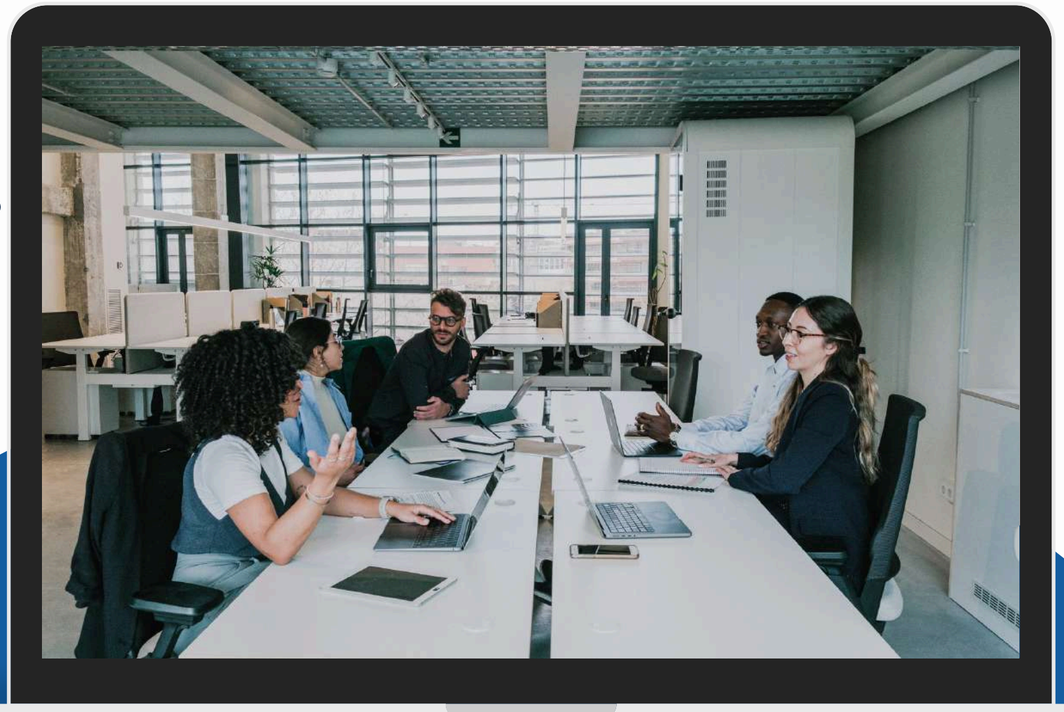
Customer Lifetime Value

Prediction using Machine Learning Algorithm



Giovanni Gunawan Wangidjaja
Capstone Project 3

Table of Contents

Customer Lifetime Value

- Metrix penting untuk **Manajemen Bisnis Modern**
- **Pengambil keputusan strategis** dalam bidang pemasaran, mempertahankan pelanggan, dan pengembangan produk.
- **Total nilai pengeluaran** pelanggan kepada perusahaan sepanjang waktu. Interaksi bisnis yang terjalin antara pelanggan dan perusahaan tersebut berlangsung secara berkelanjutan.
- **Mengakuisisi pelanggan baru** dapat memerlukan biaya hingga 25 kali lipat lebih tinggi dibandingkan dengan mempertahankan pelanggan yang sudah ada.

Kenapa CLV itu penting?

- **Meningkatkan** pendapatan Perusahaan
- **Membantu penganggaran**
- **Menganalisis Kepuasan** Pelanggan

Business Problem

- Pemasaran (Marketing): Mengarahkan promosi pada pelanggan dengan potensi keuntungan tinggi dan mengidentifikasi segmen bernilai.
- Layanan Pelanggan (Customer Service): Menyesuaikan metode pelayanan berdasarkan profil pelanggan agar sumber daya digunakan efisien.
- Keuangan & Manajemen Risiko (Finance & Risk Management): Menghitung kontribusi keuntungan tiap pelanggan dan menentukan langkah jika pelanggan berhenti.
- Pengembangan Produk (Product Development): Memahami preferensi pelanggan bernilai tinggi untuk meningkatkan loyalitas.

GOALS

Membangun sebuah sistem untuk memprediksi CLV berdasarkan data demografis dan data asuransi mobil pelanggan, sehingga proses pengolahan CLV menjadi otomatis dan mempercepat pengambilan keputusan dalam strategi pemasaran.

