Project 1
Sales Forecasting

Group: Alan Turing

September 30th, 2024

Meet the Team

DENDI SUNARDI

Your Role/Contribution

dendisunardi

MISBAHUL MUNIR

LSTM Modeling

Misbahul Munir

Chelsea Castro

Dataset Preparation & Initial Analysis

Chelsea Castro

Teddy Budiman

Your Role/Contribution

Linkedin Name

Ismail

Your Role/Contribution

Linkedin Name

HINU HARDHANTO

Model quality control

hinu hardhanto

Background & Problem Statement

Dataset ini mencakup 23 kategori produk untuk toko nomor 5, dengan **data penjualan dari 1 Januari 2013 hingga 31 Agustus 2017**. Untuk menganalisis penjualan masa lalu, tren, dan komponen musiman, kami memasukkan data hari libur dan acara khusus. Selain itu, **harga minyak harian akan menjadi faktor penting**, karena Ekuador adalah negara yang bergantung pada minyak, yang membuat kondisi ekonominya sangat rentan terhadap fluktuasi harga minyak.

Asumsi:

- **Pasar minyak beroperasi terus-menerus (24/7),** tidak seperti pasar saham, sehingga harga minyak akan diinterpolasi untuk mengisi nilai yang hilang.
- Memasukkan data promosi tidak menyebabkan kebocoran data, karena perusahaan menentukan tanggal promosi sebelumnya, dan toko sudah mengetahui tanggal-tanggal tersebut ketika membuat perkiraan.

Caveat yang tidak menjadi pertimbangan

- **Gaji di sektor publik dibayarkan dua minggu sekali,** pada tanggal 15 dan hari terakhir setiap bulan, yang mungkin berdampak pada penjualan supermarket.
- **Pada 16 April 2016, gempa berkekuatan 7,8 skala Richter melanda Ekuador**. Dalam beberapa minggu berikutnya, terjadi peningkatan signifikan dalam pembelian barang-barang kebutuhan pokok, karena masyarakat bergerak untuk mendukung upaya bantuan, yang sangat memengaruhi penjualan.

Objectives & Scope

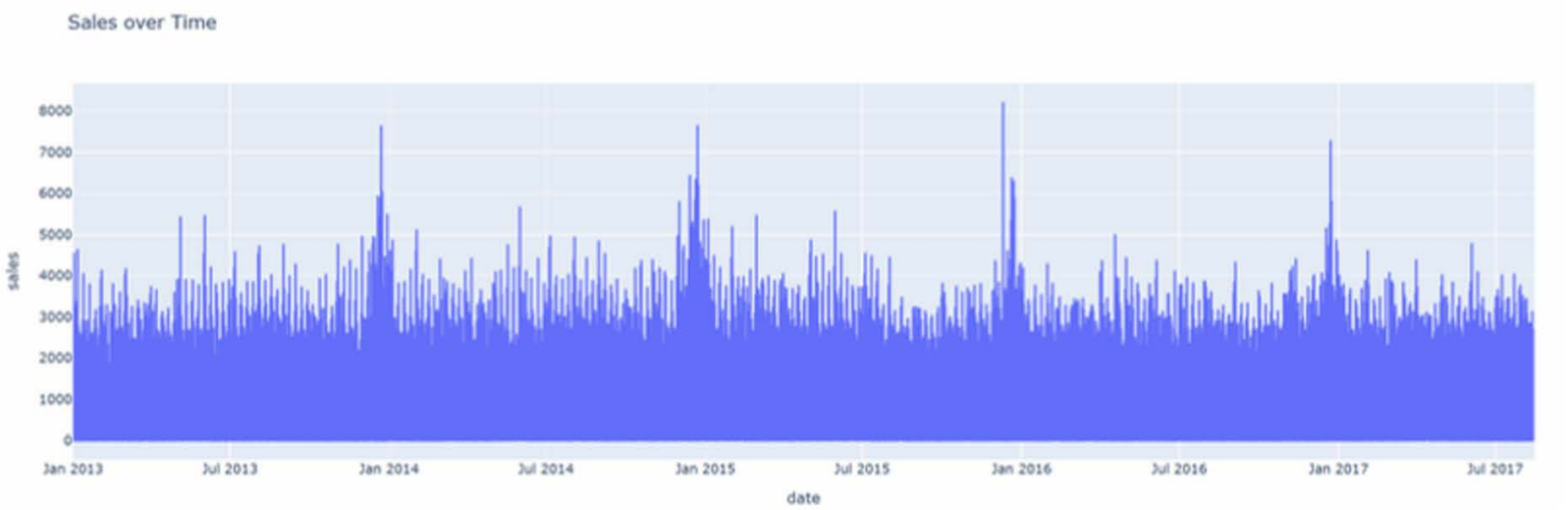
Projek ini membutuhkan pengembangan sistem cerdas berbasis AI untuk peramalan penjualan penjualan toko. Sistem ini akan membantu memperoleh nilai penjualan pada masa depan sehingga mampu menyiapkan stock sesuai permintaan.

