LAPORAN AKHIR

FINAL PROJECT

Load Prediction Based on Customer Behavior with Machine Learning Model

Diajukan untuk memenuhi persyaratan kelulusan

Bootcamp Data Science

Disusun oleh :

Dea Arighie Permatasari

Yulisnawati

Hafizh Kamaluddin Abdillah

Daniel Toby Sahala

NEVORYA

TAHUN 2025

Abstraksi

Proyek ini mengembangkan model machine learning untuk memprediksi risiko kredit nasabah guna menurunkan angka kredit macet. Dengan menggunakan dataset profil nasabah, model Random Forest berhasil mencapai akurasi 96%. Sistem diimplementasikan dalam bentuk dashboard dan website yang ditambahkan dengan fitur Generative AI untuk rekomendasi produk keuangan. Hasil proyek menunjukkan potensi signifikan dalam meningkatkan efisiensi, mengurangi risiko NPL, dan mendukung keputusan kredit secara otomatis.

Kata Pengantar

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas kelancaran dalam menyelesaikan Final Project ini. Kami mengucapkan terima kasih kepada Kak Nurul Akbar selaku mentor yang telah memberikan bimbingan serta instruktur Bootcamp Rakamin yang telah mendukung jalannya proyek ini. Tak lupa, kami juga berterima kasih kepada Tutor-tutor Bootcamp Rakamin yang telah memberikan bimbingan serta rekan satu tim yang bekerja sama dengan penuh dedikasi pada Final Project ini

Daftar Isi

Daftar lain-lain

Dapat ditambahkan berbagai daftar yang dibutuhkan seperti daftar tabel, daftar gambar, daftar algoritma, daftar padanan istilah, daftar singkatan, daftar istilah, daftar simbol. Khusus untuk daftar pustaka, dapat diletakkan setelah bab Penutup, sebelum lampiran.Jika hanya terdapat satu gambar atau satu tabel, maka tidak perlu dibuat daftar gambar atau daftar tabel. Setiap daftar, misal daftar gambar, daftar tabel, daftar istilah dan singkatan, semuanya diletakkan pada halaman terpisah.

Bab I Pendahuluan

I.1 Latar belakang

Sektor pinjaman kredit dalam industri perbankan mengalami pertumbuhan yang pesat, didorong oleh meningkatnya permintaan masyarakat. Tren ini juga mendorong banyak startup untuk turut terjun ke pasar dengan menawarkan layanan pinjaman berbasis teknologi. Pada saat yang sama, peningkatan jumlah pengajuan pinjaman dan konsumsi turut menyebabkan meningkatnya kerugian akibat kredit macet. Kredit adalah pinjaman yang diberikan oleh bank atau lembaga keuangan kepada individu atau nasabah, yang harus dilunasi dalam jangka waktu tertentu, dengan atau tanpa bunga [1]. Kredit sering digunakan untuk berbagai keperluan, seperti konsumsi, pendidikan, kesehatan, perjalanan, dan bisnis.

Bank dan lembaga keuangan perlu memanfaatkan penelitian untuk merancang model yang efektif, dengan menggunakan data yang ada guna menghasilkan prediksi yang kuat demi meminimalkan risiko kredit macet di tengah meningkatnya jumlah pengajuan pinjaman dan persaingan yang ketat. Dengan memanfaatkan berbagai teknik pemodelan prediktif modern, institusi keuangan dapat memahami kebiasaan, pola penggunaan uang, indikator risiko gagal bayar, dan karakteristik penting lainnya dari pemohon pinjaman. Berbagai penelitian telah dilakukan untuk mengidentifikasi variabel-variabel kunci yang mempengaruhi ketepatan pembayaran pinjaman, dan temuan-temuan ini penting untuk membantu bank dalam mengoptimalkan keuntungan.

I.2 Lingkup

Lingkup proyek difokuskan pada analisis dan pemodelan prediktif terhadap dataset ‘Loan Risk Prediction Based on Customer’ . Dataset yang digunakan mencakup variabel-variabel demografis, kepemilikan aset hingga pengalaman dalam bekerja

Proyek ini hanya mencakup tahapan teknis terkait data science, seperti eksplorasi data, pembersihan data, pemodelan machine learning, evaluasi model, dan visualisasi hasil. Penjabaran proses teknis yang dilakukan selama proyek dijelaskan lebih rinci pada sub bab II.3.

I.3 Tujuan

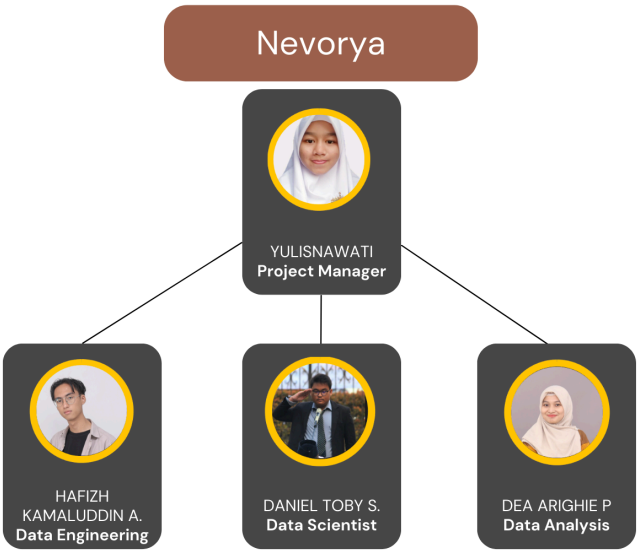
Tujuan dari pembuatan machine learning pada kasus risiko kredit ini adalah :

