2021년 1학기 산업공학종합설계 최종보고서

랜덤 포레스트, XGBoost, 서포트 벡터 머신, 다중회귀분석을 이용한 일산과 분당의 오피스텔 가격 분석 및 중요 변수 파악

이동환(2015019898)

김성우(2017036453)

김현지(2018015296)

지도 교수 : 송 재 욱 (인)

심사위원 교수 : (인)

한양대학교 산업공학과

국문 초록

이번 연구의 목적은 일산, 분당 각각의 오피스텔 데이터를 다양한 기계학습 기법을 통해 모델별로 독립변수들의 영향력을 측정하고, 모델 간 차이점 비교 분석해서 최적의 모델 선별 및 지역별로 중요변수과 특성 파악하는데 있다. 2018년 기준 가계의 평균 자산에서 부동산이 차지하는 비중이 70%이상이라는 점에서 오피스텔 가격 역시 중요할 수밖에 없다. 가격에 영향을 미치는 요인을 파악하기 위해 4가지 기계학습 기법을 이용했다. 그것은 각각 랜덤 포레스트, XGBoost, 서포트 벡터 머신, 그리고 다중회귀분석이다. 또한 서울을 중심으로 한 연구는 이전에 많이 진행되어서 경기도 일산과 분당을 연구 지역으로 선정했다. 더 좋은 결과를 얻기 위해 오피스텔을 전용면적에 따라 3집단으로 나누고 연구했다. 다양한 알고리즘과 데이터셋을 이용해 매매가격와 관련된 유의미한 결과를 도출하고 중요했던 변수들이 어떤 것인지 파악할 수 있었다.

키워드: 기계 학습, 랜덤 포레스트, 오피스텔 가격, 서포트 벡터 머신, 전용면적

목차

표 목차

1. 서론

1.1 연구의 배경 및 목적

통계청의 2019 년 인구 주택 총 조사에 의하면 1 인 가구는 전국 총 일반 가구 대비 30.2%에 해당하는 것으로 나타났다. 최근 일반 가구 대비 1 인 가구 증가율이 2015 년 27.2%부터 2019 년 30.2%까지 5 년간 꾸준히 상승해왔다. 또한, 통계청 가구 유형별 추계가구에 의하면 2045 년 1 인 가구는 전국 총 일반 가구 대비 37.3%에 달할 것으로 예측된다. 전체 가구 대비 1 인 가구의 비율이 높아짐에 따라 전용면적이 큰 아파트보다 전용면적이 작은 오피스텔의 거주 유형이 증가할 것으로 예상된다.

주로 상업지구에 위치한 오피스텔은 특성상 사업체가 많으며 교통이 편리한 역세권 중심에 위치하며 대규모 점포 인근에 있는 오피스텔의 시장이 확대되어 서울과 인근 신도시의 오피스텔 공급량이 증가하였다. (박종서, 2019)

이처럼 오피스텔의 중요성은 증가하고 있지만 아파트와 오피스텔에 비해 연구가 부족하며, 오피스텔의 선행연구는 대부분 서울을 대상으로 한 연구이며, 수도권과 신도시 지역의 오피스텔 임대료 결정 요인에 관한 연구는 부족한 실정이다. (이민경, 2021)

따라서 본 연구에서는 수도권, 신도시 지역의 오피스텔 가격에 영향을 주는 요인들을 분석해보려고 한다. 지역은 최근 IT 산업이 발달하며 스타트업, IT 회사가 많은 분당과 김포공항과 가깝고 대곡소사선과 GTX 역이 새로 신설되는 일산으로 설정했다.

더불어 오피스텔 규모마다 매매가격을 형성하는 독립변수들의 중요도가 다를 것이라는 가설을 세우고 60m^2 이하는 소형, 60m^2 초과는 중대형으로 분류하여 각각 모델을 만들었다.

본 연구는 오피스텔 매매가에 영향을 미칠 것이라고 예상되는 변수들을 추출해서 실제 거래된 매매가 자료를 바탕으로 규모별로 분류한다. 그 후 다중회귀분석, 랜덤 포레스트, 서포트 벡터 머신, XGBoost 를 이용하여 예측 모델을 만들고, 각각의 RMSE, MAPE 값을 산출하여 모델을 비교한 후 모델 별로 독립변수들의 영향력을 측정하여 모델 간의 차이점을 비교 분석하는 과정으로 진행됐다. 이를 바탕으로 오피스텔 시장을 이해하고 규모별 특성 요인이 오피스텔 매매에 어떤 영향을 미치는지를 분석하고 예측 모델 별 특성을 파악하는 것이 본 연구의 목적이다.

1.2 연구의 범위

1.2.1 공간적 범위

분당 총 87 단지와 총 3747 세대, 일산 총 67 단지와 2698 세대를 대상으로 조사하였다.

1.2.2 시간적 범위

국토교통부 실거래가 공개시스템에 공시된 2018 년 1 월부터 2020 년 12 월까지 오피스텔의 매매가 자료를 기준으로 조사하였다.

2. 선행연구 검토

우리나라의 가계자산 중 가장 높은 비중을 차지하는 것은 주택이다. 통계청의 가계금융・복지 조사 결과에 따르면 2018년 3월 기준 가계의 평균 자산은 약 4억 1,573만 원으로 그 중 부동산이 차지하는 비중은 약 2억 9,177만 원으로 약 70.2%이다. 이처럼 가계의 자산 구조가 주택에 편중되어 있어 주택가격의 급격한 변동이 발생하게 되면 국민 경제에 큰 영향을 미치게 된다. (전해정,2019) 따라서 국내에서는 부동산 가격에 대한 연구가 많은 관심을 받았고, 진행되어 왔다.

부동산이나 주식 가격 예측을 위한 연구는 전부터 계속 진행되어 왔었다. 2000년 이전에는 컴퓨터를 활용한 딥러닝(Deep Learning) 및 머신러닝(Machine Learning)이 아직 활발하게 이루어지지 않아서 그 전까지는 대부분 다중회귀분석(Multiple Linear Regression)으로 예측을 하였다. 다중회귀분석은 두개 이상의 독립변수들과 하나의 종속변수 간에 관계를 수학적인 식을 이용해 파악하는 방식이다. 이러한 다중회귀분석은 이해하기 쉽고 가격을 결정하는 요인이 무엇인지 파악하기 쉽다는 장점이 있지만 표본이나 변수의 수가 많아지면 예측력이 급격히 떨어진다는 단점이 있다. 시계열까지 가미한 다중회귀 시계열분석(ARIMA)으로 더 나아진 결과를 내긴 했지만 최근 폭발적으로 증가한 데이터의 양과 고려변수의 수로 인해 한계가 명확하게 나타난다. 이에 더 나은 결과를 위해 데이터 분석 분야에서 머신러닝이 크게 각광받고 있다. (나성호, 2021).

