회귀분석을 이용한

제주 버스 운행시간 예측

고려대학교 대학원

통계분석방법론 이재원 교수님

통계학과 응용통계학전공

2020021202 이혜원

목차

I. 서론

제주도는 한국인들 그리고 외국인 관광객들이 사랑하는 관광지 중 하나이다. 2019년 11월 기준으로 제주도민 인구가 연평균 4%의 비율로 늘고 있고, 관광객 또한 증가하면서 제주도 전체 상주 인구는 90만명을 넘을 것으로 추정된다. 또한 제주도민과 관광객의 수가 늘면서 2018년에 비해 제주 대중교통 이용객은 총 6245만명 정도로 집계될 정도로 폭발적으로 늘어났다. 본래는 불편한 대중교통 시스템으로 인해 자가용으로 여행을 하는 관광객이 많았다. 하지만 제주도의 대중교통이 개편되면서 대중교통 이용객 수 또한 많아져 전체적으로 제주도 내의 교통체증이 심화되는 상황이다. 제주특별자치도의 연구 내용에 따르면 제주 지역의 교통사고 인구 10만명당 교통사고 건수는 791건으로 전국 1위이고, 교통혼잡비용은 2016년 5000억원을 초과하여 도민 1인당 연 76만원의 교통혼잡비용을 지불해야하는 상황이다. 이러한 제주의 꼬리표처럼 따라다니는 교통체증을 해결하기 위해 대중교통을 효율적으로 운행하기 위한 방안이 필요하다.

따라서 본 보고서에서는 제주도 버스의 운행시간을 예측하는 프로젝트를 진행할 것이다. 버스의 운행시간을 예측하면 버스의 배차 간격, 버스 운행 노선마다 필요한 버스의 대수 등등을 효율적으로 관리할 수 있을 것으로 예상된다. 본 분석에서는 통계분석방법론 수업에서 학습한 다양한 회귀 모델로 버스 운행시간을 예측한 후, 가장 예측 성능이 좋은 모델을 이용해 최종 예측을 진행하고, 결과를 해석할 것이다.

II. 탐색적 데이터분석 및 특성공학

1. 활용 데이터 개요

데이터는 dacon의 버스 운행시간 관련 데이터를 사용하여 분석을 진행했다. Train data는 총 210457개의 관측 데이터와 14개의 변수로, test data는 91774개의 데이터와 13개의 변수로 구성 되어있다. 즉, 본 분석의 목표는 train data를 이용해 test data의 예상 운행 시간을 예측하는 것이다. 데이터를 구성하는 변수에 대한 정보는 아래와 같다.

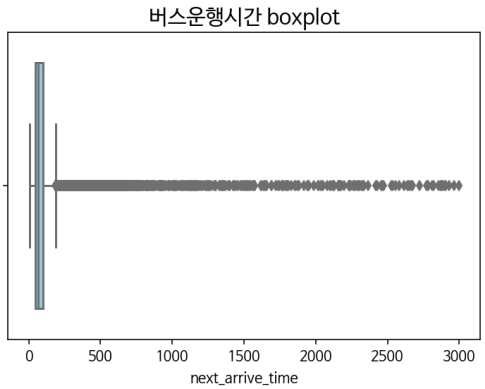
변수명	상세 설명
date	버스운행날짜
route_id	버스 노선 ID
vh_id	버스 ID
route_nm	버스 노선 실제번호
now_latitude	현재 정류소의 위도

now_longitude	현재 정류소의 경도
now_station	현재 정류소 이름
now_arrive_time	현재 정류장에 도착한 시간
distance	현재 정류장에서 다음 정류장까지 실제 이동한 거리
next_station	다음 정류소 이름
next_latitude	다음 정류소 위도
next_longitude	다음 정류소 경도
next_arrive_time (target)	다음 정류소에 도착할 때까지 걸린 시간

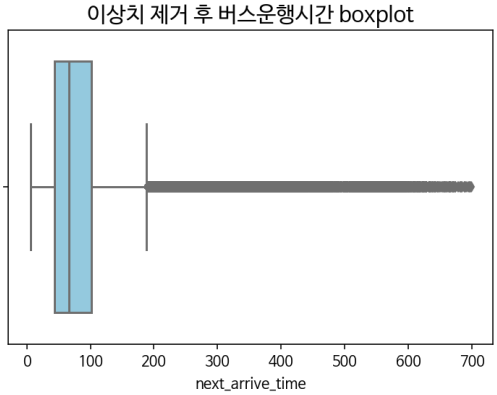
2. Data Cleansing

2.2.1 이상치 제거

본격적인 EDA를 하기 전, 데이터에 이상치와 결측값이 있는지 확인한다. 제주 버스운행 데이터에는 결측값은 없다. 하지만 예측 대상인 next_arrive_time의 분포를 보면 운행 시간이 천 단위인 지나치게 큰 값들이 존재한다. 일반적으로 버스가 한 정류장을 이동하는데 걸리는 시간이 1000분을 넘는 일은 거의 없다. 따라서 운행 시간이 너무 큰 값들은 이상치(outlier)로 판단하여 제거한다.



