고객 대출등급 분류

강찬석
권우진
이경로
정서영

CONTENTS

01. EDA

탐색적 데이터 분석하기

데이터 가져오기

train.csv | test.csv | sample_submission.csv

Views | Grid view | Hide fields | Filter | Group | Sort

	ID	대출금액	대출기간	근로기간	주택소유상태	연간소득
1	TRAIN_00000	12480000	36 months	6 years	RENT	72000000
2	TRAIN_00001	14400000	60 months	10+ years	MORTGAGE	130800000

부채_대비_소득_비율	총계좌수	대출목적	최근_2년간_연체_횟...	총상환원금	총상환이자	총연체금액	연체계좌수	대출등급
18.90	15	부채 통합	0	0	0.0	0.0	0.0	C
22.33	21	주택 개선	0	373572	234060.0	0.0	0.0	B

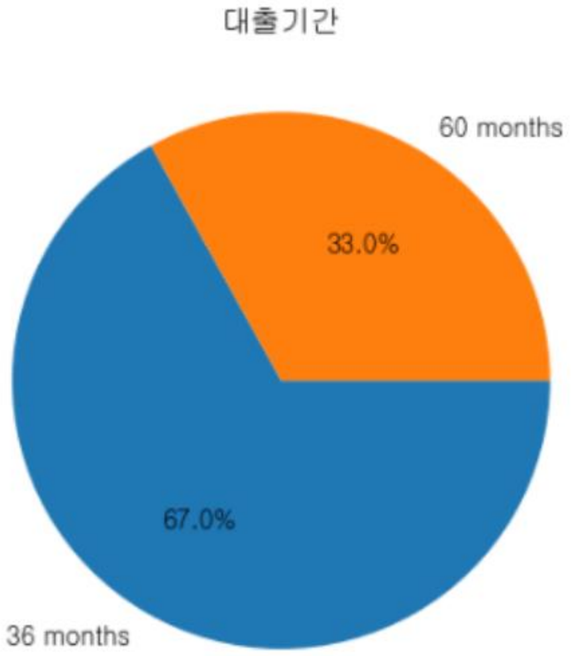
train.csv 파일

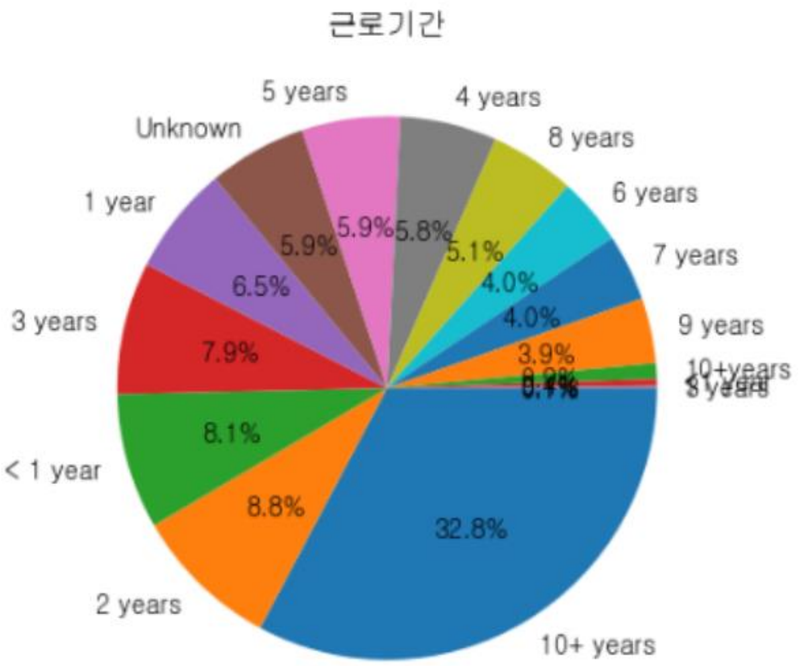
- 고객 관련 금융 정보
- ID : 대출 고객의 고유 ID
- 대출등급: 예측 목표

