

# E-Commerce Customer Churn Predict



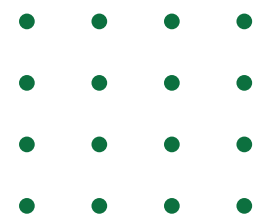
Oktar Mahardika





# Bussines Understanding

- Perusahaan ini adalah sebuah platform e-commerce yang melayani berbagai kategori produk
- Di tengah kompetisi industri e-commerce yang sangat ketat, perusahaan mulai menghadapi tantangan dalam mempertahankan pelanggannya
- Terdapat dua cara untuk mengatasi fenomena customer churn, yaitu Customer Acquisition dan Customer Retention



# Bussines Understanding

customer acquisition memiliki biaya 5 kali lipat lebih dari customer retention.

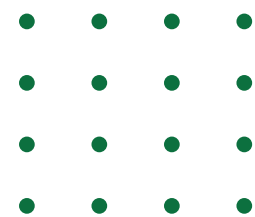
Kita asumsikan perusahaan per tahunnya menggelontorkan dana \$100.000 untuk maintain 1000 customer lama

- Retention Cost per Customer

$$\$100,000 \div 1,000 = \$100 \text{ per customer}$$

- Acquisition Cost per Customer

$$\text{Retention Cost per Customer} \times 5 = \$500 \text{ per customer}$$



# Problem

Perusahaan tidak memiliki sistem yang dapat mengidentifikasi pelanggan yang berisiko churn secara proaktif. Jika semua pelanggan diperlakukan sama tanpa segmentasi, maka strategi retensi menjadi tidak efisien dari segi biaya, waktu, dan sumber daya.





# Goal

Perusahaan ingin memiliki kemampuan untuk memprediksi kemungkinan seorang pelanggan akan churn, sehingga divisi pemasaran dapat mengambil tindakan preventif seperti memberikan penawaran khusus, diskon, atau kampanye personalisasi.





# Analitical Approach

- Melakukan analisis eksploratif pada data untuk memahami data lebih dalam.
- Melakukan feature engineering untuk meningkatkan kualitas input ke dalam model.
- Membangun model klasifikasi yang dapat memprediksi kemungkinan pelanggan churn.

# Metric Evaluation

## False Positive (FP)

- Model memprediksi pelanggan akan churn, padahal sebenarnya tidak.
- Estimasi kerugian: \$100 per pelanggan.

## False Negative (FN)

- Model memprediksi pelanggan akan tetap, padahal sebenarnya churn.
- Estimasi kerugian: \$500 per pelanggan.

Scoring : F2 Score




# Dataset

Kolom	Deskripsi
Tenure	Lama waktu (dalam bulan) pelanggan telah terdaftar.
WarehouseToHome	Jarak (kemungkinan dalam km) antara gudang dan rumah pelanggan.
NumberOfDeviceRegistered	Jumlah perangkat yang terdaftar di akun pelanggan.
PreferredOrderCat	Kategori produk yang paling sering dipesan pelanggan.
SatisfactionScore	Skor kepuasan pelanggan (kemungkinan skala 1–5).
MaritalStatus	Status pernikahan pelanggan (Single, Married, Divorced).
NumberOfAddress	Jumlah alamat yang terdaftar di akun pelanggan.
Complain	Apakah pelanggan pernah komplain (0 = tidak, 1 = ya).
DaySinceLastOrder	Jumlah hari sejak pemesanan terakhir oleh pelanggan.
CashbackAmount	Jumlah cashback yang diterima oleh pelanggan.
Churn	Apakah pelanggan berhenti menggunakan layanan (0 = tidak, 1 = ya).





