

# 선박 항차 데이터 분석

머신 러닝을 활용한 선박 대기 시간 예측







## 팀원 소개

### **Team Member**



조재경

- 데이터 분석
- PPT 제작
- 발표



최서윤

- 데이터 분석
- 문서작업

### Leader



### 이보윤

- 데이터 전처리
- 데이터 분석
- 총괄 지휘



#### 김동욱

- 데이터 전처리
- 데이터 시각화
- 데이터 분석
- 발표



박재현

- 데이터 전처리
- 데이터 분석
- PPT 제작
- 발표





- 1) 분석 배경
- 2) 프로젝트 이해

- Data
- 1) 데이터 수집
- 2) 데이터 전처리
- 3) 데이터 베이스

### **ANALYSIS**

- 1) 사용 프로그램
- 2) 머신 러닝
- 3) 모델 해석

### **CONCLUSION**

- 1) 기대 효과
- 2) 활용 방안







- 1) 분석 배경
- 2) 프로젝트 이해

- 1) 데이터 수집
- 2) 데이터 전처리
- 3) 데이터 베이스

- 1) 사용 프로그램
- 2) 머신 러닝
- 3) 모델 해석

- 1) 기대 효과
- 2) 활용 방안





### 분석 배경

### 코로나19 이후 해운 물류 현황

코로나19 이후 물류 정체로 인해 다수의 항만에서 선박 대기 시간이 길어지고, 이로 인한 물류 지연이 화두가 되고 있습니다.

특히 전 세계 물동량의 85%를 차지하는 해운 물류 분야에서 항만 정체는 큰 문제로 인식되고 있는 상황입니다.

저희 조는 접안 전에 선박이 해상에 정박하는 시간을 대기시간으로 정의하고, 선박의 제원 및 운항 정보를 활용하여 산출된 항차 데이터를 활용해 항만 內 선박의 대기 시간을 예측하는 AI 알고리즘을 개발하였습니다.





### 분석 배경

불 부산일보

항만적체 손실 갈수록 늘어

1주 전



- 카고뉴스

항만체선 현상 손실 '눈덩이'

2023. 9. 14.



말로 물류신문

선박 대기시간 줄어 경제손실비용 감소



2023. 5. 11.



[울산항 정박지 부족, 여전히 과제] 대기공간 없어 손실안고 표류하는 선박들 2023. 11. 1.



이투데이

5년간 선박 체선 2조 넘는 손실 발생



2023. 6. 25.

### 선박대기시간 감소의 중요성

기상악화 혹은 항만 사정으로 선박이 항만에 접안하지 못한 채 12시간 이상 대기하는 체선(滯船) 현상으로 인해 매년 큰 손실을 보고 있습니다.

특히 체선 현상은 정박지 부족, 기상 악화 등 다양한 이유로 일어나기 때문에 정확한 원인을 파악하는 것이 중요합니다.

보다 정확한 선박의 접안 시간 예측을 통해 선박의 대기시간을 줄임으로써 연료 절감 및 온실가스 감축 효과를 기대할 수 있습니다.





