ILLIO 팬심M 셀럽 이탈예측 시계열모형 설계

코드스테이츠 AIB 16
김정겸 심지은 이유빈 허선우 허소영

CONTENTS

01 프로젝트 개요
프로젝트 목표 및 의의

프로젝트 목표

팬심M

셀럽의 이탈 감지 시계열 모형 설계

- 이탈율 및 이탈 패턴 분석
- 이탈율 * 스티커확률 기반 유저 세그멘테이션

프로젝트 의의

이탈 감지 시계열모형을 구축하여 셀럽의 활동 로그를 통해 이탈자를 사전에 파악함으로써 선제적으로 대응하고자 함

더 나아가 이탈확률 및 스티커 받을 확률에 대한 시계열 모형을 통해 각 모형에 영향을 미치는 요소를 파악하고, 각 분류(binary)에 속할 확률에 결합하여 관리우선순위를 부여하고 각 세그먼트 별 통계량을 검정함

01 프로젝트 개요
프로젝트 프로세스

EDA

적정 이탈기간 선정
- 셀럽의 통상적인 이용패턴 분석
- 기준기간에 따른 이탈확률 분석

이탈유저 특성 분석
- 활동로그 Sankey Diagram
- 이탈율과 활동로그와의 관계
- 이탈율과 스티커여부와의 관계

모델 설계 및 선정

모델 설계
- 이탈여부예측모델
- 스티커후원여부예측모델

모델 선정 배경

모델의 특징 및 구조
- Attention based GRU-biLSTM

데이터 모델링
- 데이터 스플릿
- 데이터 스케일링

XAI(딥러닝 모델해석)
- 대리모델을 통한 SHAP

모델 결과

이탈여부예측모델 결과
- 기준모델 및 ML, 최종모델(DL)의 비교
- Accuracy, microAP, precision, recall, AUC 등

이탈여부예측모델 결과해석
- SHAP by Surrogate Model

스티커후원예측모델 결과
- 기준모델 및 ML, 최종모델(DL)의 비교
- Accuracy, microAP, precision, recall, AUC 등

스티커후원예측모델 결과해석
- SHAP by Surrogate Model

셀럽 세그멘테이션

관리우선대상군 분류
- 관리우선점수 부여 : log(이탈확률* 스티커확률)
- 점수 분포에 따른 임계치 선정 후 관리우선대상군 분류

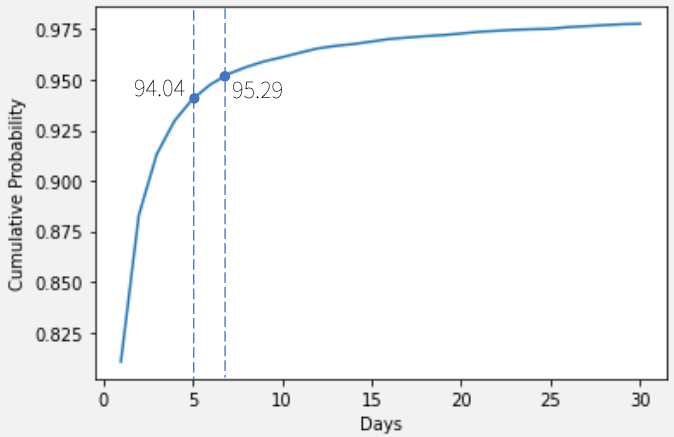
관리우선대상군의 특징 비교
- T-test를 통한 t-statistic(두 집단 간 변수별 표본평균의 차이) 비교

모델 패키징

02 EDA
적정 이탈기간 선정

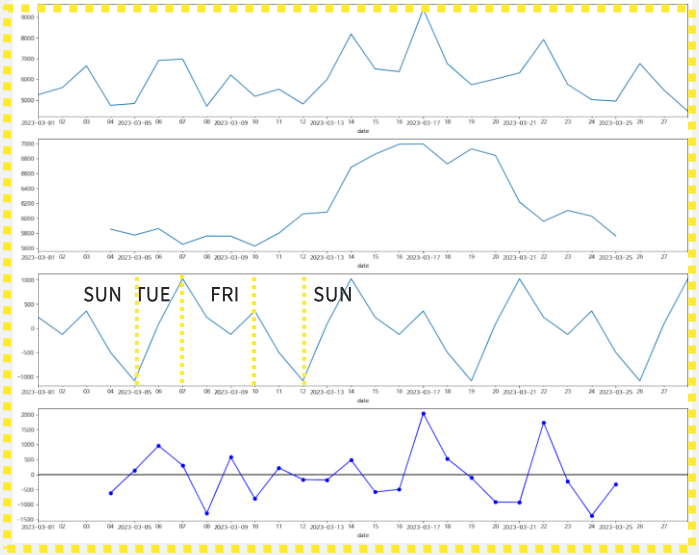
후보군(1~14일) 중 적절한 기간 선정을 위해 누적 재방문율 그래프를 살필 때, 통상적으로 셀럽들은 5~10일 이내 재방문할 확률이 높았음.
셀럽의 통상적인 이용패턴을 도식화할 때 7일 단위로 계절성 패턴(반복패턴)을 보이기에, 7일 이하로 자를 시 이탈율을 적절하게 파악하지 못할 것으로 보임.

- 오늘 이용자가 한 달 이내 재방문 확률을 판단하는 누적재방문율 그래프를 확인할 때, 셀럽들의 누적 재방문 확률이 초기에는 급격히 상승하다가 5 ~10일 사이 구간에서 한계 체증하고 7일 이내 재방문율 95%는 유의확률 5% 이내에서 유의했음. 즉, 통상적으로 이용하던 셀럽들은 5-10일 이내에 재방문할 확률이 높음.
- 셀럽의 채팅방 일일 이용현황을 분해시계열을 통해 추세trend, 계절성seasonal, 잔차residual로 분리했을 때, 일요일에 swing time, 화, 금요일에 peak time으로 반복적 패턴을 지님.
- 이는 팬케어를 위한 SNS 서비스라는 특성 상 주말보다 평일에 서비스를 통해 팬케어하고, 주말에는 방송 활동을 하는 셀럽의 활동특성과 연관이 있을 것으로 보임.
- 통상적으로 7일 단위로 반복되는 계절성을 보이는 것으로 보아, 7일 이하로 자를 때 이탈율에 대해 적절하게 파악하기 어려울 것으로 보임



오늘 이용자의 누적재방문율 그래프

- scipy.stats 라이브러리의 ttest
- 오늘 이용자가 1일 내에 재방문할 확률 81.04%
- 오늘 이용자가 5일 내에 재방문할 확률 94.04%
- 오늘 이용자가 7일 내에 재방문할 확률 95.29%



셀럽 전체의 일일 채팅메시지 로그발생건수 추이(최근4주간)

셀럽 전체의 일일 채팅메시지 로그발생건수 추이(최근20주간)

