LightGBM을 이용한 서울지역 아파트 가격 예측모형

연세대학교 미래캠퍼스 경영학부 이승준, 류호윤 지도교수 - 신택수

[개요]

본 연구에서는 LightGBM 기계학습 알고리즘을 통해 서울지역 아파트의 가격을 예측함. 변인으로 자연/환경요인, 교통/인프라요인, 교육요인, 경제요인을 사용함. 서울시 공공데이터 포털 등에서 공개된 데이터를 이용하여 분석을 수행하였으며, 변수 중요도(Feature Importance) 확인을 통해 각 변인의 중요도를 살펴봄. 지하철역의 수, 행정구역(동), 전용면적이 중요한 변인으로 확인됨. 이를 바탕으로 서울지역 아파트 가격의 예측을 수행하였으며, 확인된 LightGBM의 RMSLE score는 0.150898로 확인됨. 본 연구의 한계점은 다음과 같음. 각 sample의 특성을 반영한 요소들을 통해 각 아파트 sample에 대한 가치를 평가하는데 성공했으나, 시계열 데이터를 이용한 미래 가격에 대한 예측으로 이어지지는 못하였음. 본 연구를 통해 서울지역 아파트의 가치를 평가하는데 있어 유의미한 활용이 가능할 것으로 기대됨.

1. 서론

[연구의 배경 및 목적]

대한민국의 많은 사람들은 투자수단 중 하나로 부동산 투자를 선택함.

어떤 요소가 해당 아파트의 가격 형성에 영향을 주는지를 통해 개발이나 투자 시 시간과 노력을 줄일 수 있지 않을까 하는 기대를 가지고 다양한 요소와 선호도 간의 관계 분석 후 이를 중심으로 해당 지역의 아파트 가격에 대한 예측을 하고자 함.

2. 선행연구

▼ 부동산 가격 예측에 관한 연구

저자명 (연도)	사용 data	적용 model	성과
이현재 외 (2020)	이미지 데이터	CNN	아파트 주변의 환경변수를 통한 서울 아파트 가격예측 (설명 비율 62.5%)
김채원 외 (2020)	경제 및 금융 데이터 (시계열)	LSTM	test data 정확도(평균 96.855%), 실제 예측 정확도(평균 82.521%)
주정민 외 (2020)	실거래가, 주변 환경 요인	Linear Regression, Ridge, Xgboost, Lightgbm, Catboost	RMSE를 활용하여 각 예측 모형 간의 성능 비교 (XGBoost 95.1%의 정확도)

▼ Hyper Parameter에 관한 연구

저자명 (연도)	사용 data	적용 model	성과
최용욱 외 (2020)	Exmouth 하부분지에서 획득된 Vincent field 자료	Bayesian optimization	하이퍼 파라미터 튜닝 프로세스에 소요되는 시간을 절약
James Bergstra 외 (2011)	conves, MRBI	GP, DBNs, TPE	GP와 TPE기법에 있어 효율적인 hyperparameter의 방법 제공
Takuya Akiba1 외 (2019)	MNIST	optuna	optuna pakage를 활용한 hyper parameter 방법의 효율성 검증

3. 연구모형

[Framework]

[LightGBM]

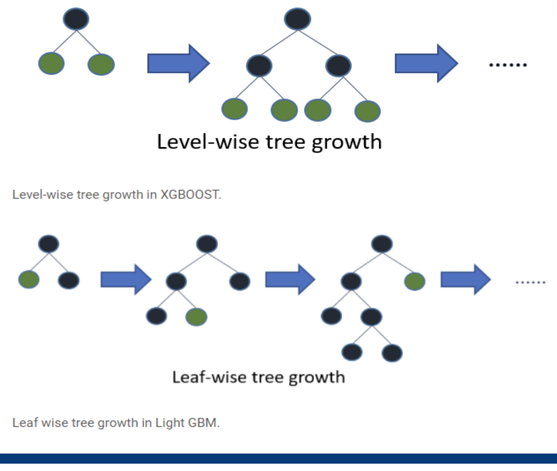
LightGbm은 Boosting기반이며, Boosting 기법의 대표적인 algorithm으로는 Gradient Boost 기법이 있음.

