

2022-여름방학세미나

3팀

김현우 정희철 이주형
채소연 박윤아 서희나

발표 한 눈에 보기



1

PRE-PROCESSING

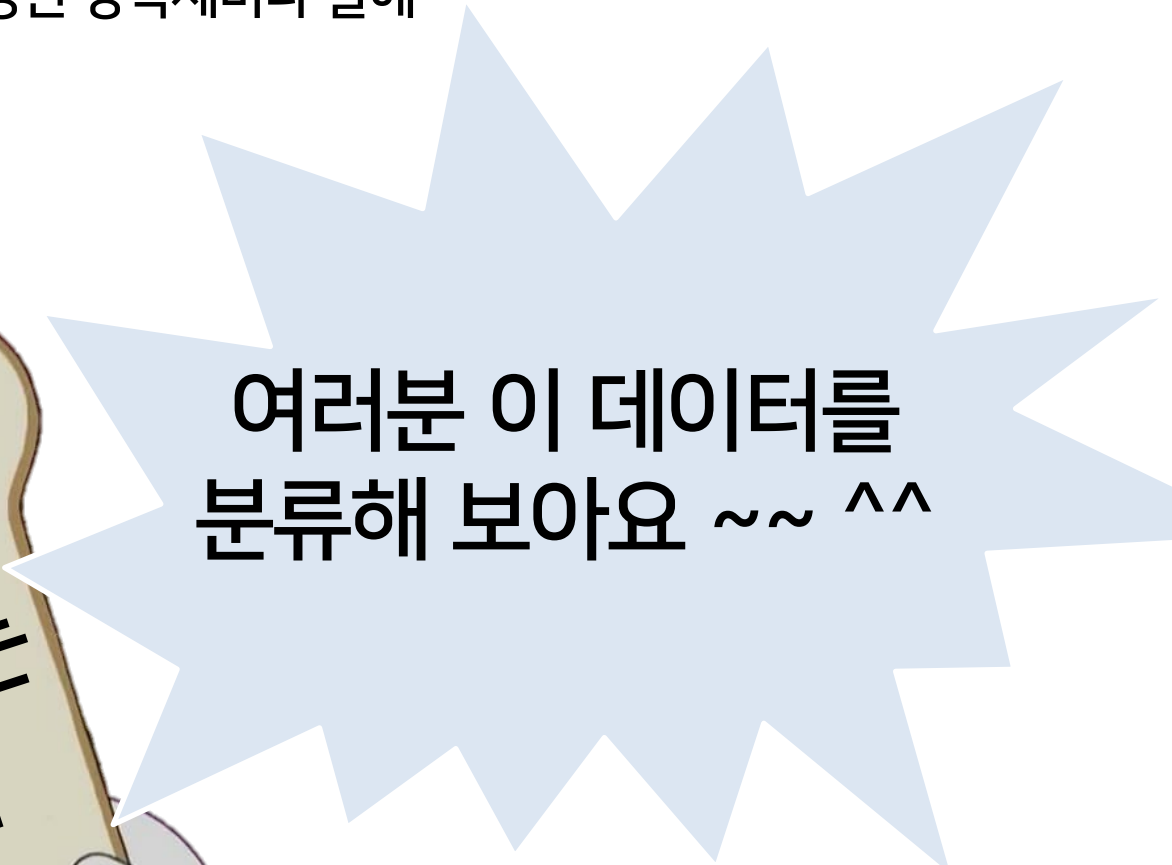
1

PRE-PROCESSING

어느 날 갑자기는 아니고.. 예정된 방학세미나 날에



정체를
알 수 없는
데이터



여러분 이 데이터를
분류해 보아요 ^^ ^^

학회장팀

EDA - 분석 목적 파악

주제

주어진 데이터를 활용하여 이진 분류 모델 만들기



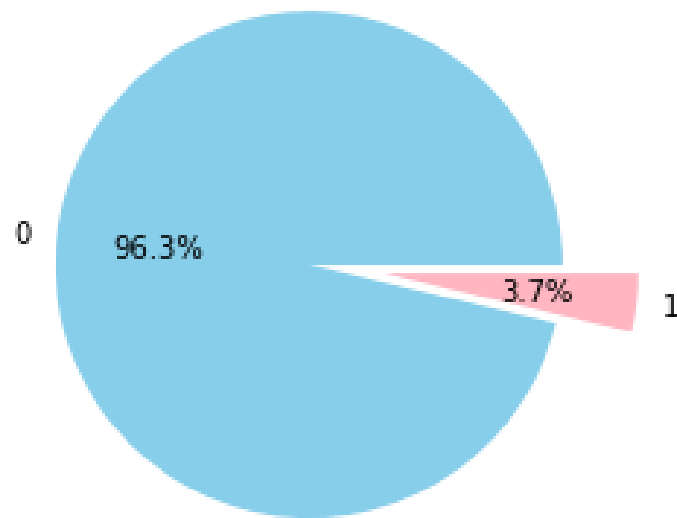
종속변수를 기준으로 먼저 EDA 진행

1

PRE-PROCESSING

EDA- 종속변수

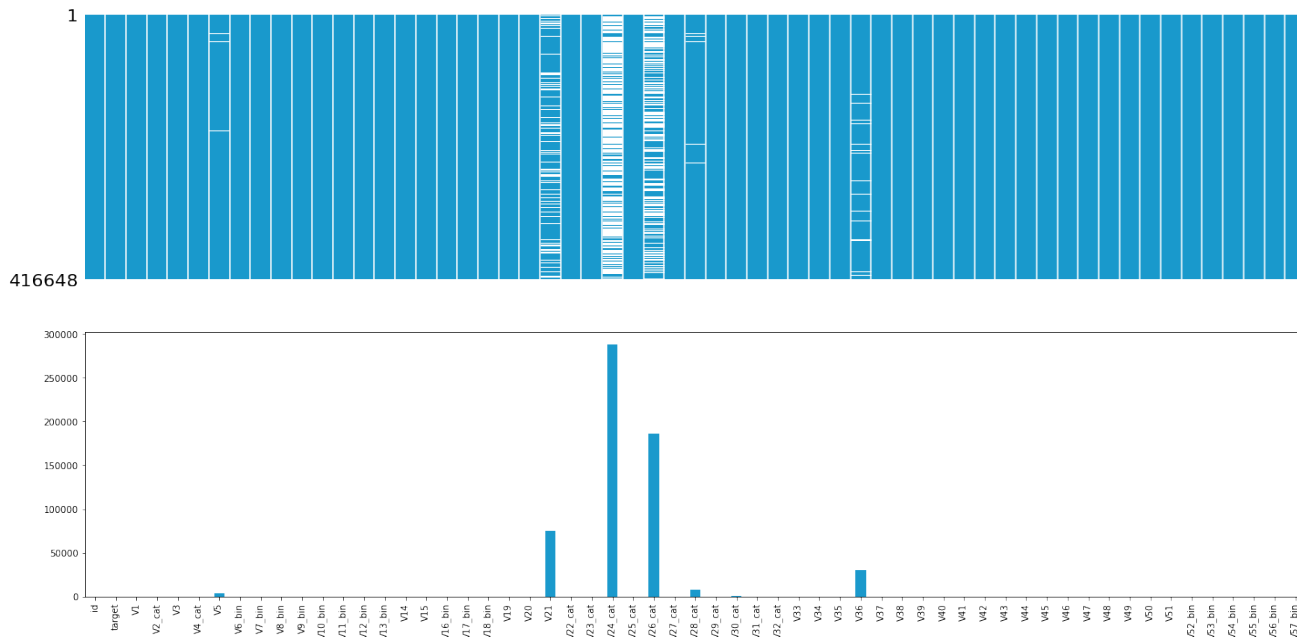
target
0
0
...
0



0과 1로 구성된 변수

96.3 : 3.7 정도의 비율로 클래스 불균형을 보임.

EDA - 설명변수



이런 데이터를 가져오기도 어렵겠다,,

- 1) 결측값이 존재하며 자료형 중 binary의 경우 결측값이 존재하지 않음을 확인
- 2) 결측값의 비중이 20% 이상인 V24_cat, V26_cat 변수는 제외 후 분석 진행

변수 분류

자료형이 표시되지 않은 변수 27개 중
고유하게 나타나는 값이 100개 이하인 변수를 Ordinal로 가정 후 분류

Binary (17)

Category (11)

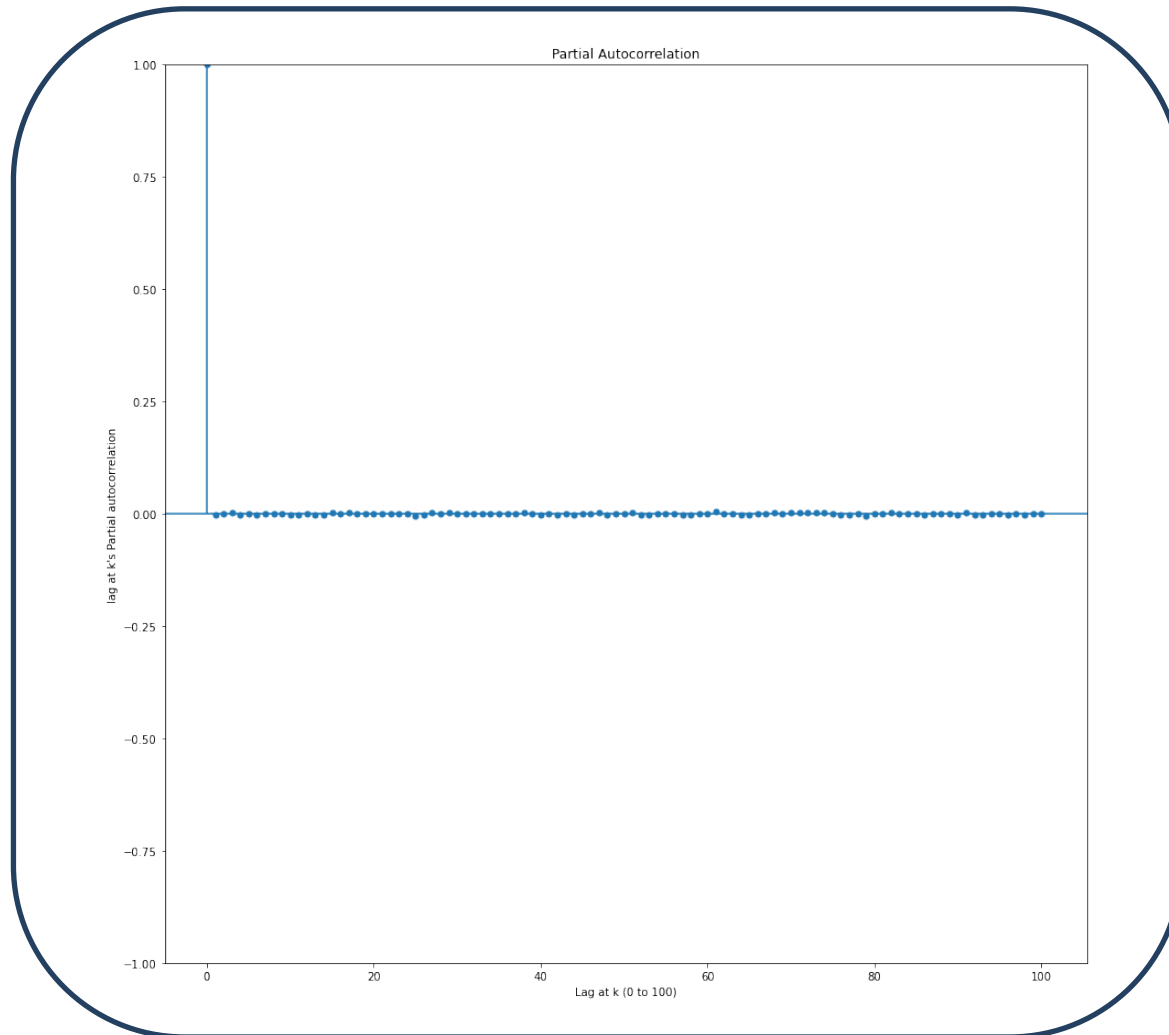
Numeric (4)

Ordinal (23)

