23/
09/
24.

E-COMMERCE CUSTOMER CHURN

ANALYSIS AND PREDICTION

M. RAHMAT HIDAYAT SYACHRUDIN

PRESENT

BUSINESS PROBLEM UNDERSTANDING

Dalam e-commerce , menjaga pelanggan sangat penting untuk kelangsungan dan pertumbuhan perusahaan. **Customer churn** mengancam profitabilitas karena kehilangan pelanggan berdampak pada pendapatan dan biaya akuisisi pelanggan baru. Strategi bisnis harus fokus pada retensi pelanggan melalui inovasi, personalisasi layanan, dan penawaran menarik untuk meningkatkan loyalitas.

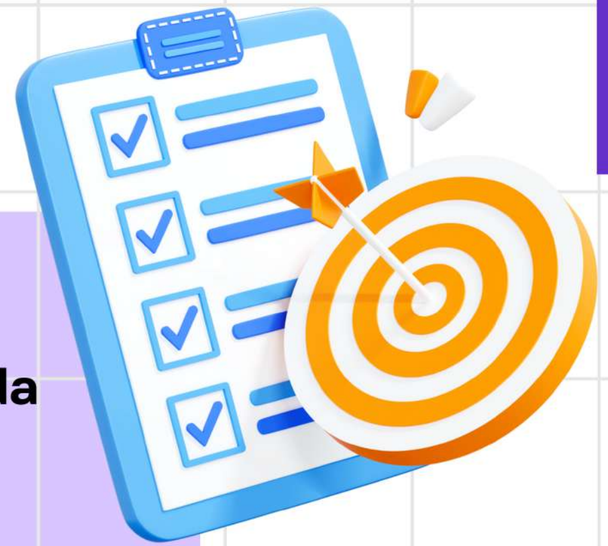
Analisis data churn membantu perusahaan mengenali pola perilaku pelanggan dan merumuskan langkah preventif seperti program loyalitas dan peningkatan layanan. Strategi ini dapat mengurangi churn, meningkatkan kepuasan pelanggan, dan mendukung keberhasilan jangka panjang di pasar e-commerce.

PROBLEM STATEMENT

Churn pelanggan adalah tantangan besar bagi bisnis e-commerce yang dapat mengancam pendapatan jangka panjang. Untuk itu, perusahaan perlu memprediksi pelanggan berisiko churn dan menawarkan promosi yang tepat.

GOALS

Tujuan analisis ini adalah mengembangkan model prediksi machine learning untuk mengidentifikasi pelanggan yang berpotensi churn. Dengan fokus pada pelanggan berisiko tinggi, perusahaan e-commerce dapat mengurangi kerugian pendapatan dan meningkatkan efektivitas biaya pemasaran.



ANALYTIC APPROACH

Actual

TN **True Negative (TN)**

pelanggan non-churn berhasil diklasifikasikan dengan benar

TP **True Positive (TP) :**

pelanggan churn dengan benar dari total pelanggan yang churn

FN **False Negative (FN):**

pelanggan churn yang salah diklasifikasikan sebagai non-churn

FP **False Positive (FP)**

pelanggan non-churn yang salah diklasifikasikan sebagai churn

Predicted

- **Recall** penting untuk memastikan semua pelanggan churn terdeteksi.
- **F1-Score** penting untuk menjaga keseimbangan antara menghindari kesalahan dan menangkap pelanggan churn dengan benar.

1. DATA UNDERSTANDING & EDA

1.1.1 Deskripsi Kolom

Kolom	Deskripsi
Tenure	Lama waktu pengguna telah berlangganan layanan
WarehouseToHome	Jarak antara gudang ke rumah pelanggan
NumberOfDeviceRegistered	Total device yang terdaftar pada satu akun pengguna
PreferedOrderCat	Kategori pesanan pilihan pelanggan dalam bulan lalu
SatisfactionScore	Skor yang menunjukkan tingkat kepuasan pelanggan terhadap layanan
MaritalStatus	Status pernikahan pelanggan
NumberOfAddress	Total alamat yang tercatat pada satu akun pengguna
Complaint	Setiap keluhan yang telah diajukan dalam bulan lalu
DaySinceLastOrder	jumlah hari yang telah berlalu sejak pesanan terakhir dilakukan oleh pelanggan
CashbackAmount	Rata-rata cashback dalam bulan lalu
Churn	Indikator apakah pengguna berhenti menggunakan layanan (1 untuk churn, 0 untuk tidak churn)

uji normalitas

	Distribusi Data
Tenure	Tidak Terdistribusi Normal
WarehouseToHome	Tidak Terdistribusi Normal
NumberOfDeviceRegistered	Tidak Terdistribusi Normal
SatisfactionScore	Tidak Terdistribusi Normal
NumberOfAddress	Tidak Terdistribusi Normal
Complain	Tidak Terdistribusi Normal
DaySinceLastOrder	Tidak Terdistribusi Normal
CashbackAmount	Tidak Terdistribusi Normal

informasi kolom

- Dalam Dataset ini terdiri dari 3941 baris dan 11 kolom
- Kolom terdiri dari 10 fitur dan 1 target
- Satu Baris di dataset menginterpretasikan 1 pelanggan