Tim Engineering diminta untuk melihat data dan melakukan eksperimen dengan ragam teknik preprocessing, juga menguji algoritma ARIMA & LSTM dan optimalkan penggunaannya (hyperparameter tuning).

Dari hasil eksperimen, lakukan evaluasi dan penarikan kesimpulan mana algoritma terbaik.

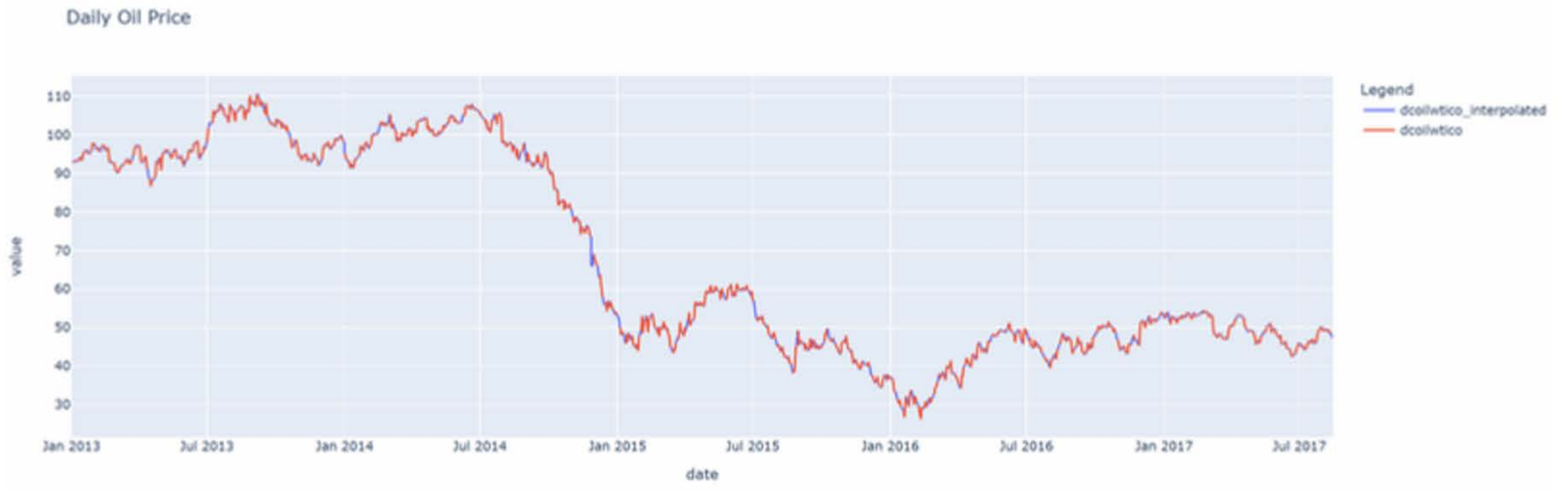
1. **Bereksperimen dengan algoritma ARIMA & LSTM.**
2. **Melakukan evaluasi dan penarikan kesimpulan.**
3. **Publikasi ke Github.**

