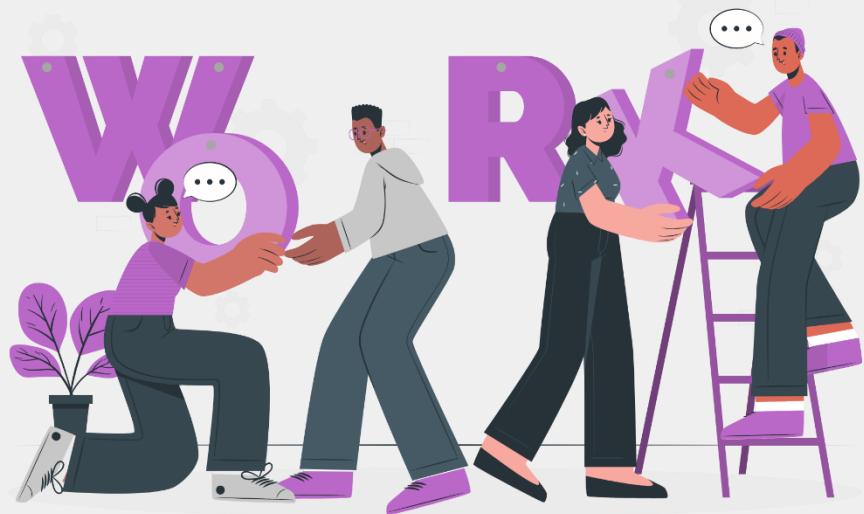


FINAL PROJECT RAKAMIN BATCH 25

CROSS SELLING INSURANCE



People Behind Project



Dwita

Project Leader/ Data Scientist



Maulana

Data Scientist



Putu

Data Scientist



Friska

Data Scientist



Anggita

Data Scientist



Pambudi

Data Scientist



Edward

Data Scientist

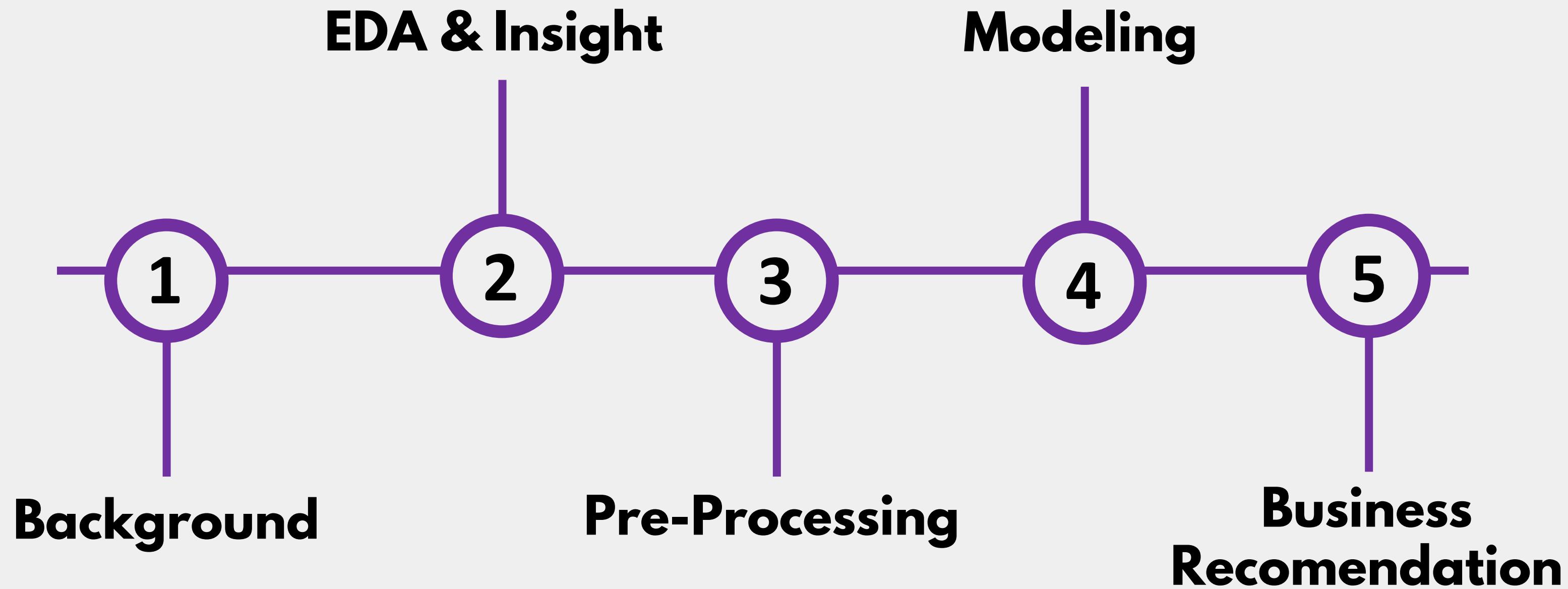


Zaki

Data Scientist

»

Outline





Problem Statement

»

Real Case

Data Kecelakaan Transportasi berdasarkan Moda Tahun 2010-2020

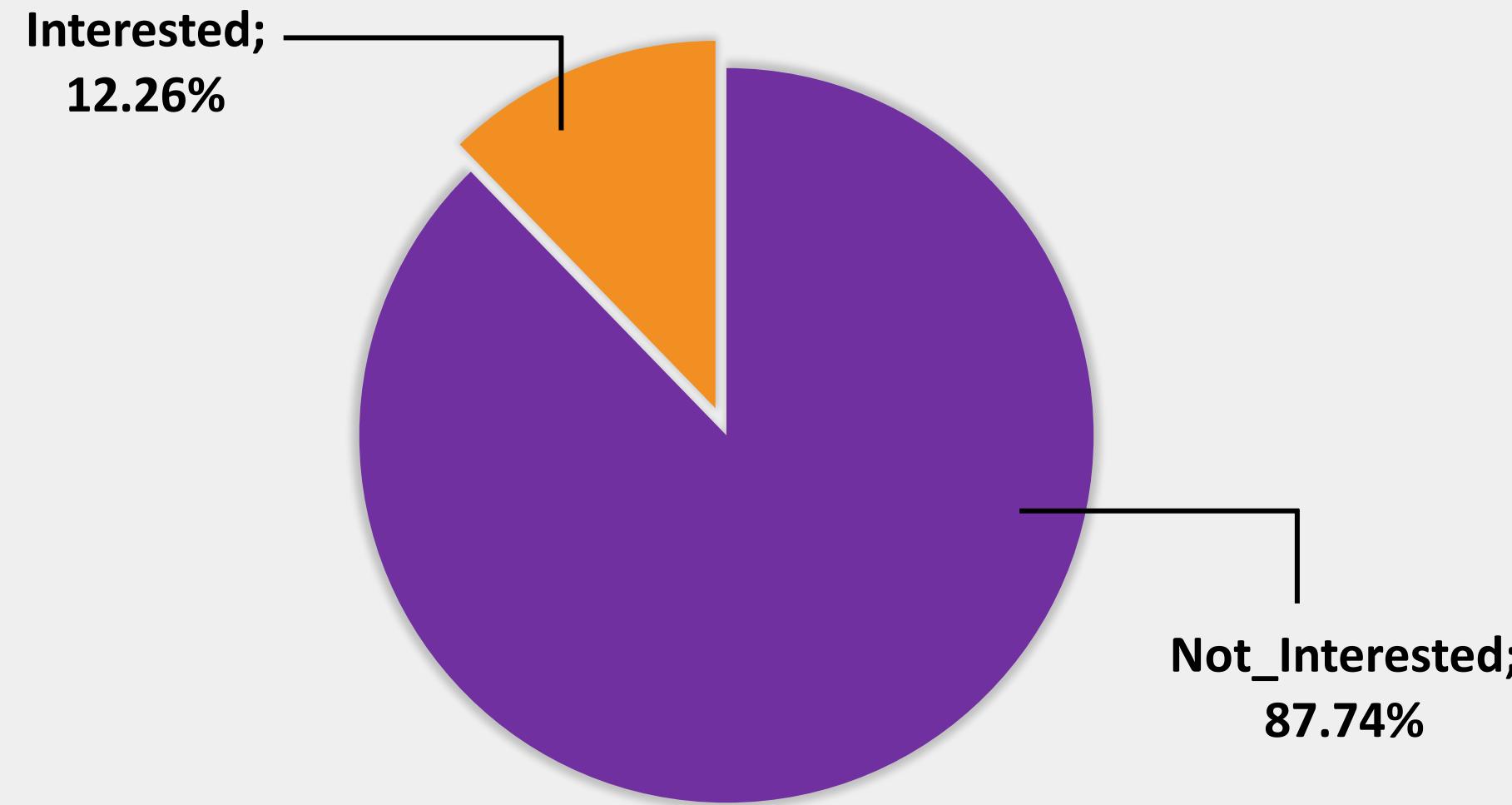
Jenis Transportasi	Jumlah Kecelakaan	Jumlah Kematian	Rata-rata kematian/miles
Pesawat	257	16	0,002
Kereta	11	56	0,31
Kapal	10	157	0,89
Bus	2.262	3.399	0,86
Mobil	5.250.837	35.766	3,5

