

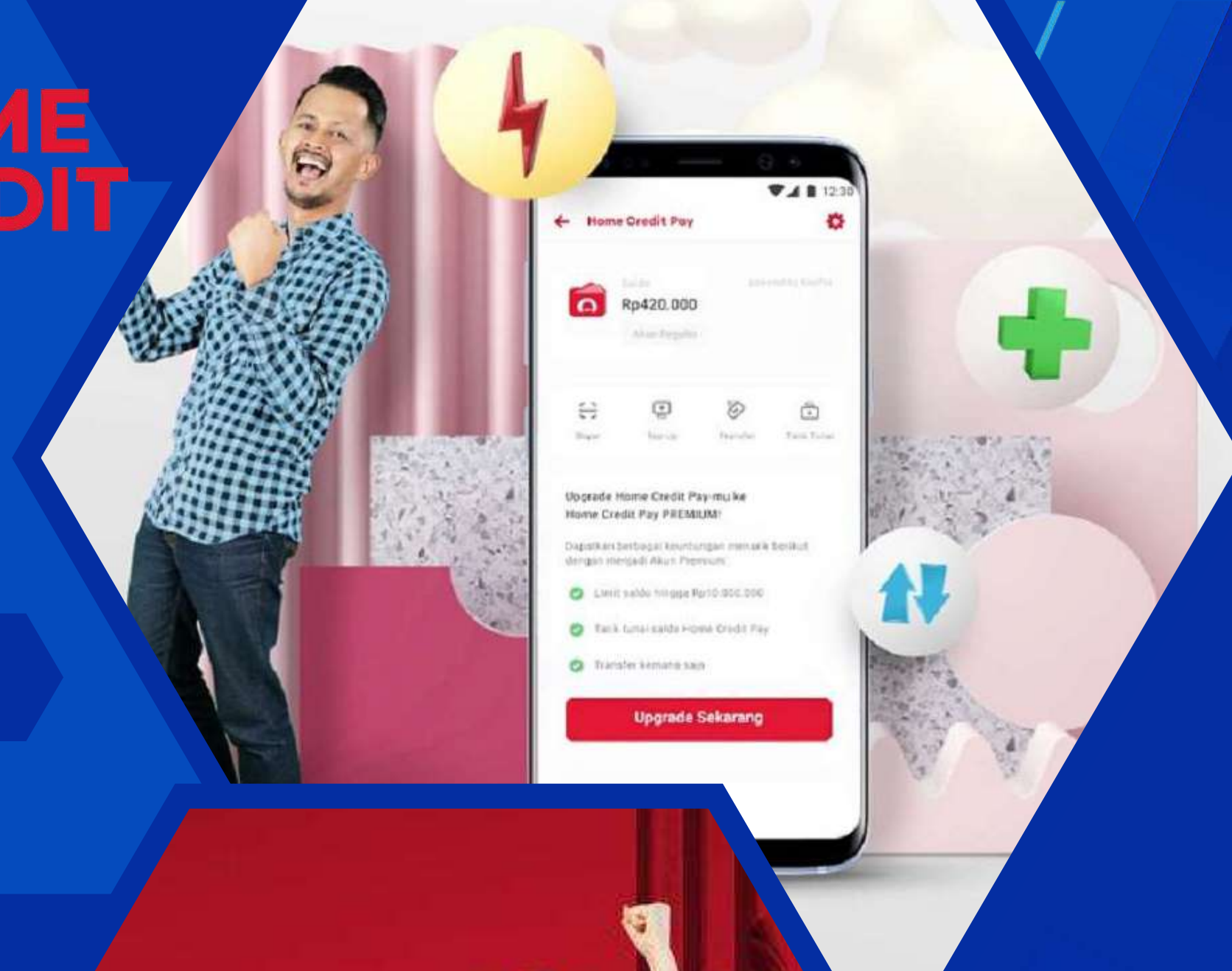
 Rakamin

HOME
CREDIT

Prediksi Risiko Gagal Bayar Nasabah Di Home Credit Indonesia

MUHAMAD RASYID ADITYA

Project Virtual Internship



Latar Belakang

Perusahaan pembiayaan seperti Home Credit menghadapi tantangan dalam menilai risiko calon nasabah. Tidak semua peminjam mampu membayar cicilan tepat waktu, sehingga diperlukan sistem yang dapat membantu memprediksi kemungkinan kegagalan pembayaran (credit default).

