliki pengalaman luas dalam siklus kredit dan manajemen proses, pengembangan skor, dan manajemen kinerja. Pengalaman gabungan kami 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 pemanfaatan solusi data analytic and decisioning (DAD) yang dikombinasikan dengan manajemen risiko dan disiplin pemasaran yang terintegrasi untuk membantu klien mengoptimalkan profitabilitas portofolio dan proses bisnis.

Layanan konsultasi komprehensif dan solusi teknologi yang ditawarkan oleh mitra id/x menjadikannya sebagai One-stop Provider.

id/x partners

DATA UNDERSTANDING

1. DATA UNDERSTANDING

Rakamin Academy

Data yang digunakan dalam proyek ini adalah data pinjaman yang diperoleh dari Rakamin Academy sebagai bagian dari program Project-based Internship.
Dataset ini mempunyai 466.285 entri dan 74 Kolom.

466.285 Entri
74 Kolom

Catatan:
1.) Dataset ini memiliki nilai hilang/NULL/NaN sebanyak 9.776.227 nilai
2.) Tidak memiliki Data duplikat
3.) Mempunyai 28 Variabel Outlier
4.) Dalam tahap ini dilakukan penyajian Statistika Deskriptif untuk kolom numerik serta ringkasan struktur data.

FEATURE ENGINEERING

2. FEATURE ENGINEERING

Rakamin Academy

Dalam tahap ini dilakukan proses drop variabel/kolom yang mempunyai proporsi nilai NULL sebesar lebih dari 30%, sehingga hasil Machine Learning dapat lebih akurat:

466.285 Entri
74 Kolom

466.285 Entri
50 Kolom

Serta melakukan drop pada variabel id dan member_id karena dataset tidak memiliki nilai duplikat jadi kedua variabel tersebut tidak digunakan.

DATA CLEANING

3. DATA CLEANING

Rakamin Academy

Dalam tahap ini dilakukan proses pembersihan data mulai dari:

1.) Mengelompokkan Status Pinjaman (loan_status)

- Kategori “Good” (1): ‘Current’ dan ‘Fully Paid’.
- Kategori “Bad” (0): Status pinjaman lain (Charged Off, Late, Default, dsb.).
- Menyimpan hasilnya di kolom baru loan_category, lalu menghapus kolom loan_status asli.

2.) Membersihkan Kolom emp_length

- Menghapus kata-kata seperti “years”, “+”, “<” agar tersisa angka saja.
- Mengubah tipe data menjadi float sehingga siap digunakan dalam analisis atau pemodelan.

3. DATA CLEANING

Rakamin Academy

3.) Mengubah Kolom term Menjadi Angka
- Sebelumnya berbentuk teks, misalnya "36 months".
- Menghapus kata "months" dan mengonversi menjadi tipe data integer agar mudah dianalisis (contoh: 36 atau 60).

4.) Memproses Kolom earliest_cr_line
- Kolom ini awalnya berbentuk teks (misalnya “Jan-85”).
- Dikoversi menjadi tipe datetime dengan format '%b-%y' agar bisa dilakukan perhitungan waktu.

5.) Menentukan Tanggal Acuan (Reference Date)
- Menggunakan ref_date = pd.to_datetime('2017-12-01') sebagai titik waktu perbandingan.

3. DATA CLEANING

Rakamin Academy

6.) Membuat Kolom mths_since_earliest_cr_line
- Menghitung selisih bulan antara earliest_cr_line_date dan ref_date.
- Hal ini membantu mengukur berapa lama nasabah telah memiliki catatan kredit sejak tanggal tersebut.

