

B2B 분석 보고서

lg aIMERS 분석

**팀원**

이재성(202184042 ICT공학부)

김성중(201904192 ICT공학부)

전병준(201904236 데이터사이언스)

우사랑(201904217 데이터사이언스)

**강의명**

기계학습과 특징추출

**작성 날짜**

2024.06.06

**목 차**

**I. 개요** **2**

1. LG AIMERS 문제 설명 2

**II. 데이터 수집** **2~3**

1. 데이터 출처 2

2. 수집된 데이터의 구조 및 특성 설명 3

**III. 데이터 전처리** **3~5**

1. 결측치 처리 3

2. 범주화 처리 4

3. 자연어 처리 5

**IV. 탐색적 데이터 분석** **6~11**

1. 데이터 분포 확인 6

2. 상관관계 분석 7

3. Feature Engineering(컬럼 생성) 8

4. 중요 변수 선택 9~11

**V. Modeling** **11~15**

1. 모델 선정 11

2. 선정된 모델 파라미터 조정 12

3. Ensemble (Voting) 13  
4. 결론 13~14

**참고문헌** **14**

**I. 개요**

1. LG Aimers문제 설명  
최근 기업은 MQL(Marketing Qualified Leads)를 생성하고 관리함으로써 잠재

고객을 찾고 모니터링하여 이를 영업의 기회로 전환하고자 합니다.  
영업 기회 전환 고객 선별을 위해 AI모델을 개발하여 미래 영업 기회를 예측하여 데이터에 기반한 의사 결정을 강화하는 것이 목적입니다.

따라서 MQL 데이터를 활용하여 영업 기회 전환을 예측하는 AI모델을 구현하고 그 성능을 비교하는 것이 목접입니다.

**II. 데이터 수집  
1. 데이터 출처**

2024 LG Aimers 4기를 참여해서 받은 데이터를 사용하였습니다.  
Test data는 LG Aimers에 문의 해보았지만, 받을 수 없다는 연락을 받고, Train데이터를 8:2비율로 나누어 Train과 Test를 직접 만들어 진행하였습니다.

**텍스트, 메뉴, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명2. 수집된 데이터의 구조 및 특성 설명**수집된 데이터는 [28 \* 47439]형태로 되어있습니다. 수치형 컬럼은 13개, 범주형 컬럼은 15개가 있으며, bool형태 1개로 구성되어 있습니다. Target은 ‘is\_converted’컬럼으로 bool형태로 영업성공에 성공/실패를 기록해둔 것 입니다.

이번 주요 과제는   
1. 불균형데이터에 대한 처리   
2. 결측치에 대한 처리  
으로 중점을 두고 분석을 진행하였습니다.

**III. 데이터 전처리**1. 결측치 처리  
(1) customer\_country

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명response\_corporate 컬럼이 LG의 해외지사 코드인 것을 이용하여, 나라의 전처리를 진행해주었습니다. 컬럼안에 중국어, 영어, 히브리어, 결측치 등 처리가 필요하여 왼쪽에 있는 지사 명을 사용하여 컬럼의 전처리하여 주었습니다.

(2) business\_unit  
train에서는 존재하지만, test에서는 존재하지않는 value값을 최빈값으로 전처리를 진행하였습니다.

(3) ver\_win\_ratio\_per\_bu  
business\_area와 business\_unit에 따라 ver\_win\_ratio\_per\_bu의 평균값을 구하여 해당 결측치를 채워주었습니다.

(4) historical\_existing\_cnt

Is\_converted가 True일 때, historical\_existing\_cnt의 평균값, Is\_converted가 False일 때, historical\_existing\_cnt의 평균값으로 결측치를 채워주었습니다.

(5) com\_ver\_win\_rate

위에서 진행한 처리방법과 동일하게 is\_converted의 값에 따라 평균값으로 결측치를 채워주었습니다.

**2. 범주화 처리**

(1) inquiry\_type, customer\_type, product\_category

<https://www.lg-informationdisplay.com/support/inquiry> 위의 사이트에서 신청서와 주어진 데이터와 유사하다는 것을 이용하여, 신청서를 기준으로 전처리를 진행하였습니다. 텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

**3. 간단한 자연어 처리**

**텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

1. ‘\_’ 제거  
   ‘\_’때문에 같은 value를 다른 value로 인식하는 것을 수정
2. 공백제거  
   공백도 ‘\_’와 같은 이유로 전처리 진행
3. 의미없는 value 결측치로 변환   
   일정에 관한 내용이 아닌 다른 내용이 들어가 있는 내용을 결측치로 변환

**IV. 탐색적 데이터 분석  
1. 데이터 분포 확인**

<데이터 분포>  
원, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명  
우리가 예측해야할 타겟 변수는 is\_converted이다. 위 파이 차트를 확인해보면, 데이터가 불균형하다는 것을 알 수 있다. False(91.9%)에 비해 True(8.1)의 비율이 매우 적다. 이는 모델링 시 불균형 데이터에 대한 처리가 필요하다는 것이다.

