Rakamin Academy

Stage 2
Health Insurance

Kaizen Team

(dipresentasikan setiap sesi mentoring)

Stage 2: Preprocessing

Pada stage 2 - Preprocessing ini yang dilakukan adalah:

1. Data Cleansing
    a. Handling missing values & duplicate data
    b. Handle outliers
    c. Feature Transformation
    d. Feature Encoding
    e. Class Imbalance
2. Feature Engineering
    a. Feature Selection
    b. Feature Extraction
    c. Feature Tambahan
3. Git

Rakamin Academy

Data Cleansing

1. Data Cleansing - Missing values & duplicate

- **Missing values & duplicate data**
  - Pada dataset Health Insurance tidak terdapat missing values & duplicate data.

missing value

```
1 X_train.isna().sum()
```

```
Gender                0
Age                   0
Region_Code           0
Previously_Insured    0
Vehicle_Age           0
Vehicle_Damage        0
Annual_Premium        0
Policy_Sales_Channel  0
Vintage               0
Age_cat               0
Annual_Membership     0
dtype: int64
```

duplicate data

```
1 column1 = ["Gender","Age_cat", "Region_Code", "Previously_Insured", "Vehicle_Age", "Vehicle_Damage", "Annual_Membership","Policy_Sales_Channel"]
2 column2 = num + nom + ordi
3 df[df.duplicated(subset= column1, keep = False)== True].sort_values(column1)
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response	Vintage_mon	Age_cat	Annual_Membership
229133	229134	Female	33	1	0.0	0	1-2 Year	No	2630.0	26.0	128	0	4.0	1	Silver
336182	336183	Female	33	1	0.0	0	1-2 Year	No	2630.0	26.0	71	0	2.0	1	Silver
2217	2218	Female	33	1	0.0	0	1-2 Year	No	2630.0	60.0	191	0	6.0	1	Silver
46962	46963	Female	31	1	0.0	0	1-2 Year	No	2630.0	60.0	148	0	5.0	1	Silver
88679	88680	Female	32	1	0.0	0	1-2 Year	No	2630.0	60.0	50	0	2.0	1	Silver
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
297795	297796	Male	80	1	50.0	1	1-2 Year	No	37241.0	8.0	265	0	9.0	5	Gold
60887	60888	Male	80	1	50.0	1	1-2 Year	No	55368.0	8.0	36	0	1.0	5	Platinum
303703	303704	Male	80	1	50.0	1	1-2 Year	No	59432.0	8.0	128	0	4.0	5	Platinum
305355	305356	Male	80	1	50.0	1	1-2 Year	Yes	37792.0	26.0	205	0	7.0	5	Gold
306859	306860	Male	80	1	50.0	1	1-2 Year	Yes	34280.0	26.0	41	0	1.0	5	Gold

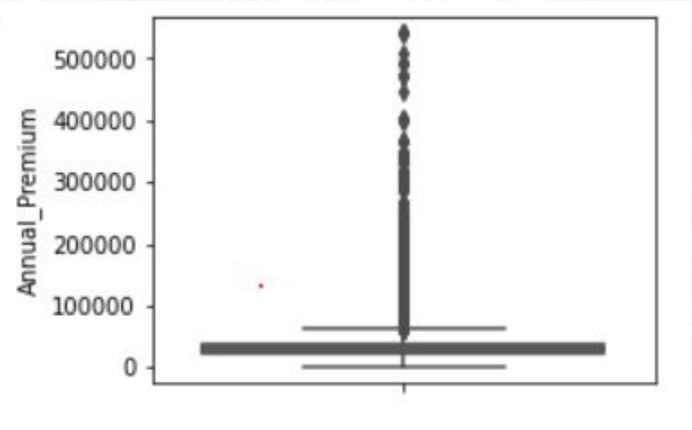
364260 rows × 15 columns

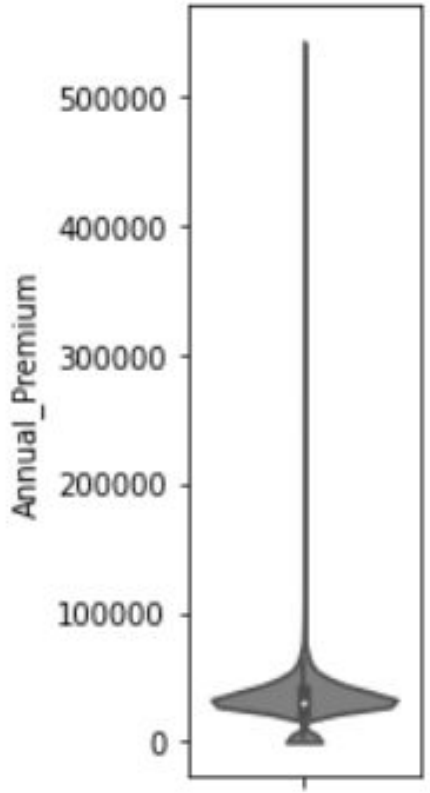
```
1 df.duplicated().sum()
```

