팀 쥬혁이

모델 개발 보고서

목차

1. Intro

2. 주요 문제 및 핵심 전략

3. 주요 데이터 분석

4. 피처 엔지니어링

5. AI 모델

6. AI 모델 학습 전략

7. AI 모델 검증 전략

8. 적용 가능성

9. 피드백

1. Intro

정확도가 높다고? 실전에도 진짜 높을까?

저의 대답은 **'NO'**입니다.

AI 모델러들은 종종 개발 단계에서 높은 정확도를 기록하면, 실제 서비스 환경에서도 모델이 잘 작동할 것이라 믿습니다. 특히 정확도는 여러 AI 모델을 앙상블만 하여도 쉽게 높아질 수 있기 때문에, 실제 서비스 성과와 직접 연결된다고 착각하기 쉽습니다.

특히 광고 클릭 예측 분야에서는 이러한 우려가 더욱 두드러집니다.

이상적인 모델 평가는 인위적 개입 없이, 실제 사용자 행동과 환경을 자연스럽게 반영한 데이터 분포에서 이뤄져야 합니다. 하지만 토스는 자체 광고 최적화 시스템의 영향으로, '잘 클릭 될' 광고 위주의 노출, 토스 통계를 기반으로 한 광고주의 전략적 소재 선택 등 데이터 자체가 이미 토스에 편향(bias)되어 있습니다. 이로 인해 오프라인 정확도가 실제 서비스에선 과대평가(overestimation)될 수 있고, 현실에서는 예측한 만큼의 클릭이 나타나지 않을 수 있습니다

예를 들어, 테스트 데이터의 'day_of_week' 피처가 7 인 샘플들로만 존재하기 때문에, 해당 특성에 맞춰 모델을 튜닝할 경우 리더보드 상에서는 성능이 높게 나오지만, 실제 서비스에서는 다양한 요일이 존재하기 때문에 오히려 과적합(Overfitting) 및 성능 저하로 이어질 수 있습니다.

이러한 문제는 단순히 모델의 성능 저하에 그치지 않고, 실제 광고 비즈니스의 성장과 전략에도 영향을 미칠 수 있습니다. 오프라인 지표만을 근거로 성공을 판단하는 접근은, 실제 환경에서는 데이터의 대표성, 다양성, 확장성 측면에서 한계가 발생할 수 있습니다. 특히, 모델이 기존 인기 광고나 자주 노출된 소재에만 성과를 집중하게 되어, 신규 광고나 다양한 광고주에게 충분한 성장 기회가 제공되지 않을 위험도 존재합니다.

따라서 제공된 데이터로 AI 를 모델링할 때 특히 두 가지를 고려해야 합니다.

1. 신규 광고/지면(콜드스타트)에서 예측력 급락
2. 노출되는 광고만 계속 노출되는 쏠림(rich-get-richer) 현상

이러한 상황은 단순한 기술적 문제가 아닌, 광고주층의 다양성 약화, 신규/소형 광고주의 서비스 이탈, 그리고 전체 플랫폼의 경쟁력 및 성장성 저하로 이어질 수 있는 비즈니스 리스크로 직결됩니다.

주식의 백 테스트(back-test)를 떠올리면 이해하기 쉽습니다. 많은 사람이 주가 예측 모델을 만들고, 백 테스트로 최고의 정확도를 자랑하지만, 실제 시장에서는 언제나 새로운 변수와 상황이 등장해 모델이 기대만큼 역할을 하지 못합니다. 광고 클릭 예측도 마찬가지입니다. 새로운 유저, 광고, UI 변화 등 현실은 끊임없이 바뀝니다.

특히 성장을 추구하는 토스라면, "정확도 높은 모델 = 잘 동작하는 모델"이라는 등식을 그대로 받아들여서는 안 됩니다. 실제로 성과를 내기 위해서는 오프라인 지표의 함정에서 벗어나, 데이터 편향을 인지하고 신규 광고와 지면에도 정확하게 예측할 수 있는 모델링을 고민해야 합니다.

이 보고서에서는 위 고민에 대한 해답으로, 오프라인 지표의 함정에서 벗어나 실제 광고 플랫폼 환경에 최적화된, '최고의 정확도와 강건함'을 동시에 갖춘 AI 모델 설계 전략과 구체적인 방법론을 다룹니다.

2. 주요 문제 및 핵심 전략

따라서 이번 테스크에서의 가장 큰 핵심은 "편향된 데이터로부터 실제 환경에서도 강건한 모델을 만드는 것"입니다.

이를 통해 일반적인 AI 모델링으로 발생하는 두 가지의 큰 문제를 해결해야합니다.

1. 신규 광고/지면(콜드스타트)에서의 예측력 급락

- 기존에 등장하지 않았던 신규 광고, 새로운 지면에서는 모델의 예측 성능이 급격히 저하될 수 있습니다.

- 이로 인해 새로운 광고의 노출 기회가 충분히 보장되지 않고, 기존 인기 소재 위주로만 노출이 지속되는 구조적 한계가 생길 수 있습니다.

2. 노출되는 광고만 계속 노출되는 쏠림(rich-get-richer) 현상

- 데이터가 '클릭률이 높은 광고' 위주로 더 많이 수집되고, 모델 또한 이러한 분포에 맞춰 학습됨에 따라 이미 잘나가는 광고에만 성과가 집중되는 경향이 심화될 수 있습니다.

- 반대로, 신규 광고주나 소형 광고 등 데이터가 충분히 확보되지 않은 집단에서는 예측 정확도 및 활용도가 떨어지는 대표성의 한계가 존재합니다.

이는 결과적으로,

1. 광고주 층의 다양성 저하,

2. 신규 및 소형 광고주의 서비스 이탈,

3. 전체 플랫폼의 성장성과 경쟁력 저하

등의 비즈니스 리스크로 이어질 수 있습니다.

본 프로젝트에서는 위 두 가지 구조적 문제를 해결하기 위해

1. Custom 손실함수 설계

2. Custom Masking validation 방법론

등을 개발하여 차별화된 광고 도메인 특화 방법론을 연구 개발했습니다.

