



Churn Prediction

Presentation by **Gretta Frizhana**



Introduction

Pelanggan memiliki hak dalam memilih provider yang sesuai dan dapat beralih dari provider sebelumnya yang diartikan sebagai Customer Churn. Peralihan ini dapat menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk ditangani.



Data

'account_length'	'area_code'	'number_vmail_messages'
'total_day_minutes'	'total_day_calls'	'total_day_charge'
'total_eve_minutes'	'total_eve_calls'	'total_eve_charge'
'total_night_minutes'	'total_night_calls'	'total_night_charge'
'total_intl_minutes'	'total_intl_calls'	'total_intl_charge'
'number_customer_service_calls'	'churn'	'avg_day_charge_call'
'avg_day_call_length'	'avg_eve_charge_call'	'avg_eve_call_length'
'avg_intl_charge'	'avg_intl_charge_call'	'intl'

Workflow

01

Load Data

02

Exploratory Data Analysis

03

Feature Engineering

04

Normalization

05

Data Encoding

06

Train-Test Split

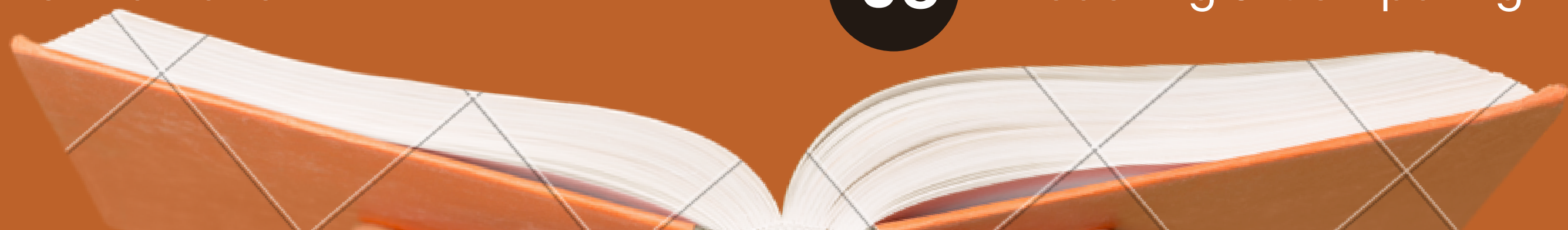
07


Resampling

08


Modelling & Comparing Model

Pre-Processing



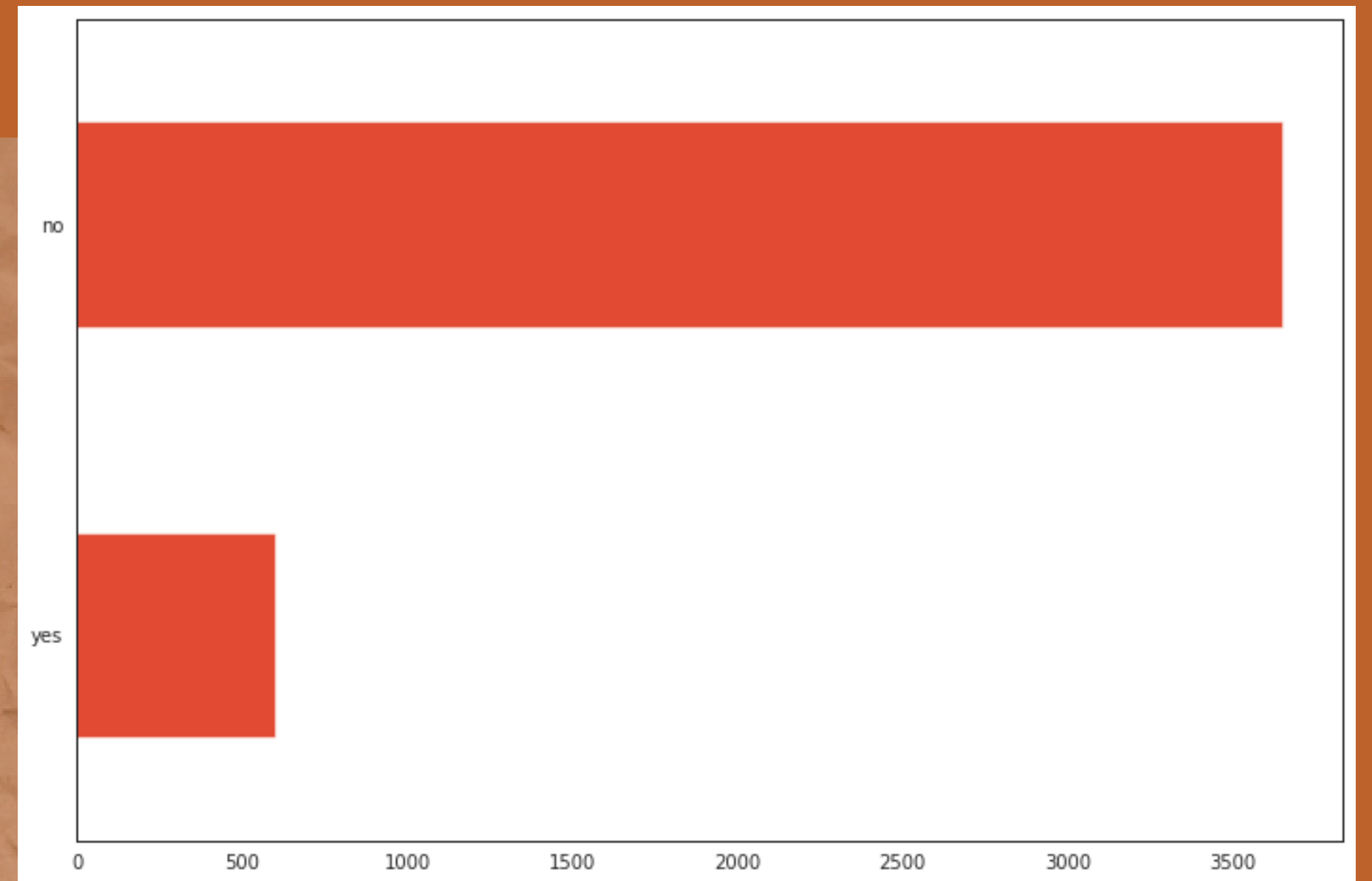


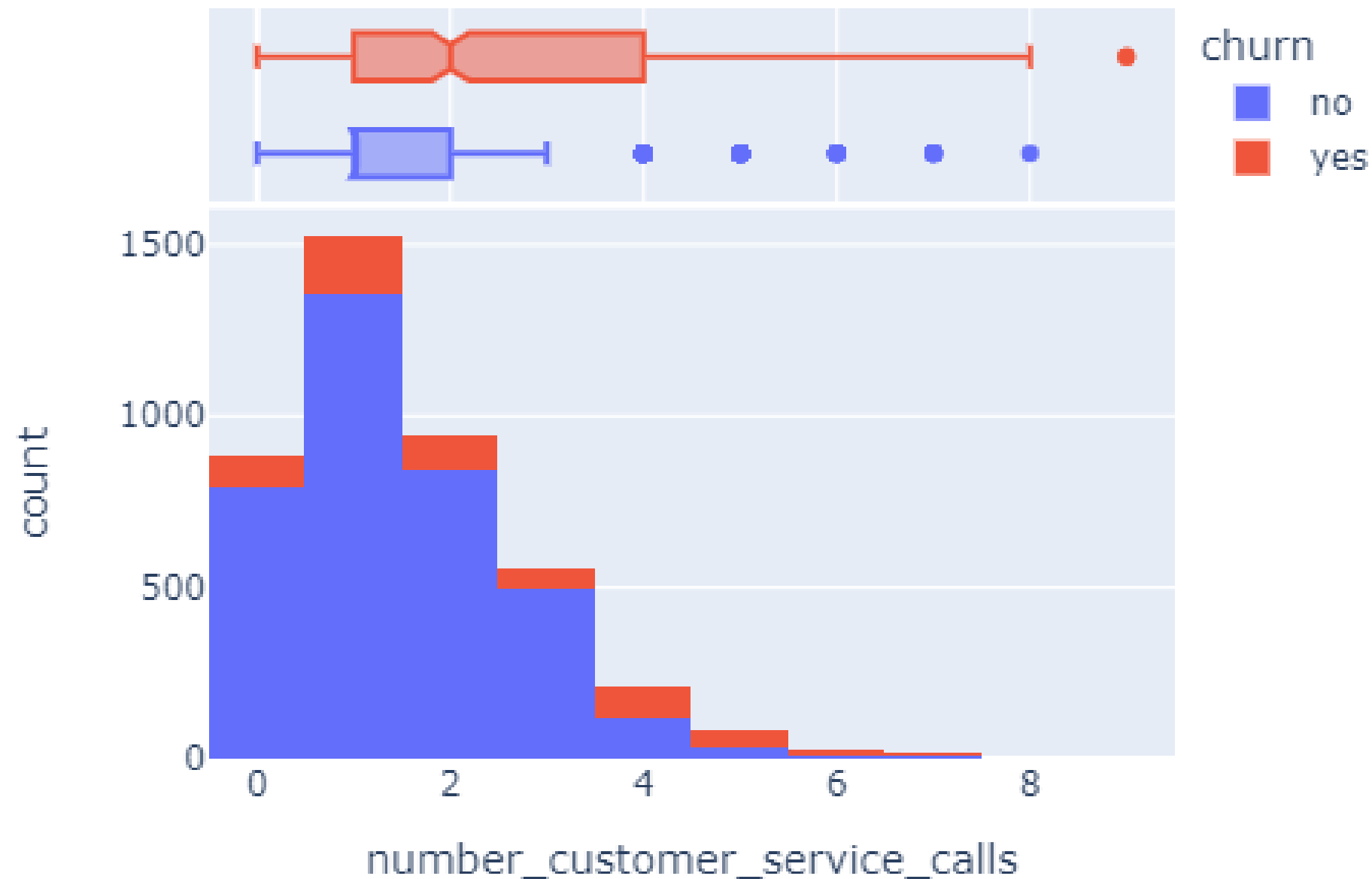
Exploratory Data Analysis (EDA)



