기상청

기상 데이터를 활용한 지하철 혼잡도 예측 및 활용방안

호선별 **Fold**기반 **XGBoost**를 활용한 지하철 혼잡도 예측 모델 개발

[TEAM 최태정원은]

기획 배경

기획 배경

1. 기상에 따른 교통수단 선택

. 폭우 시 지하철 혼잡도는 약 18.6%, 폭설 시에는 약 15.3% 증가

➜ 기상상황을 고려한 혼잡도 예측 전략 필요

2. 냉난방 민원

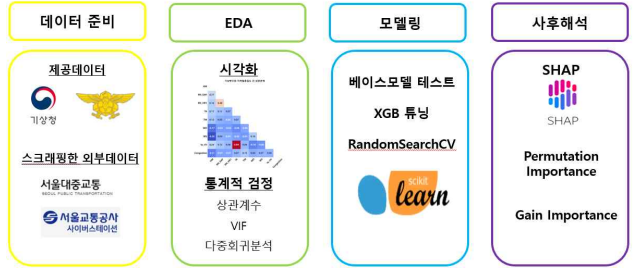
. 냉난방 민원은 2025년 5월까지 지하철 전체 민원의 75%에 달할 정도로 높음

➜ 기상 변수를 반영한 예측을 통해, 지하철 냉난방 조절 시스템에 연동해 냉난방 세기를 선제적으로 조정

분석 목표

혼잡도 예측을 통해 지하철 운영 효율화와 혼잡 완화 방안 마련에 활용 가능한 인사이트 제공

분석 프로세스



데이터 정의

제공 데이터

. 지하철 혼잡도 : 시간별 열차 내 밀집도
. 기상 변수 : 시간별 기상 관련 수치

TM	Line	station_number	station_name	...	SI	ta_chi	Congestion
2021010100	1	150	서울역		-99	-12.6	0
2021010101	1	150	서울역		-99	-9.8	0
...							
2023123123	1	150	서울역		-99	-0.2	21

외부 데이터

[지하철 주소 데이터]

. 수집 대상 : 서울시 1-8호선 지하철 주소
. 서울시 대중교통정보에서 station_name을 기준으로 검색
. 주소를 자동 스크래핑하는 알고리즘 사용
. **총 330건의 역 주소를 수집**

역명	주소
매봉	서울특별시 강남구 도곡2동464-1
새절	서울특별시 은평구 신사2동337-5
...	
진접	경기 남양주시 진접읍 경복대로 244

[환승역 데이터]

. 수집대상 : 환승역 여부 + 환승 노선 개수
. 서울교통공사 지하철 노선도에서
호선, 환승역, 해당 역을 지나는 노선의 개수
3가지를 수기로 수집.

Line	transfer_station	transfer_station
1	인천	2
2	신당	2
...		
8	모란	2

데이터 전처리

지하철 혼잡도 데이터와 기상 데이터의 전처리 및 파생변수 생성

변수명	변수 설명	변수명	변수 설명
year	탑승 기준 연도	month	탑승 기준 월
day	탑승 기준 일	hour	탑승 기준 시간
weekday	탑승 기준 요일	week_of_month	지하철 탑승 월중 주차
week_of_year	지하철 탑승 월중 주차	day_of_year	지하철 탑승 기준 연중 경과일
is_holiday	탑승 기준 휴일 여부	is_day_before_holiday	탑승 기준 휴일 전날 여부
is_day_after_holiday	탑승 기준 휴일 다음날 여부	is_weekend	주말 여부
sin_dow, cos_dow	weekday의 삼각변환 변수	sin_hod, cos_hod	hour의 삼각변환 변수
sin_wom, cos_wom	week_of_month의 삼각변환	sin_dom, cos_dom	day의 삼각변환 변수
sin_doy, cos_doy	day_of_year의 삼각변환	sin_woy, cos_woy	week_of_year의 삼각변환
transfer	환승역 여부	time_period	탑승 기준 시간대 구간
address	주소 변수	신설역.신규관측소	신설역/신규관측소 여부

