Final Project

BYTE ME

– Rakamin Academy

HOME CREDIT

Hijir Della Wirasti
Mauliddinia Iftikhar Agnany
Jericho Medion Haryono
Fakhri Dwi Nugroho
Ryan Nofandi

Hello BYTE ME

Data Science
Project Manager
Hijir Della Wirasti

Data Science
Mauliddinia Iftikhar
Agnany

Data Science
Fakhri Dwi Nugroho

Data Science
Ryan Nofandi

Data Science
Jericho Medion Haryono

Background

01	Background	This section introduces the context, goals, and objectives of the analysis related to HomeCredit.
02	EDA & Insight	Exploratory Data Analysis to identify trends, patterns, and insights within the HomeCredit dataset.
03	Pre-Processing	Data cleaning, transformation, and preparation steps to ensure the dataset is ready for modeling.
04	Technical Aspect	Implementation and evaluation of machine learning models to achieve the analysis objectives for HomeCredit.
05	Recommendation	Suggestions and actionable steps based on the insights and findings of the analysis.
06	Appendix	Supplementary material at the end of a document or book, often containing additional information, data, or references

BYTE ME

HOME CREDIT

01 Background

Home Credit adalah lembaga keuangan non-bank internasional yang menyediakan layanan pinjaman cicilan, pinjaman tunai, dan solusi pembayaran, melayani pelanggan dengan riwayat kredit yang sedikit atau bahkan tanpa riwayat kredit di 9 negara.

Problem Statement

Tantangan Utama

- Risiko gagal bayar **20%** (Smith, 2020)
- Dampak:
    - **Kerugian finansial**
    - **Gangguan arus kas**
    - **Reputasi buruk**

Solusi

- Membangun model prediksi risiko default menggunakan algoritma seperti **Random Forests** dan **XGBoost**.
- Menargetkan **AUC-ROC ≥ 0.85**.

Business Goals

1. **Mengurangi jumlah nasabah gagal bayar.**
2. **Model tidak terlalu banyak menolak nasabah yang sebenarnya layak**

HOME CREDIT

Objectives

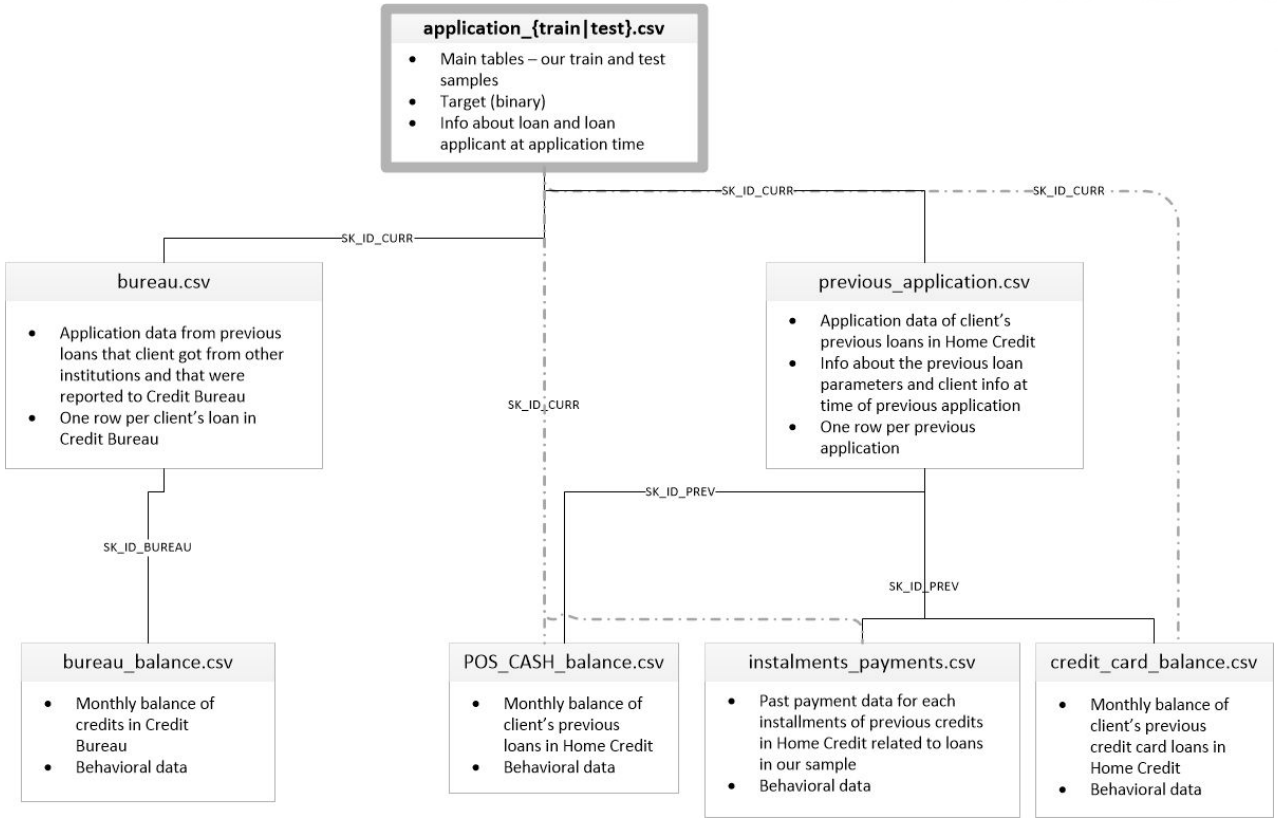
1. **Identifikasi variabel utama** (e.g., riwayat kredit, aplikasi sebelumnya).
2. **Membangun** machine learning
3. **Evaluasi hasil prediksi** dari sudut pandang bisnis.
4. **mitigasi risiko kredit** berbasis data.

Business Metrics

- **Default Prediction**: Mengukur kinerja model dalam memprediksi risiko gagal bayar.
- mengklasifikasikan nasabah berisiko tinggi atau rendah

BYTE ME

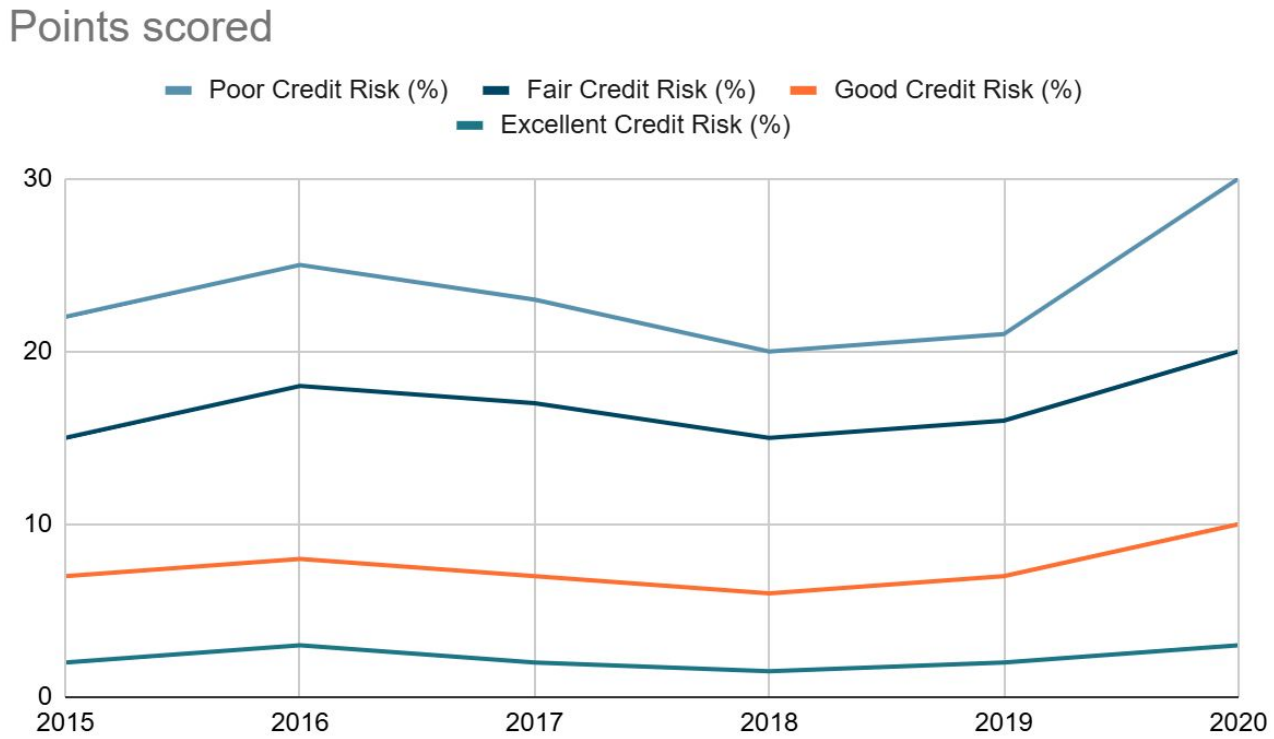
Dataset



Application Train digunakan untuk pemodelan dan data lainnya untuk EDA

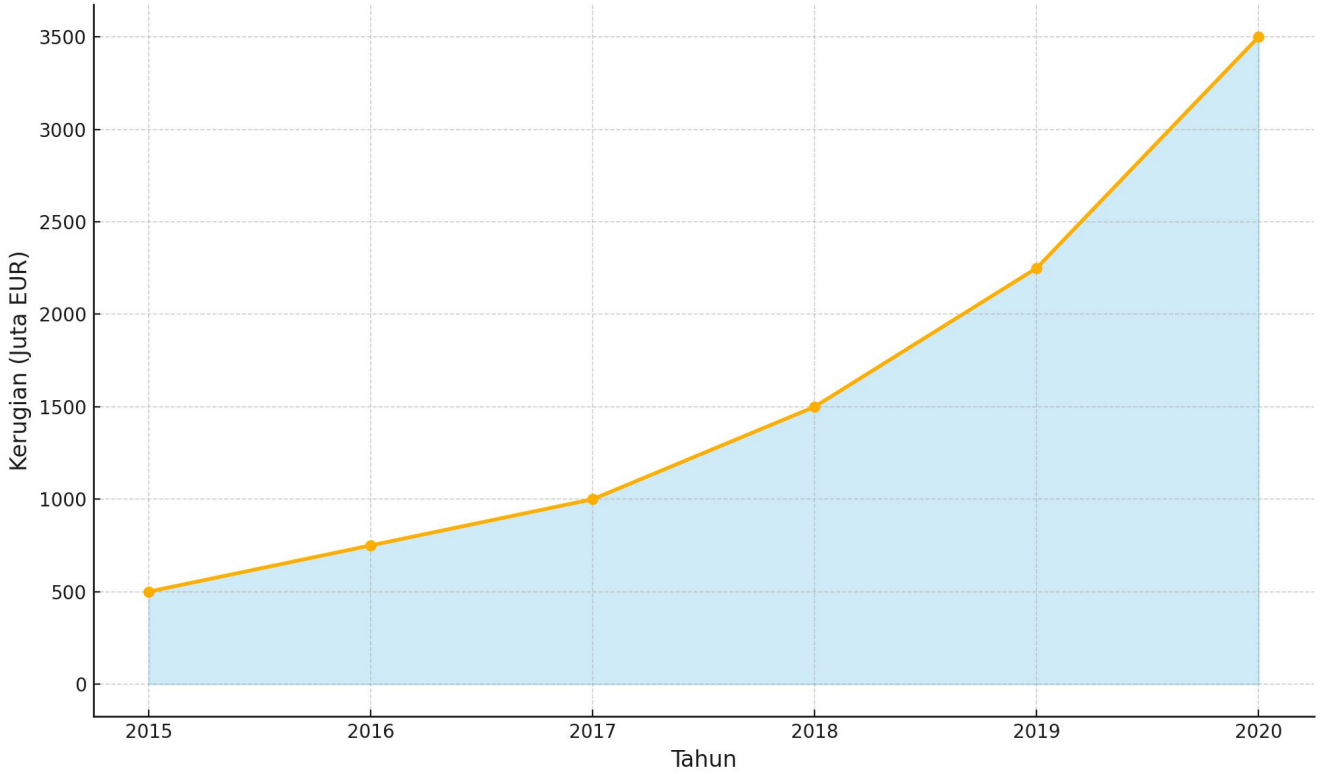
BYTE ME

Resiko Gagal Bayar Meningkat



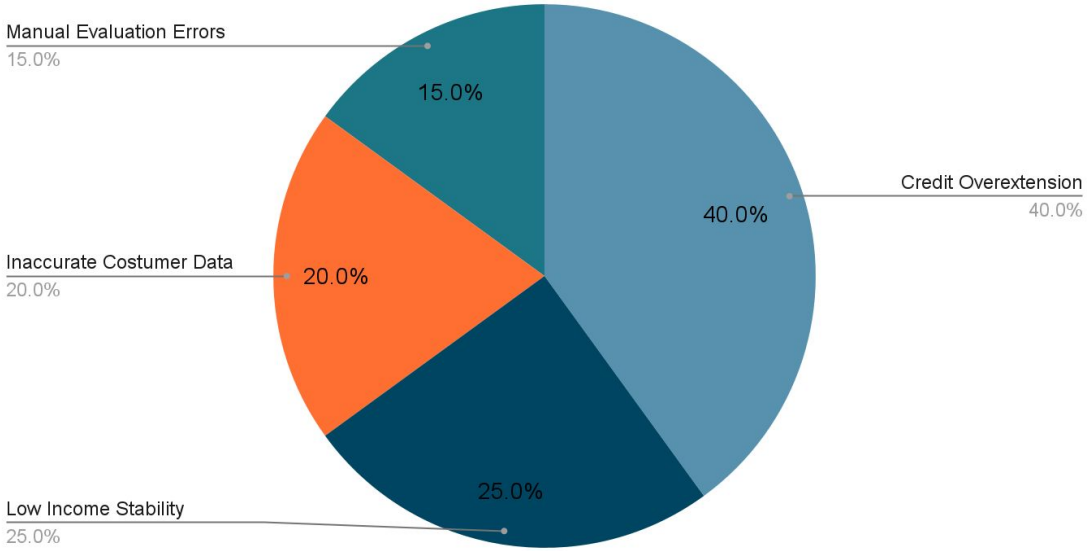
Reference: https://www.spglobal.com/ratings/en/research/articles/210407-default-transition-and-recovery-2020-annual-global-corporate-default-and-rating-transition-study-11900573

