



ibimbing

# IRANIAN CHURN



Presentasi Final Project Data Science Batch 27

**Disusun Oleh: Muhammad Haidi Nurrahman**



# Table of Content



# Project Overview

## Background

Kumpulan data ini dikumpulkan secara acak dari basis data perusahaan telekomunikasi Iran selama 12 bulan. Sebanyak 3150 baris data, masing-masing mewakili seorang pelanggan, mengandung informasi untuk 13 kolom. Atribut yang ada dalam dataset ini adalah kegagalan panggilan, frekuensi SMS, jumlah keluhan, jumlah panggilan yang berbeda, lama berlangganan, kelompok usia, jumlah biaya, jenis layanan, detik penggunaan, status, frekuensi penggunaan, dan Nilai Pelanggan.

Semua atribut kecuali atribut churn adalah data agregat dari 9 bulan pertama. Label churn adalah keadaan pelanggan pada akhir 12 bulan. Tiga bulan adalah kesenjangan perencanaan yang ditentukan.

## Objective

- Membersihkan dan mengeksplorasi dataset untuk mendapatkan insight sederhana data seperti pola pelanggan churn sebelum pemodelan.
- Membangun model prediktif untuk memproyeksikan kemungkinan pelanggan melakukan churn
- Mengidentifikasi Faktor Utama yang Mempengaruhi Churn Pelanggan

## Goals

Membantu bisnis dalam menyusun strategi retention (mempertahankan pelanggan).

# Data Understanding

FEATURE	DESKRIPSI
Call Failures	Jumlah panggilan yang gagal dilakukan pelanggan.
Complains	(0: Tidak ada keluhan, 1: Ada keluhan).
Subscription Length	Total durasi langganan pelanggan dalam bulan.
Charge Amount	Jumlah biaya yang dibebankan kepada pelanggan, dalam bentuk atribut ordinal (0: paling rendah, 10: paling tinggi).
Seconds of Use	Total durasi penggunaan layanan telepon oleh pelanggan dalam detik.
Frequency of Use	Jumlah total panggilan yang dilakukan pelanggan.
Frequency of SMS	Total jumlah pesan teks (SMS) yang dikirim pelanggan.
Distinct Called Numbers	Total jumlah nomor telepon unik yang dihubungi pelanggan.
Age Group	Kelompok usia pelanggan dalam bentuk ordinal (1: usia muda, 5: usia tua).
Tariff Plan	Jenis paket tarif yang dipilih pelanggan. (1: Pay as you go, 2: Contractual)
Age	Umur customer
Status	Biner (1: Aktif, 2: Tidak aktif).
Customer Value	Nilai pelanggan yang dihitung berdasarkan aktivitas mereka.
Churn	<b>Biner (1: churn, 0: tidak churn).</b>

- Data memiliki 3150 baris dan 14 kolom
- 14 kolom bernilai numeric semuanya 13 int , 1 float
- Churn adalah kolom target untuk prediksi

**Target**

# Data Cleaning

## Missing Value

Call Failure	0
Complains	0
Subscription Length	0
Charge Amount	0
Seconds of Use	0
Frequency of use	0
Frequency of SMS	0
Distinct Called Numbers	0
Age Group	0
Tariff Plan	0
Status	0
Age	0
Customer Value	0
Churn	0

## Duplicated

Data Duplication : 300

Data memiliki 300 duplikat( duplikat disini efek dari penghapusan kolom customer id yang sudah terhapus terlebih dahulu)

# Feature Selection

**14 FEATURE**

Correlation Heatmap														
Call Failure -	Complains -	Subscription Length -	Charge Amount -	Seconds of Use -	Frequency of use -	Frequency of SMS -	Distinct Called Numbers -	Age Group -	Tariff Plan -	Status -	Age -	Customer Value -	Churn -	Call Failure
1.00	0.15	0.17	0.59	0.50	0.57	-0.02	0.50	0.05	0.19	-0.11	0.04	0.12	-0.01	Call Failure
0.15	1.00	-0.02	-0.03	-0.10	-0.09	-0.11	-0.06	0.02	0.00	0.27	0.00	-0.13	0.53	Complains
0.17	-0.02	1.00	0.08	0.12	0.11	0.08	0.09	0.02	-0.16	0.14	-0.00	0.11	-0.03	Subscription Length
0.59	-0.03	0.08	1.00	0.45	0.38	0.09	0.42	0.28	0.32	-0.36	0.28	0.17	-0.20	Charge Amount
0.50	-0.10	0.12	0.45	1.00	0.95	0.10	0.68	0.02	0.13	-0.46	0.02	0.42	-0.30	Seconds of Use
0.57	-0.09	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	0.21	-0.45	-0.03	0.40	-0.30	Frequency of use
-0.02	-0.11	0.08	0.09	0.10	0.10	1.00	0.08	-0.05	0.20	-0.30	-0.09	0.92	-0.22	Frequency of SMS
0.50	-0.06	0.09	0.42	0.68	0.74	0.08	1.00	0.02	0.17	-0.41	0.05	0.28	-0.28	Distinct Called Numbers
0.05	0.02	0.02	0.28	0.02	-0.03	-0.05	0.02	1.00	-0.15	0.00	0.96	-0.18	-0.01	Age Group
0.19	0.00	-0.16	0.32	0.13	0.21	0.20	0.17	-0.15	1.00	-0.16	-0.12	0.25	-0.11	Tariff Plan
-0.11	0.27	0.14	-0.36	-0.46	-0.45	-0.30	-0.41	0.00	-0.16	1.00	-0.00	-0.41	0.50	Status
0.04	0.00	-0.00	0.28	0.02	-0.03	-0.09	0.05	0.96	-0.12	-0.00	1.00	-0.22	-0.02	Age
0.12	-0.13	0.11	0.17	0.42	0.40	0.92	0.28	-0.18	0.25	-0.41	-0.22	1.00	-0.29	Customer Value
-0.01	0.53	-0.03	-0.20	-0.30	-0.30	-0.22	-0.28	-0.01	-0.11	0.50	-0.02	-0.29	1.00	Churn
Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn	Call Failure