1

PRE-PROCESSING

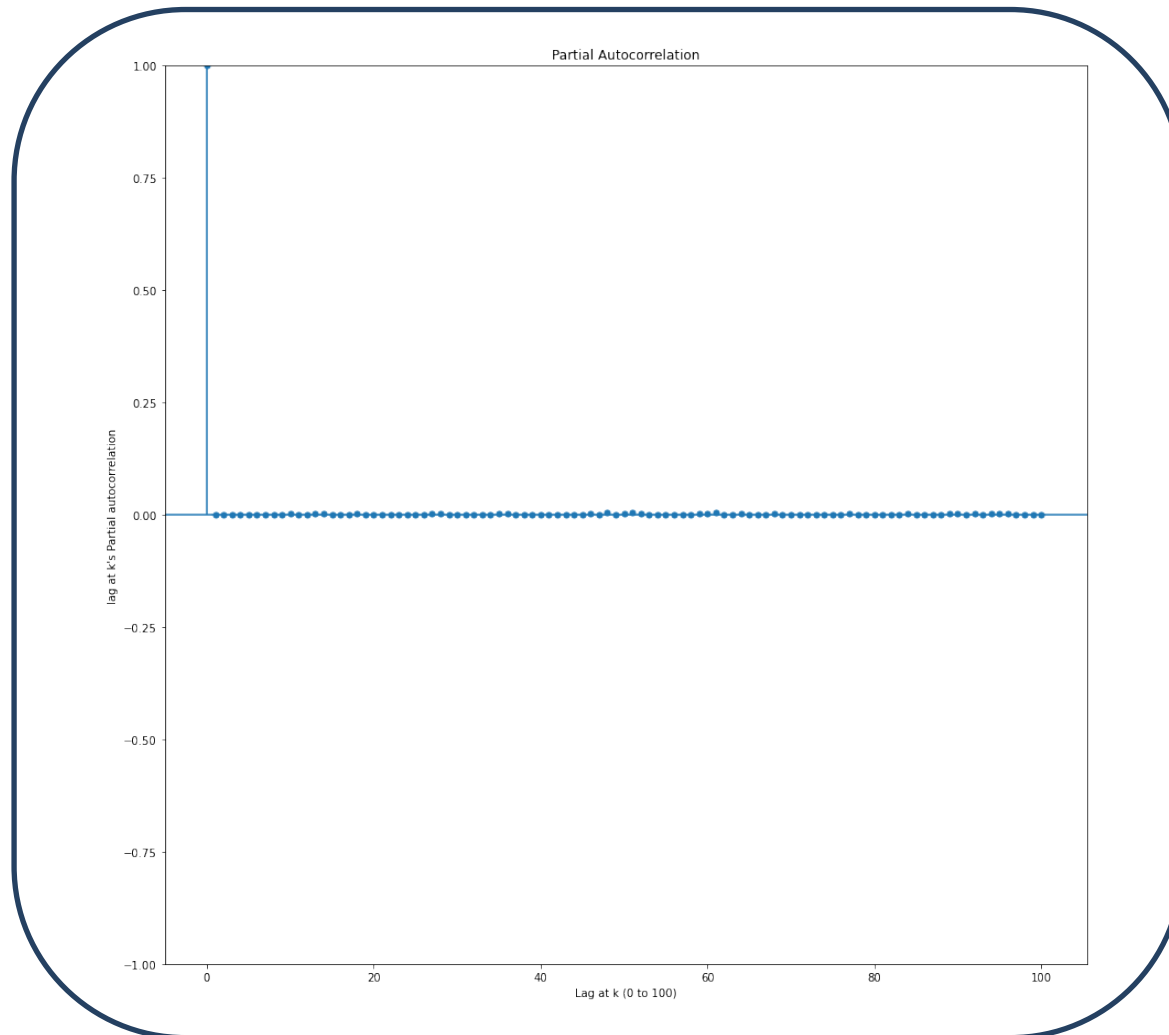
부분자기상관계수 그래프



1

PRE-PROCESSING

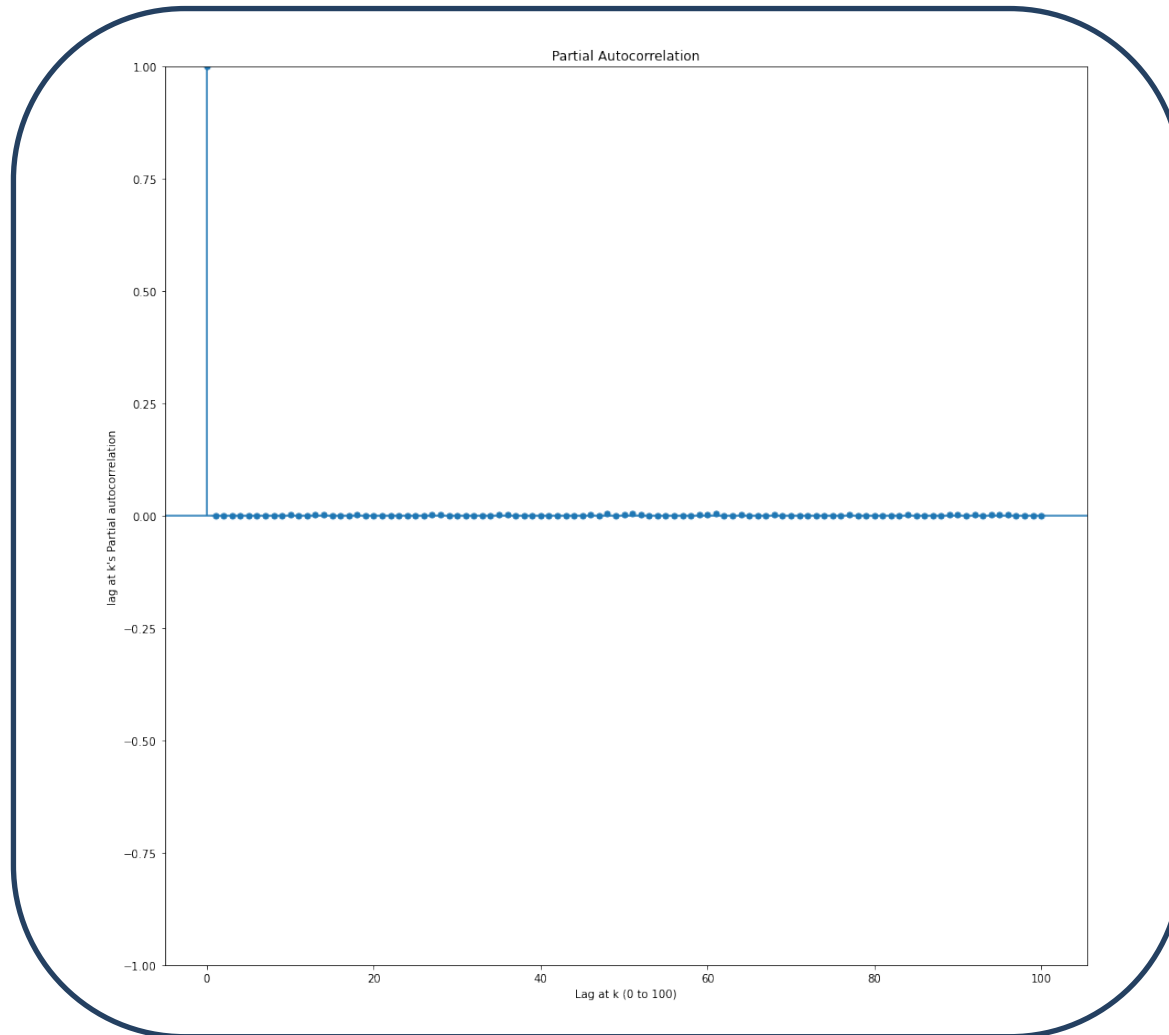
부분자기상관계수 그래프



1

PRE-PROCESSING

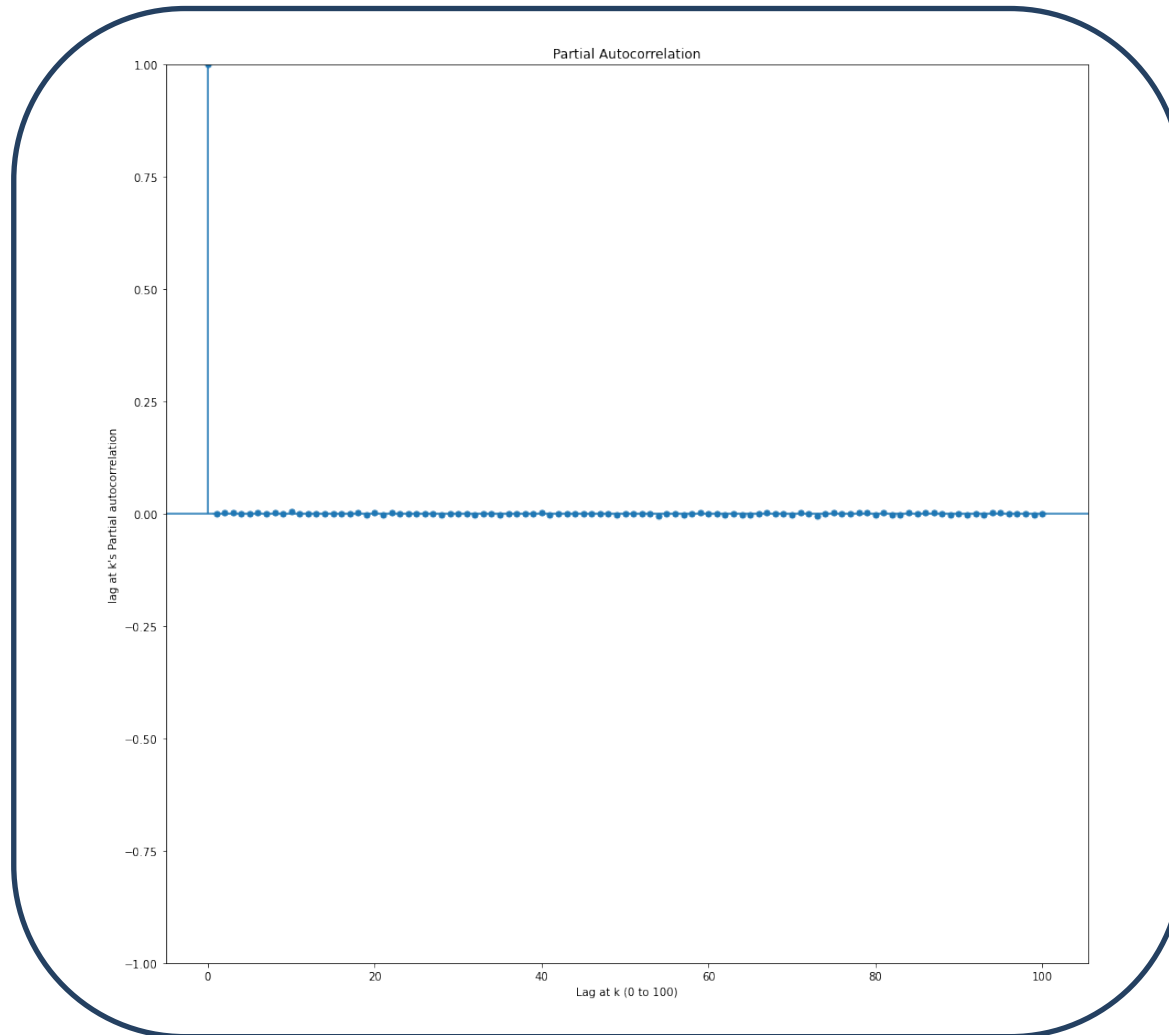
부분자기상관계수 그래프



1

PRE-PROCESSING

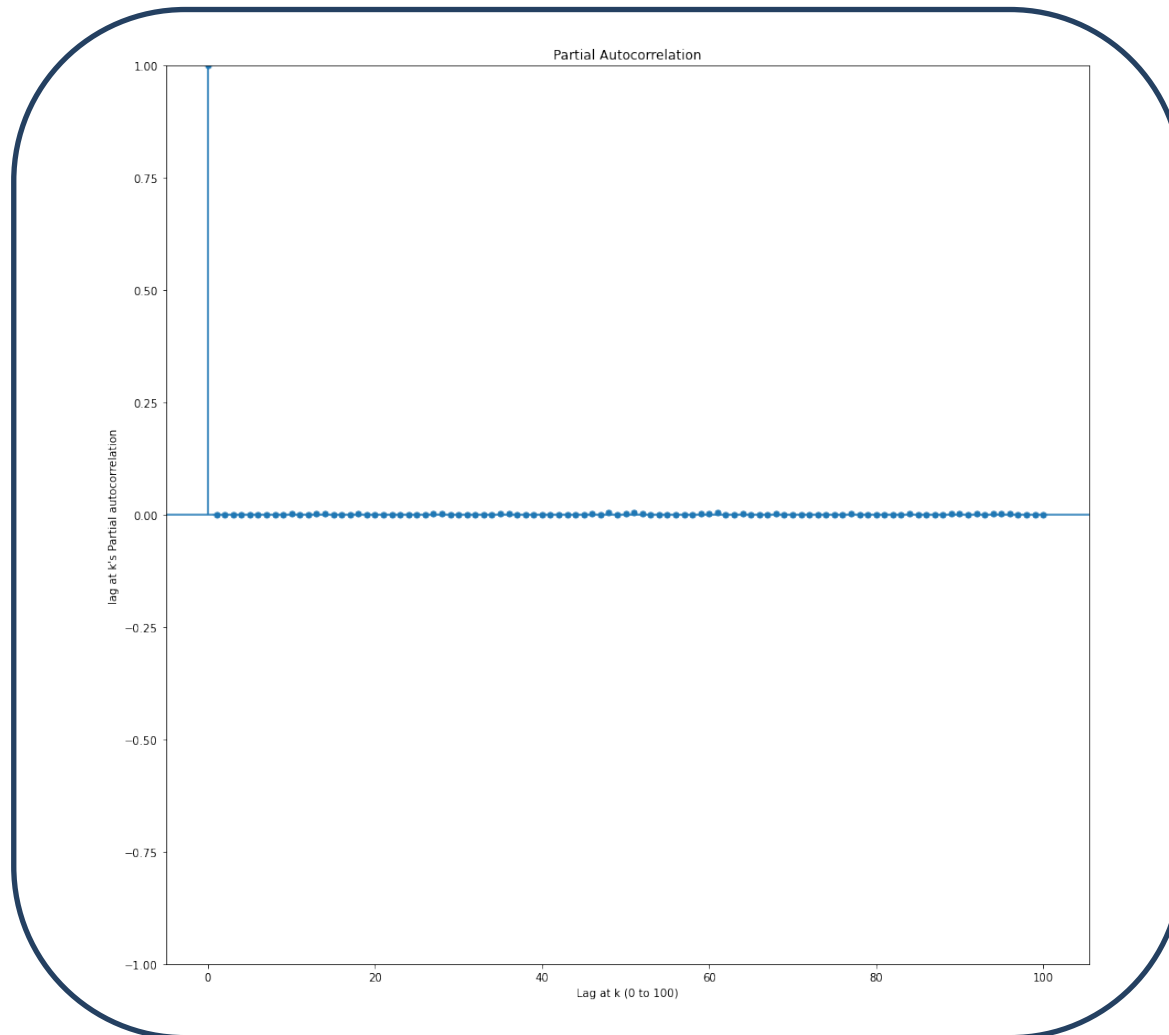
부분자기상관계수 그래프



1

PRE-PROCESSING

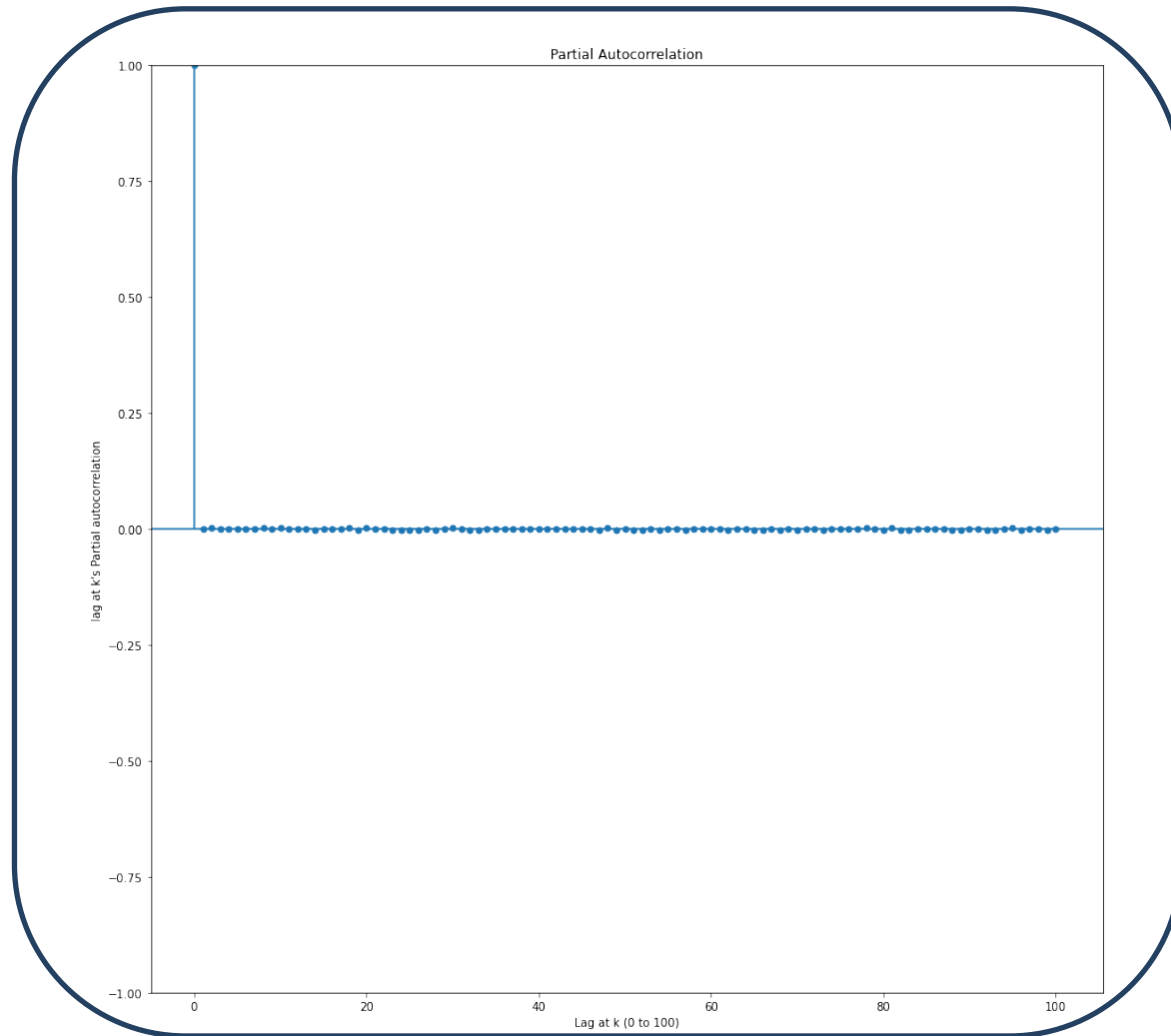
부분자기상관계수 그래프



1

PRE-PROCESSING

부분자기상관계수 그래프

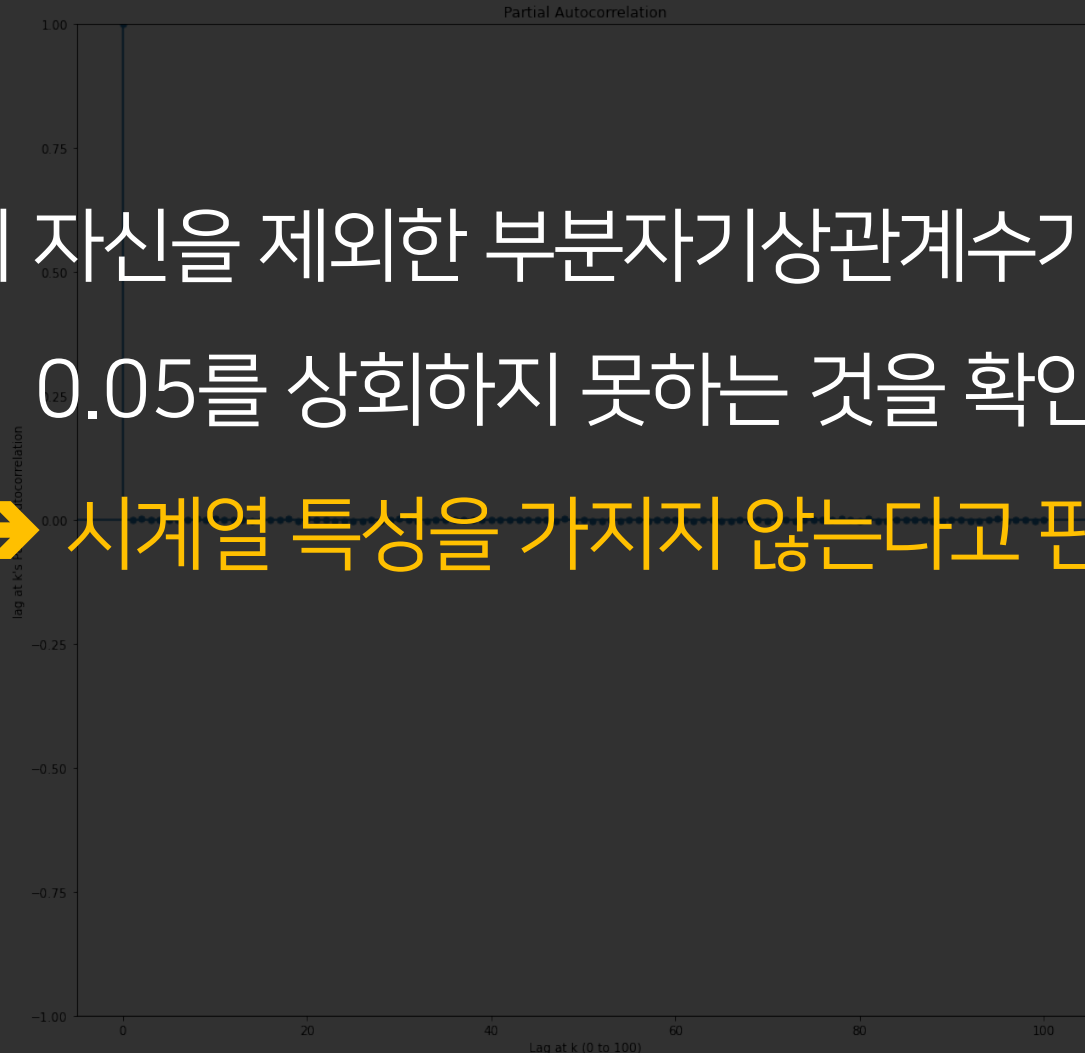


1

PRE-PROCESSING

부분자기상관계수 그래프

자기 자신을 제외한 부분자기상관계수가 모두
0.05를 상회하지 못하는 것을 확인
→ 시계열 특성을 가지지 않는다고 판단



1

PRE-PROCESSING

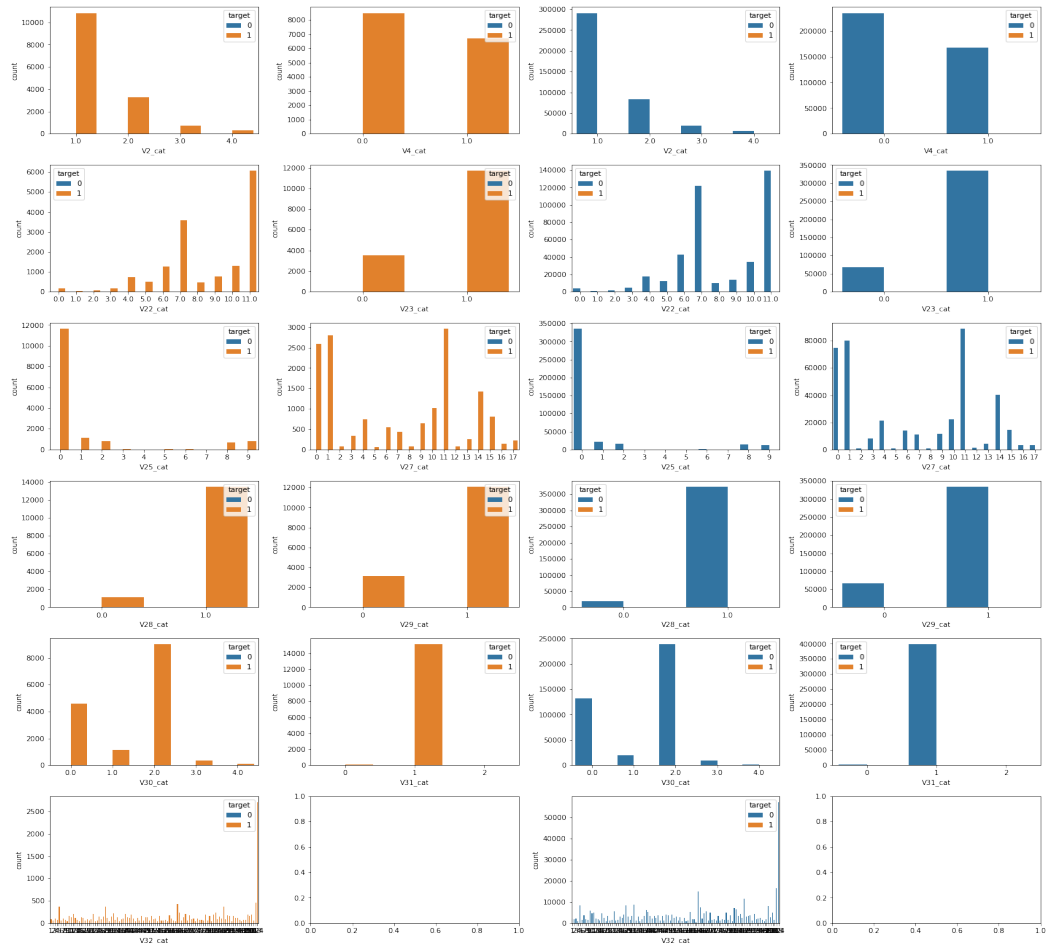
EDA - 데이터 분포 파악

Target에 따른
전체 데이터 분포 파악



Category, Numeric,
Binary, Ordinal 변수에
대해서 각각 진행

Categorical Variable



1

PRE-PROCESSING

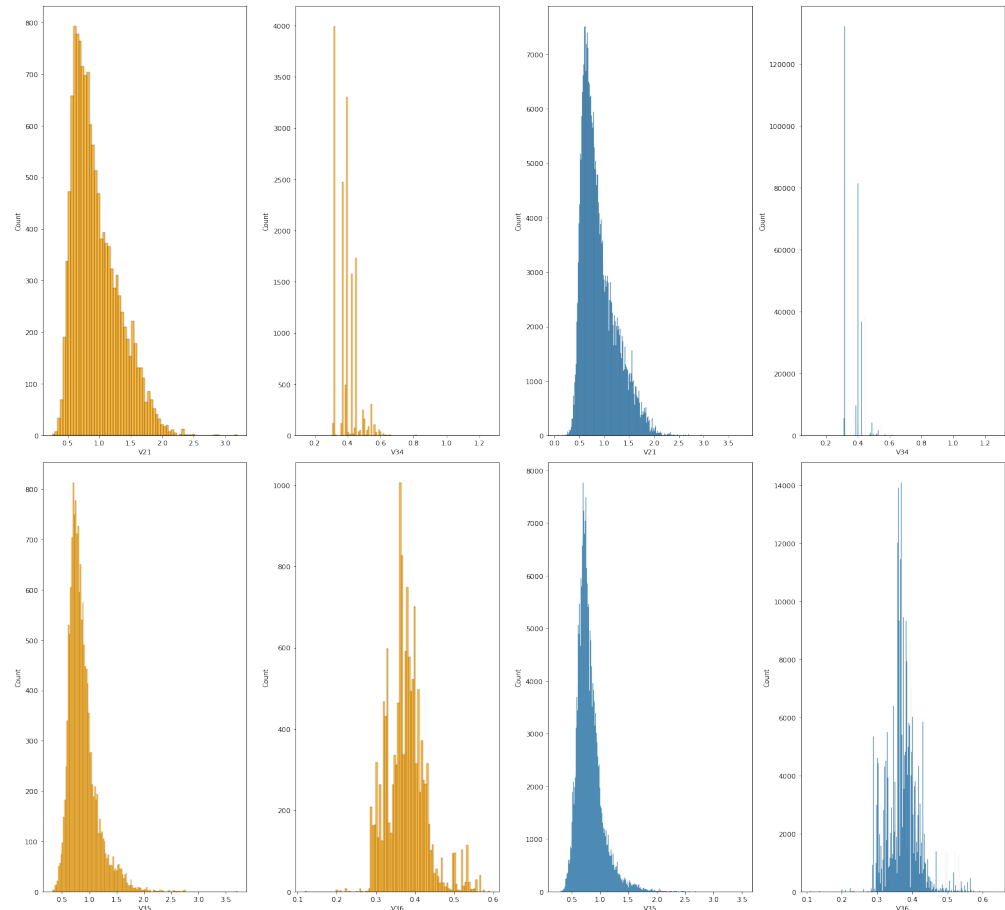
EDA - 데이터 분포 파악

Target에 따른
전체 데이터 분포 파악



Category, Numeric,
Binary, Ordinal 변수에
대해서 각각 진행

Numeric Variable



1

PRE-PROCESSING