Machine Learning Algorithm

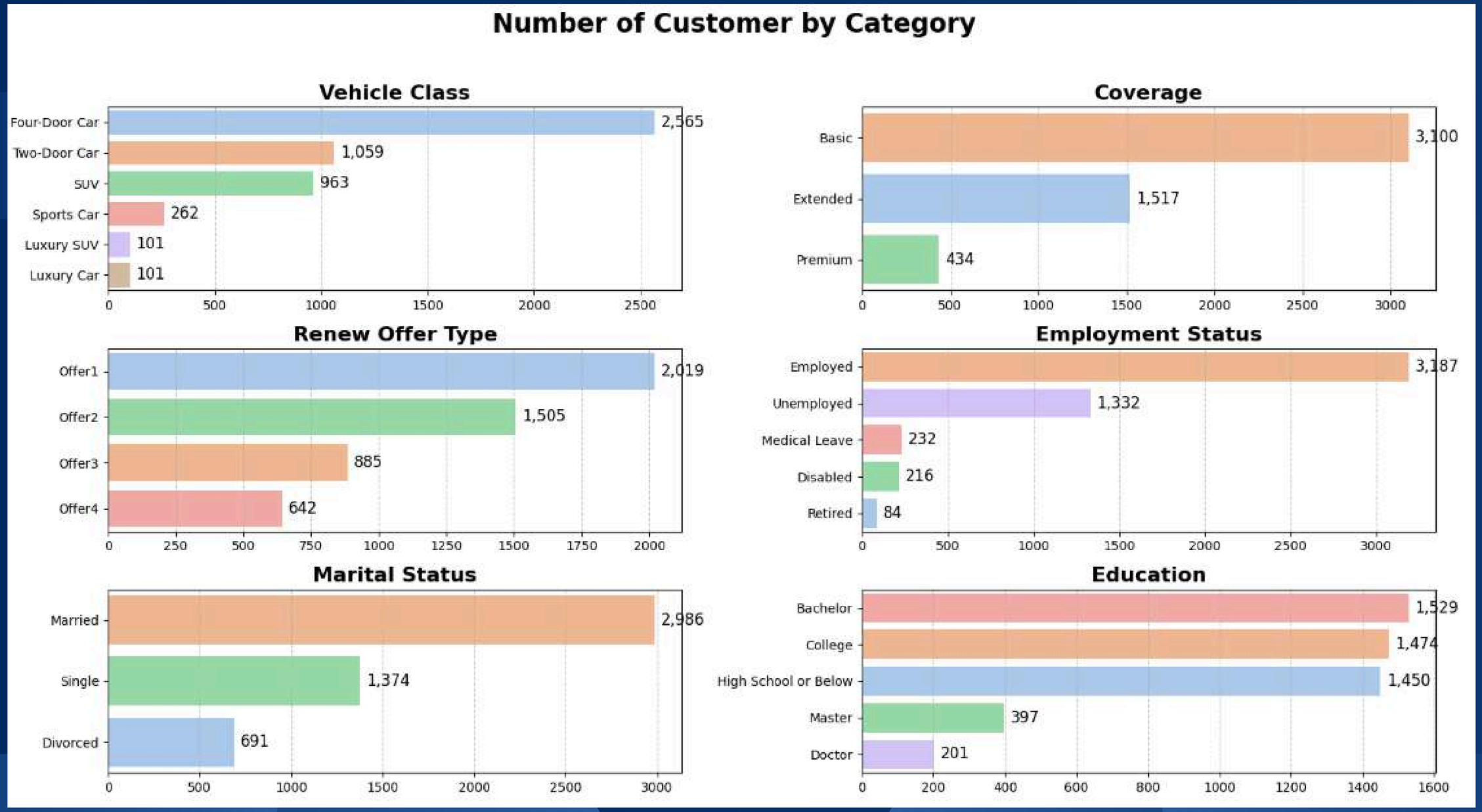
MODEL REGRESI

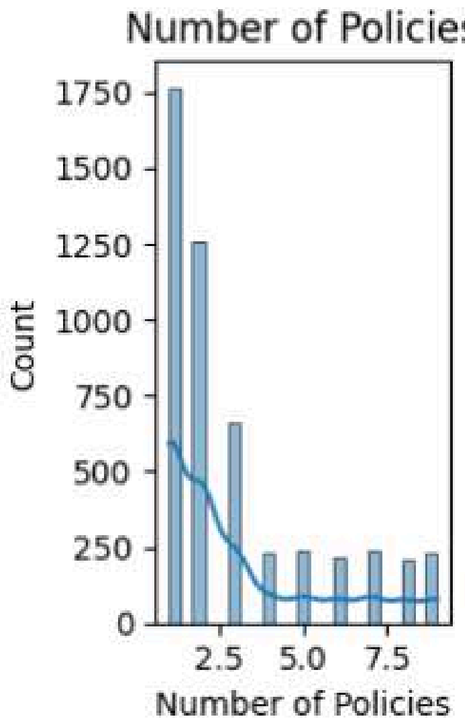
RMSE, MAE, MAPE

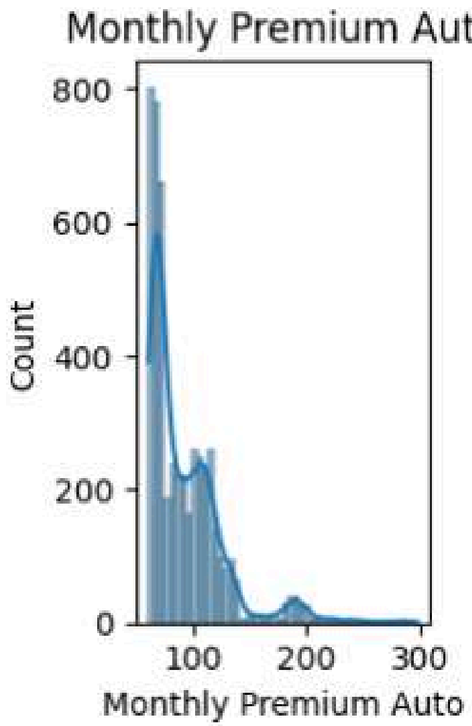
DATA UNDERSTANDING

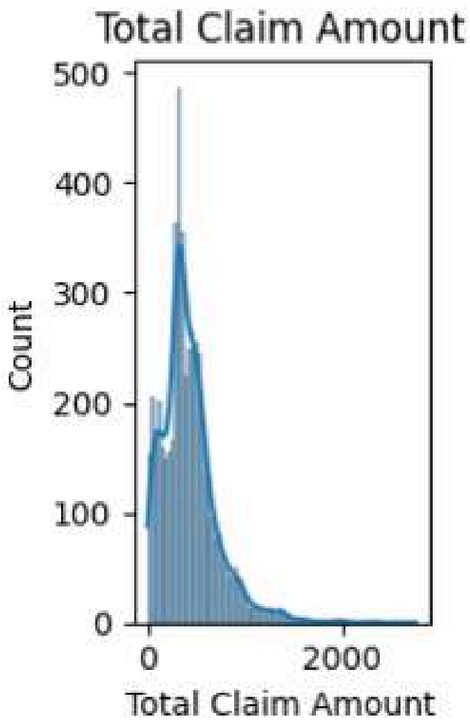
Nama Kolom	Tipe Data	Deskripsi
`Vehicle Class`	*Object*	`Kategori atau jenis kendaraan` yang dimiliki oleh pelanggan (misalnya, "*Four-Door Car*", "*Two-Door Car*").
`Coverage`	*Object*	`Tingkat atau jenis perlindungan asuransi` yang dimiliki oleh pelanggan (misalnya, "*Extended*", "*Basic*", "*Premium*").
`Renew Type Offer`	*Object*	`Jenis penawaran pembaruan` yang diberikan kepada pelanggan (misalnya, "*Offer1*", "*Offer3*").
`EmploymentStatus`	*Object*	`Status pekerjaan` pelanggan (misalnya, "*Retired*", "*Employed*", "*Disabled*", "*Medical Leave*").
`Marital Status`	*Object*	`Status pernikahan` pelanggan (misalnya, "*Divorced*", "*Married*").
`Education`	*Object*	`Tingkat pendidikan` yang telah dicapai oleh pelanggan (misalnya, "*High School or Below*", "*College*", "*Master*").
`Number of Policies`	*Float*	`Jumlah polis asuransi` yang dimiliki oleh pelanggan.
`Monthly Premium Auto`	*Float*	`Jumlah premi bulanan` yang dibayar oleh pelanggan untuk asuransi kendaraan mereka.
`Total Claim Amount`	*Float*	`Jumlah total klaim yang diajukan` oleh pelanggan.
`Income`	*Float*	`Pendapatan pelanggan` (dalam satuan moneter).
`Customer Lifetime Value`	*Float*	`Nilai seumur hidup` pelanggan, yang merupakan perkiraan total pendapatan yang dapat dihasilkan dari pelanggan ini sepanjang hubungan mereka dengan perusahaan.

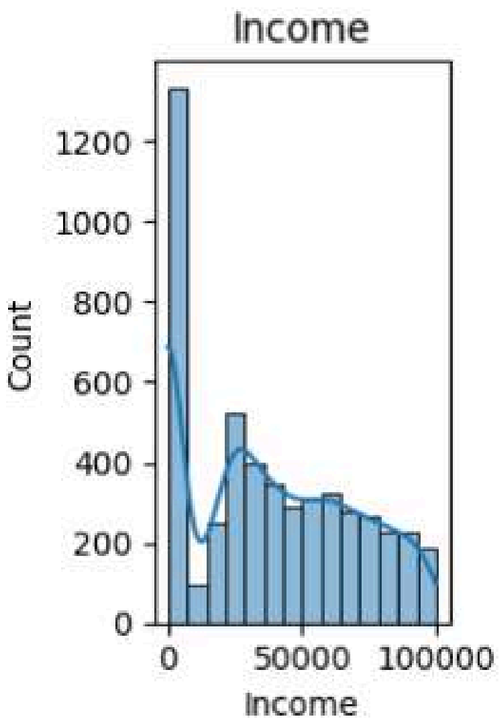
EXPLANITORY DATA ANALYSIS (EDA)

