韓國IT서비스學會誌
第21卷 第3號
2022年 6月, pp.63-72

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2022.21.3.063

인공지능 기반 빈집 추정 및 주요 특성 분석*

임규건** · 노종화*** · 이현태**** · 안재익*****

Vacant House Prediction and Important Features Exploration through Artificial Intelligence: In Case of Gunsan*

Gyoo Gun Lim** · Jong Hwa Noh*** · Hyun Tae Lee**** · Jae Ik Ahn*****

■ Abstract ■

The extinction crisis of local cities, caused by a population density increase phenomenon in capital regions, directly causes the increase of vacant houses in local cities. According to population and housing census, Gunsan-si has continuously shown increasing trend of vacant houses during 2015 to 2019. In particular, since Gunsan-si is the city which suffers from doughnut effect and industrial decline, problems regrading to vacant house seems to exacerbate.

This study aims to provide a foundation of a system which can predict and deal with the building that has high risk of becoming vacant house through implementing a data driven vacant house prediction machine learning model. Methodologically, this study analyzes three types of machine learning model by differing the data components. First model is trained based on building register, individual declared land value, house price and socioeconomic data and second model is trained with the same data as first model but with additional POI(Point of Interest) data. Finally, third model is trained with same data as the second model but with excluding water usage and electricity usage data.

As a result, second model shows the best performance based on F1-score. Random Forest, Gradient Boosting Machine, XGBoost and LightGBM which are tree ensemble series, show the best performance as a whole. Additionally, the complexity of the model can be reduced through eliminating independent variables that have correlation coefficient between the variables and vacant house status lower than the 0.1 based on absolute value. Finally, this study suggests XGBoost and LightGBM based machine learning model, which can handle missing values, as final vacant house prediction model.

Keyword : Vacant house prediction, Machine learning, XGBoost, LightGBM, Feature importance

Submitted : June 14, 2022 1st Revision : June 27, 2022 Accepted : June 28, 2022

* 이 논문은 한국국토정보공사 공간정보연구원 산학협력 R&D사업의 지원을 받아 수행된 연구임(과제명 : 인공지능 기반 빈집추정 및 가치산정에 대한 연구. 과제번호: 2021-504). 본고는 International Conference on Electronic Commerce 2022에서 발표한 내용을 기반으로 재작성한 것임.

** 한양대학교 경영대학 교수
*** 한양대학교 경영대학 석사과정
**** 한양대학교 경영대학 석사과정
***** 한양대학교 경영대학 박사 수료, 교신저자

1. 서 론

계속해서 심각해지는 수도권 집중 현상과 지방소멸 현상으로 인해 지방 도시의 구도심 슬럼화 현상은 빠르게 진행되고 있다. 이러한 현상은 빈집의 숫자를 계속해서 증가시키는 주된 요인 중의 하나이다. 실제로 2015년에서 2019년 사이 진행된 인구주택총조사에 따르면, 전국 빈집은 2015년 107만호에서 2019년 152만호로 약 50만호 증가했다(통계청 2015, 통계청 2019).

또한, 2019년 인구주택총조사에 따르면, 30년 이상 된 빈집 비율이 전북이 43.7%(4만 호)를 기록하면서 전국 지역 중에서 2번째로 높았고, 주택의 약 12.6%가 빈집인 것으로 파악됐다(통계청, 2019). 따라서, 전북 지역의 경우, 빈집 수가 많을 뿐만 아니라 노후도 역시 상당히 진행됐다는 것을 알 수 있다. 군산시 역시 조선소 운영 중단, 자동차 공장(GM) 폐쇄 등으로 인한 산업 쇠퇴와 구도심 공동화로 인해 전체 빈집 비율이 14.01%에 달한다(김동인, 2019).

이러한 현상을 계속해서 방치한다면 지방의 빈집 확산과 슬럼화가 가속화될 것으로 예상된다. 따라서 빈집이거나 빈집이 될 가능성이 큰 건축물들을 조기 발견해 빠른 행정적인 처리를 할 수 있는 시스템 개발의 필요성이 커지고 있다. 빈집과 연관성이 있는 다양한 데이터 학습을 통해 만들어진 머신러닝 모델을 기반으로 해당 시스템을 만들어진다면 효율적인 빈집탐지와 대응책 마련이 가능해질 것이다.