test.csv 파일

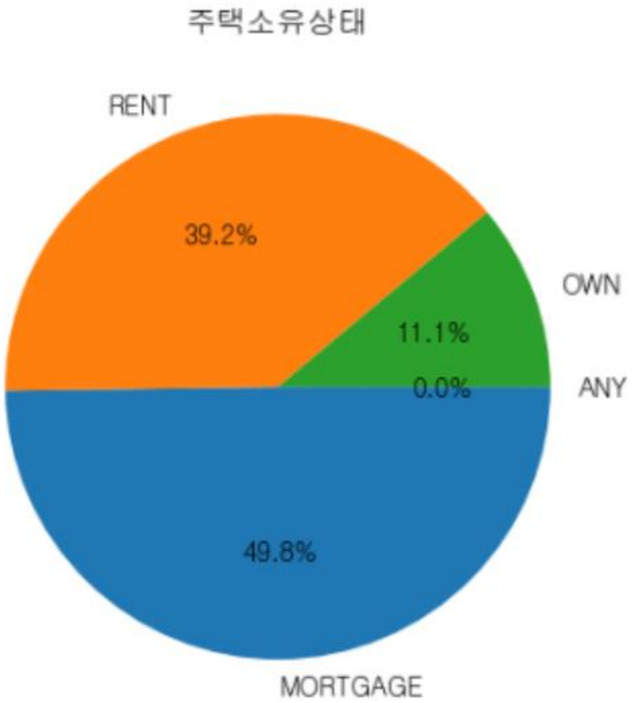
- 고객 관련 금융 정보
- ID : 대출 고객의 고유 ID
- 대출등급이 존재하지 않음

