Title: Analysis of Airbnb Pricing Factors in 10 cities of Europe

Authors:
이정환 20243070
임강현 20240335

1. Summary of questions and results

1. 다양한 위치 특성이 Airbnb 가격에 미치는 영향은 무엇인가?
- Random Forest 모델
  - Private room: 관광명소, 지하철, 시내 중심, 식당 순의 영향
  - Entire Home/Apt: 관광명소, 특정 도시, 지하철, 시내 중심 순으로 영향을 미침
- Gradient Boosting 모델
  - Private room: 관광명소, 지하철, 시내 중심, 식당 순으로 영향
  - Entire Home/Apt: 관광명소, 특정 도시, 지하철, 시내 중심 순으로 영향을 미침

2. 청결도 평가와 게스트 만족도가 가격에 미치는 영향은 무엇인가?
   - 위 2가지 요소는 다른 요소들에 비해 가격에 미치는 영향이 미미하였음

3. 연구 질문3: 주어진 위도와 경도에서 적절한 Airbnb 가격은 얼마인가?
   - 선형 회귀, 랜덤 포레스트, 그라디언트 부스팅 모델을 사용하여 예측한 결과, 15% 오차 범위 내에서 약 64%, 10% 오차 범위 내에서 약 50%의 정확도를 기록

2. Motivation

유럽 10개 도시의 Airbnb 가격 결정 요인을 이해하는 것은 호스트와 게스트 모두에게 중요하다. 이러한 요인을 파악함으로써 호스트는 경쟁력 있는 가격을 설정할 수 있고, 게스트는 가격의 적절성을 평가할 수 있다. 또한, 신규 호스트에게 최적의 가격 전략을 제공할 수 있다.

3. Data set

우리의 데이터셋은 다양한 도시에서 수집된 Airbnb 리스트를 포함하며, 각 파일은 주중과 주말 데이터를 포함하고 있습니다. 각 행은 특정 숙소의 다양한 속성(예: 방 유형, 가격(2박 3일), 위도, 경도 등)을 나타낸다. 사용된 데이터 파일들은 다음과 같다.
data set : https://zenodo.org/records/4446043

4. Methodology

데이터 전처리

1. CSV 파일 결합 및 전처리
   - 'data' 디렉토리에서 데이터를 로드
   - combine_csv_files function을 이용하여 도시별로 나누어진CSV 파일을 단일 데이터프레임으로 결합, 열 이름을 명확하게 변경, 결측값 처리
   - 새로운 열 city와 day(weekdays와 weekends를 합침)를 추가하는 전처리 작업을 수행
   - preprocess_data 함수를 사용하여 가격 데이터의 상위 25%제거(이상치)하고 city열을 one-hot encoding

