잔존가치를 고려한 유저 이탈 예측

Big Contest 2019

Analysis 분야 챔피언 리그

남조선 자료 공작단

요원1: 박상민

요원2: 허재혁

요원3: 문현종

요원4: 박용연

요원5: 이정환

INDEX

01

주제 및 데이터의 이해

01. 주제 및 데이터 이해

대회 주제

리니지 유저 활동 데이터를 활용하여 잔존가치를 고려한 이탈예측 모형 개발

대회 설계 의도

1. 시간의 변화에 강건한 모델 구축
2. 잔존가치를 고려한 모델 구축

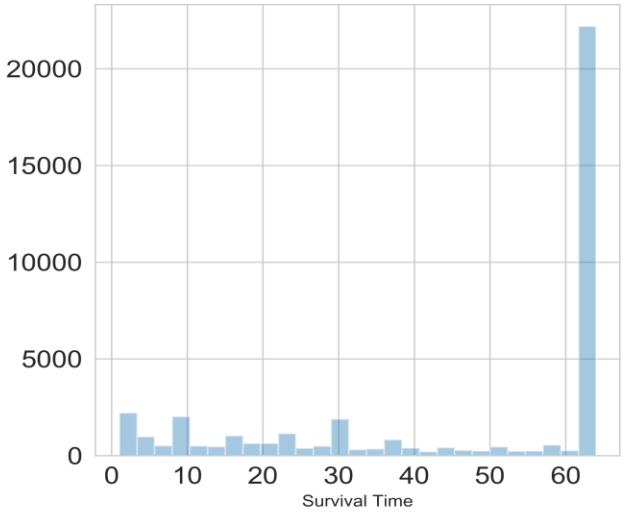
모델 평가 방법

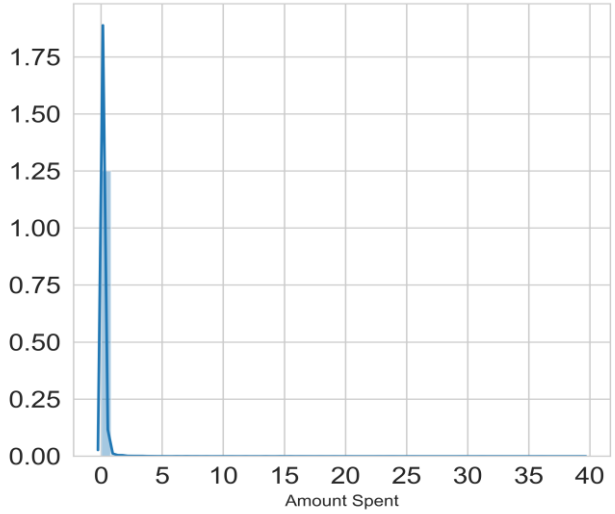
1. 예측 성능(유저별 기대 이익의 총합)
2. 재현성 테스트

01. 주제 및 데이터 이해

생존 기간, 일 평균 결제 금액

생존 기간은 1~64까지의 값을 가지며, 64가 가장 많은 빈도를 보임. 일 평균 결제 금액은 대부분의 값들이 0에 근사한 값들을 가지고 있음.





02

탐색적 자료 분석

02. 탐색적 자료 분석

데이터 이슈

대회 제공 데이터의 이슈사항 해결

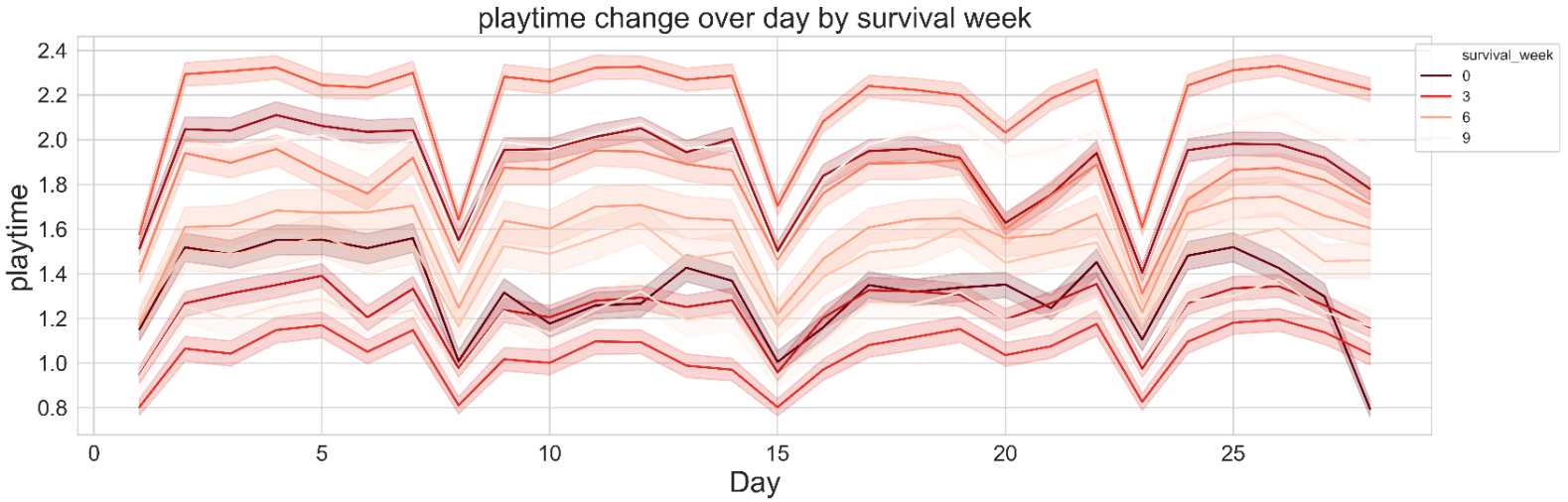
1. Fishing이 0이 아님에도 불구하고 Playtime이 0인 경우 7,145건 존재.
→ 모두 Fishing의 값으로 대체.

2. 혈맹 데이터(Pledge)에 전투 캐릭터 플레이 시간의 합 (combat_play_time)과 비전투 캐릭터 플레이 시간의 합 (non_combat_play_time)의 집계 오류로 사용변수에서 제외

02. 탐색적 자료 분석

일일 플레이 시간(Playtime)

유저들의 플레이 시간의 합계를 보았을 때, 활동일에 따라 주 단위로 구분이 되는 경향을 보임.

