DACON HD현대 AI Challenge

항만 內 선박 대기 시간 예측을 위한 선박 항차 데이터 분석 AI 알고리즘 개발

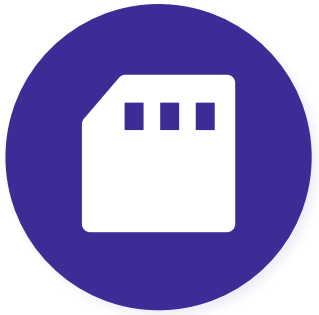
2023.09.25 ~ 2023.10.31

2018015027 김한탁
2018015030 이광진
2018000000 정민권

목차

1. 문제 배경 및 목표

2. 데이터 확인



3. EDA 및 데이터 전처리

4. 모델링 및 추론

5. 결론

6. 또다른 방법론

1. 문제 배경 및 목표

코로나 19 이후 물류정체로 인한 다수의 항만에서 선박 대기 시간이 길어지는 현상 발생

1. 항만 정체로 인한 연쇄적인 물류 지연

2. 연료 사용 증가와 연료비 증가

3. 연료 사용 증가로 인한 온실가스 배출량 증가

*** 선박 대기 시간이란?**

=〉 접안 전에 선박이 해상에 정박하는 시간을 의미

해결 방안 및 목표

선박의 제원 및 운항 정보를 활용해 산출된 항차 데이터 활용

선박의 접안 시간 예측 가능

=〉 문제점 해결 가능

2. 데이터 확인

*** 총 2개의 데이터 셋**

1. 학습 데이터(train.csv) – 21개의 독립변수와 1개의 종속변수('CI_HOUR)로 구성
2. 시험 데이터(test.csv) – 21개의 독립변수로만 구성

*** train 데이터**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
SAMPLE_ID	ARI_CO	ARI_PO	SHIP_TYPE_CATEGORY	DIST	ATA	ID	BREADTH	BUILT	DEADWEI	DEPTH	DRAUGHT	GT	LENGTH	SHIPMAN	FLAG	U_WIND	V_WIND	AIR_TEMP	BN	ATA_LT	PORT_SIZE	CI_HOUR
TRAIN_000000	SG	GIW5	Container	30.88101761	2018-12-17 21:29	Z618338	30	24	24300	10	10	16700	180	CQSB78	Panama					5	0.002615416	3.45
TRAIN_000001	IN	UJM2	Bulk	0	2014-09-23 6:59	X886125	30	13	35900	10	10	23500	180	SPNO34	Marshall Islands					12	0.000216617	0
TRAIN_000002	CN	EUC8	Container	0	2015-02-03 22:00	T674582	50	12	146000	30	20	140000	370	FNPK22	Malta					6	0.001614168	0
TRAIN_000003	JP	ZAG4	Container	0	2020-01-17 4:02	Y847238	20	18	6910	10	10	5400	120	PBZV77	Bahamas	-3.18	-1.61	6.7	2.629349642	13	0.000356118	0
TRAIN_000004	SG	GIW5	Container	27.0376502	2020-01-26 7:51	A872328	50	10	116000	20	10	96600	300	GUCE76	Liberia	-0.33	-3.28	25.6	2.495952793	15	0.002615416	253.5544444
TRAIN_000005	AU	WHH4	Bulk	49.95358543	2021-03-05 18:36	S731836	40	7	183000	20	20	94100	290	HZUO14	Japan	6.1	-2.84	28.1	4.01621746	5	0.000103275	68.39138889
TRAIN_000006	ID	REJ1	Container	42.27628053	2016-12-11 3:00	A735263	20	30	6800	10	10	4810	110	HCOS27	Indonesia					10	4.11E-05	31.70055556
TRAIN_000007	TW	JWI3	Cargo	0	2022-10-07 10:06	N531887	30	18	46600	20	10	29800	200	EIPX61	Hong Kon	2.82	0.25	28.6	2.255079469	18	0.000989649	0
TRAIN_000008	JP	HYG5	Cargo	0	2015-12-10 2:03	L124642	30	12	37200	20	10	22900	180	UOPG57	Panama					11	0.000256045	0
TRAIN_000009	CN	NGG6	Container	101.5215977	2018-11-30 19:29	S458225	50	7	124000	30	20	111000	320	YLMR26	United Kingdom					3	0.001742858	58.19305556
TRAIN_000010	CN	UVK6	Bulk	18.02249534	2021-09-23 2:48	U334123	30	11	35000	20	10	22500	180	KLJL33	Monteneg	-2.16	1.5	25.6	2.146870114	10	0.00051987	2.721944444
TRAIN_000011	SG	GIW5	Container	26.82697615	2017-09-02 1:47	J412562	30	23	30700	20	10	26000	200	ENOX58	Liberia					9	0.002615416	271.9302778
TRAIN_000012	CN	EUC8	Container	100.7199348	2016-05-30 23:13	T684332	40	15	80200	20	10	76900	300	CVDP86	Japan					7	0.001614168	68.03222222
TRAIN_000013	CN	TDA5	Bulk	21.88730703	2021-05-12 21:35	R763644	40	9	70400	20	10	54600	220	HULF38	Hong Kon	-4.88	1.55	21.9	3.347518051	5	0.000455438	7.374722222
TRAIN_000014	QA	KIU2	Bulk	5.06146197	2015-08-27 7:09	D575432	30	11	56800	20	10	33000	190	IOSS33	Hong Kong, China					10	0.000130486	105.8202778
TRAIN_000015	CN	QQW1	Cargo	0	2021-04-16 15:52	L286286	20	17	13600	10	10	9090	130	KICF32	Panama	-0.52	4.15	10.5	2.925160854	23	0.000594691	0
TRAIN_000016	TW	JWI3	Bulk	0	2016-06-30 1:00	R748821	30	10	81700	20	10	45000	230	MBAX66	Liberia					9	0.000989649	0
TRAIN_000017	CN	NGG6	Container	109.021105	2017-07-10 14:27	S764487	30	9	25300	20	10	25100	180	VNII45	Korea, South					22	0.001742858	15.03972222
TRAIN_000018	AU	NQO4	Cargo	29.62476772	2020-11-17 22:41	Q418441	30	27	55800	20	10	36000	200	GXKR25	Bahamas	-1.25	1.09	20.3	1.578840632	9	0.000110384	18.00722222
TRAIN_000019	CN	NGG6	Container	96.46359378	2017-07-17 22:45	U753134	30	14	34200	20	10	26400	210	UOQB77	Hong Kong, China					6	0.001742858	42.74194444
TRAIN_000020	TT	IVU2	Container	7.796448889	2017-03-02 8:31	K848657	20	18	13700	10	10	9960	150	EYBR24	Panama					4	4.45E-05	11.61416667
TRAIN_000021	CN	QQW1	Container	0	2016-03-16 4:53	H341333	50	10	115000	30	20	110000	330	WOLG37	Singapore					12	0.000594691	0
TRAIN_000022	AU	YDP4	Bulk	11.36544169	2016-10-08 6:16	C866378	40	17	99300	20	10	55300	250	IBFB81	Indonesia					17	2.63E-05	213.3669444
TRAIN_000023	JP	VYJ1	Bulk	9.80218885	2019-06-09 23:32	X345112	20	16	5800	10	10	3500	110	RCGL58	Japan	0	0	19	0	8	0.000264287	1.752777778
TRAIN_000024	CN	JEN5	Bulk	16.61221048	2022-05-23 22:59	D626134	30	17	30400	10	10	17900	170	UREU81	Liberia	-2.36	5.42	21.5	3.684073668	6	0.00102563	13.82777778
TRAIN_000025	CN	NGG6	Container	99.19756524	2015-08-20 5:58	M766643	20	17	13800	10	10	9950	150	HQRJ86	Panama					13	0.001742858	193.9163889
TRAIN_000026	RU	FCD5	Bulk	0	2022-01-12 9:50	I512572	10	19	3280	10	10	1980	80	AOQV51	Barbados	0.28	-11.42	-11.1	5.715578805	11	0.000398636	0
TRAIN_000027	CN	NGG6	Container	97.79634477	2020-09-25 0:38	Q641134	20	8	12400	10	10	9860	140	QJKV12	Panama	-1.05	-2.26	20.7	2.071231677	8	0.001742858	12.95694444
TRAIN_000028	SG	GIW5	Container	24.19579932	2018-02-15 13:48	Q311584	40	9	65200	20	10	51900	260	TKWC13	Marshall Islands					21	0.002615416	1510.709167
TRAIN_000029	KR	RKA2	Cargo	0	2021-02-24 23:30	L258551	20	12	7670	10	10	5140	120	DQGW56	China, Pec	-1.3	0.29	1.4	1.364128179	8	0.000148811	0
TRAIN_000030	CN	EUC8	Container	78.61684388	2022-06-22 5:49	N686823	20	11	11900	10	10	9890	150	EGOJ72	Korea, Sou	0.62	1.07	24.6	1.298255122	13	0.001614168	54.60305556
TRAIN_000031	CN	YRT6	Bulk	0	2018-07-03 14:28	F415413	30	10	48600	20	10	30200	190	HZSD87	China, People's Republic Of					22	0.000360428	0
TRAIN_000032	CN	NGG6	Container	117.5500302	2019-11-14 15:58	L753758	20	6	12300	10	10	9990	140	ENFG14	Hong Kon	-2.32	4.48	13.3	3.314675386	23	0.001742858	14.97055556

