



Machine Learning Project

Lakü: E-Commerce Churn Prediction

By: Serafino

About Us

Lakü merupakan sebuah perusahaan yang berfokus dalam dunia bisnis digital, yang menyediakan layanan E-Commerce.

Lakü sendiri merupakan E-Commerce B2C (Business to Customer).



Business

E-Commerce sendiri adalah penyebaran, pembelian, penjualan, dan pemasaran barang maupun jasa lewat media elektronik berupa internet atau jaringan komputer lainnya. Salah satu fokus utama dari sebuah E-Commerce adalah menarik banyak pelanggan baru.

Business Problem

1. Memberikan reward khusus untuk pelanggan setia.
2. Memberikan penawaran khusus seperti discount/promo untuk pelanggan yang terdeteksi Churn/ berhenti menggunakan E-Commerceny.



Discount

- Pelanggan Loyal : Rp 30.000
- Pelanggan Churn : Rp 70.000
- Kerugian Churn : Rp 700.000



Matrix

	Predicted (0)	Predicted (1)
Actual (0)	547	0
Actual (1)	107	0

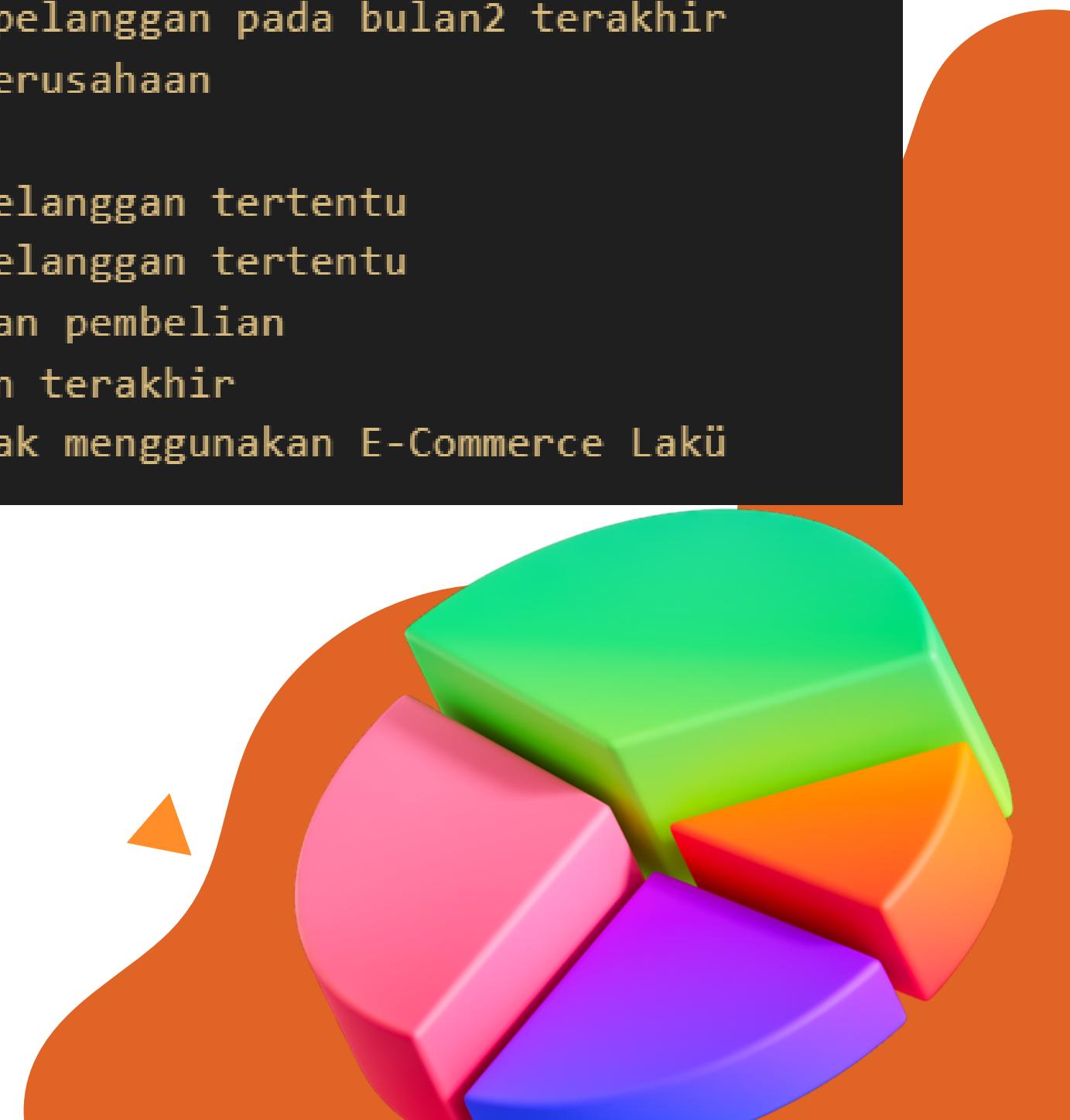
FP/False Positif: Machine Learning akan memprediksi pelanggan akan berhenti, namun pada kenyataannya pelanggan tetap loyal menggunakan E-Commerce Lakü. Jika seperti ini maka perusahaan akan mengeluarkan budget berlebih untuk pelanggan yang loyal.

FN/False Negatif: Machine Learning akan memprediksi pelanggan loyal, namun pada kenyataannya pelanggan akan berhenti menggunakan E-Commerce Lakü. Jika seperti ini maka perusahaan akan mengeluarkan budget yang lebih hemat untuk pelanggan yang akan berhenti. Dengan resiko mencari pelanggan baru yang biayanya kita asumsikan saja berdasarkan pemahaman Pak Kevin, lebih mahal 10x dari biaya memberikan penawaran khusus.