Kerugian Home Credit Akibat Debitur yang Gagal Bayar



Reference: https://www.homegroup.org.uk/about-us/corporate/transparency/financial-reports-and-annual-review

Customers gagal bayar karena Penggunaan credit analyst kurang optimal dalam memberikan credit ke nasabah*



Reference: https://ejournal2.undip.ac.id/index.php/jedsr/article/view/12530

7 Paper menunjukan Machine Learning lebih unggul dibanding model lain

LR 5%, KNN 4%, RF 3%, DT 3%, XGB 2% (Accuracy) after HT

Ismunandar, D., Firdaus, M. R., & Alkhalifi, Y. (2024)

LR 0,6, RF 0,7, GNB 0,6, DT 0,5, MLP 0,5 (ROC AUC test score) - Home Credit

Afifudin, M., & Rizki, A. (2023)

XGBoost (accuracy 82%, recall 70%, precision 92%

Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022)

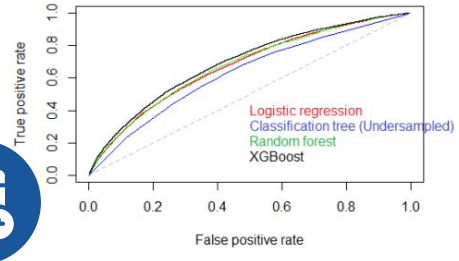
Default Prediction ML

XGB (Recall of 0.35, AUC of 0.832)

Zhou, Y. (2023)

Oversampling technique (SMOTE), SVM SMOTE, random undersampling, and ALL-KNN (accuracy of 98.6%)

Mahbobi, M., Kimiagari, S., & Vasudeva, M. (2023)



XGBoost's AUC of 0.695

Himberg, T. (2021)

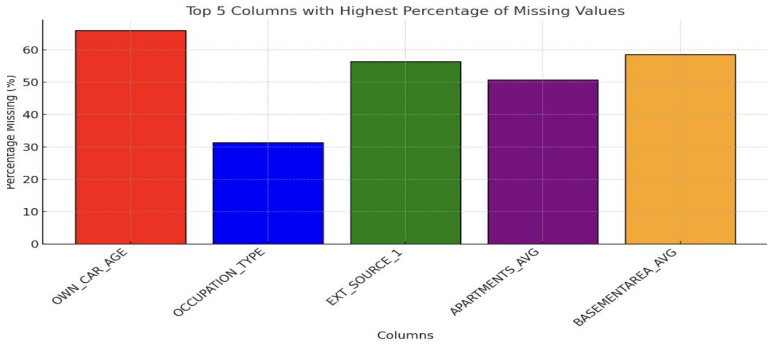
	Accuracy		F1 Score		AUC score	
	ADNI	Credit	ADNI	Credit	ADNI	Credit
SVM	0.73	0.62	0.73	0.62	0.89	0.65
XGBoost	0.72	0.68	0.72	0.67	0.88	0.74
DL	0.70	0.60	0.70	0.64	0.88	0.64

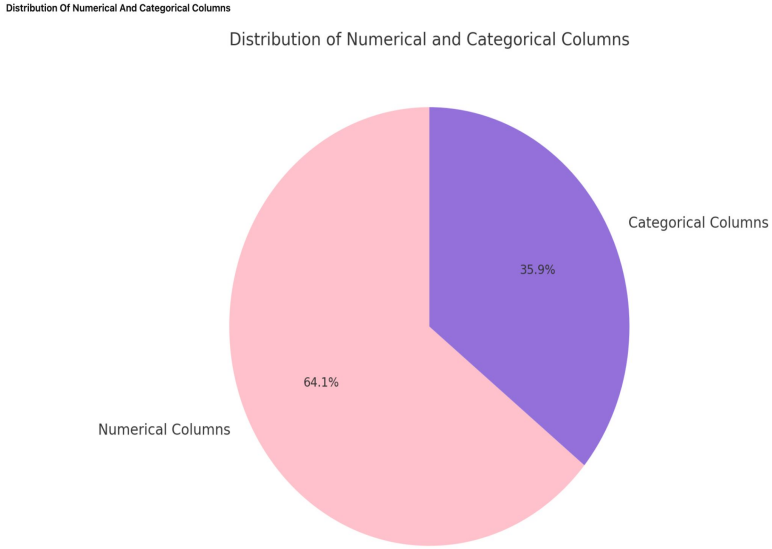
Bi, Z., Gao, R., & Fang, S. (2024)

EDA & Insight 02

Exploratory Data Analysis to identify trends, patterns, and insights within the HomeCredit dataset.

A. Data Info





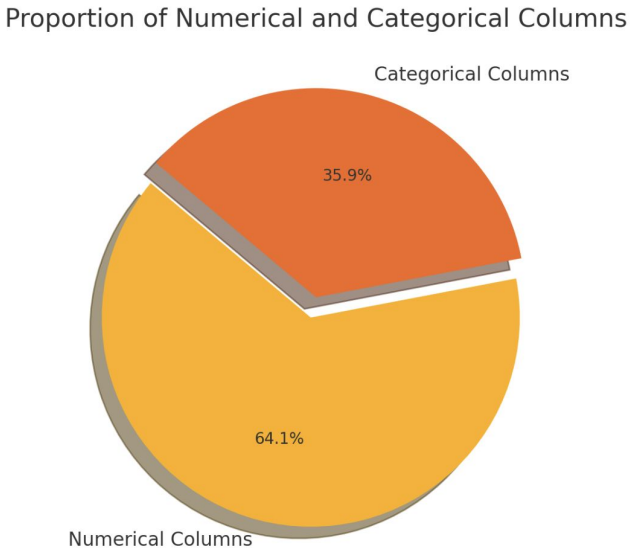
Dataset Overview

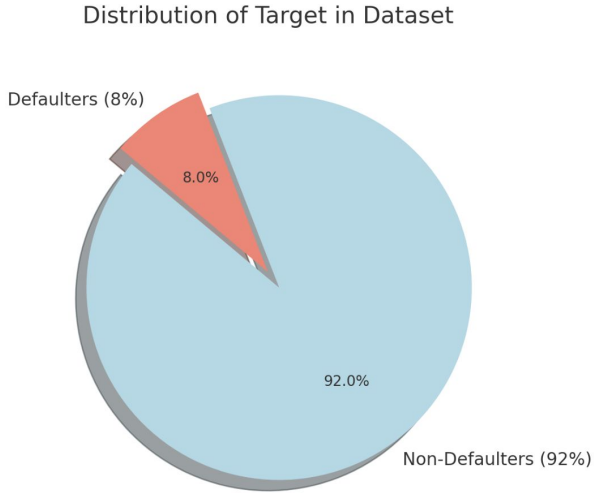
- **Dataset:** Home Credit Default Risk
- **Total Records:** 307,511
- **Total Features:** 122
  - **Numerical Columns:** 25
  - **Categorical Columns:** 14
  - **Columns with Missing Values:** 10

Missing Values Handling:

- Columns with **<5% missing:** Imputed using median.
- Columns with **5–20% missing:** Imputed using median values.
- Columns with **>20% missing:** Dropped (e.g., `OWN_CAR_AGE`, `EXT_SOURCE_1`).

B. Data Insights



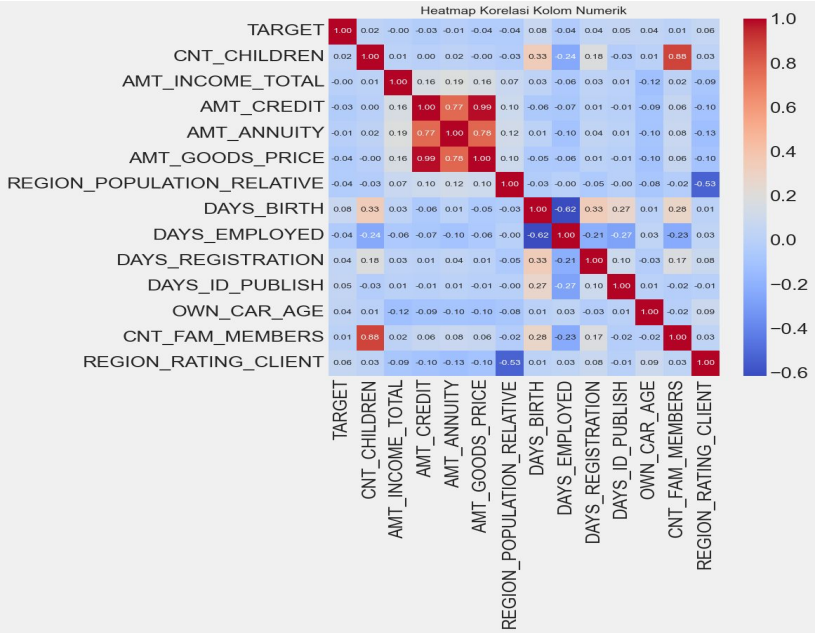


Key Observations

1. **Class Imbalance:**
   - **Class 0 (Non-Defaulters):** 92% (282,686 records)
   - **Class 1 (Defaulters):** 8% (24,825 records)
   - **Action:** Resampling needed to address imbalance.
2. **Anomalies Identified:**
   - `DAYS_EMPLOYED`: Max value of 365,243 days (over 1,000 years).
   - `OWN_CAR_AGE`: Max value of 91 years.
   - **Action:** Replace anomalies with meaningful values or remove.
3. **Missing Values Handling:**
   - Columns with **<5% missing:** Imputed using median.
   - Columns with **5–20% missing:** Imputed using median values.
   - Columns with **>20% missing:** Dropped (e.g., `OWN_CAR_AGE`, `EXT_SOURCE_1`).