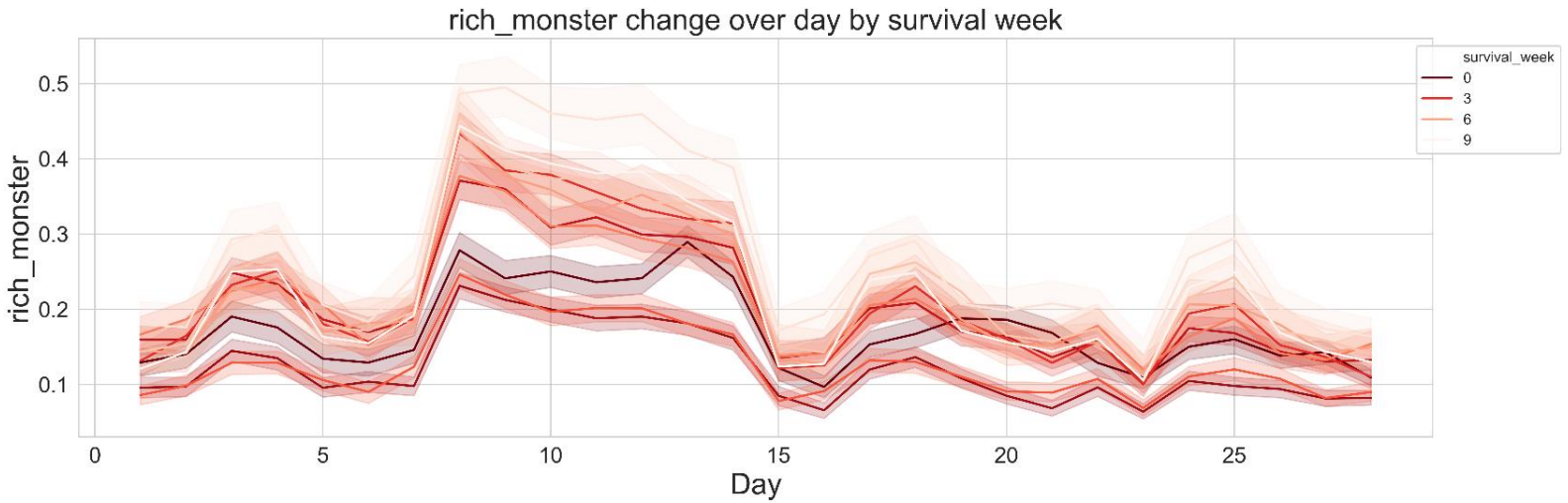


※ 생존 기간을 주별로 구분, 색이 진할수록 생존기간이 짧음

02. 탐색적 자료 분석

보스 몬스터 타격 여부(Rich Monster)

유저들의 보스 타격 여부를 보았을 때, 주별 변동의 폭이 큰 경향을 보임.



※ 생존 기간을 주별로 구분, 색이 진할수록 생존기간이 짧음

02. 탐색적 자료 분석

결과

일별 데이터를 보았을 때, 전체적으로 **시간적/통계적** 요소를 고려하여 변수를 만들 필요성이 있어 보임.

03

데이터 전처리

03. 데이터 전처리

전처리 방법

시계열/통계적 요소를 고려한 방법 적용

Method 1. 일별 데이터를 주별 데이터로 변환

Method 2. 시간에 따른 변화를 2주 단위로 표현 (첫 주는 제외)

- 이동 평균(Moving Average)
- 이동 표준편차(Moving Standard Deviation)

Method 3. 변수의 최소/최대/평균/표준편차

03. 데이터 전처리

활동 데이터(Activity)

아래 변수들에 **Method 1,2** 적용

- 일일 플레이 시간(Playtime)
- NPC를 죽인 횟수(NPC Kill)
- 솔로 사냥 획득 경험치(Solo Exp)
- 파티 사냥 획득 경험치(Party Exp)
- 퀘스트 획득 경험치(Quest Exp)
- 보스몬스터 타격 여부(Rich Monster)
- 캐릭터 사망 횟수(Death)
- 부활 횟수(Revive)
- 경험치 복구 횟수(Exp Recovery)
- 일일 낚시 시간(Fihsing)
- 일일 개인상점 운영시간(Private Shop)
- 일일 아데나 변동량(Game Money Change)
- 7레벨 이상 아이템 인첸트 시도 횟수(Enchant Count)

03. 데이터 전처리

전투 데이터(Combat)

아래 변수들에 **Method 1,2** 적용

- 레벨(Level)
- 혈맹 전투 횟수(Pledge Count)
- 무작위 공격을 행한 전투 횟수(Random Attacker Count)
- 무작위 공격자로부터 공격을 받은 전투 횟수(Random Deffender Count)
- 단발성 전투 횟수(Temp Count)
- 동일 혈맹 전투 횟수(Same Pledge Count)
- 기타 전투 횟수(ETC Count)
- 전투 상대 캐릭터 수(Num Opponent)

결제 데이터(Payment)

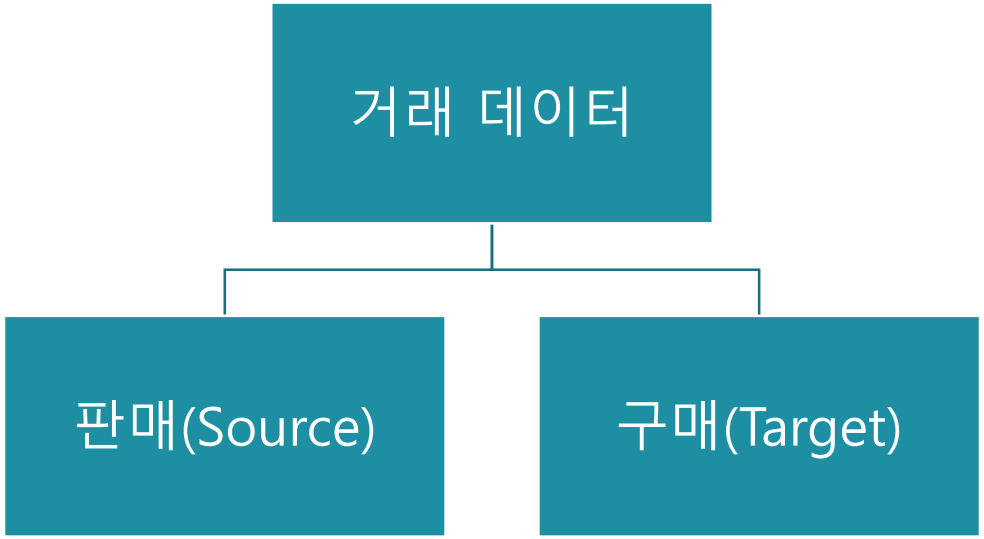
아래 변수에 **Method 1,2** 적용

- 결제 금액(Amount Spent)

03. 데이터 전처리

거래 데이터(Trade)

거래데이터는 판매와 구매 데이터로 구분하여 사용



03. 데이터 전처리

거래 데이터(Trade)

아래 변수들에 **Method 1,2** 적용

- 판매 금액(Sell Item Price)
- 판매량(Sell Item Amount)
- 구매 금액(Buy Item Price)
- 구매량(Sell Item Amount)

아래 변수들은 유저별 데이터 카운트로 만든 변수

- 판매 수(Sell Count)
- 판매일 수(Source Count)
- 구매 수(Buy Count)
- 구매일 수(Target Count)

