

PREDIKSI PURCHASE CUSTOMER

Berdasarkan pola interaksi
oleh Mulyadi



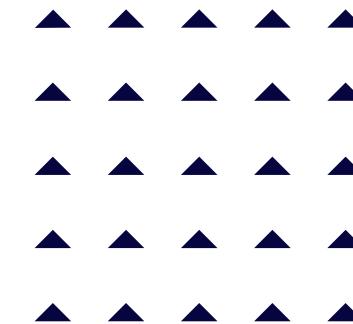
mulyadi.datasci@gmail.com



MATERI

- 1** Latar Belakang dan tujuan
- 2** Data Understanding
- 3** Data Preparation (Cleaning dan Feature Engineering)

- 4** Eksploratory Data Analytics (EDA)
- 5** Model (Insight dan Recommendation)



LATAR BELAKANG



> 70%

pengguna yang memasukkan produk ke keranjang akhirnya **tidak menyelesaikan transaksi**. Angka ini menunjukkan adanya celah besar antara ketertarikan awal dan keputusan akhir untuk membeli. Dalam konteks bisnis, setiap keranjang yang ditinggalkan adalah peluang yang hilang. Oleh karena itu, penting bagi platform e-commerce untuk dapat **mengenali pola perilaku pengguna** sejak dini, dan memprediksi siapa yang benar-benar berpotensi melakukan pembelian.

TUJUAN

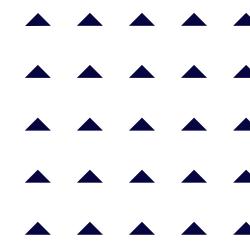


Melalui penelitian ini, saya bertujuan untuk **membangun model prediktif yang mampu mengklasifikasikan kemungkinan konversi pengguna berdasarkan pola interaksi mereka** di platform – seperti jenis aktivitas (view, cart, purchase), waktu akses (pagi, siang, malam), dan hari (weekday atau weekend).

Hasil dari prediksi ini dapat dimanfaatkan sebagai dasar dalam **pengambilan keputusan bisnis**, seperti pemberian diskon yang lebih tepat sasaran, pengiriman notifikasi strategis, hingga pengembangan sistem rekomendasi yang lebih akurat.



