UNIVERSITY OF MALAYA

WQF7023 Artificial Intelligence Research Project 2

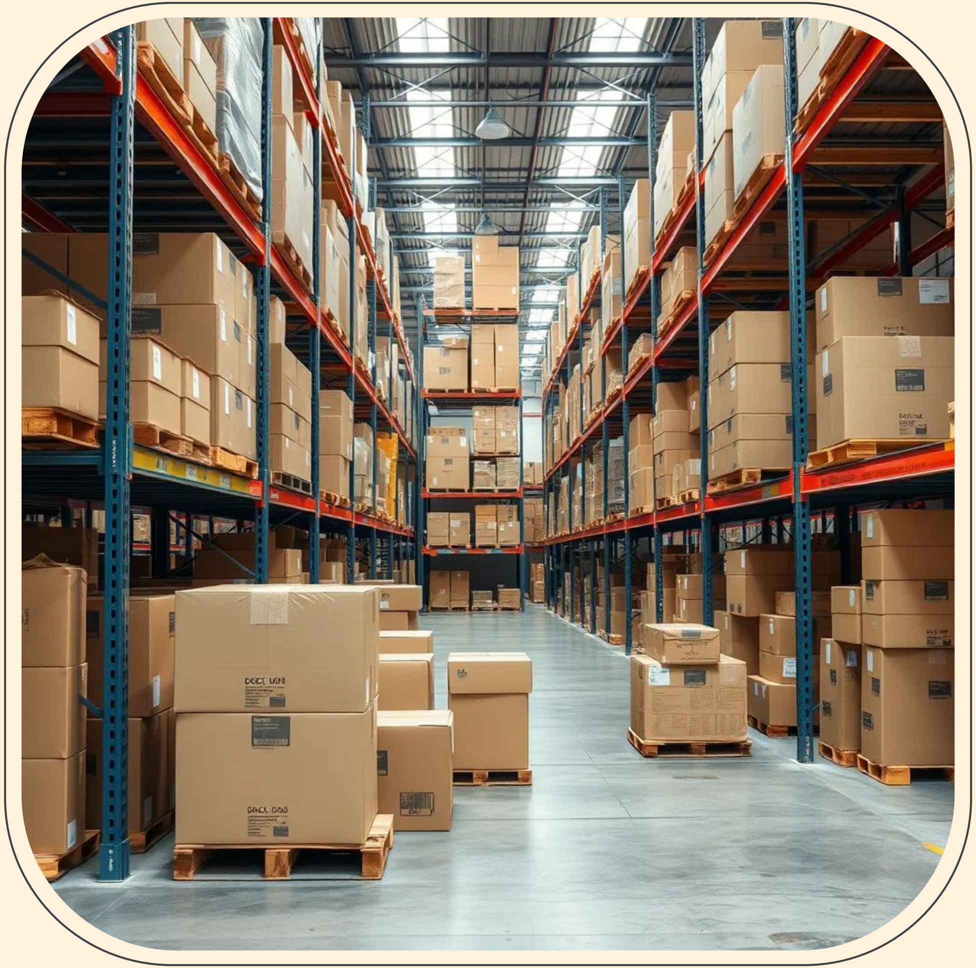
2025/2026 Semester 1

Explainable Demand Forecasting for Retailers using Graph Neural Networks and Temporal Fusion Transformers

Supervisor: ASSOCIATE PROF.TS.DR.Aznul Qalid Bin MD Sabri

Student: 23115588 Maksatbek kyzy Aizhan

Content of the Presentation



Introduction

Demand forecasting is central to retail success, enabling efficient inventory and resource management.

Complex supply chains and logistics increase the need for advanced, AI-driven forecasting models.

This research explores GNN, TFT, hybrid GNN-TFT models, prioritizing both predictive accuracy and model explainability.

Explainable AI builds trust and transparency, ensuring retail stakeholders can understand and act on model insights.

Research Questions and Objectives

1

Which approach yields higher accuracy in retail demand forecasting: GNN, TFT, or a hybrid GNN-TFT model?

To evaluate and compare the forecasting accuracy of GNN, TFT, and hybrid GNN-TFT models.

2

Which features within the retail dataset have the greatest and least impact on the predictive accuracy of these models?

To interpret model predictions by applying both model-specific and model-agnostic explainable AI (XAI) methods.

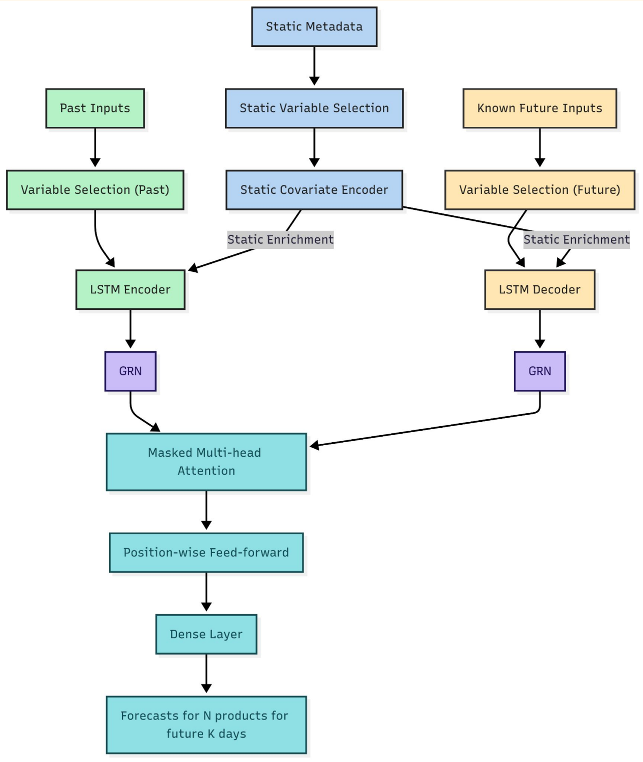
3

What architectural improvements are most effective for optimizing the performances of models?

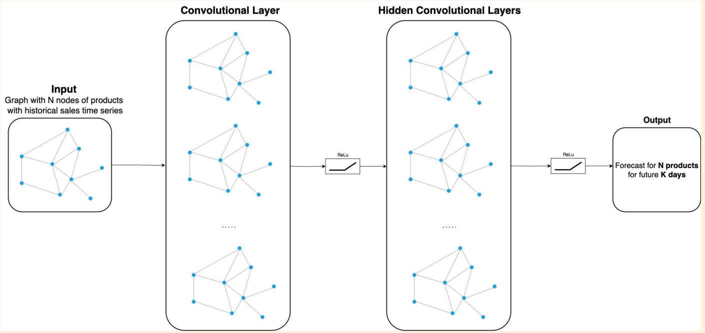
To find and implement architectural improvements, aiming to maximize predictive performance and operational value of models.