1. Meningkatkan akurasi prediksi risiko kredit dari 70% (manual) menjadi 89% dalam 12 bulan dengan machine learning.
2. Mengurangi tingkat gagal bayar (default rate) dari 12.5% (manual) ke <5% melalui model AI dalam 1 tahun.
3. Mempercepat waktu persetujuan pinjaman dari rata-rata 1 minggu menjadi beberapa detik melalui otomatisasi AI dalam 12 bulan.
4. Menurunkan kebutuhan tenaga analis kredit hingga 90% dengan otomatisasi pemrosesan data dalam 1 tahun.
5. Menghemat biaya operasional hingga Rp 4,3 Miliar per bulan dengan pengurangan kebutuhan tenaga kerja manual.

Bab II Nevorya

II.1 Struktur Organisasi

Tim proyek terdiri dari empat anggota dengan peran sebagai berikut:



II.2 Lingkup Pekerjaan

Setiap anggota tim memiliki tugas spesifik dalam pengolahan data, pengembangan model, serta interpretasi hasil dan dokumentasi. berikut lingkup pekerjaan setiap peran :

1. Project Manager : Memastikan bahwa semua dokumentasi proyek terorganisir dengan baik
2. Data Engineer : Menangani semua aspek pembersihan, transformasi, dan integrasi dari berbagai sumber dan memastikan kumpulan data terstruktur dengan baik untuk analisis
3. Data Scientist : Mengembangkan model *machine learning* atau solusi analitis tingkat lanjut berdasarkan kebutuhan proyek.
4. Data Analyst : Menentukan masalah bisnis dan tujuan analisis serta mengidentifikasi tren dan kebutuhan industri

II.3 Deskripsi Pekerjaan

1. Project Manager

   Sebagai Project Manager, tugas utama adalah memastikan kelancaran proyek dengan mengawasi timeline, koordinasi tim, dan dokumentasi. Tanggung jawab utama meliputi:

   a. Mengawasi alur kerja proyek dan memastikan setiap tugas selesai sesuai jadwal.
   b. Mengelola komunikasi dalam tim dan mendokumentasikan semua keputusan proyek.
   c. Menjamin kualitas proyek secara keseluruhan agar sesuai dengan tujuan bisnis, performa model, dan pelaporan hasil

2. Data Engineer

   Sebagai Data Engineer, fokus utama adalah menangani aspek teknis pengolahan data agar siap digunakan oleh tim Data Scientist dan Data Analyst. Tanggung jawab utama meliputi:

   a. Melakukan data cleaning, transformasi, dan integrasi dari berbagai sumber untuk memastikan dataset siap digunakan.
   b. Melakukan Exploratory Data Analysis (EDA) guna menemukan tren, hubungan, dan potensi permasalahan dalam data.
   c. Membuat visualisasi awal untuk mengkomunikasikan temuan kepada tim.

3. Data Scientist

   Sebagai Data Scientist, peran utama adalah membangun model machine learning dan memastikan model bekerja dengan optimal. Tanggung jawab utama meliputi:

   a. Mengembangkan model machine learning atau solusi analitis berdasarkan kebutuhan proyek.
   b. Melakukan eksperimen dengan berbagai algoritma (misalnya, decision trees, random forests, neural networks) untuk mendapatkan model terbaik.

c. Menggunakan data uji untuk validasi model serta mengoptimalkan akurasi menggunakan cross-validation dan evaluasi metrik.
d. Bekerja sama dengan Data Engineer untuk memastikan fitur yang digunakan dalam model dapat meningkatkan performa prediksi

4. Business / Data Analyst

   Sebagai Business / Data Analyst, peran utama adalah menerjemahkan hasil analisis data menjadi wawasan bisnis yang dapat diimplementasikan. Tanggung jawab utama meliputi:

   a. Mengidentifikasi masalah bisnis dan tujuan analisis, serta memahami tren industri.
   b. Menerjemahkan hasil analisis data menjadi rekomendasi bisnis yang dapat diimplementasikan.
   c. Menentukan indikator keberhasilan proyek (misalnya, ukuran dan metrik yang digunakan untuk menilai performa model).

II.4 Jadwal Kerja

Proyek ini berlangsung selama 5 minggu dengan pembagian kerja sebagai berikut:

- Minggu ke-1 : Stage 0 - Memahami masalah industri yang dapat dipecahkan oleh dataset dan merumuskan pertanyaan analitis
- Minggu ke-2 : Stage 1 - Melakukan prapemrosesan data, termasuk pembersihan, penggabungan set data, *feature engineering* dan visualisasi data.
- Minggu ke-3 : Stage 2 - Membangun model prediktif atau melakukan analisis eksploratif untuk memecahkan masalah
- Minggu ke-4 : Stage 3 - Memvalidasi model dengan data uji dan mengevaluasi metrik kinerja
- Minggu ke-5 : Stage 4 - Mempersiapkan presentasi akhir dan deploy model

Bab III Implementasi Model ML untuk Memprediksi Loan Risk Berdasarkan Profil Nasabah

Pada bagian ini dijelaskan mengenai pelaksanaan project Final Project Bootcamp Rakamin yang mencakup deskripsi persoalan, tahapan proses pengerjaan, serta pencapaian hasil yang diperoleh dari implementasi proyek. Penjelasan difokuskan pada langkah-langkah teknis dan analitis yang ditempuh selama proyek berlangsung.

