

Home Credit Analysis

Kelompok 2 : CONEXUS

1. Abrar Hidayat
2. Anggun Dwi
3. Benedict Caesario
4. Bramantyo Raka (Ketua)
5. Pra Setiawan Silaen
6. Siti Nur Afifah
7. Tommy Septians



Team Member

Tommy Septians

Data Scientist



Bramantyo Raka

Project Leader



Siti Nur Afifah

Business
Intelligence



Abrar Hidayat

Data Analyst



Pra Setiawan Silaen

Business Intelligence



Anggun Dwi

Data Scientist



Benedict Caesario

Data Scientist



Table of Contents

**HOME
CREDIT** ?

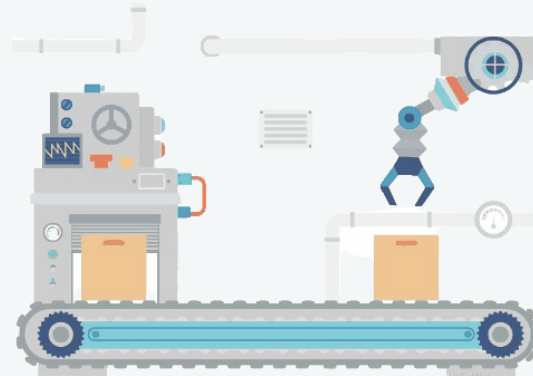
Background

EDA

Business
Recommendation

Modelling

Pre-Processing



Backgroun d

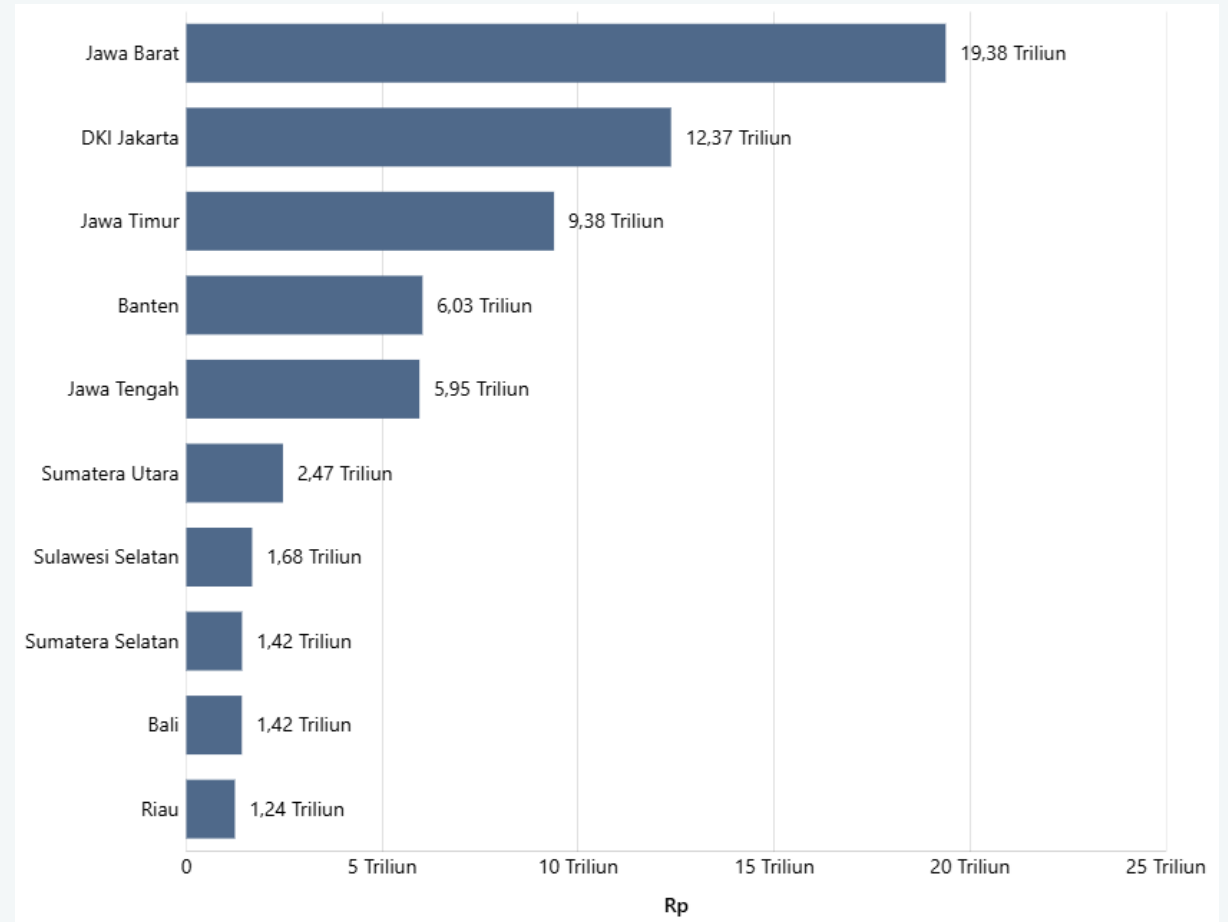


Background

HOME CREDIT

Perusahaan
pembiayaan berbasis
teknologi

Semakin tinggi jumlah pemohon
pinjaman, semakin besar potensi
keuntungan bagi Home Credit.