### 프로젝트 이해

### 프로젝트 시나리오

선박 예약 \_ . . .. 및 계획

화물 적재 및 포장

항구 진입 및 통과

화물 실적 및 보관

항구 출항

항로 운항 및 도착

항구 입항 및 양하

화물 양하 및 해제

#### Model

Regression

XGBRegressor

#### Collection

#### **Preprocessing**

#### **Visualization**

#### Modelling

RandomForest

GradientBoosting

HistGradientBoosting

LGBMRegressor

- 데이터 수집
- MariaDB로 저장

- 이상치, 결측치, 중복치 처리
- 데이터 표준화 및 정규화
- 상관관계 분석
- 주요 특성 추출

- 데이터 시각화
- EDA (탐색적 데이터 분석)
- 모델 생성
- 하이퍼파라메터 튜닝
- 모델 검증 및 예측
- 최종 모델 선정



# 프로젝트 이해





Cargo



Bulk



Container







### **INTRO**

- 1) 분석 배경
- 2) 프로젝트 이해

### Data

- 1) 데이터 수집
- 2) 데이터 전처리
- 3) 데이터 베이스

### **ANALYSIS**

- 1) 사용 프로그램
- 2) 머신 러닝
- 3) 모델 해석

### **CONCLUSION**

- 1) 기대 효과
- 2) 활용 방안





### 데이터 수집



데이터 목록								
	Α	В	С	D	Е	F	G	
1	SAMPLE_II	ARI_CO	ARI_PO	SHIP_TYPE	DIST	ATA	ID	В
2	TEST_0000	SG	GIW5	Container	1.826589	#######	K322654	
3	TEST_0000	CN	WEY7	Cargo	25.39939	#######	E376681	
4	TEST_0000	CN	NGG6	Container	111.0795	#######	F811111	
5	TEST_0000	CA	FFM2	Bulk	9.175258	#######	A737561	
6	TEST_0000	JP	QYY1	Container	0	#######	A827175	
7	TEST_0000	BR	TMW2	Bulk	0	#######	J321515	
8	TEST_0000	IN	UJM2	Bulk	15.56862	#######	N677182	
9	TEST_0000	CA	FFM2	Bulk	9.508439	#######	N786165	
10	TEST_0000	SG	GIW5	Container	24.78372	#######	D541785	
11	TEST_0000	AU	KSF1	Bulk	35.57996	#######	U444818	
12	TEST_0000	AU	WHH4	Bulk	25.48101	#######	H615556	
13	TEST_0000	CN	WAF5	Bulk	0	#######	L133228	
14	TEST_0000	MY	LHD1	Bulk	10.15348	#######	C624433	
15	TEST_0000	RU	FCD5	Container	0.001577	#######	Z638542	
16	TEST_0000	SG	GIW5	Container	31.30004	#######	U214617	
17	TEST_0000	CN	JEN5	Bulk	14.91572	#######	H126285	
18	TEST_0000	CA	BAZ5	Bulk	12.70293	#######	X867471	
19	TEST_0000	TW	JWI3	Container	6.512424	#######	D377771	
20	TEST_0000	RU	AZU6	Bulk	4.742566	#######	L545226	
21	TEST_0000	CN	EUC8	Container	0	#######	V843875	
22	TEST_0000	TW	JWI3	Bulk	187.0438	#######	K657125	
23	TEST_0000	CN	NGG6	Container	126.7541	#######	N667434	
24	TEST_0000	MY	EFG4	Container	0	#######	W617545	

- 총 391933개의 데이터 수집



### 데이터 이해 분석에 사용할 변수 목록

#### 독립 변수

항구 1곳(RGT8) 삭제

- 이유 : 데이터 null 값 (U\_WIND, V\_WIND, AIR\_TEMPERATURE, BN)

#### 종속 변수

대기시간 (CI\_HOUR)

소속국가 (ARI\_CO) 항구명 (ARI PO)

선박종류 (SHIP\_TYPE\_CATEGORY)

접안지와의 거리 (DIST)

정박시간 (ATA)

선박 일련번호 (ID)

선박의 폭 (BREADTH)

선박의 연령 (BUILT)

선박의 재화중량 (DEADWEIGHT)

선박의 깊이 (DEPTH)

흘수 높이 (DRAUGHT)

용적톤수 (GT)

선박의 길이 (LENGTH)

선박 소유주 (SHIPMANAGER)

선박의 국적 (FLAG)

풍향 u벡터 (U WIND)

풍향 v벡터 (V\_WIND)

기온 (AIR\_TEMPERATURE)

