DATAMINING FINAL REPORT

서울특별시 버스 노선 혼잡도 예측을 통한 다람쥐 버스 신규 노선 제안

제출일: 2019.06.19

과목명: 데이터마이닝

담당교수: 송종우 교수님

팀원: 고명지 박정민 유채연

TABLE OF CONTENTS

Problem Discussion

서울시의 출퇴근 및 버스 혼잡도

다람쥐 버스는 서울시에서 출퇴근 맞춤버스로 차내 혼잡이 주로 특정 구간과 특정 시간대에 극심하게 발생한다는 점에 착안해, 다람쥐 쳇바퀴 돌 듯 짧은 구간을 반복 운행하는 순환형 셔틀버스이다. 다람쥐 버스는 2017년 6월 말 처음 시행되었으며, 총 네 개의 버스 노선이 운행되었다. 각각의 신규 노선은 기존의 버스 노선 중 일부를 운행한다. 아래의 표는 각각의 운행 정보를 나타낸 것이다.

버스 노선	기존 노선	기.종점	대수	거리	배차간격	횟수	운행시간
8761	153	광흥창역 – 국회의사당역	4	8.5km	9-12 분	12	07:00 ~ 09:00 17:30 ~ 19:30
8771	702A.B	구산중학교 – 녹번역	4	7.7km	10-11 분	12	07:00 ~ 09:00
8551	500, 5535	봉천역 – 노량진역	5	12.3km	10-12 분	11	07:00 ~ 09:00
8331	3315	마천사거리 - 잠실역	5	12.1km	10-11 분	12	07:00 ~ 09:00

서울 다람쥐버스는 지속가능 교통도시평가에서 교통정책 우수사례부문 최우수정책으로 선정되었다. 네 대의 신규 노선은 반년 간의 시행으로 차내 혼잡을 39.5% 완화하는 동시에 전체 버스 이용객이 16.5%의 효과를 냈으며, 2018년에는 신규 노선이 세 개가 더 생겼다.

이번 프로젝트의 분석목표는 다람쥐 버스의 순승차인원을 예측하여 버스 노선 별 혼잡도를 계산한 후 새로운 다람쥐 버스 노선을 제안하는 것이다.

DATA 설명

데이터 설명서

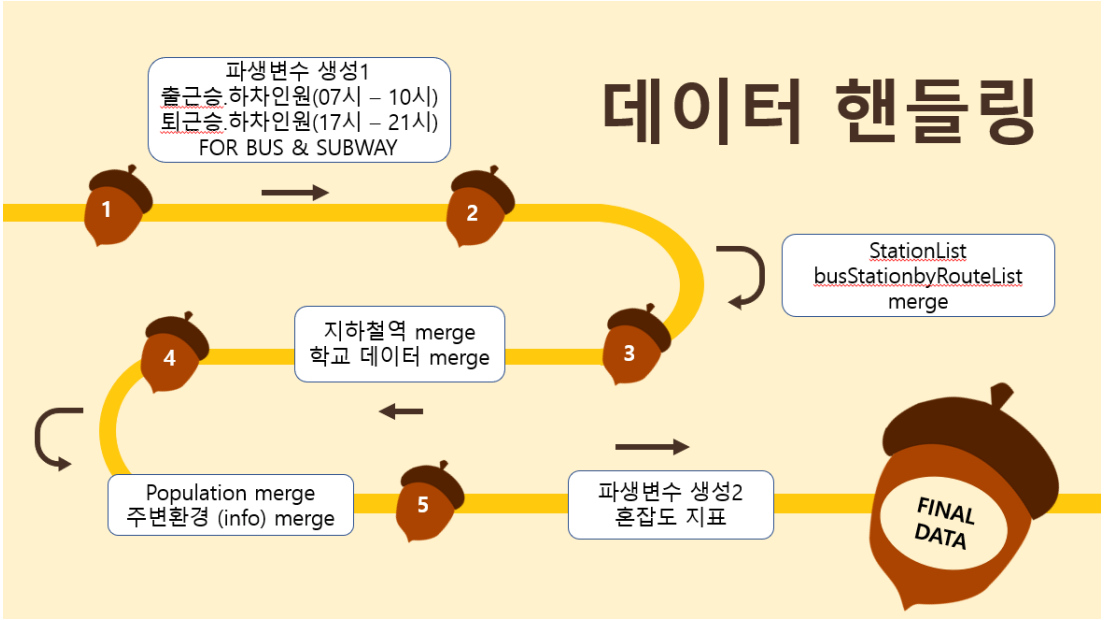
분석에 사용한 데이터는 크게 7 가지이다. 이 중 가장 기본이 되는 데이터는 공공데이터포털을 통해 얻어온 서울시 버스 노선별 정류장별 시간대별 승하차 인원 정보이며, 나머지 자료를 join 시켜 하나의 데이터 형태로 만들었다.

데이터명	Source	비고
서울시 버스노선별, 정류장별, 시간대별 승·하차 인원 정보	공공데이터포털(file)	기본 데이터
stationList	TOPIS	정류장별 위경도 자료 사용
busStationbyRouteList	공공데이터포털(API)	노선 및 정류장 별 정보 이용

서울시 동별 사업체 및 종사자 밀도 통계	서울열린데이터광장	정류장 주변 인구 밀도 정보
구/동 별 주거 지역, 녹지 지역, 상업지역	서울열린데이터광장	정류장 주변 환경 정보
지하철 역별 시간대별 승하차 인원수	공공데이터포털(file)	가까운 지하철역 승하차 수
학교 정보	공공데이터포털(file)	학교 별 학생 및 교직원 수

데이터는 다람쥐버스가 새로 도입된 이후인 2017 년 7 월부터 새로운 다람쥐버스 노선이 생기기 이전인 2018 년 3 월까지의 자료를 사용했다. 이 중 2017 년 7 월부터 12 월까지를 train 으로 2018 년 1 월부터 3 월까지의 자료를 test 로 설정하여 분석을 진행했다.

데이터 핸들링



『서울시 버스노선별, 정류장별, 시간대별 승·하차 인원 정보』 데이터에서 버스와 지하철의 출퇴근 시간대의 자료만 추출해서 월별/시간별 평균을 구한 자료를 사용하였다. 각 셀의 값은 일별 버스 또는 지하철의 승·하차 인원을 의미한다. 월별 평균을 사용한 이유는 월별로 승·하차 인원에 거의 차이를 보이지 않았기 때문이다. 다음으로는 버스정류장의 위치정보를 추가하기 위해서 『StationList』와 『busStationbyRouteList』 데이터를 사용하여 각 버스정류장의 고유번호인 ARS ID 를 기준으로 위도와 경도를 붙였다. 또한 추가한 위도와 경도를 기준으로 각 버스정류장에서 가장 가까이 위치하고 있는 지하철역을 찾고 환승 가능한 지하철노선의 개수와 지하철의 승·하차 인원을 추가하였다. 여기서는 『지하철 역별 시간대별 승·하차 인원수』 데이터를 사용하였다. 다음으로 『학교 정보』 자료를 사용하여 가장 가까이 위치하고 있는 학교를 찾고 그 학교의 재학생 수를 세서 데이터에 추가하였다. 다음으로 버스정류장이 위치하고 있는 곳의 인구정보를

추가하기 위해 『서울시 동별 사업체 및 종사자 밀도 통계』 데이터를 사용하였다. 여기서 데이터를 병합할 때 구와 동을 기준으로 병합을 하려고 하였는데 구와 동 정보가 정확히 일치하지 않아서 제대로 병합을 할 수가 없었다. 정확한 구와 동을 얻기 위해 네이버지도[1]에서 버스정류장의 ARS ID 를 입력해서 구와 동 정보를 가져올 수 있도록 하는 크롤링 코드를 짰다. 이렇게 해서 구한 구와 동을 다시 ARS ID 를 기준으로 하여 데이터셋에 병합하고 새롭게 구한 구와 동을 기준으로 하여 인구정보를 붙여주었다. 마지막으로 『구/동 별 주거 지역, 녹지 지역, 상업지역』 데이터를 이용하여 각 버스정류장의 반경 500m 이내의 지구 수를 더한 후 병합하였다.

데이터 변수 소개

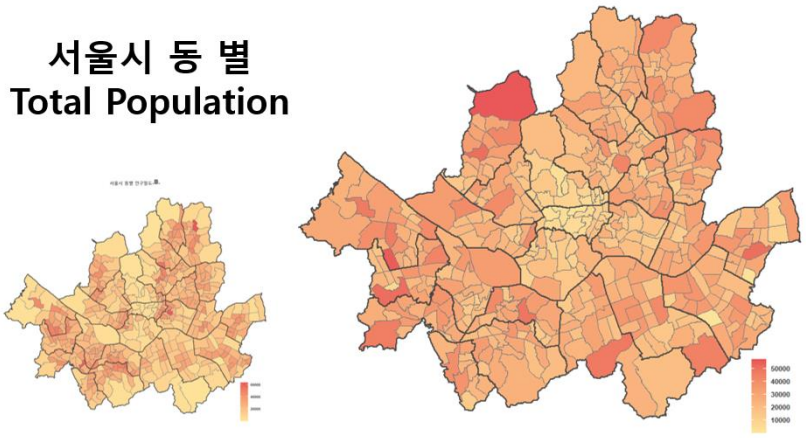
Variable	Class	데이터 설명서
busRouteNm	Factor	버스노선번호
arsId	Integer	정류장번호
lng, lat	Numeric	정류장의 위, 경도
morn_ride, morn_off	Numeric	출근 승·하차 인원 평균
even_ride, even_off	Numeric	퇴근 승·하차 인원 평균
stStationNm, edStationNm	Factor	기점 정류소 이름, 종점 정류소 이름
term	Integer	배차 간격
direction	Factor	방향
transYn	Factor	회차 여부
fullSectDist	Integer	정류소 간 거리
busnum	Numeric	시간 당 평균 운행 대수
seq	Integer	노선 순서
morn_full	Numeric	출근 승차인원 – 출근 하차 인원
even_full	Numeric	퇴근 승차인원 – 퇴근 하차 인원
cnt	Integer	정류소 당 지나가는 버스노선 개수
length	Numeric	전체 경로 거리
cong_morn	Numeric	출근 혼잡 지표

