

CAPSTONE PROJECT · MODULE 3

TRAVEL INSURANCE CLAIM PREDICTION

Machine Learning Classification Project

[Mutiaru Ayu Alzahra Ramadhani]

Data Science



Mengapa prediksi klaim penting?

Biaya Klaim

Klaim tidak terduga meningkatkan pengeluaran operasional secara signifikan

Profitabilitas

Tingginya rasio klaim menekan margin keuntungan perusahaan asuransi

Underwriting

Pricing produk perlu mencerminkan risiko aktual tiap profil nasabah

Risk Management

Antisipasi risiko klaim memungkinkan pencadangan dana yang tepat

Asuransi perjalanan menanggung risiko pembatalan, kecelakaan, kehilangan bagasi, dan darurat medis selama perjalanan. Prediksi klaim memungkinkan perusahaan beroperasi lebih efisien dan kompetitif.

01 Tidak Ada Sistem Prediksi Risiko

Perusahaan belum memiliki mekanisme otomatis untuk mengidentifikasi nasabah dengan risiko klaim tinggi sebelum perjalanan dimulai.

02 Dataset Sangat Imbalance

Hanya 1,70% data merupakan klaim (Yes). Distribusi yang sangat tidak seimbang menjadi tantangan utama dalam pemodelan klasifikasi.

03 Metrik Accuracy Tidak Tepat

Dengan data imbalance ekstrem, model yang selalu memprediksi 'Tidak Klaim' akan mendapat accuracy 98,3% — menyesatkan dan tidak berguna secara bisnis.



Identifikasi Risiko Klaim

Mendeteksi nasabah berisiko klaim tinggi sejak awal untuk intervensi dini



Bantu Underwriting & Risk Assessment

Menyediakan risk score otomatis bagi tim underwriting dalam pengambilan keputusan



Optimasi Pricing & Pencadangan Dana

Mendukung penetapan harga produk yang akurat dan cadangan klaim yang proporsional



Kurangi Risiko Tidak Terantisipasi

Menurunkan potensi kerugian akibat klaim mendadak yang tidak terdeteksi sebelumnya

DATA OVERVIEW

05

44.328

Baris Data Awal

39.661

Record Setelah Duplikat Dihapus

11

Kolom / Fitur

DISTRIBUSI TARGET (CLAIM)

Claim: No — 38.986 data (98,30%)

Claim: Yes — 675 data (1,70%)

⚠ Rasio klaim hanya 1,70% — class imbalance ekstrem yang menjadi tantangan utama pemodelan. Model baseline yang naif akan selalu memprediksi 'No Claim' dan tetap mendapat accuracy tinggi.

Duplikat Data

4.667 baris (10,53%)

Dihapus

Baris duplikat ditemukan dan di-drop agar tidak menimbulkan bias pada pemodelan.

Missing Value — Gender

27.667 baris (69,76%)

Diisi 'Unknown'

Kolom Gender diisi kategori baru 'Unknown' untuk mempertahankan data dan menghindari bias imputasi.

Nilai Negatif — Duration

Beberapa baris

Di-clip ke 0

Durasi perjalanan negatif tidak masuk akal secara bisnis, sehingga di-clip ke nilai minimum 0.

Outlier — Age

< 18 dan > 100

Di-clip ke [18, 100]

Usia di luar rentang wajar di-clip ke rentang valid 18–100 tahun.

Net Sales Negatif

Beberapa baris

Dibiarkan

Net Sales negatif kemungkinan merepresentasikan refund atau koreksi transaksi, sehingga nilai asli dipertahankan.



Agency Type Airlines memiliki claim rate lebih tinggi dibandingkan Travel Agency



Produk Annual Silver Plan dan Annual Gold Plan menunjukkan claim rate yang relatif tinggi



Destinasi tertentu (Costa Rica, UAE, Hong Kong) menunjukkan risiko klaim lebih tinggi



Kelompok usia < 25 tahun dan 25–35 tahun memiliki claim rate relatif lebih tinggi



Trip dengan durasi > 30 hari cenderung memiliki claim rate lebih tinggi

Age_Group

Categorical

Segmentasi usia ke dalam bucket (<25, 25–35, 36–50, >50)

Duration_Group

Categorical

Kategorisasi durasi perjalanan (pendek, menengah, panjang, >30 hari)

Commission_Rate

Numeric

Rasio komisi terhadap Net Sales — proxy efisiensi distribusi

Is_Long_Trip

Binary

Flag binary: 1 jika durasi > 30 hari

Is_Online

Binary

Flag binary: 1 jika pembelian dilakukan secara online

High_Risk_Destination

Binary

Flag binary: 1 jika destinasi termasuk destinasi risiko tinggi

⚠ **High_Risk_Destination** dihitung dari training set saja untuk mencegah data leakage ke test set.

