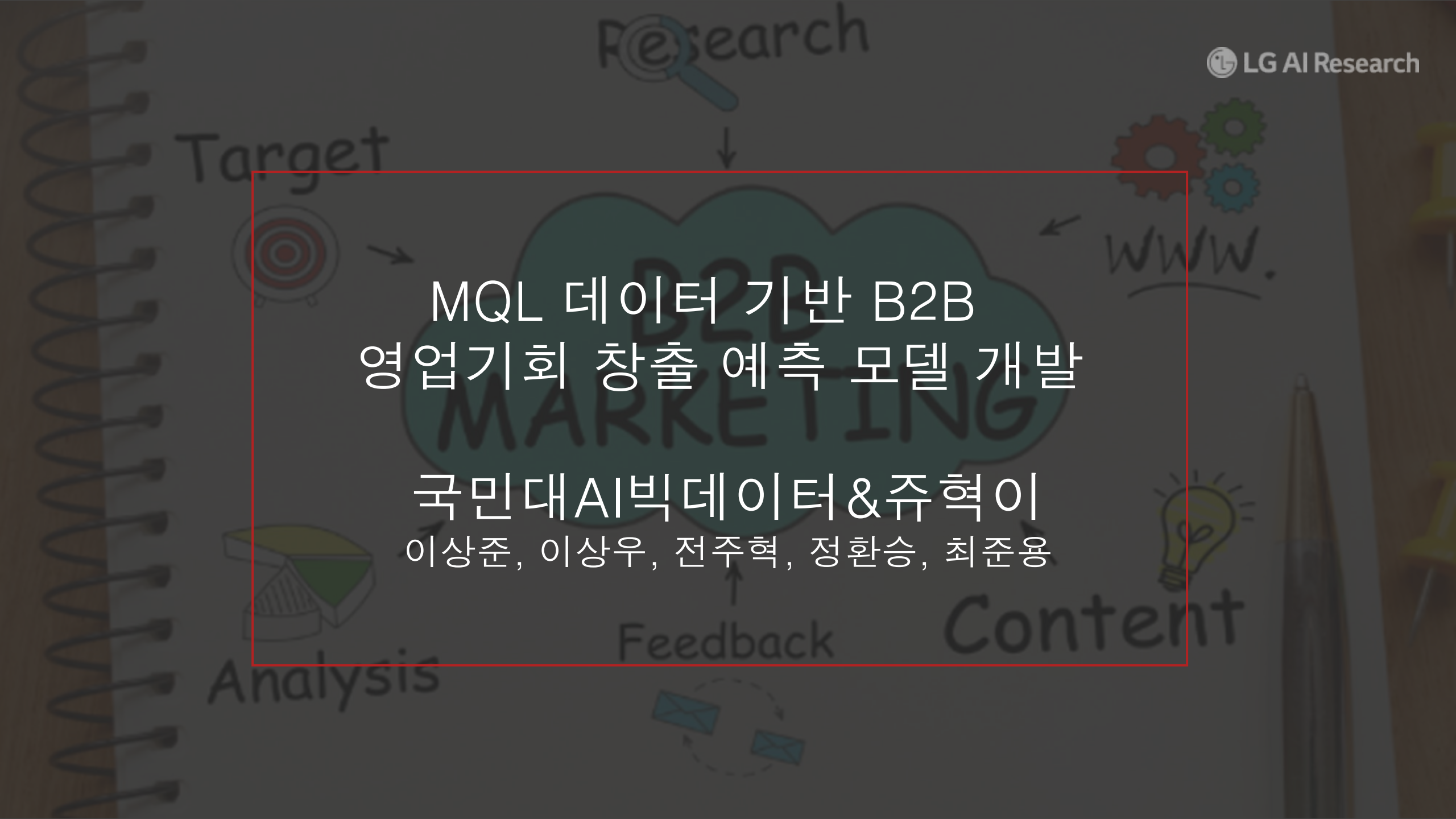


LG AI Research

MQL 데이터 기반 B2B 영업기회 창출 예측 모델 개발

국민대AI빅데이터&쥬혁이

이상준, 이상우, 전주혁, 정환승, 최준용

Contents

대회 배경 및 목표

수많은 고객 데이터들의 축적

마케팅 활동의 최적화

데이터 처리 비용 절감

MQL 데이터 기반 B2B 영업기회 창출 예측 AI 개발

제공데이터

	Columns	Remarks
Train Data	customer_idx, customer_country, customer_type, enterprise, customer_job, customer_position, customer_country	고객 정보 (Customer Information)
	bant_submit, historical_existing_cnt, lead_desc_length, inquiry_type, expected_timeline	영업 및 마케팅 (Sales & Marketing)
	business_unit, business_area, business_subarea	사업 단위 및 영역 (Business Unit & Area)
	com_reg_ver_win_rate, id_strategic_ver, it_strategic_ver, idit_strategic_ver, ver_cus, ver_pro, ver_win_rate_x, ver_win_ratio_per_bu	성능 지표 (Performance)
OUR TARGET	is_converted	영업 전환 여부

MQL 데이터 기반 B2B 영업기회 창출 예측 Task

평가 지표

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

		Actual Values: Positive	Actual Values: Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
Predicted Values	Negative	False Negative (FN)	True Negative (TN)

Why F1 SCORE?

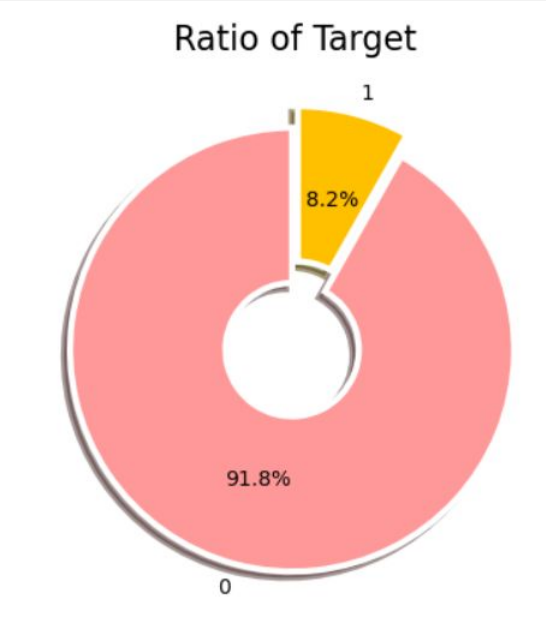
- 고가치 계약**(High Recall)**
- 리소스 효율성**(High Precision)**

Our Approach

- Understanding B2B
- How to increase F1 Score
- Causal Feature Selection
- Deal with Categorical data
- Prevent Overfitting

