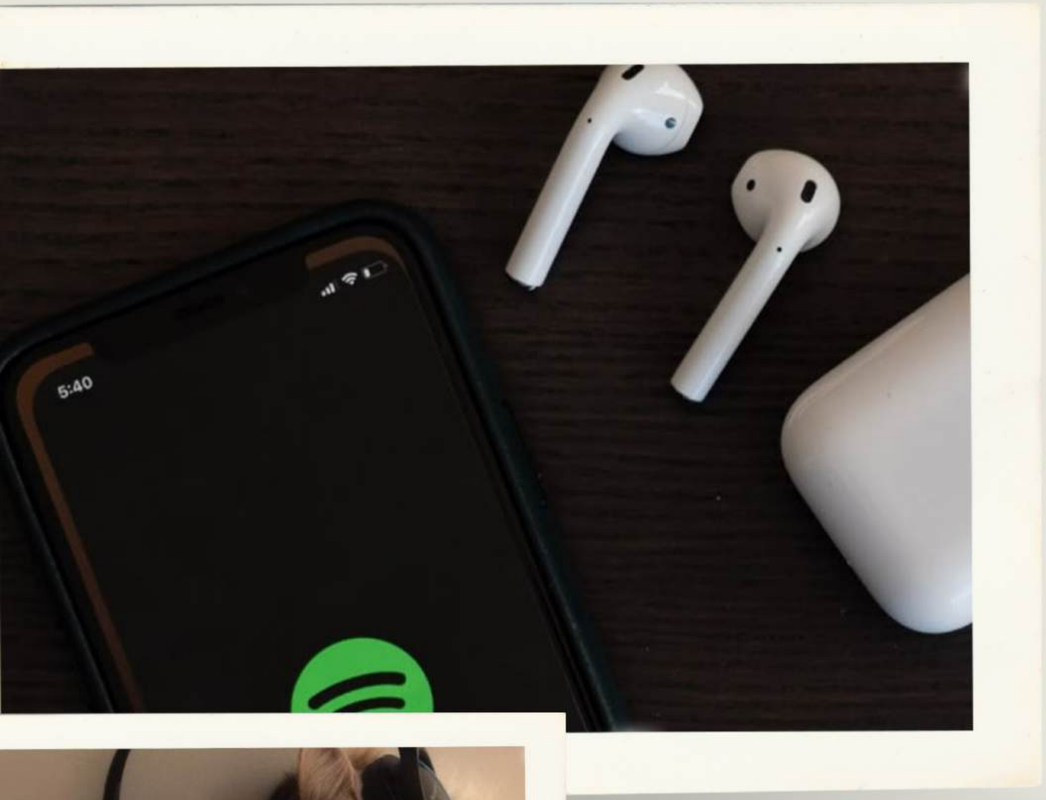
Machine Learning

Feature Engineering for Customer Churn Prediction in Music Streaming Apps

by : Faiz Haidar Halwi



Background

Dalam industri layanan digital seperti aplikasi streaming musik, kehilangan pelanggan (customer churn) secara tiba-tiba dapat menghambat pertumbuhan dan merugikan pendapatan.