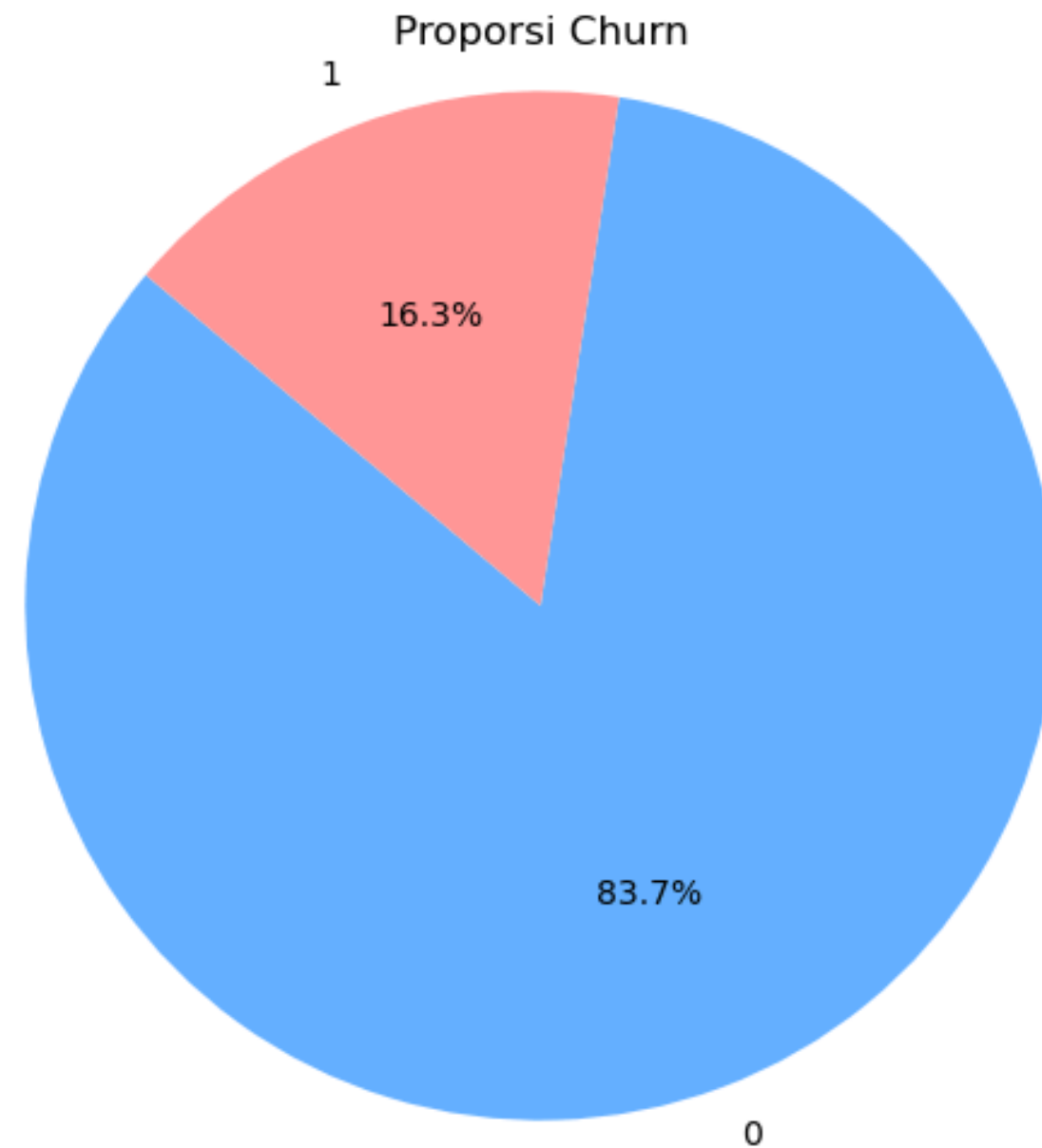
# Dataset

- Dataset terdiri dari 3941 baris dan 11 kolom
  - Mayoritas tipe data merupakan numerikal (Int/Float) namun terdapat 2 kolom kategorikal (Str/object)
  - Terdapat missing value pada kolom Tenure, WarehouseToHome, dan DaySinceLastOrder
  - Terdapat 671 data duplicate pada dataset
- 

