MQL 데이터 기반 B2B 영업기회 창출 예측 모델 개발

LLAMA

고은경

박효서

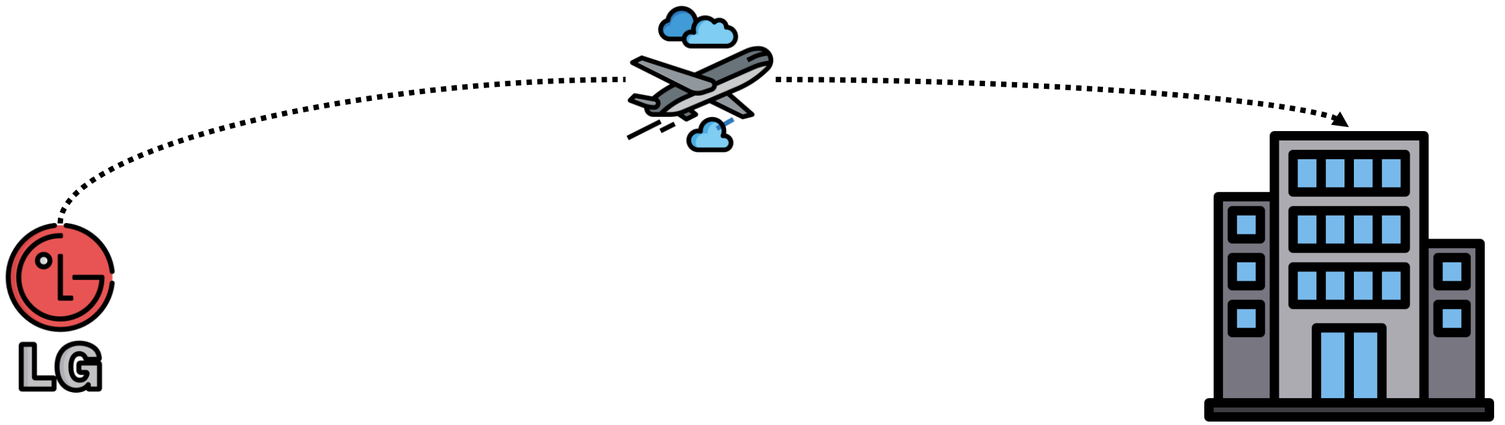
최락원

정상일

이웅희

문제 정의

- 해당 대회는 주어진 데이터로 고객의 영업전환여부를 예측하는 모델을 생성하는 것이다.
- LG 사의 영업사원 수는 정해져 있다. 따라서 많은 고객 기업들 중 영업전환 가능성이 높은 고객 기업을 대상으로 중요도를 두고 영업사원을 파견해야 한다.