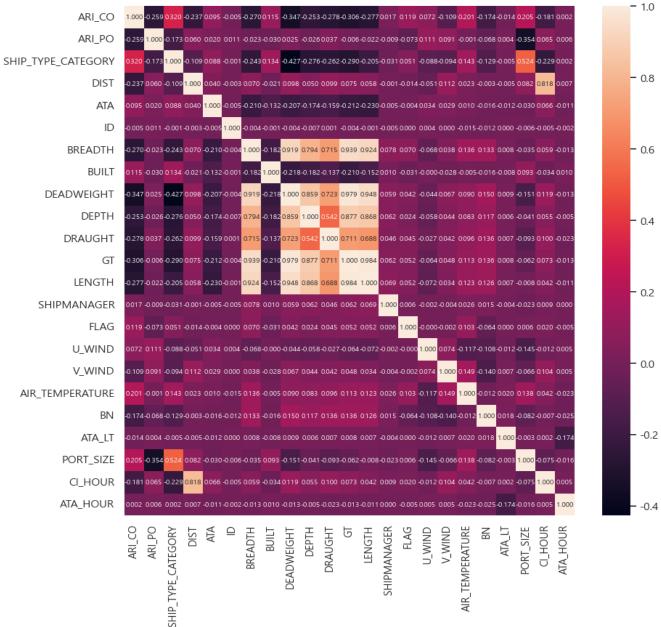
보퍼트 풍량 계급 (BN)

정박시간(ATA\_LT)

접안지 영역의 크기 (PORT SIZE)



