Rakamin Academy

DATA GEEKS


# Health Insurance Cross-Sell Prediction

Final Project Presentation

**Data Geeks**

# Problem Statement

PT Asuransi Jiwasaya

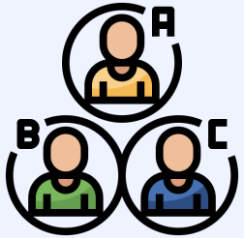
Product Expansion

Health Insurance

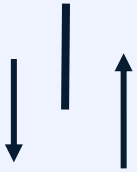
cross-selling

Vehicle Insurance

Marketing




Tertarik
12.3%



Tidak Tertarik
88%

## Goals

- Meningkatkan penjualan dari produk tambahan ‘Asuransi Kendaraan’

## Objectives

- Membangun model machine learning untuk memprediksi customer berpotensi tertarik dengan Asuransi Kendaraan.
- Meningkatkan *Conversion Rate*.
- Mengoptimalkan business model perusahaan PT Asuransi Jiwasaya sehingga efisiensi *resources*.

*Conversion Rate*

*Total Annual Revenue*

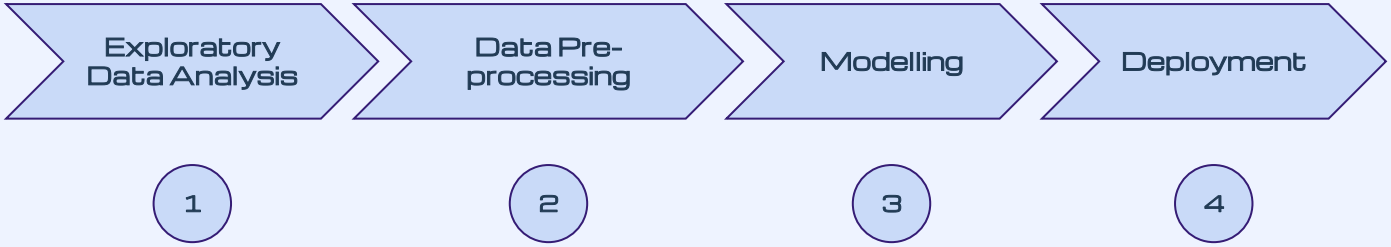
### *Conversion Rate*

Jumlah customer yang memutuskan untuk mengambil produk tambahan berupa ‘Asuransi Kendaraan’.

### Total *Revenue*

Jumlah peningkatan pendapatan.

# Modelling Flow




# About Dataset

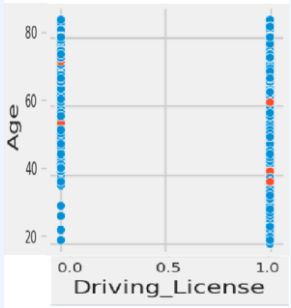
**Kaggle - Health Insurance Cross Sell Prediction**

- Dataset ini terdiri dari 12 kolom dan 381.109 baris.
- Terdapat 9 data numerik dan 3 data kategorikal.

```
<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
```

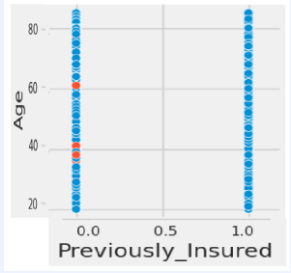
# Exploratory Data Analysis

Grafik antara **Age** dan **Driving_License** dengan variabel **Response**




Pelanggan yang **memiliki driving license** di usia **35** sampai **45** serta di usia **60** sampai **65** berpotensi ingin membeli **asuransi kendaraan**

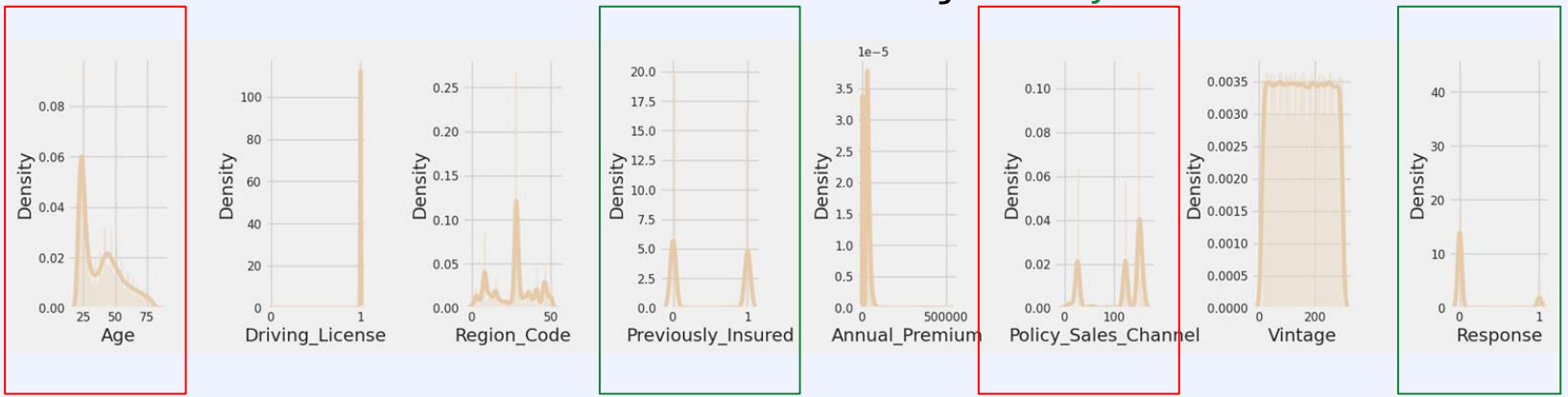
Grafik antara **Age** dan **Previously_Insured** dengan variabel **Response**




Pelanggan yang **tidak** pernah **menggunakan asuransi kendaraan** di usia **35** sampai **45** serta di usia **60** sampai **65** berpotensi tertarik membeli **asuransi kendaraan.**