머신러닝이란 인공 지능을 구현하는 방법으로 기계가 방대한 양의 데이터를 학습하고 새로운 규칙을 새우는 것을 말한다. 다룰 수 있는 데이터의 양은 다중회귀분석과 비교도 안 될 정도로 많으며, 더 정확한 결과를 내기도 한다. 머신러닝의 학습방법은 크게 지도학습(Supervised Learning), 비지도학습(Unsupervised learning), 강화학습(Reinforced Learning)으로 나뉜다. (이주미, 2021)

지도학습이란 명확한 결과 값을 갖고 있는 데이터를 이용한 학습방법이다. 현재 갖고 있는 데이터를 이용해 모델을 만들고 출력 값을 내고 그 값과 실제 값과 비교해서 그 차이를 줄여 나가는 방식으로 구동된다. 정확도 분석을 위한 오차로는 주로 MSE(Mean Squared Error) 과 RMSE(Root Mean Squared Error)이 이용된다. 본 연구에서는 RMSE와 MAPE(Mean Absolute Percent Error)를 정확도 분석을 위한 오차로 사용하였다. 본 연구에서는 지도학습, 그 중에서도 랜덤포레스트와 서포트 벡터 머신, XGBoost 위주로 사용했다. 따라서 다중회귀분석, 랜덤포레스트,

서포트 벡터 머신, XGBoost 총 4개의 알고리즘을 여러 번 이용하고 각각의 RMSE를 비교해 최선의 모델이 무엇인지 파악해볼 것이다.

부동산 가격 분석과 관련해 지금까지 많은 연구들이 진행되었다. 많은 선행연구들이 다중회귀분석, 시계열, 머신러닝을 이용하여 부동산 가격들을 분석했고, 대부분의 연구에서 머신러닝을 사용했을 때 더 우수한 결과를 얻었다. 전반적으로 보면 머신러닝 모형에 의한 추정된 예측값은 실제값과 상당히 유사한 움직임을 보이는 것을 확인 수 있다. (전해정, 2020) 또 부동산 가격과 관련한 연구들에서는 변수와 모델의 선택이 중요하다는 것을 강조하였다.

부동산 지수에 대한 예측을 진행한 논문에서는 랜덤포레스트, XGBoost, LSTM을 사용하여 비교하였고 LSTM방식이 지수 예측에서 전체적으로 좋은 성능을 보였다고 했다(이주미,2021). 전해정(2020)은 시계열 분석 모형과 머신러닝(LSTM, RNN)을 비교했을 때 머신러닝을 사용한 방법이 훨씬 우수한 결과를 얻었다고 하였다. 나성호(2019)는 다중선형회귀, 랜덤 포레스트, SVM을 사용했을 때 세가지 방법 모두 가장 크게 영향을 주는 요인이 전용면적이라는 결과를 얻었고, 랜덤포레스트의 RMSE값이 가장 작다는 결과를 얻었다.

많은 선행 연구들이 가격 연구 대상을 아파트로 정한 점과 여러 구역에 대해 같은 방식을 적용해 일관성 있는 결과가 나오는지에 대한 연구가 많지 않다는 점을 고려하여 이번 연구에서는 오피스텔 가격을 두 지역을 같은 모델을 사용하여 분석한 결과가 일관성이 있는지에 초점을 맞췄다.

3. 자료구조 및 분석 모형

3.1 자료의 구조

이번 분석의 목표는 오피스텔의 크기 별로 오피스텔 매매가격에 영향을 미치는 요인들을 찾는 것이므로, 종속변수로 2018년부터 2020년, 3년 동안 분당, 일산에서 거래된 오피스텔 매매 가격을 사용하였다. 소형 크기는 60m^2 이하, 중대형 크기는 60m^2 초과이다. 독립변수로는 전용면적, 층, 경과연수, 지하철까지의 거리, 대형쇼핑몰까지의 거리, 최고 층 수, 총 세대 수, 용적률, 건폐율, 거래 연월 총 11개를 사용하였다. 목표변수 매매가격과 입력변수 중 전용면적(m^2), 층 수, 경과연수, 거래 연월은 국토교통부 실거래가 공개시스템 공공데이터를 활용하였다. 입력변수 중 지하철 역까지의 거리, 쇼핑몰까지의 거리는 해당 오피스텔로부터 지하철 역, 대형쇼핑몰까지의 도보거리를 km로 나타낸 것이고 부동산 중개 사이트와 네이버 길 찾기를 활용하였다. 최고 층 수, 총 세대 수, 용적률, 건폐율 또한 부동산 중개 사이트와 네이버 부동산을 활용하였다.

독립변수 중 층 수와 전용면적은 매매 거래 건 별 자료이고, 나머지 경과연수, 지하철까지의 거리, 쇼핑몰까지의 거리, 최고 층 수, 총 세대 수, 용적률, 건폐율은 하나의 오피스텔이 나타내는 자료이다.

사용한 독립변수 중 층 수, 경과연수, 최고 층 수, 총 세대 수, 거래 연월은 이산형 자료이고, 나머지는 연속형 자료이다.

독립변수 중 거래 연월은 2018년 1월을 1로 설정하고 마지막 2020년 12월을 36으로 설정하여 총 1부터 36까지의 값을 갖는다.

3.2 자료의 특성

사용된 자료들의 최솟값, 중간값, 최댓값, 평균값, Shapiro-Wilk test의 p-value는 다음 표와 같다.