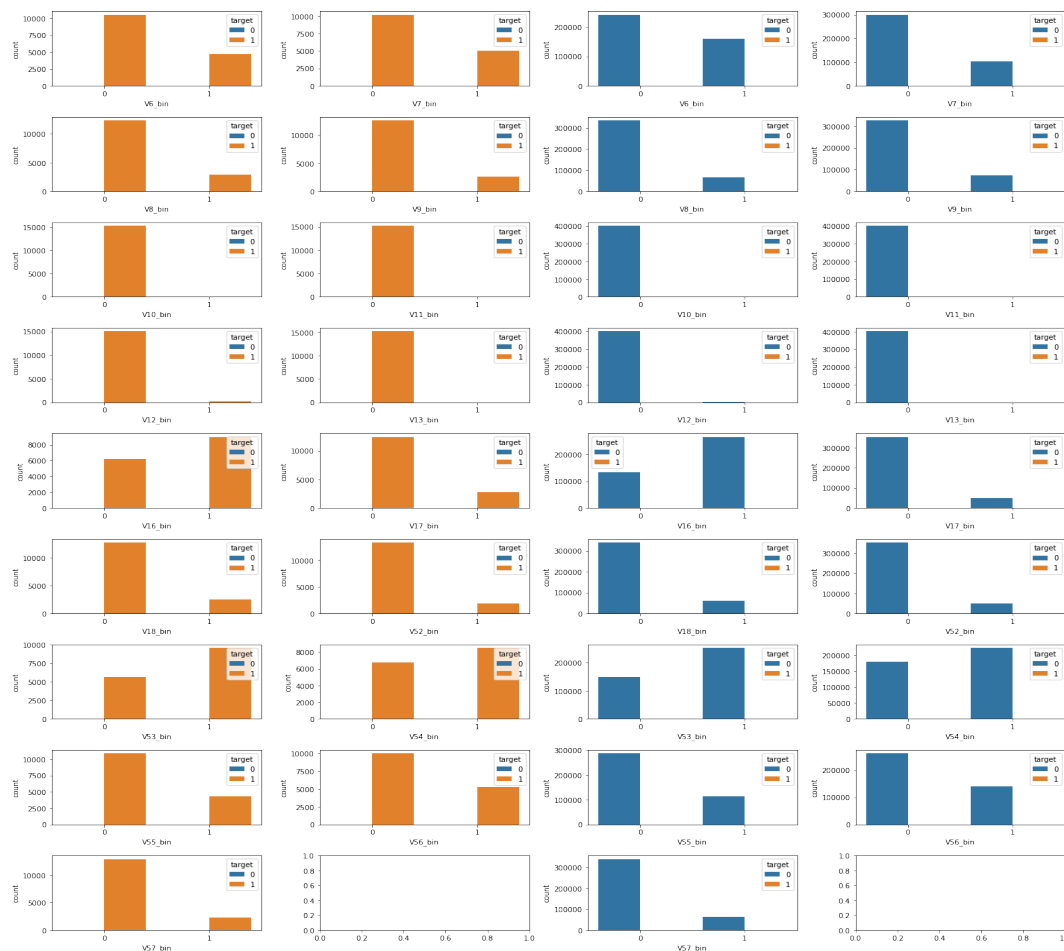
EDA - 데이터 분포 파악

Target에 따른
전체 데이터 분포 파악



Category, Numeric,
Binary, Ordinal 변수에
대해서 각각 진행

Binary Variable



PRE-PROCESSING

EDA - 데이터 분포 파악

Target에 따른
전체 데이터 분포 파악



Category, Numeric,
Binary, Ordinal 변수에
대해서 각각 진행

Ordinal Variable



1

PRE-PROCESSING

EDA - 데이터 분포 파악

전체적으로 분포가 비슷해 보이는 변수들이 많이 나타난 것으로 판단

모델링 및 예측 과정에서 결정적인 역할을 하지 못할 가능성이 높음 !

-> 동질성 검사 진행

2

검정 및 보간

동질성 검정

주어진 두 표본의 분포가 일치하는가 검정

범주형

카이 제곱 검정

KS test는 연속형 데이터에만 사용

H_0 : 비교하는 두 분포가 동질적이다

$P\text{-value} < 0.05 \rightarrow$ 비교하는 분포 이질성 존재

연속형

Kolmogorov Smirnov 검정

H_0 : 비교하는 두 분포가 동질적이다

$P\text{-value} < 0.05 \rightarrow$ 비교하는 분포 이질성 존재



만일 열 별 Target의 0/1에 대한 두 분포가 **이질적임**을 확인

\rightarrow 0/1을 결정짓는 중요한 요인으로 작용함을 기대

동질성 검정

주어진 두 표본의 분포가 일치하는가 검정

범주형

카이 제곱 검정

- Binary 변수 ('_bin')
- Category 변수 ('_cat')
- 그 외 고유값이 100개 이하인 변수는 Ordinal 변수라고 가정

연속형

Kolmogorov Smirnov 검정

그 외의 연속형 변수