Perbandingan Data Churn

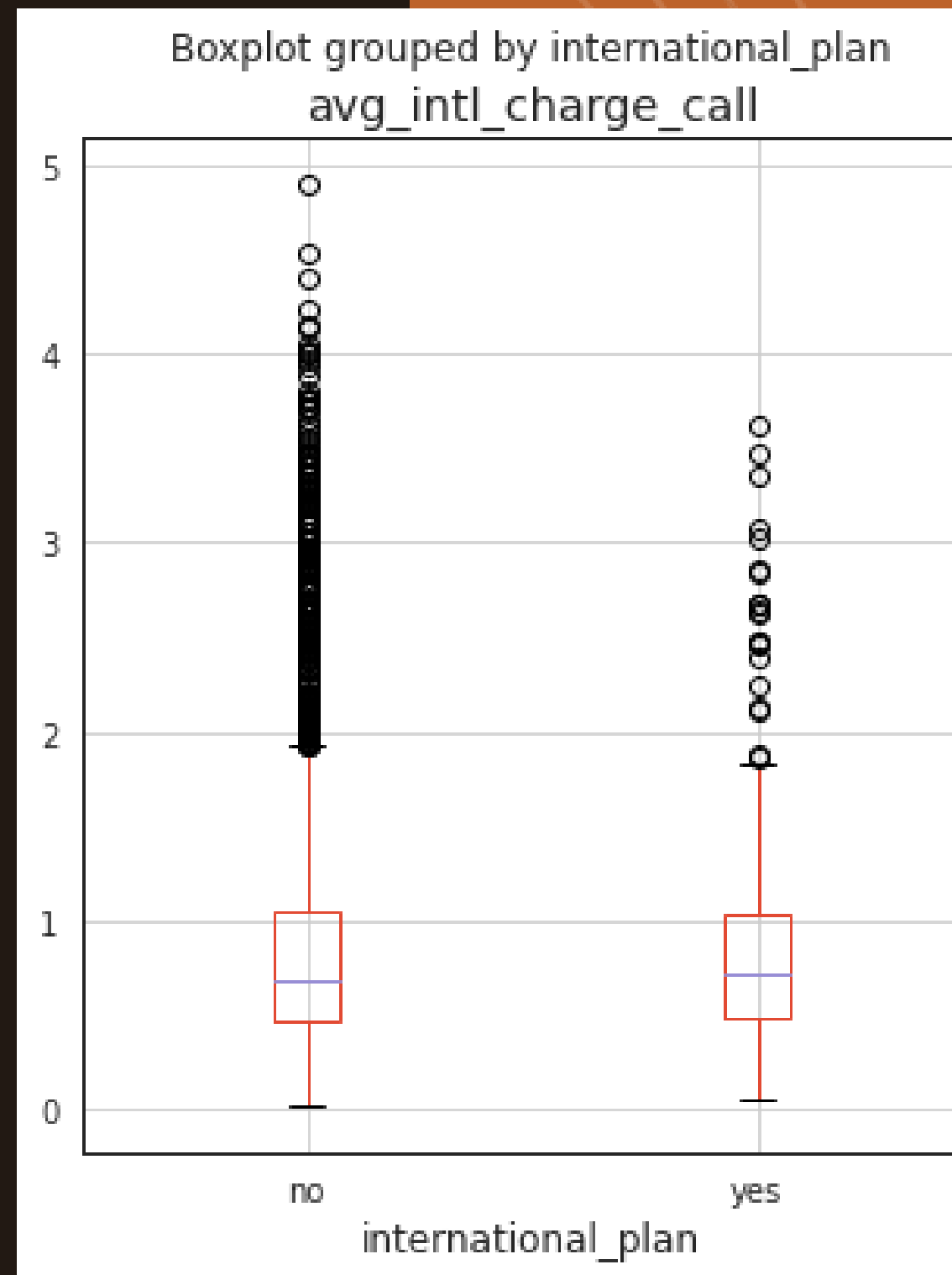
Data Churn yang menjadi target variable sangat imbalance, terdapat perbedaan yang sangat jauh antara jumlah data pelanggan churn dan tidak churn. Hal ini dapat memberikan masalah pada model klasifikasi. Untuk itu perlu dilakukan penanganan terhadap data imbalance ini dengan undersampling atau oversampling



