유료 회원 만료 예측 보고서

빅데이터 8기

이름 : 김솔미

# 1. 프로젝트 명칭 및 기간

## 가. 프로젝트 명칭

##### 유료 회원 만료 예측 프로젝트

## 나. 프로젝트의 기간

##### 2024.07.23~2024.07.25

# 2. 프로젝트의 목적 및 요약

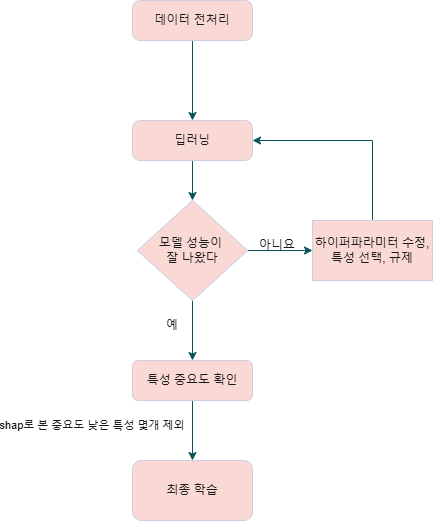
## 가. 요약

유료 회원이 언제 만료될지를 예측하는 모델을 만들기 위해 Sequential 모델을 사용하여 딥러닝을 수행했습니다. 모델의 학습 효율을 높이기 위해 특성 중요도 확인, 하이퍼파라미터 조정 등을 실시했습니다.

## 나. 목적

##### T사 유료 회원이 어느 시점에서 만료로 넘어가는지 알기 위해 만료된 회원들의 고유 아이디, 콘텐츠 아이디, 만료된 시기 등이 포함된 데이터 셋을 가지고 딥러닝을 하여 학습시키기. 언제쯤 만료가 될지 예측하는 모델을 만들고 각 회원의 만료를 막기 위한 여러 방안을 생각해 유료 회원 유지를 위함.

## <그림 1. 워크플로우>



# 3. 데이터

## 가. 데이터

##### T사 만료 회원 데이터

##### <표 1. 사용한 데이터셋 설명>

|  |  |  |
| --- | --- | --- |
| **컬럼 이름** | **컬럼 설명** | **데이터 예시** |
| **userid** | 회원고유ID | 00109005-49ee-489e-8143-1327f1ff8054 |
| **learning\_seq** | 학습순서 | 27 |
| **mcode** | 콘텐츠고유ID | T0BE00U17012 |
| **learning\_action\_seq** | 학습 중 행동순서 | 1 |
| **event\_type** | 이벤트 유형 | ViewEvent |
| **action** | 이벤트에 포함되는 행동 내용 | Viewed |
| **object\_type** | 활동 대상 | DigitalResource |
| **timestamp** | 해당 학습 중 행동이 발생한 시간 | 2022-07-02 09:58:46 |
| **gender** | 성별 (M - 남, F - 여, X - 기타, 정보없음) | M |
| **grade** | 학년 | 5 |
| **memberstatus** | 회원상태 (정회원, 준회원, 탈퇴회원 등) | 11 |
| **memberstatus\_change** | 월 중 회원상태 변화 | 11,-,11 |
| **day\_숫자\_status** | 해당일 회원상태 | 11 |
| **change\_date** | 만료된 날 | 28 |

## 나. 전처리

memberstatus\_change, day\_숫자\_status 컬럼에 있는 기호를 0으로 대체하거나 비우기 처리했습니다. 데이터 정규화를 위해 StandardScaler를 사용하고 훈련 셋과 테스트 셋으로 분할했습니다. 'gender', 'mcode', 'action', 'event\_type', 'object\_type' 컬럼을 원핫 인코딩 처리하여 컬럼 수를 225개로 늘렸습니다.