3. 주요 데이터 분석

제공된 데이터는

gender : 성별

age_group : 연령 그룹

inventory_id : 지면 ID

day_of_week : 주번호

hour : 시간

seq : 유저 서버 로그 시퀀스

l_feat_* : 속성 정보 피처 (l_feat_14 는 Ads set)

feat_e_* : 정보영역 e 피처

feat_d_* : 정보영역 d 피처

feat_c_* : 정보영역 c 피처

feat_b_* : 정보영역 b 피처

feat_a_* : 정보영역 a 피처

history_a_* : 과거 인기도 피처

clicked : 클릭 여부 (Label)

로 피처의 상세 의미는 의도적으로 공개되지 않은 비식별화된 피처입니다.
데이터 분석은 해당 도메인의 정보를 함께 녹여야 의미있는 분석이 됩니다. 예를 들어, 튀는 값은 어떤 도메인에서 이상치가 되지만 어떤 도메인에서는 특이치가 됩니다. 따라서 정확한 데이터 분석을 위해서는 비식별화된 피처들의 의미를 파악해야합니다.

l_feat_14 는 Ads set 라는 정보가 주어졌습니다. 따라서 이를 이용해서 l_feat_*내 다른 비식별화된 피처들의 정보도 유추가 가능할 것이라 생각했습니다.

이후 'toss ads 광고주 가이드'에서 추가적인 정보를 얻었습니다.

(toss ads 광고주 가이드)

해당 이미지를 보게 되면 "캠페인 > 광고 세트 > 소재 단위"의 정보가 있습니다. 즉, 광고 세트는 l_feat_14 인 Ads set 이며 Ads set 의 상위 개념은 캠페인이며 하위 개념은 소재 단위입니다.

또한 "캠페인당 광고 세트 100 개, 광고 세트당 소재 100 개 까지 생성 가능해요"라는 정보를 발견하여(https://toss-ads.gitbook.io/guide/ad/banner/creative) 이를 통해 광고 세트 피처인 l_feat_14 를 기반으로 캠페인과 소재 단위 피처 식별화 작업을 시작했습니다.

[소재 단위]

소재 단위도 마찬가지로 광고 세트당 소재가 100 개가 최대이기 때문에, 광고 세트의 종류 이하이면서 각 광고 세트당 소재 단위가 100 개 이하인 조건으로 찾아보았습니다. 또한 'l_feat_12'의 유니크 값을 2 개 이상의 'l_feat_14' 유니크 값을 가지고 있지 않았기 때문에 'l_feat_12'가 유일하게 해당 조건을 만족하였고 해당 피처를 **소재 단위** 피처로 식별하였습니다.

[inventory_id]

inventory_id(지면)별 clicked 을 분석한 결과, inventory_id 에 따라 clicked 편차가 매우 크게 나타남을 확인할 수 있었습니다. 예를 들어, 트래픽이 많은 주요 지면들 간에도 CTR 이 0.8%에서 3.8% 이상까지 차이가 났습니다. 이는 광고 클릭 여부에 있어 inventory_id 가 CTR 의 핵심 결정 변수로 작용함을 크게 시사합니다. 이러한 분석 결과를 기반으로, inventory_id 별 Segment 를 구성해 inventory_id 별 모델링을 진행한다면 다음과 같은 장점을 가질 수 있을 것이라 생각했습니다.

1. 고정된 슬롯에 대해 매우 높은 예측 성능 기대

 - 지면에 따라 클릭 확률이 명확히 다르기 때문에, 각 inventory_id 에 맞춘 모델은 높은 정확도 확보 가능

2. 베이스 CTR 이 강한 prior 역할을 함

 - 지면 자체가 클릭 성향을 크게 결정하므로, 학습에 있어 안정적인 기반 정보로 활용 가능

그러나 이러한 방식은 실제 운영에 적용하기에는 한계점이 존재합니다.

특히, 서비스 UI 는 지속적으로 개편되고, 신규 inventory_id 가 등장하기 때문에, 기존에 존재하지 않았던 inventory_id 에 대해서는 해당 모델이 적절한 예측을 수행하기 어렵습니다. 즉, 현재는 정확한 모델을 만들 수 있지만, UI 가 변화되는 미래에는 새로운 지면(Unseen inventory_id)에 대한 일반화 성능이 취약하고, 유지보수 또한 새롭게 다시 진행해야합니다. 이러한 이유로, 본 분석 결과에도 불구하고 inventory_id 별 모델링은 채택하지 않기로 결정하였습니다.

4. 피처 엔지니어링

앞서 진행한 데이터 분석에서 얻은 인사이트들을 토대로, CTR 예측이라는 도메인 특성에 적합한 모델을 만들기 위해 다양한 피처 엔지니어링을 적용했습니다.

클릭 정보 피처

먼저 클릭 정보와 관련해서는, inventory_id 별 CTR 의 신뢰성을 높이기 위해 베이지안 추정 기반 통계 피처를 도입했습니다. 광고 클릭 예측에서는 inventory_id 가 중요한 결정 변수이지만, 지면별 데이터 분포는 불균형합니다. 어떤 지면은 데이터가 풍부한 반면, 어떤 지면은 샘플 수가 극히 적어 단순한 평균 기반의 Target Encoding 을 적용할 경우 왜곡이 발생할 수 있습니다. 이를 해결하기 위해, 전체 학습 데이터의 평균 클릭률을 Global Target Encoding 값으로 설정하고, 각 inventory_id 의 실제 clicked 결과를 결합해 Global Target Encoding 에 보정된 Target Encoding 을 진행했습니다. 이와 함께, 피처의 불확실성을 나타내는 분산과 엔트로피 값도 함께 추출해 모델이 예측의 신뢰도를 고려할 수 있도록 했습니다. 또한, age_group, gender, hour 와 같이 각 범주별로 샘플 수가 일정한 피처들에 대해서는 CTR 통계를 조합해 반영하는 방식으로 Target Encoding 을 적용하여, 보다 세분화된 조건부 클릭 성향을 학습할 수 있도록 했습니다.

유저 특성 피처

유저 특성과 관련해서는, 유저의 행동 이력을 나타내는 시퀀스(seq) 길이를 통해 활성도를 반영하였고, 시간에 따른 중요도를 고려해 가장 최근의 로그를 별도 피처로 추출했습니다. 또한 '20 대 초반 남성'과 '20 대 후반 남성'처럼 실제로는 유사한 행동을 보이지만 데이터상 유니크하게 분리되는 그룹들을 일반화하기 위해, 나이대 및 성별 기반의 그룹핑 피처를 추가하여 비슷한 특성 간의 패턴 학습 가능성을 높였습니다.

