STAGE 3

Conexus Group

I. Modeling

1.1. Split Data Train & Test

Langkah pertama dalam modeling adalah membagi dataset menjadi data train dan data test untuk mengevaluasi kinerja model. Pembagian data dilakukan dengan rasio 80:20, di mana 80% data digunakan untuk melatih model, sementara 20% digunakan untuk pengujian. Teknik ini memastikan bahwa model tidak hanya menghafal data melainkan juga mampu menggeneralisasi data yang belum pernah dilihat sebelumnya.

1.2. Random Forest

Pada dataset Home Credit memiliki tujuan utama untuk memprediksi apakah seseorang beresiko mengalami default credit atau tidak, sehingga modeling yang dilakukan pada dataset ini termasuk machine learning tipe supervised learning yang klasifikasi karena terdapat target. Target 1 untuk menunjukkan pemohon yang berisiko tinggi mengalami gagal bayar, sedangkan 0 untuk menunjukkan pemohon yang dianggap aman atau tidak akan mengalami gagal bayar.

Beberapa algoritma telah dicoba untuk modeling, yaitu XGBoost, Logistic Regression, Random Forest, LightGBM, dan Support Vector Machines (SVM) - SVC. Berdasarkan hasil percobaan tersebut, hasil yang paling baik pada nilai model evaluasinya adalah metode Random Forest.

1.3. Model Evaluation

1.3.1. Metrics

Prioritas Pada Identifikasi Risiko (False Negatives): Dalam konteks Home Credit, recall menjadi penting karena berfokus pada pengurangan false negatives. False negatives di sini berarti individu berisiko yang diprediksi aman oleh model. Jika kita gagal mengidentifikasi pemohon yang sebenarnya berisiko tinggi, risiko bagi perusahaan

meningkat. Dengan memaksimalkan recall, model berfokus pada mendeteksi sebanyak mungkin individu berisiko, sehingga meminimalisir potensi kerugian finansial.

Mengurangi Kerugian Pada Kelayakan Kredit: Dataset Home Credit berurusan dengan kelayakan kredit, di mana kesalahan mengidentifikasi calon debitur berisiko rendah dapat menyebabkan kerugian yang signifikan. Dengan memilih recall sebagai fokus, model memastikan bahwa semua calon yang berpotensi gagal bayar teridentifikasi dengan baik, sehingga mengurangi potensi kerugian dan menjaga stabilitas portofolio kredit.

Keselarasan dengan Tujuan Bisnis: Dalam bisnis kredit, biasanya lebih penting untuk menghindari memberikan pinjaman kepada pemohon yang berisiko tinggi daripada sekadar meningkatkan akurasi umum model. Dengan berfokus pada recall, model memprioritaskan deteksi calon debitur berisiko, selaras dengan tujuan bisnis yang mengutamakan pengurangan risiko.

Kompromi dengan Precision yang Bisa Diterima: Mengingat konteks yang berisiko tinggi, tingkat precision yang lebih rendah bisa ditoleransi, karena dampak dari kesalahan mengidentifikasi pemohon yang aman sebagai berisiko rendah (false negatives) lebih besar dibanding kesalahan sebaliknya. Sehingga, berfokus pada recall menjadi strategi yang lebih sesuai.

Pengaruh pada Kebijakan Peminjaman: Fokus pada recall juga berdampak pada kebijakan peminjaman. Dengan mengutamakan recall, perusahaan dapat lebih berhati-hati dalam memutuskan kelayakan kredit, yang sejalan dengan pendekatan manajemen risiko ketat yang sering kali digunakan dalam industri keuangan.

1.3.2. Fitting

Dari semua model yang telah dicoba, algoritma Random Forest menunjukkan performa recall yang baik pada train set, dengan nilai recall sebesar 1,00. Hal ini menandakan bahwa model mampu mendeteksi sebagian besar kasus yang relevan dalam data pelatihan, sehingga risiko false negatives (kesalahan dalam melewatkan individu berisiko) dapat ditekan.

Namun, pada test set, nilai recall menurun menjadi 0,92. Meski penurunan ini tidak terlalu drastis, adanya selisih performa antara train set dan test set menunjukkan bahwa model mengalami overfitting terhadap data pelatihan. Hasil cross-validation pun menunjukkan tren serupa, di mana recall pada train set mencapai 1,00, sementara pada test set menurun menjadi 0,9. Perbedaan ini mengindikasikan bahwa model masih perlu dituning agar lebih mampu menggeneralisasi pada data baru tanpa kehilangan performa recall.