**Note: 1 (response yes), 0 (response No)**

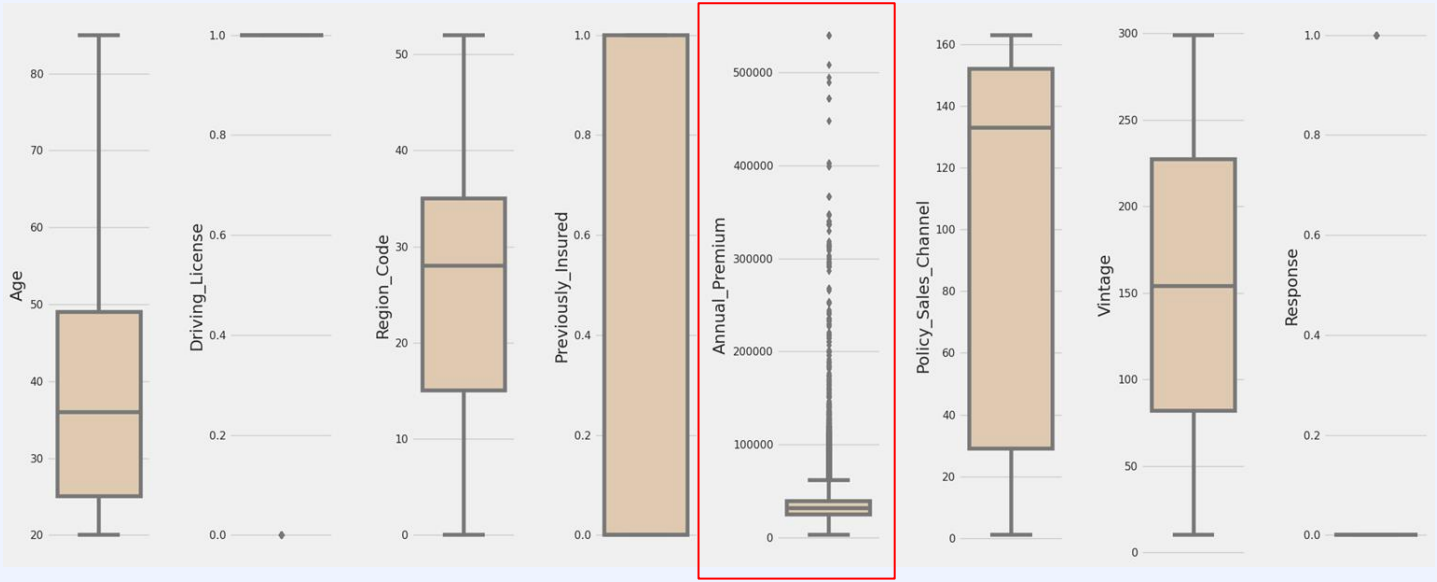
## Grafik antara kolom Numerik dengan Density




Dapat dilihat dari grafik diatas, Bahwa **hampir semua** memiliki hasil distribusi normal kecuali kolom **Age, Policy Sales Channel** yang memiliki **skewed**. Lalu pada grafik **response** dan **previously insured** memiliki grafik berupa **bimodal**

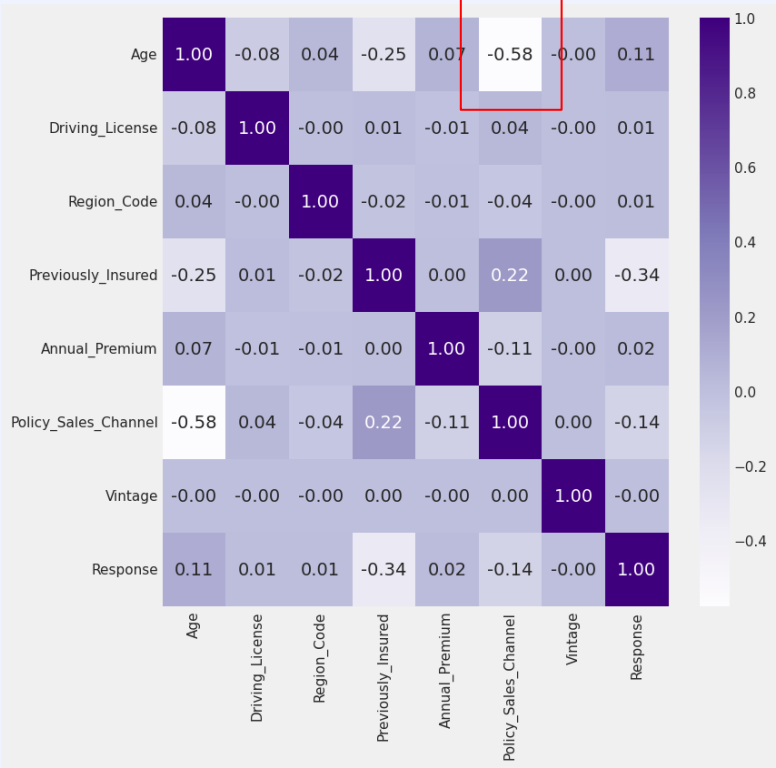
**Distribution Plot**

## Grafik kolom Numerik




Dapat dilihat dari grafik diatas, Bahwa **hampir** semua tidak memiliki hasil outlier kecuali kolom **Annual Premium.**

Grafik **Correlation Features**




- Tidak ada kolom yang memiliki **korelasi tinggi** terhadap target “response” (>0.5).
- Untuk kolom **Age** dengan **Policy Sales Channel** memiliki korelasi sekitar -0.58 menunjukkan bahwa ada **hubungan negatif** yang cukup kuat antara dua variabel

# Data Pre-processing

## 1. Pengelompokan dan Penentuan Target

Mengelompokkan kolom berdasarkan tipe datanya baik **numerik/kategorik**.

```
# Pengelompokan kolom berdasarkan jenisnya
nums = ['int64', 'int32', 'int16', 'float64', 'float32', 'float16']
nums = df.select_dtypes(include=nums)
nums.drop(columns=['id'],inplace=True)
nums = nums.columns
cats = ['Gender','Vehicle_Age','Vehicle_Damage']
```

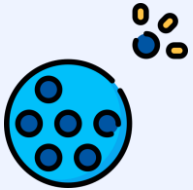
Target dari modelling classification ini adalah kolom `Response`.