P-value < 0.05인 변수들 (= Target에 따라 분포가 달라지는 변수) 선택

이를 보간하는 방식으로 NA imputation 진행

2

검정 및 보간



동질성 검정

Target 0/1 에 따른 분포가 이질적인 변수

주어진 두 표본의 분포가 일치하는가 검정

: V1, V3, V4_cat, V5, V6_bin, V7_bin, V8_bin, V9_bin, V11_bin, V12_bin, V15,
V16_bin, V17_bin, V18_bin, V19, V2_cat, V20, V21, V22_cat, V23_cat, V24_cat,
V25_cat, V27_cat, V28_cat, V29_cat, V30_cat, V32_cat, V34, V35, V36, V37,

카이 제곱 검정

Kolmogorov Smirnov 검정

- Binary 변수 ('_bin')



Target 0/1 에 따른 분포가 동질적인 변수

- 그 외 고유값이 100개 이하인 변수는

Ordinal 변수라고 가정

: V13_bin, V10_bin, V56_bin, V10_bin, V57_bin, V39, V52_bin, V48, V40, V26_cat,
V31_cat, V50, V33, V54_bin, V53_bin, V38, V46, V47, V43, V45, V49, V44, V51,
V41, V14, V42, V55_bin

P-value < 0.05인 변수들 (= Target에 따라 분포가 달라지는 변수) 선택

이를 보간하는 방식으로 NA imputation 진행

결측치 관련 탐색

범주형

?

◦ Cramer's V를 통해
상관관계를 구함

연속형

Pearson 상관계수로
상관관계를 구함

데이터를 보간하기 전, 각 변수 간의 상관관계를 확인함

⇒ 변수 간 상관 관계가 존재한다면, 이를 활용하여 결측치 보간에 활용할 수 있게 됨.

Cramer's V란?

