PURWADHIKA DIGITAL SCHOOL

Travel Insurance

By Mochammad Itmamul Wafa

Business Problem Understanding

Context

Business Problem Understanding

Problem Statement

Business Problem Understanding

Goals

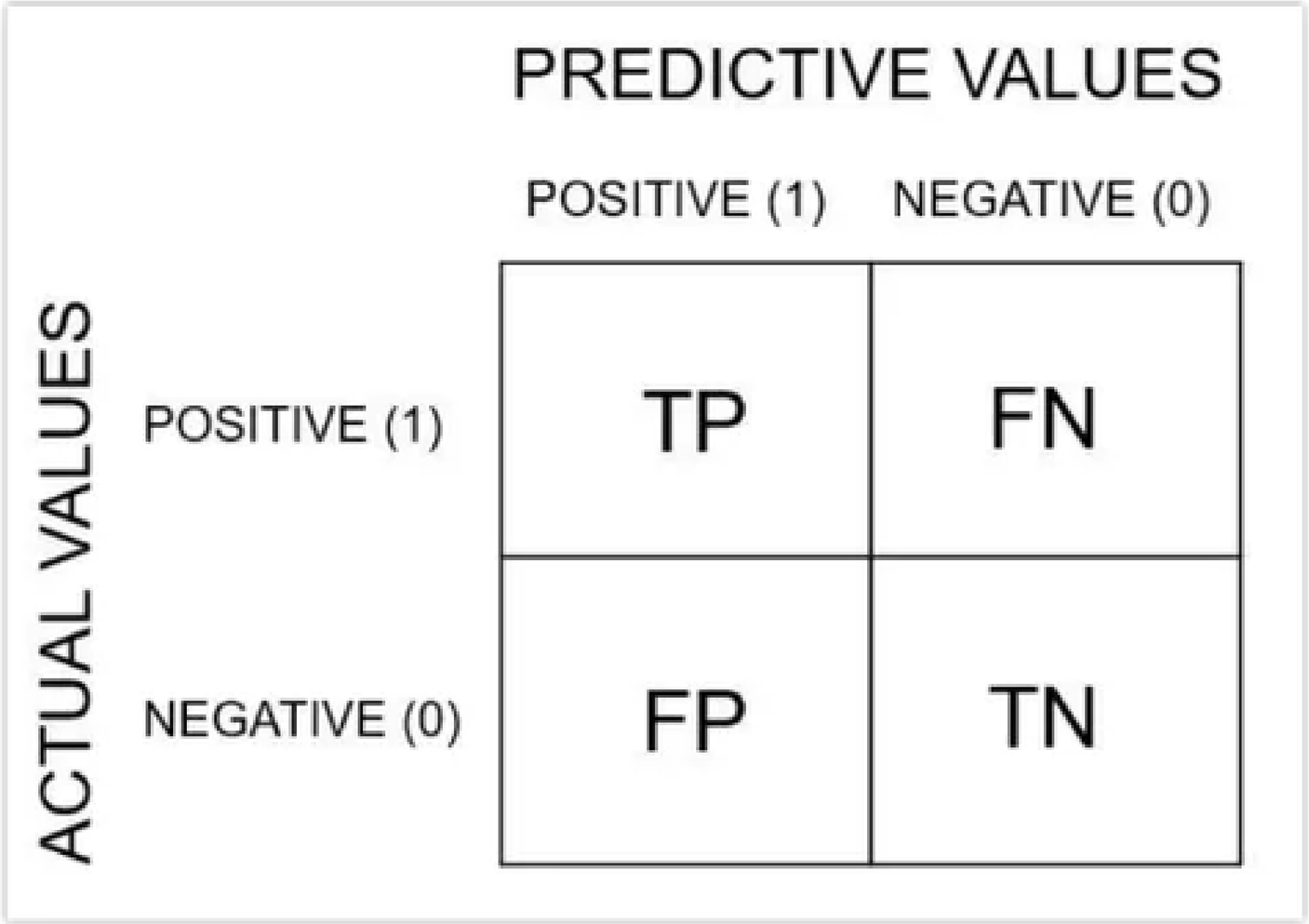
- Pembatalan polis
- Investasi premi
- Tidak adanya klaim selama masa kontrak

Business Problem Understanding

Analytic Approach

- Regresi Logistik
- Decision Tree
- KNN
- Random Forest
- XGBoost
- LightGBM

Business Problem Understanding



Metric Evaluation

Data Understanding

Attribute	Data Type, Length	Description
Agency	Text	Nama agensi
Agency Type	Text	Jenis agen asuransi perjalanan
Distribution Channel	Text	Saluran agen asuransi perjalanan
Product Name	Text	Nama produk asuransi perjalanan
Gender	Text	Jenis kelamin tertanggung
Duration	Bilangan bulat	Durasi perjalanan
Destination	Text	Tujuan perjalanan
Net Sales	Text	Jumlah penjualan polis asuransi perjalanan
Commission (in value)	Text	Komisi diterima untuk agen asuransi perjalanan
Age	Bilangan bulat	Usia tertanggung
Claim	Text	Status klaim

