

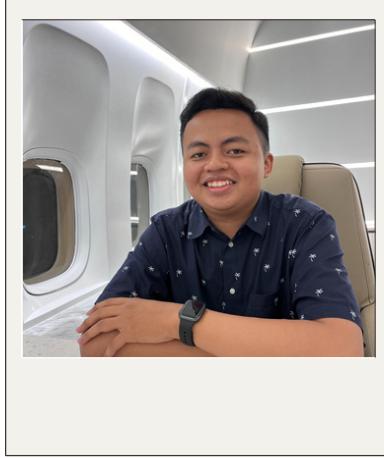


Sales Forecasting

Margareth Hamiltion

Monday, September 30th 2024

Meet the Team



Alwy Bathia R.

LSTM



Daniel Machsimus L.

EDA/VARMAX



Jason Hermawan

EDA

Background & Problem Statement

Sales forecasting adalah salah satu metode penting bagi bisnis untuk mengoptimalkan persediaan dan meningkatkan pendapatan. Forecasting membantu bisnis merencanakan promosi dan mengelola produk dengan lebih baik.

Bisnis sering menghadapi kesulitan dalam memprediksi penjualan, yang menyebabkan kelebihan atau kekurangan stok. Hal ini dapat memengaruhi keuntungan dan kepuasan pelanggan.

Solusi untuk permasalahan ini adalah model AI yang dapat membantu melakukan peramalan dengan menembukan pola dalam data dan menyesuaikan dengan tren.

Objectives & Scope

Objectives

- Meningkatkan akurasi peramalan penjualan dengan menggunakan algoritma Time Series.
- Mengidentifikasi pola penjualan dari data historis untuk membantu dalam pengambilan keputusan terkait persediaan dan promosi.
- Mempertimbangkan faktor seperti promosi dan kondisi pasar untuk memprediksi penjualan lebih tepat.

Scopes:

- Model AI ini hanya fokus pada data penjualan dari STORE 5 dan tidak mencakup data toko lain.

Data Collection & Preparation

Dataset yang digunakan merupakan bagian dari **dataset penjualan** dari perusahaan retail produk yang berbasis Ekuador, **Corporacion Favorita**. Dataset online ini adalah upaya perusahaan dalam mengajak orang-orang untuk bisa **memprediksi product sales secara akurat** menggunakan **machine learning** [1].

Pada project kali ini, **dataset dimodifikasi** sedemikian rupa sehingga kolom pada dataset adalah feature 'id', 'date', 'store_nbr', 'family', 'sales', 'onpromotion, dan 'dcoilwtico'.

Berikut keterangan dari masing-masing feature (dengan penyesuaian) [1]:

- id : penandaan yang unique untuk memberi label pada setiap barisnya
- 'date' : tanggal untuk setiap baris
- 'store_nbr' : nomor toko
- 'family' : kategori produk
- 'onpromotion' : tanda apakah sebuah produk sedang promosi atau tidak
- dcoilwtico : harga minyak (oil) harian.



Corporacion Favorita (atas). Source: Wikidata, accessed on September 30th, 2024.
Penjualan sayur-mayur di supermarket (bawah). Source: Harvard Kennedy School accessed on September 30th, 2024.

Data Collection & Preparation

family AUTOMOTIVE

	date	sales	onpromotion	dcoilwtico
count	1684	1684.000000	1684.000000	1163.000000
mean	2015-04-24 08:27:04.703088128	5.459620	0.011283	67.925589
min	2013-01-01 00:00:00	0.000000	0.000000	26.190000
25%	2014-02-26 18:00:00	3.000000	0.000000	46.390000
50%	2015-04-24 12:00:00	5.000000	0.000000	53.330000
75%	2016-06-19 06:00:00	7.000000	0.000000	95.790000
max	2017-08-15 00:00:00	19.000000	2.000000	110.620000
std	NaN	3.257265	0.111132	25.677366

family BEAUTY

	date	sales	onpromotion	dcoilwtico
count	1684	1684.000000	1684.000000	1163.000000
mean	2015-04-24 08:27:04.703088128	5.15677	0.154988	67.925589
min	2013-01-01 00:00:00	0.000000	0.000000	26.190000
25%	2014-02-26 18:00:00	3.000000	0.000000	46.390000
50%	2015-04-24 12:00:00	5.000000	0.000000	53.330000
75%	2016-06-19 06:00:00	7.000000	0.000000	95.790000
max	2017-08-15 00:00:00	25.000000	2.000000	110.620000
std	NaN	3.13061	0.391948	25.677366

family GROCERY I

	date	sales	onpromotion	dcoilwtico
count	1684	1684.000000	1684.000000	1163.000000
mean	2015-04-24 08:27:04.703088128	3125.107873	22.163302	67.925589
min	2013-01-01 00:00:00	0.000000	0.000000	26.190000
25%	2014-02-26 18:00:00	2651.986000	0.000000	46.390000
50%	2015-04-24 12:00:00	2973.500000	10.000000	53.330000
75%	2016-06-19 06:00:00	3477.500000	36.000000	95.790000
max	2017-08-15 00:00:00	7656.000000	178.000000	110.620000
std	NaN	732.038714	28.432155	25.677366

Deskripsi data:

- range feature sales: 0.0 - 19.0
- range feature dcoilwtico: 26.19 - 110.62
- range feature onpromotion: 0.0 - 2.0

Deskripsi data:

- range feature sales: 0.0 - 25.0
- range feature dcoilwtico: 26.19 - 110.62
- range feature onpromotion: 0.0 - 2.0

Deskripsi data:

- range feature sales: 0.0 - 7656.0
- range feature dcoilwtico: 26.19 - 110.62
- range feature onpromotion: 0.0 - 178.0

Seluruh data pada rentang 'date' = 2013-01-01 hingga 'date' = 2017-08-15

Data Collection & Preparation

Flow: terdapat sales = 0 pada tanggal 01-01 -> cek sales pada seluruh hari libur di Ekuador -> lihat hasil sales dari tahun ke tahun