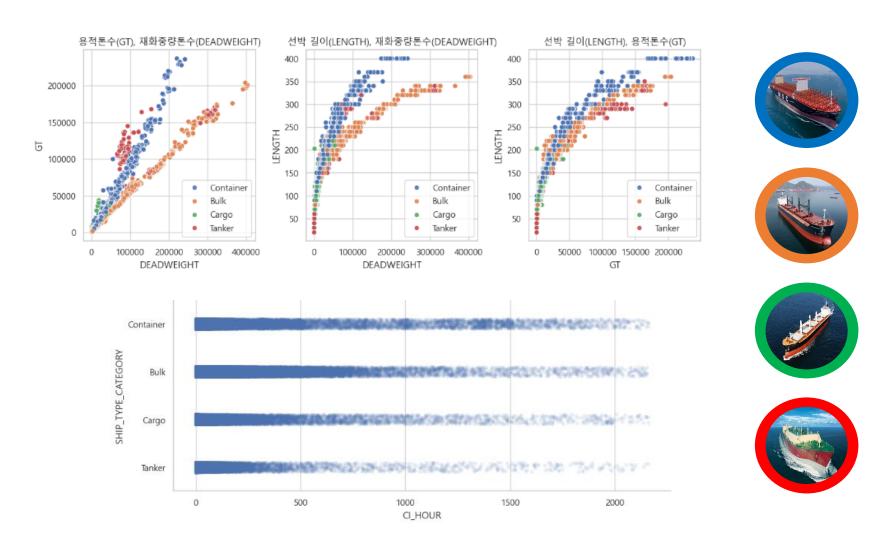
### 데이터 이해\_상관분석





## 데이터 이해

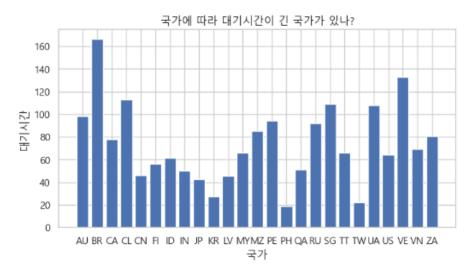
### 범주형 데이터 시각화

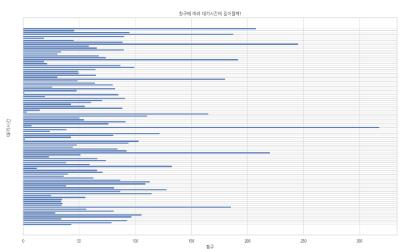




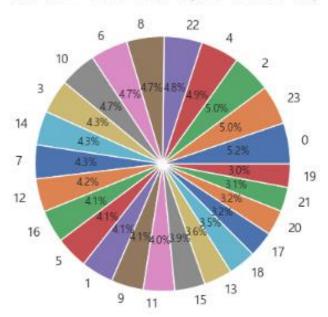
## 데이터 이해

### 범주형 데이터 시각화





#### 집중되는 도착시간이 있을까?(정박 시간대별 비율)

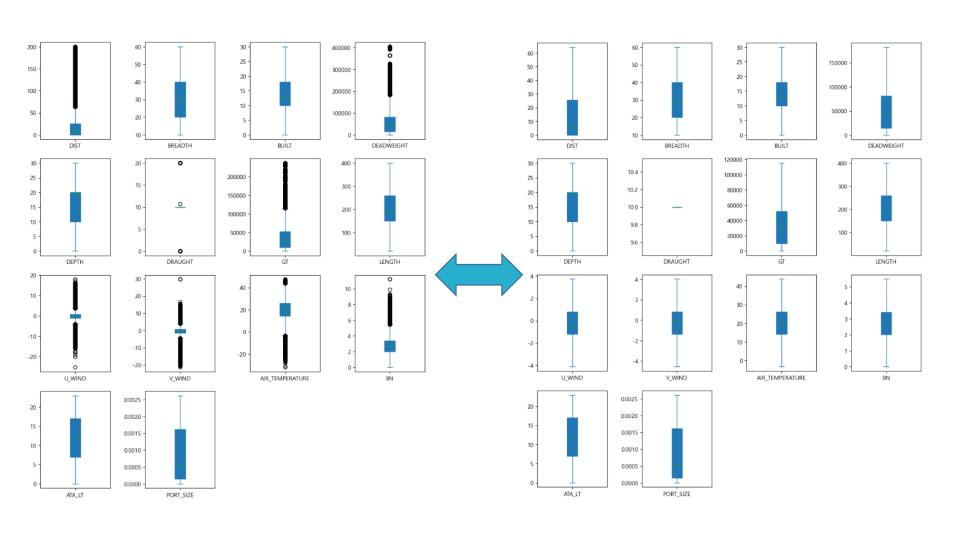






## 데이터 전처리

### 결측치 & 이상치 처리





### 데이터 베이스

### MariaDB를 활용한 데이터 공유&관리





### 목차



- 1) 분석 배경
- 2) 프로젝트 이해
- 1) 데이터 수집
- 2) 데이터 전처리
- 3) 데이터 베이스

- 1) 사용 프로그램
- 2) 머신 러닝
- 3) 모델 해석

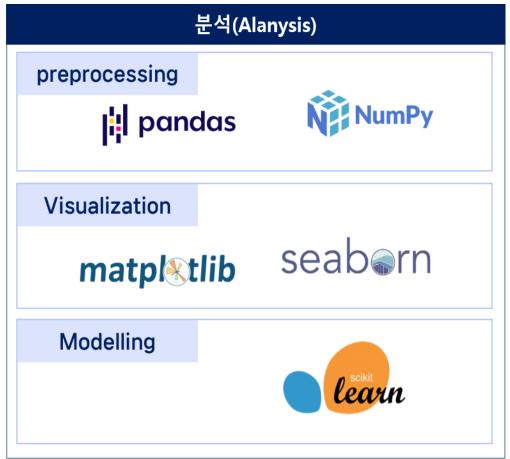
- 1) 기대 효과
- 2) 활용 방안



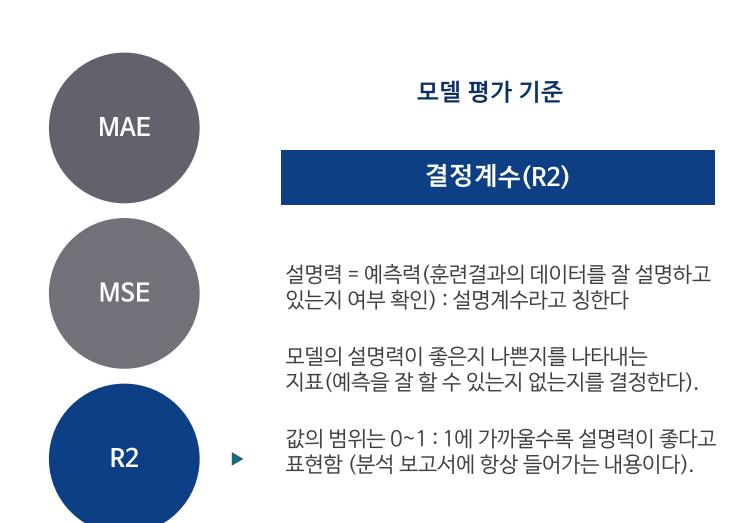


### 사용 프로그램











### (Data 1)

### 선박의 종합수명 평가

한국과학기술정보연구원 전문연구위원 이순요 (9826@reseat.re.kr)