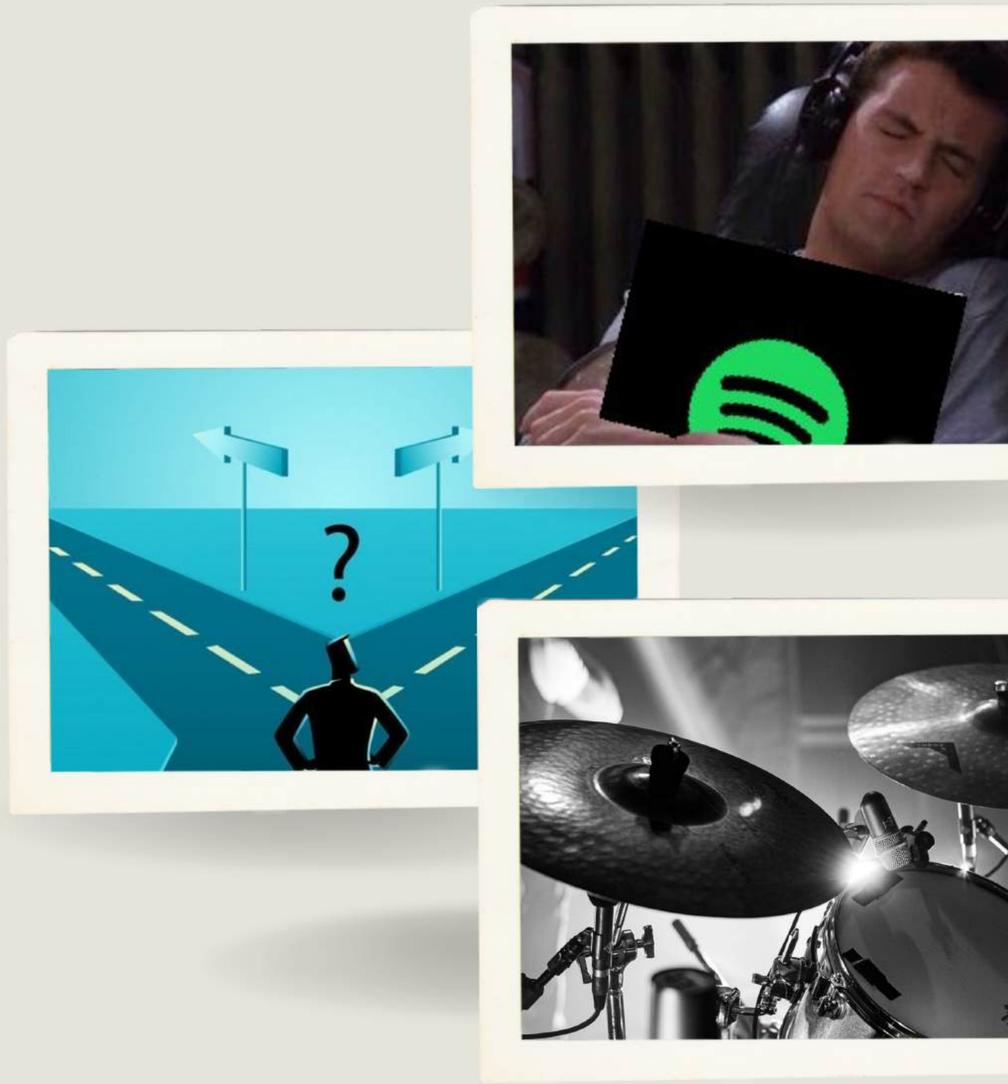
Data aktivitas pengguna yang terekam secara historis menyimpan pola penting terkait perilaku sebelum pelanggan memutuskan untuk berhenti. Dengan pendekatan feature engineering, pola tersebut bisa diekstraksi menjadi fitur-fitur bermakna untuk memprediksi kemungkinan churn.

Masalah

1. Banyak pengguna aktif harian/mingguan berhenti menggunakan aplikasi tanpa peringatan sebelumnya.

2. Data mentah dari log aktivitas, preferensi musik, dan interaksi pengguna bersifat kompleks, berisik (noise), dan tidak langsung mencerminkan potensi churn.

3. Dibutuhkan proses feature engineering yang sistematis untuk mengubah data mentah menjadi fitur-fitur yang relevan untuk prediksi churn.



