Data Pre-processing - Sest Consult

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Kelompok 5 :

Alfin Dwisatrio
Wildan Ryan
Dionisius Himando
Laurensia Vanida
Aldi Wachid Arifin
Kartika Novitasari

Dataset yang dipilih : Ecommerce Shipping Data
(https://www.kaggle.com/datasets/prachi13/customer-analytics)

Data Preprocessing

Rakamin Academy

Splitting Data

Sebelum melakukan data cleansing dan feature engineering, tim kami melakukan splitting data. Jumlah data trainingnya jadi **7.699** dari 10.999, sedangkan data testnya berjumlah **3.300**.

```
X = df[['Warehouse_block', 'Mode_of_Shipment', 'Customer_care_calls',
       'Customer_rating', 'Cost_of_the_Product', 'Prior_purchases',
       'Product_importance', 'Gender', 'Discount_offered', 'Weight_in_gms']]
y = df[['Reached.on.Time_Y.N']]

Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size=0.3, random_state=17)
```

```
print(df.shape)
print(Xtrain.shape)
print(Xtest.shape)
```

```
(10999, 12)
(7699, 10)
(3300, 10)
```

Data Cleansing

Rakamin Academy

Handle Missing Values

Untuk mengecek data yg kosong/hilang, kami menggunakan fungsi **.info()** dan **.isna().sum()**

```
Int64Index: 7699 entries, 6381 to 10863
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Warehouse_block      7699 non-null   object
 1   Mode_of_Shipment     7699 non-null   object
 2   Customer_care_calls  7699 non-null   int64
 3   Customer_rating      7699 non-null   int64
 4   Cost_of_the_Product  7699 non-null   int64
 5   Prior_purchases      7699 non-null   int64
 6   Product_importance   7699 non-null   object
 7   Gender               7699 non-null   object
 8   Discount_offered     7699 non-null   int64
 9   Weight_in_gms        7699 non-null   int64
dtypes: int64(6), object(4)
```

.info()

```
Warehouse_block      0
Mode_of_Shipment     0
Customer_care_calls  0
Customer_rating      0
Cost_of_the_Product  0
Prior_purchases      0
Product_importance   0
Gender               0
Discount_offered     0
Weight_in_gms        0
dtype: int64
```

.isna().sum()

Hasilnya tidak ada data yang yang bernilai kosong/hilang

Data Cleansing

Rakamin Academy

Handle Duplicate Data

Untuk mengecek apakah ada data yg duplikat di fitur data, kami menggunakan fungsi **.duplicated().sum()**

```
Xtrain.duplicated().sum()
```

```
0
```

Hasilnya tidak ada data yang sama/duplikat

Data Cleansing

Rakamin Academy

Handle Outliers

Di dalam data, ada 2 fitur yang memiliki outliers yaitu **Prior_purchases** dan **Discount_offered.** Untuk menghapus outlier di 2 fitur tersebut, kami menggunakan **IQR (Interquartile Range).** Berikut adalah fungsi yang kami gunakan :

```
print(f'Jumlah baris sebelum memfilter outlier: {len(Xtrain)}')

filtered_entries = np.array([True] * len(Xtrain))
for col in ['Prior_purchases','Discount_offered']:
    Q1 = Xtrain[col].quantile(0.25)
    Q3 = Xtrain[col].quantile(0.75)
    IQR = Q3 - Q1
    low_limit = Q1 - (IQR * 1.5)
    high_limit = Q3 + (IQR * 1.5)

    filtered_entries = ((Xtrain[col] >= low_limit) & (Xtrain[col] <= high_limit)) & filtered_entries

Xtrainout = Xtrain[filtered_entries]

print(f'Jumlah baris setelah memfilter outlier: {len(Xtrainout)}')
```

Hasil setelah outlier dihapus :

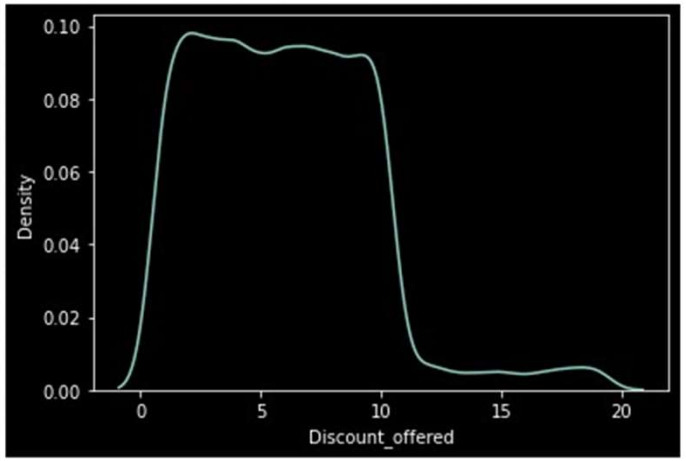
Jumlah baris sebelum memfilter outlier: 7699
Jumlah baris setelah memfilter outlier: 5557