- Moda transportasi dengan jumlah kecelakaan paling tinggi, **mobil**, bus dan pesawat
 - Moda transportasi yang paling banyak memakan korban jiwa: **mobil**, bus dan kapal
- Transportasi mobil merupakan jenis kendaraan transportasi yang paling tidak aman

Terdapat kerugian materiil bagi pemilik kendaraan mobil akibat kecelakaan tersebut

Sumber: Total Bus Crash, USA, 2020

Business Problem



Masih sedikitnya persentase pelanggan Asuransi Kesehatan yang memiliki ketertarikan untuk membeli Asuransi Kendaraan. Berdasarkan kondisi saat ini hanya sebesar 12,3% pelanggan yang tertarik membeli asuransi kendaraan.

Business Problem

Goals

- Meningkatkan **User Interested Rate sebesar 10%**

Objective

- Memprediksi nasabah yang potensial untuk asuransi Kendaraan
- Menemukan faktor penting dan karakteristik utama dari user yang tertarik dengan asuransi kendaraan
- Menemukan metode apa yang membuat customer tertarik
- Meningkatkan jumlah orang yang tertarik dengan asuransi kendaraan
- Mengimplementasikan pada sebuah simulasi bisnis untuk melihat apakah model yang dibuat memberikan dampak yang positive untuk perusahaan Asuransi.

Business Metrics

- User Interested Rate

Market share car insurance tertinggi di US diperoleh oleh Statefarm dengan menguasai sekitar 16% dari total keseluruhan Total Car Registered per 2022 sekitar 285M

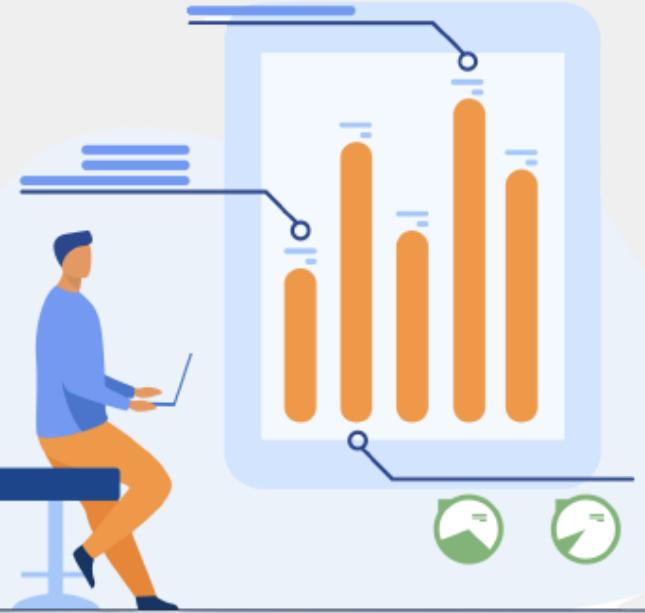
Sumber : valuepenguin



EDA & Insight

»