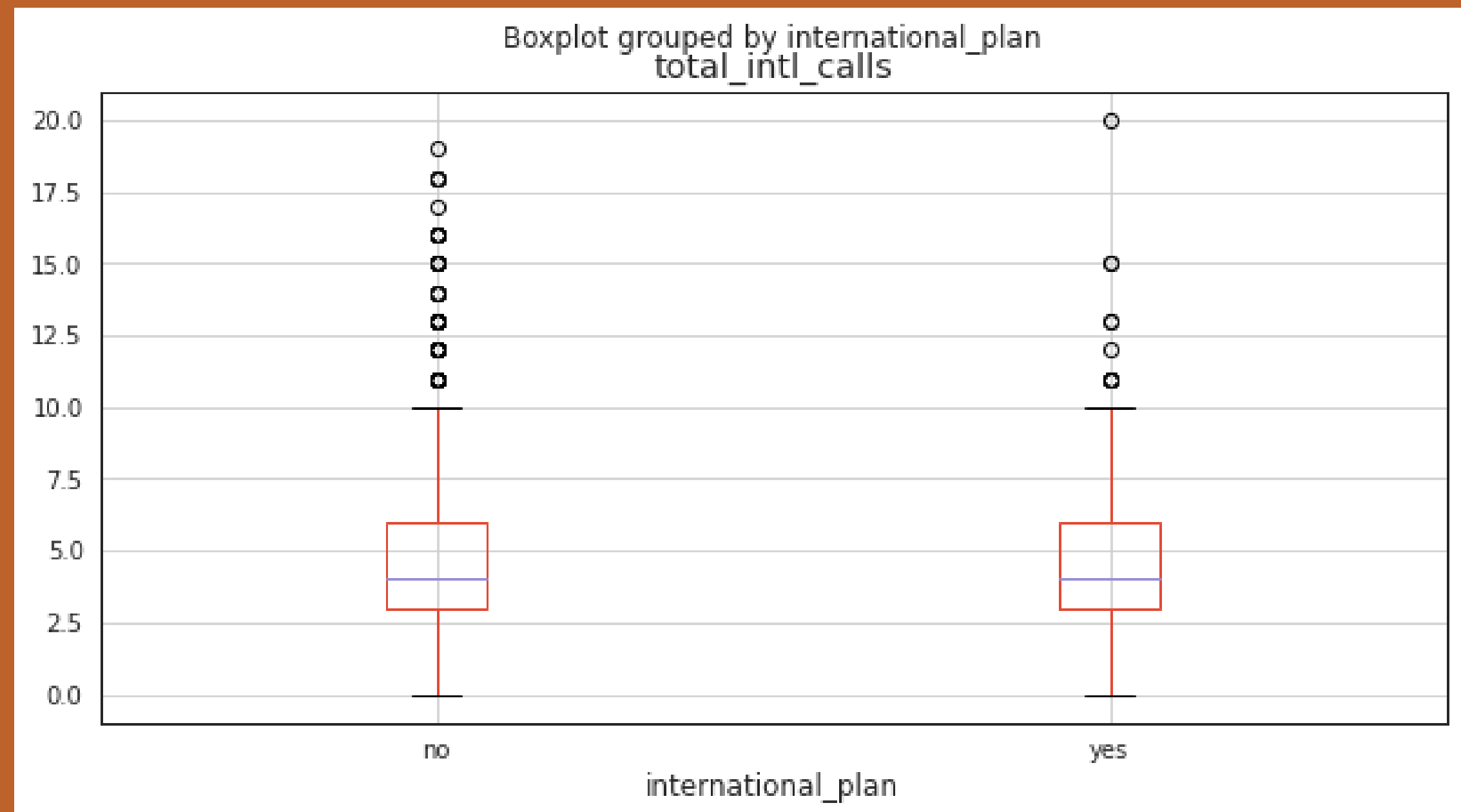


boxplot pelanggan churn (merah) lebih lebar daripada yang tidak churn (biru) , artinya pelanggan yang churn lebih sering menelpon customer service daripada yang tidak churn (lihat boxplot).

median churn adalah 2 (lebih tinggi dari tidak churn, median=1) dengan tail yang lebih panjang artinya jangkauan sebarannya lebih jauh daripada customer service call nya pelanggan churn yang cenderung berkumpul di sekitaran 1-2



Biaya per panggilan dan total panggilan untuk panggilan internasional tanpa plan lebih dan dengan plan tidak menunjukkan perbedaan distribusi. Akan tetapi pada pelanggan yang tidak menggunakan plan terdapat lebih banyak outlier di bagian atas boxplot.