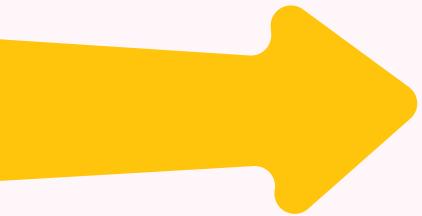
Correlation Heatmap														
Call Failure -	Complains -	Subscription Length -	Charge Amount -	Seconds of Use -	Frequency of use -	Frequency of SMS -	Distinct Called Numbers -	Age Group -	Tariff Plan -	Status -	Age -	Customer Value -	Churn -	Call Failure
1.00	0.15	0.17	0.59	0.50	0.57	-0.02	0.50	0.05	0.19	-0.11	0.04	0.12	-0.01	Call Failure
0.15	1.00	-0.02	-0.03	-0.10	-0.09	-0.11	-0.06	0.02	0.00	0.27	0.00	-0.13	0.53	Complains
0.17	-0.02	1.00	0.08	0.12	0.11	0.08	0.09	0.02	-0.16	0.14	-0.00	0.11	-0.03	Subscription Length
0.59	-0.03	0.08	1.00	0.45	0.38	0.09	0.42	0.28	0.32	-0.36	0.28	0.17	-0.20	Charge Amount
0.57	-0.09	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	0.21	-0.45	-0.03	0.40	-0.30	Seconds of Use
0.57	-0.09	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	0.21	-0.45	-0.03	0.40	-0.30	Frequency of use
-0.02	-0.11	0.08	0.09	0.10	0.10	1.00	0.08	-0.05	0.20	-0.30	-0.09	0.92	-0.22	Frequency of SMS
0.50	-0.06	0.09	0.42	0.68	0.74	0.08	1.00	0.02	0.17	-0.41	0.05	0.28	-0.28	Distinct Called Numbers
0.05	0.02	0.02	0.28	0.02	-0.03	-0.05	0.02	1.00	-0.15	0.00	0.96	-0.18	-0.01	Age Group
0.19	0.00	-0.16	0.32	0.13	0.21	0.20	0.17	-0.15	1.00	-0.16	-0.12	0.25	-0.11	Tariff Plan
-0.11	0.27	0.14	-0.36	-0.46	-0.45	-0.30	-0.41	0.00	-0.16	1.00	-0.00	-0.41	0.50	Status
0.04	0.00	-0.00	0.28	0.02	-0.03	-0.09	0.05	0.96	-0.12	-0.00	1.00	-0.22	-0.02	Age
0.12	-0.13	0.11	0.17	0.42	0.40	0.92	0.28	-0.18	0.25	-0.41	-0.22	1.00	-0.29	Customer Value
-0.01	0.53	-0.03	-0.20	-0.30	-0.30	-0.22	-0.28	-0.01	-0.11	0.50	-0.02	-0.29	1.00	Churn
Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn	Call Failure

Dengan threshold 0.8 terlihat ada beberapa feature yang berkorelasi tinggi

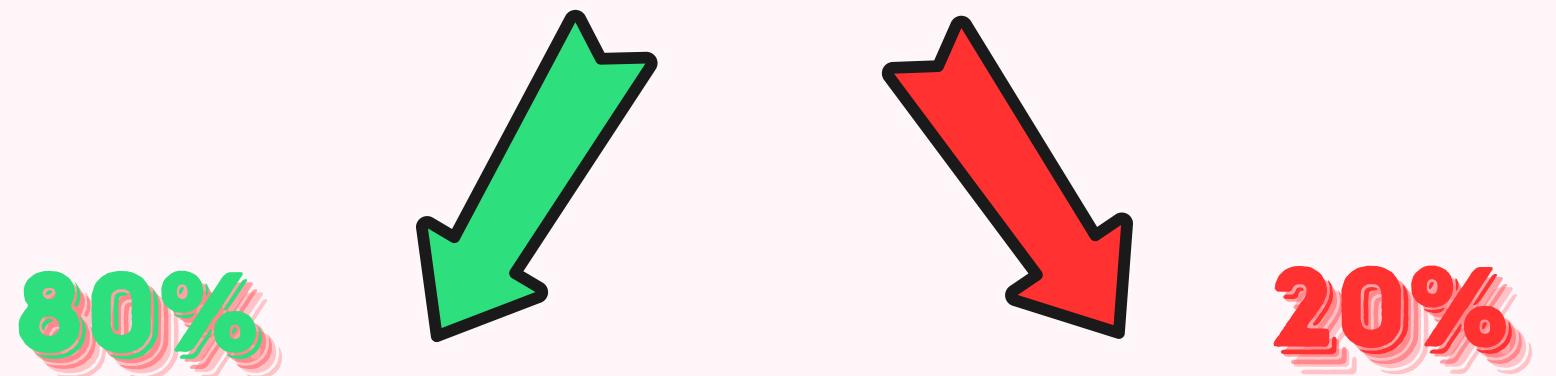
- Customer Value dan Frequency of SMS : 0.92 disini akan kita drop kolom customer value karena kolom ini tidak menjelaskan kolom secara rinci tentang perilaku pelanggan.
- Age Group dan Age : 0.96 pada kolom ini kita hapus kolom Age Group karena sudah terwakilkan oleh kolom Age yang memiliki informasi lebih rinci dari pada kolom age group.
- Second of Use dan Frequency of Use : 0.95 pada kolom ini kita akan mendrop kolom Second of use karena Dalam konteks telekomunikasi, mengetahui seberapa sering layanan digunakan lebih relevan untuk menganalisis keterlibatan pelanggan.