•

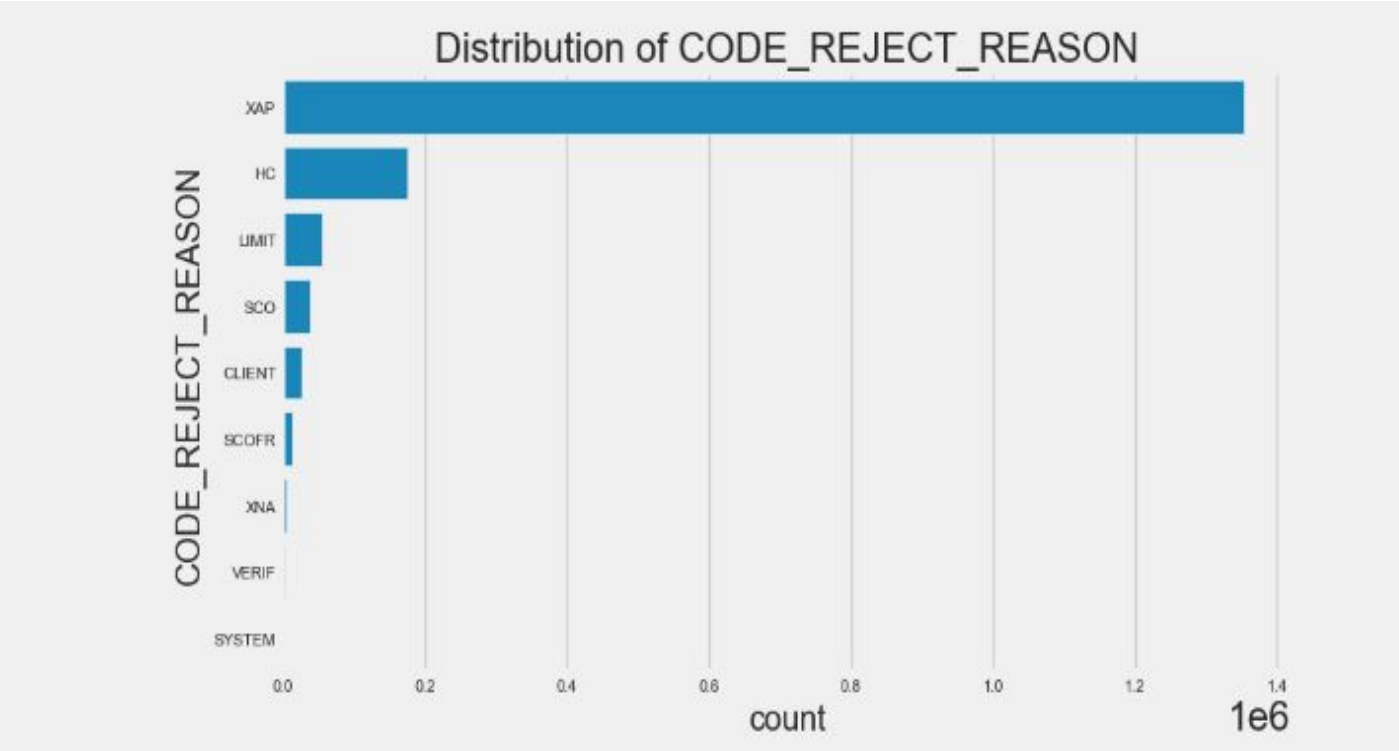
C. Correlation Insights



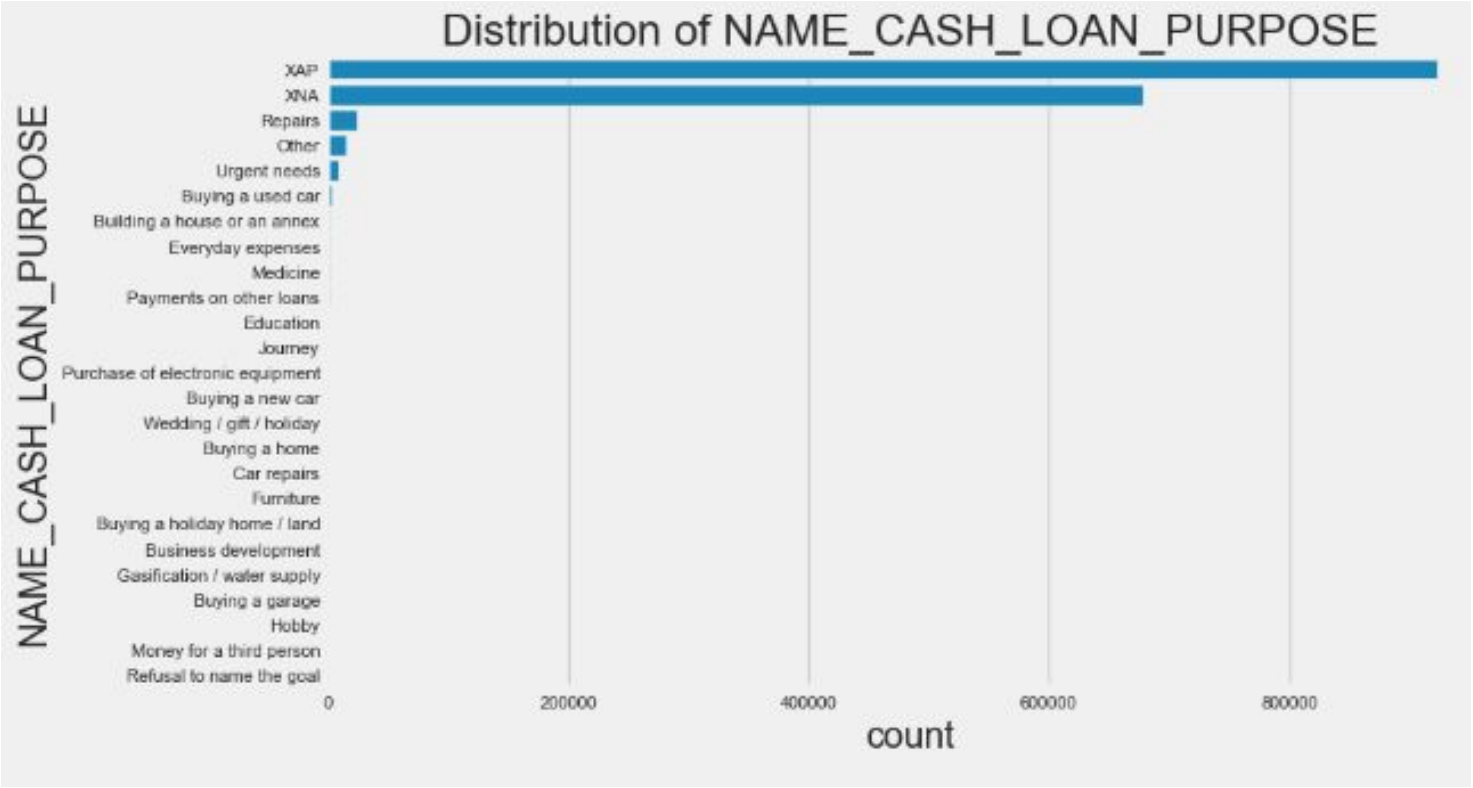
Correlation Insights

- **Negative Correlation with `TARGET`:**
  - `DAYS_BIRTH`: Younger clients have higher default risk.
  - `EXT_SOURCE_2`: Lower scores indicate higher default risk.
- **Positive Correlation with `TARGET`:**
  - `REGION_RATING_CLIENT`: Clients in higher-rated regions tend to have lower risks.

