2024-11-04 수상작 리뷰

[기업은행 혁신리그] 아파트 경매가격 예측 경진대회

https://dacon.io/competitions/official/236165/codeshare/9400

데이터

- ID : 각 경매물건별 고유 구분자
- 입찰구분 : 입찰의 형태
- 감정연월 : 감정을 수행한 날짜
- 감정가(만원) : 감정했을 당시의 가격
- 최저매각가격(만원) : 최소입찰가격
- 경매연월 : 경매가 수행된 날짜
- 경매구분 : 임의/ 강제
- 경매결과 : 경매의 결과
- 총경매횟수 : 해당 물건이 경매에 나온 횟수
- 총토지경매면적 : 경매에 부쳐진 토지의 전체 면적
- 총건물경매면적 : 경매에 부쳐진 건물의 전체 면적
- 지역 : 서울/ 부산
- 구 : xx구의 형태로 구성
- 동 : xx동의 형태로 구성
- 건물층수 : 건물의 전체 층수
- 현재층수 : 경매물건의 층수
- 가압류횟수 : 경매물건의 가압류 횟수
- 소유이전횟수 : 경매물건의 소유권 이전 횟수
- **낙찰가(만원) : 종속변수 / test 데이터에 대한 '낙찰가(만원)' 을 예측하여야 함**

코드 흐름

1) **EDA**

   Data 파악

   랜덤 시드 고정

   Train, test shape, head, info 확인

2) **Feature Engineering & Processing**

   ID drop

   날짜 형식 전처리(train, test에서 감정연월, 경매연월 날짜 형식으로 변환 / 경매연월에서 감정연월을 뺀 일수 계산 / 감정연월, 경매연월 drop)

   이상치 처리 (중앙값 대체 – IQR 기준으로 outlier 정의해서 제거 / 해당 피처의 중앙값을 계산해서 outlier를 중앙값으로 대체한다.)

   범주형 변수 수치형 처리(LabelEncoder 사용)

   다중공선성 제거(경매결과, 감정가, 총건물경매면적 drop)

정규화(왜도와 첨도를 계산한 다음, 너무 높거나 낮은 것들을 변환 대상으로 선정한다. Box-Cox 변환을 하기 위해 값들이 양의 값인지 확인하고, Box-Cox 변환을 적용한다.)
독립변수, 종속변수 지정
스케일링(MinMaxScaler)

3) **모델 학습(Linear Regression)**
Sklearn의 LinearRegression으로 학습시킨다.

4) **후처리**
예측 결과를 원래 스케일로 역변환
예측값 할당

차별점, 배울점

범주형 변수 수치형 처리로 LabelEncoder 사용할 때에, 테스트 데이터의 레이블 변환 과정에서 data leakage를 방지한 것을 관찰할 수 있었다. 테스트 데이터에서는 훈련 데이터의 Label Encoder만을 사용해 변환한다. Fit을 적용하는 것이 아니라, transform만을 수행했다는 것이다. 또한, 훈련 데이터에 없는 새로운 값이 테스트 데이터에서 등장했을 때에도 해당 값을 추가하여 숫자로 변환해서 처리한다.

또한, Box-Cox 변환이라는 것을 사용했다. Box-Cox 변환은 데이터를 정규분포에 가깝게 변환하기 위해 사용되는 것으로, 데이터의 분포가 왜곡되거나 비대칭적이면 사용한다.

이 수상작에서는 모델 학습에서 하이퍼파라미터 튜닝을 전혀 하지 않고, 전처리를 중요시 여겼다.