LG Aimers Phase2

MQL데이터를 활용하여 영업 기회 전환 고객을 선별하기 위한 AI모델 개발

송준호, 한지성, 황성주

2024.02. 27

contents

1. 대회 소개
2. EDA 및 데이터 전처리
3. 모델링
4. 결과

1. 대회 소개

[설명]
MQL데이터를 활용하여 영업 기회 전환 고객을 선별하기 위한 AI모델 개발합니다.
온라인 해커톤에서 교육생들의 문제 해결 능력을 검증하여 오프라인 해커톤에 진출할 약 100명을 선발하기 위한 과정입니다.
단, 오프라인 해커톤) 진출할 인원이 100명 미달 시, 추가 선발 가능
오프라인 해커톤은 1박 2일간 오프라인으로 진행되며, 온라인 해커톤과 주제는 동일합니다.

[주최 / 주관]
주최 : LG AI Research
주관 : 엘리스그룹
참여 : 한경닷컴

[리더보드]
평가 산식 : F1 score
Public score : 전체 테스트 데이터 샘플 중 사전 샘플링된 50%로 계산
Private score : Public score 계산에 포함되지 않은 나머지 50%의 테스트 데이터로 계산

2. EDA & 데이터 전처리

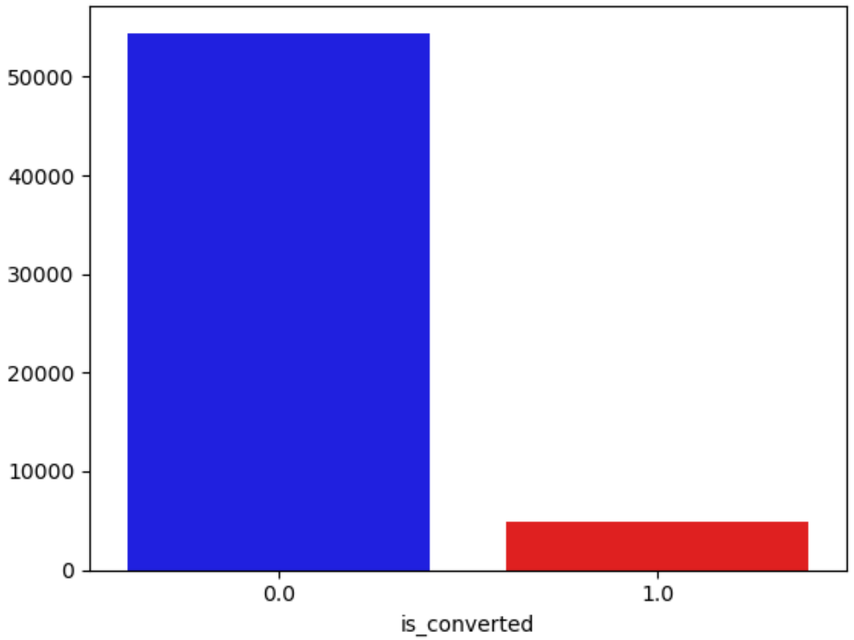
데이터 정보확인

```
df_all.isnull().sum()
```

```
bant_submit                 0
customer_country          982
business_unit               0
com_reg_ver_win_rate    48214
customer_idx                0
customer_type           45418
enterprise                  0
historical_existing_cnt 49539
id_strategic_ver        60533
it_strategic_ver        63396
idit_strategic_ver      59359
customer_job            20172
lead_desc_length            0
inquiry_type             2233
product_category        21232
product_subcategory     54542
product_modelname       54779
customer_country.1        982
customer_position           0
response_corporate          0
expected_timeline       33271
ver_cus                     0
ver_pro                     0
ver_win_rate_x          43780
ver_win_ratio_per_bu    47360
business_area           43780
business_subarea        57228
lead_owner                  0
is_converted             5271
id                      59299
dtype: int64
```

```
df_all.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 64570 entries, 0 to 5270
Data columns (total 30 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   bant_submit             64570 non-null  float64
 1   customer_country        63588 non-null  object
 2   business_unit           64570 non-null  object
 3   com_reg_ver_win_rate    16356 non-null  float64
 4   customer_idx            64570 non-null  int64
 5   customer_type           19152 non-null  object
 6   enterprise              64570 non-null  object
 7   historical_existing_cnt 15031 non-null  float64
 8   id_strategic_ver        4037 non-null   float64
 9   it_strategic_ver        1174 non-null   float64
 10  idit_strategic_ver      5211 non-null   float64
 11  customer_job            44398 non-null  object
 12  lead_desc_length        64570 non-null  int64
 13  inquiry_type            62337 non-null  object
 14  product_category        43338 non-null  object
 15  product_subcategory     10028 non-null  object
 16  product_modelname       9791 non-null   object
 17  customer_country.1      63588 non-null  object
 18  customer_position       64570 non-null  object
 19  response_corporate      64570 non-null  object
 20  expected_timeline       31299 non-null  object
 21  ver_cus                 64570 non-null  int64
 22  ver_pro                 64570 non-null  int64
 23  ver_win_rate_x          20790 non-null  float64
 24  ver_win_ratio_per_bu    17210 non-null  float64
 25  business_area           20790 non-null  object
 26  business_subarea        7342 non-null   object
 27  lead_owner              64570 non-null  int64
 28  is_converted            59299 non-null  object
 29  id                      5271 non-null   float64
dtypes: float64(9), int64(5), object(16)
memory usage: 15.3+ MB
```