데이터 전처리

결측치 처리

1-1. 풍속,일강수량,시간 강수량,기온,체감온도,상대습도

- 시간 흐름의 연속성 반영을 위하여 선형 보간법을 적용하여 결측값 보완

1-2. 일사량

- 결측치 비율이 37% 이상으로 변수 제거

시간	호선	역번호	역명	상하구분	AWS지점코드	기온	풍향	풍속	일강수량	시간 강수량	상대습도	일사량	체감온도	혼잡도
2024010100	1	150	서울역	상선	419	0.6	161	2.7	4.5	0	99	-99	-0.3	0
2024010101	1	150	서울역	상선	419	0	146	3.8	0	0	99.4	-99	-2.2	0
2024010102	1	150	서울역	상선	419	0.3	171	3.1	0	0	99.6	-99	-2.3	2.7378
2024010105	1	150	서울역	상선	419	-0.1	176	3	0	0	98.1	-99	-0.1	5.713

날짜 및 범주형 타입 변환

2-1. 시간

- datetime 형식으로 변환

2-2. 호선,역번호,역명,상하구분,AWS 지점코드

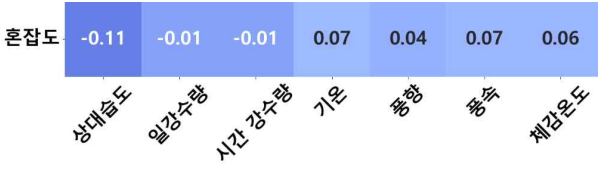
- 라벨인코딩(Label Encoding)을 사용하여 카테고리화

신규 역/관측소 식별

- 2024년 데이터의 신설역 식별
- 구리, 다산, 동구릉, 별내 등 신설역과 신규역번호를 이진변수로 처리하여 자동 식별

기상변수와 혼잡도 간의 상관관계 분석

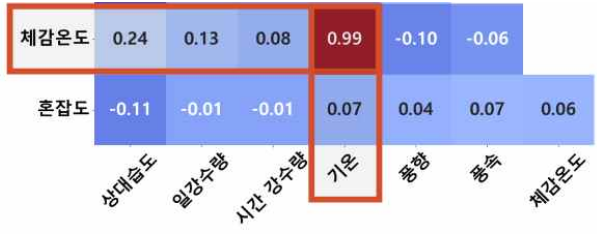
1. 혼잡도에 영향을 미치는 기상변수



- 더울수록, 바람이 불수록, 건조할수록 지하철을 이용하는 경향
- 즉, **더운 날씨에는 냉방이 되는 지하철**을 이용하고, 비가 오거나 **습한 날씨에는 외출을 줄임**

2. 기상변수 간의 높은 상관성 문제

- 다중공선성 높은 변수를 확인 후, 제거하여 모델의 안정성을 확보하고자 함
- 상관관계가 높았던 기온과 체감온도는 역시 VIF(분산팽창지수) 확인 결과, **다중공선성이 심각**



변수	VIF	해석
체감온도	73.17	매우 높음
기온	71.05	매우 높음

03 CORR 다중선형회귀를 통한 기상변수별 혼잡도 영향력 평가

기상청

. 1,614만 건의 표본을 대상으로 다중선형회귀분석을 진행

$$\text{혼잡도} = 24.9 - 0.11 * \text{상대습도} + 0.02 * \text{일강수량} + 0.06 * \text{시간강수량} + 0.00 \text{풍향} + 0.50 * \text{풍속} 0.18 * \text{기온} - 0.02 * \text{체감기온}$$

OLS Regression Results

Dep. Variable:	Congestion	R-squared:	
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	
Date:	Thu, 26 Jun 2025	Prob (F-statistic):	0.00
Time:	13:50:35	Log-Likelihood:	-7.0570e+07
No. Observations:	16143988	AIC:	1.411e+08
Df Residuals:	16143980	BIC:	1.411e+08
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	24.9056	0.023	1089.478	0.000	24.861	24.950
HM	-0.1109	0.000	-404.901	0.000	-0.111	-0.110
RN_DAY	0.0205	0.001	35.661	0.000	0.019	0.022
RN_HR1	0.0611	0.004	14.867	0.000	0.053	0.069
TA	0.1846	0.004	50.498	0.000	0.177	0.192
WD	0.0039	4.77e-05	82.782	0.000	0.004	0.004
WS	0.5000	0.004	120.604	0.000	0.492	0.508
ta_chi	-0.0205	0.003	-6.042	0.000	-0.027	-0.014