[1] https://map.naver.com/

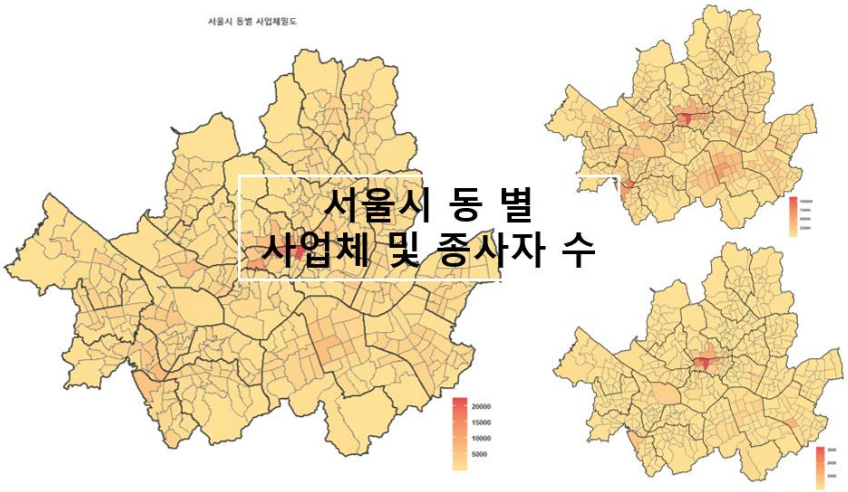
cong_even	Numeric	퇴근 혼잡 지표
gu, dong	Factor	정류소가 위치하는 구, 동
사업체수. 밀도	Numeric	정류소가 위치하는 구, 동의 사업체 수 및 밀도
종사자수. 밀도	Numeric	정류소가 위치하는 구, 동의 종업자 수 및 밀도
인구수. 밀도	Numeric	정류소가 위치하는 구, 동의 인구수 수 및 밀도
subway	Factor	정류소에서 가장 가까운 지하철역
n_subway	Integer	500m 이내 지하철역 수
mean.smr	Numeric	지하철 출근
mean_smo	Numeric	지하철 출근 하차 인원 수
mean_ser	Numeric	지하철 퇴근 승차 인원 수
mean_seo	numeric	지하철 퇴근 하차 인원 수
info 1	Integer	녹지 지역 count
info 2	Integer	상업 지역 count
info 3	Integer	주거 지역 count

DATA EDA

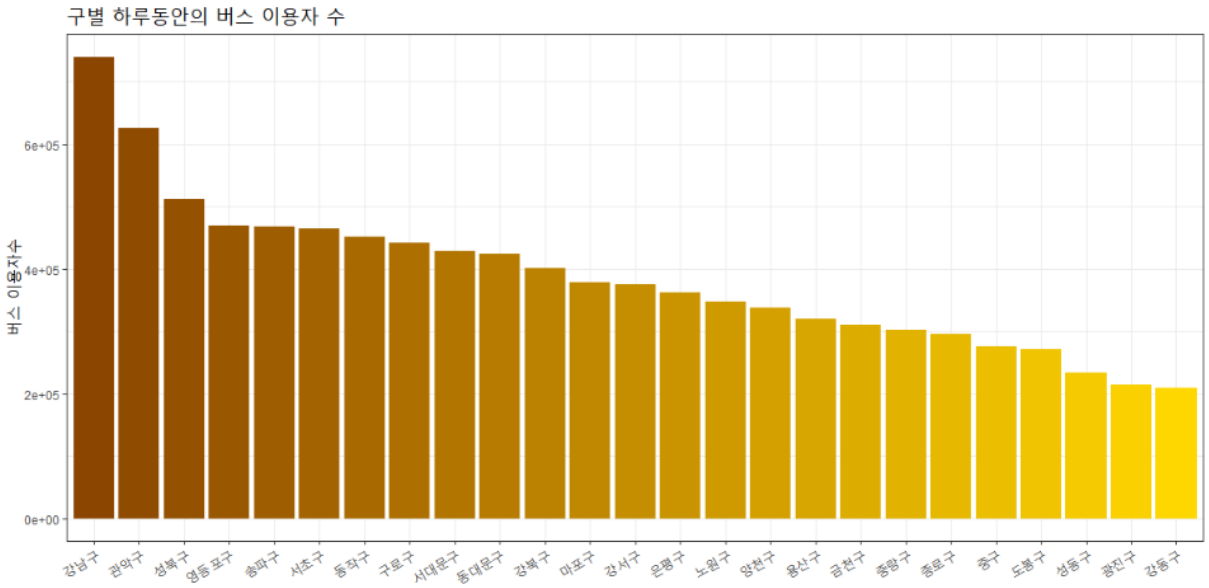
핸들링까지 마친 자료를 이용하여 몇 가지 그래프들을 통해 변수들을 표현해보았다. 크게 구, 동별 특징, 정류장 주변 환경, 버스 이용객 수를 시각화해 보고자 한다. 가장 먼저 서울시 내의 동별 인구수와 그 인구밀도를 수치화한 것을 이용해 다름 그림으로 표현하였다. 색이 짙을수록 그 밀도 및 수가 큰 것인데, 몇 지역을 살펴보자면 은평구, 강서구의 인구수가 많으며 구로구 양천구에서 인구 밀집도가 큰 것으로 보인다.



위와 동일하게 사업체 및 종사자 수와 밀도를 아래 그림으로 표현했다. 마찬가지로 색이 진할수록 그 수 또는 밀도가 큰 것인데, 이 경우 중구, 종로구, 강남구 정도가 종사자 밀집 지역으로 나타났다.

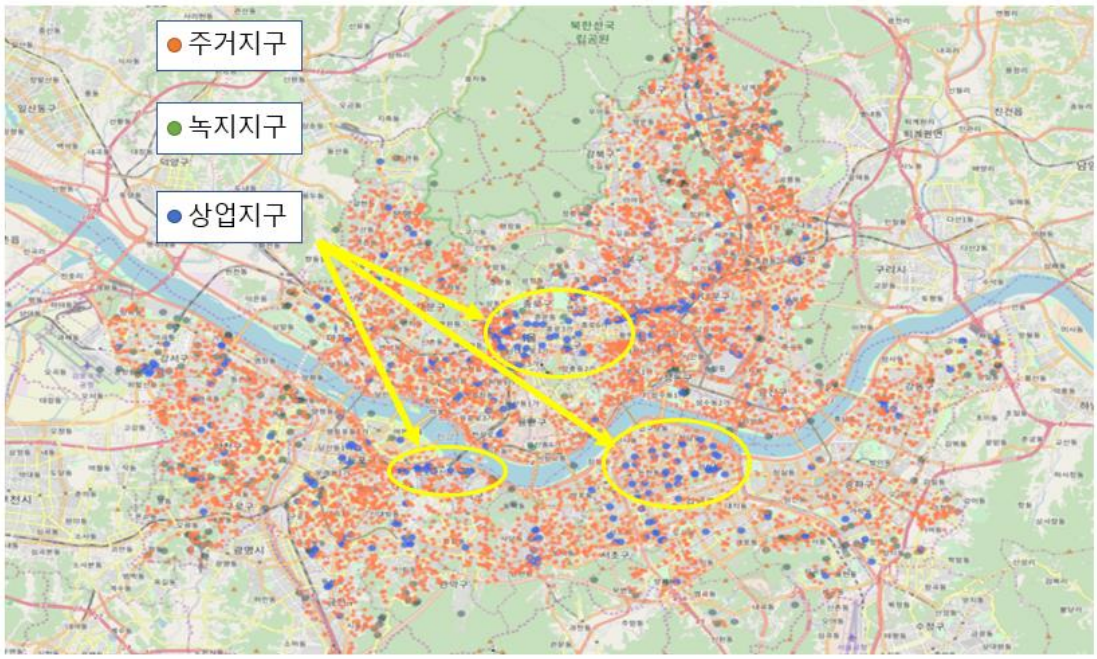


따라서 구 별 버스 이용객과 이러한 인구 통계의 연관성을 살펴보기 위해 하루 동안의 버스 이용자 수를 막대그래프로 그려보았다. 여기서는 강남구와 관악구가 가장 많은 이용자수를 보여주고 있다. 위에서 언급했던 구 중에서는 강남구 외에 특출하게 큰 이용객수를 볼 수 없으며 구 별 특징을 추출하기 어렵다고 판단된다.

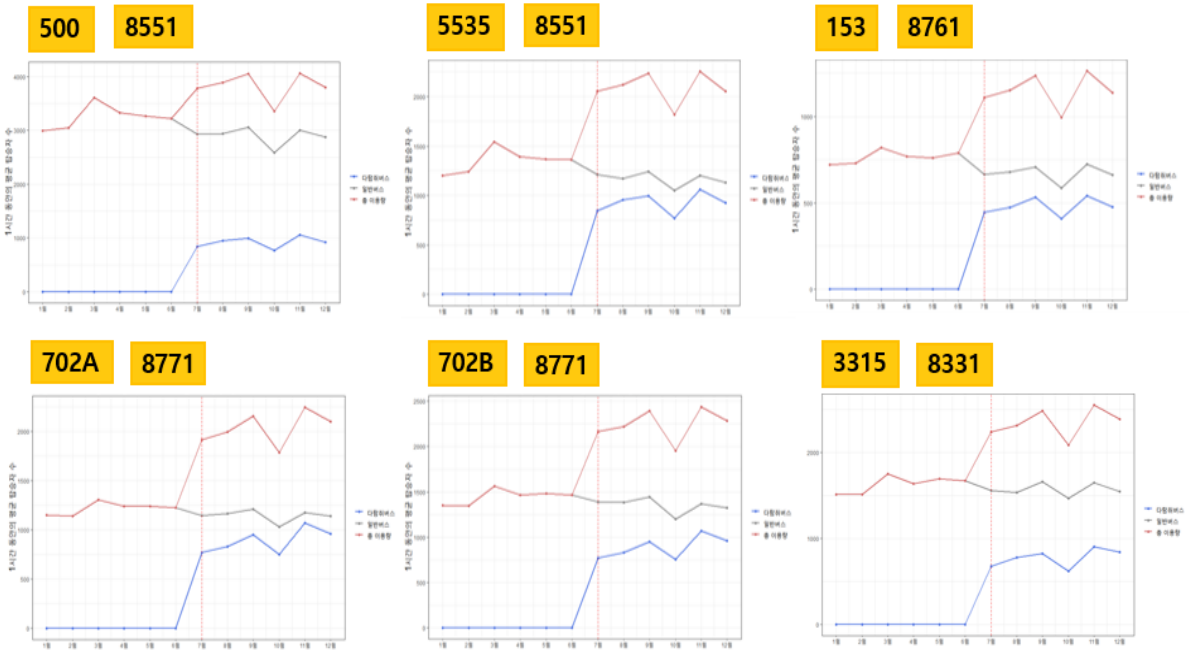


주변 환경 변수는 총 세 가지 종류의 지구로 이루어져 있다. 아파트 및 주택들에 해당하는 주거지구, 공원 등의 녹지지구, 그리고 사업체에 해당하는 상업지구가 있다. 이 세 가지를 서로 다른 색의 점으로 표현했는데 빨간색 점은 주거지구, 녹색 점은 녹지지구, 그리고 마지막으로 파란색 점은 상업 지구를 나타낸다. 이 중 파란색 점들이 매우 집중되어 찍혀 있다는 것을 볼 수 있다. 해당 그림과 같이 몇 가지 변수에 대하여 군집 형태의 데이터를 형성해 볼 수 있다는 판단 하에, 아래에서 모델링을 하기에 앞서 클러스터링을 먼저 시행해볼 수 있다.