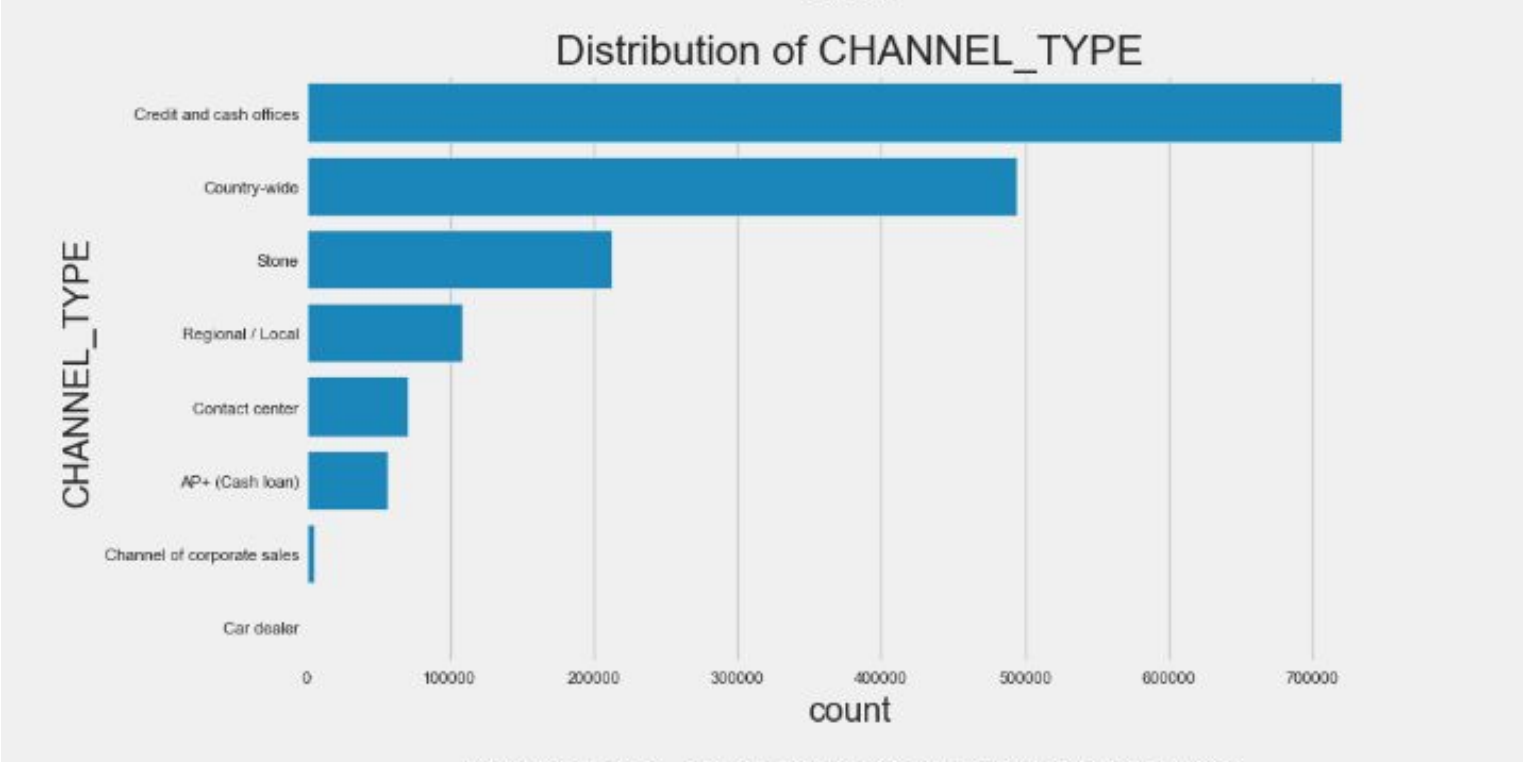
D. Distribution Insights



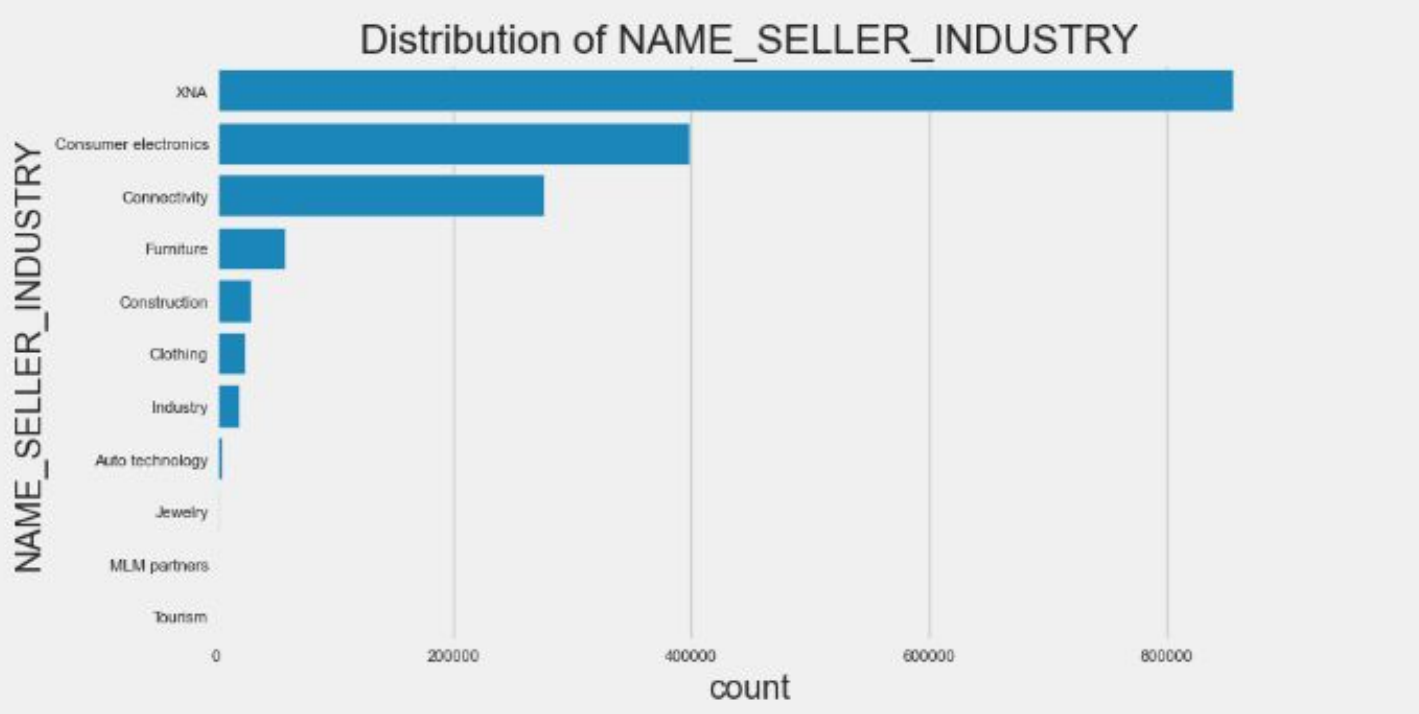
D. Distribution Insights



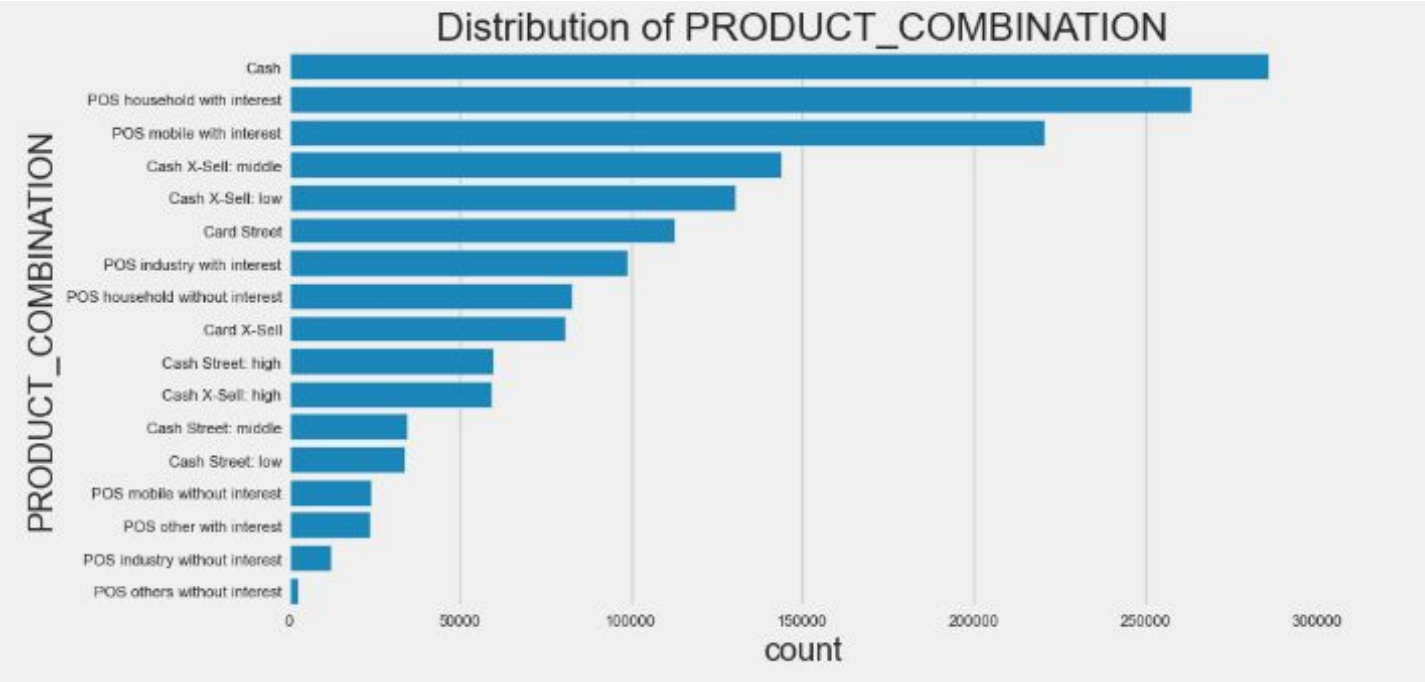
D. Distribution Insights



D. Distribution Insights

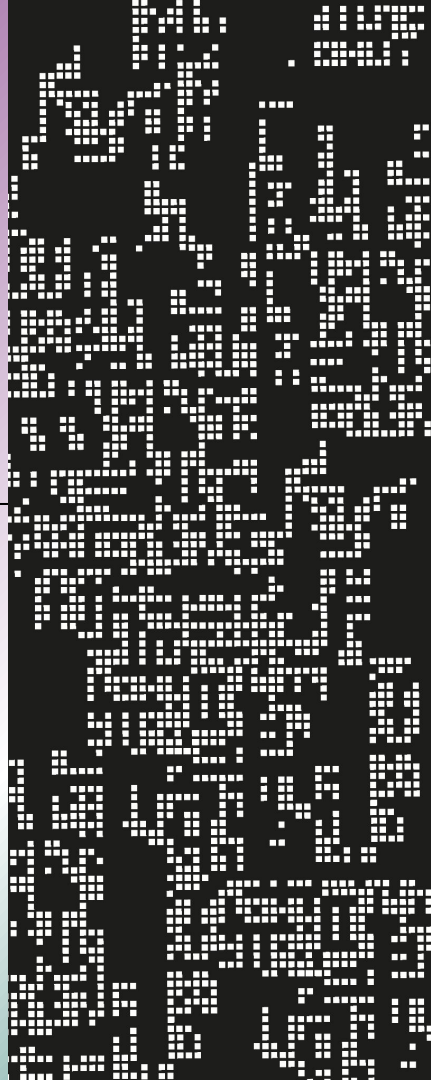


D. Distribution Insights

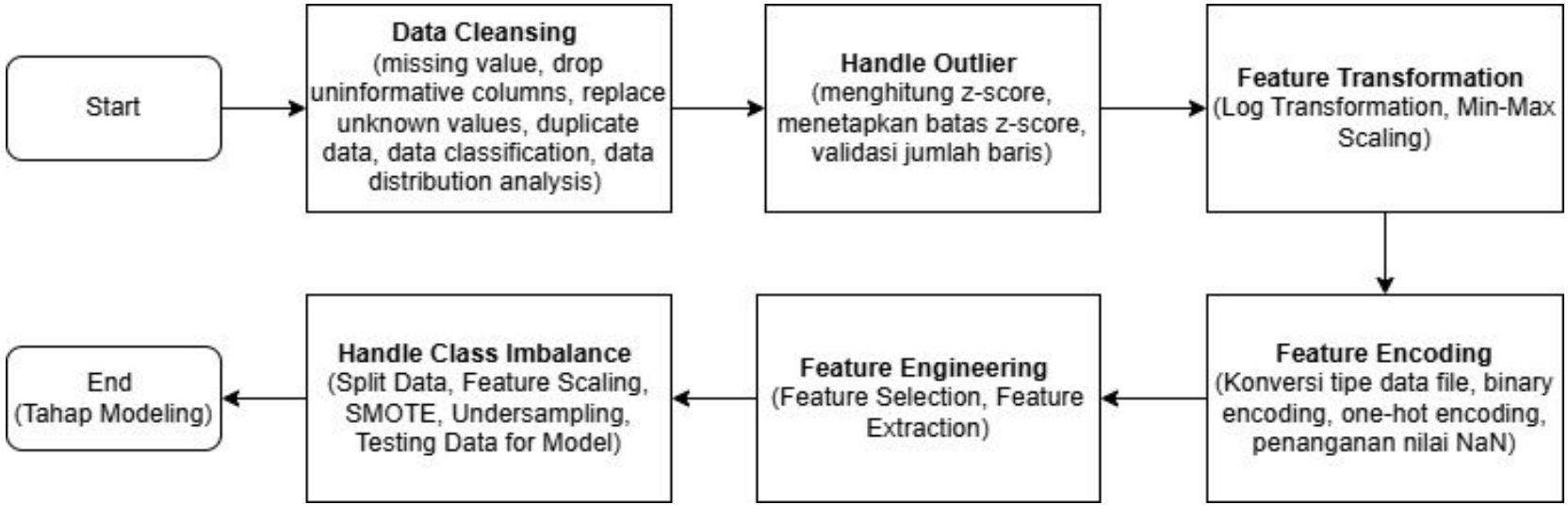


03 Pre-Processing

Data cleaning, transformation, and preparation steps to ensure the dataset is ready for modeling.