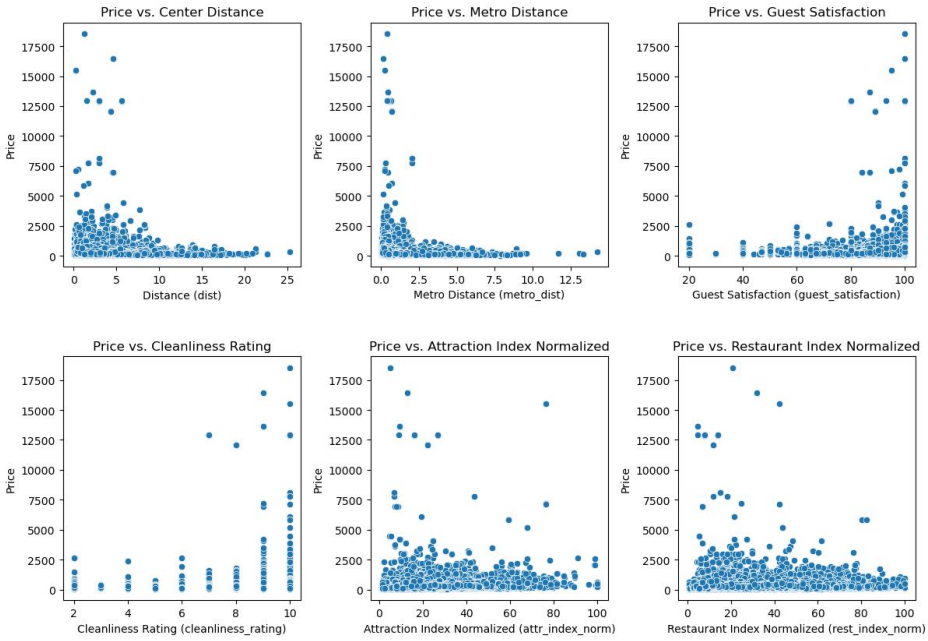
분석

1. Visualize Features-Lable
   - Matplotlib을 이용하여 다양한 특성과 가격 간의 관계를 산점도로 시각화
2. 선형 회귀 모델
   - 특정 방 유형(Private room, Entire home/apt)에 대해 데이터를 필터링하고, 독립 변수(특성)와 종속 변수(가격)를 정의했습니다. 모델을 학습하고 평가하여 특성들이 가격에 미치는 영향을 파악
3. 랜덤 포레스트 모델
   - 랜덤 포레스트 모델을 이용하여 각각의 특성들이 가격에 미치는 중요도를 분석하였다. 또한 같은 모델을 이용하여 위도, 경도를 독립변수, attr_index, rest_index, metro, center를 종속변수로 설정하고 Airbnb위치에 따른 4가지 종속변수를 예측하기 위한 모델을 학습

4. 그라디언트 부스팅 모델
    - 그라디언트 부스팅을 사용하여 선형 회귀 모델에서와 같은 학습을 시키고 다른 모델과 비교, 분석
5. 모델 적합도 판단
    - Mean Absolute Error, R-squared, Custom Accuracy를 사용하여 모델 적합도를 판단
6. 예측
    - 사용자가 기존 Airbnb의 위치, 가격 또는 경도, 위도를 편리하게 찾을 수 있도록 folium 라이브러리를 이용하여 데이터를 지도에 표현
    - 학습된 모델을 사용하여 사용자가 지도에서 원하는 위치를 찾은 후 위도, 경도, 침실 수, 도시, 수용 인원을 입력하면 Airbnb 가격을 예측

5. Result

Research Question 1: 다양한 특성이 Airbnb 가격에 미치는 영향

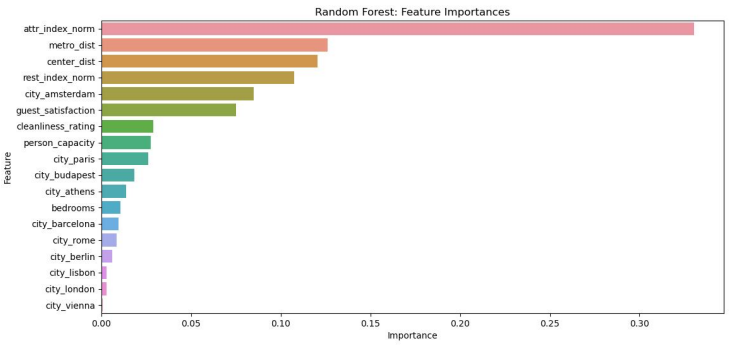


이 질문에 대해, 다양한 특성이 Airbnb 가격에 미치는 영향을 분석하기 위해 여러 특성과 가격 간의 관계를 시각화하였다. 이를 위해 도시 중심지와의 거리(center_dist), 지하철 거리(metro_dist), 게스트 만족도(guest_satisfaction), 청결도 평가(cleanliness_rating), 관광명소 지수(attr_index_norm), 식당 지수(rest_index_norm)와 가격(price) 간의 산점도를 작성하였다.