Data Pre- Processing

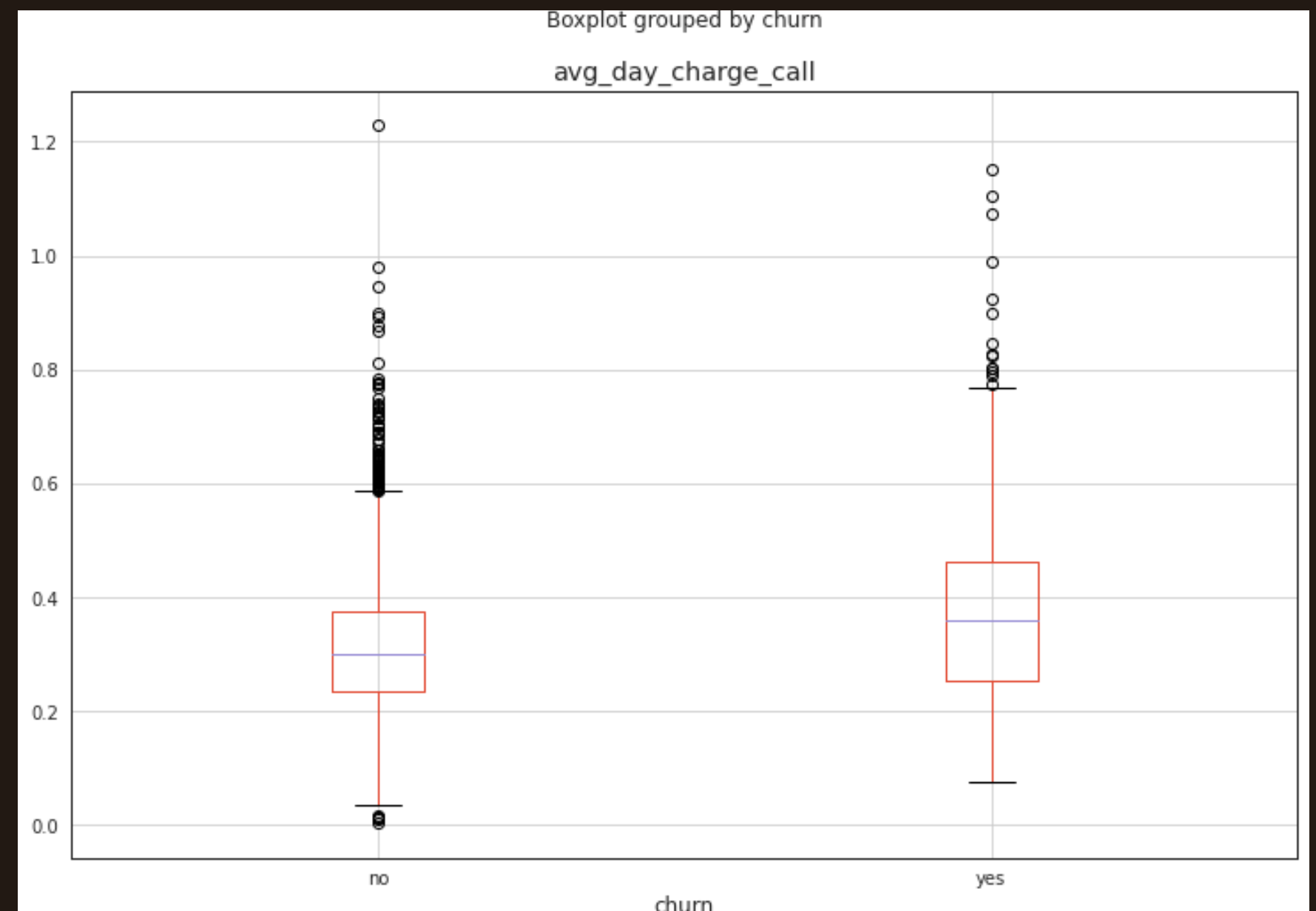


Feature Engineering

1. avg_day_charge_calls

rata-rata biaya untuk tiap panggilan di siang hari
latar: (ah kalau nelson mahal, ganti provider deh)
 $\text{total_day_charge} / \text{total_day_calls}$

dari plot terlihat bahwa pelanggan yang churn punya tail ke atas yang lebih panjang ke atas daripada yang tidak churn. Artinya pelanggan yang churn berkemungkinan memiliki rata-rata biaya panggilan di siang hari yang lebih mahal daripada yang tidak churn.

