The Risk Takers

# Customer Analysis: Predicting Churn for Home Credit

*Leveraging Data Analysis and Machine Learning to Forecast Customer Churn*


Nouval Habibie
**Mentor**

Anggun Dwi Cahyani
**Facilitator**

Kevin Raihan Yassin
**Universitas Indonesia**

Mawar Alhani
**Universitas Lampung**

Mara Priliana
**Universitas Jenderal Achmad Yani Yogyakarta**

Luqman Noor Buat
**Universitas Trunojoyo Madura**

Meirizka Maulidya A.
**Universitas Gadjah Mada**

# Context of the work

The Risk Takers

Home Credit adalah penyedia pembiayaan konsumen internasional yang ingin memperluas penawarannya ke lebih banyak pelanggan untuk meningkatkan pendapatan.

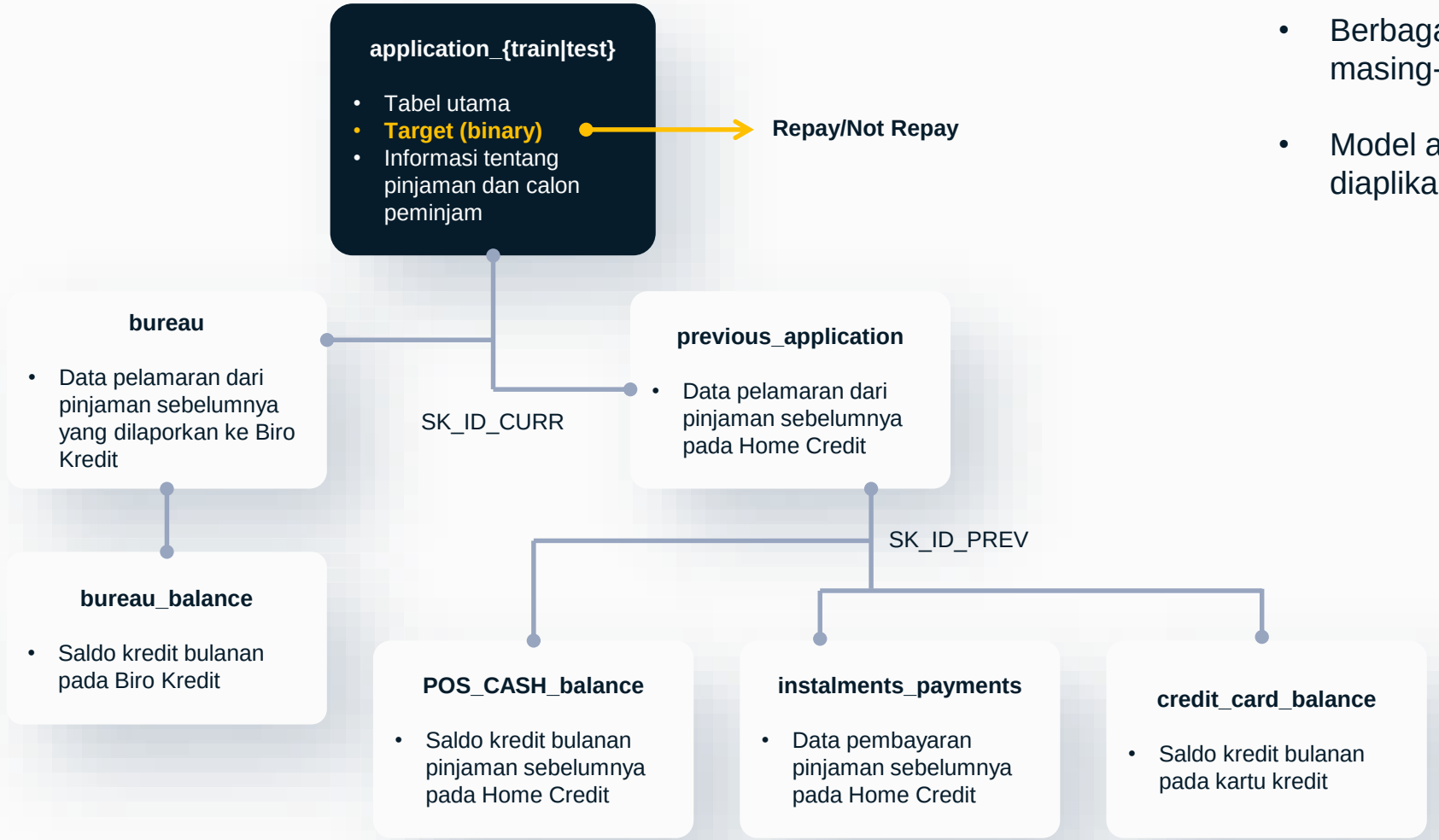
- Dengan meningkatnya penawaran pinjaman, muncul risiko gagal bayar yang lebih besar.
- Tantangannya terletak pada keragaman latar belakang dan motivasi individu yang mengajukan pinjaman.
- Memahami apakah pemohon dapat melunasi pinjaman berdasarkan karakteristiknya sangat penting untuk meminimalkan risiko gagal bayar dan memastikan pertumbuhan perusahaan.