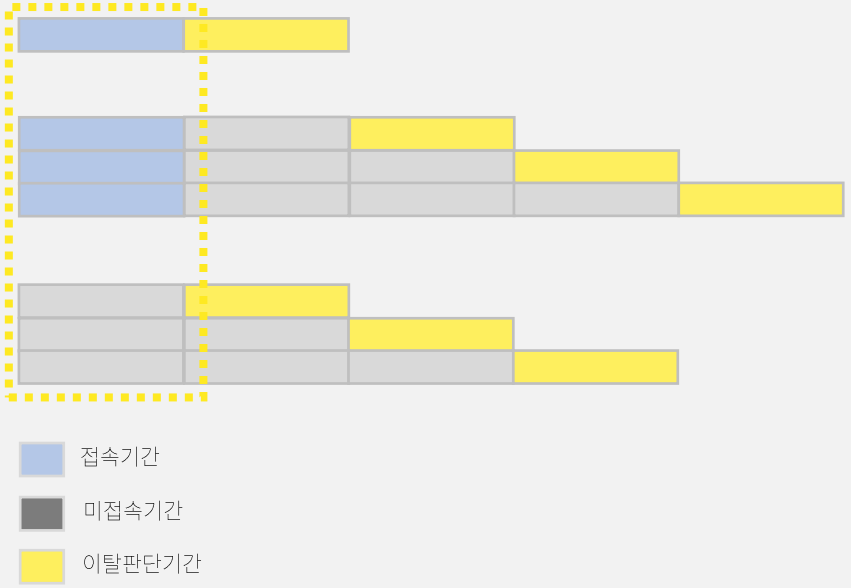
차례로 Data, Trend, Seasonal, Residual

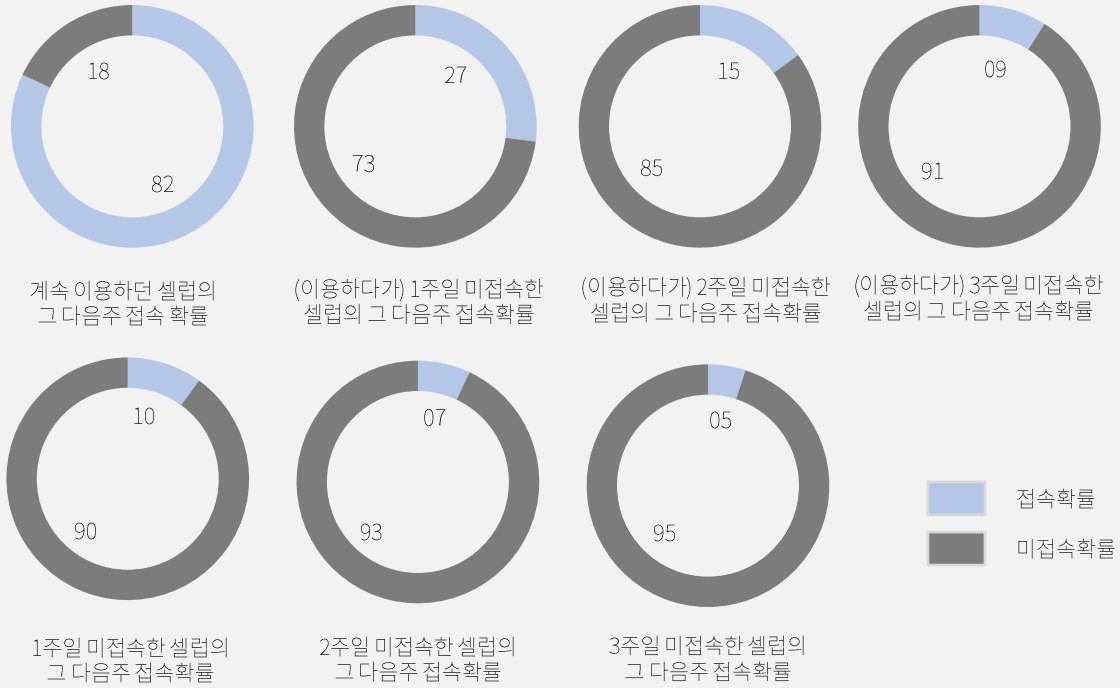
02 EDA
적정 이탈기간 선정

7일 단위로 이탈자를 파악하여 기간별 이탈 확률을 파악한 결과, 일주일 접속자는 이탈확률이 낮은 반면 일주일 미접속 시 계속 해서 이탈할 확률이 높았음. 따라서 그 다음 일주일의 이탈 확률을 예측하는 것은 선제적 대응에 도움이 될 것으로 보임.

- 잘 사용하던 사용자라도 1주일만 미접속하게 되면 다음주 미접속 확률이 73%로 상당히 높아졌고, 1주일 이상 미접속자의 경우 그 다음주 미접속 확률이 90%에 달함. 이를 통해 일주일 간의 활동로그는 그 다음주 이탈율과 유관할 것으로 판단됨.
- 그 다음 1주일, 2주일, 3주일간의 미접속 확률 현황을 살펴볼 때 한 번 안 들어오기 시작하면 계속 안 들어올 가능성이 높은 플랫폼임을 알 수 있으며, 자주 접속하는 것이 중요한 SNS 서비스인만큼 7일 간의 활동로그를 통해 그 다음 일주일의 이탈 확률을 예측할 수 있다면 선제적 대응에 도움이 될 것으로 보임.

Case 별 이탈판단기간 내 접속확률





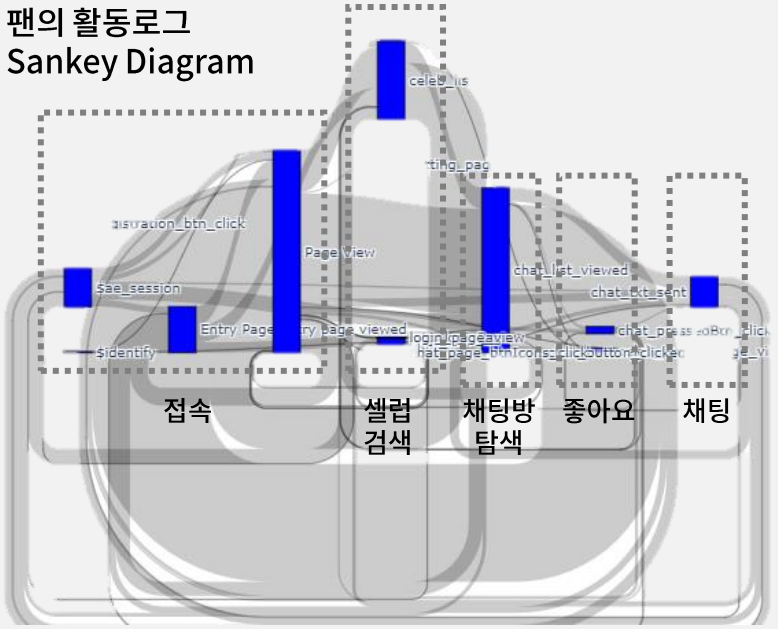
02 EDA
활동로그 Sankey Diagram

셀럽 및 팬의 서비스 이용프로세스는 총 5단계로 이루어지며, 크게 접속-셀럽검색-채팅방탐색-좋아요-채팅 등으로 구분됨.

- Sankey Diagram을 통해 10,000건 이하의 로그를 제외하고 분석한 결과, 셀럽 및 팬의 서비스 이용프로세스는 5단계로 이루어짐.
- 팬의 경우, 좋아요, 채팅 등의 소통활동보다는 셀럽 검색 및 채팅방 탐색 등의 활동에 많은 시간을 소비함

셀럽의 활동로그
Sankey Diagram

팬의 활동로그
Sankey Diagram



10,000건 이하의 로그는 제외됨

02 EDA
이탈유저 특성 분석

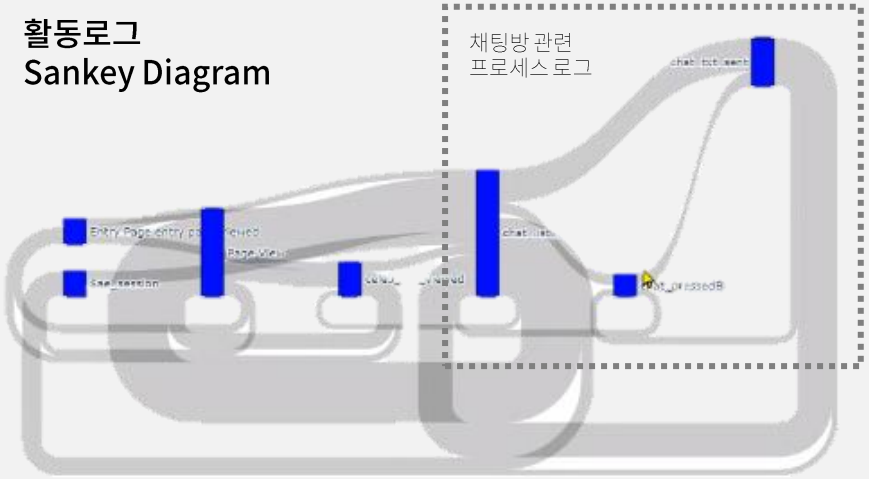
이탈자와 미이탈자의 행동패턴 양상과 땡큐스티커 수신이력여부를 살핀 결과, 두 지표는 이탈율에 영향을 미칠 것으로 추측됨.

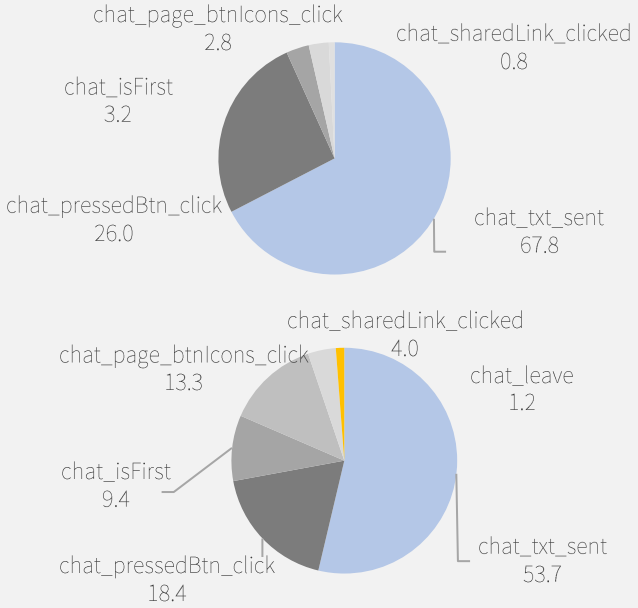
•프로세스 5단계 중 채팅방 관련 로그를 전체와 이탈자로 구분하여 살핀 결과, 이탈자는 메시지 보내기, 부가기능(좋아요 등)의 비율이 줄고 채팅방 나가기 이벤트가 발생하는 등 행동패턴이 변화하는 양상을 관찰할 수 있었음.

