Rakamin Academy

A Six Insurance

Dokumen Laporan Final Project

Background

Stage 0

1. Apa problem yang ingin diselesaikan dari dataset tersebut?

A-Six Insurance merupakan perusahaan asuransi yang berkomitmen untuk memberikan perlindungan dan keamanan finansial bagi pelanggan. Kami menyediakan jenis asuransi, termasuk Asuransi Kesehatan dan Asuransi Kendaraan, dengan penawaran premi yang kompetitif dan layanan yang unggul. Dengan pengalaman dan keahlian dalam industri asuransi, A-Six Insurance siap memberikan solusi asuransi yang tepat bagi kebutuhan individu maupun bisnis.

Berdasarkan data dari Korlantas Polri yang dipublikasikan Kementerian Perhubungan, angka kecelakaan lalu lintas di Indonesia mencapai 103.645 Kasus pada tahun 2021. Jumlah tersebut lebih tinggi dibandingkan data tahun 2020 yang sebanyak 100.028 kasus.

Lalu, berdasarkan data dari Gabungan Industri Kendaraan Bermotor Indonesia (GAIKINDO) menyatakan bahwa penjualan kendaraan bermotor meningkat 18,1% YoY menjadi 1.048.040 unit. Direktur AAUI juga menyatakan bahwa premi kendaraan bermotor meningkat sebesar 19,4% (Rp.10,9 T menjadi Rp.13 T) dari tahun 2021 ke 2022 (**kontan.co.id, 2023**).

Mempertimbangkan hal-hal tersebut, A-Six Insurance menawarkan program asuransi sesuai dengan kebutuhan pelanggan, yaitu Asuransi Kendaraan. melihat tingkat angka kecelakaan lalu lintas, program asuransi ini bisa menjadi opsi yang menarik bagi pelanggan yang ingin mengurangi dampak kerugian akibat kecelakaan yang tak terduga.

Dalam hal ini, A-Six Insurance melakukan pengembangan produk melalui strategi Cross Selling dengan jenis User Based, Cross Selling merupakan kegiatan menawarkan produk lain kepada nasabah yang sudah menggunakan produk tertentu. dimana A-Six Insurance mengidentifikasi kebutuhan atau preferensi pelanggan untuk menawarkan produk tambahan yang relevan dengan kebutuhan mereka. Nasabah A-Six Insurance yang sudah memiliki asuransi kesehatan dan atau sebelumnya memiliki asuransi kendaraan tetapi membatalkannya. (**Bold Commerce:How to use cross-sells and upsells to increase revenue on shopify**).

Untuk melakukan Cross Selling team marketing harus mengidentifikasi data nasabah dan kebutuhan mereka. Namun, team marketing mengalami kesulitan untuk mengidentifikasi data tersebut. Maka dari itu, team data talent diperlukan untuk membantu mengolah data dan melakukan klasifikasi nasabah dari data yang tersedia. Dengan demikian, A-Six Insurance dapat meningkatkan revenue melalui peningkatan penjualan produk asuransi baru melalui strategi Cross Selling yang efektif.

Stage 0

3. Apa Goal yang ingin dicapai?

Tujuan utama A-Six Insurance adalah meningkatkan revenue melalui peningkatan penjualan produk asuransi baru melalui strategi Cross Selling yang efektif berdasarkan analisis data nasabah.

4. Apa objektif yang sesuai dengan goal tersebut?

Perusahaan A Six akan membangun model Machine Learning untuk mengidentifikasi calon nasabah yang memiliki potensi untuk membeli asuransi kendaraan. Dengan demikian, perusahaan dapat lebih efektif dalam menargetkan pelanggan yang berpotensi untuk membeli produk asuransi kendaraan dan meningkatkan penjualan secara keseluruhan.

5. Apa Business Metrics yang cocok untuk data tersebut?

Berdasarkan data yang tersedia guna untuk menjangkau nasabah tertentu dan mengoptimalkan model bisnis., business metrics yang dipilih adalah Conversion Rate dimana persentase jumlah nasabah yang tertarik untuk membeli polis asuransi kendaraan dari total nasabah yang dimiliki..

Descriptive Statistics

Data Cleansing

Pertama kita cek apakah ada data memiliki baris yang kosong atau duplikat.

- Untuk memeriksa data yang kosong, kami menggunakan df.isna()sum. Dari hasil tersebut ditemukan bahwa tidak ada data yang null.

```
df.isna().sum()
```

```
id                    0
Gender                0
Age                   0
Driving_License       0
Region_Code           0
Previously_Insured    0
Vehicle_Age           0
Vehicle_Damage        0
Annual_Premium        0
Policy_Sales_Channel  0
Vintage               0
Response              0
dtype: int64
```

Tidak ada kolom yang memiliki nilai kosong pada dataset.

Descriptive Statistics

```
df.info()
```

```
<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
```

- Data terdiri dari 381109 baris dan 12 kolom
- Tidak ada kolom yang memiliki nilai kosong (terlihat dari non-null count pada setiap kolom sama, yaitu 381109).

Descriptive Statistics

```
df[nums].describe().round(2)
```

	Age	Annual_Premium	Vintage
count	381109.00	381109.00	381109.00
mean	38.82	30564.39	154.35
std	15.51	17213.16	83.67
min	20.00	2630.00	10.00
25%	25.00	24405.00	82.00
50%	36.00	31669.00	154.00
75%	49.00	39400.00	227.00
max	85.00	540165.00	299.00

Data Numerik

- Nilai Mean dan Median (50%) tidak terdapat keanehan dikarenakan nilainya tidak terlalu jauh.
- Nilai Min dan Max tidak terdapat keanehan, dikarenakan rentangnya yang masih masuk akal dari tiap fitur.