Data Collection & Preparation

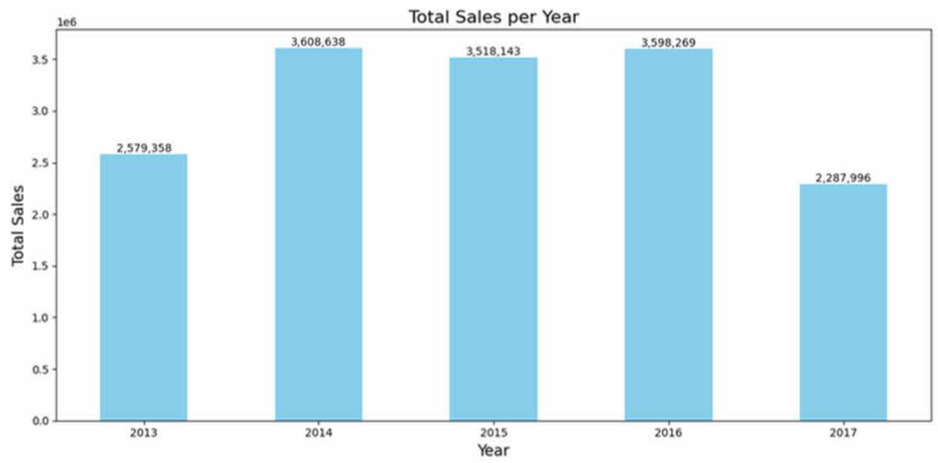


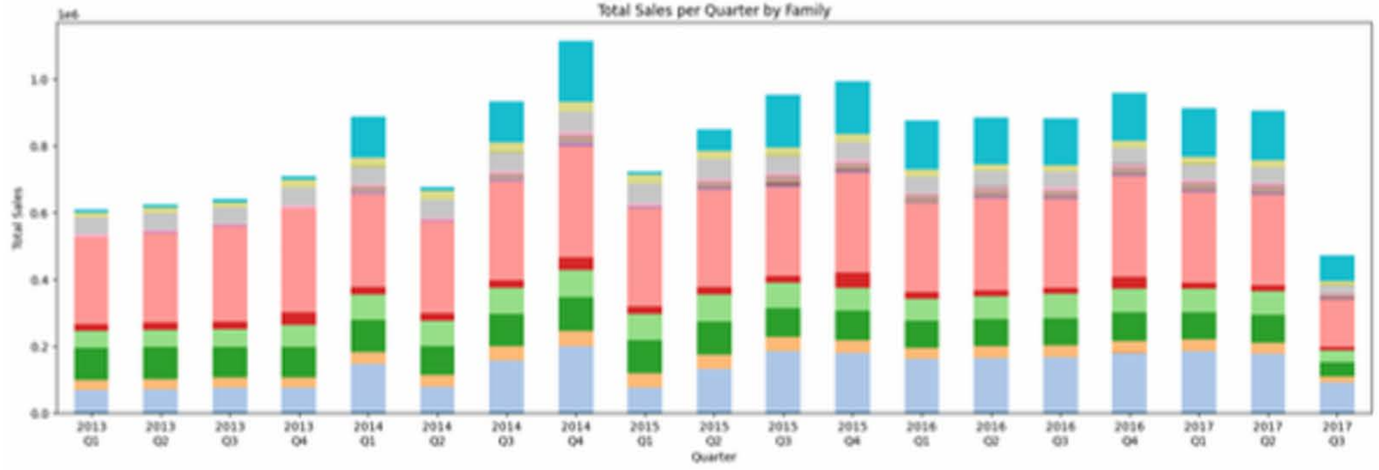
Grafik penjualan menunjukkan adanya **pola musiman dengan puncak penjualan yang terjadi setiap tahun**, terutama menjelang akhir tahun dan awal tahun berikutnya.

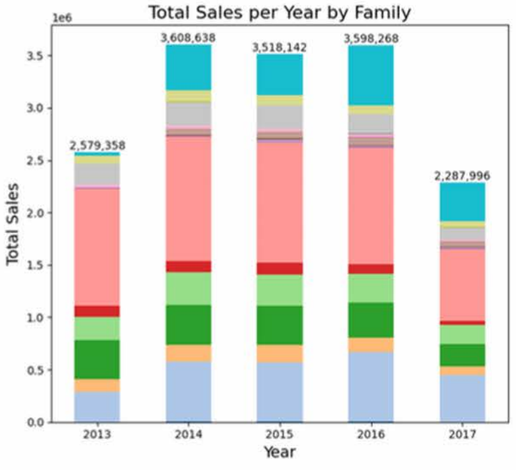
Kami menggunakan interpolasi daripada langsung menggunakan **data Jumat sebelumnya untuk akhir pekan guna memberikan kontinuitas data yang lebih smooth dan realistis** serta memberikan representasi yang lebih baik untuk model yang membutuhkan data yang konsisten dan berkelanjutan.