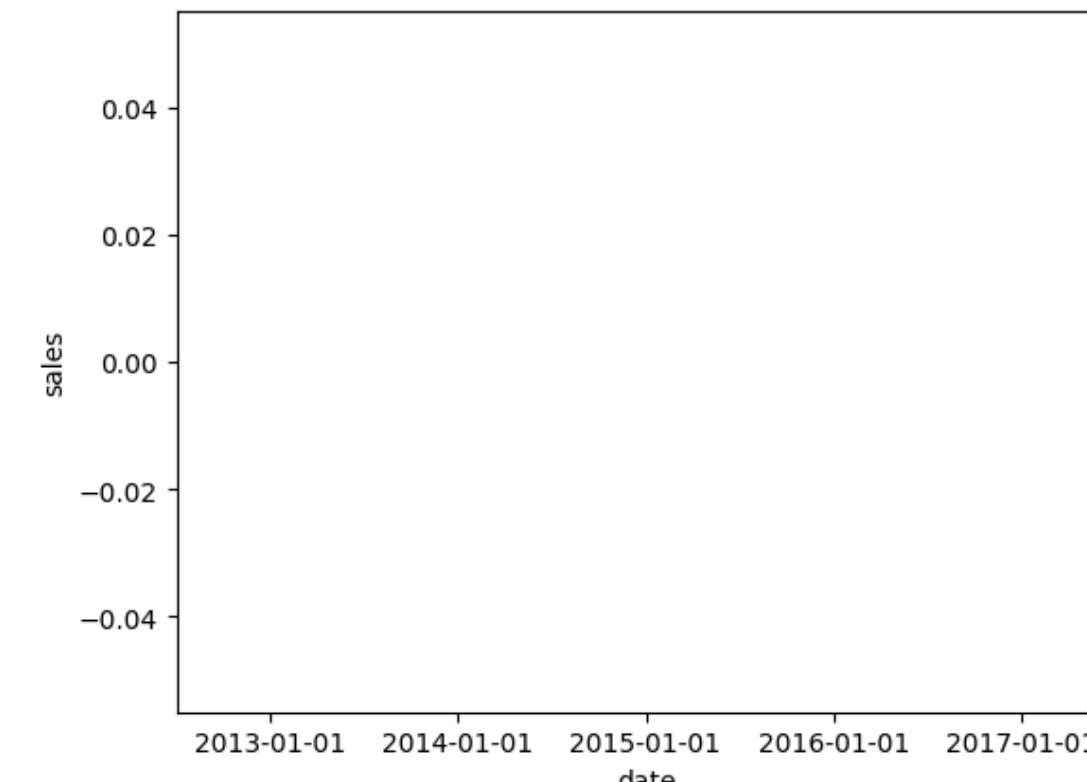
	date	family	sales	onpromotion	dcoilwtico
0	2013-01-01	AUTOMOTIVE	0.0	0	NaN
12012	2014-01-01	AUTOMOTIVE	0.0	0	NaN
24024	2015-01-01	AUTOMOTIVE	0.0	0	NaN
36036	2016-01-01	AUTOMOTIVE	0.0	0	NaN
48081	2017-01-01	AUTOMOTIVE	0.0	0	NaN

Daftar hari libur lainnya di Equador [6]:

- * Carnaval (2 hari berubah-ubah setiap tahun) : range awal Februari hingga pertengahan Maret
- * Good Friday (berubah-ubah setiap tahun) : range akhir Maret hingga pertengahan April
- * International Workers Day : selalu 1 Mei
- * The Battle of Pichincha : selalu 24 Mei
- * Declaration of Independence of Quito: selalu 10 Agustus
- * Declaration of Independence of Guayaquil : selalu 9 Oktober
- * All Souls' Day : selalu 2 November
- * Declaration of Independence of Cuenca : selalu 3 November

Kecuali tanggal 01 Januari, setiap hari libur lainnya, **sales tidak menunjukkan angka 0**, seperti ditunjukkan pada barplot di bawah.

Selain itu, dari keseluruhan tanggal, terdapat tanggal yang **tidak ada pada dataset -> 25 Desember**.