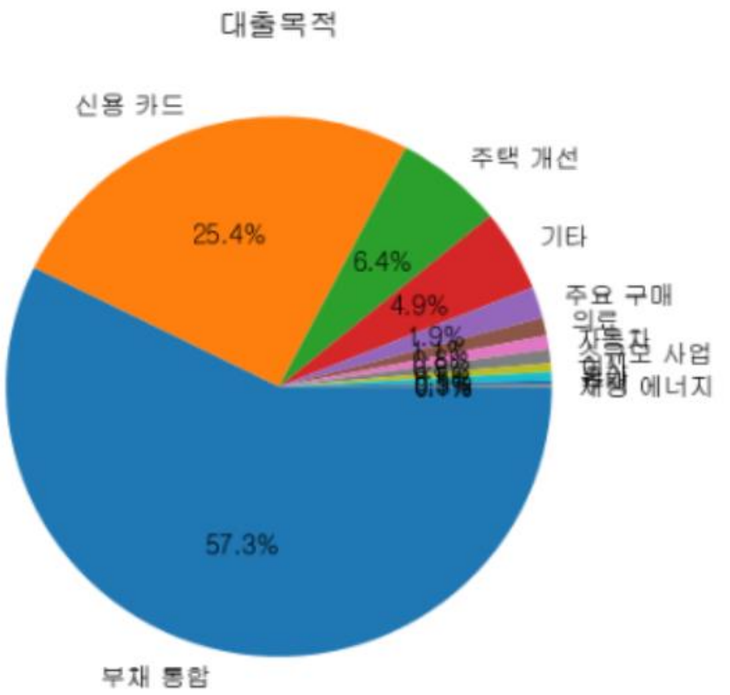
탐색적 데이터 분석


대출기간
60 months
33.0%
67.0%
36 months

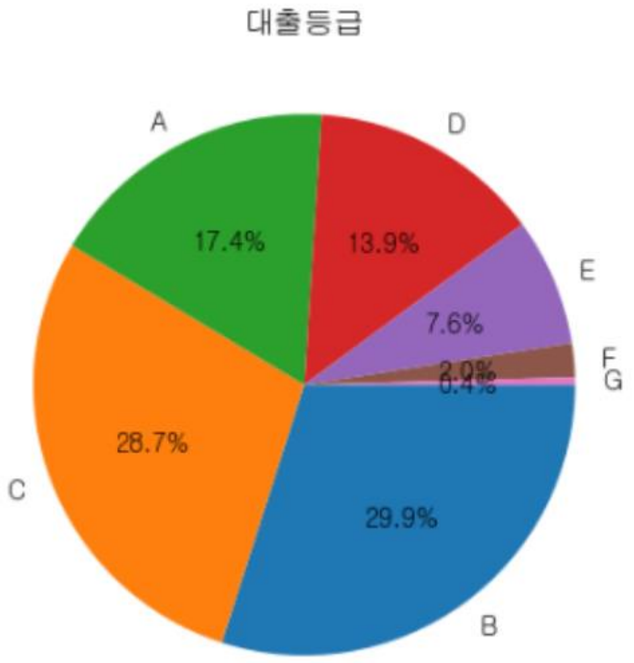

근로기간
5 years
4 years
Unknown
8 years
6 years
1 year
7 years
5.9%
5.9%
5.8%
5.1%
6.5%
4.0%
3 years
4.0%
9 years
7.9%
3.9%
10+years
8.1%
< 1 year
8.8%
32.8%
2 years
10+ years