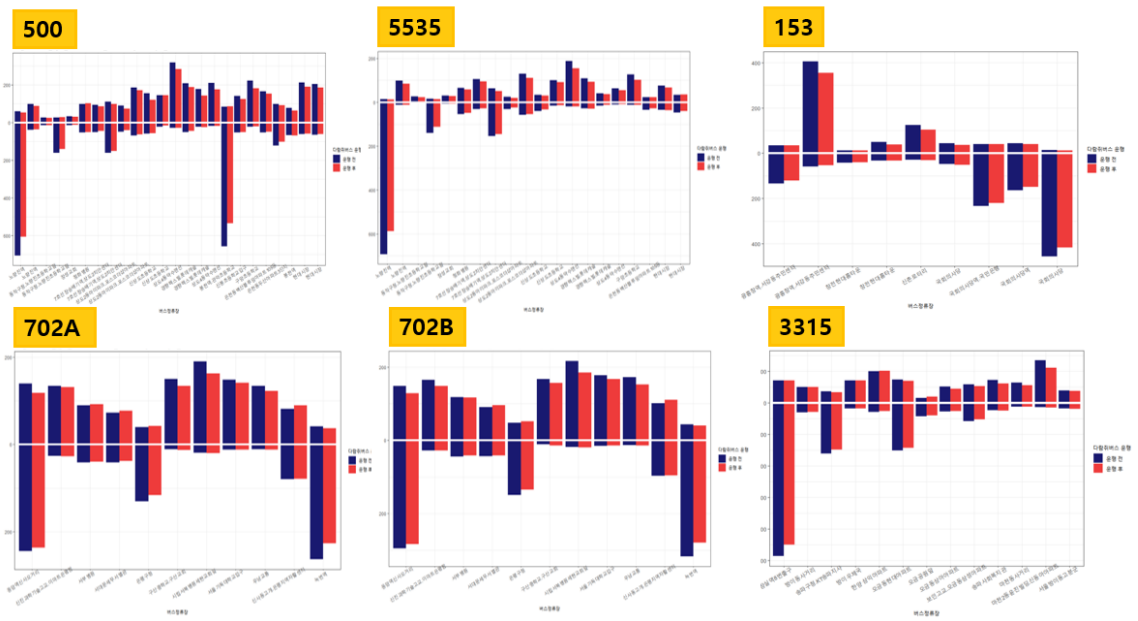
주변 환경 Information



다음은 기존 버스 노선들에 대하여 월별 버스 이용객 수를 나타낸 그래프이다. 출근 시간을 7시부터 10시라고 했을 때, 한 시간 동안 평균적으로 버스를 이용하는 사람들의 수를 월별로 나타낸 것이다. 여기서 가장 아래의 파란색 선이 6월부터 새로 도입된 다람쥐버스의 이용객, 빨간색 선이 기존 버스 노선의 이용객, 그리고 마지막으로 검정 선이 기존과 다람쥐 버스 두 노선의 이용객 합을 보여준다. 6월을 기준으로 확연히 전체 이용객 수가 늘어나고 있으며 기존 버스 노선의 이용객 수가 줄었다는 것을 육안으로 확인 가능한다. 이 중 10월에 대해서는 모든 노선이 일정 수준 감소했는데, 이는 2017년의 10월 공휴일로 인함이라고 설명 가능하다.



마지막으로 다람쥐 버스로 이용되는 기존 버스 노선의 정류소들의 Before & After를 보여주는 그림이다. 파란색 bar가 다람쥐 버스 시행 이전, 빨간색이 그 이후이며, 가운데의 흰색 구분선을 기준으로 위쪽이 승차, 아래쪽이 하차를 보여주고 있다. 전체적인 이용객 수 감소를 볼 수 있으며, 몇 개의 정류장에서 현저하게 그 이용률이 낮아지는 것을 통해, 다람쥐버스의 효과를 시각적으로 확인할 수 있다.

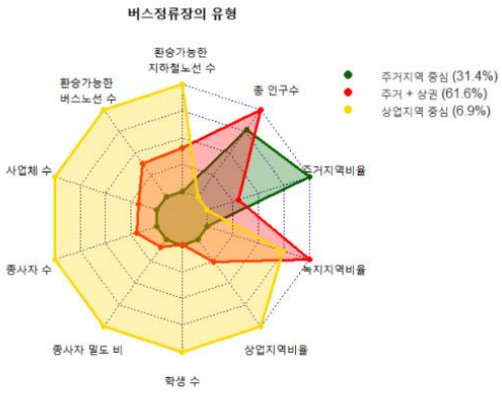


DATA MODELING

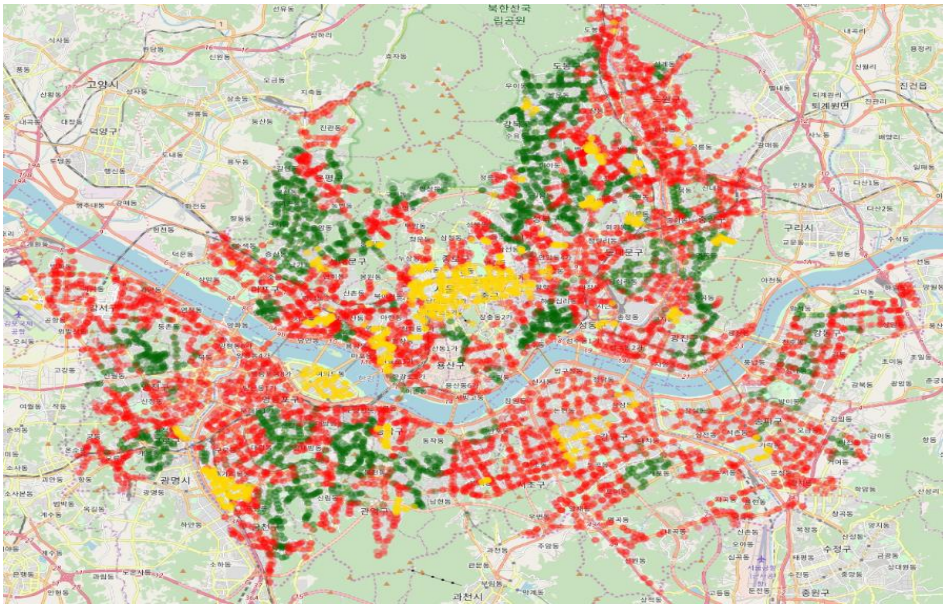
MCLUST

우리가 예측할 y변수는 순승차인원(= 승차인원 – 하차인원)이며 설명 변수로는 모델링마다 차이가 있지만 대체로 수치형(numerical) 변수들을 사용했다. 다양한 모델링 기법을 적용시키기에 앞서, 앞의 EDA에서 봤듯이 정류소별로 데이터의 군집화를 먼저 시행을 했다.

Mclust 함수를 이용하여 정류소별 인구 정보와 환경 변수들로 전체 데이터를 크게 세 개의 군집으로 나누었다. 이 때 사용된 변수는 학생수, 사업체 수, 종사자 수, 인구 수, 근처 지하철 수, 환승 가능한 버스의 대수, 그리고 주변 환경 변수이다. 각 군집의 특징을 살펴보기 위해 오른쪽의 레이더 차트를 그려보았다. 각 class의 특징을 살펴보았을 때, class 1은 주거 밀집 지역의 특징을 가진 정류소라고 할 수 있으며, class 2는 주거와 상업이 적절히 혼합된 지역, 마지막으로 class 3는 상업 밀집 지구의 특징을 가진다. 이 세 개의 클래



스를 아래의 지도에 표시해보았을 때 노란색점으로 표시된 곳이 class3이며, 위의 EDA에서 그렸던 상업지구 밀집 지역과 상당히 일치한다는 것을 알 수 있다. 강남구, 중구, 종로구, 영등포구 등 흔히 일컬어지는 '상업이 밀집된 지역'을 잘 분류했다고 볼 수 있다.



class	n_student	사업체수	종사자수	totpop	n_subway	cnt	종사자밀도비	info1	info2	info3
1	870	1283	4674	25722	1	2.60	18.3	0	0.437	15.2
2	825	2242	15542	27629	1.25	3.14	61.7	0.8	0.917	13.1
3	6426	6573	59899	19178	1.60	4.03	532	0.595	2.31	12.1

모델링에서 2017년 자료를 train, 2018년 자료를 test로 사용했고, 이전에 두 데이터 간의 차이를 보정해줄 필요가 있었다. 두 자료를 비교해 보았을 때 신규 노선 및 정류소들이 존재했기 때문에, 해당 신규 데이터에 대하여 위의 클러스터 특징들을 기준으로 세 개의 class로 직접 배정을 한 후 모델 예측에 사용하였다. 가장 먼저 종사자 밀도비가 100을 넘어가면 class 3로 넣은 뒤, 나머지 정류소들 중 info3, 즉 주거지역이 15 이상이면 class 1, 나머지는 class 2로 설정했다.

arsId	n_student	사업체수	종사자수	totpop	n_subway	cnt	종사자밀도비	info1	info2	Info3	Class
5231	656	783	5705	11889	1	2	48.19	0	1	**16**	1
9277	850	1962	8026	31274	1	2	26.50	0	2	**31**	1
12821	1013	1844	4934	26632	1	1	18.52	0	1	**18**	1
23387	270	448	8210	24927	1	2	32.80	2	0	**16**	1
5187	759	1711	8543	24738	1	3	34.94	0	1	13	2
8348	1098	921	4242	27347	1	2	15.47	3	0	11	2
10292	917	1277	7384	28694	1	1	25.56	2	0	7	2
16294	942	2657	23388	33991	1	2	68.38	2	0	15	2
18993	667	4085	24308	395456	1	1	61.66	0	0	4	2
23285	811	15590	146000	36499	1	4	**399.64**	0	3	8	3
37501	896	5271	49006	23876	1	2	**205.56**	1	0	4	3

위에서 구한 세 개의 클래스로 출근과 퇴근을 각각 나누어 총 6개의 자료를 생성했다. 이를 이용해서 아래의 Linear Regression, LASSO, Ridge, Random Forest, Bagging, 그리고 XGBoost, 총 6개의 모델링을 실시해 그 결과를 아래에 첨부하였다.

LINEAR

Linear 모형에 쓰인 변수들은 다음과 같다.

lng	lat	gu	totpop	사업체수	종사자수
평균종사자.	사업체밀도.	종사자밀도.A..	일구밀도.B.	종사자밀도비.A.B.100.	n_subway
mean.smr	mean.smo	n_student	info1	info2	info3
length	routeType	term	seq	transYn	fullSectDist
busnum	morn_full	cnt			

우선 설명변수를 하나도 넣지 않은 경우에 6개의 군집에 대한 RMSE값은 아래 표와 같다.