Descriptive Statistics

```
df[cats].describe()
```

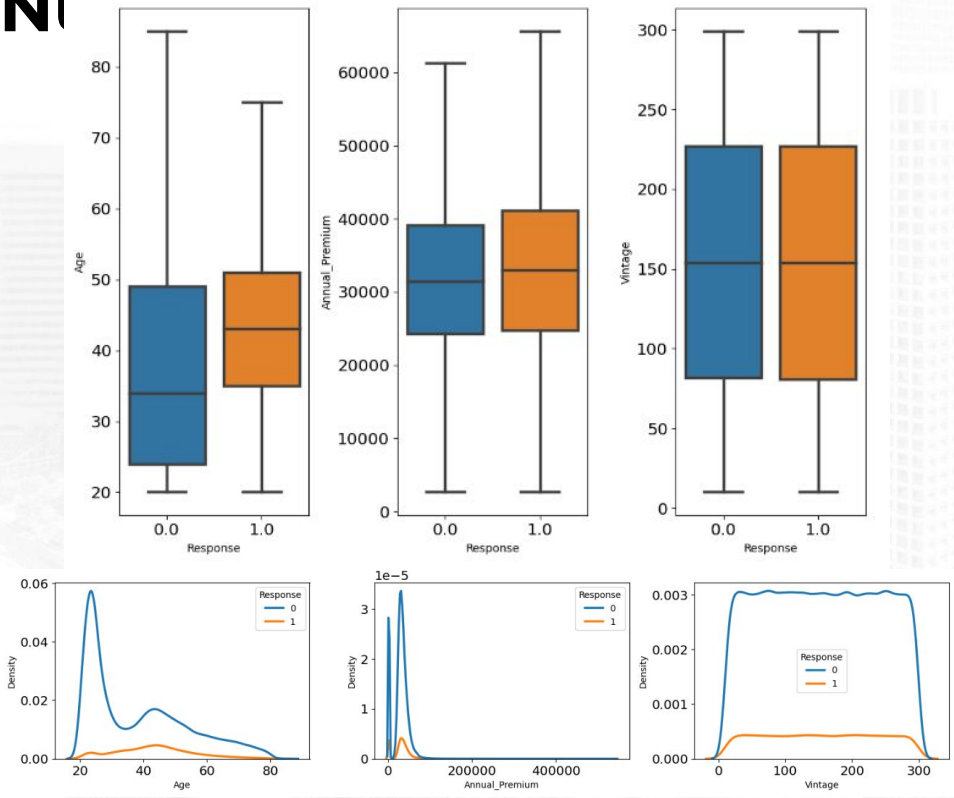
	Gender	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Policy_Sales_Channel	Response
count	381109	381109	381109.0	381109	381109	381109	381109.0	381109
unique	2	2	53.0	2	3	2	155.0	2
top	Male	1	28.0	0	1-2 Year	Yes	152.0	0
freq	206089	380297	106415.0	206481	200316	192413	134784.0	334399

Data Kategorik

- Statistik dari **Unique, Top, dan Freq** tidak terdapat keanehan, karena dari masing-masing fitur sudah memiliki kategori yang sesuai, nilai top dan frekuensi yang sesuai. Seperti **Gender** (Male dan Female; Top nya adalah Male dengan Frekuensi 206.089), **Region_Code** (ada 53 ragam daerah, dengan Top di daerah ke-58 dengan frekuensi sebanyak 106.415),
- **Driving_License, Region_Code, Previously_Insured, Policy_Sales_Channel** memiliki proporsi frekuensi top yang signifikan.

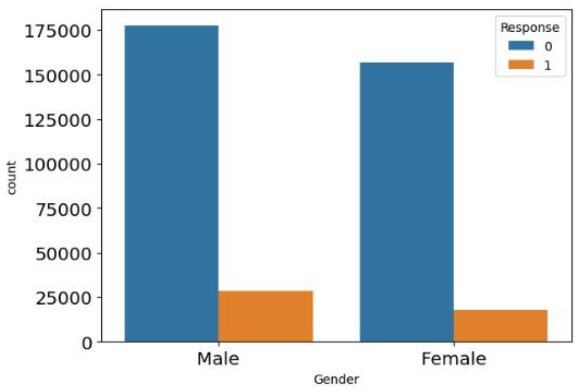
Univariate Analysis

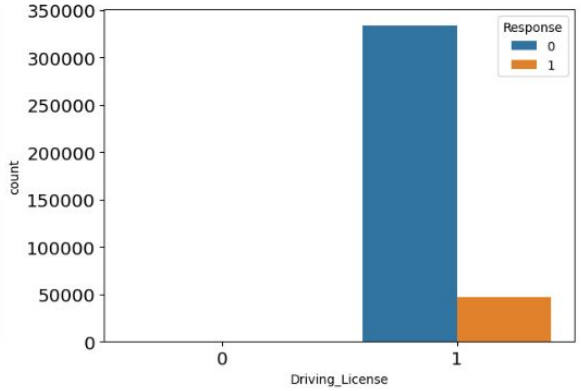
Individual Boxplot dan Displot (N

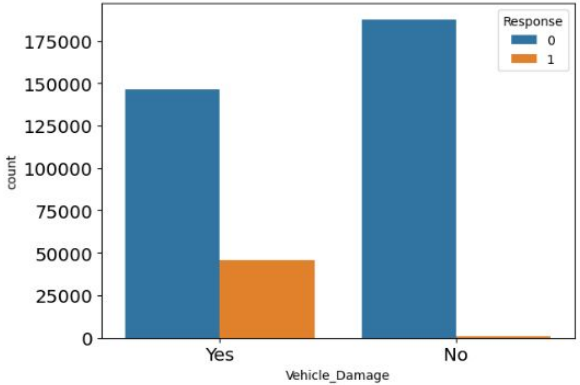


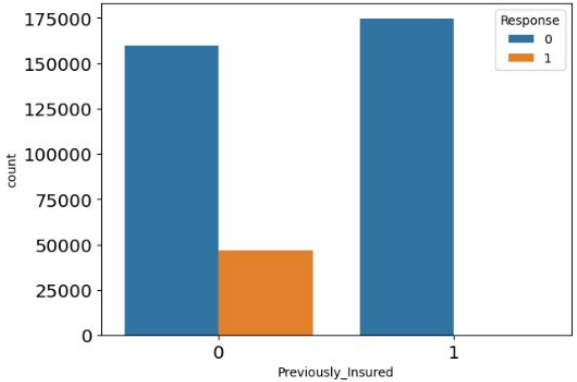
- Variabel **Age** yang tidak tertarik menggunakan asuransi kendaraan sebagian besar berada direntang umur 20-30 tahun (Response 0), sedangkan yang tertarik (Response 1) sebagian besar direntang 40-50 tahun.
- Variabel **Annual_Premium** baik yang tertarik maupun tidak tertarik berada direntang 25.000 - 40.000.
- Variabel **Vintage** memiliki distribusi yang hampir sama antara yang tertarik dengan yang tidak tertarik.

