PROJECT BASED INTERNSHIP

HOME CREDIT - DS

Gaung Taqwa Indraswara

BUSINESS UNDERSTANDING

- Home Credit saat ini sedang menggunakan berbagai macam metode statistik dan Machine Learning untuk membuat prediksi skor kredit. Sekarang, kami meminta anda untuk membuka potensi maksimal dari data kami. Dengan melakukannya, kita dapat memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman, dan pinjaman data diberikan dengan principal, maturity, dan repayment calendar yang akan memotivasi pelanggan untuk sukses. Evaluasi akan dilakukan dengan mengecek seberapa dalam pemahaman analisa yang anda kerjakan.

DATA UNDERSTANDING

Data telah dipersiapkan oleh Home Credit

Terdapat 7 sumber data yang berbeda pada data warehouse milik Home Credit, namun hanya satu yang digunakan untuk kasus ini, yaitu:

* application_train/application_test: Data utama pelatihan dan pengujian yang berisi informasi tentang setiap pengajuan pinjaman di Home Credit. Setiap pengajuan pinjaman memiliki satu baris dan diidentifikasi oleh fitur SK_ID_CURR. Pada data pelatihan terdapat kolom TARGET yang menunjukkan apakah pinjaman telah dilunasi (0) atau tidak dilunasi (1).

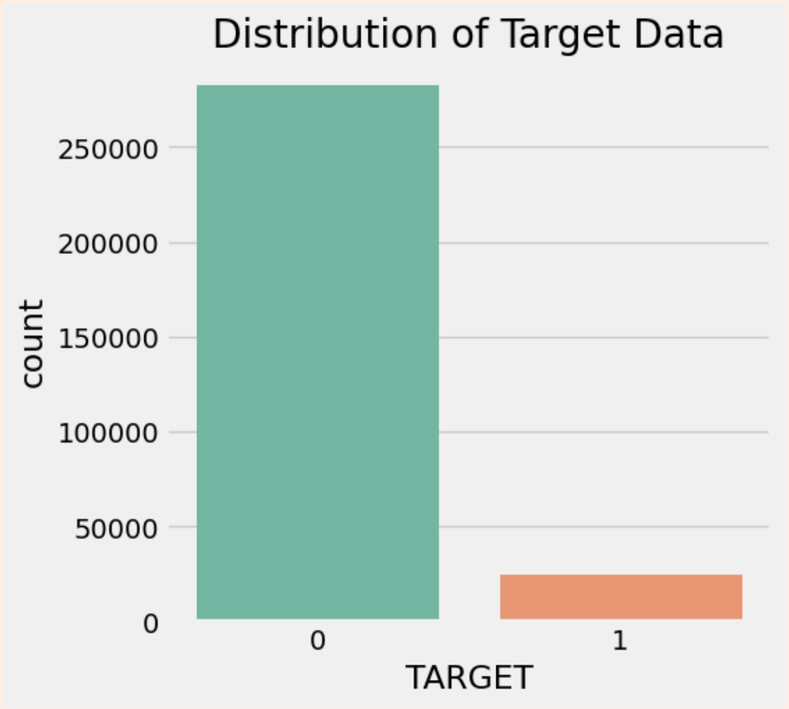
GOAL DAN METRICS

- Goal : Membuat sebuah model Machine Learning yang bisa memprediksi apakah user yang mengajukan kredit dapat membayar tepat waktu atau akan telat/bermasalah

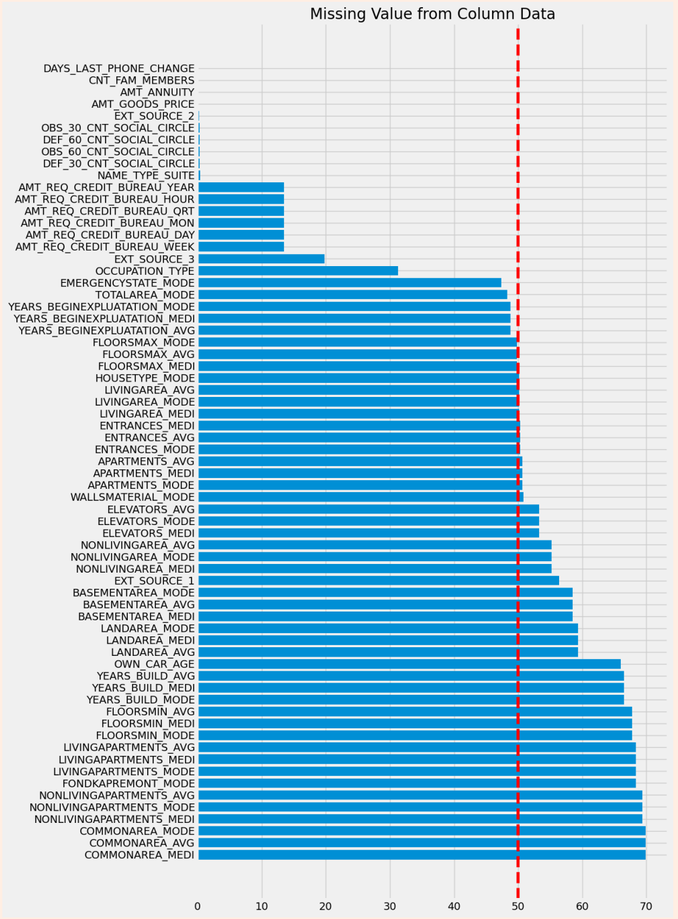
- Metrics : Yang digunakan untuk evaluasi kinerja model yaitu ROC-AUC (karena datanya imbalanced)

- Minimal score metrics yang baik yaitu 60%

INFORMASI AWAL PADA DATA :



Target variabel sangat imbalance



Terdapat banyak missing value pada dataset

```
float64    65
int64      41
object     16
dtype: int64
```