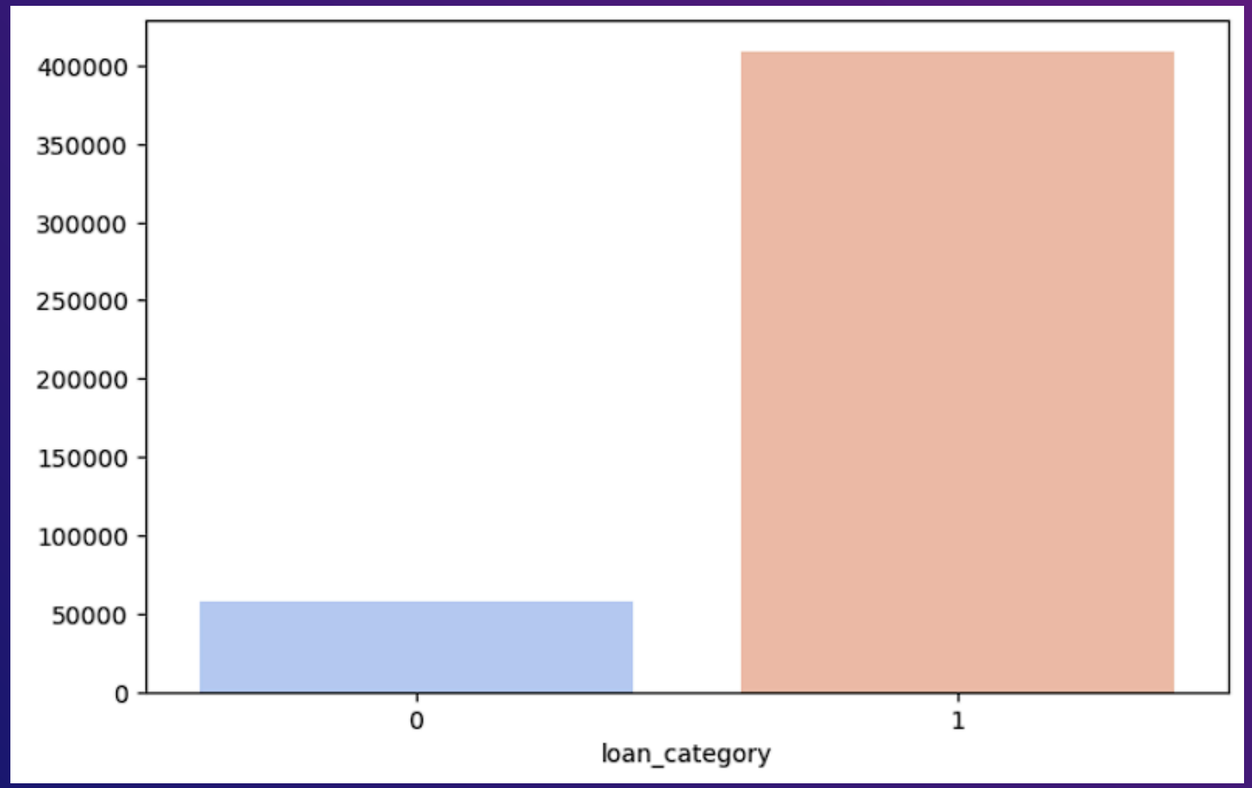
7.) Menangani Nilai Negatif pada Kolom ‘mths_since_earliest_cr_line’

8.) Mengonversi Kolom ‘issue_d’ Menjadi Tipe Datetime
- Kolom ‘issue_d’ yang semula berbentuk teks (mis. “Jan-15”) dikonversi menjadi format datetime ('%b-%y').
- Membuat kolom ‘mths_since_issue_d’ dengan menghitung selisih bulan antara ‘issue_d_date’ dan tanggal acuan (Desember 2017).

9.) Memproses Kolom ‘last_pymnt_d’
- Dikonversi menjadi datetime dengan format serupa ('%b-%y').