```
0
```

1. Data Cleansing - Handling Outliers

- **Handling Outliers**
  Pada dataset Health Insurance terdapat satu kolom yang memiliki outlier cukup banyak dan menjadi tantangan bagi kami untuk menemukan cara yang tepat menangani outlier tersebut, yaitu di kolom Annual_Premium.





1. Data Cleansing - Handling Outliers

Temuan lain yang berkaitan: dataset Health Insurance dari sumbernya (Kaggle) telah dipecah menjadi dataset train & test, namun di dataset test tidak terdapat kolom target yaitu Response, sehingga dataset train dipecah lagi menjadi train & test sehingga dataset baru menjadi 266.766 baris train dan 114.333 baris test (menggunakan 0.3 test dan random state 123)

a. **Alternatif 1: Tidak membuang outlier**
b. **Alternatif 2: Substitusi nilai dengan high limit**

1. Data Cleansing - Handling Outliers

a. Alternatif 1: Tidak membuang outlier

Jumlah outlier ditemukan dari data train adalah 3069 baris atau 2.72% sehingga tidak ada yang didrop/substitusi nilainya karena tidak ingin kehilangan data di rentang annual premium yang lebih tinggi.

```
#preview data-data untuk Annual Premium pada X_test
X_test_highlimit=X_test[X_test['Annual_Premium']>=high_limit]
X_test_highlimit.sort_values(by='Annual_Premium',ascending=False)
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage
268332	268333	Male	46	1	28.0	0	1-2 Year	Yes	540165.0	124.0	59
190154	190155	Male	47	1	28.0	0	1-2 Year	Yes	540165.0	42.0	24
11319	11320	Female	50	1	46.0	1	1-2 Year	No	508073.0	26.0	192
136304	136305	Male	50	1	28.0	0	1-2 Year	Yes	472042.0	124.0	14
275442	275443	Male	22	1	28.0	0	1-2 Year	Yes	472042.0	163.0	114
...	...	...	...	...	...	...	...	...	...	...	...
43505	43506	Female	70	1	28.0	1	1-2 Year	No	61893.0	26.0	195
249017	249018	Female	25	1	28.0	1	< 1 Year	No	61893.0	152.0	17
81886	81887	Male	23	1	28.0	0	< 1 Year	Yes	61893.0	152.0	128
147976	147977	Male	25	1	46.0	1	< 1 Year	No	61892.0	152.0	60
254849	254850	Female	52	1	28.0	0	1-2 Year	Yes	61884.0	26.0	151