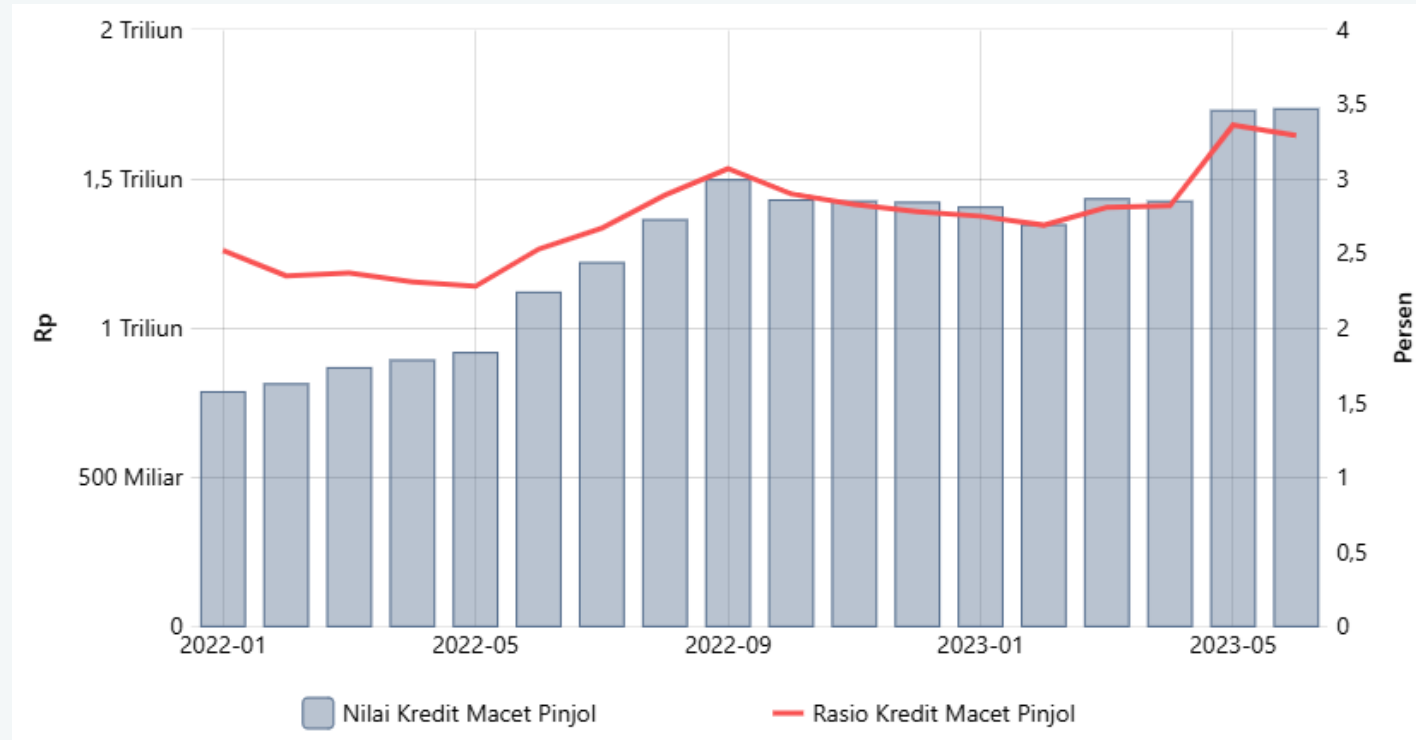


**10 Provinsi dengan Utang Pinjol Terbesar
September 2024, Jawa Barat Teratas**

Background

Data keuangan calon debitur sering kali terbatas, tidak terstruktur, atau tidak mencerminkan risiko kemampuan bayar di masa depan (Rahmah, 2016)

Akibatnya, proses pengambilan keputusan menjadi lambat dan berpotensi meningkatkan risiko kerugian karena kredit macet.



Nilai dan Rasio Kredit Macet Pinjol di Indonesia (Januari 2022–Juni 2023)

Background

Bagaimana cara meminimalkan risiko kerugian dan meningkatkan efisiensi pada home credit?

Goal



Meminimalkan risiko kerugian dan meningkatkan efisiensi operasional dalam proses pemberian kredit

Objective



Membuat model machine learning yang mampu memprediksi kemampuan nasabah untuk melunasi pinjaman

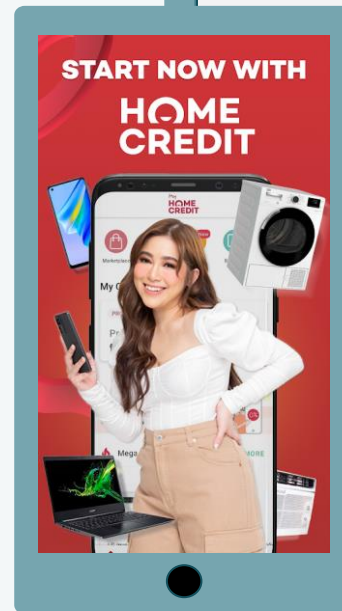
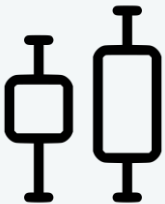
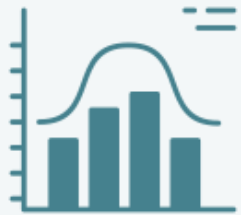
Business Metrics