기본	1	2	3
Morning	68.18492	71.42654	64.87197
Evening	55.35715	57.0989	51.55706

위의 모든 변수를 사용하여 선형모형을 만들고 RMSE를 구한 결과는 아래와 같다. 변수를 하나도 사용하지 않았을 때와 비교했을 때 RMSE값이 크게 달라지지 않았다는 것을 볼 수 있다. 이를 통해 선형 모형은 설명력이 크지 않다고 할 수 있다.

	MODEL	RMSE	RMSE.TEST
1	**morn_class1**	68.5238	65.5347
2	**morn_class2**	72.6034	69.5933
3	**morn_class3**	63.2908	61.5674
4	**even_class1**	54.9119	53.4160
5	**even_class2**	56.7045	55.4419
6	**even_class3**	53.4801	49.9626

RIDGE

Ridge 모형에서도 선형 모형에서 사용한 것과 동일한 변수를 사용하였다. 결과는 아래와 같이 나타난다. 앞서 본 선형 모형에 비해서 RMSE값이 약간 줄어들었지만 여전히 설명력이 좋지 않음을 알 수 있다.

	MODEL	RMSE	RMSE.TEST
1	**morn_class1**	66.1919	63.1130
2	**morn_class2**	69.609	66.9450
3	**morn_class3**	58.4669	56.9427
4	**even_class1**	54.0934	52.5154

5	**even_class2**	55.4906	54.1638
6	**even_class3**	52.3450	48.4804

LASSO

Lasso 모형에서도 선형 모형에서 사용한 것과 동일한 변수를 사용하였다. 결과는 아래와 같이 나타나며 앞서 살펴본 선형 모형과 Ridge 모형과 비교했을 때 RMSE값이 줄어들기는 했지만 여전히 50대에서 60대의 값으로 설명력이 좋지 않다는 것을 알 수 있다.

	MODEL	RMSE	RMSE.TEST
1	**morn_class1**	67.8731	64.8089
2	**morn_class2**	69.5668	66.9140
3	**morn_class3**	57.2459	55.8003
4	**even_class1**	54.3965	52.8427
5	**even_class2**	55.8341	54.5409
6	**even_class3**	51.5036	47.4188

RandomForest

랜덤포레스트와 배깅 모델에서는 수치형 변수들만 사용해서 모델링을 실시했다. 그리고 우선 랜덤포레스트에서는 mtry와 ntree의 tuning parameter 값을 각각 5에서 15, 50에서 200까지 그 수를 바꿔가며 6개의 데이터의 모델링을 했으며, 각각 가장 좋은 결과를 찾아 아래의 표로 입력했다. 앞선 선형 회귀, Ridge, Lasso 모형에 비해 train의 rmse값은 4-50으로 낮아진 것을 볼 수 있다. 여기서 특이한 점은, 랜덤포레스트는 오버피팅이 잘 되는 모델로 흔히 알려져있는 반면, 우리의 데이터에서는 test rmse의 값이 훨씬 작은 것으로 보인다.

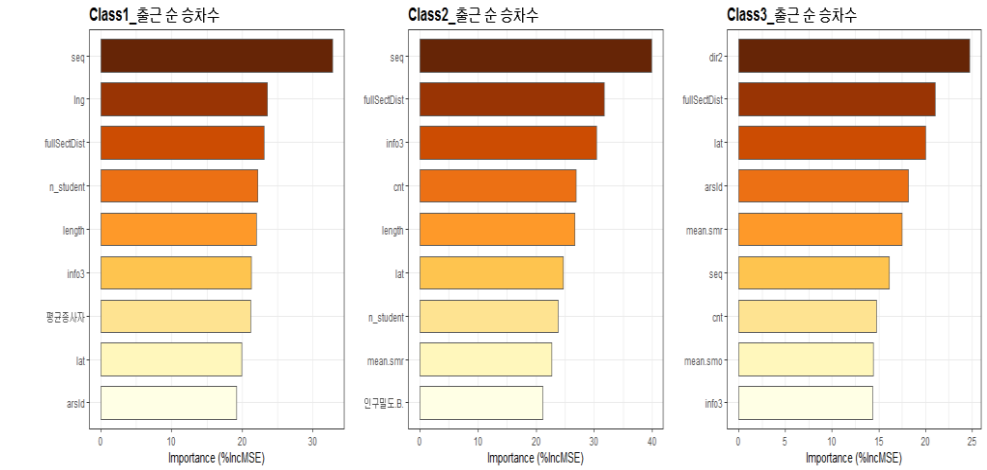
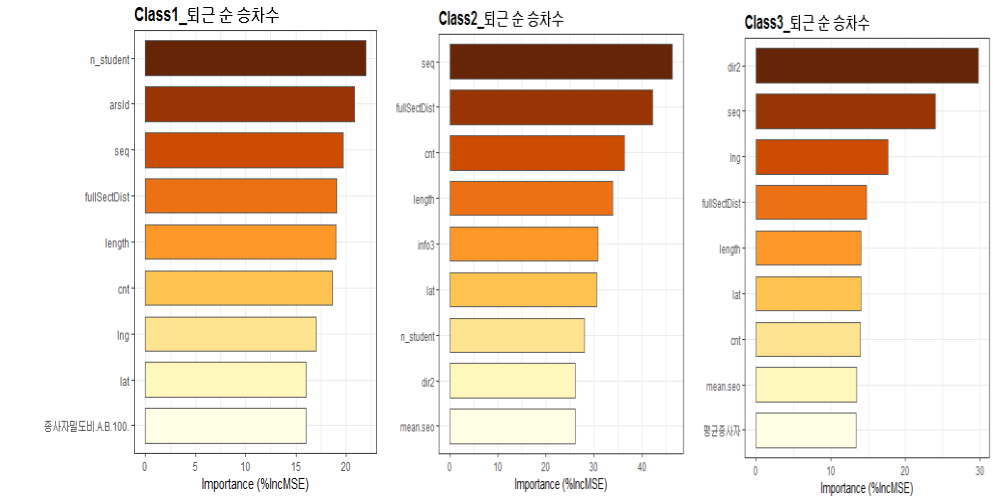
	MODEL	mtry	ntree	rmse	rmse_test
1	**morn_class1**	15	200	57.1346	26.72477
2	**morn_class2**	15	200	61.21784	28.54134
3	**morn_class3**	15	200	52.78453	26.91519
4	**even_class1**	15	100	45.94371	22.19747
5	**even_class2**	15	100	48.32353	23.66954
6	**even_class3**	15	200	44.31618	23.47896

Bagging

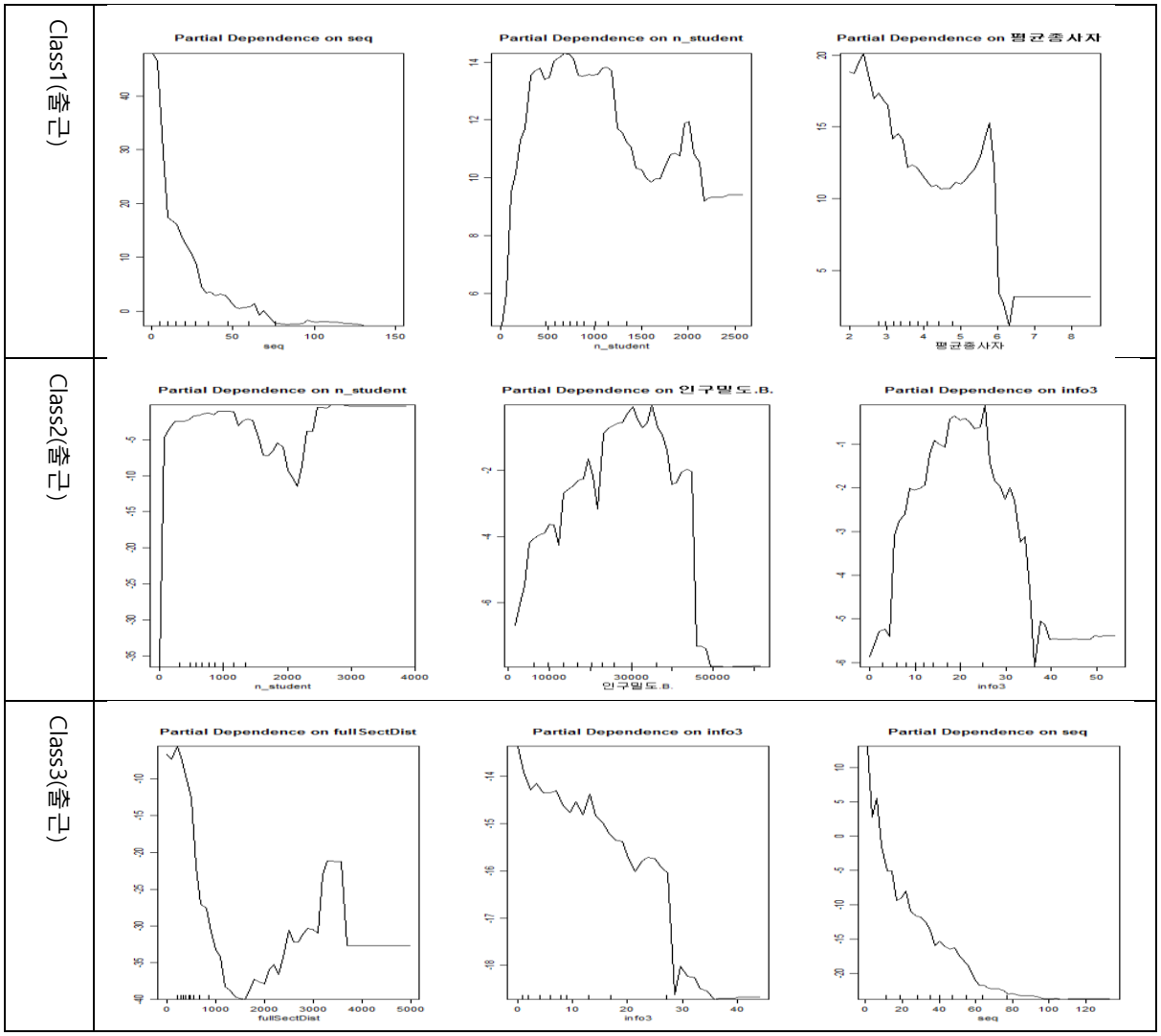
랜덤포레스트에서와 마찬가지로 수치형 변수들만을 사용하고, ntree tuning paremter를 50, 100, 200으로 바꾸어 가며 6개의 데이터의 모델링을 진행했고, 각각 가장 좋은 결과물을 아래의 표로 나타냈다. 결과적으로 위의 랜덤포레스트와 그 결과가 매우 흡사하다.