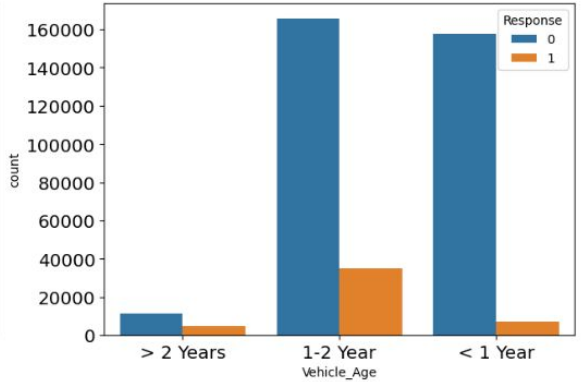
Individual Countplot (Categorical)





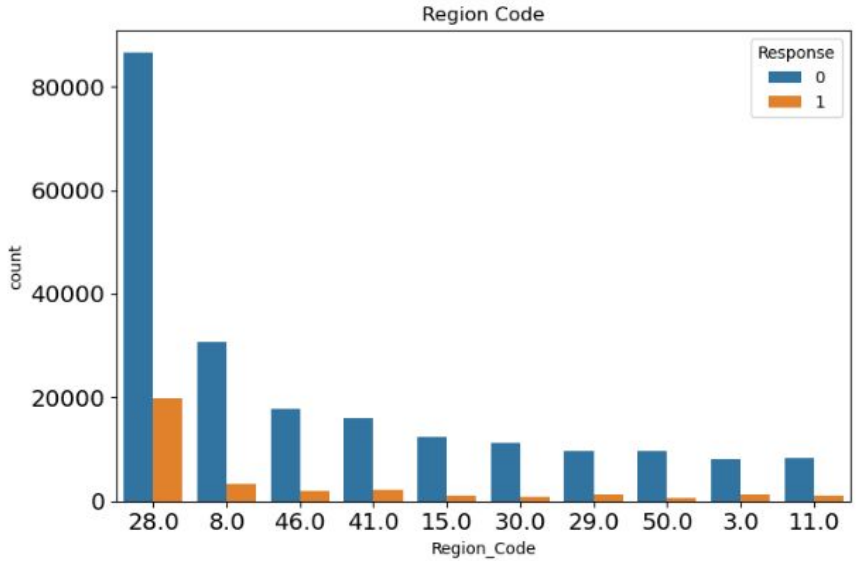


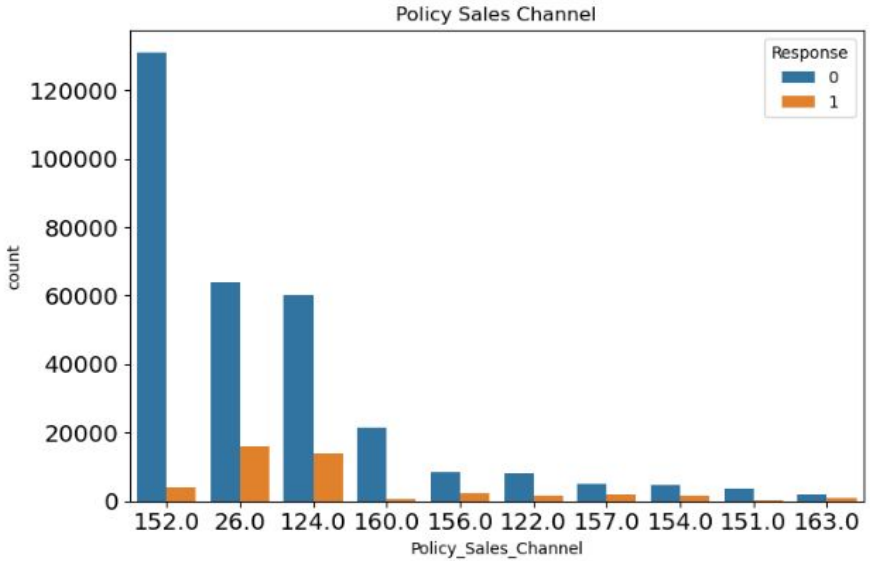




- Pada fitur **Previously_Insured**, hanya customer yang belum pernah memiliki asuransi kendaraan yang tertarik untuk mengambil asuransi kendaraan.

Individual Countplot (Categorical)





DATA VISUALISASI

Berdasarkan visualisasi data di atas dapat disimpulkan sebagai berikut:

- **Gender**: Terdapat lebih banyak pria yang tidak tertarik terhadap penawaran asuransi dibandingkan wanita.
- **Driving_License**: Hampir semua pelanggan memiliki lisensi mengemudi dan hanya sedikit yang tidak tertarik terhadap penawaran asuransi.
- **Region_Code**: Terdapat beberapa daerah yang lebih tertarik terhadap penawaran asuransi dibandingkan daerah lain.
- **Previously_Insured**: Pelanggan yang sudah memiliki asuransi di tempat lain cenderung tidak tertarik terhadap penawaran asuransi baru.
- **Vehicle_Age**: Kendaraan yang berusia 1-2 tahun lebih tertarik terhadap penawaran asuransi dibandingkan kendaraan yang lebih tua atau lebih baru.
- **Vehicle_Damage**: Pelanggan yang kendaraannya rusak lebih tertarik terhadap penawaran asuransi dibandingkan yang kendaraannya tidak rusak.

Follow-up Pre-Processing

- **Memproses outlier.**

  Untuk features yg sebagian besar datanya merupakan outlier jika seandainya pada tahap preprocessing diputuskan untuk melakukan pembersihan terhadap outlier tersebut, sebaiknya dilakukan dengan menetapkan sebuah treshold. Misalkan dengan menetapkan treshold yaitu percentile 95% atau 99%. Sehingga dengan demikian data yang terbuang akibat preprocessing outlier tidak terlalu banyak.