DATA CLEANING

	Features	DataType	Data Count	Missing value	Missing value %	Unique	UniqueSample
0	Tenure	float64	3941	194	4.92	36	[15.0, 7.0, 27.0, 20.0, 30.0, 1.0, 11.0, 17.0,...
1	WarehouseToHome	float64	3941	169	4.29	33	[29.0, 25.0, 13.0, 15.0, 16.0, 11.0, 12.0, 7.0...
2	NumberOfDeviceRegistered	int64	3941	0	0.00	6	[4, 3, 6, 2, 5, 1]
3	PreferedOrderCat	object	3941	0	0.00	6	[Laptop & Accessory, Mobile, Fashion, Others, ...
4	SatisfactionScore	int64	3941	0	0.00	5	[3, 1, 4, 2, 5]
5	MaritalStatus	object	3941	0	0.00	3	[Single, Married, Divorced]
6	NumberOfAddress	int64	3941	0	0.00	14	[2, 5, 7, 8, 3, 1, 9, 4, 10, 11, 6, 19, 22, 21]
7	Complain	int64	3941	0	0.00	2	[0, 1]
8	DaySinceLastOrder	float64	3941	213	5.40	22	[7.0, nan, 8.0, 11.0, 2.0, 1.0, 4.0, 3.0, 6.0,...
9	CashbackAmount	float64	3941	0	0.00	2335	[143.32, 129.29, 168.54, 230.27, 322.17, 152.8...
10	Churn	int64	3941	0	0.00	2	[0, 1]

```
raw_data.duplicated().sum()
```

✓ 0.0s

```
571
```

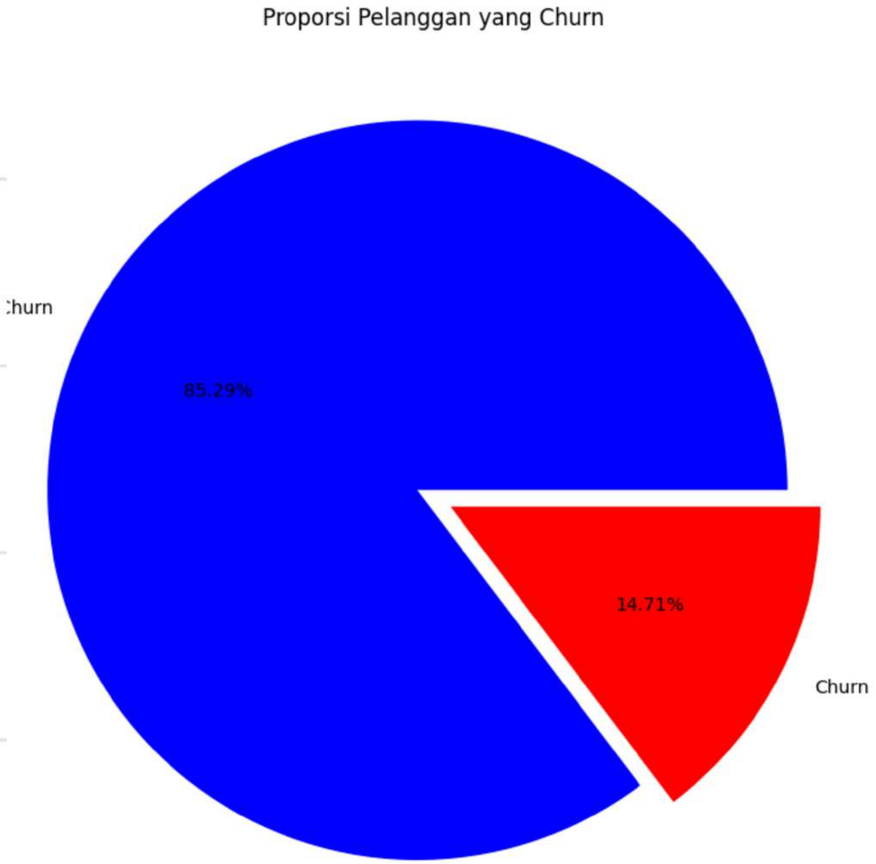
dataset di drop_duplicated memastikan setiap baris dalam dataset unik, merepresentasikan informasi yang berbeda tanpa pengulangan yang tidak diinginkan.

