Rakamin Academy X HOME CREDIT

Home Credit Loan Classification

Home Credit Indonesia - Data Scientist

Presented by
Bagas Ghulam Maulana

Rakamin Academy

Bagas Ghulam Maulana

Bachelor of Mathematics

I am a Bachelor of Mathematics, Universitas Pendidikan Indonesia with an interest into data analysis and data science. Proficient in using Microsoft Excel, SQL, Python, and Google Looker Studio. Currently learning and delving into data analysis and data science by working on projects related to these fields.

Kota Bekasi, Jawa Barat

bagasgm4@gmail.com

www.linkedin.com/in/bagasghulam

Courses and Certification

Data Analysis with Python (IBM) | Coursera	January, 2024
Crash Course on Python (Google) | Coursera	February, 2024
Excel 2019 Associate | Microsoft	July, 2024
Tech Academy - Learn Data Analytics & Software Development With AI | RevoU	Feb-July,2024
Python Fundamental for Data Science | DQLab	August, 2025
Machine Learning with Python for Beginner | DQLab	August, 2025

List of Content

- About Company
- Problem Statement & Objective
- Load Dataset & Data Preprocessing
- Exploratory Data Analysis
- Modelling & Evaluation
- Insights & Recommendations

About Company

PT Home Credit Indonesia atau yang lebih dikenal dengan Home Credit merupakan perusahaan pembiayaan multiguna multinasional. Perusahaan ini membangun layanan pembiayaan di toko (pembiayaan non-tunai langsung di tempat) untuk konsumen yang ingin membeli produk-produk seperti alat rumah tangga, alat-alat elektronik, handphone, dan furniture. Perusahaan ini juga membangun layanan pembiayaan berbasis teknologi. Didirikan pada tahun 2013 di Jakarta, saat ini Home Credit telah menjangkau lebih dari 19.000 titik distribusi yang tersebar di 144 kota di Indonesia. Hingga bulan Maret 2019, perusahaan ini telah melayani 3,4 juta pelanggan secara online maupun offline.

Problem Statement

Memprediksi risiko gagal bayar dari calon nasabah berdasarkan data aplikasi credit, untuk membantu Home Credit Indonesia dalam mengambil keputusan yang lebih akurat dan efisien.

Objective

- Membangun model machine learning untuk memprediksi status pembayaran nasabah.
- Menentukan fitur-fitur utama yang memengaruhi risiko gagal bayar.
- Menyaring nasabah dengan potensi gagal bayar tinggi untuk mengurangi risiko kerugian.
- Dikarenakan data yang bersifat imbalanced, metrik yang digunakan fokus pada kemampuan model mendeteksi kelas minoritas: ROC-AUC (utama), Recall (seberapa baik model menangkap nasabah risiko), F1-score (keseimbangan antara precision-recall), dan accuracy (pelengkap, bukan utama).

Load Dataset & Data Preprocessing

- ## Load Dataset

- Import dan load dataset dari google drive. Dataset utama yang digunakan adalah file application_train.csv.
- Dataset ini terdiri dari 307511 baris dan 122 kolom.

- ## Data Preprocessing

- Cek dataset, apakah terdapat nilai kosong pada setiap kolom nya.
- Nilai kosong yang ada akan diproses dengan cara menghapus kolom (jika nilai kosong > 50%) dan mengisi nilai kosong dengan median dan modus (jika nilai kosong =< 50%).

```
df = application_train.copy()
df
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR
0	100002	1	Cash loans	M	N
1	100003	0	Cash loans	F	N
2	100004	0	Revolving loans	M	Y
3	100006	0	Cash loans	F	N
4	100007	0	Cash loans	M	N
...	...	...	...	...	...
307506	456251	0	Cash loans	M	N
307507	456252	0	Cash loans	F	N
307508	456253	0	Cash loans	F	N
307509	456254	1	Cash loans	F	N
307510	456255	0	Cash loans	F	N

307511 rows × 122 columns

```
print("Total missing setelah imputasi:", df.isnull().sum().sum())
```

```
Total missing setelah imputasi: 0
```

Load Dataset & Data Preprocessing

- Setelah dilakukan pengecekan, terdapat 67 kolom yang memiliki nilai kosong, beberapa di antaranya memiliki nilai >50%.
- Untuk kolom yang memiliki nilai kosong >50%, kolom tersebut akan dihapus sedangkan yang <=50%, nilai kosong akan diisi dengan median dan modus.
- Mengisi nilai kosong dengan median lebih baik dari mean karena median lebih tahan outlier (untuk tipe data numerik)
- Mengisi nilai kosong dengan modus (untuk tipe data kategorikal)

missing_percent

	0
COMMONAREA_MEDI	69.872297
COMMONAREA_MODE	69.872297
COMMONAREA_AVG	69.872297
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_MEDI	69.432963
...	...
EXT_SOURCE_2	0.214626
AMT_GOODS_PRICE	0.090403
AMT_ANNUITY	0.003902
CNT_FAM_MEMBERS	0.000650
DAYS_LAST_PHONE_CHANGE	0.000325

67 rows × 1 columns