시간 정보 피처

시간 정보(hour, day_of_week)의 경우, 단순 수치로 사용할 경우 0 시와 23 시가 멀리 떨어진 것처럼 해석되는 문제가 있습니다. 하지만 시간은 본질적으로 순환적인 특성을 갖기 때문에, 이를 수치적으로 표현하기 위해 sin/cos 변환을 적용하여 AI 가 시간대 간의 연속성과 주기성을 자연스럽게 인식할 수 있도록 했습니다. 이를 통해 시간적 맥락에서의 클릭 패턴을 더 정교하게 반영했습니다.

5. AI 모델

이번 프로젝트에서는 다양한 AI 모델을 앙상블하는 방식 대신, LightGBM(LGBM) 단일 종류의 모델만을 사용하는 전략을 택했습니다. 이 같은 선택은 단순히 구현 편의성 때문이 아니라, 실제 서비스 환경에서의 실전 적합성, 운영 안정성, 예측 속도 등을 종합적으로 고려한 결과입니다.

첫째, 이번 프로젝트의 데이터셋이 지닌 편향 및 콜드스타트 리스크를 고려했을 때, 오프라인 지표만 무작정 끌어올리는 것보다 실제 현업 환경에서의 강건함과 실전 적합성이 더 중요하다고 판단했습니다. 다양한 모델을 앙상블해서 극한의 정확도를 노리는 것이 단기적으로는 좋아 보일 수 있지만, 본 데이터셋처럼 플랫폼 자체의 편향이나 신규 광고·지면에서의 콜드스타트 문제가 클 경우, 오히려 실서빙에서의 일반화나 공정성 측면에서는 불리할 수 있습니다. 특히 이번 프로젝트에서 Robustness 를 중시했기 때문에, 복잡한 앙상블보다는 LGBM 단일 모델 중심의 경량, 강건 전략이 더 적합하다고 판단했습니다.

둘째, 실시간 서빙 환경에서 예측 속도와 안정성이 매우 중요했기 때문입니다. 서로 다른 종류의 모델을 조합하면 이론적으로 오프라인 예측 정확도가 크게 향상될 수 있지만, 각 모델의 피처 파이프라인과 추론 주기가 달라 실제 서빙 단계에서 병목이나 지연이 발생하기 쉽습니다. 특히 본 과제의 핵심 목표 중 하나가 "매우 빠르게 예측"하는 것이었던 만큼, LGBM 단일 모델만 사용하면 일관된 입력 구조와 추론 경로를 가질 수 있어, 빠른 응답 속도를 안정적으로 보장할 수 있습니다.

셋째, 운영 환경의 단순화와 장기적인 유지보수 관점에서도 LGBM 단일 모델은 유리합니다. 만약 서로 다른 프레임워크와 라이브러리를 사용하는 여러 종류의 모델을 조합한다면, 당장은 동작하더라도 추후 라이브러리 업데이트, 환경 변경

시 라이브러리 간 충돌이 발생할 확률이 높아집니다. 반면 LGBM 단일 모델 체계에서는 운영 환경과 의존성이 단순해지고, 롤백이나 재학습, 재배포, 버전 관리도 훨씬 용이해집니다.

이처럼 이번 프로젝트에서는 실시간 예측 성능, 운영 안정성, 현업 적합성이라는 세 가지 관점에서 단일 종류의 LGBM 모델을 선택했습니다.

6. AI 모델 학습 전략

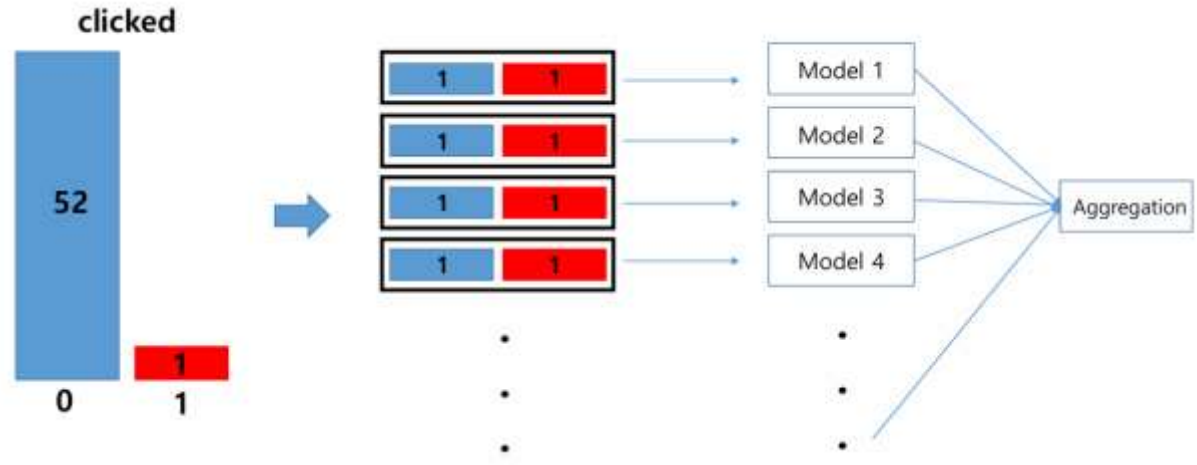
5-1. UnderSampling Based Bagging Modeling

클릭 여부를 예측하는 AI 모델을 개발하는 과정에서 데이터의 불균형 문제를 해결해야 했습니다. 학습 데이터의 경우 클릭이 발생한 양성(positive) 샘플 대비 클릭이 발생하지 않은 음성(negative) 샘플이 약 52:1 에 달할 정도로 심각하게 불균형한 분포를 보였습니다.

일반적으로는 LGBM 의 scale_pos_weight 파라미터를 조정해 클래스 불균형을 보정하는 방법이 자주 사용됩니다. 그러나 본 프로젝트의 도메인 특성과 실험 결과를 함께 고려했을 때, 단순 가중치 조정 방식으로는 한계가 존재했습니다.

