기상청

주제3. 기상과 지하철 혼잡도 상관분석·예측

기상 데이터 기반 XGBoost를 활용한 지하철 혼잡도 예측 모델

접수번호	250549
팀명	지드래곤보다용드래곤
팀원	김필준 구지민 권지민 김우용

목차

1. 서론

1.1. 공모 배경 및 분석 주제

본 분석은 기상 빅데이터와 지하철 혼잡도 데이터를 융합하여 기상 조건이 지하철 혼잡도에 미치는 영향을 정량적으로 분석하고, 이를 바탕으로 혼잡도를 예측하는 모델을 구축하는 데 목적이 있다. 특히, 온도, 강수량, 풍속, 습도 등 다양한 기상 요소와 출퇴근 시간대별 혼잡도 간의 상관관계 및 패턴을 탐색적 데이터 분석(EDA)을 통해 심층적으로 파악하였다.

본 연구에서 개발한 혼잡도 예측 모델은 실시간 기상정보를 활용해 지하철 이용자의 혼잡 상황을 사전에 예측함으로써, 교통 관리 당국과 도시 계획자들에게 혼잡 완화 및 안전 관리에 대한 의사결정 제공하여 쾌적한 이동 환경 조성에 기여할 수 있다.

1.2. 활용 데이터

변수	설명	변수	설명	변수	설명
tm	날짜 및 시각	stn	AWS 지점 코드	rn_hr1	시간 강수량(mm)
line	지하철 호선	ta	정시 기온(℃)	hm	상대습도(%)
station_number	역 번호	wd	풍향(degree)	si	일사량 (MJ/m2)
station_name	역 명	ws	풍속(m/s)	ta_chi	체감온도 (℃)
direction	지하철 상하구분	rn_day	일강수량 (mm)	congestion	열차 내 혼잡도(%)
discomfort_index	불쾌지수	is weekend or holiday	휴일/평일	time_type	출근/퇴근/ 그 외 시간

[표 1] 기상청 제공 데이터

본 분석에 사용된 데이터는 2023년 서울 지하철 1~8호선, 총 300여 개 지하철역의 1시간 단위 혼잡도 데이터와, 해당 지하철역 인근 AWS(자동기상관측시스템) 및 ASOS(자동기상관측시스템) 지점에서 관측된 1시간 단위 기상 데이터 및 객관분석자료[표 1]이다. 검증 데이터 또한 이와 유사한 구성으로 1년치 자료가 주어진다. 2023년의 데이터만 사용한 이유는 2.2에서 설명하도록 한다.

2. 데이터 전처리 및 EDA

2.1. 결측값 처리

제공된 데이터에 존재하는 -99를 결측값으로 간주하고, 이 값을 다른 피처(변수)를 활용해 랜덤 포레스트 회귀(RandomForestRegressor) 모델을 통한 예측으로 대체했다. 또한 자료에서 게시된 체감온도 공식과 데이터에서 제시하는 체감온도의 일치율이 0.49%로 자료에서 게시된 체감온도 공식으로 대체했으며, 하루의 시간 강수량의 합이 일 강수량과의 일치율이 27.92%로 이 또한 일 강수량의 합으로 대체했다.

2.2. 2023년 혼잡도 데이터 선정 이유

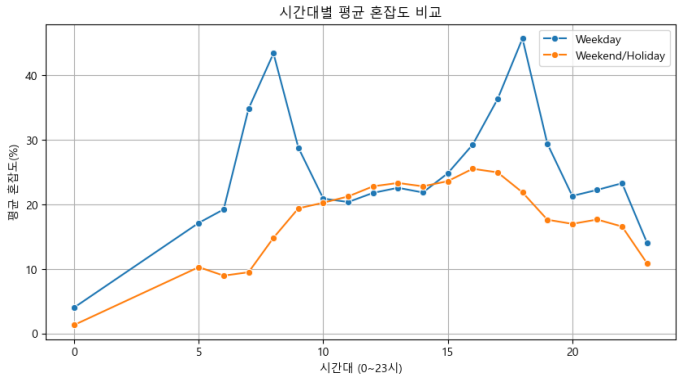
통계청의「코로나 확산에 따른 도시철도 통행량 변화 분석」에 따르면, 코로나19 발생 이후 도시철도의 통행량은 계절적인 패턴을 유지하고 있으나, 전체 통행량은 감소하는 경향을 보인 것으로 나타났다. 특히, 코로나19 확산 이전인 2019년의 월평균 통행량과 비교할 때, 통행량은 2023년에 이르러서야 평균 수준으로 회복되는 경향을 보였다. 이에 따라 2021

년과 2022년의 통행 데이터는 코로나19의 영향이 반영되어 있으며, 이는 현재의 혼잡도 수준을 정확히 반영하지 못하고 분석의 왜곡 요인으로 작용할 수 있다.