# Preprocessing Data

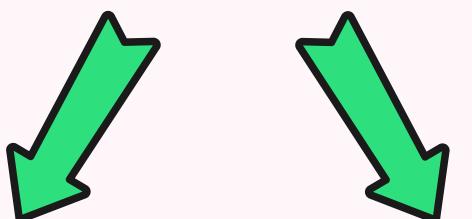
**Split Dataset**



**Scaling**



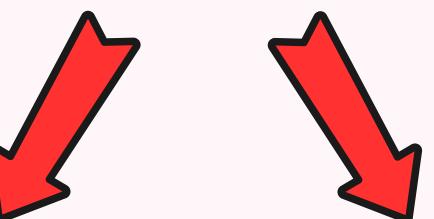
**Data Train**



**X\_train**

**y\_train**

**Data Test**



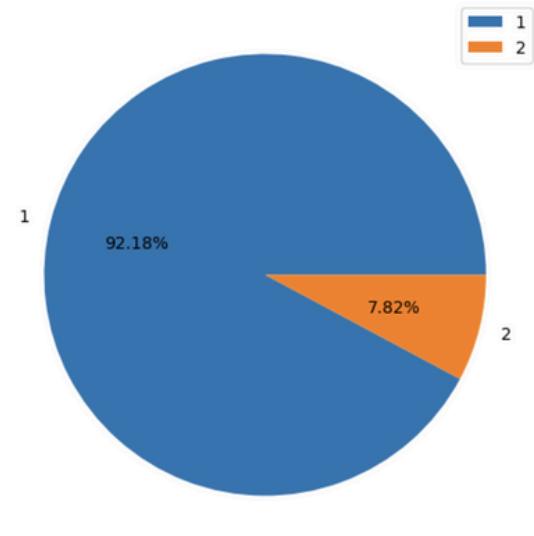
**X\_test**

**y\_test**

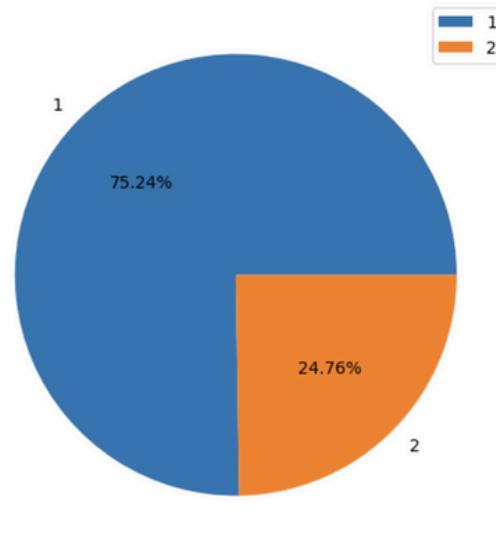
```
scaler = RobustScaler()  
X_train_scaled=scaler.fit_transform(X_train)  
X_test_scaled = scaler.transform(X_test)
```

