LG AIMERS 4기

MQL 데이터 기반 B2B 영업기회 창출 예측 모델 개발

영업 챔피언스 | 김이정 오인우 이세희 박주현

국내 최고 전문가의
AI Essential Course

LG 실제 데이터를 다루는
실전적 AI 해커톤

Phase 2 상위팀의
모델고도화 해커톤

LG

목차

1 EDA 및 데이터 전처리

2 모델 구축 및 검증

3 결과 및 제언

EDA & Data pre-processing

1. EDA 및 데이터 전처리

01. Data pre-processing

대회 소개

MQL 고객 정보를 활용하여 영업 성공 전환 여부를 예측하는 AI모델 개발

MQL 고객 정보에는
개인 정보(회사/직급 등), 구매요청 정보(제품/예산/니즈/기한) 및 영업사원, 마케팅 활동 정보등이 포함

Table data

고객 ID	직책	제품	유입채널	…	Y
A	MQL 정보 변수 약 30개				영업전환이력 성공 : 1 실패 : 0
B					
C					
…					
Z					

+

Text data

고객이 직접 작성한
요청 메시지 (영어)

예시)

we need Air Ventilation Solution
ASAP for our new building…

01 Data pre-processing

데이터 특성 파악 후
전처리를 진행

데이터 범주형 변수가 대부분 → 카테고리 재분류 & 내용 및 오타 수정

Dataset statistics

Number of variables	29
Number of observations	59299
Missing cells	654543
Missing cells (%)	38.1%
Duplicate rows	2405
Duplicate rows (%)	4.1%

Variable types

Categorical	10
Text	11
Numeric	7
Boolean	1

Processed data

customer_type / customer_job / customer_position

id_strategic_ver / it_strategic / idit_strataegic_ver

product_category

inquiry_type

expected_timeline

ver_cus / ver_pro

Dropped data

customer_country / customer_country.1

response_corporate

product_subcategory / product_modelname

business_subarea

01 Data pre-processing

LG Bussiness Solutions - Inquiry To Buy정보를 기반으로 카테고리 재할당

customer_type

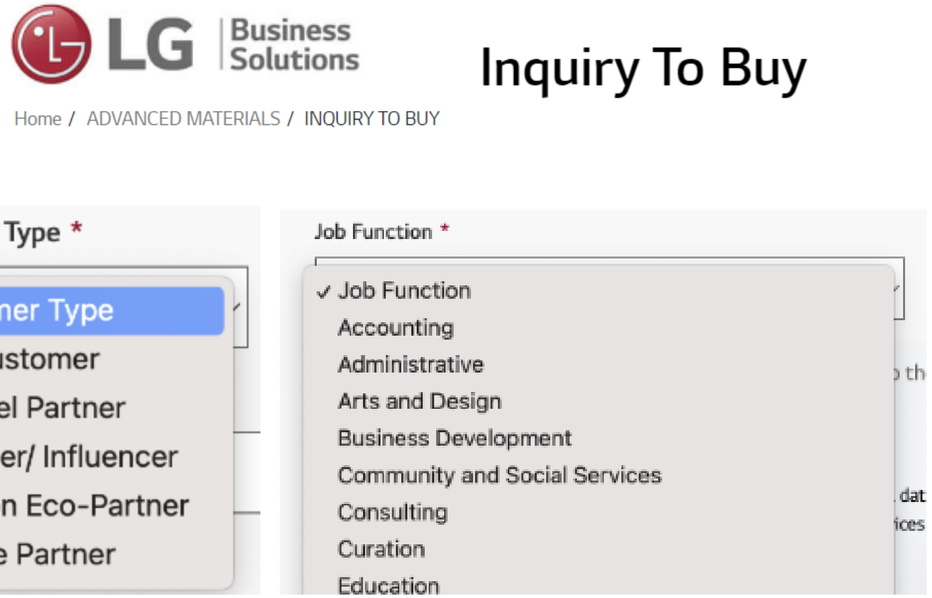
카테고리 재할당 → 5개의 카테고리화, 나머지는 그대로 유지

- End Customer','Specifier/ Influencer','Solution', 'Eco-Partner','Other','Home_Owner'

customer_job & customer_position

카테고리 재할당 → 컬럼사이 잘못 들어간 값 서로 변경

- [JOB] 총 37개의 카테고리
  accounting, administrative, arts_and_design, business_development, clinical_specialist …
- [POSITION] : 총 11개의 카테고리
  CEO/Founder, Partner, C-level Executive, Vice President, Director, Manager …