Model Architectures

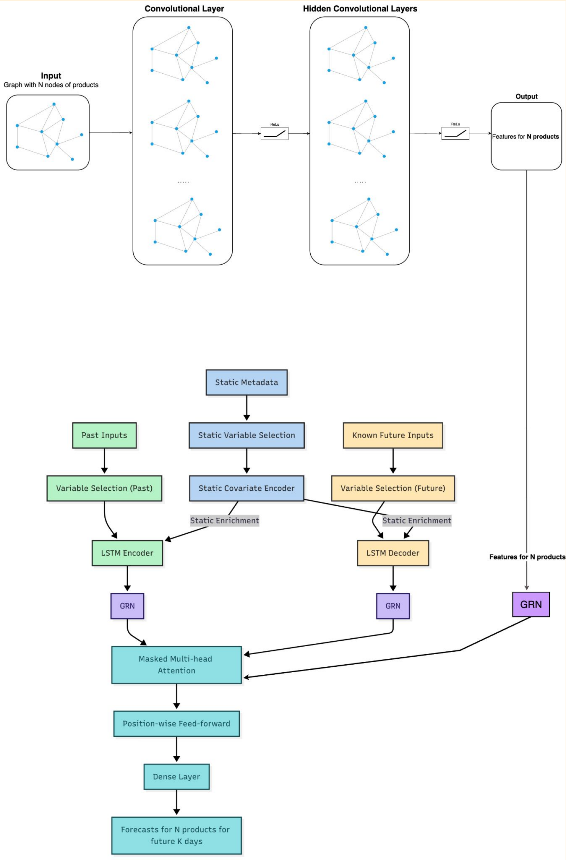
TFT



GNN



Hybrid GNN-TFT



Data - Favorita Grocery

33 - product families

54 - stores

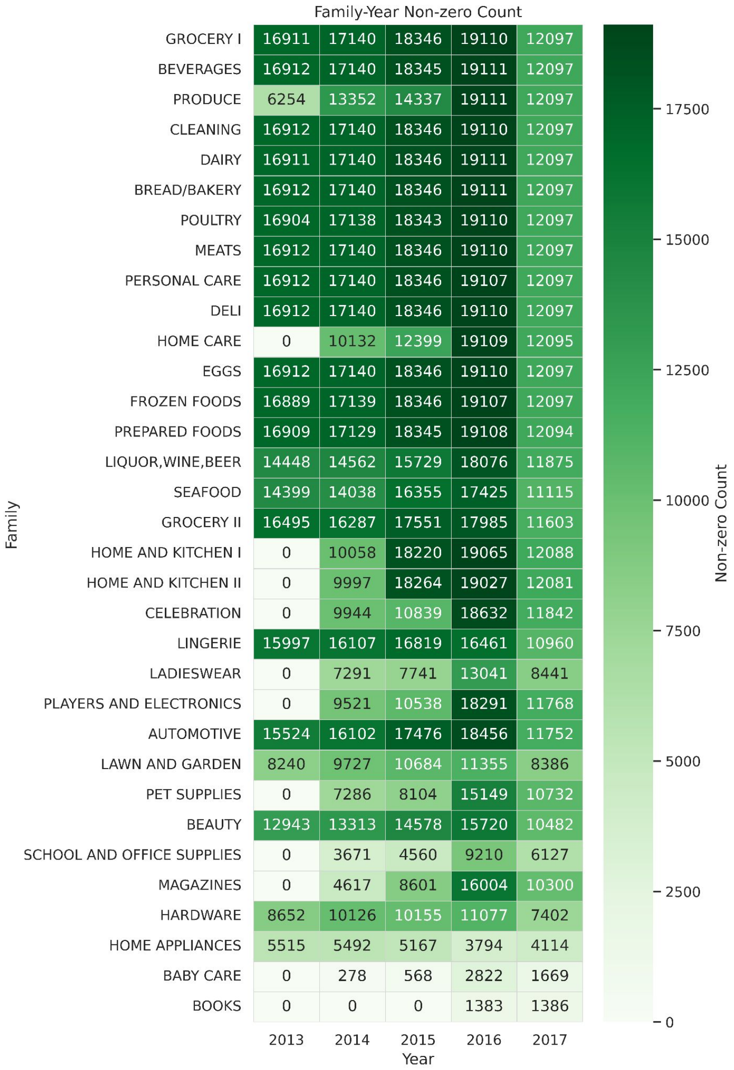
5 - store types

17 - store clusters

22 - states

grocery and household have high daily **non-zero sales** from 2013 to 2016, with a peak in 2016 and **a drop** in 2017

home care, baby care and books either **started later or have much sparser sales**

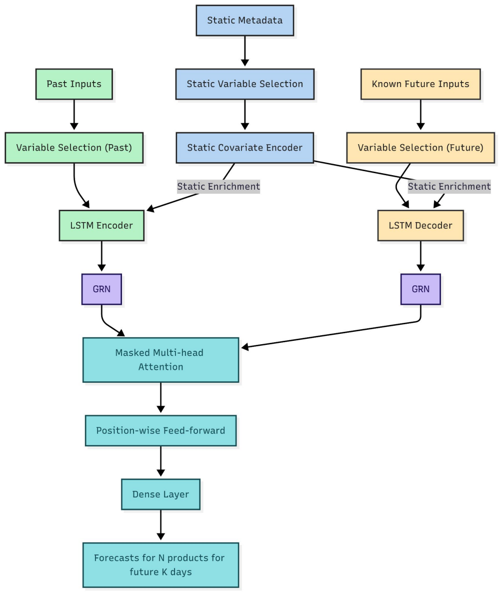


Results

Temporal Fusion Transformers

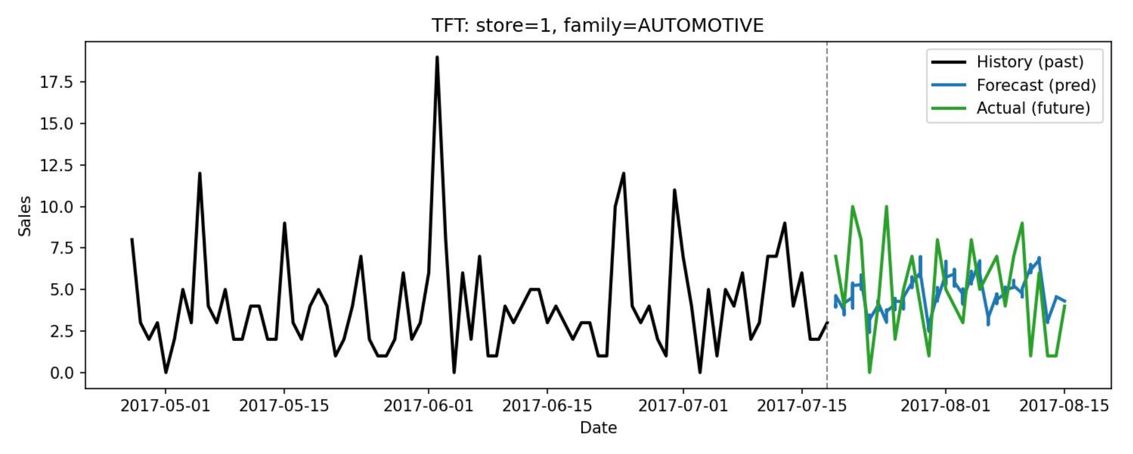
Spatio-Temporal Graph Neural Networks

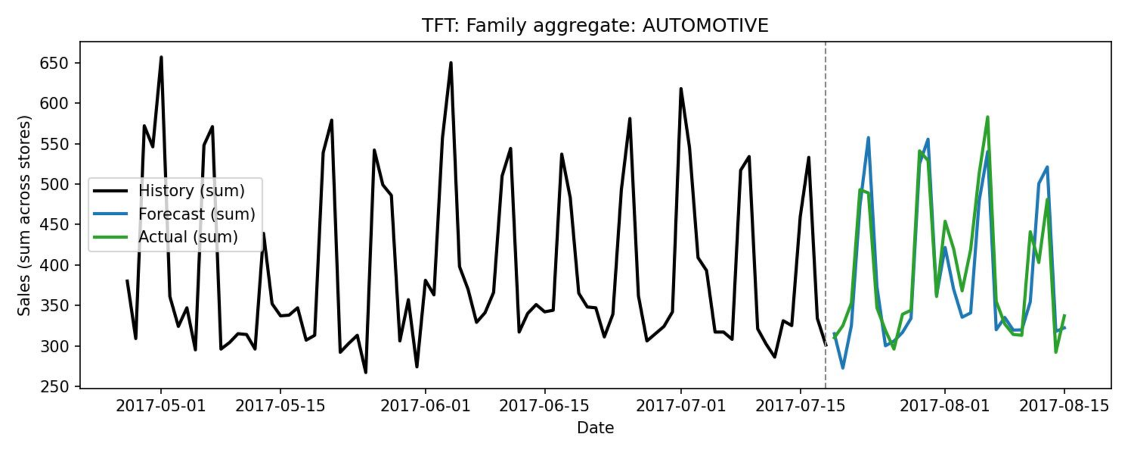
Hybrid GNN-TFT



Temporal Fusion Transformer

Temporal Fusion Transformer





TFT Input Data

Feature Category	Feature Name
Past observed (9 total)	sales transactions daily oil price promotion day of the week month week of the year holiday workday
Known future (6 total)	promotion day of the week month week of the year holiday workday
Static (4 total)	store number product family store state store cluster

