PREDIKSI RISIKO GAGAL BAYAR PADA NASABAH KREDIT

Disusun dan Dipresentasikan oleh Ardy Ansyah

AGENDA

INTRODUCTIONS 01

OVERVIEW PROJECT 02

MAIN PROJECT 03

04 EXPLORATORY DATA ANALYSIS

05 MODELING & EVALUATION

06 CONCLUSION & RECOMMENDATION

01

INTRODUCTION

TENTANG SAYA

Saya Ardy Ansyah, memiliki passion di bidang data dan meyakini bahwa data berperan penting dalam pengambilan keputusan yang tepat. Ketertarikan saya pada data analyst dan data science mendorong saya untuk terus belajar dan mengasah kemampuan dalam pengolahan data secara efektif dan profesional.

PENDIDIKAN

UNIVERSITAS INDRAPRASTA PGRI

Merancang dan membangun sistem inventori barang menggunakan java netbeans untuk mengelola stok, pencatatan masuk atau keluar barang, serta pelaporan data secara lebih efisien dan terstruktur

Dibimbing

Mempelajari tentang data preprocessing seperti data cleaning, manipulasi data, EDA dan Melakukan proses pemodelan machine learning

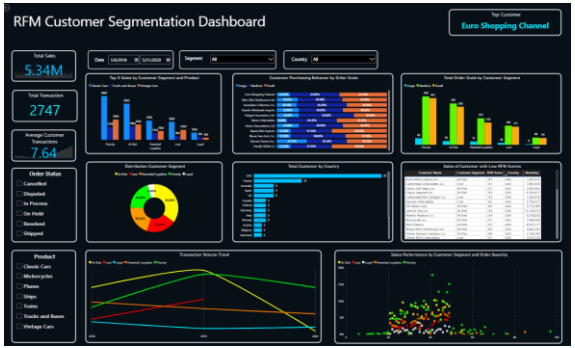
02

OVERVIEW PROJECT

PROYEK SEBELUMNYA

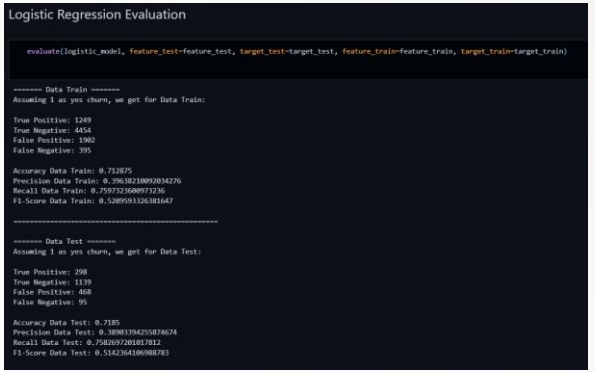
RFM Customer Segmentation

Auto Sales Data - Kaggle



Churn Analysis

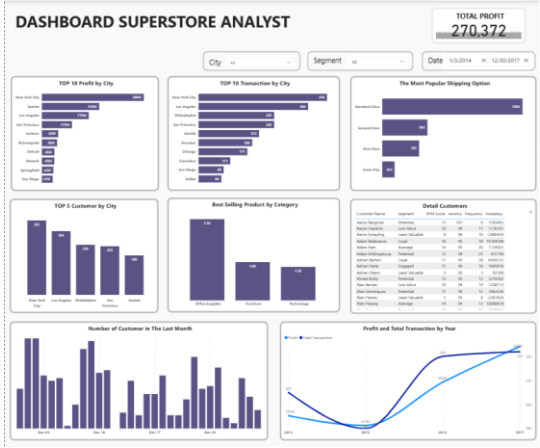
Bank Customer Churn Prediction - Kaggle



PROYEK SEBELUMNYA

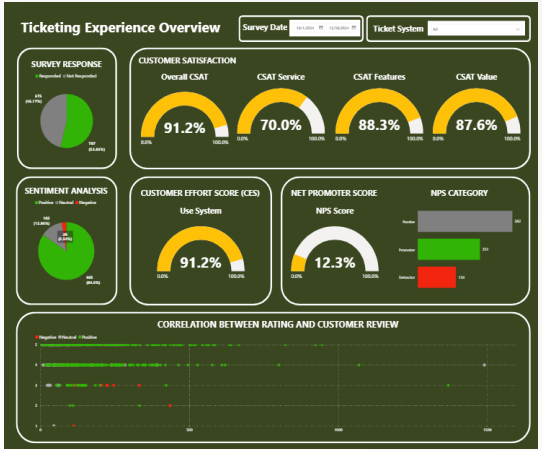
Marketing Channel Analyst

Superstore - Kaggle



Customer Satisfaction & Sentiment Analysis

Car Rental Review - Kaggle



02

MAIN PROJECT

LATAR BELAKANG PROYEK

Proyek ini berfokus pada prediksi risiko gagal bayar nasabah menggunakan *credit risk dataset* yang berisi informasi demografis, keuangan, serta riwayat pinjaman nasabah. Dengan adanya model prediksi ini, lembaga keuangan dapat mengurangi risiko kerugian, meningkatkan akurasi penilaian kredit, serta mendukung pengambilan keputusan berbasis data dalam proses pemberian pinjaman. Hasil yang diharapkan dari proyek ini adalah model prediksi yang akurat dan dapat diinterpretasikan, serta analisis faktor-faktor utama yang memengaruhi risiko gagal bayar nasabah.

