서울시(여의도, 상암 지역) 공공자전거 사용자 분류 및 운동량 예측

2022711778 윤태호

1. 서론
2. 데이터 설명 및 데이터 전처리
3. 모델 구현
   -Multivariate Gaussian Mixture Model
   -Latent Variable(속도, 긴박성)
   -Mixture of Experts
   -Latent Variable(이용목적, 의욕)

4. 모델 평가
5. 결론

1. 서론

오늘날 서울시를 돌아다니다 보면 서울시 공공자전거 '따릉이'를 많이 이용하는 시민들을 자주 볼 수 있다. 2022년 4월 25일 기준 서울시 공공자전거 누적 이용자수는 1억건을 돌파하였고 매년 서울시 공공자전거와 대여소들이 시민들의 사용증가로 인해 추가되고 있다. 서울시는 올해 말까지 자전거의 대수 약 3000대와 대여소의 개수를 약 250개까지 추가 시킬 것을 계획하였다. 이처럼 서울시 공공자전거 이용자들이 정말 많다는 것을 알 수 있고 이에 본 프로젝트는 서울시의 다양한 지역 중에 많은 사람들이 공공자전거를 이용하는 대표적인 장소인 여의도와 상암 지역의 공공자전거 이용자들을 분류하여 분류의 기준이 되는 latent variable을 찾아내고 운동량을 예측하는 것을 목표로 삼았다. 이러한 것들을 수행하기 위해 분석기법으로는 Mixute Model 기법인 Multivariate Gaussian Mixture Model과 Mixture of Experts를 사용할 것이다. 모델을 fitting한 이후에는 기존에 다른 대표적인 분석 방법인 Random Forest와 Linear Model을 fitting하여 운동량 예측 성능을 비교해 볼 것이다.

2. 데이터 설명 및 전처리

데이터는 서울시공공데이터(www.data.seoul.go.kr)에 있는 2022년 6월의 서울시 공공자전거 이용정보(시간대별) 데이터를 이용하였다. 데이터의 변수는 USE_CNT(이용건수), EXER_AMT(운동량), CARBON_AMT(탄소 절감량), MOVE_METER(이동거리)라는 4개의 수치형 변수와 RENT_DT(대여일자), RENT_ID(대여소번호), RENT_NM(대여소명), RENT_HR(대여시간), RENT_TYPE(대여구분코드), GENDER_CD(성별), AGE_TYPE(나이), , MOVE_TIME(이동시간)라는 8개의 범주형 변수로 구성되어 있고 데이터의 총 개수는 3704328개이다. 데이터 전처리는 우선 데이터의 개수가 너무 많은 관계로 '대여일자' 변수 중에 6월 1일부터 6월 8일까지의 총 2000개의 데이터만 이용하기로 하였다. 6월 1일부터 6월 8일까지의 기간은 비가 오지 않으면서 지방선거 날과 현충일이 있는 등 8일동안 평일과 휴일이 적절히 잘 섞여 있는 기간이기에 이 기간으로 설정되었다. 다음으로 '대여시간' 변수의 범주는 0시부터 23시까지 총 24개의 범주로 이루어져 있다. 0시부터 06시 사이의 이용자수들은 다른 시간대에 비해 상대적으로 이용자수가 적기 때문에 이 시간대의 있는 이용자들은 분석 대상에 포함하지 않았다. 그럼에도 범주의 수가 너무 많기 때문에 시간을 3시간 간격으로 총 6가지의 범주를 가진 변수로 처리하였다. 다음으로 RENT_SPOT(대여장소)라는 새로운 변수를 만들었다. RENT_SPOT 변수는 '서울시 공공자전거 이용특성에 관한 연구(장재인 김태형 이무영 2016)' 논문을 참고하여 공공 자전거 대여소 장소를 업무 지역 근처인 지역은 '업무'로 지하철 근처인 지역은 '지하철' 주거 지역 근처인 지역은 '주거'로 다음 표와 같이 할당하여 범주의 개수가 총 3개인 변수로 처리하였다.

여의도 지구 대여서 번호 및 대여소명	지역구분
200 국회위원회관	업무
201 진미파라곤 앞	업무
202 국민일보 앞	업무
203 국회의사당역 3 번출구 옆	지하철
204 국회의사당역 5 번출구 옆	지하철
205 산업은행 앞	업무
222 시범아파트버스정류장 옆	주거
223 진주아파트상가 앞	주거
224 롯데캐슬 앞	주거
225 앙카라공원 앞	지하철
226 샛강역 1 번출구 앞	지하철

