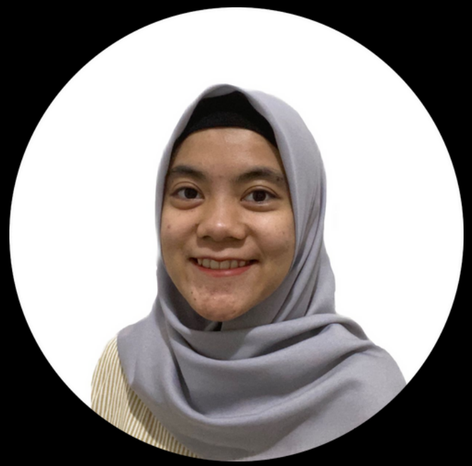
HEALTH INSURANCE CROSS SELL PREDICTION

Presentasi oleh **Data Connector**

RAKAMIN ACADEMY | 2023







Presentasi oleh **Data Connector**

RAKAMIN ACADEMY | 2023

LIST OF CONTENTS

1 BACKGROUND

- WHAT IS THE PROBLEM?
- OUR ROLE
- OBJECTIVE, GOALS, BUSINESS METRICS

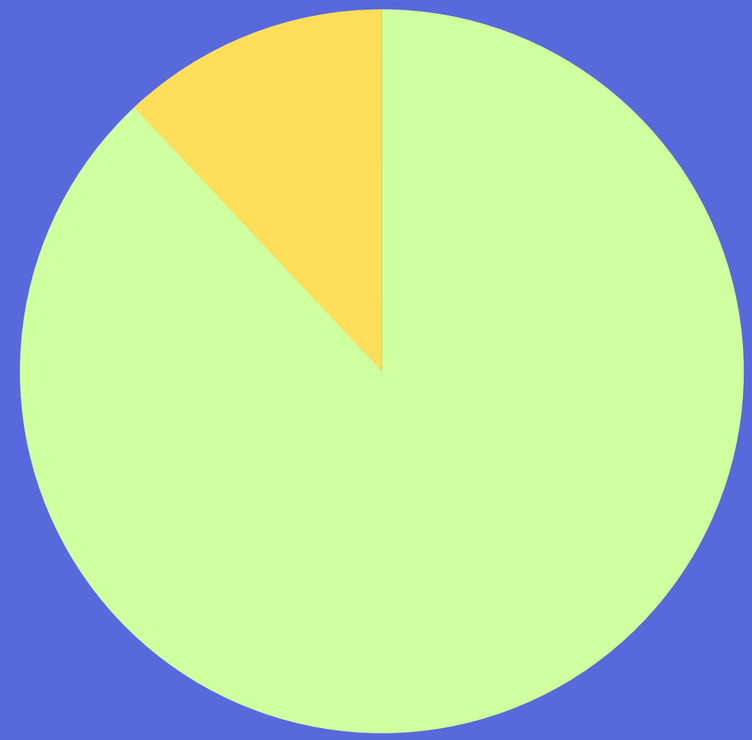
WHAT IS THE PROBLEM?

Raka Insurance
Perusahaan asuransi Kesehatan.

Menawarkan **Asuransi Kendaraan** Kepada Customer.

Diperoleh data **hanya** 12% dari 381.109 customer yang tertarik pada asuransi tersebut.

Vehicle Insurance Interested Rate



OUR ROLE

DATA CONNECTOR

EST - 2023

Data Connector

adalah tim Data Science perusahaan Raka Insurance yang bertugas :

- **Membangun Model Machine Learning**
- **Memprediksi Profil Pelanggan Asuransi**
- **Merencanakan Strategi Komunikasi**
- **Mengoptimalkan model bisnis dan pendapatan**







OBJECTIVE, GOALS, BUSINESS METRICS

Goals

Menigkatkan *Conversion Rate dan* Revenue Perusahaan

Business Metrics

- Conversion Rate
- Revenue

Objective

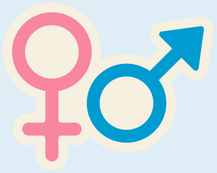
- **Membuat model** Machine Learning untuk meningkatkan conversion rate
- **Memberikan rekomendasi** analisis berdasarkan Exploratory Data Analysis

2 EXPLORATORY DATA ANALYS

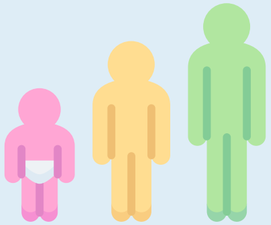
- FEATURES
- MULTIVARIATE ANALYSIS
- BUSINESS INSIGHT
- MISSING DATA & SHAPE
- DATA OUTLIER
- CORRELATION FEATURES