EXPLORATORY DATA ANALYSIS (EDA)

4. EXPLORATORY DATA ANALYSIS (EDA)

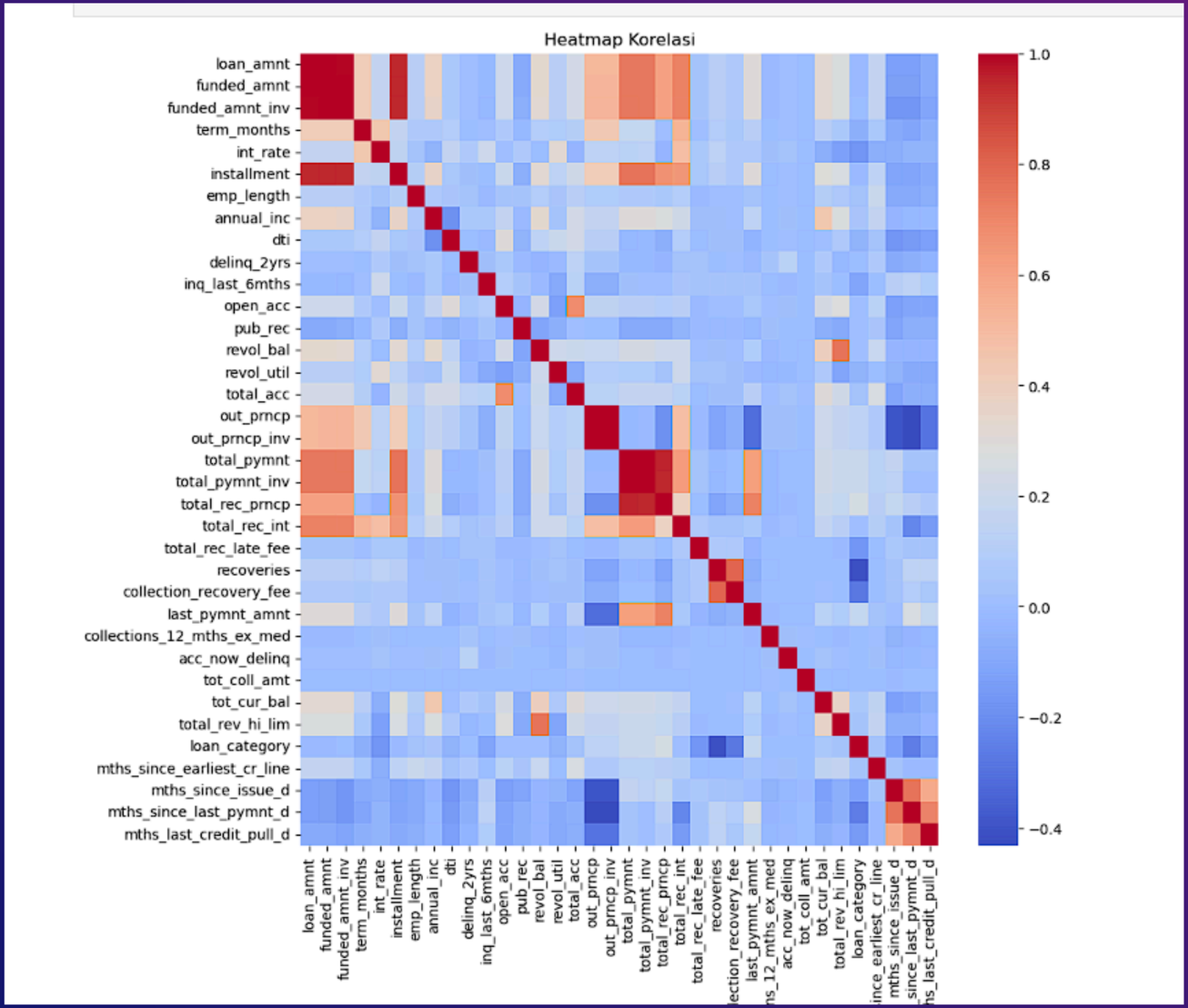


1.) Visualisasi Distribusi Pinjaman

Insight :

- Dari grafik, terlihat jumlah pinjaman kategori “Good” jauh lebih banyak dibanding “Bad”.
- Data bersifat Imbalance.
- Banyaknya pinjaman Good menunjukkan bahwa ID/X Partner memiliki proporsi nasabah yang cukup sehat secara umum.

4. EXPLORATORY DATA ANALYSIS (EDA)



2.) Visualisasi Korelasi antar Fitur

Insight :

- Adanya Kelompok Fitur yang Saling Berkaitan Erat
- Korelasi Rendah di Sebagian Besar Fitur
- Beberapa sel berwarna kebiruan menandakan korelasi negatif.