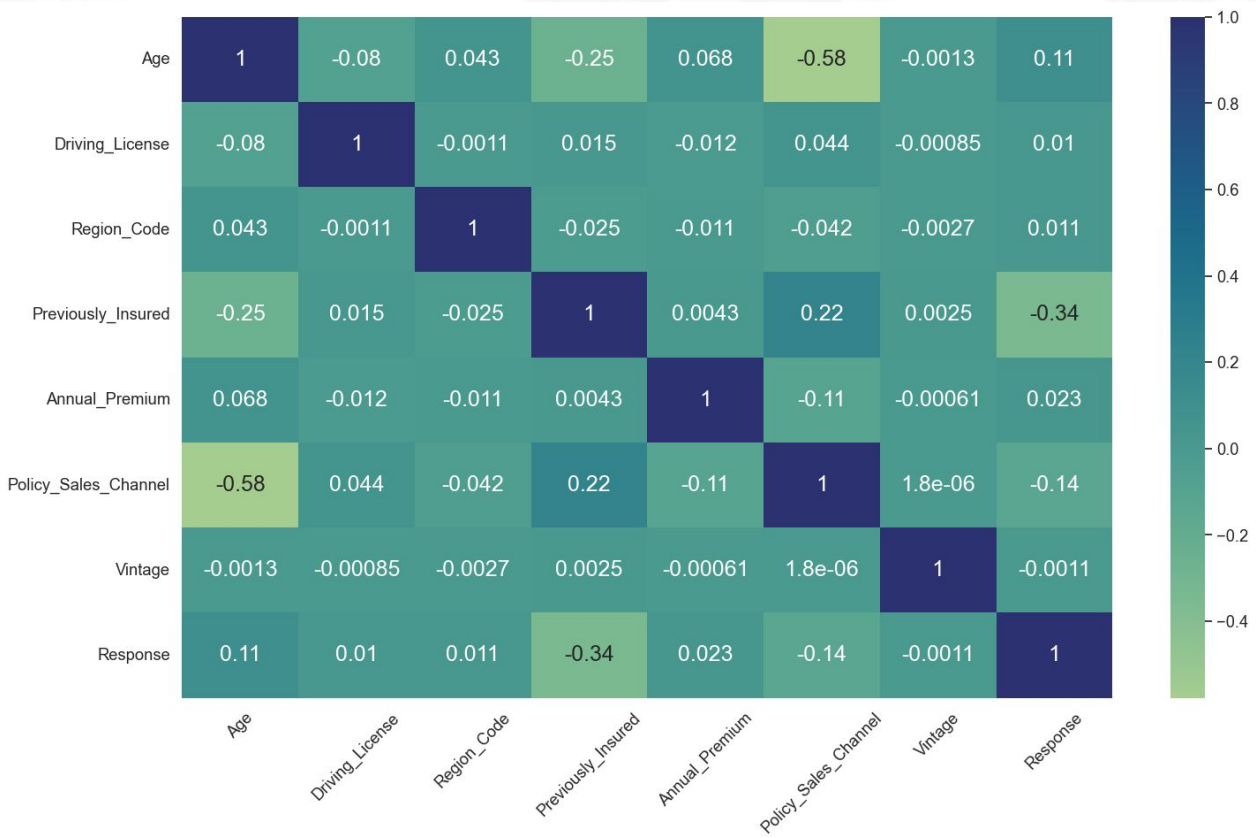
- **Menstandarisasi kolom yang jauh dari distribusi normal.**
- **Menangani data categorical dengan feature encoding.**
- **Menangani ketidakseimbangan kelas pada kolom Response sebagai target.**

  Mayoritas data customer memilih "0" untuk Response (dengan frekuensi 334399 dari 381109 data yang ada). Dimana artinya data customer yang memilih "1" hanya sekitar 12% dari total data. Hal ini tentunya menjadi suatu permasalahan dikarenakan dapat membuat model menjadi overfit. Harus dilakukan oversampling.

- **Mengambil top 10 atau top 15 dari kolom Policy_Sales_Channel dan Region_Code serta mengubah sisanya menjadi Others.**

Multivariate Analysis

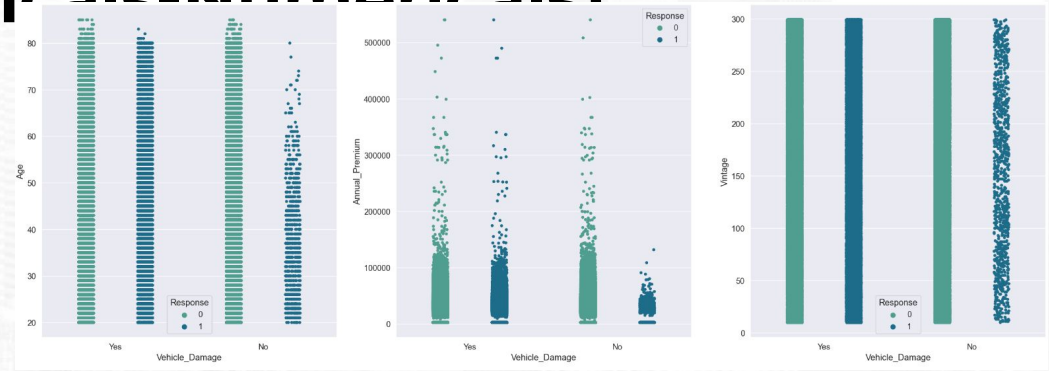
Heatmap Correlation



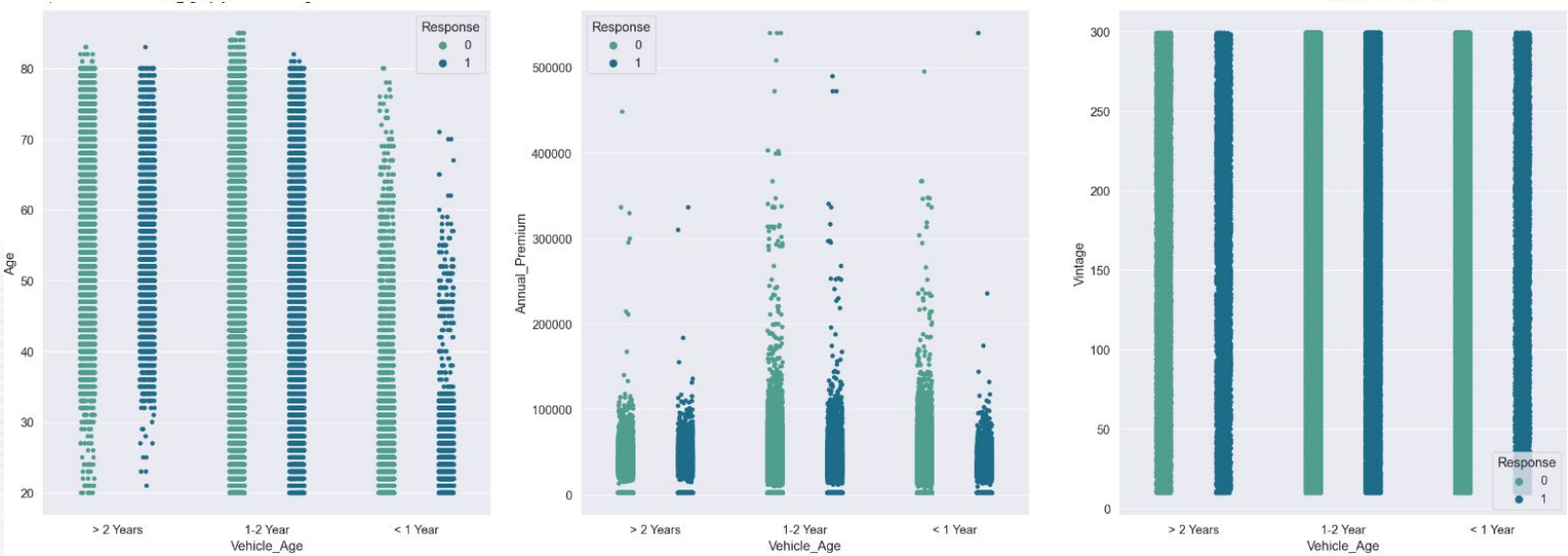
Jika dilihat dari **correlation heatmap** di atas, dapat dijelaskan sebagai berikut:

