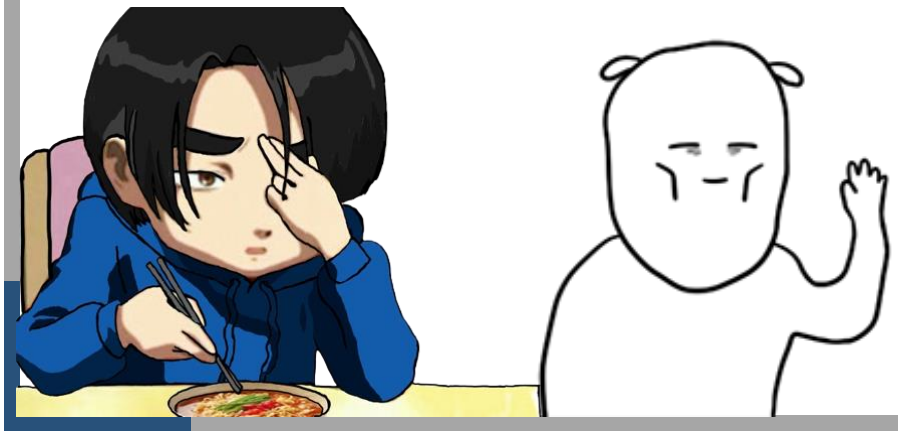
방학세미나 후기



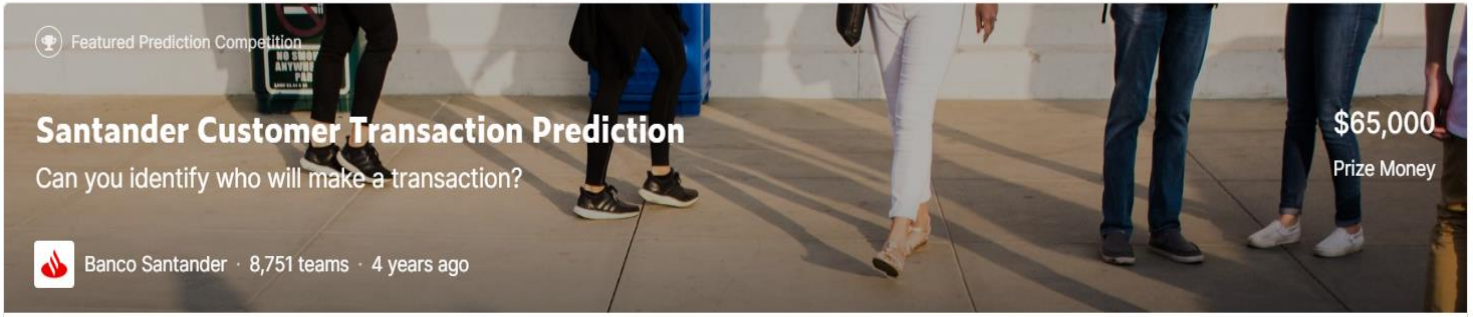
학회장팀

김현우 윤지영

INDEX

1 출제 의도

DATA



https://www.kaggle.com/competitions/santander-customer-transaction-prediction/data

Kaggle에서 진행했던 은행의 거래 예측 Competition

고객이 미래에 특정 거래를 진행 할 것인지 여부 예측

0 : 거래 X / 1 : 거래 O

목적

불균형 클래스에 대한 분류모델 생성 연습

& 변수가 많은 고차원 데이터에 대한 접근법 연습

평가지표

F1 Score(Macro)

: 각 Label당 F1 Score의 평균

$$Macro - Precision = \frac{Pecision1 + Precision2}{2}$$

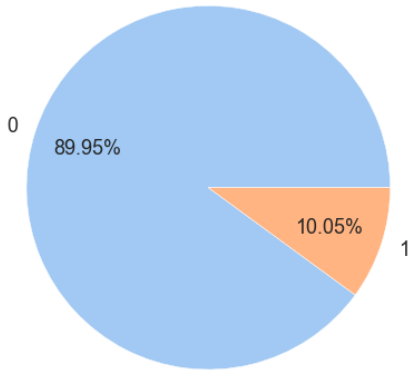
$$Macro - Recall = \frac{Recall1 + Recall2}{2}$$

$$Macro - F - Score = 2.\frac{Macro - Precision\,.\,Macro - Recall}{Macro - Precision + Macro - Recall}$$

분류 모델의 성능을 평가하는 지표로 사용

1 출제 의도

평가지표



클래스의 분포가 약 9:1인

Imbalanced Data의 분류 문제

불균형한 데이터에 대한 성능 평가에는

일반적으로 Accuracy보다

F1 Score나 Cross-Entropy를 사용

분류 모델의 성능을 평가하는 지표로 사용

1 출제 의도

평가지표

	실제 거래 O	실제 거래 X
예측 거래 O	TP	FP
예측 거래 X	FN	TN

Accuracy : (TP+TN)/(TP+FP+FN+TN)

은행의 입장에선

모델이 정확하게 작동하는 것 또한 중요하지만,

사용자가 특정 거래를 진행하는 지를

예측하는 것 또한 중요

↓

따라서 각 label(0, 1)에 대한

F1-Score의 평균인 Macro-F1을 사용!

분류 모델의 성능을 평가하는 지표로 사용

1 출제 의도

분석 과제

고차원 데이터 처리

데이터 셋 내 200개 변수에 대한 처리
변수 선택이나 변수 추출,
혹은 파생변수 생성

불균형 클래스

0과 1이 9:1의 비율로 분포
Minor Class에 대한 예측

변수 특성에 따른 모델 선택

변수의 특성에 따른
적절한 모델 선정

하이퍼파라미터 튜닝 및 시각화

파라미터 튜닝 시 평가지표
시각화의 효율성

1 출제 의도

분석 과제

Main!

