호선별 Fold 기반 XGBoost 를 활용한 지하철 혼잡도 예측 모델 개발

접수번호	250206	팀명	최태정원은	최종성능	5.231

1. 분석배경 및 목표

1.1 분석 배경

기상 데이터는 강수량, 일조량, 강설량 등 다양한 기상 요소를 포함하며, 이러한 변수들은 교통 상황에 밀접한 상관관계를 가지고 있다. 특히, 기상 변화는 대중교통 이용자의 수요에 직접적인 영향을 미치는 중요한 변수로 작용한다. 2021 년 대한교통학회에 게재된『날씨가 대중교통 이용자의 지하철 또는 버스 선택에 미치는 영향: 스마트카드 데이터를 사용하여』에 따르면, 폭우 시 지하철 혼잡도는 약 18.6%, 폭설 시에는 약 15.3% 증가하는 것으로 나타났다. 이러한 선행 연구 결과는 기상요소가 교통 수단 선택 및 혼잡도 변화에 실질적인 영향을 미친다는 점을 뒷받침한다. 이에 본 공모안은 기상 정보와 지하철 혼잡도 간의 상관 관계를 분석하고 이를 바탕으로 기상 정보를 활용한 지하철 혼잡도 예측 모델을 제안하고자 한다.

1.2 분석 목적

본 연구의 목적은 2021 년부터 2023 년까지 수집된 기상 관측치와 지하철 혼잡도 데이터를 활용하여 두 변수 간의 상관관계를 규명하고, 이를 바탕으로 예측 모델을 개발하는 데 있다. 기상 변수와 혼잡도 간의 정량적 관계를 분석함으로써, 경험적 대응에 따른 한계를 극복하고 과학적 근거에 기반한 혼잡 예측 시스템을 제시하고자 한다.

분석 범위는 서울 지하철 8 개의 노선을 대상으로 하며, 각 노선별 역에서 1 시간 단위로 수집된 혼잡도(%) 및 기상 데이터를 포함한다. 학습용 데이터는 2021 년 1 월부터 2023 년 12 월까지, 검증용 데이터는 2024 년 전체 기간으로 설정하였다. 또한 환승역 여부, 시간대, 요일 등의 시계열 특성을 함께 고려해 모델 학습에 활용함으로써 예측 정확도를 높이고자 한다.

최종 목표는 다음과 같다. 첫째, 기상변수와 호선별 혼잡도 간의 상관관계를 정량적으로 도출하여 주요 영향 인자를 식별한다. 둘째, 시간, 기상, 역별 데이터를 기반으로 효과적인 피처 엔지니어링을 수행하여 모델 입력을 최적화한다. 셋째, XGBoost 기반의 호선별 혼잡도 예측 모델을 학습, 검증하고, 검증 결과를 통해 모델의 실효성을 평가한다. 넷째, 예측 결과를 활용한 증편, 알림 서비스, 스케줄 조정 등 구체적 운영 방안을 제시하여 지하철 운영 효율성을 제고하고 승객 편의를 증진하는 정책적 제언을 제시하는 것이다.

1.3 분석 프로세스

2. 분석 데이터 및 전처리

2.1. 데이터 변수 설명

2.1.1 기존 변수

본 공모전에서 제공된 지하철 혼잡도 데이터의 변수 목록은 Table 1 과 같다.

Table 1: 지하철 혼잡도 데이터 변수 목록

순번	변수명	변수설명
1	Direction	지하철 상행 하행 구분
2	station_number	역번호
3	HM	상대습도
4	RN_DAY	일강수량
5	RN_HR1	시간 강수량
6	TA	정시 기온
7	WD	정시 10분 평균 풍향
8	WS	정시 10분 평균 풍속
9	STN	AWS 지점 코드
10	Line	지하철 호선
11	Congestion	지하철 혼잡도

2.1.2. 파생 변수 생성

지하철 혼잡도는 출근 시간대(07:00–09:00)와 퇴근 시간대(17:00–19:00)에 가장 높게 나타났으며, 평일이 주말·공휴일보다 전반적으로 더 큰 혼잡도를 보였다. 이러한 시간대별·요일별 차이를 반영하기 위해, 원천 변수인 'TM'에서 기본 시계열 변수(year, month, day, day_of_year, hour, weekday, week_of_month, week_of_year)를 추출하였다. 데이터의 순환적 패턴을 효과적으로 모델링하기 위해 삼각변환 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)를 생성하여 주기성을 반영하였다.