False : 54449
True : 4850

2. EDA & 데이터 전처리

범주형 변수

```
df_all.customer_country.value_counts()
```

```
customer_country
//India                                          3055
/São Paulo/Brazil                                1376
//United States                                  1122
//United Kingdom                                  807
//Saudi Arabia                                    719
                                                  ...
/ Mato Grosso do Sul - Campo Grande / Brazil        1
600 FREMONT ST   / LAS VEGAS / United States       1
/ São Paulo/Marilia  / Brazil                      1
/ East Delhi / Saudi Arabia                        1
via a.rosario snc   / frattaminore / Italy         1
Name: count, Length: 17480, dtype: int64
```

```
df_all.customer_job.value_counts()
```

```
customer_job
engineering                     7070
other                           4876
administrative                  3666
education                       2695
sales                           2380
                                ...
facilities and operations          1
technical / decision maker         1
installation and purchaser         1
hr posting                         1
part of video wall                 1
Name: count, Length: 562, dtype: int64
```

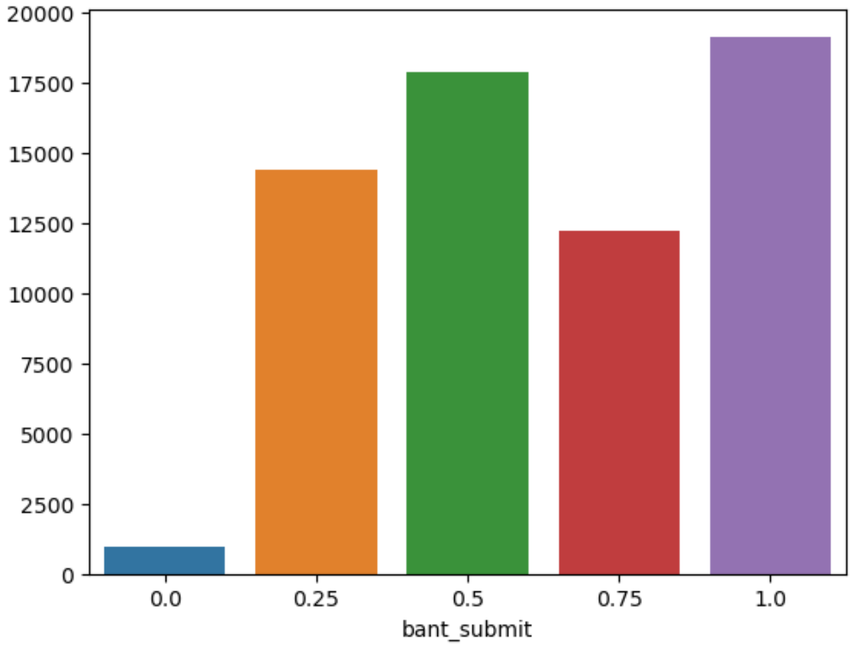
```
df_all.customer_type.value_counts()
```

