PROJECT

기상 데이터 기반 XGBoost를 활용한 지하철 혼잡도 예측 및 모델 개발

(2025.06.18 ~ 2025.06.27)

프로젝트 개요

기상 데이터와 지하철 혼잡도 데이터를 융합하여 혼잡도 예측 모델 구축을 통한 교통 관리 당국 도시 계획자들이 혼잡 완화 및 안전 관리에 대한 의사결정을 내리는 데 기여.

프로세스

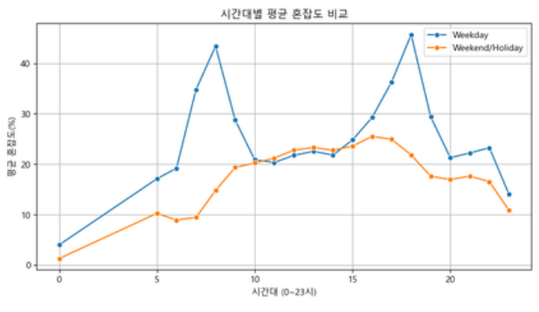
활용 데이터 : 서울 지하철역의 1시간 단위 혼잡도 데이터와 지하철 역 인근 자동기상관측시스템(AWS, ASOS) 지점에서 관측된 1시간 단위 기상 데이터

보유한 컬럼 : 날짜(공휴일 평일 구분, 날짜, 출근/퇴근/그 외 시간 구분) 지하철 정보(역 번호, 역명, 호선, 상하 구분, 혼잡도) 날씨 정보(기온, 풍향, 풍속, 시간 강수량, 일 강수량, 상대 습도, 체감온도, 불쾌 지수)

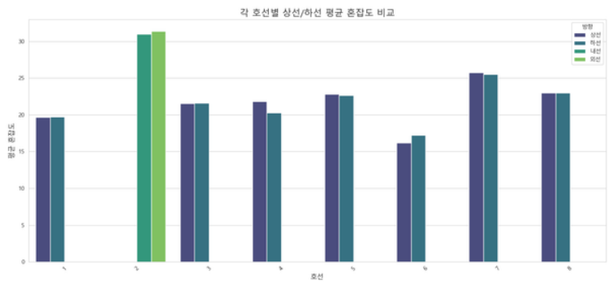
진행 과정 :