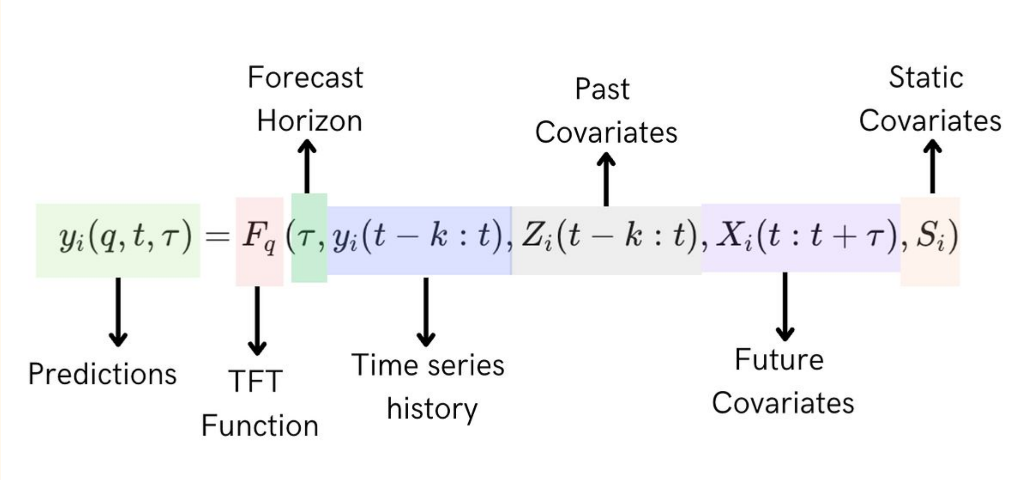


Image Reference: [Medium]

XAI

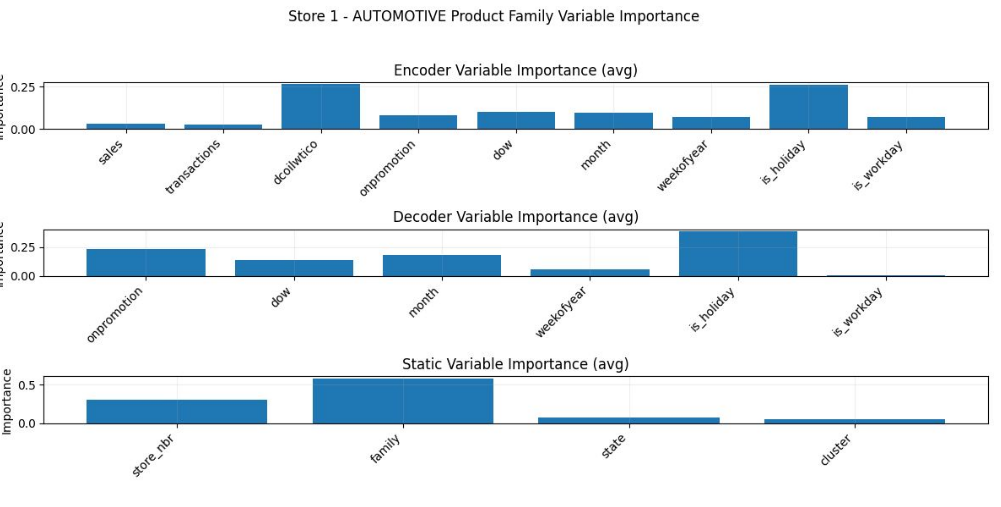
Model-Agnostic Method

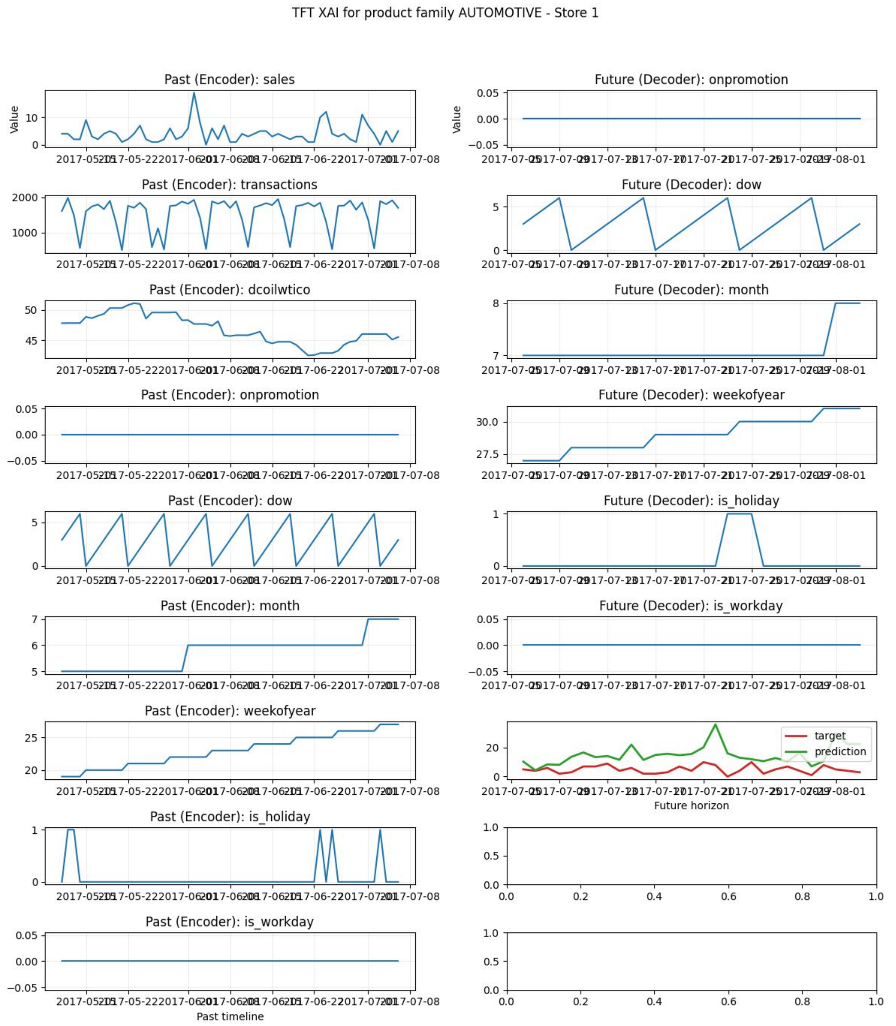
Permutation Importance

Space	Variable	ΔWAPE
Decoder (future)	**onpromotion**	**0.1053**
	dow	0.0216
	month	0.0130
	weekofyear	0.0107
	is_holiday	0.0024
	is_workday	0
Encoder (historical)	sales	0.0280
	transactions	0.0020
	onpromotion	0.0018
	is_holiday	0.0002
	dow	0.0001
	is_workday	0
	weekofyear	-0.0001
	month	-0.0004
	dcoilwtico	-0.0016
Static	state	~0
	cluster	~0
	family	~0
	store_nbr	~0