Default Rate

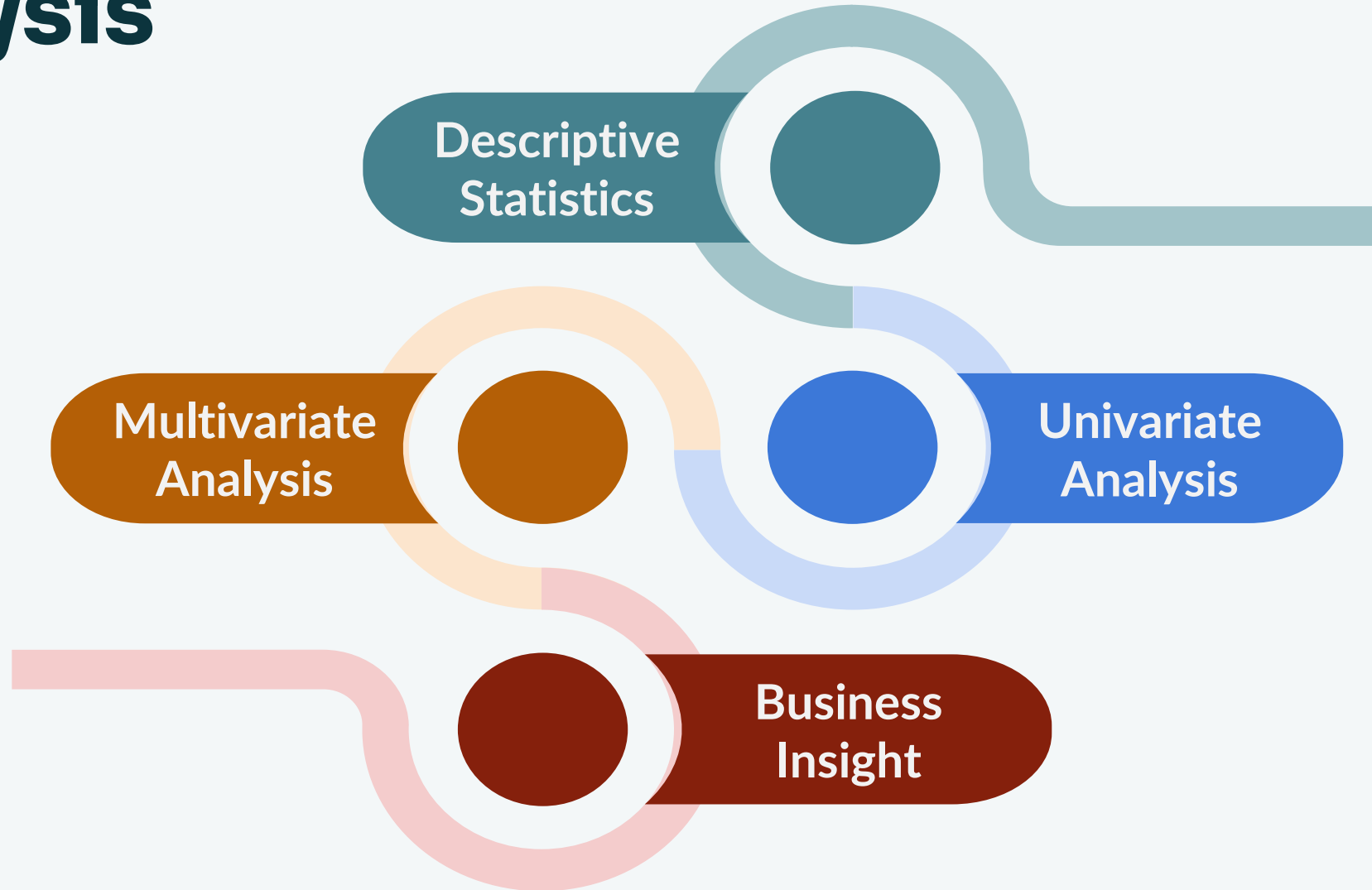
EDA

Exploratory Data Analysis

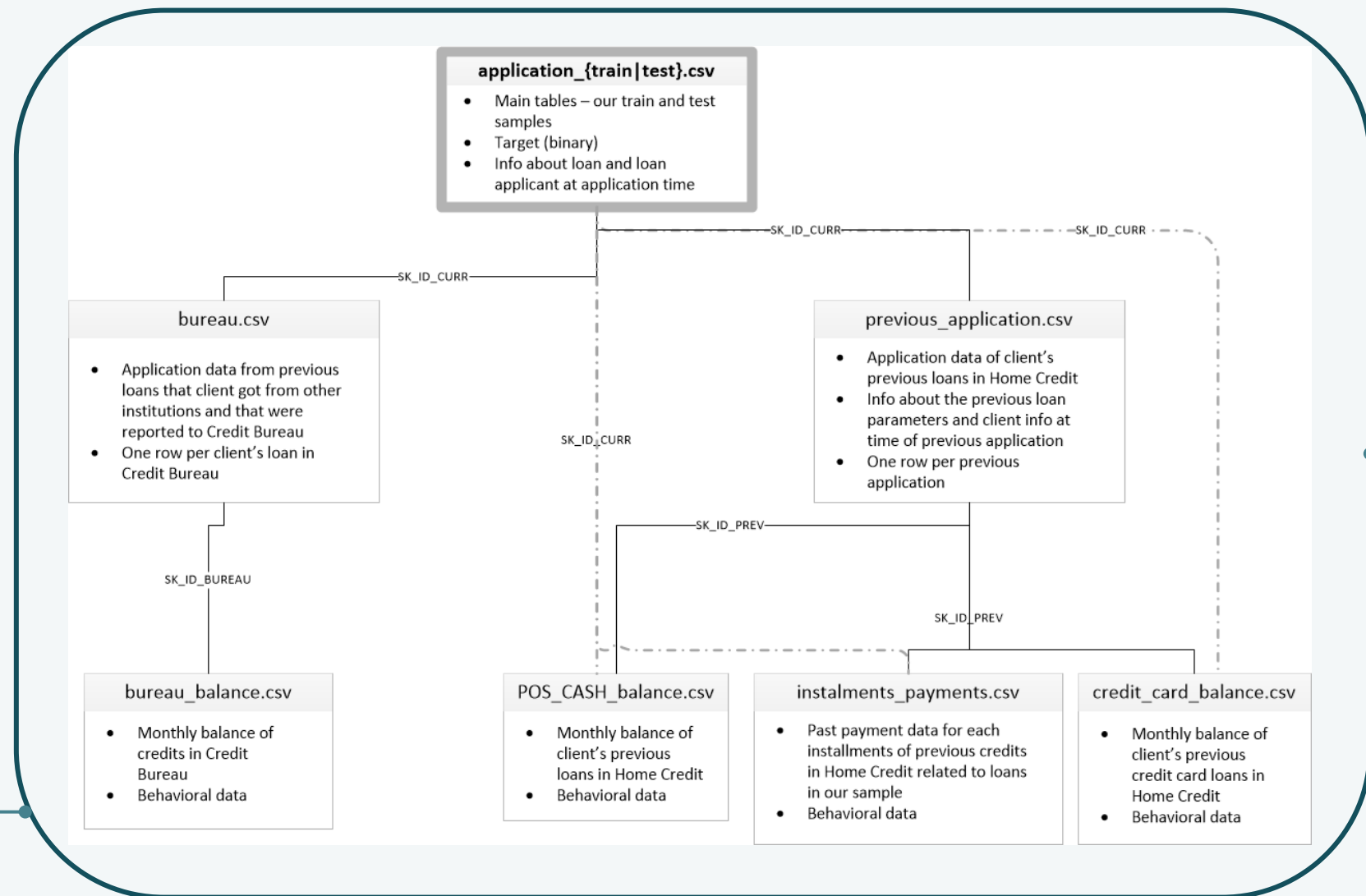


HOME CREDIT

Exploratory Data Analysis



Home Credit Dataset



Exploratory Data Analysis

❖ Descriptive Statistic

	Columns Name	Tipe Data	Null Values	Summary	
				Numerik (Outlier)	Kategori (top/mode)
1	application_train	✓	3	5	8
2	bureau	✓	7	10	-
3	credit_card_balance	✓	22	20	-
4	installments_payments	✓	2	6	-
5	POS_CASH_balance	✓	2	4	-
6	previous_application	✓	16	5	-

Exploratory Data Analysis

❖ Univariate Analysis

Dataset	Numerik			Kategori		
	Skewness			Outlier (column)	Jumlah Column	Frekuensi Tidak seimbang
	Negatif	Normal	Positif			
<u>application_train</u>	-	-	5	5	8	8
bureau	2	2	10	11	-	-
<u>credit_card_balance</u>	1	2	19	20	16	-
<u>installments_payments</u>	2	-	4	5	-	-
<u>POS_CASH_balance</u>	-	-	4	4	2	1
<u>previous_application</u>	3	2	8	8	6	2

1	Visualisasi lebih lanjut (ex: heatmap)	<ul style="list-style-type: none">• data understanding
2	data cleaning to identify	<ul style="list-style-type: none">• outliers• missing values
3	features transformation	<ul style="list-style-type: none">• on data skewed
4	features engineering	<ul style="list-style-type: none">• increase feature insight

Exploratory Data Analysis

❖ Multivariate Analysis

Dataset	Correlation			
	≥ 0.7	$0.5 \leq \text{corr} < 0.7$	$0.3 \leq \text{corr} < 0.5$	$\text{corr} < 0.3$
<u>application train</u>	3	-	-	18
bureau	7	2	1	95
<u>credit card balance</u>	14	6	22	189
<u>installments payments</u>	2	-	1	18
<u>POS CASH balance</u>	4	-	1	73
<u>previous application</u>	7	5	16	203

1

Korelasi antar fitur dan label

- hubungan non-linier
- skala unit yang berbeda
- skewed distribution atau adanya outliers

