HOME CREDIT SCORECARD MODEL

Project Based-Internship at Home Credit Indonesia
Data Science

Suryani

PROBLEM STATEMENT

Saat ini, Home Credit Indonesia sedang menggunakan berbagai metode statistik dan Machine Learning untuk membuat prediksi skor kredit calon konsumennya. Prediksi ini dibutuhkan agar pinjaman yang diberikan tepat sasaran dan tetap sesuai dengan prinsip principal, maturity, dan repayment calendar.

Lalu, konsumen seperti apakah yang berhak mendapatkan layanan pinjaman dari Home Credit Indonesia?

PROBLEM SOLUTION

Berdasarkan permasalah yang ada pada Home Credit Indonesia, maka solusi yang ditawarkan adalah membuat model prediksi kredit konsumen berdasarkan metode statistik dan machine learning

DATASET OVERVIEW

- Dataset yang digunakan adalah application_train.csv dan application_test.csv
- Dataset memiliki 30,7511 baris dan 122 kolom
- Dataset dipisahkan oleh kode unik atau ID berupa SK_ID_CURR
- Dataset memiliki target (kelas yang akan diprediksi) yaitu 0 = Clients without Difficulties Payments, dan 1 = Clients with Difficulties Payment
- Terdapat 122 kolom yang berisi categorical value dan numerical value. Masing-masing kolom memberikan informasi tentang konsumen, misalkan jenis kelamin dan usia

PROJECT WORKFLOW (1)

1) Dataset Preparation

- Import library yang dibutuhkan
- Load dataset yang akan digunakan (application_train.csv)
- Memeriksa tipe data, gambaran data secara statistik, nilai-nilai yang kosong serta duplikat data

2) Assessing Dataset

- Memperbaiki missing value dengan imputasi modus pada data kategorikal, dan median pada data numerikal
- Menghapus kolom dengan missing value > 50%

3) Exploratory Data Analysis

- Mengkategorikan umur konsumen
- Melakukan visualisasi fitur menggunakan bar chart
- Menemukan insight bisnis

4) Data Preprocessing

- Merubah kategorikal fitur menjadi numerikal fitur
- Melakukan seleksi fitur menggunakan K-Nearest Neighbor (KNN)
- Oversampling dataset menggunakan SMOTE

PROJECT WORKFLOW (2)

5) Machine Learning Modelling

- Normalisasi dataset menggunakan StandardScaller
- Membuat pemodelan Machine Learning menggunakan Logistic Regression, Random Forest dan LightGBM