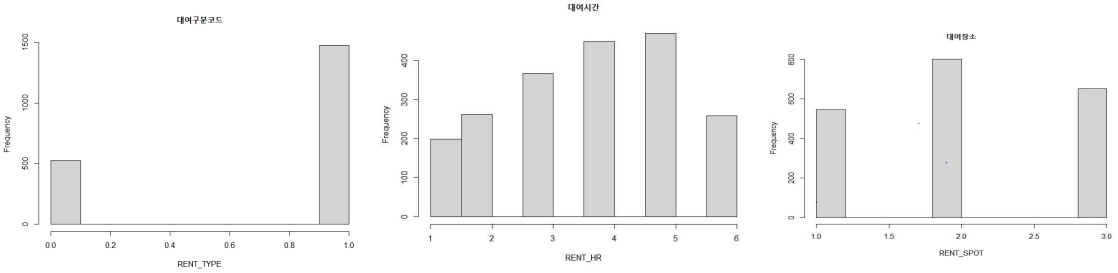
상암지구 대여소번호 및 대여소명	지역구분
400 상암 한화오벨리스크 1 차 앞	주거
401 상엄월드컵파크 10 단지 앞	주거
402 상암월드컵파크 9 단지 앞	주거
403 부엉이공원 앞	주거
405 DMC 빌 앞	업무
417 DMC 역 2 번출구 옆	지하철
418 월드컵경기장역 3 번출구 옆	지하철
419 홈플러스 앞	지하철
420 서울시 공공자전거 운영센터 옆	업무
421 마고구청 앞	주거

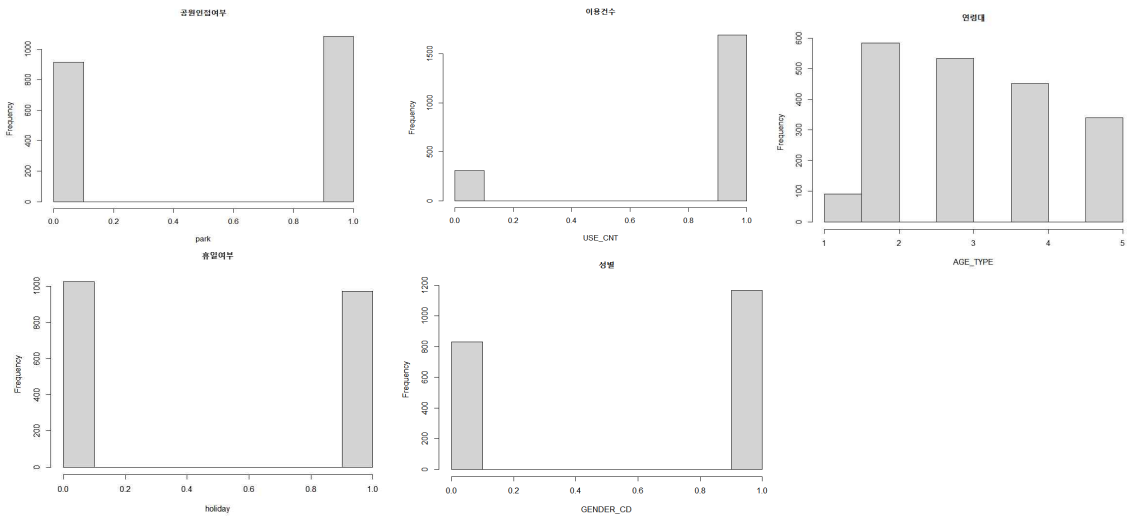
다음으로 공원인접 지역은 1, 공원 인접하지 않은 지역은 0으로 처리하여 park라는 이항 범주형 변수를 새로 만들었다.

AGE_TYPE(연령대) 변수는 10대부터 70대 이상까지 총 7개의 범주로 이루어져 있다. 7개의 범주 중 60대와 70대 이상의 공공자전거 이용자수가 다른 연령대에 비해 현저히 적기 때문에 50대와 통합하였다. 이로 인해 AGE_TYPE(연령대) 변수는 총 5개의 범주를 가진 범주형 변수가 되도록 처리하였다.

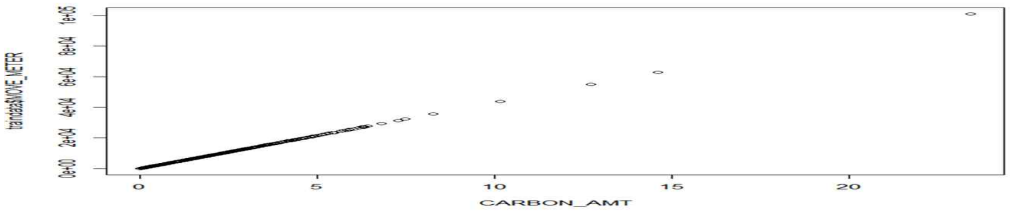
USE_CNT(이용건수) 변수는 최소값은 1, 최대값은 31인 수치형 변수이다. 수치값들을 확인한 결과 이용건수가 1번인 공공자전거 이용자들이 과반수 이상으로 많아서 이용건수가 1번인 자전거 이용자들을 1, 이용건수가 1번 초과인 자전거 이용자들은 0인 이항 범주형 변수로 처리하였다. USE_TYPE(이용권) 변수는 정기권, 단체권, 일일권, 일일권(비회원)으로 이루어진 범주형 변수이다. 변수의 범주의 비율은 정기권이 압도적으로 많기 때문에 정기권을 제외한 단체권, 일일권, 일일권(비회원)을 하나의 범주로 합쳤다. 따라서 USE_TYPE 변수는 정기권은 1, 정기권이 아닌 이용권들은 0으로 처리한 이항 범주형 변수로 처리하였다.