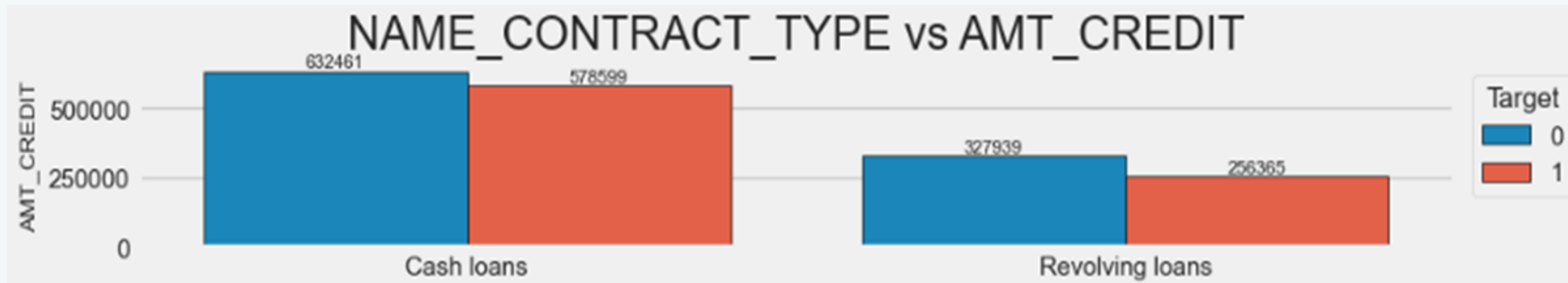
2

Korelasi antar fitur-fitur

- terdapat korelasi kuat, sedang, dan lemah

Exploratory Data Analysis

❖ Business Insight



Insight

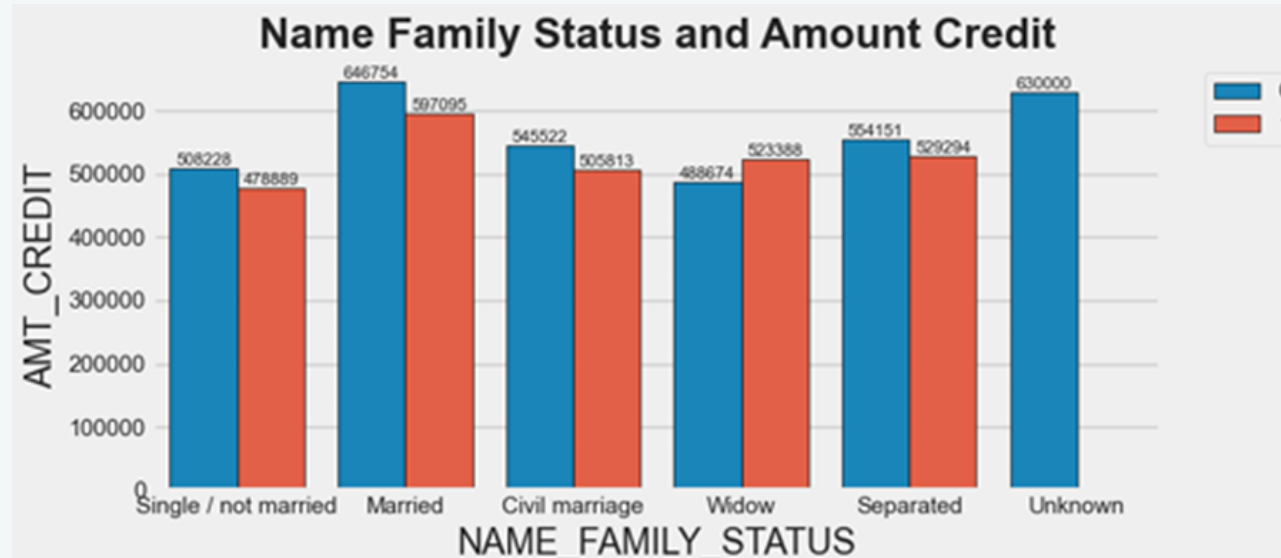
- *Cash loans* dominan:
 - Mayoritas pelanggan cenderung memilih *cash loans*
- *Revolving Loans*:
 - Jumlah kredit lebih rendah dibandingkan *cash loans*
- Secara keseluruhan nasabah yang gagal bayar jumlah kreditnya lebih rendah

Recommendation

- *Cash Loans*
 - Pertimbangkan untuk meningkatkan jumlah max pinjaman bagi nasabah dengan risiko rendah
- *Revolving Loans*
 - Pertimbangkan untuk penawaran batas kredit lebih besar bagi nasabah yang punya rekam jejak pembayaran baik.
- Program edukasi keuangan
 - Membantu mengelola pembayaran cicilan

Exploratory Data Analysis

❖ Business Insight



Insight

- Status perkawinan mempengaruhi jumlah kredit
- Secara keseluruhan, target 0 cenderung memiliki jumlah kredit lebih tinggi dibandingkan target 1, tanpa memandang status keluarga

Recommendation

- Untuk pola pembayaran yang baik, pertimbangkan untuk penawaran insentif, seperti bunga lebih rendah atau produk refinancing
- Optimasi penawaran atau penyesuaian produk:
 - Produk layanan yang disesuaikan dengan status perkawinan
 - ✓ produk kredit jangka pendek untuk lajang
 - ✓ produk pinjaman untuk perumahan
- Pengembangan produk lainnya yang dapat memenuhi kebutuhan spesifik dari segmen-segmen tersebut.

Pre-Processing



Pre-Processing

❖ Handling Missing Values

Dataset	Metode Penanganan
application_train	Imputasi Mean / Median, Modus
previous_application	Imputasi Median, Modus
POS_CASH_balance	Imputasi Mean
installments_payments	Imputasi Median
credit_card_balance	Imputasi Median
bureau	Imputasi Median

❖ Handling Duplicated Data

Dataset	Metode Penanganan
application_train	-
previous_application	-
POS_CASH_balance	-
installments_payments	-
credit_card_balance	-
bureau	-

Pre-Processing

❖ Handle Outliers

Dataset	Metode Penanganan
application_train	Trimming (Inter Quartile Range)
previous_application	Trimming (Inter Quartile Range)
POS_CASH_balance	Trimming (Inter Quartile Range)
installments_payments	Trimming (Inter Quartile Range)
credit_card_balance	Trimming (Inter Quartile Range)
bureau	Trimming (Inter Quartile Range)

❖ Aggregation

Dataset	Jenis Agregasi
previous_application	Modus
POS_CASH_balance	Max, Min, Mean, Sum
installments_payments	Mean
credit_card_balance	Mean
bureau	Mean
bureau_balance	Mean

Pre-Processing

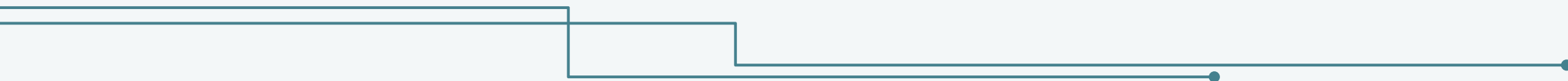


❖ Feature Transformation

Dataset	Metode
application_train	Logarithmic Transformation
previous_application	Normalization, Standardization, Logarithmic Transformation
POS_CASH_balance	Logarithmic Transformation
installments_payments	Standardization, Logarithmic Transformation
credit_card_balance	Logarithmic Transformation
bureau	Yeo-Johnson Transformation