Missing Value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55572 entries, 0 to 55571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   date        55572 non-null   datetime64[ns]
 1   family       55572 non-null   object  
 2   sales        55572 non-null   float64 
 3   onpromotion  55572 non-null   int64  
 4   dcoilwtico  38379 non-null   float64 
dtypes: datetime64[ns](1), float64(2), int64(1), object(1)
memory usage: 2.1+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1684 entries, 0 to 55539
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   date        1684 non-null   datetime64[ns]
 1   family       1684 non-null   object  
 2   sales        1684 non-null   float64 
 3   onpromotion  1684 non-null   int64  
 4   dcoilwtico  1163 non-null   float64 
dtypes: datetime64[ns](1), float64(2), int64(1), object(1)
memory usage: 143.5+ KB
```

	0
id	0
date	0
store_nbr	0
family	0
sales	0
onpromotion	0
dcoilwtico	17193
dtype: int64	

Berdasarkan deskripsi di samping, tampak bahwa ada missing value pada feature 'dcoilwtico' sebanyak 17,193 data dari total 55,572 data.

	0
date	0
family	0
sales	0
onpromotion	0
dcoilwtico	521
dtype: int64	

Karena data dianalisis untuk tiap famili nya dari total 33 family, tiap family memiliki masing-masing 521 missing value dari total 1684 data.

Missing Value

	id	date	store_nbr	family	sales	onpromotion	dcoilwtico
0	1452	2013-01-01	5	AUTOMOTIVE	0.000	0	NaN
1	1453	2013-01-01	5	BABY CARE	0.000	0	NaN
2	1454	2013-01-01	5	BEAUTY	0.000	0	NaN
3	1455	2013-01-01	5	BEVERAGES	0.000	0	NaN
4	1456	2013-01-01	5	BOOKS	0.000	0	NaN
...
55501	2997022	2017-08-13	5	POULTRY	232.897	0	NaN
55502	2997023	2017-08-13	5	PREPARED FOODS	65.945	0	NaN
55503	2997024	2017-08-13	5	PRODUCE	1487.221	3	NaN
55504	2997025	2017-08-13	5	SCHOOL AND OFFICE SUPPLIES	1.000	0	NaN
55505	2997026	2017-08-13	5	SEAFOOD	11.642	0	NaN

Jika dilihat, missing value (NaN) pasti berada pada feature 'dcoilwtico' di **tanggal-tanggal yang date-nya merupakan akhir pekan (Sabtu dan Minggu)**.

Imputasi Missing Value

