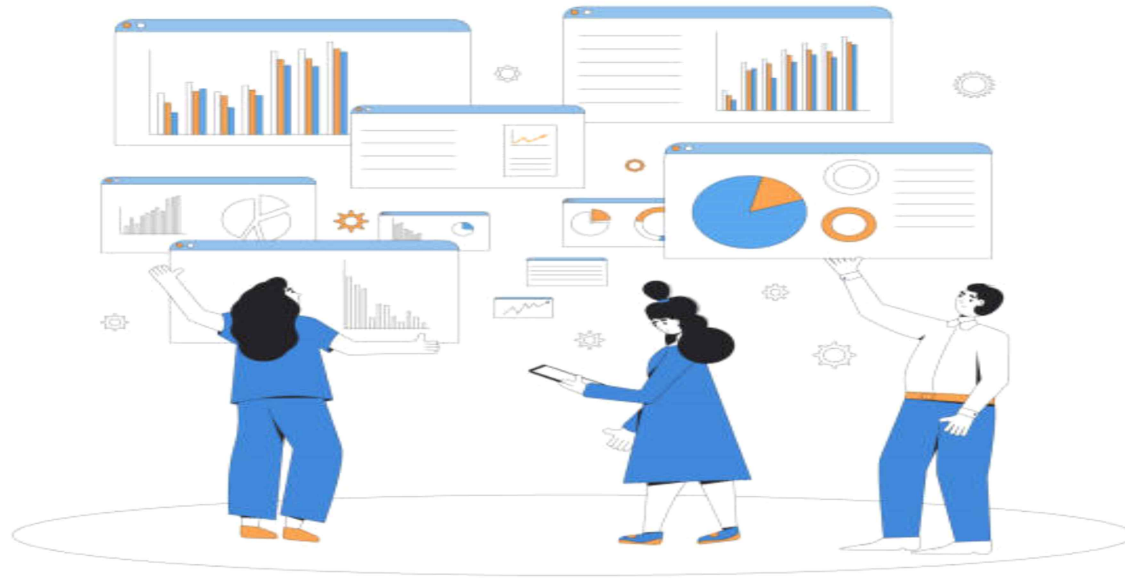


주요 프로젝트 PORTFOLIO



윤태호

성균관대학교 일반대학원 통계학과

주요 프로젝트 분야 개요

- 주요 프로젝트 분야
‘딥러닝 모델 VAE 변형 연구’, ‘머신러닝 및 딥러닝 활용 다양한 분야 예측 시도’

대학원 프로젝트

- VAE 변형 연구
- 서울시 오존 농도 예측(고급시계열분석)
- 서울시 따릉이 이용자 분류(혼합모형)
- chatgpt에 대한 tweets들의 sentiment 분류(딥러닝 토픽)
- 지구근처 소행성들의 종류 판별 및 잠재적 위험 유무 판별(일반화 선형모형)

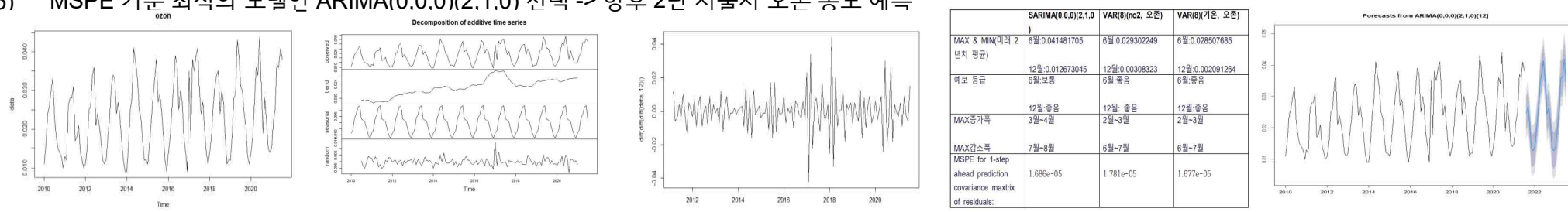
주요 프로젝트 수행 이력(대학원 프로젝트)

1. VAE(variational auto encoder) 모델 변형 연구

제목	Augmentation data with VAE that increases classification performance for sparse category
연구배경	<ul style="list-style-type: none">머신러닝 및 딥러닝 모델들은 분류에 있어 데이터의 개수가 작은 경우(예: 파산,금융사기)의 범주보다 데이터의 개수가 많은 범주를 맞추는 것에 훨씬 높은 정확성을 보인다. 데이터의 개수가 작은 범주의 낮은 정확도 문제 발생VAE에 예측모형(FFN)을 추가하여 데이터의 개수가 작은 범주의 정확도를 높이는 데이터를 VAE가 생성하도록 한다.
사용언어	R(tensorflow 이용)
연구내용	<div>1) GLM 모형에 적합한 가상 데이터 10만개 생성 후 train,test로 분류(8:2비율)[독립변수10개, 종속변수 0(데이터 개수가 많은 범주),1(데이터 개수가 적은 범주)의범주]</div> <div>2) VAE모형에 FFN모형 추가(기존의 loss function에도 binary cross entropy 추가)</div> <div>3) train data를 변형된 VAE모형에 적용하여 새로운 데이터를 생성한다.</div> <div>4) 예측모형 FFN이나 GLM에 기존 train data를 fitting시켰을 때와 변형된 새로운 데이터를 fitting(0의 범주 데이터의 상당부분을 1로 학습시킴)시켰을 때의 test data의 정확도를 비교한다. -> 1의 범주의 정확도가 62~63%대에서 75%대로 상승함 이에 따라 전체 범주의 정확도도 상승</div> <div><ul style="list-style-type: none">Reconstruction error: $-\sum_{j=1}^D (x_{i,j} - \mu_{i,j})^2$Regularization: $\frac{1}{2} \sum_{j=1}^J (\mu_{i,j}^2 + \sigma_{i,j}^2 - \ln(\sigma_{i,j}) - 1)$Binary cross entropy: $-\frac{1}{N} \sum_{i=0}^N y_i \cdot \log(\hat{y}) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$</div>
기대 효과 및 향후 계획	가상 데이터 적용 결과 데이터 개수가 적은 범주의 정확도가 상승한 것을 토대로 실제 한 범주의 개수가 적은 파산, 금융사기 데이터에 적용하여 성능이 좋아지도록 모델의 parameter를 변경한다. 연구결과를 통해 한국통계학회에 논문을 투고 하는 것을 목표로 한다.

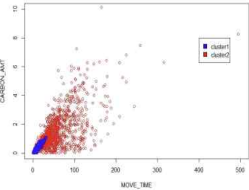
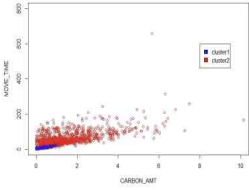
주요 프로젝트 수행 이력(대학원 프로젝트)