Data Collection & Preparation

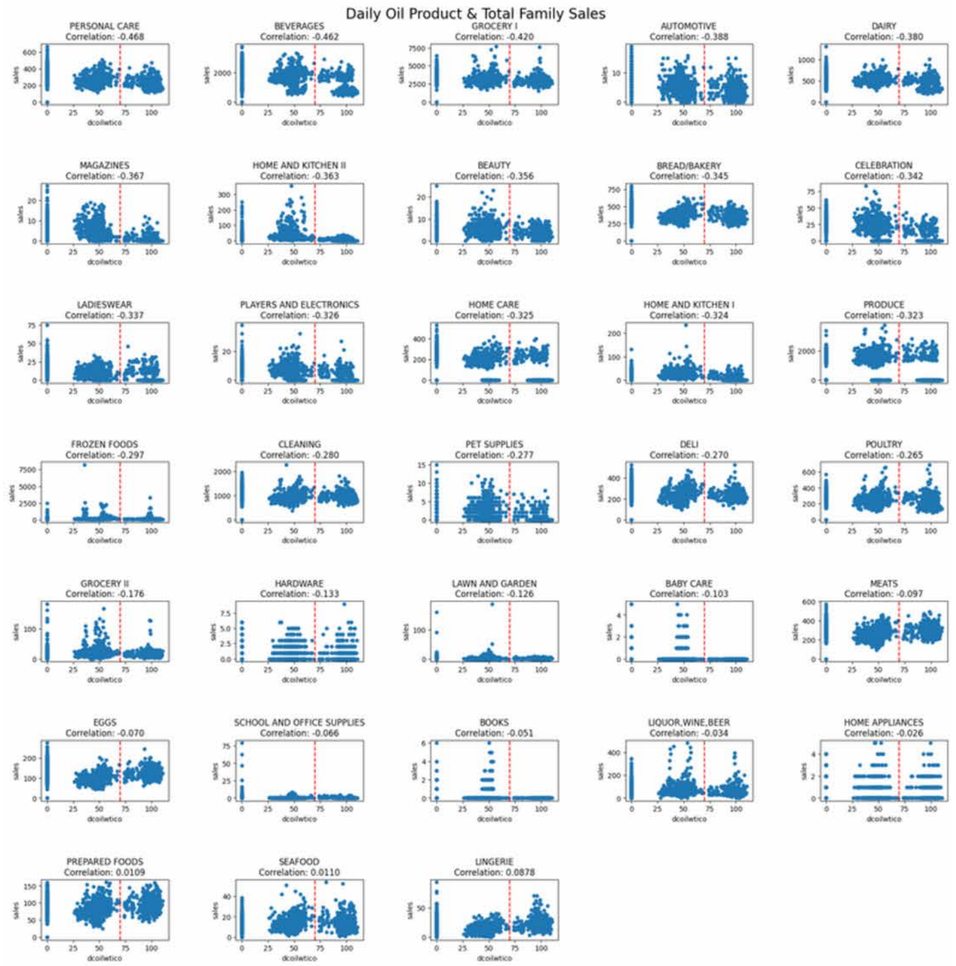






Add a little bit of body text

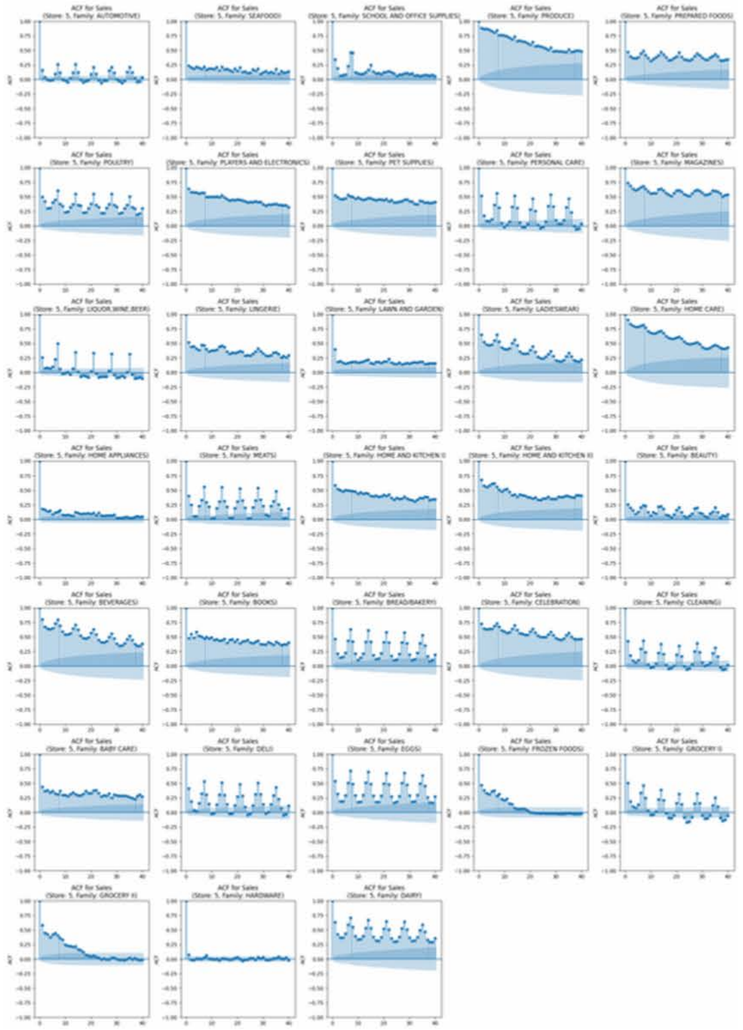
Data Collection & Preparation

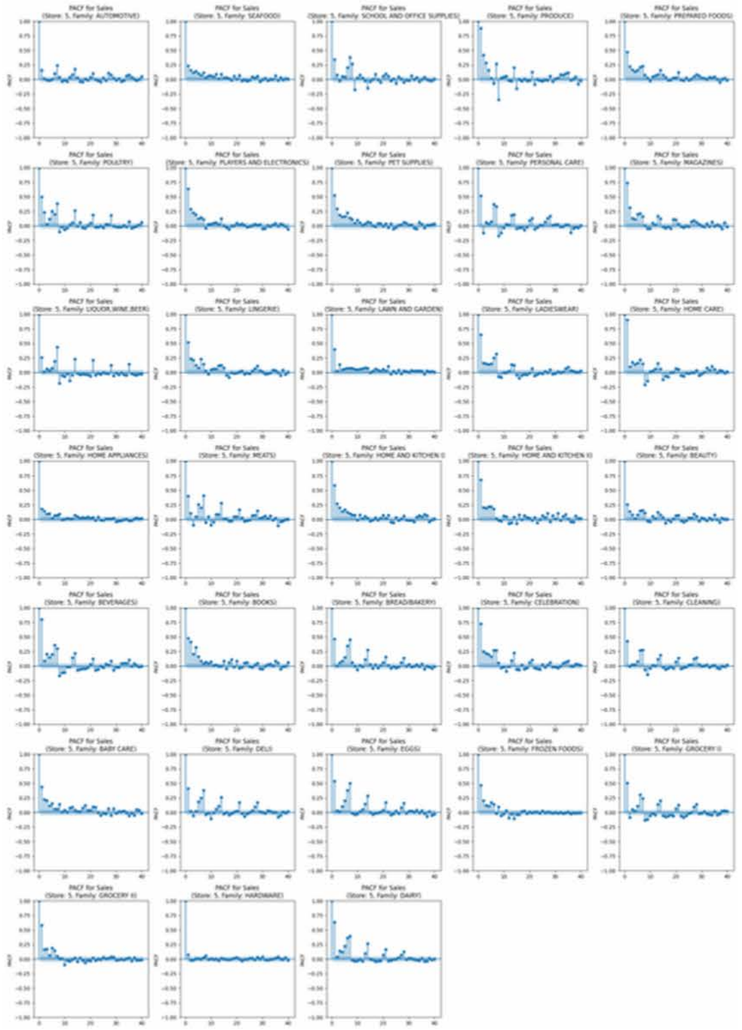


Summary :

- **Personal Care** memiliki korelasi negatif terkuat (-0.468), menunjukkan bahwa peningkatan harga minyak cenderung menurunkan penjualan di kategori ini.
- **Beverages** dan **Grocery I** juga menunjukkan korelasi negatif yang kuat, menandakan bahwa harga minyak bisa berhubungan dengan penurunan penjualan di kategori-kategori ini.
- Beberapa kategori seperti **Magazines** dan **Home and Kitchen II** menunjukkan korelasi positif, menunjukkan adanya peningkatan penjualan saat diskon minyak meningkat.
- Kategori seperti **Seafood, Prepared Foods, dan Home Appliances** memiliki korelasi yang hampir netral, menunjukkan harga minyak tidak memiliki dampak besar terhadap penjualan.
- Kategori **Ladieswear** dan **Players and Electronics** memiliki korelasi negatif moderat, menunjukkan adanya hubungan yang signifikan antara harga minyak dan penurunan penjualan.

Data Collection & Preparation





- Kategori seperti Meat, Egg, Dairy, dan Frozen Foods menunjukkan **pola penjualan berkala, dengan PACF tinggi menunjukkan siklus yang dapat diprediksi.**