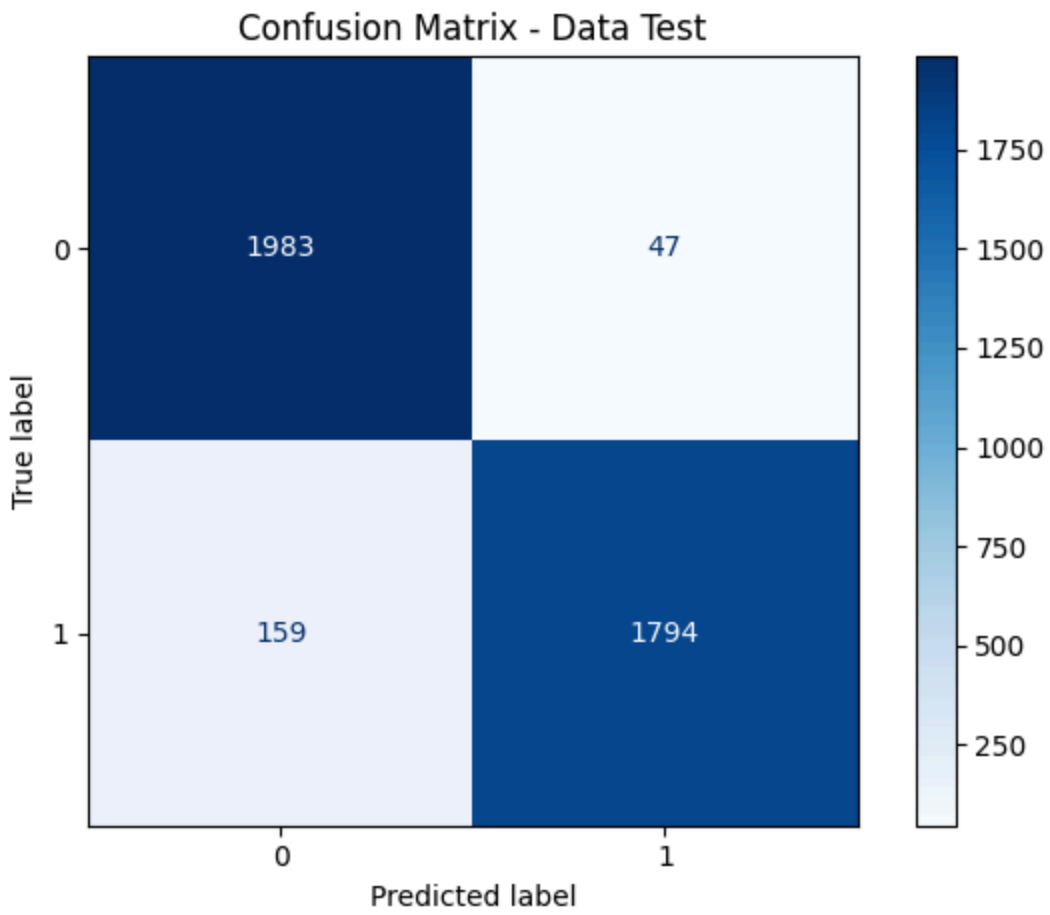
Dengan demikian, algoritma Random Forest sudah menunjukkan fit yang baik secara keseluruhan, tetapi model tersebut mengalami overfitting dan harus dilakukan hyperparameter tuning.

```
Accuracy (Train Set)    : 1.00
Accuracy (Test Set)     : 0.95
Precision (Train Set)   : 1.00
Precision (Test Set)    : 0.97
Recall    (Train Set)   : 1.00
Recall    (Test Set)    : 0.92
F1-Score (Train Set)    : 1.00
F1-Score (Test Set)     : 0.95
ROC-AUC (Train-Proba)   : 1.00
ROC-AUC (Test-Proba)    : 0.99
Recall (Crossval Train) : 1.0
Recall (Crossval Test)  : 0.9
Default rate for train set:
Predicted default count: 8003
Total count: 15929
Predicted default rate: 50.24%
Default rate for test set:
Predicted default count: 1841
Total count: 3983
Predicted default rate: 46.22%
```

1.3.3. Default Rate

Menggunakan algoritma Random Forest yang telah dilakukan hyperparameter tuning, dilakukan perhitungan default rate untuk mencari perbandingan antara jumlah orang yang gagal bayar dengan jumlah seluruh orang. Pada train set, hasil prediksi yang default adalah sebesar 7746 dari keseluruhan data sebesar 15929 yang menghasilkan persentase default rate sebesar 48.83%. Pada test set, hasil prediksi yang default adalah sebesar 1863 dari keseluruhan data sebesar 3983 yang menghasilkan persentase default rate sebesar 46.77%.