# Basic EDA

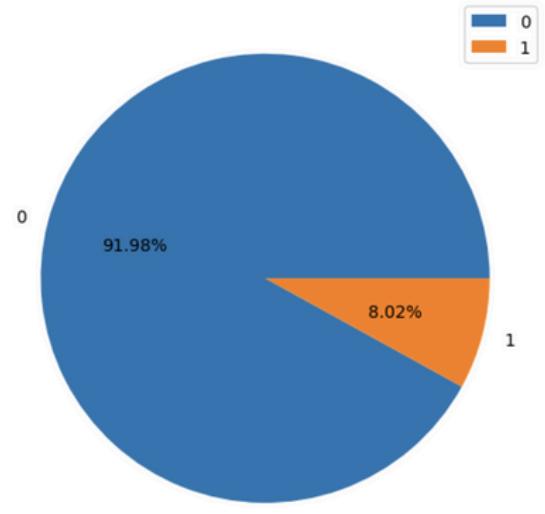
Distribution of Tariff Plan



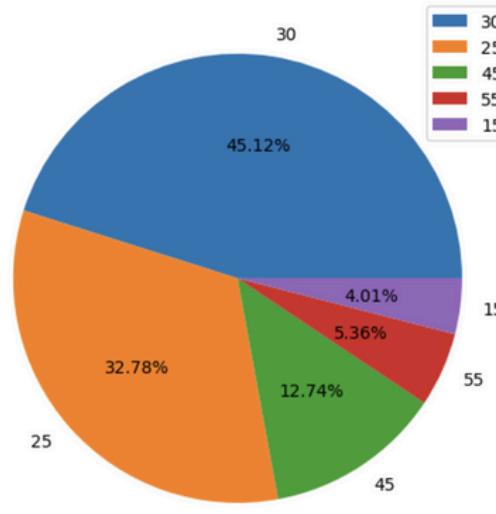
Distribution of Status



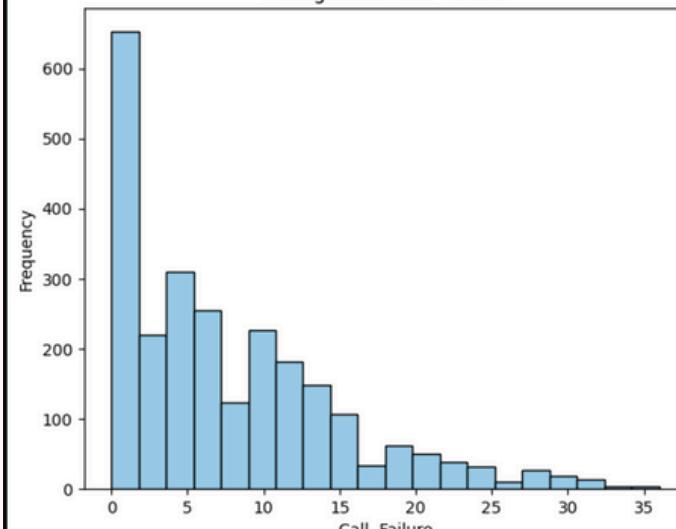
Distribution of Complains



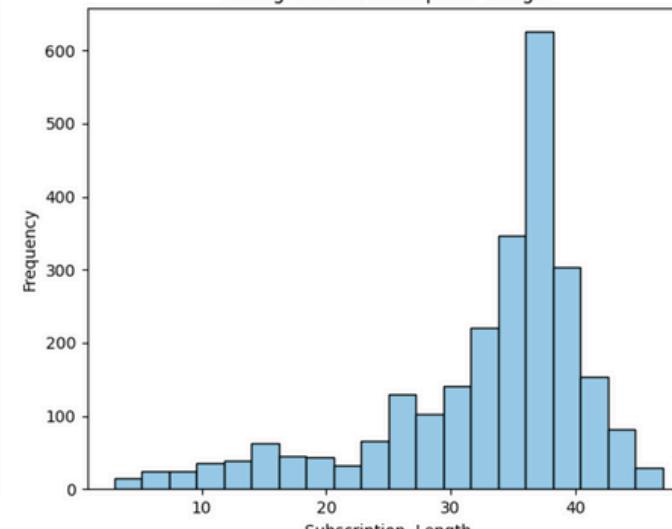
Distribution of Age



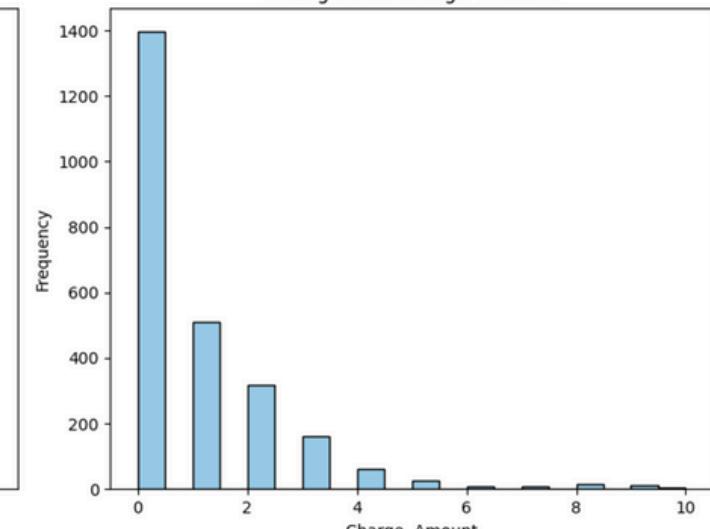
Histogram of Call Failure



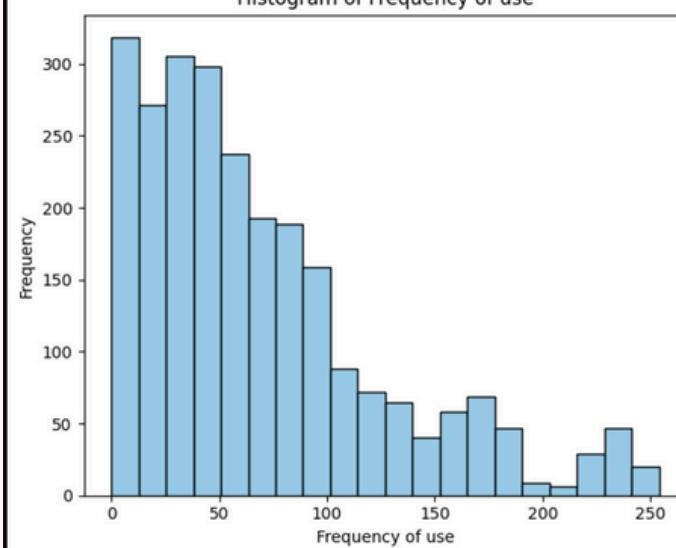
Histogram of Subscription Length



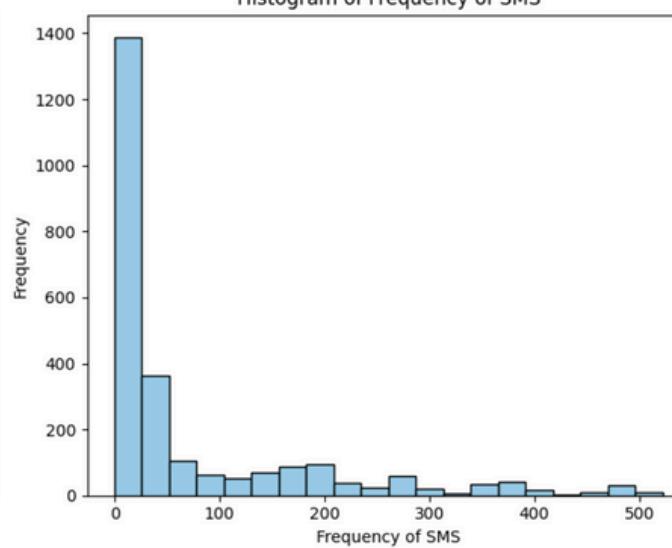
Histogram of Charge Amount



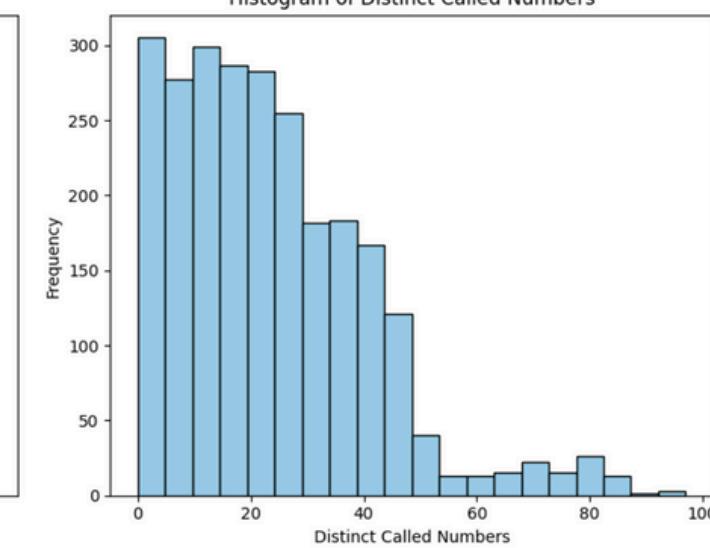
Histogram of Frequency of use



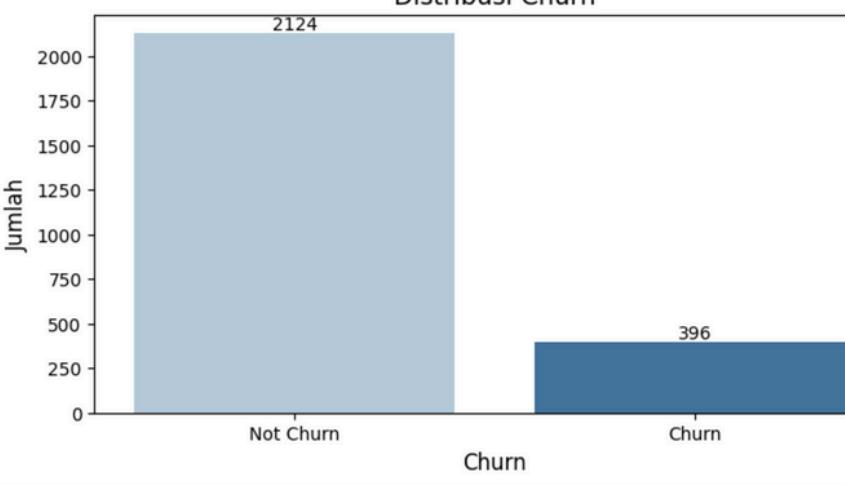
Histogram of Frequency of SMS



Histogram of Distinct Called Numbers



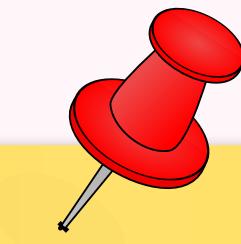