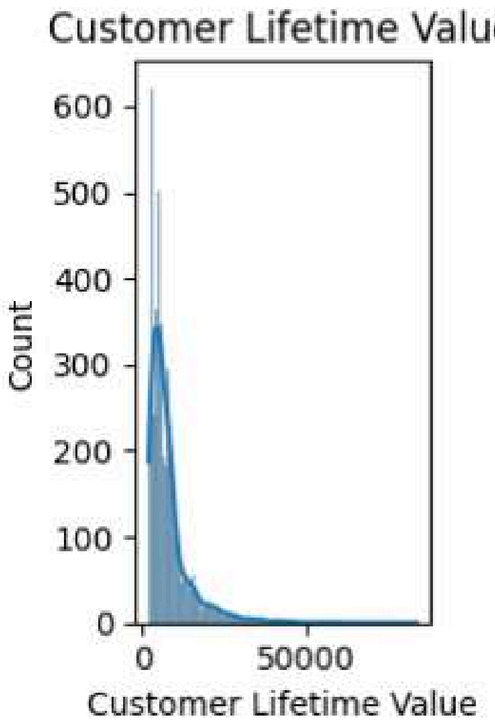

Number of Customer by Category
Vehicle Class
Four-Door Car
Two-Door Car
SUV
Sports Car
Luxury SUV
Luxury Car
2,565
1,059
963
262
101
101
0
500
1000
1500
2000
2500
Coverage
Basic
Extended
Premium
3,100
1,517
434
0
500
1000
1500
2000
2500
3000
Renew Offer Type
Offer1
Offer2
Offer3
Offer4
2,019
1,505
885
642
0
250
500
750
1000
1250
1500
1750
2000
Employment Status
Employed
Unemployed
Medical Leave
Disabled
Retired
3,187
1,332
232
216
84
0
500
1000
1500
2000
2500
3000
Marital Status
Married
Single
Divorced
2,986
1,374
691
0
500
1000
1500
2000
2500
3000
Education
Bachelor
College
High School or Below
Master
Doctor
1,529
1,474
1,450
397
201
0
200
400
600
800
1000
1200
1400
1600