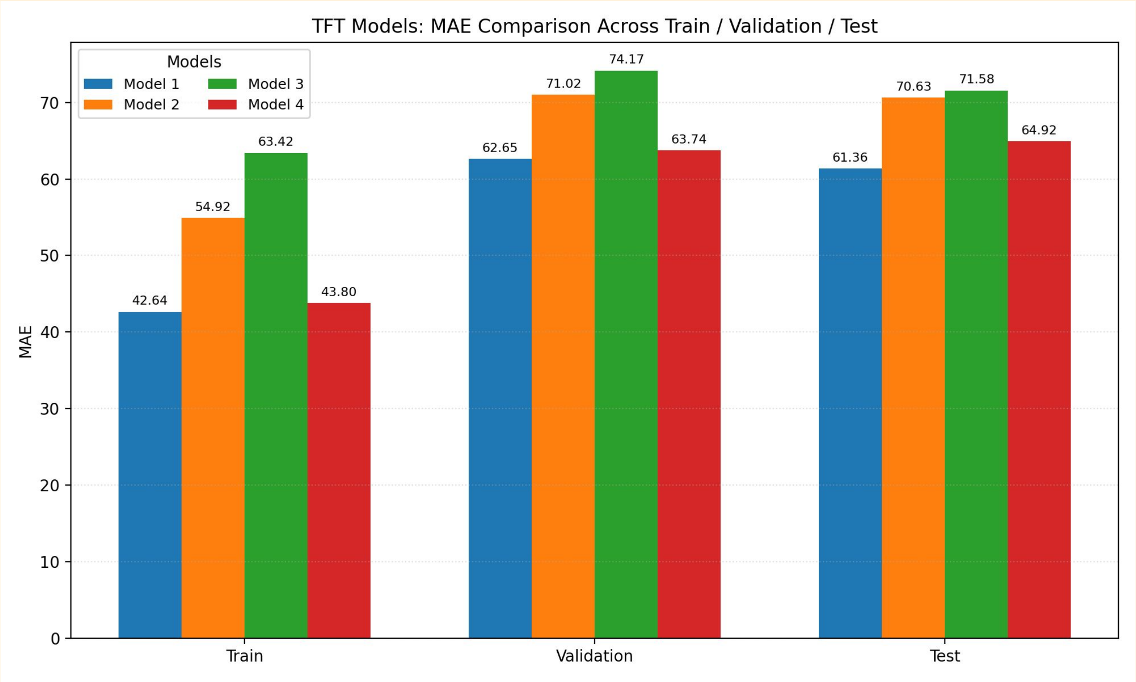
Model-Intrinsic Method

Attention and Variable Selection Weights





Model Performance



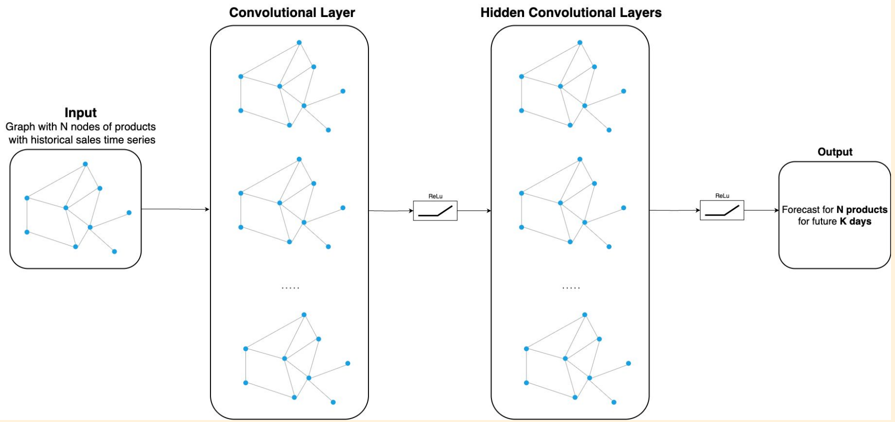
Parameters

--hidden-dim: 128
--d-model: 64
--heads: 4
--lstm-hidden: 64
--lstm-layers: 1
--dropout: 0.1

--hidden-dim: 64
--d-model: 32
--heads: 2
--lstm-hidden: 32
--lstm-layers: 1
--dropout: 0.1

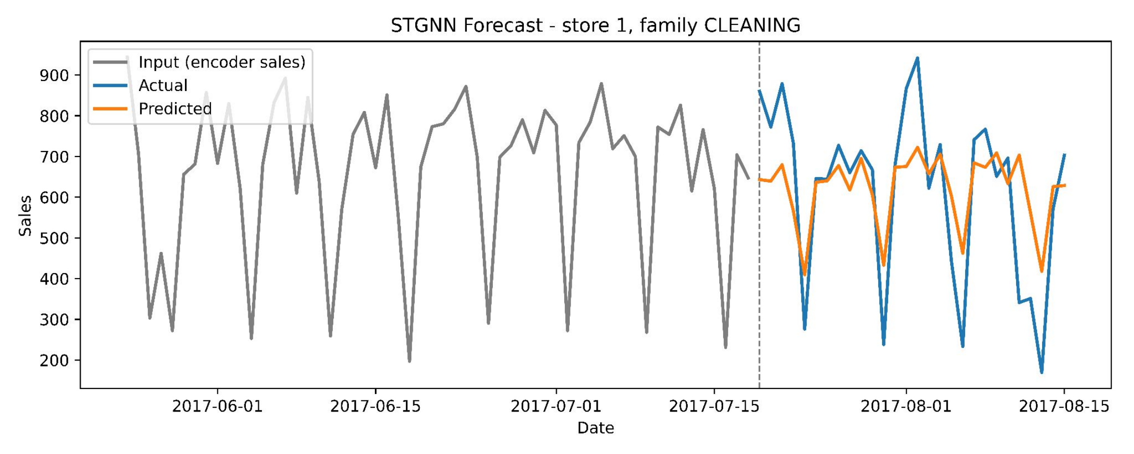
--hidden-dim: 32
--d-model: 16
--heads: 2
--lstm-hidden: 16
--lstm-layers: 1
--dropout: 0.1

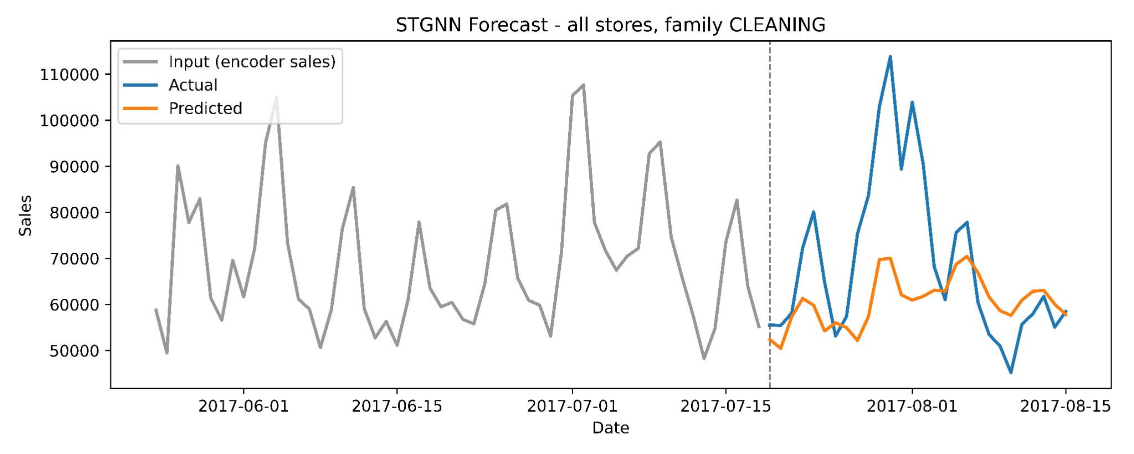
--hidden-dim: 256
--d-model: 128
--heads: 8
--lstm-hidden: 128
--lstm-layers: 1
--dropout: 0.3