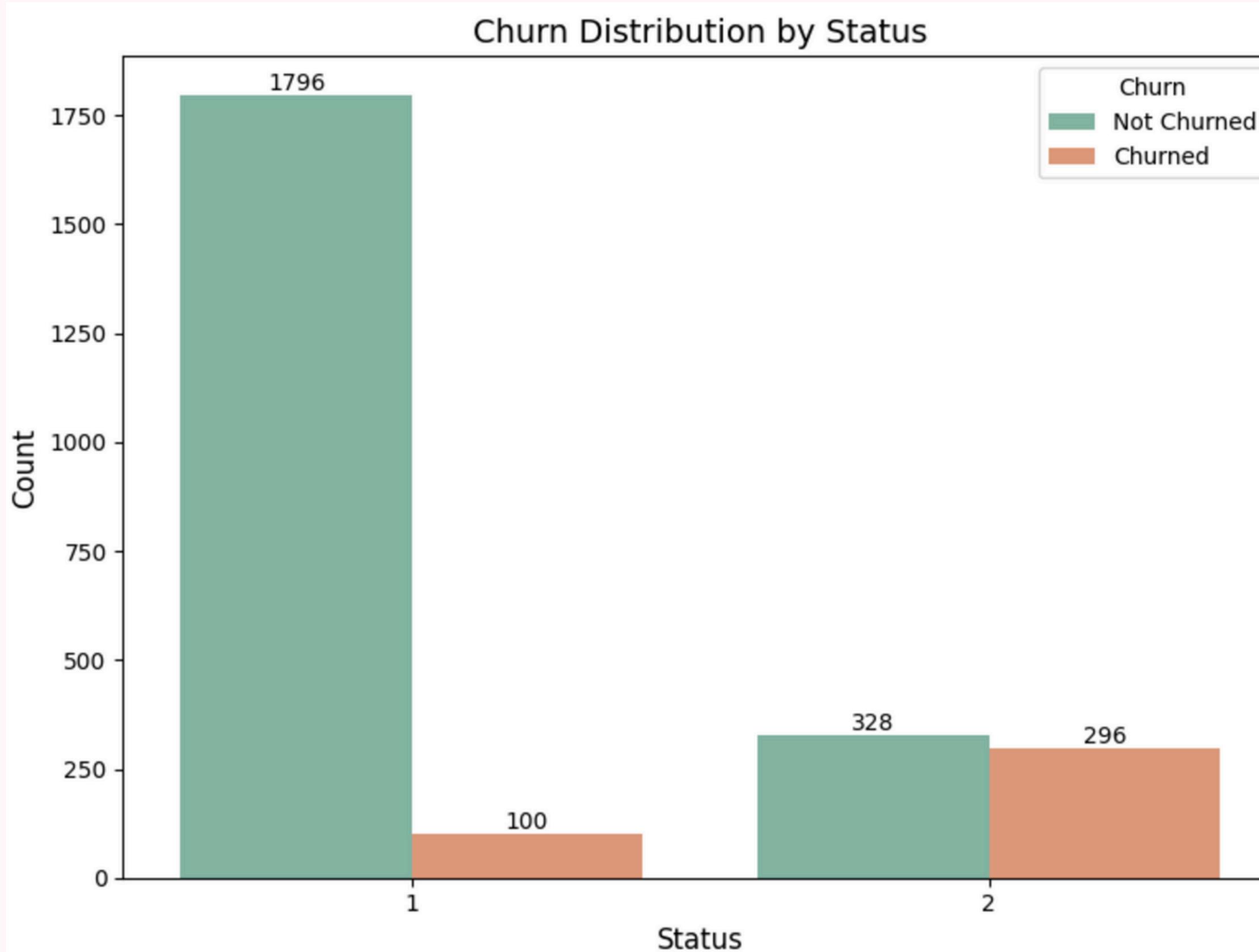
Distribusi Churn



- Dari data yang terlihat , customer lebih banyak yang tidak churn dari pada churn data termasuk imbalanced
- majoritas customer ada di umur 25 dan 30
- customer lebih dominan di label 1 (pay as you go), customer lebih suka menggunakan layanan dengan intensitas rendah karena mereka tidak terikat biaya bulanan tetap.
- jumlah complain lebih banyak yang tidak daripada yang complain , hal yang wajar mengingat distribusi tarif customer lebih dominan yang pay as you go dimana sistem ini sangat fleksibel dimana mereka hanya membayar untuk apa yang digunakan
- untuk distribusi status distribusi sebesar 75% untuk 1(active) dan 25% untuk 2(non-active) mayoritas customer masih aktif