```
num_cols = df.select_dtypes(include=['float64', 'int64']).columns
for col in num_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].median(), inplace=True)

cat_cols = df.select_dtypes(include='object').columns
for col in cat_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].mode()[0], inplace=True)
```

Exploratory Data Analysis

Walaupun sudah melalui preprocessing, kolom yang ada masih terlalu banyak sehingga akan dilakukan pemodelan awal untuk screening fitur penting (melalui Random Forest with Feature Importance).

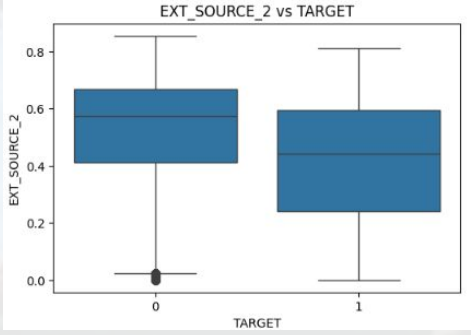
Dari grafik 2 fitur teratas yaitu ET_SOURCE_2 dan EXT_SOURCE_3, terlihat bahwa:

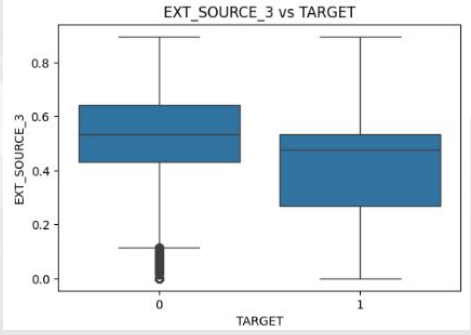
- Rata-rata skor eksternal nasabah yang tidak gagal bayar (TARGET = 0) lebih tinggi dibandingkan dengan yang gagal bayar(TARGET = 1)
- Semakin tinggi nilai skor eksternal, maka semakin rendah risiko gagal bayar
- Distribusi data untuk EXT_SOURCE_3 sedikit lebih menyebar dibandingkan EXT_SOURCE_2. Namun, kedua fitur tersebut menunjukkan tren serupa

```
# Ambil fitur dengan importance ≥ 0.01
important_features = importances_df[importance

print(f"Jumlah fitur dengan importance ≥ 0.01:
```

```
Jumlah fitur dengan importance ≥ 0.01: 22
```





Modelling & Evaluation

- Ada 5 model yang diimplementasikan yaitu Logistic Regression, Random Forest, XGBoost, CatBoost, dan Decision Tree. Metrik utama yang digunakan adalah ROC-AUC disebabkan oleh data yang imbalanced.

- Berdasarkan hasil evaluasi, CatBoost menjadi model dengan performa terbaik dengan nilai ROC-AUC sebesar 0.741. Meskipun model seperti Logistic Regression dan Decision Tree memiliki akurasi tinggi, namun gagal menangkap kelas minoritas (TARGET = 1) secara efektif, yang terlihat dari nilai recall dan F1-score yang rendah.

- Hal sebelumnya menunjukkan bahwa dalam kasus prediksi risiko gagal bayar, akurasi saja tidak cukup, dan metrik seperti ROC-AUC dan recall lebih penting untuk mengevaluasi performa model secara adil terhadap kedua kelas.

```
results_df = pd.DataFrame(model_results)
results_df.sort_values('ROC_AUC', ascending=False)
```

	Model	Accuracy	Recall_1	Precision_1	F1_1	ROC_AUC
3	CatBoost	0.919549	0.022429	0.502262	0.042940	0.741098
2	XGBoost	0.919142	0.022429	0.451220	0.042733	0.730642
1	Random Forest	0.919662	0.006668	0.568966	0.013182	0.700113
0	Logistic Regression	0.919516	0.000000	0.000000	0.000000	0.631893
4	Decision Tree	0.850219	0.163467	0.137562	0.149400	0.536892

Insights & Recommendations

- **Insights**
- Model terbaik adalah CatBoost dengan nilai ROC-AUC sebesar 0.741 dengan precision 50% dan recall 2%.
- Recall rendah artinya model belum optimal dalam mengenali nasabah gagal bayar (risiko kebocoran kredit).
- Precision 50% artinya setengah dari prediksi risiko benar (cukup efektif untuk screening awal).
- Accuracy >91% dapat menyesatkan karena dominasi nasabah lancar. Sebaiknya fokus pada deteksi risiko bukan akurasi total.
- **Recommendations**
- Gunakan CatBoost sebagai model awal evaluasi risiko.
- Fokus pada fitur-fitur penting yaitu EXT_SORUCE_2, DAYS_BIRTH, DAYS_EMPLOYED.
- Sesuaikan produk dengan segmen risiko, risiko tinggi (tenor pendek dan cicilan ringan) sedangkan risiko rendah (bunga kompetitif dan fleksibel).

Rakamin Academy

HOME CREDIT

Thank You

Repository Link

Link Github : Github