*** test 데이터**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
SAMPLE_ID	ARI_CO	ARI_PO	SHIP_TYPE	DIST	ATA	ID	BREADTH	BUILT	DEADWEI	DEPTH	DRAUGHT	GT	LENGTH	SHIPMAN	FLAG	U_WIND	V_WIND	AIR_TEMP	BN	ATA_LT	PORT_SIZE
TEST_000000	SG	GIW5	Container	1.826588579	2020-06-18 11:58	K322654	50	18	117000	30	20	109000	340	KQMD74	Panama	0.37	1.63	27.1	1.587063314	19	0.002615416
TEST_000001	CN	WEY7	Cargo	25.39938609	2021-05-26 22:20	E376681	10	13	3810	10	10	2560	80	LBYS27	Belize	-2.79	-2.33	14.2	2.663971821	6	0.001027827
TEST_000002	CN	NGG6	Container	111.0794666	2019-12-16 0:09	F811111	20	26	10900	10	10	8440	140	EKFV15	Singapore	0.04	-4.91	9.3	3.255315483	8	0.001742858
TEST_000003	CA	FFM2	Bulk	9.175258136	2015-11-16 5:30	A737561	30	9	55800	20	10	31500	190	MASW32	Panama					2	0.000181866
TEST_000004	JP	QYY1	Container	0	2018-10-24 1:11	A827175	30	19	39800	20	10	33000	220	SIEL54	Liberia					10	0.000551732
TEST_000005	BR	TMW2	Bulk	0	2019-03-19 3:59	J321515	30	7	81900	20	10	43100	230	NAED28	Marshall I	0	0	28.3	0	1	8.00E-05
TEST_000006	IN	UJM2	Bulk	15.56861608	2022-04-28 4:50	N677182	30	15	58800	20	10	32400	190	ZAKR16	Marshall I	9.86	5.63	36.3	5.692442474	10	0.000216617
TEST_000007	CA	FFM2	Bulk	9.508438896	2019-01-27 21:10	N786165	30	10	58000	20	10	32400	190	XXFC66	Hong Kon	0	0	6.4	0	17	0.000181866
TEST_000008	SG	GIW5	Container	24.78372206	2017-04-05 5:21	D541785	60	8	199000	30	20	192000	400	KQMD74	Panama					13	0.002615416
TEST_000009	AU	KSF1	Bulk	35.57995566	2015-12-12 4:14	U444818	30	11	81900	20	10	44300	230	BDOS41	Marshall Islands					15	5.75E-05
TEST_000010	AU	WHH4	Bulk	25.48101129	2014-11-22 1:50	H615556	60	9	251000	30	20	133000	330	YQFC33	Japan					12	0.000103275
TEST_000011	CN	WAF5	Bulk	0	2019-01-03 23:26	L133228	30	16	33500	10	10	20000	180	HZSD87	China, People's Republic Of					7	0.000617532
TEST_000012	MY	LHD1	Bulk	10.15347641	2023-02-07 8:45	C624433	40	12	92800	20	10	51200	230	VBDT16	Liberia	1.64	-5.73	28.5	3.704221852	16	1.55E-05
TEST_000013	RU	FCD5	Container	0.001576814	2018-02-09 5:30	Z638542	30	13	41400	20	10	35900	210	NITL57	Antigua & Barbuda					7	0.000398636
TEST_000014	SG	GIW5	Container	31.30004468	2022-01-10 20:07	U214617	40	24	68800	20	10	66300	280	TQXV21	Liberia	-0.86	-4.78	25.1	3.231663909	4	0.002615416
TEST_000015	CN	JEN5	Bulk	14.91571572	2019-01-22 17:04	H126285	30	26	46700	20	10	27000	190	NFMS48	China, Pec	-4.26	2.51	3.3	3.270457308	1	0.00102563
TEST_000016	CA	BAZ5	Bulk	12.70293409	2014-12-20 3:37	X867471	40	10	87200	20	10	48100	230	TUYZ57	Singapore					0	3.94E-05
TEST_000017	TW	JWI3	Container	6.512424005	2019-07-16 3:02	D377771	20	17	6900	10	10	5400	120	PBZV77	Bahamas	1.55	-5	28.4	3.397237248	11	0.000989649
TEST_000018	RU	AZU6	Bulk	4.742566276	2020-06-24 5:45	L545226	30	13	55100	20	10	31000	190	ZFFW62	Malta	-0.28	1	13.7	1.155550974	7	0.000176888
TEST_000019	CN	EUC8	Container	0	2021-07-28 4:57	V843875	30	12	23300	10	10	18400	180	UILI44	Portugal (	-4.53	6.68	26.6	4.534044032	12	0.001614168
TEST_000020	TW	JWI3	Bulk	187.0438299	2017-05-14 11:18	K657125	50	7	210000	20	20	107000	300	CJLN55	Panama					19	0.000989649
TEST_000021	CN	NGG6	Container	126.7541046	2018-01-02 10:17	N667434	20	28	11000	10	10	8000	140	LWMA34	Hong Kong, China					18	0.001742858
TEST_000022	MY	EFG4	Container	0	2020-05-05 5:01	W617545	20	25	5200	10	10	3840	100	CMLA18	Vanuatu	1.04	1.34	27.6	1.602700967	13	7.01E-05
TEST_000023	CN	JTD1	Bulk	21.09388328	2020-10-19 8:50	W285657	30	9	75300	20	10	40900	220	DDBP51	China, Pec	1.71	-4.48	11	3.204329282	16	0.000556558
TEST_000024	TW	EKP8	Container	0	2022-02-07 6:57	N552716	20	16	13700	10	10	9920	150	ENFG14	Hong Kon	0.06	-0.67	18.2	0.865104709	14	0.000427115
TEST_000025	JP	QYY1	Cargo	0	2022-10-28 2:27	J361314	10	17	1600	10	0	500	80	GMAU61	Japan	0.44	-1.69	16.6	1.63411088	11	0.000551732
TEST_000026	CN	NCU8	Container	0	2018-03-27 0:12	P188467	40	21	69200	20	10	65500	280	WWYY24	United Kingdom					8	0.000939155
TEST_000027	CN	EKP8	Container	0	2020-12-17 23:28	G668727	30	4	35300	20	10	26800	190	OAJG44	Hong Kon	-1.63	-9.89	-0.6	5.238494693	7	0.001660242
TEST_000028	CN	JTD1	Bulk	0	2016-11-10 20:18	Q172674	30	27	25400	10	10	15100	160	QWOE62	Cameroon					4	0.000556558