• Dataset

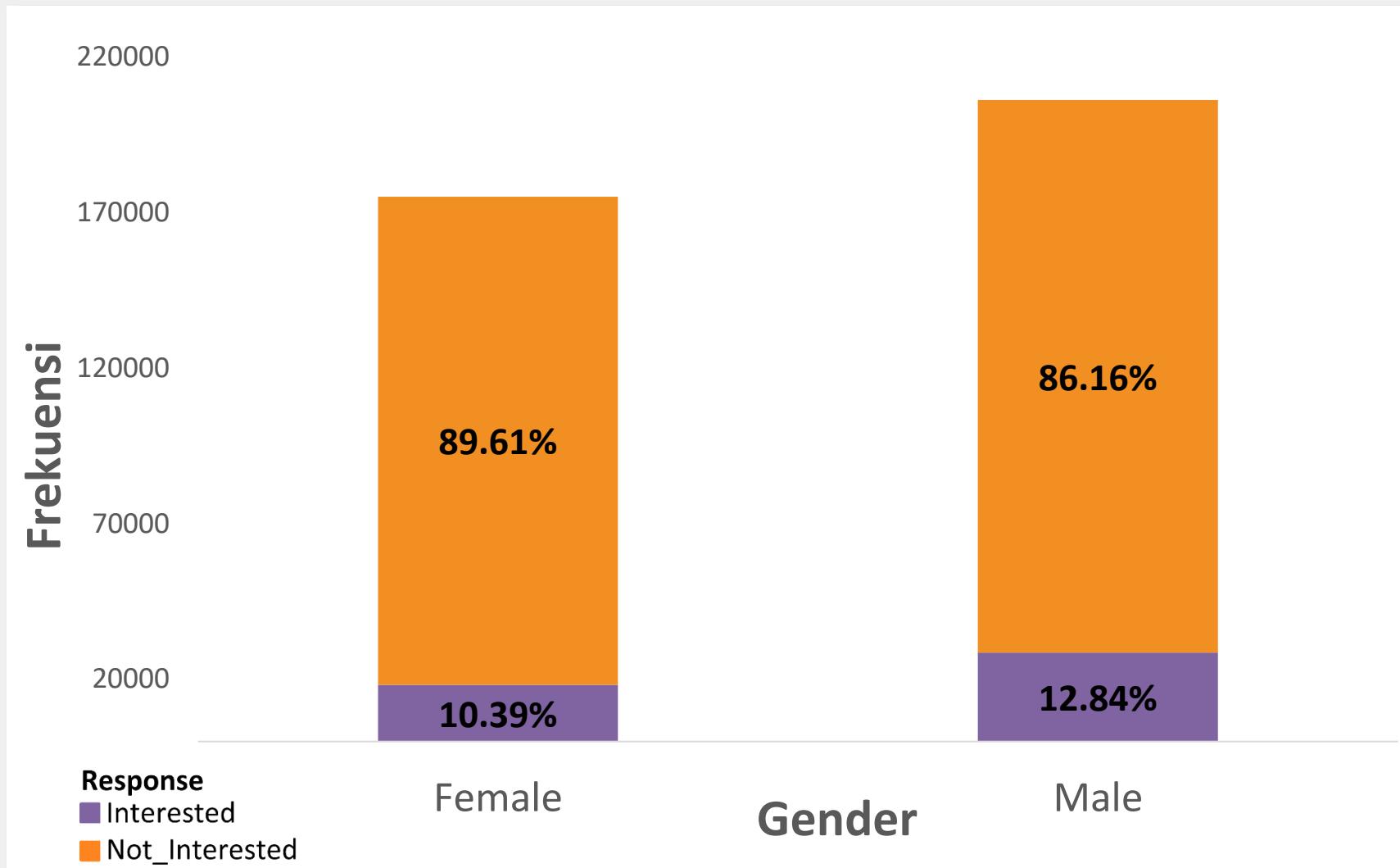


Jumlah Baris

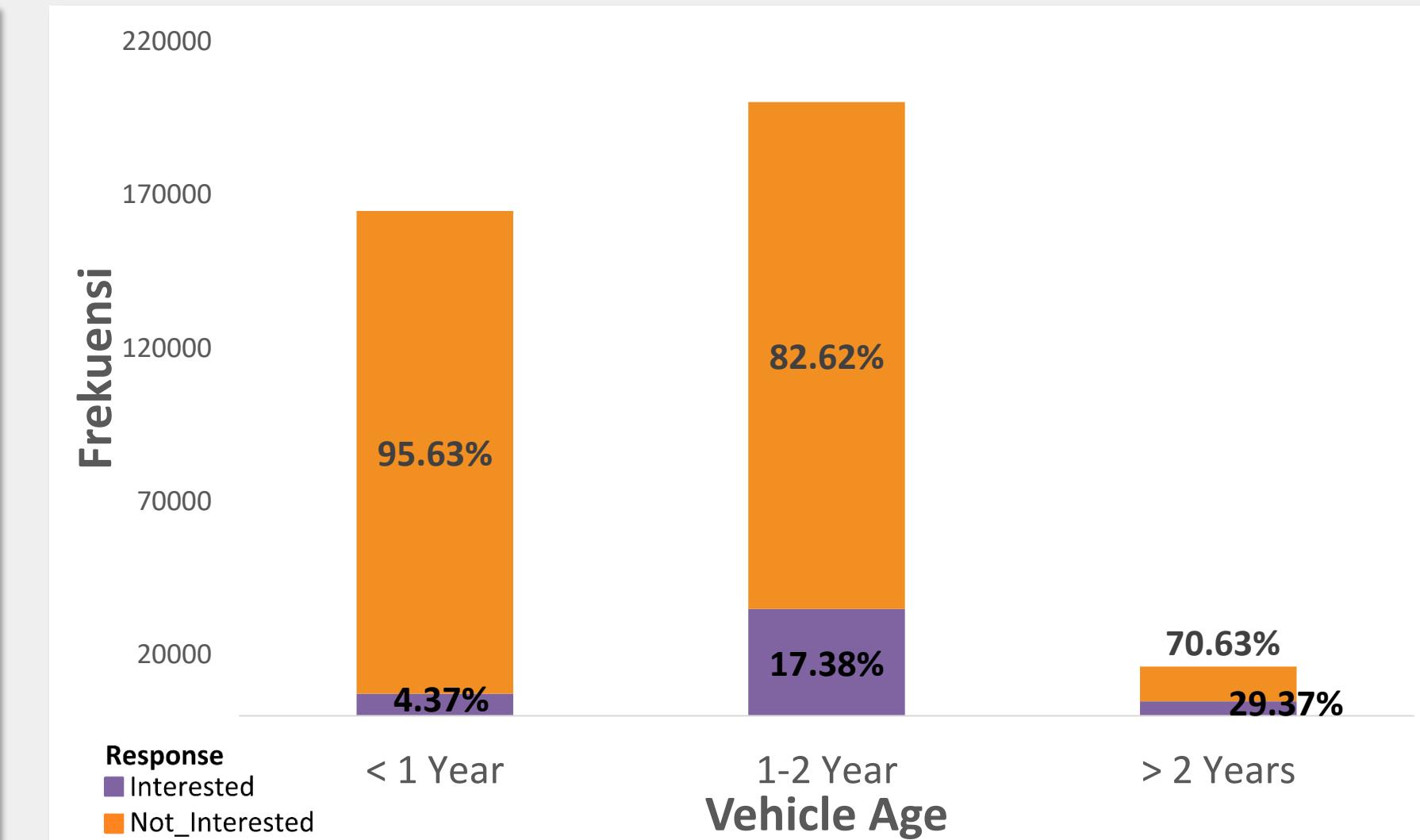
381.109

Nama Variabel	Keterangan
id	Unique ID for the customer
Gender	Gender: Gender of the customer
Age	Age: Age of the customer
Driving_License	0 : Customer does not have DL, 1: Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1: Customer already has Vehicle Insurance 0 : Customer doesn't have Vehicle Insurance
Age of the Vehicle	Age of the Vehicle
Vehicle_Damage	1: Customer got his/her vehicle damaged in the past 0 : Customer didn't get his/her vehicle damaged in the past
Annual_Premium	The amount customer needs to pay as premium in the year
PolicySalesChannel	Anonymized Code for the channel of outreach to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc
Vintage	Number of Days, Customer has been associated with the company
Response 1	1: Customer is interested 0 : Customer is not interested