Goals

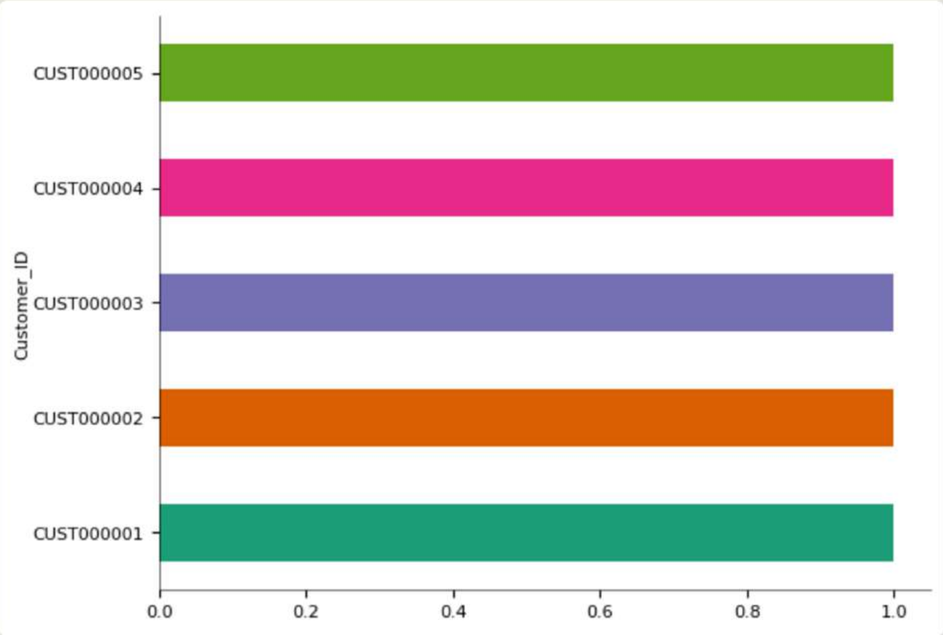
1 Membersihkan dan membentuk ulang data aktivitas pengguna agar siap digunakan untuk model prediksi churn.

2 Data mentah dari log aktivitas, preferensi musik, dan interaksi pengguna bersifat kompleks, berisik (noise), dan tidak langsung mencerminkan potensi churn.

3 Menerapkan teknik-teknik seperti: imputasi nilai kosong, encoding variabel kategorikal, normalisasi data numerik, serta penciptaan fitur perilaku dan temporal.

“Ketiga goals telah dicapai sesuai alur proses dan keterbatasan dataset yang tersedia.”

Dataset Overview



“data 5 teratas”

Customer_ID	Age	Gender
Subscription_Length	Region	Payment_Method
Support_Tickets_Raised	Satisfaction_Score	Discount_Offered
Last_Activity	Monthly_Spend	Churned

“data kolom”

kaggle

“sumber dataset”

Dataset Overview

```
Jumlah baris dan kolom: (5000, 12)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer_ID             5000 non-null   object
 1   Age                     4500 non-null   float64
 2   Gender                  5000 non-null   object
 3   Subscription_Length     5000 non-null   int64
 4   Region                  5000 non-null   object
 5   Payment_Method          5000 non-null   object
 6   Support_Tickets_Raised  5000 non-null   int64
 7   Satisfaction_Score      4500 non-null   float64
 8   Discount_Offered        5000 non-null   float64
 9   Last_Activity           5000 non-null   int64
 10  Monthly_Spend           5000 non-null   float64
 11  Churned                 5000 non-null   int64
```

Total :