Data Insight Report

```
Jumlah baris dan kolom di dataset claim adalah (44328, 11)
Range indexnya adalah range(0, 44328)
```

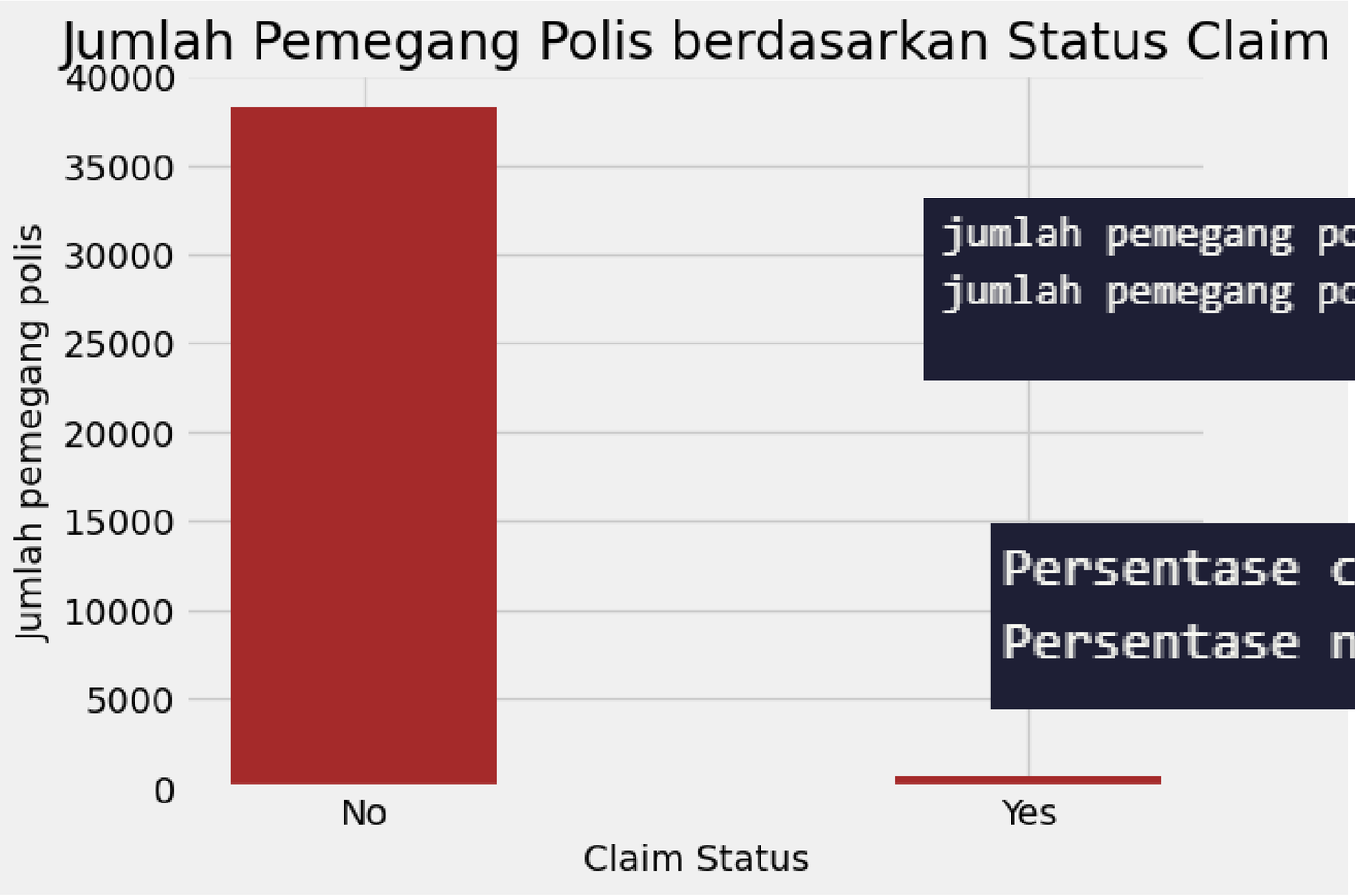
	Features	DataType	Null	NullPercentage	Unique	Count
0	Agency	object	0	0.0 %	16	44328
1	Agency Type	object	0	0.0 %	2	44328
2	Distribution Channel	object	0	0.0 %	2	44328
3	Product Name	object	0	0.0 %	26	44328
4	Gender	object	31647	71.39 %	2	12681
5	Duration	int64	0	0.0 %	437	44328
6	Destination	object	0	0.0 %	138	44328
7	Net Sales	float64	0	0.0 %	1006	44328
8	Commision (in value)	float64	0	0.0 %	915	44328
9	Age	int64	0	0.0 %	89	44328
10	Claim	object	0	0.0 %	2	44328

Data Insight Report

```
Jumlah baris dan kolom di dataset claim adalah (44328, 11)
Range indexnya adalah range(0, 44328)
```

