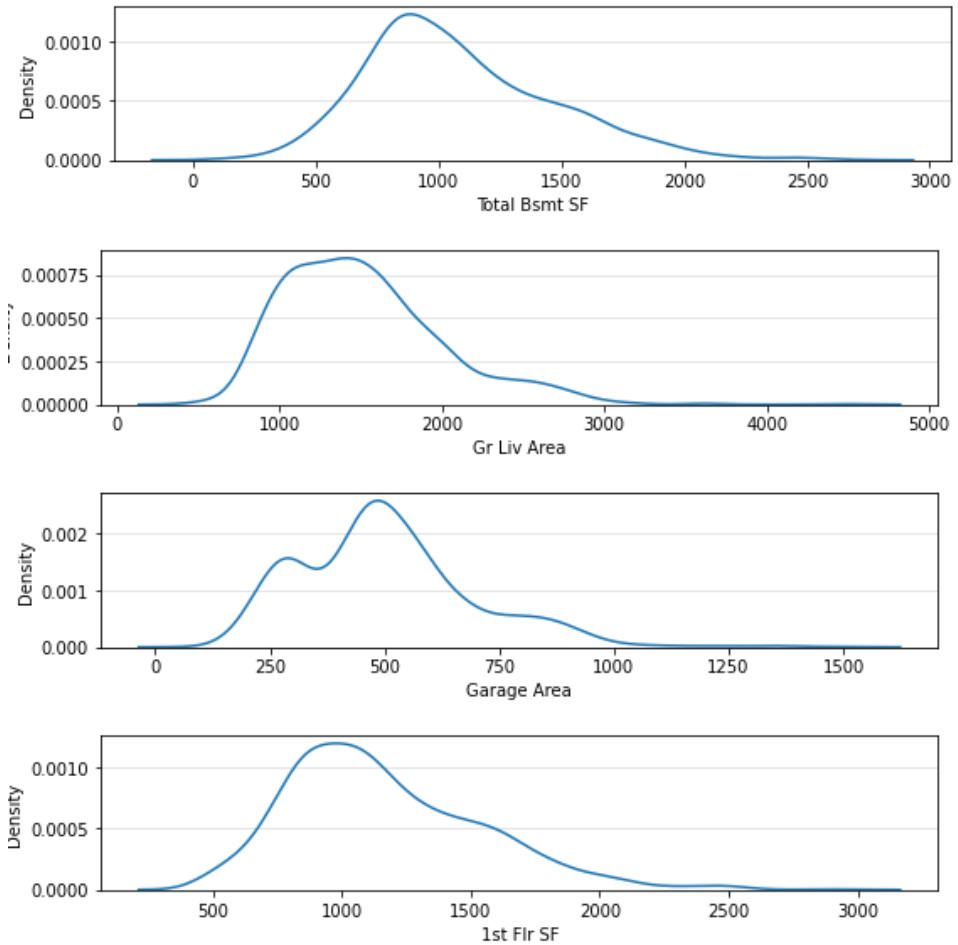
My First Dacon Competition

2022-01-28

시도해 본 것과 팁, 후기 (1)

- 데이터 클리닝
  - 애초에 클리닝 할 게 없는 데이터 (NA도 Outlier도 없는 데이터)
  - 범주형 변수의 인코딩(One-Hot? LabelEncoding?)
    - Quality 범주는 애초에 ordinal 한 범주형이므로 LabelEncoding을 해도 무방하다고 판단
- EDA
  - (연속형)변수들의 분포를 sns.kdeplot 로 확인하기 (week01 혜원님 EDA 참고)
  - 여러 matplotlib plot을 한번에 그릴 때 plt.tight_layout() 옵션을 쓰면 플롯 간에 적절하게 간격을 띄워 줌