III.1 Deskripsi Permasalahan

Suatu perusahaan memiliki permasalahan dalam mengelola risiko kredit (Non-Performing Loan/NPL) yang mencapai 12,3%. Angka ini jauh melampaui yang sudah ditetapkan oleh RBI (Reserve Bank of India) / setara dengan Otoritas Jasa Keuangan (OJK) di Indonesia, yaitu di bawah 5%. Oleh karena itu, diperlukan solusi analisis data yang mampu mengidentifikasi faktor-faktor penyebab tingginya risiko kredit dengan pendekatan machine learning berbasis data

III.2 Proses Pelaksanaan Proyek dan Penggunaan Teknologi

Proyek dilaksanakan dalam beberapa tahapan utama, yaitu:

1. Pemahaman Data
    a. Analisis struktur dan kualitas data
    b. Identifikasi outlier dan data yang tidak konsisten
2. Eksplorasi Data (EDA)
    a. Visualisasi distribusi dan hubungan antar variabel menggunakan Seaborn dan Matplotlib
    b. Analisis korelasi antar fitur terhadap target
3. Pra-pemrosesan Data
    a. Mengatasi Outlier, duplikat data dan ketidakkonsistenan data
    b. Encoding fitur kategorikal dengan metode Label Encoding dan Target Encoding

c. Normalisasi fitur numerik agar skala data seragam
d. Mengatasi ketidaseimbangan data
e. Membuat fitur tambahan : *'age_group'* dan *'experience_age_ratio'*

4. Pemodelan dan Evaluasi
    a. Membangun beberapa model prediktif: Logistic Regression, Random Forest, XGBoost, SVM, AdaBoost dan Gradient Boosting
    b. Evaluasi model menggunakan metrik akurasi, precision, recall, f1-score dan Roc Auc
    c. Melakukan hyperparameter tuning pada model terbaik
    d. Evaluasi hasil model dengan menggunakan data set validasi
5. Implementasi model dan Visualisasi
    a. Menyusun prototype dashboard sederhana untuk visualisasi hasil prediksi
    b. Membangun Website untuk penggunaan model serta menambahkan fitur Generative AI pada Website untuk memberikan rekomendasi produk keuangan berdasarkan profil nasabah

III.3 Pencapaian Hasil dan Kaitannya dengan Tujuan Proyek

Selama pelaksanaan proyek, diperoleh sejumlah pencapaian berikut:

1. Model terbaik yang dibangun (Random Forest) menghasilkan akurasi hingga 96% dalam memprediksi risiko kredit.
2. Dashboard sederhana berhasil dibuat untuk membantu visualisasi prediksi yang dapat membantu sebagai sistem pendukung keputusan.
3. Website prediksi berhasil dibuat untuk membantu perusahaan dalam melakukan otomasi prediksi risiko kredit nasabah sehingga pekerjaan menjadi lebih efisien
4. Generative AI untuk membantu memberikan rekomendasi produk keuangan untuk nasabah berhasil dibuat

Bab IV Penutup

IV.1 Kesimpulan

Kesimpulan dari hasil project kami adalah :

1. Model Random Forest berhasil menjawab permasalahan kredit macet dengan mengidentifikasi nasabah berisiko sejak awal proses.
2. Project kami memberikan dasar untuk peningkatan efisiensi dalam persetujuan kredit dan personalisasi produk berdasarkan profil risiko.
3. Machine Learning dapat memberikan nilai strategis bagi stakeholder melalui pengurangan potensi NPL dan peningkatan akurasi penawaran produk keuangan.

IV.2 Saran

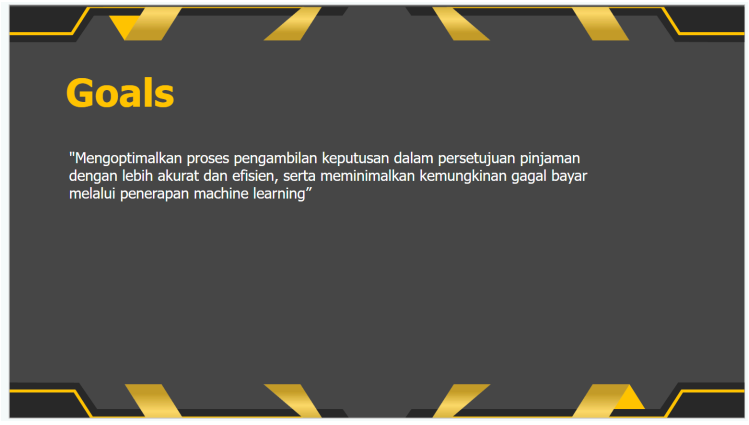
Berikut beberapa saran berupa rekomendasi bisnis untuk mendukung keberhasilan implementasi website Risk Flag Predict ke depannya :

1. Gunakan platform MLOps (seperti MLflow) untuk monitoring model, retraining, dan deployment otomatis.
2. Sediakan dokumentasi dan pelatihan ringan untuk tim bisnis atau marketing agar memahami output model.
3. Lakukan validasi rutin terhadap asumsi dan kebutuhan pengguna melalui survei, wawancara, atau A/B testing.