	MODEL	ntree	rmse	rmse_test
1	**morn_class1**	**200**	**57.19326**	**28.01503**
2	**morn_class2**	200	61.08975	29.9069
3	**morn_class3**	100	54.91622	28.40684
4	**even_class1**	200	45.54813	23.87543
5	**even_class2**	200	50.07567	25.69611
6	**even_class3**	100	47.19759	25.3292

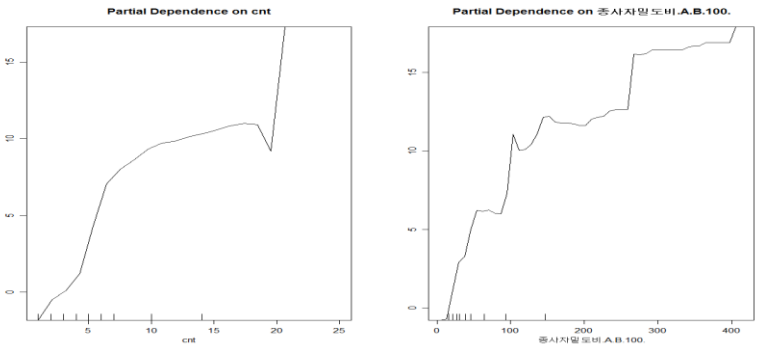
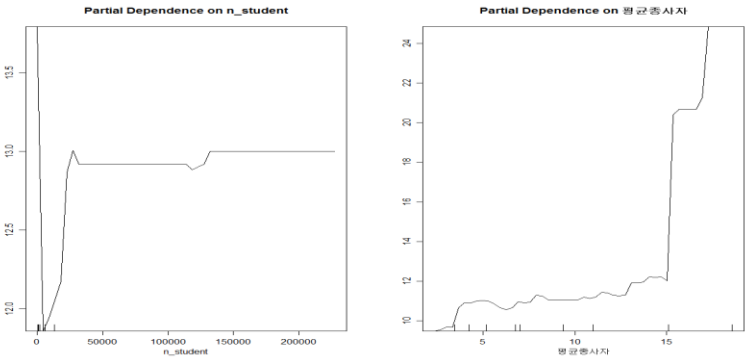
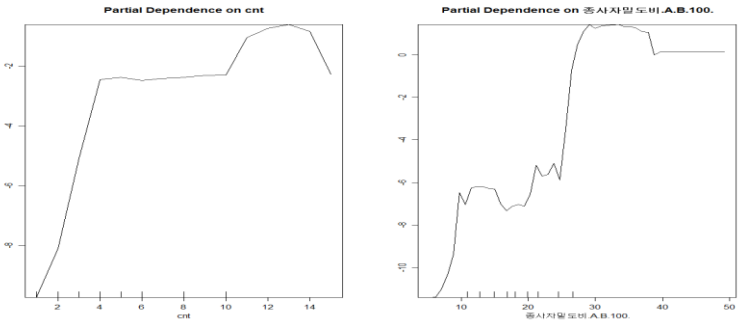
각 모델 별 변수 중요도 그래프를 그리면 다음과 같다.

출근	
퇴근	

출근 및 퇴근 시간대에서 순 승차 수를 예측할 때 대체로 seq(정류장순서), fullsectdist(정류소별 거리) 등이 중요하다고 나왔다. 각 데이터별로 PDP(Partial Dependent Plot)을 몇 개 골라서 그려보면 다음과 같다.



대체로 seq가 커질수록 순승차인원이 줄어드는 것을 볼 수 있다. 출근시간을 보면 class1에서 지하철의 수가 많아질수록 순승차인원이 커지다가 어느 정도 이후로 그 수가 줄어든다. 평균종사자의 수가 많아질수록 순승차인원이 줄어드는 것을 보아, 유의한 변수로 작용할 수 있다고 생각된다. Class2에서는 학생수, 인구밀도, 그리고 info3에 대하여 PDP를 그려봤을 때 위 그림과 같이 나왔다. 세 변수 중 인구 밀도 또는 info3에서 각각 수가 늘어날수록 순승차인원이 커졌으며, 어느 정도 이후로는 급격히 감소하는 형태를 볼 수 있다. 마지막으로 class3에서는 순서대로 정류장 간 거리, info3, 그리고 seq를 각각 그려보았을 때, 각각의 수가 커지면서 순승차인원이 줄어든 것을 볼 수 있다.

<table>
<tr><td>Class1(퇴근)</td><td> 
</td></tr>
<tr><td>Class2(퇴근)</td><td> 
</td></tr>
<tr><td>Class3(퇴근)</td><td> 
</td></tr>
</table>

퇴근 시간대 순 승차수의 변수 별 PDP 를 보면 마찬가지로 seq 및 종사자 수가 유의한 변수로 나왔다. 대체로 각각의 수가 커지면서 순승차인원이 증가함을 알 수 있다.

XGBOOST

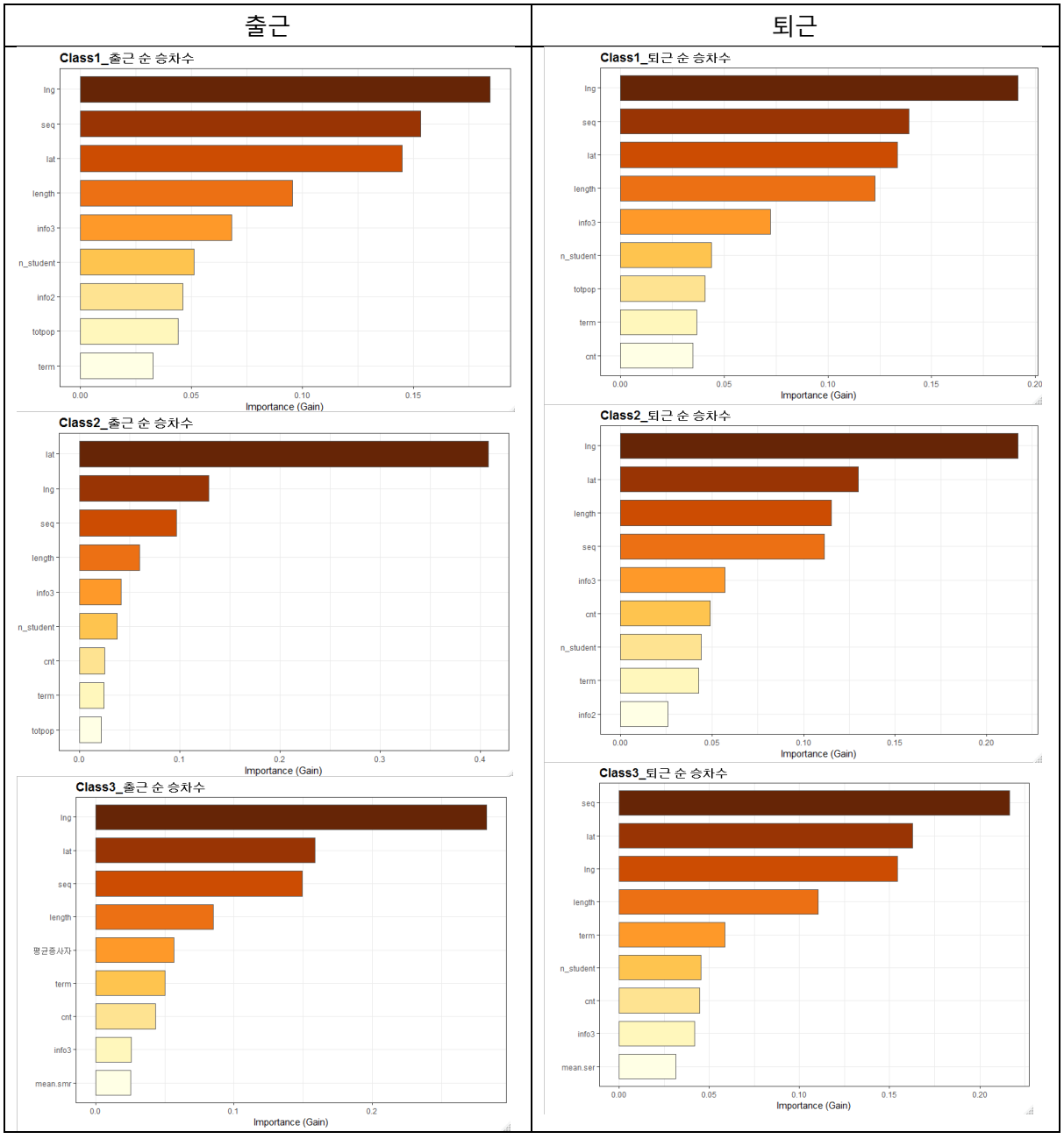
XGBoost에 쓰인 변수들은 다음과 같다.

lng	lat	사업체수	종사자수	평균종사자	사업체밀도
종사자밀도.A.	인구밀도.B.	종사자밀도비.A.B.100.	n_subway	mean.smr(mean.emr)	mean.smo(mean.emo)
n.student	info1	info2	info3	length	term
seq	busnum	cnt	routeType	totpop	