•또한 땡큐스티커 관점에서, 전체 셀럽 중 땡큐스티커 후원을 한 번이라도 받은 셀럽과 미후원 셀럽의 비율은 75:25 이었고, 땡큐스티커를 받은 셀럽 중 이탈자의 비율은 25%였던 반면 미후원 셀럽 중 이탈자의 비율은 65%였음.

•이를 통해 셀럽의 땡큐스티커 후원/미후원여부와 활동패턴이 이탈율에 영향을 미칠 것으로 추측할 수 있음.

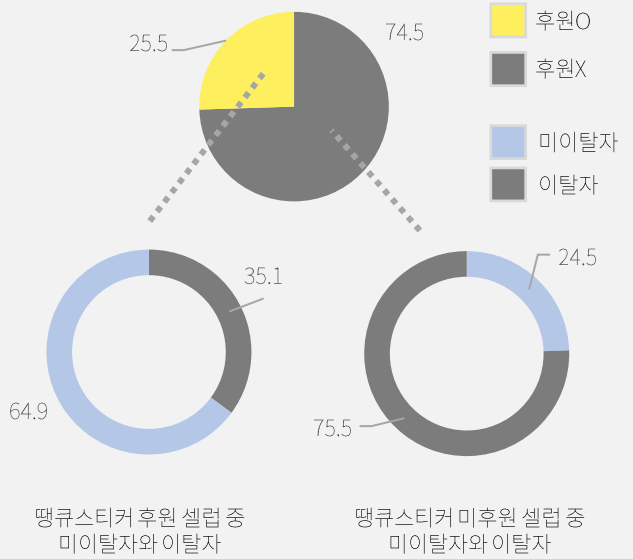
활동로그
Sankey Diagram





전체셀럽(위)와 이탈자(아래)의 활동로그 비율(마지막주 기준)

전체 셀럽 중 땡큐스티커 받은 비율

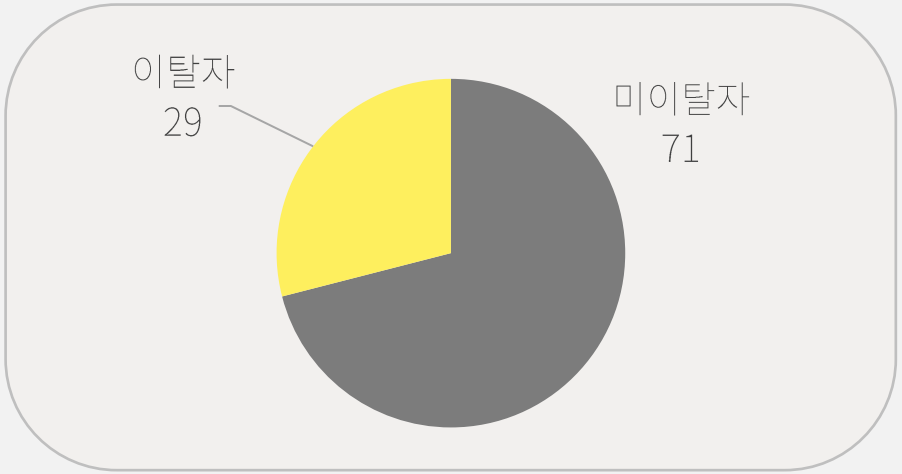


땡큐스티커 후원 셀럽 중 미이탈자와 이탈자

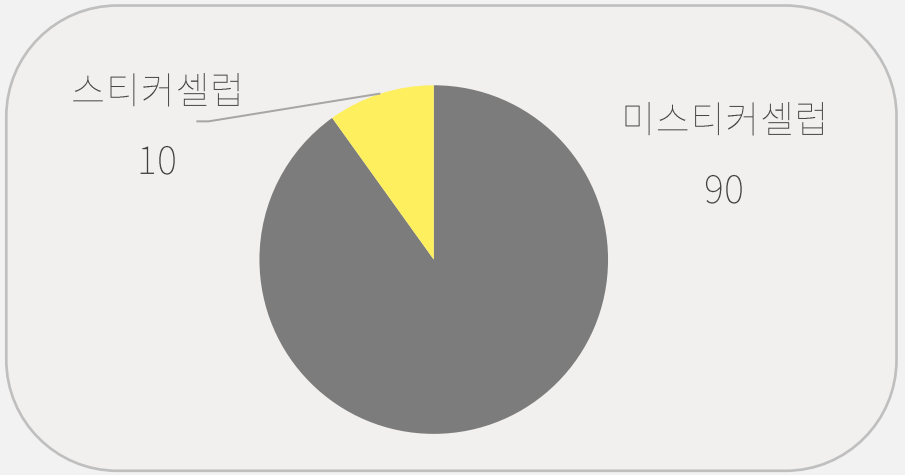
땡큐스티커 미후원 셀럽 중 미이탈자와 이탈자

03 모델 설계
모델 설계 목표

“ 1주일 간의 활동로그 및 팬/땡큐스티커 일일현황(feature)을 통해
다음 1주일 동안의 이탈여부 및 스티커여부(label)를 예측하는 시계열모델 “



x



각 모델에서 예측한 이탈확률 * 스티커받을확률에 따라 분포를 확인하여, 관리우선순위 부여 및 각 그룹별 특징 파악

03 모델 설계
시계열 데이터의 스플릿

데이터(22.11.10~23.03.28, 총 20주차)

•셀럽의 본인방에서의 각 이벤트 로그
•팬의 셀럽방에서의 각 이벤트 로그
•셀럽/팬 채팅 메시지 길이
•셀럽 별 일일 활동팬 수
•셀럽 별 일일 땡큐스티커 받은 횟수 및 총량

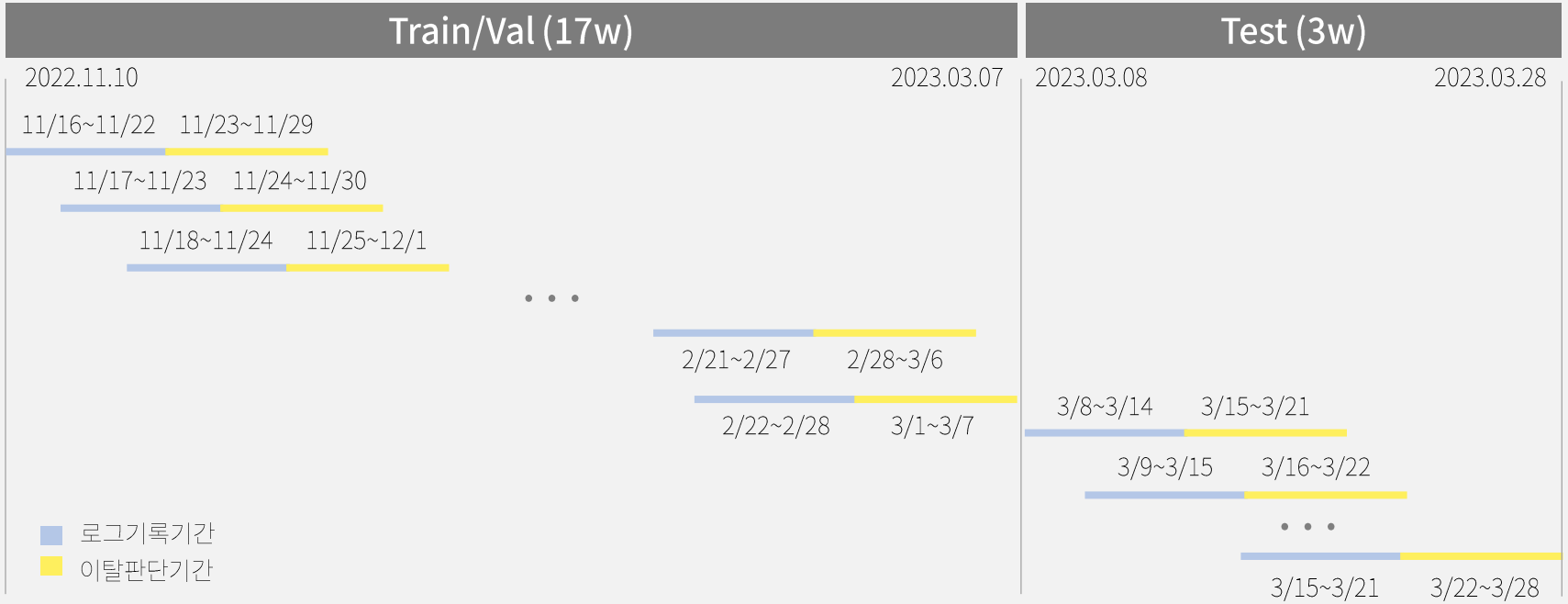
데이터 스플릿

•Train, Validation set은 최신 경향성을 반영하기 위해 20주차 중 앞 17주차(22.11.10~23.03.07)의 로그를 random-split하고, Test set은 최신 주차(18~19주차, 23.03.08~23.03.21)의 데이터로 구분하였음
•이에 따라, 각 로그의 다음 1주일 간의 접속여부(label, train 2~17주차, test 19~20주차)를 예측하는지를 기준으로 성능을 평가하였음
•최종적으로 시계열 별 분리 후 기간 내 대상자의 비율(이탈자, 스티커셀럽)은 각각 0.3, 0.1 내외였음.