# Exploratory Data Analysis (EDA)



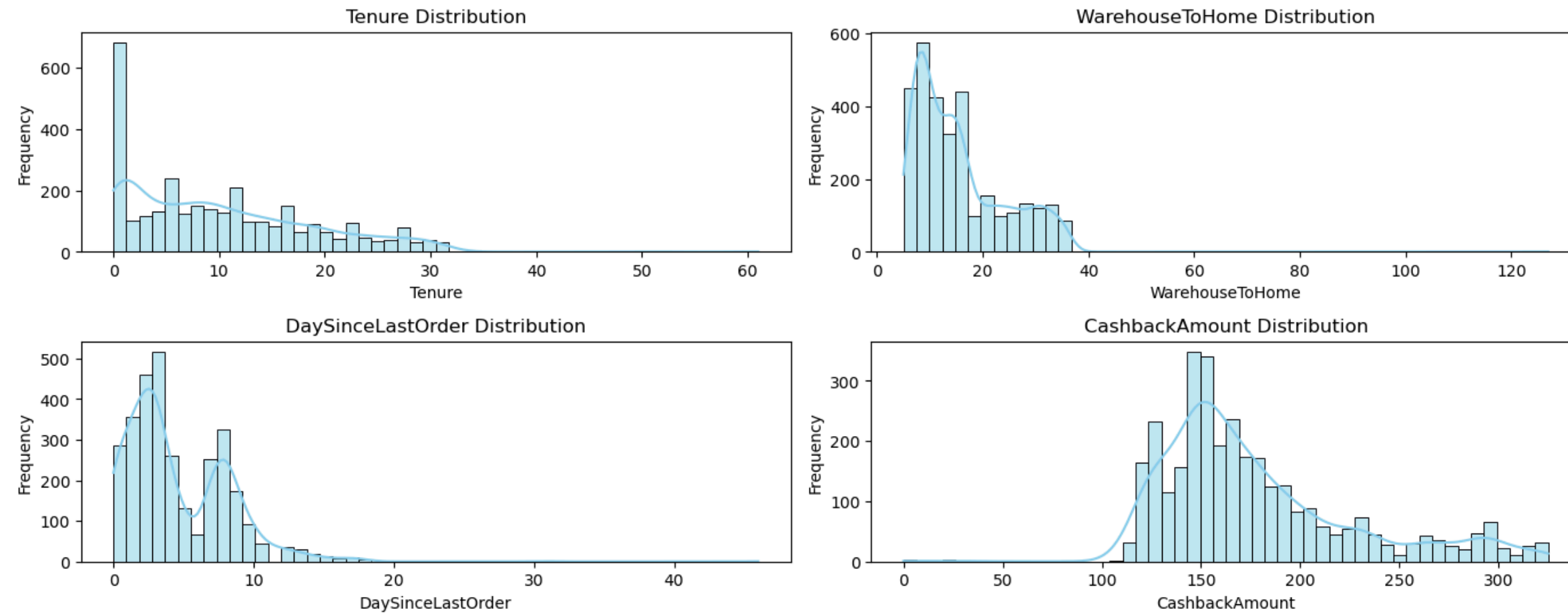
# Target Proportion



Churn (1)	534
Not Churn (0)	2736

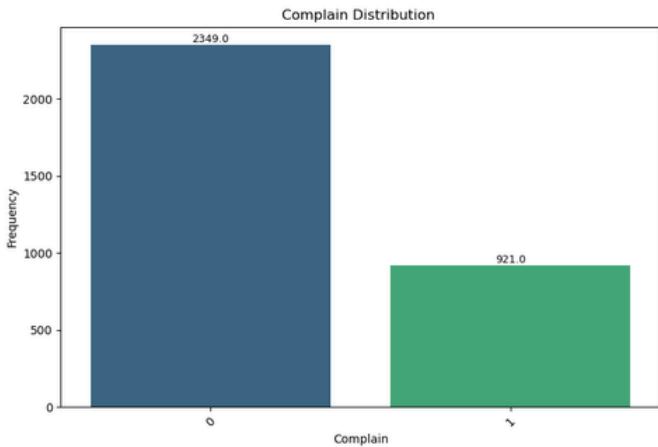
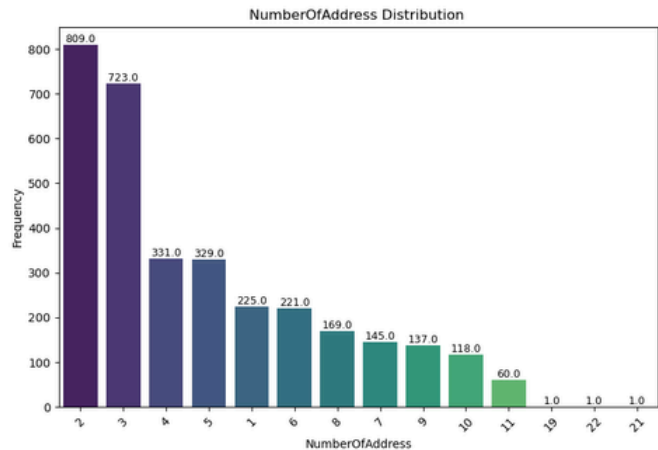