01 Data pre-processing

LG Bussiness Solutions - Inquiry To Buy정보를 기반으로 카테고리 재할당

inquiry_type

- 카테고리 재할당 → 9개의 카테고리화, 나머지는 그대로 유지

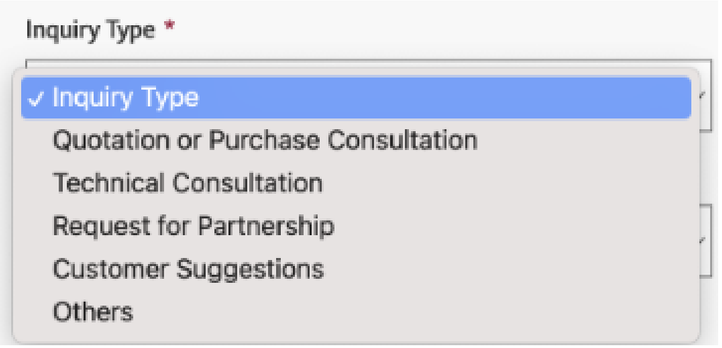
'Quotation or Purchase Consultation', ‘Usage or Technical Consultation’, ‘Request a Demo','OEM/ODM Request’,'Request for Partnership' …

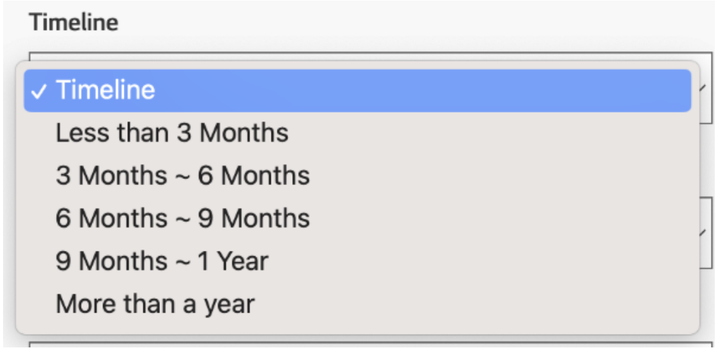
expected_timeline

- 카테고리 재할당 → 5개의 카테고리화, 결측치는 ‘Not specified’ 값

Less than 3 months
3 Months ~ 6 Months
6 Months ~ 9 Months
9 Months ~ 1 year
More than a year

LG Business Solutions





01 Data pre-processing

LG Bussiness Solutions - Inquiry To Buy정보를 기반으로 카테고리 재할당

inquiry_type

- 카테고리 재할당 → 9개의 카테고리화, 나머지는 그대로 유지

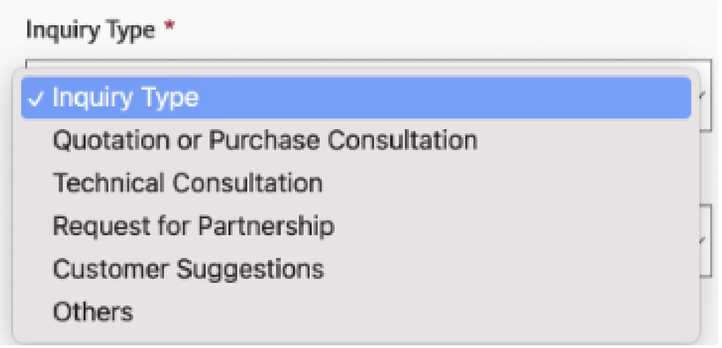
'Quotation or Purchase Consultation', ‘Usage or Technical Consultation’, ‘Request a Demo','OEM/ODM Request’,'Request for Partnership' …