- 범주형 변수들간의 연관성을 산출하기 위한 방법
- 0 ~ 1 사이에 값을 가지며, 클수록 변수간 연관이 큰 것을 의미

	1	2	...	J
1	n_{11}	n_{12}	...	n_{1J}
2	n_{21}	n_{22}	...	n_{2J}
\vdots	\vdots	\vdots	\ddots	\vdots
I	n_{41}	n_{42}	...	n_{IJ}

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{n}\right)^2}{\frac{n_{i.} \cdot n_{.j}}{n}}$$

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

결측치 관련 탐색

Category, Numeric,
Binary, Ordinal 변수 간
상관관계 확인

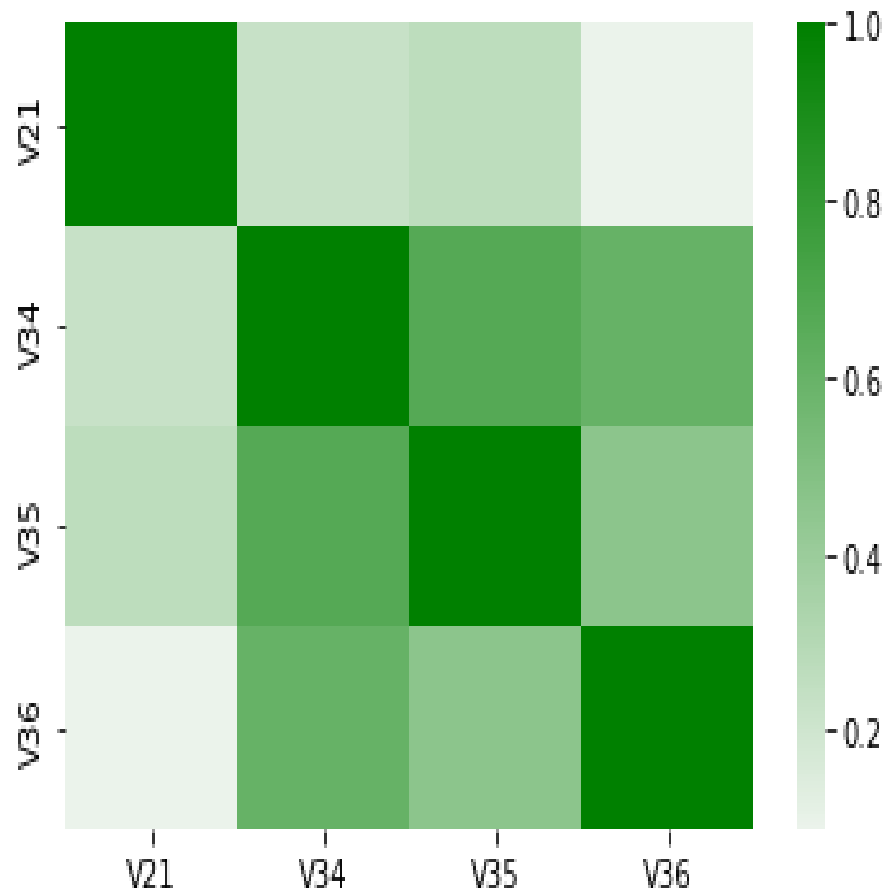


Numeric 변수간
높은 상관관계가 나타남



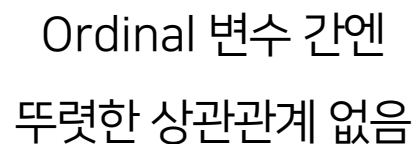
Linear한 방법으로
결측치를 보간!

Numeric Variable

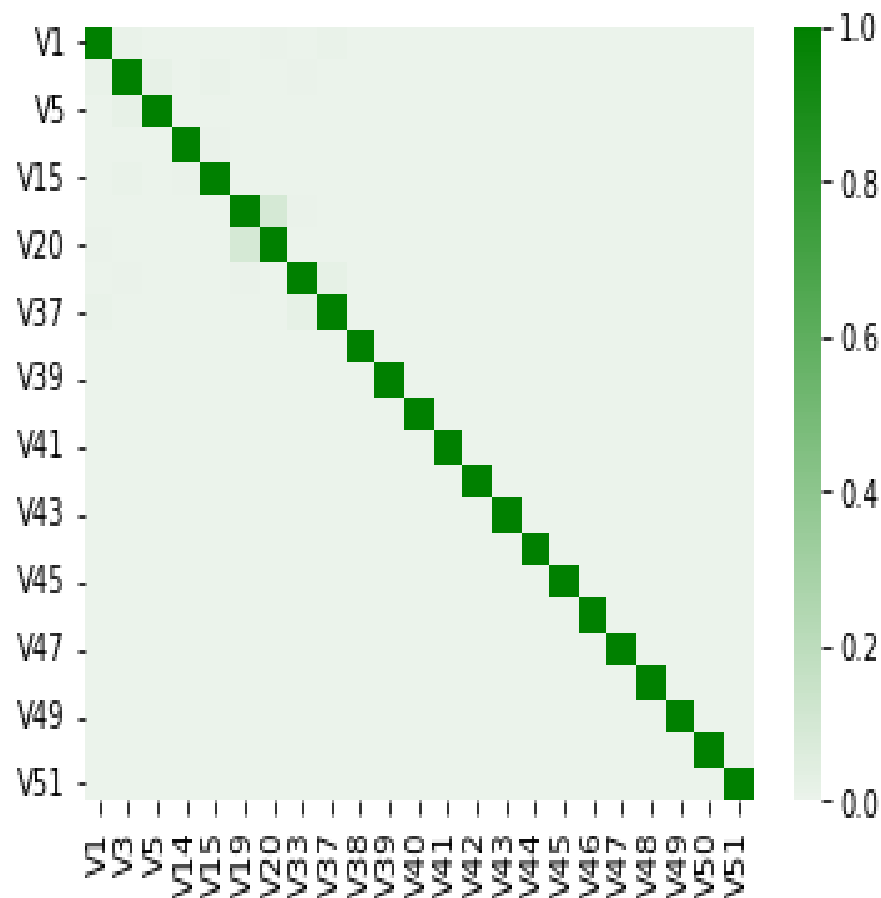


검정 및 보간

Category, Numeric,
Binary, Ordinal 변수 간
상관관계 확인

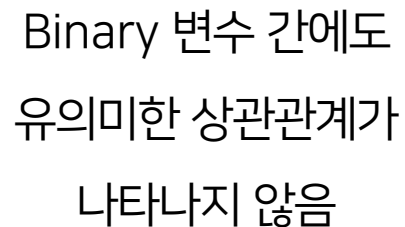


Ordinal Variable

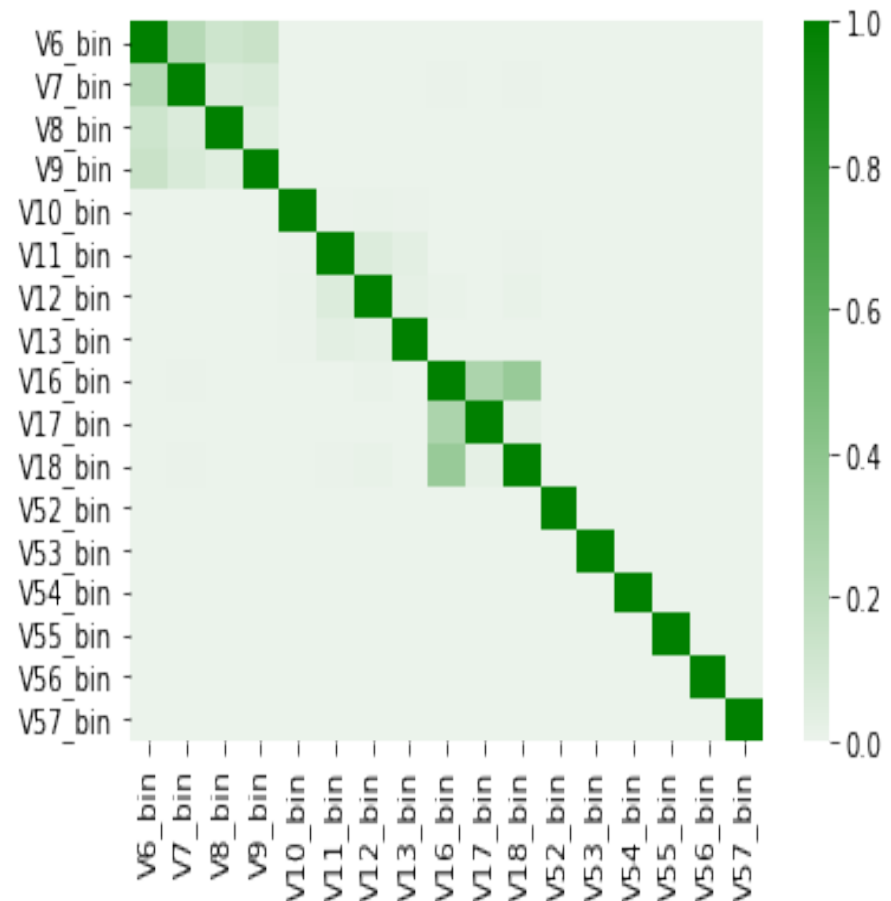


검정 및 보간

Category, Numeric,
Binary, Ordinal 변수 간
상관관계 확인

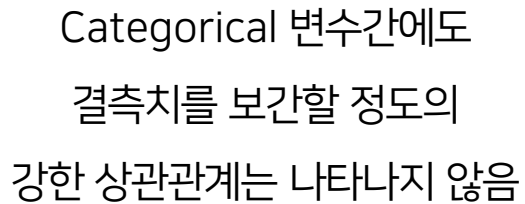


Binary Variable



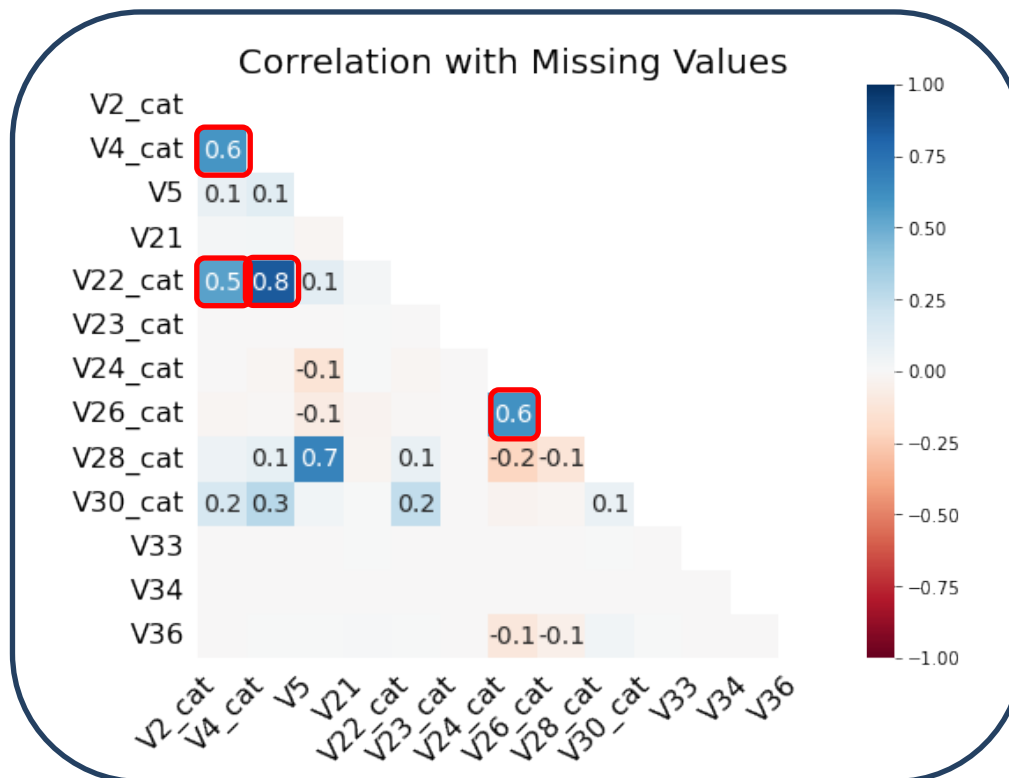
검정 및 보간

Category, Numeric,
Binary, Ordinal 변수 간
상관관계 확인



Heatmap showing the correlation matrix of the 12 V2-V32 cat visual areas. The color scale ranges from 0.0 (light green) to 1.0 (dark green). The diagonal elements are all 1.0. The off-diagonal elements show varying degrees of correlation, with V2_cat and V32_cat showing the highest correlation (approx. 0.9).

결측치 관련 탐색



결측치 간의 상관관계를 확인해본 결과 전반적으로 높은 상관관계를 보임

⇒ 순서형, 범주형 변수는 Non-Linear한 방법으로 결측치를 보간하기로 결정함

NA Imputation

KNN Imputation

KNN 분류 알고리즘을 이용하여
결측치를 최빈값, 중앙값 등으로
대체하는 방법

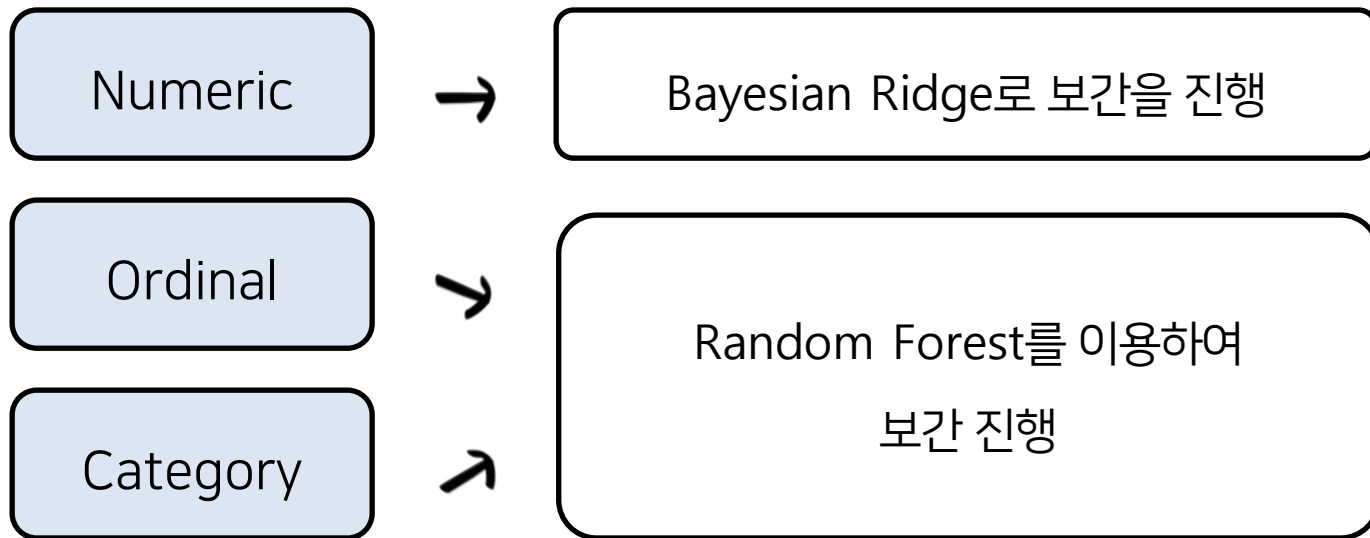
변수의 스케일이 다르고 개수가 많아
차원의 저주 가능성 존재
-> 채택하지 않음 !