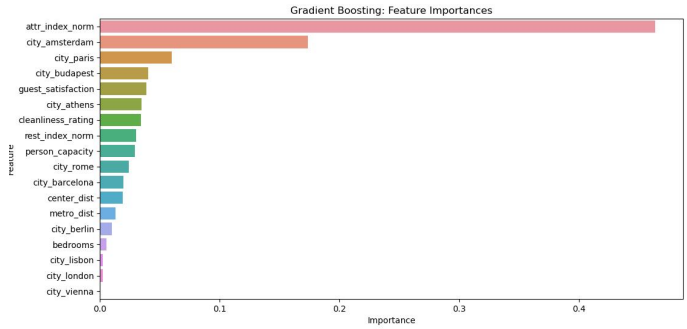
Graph Interpretation

1. Price vs. Center Distance (가격 vs. 도시 중심지와의 거리)
    - 그래프에서 볼 수 있듯이, 도시 중심지와의 거리와 가격 사이에는 반비례 관계가 있는 것으로 보인다. 즉, 도심에 가까울수록 가격이 높아지는 경향을 보인다. 이는 도심 지역이 일반적으로 더 높은 가치를 갖기 때문일 수 있다.
2. Price vs. Metro Distance (가격 vs. 지하철 거리)
    - 지하철과의 거리 역시 가격에 영향을 미친다. 지하철역에 가까울수록 가격이 높아지는 경향이 나타났는데, 이는 대중교통 접근성이 좋을수록 더 편리하기 때문일 수 있다.
3. Price vs. Guest Satisfaction (가격 vs. 게스트 만족도)
    - 게스트 만족도와 가격 간에는 비례관계가 있는 것으로 보인다. 높은 게스트 만족도는 높은 가격과 연관이 있다. 이는 높은 만족도를 제공하는 숙소가 더 높은 가격을 받을 수 있음을 시사한다.
4. Price vs. Cleanliness Rating (가격 vs. 청결도 평가)
    - 청결도 평가 역시 가격에 영향을 미치는 요인 중 하나이다. 청결도 평가가 높을수록 가격도 높아지는 경향이 나타났는데, 이는 게스트들이 청결한 숙소에 더 높은 가치를 부여함을 나타낸다.
5. Price vs. Attraction Index Normalized (가격 vs. 관광명소 지수)
    - 관광명소 지수와 가격 간의 관계도 양의 상관 관계가 있다. 관광명소에 가까운 숙소일수록 더 높은 가격을 받는 경향이 있다. 이는 관광명소 근처의 숙소가 더 많은 수요를 끌어들이기 때문이다.
6. Price vs. Restaurant Index Normalized (가격 vs. 식당 지수)
    - 식당 지수와 가격 간의 관계도 역시 비례관계를 보입니다. 식당이 많은 지역일수록 가격이 높아지는 경향이 있는데, 이는 식당과 같은 편의시설이 가까운 숙소가 더 높은 가격을 받을 수 있음을 시사한다.

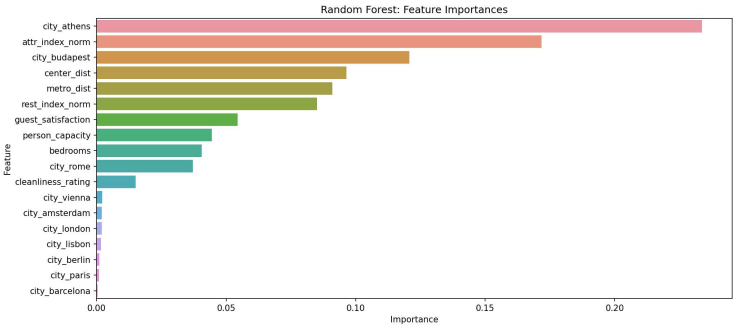
Research Question 2: 청결도 평가와 게스트 만족도가 가격에 미치는 영향은 무엇인가?



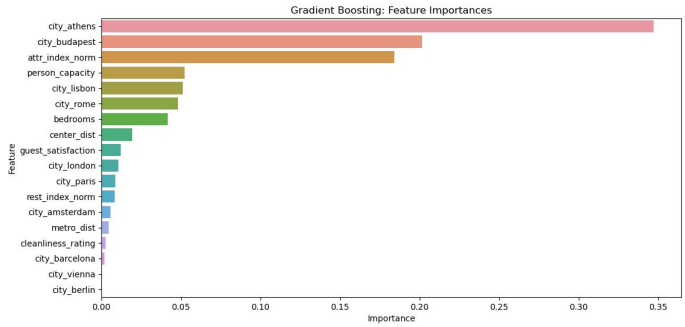
private room



private room



entire home/apt



entire home/apt

Random Forest 모델

Private Room

- 관광명소와의 근접성이 가장 큰 영향을 미쳤고, 그 다음으로 지하철 거리, 시내 중심 거리, 식당 순으로 영향을 미쳤다.

Entire Home/Apt

- 관광명소와의 근접성이 가장 큰 영향을 미쳤으나, 특정 도시(예: city_Athens, city_Budapest)의 영향력이 더 크게 나타났다.

Gradient Boosting 모델

Private Room

- 관광명소와의 근접성이 가장 큰 영향을 미쳤고, 그 다음으로 지하철 거리, 시내 중심 거리, 식당 순으로 영향을 미쳤다.