Contents

EDA

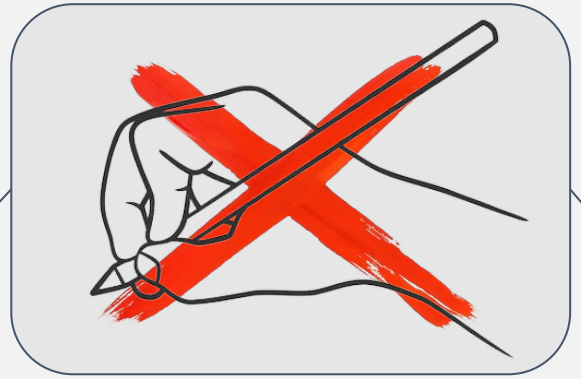


- 전체 data 중 0값이 약 92%를 차지함
- Imbalanced Target

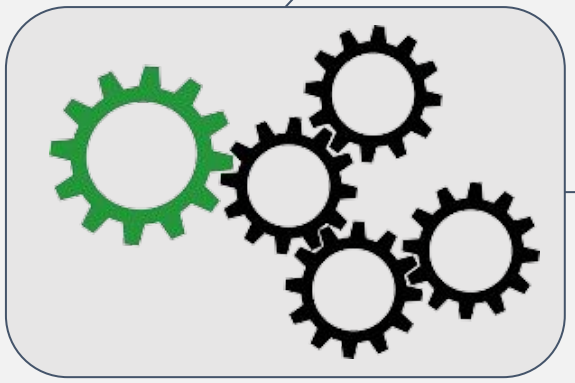
Imbalanced Target value + Sparse dataset
→ 학습시 방해요소로 작용

	number_of_null	percentage_of_null
it_strategic_ver	54634	0.980123
id_strategic_ver	52409	0.940207
idit_strategic_ver	51301	0.920329
business_subarea	50578	0.907359
product_subcategory	46740	0.838506
product_modelname	46715	0.838057
historical_existing_cnt	43380	0.778228
com_reg_ver_win_rate	41640	0.747013
customer_type	41354	0.741882
ver_win_ratio_per_bu	40868	0.733164
ver_win_rate_x	37975	0.681264
business_area	37975	0.681264
expected_timeline	28260	0.506979
product_category	17146	0.307596
customer_job	16776	0.300958
customer_country	974	0.017473
customer_country.1	974	0.017473
inquiry_type	891	0.015984

