기계학습을 이용한 아파트 매매가격 예측 연구 : 한국 아파트의 내·외적 데이터 수집과 가격 예측 중심으로

주정민*, 강선미**, 최지웅***, 한영우****
*(주저자) 동덕여자대학교 정보통계학과
**(공동저자) 고려대학교 산업경영공학부
***(공동저자) 한성대학교 컴퓨터공학과
****(교신저자) 한국예탁결제원
2019170850@korea.ac.kr, cosmos1767@naver.com, choi950830@gmail.com, ywhan@ksd.or.kr

A Study on the Prediction of Apartment Sale Price Using Machine Learning : Focused on the Collection of Internal and External Data and Price Prediction of Korean Apartments

Jeong-Min Ju*, Sun-Mee Kang**, Ji-Wung Choi***, Youngwoo Han****
*(1st Author) Dept of Statistics and Information Science, Dongduk Women's University
**(2nd Author) Division of Industrial Management Engineering, Korea University
***(3rd Author) Dept of Computer Engineering, Hansung University
****(Corresponding Author) Korea Securities Depository

요 약

본 연구에서는 아파트를 대표할 수 있는 내·외적 데이터를 수집하고 인공지능 기술들을 활용하여 아파트 가격을 예측하는 시스템을 구축하고자 한다. 구체적으로 웹크롤링 기법을 통해 수집한 아파트 내·외적 데이터의 변수들에 대한 특성 선택(Feature Selection)을 수행하였고, 다양한 인공지능 기법을 활용하여 부동산 가격 예측 모형을 개발하였다. 아파트 가격 예측 모형 생성을 위해 Linear Regression, Ridge, Xgboost, Lightgbm, Catboost 등의 기계학습 알고리즘을 사용하였고, RMSE를 사용하여 각 예측 모형 간의 성능 비교를 수행하였다. 가장 성능이 좋은 예측 모형은 Xgboost기반 예측 모형이였으며, RMSE값이 약 0.0366으로 가장 낮았으며 테스트 데이터에 대한 정확도는 약 95.1%였다.

1. 서론

1.1 연구의 배경 및 목적

최근 부동산 시장에 대한 관심도가 높아지면서 부동산 시장 정책 수립과 투자 의사결정 지원을 위한 부동산 시장 분석 및 가격 예측 모형 개발의 필요성이 부각되고 있다. 기존에 선형 회귀분석을 사용하여 아파트 가격을 예측했던 연구들이 있었지만, 복잡한 아파트 가격 산정 문제를 선형 문제로 가정한 후 단순화시켜 접근했던 방식의 한계점 때문에 현실적 아파트 가격 예측 문제에 적용하기 어렵고 과부적합 발생 및 예측 정확성이 떨어지는 단점들이 존재하였다.

이에 본 논문에서는 선형 회귀분석 모형에서 발생하는 문제를 해결하고, 정확도를 향상시키기 위한 방안으로 다양한 기계학습 기반의 아파트 가격 예측 모형을 개발하고 결과를 비교 분석하고자 한다. 아파트 내적 요인인 '전용면적, 건축년도, 총 층수, 동수' 외에도 외적 요소인 '학군, 역세권, 브랜드 인지도'를 아파트 가격 결정 요인으로 가정하고, 웹 크롤러를 개발하여 내·외적 데이터와 아파트 실거래가 데이터를 수집하고 데이터 전처리를 수행하였으며 아파트 가격 예측 모형을 구현 및 성능 비교하였다.

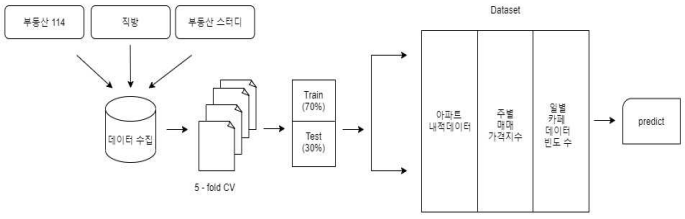
1.2 연구 범위 및 연구 방법

[그림 1]은 아파트 데이터 수집 및 가격 예측 모델 구현 구성도이다. 본 연구에서는 아파트 가격 예측 모델의 개발을 위해 서울특별시 25개구의 2012년 1월부터 2020년 8월까지의 아파트 실거래가 데이터(국토교통부 제공)를 사용하였다. 가격 예측 모형 생성을 위하여 해당 데이터 셋이 포함하고 있는 층수 등과 같은 매물의 자체 특성을 이용하였고, 웹 크롤러(Web Crawler)를 개발하여 아파트 주변 환경 요인인 교육, 교통정보, 선호도 등의 외적 변수 들을

수집 및 가공하였다.

