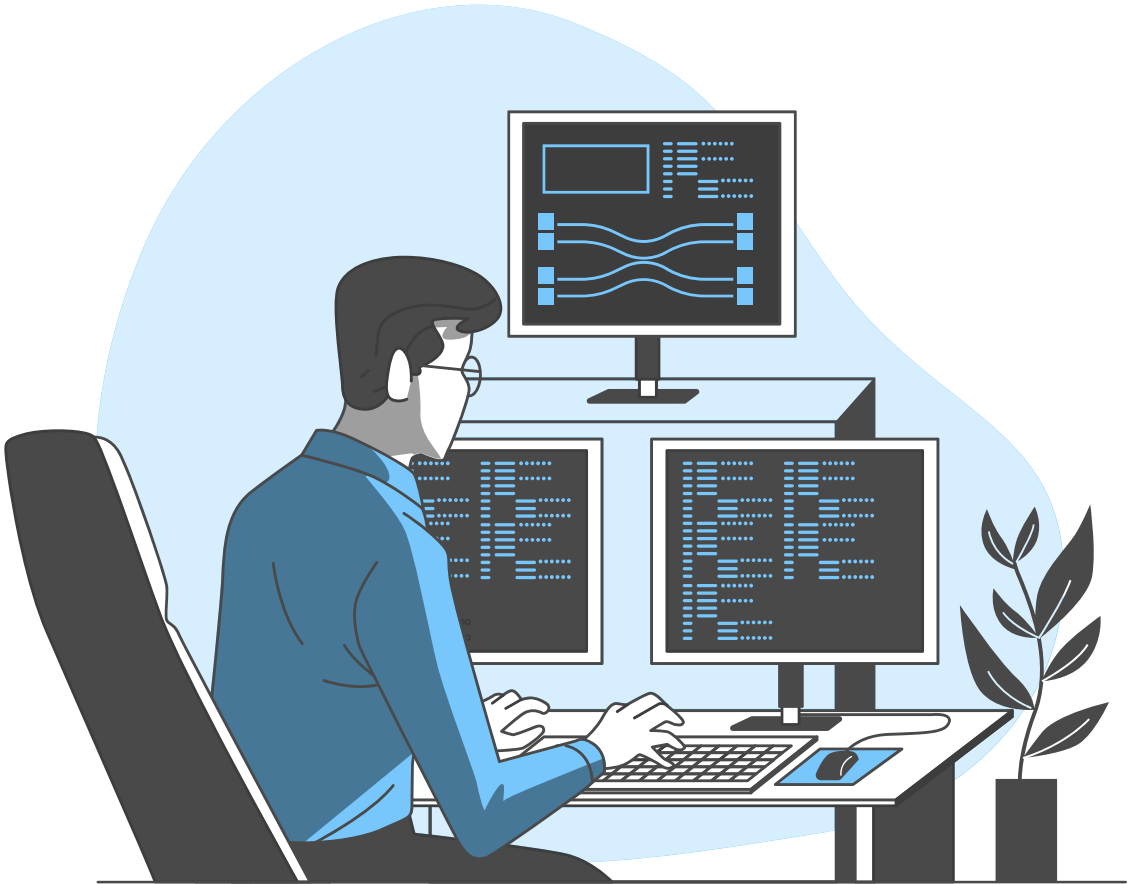
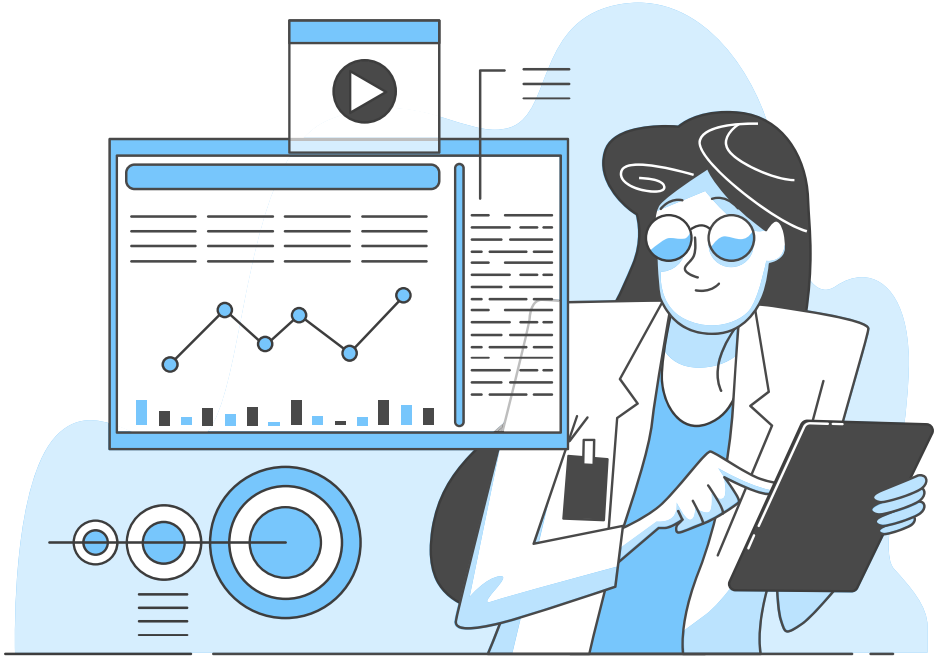