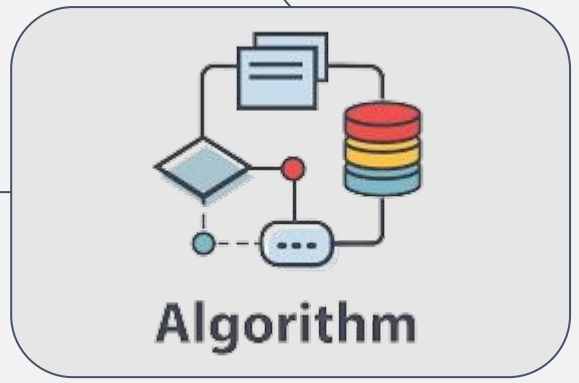
EDA_how to fill missing values



Concept 1

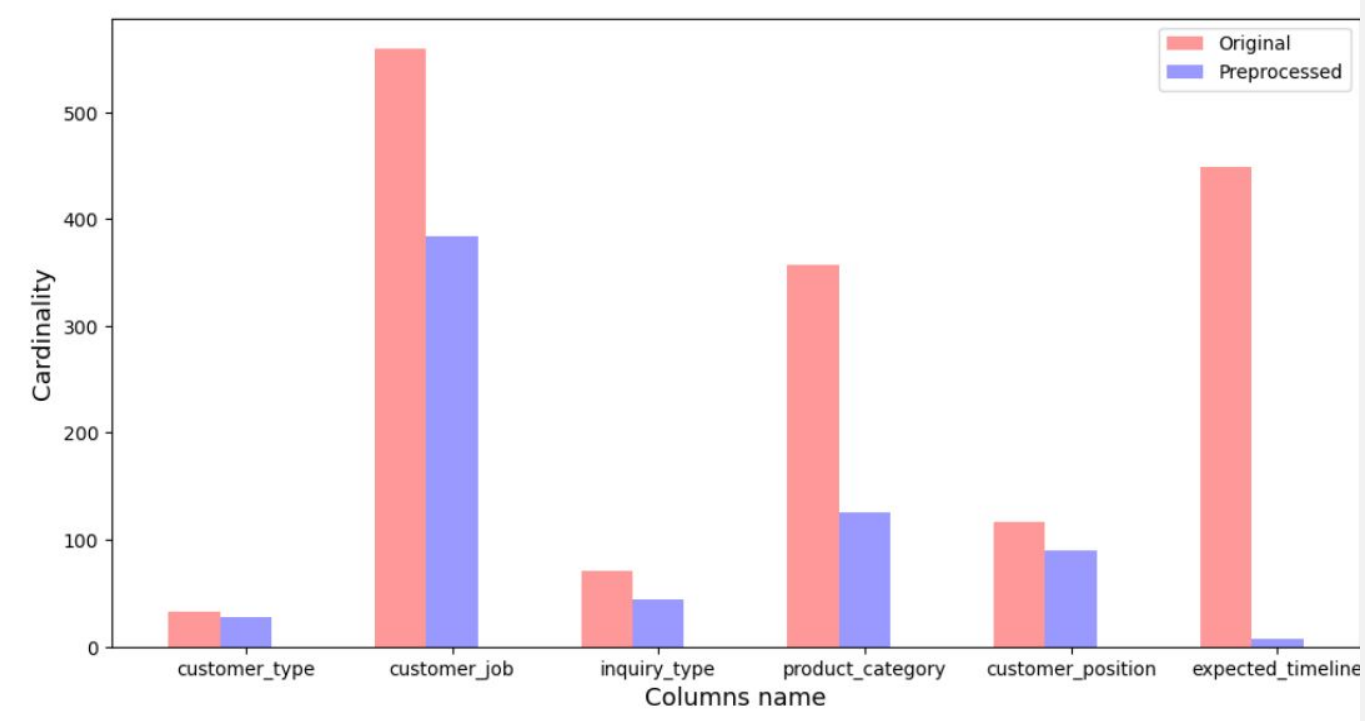


Concept 2



Concept 3

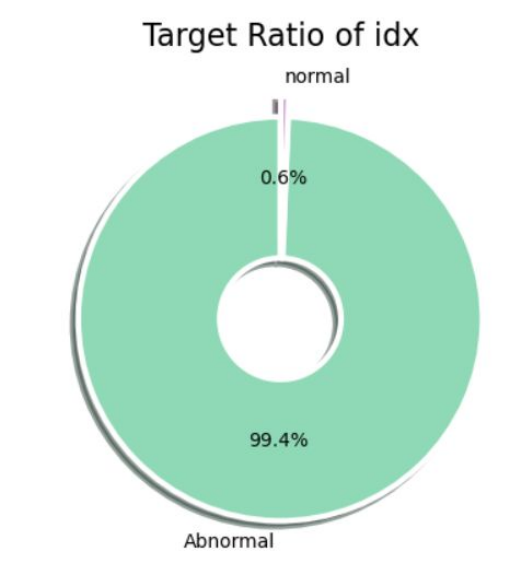
EDA



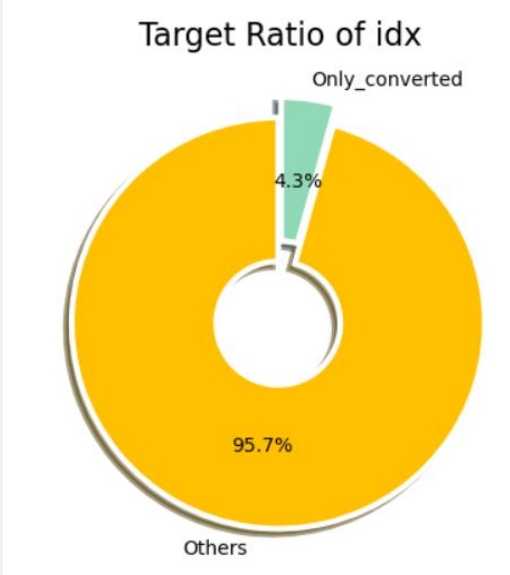
High Feature Cardinality

- Original Mean Cardinality = 264.5
- Processed Mean Cardinality = 113.0

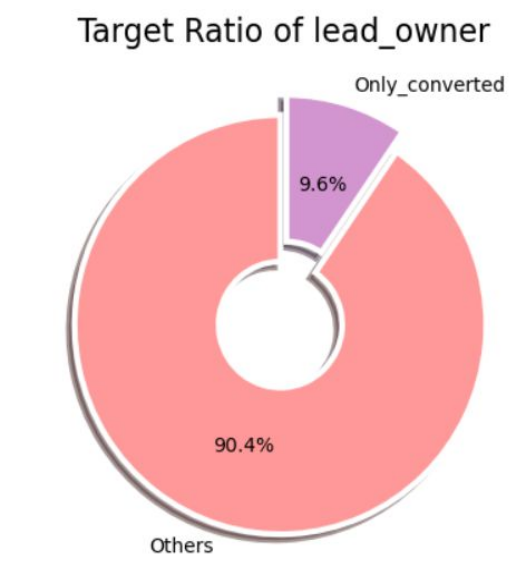
EDA



Abnormal : 영업전환율이 0 또는 1인 idx의 비율
Normal : 영업전환율이 0 과 1 사이인 idx의 비율



only_converted :
전체 idx 중 영업전환율이 1 인 idx 의 비율

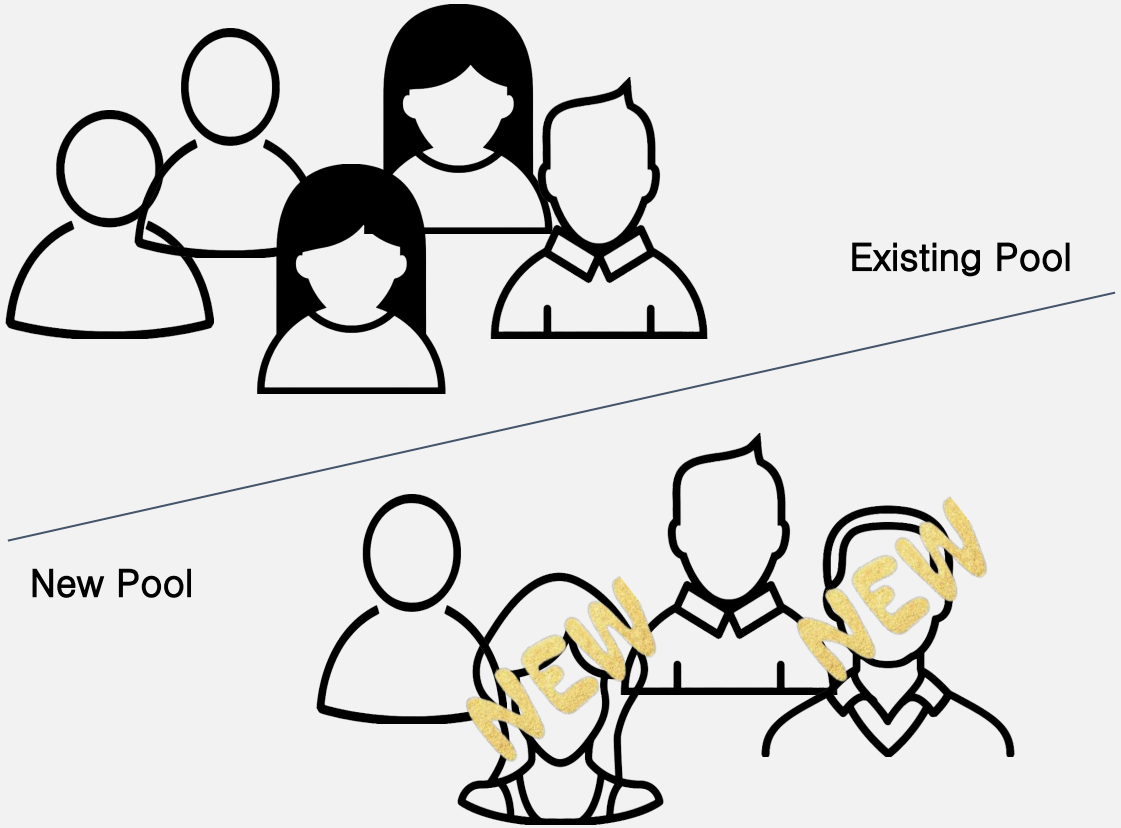


only_converted :
전체 lead_owner 중 영업전환율이 1 인 lead_owner의 비율

Customer_idx/Lead_owner
→ 모델의 편향된 예측을 초래

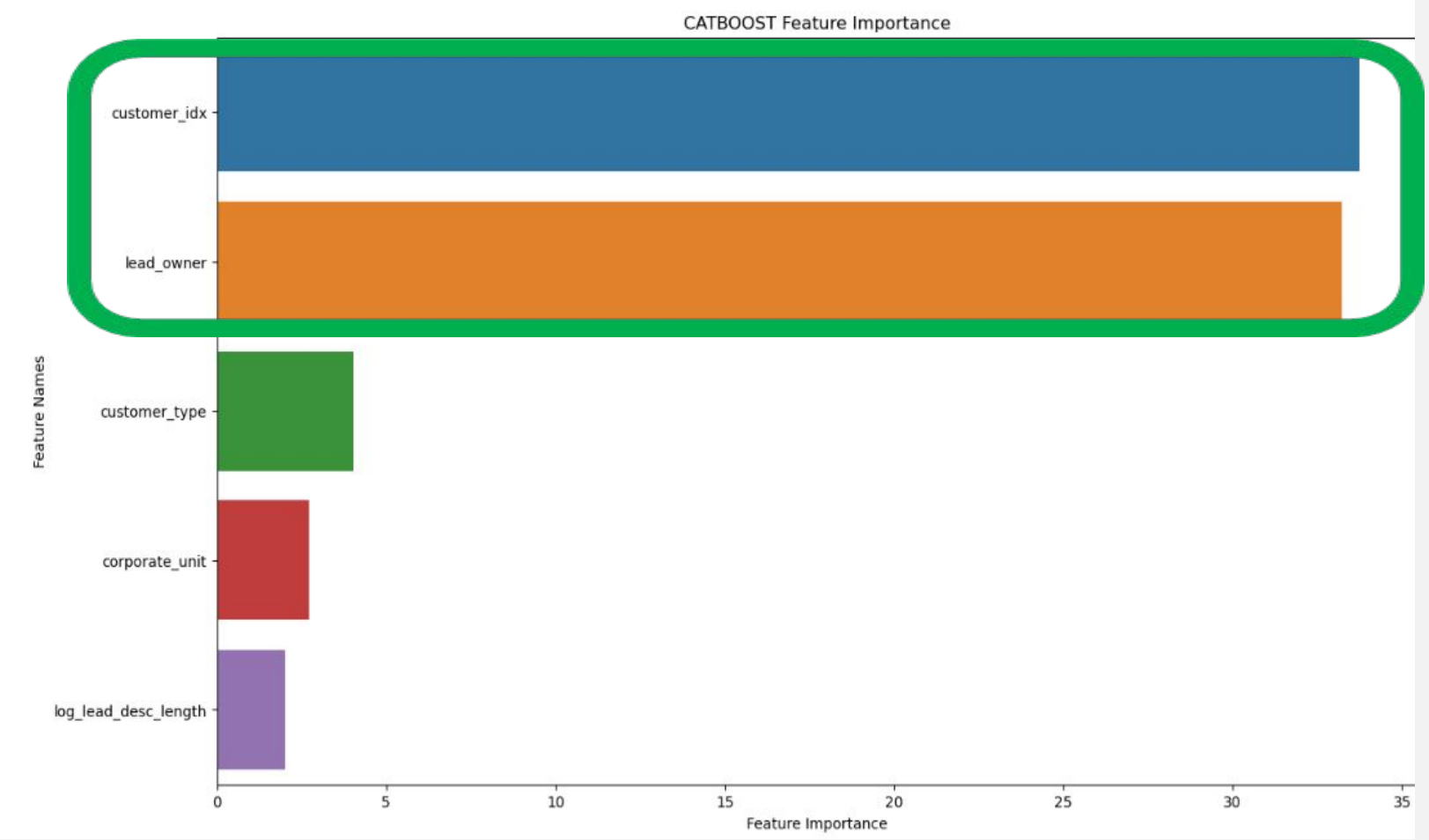
Data Drift(Feature Drift)