1.3.4. Confusion Matrix



1. **Pengertian dari Masing-masing Komponen Confusion Matrix**
    a. True Negative (TN): Jumlah kasus di mana model memprediksi negatif (tidak default) dan benar-benar negatif di data aslinya (tidak terjadi default). Dalam hasil ini, ada 1983 kasus.
    b. False Positive (FP): Jumlah kasus di mana model memprediksi positif (default) padahal sebenarnya negatif di data aslinya. Ini menunjukkan model memberikan alarm palsu. Terdapat 47 kasus di mana model salah memprediksi default padahal seharusnya tidak default.
    c. False Negative (FN): Jumlah kasus di mana model memprediksi negatif (tidak default), padahal sebenarnya positif di data aslinya (terjadi default). Ini adalah kesalahan di mana model gagal mendeteksi default yang seharusnya ada. Ada 159 kasus seperti ini.
    d. True Positive (TP): Jumlah kasus dimana model memprediksi positif (default) dan benar-benar positif di data aslinya (terjadi default). Dalam hasil ini, ada 1794 kasus.

2. **Evaluasi Berdasarkan Confusion Matrix**

    Dengan menggunakan nilai-nilai ini, dapat dihitung beberapa metrics evaluasi:

Accuracy: Persentase prediksi yang benar dari keseluruhan data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1794 + 1983}{1794 + 1983 + 47 + 159} \approx 95.14\%$$

Precision: Persentase prediksi positif yang benar-benar positif (default).

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{1794}{1794 + 47} \approx 97.44\%$$

Recall: Persentase kasus positif yang berhasil terdeteksi oleh model.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1794}{1794 + 159} \approx 91.85\%$$

F1-Score: Kombinasi precision dan recall yang seimbang.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \approx 94.55\%$$

3. **Interpretasi**

   False Positive (47 kasus): Ini adalah prediksi yang salah di mana model mengira pelanggan akan default padahal sebenarnya tidak. Meskipun nilainya relatif rendah, ini dapat menyebabkan pelanggan yang sebenarnya baik tidak mendapatkan pinjaman yang diinginkan.

   False Negative (159 kasus): Ini adalah prediksi yang salah di mana model tidak mendeteksi pelanggan yang sebenarnya akan default. Nilai ini lebih tinggi dari FP, dan berpotensi menyebabkan kerugian pada perusahaan karena pelanggan yang seharusnya dianggap berisiko malah dianggap aman.

4. **Kesimpulan**

   Model ini memiliki **accuracy yang tinggi (95.14%)** dan **precision yang sangat baik (97.44%)**, menunjukkan bahwa model ini sangat akurat dalam mengidentifikasi pelanggan yang akan default. **Recall sebesar 91.85%** menunjukkan bahwa masih ada beberapa pelanggan default yang tidak terdeteksi oleh model. Tetapi, nilai hasil recall tersebut merupakan nilai yang paling tinggi dibandingkan dengan nilai hasil dengan metode lain seperti XGBoost (91%), SVM-SVC (19%), LightGBM (90,6%), dan Regression (53%).

1.4. Hyperparameter Tuning

Berdasarkan hasil modelling pada algoritma Random Forest di atas tanpa ada hyperparameter tuning, ditemukan bahwa model tersebut overfitting. Maka, harus dilakukan hyperparameter tuning untuk mengurangi overfitting. Setelah dilakukan hyperparameter tuning, ditemukan parameter terbaik adalah sebagai berikut (criterion='entropy', max_depth=98, min_samples_leaf=6, min_samples_split=12, n_estimators=53, random_state=42). Dan berikut hasil dari hyperparameter tuning :

```
Accuracy (Train Set)    : 0.98
Accuracy (Test Set)     : 0.94
Precision (Train Set)   : 1.00
Precision (Test Set)    : 0.96
Recall    (Train Set)   : 0.97
Recall    (Test Set)    : 0.91
F1-Score (Train Set)    : 0.98
F1-Score (Test Set)     : 0.93
ROC-AUC (Train-Proba)   : 1.00
ROC-AUC (Test-Proba)    : 0.98
Recall (Crossval Train) : 0.96
Recall (Crossval Test)  : 0.89
Default rate for train set:
Predicted default count: 7746
Total count: 15929
Predicted default rate: 48.63%
Default rate for test set:
Predicted default count: 1863
Total count: 3983
Predicted default rate: 46.77%
```

Algoritma Random Forest yang telah dilakukan hyperparameter tuning menunjukkan performa recall yang baik pada train set, dengan nilai recall sebesar 0,97. Hal ini menandakan bahwa model mampu mendeteksi sebagian besar kasus yang relevan dalam data pelatihan, sehingga risiko false negatives (kesalahan dalam melewatkan individu berisiko) dapat ditekan.

