

STAGE 0

Preparation



1. Problem Statement

Home Credit, sebagai perusahaan pemberi pinjaman, menghadapi tantangan signifikan dalam menangani nasabah yang **gagal bayar**. Gagal bayar tidak hanya menyebabkan **kerugian finansial langsung**, tetapi juga **mengganggu arus kas perusahaan** dan **merusak reputasinya di pasar**. Berdasarkan laporan dari sektor keuangan, risiko gagal bayar dapat meningkat hingga **20% bagi nasabah dengan riwayat kredit yang buruk** (Smith, 2020). Untuk menyelesaikan masalah ini, Home Credit memiliki **dataset yang kaya** berisi informasi aplikasi pinjaman, riwayat kredit nasabah dari lembaga keuangan lain, serta riwayat pembayaran angsuran dan saldo bulanan dari berbagai kredit. Dataset ini menyediakan **insight penting** yang memungkinkan pengembangan **model prediksi risiko default** yang akurat. Dengan **analisis yang tepat**, perusahaan dapat mengidentifikasi nasabah berisiko tinggi dan mengambil keputusan yang lebih tepat untuk **mengurangi risiko default**, **meningkatkan arus kas**, dan **mempertahankan reputasinya** di pasar keuangan.

Data Set: home-credit-default-risk

<https://www.kaggle.com/competitions/home-credit-default-risk/data>

Highlights:

- **Tantangan utama:** Gagal bayar menyebabkan kerugian finansial, mengganggu arus kas, dan merusak reputasi.
- **Fakta pendukung:** Risiko gagal bayar dapat meningkat hingga 20% untuk nasabah dengan riwayat kredit buruk.
- **Solusi:** Pengembangan model prediksi risiko default menggunakan dataset yang kaya untuk mengurangi risiko.

Sumber:

Smith, J. (2020). "Default, Transition, and Recovery: 2020 Annual Global Corporate Default And Rating Transition Study." S&P Global Ratings. P Global.

2. Deskripsi Peran

Dalam proyek **Default Prediction** untuk Home Credit, setiap anggota kelompok memiliki peran penting yang didesain untuk mengatasi masalah gagal bayar yang telah diidentifikasi selama tahap *Business Understanding* menggunakan metodologi CRISP-DM. **Data Engineer**, **Data Scientist**, dan **Business Intelligence (BI)** berkolaborasi untuk meminimalkan risiko default melalui pembersihan data, pengembangan model prediksi, dan penerapan strategi berbasis data untuk mitigasi risiko.

A. Data Engineer

Peran **Data Engineer** membutuhkan delapan orang karena volume data yang sangat besar dan kompleksitas dalam pembersihan serta persiapan data. Dataset seperti *bureau_balance.csv*, *POS_CASH_balance.csv*, dan *installments_payments.csv* memiliki jutaan baris data yang harus dibersihkan dari missing values, duplikasi, dan outliers. Proses ini juga melibatkan penggabungan dan transformasi antar dataset untuk memastikan data siap digunakan oleh tim. Pembagian tanggung jawab kepada delapan orang meningkatkan efisiensi dan menghindari bottleneck dalam pipeline. Setiap Data Engineer bertanggung jawab untuk memastikan data berkualitas tinggi untuk mendukung analisis risiko kredit yang lebih baik (MDPI)(Finance Train).