4. EXPLORATORY DATA ANALYSIS (EDA)

3.) Mengisi Nilai Kosong (Missing Values)

- Fungsi fill_missing_values(df) memeriksa tiap kolom:
  - Kolom numerik: Nilai kosong diganti dengan rata-rata (mean).
  - Kolom kategorik: Nilai kosong diganti dengan modus (mode).

4.) Membuat Matriks Korelasi (Absolute Value)

5.) Menyimpan Bagian ‘Upper Triangle’, Menemukan Fitur dengan Korelasi Tinggi, dan Menghapus Fitur Redundan.

6.)Hasil Akhir

- Tabel sekarang memiliki 30 kolom (dari sebelumnya 50), menunjukkan pengurangan fitur yang terlalu mirip.
- Dataset menjadi lebih ringkas dan siap untuk tahap pemodelan, mengurangi risiko overfitting serta mempercepat proses pelatihan model.

DATA PREPARATION

5. DATA PREPARATION

Rakamin Academy

Pada tahap ini data dipersiapkan agar siap digunakan dalam analisis atau pemodelan. Tujuannya adalah memastikan data yang tersedia sudah dalam kondisi yang baik.
Beberapa langkah yang dilakukan dalam tahap ini diantara lain:

1.) Memisahkan Kolom Kategorik dan Numerik

2.) One-Hot Encoding pada Kolom Kategorik

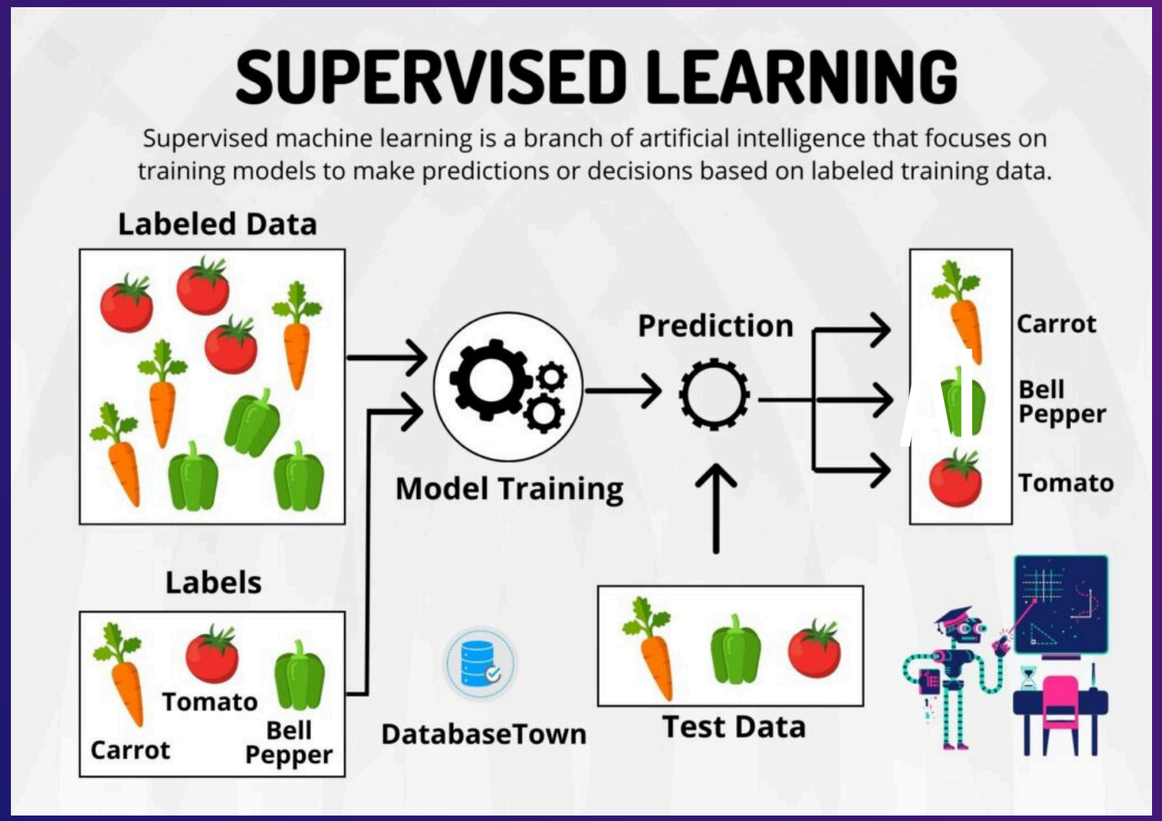
- Menggunakan pd.get_dummies(...) untuk mengonversi kolom kategorik menjadi beberapa kolom biner (0/1).
- Argumen drop_first=True digunakan agar menghindari dummy trap (mengurangi satu kolom dummy).

