#최종보고서

**3. 기대효과**

- 모형 평가

(1) Decision tree : 초기모형은 노드 8개로 만들어졌고 의미있는 변수가 ‘평균전세가’와 ‘교육시설수’ 였다면, validation과 train set의 MSE를 기준으로 적정 terminal node를 15개로 의사결정 모형을 만든 결과로는, root node인 ‘평균전세가’ 부터 ‘노후도’, ‘단지세대수’, ‘교육시설수’ 등등의 순으로 의미 있는 변수임을 알 수 있었습니다. ‘평균전세가’ 를 포함한 결과와 미포함한 결과를 분석하면, 전반적인 data set의 MSE의 변화를 통해 ‘평균전세가’가 인천아파트 ‘평균매매가’에 많은 영향을 준다는 것을 유추할 수 있습니다.

(2) Regression : 회귀모형의 최종모형 선택은 다중 공선성의 제거를 위해 X12 변수를 제거하고 stepwise을 통해 변수 선택된 모형이고, 그 결과 validation MSE도 train MSE에 비해 낮아짐을 알 수 있었습니다. 의사결정모형과는 다르게 ‘개발호재’ 변수의 coefficient가 약 18.85로 가장 높았고 ‘역세권점수’, ‘스타벅스’, ‘교육시설수’ 등의 순으로 유의한 결과가 나타났습니다.

- 기대 효과

대부분 수요가 안정적인 도시는 전세가에 따라 매매가가 연동되는 성향이 짙다고 하는데, 지역마다 다르지만 이는 보통 아파트의 매매가에서 전세가가 차지하는 비율인 전세가율이 75% 가 넘어가면 아파트 매매가도 보통 상승한다 라는 법칙에 따라 두 변수가 굉장히 상관관계가 높다고 보여집니다. 인천의 아파트의 경우 재건축 계획, 현황이 서울보다는 적은 편이라 고령화된 아파트일수록 집값도 대체로 떨어지는 추세를 가지고 있습니다. 뿐더러 새아파트들도 송도 같은 신도시에 많이 지어져 아파트 노후도가 매매가에 꽤 큰 영향을 주는 것을 알 수 있습니다.

역세권의 인기는 교통 환경이 열악한 지역일수록 높습니다. 특히 서울과 가까운 인천이나 경기도지역은 역세권이냐 아니냐 에 따라 집값이 크게 양분되는데, 화성시 청계동의 경우 동탄역을 기준으로 바로 인접해 있는 아파트와 도보로 사용하기 힘든 아파트의 가격이 3.3당 800만원 이나 차이가 나는 만큼 역세권이 아파트값에 큰 영향을 끼치는 것을 알 수 있습니다. 스타벅스의 경우 모든 매장은 본사에서 직접 운영, 관리하는 직영점으로만 운영되고 있어, 본사에서 건물 인근의 상권 등 조건을 따진 뒤 입점 가불가를 결정하기 때문에, 아파트 가격에 영향을 주기 전 영향을 받는 경우가 커 보입니다. Stepwise selection을 통해 제거된 변수 중 하나인 ‘개별세대수’ 변수는 지역에 상관없이 보통 균일한 편으로 분양면적당 50~150세대 정도 되기 때문에 아파트값에 큰 영향을 끼치지 않아 보이고 ‘문화시설수’ 변수에 포함된 공원의 개수 또한 해당 동에 많다 해서 매매가에 큰 상관이 없어 보입니다. 이러한 의미 있는 여러 변수들을 통해 인천의 아파트 매매가격을 2019년 이후로도 어느정도 잘 예측할 수 있을 것이라고 전망이 됩니다.

**4. 한계점**

(1) 예측 한계 : 세세한 아파트 단지별로 분석이나, 조사된 변수들의 자료 부족으로 완벽한 예측이 힘들 것입니다. 전문가도 예측하기 힘든 집값 조건들의 다양한 외적 변수와 예상치 못한 변수들까지는 전부 포함시키지 못하는 점과, 모델링 적용에 있어 학사 수준 분석 방법이 실제 적용하기에는 다소 무리가 있을 수 있다는 한계가 있을 수 있습니다. 또한 실거래가 경우, 시차를 두고 거래를 하는데 거래건수마다 각각 다르기 때문에 수집하기 힘들다는 부분과 최근 3년간 거래된 데이터이기 때문에 예상치 못한 미래의 사건, 사고에 대해서는 예측의 정확도의 한계가 존재합니다. 단지 내 아파트에도 층 별, 방향 별 등 세부적인 사항보다는 단지를 기준으로 예측하였습니다.

(2) 분석 한계 : 의사결정모형의 Pruning 단계에서 terminal node수가 많을수록 오분율이 계속 낮아지는 추세를 갖기에 적정한 수준의 terminal node를 선정하는데 어려움이 있었고, 회귀모형의 경우 다중 공선성을 해결하기위해 변수를 제거 및 선택을 했음에도 불구하고 test set의 MSE가 validation set의 것보다 크게 나온 것이 해석의 어려움이 있었습니다. 그 외에도 인천공항이 있는 중구의 운서동의 스타벅스가 가장 많은데 아파트값이 낮은 편이거나, 송도 같은 특정동만 집값이 높은 곳의 해석, 그리고 회귀분석시 cook’s distance에서 outlier 들이 보이는 변수들 해석의 어려움이 있었습니다.