데이터셋 당 클래스 비율

Churn 1 이탈자, Sticker 1 스티커후원셀럽

	Class	Train	Val	Test
Churn	1	0.36	0.36	0.33
	0	0.64	0.64	0.67
Sticker	1	0.09	0.10	0.13
	0	0.91	0.90	0.87



03 모델 설계
모델 선정 배경

1

데이터 형태 상
다변량 시계열 예측 모델에 해당함

•본 프로젝트에서 사용하는 데이터는 사용자 로그데이터로, 시계열 분석을 요함
•시간변수(단변량) 뿐만 아니라, 이벤트/팬/땡큐스티커 일일 현황 등 다양한 feature의 영향력을 함께 고려해야 하는 다변량 시계열의 요소를 가지고 있음.

방송회차	시청률
1	1.82
2	1.90
3	2.14

단변량 시계열

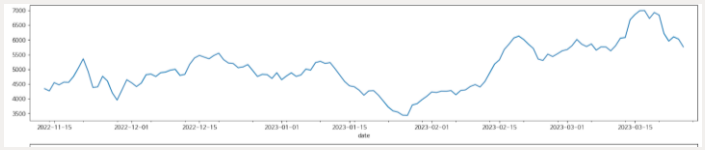
방송회차	미세먼지	시청률
1	5	1.82
2	2	1.90
3	7	2.14

다변량 시계열

2

시계열의 추세trend를 살펴볼 때
RNN 계열의 알고리즘이 적합하다고 판단됨

•시장성이 계속 성장해가는 단계에 있어 변수의 시계열적(상승)트렌드가 존재함
•이에 따라, 정상시계열이 아닐 것으로 예상되며 ARIMA 등의 시계열 모형의 예측력이 높지 않을 것으로 판단됨
•따라서 이를 분석할 수 있는 알고리즘으로는 GRU, LSTM 등 RNN 기반 모델이 적합하다고 보여짐



셀럽 전체의 일일 채팅메시지 추세 그래프Trend

3

Fancim은 빠르게 성장 중인 플랫폼 서비스

•한편, 플랫폼/SNS 서비스의 경우 고객의 행동로그가 최근 일자일 수록 미래에 더 큰 영향을 주게 됨.
•따라서, 고객의 최근 활동 트렌드에 대한 가중치를 높이기 위해 Attention Layer 활용 필요[1, 2, 3]

위의 결론을 종합하여 Attention based GRU-BiLSTM[1] 모델을 제안함

03 모델 설계
모델의 특징

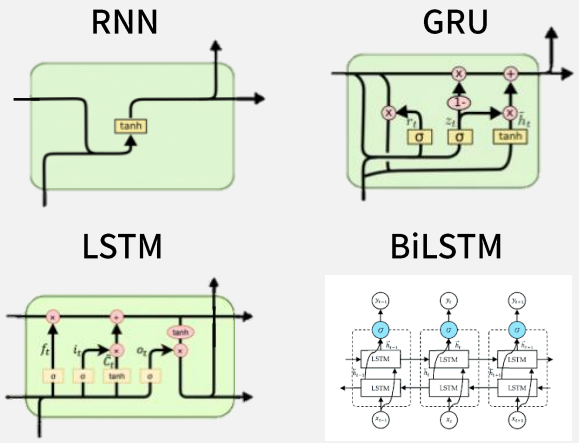
Attention based GRU-BiLSTM[1]

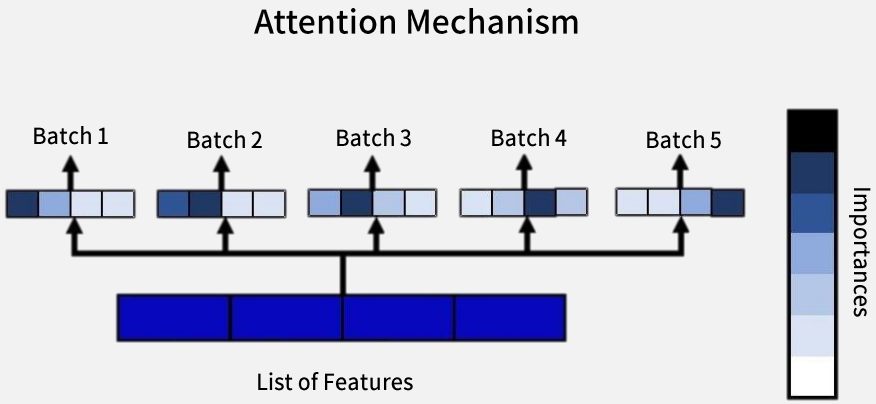
1

GRU와 BiLSTM을 조합하여 각각의 장점을 살리고 단점을 보완함
GRU는 LSTM에 비해 더 빠른 학습속도를 보이는 모델이며,
BiLSTM은 순차적인 정보를 양방향으로 학습하여, 특정 타임스텝에서 이전과 이후의 정보를 모두 고려함

2

Attention 메커니즘을 활용하여 입력 시계열 중에서 중요한 정보를 선택[3]하여 강조할 수 있음
⇒ 중요 정보 및 필요한 정보를 우선적으로 반영하여 성능을 높일 수 있음





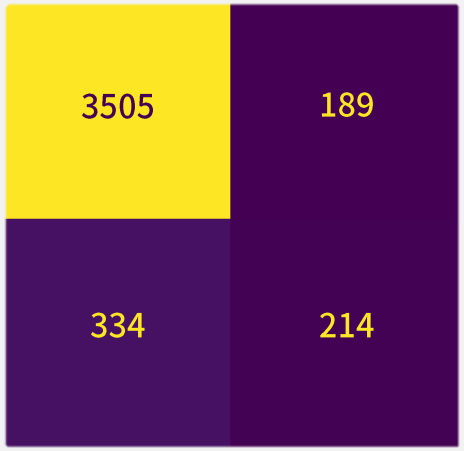
03 모델 설계
데이터 스케일링

Standard Scaling

• Vanilla 모델에 standard scaler를 적용하여 validation의 스코어를 살펴봤을 때,
Standard Scaler를 적용한 경우 이탈에 대해 보수적으로 판단하는 경향을 보였으며 accuracy와 f1 score가 근소하게 높아졌음.

•이에 따라 scaler를 적용하는 것이 영향을 줄 것으로 판단하여 스케일링 적용 후 파라미터 튜닝을 진행하였음.

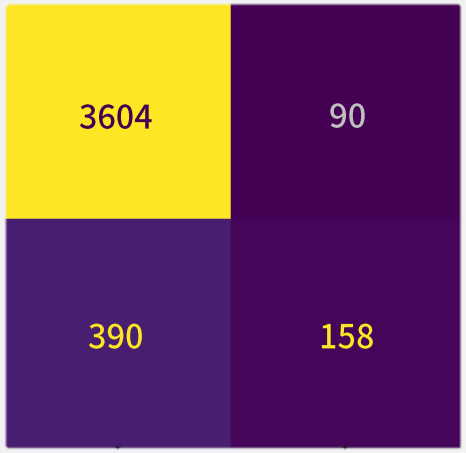
Standard Scaler 적용 이전



Accuracy 0.87
F1 score 0.86

N=4242

Standard Scaler 적용 이후



Accuracy 0.88
F1 Score 0.88

03 모델 설계
Structure of Attention based GRU-BiLSTM

Load_data

데이터 Pickle을 불러오는 함수

Sliding

시계열을 sliding하여 3차원화하는 함수(y축 = 시계열 길이)

fill-sequence

빈 시계열데이터의 값을 0으로 채우는 함수

make_array

3차원화 전, 데이터를 유저별로 분리하는 함수

make_class_2

Y_label를 2차원화하는 함수
(0 = [1, 0], 1=[0,1]

