부동산학연구 제26집 제2호, 2020. 6, pp. 23~53 http://dx.doi.org/10.19172/KREAA.26.2.2

Journal of the Korea Real Estate Analysts Association
Vol.26, No.2 2020. 6, pp. 23~53

머신 러닝 방법을 이용한 오피스 임대료 산정*

-랜덤 포레스트, 인공 신경망, 서포트 벡터 머신 활용을 중심으로-

A Study on the Office Rent Estimation by the Machine Learning Methods -Focusing on the Use of Random Forests, Artificial Neural Networks, Support Vector Machines-

정 성 훈 (Jeong, Sung-Hoon)**
진 창 하 (Jin, Changha)***

< Abstract >

We estimate office market rent using random forests, artificial neural networks, and support vector machines, which are commonly used in recent studies related to automatic evaluation models. We examine each method and compare the performances based on our sample data of the 507 offices located in Seoul, and depict PD(Partial Dependence) Plots to identify the influence of variables on predicted values. Using random forests, artificial neural networks, and support vector machines method, we estimate office market rent determinant model based on 507 office rental information data in Seoul. We attempt to identify the best performed model and also adopt the Partial Dependence(PD) Plots to examine the relative impact of research variables on predicted value. We classify the sample data into Class A, Class B, Class C, and the below Class C group. We apply the same method for each of those group. In general, we find that the support vector machines model performs best followed by random forests model and artificial neural networks model. The results indicate that the rent has a positive relation with rentable/usable ratio, floor, building age, and number of elevator. The number of parking vehicles showed the quadratic curve on rent in the model. Our study contributes to compare and allows appraisers to have a cross validation process with traditional valuation model against the finding from a machine learning method on office market.

Keyword : Office Rent Model, Machine Learning Methods, Random Forests, Artificial Neural Networks, Support Vector Machines

I. 서론

오피스 임대료는 업무 공간에 대한 가격으로, 오피스 운영 수익의 가장 큰 부분을 차지하기 때문에 이를 물건으로 한 리츠나 부동산 펀드의 기대수익률을 산정하는데 있어 중요한 정보이다. 즉 상업용 부동산 투자 분석 간 예상 현금흐름을 파악하기 위해 적정한 임대료

* 본 논문은 주저자의 석사 학위 논문을 바탕으로 작성되었음을 밝힙니다.
** 한양대학교 경상대학 응용경제학과 석사과정, jshoon@hanyang.ac.kr, 주저자
*** 본 학회 정회원, 한양대학교 경상대학 경제학부 부교수, cjin@hanyang.ac.kr, 교신저자

를 산정해야하는데, 전통적인 임대료 산정 방법으로 주로 임대사례비교법과 적산법이 사용되어 왔다. 하지만 이러한 방법들에는 몇 가지 문제가 존재하는데, 먼저 임대사례비교법은 보정치를 설정할 때 평가자의 주관이 포함되므로 평가자들마다 매우 큰 편차를 나타낼 수 있어 객관성을 유지하기 어렵다는 점과 상업용 부동산 특성 상 건물주들이 물건의 임대 정보를 공개하기 꺼려하기 때문에 평가자가 대상 물건의 주변 임대사례 정보를 수집하는 데 어려움이 있다는 문제가 존재한다. 적산법은 기초가액으로서 시장가치를 산정하는 과정에서 뿐만 아니라 기대이율을 설정하는 과정에서 평가자의 주관이 포함될 수 있으므로 이 또한 객관적인 방법이라고 보기 힘들다는 문제가 존재한다. 이렇듯 기존의 방법들을 통해 산정된 임대료에는 평가자의 주관적인 부분이 다소 포함되기 때문에 이와 상호 보완할 객관적으로 산정된 임대료가 필요하다.

축적된 데이터를 이용하여 객관적이고 효율적으로 부동산 가치를 평가하는 자동 평가 모형(automated valuation model, AVM)에 대해 지금까지 많은 연구들이 이루어져왔고, 독일과 스웨덴 등 여러 유럽 국가들은 가치 평가 프로세스와 관련한 AVM 법률을 도입하여 가치 평가 작업에 해당 모형을 활용하는 것을 법적으로 인정하고 있다. 자동 평가 모형은 부동산의 시장 가치나 시장 임대료 산정을 위해 개발된 수학적 모형(mathematical model)으로, 대부분 다중 회귀 분석(multiple regression analysis)이나 인공 신경망(artificial neural networks)이 적용되고 있다(Johnson, 2019). 하지만 전통적인 통계 기법인 다중 회귀 분석에는 갖가지 엄격한 가정들을 한다는 점, 이상치(outlier)에 민감하게 반응한다는 점, 비선형적 관계를 고려하기 힘들다는 점, 최적 회귀식을 선택하였는지에 대한 의문점이 제기될 수 있다는 점(김태훈 · 홍한국, 2004), 그리고 데이터 수에 민감하다는 점(김선주 · 이상엽, 2008) 등 많은 문제들이 지적되어 왔다. 이에 최근에는 기존 AVM에서 활용하던 인공 신경망 뿐만 아니라, 랜덤 포레스트(random forests, RF)나 서포트 벡터 머신(support vector machines, SVM)과 같은 다양한 머신 러닝(machine learning) 방법을 적용하여 주거용 부동산의 가격 산정이나 예측을 시도한 연구들이 활발히 등장하고 있다. 머신 러닝은 회귀분석과 같이 데이터 속에서 어떤 관계를 찾거나 의미있는 정보를 밝혀내는 데이터 마이닝(data mining)에 접목이 가능하고, 특히 다중 회귀 분석의 여러 문제로부터 자유롭기 때문에 오피스 시장 자료에 적용하기 적합하다고 판단되며 가치평가나 임대료 산정 등과 같은 연구에서 이들의 실제 활용 가능성을 검토해볼 필요가 있다.

최근 서울 오피스 시장에서는 권역을 벗어나는 임차인의 이동이 늘고 있어 시장 분석 시 권역보다는 등급이나 투자스타일을 기준으로 한 하위시장 관점의 분석이 더욱 중요해지고 있다. 이에 등급별 하위시장의 특성 분석과 관련한 연구들이 속속히 등장해왔으며, 현재 많은 부동산 투자 컨설팅 업체들이 권역별 시장 분석뿐만 아니라 등급별 시장 분석을 수행하고 있다. 또한 오피스 등급은 임대료 산정에 주요한 기준이 되기도 하고 투자자로 하여금 투자 판단을 위한 기초자료로 활용되어 오고 있다(장상규 · 김민철, 2018). 현재 국내에는 오피스 등급을 설정하는 통합적인 세부 지표가 존재하지 않아 각 컨설팅 업체마다 등급을 심사하는 기준이 다소 상이하지만(윤여신, 2012), 동일 권역 내에서도 건물의 규모에 따라 시장 특성이 세분화되는 모습을 보이고 있어(이재우, 2005) 컨설팅 업체들은 공통적으로 건물 규모를 주요 요소로 고려하여 오피스의 등급을 구분하고 있다. 따라서 오피스 가치 산정이나 시장 분석과 관련한 연구에서 규모를 기준으로 한 등급별 분석의 필요성 제시 등과 같은 하위시장에 대한 논의가 필요한 시점이라고 판단된다.

본 연구에서는 오피스 임대료 산정 모형과 관련하여 머신 러닝 방법의 적용 가능성을 검토하고자 하며, 이에 다음과 같은 세 가지 목적을 갖는다.

첫째, 객관적인 적정 임대료를 산정할 수 있는 최적의 머신 러닝 모형을 구축하고자 한다. 머신 러닝 모형은 다중 회귀 모형의 여러 문제점으로부터 자유롭고 예측 중심 모형으로서 객관적이고 효율적인 임대료 산정을 가능케 하기 때문이다. 본 연구는 다양한 머신 러닝 방법 중 앙상블 기법[1]인 랜덤 포레스트(random forests, RF) 방법, 인공 신경망(artificial neural networks, ANN), 그리고 서포트 벡터 머신(support vector machines, SVM)을 적용하고자 한다. 해당 알고리즘들은 사회과학 분야에 적용된 머신 러닝 방법들 중 가장 보편적이며 다양한 선행

1) 주어진 데이터로부터 여러 예측 모형들을 생성하고 이들의 예측 결과들을 종합함으로써 하나의 최종 예측치를 도출해내는 방법으로, 배깅(bagging), 부스팅(boosting), 랜덤 포레스트가 대표적이다.

연구들로부터 검증이 이루어진 방법들이다. 본 연구는 각 방법들의 초모수(hyper-parameter)를 다양하게 조정하며 일반화 오차가 가장 작은 최적 모형을 도출하고자 한다.

둘째, 각 머신 러닝 방법별 최적 모형의 성능(performance)을 비교하여 임대료 산정에 적용 가능성을 검토하고자 한다. 이때 성능이란 새로운 데이터에 대한 모형의 예측력(predictive power)을 의미하는데, 본 연구는 훈련 데이터 셋(train data set)과 검정 데이터 셋(test data set)을 일정한 비율로 나누어 훈련 데이터 셋으로 모형을 형성한 후, 검정 데이터 셋에 대한 모형의 평균제곱근오차(root mean square error, RMSE)와 평균절댓값오차(mean absolute error, MAE)를 도출하여 각 최적 모형의 성능을 비교하고자 한다.

셋째, 각 설명변수들과 산정 임대료 간의 관계를 시각화하여 제시하고자 한다. 머신 러닝 모형들은 알고리즘과 모형 자체의 복잡성 때문에 이해하기 어렵고 해석이 힘들다는 측면에서 블랙박스 모형(black-box model)으로 여겨져 기존의 관련 연구들에서 대부분 예측 중심의 모형으로 활용되어 왔다. 하지만 특정 투입변수(input variable)가 모형의 예측값과 어떠한 관계를 갖고 있는지 확인할 수 있는 부분 의존성 플롯(partial dependence plot, PD plot)과 같이 머신 러닝 모형을 해석할 수 있는 방법이 존재한다. 이에 본 연구는 각 설명변수들에 대한 PD plot을 확인하여 모형의 학습 결과를 해석하고자 한다.

마지막으로, 규모 등급별 최적의 임대료 산정 모형을 각각 구축하여 등급별로 서로 다른 임대료 산정 모형 구축의 필요성을 제시하고자 한다.

II. 선행연구

1. 머신 러닝 방법을 이용한 주거용 부동산 가치 산정 관련 연구

머신 러닝 방법을 이용하여 부동산의 가치 산정을 시도한 연구는 국내외적으로 다양하게 존재한다. 먼저, 이와 관련한 해외 연구들은 대부분 주거용 부동산 대량 감정(mass appraisal)을 위한 자동평가모형 구축에 있어 머신 러닝 방법의 적용 가능성을 확인하고자 하였다는 공통점을 가지고 있다.

Peterson and Flanagan(2009)은 부동산 가치 산정에 있어 헤도닉 모형(hedonic model)에 대한 OLS (ordinary least square) 방법은 제1종 오류(type 1 error), 제2종 오류(type 2 error), 표본 오차(sampling error), 특정화의 오류(specification error)에 노출되어 있다는 점을 지적하고 이를 대체할 방법으로 인공신경망 분석을 제시하였다. 해당 연구는 1999년부터 20005년까지 웨이크 카운티에서 발생한 주거용 부동산 거래 자료를 활용하여 다중 회귀 모형과 인공 신경망 모형의 설명력 및 예측력 차이를 비교하였고, 모형 간 가격 오차(pricing error)의 차이를 연도 간 비교하여 인공 신경망 모형의 유용성(usefulness)이 증가하고 있음을 나타냈다.

Kontrimas and Verikas(2011)는 이전서부터 선형 회귀 모형과 신경망 모형 간 예측력을 비교하고자 한 다양한 연구들이 이루어져왔지만 부동산 가치 산정 작업에 신경망을 적용하는 것에는 상당히 조심스러워야 한다고 주장하였다. 해당 연구는 리투아니아의 한 도시에서 발생한 주택 거래건 자료를 활용하여 다중 회귀 모형, 신경망 모형, 서포트 벡터 머신, 그리고 모형 통합(committees of models) 방법의 예측력을 비교하였고, 인공신경망과 다중 회귀 모형의 예측력이 다소 떨어진다는 결과를 제시하였다.

Kok et al.(2017)은 부동산 가치 평가에 상당한 비용이 소모되고 있다는 점과 평가액과 실제 거래액 간의 괴리가 다소 크다는 문제를 제기하였고, 이러한 비효율성(inefficiency)와 부정확성(imprecision)을 해결할 방법으로 머신 러닝 방법을 적용한 자동 평가 모형을 제안하였다. 해당 연구는 2011년부터 2016년 까지 캘리포니아, 플로리다, 텍사스 소재의 다가구(multifamily) 주거용 부동산의 운영 소득(net operating income, NOI) 및 거래 자료를 활용하여 OLS, 랜덤 포레스트, 그래디언트 부스티드 트리(gradient boosted tress), XGBoost(extreme gradient boosting) 방법의 설명력 및 예측력 차이를 비교하였고, 머신 러닝 모형들이 OLS 모형보다 예측력이 더 뛰어난 것으로 나타났다.

부동산 가치 평가에 머신 러닝 방법의 적용 가능성을 확인하고자 한 국내 연구들을 살펴보면, 우선 인공신경망을 활용한 연구들이 비교적 이전에 등장하였으며, 최근에 들어 다양한 머신 러닝 방법을 적용하여 모형의 성능을 비교한 연구들이 등장하고 있다.

김태훈 · 홍한국(2004)은 기존의 회귀 모형에는 엄격한 가정들이 필요하다는 점과 이상치에 민감하게 반응한다는 점, 그리고 모수의 비선형을 고려하기 힘들다는 점 등 여러 문제점들을 지적하고, 이와 상호 보완할 모형으로 인공 신경망 모형을 제시하였다. 해당 연구는 2004년 6월의 서울시 송파구와 도봉구 소재 아파트 매매 자료를 활용하여 다중 회귀 모형과 인공 신경망 모형의 예측력을 비교하였고, 인공 신경망의 예측 성능이 더 우수하게 나타났다. 하지만 인공 신경망 모형을 과용하기보다는 두 모형의 추정 가격을 적절하게 결합하여 사용하거나, 인공 신경망 모형은 해석하기 어렵다는 단점이 있으므로 인공 신경망 모형에 근거하여 가격을 추정하였을 시 회귀 모형을 이용하여 설명하는 등 두 모형의 상호 보완적인 측면을 고려해야할 것을 당부하였다.

이창로 · 박기호(2016)는 기존 사회과학 연구들이 분석의 주요 초점을 이론 또는 가설을 검증하거나 변수 간의 관계를 규명하고자 하는 것에 맞춰 설명 중심의 모형(exploratory modeling)을 설정해왔지만 설명력이 좋은 모형이 예측 성능 또한 좋은 것이 아니기 때문에 부동산의 가치를 추정할 때 모수 모형(parametric mode)을 사용하는 것은 다소 문제가 있다고 주장하였다. 이에 해당 연구는 다중 회귀 모형과 더불어, 일반가산모형(generalized additive model, GAM), 랜덤 포레스트, MARS(multivariate adaptive regression splines), 서포트 벡터 머신과 같이 머신 러닝 분야에서 고안된 비모수 모형(non-parametric model)들의 적용 가능성을 검토하고자 하였고, 2011년부터 2014년까지의 강남구 소재 단독주택의 매매건 자료를 이용하여 각 방법 간 예측력 차이를 비교한 결과 모든 머신 러닝 모형이 다중 회귀 모형보다 예측력이 뛰어난 것으로 확인되었다.

오지훈 · 김정섭(2018)은 선형 회귀 모형을 기반으로 하는 헤도닉 가격 모형은 결과가 명료하고 해석이 용이하다는 장점이 있지만, 주택 가격과 특성 간의 비선형적 관계를 포착하기 어렵다는 점을 지적하였다. 이에 해당 연구는 비선형적 관계를 효과적으로 찾아낼 수 있는 머신 러닝 방법인 MARS(Multivariate Adaptive Regression Spline)를 활용하여 서울시 주택 가격 모형을 구축하였다. 이를 통해 가격에 대한 특성 변수들의 영향력이 달라지는 변곡점을 찾아 기존 연구들이 설정하였던 더미변수들의 적절성을 검토하였다. 특히 경과연수의 경우 21년, 39년을 변곡점으로 하는 3차 함수 형태의 가격 변화를 보여 헤도닉 가격 모형에서 경과연수 3차 다항식을 도입함으로써 모형의 설명력을 개선할 수 있을 것이라고 주장하였다.

배성완 · 유정석(2018)은 앙상블 방법이나 딥 러닝(deep learning)과 같은 새로운 방법들이 개발되고 있고 하드웨어 성능 또한 계속해서 발전하고 있기 때문에, 다량의 부동산의 과세 평가에 있어 정교하고 효율적인 부동산의 가격 산정 작업에 기계 학습의 적용 가능성이 매우 높다고 주장하였다. 이에 해당 연구는 2016년의 아파트 실거래가 자료를 활용하여 다중 회귀 분석, 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스티드 트리, 심층 신경망(deep neural networks, DNN) 각각의 모형 예측력을 비교하였고, 그 결과로 다중 회귀 분석이 성능이 가장 낮은 것으로 나타났으며 머신 러닝 방법들의 성능은 대체로 비슷한 수준인 것으로 나타났다. 또한 예측력이 우수한 것으로 나타난 머신 러닝 모형을 이용하여 실거래가반영률 분석 또한 추가적으로 수행하였다.

나성호 · 김종우(2019)는 주택 가격 모형 구축에 가장 많이 활용되어 온 다중 회귀 모형은 오차의 정규성, 독립성, 등분산성 등 여러 엄격한 가정들을 충족시키기 어렵고 머신 러닝에 비해 예측 성능이 떨어진다는 문제점이 있음을 지적하였다. 이에 해당 연구는 강남구 아파트의 실거래가 자료를 바탕으로 랜덤 포레스트 모형과 서포트 벡터 머신 모형을 구축하고 이들과 다중 회귀 모형의 예측 성능을 비교하였다. 그 결과 모형의 예측 성능은 랜덤 포레스트 모형, 서포트 벡터 머신 모형, 다중 회귀 모형 순으로 나타났으며 두 머신 러닝 모형의 성능이 다중 회귀 모형의 성능을 크게 앞서는 것으로 나타났다.