Dengan memanfaatkan data historis nasabah serta teknik machine learning, perusahaan dapat membuat model prediktif untuk mengidentifikasi calon peminjam berisiko tinggi lebih dini. Model ini dapat meningkatkan efektivitas penyaluran kredit dan mengurangi potensi kerugian finansial.



Rumusan Masalah

1. Bagaimana menerapkan dan membandingkan berbagai algoritma machine learning dalam membangun model prediksi risiko gagal bayar nasabah?
2. Algoritma mana yang dapat dievaluasi dan ditentukan sebagai model dengan kinerja paling optimal dalam memprediksi risiko gagal bayar nasabah berdasarkan hasil pengujian dan metrik evaluasi?

Tujuan

1. Menerapkan dan membandingkan berbagai algoritma machine learning dalam membangun model prediksi risiko gagal bayar nasabah guna memperoleh pemahaman terhadap performa dan karakteristik masing-masing algoritma.
2. Mengevaluasi dan menentukan algoritma machine learning dengan kinerja paling optimal untuk menghasilkan model prediksi yang akurat dan andal.



Alur Kerja

- **Business Understanding**

Menentukan tujuan analisis, yaitu memprediksi nasabah berisiko gagal bayar.

- **Data Preparation**

Melakukan data cleaning, encoding, scaling, dan feature selection.

- **Evaluation**

Menggunakan metrik evaluasi seperti Accuracy, Confusion Matrix, dan ROC-AUC Score

- **Data Understanding**

Mengexplorasi data `application_train.csv` dan `application_test.csv`

- **Modeling**

Menggunakan beberapa algoritma seperti Logistic Regression, Random Forest Classifier, dan XGBoost Classifier serta Oversampling (SMOTE) untuk menangani ketidakseimbangan data



Data Understanding

Dimensi

Kolom : 122

Baris :
307.511

Tipe

Category : 16

Numerik :
106

Dataset yang digunakan dalam proyek ini adalah data set Risiko Gagal Bayar Pinjaman Home Credit, yang mencakup 307.511 entri dengan berbagai fitur terkait karakteristik pemohon pinjaman, seperti informasi penghasilan, jumlah pinjaman, status keluarga, dan riwayat pinjaman sebelumnya. Variabel target dalam data set ini adalah kolom yang bernama TARGET, yang menunjukkan apakah pemohon mengalami kesulitan dalam membayar pinjaman (1) atau tidak (0).



Data Preprocessing

- **Data Cleaning**

Menghapus Missing Value, Menghapus Duplikat data

- **Feature Engineering**

Mengubah kolom DAYS_BIRTH menjadi Usia.

