안녕하세요. 아파트 검색알고리즘을 맡은 정지원입니다. 저희는 살기좋은 집을 추천하고자 빅데이터 분석을 진행했습니다. 혹시 2반분들은 어떤 집에 살고 싶으신가요? 저는 안전하고 조용하고, 가끔씩 카페에 가서 공부하며, 출근하기엔 조금 편한 집에 살고 싶습니다. 혹시 저랑 같으신 분이 계실까요? 이러한 점에 기반하여 역세권 뿐만 아니라 스세권, 편세권 등 다양한 부동산 용어들이 생겨나고 있는 것을 보았습니다. 이러한 집들은 수요가 많기에 보통 가격이 높게 형성될 것입니다. 물론 환율에 영향도 있겠지만요.

그래서 저희는 이러한 점에 기반하여 아파트의 1제곱미터당 집의 가격을 목적 변수로 하여 다양한 변수의 데이터 분석을 진행했습니다. 그 전에 앞서 서울 지역의 2023년 평균 아파트 거래가격을 보시죠. 강남구가 가장 높은걸 볼 수 있고, 금천구가 가장 낮은 것을 볼 수 있습니다. 금천구는 가산디지털단지라는 회사가 몰려있는 지역도 있는데 왜 집값이 쌀까요? 한 블로그에서는 다음과 같이 분석했습니다. 그리고 각 지역의 편의 시설도 보시죠. 먼저 병원을 보실까요? 역시 강남구, 서초구 등 집값이 높은 곳에서 병원이 많이 위치한 것을 확인할 수 있습니다. 또한, 스타벅스도 강남, 서초, 영등포 등에서 많이 위치한 것을 확인할 수 있었습니다. 다른 약국, 편의점 등 편의시설이 유사한 경향을 따르고 있는 것을 확인했습니다.

이런 자료에 기반하여 저희는 아파트 거래 정보를 확보하여 각 주변의 편의 시설들의 위, 경도 정보를 확보했습니다. 그리고 1제곱 미터당 분포를 scatter plot으로 그려봤지만 선형 관계로 어떤 변수로도 명확히 설명할 수 없다는 것을 확인할 수 있습니다. 그래서 먼저 선형 회귀 분석을 진행해서 각 변수의 영향력을 확인해보고, 해당 모델이 선형 모델을 만족하는지 살펴봤습니다. 그리고 라이브러리를 활용해서 설명 변수 선택에 도움을 받았고 설명계수가 0.347정도 나온 것을 확인했습니다. 이 말은 해당 식이 y값을 약 20% 설명할 수 있다고 보시면 됩니다. 이 때 X들이 Y의 선형 결합임을 타당하다고 보기 위해선 선형성, 독립성, 등분산성, 정규성, 다중공산성의 부재가 있습니다. 이를 검증하기 위해서 Q-Q plot을 먼저 보고, 검정 테스트를 진행했을 때 잔차의 오차가 정규성을 따르지 않는 것을 확인했습니다. 이를 통해 Y는 X들의 비선형 결합이라고 생각했고 다른 비선형 모델을 만들었습니다. 저희는 비선형 모델 중 하나인 random forest를 사용했습니다. 그리고 각 변수의 영향력을 계산하기 위해 LIME이란 라이브러리를 사용해서 영향력을 얻을 수 있었습니다. 다음은 그 결과입니다. 훈련 테스트와 테스트 데이터를 반으로 나눠서 훈련한 후에 테스트 셋에 적용해보았고 설명 계수는 약 0.9가 나왔습니다. 예상 외로 대형 종합 병원이 너무 멀면 안좋고, 정신 병원은 멀리 있는게 좋고, 스타벅스가 가까우면 좋게 나왔습니다. 이러한 알고리즘을 기반해서 다음과 같은 식을 만들었습니다.

------------------- 참고자료------------------------------

통계자료

**1. Omnibus Test**

* **귀무가설 ((H\_0))**: 잔차가 정규 분포를 따른다.
* **대립가설 ((H\_1))**: 잔차가 정규 분포를 따르지 않는다.

**2. Jarque-Bera Test (JB)**

* **귀무가설 ((H\_0))**: 잔차가 정규 분포를 따른다.
* **대립가설 ((H\_1))**: 잔차가 정규 분포를 따르지 않는다.

**3. Skewness**

* Skewness에 대한 검정은 보통 단독으로 p-value를 제공하지 않지만, 비대칭성을 판단하는 데 사용합니다.
* **귀무가설 ((H\_0))**: 잔차의 비대칭성이 0이다 (즉, 잔차가 정규 분포를 따른다).
* **대립가설 ((H\_1))**: 잔차의 비대칭성이 0이 아니다 (즉, 잔차가 비대칭성을 가진다).

**4. Kurtosis**

* Kurtosis 역시 단독으로 p-value를 제공하지 않지만, 뾰족함(극단점에 대한 분포의 첨도)을 판단하는 데 사용됩니다.
* **귀무가설 ((H\_0))**: kurtosis가 3이다 (정규 분포에 해당).
* **대립가설 ((H\_1))**: kurtosis가 3이 아니다 (즉, 잔차 분포의 첨도가 정규 분포와 다르다).