또한 공휴일 전날 (is_day_before_holiday), 공휴일 당일(is_holiday), 공휴일 다음날(is_day_after_holiday), 주말(is_weekend) 여부를 이진 변수로 추가하고, 출퇴근 시간대를 묶은 범주형 변수(time_period)를 포함하였다. 본 항목에서 추가한 파생변수의 목록은 Table 2 와 같다.

Table 2: 파생변수 목록

순번	범주	변수명
1	기본 시계열 변수	day, day_of_year hour, month, year week_of_month week_of_year, weekday
2	삼각변환 변수	– 일: sin_dom, cos_dom – 요일: sin_dow, cos_dow – 시간대: sin_hod, cos_hod – 월중 주차: sin_wom, cos_wom – 년중 주차: sin_woy, cos_woy – 연중 일차: sin_doy, cos_doy
3	휴일/주말 변수	is_day_before_holiday is_day_after_holiday is_holiday, is_weekend
4	시간대 변수	time_period
5	환승역 여부 변수	transfer
6	주소 변수	address
7	신설역/신규관측소	신설역.신규관측소

*** 자세한 파생변수 설명은 부록 A 참조**

2.2. 외부 데이터 수집

2.2.1 지하철 주소 스크래핑

서울시 대중교통정보에서 station_name 을 기준으로 검색하여 주소를 자동 스크래핑하는 알고리즘을 사용해 총 330 건의 역 주소를 수집하여 address 라는 변수를 생성하였다. 검색이 되지 않는 역(성수 E, 자양(뚝섬한강공원), 응암 S, 당고개, 능길, 별내별가람, 서울역, 오남, 신촌(지하), 진접)은 네이버 플레이스에서 수기로 검색하여 추가하였다.

2.2.2 환승역 데이터

환승역일수록 일반역보다 혼잡도가 높을 것으로 추정하여 서울교통공사 지하철 노선도에서 호선, 환승역, 해당 역을 지나는 노선의 개수 3 가지를 수집하고, 해당 역을 지나는 노선의 개수를 새로운 컬럼(transfer) 으로 추가하였다.

2.3. 데이터 전처리

2.3.1. 기본 전처리

역별 지역 정보를 단순화하기 위해 address 컬럼의 값을 '인천' 또는 '경기'로 변경하였다. 공지사항에 따라 4 호선의 한대앞·중앙·고잔·초지·안산·신길온천·정왕·오이도역과 8 호선의 남위례역을 기간에 맞게 인스턴스 제거하였다. station_name 에 따른 혼잡도를 시각화하였을 때, 앞선 역들을 포함한 개봉역은 다른 연도에 비해 2021 년이 극심하게 낮았기 때문에 2022 년 12 월 31 일까지의 데이터를 이상치로 판단하여 해당 기간의 인스턴스를 제거하였다. 또한 station_name 통일을 위해 당고개를 불암산으로, 자양(뚝섬한강공원)을 자양으로, 신촌(지하)을 신촌으로 각각 변경하였다.

2.3.2. 날짜, 범주형 타입 변환

TM 컬럼을 연도-월-일 형식으로 변환하고 Line, station_number, STN, station_name, Direction 은 범주형 컬럼으로 카테고리로 변환하였다.

2.3.3. 결측치 처리 및 선형 보간

WD 은 풍향으로 0~360 도의 값을 가지기 때문에 0 도 미만의 값을 결측치 처리하였고, WS, RN_DAY, RN_HR1, TA, ta_chi, SI, HM 은 -99 인 값을 np.nan 값으로 대체했다. 그 결과 각 컬럼의 결측 비율은 다음과 같다.