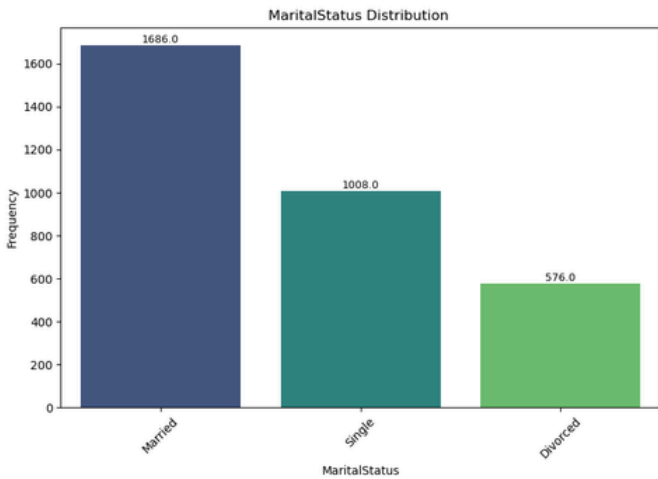
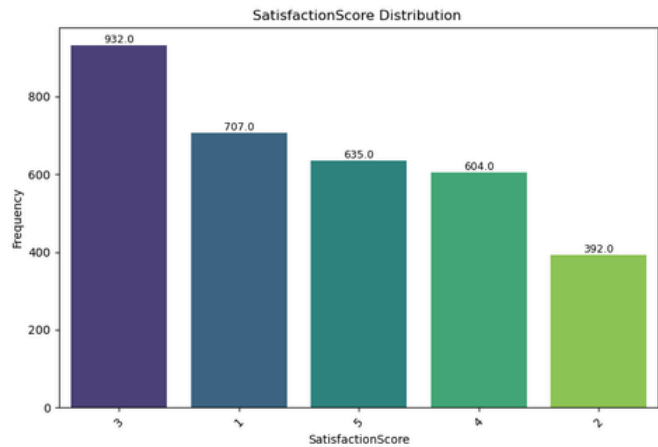
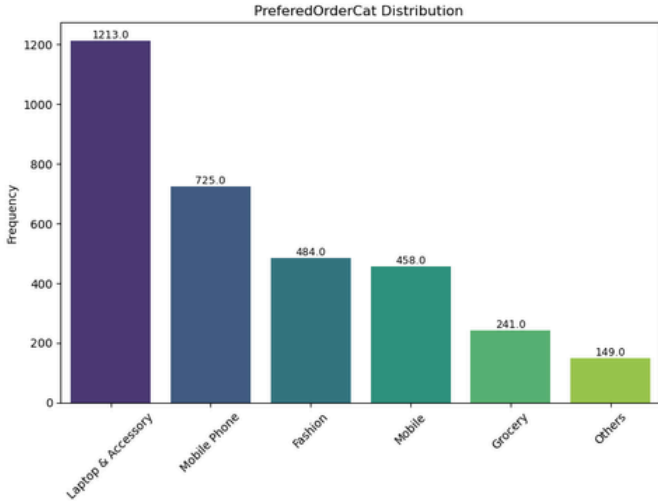
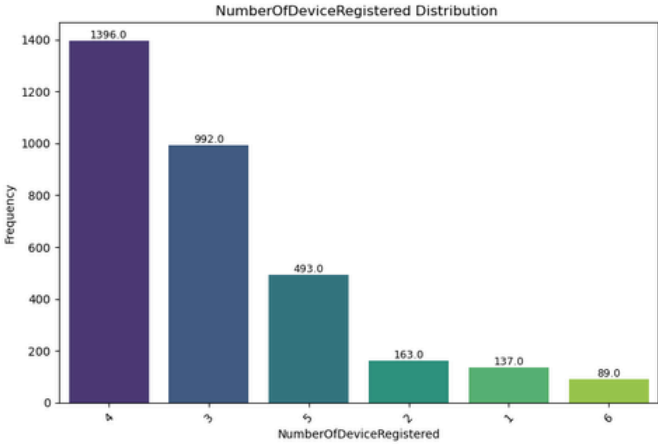
proporsi target tidak seimbang maka akan ditangani dengan Resampler