이에 본 연구는 건축물 관련 데이터, 상수도 사용량 데이터, 전기 사용량 데이터 등 빈집탐지와 연관된 자료들을 학습해 빈집 추정을 위한 미신러닝 모델 구축하고자 한다. 본 연구를 통해 개발될 머신러닝 모델은 데이터를 기반으로 빈집 위험도를 사전에 예측하고 어떠한 요인들이 빈집 생성에 영향도가 큰지를 파악해 빠른 빈집 대응 행정 절차에 이바지할 수 있을 것이다.

2. 이론적 배경

2.1 선행연구

빈집 추정과 빈집 발생 요인을 분석하기 위해 국내외에서 다양한 연구가 수행되었다.

이형석, 김승희(2018)는 강원도의 18개 지자체별 특성에 따른 군집분석을 통하여, 빈집의 유형과 특성을 살펴봤다. 이 연구에서는 분석을 위해 2015년 기준으로 인구, 사회, 가구 주택을 나타내는 지표를 19개 선정하여 지역별 특성에 따라 18개 지자체를 5개의 유형으로 군집화 및 특성 분석을 하였다.

이홍대(2018)는 빈집 및 소규모주택 정비에 관한 특례법 시행에 맞춘 지역 내 빈집이라는 자원 활용을 위해 빈집의 발생 원인을 거시적 관점에서 분석하고 활용방안을 미시적 관점으로 살펴봤다. 연구결과, 도시의 공가율에 영향을 미치는 변수는 가구수 대비 주택비율과 노후주택비율 변수인 것으로 나타났고, 비도시는 주택거래현황, 노후주택비율, 가구수 대비 주택비율 변수인 것으로 나타났다.

김현중 등(2020)은 부산시를 사례로 근린특성이 빈집이 발생하는데 미친 영향을 분석하고, 관리 및 예방을 위해 효과가 높은 근린특성을 탐색했다. 빈집은 고령자의 밀도가 높을수록 인구밀도가 낮을수록 발생 빈도가 높았으며, 빈집 예방에는 사업체 수, 주택 가격, 주택 연령 등이 큰 영향을 미친다는 것을 보여주었다.

김윤수(2020)는 인구가구, 주택, 주택거래, 경제, 지역 등 각 특성이 빈집에 미치는 영향을 분석하였다. 연구를 통해 빈집율에 영향을 미치는 특성이 주택 유형별로 다르기 때문에 각 유형별로 방안을 제시할 것을 언급하였다.

Amy et al.(2003)은 집 가격, 수도세 납부 여부 등의 데이터를 활용해 로지스틱 회귀 분석으로 빈집 위험성이 있는 집을 예측해 분석하였다.

Morckel(2014)은 multi-level regression 모델을 활용해 빈집이 될 확률이 공간 요소에 영향을 받는지에 대해서 분석하였다. 분석 결과, 공간 관계를 포함한 multi-level regression 모델이 그렇지 않은

모델보다 예측 성능이 더 높은 것을 밝혀냈다.

Porzi et al.(2015)은 CNN을 바탕으로 Google Street View 이미지를 활용해 도시 지역의 안전도 인식을 예측하는 모델을 제안하였다.

2.2 선행연구의 한계점

대부분의 선행연구는 로지스틱 회귀를 통한 예측이 주를 이루었다. 또한, 다양한 머신러닝 알고리즘을 활용해 성능 및 특성 중요도 비교 분석이 이루어진 것이 아닌 통계적 분석이 주를 이루었다. 인공지능 학습법이 통계적 분석에 우위에 있다고 할 수 없지만, 변수 간 관계성과 통계적 유의미성 확인보다 예측 성능 확보를 주목표로 한다면 인공지능 알고리즘이 유리하기 때문에 빈집 추정에 있어 인공지능을 활용한 연구가 필요하다.

본 연구에서는 빈집 추정에 활용되는 각 건축물 주소의 500미터 근처 국가관심지정정보(POI: Point of Interest)와 예측 성능에 영향력이 가장 클 것이라고 예상되는 상수도 사용량, 전기 사용량을 활용한 모델과 그렇지 않은 모델을 비교 분석해 빈집 추정 정확도에 영향력이 큰 변수를 탐색하였다.

또한, 현장에서 데이터 수집할 때 결측값 없이 데이터를 모은다는 것은 거의 불가능에 가깝다. 하지만 앞에서 언급한 선행연구에서는 결측값 처리에 관한 연구가 이루어지지 않았다. 따라서 본 연구에서는 결측값 자체 처리 기능이 있는 트리 부스팅 계열인 XGBoost와 LightGBM이 결측값이 있는 데이터를 학습해도 결측값을 제거한 모델과 성능에 차이가 없는지 비교 분석하였다. 이를 통해 현장에서 쓰일 수 있는 현실성 있는 머신러닝 알고리즘을 제시하고자 한다.