분석 결과

문제 정의

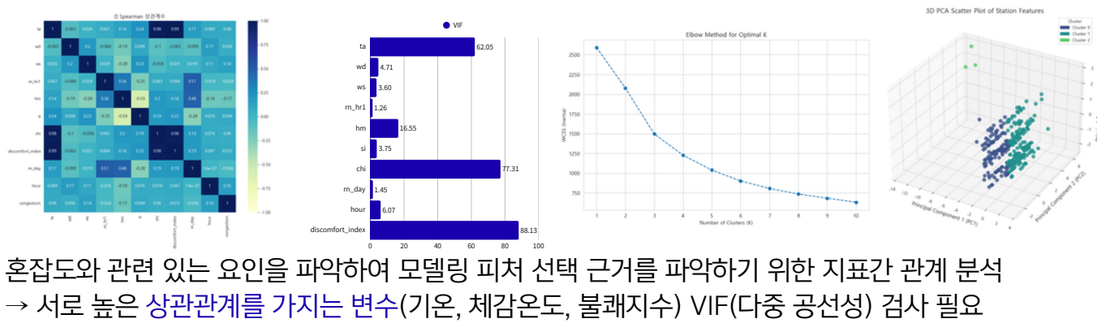


평일과 휴일의 차이가 극명.
출퇴근 시간대의 혼잡도가 큰 차이
→ 출근/퇴근/그 외 시간대 구분 필요



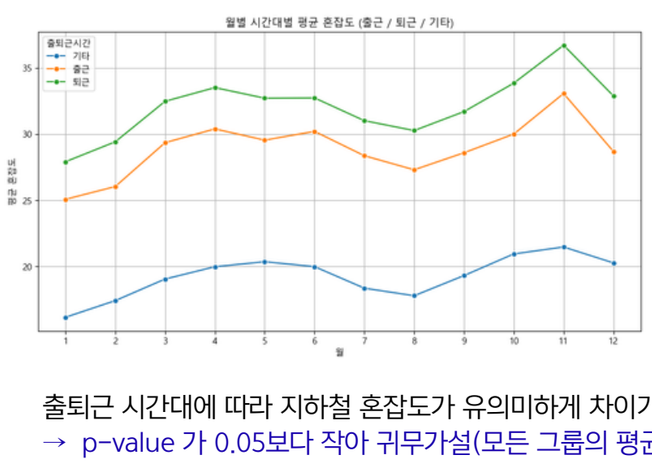
각 호선 별 상선/하선 영향 X

지표간 관계 분석



혼잡도와 관련 있는 요인을 파악하여 모델링 피처 선택 근거를 파악하기 위한 지표간 관계 분석
→ 서로 높은 상관관계를 가지는 변수(기온, 체감온도, 불쾌지수) VIF(다중 공선성) 검사 필요
→ 불쾌 지수, 온도, 체감온도 중 불쾌지수, 체감온도 제거
PCA통한 군집 분석 결과, 명확하지 않은 군집 분류(클러스터 2 의 역 수 : 3)
→ 누적 설명 비율이 100% 아니기 때문에 일부 변동 정보가 여전히 손실
→ 역 별 분류는 큰 의미 X

시간대 차이 검정

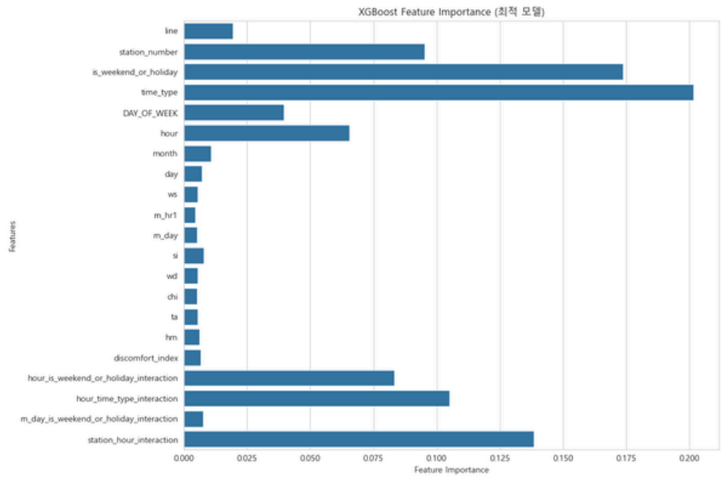



```
from scipy.stats import f_oneway

# 각 그룹의 혼잡도 추출
congestion_go = df[df['출퇴근시간'] == '출근']['congestion']
congestion_off = df[df['출퇴근시간'] == '퇴근']['congestion']
congestion_else = df[df['출퇴근시간'] == '기타']['congestion']

# 일원분산분석 (ANOVA)
f_stat, p_value = f_oneway(congestion_go, congestion_off, congestion_else)

print(f"F-statistic: {f_stat:.4f}")
print(f"P-value: {p_value:.4e}")

if p_value < 0.05:
    print("✅ 출퇴근시간에 따라 혼잡도의 유의미한 차이가 있습니다.")
else:
    print("❌ 출퇴근시간에 따른 혼잡도 차이는 통계적으로 유의하지 않습니다.")
```

```
F-statistic: 619295.3779
P-value: 0.0000e+00
✅ 출퇴근시간에 따라 혼잡도의 유의미한 차이가 있습니다.
```

출퇴근 시간대에 따라 지하철 혼잡도가 유의미하게 차이가 있는지를 검정하기 위해 ANOVA 분석 수행.
→ p-value 가 0.05보다 작아 귀무가설(모든 그룹의 평균이 같다) 기각

모델링



5000 sample x 500, 총 250만개의 반복 샘플링 기반 학습

RandomizedSearchCV 하이퍼파라미터 튜닝

5-Fold 교차검증, 범주형 변수 타겟 인코딩

RMSE : 12.09, R^2 : 0.66

활용방안 및 한계점

활용 방안

특정 시간대나 조건에서 예상되는 혼잡도를 사전에 파악 후 인력 배치 및 운영 전략 수립, 실시간 혼잡도 예측 정보 제공으로 **혼잡한 시간대 회피 유도를 통한 쾌적한 대중교통 이용 환경 조성,** 기상 변화에 따른 이용 행태를 반영한 예측으로 기상 악화 시 **특별 운행 계획 및 안내 체계 수립**.

한계점

보유한 데이터에서 날씨의 영역보단 시간 영역의 영향이 더 큼. → 외부 데이터가 있다면 보다 집중 탐구 가능.
클러스터 분리가 안된 집단에 대한 탐구 필요, 2023년의 특정 역에 행사가 발생했을 경우 다른 모델링 대입 필요
→ 시간의 한계로 진행 x