



A smartphone screen displays three team members with their names and profile pictures:

- THESION MARTA SIANIPAR
- RENDIKA NURHARTANTO S
- RIZAL RAHMAN RIZKIKA

At the bottom, it says "ROYALS OUTLAWS" and "INSTITUT TEKNOLOGI TELKOM SURABAYA".



FINAL PROJECT



BOOTCAMP
DIBIMBING.ID

GSB ACADEMY

24 SEPTEMBER 2022

DATASET

"HEALTH INSURANCE CROSS SELL PREDICTION"



01. Background
02. Problem Statements
03. Data Preparation
04. Modeling
05. Summary

01

Background



“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

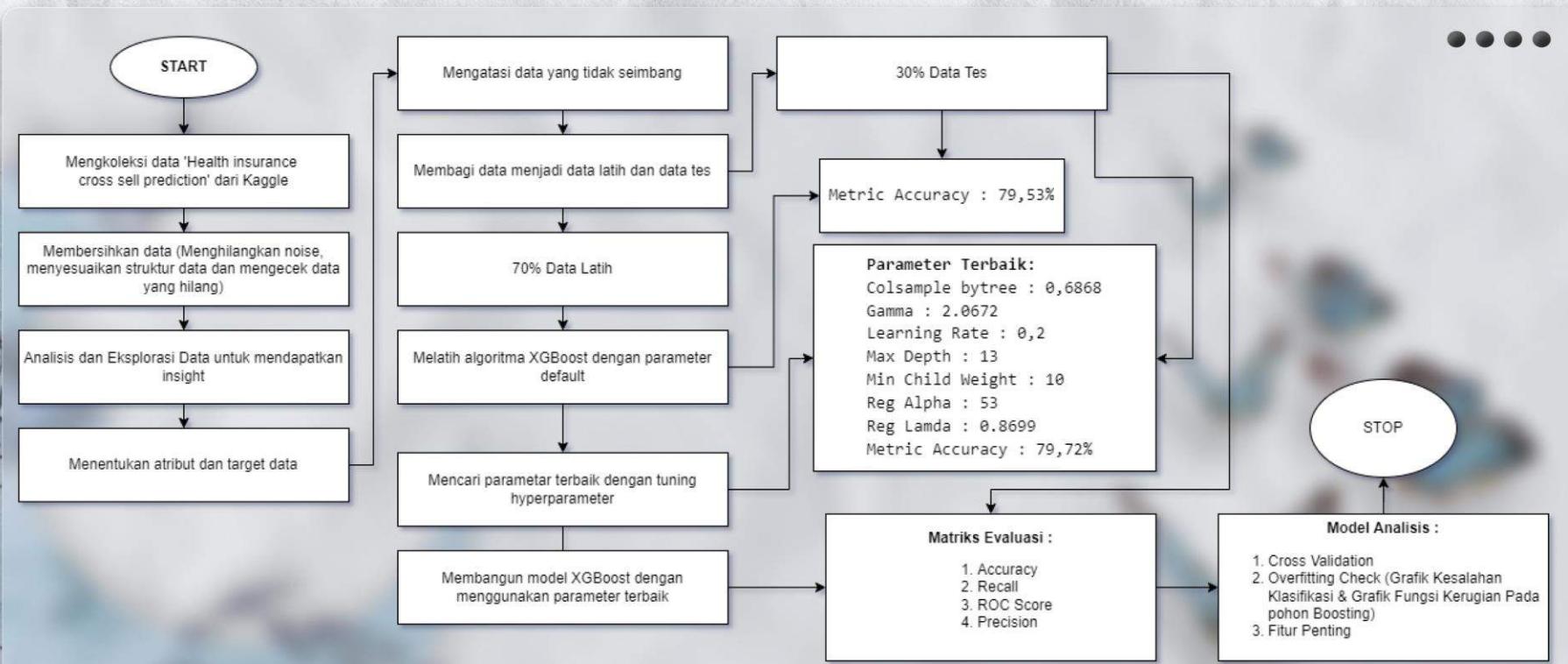
PROBLEM STATEMENT

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



03

Data Preparation (EDA & Feature Engineering)