Missing value pada data adalah pada **feature ‘dcoilwtico’**. Feature ini menyatakan harga minyak. Feature ini merupakan **variabel eksogen** karena tidak berhubungan dengan feature ‘utama’, tetapi memengaruhi nilainya. Di sini, diasumsikan feature ‘dcoilwtico’ sebagai variabel eksogen. Ekuador adalah negara yang bergantung pada minyak dan kondisinya sangat rentan terhadap perubahan harga yang drastis di pasar [1].

Harga minyak memiliki **behavior yang interpolatif**, artinya **memiliki transisi yang smooth pada harganya**, dipengaruhi oleh sentimen pasar dan berbagai kejadian. Misalnya, situasi geopolitik antara Ukraina dan Rusia bisa menyebabkan fluktuasi yang mendadak [2].

Ekuador sebagai negara yang penghasilan terbesarnya adalah ekspor minyak bumi mengikuti harga minyak bumi **Oriente crude oil price** merupakan harga minyak yang **spesifik untuk Ekuador** [3]. Akan tetapi, harga minyak Oriente **berganti setiap bulan** [3]. Artinya, ada harga minyak lain yang digunakan. **WTI (West Texas Intermediate)** adalah harga minyak yang **berganti setiap hari** dan memiliki **5 hari kerja**. Di sini, diasumsikan ‘dcoilwtico’ mengikuti harga minyak WTI juga karena karakteristik pasarnya, yaitu penjualan **minyak yang karakteristiknya ‘light dan sweet’** [4]. Light artinya punya viscosity dan density rendah. Sweet artinya kandungan sulfurnya < 0.5%, seperti karakteristik minyak bumi di pasar Oriente. Brent sebagai pesain WTI tidak digunakan karena sistemnya menggunakan 6 hari kerja [5].

Di sini, data yang hilang adalah data dcoilwtico pada weekend. Untuk itu, data yang digunakan untuk mengisi nilai NaN adalah data hari sebelumnya. Akan digunakan metode **forward fill**, yaitu metode **penggunaan nilai terakhir**. Sebagai perbandingan, **digunakan metode interpolasi** untuk menggambarkan karakteristik interpolatif dari data.

[2] Q. Zhang, Y. Hu, J. Jiao, and S. Wang, “The impact of Russia–Ukraine war on crude oil prices: an EMC framework,” *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–12, Jan. 2024, doi: <https://doi.org/10.1057/s41599-023-02526-9>.

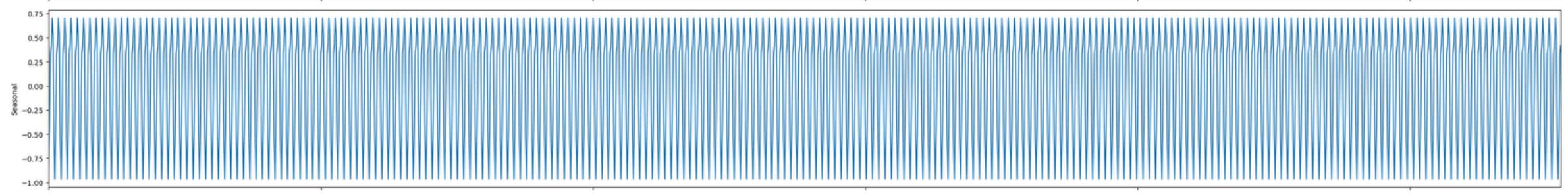
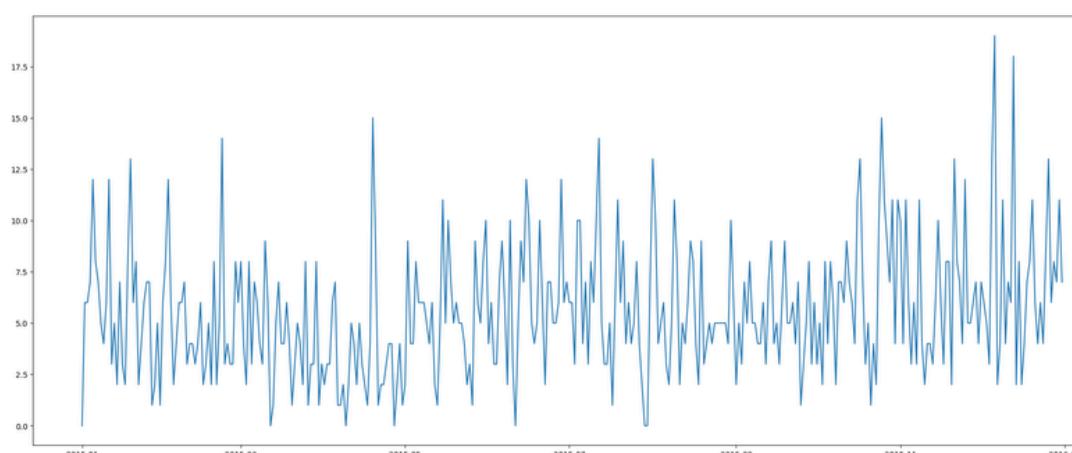
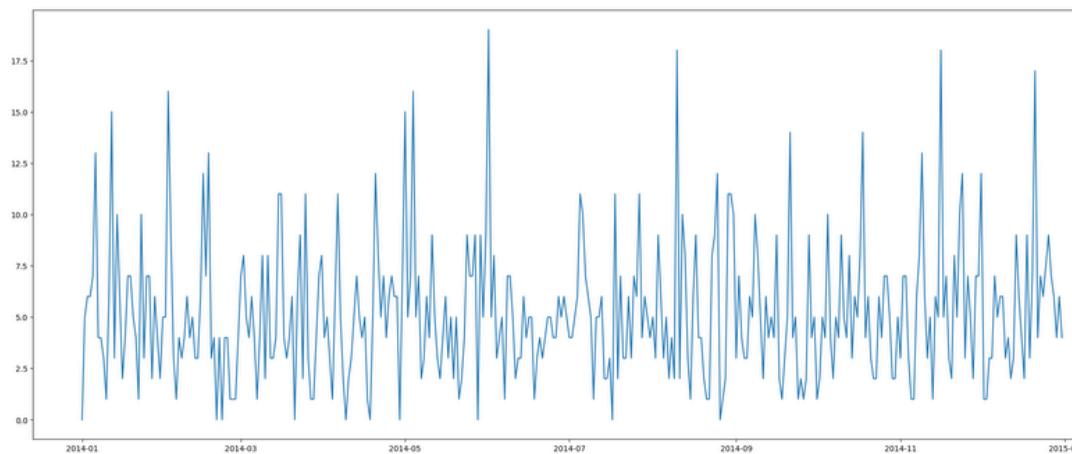
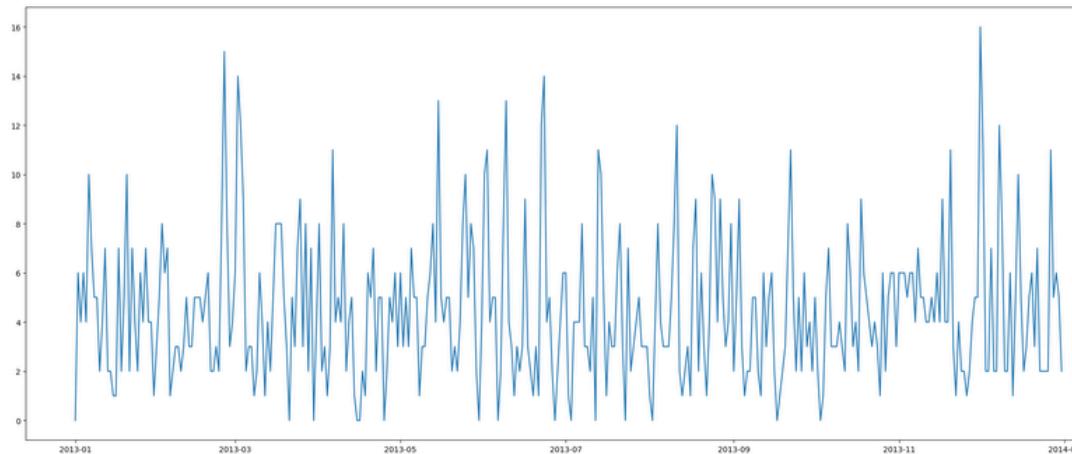
[3] “Ecuador | Crude Oil Prices | CEIC,” www.ceicdata.com/en/ecuador/crude-oil-prices (accessed Sep. 28, 2024).