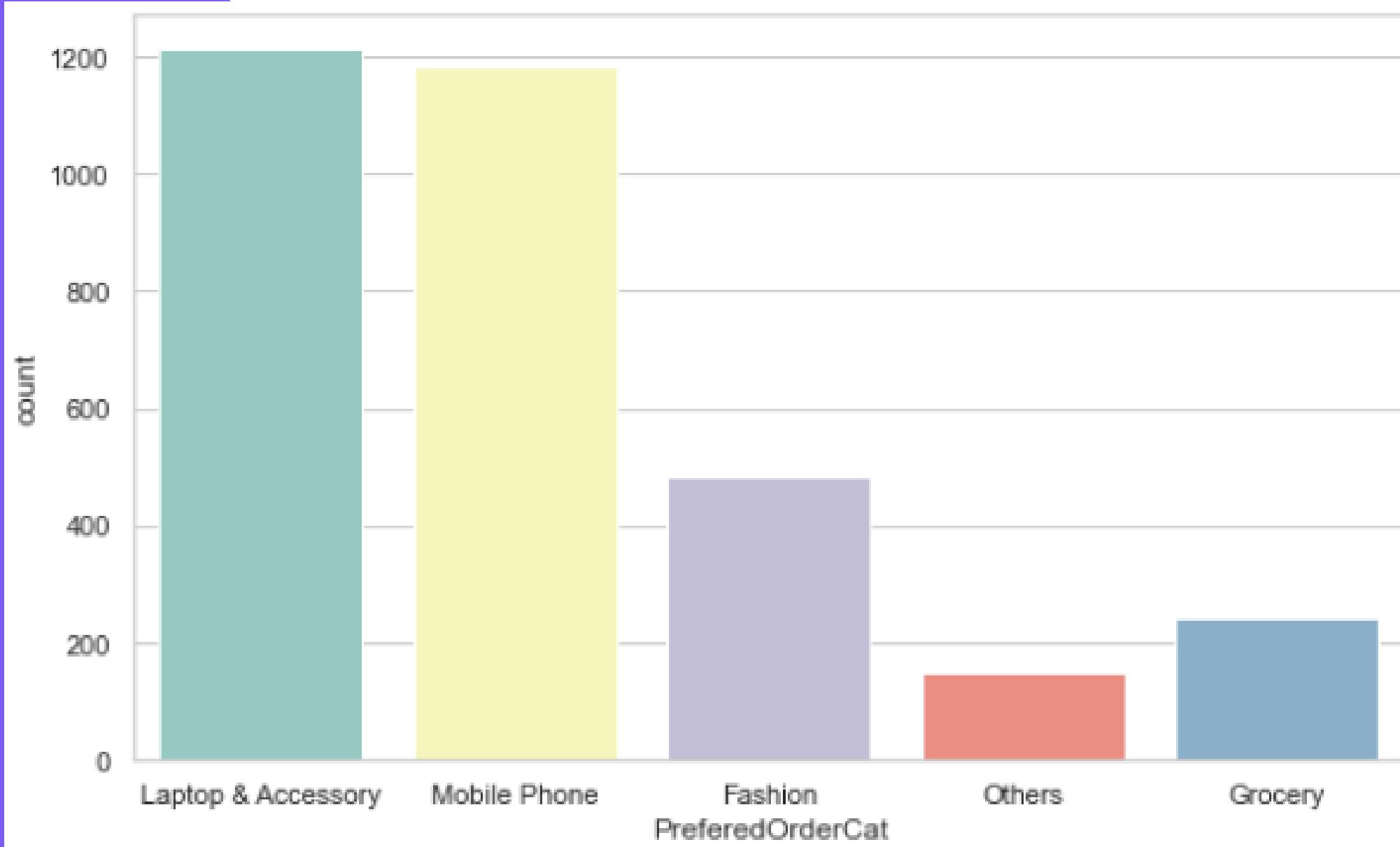
Dataset Column

- Tenure	: Masa aktif pelanggan di perusahaan
- WarehouseToHome	: Jarak antara gudang ke rumah pelanggan
- NumberOfDeviceRegistered	: Jumlah perangkat yang terdaftar oleh pelanggan tertentu
- PreferredOrderCat	: Kategori terakhir yang di order oleh pelanggan pada bulan2 terakhir
- SatisfactionScore	: Nilai kepuasan dari pelanggan untuk perusahaan
- MaritalStatus	: Status pernikahan pelanggan
- NumberOfAddress	: Jumlah alamat yang ditambahkan oleh pelanggan tertentu
- Complain	: Adakah komplain yang dilakukan oleh pelanggan tertentu
- DaySinceLastOrder	: Hari terakhir sejak pelanggan melakukan pembelian
- CashbackAmount	: Rata-rata pengembalian uang pada bulan terakhir
- Churn	: Pendekripsi apakah pelanggan sudah tidak menggunakan E-Commerce Lakü

