CUSTOMER LIFETIME VALUE PREDICTION

Presented by: Shafira Puspa

TABLE OF CONTENT

1. BUSINESS PROBLEM

- Latar Belakang
- Rumusan Masalah
- Tujuan bisnis
- Hasil yang diharapkan

LATAR BELAKANG

Dalam industri asuransi kendaraan, penting bagi perusahaan untuk memahami nilai jangka panjang dari setiap pelanggan. Nilai seumur hidup pelanggan atau Customer Lifetime Value (CLV) merupakan metrik penting yang menggambarkan profitabilitas pelanggan berdasarkan hubungan jangka panjang mereka dengan perusahaan.

RUMUSAN MASALAH

- Bagaimana membangun model prediktif yang mampu memperkirakan Customer Lifetime Value (CLV) secara akurat menggunakan data pelanggan dan informasi polis asuransi
- Bagaimana potensi kerugian yang dialami perusahaan jika tidak mempertimbangkan CLV dalam pengambilan keputusan retensi dan pemasaran
- Fitur-fitur apa saja (demografi, jenis kendaraan, premi, klaim, dll) yang paling berpengaruh dalam menentukan nilai CLV

TUJUAN BISNIS

Dengan mengetahui CLV, perusahaan dapat:

- Mengalokasikan sumber daya secara lebih efektif
- Menyusun strategi retensi pelanggan
- Meningkatkan efisiensi pemasaran dan pelayanan

HASIL YANG DIHARAPKAN

- Tersusunnya model prediktif CLV berbasis machine learning yang andal dan dapat digunakan dalam skala operasional.
- Tersedianya daftar fitur penting (feature importance) yang memengaruhi CLV untuk mendukung analisis bisnis dan segmentasi pelanggan.
- Estimasi kuantitatif terhadap potensi kerugian finansial akibat kesalahan dalam strategi retensi pelanggan, jika CLV tidak digunakan sebagai dasar pengambilan keputusan.

2. DATA UNDERSTANDING

Dataset berisi informasi pelanggan asuransi mobil, termasuk data demografis, detail polis, dan aktivitas klaim. Dataset ini digunakan untuk memprediksi nilai Customer Lifetime Value (CLV) - **CLV (Customer Lifetime Value)** dalam konteks **asuransi** adalah nilai total keuntungan (profit) yang diperkirakan akan diperoleh dari seorang pelanggan selama mereka tetap menjadi nasabah perusahaan asuransi.

KoloM	Deskripsi
`Vehicle Class`	Kategori kendaraan yang dimiliki oleh pelanggan, seperti `Two-Door Car`, `Four-Door Car`, `SUV`, `Luxury SUV`, dan `Luxury Car`
`Coverage`	Jenis cakupan asuransi yang dipilih oleh pelanggan. Contoh nilai: `Basic`, `Extended`, `Premium`
`Renew Offer Type`	Jenis penawaran yang diberikan saat pembaruan polis. Contoh nilai: `Offer 1`, `Offer 2`
`Employment Status`	Status pekerjaan pelanggan, seperti `Employed`, `Unemployed`, `Medical Leave`, atau `Retired`
`Marital Status`	Status pernikahan pelanggan, misalnya `Married`, `Single`, `Divorced`
`Education`	Tingkat pendidikan pelanggan, seperti `High School`, `Bachelor`, `Master`, `Doctor`
`Number of Policies`	Jumlah polis asuransi aktif yang dimiliki oleh pelanggan.
`Monthly Premium Auto`	Jumlah premi bulanan yang dibayarkan oleh pelanggan untuk asuransi kendaraan
`Total Claim Amount`	Jumlah total klaim yang diajukan pelanggan
`Income`	Pendapatan tahunan pelanggan
`Customer Lifetime Value`	Total nilai yang diperkirakan akan dihasilkan pelanggan selama masa hubungan mereka dengan perusahaan. Ini adalah target utama dalam analisis CLV.

3. DATA CLEANSING

- Missing Value
- Duplicated Value
- Outlier

MISSING VALUE

Data CLV memiliki 11 kolom dengan 6 kolom tipe object dan 5 kolom numerik. Data tersebut berisikan 5669 baris dan tidak memiliki nilai kosong.