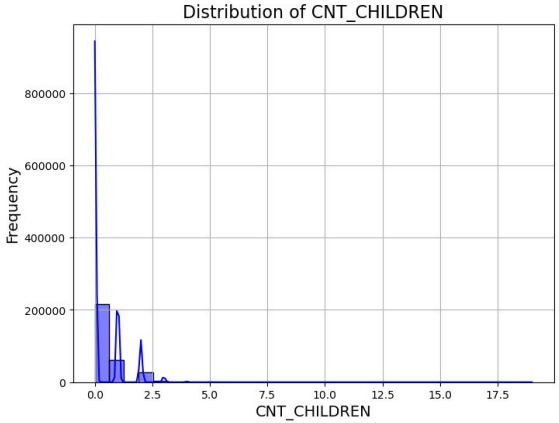


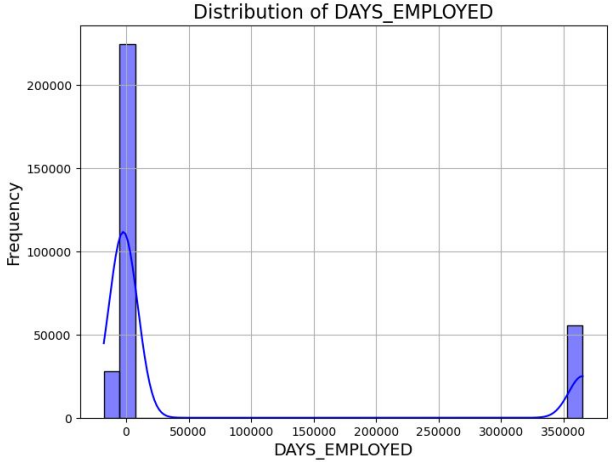
Pre-processing Flow Chart

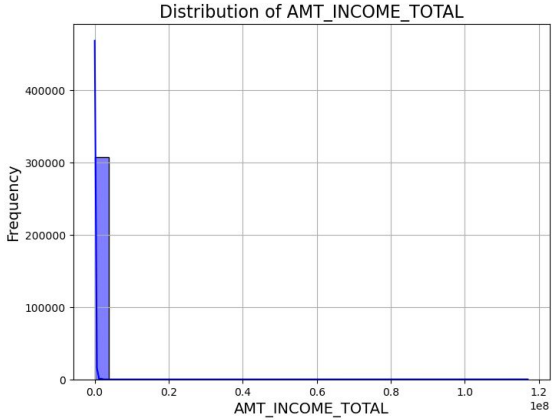


Handle Outlier

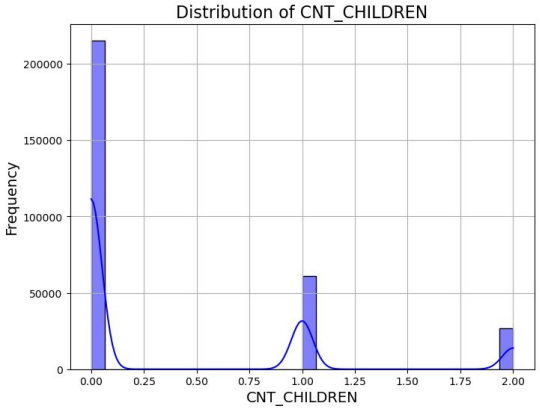
Before

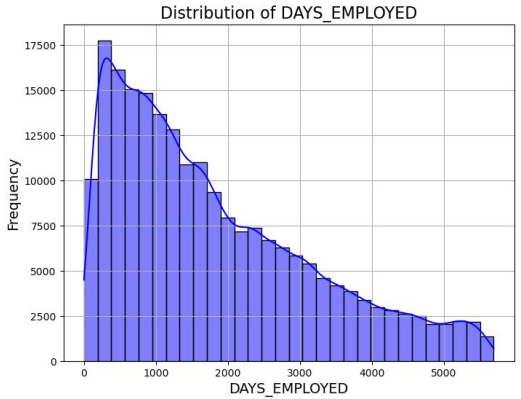


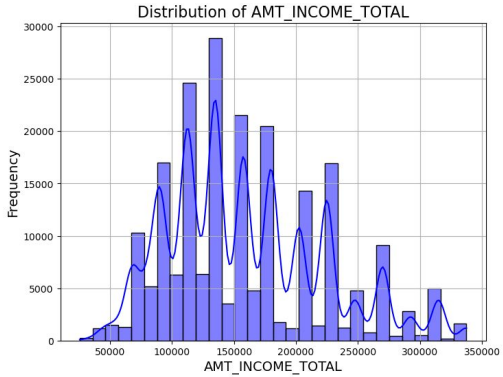




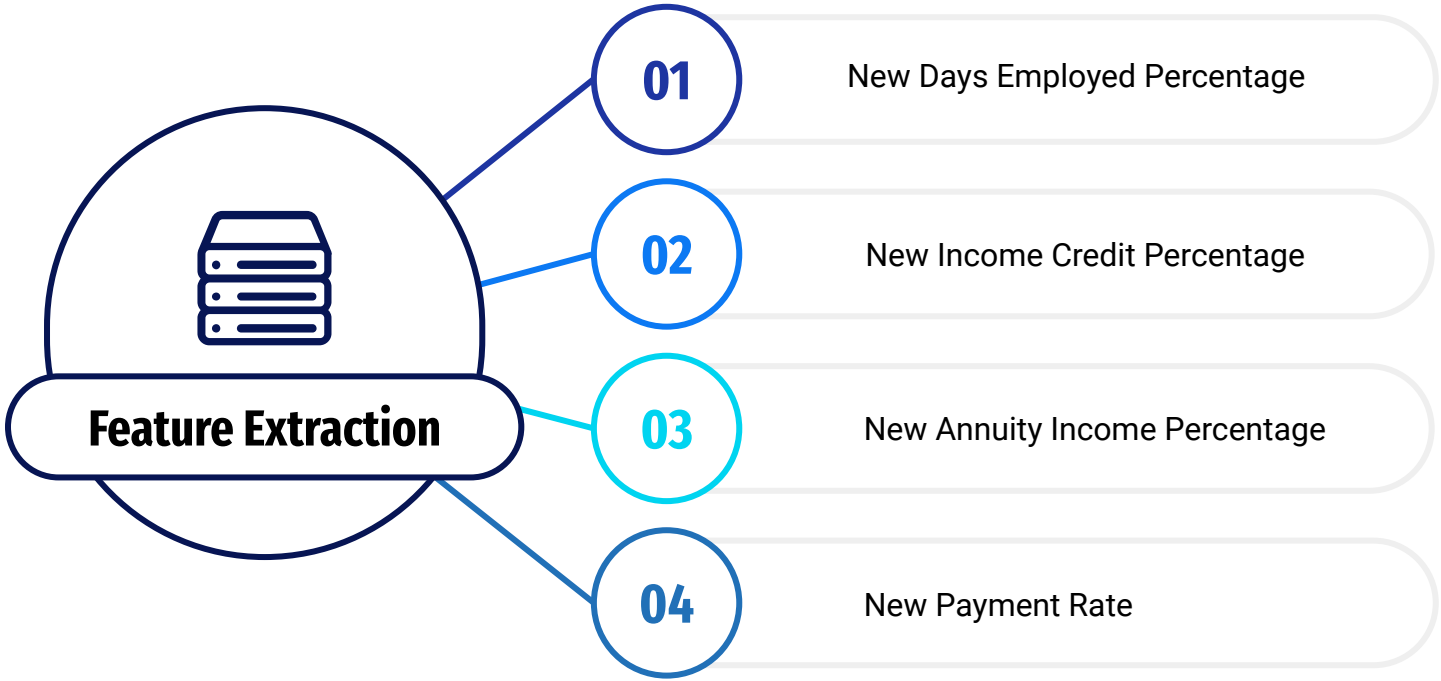
After

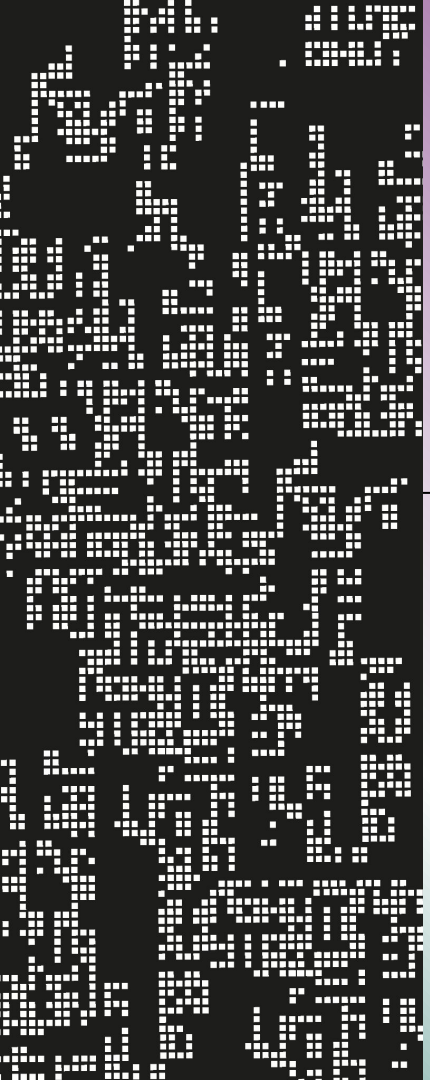






New Feature


Feature Extraction
01
New Days Employed Percentage
02
New Income Credit Percentage
03
New Annuity Income Percentage
04
New Payment Rate



Technical Aspect 04

Implementation and evaluation of machine learning models to achieve the analysis objectives for HomeCredit.

Kami Menggunakan Model

Adaboost

AdaBoost menggabungkan **weak classifiers** sederhana menjadi **strong classifier** secara adaptif, **menghasilkan akurasi tinggi** dalam klasifikasi tanpa memerlukan model kompleks.

Lin, J. (2024)

Decision Tree

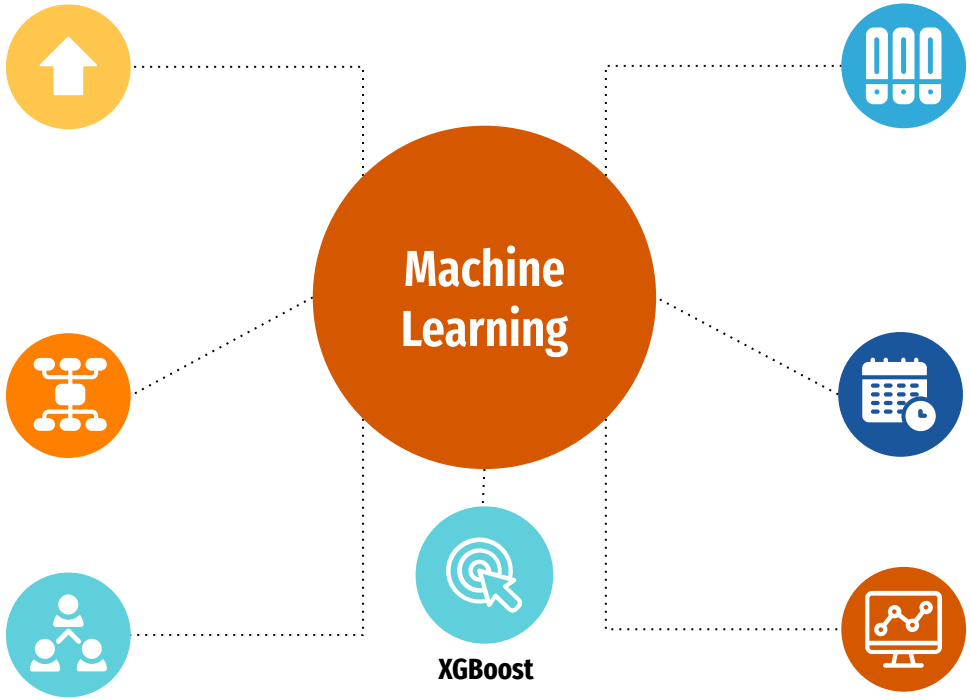
Algoritma fleksibel untuk **klasifikasi dan regresi**, yang populer karena mudah dipahami, mampu menangani data **numerik** dan **kategorikal**, serta menghasilkan **struktur pohon** yang intuitif untuk interpretasi hasil.

Madaan, Mehul, et al. (2021)

KNN

Algoritma "**lazy learning**" yang sederhana namun efektif, tidak memerlukan **model awal**, dan cocok untuk klasifikasi berdasarkan kedekatan data. Keunggulannya meliputi **fleksibilitas** untuk berbagai **dataset**, mudah dipahami, dan mampu **menangkap pola lokal** berdasarkan **tetangga terdekat**.

Nagajyothi, V. (2020)



XGBoost

XGBoost unggul karena **kecepatan tinggi**, efisiensi memproses **data sparse**, dan kemampuan **mengontrol overfitting** melalui regularisasi.

Omogbhemhe, M. I. (2021)

Random Forest

Algoritma supervised learning yang membangun **ensemble prediktor** dari **banyak decision tree** yang dilatih pada **subruang data** secara acak, cocok untuk **klasifikasi** dan **regresi**.

Madaan, Mehul, et al. (2021)

Stacking

Menggabungkan kekuatan berbagai model prediksi **(heterogeneous models)** melalui **dua lapisan—base mode**l untuk menghasilkan **prediksi awal** dan **meta-model** untuk **meningkatkan akurasi prediksi** dengan mengolah hasil tersebut

Zeng, L., Sun, J., & Zhou, Y. (2023).

Logistic Regression

Terbukti efektif dalam memprediksi loan defaults, dengan interest rate sebagai prediktor utama yang akurat.

Zhao, S., & Zou, J. (2021)

Pemilihan Metrik

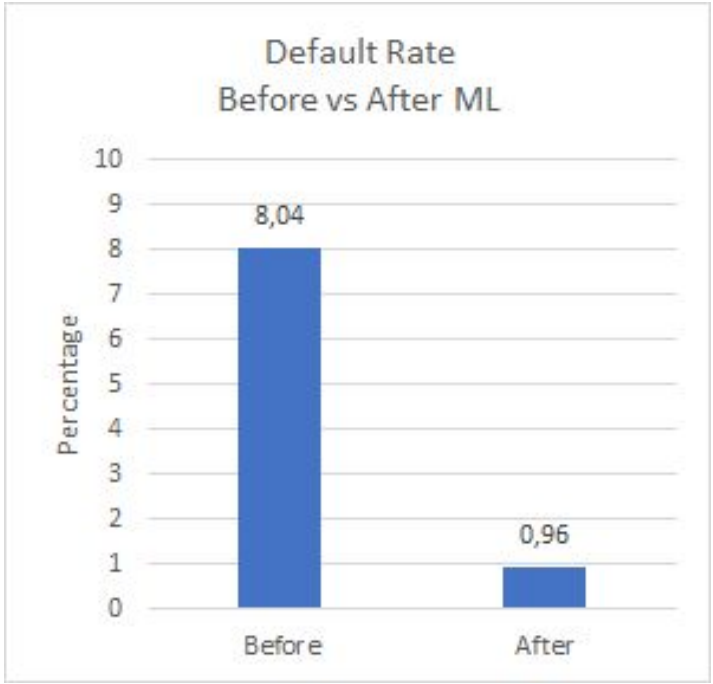
Metrik yang penting:

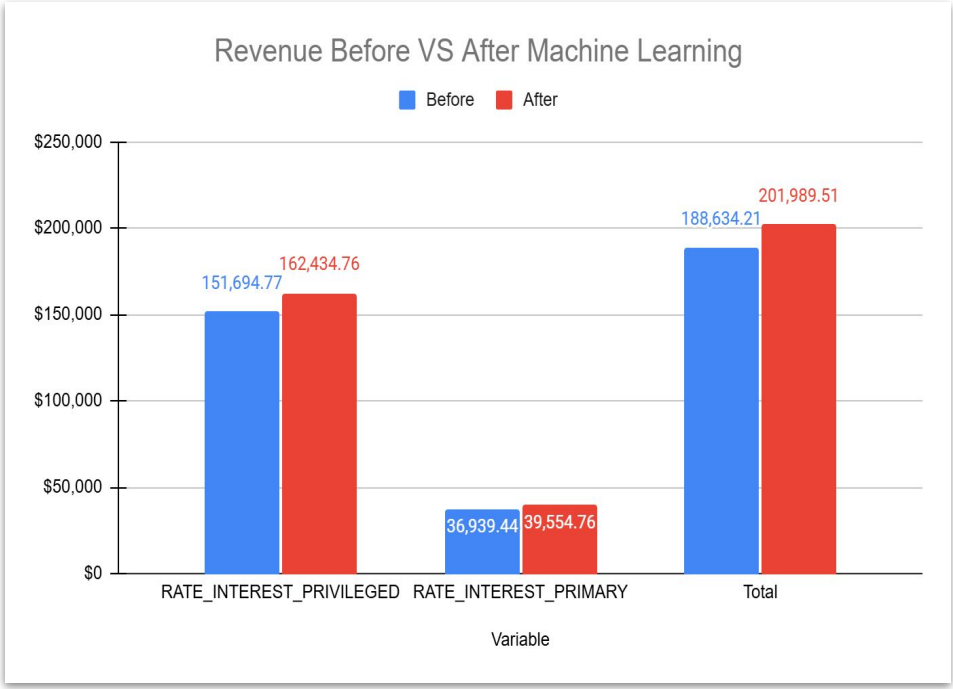
- **Recall**: recall_score, True Positive Rate, confusion_matrix -> **Sensitivity**
- **ROC AUC**: True Positive Rate (TPR), False Positive Rate (FPR), probability threshold -> **Discrimination**
- **F2-Score**: Precision, recall, weighted recall -> **Balance**
- **Accuracy**: True Positives + True Negative & confusion_matrix -> **Correctness**

Table 1. The Best Model of Oversampling (Before & After Hyperparameter)