Presentasi ini akan menjawab bagaimana Home Credit dapat secara efektif memprediksi dan mengklasifikasikan apakah pemohon akan dapat membayar kembali pinjamannya atau tidak, didasarkan pada eksplorasi data dan pembuatan model machine learning.

# Exploring the data…

The Risk Takers

*Tree diagram on dataset relationships*




*Supporting details*

- Berbagai data perilaku peminjam terdapat pada masing-masing data
- Model akan dilatih pada data train, dan diaplikasikan pada data test

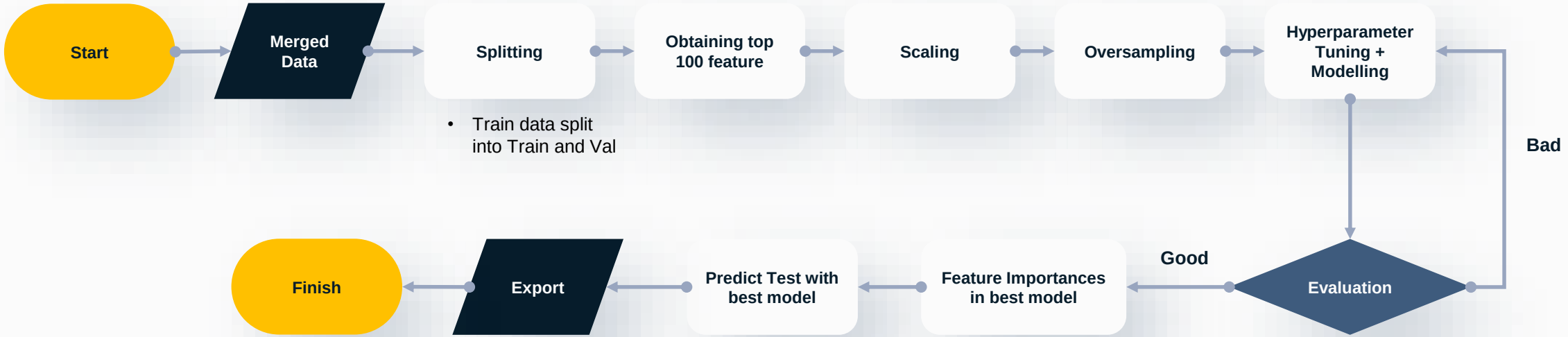
# …and what to do with it

The Risk Takers

*All individual dataset treatment flowchart*


- Train data and Test data merged separately

*Merged dataset treatment flowchart*




- Train data split into Train and Val

# Exploring predictive models to drive loan decisions

The Risk Takers

## Logistic Regression

Logistic regression populer untuk klasifikasi biner

Memodelkan hubungan antara fitur input dan probabilitas variabel target.