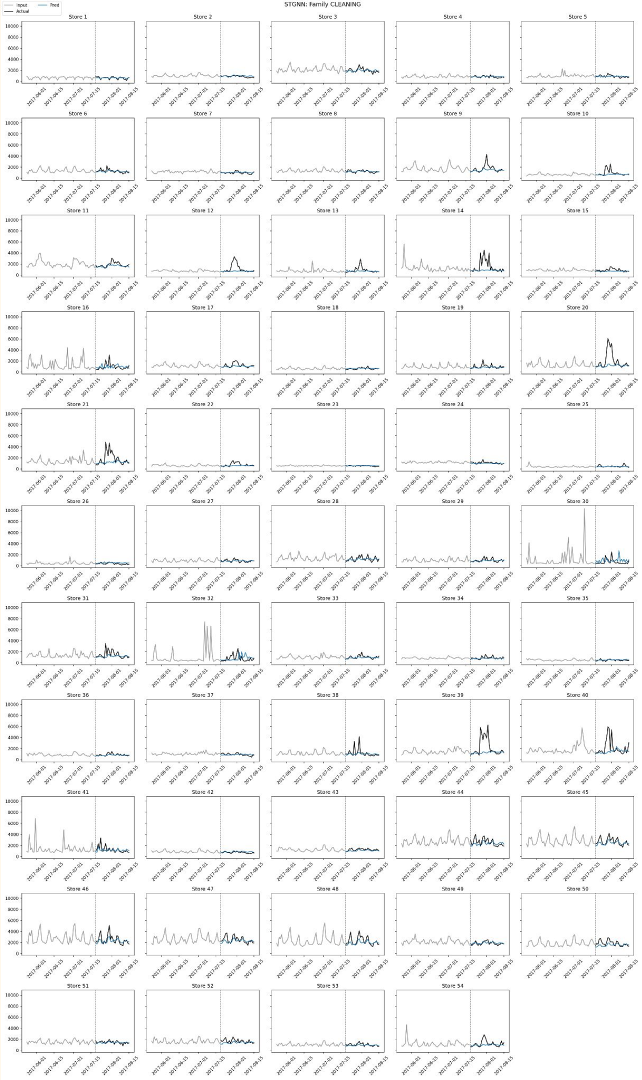


Spatio-Temporal Graph Neural Network (STGNN)

Spatio-Temporal Graph Neural Network







STGNN Data

T - Dates
N - Nodes (distinct pairs of store and family)
F - Feature columns

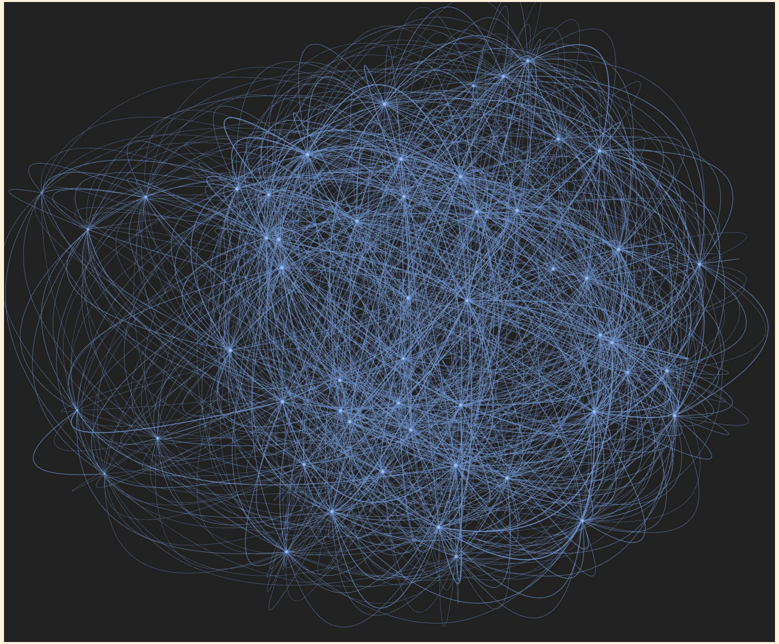


Figure 6.2.3.1 Store Graph illustration

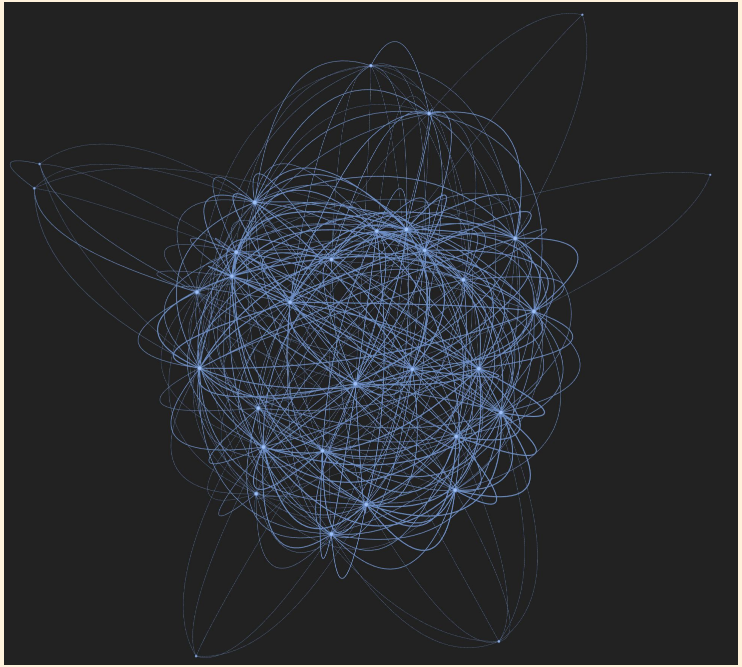


Figure 6.2.4.1 Product-Family Graph

Store Graph

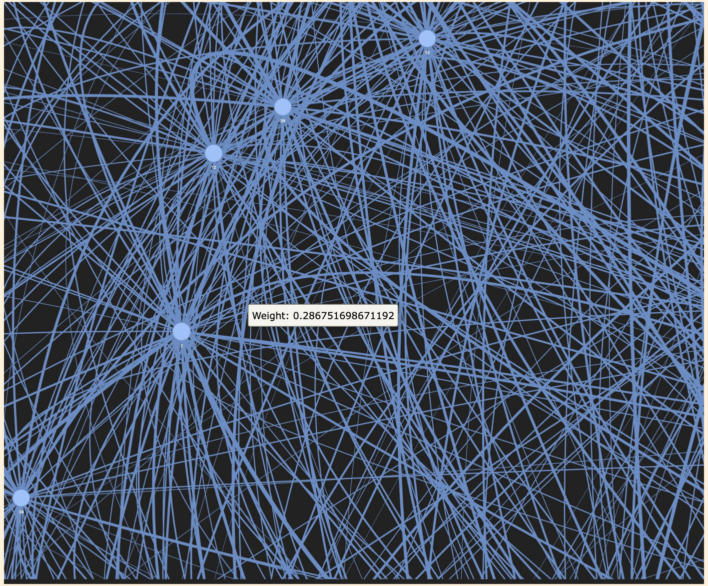
+0.5: if both stores belong to the same cluster

+0.3: if both stores are in the same state

+corr(transactions_a, transactions_b): Pearson correlation of the transaction time series

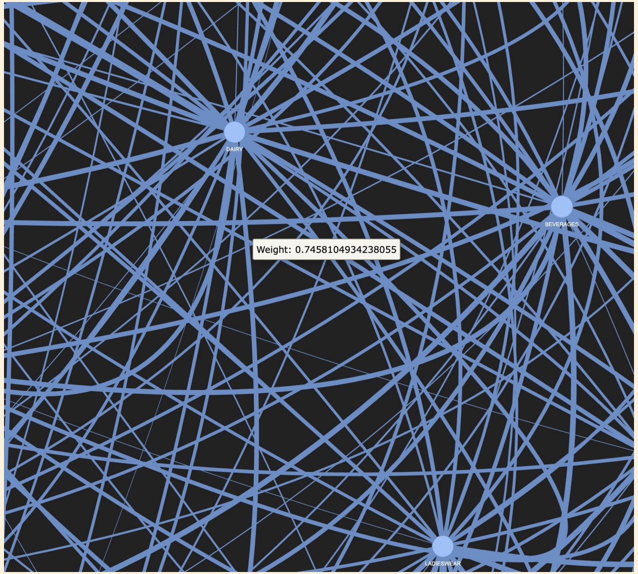
Family Graph

1. Pearson correlation coefficient is computed over a 60-day sales window for products (i) and (j)
2. If the correlation coefficient exceeds 0.3, an undirected edge is established between product (i) and (j), with the correlation value assigned as the weight