다음으로 holiday(휴일)라는 변수를 새로 만들었다. 6월 1일은 지방선거날이고 4일과 5일은 주말, 6일은 현충일로 휴일이고, 2일, 3일, 7일, 8일은 평일이다. 따라서 holiday(휴일) 변수는 휴일은 1, 평일은 0인 이항 범주형 변수로 처리하였다. 사용되는 범주형 변수들의 히스토그램은 아래와 같다.

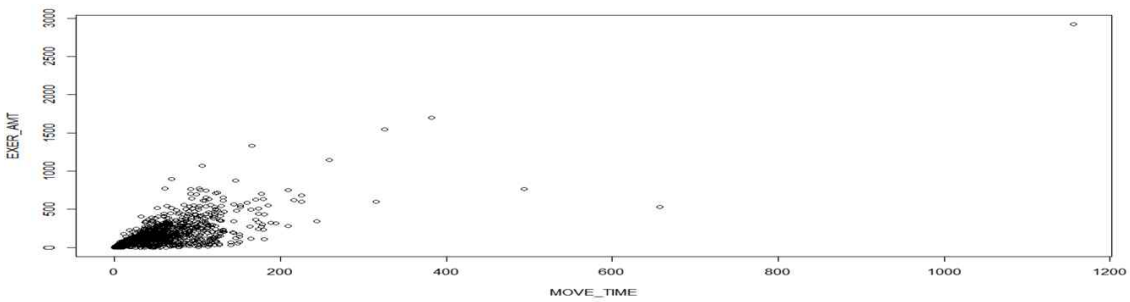




다음으로 수치형 변수를 살펴보겠다. CARBON_AMT(탄소절감량) 변수와 MOVE_METER(이동거리) 변수의 관계를 살펴보기 위해 X축이 CARBON_AMT(탄소절감량)이고 Y축을 MOVE_METER(이동거리)인 plot을 그려보았다. 그 결과는 아래와 같다.



plot을 보면 CARBON_AMT(탄소절감량)과 MOVE_METER(이동거리)는 완전히 정비례하는 것을 확인할 수 있다. 이를 통해 CARBON_AMT(탄소절감량)과 MOVE_METER(이동거리)는 거의 같은 변수로 봐도 무방하다고 할 수 있다.
다음으로 수치형 변수인 MOVE_TIME(이동시간)과 EXER_AMT(운동량)의 관계를 살펴보기 위해 X축을 MOVE_TIME(이동시간), Y축을 EXER_AMT(운동량)으로 설정하여 plot을 그려보았다.



plot을 확인해보면 앞의 그래프인 CARBON_AMT(탄소절감량)과 MOVE_METER(이동거리)처

럼 완전히 정비례하는 것은 아니지만 어느 정도 우상향하게 비례한다는 것을 확인할 수 있다.

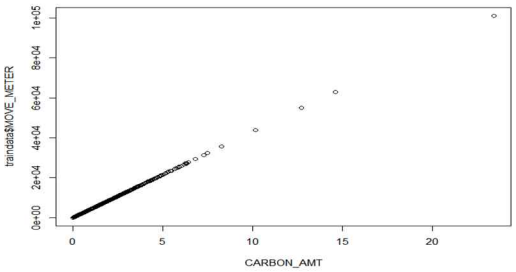
3.모델 구현

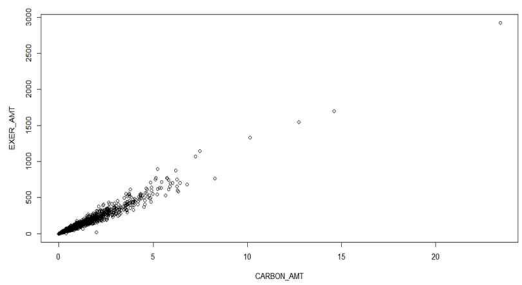
-Multivirate Gaussian Mixture Model

모델을 fitting 할 때 일단 수치형 변수만 활용해서 model을 fitting 해보겠다. 그래서 수치형 변수만을 이용해서 모델을 fitting하기 위해 선택한 혼합모형은 Multivirate Gaussian Mixture Model이다. Multivirate Gaussian Mixture Model의 모형의 pdf식은 다음과 같다.

$$f(x;\Theta) = \sum_{j=1}^{k} \frac{p_j}{(2\pi)^{\frac{p}{2}} \left|\Sigma_j\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_j)' \Sigma_j^{-1}(x-\mu_j)\right)$$

x변수로는 CARBON_AMT(탄소절감량)과 MOVE_TIME(이동시간) 총 2가지 변수만을 선택했다. 수치형 변수는 2가지만 선택한 이유는 아무래도 CARBON_AMT(탄소절감량)은 아래 두 개의 plot을 보면 알 수 있듯이 MOVE_METER(이동거리)와 EXER_AMT(운동량)과 상관관계가 너무 높기 때문에 두 변수를 분석에 사용에 있어 배제하였다.





그래서 첫 번째 X변수를 CARBON_AMT(탄소절감량)과 두 번째 X변수를 MOVE_TIME(이동시간)으로 두고 component의 개수를 2개에서 4개까지 총 3개의 모델을 fitting하였다. fitting을 한 이후의 component별로 log-likelihood, df(parameter의 개수), BIC, ICL, CLUSTERING별 개수는 아래 표와 같다.