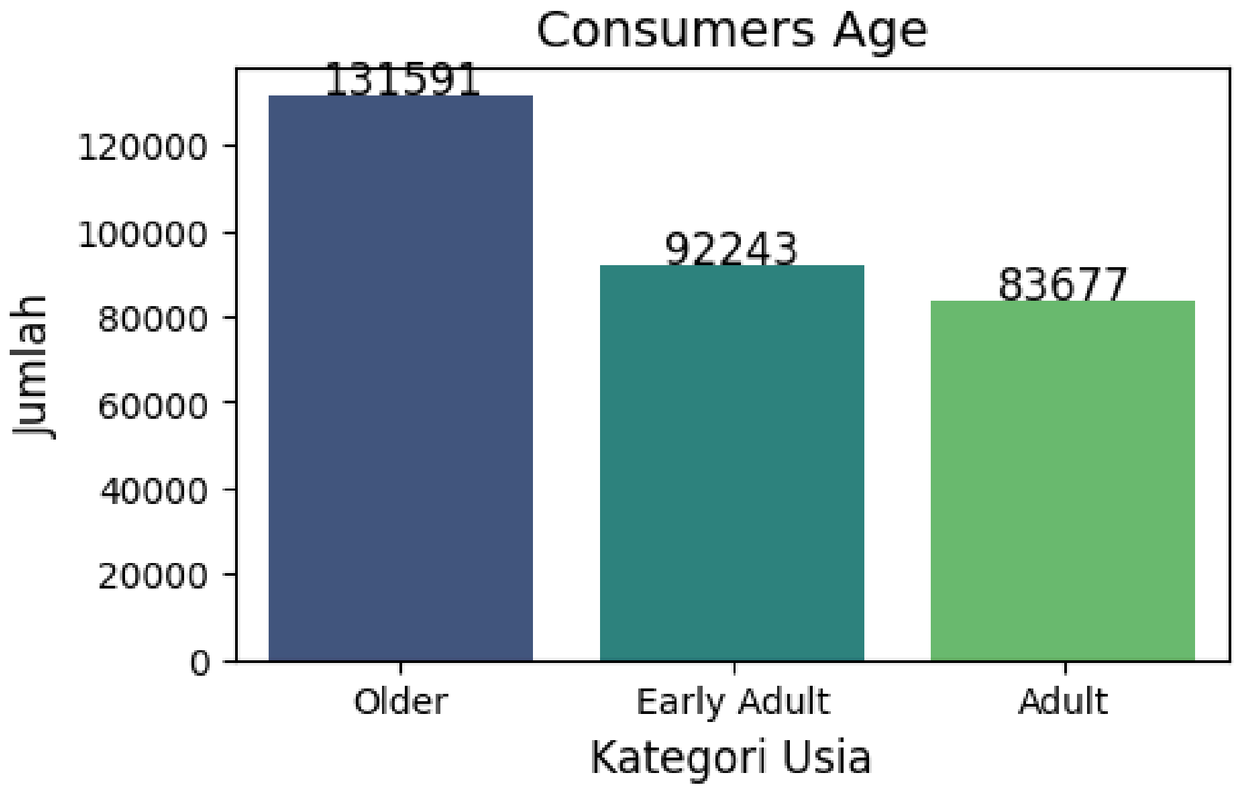
6) Model Evaluation

- Melakukan evaluasi model Machine Learning menggunakan Confusion Matrix dan nilai ROC
- Melakukan perbandingan untuk mendapatkan model dengan performa terbaik

7) Predict New Dataset

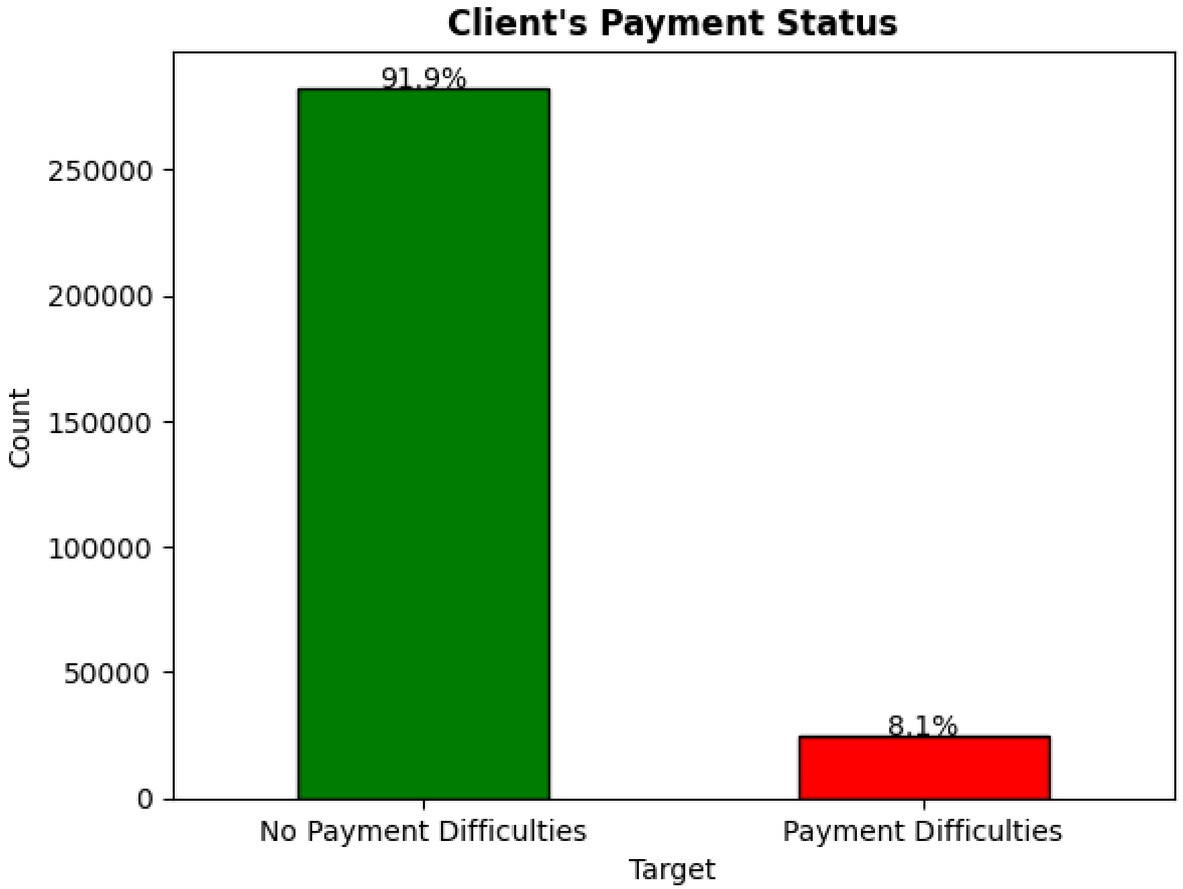
- Berdasarkan model terbaik, dilakukan prediksi terhadap data testing (application_test.csv)