INDEX

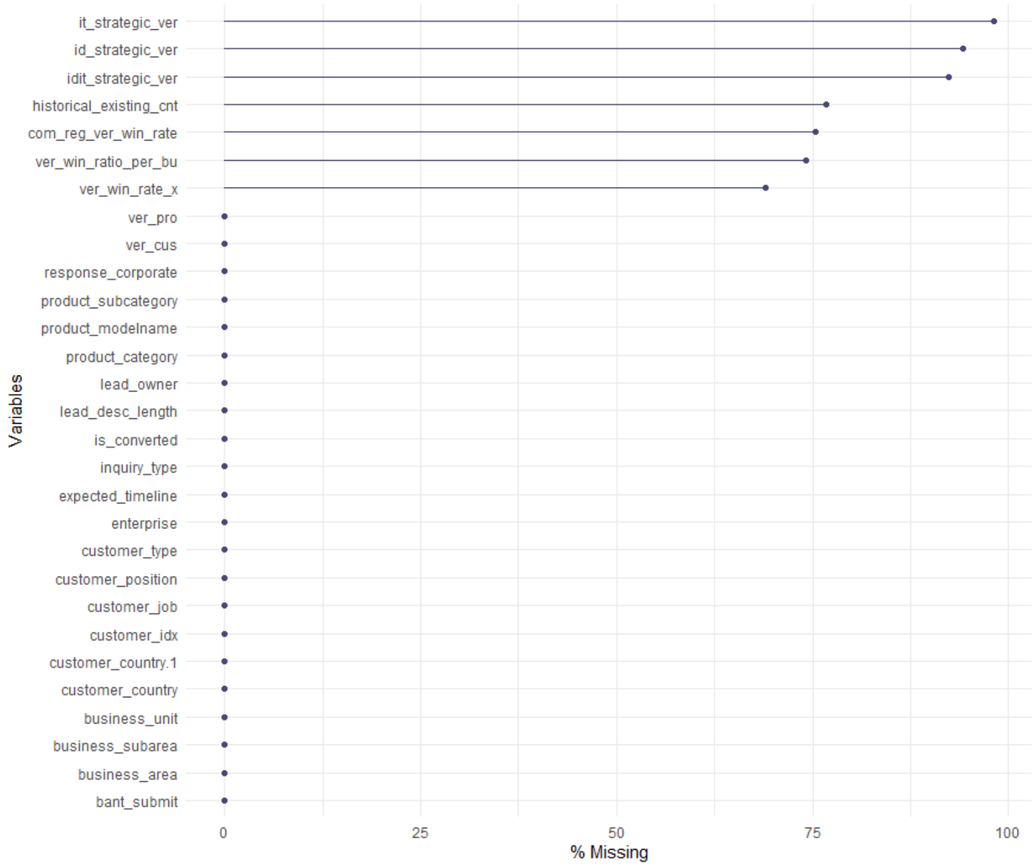
1. 데이터 탐색

(변수 정리)

Field(num)	설명	Field(cat)	설명
bant_submit	MQL 구성 요소들 중 [1]Budget(예산), [2]Title(고객의 직책/직급), [3]Needs(요구사항), [4]Timeline(희망 납기일) 4가지 항목에 대해서 작성된 값의 비율	customer_country	고객의 국적
`com_reg_ver_win_rate`	Vertical Level 1, business unit, region을 기준으로 oppty 비율을 계산	business_unit	MQL 요청 상품에 대응되는 사업부
customer_idx	고객의 회사명	customer_type	고객 유형
`historical_existing_cnt`	이전에 Converted(영업 전환) 되었던 횟수	enterprise	Global 기업인지, Small/Medium 규모의 기업인지
`id_strategic_ver`	• 변경후: (도메인 지식) 특정 사업부(Business Unit이 ID일 때), 특정 사업 영역(Vertical Level1)에 대해 가중치를 부여	customer_job	고객의 직업군
`it_strategic_ver`	• 변경후: (도메인 지식) 특정 사업부(Business Unit이 IT일 때), 특정 사업 영역(Vertical Level1)에 대해 가중치를 부여	inquiry_type	고객의 문의 유형
`idit_strategic_ver`	Id_strategic_ver이나 it_strategic_ver 값 중 하나라도 1의 값을 가지면 1 값으로 표현	product_category	요청 제품 카테고리
lead_desc_length	고객이 작성한 Lead Descriptoin 텍스트 총 길이	product_subcategory	요청 제품 하위 카테고리
`ver_cus`	특정 Vertical Level 1(사업영역) 이면서 Customer_type(고객 유형)이 소비자(End-user)인 경우에 대한 가중치	product_modelname	요청 제품 모델명
`ver_pro`	특정 Vertical Level 1(사업영역) 이면서 특정 Product Category(제품 유형)인 경우에 대한 가중치	customer_country.1	담당 자사 법인명 기반의 지역 정보(대륙)
`ver_win_rate_x`	전체 Lead 중에서 Vertical을 기준으로 Vertical 수 비율과 Vertical 별 Lead 수 대비 영업 전환 성공 비율 값을 곱한 값	customer_position	고객의 회사 직책
`ver_win_ratio_per_bu`	특정 Vertical Level1의 Business Unit 별 샘플 수 대비 영업 전환된 샘플 수의 비율을 계산	response_corporate	담당 자사 법인명
lead_owner	영업 담당자 이름	expected_timeline	고객의 요청한 처리 일정
		business_area	고객의 사업 영역
is_converted	영업 성공 여부. True일 시 성공.	business_subarea	고객의 세부 사업 영역

1. 데이터 탐색

(전체 결측치 확인)



- 전체적으로 결측치가 많은 변수가 존재한다.
- 시각화 결과 외에도 범주형 변수의 '공백' 이나 수치형 변수의 0등 명시적이지 않은 결측치도 존재한다.

