2024.08.22 머신러닝 실습 보고서 (최종)

Created	@2024년 8월 22일 오후 4:25

이름	곽태경
소속	빅데이터 9기

요약

8월 22일 실습에서는

만료 및 탈퇴한 회원들의 정보가 담긴 데이터에서 데이터를 선별하고, 로지스틱 회귀 모델을 이용해 탈퇴 상태가 아니었던 회원이 탈퇴할 확률을 예측하는 머신러닝 모델을 구현했습니다. 그리고 모델들의 예측 결과와 성능을 시각화 하였습니다.

도입

문제 정의

회원 관리 시스템에서 만료, 탈퇴 회원을 관리하는 것은 중요한 부분 중 하나인데, 만약 탈퇴 상태가 아니었던 회원들이 조기에 탈퇴 확률을 예측하고 대응할 수 있다면 탈퇴 고객의 수를 줄일 수 있을 것입니다.

기획 의도

- 고객들의 탈퇴할 확률을 예측하는 것을 목적으로 예측 모델을 만들기 위해서입니다.

개발 환경

- Vscode 1.92.2
- anaconda 1.12.3

- Python 3.12.4
- Numpy 1.26.4
- Pandas 2.2.2
- matplotlib 3.8.4
- Scikit_learn 1.4.2
- Seaborn 0.13.2
- jupyter notebook v2024.7.0

데이터 정보

파일 이름 : 만료및탈퇴회원.csv

데이터 정보

분류	값(개수)
sample	111851
features	55

features 목록

컬럼명	설명	결측치
userid	인덱스	0
gender	회원아이디	0
membertype_codename	회원구분 코드의 한글명(학교급) ex. 초등	0
grade_codename	학년 코드	0
memberstatus	회원상태 (정회원, 준회원, 탈퇴회원 등)	0
memberstatus_codename	회원상태 코드의 한글명	0
memberstatus_change	월 중 회원상태 변화	0
status_null_count	회원상태 없음 일수	0
statusgroup_10_count	임시회원 일수	0
statusgroup_20_count	무료회원 일수	0
statusgroup_30_count	유료회원 일수	0

