Walmart

Save money. Live better.

PREDICTION ON WALMART SALES DEMAND

Business Objective

- Efficiently allocate resources and accurately estimate costs and revenue based on the prediction of short-term and long-term sales performance

- Produce the 28 days ahead point forecasts for 30490 items in 42,840 time series that represent the hierarchical unit sales of Walmart, starting at the item level and aggregating to that of departments, product categories and stores in three geographical areas in US

Approaches

Machine Learning

- LightGBM

Statistics

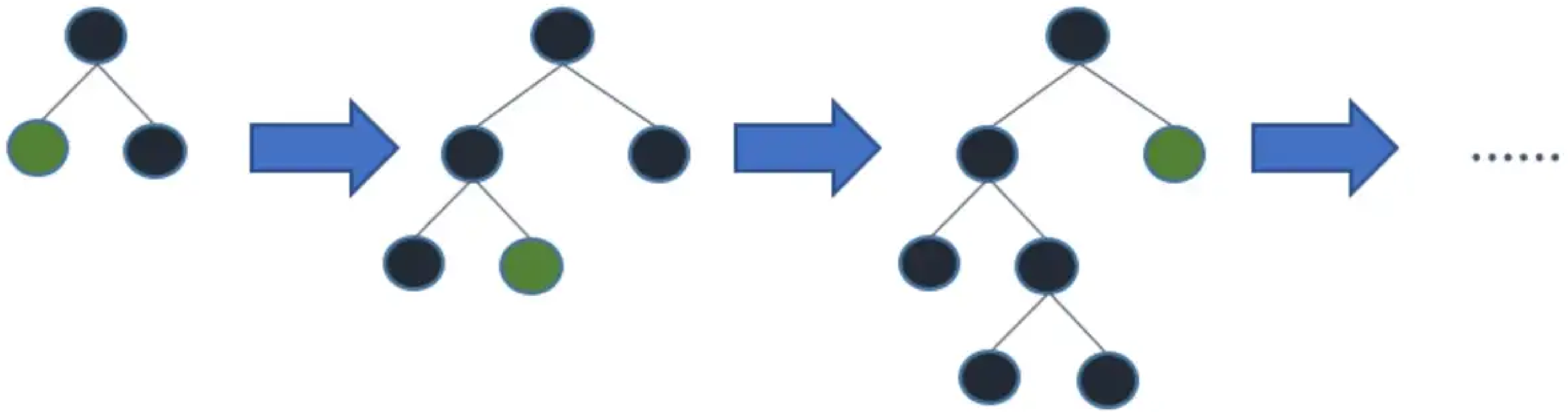
- SARIMAX
- Prophet

Deep Learning

- Seq2seq
- Transformer

LightGBM

- Type of ensemble modeling
- Uses tree-based learning algorithm
- Prefixed as 'light' because of its high speed



Leaf-wise tree growth

SARIMAX

- S: Seasonality
- AR: Autoregression
- I: Integrated
- MA: Moving Average
- X: Exogenous Variable