E-commerce Customer Churn Prediction

Indra Rivaldi Siregar

https://github.com/insersir/Churn-Prediction

Table of Contents



01

Business Understanding

Understanding the Problem

Suatu e-commerce “Blue Shine” saat ini sedang mengalami peningkatan churn rate yang nilainya sudah mencapai 18.40 %.

Understanding the Problem

Churn

Churn adalah istilah bagi customer yang berhenti berlangganan (menutup / menghapus akun)

Churn Rate

Persentase pelanggan yang berhenti berlanggan dalam kurun waktu tertentu

$$\textbf{Churn Rate} = \frac{\text{Leaving Customer}}{\text{Total Customer}}$$

Understanding the Problem

Goal

Menurunkan jumlah customer yang churn

Objective

Memprediksi customer yang akan churn

Business metrics

Customer Churn Rate

02

Exploratory Data Analysis

Dataset

CustomerID
Tenure*
PreferredLoginDevice
CityTier
WarehouseToHome*
PreferredPaymentMode
Gender
HourSpendOnApp*
NumberOfDeviceRegistered
PreferedOrderCat
SatisfactionScore
MaritalStatus
NumberOfAddress
Complain
OrderAmountHikeFromlastYear*
CouponUsed*
OrderCount*
DaySinceLastOrder*
CashbackAmount

Churn 🎯

Shape

"5630 baris dan 20 kolom"

Duplicated Data

"Tidak terdapat data duplikat"

Missing Value (*)

"Terdapat 7 kolom yang memiliki missing value"

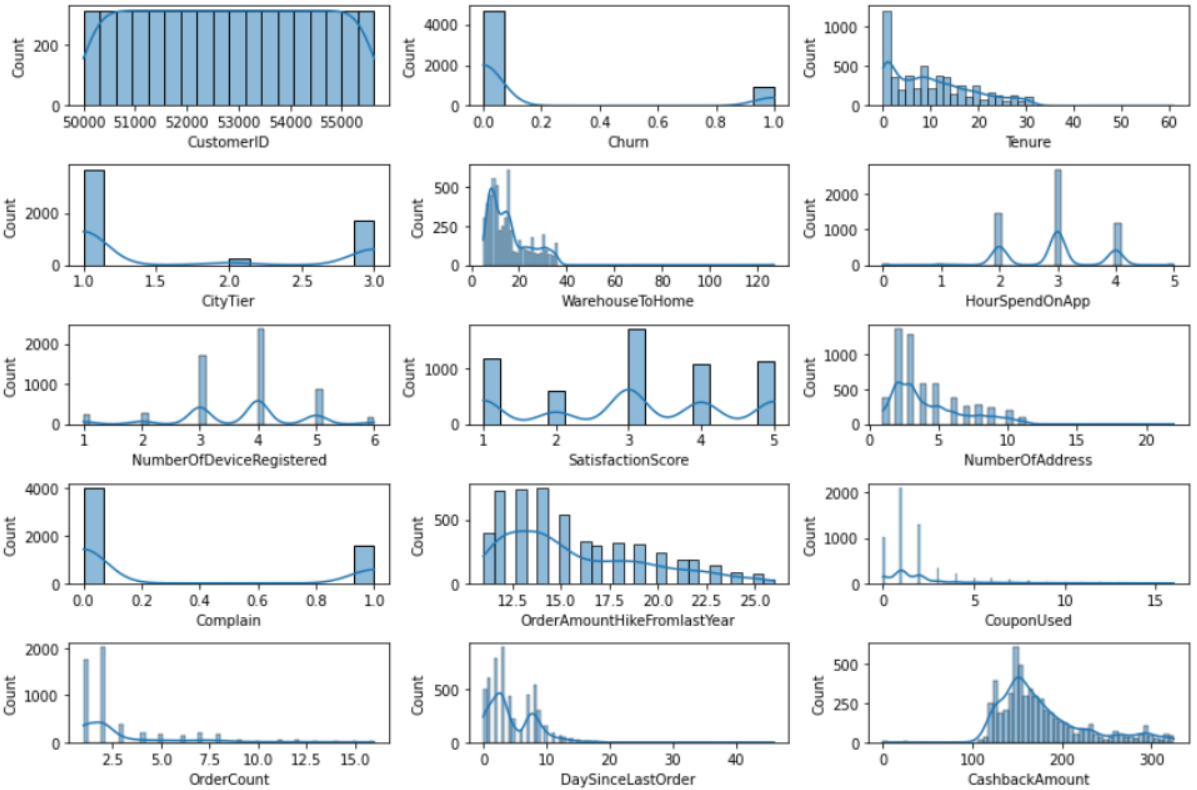
Exploratory Data Analysis (EDA)