Table 3: 컬럼별 결측 비율

컬럼명	결측치비율(%)	연속된 결측 최대 구간 길이
SI	37.046	31
ta_chi	0.002	1
HM	5.159	5209
RN_HR1	2.204	4188
RN_DAY	2.147	4188
WS	1.409	4188
WD	1.754	4188
TA	1.322	4188

SI 는 일사량으로 결측치가 37%로 30% 이상의 결측치를 가져 어떠한 결측치 대체 방법을 쓴다면 데이터에 편향이 발생할 것으로 판단하였다. 이진변수로 사용하는 방법도 고려하였지만, 일사량은 TA, hour, month 로 설명이 가능한 변수라고 판단하여 컬럼을 제거하였다.

이외의 결측치는 제공된 데이터는 기상데이터로 전후 인스턴스와 비슷한 수치를 가지는 특성이 있다. 각 컬럼에서 결측치가 연속되는 최대 구간의 길이를 구하였을 때, 5209 가 가장 높은 수치였는데 이는 실질적으로 하루 1 만 5 천 건의 30%에 해당하는 데이터다. 기상변수 특성상 시간순으로 나열하였을 때, 시간 흐름의 연속성을 반영하기 위해 선형 보간법(linear interpolation)을 적용하여 결측값을 보완하였다.

2.3.4. 신규 역/관측소 식별

검증용 2024 년 데이터에는 학습 기간(2021~2023)에는 존재하지 않았던 신설역이 포함되어 있었다. 신규 관측소는 역명 기준으로 구리, 다산, 동구릉, 별내, 암사역사공원, 장자호수공원이며, 역변호 기준으로는 2805, 2806, 2807, 2808, 2809, 2810 에 해당한다.

데이터 전처리 과정에서 학습용 데이터(2021-2023 년)로부터 추출한 고유 역번호를 기반으로, 검증용 데이터(2024 년)에 새롭게 추가된 역과 역번호를 이진변수로 처리하였다. 학습 단계에서는 모든 행에 대해 신설역과 신규역번호 컬럼을 0 으로 초기화하였고, 검증 단계에서는 1 로 설정함으로써 학습 데이터에는 없었던 역과 역번호를 자동으로 식별할 수 있도록 하였다.

3. 지하철 혼잡도와 변수들 간의 상관분석

3.1 기상변수

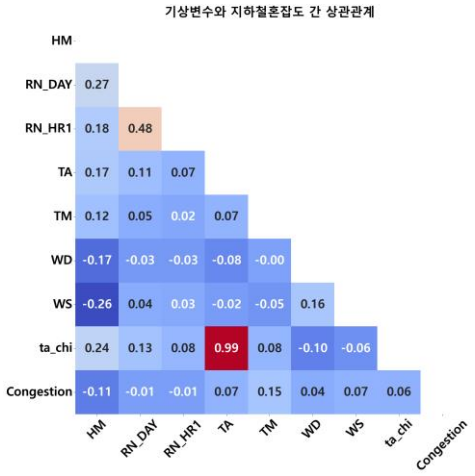


Figure 1 : 기상변수와 지하철혼잡도 간 상관관계

Congestion 과 주요 기상변수 간 피어슨 상관계수를 계산한 결과, HM 이 –0.11 로 가장 강한 음(–) 상관을 보였다. RN_DAY, RN_HR1 의

상관계수는 각각 –0.01, –0.01 로 거의 0 에 가까워 강수 관련 지표가 혼잡도에 미치는 영향은 미미한 것으로 나타났다. WS 와 ta_chi, TA 는 각각 +0.04, +0.06, +0.07 로 혼잡도와 양(+) 상관을 보여, 바람이 불고, 더울수록 지하철을 이용하는 경향이 다소 커짐을 시사한다.

3.1.1. 기상변수 간의 다중공선성

상관계수가 0.99 에 달하는 TA 와 ta_chi 는 VIF 지수 역시 매우 높아 다중공선성이 심각한 것으로 판단하여 두 변수 중 하나만 선택하고자 하였다.