Multiple Imputation

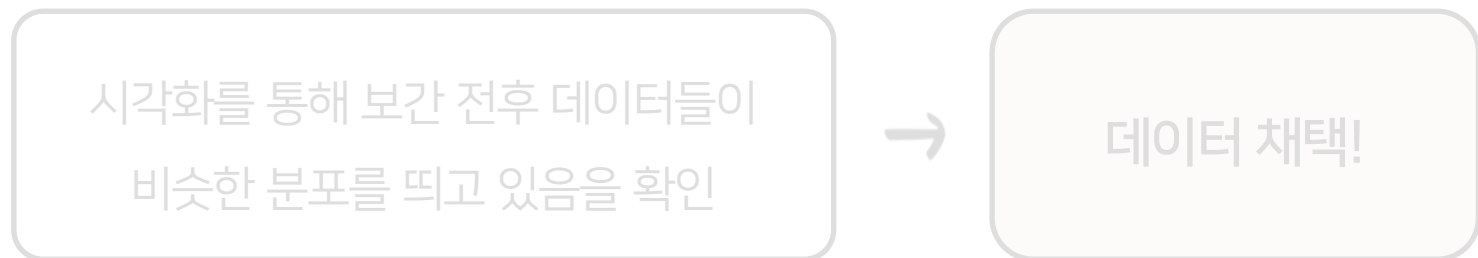
결측치가 존재하는 데이터에 대해
단순 대체법을 여러 번 시도하여
결측치를 대체하는 방법

다양한 기법들 중
Iterative Imputer를 이용하여
데이터를 보간하기로 결정

NA Imputation



앞서 본 시각화와 같은 방법 활용

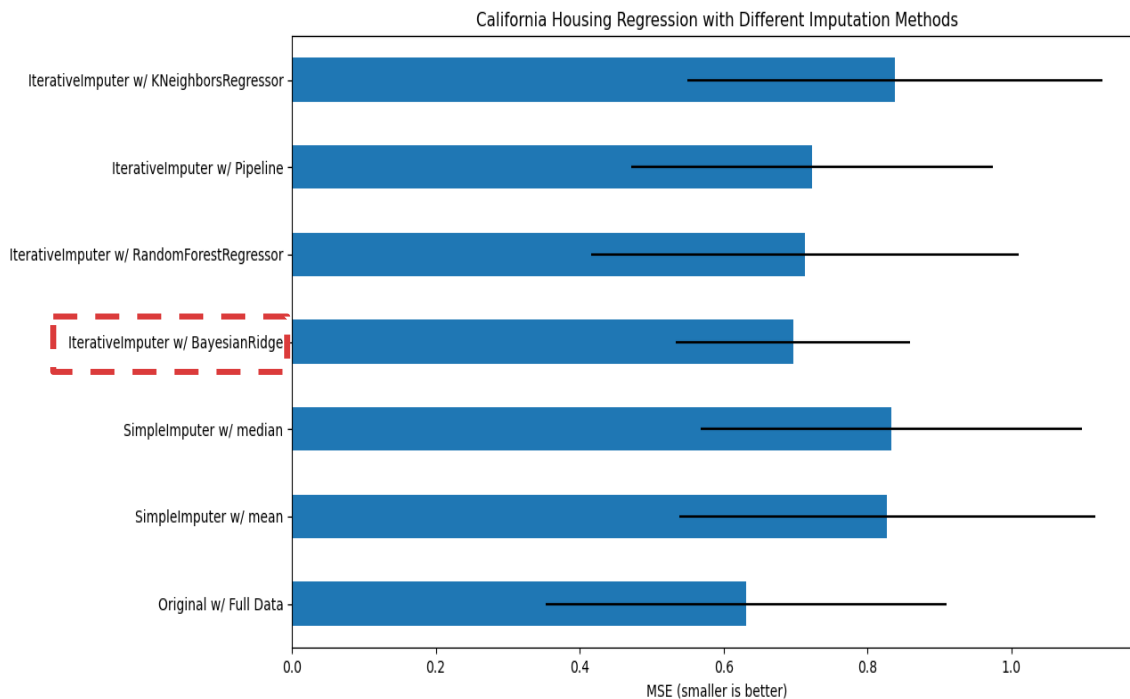


NA Imputation

Numeric



Bayesian Ridge로 보간을 진행



Scikit-learn을 이용하여

Scikit-learn에서 진행한
Imputer 성능 비교에서
Iterative Imputation
+Bayesian Ridge 보간이
가장 원본 데이터와 비슷하게
결측치를 보간해줬음

NA Imputation

Numeric

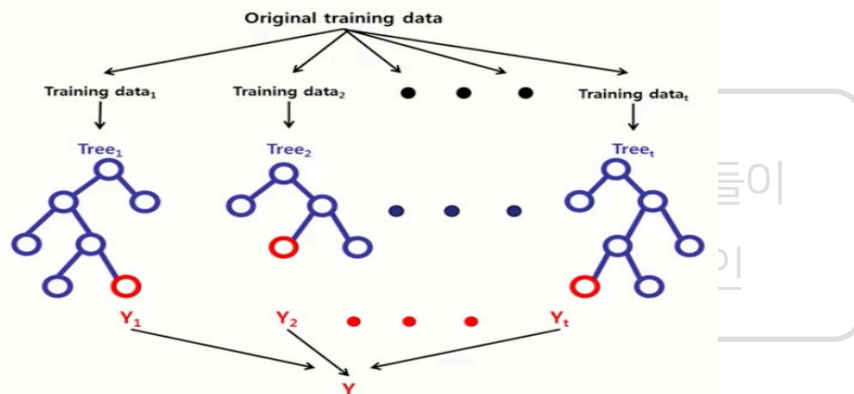


Bayesian Ridge로 보간을 진행

Ordinal

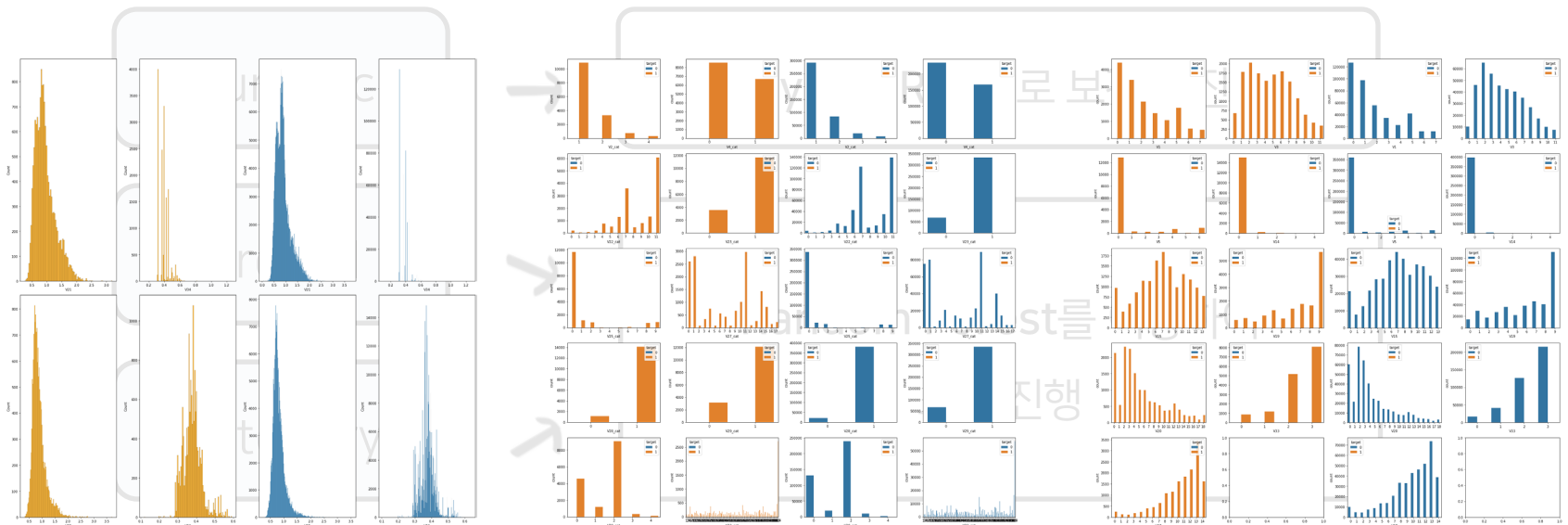


Category

Random Forest를 이용하여
보간 진행

모델이 효과적으로 데이터의
패턴을 파악하여
결측치를 보간할 수 있도록
Non-linear한 트리모델 사용

NA Imputation



앞서 본 시각화와 같은 방법 활용

시각화를 통해 보간 전후 데이터들이
비슷한 분포를 띄고 있음을 확인

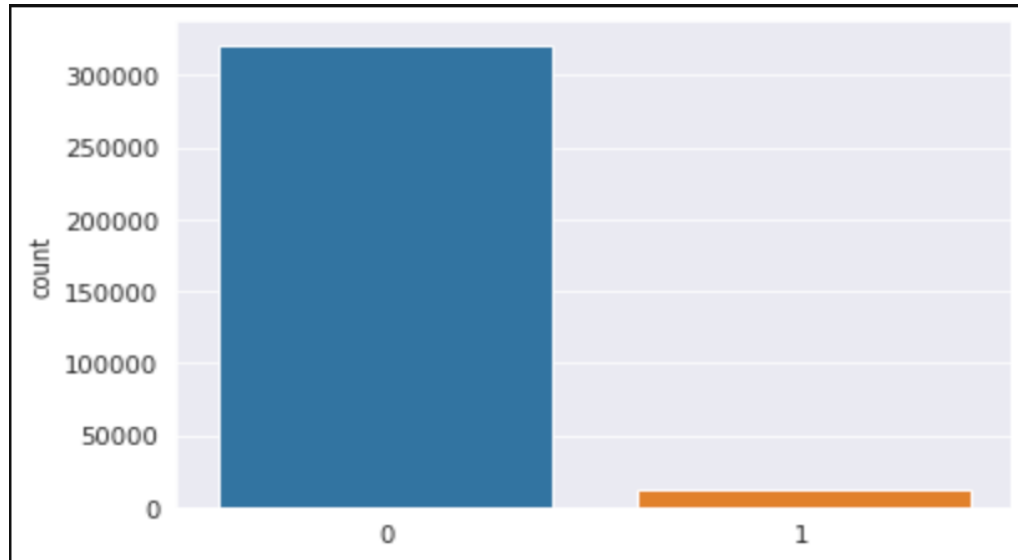


데이터 채택!