Note: 1 baris merupakan 1 pelanggan



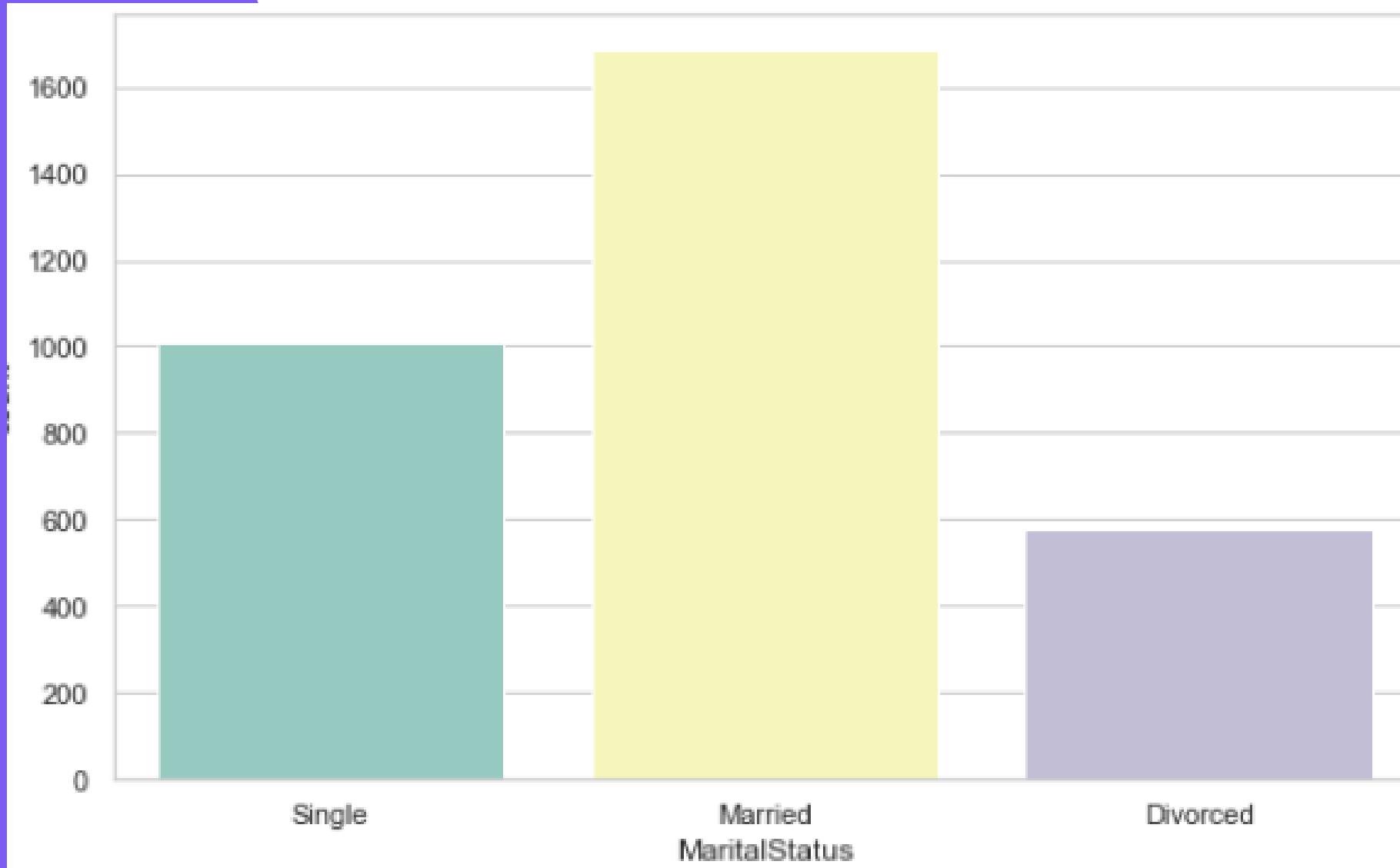
Data Understanding



Kategori yang diminati oleh pelanggan

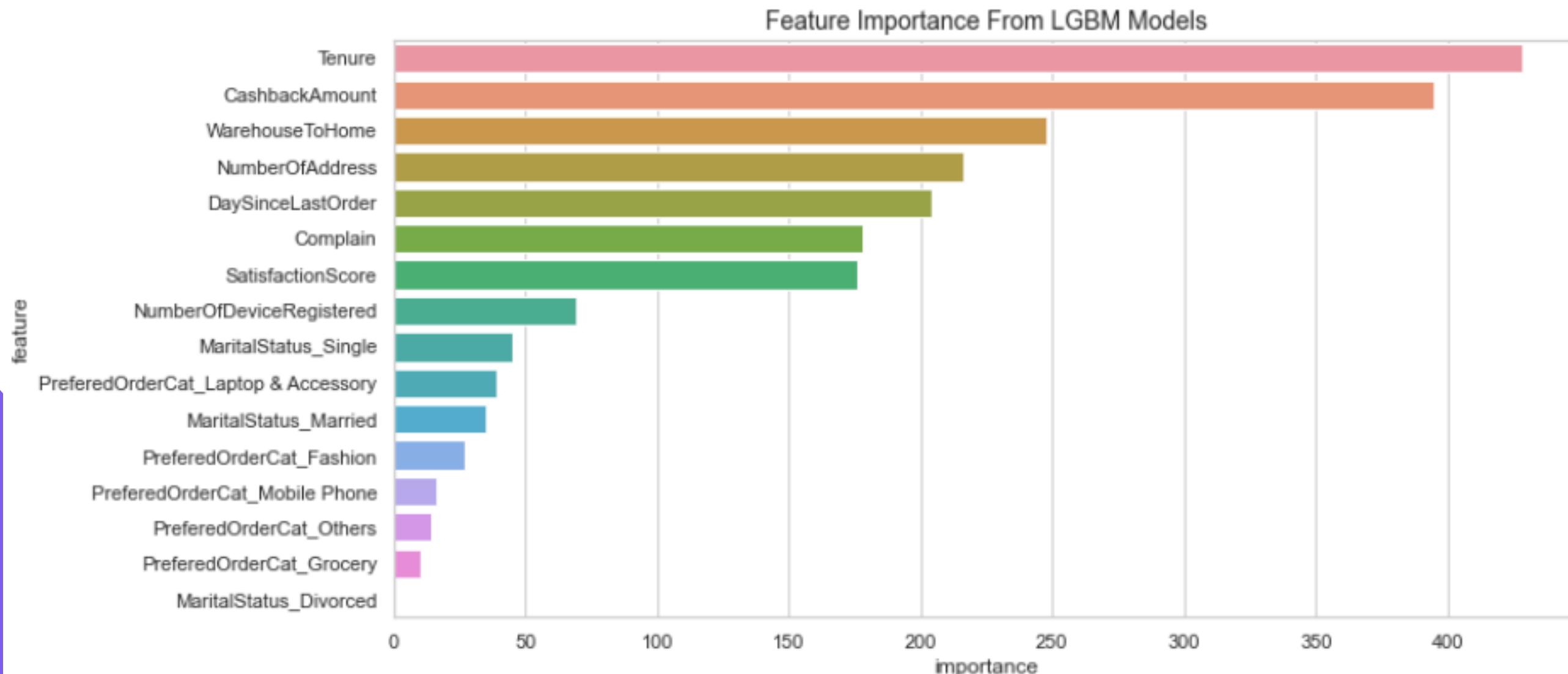


Data Understanding



Status pelanggan E-Commerce

Data Understanding



Feature Importance

Data Cleaning & Preprocessing

Duplicate:

- Before --> 3941
- After --> 3270

Encode: - OneHot Encoder:

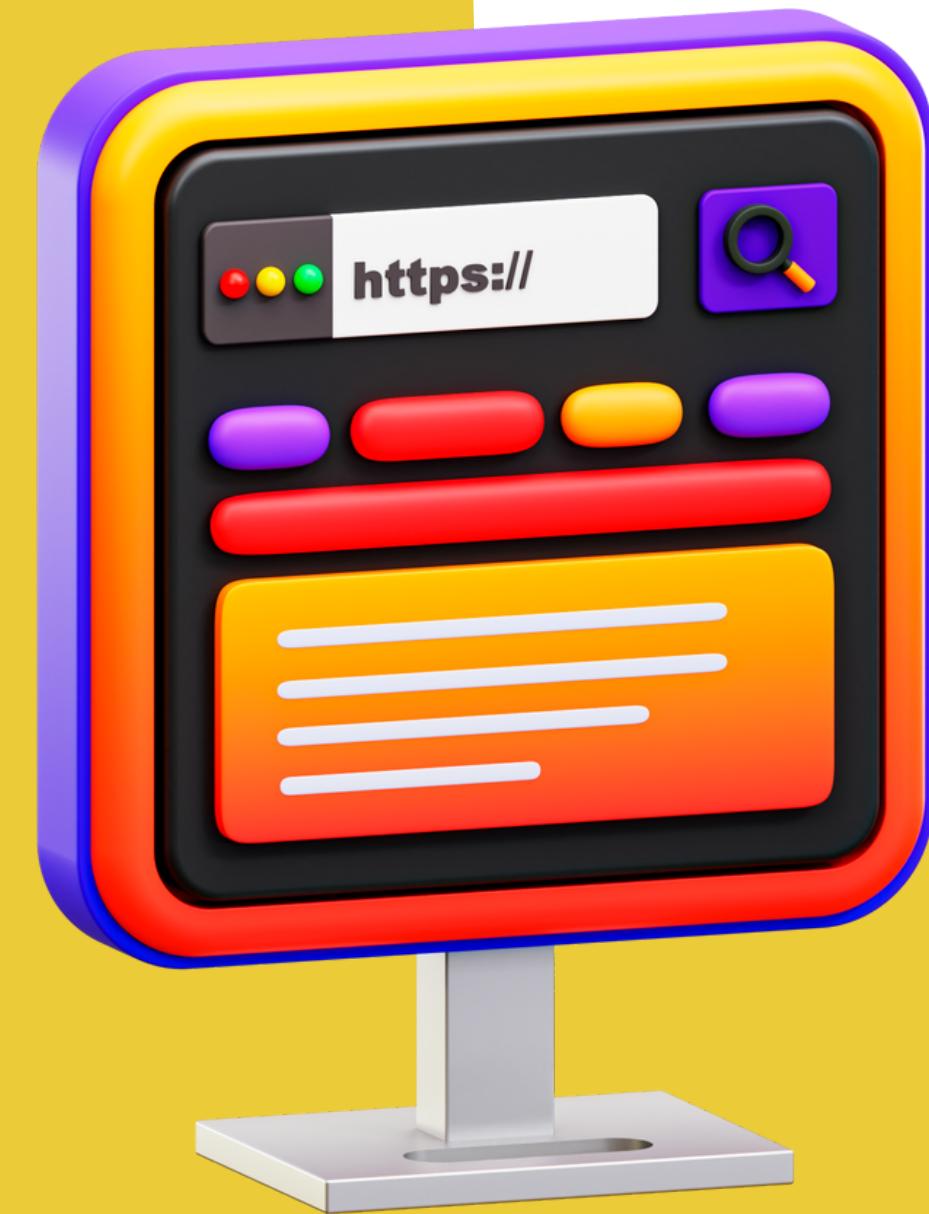
- PreferredOrderCat
- MaritalStatus

Impute Missing Value:

- Tenure
- WarehouseToHome
- DaySinceLastOrder

Scaling:

- MinMaxScaler
- RobustScaler:
- Tenure, WarehouseToHome, NumberOfDeviceRegistered, SatisfactionScore, NumberOfAddress, DaySinceLastOrder, CashbackAmount
- StandardScaler



Modeling

✓ Data Splitting

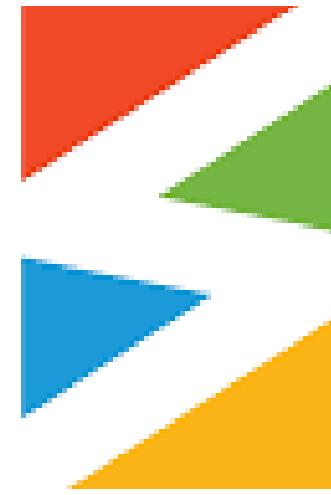
80% Data Train
20% Data Test

✓ Cross Validation

10 Algorithm
F2 Scoring
5 Fold

	🌟	algo	scoring	mean	std	all
39	👍	LGBMClassifier	make_scorer(fbeta_score, beta=2)	0.771	0.027	[0.82, 0.76, 0.748, 0.747, 0.778]
35		StackingClassifier	make_scorer(fbeta_score, beta=2)	0.733	0.032	[0.782, 0.708, 0.688, 0.738, 0.748]
3		LogisticRegression	make_scorer(fbeta_score, beta=2)	0.709	0.034	[0.747, 0.672, 0.679, 0.695, 0.751]
27		XGBClassifier	make_scorer(fbeta_score, beta=2)	0.684	0.051	[0.757, 0.659, 0.653, 0.73, 0.622]
31		VotingClassifier	make_scorer(fbeta_score, beta=2)	0.637	0.049	[0.708, 0.588, 0.576, 0.666, 0.647]
11		DecisionTreeClassifier	make_scorer(fbeta_score, beta=2)	0.617	0.060	[0.693, 0.54, 0.604, 0.679, 0.568]
23		GradientBoostingClassifier	make_scorer(fbeta_score, beta=2)	0.606	0.036	[0.603, 0.634, 0.616, 0.637, 0.538]
15		RandomForestClassifier	make_scorer(fbeta_score, beta=2)	0.602	0.028	[0.653, 0.605, 0.571, 0.6, 0.584]
19		AdaBoostClassifier	make_scorer(fbeta_score, beta=2)	0.583	0.033	[0.639, 0.543, 0.564, 0.602, 0.569]
7		KNeighborsClassifier	make_scorer(fbeta_score, beta=2)	0.380	0.047	[0.401, 0.387, 0.296, 0.379, 0.438]