2. 데이터 확인

*** 데이터 이해**

- 학습 데이터와 시험데이터는 모두 결측치를 가지고 있으며, 공통적으로 U_WIND, V_WIND, AIR_TEMPERATURE, BN변수에서 결측치를 갖는다는 공통점이 있다.
- 기본적으로 DIST변수가 0의 값을 가지면, CI_HOUR의 값 또한 0을 가진다는 것을 전제로 한다.
- PORT_SIZE는 도착항의 항구마다 다르며, 즉 항구마다 항구크기가 겹치지 않아, PORT_SIZE를 도착항을 대표한다고 할 수 있다.

Feature Name	Description	Feature Name	Description
ARI_CO	도착항의 소속국가(도착항 앞 2글자)	GT	용적톤수(Gross Tonnage)값
ARI_PO	도착항의 항구명(도착항 뒤 글자)	LENGTH	선박의 길이
SHIP_TYPE_CATEGORY	선종 통합 바탕으로 5대 선종으로 분류	SHIPMANAGER	선박 소유주
DIST	정박지(ber_port)와 접안지 사이의 거리	FLAG	선박의 국적
ATA	anc_port에 도착한 시점의 utc. 실제 정박 시각(Actual Time of Arrival)	U_WIND	풍향 u벡터
ID	선박식별 일련번호	V_WIND	풍향 v벡터
BREADTH	선박의 폭	AIR_TEMPERATURE	기온
BUILT	선박의 연령	BN	보퍼트 풍력 계급
DEADWEIGHT	선박의 재화중량톤수	ATA_LT	anc_port에 도착한 시점의 현지 정박 시각(Local Time of Arrival)(단위 : H)
DEPTH	선박의 깊이	PORT_SIZE	접안지 폴리곤 영역의 크기
DRAUGHT	흘수 높이	**CI_HOUR(target)**	**대기시간**

3. EDA 및 데이터 전처리

EDA에 앞서..

결측치 처리

앞서 확인되었던 U_WIND, V_WIND,
AIR_TEMPERATURE, BN의 결측치를
데이터가 완전한 2021년 이후의 데이터를 이용해
같은 달, 같은 시간의 데이터의 평균으로 결측치 대체

이상치 제거

DIST가 0임에도 CI_HOUR가 0이 아닌 Train 데이터 제거
(28개의 데이터 제거)

CI_HOUR에 따른 데이터 셋 분류

CI_HOUR가 0인 비율이 전체 중 30%가 넘기 때문에,
따로 나누어 모델을 학습하는 것이 예측력을 높일 수 있지만
먼저 도착한 선박의 영향을 고려하기 위해 데이터셋을
분류하지 않음