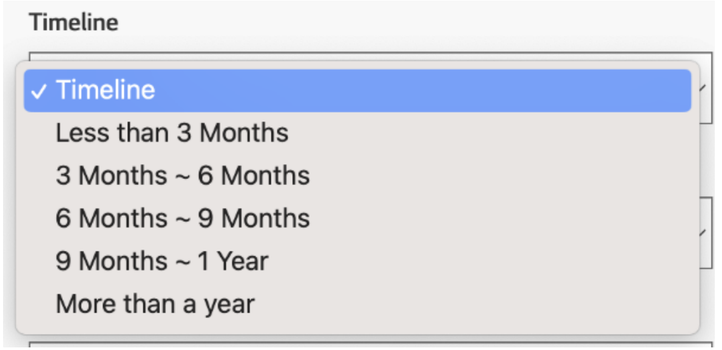
expected_timeline

- 카테고리 재할당 → 5개의 카테고리화, 결측치는 ‘Not specified’ 값

Less than 3 months
3 Months ~ 6 Months
6 Months ~ 9 Months
9 Months ~ 1 year
More than a year

LG Business Solutions





01 Data pre-processing

LG Bussiness Solutions - Inquiry To Buy정보를 기반으로 카테고리 재할당

product_categorype

- 오타수정 및 카테고리 매핑
  영어가 아닌 다른 언어로 기입된 내용 수정

```
"other" : ["other", "others", "etc.", "khác", "outros",
"commercial_tv" : ["commercial tv", "commercial tv,tv",
"heating" : ["heating", "חימום" ,"حلول التدفئة", "isıtma
"multi_split"  : ["multi-split", "פיצול מרובה", "multi s
"single_split" : ["single-split", "split tunggal", "sing
"chiller" : ["chiller", "(تشيلر) مبرد", "soğutucu", "pen
"video_wall_signage"  : ["video wall signage", "videwall
"hotel_tv" : ["hotel tv", "酒店電視"],
"hospital_tv" : ["hospital tv", "醫院電視"],
```

ver_cus & ver_pro

- train data의 ver_cus 컬럼 확인 →

  특정 business_unit을 ['corporate / office', 'retail', 'education', 'hotel & accommodation'] 로 설정

  → 위 4가지 business_unit & End_customer 이면 가중치

- train data의 ver_pro 컬럼 확인 →

  특정 category를 ['signage', 'hotel_tv'] 로 설정

  → 위 4가지 business_unit & 특정 category 해당하면 가중치

01 Data pre-processing

LG Bussiness Solutions - Inquiry To Buy정보를 기반으로 카테고리 재할당

id_strategic_ver & it_strategic_ver & idit_strataegic_ver

- 'business_unit' 컬럼의 'ID, IT' 값과 관계된 컬럼 → 가중치
- ‘business_area’ 가 'corporate / office', 'retail', 'hotel & accommodation', 'education' 라면 → 가중치

1. business_unit이 ID 이면서 business_area 가 특정 사업부인 경우 가중치(1) 부여
2. business_unit이 IT 이면서 business_area 가 특정 사업부인 경우 가중치(1) 부여
3. 'id_strategic_ver'나 'it_strategic_ver' 중 하나가 1인 경우, 'idit_strategic_ver' 에 1을 할당

Model

Catboost Model

Catboost 모델은 자체 인코딩을 진행하므로 columns_to_drop(제외할 데이터)과 cat_features(범주형 데이터)를 지정

```
columns_to_drop =
["customer_country","customer_country.1", "response_corporate",
"product_subcategory",
"product_modelname",
"business_subarea"]
```

```
cat_features =
['business_unit','customer_idx',
 'customer_type','enterprise',
 'customer_job','inquiry_type',
 'product_category','customer_position',
 'business_area','lead_owner',
 'expected_timeline']
```