Table 4 : 기상변수 간의 VIF 점수

변수	VIF	해석
HM	1.48	문제 없음
RN_DAY	1.37	문제 없음
RN_HR1	1.31	문제 없음
WS	1.15	문제 없음
WD	1.05	문제 없음
TA	71.05	매우 높은 다중공선성
ta_chi	73.17	매우 높은 다중공선성

3.1.2. 다중회귀분석으로 확인한 회귀계수

통계적 방향성을 보다 명확히 파악하기 위해 1,614 만 건의 표본을 대상으로 다중선형회귀분석을 수행했다. 표본 수가 매우 크므로 p-값 대신 계수의 부호와 크기에 초점을 맞췄다. 그 결과, HM 은 한 단위 증가할 때마다 약 0.11 만큼 혼잡도가 감소했으며, RN_DAY, RN_HR1, WD 은 모두 양(+)의 계수를 보여 비가 오거나 특정 풍향일 때 혼잡도가 증가하는 경향을 보였다. TA 는 1℃ 상승당 약 0.18 만큼, WS 는 1 m/s 증가당 약 0.50 만큼 혼잡도를 높였으나, ta_chi 는 –0.02 로 기온과 반대 방향성을 나타냈다. 상관계수 0.99 를 기록한 TA 와 ta_chi 간의 다중공선성 문제를 해결하기 위해, 최종 모델에는 TA 만을 사용하였다.

이를 종합하면, 날이 더워지고 바람이 많이 불며 습도가 낮아질수록, 또는 비가 올 때 지하철 혼잡도가 더 높아지는 경향이 있다. 이는 여름철 냉방 수단으로, 그리고 우천 시 보행 대신 대중교통을 선택하려는 이용객의 이동 패턴이 반영된 결과로 해석할 수 있다.

3.2. 기상 변수 외 혼잡도 간의 상관관계 분석

기본 시계열 변수와 Congestion 간의 상관관계를 분석한 결과, hour 이 0.21 로 가장 높은 양의 상관관계를 보였으며, 그 뒤를 이어 year 는 0.14, weekday 은 -0.11 로 나타났다. 이는 시간대가 지하철 혼잡도에 유의미한 영향을 미칠 수 있음을 시사한다.

삼각함수로 변환한 주기성 변수의 경우, cos_hod 이 -0.25 로 가장 높은 음의 상관성을 보였고, sin_hod 가 -0.12, sin_dow 가 0.11, cos_dow 가 -0.08 의 상관관계를 보였다. 이러한 결과는 특정 시간이나 요일의 주기성이 혼잡도에 일정한 패턴으로 영향을 미침을 의미한다.

휴일/주말 여부를 나타내는 변수들에서는, is_weekend 가 -0.14, is_holiday, is_day_before_holiday, is_day_after_holiday 모두 -0.1 수준의 음의 상관관계를 보여, 전반적으로 휴일과 주말이 혼잡도 감소와 연관이 있음을 알 수 있었다.

한편, 시간대를 범주화한 time_period 변수는 혼잡도와 0.30 의 상관계수를 나타내며, 전체 변수 중 가장 강한 양의 상관성을 보였다. 이는 시간대 구분이 혼잡도 예측에 있어 중요한 설명 변수가 될 수 있음을 설명한다.