• Business Insight

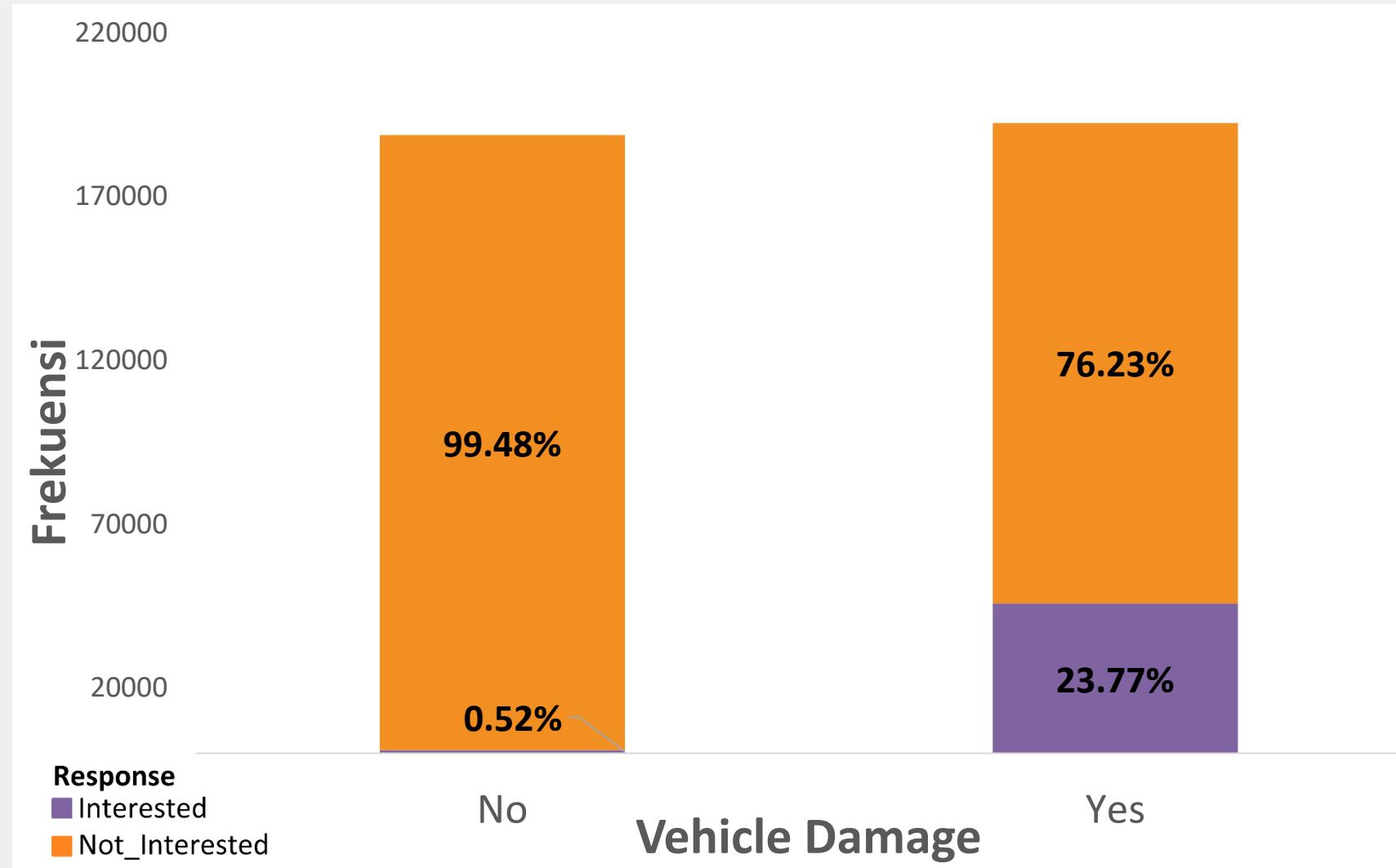


Gender Male lebih banyak tertarik menggunakan
Asuransi Kendaraan

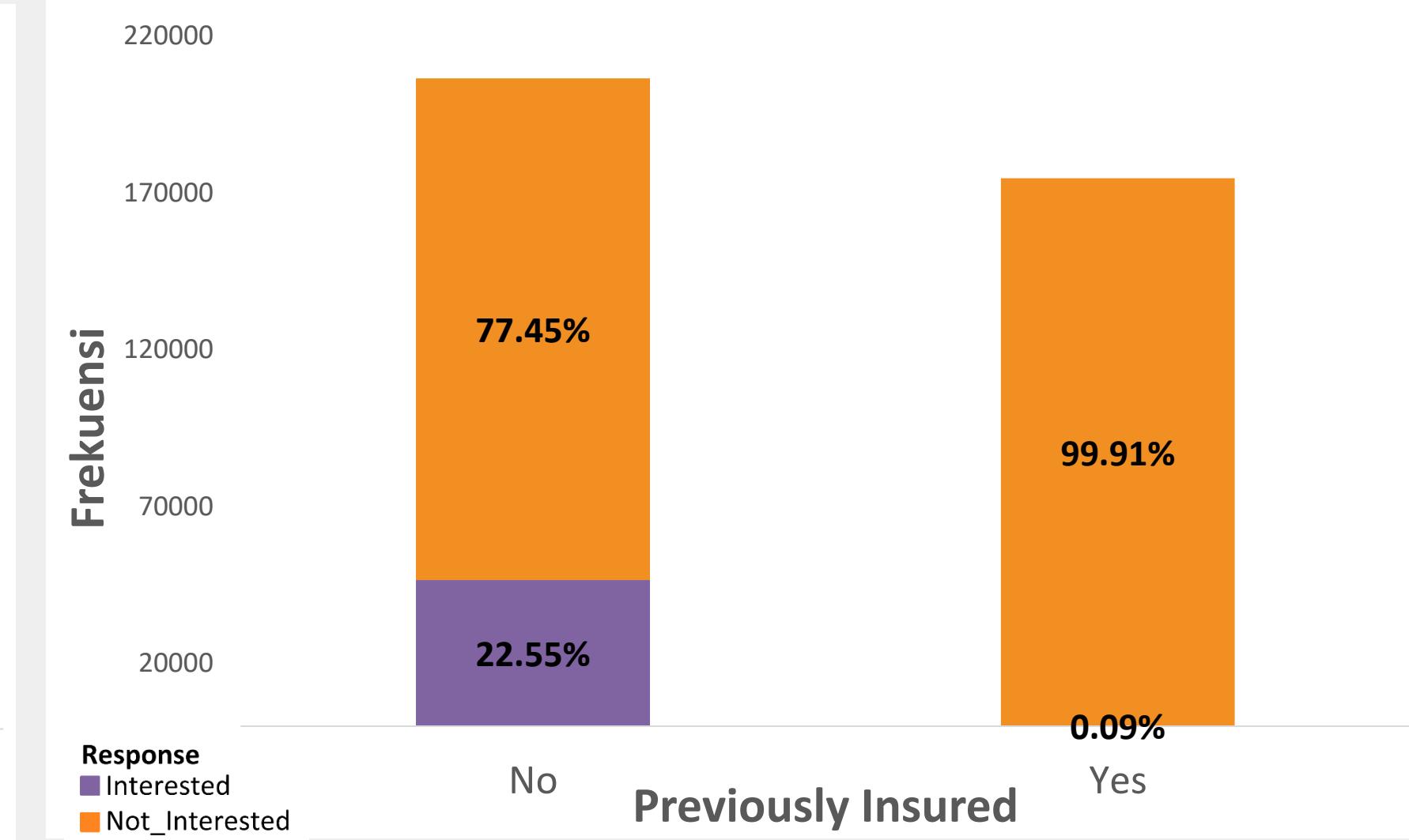


Usia Kendaraan 1-2 tahun lebih banyak tertarik
menggunakan Asuransi Kendaraan

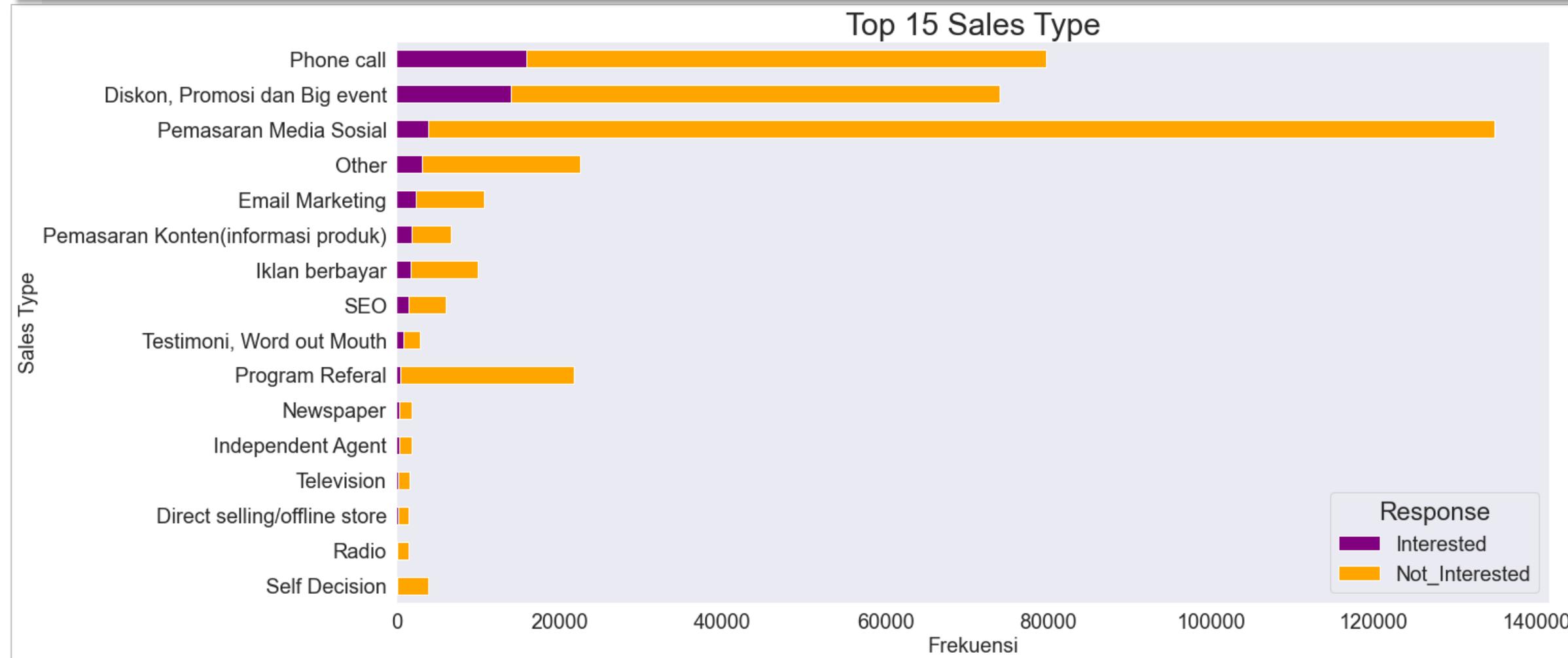
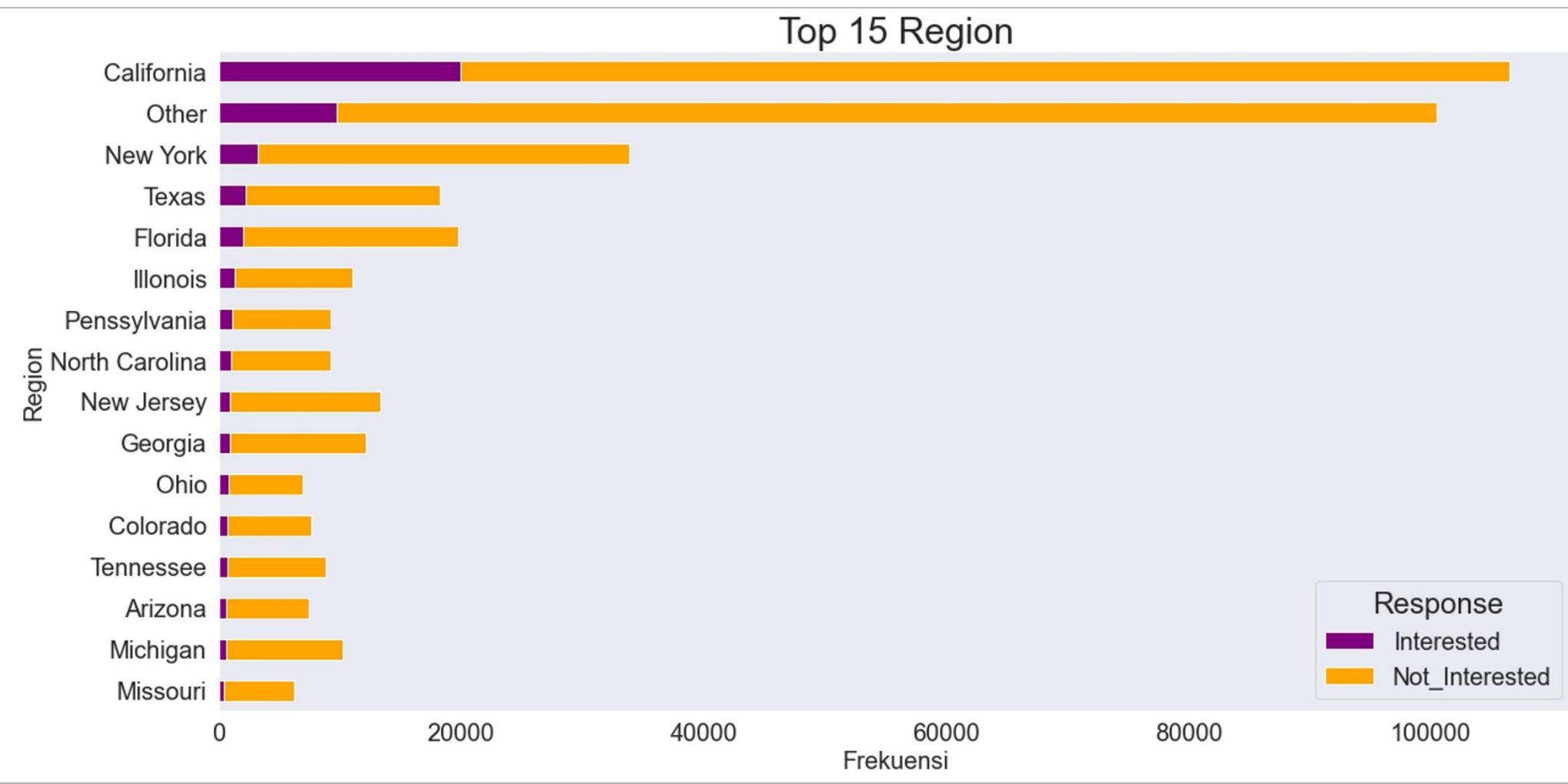
• Business Insight