Numerical Features

Terdapat **8 feature** dengan distribusi right skewed, yaitu:

- Tenure
- WarehouseToHome
- NumberOfAddress
- OrderAmountHikeFromlastYear
- CouponUsed
- OrderCount
- DaySinceLastOrder
- CashbackAmount

"TRANSFORMASI"



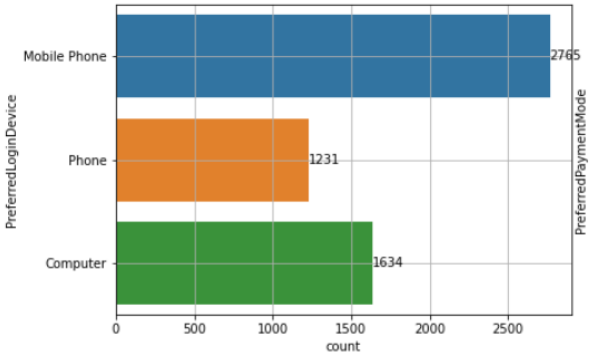
Exploratory Data Analysis (EDA)

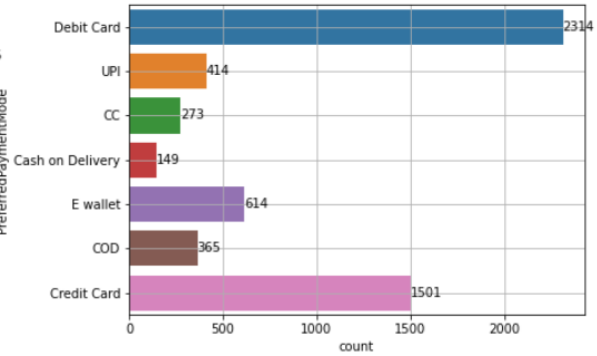
Categorical Features

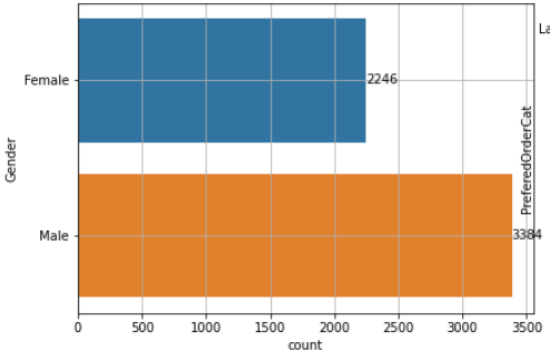
Terdapat **Unique Value** yang tidak konsisten penulisannya pada beberapa feature, yaitu:

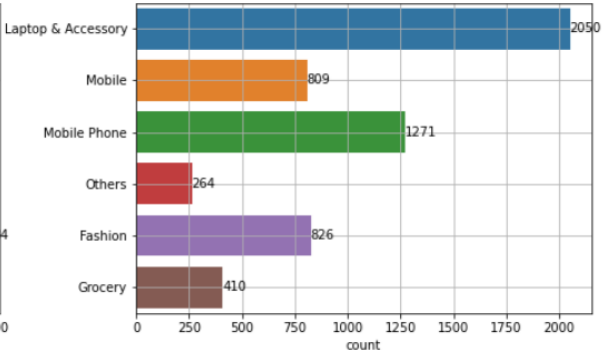
- **PreferredLoginDevice** : Mobile Phone dan Phone
- **PreferredPaymentMode**: CC dan Credit Card; Cash On Delivery dan COD
- **PreferedOrderCat**: Mobile dan Mobile Phone