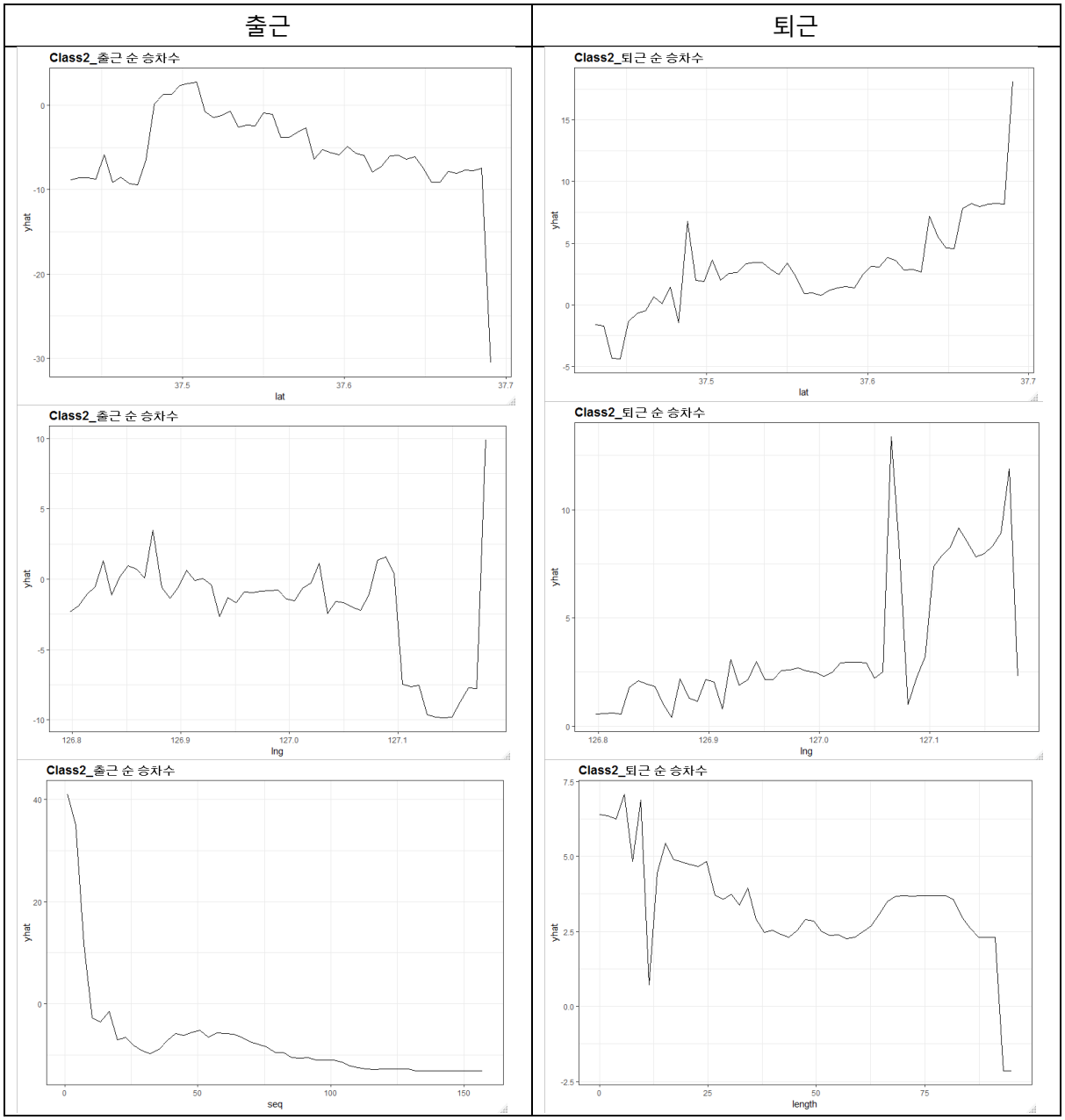
xgboost 에 쓰인 tuning parameter 로는 max.depth, eta 값을 조정해 주었다. max.depth 는 10, 30, 50 을 사용하였고, eta 값은 0 부터 1 까지 0.1 간격을 주어 사용하였다. 이렇게 총 30 개의 parameter 를 조정해주며 test error 를 비교하여 가장 작은 parameter 를 선택하였다. 그 결과는 다음과 같다.

	MODEL	MAX.DEPTH	ETA	RMSE	RMSE.TEST
1	morn_class1	30	0.2	4..114	12.452
2	morn_class2	50	0.5	0.207	16.623
3	morn_class3	30	0.1	0.904	16.569
4	even_class1	10	0.4	5.142	14.523
5	even_class2	50	0.1	0.328	11.215
6	even_class3	10	0.3	3.344	16.182

위에서 살펴본 결과 다른 모델들에 비해 train rmse와 test rmse가 월등히 작음을 확인할 수 있다. 각 모델별 변수 중요도 그래프를 그리면 다음과 같다.



출근 및 퇴근 시간대에서 순 승차 수를 예측할 때 모든 클러스터에서 위도와 경도가 굉장히 유의함을 볼 수 있다. 버스가 다닐 때 어떤 방향으로 움직이는가가 중요할 수도 있고 해당 정류소 주변 지역정보가 유의할 수도 있다. train rmse 와 test rmse 가 가장 작았던 두 번째 cluster 에 대하여 partial dependence 그래프를 그려보면 다음과 같다.

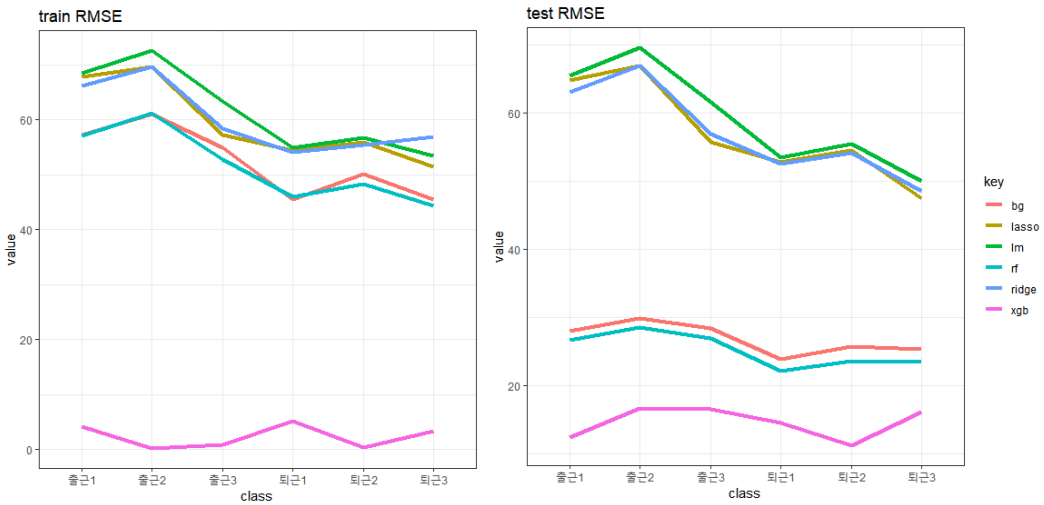


먼저 출근 시간대 순 승차수의 변수별 PDP 를 보면 위도와 경도는 대부분 값에 상관없이 비슷한 값을 가지지만 특정 값일 때 순 승차수가 많아지거나 적어짐을 볼 수 있다. 위도는 37.5 일 때 순 승차수가 급증하였다가 그 이후로는 감소하는데 37.7 에 가까워졌을 때 갑자기 하락하였다. 경도의 경우 127.1 정도까지는 거의 0 에 가까운 값을 가지다가 127.15 정도까지는 갑자기

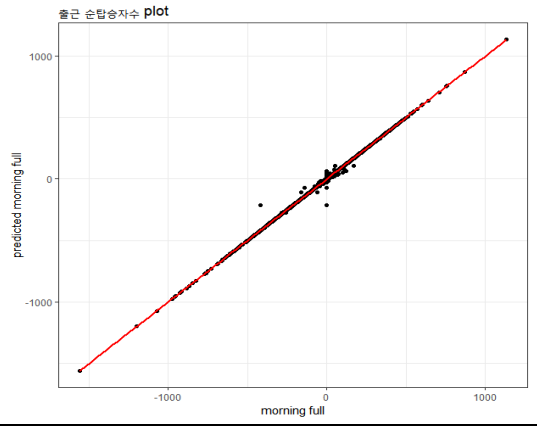
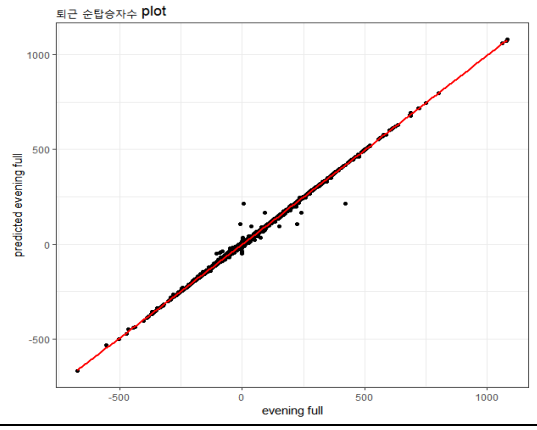
하락하고 그 이후에는 급증한다. 버스별 경유 정류소 순서인 seq 변수의 PDP 를 보면 seq 가 커질수록 순 승차수가 감소함을 볼 수 있다.

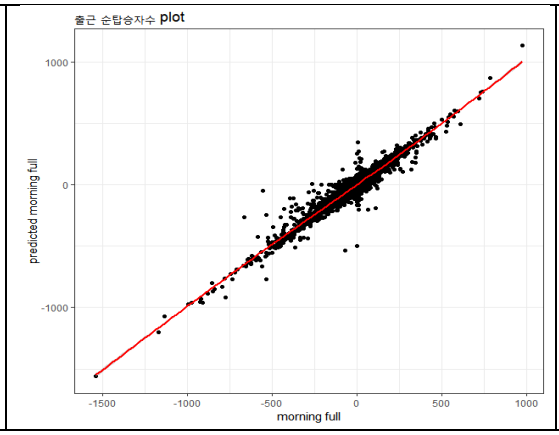
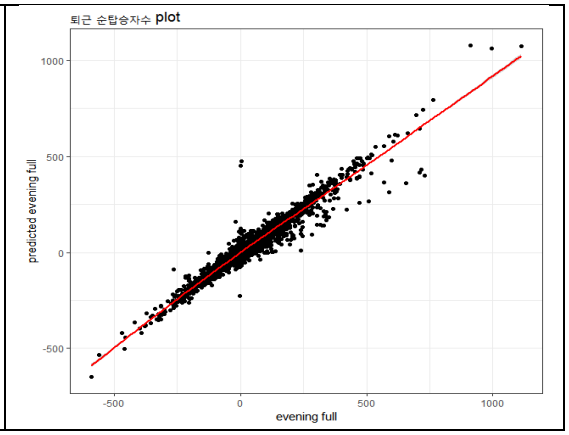
퇴근 시간대 순 승차수의 변수별 PDP 를 보면 위도와 경도의 영향이 출근 시간대보다 더 크게 나타남을 볼 수 있다. 위도 및 경도 모두 값이 커질수록 순 승차수가 증가하는데 위도의 경우 37.48, 경도의 경우 127.02 일 때 튀는 값을 볼 수 있다. 또한 퇴근 시간대의 순 승차 수에 영향을 주는 버스 노선의 길이는 커질수록 승차 수가 작아진다.

최종 모형



모델별로 train rmse 와 test rmse 를 그래프로 나타냈다. 두 경우 모두 xgboost 의 rmse 가 가장 작아 xgboost 모델을 최종모형으로 선택하였다.

	출근	퇴근
train data		

test data		

최종 모형을 이용하여 실제값(x 축) vs 예측값(y 축) 그래프를 그려보았다. train data 인 2017 년 6 월~12 월 데이터에 대해서는 거의 일직선에 가까운 형태를 보인다. test data 인 2018 년 1 월~3 월 데이터에 대해서도 어느 정도 오차를 보이긴 하지만 아주 잘 예측했다고 할 수 있다.