3. 연구방법

3.1 자료 구축

본 연구에 활용한 데이터는 한국국토정보공사(이하 LX), 군산시, 공공데이터포털, 국가통계포털(KOSIS), 국가공간정보포털(NSDI)로부터 수집했다.

각 주소의 빈집 여부는 LX로부터 제공된 현장 조사 결과 데이터이다. 활용 데이터는 <표 1>과 같다.

〈표 1〉 Data Status and Sources

Usage data	Variable name
Result of vacant house investigation	Vacant house
Water usage	Water supply_6 months_danger, Water supply_12 months_confirmed
Electricity usage	Electricity usage
Building register	Land area, Building area, Gross floor area, Main structure, Main use, Main roof, Number of households, Number of households[families], Number of ground floors, Number of basement floors, Approval of use date
Individual announced price information	Individual declared land value
Individual house price information	House price
Status of Recipients of Basic Livelihood	Basic Livelihood Recipient Status, Basic Livelihood Recipient Ratio
Population	Total population, Male population, Female population, Foreign male population, Foreign female population, People over 65, Percentage of the population over 65
Point of Interest	POI Feature 1 ~ 3

국가공간정보포털로부터 제공 받은 국가 관심지정정보(POI) 데이터는 <표 2>와 같다. 이러한 과정을 거쳐 총 6,249건의 데이터를 수집하였다.

〈표 2〉 Point of Interest Data

Variable	Variable name
POI	Elderly facility, Religious facility, Education facility, Welfare facility, Residential facility, Cultural facility, Commercial facility, Exercise facility, Transportaion facility, Rest facility, Living Convenience facility, Nature, Public institutions, Health care facility, Environmental facility, Industrial complex, Food and beverage facility, Accommodation, Financial facility, Communication facility, Tour facility

3.2 머신러닝 모델 및 평가 지표

본 연구에서 활용되는 머신러닝 방법론은 크게 선형 계열과 트리 계열, 부스팅 계열, 신경망 계열로 나뉘어서 비교 분석하였다. 선형 계열에서는 Logistic Regression, Ridge Regression, Lasso Regression, SVM, 트리 계열에서는 Decision Tree, Random Forest. 부스팅 계열에서는 Gradient Boosting, XGBoost, LightGBM, 신경망 계열에서는 Deep Neural Network (DNN)을 활용하였다.

3.3 인공지능 학습을 위한 데이터 전처리

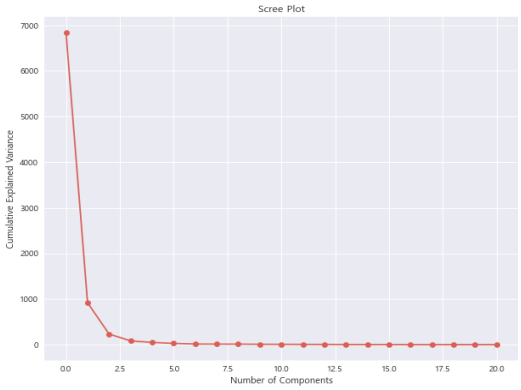
머신러닝 학습에 활용될 POI 데이터는 각 주소 주변 500미터 근처에 POI 변수에 해당하는 시설의 수를 산출한 것이다. 먼저, 빈집 여부와의 상관관계를 분석하였다. 상관관계 분석 결과를 상관계수 절댓값 기준으로 내림차순으로 상위 10개로 정렬한 결과는 <표 3>과 같다.

〈표 3〉 Correlation Coefficient of Each POI Variables

POI Variable	Correlation Coefficient
Housing facility	-0.25263
Food and Beverage facility	-0.21462
Industrial complex	-0.17454
Nature	0.150241
Commercial facility	-0.10525
Welfare facility	-0.09929
Environment facility	-0.05486
Religious facility	-0.0548
Transportation facility	0.047067
Education facility	-0.04319