**5. Durbin-Watson Test**

* **귀무가설 ((H\_0))**: 잔차 간에 자기상관이 없다 (즉, 잔차는 독립적이다).
* **대립가설 ((H\_1))**: 잔차 간에 자기상관이 존재한다.

자료해석 설명자료

Prediction for instance 0: 1021.0850697752871

Feature: 종별코드명\_상급종합 Closest Distance > 4.12, Importance: -276.7290927928991

Feature: 1.87 < 종별코드명\_정신병원 Closest Distance <= 3.08, Importance: -194.93150498285402

Feature: 종별코드명\_보건지소 Closest Distance <= 2.28, Importance: -80.00008129840653

Feature: 종별코드명\_조산원 Count Within 0.5km <= 0.00, Importance: -70.68032591114084

Feature: 0.26 < starbucksSeoul Closest Distance <= 0.44, Importance: 60.31296450777824

Feature: 종별코드명\_요양병원 Closest Distance <= 0.56, Importance: -49.95929924542477

Feature: cctvSeoul Closest Distance > 0.09, Importance: 45.049715243934415

Feature: 13.00 < supermarketSeoul Count Within 0.5km <= 19.00, Importance: -36.40513131092846

Feature: 종별코드명\_상급종합 Count Within 0.5km <= 0.00, Importance: 34.270254745884074

Feature: 종별코드명\_보건소 Count Within 0.5km <= 0.00, Importance: -33.5676019171453

Feature: 1.15 < 종별코드명\_보건소 Closest Distance <= 1.78, Importance: -27.8372637835026

Feature: supermarketSeoul Closest Distance > 0.17, Importance: 27.358263290732264

Feature: 종별코드명\_요양병원 Count Within 0.5km > 0.00, Importance: 25.135683872043952

Feature: 종별코드명\_보건지소 Count Within 0.5km <= 0.00, Importance: 23.62813387005031

Feature: laundryCoinSeoul Closest Distance > 1.26, Importance: 23.119065759479522

Feature: 75.00 < cctvSeoul Count Within 0.5km <= 109.00, Importance: -22.871402895929226

Feature: 0.36 < policeSeoul Closest Distance <= 0.56, Importance: -19.053454201690926

Feature: 종별코드명\_의원 Closest Distance > 0.25, Importance: -18.69515942423123

Feature: pharmSeoul Closest Distance > 0.23, Importance: -18.463203274129445

Feature: 0.00 < starbucksSeoul Count Within 0.5km <= 1.00, Importance: -16.842208526181352

Feature: policeSeoul Count Within 0.5km <= 0.00, Importance: -14.020819100249723

Feature: 종별코드명\_정신병원 Count Within 0.5km <= 0.00, Importance: 12.064091878628073

Feature: convSeoul Closest Distance <= 0.03, Importance: -11.975009320514273

Feature: 종별코드명\_한방병원 Count Within 0.5km <= 0.00, Importance: -11.730214644154673

Feature: 17.00 < hospitalPosEncoded Count Within 0.5km <= 33.00, Importance: -10.613175277031027

Feature: 0.11 < 종별코드명\_한의원 Closest Distance <= 0.19, Importance: 9.471311069343445

Feature: 15.00 < 종별코드명\_의원 Count Within 0.5km <= 25.00, Importance: 9.232173303832678

Feature: 0.46 < 종별코드명\_병원 Closest Distance <= 0.73, Importance: -8.823711812573144

Feature: 종별코드명\_종합병원 Closest Distance <= 1.00, Importance: 8.794810779957608

Feature: 0.00 < subwaySeoul Count Within 0.5km <= 1.00, Importance: 8.785638657984963

Feature: busstopSeoul Closest Distance > 0.15, Importance: -8.167326297243608

Feature: 6.00 < 종별코드명\_한의원 Count Within 0.5km <= 10.00, Importance: -7.88559343190598

Feature: 종별코드명\_병원 Count Within 0.5km <= 0.00, Importance: -6.780440917861496

Feature: 17.00 < busstopSeoul Count Within 0.5km <= 23.00, Importance: 5.664457816485145

Feature: 0.46 < subwaySeoul Closest Distance <= 0.67, Importance: -5.328863295800756

Feature: 5.00 < 종별코드명\_치과의원 Count Within 0.5km <= 9.00, Importance: -5.245099842120022

Feature: laundryCoinSeoul Count Within 0.5km <= 0.00, Importance: 4.7984229439113895

Feature: 1.07 < 종별코드명\_한방병원 Closest Distance <= 1.62, Importance: -4.766878016900415

Feature: 종별코드명\_치과병원 Closest Distance <= 0.84, Importance: 4.292955708168748

Feature: pharmSeoul Count Within 0.5km <= 6.00, Importance: 3.3987904069819392

Feature: 종별코드명\_치과의원 Closest Distance <= 0.09, Importance: 2.548124749349484

Feature: 16.00 < convSeoul Count Within 0.5km <= 24.00, Importance: -2.2982457003401686

Feature: 종별코드명\_종합병원 Count Within 0.5km <= 0.00, Importance: -1.9617659341181575

Feature: 3.89 < 종별코드명\_조산원 Closest Distance <= 6.10, Importance: 1.3683973860205416

Feature: 종별코드명\_치과병원 Count Within 0.5km > 0.00, Importance: 0.04728217870140293

이것으로 발표 마치겠습니다. 감사합니다.