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



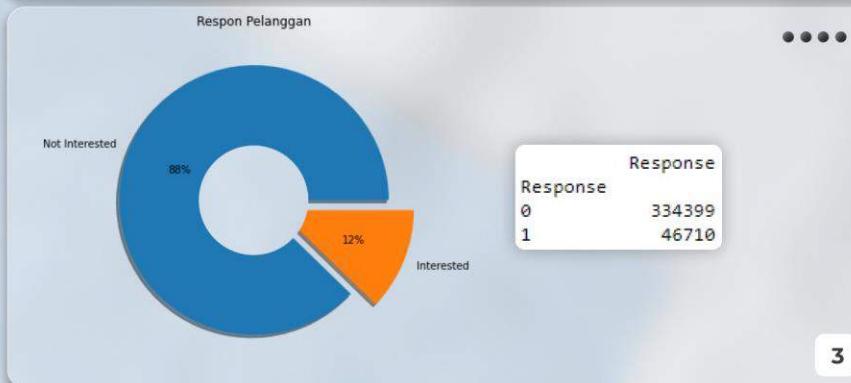
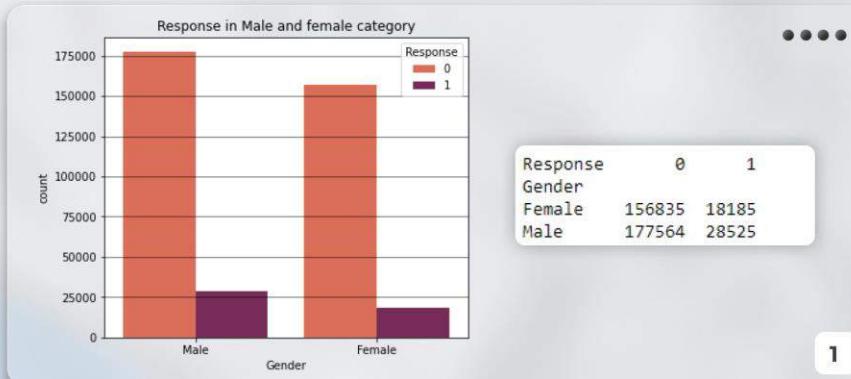
The civil...
Exchange Co...
against Goldm...
called Abacus...
bank created i...
son could line...
the value of...
prized asset c...
the late eight...
fashion every...

Morgan St...
Presidents de...
Jackson. Anot...
Libertas, defi...
Virgin Island...
JPMorgan Ch...
netar, a hedge...
from the mel...
Kellogg Sch...
it last year in...
Lewis Sach...
Department...
cently a tax...
bonds that i...
now claim we...
going and even...
deals were clo...
Bank of America too is

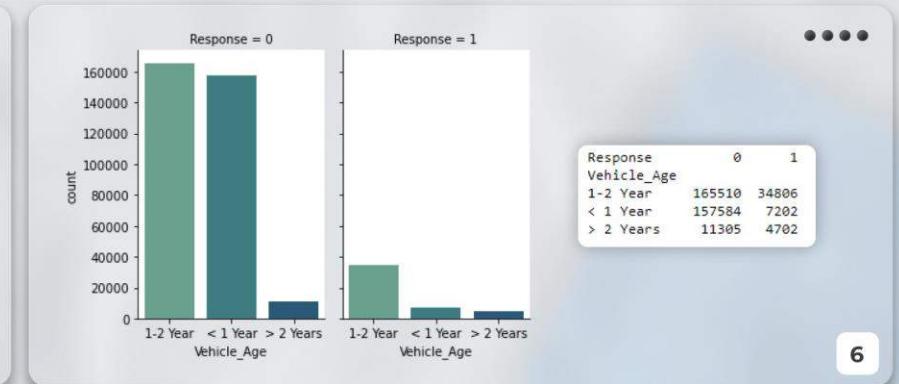
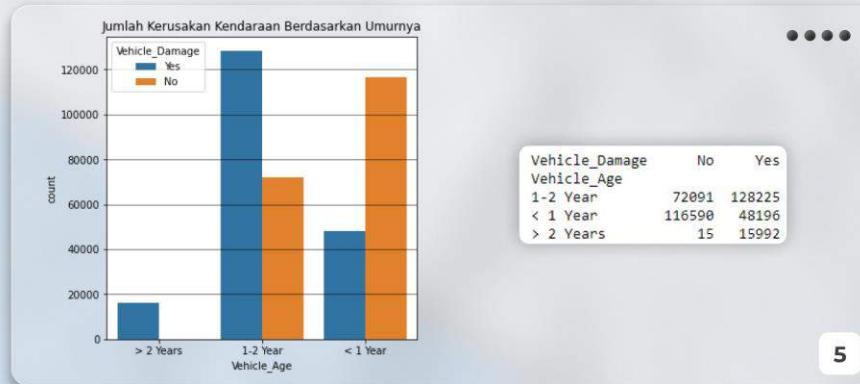
— PHILIPP MEYER, FORMER TRADER