수집 및 가공된 데이터에 대해 다양한 기계학습 모형을 적용하여 아파트 가격 예측 모형을 개발하였다. 사용한 기계학습 기법은 Linear, Ridge Regression과 Boosting 기법 (LightGBM, XGBoost, CatBoost)들이였으며 RMSE를 사용하여 예측 모델의 성능을 비교 평가하였다.

본 논문의 구성은 다음과 같다. 제1장 서론은 연구의 배경 및 목적을 설명하고, 제2장에서는 이론적 배경 및 기계학습 방법에 대해 설명한다. 제3장에서는 데이터 수집 및 전처리를 설명하고 제4장에서는 실험 결과를 설명한다. 제5장에서는 결론 및 향후 연구방향을 제시한다.



[그림 1] 아파트 데이터 수집 및 가격 예측 모델 구현 구성도

2. 이론적 배경 및 선행연구 검토

2.1 부동산 가격예측 선행연구

아파트 가격에 영향을 미치는 요인을 분석하는 연구는 과거부터 다수 진행된 바 있다. 대표적인 방법으로는 헤도닉 모형과 회귀분석이 있다.[1]

'서울 아파트 전세가격과 매매가격의 차이에 관한 연구, 성주한(2014)'과 '서울시 아파트 매매시장 유형별 가격변동 영향요인 분석, 고종완(2014)', '주택가격지수 모형의 비교연구, 임성식(2016)'은 부동산 가격 예측을 위해 다중회귀 방법을 사용하였다. [3-5]

'HEDONIC ANALYSIS OVER TIME AND SPACE: THE CASE OF HOUSE PRICES AND TRAFFIC NOISE*, Aaron Swoboba(2015)'와 'DETERMINANTS OF HOUSE PRICES IN TURKEY:A HEDONIC REGRESSION MODEL, Sibel SELIM(2008)'은 헤도닉 모형을 이용하였다. [6,7]

선형 회귀분석은 독립변수와 종속변수간의 관계성과 영향력을 수식적 쉽게 파악할 수 있기 때문에 아파트 가격 예측 연구에서 많이 사용되어 왔다. 하지만 정확도를 낮추는 몇 가지 문제점이 발견된다. 독립변수로 설정한 다양한 특성들이 실제 현실에서는 완벽히 선형적이지 않고, 독립변수들 간의 관계에서 다중공선성 문제가 발생한다. 이러한 문제를 가지고 있는 회귀분석에서는 자주 과대적합, 과소적합의 문제가 발생한다.

최근에는 기계학습을 사용하여 회귀분석의 문제점을 보완하는 연구들이 많이 증가하고 있다. 'Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Byeonghwa Park(2014)'와 'Real Estate Price Prediction with Regression and Classification, Hujia Yu(2016)'은 여러 기계학습 방법을 통해 아파트 가격을 예측하고 각 모델의 정확도를 비교하여 가장 우수한 모형을 찾는 연구가 진행되었다. [8,9]

부동산뱅크는 '미래시세 서비스'를 제공하고 있다. 미래시세 서비스는 아파트의 현재부터 1년 후의 가격을 예측하는 모형으로, 2003년부터 제공하던 부동산가격 예측 모형을 통해 아파트의 주거환경, 교통환경, 단지 및 평형등의 환경요인을 다중선형 회귀분석을 통해 예측한 평가기법이다. SPSS 통계 패키지에 기반을 두고 있으며, 29년간의 시세 데이터베이스와 각 아파트별 개별변수, 경제 변수가 포함되어 있다.

로보리포트는 2019년, 챗봇을 활용하여 부동산 매물 정보를 입력 받은 후 기계학습 기반의 추천 매물을 제공하는 방법에 관한 챗봇 특허를 등록하였다.

