Rakamin Academy

In-Vehicle Coupon Recommendation

Dokumen Laporan Final Project

DEEP LEARNING 4.0

DATA CLEANSING

HANDLE MISSING VALUES

```
Missing Values:
destination                0
passanger                  0
weather                    0
temperature                0
time                       0
coupon                     0
expiration                 0
gender                     0
age                        0
maritalStatus              0
has_children               0
education                  0
occupation                 0
income                     0
car                    12576
Bar                      107
CoffeeHouse              217
CarryAway                151
RestaurantLessThan20     130
Restaurant20To50         189
toCoupon_GEQ5min           0
toCoupon_GEQ15min          0
toCoupon_GEQ25min          0
direction_same             0
direction_opp              0
Y                          0
dtype: int64
```

```
#drop fitur car (banyak Missing value)
#---------- direction_opp (karena bernilai konstan)
#---------- toCoupon_GEQ5min  (karena multicolinearity)

df_clone.drop(columns=['car','direction_opp','toCoupon_GEQ5min'] , inplace=True)
df_clone.info()
```

```
df_clone= df.copy()
columns= ['Bar','CoffeeHouse','CarryAway','RestaurantLessThan20','Restaurant20To50']
for i in columns:
    column = df_clone[i]
    df_clone[i]= column.fillna(column.mode()[0])
df_clone.isna().sum()

#df_clone['Bar'].mode()
```

missing values diisi dengan modus ditiap kolomnya.

```
destination            0
passanger              0
weather                0
temperature            0
time                   0
coupon                 0
expiration             0
gender                 0
age                    0
maritalStatus          0
has_children           0
education              0
occupation             0
income                 0
Bar                    0
CoffeeHouse            0
CarryAway              0
RestaurantLessThan20   0
Restaurant20To50       0
toCoupon_GEQ15min      0
toCoupon_GEQ25min      0
direction_same         0
Y                      0
dtype: int64
```

DATA CLEANSING

HANDLE DUPLICATED DATA

```
df_clone.duplicated().sum()
```

```
74
```

```
df = df.drop_duplicates()
```

