머신러닝 기법을 이용한 서울시 도로 통행 속도 예측 - 강남구와 중구를 중심으로

이은진

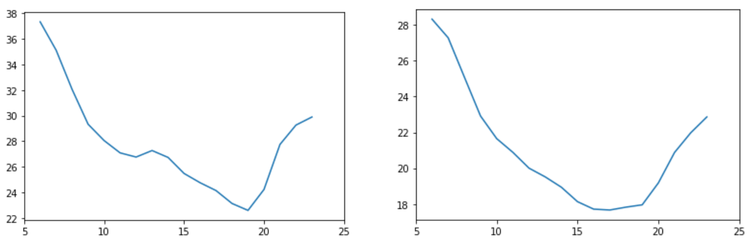
Ⅰ. 서론

- **연구의 목적 :** '교통지옥'이라는 별명을 가진 강남구는 주거지구와 상업지구, 대치동 학원가까지 밀집되어 혼잡하고 통행량이 많은 편이다. 이와는 반대로 중구의 경우, 상업밀집지역이라는 단일 특징이 있다. 강남구와 중구의 도로 통행 속도를 각각 예측해, 지역 특성과 관련해 비교하고자 한다.

Ⅱ. 자료 설명

- **데이터 수집 과정**

- 연구 대상 : 2018.01.01 ~ 2018.12.31 서울시 강남구와 중구 차량통행속도



<그림 1> 시간대별 강남구(왼쪽)와 중구(오른쪽) 평균 통행 속도

→ 서행 원인을 찾기 위해, 오전 10시부터 오후 23시까지의 차량통행속도 데이터 이용.

- 이용 데이터
  1) 2018년 기상 데이터(기온, 풍속, 강수량 등)
  2) 2018년 공기오염도 데이터(미세먼지, 초미세먼지 등)
  3) 2017년 교통사고 데이터(사망자수, 중상자수 등)
  4) TOPIS 도로특성 데이터(차선수, 거리, 방향, CCTV 개수, 신호기 개수 등)

- 최종 데이터 : 2018년 서울시 차량통행속도 데이터를 기준으로 날씨 및 공기오염도 데이터를 합칠 때에는 링크아이디, 권역구분과 일자, 시간대를 기준으로 처리하였다. 또한 교통사고가 자주 일어나는 도로에 영향을 받을 것이라 판단해, 2017년 교통사고 데이터를 활용해 도로명을 기준으로 사망자수, 중상자수 등의 변수를 병합하였다.
- 최종 데이터 수 : 강남구= 1,489,618 / 중구 = 958,080

도로명	링크아이디	시점명	종점명	거리	CCTV	어린이 보호구역
개포로	1220020500	포이사거리	구룡초교	862	1	0
개포로	1220018600	구룡초교	개포고교	651	1	0
개포로	1220015600	개포고교	개포동역	658	1	2
개포로	1220012000	개포동역	대모산입구역	889	2	1
개포로	1220008200	대모산입구역	대청역	558	1	0
개포로	1220007000	대청역	대청초교	888	0	1

<그림 2> 도로명-링크아이디(왼쪽)와 최종데이터 샘플(오른쪽)

Ⅲ. 분석 방법

- **모델링 분석 방법**

- 다중 선형회귀분석 : 2개 이상의 독립변수를 이용하는 모형으로 다음과 같은 형식을 가진다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon$$

- 릿지 회귀 모형(Ridge Regression) : MSE를 최소화하면서, 회귀계수벡터 β의 $L_2 norm$을 제약한다.

$$\hat{\beta}^{\text{Ridge}} = \text{argmin}_\beta \, (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

- 라쏘 회귀 모형(Lasso Regression) : 릿지 회귀모형식의 형태와 동일하지만 $L_1 norm$을 제약하고, $L_1 norm$은 미분이 불가능하다는 차이점이 있다. 또한 릿지 회귀모형과 다르게 변수 선택을 할 수 있다.

$$\hat{\beta}^{\text{Lasso}} = \min_\beta \, (y - X\beta)^T (y - X\beta) + \lambda |\beta|_1$$

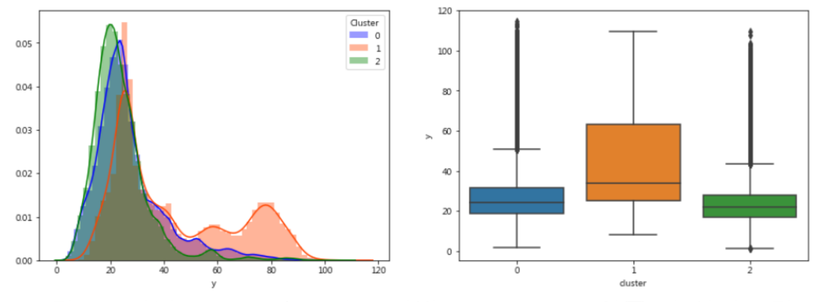
- 랜덤포레스트(Random Forest) : 의사결정 나무가 여러 개 모여서 만들어진 모형이다. Bagging 기법을 통해 데이터 셋을 생성한 후 각각 의사결정 나무를 적용한다. 그 후 선택된 의사결정 나무를 무작위로 다시 생성하고 취합해, 모델 예측력을 높인다.
- XGBoost(Extreme Gradient Boosting) : Gradient Boosting(GBM)의 속도와 성능이 향상된 모형이다. GBM은 가중치를 계산할 때 Gradient Descent를 이용하여 최적의 모수를 찾아내는데, 시간이 오래 걸리고 과적합의 가능성이 높다. 이를 보완하기 위해 XGBoost는 GBM을 병렬구조로 나누어 이런 한계를 해결한다.
- lightGBM : 2017년 Microsoft에서 발표한 모델로, XGBoost를 보완하기 위해 나온 모형이다. XGBoost가 처리하지 못하는 대용량의 데이터를 학습할 수 있으며, 히스토그램 기반 근사치를 사용해 XGBoost 대비 성능이 향상되었다.