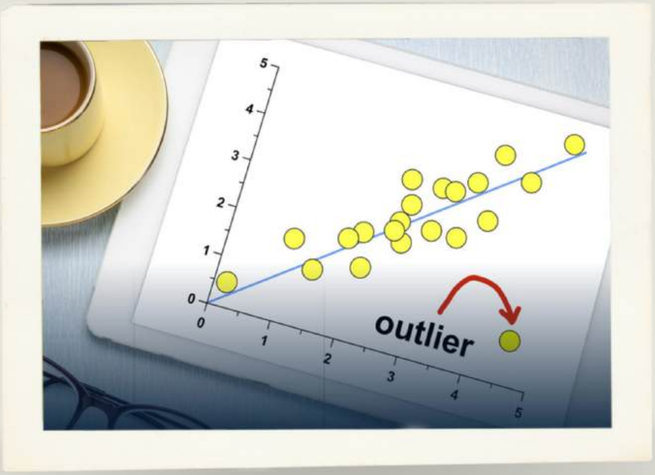
Baris berjumlah 5000

Kolom berjumlah 12

Data Cleaning

Missing Value

Duplicate Data



Menangani outliers

Missing Value & Duplicate Data

```
[ ]  # Missing values
     print("Jumlah missing values per kolom:")
     print(df.isnull().sum())

     # Duplikat
     print(f"\nJumlah data duplikat: {df.duplicated().sum()}")
```

```
Jumlah missing values per kolom:
Customer_ID                0
Age                      500
Gender                     0
Subscription_Length        0
Region                     0
Payment_Method             0
Support_Tickets_Raised     0
Satisfaction_Score       500
Discount_Offered           0
Last_Activity              0
Monthly_Spend              0
Churned                    0
dtype: int64

Jumlah data duplikat: 0
```

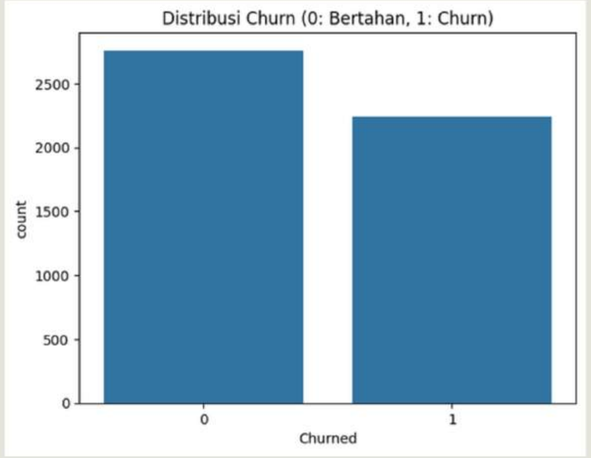
Missing Value :

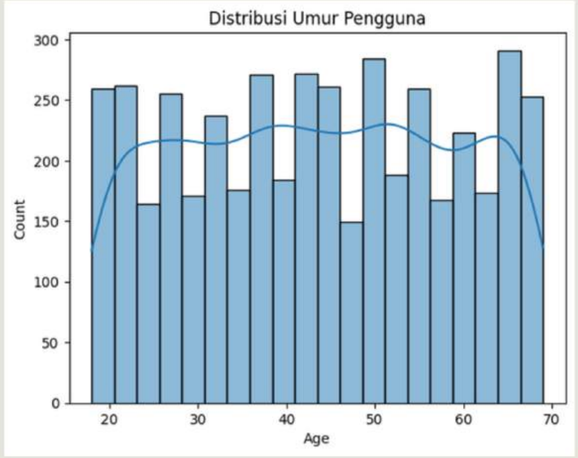
- Age: 500 nilai kosong
- Satisfaction_Score: 500 nilai kosong
- Kolom lainnya tidak memiliki missing values.

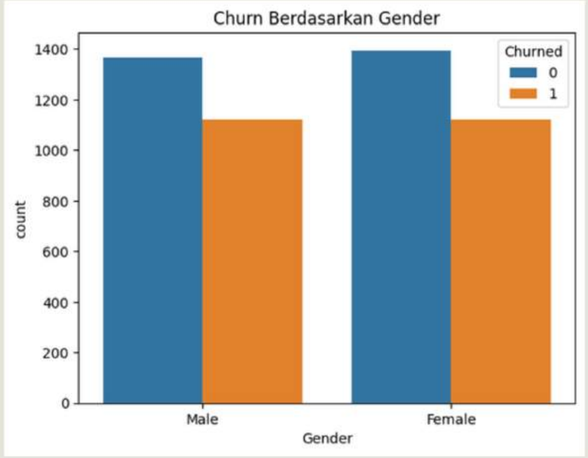
Data Duplicate :