Nasabah yang memiliki kondisi kendaraan sebelumnya rusak, banyak yang tertarik menggunakan Asuransi Kendaraan



Nasabah yang sebelumnya belum menggunakan Asuransi Kendaraan secara keseluruhan tertarik menggunakan asuransi kendaraan



- Nasabah di Region **California, New York, Texas, Florida** dan **Illinois**, banyak yang tertarik menggunakan Asuransi Kendaraan dibandingkan Region lain

- Ada 10 kode jenis **Policy_Sales_Channel** yang paling banyak dalam memberikan kontribusi ketertarikan nasabah menggunakan Asuransi Kendaraan. Top 3 diantaranya adalah dengan **Phone Call, Diskon Promosi** dan **pemasaran melalui media sosial**



Usia Nasabah dari 33 – 52 merupakan kelompok umur yang paling banyak tertarik menggunakan Asuransi Kendaraan.

Data Cleansing & Preprocessing



»

Data Cleansing & Preprocessing

Missing Value & Duplicated Data

- Tidak terdapat variabel yang memiliki missing value dan tidak terdapat duplicated data.
- Kami tidak melakukan input maupun drop terhadap data duplikatnya.

Handling Outliers

- Kami menggunakan QQ-Plot untuk menganalisa outliernya dan digunakan IQR untuk menghandle outliernya.
- Hasil analisis outlier menggunakan QQ-Plot bahwasanya Outlier tersebut merupakan collective outlier.

Feature transformation and Encoding

- Variable Annual_Premium dan Age memiliki distribusi data bimodal.
- Data preprocessing akan dilakukan scaling dan transformasi menggunakan transformasi Yeo-Jhonson.
- Feature encoding dilakukan pipeline serta membagi data menjadi tiga jenis yaitu ordinal encoding dan numeric encoding

Handle class imbalance

- Class imbalance kami handle menggunakan metode Class Weight.



Modeling



»

Modeling

Cross-Selling



Meningkatkan Cross Selling Prediction dengan memaksimalkan potensial customer

»

Membuat model yang sekecil mungkin salah prediksi pada customer yang kemungkinan tertarik

False Negative (FN) harus dikurangi agar terhindar dari salah prediksi

METRICS RECALL

$$\text{Recall} = \frac{TP}{TP + FN}$$

Semakin kecil False Negative maka semakin besar nilai Recall nya.

• Cross Validation

Data Train

Model	Training Recall	CV Recall (mean)	CV Recall (std)	Training Precision	CV Precision (mean)	CV Precision (std)	Training F1	CV F1 (mean)	CV F1 (std)	Training AUC_ROC	CV AUC_ROC (mean)	CV AUC_ROC (std)
Logistic Regression	97.63%	97.63%	0.12%	25.04%	25.04%	0.05%	39.86%	39.86%	0.05%	78.40%	82.79%	0.14%
XGB	1.45%	0.70%	0.12%	78.92%	39.38%	2.66%	2.85%	1.38%	2.66%	50.70%	84.43%	0.07%
Decision Tree	98.39%	39.42%	0.34%	65.89%	26.99%	0.34%	78.93%	32.04%	0.34%	95.64%	62.30%	0.25%
Random Forest	98.07%	39.15%	0.40%	66.30%	27.40%	0.21%	79.12%	32.24%	0.21%	95.55%	77.16%	0.15%
Naïve Bayes	97.52%	97.52%	0.12%	25.21%	25.21%	0.04%	40.07%	40.06%	0.04%	78.56%	81.63%	0.20%

Logistic Regression memiliki nilai ROC_AUC paling tinggi dan nilai CV Recall terbaik. Terlihat juga score Recall dari Cross Validation dan data training sudah best fit, namun tetap akan dioptimalkan menggunakan Tuning Hyperparameter.

• Hyperparameter Tuning

Model	Training Recall	Training Precision	Training F1	Training AUC_ROC
Logistic Regression	97.63%	25.04%	39.86%	78.40%
Logistic Regression (After Tuning)	97.71%	24.99%	39.80%	78.37%