[4] Author, “Brent Crude vs. West Texas Intermediate: The Differences,” *Investopedia*, Aug. 30, 2021. <https://www.investopedia.com/ask/answers/052615/what-difference-between-brent-crude-and-west-texas-intermediate.asp> (accessed Sep. 28, 2024).

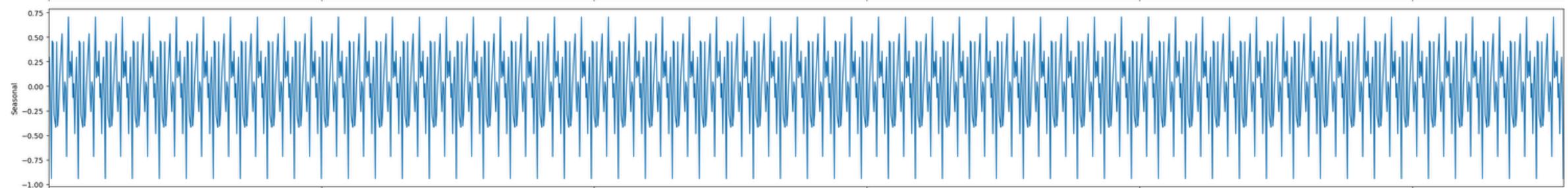
[5] “Live Crude Oil Spot Prices (Brent & WTI) + Historical Charts,” *Commodity.com*. <https://commodity.com/energy/oil/price/> (accessed Sep. 28, 2024).

Pengambilan Keputusan Komponen Seasonal

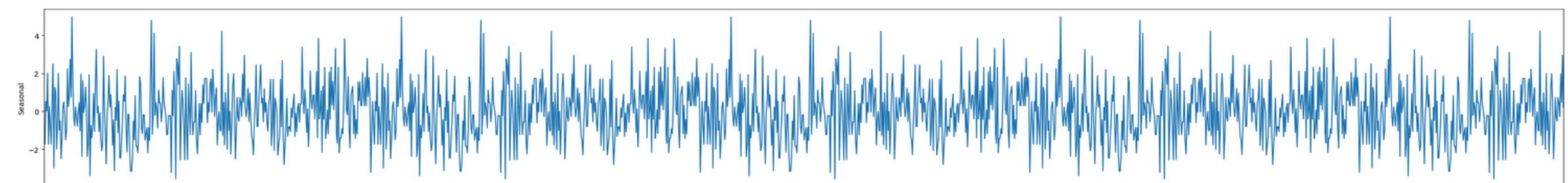
Approach : 1. Melihat lineplot dari tiap feature, 2. Melihat seasonal decomposition, 3. Melihat plot ACF family AUTOMOTIVE



Komponen seasonal weekly (period=7)



Komponen seasonal monthly (period=30)



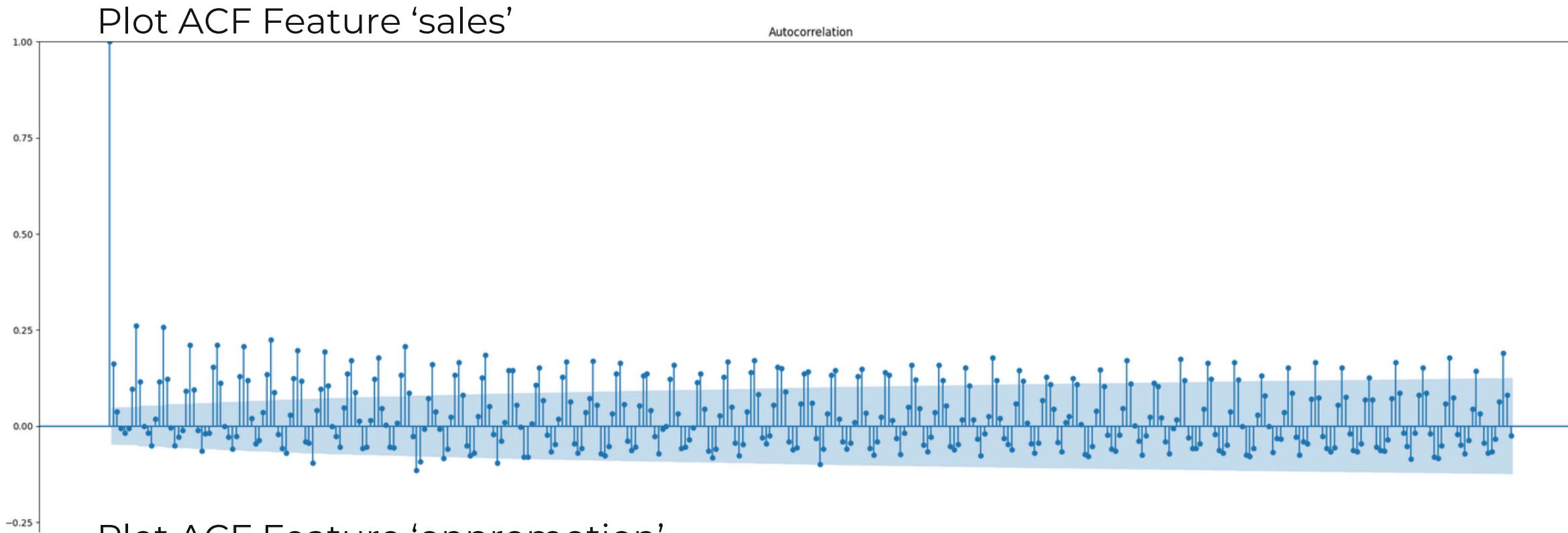
Komponen seasonal yearly (period=365)

Jika diamati, pada ketiga komponen seasonal di atas, **semuanya memiliki komponen seasonal**, bahkan pada skala yearly.

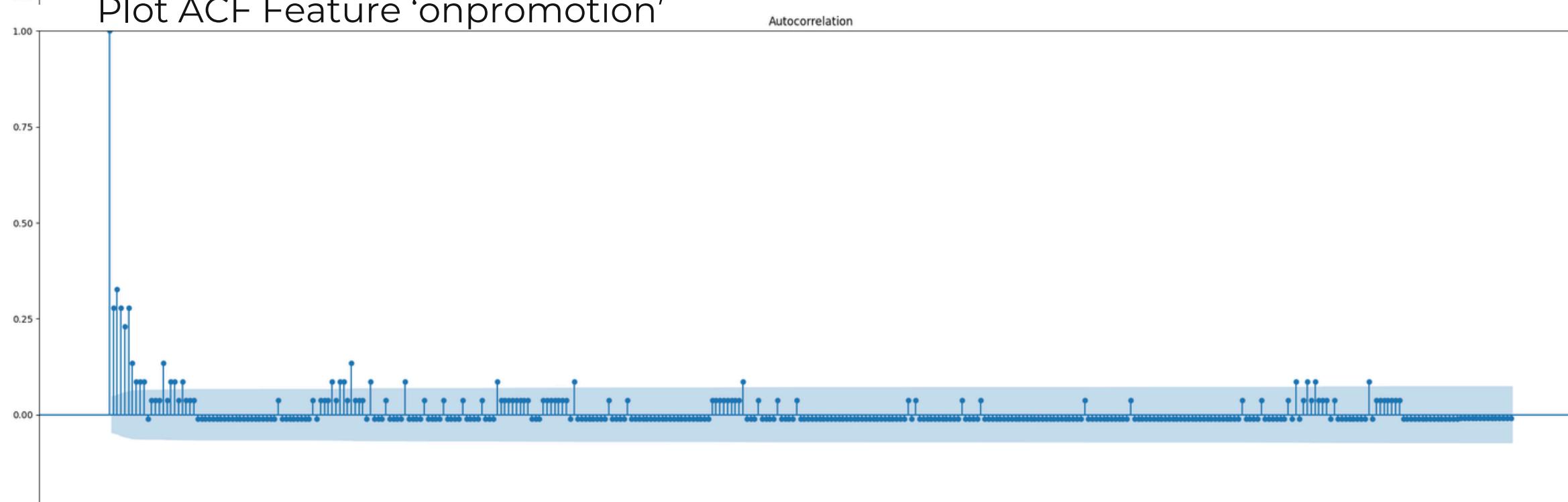
Periode seasonal **tidak ter-define dengan jelas** pada ketiga lineplot di atas (yearly)

Untuk itu, dilakukan approach lain, yaitu **melihat plot ACF**-nya.

Pengambilan Keputusan Komponen Seasonal



Pada plot ACF di kiri atas, tampak bukti bahwa **setiap 7 hari, ada garis signifikan yang keluar dari area confidence interval**. Artinya, feature 'sales' memiliki komponen seasonal yang sifatnya weekly (period=7).



Pada plot ACF di kiri bawah, tampak bukti bahwa **ada garis signifikan** yang keluar dari area confidence interval, **tetapi tidak memiliki pola tertentu**. Artinya, feature 'onpromotion' tidak memiliki komponen seasonal.

Hal ini logis karena 'onpromotion' adalah variabel pemberian promosi barang pada suatu toko yang pemberiannya sesuai dengan kebijakan toko itu.

Model Development VARMAX

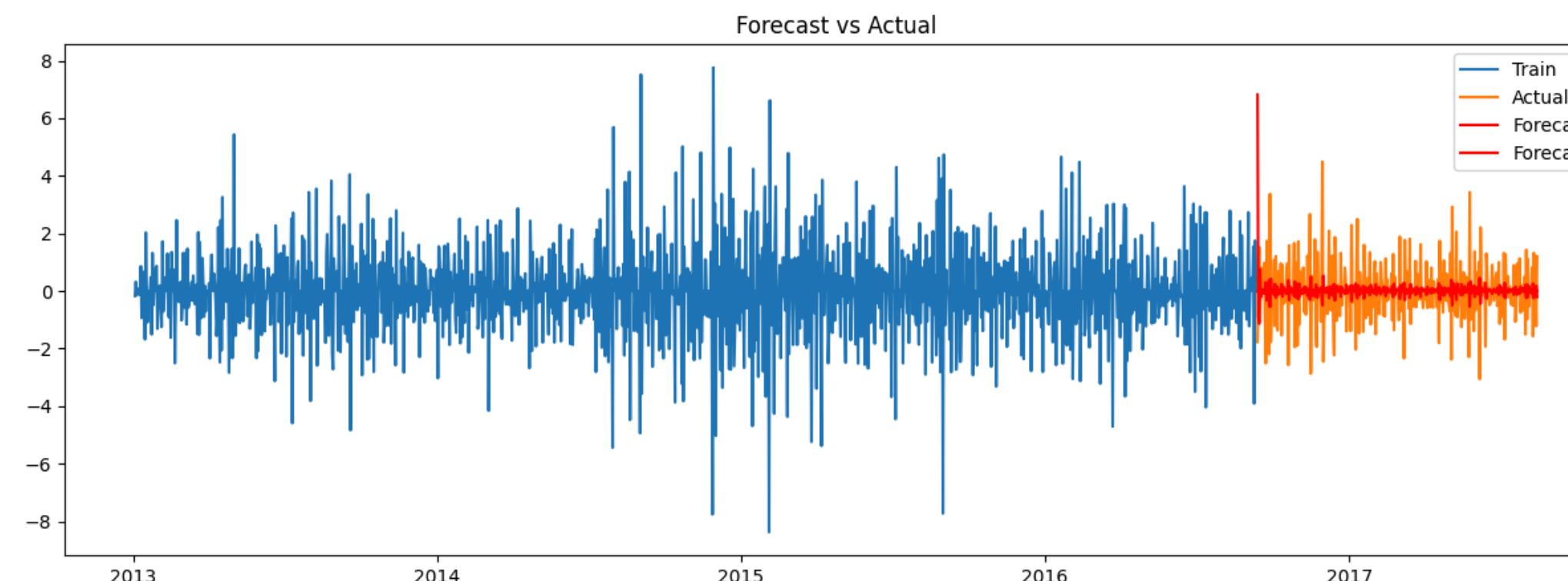
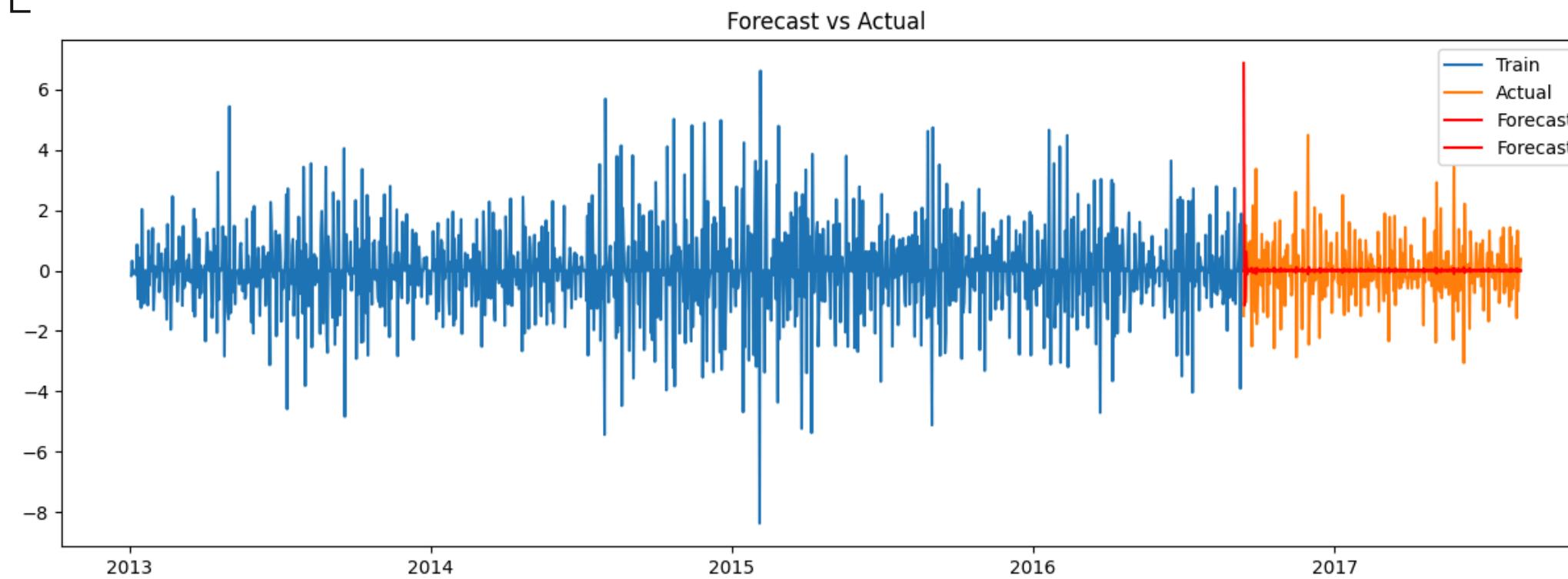
VARMAX (Vector Autoregression Moving Average with Exogenous variable) adalah algoritma pemodelan pada pengolahan data time series. Di sini, VARMAX digunakan alih-alih VAR ataupun ARIMA karena prediksi output multivariabel dan ada variabel eksogen, yaitu feature ‘dcoilwtico’.