2. 고급 시계열 분석 프로젝트

제목	서울시 오존 농도 예측																														
프로젝트 배경 및 개요	<ul style="list-style-type: none">오존 농도에는 몇 가지 예보 등급이 있다. 오존 농도 예보 위험 등급에 대비할 수 있어야 함2010년도 1월부터 2021년 7월까지의 서울시 오존 농도 데이터(공공데이터포털)을 이용하여 향후 2년 간의 오존농도 예측																														
사용언어/모델	R/SARIMA,VAR																														
프로젝트 내용	<div><div><div><div><div>1) 서울시 오존 농도 데이터를 정상성을 갖게 하기 위하여 2번 차분 실시->ADF TEST결과 정상성 따름</div><div>2) 다양한 PARAMETER를 가진 AIRIMA 모델들 중 BIC 기준으로AUTO ARIMA함수를 적용한 ARIMA(0,0,0)(2,1,0) 모형 최종선택</div><div>3) 오존과 관련있다고 판단되는 기온과 이산화질소를 활용하여 2개의 VAR(8)모형 FITTING(VAR.TEST실시 p값이 8일 때 BIC가 가장 낮음)</div><div>4) 3가지 모델 성능 비교 -> MSPE(VAR(8)(no2,오존): 1.781e-05/ VAR(8)(기온,오존):1.677E-05/ ARIMA(0,0,0)(2,1,0): 1.686e-05</div><div>5) MSPE 기준 최적의 모델인 ARIMA(0,0,0)(2,1,0) 선택 -> 향후 2년 서울시 오존 농도 예측</div></div><div><div></div><table><thead><tr><th></th><th>SARIMA(0,0,0)(2,1,0)</th><th>VAR(8)(no2, 오존)</th><th>VAR(8)(기온, 오존)</th></tr></thead><tbody><tr><td>MAX & MIN(미래 2년치 평균)</td><td>6월:0.041481705</td><td>6월:0.028302249</td><td>6월:0.028507685</td></tr><tr><td>예보 등급</td><td>12월:0.012673045 6월:보통</td><td>12월:0.00308323 6월:중음</td><td>12월:0.002091264 6월:중음</td></tr><tr><td>MAX증가폭</td><td>12월:중음 3월~4월</td><td>12월:중음 2월~3월</td><td>12월:중음 2월~3월</td></tr><tr><td>MAX감소폭</td><td>7월~8월</td><td>6월~7월</td><td>6월~7월</td></tr><tr><td>MSPE for 1-step ahead prediction</td><td>1.686e-05</td><td>1.781e-05</td><td>1.677e-05</td></tr><tr><td>covariance matrix of residuals</td><td></td><td></td><td></td></tr></tbody></table></div></div></div></div>				SARIMA(0,0,0)(2,1,0)	VAR(8)(no2, 오존)	VAR(8)(기온, 오존)	MAX & MIN(미래 2년치 평균)	6월:0.041481705	6월:0.028302249	6월:0.028507685	예보 등급	12월:0.012673045 6월:보통	12월:0.00308323 6월:중음	12월:0.002091264 6월:중음	MAX증가폭	12월:중음 3월~4월	12월:중음 2월~3월	12월:중음 2월~3월	MAX감소폭	7월~8월	6월~7월	6월~7월	MSPE for 1-step ahead prediction	1.686e-05	1.781e-05	1.677e-05	covariance matrix of residuals			
	SARIMA(0,0,0)(2,1,0)	VAR(8)(no2, 오존)	VAR(8)(기온, 오존)																												
MAX & MIN(미래 2년치 평균)	6월:0.041481705	6월:0.028302249	6월:0.028507685																												
예보 등급	12월:0.012673045 6월:보통	12월:0.00308323 6월:중음	12월:0.002091264 6월:중음																												
MAX증가폭	12월:중음 3월~4월	12월:중음 2월~3월	12월:중음 2월~3월																												
MAX감소폭	7월~8월	6월~7월	6월~7월																												
MSPE for 1-step ahead prediction	1.686e-05	1.781e-05	1.677e-05																												
covariance matrix of residuals																															
기대 효과	최종 선택 된 ARIMA(0,0,0)(2,1,0) 모델을 통해 향후 서울시 오존 농도를 예측하고 이에 따라 오존 예보 등급이 얼마진지 예측한다. 사람에게 피해를 끼칠 수 있는 오존 위험 등급에 대비																														

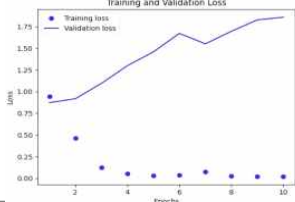
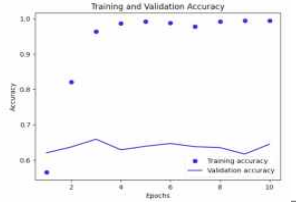
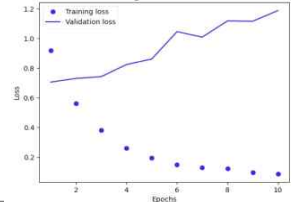
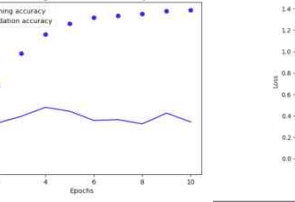
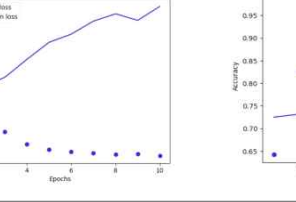
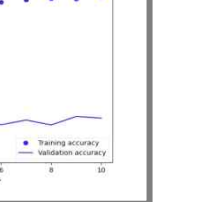
주요 프로젝트 수행 이력(대학원 프로젝트)