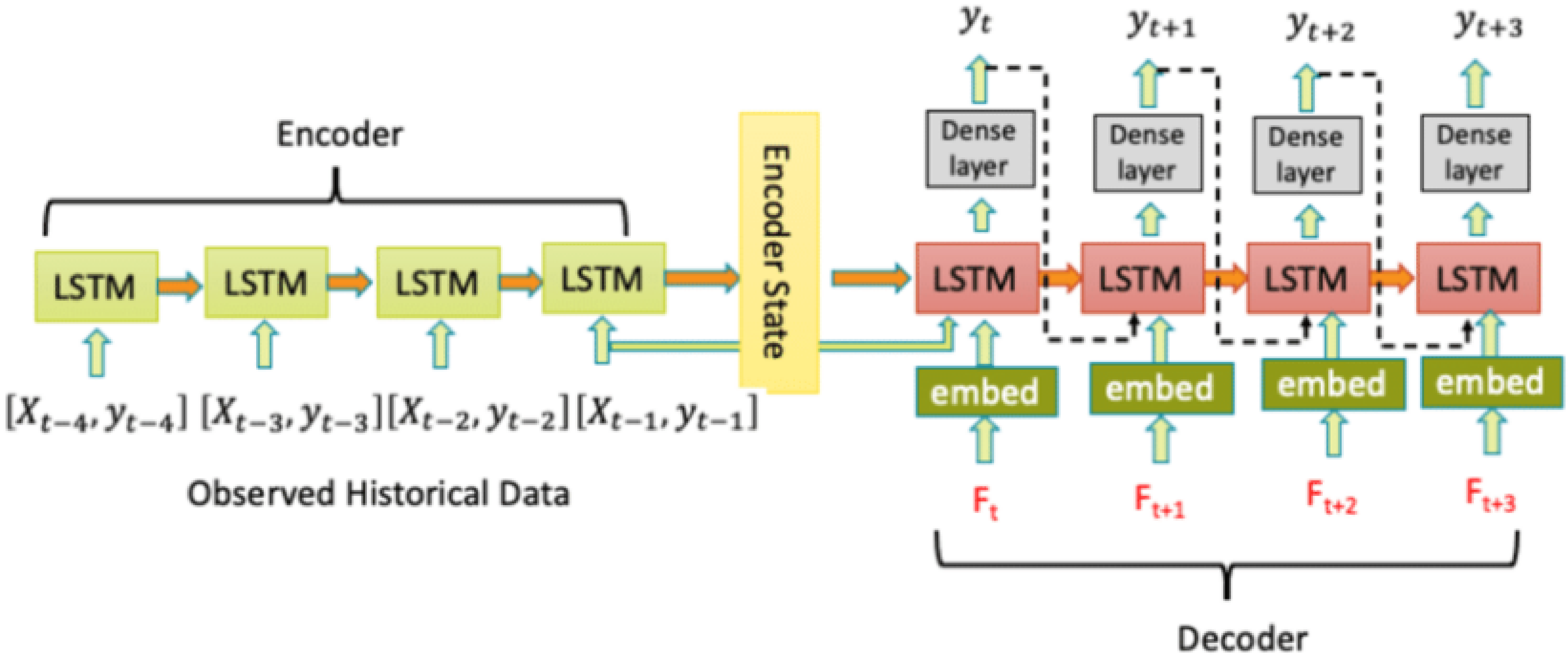
Prophet

- Trend component
- Seasonal component
- Holiday component

$$y_t = g(t) + s(t) + h(t)$$

Seq2seq

- Encoder
- Decoder