	Mininum	Median	Maximum	Average	P-value
Area for exclusive use	20.03	40	218.46	58.51	<2.2e-16
Transaction amount (10^4 won)	4000	14000	104000	20275.32	<2.2e-16
Floor	1	8	46	8.421	<2.2e-16
Construction year	1997	2004	2020	2005	<2.2e-16
Highest floor	9	15	77	13.97	<2.2e-16
Total number of households	9	499	1543	530.5	<2.2e-16
Floor area ratio (%)	326	674	972	673.6	<2.2e-16
Building-to-land ratio (%)	22	72	84	68.98	<2.2e-16
Walking distance to subway (km)	0.066	0.45	2.4	0.5054	<2.2e-16
Walking distance to hospital (km)	0.076	0.647	3.8	0.8328	<2.2e-16
Walking distance to market (km)	0.076	0.522	3.8	0.9181	<2.2e-16
Years passed	1	17	25	15.66	<2.2e-16
Months passed	1	21	36	19.27	<2.2e-16

Table 3-1. Ilsan full summary

	Mininum	Median	Maximum	Average
Area for exclusive use	20.03	35.12	59.73	35.76
Transaction amount (10^4 won)	4000	12450	46800	12632.36
Floor	2	7	36	7.542
Construction year	1997	2004	2020	2004
Highest floor	9	10	77	12.37
Total number of households	9	522	1543	563.8
Floor area ratio (%)	451	667	972	684.3
Building-to-land ratio (%)	37	72	84	70.31
Walking distance to subway (km)	0.066	0.399	2.4	0.434
Walking distance to hospital (km)	0.076	0.602	3.74	0.8279

Walking distance to market (km)	0.076	0.583	3.74	0.8988
Years passed	1	17	25	15.99
Months passed	1	18	36	17.97

Table 3-2. Ilsan small size summary

	Mininum	Median	Maximum	Average
Area for exclusive use	60.18	102.32	218.46	109.2
Transaction amount (10^4 won)	12000	34000	37268.82	104000
Floor	1	9	46	10.38
Construction year	1998	2004	2020	2006
Highest floor	10	15	49	17.52
Total number of households	24	317	1069	456.5
Floor area ratio (%)	326	689	920	649.6
Building-to-land ratio (%)	22	72	83	66.03
Walking distance to subway (km)	0.135	0.622	1.7	0.6641
Walking distance to hospital (km)	0.1	0.666	3.8	0.8436
Walking distance to market (km)	0.1	0.419	3.8	0.9611
Years passed	1	16	24	14.94
Months passed	1	25	36	22.17

Table 3-3. Ilsan large size summary

	Mininum	Median	Maximum	Average	P-value
Area for exclusive use	17	38	53.2	247	<2.2e-16
Transaction amount (10^4 won)	8000	23900	322000	31172	<2.2e-16
Floor	1	8	37	11.14	<2.2e-16
Construction year	1997	2004	2018	2006	<2.2e-16
Highest floor	4	16	37	19.14	<2.2e-16
Total number of households	28	546	1968	744.7	<2.2e-16
Floor area ratio (%)	234	687	1290	701.3	<2.2e-16
Building-to-land ratio (%)	32	72	94	69.15	<2.2e-16
Walking distance to subway (km)	0.02	0.38	4.2	0.4487	<2.2e-16
Walking distance to hospital (km)	0.11	1.5	5.1	1.474	<2.2e-16
Walking distance to market (km)	0.08	0.36	4	0.6675	<2.2e-16
Years passed	1	16	24	14.16	<2.2e-16

Months passed	1	23	36	20.58	<2.2e-16

Table 3-4. Bundang full summary

	Mininum	Median	Maximum	Average
Area for exclusive use	17	36	60	36.85
Transaction amount (10^4 won)	8000	21000	53059	22597
Floor	1	8	35	10.34
Construction year	1997	2004	2018	2007
Highest floor	4	12	37	17.89
Total number of households	28	570	1968	739.9
Floor area ratio (%)	260	604	1262	676.5
Building-to-land ratio (%)	47	72	94	70.51
Walking distance to subway (km)	0.02	0.342	3.9	0.3871
Walking distance to hospital (km)	0.11	1.4	3.9	1.412
Walking distance to market (km)	0.08	0.33	4	0.605
Years passed	1	16	24	13.63
Months passed	1	22	36	20.31

Table 3-5. Bundang small size summary

	Mininum	Median	Maximum	Average
Area for exclusive use	61	88	247	104.4
Transaction amount (10^4 won)	21000	55000	322000	58049
Floor	1	12	37	13.64
Construction year	1999	2004	2018	2004
Highest floor	5	27	37	23.07
Total number of households	32	440	1968	759.8
Floor area ratio (%)	234	877	1290	779.2
Building-to-land ratio (%)	32	68	80	64.88
Walking distance to subway (km)	0.02	0.46	4.2	0.6416
Walking distance to hospital (km)	0.11	1.7	5.1	1.666
Walking distance to market (km)	0.08	0.481	4	0.8637
Years passed	2	17	22	15.81
Months passed	1	24	36	21.43

Table 3-6. Bundang large size summary

데이터들이 정규분포를 따르는가에 대한 정규성 검정을 Shapiro-Wilk Test를 통해 실시하였다. 이 테스트의 귀무가설은 '데이터가 정규분포를 따른다'인데, 모든 변수들의 p값이 매우 작은 수준으로 귀무가설을 기각하는 것을 확인할 수 있다. 따라서 모든 변수들이 정규분포를 따르지 않는다는 것을 알 수 있고, 이것을 근거로 우리는 min-max scaler를 사용하여 데이터를 정규화 시켰다.

3.3 분석 모형에 사용된 알고리즘

3.3.1 Multiple Linear Regression

다중회귀분석(Multiple Linear Regression) 은 하나의 종속변수와 여러 독립변수 사이의 관계를 규명할때 이용되는 기법이다. 머신 러닝 기법은 아니지만, 예전부터 계속 사용되어왔던 기본적인 예측모형이다. 현재 쓰이는 대부분의 회귀모형들은 회귀분석에서 비롯되었다고 해도 될 정도로 범위가 넓다. 단순하고 데이터 숫자가 적고 단순할 때 최고의 성능을 보여주지만, 데이터 숫자가 많아지면 예측력이 떨어진다.

종속변수인 y과 독립변수인 x에 대해 식(1)으로 나타낼 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \quad \cdots (1)$$

베타(β) 값은 각 변수에 대한 회귀계수(Regression Coefficients)이다. 각 베타 값을 식(2)을 이용해 찾을 수 있다.

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \cdots (2)$$

예측값이나 중요도는 이 회귀계수에 의해 알 수 있다. 알맞은 회귀계수를 얻기 위해서는 정규분포를 이뤄야하며, 그러기 위해서 데이터셋을 정규화 시켰다. 또한 제대로 된 회귀분석을 위해서는 독립변수들 간에 다중 공선성이 존재하면 안 된다. 본 연구에서는 이를 만족하며, 종속변수는 매매가격, 독립변수는 나머지 변수로 설정했다.

LinearRegression 함수는 OLS(Ordinary Least Squares)를 이용해 최적의 모형을 구하였다. 이 모형을 이용해 마지막에 다른 기법과 비교할 때 똑같은 데이터셋에 각 알고리즘을 실행하고 유의미한 차이가 있는지를 파악했다.