Sclaer_for_3d

3차원의 Dataset을 2차원화하여 스케일링한 후 reshape하는 함수

Train_test_split

Train, Val – train_test_split

Make_Model

Keras 모델을 구축하는 함수
Churn hidden-size 40, bins 22
Sticker hidden size 40, bins 8

Model: "model_churn"

Layer (type)	Output Shape	Param #	Connected to
Input Layer	[(None, 7, 39)]	0	[]
GRU	(None, 7, 40)	9720	['input[0][0]']
Bidirectional	(None, 7, 80)	25920	['gru[0][0]']
Dense	(None, 7, 40)	3240	['bidirectional[0][0]']
Dense	(None, 7, 22)	902	['dense[0][0]']
Dot	(None, 22, 80)	0	['dense[0][0]', 'bidirectional[0][0]']
Flatten	(None, 1760)	0	['dot[0][0]']
dense (Dense)	(None, 2)	3522	['flatten[0][0]']

Total params: 43,304
Trainable params: 43,304
Non-trainable params: 0

Model: "model_sticker"

Layer (type)	Output Shape	Param #	Connected to
Input Layer	[(None, 7, 39)]	0	[]
GRU	(None, 7, 40)	9720	['input[0][0]']
Bidirectional	(None, 7, 80)	25920	['gru[0][0]']
Dense	(None, 7, 40)	3240	['bidirectional[0][0]']
Dense	(None, 7, 8)	328	['dense[0][0]']
Dot	(None, 8, 80)	0	['dense[0][0]', 'bidirectional[0][0]']
Flatten	(None, 640)	0	['dot[0][0]']
dense (Dense)	(None, 2)	1282	['flatten[0][0]']

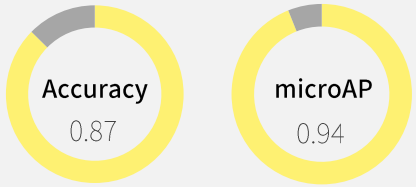
Total params: 40,490
Trainable params: 40,490
Non-trainable params: 0

04 모델 결과
이탈 감지 모델 성능 비교

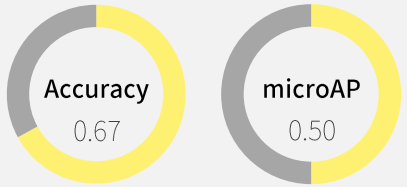
이탈감지모델은 accuracy 0.87, microAP 0.94로 다른 ML 모델에 비해 우수한 성능을 보였으며,
이는 이탈에 대해 보수적으로 판단하여 최소한의 자원으로 리텐션을 유지하는 것에 도움이 될 것으로 판단됨.

•Baseline model(최빈모델)의 accuracy(0.67)에 비해, 제안모델 Attention based GRU-BiLSTM은 accuracy 0.87으로, 87%의 정확도로 이탈자(1)/미이탈자(0)를 감지할 수 있었음.
•ML모델 LGBM classifier와 accuracy(0.87)는 비슷하나, 제안모델은 microAP에서 월등한 성능(GRU-BiLSTM 0.94, LGBM 0.86)을 보임.
•본 서비스가 스타트업의 리텐션 유지를 목적으로 하며 최소한의 자원(인적, 물적)으로 리텐션을 유지하는 데 관심이 있는 것을 고려했을 때, 이탈에 대해 보수적으로 평가하는 지표인 microAP(average precision)가 우수한 attention based GRU-BiLSTM 모델을 적절한 모델로 판단할 수 있음.

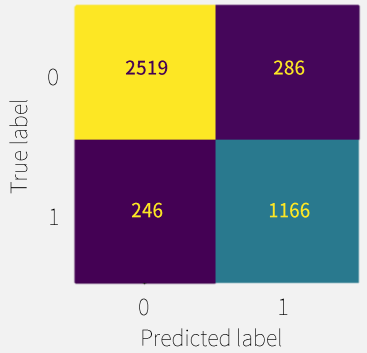
Attention Based GRU-BiLSTM



Baseline Model(최빈클래스)



Confusion Matrix of Attention Based GRU-BiLSTM
N=4217



Score Comparison of Other ML Model

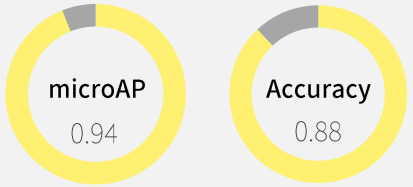
Model	Accuracy	microAP	Precisoin	Recall	F1-Score	ROC-AUC
Attention Based GRU-BiLSTM	**0.8738**	**0.9395**	**0.8030**	0.8257	0.8142	0.9308
Logistic Regression	0.8570	0.7664	0.7392	0.8852	0.8056	0.9160
Decisoin Tree Classifier	0.7585	0.5552	0.6572	0.7075	0.6814	0.7600
Random Forest Classifier	0.8705	0.8502	0.7949	0.8264	0.8104	0.9339
XGB Classifier	0.8721	0.8524	0.7851	0.8512	0.8168	0.9329
LGBM Classifier	**0.8785**	0.8617	0.7929	**0.8626**	0.8263	0.9392
Kneighbors Classifier	0.8370	0.7387	0.7241	0.8293	0.7731	0.8875
Gaussian NB	0.6990	0.5379	0.5282	0.9461	0.6780	0.7784

04 모델 결과 및 해석
스티커 예측 모델 성능 비교

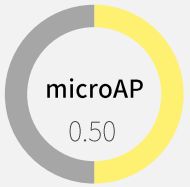
**스티커예측모델은 accuracy 0.88, microAP 0.94로 다른 ML 모델에 비해 월등히 높은 성능을 보였으며,
데이터 불균형을 고려했을 때 높은 정확도를 보여줌.**

•기간 내 스티커를 받을 셀럽(1)과 그렇지 못할 셀럽(0)의 균형비가 1:9에 가깝기 때문에, accuracy보다 데이터 불균형 시 사용하는 지표인 f1-score나 microAP가 평가에 더 적절하며, 특히 AP(average precision)는 임계치threshold 에 따른 precision-recall 을 종합적으로 판단하기에 강건한 모델을 판단하는데 더 적합한 지표임.
•제안모델은 타 ML모델에 비해 microAP 뿐만 아니라 accuracy, ROC-AUC 등 다양한 지표에서 ML모델에 비해 월등한 성능을 보였음.

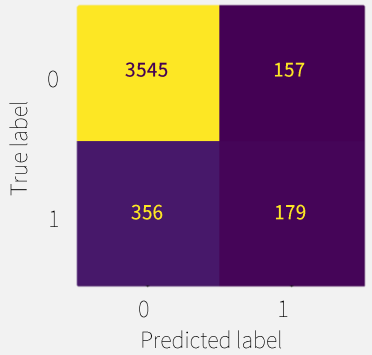
Attention Based GRU-BiLSTM



Baseline Model(최빈클래스)



Confusion Matrix of
Attention Based GRU-BiLSTM
N=4217



Score Comparison of Other ML Model

Model	Accuracy	microAP	Precisoin	Recall	F1-Score	ROC-AUC
Attention Based GRU-BiLSTM	**0.8830**	**0.9404**	0.5664	0.3345	0.4206	0.8225
Logistic Regression	0.7825	0.3628	0.3018	0.5439	0.3882	0.7943
Decisoin Tree Classifier	0.7730	0.2003	0.2769	0.4897	0.3538	0.6514
Random Forest Classifier	0.7869	0.4932	0.5183	0.4224	0.4564	0.81
XGB Classifier	0.7756	0.3925	0.2848	0.5084	0.3651	0.7817
LGBM Classifier	0.7766	0.4289	0.2846	0.5028	0.3635	0.7922
Kneighbors Classifier	0.7597	0.2423	0.2831	0.5831	0.3811	0.717
Gaussian NB	0.6874	0.2357	0.2449	0.7028	0.3623	0.7444

04 모델 결과 및 해석
SHAP by Surrogate Model

1 SHAP by Surrogate Model[1]

•XAI 기법으로 음과 양의 영향력을 둘 다 계산하는 SHAP 분석을 수행하고자 함
•딥러닝 모델의 경우 Kernel SHAP을 이용하여 SHAP value를 계산하게 되는데, 이 경우 아주 많은 시간(> 5,000h)을 소요하게 됨
•이에 따라, 기존에 학습된 모델과 유사한 트리 기반의 전역적 대리모델(Surrogate Model)을 만들어 Tree SHAP을 수행하였음

2 다중공선성 제거

•SHAP value는 다중공선성에 영향을 받는 분석 방법임
•이에 따라 설명력이 낮은 변수를 제거한 후 변수 간의 상관관계와 다중공선성을 살펴,
10 이상의 다중공선성을 가지는 변수를 제거한 후 총 39개 중 25개의 칼럼을 남김.