"Handling Inconsistent Value"









03

Data Pre-processing

Data Pre-processing

- **Handling Missing Value** : Mengisi dengan nilai median (karena feature yang right skewed distribution)
- **Feature Transformation** : Menggunakan metode yeo-johnson (bekerja dengan baik pada left and right skewed distribution)
- **Handling Outlier** : Menggunakan z-score (tidak seekstrim metode IQR dalam membuang outlier)
- **Label Encoding** : Untuk feature PreferredLoginDevice dan Gender (karena unique value berjumlah 2)
- **One Hot Encoding** : Untuk feature PreferredPaymentMode, PreferedOrderCat, MaritalStatus (karena bukan ordinal)
- **Splitting Data** : 80% train data dan 20% test data
- **Standarisasi** : dilakukan setelah splitting data agar tidak terjadi leak data dari test dataset

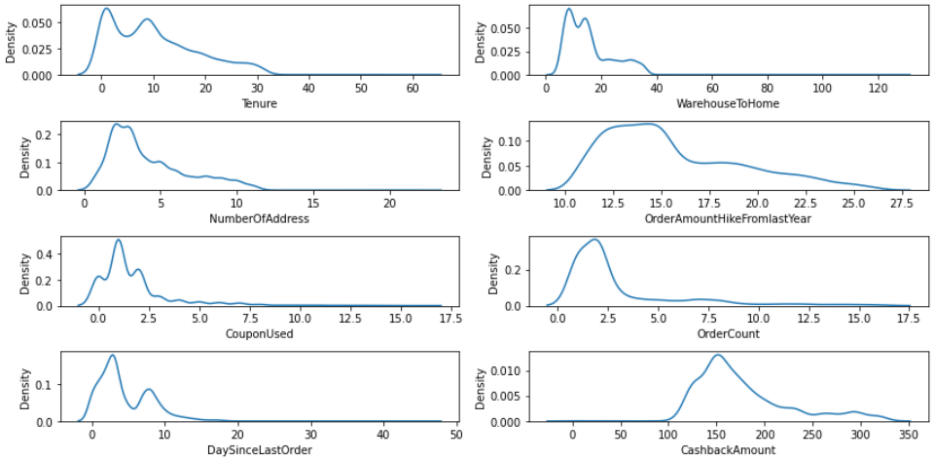
Feature Transformation

Numerical Features

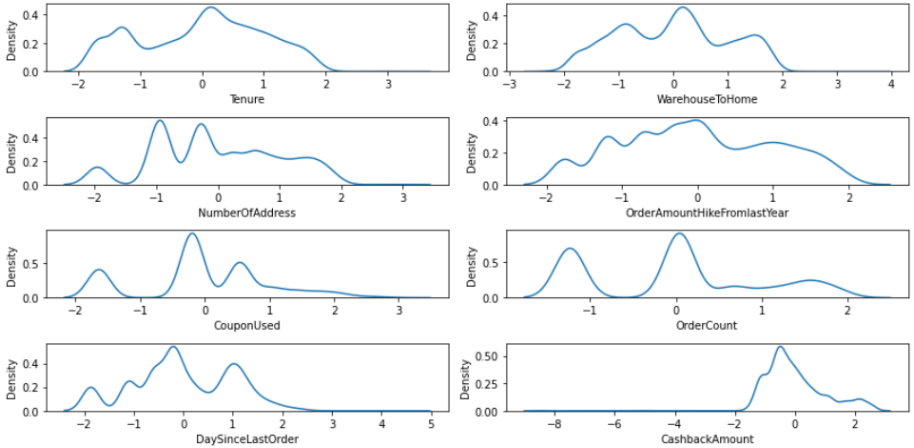
Transformasi menggunakan metode ‘yeo-johnson’

Transformasi dilakukan terlebih dahulu sebelum melakukan handling outlier. Tujuannya agar tidak banyak data yang terbuang karena diindikasi sebagai outlier akibat distribusi data yang tidak mendekati normal

“BEFORE TRANSFORMATION”



“AFTER TRANSFORMATION”