FEATURES

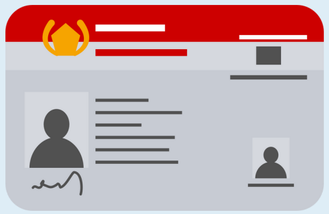
ID



GENDER



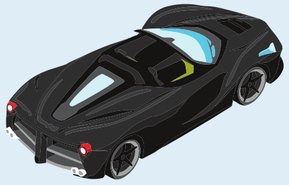
AGE



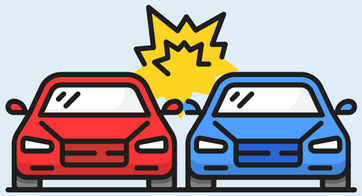
DRIVING LICENSE

REGION CODE

PREVIOUSLY INSURED



VEHICLE AGE



VEHICLE DAMAGE

ANNUAL PREMIUM

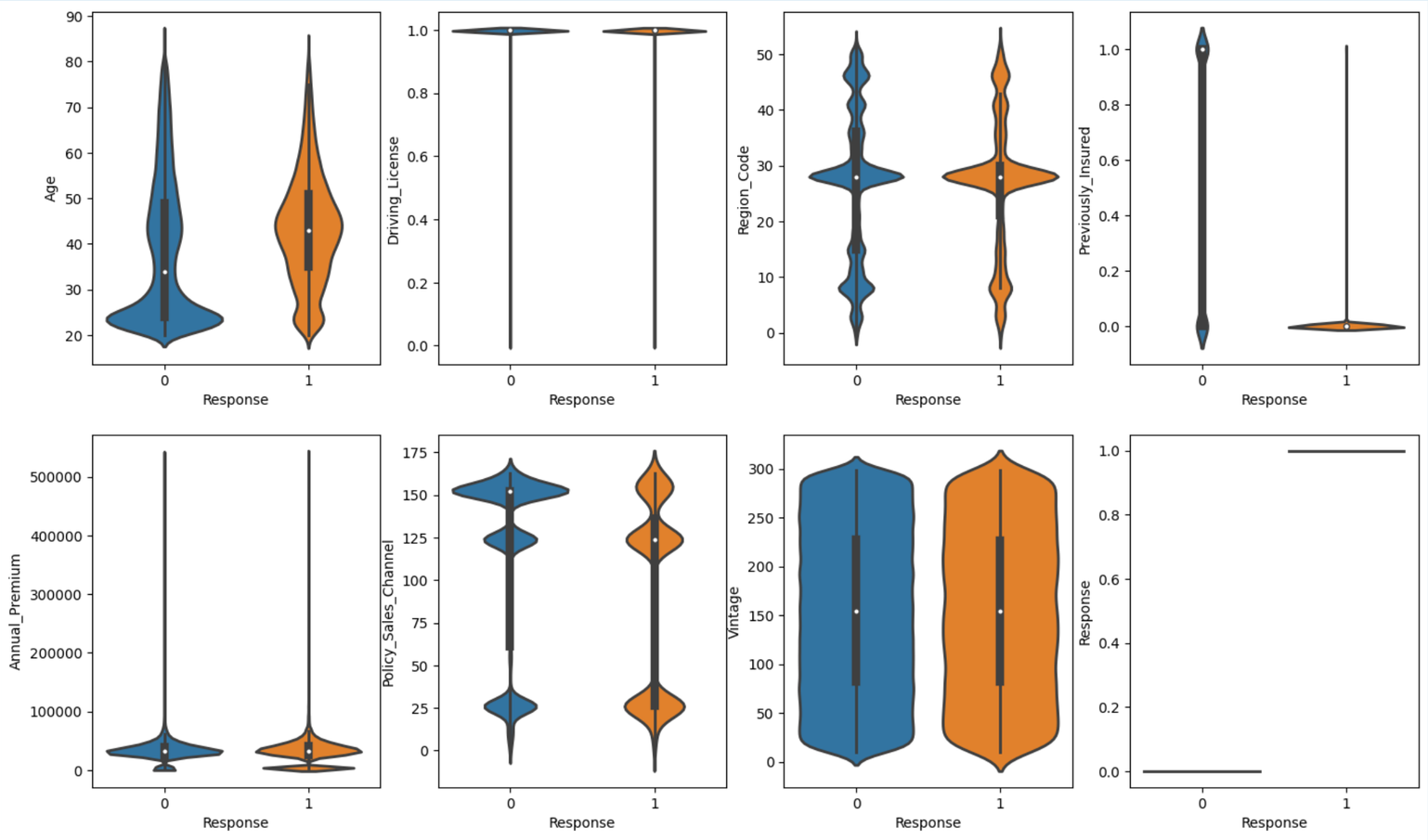
POLICY SALES CHANNEL

VINTAGE

RESPONSE

MULTIVARIATE ANALYSIS

VIOLIN PLOT



Karakteristik customer cukup seragam, antara customer yang tertarik dan tidak tertarik dengan asuransi kendaraan

Perbedaan yang dapat dihighlight:
Median umur customer yang tidak tertarik asuransi kendaraan lebih rendah dari yang tertarik dengan asuransi kendaraan.

Dan untuk chart fitur policy_sales_channel, jika dilihat perbedaan bentuk distribusi di bagian atas chart, mengindikasikan bahwa customer yang tidak tertarik asuransi kendaraan terkonsentrasi di saluran pemasaran 150-175

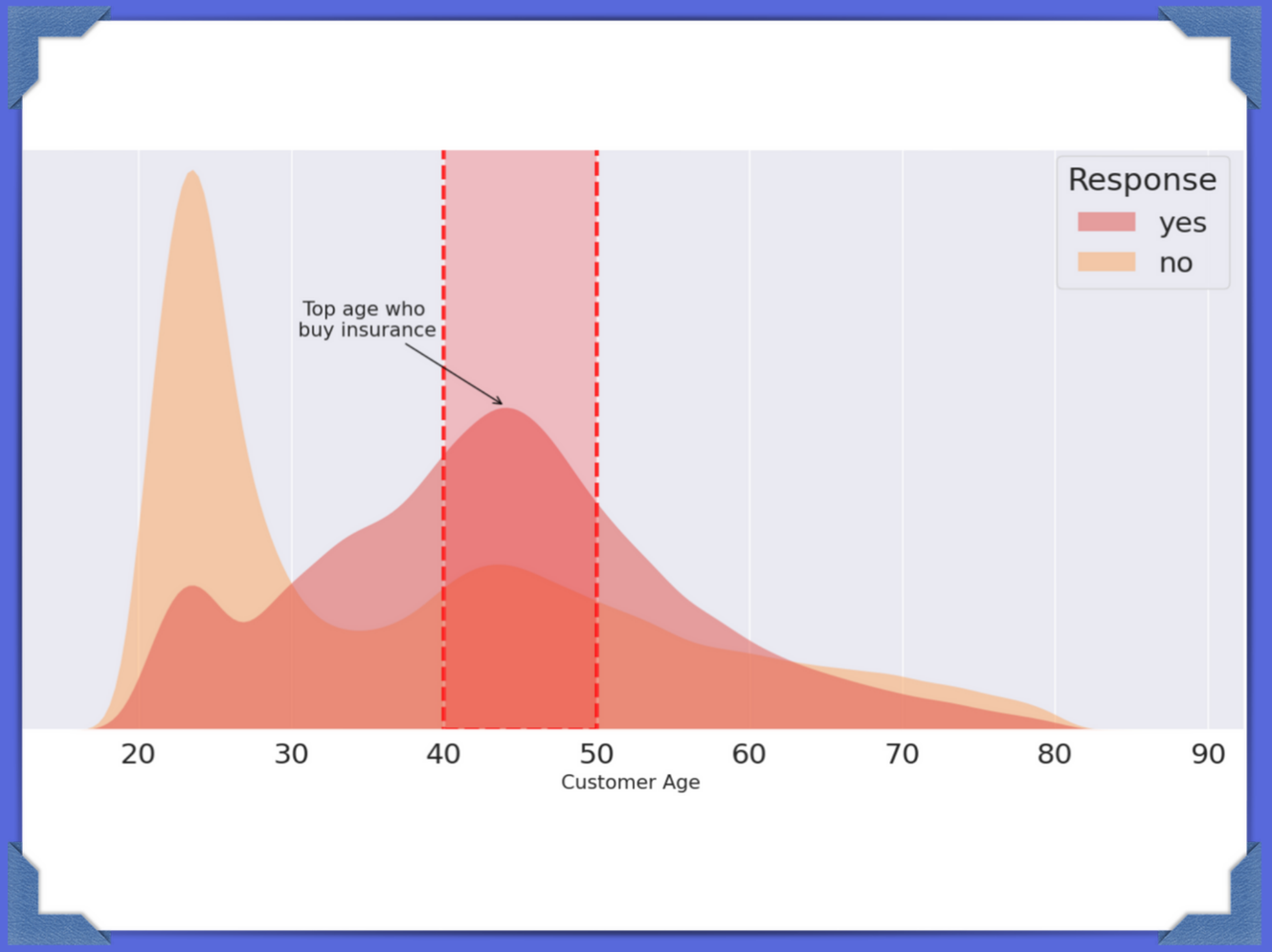
Business Insight

AGE DISTRIBUTION

Pembeli asuransi berada di rentang usia **20-50** dan semakin **menurun** pada usia **50 tahun keatas**

Saat usia **40**, umumnya keadaan **finansial sudah stabil** sehingga perlu mempersiapkan perlindungan aset.

https://investor.id/finance/332547/biar-berguna-sepanjang-hidup-ini-tips-memilih-asuransi-sesuai-usia