DATA SPLIT

Train 31.728 data (80%)

Test 7.933 data (20%)

Stratified split untuk menjaga proporsi kelas

PREPROCESSING PIPELINE

Numeric Median Imputer + StandardScaler

Tree-based OrdinalEncoder

LogReg OneHotEncoder

STRATEGI MENGATASI CLASS IMBALANCE

Resampling

(NearMiss, SMOTE)

Class Weight

class_weight='balanced'

Scale Pos Weight

untuk XGBoost

Threshold Tuning

probability cutoff

Cost-Sensitive

analisis biaya FP/FN

Metrik utama: Recall (prioritas), PR AUC, ROC AUC, Precision, F1-Score. Accuracy dikecualikan karena data imbalance.

Model	Recall	Precision	F1	PR AUC	ROC AUC	FP	FN
Baseline XGBoost	0,0000	N/A	N/A	N/A	N/A	0	135
XGBoost + NearMiss	0,7778	0,0142	N/A	N/A	N/A	7.279	30
XGBoost Penalized	0,2815	0,0611	N/A	N/A	N/A	584	97
LogReg Balanced	0,7556	0,0581	0,1079	0,1008	0,8111	1.653	33
XGBoost Hypertuned	1,0000	0,0176	N/A	N/A	N/A	7.545	0

★ LogReg Balanced unggul dalam keseimbangan Recall, Precision, PR AUC, dan ROC AUC. Akan menjadi kandidat final setelah threshold tuning.

N/A = tidak tersedia di notebook | Metrik utama: Recall & Cost Efficiency

FALSE NEGATIVE (FN)

Klaim asli tidak terdeteksi

Cost = 5.000 unit

FALSE POSITIVE (FP)

Non-klaim diprediksi sebagai klaim

Cost = 50 unit

PERBANDINGAN TOTAL COST ANTAR MODEL

Model	FP	FN	Cost FP	Cost FN	Total Cost
XGBoost Baseline	0	135	0	675.000	675.000
XGBoost + NearMiss	7.279	30	363.950	150.000	513.950
XGBoost Penalized	584	97	29.200	485.000	514.200
LogReg Balanced ★	1.653	33	82.650	165.000	247.650
XGBoost Hypertuned	7.545	0	377.250	0	377.250

LogReg Balanced memiliki total cost terendah (247.650) — jauh lebih efisien dibanding baseline (675.000) dan XGBoost Hypertuned (377.250).

THRESHOLD TUNING

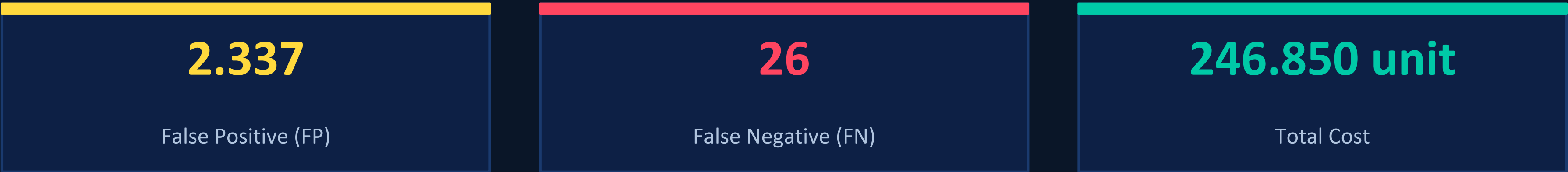
Logistic Regression Balanced — Threshold Optimal: 0.40

Default threshold classifier adalah 0.50. Dengan menurunkan threshold ke 0.40, model menjadi lebih sensitif dalam mendeteksi klaim — meningkatkan Recall namun trade-off dengan lebih banyak FP.