3. 혼합 모형 프로젝트

제목	서울시(여의도, 상암지역)공공자전거 사용자 분류 및 운동량 예측																																																																																																																																																																																																																																																																																																																																																																		
프로젝트 배경/ 목적	<ul style="list-style-type: none"> 서울시에는 공공자전거 이용자들의 이용 목적은 다양. 여의도와 상암지역에 있는 공공자전거 이용자들을 GMM과 MoE를 활용하여 분류하고 이들의 운동량을 예측. 																																																																																																																																																																																																																																																																																																																																																																		
사용언어/모델	R/ GMM, MoE																																																																																																																																																																																																																																																																																																																																																																		
프로젝트 내용	<div> <div> <div> <div>1) 2022년 6월1일~6월 8일 여의도 상암지구 이용자들 2000명 데이터만 따로 추출</div> <div>2) 범주형 변수들 전처리 및 RENT_SPOT, HOLIDAY, PARK 변수들 새로 생성</div> <div>3) BIC 기준(k=2일때 -10657.99, k=3일때 -10122.29, k=4일때 -9835.712) component가 2인 GMM 모형 fitting /추정 latent variable: 속도, 긴박성</div> <div>4) BIC 기준(k=3일때 -155586.71, k=4일때 -14672.784, k=5일때 -14601.217) component가 3인 MoE 모형 fitting / 추정 latent variable: riding purpose, willpower(desire) of riding</div> <div>5) MoE 모형 으로 나온 cluster 별 평균 운동량 예측(cluster1: 83.01738, cluster2: 80.8889, cluster3: 191.2407)</div> <div>6) 운동량의 RMSE와 MAE를 기준으로 MoE를 RANDOM FOREST와 LINEAR MODEL과 비교</div> <div>7) RMSE(randomforest: 61.07248, linear model: 30.85834, MoE: 27.87566) -> MoE가 가장 좋은 성능 보임</div> </div> <div> <div>   </div> <table> <tr> <th></th><th>K=2</th><th>K=3</th><th>K=4</th></tr> <tr> <td>Log-likelihood</td><td>-10657.99</td><td>-10122.29</td><td>-9835.712</td></tr> <tr> <td>df</td><td>11</td><td>17</td><td>23</td></tr> <tr> <td>BIC</td><td>-21399.58</td><td>-20373.8</td><td>-19846.24</td></tr> <tr> <td>ICL</td><td>-21539.32</td><td>-20722.44</td><td>-20295.07</td></tr> <tr> <td>CLUSTERING</td><td>1:1114(55.7%), 2:886(44.3%)</td><td>1:804(40.2%), 2: 934(46.7%), 3: 262(13.1%)</td><td>1: 596(29.8%), 2:596(29.8%), 3:66(3.3%), 4:742(37.1%)</td></tr> </table> <div> <table> <tr> <th></th><th colspan="5">CLUSTER1</th><th colspan="5">CLUSTER2</th><th colspan="5">CLUSTER3</th></tr> <tr> <td>MEAN(MOVE_TIME)</td><td colspan="5">31.04811</td><td colspan="5">32.12023</td><td colspan="5">86.47197</td></tr> <tr> <td>MEAN(CARBON_AMT)</td><td colspan="5">0.7330241</td><td colspan="5">0.7811445</td><td colspan="5">1.66472</td></tr> <tr> <td rowspan="2">AGE_TYPE</td><td>10대</td><td>20대</td><td>30대</td><td>40대</td><td>50대</td><td>10대</td><td>20대</td><td>30대</td><td>40대</td><td>50대</td><td>10대</td><td>20대</td><td>30대</td><td>40대</td><td>50대</td></tr> <tr> <td>4%</td><td>32%</td><td>25%</td><td>21%</td><td>15%</td><td>5%</td><td>31%</td><td>27%</td><td>17%</td><td>17.5%</td><td>3%</td><td>21%</td><td>25%</td><td>31%</td><td>17.7%</td></tr> <tr> <td rowspan="2">GENDER_CD</td><td>남</td><td>여</td><td>남</td><td>여</td><td>남</td><td>남</td><td>여</td><td>남</td><td>여</td><td>남</td><td>남</td><td>여</td><td>남</td><td>여</td><td>남</td></tr> <tr> <td>52%</td><td>47%</td><td>47%</td><td>47%</td><td>47%</td><td>39%</td><td>60%</td><td>39%</td><td>60%</td><td>39%</td><td>94%</td><td>6%</td><td>94%</td><td>6%</td><td>6%</td></tr> <tr> <td rowspan="2">RENT_SPOT</td><td>임부</td><td>지하철</td><td>주거</td><td>임부</td><td>지하철</td><td>주거</td><td>임부</td><td>지하철</td><td>주거</td><td>임부</td><td>지하철</td><td>주거</td><td>임부</td><td>지하철</td><td>주거</td></tr> <tr> <td>28.3%</td><td>55.5%</td><td>36.6%</td><td>20.4%</td><td>42.8%</td><td>36.6%</td><td>28.3%</td><td>55.5%</td><td>36.6%</td><td>28.3%</td><td>40.5%</td><td>22.7%</td><td>28.3%</td><td>55.5%</td><td>36.6%</td></tr> <tr> <td rowspan="2">RENT_HR</td><td>06</td><td>09</td><td>12</td><td>15</td><td>18</td><td>21</td><td>06</td><td>09</td><td>12</td><td>15</td><td>18</td><td>21</td><td>06</td><td>09</td><td>12</td></tr> <tr> <td>09</td><td>12</td><td>15</td><td>18</td><td>21</td><td>00</td><td>09</td><td>12</td><td>15</td><td>18</td><td>21</td><td>00</td><td>09</td><td>12</td><td>15</td></tr> <tr> <td rowspan="2">RENT_TYPE</td><td>86.7</td><td>13%</td><td>21</td><td>20</td><td>1%</td><td>14</td><td>11</td><td>17</td><td>20</td><td>2%</td><td>8.3</td><td>9.2</td><td>1%</td><td>2%</td><td>27</td></tr> <tr> <td>9%</td><td>13%</td><td>8%</td><td>9%</td><td>9%</td><td>14</td><td>11</td><td>17</td><td>20</td><td>2%</td><td>8.3</td><td>9.2</td><td>1%</td><td>2%</td><td>27</td></tr> <tr> <td rowspan="2">holiday</td><td>일</td><td>기</td><td>일</td><td>기</td><td>일</td><td>일</td><td>기</td><td>일</td><td>기</td><td>일</td><td>일</td><td>기</td><td>일</td><td>기</td><td>일</td></tr> <tr> <td>74.2%</td><td>25.7%</td><td>25.7%</td><td>26.4%</td><td>23.5%</td><td>23.5%</td><td>69.2%</td><td>69.2%</td><td>69.2%</td><td>69.2%</td><td>30.7%</td><td>30.7%</td><td>30.7%</td><td>30.7%</td><td>30.7%</td></tr> <tr> <td rowspan="2">park</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr> <td>49.4%</td><td>50.6%</td><td>50.6%</td><td>50.6%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td><td>49.4%</td></tr> <tr> <td rowspan="2">USE_CNT</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr> <td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr> <tr> <td rowspan="2"></td><td>95.3%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td></tr> <tr> <td>95.3%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td><td>4.6%</td></tr> </table> </div> </div> </div> </div>					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%)		CLUSTER1					CLUSTER2					CLUSTER3					MEAN(MOVE_TIME)	31.04811					32.12023					86.47197					MEAN(CARBON_AMT)	0.7330241					0.7811445					1.66472					AGE_TYPE	10대	20대	30대	40대	50대	10대	20대	30대	40대	50대	10대	20대	30대	40대	50대	4%	32%	25%	21%	15%	5%	31%	27%	17%	17.5%	3%	21%	25%	31%	17.7%	GENDER_CD	남	여	남	여	남	남	여	남	여	남	남	여	남	여	남	52%	47%	47%	47%	47%	39%	60%	39%	60%	39%	94%	6%	94%	6%	6%	RENT_SPOT	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거	28.3%	55.5%	36.6%	20.4%	42.8%	36.6%	28.3%	55.5%	36.6%	28.3%	40.5%	22.7%	28.3%	55.5%	36.6%	RENT_HR	06	09	12	15	18	21	06	09	12	15	18	21	06	09	12	09	12	15	18	21	00	09	12	15	18	21	00	09	12	15	RENT_TYPE	86.7	13%	21	20	1%	14	11	17	20	2%	8.3	9.2	1%	2%	27	9%	13%	8%	9%	9%	14	11	17	20	2%	8.3	9.2	1%	2%	27	holiday	일	기	일	기	일	일	기	일	기	일	일	기	일	기	일	74.2%	25.7%	25.7%	26.4%	23.5%	23.5%	69.2%	69.2%	69.2%	69.2%	30.7%	30.7%	30.7%	30.7%	30.7%	park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49.4%	50.6%	50.6%	50.6%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	USE_CNT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		95.3%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	95.3%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%
	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%)																																																																																																																																																																																																																																																																																																																																																																
	CLUSTER1					CLUSTER2					CLUSTER3																																																																																																																																																																																																																																																																																																																																																								
MEAN(MOVE_TIME)	31.04811					32.12023					86.47197																																																																																																																																																																																																																																																																																																																																																								
MEAN(CARBON_AMT)	0.7330241					0.7811445					1.66472																																																																																																																																																																																																																																																																																																																																																								
AGE_TYPE	10대	20대	30대	40대	50대	10대	20대	30대	40대	50대	10대	20대	30대	40대	50대																																																																																																																																																																																																																																																																																																																																																				
	4%	32%	25%	21%	15%	5%	31%	27%	17%	17.5%	3%	21%	25%	31%	17.7%																																																																																																																																																																																																																																																																																																																																																				
GENDER_CD	남	여	남	여	남	남	여	남	여	남	남	여	남	여	남																																																																																																																																																																																																																																																																																																																																																				
	52%	47%	47%	47%	47%	39%	60%	39%	60%	39%	94%	6%	94%	6%	6%																																																																																																																																																																																																																																																																																																																																																				
RENT_SPOT	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거	임부	지하철	주거																																																																																																																																																																																																																																																																																																																																																				
	28.3%	55.5%	36.6%	20.4%	42.8%	36.6%	28.3%	55.5%	36.6%	28.3%	40.5%	22.7%	28.3%	55.5%	36.6%																																																																																																																																																																																																																																																																																																																																																				
RENT_HR	06	09	12	15	18	21	06	09	12	15	18	21	06	09	12																																																																																																																																																																																																																																																																																																																																																				
	09	12	15	18	21	00	09	12	15	18	21	00	09	12	15																																																																																																																																																																																																																																																																																																																																																				
RENT_TYPE	86.7	13%	21	20	1%	14	11	17	20	2%	8.3	9.2	1%	2%	27																																																																																																																																																																																																																																																																																																																																																				
	9%	13%	8%	9%	9%	14	11	17	20	2%	8.3	9.2	1%	2%	27																																																																																																																																																																																																																																																																																																																																																				
holiday	일	기	일	기	일	일	기	일	기	일	일	기	일	기	일																																																																																																																																																																																																																																																																																																																																																				
	74.2%	25.7%	25.7%	26.4%	23.5%	23.5%	69.2%	69.2%	69.2%	69.2%	30.7%	30.7%	30.7%	30.7%	30.7%																																																																																																																																																																																																																																																																																																																																																				
park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																																																																																																				
	49.4%	50.6%	50.6%	50.6%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%	49.4%																																																																																																																																																																																																																																																																																																																																																				
USE_CNT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																																																																																																				
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1																																																																																																																																																																																																																																																																																																																																																				
	95.3%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%																																																																																																																																																																																																																																																																																																																																																				
	95.3%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%	4.6%																																																																																																																																																																																																																																																																																																																																																				
기대 효과	<div>1) 서울시(여의도, 상암지역)의 공공자전거 이용자 분류를 MoE를 통해 분류하여 이들이 어떤 component에 속할지 예측.</div> <div>2) 얼마만큼의 운동량을 가지는지 예측.</div>																																																																																																																																																																																																																																																																																																																																																																		