Model Selection & Validation

2. 모델 구축 및 검증

01 Catboost 모델 선정 이유

02 클래스 불균형 해결

03 Stratified K-Fold

04 Ensemble

01 Catboost 모델 선정 이유

Catboost Model

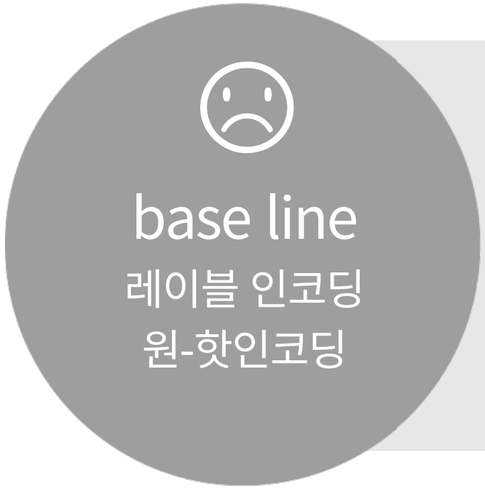
핵심 장점	
	범주형 데이터 자체 인코딩 + 자체 하이퍼파라미터 튜닝 + 상대적으로 Overfitting 적지만 좋은 성능

train data는 *20개 이상의 컬럼이 범주형 데이터*로,
복잡한 데이터 전처리 과정 없이도 자체적으로 인코딩이 가능하며,
범주형 데이터에 높은 성능을 보이는 **CatBoost 모델이 적합하다고 판단**

Variable types

Categorical	10
Text	11
Numeric	7
Boolean	1

01 Catboost 모델 선정 이유 - 범주형 데이터 전처리



base line의 레이블 인코딩과 원-핫 인코딩은 간단하지만 문제가 존재

- **레이블 인코딩**
  인코딩값이 수학적 의미를 가져 숫자의 크고 작음에 의미가 부여되어 잘못된 해석이 가능
- **원-핫 인코딩**
  레이블 인코딩의 문제는 해결 되지만 데이터의 차원이 크게 증가하여 모델의 복잡도가 크게 상승

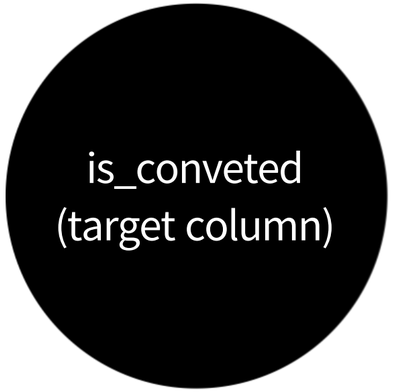
CatBoost 모델의 인코딩 과정은 Ordered Target Encoding

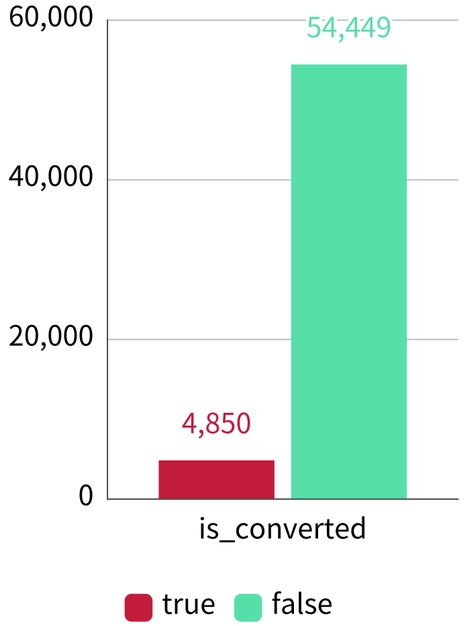
- **Ordered Target Encoding**

범주 내의 타겟 변수와의 관계를 직접적으로 반영할 수 있으며, 데이터의 차원을 증가시키지 않음
과적합을 방지할 수 있는 메커니즘을 포함하고 있음

02 클래스 불균형 해결 - scale_pos_weight

Target data value
불균형 문제 존재





True : False = 1 : 11.22xxx

parameter : scale_pose_weight

catboost 모델의 **scale_pos_weight 파라미터**는
imbalance한 데이터에서 가중치를 조절하여 **소수의 클래스에 더 집중하도록하는 역할**