IV.2.1 Referensi

[1] Anand, M., Velu, A., & Whig, P. (2021). Prediction of Loan Behaviour with Machine Learning Models for Secure Banking.

Bab V Lampiran A. PPT Final Project Presentation



Objectives

- **Meningkatkan akurasi prediksi risiko kredit** dari **70%** (manual) menjadi **89%** dalam **12 bulan** dengan machine learning.
- **Mengurangi tingkat gagal bayar (default rate)** dari **12.5% (manual) ke <5%** melalui model AI dalam **1 tahun**.
- **Mempercepat waktu persetujuan pinjaman** dari rata-rata **1 minggu menjadi 30 detik** melalui otomatisasi AI dalam **12 bulan**.
- **Menurunkan kebutuhan tenaga analis kredit** hingga **75%** dengan otomatisasi pemrosesan data dalam **1 tahun**.
- **Menghemat biaya operasional** hingga **Rp 19,37 Miliar per bulan** dengan pengurangan kebutuhan tenaga kerja manual.
- **Menghemat potensi kerugian kredit bermasalah** sebesar **Rp 5 Miliar per bulan** melalui peningkatan akurasi deteksi risiko kredit.

Business Metric

- **Akurasi model prediksi risiko kredit (%)** =89% (Seberapa tepat machine learning dalam mengklasifikasikan peminjam berisiko tinggi)
- **Default Rate (%)** = < 5% (Persentase pinjaman yang gagal bayar dibandingkan total pinjaman yang diberikan)
- **Approval Time** = 30 detik (Waktu yang dibutuhkan dari pengajuan hingga keputusan persetujuan)
- **Pengurangan SDM** Tenaga analisis kredit manual hingga 75%.
- **Biaya** Operasional = hemat 19.37 Miliar
- **Potensi kerugian kredit** (Total kerugian gagal bayar) = 5 Miliar

Dataset Exploration

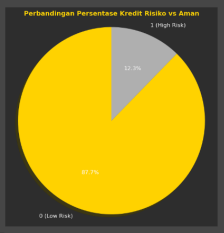
Dataset Exploration dilakukan untuk memahami dataset, menemukan pola, serta hubungan antar fitur. Dari hasil eksplorasi, dapat disimpulkan beberapa hal penting:

- Risk_Flag adalah **target variable** untuk klasifikasi (0 = rendah risiko, 1 = tinggi risiko).
- Dataset terdiri dari **252.000 data** dengan **13 kolom**, tanpa missing value.
  Terdapat kombinasi fitur **numerik** (seperti Income, Age, Experience) dan **kategori** (seperti Profession, House_Ownership, Married/Single).
- Beberapa fitur mencerminkan **stabilitas finansial** dan **gaya hidup**, yang berpotensi berpengaruh terhadap risiko.

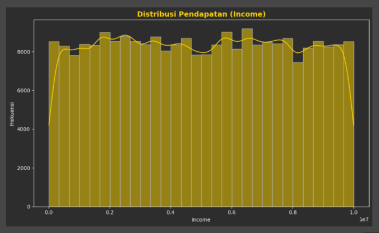
```
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Id                 252000 non-null  int64
 1   Income             252000 non-null  int64
 2   Age                252000 non-null  int64
 3   Experience         252000 non-null  int64
 4   Married/Single     252000 non-null  object
 5   House_Ownership    252000 non-null  object
 6   Car_Ownership      252000 non-null  object
 7   Profession         252000 non-null  object
 8   CITY               252000 non-null  object
 9   STATE              252000 non-null  object
 10  CURRENT_JOB_YRS    252000 non-null  int64
 11  CURRENT_HOUSE_YRS  252000 non-null  int64
 12  Risk_Flag          252000 non-null  int64
```

Distribusi Data

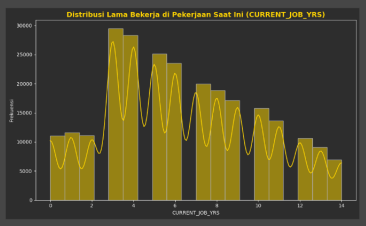
- Berdasarkan pie chart berikut, didapatkan bahwa distribusi nilai unik pada target sangat timpang dimana nilai 0 atau false lebih dominan daripada nilai 1 atau true

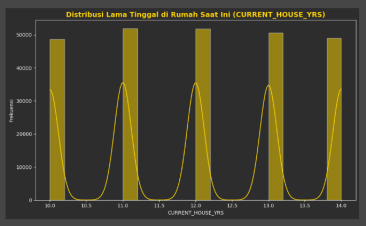


- Distribusi income: Terlihat cukup merata yang menunjukkan bahwa pendapatan tersebar secara luas di populasi data.