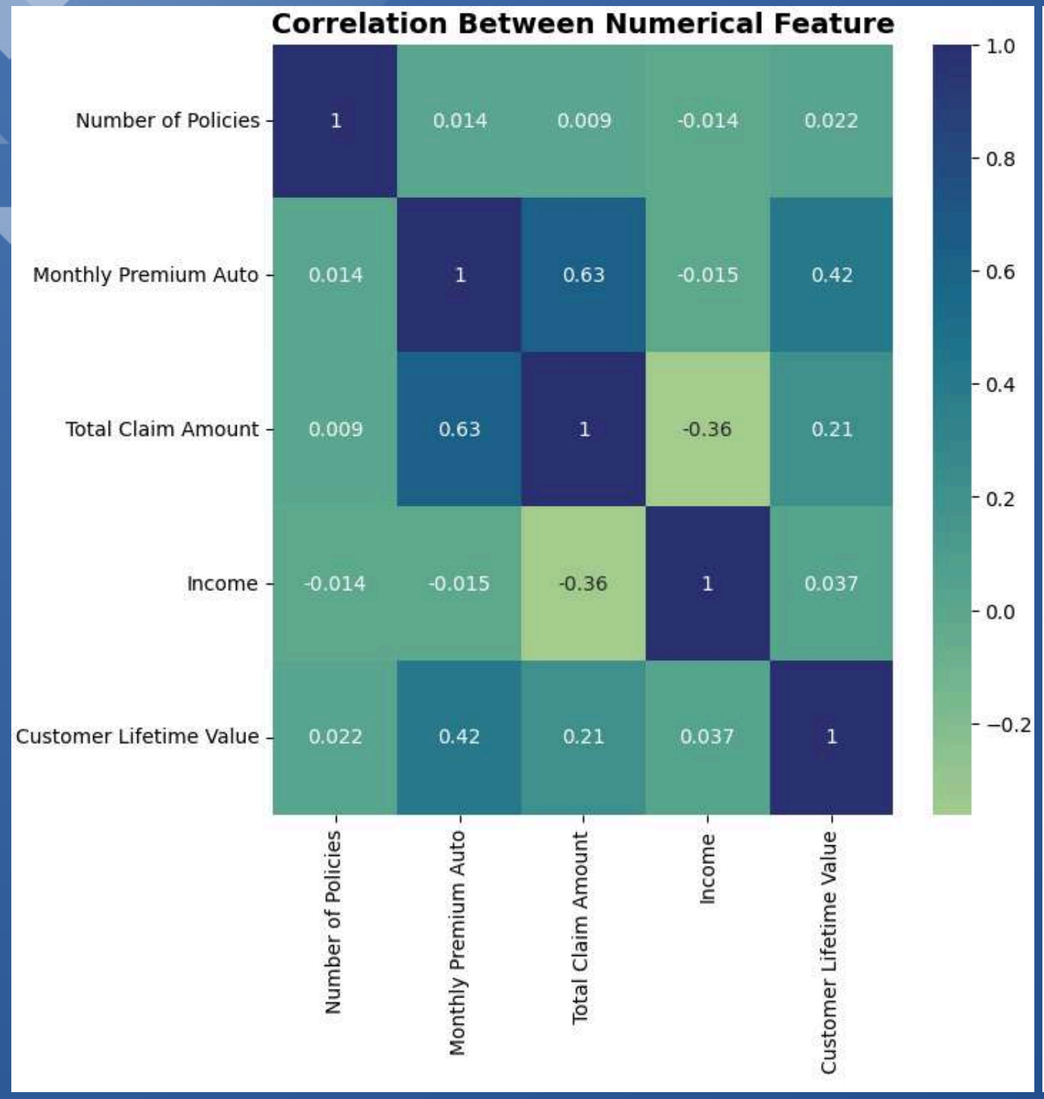


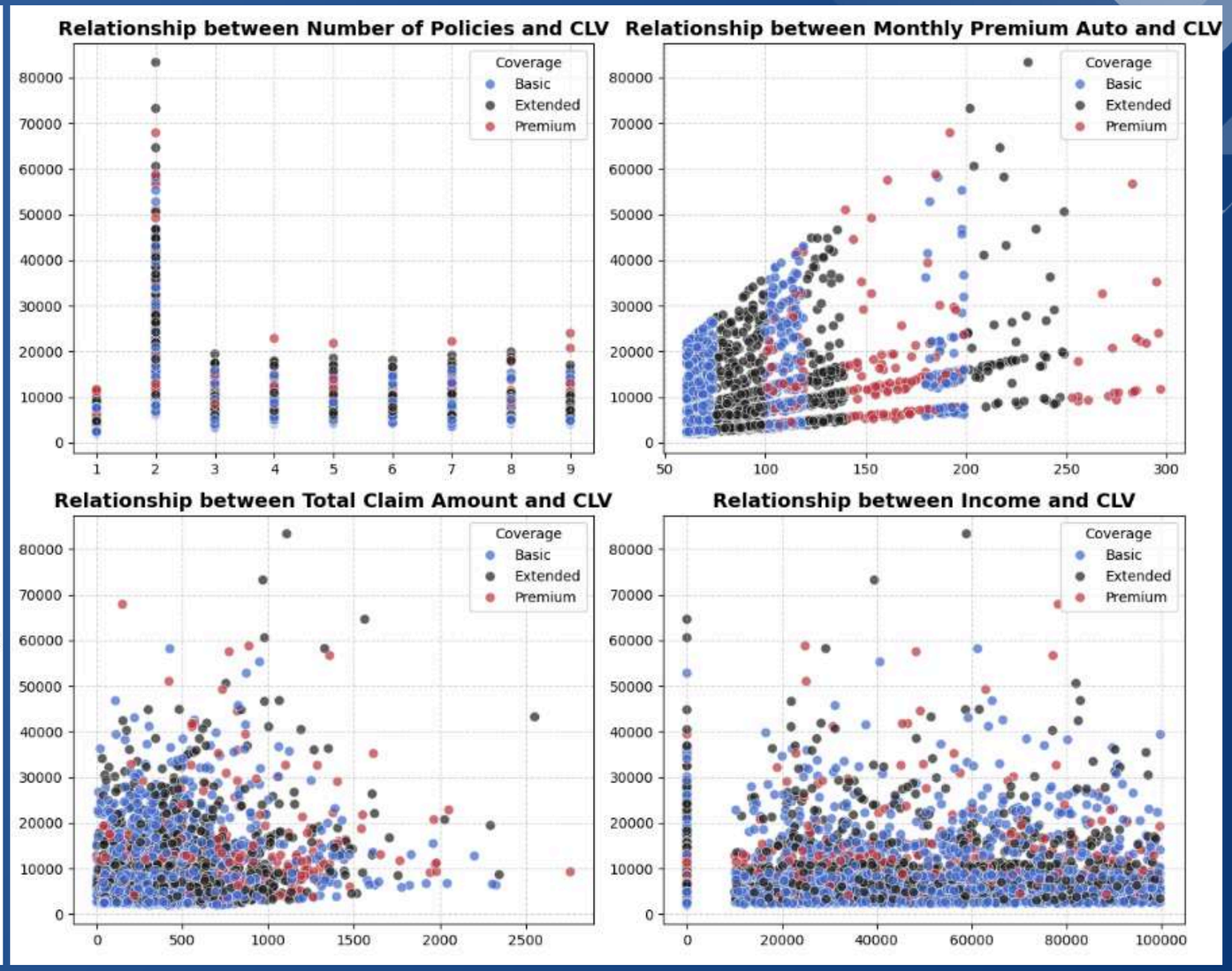




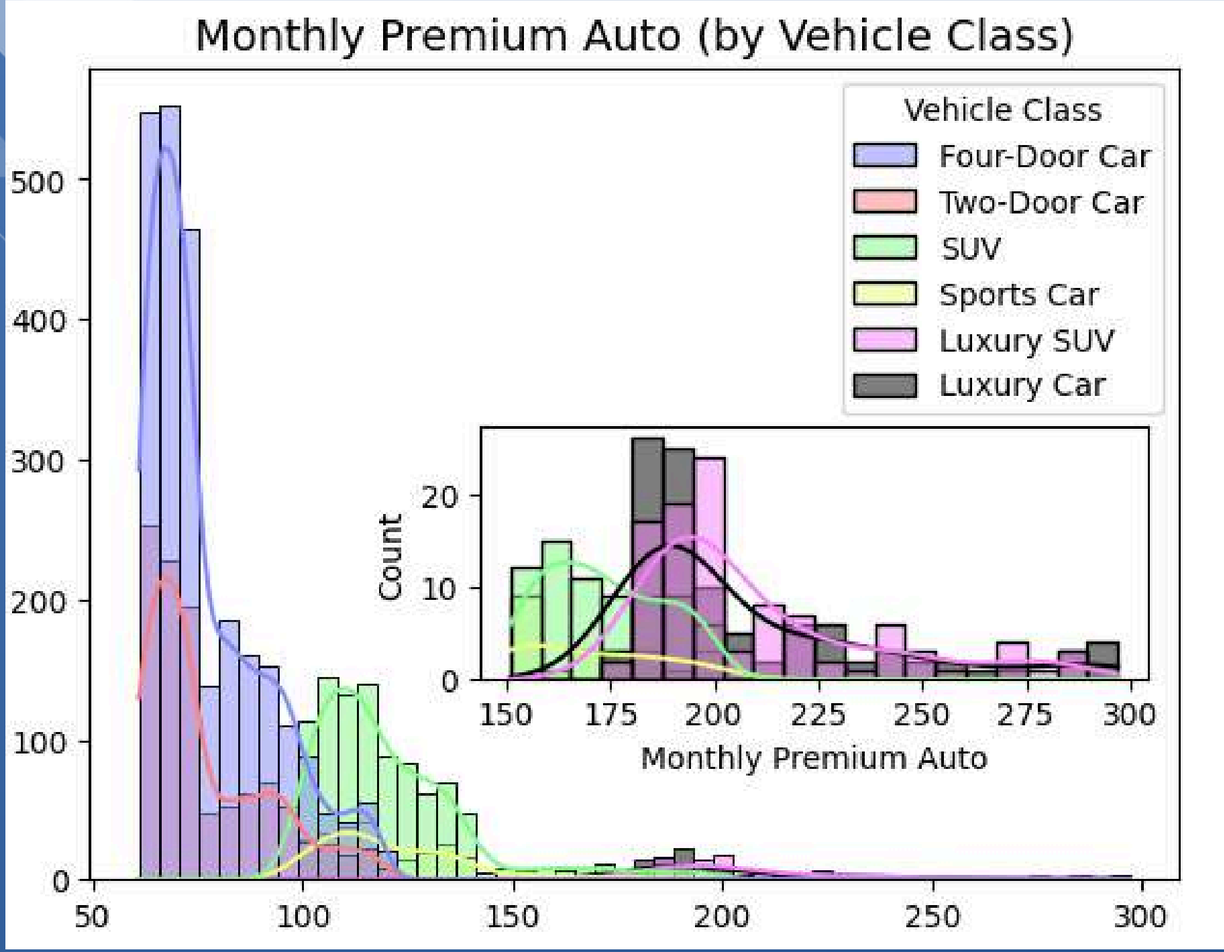


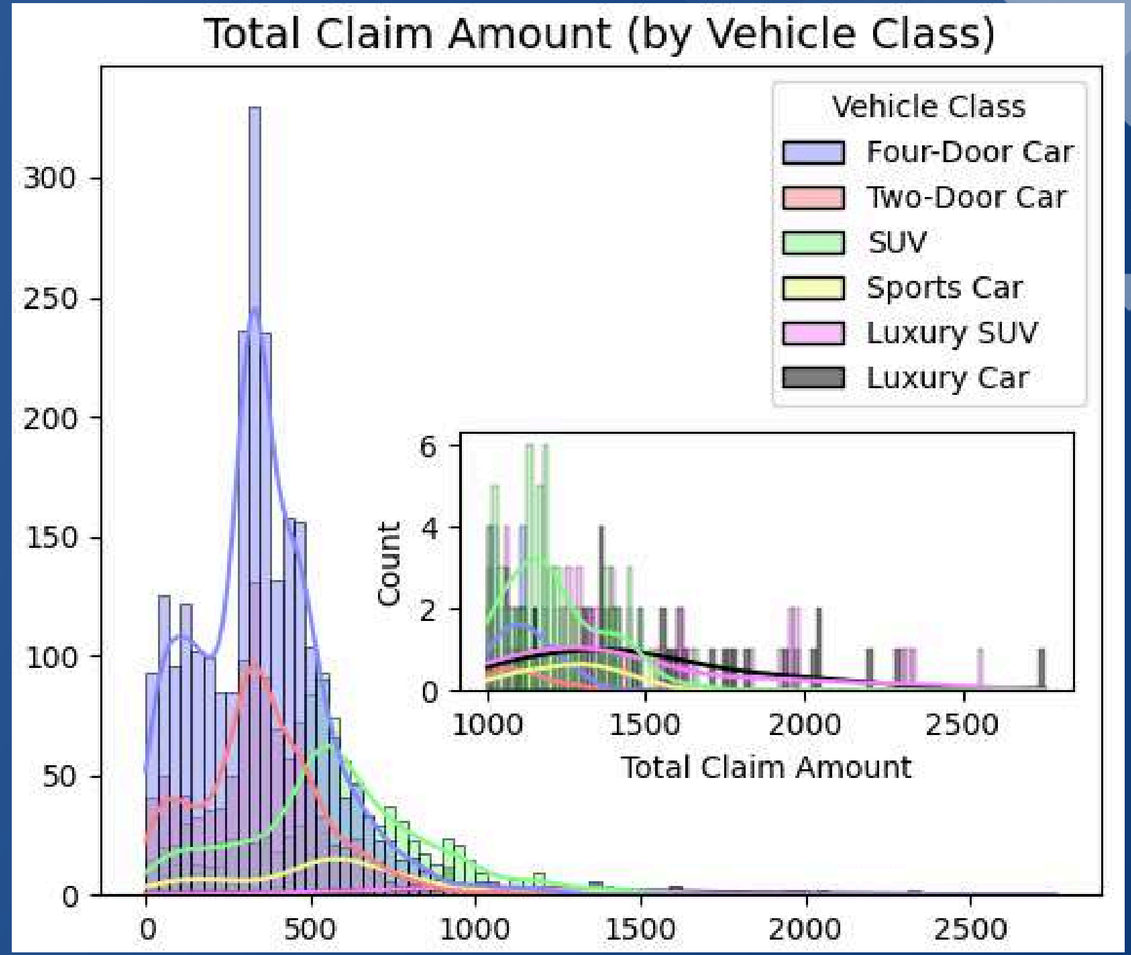
- Banyak customer dengan low CLV
- Terdapat customer dengan CLV yang ekstrim (di atas mediannya 16000)
- Terdapat **outliers** pada kolom **Number of Policies, Monthly Premium Auto, dan Total Claim Amount**