Anggota dan tugas data engineer

1. Fakhri Dwi Nugroho: *POS_CASH_balance.csv* - Mengelola data snapshot bulanan dari saldo pinjaman POS dan tunai nasabah, yang dapat memberikan indikasi kemampuan pembayaran melalui transaksi bulanan.
2. Jericho Medion: *application_test.csv* - Mengelola data aplikasi pinjaman tanpa TARGET, digunakan untuk menguji model prediksi risiko gagal bayar.
3. Johannes Sibarani: *Bureau.csv* - Mengelola riwayat kredit nasabah dari lembaga keuangan lain, yang membantu mengevaluasi riwayat kredit eksternal untuk menilai risiko gagal bayar.
4. Muhammad Naufal: *credit_card_balance.csv* - Mengelola data snapshot bulanan dari saldo kartu kredit nasabah, memberikan wawasan tentang pola penggunaan kredit dan potensi gagal bayar.
5. Mauliddinia Iftikhar Agnany: *previous_application.csv* - Mengelola riwayat aplikasi pinjaman sebelumnya, membantu memahami pola pengajuan kredit yang dapat memengaruhi keputusan risiko gagal bayar.
6. Ryan Nofandi: *bureau_balance.csv* - Mengelola data saldo bulanan kredit eksternal, memberikan wawasan tambahan tentang kemampuan pembayaran nasabah secara historis.
7. Achmad Fichri Rachmadhani: *application_train.csv* - Mengelola data aplikasi pinjaman dengan TARGET, digunakan untuk melatih model dalam mengukur risiko gagal bayar.
8. Hijir Della Wirasti: *installments_payments.csv* - Mengelola riwayat pembayaran angsuran nasabah, memberikan informasi penting tentang kemampuan nasabah untuk melakukan pembayaran tepat waktu atau mengalami keterlambatan.

B. Data Scientist

Data Scientist bertanggung jawab mengembangkan dan memimpin proses pemodelan prediksi risiko gagal bayar, memilih algoritma machine learning yang sesuai, dan mengoptimalkan model. Algoritma yang digunakan termasuk **Random Forests** yang telah terbukti efektif dalam mengidentifikasi nasabah berisiko tinggi melalui analisis

data historis. Algoritma ini secara signifikan meningkatkan akurasi prediksi risiko kredit, memungkinkan perusahaan membuat keputusan pinjaman yang lebih tepat. Model prediksi harus memiliki nilai **AUC-ROC minimal 0.85**. Nilai di atas 0.85 dianggap sangat baik untuk pengklasifikasi risiko kredit, memastikan model memiliki daya prediksi yang kuat terhadap gagal bayar. Dengan memanfaatkan pola dalam data historis, model prediktif ini memberikan insight yang mendalam tentang risiko gagal bayar nasabah (MDPI)(Finance Train).

Anggota Data Scientist

1. Hijir Della Wirasti
2. Ryan Nofandi
3. Mauliddinia Iftikhar Agnany
4. Achmad Fichri Rachmadhani

C. Business Intelligent:

Business Intelligence (BI) memainkan peran strategis dalam mengevaluasi hasil prediksi risiko gagal bayar dari perspektif bisnis. BI bertugas menyelaraskan hasil analisis teknis dengan tujuan bisnis Home Credit, memastikan bahwa model prediksi digunakan untuk mendukung profitabilitas perusahaan dan mengurangi kerugian akibat gagal bayar. BI menerjemahkan hasil prediksi model menjadi rekomendasi bisnis yang actionable, yang kemudian diintegrasikan ke dalam strategi mitigasi risiko perusahaan. Dengan demikian, BI memungkinkan perusahaan untuk merespons risiko lebih cepat dan efektif, meningkatkan efisiensi keputusan bisnis terkait pemberian pinjaman dan mengurangi potensi kerugian finansial (Finance Train)(MyGreatLearning).

Anggota Business Intelligence

1. Johannes Sibarani
2. Fakhri Dwi Nugroho
3. Jericho Medion
4. Muhammad Naufal

Sumber:

Rivera, L. A., & Herrera, J. A. (2023). *Machine Learning for Credit Risk Prediction: A Systematic Literature Review*. *Data*, 8(11), 169.

Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*.
Role of Data Science in Risk Management. *Finance Train*.

3. Goal

Goal utama dari proyek ini adalah **menurunkan jumlah nasabah yang gagal bayar**. Gagal bayar tidak hanya menyebabkan kerugian finansial langsung bagi Home Credit, tetapi juga berdampak pada arus kas perusahaan dan reputasi di pasar. Dengan mengembangkan model prediksi risiko kredit yang akurat, perusahaan dapat mengambil keputusan yang lebih baik

dalam memberikan pinjaman, meminimalkan risiko default, dan meningkatkan inklusi finansial secara bertanggung jawab.