고차원 데이터 처리

데이터 셋 내 200개 변수에 대한 처리

변수 선택이나 변수 추출,

혹은 파생변수 생성

불균형 클래스

0과 1이 9:1의 비율로 분포

Minor Class에 대한 예측

변수 특성에 따른 모델 선택

변수의 특성에 따른

적절한 모델 선정

하이퍼파라미터 튜닝 및 시각화

파라미터 튜닝 시 평가지표

시각화의 효율성

2

방법론 참고

2 방법론 참고

200개의 변수

변수선택, 차원축소 등 피쳐를 줄이는 시도를 하기 쉬움!

하지만,

이번 데이터에서는 이러한 방법이 잘 작동하지 X

변수의 특징을 찾아내어

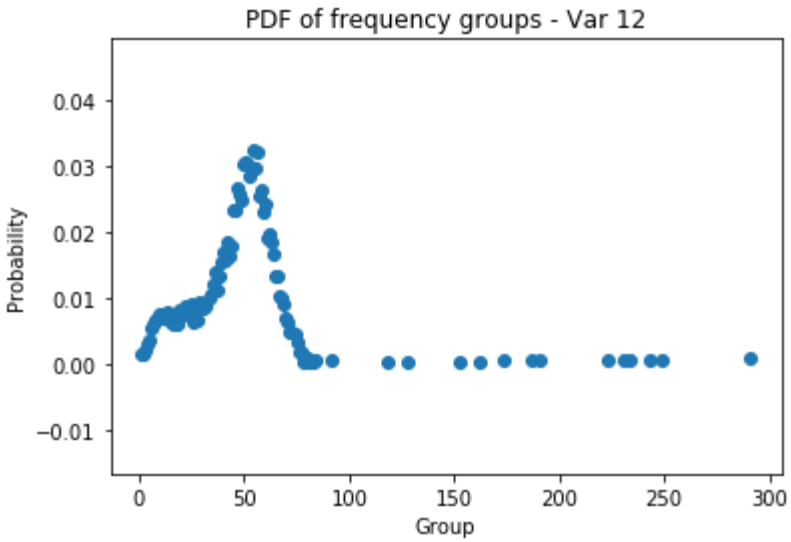
이를 충분히 이용하는 방향으로 접근

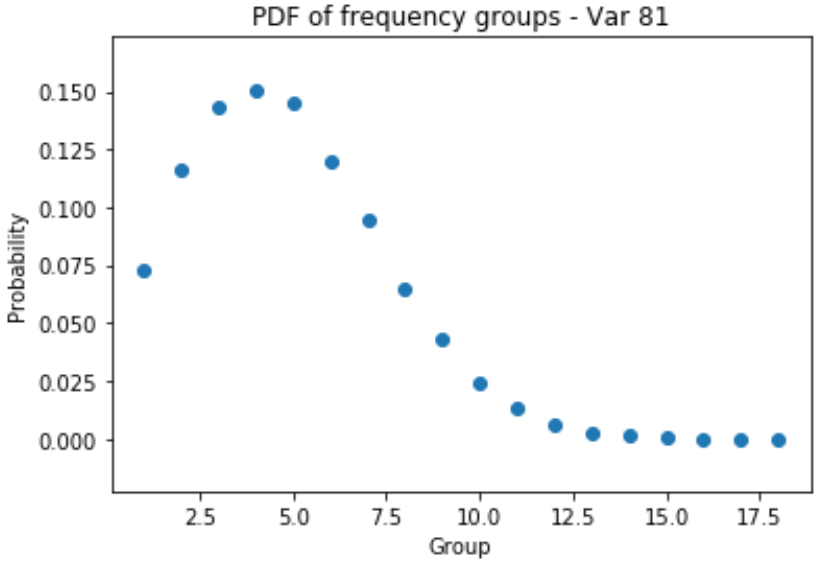
변수선택 공부에 도움되는

산탄데르 고객 만족도 예측 데이터

https://www.kaggle.com/competitions/santander-customer-satisfaction/overview