광고 클릭 예측 문제는 음성 클래스 내에 다양한 행동 패턴이 존재합니다. 하지만 복잡한 음성 집단을 한 번에 학습하게 되면, 모델은 클릭(1)에 해당하는 희귀한 패턴을 충분히 학습하지 못하고 보수적인 의사결정 경계를 형성하게 됩니다. 또한 scale_pos_weight 는 전체 클래스 간 비율만 보정할 뿐, 음성 클래스 내부의 다양성을 반영하지 못해 일부 패턴에 과도하게 편향된 예측 결과가 도출될 수 있습니다.

이러한 문제를 해결하기 위해, 본 프로젝트에서는 UnderSampling 기반 Bagging 학습 전략을 연구 개발했습니다. 전체 양성 샘플과 동일한 개수의 음성 샘플을 반복적으로 무작위로 추출하여, 1:1 로 균형잡힌 서브셋을 다수 생성한 뒤, 각각을 bagging 하는 방법입니다. 전체 음성 샘플을 소진할 때까지 이 과정을 반복하며, 최종 확률을 산출하는 Under Sampling + Bagging 구조입니다.



<0 샘플을 전체 1 샘플의 개수에 맞게 UnderSampling 후 Bagging>

이러한 접근은 다음과 같은 장점이 있습니다.

1. 의사결정 경계의 다양성 확보

반복마다 서로 다른 음성 샘플을 사용함으로써, 모델은 음성 클래스 내 다양한 행동 패턴을 접할 수 있습니다. 이를 통해 특정 음성군에 과도하게 편향되는 현상을 줄이고, 클릭(1) 클래스의 특징적인 경계를 보다 명확하게 학습할 수 있습니다. 단일 학습으로 고정된 경계가 아니라, 반복 학습을 통해 다양한 시각에서 형성된 의사결정 경계들이 결합되며, 보다 유연하고 강건한 판단 기준이 만들어집니다.

2. 음성 클래스의 다양성 반영

1:1 로 밸런싱된 서브셋을 반복적으로 학습하면, 전체 음성 클래스의 다양한 하위 집단을 더 고르게 반영할 수 있습니다. 전체 데이터를 한 번에 학습할 경우 자주 등장하는 일반적인 패턴 위주로 학습이 이뤄질 가능성이 크지만, 반복적인 샘플링을 통해 클릭이 발생하지 않는 다양한 시나리오까지 포괄할 수 있게 됩니다.

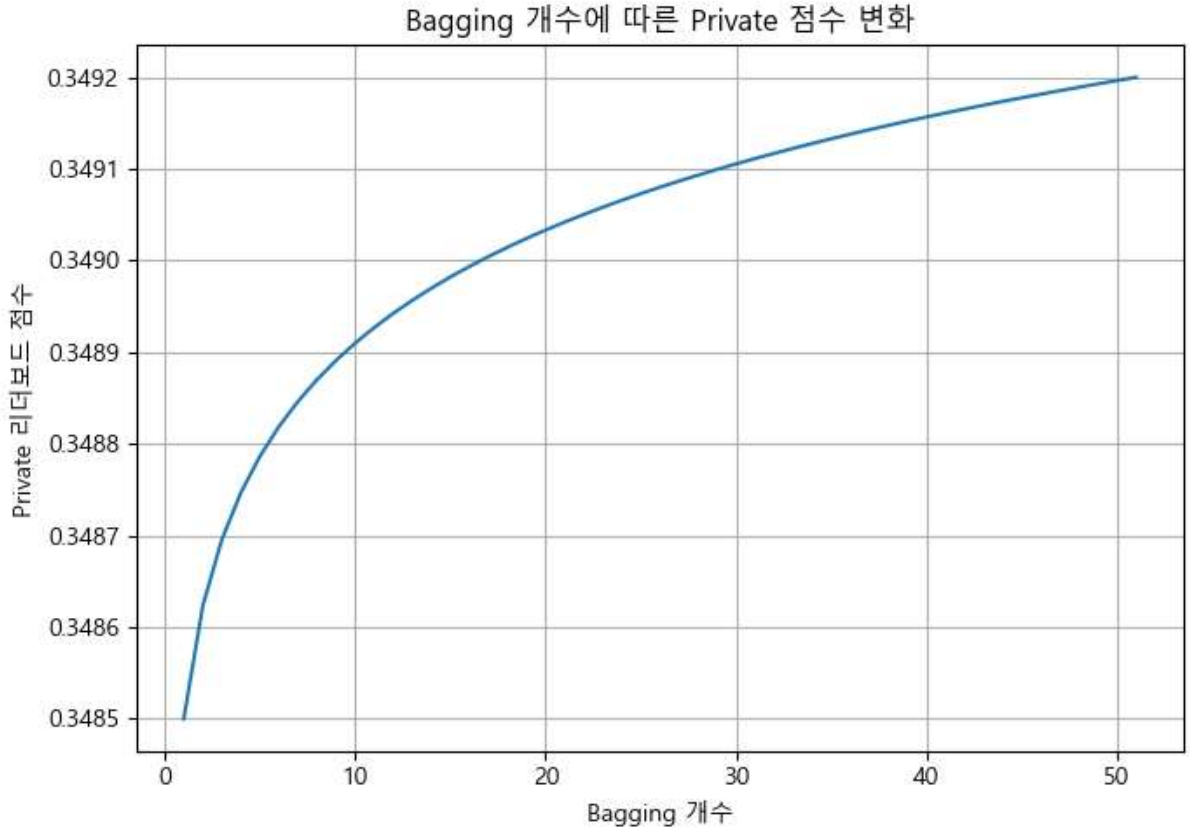
3. 불균형 환경에서의 민감도 개선

클릭(1) 클래스가 극히 희소한 상황에서는, 단일 모델이 이 신호를 효과적으로 학습하기 어렵습니다. 그러나 반복된 언더샘플링을 통해 항상 1:1 로 구성된 학습 데이터를 제공하면, 모델은 클릭 가능성이 있는 행동 패턴에 더 민감하게 반응할 수 있습니다. 이는 민감도(Recall)를 향상시키고, 희귀한 양성 신호에 대한 감지력을 높이는 데 기여합니다.