- **Label Encoding**

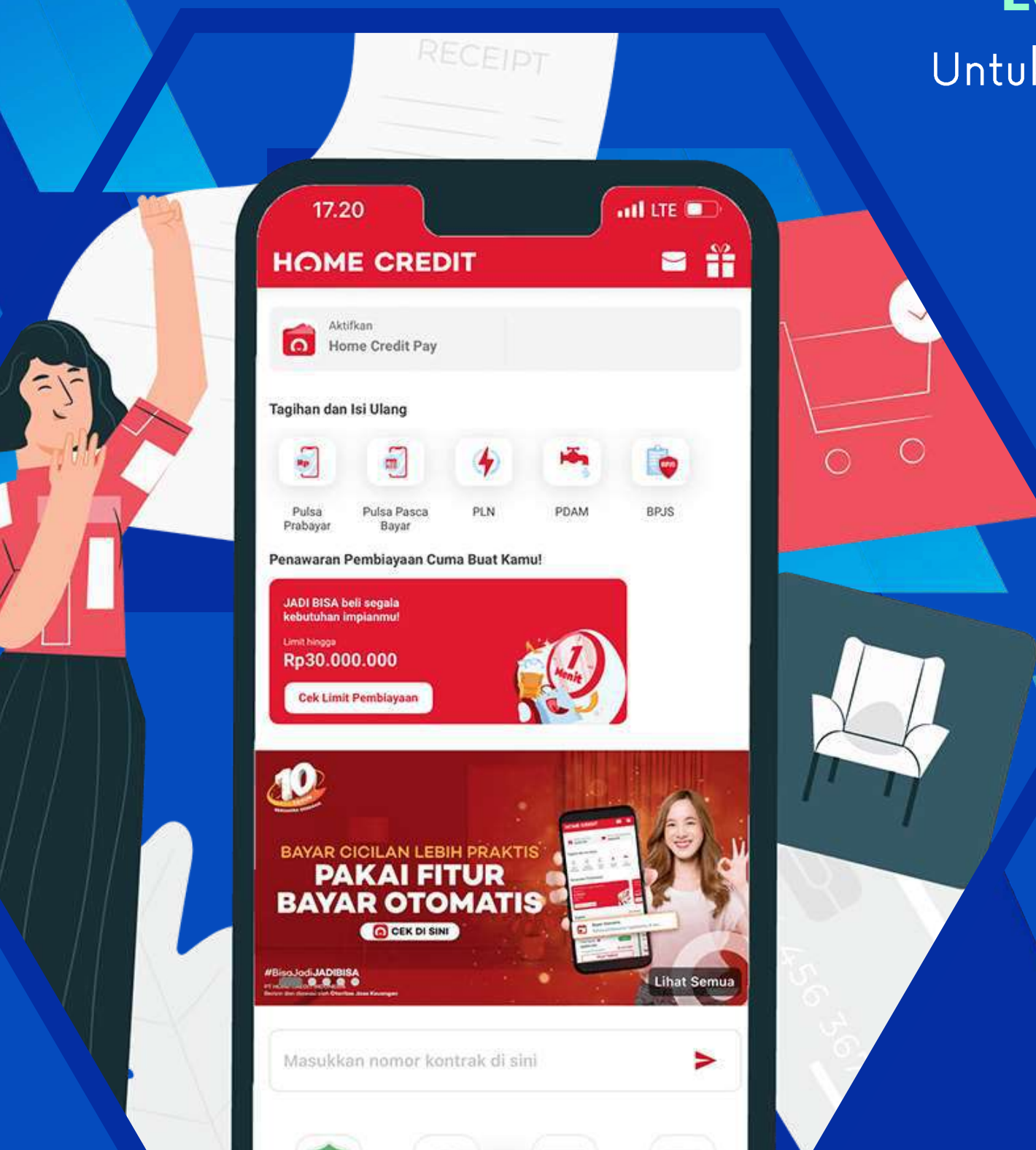
Untuk variabel kategorikal seperti CODE_GENDER, NAME_CONTRACT_TYPE, dll.

- **Normalisasi Data**

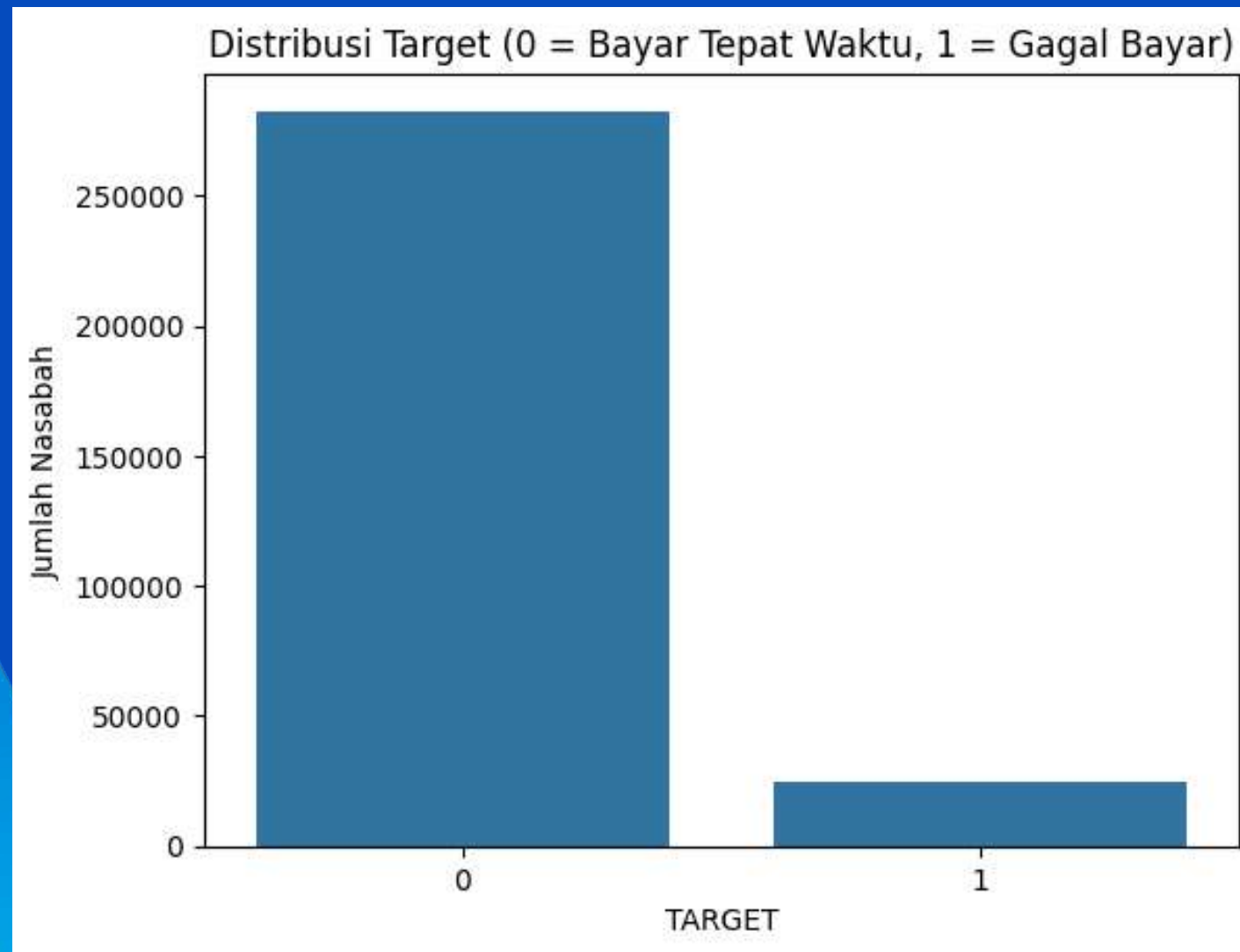
normalisasi menggunakan metode Min-Max Scaler agar seluruh fitur numerik berada pada skala yang sama, yaitu antara 0 dan 1. Tujuannya adalah untuk menghindari perbedaan skala antar variabel yang dapat memengaruhi kinerja model, mempercepat proses pembelajaran algoritma, serta meningkatkan akurasi prediksi.