```
AIC_auto_fwd = []
BIC_auto_fwd = []
for p in range(3):
    for q in range(3):
        model_auto_fwd = VARMAX(endog_train_automotive_fwd, exog=exog_train_automotive_fwd, order=(p+1, q+1))
        results_auto_fwd = model_auto_fwd.fit(disp=False)
        AIC_auto_int.append(results_auto_fwd.aic)
        BIC_auto_int.append(results_auto_fwd.bic)
```

Pada kode di atas, ingin dibandingkan nilai AIC dan BIC hasil keluaran tiap modelnya. AIC dan BIC terkecil yang akan menjadi model pilihan

Model Development VARMAX

family AUTOMOTIVE



LSTM

Model Development

Pada bagian ini, kami mencoba berbagai algoritma LSTM dan membandingkan algoritma LSTM mana yang terbaik.

Berikut Algoritma LSTM yang kami Gunakan:

- 1.Single LSTM
- 2.Stacked LSTM
- 3.GRU

Model Training & Optimization

Pada tahap ini kami menerapkan Bayesian Optimizer untuk mencari hyper parameter tuning terbaik dari untuk ke 5 model yang digunakan.

Berikut Algoritma LSTM yang kami Gunakan:

- 1.Single LSTM
- 2.Stacked LSTM
- 3.GRU

Results LSTM Automotive

LSTM