최근 머신 러닝을 통한 부동산 가치 산정에서 한 단계 더 나아가 이를 활용하여 부동산 가격 지수를 도출하고자 한 연구 또한 등장하였다. 배성완 · 유정석(2018)은 주택 가격 지수의 정확성을 위해서 표본 주택에 대한 정확한 가격 산정이 필수적이지만 기존에 활용하던 표본 주택 가격 자료에는 조사자의 성향, 경험 등에 의한 오류 또는 편의가 존재할 수 있다는 점을 지적하고 이러한 문제를 해결할 수 있는 방안으로써 머신 러닝의 활용 가능성에 주목하였다. 해당 연구는 서울 강남구 아파트 실거래가 자료를 바탕으로 랜덤 포레스트, 심층 신경망을 활용하여

주택 가격을 산정하고, 제본스 지수(Jevons index) 산정 방법을 통해 인공지능 가격에 기반한 부동산 가격 지수를 도출하였다. 분석 결과, 랜덤 포레스트 지수와 심층 신경망 지수는 시장 상황을 적절히 반영한다고 인정하기에는 어려움이 있지만, 기존의 가격 지수들에 비해 변동성이 크고 표본 주택의 가격 산출 과정에서 조사자의 주관이 개입될 여지가 없어 객관적으로 산출되기 때문에 이를 기존 부동산 가격 지수의 참고자료로써 활용할 수 있을 것이라고 주장하였다.

2. 머신 러닝 방법을 이용한 비주거용 부동산 가치 산정 관련 연구

주거용 부동산을 대상으로 한 연구들과는 달리 상업용 부동산과 같은 비주거용 부동산을 대상으로 한 관련 연구는 자료 수집의 어려움으로 국내외 모두 찾아보기 힘든 편이다.

김선주 · 이상엽(2008)은 헤도닉 함수를 활용한 오피스 임대료 결정 모형 관련 연구들이 축적됨에 따라 새로운 방법론에 대한 필요성이 제기되고 있으며, 인공신경망은 관측치 수가 적거나 불규칙한 경우에도 적용이 가능하다는 장점을 가지고 있기 때문에 오피스 임대료 추정에 관한 연구에 적합한 방법이라고 주장하였다. 이에 해당 연구는 회귀 모형과 상호 보완적인 모형으로서 인공 신경망 모형의 적용 가능성을 검토하기 위해 서울시 소재의 오피스 251동에 대한 2004년 임대사례 조사 자료를 활용하여 임대료 추정 모형을 각 방법별로 구축하였고, 이들의 설명력 및 예측력을 비교하여 인공신경망의 성능이 더 좋다는 결과를 제시하였다. 다만 김태훈 · 홍한국(2004)과 마찬가지로 인공 신경망 모형의 경우 모형에 대한 해석이 어렵기 때문에 이들을 상호 보완하여 활용하여야 한다는 점을 강조하였다.

문근식 외(2015)는 오피스 가격결정모형과 관련한 기존 연구들이 회귀 모형에 편중되어 있다는 점을 지적하고 연구방법에 대한 다양화가 필요하다고 주장하였다. 이에 해당 연구는 2006년 1월부터 2013년 8월까지 발생한 강남구 소재 중소형 오피스의 매매건 1,056건에 해당하는 자료를 활용하여 다중 회귀 모형, 의사결정나무 모형, 인공 신경망 모형을 구축하고, 모형별 중요변수 비교, 설명력 및 예측력을 비교, 이득도표(gain chart, response chart) 비교를 수행하였다. 중요변수는 모형별로 순위나 정도가 다소 상이하게 나타났으며, 설명력 및 예측력, 그리고 이득도표 비교 결과 인공 신경망, 의사결정나무, 다중 회귀 순으로 설명력 및 예측력이 우수한 것으로 확인되었으나, 각 분석 방법의 장단점이 다르기 때문에 세 가지 분석 방법을 함께 고려하는 것이 바람직하다고 주장하였다.

3. 선행연구와의 차별성

본 연구는 기존 선행연구들과 다음과 같은 점에서 차별성을 갖는다.

첫째, 다양한 머신 러닝 방법을 사용하여 최적의 오피스 임대료 산정 모형을 구축한다. 상업용 부동산에서 발생하는 주요 수익인 임대료는 해당 부동산의 가치를 결정하며, 이를 물건으로 한 직접투자나 간접투자의 기대 수익률 산정에 있어 상당히 중요한 정보이기 때문에 보다 객관적인 임대료 추정이 필요하다. 하지만 최근 국내외적으로 머신 러닝 기법을 활용한 비모수 추정 연구가 주목받고 있음에도 불구하고, 부동산 관련 연구에서는 주로 주거용 부동산의 가치 평가 문제에 머신 러닝 방법의 적용 가능성을 확인하고자 한 연구들이 주를 이루고 있기 때문에 상업용 부동산의 적정 임대료 산정에 적용 가능성을 검토하는 것 또한 의미가 있을 것으로 판단된다. 이와 관련하여 기존의 임대료 결정모형을 도출한 연구들(변기영 · 이창수, 2004; 김의준 · 김용환, 2006; 양영준 · 오세준, 2017 등)을 살펴본 결과, 이들은 결정요인 및 한계효과(marginal effect) 규명을 위해 설명 중심의 모수 모형을 설정하였다. 하지만 모수 모형은 설명변수의 독립성(independence), 자료의 정규성(normality), 종속변수와 설명변수 간의 선형성(linearity) 등 갖가지 엄격한 가정들을 하며(이창로 · 박기호, 2016에서 재인용), 이상치에 민감하게 반응하고 회귀식 선택 자체에 대한 의문이 제기될 수 있어(김태훈 · 홍한국, 2004) 이를 통해 예측된 추정 가격은 신뢰성에 한계가 존재한다. 따라서 본 연구는 이러한 제약에서 자유로운 머신 러닝 방법을 사용하여 최적의 오피스 임대료 산정 모형을 구축하며, 이때 보편적인 머신 러닝 방법인 랜덤 포레스트, 인공 신경망, 그리고 서포트 벡터 머신을 사용하여 각 방법별로 최적 모형을 구축하고 성능 비교를 통해 각각의 적용 가능성을 검토한다.

둘째, PD plot을 도출하여 각 설명변수들과 산정된

임대료 간의 관계를 확인하고 이를 해석한다. 다양한 머신 러닝 방법을 사용한 선행연구들은 대부분 예측 중심의 연구로서 각 최적 모형들의 성능을 비교 및 제시하는 것에 집중되어있다. Kok et al.(2017), 김선주 · 이상엽(2008), 문근식 외(2015) 등 몇몇 연구에서 변수의 중요도(variable importance)를 제시함으로써 모형을 해석하고자 한 시도는 있었지만, 기존의 회귀분석 연구와 같이 설명변수와 종속변수 간의 관계를 확인하고자 한 연구는 찾아보기 힘든 실정이다. 이에 본 연구는 블랙 박스 모형에서 입력 변수와 예측값 간의 관계를 직관적으로 이해할 수 있도록 도식화한 PD plot을 도출하여 모형의 학습 결과를 해석한다.

III. 방법론 및 분석자료

1. 방법론

1) 랜덤 포레스트(random forests, RF)

랜덤 포레스트는 과적합(overfitting) 문제와 같은 기존의 나무 모형(tree model)의 한계점을 개선하기 위해 Breiman(2001)이 고안한 앙상블 기법으로, 기존의 훈련 데이터로부터 추출한 부트스트랩 샘플(bootstrap samples)로 다양한 나무 모형을 형성하는 방법이다. 즉, 여러 나무 모형을 형성하는 과정에서 임의성(randomness)을 부여하여 서로 상관되지 않은(uncorrelated) 나무들을 형성하고 이들의 분류 결과나 회귀 결과를 종합하여 예측력을 향상시키는 개념으로, 나무 모형보다 예측력이 크게 개선되고 안정적인 모형을 제공해주고(배성완 · 유정석, 2018), 무작위성으로부터 비롯된 대수의 법칙(law of large numbers)에 의해 과적합 문제를 피할 수 있다는 장점을 지닌다(Breiman, 2001).

랜덤 포레스트의 알고리즘(algorithm)은 다음과 같다. 먼저, 훈련 데이터로부터 N 크기를 갖는 부트스트랩 샘플 Z^*를 복원 추출한다. 그리고 이를 이용하여 나무 T_b를 성장시키는데, 입력변수 총 p개에서 무작위적(randomly)으로 m개의 변수를 선택하고 그 중 분할 개선도(goodness of split)[2]가 가장 높은 변수와 그 분기점을 찾아 이를 기준으로 데이터를 이진 분할(binary split)하는 두 하위 노드를 형성하고, 각 노드마다 이러한 분할 과정을 거치되 단말 노드의 크기가 $n_{\min}$에 도달하면 나무 T_b의 성장이 완료된다. 그리고 위의 과정을 $b=1$부터 B까지 반복하여 나무 앙상블이 형성되면 새로운 데이터 값 x에 대한 예측값은 식 (1)과 같이 앙상블을 이루고 있는 모든 나무들의 평균값을 취해 도출된다.

$$\widehat{f}_{rf}^{B}(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x) \tag{1}$$

결국, 랜덤 포레스트에서의 학습(train)은 위와 같이 여러 나무 모형을 형성하여 하나의 숲을 만드는 것이며, 이를 위해 연구자가 설정해야 하는 주요 초모수(hyper-parameter)는 숲을 이루는 나무 수 ntree와 노드 분할 시 무작위적으로 선택하는 변수의 개수인 mtry이다. 머신 러닝과 관련된 기존 연구들에서 밝히는 최적의 초모수들은 경험법칙(rule of thumb)을 통해 도출된 것으로, 분석에 사용되는 자료의 특성에 따라 최적의 초모수 값이 상이할 가능성이 있다. 결국 궁극적으로 시행착오(trial and error)를 통해 연구에 맞는 초모수 값을 찾아 모형을 구성하는 과정을 거쳐야 한다(Heaton, 2008).

2) 인공 신경망(artificial neural networks, ANN)

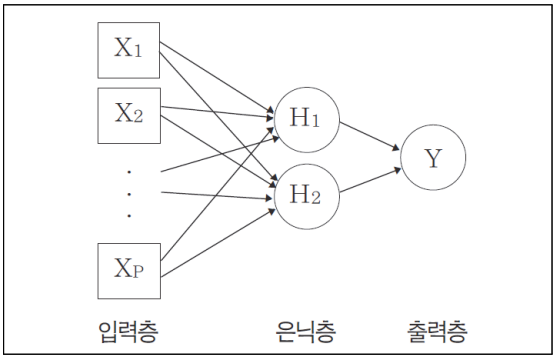
인공 신경망 모형은 뉴런(neuron)과 뉴런이 연결되어 서로 정보를 교환하는 동물의 신경계 구조를 모사한 수학적 모형으로, 1958년 Rosenblatt의 퍼셉트론(perceptron) 학습 규칙이라는 개념으로 등장하여 인공지능 개발 분야에서 큰 기대와 주목을 받았다. 그 후 1969년 Minsky and Papert에 의해 퍼셉트론이 간단한 XOR 분류조차 수행할 수 없다는 것이 수학적으로 증명되어 그 한계가 밝혔지만, 1986년 Rumelhart et al.이 은닉층을 가진 다층 퍼셉트론(multi-layer perceptron, MLP)을 학습할 수 있는 방법으로 역전파 알고리즘(backpropagation algorithm)을 소개함으로써 인공 신경망에 대한 다양한 연구와 적용이 발전되어왔다.

인공 신경망은 보편적으로 <그림 1>과 같이 다층 퍼셉트론 구조를 가진다. 입력층(input layer)은 각 입

2) 회귀 문제에서는 분할 개선도를 가장 높은 경우를 선택한다는 것은 두 하위 노드에서의 출력변수의 잔차제곱합(residual sum of squares, RSS)이 최소가 되도록 하는 분기점을 선택하는 것이다.

력변수(input variables)에 대응되는 노드로 구성되며, 이때 입력노드의 수는 입력변수의 수와 같다. 은닉층(hidden layer)은 입력층으로부터 전달되는 변수값들의 선형결합을 비선형함수로 처리하고 이를 출력층 또는 다른 은닉층에 전달하는 기능을 수행한다. 출력층(output layer)은 목표변수에 대응되는 노드로 구성된 층으로 예측 또는 분류된 결과를 나타낸다.

<그림 1> 인공 신경망 구조



자료 : 김태훈 · 홍한국(2004).

<그림 1>은 입력층과 두 개의 은닉노드를 가지는 하나의 은닉층, 그리고 출력층으로 이루어져 있는데, 해당 모형의 내부에서 이루어지는 선형결합과 비선형처리를 식 (2)와 같이 수식으로 나타낼 수 있다.

$$\begin{aligned} H_1 &= f(b_1 + w_{11}X_1 + w_{21}X_2 + \cdots w_{p1}X_p) \\ H_2 &= f(b_2 + w_{12}X_1 + w_{22}X_2 + \cdots w_{p2}X_p) \\ Y &= f(b_0 + w_{10}H_1 + w_{20}H_2) \end{aligned} \quad (2)$$

외부로부터의 입력 X가 연결 가중치 w와 곱해진 후 일들의 선형결합 값이 다음 노드로 전달된다. 그 후 활성함수(activation function) f를 통해 값을 선형결합 값을 노드의 활성화 여부 또는 활성화 정도로 변환하고 이를 신호(signal)로서 다음 층으로 전달한다. 이러한 일련의 과정을 거쳐 최종적인 출력값 Y를 얻을 수 있다.

신경망 모형에서의 학습은 비용함수(cost function)를 최소화하는 가중치 w를 찾는 최적화(optimization) 과정이다. 비용함수는 실제 y값인 t와 인공 신경망을 통해 도출된 y값인 o간의 차이(discrepancy) E를 측정하기 위해 식 (3)과 같이 오차 제곱(square error)로 정의한다.

$$L = \frac{1}{2}(t-o)^2 = E \quad (3)$$

그리고 전방향(feed-forward) 신경망 모형에서 비용함수 L을 최소화하는 w를 찾기 위해 경사하강법(gradient descent method)을 사용한다. 이는 식 (4)와 같이 초기값 w_0를 설정 한 후 매 스탭(step)에서 비용함수의 기울기(gradient)만큼 가중치를 점진적으로 변화시켜 점차 최적점(convergence)에 도달시키는 방법론이다.

$$w_{\tau+1} = w_\tau - \alpha \bullet L'(w_\tau) \qquad (0 < \alpha < 1) \quad (4)$$

이때, α는 학습률(learning rate)로, 해당 값이 클수록 가중치를 갱신(update)할 때 가중치 변화의 폭이 커져 학습속도는 빨라지지만 전역 최솟값(global minima)에 도달하지 못할 수 있고, 반대로 해당 값이 작을수록 가중치 변화의 폭이 작아져 학습속도는 느려지고 이 또한 전역 최솟값에 도달하지 못하고 지역 최솟값(local minima)에 수렴하게 될 수 있다. 이는 연구자가 선택해야 하는 초모수이다.

경사하강법 사용 시 비용함수의 기울기 $L'(w_\tau)$을 계산하는 알고리즘으로는 오류 역전파 알고리즘(error backpropagation algorithm)을 사용한다. 이는 오류를 출력층에서 입력층 방향으로, 즉 역방향으로 전파하는 방식인데, 국소적(local) 계산과 연쇄법칙(chain rule)을 통해 미분 문제를 단순화 할 수 있다는 장점이 있다.

인공 신경망의 학습 과정을 정리하면 다음과 같다. 첫째, 초기 가중치 w_0를 임의로 생성한다. 둘째, 입력층과 은닉층 사이의 w값을 이용하여 입력변수 값을 선형결합한 후 이를 활성함수에 투입하여 은닉노드의 값을 얻는다. 셋째, 은닉층과 출력층 사이의 w값을 이용하여 은닉노드의 값을 선형결합한 후 이를 활성함수에 투입하여 출력노드의 값을 얻는다. 넷째, 출력노드의 값과 실제 출력변수의 값의 차이를 줄일 수 있도록 은닉층과 출력층 사이의 w값을 업데이트(update)한다. 다섯 번째, 출력노드의 값과 실제 출력변수의 값의 차이를 줄일 수 있도록 입력층과 은닉층 사이의 w값을 업데이트(update)한다. 마지막으로, 오차가 충분히 줄어들 때까지 위의 과정을 반복한다.

인공 신경망 모형은 입력변수와 출력변수와의 상호 관계를 해석하기 어렵다는 단점이 있지만(남영우 · 이정민, 2006), 변수들의 선형결합과 이들의 비선형 처리가 다계층적으로 이루어지기 때문에 입력변수와 출력변수가 복잡한 구조를 가진 자료에서의 예측 문제를 해결할 수 있는 유연한 비선형 모형(nonlinear model)로 분류될 수 있고(김선주 · 이상엽, 2008), 활성함수와 은닉층 구조가 적절히 선택될 때 학습을 통해 모든 비선형 곡선을 상당히 정확하게 근사시킬 수 있다는 강점을 갖는다(김태훈 · 홍한국, 2004). 인공 신경망에서의 주요 초모수는 활성함수 f[3], 은닉노드 수 size, 그리고 학습률 α[4]이며, 연구자는 이를 다양하게 변화시켜 가며 일반화 오차가 가장 작은 최적 모형을 선택하는 과정을 거쳐야 한다.