4. 실전 환경에 대한 일반화 가능성

편향된 경계에 고정되지 않기 때문에, 예측 대상 분포가 일부 달라져도 더 안정적인 성능을 유지할 가능성이 큽니다. 이는 실전 환경에서 새로운 유저 유형이나 신규 광고가 등장했을 때 성능 급락을 완화하는 데 유리합니다.

결론적으로, 단순 가중치 조정만으로는 희귀한 클릭 패턴 포착과 음성 클래스의 다양성 반영을 동시에 만족시키기 어렵습니다. 반면, UnderSampling Based Bagging Modeling 전략은 모델의 일반화 성능을 높이고, 실전 환경에서도 신뢰성 있는 예측을 가능하게 해주는 실질적인 방법론입니다.



이 전략은 음성 데이터를 모두 소진할 때까지 반복되며, 결과적으로 음성 샘플 수를 양성 샘플 수로 나눈 만큼의 모델이 생성됩니다. 다만 실험 결과, 단 1 회 학습한 모델만으로도 private score 기준에서 높은 정확도를 기록했으며, Bagging 을 늘릴수록 성능이 점진적으로 향상되는 경향을 보였습니다. 또한 최종 코드에서는 5-Fold CV 까지 사용하였지만, Fold 는 전체 성능 향상에 도움이 되지 않아서 CV 는 하지 않아도 될 것 같습니다.

5-2. Custom Loss Function

이번 프로젝트에서 가장 고민이 컸던 부분 중 하나는 데이터가 구조적으로 편향돼 있다는 점이었습니다. 실제 서비스 환경에서는 광고 예산이 크거나 성과가 좋은 인기 광고 소재에 노출과 클릭이 집중되며, 반면 신규 광고주나 소형 광고주의 소재는 상대적으로 데이터가 부족합니다. 이처럼 불균형하게 수집된 데이터를 기반으로 일반적인 이진 분류 손실 함수(binary cross-entropy, logloss)를 그대로 적용하면, 모델은 자연스럽게 더 많이 데이터가 쌓인 광고

위주의 예측 성능을 끌어올리는 방향으로만 최적화됩니다. 결과적으로 "잘 노출되는 광고는 더 잘 예측하고, 그렇지 못한 광고는 무시되는 "쏠림 현상(rich-get-richer)"이 더욱 심화됩니다.

이러한 현상이 지속되면, 플랫폼 내 일부 광고만 노출 및 클릭 기회를 독점하게 되고, 신규 광고주나 다양한 소재는 예측력이 확보되지 않아 광고 생태계의 다양성과 공정성이 무너질 수밖에 없습니다. 이 문제를 해결하기 위해, 저는 학습 과정에서 각 샘플에 가중치를 차등 부여하는 커스텀 손실 함수를 설계했습니다. 구체적으로는, 'l_feat_14'로 주어진 Ads set(광고 세트)의 데이터 등장 빈도를 기준으로, 해당 광고가 많이 등장할수록 낮은 가중치를, 적게 등장할수록 높은 가중치를 부여하는 방식입니다. 이렇게 샘플별로 손실 함수의 기여도를 조정함으로써, 모델이 인기 소재뿐 아니라 데이터가 적은 비인기 또는 신규 소재에도 더 민감하게 반응하도록 유도했습니다.

결과적으로 이 Custom Loss 는 광고 데이터에 내재된 구조적 편향과 예측의 불공정성을 완화하고, 소외된 광고주와 소재도 의미 있는 예측 성능을 확보할 수 있도록 돕는 도메인 특화 방법론입니다. 단순히 정확도를 높이는 데 그치지 않고, 광고주 간 기회의 균형을 맞추고 플랫폼 생태계의 건강한 다양성을 유지하는 것이 이 설계의 궁극적인 목적이었습니다.

'데이터 쏠림'이라는 현실적인 한계를 기술적으로 보완하고, 실제 서비스에서 보다 공정하고 실질적인 예측 성과를 내기 위한 핵심 역할입니다.

7. AI 모델 검증 전략

7-1. Masking Validation Strategy

실제 광고 서비스 환경에서는 신규 광고(l_feat_14, l_feat_12)나 신규 지면(inventory_id)이 지속적으로 등장합니다. 이로 인해, 과거에 한 번도 등장하지 않았던 광고나 지면에 대해서는 모델의 예측 성능이 현저히 떨어지는, 이른바 콜드스타트(cold-start) 문제가 발생합니다. 만약 모델이 과거 데이터에만 과도하게 최적화되어 있다면, 예측 성능은 기존 인기 광고에만 집중되고, 신규 광고는 노출 기회를 얻지 못하는 악순환이 반복됩니다.

Original Validation set

l_feat_12	l_feat_14	Inventory_id	clicked
3.0	2924.0	2	0
7.0	2924.0	36	1
3991.0	2052.0	37	0
3991.0	2453.0	21	0

Masking Validation set

l_feat_12	l_feat_14	Inventory_id	clicked
3.0	2924.0	2	0
-111111	-111111	-111111	1
3991.0	2052.0	-111111	0
-111111	-111111	21	0

