ITTelkom Surabaya
Solution For The Nation

THESION MARTA SIANIPAR

RENDIKA NURHARTANTO S

RIZAL RAHMAN RIZKIKA

ROYALS OUTLAWS

INSTITUT TEKNOLOGI TELKOM SURABAYA

FINAL PROJECT

BOOTCAMP

DIBIMBING.ID

GSB ACADEMY

24 SEPTEMBER 2022

DATASET

"HEALTH INSURANCE CROSS SELL PREDICTION"

evermos

01

Background

ITTelkom Surabaya
Solution For The Nation

“HEALTH INSURANCE CROSS SELL PREDICTION”

1. Perusahaan Asuransi sedang melakukan riset untuk membuka produk baru yaitu Asuransi Kendaraan
2. Perusahaan mengalami kesulitan dalam mengamati ketertarikan costumer yang cocok untuk diberi tawaran asuransi
3. Dibutuhkan analisis mendalam mengenai karakteristik customer yang berminat terhadap produk baru tersebut.
4. Untuk memprediksi minat customer terhadap asuransi dibutuhkan model prediksi yang mumpuni dan efisien.

02

Objective and Problem Statements

OBJECTIVE AND PROBLEM STATEMENTS

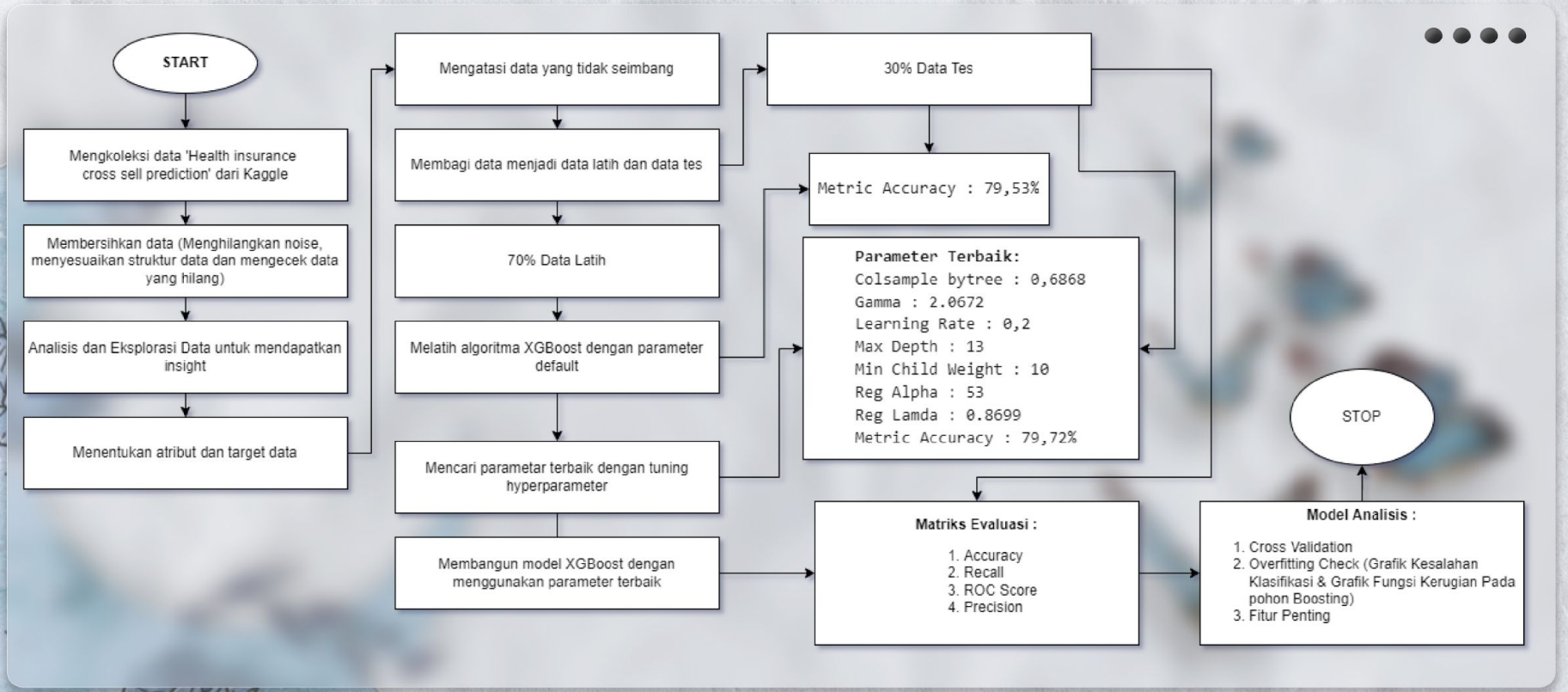
PROBLEM STATMENT

1. Bagaimana cara mendapatkan insight dari karakteristik costumer?
2. Bagaimana cara membuat model prediksi untuk menentukan ketertarikan customer?