- **Tidak ada variabel** yang memiliki **korelasi kuat** atau r >= 0.7 atau r <= -0.7
- Hanya fitur **'Age'** dengan **'Policy_Sales_Channel'** yang memiliki **korelasi** yang **cukup kuat**
- Fitur **'Previously_Insured' memiliki korelasi negatif** yang **cukup kuat dengan** fitur '**Response'** (-0.34), yang **mengindikasikan** bahwa **orang** yang telah **memiliki asuransi kendaraan di tempat lain** memiliki **kemungkinan** yang **lebih rendah** untuk **membeli asuransi kendaraan** pada **perusahaan ini**.
- Fitur **'Age' memiliki korelasi positif** yang **cukup kuat dengan** fitur **'Response'** (0.11), yang **mengindikasikan** bahwa **semakin tua usia seseorang**, **semakin besar kemungkinan** mereka akan **membeli asuransi kendaraan**.

Stripplot Analysis (Categoricals-Numericals)

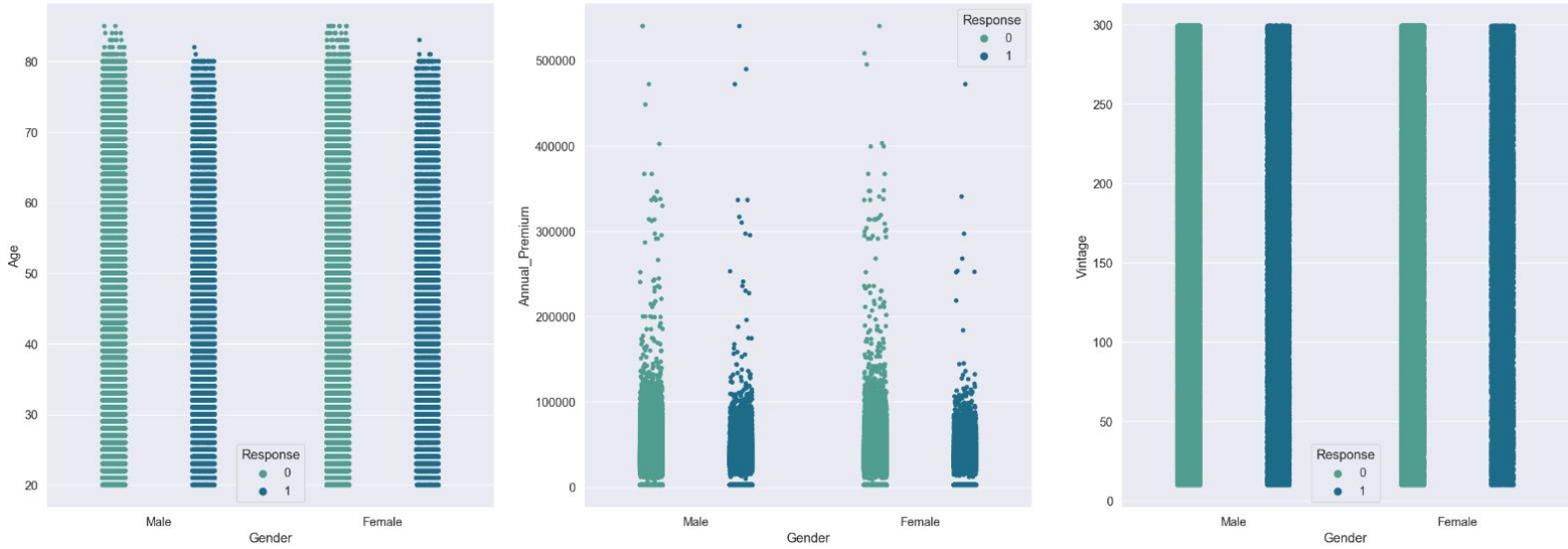


- Dari plot stripplot pertama, dapat dilihat bahwa pelanggan yang mempunyai kendaraan yang rusak (Yes) dan berusia lebih dari 50 tahun memiliki kecenderungan membeli asuransi kendaraan dibandingkan dengan pelanggan yang berusia di bawah 50 tahun. Hal ini dapat menjadi informasi yang berguna dalam menentukan target pelanggan yang akan dipasarkan.
- Pelanggan yang mengalami kerusakan pada kendaraannya memiliki kecenderungan membeli asuransi dibandingkan dengan pelanggan yang tidak mengalami kerusakan pada kendaraannya.
- Pelanggan yang membeli asuransi cenderung memiliki Annual_Premium yang lebih tinggi daripada pelanggan yang tidak membeli asuransi.
- Tidak ada perbedaan yang signifikan dalam Vintage (jumlah hari sejak pelanggan bergabung dengan perusahaan) antara pelanggan yang membeli asuransi dan pelanggan yang tidak membeli asuransi, baik pada kelompok Vehicle_Damage = Yes maupun Vehicle_Damage = No.
- Customer yang tidak memiliki kerusakan pada kendaraan cenderung tidak tertarik menggunakan produk asuransi kendaraan.



Dari plot-stripplot tersebut, dapat dilihat bahwa:

- Umumnya pelanggan yang merespons dengan 1 (membeli asuransi) memiliki usia lebih rendah daripada pelanggan yang merespons dengan 0 (tidak membeli asuransi) untuk setiap kategori umur kendaraan.
- Pelanggan dengan annual premium yang lebih tinggi cenderung merespons dengan 1 untuk setiap kategori umur kendaraan.
- Tidak ada korelasi yang jelas antara umur kendaraan dengan masa lalu pelanggan dan responsnya.



- Menunjukkan bahwa tidak terdapat perbedaan signifikan dalam distribusi usia antara gender dalam hal respon terhadap asuransi.
- Menunjukkan bahwa pria dan wanita memiliki distribusi premi tahunan yang berbeda secara signifikan dalam hal respon terhadap asuransi, dengan wanita cenderung membayar premi yang lebih rendah.
- Tidak terdapat perbedaan signifikan dalam distribusi jumlah hari vintage antara gender dalam hal respon terhadap asuransi.