모델 훈련시 "입력 데이터(Customer_idx, Lead_owner)"의 **통계적 분포** 및 **unique value**와 테스트 시/ 실제 배포 환경에서의 "입력 데이터"의 통계적 분포 및 unique value가 어떠한 변화에 의해 차이가 발생하고 있는 것을 의미



- 기존 고객 풀의 리드에 대한 영업전환예측의 경우, 높은 예측 성공률
- 새로운 고객 또는 신규 영업사원으로 구성된 리드의 경우, 예측 성공률이 급격히 저하될 것으로 예측됨

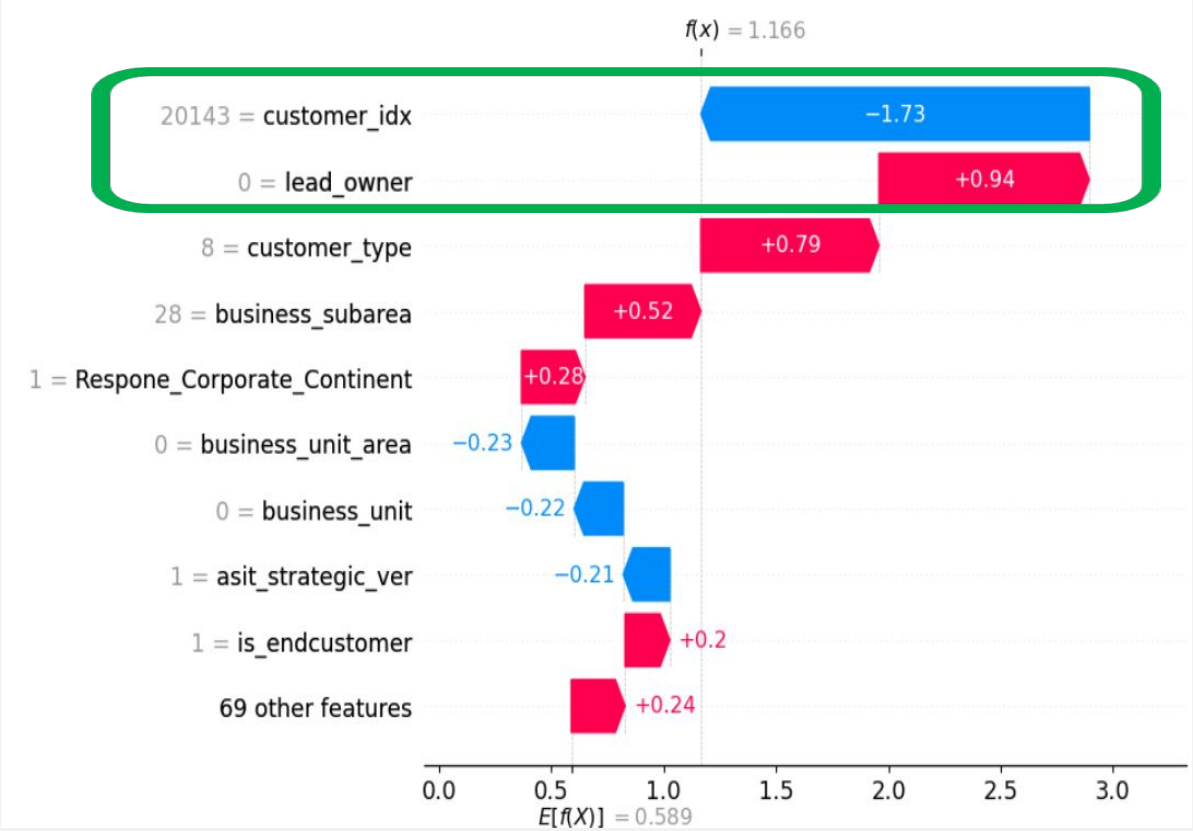
Model based(Feature Importance)