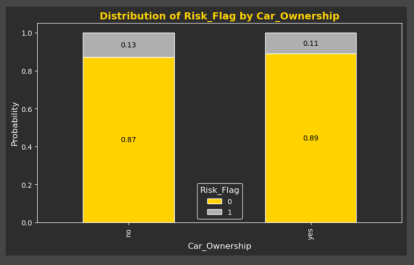
Distribusi Data





- Lebih banyak terkonsentrasi di angka yang lebih rendah, menunjukkan bahwa sebagian besar orang memiliki masa kerja yang relatif singkat.
- Ada pola berkala yang muncul (10, 12, 14 tahun) yang mungkin terkait dengan kebijakan kredit perumahan atau tren perpidahan rumah.

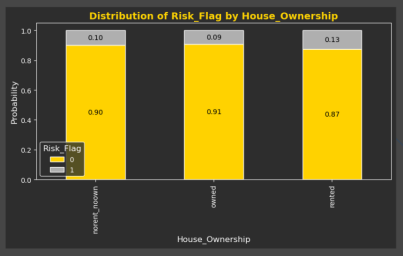
Risk Berdasarkan Kepemilikan Mobil



- Mayoritas pada kedua kategori (punya/tidak punya mobil) memiliki **risiko rendah**.
- Selisih **Risk_Flag = 1** hanya **2%** → **pengaruh kecil terhadap risiko**.
- Fitur ini **kurang signifikan** secara mandiri dalam prediksi risiko.

Risk Berdasarkan Kepemilikan Rumah

- Semua kategori menunjukkan dominasi **risiko rendah**.
- Perbedaan proporsi **Risk_Flag = 1** kecil (antara 9%–13%).
- **Kepemilikan rumah tidak terlalu menentukan** tingkat risiko.



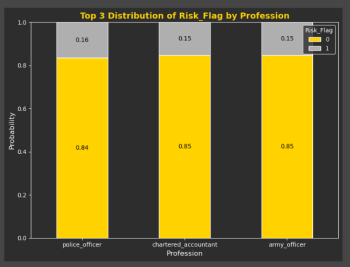
Risk Berdasarkan Profesi

Kesimpulan

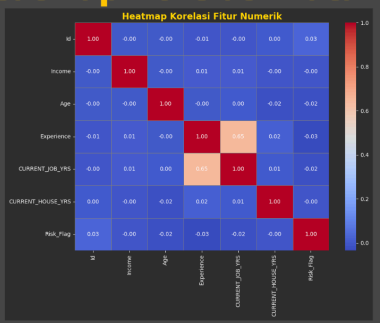
Dari analisis proporsi Risk_Flag = 1, ditemukan bahwa:

- **Police Officer** memiliki persentase risiko kredit tertinggi (16%) dibandingkan profesi lain.
- Diikuti oleh **Chartered Accountant** dan **Army Officer** (masing-masing 15%).

Temuan ini dapat menjadi pertimbangan dalam pengambilan keputusan risiko kredit berdasarkan jenis profesi.



Heatmap Korelasi Data Numerik



Korelasi rendah dengan Risk_Flag: Semua variabel numerik (Income, Age, Experience, dll.) memiliki korelasi lemah dengan Risk_Flag, menunjukkan bahwa faktor-faktor ini tidak berpengaruh besar terhadap risiko.

Korelasi kuat antara Experience dan CURRENT_JOB_YRS (0.65): Semakin lama bekerja, semakin banyak pengalaman.

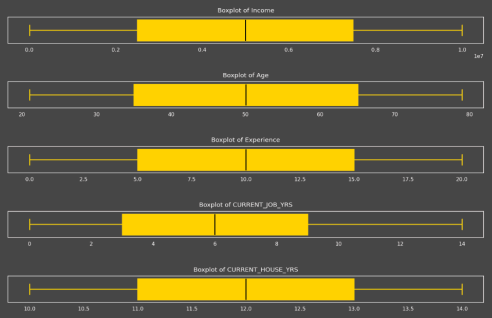
Tidak ada korelasi antara Income dengan faktor lain: Pendapatan tidak terkait signifikan dengan Age, Experience, atau Risk_Flag.

Data Pre-Processing

Proses data preprocessing sangat penting dalam proyek data science, karena data mentah biasanya belum siap digunakan langsung untuk modeling. Melalui preprocessing, kita bisa membersihkan data dari error, mengatasi data yang hilang atau duplikat, serta mengelola outlier yang bisa mengganggu hasil model.

- Handling Outlier
- Handling duplicate
- Handling Inconsistent Data
- Encode
- Handling Imbalance Data

Handling Outlier & Duplicate



Berdasarkan box plot yang didapatkan, data sudah bersih dan terlihat tidak adanya outlier pada setiap feature numerik yang ada pada dataset

Selain itu dengan fungsi remove duplicate kita melakukan penghapusan pada data yang duplikat