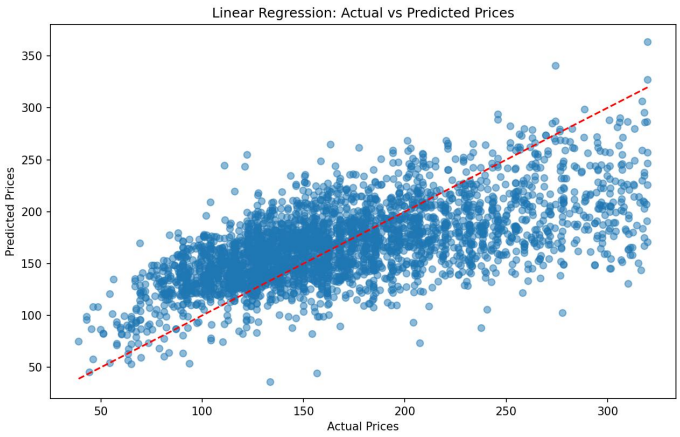
Entire Home/Apt (전체 집/아파트):

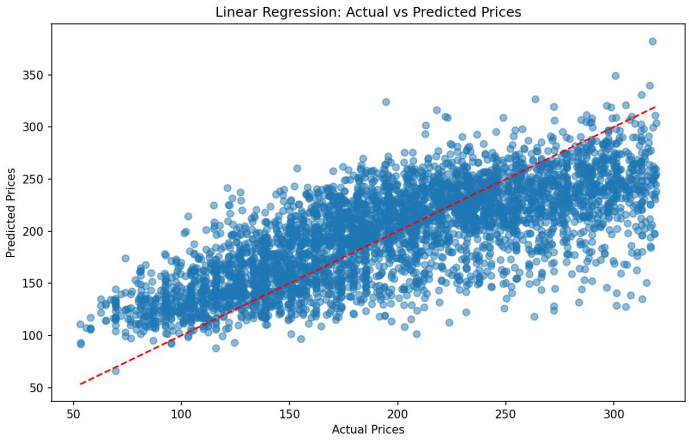
- Gradient Boosting 모델에서도 관광명소와의 근접성이 가장 큰 영향을 미쳤으며, 특정 도시(예: city_Athens, city_Budapest)의 영향력이 더 크게 나타났다.
- 관광명소 지수 (attr_index_norm), 특정 도시 (city_Athens, city_Budapest), 지하철 거리 (metro_dist), 시내 중심 거리 (center_dist)순서로 높은 영향력이 나타남.

랜덤 포레스트와 그라디언트 부스팅 모델을 사용한 결과, 청결도 평가와 게스트 만족도는 예상보다 Airbnb 가격에 미치는 영향이 미미했다. 대신 관광명소와의 근접성이 가장 큰 영향을 미쳤고, 특히 Entire home/apt의 경우 특정 도시의 영향력이 더 컸다. 이는 가족이나 단체 여행객들이 주로 이용하는 room type이기 때문에 이들이 개인 여행객보다 city특성을 중요시 한다고 볼 수 있다. 이를 통해 Airbnb 호스트가 가격 책정 시 고려해야 할 주요 위치 특성과 그 중요도를 파악할 수 있다.

Research Question3: 주어진 위도와 경도에서 적절한 Airbnb 가격은 얼마인가?