PERYATAAN MASALAH

Salah satu tantangan terbesar bagi perusahaan pembiayaan dan lembaga keuangan adalah meningkatnya risiko gagal bayar (default) dari nasabah. Gagal bayar menyebabkan kerugian finansial yang signifikan, karena jumlah pinjaman (loan amount) yang tidak kembali dapat berdampak langsung pada arus kas dan profitabilitas perusahaan. Berdasarkan data historis, sekitar **21,9% nasabah** mengalami gagal bayar, yang mengakibatkan kerugian total mencapai **$76,933,675.00**

Permasalahan utama yang ingin dipecahkan melalui proyek ini adalah bagaimana memprediksi lebih awal calon nasabah yang berpotensi gagal bayar, sehingga perusahaan dapat:

- Mengambil langkah preventif (misalnya penolakan, permintaan jaminan tambahan, atau penyesuaian bunga).
- Mengoptimalkan proses penilaian risiko kredit.
- Mengurangi potensi kerugian keuangan di masa mendatang.

TUJUAN PROYEK

- Mengidentifikasi faktor-faktor utama yang berkontribusi terhadap kemungkinan gagal bayar.
- Mengembangkan model prediktif berbasis machine learning untuk memperkirakan risiko gagal bayar setiap nasabah.
- Memberikan insight yang membantu tim analis kredit dalam mengambil keputusan berbasis data. Mengurangi potensi **kerugian perusahaan hingga jutaan dolar** dengan penerapan sistem prediksi risiko yang lebih akurat.

PEMAHAMAN DATA

Memeriksa Duplikasi Data

Terdapat duplikasi pada data dengan tingkat 99,49%

↓

Telah dilakukan handling drop duplicate

»

Memeriksa Missing Value

Terdapat missing value (NaN) pada kolom person_emp_length dan loan_int_rate

↓

Nilai yang hilang (NaN) telah ditangani menggunakan nilai median dari variable tersebut

»

Memeriksa Nilai Outlier

Terdapat outlier pada kolom person_age, person_income, person_emp_length, loan_percent_income, cb_person_cred_hist_length

↓

Telah dilakukan penanganan outlier dengan metode IQR

»