3) 서포트 벡터 머신(support vector machines, SVM)

서포트 벡터 머신은 데이터를 고차원(high dimension)의 형상 공간(feature space)로 사상(mapping)하여 최적의 의사결정 영역(decision boundary)을 구해내는 방법이다. 본 개념은 1963년 Vapnik and Chervonenkis에 의해 선형 분류기(linear classifier)로 처음 등장하였고, 1992년 Boser et al.에 의해 커널 트릭(kernel trick)이 적용된 비선형 분류기가 고안되었다. 그 후 1997년 Vapnik et al.이 ε-무감도 손실함수(ε-insensitive loss function)를 도입한 SVM 모형을 제시함으로써 회귀 문제의 영역까지 적용이 확장되었다. 서포트 벡터 머신은 구조적 위험에 대한 최소화 원칙(structural risk minimization, SRM)을 구현함으로써 과적합 문제에 자유롭다는 점, 전역 최적해를 구할 수 있다는 장점이 있다(김성진 외, 2012에서 재인용).

$$minimize\ \frac{1}{2}\parallel w\parallel^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$
$$subject\ to \begin{cases} y_i - \langle w, x_i\rangle - b \le \epsilon+\xi_i \\ \langle w, x_i\rangle + b - y_i \le \epsilon+\xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (5)$$

서포트 벡터 머신은 마진(margin)을 최대화하는 초평면(hyperplane)을 찾는 모델이다. 마진을 최대화하는 문제는 식 (5)와 같이 초평면의 w와 b를 최소화하는 문제로 나타낼 수 있다.

$$L_\epsilon = \begin{cases} 0 & \text{if } |y-f(x)| \le \epsilon \\ |y-f(x)| - \epsilon & otherwise \end{cases} \quad (6)$$

회귀 문제에서의 서포트 벡터 머신이 분류 문제에서의 서포트 벡터 머신과 다른 점은 식 (6)과 같은 ε-무감도 손실함수 개념을 추가되었다는 것이다. 이는 학습된 함수로부터의 예측값과 실제값의 차이의 크기가 ε을 초과할 경우만 이를 손실로 인정한다는 것이다. 즉 ε은 ε-무감도 지역(ε-insensitive zone)의 너비를 나타내며, 손실을 최대한으로 줄이기 위해 ε-무감도 지역에 최대한 많은 관측치가 속하도록 학습을 진행하게 된다.

식 (5)에서 목적함수의 첫 번째 항은 함수의 편평함(flatness)와 관련된 항으로(김희훈, 2012), 추정된 함수의 진동에 의해 예측력이 떨어지는 룽게현상(Runge phenomenon)을 방지하기 위해 $f(x)$를 최대한으로 편평하게 해야 한다(Deng et al., 2012). $f(x)$의 편평함은 초평면의 법선 벡터 w에 의해 결정이 되며 w의 노름(norm) $\parallel w \parallel$이 작을수록 $f(x)$는 편평해진다. 목적함수의 두 번째 항에서 ξ는 학습하는 과정에서 어느 정도 오차를 허용함으로써 모형의 유연함을 증진시키고 성능을 높이는 역할을 한다. C는 허용된 오차에 대한 패널티(penalty)를 할당함으로써, w에 의해 결정되는 모형의 복잡도와 오차 허용 정도에 대한 트레이드오프(trade-off)를 결정하는 상수로 C가 높을수록 오차 허용 정도가 낮아지므로 복잡한 모형이 형성된다.

식 (5)에 대한 계산을 용이하게 하기 위해 라그랑지 승수법(Lagrange multiplier method)을 이용하여 본 원 문제(primal problem)식을 쌍대 문제(dual problem)식으로 변환하고, Karush-Kuhn-Tucker 조건을 적용하여 최적해를 도출한다[5]. 쌍대 문제에서의 최적해인 라그랑지 승수 α와 α^*가 0이 아닌 경우와 대응되는 관측치를 서포트 벡터로 칭하며, 이를 이용하여 식 (7)과 같이 선형 추정함수를 얻을 수 있다.

3) 본 연구는 인공 신경망 연구에서 주로 사용되는 시그모이드 함수(sigmoid function)를 활성함수로 사용하였다. 본 활성함수는 S자 형태의 비선형함수로서 출력값을 0과 1사이의 값으로 변형시키는 미분가능한 함수이다. 미분한 결과를 다시 본 함수의 형태로 표현이 가능하다는 점에서 학습 과정을 용이하게 한다는 장점이 있다.

4) 다만 본 연구에서는 분석의 효율성을 위해 통계 프로그램이 자동적으로 찾은 α값을 사용하였다.

5) 자세한 풀이과정은 Smola and Schölkopf(2004), 김희훈 외(2012)를 참고하길 바란다.

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\langle x_i,x\rangle+b \quad (7)$$

그 후 식 (7)에 커널 함수(kernel function) $k(x_i,x)$를 적용함으로써 식 (8)과 같은 최종 비선형 추정함수를 도출한다. 이는 입력값을 고차원의 커널 형상 공간으로 사상하여 새로운 형상 공간에서 선형 함수를 찾는 것과 같은 효과를 가진다.

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i,x)+b \quad (8)$$

이때, 커널 함수 k로는 선형 커널(linear Kernel), 다항식 커널(polynominal Kernel), 가우시안 방사 기저함수(Gaussian radial basis function, RBF), 라플라스 방사 기저함수(Laplace radial basis function), 쌍곡탄젠트 커널(hyperbolic tangent Kernel), ANOVA 커널 등이 사용된다.

서포트 벡터 머신에서의 주요 초모수는 ε, C, 커널 함수 k이다. 또한 커널 함수 선택에 따라 추가적으로 설정해야 되는 초모수의 종류가 상이한데, 본 연구에서는 커널 함수로 가장 보편적으로 사용되는 RBF 커널 함수를 분석에 사용하기 때문에 σ가 초모수로 추가된다. 여타 머신 러닝 방법과 마찬가지로 시행 착오를 통해 연구에 맞는 최적의 초모수 값을 찾아 모형을 구축하는 과정을 거친다.

4) 부분 의존성 플롯(partial dependence plot, PD plot)

머신 러닝 방법을 통해 구축한 모형은 복잡하고 해석이 용이하지 않다는 의미로 블랙박스 모형(black-box modle)으로 여겨지는데, PD plot은 이러한 블랙박스 모형의 해석 방법 중 하나이다. 이는 모델의 예측값이 특정 입력값과 어떠한 관계를 갖고 있는지 나타내는 일원(one-way) 플롯으로, 그 개념이 간단하며 도식화를 통해 직관적인 해석을 가능케 한다. 이를 통해 머신 러닝 모형의 예측 결과값에 대한 특정한 속성(feature)의 한계 효과(marginal effect)를 확인할 수 있다(Friedman, 2001).

$$\widehat{f_{x_S}}(x_S)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S,x_C^{(i)}) \quad (9)$$

PD 함수(partial dependence function, PD function)는 식 (9)와 같다. 이때, x_S는 관심 변수, $x_C^{(i)}$는 그 외의 변수에 대한 훈련 데이터 값을 의미한다. PD값은 고정된 x_S에 대해 모든 $x_C^{(i)}$의 속성 공간(feature space)에서 예측값의 평균값이다. 그리고 x_S를 변화시켜가며 PD값을 플롯함으로써 하나의 선을 도출할 수 있으며, 이를 통해 x_S에 따른 예측값 평균의 변화 모습을 시각적으로 확인할 수 있다.

2. 분석자료

본 연구에서는 머신 러닝 방법을 사용하여 임대료 산정 모형을 구축하기 위해 서울시에 소재한 연면적 3,000평 이상의 오피스 507동을 분석대상으로 하였다. 분석에 사용된 변수 구성 및 정의는 <표 1>과 같다. 먼저 종속변수인 임대료는 각 오피스 기준층에 대한 호가 임대료이며 단위는 원/평이다. 공실률은 임대가 가능한 면적 중 공실면적이 차지하는 비율을 나타낸다. 전용률은 계약면적 중 전용면적이 차지하는 비율로 임차인이 지불한 금액이 업무공간으로 환산되는 정도를 나타낸다. 연면적은 총 바닥면적의 합으로 건물의 규모를 나타내며 단위는 평이다. 층수는 지하층을 제외한 지상층수이다. 연식은 건물 사용승인일로부터 2017년 12월 31일까지의 경과월수를 의미한다. 승강기수는 승용승강기 대수, 주차대수는 주차가능 대수를 나타낸다. 녹색건축 인증여부는 해당 오피스가 녹색건축 3등급 이상의 건물일 경우 1의 값을, 그 외에는 0의 값을 부여한 더미변수이다. 권역은 강남권역, 도심권역, 여의도권역, 기타권역으로 나누었으며, 강남권역은 해당 오피스가 강남구 또는 서초구 소재의 건물일 경우, 도심권역은 해당 오피스가 종로구 또는 중구 소재의 건물일 경우, 여의도권역은 여의도 또는 마포구 소재의 건물일 경우, 기타권역은 그 외에 소재할 경우를 나타내는 더미변수이다. 임대료와 공실률은 국내 부동산 자산관리회사의 임대 조사 자료를 통해, 전용률은 여러 부동산 중개회사 사이트 및 블로그(blog)를 통해, 그 외 특성 정보들은 서울시 건축물 대장을 통해 구득하였다.

<표 1> 변수 구성 및 정의

변수명		단위	변수설명	자료출처
임대료		원/평	평당 월 임대료(호가)	부동산 자산관리 회사
공실률		비율	층 임대면적 중 공실면적 비율	
전용률		비율	계약면적 중 전용면적이 차지하는 비율	부동산 중개회사
연면적		평	총 바닥면적의 합	서울시 건축물 대장
층수		층	지하층을 제외한 지상층수	
연식		개월	사용승인일로부터 2017년 12월 31일까지 경과월수	
승강기수		대	승용승강기 대수	
주차대수		대	주차가능 대수	
녹색건축 인증여부		1 또는 0	녹색건축 3등급 이상의 건물 = 1, 그 외 = 0	
권역	강남 권역	1 또는 0	강남구 · 서초구 소재의 건물 = 1, 그 외 = 0	
	도심 권역		종로구 · 중구 소재의 건물 = 1, 그 외 = 0	
	여의도 권역		여의도 · 마포구 소재의 건물 = 1, 그 외 = 0	
	기타 권역		기타 서울지역 소재의 건물 = 1, 그 외 = 0	

전체 표본에 대한 기초통계량은 <표 2>에 제시되어 있다. 먼저 임대료는 평당 평균적으로 약 66,191원이며, 공실률은 평균적으로 약 9%인 오피스로 표본이 구성되어 있다. 전용률은 평균 약 55%, 연면적은 평균 약 11,122평, 층수는 평균 약 18층, 연식은 평균 약 256개월, 승강기수는 평균 약 6대, 주차대수는 약 266대인 것으로 나타났다. 전체 표본 중 약 4.5%의 오피스가 녹색건축 인증을 받은 것으로 확인되었으며, 강남권역에 소재한 오피스가 약 36%, 도심권역에 소재한 오피스가 약 26%, 여의도권역에 소재한 오피스가 약 18%, 그리고 그 외의 권역에 소재한 오피스가 약 21%인 것으로 확인되었다.

<표 2> 기초통계량 - 전체

변수명(단위)		관측수	평균값	표준 편차	최솟값	최댓값
임대료(원/평)		507	66,191	22,419	25,000	146,500
공실률(비율)		507	0.09	0.14	0	0.89
전용률(비율)		507	0.55	0.07	0.31	0.99
연면적(평)		507	11,122	10,118	3,002	98,959
층수(층)		507	18.02	8.23	5	123
연식(월)		507	256.09	130.13	1	659
승강기수(대)		507	5.58	5.14	1	36
주차대수(대)		507	266	389	5	4,032
녹색건축 인증여부 (1=녹색인증)		507	0.045	0.208	0	1
권역 (1=해당지역)	강남 권역	507	0.36	0.48	0	1
	도심 권역	507	0.26	0.44	0	1
	여의도 권역	507	0.18	0.38	0	1
	기타 권역	507	0.21	0.41	0	1

본 연구는 규모 등급별 임대료 산정 모형을 각각 구축하기 위해 전체 표본을 규모를 기준으로 하여 총 4개의 등급(초대형, 대형, 중대형, 중형)으로 나누었다. 먼저 연면적이 20,000평 이상인 오피스를 초대형 등급으로 분류하였고 총 65동이 해당 등급에 속한다. 연면적이 10,000평 이상, 20,000평 미만인 오피스를 대형 등급으로 분류하였고 총 136동이 해당 등급에 속한다. 연면적이 5,000평 이상, 10,000평 미만인 오피스를 중대형 등급으로 분류하였고 총 176동이 해당 등급에 속한다. 마지막으로 연면적이 3,000평 이상, 5,000평 미

만인 오피스를 중형 등급으로 분류하였고 총 130동이 해당 등급에 속한다. 위 기준은 자산관리회사 젠스타의 오피스 등급 분류 기준을 참고한 것이다.

<표 3> 기초통계량 - 초대형 오피스

변수명(단위)		관측수	평균값	표준편차	최솟값	최댓값
임대료(원/평)		65	88,172	28,793	31,800	146,500
공실률(비율)		65	0.12	0.19	0	0.83
전용률(비율)		65	0.54	0.07	0.43	0.71
연면적(평)		65	32,116	13,452	20,005	98,959
층수(층)		65	28.71	15.57	7	123
연식(월)		65	196.52	145.91	1	564
승강기수(대)		65	15.43	8.15	4	36
주차대수(대)		65	819	761	167	4,032
녹색건축 인증여부 (1=녹색인증)		65	0.12	0.33	0	1
권역 (1=해당지역)	강남권역	65	0.23	0.42	0	1
	도심권역	65	0.34	0.48	0	1
	여의도권역	65	0.15	0.36	0	1
	기타권역	65	0.28	0.45	0	1

<표 4> 기초통계량 - 대형 오피스

변수명(단위)		관측수	평균값	표준편차	최솟값	최댓값
임대료(원/평)		136	72,224	21,051	31,000	138,000
공실률(비율)		136	0.09	0.12	0	0.64
전용률(비율)		136	0.53	0.08	0.31	0.99
연면적(평)		136	13,224	2,622	10,007	19,735
층수(층)		136	20.31	4.67	9	37
연식(월)		136	233.28	130.90	9	593
승강기수(대)		136	6.33	2.02	3	14
주차대수(대)		136	296	230	50	2,566
녹색건축 인증여부 (1=녹색인증)		136	0.07	0.25	0	1
권역 (1=해당지역)	강남권역	136	0.32	0.47	0	1
	도심권역	136	0.33	0.47	0	1
	여의도권역	136	0.14	0.35	0	1
	기타권역	136	0.21	0.41	0	1

<표 5> 기초통계량 - 중대형 오피스

변수명(단위)		관측수	평균값	표준편차	최솟값	최댓값
임대료(원/평)		176	61,667	17,060	27,778	104,600
공실률(비율)		176	0.08	0.14	0	0.89
전용률(비율)		176	0.56	0.07	0.38	0.78
연면적(평)		176	7,060	1,441	5,023	9,999
층수(층)		176	15.72	3.76	5	26
연식(월)		176	265.58	120	2	530
승강기수(대)		176	3.73	1.07	1	7
주차대수(대)		176	166	190	34	1,832
녹색건축 인증여부 (1=녹색인증)		176	0.02	0.15	0	1
권역 (1=해당지역)	강남권역	176	0.36	0.48	0	1
	도심권역	176	0.24	0.43	0	1
	여의도권역	176	0.19	0.39	0	1
	기타권역	176	0.21	0.41	0	1

<표 6> 기초통계량 - 중형 오피스

변수명(단위)		관측수	평균값	표준편차	최솟값	최댓값
임대료(원/평)		130	55,014	16,309	25,000	96,429
공실률(비율)		130	0.08	0.12	0	0.81
전용률(비율)		130	0.56	0.07	0.4	0.8
연면적(평)		130	3,926	597	3,002	4,980
층수(층)		130	13.4	3.43	5	20
연식(월)		130	296.91	119.38	3	659
승강기수(대)		130	2.38	0.64	1	4
주차대수(대)		130	92	47	5	480
녹색건축 인증여부 (1=녹색인증)		130	0.02	0.12	0	1
권역 (1=해당지역)	강남권역	130	0.46	0.5	0	1
	도심권역	130	0.15	0.36	0	1
	여의도권역	130	0.22	0.41	0	1
	기타권역	130	0.17	0.38	0	1

각 규모 등급별 기초통계량은 <표 3>부터 <표 6>에 제시되어 있다. 초대형 등급 오피스의 임대료는 평당 평균 약 88,172원, 공실률은 평균 약 12%로, 대형 등급 오피스의 임대료는 평당 평균 약 72,224원, 공실률은 평균 약 9%, 중대형 등급 오피스의 임대료는 평당 평균 약 61,667원, 공실률은 평균 약 8%, 중형 등급 오피스의 임대료는 평당 평균 약 55,014원, 공실률은 평균 약 8%로 나타났다. 규모 등급이 높을수록 임대료 수준과 공실률 수준이 높은 것 외에도, 규모 등급이 높을수록 연식이 낮고, 주차대수가 많았으며, 녹색건축 인증 오피스 비율이 높은 것으로 확인되었다. 권역 분포 또한 등급별로 다소 상이한 모습을 보였다. 초대형 등급 오피스는 도심권역에 가장 많이(약 34%) 분포되어 있는 것으로 나타났고, 그 다음 기타권역(약 28%), 강남권역(약 23%), 여의도권역(약 15%) 순으로 분포되어 있는 것으로 나타났다. 대형 등급 오피스는 도심권역에 가장 많이(약 33%) 분포되어 있는 것으로 나타났고, 또한 강남권역에 많이(약 32%) 분포되어 있는 것으로 나타났으며, 그 다음 기타권역(약 21%), 여의도권역(약 14%) 순으로

분포되어 있는 것으로 나타났다. 중대형 등급 오피스는 강남권역에 가장 많이(약 36%) 분포되어 있는 것으로 나타났고, 그 다음 도심권역(약 24%), 기타권역(약 21%), 여의도권역(약 19%) 순으로 분포되어 있는 것으로 나타났다. 중형 등급 오피스는 강남권역에 가장 많이(약 46%) 분포되어 있는 것으로 나타났으며, 그다음으로 여의도권역(약 22%), 기타권역(약 17%), 도심권역(약 15%) 순으로 분포되어 있는 것으로 나타났다.

3. 분석방법

본 연구에서는 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신 방법 각각에 대한 최적의 임대료 산정 모형을 구축하고, 각 최적 모형들의 성능을 비교한다.

먼저, 머신 러닝에 사용되는 변수들의 단위가 상이할 경우 학습되는 모형의 성능이 저하될 수 있으므로 모든 변수를 0과 1사이의 값을 갖도록 정규화(normalization)를 한다.