4. 분석결과 및 검증

4.1. 혼잡도 분포

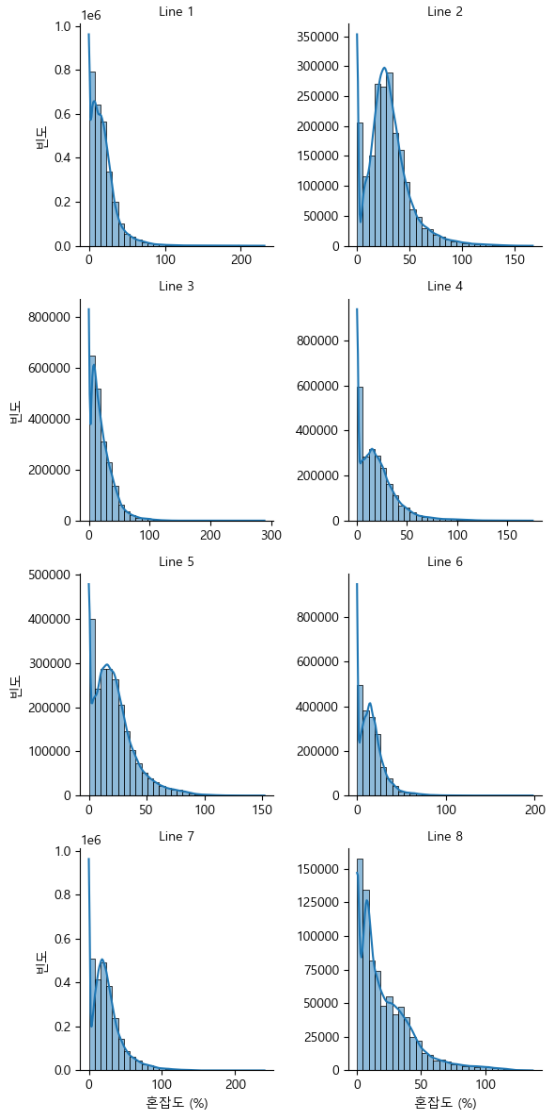


Figure 2: 호선별 혼잡도 분포

Figure 2: 혼잡도가 호선별로 차이가 심해 모델이 호선별 특성을 반영하도록 분석하였다.

4.2. 분석 방법

학습 데이터는 최종 예측 대상인 변수 Congestion 을 타깃으로 설정하고, 나머지 모든 관측 항목과 부록에서 정의된 파생변수를 피처로 활용하였다. 원본으로 제공된 기상 관측치와 역별, 시간별 기본 속성 외에도, 주기성 변수 등 약 18 종의 파생변수를 추가로 생성하여 모델 입력의 표현력을 높였다.

피처 전처리 과정에서는 먼저 기상 변수의 결측치 및 이상치를 선형보간 방식으로 처리하였으며 범주형 변수는 One-Hot Encoding 을 적용하였다. 수치형 입력 변수는 MinMaxScaler 를 적용하여 각 변수의 상대적 대소 관계를 유지하면서, 스케일 차이로 인한 학습 편중을 방지하였다.

데이터 분할은 표준편차를 일반화한 척도로서 실제 값과 추정 값과의 차이를 나타내는 평가 지표인 Root Mean Squared Error(RMSE)를 안정적으로 평가 하기 위해 호선별로 시간순으로 정렬한 뒤 앞 80%를 학습 세트, 뒤 20%를 검증 세트로 구성하였다. 이때 단순 랜덤 분할을 지양하고 시계열 순서를 유지함으로써 시간적 누수를 막고, 실제 예측 환경과 유사한 평가를 수행하였다. 교차검증이 아닌 단일 분할 방식을 채택했으나, 후속 하이퍼파라미터 튜닝 시 TimeSeriesSplit 기반 RandomSearchCV 를 병행하여 모델의 일반화 성능을 확보하였다.

최종적으로 준비된 X_train, X_val, y_train, y_val 데이터를 이용해 XGBoost 모델을 학습, 검증함으로써, RMSE 기준에서 성능을 점진적으로 개선해 나갔다.

4.2.1. 베이스라인 모델 테스트

베이스라인 모델로 변수 간의 다중공선성에 강한 ARDRegression, 대용량·고차원 데이터에 강한 XGBRegressor, 대용량 데이터에서도 학습 속도·메모리 효율이 우수한 LGBMRegressor, 범주형 변수를 포함한 혼합형 데이터에서 안정적인 CatBoostRegressor 를 선정하여 성능을 테스트하였다. 테스트한 성능은 Table 5 의 값과 같다.

Table 5 : 베이스라인 모델 성능 비교표

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

성능 비교 결과 평균 RMSE 가 6.61 로 가장 우수한 성능을 보인 XGBRegressor 모델을 선정하고, 하이퍼파라미터 튜닝을 진행하였다.

4.2.2. XGBoost

기상 및 시간 정보를 바탕으로 지하철 혼잡도를 예측하기 위해 XGBoost(eXtreme Gradient Boosting) 모델을 활용하였다. XGBoost 는 수치형·범주형·시계열 파생 변수가 혼합된 복잡한 구조의 데이터를 효과적으로 처리하며, 변수 간의 비선형 관계와 상호작용을 잘 반영하는 트리 기반 모델이다. 정규화 항과 조기 종료를 통해 과적합을 방지하고, 결측치 자동 처리 및 병렬 학습 구조로 빠르고 안정적인 학습이 가능하다. 히스토그램 기반 트리 분할(tree_method='hist')은 대규모 시계열 데이터에 효율적으로 작동하며, 변수 중요도 도출을 통해 주요 영향 요인 해석도 가능하다.