주거시설, 식음료시설, 산업단지, 자연, 상업시설을 제외하고는 대부분의 POI 변수의 상관계수가 절댓값 기준으로 0.1보다 낮은 것을 볼 수 있다. 따라서, POI 데이터 전반적으로 빈집 여부에 큰 상관관계를 보이지 않았다. 이러한 데이터를 그대로 반영할 경우, 열의 개수가 불필요하게 많아지고 특정한 POI 변수는 0값이 많으므로 학습 시간이 길어지고 성능 저하의 원인이 될 수 있다. 따라서, POI 데이터를 종합적으로 학습 데이터에 반영하기 위해 주성분 분석(PCA) 기법을 활용해 차원 축소를 실행하여 재가공하였다. 이때, 적절한 주성분 수를 선정하기 위해 고윳값-주성분 분산 변화를 보는 그래프인 Scree plot을 활용하였다. [그림 1]과 같이 고윳값이 3차원에서 4차원으로의 고윳값 변환율이 완만해지기 시작하는 지점이 적절한 주성분의 수로 파악된다. 이러한 분석을 바탕으로, 총 21차원을 가진 POI 데이터를 3차원의 데이터로 축소하였다.

건축물의 노후도를 반영하기 위해 날짜 형식으로 구성된 사용승인일(Approval of use date) 데이터를 현재 연도에서 사용승인일 날짜의 연도를 빼서 학습에 활용했다.



[그림 1] POI 데이터 Scree Plot

〈표 4〉 Skewness of Each Variables

Variable	Skewness
Water supply_6 months_danger	33.839329
Water supply_12 months_confirmed	32.518073
Electricity usage	17.551314
Number of households	14.539291
Gross Floor area	11.355888
Building area	9.523057
Land area	7.998586
Number of basement floors	3.231410

선형기반 알고리즘의 경우, 변수들의 분포가 정규분포를 따라야지만 분석의 정확도를 높일 수 있으므로, 각 변수의 왜도 값(비대칭도)을 구해 치우침이 심한 데이터는 로그 변환을 통해 정규성을 부여했다. 왜도 값이 절댓값 기준으로 2보다 큰 값을 가진 변수는 <표 4>와 같다.

이렇게 왜도 값이 절댓값 기준으로 2보다 큰 변수는 치우침이 크다고 판단하여 로그 변환을 실시했다. 다만 지하층수(Number of basement floors) 변수 같은 경우는 스케일이 크지 않기 때문에 로그변환을 하지 않았다. 그 후에는, 전체 데이터의 분포 범위를 모두 통일하기 위해 Scikit-learn의 MinMaxScaler를 통해 연속형 변수는 모두 0과 1 사이의 값을 가지도록 했다. 이 과정은 선형기반 알고리즘과 DNN 알고리즘 학습 시에만 적용됐다.

건축물의 구조와 용도를 나타내는 변수인 주구조(Main structure), 주용도(Main use) 그리고 주지붕(Main roof) 데이터는 원-핫 인코딩(One-hot encoding)을 적용했다. 이 과정은 모든 알고리즘에 공통으로 적용됐다.

본 연구에서는 첫 번째 실험 단계에서는 결측값이 포함한 행은 모두 제거하고, 첫 번째 모델(Model 1)에서는 <표 1> 데이터에서 POI 데이터를 제외하고 학습하였다. 두 번째 모델(Model 2)에서는 <표 1>의 모든 데이터를 학습했으며, 세 번째 모델(Model 3)은 <표 1>에서 상수도 누적 사용량 데이터(Water supply_6_months_danger, Water supply_12_months_confirmed) 데이터와 전기 12개월 누적 사용량 데이터 (Electricity usage) 데이터를 제거해 학습했다. 이렇게 학습 데이터를 달리해 POI 데이터와 상수도, 전기 사용량 데이터가 빈집추정 정확도에 끼치는 영향을 비교 분석했다.

최종적으로 전처리한 결과, 첫 번째 모델과 두 번째 모델은 2,896건의 데이터를 학습했고, 세 번째 모델은 4,002건의 데이터를 학습했다. 모든 모델에 대해 Train Set과 Test Set은 7:3의 비율로 학습 및 성능 평가를 실행했으며 Cross Validation을 통해 Overfitting 문제를 최소화 하고자 했다.

4. 결과 및 고찰

4.1 빈집 추정 모델 실험 결과

첫 번째, 두 번째, 세 번째 모델의 머신러닝 별 알고리즘 결과는 <표 5>와 같다.