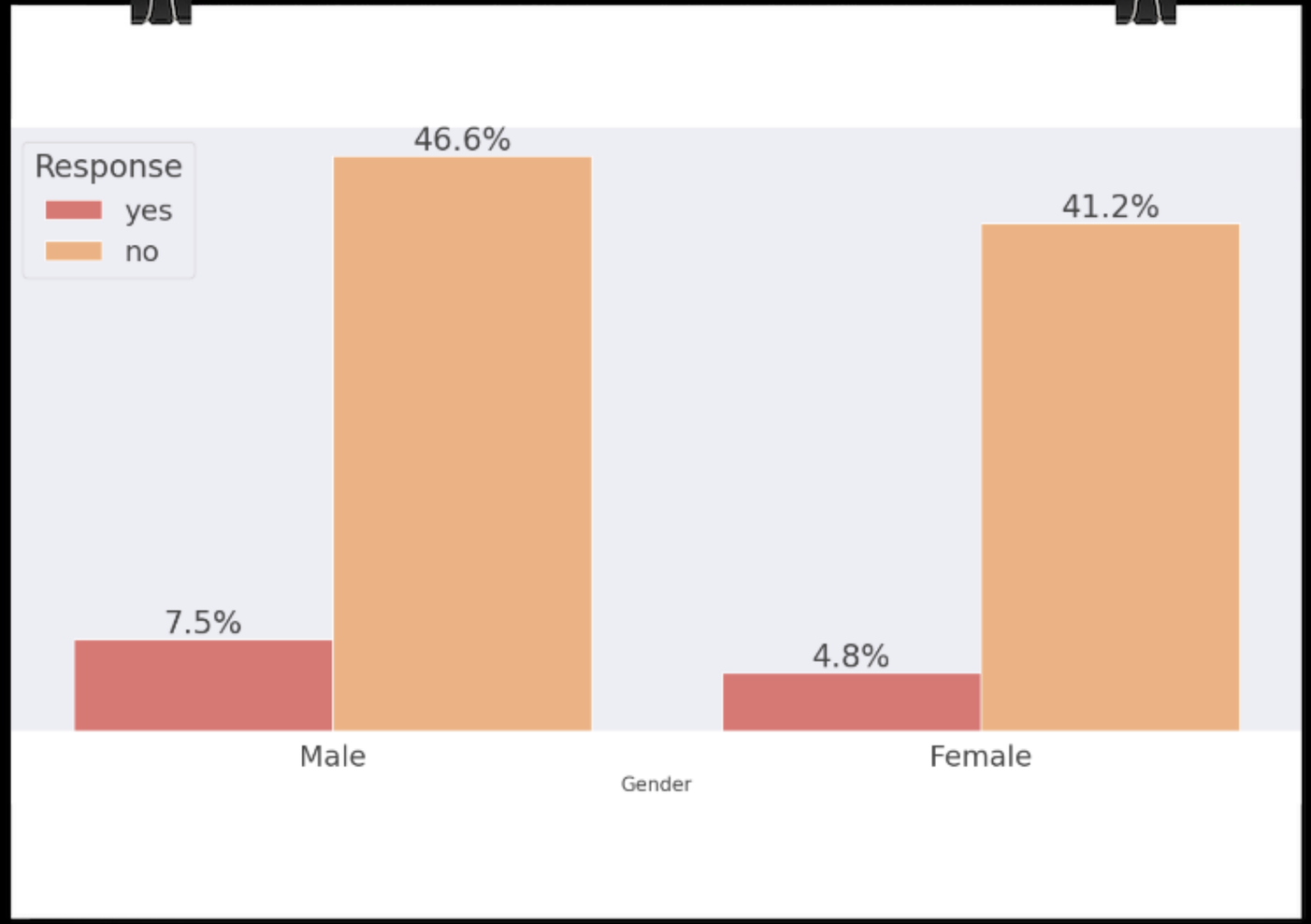


Business Insight

GENDER DISTRIBUTION

Customer asuransi pria lebih banyak dibandingkan wanita.

Menurut penelitian dari **AAA Foundation for Traffic Safety**, pria lebih banyak memiliki asuransi karena pria cenderung lebih berisiko untuk mengalami kecelakaan karena faktor seperti perilaku mengemudi yang agresif, mengantuk, mabuk, tidak memakai sabuk pengaman.



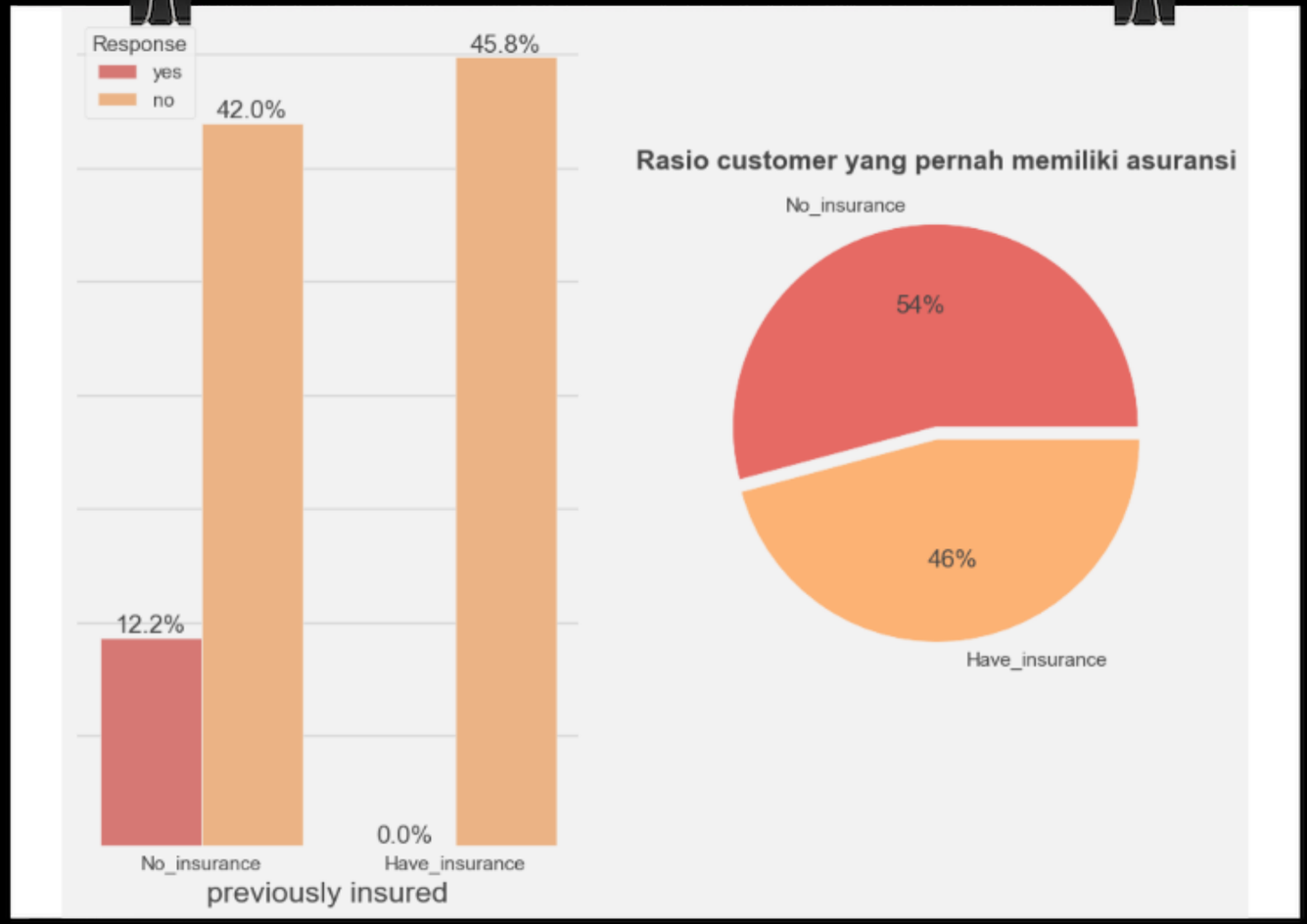
Business Insight

PREVIOUSLY INSURED DISTRIBUTION

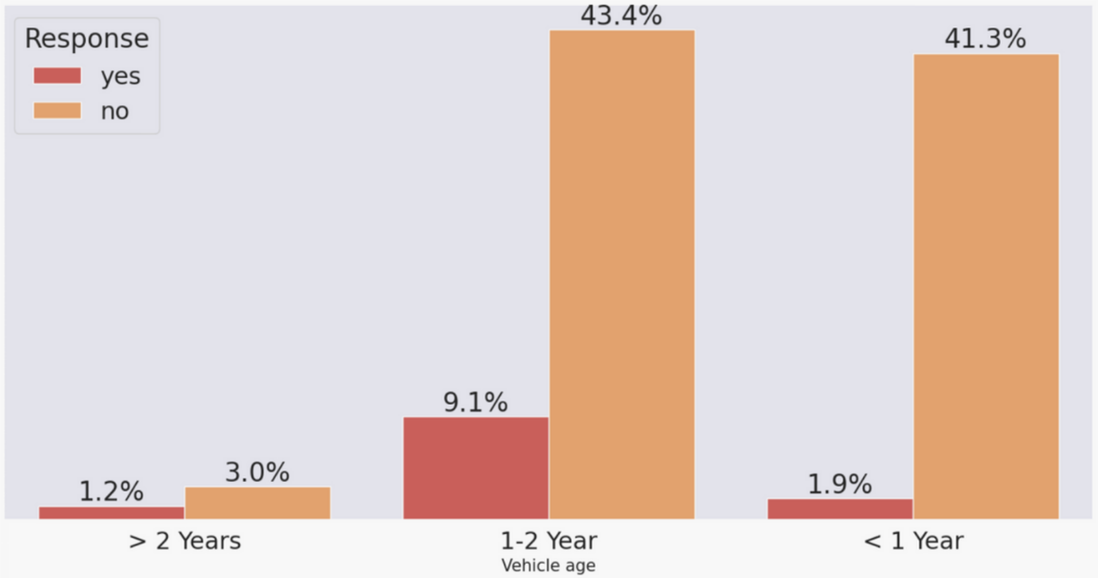
Sekitar 46% customer dalam data ini pernah belangganan asuransi sebelumnya.

menurut artikel **analyticsvidhya.com yang berjudul "Cross-Sell Prediction Using Machine Learning in Python"** Pada dasarnya customer hanya ingin memiliki satu asuransi kendaraan saja, karena apabila customer sudah memiliki asuransi kendaraan, customer tidak akan mempertimbangkan untuk membeli asuransi.

maka perusahaan bisa untuk lebih berfokus pada customer yang belum memiliki asuransi kendaraan, untuk dapat memaksimalkan usaha strategi komunikasi perusahaan dalam menjangkau potensi revenue.