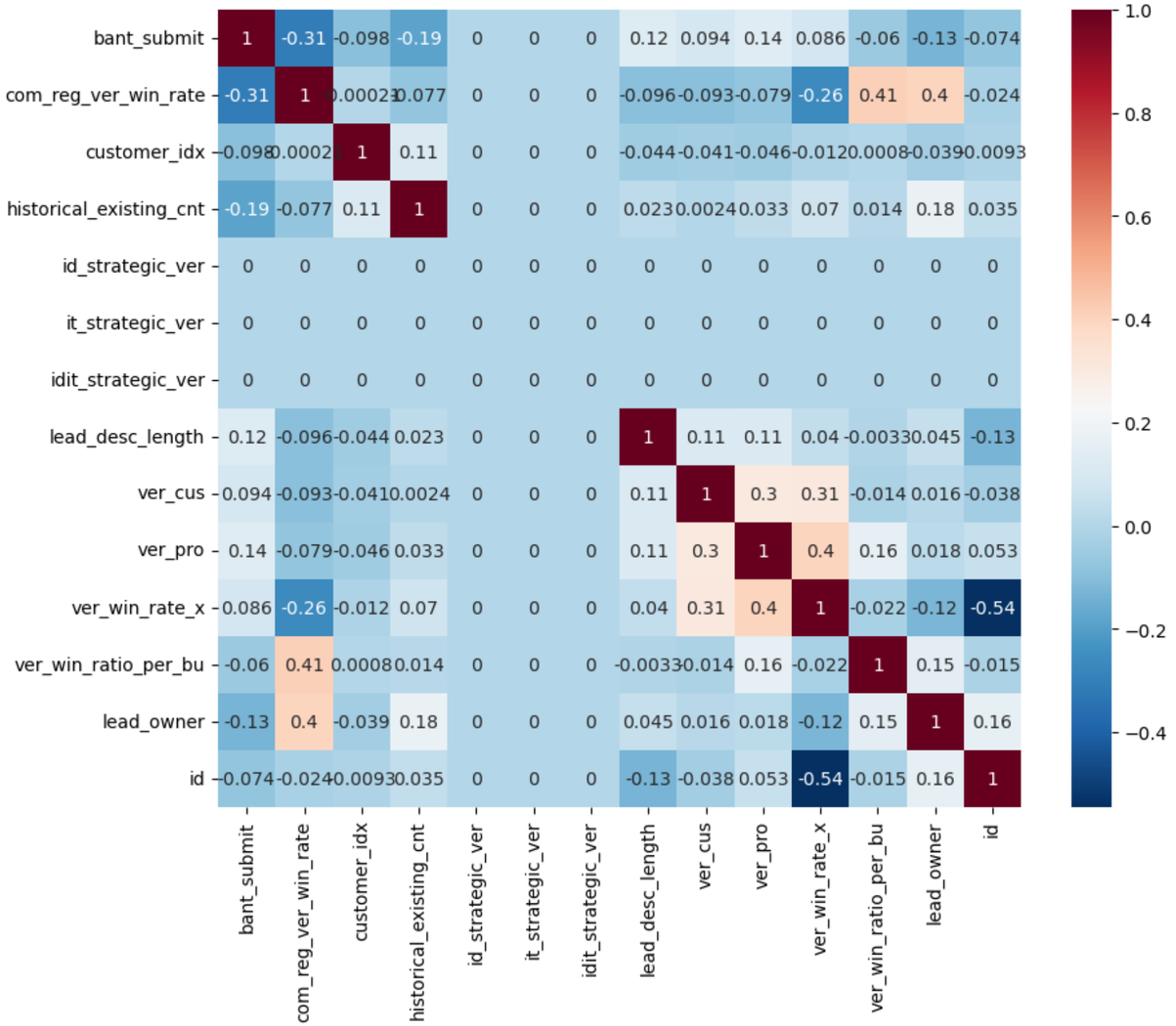
```
customer_type
End-Customer                    6648
End Customer                    6449
Specifier/ Influencer           3313
Channel Partner                 1695
Service Partner                  447
Solution Eco-Partner             292
Installer/Contractor              52
Specifier / Influencer            43
Corporate                         31
HVAC Engineer                     23
Engineer                          20
Developer                         18
Technician                        16
Consultant                        15
Home Owner                        10
Other                             10
End-user                           8
Manager / Director                 8
Software/Solution Provider         7
Etc.                               6
Reseller                           5
Homeowner                          5
Architect/Consultant               5
Interior Designer                  5
Installer                          5
Distributor                        4
Others                             4
System Integrator                  2
Dealer/Distributor                 2
Technical Assistant                1
Software / Solution Provider       1
Commercial end-user                1
Administrator                      1
Name: count, dtype: int64
```