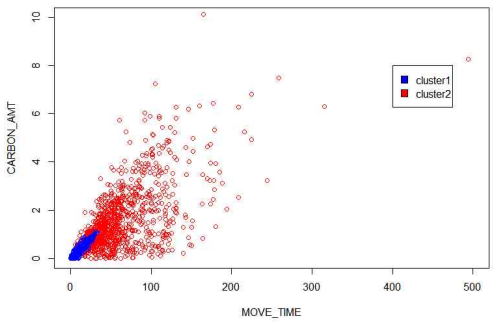
	K=2	**K=3**	**K=4**
Log-likelihood	-10657.99	-10122.29	-9835.712
df	11	17	23
BIC	-21399.58	-20373.8	-19846.24
ICL	-21539.32	-20722.44	-20295.07
CLUSTERING	1:1114(55.7%), 2:886(44.3%)	1:804(40.2%), 2: 934(46.7%), 3: 262(13.1%)	1: 596(29.8%), 2:596(29.8%), 3:66(3.3%), 4:742(37.1%)

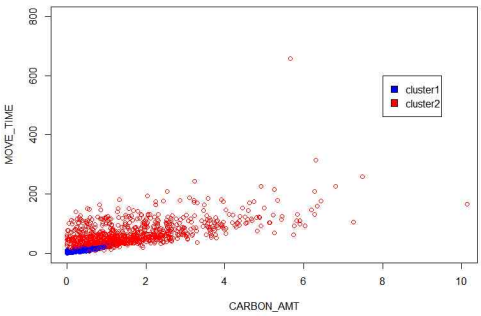
log-likelihood는 component의 개수가 2개일 때 가장 낮고 개수가 4개일 때 가장 높은 것을 알 수 있다. 또한 BIC와 ICL의 경우는 모두 component의 개수가 2개일 때 가장 낮은 것을 확인할 수 있다. 따라서 BIC와 ICL 기준으로 component의 개수가 2개인 모델을 최종적으로 선택하였다. component가 2개일 때의 첫 번째 clustering의 개수는 1114개 두 번째 clustering의 개수는 886개인 것을 확인할 수 있다. 최종 선택된 모델의 평균과 분산의 parameter는 아래와 같다.

$$m1 = \begin{pmatrix} 0.3954549 \\ 12.2312660 \end{pmatrix} \quad m2 = \begin{pmatrix} 1.744279 \\ 69.841926 \end{pmatrix}$$

$$\sigma_1^2 = \begin{pmatrix} 0.7257567 & 1.926341 \\ 1.92634072 & 63.096600 \end{pmatrix} \quad \sigma_2^2 = \begin{pmatrix} 2.62635 & 67.18486 \\ 67.18486 & 3507.56188 \end{pmatrix}$$

두번째 cluster의 평균과 분산이 첫 번째 cluster의 평균과 분산보다 상대적으로 큰 것을 통해 두 번째 cluster가 상대적으로 탄소절감량과 이동거리가 많고 탄소절감량과 이동거리에 있어서 더 분산이 큰 집단인 것을 확인할 수 있다. 다음으로 탄소절감량과 이동거리를 X축과 Y축으로 한 plot을 살펴보겠다.





plot을 살펴본 결과 첫 번째 cluster인 파란색의 산점도를 보면 확실히 두 번째 cluster인 빨간색의 산점도에 비해 탄소절감량과 이동거리 모두 상대적으로 적은 곳에 분포해 있는 것을 확인할 수 있다. 이에 반면 두 번째 cluster는 첫 번째 클러스터보다 상대적으로 탄소절감량과 이동거리가 크고 다양하게 분포해 있는 것을 확인할 수 있다.

-latent variable(속력, 긴박성)

위 두 개의 plot을 보고 latent variable은 속력과 긴박성으로 설정하였다. 우선 탄소절감량이란 변수는 앞선 plot을 보았을 때 이동 거리와 같은 변수로 봐도 무방하다. 그래서 두 번째 plot을 보면 똑같은 탄소절감량을 가진 상태, 즉 똑같은 거리를 이동한 상태에서 두 번째 cluster의 산점도가 첫 번째 cluster에 비해 상대적으로 이동 시간이 더 크게 분포되어 있는 것을 확인할 수 있다. 이런 결과를 통해 latent variable을 속력과 긴박성으로 추정하였다. 그 이유는 자전거를 똑같은 거리를 이동했다고 가정했을 때 첫 번째 cluster가 두 번째 cluster에 비해 이동 시간이 짧다는 것은 첫 번 째 cluster의 공용 자전거 이동 속력이 상대적으로 더 빠르다는 것을 의미하기 때문이다. 이를 통해 첫 번째 cluster에 속하는 사람들은 짧은 거리를 빠르게 이동해야 하는 사람이라고 생각할 수 있다. 따라서 첫 번째 cluster에 속하는 사

람들은 아침에 빠른 시간 안에 출근을 해야하는 사람들이지 않을까 생각된다. 그러므로 첫 번째 cluster에 속하는 사람들은 긴박성의 측면에서 두 번째 cluster보다 상대적으로 긴박할 수 있다고 추정한다. 이에 반해 두 번째 cluster에 속하는 사람들은 상대적으로 첫 번째 cluster에 비해 천천히 많이 이동하는 사람들이라 할 수 있다. 이런점에서 긴박성의 측면에서 첫 번째 cluster에 비해 상대적으로 덜 긴박할 것이라고 생각된다. 또한 자전거를 타는 목적에 대해서도 더 다양하다고 할 수 있다. 아무래도 더 긴거리를 천천히 이동하는 사람들이기 때문에 자전거를 이용해 산책을 하는 사람들이라던지, 운동을 하는 사람들이라던지, 여행을 하는 사람들이 주로 두 번 째 cluster에 속할 것이라고 예상된다.