'부동산 카페 지표'는 부동산 관련 카페의 활성화 정도를 나타내는 지표로써, 최근 주택시장 방향성을 가늠하는 데에 사용될 수 있다. 해당 지표를 통해 부동산 거래 활성화 정도를 간접적으로 파악할 수 있다. 또한 '호갱노노' 같은 전문 부동산 정보 사이트의 지역별 방문자 순위 및 실시간 방문자 분석 정보 덕분에 점점 많은 사람들이 빅데이터 및 인공지능을 활용한 아파트 통계 분석과 가격 분석 방법에 대한 관심이 높아지고 있다.

2.2 기계학습 개념

[표 1]은 본 연구에서 사용한 기계학습 기법의 특징이다. linear, ridge의 두가지 회귀모형과 xgboost, lightgbm, catboost의 부스팅 알고리즘을 선택하였다. 각 모형별 특징이 [표 1]에 서술되어 있다.

모형의 종류	특징
linear regression	독립변수와 종속변수의 데이터로 간단한 방법으로 추정가능
ridge regression	종속변수에 미치는 독립변수의 영향력이 큰 경우 용이함
xgboost	의사결정트리에서 파생된 모형으로 병렬처리를 통해 빠른 학습이 가능하며, 과적합을 방지할 수 있음
lightgbm	기존 부스팅 계열의 방식과 다르게 리프 중심 트리 분할방식을 사용하여 예측 오류 손실을 최소화함
catboost	범주형 변수를 수치형변수로 처리하는데 유용함

[표 1] 기계학습 기법의 특징

3. 데이터 수집 및 전처리

3.1 데이터 수집

본 연구는 2012년부터 2020년까지 8년 동안 서울특별시에서 거래된 아파트를 대상으로 한다. 부동산 114, 직방, 네이버 카페 부동산 스터디 웹사이트를 통해서 다양한 부동산 데이터를 수집하였다. 또한 연구에서 정한 기간에 등록된 매매 거래 데이터(전, 월세 미포함)를 국토교통부 실거래가 공개시스템에서 찾아 이용하였다. 같은 기간의 뉴스 및 카페 텍스트 분석을 위해 아파트명이 포함된 게시글의 '제목, 본문, 언론사, 인물, 위치, 기관, 키워드, 특성'등의 정보를 크롤링을 통하여 수집하였다. 부동산 웹에서 제공하는 정보들 중 아파트 가격에 영향을 미치는 요인으로 아파트 내부 변수와 교육, 역세권 정보 등 아파트의 외적 변수들을 선정하여 수집하였다. 또한 건설취업포털 건설워커에서 2014년부터 2020년도의 시공능력순위 데이터를 수집하였다.

3.2 변수 선정

[표 2]는 본 실험에서 사용한 변수 타입과 설명을 나타낸다. 예측 모형 생성시 사용한 독립 변수는 전용면적, 층, 건축년도, 학군, 역, 총층수, 동수 그리고 시공사이고, 종속 변수를 거래금액이다.

변수명	타입	설명	출처
전용면적	FLOAT	아파트 면적	국토교통부
거래금액	INT	아파트 매매가격	
층	INT	거래된 아파트의 층	
건축년도	INT	아파트 건축년도	부동산웹사이트
학군	INT	학교까지 거리 및 도보시간	
역	INT	역까지 거리 및 도보시간	
총층수	INT	아파트의 가장 높은 층	
동수	INT	아파트 단지의 동 수	
시공사	STR	아파트 시공사	

[표 2] 변수 타입과 설명

3.3 데이터 전처리

3.2.1 더미 변수 생성

1) 지하철과의 거리

역세권 개념을 바탕으로 아파트별 지하철과의 거리의 역세권 기준을 선정하였다. 가변수를 활용하여 범주형 변수를 양적데이터로 변환하였다. 비역세권은 반경 1,500m(도보 10분 이상)으로 지정하여 0으로 가변수화 했으며, 역세권은 반경 500m(도보 10분 이내)로 지정하여 1로 가변수화 하였다. 초역세권은 반경 200m(도보 5분 이내)로 지정하여 2로 가변수화를 수행하였다.

2) 시공사

건설워커 홈페이지에서 제공하는 2014년부터 2020년까지의 시공사 순위를 사용하였다. 연도별 시공사(서울 한정) 순위 1위부터 10위까지와 '한국주택공사'를 포함한 11개의 시공사에 포함되는 경우 1을, 아닌 경우 0으로 바꾼 레이블 인코딩을 적용하였다.