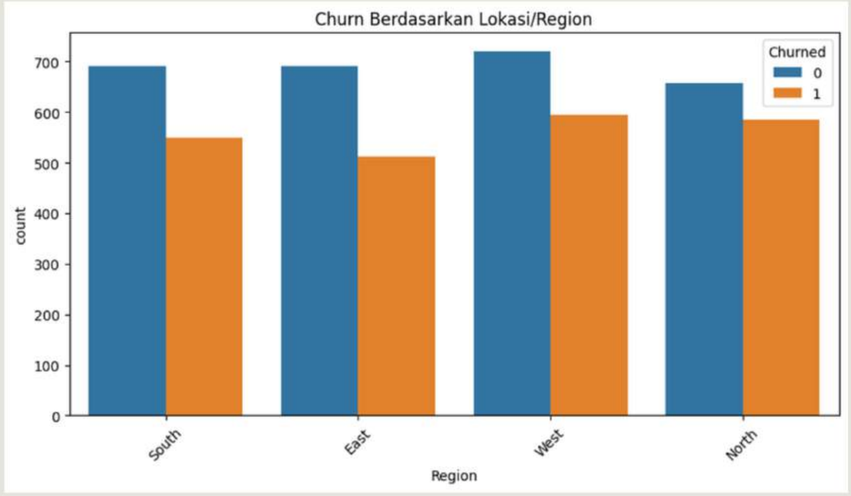
Hasil nya tidak ada data duplicate

Exploratory Data Analysis (EDA) Sederhana



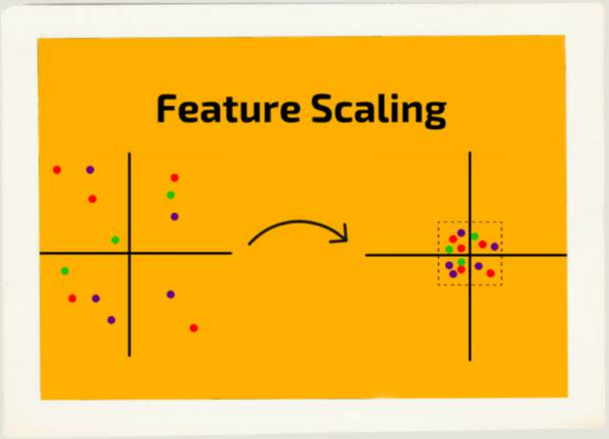