같은 의미의 데이터가
대소문자나 다른 용어로
다르게 표현됨

2. EDA & 데이터 전처리

수치형 변수

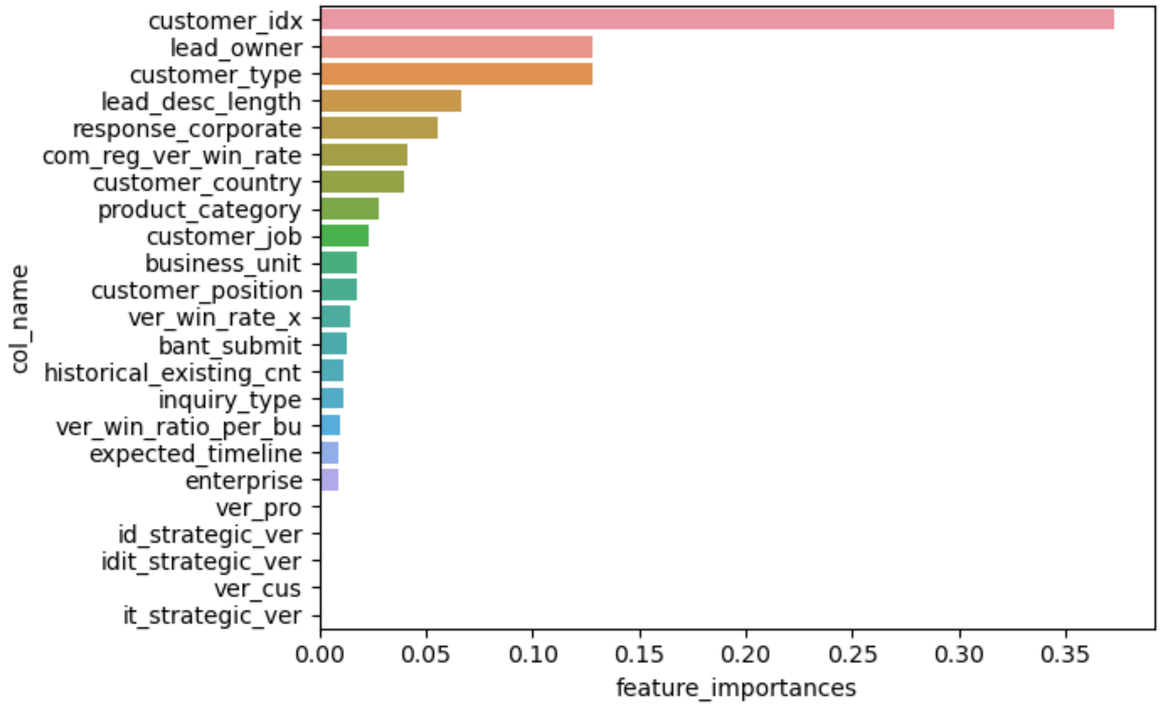




2. EDA & 데이터 전처리

1. 열 삭제

중복 변수 제거 : customer_country.1
결측치가 과반인 변수는 제거 : product_subcategory, product_modelname, business_area, business_subarea, ver_cus
변수 중요도가 매우 낮은 변수 제거 : id_strategic_ver, it_strategic_ver, idit_strategic_ver, ver_cus, ver_pro



2. EDA & 데이터 전처리

2. 같은 의미의 다른 데이터는 같은 범주로 처리

같은 의미인데 다른 용어로 된 변수 통일
ex) etc. , other, others -> etc // end-customer, end customer, end-user -> end_user