이상치를 제거하는 과정은 다음과 같다. 본 데이터의 distance와 next_arrive_time 변수는 상관계수가 약 0.44 로 약한 양의 상관관계를 갖고 있다. 그리고 일반적으로 거리가 멀면 버스가 운행하는 시간은 더 길어진다. 따라서 distance가 최대값인 데이터 중 next_arrive_time의 최대값을 찾는다. 이 next_arrive_time의 최댓값은 664이다. 이를 이용하여 train data에서 next_arrive_time이 700 이상인 데이터들을 제거한다. 해당 기준 최댓값은 664이지만, 좀 더 유연한 처리를 위해 700으로 기준을 설정했다. 이상치를 제거한 후 운행시간의 분포는 오른쪽과 같다.



2.2.2 Train 및 test dataset에 없는 데이터 제거

기준 변수는 'now_arrive_time'이다. Train set에는 06시부터 00(24)시까지의 데이터가 있지만, test set에는 06시부터 23시까지의 데이터만 존재한다. 원활한 예측을 위해, train set에서 now_arrive_time이 00시인 데이터를 제거한다. 해당 데이터는 5개이기 때문에 제거하도 전체 성능에 영향을 미치지 않는다.

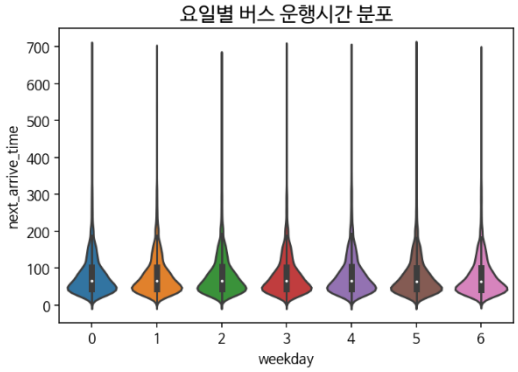
3. 탐색적 데이터분석(EDA) 및 특성공학(Feature Engineering)

이 섹션에서는 데이터를 구성하는 변수들과 target 값인 버스 운행시간이 어떤 관계에 있는지를 확인한다. 그리고 필요한 파생변수를 생성한다.

(1) Date: 버스 운행 날짜

본 데이터의 train set의 버스 운행 날짜는 2019년 10월 15일부터 10월 28일까지이고, test set의 날짜는 2019년 10월 29일부터 11월 5일까지이다. 즉, 본 분석의 목표는 2019년 10월 15일부터 28일까지의 데이터를 이용해 10월 29일~11월 5일의 제주 버스 운행시간을 예측하는 것이다. 따라서 본 데이터의 총 기간은 한 달이 되지 않기 때문에 연도와 월(month)을 따로 추출하여 분석하는 것은 예측에 의미가 없다고 판단했다. 대신, 버스운행시간에는 요일, 평일/주말 여부가 영향을 미친다고 예상했다. 출근/등교를 하는 평일에는 교통체증이 많기 때문에 버스 운행시간 또한 증가한다고 판단했다.

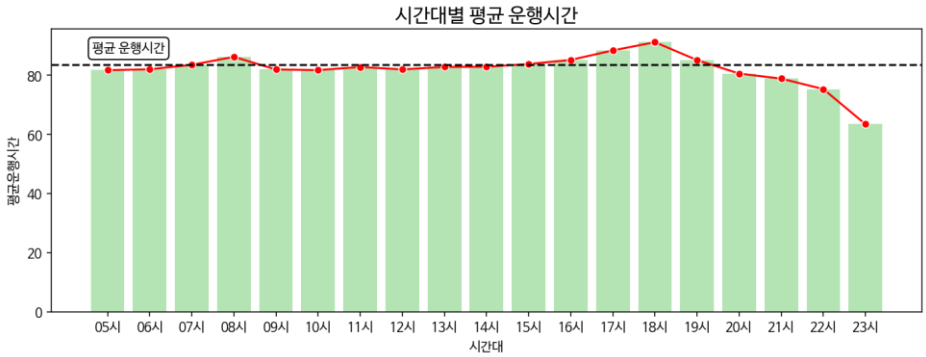
따라서 train, test set 모두 date 변수를 날짜형으로 변환한 뒤, 요일을 추출하고, 요일을 기준으로 평일/주말 여부를 binary로 나타내어 weekend라는 변수를 생성하였다.



하지만 요일별로 버스 운행시간의 violin plot을 그렸을 때(오른쪽 그림), 요일별 운행시간은 거의 차이를 보이지 않았다. 따라서 요일 변수는 제거하고, weekend 변수만 모델링에 반영하였다.

