Exploratory Data Analysis

The dataset used in this project includes various environmental, operational, and temporal features collected over time from HVAC system. The primary goal of the analysis is to understand how these features relate to the target variable: **HVAC_Level**, which indicates the intensity or operational load of the HVAC system at a given time.

Overview of Dataset

- **Number of observations**: *444*
- **Number of features**: *36*
- **Target variable**: *HVAC_Level*

	Date	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0	2020-12-01	51.354	4.050	8.748	3.402	225.666	232.956	316.224	133.164	517.590	57.510	43.578	60.912	5.670	145.152	495.072
1	2020-12-02	47.304	4.050	8.424	3.564	228.582	231.498	300.348	126.036	529.254	55.566	41.472	63.990	5.184	142.560	470.610
2	2020-12-03	50.706	3.726	9.072	3.564	375.678	292.410	283.986	181.116	646.380	66.420	47.628	67.716	9.720	144.504	485.190
3	2020-12-04	53.622	4.050	8.910	3.564	411.156	318.168	319.464	187.272	668.250	67.554	48.276	68.526	9.072	135.270	456.840
4	2020-12-05	52.650	4.050	9.396	3.726	382.968	314.118	339.876	190.998	690.120	65.286	45.198	59.454	8.100	133.164	435.942
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
439	2022-03-18	49.086	3.888	9.396	3.402	301.320	297.270	445.824	209.304	618.192	67.878	49.572	77.598	9.234	137.862	649.296
440	2022-03-26	50.382	3.726	9.558	3.402	308.286	303.588	439.020	208.980	702.918	74.682	48.762	76.950	9.234	138.672	644.922
441	2022-03-27	47.466	3.564	9.720	3.726	305.208	291.438	429.138	200.556	718.308	69.174	43.416	74.520	9.882	145.800	627.102
442	2022-03-28	34.668	3.726	6.480	3.402	142.884	144.990	304.074	104.976	458.136	34.344	26.082	61.236	4.698	125.226	469.638
443	2022-03-29	39.852	4.050	6.966	3.402	147.744	147.420	277.992	99.144	453.600	36.126	28.512	62.694	4.860	116.802	476.442

Figure 1: HVAC_Level.csv DataFrame

	A	B	C	D	E	F	G	H	I	J	K
count	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000	444.000000
mean	40.389081	4.047811	9.117243	3.497959	226.588014	222.872595	353.505892	134.480432	508.800405	58.380568	42.760703
std	7.586832	0.370883	1.187097	0.235882	91.024999	61.080753	45.617579	41.826592	107.403550	11.143500	8.001014
min	19.764000	2.430000	5.346000	2.268000	90.882000	104.976000	171.396000	67.716000	302.616000	27.216000	19.116000
25%	35.964000	3.888000	8.586000	3.402000	166.698000	185.611500	333.396000	102.505500	427.963500	54.594000	41.431500
50%	38.232000	4.050000	9.234000	3.564000	187.353000	204.768000	361.503000	117.288000	468.504000	59.859000	45.198000
75%	44.712000	4.212000	9.882000	3.564000	263.938500	256.527000	384.912000	171.882000	601.020000	65.488500	47.790000
max	61.560000	5.184000	11.016000	4.374000	472.554000	363.690000	450.360000	225.342000	764.316000	86.184000	55.080000

Figure 2: Summary of the dataset in the HeartRisk.csv

Top Features

Based on correlation value, the top 10 features will be used in model are: 'A', 'CC', 'H', 'I', 'Z', 'E', 'F', 'M', 'BB', 'HVAC_Level_t-1' (back-shift 1 day).