Feature Engineering

Encoding variabel kategorikal



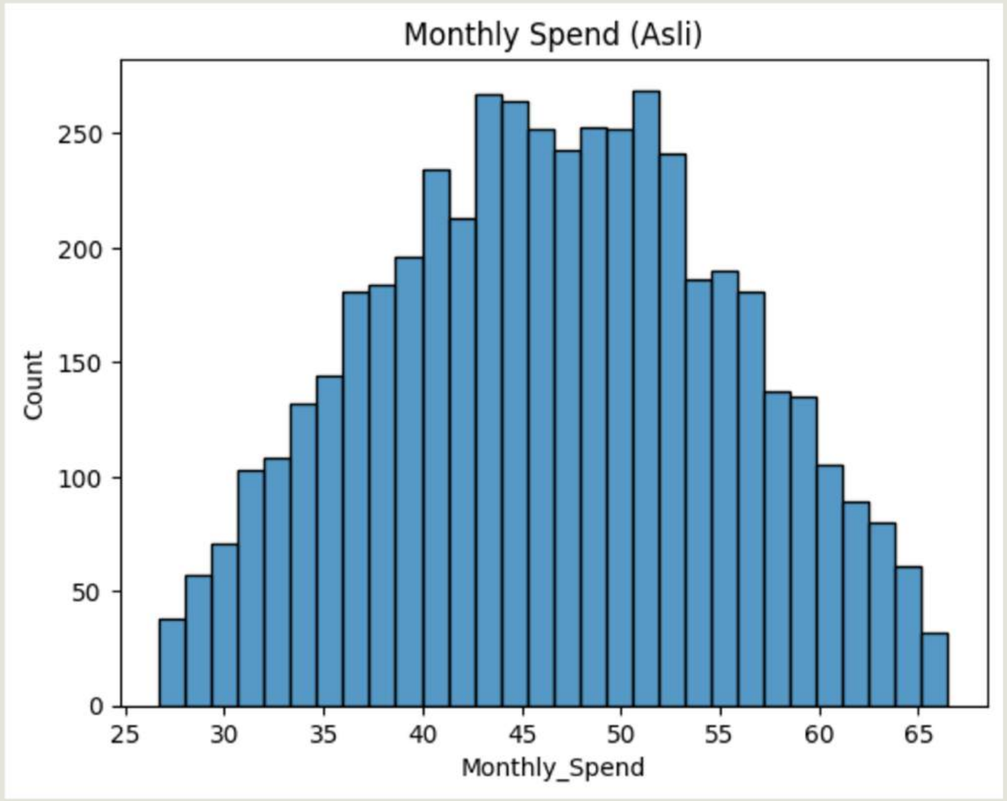
Scaling fitur numerik

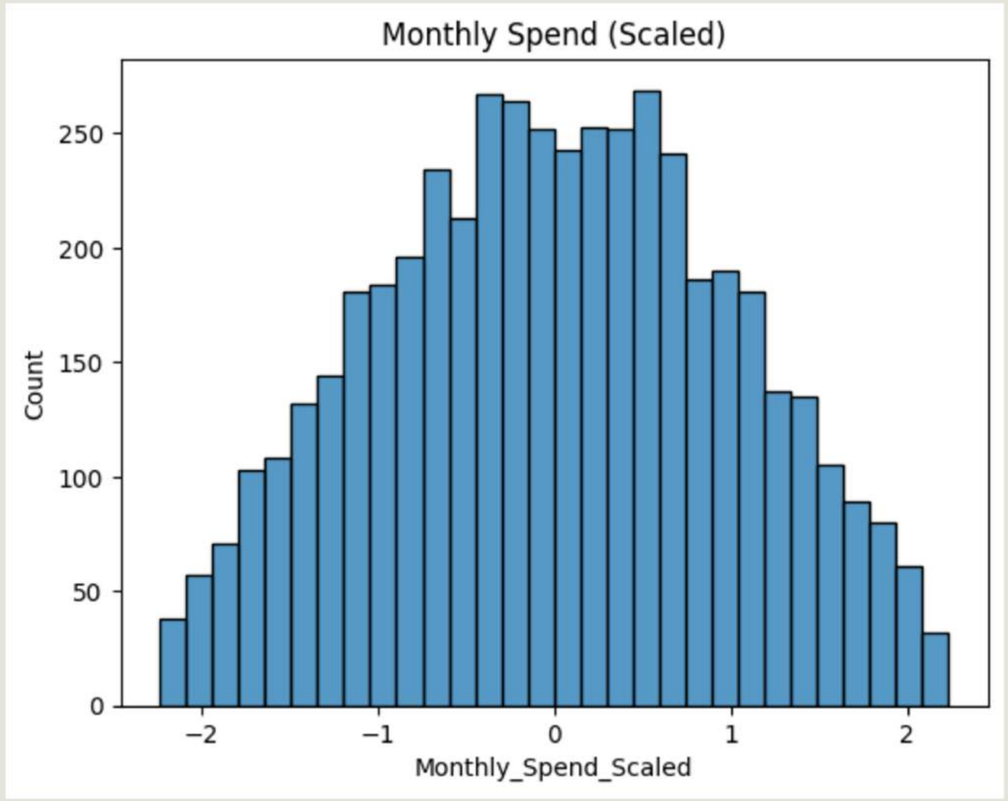
Fitur yang dibuat :