# 4. 딥러닝

## 가. 초기 딥러닝

change\_date를 제외한 모든 특성을 X로, y는 change\_date로 설정했습니다. LabelEncoder를 통해 문자열 데이터를 수치형으로 변환한 후, 텐서플로우를 사용하여 신경망 모델을 정의했습니다. Sequential 모델을 생성하고 Dense 레이어와 Dropout을 통해 과적합을 방지했습니다. Adam 옵티마이저를 사용하여 최적화하였으며, 손실 함수로는 mean squared error를 사용하고, 평가 지표로 mean absolute error를 사용했습니다. 에포크 수는 100, 150, 200을 시도한 후 150으로 선택했습니다.

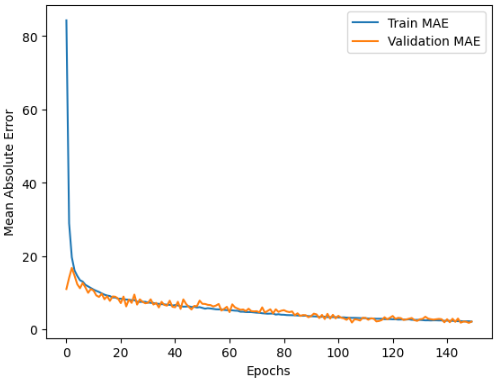
<표 2. model 설명>

|  |  |  |
| --- | --- | --- |
| Layer(type) | Output Shape | Param |
| dense (Dense) | (None, 128) | 5760 |
| dropout (Dropout) | (None, 128) | 0 |
| dense\_1 (Dense) | (None, 64) | 8256 |
| dropout\_1 (Dropout) | (None, 64) | 0 |
| dense\_2 (Dense) | (None, 32) | 2080 |
| dropout\_2 (Dropout) | (None, 32) | 0 |
| dense\_3 (Dense) | (None, 1) | 33 |

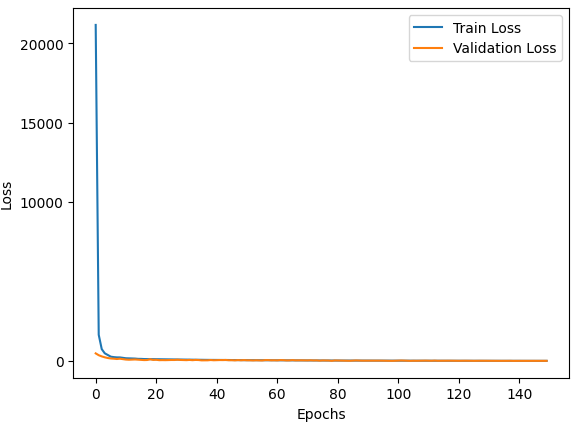
##### <표 3. 1차 딥러닝 MAE, R^2 스코어>

|  |  |
| --- | --- |
| MAE | 2.27 |
| R^2 스코어 | 0.57 |

<그림 2. MAE 시각화 그래프>



<그림 3. loss 시각화 그래프>



## 나. 2차 딥러닝

X에서 event\_type, day\_숫자\_status, userid, timestamp, grade를 제외했습니다. 문자열 데이터 간 상관이 없기 때문에 라벨링 인코딩에서 원핫 인코딩으로 변환하였고, 과적합 방지를 위해 L1, L2 규제를 추가했습니다. 에포크 수는 150으로 설정했습니다.

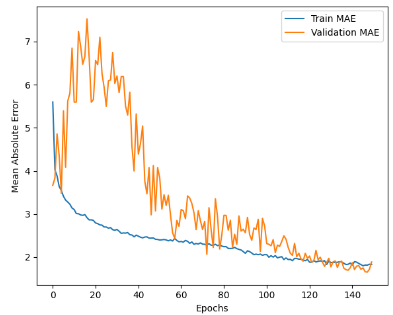
<표 4. model 설명>

|  |  |  |
| --- | --- | --- |
| Layer(type) | Output Shape | Param |
| dense (Dense) | (None, 128) | 28928 |
| dropout (Dropout) | (None, 128) | 0 |
| dense\_1 (Dense) | (None, 64) | 8256 |
| dropout\_1 (Dropout) | (None, 64) | 0 |
| dense\_2 (Dense) | (None, 32) | 2080 |
| dense\_3 (Dense) | (None, 1) | 33 |