DUPLICATED VALUE

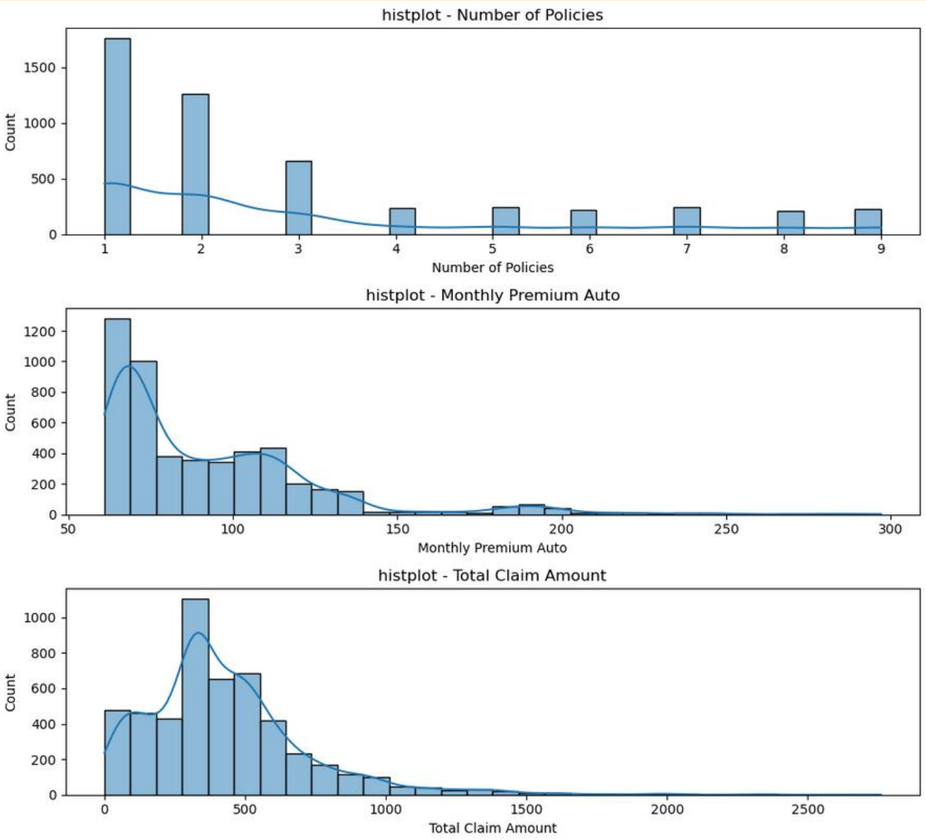
Terdapat 618 data duplikat. Dilakukan drop duplikat karena dengan asumsi tidak ada pelanggan yang benar benar identik. Jumlah baris setelah dilakukan drop duplikat hanya tinggal 5051 baris.

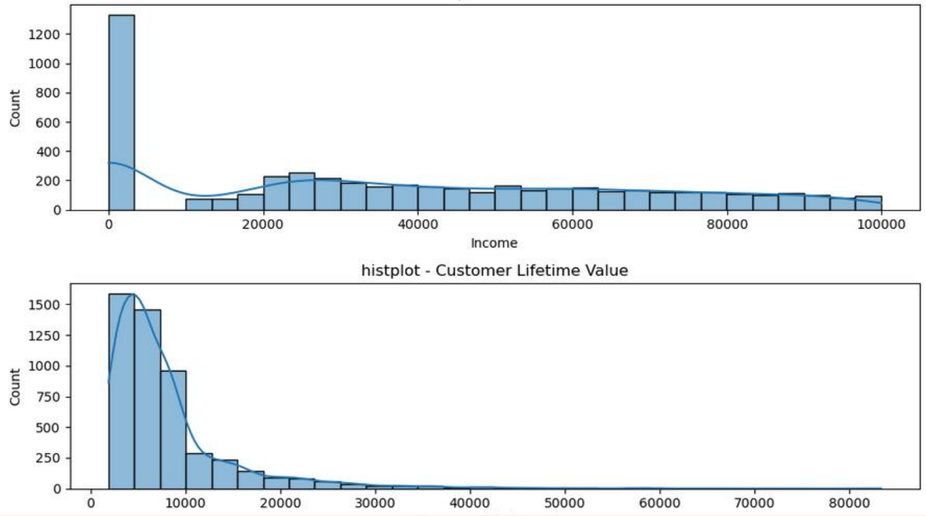
OUTLIER / DATA ANOMALI

Outlier pada data dapat dikatakan masuk akal,sehingga tidak dilakukan pengapusan outlier.

4. EDA

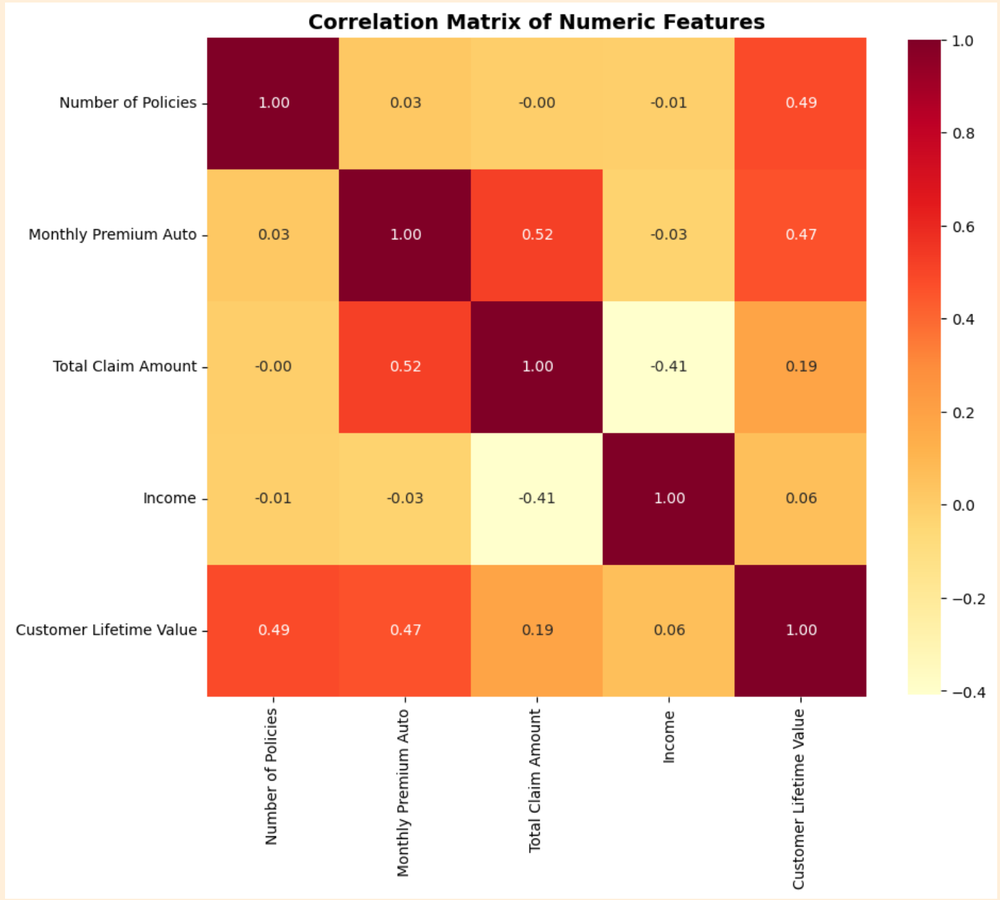
- Analisis Kolom Numerik
- Analisis Kolom Kategorik