4.2.3. 하이퍼 파라미터 튜닝

혼잡도 예측 모델의 일반화 능력을 극대화 하기 위해, XGBoost 의 주요 하이퍼파라미터(n_estimator, learnin_rate, max_depth, ...)에 대해 단계적 탐색 전략을 적용하였다. 전역 탐색을 위해 RandomSearchCV 를 사용하여 10 번의 임의 조합을 평가하였다. 모든 탐색에서 시계열 데이터의 순서 보존을 위해 TimeSeriesSplit(n_split=3)을 검증 분할 전략으로 사용하였으며, 평가지표는 RMSE 를 채택하였다. 최종적으로 제출한 모델의 하이퍼파라미터는 다음과 같다.

Table 6 : 모델 하이퍼파라미터

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

4.2.4. 전체데이터로 모델 학습 결과

연도별 혼잡도(Congestion)를 분석한 결과, 2021 년부터 2023 년까지 혼잡도가 꾸준히 증가하는 추세를 보였다. 2021~2023 년 데이터를 8:2 비율로 훈련 및 검증 세트로 분할하고, 2024 년 데이터를 독립적인 테스트 데이터로 활용하여 모델 성능을 평가하였다.

최종적으로는 앞에서 최고 성능을 보인 모델과 동일한 구조로 2021~2023 년 전체 데이터를 모델 훈련에 활용하고 2024 년 데이터를 테스트용으로만 사용하였다. 동일한 구조에 모델 훈련 데이터를 확장 적용한 결과, 전반적으로 호선별 예측 성능이 향상되었으며, 특히 8 호선 모델이 가장 우수한 성능을 나타냈다. 최종적으로 RMSE 5.231 을 기록하며 6 월 27 일 기준으로 전체 8 위를 기록했다.

Table 7 : 전체 데이터 모델 RMSE

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

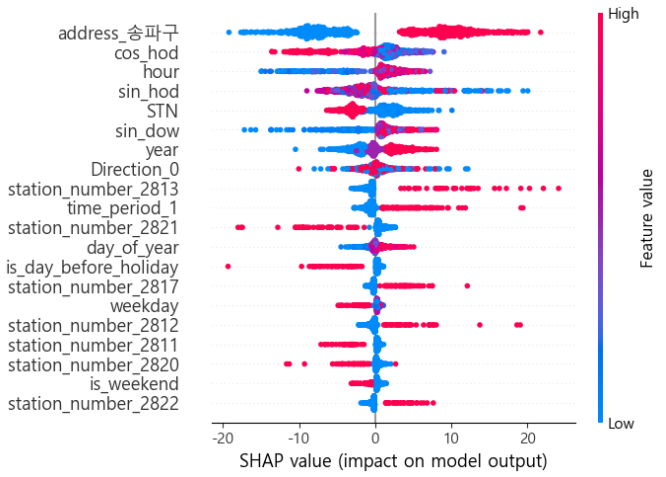
5. 모델 사후 해석

모델 학습 단계에서 최고 성능을 보인 호선은 RMSE 0.531, 결정계수 0.99 로 8 호선이었다. 따라서, 8 호선의 모델을 기준으로 사후해석을 진행했다.

개별 예측값에 대해 특성이 기여한 정도/방해한 정도를 정량적으로 계산해주는 SHAP 기법, 학습된 모델에서 특정 피처 값을 무작위로 섞었을 때, 성능 저하가 얼마나 발생하는지로 중요도를 측정하는 Permutation Importance, 피처가 트리의 분할 노드에서 사용될 때, 획득한 정보량을 누적해서 측정하는 Gain Importance 를 활용하였다.