날짜 형식의 ATA 변수로 파생변수 생성

EDA에 이용하기 위해, 날짜 형식의 ATA 변수로
년, 월, 일, 시간, 분을 나타내는 파생변수를 생성하고,
ATA 변수 제거

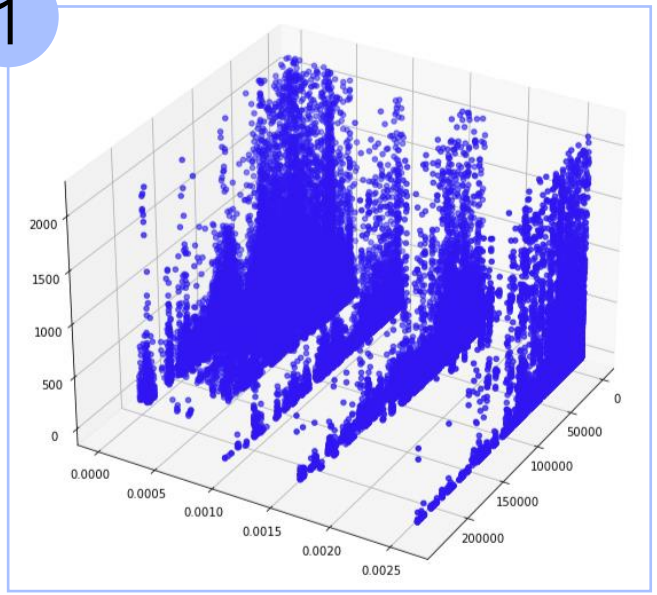
3. EDA 및 데이터 전처리

*** EDA 결과**

- 다양한 EDA로 특징을 발견하기 위한 노력이 있었으나, 아래의 세가지 정도만 유의하다고 생각됨

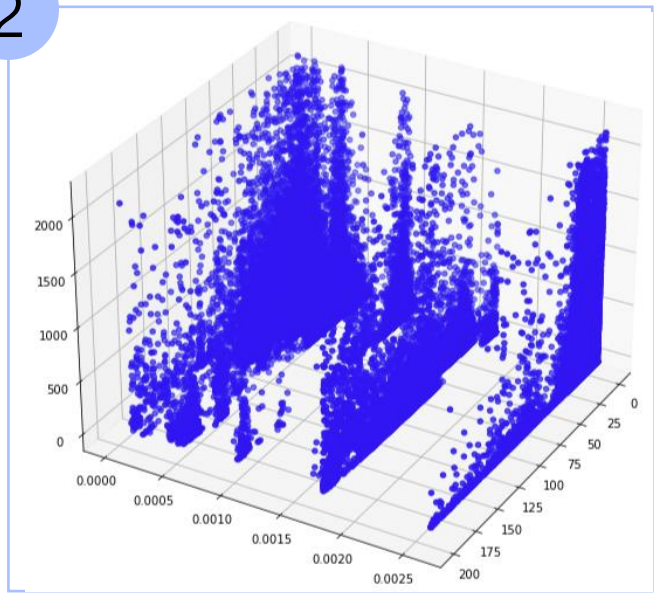
=〉 EDA의 결과에 따라, PORT_SIZE에 따라 3개의 데이터셋에 대한 각각의 모델 생성하기로 결정 (PORT_SIZE 〈= 0.0005, 0.0005 〈 PORT_SIZE 〈 0.0020, PORT_SIZE 〉 0.0020)

1



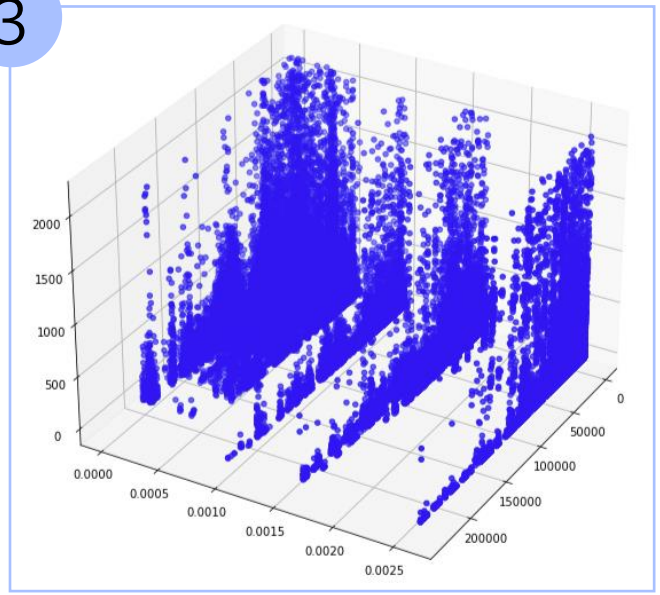
x축 GT, y축 PORT_SIZE, z축 CI_HOUR

2



x축 DIST, y축 PORT_SIZE, z축 CI_HOUR

3



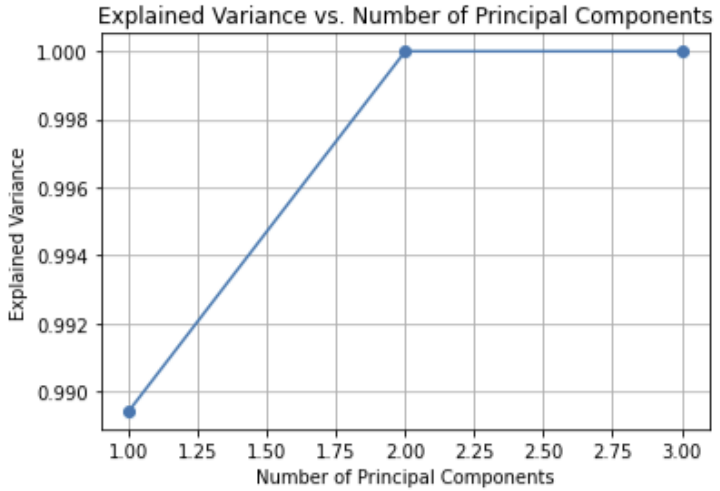
x축 DIST, y축 PORT_SIZE, z축 CI_HOUR
(SHIP_TYPE_CATEGORY = Container 경우)

3. EDA 및 데이터 전처리

*** 파생변수 생성 및 PCA로 변수 변환**

GT/Port_Size	GT / 항구별 PORT_SIZE 항구의 화물처리능력을 의미
ARI_CO_mean	도착 나라별 CI_HOUR 평균
ARI_PO_mean	도착 항구별 CI_HOUR 평균
Weekday	월~일을 나타내며 1~7로 표현됨, 휴일의 영향을 고려
SHIP_PHYSICAL	서로 상관계수가 높은 GT. DEADWEIGHT, LENGTH, BREADTH, DEPTH, DRAUGHT를 PCA를 통해 하나의 변수로 변환