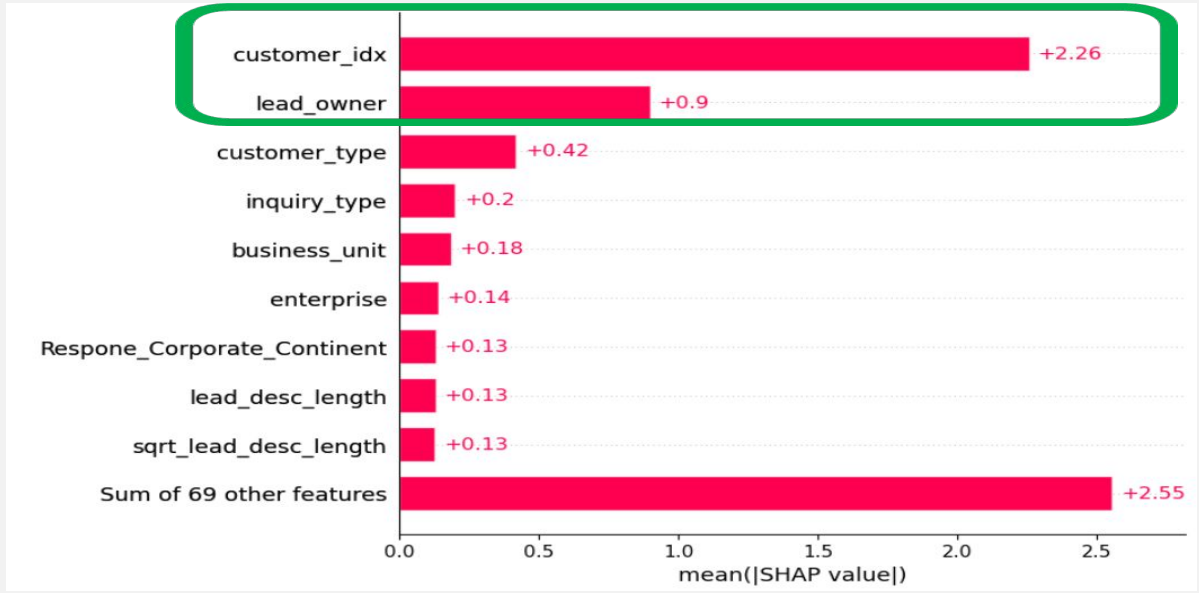


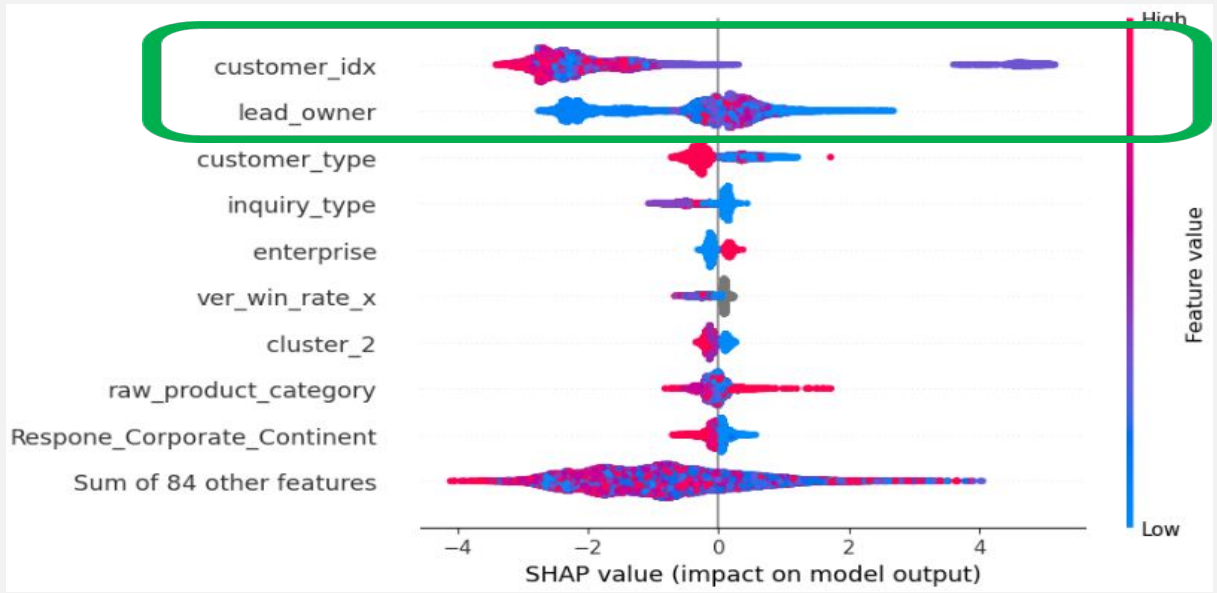
- 특정 Customer_idx 및 Lead_owner가 모델 학습에 **편향성을 가중**시키고 있는 모습
- Customer_idx 및 Lead_owner가 결측치인 Lead에 대한 예측력이 현저히 떨어질 것으로 예측됨

LG AI Research

Shap value

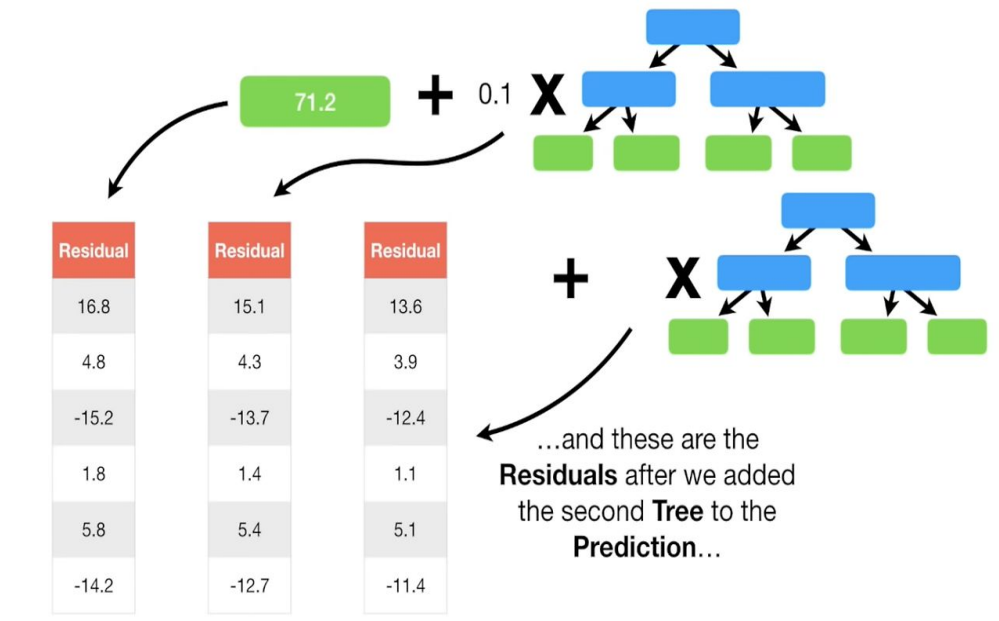






Contents

CATBOOST



특징

- Ordering Boosting & Random Permutation
→ Prevent Overfitting

- Auto Ordered Target Encoding & One-hot Encoding

- Categorical Feature Combinations
→ Feature Cardinality down
→ 연속형 변수가 아닌 수치형 변수는 범주형 변수로 볼 수 있음