#### 1. 머리말

- 일반적으로 선박은 설계수명 25년 문 표준으로 하여 설계되고 건조된다. 그렇지만 강도 혹은 기능의 열화에 의한 선박의 물리적인 수명은 취항 후의 취항해역, 조선 조건과 보전관리 조건 등에 따라 해마다 달라지므 로 설계수명을 맞이했다고 해서 선박이 수명을 다했다고는 말할 수 없 다.
- 가령 LNG선과 같이 초기단계부터 잘 짜여진 스펙(spec.)에 의해 설계되고 건조되며, 또한 취항 후에도 보전관리가 잘 되고 있는 배의 경우 설계수명을 넘어 더욱 수명이 긴 운행계획이 지향된다. 이렇게 긴 수명을 유지하려면 선박이 안전하면서도 안정적인 운행을 해야 하는데, 선박의 물리수명에 영향을 주는 부식과 피로라고 하는 경년 요인에 대한 배려가필요하다.
- 이러한 경년 열화요인에 대해 선체 구조와 기기를 충분히 안전한 레벨로 유지하는 것이 필요하지만, 한편으로는 선박의 수명주기 비용을 적정하게 하는 것도 중요하다. 이러한 일을 실현하기 위해서는 첫째, 열화상태의 감시와 정확한 파악, 둘째, 잔여수명 혹은 잔여강도의 적절한 평가, 셋째, 엄정하고도 정확한 점검과 검사계획에 따르는 효과적인 보전계획의 수립 과 실시가 필요하다.



제거한 컬림



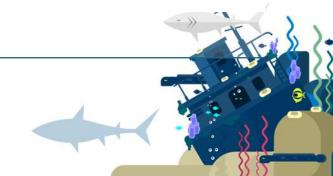
1딩



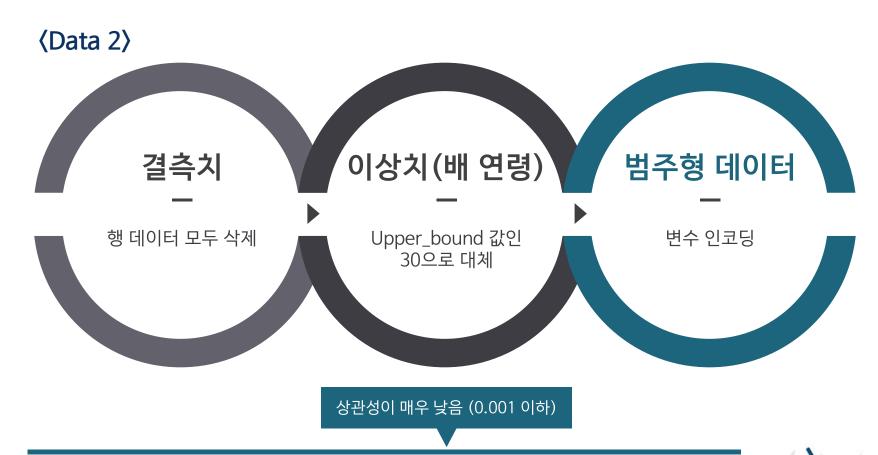


### ⟨Data 1⟩

스케일링	모델	MAE	MSE	R2	
MinMaxScaler	HGB	55.010970	20873.429794	0.281361	
RobustScaler	HGB	55.051707	20994.448045	0.277194	
StandardScaler	HGB	54.994847	20928.868991	0.279452	







제거한 컬럼 : ID, SHIPMANAGER, BN, ATA\_LT, month, day, hour, minute, ARI\_PO, V\_WIND