3069 rows × 11 columns

1. Data Cleansing - Handling Outliers

b. Alternatif 2: Substitusi nilai dengan high limit

Mengganti 2.72% data ‘annual_premium’ dengan data maksimal (high limit) pada feature tersebut

```
## Change Outlier on AP with Q3
def ChangeByQ (x, col, low_limit, high_limit, Q3, Q1) :
    if x[col] < low_limit :
        output = Q1
    elif x[col] > high_limit :
        output = Q3
    else :
        output = x[col]
    return output

def Handling_Outlier(df_HO, col) :
    Q1 = df_HO[col].quantile(0.25)
    Q3 = df_HO[col].quantile(0.75)
    IQR = Q3 - Q1
    low_limit = Q1 - (IQR * 1.5)
    high_limit = Q3 + (IQR * 1.5)
    colname = col + "1"
    df_HO_copy = df_HO.copy()
    df_HO_copy[colname] = df_HO_copy.apply(lambda x : ChangeByQ(x,col, low_limit, high_limit, Q3, Q1), axis = 1)
    return df_HO_copy

X_train_HO = Handling_Outlier(X_train,"Annual_Premium")
X_train_HO = X_train_HO.drop(["Annual_Premium"], axis = 1).rename(columns={"Annual_Premium1":"Annual_Premium"})
```

Dari kedua alternatif handling outlier tersebut, kami memutuskan akan menggunakan alternatif pertama (tidak membuang outlier) untuk kedepannya, namun kami tetap akan menyimpan kode alternatif kedua untuk keperluan pengetesan model sehingga bisa dibandingkan hasil mana yang lebih akurat.

1. Data Cleansing - Feature Transformations

Feature transformations & encoding dilakukan bersamaan dalam pipeline.
Ada 4 pipeline:

- **Numeric pipe** untuk feature Age, Annual Premium
- **Categoric pipe** untuk feature Gender, Vehicle Damage, Previously Insured
- **Binary pipe** untuk Policy Sales Channel & Region Code
- **Ordinal pipe**, ada 3 pipeline terpisah untuk masing-masing feature yaitu Vehicle Age, Age cat, Annual Membership

Feature transformations dilakukan pada feature numerical yaitu Age & Annual Premium.
Preprocessing Stepsnya adalah menggunakan standardisasi karena lebih robust terhadap outlier, lalu menggunakan Power Transformer yang bekerja dengan nilai negatif maupun positif.

1. Data Cleansing - Feature Encoding

Feature transformations & encoding dilakukan bersamaan dalam pipeline.

Feature encoding dilakukan pada feature kategorik yaitu:
-Gender, Vehicle Damage & Previously Insured diencoding menggunakan one hot encoder agar menghindari multikolinearitas.

-Policy Sales Channel dan Region Code diencoding dengan binary encoder. Feature awalnya yang terdiri dari banyak kategori diencoding menjadi jumlah yang lebih sedikit.

-Vehicle Age, Age cat dan Annual Membership menggunakan ordinal encoder karena tipe datanya merupakan ordinal.