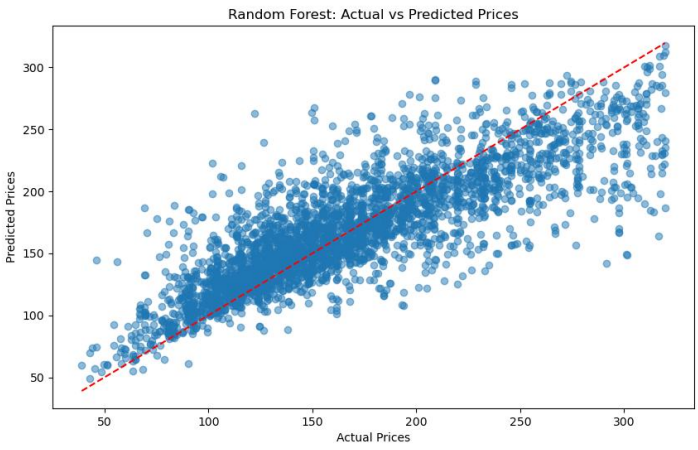
선형 회귀: 실제 가격 vs 예측 가격

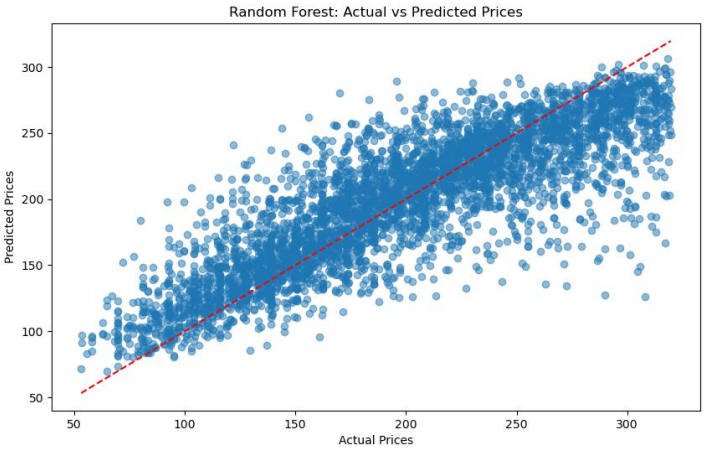




- 두 그래프는 각각 Private room과 Entire home/apt에 대해 실제 가격과 예측 가격을 비교한 것이다. 빨간 점선은 완벽한 예측을 나타내며, 데이터 점들이 이 선에 가까울수록 모델의 예측이 정확함을 의미한다.
- 선형 회귀 모델은 중간 가격대에서는 비교적 정확한 예측을 보였지만, 고가와 저가의 가격대에서는 예측이 다소 분산되는 경향이 있었다.
- 평균 절대 오차 (MAE): 35.56
- R^2 스코어: 0.396
- Custom accuracy(within 15%): 40.99%

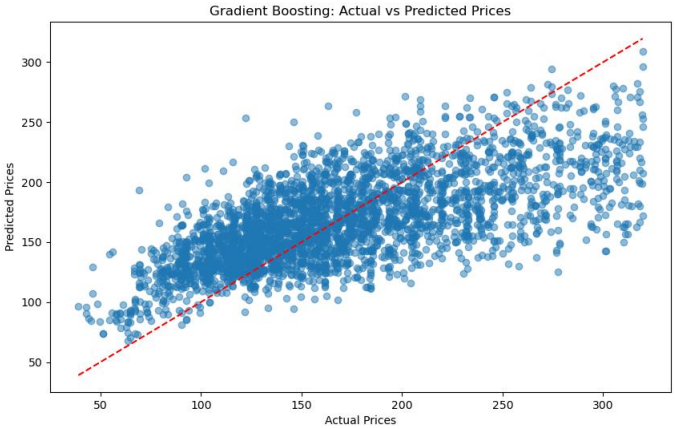
랜덤 포레스트: 실제 가격 vs 예측 가격

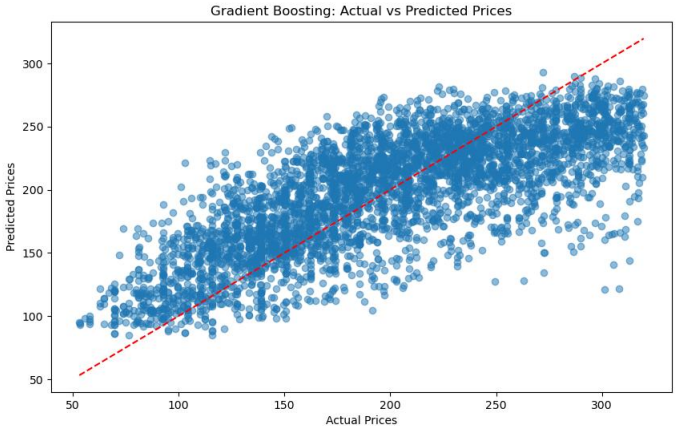




- 두 그래프는 각각 Private room과 Entire home/apt에 대해 실제 가격과 예측 가격을 비교한 것이다. 랜덤 포레스트 모델은 다양한 가격대에서 예측의 정확도를 높이기 위해 사용되었다.
- 개인실의 경우, 랜덤 포레스트 모델은 중간 가격대에서 높은 예측 정확도를 보였으며, 고가와 저가에서도 비교적 일관된 예측 성능을 보여주었다.
- Entire home/apt의 경우, 랜덤 포레스트 모델은 전반적으로 높은 예측 정확도를 보였으며, 특히 중간 가격대에서 매우 높은 성능을 나타냈다.
- 평균 절대 오차 (MAE): 23.46
- R^2 스코어: 0.688
- Custom accuracy(within 15%): 64.52%