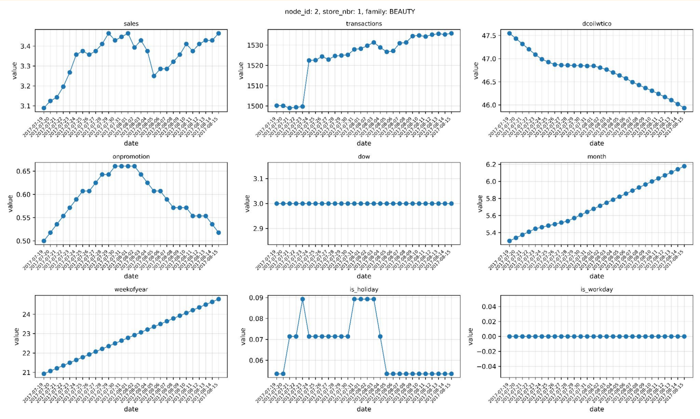
Kronecker-Sum

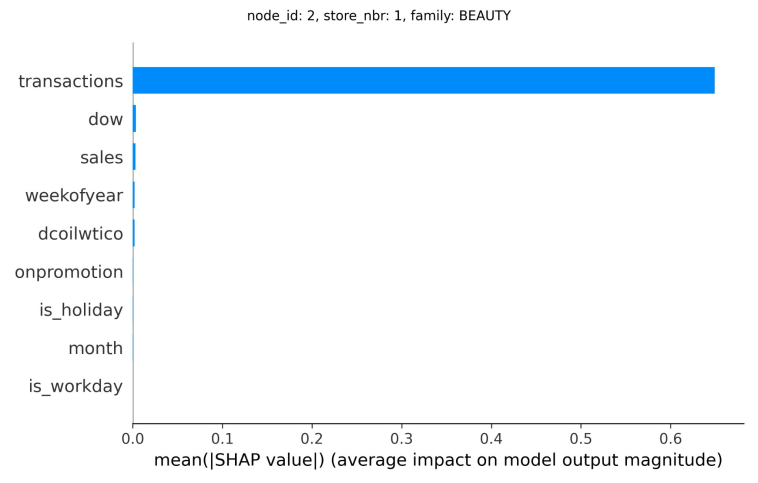
$$A \oplus B = A \oplus I_b + I_a \oplus B$$



XAI

Feature-Level XAI: **SHAP** DeepExplainer



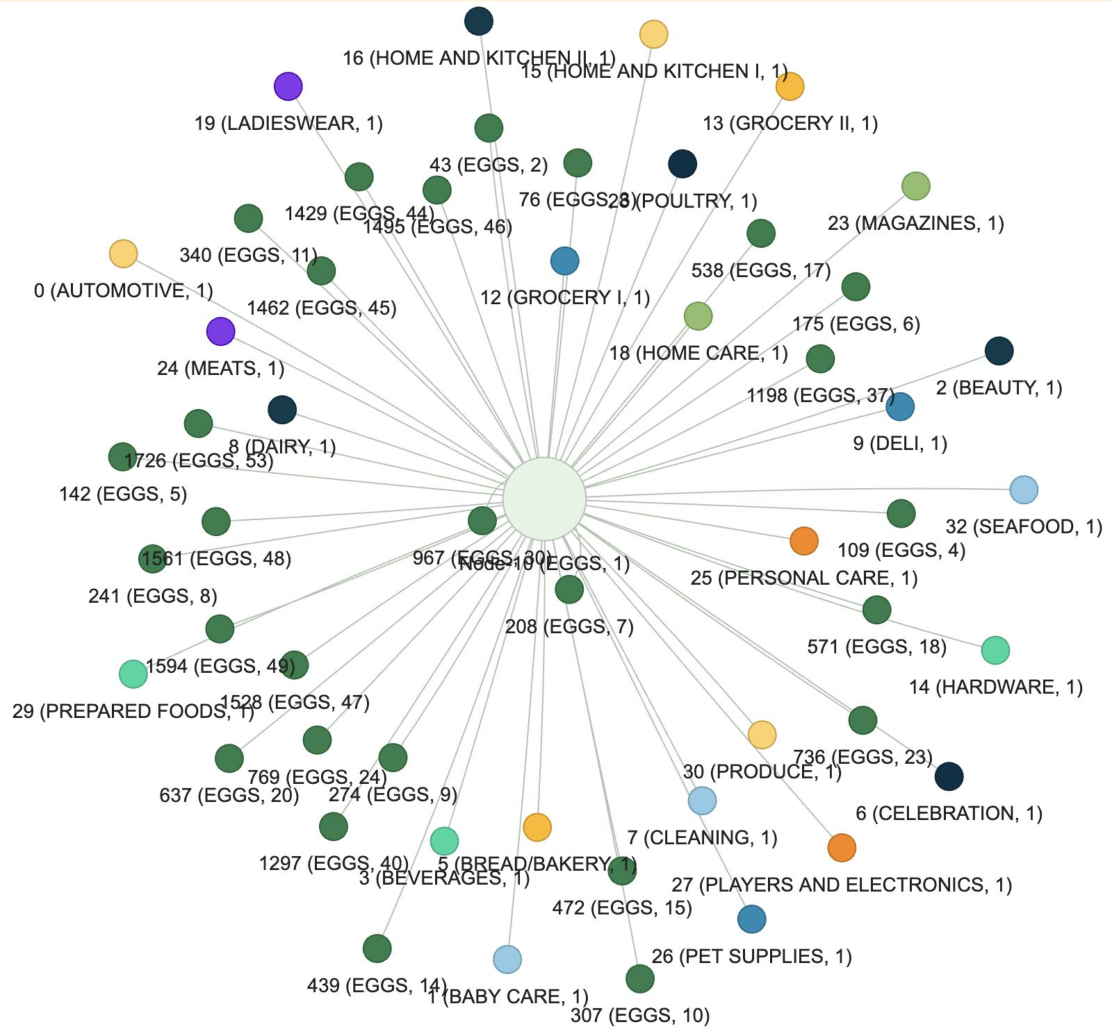


Feature	Mean absolute Shapley value
transactions	**0.6492**
dow	0.0033
sales	0.0031
weekofyear	0.0021
dcoilwtico	0.0020
onpromotion	0.0002
is_holiday	0.0002
month	0.0002
is_workday	0.0