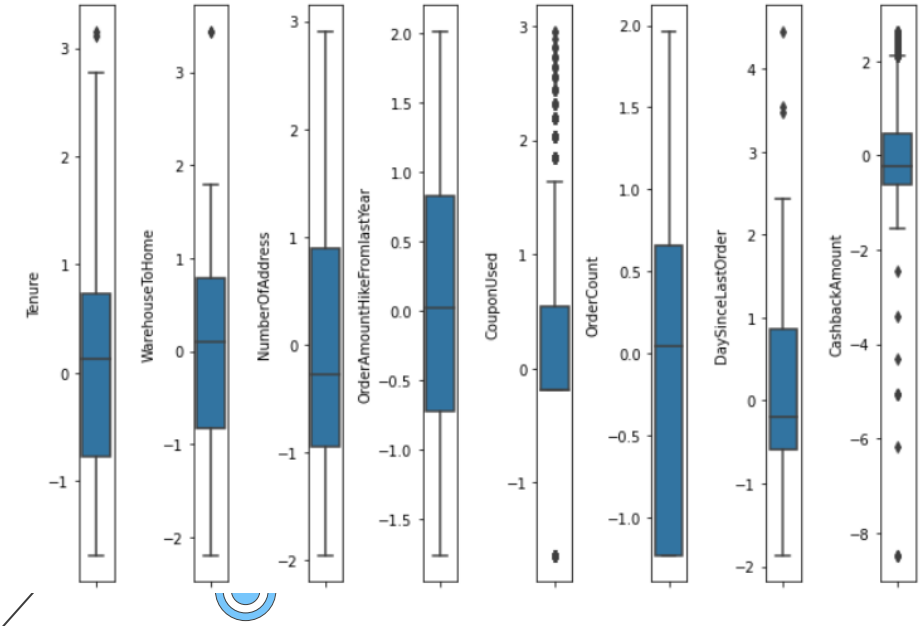
Handling Outliers

Handling Outliers menggunakan Metode Z-Score

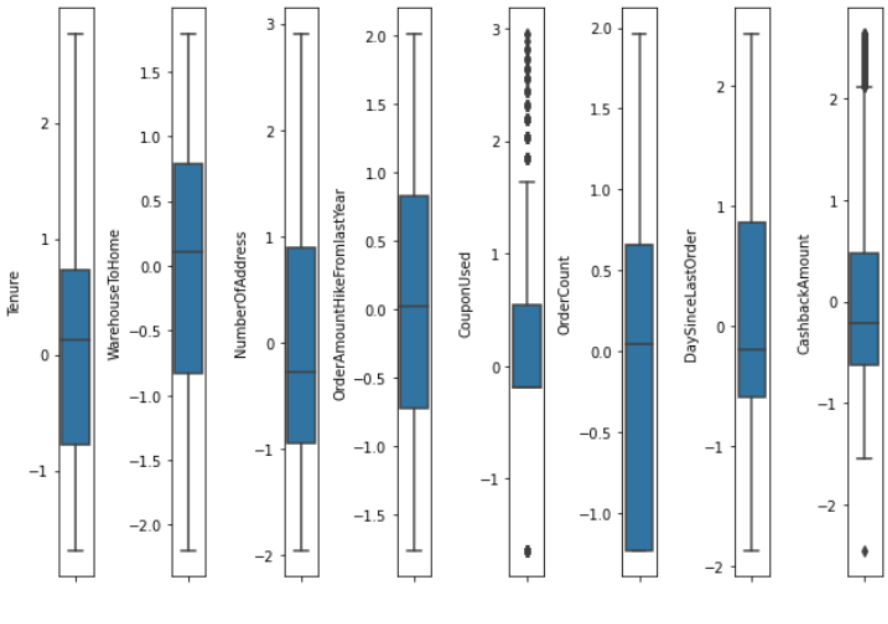
Numerical Features

Terlihat bahwa sebagian besar feature yang sebelumnya mengandung outlier kini tidak lagi mengandung outlier. Meskipun begitu, outlier masih teridentifikasi pada CoupenUsed dan CashbackAmount

"BEFORE HANDLING OUTLIER"



"AFTER HANDLING OUTLIER"



03

Modeling and Evaluation

Metrics Evaluation

- **Recall (True Positive Rate)**: Tidak memperbolehkan False Negative yang besar. Artinya, kita fokus pada customer yang churn.
- **ROC_AUC**: karena target pada kasus ini bersifat imbalance, maka perlu dicek juga nilai ROC_AUC yang merepresentasikan apakah model mampu membedakan kelas dengan baik. Dikuatirkan model justru memprediksi semua data menjadi Churn (1) maupun Not Churn (0).