EXPLORATORY DATA ANALYSIS (EDA)



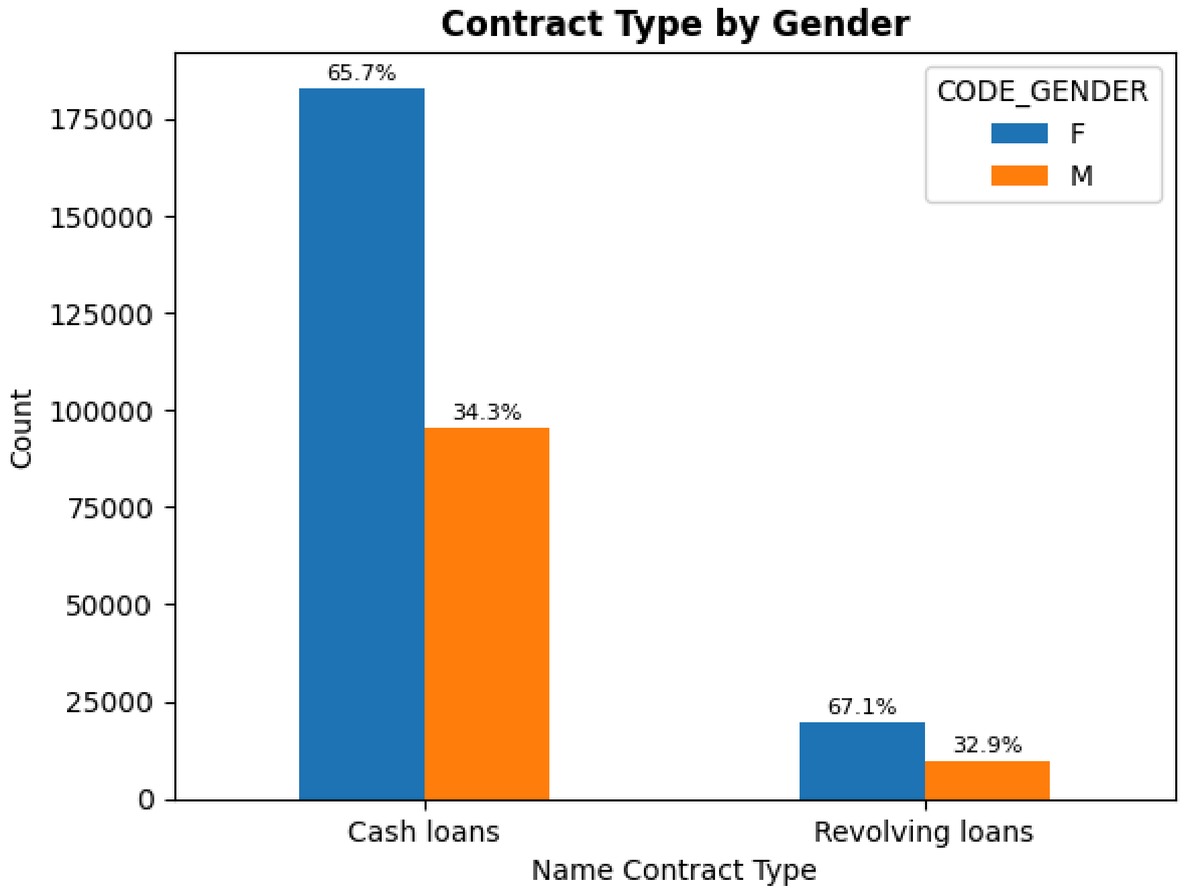
- Persentase usia konsumen Home Credit Indonesia didominasi oleh Older > 45 tahun
- Kategori Early adult <= 35 tahun, dan
- Kategori adult <= 45 tahun