주요 프로젝트 수행 이력(대학원 프로젝트)

4. 딥러닝 토픽 프로젝트

제목	chatgpt에 대한 tweets의 sentiment 분류
프로젝트 배경 및 개요	<ul style="list-style-type: none">• chatgpt 출시 이후 이에 대한 다양한 의견을 나타내는 tweets들이 생겨나고 있다.• chatgpt 입장에서는 향후 개선을 위해 다양한 의견을 파악하는 것이 중요하다고 판단한다.• RNN,LSTM,GRU 모델을 활용하여 chatgpt에 대한 tweets의 sentiment(긍정,중립,부정)가 무엇인지 예측한다.
사용언어/모델	PYTHON/ RNN, LSTM, GRU
프로젝트 내용	<div><div><div><div>1) kaggle에서 12000개의 chatgpt에 대한 tweets 데이터 추출 / 데이터 전처리: tweets에서 특수문자를 제거하여 text부분만 따로 남긴다. text를 정수로 인코딩한 후 시퀀스의 길이를 똑같이 맞춘다.</div><div>2) RNN 모형 FITTING(epochs=10, batch size= 32, optimizer=adam, loss=categorical entropy, activation=softmax, 유닛수=64, bidirectional(return_sequences=True) 1번,bidirectional 1번 dropout(0.2)옵션 1번 추가)</div><div>3) LSTM 모형 FITTING(epochs=10, batch size 32, optimizer=adam, loss=categorical entropy, activation=softmax, 유닛수=128,bidirectional(return_sequences=True) 1번, dropout(0.2)옵션 1번 추가)</div><div>4) GRU 모형 FITTING(epochs=10, batch size 32, optimizer=adam, loss=categorical entropy, activation=softmax, 유닛수=128,bidirectional(return_sequences=True, recurrent_dropout=0.2) 1번, bidirectional(recurrent_dropout=0.2) 1번 dropout(0.2)옵션 1번 추가)</div><div>5) test data accuracy 기준(RNN: 64.46%, LSTM: 69.29, GRU: 72.29%) GRU 모형 최종 선택</div></div><div><div><div><div><div><div>Training and Validation Loss</div><div>Accuracy</div></div><div><div><div>Training and Validation Loss</div><div>Accuracy</div></div><div><div><div>Training and Validation Loss</div><div>Accuracy</div></div></div></div></div></div></div></div></div></div>
기대 효과	향후 chatgpt에 대한 의견을 나타내는 tweets이 앞으로도 계속 생겨날 것이고 이 모델을 통해 chatgpt에 대한 전반적인 여론이 어떠한지 파악할 수 있다.

주요 프로젝트 수행 이력(대학원 프로젝트)

5. 일반화 선형모형 프로젝트

제목

지구근처 소행성 종류 분류 및 잠재적 위험 유무 판별

프로젝트 배경 및 개요

- NASA에 따르면 지구 근처에 있는 잠재적 충돌 위험이 있는 소행성들이 존재한다고 한다.
- 지구 근처에 있는 소행성들의 종류(Apollo, Amor, Aten)를 판별 & 잠재적 충돌 위험성이 있는지 판별

사용언어

R

프로젝트 내용

1) NASA에 있는 15619개의 소행성 데이터들을 추출하고 결측치를 제거 및 범주형 변수를 수치형으로 처리한다. (11개의 설명변수)

2) 소행성 분류를 위해 'VGAM' 패키지 안에 있는'step4vglm' 함수를 사용하여 변수 선택(AIC 기준)후(변수(X2, X7, X9)변수가 선택) baseline category logit model fitting(결과: residual deviance는 1507.743, AIC는 1523.743, Orbit.Eccentricity(X2)와 Perihelion.Distance..AU(X7)가 유의한 변수)

3) 잠재적 위험 유무 판별을 위해 R 내장 함수인 step function을 사용(AIC를 기준으로 변수 선택) orbit axis, Asteroid.Magnitude(X11),Minimum.Orbit.Intersection.Distance(X10) 변수로 link function을 logit으로 설정 후 binomial glm model fitting(결과: Null deviance: 11056.0, Residual deviance: 3191.7 on 15606 degrees of freedom, AIC: 3203.4)

4) R 내장 함수인 step function을 사용(AIC를 기준으로 변수 선택) Orbit Axis..AU.(X1), Minimum.Orbit.Intersection.Distance..AU.(X10), Asteroid.Magnitude(X11) 변수로 link function을 probit으로 설정 후 binomial glm model fitting(결과: Null deviance: 11056.0, Residual deviance: 3187.3, AIC: 3195.3)

Category	Count
Amor Asteroid	5918
Amor Asteroid (Hazard)	99
Apollo Asteroid	6940
Apollo Asteroid (Hazard)	1520
Aten Asteroid	987
Aten Asteroid (Hazard)	155

Variable	mean	variance	standard deviation	min	max
Orbit.Axis..AU.	1.7821	0.363095	0.60257	0.5798	21.3954
Orbit.Eccentricity	0.4498	0.03997964	0.1760103	0.0044	0.9695
Orbit.Inclination..deg.	12.9349	127.7286	11.30171	0.0147	154.3751
Perihelion.Argument..deg.	181.4619	10786.54	103.8583	0.0081	359.9942
Node.Longitude..deg.	172.681	19684.74	140.367	0.007	359.998
Mean.Annomaly..deg.	172.8232	13490.38	116.1481	0.0031	359.9982
Perihelion.Distance..AU.	0.9154	0.05753963	0.2398742	0.0707	1.3000
Aphelion.Distance..AU.	2.649	1.304093	1.141969	0.990	41.540
Orbital.Period..yr.	2.475	2.30184	1.517182	0.440	98.970
Minimum.Orbit.Intersection.Distance..AU.	0.1019	0.0112804	0.1059736	0.0000	0.7069
Asteroid.Magnitude	22.29	9.093577	3.015536	9.45	33.20