DATA UNDERSTANDING



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 885129 entries, 0 to 885128
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   event_time       885129 non-null   object  
 1   event_type       885129 non-null   object  
 2   product_id       885129 non-null   int64  
 3   category_id      885129 non-null   int64  
 4   category_code    648910 non-null   object  
 5   brand            672765 non-null   object  
 6   price            885129 non-null   float64 
 7   user_id          885129 non-null   int64  
 8   user_session     884964 non-null   object  
dtypes: float64(1), int64(3), object(5)
memory usage: 60.8+ MB
```

Memiliki **800k** data,
dengan **9 kolom**

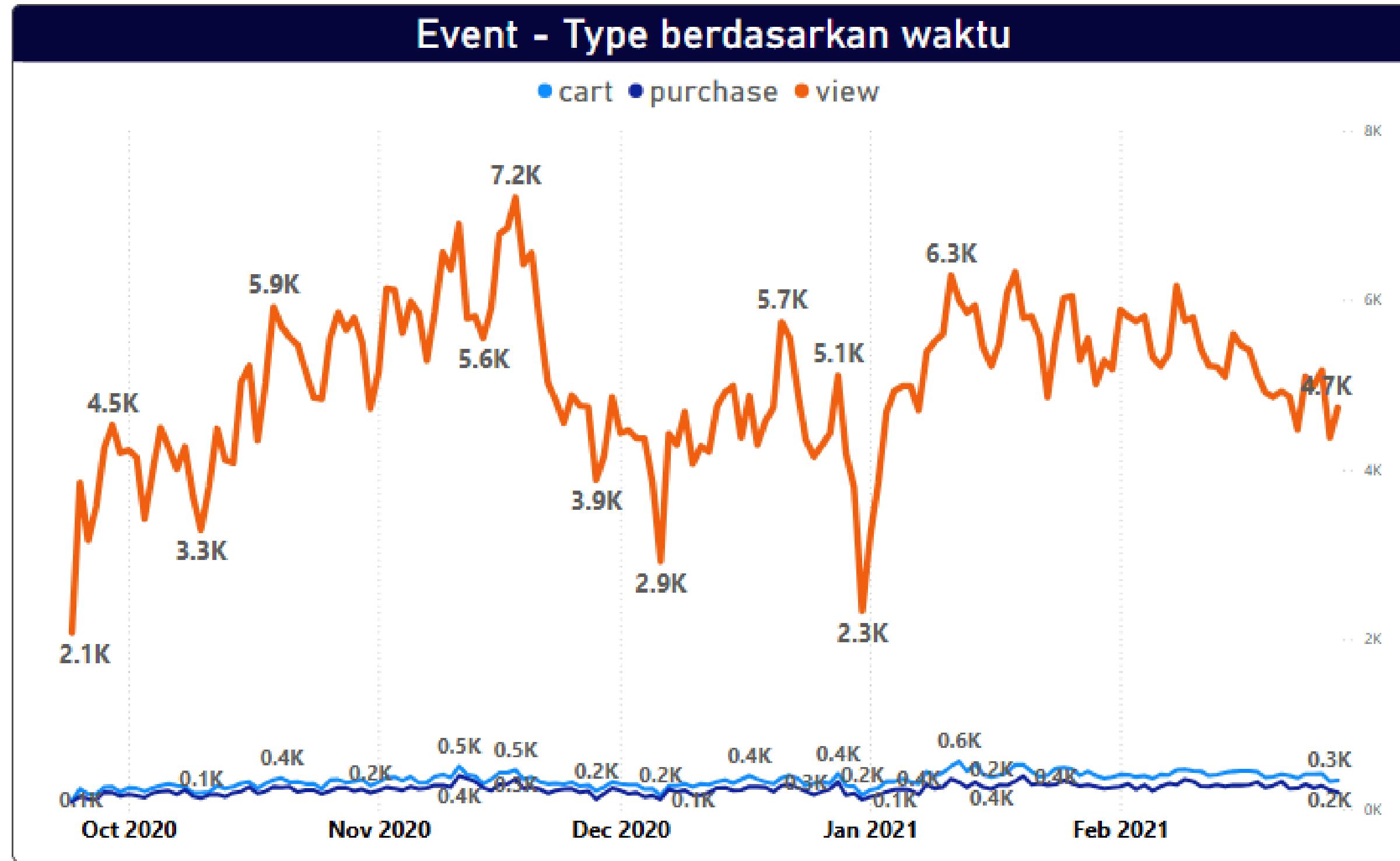
7.4% data duplicate

data null

- **26.6%** pada category_code
- **23.9%** pada brand
- **1.8%** pada user-session

Event_Type

- **89.6%** view
- **6.1%** purchase
- **4.2%** cart



- Event **view** (ditandai dengan garis oranye) memiliki **volume yang jauh lebih tinggi** dibandingkan cart dan purchase.
- Rata-rata harian mencapai ribuan, dengan **puncaknya pada awal November 2020 sekitar 7.2K views.**
- Ini menunjukkan **ketertarikan** pengguna terhadap produk cukup tinggi, namun **tidak berbanding lurus** dengan tindakan pembelian.

DATA PREPARATION



Cleaning

Drop Duplicate (655)