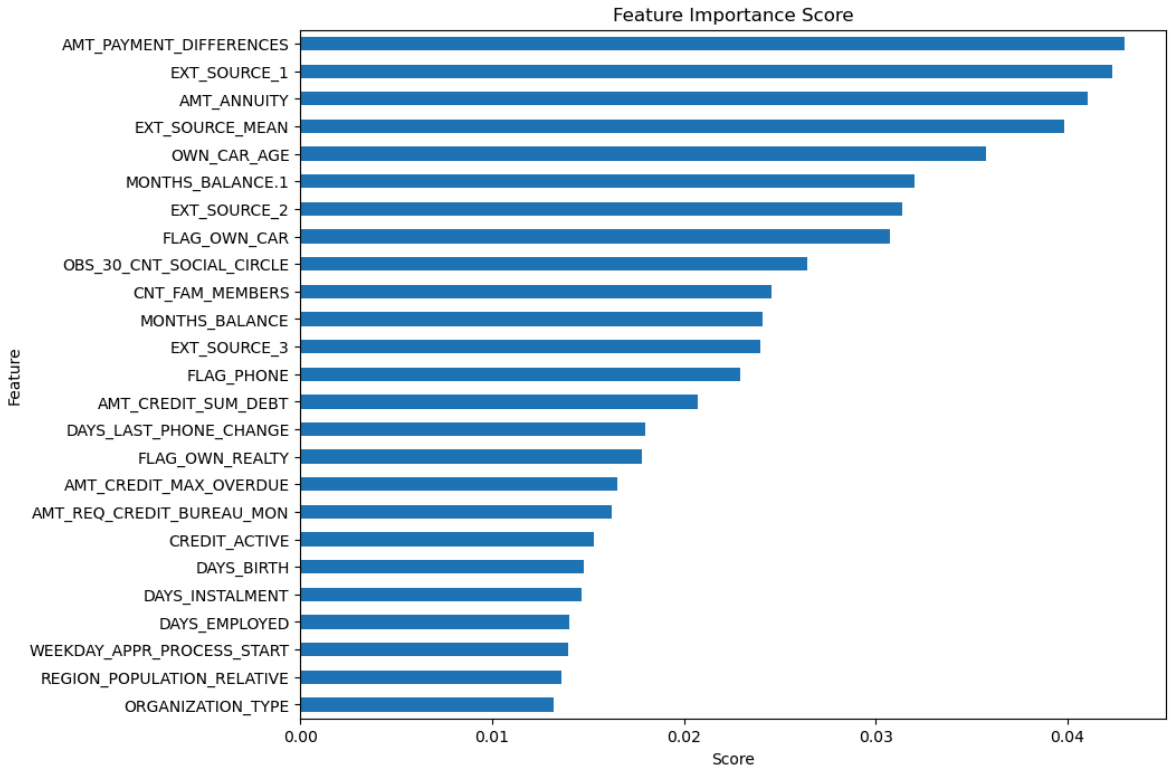
Namun, pada test set, nilai recall menurun menjadi 0,91. Meski penurunan ini tidak terlalu drastis, adanya selisih performa antara train set dan test set menunjukkan bahwa model sedikit mengalami overfitting terhadap data pelatihan. Hasil cross-validation pun menunjukkan tren serupa, di mana recall pada train set mencapai 0,96, sementara pada test set menurun menjadi 0,89. Perbedaan ini mengindikasikan bahwa model masih perlu ditingkatkan agar lebih mampu menggeneralisasi pada data baru tanpa kehilangan performa recall.

Fokus pada recall menjadi penting dalam konteks proyek ini karena model dituntut untuk mendeteksi sebanyak mungkin kasus berisiko, sehingga mengurangi potensi kerugian akibat false negatives. Dengan demikian, meskipun algoritma Random Forest sudah menunjukkan fit

yang baik secara keseluruhan, penyetelan tambahan mungkin diperlukan untuk meningkatkan recall pada test set dan mengurangi potensi overfitting.