4. Objective

- 1) **Mengidentifikasi faktor-faktor yang mempengaruhi gagal bayar:** Setelah data bersih dan siap digunakan, analisis awal dilakukan untuk mengidentifikasi variabel-variabel yang berkaitan dengan risiko gagal bayar, seperti riwayat kredit, pembayaran angsuran, dan aplikasi pinjaman sebelumnya.
- 2) **Mengembangkan model machine learning:** Berdasarkan dataset yang telah diolah, model machine learning dikembangkan untuk memprediksi risiko default. Pemilihan algoritma yang tepat seperti **Random Forests**, **Gradient Boosting**, dan **Support Vector Machines** dilakukan untuk mencapai hasil prediksi yang akurat. **Akurasi model** diukur menggunakan nilai **AUC-ROC**, di mana model prediksi harus memiliki nilai minimal **0.85**. Nilai AUC-ROC ini menandakan kemampuan model dalam memisahkan nasabah berisiko tinggi dan rendah dengan akurasi tinggi. Nilai di atas 0.85 dianggap sangat baik untuk pengklasifikasi risiko kredit, memastikan model memiliki daya prediksi yang kuat dalam mengidentifikasi risiko gagal bayar secara akurat ([Glass Box](#))([MyGreatLearning](#)).
- 3) **Mengevaluasi hasil prediksi dari perspektif bisnis:** Setelah model dioptimalkan, hasil prediksi dievaluasi dari perspektif bisnis. Tujuannya adalah memastikan model relevan dengan kebutuhan bisnis Home Credit dan membantu dalam pengambilan keputusan pemberian pinjaman.
- 4) **Membuat rekomendasi bisnis berdasarkan faktor-faktor penyebab gagal bayar:** Berdasarkan hasil analisis dan model prediksi, rekomendasi bisnis dibuat untuk membantu Home Credit menyusun strategi mitigasi risiko kredit, misalnya dengan menawarkan pinjaman yang lebih selektif kepada nasabah dengan risiko tinggi.
- 5) **Mengurangi risiko gagal bayar:** Dengan implementasi model dan rekomendasi bisnis yang tepat, Home Credit diharapkan dapat mengurangi risiko gagal bayar secara signifikan, meningkatkan efisiensi dalam pemberian pinjaman, dan memperbaiki arus kas serta reputasi perusahaan.
- 6) **Menghasilkan laporan analisis risiko dan visualisasi data:** Setelah implementasi model dan melakukan monitoring, tim akan melakukan *compile* dokumen yang merangkum faktor-faktor utama yang berkontribusi terhadap risiko gagal bayar berdasarkan analisis data. Selain itu tim akan menyusun visualisasi data dengan membuat Grafik dan atau dashboard interaktif menunjukkan distribusi data nasabah berdasarkan tingkat risiko gagal bayar, hubungan antara variabel-variabel utama, dan hasil prediksi model.

5. Business Metrics

Main Metrics:

- a) **Default Prediction**

Default Prediction adalah metrik utama yang digunakan untuk mengukur risiko gagal bayar nasabah. Prediksi ini membantu Home Credit menentukan kemungkinan seorang nasabah tidak mampu membayar pinjamannya berdasarkan profil risiko. Dengan menggunakan algoritma machine learning, model prediksi akan memberikan **skor risiko** untuk setiap nasabah, mengklasifikasikannya sebagai berisiko tinggi atau rendah untuk gagal bayar. Ini memungkinkan perusahaan untuk membuat keputusan yang lebih baik terkait pemberian pinjaman, berdasarkan tingkat risiko setiap nasabah ([SpringerLink](#))(MDPI).