3

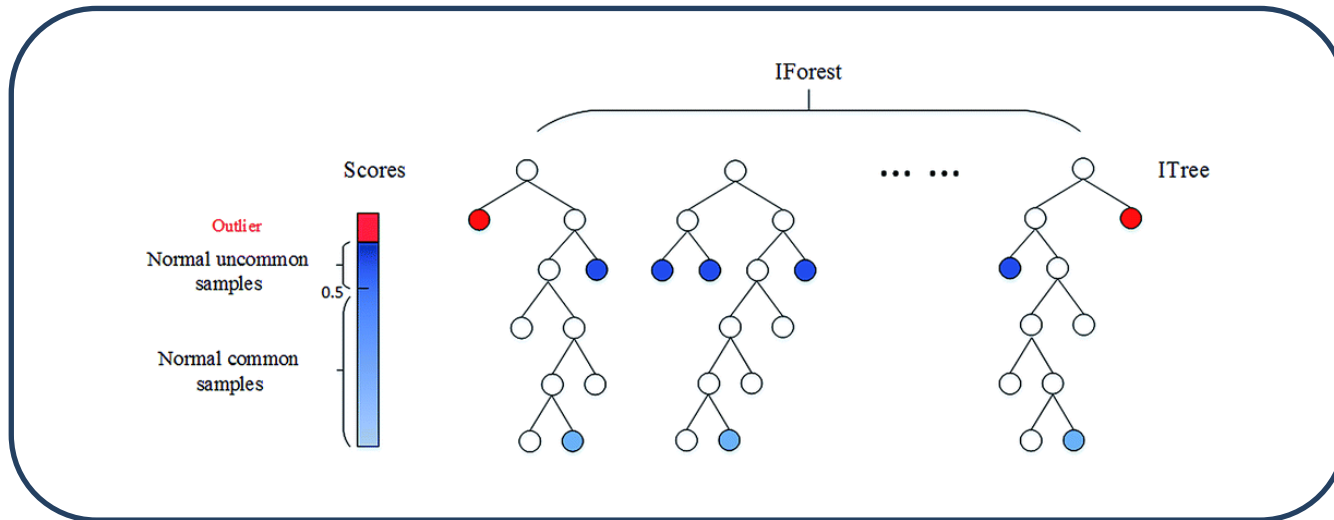
MODELING

1. Isolation Forest



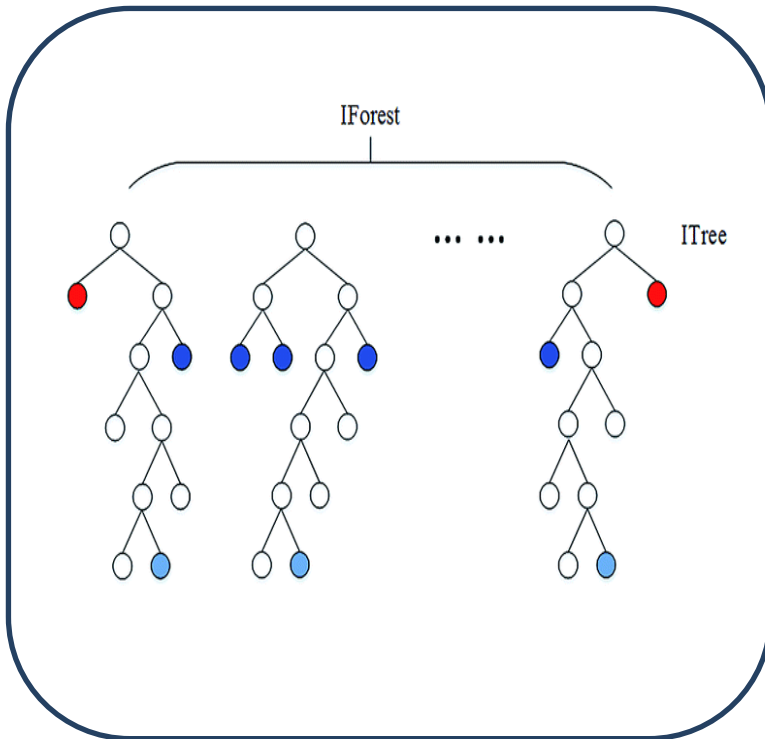
- Y의 클래스 불균형 확인
- 일반적인 이진분류 문제가 아닌 이상치 탐지 문제로 생각하고 접근
- 대표적인 이상치 탐지모델 Isolation Forest를 선택

1. Isolation Forest



- Decision Tree에서 파생된 모델로 비정상 데이터를 Tree의 가장 가까운 깊이에서 고립되게 만드는 모델
- 특정한 샘플이 고립되는 leaf 노드까지의 거리를 Outlier Score로 정의
- Root 노드까지의 평균거리가 짧을수록 Outlier Score가 높아지는 원리

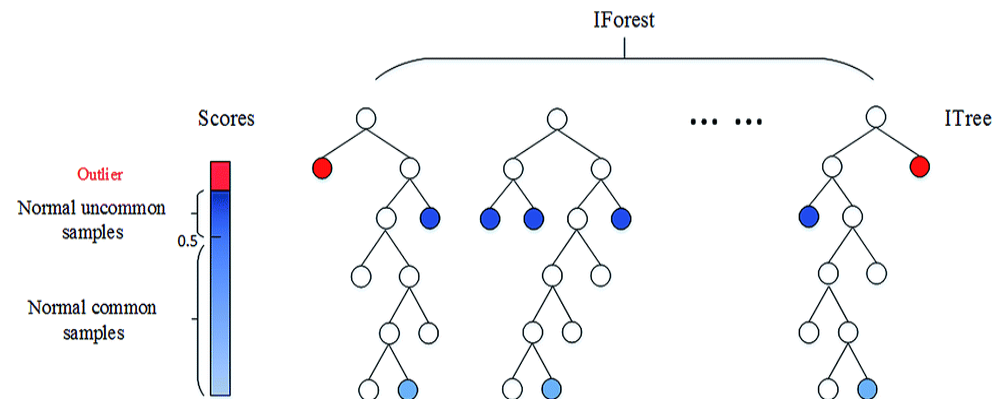
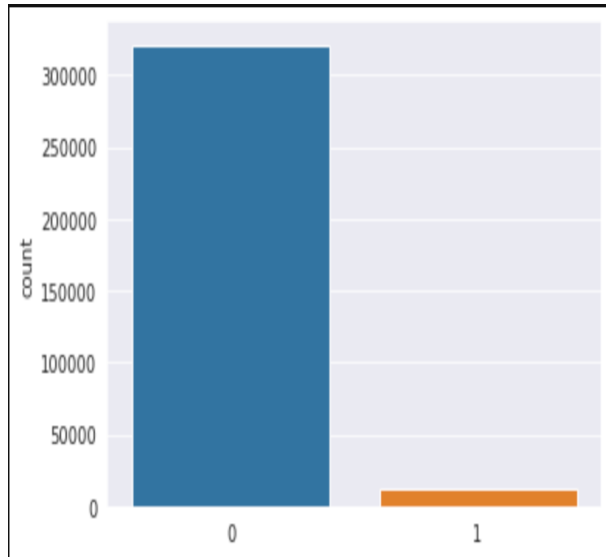
1. Isolation Forest



- 샘플링된 데이터와 선택된 변수를 활용하여 다양한 Tree를 생성해서 분기
→ 각 나무별로 각 데이터의 leaf 노드를 알 수 있음
- 만약 leaf 노드에 포함된 데이터가 1일 경우
→ 고립
- 1보다 크다면 최대 깊이로 한정되어 더 이상 분기하지 못한 것

3 MODELING

1. Isolation Forest



앞서 말한 것처럼 Y의 클래스 불균형이 심하여
하이퍼 파라미터 튜닝 없이도 0.508이라는 높은 f1-score를 얻어냄

→ 이후 파라미터 튜닝을 진행하여 모델의 성능을 높일 예정

2. Mixed Naive Bayes

Naive Bayes Classifier

The diagram shows the Naive Bayes formula with arrows pointing from descriptive labels to the corresponding terms in the equation:

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

Labels and their corresponding terms:

- Likelihood: $P(x | c)$
- Class Prior Probability: $P(c)$
- Posterior Probability: $P(c | x)$
- Predictor Prior Probability: $P(x)$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)$$

베이즈 정리를 바탕으로, 각 변수들 간의 조건부 독립을 가정하여
베이즈 정리의 복잡한 조건을 완화한 확률 모델

2. Mixed Naive Bayes

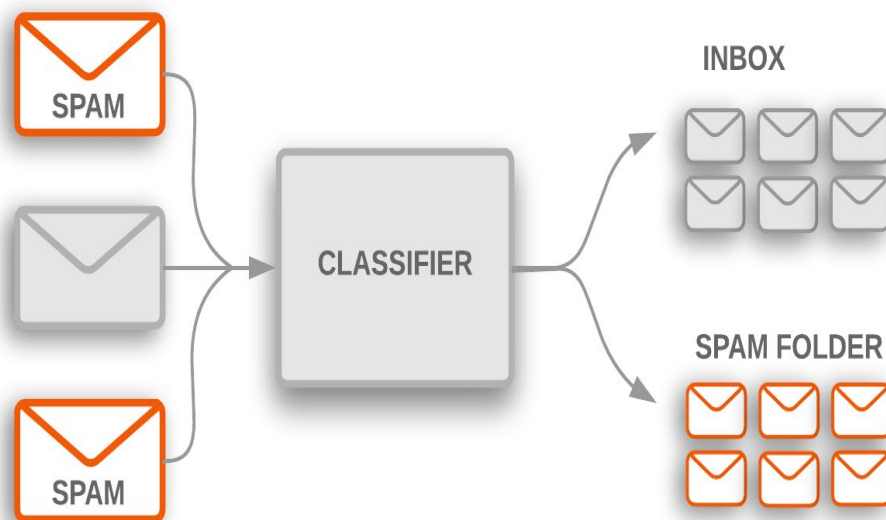
Mixed-Naive Bayes Classifier

$$\begin{aligned} p(y|X_1, X_2, \dots, X_n) \\ &\propto p(y, X_1, X_2, \dots, X_n) \\ &\propto p(y)p(X_1|y)p(X_2|y)p(X_3|y) \cdots \\ &\propto p(y) \prod_{i=1}^n p(X_i|y) \end{aligned}$$

변수들 간의 조건부 독립을 가정하여
설명변수의 값이 주어졌을 때의 특정 라벨이 나타날 확률을
단순 확률의 곱연산으로 쉽게 계산가능함

2. Mixed Naive Bayes

Naive Bayes Classifier



주로 스팸 메일 분류기 등 텍스트 분류에 많이 사용되며,
데이터에 변수가 많아도 변수간 조건부 독립 가정만 만족한다면 뛰어난 성능을 보임

2. Mixed Naive Bayes

Naive Bayes Classifier

연속형 - 정규분포

$$x_i | y \sim N(\mu_y, \sigma_y)$$

$$P(x_i | y) = \frac{1}{2\pi\sigma_y^2} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

MLE를 통해

μ_y, σ_y 를 추정

범주형 - 다항분포

$$x_i | y \sim \text{Multinomial}(\theta_{yi})$$

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

MLE를 통해

N_{yi}, N_y 를 추정

나이브 베이즈에선 $P(x_i | y)$ 에 대해

연속형에선 정규분포를 가정하고, 범주형에선 다항분포를 가정함.

2. Mixed Naive Bayes

Naive Bayes Classifier의 장점

장점 ①

MLE 추정량이 **단순 등장 빈도** 내지는 **확률 계산**으로 이루어짐



단순히 데이터에서 특정 값이 나타난 횟수만 세어주거나
확률 계산만 하면 되어 모델이 가볍고 학습·예측 속도가 빠름

장점 ②

설명변수의 수가 많고 **이산형 변수**가 많을수록 **효과적**임



데이터 셋의 설명변수 대부분을 차지하는 이산형 변수들을
최대한 활용하여 분류를 진행할 수 있게 됨.

2. Mixed Naive Bayes

Naive Bayes Classifier의 단점

한계 ①

변수들 간의 조건부 독립이라는 가정을 만족하지 못하면
모델의 성능이 저하됨.



그러나 데이터 간 상관관계가 잘 나타나지 않아
충분히 모델이 성능을 발휘할 수 있을 것으로 기대됨

한계 ②

학습데이터에 없는 값이 들어왔을 때,
확률값이 0이 되어 분류가 제대로 진행되지 않을 수도 있음.