PCA 결과



PCA로 생성된 한 개의 주성분(SHIP_PHYSICAL)으로도 나머지 6개의 변수 GT. DEADWEIGHT, LENGTH, BREADTH, DEPTH, DRAUGHT를 잘 설명할 수 있음을 보임

3. EDA 및 데이터 전처리

* Label Encoding 및 Min-Max scaling 수행

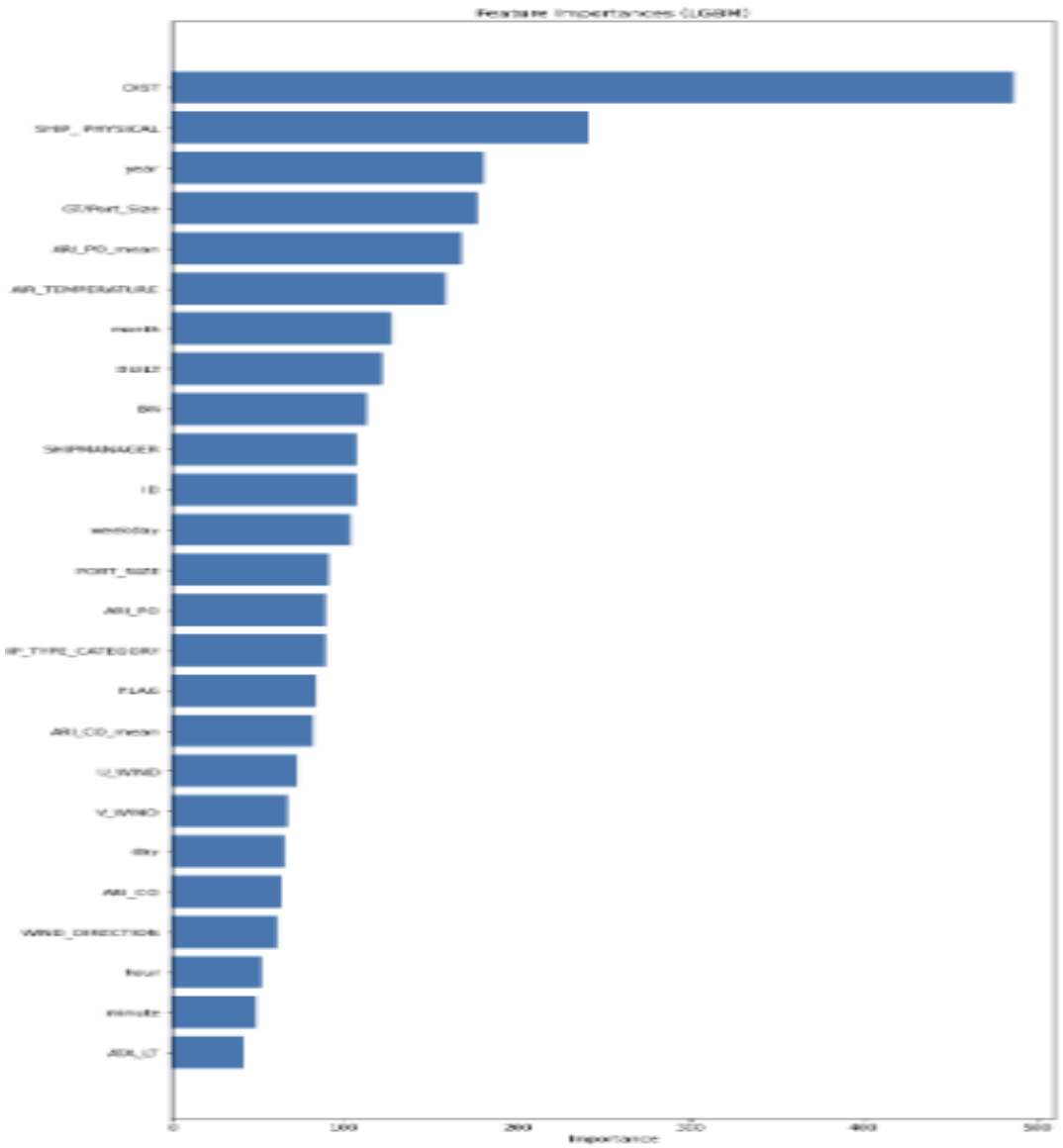
Label Encoding

- 머신 러닝과 데이터 처리에서 범주형 데이터 (카테고리형 데이터)를 숫자형으로 변환

Min-Max Scale

- 데이터를 정해진 범위 내로 스케일링하는 방법 중 하나
- 이상치(outliers)에 민감하다는 특징을 가짐
- 일반적으로 신경망과 같은 기계 학습 모델에서 입력 데이터를 전처리할 때 유용하게 사용