## 2. Handling Missing Value

Dari 12 kolom **tidak ditemukan nilai kosong**.

## 3. Handling Duplicates Value

**Tidak terdapat data duplikasi** pada dataset



## 4. Handling Outliers

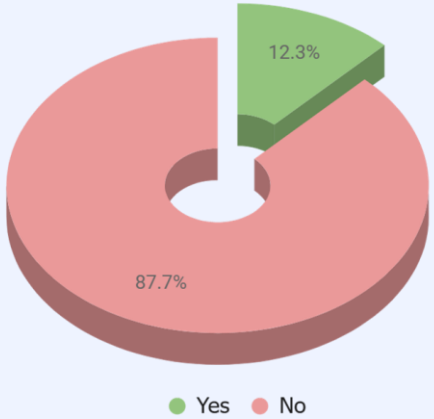
Terdapat kolom Annual_Premium (memiliki outliers) yang merupakan hal yang **normal jika terdapat outliers** sehingga **tidak dilakukan penghapusan outliers**. Hal ini juga didasarkan dengan pertimbangan **pembuatan model yang robust terhadap outliers**.

## 5. Feature Encoding

- **Vehicle_Damage** boolean -> integer (change data type)
- **Vehicle_Age** string -> integer (Label Encoding)
- **Gender** string -> integer (One Hot Encoding)

## 6. Class Imbalance

Class Imbalance pada data ini tergolong ***Moderate Imbalance***. Dilakukan ***undersampling***.




## 7. Feature Engineering

**Feature Extraction**, 4 features baru yakni **Age_Group**, **Premium_cat**, **Policy_cat**, dan **Region_cat**.

## 8. *Features Selection*

Feature yang dipilih untuk modeling adalah **Vehicle_Age, Vehicle_Damage, Region_cat, Previously_Insured, Age_Group, Policy_Sales_Channel, Gen_Female, Gen_Male.**

# Modelling

Pada modelling digunakan 7 algoritma klasifikasi yakni,

- Logistic Regression
- K-Nearest Neighbor
- Decision Tree
- XGBoost
- Random Forest
- LightGBM
- Gradient Boost

| Model | Accuracy Test | Accuracy Train | Precision Test | Precision Train | Recall Test | Recall Train | F1 Test | F1 Train | ROC AUC Test | ROC AUC Train | ROC AUC CrossVal Test | ROC AUC CrossVal Train |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logistic | 0.79 | 0.78 | 0.71 | 0.71 | 0.98 | 0.98 | 0.82 | 0.82 | 0.82 | 0.82 | 0.99 | 0.80 |
| KNN | 0.76 | 0.81 | 0.72 | 0.77 | 0.82 | 0.88 | 0.77 | 0.82 | 0.81 | 0.89 | 0.99 | 0.80 |
| Decision Tree | 0.72 | 0.95 | 0.71 | 0.93 | 0.75 | 0.97 | 0.73 | 0.95 | 0.73 | 0.99 | 0.99 | 0.80 |
| XGBoost | 0.79 | 0.80 | 0.72 | 0.73 | 0.93 | 0.94 | 0.81 | 0.82 | 0.83 | 0.87 | 0.99 | 0.80 |
| Random Forest | 0.73 | 0.95 | 0.71 | 0.92 | 0.76 | 0.98 | 0.73 | 0.95 | 0.81 | 0.99 | 0.99 | 0.80 |
| LightGBM | 0.79 | 0.79 | 0.72 | 0.73 | 0.93 | 0.94 | 0.81 | 0.82 | 0.84 | 0.85 | 0.99 | 0.80 |
| Gradient Boost | 0.79 | 0.79 | 0.72 | 0.72 | 0.93 | 0.94 | 0.82 | 0.82 | 0.84 | 0.84 | 0.99 | 0.80 |

Pada pembuatan model pertama dengan features, dimana dengan features ini **score recall** sudah baik namun **score AUC/ ROC** cukup *overfitting*.

```
features = ['Vehicle_Age','Vehicle_Damage','Previously_Insured','Gen_Female','Gen_Male','Age_Group','Region_cat','std_Annual_Premium']
target = ['Response']
```

Walaupun telah dilakukan *regularization* dan *hyperparameter tuning* lain **score AUC ROC** masih *overfitting*.

Sehingga dilakukan:

- ***features selection*** ulang
- penambahan data dengan ***sampling***.

```
# New Feature Selection
features_new = ['Vehicle_Age','Vehicle_Damage','Previously_Insured','Age_Group','Region_cat','Policy_Sales_Channel','Gen_Female','Gen_Male']
target_new = ['Response']
```

## Final Score Modelling - Hyperparameter Tuning,

| Model | Accuracy Test | Accuracy Train | Precision Test | Precision Train | Recall Test | Recall Train | F1 Test | F1 Train | ROC AUC Test | ROC AUC Train | ROC AUC CrossVal Test | ROC AUC CrossVal Train |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logistic | 0.78 | 0.79 | 0.71 | 0.71 | 0.96 | 0.95 | 0.82 | 0.82 | 0.87 | 0.87 | 0.93 | 0.92 |
| Decision Tree | 0.79 | 0.79 | 0.73 | 0.72 | 0.94 | 0.94 | 0.82 | 0.82 | 0.86 | 0.86 | 0.93 | 0.92 |
| XGBoost | 0.82 | 0.82 | 0.76 | 0.76 | 0.94 | 0.94 | 0.84 | 0.84 | 0.92 | 0.92 | 0.93 | 0.92 |
| Random Forest | 0.81 | 0.81 | 0.74 | 0.74 | 0.95 | 0.95 | 0.83 | 0.83 | 0.89 | 0.89 | 0.93 | 0.92 |
| LightGBM | 0.82 | 0.82 | 0.77 | 0.77 | 0.92 | 0.92 | 0.84 | 0.84 | 0.92 | 0.92 | 0.93 | 0.92 |
| Gradient Boost | 0.82 | 0.82 | 0.76 | 0.76 | 0.93 | 0.93 | 0.84 | 0.83 | 0.91 | 0.91 | 0.93 | 0.92 |