3.) Menstandarisasi Kolom Numerik

- Menggunakan StandardScaler untuk mengubah nilai tiap kolom numerik agar memiliki mean = 0 dan standar deviasi = 1.
- Membantu model machine learning (terutama algoritma berbasis jarak atau gradient-based) agar konvergensi lebih cepat dan hasil lebih stabil.

DATA MODELING

6. DATA MODELING



Credit:
https://databasetown.com/supervised-learning-algorithms/

Algoritma Machine Learning yang digunakan adalah:

1.) Logistic Regression
2.) Random Forest
3.) Gradient Boosting

6. DATA MODELING

Rakamin Academy

Hasil Machine Learning :

Algoritma Machine Learning	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.9270	0.9245	0.9983	0.9600	0.8345
Random Forest	0.9338	0.9309	0.9987	0.9636	0.8510
Gradient Boosting	0.9341	0.9317	0.9981	0.9637	0.8573

7. CONCLUSION

- Gradient Boosting cocok bila perusahaan menginginkan performansi tertinggi (accuracy, F1, dan ROC-AUC) dan memiliki sumber daya untuk melakukan tuning dan interpretasi lanjutan.
- Random Forest bisa menjadi pilihan apabila prioritasnya meminimalkan kesalahan dalam mengklasifikasikan pinjaman bermasalah (karena Recall tertinggi) dan tetap mempertahankan akurasi tinggi.
- Logistic Regression masih relevan jika interpretabilitas menjadi kunci (misalnya, untuk menjelaskan faktor risiko ke regulator atau stakeholder).

Secara keseluruhan, Gradient Boosting menawarkan performa paling unggul di hampir semua metrik, tetapi Random Forest dan Logistic Regression juga menunjukkan hasil yang sangat baik. Pemilihan model ideal tergantung tujuan bisnis (minimasi risiko, kebutuhan interpretasi, atau ketersediaan sumber daya).

BUSINESS RECOMENDATION

8. BUSINESS RECOMENDATION

1. Implementasi Model Ensemble (Gradient Boosting atau Random Forest)
    a. Gunakan model dengan performa terbaik secara keseluruhan (Gradient Boosting) atau recall tertinggi (Random Forest), sesuai prioritas perusahaan.
    b. Lakukan hyperparameter tuning agar performa model optimal.
2. Gunakan Logistic Regression sebagai Pendamping
    a. Untuk memudahkan interpretasi dan menjelaskan faktor-faktor penentu risiko kredit kepada stakeholder, terutama jika dibutuhkan oleh regulator atau tim manajemen.
3. Siapkan Proses Monitoring
    a. Buat pipeline end-to-end untuk pemrosesan data, scoring, dan monitoring performa model secara berkala.
    b. Lakukan retraining jika data nasabah atau tren pasar berubah signifikan.

Dengan strategi ini, IDX Partner dapat meningkatkan keakuratan penilaian risiko kredit, meminimalkan potensi kerugian, dan memperkuat kepercayaan stakeholder terhadap proses penyaluran kredit.

Rakamin Academy × id/x partners

TERIMA KASIH

Presented by :
Aldrian Syafril Lubis