3 Boruta-SHAP[2] 활용

SHAP 분석의 유용성을 높이고 변수의 중요도가 분할되어 under-estimated 되지 않도록
중요하지 않은 변수를 삭제하고 나머지 변수로 기존모형을 최대한 모사하도록 만들 필요가 있음.
이에 따라 Boruta-SHAP 알고리즘(percentile 70)을 적용한 결과, 25*7개 변수 중 이탈 모형에선 39, 스티커 모형에선 45개 변수가 통계적으로 유의한 것으로 나타남.

결과, Surrogate Model의 유사도는 이탈모델 accuracy 0.96, 스티커모델 microAP 0.94 이었음

**다중공선성이 높은 변수 및 설명력이 낮은 변수를 제거하고 Boruta-SHAP을 통해 통계적으로 유의한 변수만 남긴 후,
기존 모델을 최대한 모사하는 Surrogate Model을 통해 모델을 해석했음.**

* 다중공선성 : 어떤 독립 변수가 다른 독립 변수의 조합으로 표현될 수 있는 경우

04 모델 결과 및 해석

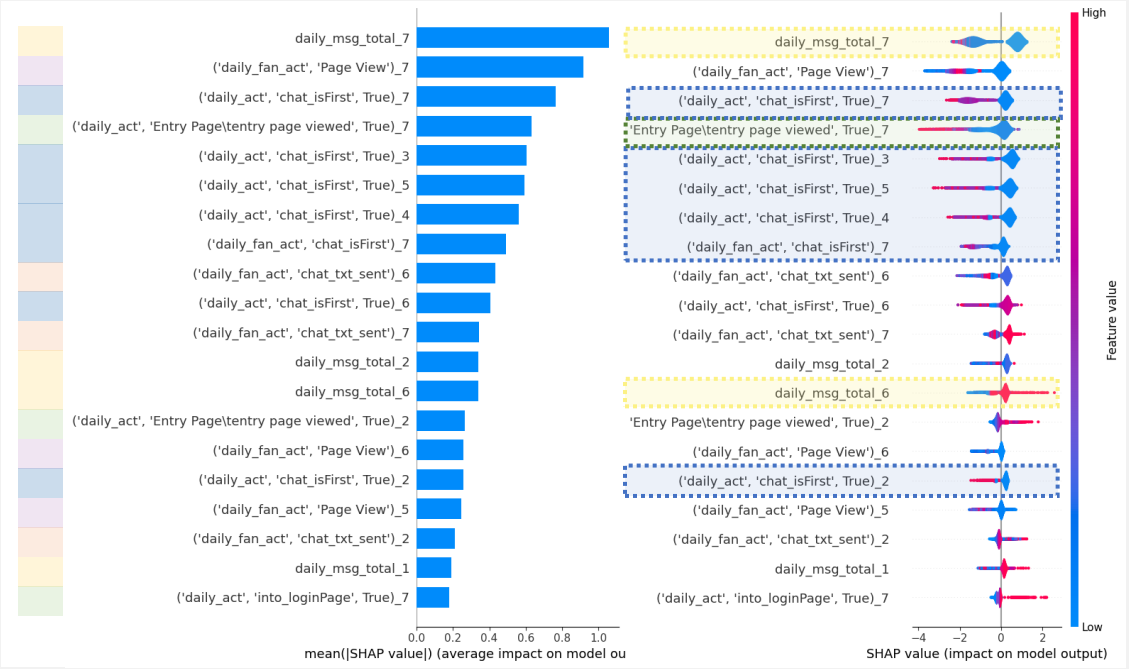
이탈 감지 모델 결과 해석(XAI) – SHAP with Surrogate Model

이탈 모형에서는 이전 일자의 셀럽의 일일 메시지 수, 팬의 채팅 횟수, 입장한 팬 수, 셀럽의 로그인 횟수 등이 중요 변수로 꼽힘.
특히 최근 일자의 기록을 우선적으로 반영했음을 알 수 있으며, 이전날 접속이 적을 때나 최근 입장 팬이 적을 때 이탈자로 판단할 확률이 높아졌음.

•이탈 모형에서는 이전 일자의 팬과의 채팅이력 횟수(chat_txt_sent), 셀럽의 일일 메시지 수(daily_msg_total), 입장한 팬 수(chat_isFirst), 셀럽의 로그인 이력 횟수(login succeeded, into_loginPage) 등이 중요 컬럼으로 꼽힘
•Feature importance를 통해 최근 일자의 기록을 우선적으로 반영했다는 것을, SHAP을 통해 이전날 접속(Entry Page)이 적을 때, 최근 입장 팬 수(chat_isFirst)가 적을 때, 최근 로그인 이력(login_succeeded, into_loginPage)이 많을 때, 이탈자로 판단할 확률이 높아진다는 것을 알 수 있음.
•이전일의 일일 메시지 수(daily_msg_total_7) 는 이탈판단확률을 낮추고 그 전일의 일일 메시지 수(6)는 이탈확률을 높였는데, 이전일과 그전일의 대비를 통해서 이탈자를 판단한 것으로 추측됨.

	Feature	Importance		Feature	Importance
1	{daily_fan_act. chat_txt_sent)_7	0.2999	11	{daily_act. chat_isFirst)_5	0.0216
2	daily_msg_total_7	0.1577	12	{daily_act. chat_isFirst)_4	0.0215
3	daily_msg_total_6	0.0669	13	daily_msg_total_3	0.0214
4	{daily_fan_act. chat_txt_sent)_6	0.0457	14	{daily_act. chat_isFirst)_2	0.0192
5	{daily_fan_act. chat_txt_sent)_5	0.0457	15	{daily_act. login_succeeded, True)_4	0.0152
6	daily_msg_total_4	0.0372	16	{daily_fan_act. chat_isFirst, True)_7	0.0151
7	daily_msg_total_5	0.0365	17	daily_msg_total_2	0.0133
8	{daily_act. chat_isFirst, True)_3	0.0314	18	{daily_fan_act. chat_txt_sent)_2	0.0084
9	{daily_act. chat_isFirst, True)_7	0.0291	19	{daily_act. login_succeeded, True)_5	0.0083
10	{daily_act. chat_isFirst, True)_6	0.0231	20	{daily_act. login_succeeded, True)_3	0.0080

TOP 20 of Feature Importance of Churn Surrogate Model



SHAP of Churn Surrogate Model

04 모델 결과 및 해석

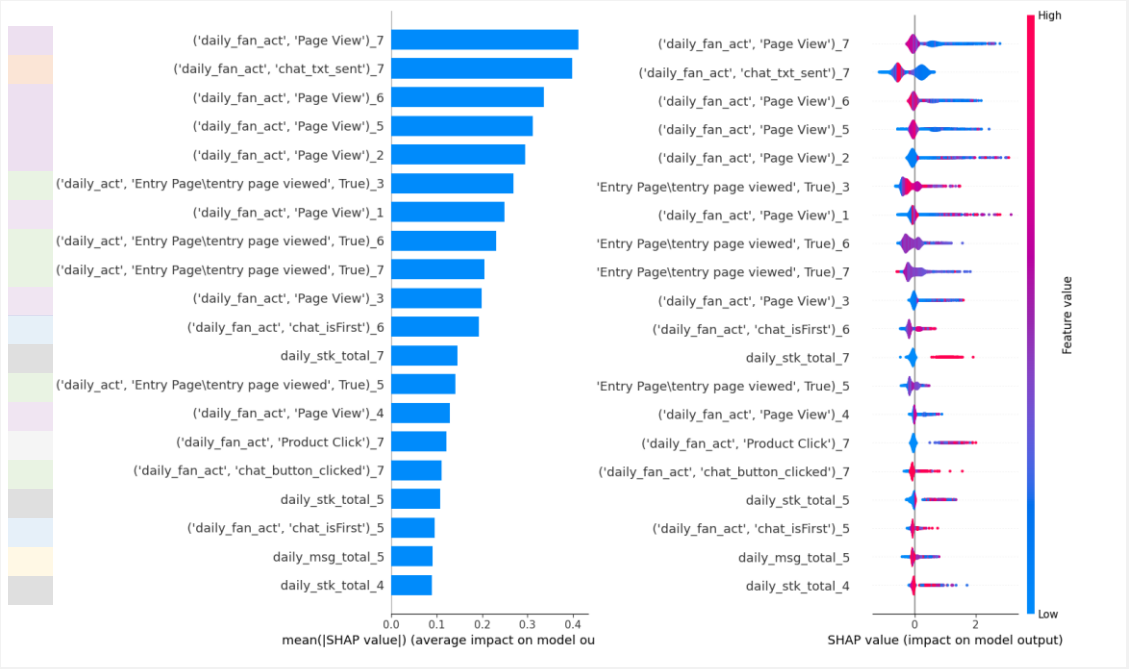
스티커 예측 모델 결과 해석(XAI) – SHAP with Surrogate Model

스티커 예측 모형에서는 이전 일자의 팬의 샷 구매횟수, 일일 땡큐스티커 합계, 셀럽의 접속 횟수, 팬의 채팅 횟수 등이 중요 변수로 꼽혔음.
특히 입장 팬 수와 팬의 채팅방활동이력이 스티커 확률에 영향을 주었기에, 팬의 수와 상호작용 역시 스티커 확률에 영향을 미침을 알 수 있음.