# Deep Dive

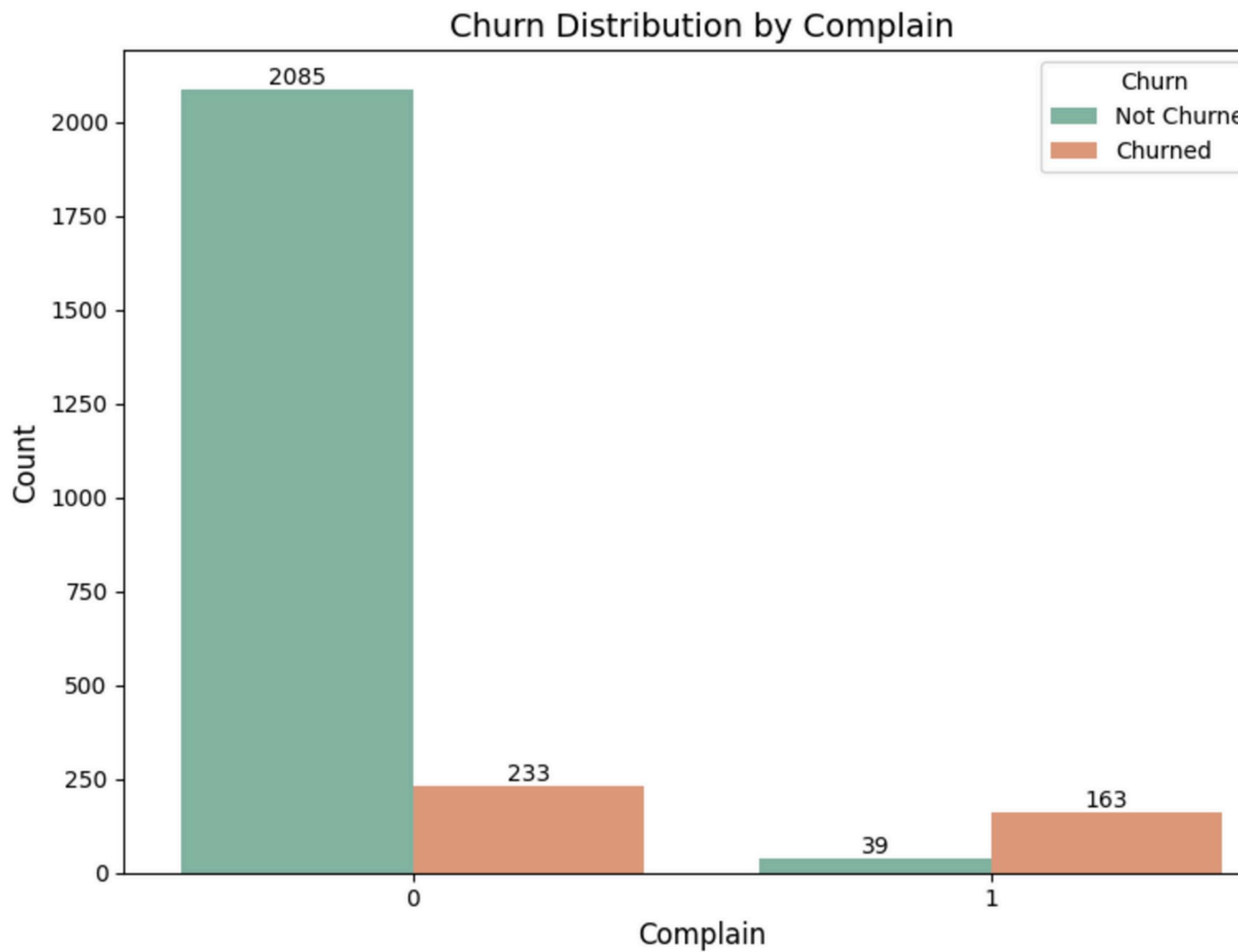
# Apakah semua pelanggan yang tidak aktif pasti Churn?



## INSIGHT :

- Pelanggan dengan status aktif (1) memiliki kemungkinan sangat rendah untuk churn. Ini menunjukkan bahwa pelanggan yang aktif lebih terlibat dengan layanan dan lebih cenderung loyal.
- Pelanggan dengan status tidak aktif (2) memiliki kemungkinan sangat tinggi untuk churn. Ini menunjukkan bahwa pelanggan tidak aktif adalah kelompok yang paling berisiko untuk churn.

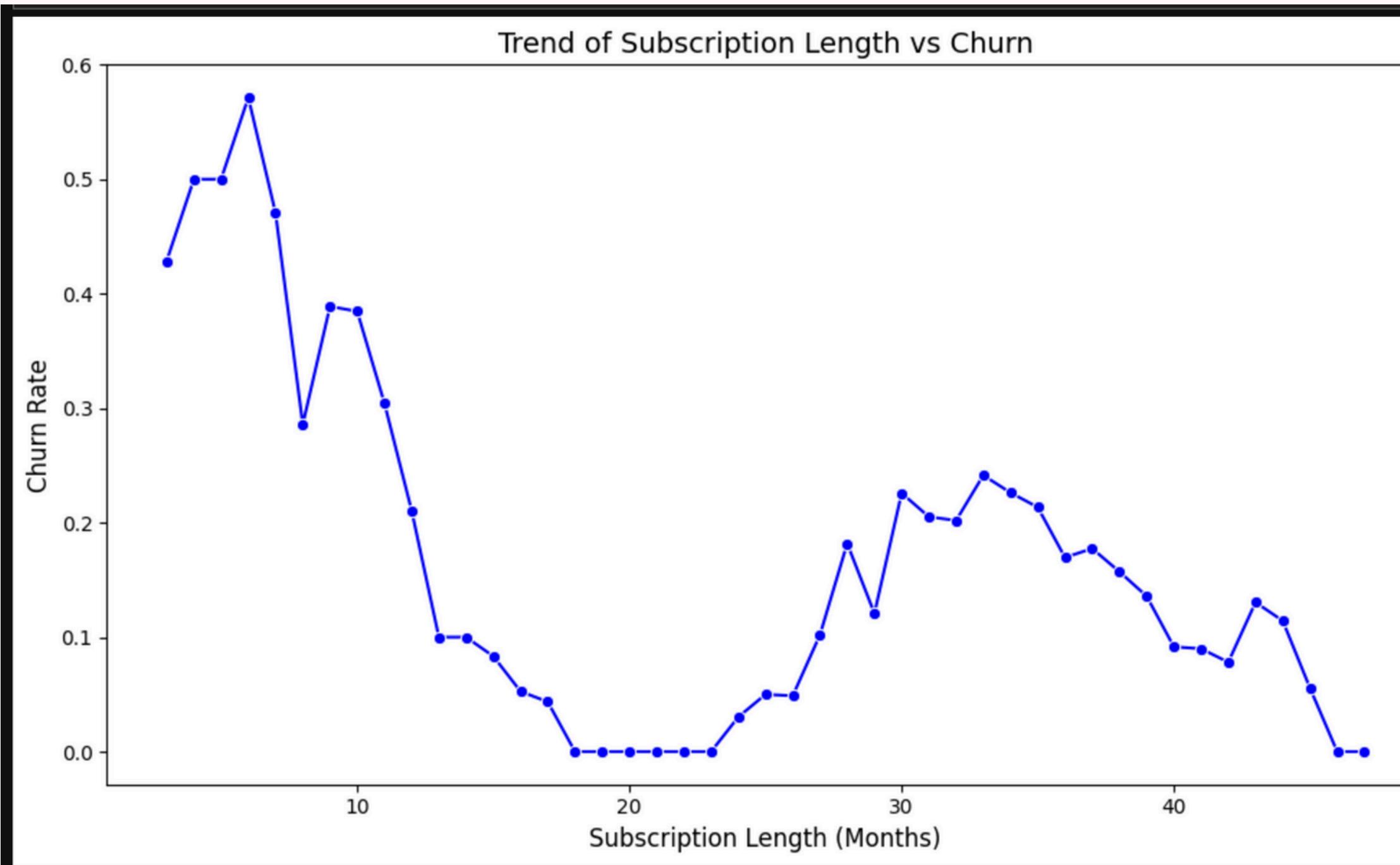
# Apakah pelanggan yang sering mengajukan keluhan lebih mungkin churn?