따라서 본 분석에서는 코로나19의 영향이 상대적으로 적고 통행량이 정상화된 2023년의 도시철도 혼잡도 데이터를 기반으로 분석을 수행하였다.

2.3. EDA

2.3.1. 평일/공휴일 시간대별 평균 혼잡도 비교

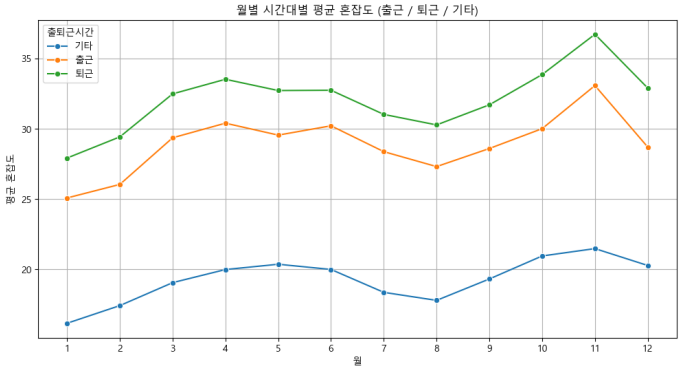


[그림1] 시간대별 평균 혼잡도 비교 (평일 vs 주말/공휴일)

[그림1]은 시간대별 평균혼잡도를 평일(Weekday)과 주말/공휴일(Weekend/Holiday)로 구분하여 나타낸 것이다. 평일에는 두 개의 뚜렷한 피크(혼잡도 급등 구간)가 존재하는데, 각각 오전 8시 전후와 오후 6시 전후로, 이는 일반적인 출근 및 퇴근 시간대에 해당한다. 특히 오전 8시 시간대는 전체 시간 중 가장 높은 평균 혼잡도(약 44%)를 기록하였다. 반면, 주말 및 공휴일에는 이러한 뚜렷한 피크 없이 전반적으로 완만한 혼잡도 곡선을 그리며, 10~17시 사이에 비교적 균일한 수준의 혼잡도를 유지하였다.

이러한 결과는 평일과 주말/공휴일 간의 도시철도 이용 행태가 뚜렷하게 다름을 보여주기에, 시간대뿐만 아니라 요일 정보 또한 반드시 반영되어야 함을 시사한다.

2.3.2. 출퇴근시간 여부에 따른 월별, 시간대별 평균 혼잡도 비교



[그림2] 월별 시간대별 평균 혼잡도 (출근 / 퇴근 / 기타)

[그림2]에서는 월별로 출근, 퇴근, 기타 시간대의 평균 혼잡도를 비교하였다. 분석 결과, 출근(오전 7~9시), 퇴근(오후 5~7시) 시간대의 혼잡도가 기타 시간대에 비해 지속적으로 높은 수준을 보였고, 이러한 패턴은 계절과 무관하게 연중 일관되게 나타났다. 특히 퇴근 시간대의 혼잡도가 출근 시간대보다 전반적으로 더 높게 나타났으며, 이는 퇴근 시에는 개인 일정이 다양하게 분포하는 점에서 비롯된 것으로 해석된다.

이를 바탕으로 혼잡도 예측 모델에 시간 정보를 보다 직관적으로 반영하기 위해 "출근", "퇴근", "기타"의 세 가지 범주로 시간대를 구분하였다. 단순히 시(hour)를 연속형 변수로 처리하는 대신, 출퇴근 시간대의 구조적 차이를 반영한 범주형 변수로 처리함으로써 모델이 시간대별 혼잡도 패턴을 보다 효과적으로 학습할 수 있도록 하였다.

2.4. 파생변수 생성

도시철도 혼잡도는 다양한 요인의 영향을 받기 때문에, 본 분석에서는 예측 모델의 성능을 높이고 혼잡도의 시간적·맥락적 변동성을 효과적으로 반영하기 위해 여러 파생 변수를 생성하였다.

우선, 단순한 기온이나 습도와 같은 개별 기상 요소보다, 사람들이 체감하는 불쾌감 수준이 실제 이동 패턴에 더 큰 영향을 미칠 수 있다는 가설에 기반하여 불쾌지수(discomfort index) 변수를 도입했다.