✓ 0.1s

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

✓ 0.1s

```
0
```

```
df.shape
```

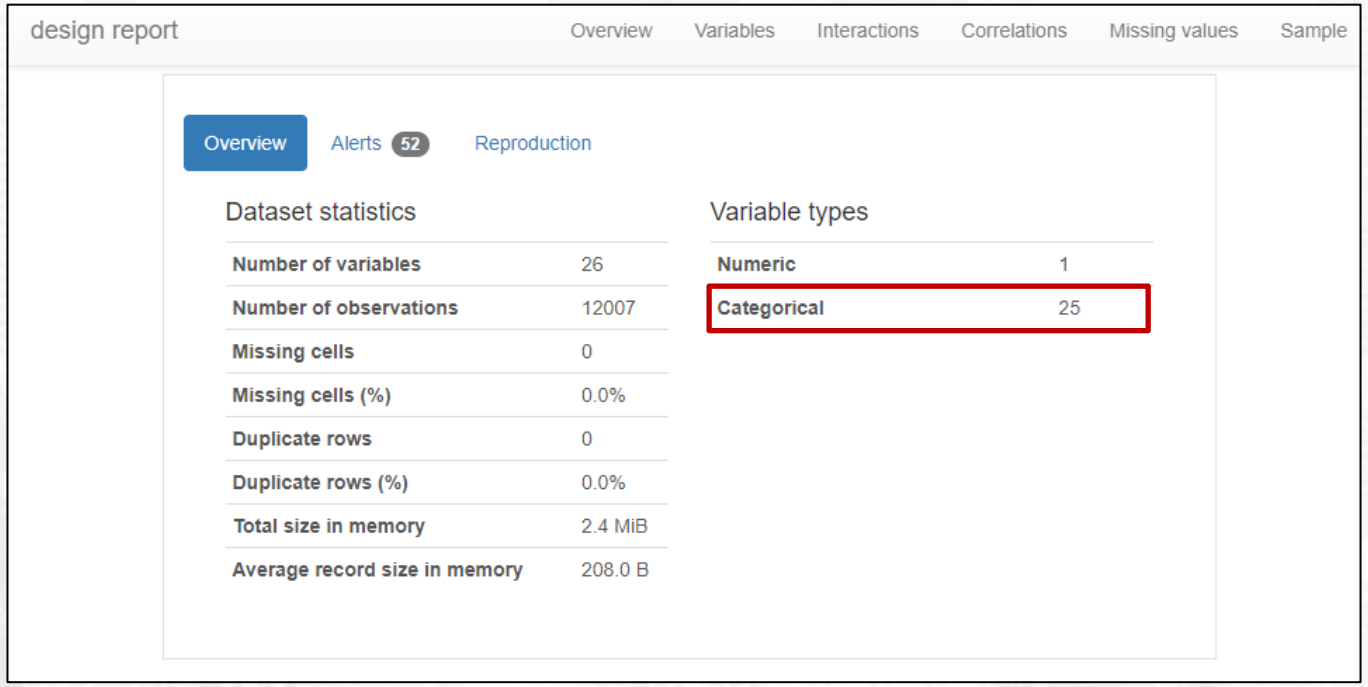
✓ 0.1s

```
(12610, 23)
```

Pada dataset ini ditemukan sebanyak 74 baris data yang terduplikasi. Data duplicated ini didrop untuk meningkatkan variasi fitur untuk model machine learning

DATA CLEANSING

HANDLE OUTLIERS



Tidak dilakukan handle outliers, karena seluruh fitur pada dataset ini merupakan kategorikal. Selain itu, unique value yang terdapat pada fitur masih masuk akal dan nyata angkanya

DATA CLEANSING

FEATURE TRANSFORMATION

```
#checking age unique value
df['age'].value_counts()
```

```
21          2653
26          2559
31          2039
50plus      1788
36          1319
41          1093
46           686
below21      547
Name: age, dtype: int64
```

```
#categorize it to be more simple
age_list = []
for i in df['age']:
    if i == 'below21':
        age = '<21'
    elif i == '21' or i == '26':
        age = '21-30'
    elif i == '31' or i == '36':
        age = '31-40'
    elif i == '41' or i == '46':
        age = '41-50'
    else:
        age = '>50'
    age_list.append(age)
df['age'] = age_list
```

```
#checking new age unique value
df['age'].value_counts()
```

```
21-30    5212
31-40    3358
>50      1788
41-50    1779
<21       547
Name: age, dtype: int64
```

Fitur age sebelumnya memiliki 7 unique values, dilakukan penyusutan menjadi 5 unique values

DATA CLEANSING

FEATURE ENCODING

```
# Implement the label encoding for column expiration, gender, age, education, Bar, CoffeeHouse, CarryAway, Re
df_2 = df_1.replace({'expiration':{'2h': 0, '1d' : 1},
                     'gender':{'Male': 0, 'Female' : 1},
                     'age':{'<21': 0, '21-30': 1, '31-40': 2, '41-50': 3, '>50': 4},
                     'education':{'Some High School': 0, 'High School Graduate': 1, 'Some college - no degree'
                     'Bar':{'never': 0, 'less1': 1, '1~3': 2, '4~8': 3, 'gt8': 4},
                     'CoffeeHouse':{'never': 0, 'less1': 1, '1~3': 2, '4~8': 3, 'gt8': 4},
                     'CarryAway':{'never': 0, 'less1': 1, '1~3': 2, '4~8': 3, 'gt8': 4},
                     'RestaurantLessThan20':{'never': 0, 'less1': 1, '1~3': 2, '4~8': 3, 'gt8': 4},
                     'Restaurant20To50':{'never': 0, 'less1': 1, '1~3': 2, '4~8': 3, 'gt8': 4},
                     'temperature':{30: 0, 55: 1, 80: 2},
                     'income':{'Less than $12500':0, '$12500 - $24999':1, '$25000 - $37499':2, '$37500 - $4999
                     'time':{'7AM':0, '10AM':1, '2PM':2, '6PM':3, '10PM':4}})
```

```
# Implement One Hot Encoding for column destination, passanger, marital status, occupation, coupon, & weather
oh_list = ['destination', 'passanger', 'maritalStatus', 'occupation', 'coupon', 'weather']
df_oh= pd.get_dummies(df_2[oh_list], columns=oh_list)
```

```
#merging label encoding columns and one hot encoding columns
df_le = pd.concat([df_oh, df_2], axis = 1)
df_le = df_le.drop(columns=['destination', 'passanger', 'maritalStatus', 'occupation', 'coupon', 'weather'])
```