변수별 빈도수 확인



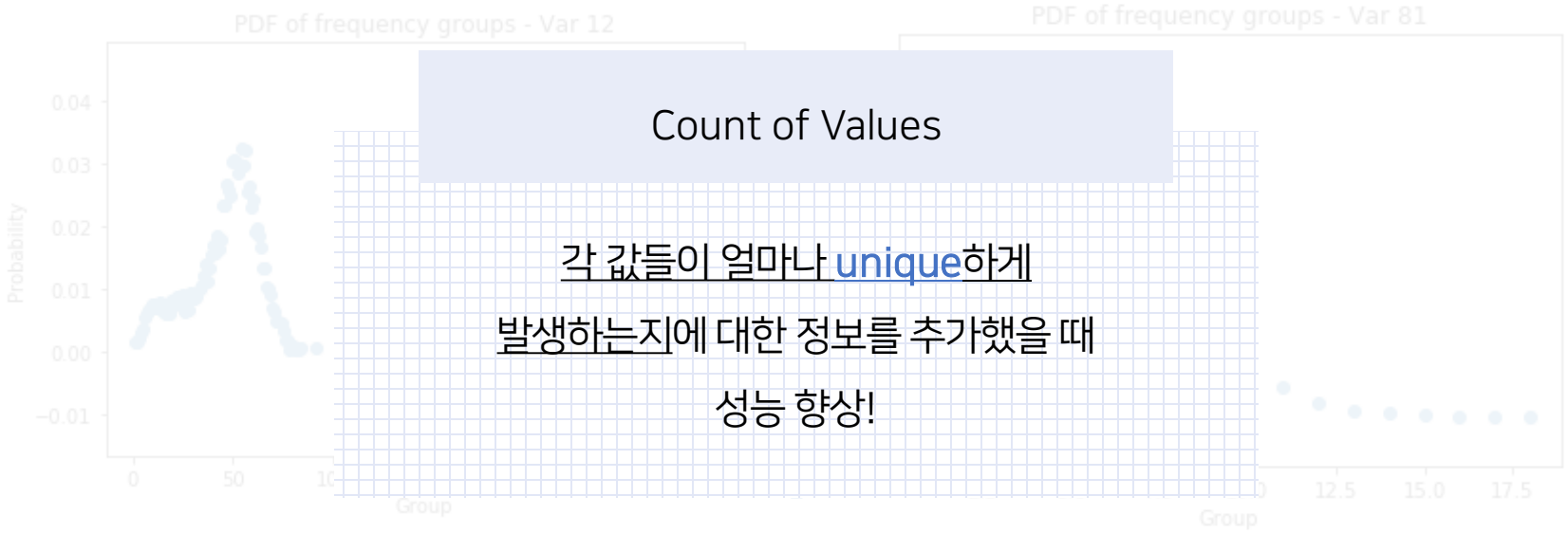


각 값이 발생하는 빈도수대로 그룹을 만들어 확인

→ 특정 변수에서는 매우 많은 그룹 존재

2 방법론 참고

변수별 빈도수 확인



Count of Values

각 값들이 얼마나 unique하게
발생하는지에 대한 정보를 추가했을 때
성능 향상!

이번 데이터처럼 연속적인 값들의 빈도수가
유의미하게 작동되는 것은 이례적인 일이긴 하나,
범주형 변수의 경우 이런 빈도수를 확인하는 것은
매우 기본적인 절차라고 합니다~

2 방법론 참고

빈도 변수 생성

EXAMPLE

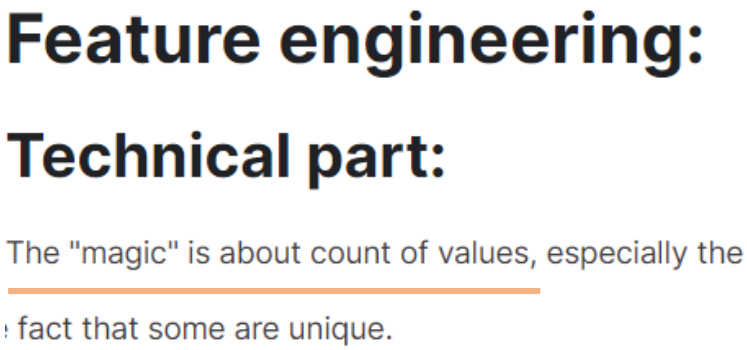
빈도 변수를 이용한 여러 접근법

① 빈도수가 발생하는 5가지의 상황으로
범주형 변수 생성

② 빈도수 변수 추가 후 인코딩

③ 빈도수가 1이면 0, 나머지는 1로 처리

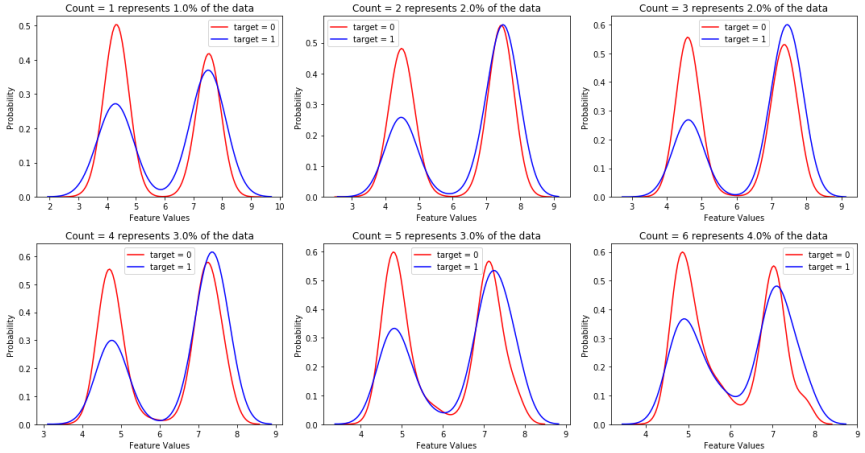
. . .