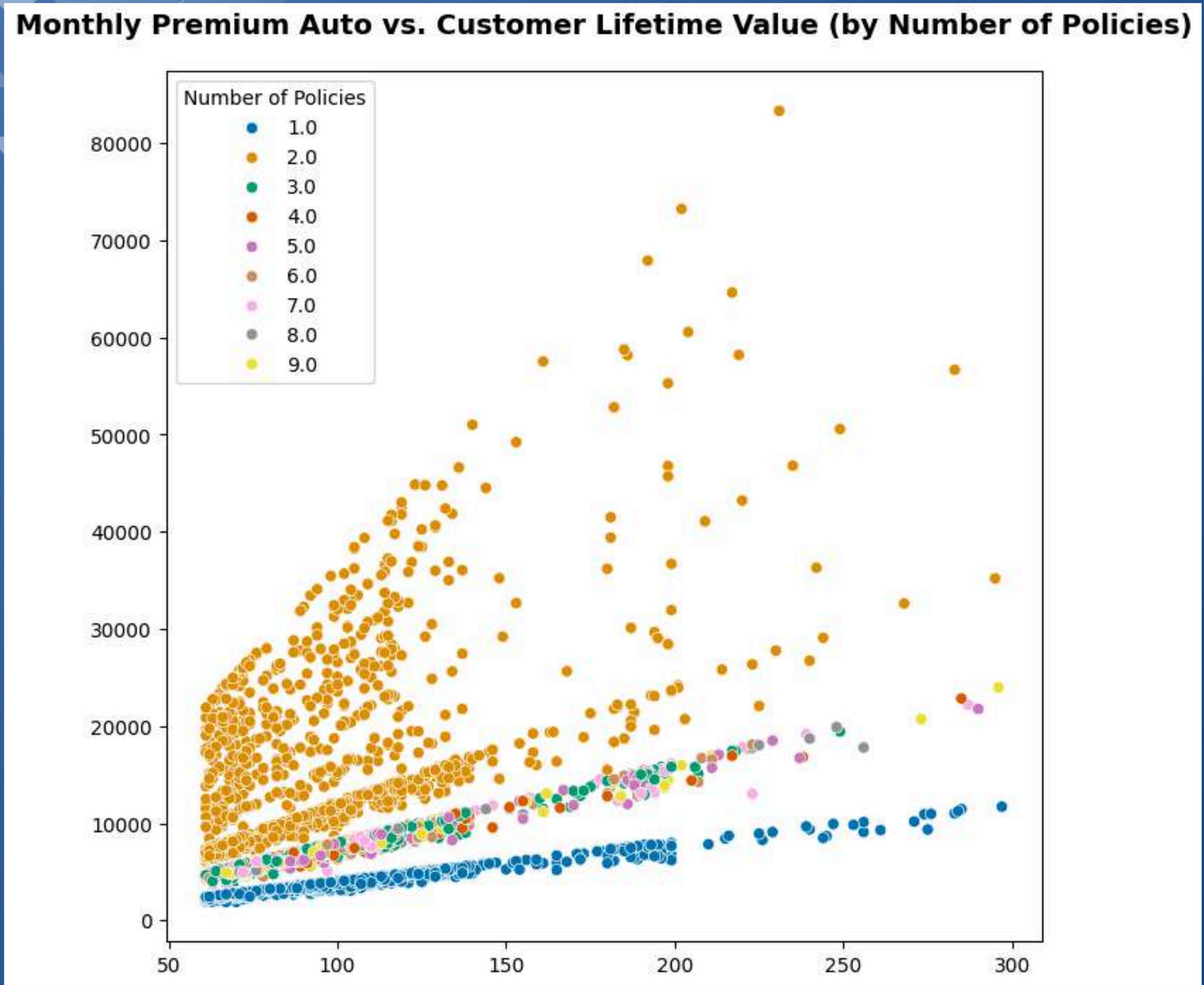


- Kenaikan CLV sejalan dengan kenaikan Number of Policies, Monthly Premium Auto, dan Total Claim Amount





- Mayoritas pelanggan membayar premi bulanan di bawah 150, dengan konsentrasi tertinggi pada kisaran 60-100, sementara premi tinggi (>150) hanya dibayar oleh sebagian kecil pelanggan.
- Premi dan klaim bervariasi berdasarkan kelas kendaraan: mobil mewah dan sport cenderung membayar premi dan mengajukan klaim dengan nilai lebih tinggi dibandingkan kendaraan lain seperti SUV dan 4 pintu.
- Sebagian besar klaim berada di kisaran rendah hingga menengah, namun terdapat klaim outlier bernilai sangat tinggi dari kendaraan mewah yang perlu diperhatikan untuk pengelolaan risiko dan strategi penentuan premi.
- Variasi premi dan klaim antar kelas kendaraan ini penting sebagai dasar segmentasi pasar dan penyesuaian strategi pemasaran serta pengembangan produk asuransi.



- Customer Lifetime Value (CLV) cenderung meningkat seiring kenaikan Monthly Premium Auto dalam kelompok Number of Policies yang sama.
- Pelanggan dengan dua polis menunjukkan nilai CLV tertinggi dibanding kelompok polis lainnya, sementara pelanggan dengan satu polis berada di posisi lebih rendah.
- Nilai CLV berkisar antara sekitar -3.632 hingga 16.625, yang menjadi batasan untuk analisis dan pengembangan model prediksi.

Insight from EDA

- Fitur yang kemungkinan besar berkontribusi pada nilai Customer Lifetime Value adalah Number of Policies dan Monthly Premium Auto.
- Fitur Employment Status, Marital Status, dan Education tidak menunjukkan perbedaan median CLV yang signifikan, sehingga kemungkinan pengaruhnya terhadap model relatif kecil.
- Fitur Vehicle Class, Coverage, dan Renew Offer Type memang menunjukkan variasi median, namun perbedaan ini mungkin dipengaruhi oleh Monthly Premium Auto, sehingga dampaknya terhadap model mungkin terbatas.
- Sebagian besar fitur numerik selain Monthly Premium Auto memiliki nilai korelasi yang rendah (di bawah 0,3), yang mengindikasikan kemungkinan pengaruh yang minimal terhadap model.