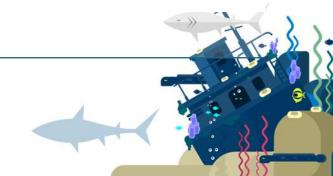
### ⟨Data 2⟩

스케일링	모델	MAE	MSE	R2
MinMaxScaler	HGB	61.197034	246104765	0.252483
	XGB (랜덤1)	60.037728	23507.862577	0.276062
	XGB (랜덤2)	59.499774	23721.862577	0.269470
	XGB (랜덤3)	59.690386	23836.138246	0.265951
StandardScaler	HGB	60.981814	24318.273863	0.251104
	XGB (랜덤1)	60.139816	24362.103766	0.249754
	XGB (랜덤2)	61.004618	23725.901519	0.269346
				· »

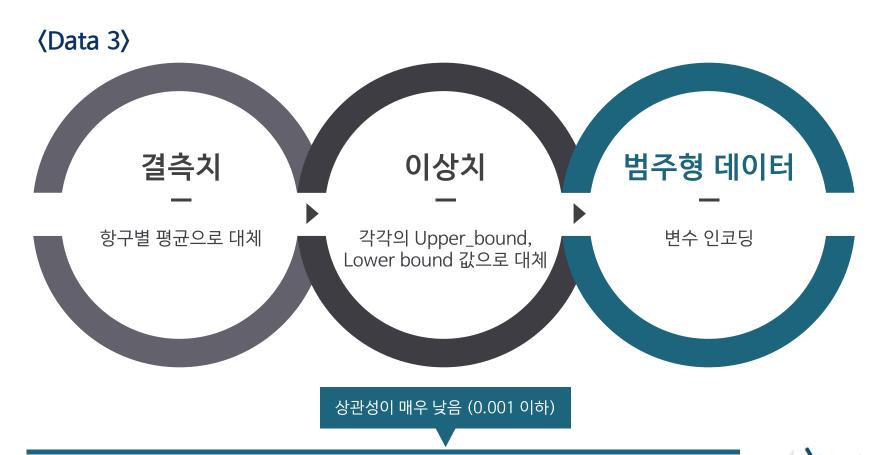


### ⟨Data 2⟩

스케일링	모델	MAE	MSE	R2
RobustScaler	HGB	60.933237	24125.222525	0.257049
	XGB (랜덤1)	64.699298	24563.976314	0.243537
	XGB (랜덤2)	59.387254	23990.679197	0.261192
	XGB (랜덤3)	60.219166	23543.196341	0.274973
	GBR (랜덤1)	61.111228	24353.282403	0.250025
	GBR (랜덤2)	63.408835	25353.374114	0.219227





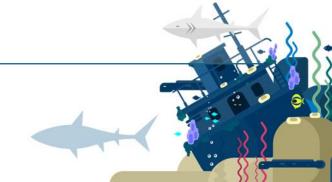


제거한 컬럼 : ID, SHIPMANAGER, BN, ATA\_LT, month, day, hour, minute, ARI\_PO, V\_WIND



### ⟨Data 3⟩

스케일링	모델	MAE	MSE	R2
MinMaxScaler	HGB	17.856958	801.199909	0.576096
	XGB	17.212070	773.719893	0.590635
RobustScaler	HGB	17.871116	802.476394	0.590684
	XGB	17.223823	773.627572	0.590684
StandardScaler	HGB	17.891397	804.434210	0.574385
	XGB	17.215817	774.059990	0.590455





#### 주성분 분석

Standard / 15

**Model\_Name** = GradientBoostingRegressor

Train Score: 0.3959 / Test Score: 0.3900 / Train - Test = 0.00598261232630537 / Train MAE: 24.4758 / Test MAE: 24.5622

<u>Train MSE: 1148.0904 / Test MSE: 1153.0116</u>

**Train R2**: 0.3959 / **Test R2**: 0.3900

**Model\_Name** = HistGradientBoostingRegressor

Train Score: 0.4617 / Test Score: 0.4485 /
Train - Test = 0.01320343119862144 /
Train MAE: 22.2648 / Test MAE: 22.5251
Train MSE: 1023.1217 / Test MSE: 1042.3856