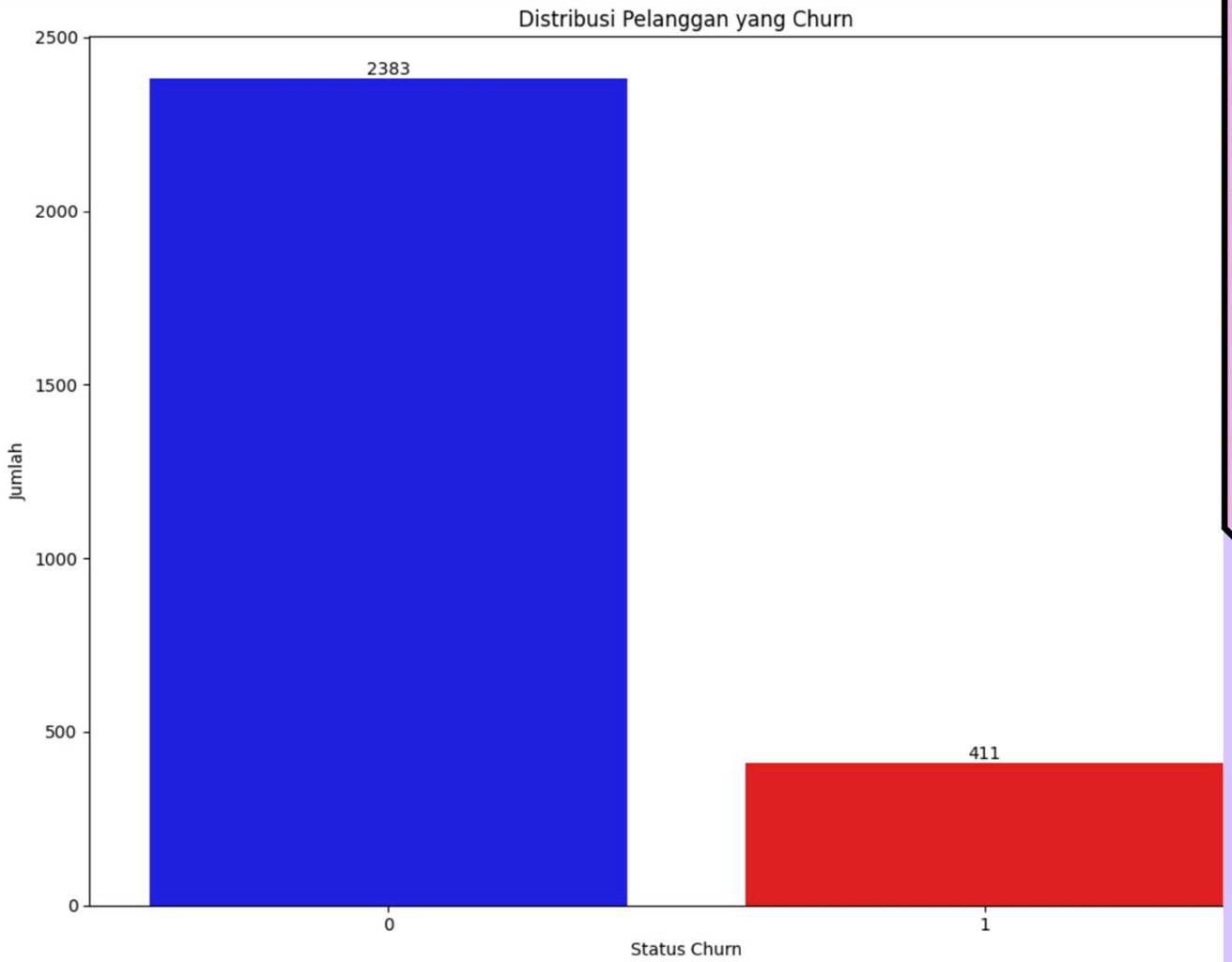
Karena nilai Missing value cukup kecil (dibawah 6%), maka akan di drop setiap baris yang mengandung Missing value.

	Column	Total Outliers	Percentage Outliers(%)	Lower Bound	Upper Bound
0	Tenure	4	0.14	-16.500	35.500
1	WarehouseToHome	1	0.04	-9.000	39.000
2	NumberOfDeviceRegistered	194	6.94	1.500	5.500
3	SatisfactionScore	0	0.00	-1.000	7.000
4	NumberOfAddress	2	0.07	-4.000	12.000
5	Complain	0	0.00	-1.500	2.500
6	DaySinceLastOrder	9	0.32	-7.000	17.000
7	CashbackAmount	340	12.17	81.825	259.585
8	Churn	411	14.71	0.000	0.000

Outliers akan dipertahankan karena mereka mengandung informasi penting yang dapat mempengaruhi prediksi churn dan akan dioptimalkan melalui teknik preprocessing untuk memastikan model dapat menangani variasi data dengan baik.