	Features	DataType	Null	NullPercentage	Unique	Count
0	Agency	object	0	0.0 %	16	44328
1	Agency Type	object	0	0.0 %	2	44328
2	Distribution Channel	object	0	0.0 %	2	44328
3	Product Name	object	0	0.0 %	26	44328
4	Gender	object	31647	71.39 %	2	12681
5	Duration	int64	0	0.0 %	437	44328
6	Destination	object	0	0.0 %	138	44328
7	Net Sales	float64	0	0.0 %	1006	44328
8	Commision (in value)	float64	0	0.0 %	915	44328
9	Age	int64	0	0.0 %	89	44328
10	Claim	object	0	0.0 %	2	44328

Data Imbalace Claim




```
jumlah pemegang polis yang melakukan claim: 677
jumlah pemegang polis yang tidak melakukan claim: 43651
```

```
Persentase claim: 1.527251398664501 %
Persentase non claim: 98.4727486013355 %
```

Data Preparation

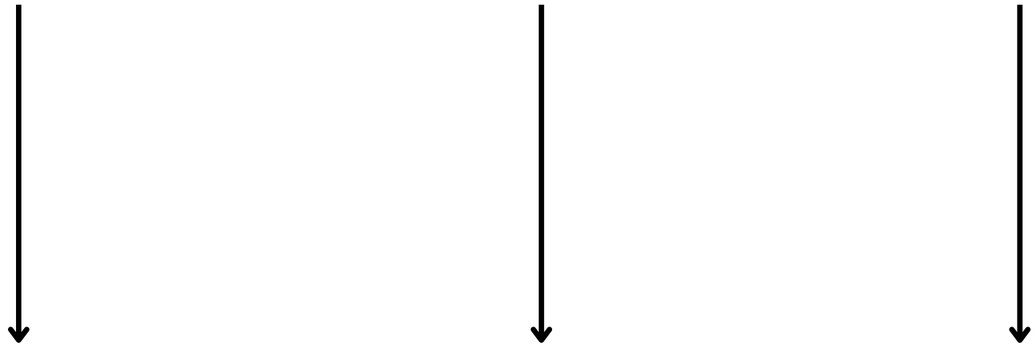
1. Handling Missing Value

	variable	null count	null percentage
0	Agency	0	0.000000
1	Agency Type	0	0.000000
2	Distribution Channel	0	0.000000
3	Product Name	0	0.000000
4	Gender	31647	71.392799
5	Duration	0	0.000000
6	Destination	0	0.000000
7	Net Sales	0	0.000000
8	Commision (in value)	0	0.000000
9	Age	0	0.000000
10	Claim	0	0.000000

Data Preparation

2. Handling Duplicates

```
number of rows: 44328
number of columns: 10
Jumlah data yang memiliki duplikat: 5004
```



```
number of rows: 39324
number of columns: 10
```

Data Preparation

3. Handling Invalid Data

	Agency	Agency Type	Distribution Channel	Product Name	Duration	Destination	Net Sales	Commision (in value)	Age	Claim
87	SSI	Airlines	Online	Ticket Protector	4736	SINGAPORE	0.32	0.09	48	No
4678	SSI	Airlines	Online	Ticket Protector	4857	SINGAPORE	0.32	0.09	48	No
10172	SSI	Airlines	Online	Ticket Protector	4815	SINGAPORE	0.32	0.09	48	No
12140	SSI	Airlines	Online	Ticket Protector	4652	SINGAPORE	0.32	0.09	48	No
19497	SSI	Airlines	Online	Ticket Protector	4881	SINGAPORE	0.13	0.04	48	No
20038	SSI	Airlines	Online	Ticket Protector	4844	SINGAPORE	0.32	0.09	48	No
21551	C2B	Airlines	Online	Annual Silver Plan	740	SINGAPORE	0.00	54.19	36	No
22215	SSI	Airlines	Online	Ticket Protector	4580	SINGAPORE	0.32	0.09	48	No
25606	SSI	Airlines	Online	Ticket Protector	4685	SINGAPORE	0.32	0.09	48	No
28032	SSI	Airlines	Online	Ticket Protector	4829	SINGAPORE	0.32	0.09	48	No
37207	SSI	Airlines	Online	Ticket Protector	4609	SINGAPORE	0.32	0.09	48	No
38536	SSI	Airlines	Online	Ticket Protector	4784	SINGAPORE	0.32	0.09	48	No
42584	SSI	Airlines	Online	Ticket Protector	4831	SINGAPORE	0.32	0.09	48	No
42717	SSI	Airlines	Online	Ticket Protector	4847	SINGAPORE	0.32	0.09	48	No

	Agency	Agency Type	Distribution Channel	Product Name	Duration	Destination	Net Sales	Commision (in value)	Age	Claim
26228	JZI	Airlines	Online	Basic Plan	-1	MALAYSIA	18.0	6.3	118	No
26494	JZI	Airlines	Online	Basic Plan	-1	BRUNEI DARUSSALAM	18.0	6.3	118	No