- **군집화 알고리즘**

- K-means 군집화 알고리즘을 적용해, 도로 특성변수로만 강남구, 중구를 각각 3개의 군집으로 나눴다. 위의 6가지 모델을 적용하고 test set에서 RMSE가 가장 작은 모델을 최종 모형으로 선택한다(train:test=7:3)

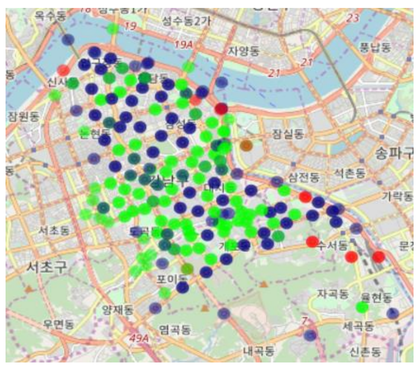
Ⅳ. 분석 결과

- **강남구 분석 결과**



<그림 3> 강남구 도로 속도의 군집별 density plot(왼쪽)와 상자그림(오른쪽)

→ 반응변수인 도로 속도의 군집별 density plot과 상자그림을 확인했을 때, 군집 0과 군집 2의 속도가 비슷하고, 군집 1이 다른 군집에 비해 속도가 높은 것을 알 수 있다.

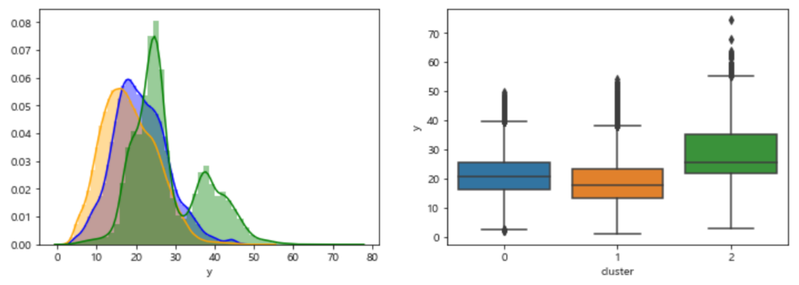


<그림 4> 강남구 군집별 지도 시각화

구분	구간 개수	데이터 수	최종 모형	RMSE
상업밀집 지역(군집0)	131	659,526	랜덤포레스트	3.287
강남구 외곽 도로(군집1)	14	70,040	lightGBM	7.616
주거밀집 지역(군집2)	153	760,052	lightGBM	2.518

- 상업밀집지역(군집 0) 주요 변수 : 신호기 개수, 도로 길이, CCTV 개수, 사망자수,시간대
  → 주요 변수 10개 중 날씨 관련 변수 0개
- 강남구 외곽도로(군집 1) 주요 변수 : 기온, 시간대, 미세먼지 농도, NO2 농도, 오존농도
  → 주요 변수 10개 중 날씨 관련 변수 6개
- 주거밀집지역(군집 2) 주요 변수 : 도로 길이, 기온, 시간대, 상/하행, 미세먼지 온도
  → 주요 변수 10개 중 날씨 관련 변수 5개

- **중구 분석 결과**



<그림 5> 중구 도로 속도의 군집별 density plot(왼쪽)와 상자그림(오른쪽)

→ 강남구와 마찬가지로 반응변수인 도로 속도의 군집별 density plot과 상자그림을 확인했을 때, 군집 0과 군집 1의 속도가 비슷하고, 군집 2이 다른 군집에 비해 속도가 높은 것을 알 수 있다.

<그림 6> 강남구 군집별 지도 시각화

구분	구간 개수	데이터 수	최종 모형	RMSE
중구 외곽 (군집0)	66	309.926	랜덤포레스트	2.281
중구 내부 (군집1)	131	611,122	랜덤포레스트	2.288
대로 및 남산 터널 (군집2)	8	37,032	lightGBM	2.793

- 중구 외곽 도로(군집 0) 주요 변수 : 신호기 개수/도로길이, 도로 길이, 시간대, 주말 여부, 방향
  → 주요 변수 10개 중 날씨 관련 변수 0개
- 중구 내부 도로(군집 1) 주요 변수 : 제한속도, 시간대, 도로길이, 주말 여부, 사고심각도
  → 주요 변수 10개 중 날씨 관련 변수 0개
- 대로 및 남산터널(군집 2) 주요 변수 : 시간대, 도로길이, 온도, 미세먼지 수치, 공휴일 여부, 풍속
  → 주요 변수 10개 중 날씨 관련 변수 6개

Ⅴ. 결론

- 강남구 상업 밀집 지역 도로 속도에 영향을 주는 상위 10개 변수에는 기상, 공기오염도 변수가 포함되지 않아, 출퇴근에 영향을 주지 않는 것으로 예상된다. 반면 외곽 도로와 주거 밀집 지역에는 기상, 공기오염도 변수가 포함되어 기상 상황이 좋지 않을 때 자차를 많이 이용해 날씨 상황에 영향을 받는 것으로 예상된다.
- 중구의 경우 대부분의 도로 속도가 기상 상황, 미세먼지 등에 영향을 받기보다 도로 변수, 시간 변수에 영향을 많이 받는 것으로 나타났다.
- 강남구와는 다르게 중구 중요 변수에는 주말 여부나 공휴일 여부가 포함되었는데, 중구 특성상 상업 밀집 지역이라는 단일 특성 때문인 것으로 예상된다.