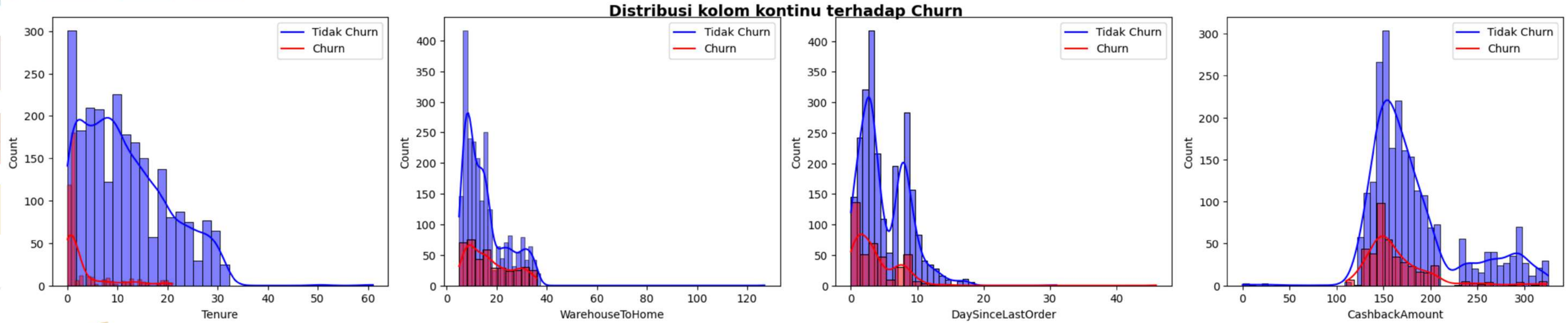
TARGET





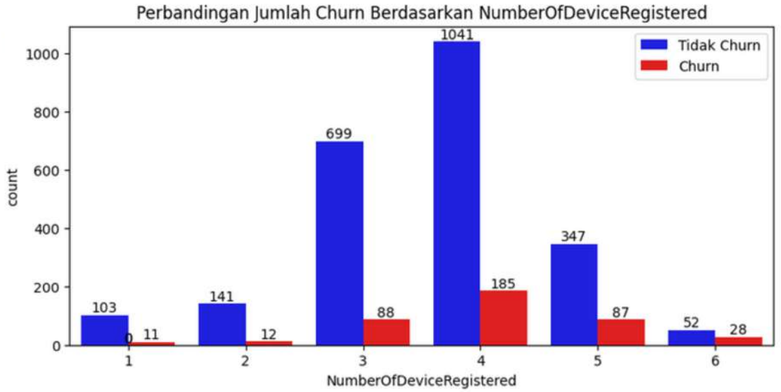
Visualisasi menunjukkan **Data Imbalance dengan dominasi pelanggan yang tidak churn: 411 pelanggan churn (14,71%) dan 2383 pelanggan tidak churn (85,29%). Imbalance ini dapat mempengaruhi model prediksi, terutama dalam mengidentifikasi kelas minoritas.**

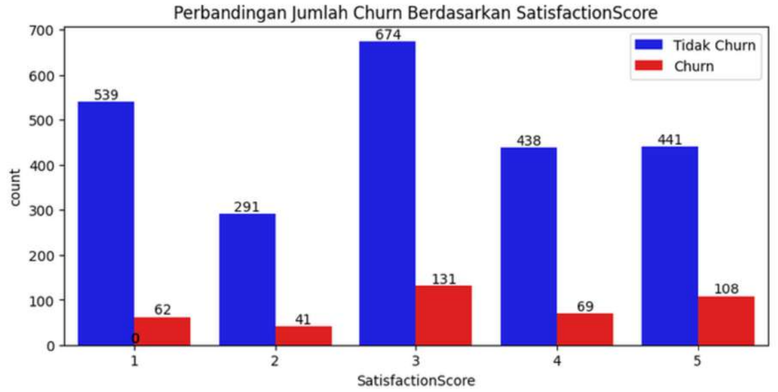
FEATURE VS TARGET

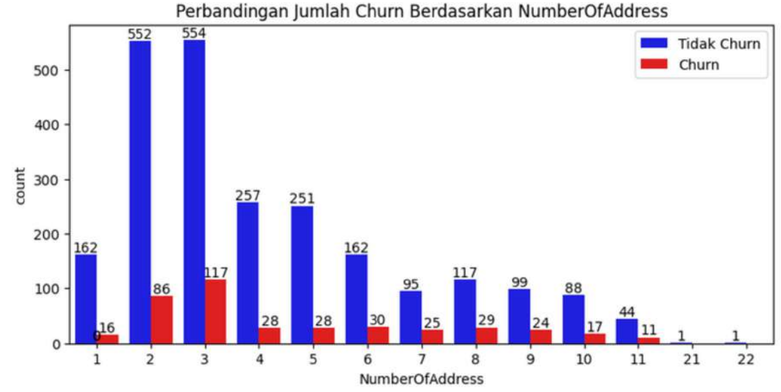


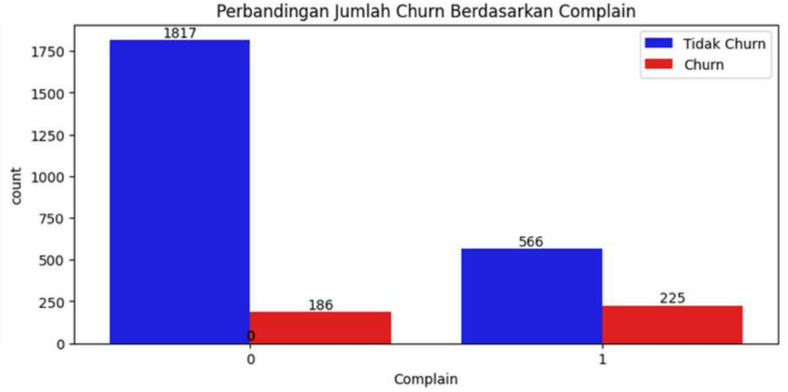
Grafik menunjukkan bahwa Tenure, DaySinceLastOrder, WarehouseToHome, dan CashbackAmount memengaruhi churn, dengan pelanggan baru atau yang lama tidak bertransaksi serta yang memiliki jarak pengiriman lebih jauh lebih cenderung churn, sementara cashback yang lebih tinggi membantu mempertahankan pelanggan.