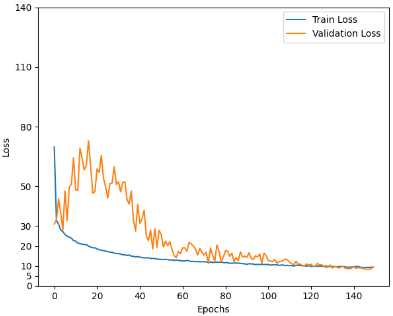
<표 5. 2차 딥러닝 MAE, R^2 스코어>

|  |  |
| --- | --- |
| MAE | 1.76 |
| R^2 스코어 | 0.68 |

<그림 4. MAE 시각화 그래프>

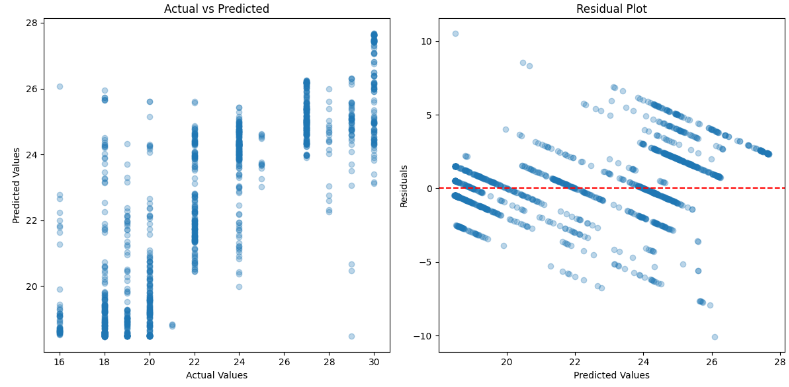


<그림 5. loss 시각화 그래프>



## 다. 2차 딥러닝 실제 값과 예측값 비교, 잔차 플롯

<그림 6. 값 비교, 잔차 시각화 그래프>