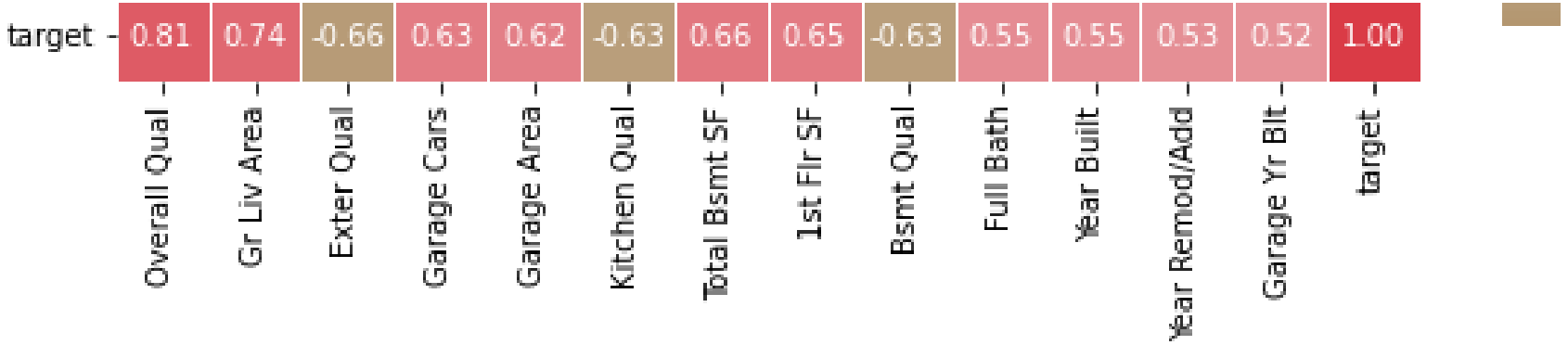
시도해 본 것과 팁, 후기 (2)



- 면적과 관련된 지표들이 일관된 분포를 보이고 있음
- 그렇다고 로그 스케일로 변환했더니 너무 오른쪽으로 치우쳐버림
- 지하실이 크면 1층도 크고, 차고도 큰 경향이 있지 않을까?

시도해 본 것과 팁, 후기 (3)

- Feature Engineering
  - Target 포함 13개의 feature
  - 필사와 실제 feature engineering이 다르다고 느낀 부분(아이디어)
  - Overall Qual은 가장 높은 상관관계를 보이고 있고, Area 관련 변수들은 서로 묶였을 때 영향력이 강할 것으로 예상함



시도해 본 것과 팁, 후기 (4)

- 추가로 만들 수 있는 feature가 있을까?
  - Location? Time? 없음
  - Polynomial features (week07에서 처음 알게된 방법)
  - 모든 polynomial feature를 쓰면 overfitting이 될 것이므로, FE 전 가장 상관관계가 높은 feature의 상관계수인 0.81보다 높은 feature만 사용하기로 함

시도해 본 것과 팁, 후기 (5)

- Baseline으로 XGBRegressor를 사용
- Hyperparameter Tuning을 위해 Optuna를 사용(week10에서 소개한 방법)

```
xgb_model = XGBRegressor(**trial_params)
xgb_model.fit(X, y)
xgb_model_pred = xgb_model.predict(test_all)
```

최적의 하이퍼파라미터를 XGBRegressor에 사용

Train data의 X_test, y_test 사용

Test data에서 predict

시도해 본 것과 팁, 후기 (6)

- sample_submission.csv 파일을 읽은 다음
- ‘target’ 컬럼에 예측값(xgb_model_pred)을 채운 다음
- Submission.csv 으로 내보내서 제출

시도해 본 것과 팁, 후기 (6)

- 1회 제출, 점수 0.10964, Public 100위
- 0.9xxx 대로 떨어지기 위해서는 ensemble이 반필수적?

시도해 볼 것

- Ensemble (RF + XGBRegressor + LGBMRegressor)
- Target 변수도 왼쪽으로 skew되어 있음 → log scale로 변환한 다음에 일련의 과정을 다시 진행해 보기