- Kategori seperti Seafood, Poultry, dan Electronics **mudah diprediksi dalam jangka pendek**, sementara kategori seperti Automotive, Cleaning, dan Baby Care **sulit diprediksi.**
- Kategori seperti Alcohol, Lawn & Garden, dan Home Care menunjukkan **pola penurunan bertahap, mengindikasikan tren berkelanjutan jangka panjang.**

Model Development

Untuk pemodelan menggunakan LSTM, ada beberapa jenis arsitektur yang digunakan untuk nantinya dilakukan perbandingan, yaitu :

1. Single Layer LSTM

- 50 units
- 0.2 dropout rate
- adam optimizer
- 0.001 learning rate
- 100 epochs
- 32 batch size

2. Bidirectional LSTM

- 50 units
- 0.2 dropout rate
- adam optimizer
- 0.001 learning rate
- 100 epochs
- 32 batch size

3. Stacked LSTM

- 50 units
- 0.2 dropout rate
- adam optimizer
- 0.001 learning rate
- 100 epochs
- 32 batch size
- 2 layers

4. GRU

- 50 units
- 0.2 dropout rate
- adam optimizer
- 0.001 learning rate
- 100 epochs
- 32 batch size

Training & Optimization

Proses training model dilakukan menggunakan dataset dengan pembagian :

Training data	**Validation data**	**Testing data**
2013-01-01 - 2015-12-31	2016-01-01 - 2016-12-31	2017-01-01 - 2017-08-15

Menggunakan **Min-Max Scaler**:

- Memastikan skala yang konsisten untuk semua fitur.
- Membantu optimizer bekerja lebih efisien.
- Mempercepat konvergensi model.

Fitur :

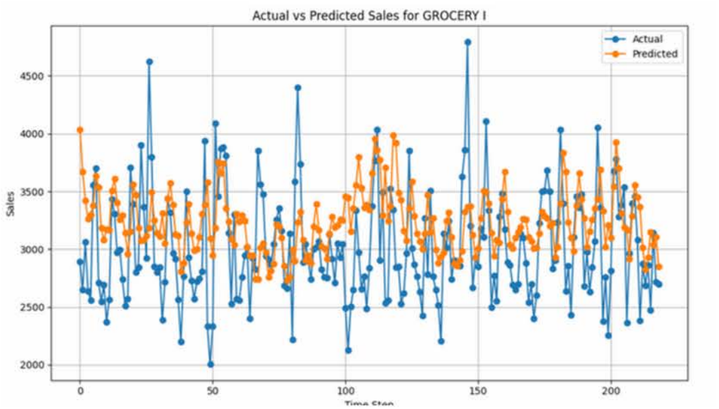
- sales.
- onpromotion.
- dcoilwtico.