Terdapat nilai **Nan** pada lebih dari 25% dari **dataset**. Namun, karena nilai-nilai tersebut tidak **digunakan** dalam proses modeling, maka tidak dilakukan proses drop agar dataset tetap memiliki kualitas dan jumlah data yang memadai untuk pelatihan model.



Feature Engineering

	day_type	time_period
0	weekday	siang
1	weekend	sore
2	NaN	malam
3	NaN	pagi

Menghapus satuan waktu “UTC” pada kolom `event_time`

Mengganti type data menjadi “datetime” pada kolom `event_time`

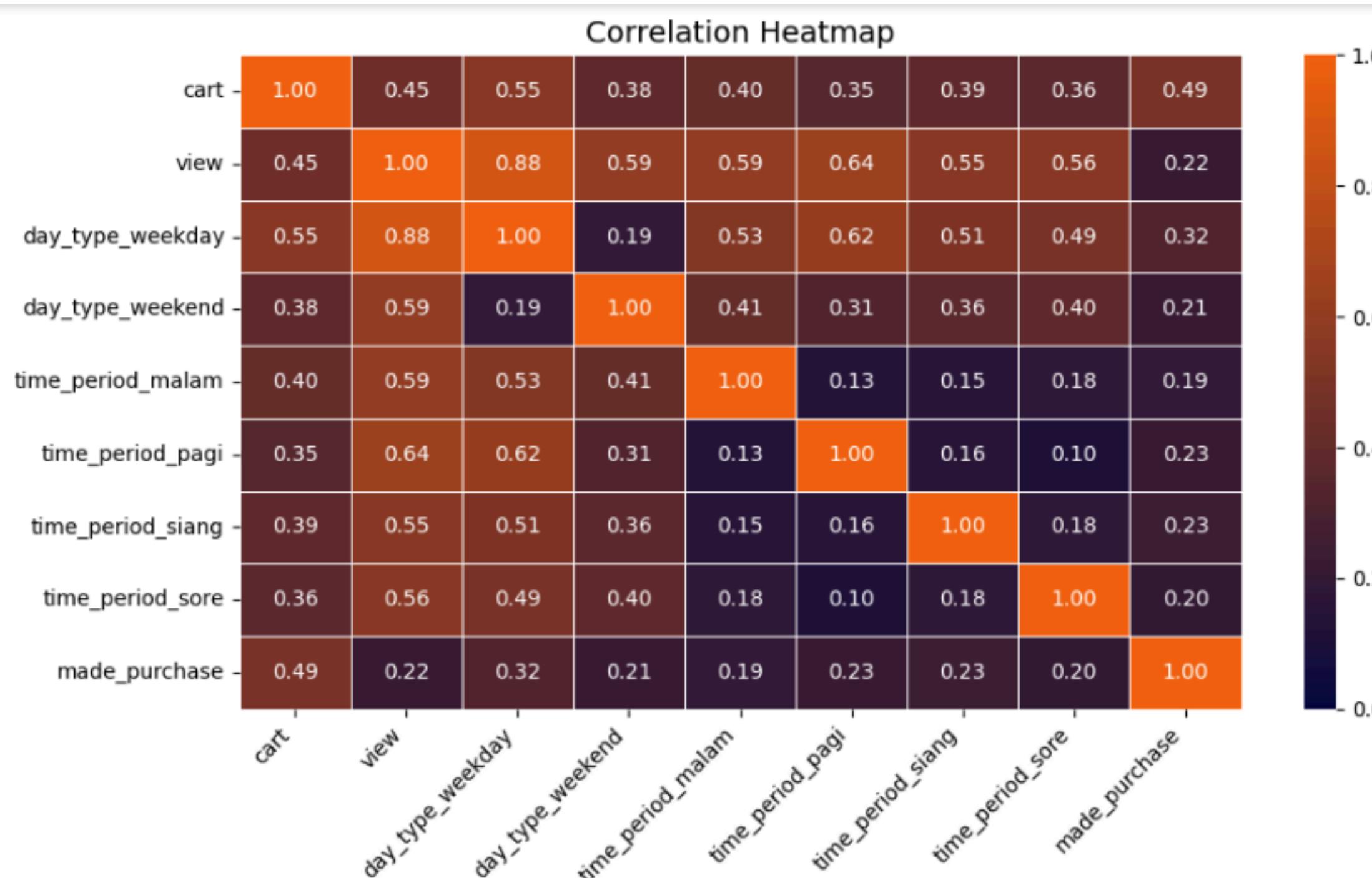
Menurunkan kolom baru dari `event_time` menjadi kolom baru “weekday” dan “weekend”, dan kolom `time_period` “pagi,siang,sore,malam”

Feature Engineering

	user_id	cart	purchase	view	day_type_weekday	day_type_weekend	time_period_malam	time_period_pagi	time_period_siang	time_period_sore	made_purchase