	Metric	XGBoost Before Tuning	XGBoost After Tuning
1.	ROC AUC (Train Set)	0.9646	0.9861
2.	ROC AUC (Test Set)	0.9469	0.9470
3.	Accuracy (Train Set)	0.9406	0.9460
4.	Accuracy (Test Set)	0.9387	0.9397
5.	F2-Score (Test Set)	0.8483	0.8519
6.	Recall (Cross-Validation Train)	0.8197	0.8239
7.	Recall (Test Set)	0.8176	0.8220
8.	Recall (Cross-Validation Test)	0.8165	0.8208

Dampak Penerapan Machine Learning

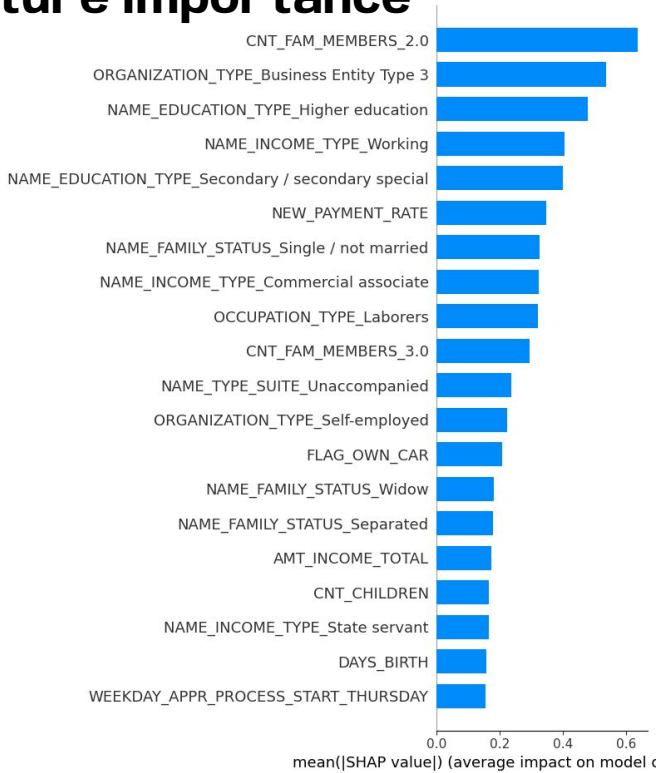


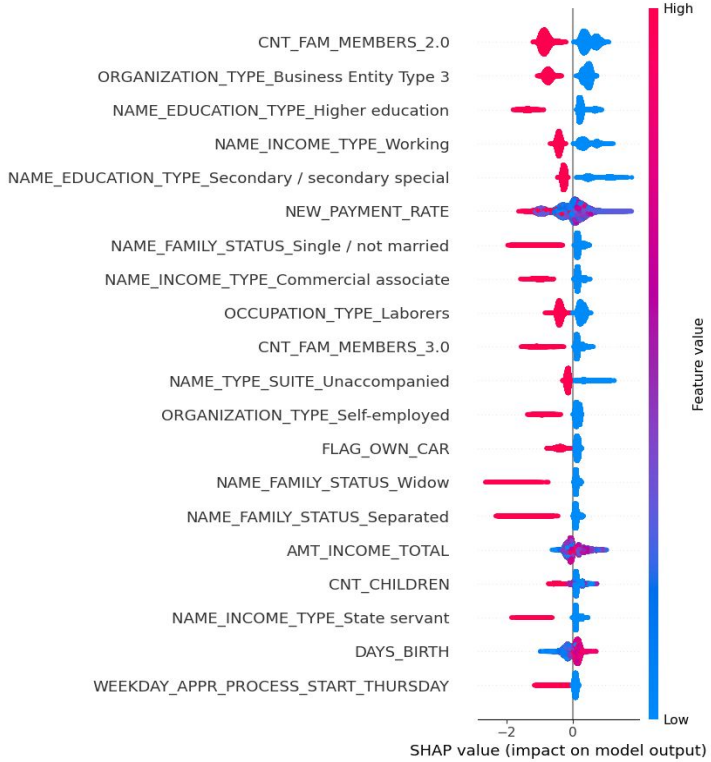


BYTE ME

Reference: https://docs.google.com/spreadsheets/d/14H5pG8FOO4Oz6hZFL-zXD-QuKXPERRlUguRunnBAdTQ/edit?usp=sharing

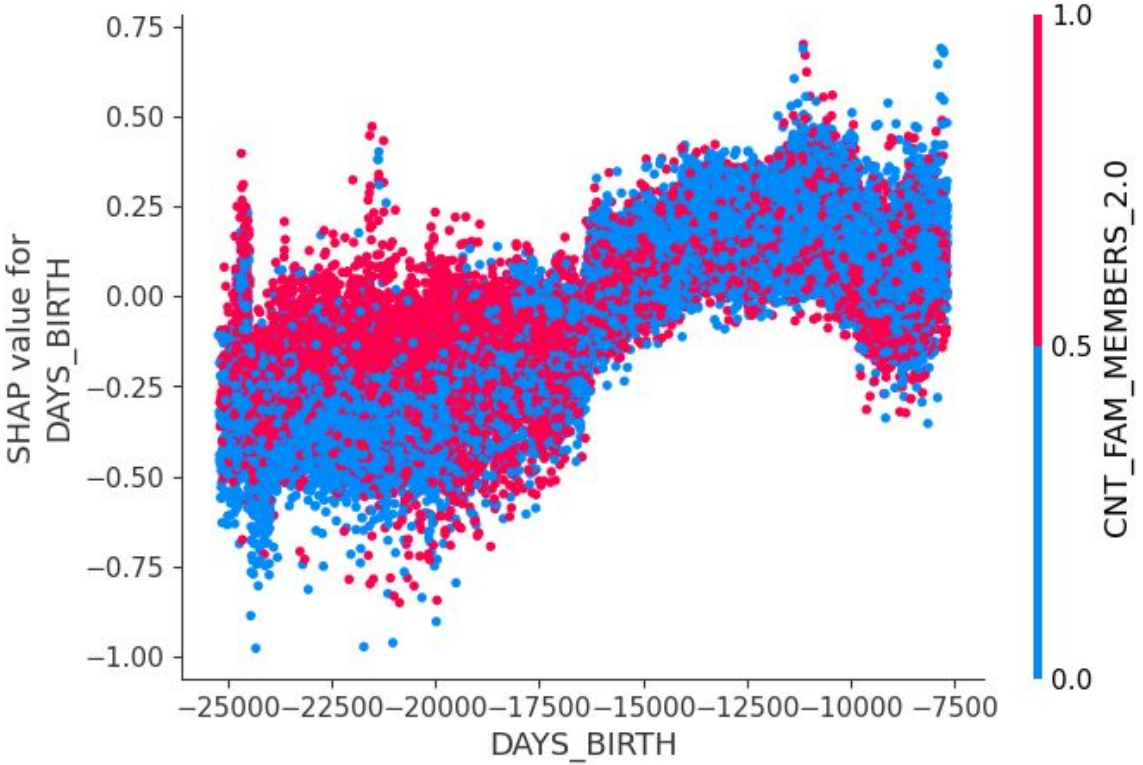
Feature Importance





Feature yang mempengaruhi gagal bayar (default rate) tertinggi dipengaruhi oleh faktor demografis seperti (jumlah anggota keluarga, jenis pekerjaan dan status pendidikan).

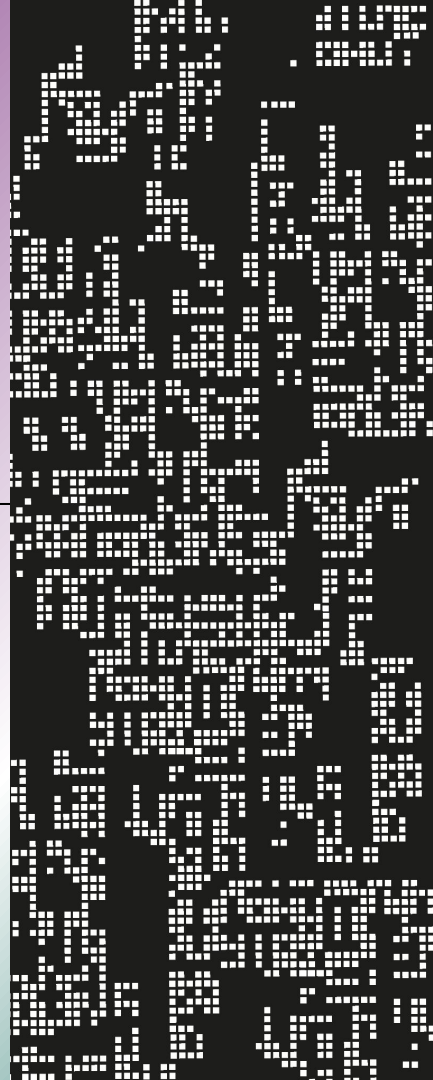
Feature Importance



Berdasarkan grafik tersebut didapatkan bahwa usia yang lebih tua dan dengan 2 anggota keluarga memiliki resiko gagal bayar yang lebih rendah.

05 Recommendation

Suggestions and actionable steps based on the insights and findings of the analysis.



Business Recommendations

Recomm.	Insights	Actionable Items
Focus on Customer Segmentation and Retargeting	Pengajuan kredit banyak berasal dari sektor Consumer Electronics, Connectivity, dan pembiayaan kebutuhan rumah tangga.	1. Lakukan segmentasi pelanggan berdasarkan kebutuhan kredit dan perilaku pembayaran. 2. Rancang promosi dan kampanye retargeting yang ditargetkan untuk kategori populer seperti elektronik konsumen dan kebutuhan rumah tangga. 3. Gunakan data pengajuan XNA untuk memahami peluang yang belum dimanfaatkan.
Enhance Default Prediction Accuracy	Feature penting dalam memprediksi gagal bayar meliputi faktor demografis seperti jumlah anggota keluarga (CNT_FAM_MEMBERS), jenis pekerjaan (OCCUPATION_TYPE), dan status pendidikan (NAME_EDUCATION_TYPE).	1. Prioritaskan edukasi keuangan untuk nasabah dengan profil berisiko tinggi berdasarkan fitur seperti pendidikan dan pekerjaan. 2. Integrasikan analisis berbasis SHAP ke dalam strategi evaluasi risiko untuk meningkatkan transparansi model. 3. Fokus pada segmentasi yang lebih baik berdasarkan faktor demografis untuk intervensi dini. 3. Menerapkan model XGB kedalam proses penilaian kredit sehinga dapat memprediksi calon nasabah yang beresiko.
Flexible Credit Policies	Nasabah dengan keluarga kecil, pendidikan tinggi, dan rasio pembayaran tinggi memiliki risiko lebih rendah.	1. Tawarkan kredit kompetitif dengan suku bunga lebih rendah untuk nasabah profil ini. 2. Tinjau ulang kebijakan kredit untuk memberikan insentif pada kategori ini untuk meningkatkan loyalitas.

Kesimpulan

Model prediksi berhasil menurunkan default dari **8.04%** menjadi **0.96%** (penurunan kesalahan 7.08%).

Meningkatkan:

- **Revenue meningkat dari** 188,634.21 ke 201,989.51 (107,08%)
- **Efisiensi operasional** dengan pengurangan risiko kredit.
- **Kepuasan nasabah** berkat klasifikasi yang lebih akurat.
- **Keuntungan bisnis** melalui akses kredit yang lebih adil.

Thank You!

Do you have any questions?