Omnibus:	6482932.389	Durbin-Watson:	0.355
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35510442.394
Skew:	1.872	Prob(JB):	0.00
Kurtosis:	9.227	Cond. No.	1.05e+03

1. 기온 vs 체감기온

- 다중공선성으로 인한 왜곡 가능성으로 단순 상관에서는 모두 양의 관계를 보임 하지만, 다중회귀에서는 체감기온이 음의 계수를 가짐
- **두 변수 중 실제 혼잡도에 더 큰 영향을 주는 기온을 남겨 모델의 안정성을 높임**

2. 기상변수 영향력 정리

- 더워지고 바람이 많이 불며 습도가 낮아질수록, 또는 비가 올 때 혼잡도 상승
- 여름철 냉방 수단으로, 그리고 우천 시 보행 대신 대중교통을 선택하려는 이용객의 이동 패턴이 반영된 결과로 해석됨
- 기상변수 단독으로는 모델 설명력이 결정계수 0.023로 매우 낮음
- **혼잡도를 정확하게 예측하기 위해서는 기상 외의 변수들이 필요함**

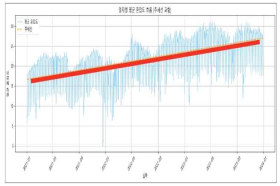
03 CORR 기상 외 변수 혼잡도 간의 상관관계

기상청

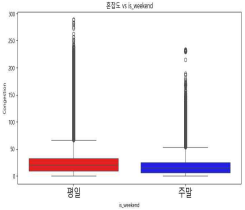
1. 시계열 변수

- 연도, 요일, 시간대의 시계열 변수가 영향을 미침

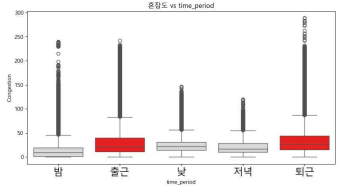
연도별 혼잡도 상승 추세



주말여부 : CORR -0.14

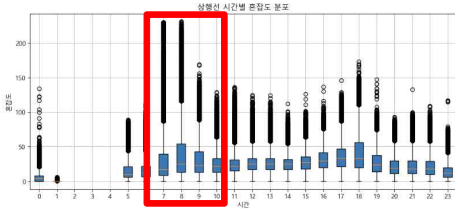


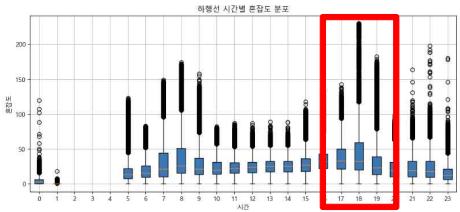
출퇴근시간여부 : CORR 0.30



2. 시간대 + 상행하행

- 상행선은 출근시간대, 하행성은 퇴근시간대에 혼잡함



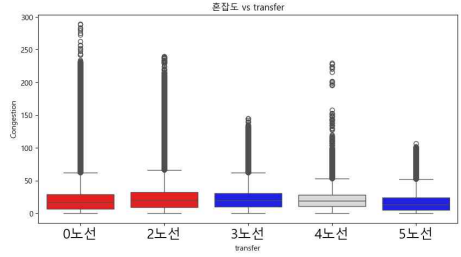


3. 지하철역 위치

- 지하철역의 위치에 따라 평균 혼잡도의 큰 차이

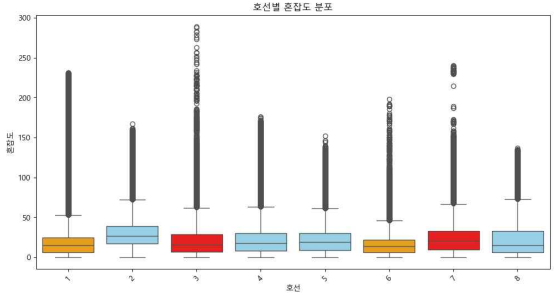
순위	위치	평균 혼잡도
가장 혼잡	관악구	35.338
두 번째로 혼잡	구로구	31.609
두 번째로 여유	인천	11.740
가장 여유	강서구	10.791