OBJECTIVE

1. Mengeksplorasi data untuk mendapatkan insight dari karakteristik customer
2. Membangun model klasifikasi dengan menggunakan metode extreme gradient boosting untuk memprediksi ketertarikan customer

METHODOLOGY RESEARCH



evermos

03

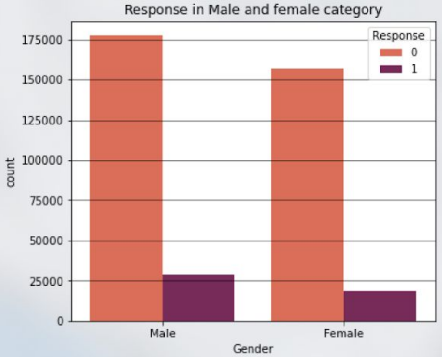
Data Preparation (EDA & Feature Engineering)

ITTelkom Surabaya
Solution For The Nation

DATA SET TERDIRI DARI 12 ATRIBUT KLASIFIKASI DATA

ID (IDENTITY NUMBER)	GENDER	AGE
DRIVING LICENSE	REGION CODE	PREVIOUSLY INSURED
VEHICLE AGE	VEHICLE DAMAGE	ANNUAL PREMIUM
POLICY SALES CHANNEL	VINTAGE	RESPONSE

EDA (EXPLORATORY DATA ANALYSIS)

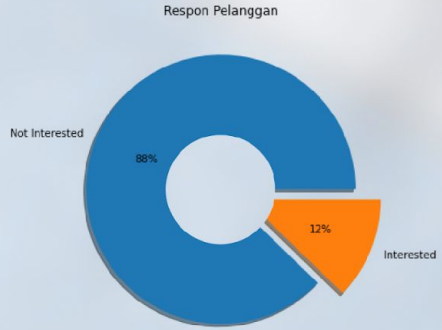


Response	0	1
Gender		
Female	156835	18185
Male	177564	28525

1

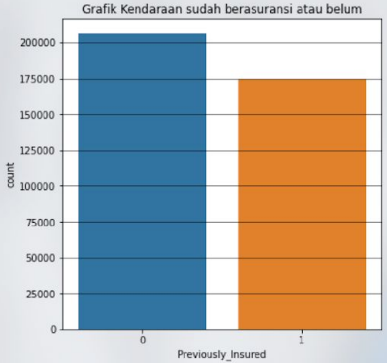
Vehicle_Age	
1-2 Year	200316
< 1 Year	164786
> 2 Years	16007

2



	Response
Response	
0	334399
1	46710

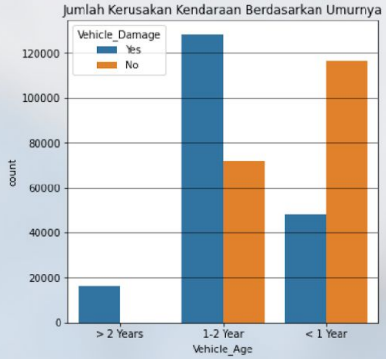
3



Previously_Insured	
0	206481
1	174628

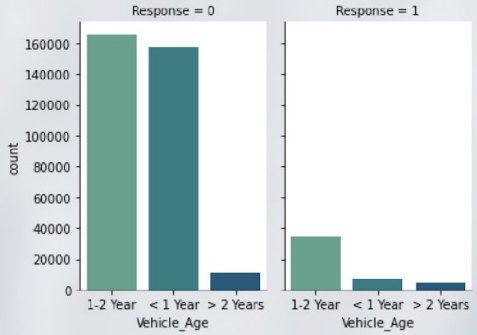
4

EDA (EXPLORATORY DATA ANALYSIS)