-Mixture of Experts

다음으로 수치형 변수뿐만 아니라 범주형 변수까지 함께 고려해 모형을 fitting을 해보고 싶어서 범주형 변수와 수치형 변수를 함께 고려해 볼만한 mixture model 중 Mixtue of Experts를 이용해 model을 fitting 해보기로 하였다. Mixture of Experts의 pdf의 식은 다음과 같다.

$$p(y|x) = \sum_{k=1}^{K} \pi_k(x;\alpha)\phi_1(y;\beta_{0k} + x^T\beta_k, \sigma_k^2) = \sum_{k=1}^{K} \frac{\exp(\alpha_{0k} + x^T\alpha_k)}{\sum_{j=1}^{K}\exp(\alpha_{0j} + x^T\alpha_j)}\phi_1(y;\beta_{0k} + x^T\beta_k, \sigma_k^2)$$

식의 분모 부분은 gating part를 의미하고 식의 가장 오른쪽 부분은 expert part 부분을 의미한다. gating part에 속하는 X변수들은 잠재변수 Z에 영향을 미치는 변수들이고 expert part 부분의 X변수들은 Y값에 영향을 미치는 변수들이다. 따라서 공공자전거 이용자들의 EXER_AMT(운동량)을 예측 하기 위해서 Y값은 EXER_AMT(운동량)으로 설정하였다. 다음으로 gating part의 X변수들은 자전거 이용자들을 분류해 내기 위한 latent variable에 영향을 준다고 판단하여 AGE_TYPE(연령대), RENT_HR(이용시간), RENT_TYPE(대여구분코드), RENT_SPOT(대여 장소), park(공원인접여부), holiday(휴일 여부), GENDER_CD(성별), USE_CNT(이용건수) 들로 설정하였다. 다음으로 expert part 부분의 X변수들은 Y변수인 운동량에 영향을 미친다고 할 수 있는 변수들인 CARBON_AMT(탄소절감량)와 MOVE_TIME(이동시간)으로 설정하였다. 여기서 MOVE_METER(이동거리)를 X변수에 추가하지 않은 이유는 아무래도 탄소절감량과의 상관성이 너무 높아서 이 변수를 넣었을 경우 올바른 model fitting이 되지 않을 것이라 판단하여 추가하지 않았다.

그래서 model의 component의 개수를 정하기 위해 k=2부터 5까지의 Mixture of Experts model 총 4개의 model을 fitting 하였다. fitting을 하고난 이후의 log-likelihood와 parameter의 개수와 AIC, BIC, CLUSTER별 개수는 다음의 표와 같다.

	K=2	K=3	K=4	K=5
Log-likelihood	-8396.033	-7618.534	-7081.762	-6966.169
df	25	46	67	88
BIC	-16982.09	-155586.71	-14672.784	-14601.217
AIC	-16842.06	-15329.068	-14297.524	-14108.338
ICL	-17369.92	-16147.37	-15023.26	-15091.68
CLUSTERING	1:817(40.8%), 2:1183(59.1%)	1:582(29.1%), 2:865(43.2%), 3: 553(27.6%)	1:573(28.6%), 2: 698(34.9%), 3: 260(13%), 4: 469(23.4%)	1: 678(33.9%), 2: 570(28.5%), 3: 133(6.6%), 4: 433(21.6%), 5: 186(9.3%)

위 표를 보면 log-likelihood의 값과 parameter의 개수는 component의 개수가 많아질수록 높아지는 것을 확인할 수 있다. AIC와 BIC, ICL의 값을 살펴보면 모두 component의 개수가 2개일 때 가장 작은 것을 확인할 수 있다. 이를 통해 AIC, BIC, ICL 어떤 기준을 선택하더라고 최적의 Mixture of Experts는 component의 개수가 2개인 모형이라고 할 수 있다. 그러나 component의 개수가 2개인 모형을 fitting을 한 결과 latent variable을 찾는 부분에 있어 해석에 어려움이 생겨 component의 개수가 2개인 모형을 제외한 모형들 중에 가장 BIC, AIC, ICL 값이 작은 component의 개수가 3개인 모형을 최종 모형으로 선택하였다. component의 개수가 3개인 모형으로 fitting한 model의 parameter 값은 아래와 같다.