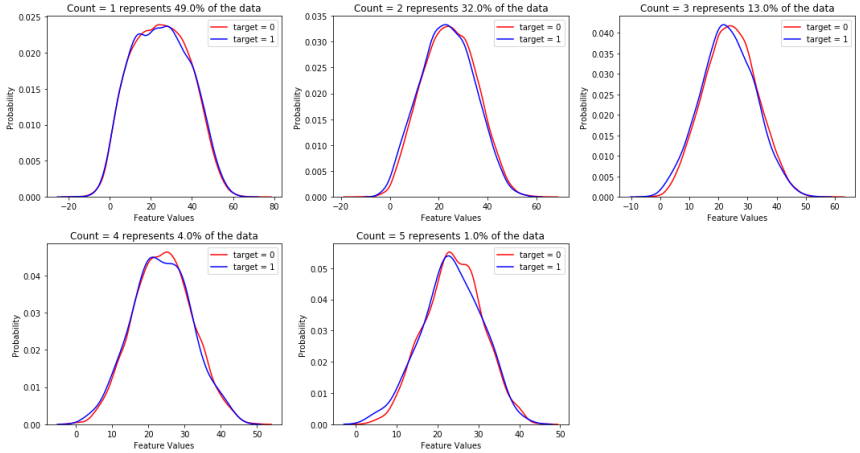
실제 대회에서 상위팀 모두 빈도 변수 이용

2 방법론 참고

빈도 변수 생성



Var111의 빈도수별 분포

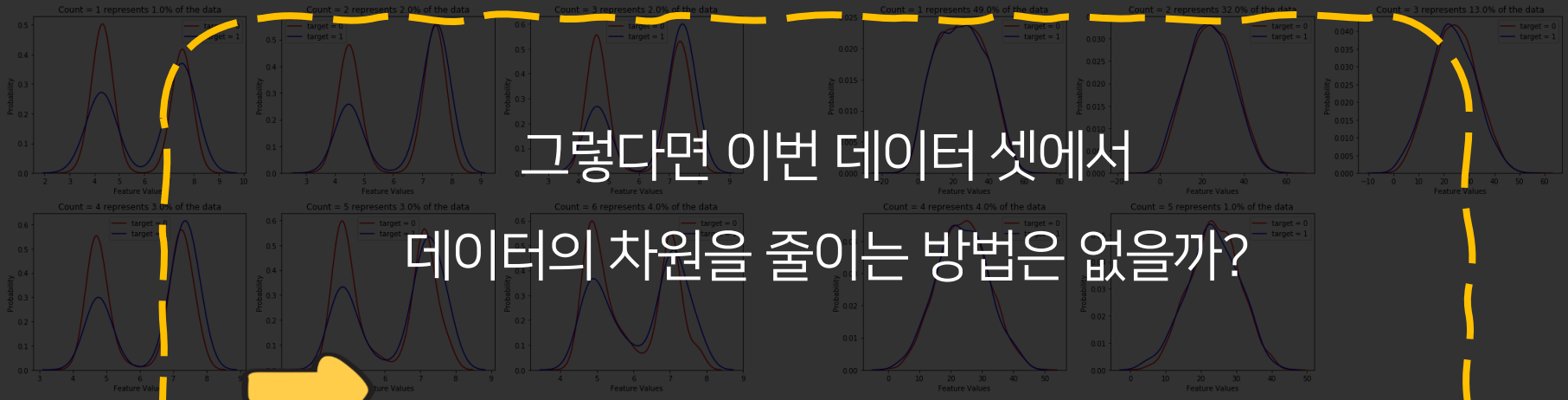


Var120의 빈도수별 분포

즉, EDA를 통해 변수별로 차이점을 나타내는 **고유한 특징**을 찾아
그 요소를 이용할 수 있는 변수를 만들어보는 것이 중요

2 방법론 참고

빈도 변수 생성



그렇다면 이번 데이터 셋에서

데이터의 차원을 줄이는 방법은 없을까?

Var111의 빈도수별 분포

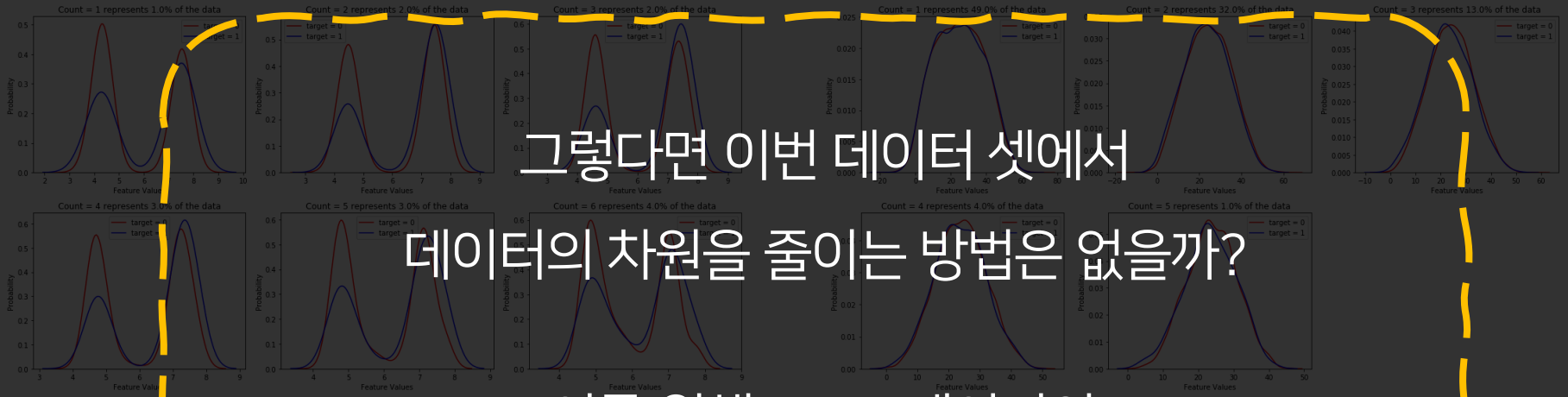
Var120의 빈도수별 분포

즉, EDA를 통해 변수별로 차이점을 나타내는 **고유한 특징**을 찾아

그 요소를 이용할 수 있는 변수를 만들어보는 것이 중요

2 방법론 참고

빈도 변수 생성



그렇다면 이번 데이터 셋에서

데이터의 차원을 줄이는 방법은 없을까?