전체적인 결과를 살펴보면, <표 1>에 있는 모든 데이터를 토대로 학습한 Model 2가 제일 좋은 성능을 보여줬다. POI를 학습하지 않은 Model 1은 Model 2보다 SVM을 제외하고 F1-score가 소폭 감소한 것을 볼 수 있다. 또한, Accuracy 측면에서

〈표 5〉 Accuracy and F1-score of Each Models by Algorithms

Algorithms	Model 1		Model 2		Model 3	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Logistic Regression	0.93	0.80	0.94	0.83	0.83	0.67
Ridge Regression	0.91	0.76	0.92	0.78	0.84	0.68
Lasso Regression	0.94	0.82	0.94	0.84	0.83	0.66
SVM	0.94	0.82	0.94	0.82	0.84	0.68
Decision Tree	0.95	0.86	0.95	0.89	0.81	0.68
Random Forest	0.96	0.89	0.97	0.91	0.85	0.62
Gradient Boosting Machine	0.96	0.89	0.97	0.93	0.85	0.71
XGBoost	0.96	0.90	0.97	0.91	0.83	0.72
LightGBM	0.97	0.91	0.97	0.92	0.83	0.72
DNN	0.94	0.84	0.95	0.86	0.83	0.71

보면, Lasso Regression, SVM, Decision Tree, LightGBM은 Model 1과 Model 2랑 비슷하지만, 다른 알고리즘에서는 Model 1이 Model 2보다 소폭 감소한 것을 볼 수 있다. 전기, 상수도 데이터를 제외하고 학습한 Model 3에서는 F1-score와 Accuracy가 Model 1, Model 2보다 대폭 하락한 것을 볼 수 있다. 이를 통해 전기, 상수도 데이터가 빈집추정 성능 향상에 중요한 역할을 하고 있다는 것을 파악할 수 있다.

또한, 전체적으로 트리 앙상블 계열인 Random Forest, Gradient Boosting Machine, XGBoost, LightGBM이 제일 좋은 성능을 보여주고 있다.

Model 2는 다른 모델에 비해 추정 성능이 좋지만 <표 1>의 모든 변수를 사용해 학습된 모델인 만큼 모델 복잡도에 대한 우려가 있을 수 있다. 또한, 예측 성능에 도움이 되지 않을만한 변수들을 제거해 학습한다면, 머신러닝 모델 학습을 위해 현장에서 수집해야 할 데이터 수도 줄일 수 있을 것이다. 따라서, <표 1>에 해당하는 데이터와 빈집 여부 간의 상관관계를 분석해 상관계수가 절댓값 기준으로 0.1보다 낮은 변수들과 다른 변수들과 상관관계가 높은 변수들은 제거했다. 독립변수와 빈집 여부 간의 상관계수는 <표 6>과 같다.

〈표 6〉 Correlation Coefficient of Each Variables

Variables	Correlation Coefficient
Main use_apartment	0.38
Main use_detached house	-0.38
House price	-0.34
Number of households	-0.28
Main roof_concrete	-0.23
Main roof_slate	0.22
Main structure_wood	0.20
POI feature 2	-0.20
POI feature 1	-0.18
Main structure_block	0.17
Approval of use date	-0.16
Main roof_tile	0.15
Number of ground floors	-0.14
Basic Livelihood Recipient Status	-0.14
Electricity useage	-0.13
Building area	0.13
Foreign female population	-0.13

상관 분석을 통해 주로 건축물의 구조에 대한 정보를 담고 있는 주용도(Main use), 주지붕(Main roof), 주구조(Main structure) 데이터가 빈집 여부와 높은 상관관계를 보여주고 있는 것을 알 수 있다. 또한, 건축물의 구조와 관련된 정보 이외에도 주택가격, 사용승인일, 전기 사용량 최종과 같은 데이터들도 빈집 현황과의 상관관계가 0.1보다 높은 수치를 보여주고 있다.

모델 최적화를 위해 상관계수 절댓값이 0.1보다 낮은 변수는 제거하고 인구 관련 데이터는 총인구(Total population) 데이터와 높은 상관계수가 계산되므로 총인구 데이터를 제외한 관련된 모든 인구 데이터는 제거하였다. 또한, 상수도 데이터 역시 상관계수가 -0.02를 기록하였으므로 제거하였다. 이런 과정을 통해 구축된 최종 학습 데이터는 <표 7>과 같다.