04

모델링

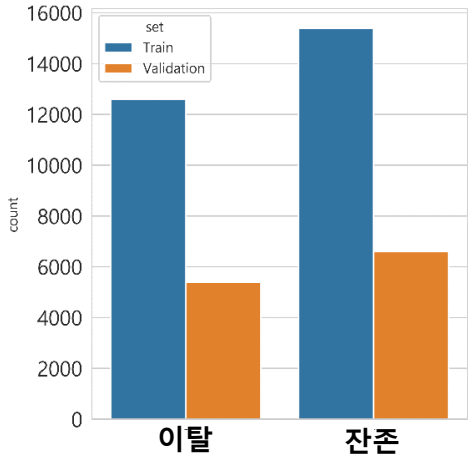
04. 모델링

Validation 방법

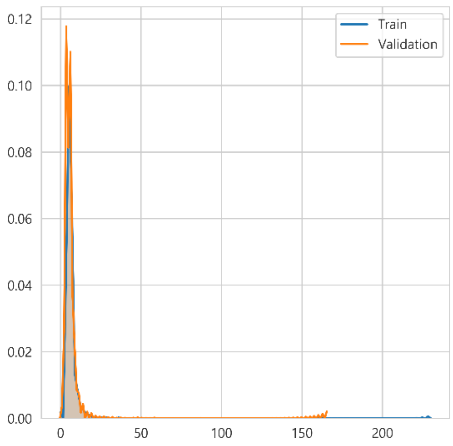
분포를 고려하여 학습데이터와 검증데이터를 분리

Train Set : 70 %	Validaiton Set : 30%

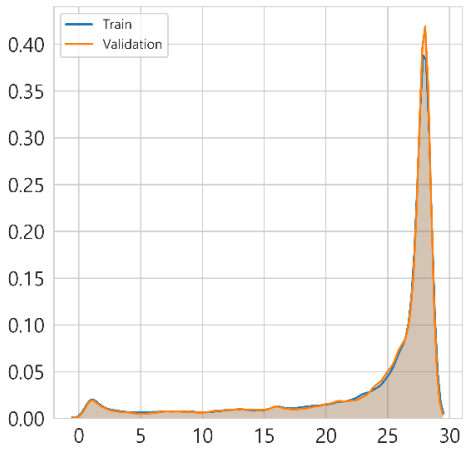
이탈/잔존



캐릭터 수



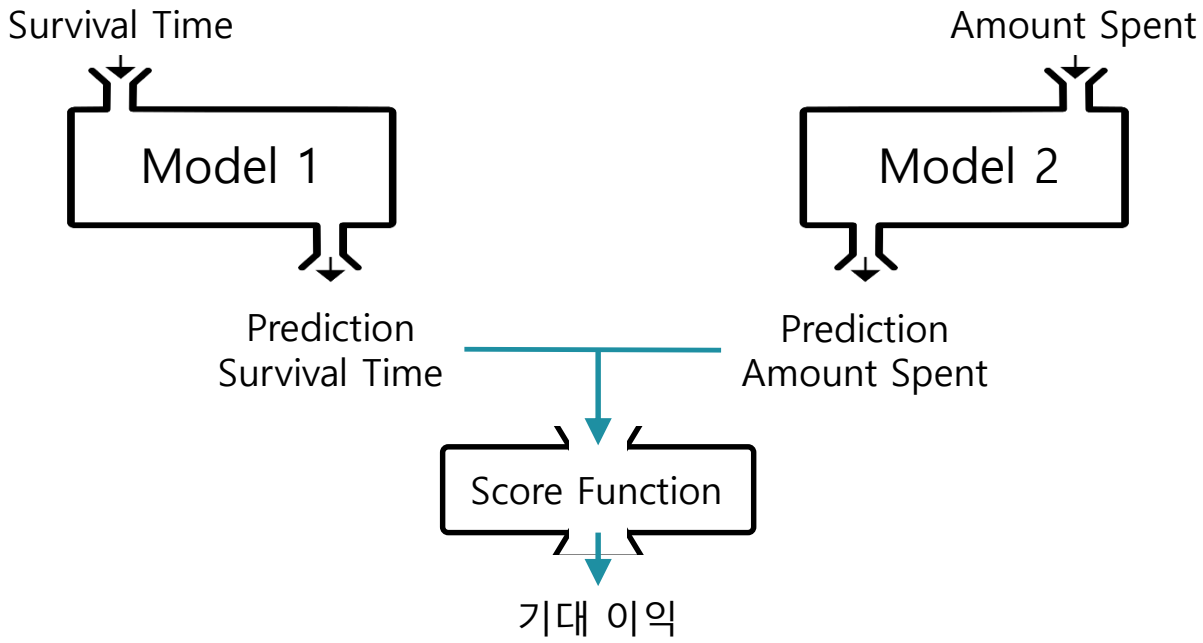
활동일 수



04. 모델링

모델링 과정

생존기간과 일 평균 결제 금액에 맞는 두 가지 모델 구성



04. 모델링

모델링 과정

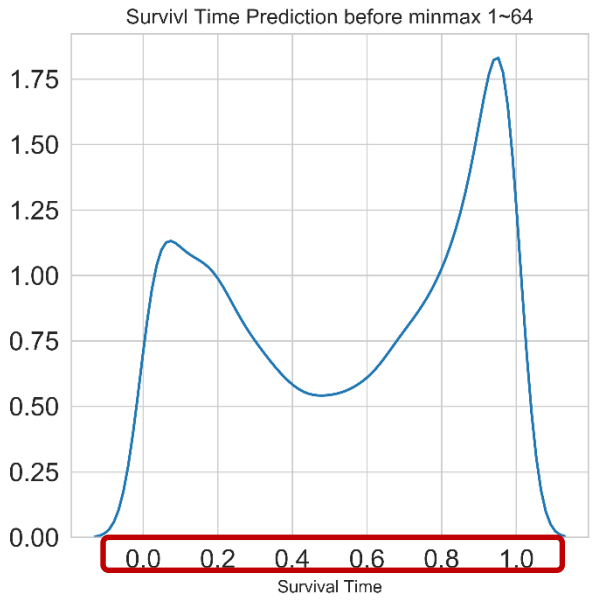
모델 성능을 최적화 하기 위해 목표변수를 변환하여 사용

04. 모델링

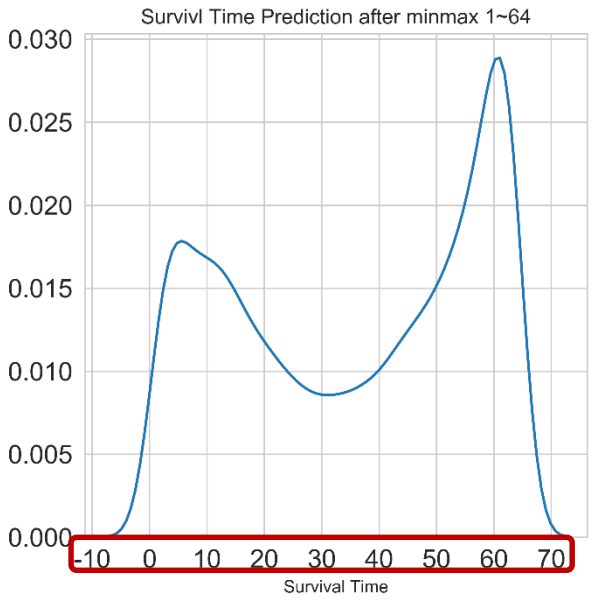
Model 1: 생존 기간(Survival Time)