Vehicle_Damage	No	Yes
Vehicle_Age		
1-2 Year	72091	128225
< 1 Year	116590	48196
> 2 Years	15	15992

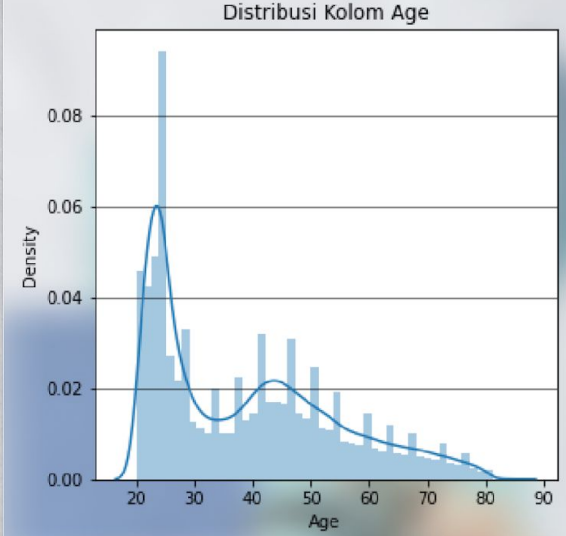
5

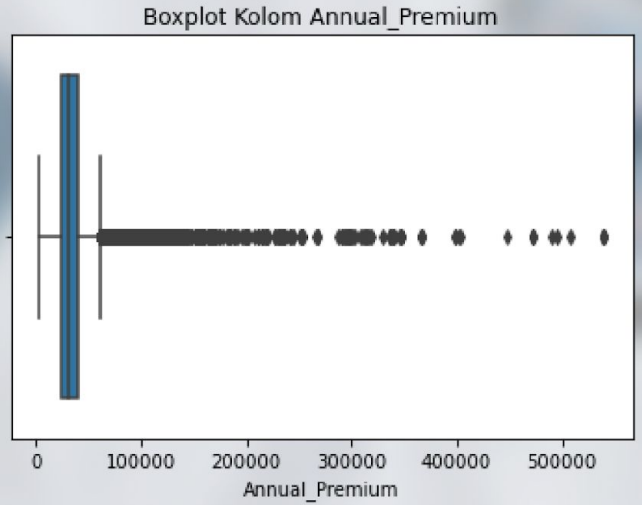


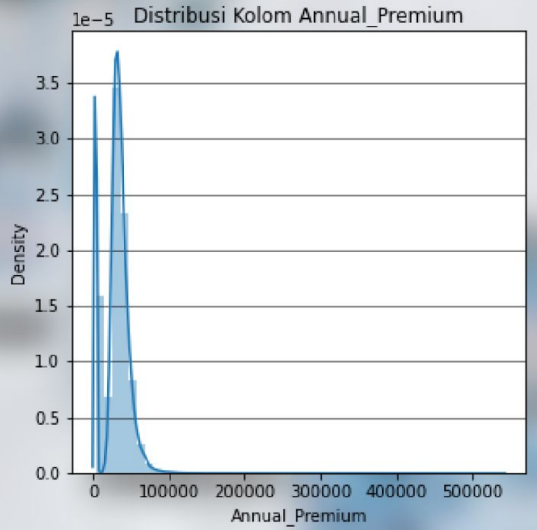
Response	0	1
Vehicle_Age		
1-2 Year	165510	34806
< 1 Year	157584	7202
> 2 Years	11305	4702