→ 기본적으로 수치형 변수의 경우 '0'으로, 범주형 변수의 경우 '공백 ' 으로 결측치를 처리한다.

(예외적인 사항들은 후술한다)

2. 결측치 처리

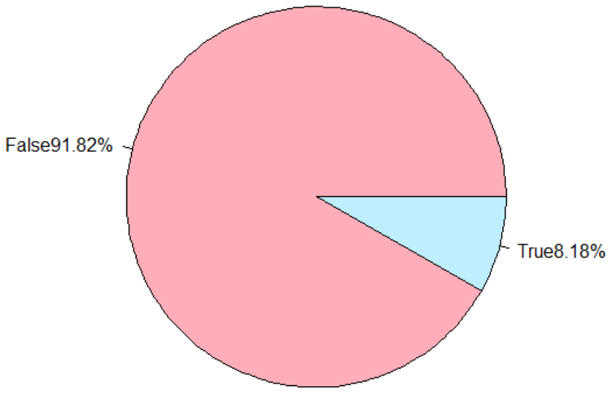
(결측치가 너무 많은 행 삭제)

- 데이터에 많은 결측치가 존재하며 단순 제거 시 많은 데이터가 소실.
- Target 변수의 True와 False의 불균형이 심함.

영업 전환이 False이면서 결측치가 많은 행은 학습에 도움이 되지 않는다 판단.

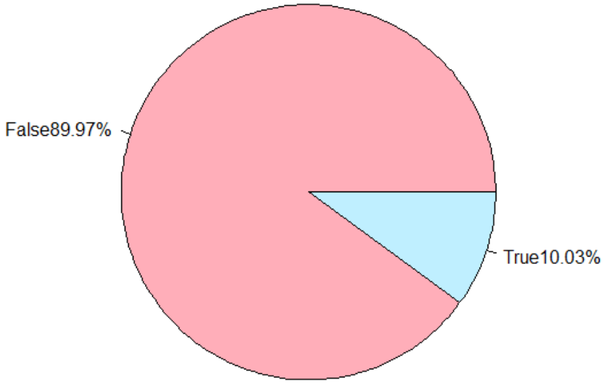
Target 변수가 False인 데이터를 대상으로 row 별(axis=1) 결측치의 비율이 30%가 넘는 경우 해당 행을 삭제

Before



is_converted	
False	True
54449	4850

After

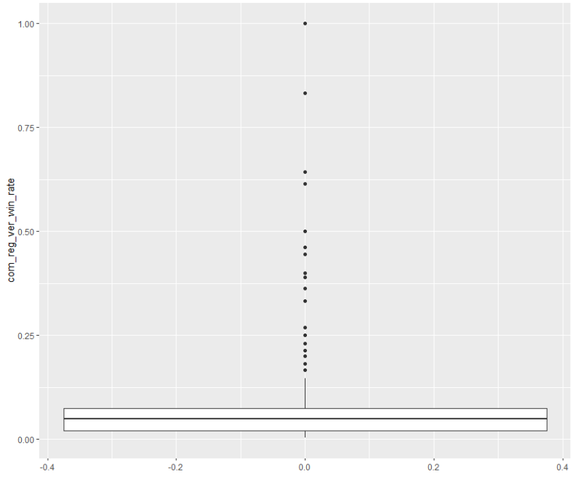


is_converted	
False	True
40062	4464

→ 비율적으로는 큰 차이가 없으나 의미 없는 row를 제거하여 모델 성능과 학습의 효율을 향상시킨다.