생존기간의 이탈/잔존으로 분류로 학습한 후 MinMax Scaling으로 변환

이탈/잔존으로 학습



MinMax Scaling(1~64)

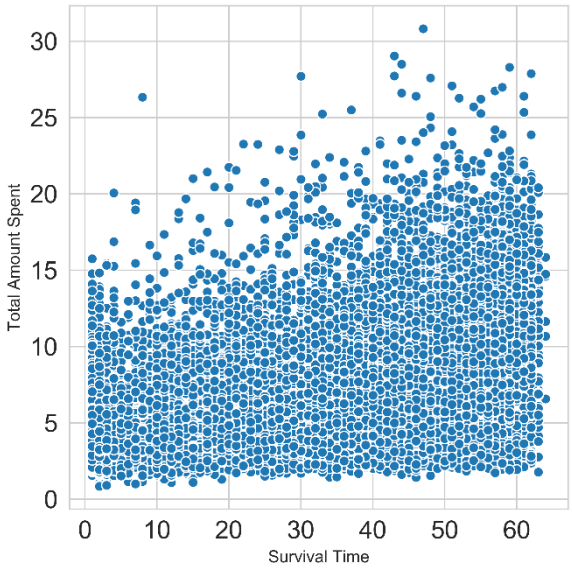


04. 모델링

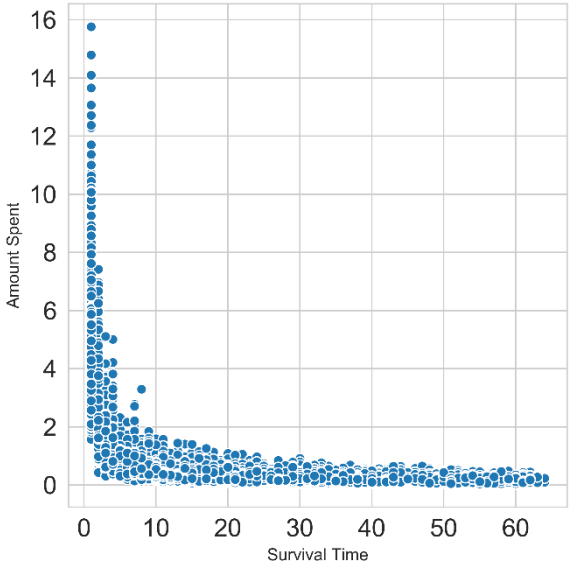
Model 2: 일 평균 결제 금액(Amount Spent)

총 결제금액을 학습한 후 예측한 생존일수로 일 평균 결제 금액 계산

$$Total\ Amount\ Spent(TR) = T \times R$$



$$Amount\ Spent(\hat{R}) = \frac{\widehat{TR}}{\hat{T}}$$



T : True Survival Time R : True Amount Spent $\hat{T}$: Prediction of Survival Time $\hat{R}$: Prediction of Amount Spent

04. 모델링

사용 모델

Light GBM

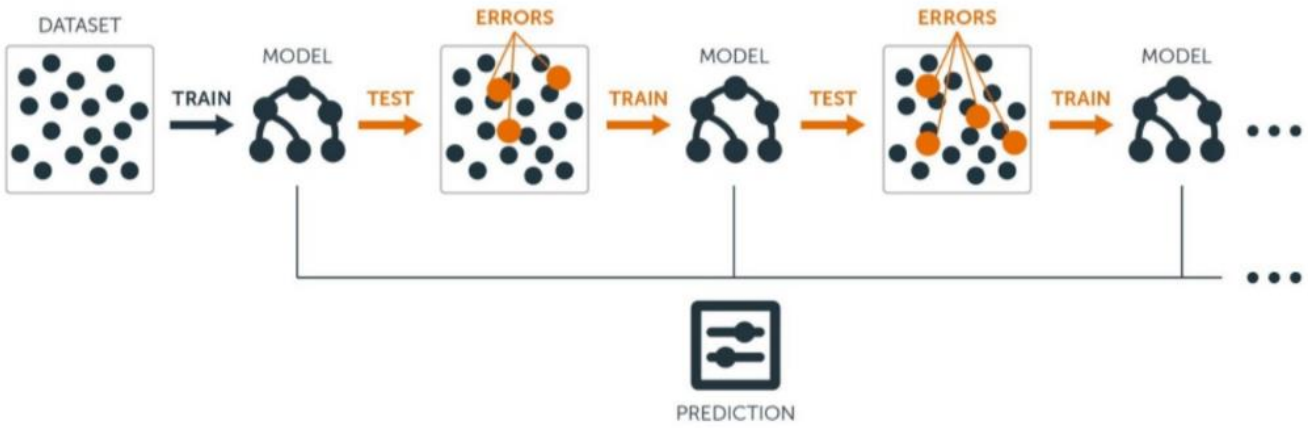
Why?

04. 모델링

Boosting 모델 사용 이유

높은 예측력을 보장, 변수간의 상관성을 고려하지 않음

- 많은 변수를 사용하여 예측력을 높이기 위해
  (Gradient boosting 기법은 오차를 예측해서 예측력을 높인다.)
- 변수 간의 다중공선성을 고려하지 않는 모델을 사용하기 위해

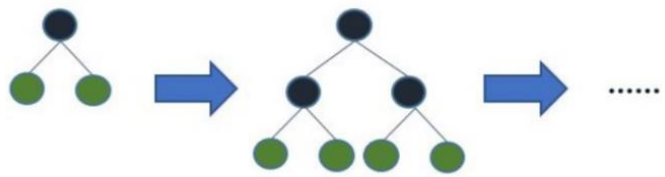
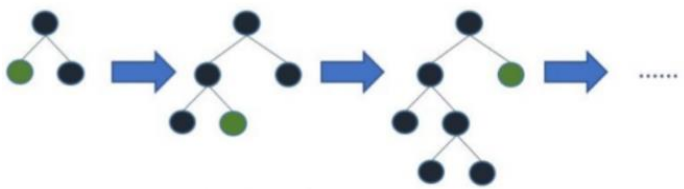