•스티커 모형에서는 이전 일자의 팬의 샷 구매횟수(Product Click), 셀럽의 일일 땡큐스티커 합계(daily_stk_total), 팬의 채팅방기능 사용횟수(Page View), 입장한 팬 수(chat_isFirst), 셀럽의 접속횟수(Entry Page), 팬의 채팅횟수(fan/chat_txt_xent) 등이 중요 컬럼으로 꼽힘
•스티커 모형 역시 최근 일자의 기록을 우선적으로 반영했고, 팬의 채팅방기능 사용 이력(Page View), 입장 팬 수(chat_isFirst), 그리고 이전일자에 팬이 샷을 구매(Product click)했거나 땡큐스티커를 받았던(daily_stk_total) 셀럽은 그 다음주에도 받을 확률에 영향을 미친다는 것을 알 수 있음.

	Feature	Importance		Feature	Importance
1	daily_fan_act, Product Click_ 7	0.5873	11	daily_fan_act, Product Click_ 6	0.0090
2	daily_stk_total_7	0.0579	12	daily_fan_act, Page View_ 3	0.0075
3	daily_stk_total_6	0.0356	13	daily_fan_act, chat_button_clicked_7	0.0073
4	daily_fan_act, Page View_ 6	0.0287	14	daily_fan_act, chat_isFirst_ 5	0.0070
5	daily_fan_act, Page View_ 7	0.0215	15	daily_fan_act, chat_isFirst_ 7	0.0069
6	daily_fan_act, Page View_ 5	0.0196	16	daily_fan_act, Page View_ 1	0.0067
7	daily_stk_total_4	0.0172	17	daily_fan_act, Page View_ 4	0.0065
8	daily_fan_act, chat_isFirst_ 6	0.0167	18	daily_fan_act, chat_txt_sent_3	0.0053
9	daily_stk_total_5	0.0135	19	Daily_act, Entry page viewed, 7	0.0053
10	daily_fan_act, chat_txt_sent_7	0.0109	20	Daily_act, Entry page viewed, 6	0.0052

TOP 20 of Feature Importance of Sticker Surrogate Model

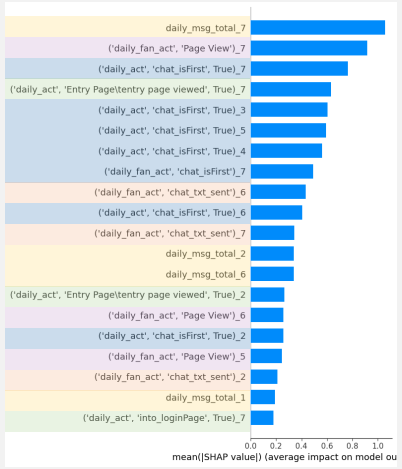


SHAP of Sticker Surrogate Model

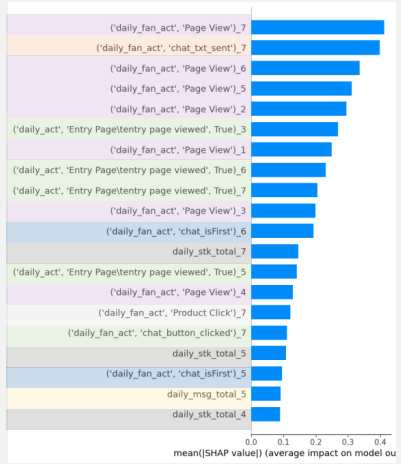
04 모델 결과

두 모델에 영향을 주는 공통 변수 분석

	Feature		Feature
1	{daily_fan_act. chat_txt_sent)_7	11	{daily_act. chat_isFirst)_5
2	daily_msg_total_7	12	{daily_act. chat_isFirst)_4
3	daily_msg_total_6	13	daily_msg_total_3
4	{daily_fan_act. chat_txt_sent)_6	14	{daily_act. chat_isFirst)_2
5	{daily_fan_act. chat_txt_sent)_5	15	{daily_act. login_succeeded, True)_4
6	daily_msg_total_4	16	{daily_fan_act. chat_isFirst, True)_7
7	daily_msg_total_5	17	daily_msg_total_2
8	{daily_act. chat_isFirst, True)_3	18	{daily_fan_act. chat_txt_sent)_2
9	{daily_act. chat_isFirst, True)_7	19	{daily_act. login_succeeded, True)_5
10	{daily_act. chat_isFirst, True)_6	20	{daily_act. login_succeeded, True)_3



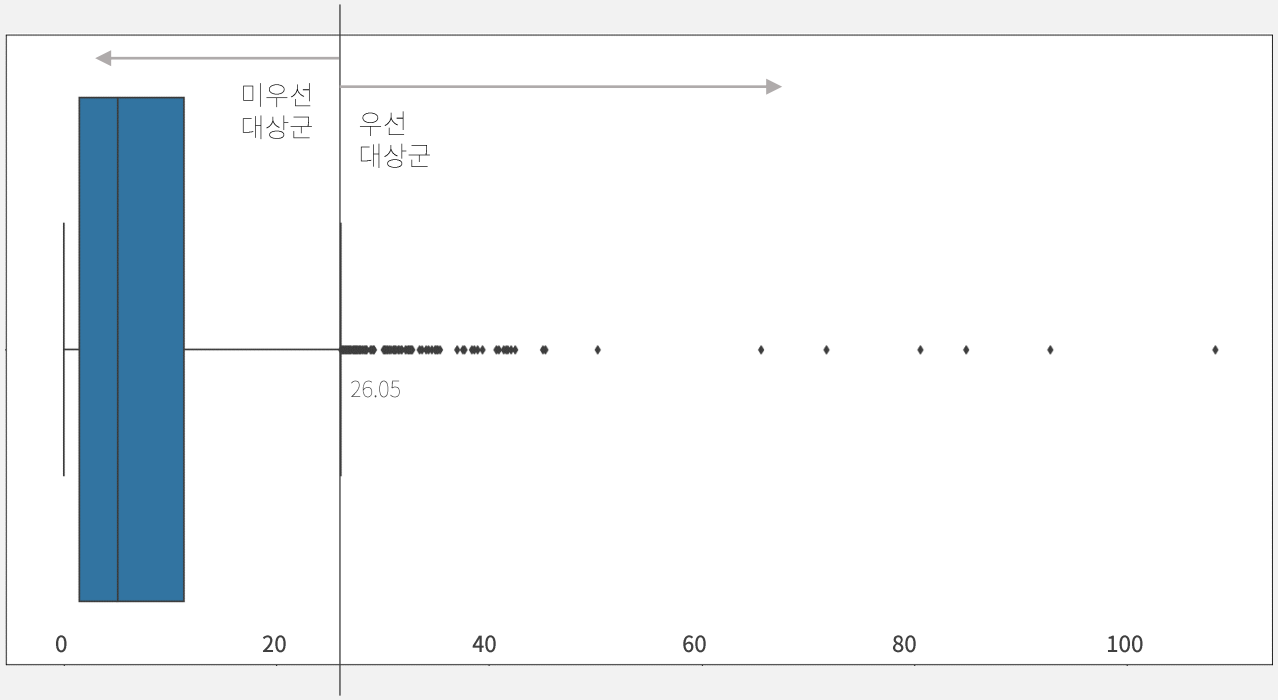
	Feature		Feature
1	daily_fan_act, Product Click_ 7	11	daily_fan_act, Product Click_ 6
2	daily_stk_total_7	12	daily_fan_act, Page View_ 3
3	daily_stk_total_6	13	daily_fan_act, chat_button_clicked_7
4	daily_fan_act, Page View_ 6	14	daily_fan_act, chat_isFirst_ 5
5	daily_fan_act, Page View_ 7	15	daily_fan_act, chat_isFirst_ 7
6	daily_fan_act, Page View_ 5	16	daily_fan_act, Page View_ 1
7	daily_stk_total_4	17	daily_fan_act, Page View_ 4
8	daily_fan_act, chat_isFirst_ 6	18	daily_fan_act, chat_txt_sent_3
9	daily_stk_total_5	19	Daily_act, Entry page viewed, 7
10	daily_fan_act, chat_txt_sent_7	20	Daily_act, Entry page viewed, 6



결론

- 이탈모델에서의 주요컬럼은 이전 일자의 셀럽의 일일 메시지 수, 팬의 채팅 횟수, 입장한 팬 수, 셀럽의 로그인 횟수였음.
- 스티커모델에서의 주요컬럼은 이전 일자의 팬의 샷 구매횟수, 일일 땡큐스티커 합계, 셀럽의 접속 횟수, 팬의 채팅방 활동이력이었음.
- 전체적으로 최근 일자의 시계열 변수가 모델에 좀 더 큰 영향력을 끼쳤음.
- 두 모델에 공통적으로 중요하게 작용했던 컬럼은 팬의 채팅 횟수와 입장 팬 횟수로, 직접적인 기록(이전 일자의 셀럽의 메시지 이력, 샷 받은 이력) 외에도 팬과의 상호작용 역시 이탈율과 스티커확률에도 영향을 줌을 알 수 있음.
- 셀럽의 리텐션 유지를 위해선 팬의 영향력이 중요하게 작용함을 알 수 있음.