<gating part의 parameter값>

(intercept)	AGE_TYPE2	AGE_TYPE3	AGE_TYPE4	AGE_TYPE5	RENT_HR2	RENT_HR3	RENT_HR4	RENT_HR5
1.8231830	-0.4231997	-0.3373855	-0.5493095	-0.1454473	-0.3474 5963	-0.6641650	-0.4667648	-0.3340197
-0.8814753	0.4404803	0.9031536	1.1422028	1.0327297	0.0699 7129	0.1549685	0.3602649	0.5662227
RENT_HR6	RENT_TYPE1	park1	holiday1	RENT_SPOT2	RENT_SPOT3	GENDER_CD1	USE_CNT1	
-0.6667 6430	0.1315269	0.1711044	0.0774698 1	0.4341394 80	0.3624938	-0.6221183	-0.9029996	
0.0333357 5	-0.2672219	0.2912679	0.0533285 79	0.0077724 881	-0.4494 544	3.2720818	-2.9198973	

<expert part의 parameter값>

cluster1		
(intercept)	CARBON_AMT	MOVE_TIME
0.025831221	110.984818549	-0.001192848
cluster2		
(intercept)	CARBON_AMT	MOVE_TIME
4.782200154	92.275657056	-0.002849478
cluster3		
(intercept)	CARBON_AMT	MOVE_TIME
3,57816047	126.76086739	-0.08268033

다음으로 fitting한 최종 model의 예측된 Y값(운동량)의 평균값과 pair plot은 아래와 같다.

cluster1	cluter2	cluster3
83.01738	80.8889	191.2407

결과를 보면 세 번째 cluster의 예측된 운동량의 평균값이 다른 두 개의 cluster보다 2배 이상 더 많은 것을 알 수 있다. pair plot을 보더라도 세 번째 cluster가 초록색을 나타내는데 Y값이 운동량을 나타내는 맨위의 plot들을 살펴보았을 때 확실히 모든 plot들이 초록색이 가장 위에 분포해 있는 것을 확인할 수 있다.
다음으로 latent variable을 추정하기 위해 cluster 별로 수치형 변수들의 평균값과 범주형 변수들의 범주별 비율을 다음의 표로 나타내 보았다.

	CLUSTER1					CLUSTER2					CLUSTER3				
MEAN(MOVE_TIME)	31.04811					32.12023					56.47197				
MEAN(CARBON_AMT)	0.7330241					0.7811445					1.66472				
AGE_TYPE	10대 이하	20대	30대	40대	50대 이상	10대 이하	20대	30대	40대	50대 이상	10대 이하	20대	30대	40대	50대 이상
	4%	32%	25%	21%	15%	5%	31%	27%	17%	17.5%%	3%	21%	25%	31%	17.7%
GENDER_CD	남		여			남		여			남		여		
	52%		47%			39%		60%			94%		6%		
RENT_SPOT	업무		지하철		주거	업무		지하철		주거	업무		지하철		주거
	28.3%		35.5%		36.6%	20.4%		42.8%		36.6%	36.8%		40.3%		22.7%
RENT_HR	06-09	09-12	12-15	15-18	18-21	21-00	06-09	09-12	12-15	15-18	18-21	21-00	06-09	09-12	12-15	15-18	18-21	21-00
	8.7%	13.7%	21.8%	20.9%	19.9%	14.7%	11.6%	15%	17.4%	20.6%	23.1%	12%	8.3%	9.2%	16%	26.4%	27.6%	12.2%
RENT_TYPE	정기권		그외			정기권		그외			정기권		그외		
	74.2%		25.7%			76.4%		23.5%			69.2%		30.7%		
holiday	O		X			O		X			O		X		
	49.3%		50.6%			50%		49.9%			45.9%		54%		
park	O		X			O		X			O		X		
	49.4%		50.5%			52.3%		47.6%			62%		37.9%		
USE_CNT	1		>1			1		>1			1		>1		
	95.3%		4.6%			89.7%		10.2%			65%		34.9%		

표의 결과를 보면 첫 번째 cluster와 두 번째 cluster의 평균 이동시간을 보면 약 30분으로 비슷한 것을 알 수 있다. 반면 세 번째 cluster의 평균 이동시간은 56분으로 다른 두 cluster에 비해 약 두 배 많은 것을 알 수 있다. 탄소 절감량도 마찬가지로 첫 번째 cluster와 두 번째 cluster는 비슷한 수치를 보였고 세 번째 cluster는 다른 cluster에 비해 두 배 많은 수치를 보인다. 다음으로 연령대이다. 연령대 모두 어느 한 연령대로 치우치지 않고 고른 비율을 보인다는 것을 알 수 있다. 첫 번째 cluster와 두 번째 cluster는 20대가 가장 많고 그 다음이 30대가 차지하고 있다. 반면 세 번째 cluster는 40대가 가장 많은 비율을 차지한다. 또한 이례적으로 50대 이상 비율이 다른 cluster에 비해 상대적으로 가장 높다는 것을 확인할 수 있다. 다음으로 성별을 살펴보면 첫 번째 cluster는 남녀 비율이 비슷하다는 것을 알 수 있고 두 번째 cluster는 4:6 비율로 여성이 더 많다는 것을 알 수 있다. 세 번째 cluster는 남성의 비율이 94%로 남성의 비율이 압도적으로 많다는 것을 알 수 있다. 다음으로 RENT_SPOT을 살펴보면 첫 번째 cluster는 주거, 지하철 지역 순으로 많은 것을 알 수 있고 두 번째와 세 번째 cluster는 지하철 지역이 가장 많은 비율을 차지하고 있는 것을 확인할 수 있다. 다음으로 휴일 여부에 대해서는 세 cluster 모두 휴일 여부에 대한 범주가 약 5대5 비율로 비슷하다는 것을 확인할 수 있다. 다음으로 공원 여부에 대해서는 첫 번째 cluster와 두 번째 cluster는 두 범주의 비율이 비슷한 반면 세 번째 cluster는 공원과 인접한 지역의 비율이 62%로 과반수 이상을 차지하는 것을 확인할 수 있다. 마지막으로 이용 건수 변수는 세 cluster 공용자전거를 1번 이용한 사람이 1번 넘게 이용한 사람들보다 더 많다는 것을 확인할 수 있었다. 이례적인 것은 1번 넘게 이용한 사람들의 비율은 세 번째 cluster가 가장 많다는 것이라 알 수 있다.