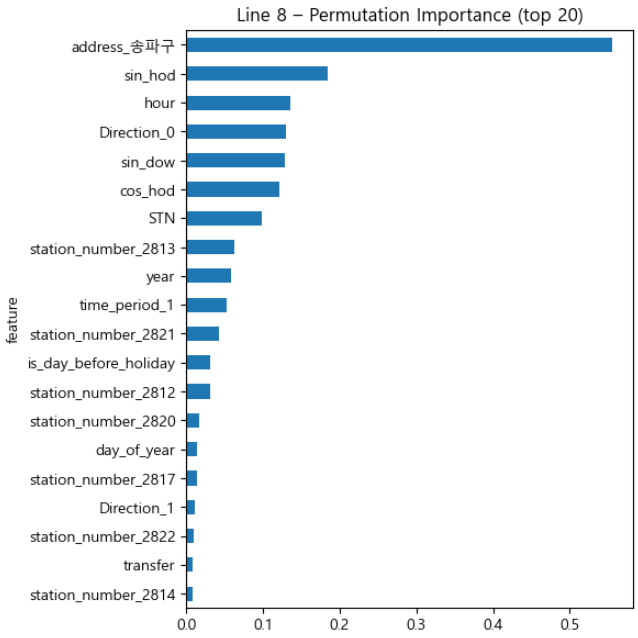
5.1. SHAP

SHAP 분석 결과, adddress_송파구, cos_hod, hour, sin_hod, STN, sin_dow, year, Direction_0 등의 변수에 민감하게 반응하였다. 역이 위치한 주소, 하루 중 시간대에 따른 주기성 패턴, 그리고 연도에 따른 혼잡도 상승이 예측에 주요한 결정 요인으로 작용했다. 또한, EDA 로 확인하였던 상행선은 출근시간대에 Congestion 이 높고, 하행선은 퇴근시간에 혼잡도가 높은 사실이 SHAP 와 일치한다.



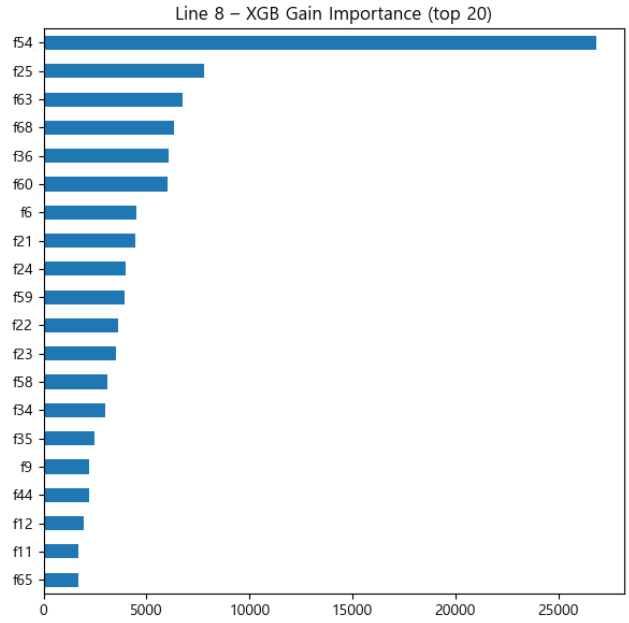
5.2. Permutation Importance

Permutation Importance 분석 결과 역시 유사한 양상을 보였다. address_송파구, sin_hod, hour, Direction_0 등의 변수가 높은 중요도를 나타냈으며, 이는 해당 변수들이 모델 예측에서 제거되었을 때 성능이 크게 저하됨을 의미한다. 모델이 시간적 패턴과 공간적 위치 정보를 효과적으로 학습하고 있음을 뒷받침한다.



5.3. Gain Importance

Gain Importance 에서는 피처명이 자동 인코딩되어 있어, 직접적인 변수명 확인은 어려웠으나, f54, f25, f63 등의 변수가 높은 정조 이득(Gain)값을 가졌다. 트리 계열인 XGboost 모델 내부에서 반복적으로 주요 분할점으로 사용되었음을 의미하며, SHAP, **Permutation Importancn** 의 결과를 뒷받침한다.