1. Data Cleansing - Feature Encoding

Hasil setelah dilakukan pipeline menghasilkan 21 kolom sebagai berikut:

```
[82] X_train_result=pd.DataFrame(preprocessor.fit_transform(X_train))
     X_train_result
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	-0.946879	-0.261516	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.366362	-0.028939	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0
2	0.236563	-0.430851	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0
3	-0.750448	-0.056116	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
4	-1.364742	1.463324	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
266771	0.708598	0.240644	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	2.0
266772	0.427890	0.635720	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	2.0
266773	-1.048260	-1.739779	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0
266774	-1.151714	0.576127	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	2.0
266775	0.545133	-1.739779	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0

266776 rows × 22 columns

Keterangan:

Kolom 0: Age

Kolom 1: Annual_Premium

Kolom 2: Gender

Kolom 3: Vehicle_Damage

Kolom 4: Previously_Insured

Kolom 5-12: Policy_Sales_Channel

Kolom 13-18: Region_Code

Kolom 19: Vehicle_Age

Kolom 20: Age_cat

Kolom 21: Anual_Membership

1. Data Cleansing - Class Imbalance

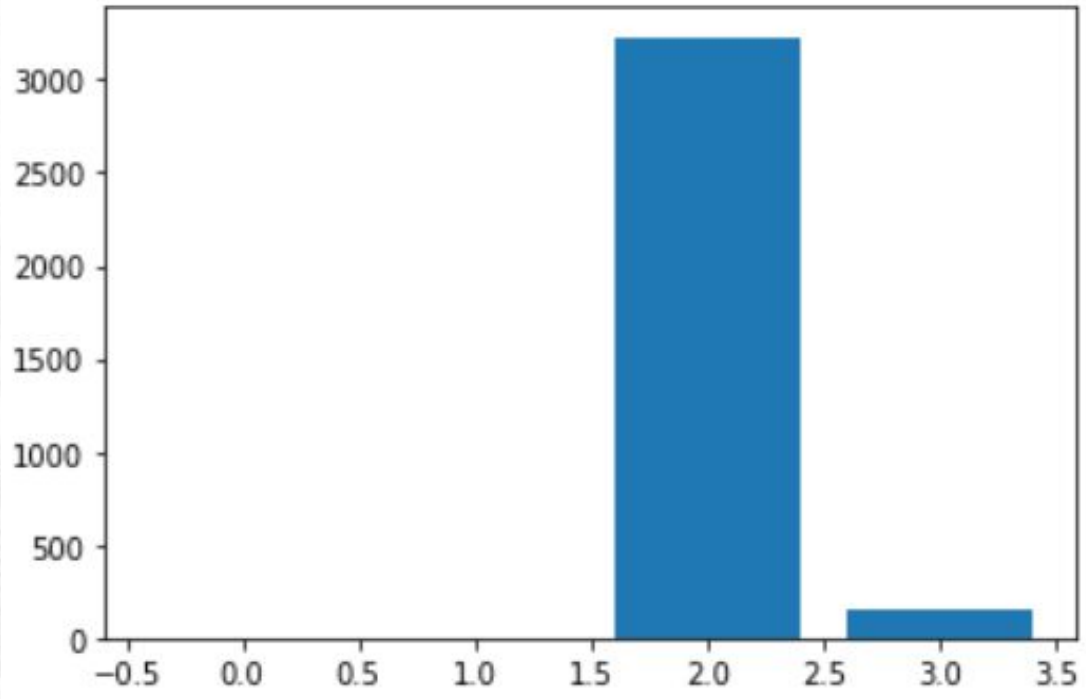
Handling imbalance data akan dilakukan di stage model machine learning menggunakan class weight feature (**sesuai dengan arahan mentor*)

Rakamin Academy

Feature Engineering

2. Feature Engineering - Feature Selection

1. Anova



Anova digunakan untuk feature numerik dengan target kategorik. Feature numerik pada dataset Health insurance yaitu Annual_Premium, Age, Vintage dan Vintage_mon.

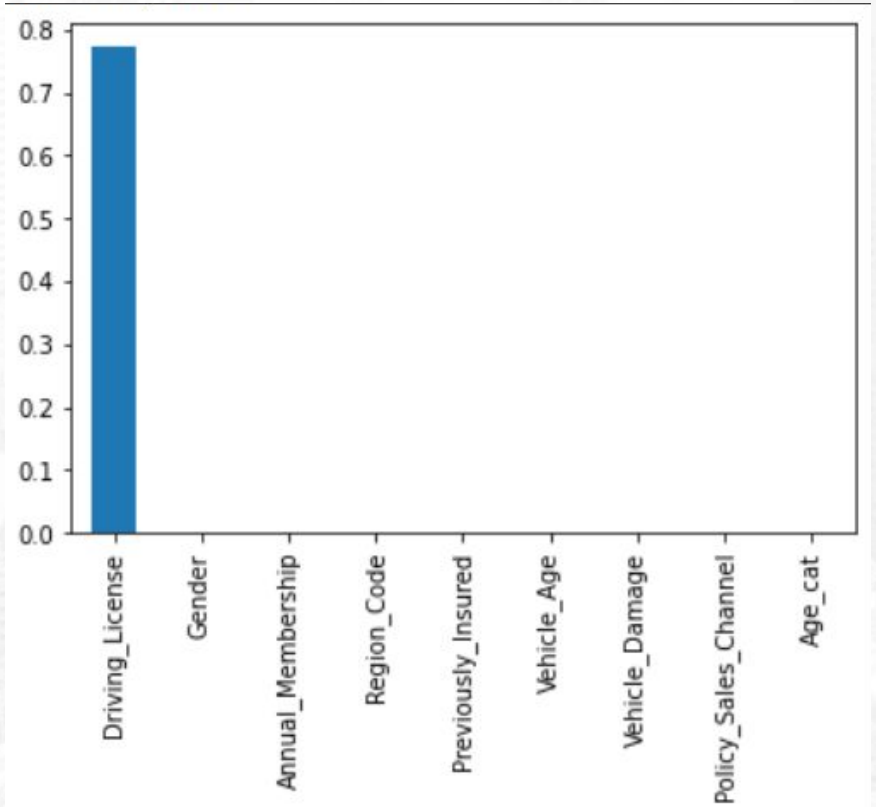
Dari hasil grafik di samping, feature yang akan digunakan adalah:

- annual_premium
- age

Feature yang tidak digunakan adalah Vintage dan Vintage_mon.

2. Feature Engineering - Feature Selection

2. Chi Square



Chi Square digunakan untuk feature kategorik dengan target kategorik.

Dari hasil grafik di samping, feature yang akan digunakan adalah:

- Gender
- Annual_Membership
- Region_Code
- Vehicle_Damage
- Previously_Insured
- Vehicle_Age
- Policy_Sales_Channel
- Age_cat

Feature yang tidak digunakan adalah Driving_License.

2. Feature Engineering - Feature Extraction

Feature extraction yang dilakukan pada dataset:
1. Mengubah kolom Vintage dari yang tadinya memiliki value hari menjadi bulan.
2. Melakukan binning pada kolom Age dan Annual_Premium.

	Vintage	Vintage_mon
0	217	7.0
1	183	6.0
2	27	1.0
3	203	7.0
4	39	1.0
...	...	...
381104	88	3.0
381105	131	4.0
381106	161	5.0
381107	74	2.0
381108	237	8.0

381109 rows × 2 columns