0	1515915625353226922	0	0	1		1	0	0	0	1	0
1	1515915625353230067	0	0	1		1	0	0	1	0	0
2	1515915625353230683	0	0	13		12	1	0	13	0	0
3	1515915625353230922	0	0	1		1	0	0	1	0	0
4	1515915625353234047	0	0	36		28	8	0	32	0	4

menggunakan **pivot** berdasarkan user_id pada kolom event_type, day_type, dan time_period lalu menambahkan kolom made_purchase yang akan menjadi target dari pemodelan menggunakan dataframe baru ini

EDA



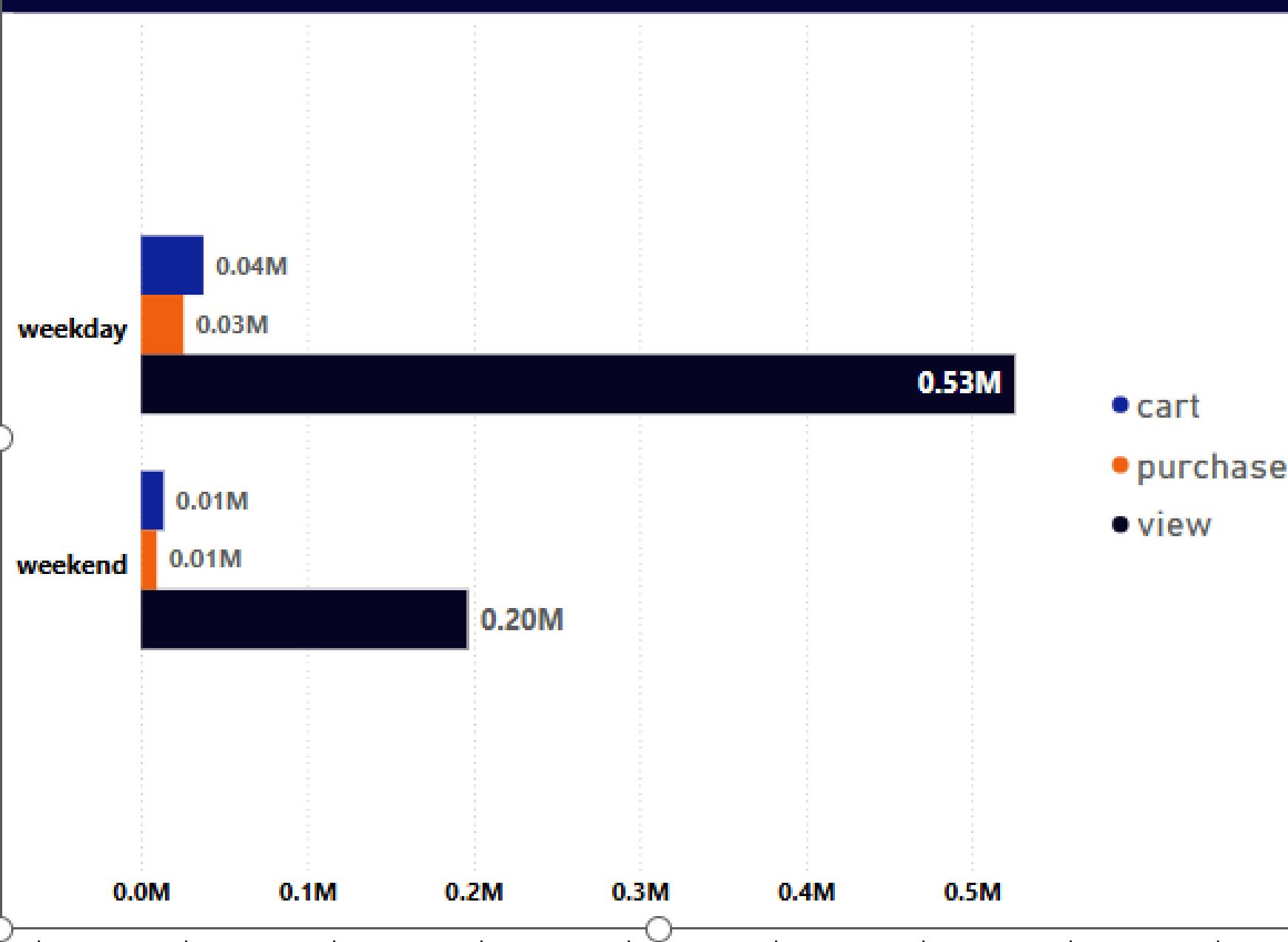
Correlation

Cart memiliki korelasi tertinggi dengan made_purchase sebesar 0.49 -

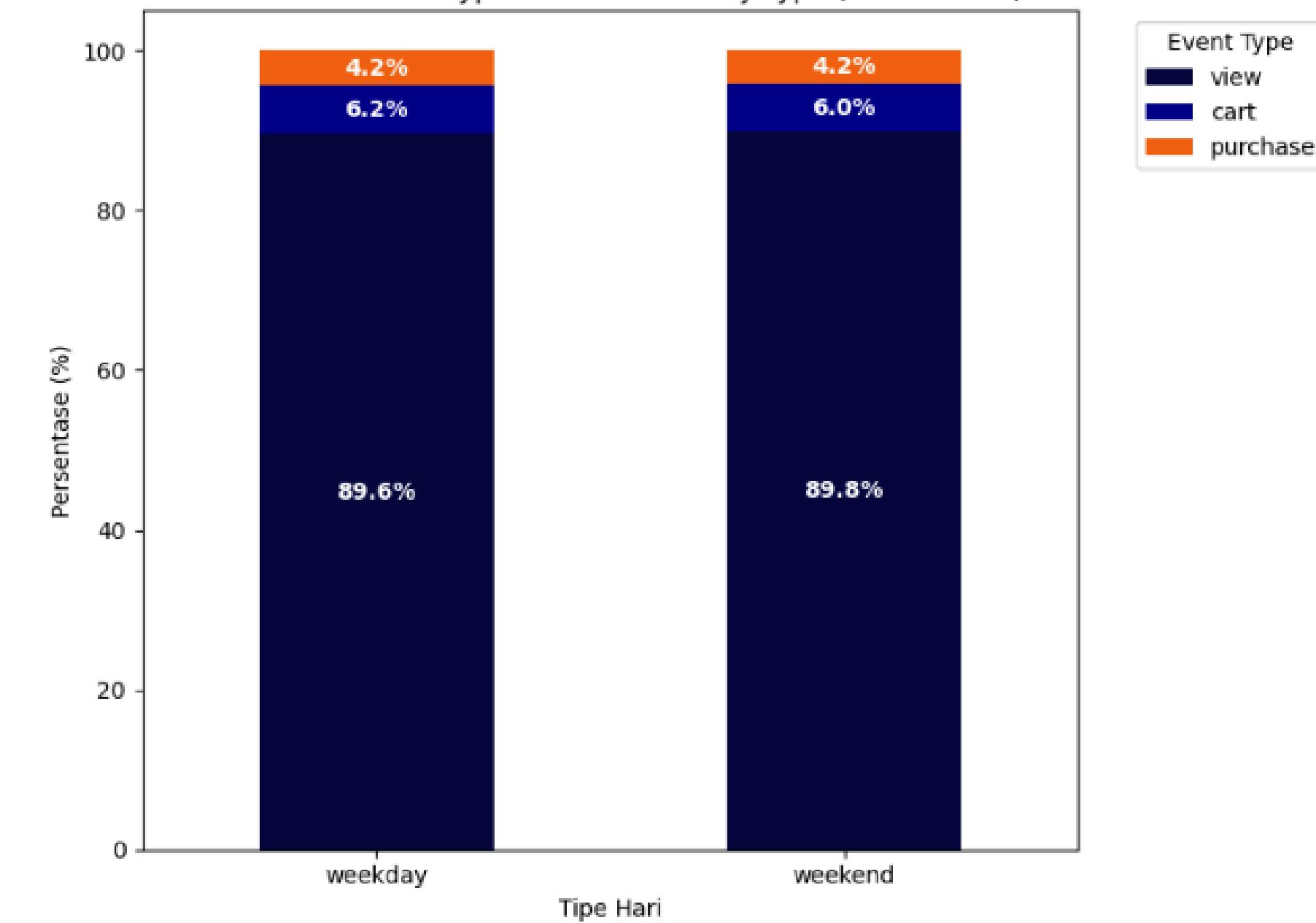