Modeling and Evaluation

No	Model	Train		Test	
		ROC_AUC	Recall	ROC_AUC	Recall
1	Decision Tree	1.0	1.0	0.93	0.88
2	KNN	0.94	0.9	0.83	0.68
3	SVM	0.85	0.7	0.82	0.66
4	Random Forest	1.0	1.0	0.95	0.91
5	Gradient Boosting	0.83	0.68	0.8	0.64
6	XGBoost	0.81	0.65	0.78	0.6
7	Catboost	0.98	0.95	0.92	0.85
8	Extra Tree Classifier	1.0	1.0	0.95	0.9

Tiga model terbaik adalah Decision Tree, Random Forest dan Extra Tree Classifier

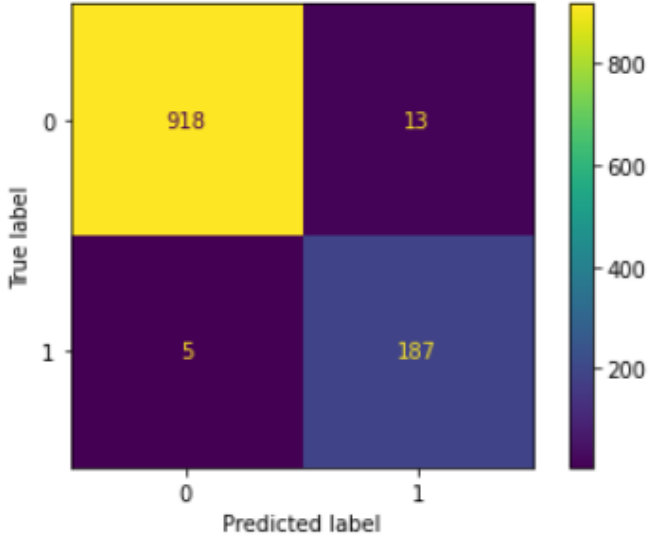
Hyperparamater Tuning

	Model	Train		Test	
		ROC_AUC	Recall	ROC_AUC	Recall
Before Hyperparameter Tuning	Decision Tree	1.0	1.0	0.93	0.88
After Hyperparameter Tuning		1.0	1.0	0.91	0.86
Before Hyperparameter Tuning	Random Forest	1.0	1.0	0.95	0.91
After Hyperparameter Tuning		1.0	1.0	0.93	0.86
Before Hyperparameter Tuning	Extra Tree Classifier	1.0	1.0	0.95	0.9
After Hyperparameter Tuning		1.0	1.0	0.96	0.94

Model terbaik setelah hyperparameter tuning adalah Extra Tree Classifier dengan RECALL dan ROC_AUC tertinggi pada testing data

Confusion Matrix

Recall score dari model Extra Tree Classifier setelah hyperparameter tuning adalah 94%. Artinya, dari seluruh customer yang churn, **hanya 6% saja** yang salah diprediksi sebagai not churn (false negative). Ini menunjukkan bahwa potensi yang cukup kecil dalam melakukan kesalahan prediksi customer yang churn.