Hijir Della Wirasti | LinkedIn

Mauliddinia Iftikhar Agnany | LinkedIn

Fakhri Dwi Nugroho | LinkedIn

Ryan Nofandi | LinkedIn

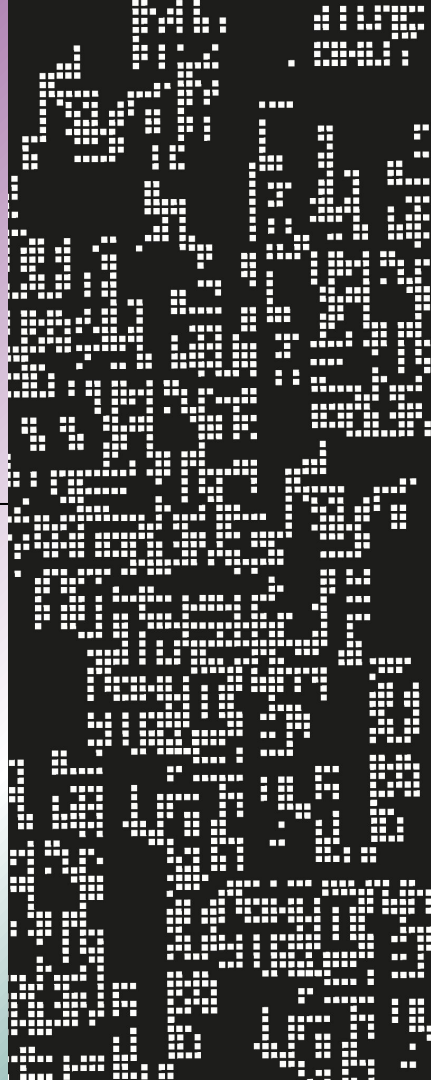
Jericho Medion Haryono | LinkedIn

FINAL PROJECT DRIVE HERE

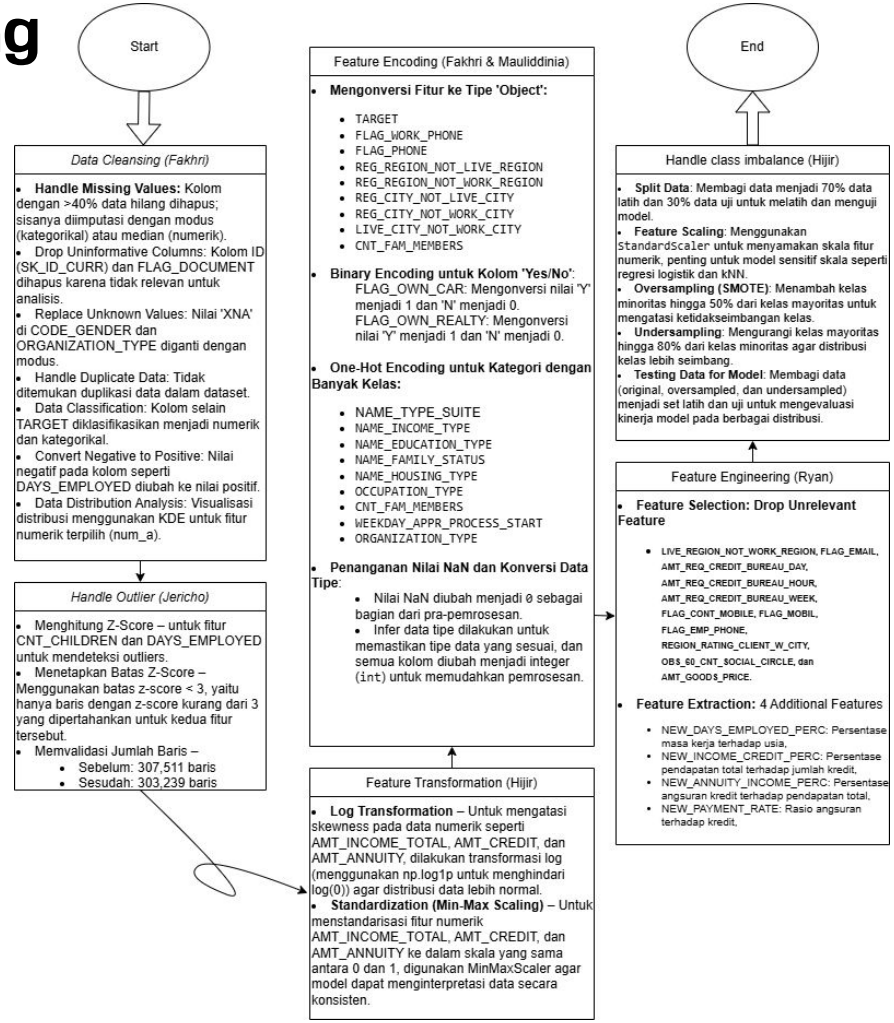
BYTE ME

06 Appendix

Suggestions and actionable steps based on the insights and findings of the analysis.



Pre-Processing Flow Chart



Pre-Processing steps:

1.

Data Cleansing (Fakhri)

- **Handle Missing Values:** Kolom dengan >40% data hilang dihapus; sisanya diimputasi dengan modus (kategorikal) atau median (numerik).
- Drop Uninformative Columns: Kolom ID (SK_ID_CURR) dan FLAG_DOCUMENT dihapus karena tidak relevan untuk analisis.
- Replace Unknown Values: Nilai 'XNA' di CODE_GENDER dan ORGANIZATION_TYPE diganti dengan modus.
- Handle Duplicate Data: Tidak ditemukan duplikasi data dalam dataset.
- Data Classification: Kolom selain TARGET diklasifikasikan menjadi numerik dan kategorikal.
- Convert Negative to Positive: Nilai negatif pada kolom seperti DAYS_EMPLOYED diubah ke nilai positif.
- Data Distribution Analysis: Visualisasi distribusi menggunakan KDE untuk fitur numerik terpilih (num_a).

Output:

1. The number of columns before data cleansing: 122, the number after data cleansing: 52.
2. Visualizations of the data distribution for each column to identify outliers.

2.

Handle Outlier (Jericho)

- Menghitung Z-Score – untuk fitur CNT_CHILDREN dan DAYS_EMPLOYED untuk mendeteksi outliers.
- Menetapkan Batas Z-Score – Menggunakan batas z-score < 3, yaitu hanya baris dengan z-score kurang dari 3 yang dipertahankan untuk kedua fitur tersebut.
- Memvalidasi Jumlah Baris –
  - Sebelum: 307,511 baris
  - Sesudah: 303,239 baris

BYTE ME

Pre-Processing steps (2):

3.

Feature Transformation (Hijir)

- **Log Transformation** – Untuk mengatasi skewness pada data numerik seperti AMT_INCOME_TOTAL, AMT_CREDIT, dan AMT_ANNUITY, dilakukan transformasi log (menggunakan np.log1p untuk menghindari log(0)) agar distribusi data lebih normal.
- **Standardization (Min-Max Scaling)** – Untuk menstandarisasi fitur numerik AMT_INCOME_TOTAL, AMT_CREDIT, dan AMT_ANNUITY ke dalam skala yang sama antara 0 dan 1, digunakan MinMaxScaler agar model dapat menginterpretasi data secara konsisten.

4.

Feature Encoding (Fakhri & Mauliddinia)

- **Mengonversi Fitur ke Tipe 'Object':**
  - TARGET
  - FLAG_WORK_PHONE
  - FLAG_PHONE
  - REG_REGION_NOT_LIVE_REGION
  - REG_REGION_NOT_WORK_REGION
  - REG_CITY_NOT_LIVE_CITY
  - REG_CITY_NOT_WORK_CITY
  - LIVE_CITY_NOT_WORK_CITY
  - CNT_FAM_MEMBERS
- **Binary Encoding untuk Kolom 'Yes/No'**:
  FLAG_OWN_CAR: Mengonversi nilai 'Y' menjadi 1 dan 'N' menjadi 0.
  FLAG_OWN_REALTY: Mengonversi nilai 'Y' menjadi 1 dan 'N' menjadi 0.
- **One-Hot Encoding untuk Kategori dengan Banyak Kelas:**
  - NAME_TYPE_SUITE
  - NAME_INCOME_TYPE
  - NAME_EDUCATION_TYPE
  - NAME_FAMILY_STATUS
  - NAME_HOUSING_TYPE
  - OCCUPATION_TYPE
  - CNT_FAM_MEMBERS
  - WEEKDAY_APPR_PROCESS_START
  - ORGANIZATION_TYPE
- **Penanganan Nilai NaN dan Konversi Data Tipe**:
  - Nilai NaN diubah menjadi 0 sebagai bagian dari pra-pemrosesan.
  - Infer data tipe dilakukan untuk memastikan tipe data yang sesuai, dan semua kolom diubah menjadi integer (`int`) untuk memudahkan pemrosesan.

BYTE ME

Pre-Processing steps (3):

Reasons to drop features:
1. chi-square test where columns have 'P-value >= 0.05 and ['Chi2 %'] < 1.
2. Correlation matrix where columns have a correlation number >0.9.

5.

Feature Engineering (Ryan)

- **Feature Selection: Drop Unrelevant Feature**
  - LIVE_REGION_NOT_WORK_REGION, FLAG_EMAIL, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_WEEK, FLAG_CONT_MOBILE, FLAG_MOBIL, FLAG_EMP_PHONE, REGION_RATING_CLIENT_W_CITY, OBS_60_CNT_SOCIAL_CIRCLE, dan AMT_GOODS_PRICE.
- **Feature Extraction:** 4 Additional Features
  - NEW_DAYS_EMPLOYED_PERC: Persentase masa kerja terhadap usia,
  - NEW_INCOME_CREDIT_PERC: Persentase pendapatan total terhadap jumlah kredit,
  - NEW_ANNUITY_INCOME_PERC: Persentase angsuran kredit terhadap pendapatan total,
  - NEW_PAYMENT_RATE: Rasio angsuran terhadap kredit,

6.

Handle class imbalance (Hijir)

- **Split Data**: Membagi data menjadi 70% data latih dan 30% data uji untuk melatih dan menguji model.
- **Feature Scaling**: Menggunakan `StandardScaler` untuk menyamakan skala fitur numerik, penting untuk model sensitif skala seperti regresi logistik dan kNN.
- **Oversampling (SMOTE)**: Menambah kelas minoritas hingga 50% dari kelas mayoritas untuk mengatasi ketidakseimbangan kelas.
- **Undersampling**: Mengurangi kelas mayoritas hingga 80% dari kelas minoritas agar distribusi kelas lebih seimbang.
- **Testing Data for Model**: Membagi data (original, oversampled, dan undersampled) menjadi set latih dan uji untuk mengevaluasi kinerja model pada berbagai distribusi.

Size Differences (Before and After):

BYTE ME

Modeling Result (1)

Table 1. All Model (Data Biasa)

	Metric	AdaBoost	Decision Tree	kNN	Logistic Regression	Random Forest	Stacking	XGBoost
1.	Precision (Test)	1.0000	0.1121	0.1484	0.5000	1.0000	0.3299	0.5385
2.	Accuracy (Test)	0.9200	0.8469	0.9146	0.9200	0.9200	0.9196	0.9200
3.	Accuracy (Train)	0.9194	1.0000	1.0000	0.9194	1.0000	0.9612	0.9195
4.	ROC AUC (Test)	0.6933	0.5206	0.5435	0.6799	0.6544	0.6847	0.7109
5.	ROC AUC (Train)	0.6910	1.0000	1.0000	0.6770	1.0000	0.9998	0.7726
6.	F1-Score (Test)	0.0003	0.1213	0.0258	0.0003	0.0003	0.0087	0.0019
7.	F2-Score (Test)	0.0002	0.1276	0.0173	0.0002	0.0002	0.0055	0.0012
8.	Recall (CV Train)	0.0002	0.1260	0.0147	0.0002	0.0002	0.0066	0.0006
9.	Recall (CV Test)	0.0001	0.1387	0.0148	0.0001	0.0000	0.0036	0.0025
10.	Recall (Test)	0.0001	0.1322	0.0142	0.0001	0.0001	0.0044	0.0010

1 - 10 / 10

Recall yang sangat rendah (0.0044), menunjukkan banyaknya FN

Table 2. All Model (Oversampling)

	Metric	AdaBoost	Decision Tree	kNN	Logistic Regres...	Random Forest	Stacking	XGBoost
1.	Precision (Test)	0.9669	0.1121	0.5996	0.5775	0.9996	0.9817	0.9985
2.	ROC AUC (Train)	0.9214	1.0000	1.0000	0.7071	1.0000	1.0000	0.9647
3.	ROC AUC (Test)	0.9199	0.5206	0.9445	0.7042	0.9655	0.9658	0.9471
4.	Accuracy (Train)	0.8974	1.0000	1.0000	0.6935	1.0000	1.0000	0.9408
5.	Accuracy (Test)	0.8967	0.8469	0.7757	0.6922	0.9370	0.9443	0.9390
6.	F1-Score (Test)	0.8221	0.1213	0.7460	0.3856	0.8959	0.9104	0.8995
7.	F2-Score (Test)	0.7543	0.1276	0.8741	0.3215	0.8434	0.8725	0.8489
8.	Recall (CV Test)	0.7168	0.1387	0.8955	0.2874	0.7778	0.8297	0.8175
9.	Recall (Test)	0.7150	0.1322	0.9870	0.2894	0.8117	0.8489	0.8183
10.	Recall (CV Train)	0.7144	0.1260	0.9755	0.2908	0.8076	0.8470	0.8205

1 - 10 / 10

Stacking dan **XGBoost** adalah model terbaik dalam pendekatan ini dengan F2-Score yang tinggi,

Table 3. All Model (Undersampling)