Terdapat korelasi positif moderate antara pengguna yang memasukkan item ke keranjang dengan tindakan melakukan pembelian.

Insight: Fitur cart bisa jadi indikator kuat untuk prediksi pembelian.

Jenis Aktifitas berdasarkan Hari



Percentase Event Type Berdasarkan Day Type (Stacked Bar)



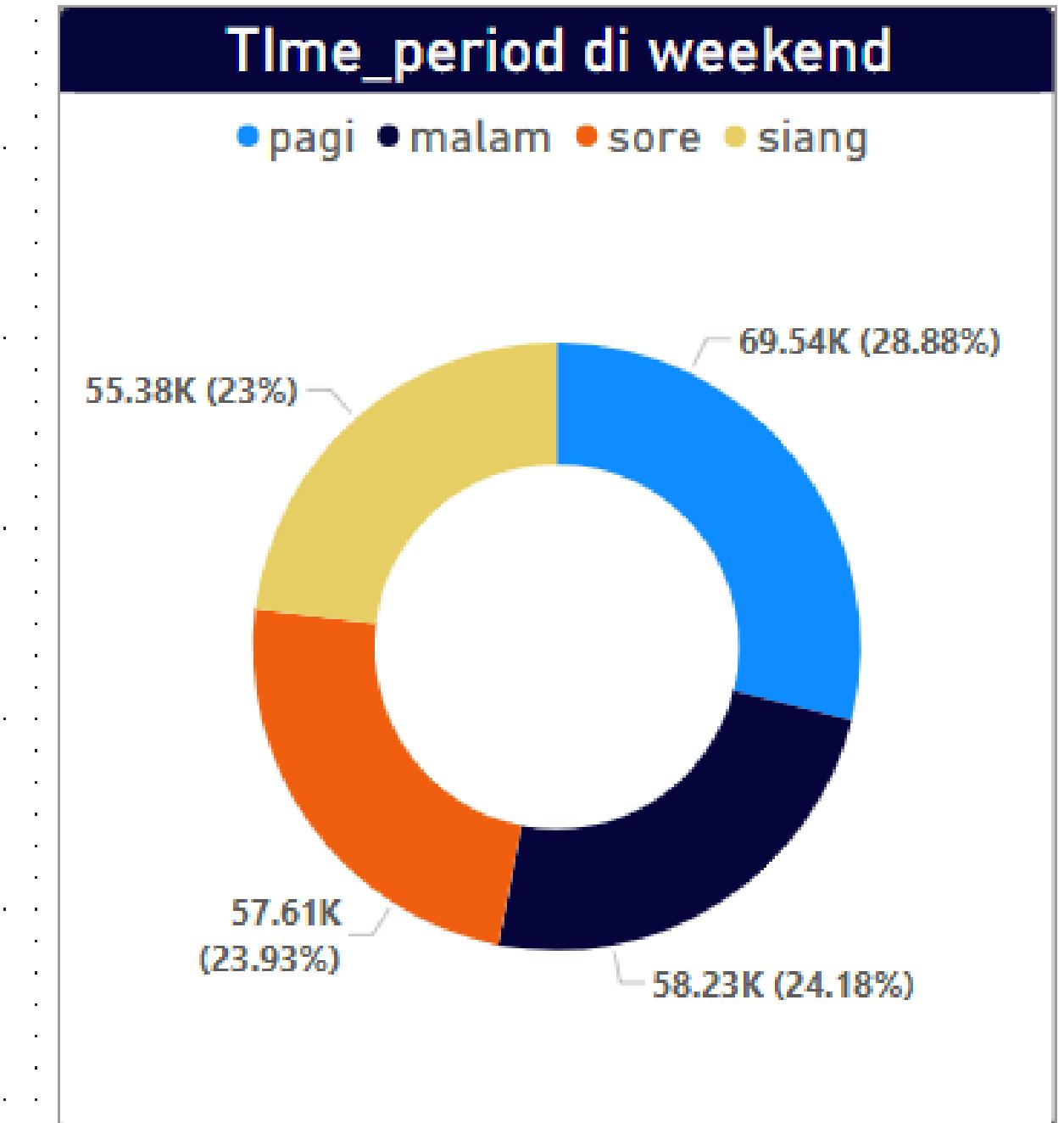
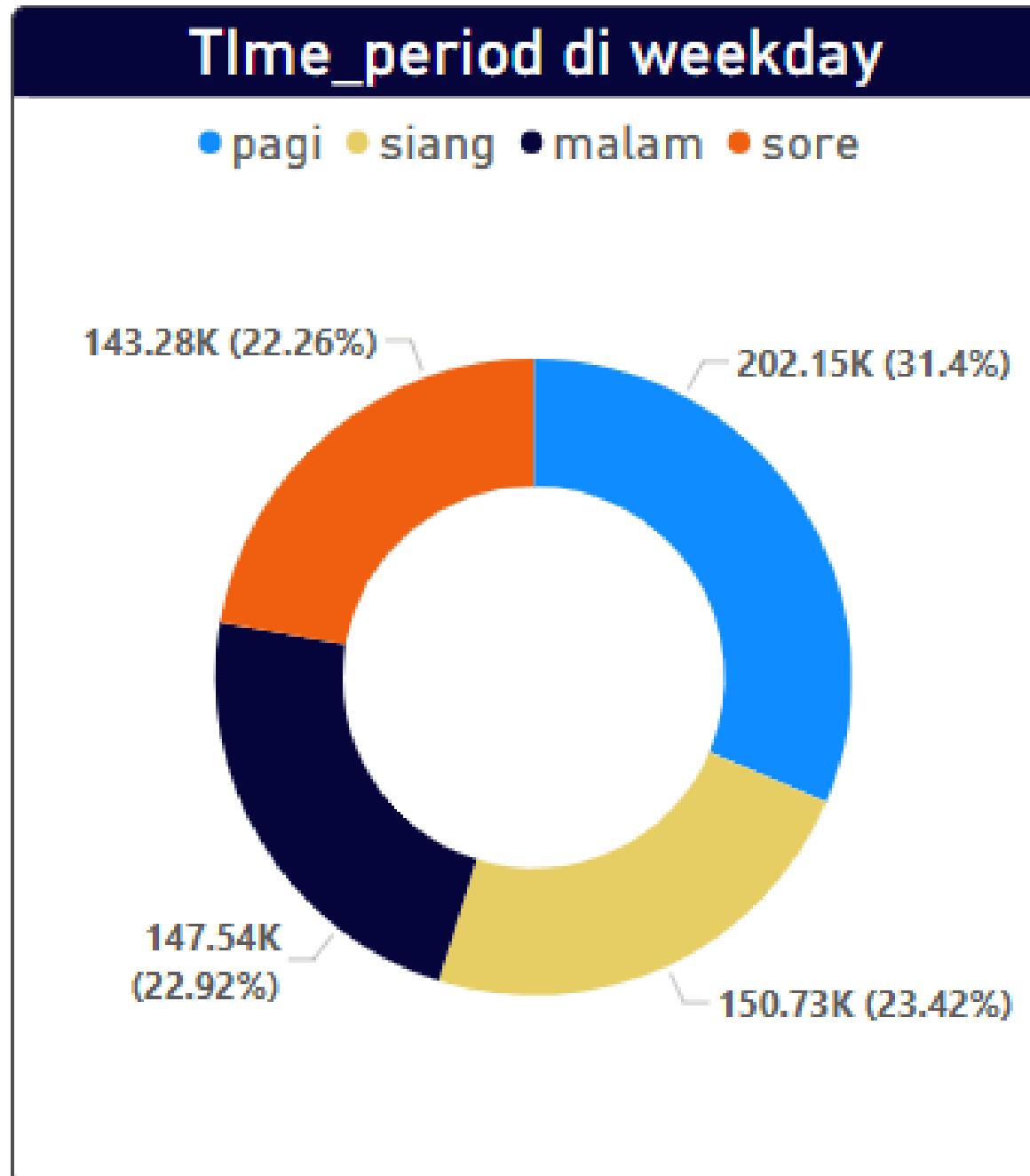
Event_type vs Day_type

tingkat **purchase** antara weekday dan weekend masih sama sekitar **4.2 %** sedangkan customer yang hanya melihat sangat tinggi **89.4 %**

Time_period vs Day_type

dihari kerja 31.4% aktifitas interaksi terjadi pada pagi hari,

berbeda dengan weekend yang terbagi merata di pagi, siang, sore dan malam hari sekitar 25%



MODELING ••••



Feature dan target