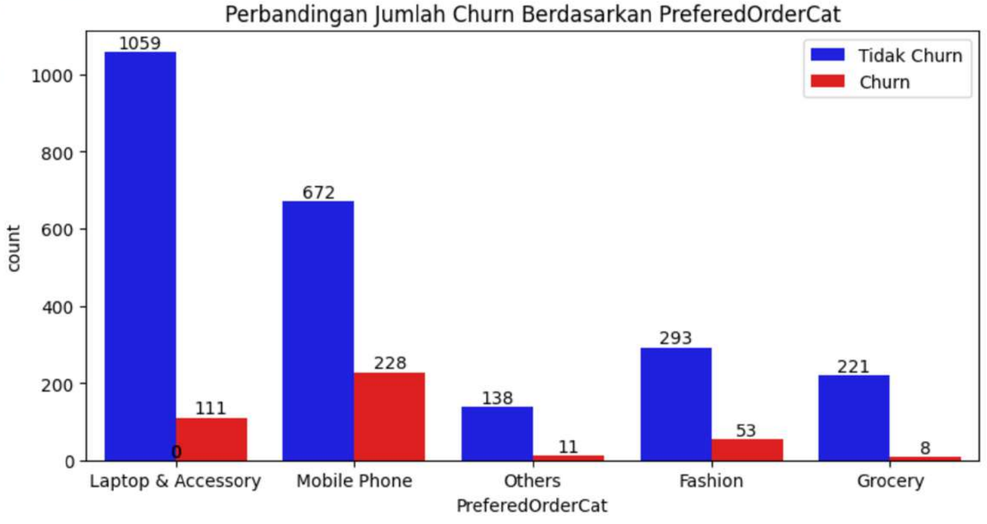
Dari grafik , terlihat bahwa pelanggan dengan banyak perangkat terdaftar, skor kepuasan rendah, keluhan, dan status pernikahan single lebih cenderung churn, sementara kategori produk dan jumlah alamat juga mempengaruhi keputusan churn.

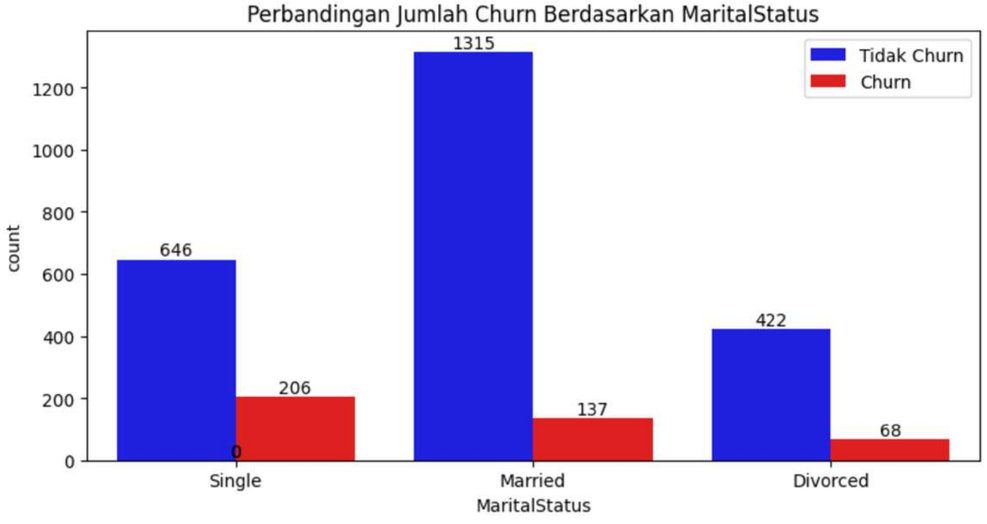












MACHINE LEARNING MODELING

	model	mean_f1_with_SMOTETomek	mean_f1_without_sampling	std_f1_with_SMOTETomek	std_f1_without_sampling
4	GradienBoost	0.901247	0.902498	0.003869	0.004767
5	XGBoost	0.892332	0.897308	0.005255	0.005124
3	Random Forest	0.818687	0.817988	0.008415	0.006343
2	Decision Tree	0.818417	0.809768	0.007580	0.006163
0	Logistic Regression	0.805371	0.805386	0.012370	0.013030
1	KNN	0.766097	0.767524	0.010694	0.009662

Model GradientBoost dan XGBoost menunjukkan performa terbaik dalam memprediksi churn dengan mean F1 score tertinggi dan stabilitas yang kuat, terutama setelah data di-scale menggunakan MinMaxScaler dan diatasi ketidakseimbangannya dengan SMOTETomek.