	Metric	AdaBoost	Decision Tree	kNN	Logistic Regres...	Random Forest	Stacking	XGBoost
1.	ROC AUC (Train Set)	0.6916	1.0000	1.0000	0.6843	1.0000	1.0000	0.8808
2.	ROC AUC (Test Set)	0.6876	0.5484	0.5640	0.6762	0.6717	0.6706	0.6898
3.	Accuracy (Train Set)	0.6455	1.0000	1.0000	0.6405	1.0000	1.0000	0.7960
4.	Accuracy (Test Set)	0.6449	0.5533	0.5556	0.6391	0.6364	0.6325	0.6431
5.	Precision (Test Set)	0.6266	0.4996	0.5023	0.6157	0.6254	0.6146	0.6150
6.	F1-Score (Test Set)	0.5596	0.5009	0.4867	0.5575	0.5315	0.5349	0.5728
7.	F2-Score (Test Set)	0.5259	0.5018	0.4779	0.5275	0.4876	0.4963	0.5502
8.	Recall (Cross-Valid...	0.5159	0.5184	0.4737	0.5008	0.4522	0.4649	0.5348
9.	Recall (Cross-Valid...	0.5118	0.4945	0.4808	0.5005	0.4554	0.4724	0.5343
1...	Recall (Test Set)	0.5056	0.5023	0.4721	0.5093	0.4622	0.4735	0.5361

1 - 10 / 10

Pendekatan ini lebih baik daripada data biasa, tetapi masih menghasilkan FN yang cukup tinggi dan cenderung overfit pada kelas defaulter

Modeling Result (2)

Table 4. All Model (Hyperparameter Data Biasa)

	Metric	AdaBoost	Decision Tree	kNN	Logistic Regres...	Random Forest	Stacking	XGBoost
1.	Accuracy (Test)	0.9190	0.8535	0.9033	0.9200	0.9200	0.9196	0.9192
2.	Accuracy (Train)	0.9189	0.9998	1.0000	0.9194	0.9998	0.9627	0.9251
3.	ROC AUC (Train)	0.7329	1.0000	1.0000	0.6776	1.0000	0.9998	0.9016
4.	ROC AUC (Test)	0.6969	0.5242	0.5334	0.6805	0.6564	0.6848	0.6893
5.	Precision (Test)	0.3168	0.1210	0.1213	0.5000	0.0000	0.3391	0.3590
6.	Recall (CV Test)	0.0242	0.1340	0.0393	0.0001	0.0000	0.0040	0.0197
7.	F1-Score (Test)	0.0220	0.1266	0.0525	0.0003	0.0000	0.0106	0.0260
8.	F2-Score (Test)	0.0141	0.1302	0.0392	0.0002	0.0000	0.0067	0.0167
9.	Recall (CV Train)	0.0134	0.1304	0.0386	0.0002	0.0003	0.0068	0.0138
10.	Recall (Test)	0.0114	0.1327	0.0335	0.0001	0.0000	0.0054	0.0135

1 - 10 / 10

Table 5. Best Model of Undersampling (Before & After Hyperparameter)

	Metric	XGBoost Before Tuning	XGBoost After Tuning
1.	Precision (Test Set)	0.9985	0.6385
2.	ROC AUC (Train Set)	0.9647	0.7697
3.	ROC AUC (Test Set)	0.9471	0.7112
4.	Accuracy (Train Set)	0.9408	0.7004
5.	Accuracy (Test Set)	0.9390	0.6578
6.	F1-Score (Test Set)	0.8995	0.5839
7.	F2-Score (Test Set)	0.8489	0.5553
8.	Recall (Cross-Validation Train)	0.8205	0.5398
9.	Recall (Test Set)	0.8183	0.5378
10.	Recall (Cross-Validation Test)	0.8175	0.5412

1 - 10 / 10

Table 6. Best Model: Oversampling Without Tuning

	Metric	Stacking	XGBoost
1.	Precision (Test)	0.9817	0.9985
2.	ROC AUC (Train)	1.0000	0.9647
3.	ROC AUC (Test)	0.9658	0.9471
4.	Accuracy (Train)	1.0000	0.9408
5.	Accuracy (Test)	0.9443	0.9390
6.	F1-Score (Test)	0.9104	0.8995
7.	F2-Score (Test)	0.8725	0.8489
8.	Recall (CV Train)	0.8470	0.8205
9.	Recall (Test)	0.8489	0.8183
10.	Recall (CV Test)	0.8297	0.8175

1 - 10 / 10

Peningkatan yang signifikan dalam metrik utama, terutama **recall, F1-score, dan ROC AUC**. Model ini sangat cocok untuk memprediksi risiko default, memastikan jumlah kasus default yang lebih tinggi dapat diidentifikasi tanpa mengorbankan presisi.

Table 7. The Best Model of Oversampling (Before & After Hyperparameter)

	Metric	XGBoost Before Tuning	XGBoost After Tuning
1.	Precision (Test Set)	0.9985	0.9966
2.	ROC AUC (Train Set)	0.9646	0.9861
3.	ROC AUC (Test Set)	0.9469	0.9470
4.	Accuracy (Train Set)	0.9406	0.9460
5.	Accuracy (Test Set)	0.9387	0.9397
6.	F1-Score (Test Set)	0.8990	0.9009
7.	F2-Score (Test Set)	0.8483	0.8519
8.	Recall (Cross-Validation Train)	0.8197	0.8239
9.	Recall (Test Set)	0.8176	0.8220
10.	Recall (Cross-Validation Test)	0.8165	0.8208

1 - 10 / 10

```
df['TARGET'].value_counts()
```

```
TARGET
0    278843
1     24396
Name: count, dtype: int64
```

```
df['predictions'].value_counts()
```

```
predictions
0    300338
1      2901
Name: count, dtype: int64
```

24,396 nasabah (8.04%) -> 2,901 nasabah (0.96%) Penurunan jumlah default ini sejalan dengan **tujuan bisnis** Home Credit untuk **mengurangi kesalahan 7,08%** dalam mengidentifikasi nasabah default.

Model berhasil **meningkatkan akurasi** keputusan pemberian pinjaman dengan **meminimalkan kasus** di mana nasabah diklasifikasikan secara salah sebagai default **(false positives)**.

Revenue Calculation

Variable	Mean	AMT_CREDIT	Revenue
RATE_INTEREST_PRIVILEGED	0.773503	196114	151694.7673
RATE_INTEREST_PRIMARY	0.188357	196114	36939.4447

	Revenue	
Variable	Before (100%)	After (107,08%)
RATE_INTEREST_PRIVILEGED	151,694.77	162,434.76
RATE_INTEREST_PRIMARY	36,939.44	39,554.76
Total	188,634.21	201,989.51

Business Recommendations

Recomm.	Insights	Actionable Items
Focus on Customer Segmentation and Retargeting	Pengajuan kredit banyak berasal dari sektor Consumer Electronics, Connectivity, dan pembiayaan kebutuhan rumah tangga.	1. Lakukan segmentasi pelanggan berdasarkan kebutuhan kredit dan perilaku pembayaran. 2. Rancang promosi dan kampanye retargeting yang ditargetkan untuk kategori populer seperti elektronik konsumen dan kebutuhan rumah tangga. 3. Gunakan data pengajuan XNA untuk memahami peluang yang belum dimanfaatkan.
Optimize Regional Loan Applications	Mayoritas pengajuan berasal dari Credit and cash offices dan Country-wide, sedangkan Regional/Local kurang dimanfaatkan.	1. Perluas aksesibilitas layanan di area regional/lokal melalui peningkatan infrastruktur dan strategi promosi. 2. Luncurkan kampanye untuk meningkatkan pengajuan kredit di lokasi yang kurang terjangkau. 3. Fokuskan pengembangan produk untuk kebutuhan lokal yang spesifik.
Develop Tailored Loan Products	Data menunjukkan banyak nasabah menggunakan pinjaman untuk kebutuhan rumah tangga dan elektronik melalui skema cicilan berbunga.	1. Buat produk pinjaman yang sesuai dengan kebutuhan utama seperti renovasi rumah atau pembelian elektronik. 2. Berikan penawaran spesial berbunga rendah untuk kategori populer. 3. Dorong nasabah menggunakan cicilan dengan insentif tambahan seperti cashback atau diskon.
Enhance Default Prediction Accuracy	Feature penting dalam memprediksi gagal bayar meliputi faktor demografis seperti jumlah anggota keluarga (CNT_FAM_MEMBERS), jenis pekerjaan (OCCUPATION_TYPE), dan status pendidikan (NAME_EDUCATION_TYPE).	1. Prioritaskan edukasi keuangan untuk nasabah dengan profil berisiko tinggi berdasarkan fitur seperti pendidikan dan pekerjaan. 2. Integrasikan analisis berbasis SHAP ke dalam strategi evaluasi risiko untuk meningkatkan transparansi model. 3. Fokus pada segmentasi yang lebih baik berdasarkan faktor demografis untuk intervensi dini. 3. Menerapkan model XGB kedalam proses penilaian kredit sehinga dapat memprediksi calon nasabah yang beresiko.

Business Recommendations

Recomm.	Insights	Actionable Items
Implement Risk-Based Segmentation	Nasabah di sektor risiko tinggi seperti Business Entity Type 3 membutuhkan perhatian khusus.	1. Gunakan variabel kunci seperti `CNT_FAM_MEMBERS`, `NEW_PAYMENT_RATE`, dan `NAME_EDUCATION_TYPE` untuk segmentasi risiko yang lebih akurat. 2. Fokuskan kebijakan kredit dengan verifikasi tambahan untuk sektor berisiko atau dengan pola pengeluaran tidak stabil. 3. Meningkatkan frekuensi komunikasi dengan nasabah berisiko tinggi.
Flexible Credit Policies	Nasabah dengan keluarga kecil, pendidikan tinggi, dan rasio pembayaran tinggi memiliki risiko lebih rendah.	1. Tawarkan kredit kompetitif dengan suku bunga lebih rendah untuk nasabah profil ini. 2. Tinjau ulang kebijakan kredit untuk memberikan insentif pada kategori ini untuk meningkatkan loyalitas.

Kesimpulan

Model prediksi berhasil menurunkan default dari **8.04%** menjadi **0.96%** (penurunan kesalahan 7.08%).

Meningkatkan:

- **Revenue meningkat dari** 188,634.21 ke 201,989.51 (107,08%)
- **Efisiensi operasional** dengan pengurangan risiko kredit.
- **Kepuasan nasabah** berkat klasifikasi yang lebih akurat.
- **Keuntungan bisnis** melalui akses kredit yang lebih adil.

References

1. Afifudin, M., & Rizki, A. M. (2023). Analisis Perbandingan Penggunaan Model Machine Learning pada Kasus Deteksi Kemampuan Calon Klien dalam Membayar Kembali Pinjaman. Scan, XVIII(2). Universitas Pembangunan Nasional "Veteran" Jawa Timur.
2. Bi, Z., Gao, R., & Fang, S. (2024). A general framework for visualizing machine learning models. Preprints. https://doi.org/10.20944/preprints202402.0798.v1
3. Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022). Perbandingan Algoritma SVM, Random Forest dan XGBoost untuk Penentuan Persetujuan Pengajuan Kredit. Jurnal Nuansa Informatika, 16(1), 141. https://journal.uniku.ac.id/index.php/ilkom
4. Himberg, T. (2021). Loan Default Prediction with Machine Learning (Master's thesis). Åbo Akademi University, Faculty of Social Sciences, Business and Economics. https://urn.fi/URN:NBN:fi-fe2021120359162
5. Lin, J. (2024). Research on loan default prediction based on logistic regression, random forest, XGBoost, and AdaBoost. SHS Web of Conferences, 181. https://doi.org/10.1051/shsconf/202418102008
6. Ismunandar, D., Firdaus, M. R., & Alkhalifi, Y. (2024). Penerapan Hyperparameter Machine Learning dalam Prediksi Gagal Pinjam. INTI Nusa Mandiri, 19(1), 62–70. https://doi.org/10.33480/inti.v19i1.5612
7. Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering, 1022(1), 012042. https://doi.org/10.1088/1757-899X/1022/1/012042
8. Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2021). Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. Annals of Operations Research, 330, 1–29. https://doi.org/10.1007/s10479-021-04114-z
9. Nagajyothi, V. (2020). Loan approval prediction using KNN, decision tree, and Naïve Bayes models. International Journal of Engineering in Computer Science, 2(1), 32–37. https://doi.org/10.33545/26633582.2020.v2.i1a.30
10. Omogbhemhe, M. I. (2021). Model for Predicting Bank Loan Default using XGBoost. International Journal of Computer Applications, 183(32), 1-6. https://doi.org/10.5120/ijca2021921738
11. Zeng, L., Sun, J., & Zhou, Y. (2023). Auto loan default prediction based on Stacking model. Proceedings of the International Conference, 31, 1-6. https://doi.org/10.2991/978-94-6463-270-5_31
12. Zhao, S., & Zou, J. (2021). Predicting loan defaults using logistic regression. Journal of Student Research, 10(1). https://doi.org/10.47611/jsrhs.v10i1.1326
13. Zhou, Y. (2023). Loan default prediction based on machine learning methods. In Proceedings of the 3rd International Conference on Big Data Economy and Information Management (BDEIM 2022) (pp. [pages if available]). Zhengzhou, China: EAI. https://doi.org/10.4108/eai.2-12-2022.2328740