EXPLORATORY DATA ANALYSIS (EDA)



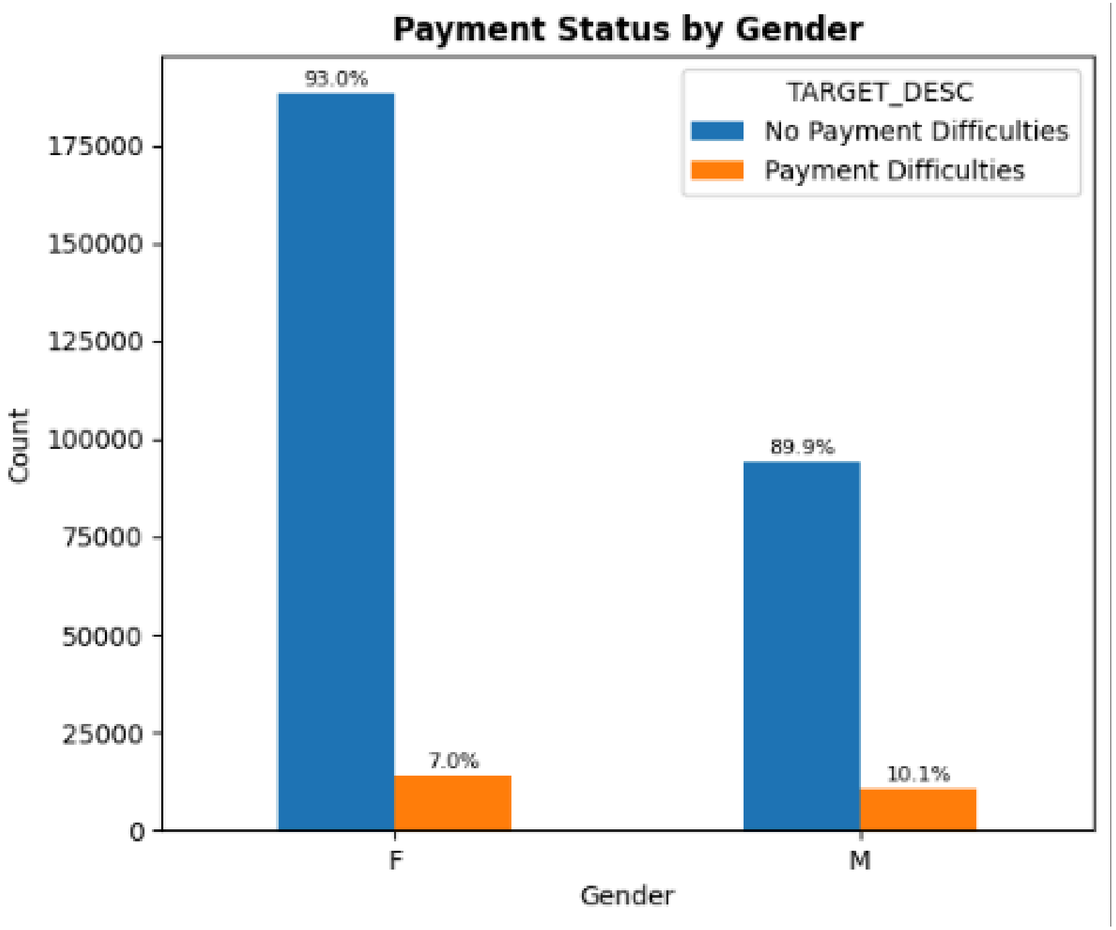
- Persentase konsumen dengan status 0 atau No Payment Difficulties mendominasi 91,9%
- Sedangkan konsumen dengan status 1 atau Payment with Difficulties hanya 8.1%