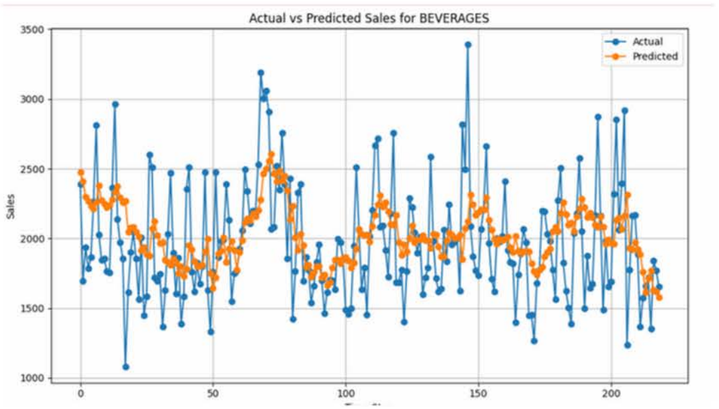
Training & Optimization

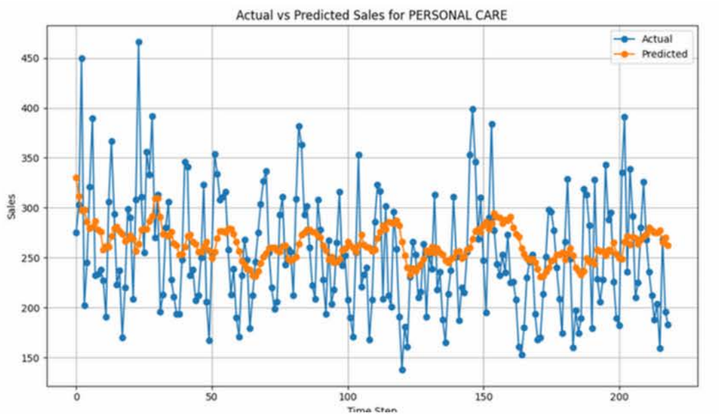
Hasil Evaluasi Training Model dengan Initial Parameters

Family	Metric	bi_lstm	gru	lstm	stacked_lstm
BEVERAGES	MAE	346.060924	312.716257	335.970584	353.768471
	MAPE	19.179210	16.891053	18.435617	19.602411
	MSE	181054.364934	156640.114877	175614.328483	184007.197030
	RMSE	425.504835	395.777861	419.063633	428.960601
GROCERY I	MAE	568.778399	353.286527	355.167334	384.856903
	MAPE	20.344996	12.087299	11.903998	13.002440
	MSE	484310.067419	201741.315672	204113.454598	234948.152302
	RMSE	695.923895	449.156226	451.789171	484.714506
PERSONAL CARE	MAE	58.430737	49.034552	57.725213	49.212728
	MAPE	26.302326	21.404039	25.956629	20.605745
	MSE	4710.256430	3405.241151	4596.338542	3555.399595
	RMSE	68.631308	58.354444	67.796302	59.627172