그라디언트 부스팅: 실제 가격 vs 예측 가격

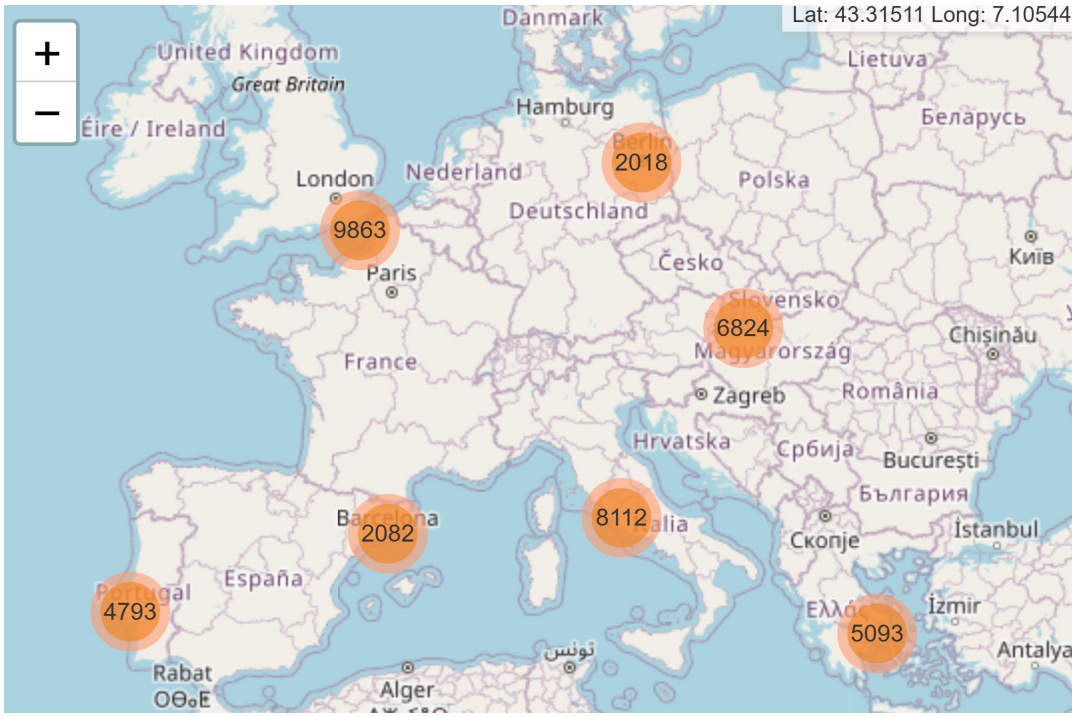




두 그래프는 각각 Private room과 Entire home/apt에 대해 실제 가격과 예측 가격을 비교한 것이다. 그라디언트 부스팅 모델은 랜덤 포레스트와 유사하게 중간 가격대에서 높은 예측 정확도를 보였으며, 고가와 저가에서도 비교적 일관된 예측 성능을 보여주었다.

- Private room의 경우, 그라디언트 부스팅 모델은 랜덤 포레스트와 비슷한 예측 성능을 보였으나, 고가와 저가에서의 예측 분산이 다소 줄어든 모습을 보였다.
- Entire home/apt의 경우, 그라디언트 부스팅 모델은 모든 가격대에서 비교적 일관된 예측 성능을 나타내었으며, 특히 중간 가격대에서 매우 높은 예측 정확도를 보였다.
- 평균 절대 오차 (MAE): 32.32
- R^2 스코어: 0.574
- Custom accuracy(within 15%): 52.14%

세 가지 모델 모두에서 예측 가격이 실제 가격과 대체로 잘 일치하는 것을 확인할 수 있다. 그러나 고가와 저가에서의 예측 정확도는 다소 떨어지는 경향이 있어, 향후 모델 개선이 필요할 수 있다. 선형 회귀 모델은 중간 가격대에서 좋은 성능을 보였으나, 극단적인 가격대에서는 예측이 부정확했다. 랜덤 포레스트 모델과 그라디언트 부스팅 모델은 전체적으로 더 나은 예측 성능을 보였으며, 특히 랜덤 포레스트 모델이 가장 높은 정확도를 보여주었다. 그라디언트 부스팅 모델은 모든 가격대에서 일관된 성능을 보여주었지만, 예측 오차가 랜덤 포레스트 모델보다 다소 큰 편이었다. 이러한 결과를 바탕으로, 다양한 가격대에서 더 정확한 예측을 위해 랜덤 포레스트 모델을 사용하는 것이 바람직할 것으로 보인다.



