

주택가격-고급회귀기법

다중 공선성과의 LASSO기법

모델

데이터를 읽기

데이터 전처리

nan값이 존재하는가?

상관계수와 데이터 변수의 패턴이나 유사성을 통하여 nan값을 추론할수있는가??

선형이면 선형회귀

비선형이면 의사결정트리, 랜덤포레스트, 그래디언트, 신경망

랜덤포레스트로 nan값을 채웠는데 0.66이 나왔다

Light GBM은 트리 리프방향으로 성장 수직성장

다른알고리즘은 수평성장

왜도 첨도를 보기

꼭 feature_importance를 보자!!

데이터증강

<https://www.kaggle.com/code/jiweiliu/lgb-2-leaves-augment>

<https://www.kaggle.com/code/yag320/list-of-fake-samples-and-public-private-lb-split>

skf.split(df_train, df_train['target']) 클래스 레이블을 기준으로...나누어준다

웬만하면 k-fold로 하고..

k-fold로 할때., 누적해서 값을 더한다음 k-fold 인자 n으로 나누어서 최종 예측값이랑 비교하자

EDA

eda를 통하여 모델의 방향성을 정하고 필요없는 컬럼을 없앤다..

데이터 종류나 도메인에 따른 eda와 전처리를 정해보자...

랜덤포레스트

보우팅...

다중대체

작업을 할 때 이걸 생각했었다..하나의 컬럼이 빈값이라면 다른 컬럼을 사용하여 그 빈값을 추론할텐데, 만약에 다른컬럼들에도 동시에 빈값이 존재한다면 어떻게 해야하나?

M.I.는 시뮬레이션을 사용하고 데이터에 불규칙 잡음 (random noises)들을 더하기 때문에 초기에 M.I.가 받아들여지지 않게 될 수 (unacceptable) 있다...

현실적이고 신뢰할수있는 값인가???

vs

완전케이스분석

다중대체 동시추정법..