(2) Now_arrive_time: 현재 정류장에 도착한 시간

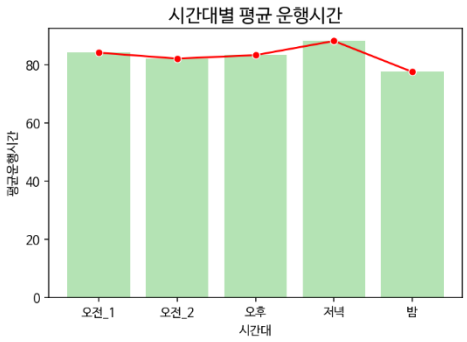
이 변수는 앞서 밝혔듯이 현재 정류장에 도착한 시간인 06시부터 23시까지를 시간 단위로 나타낸다.



위 그래프는 시간대별로 평균 버스 운행시간을 나타낸다. 평균 운행시간이 가장 긴 시간대는 오전 8시, 오후 5시-7시이다. 오전 8시는 사람들이 주로 등교를 하거나 출근을 하는 시간대이고, 오후 5시-7시는 사람들이 하교하거나 퇴근을 하는 시간이기 때문에 다른 시간대에 비해 평균 운행시간이 긴 것으로 보인다.

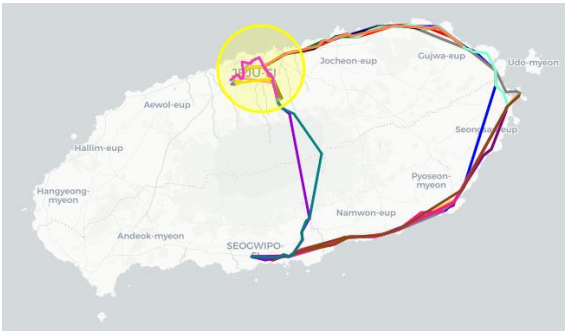
하지만 19개의 시간대를 모두 one-hot encoding을 하게 되면 모델링을 할 때, feature의 수가 너무 많아져 모델 성능을 저하시킬 수 있다. 따라서 19개의 시간대를 크게 5개의 시간대로 나눈다. 시간대를 나누는 기준은 다음과 같이 설정했다.

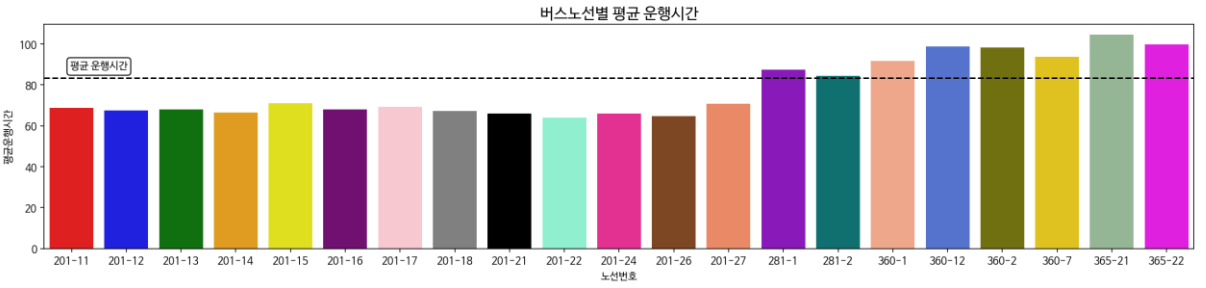
- 오전_1: 05시 ~ 08시
- 오전_2: 09시 ~ 11시
- 오후: 12시 ~ 16시
- 저녁: 17시 ~ 19시
- 밤: 21시 ~ 23시

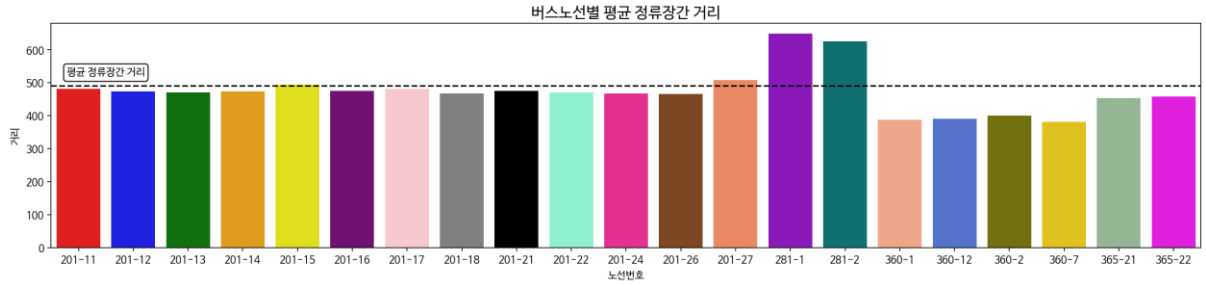


(3) Route_nm: 버스 노선 번호

본 데이터에는 총 21개의 노선이 존재한다. 각 노선이 어떤 경로로 운행하는지 알아보기 위해 folium을 이용해 제주 버스 노선을 시각화했다.