Var111의 빈도수별 분포

이를 위해 Train 데이터의

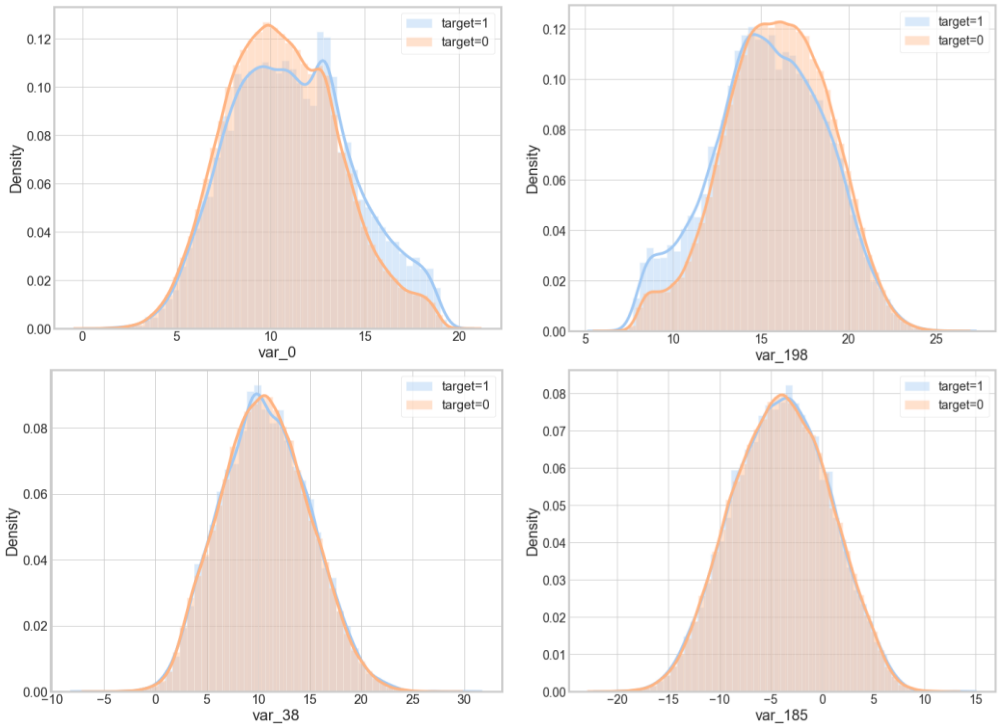
Var120의 빈도수별 분포

Target 값에 따른 분포를 확인!

즉, EDA를 통해 변수별로 차이점을 나타내는 **고유한 특징**을 찾아

그 요소를 이용할 수 있는 변수를 만들어보는 것이 중요

Target 별 변수 분포 시각화

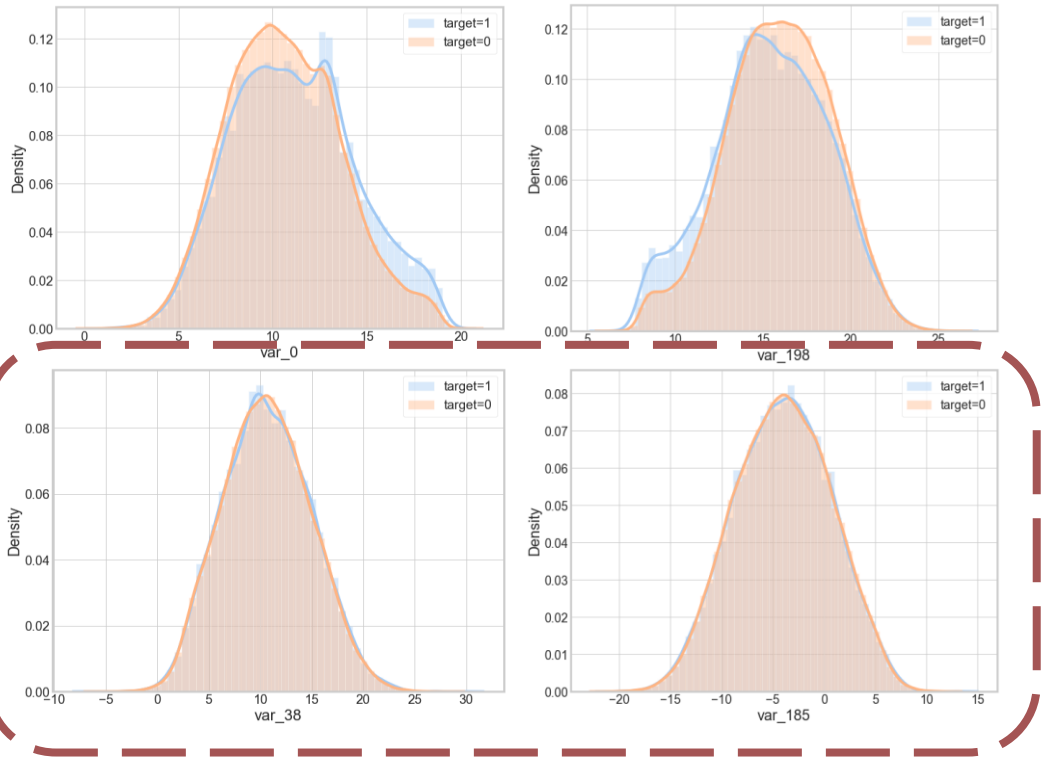


Target에 따른 변수의 분포를 확인해보았을 때,

대체로 Target값에 따라 분포가 달라 보이지만,

Target값에 관계없이 비슷해 보이는 분포가 나타남을 확인 가능

Target 별 변수 분포 시각화



제거 !