Data Preparation

4. Handling Outliers

	Agency	Agency Type	Distribution Channel	Product Name	Duration	Destination	Net Sales	Commision (in value)	Age	Claim
99	ART	Airlines	Online	Value Plan	7	MALAYSIA	29.0	10.15	118	No
25513	JWT	Airlines	Online	Value Plan	31	INDIA	120.0	48.00	118	No
29674	JWT	Airlines	Online	Value Plan	41	INDIA	31.0	12.40	118	No
29660	EPX	Travel Agency	Offline	Cancellation Plan	72	MALAYSIA	0.0	0.00	118	No
29455	JWT	Airlines	Online	Value Plan	19	INDIA	72.0	28.80	118	No
...	...	...	...	...	...	...	...	...	...	...
12881	ART	Airlines	Online	Value Plan	27	UNITED KINGDOM	96.0	33.60	118	No
12588	JWT	Airlines	Online	Value Plan	102	INDIA	-78.0	31.20	118	No
12519	JWT	Airlines	Online	Value Plan	24	INDIA	50.0	20.00	118	No
12450	CCR	Travel Agency	Offline	Comprehensive Plan	6	THAILAND	29.0	9.57	118	No
44140	JWT	Airlines	Online	Value Plan	84	INDIA	93.0	37.20	118	No