Memeriksa Inkonsistensi data

Terdapat inkonsistensi data pada kolom person_emp_length = 123

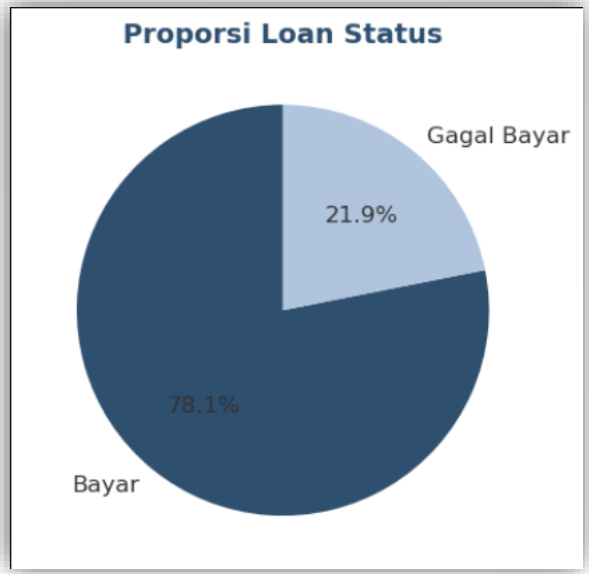
↓

Hapus baris yang berisikan inkonsistensi data tersebut

DATASET SUMMARY

1. Credit Risk Dataset diperoleh dari Kaggle
2. Dataset ini terdapat 32.581 baris dan 12 kolom
3. Terdapat missing value pada:
    - Person_emp_length: 895 data kosong
    - Loan_int_rate: 3.116 data kosong
4. Tipe data:
    - Numerik: person_age, person_income, person_emp_length, loan_amnt, loan_int_rate, loan_percent_income, cb_person_cred_hist_length, loan_status
    - Kategorikal: person_home_ownership, loan_intent, loan_grade, cb_person_default_on_file

EXPLORATORY DATA ANALYSIS

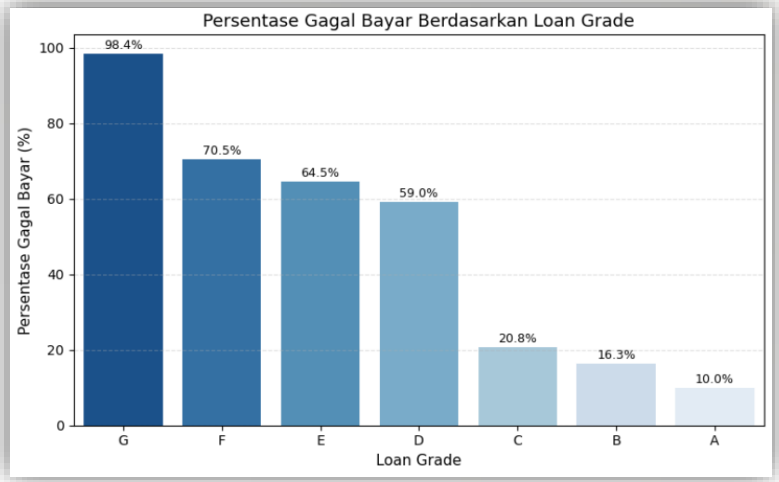


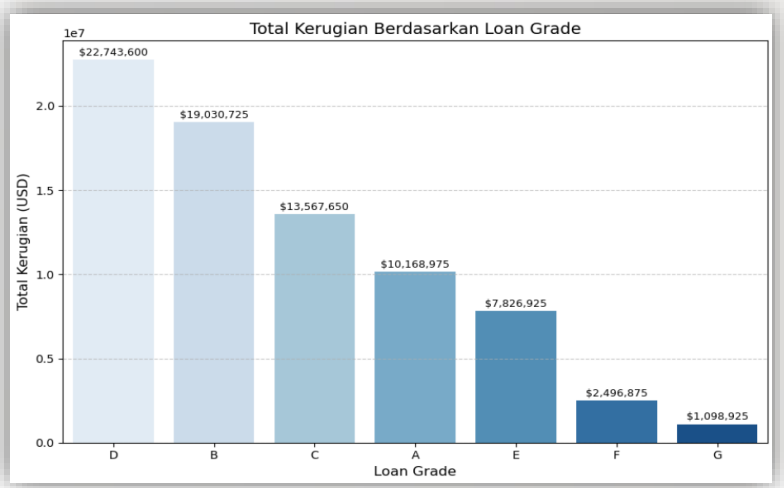

```
Jumlah nasabah gagal bayar : 7,088
Jumlah nasabah yang membayar     : 25,326
```

Sebagian Besar Nasabah Melunasi Pinjaman, Namun Risiko Kerugian Tetap Ada

Berdasarkan hasil visualisasi, dapat dilihat bahwa distribusi nasabah yang bayar berjumlah **25.326 (78.1%)** lebih banyak dari pada yang gagal bayar dengan jumlah **7.088 nasabah (21.9%)**