6

FEATURE ENGINEERING







evermos

04

Modeling

MODELING

```
train['Response'].value_counts()
```

```
0    334399
1     46710
Name: Response, dtype: int64
```

```
X = train.drop(['Response'],axis=1)
y = train['Response']
X_under, y_under = under_sampling.RandomUnderSampler().fit_resample(X, y)

df_undersampling = pd.concat([X_under, y_under], axis=1)
df_undersampling.head(3)
df_undersampling.shape
```

```
(93420, 11)
```

```
y_under.value_counts()
```

```
0    46710
1    46710
Name: Response, dtype: int64
```

```
model_xgb = XGBClassifier()
model_xgb.fit(X_train, y_train)
pred_xgb = model_xgb.predict(X_test)
accuracy_xgb = accuracy_score(y_test, pred_xgb)
print("Accuracy:" ,accuracy_xgb)
```

```
[21:55:37] WARNING: C:\Windows\Temp\abs_557yfx631l\croots\recipe\xgboost-split_1659548953302\work\src\learner.cc:1115: Starting
in XGBoost 1.3.0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'loglos
s'. Explicitly set eval_metric if you'd like to restore the old behavior.
Accuracy: 0.7953329051594947
```

MODELING

No	Score
1	0.7982230785698994
2	0.7980446728038251
3	0.7979376293441804
4	0.7978662670377507
5	0.7978305858845358
6	0.797759223578106
7	0.7977235424248912
8	0.7976521801184614
9	0.7975094555056019
10	0.7973667308927425

10 Tertinggi dari hasil uji coba Hyperparameter

No	Best Parameter
1	colsample_bytree= 0.6867988935946676
2	gamma= 2.0672479277514437
3	learning_rate= 0.2
4	max_depth= 13
5	min_child_weight= 10
6	reg_alpha= 53.0
7	reg_lambda= 0.8698526815895538
8	n_estimators = 100

MODELING

```
model_xgb_bestparams = XGBClassifier(colsample_bytree= 0.6867988935946676, gamma= 2.0672479277514437, learning_rate= 0.2,
                                     max_depth= 13, min_child_weight= 10, reg_alpha= 53.0,
                                     reg_lambda= 0.8698526815895538,n_estimators = 100 )
eval_set = [(X_train, y_train),(X_test,y_test)]
model_xgb_bestparams.fit(X_train, y_train,eval_metric=['error','logloss'],eval_set = eval_set)

pred_xgb = model_xgb_bestparams.predict(X_test)
predictions_xgb = [round(value) for value in pred_xgb]
accuracy_xgb = accuracy_score(y_test, predictions_xgb)
print("Accuracy:" ,accuracy_xgb)
```

```
[91]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[92]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[93]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[94]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[95]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[96]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[97]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[98]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
[99]    validation_0-error:0.20202      validation_0-logloss:0.42296    validation_1-error:0.20278      validation_1-loglos
s:0.42536
Accuracy: 0.797224006279883
```

MODELING

```
metrics.recall_score(y_test,predictions_xgb)
```

```
0.9317564670573103
```

```
metrics.roc_auc_score(y_test,predictions_xgb)
```

```
0.7974061696746072
```

```
metrics.precision_score(y_test, predictions_xgb)
```

```
0.7338886700061913
```

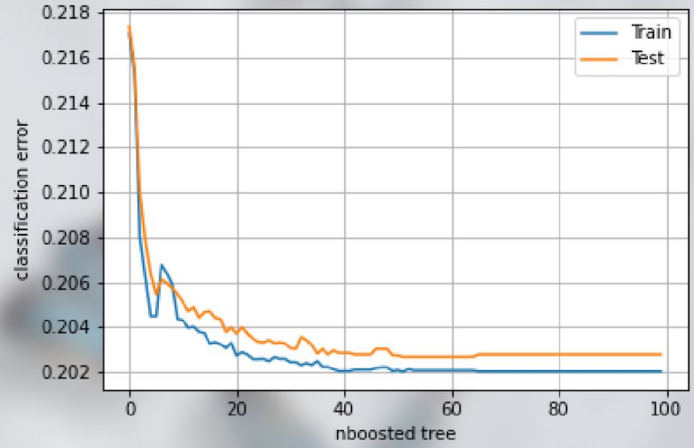
```
confusion_matrix(y_test,predictions_xgb)
```