ANALISI KOLOM NUMERIK

Bila dilihat dari grafik setiap kolom tidak terdistribusi normal. Setiap kolom cenderung right skewed yang artinya data cenderung banyak didata kecil dan terdapat outlier didata yang besar.

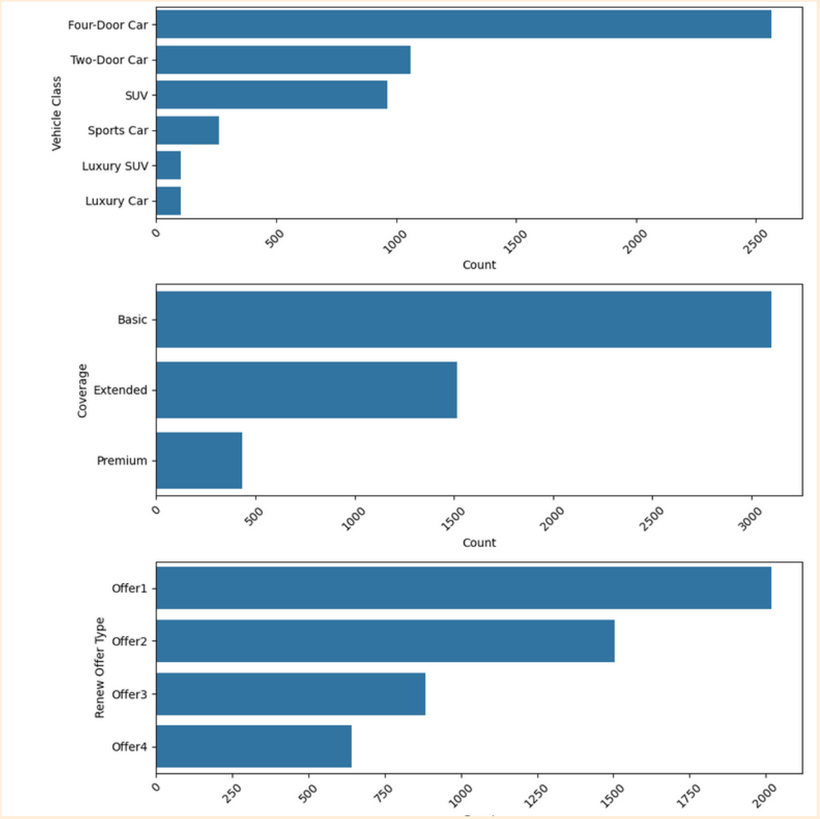


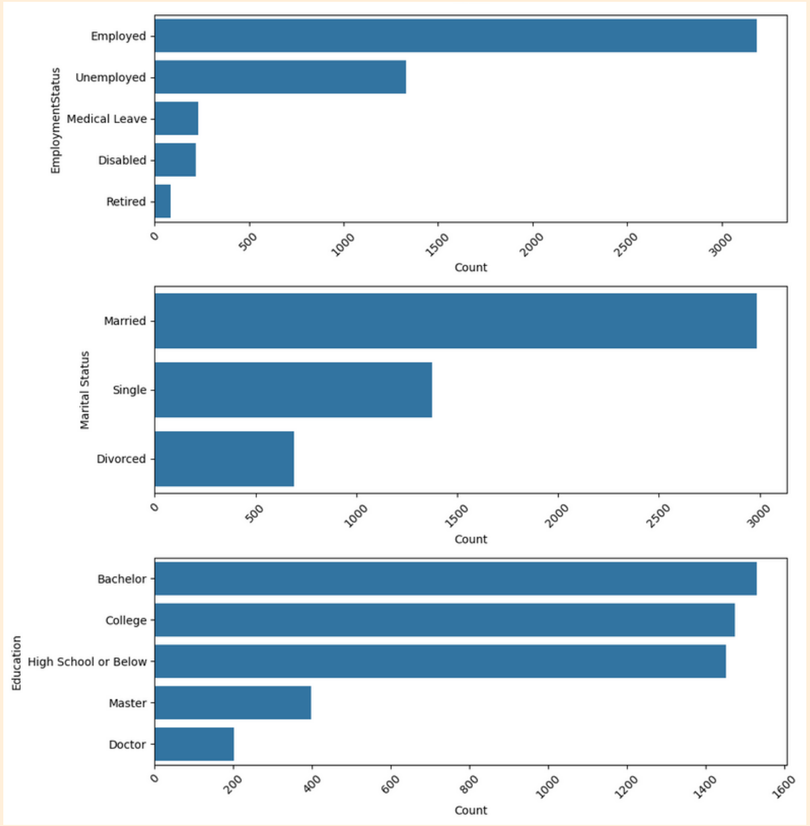
ANALISI KOLOM NUMERIK

Dilihat dari nilai korelasi didapat insight sebagai berikut:

- Fokus utama untuk prediksi CLV dapat diletakkan pada Monthly Premium Auto karena kontribusinya cukup signifikan terhadap nilai CLV.