Handling Inconsistent Data

```
Inconsistent Entries Check:
CITY:
CITY
Vijayanagaram      1259
Bhopal             1208
Bulandshahr        1185
Saharsa[29]        1180
Vijayawada         1172
                   ...
Ujjain              486
Warangal[11][12]    459
Bettiah[33]         457
Katni               448
Karaikudi           431
```

```
Inconsistent Entries Check:
STATE:
STATE
Uttar_Pradesh      28400
Maharashtra        25562
Andhra_Pradesh     25297
West_Bengal        23483
Bihar              19780
Tamil_Nadu         16537
Madhya_Pradesh     14122
Karnataka          11855
Gujarat            11408
Rajasthan           9174
Jharkhand           8965
Haryana             7890
Telangana           7524
Assam               7062
Kerala              5805
Delhi               5490
Punjab              4720
Odisha              4658
Chhattisgarh        3834
Uttarakhand         1874
Jammu_and_Kashmir   1780
Puducherry          1433
Mizoram              849
Manipur              849
Himachal_Pradesh     833
Tripura              809
Uttar Pradesh[5]     743
```

Terdapat inconsistency data pada feature CITY dan STATE, dimana pada nama city dan state mengandung format "..[11]"

hal ini dapat menyebabkan 1 nama lokasi sebagai dua entitas yang berbeda. Maka dari itu kita melakukan handling inconsisten data untuk feature CITY dan STATE

Feature Engineering

1. age_group

Penambahan feature "age_group" dilakukan untuk mengelompokkan nilai umur ke dalam kategori seperti young, adult, middle_aged, dan senior. Tujuannya adalah untuk menyederhanakan data numerik menjadi bentuk kategorikal yang lebih mudah dianalisis dan diinterpretasikan

Rentang Umur	Label
<25	young
25 - 39	adult
40 - 59	middle_aged
60 ≥	senior

Feature Engineering

1. age_group

Pembagian kelompok usia tidak berdasarkan pada satu teori tunggal, tapi merupakan gabungan dari beberapa teori yakni:

1. Teori Perkembangan Psikososial-Erikson
(Sumber: Kategori Umur Menurut Depkes RI | PDF)
2. Klasifikasi Usia Menurut WHO dan Kementerian Kesehatan RI
(Sumber: Kategori Umur Menurut Depkes RI | PDF)
3. Klasifikasi Usia dalam penelitian Demografi
(Sumber: Klasifikasi Kelompok Umur Manusia Berdasarkan Analisis Dimensifraktal Box Counting Dari Citra Wajah Dengan Deteksi Tepi Canny - Neliti)

yang mana pembagian kelompok usia in merupakan konvensi umum yangi sering dipakai dalam berbagai bidang terutama dalam bidang: sosiologi, psikologi perkembangan, demografi serta pemasaran dan kebijakan publik.

Feature Engineering

1. age_group

Young (Muda):

- Biasanya mencakup usia **0-18 tahun**, termasuk bayi, anak-anak, dan remaja.
- Masa ini sering dikaitkan dengan pertumbuhan fisik dan perkembangan mental yang cepat.

Adult (Dewasa):

- Usia **19-40 tahun** sering dianggap sebagai dewasa awal.
- Masa ini ditandai dengan kemandirian, produktivitas, dan pembentukan identitas sosial.

Middle Age (Paruh Baya):

- Usia **41-60 tahun** biasanya masuk dalam kategori ini.
- Masa ini sering dikaitkan dengan stabilitas karier, keluarga, dan refleksi terhadap pencapaian hidup.

Older Age (Lansia):

- Usia **61 tahun ke atas** termasuk dalam kategori ini.
- Fokus pada kesehatan, kebijaksanaan, dan kontribusi sosial.

Feature Engineering

2. experience_age_ratio

Penambahan feature "experience_age_ratio" dilakukan untuk merepresentasikan proporsi pengalaman kerja seseorang terhadap usianya. Rasio ini memberikan insight seberapa besar bagian hidup seseorang yang telah dihabiskan untuk bekerja

age	experience
23	3
40	10
66	4
41	2
47	11

experience_age_ratio
0.130435
0.250000
0.060606
0.048780
0.234043

Feature Engineering

2. experience_age_ratio

Efek Penambahan Fitur ini ke Model:

1. Meningkatkan signal prediksi stabilitas finansial .
2. Mendeteksi inkonsistensi (misalnya usia 25 tapi pengalaman 20 tahun = data error)
3. Bisa membantu memecahkan ambiguitas ketika:
   a. Income tinggi tapi umur muda
   b. Usia tua tapi pengalamannya rendah

Encoding Data

1. Label Encoding pada feature "Married/Single", "House_Ownership", "Car_Ownership" dan "age_group"

Married/Single		House_Ownership		Car_Ownership		age_group	
married	1	rented	2	no	1	adult	0
single	0	owned	1	ye	0	middle_aged	1
		norent_noown	0			senior	2
						young	3

Encoding Data

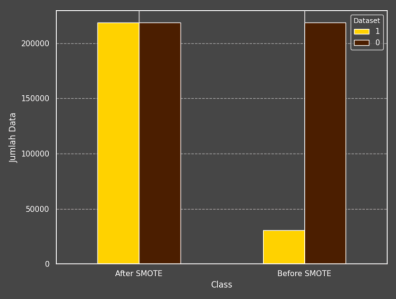
2. Target Encoding pada feature "profession, "state", dan "city"

Target encoding digunakan pada kolom kategorikal seperti profession, city, dan state untuk mengubah nilai kategori menjadi representasi numerik berdasarkan rata-rata target (risk_flag) pada masing-masing kategori.