Data Preparation

5. Manage Net Sales & Commission

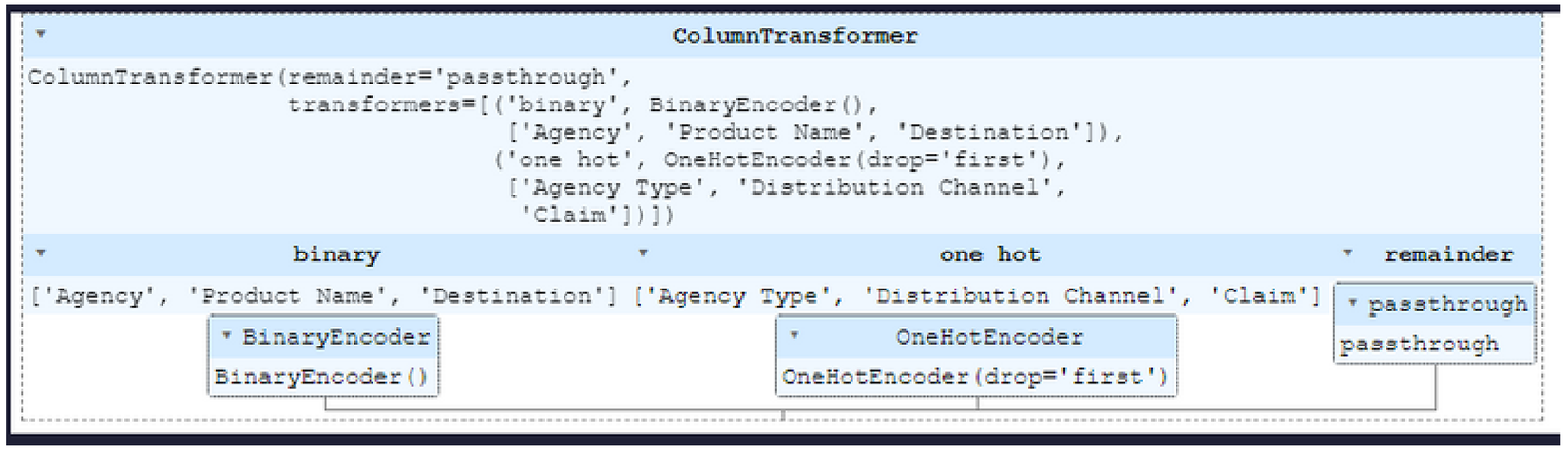
Data Preparation

6. Recheck duplicate, null value, data type dan imbalance data

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38887 entries, 0 to 44327
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Agency                38887 non-null  object
 1   Agency Type           38887 non-null  object
 2   Distribution Channel  38887 non-null  object
 3   Product Name          38887 non-null  object
 4   Duration              38887 non-null  int64
 5   Destination           38887 non-null  object
 6   Net Sales             38887 non-null  float64
 7   Commision (in value)  38887 non-null  float64
 8   Age                   38887 non-null  int64
 9   Claim                 38887 non-null  object
dtypes: float64(2), int64(2), object(6)
memory usage: 3.3+ MB
```

Data Preparation

7. Feature Engineering - Encoding



Data Preparation

7. Feature Engineering - Scaling

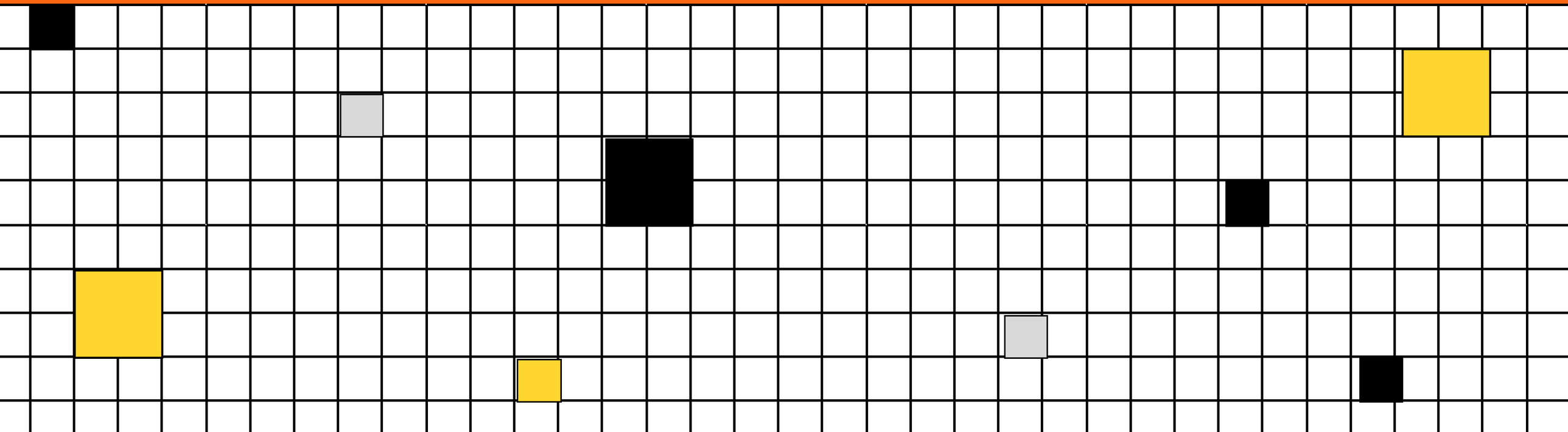
Scaling