HASIL FINAL (THRESHOLD = 0.40)



CONFUSION MATRIX & COST



MODEL INTERPRETABILITY

13

Top Fitur Berpengaruh — Logistic Regression Balanced

● MENINGKATKAN KEMUNGKINAN KLAIM

Destination_COSTA RICA

Destinasi dengan risiko perjalanan tinggi, berkorelasi kuat dengan klaim

Product_Spouse or Parents Comprehensive Plan

Produk dengan cakupan luas cenderung lebih sering diklaim

Destination_UNITED ARAB EMIRATES

Destinasi UAE teridentifikasi sebagai high-risk destination

Destination_HONG KONG

Hong Kong termasuk destinasi dengan claim rate lebih tinggi

● MENURUNKAN KEMUNGKINAN KLAIM

Agency_ART

Agen ART berkorelasi negatif dengan klaim — nasabah dari agen ini cenderung tidak mengajukan klaim

MODEL FINAL YANG DIREKOMENDASIKAN

Logistic Regression Balanced | Threshold = 0.40

Recall: 0.8074 · PR AUC: 0.1008 · ROC AUC: 0.8111 · Total Cost: 246.850 unit

ALASAN PEMILIHAN

- ✓ Recall cukup tinggi (0.8074) — 81% klaim asli berhasil terdeteksi
- ⚖ Lebih seimbang secara operasional — FP tidak terlalu membebani tim verifikasi
- 🔍 Lebih interpretable dibanding XGBoost — koefisien mudah dijelaskan ke stakeholder bisnis
- 🚀 Cocok sebagai model produksi pertama — dapat diandalkan dan mudah di-maintain
- ✗ XGBoost Hypertuned ditolak meski Recall 1.0 — FP 7.545 terlalu tinggi, membebani operasional

REKOMENDASI BISNIS



Deploy model sebagai risk scoring tool untuk sistem underwriting



Buat dashboard monitoring klaim untuk tim risk & underwriting



Monitor data drift dan concept drift setiap 3 bulan



Retraining model setiap 6 bulan atau saat performa menurun

PENGEMBANGAN SELANJUTNYA



Tambahkan data kondisi kesehatan nasabah



Integrasikan riwayat klaim masa lalu



Informasi jenis transportasi yang digunakan



Data musim & kondisi cuaca di destinasi



Histori perjalanan nasabah sebelumnya

Model ini adalah titik awal — dengan lebih banyak data dan fitur yang relevan, akurasi prediksi klaim dapat terus ditingkatkan untuk mendukung keputusan bisnis yang lebih baik.

- Dataset memiliki class imbalance ekstrem, dengan Claim hanya 1,70%, sehingga accuracy tidak cocok sebagai metrik utama. Fokus evaluasi diarahkan pada Recall dan cost-benefit analysis karena False Negative lebih berisiko bagi perusahaan.
- Model final yang direkomendasikan adalah Logistic Regression Balanced dengan threshold 0,40. Model ini menghasilkan Recall 0,8074, False Negative 26, dan total cost 246.850 unit, sehingga lebih seimbang secara bisnis dan operasional.
- Perusahaan disarankan menggunakan model ini sebagai risk scoring tool untuk membantu underwriting, risk management, claim operation, serta pricing. Nasabah high risk tidak perlu langsung ditolak, tetapi dapat diarahkan ke proses verifikasi tambahan.
- Ke depannya, perusahaan perlu membuat dashboard monitoring, mengevaluasi performa model secara berkala, serta menambahkan data seperti riwayat klaim, kondisi kesehatan, jenis transportasi, musim perjalanan, dan aktivitas perjalanan.

A person wearing a brown hoodie and dark pants is walking away from the camera on a city street at night. They are pulling a dark blue rolling suitcase and carrying a teal backpack. To their left, a cyclist wearing a helmet and a dark jacket is riding a bicycle with a black bag that has "THE BIKING LAWYER LLP" written on it. The background shows a city street with cars, buildings, and streetlights. The text "Thank You" is overlaid in the center of the image.

Thank You