- Avg_Listen_Per_Session
- Usage_Score
- Days_Since_Last_Active (simulasi)

Catatan fitur tidak tersedia: genre, aktivitas granular, dll

Visualisasi Sebelum vs Sesudah Feature Engineering





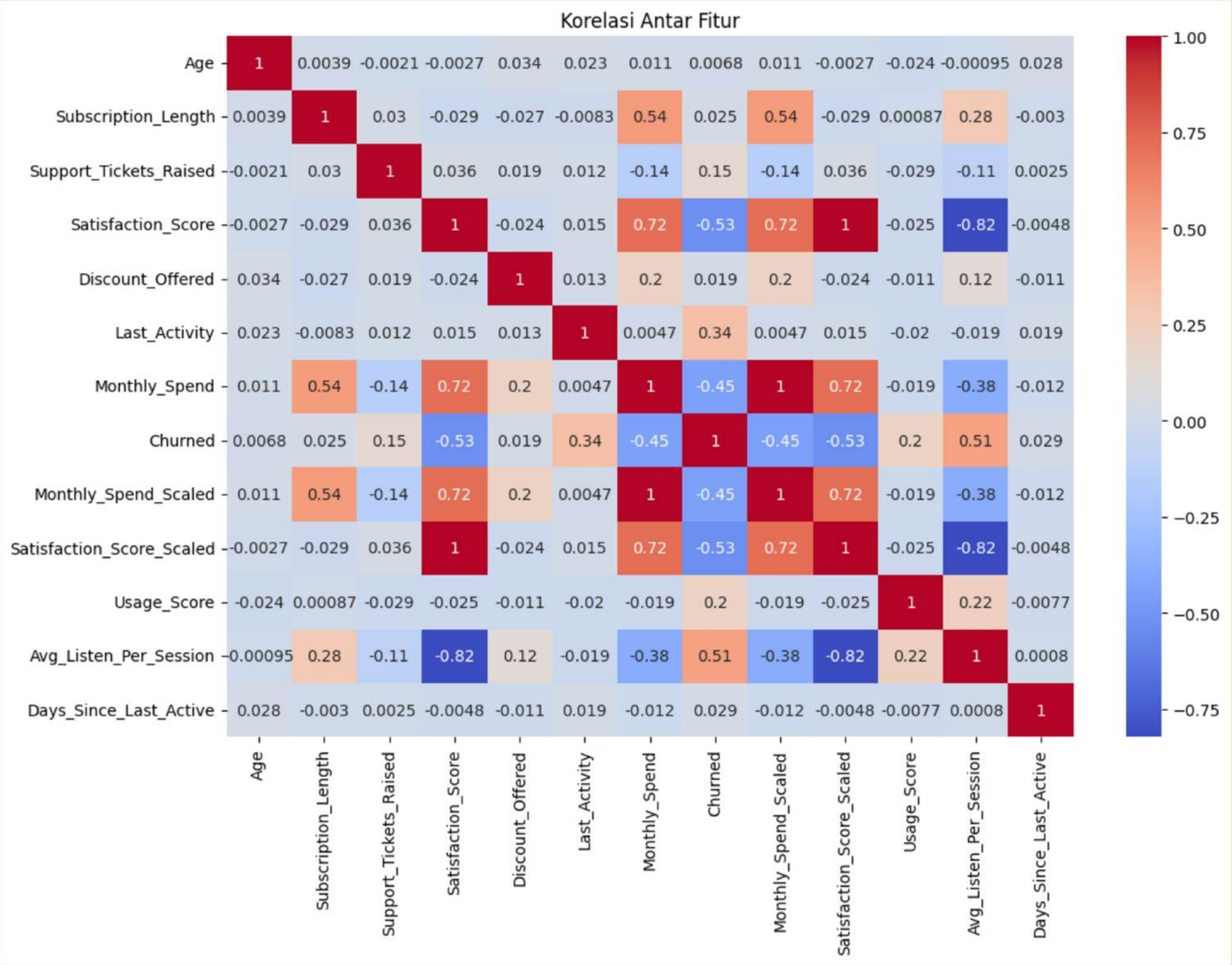
Visualisasi Sebelum vs Sesudah Feature Engineering