ANALISI KOLOM KATEGORIK





Dari keenam kolom kategorik. Pengguna asuransi terbanyak memiliki tipe mobil 4 pintu, jenis asuransi basic, dan jenis penawaran yang paling banyaka diambil adalah offer 1. Identitas mayoritas pengguna asuransi adalah seorang pekerja, seorang yang menikah dan seorang sarjana.

5. MODELING

DATA PREPROCESSING

ORDINAL ENCODING

Kolom-kolom berikut diencode secara **Ordinal** karena memiliki urutan/logika hierarki yang jelas. Kolom yang dilakukan ordinal yaitu:

- `**Coverage**`: Tingkat cakupan asuransi
- `**Education**`: Tingkat pendidikan

Nilai dalam kolom ini memiliki makna relatif (tinggi-rendah) sehingga ordinal encoding **mempertahankan informasi urutan.**

ONEHOT ENCODING

Kolom-kolom berikut diencode menggunakan **One-Hot Encoding** karena merupakan variabel kategorikal nominal (tidak memiliki urutan hierarki), kolom yang dilakukan one hot encoding yaitu:

- `**Renew Offer Type**`: Jenis penawaran pembaruan
- `**Marital Status**`: Status pernikahan
- `**Vehicle Class**`: Kelas kendaraan
- **EmploymentStatus**`: Status pekerjaan

One-Hot Encoding cocok untuk kategori nominal karena mengonversi setiap nilai unik menjadi kolom biner terpisah (0/1) **tanpa membuat asumsi urutan.**

DATA PREPROCESSING

ROBUST SCALING

Kolom-kolom berikut diskalakan menggunakan **Robust Scaler** karena mengandung outlier atau distribusi non-normal:

- **Number of Policies:** Jumlah polis.
- **Monthly Premium Auto:** Premi bulanan.
- **Total Claim Amount:** Total klaim.
- **Income:** Pendapatan.

MODEL BANCHMARK

BEST MODEL

Model	Nilai MAPE	Nilai MAE
1. Gradient Boosting	0.146	1793.27
2. CatBoost	0.172	1917.89
3. Decisiom tree	0.144	2027.1
4. Random Forest	0.124	1706.00

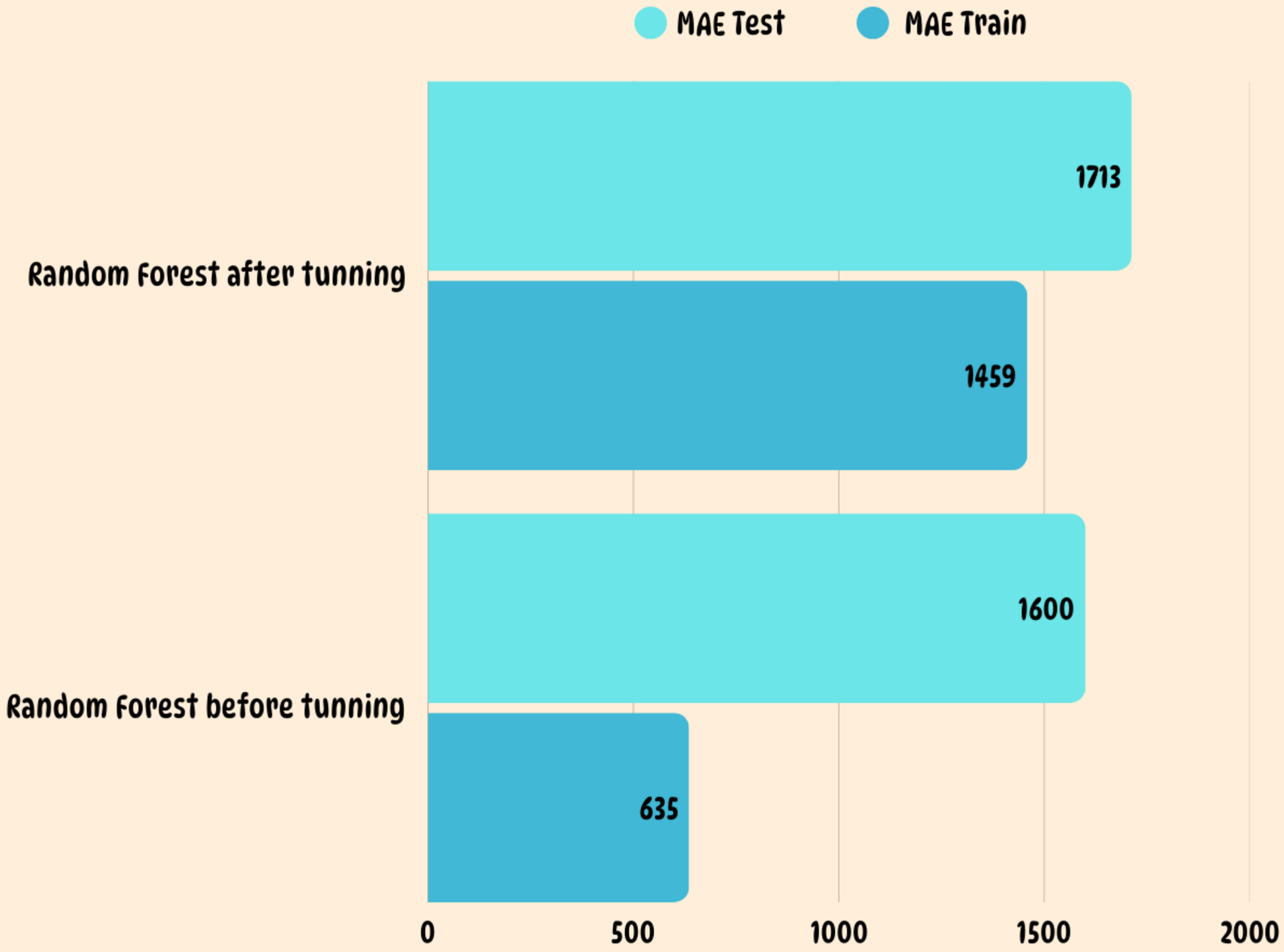
Hanya menggunakan MAE dan MAPE karena:

- MAE dan MAPE memberikan **insight praktis** yang mudah diinterpretasikan oleh tim non-teknis (seperti marketing dan manajemen).