3. 개수가 1개인 범주들을 기타 처리

변수가 1개인 데이터들은 학습하면 범주의 개수가 많아져 학습 과정에 과적합이나 노이즈 등의 문제로 예측 성능에 안좋은 영향을 미칠 수 있음
개수가 1개인 범주들을 '기타' 범주로 분류

4. 결측치 처리

수치형 데이터는 변수 설명을 보고 0으로 대체해도 무방하다고 판단(비율, 이전 거래 수)
범주형 데이터는 'None' 범주로 처리

2. EDA & 데이터 전처리

범주형 변수

```
df_all.customer_country.value_counts()
```

```
customer_country
//India                                          3055
/São Paulo/Brazil                                1376
//United States                                  1122
//United Kingdom                                  807
//Saudi Arabia                                    719
                                                  ...
/ Mato Grosso do Sul - Campo Grande / Brazil        1
600 FREMONT ST   / LAS VEGAS / United States       1
/ São Paulo/Marilia  / Brazil                      1
/ East Delhi / Saudi Arabia                        1
via a.rosario snc   / frattaminore / Italy         1
Name: count, Length: 17480, dtype: int64
```

```
df_all.customer_job.value_counts()
```

```
customer_job
engineering                     7070
other                           4876
administrative                  3666
education                       2695
sales                           2380
                                ...
facilities and operations          1
technical / decision maker         1
installation and purchaser         1
hr posting                         1
part of video wall                 1
Name: count, Length: 562, dtype: int64
```

```
df_all.customer_type.value_counts()
```

```
customer_type
End-Customer                    6648
End Customer                    6449
Specifier/ Influencer           3313
Channel Partner                 1695
Service Partner                  447
Solution Eco-Partner             292
Installer/Contractor              52
Specifier / Influencer            43
Corporate                         31
HVAC Engineer                     23
Engineer                          20
Developer                         18
Technician                        16
Consultant                        15
Home Owner                        10
Other                             10
End-user                           8
Manager / Director                 8
Software/Solution Provider         7
Etc.                               6
Reseller                           5
Homeowner                          5
Architect/Consultant               5
Interior Designer                  5
Installer                          5
Distributor                        4
Others                             4
System Integrator                  2
Dealer/Distributor                 2
Technical Assistant                1
Software / Solution Provider       1
Commercial end-user                1
Administrator                      1
Name: count, dtype: int64
```

같은 의미의 데이터가
대소문자나 다른 용어로
다르게 표현됨

3. 모델링

모델 선택

autoML - pycaret 사용

pycaret
ML workflow을 자동화 하는 opensource library로 여러 머신러닝 task에서 사용하는 모델들을 하나의 환경에서 비교하고 튜닝하는 등 간단한 코드를 통해 편리하게 사용할 수 있도록 자동화한 라이브러리

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9029	0.0000	0.9203	0.8833	0.9012	0.8057	0.8068	2.0300
catboost	CatBoost Classifier	0.9018	0.0000	0.9220	0.8803	0.9005	0.8036	0.8048	2.8190
xgboost	Extreme Gradient Boosting	0.9015	0.0000	0.9141	0.8855	0.8993	0.8029	0.8037	0.1280
rf	Random Forest Classifier	0.8925	0.0000	0.8989	0.8806	0.8895	0.7849	0.7853	0.4100
gbc	Gradient Boosting Classifier	0.8849	0.0000	0.8994	0.8672	0.8826	0.7698	0.7710	0.4790
et	Extra Trees Classifier	0.8634	0.0000	0.8616	0.8560	0.8586	0.7265	0.7269	0.3250
dt	Decision Tree Classifier	0.8408	0.0000	0.8418	0.8307	0.8358	0.6814	0.6821	0.0410
ada	Ada Boost Classifier	0.8400	0.0000	0.8311	0.8364	0.8335	0.6795	0.6799	0.1810
ridge	Ridge Classifier	0.7497	0.0000	0.7017	0.7604	0.7295	0.4973	0.4991	0.0280
lda	Linear Discriminant Analysis	0.7491	0.0000	0.7017	0.7594	0.7291	0.4962	0.4979	0.0290
qda	Quadratic Discriminant Analysis	0.7382	0.0000	0.5932	0.8136	0.6854	0.4707	0.4884	0.0290
lr	Logistic Regression	0.7312	0.0000	0.6859	0.7381	0.7105	0.4602	0.4618	0.2750
nb	Naive Bayes	0.7205	0.0000	0.6791	0.7241	0.7005	0.4391	0.4402	0.0280
knn	K Neighbors Classifier	0.6697	0.0000	0.6497	0.6606	0.6544	0.3381	0.3388	0.0470
svm	SVM - Linear Kernel	0.5276	0.0000	0.5390	0.5498	0.4429	0.0570	0.0560	0.0510
dummy	Dummy Classifier	0.5184	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0250