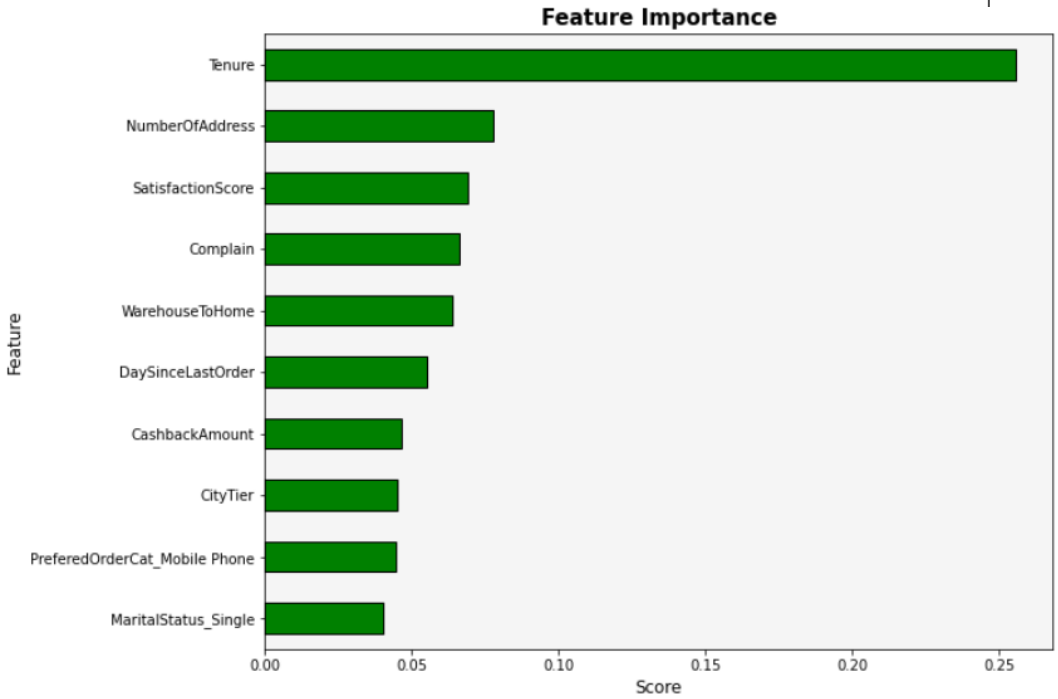
03

Business Recomendation

Feature Importance

Feature Importance merepresentasikan fitur-fitur yang paling berpengaruh dari suatu model dalam memprediksi suatu target. Dalam kasus ini, feature importance diekstrak dari model Extra Tree Classifier yang berfungsi untuk memprediksi target Churn.

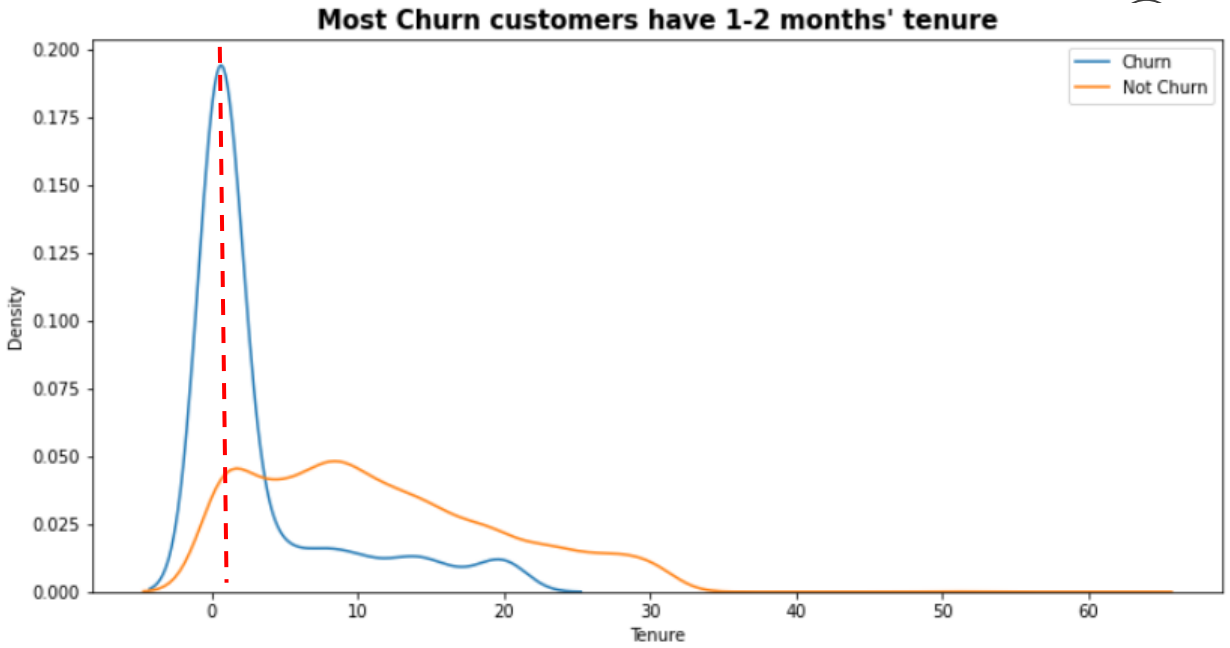
Berdasarkan **feature importance** di samping, dapat diamati beberapa feature penting yang sangat mempengaruhi seorang customer untuk melakukan churn, seperti **tenure**, **jumlah alamat yang terdaftar**, **skor kepuasan**, **komplain**, dan lainnya. Ini dapat digunakan sebagai **bussiness insight** dan **business recommendation**.