04. 모델링

LightGBM 사용 이유

대용량 데이터를 다른 부스팅 모델보다 더 빠른 시간에 학습 가능

다른 부스팅 모델들은 Level-wise tree growth기반으로 과적합되기 쉬운 경향이 있기 때문에 더 적은 메모리를 사용하며 Leaf-wise tree growth기반인 Light GBM 모델이 대용량 데이터 세트를 더 빠른 처리를 하기 때문에 채택

XGBoost	LightGBM
	

05

이탈 징후 분석 및 결론

05. 이탈 징후 분석 및 결론

모델 해석

학습 중요도, SHAP 그리고 Partial Dependence Plot을 통해 비교

학습중요도 → SHAP → Partial Dependence Plot

1. 학습 중요도(Featrue importance)
- 변수별 목표변수 분류에 영향력이 큰지 정리
- But, 학습 중요도는 다른 변수와 상호작용을 고려하지 않는 한계가 있음

2. SHAP
- 입력값이 결과값에 얼마나 영향을 미치는지 계산

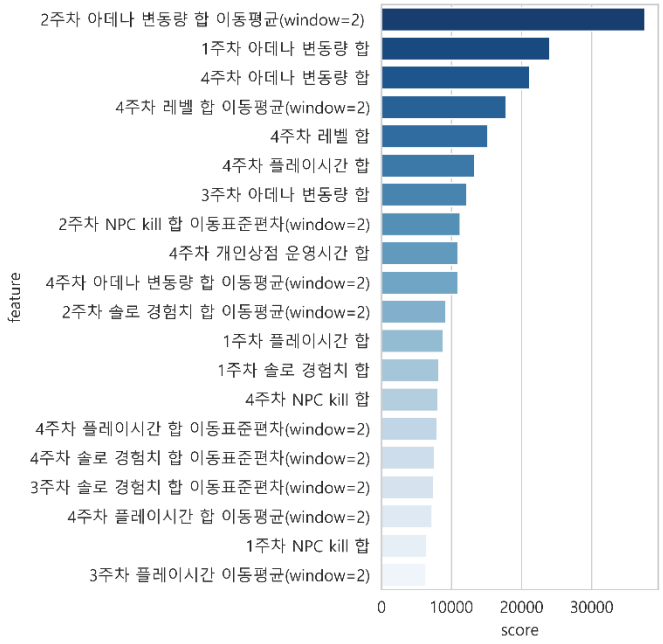
3. Partial Dependence Plot
- 입력 변수값 변화에 따라 예측 모델 결과값의 평균적인 변화를 시각화

※ 주의할 점은 모델 해석 결과를 이탈원인으로 설명하는 것과는 별개이다.

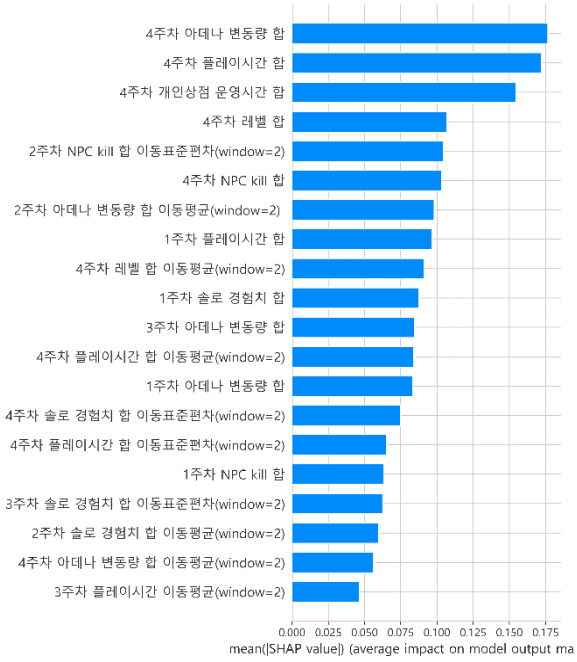
05. 이탈 징후 분석 및 결론

Model 1: 생존 기간(Survival Time)

아데나 변동량과 **4주차 데이터**들이 결과값에 큰 영향력을 가짐



변수 중요도

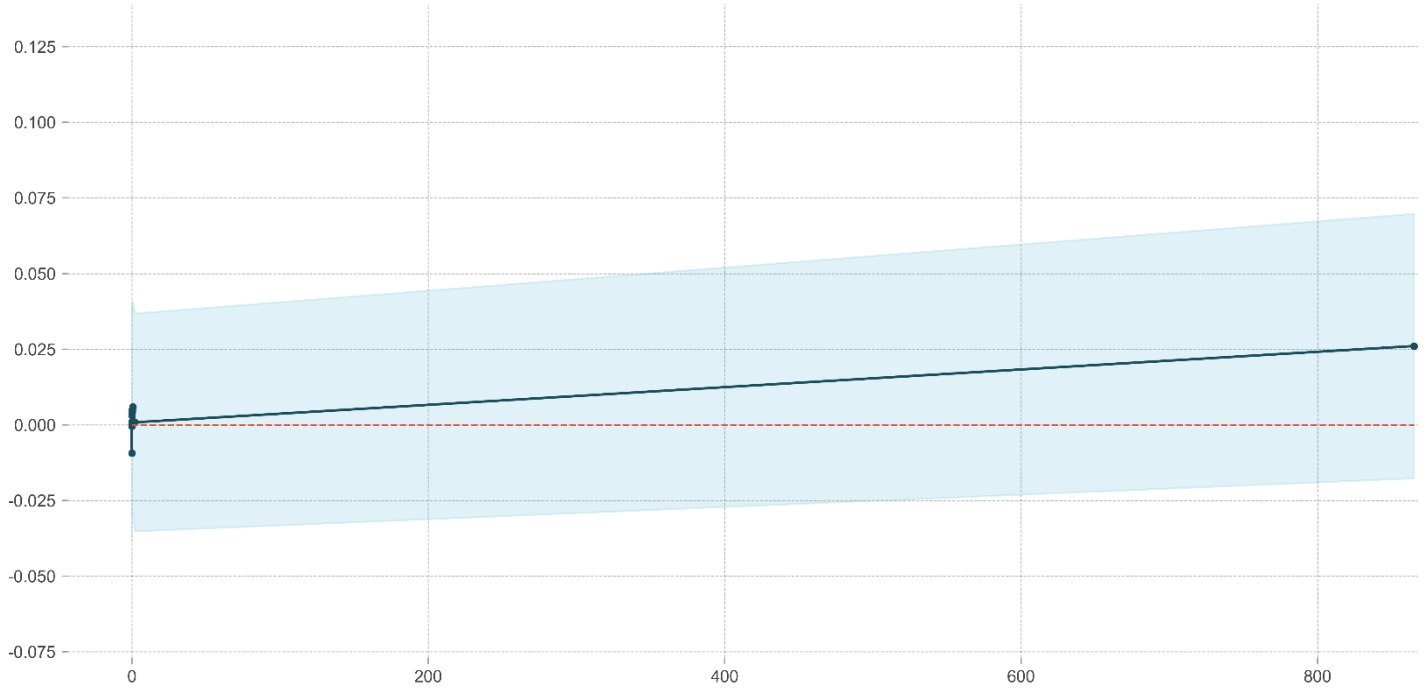


SHAP

05. 이탈 징후 분석 및 결론

Model 1: 생존 기간(Survival Time)