Data Cleansing

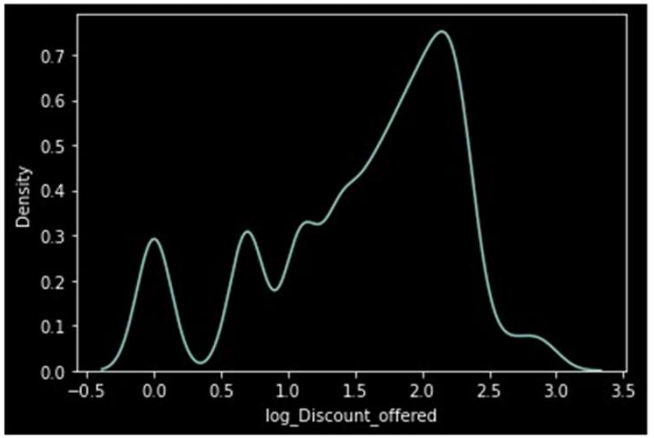
Rakamin Academy

Feature Transformation

Feature Transformation kami lakukan pada fitur **Discount_offered** yang distribusi datanya cenderung **right-skewed** dengan fungsi **np.log.** Setelahnya, kami menghapus fitur data Discount_offered dan menggunakan data setelah fungsi log digunakan.



BEFORE



AFTER

Hasilnya, distribusi fitur Discount_offered terlihat normal, tidak skewed lagi.

Data Cleansing

Rakamin Academy

Feature Encoding

Karena data bertipe kategorikal maka kami melakukan **Label Encoding** dan **One Hot Encoding (OHE)**.

Label Encoding digunakan pada fitur **Gender** dan **Product_Importance**, sedangkan OHE pada fiture **Mode_of Shipment**

Gender	Weight_in_gms	log_Discount_offered	enc_gender
M	1327	2.197225	0
F	1522	2.397895	1
M	4539	2.079442	0
F	4766	0.693147	1
M	5659	0.693147	0

Hasil label encoding pada fitur Gender. Male : 0, Female : 1

Data Cleansing

Rakamin Academy

Product_importance	Gender	Weight_in_gms	log_Discount_offered	enc_gender	enc_Product_importance
medium	M	1510	1.791759	0	1
medium	F	1113	2.302585	1	1
low	F	4742	2.079442	1	0
high	F	4260	1.791759	1	2
low	F	5219	2.079442	1	0

Hasil label encoding pada fitur Product Importance. Low : 0, Medium : 1, High : 2

Mode_of_Shipment	mode_Flight	mode_Road	mode_Ship
Ship	0	0	1
Ship	0	0	1
Flight	1	0	0
Flight	1	0	0
Ship	0	0	1

Hasil OHE pada fitur Mode_of Shipment.

Data Cleansing

Rakamin Academy

Handle Class Imbalance

Untuk mengecek adanya class imbalance, kami mengecek distribusi di fitur **Reached.on.Time_Y.N.**

