목차

1. 서론

    A. 분석 배경

        - 카드 연체자들의 패턴을 데이터를 통해 분석하고 앞으로 나올 연체자들을 사전에 방지하기위해~

    B. 분석 목표

        i. ANOVA를 통한 실험계획법 분석 방법 쓸 거 미리 소개

        - ANOVA를 통해 문제점 언급할 거라 미리 말하기

        - 변수 PAY1만 썻을때랑 PAY1 ,2 두개 썻을 때를 ANOVA로 비교

2. 데이터 설명

- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

    A. 응답변수 설명

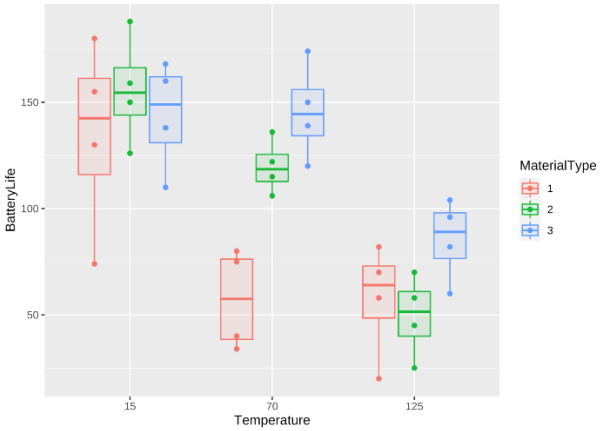
        i. 응답변수 분포

            1. 데이터 불균형 설명

    B. 반응변수 설명

        i. 수치형 20개

            1. 기초 통계량



(이런 느낌으로 시각화 할거임)

2. 분포
    A. 히스토그램
    B. 파이 차트
3. 상관계수
    A. 다중 공선성
        i. 파생변수 생성
            1. 한계점 제시: 데이터 셋에 대해 구체적으로 아는 바가 없어 파생변수 생성과 관련해선 보수적으로 접근할 필요 있다
        ii. 변수 제거
            1. PAY_1, 2 데이터 중 1 하나만 VS 1, 2 둘 다 쓰는 거 비교
                A. 피쳐 임포턴스가 근거

	feat	score
5	PAY_1	0.722666
6	PAY_2	0.145870
19	PAY_AMT3	0.049608
11	BILL_AMT1	0.024832
7	PAY_3	0.021724
10	PAY_6	0.018298
2	EDUCATION	0.006637
0	LIMIT_BAL	0.003905
8	PAY_4	0.003220
18	PAY_AMT2	0.002195
12	BILL_AMT2	0.001045
9	PAY_5	0.000000
4	AGE	0.000000
1	SEX	0.000000
13	BILL_AMT3	0.000000
14	BILL_AMT4	0.000000

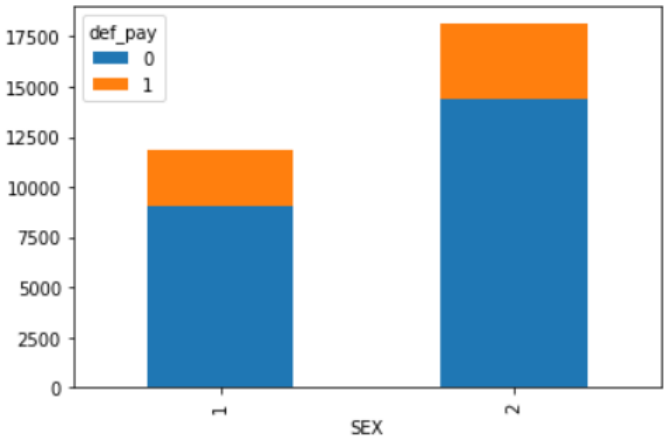
(옆 그림이 FEATURE IMPORTANTS)