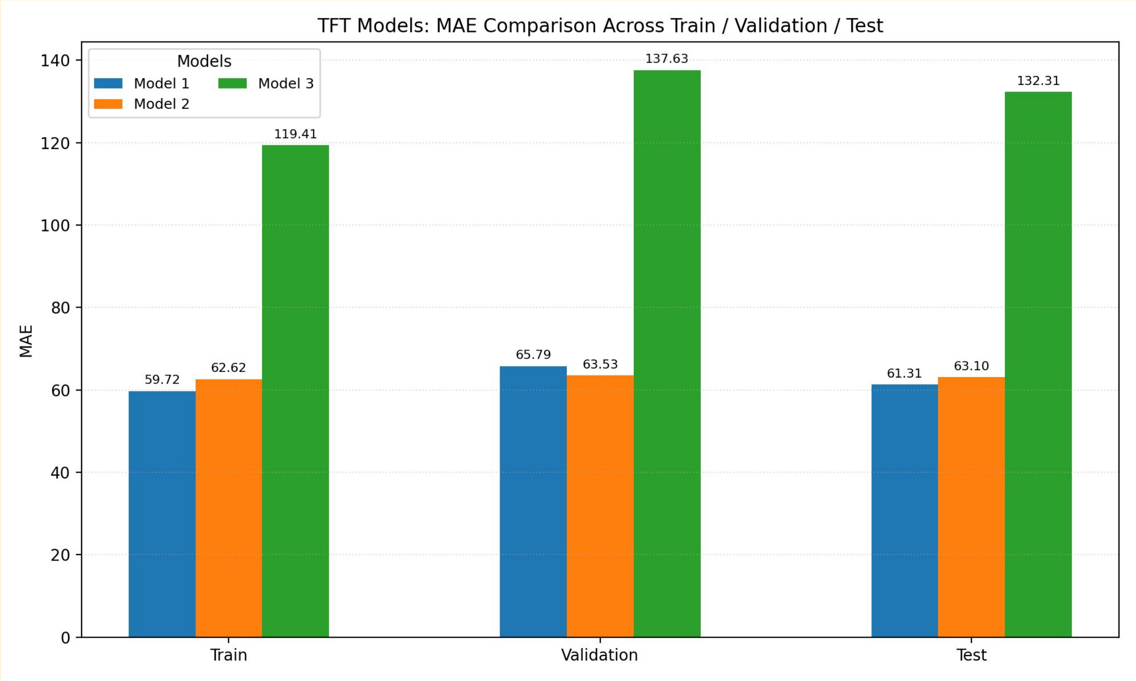
Structure-Level XAI

Custom Edge-Mask Explainer

Decision Influencing Neighbours

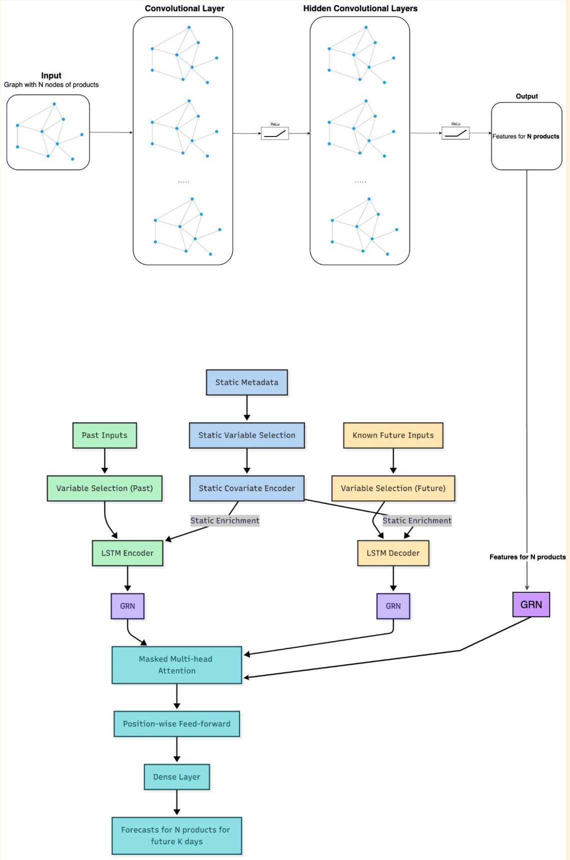


Performance



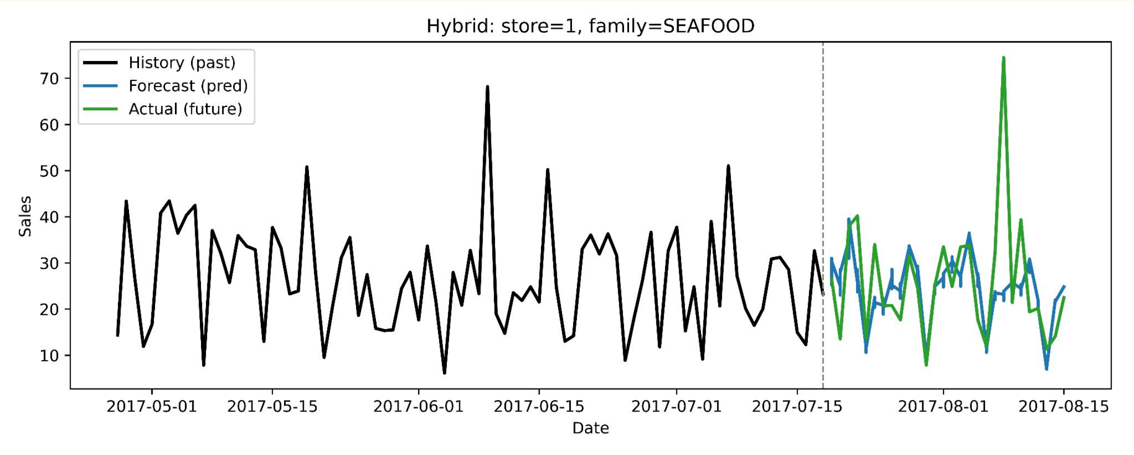
Parameters
--hidden: 64 --blocks: 3 --kernel: 3
--hidden: 32 --blocks: 3 --kernel: 3
--hidden: 16 --blocks: 1 --kernel: 1

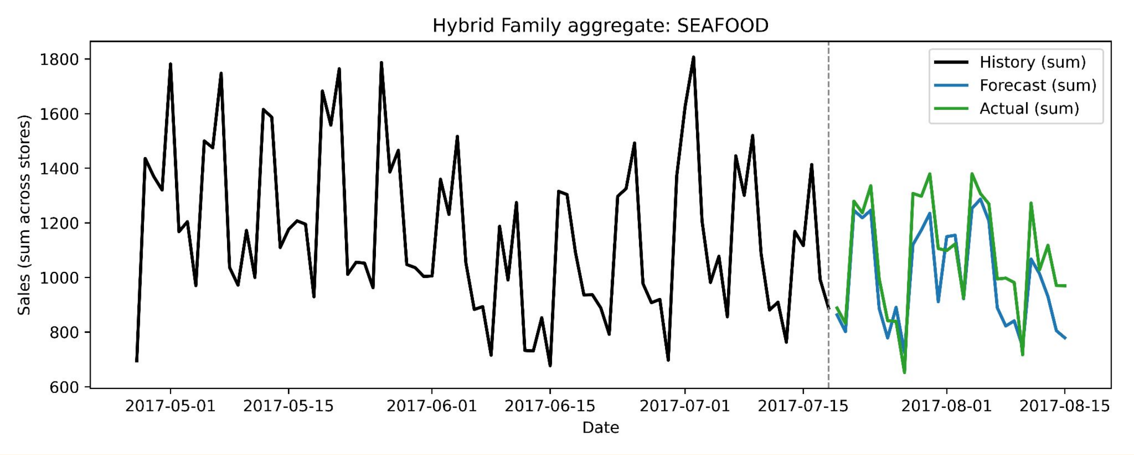
Note from computational standpoint: configuring the model with hidden size of 128 or greater leads to GPU A100 memory allocation errors. Consequently, the model's hidden size is restricted to the maximum feasible model capacity within the available computational resources.



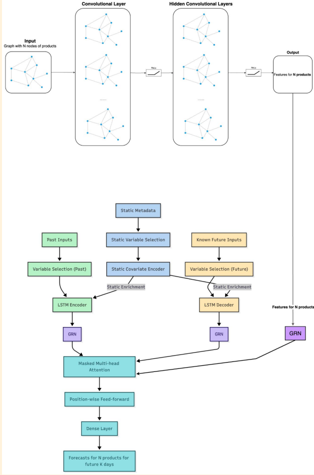
Hybrid GNN-TFT

Hybrid Model





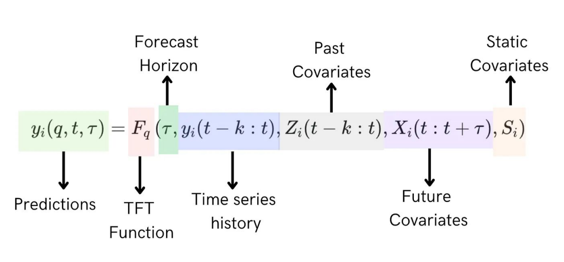
Hybrid Model Data



Exploits both **relational** and **temporal patterns** in the input data

Composed of two primary components: (1) Relational Encoder and (2) TFT

The RE employs a GNN to create node embeddings, thereby capturing the relational structure among product family and store number entities.