따라서, **동질성 검정**을 진행하여

Target값에 관계없이 비슷해 보이는 분포를 **제거**

2 방법론 참고

동질성 검정

주어진 **두 표본의 분포가 일치**하는가 검정

범주형

카이 제곱 검정

KS test는 연속형 데이터에만 사용

H_0 : 비교하는 두 분포가 동질적이다

P-value <0.05 → 비교하는 분포 이질성 존재

연속형

Kolmogorov Smirnov 검정

H_0 : 비교하는 두 분포가 동질적이다

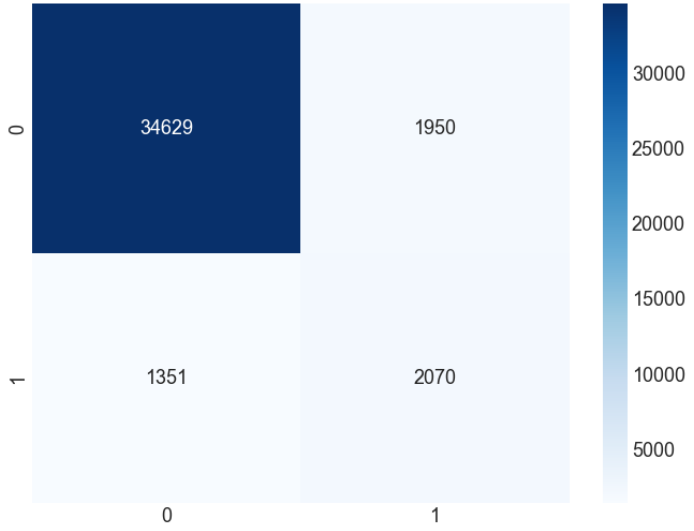
P-value <0.05 → 비교하는 분포 이질성 존재

만일 열 별 Target의 0/1에 대한 두 분포가 이질적임을 확인

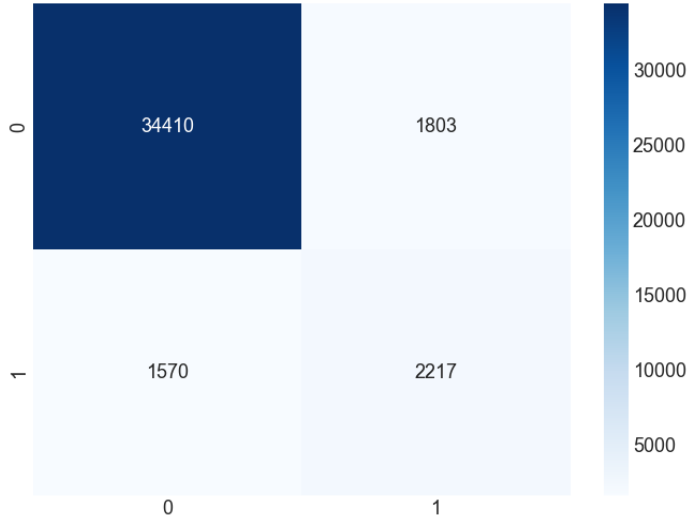
→ 0/1을 결정짓는 중요한 요인으로 작용함을 알 수 있음

2 방법론 참고

동질성 검정



변수선택



변수 선택 전 F1-Score : 0.7556

변수 선택 후 F1-Score : 0.7606

KS검정에서 동질성을 보이는 분포(P-value >=0.05)를 가지는
변수를 제거해주었을 때, 유의미한 성능향상이 나타남을 확인

추가적 성능 향상 방법: 앙상블

단일 모델들을 세운 후, **앙상블** 사용

예측이 틀리는 경우는 흔하게 발생하지 않기 때문에,
여러 모델의 예측값들을 함께 이용하여 오분류된 데이터를 보완가능

↓

각 모델의 단일 예측률이 높을수록

모델 간 상관관계가 낮을수록(다양성)

앙상블 통한 error 보완력 ↑

2 방법론 참고

앙상블 대표적 예시

Weighted Voting Ensemble

성능이 좋은 모델에 가중치를 주며 **다수결**로 예측값을 결정하는 형식

6개의 모델이 있는 경우

EXAMPLE

'가장 잘 예측하는 모델 ' 에게 3표를 주고,
나머지 5개의 모델에게 한 표씩 주었을 때
가장 좋은 모델의 몇 개의 오분류를
바로잡을 가능성 부여