## Why **Recall** and **AUC ROC** score?

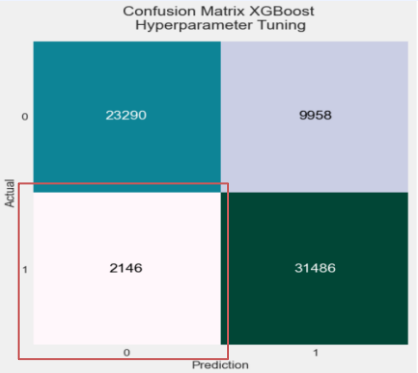
Dari data train tidak memiliki *class imbalance* serta business metrics yang ada,

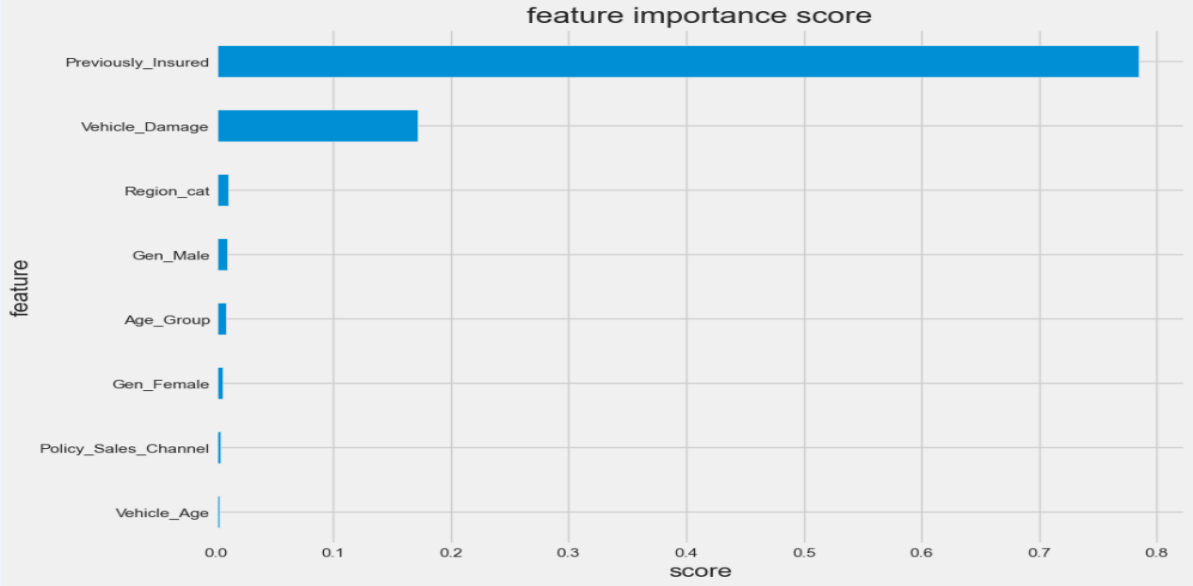
- Kita perlu **meminimalisir** ***False Negative*** (recall score).
- ROC AUC score untuk menilai sejauh mana **kepekaan model** dalam membedakan kelas (TPR dan FPR) meskipun *class* sudah seimbang.

# The Best Fit Model

## XGBoost Model

Recall : **0.94**
AUC ROC : **0.92**
AUC ROC CV : **0.91**
Model tidak overfit maupun underfit yang dapat disebut sebagai ***model best fit***.



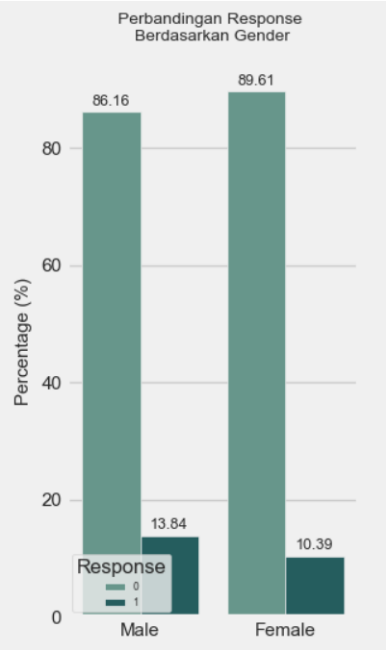

2 features yang memiliki importance terbesar adalah

- Previously_Insured
- Vehicle_Damage

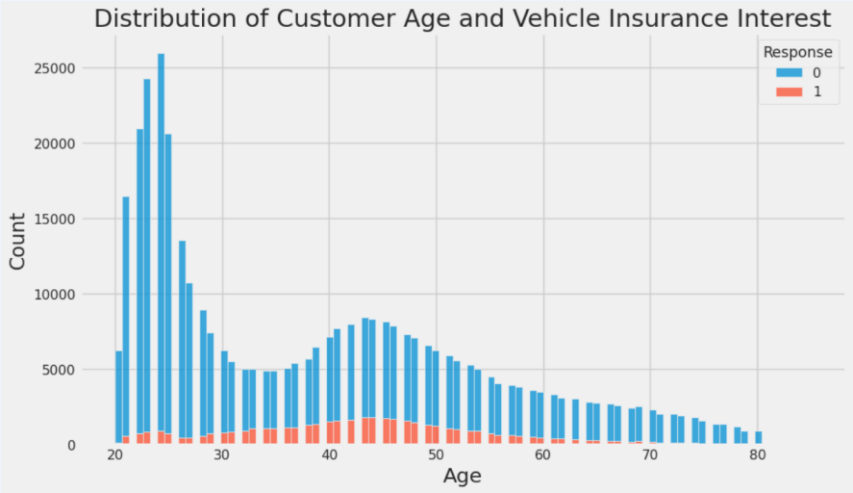
Features yang paling rendah

- Policy_Sales_Channel
- Vehicle_Age.