Berdasarkan nilainya,
Not on time 4605 – **59, 81%**
On time 3094 – **40,19%**

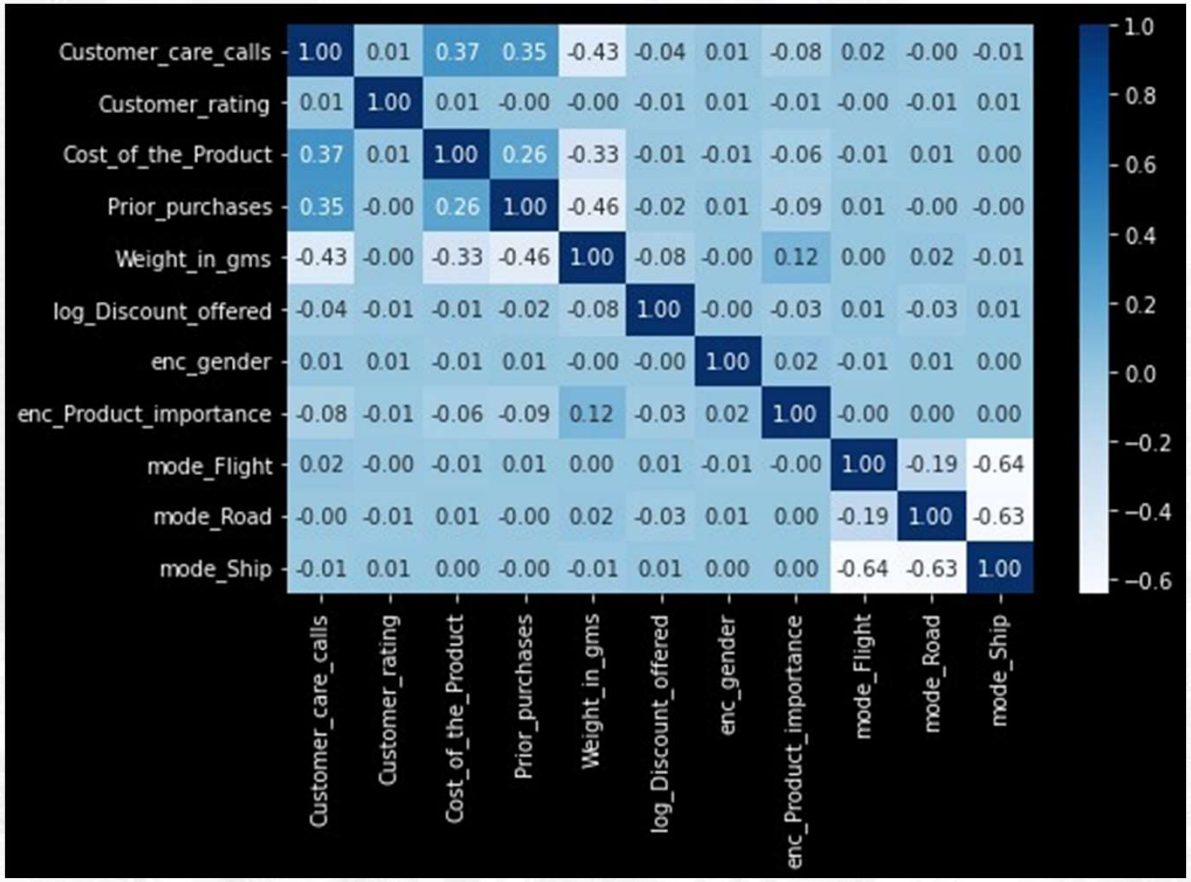
Berdasarkan derajat ketimpangan data, data kami tidak termasuk dalam kategori Mild / Moderate / Extreme. Maka, data kami tidak termasuk dalam kategori imbalance.

Feature Engineering

Feature Selection

Berdasarkan heatmap, **tidak ada fitur yang redundan, sehingga tidak perlu di drop**.

Maka, kami memutuskan untuk menggunakan semua fitur karena fitur yang ada di data kami tidak banyak.



Feature Engineering

Feature Extraction

Kami menambahkan fitur **weight_category** dari fitur **Weight_in_gms.** Kami mengelompokan berdasarkan nilai minimal dan maksimal dari fitur ini. Berikut adalah pembagian berdasarkan beratnya :

- 0-2000 gr : **Light**
- 2000-5000 gr : **Medium**
- di atas 5000 gr : **Heavy**

Berikut adalah contoh dari pembagian kategori berdasarkan beratnya

Weight_in_gms	weight_category
4967	medium
4432	medium
2381	medium
4808	medium
4867	medium
...	...
5331	heavy
4958	medium
1906	light
4010	medium
5345	heavy

```
medium    2277
heavy     1843
light     1437
Name: weight_category, dtype: int64
```

Setelah dilakukan pembagian kategori, mayoritas termasuk dalam **Medium** category.

Feature Engineering

Rakamin Academy

Feature Addition

Di dalam tahap ini, berikut adalah fitur-fitur yang kami bisa tambahkan untuk membantu maksimalisasi penggunaan data di dalam ecommerce :

1. **Shipment Date (Tanggal pengiriman barang)**
   Fitur ini bisa membantu customer dan perusahaan shipping untuk mengecek kapan barang dikirim, bisa juga untuk memberi estimasi barang akan sampai

1. **Revenue (Cost - Discount offer)**
   Fitur ini untuk melihat seberapa besar pendapatan ecommerce setelah memberikan diskon untuk customer,sehingga bisa digunakan untuk pertimbangan pemberian diskon di periode berikut

1. **Order Date (Tanggal pemesanan sekaligus pembayaran)**
   Fitur ini bisa membantu perusahan mengecek seberapa efektif dan efisien proses packing barang setelah customer melakukan pemesanan dan pembayaran.

1. **Membership status (status customer berdasarkan banyaknya transaksi)**
   Fitur ini untuk melihat loyalitas setiap customer berdasarkan pengulangan transaksinya. Perusahaan bisa memberikan reward untuk customer yang melakukan transaksi berkali2 di ecommerce.

Git

- GITHUB KELOMPOK 5 :

  https://github.com/aldiwachid/E-Commerce-Shipping-Data-.git