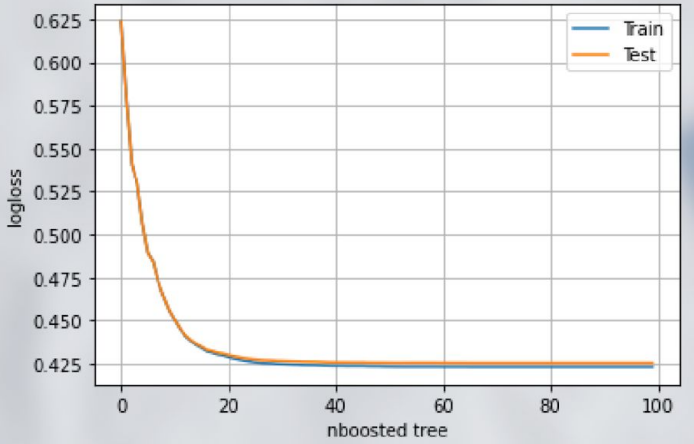
```
array([[ 9304,  4728],
       [  955, 13039]], dtype=int64)
```

```
print('akurasi cross validation = %.2f%%(%.2f%%)' %(result.mean()*100,result.std()*100))
```

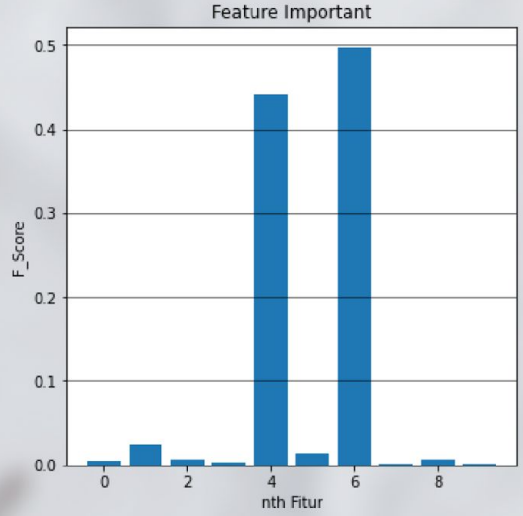
```
akurasi cross validation = 79.64%(0.43%)
```

MODELING


0.218
0.216
0.214
0.212
0.210
0.208
0.206
0.204
0.202
classification error
0
20
40
60
80
100
nboosted tree
Train
Test


0.625
0.600
0.575
0.550
0.525
0.500
0.475
0.450
0.425
logloss
0
20
40
60
80
100
nboosted tree
Train
Test

MODELING



0	1	2	3	4	5	6	7	8	9	10
Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Respons

evermos

05

Summary

SUMMARY

Mulai dari memuat dataset untuk memahami data terlebih dahulu, kemudian melakukan Exploratory data analysis, data cleaning, data manipulation dan encoding. Setelah itu mengatasi imbalanced handling dengan menggunakan metode under sampling lalu uji coba menggunakan model XGBoost dengan hyperparameter default dan menghasilkan akurasi 79.5%. Kemudian menguji model xgboost dengan tuned hyperparameter dan mendapatkan skor 79,6%. Dari hasil hyperparameter yang mendapatkan metric evaluasi, didapatkan recall dengan skor 0.927852708199529, ROC_AUC dengan skor 0.7927638621280241, precision dengan skor 0.7304904769930896. Dari hasil confusion matrix yang perlu diperharikan adalah False Negative yaitu ketika response pelanggan diprediksi tidak tertarik tetapi sebenarnya pelanggan tersebut tertarik. Untuk mengevaluasi kinerja model dilakukan cross validation yang menghasilkan akurasi 79.48%. Dan hasilnya tidak jauh beda dengan akurasi dari model XGBoost yang menggunakan hyperparameter. Berdasarkan hasil pemeriksaan overfitting, tidak terdapat overfitting pada hasil pemodelan dapat dilihat dari grafik yang ada. Berdasarkan kepentingan fitur, kolom Previous_Insured & Vehicle_Damage adalah kolom fitur yang paling berpengaruh dalam pemodelan prediktif. Model prediksi ini dapat digunakan untuk memperkirakan response customer Terhadap produk asuransi yang baru.

THANK YOU

FOR YOUR ATTENTION

BOOTCAMP

DIBIMBING.ID

ROYALS OUTLAWS

INSTITUT TEKNOLOGI TELKOM SURABAYA