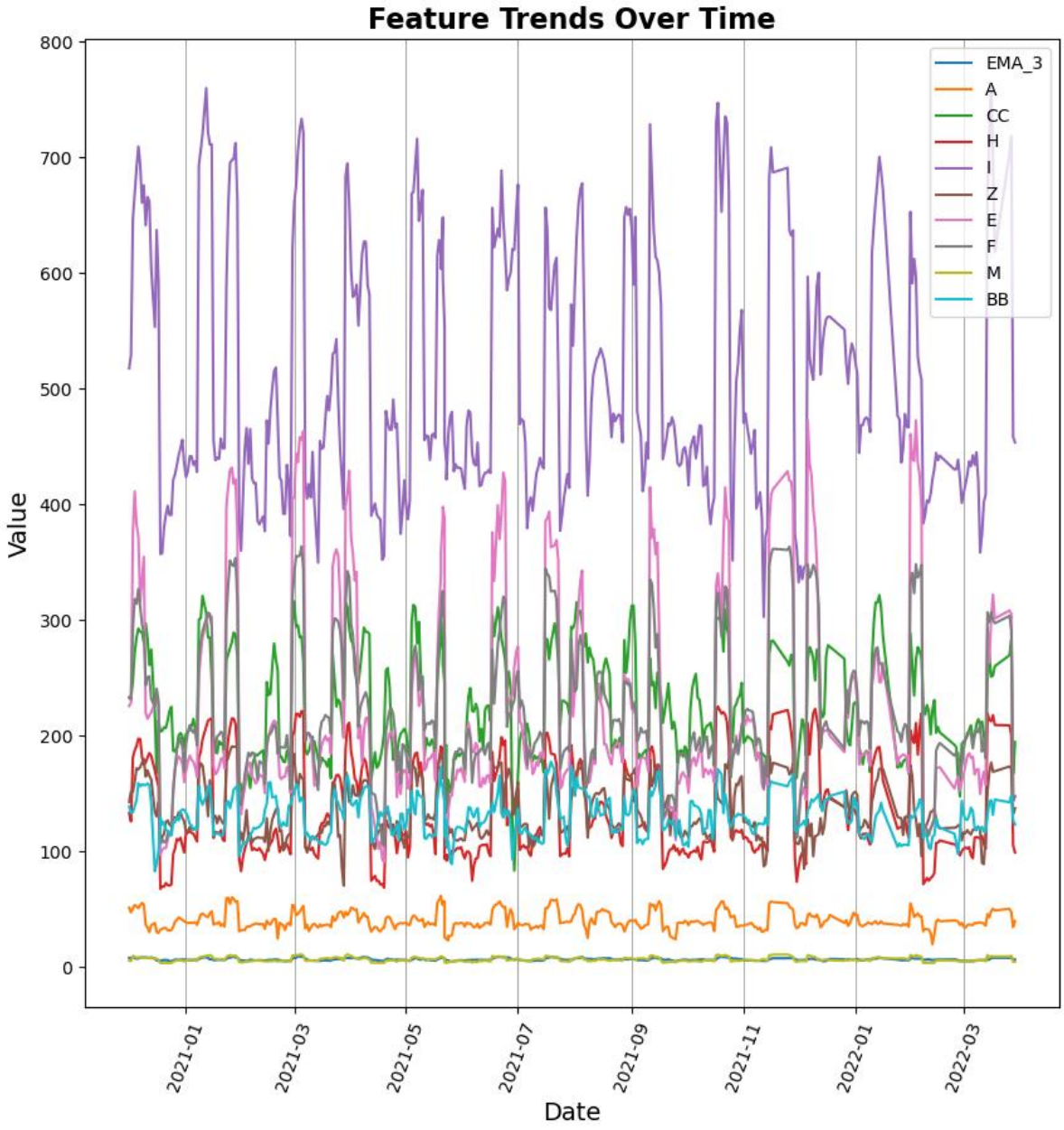

Figure 3: Trends of predictor variables and HVAC Level Over Time

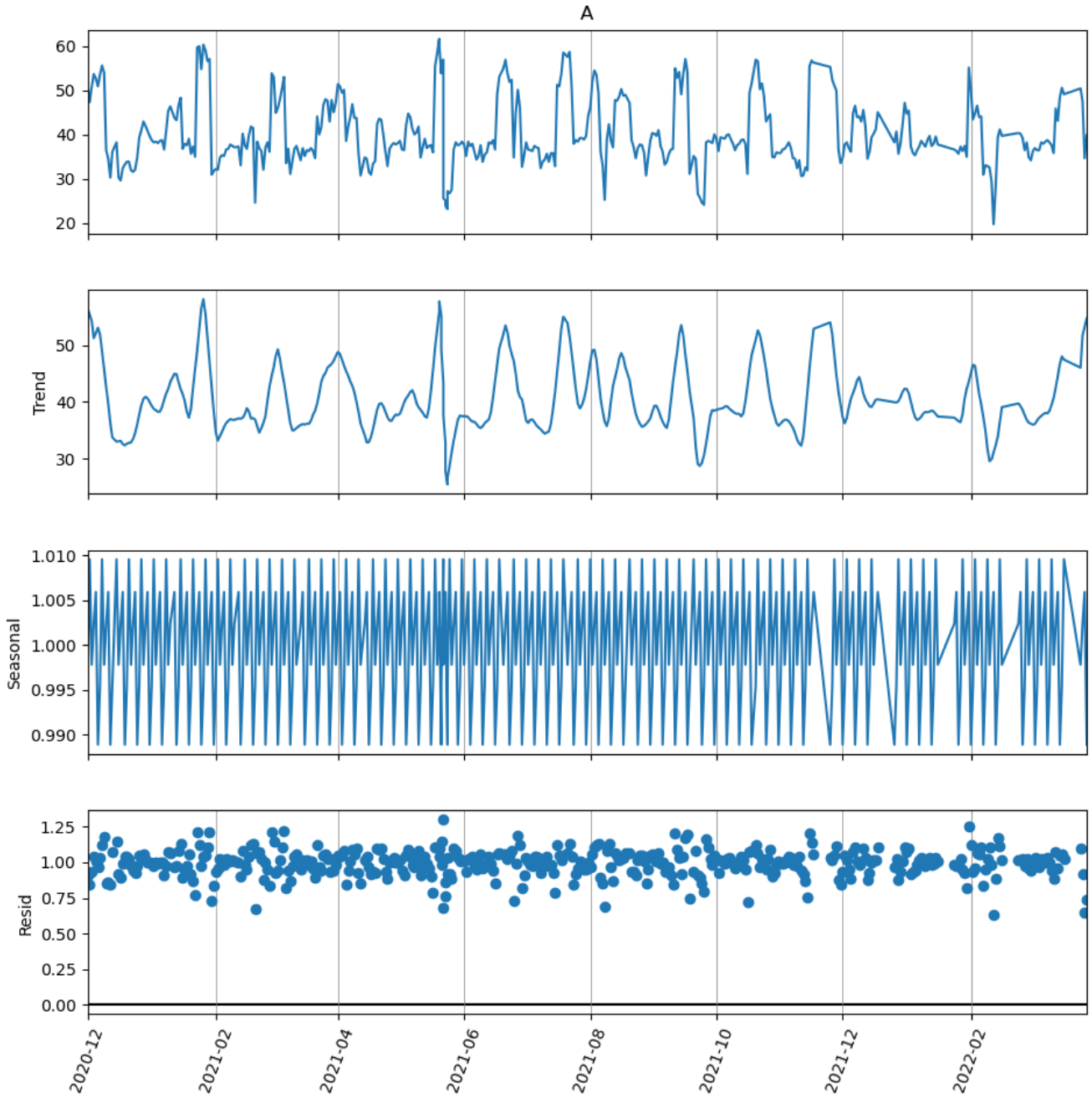

Figure 4: Trends, cycles, seasonal movements and residual of the best predictor variables (A)

Correlation

This section will illustrate the correlation between the target and all predictor variables which are used to predict.