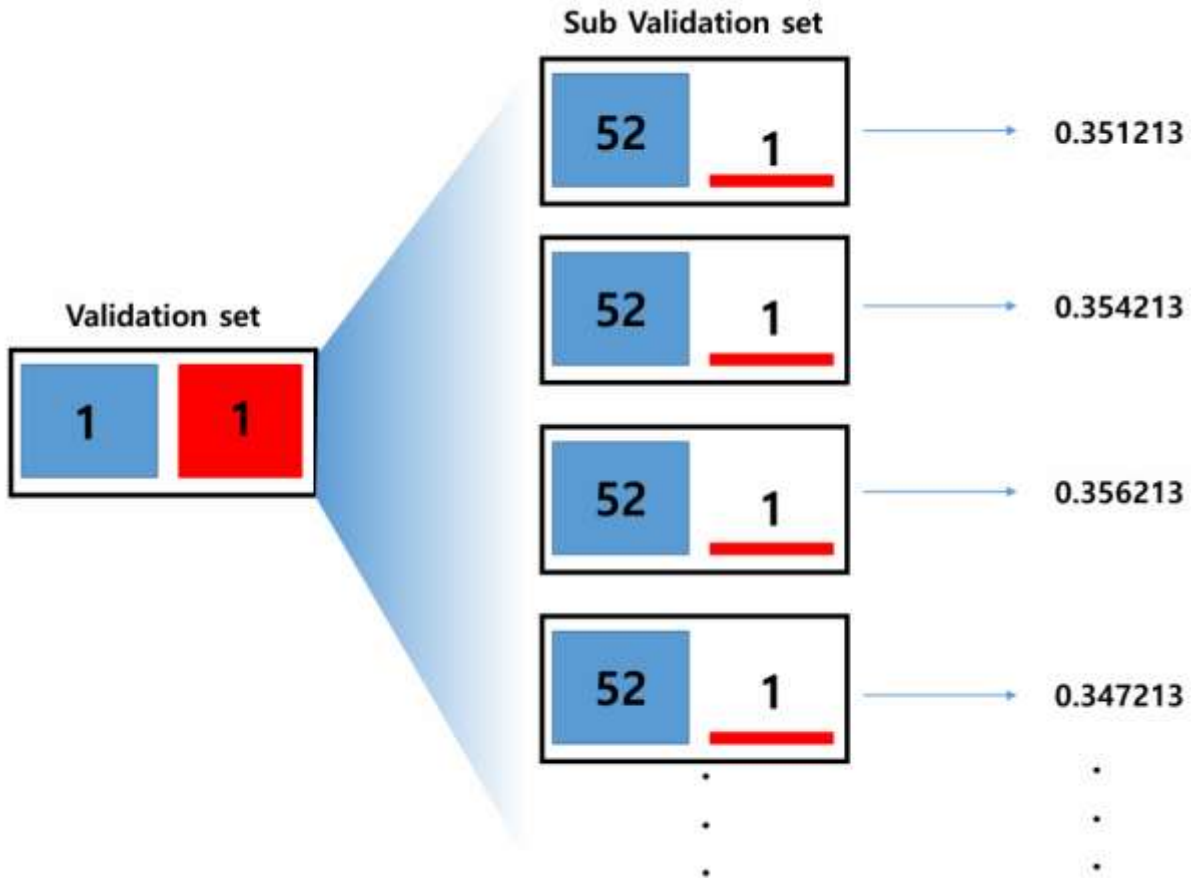
이러한 문제를 검증 단계에서부터 반영하기 위해, Masking Validation 전략을 도입했습니다. 구체적으로는 검증 데이터에서 일부 무작위 샘플의 l_feat_14(Ads set), l_feat_12(소재 단위) 또는 inventory_id(지면) 피처 값을 특수 마스킹 값(-111111)으로 치환하여, 모델이 "새로운 Ads set 와 소재 단위"에서도 일반화된 예측을 할 수 있도록 유도했습니다. 즉, 훈련 과정에서는 해당 정보를 활용하되, 검증(Validation)에서는 해당 정보를 인위적으로 제거(masking)하여, 신규 광고에 대한 모델의 예측 성능 저하(콜드스타트 리스크)를 사전에 평가하고, 더 강건한 일반화 성능을 가진 모델을 만들고자 했습니다.

7-2. Real-world Based Validation Strategy

모델 학습 과정에서 clicked 의 0 과 1 비율을 1:1 로 맞춰주는 언더샘플링 작업을 진행했습니다. 그런데 이 상태에서 그대로 validation set 을 구성하면 일반적으로

validation set 또한 clicked 의 0 과 1 비율이 1:1 로 들어가게 됩니다. 하지만 실제 전체 데이터에서의 비율은 0:1 이 약 52:1 에 이를 정도로 클릭(1)은 극히 드문 이벤트입니다. 즉, 자연 상태의 데이터 분포, 다시 말해 실제 광고 클릭 예측이라는 특성을 제대로 반영하려면 0:1 의 불균형(약 52:1)이 유지된 상태에서 모델의 성능을 평가해야 합니다.

만약 언더샘플링된 1:1 비율의 검증셋을 그대로 사용해 모델을 평가할 경우, 예측 성능이 과대평가되는 현상이 발생할 수 있습니다. 이는 특히, 대다수 샘플이 음성(비클릭)인 '극단적 불균형' 상황에서 실제 서비스 적용 시의 마주하게 될 극단적 불균형 상황에서의 성능 저하를 충분히 감지하지 못하게 만듭니다.



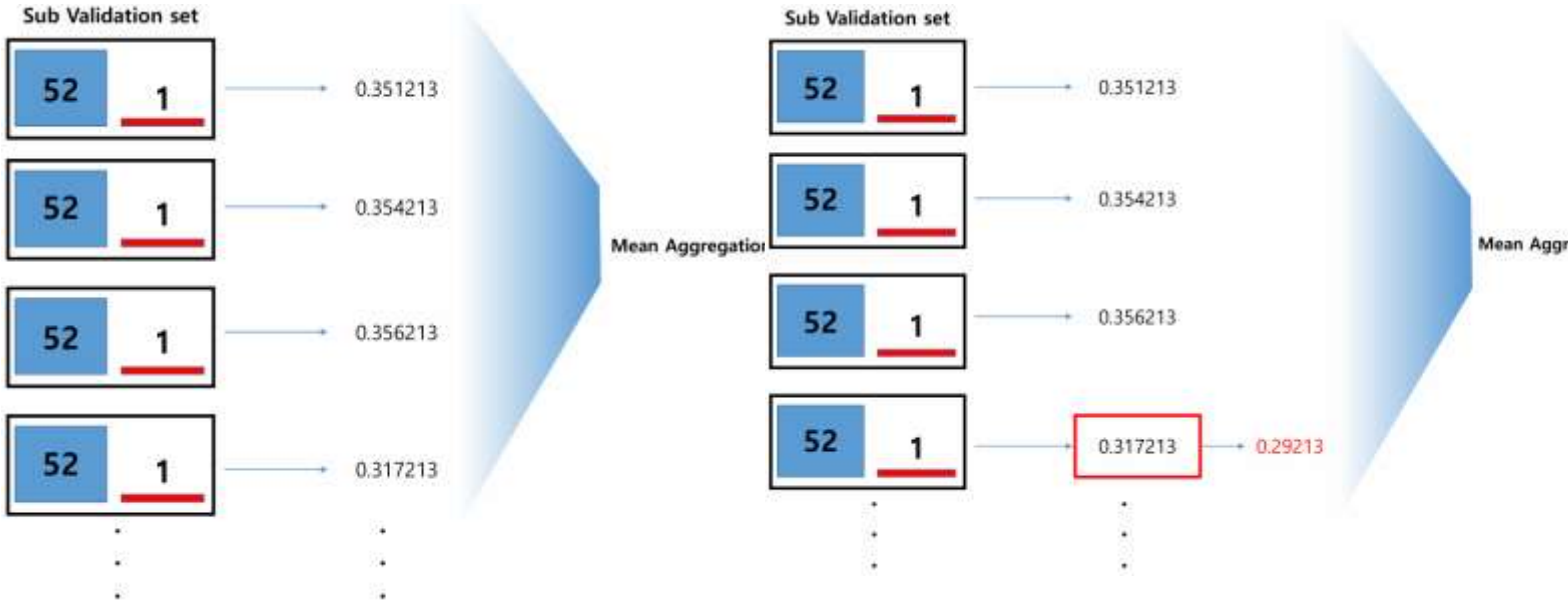
따라서, 본 프로젝트에서는 validation set 역시 실제 데이터의 자연스러운 비율(약 52:1)을 충실히 반영할 수 있도록 별도의 검증셋 로직을 설계하였습니다. 음성(0) 샘플을 모두 포함시키고, 양성(1) 샘플을 실제 분포에 맞춰 랜덤하게 여러 개의 검증셋(valid set)으로 나누는 방식입니다. 각 검증셋마다 음성:양성 샘플의 비율이