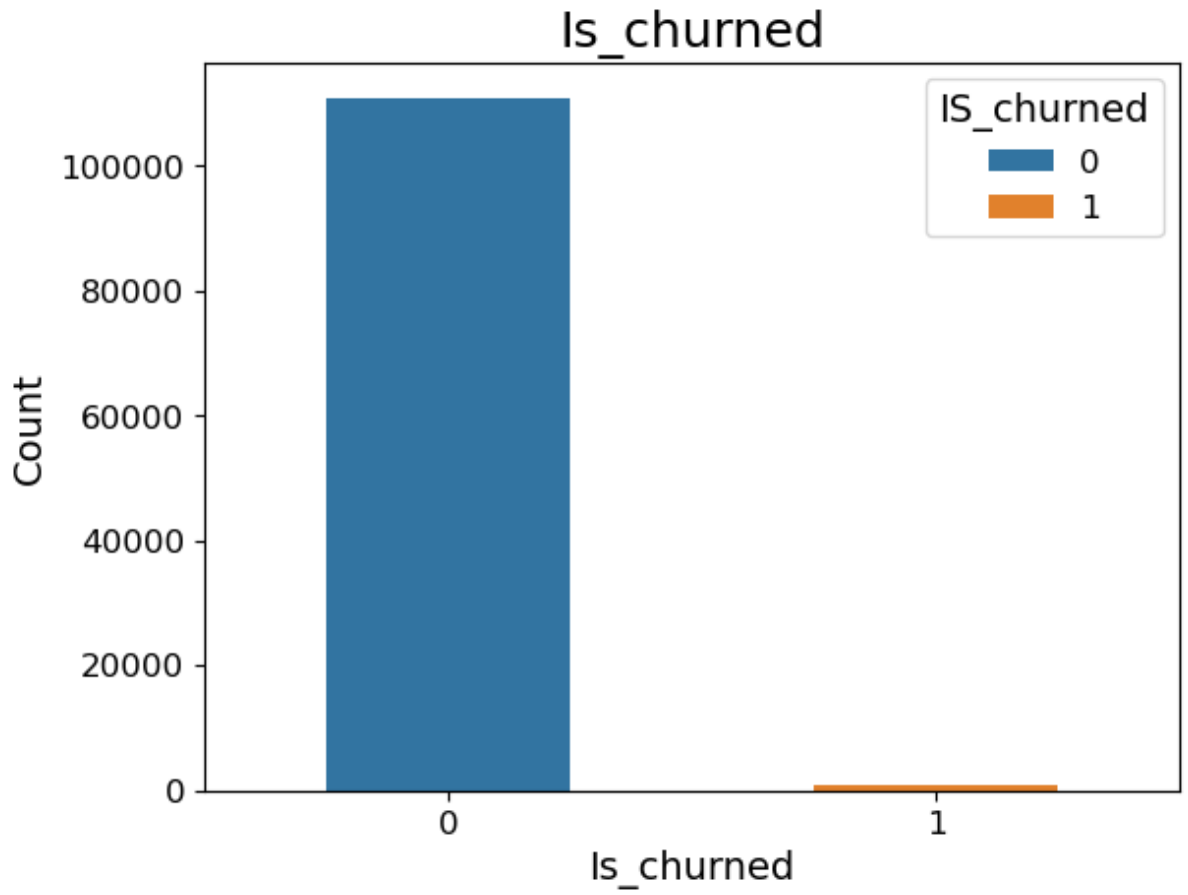
statusgroup_40_count	중지회원 일수	0
statusgroup_50_count	만료회원 일수	0
statusgroup_90_count	해지회원 일수	0
point_gain_activeday_count	포인트 획득일수	3955
point_gain_count	포인트 획득 횟수	3955
point_gain	획득 포인트	3955
point_loss_activeday_count	포인트 차감일수	3955
point_loss_count	포인트 차감 횟수	3955
point_loss	차감 포인트	3955
tablet_activeday_count	기기 활성 횟수	69893
tablet_moved_menu_count	기기 메뉴이동 횟수	69893
tablet_leave_count	기기 물리적 종료 횟수	69893
tablet_resume_count	기기 물리적 재개 횟수	69893
tablet_login_count	로그인 횟수	69893
tablet_logout_count	기기 로그아웃 횟수	69893
study_activeday_count	학습 활성일 횟수	77821
study_count	학습 횟수	77821
study_notcompleted_count	학습 미완료 횟수	77821
study_completed_count	학습 완료 횟수	77821
study_restart_count	학습 재시작 횟수	78457
total_system_learning_time	학습 시간합계(시스템)	78584
total_caliper_learning_time	학습 시간합계(캘리퍼)	78496
media_activeday_count	미디어 활동 활성 일수	79605
media_count	미디어 학습 횟수	79605
video_action_count	비디오 활동 횟수	80735
video_start_count	비디오 시작 횟수	80735
video_restart_count	비디오 재시작 횟수	80735
video_pause_count	비디오 일시정지 횟수	80735
video_jump_count	비디오 점프 횟수	80735
video_resume_count	비디오 재개(일시정지 후 횟수)	80735
video_speed_count	비디오 속도 조절 횟수	80735
video_volume_count	비디오 볼륨 조절 횟수	80735

video_end_count	비디오 종료 횟수	80735
test_activeday_count	평가 활성 일수	82852
test_count	평가 횟수	82852
test_average_score	평가 평균 점수	82852
test_item_count	평가 문항 개수	83472
test_correct_count	평가 정답 개수	83472
wrong_count	오답 노트 진입 횟수	104408
wrong_item_count	오답 노트 문항 개수	104408
wrong_correct_count	오답 노트 정답 개수	104408
yyyy	년	0
mm	월	0

개발 과정

1. EDA

먼저 전체 회원에서 탈퇴 회원의 비율을 알아보기 위해 데이터의 column에 탈퇴한 회원들은 1, 그렇지 않으면 0의 값을 가지는 '**is_churned**'라는 column을 만들었습니다.



그리고 'is_churned' 열을 target으로 두고, feature는

- 'statusgroup_20_count' : 무료 회원 일수
- 'stautsgroup_30_count' : 유료 회원 일수
- 'statusgroup_40_count' : 중지 회원 일수
- 'statusgroup_50_count' : 만료 회원 일수

으로 선정해서 한 달간 회원의 상태 변화에 따른 탈퇴 여부를 예측하는 머신러닝 모델을 구현하기로 결정했습니다.

('statusgroup_10_count'는 전체 회원중 아무도 임시회원이었던 기간이 없으므로 포함하지 않았습니다.)

2. 모델 학습

학습 데이터와 테스트 데이터를 train_test_split 함수를 이용하여 분리했고, test_size=0.3 으로 두었습니다.