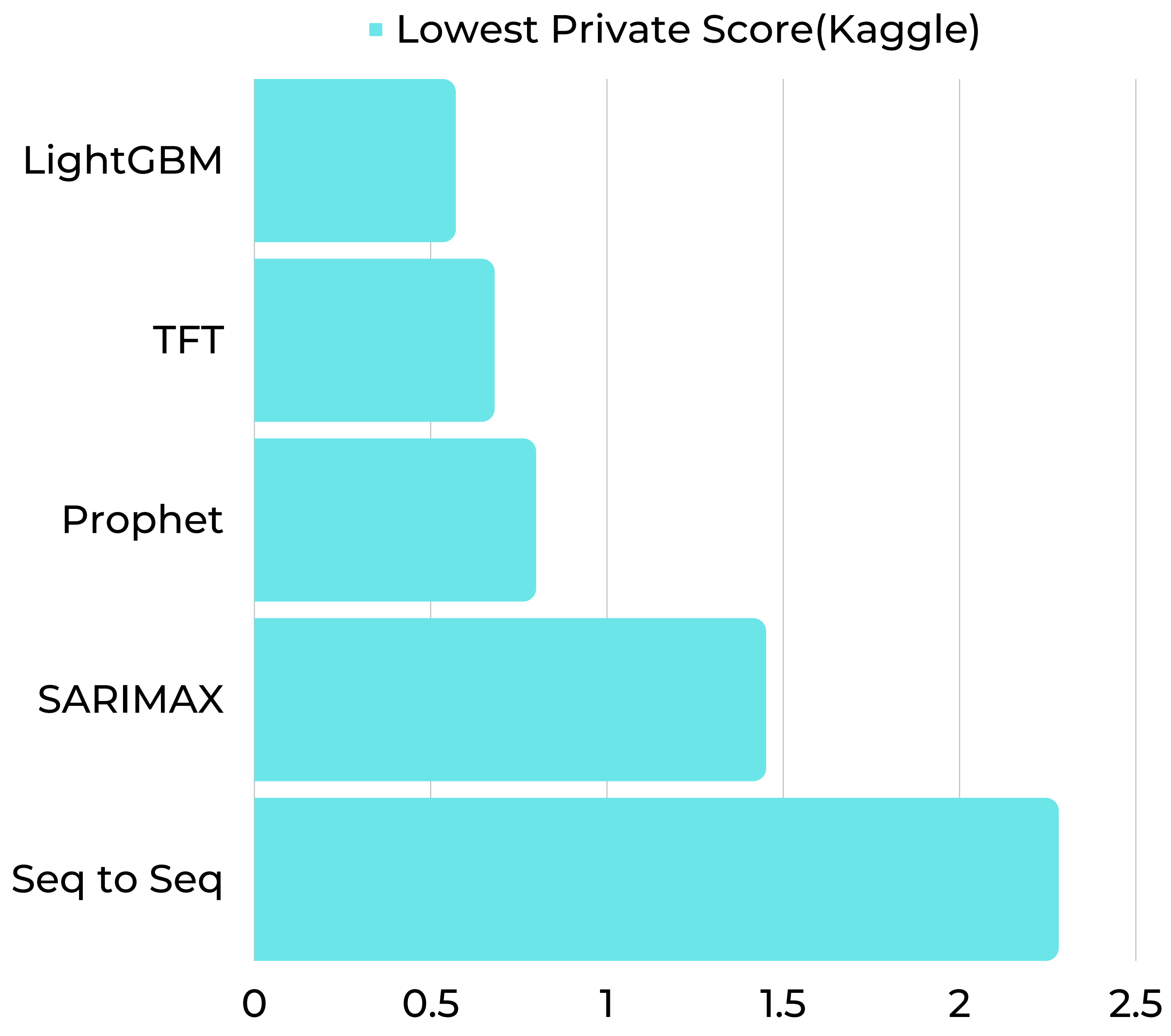
Temporal Fusion Transformer

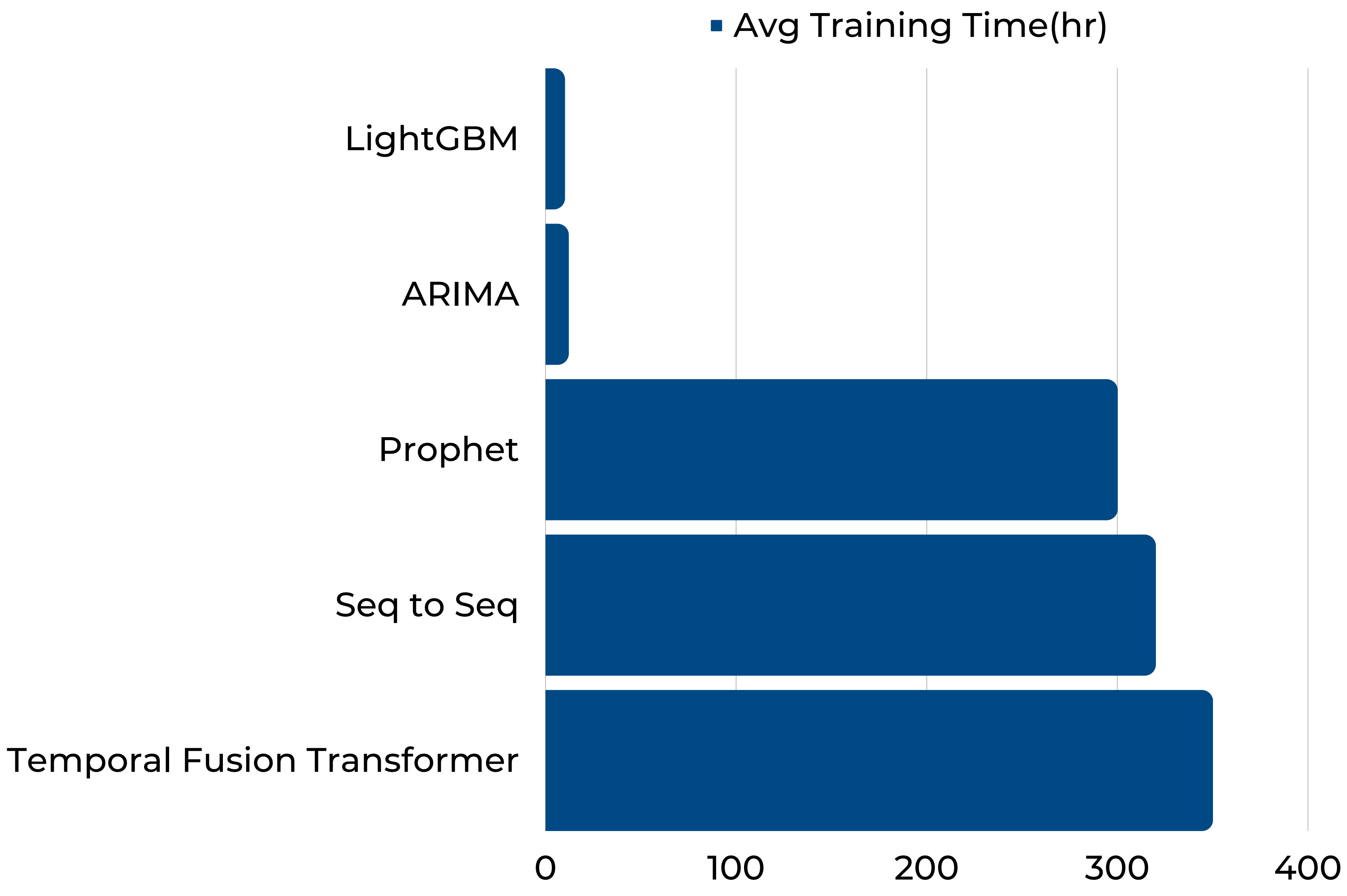
- Gating mechanisms
- Variable selection networks
- Static covariate encoder
- Interpretability

