클린업 2주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1-3주차 패키지 문제의 조건 및 힌트는 Python을 기준으로 하지만, R을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 1 : Data Preprocessing

2주차 패키지의 분석 과제는 Decision Tree 기반의 Boosting에 대해서 다뤄보고자 합니다. Boosting이란 기존의 모델을 지속적으로 업데이트를 시켜 성능을 높이는 방법을 의미합니다. 대표적으로 XGboost와 LightGBM이 있습니다. XGboost와 LightGBM은 Kaggle이나 Dacon에서 매우 뛰어난 성능을 보여주고 있는 모델이고, 특히 10000개 이상의 많은 데이터를 활용하여 예측을 하는 경우에는 매우 좋은 성능을 보여줍니다.

두 모델을 사용하여 성능이 높은 회귀 모델을 모델링하는 것을 목표로 하며, 트리모델의 구조와 코드를 짜는 과정에 대해서 이해해보도록 합시다!

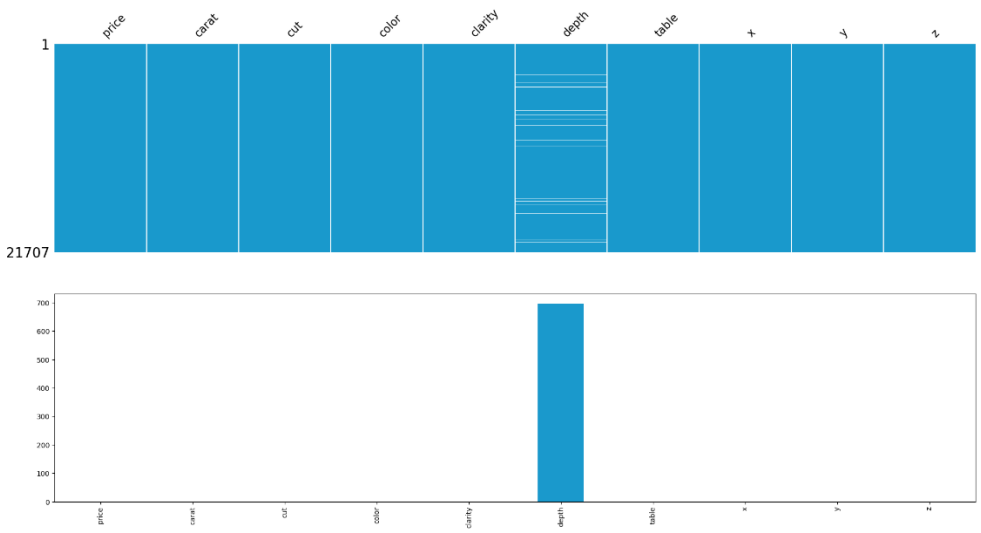
문제1. Train, test 데이터를 불러온 뒤 데이터의 구조를 파악하세요.

문제2. Train과 Test 데이터에서 ID_code는 필요하지 않으니 해당 열을 삭제해주세요.

(HINT) pandas에서 drop기능을 활용하여 해당 열을 삭제 가능합니다.