혼잡 지수(chi)와 시간(hour), 주말 여부 간의 상호작용 변수도 함께 생성하였다. 구체적으로는 &chi × 시간대&, &chi × 주말/공휴일 여부&와 같은 변수들로, 특정 시간대나 특정 요일 조건에서 혼잡도의 변화가 어떻게 달라지는지를 모델이 더 정밀하게 학습할 수 있도록 설계하였다.

마지막으로, 혼잡도의 시계열적 연속성을 반영하기 위해 이전 시간대의 혼잡도 정보도 변수로 포함하였다. 직전 1시간의 혼잡도(lag_1), 직전 3일간의 평균 혼잡도(roll_mean_3), 직전 7일간의 평균 혼잡도(roll_mean_7) 등을 통해 과거 혼잡 패턴이 현재에 미치는 영향을 모델이 학습하도록 하였다. 이를 통해 일시적인 급등락뿐만 아니라, 혼잡도의 점진적인 변화까지 고려할 수 있도록 하였다.

이와 같이 설계된 파생 변수들은 도시철도 혼잡도를 보다 현실적으로 반영하고, 날씨 및 시간적 요인과의 상호작용을 모델이 학습할 수 있도록 구성되었다.

3. Features 선택

3.1. Pearson 상관계수와 Spearman 상관계수

혼잡도 예측 모델의 입력 변수 선정을 위해 피어슨 상관계수와 스피어만 상관계수를 분석하였다.

피어슨 상관계수는 변수 간 선형 관계를, 스피어만 상관계수는 순위 기반의 비선형 관계를 반영한다. 실제 기상 및 혼잡도 데이터는 이상치나 비선형성이 존재할 가능성이 높기 때문에, 스피어만 계수를 주요 기준으로 활용하였다.

분석 결과, hour, ta, chi, discomfort_index는 타겟 변수(congestion)와 비교적 높은 양의 상관을 보여 주요 변수로 선택하였다. 반면 rn_hr1, hm, rn_day 등은 상관관계가 낮거나 방향성이 불명확하여 제외를 고려하였다. 기온 관련 변수(ta, chi, discomfort_index)는 서로 간의 상관이 높아 다중공선성 문제가 발생할 수 있어, 추후 VIF 검토를 통해 일부 조정할 계획이다. 또한 ws(풍속), wd(풍향), si(일사량) 등은 상관계수는 낮지만, 다른 변수와의 상호작용 가능성을 고려해 모델 학습에 포함시킨 뒤, 중요도 기반으로 최종 판단하고자 한다.

최종 변수 선택은 상관계수 분석을 바탕으로 하되, 모델 성능과 해석 가능성까지 함께 고려하여 결정하였다.

4. 군집분석

4.1. 수행 목적 및 최적의 클러스터 수 선정

클러스터링은 역별 주요 혼잡 특성, 불쾌지수, 기상 변수 등 다차원 데이터를 요약하여 서로 유사한 역끼리 그룹화한 후, 군집 간 패턴과 특성을 탐색하기 위해 수행하였다. 또한 군집마다 혼잡도 영향 요인이 다를 수 있으므로, 군집별 모델을 따로 제안할 경우 모델의 복잡성 완화 및 RMSE, R^2 지표 성능 개선 등의 효과를 기대하였다.

클러스터 수 결정은 엘보우 방법(Elbow Method)을 통해 진행하였으며, Within-Cluster-Sum-of-Squares(WCSS)의 감소율이 급격히 완만해지는 지점에서 최적의 클러스터 수(K)를 3으로 설정하였다. 이는 데이터 설명력을 유지하면서 군집 간의 명확한 구분을 가능하게 하는 지점이다.

4.2. 군집별 특징 및 해석, 제거 이유

군집분석 결과 3개의 클러스터로 나누어진 역들의 특성은 다음과 같다.

