결측치가 적은 변수들에 대해 중앙값으로 대체하는 이유는 다음과 같습니다:

왜곡 최소화: 중앙값은 이상치(outliers)에 의한 왜곡이 덜하다는 장점이 있습니다. 데이터에 이상치가 많은 경우 평균값은 이러한 이상치의 영향을 크게 받지만, 중앙값은 이상치에 덜 민감하여 데이터의 중앙적 경향을 더 잘 나타냅니다.

대표성: 중앙값은 데이터의 중간 지점에 위치한 값으로, 전체 데이터를 대표하는 경향이 있습니다. 따라서 중앙값으로 결측치를 대체하면 원래 데이터의 전반적인 특성을 유지하는 데 도움이 됩니다.

안정성: 특히 결측치가 적은 경우, 중앙값으로 대체하는 것은 전체 데이터 분포에 큰 영향을 미치지 않으면서도 결측치 문제를 효과적으로 해결할 수 있는 방법입니다.

수치형 데이터에 적합: 수치형 데이터의 경우, 중앙값이나 평균값과 같은 측정치를 사용하여 결측치를 대체하는 것이 일반적입니다. 중앙값은 데이터의 분포가 정규 분포를 따르지 않는 경우에도 적합한 대체값을 제공합니다.

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

CI_HOUR	대기시간

XGBoost 모델의 피처 중요도(feature importance) 평가를 통해 얻을 수 있는 인사이트는 다음과 같습니다:

인사이트:
중요한 특징 식별: 'DIST', 'PORT_SIZE', 'LENGTH', 'SHIP_TYPE_CATEGORY' 등이 항만 내 선박 대기시간 예측에 가장 중요한 특징으로 나타났습니다. 이는 이러한 변수들이 대기시간에 가장 큰 영향을 미칠 수 있는 요소라는 것을 시사합니다.

영향력 있는 변수의 이해:

'DIST' (거리): 선박이 항만에 접근하는 거리가 대기시간 예측에 큰 영향을 미칩니다.
'PORT_SIZE' (항만 크기): 항만의 크기나 용량이 선박 대기시간에 중요한 요소임을 나타냅니다.
'LENGTH' (선박 길이): 선박의 길이가 대기시간에 영향을 미칠 수 있음을 보여줍니다.
'SHIP_TYPE_CATEGORY' (선박 유형): 선박의 유형이 대기시간에 영향을 줄 수 있음을 의미합니다.
전략적 의사결정 지원:

이러한 결과를 바탕으로 항만 운영자나 정책 결정자는 선박 대기시간을 줄이기 위한 효율적인 관리 전략을 수립할 수 있습니다.
예를 들어, 특정 거리에서 오는 선박에 대한 우선 순위 조정, 항만 크기 확장 계획, 다양한 선박 유형에 대한 처리 절차 최적화 등이 있습니다.
피처 중요도의 기준 설정:
피처 중요도의 기준을 정하는 것은 주관적일 수 있으며, 종종 데이터와 모델의 특성에 따라 달라질 수 있습니다. 중요도 기준을 설정하는 몇 가지 방법은 다음과 같습니다:

임계값 설정: 피처 중요도가 특정 임계값(예: 상위 10%, 0.05 등) 이상인 피처만 중요하다고 간주할 수 있습니다.

상위 N개 피처 선택: 중요도가 가장 높은 상위 N개의 피처를 중요한 피처로 간주할 수 있습니다.

데이터 기반 결정: 데이터의 분포나 특성을 고려하여 기준을 설정할 수 있습니다. 예를 들어, 중요도가 눈에 띄게 높은 지점에서 기준을 설정할 수 있습니다.

실용적 접근: 실제 문제 해결에 있어서 중요도가 높은 피처가 실질적으로 얼마나 기여하는지를 고려하여 기준을 정할 수 있습니다.

발표문에서는 이러한 피처 중요도의 분석 결과와 함께, 어떻게 이 인사이트를 통해 항만 내 선박 대기시간을 효과적으로 관리하고 최적화할 수 있는지에 대한 전략을 제시하는 것이 좋을

것입니다.

ANOVA 분석을 통해 얻은 결과를 바탕으로, 항만 내 선박 대기시간 예측과 관련된 주요 인사이트는 다음과 같습니다:

중요한 특징 식별: ANOVA 분석 결과, 'ARI_CO', 'ARI_PO', 'SHIP_TYPE_CATEGORY', 'DIST', 'BREADTH', 'BUILT', 'DEADWEIGHT', 'DRAUGHT', 'ATA_LT', 'PORT_SIZE'가 항만 내 선박 대기시간 예측에 가장 중요한 피처로 나타났습니다. 이는 이러한 변수들이 대기시간에 가장 큰 영향을 미칠 수 있는 요소라는 것을 시사합니다.

영향력 있는 변수의 이해:

'ARI_CO'와 'ARI_PO': 출발 및 도착 항만의 위치가 대기시간에 영향을 미칠 수 있음을 나타냅니다. 특정 지역이나 노선이 대기시간에 영향을 줄 수 있습니다.
'SHIP_TYPE_CATEGORY': 선박의 유형에 따라 대기시간이 달라질 수 있음을 의미합니다. 예를 들어, 컨테이너선과 화물선의 대기시간이 다를 수 있습니다.
'DIST': 선박이 항만에 접근하는 거리가 대기시간에 영향을 줄 수 있습니다.
'BREADTH', 'BUILT', 'DEADWEIGHT', 'DRAUGHT': 선박의 물리적 특성이 대기시간에 영향을 미침을 시사합니다.
전략적 의사결정 지원:

이러한 결과를 바탕으로 항만 운영자는 특정 유형의 선박, 특정 노선, 또는 특정 크기의 선박에 대한 대기시간 관리 전략을 개발할 수 있습니다.
예를 들어, 대기시간이 긴 특정 노선에 대해 우선순위를 조정하거나, 특정 유형의 선박에 대한 처리 절차를 개선할 수 있습니다.
향후 연구 방향 설정:

추가적인 분석을 통해 이러한 변수들이 대기시간에 어떤 방식으로 영향을 미치는지 더 깊이 이해할 수 있습니다.
또한, 다른 잠재적인 영향 요인들에 대한 연구를 통해 더 정확한 예측 모델을 개발할 수 있습니다.
이러한 인사이트는 항만 내 선박 대기시간을 줄이고, 항만 운영의 효율성을 높이는 데 기여할 수 있으며, 발표문에서 이러한 점들을 강조할 수 있습니다.