

主要思路

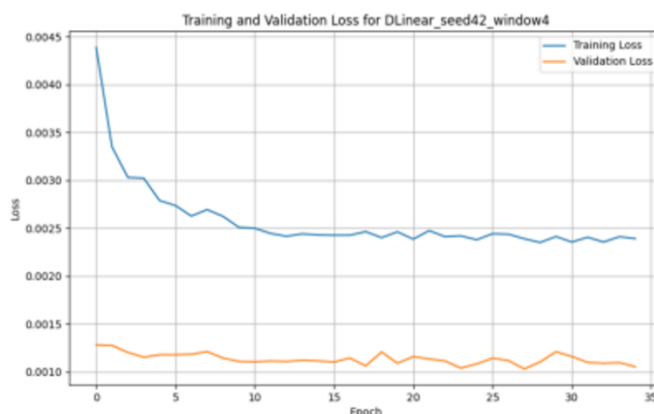
因为是多步预测，首先考虑采用神经网络。从自相关图可以看出变量明显的降低趋势，日频波动，考虑有趋势季节分解的 Dlinear 时序模型，它不仅能很好捕捉趋势项和季节项，在该规模的样本上还有不容易过拟合的优势。除此之外，考虑到强自相关性，可以考虑采用 TCN 或者 GRU 捕捉，但是两者各有优缺点：TCN 可以通过扩张卷积增大感受野，从而捕捉比较长的趋势，但是缺点在于卷积核参数可能需要优化调整；GRU 可以捕捉相关性，但比较难捕捉长期趋势。对于其他模型，如 transformer，这个规模的数据非常容易过拟合。可以考虑由实验确定模型。

在进行神经网络实验之后感觉神经网络未能很好地捕捉目标的强自相关性，考虑直接进行特征工程，加入滞后、滚动等特征后采用 lgbm 模型建模特征间的非线性关系并用 SHAP 进行特征筛选。

