

데이터 분석을 통한 최적의 행복주택 입지 선정

서지수
아주대학교

xtjwltnx@ajou.ac.kr

Optimal Happy-House Location Selection through Data Analysis

Seo Ji Soo
Ajou Univ.

요 약

정부에서 행복주택 건설을 포함한 주택 지원 정책을 실시하고 있지만, 젊은 층의 주거불안과 고령층의 주거 부족 문제는 심각하다. 지역별 정확한 수요예측 지표가 없는 주거 지원은 건설 이후 수급 불일치 현상을 야기할 뿐 아니라 보급이 원활히 이루어지지 않을 수 있다. 따라서 본 논문에서는 인구, 세대 수, 소득 등의 데이터를 이용하여 연령별 주거지역 선호요소를 도출하고 지역 특성과 비교하여 가장 효율적인 주거지원을 위한 수요예상 지표를 제공한다.

I. 서 론

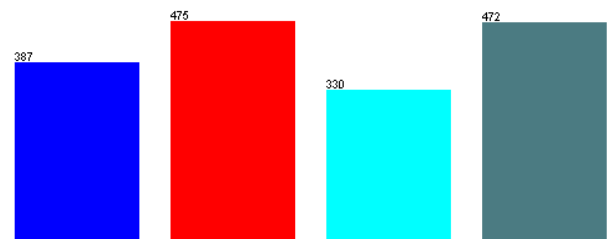
전세가격 폭등, 임대주택 부족 현상, 높은 실업률 등으로 인해 젊은 층의 주거 불안이 가중되고 있다. 이 문제를 해결하기 위해 현 정부는 임차료가 저렴한 도심형 아파트인 행복주택 건설을 추진하고 있다.[1] 최근 행복주택은 높은 경쟁률을 보이고 있다. 그러나 계약율이 저조하다는 문제가 있다. 그 이유로는 주거지가 연령별 타겟층에 적합하지 않다는 점, 자금 마련이 어렵다는 점이 있다. 행복주택의 인지도와 조건에 따른 만족도는 입주자의사에 통계적으로 영향을 미치지 않는다.[2] 지역별 정확한 수요예측지표가 없는 주거지원의 건설 이후 수급불일치로 보급이 원활히 이루어지지 않을 수 있다.

이에 본 연구에서는 데이터 분석 기법을 활용하여 연령별 행복주택 입지 조건에 가장 큰 영향을 미치는 요인을 찾아본다. 인구이동 데이터를 이용하여 연령별 주거지역 선호요소를 도출한다. 이를 지역특성과 비교하여 연령별 최적의 입지 선정을 위한 모델을 만들고자 한다.

총 인구 및 소득에 해당하는 비율로 채워 넣었다. 데이터 사이의 간격이 커서 한쪽으로 편중되거나 수치가 큰 문제가 있었다. 표준화 과정을 통해 모든 데이터를 통합시킴으로써 편중 현상을 해결하고 간소화했다.

2. 데이터 마이닝

행복주택 지원 가능 계층을 목표치(Target Value)로 설정하기 위해 연령층을 구분할 필요가 있었다. (1)15~29 세, (2)30~49 세, (3)50~64 세, (4)65 세 이상 4 개의 계층으로 구분하였다. 하나의 동에서 가장 큰 비율을 차지하는 계층을 목표치로 설정하였다.



[그림 1] 목표치 데이터 분포도(왼쪽부터 1,2,3,4)

II. 본론

1. 데이터 수집 및 전처리

본 연구를 하는데 필요한 데이터를 수집했다. ‘서울 열린데이터 광장’에서 소득소비, 주민등록인구분포, 혼인, 결혼나이, 전용면적별 세대 수, 세대수별 주거 인구, 집객시설, 주거 환경 심각도, 도로구간 위치정보 데이터를 가져왔다. 데이터를 통합시키기 위한 주요 속성은 구주소의 동 단위로 하였다. 새주소는 도로구간 위치정보 데이터를 활용하여 구주소로 변환하였다. 세대수별 주거 인구, 소득소비 데이터는 결국 값을 포함하고 있었다. 해당 데이터는 누락된 것으로 판단하여

위 목표치를 기준으로 데이터를 분류 및 학습 시키기 위해 ‘지원 벡터 기계(Support Vector Machine)’, ‘나이프 베이즈(Naïve Bayes)’, ‘최근접 이웃(K-Nearest Neighbor)’ 알고리즘을 사용하였다.