Kita tidak menggunakan one-hot encoding, karena bisa menambah dimensi secara signifikan terutama jika jumlah kategori banyak, dengan menggunakan target encoding tetap efisien karena hanya menghasilkan satu kolom numerik per fitur.

Handling Imbalance Data

SMOTE(Oversampling)



Before SMOTE	After SMOTE
218490	218490
30568	218490

Penanganan yang paling direkomendasikan adalah menggunakan oversampling. Teknik ini dapat meningkatkan jumlah untuk kelas yang minoritas

Model Selection

	Regresi Logistik	Random Forest	SVM	XGBoost	AdaBoost	Gradient Boosting
Accuracy	0.62	0.92	0.87	0.89	0.88	0.88
Precision	0.17	0.58	0.00	0.63	0.00	0.82
Recall	0.56	0.99	0.00	0.25	0.00	0.01
f1-score	0.27	0.73	0.00	0.36	0.00	0.03
Roc Auc	0.63	0.96	0.49	0.89	0.70	0.78

Model Selection

Kami memilih Model Random Forest karena:

- Setelah kami mencoba melatih data dengan model Regresi logistik, SVM, Random Forest, XGBoost, GradientBoosting. Hasilnya menunjukkan bahwa Random forest memiliki performa model terbaik. dengan Recall: 0.99 dan F1 Score= 0.73 serta ROC AUC: 0.96.
- Recall 0.99 bermakna bahwa 99% data berhasil diklasifikasikan dengan benar oleh model.
- ROC AUC = 0.96 menunjukkan bahwa model mampu membedakan antara kelas positif dan negatif . model sangat baik dalam membedakan kedua kelas secara umum.

Model Tuning dan Optimization

hyperparameter yang digunakan adalah RandomizedSearchCV

Best Parameter :

- max_depth': None
- 'min_samples_leaf': 2
- 'min_samples_split': 3
- 'n_estimators': 93

Hasil Evaluasi menggunakan Hyperparameter

Accuracy	Precision	Recall	F1 Score
0.9627	0.9615	0.9627	0.9616

Model Performance

DataSet	F1 Score
Train	0.9615
Test	09616

Kesimpulan: kedua model sudah cukup baik dan tidak mengalami overfitting/ underfitting. Model Random forest memiliki performance lebih unggul dibandingkan dengan gradient boosting

Model Evaluasi

Melakukan perbandingan performa dengan data validasi

Data Set	Accuracy	Precision	Recall	F1-score
Uji	0.9627	0.9615	0.9627	0.9616
validation	0.8967	0.9008	0.8967	0.8586

- Model memiliki keseimbangan yang baik antara menggunakan data validasi dan data uji, menunjukkan generalisasi yang baik
- Meskipun Matriks memiliki penurunan, namun hasil tersebut masih cukup baik
- Pada Kasus ini model Random Forest cukup baik dalam menentukan pengguna yang mungkin benar-benar mengalami Risk Flag

Error Analysis

Model mengalami kesulitan pada kelompok profesi, kota, dan negara bagian tertentu. Misalnya, city_enc = 0.219178 muncul dalam banyak kesalahan, menunjukkan bahwa kota ini sering salah diklasifikasikan

Distribusi Fitur dalam Kesalahan Prediksi:

profession_enc	city_enc	state_enc	
0.134078	0.219178	0.128553	8
0.127400	0.219178	0.128553	7
0.148571	0.161290	0.131234	7
0.152000	0.140000	0.128553	7
	0.171429	0.142166	6
0.148571	0.136986	0.121248	6
0.123239	0.144144	0.134809	6
0.115873	0.089744	0.133956	6
0.132184	0.174603	0.131234	6
0.152878	0.216216	0.117050	6

Error Analysis

Berdasarkan Ground Truth 99.01% dari total kesalahan prediksi terdapat pada kelas 1, dimana terdapat 10 contoh kasus salah prediksi pada kelas tersebut

Contoh Kasus Salah Prediksi:

	Actual	Predicted	Error
2	1	0.0	1
9	1	0.0	1
18	1	0.0	1
26	1	0.0	1
32	1	0.0	1
33	1	0.0	1
43	1	0.0	1
44	1	0.0	1
46	1	0.0	1
56	1	0.0	1

Webs Apps

Teknologi yang digunakan

- **Back End : Fast API**
- **Front End : Streamlit**
- **DataBase : MySql**

Feature yang tersedia

- **Register**
- **Login**
- **Predict**
- **Logs**
- **Financial Recommendation**

Link Streamlit :
https://predict-riskflag.streamlit.app/

Business Impact

- **Menurunkan Risiko Kredit Macet (NPL):**
  Dengan memahami pola risiko berdasarkan fitur pelanggan, perusahaan dapat meminimalkan potensi gagal bayar.
- **Meningkatkan Efisiensi Proses Kredit:**
  Pengambilan keputusan kredit menjadi lebih cepat dan akurat melalui analisis data historis dan segmentasi risiko.
- **Personalisasi Penawaran Produk Keuangan:**
  Pemanfaatan generative AI memungkinkan pembuatan produk yang lebih relevan berdasarkan karakteristik nasabah.