탐색적 데이터 분석

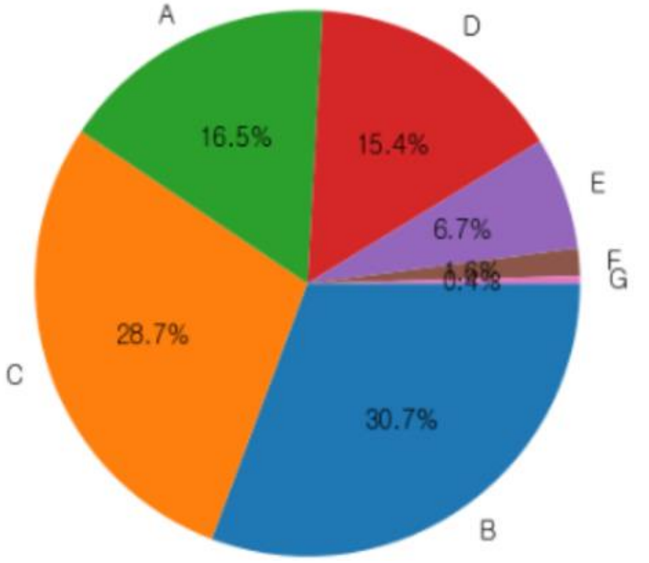

주택소유상태
RENT
39.2%
OWN
11.1%
ANY
0.0%
49.8%
MORTGAGE


대출목적
신용 카드
25.4%
주택 개선
6.4%
기타
4.9%
주요 구매
사업
에너지
57.3%
부채 통합

탐색적 데이터 분석

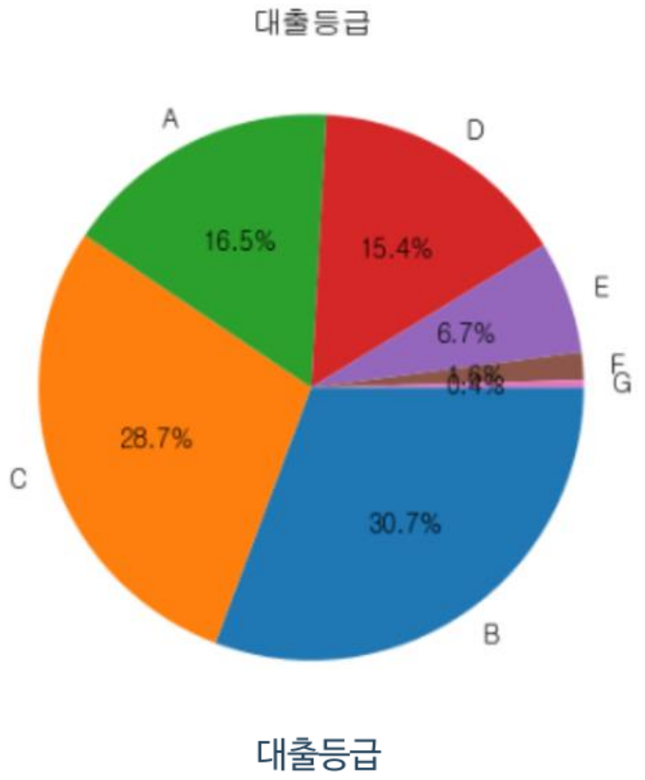


대출등급



근로기간이 Unknow인 사람들의 대출등급 분포

탐색적 데이터 분석



전체 데이터의 대출등급과
근로기간이 Unknown인 데이터의
대출등급 비율이 거의 일치

▶ 근로기간이 Unknwon인 데이터를 제외

탐색적 데이터 분석

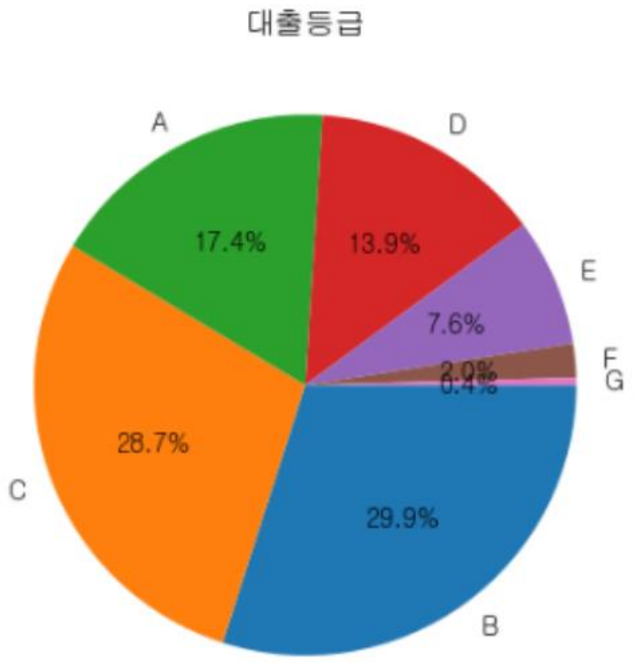
```
col = [   '대출기간',
          '주택소유상태',
          '총계좌수',
          '대출목적',
          '최근_2년간_연체_횟수',
          '연체계좌수'  ]


for name in col :
    train_df = train_df.join(pd.get_dummies(train_df[name],
                                            drop_first=True,
                                            prefix=name,
                                            dtype=float))
```

명목형 변수와 이산형 변수를 가변수 처리

```
train_df = train_df.drop(columns=col)
```

기존에 있는 컬럼 삭제

탐색적 데이터 분석



대출등급 데이터의 불균형

02. 모델선정 및 평가

```
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier
GradientBoostingClassifier()
```

```
pred_gbc = gbc.predict(X_test)
```

```
conf_mat = confusion_matrix(y_test, pred_gbc)
conf_mat
```

```
array([[3748,  906,   49,    0,    0,    0,    0],
       [ 352, 6501, 1198,    4,    0,    0,    1],
       [  51,  931, 6696,  147,    5,    3,    2],
       [   9,   98, 1595, 1882,  229,    5,    3],
       [   3,   20,  320,  520, 1226,   37,    4],
       [   0,    3,   33,   55,  131,  286,   22],
       [   0,    1,    5,   10,    4,   38,   54]], dtype=int64)
```