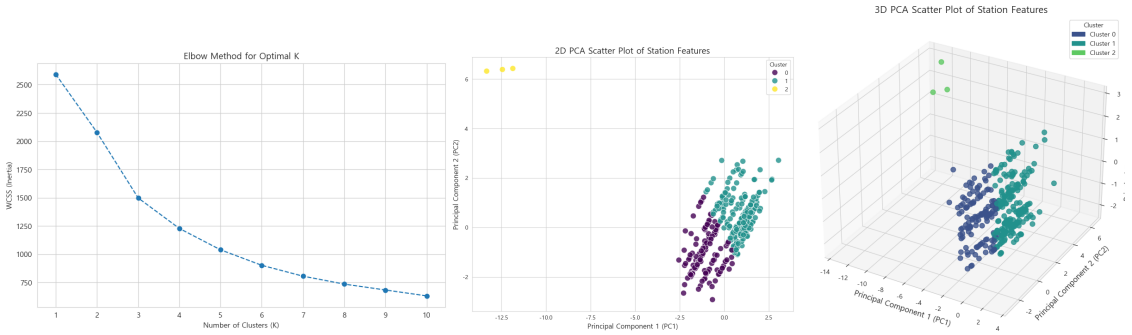
클러스터 0 (중간 혼잡도·중간 불쾌지수): 총 125개 역이 포함되어 있으며 평균 혼잡도는 약 14.8, 평균 불쾌지수는 약 57.5이다. 상대적으로 관리 우선순위는 높지 않으나 지속적 모니터링을 통한 점검이 필요하다.

클러스터 1 (고혼잡도·고불쾌지수): 196개 역으로 구성되며, 평균 혼잡도는 약 33.4로 가

장 높고, 평균 불쾌지수도 58.5로 높은 편이다. 이러한 역들은 서울의 중심지나 주요 혼잡 구간에 위치할 가능성이 높으며, 혼잡 완화와 환경 개선을 위한 최우선 관리 대상이다.

클러스터 2 (저혼잡도·저불쾌지수): 3개 역(연천, 전곡, 청산)으로 구성되어 있으며 평균 혼잡도는 5.8, 불쾌지수는 27.1로 가장 낮다. 이 역들은 상대적으로 혼잡이 적고 쾌적한 환경을 유지하고 있으나, 극히 소수이므로 개별적인 역 특성 분석이 추가로 요구된다.

그러나 군집 간 구분이 명확하지 않고 일부 클러스터가 너무 소규모로 형성되는 등의 한계가 있어, 본 분석에서는 군집 분석 결과를 최종적으로 활용하지 않고 제거하였다.



[그림3] 엘보우 방법, PCA 2차원, PCA 3차원 시각화

5. 모델링 전 전처리

5.1. 수치형 변수 정제 및 다중공선성 검토

지하철 혼잡도를 예측하기 위한 수치형 변수의 적절성을 판단하고자 다중 공선성 검토를 진행하였다. 초기 단계에서는 ta(기온), chi(체감온도), discomfort_index(불쾌지수) 등 온도 관련 변수들 간에 높은 상관관계가 관찰되었고, VIF(Variance Inflation Factor) 값이 60~80 이상으로 나타나 다중공선성이 심각하다는 판단을 내릴 수 있었다.

이에 따라 가장 높은 VIF를 기록한 discomfor_index를 먼저 제거한 뒤, chi(72.47)와 ta(58.68)의 VIF가 여전히 높게 유지되었다. 최종적으로 discomfort_index와 chi를 모두 제거한 후 VIF를 다시 계산한 결과, 모든 변수의 VIF가 5 이하로 낮아지며 다중공선성 문제가 해소된 것으로 판단하였다.

이와 함께 모델의 예측력을 향상시키기 위해 변수 간의 상호작용을 반영한 파생 변수를 생성하였다. 예를 들어, 시간대와 주말 여부를 곱한 hour * is_weekend_or_holiday, 시간대와 출퇴근 여부를 결합한 hour * time_type, 역 번호와 시간대를 곱한 station_number * hour, 강수량과 주말 여부 간 상호작용인 rn_day * is_weekend_or_holiday 등이 있다. 이들은 모델의 Feature Importance 분석 결과에서 상위에 위치하며 유의미한 예측 변수로 작용하였다.

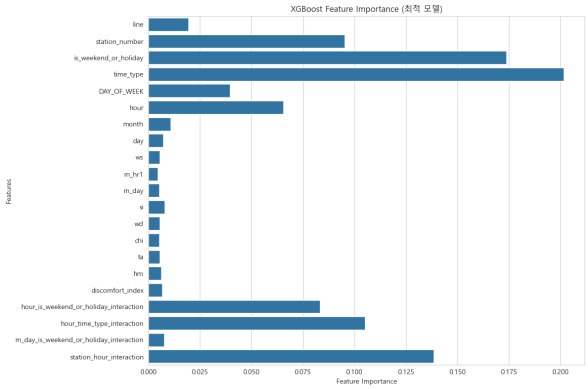
5.2. 모델 평가

5.2.1. 반복 샘플링 기반 학습 및 하이퍼파라미터 튜닝

전체 데이터의 규모가 매우 컸기 때문에 한 번에 학습시키기보다는, 5,000개씩 샘플링하여 500회 반복 학습을 수행하였다. 이를 통해 약 250만 개의 샘플을 바탕으로 평균적인 예측 성능을 평가하였으며, 과적합을 방지하고 일반화 가능성을 높이고자 하였다. 그 결과, 평균 RMSE는 약 12.0930, 평균 결정계수(R^2)는 약 0.6649로 나타나 비교적 안정적인 예측 성능을 확보하였다.