Personalisasi Produk Keuangan

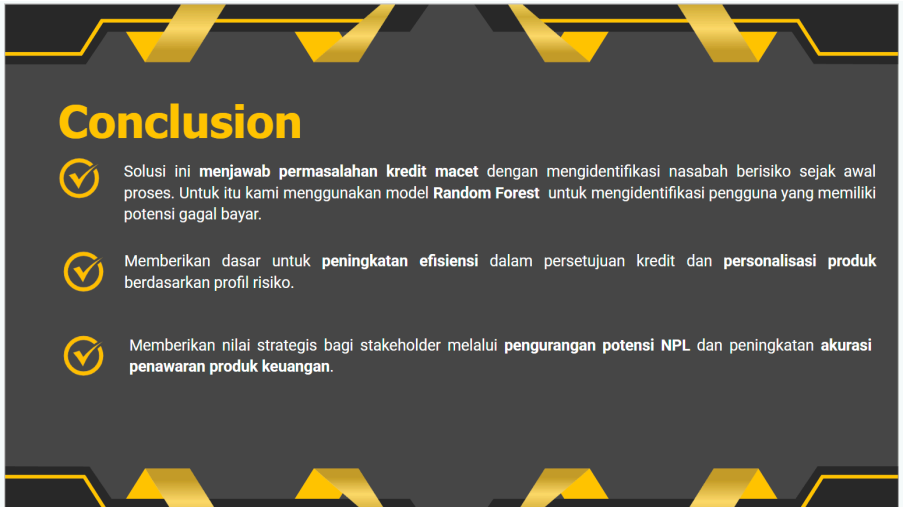
Menurunkan Risiko Kredit Macet

Meningkatkan efisiensi proses kredit

Business Recommendation

- **Implementasi Model Risiko dalam Proses Kredit Awal:**
  Gunakan model prediksi *Risk_Flag* sebagai filter awal dalam proses approval kredit untuk menghindari calon nasabah berisiko tinggi.
- **Integrasi Generative AI untuk Pengembangan Produk:**
  Manfaatkan AI untuk menciptakan varian produk keuangan yang menyesuaikan kebutuhan dan profil risiko nasabah secara otomatis.
- **Peningkatan Awareness Tim Kredit:**
  Berikan pelatihan pada tim analis kredit agar mampu menginterpretasi output model secara akurat sebagai bagian dari pengambilan keputusan.


Project Challenges faced during the project
Tantangan-tantangan yang kami hadapi selama baik yang teknis dan non-teknis pada project ini adalah
Meski menggunakan google colab yang sama namun percobaan setiap anggota berbeda
Distribusi data yang terlihat buatan
Perlu berhati-hati pada inconsisten data pada feature CITY dan STATE


Conclusion
Solusi ini menjawab permasalahan kredit macet dengan mengidentifikasi nasabah berisiko sejak awal proses. Untuk itu kami menggunakan model Random Forest untuk mengidentifikasi pengguna yang memiliki potensi gagal bayar.
Memberikan dasar untuk peningkatan efisiensi dalam persetujuan kredit dan personalisasi produk berdasarkan profil risiko.
Memberikan nilai strategis bagi stakeholder melalui pengurangan potensi NPL dan peningkatan akurasi penawaran produk keuangan.

Bab VI Lampiran B. Notulensi Mentoring

Pada Bagian ini berisi log activity atau notulensi mentoring dengan format sebagai berikut

Minggu/Tgl	Kegiatan	Hasil
1 Febuary 2025	Membahas mengenai indentifikasi industri, matriks bisnis serta masalah-masalah yaang dapat diselesaikan oleh model *machine learning*	Perlu ditambahkan detail hasil perbandingan ketika sebelum menggunakan hasil model dengan sesudah menggunakannya terutama untuk Industry Problem dan penentuan metriks bisnis, untuk poin "menyesuaikan penawaran produk keuangan" perlu mengunakan Gen AI
15 February 2025	Membahas mengenai penyediaan formatting data, penggunaan median untuk replace missing values, class imbalance data untuk risk_flag dan segmentasi	Merapihkan keseragaman dari kolom data, melakukan over sampling untuk handling imbalance data dan melakukan feature engineering
8 Maret 2025	Membahas mengenai inovasi yang ingin di gunakan pada project ini, matriks evaluasi untuk data imbalance, penggunaan Streamlit untuk front end	Revisi laporan PPT, menambahkan otentikasi untuk Fast API

22 Maret 2025	Membahas mengenai impact bisnis dari model yang telah dibuat, Analysis Error, Visualisasi	Api, streamlit sudah dibuat, tidak ada perubahan pada feature, hanya ganti ML modelnya jadi Random forest, bisnis insight nya dipaparkan

Bab VII Lampiran C. Dokumen Teknik

- Link Github : https://github.com/danieltoby7/NEVORYA
- Sumber referensi:

(1) McKinsey & Company. (2020). *AI in banking: Can banks meet the challenge?* Retrieved from https://www.mckinsey.com/. Mengungkap bahwa adopsi AI/ML di sektor keuangan bisa memangkas biaya operasional hingga **20–25%**.

(2) PricewaterhouseCoopers. (2019). *AI in financial services*. Retrieved from https://www.pwc.com/. Menyatakan bahwa penggunaan AI & ML bisa mengurangi waktu pemrosesan kredit hingga **90%** dan meningkatkan akurasi penilaian risiko.