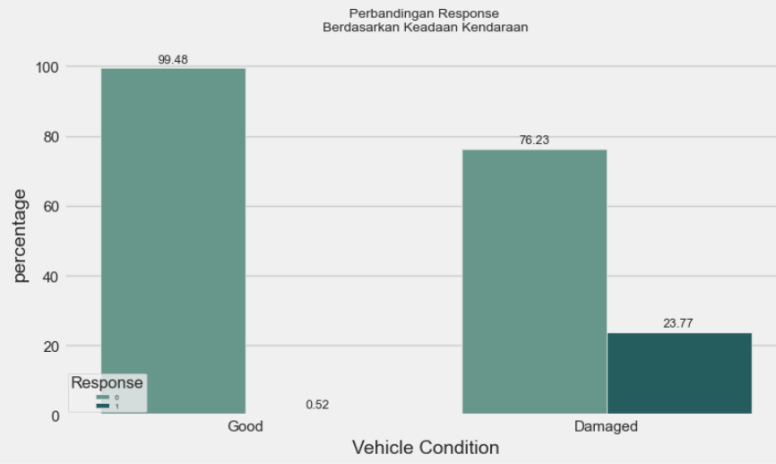
# Insight



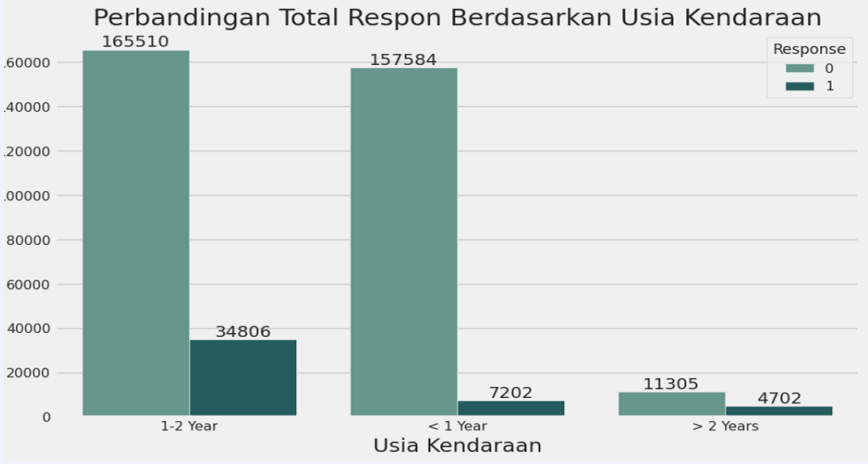

**Asuransi kendaraan** lebih diminati oleh pelanggan **laki-laki** sebanyak **13.84%** dibandingkan **perempuan** hanya **10.39%.**



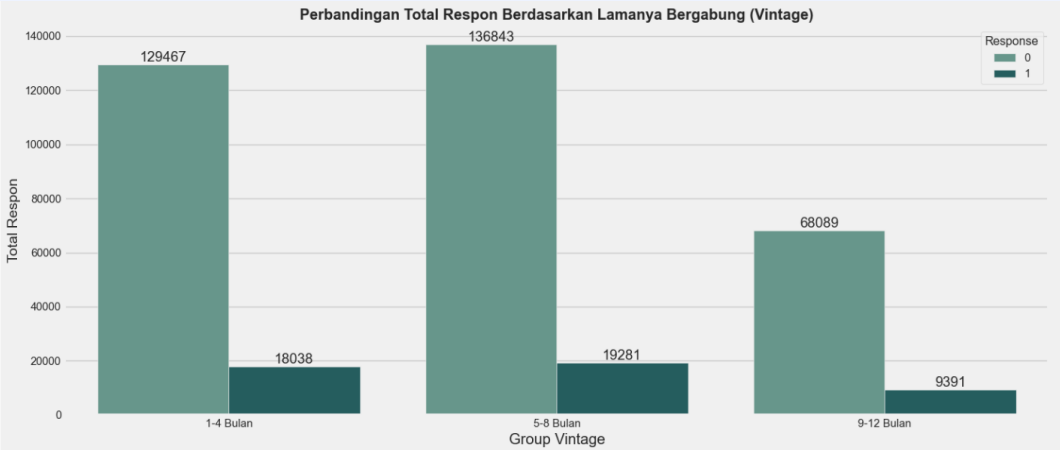

**Asuransi kendaraan** lebih diminati oleh pelanggan berusia **30-50** tahun.



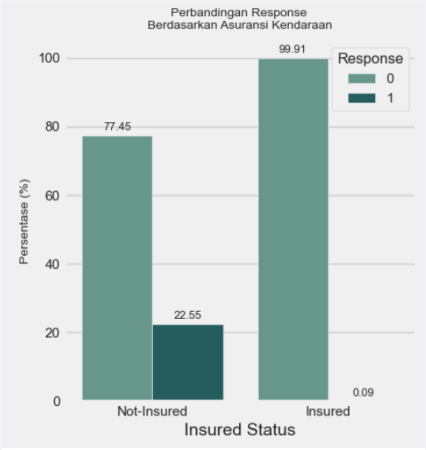

Dapat dilihat dari grafik, pelanggan yang memiliki **kondisi** kendaraan yang **"kurang baik"** lebih **berminat** terhadap **'Asuransi Kendaraan'** dengan total **23.77%** dibandingkan pelanggan yang memiliki **kendaraan** tergolong **bagus**, hanya **0.52%.**