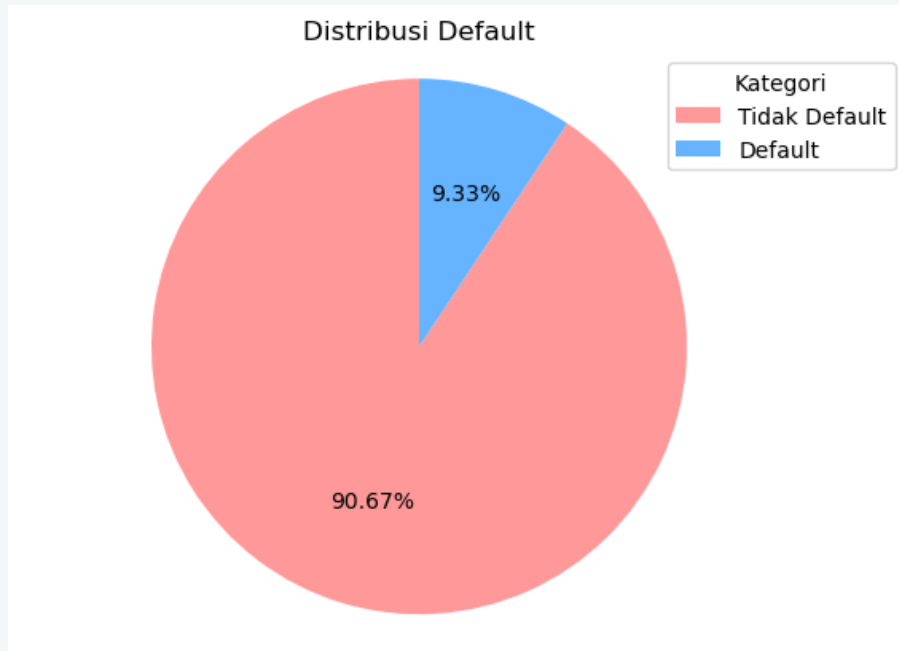
❖ Feature Encoding

Dataset	Metode
application_train	Label Encoding
previous_application	One Hot Encoding
POS_CASH_balance	One Hot Encoding
installments_payments	-
credit_card_balance	One Hot Encoding
bureau	Ordinal Encoding, Frequency Encoding



Pre-Processing

❖ Handle Class Imbalance



Pada dataset utama yaitu `application_train`, perlu dilakukan penanganan Imbalance Class dengan menggunakan metode undersampling.

Pre-Processing

❖ Feature Selection

Dataset	Banyak Kolom Awal	Jumlah Fitur yang Dipilih
application_train	120	65
previous_application	35	3
POS_CASH_balance	6	4
installments_payments	6	5
credit_card_balance	20	15
bureau	15	13
bureau_balance	2	1

❖ Feature Extraction

- NUM_DOCUMENTS
Total dokumen yang dilampirkan
- IS_WEEKEND_APPR_PROCESS_START
Apakah aplikasi diproses saat akhir pekan atau bukan
- EXT_SOURCE_MEAN
Rata-rata nilai sumber eksternal

Modelling



**HOME
CREDIT**

Modelling

❖ Metrics Evaluation

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Recall mengukur seberapa baik model dalam mengidentifikasi target positif.
Dimana, **target** : menentukan nasabah yang default/gagal bayar.

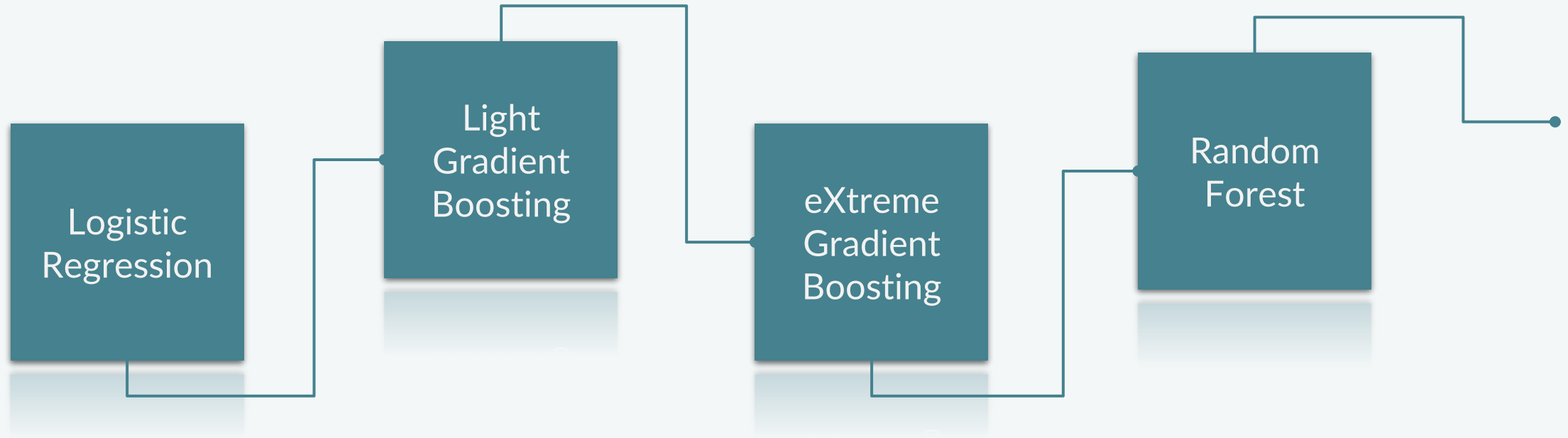
Maka,

True Positives (TP): Data yang benar-benar **default (1)** dan diprediksi sebagai **default (1)** oleh model.

False Negatives (FN): Data yang memiliki **default (1)**, tetapi diprediksi sebagai **tidak default (0)** oleh model.

Modelling

❖ Model Machine Learning



Modelling

❖ Logistic Regression

Model yang digunakan untuk memprediksi nilai kontinu (regresi) berdasarkan hubungan linear antara variabel independen (fitur) dan variabel dependen (target).

Logistic Regression			
Pre-processing	Model Machine Learning	Recall	
		Train Set	Test Set
-	Logistic Regression	0%	0%
Undersampling	Logistic Regression	50%	48%
Class Weight	Logistic Regression	50%	46%
Handling Outliers (IQR)	Logistic Regression	0%	0%
Handling Class Imbalance (SMOTE)	Logistic Regression	0%	0%
IQR & SMOTE	Logistic Regression	0%	0%
IQR, SMOTE & Hyperparameter Tuning	Logistic Regression	70%	62%

Kesimpulan : Model Logistic Regression tidak cukup baik karena hanya terbatas pada hubungan antara fitur dan target yang bersifat linear dan sederhana. Sedangkan dataset Home Credit bersifat lebih complex.

Modelling

❖ Light Gradient Boosting Machine (LGBM)

Algoritma pembelajaran mesin berbasis **gradient boosting**, yang dirancang untuk meningkatkan kecepatan dan efisiensi dibandingkan metode boosting tradisional seperti XGBoost.

Light GBM (Light Gradient Boosting Machine)			
Pre-processing	Model Machine Learning	Recall	
		Train Set	Test Set
-	Light Gradient Boosting Machine	68%	3%
Undersampling	Light Gradient Boosting Machine	100%	58%
Class Weight	Light Gradient Boosting Machine	100%	29%
Handling Outliers (IQR)	Light Gradient Boosting Machine	67%	4%
Handling Class Imbalance (SMOTE)	Light Gradient Boosting Machine	93%	5%
IQR & SMOTE	Light Gradient Boosting Machine	93%	6%
IQR, SMOTE & Hyperparameter Tuning	Light Gradient Boosting Machine	100%	4%

❖ XGBoost (eXtreme Gradient Boosting)

Algoritma berbasis **gradient boosting** yang dirancang untuk menghasilkan prediksi akurat dengan memanfaatkan kombinasi banyak pohon keputusan. XGBoost sangat efektif untuk dataset besar dan aplikasi yang membutuhkan performa prediksi tinggi.