- Optimized Parameter tuning

Modeling based on CONTEXT

- ALL Feature Model

- DROP customer_idx Model

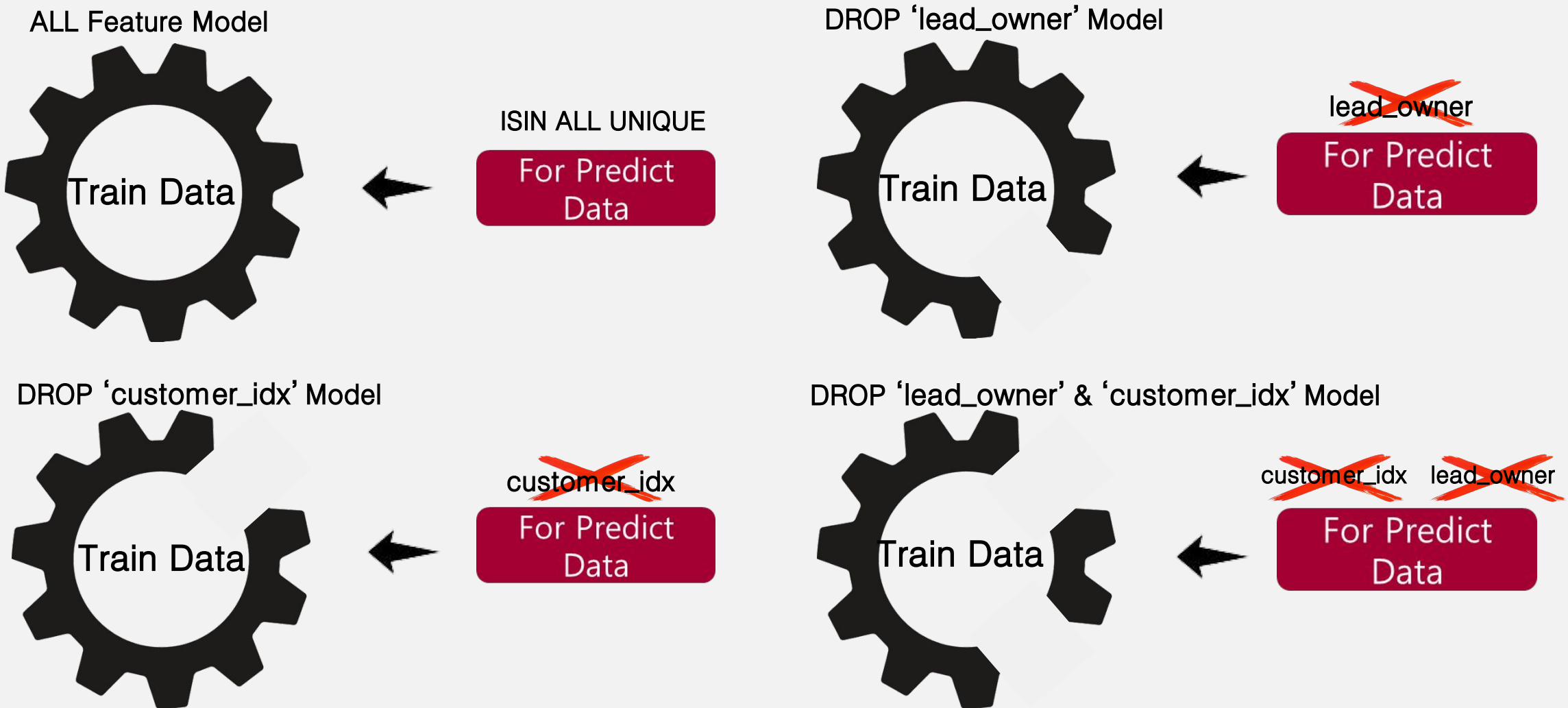
- DROP lead_owner Model

- DROP lead_owner & customer_idx Model

기대효과

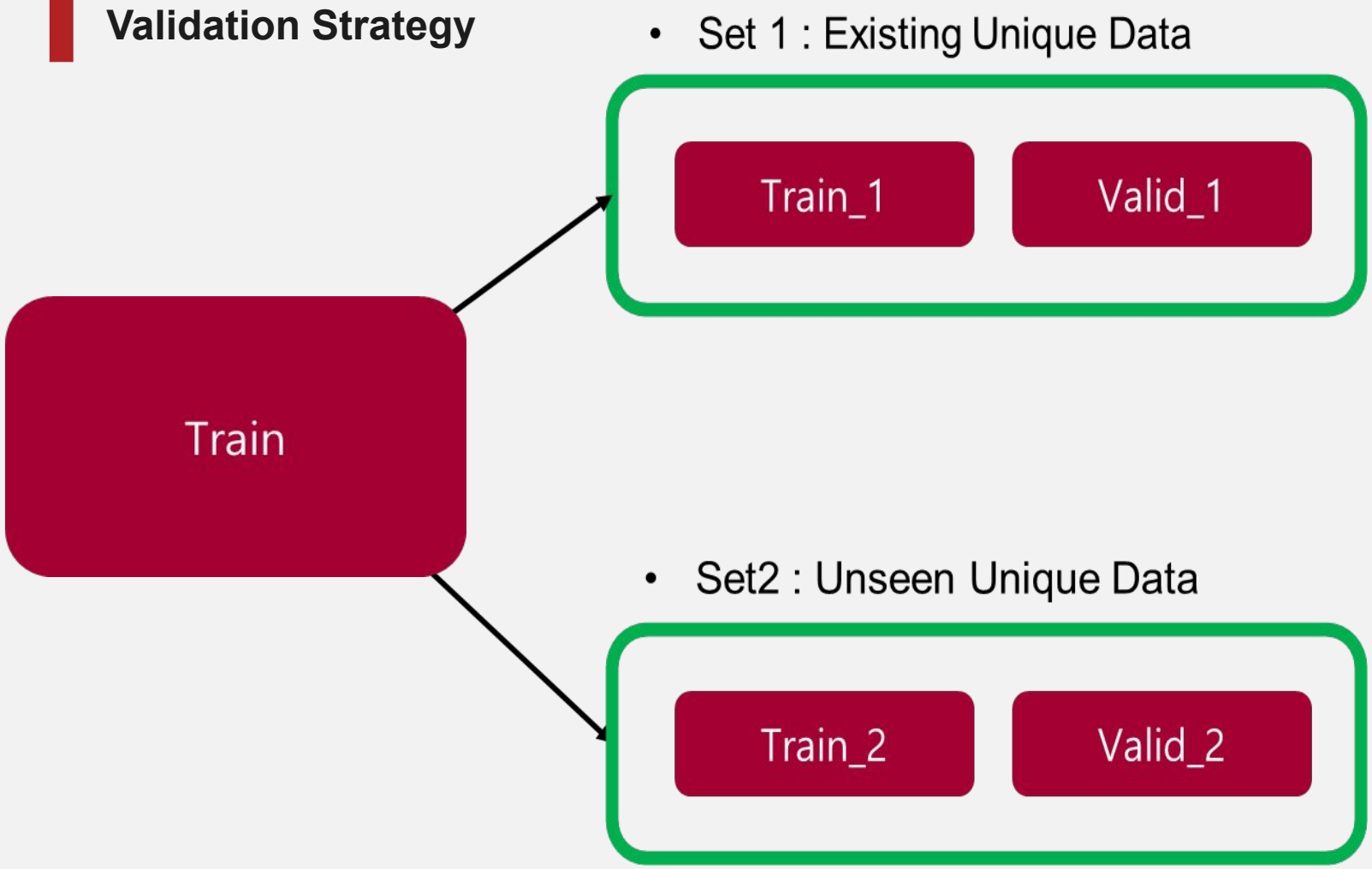
- Data Drift 로 인한 train data 에 존재하는 customer_idx, lead_owner 에 대한 과적합을 막을 수 있음
- test data 에 unseen data(customer_idx, lead_owner)가 등장해도 예측을 잘해낼 수 있음

Modeling based on CONTEXT OVERALL FLOW



Contents

Validation Strategy



- 앞서 언급한 Unseen Unique 값에 대한 오버피팅을 증명하는 방법
- F1 Score 를 validation score 로 사용
- 현실 상황에서도 새로운 상황 및 오타와 같은 여러 상황을 고려했을 때 Unseen Unique 에 대한 고려가 필수적
- Overfitting 방지 및 최종적으로 고객 전환률 예측에 활용