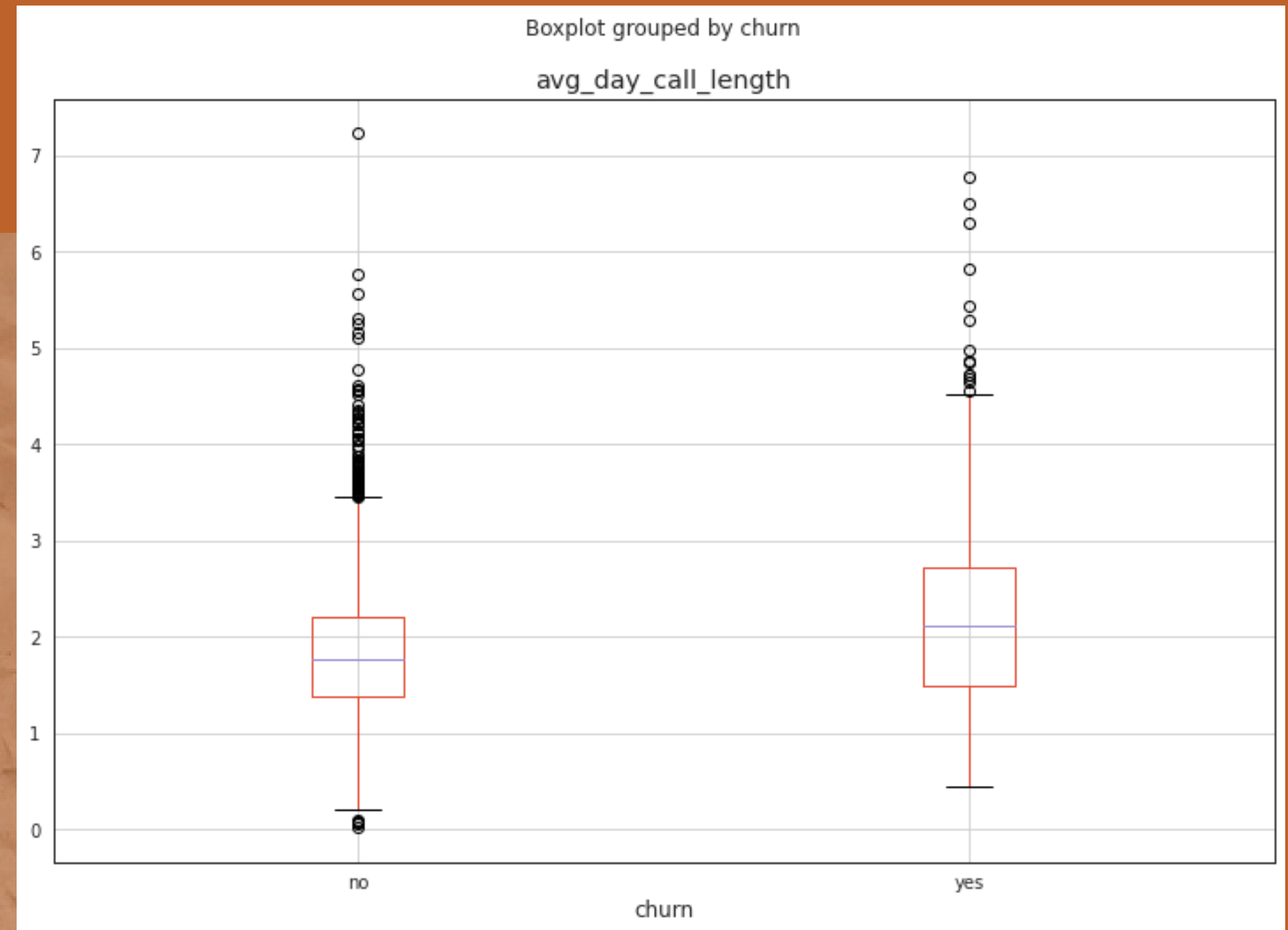


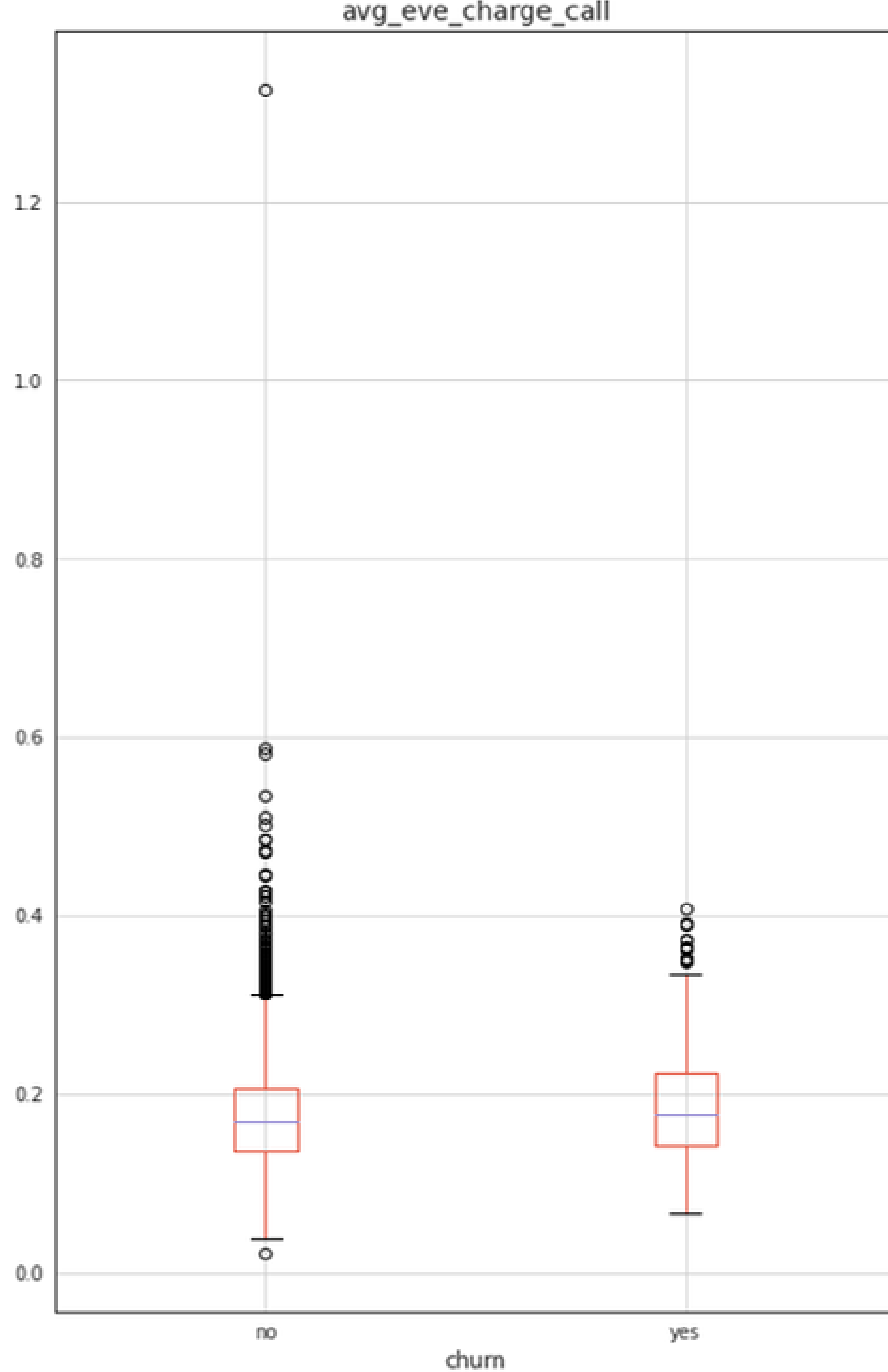
2. avg_day_call_length

rata-rata durasi tiap panggilan di siang hari
latar:

$\text{total_day_minutes} / \text{total_day_calls}$

dari plot terlihat bahwa pelanggan yang churn punya box dan tail yang lebih cenderung ke atas dibandingkan dengan yang tidak churn. Artinya pelanggan yang churn lebih cenderung memiliki durasi panggilan yang lebih lama daripada pelanggan yang tidak churn.





3. avg_eve_charge_calls

rata-rata biaya untuk tiap panggilan di malam hari

latar:

$\text{total_eve_charge} / \text{total_eve_calls}$

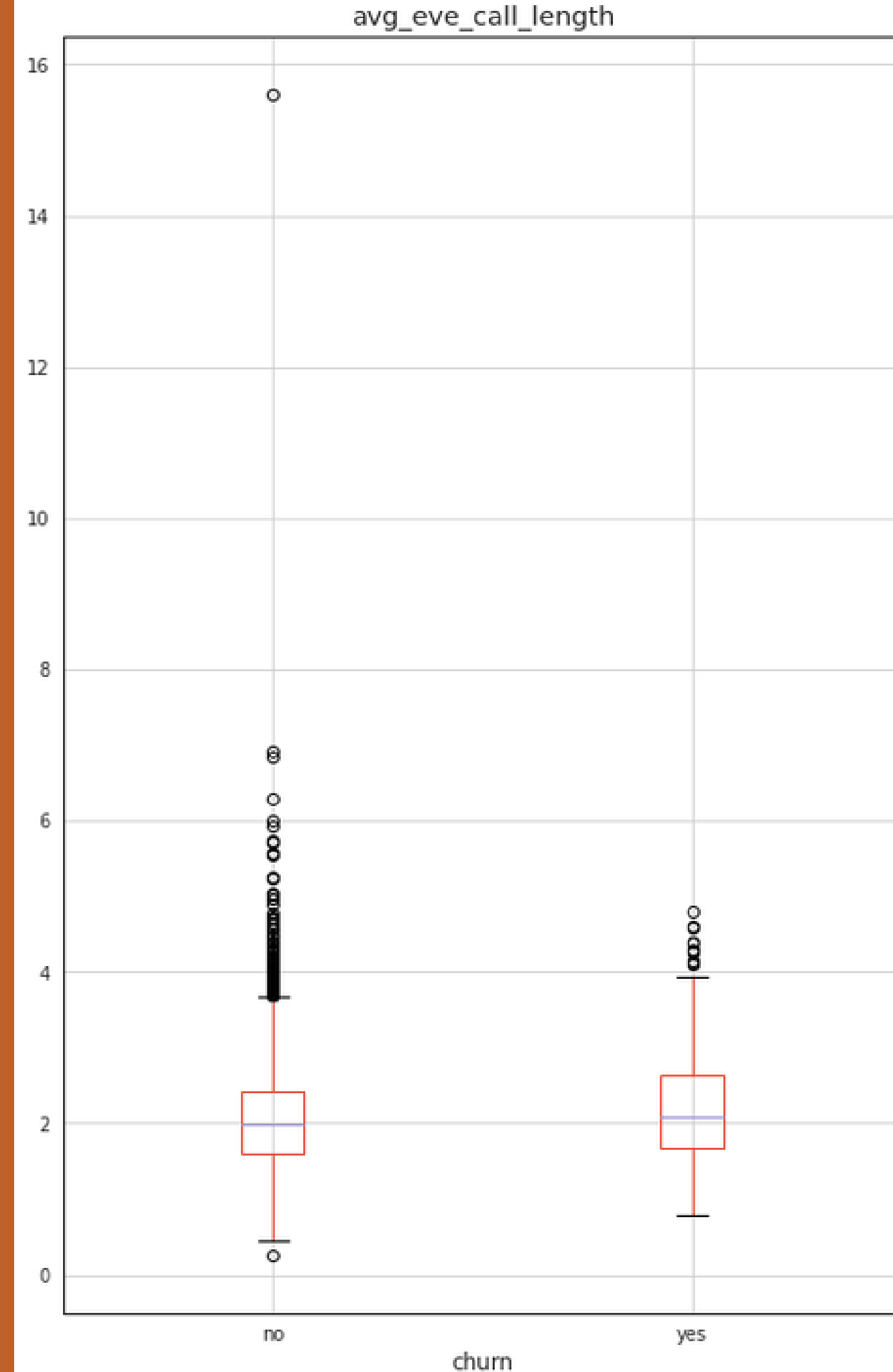
dari plot terlihat bahwa pelanggan yang churn boxplotnya sedikit lebih tinggi daripada yang tidak churn. pelanggan yang tidak churn memiliki banyak outlier yang biaya per panggilannya lebih mahal daripada pelanggan churn. Akan tetapi mayoritas biaya per panggilan sedikit cenderung lebih rendah daripada yang churn

4. avg_eve_call_length

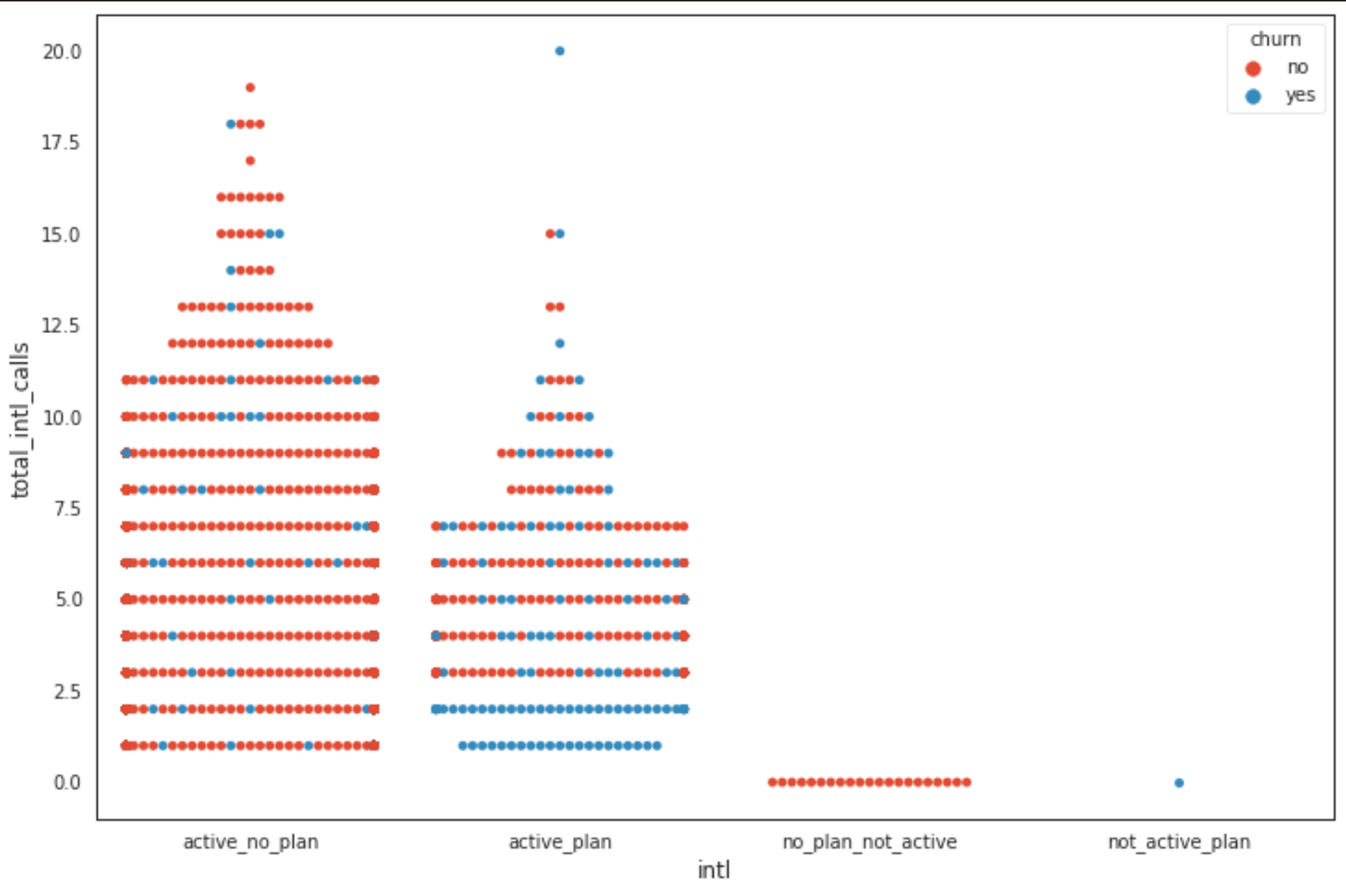
rata-rata durasi tiap panggilan di malamhari
latar:

$\text{total_eve_minutes} / \text{total_eve_calls}$

dari plot terlihat bahwa pelanggan yang churn punya box yang sedikit lebih tinggi dan lebih lebar dari pada yang tidak churn. pelanggan yang tidak churn terdapat banyak outlier yang durasi panggilannya jauh lebih lama dibandingkan pelanggan yang churn



5. International (intl)



pelanggan yang churn cenderung mempunyai plan panggilan internasional dengan pemakaian yang sedikit. Kategorisasi dilakukan dengan aturan sebagai berikut:

total_intl_calls	international_plan	new_cat
0	yes	no_plan_not_active
0	no	not_active_plan
> 0	yes	active_plan
> 0	no	active_plan

Normalization

Fungsi

Normalisasi dilakukan karena data memiliki skala yang berbeda-beda (menit, bulan, uang, dsb) dan agar rentang data tidak terlalu jauh

Library

normalisasi dengan menggunakan MinMaxScaler

Encoding



Mengubah data kategorik menjadi numerik. One Hot Encoding mengubah variabel kategorik yang setiap kategorinya membentuk kolom baru yang berisikan variabel dummy (1 dan 0)

Train-Test Split



Membagi data train menjadi data train dan data validasi untuk menguji kualitas model. Digunakan rasio 7:3



Resampling

(handling imbalance data)

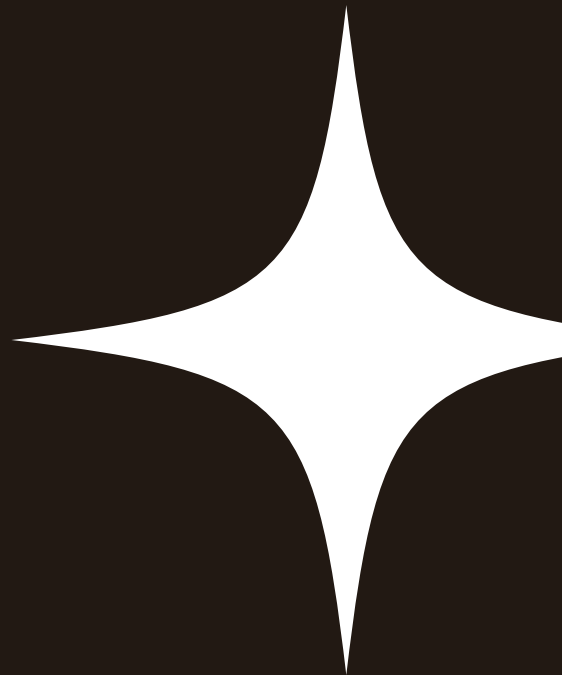
Undersampling

Dilakukan random sampling pada data mayoritas (not churn) sehingga jumlahnya sama dengan data minoritas (churn).
Dengan kata lain data mayoritas dipotong

Oversampling > SMOTE

Dilakukan duplikasi pada data minoritas(churn) sehingga rasio churn dan not churn sama.

Modelling



Model	Tipe Data	roc_auc_score_train			Best
		Tanpa Resampling	Undersampling	Oversampling	
knn	Train	74.61%	87.08%	92.65%	Undersampling
	Test	65.93%	80.39%	78.43%	
Decision Tree	Train	100%	100%	100%	Oversampling
	Test	82.49%	81.61%	83.41%	
Random Forest	Train	100%	100%	100%	All
	Test	83.41%	83.4%	83.4%	
Logistic Regression	Train	58.18%	80.02%	78.33%	Undersampling
	Test	60.31%	78.06%	75.92%	

Kesimpulan



Model knn, decision tree, dan random forest memberikan hasil yang overfitting terutama decision tree dan random forest. Sedangkan score model logistic regression cenderung mirip pada data train dan data validasi

Model terbaik adalah model knn dengan undersampling. Model ini menghasilkan roc_auc_score sebesar 87.8% pada data train dan 80.39% pada data validasi

Thank
You!