XGBoost (eXtreme Gradient Boosting)			
Pre-processing	Model Machine Learning	Recall	
		Train Set	Test Set
-	XGBoost (eXtreme Gradient Boosting)	100%	6%
Undersampling	XGBoost (eXtreme Gradient Boosting)	100%	63%
Class Weight	XGBoost (eXtreme Gradient Boosting)	100%	14%
Handling Outliers (IQR)	XGBoost (eXtreme Gradient Boosting)	100%	6%
Handling Class Imbalance (SMOTE)	XGBoost (eXtreme Gradient Boosting)	35%	2%
IQR & SMOTE	XGBoost (eXtreme Gradient Boosting)	34%	5%
IQR, SMOTE & Hyperparameter Tuning	XGBoost (eXtreme Gradient Boosting)	100%	3%

Kesimpulan : Model **XGBOOST** dan **LGBM** tidak cukup optimal karena menghasilkan model yang overfitting dengan data latih, tetapi tidak dapat melakukan generalisasi pada data uji. Model tidak efektif untuk memprediksi kasus positif dalam data baru.

Modelling

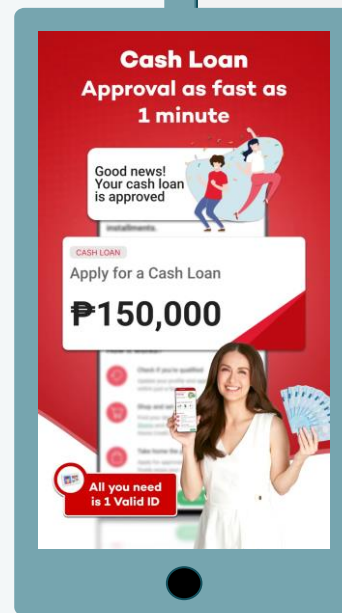
❖ Random Forest

Random Forest adalah model machine learning yang menggunakan ensemble method, model ini akan mengkombinasikan beberapa model Decision Tree.

Pre-Processing	Recall	
	Train Set	Test Set
-	100%	1.4%
Handling Class Imbalance (SMOTE)	100%	5%
Handling Class Imbalance (SMOTE) + Hyperparameter Tuning	83%	42%
Handling Class Imbalance (Undersampling)	100%	63%
Handling Class Imbalance (Undersampling) + Hyperparameter Tuning	67.5%	66%

Kesimpulan : Random forest dengan handling class imbalance undersampling dan hyperparameter tuning adalah model dengan performa terbaik dari model-model yang telah dicoba.

Business Recommendation



HOME CREDIT

Recommendation Automation



Reduce Manual Workload



Minimize Risk of Decision



**Manual Checking every 4
months**

Recommendation

Focus on False Negatives



WHY?

Failure to detect potential defaults, could lead to a harm towards the institution



HOW?

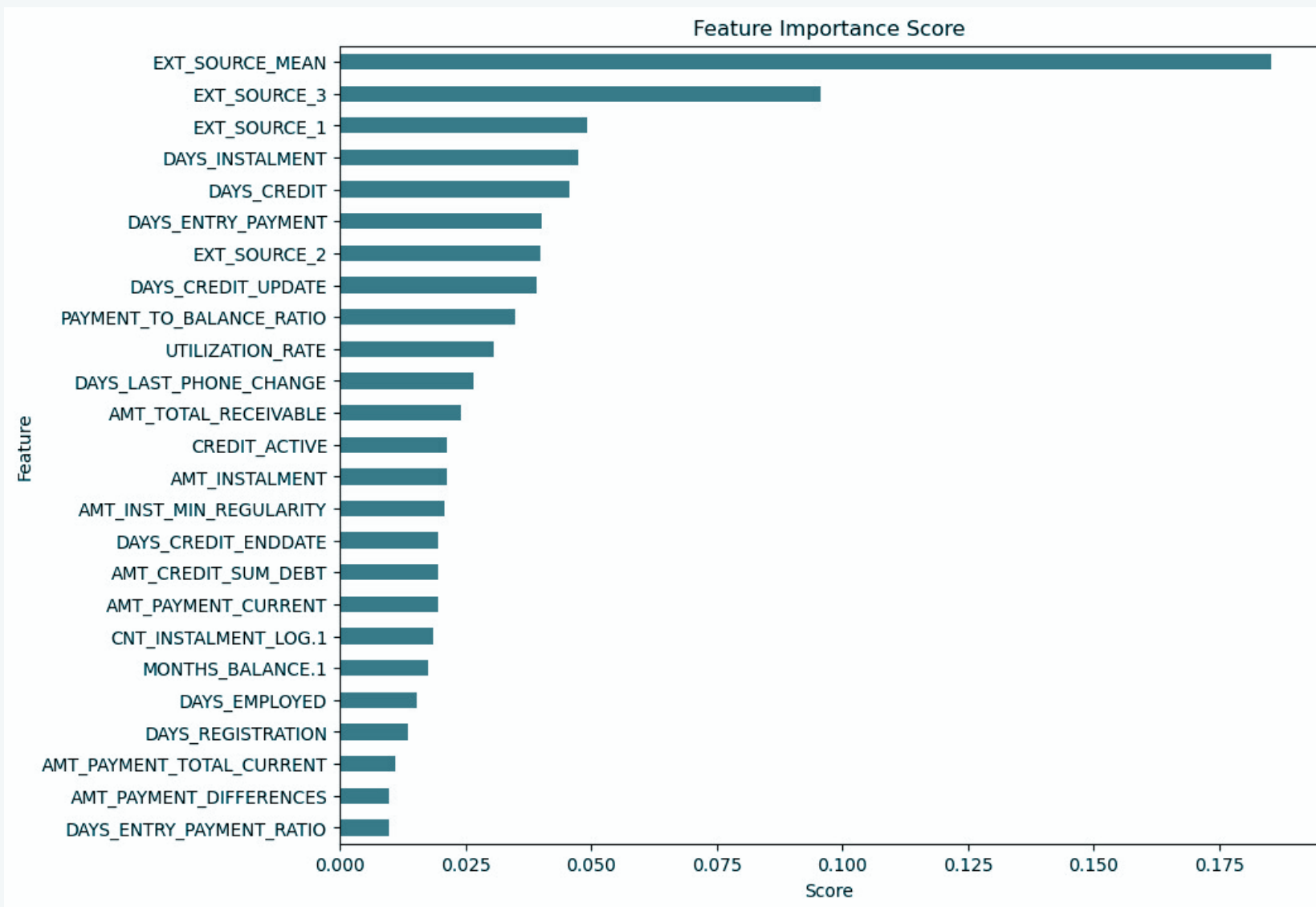
Feature importance and model parameter



HOW?