그 후, 모형의 과적합 문제를 피하기 위해 전체 데이터 셋에서 관측치의 약 70%를 훈련 데이터 셋(train data set)으로 모형을 학습한다. 이때 k겹 교차 검증(k-fold cross validation)을 수행하는데, k겹 교차 검증이란 훈련 데이터를 k등분한 후, 등분된 데이터 중 k-1개를 학습에 사용하고 나머지 1개를 모형에 투입하여 학습된 모형의 오차를 측정하고, 이러한 학습과 검증 과정을 k번 반복함으로써 도출된 k개의 오차를 평균하여 일반화 오차를 측정하는 방법이다. 오차는 검증 데이터(validation data)에 대한 평균제곱근오차(root mean square error, RMSE)와 평균절댓값오차(mean absolute error, MAE)로 측정하며, 본 연구는 각 방법별 초모수를 다양하게 변화시켜가며 10겹 교차 검증을 수행하여 가장 일반화 오차가 작은 최적의 모형을 구축한다[6].

이어, 모형의 학습에 사용되지 않은 남은 데이터 셋인 실험 데이터 셋(test data set)으로 최적 모형의 성능을 검증한다. 이는 새로운 데이터가 주어졌을 때 학습된 모형이 얼마만큼 예측 또는 추정을 정확하게 수행하는지 확인하기 위한 것이다. 성능은 실험 데이터에 대한 RMSE와 MAE로 측정하고, 각 최적 모형별 성능 비교를 통해 임대료 산정 모형 구축에 있어 각 머신 러닝 방법들의 적용 가능성을 검토한다[7].

가장 성능이 좋은 최종 모형이 선정되면 모형 해석을 위한 PD plot을 도출한다. 해당 방법은 블랙박스 모형을 해석하는 대표적인 방법으로, 모형의 예측값과 특정 투입변수와의 관계를 시각적으로 나타내어 결과가 기존 선행연구나 직관에 어긋나지 않는지 확인한다.

추가적으로, 표본을 규모 등급별로 나누어 오피스 등급별 최적 모형 구축, 성능 비교, PD plot 도출 또한 수행한다.

IV. 분석결과

1. 오피스 임대료 산정 모형 구축 결과

1) 최적 모형의 성능 비교

랜덤 포레스트 방법을 이용하여 최적 모형을 도출하기 위해 ntree(나무 수)를 10, 20, 50, 100, 200, 300, 400, 500, 1000, 2000, 5000으로, mtry(노드 분할 시 무작위적으로 선택하는 변수의 개수)를 1부터 12까지 조절하였고, 그 결과로 ntree가 1,000, mtry가 11인 경우와 ntree가 2,000, mtry가 12인 경우에 가장 일반화 오차가 낮은 것으로 나타나 본 연구는 해당 두 모형을 랜덤 포레스트 방법의 최적 모형으로 선정하였다.

인공 신경망 방법을 이용하여 최적 모형을 도출하기 위해 size(은닉노드 수)를 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70으로 조절하였고, 그 결과로 size가 5인 경우와 size가 60인 경우에 가장 일반화 오차가 낮은 것으로 나타나 본 연구는 해당 두 모형을 인공 신경망 방법의 최적 모형으로 선정하였다.

서포트 벡터 머신을 이용하여 최적 모형을 도출하기 위해 ε(ε-무감도 지역의 너비)을 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9로, C(허용된 오차에 대한 페널티 상수)를 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32로, σ(RBF 커널 함수의 초모수)를 2^(-9), 2^(-8), 2^(-7), 2^(-6), 2^(-5), 2^(-4), 2^(-3), 2^(-2), 2^(-1)로 조절하였고, 그 결과로 ε이 0.5, C가 2, σ가 2^(-6)인 경우와 ε이 0.09, C가 2, σ가 2^(-6)인 경우에 가장 일반화 오차

6) 본 연구는 논의의 편의성을 위해 RMSE를 기준으로 최적 모형을 선정한다.

7) 최적 모형 선정과 마찬가지로, 본 연구는 논의의 편의성을 위해 RMSE를 기준으로 모형의 성능을 비교한다.

가 낮은 것으로 나타나 본 연구는 해당 두 모형을 서포트 벡터 머신 방법의 최적 모형으로 선정하였다.

<표 7>은 실험 데이터 셋에 대한 최적 모형들의 성능을 비교한 결과를 나타낸 것이다. RMSE를 기준으로 할 때, 먼저 랜덤 포레스트 최적 모형의 경우 ntree가 2,000, mtry가 12인 최적 모형의 성능이 더 좋은 것으로 나타났고, 인공 신경망 최적 모형의 경우 size가 5인 최적 모형의 성능이 더 좋은 것으로 나타났으며, 서포트 벡터 머신 최적 모형의 경우 ε이 0.09, C가 2, σ가 2^(-6)인 최적 모형의 성능이 더 좋은 것으로 나타났다.

<표 7> 최적 모형별 성능 비교

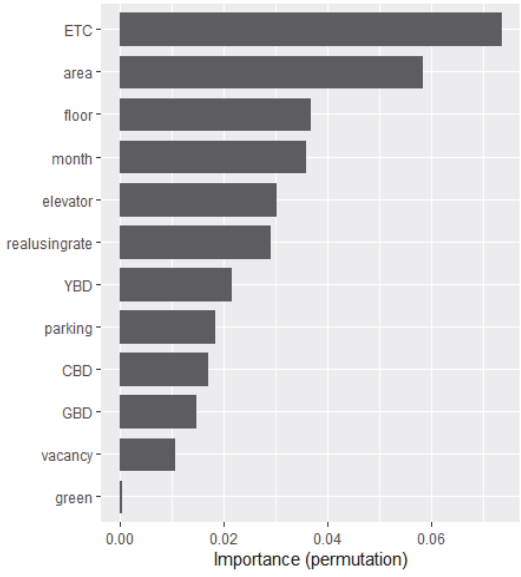
구분	최적 모형별 초모수	일반화 오차		성능 검증	
		RMSE	MAE	RMSE	MAE
RF	ntree=1,000, mtry=11	0.11796	0.09335	0.13348	0.09730
	ntree=2,000, mtry=12	0.11798	0.09316	0.13289	0.09689
ANN	size=5	0.13066	0.10245	0.13416	0.09985
	size=60	0.13190	0.10363	0.13610	0.10036
SVM	ε=0.5, C=2, σ=2^(-6)	0.11971	0.09722	0.13015	0.09837
	ε=0.09, C=2, σ=2^(-6)	0.11973	0.09386	0.12723	0.09577

각 방법별로는 서포트 벡터 머신의 최적 모형의 성능이 가장 우수하였고, 그 다음으로는 랜덤 포레스트의 최적 모형의 성능이 좋게 나타났으며, 인공 신경망의 최적 모형의 경우 다른 방법의 최적 모형들보다는 성능이 상대적으로 떨어지는 것으로 확인되었다.

2) 속성 선택(feature selection) 및 최종 모형

<그림 2>는 ntree가 2,000, mtry가 12인 랜덤 포레스트 모형, size가 5인 인공 신경망 모형, 그리고 ε이 0.09, C가 2, σ가 2^(-6)인 모형으로부터 각각 변수 중요도(variable importance)를 도식화한 결과이다. 이는 다양한 머신 러닝 모형에 일관적으로 적용할 수 있는 변수 중요도 도출 방법인 순열 특성 중요도(permutation importance)[8]를 나타낸 것이다.

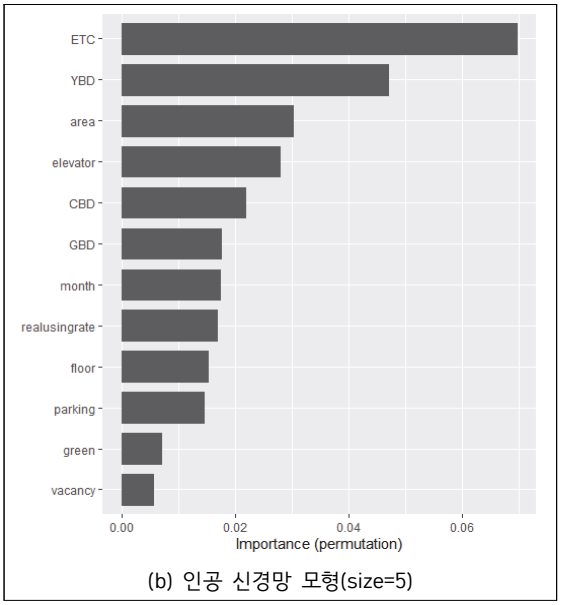
<그림 2> 변수 중요도



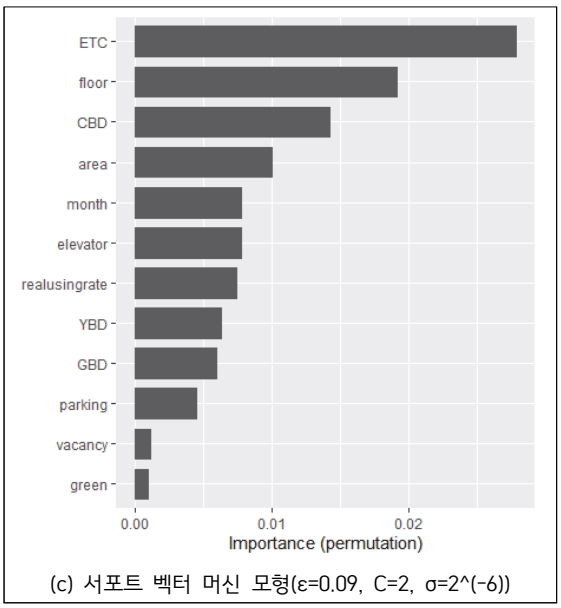
(a) 랜덤 포레스트 모형(ntree=2.000, mtry=12)

8) 순열 특성 중요도는 Breiman(2001)이 랜덤 포레스트 모형에서의 변수 중요도를 도출하기 위해 제안한 방법으로, Fisher et al.(2018)이 이 외의 다양한 예측 모형에도 적용이 가능한 모형-불구속적(model-agnostic)인 개념으로 발전시켰다. 이는 기존의 오류와 특정한 특성(feature)을 셔플링(shuffling)하여 훈련된 모형에 투입하였을 때의 오류의 차이 또는 그 변화 정도로 변수 중요도를 측정하는 것으로, 만약 모형의 예측 오류가 크게 증가한다면 이는 모형의 성능이 해당 변수에 상당히 의존한다는 의미이기 때문에 해당 변수를 중요한 변수로 판단하는 것이다. 본 방법은 다양한 머신 러닝 모형에 일관적인 방법으로 적용이 가능하다는 점과 최적화된 모형에 곧바로 적용이 가능하여 효율적으로 결과를 얻어낼 수 있다는 장점을 가진다.

<그림 2> 계속



(b) 인공 신경망 모형(size=5)



(c) 서포트 벡터 머신 모형(ε=0.09, C=2, σ=2^(-6))

<그림 2>에 따르면 랜덤 포레스트 모형의 경우 기타권역(ETC), 연면적(area), 층수(floor), 연식(month), 승강기수(elevator), 전용률(rentable/usable ratio), 여의도권역(YBD), 주차대수(parking), 도심권역(CBD), 강남권역(GBD), 공실률(vacancy), 녹색건축 인증여부(green) 변수 순으로 변수 중요도가 도출되었다. 인공신경망 모형의 경우 기타권역, 여의도권역, 연면적, 승강기수, 도심권역, 강남권역, 연식, 전용률, 층수, 주차대수, 녹색건축 인증여부, 공실률 순으로 변수 중요도가 도출되었다. 서포트 벡터 머신 모형의 경우 기타권역(ETC), 층수, 도심권역, 연면적, 연식, 승강기수, 전용률, 여의도권역, 강남권역, 주차대수, 공실률, 녹색건축 인증여부 순으로 변수 중요도가 도출되었다. 세 모형 모두 공통적으로 기타권역 더미변수(ETC)가 모형들의 성능에 가장 큰 영향을 주는 것으로 나타났고, 연면적 변수(area) 또한 모형의 성능에 큰 영향을 주는 것으로 나타났다. 하지만 녹색건축 인증여부 더미변수(green)와 공실률(vacancy) 변수는 모형들의 성능에는 큰 영향을 주지 않는 것으로 나타났다. 이에 본 연구는 녹색건축 인증여부와 공실률을 투입변수에서 제외하고 나머지 변수로 다시 모형을 학습하는 속성 선택(feature selection) 과정을 수행하였으며, 이를 통해 구축된 최적 모형의 성능을 <표 8>과 같이 나타내었다.

<표 7>과 <표 8>에 따르면 세 가지 머신 러닝 방법 모두 녹색건축 인증여부 변수를 제외하고 모형을 학습시켰을 때 최적 모형의 성능이 개선되는 것으로 나타났다. 인공 신경망의 경우에는 속성 선택 전 최적 모형(size=5)의 RMSE가 0.134156이었던 반면에 속성 선택 후 최적 모형(size=45)의 RMSE는 0.124274으로 성능 개선 정도가 가장 크게 나타났다. 서포트 벡터 머신의 경우에는 속성 선택 전 최적 모형(ε=0.09, C=2, σ=2^(-6))의 RMSE가 0.127234이었고, 속성 선택 후 최적 모형(ε=0.2, C=4, σ=2^(-6))의 RMSE는 0.122910으로 성능 개선이 다소 이루어진 것으로 나타났다. 랜덤 포레스트의 경우에는 속성 선택 전 최적 모형(ntree=2,000, mtry=12)의 RMSE가 0.132893, 속성 선택 후 최적 모형(ntree=300, mtry=10)의 RMSE는 0.132509로 성능 개선 정도가 미미한 것으로 나타났다. 녹색건축 인증여부와 공실률을 함께 제외한 경우에도 세 방법 모두 성능이 다소 개선되는 것으로 나타났지만, 녹색건축 인증여부만을 제외한 후 구축된 최적 모형의 성능이 가장 좋은 것으로 확인되었다. 이에 본 연구는 녹색건축 인증여부를 투입변수에서 제외하였을 때의 최적 모형 중 랜덤 포레스트는 ntree가 300, mtry가 10인 모형, 인공 신경망은 size가 45인 모형, 서포트 벡터 머신은 ε이 0.2, C가4, σ가2^(-6)인 모형을 각 방법별 최종 임대료 산정 모형으로 선정하였다.

<표 8> 속성 선택 후 구축된 최적 모형의 성능

구분	제외변수	최적 모형별 초모수	성능 검증	
			RMSE	MAE
RF	녹색건축 인증여부	ntree=300, mtry=10	0.132509	0.097491
		ntree=400, mtry=10	0.133104	0.097754
	녹색건축 인증여부, 공실률	ntree=400, mtry=9	0.132938	0.097000
		ntree=500, mtry=9	0.132837	0.096685
ANN	공실률	size=70	0.139113	0.101837
		size=65	0.140969	0.103090
	녹색건축 인증여부	size=5	0.127030	0.095667
		size=45	0.124274	0.096639
	녹색건축 인증여부, 공실률	size=50	0.136275	0.101465
		size=55	0.125686	0.097950
SVM	녹색건축 인증여부	ε=0.2, C=4, σ=2^(-6)	0.122910	0.092384
		ε=0.5, C=8, σ=2^(-7)	0.127856	0.096438
	녹색건축 인증여부, 공실률	ε=0.5, C=2, σ=2^(-5)	0.128763	0.096779
		ε=0.2, C=4, σ=2^(-6)	0.125172	0.093847

본 연구는 각 최종 임대료 산정 모형의 적용 가능성을 검토하기 위해 모형별 성능 비교와 더불어, 자동 평가 모형에 주로 적용되어 왔던 다중 회귀 모형과의 비교를 실시하였고 그 결과는 <표 9>에 제시되어 있다.

<표 9> 각 방법별 최종 모형의 성능 비교

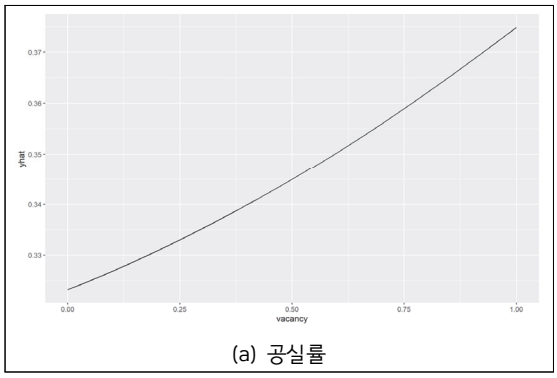
구분	제외변수	모형별 초모수	RMSE	MAE
RF	녹색건축 인증여부	ntree=300, mtry=10	0.132509	0.097491
ANN		size=45	0.124274	0.096639
SVM		ε=0.2, C=4, σ=2^(-6)	0.122910	0.092384
다중 회귀	-	-	0.127708	0.099873

실험 데이터에 대한 성능은 서포트 벡터 머신 방법으로 구축된 최종 모형이 가장 좋은 것으로 나타났고, 그 다음으로 인공 신경망 방법으로 구축된 최종 모형의 성능이 좋은 것으로 나타났다. 두 모형의 RMSE 모두 다중 회귀 모형의 RMSE인 0.127708보다 낮은 것으로 나타난 반면에, 랜덤 포레스트 방법으로 구축된 최종 모형의 RMSE는 다중 회귀 모형의 RMSE 보다 높은 것으로 나타났다. 하지만 MAE를 기준으로 성능을 비교할 시 랜덤 포레스트의 최종 모형이 다중 회귀 모형보다 더 좋은 성능을 보인다고 할 수 있기 때문에 실제 적용 가능성이 떨어진다고는 판단할 수는 없다.

3) Partial Dependence(PD) plot

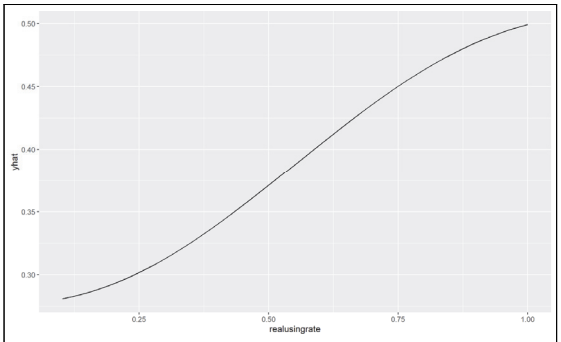
본 연구는 <표 9>에서 가장 성능이 좋은 것으로 확인된 서포트 벡터 머신 모형으로부터 PD plot을 도출하였고, 그 결과는 <그림 3>과 같다.