LG AI Research

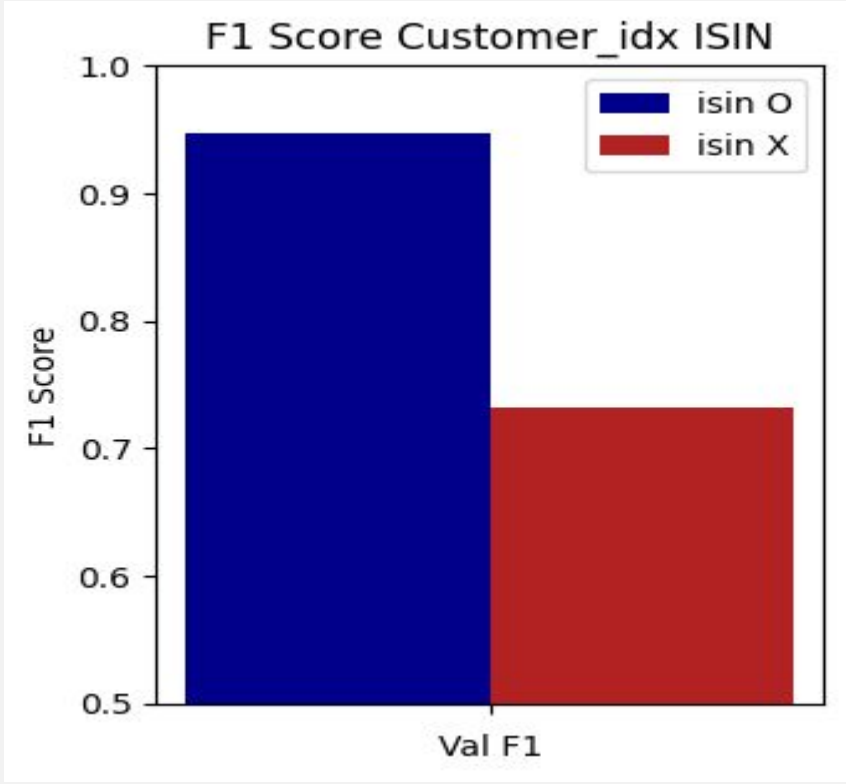
Prevent Over-Fitting1 (Overlapping Unique Values)
- customer_idx

Valid1 : Existing Unique Data

```
Validation F1 score: 0.9475139001561806
```

Valid2 : Unseen Unique Data

```
Validation F1 score: 0.7320472759480329
```



- (Valid1: Existing Unique) VS (Valid2: Unseen Unique) F1 스코어 비교 ->모델의 일반화 능력 평가
- Unseen Unique Data를 검증 데이터로 사용했을 때의 F1 스코어 저하는, 모델이 새로운 customer_idx unique 값에 대한 예측에서 극도의 과적합을 보임을 증명

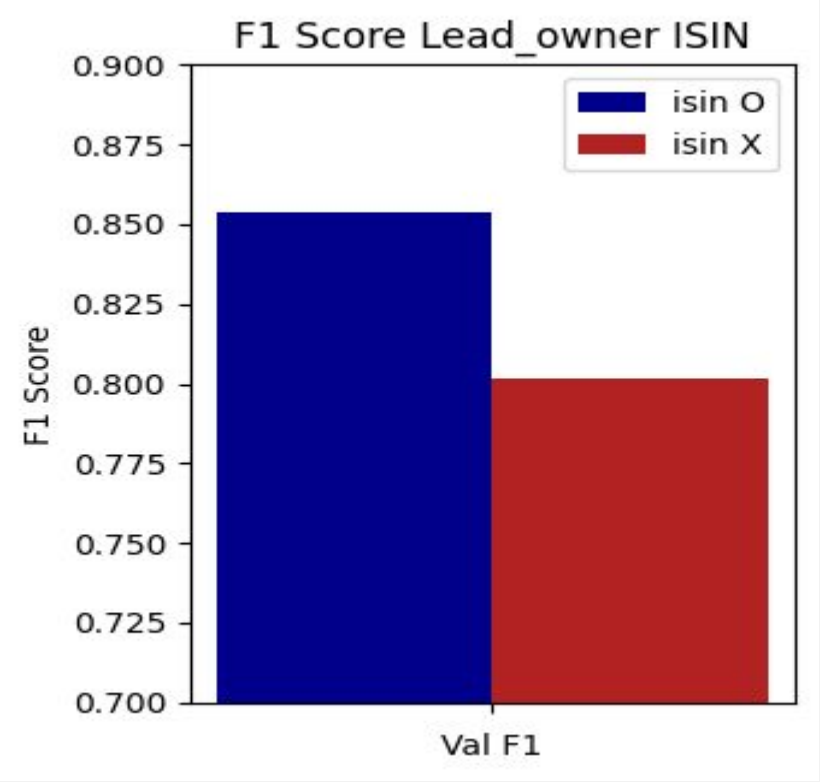
Prevent Over-Fitting2 (Overlapping Unique Values) - lead_owner

Valid1 : Existing Unique Data

```
Validation F1 score: 0.8540386804366285
```

Valid2 : Unseen Unique Data

```
Validation F1 score: 0.80180095552558827
```



- (Valid1: Existing Unique) VS (Valid2: Unseen Unique) F1 스코어 비교 ->모델의 일반화 능력 평가
- Unseen Unique Data를 검증 데이터로 사용했을 때의 F1 스코어 저하는, 모델이 새로운 lead_owner unique 값에 대한 예측에서 극도의 과적합을 보임을 증명

강건한 모델 검증

- (4개 모델 **VALIDATION SCORE**와 실제 **PRIVATE SCORE** 비교)