3.3.2 Random Forest

RF (Random Forest)는 여러 개의 나무를 생성하고 결과를 종합하는 앙상블(Ensemble) 알고리즘 중 하나이며 의사결정나무(Decision Tree)를 생성하는데 부트스트랩(Bootstrap) 기법을 이용한다 (Min & Seo, 2017). 의사결정나무는 설명변수 $\mathrm{X1},\mathrm{X2},\cdots,\mathrm{X}p$ 를 순서대로 분류하고 최종 결과치를 제시하고 모든 나무들의 평균을 예측치로 제시한다. RF는 impurity measure 중 하나인 잔차제곱합 (Residual Sum of Squares)이 최소화되도록 분할하고 가지를 쳐내는 작업을 하는데 이것은 식(3)을 최소화하는 작업이다(Lee, 2015). 분류 작업의 경우에는 분류 값들 중 다수의 의견을 따른다.

$$\sum_{m=1}^{|T|}\sum_{x_i\in R_m}(y_i-\widehat{y_{R_m}})^2+\alpha|T| \quad \cdots (3)$$

설명변수에는 평수, 층 등이 입력되었으며 종속변수는 매매가격이다. 본 연구에서는 Relaxed Search를 이용해 최적의 모델을 찾았다. RandomForestRegressor에 쓰일 파라미터(Parameter)는 나무의 숫자를 나타내는 n_estimators 은 5,10,50,100,250,500,750으로, 그리고 나무의 최대 깊이를 나타내는 max_depth 은 1,3,5,7으로 총 28가지 경우의 수를 모두 실행시켰다. 그리고 그 중에서 과적합이 가장 적게 일어난 5개 파라미터에 대해 추가적으로 연구를 했다.

3.3.3 XGBoost

XGB(XGBoost)는 RF을 응용한 것으로, RF와 똑같이 다수의 결정트리를 생성하고 평균치 또는 다중치를 반환하는 머신 러닝 기법이다. 하지만 XGB에서의 Boost 과정은 RF 과 다르게 이전 의사결정트리를 차례로 학습하여 선행 트리의 오류를 개선해 나가는 과정이다. 따라서 이전 단계의 트리 모양에 많은 영향을 받게 된다. XGB의 부스팅(Boosting)을 이용하여 오차를 줄여나가면서 Gain 값이 최대인 최적의 트리를 찾는다. 이러한 최적의 트리들을 조합하여 모형을 만든다(Park, 2019). 아래의 식(4) 와 식(5)는 각각 초기 모델이며 x는 설명변수, y는 종속변수를 의미한다. $L(y,F(x))$는 미분가능한 손실함수(Loss Function) 이며 식(6)에서는 유사잔차(Pseudo-residuals)를 m번 반복한다.

$$F_0(\chi)=argmin\,\gamma\sum_{i=1}^{n}L(y_i,\gamma) \quad \cdots (4)$$

$$\gamma_{im}=-\left[\frac{\delta L(y_i,F(x_i))}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \cdots (5)$$

이후 계산된 유사잔차에 대하여 기본 학습자인 $h_m(x)$를 적합한 후 잔차를 m번 업데이트 한다. 아래의 식(7)과 식(8)에 표현되어 있다.

$$\gamma_m=argmin\,\gamma\sum_{i=1}^{n}L(y_i,F_{m-1}(x_i))+\gamma h_m(x_i) \quad \cdots (6)$$

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x) \quad \cdots (7)$$

위의 RF 와 똑같이 Relaxed Search를 이용해서 최적의 파라미터 값들을 찾았다. XGBRegressor에

쓰일 파라미터는 사용 변수의 비율을 나타내는 colsample_bytree을 0.2,0.4,0.6, 학습율을 나타내는 learning_rate으로 0.1,0.2,0.3,0.4, 나무의 깊이를 나타내는 max_depth은 5,7,9,11, 그리고 나무의 숫자는 50,75,100,250,500으로 총 240가지 경우의 수를 실행했다. 이 중에 RMSE와 MAPE를 이용해 오차를 구하고 최적의 모델은 과적합이 가장 안 된 모델 5개를 선정했다.

3.3.4 Support vector regressor

기계학습 기법인 SVR은 분류 알고리즘의 하나인 SVM (Support Vector Machine)을 일반화한 기법으로, 서포트 벡터간의 거리내에 최대한 많은 데이터를 포함하게 하는 최적의 초평면을 찾아 데이터를 예측하는 기법이며 수요예측뿐만 아니라 기상, 금융 등 다양한 예측분야에서 활용되고 있다.

주어진 자료를 $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n}$로 표현하기로 하고, 이때의 $\boldsymbol{x_i}$는 독립 변수 벡터, y_i는 종속 변수 값이다. SVR을 학습시키기 위한 기본 모델식은 식 (8)과 같고, $\boldsymbol{x_i}$로부터 y_i를 가장 잘 예측하는 함수 $f(x)$를 추정하는 것을 목적으로 한다.

$$f(x) = \boldsymbol{w}^T \cdot \boldsymbol{x} + b, \;\; w \in R^m, b \in R \quad \cdots (8)$$

과적합 방지를 위해 소프트마진 형식의 SVR을 구현하기 위해 오류를 허용하는 변수인 ξ, ξ^*를 추가하여 학습시키고, 실측 데이터와 예측 데이터의 편차가 ε 이내에 들어오면서 w함수를 최소화하게 한다. (오병찬, 2019)

$$argmax \; \frac{1}{2} w^T \boldsymbol{w} + C \sum (\xi_i + \xi^*{}_i) \quad \cdots (9)$$

$$s.t. \quad y_i - (w^T \boldsymbol{x}_i + b) \leq \varepsilon + \xi_i \quad \cdots (10)$$

$$-y_i + (w^T \boldsymbol{x}_i + b) \leq \varepsilon + \xi^*{}_i \quad \cdots (11)$$

$$C, \xi_i, \xi^*{}_i \geq 0 \quad for\ all\ i \quad \cdots (12)$$

위의 선형식을 모든 $\{(\boldsymbol{x_i}, y_i)\}$가 만족해야 하며, 매개변수 C를 통해 과적합을 조절할 수 있다. C 값이 작으면 오분류를 허용하고, 또 C값이 크면 오차를 작게 학습한다.