Business Insight

Business Insight

Business Insight 1

Insight: Pelanggan dengan polis asuransi kendaraan yang lebih lama (Vehicle_Age > 2 tahun) cenderung lebih jarang merespon tawaran asuransi kendaraan baru dibandingkan pelanggan dengan polis kendaraan yang lebih baru.
Rekomendasi: Perusahaan asuransi perlu melakukan kampanye pemasaran khusus untuk pelanggan dengan polis kendaraan yang lebih lama untuk mendorong mereka untuk mempertimbangkan untuk membeli polis asuransi kendaraan baru. Kampanye tersebut bisa dilakukan melalui media sosial, email marketing, atau telepon langsung.

Business Insight 2

Insight: Pelanggan dengan kecelakaan kendaraan di masa lalu cenderung lebih memilih untuk membeli polis asuransi kendaraan yang lebih lengkap dengan manfaat tambahan seperti asuransi kesehatan dan asuransi kecelakaan diri.
Rekomendasi: Perusahaan asuransi dapat menargetkan pelanggan yang pernah mengalami kecelakaan kendaraan dengan paket asuransi kendaraan yang lebih lengkap. Perusahaan dapat menambahkan manfaat tambahan seperti asuransi kesehatan dan asuransi kecelakaan diri pada paket ini untuk menarik minat pelanggan. Selain itu, perusahaan juga dapat mempertimbangkan untuk menawarkan diskon atau promosi khusus untuk pelanggan yang membeli paket asuransi kendaraan lengkap.

Business Insight

Business Insight 3

Insight: Wilayah-wilayah dengan Region_Code 28.0, 41.0, 8.0, 46.0, dan 29.0 memiliki tingkat minat yang lebih tinggi dalam menggunakan asuransi kendaraan.

Rekomendasi: Untuk meningkatkan kepercayaan dan minat masyarakat dalam menggunakan asuransi kendaraan di wilayah-wilayah tersebut, perusahaan dapat melakukan kampanye pemasaran yang lebih intensif dan menawarkan program-program khusus yang menarik bagi nasabah di wilayah tersebut. Selain itu, perusahaan juga perlu melakukan analisis lebih lanjut terkait faktor-faktor apa yang membuat nasabah di wilayah-wilayah tersebut lebih tertarik pada asuransi kendaraan. Analisis tersebut dapat digunakan sebagai dasar dalam pengembangan produk dan strategi pemasaran yang lebih efektif.

Business Insight 4

Insight: Lebih banyak kendaraan yang diasuransikan berusia antara 1-2 tahun dibandingkan kendaraan dengan usia lebih dari 2 tahun. Hal ini menunjukkan bahwa kendaraan dengan usia tersebut masih banyak digunakan dan dianggap berharga oleh pemiliknya.

Rekomendasi: Perusahaan dapat menargetkan poromsi kepada pemilik kendaraan yang kendaraannya berusia 1-2 tahun.

Business Insight

Business Insight 5

Insight: Customer dengan usia antara 20-30 tahun lebih sering menolak penawaran asuransi daripada responden dengan usia di atas 30 tahun.

Rekomendasi: Hal ini dapat menjadi pertimbangan bagi perusahaan asuransi untuk menyesuaikan penawaran asuransi sesuai dengan profil usia pelanggan.

Business Insight 6

Insight: Lebih banyak kendaraan yang diasuransikan berusia antara 1-2 tahun dibandingkan kendaraan dengan usia lebih dari 2 tahun. Hal ini menunjukkan bahwa kendaraan dengan usia tersebut masih banyak digunakan dan dianggap berharga oleh pemiliknya.

Rekomendasi: Perusahaan dapat menargetkan promosi kepada pemilik kendaraan yang kendaraannya berusia 1-2 tahun.

Business Insight

Business Insight 7

Insight: Customer dengan jenis kelamin laki-laki cenderung membayar premi asuransi yang lebih tinggi daripada responden perempuan. Ini mungkin karena perempuan cenderung lebih hati-hati dalam mengemudi dan memilih kendaraan yang lebih aman.

Rekomendasi: Lebih fokus pada pelanggan laki-laki untuk meningkatkan pendapatan perusahaan.

Data Cleansing

A. Handle missing values

Dataset terdiri dari 11 kolom dengan jumlah baris tidak diketahui (tidak disebutkan). Setiap kolom memiliki tipe data yang berbeda-beda, dengan jumlah missing value (nilai kosong) pada setiap kolom bernilai 0.

```
# jumlah entry NULL di setiap kolom
df.isna().sum()
```

```
id                    0
Gender                0
Age                   0
Driving_License       0
Region_Code           0
Previously_Insured    0
Vehicle_Age           0
Vehicle_Damage        0
Annual_Premium        0
Policy_Sales_Channel  0
Vintage               0
Response              0
dtype: int64
```

B. Handle duplicated data

Dataset yang diberikan terdiri dari 11 kolom dengan jumlah baris tidak diketahui (tidak disebutkan). Setiap kolom memiliki tipe data yang berbeda-beda, dengan jumlah data duplikat pada setiap kolom bernilai 0.

```
# cek jumlah duplicated rows
# dari semua kolom
df.duplicated().sum()
```

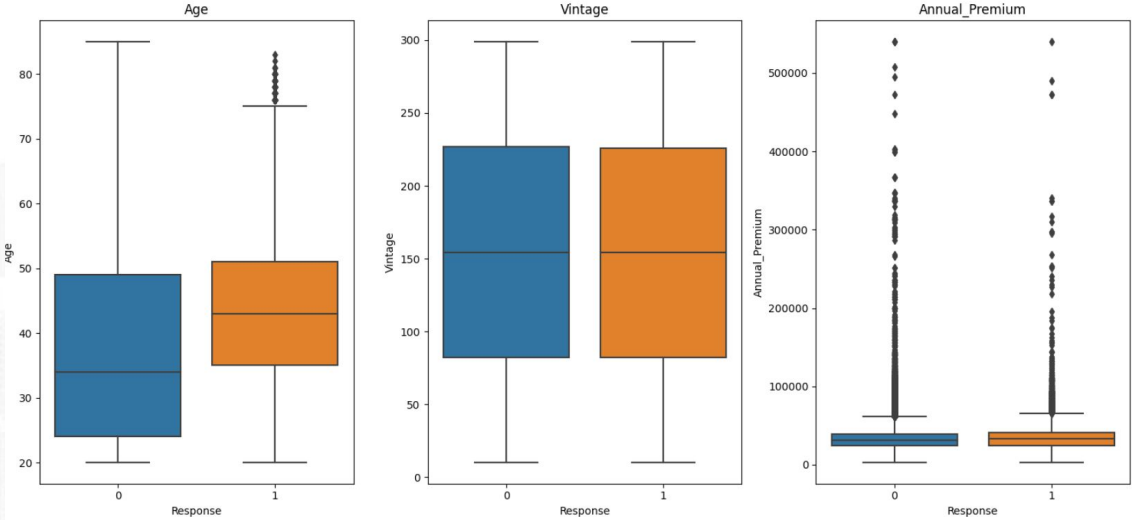
```
0
```

C. Handle outliers

Fitur yang termasuk nums kolom Annual_Premium perlu dilakukan handle outliers.