Variable	Estimate	Std. Error	z value	Pr(> z)
(Intercept):1	7.2878	0.9398	7.755	8.83 × 10 ⁻¹⁵ ***
(Intercept):2	20.5020	0.8270	24.792	< 2 × 10 ⁻¹⁶ ***
Orbit.Eccentricity:1	10.3833	0.5143	20.188	< 2 × 10 ⁻¹⁶ ***
Orbit.Eccentricity:2	10.2610	0.4721	21.733	< 2 × 10 ⁻¹⁶ ***
Perihelion.Distance..AU.:1	45.1189	1.2302	36.677	< 2 × 10 ⁻¹⁶ ***
Perihelion.Distance..AU.:2	13.9332	0.6289	22.154	< 2 × 10 ⁻¹⁶ ***
Orbital.Period..yr.:1	-0.8533	0.5862	-1.456	0.145
Orbital.Period..yr.:2	-0.7863	0.5173	-1.520	0.129

Variable	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.38568	0.13154	-40.944	< 2 × 10 ⁻¹⁶ ***
Orbit.Axis..AU.	-0.06237	0.02587	-2.411	0.0159 *
Asteroid.Magnitude	-2.33622	0.05471	-42.702	< 2 × 10 ⁻¹⁶ ***
Minimum.Orbit.Intersection.Distance..AU.	-6.28228	0.17090	-36.759	< 2 × 10 ⁻¹⁶ ***



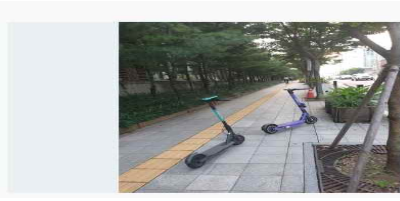

- Null deviance: 11056.0 on 15617 degrees of freedom
- Residual deviance: 3187.3 on 15614 degrees of freedom
- AIC: 3195.3

기대 효과

지구 근처의 소행성들이 어떤 종류의 소행성인지 판별
소행성이 잠재적 충돌 위험이 있는지 판별

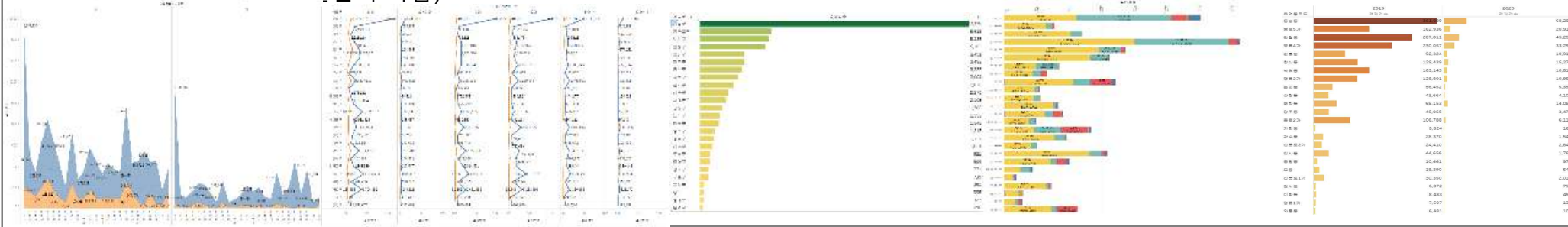
주요 프로젝트 수행 이력(대외활동 프로젝트)

6. 경기미래기술학교

제목	헬멧 착용 여부 & 전동 킥보드 불법 주정차 여부 판별
프로젝트 배경	<ul style="list-style-type: none">전동킥보드 이용 시 헬멧 미착용 문제, 불법 주정차 문제가 존재문제 해결을 위해 이용 전 헬멧 착용 여부, 이용 후 불법 주정차 여부를 판별하는 서비스 필요전동킥보드를 타기 전에 어플로 사진을 찍어 헬멧 착용을 인증하는 과정과, 타고난 후 주정차 부분을 사진을 찍어 인증해야 하는 부분이 필요하다고 생각
사용언어	PYTHON
프로젝트 내용	<div><div><div><div>1) 웹크롤링(헬멧 착용&미착용 이미지 : 8000장, 전동킥보드 주차 이미지 : 4000장)</div><div>2) ROBOFLOW라는 툴을 활용하여 이미지 라벨링 작업 실시(헬멧 미착용:head, 헬멧착용: helmet, 불법주정차o: illegal, 불법주정차x: legal)</div><div>3) YOLOV5라는 모델을 활용하여 헬멧 착용 여부 & 불법 주정차 여부 판별 모델 총 2개 fitting</div><div>4) epoch, batch size 등 MAP기준 가장 좋은 parameter 선택(batchsize: 40, epochs: 40)</div><div>5) 학습데이터의 map는 78%, precision,recall은 약 96%의 결과를 보임</div></div><div><div></div><div></div><div></div><div></div></div></div></div>
기대 효과	<div><div>1) 공유 킥보드 사용자:안전사고 예방, 안전 자가진단</div><div>2) 지방자치단체: 시민 보행불편 경감, 민원 처리부담 완화</div><div>3) 공유 킥보드 업체: 공유킥보드 인식 제고, 불법주차 견인비용 절감</div></div>

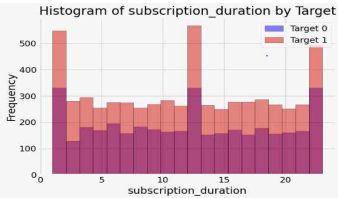
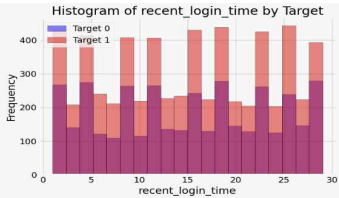
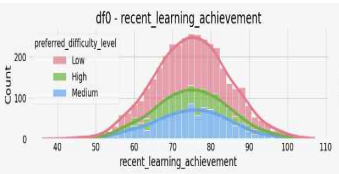
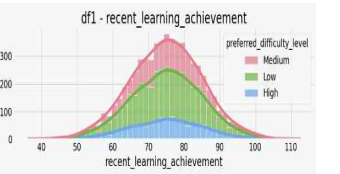
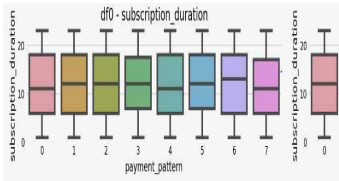
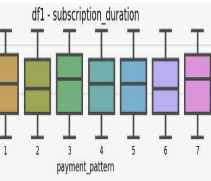
주요 프로젝트 수행 이력(대외활동 프로젝트)