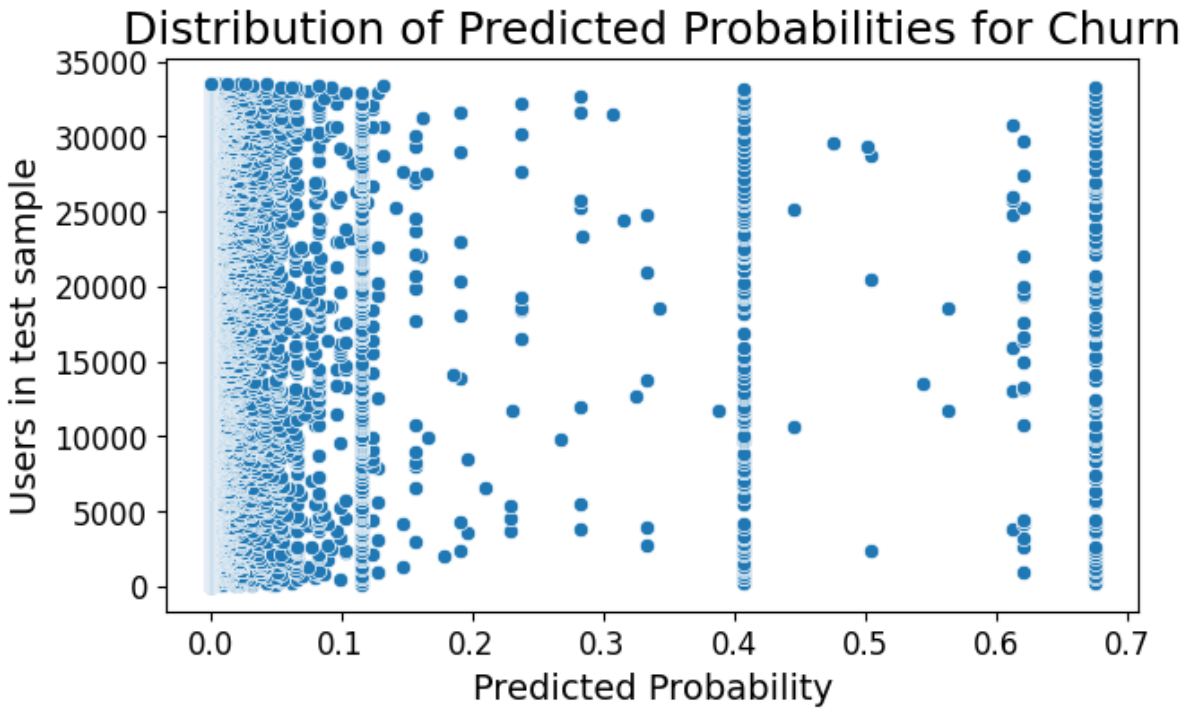
(학습 데이터 샘플 : 78295개, 테스트 데이터 샘플 33556개 입니다.)

로지스틱 회귀는 선형 회귀와 달리, 입력 값에 Sigmoid 함수를 적용하여 결과를 0과 1 사이의 값으로 제한하기 때문에 이진 분류에 사용하기에 적합하다고 판단했습니다. 그러므로 이번 실습에 적합하다고 판단해 로지스틱 회귀를 사용하기로 했습니다.

그래서 로지스틱 회귀 모델에 학습 데이터 샘플을 학습시켰고, 회원들의 상태 변화에 따른 탈퇴 여부를 예측하는 모델을 구현했습니다.

예측값
0.0005154
0.00016446
0.0005154
...
0.00107815
0.00039083
0.0005154

그리고 테스트 데이터들의 탈퇴 확률 예측값들의 양상을 확인하기 위해서 산점도로 시각화 해 보았습니다.



그래프를 확인해 보니, 확률이 0.1%를 넘는 경우가 테스트 데이터 샘플 데이터 33556명 중 616명뿐이고, 똑같은 예측값을 가지는 군집들이 그래프 상으로 확인되어 모델 성능을 시험해보기 위해 모델을 검증하는 과정을 거쳤습니다.