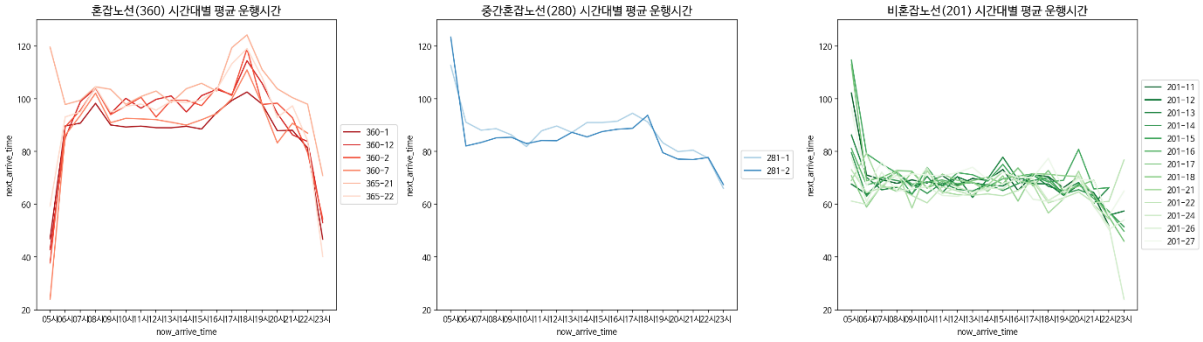
첫번째 그래프는 21개의 노선을 각기 다른 색으로 나타낸 지도이다. 그리고 두번째 그래프는 버스노선별 평균 버스 운행시간을 계산하여 나타낸 그래프이다. 세번째 그래프는 노선별 평균 정류장 간 거리를 나타낸다.

두번째 그래프를 보면 360으로 시작하는 버스들이 평균 운행시간이 가장 길고, 그 다음으로는 280으로 시작하는 버스들의 운행시간이 길다. 그리고 360 버스들은 첫번째 그래프의 노란색 원으로 표시했는데, 이는 360 버스들은 제주시내에서 운행한다는 것을 나타낸다. 제주도의 제주시는 제주도 내에서 가장 인구가 밀집한 지역이다. 많은 인구는 곧 교통이 혼잡하다는 것을 의미하므로 제주시내에서 운행하는 360 버스들의 평균 운행시간이 길 것이다. 그리고 흥미로운 사실은 360 버스들의 평균 정류장간 거리를 제일 짧지만 평균 운행시간은 가장 길다는 것이다. 이는 제주시내의 교통이 가장 혼잡하다는 것을 나타낸다.

그리고 그 다음으로 평균 운행시간이 긴 280버스에 해당하는 지도를 보면, 제주시에서 한

라산 주변을 통과해 서귀포시로 가는 노선이다. 이 노선의 경우 한라산을 지나기 때문에 정류장 간의 거리가 가장 멀기 때문에 비교적 평균 운행시간이 긴 것으로 보인다.

위 분석을 통해 우리는 제주 21개의 버스 노선이 크게 세 가지 그룹으로 구분된다는 것을 알 수 있다. 360번 버스, 280번 버스, 201번 버스 그룹으로 나눌 수 있다. 아래 그래프는 세 그룹의 시간대별 평균 운행시간을 나타낸 것이다. 혼잡노선인 360번 버스들이 등하교 및 출퇴근 시간대에 평균 운행시간이 증가하는 것에 반해 280번과 201번 버스는 뚜렷한 패턴을 보이지 않는다.



위 EDA결과에 착안하여, 버스 노선을 세 개의 그룹으로 나누는 파생 변수 crowded_bus를 생성했다.

(4) 추가 전처리

- Distance 로그변환: distance 변수의 왜도(skewness)는 8이 넘기 때문에 왜도가 크다고 볼 수 있다. 왜도가 큰 변수의 경우, 밀집된 데이터는 모델링에 반영이 잘 되지만 그렇지 않은 데이터는 반영이 되지 않는 상황이 발생할 수 있다. 따라서 왜도가 큰 distance를 로그변환 한다.
- 카테고리형 변수는 one-hot encoding하고, 이중에서 정류장을 나타내는 now_station과 next_station은 label encoding을 한다. 이 두 변수는 unique한 값이 모두 350개로 이 두 변수에 대해 one-hot encoding을 하면 feature의 수가 350개 생성이 되어 모델 성능 저하의 요인이 될 수 있다. 따라서 label encoding을 진행한다.

III. 회귀분석 모델링

제주 버스 운행시간을 예측하기 위한 회귀 모델에는 baseline model로 Linear Regression(선형회귀), Robust Regression(로버스트 회귀), Random Forest Regressor(랜덤포레스트회귀)의 세 종류를 사용하였다. 모델링에 앞서, 예측 대상인 next_arrive_time을 로그변환했다. 이 변수 또한 왜도가 3 이상이기 때문에 위 distance와 비슷한 이유로 로그변환을 하는 것이 타당하다고 판단했다.