**INSIGHT :**

- Keluhan dan Churn Terkait Positif: Pelanggan yang mengeluh lebih cenderung untuk churn, yang menyoroti pentingnya menangani keluhan pelanggan dengan cepat dan efektif untuk mencegah churn.
- Pelanggan Tanpa Keluhan Bisa Juga Churn: Walaupun pelanggan tidak mengeluh, mereka masih bisa churn, yang menunjukkan bahwa faktor selain keluhan langsung, seperti harga atau pengalaman pengguna secara keseluruhan, dapat mempengaruhi keputusan churn.

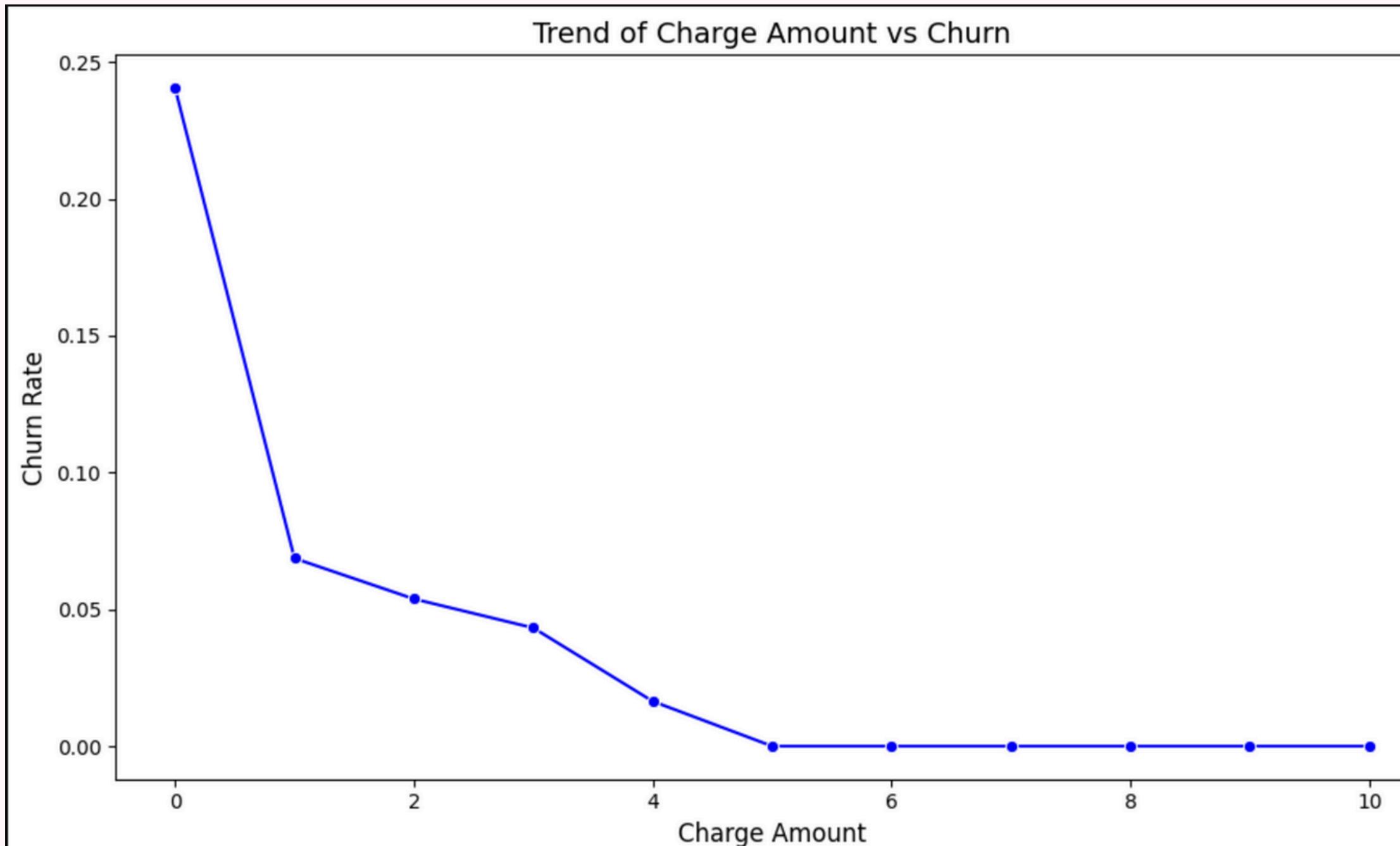
# Tren pola churn berdasarkan durasi langganan.



## INSIGHT :

Secara keseluruhan, rasio churn tampaknya berkurang seiring dengan bertambahnya waktu langganan. Hal ini bisa menunjukkan bahwa semakin lama pelanggan berlangganan, semakin besar kemungkinan mereka akan tetap bertahan, mungkin karena mereka sudah terbiasa dengan layanan atau merasa mendapatkan nilai lebih.

# Pola churn pelanggan berdasarkan tingkat biaya yang dihabiskan pelanggan



## INSIGHT :

Charge Amount = 0: Pelanggan yang tidak mengeluarkan uang memiliki rasio churn tertinggi

Charge Amount = 1-4: Ketika pengeluaran meningkat sedikit, rasio churn cenderung menurun secara signifikan.