and the hedge funds that

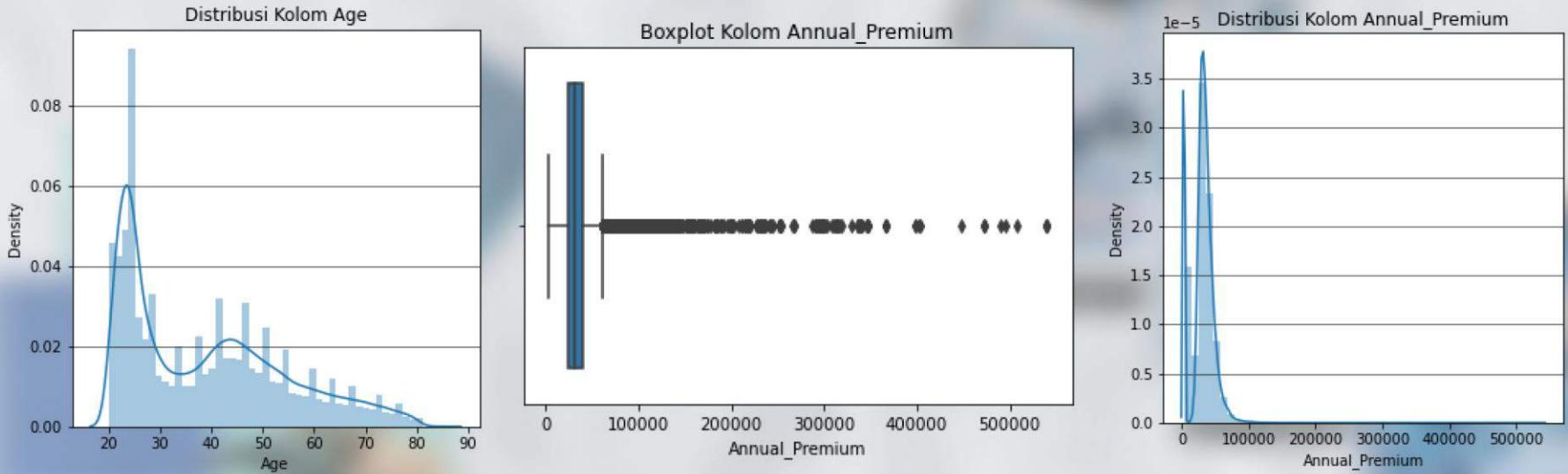
EDA (EXPLORATORY DATA ANALYSIS)



EDA (EXPLORATORY DATA ANALYSIS)



FEATURE ENGINEERING



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_165954895302\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 'logloss'.
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
```

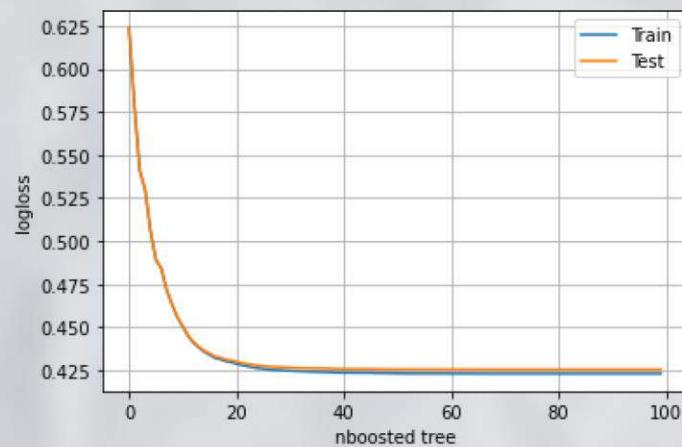
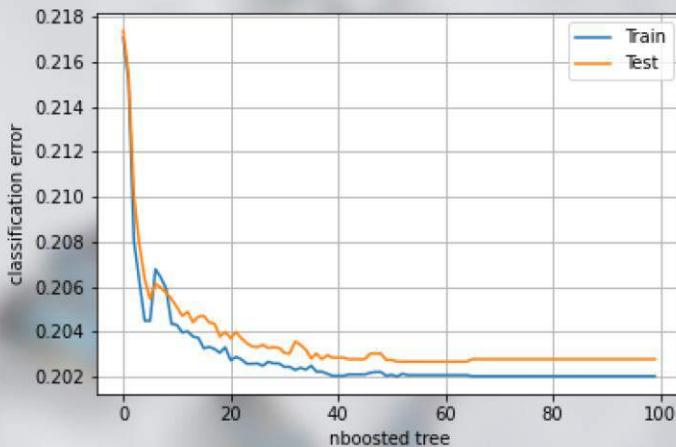
```
confusion_matrix(y_test,predictions_xgb)
```

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

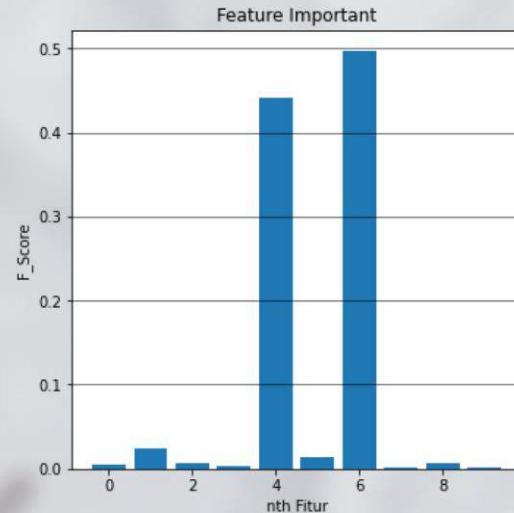
```
print('akurasi cross validation = %.2f%%(% .2f%%)' %(result.mean()*100,result.std()*100))
```

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

MODELING



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

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 diperhatikan 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