-Latent Variable(riding purpose)

위의 cluster별 수치형 변수들의 평균 값과 범주형 변수들의 범주별 비율을 통해 latent variable을 추정해 보았을 때 이용 목적이라는 변수로 추정하게 되었다. 우선 세 번째 cluster를 살펴보게 되면 이용 시간이 약 1시간 정도 되면서 40대의 비율이 가장 높다. 그리

고 50대 이상의 비율이 다른 cluster에 비해 가장 높다. 성별은 남성의 비율이 압도적으로 높고 주로 저녁 시간대에 이용자들이 많이 분포해 있다. 또한 공원과 인접한 대여소가 과반수 이상을 차지하면서 1번 넘게 자전거를 이용한 이용자들의 비율이 다른 cluster에 비해 높다. 이런 점을 미루어 보았을 때 세 번째 cluster에 속하는 사람들은 공공자전거를 운동의 목적으로 이용하는 사람들이라 판단된다. 예를 들어 주로 퇴근 후 일과를 마친 저녁 시간 때에 반복적으로 운동을 하기 위해 자전거를 이용하는 중년 및 노년층의 남성분들 대다수가 첫 번째 cluster에 속할 것이라 예상된다. 다음으로 두 번째 cluster를 살펴보면 평균 이용시간은 약 30분 정도이면서 20~30대 비율이 가장 높고 여성의 비율이 더 높으면서 지하철역 근처에서 대여가 많고 출퇴근 시간대에 이용 비율이 다른 cluster들에 비해 높다는 것을 확인할 수 있다. 이를 통해 두 번째 cluster의 이용 목적은 통근이라 할 수 있고 주로 공공자전거를 이용하여 출퇴근 하는 여성분들 대다수가 두 번째 cluster에 속할 것이라고 예상한다. 다음으로 첫 번째 cluster는 평균 이용 시간이 약 30분이면서 20~30대 이용자들이 가장 많고 주로 12시 이후 오후 시간대에 많이 이용하는 이용자가 많다. 이를 통해 점심시간 이후 남는 여유 시간 때 시간을 간단히 보내기 위해 공공자전거를 통해 가벼운 산책을 하거나 여가용으로 이용하는 사람들 대다수가 첫 번째 cluster에 속할 것으로 예상된다.

-Latent Variable(Willpower(Desire) of Riding)

다음으로 추정한 latent variable은 자전거를 타고자 하는 의욕, 욕망으로 보았다. 왜냐하면 앞서 추정한 latent variable인 이용 목적에 근거하면 세 번째 cluster의 같은 경우는 아무래도 퇴근 후에 운동을 목적으로 이용하시는 중장년층분들이 대다수이다. 보통 이런 사람들은 운동을 하기 전에 계획을 세우기도 하고 공공자전거 이용에 있어서 운동이라는 뚜렷한 목적이 있기 때문에 아무래도 세 cluster 중에서는 자전거를 타는 것에 대한 의욕이 가장 높을 것이라고 예상된다. 반대로 두 번째 cluster의 경우 공용자전거를 타는 것에 대한 의욕이 가장 낮을 것이라고 판단된다. 왜냐하면 두 번째 cluster에 속하는 사람들이 지하철역 근처에서 퇴근 시간 때 지하철을 타지 않고 공공자전거를 이용하는 이유는 아무래도 지하철 혼잡도를 피하기 위해서라고 판단된다. 이러한 이유로 공공자전거를 타는 사람들은 공공자전거를 타고 싶어서 타는 것이 아니라 지하철 혼잡도를 피하기 위해서 어쩔 수 없이 타는 분들이라고 생각한다. 이렇기 때문에 세 cluster중 공공자전거를 타는 것에 대해 의욕적인 측면에서 가장 약할 것이라고 예상된다. 마지막으로 첫 번째 cluster의 이용자들은 대학생의 경우 수업 사이의 남는 시간 때나 직장인의 경우 점심시간 이후 남는 시간 때 잠깐 시간을 보내기 위해 공용자전거를 이용하는 사람들이라 추정된다. 또한 공공자전거 대여소를 발견하고 호기심에 타보는 서울시에 관광 온 외국인이거나 대중교통을 이용하려다가 공공자전거 대여소를 발견하고 충동적으로 공공자전거를 이용하는 사람들이 주로 첫 번째 cluster에 속할 것이라고 예상한다. 이러한 사람들은 사실 공공자전거를 타든 안타든 그렇게 큰 상관이 없는 사람들이기 때문에 의욕적인 측면에서는 세 번째 cluster와 두 번째 cluster 사이에 있지 않을까 생각된다.