- 환승 노선 수에 따라 혼잡도의 분포 차이



1. 호선별 모델 학습의 필요성

호선별로 평균 혼잡도의 차이가 15.9에서 29.7로 폭이 매우 크며, 호선별로 사용 목적이 다른 만큼 **호선별 패턴 학습 필요**



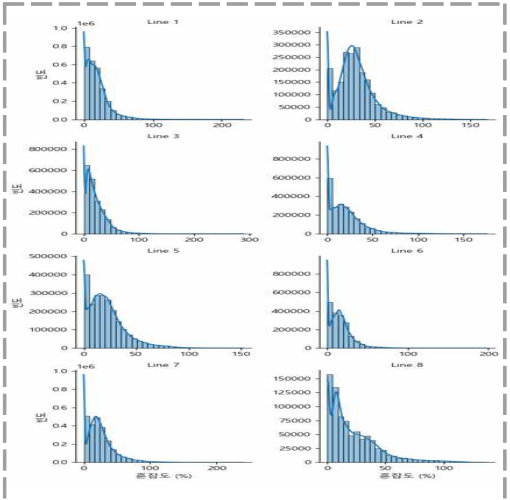
호선	평균 혼잡도
1호선	18.7
2호선	29.7
3호선	20.5
4호선	21.8
5호선	21.6
6호선	15.9
7호선	24.4
8호선	22

2. 모델에 사용된 최종 피처

종류	변수
기상변수	상대습도/ 일강수량/ 시간 강수량/ 기온/ 풍향/ 풍속
지하철 변수	환승역 여부/ 신설역/ 신규관측소/ 역번호/ 주소/ 상행하행/ AWS 지점 코드
주기성 변수	sin_dom , cos_dom , sin_dow , cos_dow , sin_hod , cos_hod sin_wom , cos_wom , sin_woy , cos_woy , sin_doy , cos_doy
시계열 변수	(탑승 기준) 연/ 월/ 일/ 시간/ 시간대 구간/ 연중 경과일/ 휴일여부/ 다음날 휴일 여부/ 휴일 여부/ 주말 여부/ 월중 주차/ 연중 주차/ 요일

04 MODEL 분석 방법론

기상청



호선마다 상이한 이용 패턴과 특성이 존재하므로 **각 호선별로 개별 모델 구축**

네 개의 모델을 적용하여 RMSE를 기준으로 비교한 결과

Line	ARD	LGBM	CatBoost	XGB
1	15.08	7.87	5.47	6.85
2	15.61	10.43	11.66	7.88
3	13.59	5.82	6.42	5.66
4	15.09	7.91	5.38	6.82
5	13.58	7.16	4.99	6.08
6	11.45	6.8	7.71	5.28
7	17.96	9.99	6.82	8.44
8	15.1	6.82	6	5.88
Average	14.68	7.85	6.8	6.61

평균 RMSE 6.61로

XGBRegressor가 가장 우수한 성능 보임

모델 선정 이유

- 수치형·범주형·시계열 파생 변수가 혼합된 복잡한 구조의 데이터를 효과적으로 처리
- 변수 간의 비선형 관계와 상호작용을 잘 반영하는 트리 기반 모델
- 정규화 항과 조기 종료를 통해 과적합 방지 및 결측치 자동 처리 가능

기상청

하이퍼파라미터 튜닝

1. 단계적 탐색 전략 적용

. XGBoost 모델의 주요 하이퍼파라미터에 대해 효율적인 탐색을 위해 **RandomSearchCV** 적용

. 과적합 방지 및 탐색 시간 효율을 위해 10개의 파라미터 조합을 샘플링하여 탐색

2. 검증방법 선택

2-1.시계열 데이터 분할

. **TimeSeriesSplit(n_split=3)**을 검증 분할 전략으로 사용

2-2.평가기준

. 회귀 모델 성능 지표로 **RMSE** 채택

3. 튜닝 결과 및 효과

. 최적 파라미터 조합 도출

no	Hyperparameters	units
1	n_estimators	2000
2	learning_rate	0.05
3	max_depth	12
4	subsample	0.9
5	colsample_bytree	0.8
6	reg_alpha	0.3
7	reg_lambda	0.8
8	min_child_weight	3
9	gamma	0