<그림 3> PD plot

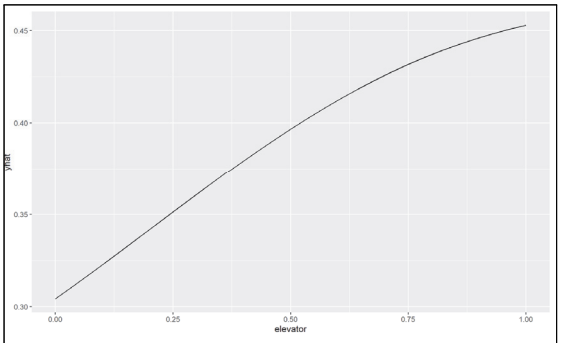


(a) 공실률

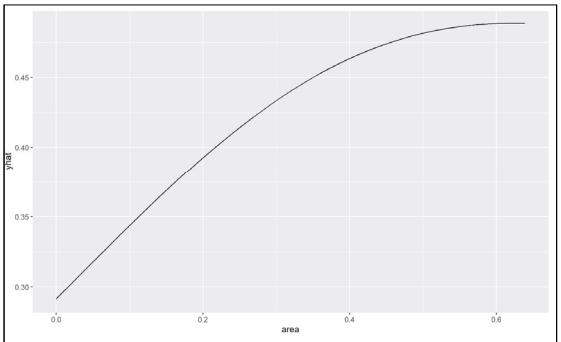
<그림 3> 계속



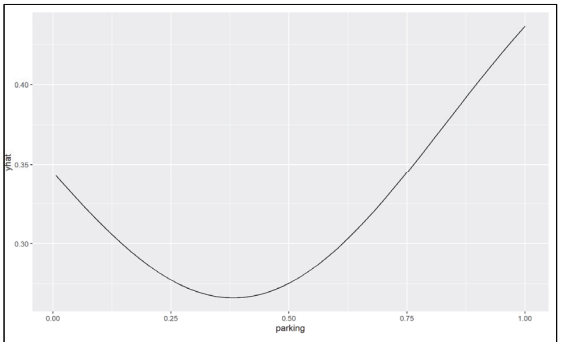
(b) 전용률



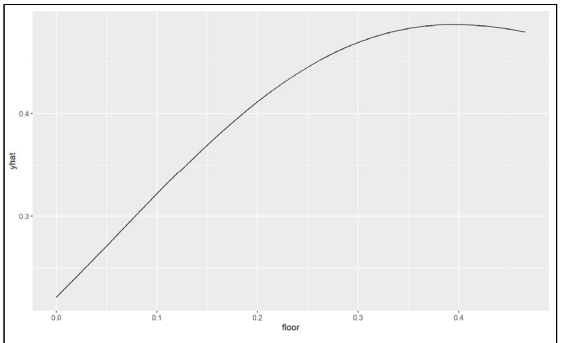
(f) 승강기수



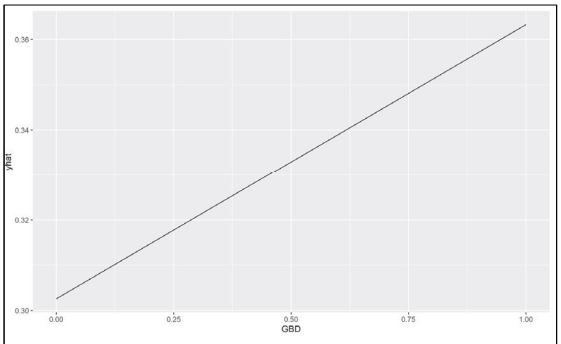
(c) 연면적



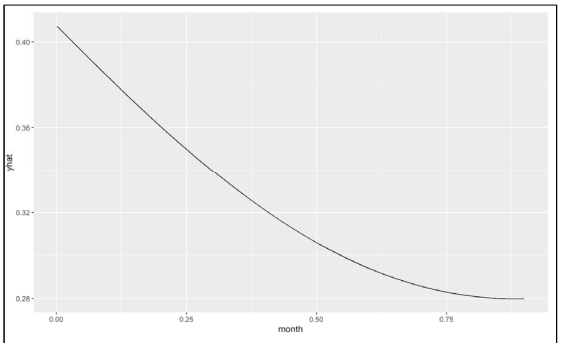
(g) 주차대수



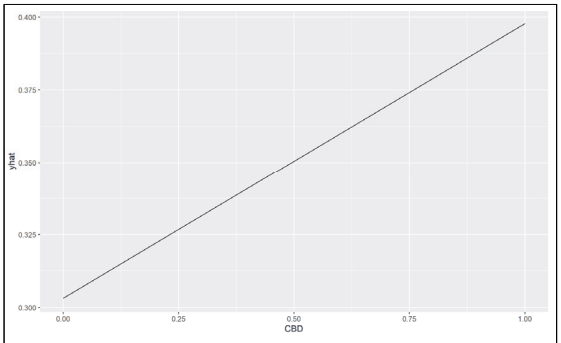
(d) 층수



(h) 강남권역

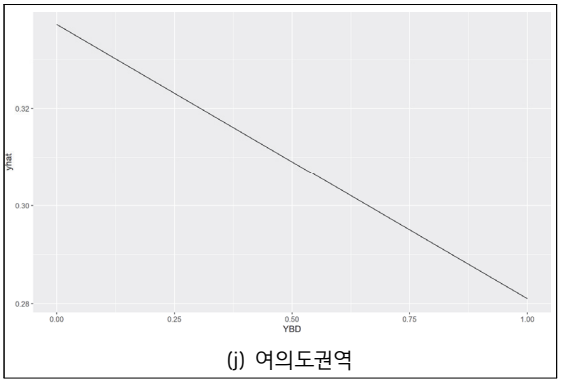


(e) 연식

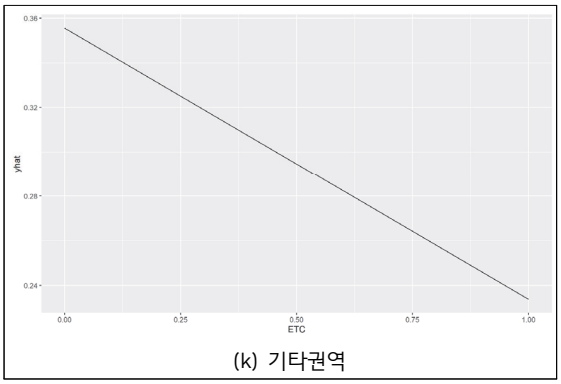


(i) 도심권역

<그림 3> 계속



(j) 여의도권역



(k) 기타권역

먼저, 공실률은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 이는 투입되는 공실률 값이 높을수록 모형이 높은 임대료를 산출한다는 것인데, 현재 서울시 오피스 임대시장에서 건물의 공실이 발생하더라도 임대인들이 물건의 가치 하락을 막기 위해 표면적인 순운영소득을 어느 정도 유지하고자 오히려 호가 임대료를 올리는 그러한 행태를 모형이 학습한 것으로 추정된다. 만약 렌트 프리(rent free) 제공이나 실내 인테리어 공사비 지원(tenant improvement) 등과 같이 공실을 줄이기 위한 여러 노력에 따른 비용을 임대료에서 제한다면 이와 반대되는 결과가 나타날 것으로 예상된다.

전용률은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 이는 임차인이 임대료가 실제 업무공간으로 환산되는 효율이 높은 건물을 선호하기 때문에 이에 대한 지불의사금액이 높게 나타나는 것을 모형이 학습한 것으로 추정된다.

이 외에도 건물의 규모를 나타내는 연면적은 모형의 예측값과 양의 관계, 건물의 랜드마크적 특성을 나타내는 층수는 모형의 예측값과 양의 관계, 건물의 물리적 노후화 정도의 대리변수인 연식은 모형의 예측값과 음의 관계, 건물의 이용 편의성과 관련된 변수인 승강기수는 모형의 예측값과 양의 관계를 갖는 것으로 나타났다. 다만, 주차대수는 예측값과 2차 곡선 형태의 관계를 갖는 것으로 나타났다. 이는 임대인의 입장에서 주차비가 기타수익으로 발생하지만, 주요 주차공간이 지하 주차장인 것을 고려한다면 주차 공간 마련에 따른 공사비 증가 수준이 계속해서 상승하기 때문에 임대료 증가 수준 또한 커지는 것을 모형이 학습한 것으로 추정된다.

권역 변수의 PD plot을 살펴보면, 해당 모형은 강남권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 도심권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 여의도권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게, 기타권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게 산정하는 것으로 나타났다.

<표 10> 다중 회귀 결과

구분		계수	표준오차	P값
상수항		20,761.99***	7,750.59	0.008
공실률		6,407.83	6,519.16	0.326
전용률		44,696.74***	13,026.65	0.001
연면적		0.203	0.223	0.362
층수		574.57***	196.43	0.004
연식		-32.26***	8.43	0.000
승강기수		1,295.27***	400.08	0.001
주차대수		-8.63**	3.67	0.019
녹색건축 인증여부		-969.80	4,348.84	0.824
권역	강남 권역	16,358.81***	2,557.76	0.000
	도심 권역	22,899.91***	2,767.07	0.000
	기타 권역	-7,324.78**	2,979.60	0.015
Adj R^2		0.5409		
관측수		355		
F값		34.85		

주: 1) ***, **, * : 1%, 5%, 10% 이내에서 통계적으로 유의함
2) 전체 데이터 셋 관측치의 70%를 분석에 활용함
3) 권역변수의 기준은 여의도권역이며 이는 생략(omit)됨

본 연구는 추가적으로 PD plot과 다중 회귀 결과를 비교하였다. 다중 회귀 결과는 <표 10>에 제시되어 있다. PD plot에서 나타난 영향 관계의 방향성은 다중 회귀 계수의 부호와 대부분 유사하며 해석에 있어 주목해야 할 부분이나 기존 연구들과의 큰 차이점은 없는 것으로 나타났는데, 이는 구축된 머신 러닝 모형이 데이터로부터 나타나는 오피스 임대 시장의 전반적인 특성을 적절히 반영하고 있음을 의미한다. 하지만 종속변수와 설명변수 간의 선형 관계를 가정하는 다중 회귀 모형에서는 발견할 수 없는 비선형적 관계가 PD plot을 통해 확인되었다. 더미 변수인 권역 변수들을 제외한 모든 특성 변수들은 예측값과 완만한 곡선 형태의 관계를 갖는 것으로 나타났는데, 특히 다중 회귀 결과 주차대수의 추정계수는 음의 값으로 나타난 반면에 머신 러닝 모형의 PD plot에서는 2차 곡선 형태를 보여 다중 회귀 계수로부터 단순히 해당 변수가 임대료와 음의 영향 관계를 갖는다고 해석하기에는 문제가 있는 것으로 확인되었다. 결과적으로 다중 회귀 모형은 선형 가정으로 인해 계수가 과대 또는 과소 추정되어있을 가능성이 있는 반면에 머신 러닝 모형은 이러한 비선형적 관계를 반영하고 있어 상대적으로 성능이 더 우수하게 나타난 것으로 판단된다.

2. 규모 등급별 오피스 임대료 산정 모형 구축 결과

1) 초대형 오피스

초대형 등급 오피스에 대한 임대료 산정 모형의 구축 결과는 <표 11>에 제시되어 있다.

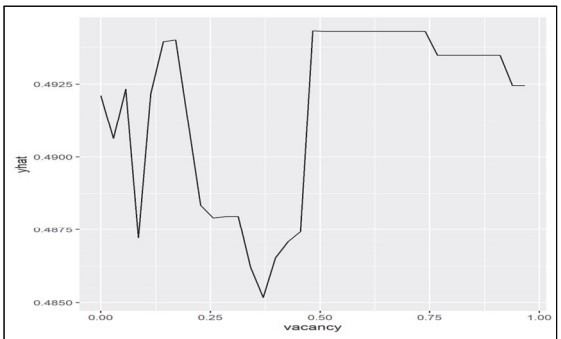
먼저 랜덤 포레스트의 경우 변수 중요도가 낮은 변수들 중 여의도권역, 녹색건축 인증여부, 강남권역 변수를 제외하였을 때 ntree가 200, mtry가 4인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.177713인 것으로 나타났다. 인공 신경망의 경우 변수 중요도가 낮은 변수들 중 연면적, 녹색건축 인증여부를 제외하였을 때 size가 10인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.245050으로 나타났다. 서포트 벡터 머신의 경우 변수 중요도가 낮은 변수들 중 녹색건축 인증여부를 제외하였을 때 ε이 0.5, C가 4, σ가 2^(-5)인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.184918으로 나타났다.

<표 11> 초대형 등급 오피스 임대료 산정 모형 구축 결과

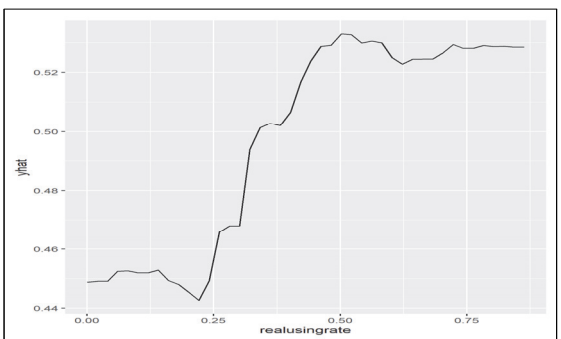
구분	제외변수	모형별 초모수	RMSE	MAE
RF	여의도 권역, 녹색건축 인증여부, 강남권역	ntree=200, mtry=4	0.177713	0.146459
ANN	연면적, 녹색건축 인증여부	size=10	0.245050	0.193475
SVM	녹색건축 인증여부	ε=0.5, C=4, σ=2^(-5)	0.184918	0.137292
다중 회귀	-	-	0.219610	0.170403

머신 러닝 모형별 성능의 순위는 랜덤 포레스트 모형, 서포트 벡터 머신 모형, 인공 신경망 모형 순으로 나타났으며, 인공 신경망 모형은 다른 머신 러닝 모형과 다중 회귀 모형보다 성능이 크게 떨어져 적용 가능성이 낮은 것으로 확인되었다. 이에 본 연구는 가장 성능이 좋은 것으로 확인된 랜덤 포레스트 모형으로부터 PD plot을 도출하였고 그 결과는 <그림 4>와 같다.

<그림 4> PD plot - 초대형 오피스

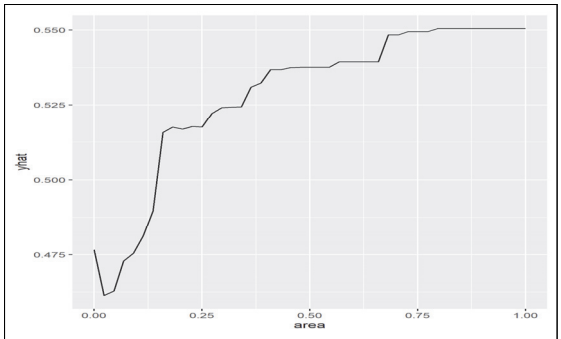


(a) 공실률

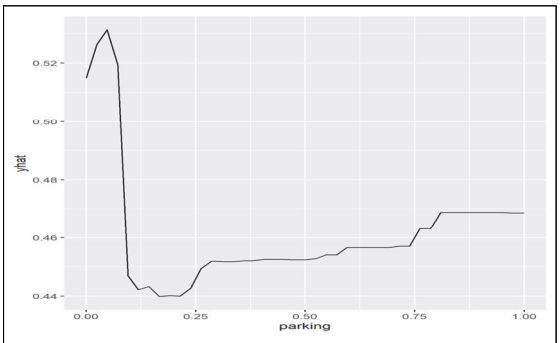


(b) 전용률

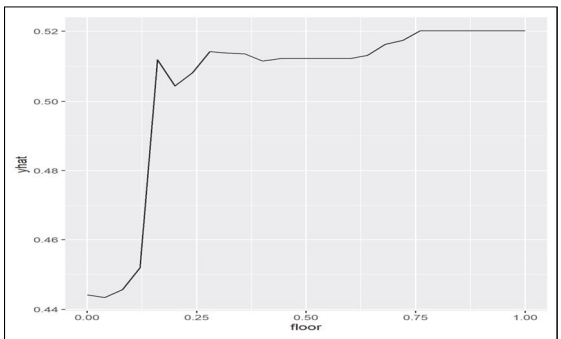
<그림 4> 계속



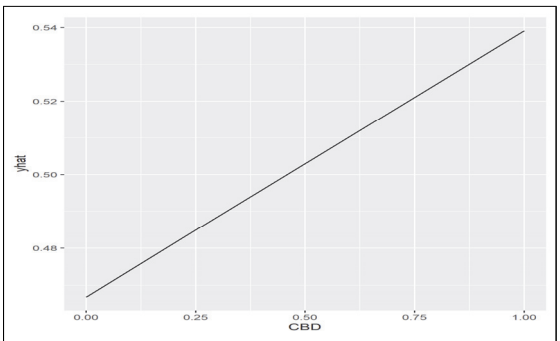
(c) 연면적



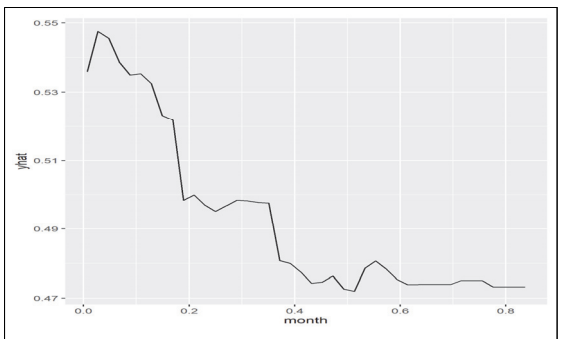
(g) 주차대수



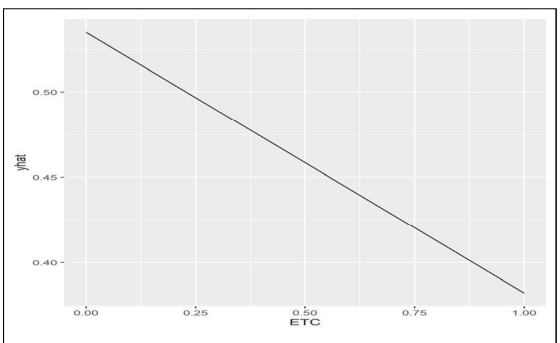
(d) 층수



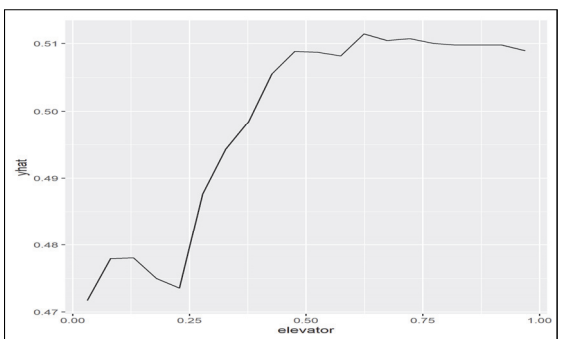
(h) 도심권역



(e) 연식



(i) 기타권역



(f) 승강기수

먼저, 공실률과 모형의 예측값 간의 관계는 특정할 수 없는 것으로 나타났다. 이는 초대형 오피스들의 특질(characteristics)이 상당히 상이하기 때문에 임대시장에서 임대인들이 공실에 대해 각기 다른 반응과 전략을 취하고 있다는 것을 모형이 학습했기 때문이다. 전용률과 모형의 예측값 간의 관계는 구간마다 다르지만 대략적으로 양의 관계를 갖는 것으로 확인되었다. 연면적은 모형의 예측값과 대략적으로 양의 관계를 갖는 것으로 확인되었다. 층수는 모형의 예측값과 대략적으로 양의 관계를 갖는 것으로

확인되었다. 연식은 모형의 예측값과 대략적으로 음의 관계를 갖는 것으로 확인되었다. 승강기수와 모형의 예측값 간의 관계는 구간마다 다르지만 대략적으로 양의 관계를 갖는 것으로 확인되었다. 주차대수와 모형의 예측값 간의 관계는 일정구간에서 음의 관계를 가진 후 다시 대략적으로 양의 관계를 갖는 것으로 확인되었다. 권역 변수의 PD plot을 살펴보면, 해당 모형은 도심권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 기타권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게 산정하는 것으로 나타났다.