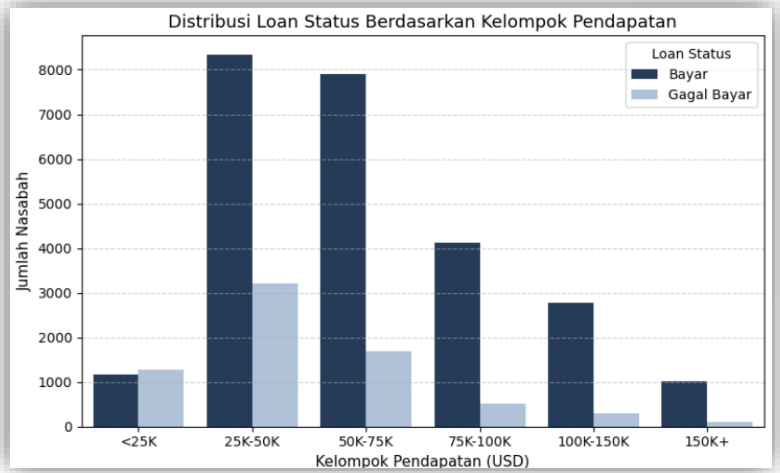
Meskipun Loan Grade G Paling Berisiko, Kerugian Terbesar Justru Terjadi pada Loan Grade D





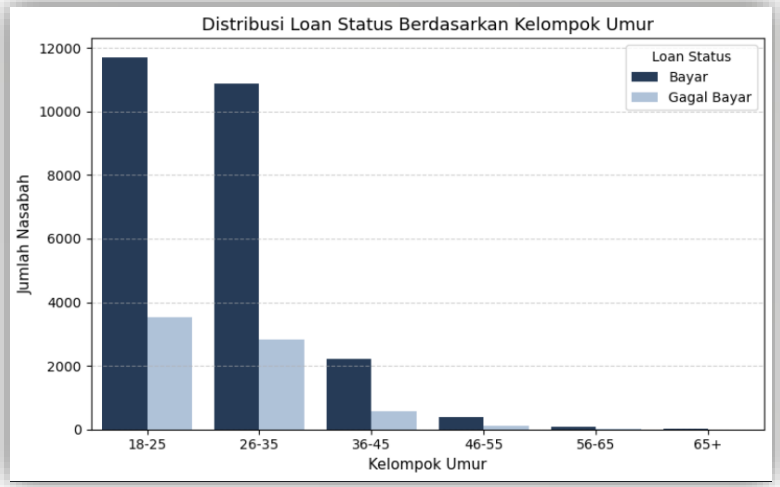
Dari visualisasi terlihat bahwa **Loan Grade G** memiliki tingkat gagal bayar tertinggi **(98.4%)**, menandakan risiko kredit yang sangat besar. Namun, secara mengejutkan, **Loan Grade D** memberikan kerugian finansial terbesar mencapai **$22,743,600**.

Nasabah Berpendapatan Menengah Paling Banyak Melunasi, Tapi Risiko Gagal Bayar Masih Tinggi



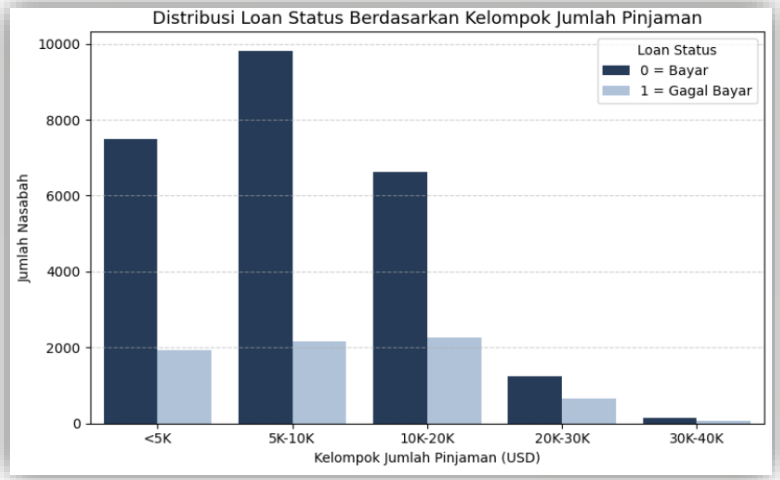
Berdasarkan hasil visualisasi, terlihat bahwa kelompok pendapatan **25K–50K & 50k-75k** mendominasi jumlah nasabah baik yang membayar maupun gagal bayar. Meskipun mayoritas nasabah pada kelompok ini berhasil melunasi pinjaman, proporsi gagal bayar juga cukup tinggi, menunjukkan bahwa pendapatan menengah tidak selalu menjamin kemampuan pelunasan yang stabil.