2. 결측치 처리

(com_reg_ver_win_rate)



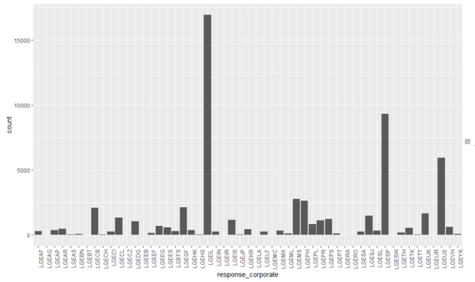
0.0037878787878787	0.0039370078740157	0.004	0.0109890109890109	0.0118577075098814	0.0135135135135135
713	803	434	104	423	242
0.0151515151515151	0.0169491525423728	0.0172413793103448	0.0175438596491228	0.0181818181818181	0.0196078431372549
118	221	63	96	65	107
0.0199004975124378	0.0202020202020202	0.0227272727272727	0.025	0.0289256198347107	0.0289855072463768
316	110	58	54	253	116
0.0311958405545927	0.032258064516129	0.0327868852459016	0.0330578512396694	0.036036036036036	0.037037037037037
609	106	422	128	119	75
0.04	0.0408163265306122	0.0416666666666666	0.0422535211267605	0.043103448275862	0.0434782608695652
276	155	89	100	153	54
0.0446428571428571	0.0476190476190476	0.0485436893203883	0.0491803278688524	0.0496894409937888	0.0531914893617021
292	141	109	214	172	99
0.0538922155688622	0.0544217687074829	0.0555555555555555	0.0575342465753424	0.0666666666666666	0.0677966101694915
193	175	22	399	216	240
0.0681818181818181	0.0695652173913043	0.0714285714285714	0.0732484076433121	0.0749486652977412	0.075
103	251	82	791	1130	274
0.0806916426512968	0.0833333333333333	0.0843373493975903	0.0869565217391304	0.0888888888888888	0.105263157894737
410	17	166	62	120	31
0.113636363636364	0.116279069767442	0.118421052631579	0.11864406779661	0.124121779859485	0.125
57	71	141	118	431	13
0.136363636363636	0.147058823529412	0.166666666666667	0.181818181818182	0.2	0.214285714285714
60	46	17	33	5	60
0.230769230769231	0.25	0.269230769230769	0.333333333333333	0.363636363636364	0.390243902439024
40	16	44	110	13	95
0.4	0.444444444444444	0.461538461538462	0.5	0.615384615384615	0.642857142857143
12	12	16	34	24	782
0.833333333333333	1	NA's			
17	10	44731			

- com_reg_ver_win_rate은 business_area와 business_unit별 영업전환율을 의미한다.

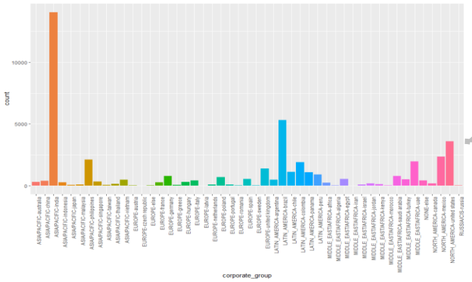
결측치가 아닌 영업전환율을 business_area와 business_unit 별로 구분하고, 결측치를 해당하는 기준의 최빈값으로 채운다.

3. 범주형 변수 처리

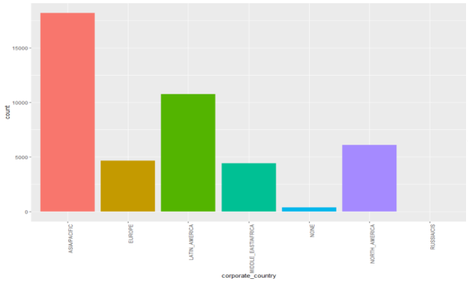
(희소 범주 처리)



<홈페이지를 통한 지사별 매핑>



더 큰 범주(국가별)로 매핑



- 많은 범주형 변수들이 unique한 범주를 포함한다.