또한, 모델 성능 향상을 위해 XGBoost에 대해 RandomizedSearchCV를 활용한 하이퍼파라미터 튜닝을 실시하였다. 50만 개의 샘플을 기반으로 5-Fold 교차 검증을 수행하였으며, 다음과 같은 최적의 파라미터 조합을 도출하였다.

5.2.2. 최종 모델 평가



[그림3] XGBOOST Feature Importance (최적 모델)

하이퍼파라미터 튜닝을 통해 얻은 최적의 XGBoost 모델을 바탕으로 전체 데이터를 학습하고 최종 테스트 세트를 활용하여 모델의 성능을 검증하였다. 최종 RMSE는 12.0930, 결정계수(R^2)는 0.6649로 나타나 반복 학습 시 얻은 평균 성능과 유사한 수준을 보였다. 이는 모델이 비교적 일관되게 혼잡도를 예측할 수 있음을 시사한다.

특히, 상호작용 변수들이 상위 중요 변수로 작용하면서 시간대, 요일, 기상 조건 등 다양한 요소가 혼잡도에 복합적으로 영향을 미친다는 점을 확인할 수 있었다. 이로써 단순 변수만을 사용하는 모델보다 예측 정밀도가 크게 향상되었다.

6. 결론

6.1. 분석 요약

본 프로젝트의 목표는 기상 조건, 시간대, 역별 특성 등 다양한 요인이 지하철 혼잡도에 미치는 영향을 분석하고 이를 기반으로 혼잡도를 예측하는 모델을 구축하는 것이었다. 데이터 전처리 단계에서는 다중공선성 문제를 해결하기 위해 VIF 분석을 수행하였으며, 모델 성능 향상을 위해 파생 변수 생성 및 범주형 변수에 대한 타겟 인코딩을 진행하여 입력 변수를 정제하였다. 이후 반복 샘플링 기반 학습을 통해 모델의 예측 성능을 안정적으로 평가하였으며, 하이퍼파라미터 튜닝을 통해 최적의 XGBoost 모델을 확보하였다. 최종적으로 구축된 모델은 다양한 요소를 통합적으로 반영하여 RMSE 12.0930, R^2 0.6649의 성능을 달성하였다.

6.2. 활용 방안 및 기대효과

본 예측 모델은 시간대별 혼잡도 변화뿐만 아니라 기상 조건과 주말 여부 등의 외생 변수까지 통합적으로 고려할 수 있는 구조로 설계되었다. 이를 통해 다음과 같은 활용이 가능하다.

첫째, 지하철 운영 기관은 본 모델을 활용해 특정 시간대나 조건에서 예상되는 혼잡도를 사전에 파악하고, 이에 따른 인력 배치 및 운영 전략을 수립할 수 있다.

둘째, 시민들에게는 실시간 혼잡도 예측 정보를 제공함으로써 혼잡한 시간대를 회피하도록 유도할 수 있으며, 이는 궁극적으로 쾌적한 대중교통 이용 환경을 조성하는 데 기여할 수 있다.

셋째, 기상 변화에 따른 이용 행태를 반영한 예측이 가능해짐에 따라 기상 악화 시 특별 운행 계획이나 안내 체계를 효율적으로 마련할 수 있다. 또한, 향후 버스, 자전거 등 다른 교통수단과의 연계 분석을 통해 종합적인 도시 교통 운영 방안으로 확장하는 것도 가능하다.

궁극적으로 본 모델은 도시 내 대중교통 수요 관리를 보다 정밀하고 체계적으로 수행하는 데 있어 핵심적인 기초자료로 활용될 수 있으며, 시민의 이동 편의성과 도시 운영의 효율성을 동시에 높이는 데 기여할 것으로 기대된다.

7. 참고문헌

통계청 - [코로나19 확산에 따른 도시철도의 통행량 변화]