```
MSE Y1: 10.2609963916165  
RMSE Y1: 3.20327900620856  
MAE Y1: 2.5209291561444602  
  
MSE Y2: 0.00993773791228817  
RMSE Y2: 0.0996882034760792  
MAE Y2: 0.02480681628609697
```

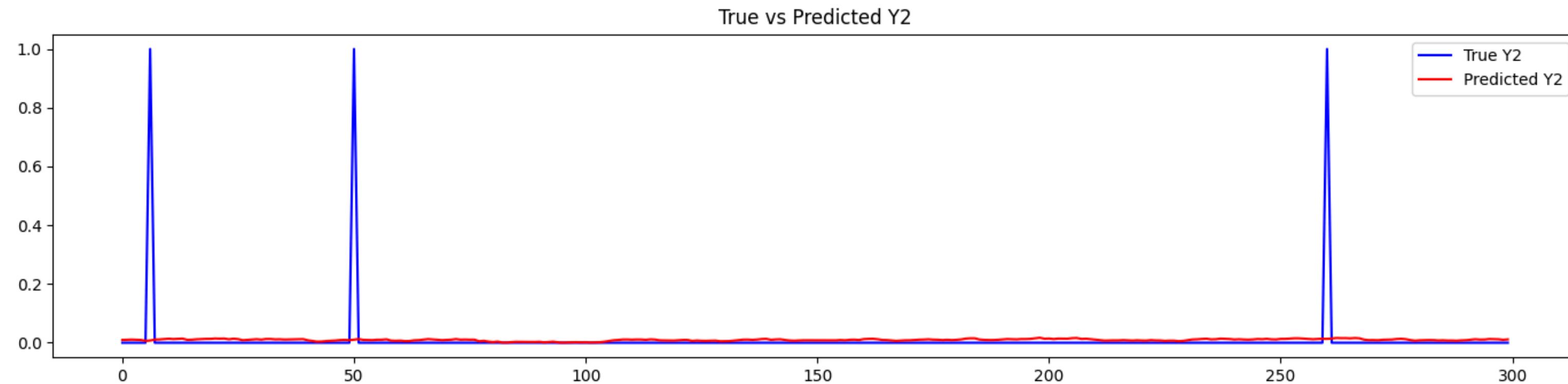
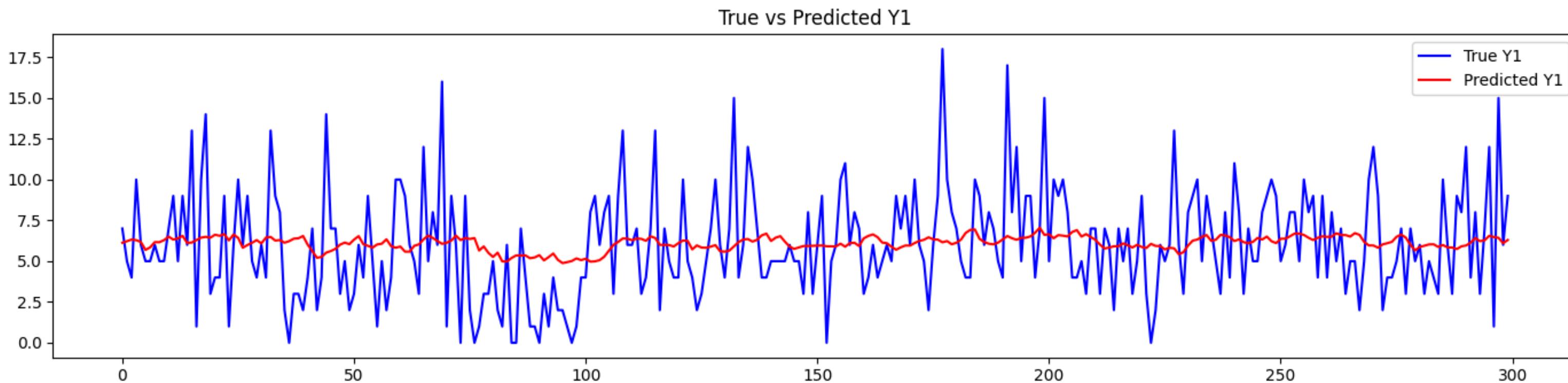
STACKED LSTM

```
MSE Y1: 10.429980638230793  
RMSE Y1: 3.2295480547950968  
MAE Y1: 2.5710357999801636  
  
MSE Y2: 0.009897322877822181  
RMSE Y2: 0.09948528975593418  
MAE Y2: 0.01939831764359648
```

GRU

```
MSE Y1: 10.425120469443348  
RMSE Y1: 3.2287955137238638  
MAE Y1: 2.502937143643697  
  
MSE Y2: 0.010086169353222197  
RMSE Y2: 0.10042992259890574  
MAE Y2: 0.014172124973071428
```

Results LSTM Automotive



Results LSTM GROCERY 1

LSTM

MSE Y1: 352548.77259261766
RMSE Y1: 593.758176863795
MAE Y1: 446.124541015625

MSE Y2: 230.02043617361798
RMSE Y2: 15.166424633829095
MAE Y2: 9.69217856725057

STACKED LSTM

MSE Y1: 352983.57617333176
RMSE Y1: 594.1242093816172
MAE Y1: 431.1924104817708

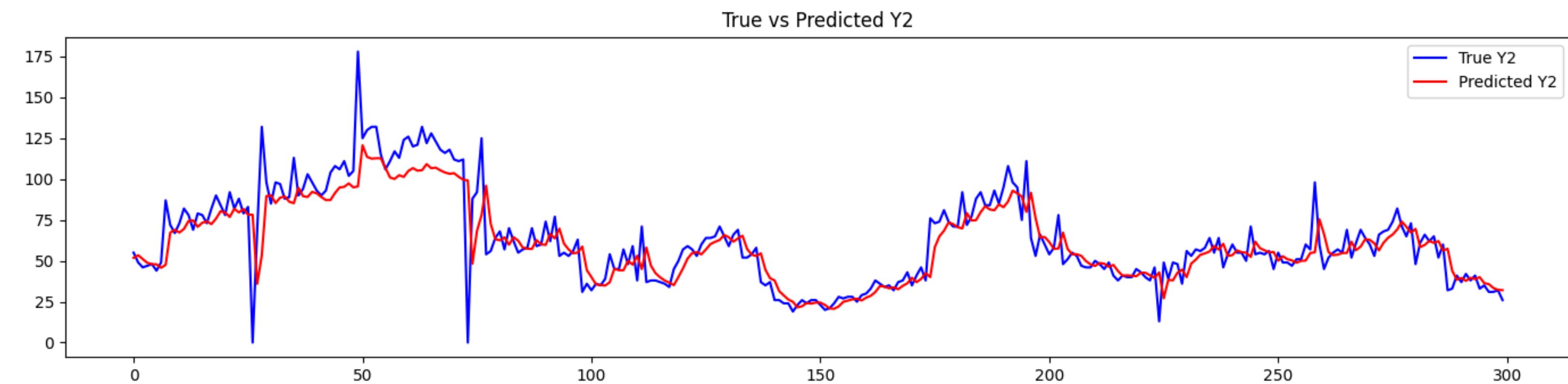
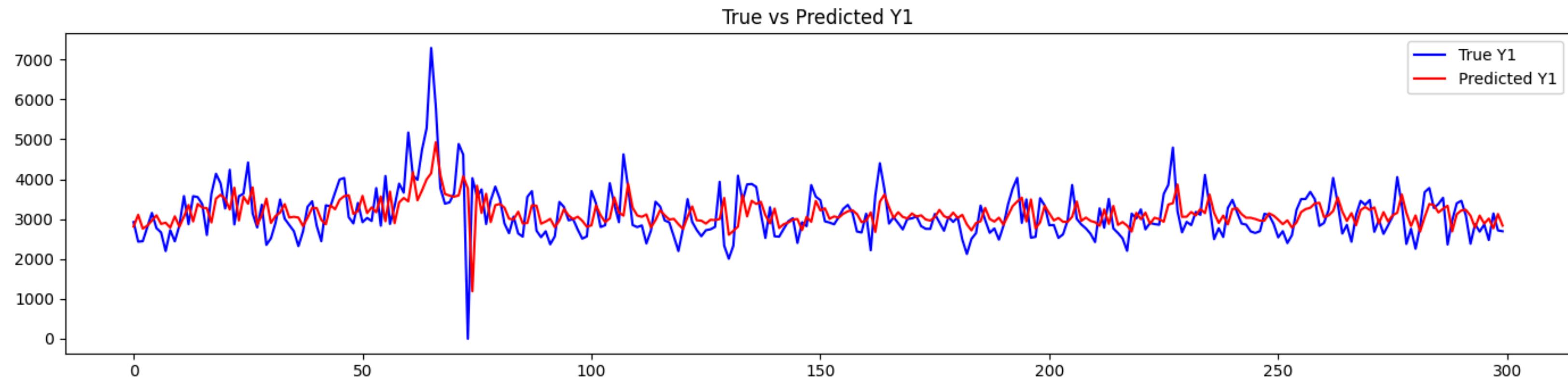
MSE Y2: 338.5816485299347
RMSE Y2: 18.400588265866247
MAE Y2: 13.1119189453125

GRU

MSE Y1: 343758.5173530601
RMSE Y1: 586.3092335560307
MAE Y1: 415.21027913411456

MSE Y2: 226.99956513133054
RMSE Y2: 15.066504741688782
MAE Y2: 9.029496828715006

Results LSTM Grocery 1



Results LSTM BEAUTY

LSTM