Model 1: ALL Feature : 0.861483

Model 2: DROP 'lead_owner' : 0.813921

Model 3: DROP 'customer_idx' : 0.787294

Model 4: DROP 'lead_owner' & 'customer_idx' : 0.715691

4 MODEL MEAN F1 SCORE : 0.794597

Contents

현업 적용 가능성

범주형 데이터에 적합한 Algorithm Model

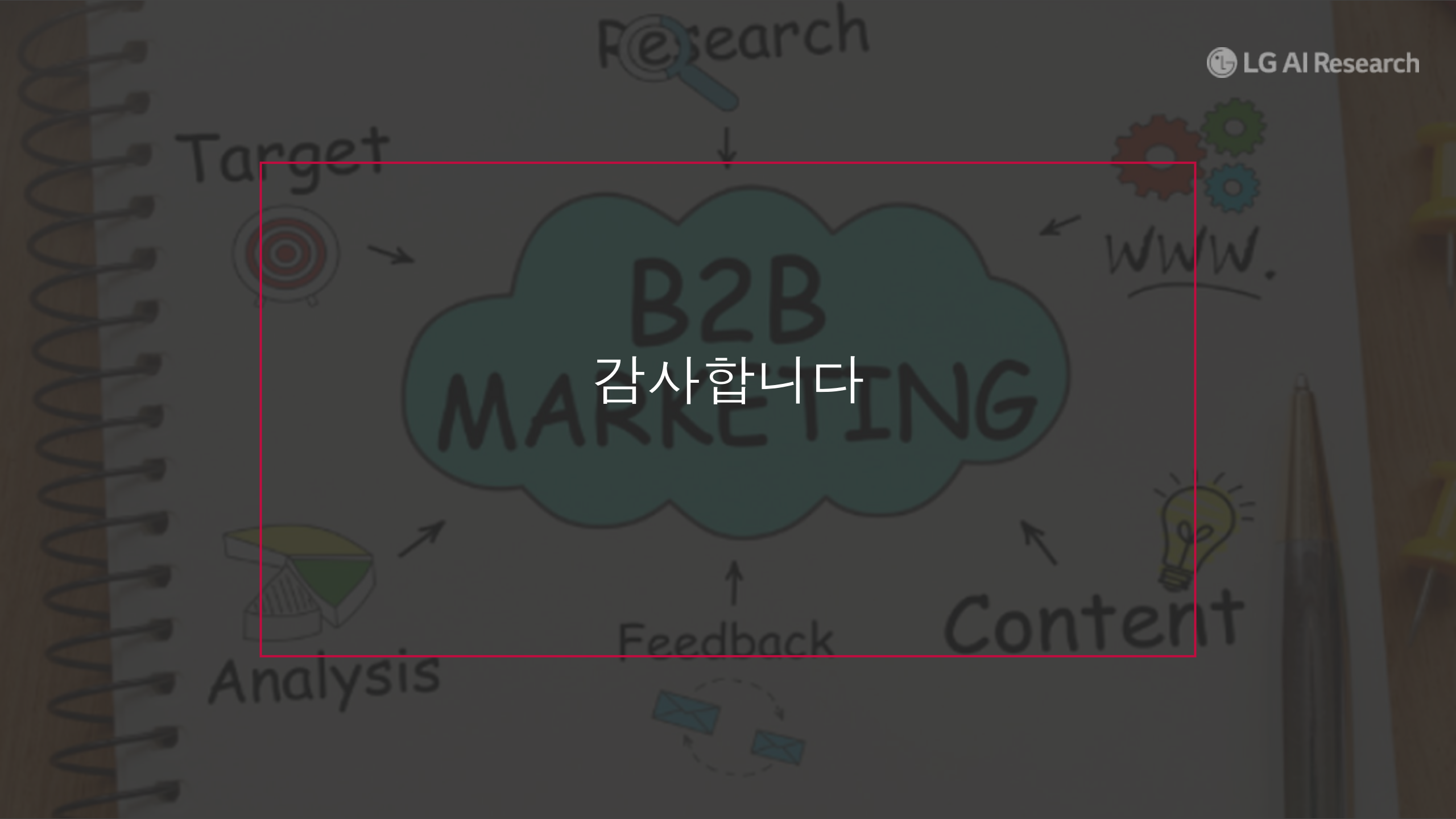
범주형 변수가 많은 B2B 마케팅 데이터에 다양한 범주형 변수를 학습할 수록 성능이 좋아지는 CatBoost 모델

Imbalanced Dataset 에 Robust

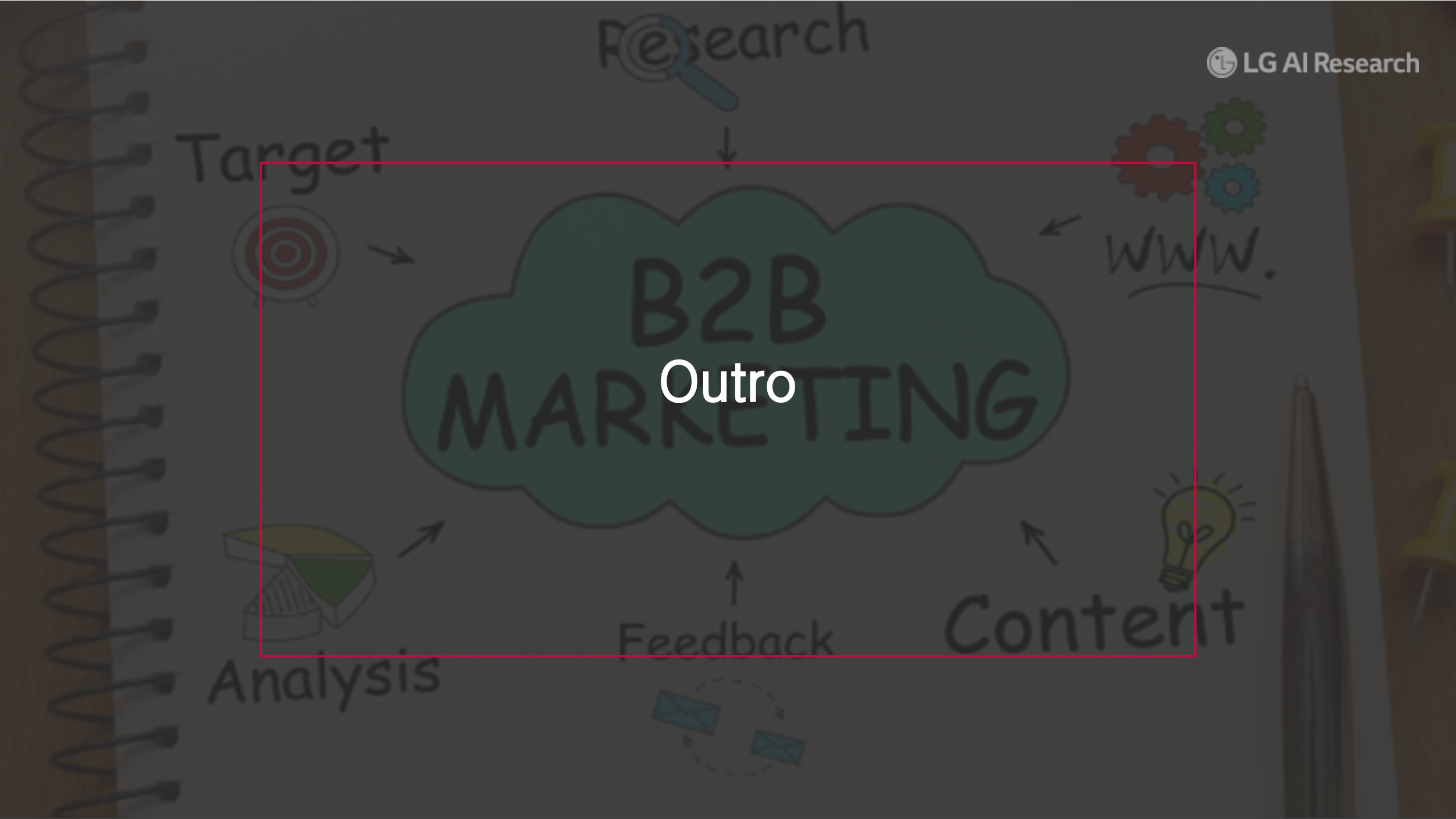
특정 변수에 과적합 되지 않고 불균형이 심한 데이터에 대해 명확한 기준(scale_pos_weight)이 있어 threshold 값을 별도로 설정할 필요가 없음

Tunning이 필요없는 Model

파라미터 튜닝 및 복잡한 모델 구조를 띄지 않아 상대적으로 가볍고 Robust한 모델임.


LG AI Research
Research
Target
WWW.
B2B
MARKETING
Analysis
Feedback
Content

감사합니다


Research
Target
WWW.
B2B
MARKETING
Analysis
Feedback
Content

Outro