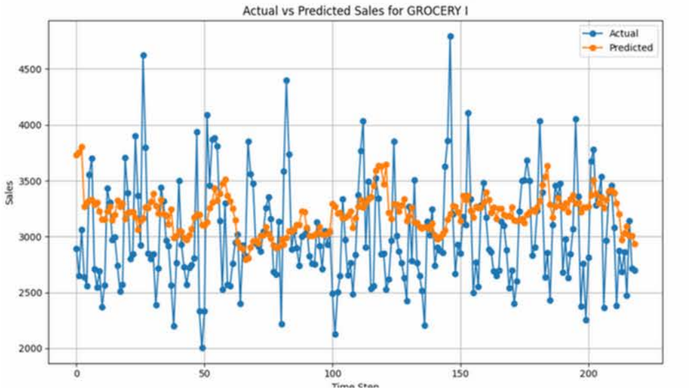
Training & Optimization

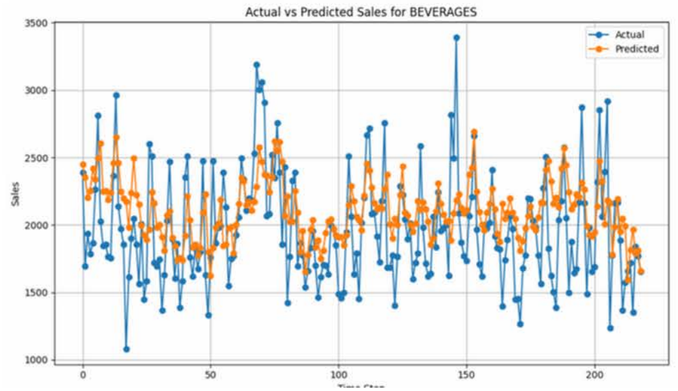


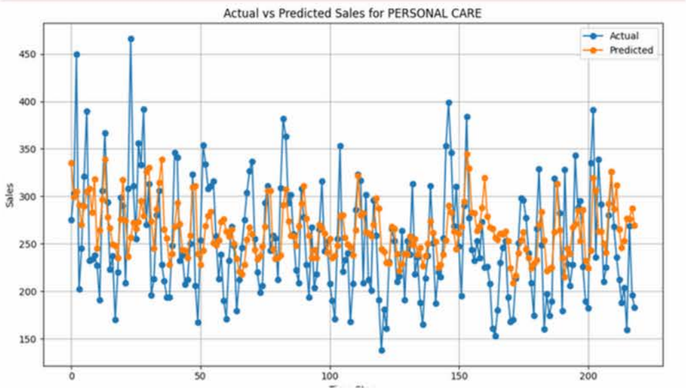




Single LSTM

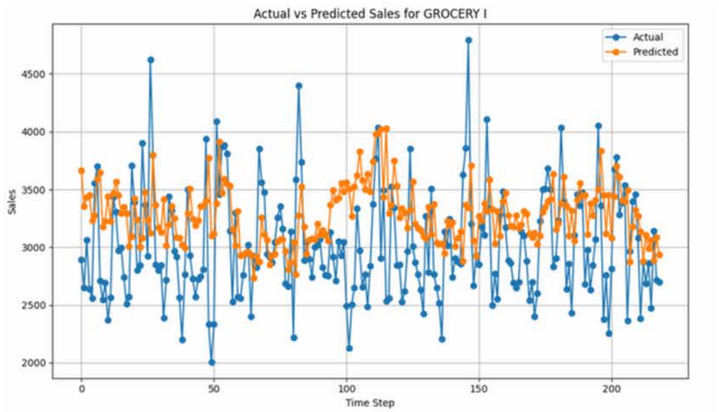


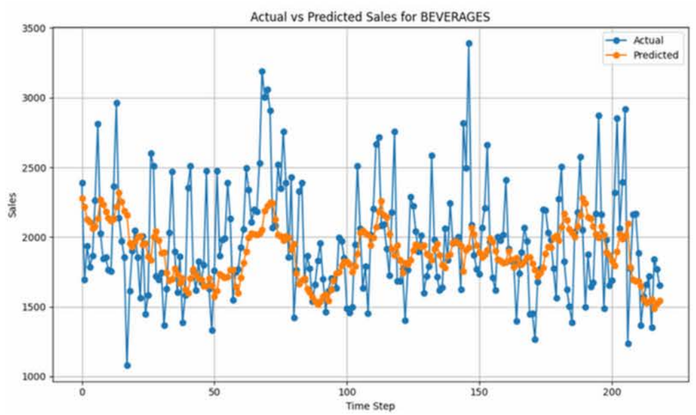


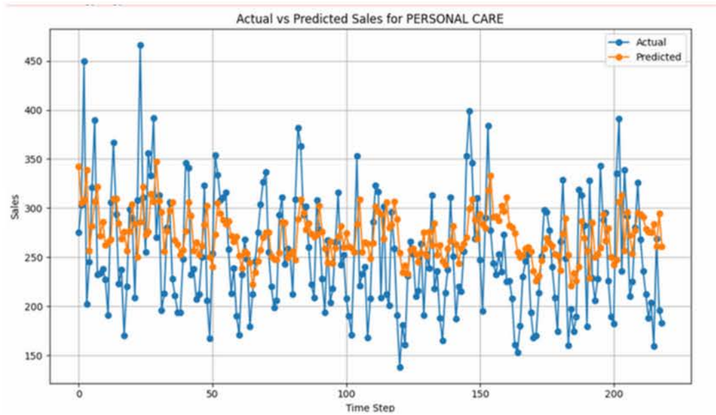


Bi LSTM

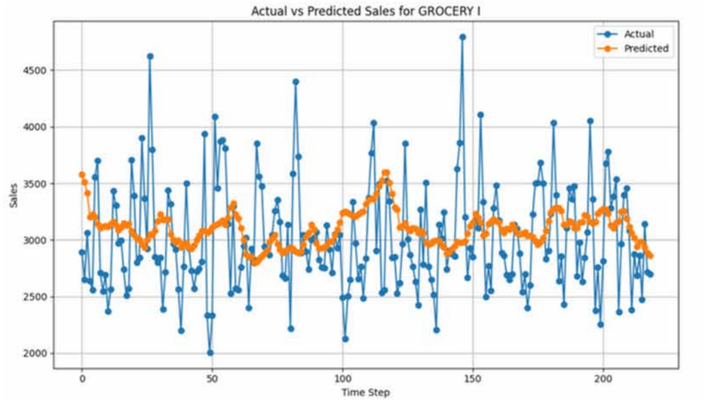
Training & Optimization

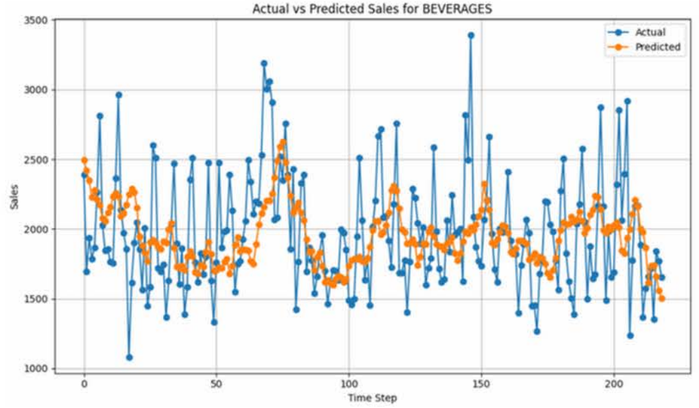


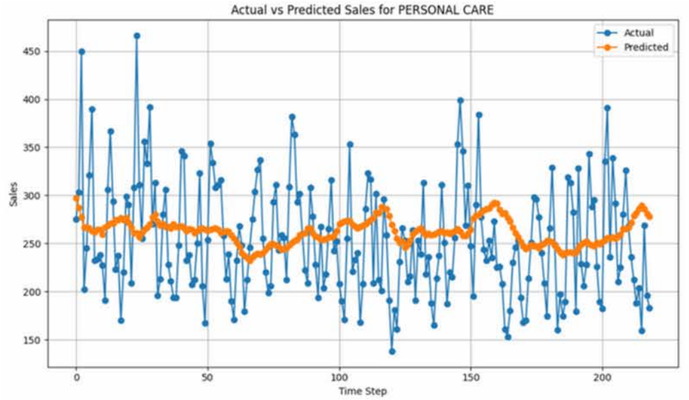




GRU







Stacked LSTM

Training & Optimization

Menggunakan metode Bayesian Optimization untuk melakukan Hyperparameter Tuning

Parameter	Bound
units	50 - 150
dropout_rate	0.2 - 0.5
learning_rate	0.0001 - 0.01
batch_size	16 - 64
epochs	50 - 200
n_layers	2 - 5

Training & Optimization

Best Hyperparameter

Family	Model / Param	bi_lstm	gru	lstm	stacked_lstm
BEVERAGES	best_batch_size	21.000000	25.000000	24.000000	16.000000
	best_dropout_rate	0.257437	0.200000	0.200000	0.200000
	best_epochs	111.000000	186.000000	179.000000	139.000000
	best_learning_rate	0.009571	0.010000	0.010000	0.010000
	best_n_layers	NaN	NaN	NaN	2.000000
	best_units	114.000000	115.000000	142.000000	142.000000
GROCERY I	best_batch_size	38.000000	33.000000	64.000000	29.000000
	best_dropout_rate	0.200000	0.500000	0.253963	0.327222
	best_epochs	146.000000	164.000000	124.000000	138.000000
	best_learning_rate	0.010000	0.010000	0.010000	0.003454
	best_n_layers	NaN	NaN	NaN	2.000000
	best_units	72.000000	51.000000	83.000000	51.000000
PERSONAL CARE	best_batch_size	23.000000	45.000000	24.000000	43.000000
	best_dropout_rate	0.217425	0.241848	0.291273	0.500000
	best_epochs	179.000000	93.000000	128.000000	70.000000
	best_learning_rate	0.006051	0.003727	0.004376	0.008477
	best_n_layers	NaN	NaN	NaN	2.000000
	best_units	120.000000	95.000000	79.000000	50.000000

Training & Optimization