또한, 기상 변수와 혼잡도(Congestion) 간의 상관분석 결과에서 확인된 바와 같이, 기상 변수는 모델 예측에서 상대적으로 낮은 중요도를 보였다. 이는 기상 변수 단독으로는 혼잡도에 주요한 영향을 미치지 않으며, 예측 변수 중 후순위에 해당함을 사후 해석 결과를 통해 다시 한번 확인할 수 있었다.

6. 활용방안 및 기대효과

6.1 기상 기반 지하철 혼잡도 예측을 통한 운영 효율화 방안

서울 지하철은 전 연령층이 이용하는 교통수단이며, 객실 내 동일한 온도라도 승객 수나 습도 등 기상 조건에 따라 체감 온도는 크게 달라진다. 이러한 차이는 냉·난방 관련 민원으로 이어지며, 2025 년 5 월까지 전체 민원의 75%를 차지할 만큼 비중이 높다. 특히 승객이 집중되는 2 호선에서 냉방 민원이 다수 발생하였다.

서울교통공사와 SK 텔레콤은 2021 년에 교통카드 데이터와 이동통신 빅데이터를 결합해 칸별 실시간 혼잡도를 산출하는 모델을 시범 운영한 결과, 2 호선 서울대입구역에서 교대역 구간의 시간대별 혼잡도 표준편차가 최대 30% 감소했다. 혼잡도 정보만으로도 승객 분산 효과가 확인되지만 기상 변수와 체감온도는 아직 혼잡도 예측에 반영되지 않는다.

본 예측 모델은 기상 변수를 반영한 예측을 통해, 지하철 냉난방 조절시스템에 연동해 냉·난방 세기를 선제적으로 조정할 수 있다. 전광판 및 모바일 앱 등 채널을 통해 승객에게 혼잡 예측 정보를 사전 제공할 수 있으며, 이를 통해 승객이 비교적 여유 있는 시간대를 선택하도록 유도할 수 있다. 이러한 활용은 쾌적한 탑승 환경 조성과 민원 감소에 기여할 수 있다.

6.2 교통 시스템 개선

기상 기반 혼잡도 예측 데이터를 장기적으로 축적하면, 기후 조건에 따라 혼잡이 반복적으로 발생하는 시간대나 구간을 식별할 수 있다. 이는 역별 플랫폼 확장, 열차 증편, 특정 시간대 집중 대응 등 인프라 투자와 도시 교통 정책 수립의 정량적 근거로 활용될 수 있다.

또한, 공공 API 를 개방하면, 공유 모빌리티 사업자가 경로를 자동 재계산해 환승 수요를 분산시킬 수 있으며, 보다 효율적인 교통 시스템 구축에 이바지할 수 있다.

[참고문헌]

이승연, 이영인. (2021-11-10). 날씨가 대중교통 이용자의 지하철 또는 버스 선택에 미치는 영향 :스마트카드 데이터를 사용하여. 대한교통학회 학술대회지, 제주.

[부록]

A. 파생변수 목록

순번	변수명	변수 설명
1	year	탑승 기준 연도
2	month	탑승 기준 월
3	day	탑승 기준 일
4	hour	탑승 기준 시간
5	weekday	탑승 기준 요일
6	week of month	지하철 탑승 월중 주차
7	week_of_year	지하철 탑승 연중 주차
8	day of year	지하철 탑승 기준 연중 경과일
9	is_holiday	탑승 기준 휴일 여부
10	is_day_before_holiday	탑승 기준 휴일 전날 여부
11	is_day_after_holiday	탑승 기준 휴일 다음날 여부
12	sin_hod,cos_hod	hour의 삼각변환 변수
13	sin_dow,cos_dow	weekday의 삼각변환 변수
14	sin_dom,cos_dom	day의 삼각변환 변수
15	sin_wom,cos_wom	week_of_month의 삼각변환 변수
16	sin_woy,cos_woy	week_of_year의 삼각변환 변수
17	sin_doy,cos_doy	day_of_year의 삼각변환 변수
18	time_period	탑승 기준 시간대 구간 변수 (밤,출근,낮,저녁,퇴근)

B. 외부데이터

서울시 대중교통정보(https://bus.go.kr/)

서울교통공사 지하철 노선도

(http://www.seoulmetro.co.kr/kr/cyberStation.do))