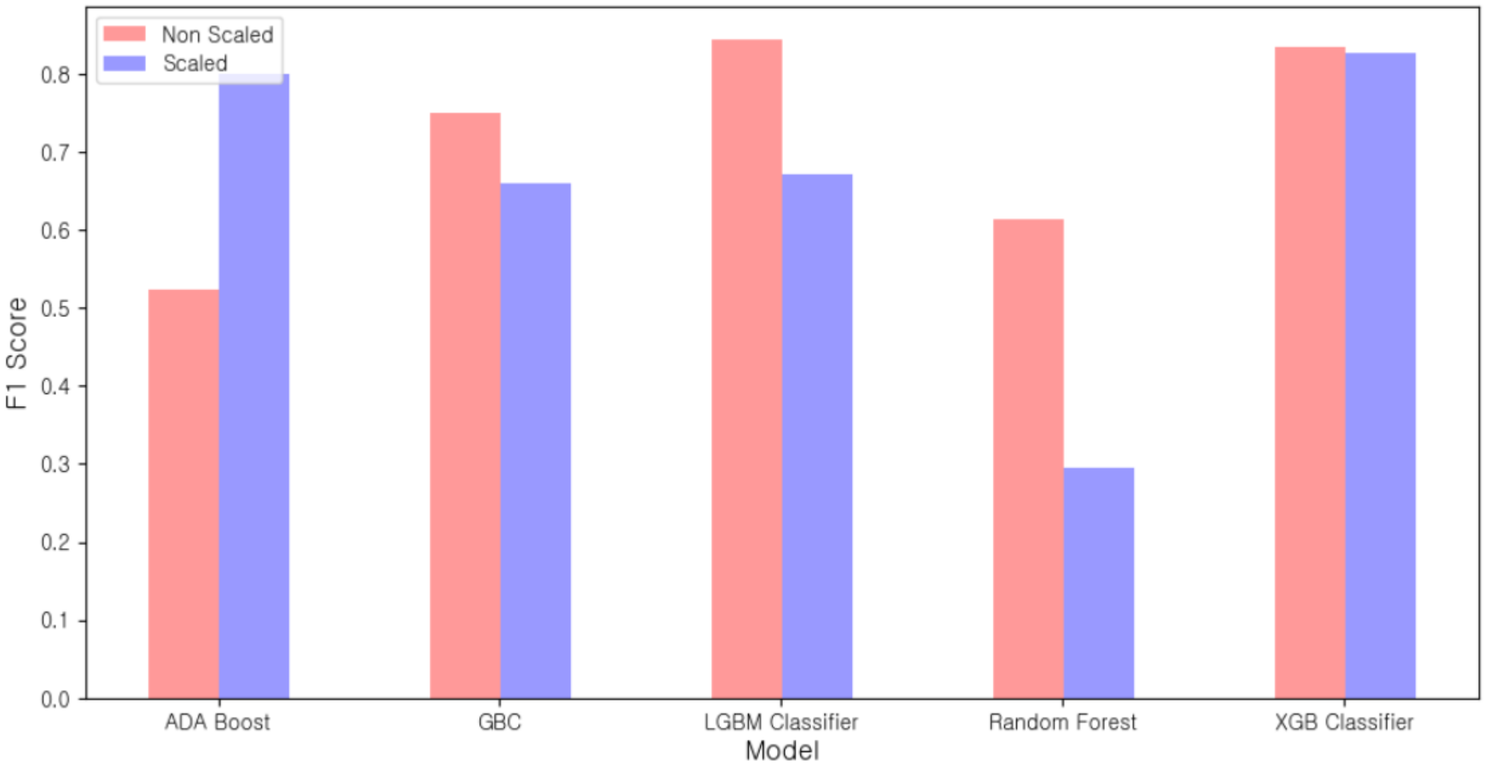
```
for i in range(conf_mat.shape[0]) :
    tp = conf_mat[i, i]
    fp = conf_mat[:, i].sum() - tp
    fn = conf_mat[i, :].sum() - tp
    tn = conf_mat.sum() - (fp + fn + tp)

    print(f'Class{i} : TP - {tp}, FP - {fp}, FN - {fn}, TN - {tn}')
print(classification_report(y_test, pred_gbc))
```

```
Class0 : TP - 3748, FP - 415, FN - 955, TN - 22069
Class1 : TP - 6501, FP - 1959, FN - 1555, TN - 17172
Class2 : TP - 6696, FP - 3200, FN - 1139, TN - 16152
Class3 : TP - 1882, FP - 736, FN - 1939, TN - 22630
Class4 : TP - 1226, FP - 369, FN - 904, TN - 24688
Class5 : TP - 286, FP - 83, FN - 244, TN - 26574
Class6 : TP - 54, FP - 32, FN - 58, TN - 27043
              precision    recall  f1-score   support

           A       0.90      0.80      0.85      4703
           B       0.77      0.81      0.79      8056
           C       0.68      0.85      0.76      7835
           D       0.72      0.49      0.58      3821
           E       0.77      0.58      0.66      2130
           F       0.78      0.54      0.64       530
           G       0.63      0.48      0.55       112

    accuracy                           0.75     27187
   macro avg       0.75      0.65      0.69     27187
weighted avg       0.76      0.75      0.75     27187
```

```
metrics.f1_score(y_test, pred_gbc, average='micro')
```

```
0.7501011512855409
```

	model	F1 Score	F1 Score_std
0	ADA Boost	0.524	0.800
1	GBC	0.750	0.660
2	LGBM Classifier	0.845	0.672
3	Random Forest	0.613	0.295
4	XGB Classifier	0.834	0.828