Hasil Evaluasi Metrik menggunakan Best Hyperparameter

Model		bi_lstm	gru	lstm	stacked_lstm
Family	Metric				
BEVERAGES	MAE	5.382848e+02	6.575488e+02	541.362035	276.182267
	MAPE	3.027886e+01	3.561089e+01	30.177188	14.253791
	MSE	5.232762e+05	6.808138e+05	408828.401509	132921.142188
	RMSE	7.233783e+02	8.251144e+02	639.396905	364.583519
GROCERY I	MAE	3.855001e+03	7.846033e+02	519.476014	605.731268
	MAPE	1.330224e+02	2.755922e+01	18.400411	20.987980
	MSE	2.651849e+07	1.205010e+06	397515.897060	562411.272975
	RMSE	5.149611e+03	1.097729e+03	630.488618	749.940846
PERSONAL CARE	MAE	6.273058e+01	5.501547e+01	73.679581	46.947010
	MAPE	2.624090e+01	2.431403e+01	32.547958	19.807926
	MSE	7.626927e+03	4.475535e+03	8930.070856	3351.631215
	RMSE	8.733228e+01	6.689944e+01	94.499052	57.893274

Real-world Application

- Forecasting dapat digunakan sebagai:
  - Sistem stock barang
  - Pricing Adjustment
  - Monitoring Expiry products
  - Insight untuk promotion strategy

Future Improvement

- Hari gajian dan tanggal ganda, yang belum termasuk dalam data saat ini, mungkin dapat memengaruhi peningkatan kualitas data di masa mendatang.
- Menambahkan metode hyperparameter tuning lainnya, seperti Grid Search dengan Cross-Validation, Bayesian Optimization, dll.

Conclusion

Data penjualan diproses melalui pembersihan dan normalisasi, kemudian dianalisis untuk mengidentifikasi pola musiman dan tren. Model yang digunakan mencakup Single LSTM, LSTM, Bi-LSTM, Stack LSTM, dan GRU, dengan hasil evaluasi melalui MSE yang menunjukkan akurasi yang memadai.

Variabel eksternal penting, seperti hari gajian dan tren belanja, belum dimasukkan, namun dapat meningkatkan kualitas prediksi. Untuk perbaikan, disarankan untuk menambahkan variabel tersebut dan melakukan tuning hyperparameter guna mengoptimalkan model.

Terima Kasih