Model Comparison

Appraoches	Model	Pros	Cons
Machine Learning	**LightGBM**	• Faster training speed and higher efficiency • Lower memory usage • Better accuracy than any other boosting algorithm • Compatibility with Large Datasets	• Overfitting • Sensitive to noise data
Statistics	**SARIMAX**	• Brings seasonality as a parameter • Supports exogenous variabls	• Required stationary data • Linear relationship only
	Prophet	• Accommodates seasonality with multiple periods • Resilient to missing values • Fitting the model is fast • Enable hyperparameter tuning	• Dependence on past realizations is completely ignored • Variance is presumed to be constant
Deep Learning	**Seq2seq**	• Much less human engineering effort	• less interpretable, hard to debug • incapability of long sentences memory
	Transformer	• Attention allow decoder to focus on certain parts of the source • Able to capture long-range dependencies and interactions	• High memory usage • Inability to process input sequentially

Performance Comparison





Features

CATEGORICAL

Store Id
Dept Id
Item Id

TEMPORAL

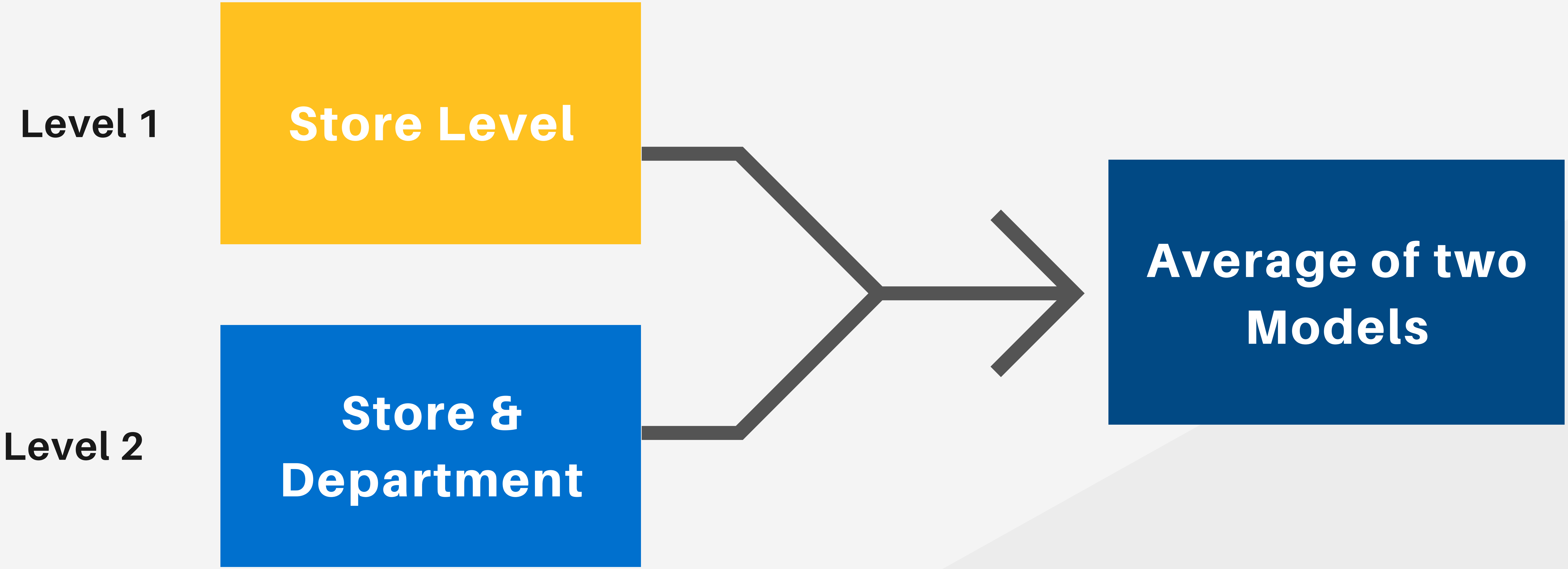
Weekday
week_month
month, day
Events
SNAP
Sell Price