1. 유럽 전체 Airbnb 분포 지도

- 첫 번째 지도는 유럽 전역의 주요 도시에서 제공되는 Airbnb의 분포를 나타내고 있다. 각 도시에 표시된 원은 해당 도시에서 제공되는 Airbnb의 수를 나타내며, 이를 통해 특정 도시의 Airbnb 밀집도를 파악할 수 있다.
- 예를 들어, 런던은 9863개의 Airbnb 숙소를 보유하고 있으며, 이는 유럽 내에서 가장 많은 숙소 수를 나타낸다. 이 지도는 사용자가 유럽 내 여러 도시에서 Airbnb 숙소의 분포와 밀집도를 비교하는 데 유용하다.

2. 지도 시각화: 로마의 Airbnb 위치 및 가격 예시

- 두 번째 지도는 로마의 Airbnb 위치와 가격을 나타내고 있다. 지도 상의 마커를 클릭하면 해당 위치의 가격, 방 유형, 도시에 대한 정보가 표시된다. 예를 들어, 로마에 있는 특정 Airbnb는 2박3일에 $218의 가격을 가지고 있으며, 방 유형은 entire home/apt이다.
- 이러한 시각화는 특정 위치의 Airbnb 가격을 직관적으로 파악할 수 있게 해주며, 사용자가 원하는 지역의 가격 정보를 쉽게 얻을 수 있도록 도와준다.

6. Impact and Limitations:

Airbnb 호스트

- 이 분석을 통해 경쟁력 있는 가격을 설정하는 데 도움을 받을 수 있다. 비슷한 위치와 조건을 가진 다른 숙소와 비교하여 자신의 숙소에 적합한 가격을 책정할 수 있다.

게스트

- 분석 결과를 통해 가격의 적절성을 평가할 수 있다. 원하는 지역의 평균 가격과 비교하여 숙소의 가격이 합리적인지 여부를 판단할 수 있다.

부동산 투자자나 시장 분석가

- Airbnb 가격 변동 패턴을 이해하고, 이를 통해 투자 결정을 내릴 수 있다.

데이터 편향

- 수집된 데이터가 모든 요인을 완전히 반영하지 못할 수 있으며, 특정 도시나 지역에서의 데이터 부족은 그 지역의 가격 예측 정확도를 낮출 수 있다. 예를 들어, 특정 도시에서 Airbnb 데이터가 충분하지 않을 경우 해당 지역의 가격 예측이 부정확할 수 있다.

일반화의 어려움

- 우리의 모델은 특정 시점의 데이터를 기반으로 하므로, 시간이 지남에 따라 시장 변화에 적응하지 못할 수 있다. 또한, 다른 시기나 상황에서 동일한 결과를 보장하지 않는다.

특정 사용자 그룹의 제외

- 데이터 접근성이 제한된 사용자나, 분석에 반영되지 않은 특정 조건을 가진 숙소 소유자들은 이 분석의 혜택을 받지 못할 수 있다.

피해 가능성

- 이런 요소들이 야기하는 잘못된 가격 설정으로 인해 호스트는 손해를 볼 수 있으며, 게스트는 부당한 가격을 지불할 위험이 있다.

이와 같은 한계를 인식하고, 우리의 결론을 사용할 때 이러한 제한 사항을 고려해야 한다. 결론을 맹신하기보다는 참고 자료로 사용하고, 추가적인 연구와 분석을 통해 보완하는 것이 중요하다.