2 Algoritma Dengan Nilai Terbaik:
LGBM Classifier & Stacking Classifier



LightGBM

Light GBM merupakan implementasi Gradient Boosting Decision Tree (GBDT). Pada proses pelatihan setiap individual decision tree akan melakukan pemisahan data. LightGBM menggunakan dua strategi yaitu gradient-based one-side sampling (GOSS) dan leaf-wise growth.

Konsep leaf-wise growth merupakan salah satu Teknik untuk membatasi depth dari model LightGBM, proses ini dilakukan untuk mencari node dengan splitting gains terbesar. Proses yang dilakukan selanjutnya adalah memecahkan node tersebut dan meneruskan untuk node yang baru. Model Light GBM tidak perlu menambahkan kedalaman model untuk menghindari penggunaan daya komputasi yang lebih besar dan juga mengurangi overfitting.

Keuntungan:

1. Training speed yang lebih cepat dan efisien
2. Penggunaan memori yang minimal
3. Tingkat akurasi yang baik
4. Bisa mengatasi data berukuran besar
5. Support untuk pembelajaran menggunakan GPU
6. Mengatasi data tidak seimbang



Modeling (F2 Scoring)

✓ Train Set Base Model

- LGBM Classifier : 77.1
- Stacking Classifier : 73.3



- N_Estimators
- Max_depth
- Learning_Rate
- Num_Leaves
- Boosting_Type
- Encoder: OneHot & Binary
- Scaling: Robust, MinMax, & Standart
- Imputer: Iterative, Simple, & KNNImputer
- Resampler: None, RandomOver, RandomUnder, Nearmiss, & SMOTE