EXPLORATORY DATA ANALYSIS (EDA)



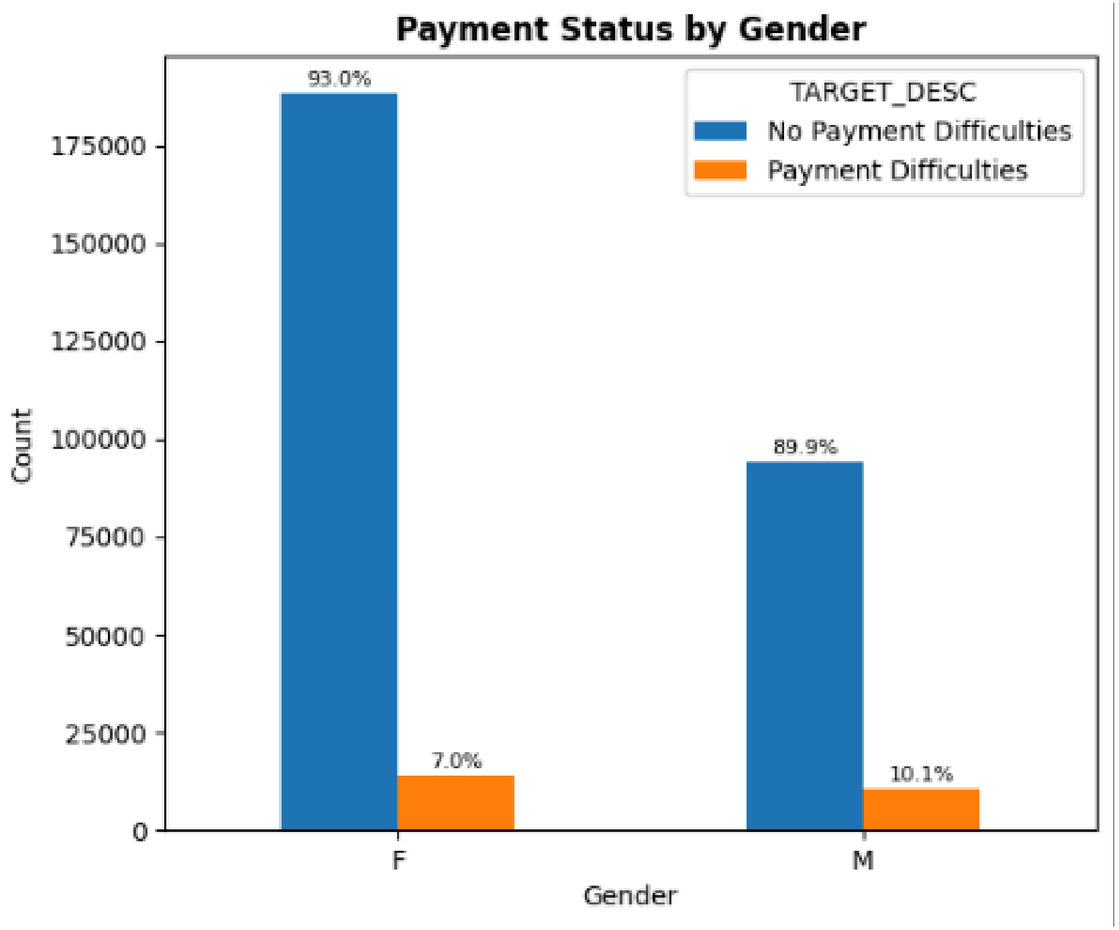
- Jenis contract yang ditawarkan kepada para client terbagi menjadi 2 jenis, yaitu cash loans dan revolving loans dimana pada kedua jenis didominasi oleh **konsumen wanita**, yaitu sebesar 65,7% pada cash loans, dan 67,1% pada revolving loans