Nasabah Usia Muda Mendominasi Pinjaman, Namun Juga Menanggung Risiko Gagal Bayar Tertinggi



Berdasarkan hasil visualisasi, terlihat bahwa kelompok usia **18–25 tahun dan 26–35 tahun** merupakan **nasabah paling dominan** dalam pengajuan pinjaman. Meskipun mayoritas nasabah muda ini berhasil melunasi pinjamannya, **proporsi gagal bayar juga cukup tinggi** pada rentang usia tersebut. Hal ini menunjukkan bahwa **nasabah usia muda cenderung lebih berani mengambil risiko finansial**, namun **stabilitas kemampuan bayar mereka masih perlu diwaspadai**

Jumlah Pinjaman Rendah Tidak Menjamin Risiko Rendah



Berdasarkan hasil visualisasi, terlihat bahwa kelompok pinjaman dengan jumlah **<5K USD dan 5K–10K USD** merupakan segmen dengan jumlah **nasabah tertinggi**, baik untuk yang **melunasi** maupun yang **gagal bayar**. Meskipun mayoritas nasabah pada kelompok ini berhasil membayar kewajibannya, **proporsi gagal bayar juga cukup signifikan**, menandakan bahwa **pinjaman kecil tidak selalu menjamin risiko rendah**.

DATA PRE-PROCESSING

SPLIT DATASET

Memisahkan dataset df menjadi variabel fitur (feature) dan target (target), lalu membaginya menjadi data latih dan data uji (80:20)

FEATURE ENGINEERING

Melakukan encoding pada variabel train dan test, lalu melakukan pengecekan multicollinearity dan scaling

MODEL SELECTION

Model	Data	Metrics			
		Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	Train	0.738349	0.462214	0.568526	0.847939
	Test	0.734282	0.442775	0.552432	0.848173
KNN	Train	0.785333	0.594728	0.676868	0.930743
	Test	0.647228	0.484263	0.554010	0.802690
Decision Tree	Train	1.000000	0.999649	0.999824	1.000000
	Test	0.756606	0.778255	0.767278	0.854712
Random Forest	Train	1.000000	0.999473	0.999736	1.000000
	Test	0.979671	0.723891	0.832579	0.931043
XGBoost	Train	0.990245	0.802812	0.886732	0.987407
	Test	0.966574	0.744635	0.841212	0.947291
LightGBM	Train	0.989946	0.744112	0.849604	0.974099
	Test	0.983591	0.728898	0.837305	0.946609

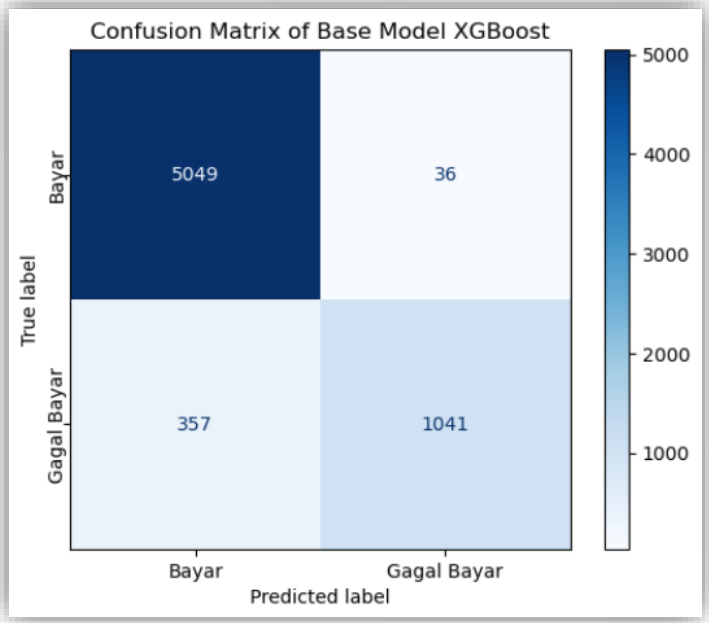
Berdasarkan hasil evaluasi base model yang disajikan, XGBoost terbukti sebagai base model dengan performa terbaik. Model ini mencapai F1-Score tertinggi pada data test (0.8412) dan sekaligus ROC AUC tertinggi (0.9472), yang menunjukkan keseimbangan optimal antara precision dan recall serta kemampuan klasifikasi yang sangat baik.