MAE terbaik di model random forest dan MAPE terbaik di Gradient boosting. Oleh sebab itu hyperparameter dilakukan di keedua model tersebut

Perbandingan hasil mae di train dan test dalam model random forest

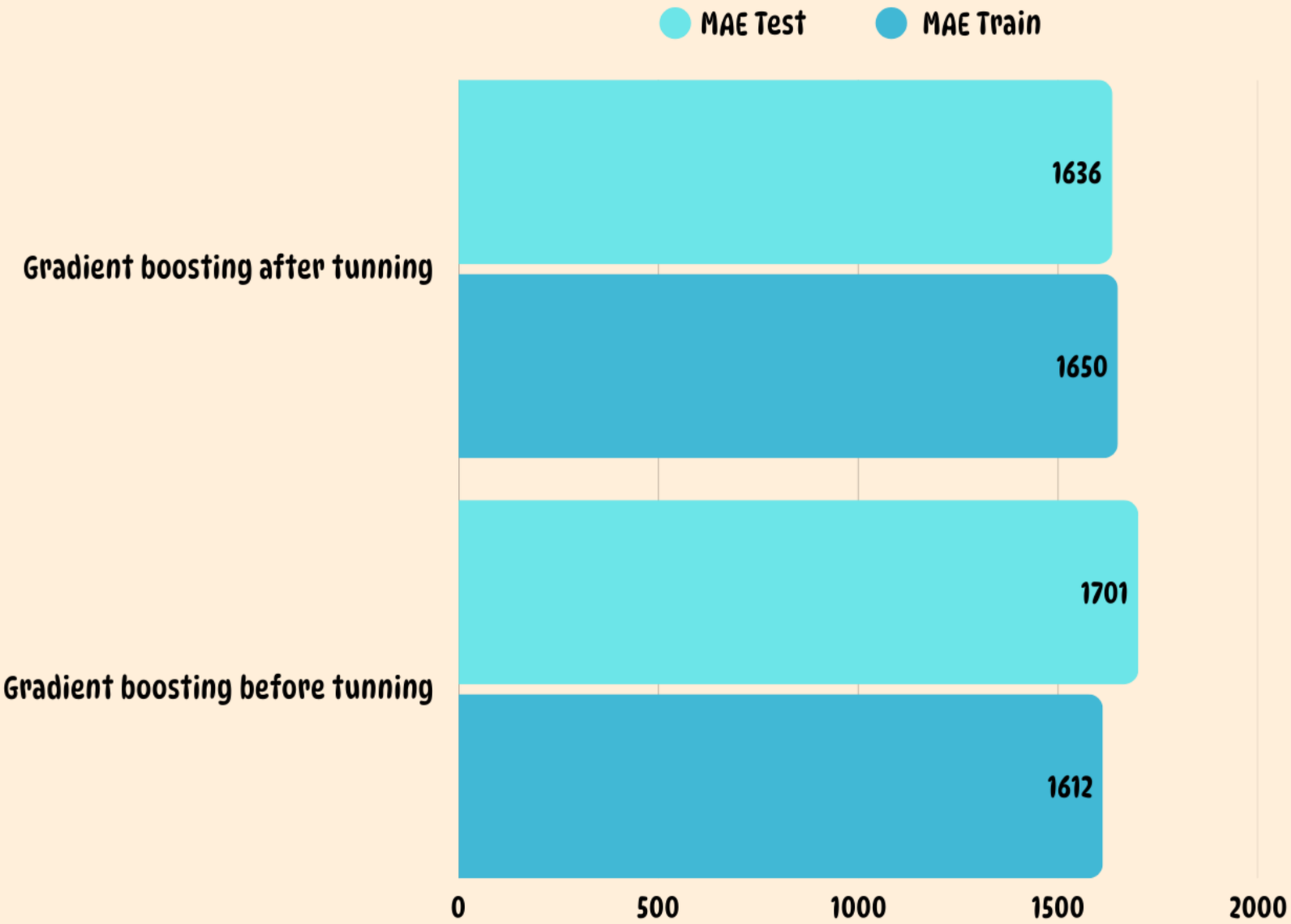
Model	MAE Test	MAE Train	MAPE Test	MAPE Train
Random Forest after tunning	1712.683534	1458.894361	0.141554	0.115170
Random Forest before tunning	1600.321820	635.340340	0.124230	0.046543



Meskipun performa test set terlihat menurun, hyperparameter tuning berhasil mengurangi overfitting secara signifikan

Perbandingan hasil MAE dan MAPE di train dan test dalam model Gradient Boosting

Model	MAE Test	MAE Train	MAPE Test	MAPE Train
Gradient boosting after tunning	1636.253195	1649.556057	0.135514	0.130715
Gradient boosting before tunning	1700.790328	1611.949553	0.124230	0.130031