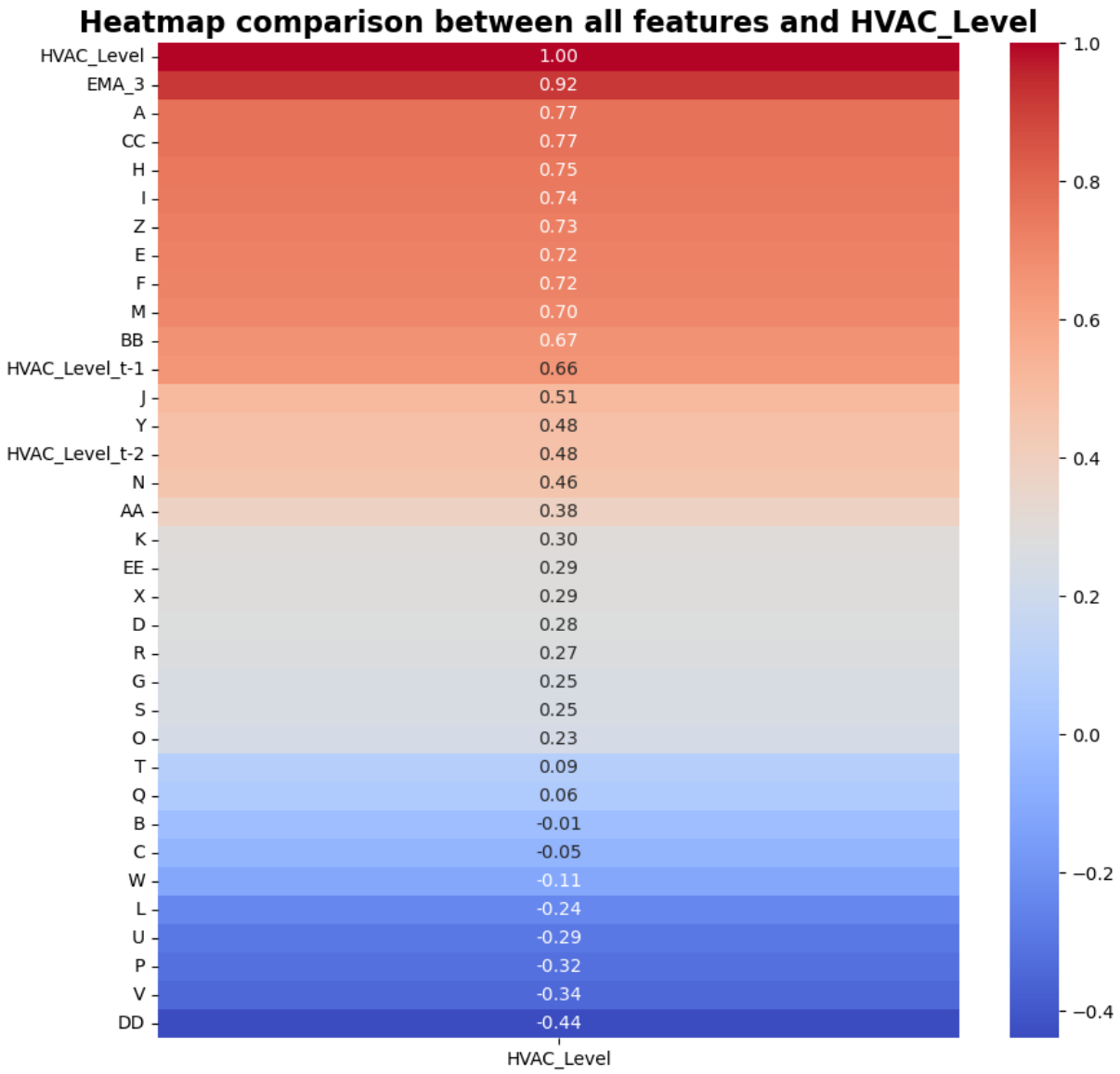

Figure 5: Correlation heat map between all features and HVAC Level

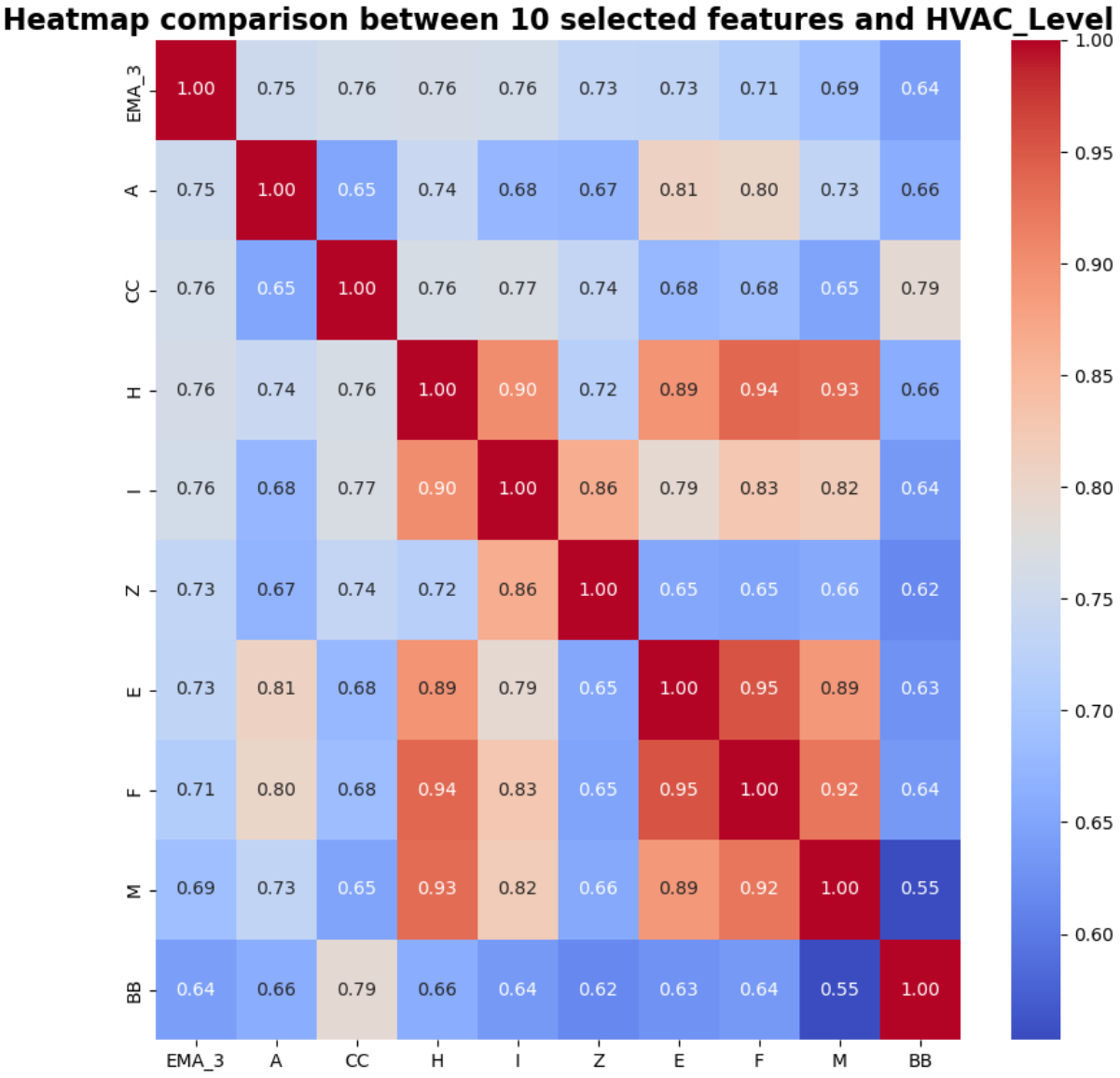

Figure 6: Correlation heat map between predictor variables and HVAC Level

Model Evaluation

Model Type	Backtest 1 RMSE	Backtest 2 RMSE	Selected Features	Number of selected features	Data Treatment
OLS	0.7610	0.7885	['CC' 'A' 'H' 'I' 'Z' 'E' 'F' 'M' 'BB' 'HVAC_Level_t-1']	10	Backshifting, Exponential Moving Average, Feature Selection, Drop Outliers, Scaling Constant Add (For Holt-Winters also have Seasonal Handling, Smoothing)
Holt-Winters	1.2922	0.8472			
ARIMA (1,1,1)	1.1292	0.6636			

The OLS model perform best overall, as both 2 back-testing has relatively low and stable RMSE values (Backtest 1 – 0.7610, and Backtest 2 – 0.7885).

In my hvac_level_predictions.py, I use OLS Model to predict the HVAC Level for the next 6 days from 30/3/2022 to 3/4/2022 (HVAC_Level.csv), then export the final result, and the new predicted values are near the mean of HVAC_Level proving that this OLS Model works efficiently

	Date	HVAC_Level
0	2022-03-30	7.332
1	2022-03-31	7.173
2	2022-04-01	7.835
3	2022-04-02	8.354
4	2022-04-03	8.114
5	2022-04-04	8.108