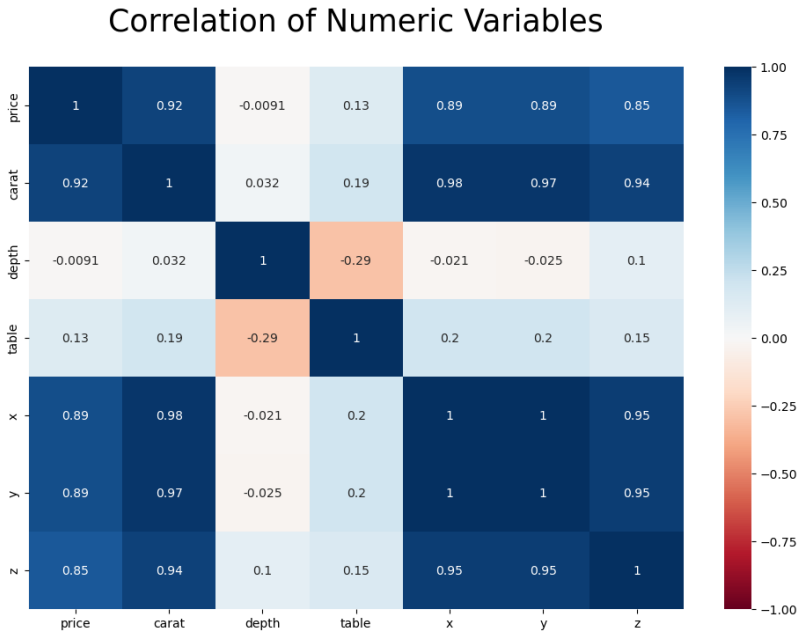
문제 3. Train 과 Test 데이터에서 결측치가 존재하는지 확인한 후, 이를 다음과 같이 시각화해주세요.

(HINT) Python의 missingno 패키지를 사용하여 결측치를 시각화 할 수 있습니다.



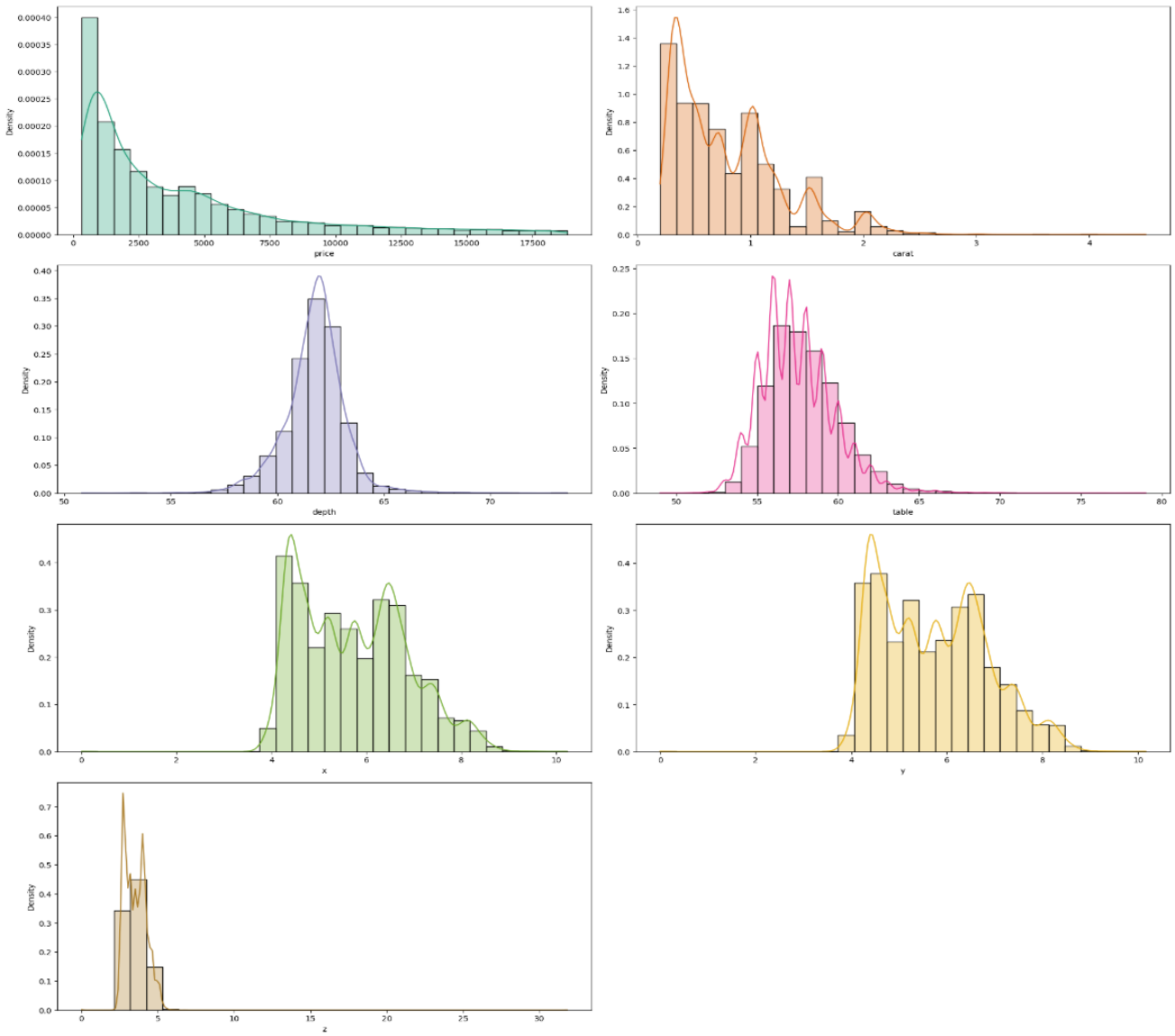
문제4. 문제 1부터 문제 3에서 얻어낸 정보들을 바탕으로, 데이터에서 각 변수들이 범주형 변수인지 수치형 변수인지 판단해보고, 그 이유에 대해서 간략하게 서술해주세요.