7. Challenge goals:

- Multiple Datasets: 수집한 데이터에서 weekdays, weekends가 분리되어 있었기 때문에 우리는 데이터를 결합하여 분석을 진행하였다. 이 데이터들을 결합함으로써 더 풍부한 분석을 수행할 수 있었다.
- Machine Learning: 다양한 머신 러닝 모델(선형 회귀, 랜덤 포레스트, 그라디언트 부스팅)을 사용하여 Airbnb 가격을 예측하였고. 각 모델의 성능을 비교하여 최적의 예측 모델을 선택하였으며, 모델의 성능을 평가하기 위해 평균 절대 오차(MAE), R^2 스코어, Custom accuracy(within 15%) 를 사용하였다.
- New Library: Folium 라이브러리를 사용하여 Airbnb 숙소의 위치와 가격 정보를 시각화하였다. 이 새로운 라이브러리를 활용하여 지도를 기반으로 한 시각화를 구현함으로써, 데이터의 공간적 분포와 가격 패턴을 직관적으로 파악할 수 있었다.

9. Work Plan Evaluation:

1. 데이터 전처리:

- 예상 시간: 3시간
- 실제 시간: 3시간
- 작업 내용: 여러 CSV 파일을 하나의 DataFrame으로 결합하고, 컬럼 이름을 변경하고, 불필요한 컬럼을 삭제하며, 결측값을 처리했다. 이 단계는 비교적 예상과 비슷한 시간이 걸렸다. 데이터 파일을 DataFrame으로 변환하고, 'city'와 'day' 컬럼을 추가하며, 컬럼 이름을 변경하고 결측값을 제거하는 과정이 포함되었다.

2. 머신러닝 및 코드 작성:

- 예상 시간: 10시간
- 실제 시간: 10시간
- 작업 내용: 머신러닝 모델(Linear Regression, Random Forest, Gradient Boosting)을 훈련시키고 평가하는 작업을 수행했다. 모델의 성능을 평가하기 위해 다양한 지표를 사용하고, 하이퍼파라미터 튜닝을 통해 최적의 모델을 찾았다.

3. 결과 분석:

- 예상 시간: 2시간
- 실제 시간: 5시간
- 작업 내용: 모델의 예측 결과를 분석하고, 정확도를 높이기 위해 추가적인 작업을 수행했다. 초기 예상보다 시간이 더 소요된 이유는 모델의 정확도를 높이기 위해 다양한 시도를 해야 했기 때문이다.

4. 프로젝트 발표자료 준비:

- 예상 시간: 3시간
- 실제 시간: 5시간
- 작업 내용: 프로젝트의 결과를 효과적으로 전달하기 위해 발표 자료를 준비했다. 중요한 정보를 명확하고 간결하게 전달하기 위해 추가적인 시간이 필요했다.

전반적으로, 예측한 작업 시간과 실제 소요된 시간은 대부분 일치했지만, 결과 분석과 발표자료 준비에서 예상보다 더 많은 시간이 소요되었다. 이는 모델의 성능을 최적화하고 결과를 명확하게 전달하기 위한 추가적인 작업이 필요했기 때문이다.

10. Testing:

테스트는 개발 과정 전반에 걸쳐 수행되어 버그를 조기에 발견하고 코드의 정확성을 보장했다. 우리는 Assert statements, Manual inspection을 결합하여 결과를 검증했다.

우선, 주요 함수와 모델 출력이 예상 범위 내에 있는지 확인하기 위해 Assert statements를 사용했다. 예를 들어, 데이터 전처리 단계에서는 결과 데이터 프레임에 예상되는 column이 포함되어 있는지 확인했다. 이는 데이터가 모델 학습에 사용되기 전에 올바르게 처리되고 있는지 보장하는 데 중요했다.

또한, 머신러닝 모델의 성능을 평가하기 위해 Assert statements을 사용했다. 선형 회귀, 랜덤 포레스트, 그라디언트 부스팅 각각의 모델에 대해 정확도 임계값을 설정하고, 각 모델의 정확도가 사전 정의된 임계값을 초과하면 Assertion 오류가 발생하도록 했다. 이 방법을 통해 모델이 허용 가능한 범위 내에서 성능을 발휘하고 있음을 보장했다.

Manual inspection은 테스트 과정에서 중요한 역할을 했다. 출력 데이터의 column과 예측 값을 수동으로 확인함으로써 결과가 일치하는 지 확인했다.

전반적으로 Assert statements, Manual inspection을 포함한 우리의 종합적인 테스트 접근 방식은 분석 결과와 결론의 정확성과 신뢰성을 확인하는 데 도움이 되었다.

11. Collaboration:

Chat-gpt를 활용하여 팀원끼리만 작업을 진행하였음.