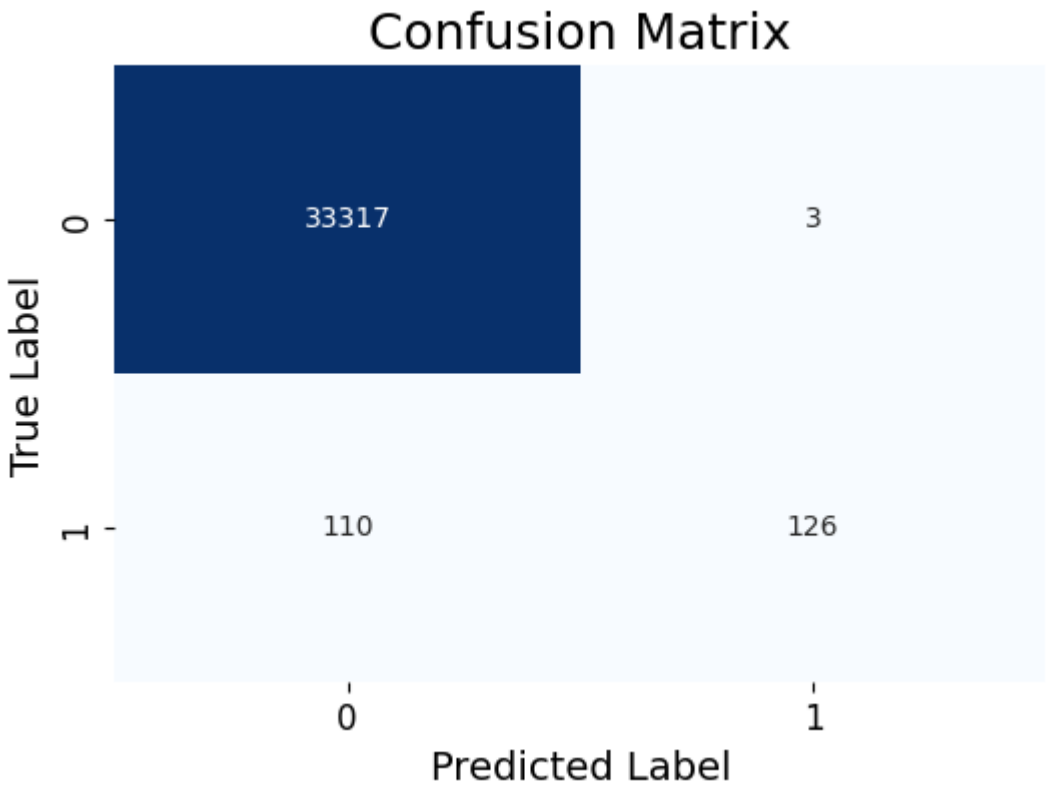
3. 모델 검증

학습 데이터 정확도와 테스트 데이터 정확도를 계산했습니다.

