Rakamin Batch 19A Final Project -- Data Preprocessing

Group 1 Members :
Yosafat Respati
Ridho Fajar
Li'izza Diana M
Teguh Tri A
Vito Rihaldijiran
M Supian Noor
Rexy Anggala Putra

1. Data Cleansing

A. Handle missing values

Pada E Commerce Dataset, ditemukan 7 feature yang memiliki missing value, yaitu Tenure, WarehouseToHome, HourSpendOnApp, DaySinceLastOrder, OrderCount, CouponUsed, dan OrderAmountHikeFromlastYear. Dalam menghandle missing value tersebut, digunakan metode imputasi untuk mengisi missing value pada feature tersebut. Imputasi yang dipilih untuk mengisi semua semua feature tersebut adalah dengan median karena terdapat outlier pada feature tersebut dan distribusi pada feature Tenure, WarehouseToHome, DaySinceLastOrder, OrderCount, CouponUsed, dan OrderAmountHikeFromlastYear, merupakan distribusi right-skewed. Sedangkan, pada feature HourSpendOnApp tetap digunakan median walaupun data berdistribusi normal karena feature tersebut memiliki tipe data integer.

B. Handle duplicated data

Pada E Commerce Dataset tidak ditemukan data duplikat maka tidak diperlukan metode apapun untuk menghandle data duplikat.

C. Handle outliers

Pada tahap Handle Outliers, digunakan metode menghapus outlier berdasarkan IQR. Penghapusan outlier tersebut dilakukan pada kolom yang memiliki outlier, yaitu Tenure, WarehouseToHome, HourSpendOnApp, SatisfactionScore, CouponUsed, OrderCount, dan DaySinceLastOrder. Jumlah baris sebelum dilakukan handle outliers sebanyak 5630 kemudian setelah dilakukan handle outliers jumlah baris yang tersisa sebanyak 3827.

D. Feature transformation

Metode yang digunakan untuk feature transformation yaitu kombinasi antara metode MinMaxScaler dan StandarScaler. Untuk metode MinMaxScaler diimplementasikan pada feature kolom yang memiliki distribusi normal yaitu pada kolom HourSpendOnApp dan SatisfactionScore. Sedangkan untuk feature kolom yang memiliki distribusi positif skewed metode feature transformation yang diimplementasikan yaitu StandarScaler pada kolom numerikal selain kolom HourSpendOnApp dan SatisfactionScore (Tenure, WarehouseToHome, NumberedDeviceRegistered, NumberofAdress, OrderAmountHikeFromLastYear, CouponUsed, OrderCOunt, DaySinceLastOrder, dan CashbackAmount)

E. Feature encoding

Pada step feature encoding, kolom kategorikal Gender akan diproses menggunakan metode label encoding karena hanya memiliki 2 nilai. Sedangkan untuk kolom kategorikal PreferredLoginDevice, PreferredPaymentMode, PreferredOrderCat, dan MaritalStatus akan dihandle menggunakan metode One Hot Encoding dikarenakan memiliki lebih dari 2 value.

F. Handle class imbalance

Target nilai churn dan tidak churn pada kolom target memiliki perbandingan yang cukup timpang. Metode yang digunakan untuk mengatasi class imbalance yaitu dilakukan dengan menggunakan metode SMOTE dengan rasio SMOTE sebesar 0,9

2. Feature Engineering

A. Feature selection

Pada tahap ini, kami tidak menghapus suatu feature sebagai pertimbangan penggunaan semua feature pada tahap modelling

B. Feature extraction

Pada feature extraction ini kami menambahkan kolom avg_totalbelanja, aov dan gmv, dengan penjelasan sebegai berikut:

- avg_totalbelanja = rata-rata total uang belanja yang harus dibayarkan sebelum coupon/voucher digunakan
- aov (average order value)= rata-rata jumlah uang yang dibelanjakan setiap customer tiap bulan
- gmv (gross merchandise value)= total pembelian yg terjadi tiap bulan

Pada case ini kamu mengamsumsikan voucher/coupun yang diberikan ecommerce sebesar 10%