```
1    181876
2    104426
3     63947
4     29822
5      1038
Name: Age_cat, dtype: int64
```

```
Gold        137707
Bronze      133461
Silver       66554
Platinum     43387
Name: Annual_Membership, dtype: int64
```

Feature Extraction

2. Feature Engineering - Ide Feature Tambahan

1. **Jumlah kendaraan yang dimiliki**

Semakin banyak kendaraan semakin banyak biaya operasional yang perlu dikeluarkan, sehingga kemungkinan akan mempengaruhi keputusan pelanggan untuk membeli asuransi kendaraan.

2. **Kepemilikan utang/loan/credit score**

Semakin banyak utang pelanggan kemungkinan akan mempengaruhi keputusannya untuk membeli asuransi kendaraan.

3. **Jumlah anak**

Semakin banyak anak semakin banyak biaya operasional yang perlu dikeluarkan, sehingga kemungkinan akan mempengaruhi keputusan pelanggan untuk membeli asuransi kendaraan.

4. **Tipe mobil**

Model mobil yang bergengsi /prestigious memiliki kemungkinan akan mempengaruhi keputusan pelanggan untuk membeli asuransi kendaraan.

5. **Pekerjaan**

Pelanggan yang memiliki tipe pekerjaan yang *mobile* dengan kendaraan memiliki kemungkinan untuk membeli asuransi kendaraan.

Rakamin Academy

Git

3. Git

Berikut adalah link repository git kelompok Kaizen:

https://github.com/nengnisye/kaizen_rakamin.git