Secondary Metrics :

b) Loss Ratio

Loss Ratio adalah metrik sekunder yang digunakan untuk mengukur dampak finansial dari kredit macet. Ini dihitung sebagai persentase kerugian yang disebabkan oleh gagal bayar dibandingkan dengan total pinjaman yang diberikan. Metrik ini sangat penting bagi Home Credit karena mencerminkan dampak gagal bayar pada arus kas dan profitabilitas perusahaan. **Loss Ratio** digunakan untuk menilai efektivitas model prediksi dan strategi mitigasi risiko yang diterapkan ([Risk.net](#)).

$$\text{Loss Ratio} = \frac{\text{Total Kerugian dari Gagal Bayar}}{\text{Total Pinjaman yang Diberikan}}$$

Semakin rendah loss ratio, semakin baik kinerja perusahaan dalam mengelola risiko kredit. Idealnya, implementasi model prediksi default yang akurat akan membantu menurunkan loss ratio, karena Home Credit dapat lebih selektif dalam memberikan pinjaman hanya kepada nasabah dengan risiko rendah. Selain itu, perusahaan juga dapat menggunakan loss ratio untuk memantau efektivitas strategi mitigasi risiko dan kebijakan kredit yang diterapkan berdasarkan hasil prediksi.

Hubungan Antara Main dan Secondary Metrics

Default Prediction berfungsi sebagai alat pencegahan, membantu mengidentifikasi nasabah yang berisiko tinggi gagal bayar, sedangkan **Loss Ratio** mengukur dampak keuangan dari keputusan tersebut. Sinergi antara kedua metrik ini membantu Home Credit menjaga **profitabilitas** dan **stabilitas arus kas** perusahaan.

Sumber:

Draelos, R. (2021). *Measuring Performance: AUC (AUROC). Glass Box Medicine*. Diakses dari <https://glassboxmedicine.com> ([Glass Box](#)).

What is AUC - ROC in Machine Learning. MyGreatLearning. Diakses dari <https://www.mygreatlearning.com/>

([MyGreatLearning](#)).

Validation techniques and performance metrics for loss given default models. Journal of Risk Model Validation. Diakses dari <https://www.risk.net/>([Risk.net](#)).

Machine learning-driven credit risk models. Neural Computing and Applications. Diakses dari <https://link.springer.com/>([SpringerLink](#)).

6. Portfolio Outputs

1. Kriteria Kualitas

Untuk memastikan output proyek **Default Prediction** berkualitas tinggi, tim kami menetapkan beberapa kriteria yang penting untuk dipenuhi. Berikut adalah kriteria kualitas yang digunakan, beserta sumber-sumber pendukung:

1. **Akurasi Model:** Model prediksi harus memiliki nilai **AUC-ROC minimal 0.85**. Nilai AUC-ROC ini menandakan kemampuan model dalam memisahkan nasabah berisiko tinggi dan rendah secara akurat. Nilai di atas 0.85 dianggap sangat baik untuk pengklasifikasi risiko kredit, memastikan model memiliki daya prediksi yang kuat terhadap gagal bayar (Glass Box)(MyGreatLearning).
2. **Data yang Bersih dan Siap Analisis:** Data yang digunakan harus melalui proses pembersihan yang menyeluruh untuk menghilangkan **missing values**, duplikasi, dan **outliers**. Kualitas data yang buruk dapat menyebabkan bias atau kesalahan dalam hasil prediksi, sehingga data yang bersih adalah prasyarat penting dalam pengembangan model machine learning (MyGreatLearning).
3. **Relevansi Bisnis:** Setiap output harus relevan dengan tantangan bisnis Home Credit, terutama dalam konteks mengurangi risiko gagal bayar dan meningkatkan inklusi finansial. Hasil prediksi dan analisis harus dapat diimplementasikan secara langsung ke dalam strategi bisnis untuk memastikan efektivitas solusi berbasis data(Glass Box).
4. **Visualisasi yang Informatif:** Visualisasi data harus jelas dan mudah dimengerti, dengan fokus pada insight yang mendukung pengambilan keputusan. Visualisasi yang efektif akan membantu tim bisnis memahami distribusi risiko nasabah dan hubungan antar variabel penting yang ditemukan dalam analisis (MyGreatLearning).