1. 모델링 과정

Step 1 파라미터 튜닝이 필요한 경우 GridSearchCV로 가장 좋은 성능을 내는 파라미터 선택
Step 2 최적 파라미터를 적용한 모델(혹은 기본 모델)에 5-fold 교차검증 시행 후 각 fold마다 RMSLE score 출력

Step 3 5-fold의 평균 RMSLE를 기준으로 성능이 좋은 모델로 test set 예측

사용모델: Linear Regression, Huber Regression, Random Forest Regressor

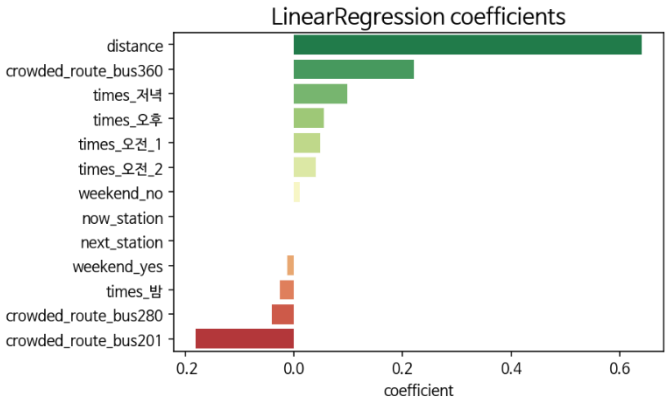
2. 모델링 시행

모델	파라미터	평균 RMSLE
Linear Regression	x	0.4829
Huber Regression	x	0.4836
RandomForest Regressor	n_estimators: 500 max_depth: 10	0.3421

(1) 선형회귀(Linear Regression)

Baseline model로는 기본적인 선형회귀 모델을 사용했다. 선형회귀를 이용해 전처리한 데이터를 학습하여 평가한 결과, 평균 RMSLE는 0.4829가 나왔다. 선형회귀는 이론적으로 반응변수와 설명변수의 선형성이 만족될 때 성능이 좋다. 하지만 본 데이터의 경우, distance와 next_arrive_time의 피어슨 상관계수가 0.44로 약한 선형성을 띄고 있고, 카테고리형 변수들이 많기 때문에 전체적으로 반응변수 next_arrive_time과 다른 feature들이 선형관계에 있다고 보기 어렵다. 이러한 이유로 Random Forest Regressor보다 RMSLE가 높게 나왔다고 해석할 수 있다.

[선형회귀계수 해석]



왼쪽의 그래프는 선형회귀 모델에 데이터를 학습시킨 결과 도출한 설명변수별 회귀계수이다.

- Distance의 회귀계수는 0.6404로 버스 운행시간 예측에 가장 큰 영향을 미쳤다. Distance의 회귀계수의 p-value가 유의수준 5%보다 작았기 때문에, 우리는 distance가 1단위 증가하면 버스 운행시간이 0.6404 단위만큼 증가한다고 해석할 수 있다.

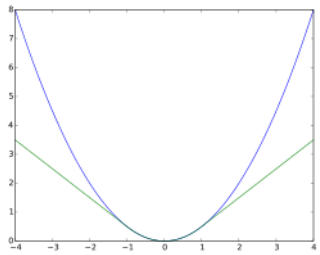
- 360번 버스 여부의 회귀계수는 0.2530으로 360번 버스이면 버스 운행시간이 증가한다는 것을 의미한다. 이는 앞서 EDA파트에서 분석한 결과와 일치한다. (EDA 파트에서 360번 버스의 평균 운행시간이 가장 길었다.) 비슷한 맥락으로 201번 버스 여부의 회귀계수는 -0.2212로 버스 번호가 201로 시작하면 버스 운행시간이 감소한다는 해석을 할 수 있다.

- 시간대를 4 구간으로 나눈 변수들도 다른 변수들에 비해 버스 운행시간 예측에 유의미한 영향을 끼쳤다. 해당 회귀계수의 p-value는 모두 유의수준 5%보다 작다. 회귀계수는 절댓값 기준으로 저녁>오후>오전_1>오전_2>밤 순으로 컸다. 선형회귀모델은 저녁시간대에 평균 버스 운행시간이 가장 길고, 밤에 가장 짧다는 것을 반영하였다. 오후, 오전_1, 오전_2의 경우 평균 운행시간의 차이가 크지 않았기 때문에 회귀계수 절댓값의 차이가 거의 없었다. 필자는 밤 시간대에 해당하는 회귀계수가 오후, 오전보다는 클 것이라고 예상했지만, 선형회귀 모델은 예상과 달리 밤 시간대의 회귀계수를 작게 예측했다.

- 출발 정류장(now_station)과 도착 정류장(next_station)의 회귀계수는 거의 0에 가까웠다. 따라서 이 두 변수는 선형회귀모델에서는 버스 운행시간에 미치는 영향이 거의 없다.