학습 데이터 정확도 : 0.9955424995210422

테스트 데이터 정확도 : 0.996632494933842

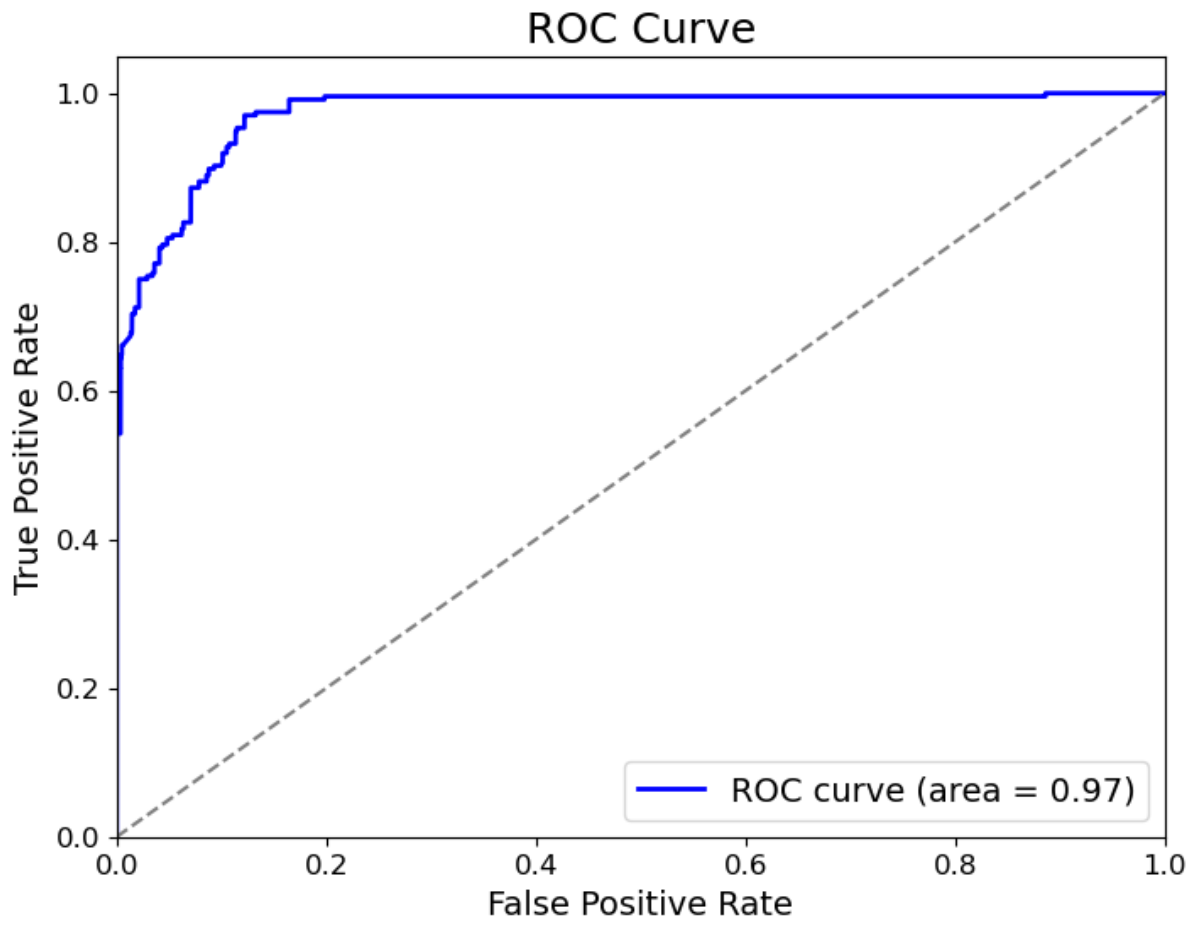
혼동 행렬을 metrics.confusion_matrix 메서드로 계산했고, heatmap으로 시각화했습니다.