# Numerical Data Distribution



- Berdasarkan grafik hisplot pada data numerikal, terlihat seluruh data tidak terdistribusi normal.
- Hal ini juga dibuktikan pada uji statistik, dimana seluruh nilai P-Value kurang dari 0.05 yang berarti data tidak terdistribusi normal

# Categorical Data Distributions



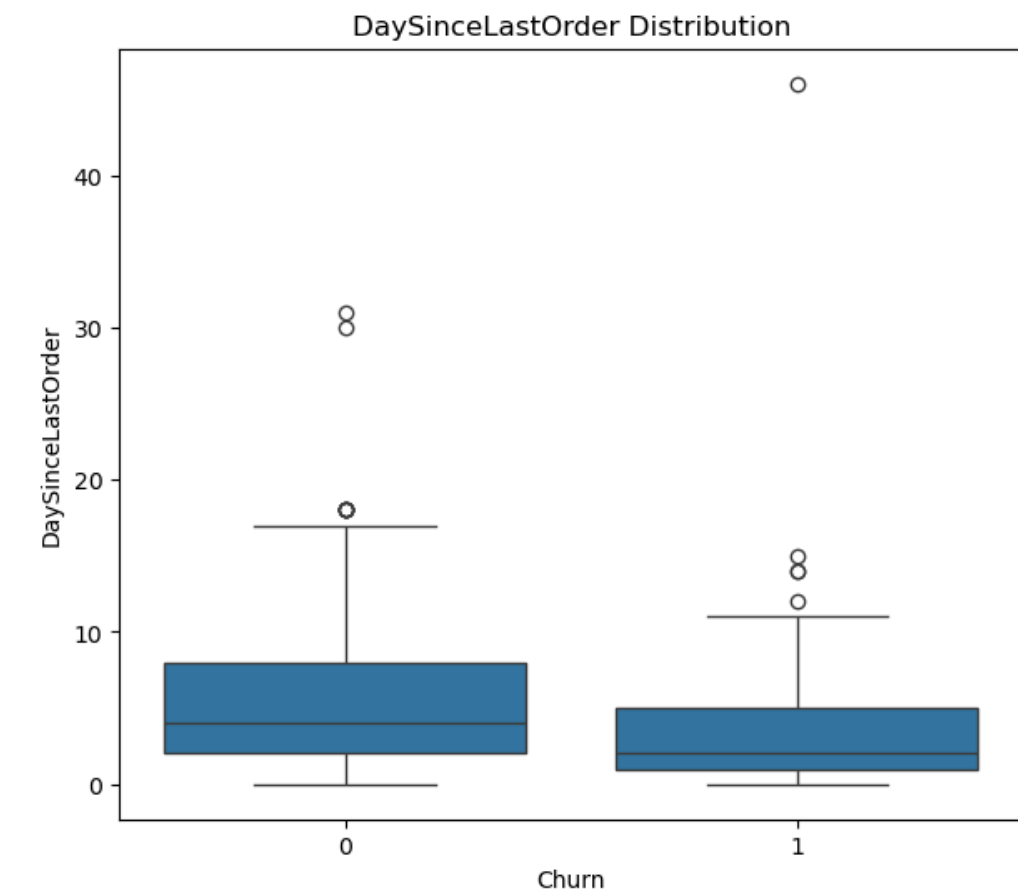
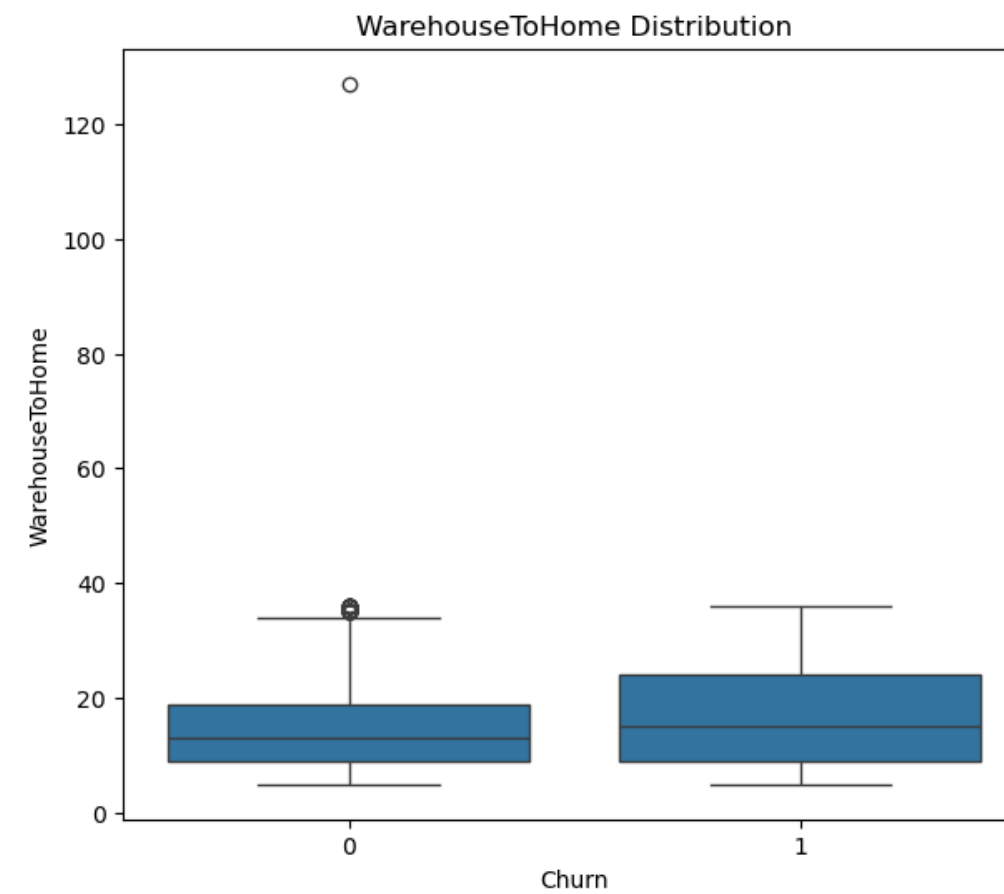
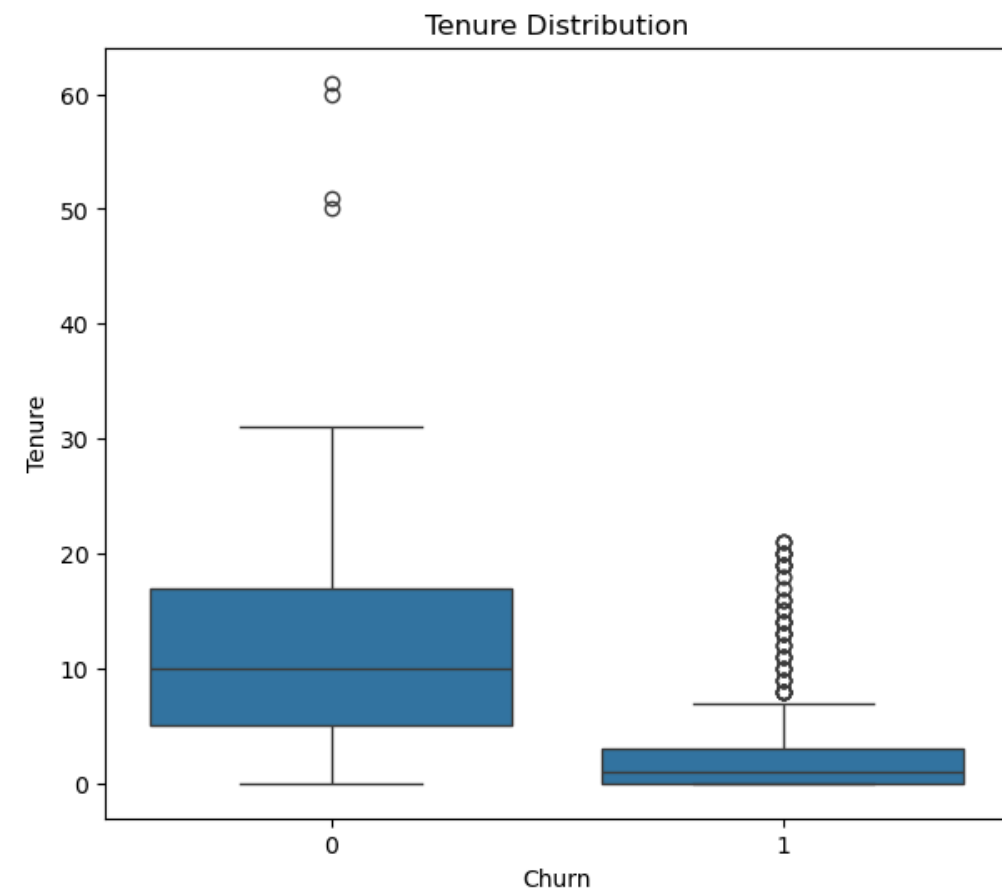
- Pada kolom NumberOfAddress, ditemukan bahwa terdapat orang dengan jumlah alamat 19, 20, dan 21 yang masing-masing hanya muncul sebanyak satu kali.
- Hal ini dapat mempengaruhi prediksi Machine Learning jadi akan dihapus saja