OPTIMIZING XGBOOST PARAMETERS

Model	Metrics			
	Precision	Recall	F1-Score	ROC - AUC
Base XGBoost	0.966574	0.744635	0.841212	0.947291
Tuned XGboost	0.984496	0.726753	0.836214	0.947167
Tuned XGBoost With SMOTE	0.963585	0.738197	0.835966	0.946877

Base XGBoost Model menunjukkan performa terbaik tanpa perlu hyperparameter tuning atau SMOTE. Upaya tuning dan handling class imbalance justru menurunkan keseimbangan overall model, khususnya dalam hal recall dan F1-Score yang krusial untuk use case prediksi gagal bayar.

MODEL ANALYSIS



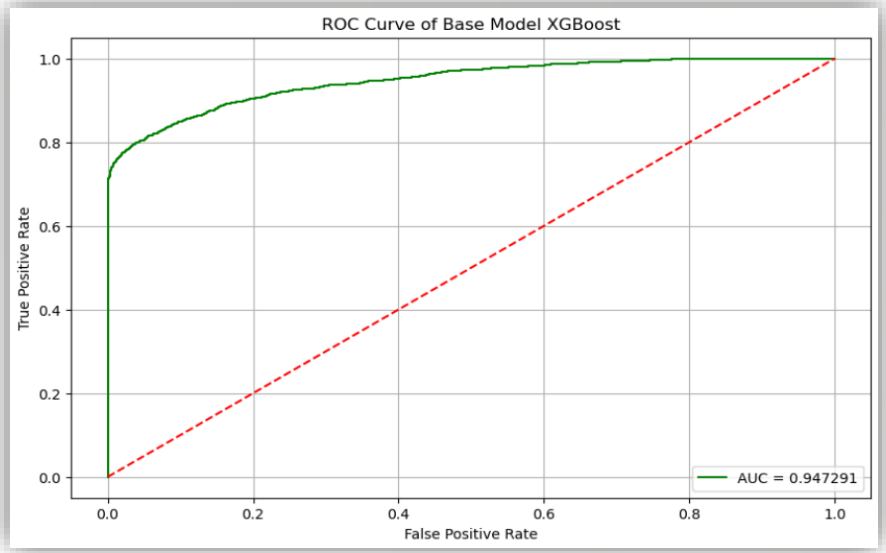
5.049 adalah jumlah prediksi yang benar untuk kategori bayar (model memprediksi bayar dan benar).

36 adalah jumlah prediksi yang salah untuk kategori bayar (model memprediksi gagal bayar padahal seharusnya bayar).

357 adalah jumlah prediksi yang salah untuk kategori gagal bayar (model memprediksi bayar padahal seharusnya gagal bayar).