2) 대형 오피스

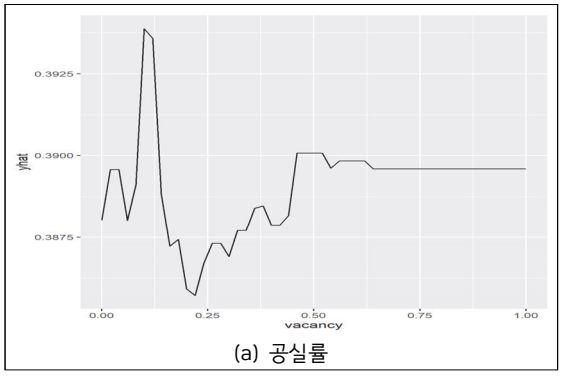
대형 등급 오피스에 대한 임대료 산정 모형의 구축 결과는 <표 12>에 제시되어 있다. 세 가지 머신 러닝 방법 모두 제외한 변수가 없을 때의 최적 모형이 가장 성능이 좋은 것으로 나타났다. 먼저 랜덤 포레스트의 경우 ntree가 50, mtry가 8인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.129731인 것으로 나타났다. 인공 신경망의 경우 size가 5인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.158508인 것으로 나타났다. 서포트 벡터 머신의 경우 ε이 0.1, C가 8, σ가 2^(-7)인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.147432로 나타났다.

<표 12> 대형 등급 오피스 임대료 산정 모형 구축 결과

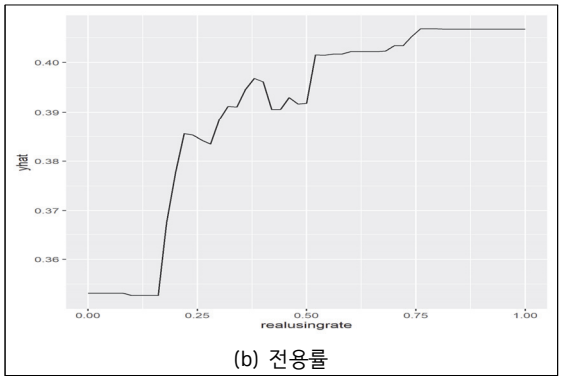
구분	제외변수	모형별 초모수	RMSE	MAE
RF	-	ntree=50, mtry=8	0.129731	0.103064
ANN	-	size=5	0.158508	0.119305
SVM	-	ε=0.1, C=8, σ=2^(-7)	0.147432	0.112890
다중 회귀	-	-	0.187580	0.130274

머신 러닝 모형별 성능의 순위는 랜덤 포레스트 모형, 서포트 벡터 머신 모형, 인공 신경망 모형 순으로 나타났으며, 세 모형 모두 다중 회귀 모형보다 성능이 좋은 것으로 확인되었다. 특히 랜덤 포레스트 모형의 성능은 다른 모형의 성능보다 크게 앞서는 것으로 나타났다. 이에 본 연구는 가장 성능이 좋은 것으로 확인된 랜덤 포레스트 모형으로부터 PD plot을 도출하였고 그 결과는 <그림 5>와 같다.

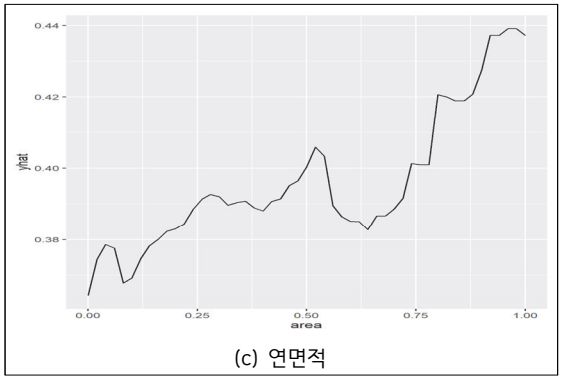
<그림 5> PD plot - 대형 오피스



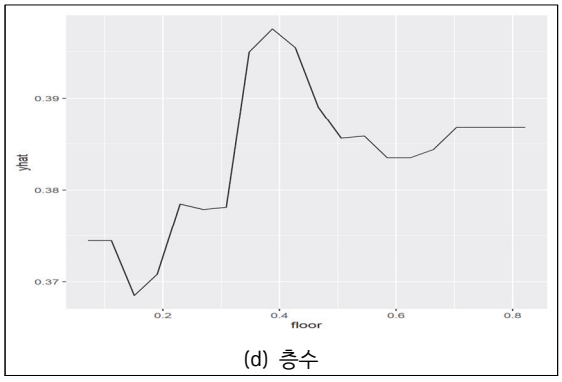
(a) 공실률



(b) 전용률

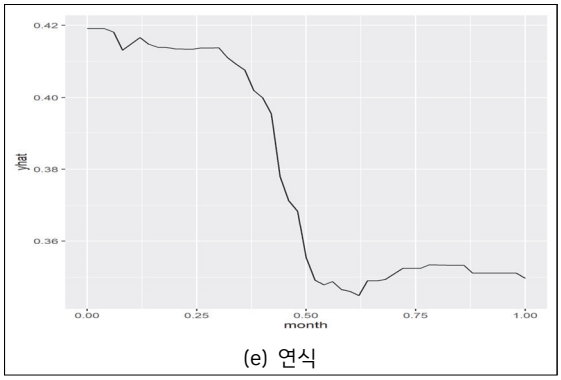


(c) 연면적

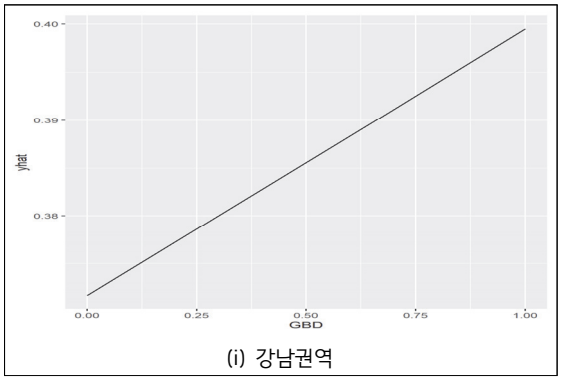


(d) 층수

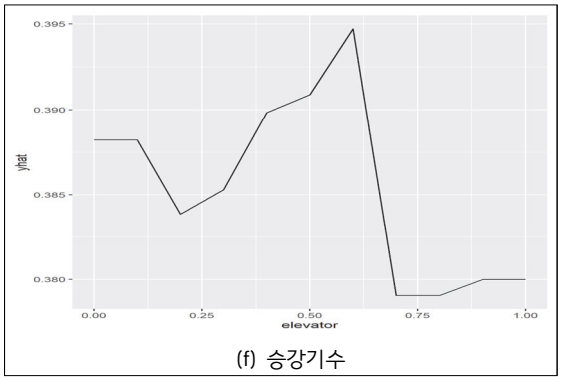
<그림 5> 계속



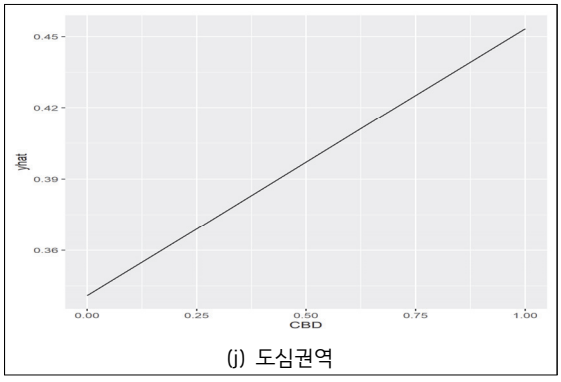
(e) 연식



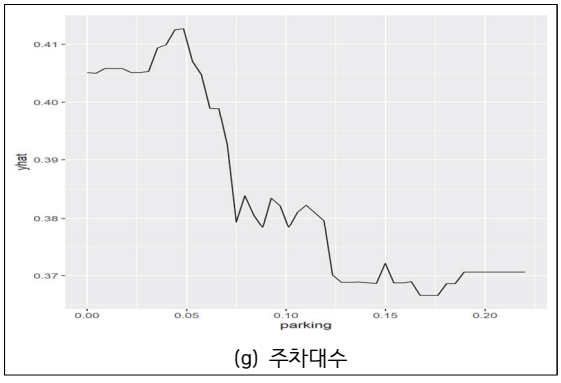
(i) 강남권역



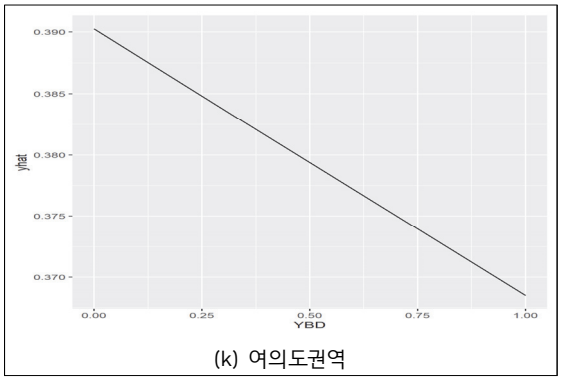
(f) 승강기수



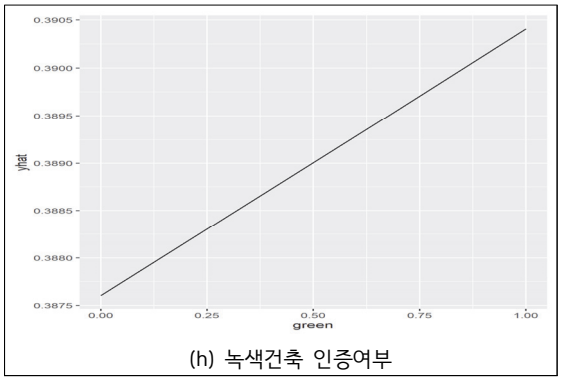
(j) 도심권역



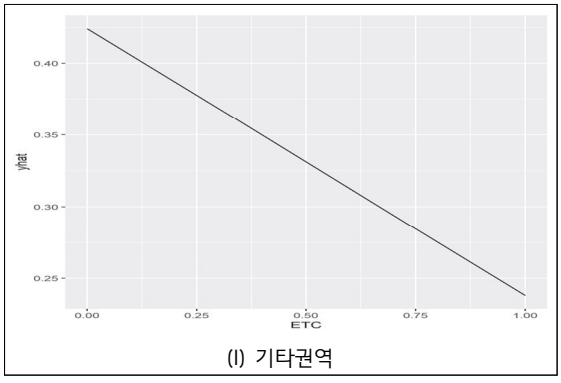
(g) 주차대수



(k) 여의도권역



(h) 녹색건축 인증여부



(l) 기타권역

먼저, 공실률과 모형의 예측값 간의 관계는 특정할 수 없는 것으로 나타났다. 이는 초대형 오피스와 마찬가지로 대형 오피스의 임대인들이 물건의 공실에 대해 저마다 상이한 태도를 보이고 있다는 것을 의미한다. 전용률은 모형의 예측값과 대략적으로 양의 관계를 갖는 것으로 확인되었다. 연면적과 모형의 예측값 간의 관계는 음인 구간도 몇몇 존재하지만 대략적으로 양의 관계를 갖는 것으로 확인되었다. 층수와 모형의 예측값 간의 관계는 대략적으로 양의 관계를 보이다 약 0.387218지점 이후 음의 관계를 보이는 것으로 나타났다. 연식은 모형의 예측값과 대략적으로 음의 관계를 갖는 것으로 확인되었다. 승강기수와 모형의 예측값 간의 관계는 대략적으로 양의 관계를 보이다 0.6지점에서 단절이 발생하는 것으로 나타났다. 주차대수는 모형의 예측값과 대략적으로 음의 관계를 갖는 것으로 확인되었다. 해당 모형은 녹색건축 인증 오피스의 임대료를 비인증 오피스의 임대료보다 더 높게 산정하는 것으로 나타났다. 권역 변수의 PD plot을 살펴보면, 해당 모형은 강남권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 도심권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 여의도권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게, 기타권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게 산정하는 것으로 나타났다.

3) 중대형 오피스

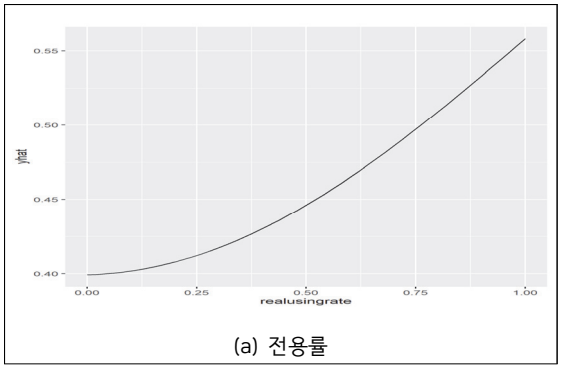
중대형 등급 오피스에 대한 임대료 산정 모형의 구축 결과는 <표 13>에 제시되어 있다. 먼저 랜덤 포레스트의 경우 변수 중요도가 낮은 변수들 중 녹색건축 인증여부, 도심권역 변수를 제외하였을 때 ntree가 100, mtry가 10인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.161520인 것으로 나타났다. 인공 신경망의 경우 변수 중요도가 낮은 변수들 중 녹색건축 인증여부, 주차대수를 제외하였을 때 size가 5인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.174484인 것으로 나타났다. 서포트 벡터 머신의 경우 변수 중요도가 낮은 변수들 중 공실률, 녹색건축 인증여부를 제외하였을 때 ε이 0.3, C가 8, σ가 2^(-6)인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.148856로 나타났다.

<표 13> 중대형 등급 오피스 임대료 산정 모형 구축 결과

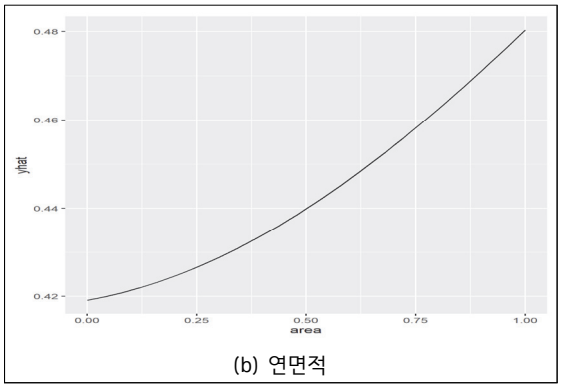
구분	제외변수	모형별 초모수	RMSE	MAE
RF	녹색건축 인증여부, 도심권역	ntree=100, mtry=10	0.161520	0.130699
ANN	녹색건축 인증여부, 주차대수	size=5	0.174484	0.136855
SVM	공실률, 녹색건축 인증여부	ε=0.3, C=8, σ=2^(-6)	0.148856	0.121768
다중 회귀	-	-	0.202637	0.145645

머신 러닝 모형별 성능의 순위는 서포트 벡터 머신 모형, 랜덤 포레스트 모형, 인공 신경망 모형 순으로 나타났으며, 세 모형 모두 다중 회귀 모형보다 성능이 좋은 것으로 확인되었다. 특히 서포트 벡터 머신 모형의 성능은 다른 모형의 성능보다 크게 앞서는 것으로 나타났다. 이에 본 연구는 가장 성능이 좋은 것으로 확인된 서포트 벡터 머신 모형으로부터 PD plot을 도출하였고 그 결과는 <그림 6>과 같다.