# Outlier



- Terdapat 3 kolom numerikal dengan outlier paling signifikan
- Outlier pada kolom tenure akan dihapus, sedangkan yang lainnya akan dibiarkan



# Machine Learning



# Define X, y and Splitting

- **Feature (X):** Tenure, WarehouseToHome, NumberOfDeviceRegistered, PreferredOrderCat, SatisfactionScore, MaritalStatus, NumberOfAddress, Complain, DaySinceLastOrder, CashbackAmount
- **Target (y) :** Churn
- Data di split dengan 80% data train serta 20% data test
- Kemudian digunakan stratify agar proporsi target dibagi menjadi rata



# Data Preprocessing

## **Imputer :**

- Iterative :Tenure, WarehouseToHome, DaySinceLastOrder

## **Encoding :**

- OneHot :PreferredOrderCat, MaritalStatus

## **Scaling :**

- Robust : NumberOfAddress, NumberOfDeviceRegistered, CashbackAmount, Tenure, WarehouseToHome, DaySinceLastOrder

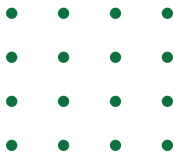
**pass** : SatisfactionScore, Complain



# Cross Validation

## Top 3 F2 Score

Model	Mean F2	Std
LGBMClassifier	0.666251	0.038162
XGBClassifier	0.665740	0.045977
DecisionTree	0.595214	0.053707