MACHINE LEARNING MODELING

Anomali Handling

Missing Value
Outlier
Duplicated
Encoding (categorical feature)

Evaluation Metrics

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

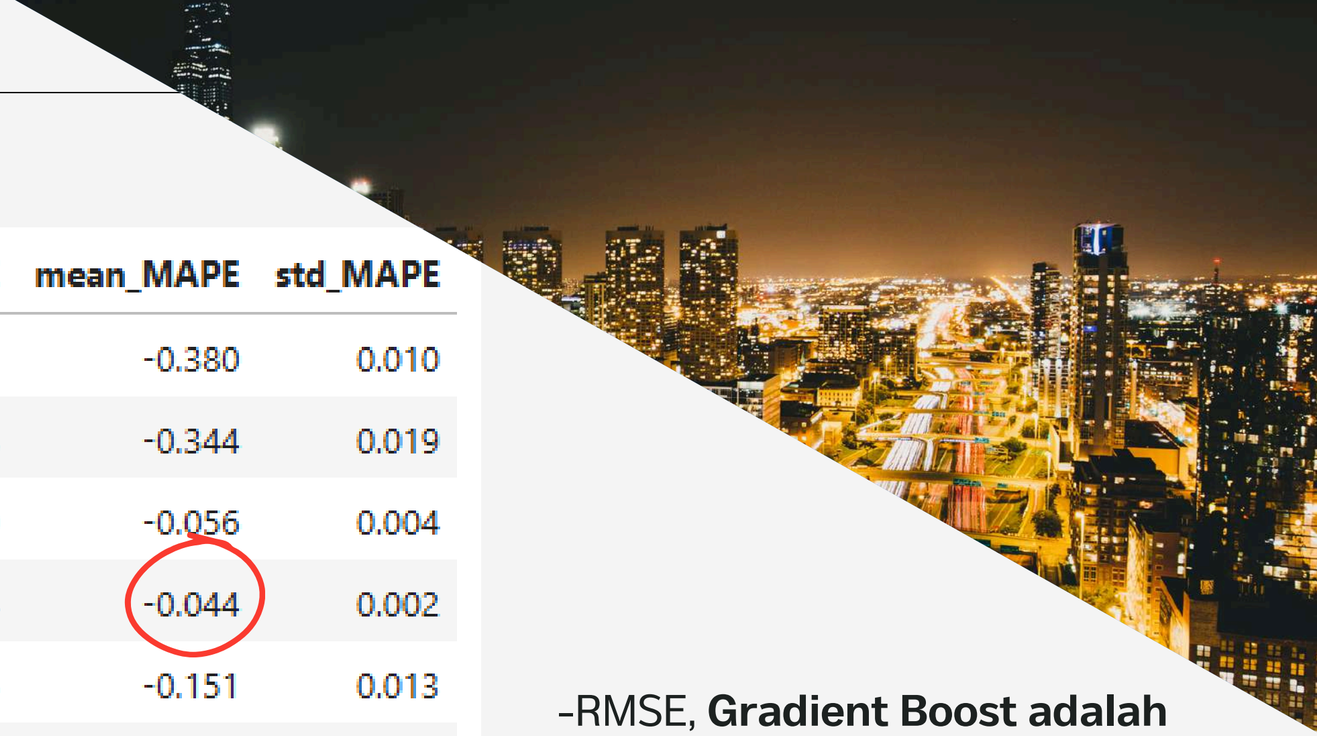
Where:
$\hat{y}_i$ = Predicted value for the i^{th} data point
y_i = Actual value for the i^{th} data point
n = number of observations

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Where:
$\hat{y}_i$ = Predicted value for the i^{th} data point
y_i = Actual value for the i^{th} data point
n = number of observations

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

Where:
$\hat{y}_i$ = Predicted value for the i^{th} data point
y_i = Actual value for the i^{th} data point
n = number of observations



Benchmark Model

	model	mean_RMSE	std_RMSE	mean_MAE	std_MAE	mean_MAPE	std_MAPE
0	Linear Regression	-2764.679	164.294	-2082.132	88.311	-0.380	0.010
1	KNN	-2746.820	111.056	-1876.383	70.298	-0.344	0.019
2	Decision Tree	-1236.737	54.495	-443.326	25.760	-0.056	0.004
3	Random Forest	-927.836	87.652	-362.320	27.474	-0.044	0.002
4	AdaBoost	-1378.767	82.154	-993.204	76.116	-0.151	0.013
5	XGBoost	-1009.139	79.563	-452.196	28.822	-0.068	0.003
6	Gradient Boost	-885.588	124.247	-382.775	32.570	-0.050	0.002
7	Lasso	-2764.678	164.292	-2082.131	88.309	-0.380	0.010
8	Ridge	-2764.679	164.294	-2082.131	88.310	-0.380	0.010

-RMSE, **Gradient Boost adalah model terbaik (-885.558)**

MAE dan MAPE, **RandomForest memiliki nilai paling rendah (-362.320 dan -0.044).**

Predict to Test Set

	RMSE	MAE	MAPE
rf	970.722	376.452	0.045
gbr	942.598	394.817	0.050

Gradient Boost menjadi model terpilih

Hyperparameter Tuning

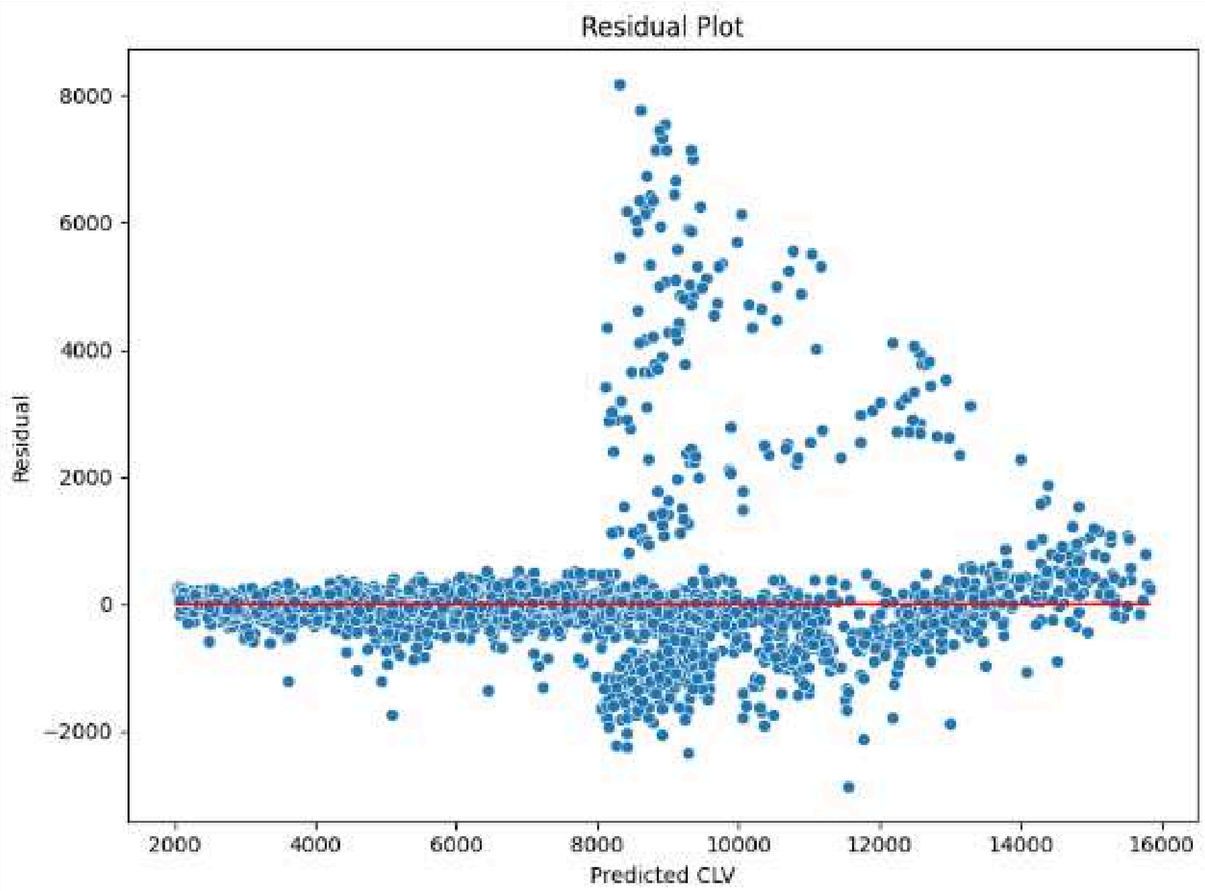
Condition	RMSE	MAE	MAPE
Before Tuning	-885.588	-382.775	-0.050
After Tuning (RandomizedSearch)	-888.020	-371.801	-0.045
After Tuning (GridSearch)	-884.353	-371.801	-0.045