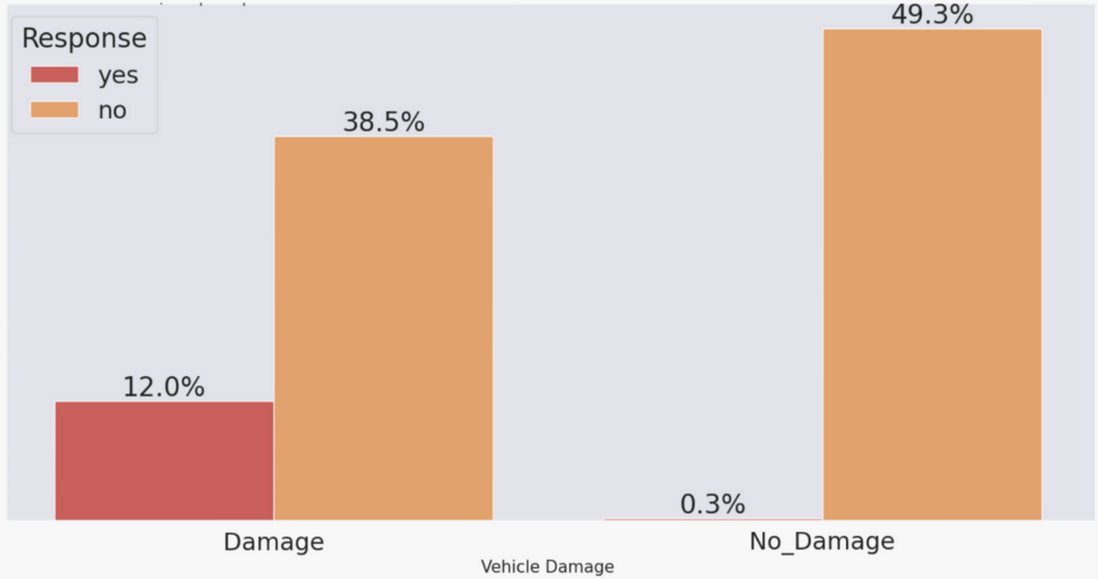
VEHICLE AGE DISTRIBUTION



Customer dengan usia kendaraan kurang dari 2 tahun (1-2 Year dan < 1 Year)

Artikel berjudul “advisor/car-insurance/new-car-replacement” mengenai asuransi kendaraan membuat customer lebih tertarik karena kendaraan lebih terjamin saat menggunakan asuransi.

VEHICLE DAMAGE



customer yang kendaraan nya pernah mengalami kerusakan cenderung lebih tertarik untuk membeli asuransi

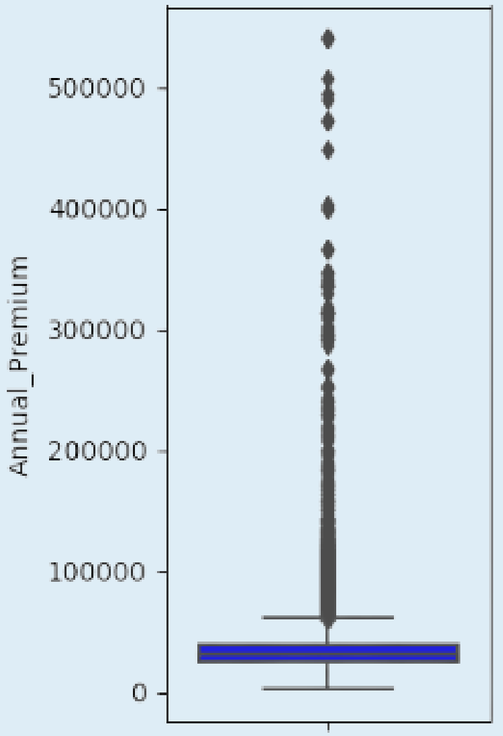
analyticsvidhya.com dalam artikelnya “Cross-Sell Prediction Using Machine Learning in Python” menyatakan customer yang kendaraan nya pernah rusak cenderung lebih memiliki pengalaman dan pengetahuan dari segi biaya perbaikan.