〈표 7〉 Final Selected Variables

Usage data	Variable name
Electricity usage	Electricity usage
Building register	Building area, Number of households, Number of households(families), Number of ground floors, Number of basement floors, Approval of use date, Main structure (block, wood, concrete), Main use(apartment, detached house), Main roof (concrete, tile, slate)
Individual announced price information	Individual declared land value
Individual house price information	House price
Population by Eup/Myeon/Dong in Gunsan-si	Total population
Point of Interest	POI Feature 1 ~ 3

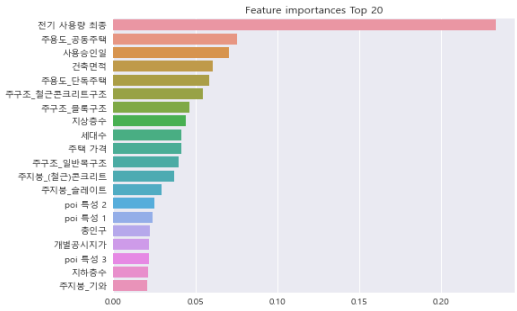
또한, 이렇게 선택된 변수 중에서 결측값이 전체 데이터 중 35% 비율 이내로 존재하는 변수가 다수 존재한다. 예를 들어, 주택가격(House price)은 전체 데이터 중 34% 결측값이 존재했고, 가장 영향력이 클 것이라고 예상되는 전기 사용량 최종(Electricity usage) 데이터도 21% 정도 결측값이 존재했다. 결측값이 존재하는 행을 제거하게 된다면 총 6249개 데이터에서 학습할 수 있는 데이터 수가 줄어들기 때문에 정보 손실에 대한 문제도 생길 수 있다. 따라서 <표 5>에서 좋은 성능을 보여주었고 결측값 대체 기능이 있는 XGBoost와 LightGBM을 토대로 <표 7>에 있는 데이터를 토대로 학습을 진행하였다. 이때, 결측값을 제거하지 않아, 수집한 총 6,249개 데이터를 모두 활용할 수 있었다. 이렇게 재구성된 데이터로 수행한 학습 결과는 <표 8>과 같다. 여기서 Model 2_D는 전기 데이터를 포함해서 학습한 모델이고 Model 3_D는 전기 데이터를 제거하고 학습한 모델이다.

〈표 8〉 Accuracy and F1-score of XGBoost and LightGBM Model Trained with Missing Values

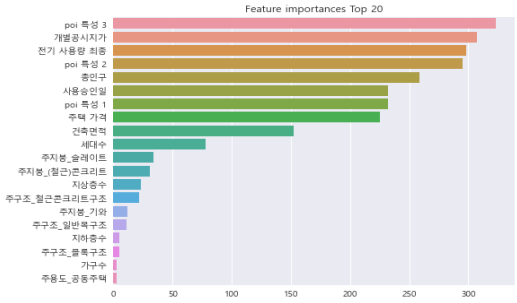
Algorithms	Model 2_D		Model 3_D	
	Accuracy	F1-score	Accuracy	F1-score
XGBoost	0.92	0.90	0.78	0.75
LightGBM	0.92	0.90	0.79	0.75

<표 8>에서 볼 수 있듯이, 결측값이 35% 내외로 존재하는 변수가 존재하는 데이터셋을 활용했음에도 불구하고, <표 5>에서 Model 2의 XGBoost와 LightGBM과 비슷한 성능을 보여주고 있다. 이는 결측값이 다소 존재하더라도 빈집추정의 데이터로 활용될 수 있다는 것을 함의한다고 볼 수 있다. 또한, 상관계수가 낮아 상수도 데이터를 제거한 <표 8>에서의 Model 2_D와 상수도 데이터가 포함된 <표 5>의 Model 2와 F1-score 측면에서 성능에 큰 차이가 없고, 결측값은 제거하지 않고 전기 데이터를 제거한 <표 8>의 Model 3_D가 <표 8>의 Model 2보다 성능이 하락한 것을 볼 수 있다. 이를 통해, 빈집추정에 제일 결정적인 역할을 하는 것은 전기 사용량 데이터라는 결론을 내릴 수 있었다.

또한, XGBoost와 LightGBM은 다른 트리 계열 알고리즘과 마찬가지로 특성 중요도 분석을 수행할 수 있다. 해당 알고리즘의 특성 중요도 분석 결과는 [그림 2]와 [그림 3]과 같다. XGBoost의 경우, 전기 사용량 데이터가 제일 중요한 변수로 파악되었고, LightGBM의 경우, 전기 사용량 데이터가 3번째로 중요한 변수로 파악되었다. 결론적으로, 특성 중요도 분석을 통해 전기 사용량 데이터가 빈집추정에 있어서 가장 중요한 역할을 한다는 것을 다시 한 번 확인할 수 있었다.