Using formula below:

$$\text{AMT_False_Negative_Credit}(\%) = \left(\frac{\text{Total Kredit False Negative}}{\text{Total Kredit}} \right) \times 100\%$$

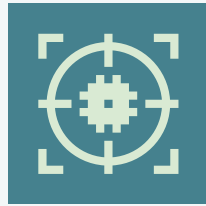


BUSINESS INSIGHT



DAYS_INSTALLMENT, DAYS_CREDIT_UPDATE

The higher days indicates **increasing of possibility of default**



PAYMENT_TO_BALANCE _RATIO

The higher values indicates customer **able to pay**.

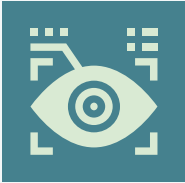


UTILIZATION_RATE

Customers who have a high utilization rate tend to have potential to **default**

Recommendation

Default Rate Management Continuous Monitoring



Track credit performance



Product Strategy

Providing credit limits based on customer capabilities



Objective

Maintain a low proportion of loan defaults



Utilizing Historical Data

Detect patterns in defaulted loans in Indonesia



Formula

$$\text{Default Rate} = \left(\frac{\text{Total Kredit Default}}{\text{Total Kredit}} \right) \times 100\%$$

“Tujuan utama dari semua langkah ini adalah untuk meminimalisir risiko kerugian dan meningkatkan efisiensi operasional dalam proses pemberian kredit”

— **Conexus**

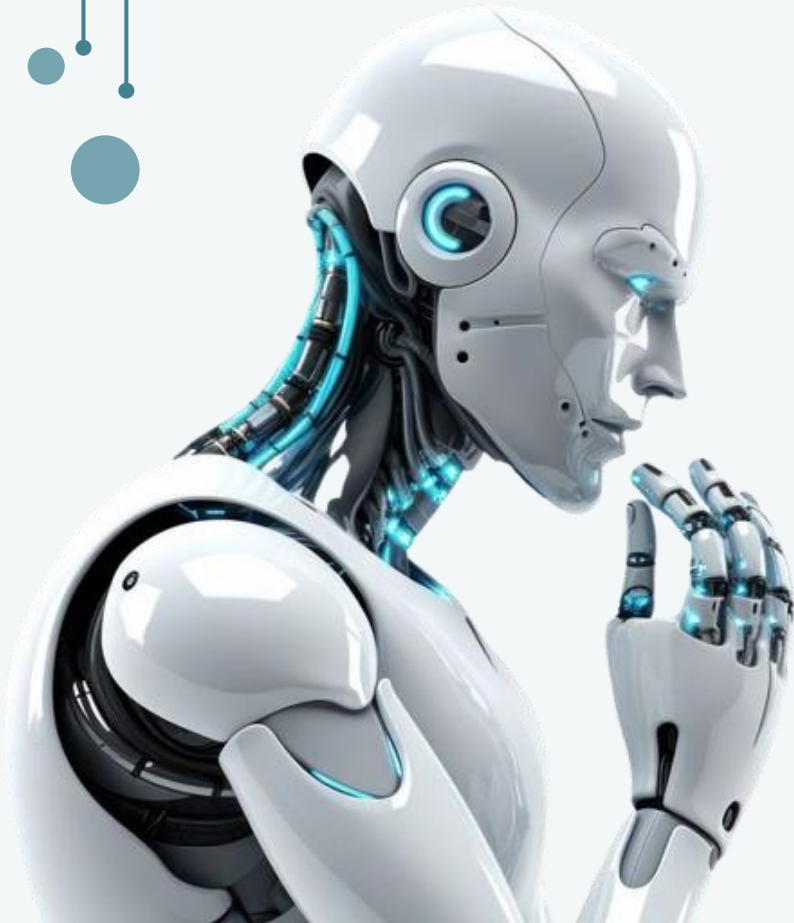


Thank You!

Any Question?

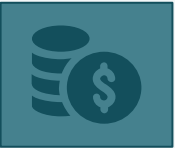
Let's Connect! 

Abrar	: Abrar Hidayat
Anggun	: Anggun Dwi Lestari
Ben	: Benedict C.
Bram	: Bramantyo Raka Adi Nugroho
Wawan	: Pra Setiawan Silaen
Ifa	: Siti Nur Afifah
Tommy	: Tommy Septians

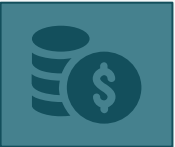


Appendix

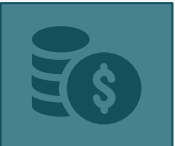
Source



1. **“10 Provisi dengan Utang Pinjol Terbesar September 2024, Jawa Barat Teratas”**
<https://databoks.katadata.co.id/keuangan/statistik/673eb7d8eba86/10-provinsi-dengan-utang-pinjol-terbesar-september-2024-jawa-barat-teratas>



1. **“Tren Kredit Macet Pinjol Meningkat Pada Semester I 2023”**
<https://databoks.katadata.co.id/keuangan/statistik/e691191a1332880/tren-kredit-macet-pinjol-meningkat-pada-semester-i-2023>



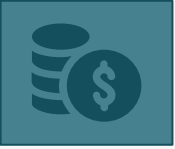
1. **ST RAHMAH IB (2016). ANALISIS TEKNIK PENYELESAIAN KREDIT MACET DAN PENGARUHNYA TERHADAP LAPORAN KEUANGAN PADA BANK MANDIRI Tbk MAKASSAR. *Skripsi*. Universitas Muhammadiyah Makassar.**
https://digilibadmin.unismuh.ac.id/upload/3781-Full_Text.pdf



1. **“What Can Big Data Tell Us About Loan Default, Lending Rate and Loan Amount in Financial Technology Peer-to-Peer Lending? Case of Indonesia”. (2024, November). Otoritas Jasa Keuangan. https://ojk.go.id/id/data-dan-statistik/research/working-paper/Documents/OJK_WP.21.01.pdf**

Appendix

Source



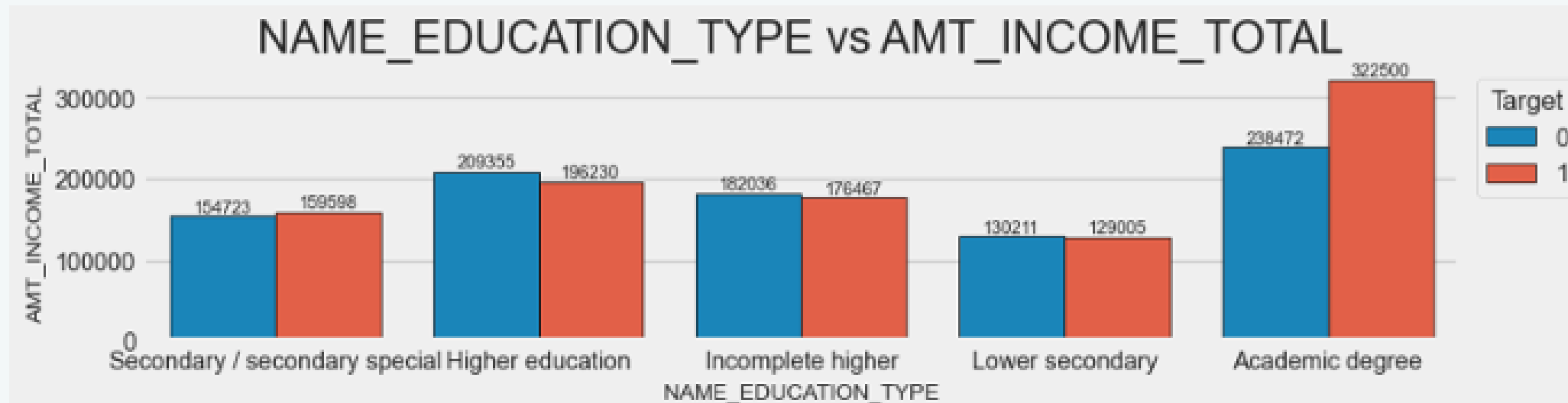
5. “Default Rate – Overview, Formula, Importance”