2. Ketepatan Sasaran (Alignment)

Setiap output dalam proyek **Default Prediction** ini akan selaras dengan tujuan bisnis yang telah ditetapkan selama tahap *Business Understanding* dalam metodologi CRISP-DM. Misalnya, model prediksi risiko default yang dikembangkan akan membantu Home Credit dalam mengidentifikasi nasabah berisiko tinggi gagal bayar. Laporan analisis faktor risiko serta rekomendasi bisnis akan memfasilitasi pengambilan keputusan strategis untuk mengurangi kerugian finansial dan memperbaiki arus kas perusahaan. Setiap output akan memberikan hasil yang relevan untuk meningkatkan efektivitas operasional perusahaan dan mendukung tujuan bisnis secara keseluruhan(Glass Box)(MyGreatLearning).

Sumber:

Draeos, R. (2021). *Measuring Performance: AUC (AUROC). Glass Box Medicine. What is AUC - ROC in Machine Learning. MyGreatLearning.*

Dalam proyek ini, *Ketepatan Sasaran* merujuk pada kesesuaian antara hasil analisis teknis dan tujuan bisnis Home Credit, khususnya dalam mengurangi risiko gagal bayar. Untuk mencapai ketepatan sasaran, proyek ini harus berfokus pada tiga elemen utama: pemodelan

prediktif risiko gagal bayar, implementasi hasil analisis dalam strategi bisnis, serta dampak nyata pada arus kas dan profitabilitas perusahaan.

1. Keselarasan dengan Tantangan Bisnis:

- Risiko gagal bayar menjadi masalah utama yang dihadapi Home Credit, sehingga fokus utama proyek ini adalah pada pengembangan model prediksi yang mampu mengidentifikasi nasabah berisiko tinggi.
- Metrik *Default Prediction* dan *Loss Ratio* dipilih karena secara langsung terkait dengan tantangan utama perusahaan dalam mengelola risiko kredit dan kerugian finansial.

2. Penggunaan Data yang Tepat:

- Dataset yang digunakan berasal dari informasi riwayat kredit dan pembayaran nasabah yang sangat relevan untuk menilai kemampuan pembayaran. Variabel-variabel seperti *application_train.csv*, *previous_application.csv*, dan *installments_payments.csv* menyediakan data penting yang memungkinkan analisis mendalam tentang risiko kredit nasabah.
- Proses pembersihan dan integrasi dataset oleh tim Data Engineer memastikan data berkualitas tinggi, yang akan meningkatkan ketepatan hasil prediksi model.

3. Pemodelan Machine Learning yang Mendukung Keputusan Bisnis:

- Algoritma machine learning seperti Random Forests, Gradient Boosting, dan XGBoost yang digunakan memiliki daya prediksi kuat dan telah terbukti efektif dalam pemodelan risiko gagal bayar di industri keuangan. Nilai AUC-ROC minimal 0.85 memastikan model memiliki akurasi yang memadai untuk diimplementasikan dalam keputusan bisnis.
- Model prediktif yang dihasilkan memberikan skor risiko pada setiap nasabah, memudahkan Home Credit dalam menentukan nasabah mana yang layak mendapatkan pinjaman.

4. Implementasi Hasil dalam Strategi Bisnis:

- Business Intelligence (BI) bertugas menerjemahkan hasil prediksi model ke dalam strategi mitigasi risiko yang actionable. Dengan demikian, hasil analisis teknis diselaraskan dengan prioritas bisnis Home Credit, seperti mengurangi risiko gagal bayar, memperbaiki arus kas, dan mempertahankan reputasi di pasar keuangan.

5. Dampak pada Profitabilitas dan Inklusi Finansial:

- Dengan model prediksi yang akurat dan rekomendasi bisnis berbasis data, Home Credit dapat lebih selektif dalam memberikan pinjaman kepada nasabah berisiko tinggi, sehingga risiko gagal bayar berkurang.
- Selain itu, proyek ini tidak hanya bertujuan untuk mengurangi risiko kredit tetapi juga meningkatkan inklusi finansial dengan memberikan akses kredit yang lebih bertanggung jawab kepada nasabah berisiko rendah.

Secara keseluruhan, *Ketepatan Sasaran* dalam proyek ini tercapai melalui penyalarsan yang kuat antara analisis teknis, rekomendasi bisnis, dan kebutuhan strategis Home Credit dalam mengelola risiko gagal bayar.