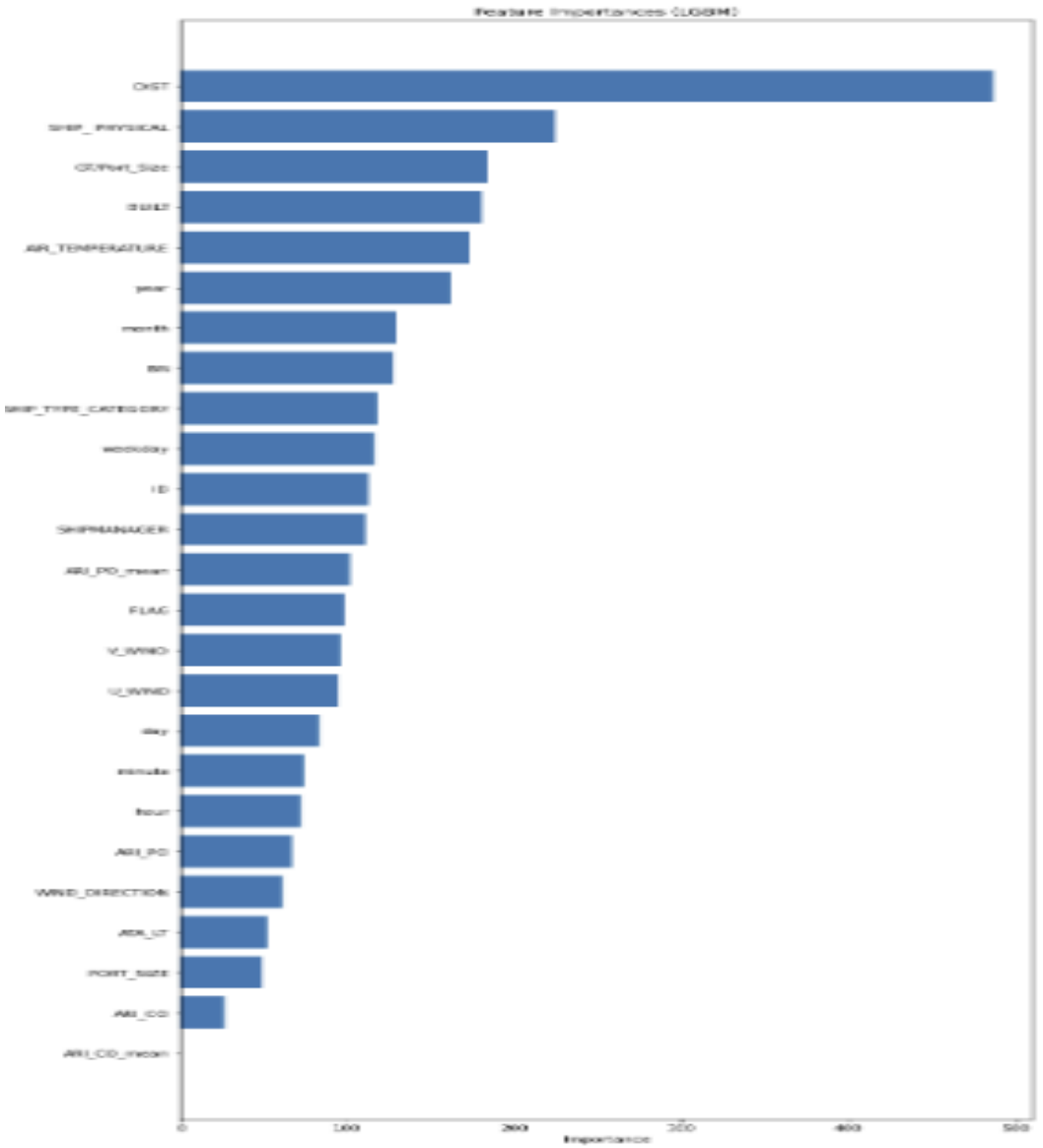
4. 모델링 및 추론



* 모델1(PORT_SIZE 〈 = 0.0005)의 학습에 이용할 변수 선택
  => LGBM 기준으로 Feature Importance가 95이상인 변수들 선정

```
Data columns (total 12 columns):
 #   Column           Non-Null Count
---  ------           --------------
 0   DIST             186615 non-null
 1   ID               186615 non-null
 2   BUILT            186615 non-null
 3   SHIPMANAGER      186615 non-null
 4   AIR_TEMPERATURE  186615 non-null
 5   BN               186615 non-null
 6   year             186615 non-null
 7   month            186615 non-null
 8   GT/Port_Size     186615 non-null
 9   ARI_PO_mean      186615 non-null
 10  weekday          186615 non-null
 11  SHIP_ PHYSICAL   186615 non-null
```