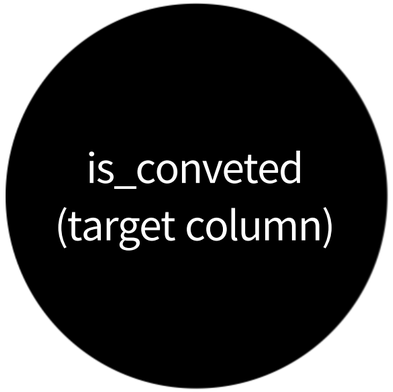
scale_pos_weight=11.22xxx로 설정 overfitting 이슈
최종적으로 완화된 수치인 11로 설정

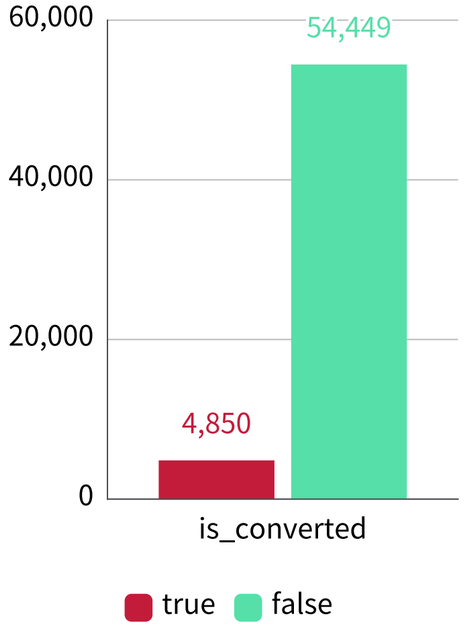
```
# 클래스 0과 클래스 1의 비율에 따라 scale_pos_weight 설정
df_train['is_converted'].value_counts()

scale_pos_weight = 54449/4850
scale_pos_weight
```

```
11.22659793814433
```

02 클래스 불균형 해결 - scale_pos_weight

Target data value
불균형 문제 존재





True : False = 1 : 11.22xxx

parameter : scale_pose_weight

catboost 모델의 **scale_pos_weight 파라미터**는
imbalance한 데이터에서 가중치를 조절하여 **소수의 클래스에 더 집중하도록하는 역할**

scale_pos_weight=11.22xxx로 설정 overfitting 이슈
최종적으로 완화된 수치인 11로 설정

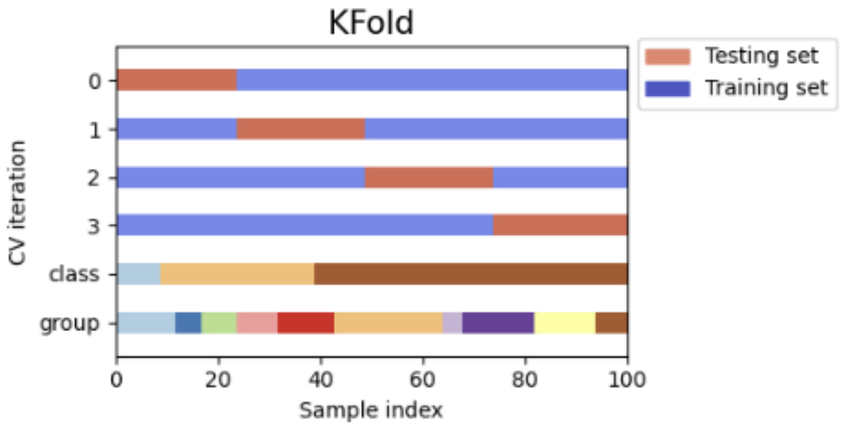
해당 파라미터 설정 이후 성능이 유의미하게 상승

	scale_pos_weight 설정 전	scale_pos_weight=11 설정 후
F1 score	0.612153	0.757121
리더보드	168위	리더보드 21위/1260위