✓ Test Set Base Model

- LGBM Classifier : 78.7
- Stacking Classifier : 76.4



✓ Train Set Tune Model

- LGBM Classifier : 79.2 dan 79.4
- Stacking Classifier : 75.2

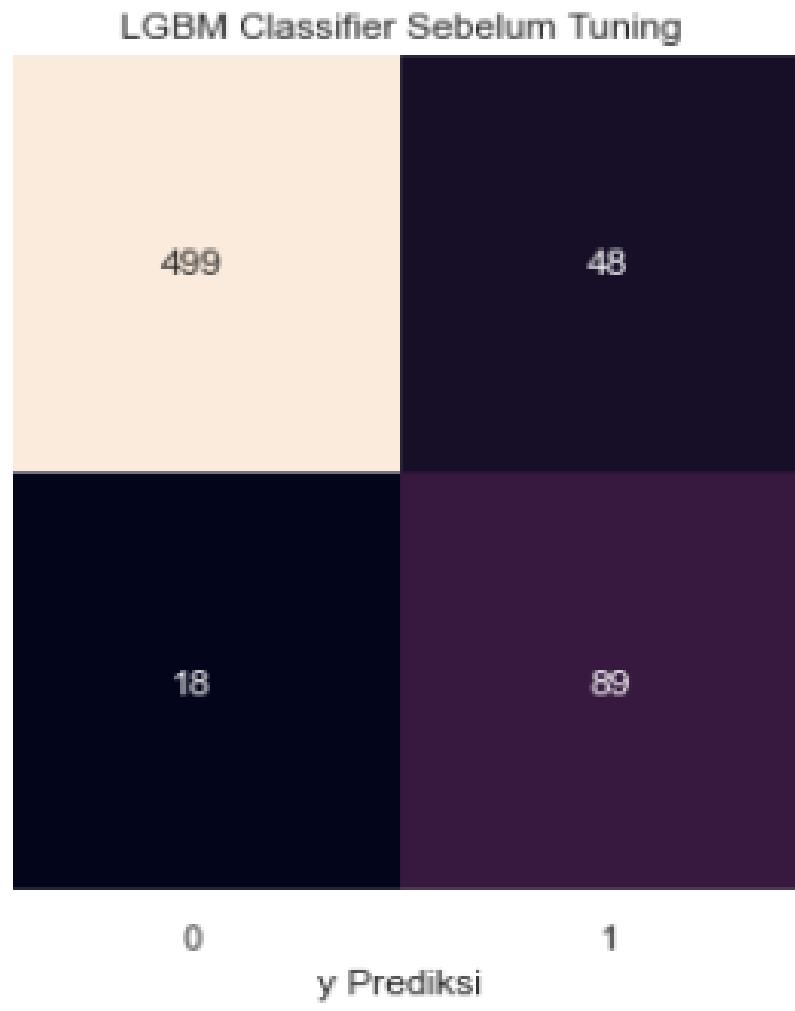


✓ Test Set Tune Model

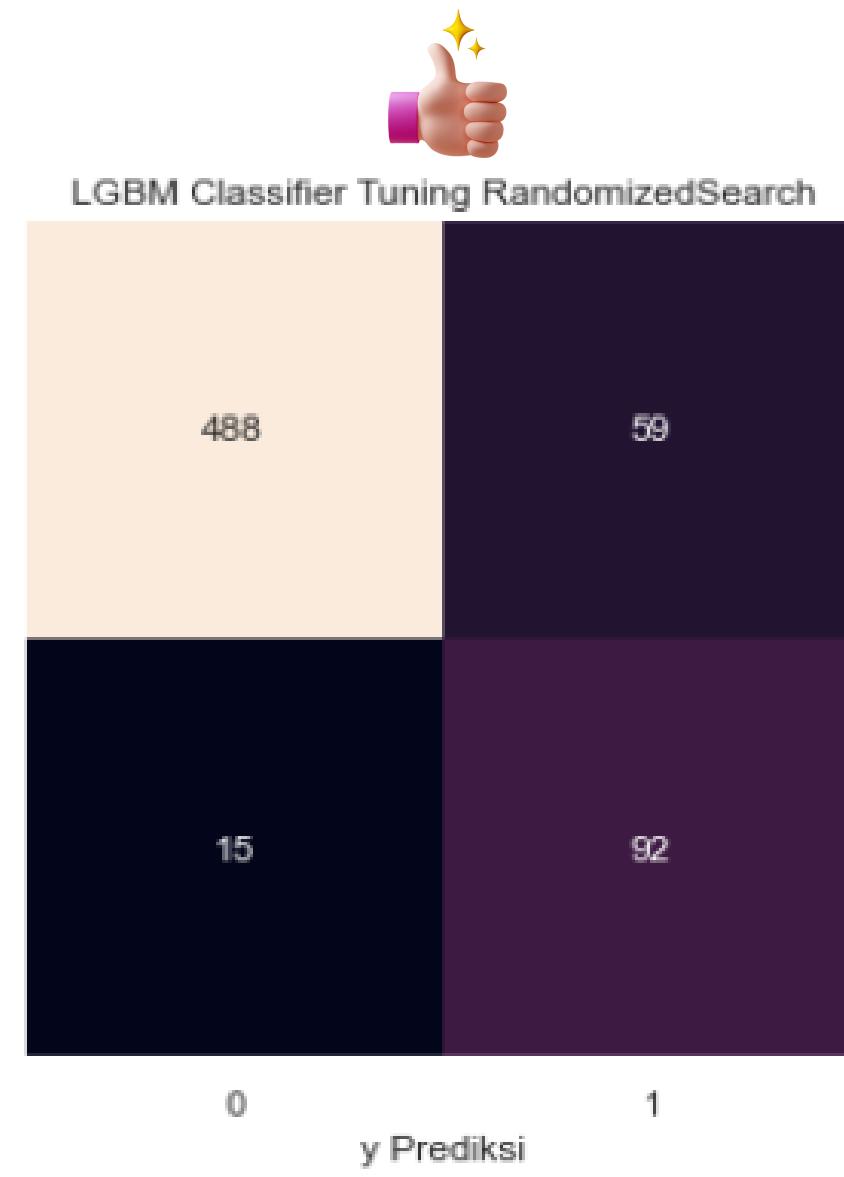
- LGBM Classifier : 79.4
- Stacking Classifier : 76.4



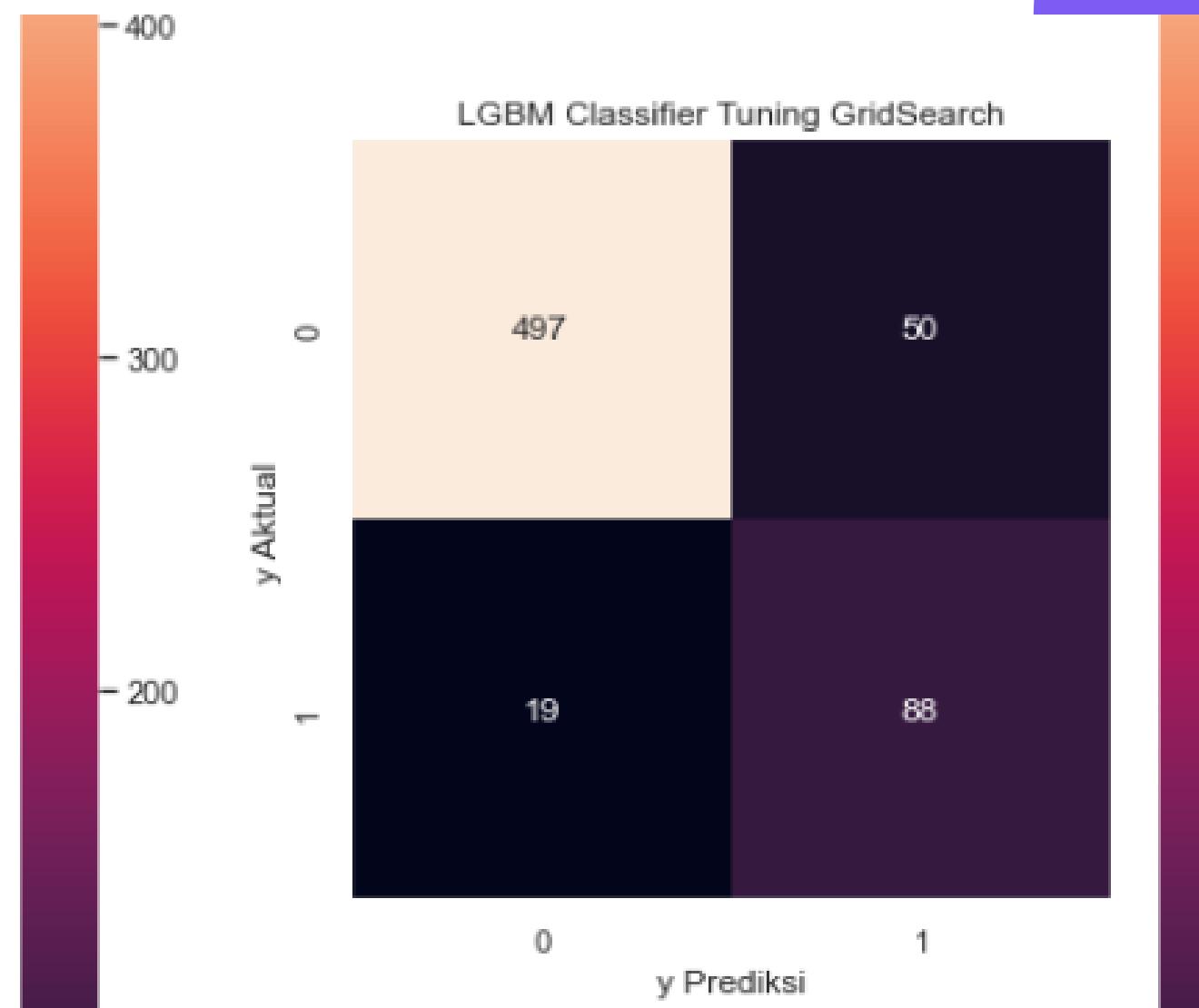
Model Matrix



Base Model



Randomized Tuning



Grid Tuning

No Machine Learning

Vs

With Machine Learning

No Machine Learning

- Perusahaan untuk semua pelanggan adalah $(547 + 107) \times \text{Rp } 30.000 = \text{Rp } 19.620.000$
- Kerugian perusahaan untuk pelanggan hilang karena salah prediksi $107 \times \text{Rp } 700.000 = \text{Rp } 74.900.000$ (FN)
- Kerugian perusahaan $\text{Rp } 19.620.000 - \text{Rp } 74.900.000 = \text{Rp } 55.280.000$