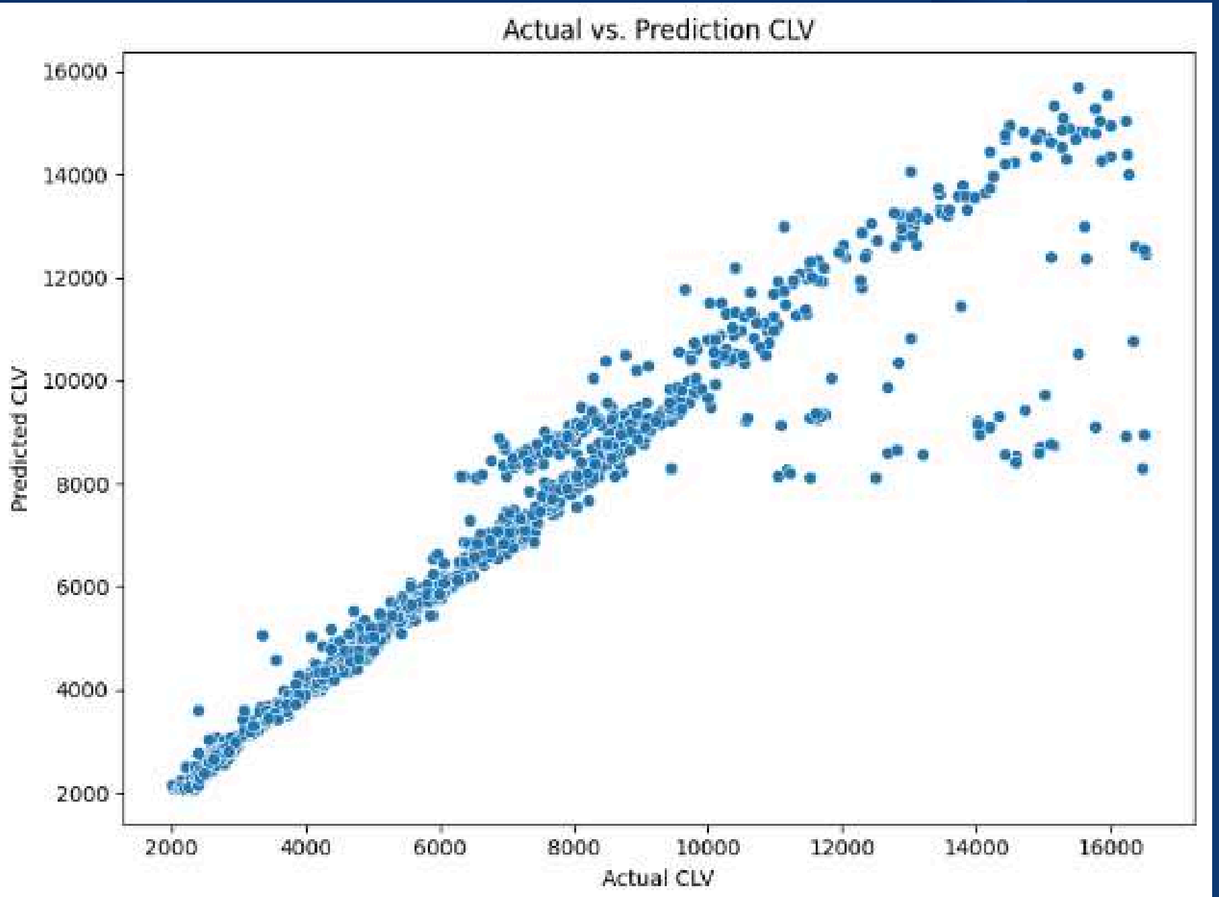
Tuning menggunakan **GridSearch menghasilkan performa model yang lebih baik**.

Predict to Test Set with Tuned Model

- RMSE, MAE & MAPE sebelum tuning: 942.598, 394.817, dan 0.05
- RMSE, MAE & MAPE setelah tuning: 941.003, 380.672, dan 0.047

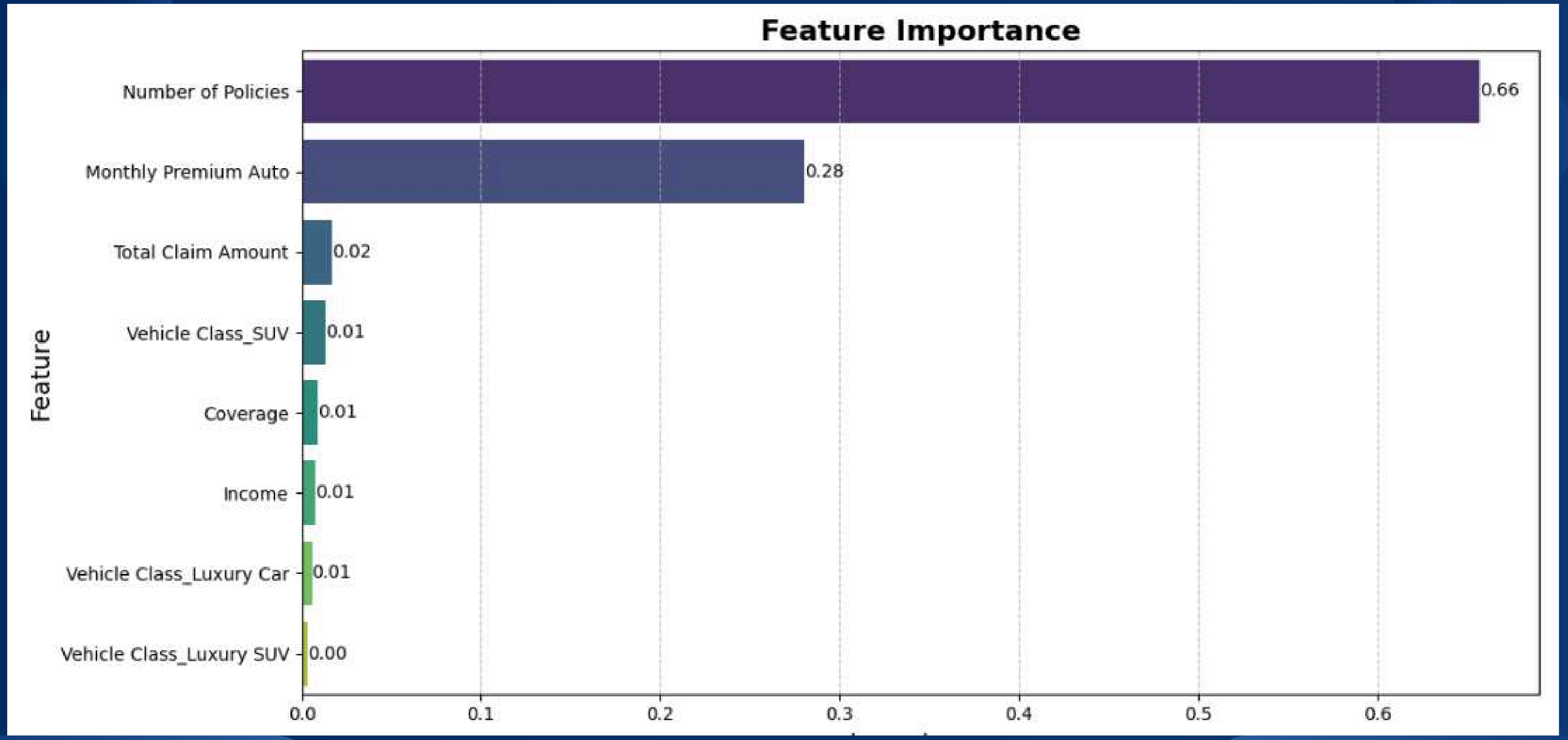
Evaluation Hasil Prediksi by Residual Plot