03. REVIEW

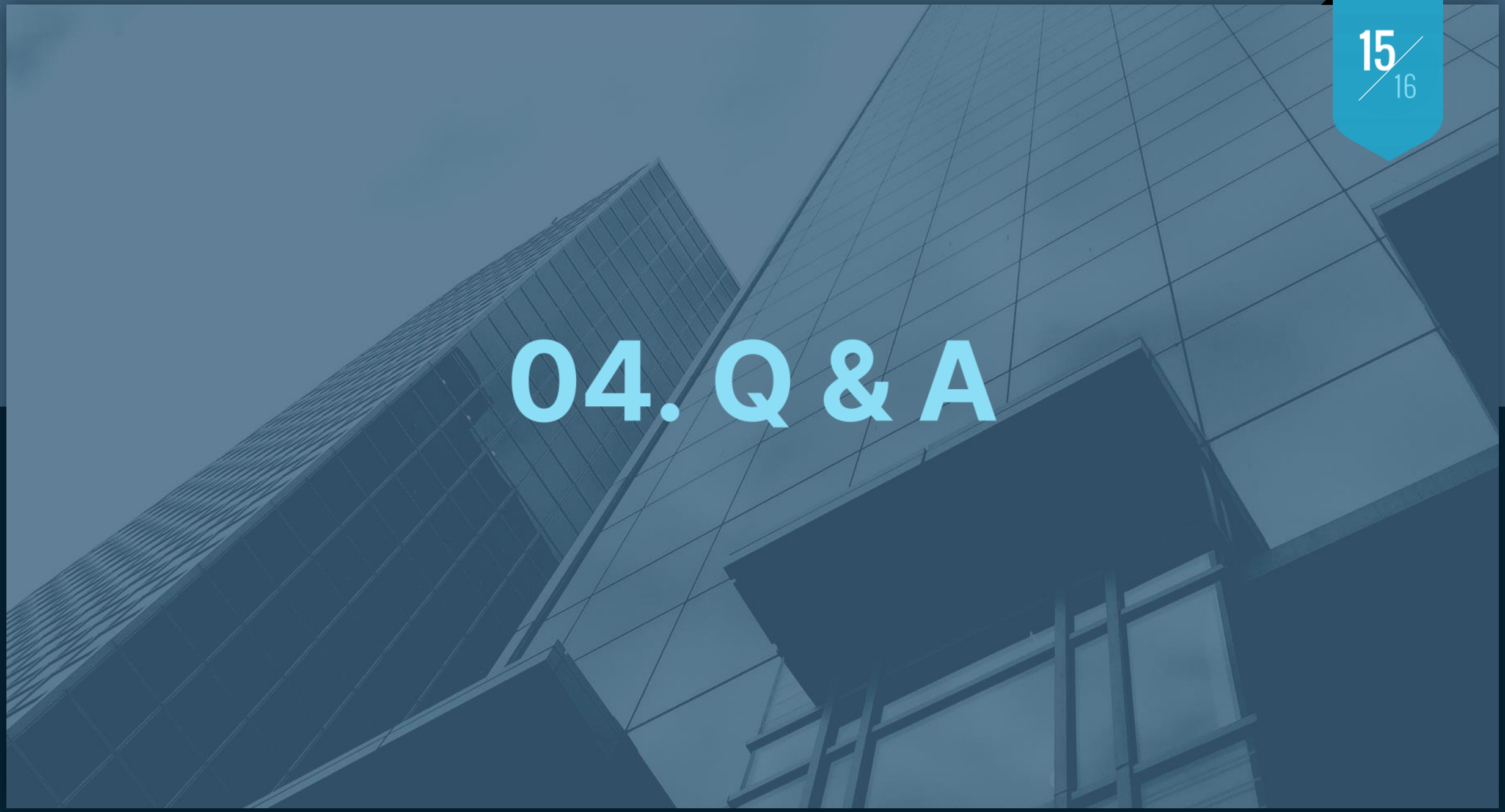
REVIEW

01 폐암과 흡연 데이터의 신뢰도 부족으로 주제 변경

02 신용등급 데이터 전처리 미흡

03 신용등급 산정방법에 대한 배경지식이 부재

04 GridSearchCV, RandomizedSearchCV를 통해 최적의 하이퍼 파라미터를 찾기에는 시간이 부족



04. Q & A

감사합니다

데이터 출처: Dacon 고객 대출등급 분류 해커톤