<표 6. 2차 딥러닝 실제 값, 예측 값 비교>

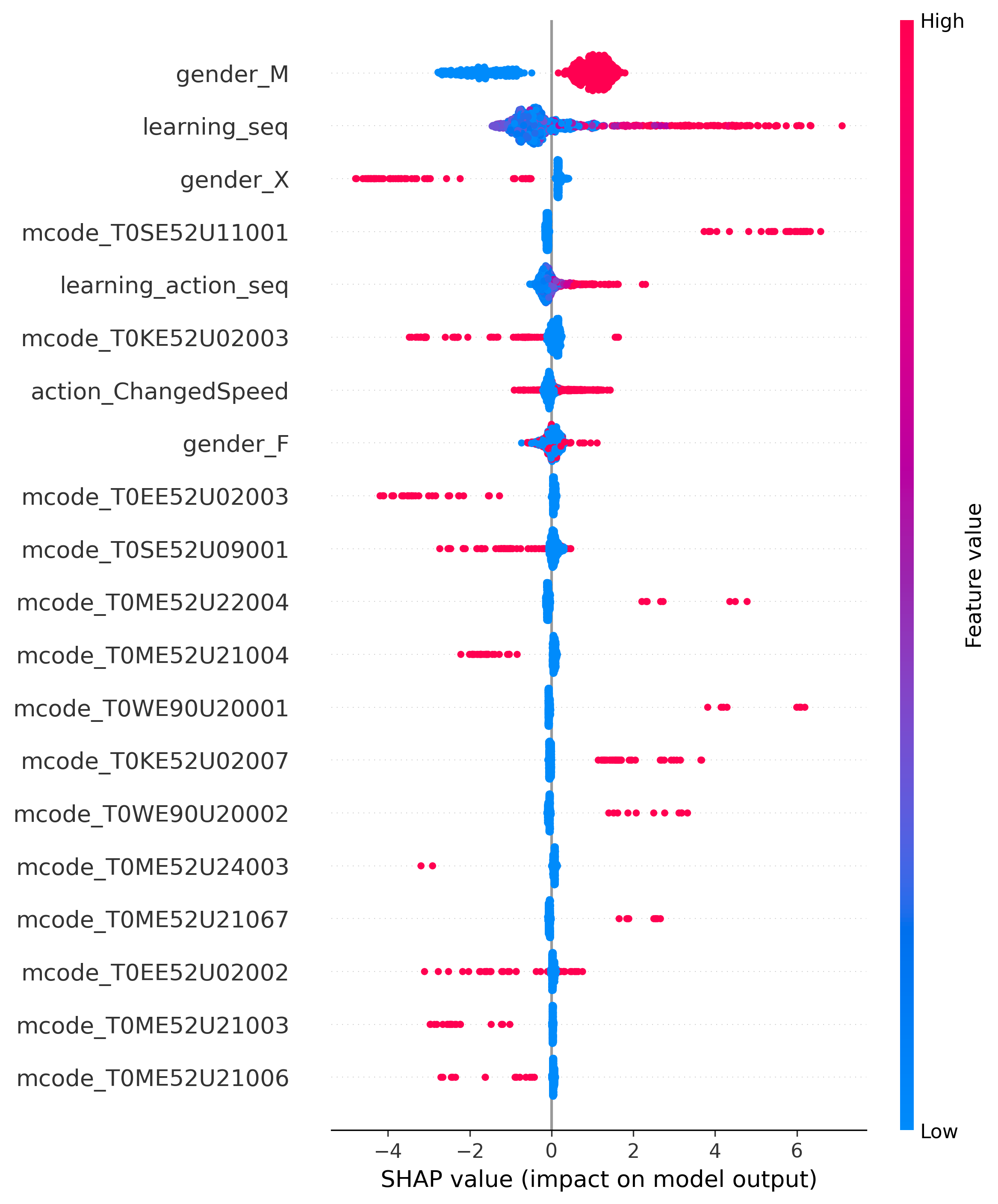
|  |  |  |
| --- | --- | --- |
| 랜덤 인덱스 | 실제 값 | 예측 값 |
| 1 | 27 | 26 |
| 2 | 19 | 20 |
| 3 | 27 | 25 |
| 4 | 20 | 20 |
| 5 | 18 | 19 |
| 6 | 22 | 21 |
| 7 | 20 | 20 |
| 8 | 22 | 21 |
| 9 | 27 | 25 |
| 10 | 24 | 24 |