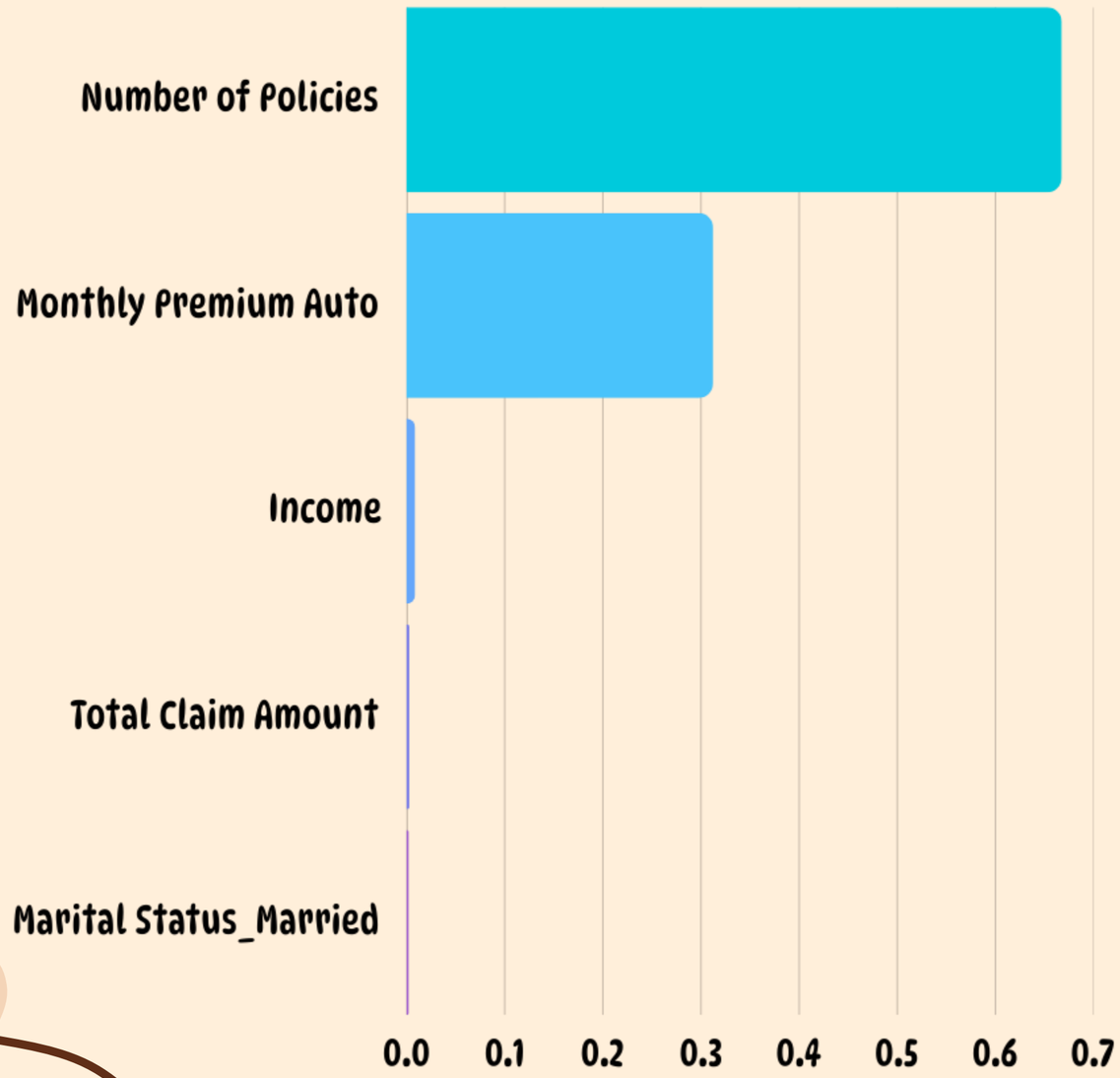


- Model **sesudah tuning** menunjukkan **performa yang lebih baik pada data uji** yang berarti generalisasi model membaik.
- **Slight overfitting sebelumnya berkurang,** ditunjukkan dengan selisih error train dan test yang lebih seimbang

Hypermarameter Tuning
Model Gradient Boosting

```
hyperparameter = {
    'model__n_estimators': [100, 200],
    'model__learning_rate': [0.01, 0.05, 0.1],
    'model__max_depth': [3, 5],
    'model__min_samples_split': [10, 20],
    'model__min_samples_leaf': [5, 10],
    'model__subsample': [0.8, 1.0],
    'model__loss': ['squared_error']
```