라플라스 평활법(Laplace Smoothing)을
적용하여 이를 방지할 수 있음

2. Mixed Naive Bayes

Mixed Naive Bayes Classifier

연속형 변수와 이산형 변수가 섞여 있을 때

- ① 연속형 변수는 **정규분포**를 따른다고 가정
- ② 이산형 변수는 **다항분포**를 따른다고 가정
- ③ 이후 설명변수의 값이 주어졌을 때의 특정 라벨이 나타날 확률을
①과 ②를 계산하고 곱해주어 가장 나타날 확률이 높은 라벨로 예측

이를 바탕으로 현재 데이터 셋과 같은

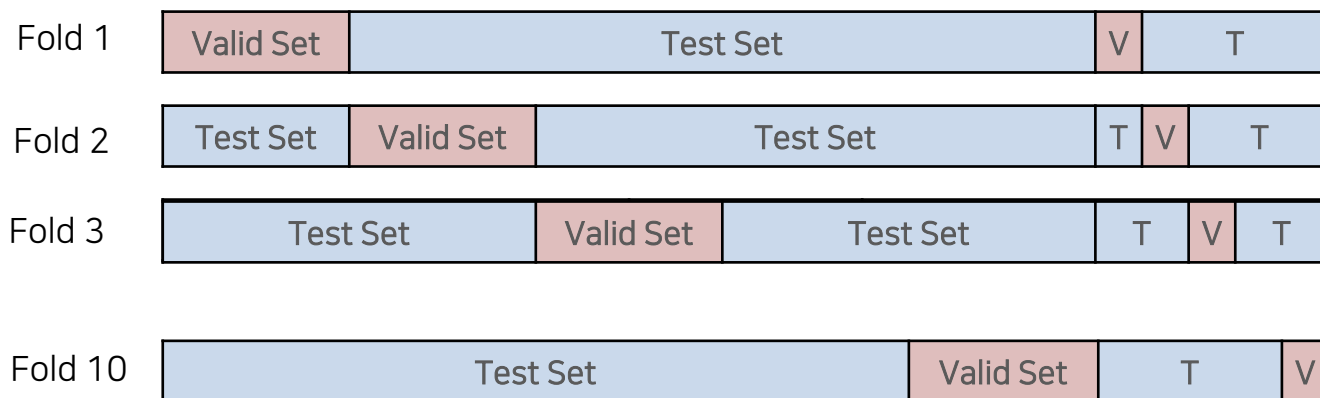
연속형 변수와 이산형 변수가 혼합된 데이터에서도 효과적으로

나이브 베이즈 모델을 통해 데이터 분류를 진행할 수 있음

Modeling Strategies

Stratified K-Fold CV

학습데이터	'target'==0	'target'==1
-------	-------------	-------------



Test F1_score:

$$\frac{1}{10} \sum_{i=1}^k f1_score_{(i)}$$

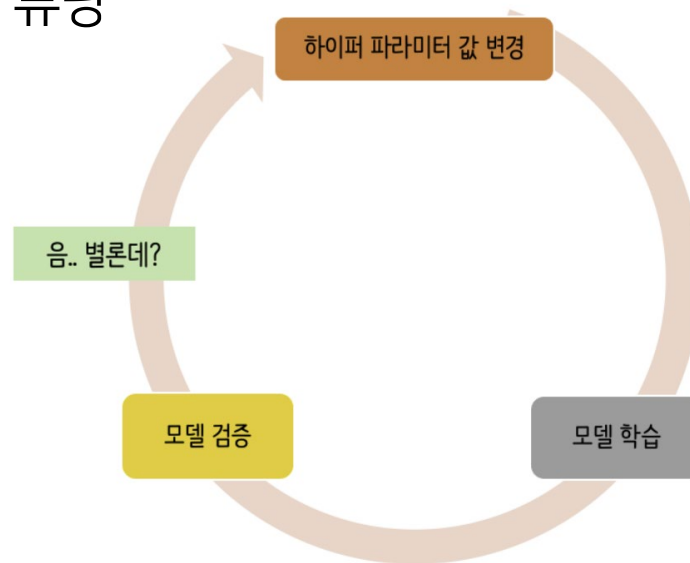
최대한 전체 데이터를 학습한 모델과 비슷한 모델로

분석 모델의 평가를 진행하기 위해

클래스 비율을 맞춰주는 Stratified K-Fold를 진행하여 모델을 평가함

Modeling Strategies

하이퍼 파라미터 튜닝

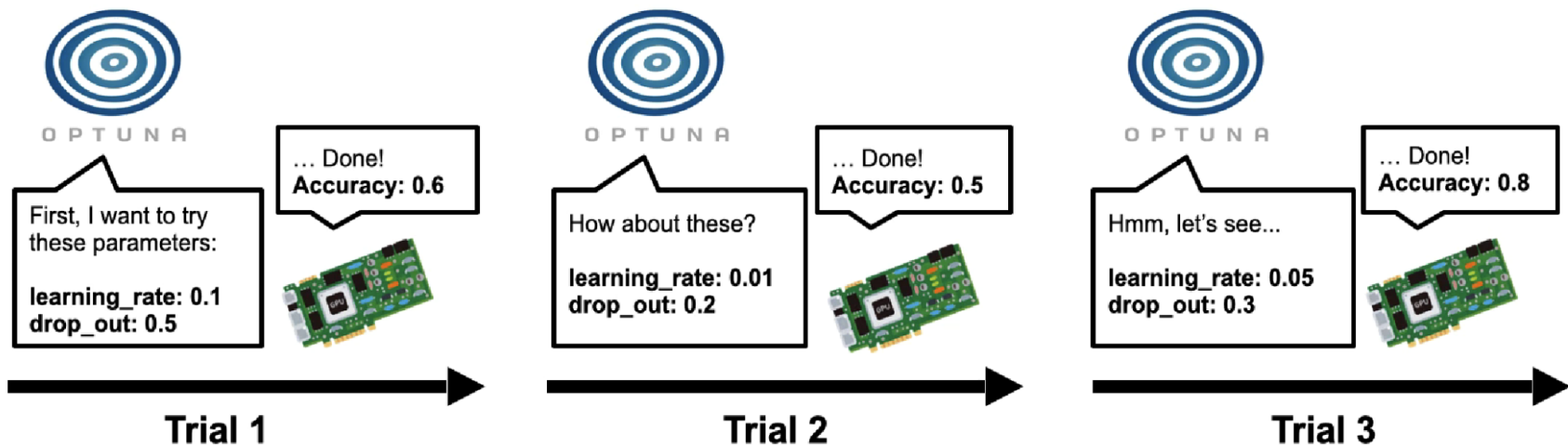


- 하이퍼 파라미터 : 최적의 훈련 모델 구현을 위해 모델에 설정하는 변수
(예: 학습률(Learning Rate) , 에포크 수 (훈련 반복 횟수), 가중치 초기화 유무 결정 등)
- 하이퍼 파라미터 튜닝의 종류
: Grid Search , Random Search , Bayesian Optimization

OPTUNA

: 머신러닝을 위해 설계된 자동 하이퍼 파라미터 최적화 software framework

- Sampler 로 각 하이퍼 파라미터의 값 선택 → 해당 조건에서 실험(trial)을 수행
- 해당 방향으로의 조정을 계속하는 것이 좋은지를 판단하여 최적의 하이퍼 파라미터 값을 찾아냄
- Pytorch, TensorFlow, Keras 등 여러 머신러닝 프레임워크와 함께 사용될 수 있음
- 시각화에 용이한 프레임 워크



Objective Function

: Optuna는 objective 함수의 결과값을 통해 파라미터의 성능을 평가하고 향후 실험(trial)에서 샘플링 할 위치를 결정함

```
def objective( trial ) :  
  
    params = {  
  
        '파라미터 1' : ( trial 의 범위 )  
        '파라미터 2' : ( trial 의 범위 )  
        '파라미터 3' : ( trial 의 범위 )  
    }  
  
    model=(** params)  
    (데이터 학습 코드 )  
  
    return 평가지표
```

Objective Function

: Optuna는 objective 함수의 결과값을 통해 파라미터의 성능을 평가하고 향후 실험(trial)에서 샘플링 할 위치를 결정함

```
def objective( trial ) :
```

```
    params = {
```

```
        '파라미터 1' : ( trial 의 범위 )
```

```
        '파라미터 2' : ( trial 의 범위 )
```


```
        '파라미터 3' : ( trial 의 범위 )
```

```
    }
```

```
    model=(** params)
```

```
    (데이터 학습 코드 )
```

```
    return 평가지표
```



모델에 설정할
파라미터 선언

Objective Function

: Optuna는 objective 함수의 결과값을 통해 파라미터의 성능을 평가하고 향후 실험(trial)에서 샘플링 할 위치를 결정함

```
def objective( trial ) :
```

```
    params = {
```

```
        '파라미터 1' : ( trial 의 범위 )
```

```
        '파라미터 2' : ( trial 의 범위 )
```


```
        '파라미터 3' : ( trial 의 범위 )
```

```
    }
```

```
    model=(** params)
```

```
    (데이터 학습 코드 )
```

```
    return 평가지표
```



파라미터의 범위에
해당하는 수 지정

Objective Function

: Optuna는 objective 함수의 결과값을 통해 파라미터의 성능을 평가하고 향후 실험(trial)에서 샘플링 할 위치를 결정함

```
def objective( trial ) :
```

```
    params = {
```

```
        '파라미터 1' : ( trial 의 범위 )
```

```
        '파라미터 2' : ( trial 의 범위 )
```


```
        '파라미터 3' : ( trial 의 범위 )
```

```
    }
```

```
    model=(** params)
```

```
    (데이터 학습 코드 )
```

```
    return 평가지표
```



시도하는 파라미터에
따른 모델 학습

Objective Function

: Optuna는 objective 함수의 결과값을 통해 파라미터의 성능을 평가하고 향후 실험(trial)에서 샘플링 할 위치를 결정함

```
def objective( trial ) :
```

```
    params = {
```

```
        '파라미터 1' : ( trial 의 범위 )
```

```
        '파라미터 2' : ( trial 의 범위 )
```

```
        '파라미터 3' : ( trial 의 범위 )
```

```
    }
```

```
    model=(** params)
```

```
    (데이터 학습 코드 )
```

```
    return 평가지표
```

← 학습의 결과를
평가지표로 반환

Isolation Forest의 최적의 파라미터 구하기

: 성능 스코어를 최대화/ 최소화하는 파라미터 값을 OPTUNA 가 반환

IsolationForest 모델의 파라미터

Parameters	Description	Value
n_estimators	나무의 개수 (디폴트 10)	148
max_samples	지정한 비율만큼 데이터 샘플링	0.637109665355488
contamination	전체 데이터에서 이상치의 비율	0.040562698102812676
max_features	사용할 feature의 개수	0.7392189971666111

→ 최종 성능 : 0.52160로
교차검증에서 높은 성능을 보임

2. Mixed Naive Bayes

Naive Bayes Model 평가

Model A

전처리를 통해
얻어낸 변수 중
이산형 변수만 사용한
Multinomial
Naive Bayes

Model B

Model A 변수
+ 결측치가 나타나지 않은
모든 이산형 변수를 사용한
Multinomial
Naive Bayes

Model C

Model B 변수
+ 연속형 변수도 사용한
3. Mixed Naive Bayes

Naïve Bayes 모델은 평가를 위하여 하이퍼 파라미터 튜닝대신
3가지 Naive Bayes 모델을 준비하고,
3개의 모델의 평균 f1-score를 비교하였음

2. Mixed Naive Bayes

Naive Bayes Model 평가

Model	Model A (Multinomial NB)	Model B (Multinomial NB)	ModelC (Mixed NB)
평균 Valid f1_score	0.51631	0.51638	0.52024

평가 결과 Mixed Naive Bayes 모델이
성능이 가장 좋게 나온 것을 확인하게 됨

그외 사용한 모델들

LGBM

- F1 Score -

0.490997

Random Forest

- F1 Score -

0.490692

XGBoost

- F1 Score -

0.50800

Gradient Boosting

- F1 Score -

0.490691

그외 사용한 모델들

LGBM

- F1 Score -

0.499997

Random
Forest

- F1 Score -

0.490692

비교적 F1 Score가 작게 나와

사용 안함

XGBoost

- F1 Score -

0.50800

Gradient
Boosting

- F1 Score -

0.490691

4

PREDICTION

모델 비교 및 예측

교차검증 결과 정리 및 모델 평가

해당 모델을 사용하여 예측!

Model	Isolation Forest	Isolation Forest + Optuna	Multinomial NB	Mixed NB
평균 Valid f1_score	0.508	0.52160	0.51638	0.52024

모델링 과정에서 하이퍼 파라미터 튜닝을 진행한 Isolation Forest와 Mixed Naive Bayes가 높은 성능을 보였음을 확인하였음

➡ 2개의 모델을 둘 다 채택하여 각 모델로 테스트 데이터의 예측을 진행함

모델 비교 및 예측

모델의 최종 예측 결과 평가

Model	Isolation Forest	Isolation Forest + Optuna	Multinomial NB	Mixed NB
평균 Valid f1_score	0.508	0.52160	0.51638	0.52024



Predict

Model	Mixed Naive Bayes	Isolation Forest + Optuna
Test f1_score	0.51790	0.51720

두 모델 다 테스트 데이터에 대해서도 높은 f1-score를 보여주었으나,
테스트 데이터에서 높은 f1-score를 보이고 학습 및 예측 속도가 빠른
Mixed Naive Bayes 모델을 최종 모델로 선정하게 됨.

감사합니다