Figure 7: The predictions of OLS Model for the next 6 days

Code Appendix

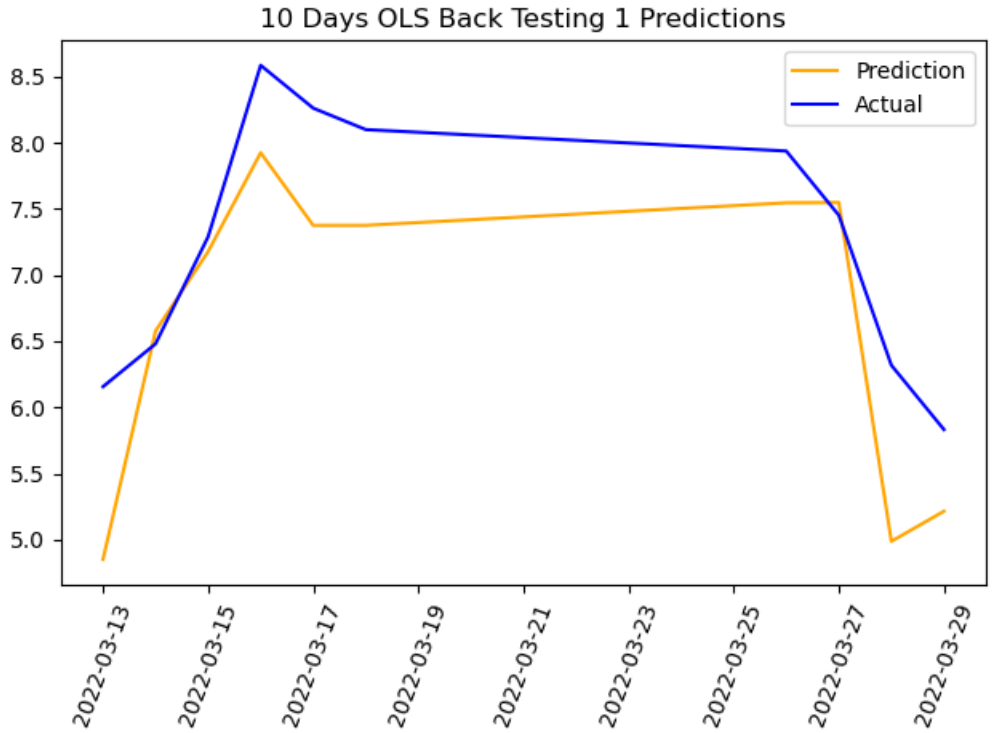

Figure 8: The prediction and actual values of OLS Model for 10 testing days with Backtest 1

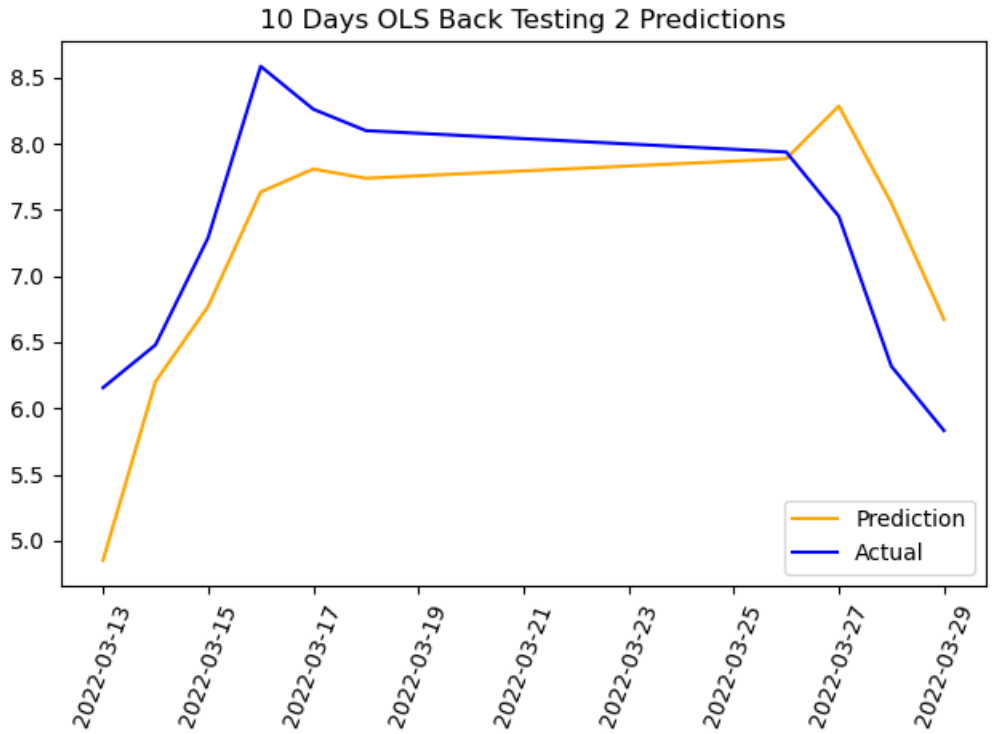

Figure 9: The prediction and actual values of OLS Model for 10 testing days with Backtest 2