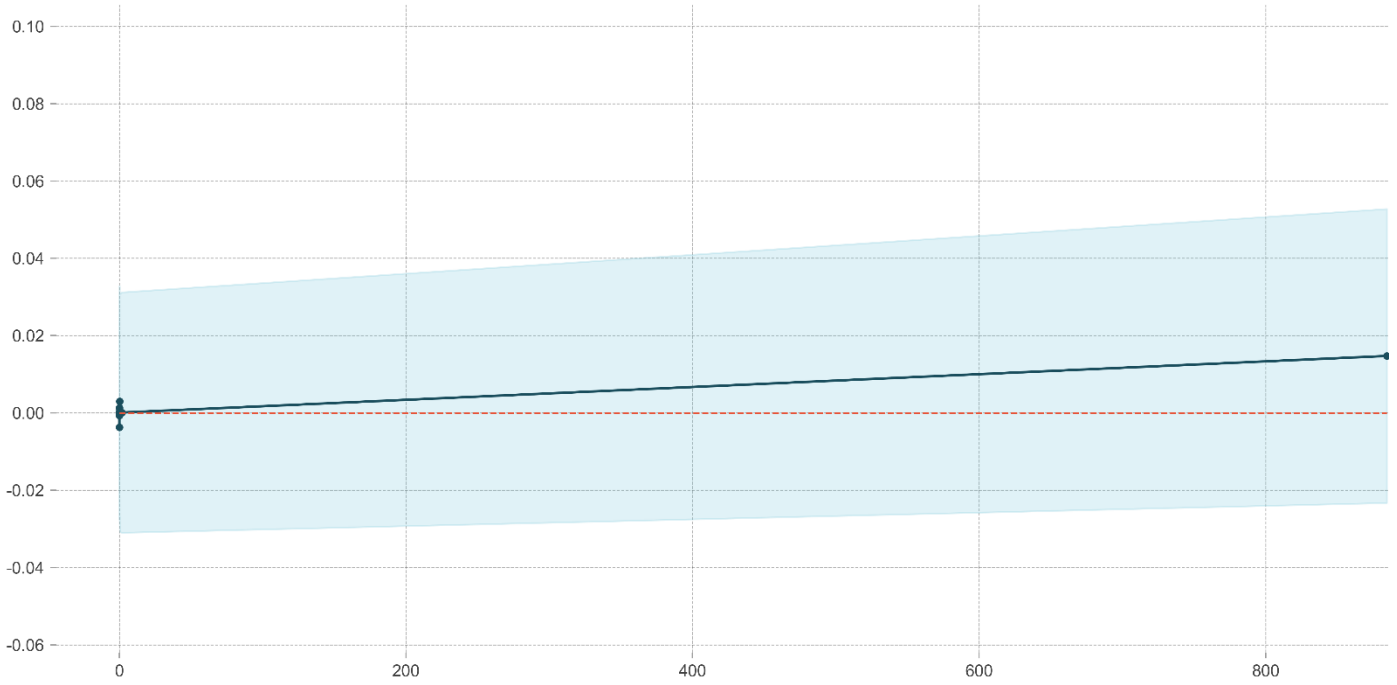
4주차 아데나 변동량 합은 생존 기간에 지속적 영향을 보임



05. 이탈 징후 분석 및 결론

Model 1: 생존 기간(Survival Time)

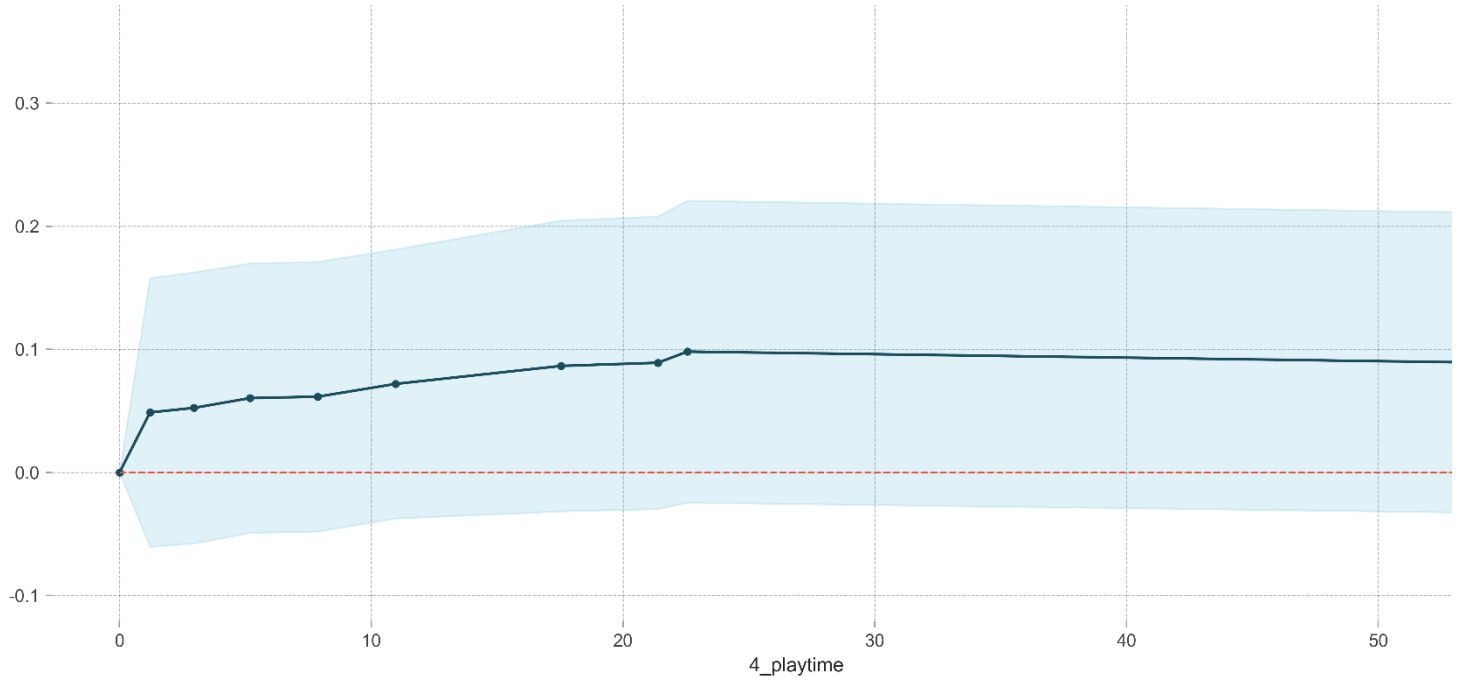
2주차 아데다 변동량 합 이동평균은 생존 기간에 지속적 영향을 보임



05. 이탈 징후 분석 및 결론

Model 1: 생존 기간(Survival Time)