FEATURE SELECTION

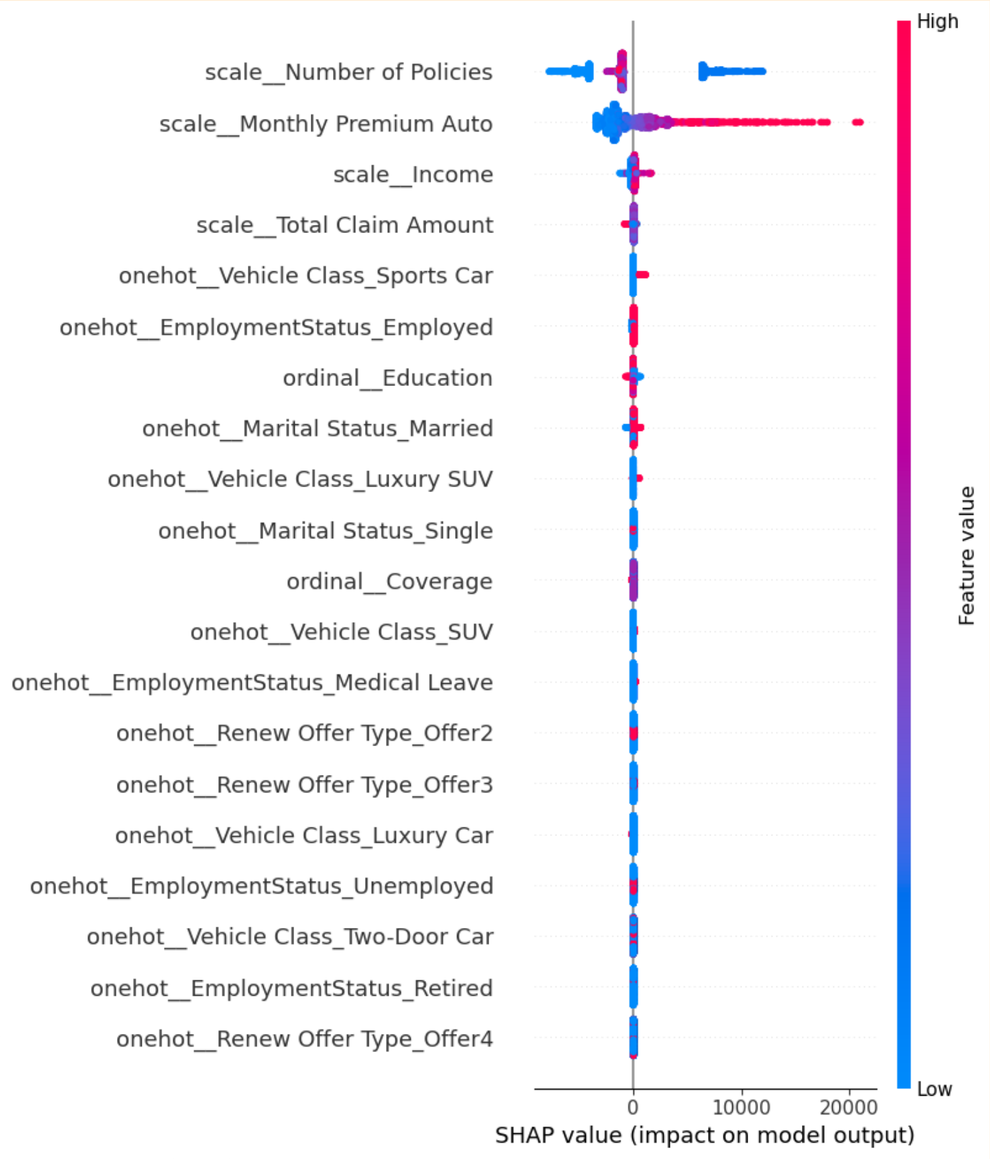


FEATURE IMPORTANCE

Bila dilihat pada grafik terdapat 3 fitur yang cukup berpengaruh pada target. Ketiga feture tersebut adalah `**Number of policies**`, `**monthly premium auto**` dan `**income**`.

Feature pada model bisa diubah menjadi 3. Namun,secara umum Gradient Boosting (seperti GradientBoostingRegressor, XGBoost, LightGBM, dll.) tidak wajib melakukan feature selection.