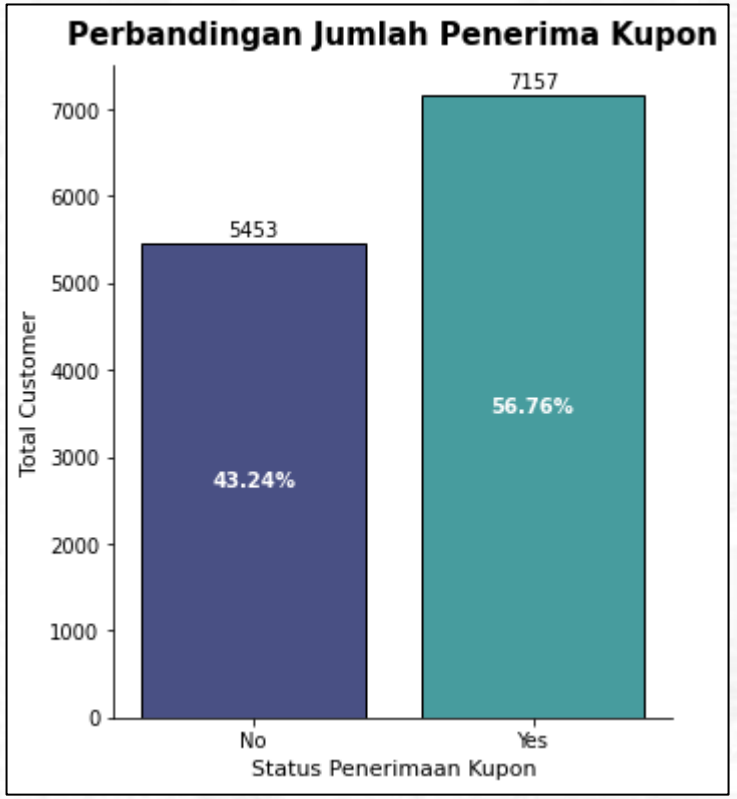
```
df_encode.shape
✓ 0.1s
(12610, 62)
```

- Label encoding dilakukan terhadap fitur, seperti ‘expiration’, ‘gender’, ‘age’, ‘education’, ‘Bar’, ‘CoffeeHouse’, ‘CarryAway’, ‘RestaurantLessThan20’, ‘Restaurant20To50’, ‘temperature’, ‘income’, ‘time’.
- One hot encoding dilakukan terhadap fitur, seperti: ‘destination’, ‘passanger’, ‘maritalStatus’, ‘occupation’, ‘coupon’, ‘weather’.
- Setelah dialkukan feature encoding, jumlah fitur yang dimiliki saat ini menjadi 65 fitur.