```
MSE Y1: 10.2609963916165  
RMSE Y1: 3.20327900620856  
MAE Y1: 2.5209291561444602  
  
MSE Y2: 0.00993773791228817  
RMSE Y2: 0.0996882034760792  
MAE Y2: 0.02480681628609697
```

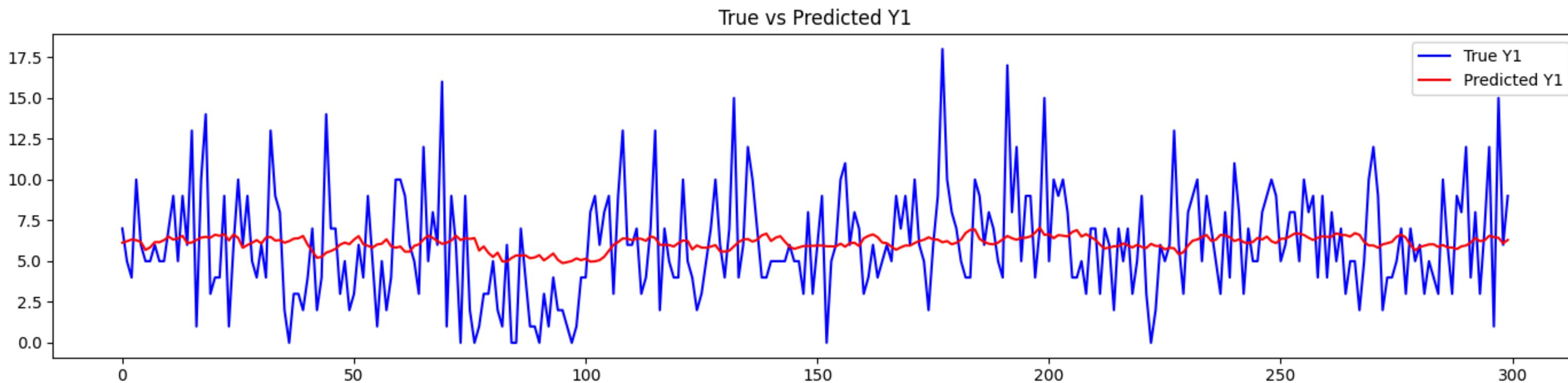
STACKED LSTM

```
MSE Y1: 10.429980638230793  
RMSE Y1: 3.2295480547950968  
MAE Y1: 2.5710357999801636  
  
MSE Y2: 0.009897322877822181  
RMSE Y2: 0.09948528975593418  
MAE Y2: 0.01939831764359648
```

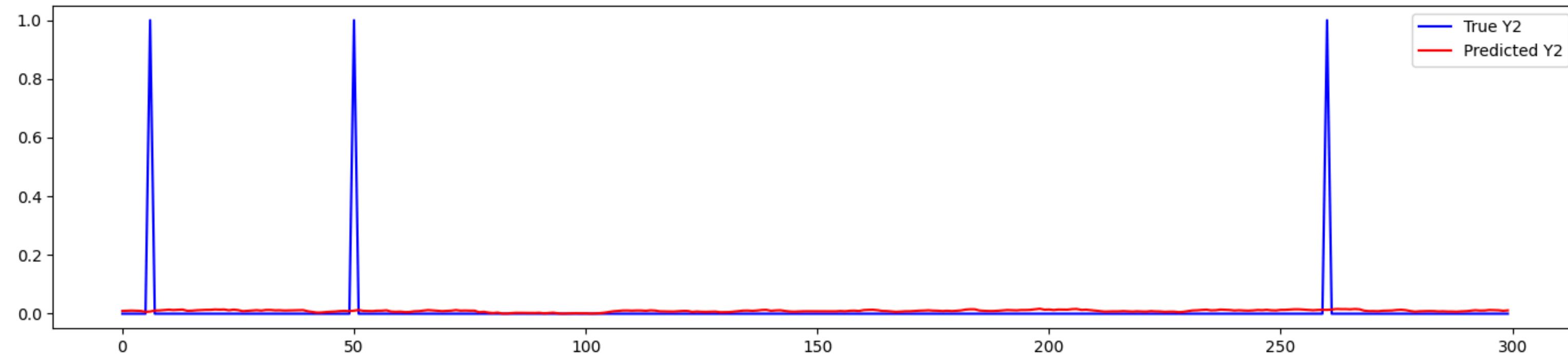
GRU

```
MSE Y1: 10.425120469443348  
RMSE Y1: 3.2287955137238638  
MAE Y1: 2.502937143643697  
  
MSE Y2: 0.010086169353222197  
RMSE Y2: 0.10042992259890574  
MAE Y2: 0.014172124973071428
```

Results LSTM Automotive



True vs Predicted Y2



Real-world Application

Examples of how your AI solution has been or could be applied in real scenarios.

Whether your AI solution would be deployed as web application, edge device, CCTV etc.

Future Improvement

Potential limitations of the current solution.

Ideas for further development and improvement.

1. Di sini, penggunaan dcoilwtico sebagai variabel eksogen tidak memiliki dasar statistik yang kuat. Untuk itu, bisa digunakan metode statistik untuk melihat ke-'pantas'-an feature ini sebagai variabel eksogen.
2. Perdalam pemahaman tentang data dan hal-hal lain yang memungkinkan untuk dijadikan variabel eksogen.
3. Explorasi Hyperparameter tuning untuk mengoptimalkan model.

Conclusion

Recap of the key points presented.

Emphasis on the positive impact your AI solution brings.

Invitation for questions and discussions.



Contact Us

Don't hesitate to contact us for further inquiries or any collaborations.

Alwy



Daniel



Your Full Name