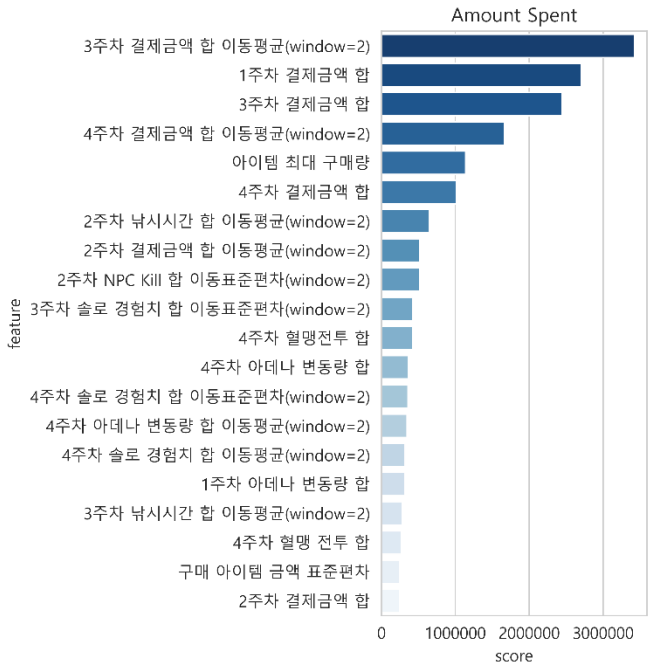
4주차 플레이 시간 합은 생존 기간에 지속적 영향을 보임



05. 이탈 징후 분석 및 결론

Model 2: 일 평균 결제 금액(Amount Spent)

가장 영향력있는 변수는 **결제금액의 합**



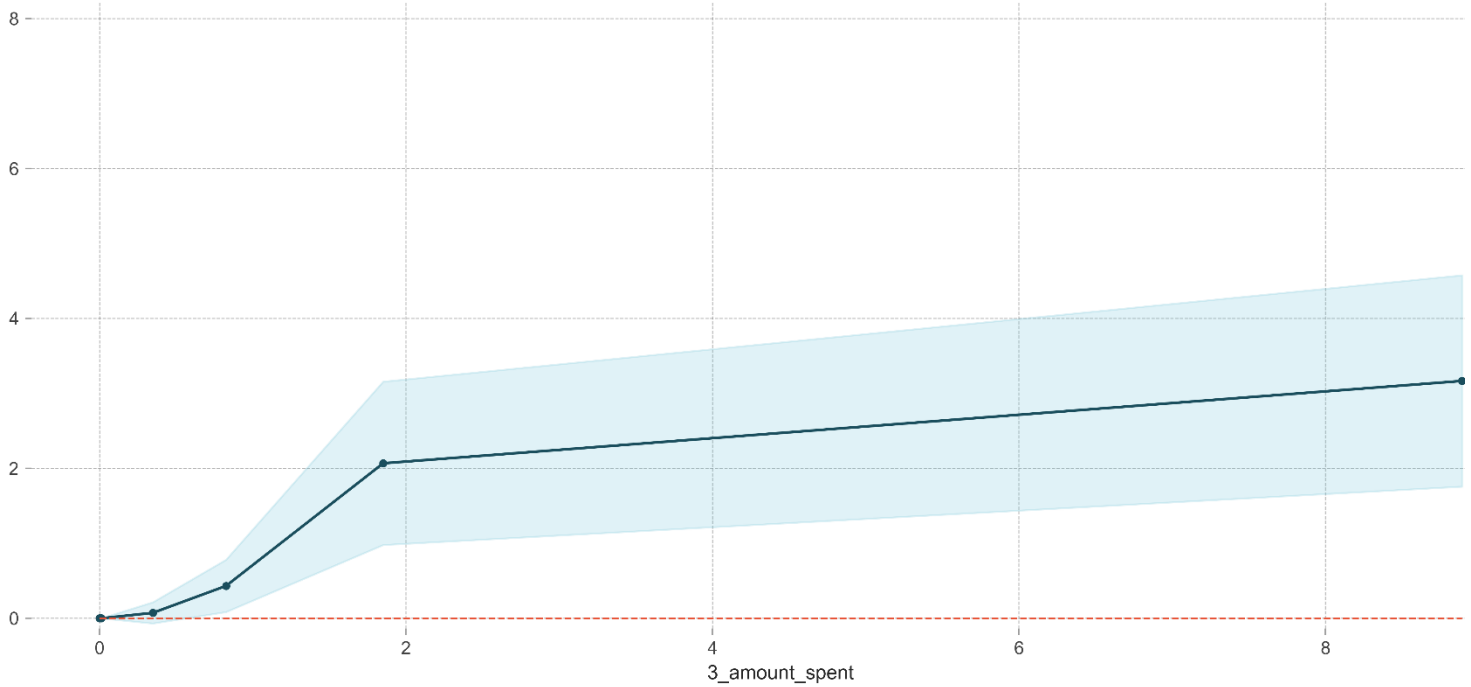
변수 중요도

SHAP

05. 이탈 징후 분석 및 결론

Model 2: 일 평균 결제 금액(Amount Spent)