(2) Huber Regression

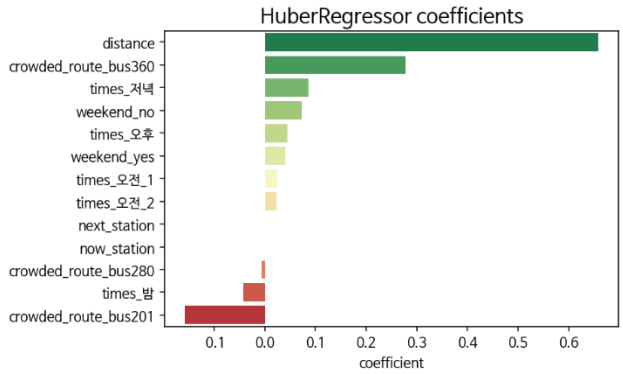
Huber Regression은 로버스트 회귀의 일종으로, 손실함수로 huber loss function을 이용하는 회귀분석 기법이다. 일반 선형회귀와 달리 huber loss function은 x축의 절대값이 |a| 보다 작으면 이차함수, |a| 보다 크면 일차함수로 정의한다. 이러한 손실함수의 특성 때문에 로버스트 회귀는 일반 선형 회귀보다 이상치(outlier)에 덜 민감하다.



[Huber loss function]

제주 버스 데이터를 huber regression 모델로 validation을 하면 세 가지 모델 중 RMSLE가 가장 높았다. 즉 이 모델이 성능이 가장 낮았다는 뜻이다. Huber regression의 성능이 일반 선형회귀보다 낮은 이유는 무엇일까? Data cleansing 파트에서 예측 성능을 높이기 위해 next_arrive_time이 700보다 큰 데이터를 이상치로 판단해 제거했다. 따라서 전처리가 된 데이터에는 이미 이상치가 없기 때문에 huber regression이 선형회귀보다 두드러지게 성능을 좋게 낼 수 없을 것이다. 위 표의 결과를 보면 huber regression과 선형 회귀의 성능 차이는 거의 비슷한 것을 볼 수 있다.

[Huber Regression 회귀계수 해석]



- 선형회귀와 마찬가지로 distance와 360 버스 여부의 회귀계수가 가장 컸다. 이에 대한 해석은 선형회귀와 동일하다.

- Huber regression은 다른 모델과 달리 평일/주말 구분 변수인 weekend의 회귀계수를 비교적 크게 예측했다. 하지만, EDA에서 weekend 여부에 따른 버스 운행 시간 차이는 거의 없다고 해석했다. 이러한 이유로 huber regression의 성능이 다른 모델보다 떨어진 것이라고 예상할 수 있다.

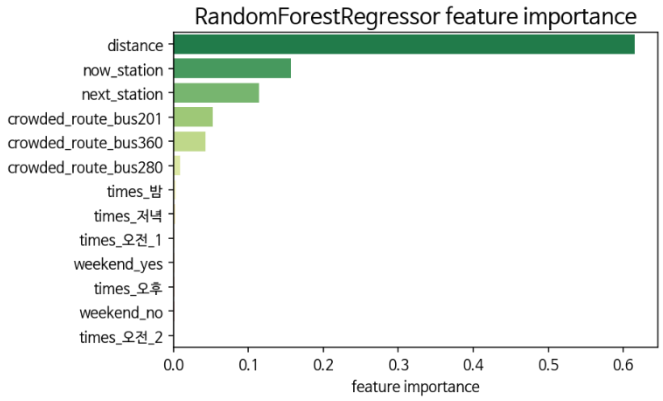
- 선형회귀와 달리 시간대를 나눈 변수 times에서 밤 시간대에 해당하는 회귀계수의 절댓값은 오전 시간대보다는 크고, 오후와는 비슷했다.

(3) Random Forest Regression

Random Forest Regressor는 앞의 두 모델이 회귀 기반인 것과 달리 트리 기반의 모델이다. 랜덤포레스트 분류와 알고리즘은 차이가 없지만, 회귀 트리는 리프 노드에 속한 데이터의 평균값을 구해 예측값을 구한다. 랜덤포레스트 회귀는 다른 두 모델보다 RMSLE가 작다. 그 이유는 linear regression 모델들과 달리 랜덤포레스트 회귀는 반응변수와 설명변수 사이에 선형성이 만족하지 않아도 되기 때문이다.

Random forest regressor로 예측을 하기 전, GridSearchCV를 이용해 하이퍼파라미터 튜닝을 진행했다. 여러 개의 파라미터 후보 중 n_estimatos=50, max_depth=10이 가장 높은 성능을 나타내는 파라미터였다. 이 때, max_depth를 지정해줌으로써 트리가 깊게 가지치기하는 것을 방지해 과적합(overfitting)을 막을 수 있다. 그리고 선형회귀와 달리 회귀계수가 아닌, feature importance(변수 중요도)를 추출하여 분석할 수 있다.