**Confusion Matrix Logreg Training
(Before Tuning Hyperparameter)**



		Prediksi	
		No	Yes
Realita	No	158,311	109,208
	Yes	886	36,482

**Confusion Matrix Logreg Training
(After Tuning Hyperparameter)**

		Prediksi	
		No	Yes
Realita	No	157,928	109,591
	Yes	856	36,512

Recall mencapai **97.71%**, dimana jumlah nasabah yang terprediksi FN turun

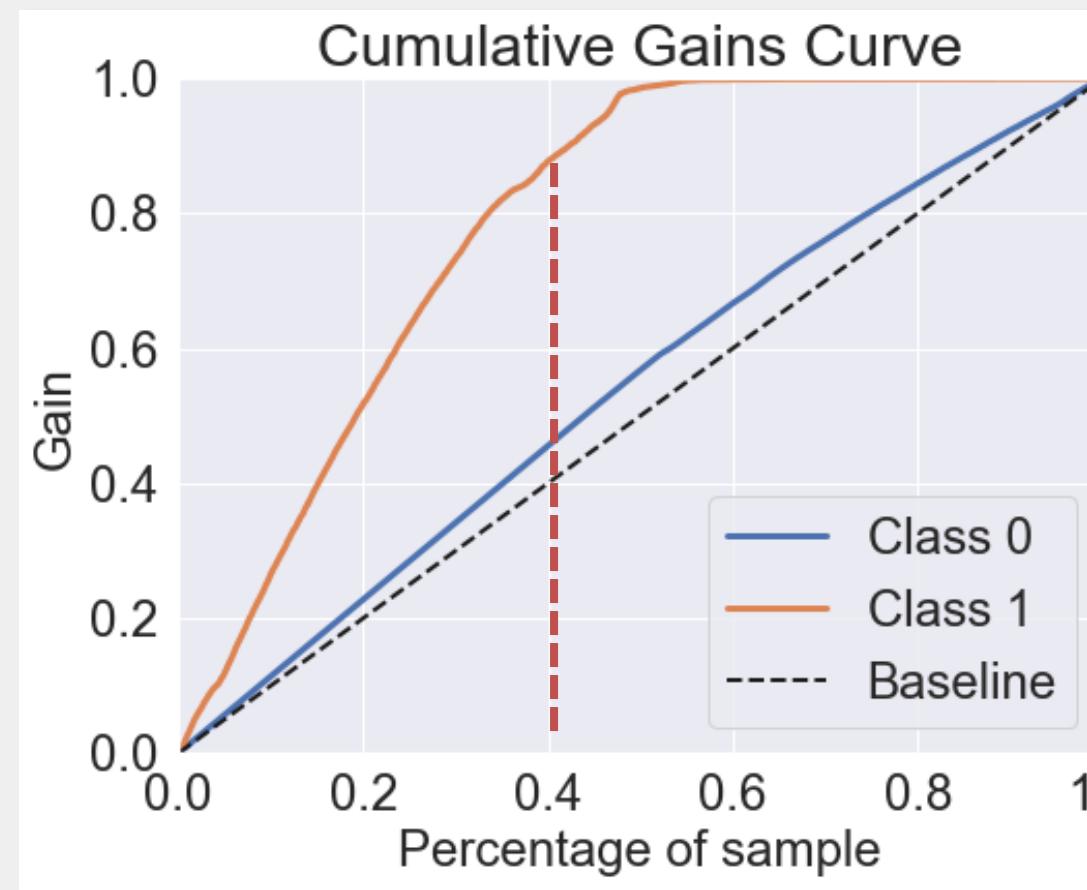


Data Test

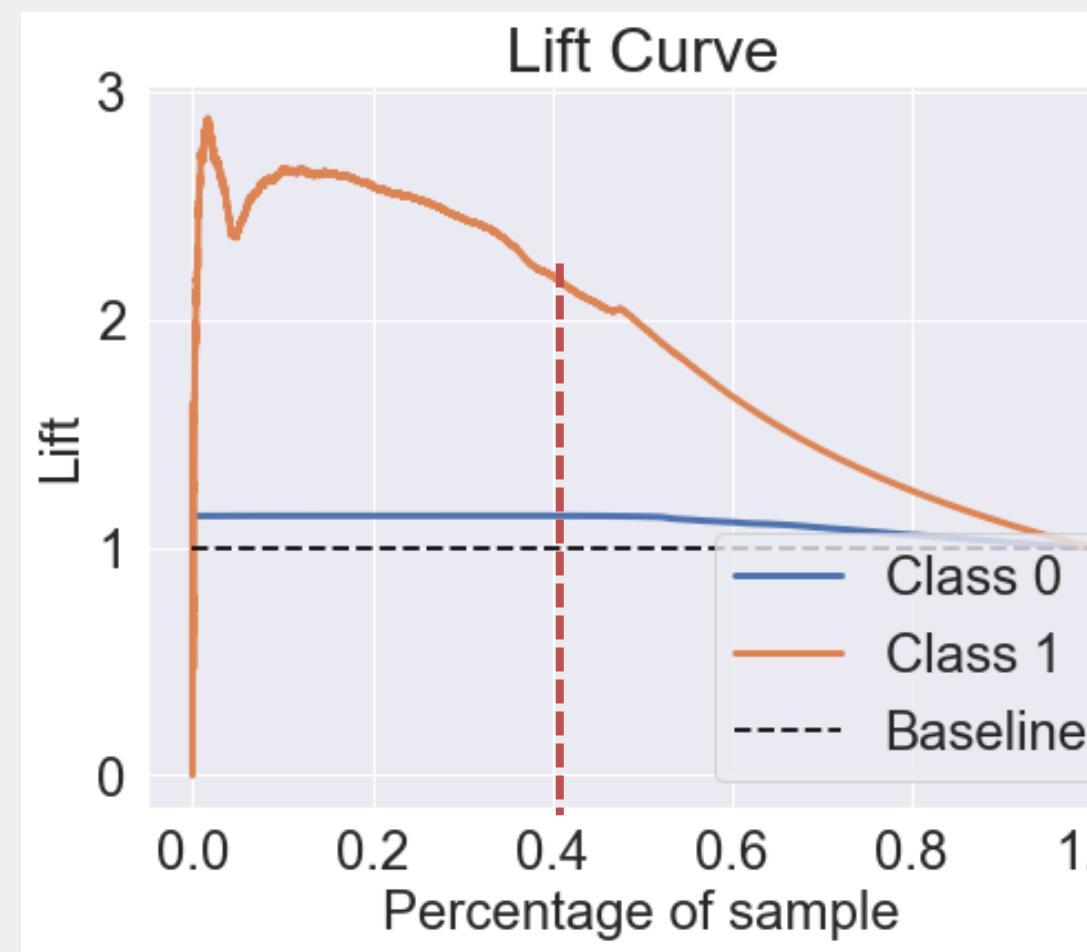
		Confusion Matrix Logreg Test (After Tuning Hyperparameter)	
Realita	No	39,568	27,312
	Yes	212	9,130
		No	Yes
		Prediksi	

Model	Test Recall	Test Precision	Test F1	Test AUC_ROC	
Logistic Regression (After Tuning)	97.73%	25.05%	39.88%	78.45%	

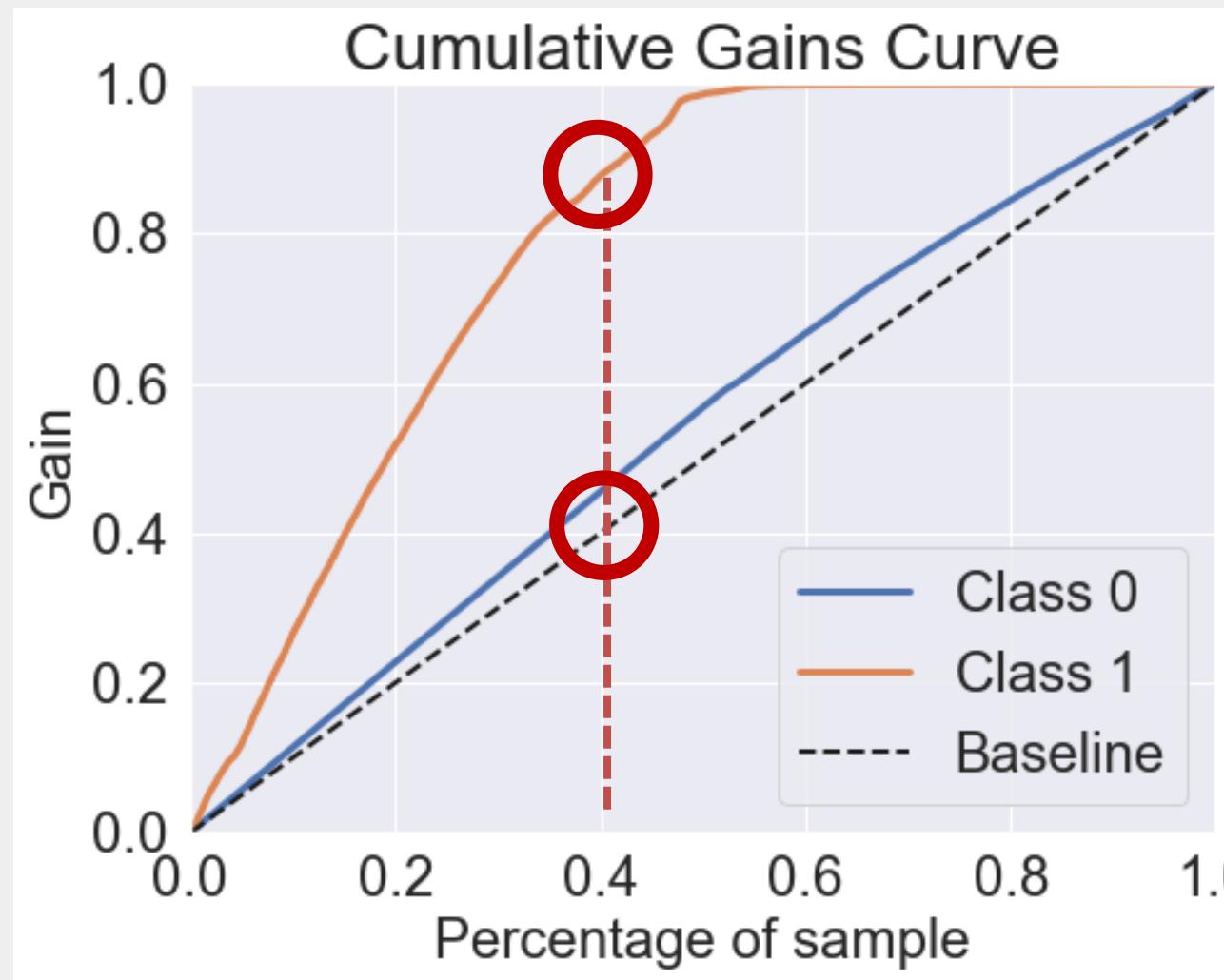
- Rate Interested Customer naik dari sebelumnya **12.26%** menjadi **25.05%**
- Secara keseluruhan apabila diliat dari matrix klasifikasi, kemampuan model untuk memprediksi pelanggan yang tertarik menggunakan Asuransi kendaraan cenderung rendah
- Untuk melihat apakah model masih bermanfaat jika diterapkan dalam bisnis, akan dilihat menggunakan analisis Gain dan Lift