- **SMOTE dan Undersampling**

Untuk mengatasi ketidakseimbangan kelas dan memastikan model dapat belajar dengan baik dari kelas minoritas.



Data Visualization and Business Insight



Berdasarkan visualisasi data diatas terdapat Risiko gagal bayar tergolong rendah, menunjukkan mayoritas nasabah memiliki profil pembayaran baik, namun proporsi yang tidak seimbang perlu diwaspadai agar model prediksi tidak bias terhadap nasabah lancar bayar.



pekerjaan berisiko tinggi seperti buruh dan sopir lebih rentan gagal bayar karena pendapatan tidak stabil, sehingga perusahaan perlu menyesuaikan kebijakan kredit dan bunga sesuai tingkat risiko pekerjaan.

Modelling dan Evaluation

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0,608	0,880	0,610	0,700	0,656
Random Forest	0,914	0,860	0,910	0,880	0,633
Gradient Boosting	0,914	0,860	0,910	0,880	0,634

Berdasarkan tabel diatas didapatkan, Jika fokus pada akurasi dan kestabilan performa keseluruhan, maka Random Forest atau Gradient Boosting merupakan pilihan terbaik, kemudian Jika tujuan utama adalah kemampuan membedakan antar kelas (AUC-ROC), maka Logistic Regression sedikit lebih unggul, namun dengan performa prediksi yang jauh lebih rendah secara keseluruhan.



Business Recommendation

1. Implementasikan model machine learning hasil analisis (seperti Random Forest atau Gradient Boosting) ke dalam sistem penilaian kredit Home Credit untuk membantu mengidentifikasi calon nasabah dengan risiko gagal bayar tinggi sejak awal proses pengajuan.
2. Gunakan hasil model untuk membuat segmentasi risiko nasabah. Misalnya dengan Risiko tinggi: diberikan batas pinjaman lebih kecil atau bunga lebih tinggi, dan risiko rendah: dapat menerima penawaran pinjaman dengan tenor lebih panjang atau bunga lebih rendah.
3. Model dapat dimanfaatkan untuk pemantauan berkelanjutan, dengan memberikan peringatan dini terhadap nasabah yang menunjukkan potensi keterlambatan pembayaran, sehingga tim penagihan dapat mengambil tindakan preventif lebih cepat.
4. Berdasarkan hasil visualisasi, profesi seperti buruh dan sopir memiliki tingkat gagal bayar lebih tinggi. Perusahaan dapat melakukan penyesuaian kebijakan pinjaman untuk kategori pekerjaan tersebut serta memberikan edukasi finansial tambahan atau pendampingan manajemen keuangan bagi kelompok ini.



Thank You

GITHUB:
https://github.com/rsydadtya/Final-Task-Data-Scientist-HOME-CREDIT_Muhamad-Rasyid-Aditya