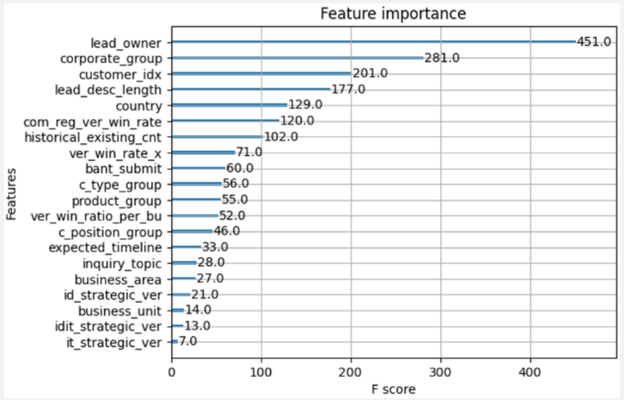
- 최대한 도메인을 이용하여 unique한 범주를 더 큰 범주에 매핑한다.
- 대표적인 예로 좌측의 그림은 response_group을 LG 전자의 홈페이지를 참고하여 각 지사에 맞게 매핑한 결과이다.

→ 세부 범주를 도메인을 통해 더 큰 범주에 매핑하여 더 큰 데이터의 패턴에 편승하도록 하였기 때문에 private 데이터에서 더 강건한 모델을 형성할 수 있다.

4. 파생변수 생성

(historical_true_mean)

문제 인식 & 파생변수 생성



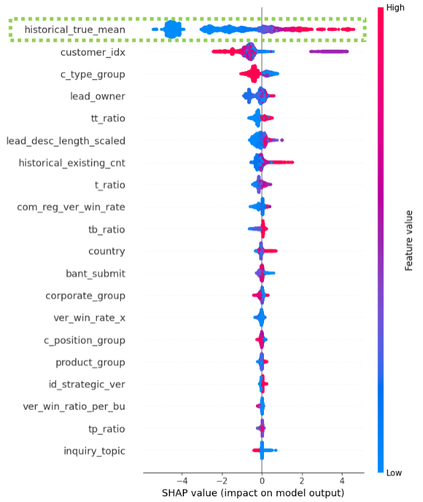
- lead_owner(고객 기업) 변수의 중요도가 크기에 이를 활용하여 파생변수 생성을 시도할 수 있다.

$\frac{True\ his\_cnt}{Total\ his\_cnt}$ $\frac{False\ his\_cnt}{Total\ his\_cnt}$

	lead_owner	count	true_cnt	false_cnt	historical_true_mean	historical_false_mean
0	0	338	20	318	0.059172	0.940828

- lead_owner별로 historical_existing_cnt의 is_converted 비율을 계산한다.
  → 고객 기업별 이전에 영업전환의 성공/실패의 여부를 비율로 계산함을 의미한다.
  → True의 비율을 historical_true_mean으로, False의 비율을 historical_false_mean을 계산한다.
  → 최종적으로 historical_true_mean을 파생변수로 추가한다.
- $historical\_true\_mean = \frac{true\_cnt}{true\_cnt+false\_cnt}$
- $historical\_false\_mean = \frac{false\_cnt}{true\_cnt+false\_cnt}$

결과 확인



- SHAP을 통해 변수 중요도를 확인한 결과 파생변수(historical_true_mean)의 중요도가 가장 높은 것을 확인할 수 있다.