PROPOSITION

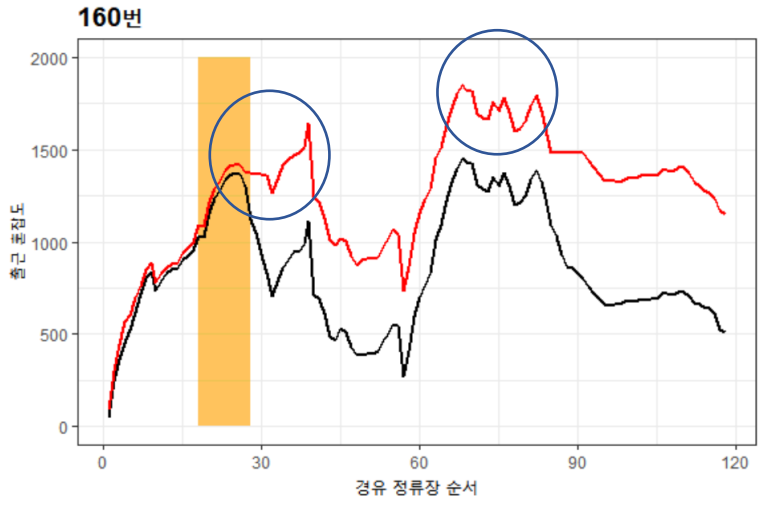
새로운 다람쥐버스에 대해 제안하기 위해 2018년 1~3월 데이터로 예측한 출근, 퇴근 순 승차수로 혼잡도를 계산하고 버스별로 상위 5%의 혼잡도(출근 : 1158.664, 퇴근 : 834.1195)가 연속된 구간의 개수를 세어보았다.

출근	노선번호	연속구간
1	720	93
2	**150**	**60**
3	**160**	**57**
4	**340**	**44**

퇴근	노선번호	연속구간
1	7211	41
2	110B	39
⋮	⋮	⋮
10	**340**	**20**

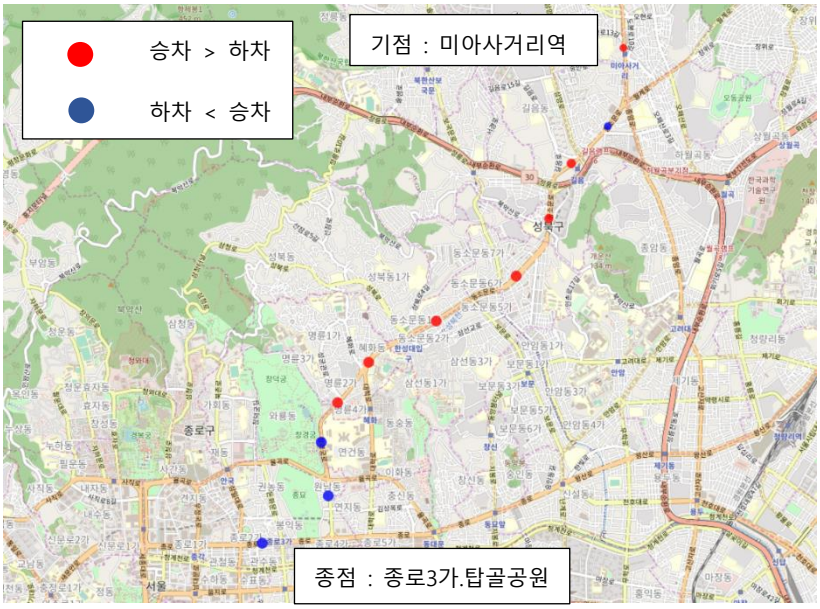
먼저 출근 시간대에서 가장 혼잡한 노선인 720번 버스는 연속 구간이 93개로, 전체 107개의 구간 중 대부분이 혼잡한 구간이라고 예측하였다. 하지만 720번의 혼잡한 구간은 2017년에 신설된 702A, 702B 버스에서 신설된 8771번, 2018년에 신설된 8221번과 비슷하기 때문에 새로운 노선 제안에서 제외하였다. 그 다음으로 혼잡도 연속 구간이 가장 많은 150번, 160번은 대체로 비슷한 정류소를 경유하기 때문에 혼잡구간이 겹치는 곳으로 새로운 다람쥐 버스 노선을 제안하는 후보로 선정하였다. 그 다음으로 많은 340번 버스는 퇴근 구간에서도 상위 10번 째에 포함되어 있기 때문에 출근, 퇴근 시간대에서 모두 운행하는 새로운 다람쥐 버스 노선으로 제안하였다.

출근 시간대 – 160

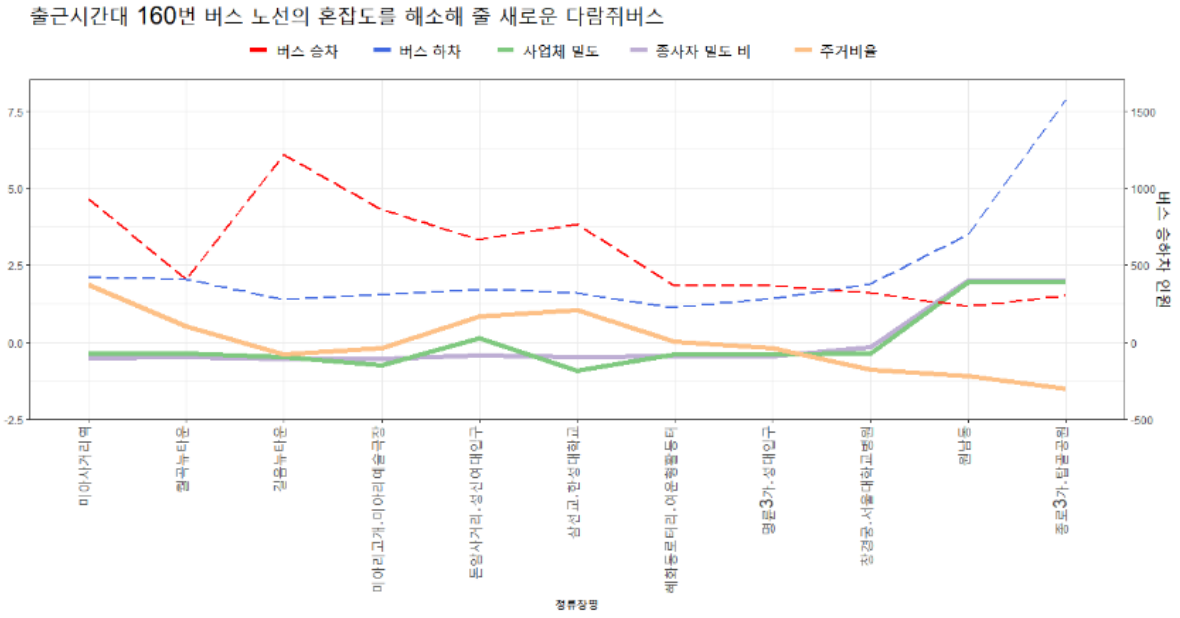


먼저 160번 버스가 지나다니는 정류소 순서에 대하여 출근 혼잡도를 그래프로 나타내면 다음과 같다. 검정색은 실제 값, 빨간색은 예측한 값이다. 예측한 값이 실제 값보다 더 크게 예측하는 경향이 있지만, 혼잡도의 경향은 잘 예측하기 때문에 문제는 없어 보인다. 혼잡도가 많은 구역은 두 군데로 보이는데, 출근 버스를 제안하는 것이기 때문에 2번 클러스터에서 3번 클러스터로 가는

방향을 제안하였다. 중간 지점인 60~80번 구간도 3번 클러스터로 많이 가는 구간이 보이긴 하지만 18번에서 28번 구간이 상대적으로 이용할 수 있는 지하철이나 버스가 많이 다님에도 불구하고 혼잡도가 높게 나타나 이 구간을 새로운 노선으로 정하였다. 구간을 지나는 정류소를 지도에 그려보면 다음과 같다.



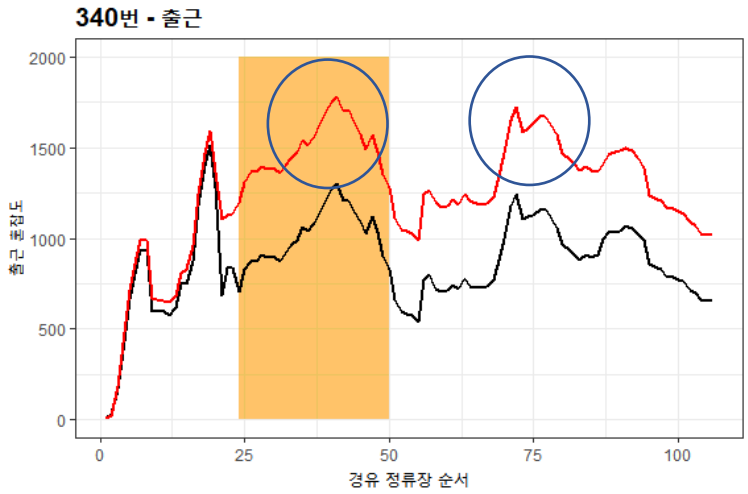
1	미아사거리역
2	월곡뉴타운
3	길음뉴타운
4	미아리고개.미아리예술극장
5	돈암사거리.성신여대입구
6	삼선교.한성대학교
7	명륜 3 가.성대입구
8	창경궁.서울대학교병원
9	원남동
10	종로 3 가.탑골공원



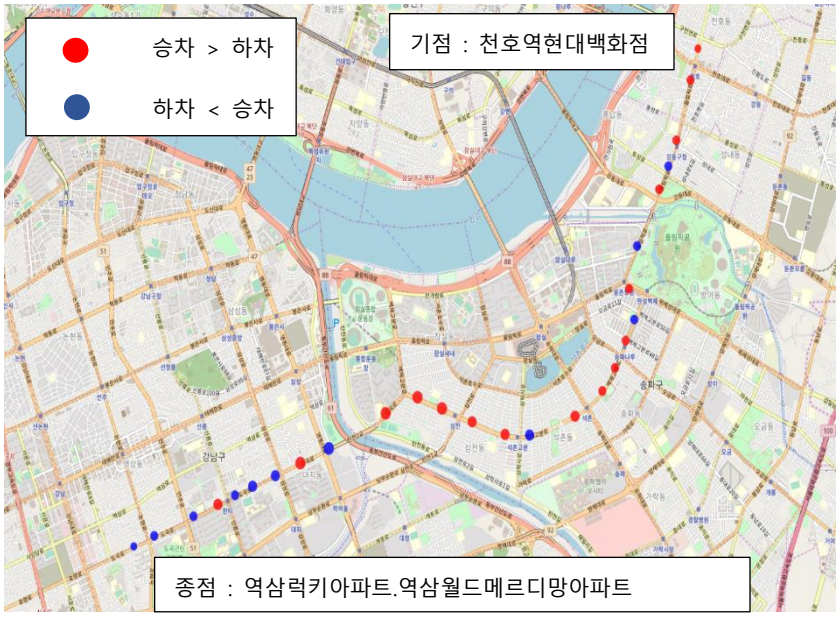
총 7.5km 로 빨간색 점은 승차가 많은 지점, 파란색 지점은 하차가 더 많은 지점이다. 주로 앞부분에는 승차가 많고, 뒤쪽 부분에는 하차가 많이 몰려있다. 이 정류소에 대한 정보를 나타내기 위해 정류소별로 지역정보를 시각화하였다. 먼저 종사자 밀도비와 사업체 밀도가 종점지역으로 갈수록 증가하여 하차가 많고, 주거비율이 높은 정류장일수록 승차가 많아 보인다. 길음뉴타운역에서는 주거지역이 적음에도 불구하고 승차 인원이 많은데, 주변에 길음역을

이용하는 사람들이 환승을 많이 하는 영향이 있을 것으로 보인다. 그리고 종로 3 가 탑골공원은 720 번도 지나는 정류소로, 이 버스를 신설하면 720 번의 혼잡도도 줄어들 것이라고 예상한다.