Business Insight and Business Recommendation

Tenure

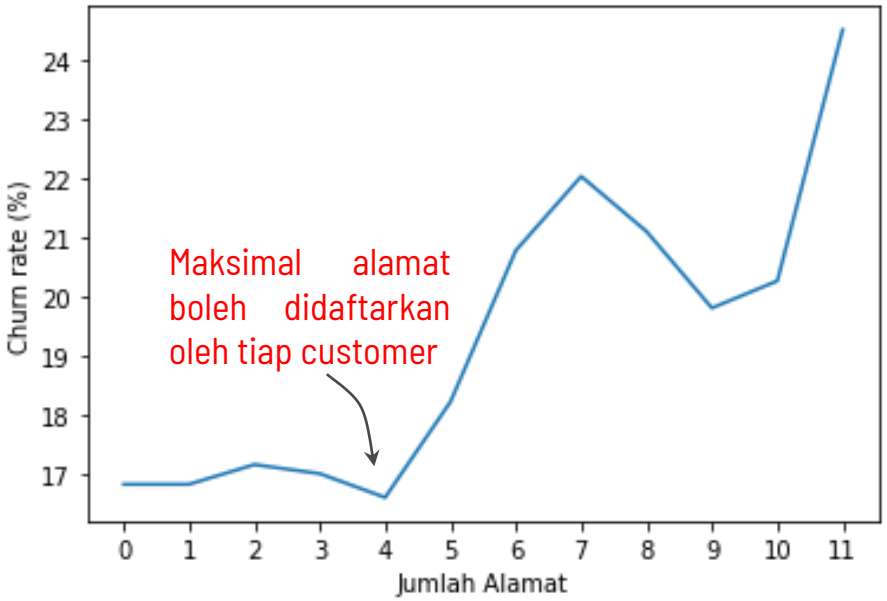
Berdasarkan kde plot di samping, diketahui bahwa customer yang churn biasanya memiliki Tenure yang rendah yakni sekitar 1-2 bulan. Artinya, kita perlu membuat para customer untuk bisa bertahan lebih > 2 bulan agar potensi untuk Churn-nya menjadi lebih rendah. Beberapa rekomendasi yang dapat diberikan seperti cashback atau promo.



Business Insight and Business Recommendation

NumberofAddress

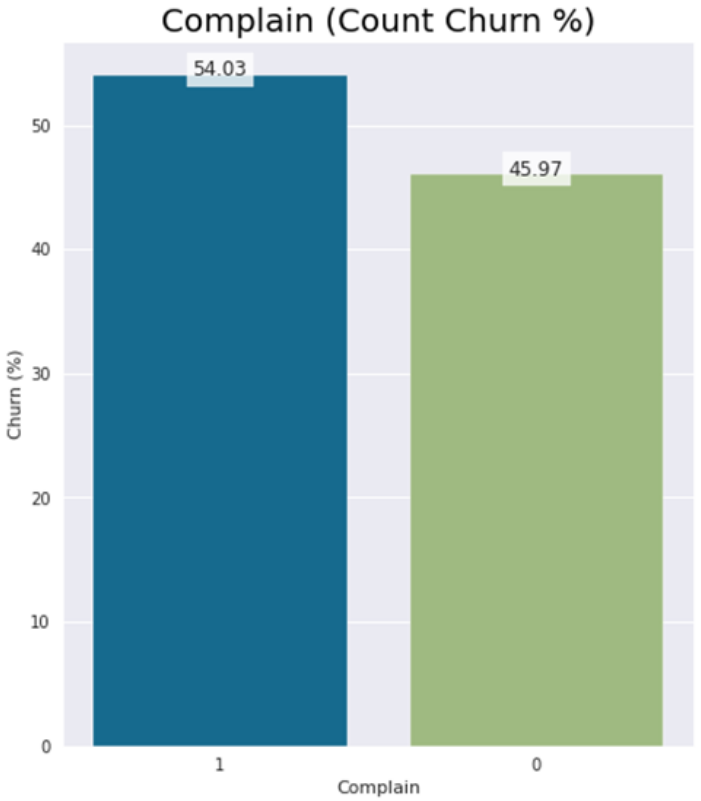
Berdasarkan grafik di samping, Churn Rate akan meningkat secara signifikan ketika jumlah alamat > 4. Oleh karena itu, disarankan agar customer hanya boleh mendaftar alamatnya pada e-commerce <= 4.



Business Insight and Business Recommendation

Complain

Semakin banyak Complain oleh Customer, maka Churn Rate semakin meningkat. Oleh karena itu, Complain dari customer harus segera ditanggapi agar customer tidak melakukan Churn kedepannya.



Thanks!

Do you have any questions?

indrarivaldisiregar@gmail.com
+62 81322487097