<그림 6> PD plot - 중대형 오피스

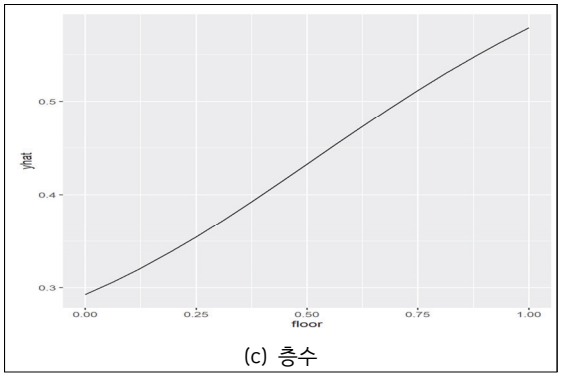


(a) 전용률

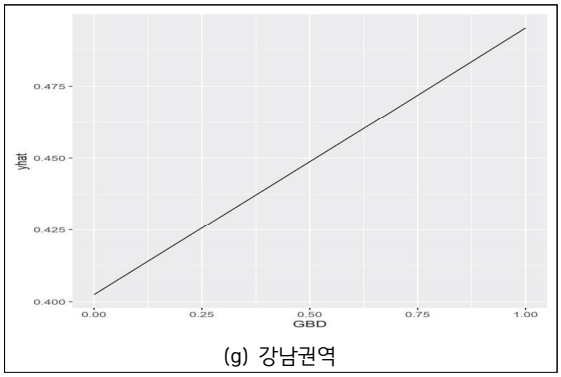


(b) 연면적

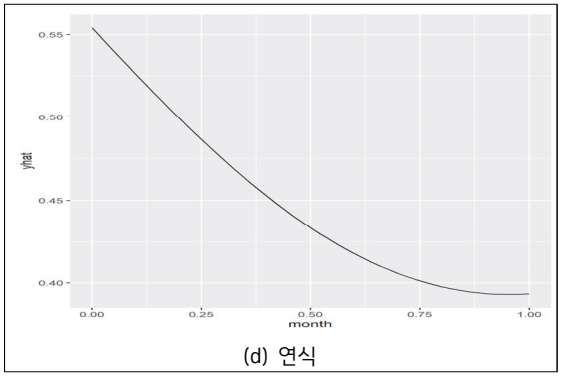
<그림 6> 계속



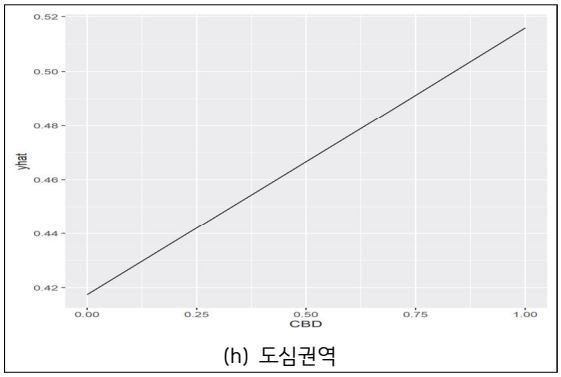
(c) 층수



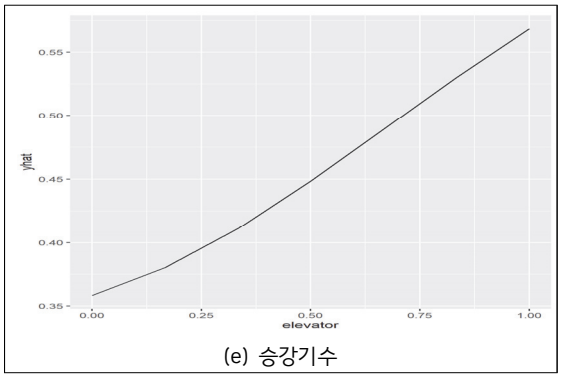
(g) 강남권역



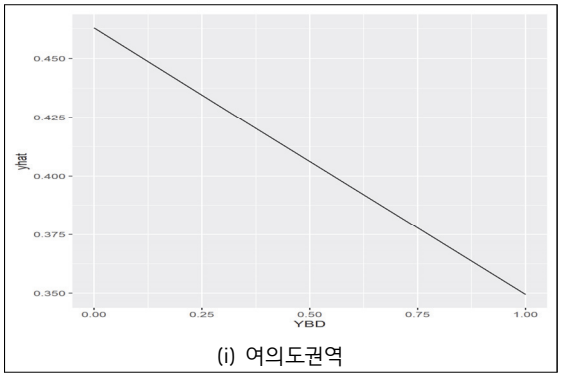
(d) 연식



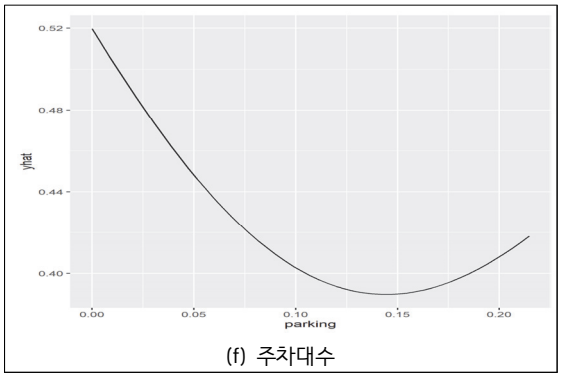
(h) 도심권역



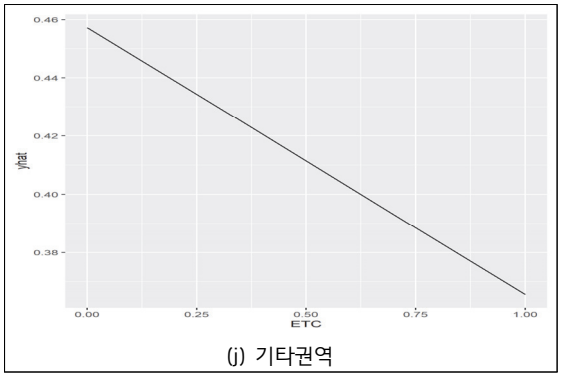
(e) 승강기수



(i) 여의도권역



(f) 주차대수



(j) 기타권역

먼저, 전용률은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 연면적은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 층수는 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 연식은 모형의 예측값과 음의 관계를 갖는 것으로 확인되었다. 승강기수는 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 주차대수는 모형의 예측값과 2차 곡선 형태의 관계를 갖는 것으로 확인되었다. 권역 변수의 PD plot을 살펴보면, 해당 모형은 강남권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 도심권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 여의도권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게, 기타권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게 산정하는 것으로 나타났다.

4) 중형 오피스

중형 등급 오피스에 대한 임대료 산정 모형의 구축 결과는 <표 14>에 제시되어 있다.

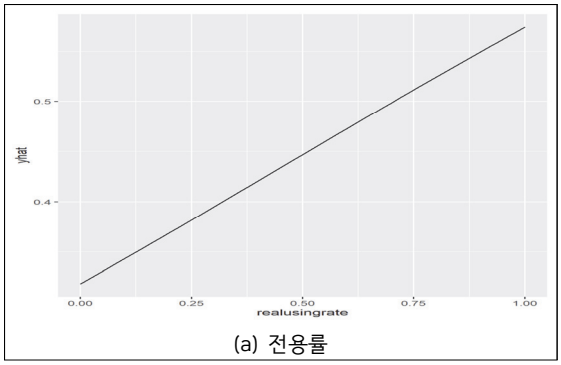
<표 14> 중형 등급 오피스 임대료 산정 모형 구축 결과

구분	제외변수	모형별 초모수	RMSE	MAE
RF	녹색건축 인증여부	ntree=500, mtry=2	0.140644	0.121333
ANN	주차대수	size=50	0.206544	0.166937
SVM	주차대수, 녹색건축 인증여부, 승강기수, 공실률	ε=0.1, C=16, σ=2^(-9)	0.133084	0.107672
다중 회귀	-	-	0.141731	0.112353

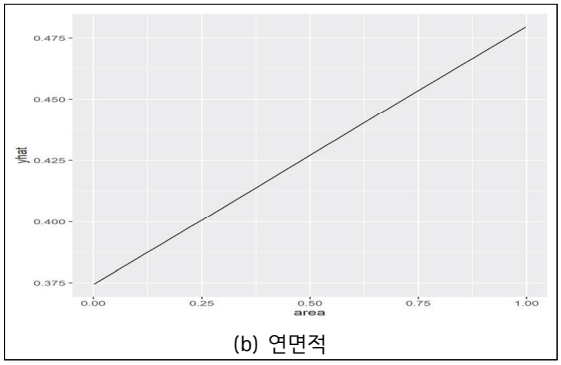
먼저 랜덤 포레스트의 경우 변수 중요도가 낮은 변수들 중 녹색건축 인증여부를 제외하였을 때 ntree가 500, mtry가 2인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.140644인 것으로 나타났다. 인공 신경망의 경우 변수 중요도가 낮은 변수들 중 주차대수를 제외하였을 때 size가 50인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.206544인 것으로 나타났다. 서포트 벡터 머신의 경우 변수 중요도가 낮은 변수들 중 주차대수, 녹색건축 인증여부, 승강기수, 공실률을 제외하였을 때 ε이 0.1, C가 16, σ가 2^(-9)인 최적 모형의 성능이 가장 좋은 것으로 확인되었으며, 해당 모형의 RMSE 값은 약 0.133084로 나타났다.

머신 러닝 모형별 성능의 순위는 서포트 벡터 머신 모형, 랜덤 포레스트 모형, 인공 신경망 모형 순으로 나타났으며, 인공 신경망 모형은 다른 머신 러닝 모형과 다중 회귀 모형보다 성능이 크게 떨어져 적용 가능성이 낮은 것으로 확인되었다. 이에 본 연구는 가장 성능이 좋은 것으로 확인된 서포트 벡터 머신 모형으로부터 PD plot을 도출하였고 그 결과는 <그림 7>과 같다.

<그림 7> PD plot - 중형 오피스

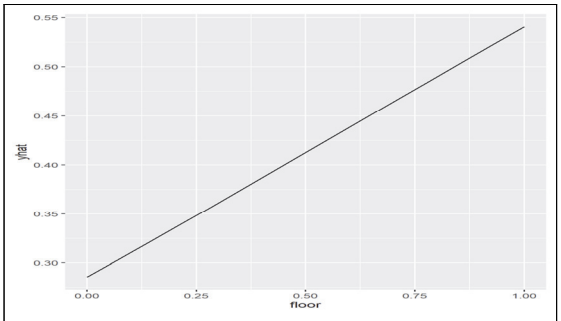


(a) 전용률

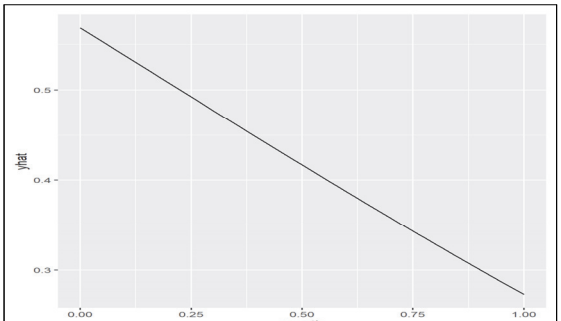


(b) 연면적

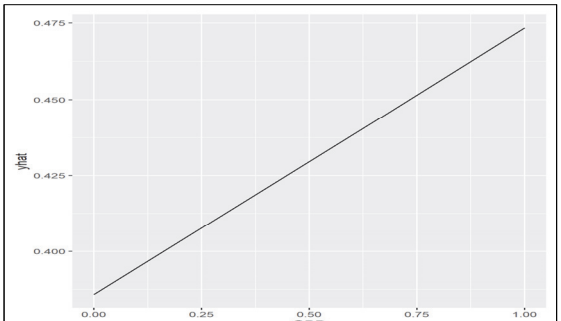
<그림 7> 계속



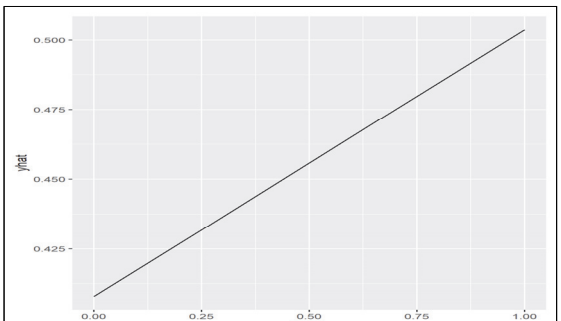
(c) 층수



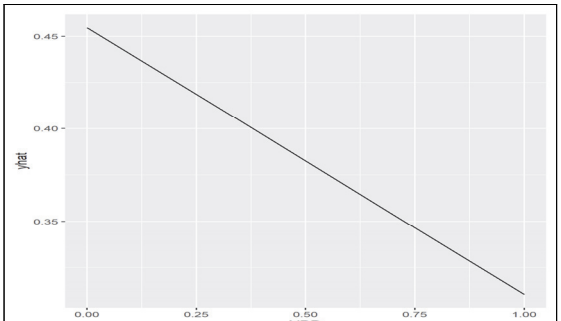
(d) 연식



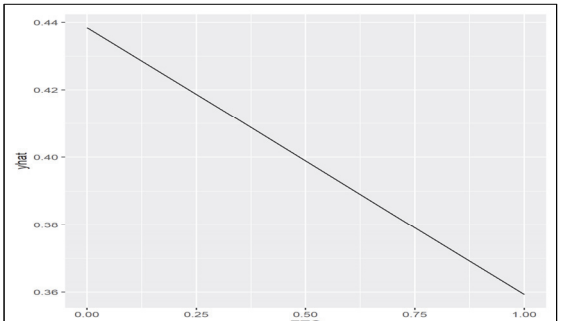
(e) 강남권역



(f) 도심권역



(g) 여의도권역



(h) 기타권역

먼저, 전용률은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 연면적은 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 층수는 모형의 예측값과 양의 관계를 갖는 것으로 확인되었다. 연식은 모형의 예측값과 음의 관계를 갖는 것으로 확인되었다. 권역 변수의 PD plot을 살펴보면, 해당 모형은 강남권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 도심권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 높게, 여의도권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게, 기타권역에 소재한 오피스의 임대료를 그 외 오피스의 임대료보다 더 낮게 산정하는 것으로 나타났다.

<표 15>는 등급별 모형별 PD plot으로부터 확인 가능한 예측값과 투입변수 간 영향 관계를 종합 · 비교한 표이다. 먼저 공실률의 경우 PD plot이 진동하는 형태를 보여 예측값과의 관계를 특정할 수 없거나 사전에 속성 선택 과정에서 제외되었다. 이는 공실에 대해 임대인들이 각기 상이한 전략을 취하고 있기 때문이며 본 연구에서 사용한 분석자료의 임대료 정보가 호가

<표 15> 등급별 모형별 PD plot 결과 종합

변수명		초대형 오피스	대형 오피스	중대형 오피스	중형 오피스
공실률		특정할 수 없음	특정할 수 없음	제외	제외
전용률		+	+	+	+
연면적		+	+	+	+
층수		+	+ → -	+	+
연식		-	-	-	-
승강기수		+	+ 이후 단절	+	제외
주차대수		- → +	-	- → +	제외
녹색건축 인증여부		제외	+	제외	제외
권역	강남 권역	제외	+	+	+
	도심 권역	+	+	+	+
	여의도 권역	제외	-	-	-
	기타 권역	-	-	-	-

임대료라는 한계로부터 기인한 것으로 추정된다. 임차인이 지불하는 임대료가 실제 업무공간으로 치환되는 정도를 나타내는 전용률과 건물의 규모를 나타내는 연면적은 네 모형에서 모두 예측값과 양의 관계를 갖는 것으로 확인되었고 이는 기존 연구들의 분석 결과 및 직관에 어긋나지 않는 결과이다. 층수의 경우 대형 오피스 모형을 제외하고 예측값과 양의 관계를 갖는 것으로 나타났고 대형 오피스 모형에서는 양의 관계를 보이다 특정 지점 이후 음의 관계를 보이는 것으로 나타났다. 이는 대형 오피스가 중대형 오피스와 중형 오피스에 비해 건물 규모가 더 크기 때문에 점유자(occupier) 수가 많아 층수 증가에 따른 랜드마크적 효과보다는 승강기 이용 증가에 따른 불편성이 임차 수요에 더 큰 영향을 미치고 있음을 모형이 학습한 것으로 추정되며, 초대형 오피스 모형의 PD plot에서도 초기 구간을 제외하고는 층수 증가에 따른 예측값 증가 오름세가 크지 않은 것도 이러한 점에 연유하는 것으로 추정된다. 건물의 노후화 정도를 나타내는 연식은 네 모형에서 모두 예측값과 음의 관계를 갖는 것으로 확인되었고 이 또한 기존 연구들의 분석 결과 및 직관에 어긋나지 않는 결과이다. 승강기수는 초대형 오피스 모형과 중대형 오피스 모형에서 예측값과 양의 관계를 갖는 것으로 확인되었다. 대형 오피스 모형의 경우 승강기수는 예측값과 양의 관계를 보이다 약 0.39지점에서 단절이 발생하였고 이를 역변환하여 승강기수 약 7대에서 단절이 이루어졌다고 볼 수 있다. 이는 적정 수준 이상의 승강기수는 대형 오피스에 대한 임차 수요에 영향을 미치지 못하며 초대형 오피스 모형의 승강기수 PD plot에서도 이러한 점을 발견할 수 있다. 중형 오피스의 경우 층수가 상대적으로 높지 않기 때문에 승강기수 변수의 중요도가 떨어져 속성 선택 과정에서 제외되었다. 주차대수의 경우 초대형 오피스 모형과 중대형 오피스 모형에서는 예측값과 음의 관계를 보이다 양의 관계로 전환되는 것으로 나타났고, 이는 임대인에게 주차비가 기타수익으로 발생하는 반면에 주차 공간 마련을 위한 지하 주차장 공사비 증가 수준이 계속해서 상승하여 임대료 수준 또한 상승하기 때문인 것으로 추정된다. 대형 오피스 모형에서는 주차대수가 전반적으로 예측값과 음의 관계를 보였지만 PD plot상 초기 구간의 하락세 이후 큰 하락세가 나타나지 않는 것도 이러한 점에서 연유하는 것으로 추정된다. 중형 오피스의 경우 주차대수 변수의 중요도가 떨어져 속성 선택 과정에서 제외되었다. 녹색건축 인증여부 변수는 대형 오피스 모형을 제외한 세 모형에서 변수의 중요도가 떨어지는 것으로 나타나 속성 선택 과정에서 제외되었다. 대형 오피스 모형에서는 녹색건축 인증여부가 예측값과 양의 관계를 갖는 것으로 나타났는데, 이는 대형 오피스 임대 시장에서는 친환경 건물에서 나타나는 에너지 효율성에 따른 관리비 감소분을 임대인들이 차지하고 있음을 의미한다. 권역 변수의 경우 공통적으로 강남권역 변수는 예측값과 양의 관계를, 도심권역 변수는 예측값과 양의 관계를, 여의도권역 변수는 예측값과 음의 관계를, 기타권역 변수는 예측값과 음의 관계를 갖는 것으로 나타났다. 다만 초대형 오피스 임차 수요자들의 경우 권역을 벗어나는 이동이 늘고 있어 강남권역, 여의도권역에 위치한 것에 따른 임대료 프리미엄은 존재하지 않는 것으로 나타났다.