상위 5개 모델을 후보로 놓고
Voting 하여 성능 측정

3. 모델링

과적합을 해결하기 위한 방법

1. 언더샘플링
   클래스 불균형이 심한 데이터를 그대로 학습하게 되면 다수 클래스에 편향된 모델이 됨
   데이터셋의 크기가 줄어들기 때문에 학습시간이 감소됨
   정보손실의 위험이 있음 --> 앙상블 + 보팅으로 문제 해결
   public score 0.2.. 에서 0.6 대로 상승

2. 앙상블
   여러개의 예측 모델을 결합하여 과적합을 줄이고 모델을 일반화하는 방법
   앞서 고른 상위 5개 모델을 앙상블하여 모델 일반화

3. 모델 학습 시 편향되어 학습되는 요인 찾기
   train 데이터에서 customer_idx = 25096 의 경우 영업 횟수 2421 모두 성공한 것으로 관측됨
   train 데이터의 True 개수가 4850개 임을 생각하면 위 idx에 편향되어 학습된다고 판단됨
   test 셋에 위 idx가 없는 것을 확인하였고 위 2421개 중 일부 추출하여 사용하여 과적합을 줄임(100개 추출)
   public score 0.7 대로 상승

4. Voting
   언더샘플링 시 정보손실의 문제가 있음
   False 데이터 54449 개를 랜덤셔플 후 20등분하고 True 와 합쳐 데이터셋 생성
   각각 데이터셋의 모델에서의 결과를 확률로 받은 후 0, 1 클래스의 확률을 평균을 내어 최종 결과로 생성(Soft voting)
   Public score 0.02 정도의 상승을 보임

3. 모델링

AB test

A	A Public Score	B Public Score	B
IDX == 25096 샘플링 X	0.7188	0.7253	IDX == 25096 100개 샘플링
product_category, customer_country 열 삭제	0.7206	0.7253	product_category, customer_country 라벨 인코딩
Customer_job 전처리	0.7059	0.7253	Customer_job 기존 전처리
Customer_job 삭제	0.7202	0.7253	Customer_job 기존 전처리
언더샘플링 비율 1.5:1	0.7226	0.7253	언더샘플링 비율 1:1
Value_counts = 1 그대로	0.7211	0.7253	Value_counts = 1 etc로 대체
결측치 etc	0.7253	0.7445	결측치 none으로
Expected_timeline other로 묶기	0.7344	0.7445	Expected_timeline 그대로

모델 학습

1. GridSearchCV
   최적의 하이퍼파라미터 튜닝을 위해 GridSearchCV를 이용
   시간이 오래걸리지만 최적의 파라미터를 찾을 가능성이 높음

4. 결과

모델 선택

앞서 선택한 5개 모델 중 5개, 3개, 1개로 나누어 앙상블하고 public score가 가장 높았던 모델 선택
모델 1개만 사용하였을 때 성능이 가장 좋았음 - xgb

최종 결과

Public score : 0.75611
Final score : 0.76485

844팀 중 63위 (30위 팀 Final Score : 0.78086)

아쉬웠던 점, 좋았던 점

실제 현업에서 사용하는 데이터는 전처리에 많은 시간을 쏟아야 한다는 것을 느낌
과적합을 해결하기 위해 많은 고민을 했고 그 과정에서 데이터에 대한 이해와 모델에 대한 이해를 키울 수 있었음

AutoML 의 pycaret 을 일찍 적용했다면 시간을 더 효율적으로 사용했을 것 같음
임의로 설정한 값들에 대해 정확한 튜닝을 하지 못하였음(시간 및 제출 횟수 부족)

감사합니다

2024. 02. 27