[표 1] 모델 성능 비교 평가

	SVM	Naïve Bayes	K-NN
--	-----	-------------	------

ROC Area	0.770	0.784	0.785
F-Measure	0.549	0.543	0.571

해당 모델 알고리즘을 비교 해 보았을 때, ‘나이브 베이즈’가 가장 좋은 성능을 보여줌을 알 수 있다. 따라서 행복주택의 최적의 입지 선정에 위해 ‘나이브 베이즈’ 알고리즘이 가장 적합하다.

3. 데이터 분석

4 개의 목표치에서 지원 대상인 신혼부부를 추출해 내기 위해 수식 (1)을 이용하였다.

$$\frac{\text{한 동에 해당하는 하나의 목표치}}{\text{전체인구}} * (\text{하나의 목표치에 해당하는 신혼부부 수})$$

(1)

‘목표치 1’의 한 동에 해당하는 데이터에서 신혼부부 수를 뺀 값을 청년으로 간주 하였다. 데이터를 쉽게 분석하기 위해 목표치 속성을 제외한 모든 수치 데이터를 [그림 2]와 같이 순위 비율에 따라 재정의했다. 신혼부부 속성은 0 에 가까울수록 비율 및 절대적 수치가 높은 것을 의미한다. 소음 및 범죄 속성은 0 에 가까울수록 해당 문제가 적게 발생함을 의미한다.

구	동	목표치	신혼부부	소음	범죄
은평구	진관동	2	0	0.166	0.125
강서구	화곡1동	2	0	0.375	0.5
강동구	길동	3	0	0.458	0.916

[그림 2] 순위 비율에 따라 재정의 된 데이터

은평구의 진관동, 강서구의 화곡 1 동, 강동구의 길동은 가장 많은 신혼부부가 거주하고 있었다. 하지만, 진관동과 화곡 1 동은 ‘목표치 2’에, 길동은 ‘목표치 3’에 해당하는 인구가 많이 분포했다. 신혼부부가 많이 거주하는 상위 10% 지역은 ‘목표치 2’에 해당하는 연령층이 56%로 주를 이루었다. ‘목표치 4’에 해당하는 연령층은 찾아볼 수 없었다. 진관동과 화곡 1 동은 소음, 범죄 측면에서 좋은 수치를 보여 주었다. 길동은 서울시에서 높은 범죄율을 보였다. 또한 신혼부부가 많이 거주하는 곳은 2~5 인 가구가 많이 거주했다. 1 인가구와는 관련을 보이지 않았다.

평창동, 해화동 등을 포함하는 종로구는 소음과 범죄율이 높았다. 신혼부부의 비율은 높지 않았고, 학생 및 ‘목표치 3,4’연령대가 많이 거주했다. 종로구는 대체적으로 적은 가구원들이 거주했지만, 평창동과 부암동에는 많은 가구원의 분포를 보여주었다. 특히 평창동에서는 5 인 이상 가구원이 함께 거주하는 경우가 주를 이루었다.

해당 데이터 분석 끝에 소음과 범죄를 포함하는 주변 환경 데이터는 각각 유사함을 보였고, 소득과는 관계를 찾을 수 없었다. 1 인 가구가 많은 곳은 2 인 가구도 대체적으로 많았고, 3 인~7 인 이상 가구는 각각 유사함을 보였다.

III. 결론

본논문에서는 데이터 마이닝 기법을 포함하여 데이터 분석을 진행하였다. 데이터의 정확성이 떨어져 정확도는 높지 않았다. 지출, 인구, 세대 수 등의 데이터는 완벽한 선형을 보여주었다. 해당 데이터에서 상호 연관성을 제거 하기엔 데이터가 부족했다. 이러한 이유로 최적 입지 선정에 큰 기여를 하지 못했다고 본다. 하지만, 새로운 방법으로 분석을 시도 해 보았다는 점에서 기여 했다고 볼 수 있다. 또한, 추후 더 많은 데이터로 상호 연관 관계가 높은 데이터 속성을 제거하고, 인공 신경망 알고리즘을 이용한다면 높은 정확도를 내는 데 기여할 수 있다고 본다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업의 연구결과로 수행되었음"(2015-0-00908)

참 고 문 헌

- [1] 김성연, 권성문, 서원석, 양광식." A Study on Happy-House Construction Program Policy Factors ", 국토연구, p. 3-15.
- [2] 권요규, 강은택, " University Student's Determinants of Happy-House Choice - Focused on Gimcheon.", 대한부동산학회, p.161-176