```
x = features.drop(columns=['user_id', 'made_purchase', 'purchase'])  
y = features['made_purchase']
```

split dataset

```
x_train, x_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.3, random_state=42, stratify=y)
```

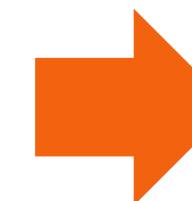
Smote

```
smote = SMOTE(random_state=42)  
x_train_resampled, y_train_resampled = smote.fit_resample(x_train, y_train)
```

Bentuk model

```
model = RandomForestClassifier(random_state=42)
model.fit(X_train_resampled, y_train_resampled)

+ RandomForestClassifier
  RandomForestClassifier(random_state=42)
```



Terapkan pada datatrain

	precision	recall	f1-score	support
0	1.00	1.00	1.00	270185
1	1.00	1.00	1.00	270185
accuracy				
macro avg				
weighted avg				
ROC-AUC: 0.999901919055462				

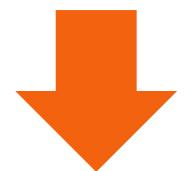
Skor ROC-AUC yang mendekati 1 (0.999) menandakan model mampu memisahkan dengan sangat baik antara pengguna yang akan melakukan pembelian dan yang tidak, di dalam data pelatihan.

Terapkan pada datatest

	precision	recall	f1-score	support
0	1.00	0.99	1.00	115794
1	0.99	0.98	0.94	6391
accuracy			0.99	122185
macro avg	0.95	0.99	0.97	122185
weighted avg	0.99	0.99	0.99	122185

ROC-AUC: 0.9994072153888163

Skor ROC-AUC pada datatest yang didapatkan masih mendekati 1 menandakan model sangat bagus, tetapi kita akan melakukan pengecekan apakah model ini **overfitting**?



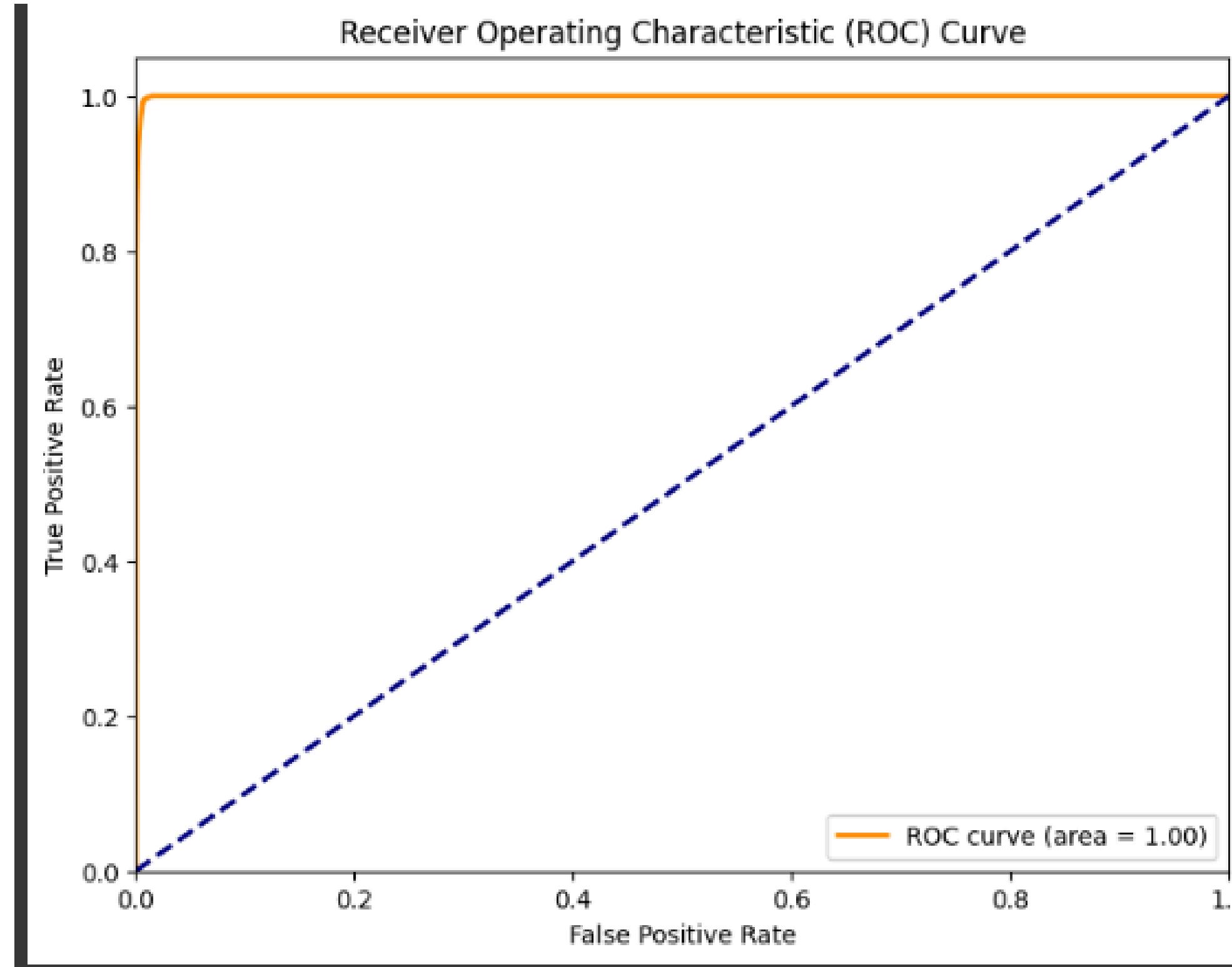
Evaluasi menggunakan CV-Score