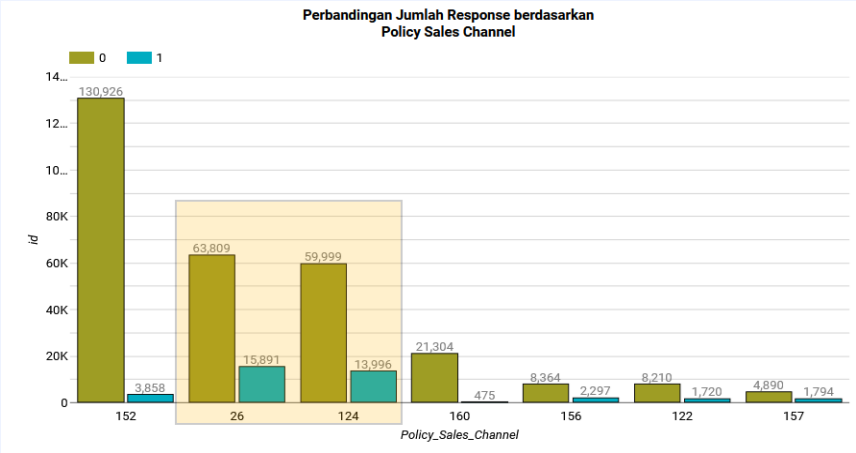

Berdasarkan grafik diatas, pelanggan dengan usia **kendaraan 1-2 tahun** memiliki ketertarikan untuk membeli asuransi kendaraan.

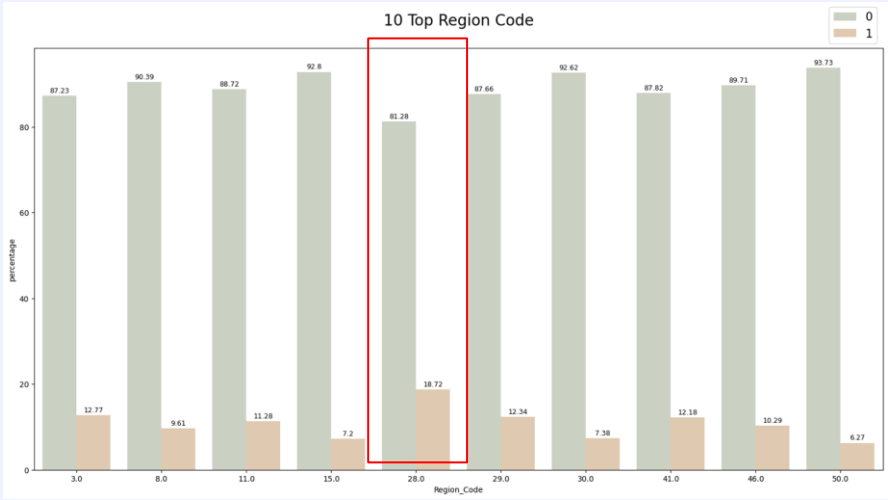



Pelanggan dengan vintage selama **5-8 bulan memiliki ketertarikan** lebih **banyak** dibandingkan dengan yang lain.




Dapat dilihat dari grafik, Pelanggan yang **belum** memiliki **'Asuransi Kendaraan'** memiliki minat terhadap produk cross-selling sekitar 20% dari total pelanggan yang belum memiliki asuransi kendaraan.




Berdasarkan column policy sales channel didapatkan **3 sales channel yang banyak digunakan** yaitu channel 124, 26 serta 152.

Berdasarkan Top 10 Region Code **potential customers terbanyak ada di Region 28** lalu diikuti dengan 8, 46, 41, 15, 30, 29, 50, 3, dan 11.

# Business Recommendation

## 01

Perusahaan dapat bekerja sama dengan bengkel-bengkel dalam pemasaran dan membuat penawaran produk yang bervariatif berdasarkan kebutuhan customers.

## 02

Memberikan 2x/tahun discount service berkala/sparepart kepada 50 customers utama dan diberikan rincian perbandingan biaya service berkala/risk yang lain antara yang memiliki asuransi dan tidak.

## 03

Fokus pada segmen pasar Middle Ages 31 - 45, untuk menawarkan produk yang sesuai dengan kebutuhan dan preferensi mereka. Misalnya, produk asuransi yang memberikan perlindungan lebih luas, fleksibilitas pembayaran, dan kemudahan klaim.

## 04

Pemasaran fokus pada Top 10 Region (28,  8, 46, 41, 15, 30, 29, 50, 3, dan 11) serta menggunakan media pemasaran di *channel* 152, 26 dan 124.

## 05

Melakukan *campaign* penawaran produk dan edukasi untuk meningkatkan awareness akan asuransi kendaraan untuk jangka panjang.

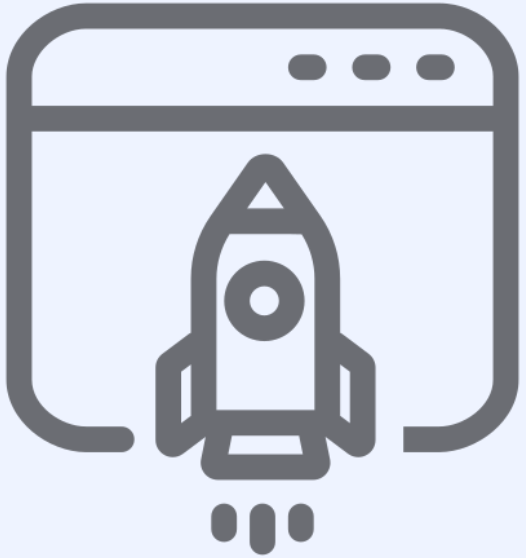
## 06

Melakukan *campaign* mengenai pentingnya asuransi kendaraan kepada customers yang belum memiliki 'Asuransi Kendaraan'.