## Random Forest

Random Forest adalah metode pembelajaran ensemble yang menggabungkan beberapa *decision tree*.

Model ini dapat menangani hubungan non-linier dan menangkap interaksi fitur secara efektif.

## Naive Bayes

Naive Bayes adalah algoritma klasifikasi probabilistik berdasarkan teorema Bayes.

Model ini mengasumsikan bahwa fitur-fiturnya independen secara kondisional mengingat variabel target.

## Linear Discriminant Analysis

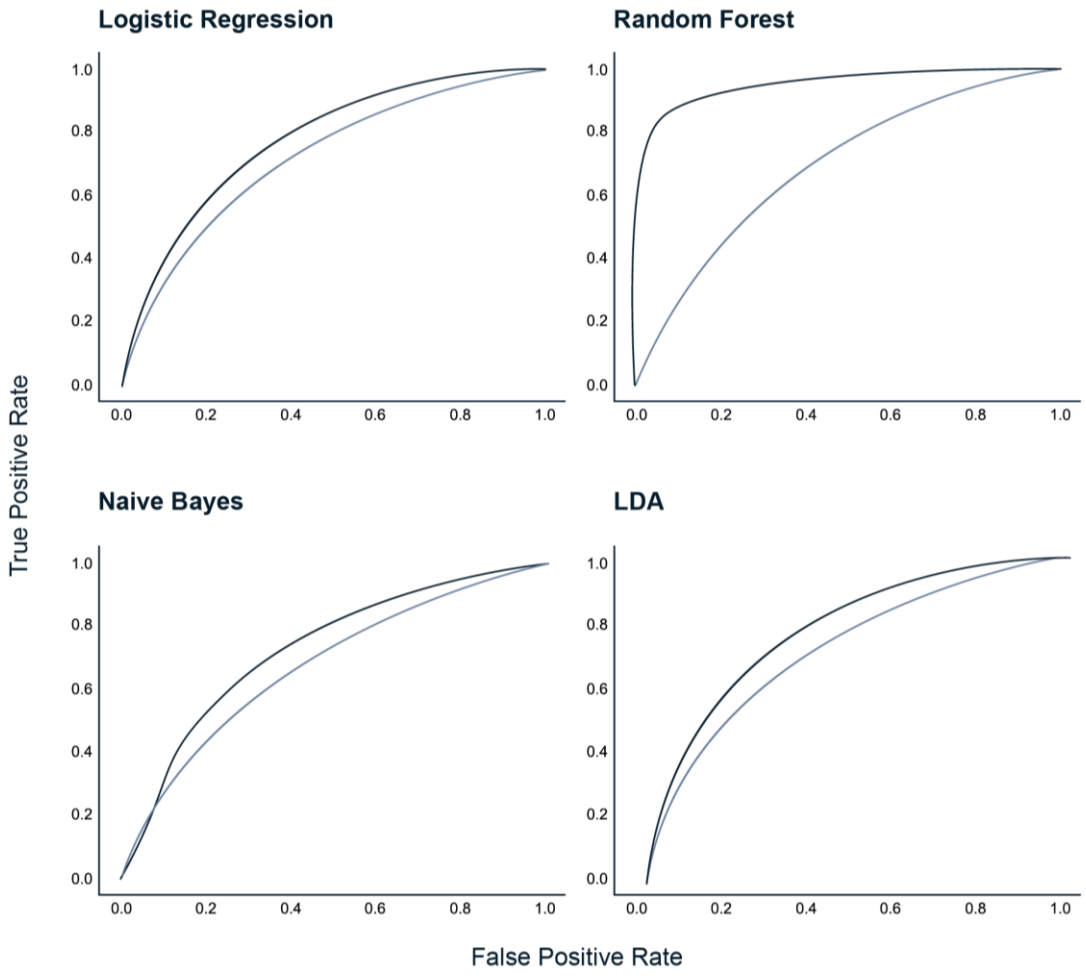
LDA adalah teknik pengurangan dimensi dan klasifikasi.