7. 데이터 청년 캠퍼스

제목	코로나 전후 문화/여가 업종 주말 결제건수 결제금액 비교
프로젝트 배경 및 개요	<ul style="list-style-type: none"> 코로나 이후 문화/여가 업종은 매출 부분에서 많은 타격을 입었다. 코로나 전후 결제건수와 결제금액의 변화는 자치구별로 다르고 이에 따라 변화의 정도에 따라 지원해야 하는 정도는 다르다.
사용도구	EXCEL, Tableau, R
프로젝트 내용	<p>1) 2019년 3월~5월과 2020년 3월~5월의 결제건수와 결제금액을 성별, 연령대별, 자치구별, 업종별로 시각화하여 살펴본다.</p> <p>2) 이러한 결제건수와 결제금액의 차이가 유의한지 t-test를 실시한다.(p-value 값이 0.05보다 작아 유의하다는 결론이 나옴)</p> <p>3) 문화/여가 업종이 많은 종로구의 음식점들의 결제건수와 결제금액도 시각화하고 이들의 차이가 유의한지 t-test를 실시한다. (p-value 값이 0.05보다 작아 유의하다는 결론이 나옴)</p> 
결과	<p>1) 분석 결과 코로나 19로 인해 서울특별시 문화여가 업종의 주말/공휴일 결제건수 및 결제금액이 감소하였으며, 업종 중분류 중 극장, 스포츠, 음악의 소비가 크게 감소하였음</p> <p>2) 업종 분류에 따른 소비 비율은 오락부분이 크게 증가 -> 사회적 거리두기로 소비 활동의 범위가 좁혀지고 소규모 소비 활동이 활성화되었기 때문</p> <p>3) 종로구의 경우 코로나19로 휴·폐업한 공연장이 많아 결제금액이 크게 감소하였고, 이로 인해 방문객이 줄어들어 근처 음식점의 소비 또한 감소하였음</p>

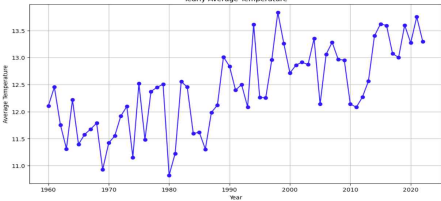
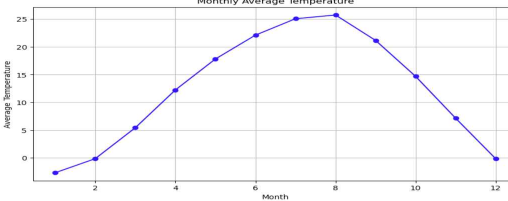
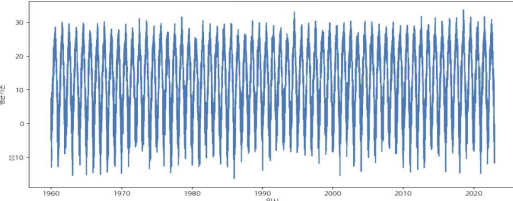
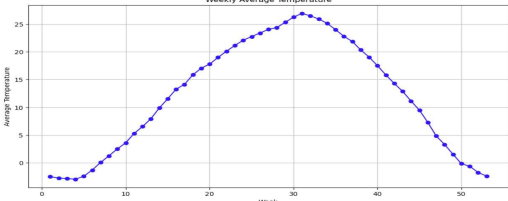
주요 프로젝트 수행 이력(경진대회 프로젝트)

8. 데이콘 경진대회

제목	학습 플랫폼 이용자 구독 갱신 예측 해커톤
프로젝트 배경 및 개요	<ul style="list-style-type: none">• 학습 플랫폼 입장에서는 다음달에도 이용자들이 구독을 할지 안할지 미리 파악하는 것이 중요하다• 학습 플랫폼 이용자들이 다음달 구독 여부를 판별한다
사용도구	PYTHON
프로젝트 내용	<div><div><div><div><div>1) 변수들의 다양한 조합으로 EDA를 수행한다.</div><div>2) 수치형 변수들의 조합하여 8개 추가 변수를 생성한다. (feature engineering)</div><div>3) lower bound와 upper bound 설정을 통해 수치형 변수 이상치 제거</div><div>4) subscription_type, payment_pattern 원핫 인코딩, preferred_difficult_level 라벨 인코딩 수행</div><div>5) 수치형 변수에 루트 값을 씌운 변수와 루트 1/3을 씌운 변수를 추가 생성</div><div>6) 수치형 변수 전체 min max scaler로 표준화 수행(정규분포를 띄도록 만들기 위해) -> 2차 다항식 변수 추가 생성</div><div>7) votingclassifier(xgboost+catboost+lightgb), stacking(xgboost+catboost), lightgbm, catboost, xgboost 등 다양한 모델 적용 결과 macrof1(0.51) 기준 가장 좋은 성능의 모델을 선택한다.</div></div><div><div>catboost(hyper parameter tuning: optuna)</div><div><div><div><div>Histogram of subscription_duration by Target</div><div><div>Target 0</div><div>Target 1</div></div></div><div><div>Histogram of recent_login_time by Target</div><div><div>Target 0</div><div>Target 1</div></div></div></div><div><div><div>df0 - recent_learning_achievement</div><div><div>preferred_difficult_level</div><div>Low</div><div>High</div><div>Medium</div></div></div><div><div>df1 - recent_learning_achievement</div><div><div>preferred_difficult_level</div><div>Low</div><div>High</div><div>Medium</div></div></div></div><div><div><div>df0 - subscription_duration</div><div><div>payment_pattern</div><div>0</div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div><div>6</div><div>7</div></div></div><div><div>df1 - subscription_duration</div><div><div>payment_pattern</div><div>0</div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div><div>6</div><div>7</div></div></div></div></div></div></div></div></div>
기대 효과	아무래도 처음 나가는 경진대회이다 보니 EDA에서도 유의미하지 않은 시각화를 실시했다고 생각한다. 다음부터는 인사이트를 발견할 수 있도록 유의미한 시각화를 실시해야 한다는 걸 느꼈다. 전처리 또한 어느 정도 전처리가 완료 되었음에도 다른 방법으로 다시 표준화를 실시하는 건 잘못된 수행방법이라 생각한다. 아무래도 처음이다보니 여러가지 방법을 적용해 보고 싶어서 너무 무리하여 전처리를 진행했다. 모델링 같은 경우에도 좀 더 다양한 방법으로 시도해봐야 했었다는 아쉬움이 남는다.

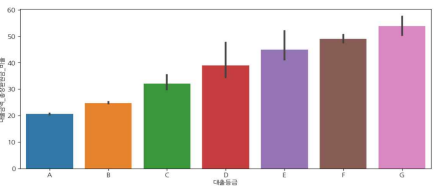
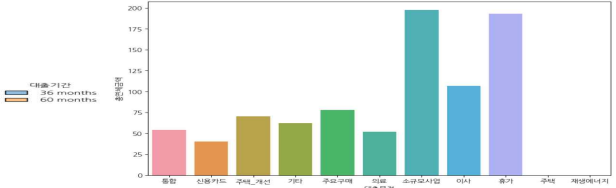
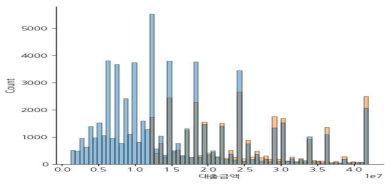
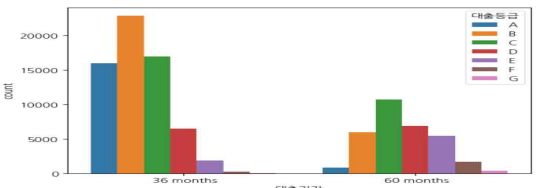
주요 프로젝트 수행 이력(경진대회 프로젝트)

9. 데이콘 경진대회

제목	서울시 평균기온 예측 해커톤
프로젝트 목표	<ul style="list-style-type: none">1960년부터 2022년의 서울시 기후 데이터를 이용하여 2023년의 평균 기온 예측
사용도구	PYTHON
프로젝트 내용	<div><div><div>1) 데이터 전처리: 최고기온,최저기온(선형보간법), 일조합(일조율을 독립변수로 한 선형모형), 일사합(일조합을 독립변수로 한 선형모형), 평균풍속(월별 중앙값), 강수량(bfill)</div><div>2) feature engineering: EDA후 년,월,일 ,주차, 계절(one-hot encoding), daysin,daycos(주기함수) 생성// test data는 날짜 변수 밖에 없어서 일교차,일조율, 평균습도, 강수량은 catboost로 예측하여 생성</div><div>3) MAE(2.93) 기준 최적의 머신러닝 모델 선택 catboost</div></div><div><div></div><div></div><div></div><div></div></div></div>
느낀점	향후 서울시의 평균 기온을 예측submission data의 날짜 변수밖에 존재하지 않아 기존 train data의 변수들과 똑같은 변수로 맞추기 위해 모델링을 활용하여 새로운 변수를 생성하였던 것이 좋은 결과로 이어졌다고 생각한다. 이번 대회를 통해 feature engineering의 중요성을 확실하게 느끼게 되었고 다음 대회 때는 더 성능을 높일 수 있는 방안을 더 많이 생각해봐야 될 것 같다.