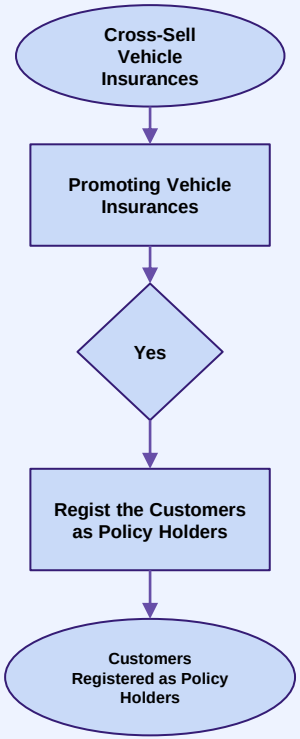
# Business Recommendation

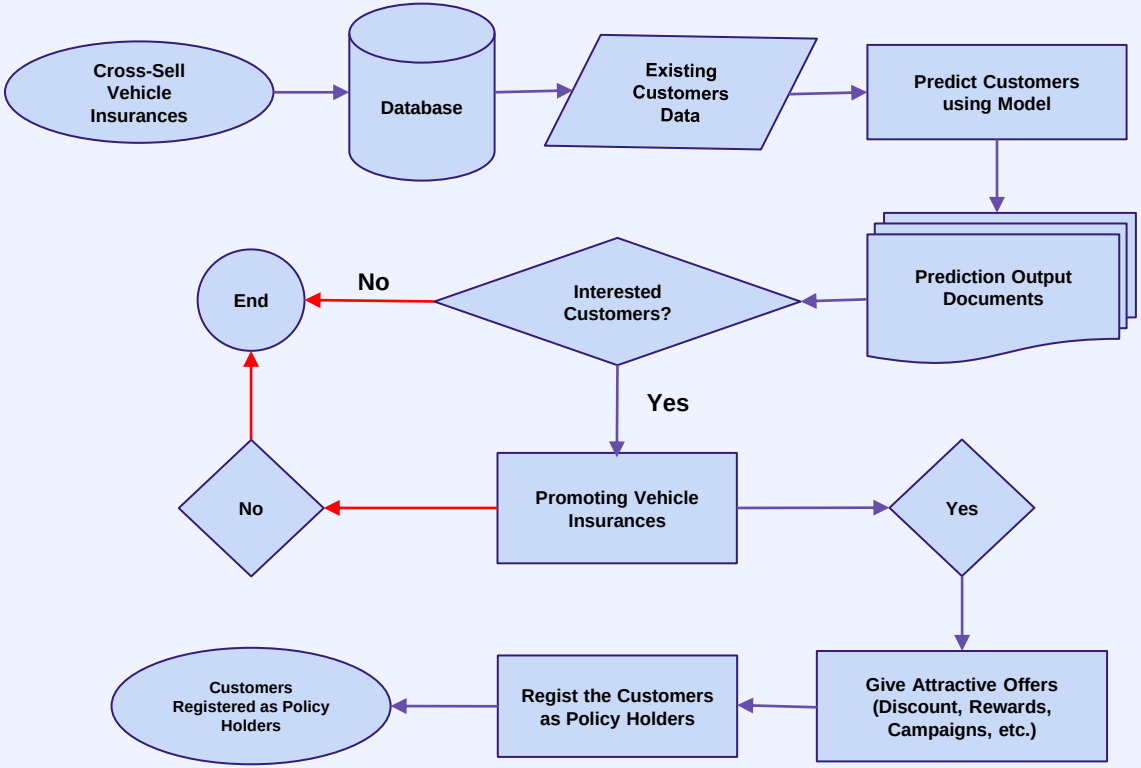
## Deployment Model ke Sistem Asuransi

- Model dapat digunakan untuk mendeteksi seorang customer yang mempunyai kriteria sebagai potential buyer.
- Dengan adanya prediksi tersebut akan meningkatkan Jumlah customer yang memutuskan untuk mengambil produk tambahan berupa ‘Asuransi Kendaraan’.
- Model dapat mengurangi biaya tenaga kerja.
- Model dapat memprediksi potential buyer dengan tepat, tidak perlu untuk mengiklankan ke semua customer, hanya mengiklankan ke customer potential buyer saja.

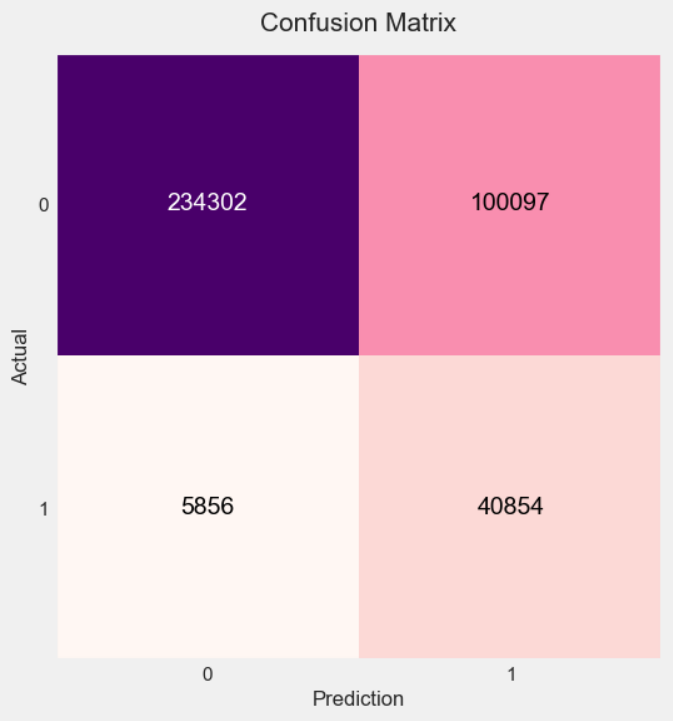


# Business Flow Simulation



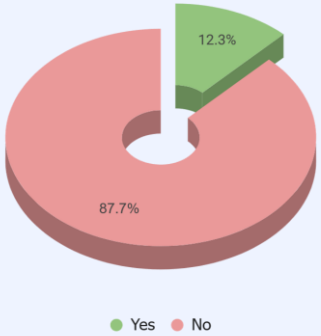



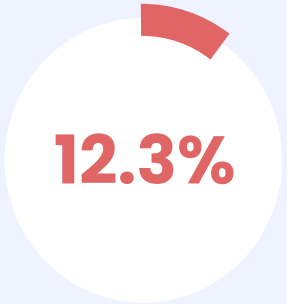
# Hasil prediksi dari penerapan model XGBoost terhadap data 381.109 existing customers.