Model ini berusaha untuk menemukan kombinasi linear dari fitur yang paling baik memisahkan kelas yang berbeda.

# ROC curves explains model quality

The Risk Takers

*Comparing ROC curves from each model*




*AUC score comparison*

| Model Name | Train AUC | Val AUC |
|---|---|---|
| Logistic Regression | 0.78 | 0.72 |
| Random Forest | 0.95 | 0.69 |
| Naive Bayes | 0.73 | 0.66 |
| LDA | 0.78 | 0.72 |

*Observations:*

- ROC Curve milik model Random Forest diyakini terjadi overfitting pada data, model memberikan prediksi akurat untuk data train tetapi tidak untuk data Val.
- Val AUC bernilai 0.72 pada Logistic Regression dan LDA, menunjukkan bahwa model diterapkan dengan baik pada unseen data

**Which one is the best model?**

# LDA ascends to the pinnacle of performance by achieving highest recall and AUC scores on a tiny margin

The Risk Takers

*Final comparison to determine the best model*

| Model Name | Val Accuracy | Val Recall | Val AUC |
|---|---|---|---|
| Logistic Regression | 69.29 | 63.07 | 72.09 |
| Random Forest | **85.98** | 25.53 | 67.62 |
| Naive Bayes | 62.60 | 60.31 | 65.70 |
| LDA | 69.33 | **63.18** | **72.23** |

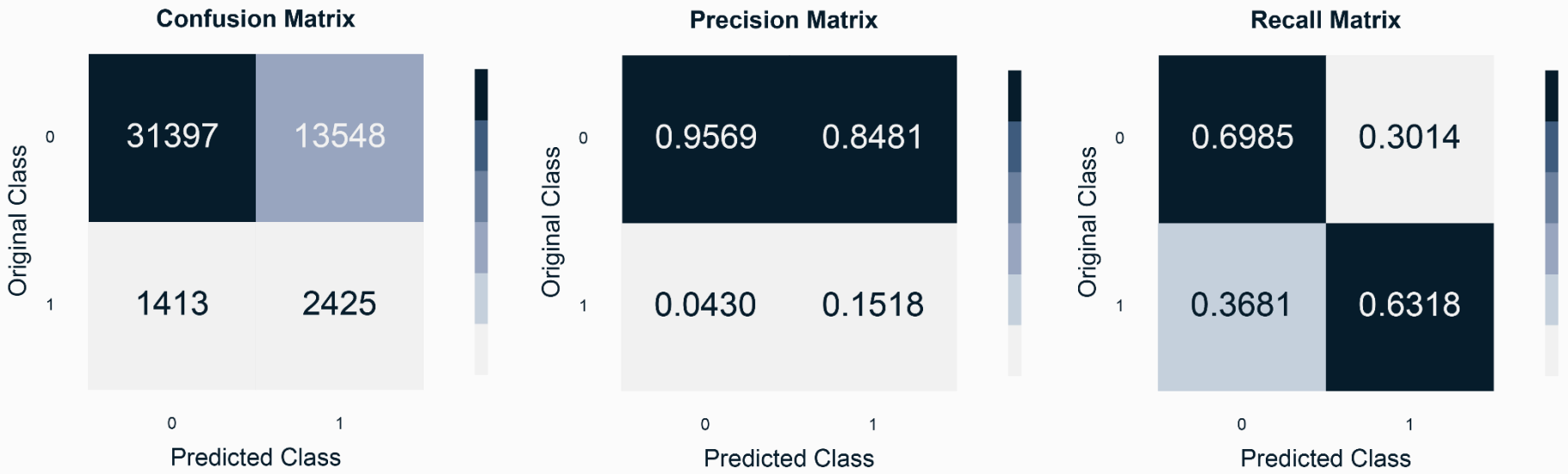
*Supporting details*

- Meskipun Random Forest memberikan nilai akurasi tertinggi pada dataset, kebutuhan objektif tidak terpaku pada nilai akurasi.
- Evaluasi akhir difokuskan pada nilai *recall*, yang mengevaluasi positif palsu.