LAG

Past item sales
- Item level
- Dept level
- Store level

Past Sell Price
4wks, 5wks, 6wks

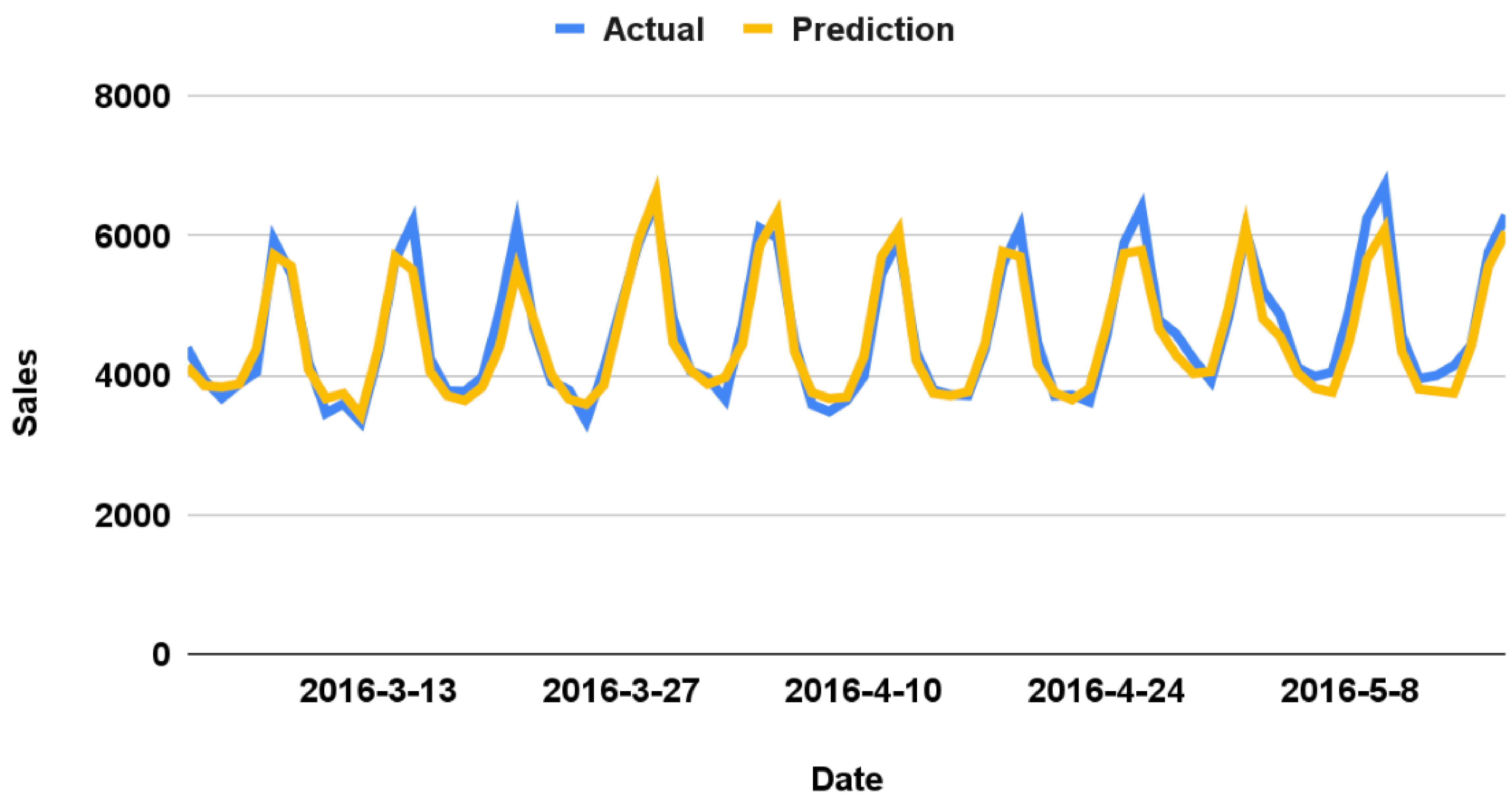
Modeling Process



Model Outcome

Test Metrics: Mean Absolute Percentage Error = 4.97%

On average our predictions have less than **5%** error from the actual sales