<결측치 확인>  
텍스트, 스케치, 책, 흑백이(가) 표시된 사진

자동 생성된 설명  
위 그림의 하얀색 부분이 결측치가 있는 분이다. 결측치가 아주 많다는 것을 알 수 있다. 총 컬럼의 갯수는 28개이고, 결측치가 없는 컬럼은 11개이다. 하지만 ‘NA’, ‘NONE’ 등 문자열로 되어있는 결측치가 존재할 수 있으므로 확인 후 처리를 진행해야 할 것이다. 결측치 처리 방법으로는 삭제, 다른 겂으로 대체 등으로 각 컬럼의 특성에 맞추어 처리를 해야 한다.

2. 상관관계 분석

<상관관계 분석>  
텍스트, 스크린샷, 패턴, 사각형이(가) 표시된 사진

자동 생성된 설명  
위 그림은 변수들 간 상관관계를 나타낸 그림이다. 우리가 주의깊게 보아야 할 부분은 타겟변수인 is\_converted와 다른 변수들 간의 상관관계이다. com\_reg\_ver\_win\_rate, lead\_cor, win\_ratio가 비교적 높은 상관관계가 나왔고 각각의 상관계수는 0.58, 0.62, 0.69이다. res\_cor, cus\_cor, cus\_res는 높지는 않지만 각각 0.21, 0.28, 0.33으로 약간의 상관관계가 있는 것으로 나타났다.

3. Feature Engineering(컬럼 생성)  
(1) as\_strategic\_ver  
id\_strategic\_ver, it\_strategic\_ver는 business\_unit컬럼에서 id or it일때, 특정사업 영역에 대해 가중치를 부여한 컬럼이다. 두 개의 컬럼을 이용하여 특정사업영역을 특정지을 수 있었고, business\_unit이 as일 때 특정사업영역에 대해서 가중치를 부여하여 주었습니다.

(2) res\_cor  
response\_corporate컬럼은 결측치가 없는 컬럼 중 하나입니다. True의 예측을 더 잘할 수 있도록 response\_corporate의 value값에 대한 True 비율을 측정하여 새로운 res\_cor컬럼을 만들어 주었습니다.

(3) cus\_cor  
res\_cor과 동일한 목적으로 customer\_idx컬럼 value값에 대한 True 비율을 측정하여 새로운 cus\_cor 컬럼을 만들어 주었습니다.

(4) cus\_res  
cus\_cor 과 res\_cor 을 곱하여서 cus\_res컬럼을 만든 것입니다. Overfiting을 방지하기 위해서 두 가지의 True비율을 곱하여 컬럼을 만들어 주었습니다.  
  
(5) lead\_cor  
lead\_owner컬럼에서 True비율을 구하여 만들어준 컬럼입니다. Lead\_owner컬럼은 3가지 모델의 피쳐중요도에서 상위권에 있는 것을 보여주었습니다. 중요한 컬럼으로 인지하고, True비율을 계산하여 주었습니다.

(6) win\_ratio  
com\_reg\_ver\_win\_rate, ver\_win\_rate\_x, ver\_win\_ratio\_per\_bu의 3가지 컬럼은 영업전환에 관한 비율을 기록하고 있는 컬럼입니다. 하지만 결측지가 너무 많아 제대로 사용을 할 수 가 없는 상황이라서 3가지의 컬럼의 True 비율을 계산하고, 더하여 새로운 win\_ration 컬럼을 만들어 주었습니다.

4. 피쳐 중요도 확인

<피쳐 중요도 확인>  
- LGBM  
텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명  
lgbm의 경우 다른 두 모델보다 많은 변수들이 주요 변수들로 작용했다. 다수의 변수를 효과적으로 활용하여 예측 성능을 높였고, 이로 인해 더 많은 변수가 모델의 예측에 기여하게 되었다. 이는 모델의 복잡도를 높이지만 성능을 향상시키는 요인으로 작용 될 것으로 보인다.

- XGB  
텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명  
xgb의 경우 다른 두 모델보다 적은 변수가 주요 변수로 작용했다. 이는 모델의 복잡도를 줄이고 모델의 해석 용이성을 높이는 데에 도움을 준다.

- CATBoost  
텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명  
catboost의 경우 적절한 수의 변수가 주요 변수로 작용한 것을 볼 수 있다. 이는 모델의 복잡도를 조절하고, 모델의 효율성을 높였다고 볼 수 있다.