3.2.2 정규화

모델의 최적화 과정에서의 안정성 및 수렴속도 향상을 위하여 모든 연속형 변수의 데이터를 0에서 1 사이의 값으로 정규화 처리하였다.

3.2.3 데이터 분할

모형의 적합성을 위해 Dataset을 train data와 test data 7:3 비율로 분리하였다. test data에 과적합(overfiting) 문제를 방지하기 위해 k- 겹 교차검증(cross-validation)을 실시하였다. 데이터를 5개의

data fold로 분할하고, 총 5개의 data fold set를 구성하였다. 각 데이터 세트의 수행 결과에 대한 평균값을 사용하여 최종적인 검증 결과를 도출하였다.

4. 실험 결과

서울특별시 467개의 법정동별 아파트를 기계학습하여 전체 RMSE 평균값으로 계산하였다. [표 3]은 기계학습 모형의 RMSE값을 나타낸다.

Model	Linear regression	Lasso regression	Xgboost	Lightgbm	Catboost
RMSE	0.059185	0.059689	0.036559	0.083192	0.038034

[표 3] 기계학습 모형의 RMSE 값

본 논문은 최적의 예측모형을 만들기 위해 회귀 알고리즘 Linear regression, Ridge regression, Xgboost, Lightgbm, Catboost를 총 5개의 알고리즘 성능에 대해 정확성을 비교하였다. 모델 성능 평가 지표로는 Root mean square Error(RMSE)를 사용하였다.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

5개의 예측 알고리즘 중 RMSE 값을 비교한 결과, 0.0305로 가장 높은 Xgboost를 최종 모델로 선정하여 유의미한 예측 결과를 도출하였다. Xgboost의 테스트 데이터에 대한 정확도는 0.950796이다.

5. 결론 및 향후 연구방향

5.1 결론

웹크롤러를 개발하여 아파트 내·외적 데이터를 수집하고 인공지능 기술을 활용하여 다양한 아파트 가격 예측 모형을 만들어 RMSE값이 0.0305이며 데스트 데이터의 정확도가 약 95.1%인 유의미한 결과를 도출 하였다.

5.2 시사점 및 향후 연구방향

본 연구는 아파트 내적 지표와 커뮤니티 사이트의 빈도 지표를 추가하여 최근 이슈가 되고 있는 부동산 아파트 시장의 동향 및 가격을 예측해볼 수 있다. 향후 빅데이터 분석 플랫폼(Apache spark)와 개발한 부동산 웹사이트 크롤러를 연동하여 다양한 부동산 프로그램을 개발할 수 있다. 본 연구 결과가 향후 부동산 투자 컨설팅 업무에 활용될 수 있을 것으로 기대된다.

사 사

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링프로젝트 결과물입니다.

참고문헌

[1] 이진성, 지역별 주택가격 변동률에 영향을 미치는 요인 규명에 관한 연구, 한국부동산학회, 부동산학보 55권0호, 266-278(13pages), 2013

[2] 이강배, 기계 학습을 이용한 아파트 가격결정요인 분석 : 부산 지역을 사례로, 동아대학교 대학원, 2018

[3] 성주한,서울 아파트 전세가격과 매매가격의 차이에 관한 연구,한국부동산학회,不動産學報 第57輯 , 108~122쪽,2014

[4]고종완,서울시 아파트 매매시장 유형별 가격변동 영향요인 분석,한국부동산학회,부동산학보 58권 58호, 116-128p,2014

[5]임성식,한국데이터정보과학회,한국데이터정보과학회지 27권 6호,1573p ~ 1583p,2016

[6]Aaron Swoboda,HEDONIC ANALYSIS OVER TIME AND SPACE: THE CASE OF HOUSE PRICES AND TRAFFIC NOISE,Journal of Regional Science 55권 4호, 644p ~ 670p, 2015

[7]Sibel SELİM,Determinants of House Prices in Turkey : A Hedonic Regression Model = Türkiye'de Konut Fiyatlarının Belirleyicileri : Hedonik Regresyon Modeli,Dogus University Journal 9권 1호, 65p ~ 76p, 2008

[8]Byeonghwa Park, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,Expert Systems with Applications 42권 6호,2928p ~ 2934p,2014

[9]Hujia Yu, Real Estate Price Prediction with Regression and Classification, 2016