식 (12)를 라그랑주 승수법을 사용하여 해결할 수 있고, 식(12)을 dual로 바꾸면 식(13)과 같다.

$$L = C \sum (\xi_i + \xi^*{}_i) + \frac{1}{2} ||\boldsymbol{w}||^2 - \sum (\eta_i \xi_i + (\eta_i{}^* \xi_i{}^*) - \sum \alpha_i (\varepsilon + \xi_i - y_i + w x_i + b) - \sum \beta_i (\varepsilon + \xi^*{}_i + y_i - w x_i - b) \quad \cdots (13)$$

이때 식(13)을 각 변수에 대해 편미분하여 식(13)을 변형하면 식(14)와 같다.

$$argmax \; -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha^*{}_i)(\alpha_j - \alpha^*{}_j) x^{T_i} x_j + \sum y_i (\alpha_i - \alpha^*{}_i) - \varepsilon \sum (\alpha_i + \alpha^*{}_i) \quad \cdots (14)$$

$$s.t.\quad 0 \le \alpha_i, \alpha^*{}_i \le C,\quad \sum(\alpha_i - \alpha^*{}_i) = 0 \qquad \cdots (14)$$

이를 이용하여 최종 예측 모델을 구하면 식(15)와 같다.

$$f(x) = \sum(\alpha_i - \alpha^*{}_i)x^T{}_i x_i + b\ , b == y_i - \varepsilon\sum(\alpha_i + \alpha^*{}_i)x^T{}_i x_i \qquad \cdots (15)$$

SVR이 선형예측 기법이기 때문에 $\Phi: x \to \varphi(x)$를 통해 고차원 공간으로 데이터 포인트들을 이동시켜야 한다. 우리가 이번 연구에 사용한 커널은 RBF(=Gaussian)커널이고 커널식은 식(16)과 같다.

$$K(x_i, x_j) = \exp\left(-\frac{\left||x_i - x_j|\right|^2}{2\sigma^2}\right) \qquad \cdots (16)$$

오피스텔 가격 모형의 알고리즘 간의 성능 비교를 위해 다중회귀분석, 랜덤 포레스트, 서포트 벡터 머신 세가지의 회귀 방식을 사용하여 RMSE(수식 추가) 값을 비교하였다. 각 회귀 모형에 적합을 위해 트레이닝 셋과 테스트 셋의 비율을 바꾸어 가며 어떤 비율에서 가장 최적의 값이 나오는지 찾았다. (오병찬, 2019)

3.3.5 평가지표

오차는 식(17) 2 에서의 평균제곱근오차(RMSE, Root Mean Square Error)와 식(18) 3 평균 절대 백분율 오차(MAPE, Mean Absolute Percentage Error)을 관찰했다. 이때 훈련 데이터(train set)에 비해 실험 데이터(test set)의 RMSE 또는 MAPE 가 너무 높게 나오면 과적합(Overfitting) 이 된 것인데 과적합 정도는 식(19) 4의 overfit 수식으로 알아봤다. 이 수치가 작을수록 훈련 데이터의 RMSE 와 실험 데이터의 RMSE 간의 차이가 크지 않음을 얘기한다. 여기서 y_i 은 각각의 예측값이며 $\hat{y_i}$ 은 전체 평균값이다. 그리고 $y_{train,i}$, $\widehat{y_{train,i}}$은 각각 훈련 셋에서의 예측값과 평균값이며, $y_{test,i}$, $\widehat{y_{test,i}}$은 각각 실험 셋에서의 예측값과 평균값이다.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y_i})^2}{n}} \qquad \cdots (17)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \hat{y_i})}{y_i}\right| \qquad \cdots (18)$$

$$overfit = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left(\frac{(y_{test,i} - \widehat{y_{test,i}})}{y_{test,i}} - \frac{(y_{train,i} - \widehat{y_{train,i}})}{y_{train,i}}\right)}{\frac{(y_{test,i} - \widehat{y_{test,i}})}{y_{test,i}}} \qquad \cdots (19)$$

4. 결과

본 연구에서는 랜덤 포레스트, 서포트 벡터 머신, 다중선형회귀, XGBoost를 통해 3가지 모델을 만들었다. 하이퍼 파라미터가 필요한 랜덤 포레스트, 서포트 벡터 머신, XGBoost는 하이퍼 파라미

터 후보를 임의로 설정하고 모든 조합을 만들어 본 뒤 MAPE를 최소로 하는 파라미터 셋을 찾았다. 랜덤 포레스트는 n_estimators를 5개, 10개, 50개, 100개, 250개, 500개 ,750개 총 7가지, max_depth를 1, 3, 5, 7 총 4가지로 설정해 총 28가지의 조합 중 최적의 파라미터 셋을 찾았다. 서포트 벡터 머신은 이번 연구에서는 방사기저함수(Radical Basis Function)가 사용되어, gamma를 0.00005, 0.0001, 0.000125, 0.00015, 0.000175, 0.0002 총 4가지, C값을 5000, 10000, 12500, 15000, 75000 총 5가지로 설정해 총 20가지의 조합 중 최적의 하이퍼 파라미터 셋을 찾았다. XGBoost는 colsample_bytree를 0.2, 0.4, 0.6, learning_rate를 0.1, 0.2, 0.3, 0.4, max_depth를 5, 7, 9, 11, n_estimators를 50, 75, 100, 250, 500로 설정한 뒤 최적의 하이퍼 파라미터 셋을 찾았다. 모델을 만든 뒤 잔차분석을 통하여 회귀모형에 대한 가정들의 충족 여부를 검토하였다.

또한, 이에 그치지 않고 전체 데이터 셋을 60m^2 이하는 소형, 60m^2초과는 중대형으로 분류하여 위에 기술한 방법을 이용해 분석을 한 번 더 반복하였다. 최종적으로 랜덤 포레스트, 서포트 벡터 머신, 선형회귀, XGBoost 네가지 모델 간의 성능 비교를 통해 어떤 모델이 가장 적합한지 결정하였다. 더불어 독립 변수들의 중요도를 파악하였다.