EXPLORATORY DATA ANALYSIS (EDA)



- Berdasarkan gender, diketahui bahwa **konsumen wanita lebih disiplin** dalam pelunasan pinjaman daripada konsumen pria.
- Hanya **7% konsumen wanita** yang kesulitan dalam pelunasan, sedangkan **10.1% konsumen laki-laki** mengalami kesulitan yang sama

EXPLORATORY DATA ANALYSIS (EDA)



- Berdasarkan gender, diketahui bahwa **konsumen wanita lebih disiplin** dalam pelunasan pinjaman daripada konsumen pria.
- Hanya **7% konsumen wanita** yang kesulitan dalam pelunasan, sedangkan **10.1% konsumen laki-laki** mengalami kesulitan yang sama
- Berdasarkan statistik, konsumen dengan latar belakang pekerjaan sebagai **Realty Agents, HR Staff** dan **IT Staff** adalah konsumen urutan tiga terbawah dengan jumlah kurang dari 800 orang.

MACHINE LEARNING ANALYSIS

Model	Acuracy	ROC
Logistic Regression	81%	0,57
Random Forest	92%	0,67
LightGBM	93%	0,69

- Pemodelan Machine Learning dilakukan menggunakan tiga algoritma berbeda
- Dataset dibagi menjadi 70% data training, dan 30% data testing
- Performa terbaik dihasilkan oleh model LightGBM dengan **akurasi 93%** dan nilai **ROC 0,69**
- Nilai ROC yang semakin dekat kepada 1 menunjukkan performa model yang baik
- Model LightGBM akan digunakan untuk melakukan prediksi dataset baru (application_test.csv) untuk mengetahui score catd komsumen Home Credit Indonesia

BUSINESS RECOMENDATION

- Berdasarkan analisis statistik, terdapat perbedaan yang sangat signifikan antara kemampuan konsumen wanita dan pria dalam melakukan pelunasan pinjaman. Home Credit Indonesia dapat mempertimbangkan untuk memberikan pinjaman kepada konsumen pria kedepannya. Pertimbangan ini dapat didasarkan pada aspek lain, misalkan status pekerjaannya dan kepemilikan aset pribadi.

- Minoritas konsumen Home Credit Indonesia adalah pekerja Realty Agents, IT Staff, dan HR Staff. Home Credit Indonesia dapat mempertimbangkan kampanye atau promosi khusus untuk meningkatkan jumlah konsumen dari latar belakang pekerjaan tersebut

- Konsumen yang memiliki kemampuan baik dalam menyelesaikan pinjaman dapat dipertimbangkan untuk diberikan bonus atau penawaran khusus sebagai bentuk loyalitas perusahaan.

THANK YOU

https://github.com/suryani622/HCI-Scorecard-Prediction