I. 서론

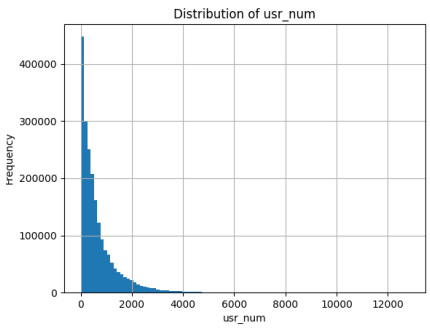
2024년 7월 부산시 지하철역별 시간대별 탑승객 수를 예측하기 위해서는 해당 자료가 시공간 데이터라는 점을 고려해야 한다. 본 분석에서는 이를 반영하기 위해 시계열 데이터라는 점을 감안하여 월, 일, 시간에 대해 cos/sin 주기성 정보를 변수로 삽입하였고, 서면역, 센텀시티역 등 유동인구가 많은 역이 존재한다는 점을 감안하여 주어진 경위도 좌표를 중부원점 좌표계로 변환하여 분석에 사용하였다. 평면 직교 좌표계로의 변환을 통해 비교적 작은 지역 내에서 보다 정확한 위치 정보를 반영하였다.

II. 데이터 전처리

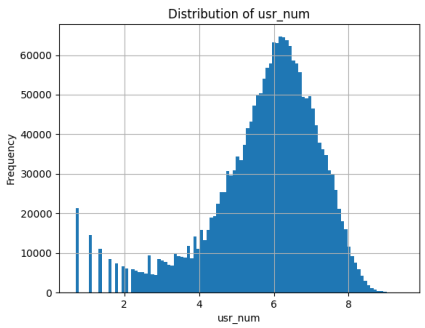
추가 데이터 수집을 통해 X1(기온), X2(강수량), X3(환승역), X4(공휴일/주말), X5(출퇴근 시간대), X6(계절), X7(축제 및 행사 등 이벤트), X8~X10(년/월/일), X11~16(주기성 정보), X17~X18(중부원점 좌표)의 변수를 추가하였다. X3~X10은 범주형 변수로 X3은 0, 1, 2의 값을, X6은 겨울(1), 봄(2), 여름(3), 가을(4)의 값을, X4, X5, X7은 이진값을 갖는다. 년월일은 개별 숫자의 크기가 의미있지 않으므로 범주화하여 사용하였다. 이는 추후 분석에서 독립변수로써 Y(종속변수)의 탑승객수를 예측하는 데 사용된다.

본 분석에 사용될 회귀 모델은 회귀 함수 기반이 아니므로 탑승객 수의 정규성을 확인하는 것이 필수적이지는 않지만, 모델의 성능 향상을 위해 탑승객 칼럼의 정규성을 확인하였고 [그림1]과 같이 왜곡이 심하여 로그 변환을 수행하였다.

[그림 1] 로그 변환 전 히스토그램



[그림 2] 로그 변환 후 히스토그램



범주형 데이터를 분석에 이용하기 위해서 인코딩이 필수적인데 대표적인 인코딩 방식인 원-핫 인코딩 (One-Hot Encoding)을 사용하게 되면 변수 간 상관관계가 높은 것을 확인하였다. 이에 다중공선성을 우려하여 타겟 인코딩(Target Encoding)을 수행하여 각 범주를 해당 범주의 Y 평균값으로 변환하여 분석에 사용하였다. 이는 과적합의 위험을 수반하는데 3장에서 소개될 모델의 하이퍼 파라미터 조정을 통하여 과적합 문제를 완화하였다.

추가적으로 변수들간의 상관계수를 확인하여 계절 변수를 제거 후 분석을 시행하였다.

Ⅲ. 분석

본 분석에서는 회귀 트리와 부스팅 알고리즘을 기반으로 한 머신러닝 모델인 XGBoost를 사용하여 예측을 진행하였다. 부스팅 알고리즘은 개별 약한 학습기(weak learner)에 가중치를 부여하면서 이를 결합해 학습/예측하는 방식으로, 가중치 부여에는 모델의 손실함수(loss function)을 최소화하도록 가중치를 업데이트하는 경사 하강법(Gradiant Descent)이 사용된다. 회귀 트리의 경우 직선으로 예측 회귀선을 표현하는 선형 회귀와는 달리 분할되는 데이터 지점에 따라 브랜치(branch)를 만들며 계단 형태로 회귀선을 만드는 방식이다.

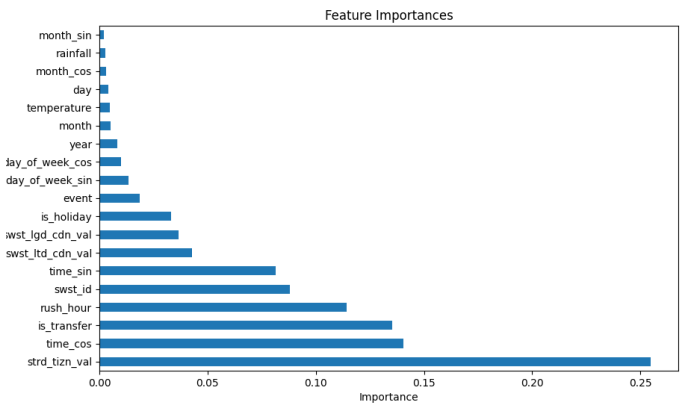
GridSearch 방법을 통해 모델의 학습기 개수와 학습률, 트리 생성을 위한 데이터의 샘플링 비율 등의 하이퍼 파라미터 튜닝을 시행하였고 이를 통해 모델의 과적합 위험을 줄이면서 예측 성능을 향상시켰다.

2022년 1월부터 2024년 6월까지의 데이터를 80:20의 비율로 나누어 train/test set을 생성하였고 이를 통해 모델의 RMSE를 측정하여 모델의 적합성을 평가하였다. RMSE : 179.444로 실제 탑승객 수 데이터의 평균이 661.006인 점을 감안하면 무난한 값으로 모델이 더 개선될 여지가 있음을 확인해볼 수 있다.

Ⅳ. 결론

피처 중요도를 통해 각 변수들이 모델의 예측 성능에 기여하는 정도를 확인하여 [그림3]과 같은 결과를 얻었다. 시간대와, 환승역 여부, 출퇴근 시간대 등이 탑승객 수에 크게 영향을 미치는 변수인 것을 알 수 있으며 이는 직관적으로 납득할만한 결과이다.

[그림 3] 피처 중요도 그림



향후 분석에서는 데이터의 시공간 의존성에 관한 고차원적 탐색과 독립변수의 엄선을 통해 보다 향상된 예측 능력을 가진 모델을 이끌어낼 수 있을 것으로 기대한다.

참고문헌

- 권철민, 2022, 파이썬 머신러닝 완벽 가이드, 개정2판, 위키북스, 서울
- Cho, S., Kim, B., Kim, N., & Song, J. (2019). 서울의 지하철 역 이용객 수에 대한 연구.
- 이현상, & 오세환. (2020). 시계열 예측을 위한 LSTM 기반 딥러닝: 기업 신용평점 예측 사례. 정보시스템연구, 29(1), 241-265.
- 김성연, 이정형, 김영근, 이규용, & 권치명. (2007). 경영계획과 지역개발을 위한 지하철 수송수요 예측의 통계적 방법. Journal of The Korean Data Analysis Society, 9(2), 835-844.

정시원, 임선우, 주원우