Ⅴ. 결론

전통적인 오피스 임대료 산정 방식인 임대사례비교법과 적산법은 평가자의 주관이 다소 반영된다는 문제점이 있어 이와 상호 보완할 객관적인 산정 방법의 필요가 제기되고 있다. 이에 본 연구는 최근 주거용 부동산을 대상으로 한 AVM 관련 연구에서 보편적으로 활용되고 있는 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신을 사용하여 오피스 임대료 산정 모형을 구축하고 이들의 적용 가능성을 검토하였다.

먼저, 표본 전체 507동의 오피스를 대상으로 하여 각 방법별로 초모수를 다양하게 조절함으로써 최적 모형을 구축한 결과, 최적 모형별 성능의 순위는 서포트 벡터 머신 모형, 랜덤 포레스트 모형, 인공 신경망 모형 순으로 나타났다. 이어 각 최적 모형별로 변수 중요도가 상대적으로 낮은 변수들을 제외하는 속성 선택을 수행하였고, 그 결과 인공 신경망의 성능 개선 정도가 가장 크게 발생한 반면에 랜덤 포레스트의 성능 개선 정도는 미미하여 성능의 순위가 서포트 벡터 머신 모형, 인공 신경망 모형, 랜덤 포레스트 모형 순으로 바뀌었다. 다만 랜덤 포레스트 모형의 경우 다중 회귀 모형보다 RMSE가 다소 높은 것으로 나타났는데, MAE는 더 낮은 것으로 나타나 실제 적용 가능성이 떨어진다고 판단할 수는 없었다.

또한 모형의 학습 결과를 해석하기 위해 가장 성능이 좋은 것으로 나타난 서포트 벡터 머신 모형을 대상으로 PD plot을 도출하였다. 해당 모형은 공실률이 높을수록, 전용률이 높을수록, 연면적이 클수록, 층수가 높을수록, 연식이 낮을수록, 승강기수가 많을수록 임대료를 높게 산출하는 것으로 나타났고, 주차대수가 높을수록 임대료를 낮게 산출하다 그 기울기가 점차 증가하여 특정 지점 이후 높게 산출하는 2차 곡선의 형태를 보이는 것으로 나타났다. 그리고 강남권역과 도심권역에 소재한 오피스의 임대료를 더 높게 산출하는 것으로, 여의도권역과 기타권역에 소재한 오피스의 임대료를 더 낮게 산출하는 것으로 나타났다.

추가적으로 표본을 초대형, 대형, 중대형, 중형 등급으로 나누어 각 등급별 임대료 산정 모형을 각각 구축하였다. 초대형 등급의 경우 인공 신경망 모형은 성능이 크게 떨어져 적용 가능성이 낮은 것으로 확인되었고, 랜덤 포레스트 모형의 성능이 가장 우수한 것으로 나타났다. 본 모형에서 전용률, 연면적, 층수, 승강기수는 예측값과 대략적으로 양의 관계를, 연식은 예측값과 대략적으로 음의 관계를 갖는 것으로 나타났고, 주차대수는 음의 관계를 가진 후 다시 대략적으로 양의 관계를 갖는 것으로 나타났다. 공실률과 예측값 간의 관계는 특정할 수 없었다. 그리고 도심권역에 소재한 오피스의 임대료를 더 높게 산출하는 것으로, 기타권역에 소재한 오피스의 임대료를 더 낮게 산출하는 것으로 나타났다.

대형 등급의 경우 세 가지 머신 러닝 모형 모두 다중 회귀 모형보다 성능이 좋은 것으로 나타났고, 특히 랜덤 포레스트 모형의 성능은 다른 모형의 성능보다 크게 앞서는 것으로 나타났다. 본 모형에서 전용률과 연면적은 예측값과 대략적으로 양의 관계를, 연식과 주차대수는 음의 관계를 갖는 것으로 나타났다. 층수는 예측값과 대략적으로 양의 관계를 보이다 특정 지점 이후 음의 관계를 갖는 것으로 나타났고, 승강기수는 예측값과 대략적으로 양의 관계를 보이다 특정 지점에서 단절이 발생하는 것으로 나타났다. 공실률과 예측값 간의 관계는 특정할 수 없었다. 그리고 녹색건축 인증 오피스의 임대료를 더 높게 산출하는 것으로, 강남권역과 도심권역에 소재한 오피스의 임대료를 더 높게 산출하는 것으로, 여의도권역과 기타권역에 소재한 오피스의 임대료를 더 낮게 산출하는 것으로 나타났다.

중대형 등급의 경우 세가지 머신 러닝 모형 모두 다중 회귀 모형보다 성능이 좋은 것으로 나타났고, 특히 서포트 벡터 머신 모형의 성능은 다른 모형의 성능보다 크게 앞서는 것으로 나타났다. 본 모형에서 전용률, 연면적, 층수, 승강기수는 예측값과 양의 관계를, 연식은 음의 관계를 갖는 것으로 나타났고, 주차대수는 2차 곡선 형태의 관계를 갖는 것으로 나타났다. 그리고 강남권역과 도심권역에 소재한 오피스의 임대료를 더 높게 산출하는 것으로, 여의도권역과 기타권역에 소재한 오피스의 임대료를 더 낮게 산출하는 것으로 나타났다.

중형 등급의 경우 인공 신경망 모형은 성능이 크게 떨어져 적용 가능성이 낮은 것으로 확인되었고, 서포트 벡터 머신 모형의 성능이 가장 우수한 것으로 나타났다. 본 모형에서 전용률, 연면적, 층수는 예측값과 양의 관계를, 연식은 음의 관계를 갖는 것으로 나타났다. 그리고 강남권역과 도심권역에 소재한 오피스의 임대료를 더 높게 산출하는 것으로, 여의도권역과 기타권역에 소재한 오피스의 임대료를 더 낮게 산출하는

것으로 나타났다.

본 연구의 의의는 다음과 같다. 첫째, 다양한 머신 러닝 방법을 사용하여 오피스 임대료 산정 모형을 구축하고자 하였다는 점이다. 회귀 문제에 적용이 가능한 머신 러닝 방법들을 이용하여 주거용 부동산의 가치 산정을 시도한 연구들이 최근 활발히 이루어졌던 반면에 상업용 부동산을 대상으로 한 관련 연구는 찾아보기 힘든 실정이다. 이에 본 연구는 객관적으로 오피스 임대료를 산정할 방법으로 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신의 적용 가능성을 검토하고자 하였다는 점에서 의의를 갖는다. 시장 참여자들은 다양한 머신 러닝 방법을 통해 임대료를 산정하고, 이를 전통적인 방법으로 산정된 임대료와 적절히 상호 보완하여 활용함으로써 다양한 의사결정을 내릴 수 있을 것이다. 둘째, PD plot을 통해 모형의 학습 결과에 대한 해석을 시도하였다는 점이다. 기존의 선행연구들이 대부분 예측 중심의 연구였다면, 본 연구는 PD plot을 통해 모형의 예측값과 특정 투입변수 간의 관계를 시각적으로 나타내어 이를 해석함으로써 학습이 제대로 이루어졌는지 확인하였다는 점에서 의의를 갖는다. 마지막으로, 규모 등급별 임대료 산정 모형을 각각 구축하여 그 결과를 제시하였다는 점이다. 본 연구는 건물의 규모에 따라 시장 특성이 세분화되고 있다는 것에 착안하여 규모를 기준으로 하위시장을 나누고 각 하위시장별로 서로 다른 산정 모형을 구축하고자 하였다. 그 결과, 각 등급별로 가장 좋은 성능을 보였던 방법과 그에 대한 초모수 및 제외변수가 각기 상이하였고 PD plot 또한 차이를 보였다는 점을 통해 규모 등급별 분석의 필요성을 입증하였다는 점에서 의의를 갖는다.

본 연구는 다양한 머신 러닝 방법 중 3가지 알고리즘 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신을 활용하여 오피스 임대료 산정 모형을 구축하고자 하였다. 하지만 최근 그래디언트 부스티드 트리, XGBoost, MARS, DNN 등 새로운 머신 러닝 알고리즘들이 속속히 등장하고 있기 때문에 이러한 신규 방법론들의 적용 가능성을 검토하는 후속 연구가 필요하다.

또한 본 연구에서 모형의 해석에 활용한 PD plot은 방법론 자체의 한계점이 존재한다. PD값은 고정된 관심 변수 x_S에 대해 모든 $x_C^{(i)}$의 속성 공간에서의 예측값의 평균값이다. 즉 PD값을 플롯함으로써 도출된 그래프인 PD plot은 평균 효과만을 나타내기 개별 관측치별 이질적 효과를 고려하지 못하고, 평균을 취하는 과정에서 효과가 상쇄되었을 가능성 또한 존재한다. 따라서 ICE plot(Individual Conditional Expectation Plot)을 통해 관측치별 이질적인 효과를 검토하는 후속 연구 또한 필요하다.

머신 러닝 방법은 데이터 특성에 따라 최적 모형의 초모수와 성능이 상당히 상이할 수 있다. 즉 연구자마다 최적 모형을 구축하기위한 일련의 과정들을 반드시 거쳐야 하며, 이에 각 방법별로 변수 중요도, 성능 순위, PD plot의 결과 또한 달라질 수 있다는 한계점을 가진다. 따라서 연구자들은 초모수를 최대한으로 세밀하게 조절해가며 모형을 구축해야 할 것이다. 또한 각 머신 러닝 방법별로 특성이 다르기 때문에 성능을 기준으로 최종 모형을 선택하는 것에서 더 나아가, 모형별 결과를 가중평균 하는 등 여러 방법들의 결과를 함께 적절히 활용하여 의사결정을 내릴 필요가 있다.

논문접수일 : 2020년 2월 21일
논문심사일 : 2020년 2월 24일
게재확정일 : 2020년 4월 6일

참고문헌

1. 김선주 · 이상엽, “오피스 임대료 결정 모형에 관한 연구 – 회귀분석과 신경망 이론을 중심으로 –”, 「지역연구」 제24권 제2호, 2008, pp. 3-26
2. 김성진 · 유은정 · 정민규 · 김재경 · 안현철, “감정예측모형의 성과개선을 위한 Support Vector Regression 응용”, 「지능정보연구」 제18권 제3호, 2012, pp. 185-202
3. 김의준 · 김용환, “서울시 오피스 임대료 결정요인의 변화분석”, 「지역연구」 제22권 제2호, 2006, pp. 79-96
4. 김태훈 · 홍한국, “회귀모형과 신경망모형을 이용한 아파트 가격 모형에 관한 연구”, 「The Korea Spatial Planning Review」 제43권, 2004, pp. 183-200
5. 나성호 · 김종우, “공공데이터를 활용한 아파트 매매 가격 결정 모형의 예측능력 비교”, 「한국지적정보학회지」 제21권 제1호, 2019, pp. 3-12
6. 남영우 · 이정민, “아파트시장예측을 위한 신경망분석 적용가능성에 대한 연구”, 「한국건설관리학회논문집」 제7권 제2호, 2006, pp. 162-170
7. 문근식 · 최재규 · 이현석, “데이터마이닝기법을 활용한 강남구 중소형 오피스빌딩의 매매가격 결정요인 분석”, 「한국콘텐츠학회논문지」 제15권 제7호, 2015, pp. 414-427
8. 배성완 · 유정석, “기계 학습을 이용한 공동주택 가격 추정: 서울 강남구를 사례로”, 「부동산학연구」 제24권 제1호, 2018, pp. 69-85
9. 배성완 · 유정석, “표본 주택 가격 기반 부동산 가격지수 산정: 머신 러닝 방법의 활용을 중심으로”, 「주택연구」 제26권 제4호, 2018, pp. 53-74
10. 변기영 · 이창수, “서울시 오피스 임대료 결정구조에 관한 연구”, 「국토계획」 제39권 제3호, 2004, pp. 205-219
11. 양영준 · 오세준, “서울시 오피스의 임대료 결정요인 분석 – 호가임대료와 실질임대료를 대상으로–”, 「부동산학보」 제71권, 2017, pp. 134-146
12. 오지훈 · 김정섭, “머신러닝을 활용한 서울시 아파트 물리적 특성 변수들의 비선형 영향 분석: 다변량 적응 회귀 스플라인 모형의 적용”, 「감정평가학 논집」 제17권 제3호, 2018, pp. 5-26
13. 윤여신, “오피스 빌딩의 등급 표준화 필요성에 관한 소고”, 「부동산연구원 부동산포커스」 제54권, 2012, pp. 39-50
14. 이재우, “빌딩규모 구분에 의한 서울 오피스시장 현황과 특성차이”, 「부동산연구」 제15권 제2호, 2005, pp. 51-65
15. 이창로 · 박기호, “단독주택가격 추정을 위한 기계학습 모형의 응용”, 「대한지리학회지」 제51권 제2호, 2016, pp. 219-233
16. 장상규 · 김민철, “오피스 빌딩 임대관리 Insight”, 보성인쇄기획, 2018
17. Breiman, L., “Random forests”, *Machine Learning*, vol. 45 no. 1, 2001, pp. 5-32
18. Deng, N., Tian, Y., and Zhang, C., “Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions”, CRC Press, 2012
19. Fisher, A., Rudin, C., and Dominici, F., “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously”, *Journal of Machine Learning Research*, vol. 20 no. 177, 2019, pp. 1-81
20. Friedman, J. H., “Greedy function approximation: A gradient boosting machine”, *Annals of Statistics*, vol. 29 no. 5, 2001, pp. 1189-1232
21. Hastie, T., Tibshirani, R., and Friedman, J., “The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition”, Springer Series in Statistics, 2009
22. Heaton, J., “Introduction to Neural Networks with Java, 2nd Edition”, Heaton Research, 2008
23. Johnson, L., AVMs: An international comparison of the opportunities and challenges, In 26th Annual European Real Estate Society Conference, 2019
24. Kim, H. H., Kang, S. H., Park, J. H., Ha, H. H., and Lim, D. H., “Noise Removal using Support Vector Regression in Noisy Document Images”, *Korean Journal of Applied Statistics*, vol. 25 no. 4, pp. 2012, pp. 669-680
25. Kok, N., Koponen, E. L., and Martinez-Barbosa, C. A., “Big Data in Real Estate?”, *The Journal of Portfolio Management*, vol. 43 no. 6, 2017, pp. 202-211
26. Kontrimas, V. and Verikas, A., “The mass appraisal of the real estate by computational intelligence”, *Applied Soft Computing Journal*, vol. 11 no. 1, 2011, pp. 443-448
27. Peterson, S. and Flanagan, A. B., “Neural network hedonic pricing models in mass real estate appraisal”, *Journal of Real Estate Research*, vol. 31 no. 2, 2009, pp. 147-164
28. Rumelhart, D. E., Hinton, G. E., and Williams, R. J., “Learning internal representations by error propagation”, *parallel distributed processing: explorations in the microstructure of cognition*, vol. 1, 1986, pp. 318-362
29. Smola, A. J. and Scholkopf, B., “A tutorial on support vector regression”, *Statistics and Computing*, vol. 14, 2004, pp. 199-222
30. Vapnik, V., Golowich, S. E., and Smola, A., “Support vector method for function approximation, regression estimation, and signal processing”, *Advances in Neural Information Processing Systems*, vol. 9, 1997, pp. 281-287

<국문요약>

머신 러닝 방법을 이용한 오피스 임대료 산정
-랜덤 포레스트, 인공 신경망, 서포트 벡터 머신 활용을 중심으로-

정 성 훈 (Jeong, Sung-Hoon)
진 창 하 (Jin, Changha)

본 연구는 최근 자동 평가 모형과 관련한 연구에서 보편적으로 활용되고 있는 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신을 사용하여 오피스 임대료 산정 모형을 구축하고 이들의 적용 가능성을 검토하였다. 이를 위해 서울시에 소재한 507동 오피스의 임대 조사 자료를 활용하여 각 방법별 최적 모형을 구축하고 이들의 성능을 비교하였고, 가장 좋은 성능을 보인 모형으로부터 PD plot을 도출하였다. 표본 전체를 대상으로 하여 임대료 산정 모형을 구축한 결과, 모형별 성능의 순위는 서포트 벡터 머신 모형, 인공 신경망 모형, 랜덤 포레스트 모형 순으로 나타났고, 서포트 벡터 머신 모형으로부터 PD plot을 도출하여 해당 모형이 공실률이 높을수록, 전용률이 높을수록, 연면적이 클수록, 층수가 높을수록, 연식이 낮을수록, 승강기수가 많을수록 임대료를 높게 산출하는 것을 확인하였다. 주차대수는 모형의 예측값과 2차 곡선의 형태를 보이는 것으로 나타났다. 추가적으로 표본을 초대형, 대형, 중대형, 중형 등급으로 나누어 각 규모 등급별 임대료 산정 모형을 각각 구축하고 이들의 PD plot을 도출하였다. 본 연구는 다양한 머신 러닝 방법을 사용하여 오피스 임대료 산정 모형을 구축하고자 하였다는 점, PD plot을 통해 모형의 학습 결과에 대한 해석을 시도하였다는 점, 규모 등급별 임대료 산정 모형을 각각 구축하여 그 결과를 제시함으로써 규모 등급별 분석의 필요성을 입증하였다는 점에서 의의를 갖는다.

주 제 어 : 오피스 임대료 모형, 머신 러닝 방법, 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신