03 Validation – Stratified K-Fold

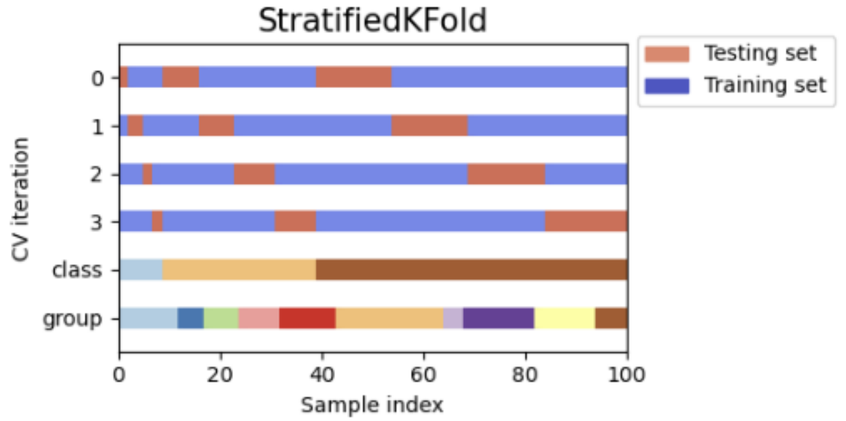
Validation 전략의 변화 | K-Fold에서 Stratified K-Fold로 변환

K-Fold



데이터셋을 k개의 폴드로 나누는 과정에서
각 폴드는 동일한 비율의 데이터를 가지도록 랜덤하게 선택
과적합을 방지하고 일반화 성능을 향상시키기 위해 사용

Stratified K-Fold



K-Fold를 진행할 때 각 폴드 내의 클래스 비율이
전체 데이터셋과 비슷하도록 보장
train data와 같이 클래스 간의 비율이 크게 다를 때 유용하게 사용
일반화 성능을 상승
각 클래스의 샘플을 충분히 반영해 모델의 편향 방지

04 Ensemble

Ensemble 전략 변화

처음에 k-fold로 5개 모델 학습 후 앙상블 진행

5개중에 2개 이상 모델이 True로 예측 할 경우 최종 예측을 True로 판단 (threshold = 0.4)

10개 중에 5개로 바꾼 이유 : 예측 결과의 신뢰도를 더 올리기 위해서 threshold를 50%로 설정

5개 모델을 쓸 경우 40%와 60%만 가능 → 10개로 늘려서 50%로 조정 가능

```
# 앙상블을 통한 최종 예측/10개 모델 중 5개 이상 모델이 True로 예측할 경우
final_test_preds = []

for i in range(len(test_preds[0])):
    combined_test_preds = [test_preds[j][i] == 'True' for j in range(len(test_preds))]
    num_ones = sum(combined_test_preds)  # True로 분류된 모델의 개수를 계산
    if num_ones >= 5:  # 5개 이상의 모델이 True로 분류했을 경우
        final_test_preds.append(True)   # 최종 예측을 Treu로 예측
    else:
        final_test_preds.append(False)  # 그 외의 경우에는 False로 예측
```

Model Selection & Validation

3. 결과 및 제언

01 최종 결과

02 느낀 점

03 추가적인 성능향상 방법

결과 및 제언

최종 결과

Final Score

	Public	Final
F1 score	0.761333	0.792754
리더보드	145위	15위(+30위)

느낀 점

- **1. 전처리의 필요성**

  개별적으로 인코딩을 시도하려고 했으나 너무 많은 경우의 수가 존재하고 컬럼이 너무 많아서 최적의 방법을 찾기 어려움 -> 캣부스트라는 모델을 통해 해결

- **2. 자체 검증에 대한 확신**

  제출 직전 순위는 안정권 밖이였으나 , 이론적으로 전처리를 진행하고 캣부스트의 파라미터 튜닝을 진행한다면 final score오를 것으로 예상, 30위 상승 의 결과

추가적인 성능향상 방향 (오프라인 해커톤)

1. 전처리 시 unkown처리 후 인코딩

2. 아직 해결하지 못한 가중치에 대한 해석

3. translate모델 사용 for 대용량 데이터 처리
   컬럼의 딕셔너리 매핑방식을 대체
   - category, position, job 등

. pandas 기능을 활용한 EDA 참고자료

MQL 데이터 기반

B2B 영업기회 창출 예측 모델 개발

감사합니다

영업 챔피언스 | 김이정 오인우 이세희 박주현

LG