4.1 일산과 분당 각각의 전체 데이터 셋

4.1.1. 모델별 결과 분석

Ilsan				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	4093.8	2507.1	2830.05	4400.65
RMSE Mean(test)	4063.35	3194.4	3375.55	4518.15
RMSE STD(train)	100.21	59.32	97.7	104.35
RMSE STD(test)	259.79	257.36	534.17	407.93
MAPE Mean(train)	14.52	8.36	5.12	13.1
MAPE Mean(test)	14.7	9.33	6.76	13.466
MAPE STD(train)	0.38	0.13	0.07	0.54
MAPE STD(test)	0.8	0.42	0.58	1.04
Bundang				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	7797.35	2778.4	4562.8	7907.35

RMSE Mean(test)	8918.25	5113.7	4370	7458.45
RMSE STD(train)	285.56	88.3	453.54	257.29
RMSE STD(test)	1345.73	1632.9	1817.89	984.63
MAPE Mean(train)	16.35	5.87	4.02	16.549
MAPE Mean(test)	16.46	7.22	4.9	16.399
MAPE STD(train)	0.24	0.13	0.04	0.17
MAPE STD(test)	0.51	0.24	0.2	0.55

Table 4-1. RMSE and MAPE values for each algorithm and region with full data.

test_size는 0.2로 고정했으며 하이퍼파라미터 전체 조합 중에 과적합이 가장 적게 일어난 순서대로 5가지 조합을 뽑은 뒤, 20번 반복 시행해서 RMSE, MAPE의 평균과 표준편차를 구했다. 그리고 최적의 파라미터는 아래와 같이 나타났다.

	Ilsan	Bundang
Random Forest – (n_estimators, max_depth)	(5,3)	(10,3)
XGboost – (colsample_bytree, learning_rate, max_depth, n_estimators)	(0.4,0.1,5,50)	(0.2, 0.3, 5, 75)
Support Vector Machine – (gamma, C)	(0.0001,75000)	(0.000175,75000)

Table 4-2. Best Parameters for each algorithm without linear regression.

오피스텔 매매가의 변동이 크기 때문에 모델의 예측력의 판단 척도로 MAPE를 선택했다. 모델 간의 차이가 유의미하게 나타나는지 알아보기 위해 추가 검정을 시행하였다. 20회를 반복한 MAPE의 값의 정규성검정을 했을 때, 정규분포를 따르지 않는 것을 알 수 있었다. Wilcoxon test를 진행했을 때, 분당, 일산 모두에서 유의미하게 SVM 모델의 MAPE가 가장 낮았고, 즉 가장 좋은 예측력을 보여주었다. 결론적으로 전체 데이터 셋에 대해서 가장 적합한 모델은 SVM 모델임을 알 수 있다.

4.1.2. 변수중요도

	Ilsan	Bundang
Y-intercept	0.000201	0.0000115
Area for exclusive use	0.000491	4208.7283
Floor	5860.6412	4193.895

Highest floor	0.000419	-5059.8589
Total number of households	-2332.375	9745.909
Floor area ratio (%)	-4970.5247	-0.000187
Building-to-land ratio (%)	-3841.471	-0.000133
Walking distance to subway (km)	9621.4426	0.00016
Walking distance to hospital (km)	-1338.1174	-3831.7446
Walking distance to market (km)	-1796.8129	-0.000238
Years passed	-0.000139	5734.4858
Months passed	776.3722	0.0000115

Table 4-3. Standardized Regression Coefficients for each region

Unit: %

	Ilsan	Bundang
Area for exclusive use	87.79	89.92
Floor	0.08	0.68
Highest floor	0	3.88
Total number of households	0	0
Floor area ratio (%)	0	0
Building-to-land ratio (%)	0.05	0
Walking distance to subway (km)	0.05	0
Walking distance to hospital (km)	0.39	3.84
Walking distance to market (km)	0	0.17
Years passed	10.46	1.48
Months passed	1.13	0

Table 4-4. Random Forest Feature Importance

Unit: Fscore

	Ilsan	Bundang
Area for exclusive use	203	208

Floor	248	197
Highest floor	74	118
Total number of households	181	172
Floor area ratio (%)	155	142
Building-to-land ratio (%)	99	98
Walking distance to subway (km)	124	118
Walking distance to hospital (km)	211	156
Walking distance to market (km)	178	151
Years passed	45	56
Months passed	108	100

Table 4-5. XGBoost Feature Importance

다중선형회귀, 랜덤포레스트, XGBoost 알고리즘으로 회귀모형을 적합하고 각 모형별 변수 중요도가 높은 순서대로 정리해 다음과 같은 결과를 얻었다. 다중선형회귀모형은 P-value값이 낮은 순서대로 정리하였고, 랜덤 포레스트는 어떤 변수를 사용한 노드가 (전체 트리에 대해)평균적으로 불순도를 얼마나 감소시키는지 확인하여 중요도를 측정했다. XGBoost의 변수 중요도 단위는 F score로 Tree를 분할할 시 얼마나 해당 feature가 자주 사용되었는가를 나타내는 지표이다. F score가 높을 수록 Tree 분할에 더 잘 사용된다.

다중선형회귀와 랜덤포레스트의 변수중요도를 보면 소수의 변수에 중요도가 편향되어 있다. 따라서 비교적 좋은 예측력을 보여준 XGBoost의 변수 중요도를 중점적으로 보았다. 분당과 일산에서 공통적으로 전용면적이 중요변수로 작용했다. 분당에서는 일산과 비교해 최고층의 중요성이 높고 일산에서는 비교적 대형병원과의 거리, 그리고 해당 오피스텔 층의 중요성이 높다.

다중선형회귀 변수중요도를 살펴봤을 때, 분당에서는 용적률, 경과월수, 총세대수 순으로 중요도가 높았고, 일산에서는 지하철부터의 거리, 해당 오피스텔 층, 용적률 순으로 중요도가 높았다. 랜덤포레스트 변수중요도를 보면, 분당에서는 전용면적, 최고층, 대형병원으로부터의 거리 순서로 중요도가 높음을 알 수 있고 일산에서는 전용면적, 경과년수, 경과월수 순서로 중요도가 높다. XGBoost 변수중요도를 보면, 전용면적, 해당오피스텔 층, 총 세대수 순으로 변수 중요도가 높으며 일산에서는 해당 오피스텔 층, 대형병원부터의 거리, 전용면적 순으로 변수 중요도가 높음을 알 수 있다.