[그림 2] Feature Importance of Final XGBoost



[그림 3] Feature Importance of Final LightGBM

최종적으로, <표 7>에 있는 데이터를 학습하고 결측값이 대체 기능이 있는 XGBoost와 LightGBM으로 학습한 모델이 현실적으로 최종 빈집추정 문제에 있어서 제일 적합한 머신러닝 모델이라는 결론을 내릴

수 있게 된다.

5. 결 론

최근 우리 사회는 수도권 집중 현상과 지방소멸 현상으로 인해 지방의 구도심 슬럼화와 함께 빈집 발생률이 높아지고 있다. 이를 방지하고 속도를 낮추기 위해서는 빈집 발생 확률이 높은 건축물들을 조기에 발견하고 적절한 처분이 필요할 것이다. 이에 본 연구에서는 건축물 데이터, 전기·상수도 데이터 등을 활용하여 조기에 빈집을 추정할 수 있는 예측 모델을 구축하고자 하였다.

전체 데이터를 학습한 Model 2가 모든 알고리즘에서 POI 데이터를 제외한 Model 1하고 상수도, 전기 사용량 데이터를 제외하고 학습한 Model 2보다 좋은 성능을 보여줬다. 하지만, 전체 데이터 중 불필요한 데이터를 제거해 더 효율적인 머신러닝 모델을 만들 수 있다.

따라서 상관계수가 절댓값 기준으로 0.1보다 낮은 변수를 제거하고 다른 독립변수 간에 상관관계가 높은 변수들을 제거해 최종 학습 데이터셋을 구축하였다. 또한, 최종 학습 데이터셋에 있는 데이터 중에서 결측값이 35% 이내로 존재하는 특성 데이터가 다수 존재했다. 이러한 데이터가 포함된 행을 모두 제거하게 된다면 정보 손실의 우려가 있기 때문에 결측값 대처 처리가 가능한 XGBoost와 LightGBM을 활용하였다. 그 결과, 결측값을 제거한 데이터셋으로 학습한 결과도 전체 데이터로 학습한 Model 2와 비슷한 성능을 보여줬다. 이는 결측값이 다소 존재하더라도 빈집추정의 자료로 활용될 수 있다는 것을 함의한다고 볼 수 있다.

이를 통해, 현장에서는 결측값 없이 데이터를 수집하는 것은 거의 불가능에 가깝기 때문에 XGBoost와 LightGBM이 빈집추정에 가장 현실적인 머신러닝 모델이라고 결론을 내릴 수 있다.

또한, 최종적으로 제안된 모델에서는 상수도 데이터를 제거한 데이터셋을 학습한 모델임에도 불구하고, 상수도 사용량 데이터를 학습한 Model 2와 비슷한 성능을 보여줬다. 또한, 최종 모델에서 XGBoost와 LightGBM의 특성 중요도 분석 결과 두 알고리즘 모두 전기 사용량 최종 데이터가 중요한 변수로 파악됐다. 따라서 전기 사용량 데이터가 빈집 추정하는 데 결정적인 역할을 한다고 볼 수 있다. 따라서 추후 정확한 빈집추정 머신러닝 모델을 개발하기 위해서는 <표 7>에 속한 최종 데이터와 그중에서도 특히 전기 사용량 데이터를 정확하게 수집할 수 있는 시스템을 구축하는 것이 가장 중요하다고 판단된다.

본 연구의 한계점으로는 군산시 지역의 데이터만 활용하여 학습된 예측 모델이기 때문에 타 지역에 동일 모델로 빈집을 추정할 시 오차가 발생할 수 있다.

향후 연구에서는 여러 지역의 데이터 확보를 통해 전 지역에서 활용할 수 있는 예측 모델을 시도해야 할 것이다. 또한 빈집 등급 예측, 빈집 가치 평가, 시계열 변화에 따른 빈집 행태 등 빈집 추정을 넘어선 빈집 문제를 직접적으로 해결하기 위한 여러 연구들이 필요할 것이다.

참고문헌

김동인, “빈집에 울려퍼지는 지방도시의 신음. 시사IN, https://www.sisain.co.kr/news/articleView.html?idxno=40514, 2020.

김윤수, “경기도 주택유형별 빈집발생에 영향을 미치는 특성분석”, 국내석사학위논문, 한양대학교 도시대학원, 2020.