문제5. 데이터에서 수치형 변수들만을 사용하여, 수치형 변수들 간의 상관관계를 다음과 같은 상관관계 Plot을 통해 확인해주세요. 그리고 그 결과에 대해서 간단히 해석해주세요.



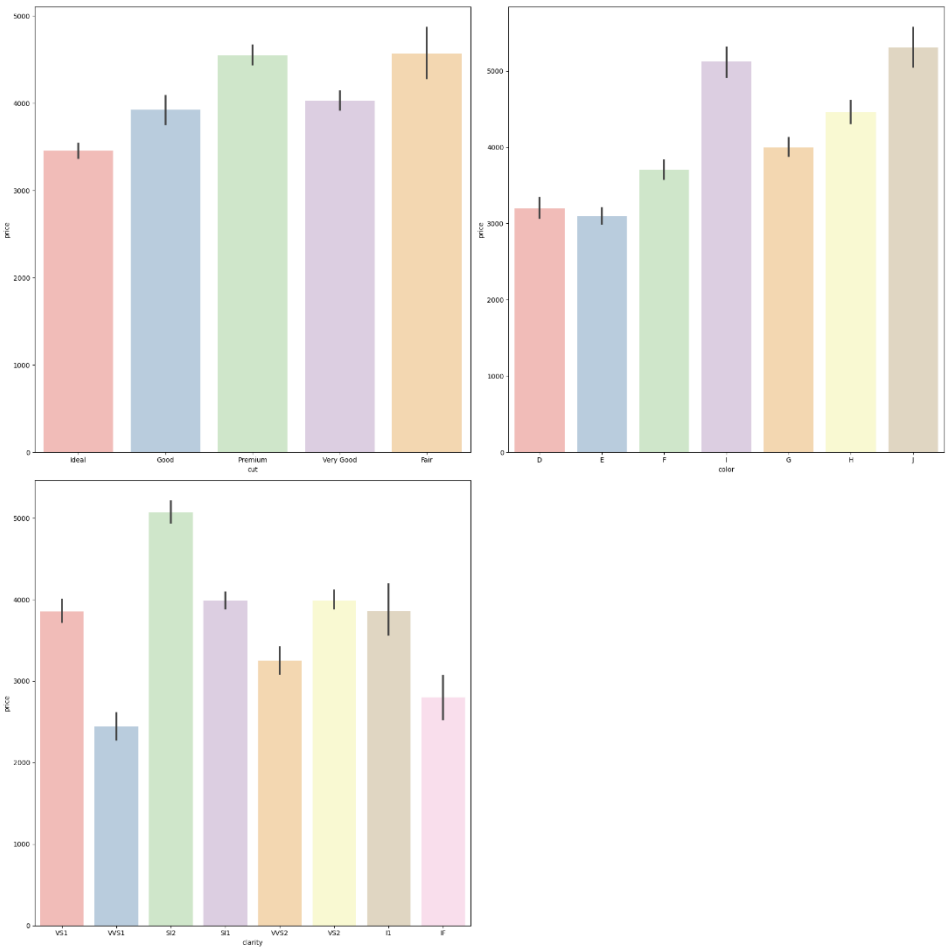
(+ 보너스 문제 1) 범주형 자료에 대해서도, 범주형 자료간의 상관관계를 상관관계 Plot을 통해 확인해주세요.