MACHINE LEARNING MODELING

01 Hyperparameter Tuning

```
# Hyperparameter space untuk GridSearchCV
hyperparam_space = [
    {'model': [xgb],
     'model__learning_rate': [0.05, 0.1, 0.25],
     'model__n_estimators': [50, 100, 200],
     'model__max_depth': [3, 4, 5]},

    {'model': [gbc],
     'model__learning_rate': [0.05, 0.1, 0.25],
     'model__n_estimators': [50, 100, 200],
     'model__max_depth': [3, 4, 5]}
]
```

Ini bertujuan untuk meningkatkan performa model dengan mengoptimalkan hyperparameter

02 Best Parameters from XGBoostClassifier

```
# Tampilkan hasil terbaik Best_Model
grid.best_params_, grid.best_score_
```

✓ 0.0s

```
({'model': XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.1, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=3, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, random_state=42, ...),
  'model__learning_rate': 0.25,
  'model__max_depth': 4,
  'model__n_estimators': 200},
 0.9441092580484977)
```

Hasil tuning XGBoostClassifier, model terbaik ditemukan dengan parameter sebagai berikut:

- **learning_rate: 0.25**
- **max_depth: 4**
- **n_estimators: 200**

03 Before Tunning

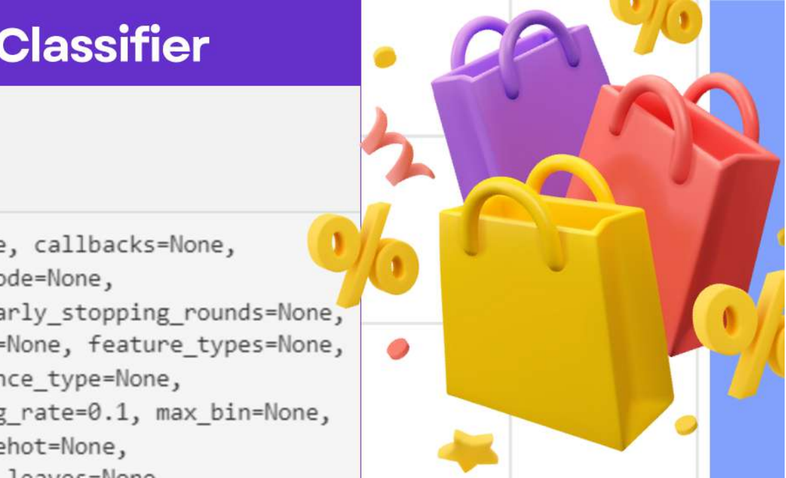
Classification Report Default XGBClassifier:

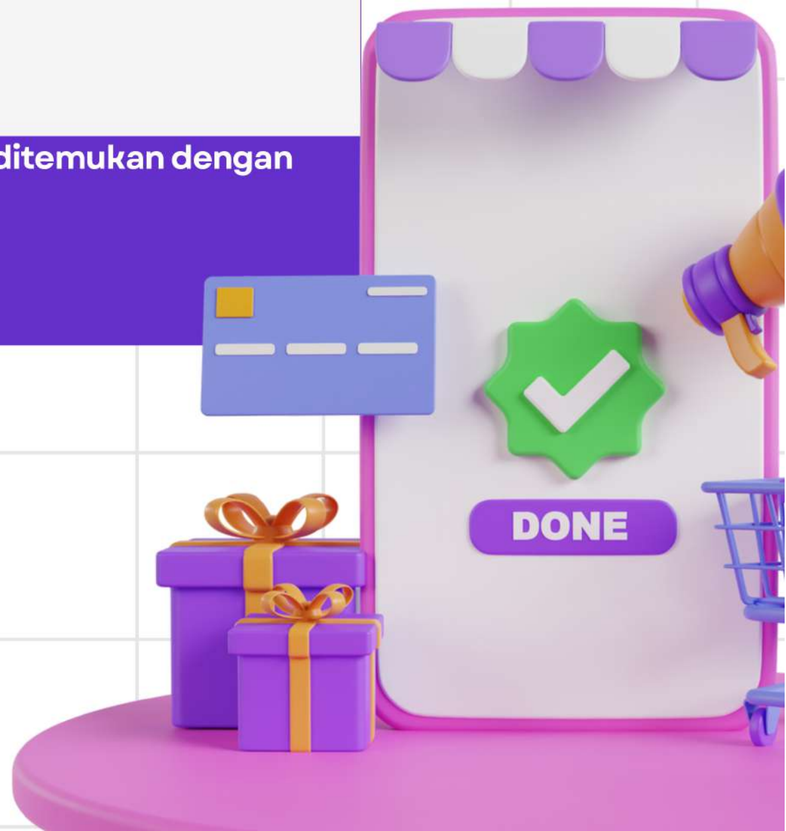
	precision	recall	f1-score	support
0	0.88	0.90	0.89	711
1	0.90	0.87	0.89	715
accuracy			0.89	1426
macro avg	0.89	0.89	0.89	1426
weighted avg	0.89	0.89	0.89	1426