[Random forest regressor 변수 중요도 해석]



- 다른 모델들과 마찬가지로 distance의 중요도가 가장 크다. Distance 값이 버스 운행시간과 연관성이 크기 때문에 나타난 결과이다.

- 앞선 linear, huber regression과 차이나는 특징 중 하나는 랜덤 포레스트가 new_station과 next_station을 distance 다음으로 중요한 변수로 인식했다는 것이다. 정류장 정보는 버스 운행시간 예측에 중요하다. 본 분석에 이용한 데이터 자체가 정류장간 버스 운행시간을 예측하는 것이고, 각 정류장은 버스 노선 정보를 담고 있기 때문이다. 따라서 랜덤포레스트가 정류장 변수의 중요성을 파악했기 때문에 다른 두 모델보다 성능이 높은 것으로 예상된다.

- 버스 노선에 따른 중요도를 살펴보면 201번>360번>280번 순으로 크다. 이 역시 EDA 결과를 적절히 반영한 결과이다. 201번으로 시작하는 버스들은 평균 버스 운행시간이 타 버스들에 비해 짧고, 360번 버스들은 타 버스들에 비해 운행시간이 길다. 반면 280번은 평균 운행시간이 평균 부근에 있었는데, 랜덤포레스트의 변수 중요도 추출 결과는 이를 잘 반영한 것이라고 해석할 수 있다.

- 랜덤포레스트 모델이 선형 회귀 모델과 다른 점 중 두드러지는 특징은 시간대 정보를 거의 반영하지 않았다는 것이다. EDA 파트에서는 저녁시간대, 오전_1 시간대에 평균 버스 운행시간이 가장 길고, 밤에는 짧다고 분석을 했었다. 따라서 시간대 변수가 버스 운행시간 예측에 유의미한 영향을 끼칠 것으로 예상했으나 랜덤포레스트에서는 이 변수들을 중요하지 않다고 판단했다.

(4) 모델링 결과 요약

제주 버스 운행 데이터를 분석하기 위해 세 가지 모델을 사용했다. 그리고 그 성능을 분석한 결과, 본 데이터에는 Random Forest이 가장 우수한 성능을 보였기 때문에 최종 test set 예측에도 Random Forest Regressor을 사용하였다. Feature 수가 매우 많다면 상대적으로 많은 feature을 다루는데 유리한 선형 회귀가 더 성능이 좋았겠지만, 본 데이터에서는 카테고리형 변수들이 대부분이었기 때문에 Random Forest 모델이 다른 두 모델보다 성능이 좋은 것이라고 결론지을 수 있다.

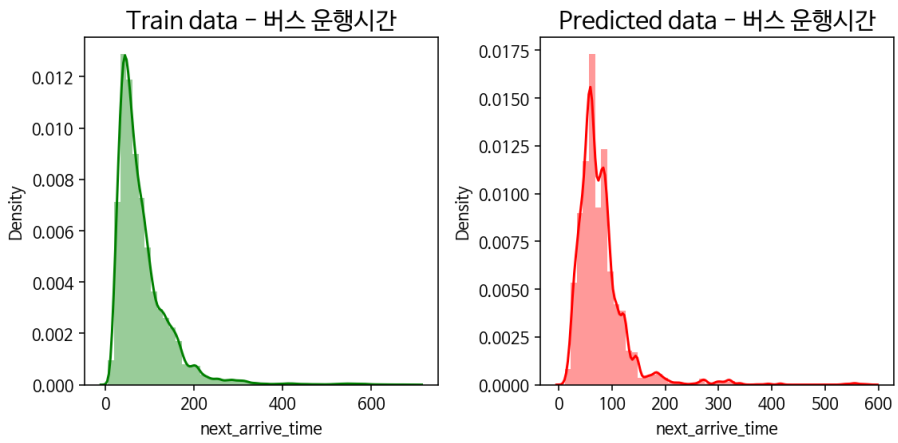
결과 비교에 앞서, feature engineering 파트에서 next_arrive_time에 로그변환을 하여 모델링을 했기 때문에 np.expm1 을 이용해 다시 원래 범위로 되돌린다. 이 때, 기존 next_arrive_time은 모두 자연수이므로, 예측된 next_arrive_time도 일의자리까지 반올림하여 형식을 맞춰준다.