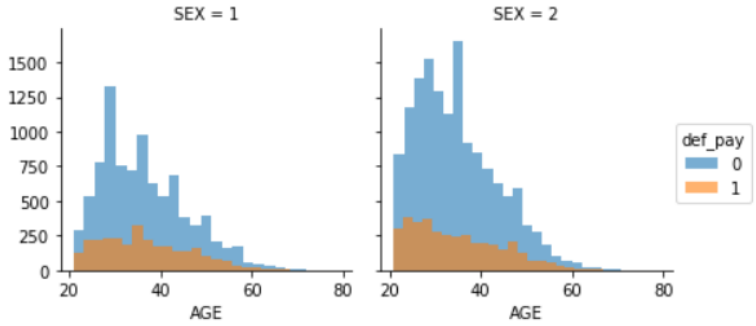
ii. 범주형 3개

1. 성별, 결혼여부, 교육수준

A. 디폴트 비율 차이 시각화

i. 파이 차트





2. 나이

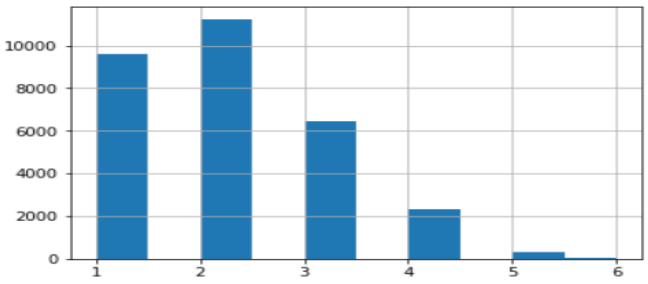
A. 수치형으로 따지기엔 비선형적일 것 같다

i. 때문에 범주화해서 디폴트 비율 차이

1. 20대, 30대, 40대, 50대, 60세 이상

```
df['AgeBin'] = 0 #creates a column of 0
df.loc[((df['AGE'] > 20) & (df['AGE'] < 30)) , 'AgeBin'] = 1
df.loc[((df['AGE'] >= 30) & (df['AGE'] < 40)) , 'AgeBin'] = 2
df.loc[((df['AGE'] >= 40) & (df['AGE'] < 50)) , 'AgeBin'] = 3
df.loc[((df['AGE'] >= 50) & (df['AGE'] < 60)) , 'AgeBin'] = 4
df.loc[((df['AGE'] >= 60) & (df['AGE'] < 70)) , 'AgeBin'] = 5
df.loc[((df['AGE'] >= 70) & (df['AGE'] < 81)) , 'AgeBin'] = 6
df.AgeBin.hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa21adb67b8>
```