문제6. 수치형 변수에 대해서 다음과 같이 시각화를 진행해주세요. 이후, 시각화한 Plot을 통해 확인할 수 있는 점을 간단히 서술해주세요.



- Palette는 “Dark2”를 사용합니다.
- Target 변수인 ‘price’에 대한 해석은 꼭 들어가야합니다.

문제7. 범주형 변수에 대해서 다음과 같이 시각화를 진행해주세요. 이후, 시각화한 Plot을 통해 확인할 수 있는 점을 간단히 서술해주세요.



- Palette는 "Pastel1"을 사용합니다.

(+ 보너스 문제 2) 지난주 패키지에서 풀어보았던 것처럼, 범주에 따른 다양한 분포의 시각화를 통해 쉽게 다양한 정보를 데이터에서부터 얻어낼 수 있었습니다. 추가적인 전처리 및 검정을 자유롭게 진행하고, 이 과정에서 얻어낸 인사이트에 대해서 서술해보세요.

문제8. 이후 데이터를 전처리하는 과정에서, Test Data에 전처리를 진행할 때 Train Set과 동일한 전처리를 진행해줘야 합니다. 이 이유와 더 나아가서 'Data Leakage'가 무엇인지 알아보고 간단히 서술하세요.

문제9. 문제 5부터 문제 7에서 얻어낸 정보들을 바탕으로, 앞서 Train과 Test 데이터에 나타나는 결측치를 적절하게 채워주세요. 이후, 결측치가 잘 대체되었는지 다시 한번 확인해보세요.

문제10. 범주형 변수들에 대해서 LabelEncoding을 진행하고, 이를 Train과 Test 모두에 적용해주세요.

(HINT) sklearn.preprocessing의 LabelEncoder() 함수를 사용하면 범주형 자료들을 인코딩할 수 있습니다.

(+ 보너스 문제 3) 결측치 보간을 위해 굉장히 많은 방법들이 고안되었습니다. 결측치 보간법에 대해서 알아본 후, 결측치 보간법의 종류와 그 방법에 대해서 간단히 서술해보세요.

(+ 보너스 문제 4) Chapter 1의 문제들을 풀어보면서 얻어낸 인사이트를 바탕으로, 데이터의 특징을 고려하여 가격 예측에 활용할 수 있는 파생변수를 생성해보세요.

Chapter 2 : XGBoost & LightGBM

가장 많이 사용되는 트리 기반 모델이라고 할 수 있는 XGBoost와 LightGBM을 사용해보도록 하겠습니다. XGboost와 LightGBM은 의사결정나무를 기반으로 한 형태의 모델이지만, 의사결정나무의 단점인 과적합 문제를 해결할 수 있는 앙상블 기법을 통해 문제를 해결합니다. Boosting의 경우 매우 약한 의사결정나무를 여러 개 사용하여 계속하여 오차를 줄여가는 방식으로 학습하여 값을 출력합니다.