Kode di sampingakan menghapus data pada kolom Annual_Premium dan Age yang berada di luar rentang lower_limit dan upper_limit.

Rentang ini ditentukan berdasarkan IQR (Interquartile Range) dari kolom tersebut. Data yang dihapus adalah data yang dianggap sebagai outliers.




```
print(f'Jumlah baris sebelum memfilter outlier: {len(df)}')

Q1_ap = df['Annual_Premium'].quantile(0.25)
Q3_ap = df['Annual_Premium'].quantile(0.75)
IQR_ap = Q3_ap - Q1_ap
lower_limit_ap = Q1_ap - 1.5*IQR_ap
upper_limit_ap = Q3_ap + 1.5*IQR_ap
df = df[(df['Annual_Premium'] > lower_limit_ap) & (df['Annual_Premium'] < upper_limit_ap)]

Q1_age = df['Age'].quantile(0.25)
Q3_age = df['Age'].quantile(0.75)
IQR_age = Q3_age - Q1_age
lower_limit_age = Q1_age - 1.5*IQR_age
upper_limit_age = Q3_age + 1.5*IQR_age
df = df[(df['Age'] > lower_limit_age) & (df['Age'] < upper_limit_age)]

print(f'Jumlah data setelah handling outlier: {len(df)}')
```

```
Jumlah baris sebelum memfilter outlier: 381109
Jumlah data setelah handling outlier: 370779
```

D. Feature transformation

Normalization/Standardization:

Fitur Age, Annual_Premium dan Vintage memiliki rentang nilai yang jauh berbeda range-nya.

Kolom Age memiliki rentang nilai antara 15 hingga 84, lalu kolom Annual_Premium memiliki rentang nilai yang besar juga, dan fitur Vintage juga memiliki rentan nilai yang jauh berbeda.

Oleh karena itu, sebaiknya dilakukan normalisasi/standarisasi pada kedua kolom ini agar memiliki rentang nilai yang tidak terlalu jauh.

	id	Age	Annual_Premium	Vintage
count	370779.00	370779.00	370779.00	370779.00
mean	190535.40	0.29	0.45	0.00
std	110037.57	0.24	0.25	1.00
min	1.00	0.00	0.00	-1.73
25%	95217.50	0.08	0.36	-0.86
50%	190530.00	0.25	0.48	-0.00
75%	285824.50	0.45	0.61	0.87
max	381109.00	1.00	1.00	1.73

Dari hasil transformasi diatas yang dilakukan, terlihat bahwa:

1. Fitur Age memiliki rata-rata 0.29 dan standar deviasi 0.24 setelah dilakukan normaisasi
2. Fitur Annual_Premium memiliki rata-rata 0.45 dan standar deviasi 0.25 setelah dilakukan normalisasi
3. Fitur Vintage memiliki rata-rata 0.00 dan standar deviasi 1.00 setelah dilakukan standarisasi

E. Feature encoding

Strategi encoding

- `Vehicle_Damage` & `Vehicle_Age` \: Label Encoding, atau Mapping.
- `Gender` : One Hot Encoding

```
df.head()
```

	id	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response	Gender_Female	Gender_Male
0	1	0.375000	1	28.0	0	2	1	0.638250	26.0	0.748828	1	0	1
1	2	0.875000	1	3.0	0	1	0	0.521515	26.0	0.342470	0	0	1
2	3	0.421875	1	28.0	0	2	1	0.601802	26.0	-1.521996	1	0	1
3	4	0.015625	1	11.0	1	0	0	0.438544	152.0	0.581504	0	0	1
4	5	0.140625	1	41.0	1	0	0	0.419594	152.0	-1.378575	0	1	0

F. Handle class imbalance

```
response_counts = df['Response'].value_counts()
print(response_counts)
```

```
0    325624
1     45155
Name: Response, dtype: int64
```

Setelah dilakukan analisis lebih lanjut, dan berdasarkan hasil diskusi maka ditentukan untuk metode yang digunakan supaya data lebih balance menggunakan metode Undersampling.

```
from sklearn.preprocessing import LabelEncoder
from imblearn.over_sampling import SMOTE
import pandas as pd
from imblearn.under_sampling import RandomUnderSampler

le = LabelEncoder()
y = le.fit_transform(y)

rus = RandomUnderSampler()

# apply RandomUnderSampler to the dataset
X_rus, y_rus = rus.fit_resample(X, y)

# convert y_rus to pandas Series
y_rus = pd.Series(y_rus)

# check the class distribution
print(y_rus.value_counts())
```

```
0    45155
1    45155
dtype: int64
```

Output yang dihasilkan menunjukkan bahwa setelah dilakukan undersampling, jumlah data pada kedua kelas menjadi sama dan seimbang, yaitu masing-masing 45155.

Feature Engineering

A. Feature selection

Dari dataset yang diberikan, beberapa fitur yang mungkin perlu dibuang atau tidak terpakai yakni :

1. id: Fitur ini seharusnya tidak memberikan pengaruh pada prediksi apapun, karena hanya merupakan sebuah identifikasi untuk setiap data dan tidak berkaitan dengan masalah prediksi.
2. Region_Code: Fitur ini hanya menunjukkan kode wilayah dari pelanggan. Namun, tidak jelas apakah informasi ini dapat memberikan kontribusi signifikan terhadap prediksi.
3. Policy_Sales_Channel: Fitur ini merupakan channel yang digunakan oleh pelanggan untuk membeli asuransi. Namun, terdapat 155 jenis channel yang berbeda, yang membuatnya sulit untuk diolah dan memerlukan preprocessing khusus. Selain itu, informasi ini mungkin tidak memiliki korelasi yang kuat dengan masalah prediksi.

B. Feature extraction

Untuk menambah fitur, perlu analisa lebih lanjut sesuai keadaan atau kondisi. Karena, penambahan fitur juga dapat mempengaruhi hasil model. Untuk sementara pada kasus ini, disimpulkan tidak perlu ditambahkan fitur.