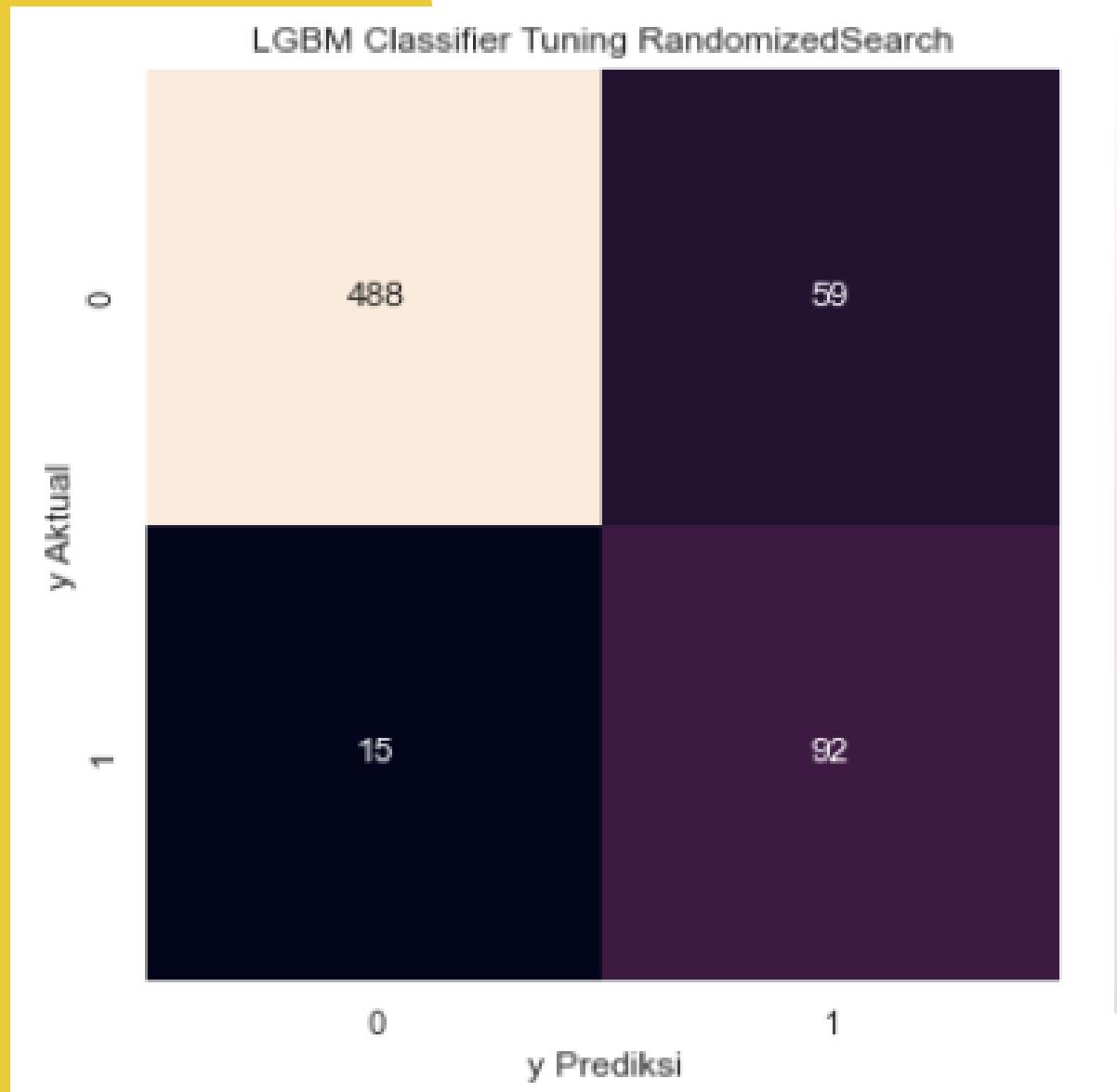
Tune Model

- Perusahaan salah memprediksi pelanggan Loyal menjadi Churn $48 \times \text{Rp } 70.000 = \text{Rp } 3.360.000$, seharusnya $48 \times \text{Rp } 30.000 = \text{Rp } 1.440.000$ (sia-sia Rp 1.920.000) (FP)
- Perusahaan kehilangan pelanggan karena tidak terdeteksi Churn $18 \times \text{Rp } 700.000 = \text{Rp } 12.600.000$ (FN)
- Kerugian perusahaan dengan ML Base Model $\text{Rp } 12.600.000 + \text{Rp } 1.920.000 = \text{Rp } 14.520.000$