김현중, 성은영, 여관현, “빈집의 선제적 관리를 위한 근린환경 요인 탐색: 부산광역시를 사례로”, *한국도시설계학회지*, 제21권, 제6호, 2020, 137-150.

이형석, 김승희, “빈집의 지역별 유형과 특성: 강원도 18개 시·군을 중심으로”, *사회과학연구*, 제57권, 제2호, 2018, 37-64.

이홍대, “빈집의 발생 원인에 따른 지역별 활용방안에 관한 연구”, 국내박사학위논문, 공주대학교 대학원, 2018.

통계청, “2015 인구주택총조사 표본집계결과(인구,

가구, 주택 기본특성항목) 보도자료", https://kostat.go.kr/portal/korea/kor_nw/1/2/2/index.board?bmode=read&aSeq=358170&pageNo=&rowNum=10&amSeq=&sTarget=&sTxt=, 2015.

통계청, "2019년 인구주택총조사 보도자료 집계결과(배포용)", https://kostat.go.kr/portal/korea/kor_nw/1/2/2/index.board?bmode=read&aSeq=384690, 2019.

Chen, T. and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", https://doi.org/10.48550/arXiv.1603.02754, 2016.

Hillier, A.E., D.P. Culhane, and T.E. Smith, Tomlin, C. D., "Predicting Housing Abandonment with the Philadelphia Neighborhood Information System", *Journal of Urban Affairs*, Vol.25, No.1, 2003, 91-105.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, 3149-3157.

Morckel, V. C., "Spatial Characteristics of Housing Abandonment", *Applied Geography*, Vol.48, 2014, 8-16.

Natekin, A. and A. Knoll, "Gradient Boosting Machines, a Tutorial. Frontiers in Neurorobotics", https://doi.org/10.3389/fnbot.2013.00021, 2013.

Porzi, L., S. Rota Bulò, B. Lepri, and E. Ricci, "Predicting and Understanding Urban Perception with Convolutional Neural Networks", *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, 139-148.

Xu, F., H. C. Ho, G. Chi, and Z. Wang, "Abandoned Rural Residential Land: Using Machine Learning Techniques to Identify Rural Residential Land Vulnerable to Be Abandoned in Mountainous Areas", *Habitat International*, Vol.84, 2019, 43-56.

About the Authors

임 규 건 (gglim@hanyang.ac.kr)

한양대학교 경영대학 임규건 교수는 KAIST 전산학 학사, POSTECH 컴퓨터 석사, KAIST 경영공학 박사학위를 취득하였고, 삼성전자, KT, 국제전자상거래 연구센터(ICEC) 연구위원, 세종대학교 경영학과 교수를 역임하였다. 관심분야는 혁신비즈니스모델, IT서비스 혁신, 인공지능과 경영, e-Business 등이며, 2018년 IT서비스 우수연구인상을, 2009년 IT Innovation 유공자 지식경제부 장관 표창과 2007년 SW산업발전 유공자 정통부 장관 표창을 수여하였다. 주요 저서로는 '경영을 위한 정보기술', 'e-비즈니스 경영', '디지털경제시대의 경영정보시스템' 등 전문서적과 다수의 논문과 특허가 있다. 또한, 아시아최초 상용인터넷인 KORNET 상용화, 중국 Shanghai Telecom SI사업전략, 한국영화 기술 로드맵, KTI 사업전략, 나라장터 (G2B) 효과평가, 행정정보화(G4C) 성과분석, 국가정보보호지수개발, 국방정보화수준평가모형, IT혁신인력양성 종합대책, 국가디지털식별체계(UCI), 저작권정품인증제도, SW사업자신고 제도개선, SW기술자신고제도개선 등 다양한 IT혁신분야의 프로젝트를 수행하였다.

노 종 화 (jonghwa0315@hanyang.ac.kr)

한양대학교 경영대학 비즈니스인포매틱스학과 석사과정에 재학 중이다. 관심분야는 빅데이터 분석, 디지털 무역, 혁신비즈니스모델 등이다.

이 현 태(leeht0113@hanyang.ac.kr)

한양대학교 경영대학 비즈니스인포매틱스학과 석사과정에 재학 중이다. 주요 관심분야는 감성분석, 자연어처리 그리고 추천 시스템이다.

안 재 익(anssame@hanyang.ac.kr)

한양대학교 비즈니스인포매틱스학과 박사과정을 수료하였다. 주요 관심분야는 빅데이터 분석, Machine Learning, 디지털마케팅, 혁신비즈니스모델, e-비즈니스이다.