모든 변수를 사용하여 작성한 모델의 성능이 반드시 높다고는 할 수 없다 왜냐? 만약 특징이 없는 변수를 포함시켜 모델을 작성하면 계산량이 많아질 뿐아니라 성능이 떨어질 가능성이 있기 떄문에 특징이 있는 변수만 사용해서 성능이 좋은 모델을 작성해야함. 즉 특징이라하면 가격상승 및 하락에 영향을 주는 변수만을 선별 후 그 변수들 사이에서 재선별하는 후진제거법(back elimination) 또는 우리가 listup해놓은 변수들을 모델의 데이터셋에 추가해가면서 성능 측정을 반복하고 성능이 최고로 높아졌을 떄 반복을 중지하고 남은 변수를 모델 작성 변수로 선택하는 전진선택법(Forward Selection) 둘 중 어느 기법을 활용할 것인지를 선택할 필요가있음. 우리의 목표는 상승 및 하락 예측모델의 성능을 높이기 위해서는 parameter optimization(모델의 성능이 높아질 수 있도록 파라미터 조정 및 최적화)이 절대적으로 필요하다. 다양한 지도학습 모델을 비교(선형 회귀/결정트리/랜덤 포레스트/XGBoost,SVM 신경망)하여 최적의 모델선정할 필요가 있으며, 비지도학습을 보조적인 측면에서 활용해 볼 수 있는데, 클러스터링을 통해 유사한 지역의 부동산을 군집화하여 특성을 추가할 수 있다.[가령, 울산과 부산의 유사한 지리적 특성에 따른 비슷한 인프라(일자리/주거환경)가 형성되므로 군집화시킴] 부동산 가격을 예측하는 경우 레이블링된 데이터(금리/가구수/조세정책에 따른 세율)를 활용하여 다중회귀모델(기본적으로 feature(변수)간의 '독립성'을 전제로함)과 같은 지도학습을 수행하여 예측해볼 수 있는데 이때 입력 변수들 간에 상관관계가 높게 나타나면 다중공선성(Multicollinearity) 문제(즉, 독립변수가 서로 의존하게 되면 'over-fitting'과적합문제가 발생한다)로 모델의 예측 성능이 저하 될 수 있기 떄문에 주성분분석(feature extraction)을 통해 회귀분석의 전제조건(변수간 '독립성')을 달성하기 위해 비지도학습인 클러스터링 및 PCA분석을 지도학습 모델과 더불어 활용하면 예측력을 높이는데 상당한 도움이 될 것으로 기대하고 있습니다.

thick data(추상적 데이터: 사람들의 실제 경험과 행위에 담긴 맥락과 의미를 설명해 주는 데이터)를 우리는 (A~E)등급을 부여 후 순서대로 가중치(가격상승에 긍정영향) 1부터 (하락에 긍정영향) -1를 매기는 과정같이 data labeling를 통해 지도학습(다중회귀)의 변수로서 활용 될 수 있도록 모델을 설계할 필용성을 느껴 기본 수치화 가능한 빅데이터 말고 인간의 욕망과 여러 행동맥락적 성격이 강한 thick데이터에 대해서 우리는 위와같은 방향성을 갖고 프로젝트 를 진행할 예정입니다.

결국 파멸적인 집값상승으로 살 타이밍을 놓친 자가가 없는 무주택자들이 서비스 대상이며 개인의 최적화된(인프라 고려순위[1위 : 의료(종합병원 도보 5분거리)/2위 : 여가(호수공원 도보10분거리)/ 3위 : 안전(경찰서 도보 15분거리) 와 개인 별 budget range에 따른 가성비 집을 bigdata와 thickdata를 활용하여 상승 및 하락을 예측하여 최적의 매수 타이밍 알림 서비스로 가야겠다