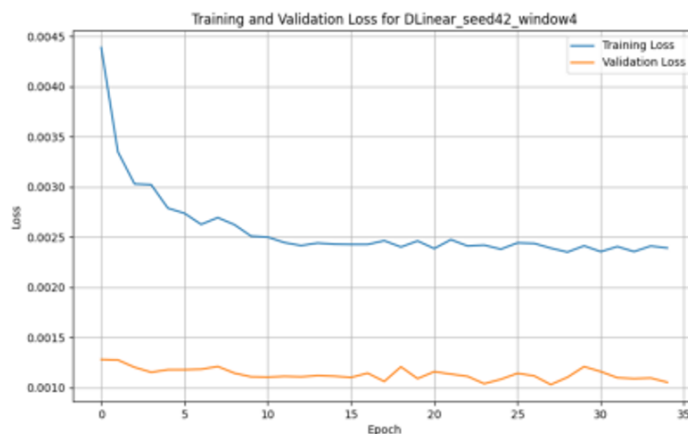
实验过程

完整训练了 Dlinear 和 TCN 网络，TCN 与 Dlinear 的 loss 曲线保存在工作目录下：

TCN 滚动窗口 4 的 loss 曲线

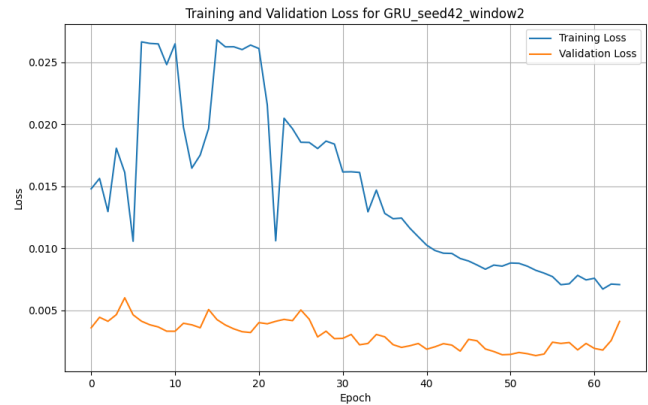


Dlinear 滚动窗口 4 的 loss 曲线



然而，Transformer 和 GRU 都收敛较慢甚至在规定 epoch 内不收敛（或是代码有问题，不过训练太慢没有进一步检查），由于时间不足就不进一步实验：

GRU 滚动窗口 2 的 loss 曲线

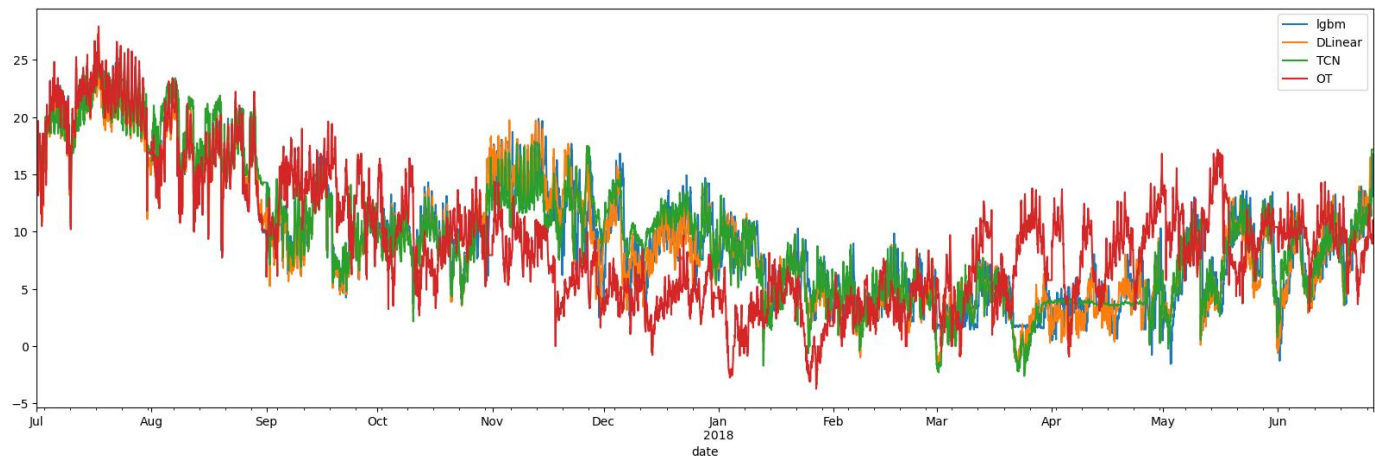


Lgbm 相关实验：在每次时间窗口，进行特征工程后用 SHAP 筛选特征，再用 Optuna 调优，详见代码。

GRU 与 Tranformer 相关模型保存在 log\20250523_195644 中，DLinear 与 TCN 相关模型保存在 log\ 20250523_190216 中。

实验结果

模型预测与真实标签对比（滚动训练+预测后拼接）



MSE 损失表格：

	train_interval	test_interval	DLinear	TCN	lgbm
0	[2016.07.01--2017.07.01)	[2017.07.01--2017.08.01)	1.888094	2.583337	0.203252
1	[2016.07.01--2017.08.01)	[2017.08.01--2017.09.01)	1.886668	8.524174	0.148163
2	[2016.07.01--2017.09.01)	[2017.09.01--2017.10.01)	2.680925	2.902842	0.131616
3	[2016.07.01--2017.10.01)	[2017.10.01--2017.11.01)	2.241481	5.730776	0.124737
4	[2016.07.01--2017.11.01)	[2017.11.01--2017.12.01)	1.781961	2.034744	0.232995
5	[2016.07.01--2017.12.01)	[2017.12.01--2018.01.01)	1.808082	1.619285	0.304125
6	[2016.07.01--2018.01.01)	[2018.01.01--2018.02.01)	2.473879	1.803763	2.474177
7	[2016.07.01--2018.02.01)	[2018.02.01--2018.03.01)	2.431495	2.496953	0.107780
8	[2016.07.01--2018.03.01)	[2018.03.01--2018.04.01)	4.434882	4.295530	0.085149
9	[2016.07.01--2018.04.01)	[2018.04.01--2018.05.01)	4.368138	3.387281	0.108999
10	[2016.07.01--2018.05.01)	[2018.05.01--2018.06.01)	4.433099	5.151728	0.104045
11	[2016.07.01--2018.06.01)	[2018.06.01--2018.07.01)	3.984947	4.380662	0.070747

MAE 损失表格：

	train_interval	test_interval	DLinear	TCN	lgbm
0	[2016.07.01---2017.07.01)	[2017.07.01---2017.08.01)	1.154743	1.244432	0.356254
1	[2016.07.01---2017.08.01)	[2017.08.01---2017.09.01)	0.966476	2.430797	0.299137
2	[2016.07.01---2017.09.01)	[2017.09.01---2017.10.01)	1.193133	1.377660	0.270517
3	[2016.07.01---2017.10.01)	[2017.10.01---2017.11.01)	1.188132	1.943473	0.275752
4	[2016.07.01---2017.11.01)	[2017.11.01---2017.12.01)	0.987486	1.078804	0.367762
5	[2016.07.01---2017.12.01)	[2017.12.01---2018.01.01)	1.037892	0.977215	0.403209
6	[2016.07.01---2018.01.01)	[2018.01.01---2018.02.01)	1.282263	1.059912	0.964470
7	[2016.07.01---2018.02.01)	[2018.02.01---2018.03.01)	1.228448	1.266313	0.251990
8	[2016.07.01---2018.03.01)	[2018.03.01---2018.04.01)	1.603905	1.542884	0.219479
9	[2016.07.01---2018.04.01)	[2018.04.01---2018.05.01)	1.640234	1.426104	0.246963
10	[2016.07.01---2018.05.01)	[2018.05.01---2018.06.01)	1.597528	1.723403	0.240174
11	[2016.07.01---2018.06.01)	[2018.06.01---2018.07.01)	1.616348	1.673739	0.195893

可以看到 lgbm 是显然优于神经网络的