Missing Data & Shape

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   id                    381109 non-null  int64  
 1   Gender                381109 non-null  object 
 2   Age                   381109 non-null  int64  
 3   Driving_License       381109 non-null  int64  
 4   Region_Code           381109 non-null  float64
 5   Previously_Insured    381109 non-null  int64  
 6   Vehicle_Age           381109 non-null  object 
 7   Vehicle_Damage        381109 non-null  object 
 8   Annual_Premium        381109 non-null  float64
 9   Policy_Sales_Channel  381109 non-null  float64
 10  Vintage               381109 non-null  int64  
 11  Response              381109 non-null  int64  
dtypes: float64(3), int64(6), object(3)
memory usage: 34.9+ MB
```

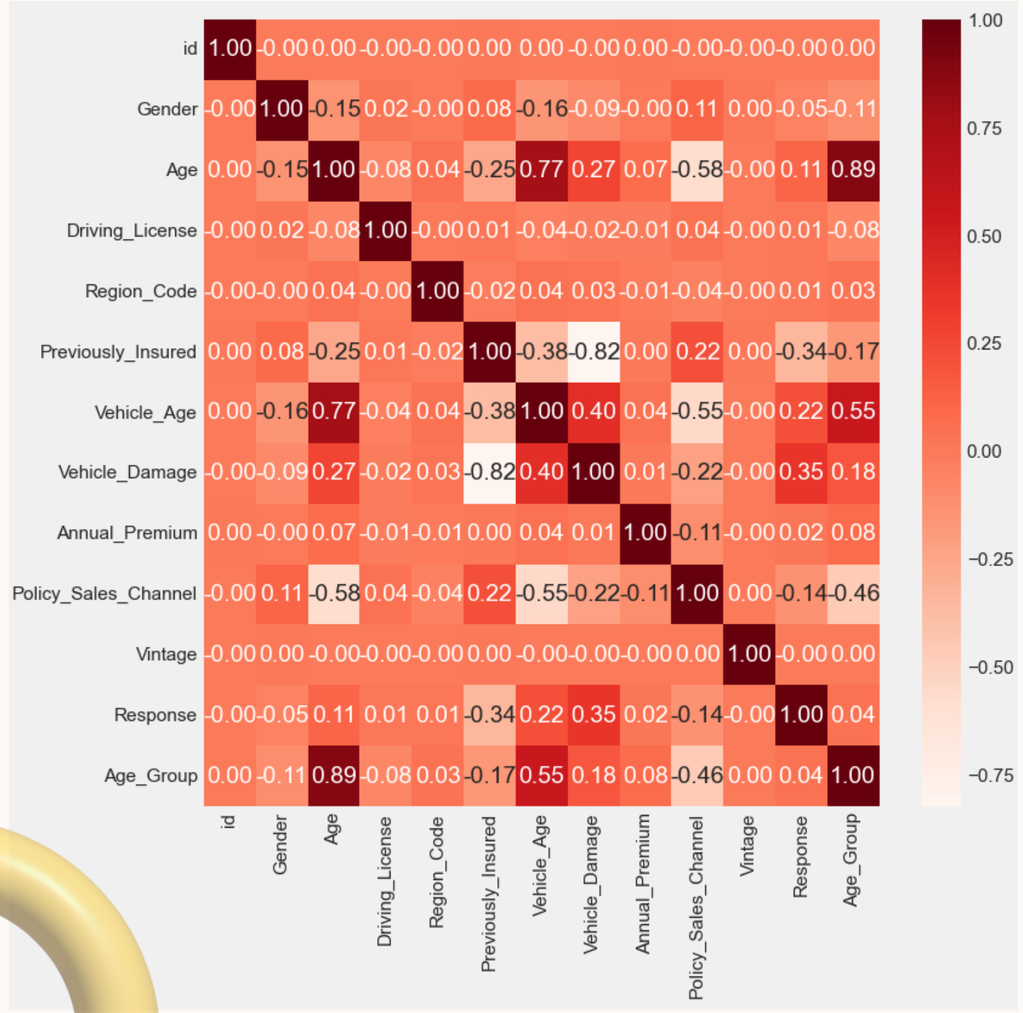
- Tidak ada missing data
- 381109 Baris rows & 12 kolom (10 fitur dan 1 target)

Outlier Check



Outlier pada dataset ini ditemukan pada feature Annual_Premium.

CORRELATION FEATURES



- Feature **Age, previously insured, dan policy sales channel** memiliki nilai korelasi yang **lebih tinggi**
- Feature **age dan policy sales channel** memiliki pola yang cukup **menarik**
- **Uji multikoleniaritas** untuk feature dengan hubungan yang **cukup kuat**

3. PRE-PROCESSING

- MISSING VALUE
- DUPLICATED DATA
- HANDLING OUTLIER
- FEATURE ENGINEERING
- FEATURE SELECTION

DATA CLEANSING
&
PRE PROCESSING

MISSING VALUE & DUPLICATED DATA

- Tidak ada nilai null atau missing value, maka tidak perlu dilakukan handling
- Tidak ada data terduplikasi, maka tidak perlu dilakukan handling

FEATURE ENGINEERING

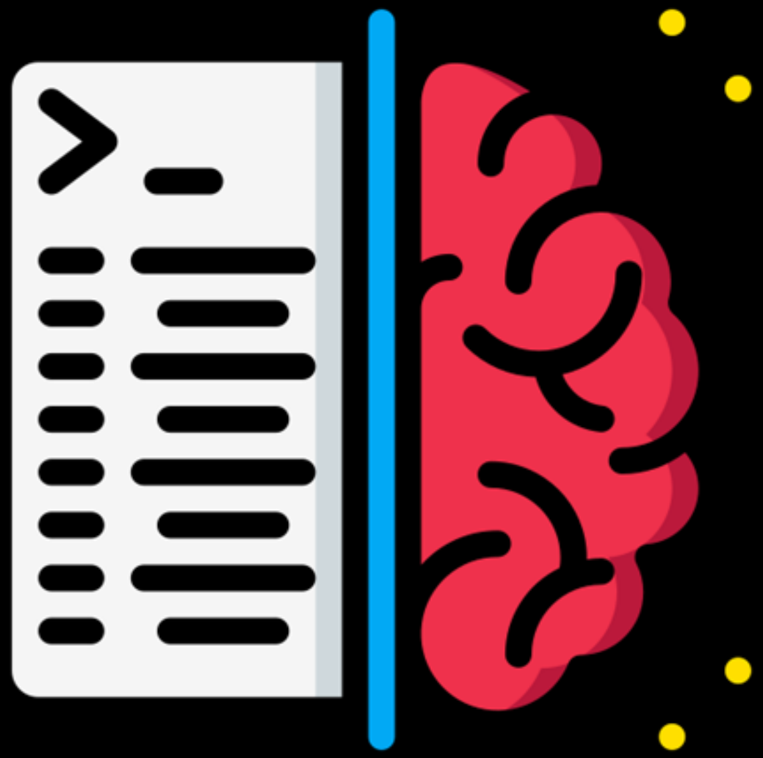
- menambah fitur baru age_group
- melakukan encoding pada fitur kategorikal
- melakukan uji multikolonieritas
- melakukan split train & test
- Melakukan scaling dengan menggunakan standart scaler.
- Handle imbalance data menggunakan random under sampling.

HANDLING OUTLIER

- Dari observasi dan analisis IQR dan QQ plot dapat disimpulkan outlier Annual_Premium merupakan collective outlier,sehingga hal ini masih batas wajar.
- Berdasarkan analisis diatas, tidak akan dilakukan handling outliers lebih lanjut.

FEATURE SELECTION

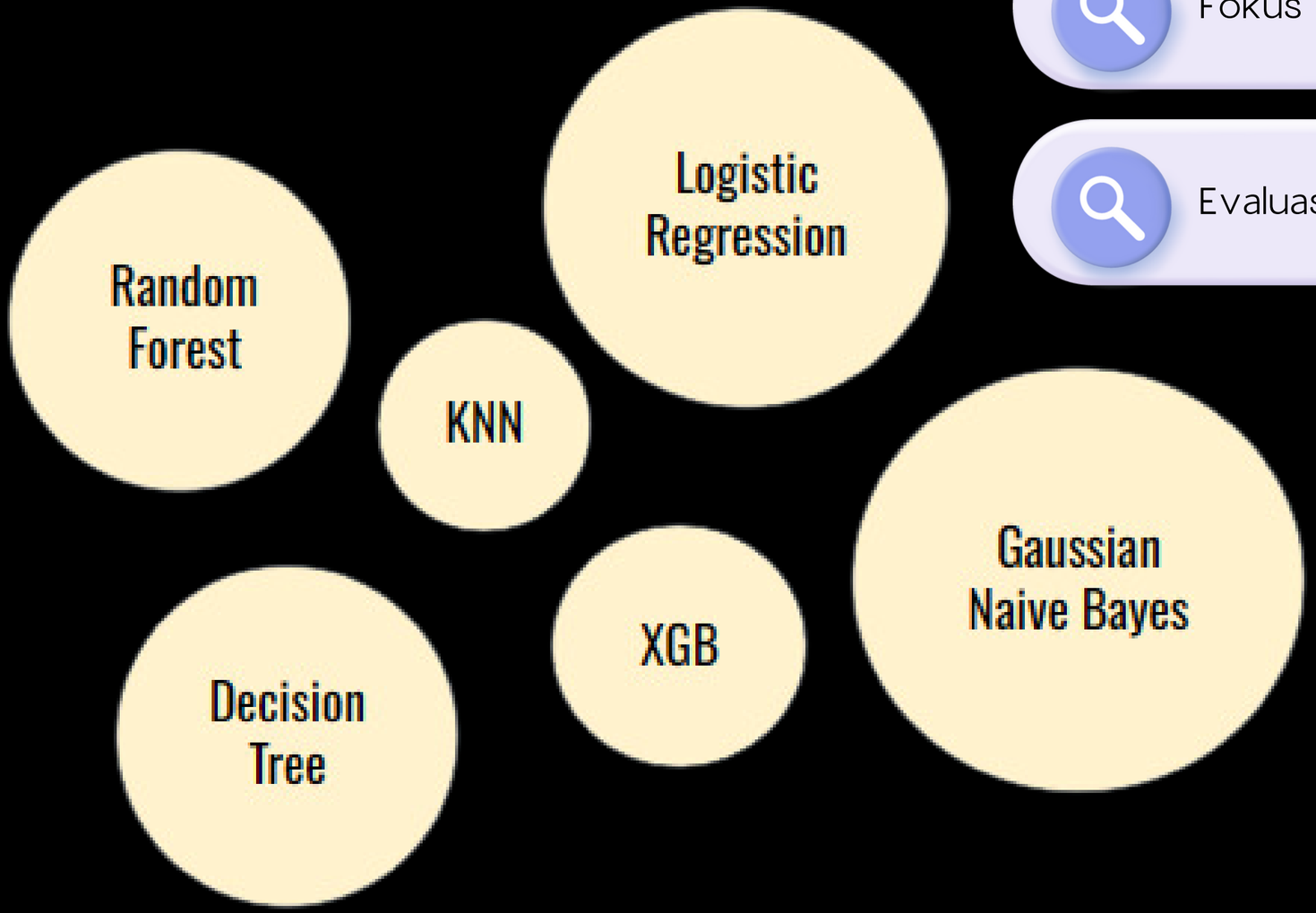
Feature yang dipilih untuk modelling adalah sebagai berikut:
Gender, Previously_Insured, Vehicle_Age, Vehicle_Damage, Annual_Premium, Age_Group, Response



4. MODELLING

- MODELLING EVALUATION
- HYPERPARAMETER TUNING
- CONFUSION MATRIX
- FEATURE IMPORTANCE
- SHAP VALUE

MODELLING



- Fokus pada **potensial user** untuk berlangganan asuransi
- Evaluasi metric yang diperhatikan adalah **recall**

MODELLING
EVALUATION

Model	Training Recall	CV Recall test	Training Precision	CV Precision test	Training F1	CV F1 test	Training AUC_ROC	CV AUC_ROC test
Logistic Regression	97.60%	97.60%	70.40%	70.40%	81.80%	81.80%	83.30%	83.30%
XGB	94.50%	93.20%	72.30%	71.40%	81.90%	80.80%	85.30%	83.00%
Decision Tree	96.00%	74.30%	92.30%	70.60%	94.10%	72.40%	99.00%	72.10%
Random Forest	97.90%	75.70%	90.70%	70.70%	94.20%	73.10%	98.40%	79.80%
Naive Bayes	97.70%	97.70%	70.40%	70.40%	81.80%	81.80%	81.10%	81.10%
KNN	73.80%	67.90%	78.60%	72.20%	76.10%	70.00%	81.70%	74.80%

logistic dan **naive bayes** best model dengan Recall score 97% dan score Auc Roc stabil

HYPERPARAMETER
T U N I N G

Dari hasil Hyperparameter disamping menunjukan bahwa **logistic** mengalami peningkatan, maka akan menggunakan **logistic** untuk modeling.

Model	Training recall	CV Recall test	Training Precision	CV Precision test	Training F1	CV F1 test	Training AUC_ROC	CV AUC_ROC test
Logistic regession	97.70%	97.70%	70.40%	70.40%	81.80%	81.80%	83.40%	83.30%
GNB	97.70%	97.70%	70.40%	70.40%	81.80%	81.80%	81.10%	81.10%

CONFUSION MATRIX

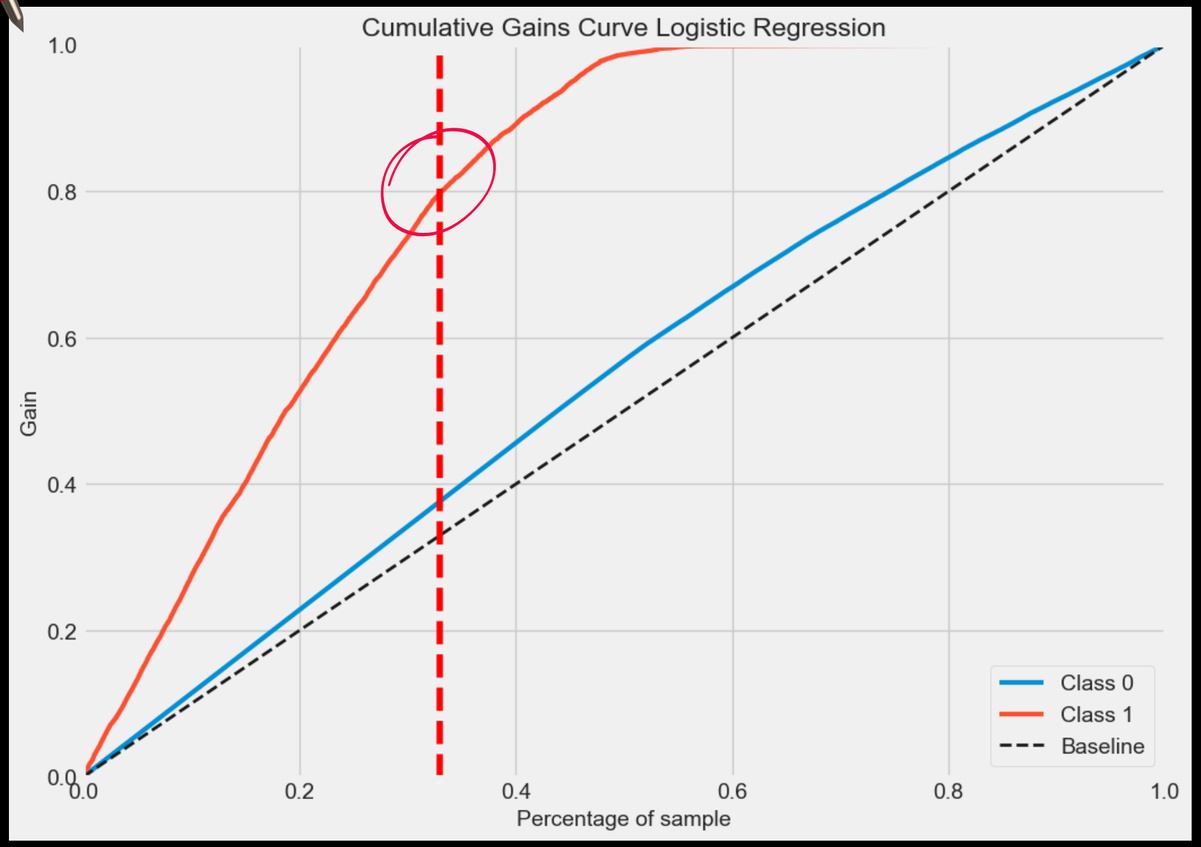
logistic regression before tuning

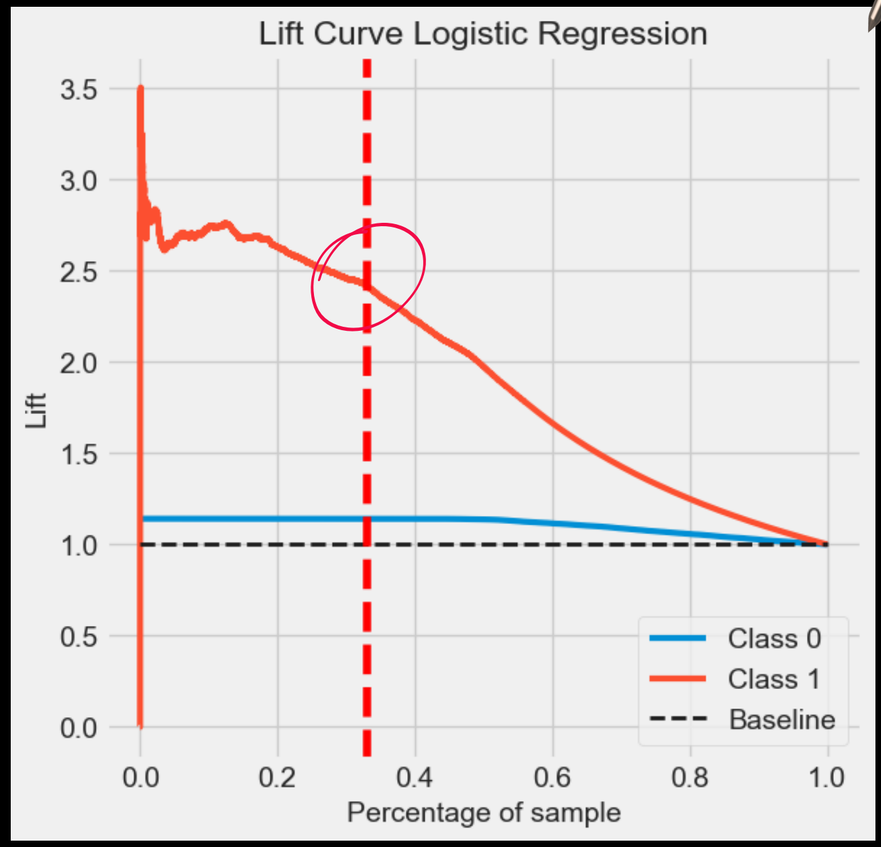
Actual		
	19196	13376
	778	31794
	Predicted Label	

logistic regression after tuning

Actual		
	19191	13381
	753	31819
	Predicted Label	

- Hasil dari hyperparameter tuning logistic regression berhasil menaikan true positive dengan menurunkan false negative.
- Karena ingin melihat model bekerja dengan baik/tidak, kita akan mencoba melakukan analisis cumulative gain curve dan lift curve.



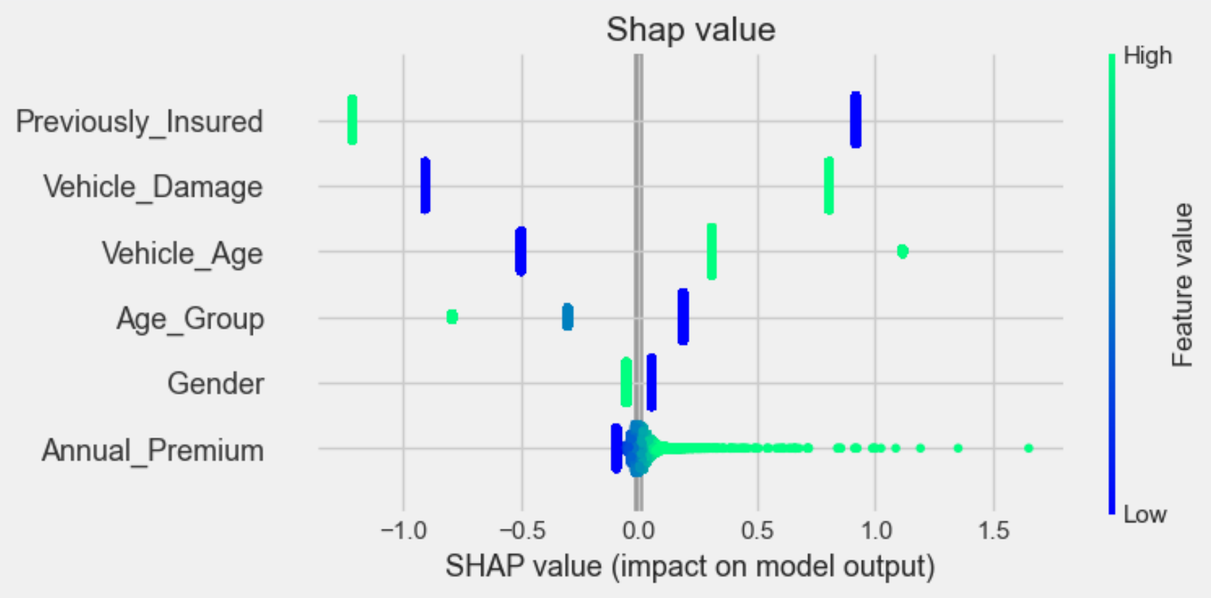


Pada Cumulative gain curve model mampu memprediksi 80% customer yang interest, dengan menggunakan 30% sample data. Hal ini menunjukan menggunakan model lebih baik, daripada melakukan prediksi random/manual.

pada lift curve jika kita memakai 33% sample maka performa machine learning akan meningkat 2.4X lebih baik daripada melakukan prediksi random/manual