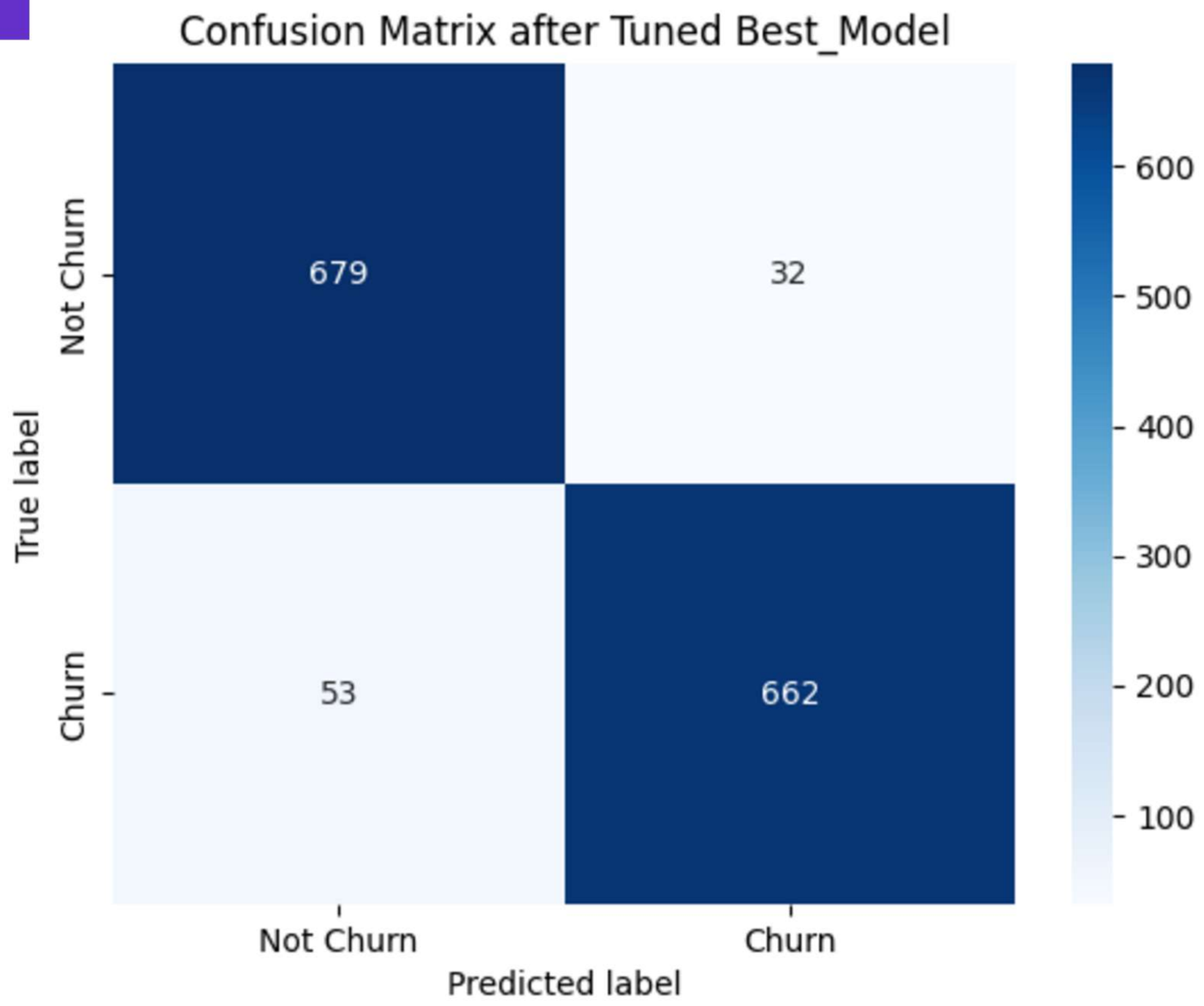
04 After Tunning

Classification Report Tuned Best_Model:

	precision	recall	f1-score	support
0	0.93	0.95	0.94	711
1	0.95	0.93	0.94	715
accuracy			0.94	1426
macro avg	0.94	0.94	0.94	1426
weighted avg	0.94	0.94	0.94	1426