# 63.18% recall score demonstrates model's aptitude in churn prediction

The Risk Takers

*Performance analysis on LDA*



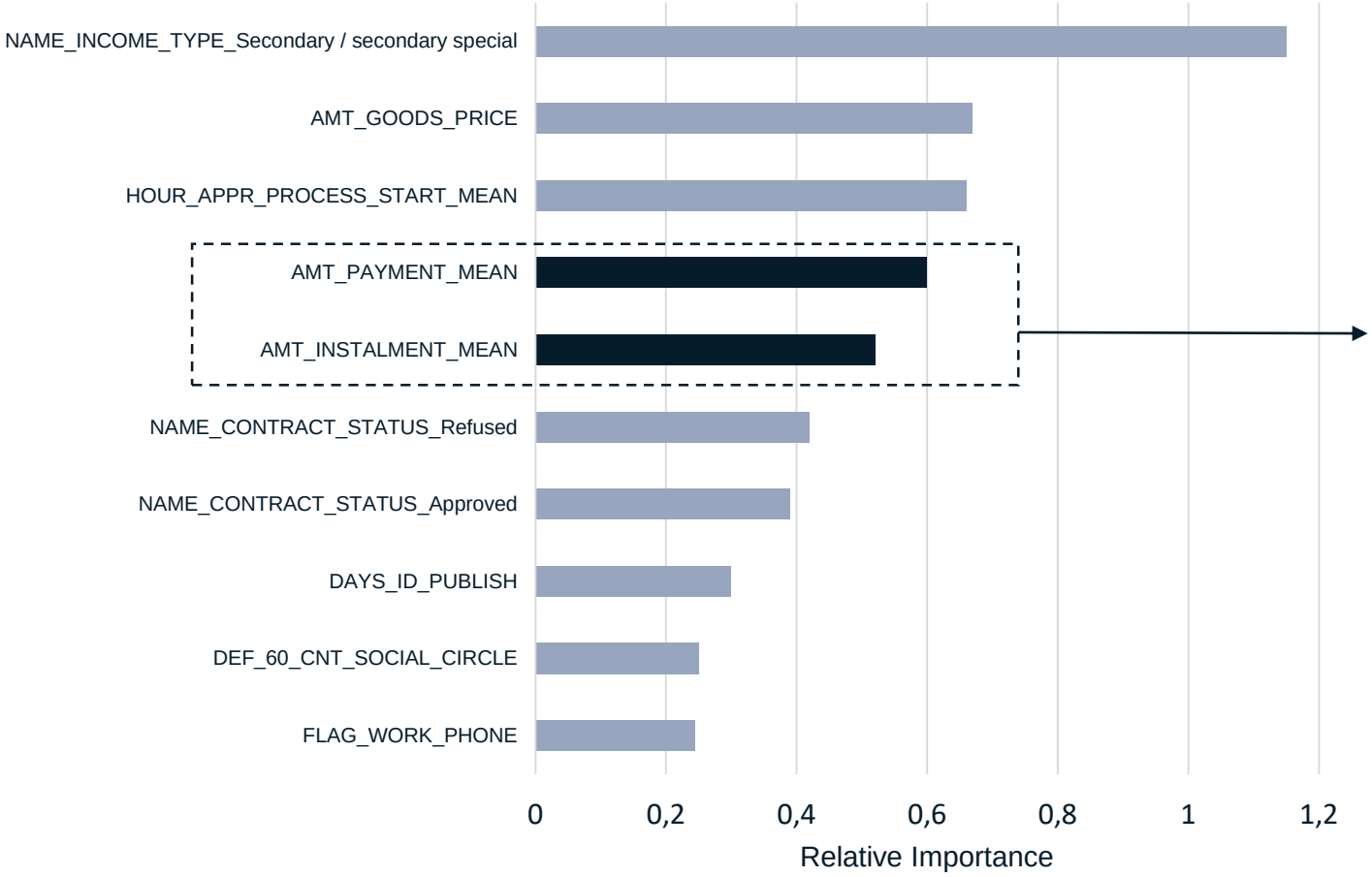

*Observations:*

- Menurut Confusion Matrix, nilai akurasi adalah: (31397 + 2425)/48783 x 100 = 69.33%
- Pada Precision Matrix, dapat dikatakan bahwa dari semua titik yang diprediksi termasuk dalam kelas 0, 95.69% di antaranya benar-benar termasuk dalam kelas 0 dan 4.3% di antaranya termasuk dalam kelas 1
- Pada Recall Matrix, model memperkirakan 69.85% di antaranya termasuk dalam kelas 0 dan 30.14% di antaranya termasuk dalam kelas 1. Begitu pula dengan semua titik yang semula termasuk kelas 1, 63.18% dari titik tersebut telah diprediksi oleh model menjadi kelas 1 dan 36.81% menjadi kelas 0.

# The top predictive factors impacting loan repayment

The Risk Takers

*Feature importances gained from the top model*




*Observations:*

- NAME_INCOME_TYPE_Secondary / secondary special menjadi top feature dalam menentukan kemampuan pembayaran pinjaman oleh peminjam

*Key Findings:*

- Diantara 5 fitur teratas, fitur yang mencerminkan kemampuan dan perilaku calon peminjam dalam pembayaran kredit adalah **AMT_INSTALMENT**, merupakan jumlah angsuran yang ditentukan dari kredit sebelumnya, dan **AMT_PAYMENT**, yang merupakan jumlah yang sebenarnya dibayarkan pada angsuran tersebut

The Risk Takers

# Summary

## Model Performance

Secara keseluruhan, semua model menunjukkan berbagai tingkat performa, dengan LDA mencapai *recall* tertinggi (63.18%) dan AUC score tertinggi (72.23%).