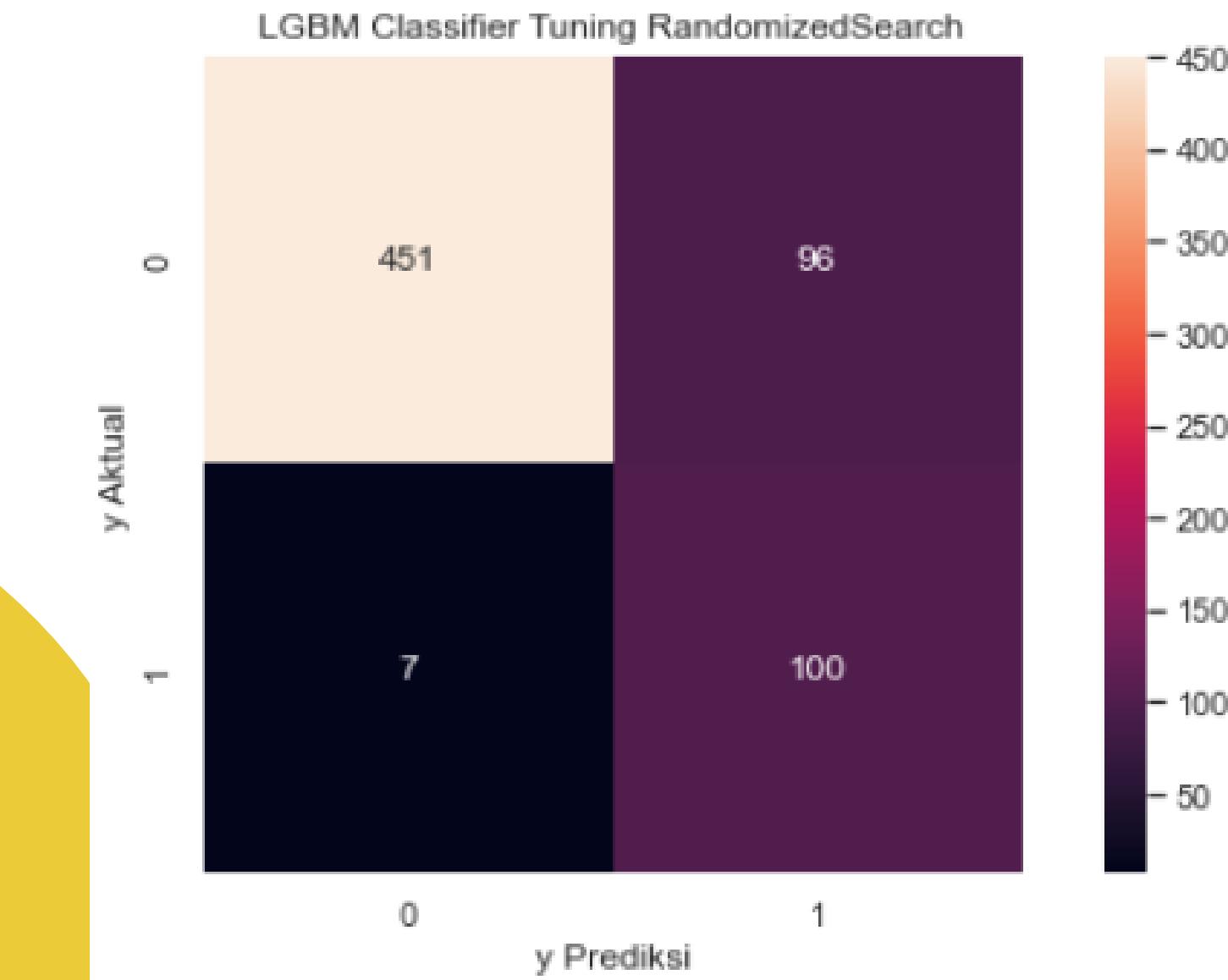
Tune & Optimized Threshold

- Perusahaan salah memprediksi pelanggan Loyal menjadi Churn $96 \times \text{Rp } 70.000 = \text{Rp } 6.720.000$, seharusnya $96 \times \text{Rp } 30.000 = \text{Rp } 2.880.000$ (sia-sia Rp 3.840.000) (FP)
- Perusahaan kehilangan pelanggan karena tidak terdeteksi Churn $7 \times \text{Rp } 700.000 = \text{Rp } 4.900.000$ (FN)
- Kerugian perusahaan dengan ML after tuning Model & Optimized Treshold $\text{Rp } 4.900.000 + \text{Rp } 3.840.000 = \text{Rp } 8.740.000$

WITH MACHINE LEARNING



76% Rp 14.520.000
MENGHEMAT BIAYA



84% Rp 8.740.000
MENGHEMAT BIAYA

KESIMPULAN

F2 Scoring

Kerugian FN lebih besar dari pada kerugian FP, sehingga jumlah hasil FN harus diminimalkan. Namun juga kita tetap mempertahankan FP untuk tidak sangat tinggi.

BEST MODEL

LGBM Classifier tune & Optimized Threshold

No ML & With ML

- Kerugian Perusahaan Sebelum ML : Rp. 55.280.000
- Kerugian Perusahaan dengan ML Base Model : Rp. 14.520.000
- Kerugian Perusahaan dengan ML Tune Model : Rp. 12.860.000
- Kerugian Perusahaan dengan ML Tune & Optimize Treshold : Rp 8.740.000

Feature Importance

CashbackAmount -->
perusahaan sudah mengambil
langkah bagus untuk
memberikan penawaran
menarik kepada setiap
pelanggannya.



Recommendation

1. Data Scientist yang mengerjakan model ini, memberikan saran kepada Data Scientist selanjutnya yang akan mengerjakan model ini. Akan lebih baik, jika dicoba menggunakan model lain selain 10 model yang digunakan saat ini. Supaya mendapatkan perbandingan hasil model yang lebih baik.
2. Model saat ini sebetulnya masih banyak keterbatasan, karena hyperparameter tuning yang dilakukan kurang maksimal akibat hardware yang kurang mendukung. Disarankan untuk Data Scientist selanjutnya melakukan banyak kombinasi hyperparameter dengan GridSearch, untuk menemukan score yang maksimal.
3. Feature yang ditemukan saat ini, jika menurut Data Scientist selanjutnya tidak relevan. Akan lebih baik, jika feature tersebut dibuang. (Contoh: Feature dengan score 0 di feature importance)

Rekomendasi Perusahaan, Bisa terlihat nilai ketidakpuasan perusahaan Lakü yaitu 61%, memang sudah bagus perusahaan memperlakukan pelanggannya dengan baik yaitu dengan cara memberikan promo, namun perusahaan harus tetap memperhatikan faktor-faktor lain yang mempengaruhi 61% pelanggan tidak puas tersebut.