FEATURE IMPORTANCE

SHAP VALUE



Pada Feature Importance dan Shap Value terdapat 3 fitur dominan (Vehicle_Damage, Vehicle_Age, dan Previously_Insured). Fitur tersebut memiliki **importance score tinggi** dan memiliki **hubungan kuat** pada **conversion rate**.

5. RECOMMENDATION

- BUSINESS INSIGHT
- ADDITIONAL DATA
- STRATEGIC TREATMENT
- MODEL DEPLOYMENT SIMULATION
- REVENUE SIMULATION
- BUSINESS FLOW SIMULATION

INSIGHT

Previously Insured

Customer yang sudah memiliki asuransi tidak akan berlanggan asuransi lagi

University of San Diego “A Classification Problem: Health Insurance Cross Sell Prediction“ mengatakan bahwa: “Customers who were previously insured tended to accept the cross-selling offer if the company made a better offer to them”

Vehicle Damage

Jika customer pernah mengalami kerusakan pada kendaraan maka customer memiliki kemungkinan yang tinggi untuk melakukan pembelian asuransi

University of San Diego “A Classification Problem: Health Insurance Cross Sell Prediction“, mengatakan bahwa: “people who have experienced vehicle damage before are very likely to buy insurance to avoid out-of-pocket costs ”

Vehicle Age

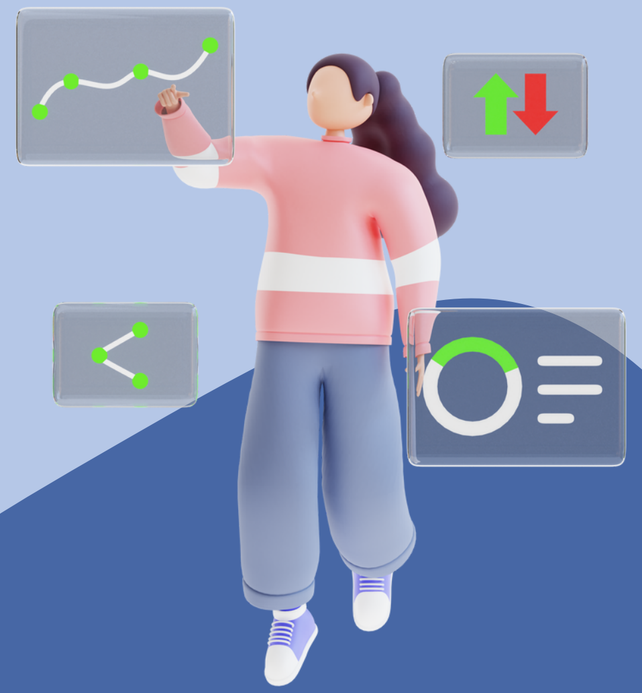
Untuk kendaraan baru customer cenderung melakukan perawatan maksimal dengan membeli asuransi

Kunming University of Science and Technology “Research on the Features of Car Insurance Data Based on Machine Learning” mengatakan bahwa: “Lower Car/Vehicle age have a greater impact on whether to buy/renew insurance or not”

RECOMMENDATION
ADDITIONAL DATA

- LOAN
- DEPENDENT
- VEHICLE TYPE
- VEHICLE BRAND
- VEHICLE OWNED
- OCCUPATION
- INCOME
- PAYMENT METHOD

Data tambahan dapat bermanfaat untuk melihat faktor lain yang mempengaruhi kinerja model dan ketertarikan customer dengan vehicle insurance yang ditawarkan



RECOMMENDATION
STRATEGIC TREATMENT

Targeted Campaign

Berfokus pada customer yang sudah mempunyai asuransi, pernah mengalami kerusakan kendaraan, dan mempunyai usia kendaraan kurang dari 2 tahun

Add-on Services

Menawarkan layanan add-on gratis khusus kepada customer yang sudah pernah mengalami kerusakaan pada kendaraan dengan opsi layanan add on seperti **"Zero-Depreciation", "Engine Protection", dan "24x7 Roadside Assistance".**