<주요 변수>  
위 과정을 통해 피쳐 중요도를 알아보았다. 이 과정에서 공통적으로 보이는 컬럼은 historical\_existing, win\_ratio, com\_reg\_ver\_win\_rate, lead\_cor, ver\_win\_rate\_x, ver\_win\_ratio\_per\_bu가 있다. 이 변수들은 타겟 변수인 is\_converted를 예측하는 데에 중요한 역할을 한다는 것을 의미한다.

**V. Modeling**1. 모델 선정  
모델은 데이터 공모전에서 우수한 성적을 거두는 Boosting모델 3가지를 사용하였습니다. LGBM, XGB, Catboost 모델을 사용하였습니다. 평가지표는 분류모델에서 자주 사용하는 F1-score로 선정하였고, 혼동행렬을 같이 보면서 모델링을 진행하였습니다.

**텍스트, 폰트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명**

왼쪽부터 순서대로 LGBM, XGB, Catboost 모델입니다. 3가지 모델의 F1\_Score점수를 확인한 결과 LGBM가 제일 좋은 성적을 거두었지만, 3가지 모델 전부 0.97이라는 수치로 좋은 성적을 거둔 것을 볼 수 있습니다.

2. 파라미터 설정  
모델의 파라미터를 default 값으로 진행하였기 때문에 optuna를 진행하여 파라미터를 선정하였습니다.  
random\_state : 랜덤 시드를 설정하여 결과의 재현성을 보장합니다.  
learning\_rate : 학습률로 모델이 업데이터되는 속도를 조절합니다.  
n\_estimators : 학습 반복 횟수를 설정합니다. 값이 클수록 오래 학습하고 복잡한 모델을 만듭니다.  
objective : ‘binary’로 설정하여 이진분류문제로 풀 수 있겠금 설정합니다.  
metric : ‘binary\_logloss’ 평가지표를 이진 분류 문제에서 사용하는 로그 손실로 설정합니다.  
reg\_alpha : L1 정규화 강도를 설정합니다. 정규화 강도를 설정하여 overfiting을 방지합니다.  
reg\_lambda : L2 정규화 강도를 설정합니다. 동일하게 overfiting을 방지합니다.  
max\_depth : 트리의 최대 깊이를 설정합니다. 모델의 복잡성을 의미합니다.  
num\_leaves : 하나의 트리가 가질 수 있는 최대 리프 노드 수를 의미합니다. 동일하게 모델의 복잡성을 의미합니다.  
colsample\_bytree : 트리를 만들 때 사용할 피처의 비율을 설정합니다.  
subsample : 각 트리를 만들 때 사용할 데이터 샘플의 비율을 설정합니다.  
subsample\_freq : subsample파라미터를 사용할 빈도를 설정합니다.  
min\_child\_samples : 리프 노드가 최소한으로 가져야 하는 샘플 수를 설정합니다.  
max\_bin : 피처를 이산화할 최대 빈의 수를 설정합니다.  
위의 설명을 토대로 파라미터값을 조정하여 optuna를 설정하여 최적의 파라미터 값을 찾았습니다.

텍스트, 스크린샷, 폰트, 디자인이(가) 표시된 사진

자동 생성된 설명3. Ensemble(voting)  
모델의 F1-score를 더 향상시키기 위해서 3가지의 Boosting모델을 voting하였습니다. voting할 때, 불균형데이터에 대한 문제를 해결하고자 StratifiedKFold를 진행하여 데이터를 균형있게 분리하여 학습할 수 있도록 진행하였습니다.  
텍스트, 폰트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명

왼쪽의 그림은 Catboost 모델의 혼동행렬과 F1스코어입니다. 오른쪽그림은 Voting을 진행한 모델의 혼동행렬과 F1스코어입니다. 보이는 것과 같이 FP, FN의 수치가 줄어든 것을 볼 수 있고, F1스코어가 0.0012 향상한 것을 볼 수 있습니다.

4. 결론  
voting한 모델로 Test를 예측하였을 때, 결과는 F1스코어가 0.26510로 좋지 못한 것을 볼 수 있다. 과대적합의 문제가 일어났을 것으로 생각하며, 성능이 두번째로 좋았던 모델인 LGBM을 사용하여 Test를 예측해보았지만 동일하게 성능은 0.4148로 과대적합이 일어난 것을 볼 수 있었다.   
이에 컬럼중요도의 그래프를 보면 훈련데이터의 True 비율을 가지고 만든 컬럼들이 상위권에 있다는 것을 볼 수 있는데, 오히려 이것이 Train에 과대적합한 모델을 만든 것이라고 생각을 한다. 그리고 Test데이터를 받지 못하여 Train을 split하여 만든 영향도 있다고 생각한다.

**참고문헌**

LG Information Display, “LG Inquiry to Buy”, <https://www.lg-informationdisplay.com/support/inquiry>, (2024.04.20)

사업장 | LG전자, “LG전자 사업장”, <https://www.lge.co.kr/company/info/overseas>, (2024.4.20)