DATA CLEANSING

HANDLE CLASS IMBALANCE

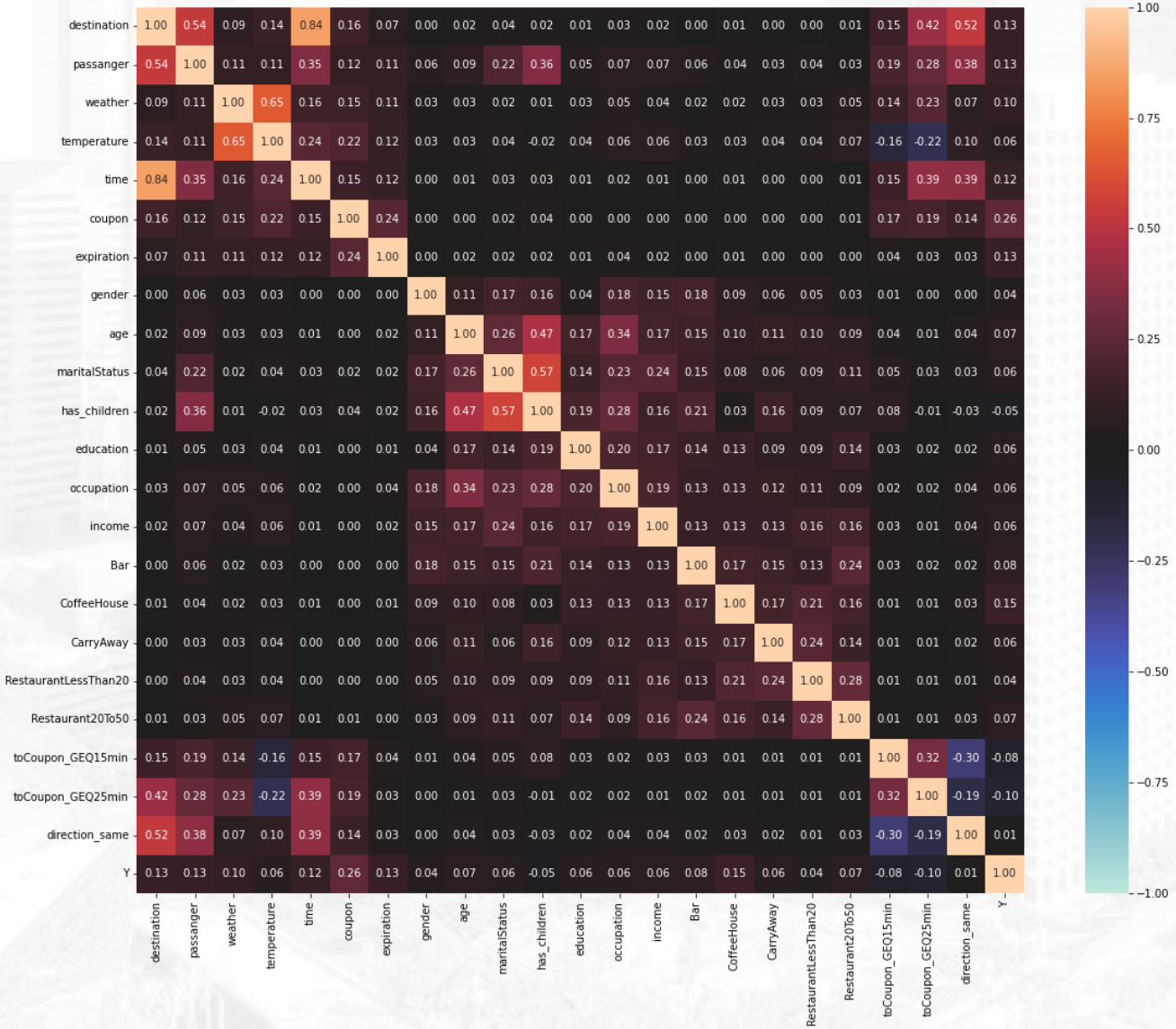
	Y	customers	total_customer	percentage
0	0	5453	12610	43.24
1	1	7157	12610	56.76



tidak dilakukan handle class imbalance karena perbandingan ratio pada class tidak lah ekstrim, yakni 57:43.