05 셀럽 세그멘테이션
관리우선점수 부여 및 대상군 도출



이탈확률*스티커확률의 분포 Box Plot

- 각 모델에서 1로 분류될 확률을 도출 후 log를 취해 확률 분포의 왜도를 완화하고 두 값을 곱하여 분포를 확인함
- 값이 높을 수록 이탈확률도 높고 스티커 받을 확률도 높은 대상자임.
- 전체 대상자 = 4217
- 75% 기준 대상자(>11.29) = 1054
- Maximum line 기준 대상자(>26.05) =109
- 75%(1,054)로 지정 시, 너무 많은 대상자가 포함되어 management의 한계를 넘어섬.
- 따라서 기준값을 maximum line의 값으로 지정하여, 그 이상으로 분류된 유저를 관리우선대상군으로 지정함

각 모델에서 도출된 이탈확률*스티커확률probability에 각각 로그를 취하여 곱한 값을 value로 부여한 후,
기준값 이상의 유저를 관리우선대상군으로 분류하여 두 세그먼트의 특징을 도출

05 셀럽 세그멘테이션
관리우선대상군의 특징 비교

T-test를 통해 두 집단의 변수 별 표본평균 차이(t검정량)를 확인한 결과, 셀럽의 접속 및 기타 페이지 열람 횟수, 타 셀럽의 채팅방 접속 횟수, 프로필 확인 횟수, 입장 팬 수, 채팅방 링크 외부 공유 횟수 등에서 유의한 차이를 보임.

•두 집단의 배치별 각 컬럼 별 평균값을 t-test하여 평균적으로 얼마나 차이나는지(T-통계량)와 유의확률(p-value)을 검정한 결과, 모든 배치에서 유의한 차이를 보이는 컬럼은 셀럽의 첫페이지 접속 횟수, 기타 페이지 접속 횟수, 타 셀럽 리스트 확인 횟수, 타 채팅방 접속 횟수, 셀럽 프로필 확인 횟수, 입장한 팬 수, 채팅방 링크 외부 공유 횟수 등이 있었음.
•이를 통해 관리우선대상군은 다른 셀럽에 많은 관심을 보이며, 시계열적으로 채팅방 공유 횟수가 점점 줄어드는 등 추가적인 특징이 있음을 확인할 수 있음.
•특히 일주일 사이에 평균 입장 팬 수가 약 3.3명 하락하는 추세를 보였음.

Features

Batch of Time series

	1	2	3	4	5	6	7
Daily_act, Entry Page entry page viewed	3.60	2.86	3.09	2.71	2.70	2.46	2.10
Daily_act, Page View	3.28	2.92	3.28	2.83	2.46	1.85	1.38
Daily_act, celeb_list_viewed	4.30	4.02	4.39	4.07	3.72	3.06	2.49
Daily_act, celeb_profile_page_viewed	3.95	4.54	5.30	6.16	4.67	3.07	2.19
Daily_act, chat_button_clicked	5.92	5.92	5.92	4.06	8.29	5.37	3.91
Daily_act, chat_isFirst	10.00	9.07	9.33	8.56	7.16	7.57	6.70
Daily_act, share_chatLink_clicked	4.82	4.03	3.83	2.95	3.50	0.89	1.50

T-Statistic

	1	2	3	4	5	6	7
Daily_act, Entry Page entry page viewed	0.00	0.00	0.00	0.01	0.01	0.01	0.04
Daily_act, Page View	0.00	0.00	0.00	0.00	0.01	0.06	0.17
Daily_act, celeb_list_viewed	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Daily_act, celeb_profile_page_viewed	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Daily_act, chat_button_clicked	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Daily_act, chat_isFirst	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Daily_act, share_chatLink_clicked	0.00	0.00	0.00	0.00	0.00	0.37	0.13

P-value

관리우선대상군과 비대상군의 변수 별 평균의 통계적 차이 검정 결과(T-test)

06 프로젝트 마무리
프로젝트 회고

[이후 시도해볼 수 있는 것들]

코호트 분석을 통한 분석 고도화
간단하게 일주일접속집단/ 미접속집단으로 나누어 이탈율을 분석했지만,
가입시기 별로도 이탈율을 분석하면 또다른 인사이트를 발견할 수 있을 것이라 생각됨.

세그먼테이션 고도화
Kmeans를 시도했지만 비지도학습의 결과를 해석하는데 있어 설명 가능한 결과를 내지 못해 자체점수를 부여하여 t-test하는 방식으로 변경함.
점수를 좀 더 변별력 있게 부여하는 다른 방식이나, 유저 세그멘테이션 후 인사이트를 도출하는 다른 방법에 대해 고민해볼 필요가 있음.

도메인과 마케팅 방법론 등 다양한 지식의 필요성
모델 결과 해석에 있어서 모델 지식 뿐만 아니라 도메인 지식과 마케팅 전략 등 도메인에 대한 한계를 느꼈음.
마케팅 전략을 이렇게 수정하면 좋겠다 등 방법론적인 부분을 더 자세히 알고 있으면 모델 해석 역시 더 다양한 방면으로 해석할 수 있을 듯 하여
예시를 많이 살펴볼 필요성을 느꼈음.

이상탐지모델로의 전환
전체적으로 이탈감지모델은 이상탐지모델과도 구조나 목표가 비슷하다고 느꼈음.
동일한 모델 구조로 이상탐지와 관련된 다른 시계열데이터셋에도 응용해 볼 수 있을 듯 함.

06 프로젝트 마무리
참고문헌

11, 12p

[1] Britto, M. M. J., & Gobinath, D. R. (2021). Improved Churn Prediction Model In Banking Industry And Comparison Of Deep Learning Algorithms. Int. J. of Aquatic Science, 12(2), 2521-2529.

[2] 박수연. "고객 이용 로그와 순환신경망을 활용한 이커머스 고객 이탈 예측." 국내석사학위논문 고려대학교 컴퓨터정보통신대학원, 2023. 서울

[3] 이준석, Lecture 9. Attention Mechanism & Transformers https://www.youtube.com/watch?v=iiXRuCZ0ww8&t=2654s

17p

[1] Surrogate Model Ahn, J. H., XAI, Dissects Artificial Intelligence, Wiki Books, 2020.

[2] Boruta-SHAP Keany, E., Boruta-Shap: A Tree Based Feature Selection Tool which Combines Both the Boruta Feature Selection Algorithm with Shapley Values, 2019. [Website] (2023, Apr .17). https://github.com/Ekeany/Boruta-Shap.

그 외
"[NDC] 실전 이탈 예측과 유의사항", 유튜브 비디오, 00:23:54, 게시자 "NDC", 2018.07.02. https://www.youtube.com/watch?v=kcE_1n41xdk&t=3s

06 프로젝트 마무리
프로젝트 구성

김정겸	- 데이터 EDA 및 Preprocessing - 모델 구축 : 하이퍼파라미터 튜닝
이유빈	- 데이터 EDA 및 Preprocessing
허소영	- 서비스 관찰 및 데이터 EDA
심지은	- 데이터 EDA 및 Preprocessing : 분해시계열, heuristic mining - 모델 구축 : 모델 실험, 하이퍼파라미터 튜닝(baysian search) - 모델 해석 : surrogate model 튜닝, 결과 도출 및 해석 - 유저 세그멘테이션 - 협업툴(노션, 구글 ppt) 템플릿 구성 및 최종 보고서 작성
허선우	- 데이터 EDA 및 Preprocessing : Sankey Diagram, 누적확률분포분석, sliding window, 전처리 코드 모듈화(.py) - 모델링 방법론 제안 : Boruta-SHAP을 활용한 Feature Selection, 대리모델을 활용한 SHAP, 다중공선성 제거 - 모델 구축 : Attention based GRU-BiLSTM

감사합니다