IV. 결론 및 의의

1. 예측 결과 해석

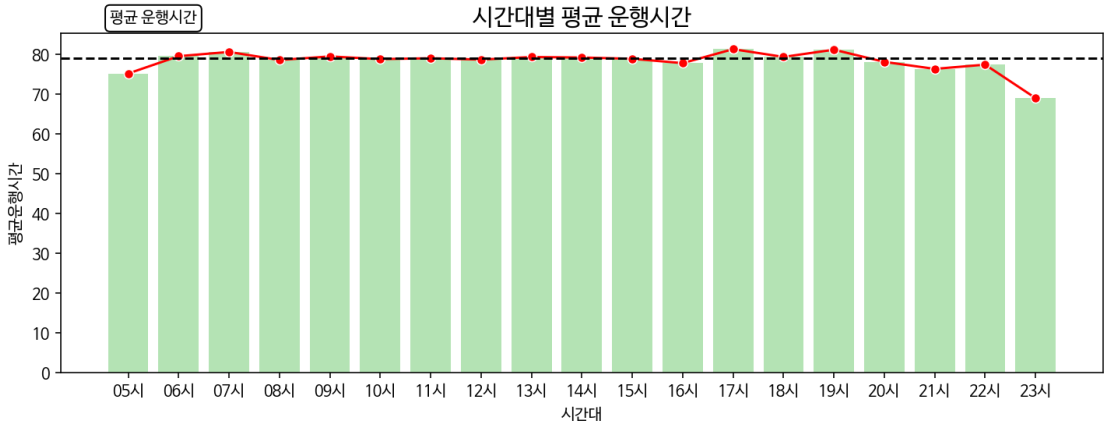
(1) 전체적인 버스 운행시간 분포 비교

Random Forest Regressor로 test set을 예측한 결과와 원본 train set의 버스 운행시간을 비교해보자. 아래 그림은 next_arrive_time에 대해 distribution plot을 그린 결과이다. 분포가 비슷하기 때문에 대체적으로 버스 운행시간을 잘 예측한 것으로 볼 수 있다. 하지만 train data에서 가장 밀도가 높은 지점이 예측된 test set에서는 다른 분포 모양을 보인다.



(2) 시간대별 버스 운행시간

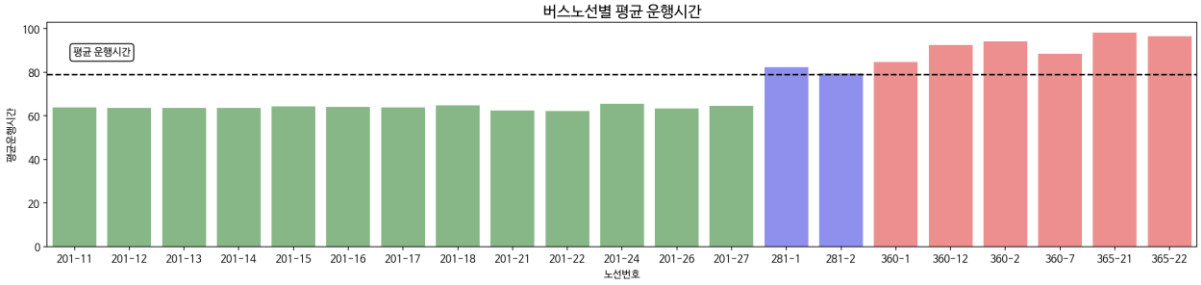
예측한 test set의 시간대별 버스 운행시간을 plot하고 그 결과를 train data와 비교해보자.



시간대별 평균 운행시간을 살펴보면 역시 오전 7시, 오후 5-7시에 제주 평균 버스 운행시간이 다른 시간대보다 높게 예측한 것을 알 수 있다.

(3) 버스 노선별 평균 운행시간

버스 노선별 평균 운행시간을 살펴보자. 이 그래프에서 201번 버스는 모두 초록색, 280번은 파란색, 360번 버스는 빨간색으로 나타냈다. 앞선 EDA 분석과 마찬가지로, 360번 노선의 평균 운행시간이 가장 길었고, 그 다음으로는 280번, 201번이 순서대로 길었다.



결론적으로 Random Forest Regressor을 이용해 test set을 예측한 결과, train set의 데이터들이 띄는 특징을 비슷하게 유지한 것을 알 수 있다.

2. 분석의 의의 및 한계

(1) 의의

- 다양한 종류의 회귀분석 모델을 사용해 예측 결과를 비교분석 했다.
- 제주 버스 운행 데이터를 이용해 다양한 파생 변수를 생성함으로써 예측에 유의미한 영향을 주는 feature을 발굴할 수 있었다.
- 랜덤포레스트를 이용한 예측 결과가 원본 데이터의 특징을 대체로 잘 유지했기 때문에 모델링 결과가 좋았다고 판단했다.

(2) 한계

- 전체적으로 모델들이 원본 train set보다 버스 운행시간을 더 길게 예측한 경우가 많았다. 이를 해결하기 위해 scaling 등 다양한 방법을 시도했지만 해결을 하지 못했다.

Appendix

1. 데이터 출처

https://dacon.io/competitions/official/229611/overview/

2. 참고문헌

권철민. 파이썬 머신러닝 완벽 가이드(위키북스)

허명회. 응용데이터분석(자유아카데미)

“제주 대중교통 이용객 6천만명 돌파…2000년대 들어 처음”, 2019년 1월 27일
https://www.yna.co.kr/view/AKR20190125062400056