## Recall and Precision Trade-off

LDA mencapai skor *recall* 63.18%, menunjukkan kemampuannya untuk mengidentifikasi dengan benar proporsi kasus positif aktual yang relatif lebih tinggi.

## AUC Scores

Skor AUC pada data train secara konsisten lebih tinggi daripada skor AUC pada data cross-validation (Val) untuk semua model.

## Feature Importances

Analisis *feature importances* mengungkapkan faktor signifikan yang mendorong prediksi di setiap model.

# Three actions to improve loan approval processes, minimize default risk, and enhance decision-making in Home Credit's lending operations

The Risk Takers

| Recommendation | Cause | How | Benefit |
|---|---|---|---|
| **1. Implement Risk-Based Pricing** | Analisis dan pemodelan telah mengungkapkan variasi yang signifikan pada profil risiko pemohon pinjaman. | Menyesuaikan persyaratan pinjaman dan suku bunga berdasarkan penilaian risiko individual. | Dapat mengoptimalkan profitabilitas sekaligus memastikan harga yang adil bagi pelanggan. |
| **2. Enhance Customer Segmentation** | Analisis telah mengidentifikasi pola dan karakteristik yang berbeda di antara kelompok pelanggan yang berbeda. | Segmentasi pelanggan berdasarkan atribut dan perilaku mereka | Lebih baik menangani kebutuhan dan preferensi unik segmen pelanggan tertentu, yang mengarah pada peningkatan kepuasan pelanggan dan peningkatan tingkat pembayaran. |
| **3. Fraud Detection Mechanisms** | Model telah menunjukkan kemampuannya untuk memprediksi dan mengklasifikasikan pemohon pinjaman secara akurat | Mengintegrasikan algoritma deteksi penipuan tambahan dan sumber data | Dapat mengidentifikasi dan mengurangi aplikasi pinjaman bodong, mengurangi risiko kerugian finansial. |


The Risk
Takers