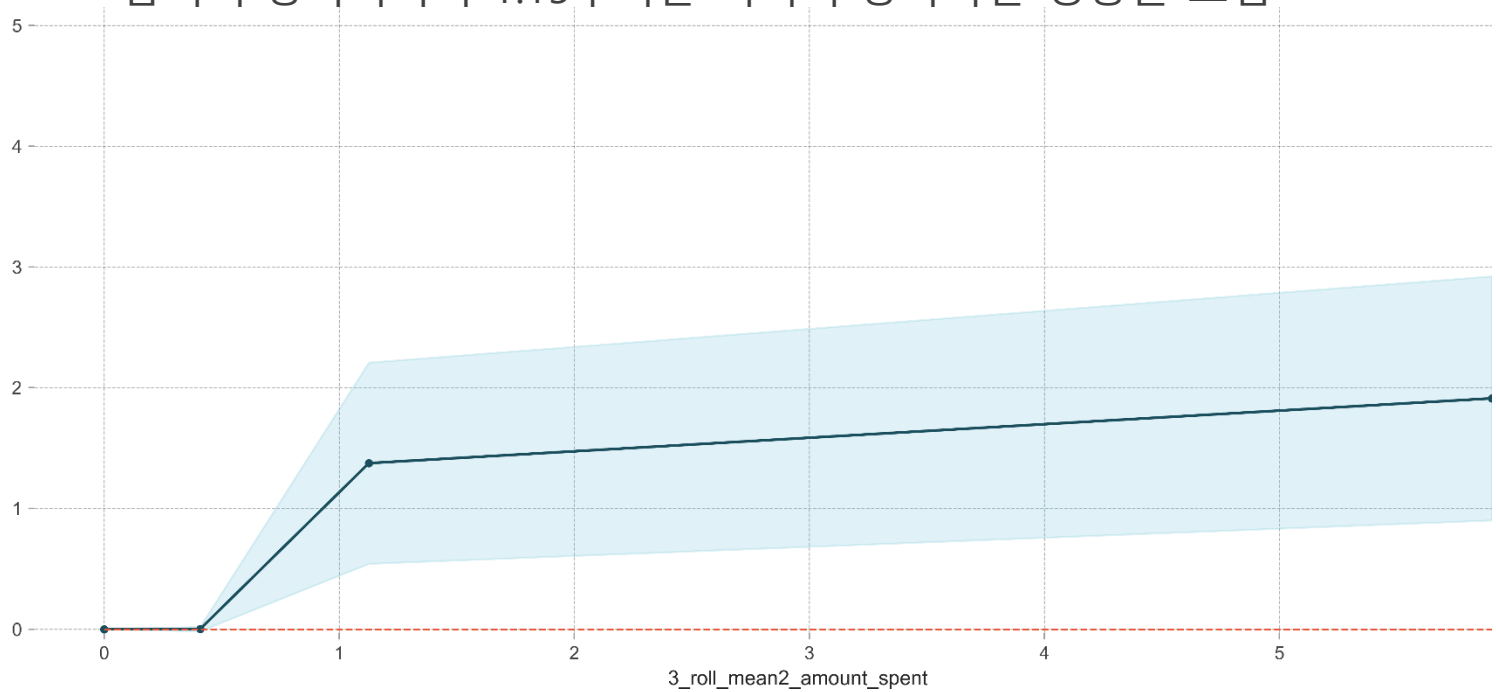
일 평균 결제 금액은 3주차 결제금액 합 이동평균이 0.4부터 1.85 까지
급격히 증가하다가 1.85부터는 서서히 증가시키는 경향을 보임



05. 이탈 징후 분석 및 결론

Model 2: 일 평균 결제 금액(Amount Spent)

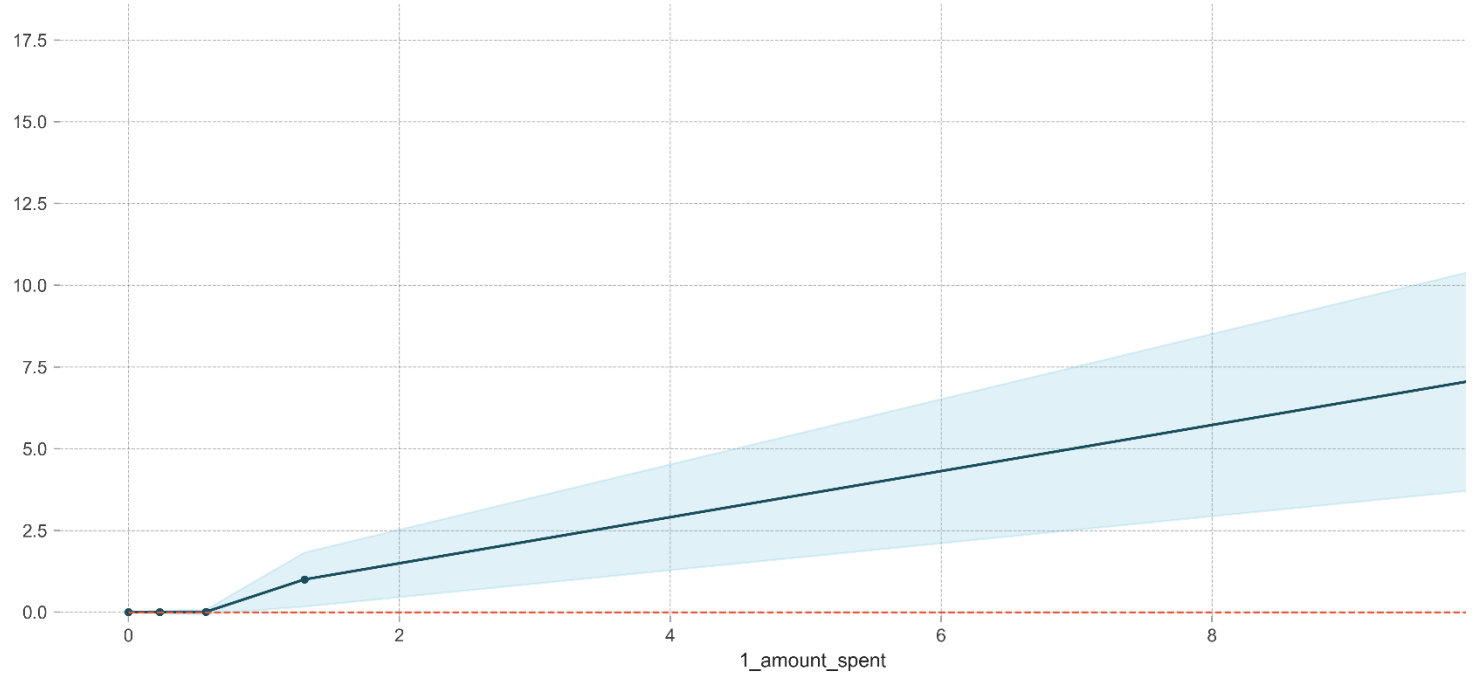
일 평균 결제 금액은 3주차 결제금액 합 이동평균이 0.4부터 1.15까지
급격히 증가하다가 1.15부터는 서서히 증가하는 경향을 보임



05. 이탈 징후 분석 및 결론

Model 2: 일 평균 결제 금액(Amount Spent)

일 평균 결제 금액은 1주차 결제금액 합이 0.8부터 증가하는 경향을 보임



05. 이탈 징후 분석 및 결론

Model 1과 Model 2 해석

변수 중요도가 높은 변수들을 PDP를 통해 확인해보니
목표변수에 영향을 미치는 변수들이라는 것을 확인했음.

Validation Score와 Test1,2 Public Score

Validation	Test 1	Test 2
11431.92	11530.2	3549.12

05. 이탈 징후 분석 및 결론

EDA와 모델 해석 결과를 통해 발견한 이탈 징후

플레이시간이 급격히 감소하는 경우

아데나가 큰 폭으로 변동하는 경우

결제 금액의 합이 급격히 변동하는 경우

05. 이탈 징후 분석 및 결론

결론

실제 기업에서 활용될 수 있는 분석 자료로 구성되기 위해
전처리부터 모델 평가까지 Light하게 구성하여 15분만에
예측 결과를 낼 수 있으며 리더보드 기준상 예측 성능이 높았다