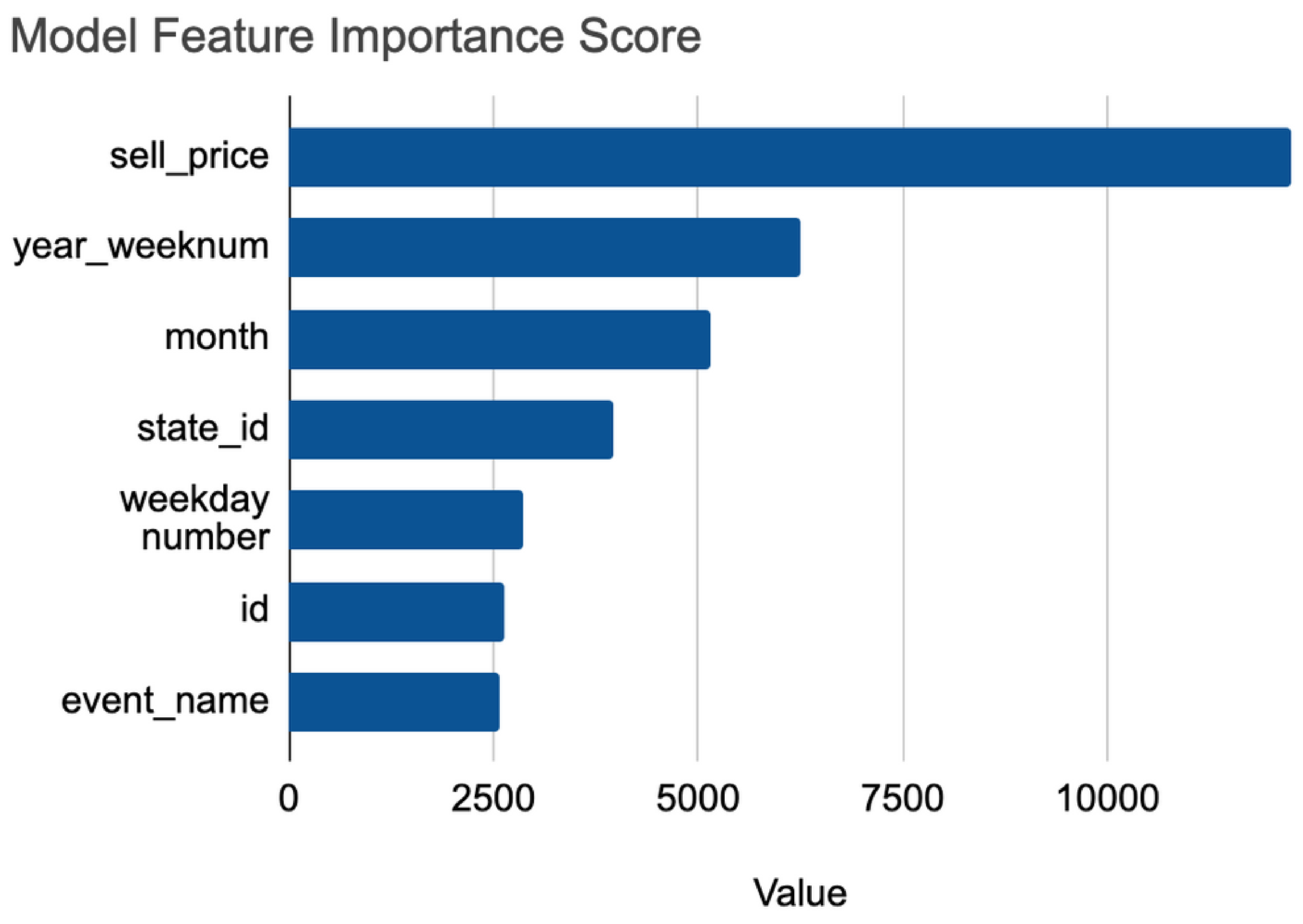


Findings/Conclusions

Sell price is the #1 factor that impacts sales

There exist distinct patterns that follow along the calendar which are easily predictable using the historical data

The demand for the same item can vary by state as states can have different customer bases



THANK YOU