# Hyperparameter Tunning

## Best parameter

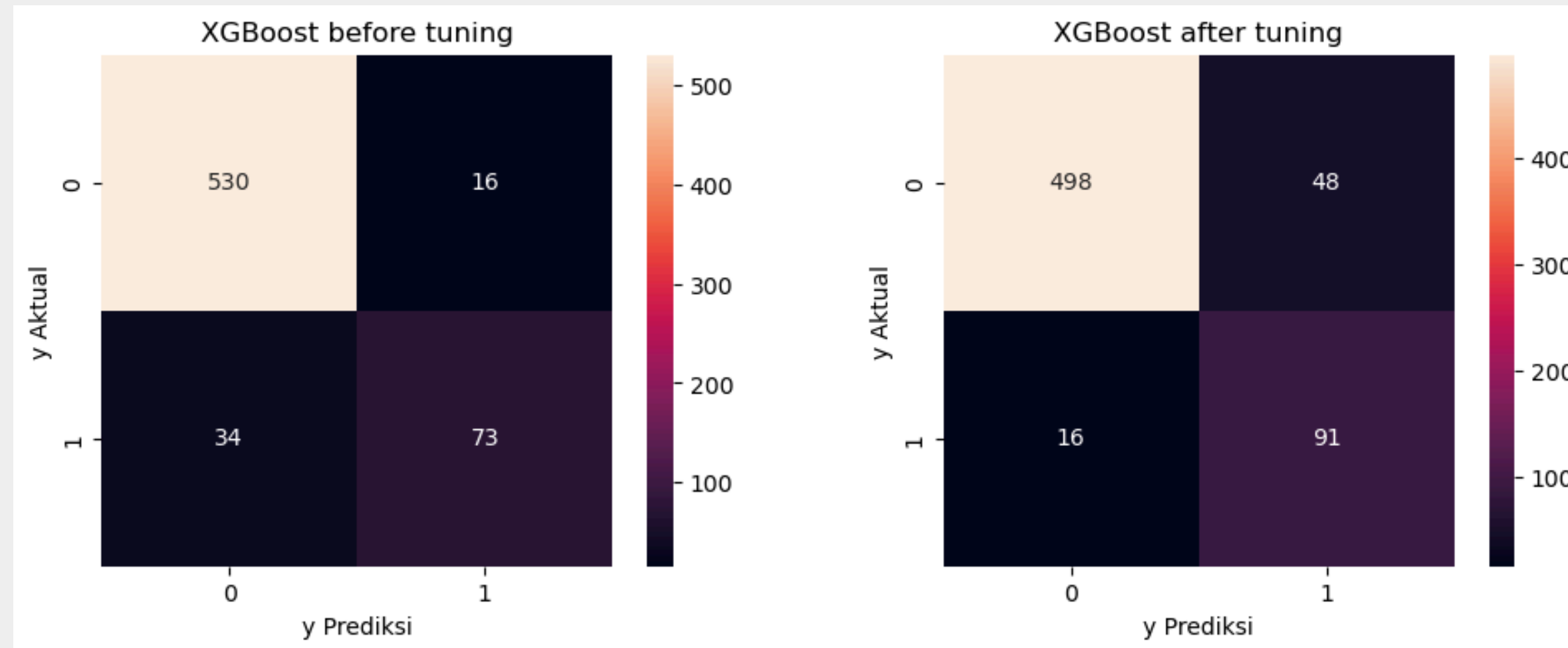
- {'modeling\_\_booster': 'gbtree',  
'modeling\_\_max\_depth': 3,  
'modeling\_\_n\_estimators': 150,  
'modeling\_\_scale\_pos\_weight': 4}
- **Training Score** : 0.7284