FEATURE ENGINEERING

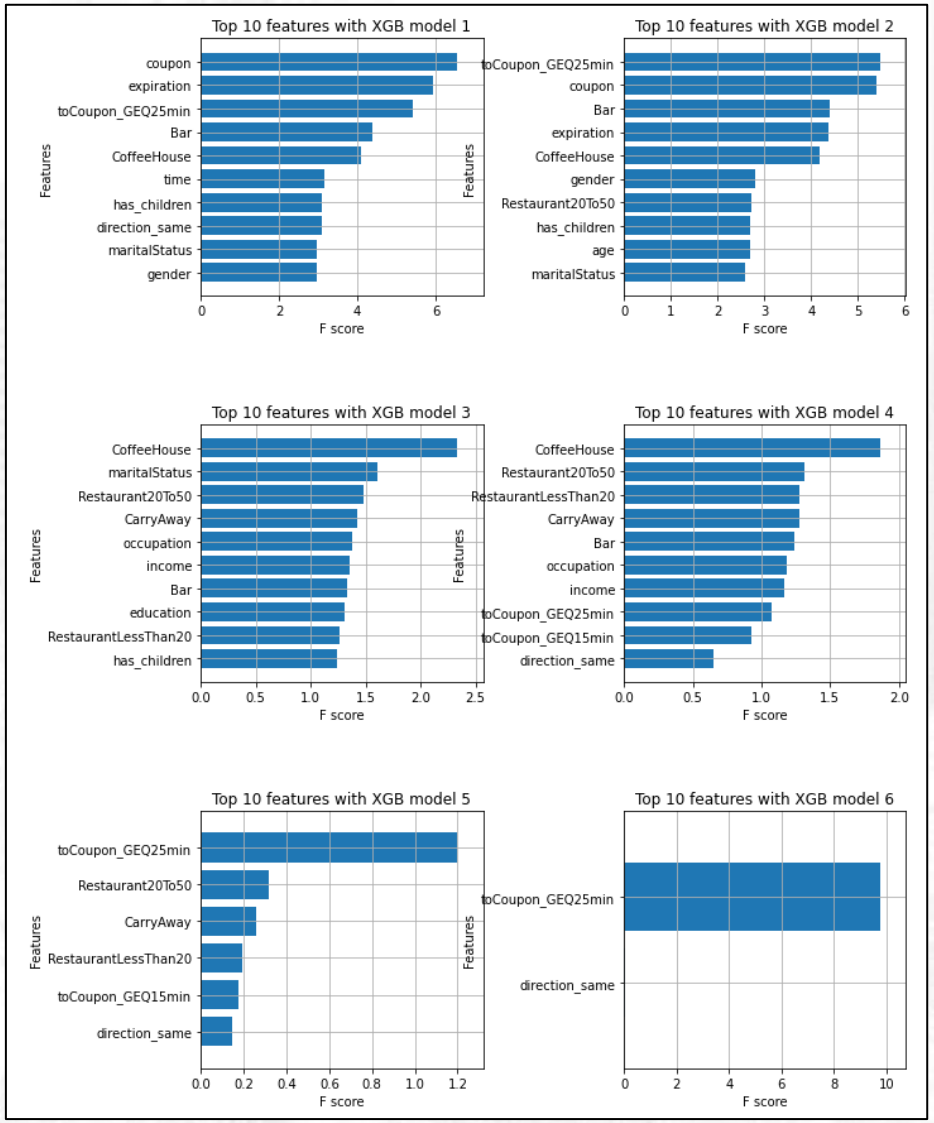
FEATURE SELECTION



- Heatmap correlation dihasilkan menggunakan metode cramer

FEATURE ENGINEERING

FEATURE SELECTION




```
Selected 19 important features:
['toCoupon_GEQ25min', 'coupon', 'expiration', 'CoffeeHouse', 'Bar', 'time',
'direction_same', 'maritalStatus', 'has_children', 'passanger', 'gender', 'occupation',
'Restaurant20To50', 'CarryAway', 'income', 'education', 'RestaurantLessThan20', 'age',
'toCoupon_GEQ15min']
```

- Metode ini dilakukan menggunakan pyhton library featurewiz.
- Daftar fitur penting yang dihasilkan ini berawal dari fitur di dataset yang disortir menggunakan algoritma SULOV (Searching for Uncorrelated List of Variables) lalu diuji sebanyak 6x pengulangan dengan XGBoost.