1. Monthly Spend (Asli)

- Distribusi data berada dalam rentang sekitar 25 hingga 65 USD.
- Data tampak simetris dan menyerupai distribusi normal.
- Skala nilai asli digunakan, sehingga model bisa berat sebelah (bias) jika ada variabel lain dengan skala berbeda.
- Kurang ideal untuk model berbasis jarak seperti KNN, SVM, atau bahkan neural network.

2. Monthly Spend (Scaled)

- Telah dilakukan Standard Scaling sehingga nilai-nilainya memiliki: Mean = 0, Standar deviasi = 1
- Distribusi tetap sama bentuknya (simetris), tetapi skala berubah menjadi unit baku (standard unit).
- Membantu model untuk lebih stabil dan adil dalam memproses fitur ini bersama fitur lainnya.



Korelasi Fitur terhadap Churn

Korelasi Fitur terhadap Churn

Fitur	Korelasi dengan Churned	Interpretasi
Satisfaction_Score	-0.53	Semakin puas pengguna, semakin kecil kemungkinan mereka churn.
Monthly_Spend	-0.45	Pengguna yang mengeluarkan lebih banyak cenderung lebih loyal.
Avg_Listen_Per_Session	-0.38	Durasi sesi yang lebih tinggi berkorelasi negatif dengan churn.
Last_Activity	plus 0.34	Semakin akhir aktivitasnya, semakin besar kemungkinan churn (kemungkinan cut-off waktu).
Usage_Score	plus 0.22	Korelasi lemah ke sedang; pengguna dengan skor rendah lebih mungkin churn.

Fitur dengan Korelasi Lemah (mendekati 0):

- Age, Discount_Offered, Region, Gender, Support_Tickets_Raised, Days_Since_Last_Active itu semua memiliki korelasi rendah terhadap churn.
- Artinya, fitur-fitur ini kurang informatif secara langsung dalam memprediksi churn.

Insight Utama:

- Fitur berbasis perilaku pengguna dan kepuasan jauh lebih penting daripada data demografis.
- Variabel hasil feature engineering seperti Usage_Score dan Avg_Listen_Per_Session memberikan kontribusi signifikan dan meningkatkan kekuatan prediktif dataset.
- Variabel numerik asli seperti Satisfaction_Score dan Monthly_Spend terbukti sebagai penentu utama churn.

Dampak terhadap Kompleksitas & Model (Preview)

	precision	recall	f1-score	support
0	0.99	1.00	0.99	532
1	1.00	0.99	0.99	448
accuracy			0.99	980
macro avg	0.99	0.99	0.99	980
weighted avg	0.99	0.99	0.99	980

Evaluasi Goals

1. Data dibersihkan & disiapkan
2. Data kompleks disederhanakan
3. Teknik feature engineering diterapkan

Fitur tambahan disesuaikan karena keterbatasan data (genre, aktivitas granular)

Kesimpulan & Next Step

1. Data telah siap digunakan untuk modeling lebih lanjut

2. Feature engineering memberi insight yang lebih baik

3. Next step (opsional):
   - Tambah data temporal
   - Eksperimen dengan model lain (XGBoost, SVM)
   - Evaluasi lebih lanjut (ROC-AUC, Confusion Matrix)

Thank You For Watching