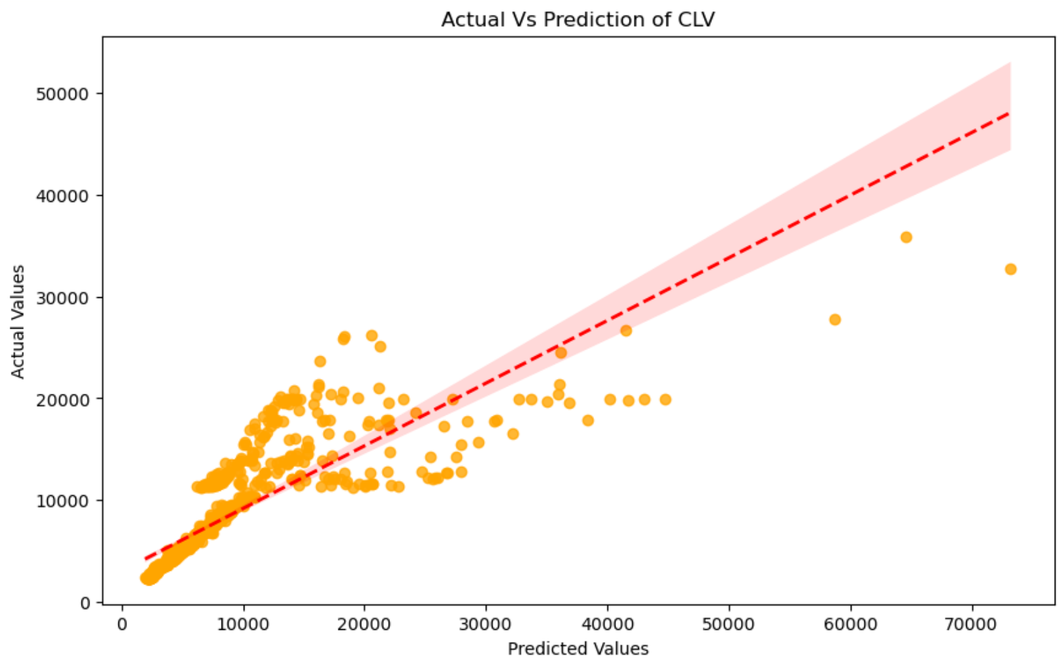
MODEL INTERPRETATION

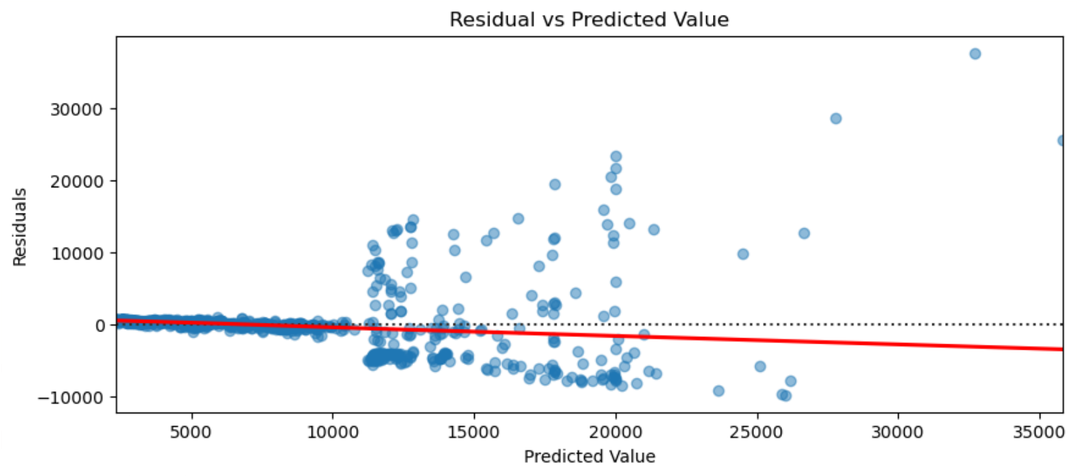


ANALYSIS

`Number of Policies` dan **`Monthly Premium Auto`** menampilkan variasi SHAP value yang signifikan, mengindikasikan pengaruh kuat terhadap hasil prediksi model. Fitur-fitur lain seperti **`income`**juga memberikan kontribusi meskipun tidak sebesar kedua fitur utama tersebut.

Temuan ini dapat dimanfaatkan untuk menyusun strategi bisnis yang lebih efektif, contohnya dengan mengoptimalkan penawaran produk berdasarkan Monthly Premium Auto atau merancang program insentif khusus bagi pelanggan yang memiliki lebih banyak polis asuransi.





ANALYSIS

Analisis Error Prediksi

Plot residual di atas memperlihatkan perbedaan antara nilai sebenarnya (y_test) dengan hasil prediksi model (y_pred_Test_after). Berikut temuan kunci dari visualisasi tersebut:

Pola Sebaran Error

- Untuk nilai CLV rendah hingga sedang, error cenderung terkumpul di sekitar nol, menunjukkan akurasi prediksi yang baik pada rentang ini
- Namun pada nilai CLV tinggi, error menjadi lebih tersebar dan tidak konsisten

Masalah Heteroskedastisitas

- Terdapat kecenderungan error yang melebar seiring peningkatan nilai prediksi CLV

CONCLUSION

Project ini telah membangun model Machine Learning untuk memprediksi target Customer Lifetime Value (CLV), yang bertujuan membantu perusahaan dalam mengoptimalkan strategi bisnis berbasis data dengan. Berikut adalah uraiannya;

1. Akurasi Model (MAE dan MAPE)

Model Gradient Boost setelah hyperparameter tunning merupakan yang terbaik jika ditinjau dari semua metrics penilaian dengan hasilnya adalah sebagai berikut:

- MAE Test : $1636.25
- MAE Train : $1649.55
- MAPE Test: 13.55%
- MAPE Train: 13.07%

yang artinya prediksi kesalahan model sekitar $1,636–1,649 dari nilai asli atau sekitar 13% dari nilai aktual.

2. Fitur yang paling mempengaruhi nilai target (CLV)

Dari analisis yang dilakukan, berdasarkan SHAP dan feature importance kita dapat mengidentifikasi faktor-faktor yang paling berpengaruh terhadap CLV adalah jumlah `**Number of Policies**`, `**Monthly Premium Auto**`, dan `**Income**`.

3. Limitasi/batasan Model

- Untuk nilai CLV rendah hingga sedang, error cenderung terkumpul di sekitar nol, menunjukkan akurasi prediksi yang baik pada rentang ini
- Namun pada nilai CLV tinggi, error menjadi lebih tersebar dan tidak konsisten
- Model mungkin kurang akurat dalam memprediksi pelanggan dengan CLV ekstreme yang nilainya terlalu tinggi (> 15.000).

REKOMENDASI

Rekomendasi Strategis yang Menjawab Rumusan Masalah dan Tujuan Bisnis

1. Gunakan Model Prediktif CLV yang Andal
   - Model Gradient Boosting yang telah dituning menunjukkan performa baik (MAE: $1636.25) dan akurasi stabil pada data test.
   - Model ini cocok digunakan dalam skala operasional untuk memperkirakan nilai pelanggan secara akurat.

2. Manfaatkan Fitur-Fitur Paling Berpengaruh dalam Strategi Pemasaran
   Berdasarkan analisis SHAP dan feature importance, fitur yang paling berpengaruh terhadap CLV adalah:
   - Number of Policies: Pelanggan dengan banyak polis dapat diberi penawaran bundling.
   - Monthly Premium Auto: Pelanggan dengan premi tinggi perlu strategi retensi lebih kuat.
   - Income: Penyesuaian produk dan premi sesuai daya beli meningkatkan efektivitas pemasaran.

3. Minimalkan Kerugian Finansial dengan Integrasi Model CLV
   - Tanpa CLV, perusahaan berisiko salah sasaran dalam program retensi.
   - Estimasi kerugian menggunakan model. Penerapan model jelas menjadi solusi penghematan, menerapkan model terbaik juga akan mempengaruhi penghematan

THANK YOU