5. 모델링

모델 후보군

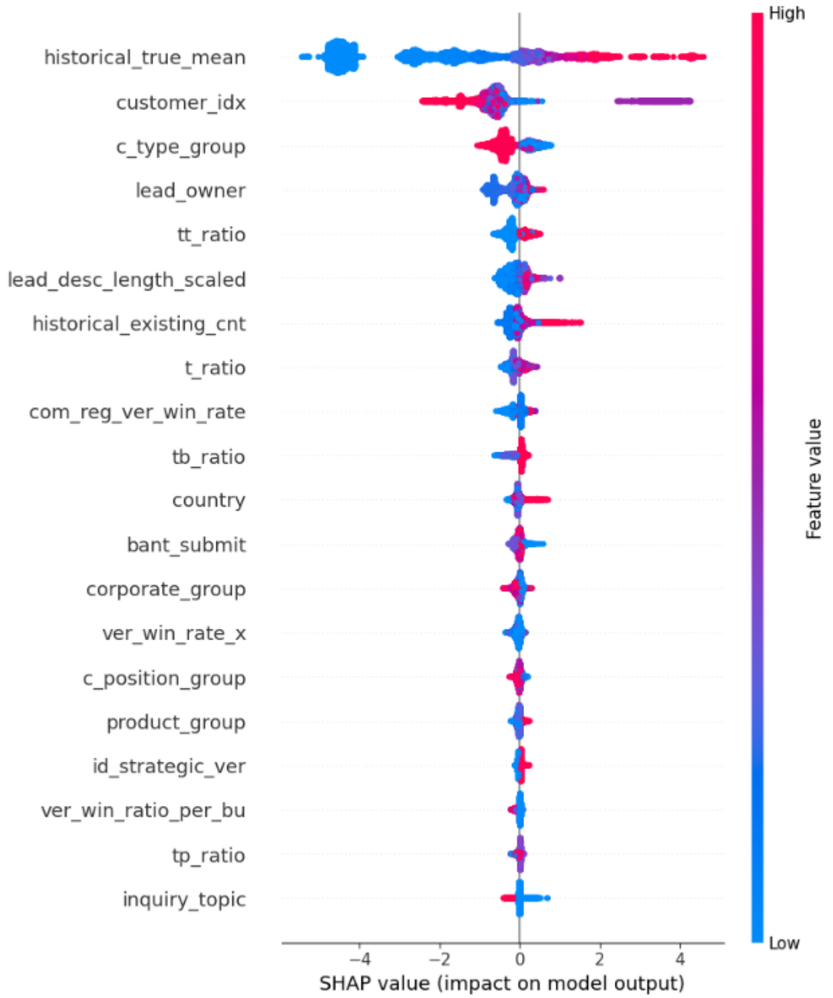
• Decision tree	대회에서 주어진 base model. 직관적이고 해석이 용이하다는 장점이 있으나 bagging과 boosting 계열의 모델에 비해 성능이 좋지 못하다는 단점이 존재한다.
• Random forest	Bagging을 활용한 tree 계열 모델. decision tree 보다 성능이 높고, bagging을 통해 과적합을 방지할 수 있다는 장점이 있다. 그러나 희소 데이터에서 성능이 떨어지는 단점이 있다.
• LGBM	Leaf-wise 방식으로 tree를 확장하기 때문에 XGB 모델모다 속도가 빠른 장점이 있다. 그러나 현제 데이터는 약 50,000 개로 속도가 갖는 이점이 크지 않다.
• CatBoost	범주형 변수가 많을 때 장점을 갖는 boosting 계열의 모델이다. 다만 현제 EDA를 완료한 우리 팀의 데이터에 적합하지 않아 XGB보다 성능이 떨어진다.

최종 모델

• XGB	a. Tree 계열 모델 중 boosting을 사용한 모델로 가장 성능이 좋으며 현제 문제에서도 가장 고성능을 보이기에 최종 모델로 채택. b. 파라미터는 tree의 depth와 leaf와 관련된 튜닝 외에 큰 변경을 하지 않음. c. train set과 test set에서 모델 성능의 편차가 존재하기에 파라미터 최적화를 하지 않음. d. 위와 같은 이유로 앙상블(ensemble) 또한 진행하지 않았으며 단일 모델로 구현.

6. Feature Importance using SHAP

& Total summary



SHAP을 해석하는 법

- 변수의 중요도가 클수록 상단에 배치된다.
- 붉은색에 가까운 것은 변수의 값이 큼을, 보라색은 중간을, 파란색은 작은 값을 의미한다.

  예를 들어 historical_true_mean은 값이 클수록(붉은색) True로 분류되고, 작을수록(파란색) False로 분류된다.

Summary

- customer_idx와 lead_owner는 연속형 변수(int)이다. 연속형 변수가 분류 기준으로 자주 사용되는 것은 맞으나, 분류에도 유의미한 영향을 주는 것으로 보아 고객 영업전환에 있어 대상 고객과 해당 고객이 속한 기업이 영업전환과 연관성이 높음을 알 수 있다.
- 도메인을 활용하여 모델을 강건하게 만들고자 하였다. 결과적으로 public data와 private data에서 모두 0.75로 강건한 모델을 만들 수 있었다.
- 아쉬운 점은 customer_idx와 lead_owner를 통한 파생변수를 더 고려하여 부족한 정보량을 채울 수 있었을 것이라 생각한다.

LLAMA.