II. Feature Importance

2.1. Interpretasi



1. **AMT_PAYMENT_DIFFERENCES**: Fitur ini menunjukkan perbedaan antara jumlah yang diminta dan jumlah yang dibayar. Fitur ini memiliki kontribusi tertinggi dalam model, mengindikasikan bahwa gap pembayaran mungkin penting untuk memprediksi kelayakan kredit.
2. **EXT_SOURCE_1, EXT_SOURCE_2, dan EXT_SOURCE_3**: Variabel eksternal ini sering berasal dari pihak ketiga yang menyediakan skor kredit atau informasi terkait risiko pelanggan. Peringkat ini memiliki pengaruh kuat terhadap prediksi, menandakan pentingnya data eksternal dalam mengukur risiko kredit.
3. **AMT_ANNUITY**: Jumlah anuitas atau cicilan yang harus dibayar oleh pelanggan setiap periode juga merupakan faktor penting. Hal ini menggambarkan kemampuan pelanggan dalam menangani pembayaran berkala.
4. **OWN_CAR_AGE**: Usia mobil yang dimiliki pelanggan menunjukkan status ekonomi dan kemampuan finansial mereka, yang juga penting dalam evaluasi risiko kredit.
5. **MONTHS_BALANCE dan MONTHS_BALANCE_1**: Durasi historis pelanggan dengan perusahaan atau saldo bulan sebelumnya menunjukkan stabilitas finansial.

2.2. Business Insight

1. **AMT_PAYMENT_DIFFERENCES** yang tinggi mungkin menandakan ketidaksesuaian antara jumlah yang diminta dan kemampuan membayar pelanggan. Pelanggan dengan gap besar antara permintaan dan pembayaran mungkin berisiko lebih tinggi gagal bayar.
2. **EXT_SOURCE_1, EXT_SOURCE_2, dan EXT_SOURCE_3** memberikan wawasan tambahan dari pihak ketiga. Jika skor dari sumber eksternal rendah, kemungkinan besar pelanggan memiliki risiko lebih tinggi.
3. **AMT_ANNUITY** yang tinggi dibandingkan dengan pendapatan pelanggan dapat menunjukkan ketidakmampuan untuk membayar cicilan, sehingga berisiko terhadap kegagalan bayar.
4. **OWN_CAR_AGE** yang lebih tua mungkin menunjukkan aset yang berkurang nilainya, yang bisa dihubungkan dengan risiko yang lebih tinggi jika tidak diiringi dengan kepemilikan aset lain.
5. **MONTHS_BALANCE** menunjukkan riwayat kredit pelanggan dalam beberapa bulan terakhir. Riwayat yang stabil menunjukkan kelayakan kredit yang lebih baik.

2.3. Actionables

1. **Tingkatkan Penilaian Berdasarkan AMT_PAYMENT_DIFFERENCES:**
   Lakukan analisis lebih dalam pada pelanggan dengan nilai AMT_PAYMENT_DIFFERENCES yang tinggi. Penyesuaian jumlah kredit yang ditawarkan atau peninjauan ulang persyaratan pinjaman dapat dilakukan untuk pelanggan dengan gap besar antara permintaan dan pembayaran.
2. **Gunakan Data Eksternal dalam Pengambilan Keputusan:**
   Data dari EXT_SOURCE sangat berguna dalam memberikan gambaran tambahan terkait risiko pelanggan. Kembangkan kolaborasi lebih lanjut dengan lembaga pihak ketiga untuk memperbaharui data risiko secara berkala.
3. **Pantau Kemampuan Pembayaran Melalui AMT_ANNUITY:**
   Bagi pelanggan yang memiliki AMT_ANNUITY tinggi, pertimbangkan untuk menawarkan program restrukturisasi pembayaran yang memungkinkan cicilan lebih ringan atau tenor lebih panjang guna meningkatkan peluang pembayaran tepat waktu.
4. **Kaji Kembali Profil Risiko Pelanggan Berdasarkan Aset yang Dimiliki:**
   Gunakan OWN_CAR_AGE sebagai indikator tambahan untuk menilai stabilitas ekonomi pelanggan. Jika usia aset terlalu tua, mungkin diperlukan jaminan tambahan untuk menurunkan risiko.
5. **Gunakan Riwayat Bulanan Sebagai Indikator Stabilitas:**

Evaluasi MONTHS_BALANCE secara periodik untuk melihat stabilitas pembayaran pelanggan. Jika riwayat bulanan menunjukkan kestabilan, pertimbangkan untuk memberikan kredit lebih lanjut atau menaikkan batas kredit untuk pelanggan tersebut.

III. LAMPIRAN (LINK)

Repository :
https://github.com/Bramasta66/Home-Credit-Default-Risk/blob/d8bb5295677971ad9d861e0afeb379d71141e978/modeling_all.ipynb
Google Collab : modeling_all.ipynb
Google Docs : Supervised - Conexus