```
# Define Scaler
scaler = RobustScaler()
```

✓ 0.1s

Modeling & Evaluation

Logistic Regression,KNN, Decision Tree, XGBoost dan LightGBM



Cross Validation

ROC-AUC without feature selection

	model	mean_rocauc_with_sampling	mean_rocauc_without_sampling	std_rocauc_with_sampling	std_rocauc_without_sampling
0	Logistic Regression	0.814232	0.817617	0.026889	0.024152
1	KNN	0.661572	0.592402	0.015644	0.004658
3	LightGBM	0.803535	0.810924	0.020263	0.014529
2	XGBoost	0.789781	0.781210	0.022319	0.013004
1	Random Forest	0.726795	0.687611	0.021515	0.012006
0	Desicion Tree	0.532399	0.522206	0.008438	0.013788

ROC-AUC with feature selection

	model	mean_rocauc_with_sampling	mean_rocauc_without_sampling	std_rocauc_with_sampling	std_rocauc_without_sampling
0	Logistic Regression	0.820615	0.821497	0.022901	0.021905
1	KNN	0.661699	0.609798	0.026830	0.012776
3	LightGBM	0.809756	0.806682	0.023716	0.022060
2	XGBoost	0.795564	0.795167	0.027623	0.020855
1	Random Forest	0.685810	0.667192	0.029237	0.013125
0	Desicion Tree	0.579932	0.570697	0.015448	0.015157

Cross Validation

Balanced accuracy without feature selection

	model	mean_acc_with_sampling	mean_acc_without_sampling	std_acc_with_sampling	std_acc_without_sampling
0	Logistic Regression	0.750165	0.499984	0.020601	0.000033
1	KNN	0.612547	0.500535	0.007438	0.001858
3	LightGBM	0.574642	0.500787	0.018336	0.001781
2	XGBoost	0.533460	0.500714	0.022176	0.001964
1	Random Forest	0.533022	0.504172	0.009496	0.007648
0	Desicion Tree	0.530406	0.521705	0.009726	0.014221

Balanced accuracy with feature selection

	model	mean_acc_with_sampling	mean_acc_without_sampling	std_acc_with_sampling	std_acc_without_sampling
0	Logistic Regression	0.761167	0.499984	0.016733	0.000033
1	KNN	0.598478	0.505176	0.018497	0.003539
3	LightGBM	0.620180	0.499804	0.013563	0.000122
2	XGBoost	0.573361	0.502584	0.016296	0.002314
1	Random Forest	0.557729	0.512504	0.009695	0.005576
0	Desicion Tree	0.555190	0.518245	0.008589	0.008865

Model Stability

without oversampling and without feature selction

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
Logistic Regression	0.499984	0.817617	0.499935