0에 대한 예측력은 높지만 1에 대해서 50%로
찍고 있는 경우 더욱 잘 작동

2 방법론 참고

앙상블 대표적 예시

예측값이 연속적인 경우

Weighted Average Ensemble

여러 모델의 예측값들의 **가중평균**을 사용

EXAMPLE

이번 데이터의 기존 대회는 분류확률을 제출

LGBM & NN 모델의 예측력이 높다는 사실 발견

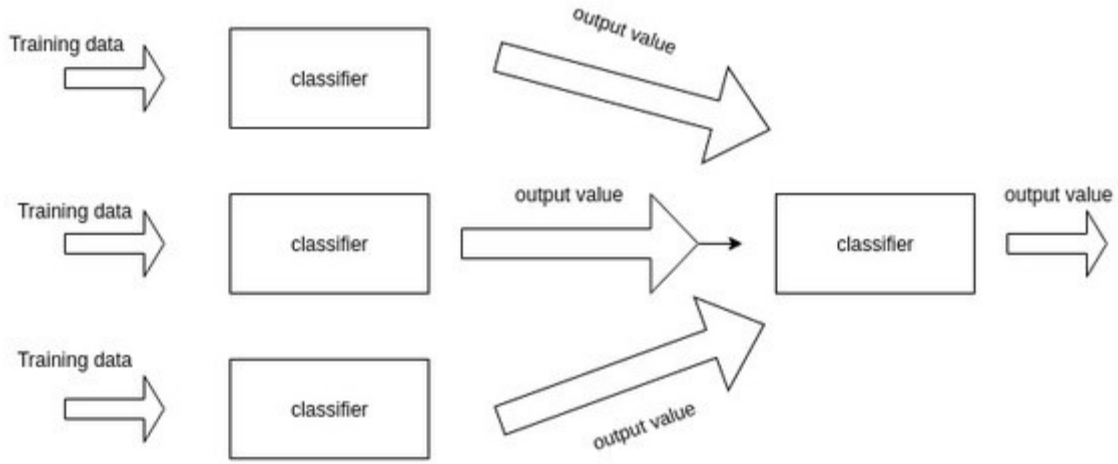
1, 2위팀 모두lgbm, nn모델의 예측값을

ex) 6:4 의 비율로 평균내어 사용

앙상블 대표적 예시

Stacking Ensemble

개별 모델의 예측 결과들로 메타 데이터셋을 만들고
최종 모델에 적용하여 예측을 수행하는 방법

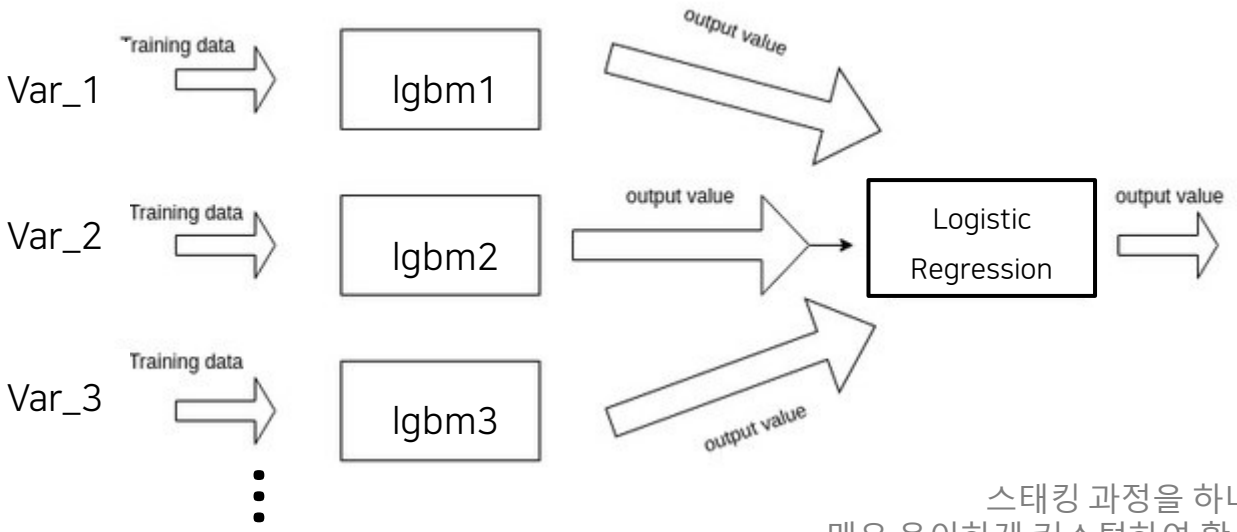


2 방법론 참고

앙상블 대표적 예시

EXAMPLE : Gold Medal solution

각 변수별로 학습을 진행하여 200개의 모델로 예측을 한 후
로지스틱 회귀모델로 최종 예측 진행



스태킹 과정을 하나의 클래스로 짜둔 커널로,
매우 용이하게 커스텀하여 활용할 수 있으니 참고하세용

https://www.kaggle.com/code/yekenot/simple-stacker-lb-0-284

앙상블 대표적 예시

스태킹 기법 - 다중계층스태킹, 메타 특징 변수 생성 등

홀드아웃 데이터를 이용한 블렌딩 기법 등

이외에도 다양한 앙상블 기법 존재

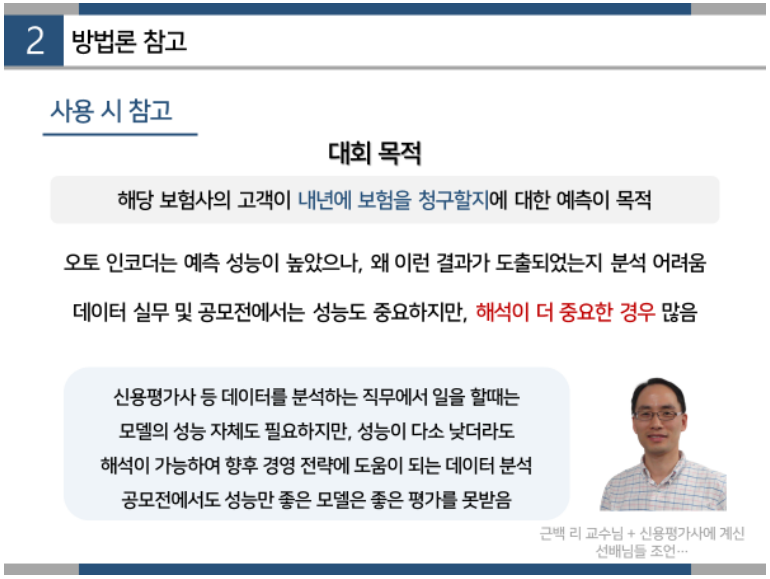
데이터/경진대회의 성격, 평가지표 등에 따라 효과가 달라짐

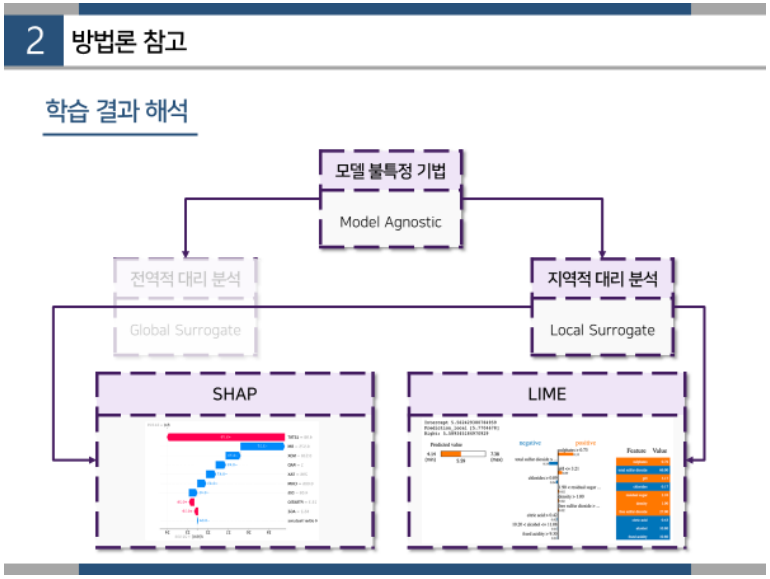
→ 상황에 따라 적용하는 것이 필요

가중평균을 보완하여 사용한 사례
https://www.kaggle.com/c/home-credit-default-risk/discussion/64722

2 방법론 참고

기타 참고 사항





자세한 내용은 지난 방세 PPT 참고!