	Linear Regression	Random Forest	XGBoost
Ilsan	지하철부터 거리	전용면적	층
	층	경과년수	대형병원부터 거리
	용적률	경과월수	전용면적
Bundang	용적률	전용면적	전용면적
	경과월수	최고층	층
	총세대수	대형병원부터 거리	총 세대수

Table 4-6. Final important variables for each region with full data

4.2. 분당&일산 소형&중대형 셋

4.2.1. 모델별 결과 분석

Ilsan small				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	1793.75	840.05	1204.35	1609.2
RMSE Mean(test)	1936.00	1225	1423.4	1729.35
RMSE STD(train)	27.65	12.84	74.77	50.64
RMSE STD(test)	208.41	251.01	328.22	332.1
MAPE Mean(train)	11.39	4.91	5.23	9.04
MAPE Mean(test)	11.82	6.04	6.13	9.6
MAPE STD(train)	0.18	0.06	0.06	0.19
MAPE STD(test)	0.49	0.22	0.32	0.95
Bundang small				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	3213.60	1125.25	1948.85	3406.35
RMSE Mean(test)	3271.10	1530.85	1976.35	3514.7
RMSE STD(train)	56.10	27.88	34.27	28.95
RMSE STD(test)	142.63	112.1	194.6	111.45
MAPE Mean(train)	11.34	3.56	4.51	11.92

MAPE Mean(test)	11.46	4.53	4.89	12.21
MAPE STD(train)	0.18	0.05	0.03	0.08
MAPE STD(test)	0.43	0.16	0.22	0.44
Ilsan large				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	10329.05	4660.7	5333	6212.35
RMSE Mean(test)	10398.50	6286.6	5765.7	6291.75
RMSE STD(train)	168.02	138.02	168.25	101.95
RMSE STD(test)	535.75	795.57	860.62	414.84
MAPE Mean(train)	24.58	8.44	6.45	11.45
MAPE Mean(test)	25.25	10.48	7.96	11.67
MAPE STD(train)	0.37	0.14	0.13	0.22
MAPE STD(test)	1.71	0.53	0.66	0.69
Bundang large				
	Random Forest	XGB	SVM	Linear Regression
RMSE Mean(train)	11220.15	6057.15	11023.05	11727.7
RMSE Mean(test)	14523.95	10644.05	10364.4	11529.45
RMSE STD(train)	562.85	204.85	1424.84	720.59
RMSE STD(test)	2894.17	3826.03	4441.61	2474.33
MAPE Mean(train)	15.42	6.4	5.96	13.3
MAPE Mean(test)	16.11	8.28	7.08	13.59
MAPE STD(train)	0.33	0.14	0.13	0.28
MAPE STD(test)	1.32	0.55	0.6	0.84

Table 4-7. RMSE and MAPE values for each algorithm and region with small and large data.

	Ilsan small	Bundang small	Ilsan large	Bundang large
Random Forest –	(50,3)	(750,3)	(10,1)	(750, 3)

(n_estimators, max_depth)				
XGboost – (colsample_bytree, learning_rate, max_depth, n_estimators)	(0.6, 0.1, 5, 50)	(0.4, 0.1, 5,100)	(0.2, 0.1, 5, 50)	(0.2, 0.1, 5, 75)
Support Vector Machine – (gamma, C)	(0.000125, 12500)	(0.000125, 15000)	(0.000175, 15000)	(0.0002, 10000)

Table 4-8. Best Parameters for each algorithm without linear regression.

분당, 일산 소형, 중대형 모든 데이터 셋에서 SVM과 XGB가 가장 낮은 MAPE를 보여주고 있다. 차이점은 SVM의 결과가 훈련 셋과 테스트 셋에서의 결과값 차이가 상대적으로 적다는 점이다. 결과적으로 SVM과 XGB 중 SVM이 상대적으로 안정적인 결과를 낸다고 판단하였다.

4.2.2. 변수중요도

	Ilsan small	Bundang small	Ilsan large	Bundang large
Y-intercept	0.000163	0.000228	0.000338	0.000497
Area for exclusive use	0.000109	0.000264	0.00038	0.000784
Floor	2570.7331	1691.9356	2553.1651	9740.025
Highest floor	4982.9245	1878.2844	0.000178	0.000143
Total number of households	316.2037	-2730.847	-9408.85	-0.000128
Floor area ratio (%)	-1019.651	4492.8997	1029.7699	0.000167
Building-to-land ratio (%)	-1843.21	-6589.796	4278.8726	-0.000254
Walking distance to subway (km)	3027.9855	-0.000155	6476.1775	-1905.869
Walking distance to hospital (km)	-1626.46	3297.9057	0.000109	0.000366
Walking distance to market (km)	76.5526	5342.5751	7768.6545	-0.000335
Years passed	-0.0001	-0.00024	-0.000367	-0.000333
Months passed	-1808.179	2858.7257	6680.2324	0.000183

Table 4-9. Standardized Regression Coefficients for small and large regions

Unit: %

	Ilsan small	Bundang small	Ilsan large	Bundang large
Area for exclusive use	59.67	43.81	0	65.12

Floor	0.09	0	0	4.52
Highest floor	2.55	6.01	30	3
Total number of households	0.03	0.01	0	0.24
Floor area ratio (%)	0.57	0.05	0	15.86
Building-to-land ratio (%)	5.83	0	0	0.21
Walking distance to subway (km)	0.61	0.01	0	0.07
Walking distance to hospital (km)	0.03	0.25	0	0.73
Walking distance to market (km)	0.39	24.95	0	0.76
Years passed	30.13	24.84	70	6.6
Months passed	0.05	0.02	0	2.84

Table 4-10. Random Forest Feature Importance

Unit: Fscore

	Ilsan small	Bundang small	Ilsan large	Bundang large
Area for exclusive use	366	371	67	151
Floor	237	328	154	173
Highest floor	65	198	41	82
Total number of households	215	255	46	141
Floor area ratio (%)	241	254	81	105
Building-to-land ratio (%)	155	134	62	98
Walking distance to subway (km)	170	183	52	93
Walking distance to hospital (km)	184	158	131	121
Walking distance to market (km)	125	156	112	119
Years passed	128	159	13	37
Months passed	248	277	58	88

Table 4-11. XGBoost Feature Importance

소형, 중대형으로 나누어 분석을 진행했을 때, 전체 데이터 셋에서의 결과와 비교해서 여러 변수에 중요도가 분배되어진 측면이 있지만 여전히 소수의 변수의 집중되어있다. XGBoost에서의 변

수중요도를 보면 소형에서 변수들의 Fscore 수치가 중대형과 비교해 전체적으로 높다. 소형에서 각각의 변수들이 tree 분할에 더 많이 사용되었음을 알 수 있다. 추가적으로 규모별로 비교하기 위해 XGBoost 범주별 변수 중요도 표를 아래와 같이 만들었다.