Jenis tipe data pada setiap kolom dataset

PROSES DATA CLEANSING

Proses Data Cleansing yang dilakukan yaitu:

1. Melakukan Label Encoding untuk kolom dataset yang hanya memiliki 2 kategori
2. Melakukan One-Hot Encoding untuk kolom dataset yang lebih dari 2 kategori
3. Melakukan imputasi pada missing value dengan menggunakan nilai median

PENGGALIAN INSIGHT PADA DATA

Beberapa hasil pencarian informasi pada data yaitu:

1. Terdapat indikasi outlier pada kolom ‘DAYS_EMPLOYED’, sehingga nilai yang outlier diganti dengan nilai bukan angka dan pembuatan kolom baru sebagai penanda nilai tersebut merupakan anomaly atau bukan dengan alasan data yang memiliki nilai tersebut memiliki tingkat gagal bayar lebih rendah.
2. Korelasi setiap kolom terhadap variabel ‘TARGET’ menunjukkan nilai yang rendah
3. Hasil eksplorasi terkait pengaruh usia menyatakan bahwa klien dengan usia muda cenderung tidak membayar kembali pinjaman dengan tingkat kegagalan membayar kembali di atas 10% untuk 3 kelompok usia termuda

FEATURE ENGINEERING

Terdapat beberapa fitur baru yang ditambahkan pada data untuk melihat pengaruh penambahan fitur ini terhadap hasil pemodelan, diantaranya:

- CREDIT_INCOME_PERCENT: persentase jumlah kredit relatif terhadap pendapatan klien
- ANNUITY_INCOME_PERCENT: persentase anuitas pinjaman relatif terhadap pendapatan klien
- CREDIT_TERM: jangka waktu pembayaran dalam bulan (karena anuitas adalah jumlah yang harus dibayarkan setiap bulan)
- DAYS_EMPLOYED_PERCENT: persentase hari kerja relatif terhadap usia klien

HASIL EVALUASI PEMODELAN

Pemodelan dilakukan dengan 3 skenario, yaitu:

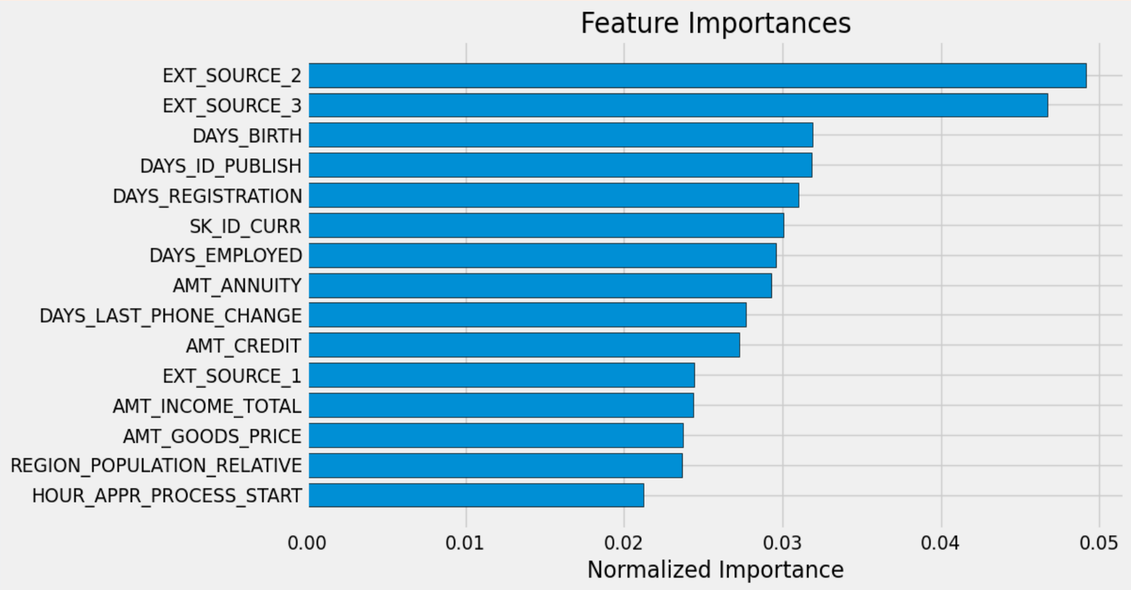
1. Menggunakan Algoritma Logistic Regression
2. Menggunakan Algoritma Random Forest
3. Menggunakan Algortima terbaik dari 2 pilihan dengan penambahan fitur yang telah ditambahkan sebelumnya

Hasil Pemodelan dievaluasi dengan menggunakan ROC-AUC, hasilnya menyatakan bahwa:

1. Algoritma Logistic Regression : 0.69
2. Algoritma Random Forest : 0.71
3. Random Forest dengan penambahan fitur pada data : 0.70

Hasil terbaik diperoleh dengan hanya menggunakan algoritma Random Forest

REKOMENDASI BISNIS



Fitur EXT_SOURCE memiliki pengaruh tertinggi, menunjukkan bahwa informasi dari sumber eksternal sangat berguna untuk prediksi kemampuan pembayaran. Perusahaan dapat meningkatkan kerja sama dengan penyedia data eksternal untuk memperkaya data mereka atau memperbarui data ini secara rutin.

Usia yang lebih tua atau lebih muda dapat berhubungan dengan tingkat risiko yang berbeda. Perusahaan dapat mempertimbangkan kebijakan pinjaman yang berbeda berdasarkan kelompok usia, seperti menawarkan bunga lebih rendah untuk usia tertentu yang dianggap kurang berisiko.

LINK GITHUB REPOSITORY

https://github.com/arawsardni/Final-Task---Home-Credit-Scorecard-Model