4. 모델 평가

마지막으로 component가 2개인 Mixture of Experts model의 운동량을 예측하는 성능이 다른 모델과 비교해 봤을 때 어떤지 확인해보고자 하였다. 그래서 독립변수로 범주형 변수와 수치형 변수를 모두 활용가능한 model인 Random Forest 모델을 fitting하기로 하였다. 나무의 개수는 500개로 하고 각 split마다 사용되는 변수의 개수를 2개로 설정하여 Y변수를 운동

량으로 설정한 model을 fitting 하였다. 다음으로 3개 이상의 범주를 가진 범주형 변수들을 모두 더비 변수로 처리한 상태에서 나머지 변수들과 함께 독립변수로 설정하고 Y변수를 운동량으로 설정하여 Linear Model을 fitting하였다. 이렇게 하여 총 세 가지 model의 RMSE와 MAE를 비교해보고자 하였다. testdata로는 동일한 6월 1일부터 6월 8일까지의 총 200개의 dataset으로 이용하였다. 결과는 아래 표와 같다.

	Random Forest	Linear Model	Misture of Experts
RMSE	61.07248	30.85834	27.87566
MAE	28.00866	16.86678	9.811386

표를 확인한 결과 RMSE와 MAE 모두 Mixture of Experts가 가장 작은 것을 확인할 수 있다. 이를 통해 운동량을 예측하는 성능에 대해서는 Mixture of Experts가 세 model중 가장 좋다고 할 수 있다.

5. 결론

2022년 4월 기준 서울시 공공자전거 누적 이용자수가 1억명을 돌파할 정도로 서울시 공공자전거 이용자들의 공공자전거 수요가 오늘날 점점 증가하고 있다. 이러한 증가에 더불어 본 프로젝트는 서울시에서 공공자전거가 많이 사용되는 여의도와 상암지역의 공공자전거 이용자들을 mixture model을 이용하여 clustering하고 clustering의 기준이 되는 latent variable을 추정하는 것을 목적으로 두었다. 이를 위해 Multivariate Gaussian Mixtue Model과 Mixture of Experts를 이용하였다. Multivariate Gaussian Mixture Model을 이용한 결과 총 2개의 component를 가진 model이 fitting 되었고 추정된 latent variable은 속력과 긴박성이었다. 첫 번째 cluster는 짧은 거리를 빠르게 이동하는 주로 출근을 목적으로 공공자전거를 이용하는 사람들로 추정되었고 두 번째 cluster는 이에 반해 더 긴 거리와 긴 시간을 이용하는 이용자들이었다. 아마 자전거를 통해 산책과 같은 여가 시간을 보내거나 운동을 한다거나 여행을 한다던지 다양한 이용 목적을 가진 사람들이 이 두 번째 cluster에 속할 것이라고 예측된다. 다음으로 Mixture of Experts로 추정된 model의 component 개수는 3개였고 추정된 latent variable은 이용 목적과 공공자전거를 타는 것에 대한 의욕이었다. 세 번째 cluster의 이용 목적은 운동이라 추정하였고 운동을 목적으로 공공자전거를 타는 사람들은 자전거를 타는 것에 대한 의욕적인 측면에서 가장 높은 의욕을 가질 것이라고 보았다. 다음으로 두 번째 cluster의 이용 목적은 출퇴근으로 추정하였고 공공자전거를 출퇴근용으로 타는 사람들은 지하철의 혼잡도를 피하기 위해서 억지로 타는 경향이 있다고 판단하여 의욕적인 측면에서는 가장 낮을 것이라고 생각하였다. 다음으로 첫 번째 cluster의 이용 목적은 여가, 산책, 간단하게 시간 보내기로 추정하였다. 이들은 공용자전거를 이용하는 목적에 있어서 이용을 하던 안하던 크게 상관이 없는 사람들이라 의욕적인 측면에서 세 번째 cluster와 두 번째 cluster 중간 정도일 것이라 예측된다. 이밖에도 Mixture of Experts의 운동량 예측 성능을 파악하기 위해 Random Forest 모델과 Linear Model 모델과의 RMSE와 MAE를 test data 200개를 이용하여 구하였다. 그 결과 Mixture of Experts의 RMSE와 MAE가 가장 작게 나

온 것을 확인할 수 있었다. 마지막으로 프로젝트를 진행하면서 느꼈던 문제점은 변수를 추가적으로 더 만들었음에도 불구하고 CLUSTERING을 하여 latent variable을 추정하는 있어 어려움이 존재한다고 볼 수 있었다. 이로 인해 latent variable과 각 cluster별 특징을 잡는데 있어 과도한 가정을 했다고 볼 수 있다. 따라서 model이 더 발전되기 위해서는 여의도와 상암 지역 뿐만 아니라 서울시 전체를 포괄할 수 있도록 대여소들을 설정하고 올바른 latent variable을 찾을 수 있도록 더 많은 변수가 생성되어야 한다고 본다.