	Small		Large	
Area for exclusive use	737	16%	218	11%
Floor	565	12%	327	16%
Highest floor	263	6%	123	6%
Total number of households	470	10%	187	9%
Floor area ratio (%)	495	11%	186	9%
Building-to-land ratio (%)	289	6%	160	8%
Walking distance to subway (km)	353	8%	145	7%
Walking distance to hospital (km)	342	7%	252	12%
Walking distance to market (km)	281	6%	231	11%
Years passed	287	6%	50	2%
Months passed	525	11%	146	7%

Table 4-12. Categorical XGBoost Feature Importance

소형에서는 전용면적, 해당 오피스텔 층, 경과월수 순으로 영향력이 크다. 중대형에서는 해당 오피스텔 층, 대형병원까지의 최소거리, 대형마트까지의 최소거리 순으로 영향력이 크다. 공통적으로 해당 오피스텔 층이 중요한 변수임을 알 수 있다. 소형에서는 오피스텔 건물 자체의 측면이 부각되었다면 중대형에서는 오피스텔의 외적인 요소, 편의시설 접근성과 같은 측면이 중요하게 작용했다.

5. 결론 및 향후 연구 방향

위 연구는 다양한 변수와 분당, 일산 오피스텔 매매가의 변수 중요도 관계를 알아보고 여러 모델 중 적합한 모델을 찾아보고자 2018년~2020년 기간에 걸친 분당, 일산의 오피스텔 관련 자료를 활용하였다. 본 연구에서는 선행연구와 이전보다 세분화한 오피스텔 자료를 활용함으로써 차

별화된 연구결과를 얻을 수 있었다. 분석을 위해 오피스텔 매매가를 종속변수 y로, 전용면적, 층, 경과년수, 지하철로부터의 최소거리, 대형마트로부터의 최소거리, 대형병원으로부터의 최소거리, 최고층, 총세대수, 용적률, 그리고 건폐율을 변수로 고려하였다. 분석에서 도출된 주요결과를 요약하면 다음과 같다.

MAPE값과 과적합의 정도를 통해 각 모델 별 예측력을 판단할 수 있었다. 다중회귀분석, 랜덤포레스트, SVM, XGBoost 네 개의 모델 중 다중회귀분석과 랜덤포레스트가 비슷한 정도로 비교적 MAPE값이 높아서 좋지 않은 결과를 나타내었다. SVM과 XGBoost 중에서 SVM은 과적합의 정도가 낮은 반면에 XGBoost에서는 비교적 높았다. 결론적으로 SVM이 MAPE값과 과적합 정도에서 가장 좋은 예측력을 보여주었다.

예측력이 가장 좋았던 SVM으로 변수중요도 분석을 하는 것이 맞으나, rbf커널의 특성상 변수중요도 파악이 어려워 예측력이 비슷했던 XGB로 변수중요도 분석을 진행하였다. 예측력이 좋지 않았던 랜덤포레스트와 다중회귀분석은 소수의 변수에 중요도가 치우쳐진 경향이 있었다. 따라서 XGBoost의 변수중요도를 중점적으로 분석하였다. 공통적으로 전용면적이 중요변수로 작용했다. 지역별로는 분당에서는 일산과 비교해 최고층의 중요성이 높고 일산에서는 비교적 대형병원과의 거리, 그리고 해당 오피스텔 층의 중요성이 높다. 규모별로는 소형에서는 오피스텔 건물 자체의 측면이 부각되었다면 중대형에서는 오피스텔의 외적인 요소, 편의시설 접근성과 같은 측면이 중요하게 작용했다.

앞으로 연구를 진행할 때 추가적으로 고려해야 할 사항으로는 더 많은 데이터와 더 다양한 변수를 사용하는 방법을 생각해볼 수 있다. 현재의 연구는 데이터에 대한 접근 제한 등 현실적인 제약들 때문에 3년 치의 데이터만을 사용했으며 사용한 변수 역시 오피스텔 매매가에 영향을 주는 모든 요소를 고려하지는 못했기에 예측력이 더 좋아질 여지가 있다. 따라서 더 많은 기간의 데이터와 시간에 따른 금리변화, 부동산 지수, 부동산 정책 등을 추가적으로 고려할 수 있다면 예측 정확도를 높일 수 있을 것이라고 생각된다.

본 논문에서는 오피스텔 매매가 분석에 적합한 모델과 중요한 변수를 분석하고, 이에 기반하여 효과적이고 효율적인 부동산 예측 모델 형성의 필요성을 피력하는데 목적이 있다. 향후 연구에서는 앞서 말한 한계점을 보완하여 보다 근본적인 오피스텔 매매가 예측 모델 형성을 위한 연구가 수행되는 것이 필요하다고 여겨진다.

6. 참고 문헌

Lee J. M., Park S. H., Choi S.H., Kim J. H. (2021), Comparison of Models to Forecast Real Estates Index Introducing Machine Learning , *Journal of the Architectural Institute of Korea,* 37(1), 191-199.

Na S. H., Kim J. W. (2019), A study on the sales price of apartment using public data : The apartment in Gangnam-gu Seoul*, Journal of the Korean Cadastre Information Association*, 21(1), 3-12.

Chun H. J., Prediction of Housing Price Using Time Series Analysis and Machine Learning Methods, *Residual Environment : Journal of the Residential Environment Institute of Korea*, 18(1), 49-65.

Chun H. J., Yang H. S., A Study on Prediction of Housing Price Using Deep Learning, *Residual Environment : Journal of the Residential Environment Institute of Korea* , 17(2), 37-49.

Oh B. C., Kim S. Y., Development of SVR based Short-term Load Forecasting Algorithm, The Transaction of the Korean Institute of Electrical Engineers, 68(2), 95-99.

Choi B. M., Kim J. H., An Empirical Analysis of the Dynamic Correlation between Officetel Supply and Apartment Price*, Journal of the Korean Regional Develipment Association*, 30(5), 55-74.

Kwon, M. J., Kim J. C., Forecasting Seoul Apartment Price Index based on a Deep Learning Model, *Korean Institute of Industrial Engineers*, 101-106.

Choi Y., Kim H.J., Yeo J. H., A Study on the Factors Determining Officetel Price in Busan, *Journal of the Korean Society of Civil Engineers*, 35(3), 725-735.

Abstract

The aim of this research is to discover factors that have an important affect on studio prices in Ilsan and Bundang and distinguish machine learning algorithms that have performed well in the process. When considering the fact that over 70% of the public's asset comes from real estates, studio prices are also an important issue. To find such factors regarding studio prices we have used four machine learning techniques, each being random forest, XGBoost, support vector machines and multiple linear regression. We also chose our research area to be Ilsan and Bundang because there are plenty of research regarding house pricing in the capital city Seoul. To yield better results, we divided the data into three groups according to total area. With various algorithms and datasets, we have produced meaningful results and singled out variables that have had most meaning in the pricing.

Keywords: Machine Learning, Random Forest, Studio Prices, Support Vector Machine, Total Area