- Model prediksi cukup akurat untuk nilai CLV di bawah 8000, dengan error yang mendekati nol.
- Untuk CLV di atas 8000, variansi residual tidak seragam atau konsisten.
- Grafik Actual vs Predicted CLV menunjukkan pola mendekati garis lurus, menandakan prediksi yang baik.
- Terdapat beberapa outlier mulai dari kisaran ±9000 yang mempengaruhi hasil.

Important Feature

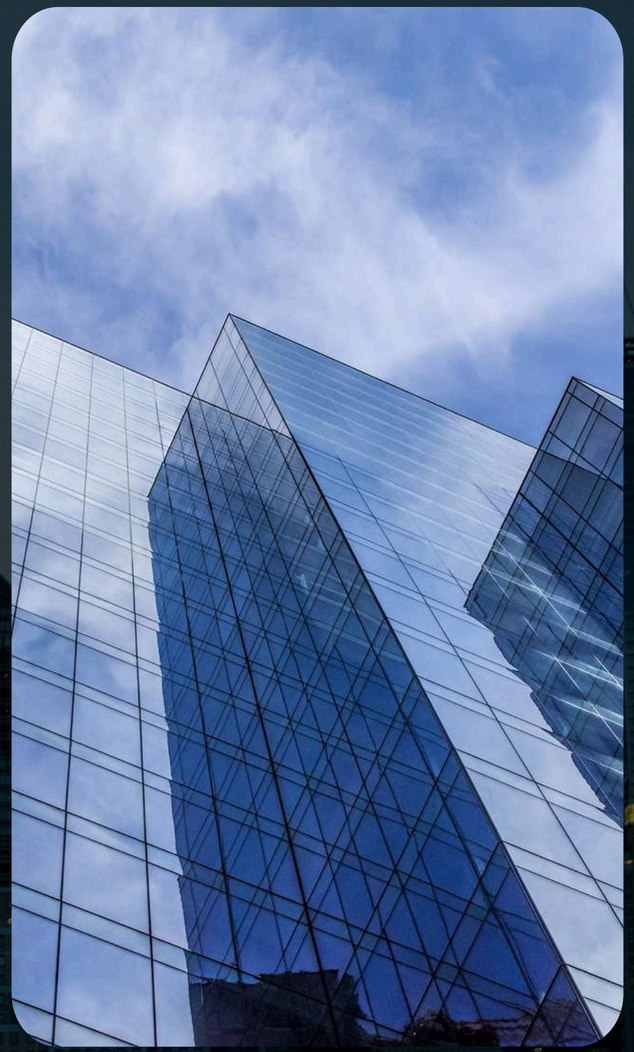


Number of Policies fitur paling penting

KESIMPULAN DAN RECOMENDATION

Kesimpulan

1. Model Terbaik
   - Gradient Boosting dipilih sebagai model akhir dengan performa terbaik setelah hyperparameter tuning.
   - MAE terendah: ±364
   - MAPE terendah: ±4,4%
   - RMSE relatif stabil meskipun sedikit meningkat setelah tuning.
2. Faktor Penting Penentu CLV
   - Monthly Premium Auto, Number of Policies, dan Total Claim Amount berpengaruh signifikan.
   - Pelanggan dengan premi tinggi, polis banyak, dan klaim besar cenderung memiliki CLV lebih tinggi.
3. Distribusi Pelanggan
   - Segmen Premium mendominasi kelompok dengan CLV tinggi.
   - Dua polis merupakan titik optimum CLV tertinggi, meski terdapat beberapa outlier.
4. Kelemahan Model
   - Model akurat untuk CLV < 8000, namun error membesar untuk CLV tinggi (>8000).
   - Mengindikasikan adanya heteroskedastisitas dan outlier yang mempengaruhi kestabilan prediksi.



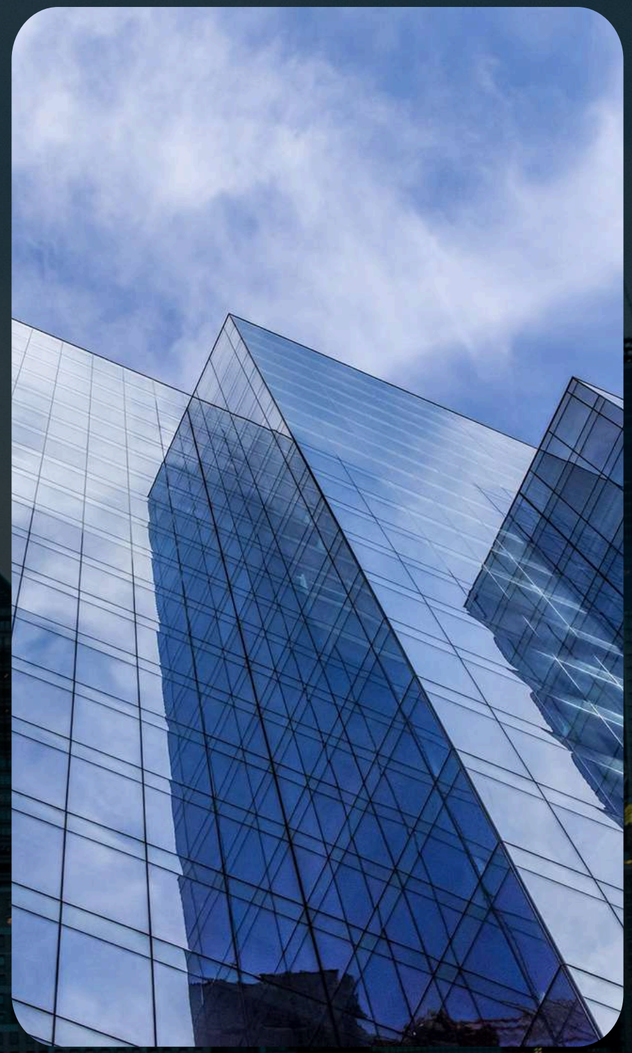
Recomendation

Pengembangan Model Machine Learning

- Analisis prediksi dengan error terbesar untuk identifikasi pola dan fitur penyebab.
- Tambahkan fitur relevan seperti lama menjadi nasabah, riwayat klaim, lokasi, jenis polis, dan interaksi pelanggan.
- Perbanyak data pelatihan untuk representasi perilaku pelanggan lebih baik.
- Gunakan Gradient Boosting sebagai baseline, eksplorasi ensemble stacking dan clustering untuk segmentasi pelanggan.

Strategi Bisnis

- Sesuaikan penawaran berdasarkan Number of Policies dan Monthly Premium Auto untuk optimalkan anggaran pemasaran dan jaga loyalitas pelanggan bernilai tinggi.
- Manfaatkan prediksi CLV untuk strategi upselling dan cross-selling guna tingkatkan lifetime value dan profitabilitas.




Thank You