```
cv_scores = cross_val_score(model, X_train_resampled, y_train_resampled, cv=5, scoring='roc_auc')
print(f"Cross-validation Scores: {cv_scores}")

Cross-validation Scores: [0.9998542  0.9998691  0.9999137  0.99982683 0.999933 ]
```

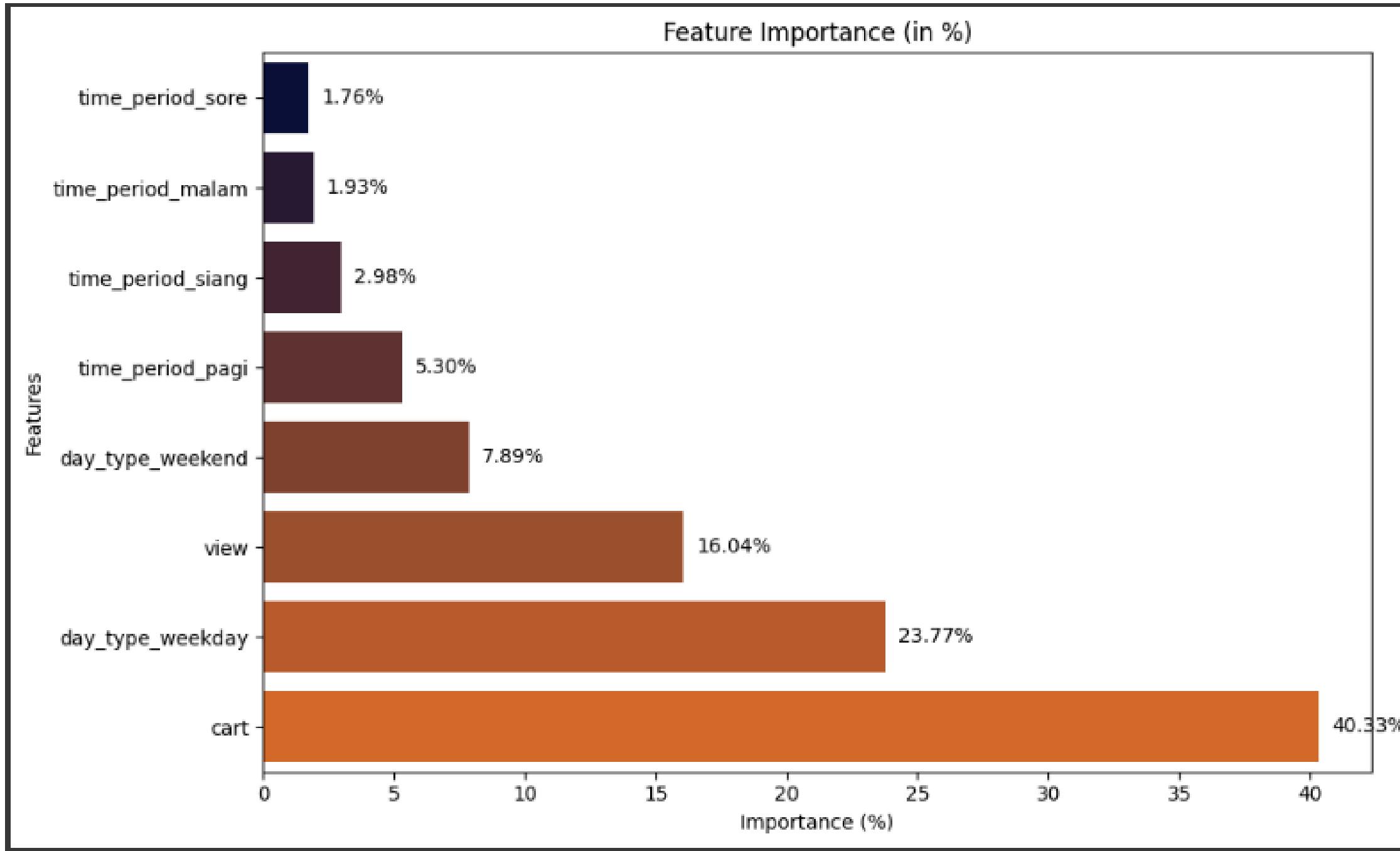
Di sini, skor training dan CV sama-sama sangat tinggi dan seragam, artinya **tidak** ada indikasi **overfitting** saat ini.

Evaluasi menggunakan ROC-Curve



- Kurva ROC-nya sangat dekat dengan titik (0, 1) – artinya model memiliki **True Positive Rate (TPR)** tinggi dan **False Positive Rate (FPR)** sangat rendah di hampir semua threshold.
- Area Under Curve (AUC) = 1.00 – ini adalah nilai maksimum. Artinya:
 - Model memisahkan kelas positif dan negatif dengan sempurna pada data yang diuji.
 - Tidak ada kesalahan klasifikasi pada threshold optimal.

Feature Importance



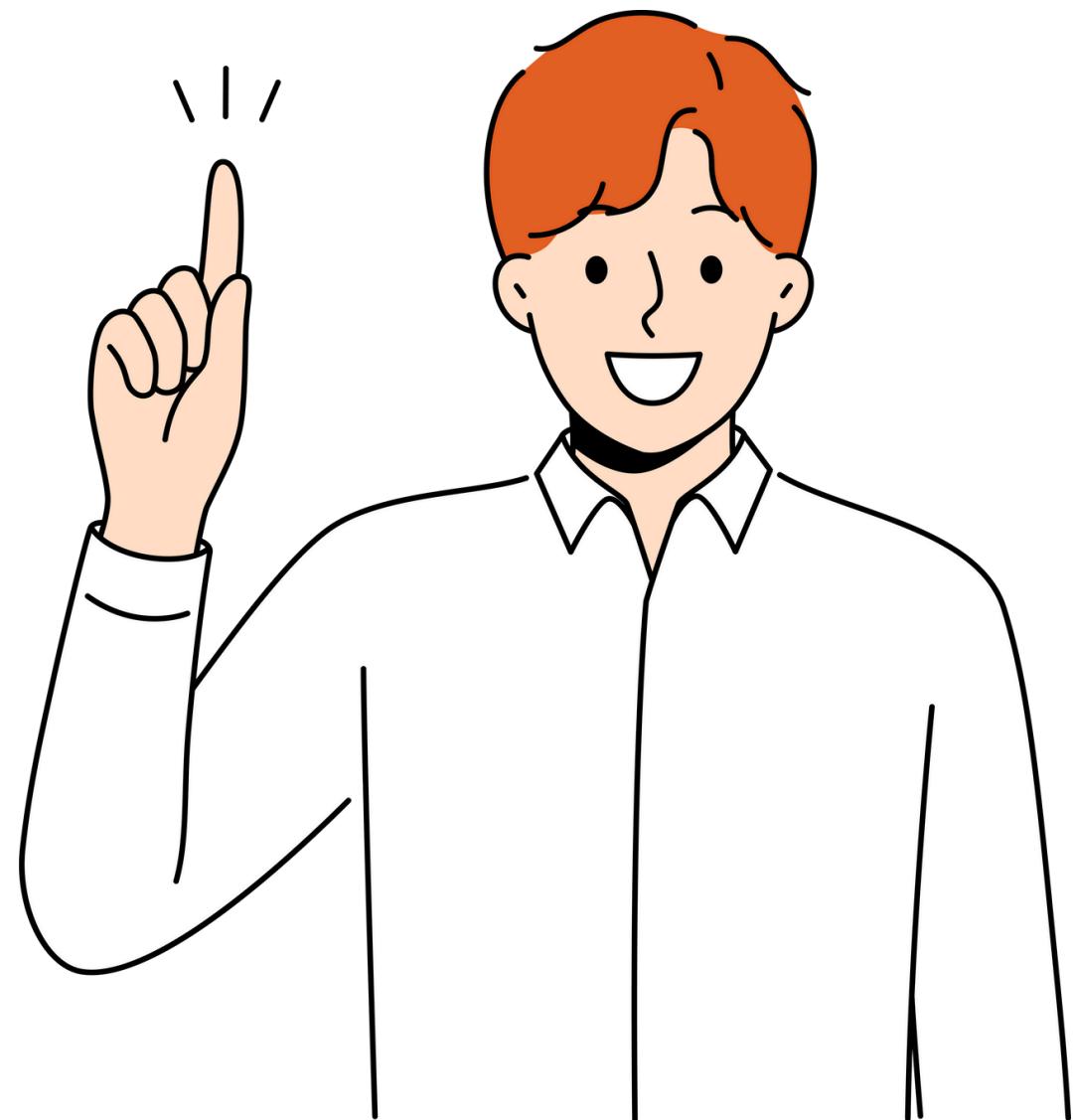
Cart dengan 40.3 % merupakan indikator terkuat dalam memprediksi pembelian. Artinya, pengguna yang menambahkan produk ke keranjang sangat berkorelasi dengan peluang melakukan pembelian.

Hari kerja memiliki pengaruh besar terhadap keputusan pembelian sebesar 23%. Menunjukkan bahwa mayoritas transaksi mungkin terjadi di hari kerja

RECOMMENDATION



- **Cart:** Prioritaskan pengguna yang menambahkan produk ke keranjang dengan strategi retargeting dan insentif checkout.
- **Day Type Weekday:** Fokuskan promosi utama di hari kerja karena tingkat konversinya lebih tinggi.
- **View:** Tingkatkan konversi dari pengguna yang melihat produk dengan CTA (Call to Action) yang kuat dan rekomendasi produk.
- **Day Type Weekend:** Manfaatkan akhir pekan untuk kampanye santai dan promosi bertema liburan.



STREAMLIT APP



Prediksi Customer Akan Melakukan Purchase ↴

Masukkan informasi interaksi customer:

Jumlah Cart

1

Jumlah View

57

Jumlah penggunaan di hari biasa (senin-jumat)

1

Jumlah penggunaan di hari libur minggu (sabtu-minggu)

0



Kunjungi [Link](#) berikut

TERIMA KASIH

telah membaca



mulyadi.datasci@gmail.com

.....