<https://corporatefinanceinstitute.com/resources/commercial-lending/default-rate/>



Appendix

EDA



Insight

- Rata-rata pendapatan meningkat seiring dengan tingkat pendidikan yang lebih tinggi.
- Kelompok berpendidikan tinggi (Higher education, Academic degree) tetap memiliki risiko default meskipun pendapatan mereka lebih tinggi. Hal ini menunjukkan bahwa pendapatan besar saja tidak selalu menjamin stabilitas finansial.

Recommendation

- Untuk peminjam dengan tingkat pendidikan rendah (Lower secondary, Secondary/special), tetapkan batas kredit yang lebih konservatif karena risiko default yang lebih sulit diprediksi berdasarkan pendapatan.
- Lakukan edukasi keuangan kepada kelompok berpendapatan tinggi namun tetap memiliki risiko default (terutama di tingkat Higher education dan Academic degree). Fokuskan pada pengelolaan utang dan stabilitas keuangan jangka panjang.
- Integrasikan tingkat pendidikan sebagai salah satu variabel penilaian risiko. Gunakan kombinasi tingkat pendidikan dan rasio pendapatan terhadap utang untuk memprediksi risiko default dengan lebih akurat.

Appendix

Pre-Processing

❖ Yeo-Johnson Transformation

Yeo-Johnson Transformation adalah teknik mengubah distribusi data agar menjadi lebih mendekati distribusi normal dengan menangani data yang mengandung nilai negatif dan nol.

$$y = \ln(x), \text{ jika } \lambda = 0$$

$$y = -\ln(-x), \text{ jika } \lambda = 0$$

Langkah-langkah Yeo-Johnson Transformation :

1. Identifikasi apakah data mengandung nilai negatif, nol, atau positif.
2. Pilih parameter λ . Nilai λ dapat dipilih menggunakan metode optimasi, seperti Maximum Likelihood Estimation (MLE), untuk meminimalkan deviasi antara data yang sudah ditransformasi dengan distribusi normal.
3. Lakukan transformasi sesuai dengan kondisi nilai x (positif atau negatif) dan nilai λ yang dipilih.
4. Evaluasi distribusi data setelah transformasi untuk memastikan apakah data sudah lebih mendekati distribusi normal.

Appendix

Background/Business Recommendation

Kredit bermasalah dapat disebabkan oleh faktor internal dan eksternal. Faktor internal penyebab kredit bermasalah yaitu kebijakan perkreditan yang ekspansif penyimpangan dalam pelaksanaan prosedur perkreditan, itikad kurang baik dari pemilik, pengurus atau pegawai bank, lemahnya sistem informasi kredit bermasalah. Sedangkan faktor eksternal penyebab kredit macet adalah: kegagalan usaha debitur, pemanfaatan iklim persaingan perbankan yang tidak sehat oleh debitur serta menurunnya kegiatan ekonomi dan tingginya suku bunga kredit. (Iswi Hariyani. 2010: 35 - 38).

Appendix

Modelling

❖ Logistic Regression

Logistic Regression					
Pre-processing	Model Machine Learning	Recall		Default Rate	
		Train Set	Test Set	Train Set	Test Set
-	Logistic Regression	0.00	0.00	0.01%	0.00%
Undersampling	Logistic Regression	0.50	0.48	42.41%	36.40%
Class Weight	Logistic Regression	0.50	0.46	36.52%	34.99%
Handling Outliers (IQR)	Logistic Regression	0.00	0.00	0.00%	0.00%
Handling Class Imbalance (SMOTE)	Logistic Regression	0.00	0.00	0.00%	0.00%
IQR & SMOTE	Logistic Regression	0.00	0.00	0.00%	0.00%
IQR, SMOTE & Hyperparameter Tuning	Logistic Regression	0.70	0.62	50.90%	35.22%

Appendix

Modelling

❖ Light Gradient Boosting Machine (LGBM)

Light GBM (Light Gradient Boosting Machine)					
Pre-processing	Model Machine Learning	Recall		Default Rate	
		Train Set	Test Set	Train Set	Test Set
-	Light Gradient Boosting Machine	0.6878	0.03	6.69%	0.54%
Undersampling	Light Gradient Boosting Machine	1.00	0.58	39.03%	36.40%
Class Weight	Light Gradient Boosting Machine	1.00	0.29	14.31%	12.54%
Handling Outliers (IQR)	Light Gradient Boosting Machine	0.67	0.04	6.70%	1.00%
Handling Class Imbalance (SMOTE)	Light Gradient Boosting Machine	0.93	0.05	46.69%	1.67%
IQR & SMOTE	Light Gradient Boosting Machine	0.93	0.06	46.66%	1.95%
IQR, SMOTE & Hyperparameter Tuning	Light Gradient Boosting Machine	1.00	0.04	50.00%	1.36%

Appendix

Modelling

❖ XGBoost (eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting)					
Pre-processing	Model Machine Learning	Recall		Default Rate	
		Train Set	Test Set	Train Set	Test Set
-	XGBoost (eXtreme Gradient Boosting)	1.00	0.06	9.98%	1.49%
Undersampling	XGBoost (eXtreme Gradient Boosting)	1.00	0.63	50.00%	37.85%
Class Weight	XGBoost (eXtreme Gradient Boosting)	1.00	0.14	9.94%	5.34%
Handling Outliers (IQR)	XGBoost (eXtreme Gradient Boosting)	1.00	0.06	10.01%	1.45%
Handling Class Imbalance (SMOTE)	XGBoost (eXtreme Gradient Boosting)	0.35	0.02	17.52%	1.04%
IQR & SMOTE	XGBoost (eXtreme Gradient Boosting)	0.34	0.05	16.97%	1.45%
IQR, SMOTE & Hyperparameter Tuning	XGBoost (eXtreme Gradient Boosting)	1.00	0.03	50.00%	1.18%

Appendix

Modelling

❖ Random Forest

Best Parameter-Undersampling

```
'n_estimators': 66,  
'min_samples_split': 11,  
'min_samples_leaf': 11,  
'max_leaf_nodes': 6,  
'max_features': 'log2',  
'max_depth': 28,  
'criterion': 'gini'
```