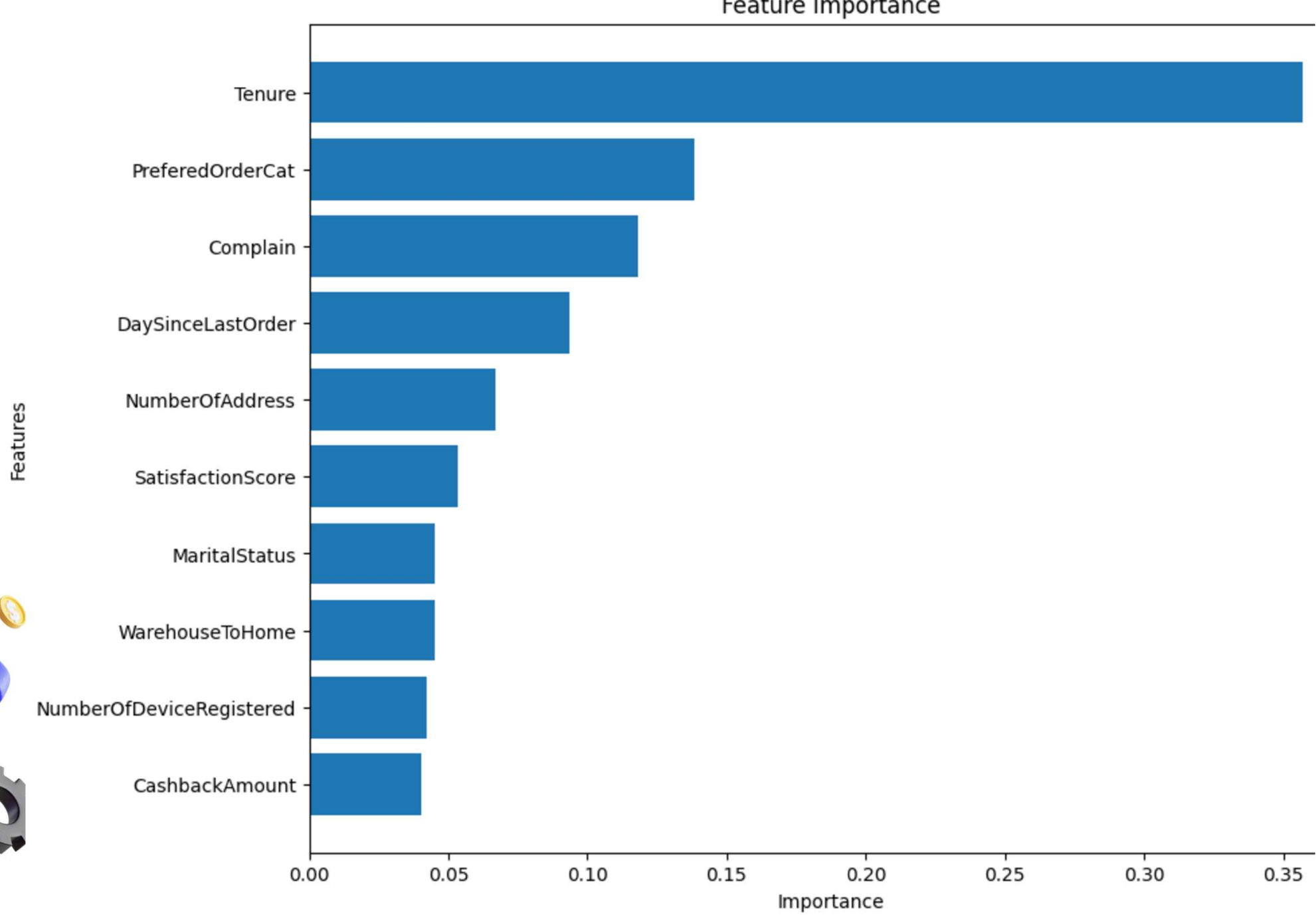




CONFUSION MATRIX AFTER TUNED BEST_MODEL

- **True Positive (TP) : 662 pelanggan churn dengan benar dari total pelanggan yang churn**
- **True Negative (TN) : 679 pelanggan non-churn berhasil diklasifikasikan dengan benar**
- **False Positive (FP): 32 pelanggan non-churn yang salah diklasifikasikan sebagai churn**
- **False Negative (FN): 53 pelanggan churn yang salah diklasifikasikan sebagai non-churn**

FEATURE IMPORTANCE



PERAN MACHINE LEARNING

Tanpa ML :

- Promosi untuk seluruh pelanggan: 1.426 pelanggan x 50 USD = 71.300 USD
- Biaya tepat sasaran untuk pelanggan churn: 715 pelanggan x 250 USD = 178.750 USD

- Total kerugian: 71.300 - 178.750 = - 107.450 USD

Dengan ML :

- FP (False Positive): 654 pelanggan x 50 USD = 32.700 USD
- TP (True Positive): 1.426 pelanggan x 50 USD = 71.300 USD
- FN (False Negative): 53 pelanggan x 250 USD = 13.250 USD
- Total biaya dengan ML: 32.700 + 71.300 + 13.250 = 117.250 USD
- Promosi tepat sasaran: 662 pelanggan x 250 USD = 165.500 USD

- Total Kerugian : 177.250 USD - 165.500 USD = - 11.750 USD

cost_saving = [((biaya_tanpa_ml - biaya_dengan_ml) /biaya_tanpa_ml)] * 100 %
cost_saving = [((-107.450 - (-11.750))/-107.450] * 100 %
cost_saving = 89.04 %

Cost saving-nya adalah 89.04%, yang berarti dengan menggunakan Machine Learning, perusahaan menghemat sekitar 89.04% dari kerugian yang seharusnya terjadi tanpa ML.

KESIMPULAN DAN REKOMENDASI

1. Fokus pada Retensi Pelanggan Baru
   - Prioritaskan diskon berkelanjutan dan program loyalitas untuk pelanggan dengan tenure rendah.
2. Tingkatkan Penanganan Keluhan
   - Sediakan solusi cepat dan efektif untuk mengurangi churn dari pelanggan yang mengajukan keluhan.
3. Optimalkan Pengalaman Berbelanja
   - Fokus pada peningkatan penawaran dan layanan di kategori produk utama seperti Laptop & Accessories dan Mobile Phone.
4. Retargeting Pelanggan Tidak Aktif
   - Lakukan retargeting pada pelanggan yang tidak aktif dengan email marketing atau penawaran khusus.
5. Perbaiki Layanan Pengiriman
   - Tingkatkan efisiensi pengiriman dan berikan opsi pengiriman lebih cepat untuk pelanggan dengan jarak jauh.

THANK YOU!