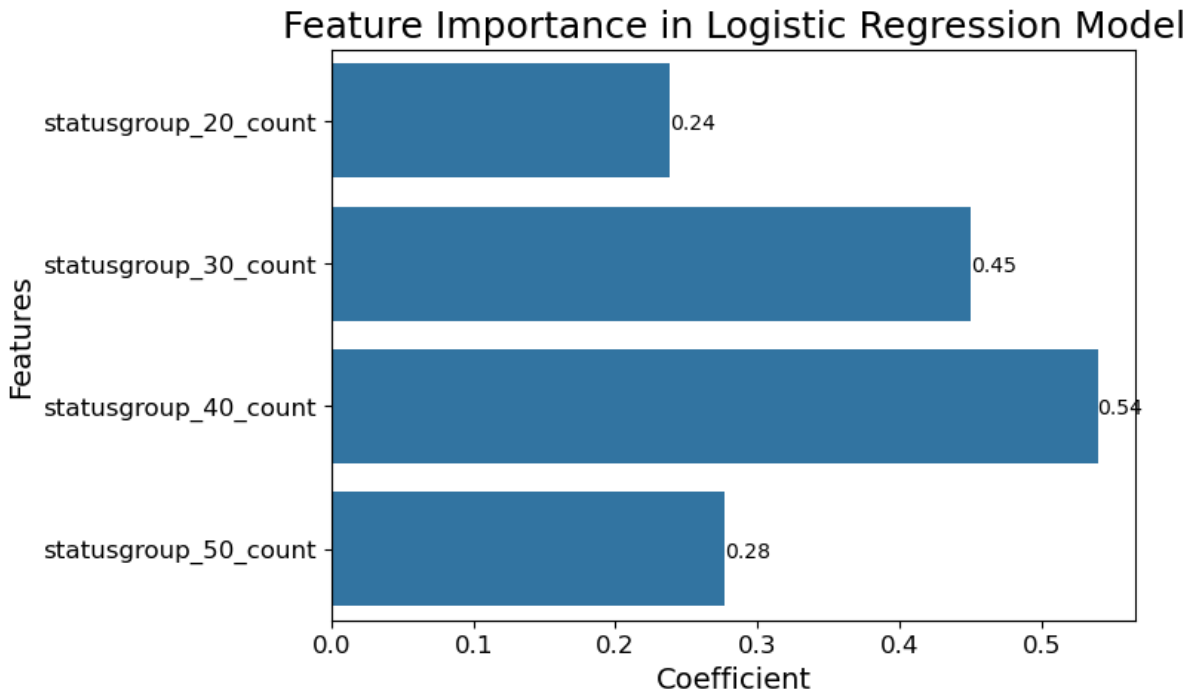


학습 데이터와 테스트 데이터 정확도를 보니, 학습 데이터는 99.55%, 테스트 데이터는 99.66%를 기록하여 정확도가 아주 높은 것으로 보이고, 혼동행렬도 정답 비율 : 오답 비율 = 99.3% : 0.7% 비율으로 정답 비율이 아주 우세해서 좋은 모델이 나왔다고 판단했습니다.

ROC 커브와 AUC값, features importance를 시각화하였습니다.

(features importance는 coef_를 사용하여 계산했고, feature들 각각의 영향력만을 비교하기 위해 회귀 계수의 절댓값을 시각화했습니다.)





ROC 커브를 보면 좌상단으로 볼록하게 휘어진 그래프 양상을 확인할 수 있으며, 이를 통해 모델이 효과적으로 학습된 것을 알 수 있었습니다. AUC(Area Under the Curve) 값이 1에 아주 가까운 0.97이라는 것도 모델의 학습이 아주 효과적이었다는 것을 나타냅니다.

그리고 feature importance 그래프를 보면 유료 회원과 중지 회원이었던 기간이 무료 회원이나 만료 회원이었던 기간에 비해 탈퇴할 확률에 2배 가량 영향을 더 많이 주는 것으로 확인되었습니다.

결론

이번 실습에서는 한 달 동안 회원 상태 변화에 따른 탈퇴 여부를 예측해 보았습니다. EDA 단계에서 데이터의 특성을 파악하고, 회원 상태와 탈퇴 간의 관계를 잘 설명할 수 있는 feature들을 선정했습니다. '무료 회원 일수', '유료 회원 일수', '중지 회원 일수', '만료 회원 일수'를 주요 feature으로 선정하여 이들이 탈퇴 여부에 미치는 영향을 모델로 구현했습니다.

이번 실습에서 로지스틱 회귀 모델을 사용한 이유는 입력 값에 시그모이드 함수를 적용해 결과를 0과 1 사이의 값으로 제한할 수 있다는 점에서 이진 분류에 용이하기 때문입니다. 예측 결과를 시각화하여 본 결과 같은 예측값들을 가지는 3개 이상의 군집이 확인됐으나, 결과적으로 모델은 학습 데이터와 테스트 데이터에서 각각 99.55%와 99.66%의 높은 정확도를 기록했으며, 혼동 행렬과 ROC 커브 분석을 통해 모델이 효과적으로 학습되었음을 확인했습니다.

그리고 feature importance 분석을 통해 '유료 회원 일수'와 '중지 회원 일수'가 탈퇴 여부에 두 배 정도 더 큰 영향을 미친다는 사실을 발견했습니다. 그래프에 따르면, 비즈니스 측면에서 중지 회원 상태와 유료 회원 상태에 있는 회원들을 대상으로 한 구독 이벤트 등의 대응이 더욱 중요하다는 결론을 도출할 수 있었습니다.

개선점

데이터가 담은 회원들은 결론적으로 만료되거나 탈퇴한 회원들이고, 탈퇴하지 않은 회원들이 109892명, 탈퇴한 회원이 892명으로 10만명 이상 차이나기 때문에, 데이터의 편향이 아주 심했습니다. 그래서 결론에서 언급한 비즈니스 측면에서의 효과는 현실적으로 실행시키기는 어렵습니다. 그래서 언더 샘플링(비탈퇴 회원을 줄임)이나 오버 샘플링(탈퇴 회원을 늘림) 등의 기법으로 샘플 데이터의 불균형을 줄이거나, 서비스 구독을 지속하고 있는 회원들을 데이터 샘플에 포함하는 등의 추가 데이터 수집을 통해 더 현실적인 모델을 만들어 낼 수 있을 것입니다.

소스 코드

과제 제출 폴더에 데이터셋 파일과 같이 첨부하겠습니다.

파일 이름 : member_churned_predict(수정).ipynb (데이터셋 파일 경로는 ./data/만료및탈퇴회원.csv)

참고 문헌

수업에서 배운 내용을 토대로 실습했습니다.