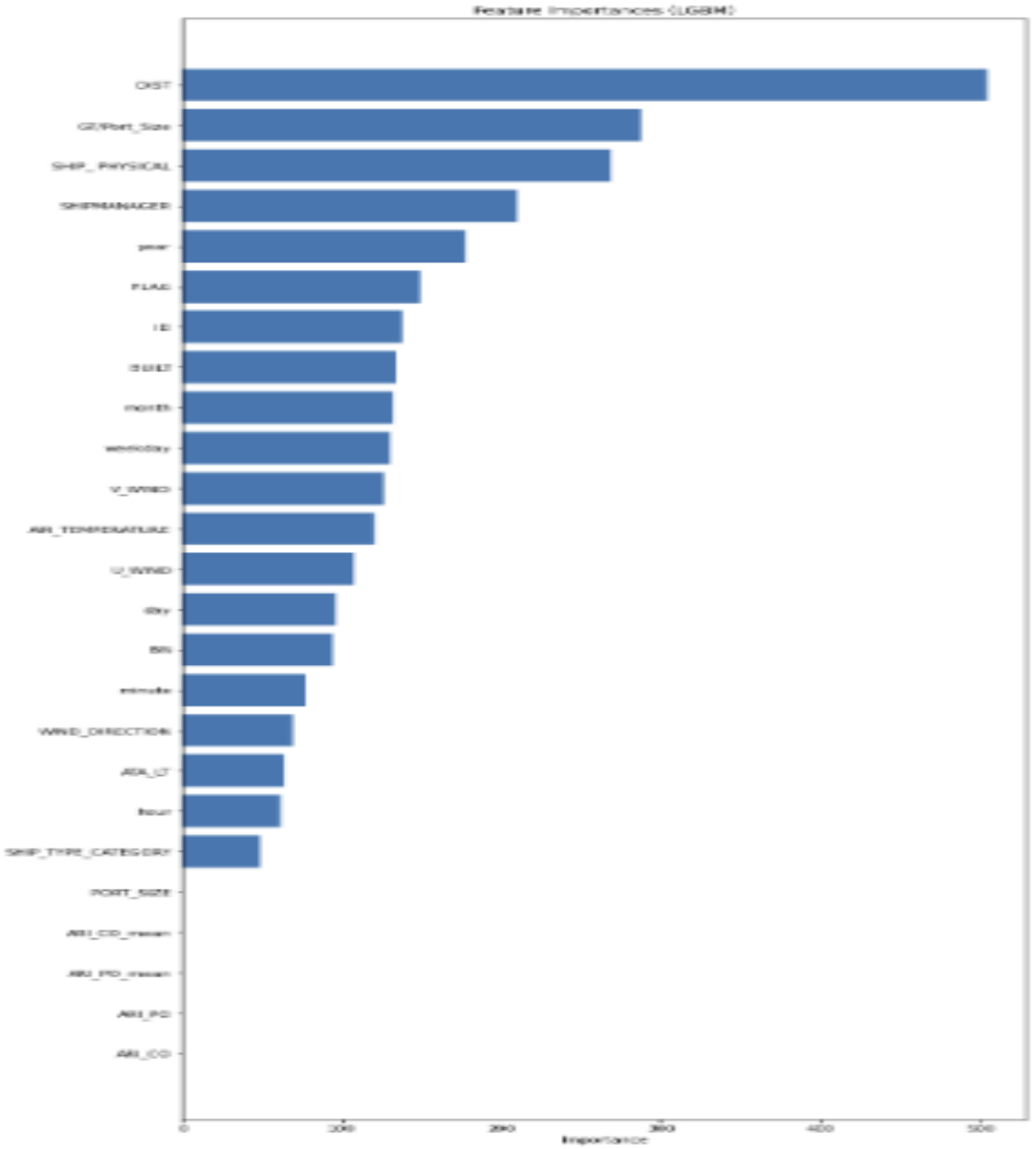
4. 모델링 및 추론



* 모델2(0.0005 〈 PORT_SIZE 〈= 0.0020)의 학습에 이용할 변수 선택
  => LGBM 기준으로 Feature Importance가 95이상인 변수들 선정

```
Data columns (total 16 columns):
 #   Column               Non-Null Count
---  ------               --------------
 0   SHIP_TYPE_CATEGORY   161435 non-null
 1   DIST                 161435 non-null
 2   ID                   161435 non-null
 3   BUILT                161435 non-null
 4   SHIPMANAGER          161435 non-null
 5   FLAG                 161435 non-null
 6   U_WIND               161435 non-null
 7   V_WIND               161435 non-null
 8   AIR_TEMPERATURE      161435 non-null
 9   BN                   161435 non-null
 10  year                 161435 non-null
 11  month                161435 non-null
 12  GT/Port_Size         161435 non-null
 13  ARI_PO_mean          161435 non-null
 14  weekday              161435 non-null
 15  SHIP_ PHYSICAL       161435 non-null
```

4. 모델링 및 추론



* 모델3(PORT_SIZE 〉 0.0020)의 학습에 이용할 변수 선택
  =〉 LGBM 기준으로 Feature Importance가 115이상인 변수들 선정

```
Data columns (total 12 columns):
 #   Column            Non-Null Count
---  ------            --------------
 0   DIST              43860 non-null
 1   ID                43860 non-null
 2   BUILT             43860 non-null
 3   SHIPMANAGER       43860 non-null
 4   FLAG              43860 non-null
 5   V_WIND            43860 non-null
 6   AIR_TEMPERATURE   43860 non-null
 7   year              43860 non-null
 8   month             43860 non-null
 9   GT/Port_Size      43860 non-null
 10  weekday           43860 non-null
 11  SHIP_ PHYSICAL    43860 non-null
```

4. 모델링 및 추론

*** 모델링 – Autogluon**

- AutoGluon은 AutoML(Automated Machine Learning) 라이브러리로, 다양한 기계 학습 작업을 자동화하고 최적화하는 데 사용됨
- AutoGluon은 여러 가지 모델을 조정하고 최상의 모델과 하이퍼 파라미터를 선택하여 모델을 구축하고 튜닝하는 작업을 자동화하기 때문에 쉽고 간편하게 이용 가능
- AutoGluon은 다음과 같은 주요 기능을 제공함

1. **다양한 모델 선택 :** 다양한 종류의 기계 학습 모델들을 지원함
2. **하이퍼 파라미터 튜닝 :** 모델링에 사용되는 하이퍼 파라미터들을 자동으로 튜닝함
3. **자동 특성 엔지니어링 :** 데이터로부터 유용한 특성을 추출하거나 생성하여 모델의 성능을 향상시킴
4. **분산 학습 및 스태킹**: 여러 머신에서 학습을 병렬로 수행하거나, 여러 모델을 결합하여 더 나은 예측을 가능케 하는 앙상블 방법을 지원함

=〉 실제 모델링에서는 regression 옵션을 통해, 회귀에 적합한 모델들만 사용하였고, 앙상블 옵션과, 스태킹 옵션 또한 이용함.
추가로 평가지표를 선택할 수 있는 옵션이 제공되어, 평가지표인 MAE를 기준으로 validatio의 MAE를 확인할 수 있다.

*** 평가지표 – MAE**

- MAE(평균 절대 오차, Mean Absolute Error)는 회귀(Regression) 모델의 성능을 측정하는 평가 지표 중 하나
- 모델이 예측한 값과 실제 값 사이의 차이를 절대값으로 계산하고, 이를 모든 데이터 포인트에 대해 평균을 내어 구함
- MAE값이 낮을수록 예측 성능이 좋다고 할 수 있음

4. 모델링 및 추론

```
Fitting model: LightGBMXT_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -56.995  = Validation score   (-mean_absolute_error)

        6.4s     = Training   runtime
        4.19s    = Validation runtime
Fitting model: LightGBM_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -56.9438         = Validation score   (-mean_absolute_error)
        7.74s    = Training   runtime
        3.58s    = Validation runtime
Fitting model: RandomForestMSE_BAG_L3 ...
        -60.3938         = Validation score   (-mean_absolute_error)
        226.86s  = Training   runtime
        7.41s    = Validation runtime
Fitting model: CatBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -56.9568         = Validation score   (-mean_absolute_error)
        13.38s   = Training   runtime
        0.12s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L3 ...
        -59.71   = Validation score   (-mean_absolute_error)
        27.9s    = Training   runtime
        6.07s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -54.3366         = Validation score   (-mean_absolute_error)
        226.98s  = Training   runtime
        9.2s     = Validation runtime
Fitting model: XGBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -52.0631         = Validation score   (-mean_absolute_error)
        9.02s    = Training   runtime
        1.46s    = Validation runtime
Fitting model: NeuralNetTorch_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -50.2121         = Validation score   (-mean_absolute_error)
        201.07s  = Training   runtime
        3.68s    = Validation runtime
Fitting model: LightGBMLarge_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -57.1414         = Validation score   (-mean_absolute_error)
        9.57s    = Training   runtime
        7.6s     = Validation runtime
Fitting model: WeightedEnsemble_L4 ...
        -50.2121         = Validation score   (-mean_absolute_error)
        3.66s    = Training   runtime
        0.0s     = Validation runtime
```