전체 데이터의 실제 분포(target_ratio ≒ 0.019, 즉 약 52:1)와 거의 동일하게 맞춰지도록 했습니다.

이렇게 구성된 현실 기반의 검증셋을 통해, 모델의 평가가 보다 실전적인 환경에 가까워졌으며, 언더샘플링 학습에서 발생할 수 있는 평가 왜곡을 효과적으로 방지할 수 있었습니다. 결과적으로, 모델이 실제 운영 환경에서도 안정적이고 신뢰도 높은 성능을 발휘할 수 있는지를 더 효과적으로 검증할 수 있는 기반을 마련했습니다.

7-3. Downside Risk-Evaluation Strategy

7-2 의 방법론으로, 하나의 Validation set 이 여러 개의 Sub Validation 으로 구성된 데이터로 모델의 성능을 평가할 때 모델의 평가 역시 단순한 평균 점수만을 사용하지 않고, 낮은 평가 점수를 고려하는 Robust Aggregation 방식을 적용하였습니다. 평균 점수는 높지만 일부 검증셋에서 심각하게 성능이 저하되는 불안정한 모델이 있다면, 이 모델이 선택되는 것을 방지하고자 하였습니다.



이는 평균 성능이 높더라도, 특정 검증셋에서 유난히 성능이 저하되는 경우에는 의도적으로 점수에 패널티를 부여해, 해당 모델이 선택되지 않도록 하는 전략입니다. 즉, 강건하지 못한 모델이 평균 수치만으로 높은 평가를 받는 것을

방지하고, 다양한 상황에서 안정적으로 성능을 유지할 수 있는 모델을 우선적으로 선택하기 위함입니다.

이러한 평가 전략은 앞서 설명한 Masking Validation 이나 Custom Loss 와 함께 작동하여, 모델이 특정 광고에만 과도하게 성능을 집중하는 것을 막고, 신규 광고나 신규 지면과 같은 실전 환경에서도 정확하고 강건한 예측을 할 수 있도록 했습니다.

8. 적용 가능성

8-1. 확장가능성

이번 토스 NEXT ML CHALLENGE 의 목표는 단순히 지표에서 높은 정확도를 얻는 것이 아니라, 실제 서비스 환경에서도 강건하게 작동하는 모델을 만드는 것이었습니다. 이 같은 방향성 덕분에, 신규 광고나 신규 지면이 등장하는 상황에서도 예측 정확도가 안정적으로 유지되는 구조를 갖출 수 있었습니다.

(a) UnderSampling Based Bagging Modeling 으로 극단적인 클래스 불균형 환경에서도 다양한 음성 패턴을 반영하고, 클릭 신호에 민감한 모델을 구현했습니다.

(b) 빈도 기반 Custom Loss Function 을 통해 특정 광고에만 성능이 집중되는 편향을 완화했고,

(c) Masking Validation 을 적용해 새로운 광고 소재, 지면 상황에서도 일반화 능력을 평가하고 강화했으며,

(d) 실제 데이터 분포(약 52:1) 를 반영한 검증셋 구성으로 현실적인 성능 평가가 가능했고,

(e) Downside Risk-Evaluation Strategy 를 통해 일부 케이스에서 성능이 급락하는 불안정한 모델을 선제적으로 배제할 수 있었습니다.

이러한 전략의 조합은 광고 UI 개편, 신규 inventory_id 등장, 아직 등장하지 않은 광고 세트나 소재의 유입 등, 지속적으로 변화하는 환경에서도 모델의 예측력을 유지할 수 있게 합니다. 결과적으로, 이번 프로젝트는 변화가 많은 실전 비즈니스 환경에서도 높은 적응력과 유지보수 효율성을 함께 갖춘 모델을 구현했다는 점에서, 확장 가능성 측면에서 매우 실용적인 성과를 도출했다고 볼 수 있습니다.

8-2. 유저에 대한 광고 최적화

본 모델은 유저별 광고 노출 최적화에도 효과적으로 활용될 수 있도록 설계되었습니다. 운영 단계에서는 I_feat_14(광고 세트) 또는 I_feat_12(광고 소재)의 모든 유효 조합을 생성한 뒤, 개별 유저의 컨텍스트를 반영해 각 조합별 클릭 확률을 한 번에 배치 추론합니다. 이후 예측된 확률값을 기준으로 광고를 랭킹화하고, 가장 높은 확률을 가진 조합(Top-prob)에 대해 더 높은 입찰가 또는 단가를 설정하여 노출 우선순위를 조정합니다. 데이터 편향을 효과적으로 보정했기 때문에, 인기 광고만 과도하게 밀어주는 현상 없이 신규 소재나 롱테일 광고에도 노출 기회를 제공할 수 있습니다.