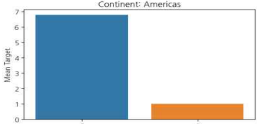
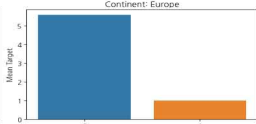
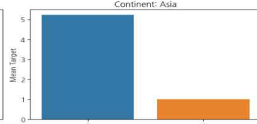
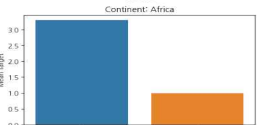
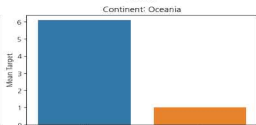
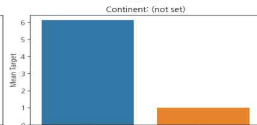
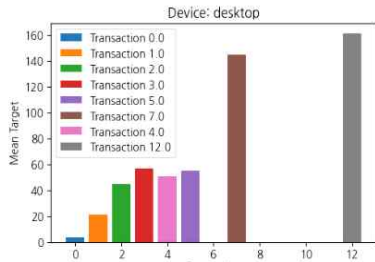
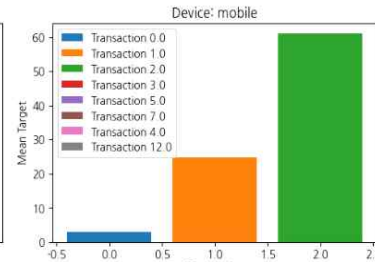
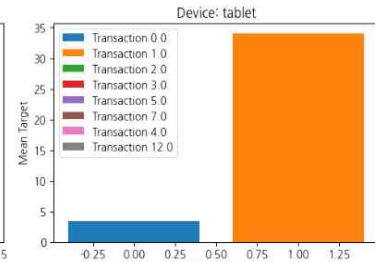
주요 프로젝트 수행 이력(경진대회 프로젝트)

10. 데이콘 경진대회

제목	고객 대출등급 분류 AI 해커톤
프로젝트 목표	<ul style="list-style-type: none">주어진 금융 데이터를 활용하여 고객의 대출등급을 예측
사용도구/최종모델	PYTHON/ stacking model(decisiontree+xcboost+lgbm/ 최종모델:randomforest)
프로젝트 내용	<div><div><div>1) 다양한 변수의 조합을 통해 EDA 수행</div><div>2) 데이터 전처리: 월_대출금액, 월_대출대비_소득비율, 계좌수, 대출금액_총상환이자_비율, 대출금액_총상환원금_비율, 상환이자_상환원금, 총상환액, 소득대비_총상환액_비율, 대출대비_총상환액_비율, 기간대비_총상환액_비율, 대출대비_총상환원금_비율, 대출대비_총상환이자_비율, 소득대비_총상환원금_비율, 소득대비_총상환이자_비율, 기간대비_총상환원금_비율, 기간대비_총상환이자_비율, 월_이자_지불액 변수 feature engineering으로 새로 생성</div><div>3) 대출목적 변수 binary encoding 수행/수치형 변수 모두 standard scaler를 통해 표준화 처리/ lda를 통해 나온 중요 변수 2개 독립변수로 추가</div><div>4) decision tree parameter(max_depth=15, min_samples_split=5, min_samples_leaf=2, random_state=2024), xcboost parameter(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=2024), lgbm parameter(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=2024)를 설정하고 stacking model(최종모형 randomforest(n_estimators=100,random_state=2024)) fitting(macro f1: 약 0.95) -> 최종 모델 선정</div></div><div><div></div></div></div>
느낀점	향후 고객 대출등급 예측 이 데이터는 수치형 변수 데이터가 많은데 이 수치형 변수들의 다양한 조합으로 feature engineering을 수행하여 다양한 변수를 생성했기 때문에 좋은 결과가 날로 수 있었다고 생각한다. 다음 대회 때도 다양한 변수들의 조합으로 적절한 feature engineering을 해야겠다는 생각을 했다.

주요 프로젝트 수행 이력(경진대회 프로젝트)

11. 데이콘 경진대회

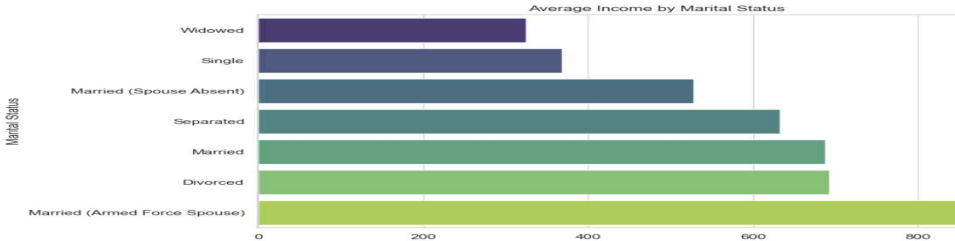
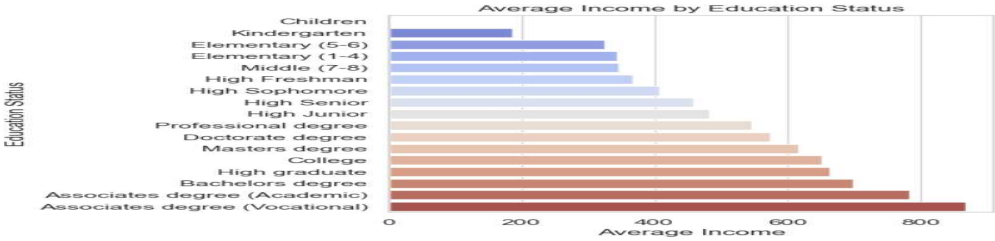
제목	웹 로그 기반 조회수 예측 해커톤
프로젝트 목표	<ul style="list-style-type: none">웹 로그 기반 데이터를 활용하여 조회수를 예측
사용도구/최종모델	PYTHON/ lgbm
프로젝트 내용	<div><div><div>1) 다양한 변수의 조합을 통해 EDA 수행</div><div>2) 데이터 전처리: 'browser', 'OS', 'device', 'country', 'continent', 'subcontinent', 'traffic_source', 'traffic_medium', 'referral_path', 'keyword' 범주형 변수들 원핫인코딩 처리/'quality', 'duration', 'transaction', 'transaction_revenue' 변수 standard scaler를 활용하여 표준화/ 수치형 변수들의 조합으로 22개 변수 추가 생성(feature engineering)</div><div>3) parameter(random_state=2024, learning_rate=0.05, max_depth=20, min_child_samples=30, n_estimators=180, num_leaves=40)를 설정하여 lgbm 모델 fitting(RMSE: 2.95) -> 최종모델 선정</div></div><div><div><div><div>Continent: Americas</div></div><div><div>Continent: Europe</div></div><div><div>Continent: Asia</div></div></div><div><div><div>Continent: Africa</div></div><div><div>Continent: Oceania</div></div><div><div>Continent: (not set)</div></div></div></div><div><div><div>Device: desktop</div></div><div><div>Device: mobile</div></div><div><div>Device: tablet</div></div></div></div>

느낀점

단일 모델이 아닌 ENSEMBLE방식을 다양하게 활용하여 모델링을 진행했으면 더 좋은 성과가 나오지 않았을까 생각한다. 다음 대회에 참여하게 된다면 다양한 ENSEMBLE 방법을 적용하여 더 성능이 높은 모델을 만들도록 해야겠다.