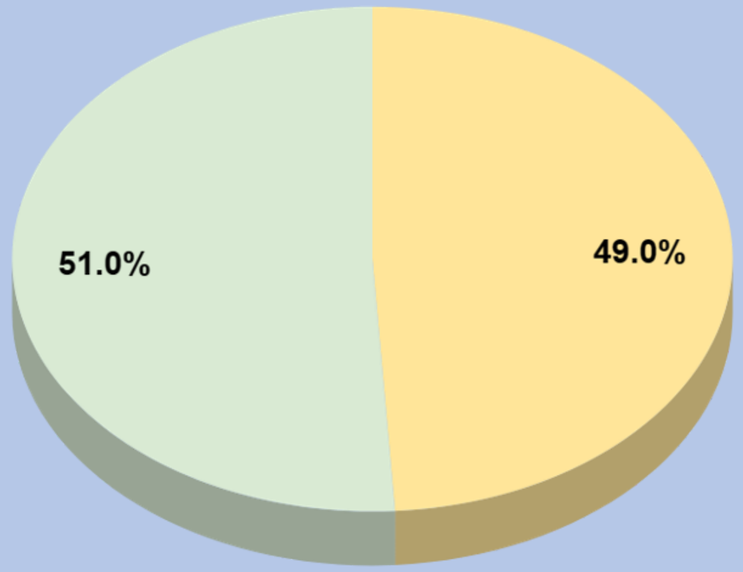
Discount

Menawarkan diskon khusus kepada customer yang sudah mempunyai asuransi sebelumnya seperti diskon untuk **"Safety Driver", "Young Driver" dan "Mature Driver".**

Showroom / Dealer Partnership

Melakukan bundling asuransi kendaraan dengan berpartner dengan Showroom/Dealer untuk customer yang membeli kendaraan

Prediction Result



Meningkat sebanyak 37%
12% ➡ 49% dari total
381.109

Recommendation

Business Simulation Model Deployment

Pada model logistic regression kita berhasil melakukan prediksi pada customer yang berminat untuk melakukan pembelian asuransi, Dengan model logistic regression kita bisa menemukan potensial customer sebanyak 49%. Sehingga dengan ini direkomendasikan untuk melakukan deploy model machine learning

Recommendation
Business Simulation Revenue

Customer Base	
Total Customer	381,109
Interested Customer Without Model (12%)	45,733
Interested Customer With Model (49%)	186,743

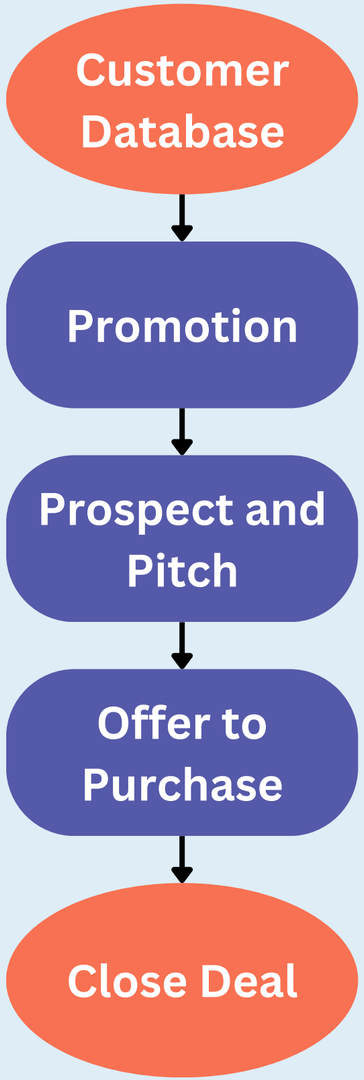
Revenue/Profit Simulation					
Description	$ Per Person	Without Model	With Model	Increment (in $)	Increment (in %)
Premium (Revenue)	$468	$21,403,081	$87,395,916	$65,992,834	308.33%
Campaign (7% per Premium)	$33	-$1,509,192	-$6,162,533	-$4,653,341	308.33%
Operational Expense (18% per Premium)	$84	-$3,841,579	-$15,686,446	-$11,844,868	308.33%
Claim (68% per Premium)	$318	-$14,543,119	-$59,384,404	-$44,841,285	308.33%
Profit Investment (16% per Premium)		$3,424,493	$13,983,347	$10,558,854	308.33%
Profit		$4,933,685	$20,145,879	$15,212,194	308.33%

Apabila menggunakan model yang sudah dibuat, revenue akan mengalami kenaikan yang sangat drastis sebanyak 308%. Hal ini sesuai dengan business metrics yang disebutkan di awal untuk menaikkan revenue dari asuransi kendaraan, Hal ini juga memperkuat argumen untuk kami merekomendasikan melakukan deployment model machine learning yang telah dibuat.

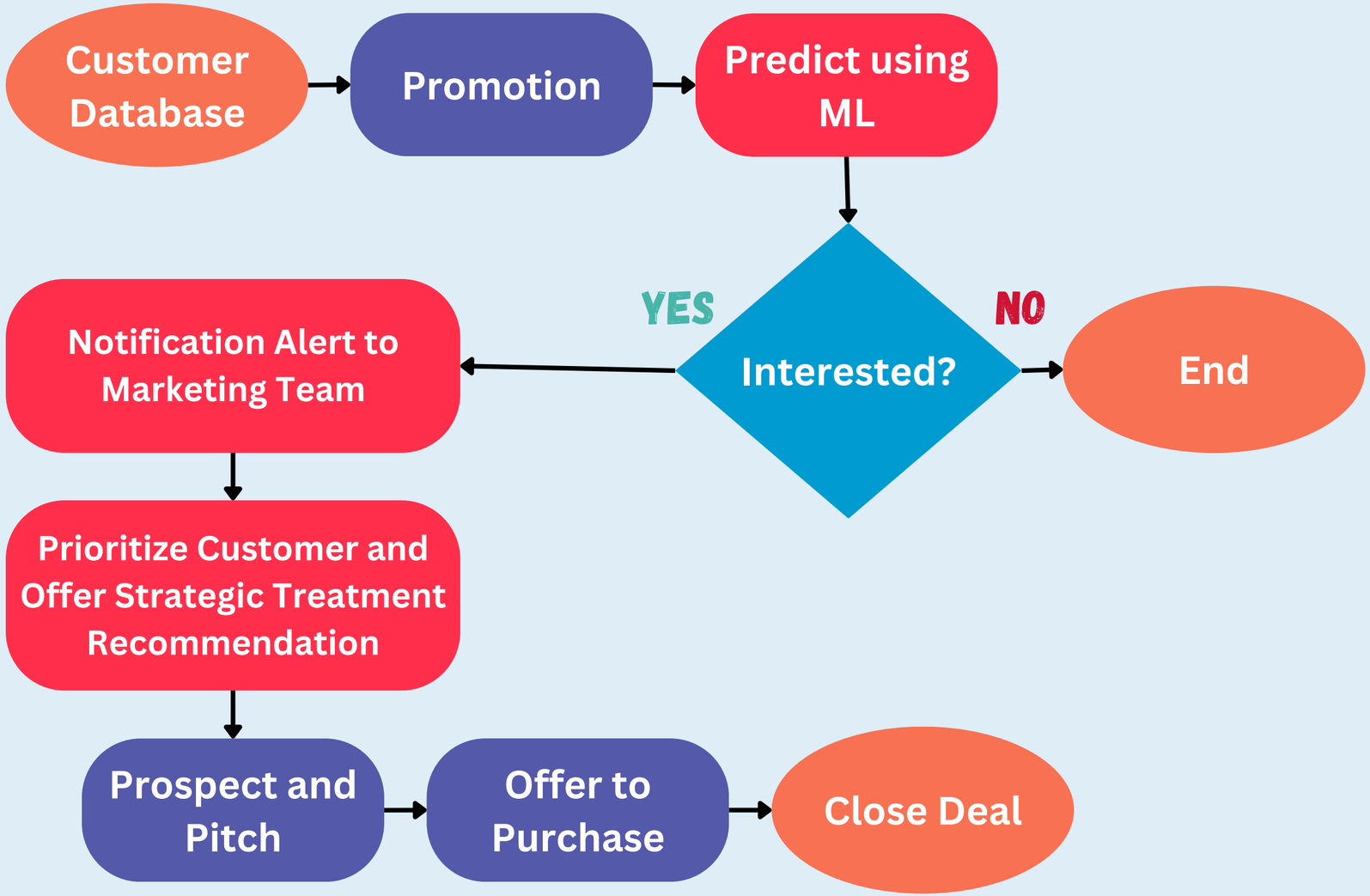
Sumber: 1.winsurtech 2. carinsurance 3. Insurance Information Institute
Asumsi: Variabel Revenue dan Expense diasumsikan Konstan dan dikalikan dengan interested customer untuk simulasi revenue ini.

Business Flow Simulation

Before Model Deployment



After Model Deployment



THANK YOU!

Presentasi oleh **Data Connector**

RAKAMIN ACADEMY | 2023