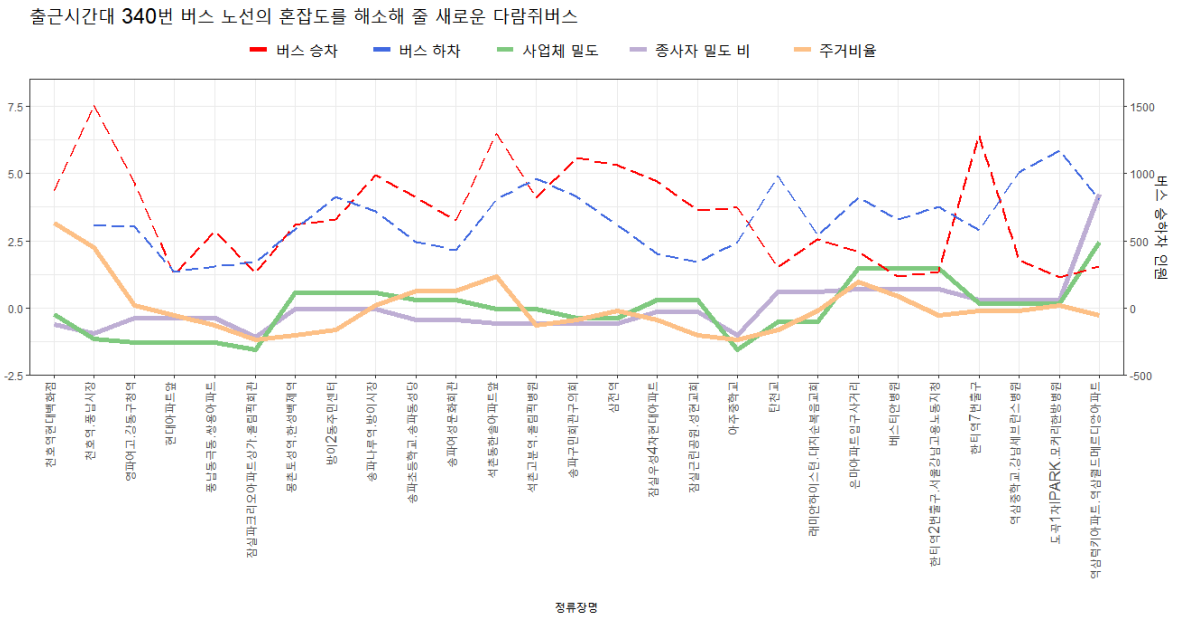
출근 시간대 – 340



160 번과 마찬가지로 혼잡도가 큰 구간이 크게 두 군데로 나오는데, 종점 쪽의 클러스터가 주거비율이 높은 1 번 클러스터보다는 2 번, 3 번이 높은 곳으로 선택하였다. 24 번부터 50 번까지의 정류소를 지도에 나타내면 다음과 같다.

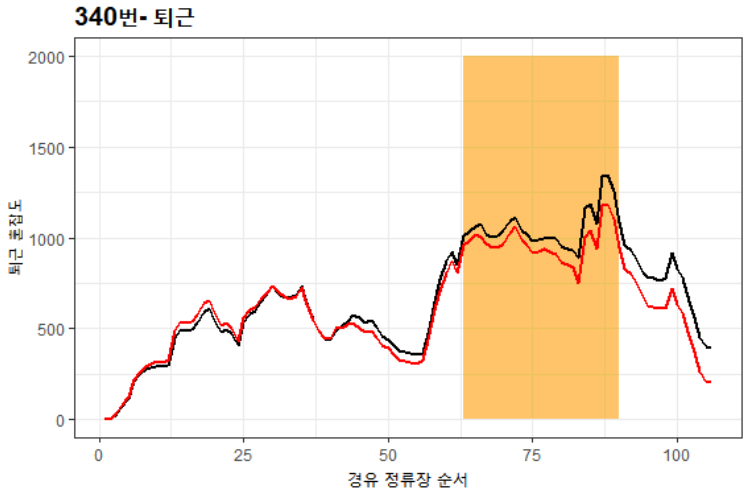


1	천호역현대백화점
2	천호역.풍납시장
3	영파여고.강동구청역
4	현대아파트앞
5	풍납동극동.쌍용아파트
6	잠실파크리오아파트상가
7	몽촌토성역.한성백제역
8	방이 2 동주민센터
9	송파나루역.방이시장
10	송파초등학교.송파동성당
11	송파여성문화회관
12	석촌동한솔아파트앞
13	석촌고분역.올림픽병원
⋮	⋮
27	역삼럭키아파트



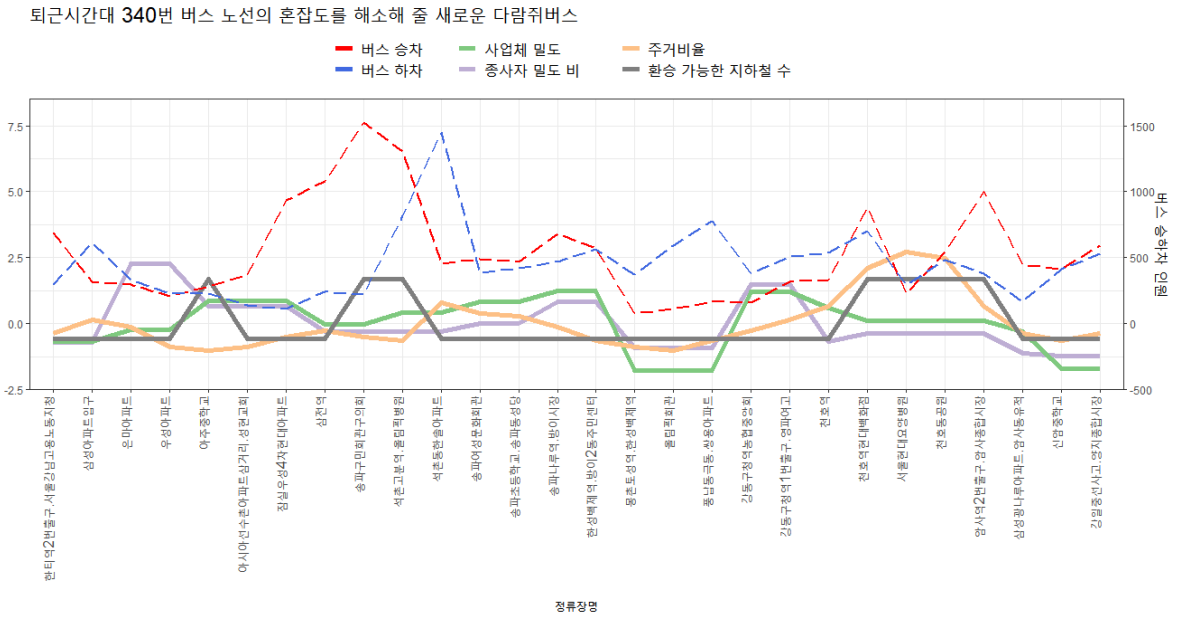
총 11.6km 로 승차가 많은 지역, 하차가 많은 지역이 번갈아 나타나고, 기점에서는 승차, 종점에서는 하차 지역이 많아 보인다. 각각의 정류소에 대한 지역정보 그래프를 보면, 주거비율이 높을수록 승차가 많고, 사업체 밀도와 종사자밀도비가 높은 것을 볼 수 있다. 그리고 기점에서는 사업체와 종사자 밀도비보다는 주거비율, 종점에서는 주거비율보다는 사업체밀도, 종사자밀도비가 더 높게 나타나 출근버스에 적합한 버스라고 할 수 있다.

퇴근 시간대 - 340



퇴근 혼잡도에 대해서는 혼잡도가 높은 구간이 뚜렷하여 이 구간을 새로운 퇴근버스 노선으로 선정하였다. 주로 2 번 클러스터에서 1 번 클러스터로 이동하여 퇴근버스에 맞게 주거지역으로 이동한다. 이 구간을 지도에 나타내면 다음과 같다.

1	한티역 2 번출구
2	삼성아파트입구
3	은마아파트
4	우성아파트
5	아주중학교
6	아시아선수촌아파트삼거리
7	잠실우성 4 차현대아파트
8	삼전역
9	송파구민회관구의회
10	석촌동한솔아파트
11	송파여성문화회관
12	송파초등학교.송파동성당
13	송파나루역.방이시장
⋮	⋮
26	강일중선사고.양지종합시장



총 12.5km로 기점에서 하차, 종점에서 승차가 더 많이 나타나는 것 같지만 지역정보를 나타내는 그래프를 보면 기점과 종점의 승차가 크게 보이지는 않아 큰 영향은 없을 것으로 보인다. 이 구간에 대해서는 지하철 역이 많아 승차와 하차 지점이 지역정보보다는 환승 가능한 지하철 역의 개수가 더 의미 있어 보여 위 그래프에 이 변수를 추가하였다. 사업체 밀도 및 종사자 밀도비가 높은 정류소에서 승차가 많고, 주거비율이 높을수록 하차가 많아 보이는데 이외에도 환승 가능한 지하철 역의 수가 많을 때, 승차 및 하차의 수가 더 많아 보인다.

CONCLUSION

이번 프로젝트를 통해 순 승차 수를 예측하고 이것을 이용하여 혼잡도 지표를 계산한 후 새로운 다람쥐 버스를 제안하였다. 제안한 다람쥐 버스를 운행하였을 때 혼잡도의 감소량은 예측 불가능하다는 한계가 있지만 지금까지도 계속되고 있는 출근 및 퇴근 시간의 버스 혼잡도를 완화할 수 있고 이로 인해 대중교통 이용률이 증가하는 데 의의가 있다. 다시 한번 제안하는 다람쥐 버스를 정리하면 다음과 같다.

버스 노선	기.종점	대수	거리	배차간격	횟수	운행시간
8661	미아사거리역 ~ 종로 3 가.탑골공원	5	7.5km	10~12 분	11	07:00~09:00
8432(출)	천호역현대백화점 ~ 역삼럭키아파트	4	11.6km	10~11 분	12	07:00~09:00
8432(퇴)	한티역 2 번 출구 ~ 강일중선사고	5	7.5km	10~12 분	11	17:30~19:30