(Charge Amount  $\geq 5$ ) menunjukkan rasio churn yang sangat rendah atau bahkan 0%. Ini menunjukkan bahwa pelanggan yang lebih banyak mengeluarkan uang cenderung lebih puas atau lebih terikat pada layanan tersebut,

# Modelling

# Modelling

## Interpretasi :

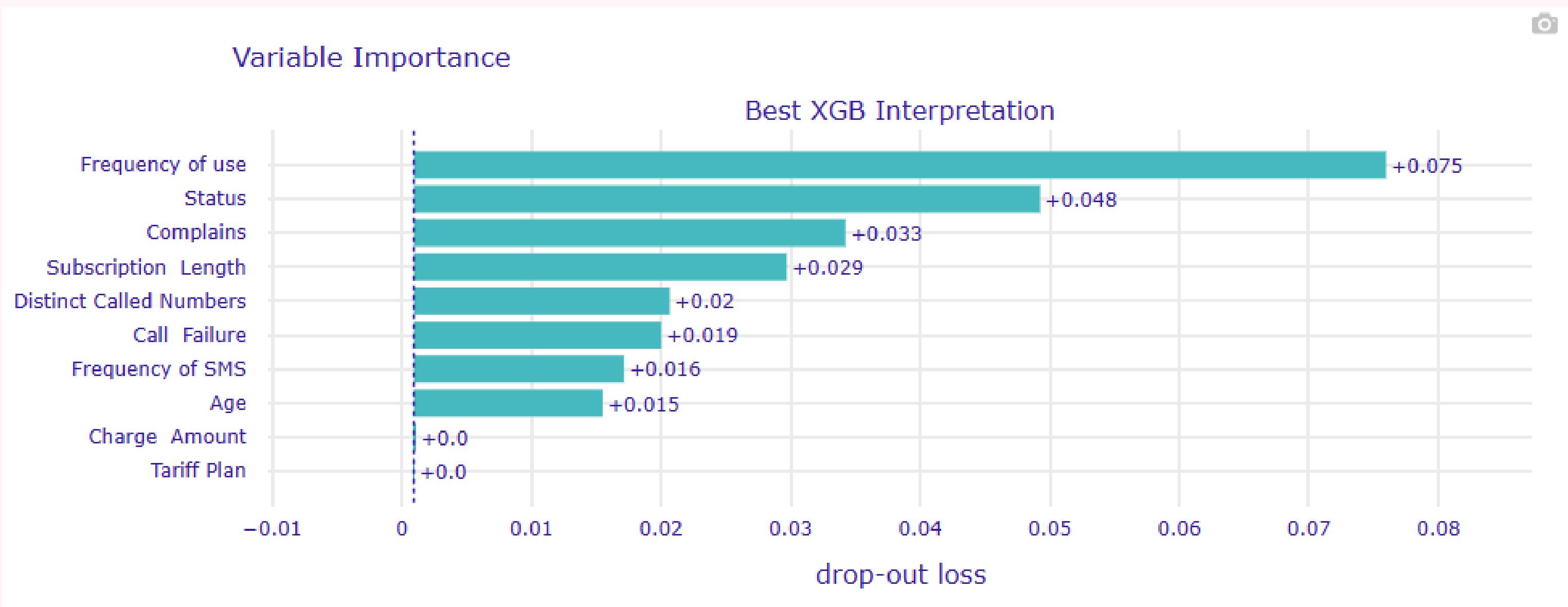
Model	Data Train F1 Score	Data Test F1 Score	ROC-AUC
Logistic Regression	54%	55%	91%
Random Forest	97%	87%	98%
XGboost	96%	87%	99%

1. Evaluasi Score model berdasarkan f1 score model dikarenakan ingin mempertahankan kinerja optimal di kedua sisi kesalahan tanpa terlalu fokus pada salah satunya."
2. Perbedaan data train dan tes tidak sampai lebih dari 10% karena mengindikasikan overfitting
3. XGboost best model
4. ROC-AUC XGboost model lebih bagus dari model yang lain, ini menunjukkan bahwa model sangat baik dalam membedakan antara kelas positif (churn) dan kelas negatif (tidak churn).
5. makna roc-auc 99% ini kita contoh kan : misal kita punya 100 data 60 tidak churn dan 40 yang churn , nah karena model 99% maka, 39 dari 40 churn diklasifikasikan dengan benar sebagai churn. Hanya 1 dari 60 non-churn salah diklasifikasikan sebagai churn.



# Tuning Best Model

Model	Data Train F1 Score	Data Test F1 Score	ROC-AUC	Feature Importance
XGBoost	95%	91%	99%	Frequency of use, Status, Complains, dan Subscription Length



Highlight top  
4 produk

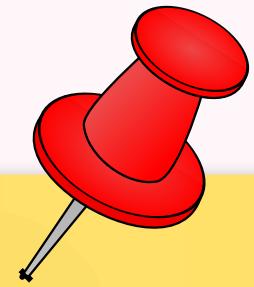


# Conclusion



- Model XGBoost yang telah dioptimasi dengan hyperparameter tuning menunjukkan performa yang sangat baik. Dengan F1-Score tinggi dan ROC-AUC mendekati 1.0,
- F1 Score test model meningkat sebanyak 4 %
- beda train test turun menjadi 4 %
- model ini dapat digunakan untuk mengambil tindakan preventif terhadap pelanggan yang berisiko churn
- Fitur yang berpengaruh terhadap prediksi churn mencakup: Frequency of Use, Status, Complains, dan Subscription Length. Fitur-fitur ini memiliki kontribusi besar terhadap keputusan model dalam memprediksi churn,

# Recomendation



- Kirim kampanye reaktivasi kepada pelanggan tidak aktif, seperti diskon atau bonus.
- Berikan penghargaan kepada pelanggan aktif dan royal untuk mendorong keterlibatan yang berkelanjutan.
- Lebih fokus pada retensi pelanggan baru dengan meningkatkan pengalaman mereka serta memahami manfaat layanan agar tetap bertahan lebih lama,
- Perkenalkan fitur atau layanan baru yang dapat meningkatkan nilai layanan untuk pelanggan.
- Buat survei atau sistem umpan balik untuk memastikan keluhan pelanggan telah diselesaikan dengan memuaskan.
- Percepat waktu penyelesaian keluhan pelanggan untuk meningkatkan kepuasan, Identifikasi penyebab utama keluhan yang berulang dan lakukan perbaikan pada produk atau layanan.





# Link Dataset

<https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset>

# Link Collab(Code)

<https://drive.google.com/file/d/1OmKQ3zS-V1a0gUVq2VKpMMdlzQrX3jlr/view?usp=sharing>



# Terima Kasih