DATA CLEANSING

FEATURE EXTRACTION

```
Unemployed                                  1861
Student                                     1575
Computer & Mathematical                     1390
Sales & Related                             1088
Education&Training&Library                   939
Management                                   821
Office & Administrative Support              638
Arts Design Entertainment Sports & Media     627
Business & Financial                         543
Retired                                      493
Food Preparation & Serving Related           298
Healthcare Practitioners & Technical         244
Healthcare Support                           242
Community & Social Services                  239
Legal                                        219
Transportation & Material Moving             218
Architecture & Engineering                   175
Personal Care & Service                      175
Protective Service                           174
Life Physical Social Science                 169
Construction & Extraction                    154
Installation Maintenance & Repair            133
Production Occupations                       108
Building & Grounds Cleaning & Maintenance     44
Farming Fishing & Forestry                    43
Name: occupation, dtype: int64
```

```python
occupation_list = []
for i in df['occupation']:
    if i == 'Installation Maintenance & Repair' or i == 'Transp
        occupation = 'Crafts'
    elif i == 'Architecture & Engineering' or i == 'Education&T
        occupation = 'Professionals'
    elif i == 'Retired':
        occupation = 'Retired'
    elif i == 'Sales & Related' or i == 'Personal Care & Servic
        occupation = 'Service and sales'
    elif i == 'Student':
        occupation = 'Student'
    elif i == 'Healthcare Support' or i == 'Life Physical Socia
        occupation = 'Technicians'
    elif i == 'Unemployed':
        occupation = 'Unemployed'
    else:
        occupation = 'Others'
    occupation_list.append(occupation)
df_extract['occupation'] = occupation_list
```

✓ 0.1s

```python
#checking occupation unique value
df_extract['occupation'].value_counts()
```

✓ 0.1s

```
Professionals        4958
Unemployed           1861
Student              1575
Service and sales    1437
Technicians           804
Others                789
Crafts                693
Retired               493
Name: occupation, dtype: int64
```

Melakukan ekstraksi fitur occupation dengan mengkategorisasikannya menjadi lebih sedikit sebanyak 8 kategori

FEATURE ENGINEERING

FEATURE EXTRACTION

```
df_extract = df_dummy.copy()
```

```
df_extract['passanger_destination'] = df_extract['passanger'].astype(str) + '-' + df_extract['destination'].astype(str)
df_extract['marital_hasChildren'] = df_extract['maritalStatus'].astype(str) + '-' + df_extract['has_children'].astype(str)
df_extract['temperature_weather'] = df_extract['temperature'].astype(str) + '-' + df_extract['weather'].astype(str)
```

```
df_extract = df_extract.drop(columns=['passanger', 'destination', 'maritalStatus', 'has_children', 'temperature','weather'])
```

passanger_destination	marital_hasChildren	temperature_weather
Alone-No Urgent Place	Unmarried partner-1	55-Sunny
Friend(s)-No Urgent Place	Unmarried partner-1	80-Sunny
Friend(s)-No Urgent Place	Unmarried partner-1	80-Sunny

Feature extraction dilakukan terhadap:

- fitur ‘passanger’, ‘destination’ menjadi ‘passanger_destination’,
- fitur ‘maritalStatus’, ‘has_children’ menjadi ‘marital_hasChildren’,
- fitur ‘temperature’, ‘weather’ menjadi ‘temperature_weather’

FEATURE ENGINEERING

FEATURE TAMBAHAN

1. Operating System GPS mobil/handphone

Operating system yang dimaksud disini adalah versi system update terbaru dari handphone dimana bisa mempengaruhi terhadap fitur GPS dan berpengaruh pada penerimaan kupon.

2. Design e-coupon

Design e-coupon yang menarik akan menambah keinginan para pengendara untuk menerima kupon tersebut.

3. Email user

Dengan mengetahui email user maka bisa dilakukan promo terhadap masing-masing user dan bisa diidentifikasi untuk lebih tepat dalam pemberian kupon sehingga kupon yang diterima juga akan semakin banyak.

4. Internet Service Provider

Perbedaan sinyal dan internet service provider akan berpengaruh pada kecepatan pembukaan GPS dan dengan berbedanya internet service provider maka promosi yang akan diberikan juga berbeda.