이 외에도 SHAP이나 LIME들을 활용해서

각 변수들이 어떻게 모델에서 작용하는지,

그 해석에도 힘 쓸 수 있음

3

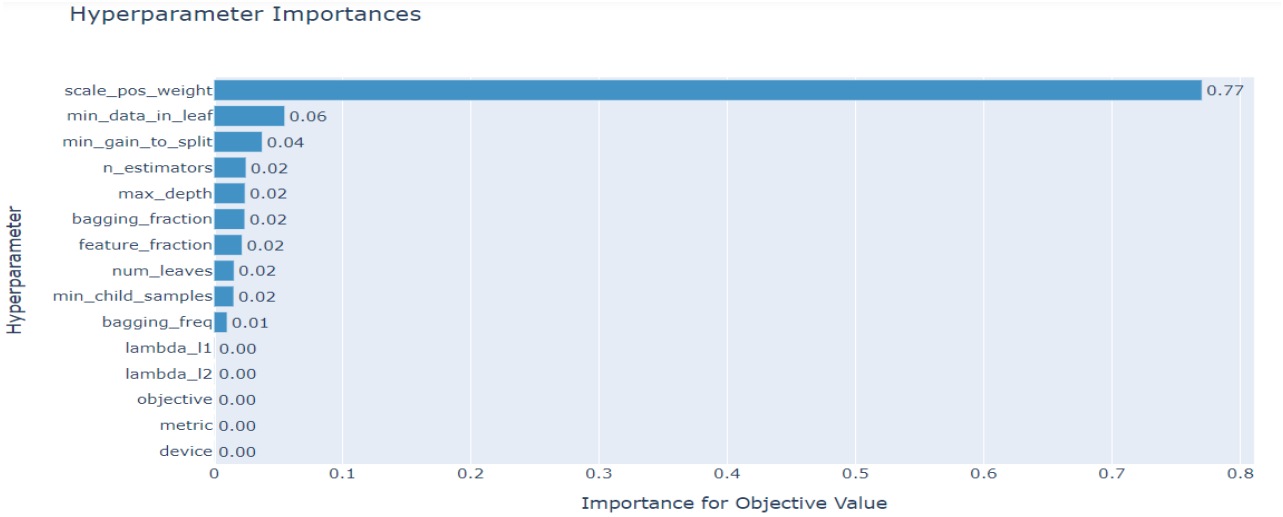
공통 피드백

공통 피드백

EDA 과정에서 대부분의 팀들이 본인들이 실패한 내용을 제외하고 제출

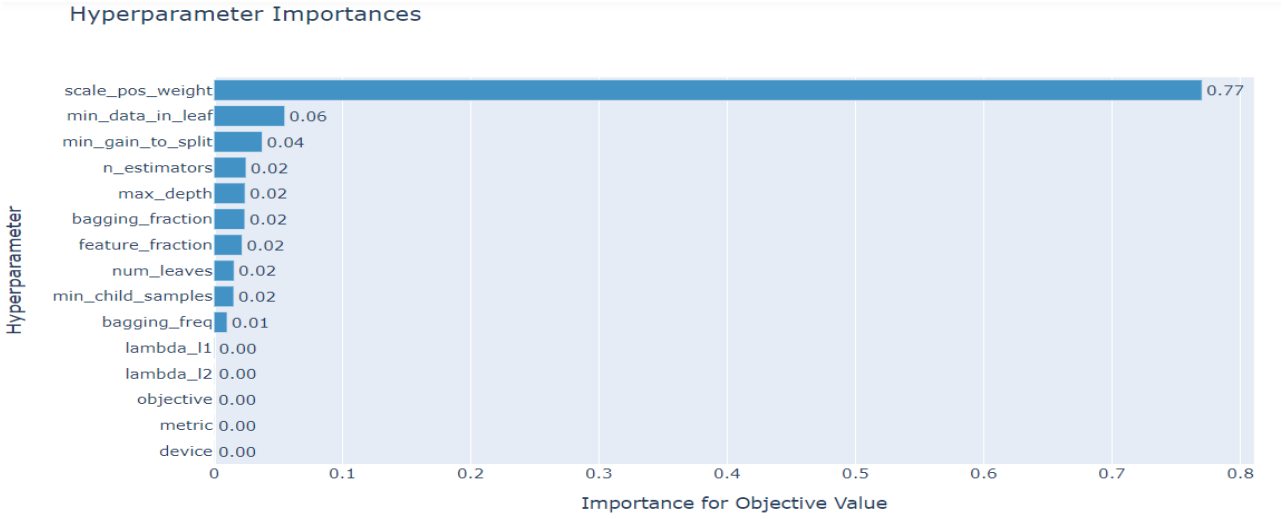
EDA 과정에서 인사이트를 찾아내지 못하거나, 적용 결과 큰 의미가 없더라도
이 내용을 잘 정리한다면, 해당 기법을 사용하지 않은 이유에 대한
당위성을 부여할 수 있음

공통 피드백



많은 팀에서 Optuna를 활용하여 하이퍼파라미터 튜닝을 진행한 것을 확인

공통 피드백



Optuna의 경우에는 하이퍼파라미터의 중요도를

plot_param_importances() 함수를 통해 플랏을 그려 확인 가능하고,

이를 바탕으로 어떤 변수에 중점을 둬서 튜닝할 지 고려해볼 수 있을 것임

2 공통 피드백

Q&A 및 모델에 대한 설명

4

1등팀 발표

2 1등팀 발표

3팀

김수빈 김민 안은선 조건우

다른 팀들도 모두 수고하셨습니다!