오늘 다루게 모델 안에는 다양한 하이퍼파라미터들이 포함이 되어있는데, 해당 파라미터들을 어떻게 설정하느냐에 따라서 모델의 성능이 달라질 수 있습니다. 최대한 좋은 성능을 갖추기 위해서 적절한 파라미터 값들을 찾아주는 방법으로 교차검증이 사용됩니다. 교차검증의 대표적인 방법인 K-Fold CV를 사용해보겠습니다.

문제1. 항상 동일한 결과를 얻기 위하여 Seed를 고정해주세요 (Seed: 3031)

문제2. Train 데이터를 5개의 Fold로 분리하세요.

(HINT) sklearn.model_seleciton의 KFold() 함수를 사용하면 K개의 데이터의 Fold의 인덱스번호를 반환받습니다.

문제2. XGboost와 LightGBM이 어떤 모델인지 살펴보고, 트리를 분기함에 있어 어떠한 특징을 가지는지 간단하게 설명해주세요.

문제3. XGBoost와 LightGBM 패키지를 불러와주세요.

문제4. XGBoost와 LightGBM모델에서는 굉장히 하이퍼파라미터가 사용되고 있습니다. XGBoost와 LightGBM에 사용되는 하이퍼파라미터에는 무엇이 있으며, 각 하이퍼파라미터가 의미하는 바는 무엇인지 서술해주세요.

문제5. GridSearch를 통해 하이퍼파라미터 튜닝을 진행하도록 하겠습니다.

- params ={'n_estimators':[300, 400, 500], 'max_depth':[2,4,6],
  'learning_rate' : [0.05,0.1,0.2]} 으로 설정해주세요.
- 평가지표로는 RMSE(Root Mean Squared Error)를 사용합니다.
- XGBoost, LightGBM 모두 같은 params로 param_grid를 설정합니다.

(HINT) sklearn.model_seleciton의 GridSearchCV() 함수를 사용하면 해당 하이퍼파라미터 조합을 모두 사용하여 GridSearch를 진행할 수 있습니다.

(HINT) sklearn.metrics의 mean_squared_error() 함수를 사용하면 평가지표로 사용되는 RMSE를 쉽게 계산할 수 있습니다.

문제7. GridSearch 결과 RMSE가 낮은 파라미터 조합을 각 모델별로 찾고, 해당 결과로 모델을 다시 학습시키세요.

문제8. 각 모델 별 변수의 중요도를 Importance plot으로 시각화하고, 어떤 변수가 보석의 크기에 영향을 미치는지 서술해주세요.

(HINT) 모델을 Train Data에 적합시킨 후 '._feature_importances_'를 통하여 XGBoost와 LightGBM에서의 Feature Importance를 확인할 수 있습니다.

문제9. 학습시킨 모델으로 Test Data를 예측하고, 이를 'Sample_submission.csv'의 price 열에 예측 값들을 채워주세요.
그 후, 'Answer.csv'에 있는 Test Data의 실제 price 값과의 Test RMSE를 구한 후, CV를 진행했을 때 얻었던 Train RMSE와 함께 각 모델별로 비교해보세요.

(+ 보너스 문제 5) 앙상블 기법은 말 그대로 여러가지 모델을 합쳐서 사용한다는 것이기 때문에, 이렇게 트리 기반 모델이 아니어도 사용이 가능합니다. 앙상블 기법에 대해서 찾아본 후, 앙상블 기법이 무엇인지, 그리고 앙상블 기법의 종류에는 어떤 것들이 있는지 알아보세요.

(+ 보너스 문제 6) 앞서 얻어낸 XGBoost의 예측 값과 LightGBM의 예측값의 평균을 구한 후, 이 값과 Test Data의 실제 price 값과의 RMSE 값을 계산해보세요. 그리고 교차검증을 진행하면서, XGboost의 예측값과 LightGBM의 예측 값들의 반영 비율을 조정해보면서 RMSE를 최소화시켜주세요.