# 5. 특성 중요도

## 가. 특성 중요도 확인

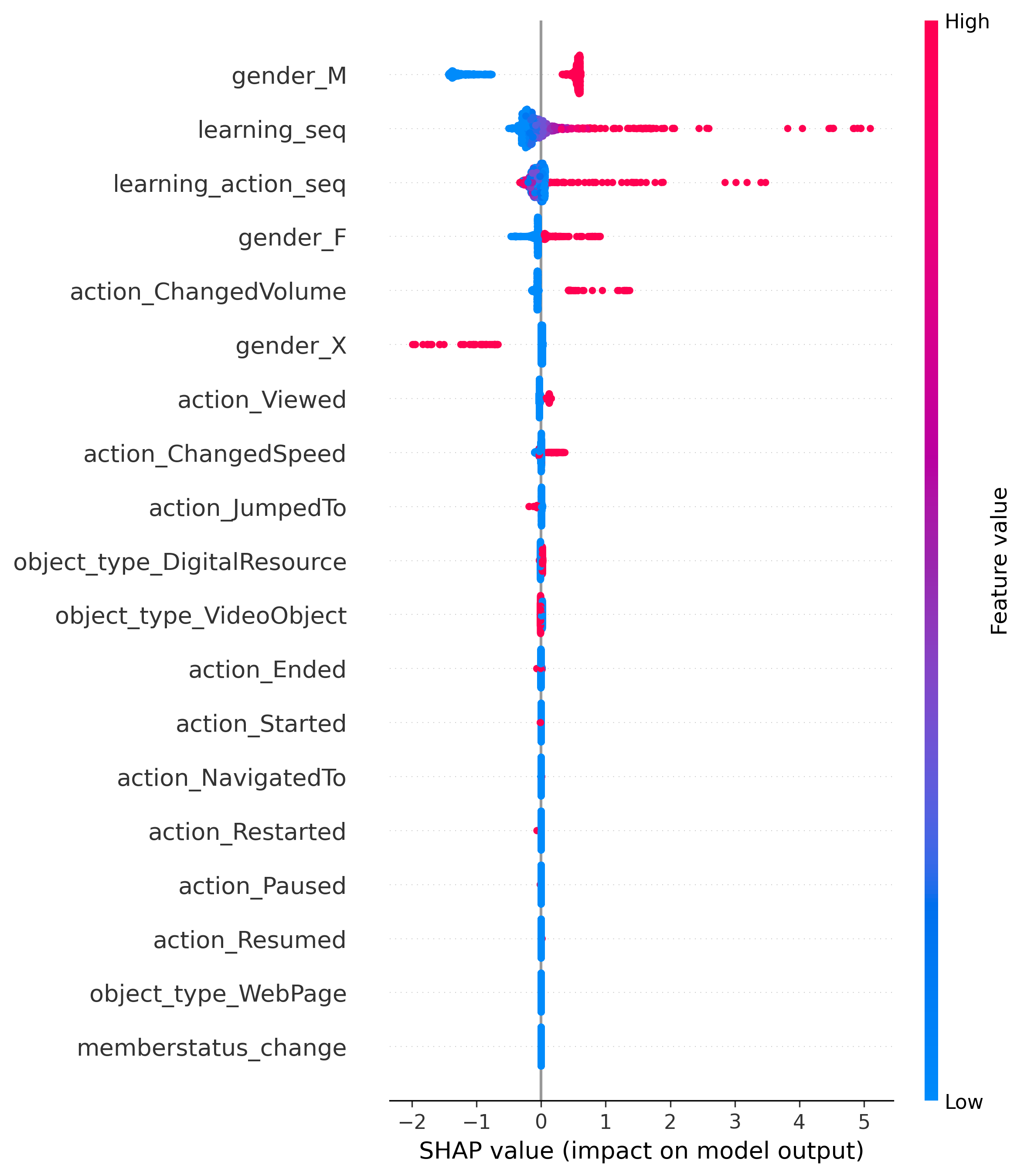
20개의 특성을 확인할 수 있는 shap 라이브러리를 사용했습니다. mcode를 포함한 상위 20개 shap, mcode를 제외한 전체 특성 보이는 shap로 2번 진행하였습니다. 원핫인코딩한 mcode의 컬럼이 너무 많아 나눠서 진행하였습니다. gender\_M, learning\_seq, learning\_action\_seq, action\_ChangedVolume 등의 특성이 높게 나왔습니다. 같은 gender, action의 다른 컬럼도 포함하여 딥러닝을 진행했습니다.

<그림 7. shap plot(20개)>



mcode를 제외하고 특성중요도를 확인. memberstatus\_change 특성을 제외하기로 결정.

<그림 8. shap plot(20개)>



# 6. 최종 딥러닝

## 가. 딥러닝

특성중요도에서 0으로 나오는 memberstatus\_change 제외. X에 'gender', 'learning\_seq', 'action', 'mcode', 'learning\_action\_seq', object\_type' 입력

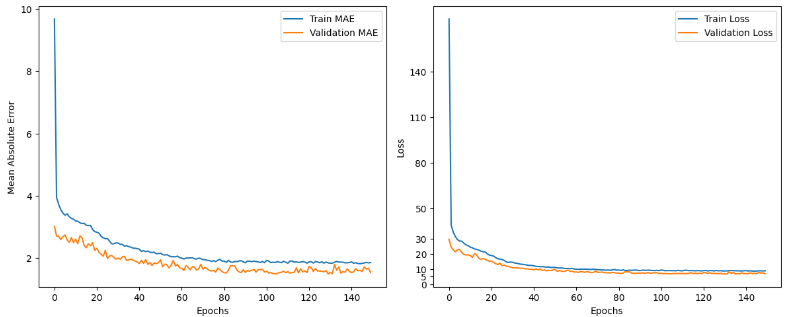
<표 7. model 설명>

|  |  |  |
| --- | --- | --- |
| Layer(type) | Output Shape | Param |
| dense (Dense) | (None, 128) | 28672 |
| dropout (Dropout) | (None, 128) | 0 |
| dense\_1 (Dense) | (None, 64) | 8256 |
| dropout\_1 (Dropout) | (None, 64) | 0 |
| dense\_2 (Dense) | (None, 32) | 2080 |
| dense\_3 (Dense) | (None, 1) | 33 |