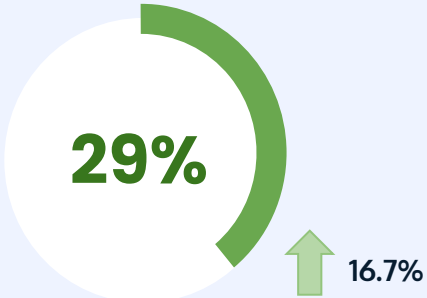
# Conversion Rate

## Response ‘Yes’ terhadap penawaran yang dilakukan

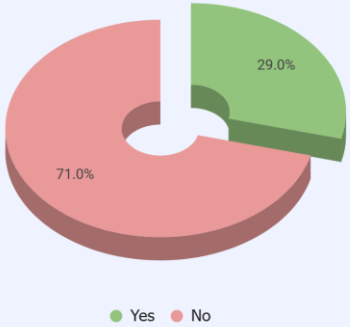



Before




After




Dengan menawarkan ke target yang memang terprediksi sebagai potential customers maka conversion rate meningkat menjadi 29%.

# Total Revenue

## Melakukan pemasaran kepada seluruh Customer Existing.

| No | Keterangan | Jumlah | Biaya | Total |
|---|---|---|---|---|
| 1 | Gaji Pegawai PIC / 1000 | 385 | Rp3,000,000 | Rp1,155,000,000 |
| 2 | Biaya Campaign CPL | 381109 | Rp100,000 | Rp38,110,900,000 |
| 3 | Biaya Promosi | 1 | Rp10,000,000 | Rp10,000,000 |
| 4 | Promo 50 Customers Pertama | 50 | Rp400,000 | Rp20,000,000 |
| 5 | Biaya Listrik, Pulsa, dll. per PIC | 385 | Rp500,000 | Rp192,500,000 |
| 6 | Perlengkapan | 1 | Rp2,000,000 | Rp2,000,000 |
| 7 | Penyusutan | 1 | Rp1,000,000 | Rp1,000,000 |
| 8 | Pemeliharaan | 1 | Rp1,000,000 | Rp1,000,000 |
| | | | **Total Anggaran** | Rp39,492,400,000 |

Premi yang dibayarkan oleh customers sebesar Rs 5000 (Rp 950.000)

| Jumlah Customers | Harga | Pendapatan | Pengeluaran | Revenue |
|---|---|---|---|---|
| 46710 | Rp950,000 | Rp44,374,500,000 | Rp39,492,400,000 | **Rp4,882,100,000** |

# Total Revenue

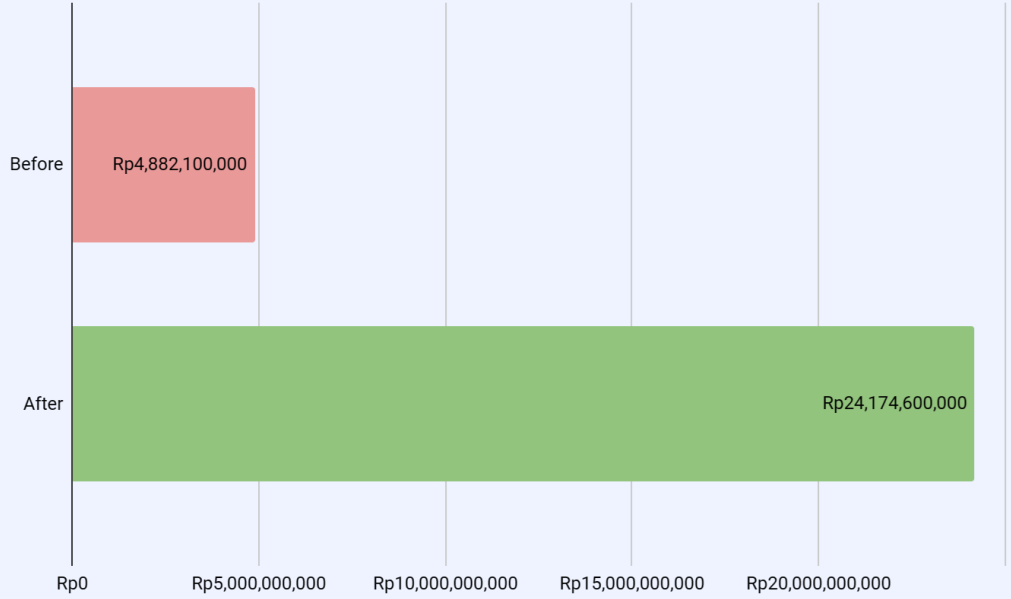
## Melakukan pemasaran kepada Potential Customers di hasil prediksi sebelumnya.

| No | Keterangan | Jumlah | Biaya | Total |
|---|---|---|---|---|
| 1 | Gaji Pegawai PIC / 1000 | 145 | Rp3,000,000 | Rp435,000,000 |
| 2 | Biaya Campaign CPL | 140952 | Rp100,000 | Rp14,095,200,000 |
| 3 | Biaya Promosi | 1 | Rp10,000,000 | Rp10,000,000 |
| 4 | Promo 50 Customers Pertama | 50 | Rp400,000 | Rp20,000,000 |
| 5 | Biaya Listrik, Pulsa, dll. per PIC | 145 | Rp500,000 | Rp72,500,000 |
| 6 | Perlengkapan | 1 | Rp2,000,000 | Rp2,000,000 |
| 7 | Penyusutan | 1 | Rp1,000,000 | Rp1,000,000 |
| 8 | Pemeliharaan | 1 | Rp1,000,000 | Rp1,000,000 |
| | | | **Total Anggaran** | Rp14,636,700,000 |

Premi yang dibayarkan oleh customers sebesar Rs 5000 (Rp 950,000)

| Jumlah Customers | Harga | Pendapatan | Pengeluaran | Revenue |
|---|---|---|---|---|
| 40854 | Rp950,000 | Rp38,811,300,000 | Rp14,636,700,000 | **Rp24,174,600,000** |

# Total Revenue




Dengan melakukan pemasaran kepada potential customers, biaya yang digunakan lebih efisien dan total revenue meningkat hingga 83%.

# Thank you

Arini Arumsari
GitHub, LinkedIn