결과적으로, 사용자는 더 관련성 높은 광고를 경험하고, 광고주는 단가 대비 성과를 개선할 수 있으며, 플랫폼 입장에서도 클릭 수, 수익, 콘텐츠 다양성을 동시에 추구할 수 있는 전략적 기반이 마련됩니다. 이 모델은 단기 성과뿐만 아니라 장기적 광고 생태계의 효율성과 지속 가능성을 함께 고려한 설계라 할 수 있습니다.

8-3. 예측 시간

실시간 광고 서빙에 필요한 예측 시간을 보다 구체적으로 파악하기 위해, 두 가지 시나리오에 따라 LGBM 모델의 추론 시간을 측정했습니다. 측정은 편차 보정을 위해 총 300 번 수행했습니다.

8-3-1. 유저별 타겟 광고 추론 시간

첫 번째로, 유저 단위 개별 추론 시나리오에서는 한 명의 유저에 대해 AI 예측에 평균 0.003 초가 소요되었습니다.

8-3-2. 유저별 전체 광고 클릭 확률 추론

두 번째로, 8-2 에서 기술한 유저에 대한 광고 최적화 시나리오입니다. 유저 단위 개별 전체 광고 조합(약 5,000 개)에 대해 클릭 확률을 모두 예측한 뒤, 그중 가장 높은 확률을 가진 광고를 선정하는 시나리오에서는 평균 0.05 초의 추론 시간이 측정되었습니다.

실제 예측 속도가 매우 빠르다는 점 외에도, 모델 구조는 자원 상황에 따라 유연하게 운영 가능하도록 설계되어 있습니다. 환경이 충분하다면 모든 Bagging 모델을 병렬 처리로 추론할 수 있지만, 만약 연산 자원이 제한적일 경우, Bagging 별로 1 개의 폴더 모델만 사용하거나, Bagging 모델 중 일부만 샘플링하여 사용해도 높은 예측 성능을 안정적으로 유지할 수 있습니다.

결과적으로, 성능과 속도, 유연한 운영 가능성까지 모두 고려하여 실시간 광고 환경에서도 안정적으로 적용할 수 있습니다..

9. 피드백

9-1. 평가지표에 대한 피드백

현재 테스크의 평가지표(AP + WLL 조합 스코어)는 클릭 예측 문제의 불균형성, 실전 노출의 중요도, 확률 예측의 신뢰성을 모두 고려한 좋은 지표라고 생각합니다. 그러나 실제 서비스 환경과 비즈니스 목적에 더욱 부합하는 모델 평가를 위해서는, 데이터의 구조적 편향과 광고별(특히 l_feat_14, 광고세트 단위 등) 기회 균형까지 함께 반영하는 지표 설계가 필요하다고 생각합니다.

이번 모델링에서는, 플랫폼 내에서 특정 인기 광고 세트(l_feat_14)에 데이터가 집중되어 발생하는 쏠림 현상을 완화하고, 최대한 공정하게 기회가 돌아갈 수 있도록 l_feat_14 등장 빈도에 따라 샘플별 가중치를 부여한 커스텀 손실 함수를 적용했습니다. 이 전략의 핵심은, 단순히 전체적인 예측 정확도만 올리는 것이 아니라, 데이터가 적은 광고세트에서도 의미 있는 예측 성능을 확보해 플랫폼의 다양성과 공정성을 보장하는 데 있습니다.

따라서 모델의 성과를 평가하는 평가지표 역시 이러한 광고별 기회 균형을 반영하는 방식이 되어야, 실제로 의도한 "공정하고 다양한 광고 생태계" 구현 여부를 측정할 수 있다고 생각합니다. 예를 들어, l_feat_14 단위별로 가중치를 적용한 가중 AP/WLL 을 별도로 도입하는 방식이 효과적이라 생각합니다.

실제로, 현행 평가지표로는 데이터가 많이 쌓인 광고 세트에서만 성능이 좋아도 전체 점수가 높게 나오므로, 신규 광고주, 소형 광고주 등 데이터가 적은 집단의 예측력 개선 여부가 제대로 반영되지 않는 한계가 있습니다. 제가 Custom Loss 를 적용한 근본적 이유 역시, 이러한 광고별 불균형을 보완하고자 했던 것입니다.

따라서, 평가지표 설계도 "플랫폼 공정성, 광고주 다양성" 관점의 목표와 부합하도록 개선된다면, 모델 개발과 평가, 그리고 실제 비즈니스 인사이트 간의 일관성을 더욱 높일 수 있을 것이라 생각합니다.

9-2. 테스트셋 구성에 대한 제안

테스트 데이터셋에 Unseen 데이터(즉, 학습에 등장하지 않은 신규 광고·지면·유저 등)가 많이 포함되어 있을수록, 실제 서비스 환경에서 더 강건하고 실질적으로 정확한 모델을 평가할 수 있다고 생각합니다.

실제 광고 플랫폼 운영 환경에서는 새로운 광고, 신규 지면, 이전에 관측되지 않은 유저 행동 등이 빈번하게 발생합니다. 만약 테스트 데이터가 학습 데이터와 거의 동일한 분포이거나, 기존에 자주 등장했던 광고·지면 위주로만 구성되어 있다면, 모델이 단순히 과거 패턴만 암기(Overfitting)해도 높은 점수를 받을 수 있습니다. 이런 경우, 실 서비스에서는 성능이 급격히 저하될 위험이 크며, 실제 환경에서의 강건함과 일반화 능력을 제대로 평가하지 못하게 됩니다.

반대로, 테스트 데이터셋에 Unseen 데이터가 충분히 포함된다면, 모델이 신규 광고나 변화된 상황에도 얼마나 잘 적응하고, 기존 인기 광고에만 치우치지 않는지, 즉 실제 환경에서의 실질적 예측력과 Robustness 를 더 엄밀히 검증할 수 있습니다.

특히 이번 과제처럼 콜드스타트와 데이터 편향 문제가 중요한 도메인에서는, Unseen 데이터 기반의 평가가 모델의 진짜 경쟁력을 가려내는 핵심 기준이 될 수 있습니다.

결론적으로, 테스트셋을 설계할 때 Unseen 데이터의 비율을 충분히 확보하는 것이, 실전형 AI 모델 개발 및 평가에 큰 도움이 될 것이라고 생각합니다.

좋은 기회 주셔서 감사드립니다.

팀 쥬혁이

전주혁