- Dengan mengambil populasi 40% Probability teratas model mampu memprediksi hingga 90% customer yang tertarik asuransi dibandingkan Random Choice



- Ketika mengambil populasi 40% data berdasarkan model, perusahaan dapat menemukan nasabah yang berpotensi tertarik menggunakan asuransi kendaraan 2.2x lebih baik dibandingkan dengan random choice



Keterangan	Informasi
Populasi	76.222
Treat	40% x populasi = 30.489
Model based on Gain Curve	90%
Random Choice	40%
Nasabah Tertarik	Prediction * treat
Tidak tertarik	Treat – Nasabah tertarik

Keterangan	Model	Random Choice
Tertarik	27,440	12,196
Tidak Tertarik	3,049	18,293

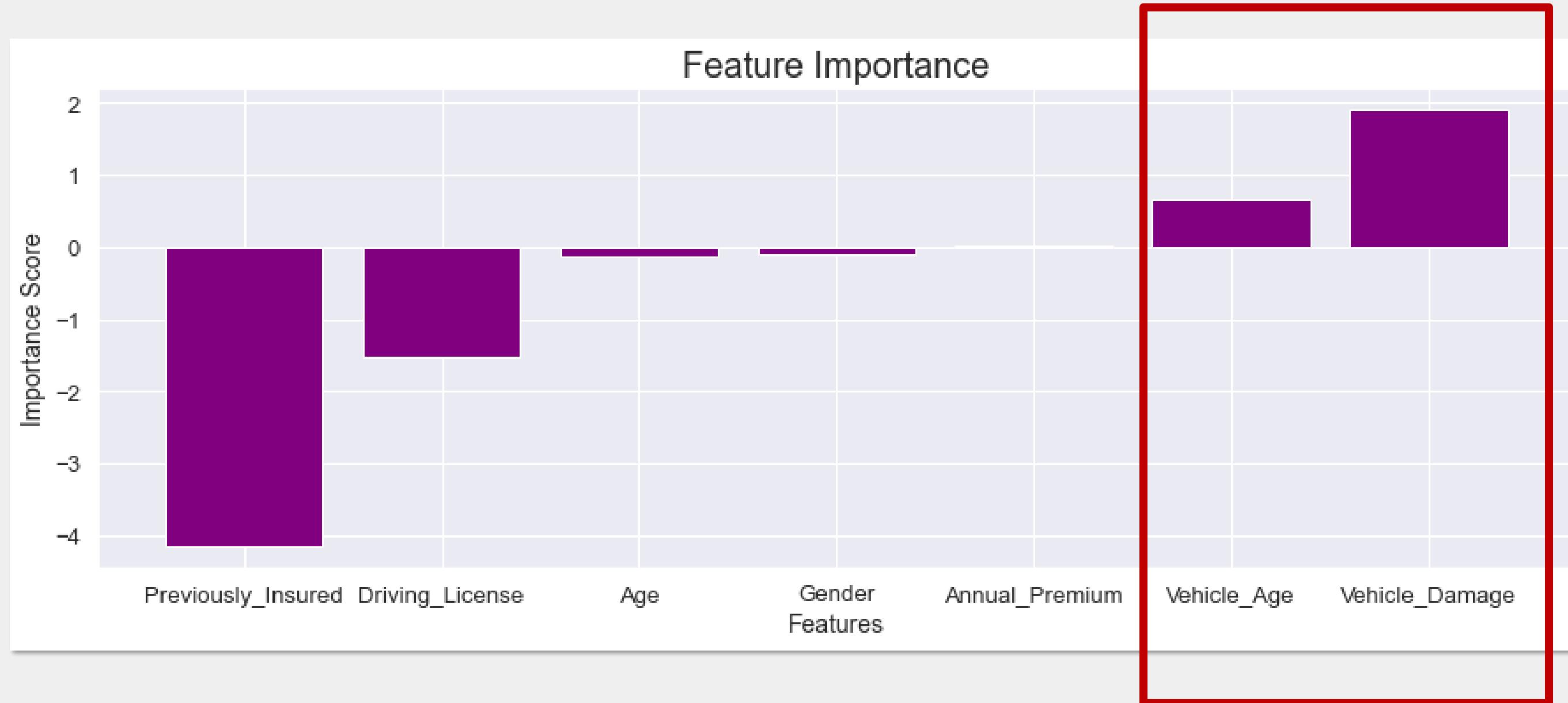
• Business Simulation

Dengan Model			
Jumlah Nasabah di Telfon			36,442
Yes			9,130
Premi	\$ 468	\$ 4,272,840	
Pengeluaran untuk campaign/ nasabah 7%	\$ 33	\$ -1,193,840	
Biaya Operational 18%	\$ 84	\$ -769,111	
Klaim 68%	\$ 318	\$ -2,905,531	
Profit Investmen 16%		\$ 683,654	
Profit		\$ 88,012	
Tanpa Model			
Jumlah Nasabah di Telfon			76,222
Yes			9,342
Premi	\$ 468	\$ 4,372,056	
Pengeluaran untuk campaign/ nasabah	\$ 33	\$ -2,497,033	
Biaya Operational 18%	\$ 84	\$ -786,970	
Klaim 68%	\$ 318	\$ -2,972,998	
Profit Investmen 16%		\$ 699,529	
Profit		\$ -1,185,416	

Source :

1. winsurtech
2. Howstuffworks (source: Insurance Information Institute])
3. website carinsurance