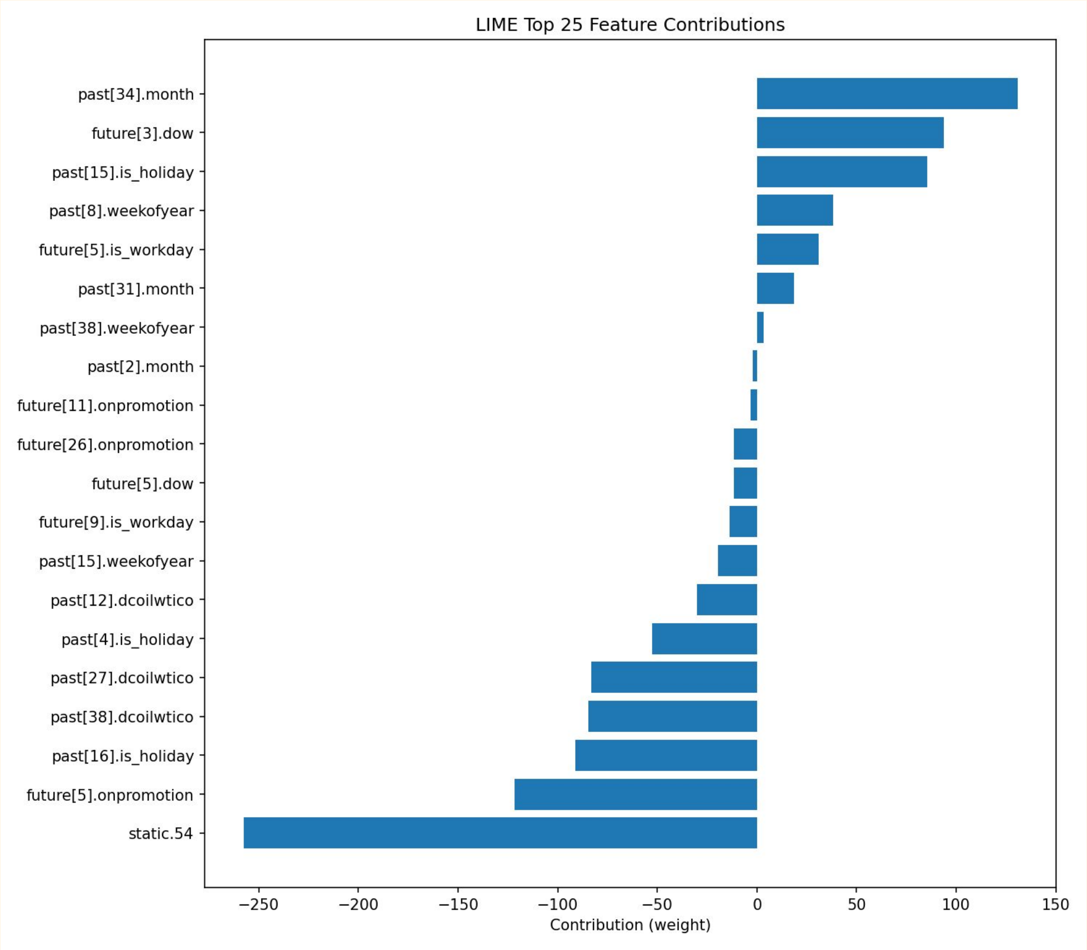
The node embeddings are concatenated with static input features, and then passed to the TFT block static covariate.

Node embeddings creation comprises the Pearson correlation coefficient computation

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

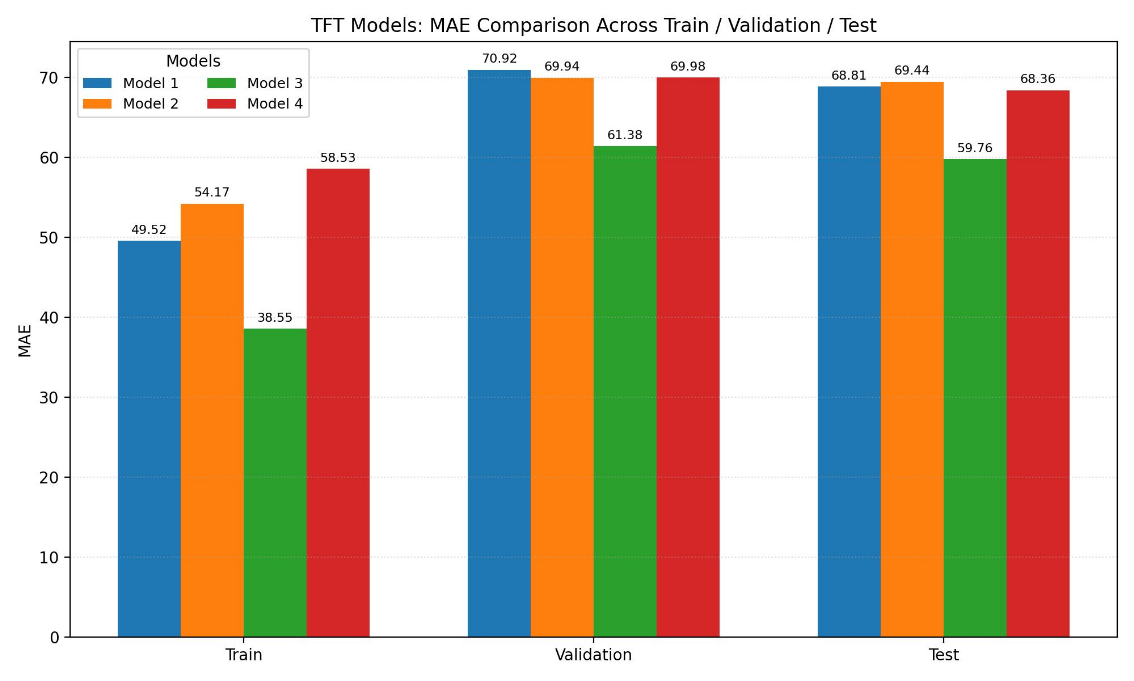
XAI

Hybrid XAI: LIME



Feature	Weight
past[34].month	**130.8142**
future[3].dow	93.7328
past[15].is_holiday	85.4226
past[8].weekofyear	38.0508
future[5].is_workday	31.0180
past[31].month	18.6158
past[38].weekofyear	3.4062
past[2].month	-2.2726
future[11].onpromotion	-3.1617
future[26].onpromotion	-11.6736
future[5].dow	-11.7301
future[9].is_workday	-13.8870
past[15].weekofyear	-19.5332
past[12].dcoilwtico	-30.0725
past[4].is_holiday	-52.7784
past[27].dcoilwtico	-83.1999
past[38].dcoilwtico	-84.6973
past[16].is_holiday	-91.2767
future[5].onpromotion	-121.6794
static.54	**-257.7340**

Hybrid Model Performance



Parameters
--hidden-dim: 128 --d-mode: 64 --heads: 4 --lstm-hidden: 64 --lstm-layers: 1 --gnn-hidden: 64 --gnn-embed: 32
--hidden-dim: 64 --d-mode: 32 --heads: 4 --lstm-hidden: 32 --lstm-layers: 1 --gnn-hidden: 32 --gnn-embed: 16
--hidden-dim: 256 --d-model: 128 --heads: 8 --lstm-hidden: 64 --lstm-layers: 3 --gnn-hidden: 128 --gnn-embed: 64
--hidden-dim: 32 --d-model: 16 --heads: 2 --lstm-hidden: 32 --lstm-layers: 1 --gnn-hidden: 16 --gnn-embed: 8

Summary

The highest performance - Hybrid model with **59.76 MAE**

Difference - (1) 1,6 from TFT (2) 1,55 difference from GNN

The **highly** impacting features - future promotion, holiday, oil price, product family type, transactions, month, day-of-week, holidays

The **least** impacting features - future workday, static metadata like state, cluster, family, store number, and the concatenated static features (node embeddings and static attributes)

Although the performance of the hybrid model was high using concatenated static features, XAI analysis showed that **static features had the greatest negative** impact on the results

Experimental evaluation of key structural parameters demonstrates that tuning hidden layer dimensions, embedding sizes, attention heads, and network depth **can markedly impact the predictive accuracy**

Thank you for your attention!

Q & A