## Predict To Test :

- Score Before Tunning : 0.7059
- Score After Tunning : 0.8024



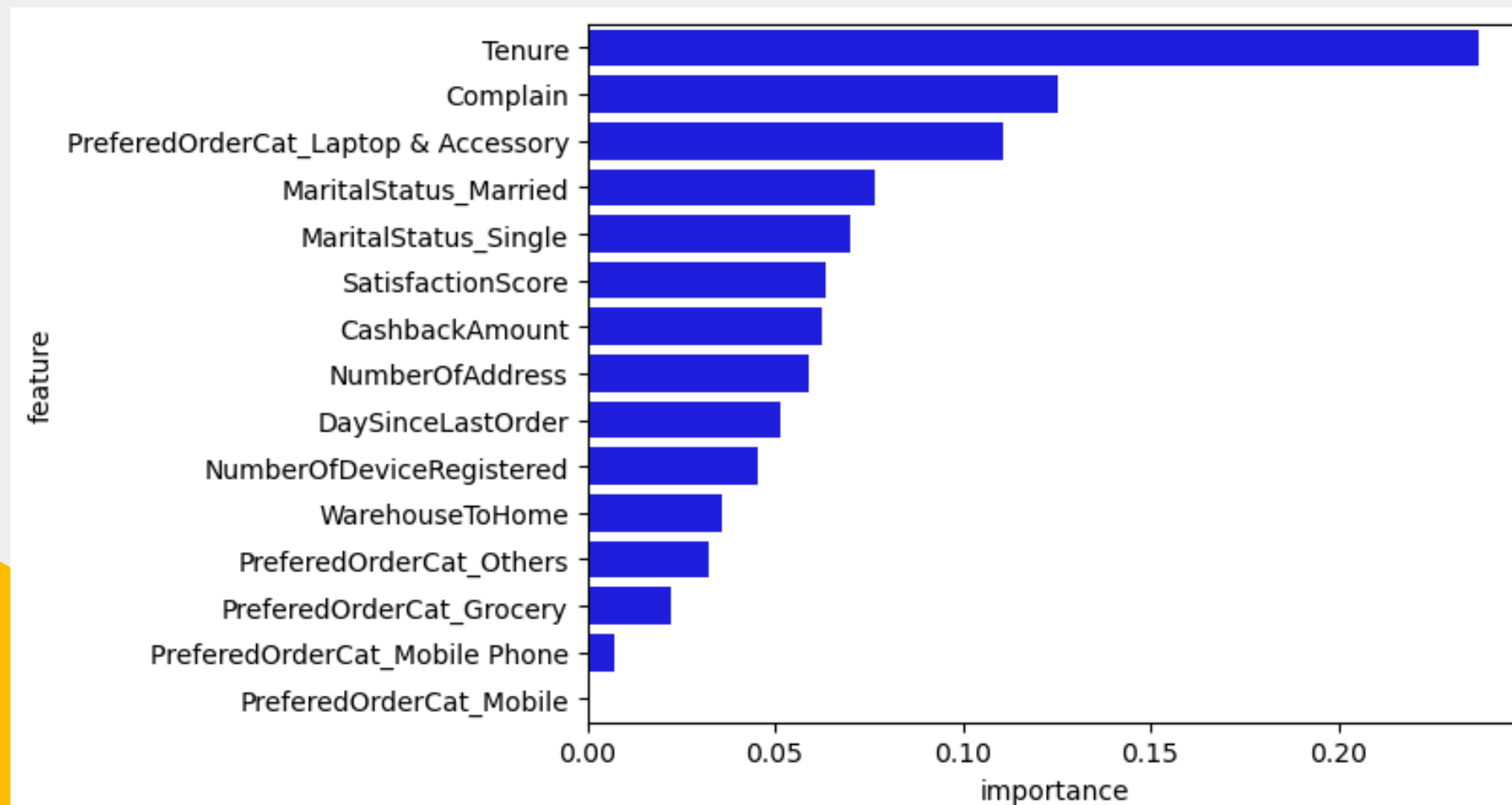
# Confusion Metric



- Pada Garfik confusion matrix terlihat model before tuning salah memprediksi FN sebanyak 34 Data sedangkan FP sebanyak 16 Data
- Sedangkan Pada Garfik confusion matrix terlihat model after tuning salah prediksi FN turun menjadi 16 data, sedangkan FP nya naik menjadi 48 Data



# Feature Importance



Terlihat bahwa feature yang paling berpengaruh adalah :

- Tenure,
- Complain
- PreferedOrderCat\_Laptop & Accessory.

Sedangkan feature yang lain tidak berpengaruh signifikan



# Streamlit Simulation



OR

[Click Here](#)

# Kesimpulan

Berdasarkan hasil metric evaluation Machine Learning berhasil memperoleh F2 Score sebesar 80% untuk memprediksi pelanggan Churn atau Not Churn dari data test.

F2 Score adalah metrik evaluasi yang mengutamakan recall lebih tinggi daripada precision (karena recall diberi bobot lebih besar).

Data test sendiri terdiri dari 653 data dimana secara metric Machine Learning salah memprediksi 16 Data FN, Serta 48 Data FN.





# Kesimpulan

## Tanpa ML

	Predict (0)	Predict (1)
Actual (0)	0	546
Actual (1)	0	107

Pengeluaran perusahaan untuk promosi

- $\$100 \times 653 = \$65,300$

Promosi yang tepat sasaran

- $\$100 \times 107 = \$10,700$

Biaya promosi sia-sia

- $\$65,300 - \$10,700 = \$54,600$

## Dengan ML

	Predict (0)	Predict (1)
Actual (0)	500	48
Actual (1)	16	91

salah promosi ke customer loyal (FP)

- $\$100 \times 46 = \$4,800$

Tidak terprediksi akan churn (FN)

- $\$500 \times 16 = \$8000$

Total Kerugian

- $\$4,800 + \$8000 = \$12,800$



# Kesimpulan

- Kerugian sebelum pakai ML: \$54,600
- Kerugian setelah pakai ML: \$12,800
- ML berhasil menurunkan kerugian perusahaan sebesar 75% -->  $(\$54,600 - \$12,600) / \$159,300$

# Rekomendasi

- Lebih baik data yang diambil dengan lebih teliti lagi agar tidak terdapat missing value serta membuat prediksi akan lebih akurat dengan data real
- Menambahkan feature-feature yang lain mungkin seperti jumlah order barang, umur, dll
- Menambahkan parameter yang lebih banyak pada bagian Hyperparameter Tuning.



The background features four isometric cubes in the corners, each composed of two overlapping hexagons in green and yellow. A 4x4 grid of small yellow dots is positioned to the left of the text.

# Terima Kasih