iii. 문제 해결 방안 소개

1. 다중 공선성 관련해서 변수 하나 남기고 없앨지, 두 개 남기고 없앨지

2. 오버 샘플링 했을 때와 안 했을 때 비교

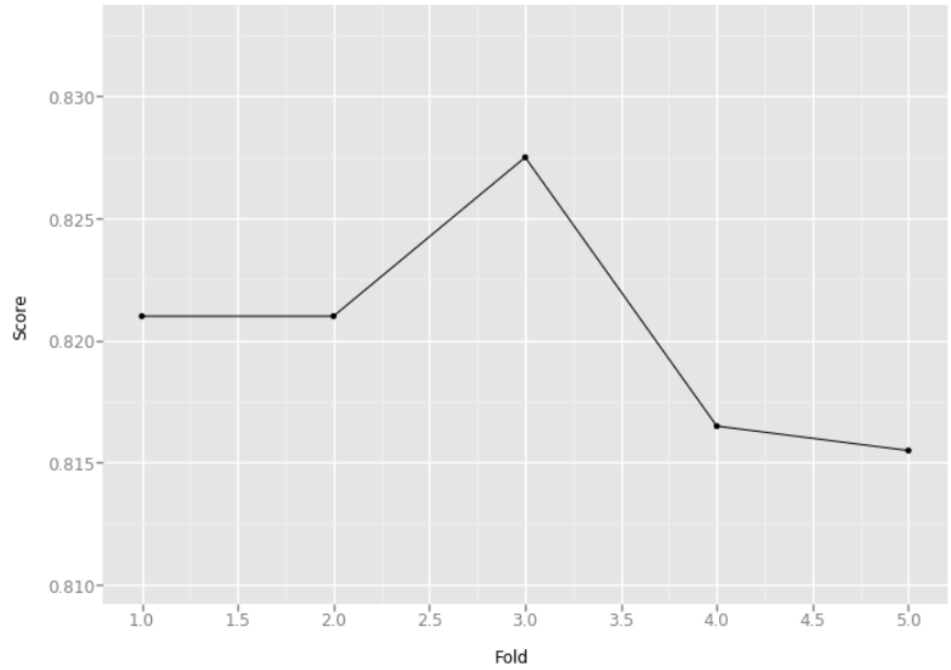
iv. 분류 모델 소개

1. 파라미터 튜닝 관련 소개

A. GridSearchCV 패키지 써서 최적 튜닝 찾기

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
            max_features=None, max_leaf_nodes=20,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=20,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
0.821291666667
```

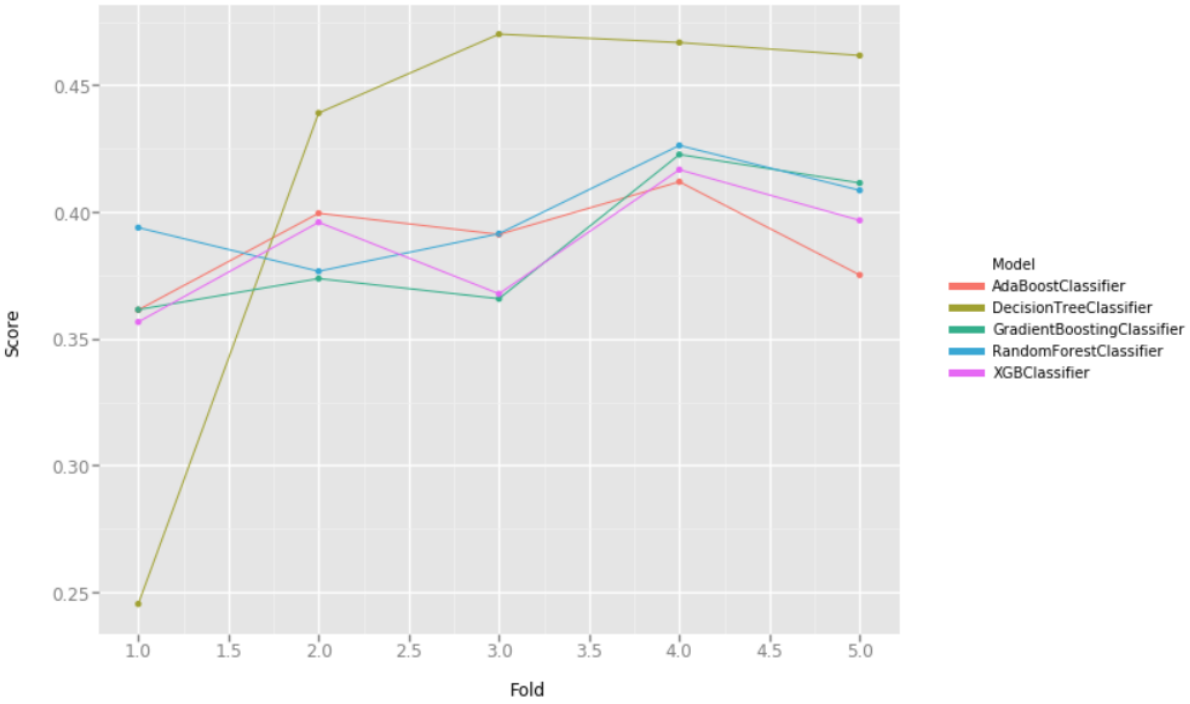
B. KFold 기법 사용



2. LinearSVC
   A. 소개, 장단점과 데이터 특성 연결
   B. 결과
3. Logistic Regression
   A. 소개, 장단점과 데이터 특성 연결
   B. 결과
4. Random Forest
   A. 소개, 장단점과 데이터 특성 연결
   B. 결과
5. Neural Network
   A. 소개, 장단점과 데이터 특성 연결

B. 결과

6. 모델을 소개할 때 시각화한 자료와 엮어서 설명할 것



v. ANOVA

1. 교호가 발생, 왜 나타나는지 쓰기