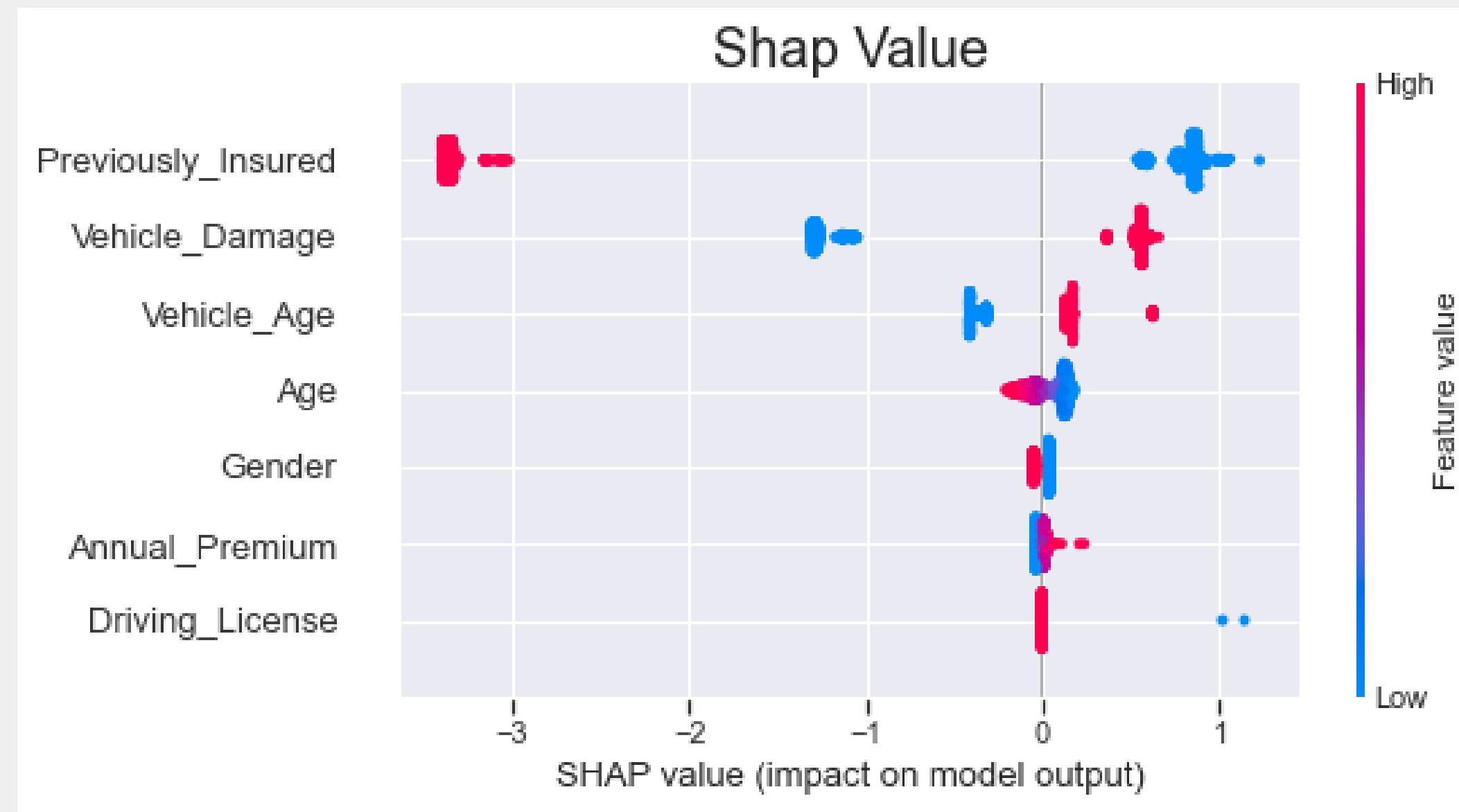
• Feature Importance



- Feature yang paling berpengaruh dari model yang dihasilkan menggunakan Logistic Regression. **Fitur Vehicle Damage dan Vehicle Age merupakan 2 fitur paling penting**
- Hal ini menggambarkan bahwa Kondisi Kendaraan nasabah (Rusak atau tidaknya) dan usia kendaraan nasabah berpengaruh terhadap keputusan nasabah dalam menggunakan asuransi kendaraan.

»

• Shap Values



- Jika seorang nasabah **pernah menggunakan asuransi kendaraan**, semakin menyebabkan nasabah tersebut **tidak mau** mengambil asuransi kendaraan
- Jika nasabah yang memiliki **kondisi kendaraan rusak** maka semakin menyebabkan nasabah tersebut **ingin** menggunakan asuransi kendaraan

»

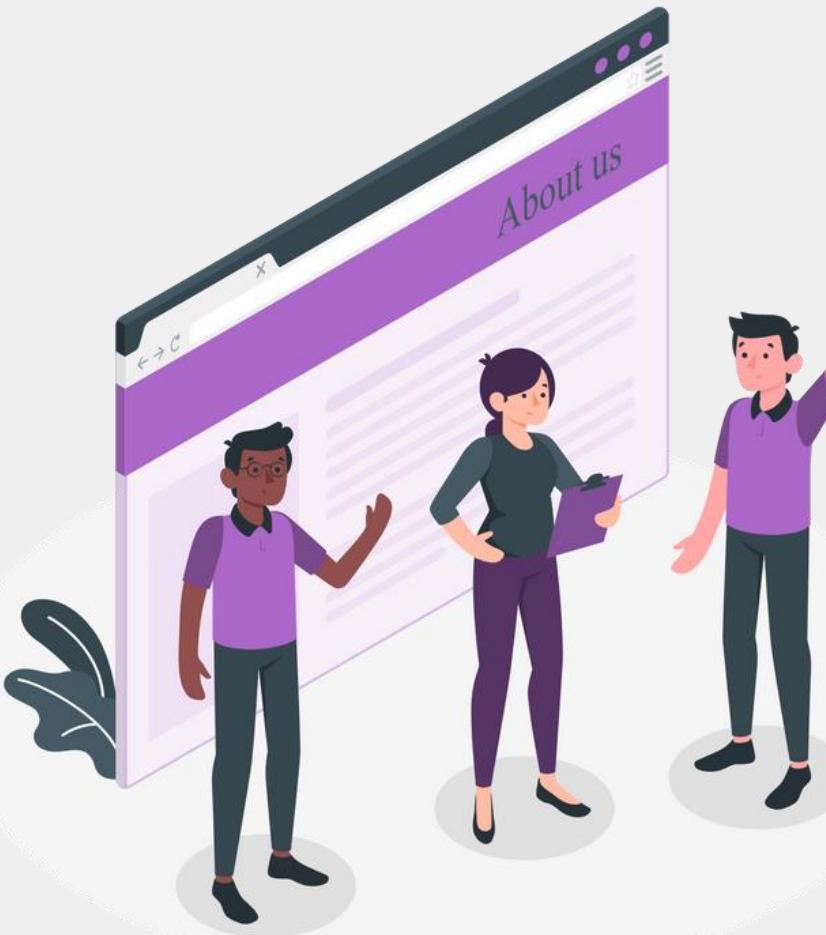
Business Recommendation



»

• Business Recommendation

1. Perusahaan dapat memfokuskan pada nasabah yang sebelumnya belum memiliki Asuransi Kendaraan dan kondisi kendaraannya rusak
2. Memberikan diskon pada beberapa kategori seperti :
 - drive safe: semakin sering berkendara semakin mendapat diskon
 - teen driver: diskon sekitar 25% untuk pengguna sebelum usia 25 tahun
 - steer clear: usia dibawah 25 tahun dan tidak ada catatan kecelakaan dapat diskon hingga 15%
3. Berpartisipasi dalam kegiatan sosial masyarakat dan memberikan edukasi bahwa tujuan asuransi adalah membantu memberikan rasa aman dan nyaman kepada masyarakat terkait resiko yang akan terjadi kedepannya terkait penggunaan kendaraan. Kemudian membuat kampanye kegiatan sosial untuk membantu perusahaan membangun citra positive yang lebih kuat dihadapan masyarakat
4. Bekerjasama dengan dealer dan showroom untuk menawarkan asuransi kepada konsumen mereka yang membeli mobil dengan cash terutama pada usia kendaraan 1-2 tahun



• Saran Penelitian Selanjutnya

Menambah beberapa fitur yang relevan seperti :

- Feature Vintage (lamanya customer dalam berasuransi)
- Loan (apakah nasabah memiliki pinjaman atau tidak)
- Jumlah tanggungan keluarga
- Jenis Pekerjaan
- Jenis Kendaraan dan Merk Kendaraan
- Jumlah mobil yang dimiliki nasabah



Hal tersebut dilakukan untuk melihat factor-faktor lain yang mempengaruhi kinerja model dan ketertarikan nasabah dalam merespon campaign car insurance ini

Diharapkan dengan menambah fitur-fitur tersebut, metric-metric lain yang nilainya masih kecil dapat diimprove.

Terimakasih

»