주요 프로젝트 수행 이력(경진대회 프로젝트)

12. 데이콘 경진대회

제목	소득 예측 AI 해커톤																																																					
프로젝트 목표	<ul style="list-style-type: none">개인 정보 데이터를 활용하여 소득 예측																																																					
사용도구/최종모델	PYTHON/ voting regressor(lgbm+catboost+hgb+gb)																																																					
프로젝트 내용	<div><div><div>1) 다양한 변수의 조합을 통해 EDA 수행</div><div>2) "Employment_Status", "Industry_Status", "Occupation_Status", "Race", "Hispanic_Origin", "Marital_Status", "Household_Summary", "Citizenship", "Tax_Status", "Income_Status", "Education_Status", "Birth_Country" 범주형 변수들을 Label Encoding 처리</div><div>3) Gender 변수 M=1, F=0으로 처리</div><div>4) 수치형 변수들의 조합으로 'Age_Working_Week', 'Gains_Losses', 'Age_Dividends', 'Age_Losses', 'Working_Week_Dividends', 'Gains_Dividends', 'Working_Week_Losses', 'Age_Working_Week_Gains', 'Working_Week_Gains_Losses' Feature Engineering을 통해 생성</div><div>5) optuna 활용한 parameter로 voting regressor(lgbm+catboost+hgb+gb)model fitting(RMSE: 541.85_ <- 최종 모델 선정</div></div><div><div><div>Average Income by Marital Status</div><table border="1"><thead><tr><th>Marital Status</th><th>Average Income</th></tr></thead><tbody><tr><td>Widowed</td><td>~350</td></tr><tr><td>Single</td><td>~400</td></tr><tr><td>Married (Spouse Absent)</td><td>~550</td></tr><tr><td>Separated</td><td>~650</td></tr><tr><td>Married</td><td>~700</td></tr><tr><td>Divorced</td><td>~700</td></tr><tr><td>Married (Armed Force Spouse)</td><td>~800</td></tr></tbody></table></div><div><div>Average Income by Education Status</div><table border="1"><thead><tr><th>Education Status</th><th>Average Income</th></tr></thead><tbody><tr><td>Children</td><td>~200</td></tr><tr><td>Kindergarten</td><td>~300</td></tr><tr><td>Elementary (5-6)</td><td>~350</td></tr><tr><td>Elementary (1-4)</td><td>~400</td></tr><tr><td>Middle (7-8)</td><td>~450</td></tr><tr><td>High Freshman</td><td>~500</td></tr><tr><td>High Sophomore</td><td>~550</td></tr><tr><td>High Senior</td><td>~600</td></tr><tr><td>High Junior</td><td>~650</td></tr><tr><td>Professional degree</td><td>~700</td></tr><tr><td>Doctorate degree</td><td>~750</td></tr><tr><td>Masters degree</td><td>~780</td></tr><tr><td>College</td><td>~800</td></tr><tr><td>High graduate</td><td>~800</td></tr><tr><td>Bachelors degree</td><td>~800</td></tr><tr><td>Associates degree (Academic)</td><td>~800</td></tr><tr><td>Associates degree (Vocational)</td><td>~800</td></tr></tbody></table></div></div></div> <td><p>다양한 모델링 방법과 앙상블 과정을 통해 코드 제출을 했지만 생각보다 높은 순위를 내지 못했다. EDA를 통해 적절한 전처리를 수행했어야 하는데 그러지 못해 뭔가 Feature Engineering하여 전처리 하는 과정에서 유의미한 결과를 내지 못했기 때문인 것 같다. 다음에는 EDA를 활용하여 유의미한 인사이트를 발굴하도록 더 데이터 시각화에 중점을 두어서 분석을 진행해 봐야 겠다.</p></td>	Marital Status	Average Income	Widowed	~350	Single	~400	Married (Spouse Absent)	~550	Separated	~650	Married	~700	Divorced	~700	Married (Armed Force Spouse)	~800	Education Status	Average Income	Children	~200	Kindergarten	~300	Elementary (5-6)	~350	Elementary (1-4)	~400	Middle (7-8)	~450	High Freshman	~500	High Sophomore	~550	High Senior	~600	High Junior	~650	Professional degree	~700	Doctorate degree	~750	Masters degree	~780	College	~800	High graduate	~800	Bachelors degree	~800	Associates degree (Academic)	~800	Associates degree (Vocational)	~800	<p>다양한 모델링 방법과 앙상블 과정을 통해 코드 제출을 했지만 생각보다 높은 순위를 내지 못했다. EDA를 통해 적절한 전처리를 수행했어야 하는데 그러지 못해 뭔가 Feature Engineering하여 전처리 하는 과정에서 유의미한 결과를 내지 못했기 때문인 것 같다. 다음에는 EDA를 활용하여 유의미한 인사이트를 발굴하도록 더 데이터 시각화에 중점을 두어서 분석을 진행해 봐야 겠다.</p>
Marital Status	Average Income																																																					
Widowed	~350																																																					
Single	~400																																																					
Married (Spouse Absent)	~550																																																					
Separated	~650																																																					
Married	~700																																																					
Divorced	~700																																																					
Married (Armed Force Spouse)	~800																																																					
Education Status	Average Income																																																					
Children	~200																																																					
Kindergarten	~300																																																					
Elementary (5-6)	~350																																																					
Elementary (1-4)	~400																																																					
Middle (7-8)	~450																																																					
High Freshman	~500																																																					
High Sophomore	~550																																																					
High Senior	~600																																																					
High Junior	~650																																																					
Professional degree	~700																																																					
Doctorate degree	~750																																																					
Masters degree	~780																																																					
College	~800																																																					
High graduate	~800																																																					
Bachelors degree	~800																																																					
Associates degree (Academic)	~800																																																					
Associates degree (Vocational)	~800																																																					
느낀점																																																						

보유 직무 역량

보유 SW 직무 역량

	사용 가능 Tool	활용능력	활용 내역
데이터 추출	SQL, EXCEL	중	프로그래머스 고득점 kit 풀이
데이터 시각화	Tableau, R, PYTHON	중	데이터 청년 캠퍼스 데이터 시각화 교육/ 한국지능정보시스템 학회 참여 발표 수행
머신러닝, 딥러닝	R, PYTHON	중상	대학원 연구 및 경진대회 참가

직무 관련 교육

<인공지능기초다지기>코칭스터디 (네이버커넥트재단)	<ul style="list-style-type: none">python 기초 문법 및 라이브러리(numpy,pandas) 학습 및 실습머신러닝/딥러닝 기초학습
<beyondAI BASIC> 코칭스터디 (네이버커넥트재단)	<ul style="list-style-type: none">python 활용 EDA, 데이터 전처리 학습머신러닝(트리모델), 하이퍼파라미터 튜닝, 앙상블 기법 학습딥러닝(pytorch) 모델 (CNN,RNN,GENERATIVE MODEL) 학습
처음하는 SQL과 데이터베이스 (MYSQL)부트캠프(잔재미코딩)	<ul style="list-style-type: none">SQL 기초문법SQL 데이터 분석