KNN	0.515768	0.592402	0.503171

without oversampling & with feature selction

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
Logistic Regression	0.500905	0.817617	0.499935
KNN	0.522073	0.592402	0.506734

with oversampling & without feature selection

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
Logistic Regression	0.762479	0.814232	0.737987
KNN	0.960829	0.661572	0.628980

with oversampling & with feature selection

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
Logistic Regression	0.763236	0.820615	0.741127
KNN	0.826675	0.661699	0.557716

Model Stability

without oversampling and without feature selction

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
LightGBM	0.535647	0.810924	0.503694
XGBoost	0.542214	0.781210	0.499738
Random Forest	0.960551	0.687611	0.501470
Desicion Tree	0.959662	0.522206	0.526803

with oversampling & without feature selection

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
LightGBM	0.663517	0.803535	0.578550
XGBoost	0.654917	0.789781	0.535860
Random Forest	0.978097	0.726795	0.535891
Desicion Tree	0.959662	0.532399	0.532849

without oversampling & with feature selction

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
LightGBM	0.512179	0.806682	0.499804
XGBoost	0.525296	0.795167	0.499542
Random Forest	0.756758	0.667192	0.508531
Desicion Tree	0.750371	0.570697	0.502941

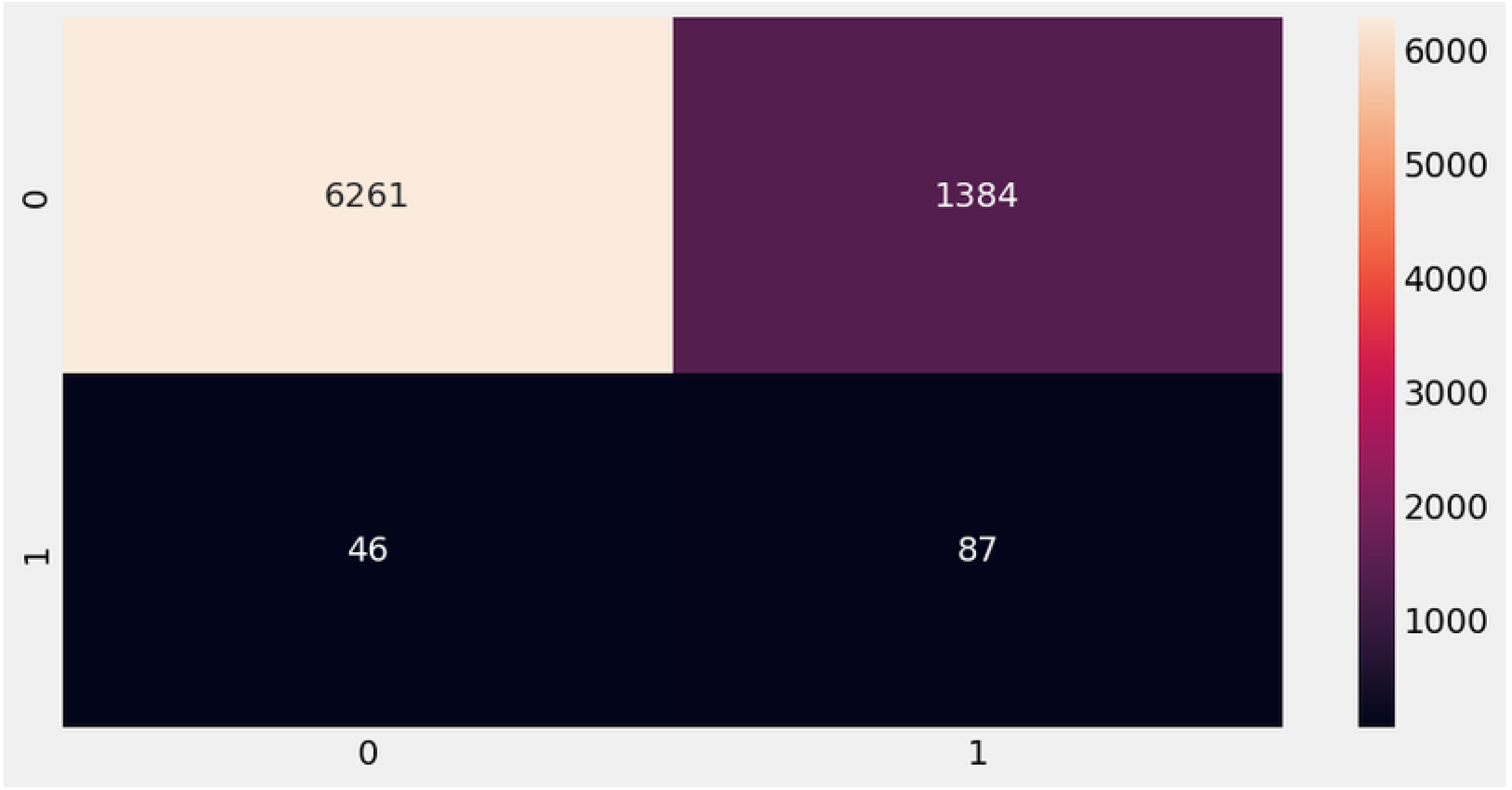
with oversampling & with feature selection

model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
LightGBM	0.734078	0.809756	0.642520
XGBoost	0.697599	0.795564	0.563447
Random Forest	0.833656	0.685810	0.552791
Desicion Tree	0.805376	0.579932	0.544749

Model performance comparison

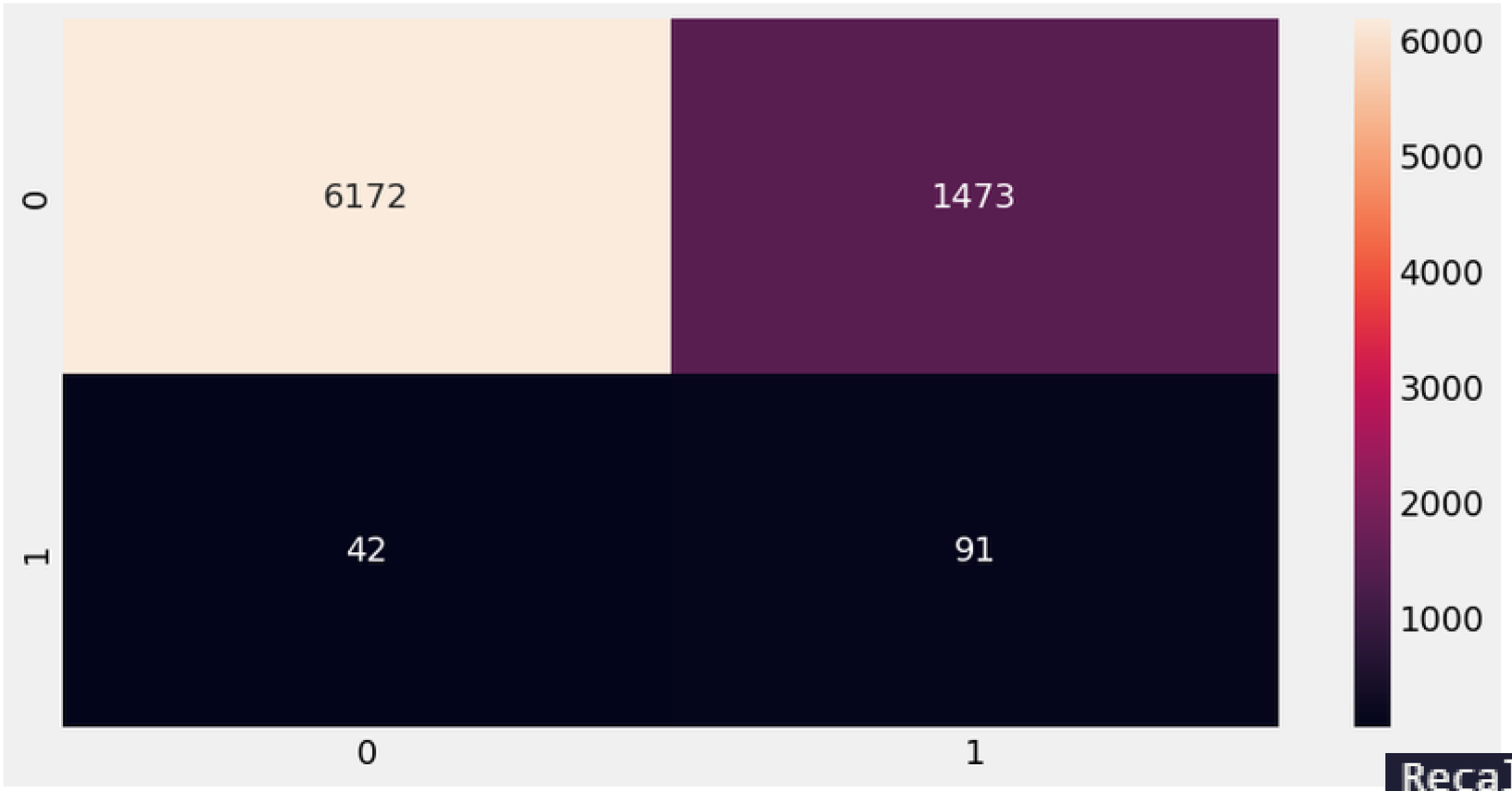
model	rocauc score (train set)	rocauc score (validation set)	rocauc score (test set)
Logistic Regression	0.763236	0.820615	0.741127
LightGBM	0.734078	0.809756	0.642520
XGBoost	0.697599	0.795564	0.563447
Random Forest	0.833656	0.685810	0.552791
KNN	0.826675	0.661699	0.557716
Desicion Tree	0.805376	0.579932	0.544749

Comparison model before tuning



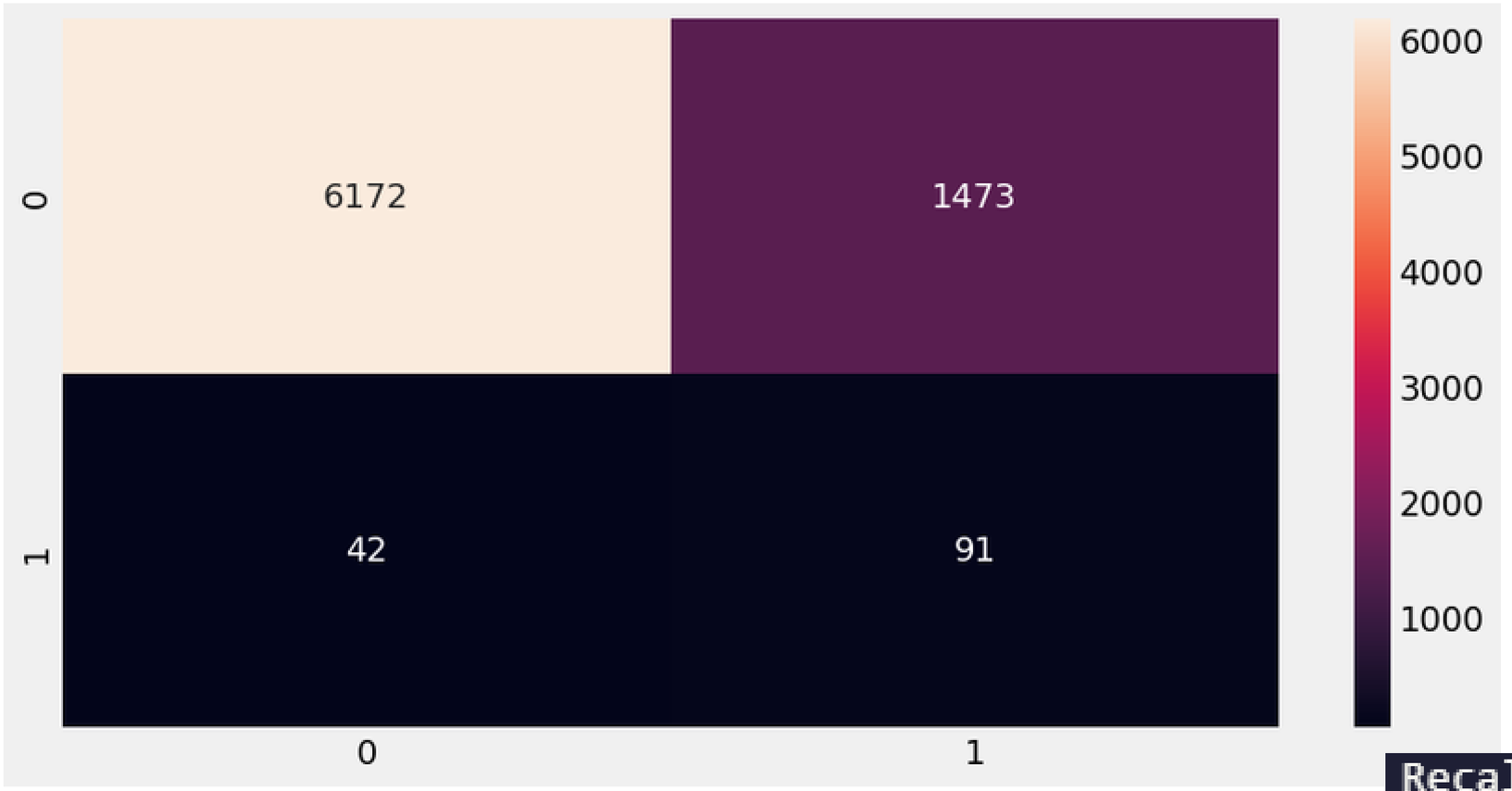