Train R2: 0.4617 / Test R2: 0.4485

**Model\_Name** = XGBRegressor

Train Score: 0.5179 / Test Score: 0.4727 / Train - Test = 0.04515300588040283 / Train MAE: 20.7078 / Test MAE: 21.6243 Train MSE: 916.3650 / Test MSE: 996.6091

**Train R2**: 0.5179 / **Test R2**: 0.4727

Standard / 14

**Model\_Name** = GradientBoostingRegressor

Train Score: 0.5073 / Test Score: 0.5061 / Train - Test = 0.0012198898720925389 / Train MAE: 20.5677 / Test MAE: 20.5552 Train MSE: 936.3825 / Test MSE: 933.4795

**Train R2**: 0.5073 / **Test R2**: 0.5061

**Model\_Name** = HistGradientBoostingRegressor

Train Score: 0.5566 / Test Score: 0.5496 / Train - Test = 0.007011611826404751 / Train MAE: 18.5502 / Test MAE: 18.6724 Train MSE: 842.6463 / Test MSE: 851.2113

**Train R2**: 0.5566 / **Test R2**: 0.5496

**Model\_Name** = XGBRegressor

Train Score: 0.5941 / Test Score: 0.5661 / Train - Test = 0.02801211874677667 / Train MAE: 17.7026 / Test MAE: 18.2793 Train MSE: 771.3974 / Test MSE: 820.0508

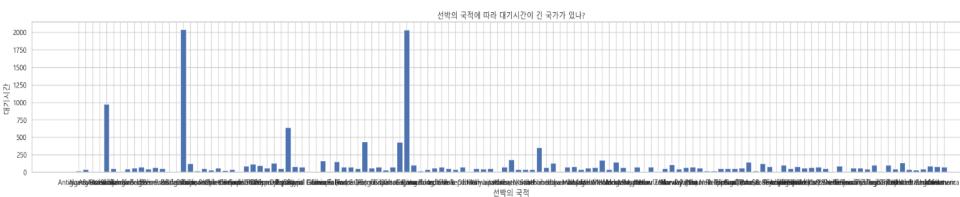
**Train R2**: 0.5941 / **Test R2**: 0.5661