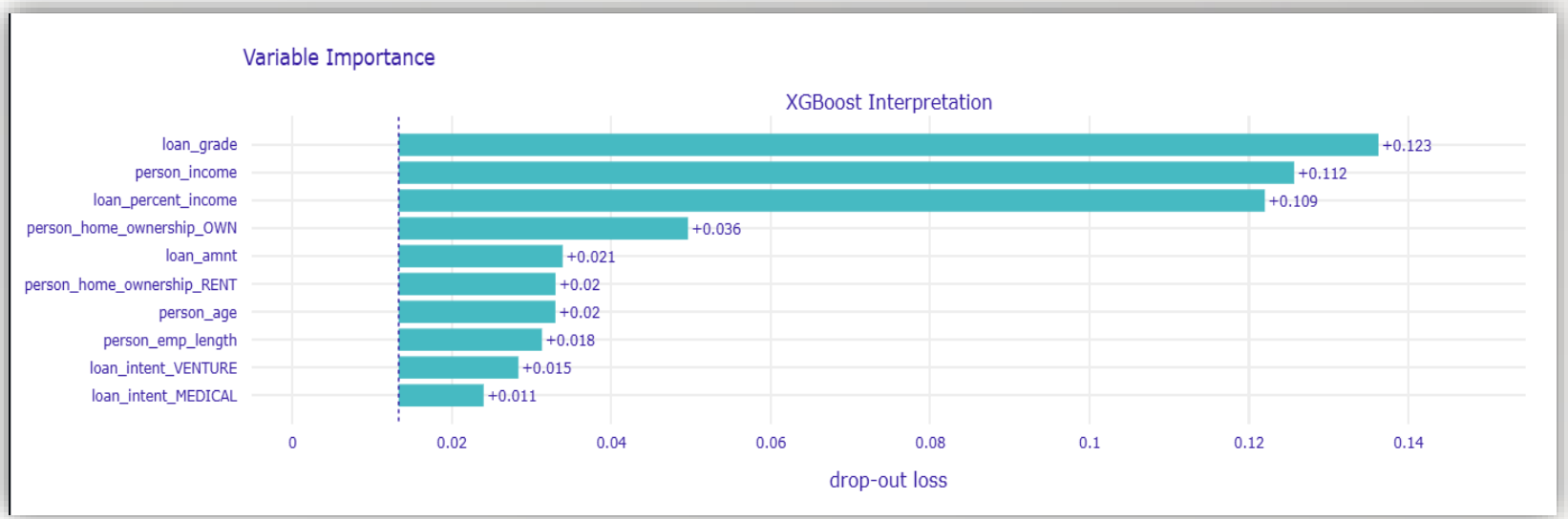
1.041 adalah jumlah prediksi yang benar untuk kategori gagal bayar (model memprediksi gagal bayar dan benar).

MODEL ANALYSIS



Berdasarkan ROC Curve dengan nilai AUC sebesar 0.947, model XGBoost terbukti sangat baik dalam memprediksi risiko gagal bayar pinjaman, menunjukkan keseimbangan yang kuat antara sensitivitas (recall) dan precision.

MODEL ANALYSIS



loan_grade menjadi variabel paling penting karena sudah merepresentasikan tingkat risiko kredit peminjam secara menyeluruh. Loan_grade menyatukan informasi dari banyak aspek keuangan lain seperti pendapatan, jumlah pinjaman, dan stabilitas pekerjaan.

CONCLUSION

Risiko Kredit Signifikan Meski Mayoritas Bayar

- Meskipun 78.1% nasabah berhasil melunasi pinjaman, terdapat 21.9% gagal bayar yang merepresentasikan risiko kerugian yang tidak kecil.

Pemisahan antara Risiko Tertinggi vs Kerugian Terbesar

- Grade G memiliki risiko gagal bayar tertinggi (98.4%), namun Grade D justru menyumbang kerugian finansial terbesar ($22.7 juta).
- Fokus tidak hanya pada probabilitas gagal bayar, tetapi juga pada potensi exposure kerugian.

Segmentasi Berisiko Tinggi

- Nasabah usia muda (18-35 tahun) dominan namun memiliki proporsi gagal bayar tinggi
- Pendapatan menengah ($25K-75K) tidak menjamin kemampuan bayar stabil
- Pinjaman kecil (<$10K) tetap berisiko signifikan karena volume tinggi

RECOMMENDATION

- **Tingkatkan monitoring** khusus untuk loan **Grade D** karena potensi kerugian terbesar
- **Perketat analisis kredit** untuk nasabah usia muda dan pendapatan menengah
- **Implementasi risk-based pricing** yang mempertimbangkan tidak hanya grade tapi juga total dana yang dipinjamkan/diinvestasikan yang berisiko hilang
- **Develop early warning system** untuk deteksi dini potensi gagal bayar

THANK YOU!


https://www.linkedin.com/in/ardyansyah99/


ardyansyah3199@gmail.com


https://github.com/ardyous