- 모델1(PORT_SIZE 〈 = 0.0005)의 Autogluon 결과
  => WeightedEnsemble_L4 모형의 Validation MAE가 50.2121로 가장 낮게 나타남

Model	Validation MSE
LightGBMXT_BAG_L3	56.995
LightGBM_BAG_L3	56.938
RandomForestMSE_BAG_L3	60.3938
CatBoost_BAG_L3	56.9568
ExtraTreesMSE_BAG_L3	59.71
NeuralNetFastAI_BAG_L3	54.3366
XGBoost_BAG_L3	52.0631
NeuralNetTorch_BAG_L3	50.2121
LightGBMLarge_BAG_L3	57.1414
WeightedEnsemble_L4	50.2121

4. 모델링 및 추론

```
Fitting model: LightGBMXT_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -30.3741        = Validation score   (-mean_absolute_error)
        7.12s   = Training   runtime
        3.57s   = Validation runtime
Fitting model: LightGBM_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -30.2323        = Validation score   (-mean_absolute_error)
        6.65s   = Training   runtime
        2.97s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L3 ...
        -32.8365        = Validation score   (-mean_absolute_error)
        197.39s  = Training   runtime
        6.0s    = Validation runtime
Fitting model: CatBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -30.2057        = Validation score   (-mean_absolute_error)
        11.82s   = Training   runtime
        0.11s   = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L3 ...
        -32.628  = Validation score   (-mean_absolute_error)
        29.06s   = Training   runtime
        5.32s   = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -28.7976        = Validation score   (-mean_absolute_error)
        200.25s  = Training   runtime
        9.1s    = Validation runtime
Fitting model: XGBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -27.0164        = Validation score   (-mean_absolute_error)
        9.79s   = Training   runtime
        1.15s   = Validation runtime
Fitting model: NeuralNetTorch_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -25.9751        = Validation score   (-mean_absolute_error)
        154.19s  = Training   runtime
        3.24s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -30.4952        = Validation score   (-mean_absolute_error)
        8.54s   = Training   runtime
        5.51s   = Validation runtime
Fitting model: WeightedEnsemble_L4 ...
        -25.9689        = Validation score   (-mean_absolute_error)
        3.22s   = Training   runtime
        0.01s   = Validation runtime
```

* 모델2(0.0005 〈 PORT_SIZE 〈= 0.0020) 의 Autogluon 결과
=〉 WeightedEnsemble_L4 모형의 Validation MAE가 25.9751로 가장 낮게 나타남

Model	Validation MSE
LightGBMXT_BAG_L3	30.3741
LightGBM_BAG_L3	30.2323
RandomForestMSE_BAG_L3	32.8365
CatBoost_BAG_L3	30.2057
ExtraTreesMSE_BAG_L3	32.628
NeuralNetFastAI_BAG_L3	28.7976
XGBoost_BAG_L3	27.0164
NeuralNetTorch_BAG_L3	25.9751
LightGBMLarge_BAG_L3	30.4952
WeightedEnsemble_L4	25.9751

4. 모델링 및 추론

```
Fitting model: LightGBMXT_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -100.6855       = Validation score   (-mean_absolute_error)
        4.66s    = Training   runtime
        1.39s    = Validation runtime
Fitting model: LightGBM_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -100.475        = Validation score   (-mean_absolute_error)
        4.67s    = Training   runtime
        0.94s    = Validation runtime

Fitting model: RandomForestMSE_BAG_L3 ...
        -107.0598       = Validation score   (-mean_absolute_error)
        46.04s   = Training   runtime
        2.01s    = Validation runtime
Fitting model: CatBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -100.1354       = Validation score   (-mean_absolute_error)
        7.79s    = Training   runtime
        0.05s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L3 ...
        -107.6784       = Validation score   (-mean_absolute_error)
        8.06s    = Training   runtime
        1.93s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -95.9504        = Validation score   (-mean_absolute_error)
        58.37s   = Training   runtime
        2.28s    = Validation runtime
Fitting model: XGBoost_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -93.0233        = Validation score   (-mean_absolute_error)
        7.22s    = Training   runtime
        0.3s     = Validation runtime
Fitting model: NeuralNetTorch_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -81.0231        = Validation score   (-mean_absolute_error)
        75.63s   = Training   runtime
        1.08s    = Validation runtime
Fitting model: LightGBMLarge_BAG_L3 ...
        Fitting 6 child models (S1F1 - S3F2) | Fitting with ParallelLocalFoldFittingStrategy
        -101.6536       = Validation score   (-mean_absolute_error)
        6.42s    = Training   runtime
        1.74s    = Validation runtime
Fitting model: WeightedEnsemble_L4 ...
        -81.0115        = Validation score   (-mean_absolute_error)
        0.87s    = Training   runtime
        0.0s     = Validation runtime
```

* 모델3(PORT_SIZE 〉 0.0020) 의 Autogluon 결과

=〉 WeightedEnsemble_L4 모형의 Validation MAE가 81.0115로 가장 낮게 나타남

Model	Validation MSE
LightGBMXT_BAG_L3	100.6855
LightGBM_BAG_L3	100.475
RandomForestMSE_BAG_L3	107.0598
CatBoost_BAG_L3	100.1354
ExtraTreesMSE_BAG_L3	107.6784
NeuralNetFastAI_BAG_L3	95.9504
XGBoost_BAG_L3	93.0233
NeuralNetTorch_BAG_L3	81.0231
LightGBMLarge_BAG_L3	101.6536
WeightedEnsemble_L4	81.0115

5. 결론

결론 및 아쉬운 점

*모델 1과 모델 2는 Validation MAE가 각각 약 50, 25로 양호했던 반면, 모델 3는 약 81로 좋지 못한 예측력을 나타냈음
 => 모델 3 즉, PORT_SIZE가 0.0020보다 큰 항구에 대해선 모델 1, 2와 결합하거나 또다른 방식으로 학습을 하는 것이 타당해 보임

*최종적으로 발표한 모델을 통해 public 45.18991, private 44.96305를 얻을 수 있었음
 => public보다 private에서 점수가 조금 더 높았던 이유는 배깅과 스태킹을 통한 앙상블로 인해 과적합이 방지되었을 것으로 생각됨

* 데이터를 통해 특정한 패턴을 확인하지 못했고, 도메인 지식이 너무 부족했다는 것이 대회 내내 느껴져서 아쉬웠음

감사 합니다