Perlu dipertimbangkan untuk ditambah, ekstraksi dari fitur :

1. Age : Kategorikan jadi Muda, Tua dst.
2. Vintage : Kategorikan jadi New User, Retain User.
3. Annual_Premium: Kategorikan jadi rendah, sedang, tinggi (atau bisa menjadi status: Bronze, Gold, Platinum).

C. Feature Tambahan

1. Monthly Salary

Pada kategori gaji tertentu kemungkinan kepedulian dan ketertarikan untuk mengasuransikan kendaraan akan cenderung lebih tinggi.

2. Number of vehicles (owned)

Jumlah kendaraan yang dimiliki kemungkinan berpengaruh terhadap keinginan mengasuransikan kendaraan.

3. Frequent use of private vehicles

Keseringan menggunakan kendaraan pribadi untuk aktifitas sehari-hari.

Kategori: 1 - setiap hari; 2 - beberapa kali seminggu; 3 - sekali seminggu; 4 - beberapa kali sebulan; 5 - sebulan sekali.

Dengan adanya pengelompokkan ini, akan terlihat kecenderungan kelompok mana yang akan mengasuransikan kendaraan mereka.

4. Profession

Profesi calon pelanggan. Kemungkinan orang yang berprofesi yang berhubungan dengan kendaraan akan mengasuransikan kendaraan mereka.

Modeling

Kami melakukan 8 algoritma jenis machine learning, untuk dicari mana yang paling sesuai dengan tujuan bisnis kami. Dengan 2 skenario data preprocessing.

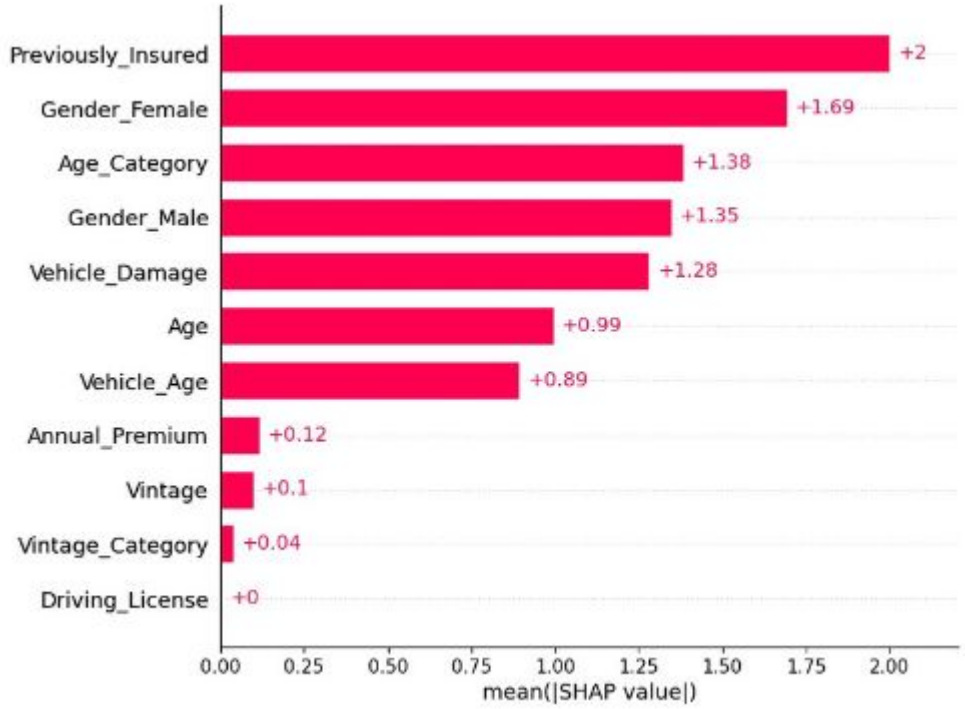
- Naive Bayes
- KNN
- Logistic Regression
- Decision Tree
- Adaboost
- Random Forest
- XGBoost
- CatBoost

Model terbaik yang kami pilih adalah model machine learning Cat Boost dengan nilai accuracy sebesar 0.85 dan nilai Recall 0.88

Pemilihan model ini menggunakan model evaluation recall, karena sesuai dengan goals dan tujuan bisnis kami yang ingin menargetkan pelanggan yang berpotensi untuk membeli produk asuransi kendaraan

Classifier	Naive Bayes	KNN	Decision Tree	RandomForest	CatBoost
Akurasi	66%	72%	87%	88%	85%
Presisi	40%	58%	80%	87%	73%
Recall	**2%**	**61%**	**80%**	**77%**	**88%**
F1-Score	4%	59%	80%	81%	80%
AUC	50%	69%	85%	85%	86%

Feature Importance



Fitur Previously_Insured adalah fitur yang paling mempengaruhi pada dataset yang dimiliki.

Sedangkan Driving_License memiliki pengaruh paling kecil atau tidak berpengaruh sama sekali.

Summary and Recommendation

Executive Summary

Sesuai dengan tujuan dari modeling yang dilakukan diatas, dapat menggunakan strategi sebagai berikut:

mengurangi marketing cost dengan melakukan penawaran kepada calon pengguna potensial berdasarkan prediksi dari model untuk membeli polis asuransi kendaraan.

Recommendation

1. Paket bundling asuransi kesehatan dan kendaraan - sesuai dengan dataset, adanya potensial untuk nasabah memiliki 2 jenis asuransi sekaligus, sehingga bundling paket asuransi kesehatan dengan asuransi kendaraan ini dapat meningkatkan pelanggan perusahaan.
2. Memberikan penawaran (misalkan cashback ataupun angsuran gratis 1-3 bulan pertama) kepada pelanggan yang kendaraan pernah mengalami kerusakan rusak ataupun yang berumur tua maupun yang pernah mengasuransikan kendaraan sebelumnya, hal ini didasari data yang ada, bahwa sebagian besar orang akan mengasuransikan kendaraan mereka yang pernah mengalami kerusakan ataupun berumur lebih tua ataupun mereka yang pernah mengasuransikan kendaraan mereka.
3. Memaksimalkan uang yang mengendap agar dapat diputarkan untuk mendapatkan revenue lebih (estimasi ada penambahan 5%), ini sesuai dengan sistem perbankan, yang dimana dapat mengelola dana yang diterima dari nasabah untuk mendapatkan keuntungan lebih.
4. Meningkatkan CASA sebesar 12% dari pelanggan yang telah berlangganan asuransi.

Rakamin Academy

Terima kasih!

Oleh : A Six