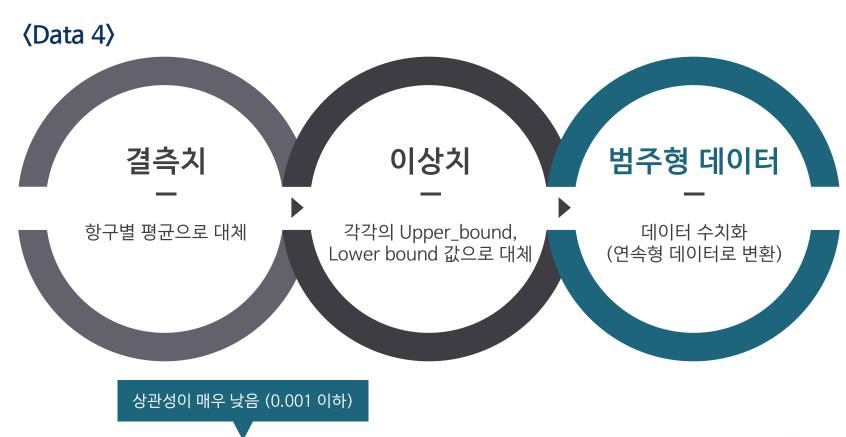
### 범주형 데이터 수치화



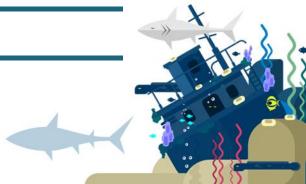
도착항의 항구명







제거한 컬럼: ID, SHIPMANAGER





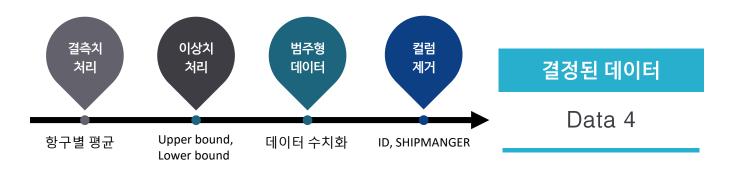
### ⟨Data 4⟩

스케일링	모델	MAE	MSE	R2
MinMaxScaler	LGBM	17.482418	779.105775	0.587786
	XGB	17.109078	757.924144	0.598993
	HGB	17.727757	788.114363	0.583019
StandardScaler	LGBM	17.503978	779.611762	0.587518
	XGB	17.118151	758.266164	0.598812
	HGB	17.661751	785.756841	0.584267
	Stacking	16.684948	754.867434	0.600610
				·» · · · ·

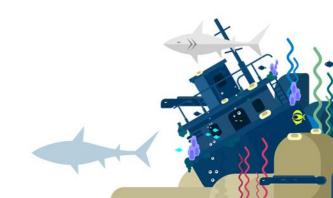


## 모델 해석

### 최종 결정 데이터&모델



스케일링	모델	MAE	MSE	R2
StandardScaler	Stacking	16.684948	754.867434	0.600610





### 목차



- 1) 분석 배경
- 2) 프로젝트 이해
- 1) 데이터 수집
- 2) 데이터 전처리
- 3) 데이터 베이스

- 1) 사용 프로그램
- 2) 머신 러닝
- 3) 모델 해석

- 1) 기대 효과
- 2) 활용 방안





### 기대 효과

#### 1. 대기 시간 최적화

선박의 항구 대기 시간 최소화를 통한 물류 및 운송 효율성 향상 목표

#### 2. 정확한 예측 제공

다양한 요인을 고려하여 신뢰할 수 있는 선박 대기 시간 예측 모델 개발로 항구 운영자 및 물류 관리자에게 정확한 정보 제공

#### 3. 효율적인 자원 할당

선박 도착 및 출발 시간을 더욱 정확하게 예측함으로써 항구 내 자원의 효율적 할당 및 운영 비용 절감

#### 4. 서비스 수준 향상

물류 및 운송 관련 기업 및 단체들에게 대기 시간 예측 서비스를 제공함으로써 고객 만족도 및 서비스 수준 향상

#### 5. 장기적인 이점 확보

정확한 대기 시간 예측을 통해 더욱 효율적인 운송 및 물류 체인을 구축함으로써 장기적인 경쟁력확보 및 기업 가치 향상



### 활용 방안

#### 1. 운영 최적화

- 선박 운항 일정 최적화
- 예측된 대기 시간을 고려하여 출발 시간 조절 및 항로 변경으로 선박 운항에 효율성 극대화

#### 2. 리스크 관리

- 대기 시간 예측을 통해 지연으로 발생하는 비용 감소

#### 3. 자원 할당

- 예측된 대기 시간에 따라 인력, 연료 등의 자원을 적절하게 배분함으로써 비용 절감 및 생산성 향상

#### 4. 고객 서비스 향상

- 예측 결과 고객 서비스에 활용
- 정확한 예측을 통해 고객들에게 출발 및 도착 시각을 더 정확하게 안내함으로써 고객 만족도 향상

#### 5. 정책 개발

- 항로 혼잡도 예측, 이를 기반으로 한 정책 개발로 선박 운영 환경 개선





### 참고 문헌

#### 1. 데이터수집

- DACON
- <a href="https://hwi-doc.tistory.com/entry/%EC%8A%A4%ED%83%9C%ED%82%B9Stacking-%EC%99%84%EB%B2%BD-%EC%A0%95%EB%A6%AC">https://hwi-doc.tistory.com/entry/%EC%8A%A4%ED%83%9C%ED%82%B9Stacking-%EC%99%84%EB%B2%BD-%EC%A0%95%EB%A6%AC</a>

#### 2. 선박의 종합수명 평가

https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=KAR2008030803#;

#### 3. 스테킹

 https://hwi-doc.tistory.com/entry/%EC%8A%A4%ED%83%9C%ED%82%B9Stacking-%EC%99%84%EB%B2%BD-%EC%A0%95%EB%A6%AC