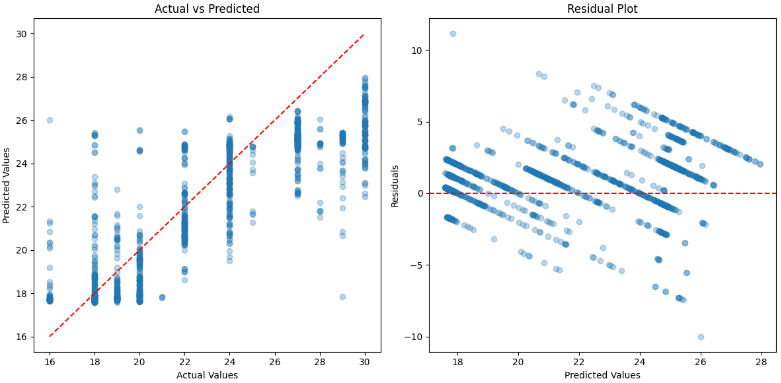
<표 8. 최종 딥러닝 MAE, R^2 스코어>

|  |  |
| --- | --- |
| MAE | 1.53 |
| R^2 스코어 | 0.75 |

<그림 9. MAE, loss 시각화 그래프>

****

<그림 10. 값 비교, 잔차 시각화 그래프>



# 7. 결론

본 보고서에서는 서비스 이용자가 언제 만료될지를 예측하기 위해 딥러닝 모델을 개발하고 성능을 평가했습니다. 초기 모델에서는 모든 특성을 사용하여 학습을 진행하였고, 2차에서는 몇몇 특성을 제외하고 규제를 적용하였습니다. 이후 shap 분석을 통해 각 특성이 어떻게 기여하는지를 확인하고 3차 딥러닝을 진행했습니다. 이 과정에서 MAE 값이 2.27에서 1.76, 최종적으로 1.53으로 감소하였고, R² 값이 0.57에서 0.68, 최종적으로 0.75로 증가하여 모델의 성능이 향상된 것을 확인할 수 있었습니다.

모든 특성을 사용했을 때보다 몇 가지 특성을 제외한 후 모델 성능이 더욱 향상된 것을 확인했습니다. 특히, 중요도가 낮았던 특성들을 제외함으로써 모델의 예측 성능이 개선된 것을 시각적으로 확인할 수 있었습니다.

학습 과정에서 드롭아웃과 L1, L2 규제를 적용하여 과적합을 방지했습니다. 잔차 플롯을 통해 모델의 예측 값이 전반적으로 실제 값과 일치하는 경향이 더욱 강화된 것을 확인할 수 있었습니다.

학습 과정에서 손실 함수 값이 안정적으로 감소하고, 예측 성능이 개선된 결과를 바탕으로, 이러한 딥러닝 모델이 실제 서비스에서도 효과적으로 활용될 수 있습니다. 서비스 이용자의 행동 예측에서 딥러닝 모델을 활용함으로써 예측 정확성을 높일 수 있습니다.

# 8. 향후 과제

더 많은 데이터를 추가로 수집하고, 새로운 특성을 포함시켜 모델의 예측 성능을 더욱 향상시킬 수 있습니다. 또한 앙상블 기법 등을 활용해 모델 성능을 높일 수 있습니다. R^2값이 0.75로 나왔는데 하이퍼 파라미터 조정과 최적의 모델을 찾아 예측 성능이 더 높은 모델을 구축할 수 있습니다.