```
Recall: 0.6541353383458647
Precision: 0.059143439836845 68
ROC: 0.7365509916058951
Accuracy: 0.736550991605895
```

Comparison model after tuning




```
Recall: 0.6842105263157895
Precision: 0.05818414322250639
ROC: 0.7457677876837285
Accuracy: 0.7457677876837286
```

Comparison model after tuning




```
Recall: 0.6842105263157895
Precision: 0.05818414322250639
ROC: 0.7457677876837285
Accuracy: 0.7457677876837286
```

End Session

Kesimpulan

Model ini hanya dapat digunakan pada data dengan nilai maximum dan minimum pada tiap variabel berikut / penggunaannya menjadi tidak valid jika nilai datanya diluar jangkauan ini:

- Age: 0 hingga 88
- Net sales: -357.5 hingga 682.0
- Duration: 0 hingga 547
- Commision : 0.0 hingga 262.76

Rekomendasi

Rekomendasi untuk penjualan produk:

- Tetap lakukan penjualan produk kepada semua rentang umur
- Lanjutkan pemasaran produk asuransi perjalanan melalui agensi penerbangan dan agensi perjalanan
- Perhatikan produk-produk dengan rasio klaim yang tinggi seperti Annual Silver Plan, Annual Gold Plan, Annual Travel Protect Gold, Single Trip Travel Protect Platinum, dan Spouse or Parents Comprehensive Plan
- Perusahaan perlu mempertimbangkan untuk memperluas daftar destinasi yang dapat dicover oleh produk asuransinya

Untuk mengembangkan model yang lebih baik:

- Tambahkan fitur atau kolom baru yang berhubungan dengan status klaim
- Coba menggunakan algoritma Machine Learning yang berbeda dan lakukan hyperparameter tuning yang lebih luas
- Perluas penggunaan metrik dan algoritma classification lainnya seperti CATboost, Gradient Boosting, ADAboost, dan sebagainya
-

PURWADHIKA SCHOOL DIGITAL

Thank you for your time!