. **예측 성능 및 일반화 향상**

기상청

전체 데이터로 모델 학습

1. 전체 데이터 학습 전략

- 2021~2023년 혼잡도는 연도별로 꾸준히 증가하는 추세
- 2021~2023년 데이터를 8:2로 분할하여 훈련 및 검증 수행
- 2024년은 독립적인 테스트셋으로 활용해 성능 평가
- 최종 모델은 **2021~2023년 전체 데이터를 학습에 활용**

2. 종합 성과 요약

- 전반적으로 호선별 예측 성능이 향상, 특히 **8호선이 가장 우수**
- **최종 RMSE 5.231**, 6월 27일 기준 **전체 8위** 달성

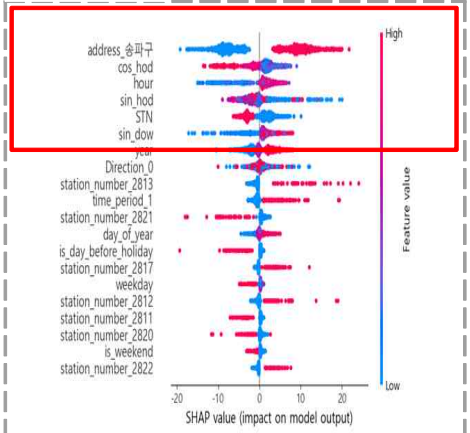
3. 호선별 성능 비교

- 호선별 예측 성능 확인 결과, 8호선이 가장 우수
- **8호선을 기준으로 모델 사후해석**

Line	RMSE	결정계수
1	0.94	0.99
2	0.90	0.99
3	0.74	0.99
4	0.88	0.99
5	0.83	0.99
6	0.63	0.99
7	1.01	0.99
8	0.53	0.99

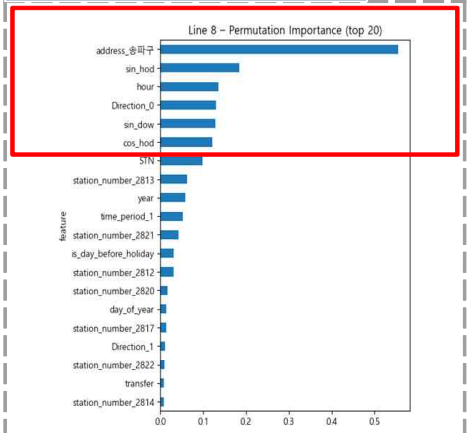
기상청

SHAP 분석



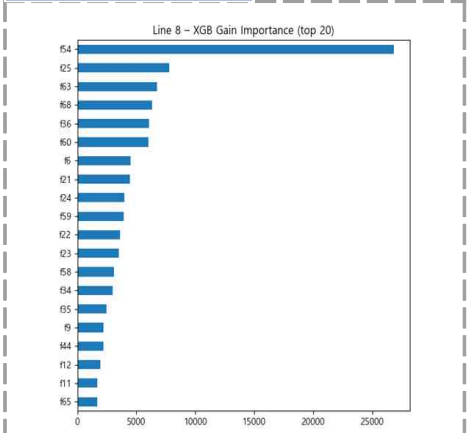
주소, 시간대 주기성, 연도에 따른 혼잡도 상승이 예측에 주요 결정 요인으로 작용

Permutation Importance



주소, 시간 변수가 높은 중요도를 보이며, 제거 시 성능이 크게 저하됨

Gain Importance



트리 모델 내부에서 주요 분할점으로 사용된 변수들이 높은 정보 이득값 기록

시간대와 위치 정보는 혼잡도 예측에 **핵심변수**로 작용한 반면,

기상 변수는 핵심 변수에 비해 낮은 영향력을 나타냈음

활용방안 및 기대효과

교통 시스템 개선

혼잡 구간·시간 식별
인프라 투자 근거 제공

운영 효율화

기상 변수 반영한
냉난방 조절 시스템 연동

승객 편의성 증진

채널을 통해
혼잡 예측 정보 사전 제공으로
분산 유도

민원 감소

쾌적한 탑승 환경으로
냉난방 관련 **민원 감소**

감사합니다.