장점으로는 큰 사이즈의 데이터를 다루기 용이하고 적은 메모리를 차지하는 특징, 단점으로 10,000개 이하의 dataset을 사용할 경우 과적합 가능성이 존재함.

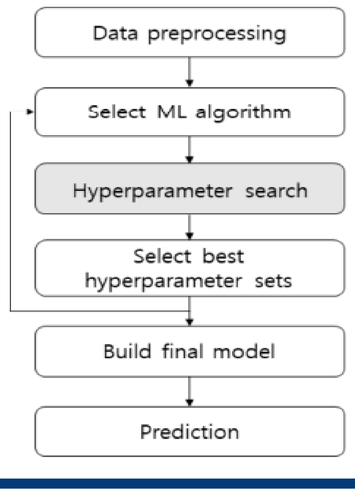
[Hyper Parameter]

본 연구에서는 Optuna, 그중에서도 Tree-structured Parzen Estimator라는 베이지안 최적화 알고리즘을 사용하여 Hyper Parameter를 추정함.

▼ LightGBM model Structure



▼ Hyper Parameter Flowchart



4. 실증분석

[사용 data]

#	Column	Info
1	exclusive_use_area	전용면적
2	floor	층
3	transaction_real_price	실거래가
4	hangang	한강 인접 여부
5	CBD	중심업무지구 인접
6	park	공원의 수
7	base_money_rate	기준금리 (연)
8	number_subway_stations	지하철역의 수
9	department_stores	백화점의 수
10	age	준공 경과기간
11	is_rebuild	재건축
12	special_high_school	특목고의 수
13	middle_school	중학교의 수
14	elementary_schools	초등학교의 수

[분석 수행]

▼ RMSLE Score

#	model	score
1	Linear regression	0.281969
2	Ridge	0.281969
3	Lasso	0.298397
4	ElasticNet	0.294288
5	Decision tree Regressor	0.266754
6	Random forest Regression	0.215594
7	LightGBM Regressor	0.150898
8	XGBoost Regressor	0.131689

▼ Hyper Parameter Tuning

best trial score	0.15699554814820735
max_depth	15
learning_rate	0.009770823729346987
n_estimators	2984
min_child_samples	36
subsample	0.8832752415247064

▼ Feature Importance

#	column	importance
1	전용면적	27889
2	준공 경과기간	16597
3	지하철 역 수	8378
4	기준 금리	6821
5	공원	6172
6	중학교의 수	5792
7	초등학교의 수	4044
8	층	3970
9	백화점의 수	3794
10	한강 인접 여부	2775
11	중심 업무 지구 인접 여부	1794
12	특목고의 수	1493

5. 결론

[결론]

LightGBM 을 활용한 분석 결과에서는 전용면적, 아파트의 준공 경과 기간이 높은 수치를 나타냄.

이는 샘플이 가지고 있는 특성이 가격에 중요한 영향을 미쳤다는 것을 의미함.

교통 요인인 지하철의 수가 세 번째로 중요한 요인으로 선택되었다는 점은 아파트 가격에 교통이 중요한 요소가 되고 있다는 의미.

따라서 아파트 입지분석 시 교통 요소를 중점적으로 확인하는 것이 적합.

[한계점]

이번 프로젝트에서 진행하고자 했던 아파트 가격 예측은 아파트 가격의 가치 평가에 그침.

미래 가격 예측을 위해서는 각 샘플이 되는 아파트 가격에 대한 시간 데이터가 필요하지만 보유하고 있던 데이터는 그렇지 못했음.

미래 예측을 위해서는 각 충분한 시간 시퀀스를 가지고 있는 아파트 샘플들이 갖춰져야 함.

따라서 후행 연구에서는 적합한 형태를 가진 시퀀스 데이터를 수집하여 미래 예측이 가능하며 높은 정확도를 확보할 수 있는 모델을 이용하여 분석을 수행해야 할 것으로 보여짐.

[향후 활용 방안]

아파트 샘플에 대한 가격 예측 결과를 바탕으로 모델이 도출한 가치는 시장의 평가를 토대로 진행됨.

그리고 본 프로젝트에서 활용한 모델에서는 변수 중요도 (Feature Importance)를 확인할 수 있었음. 이를 통해 예측 가격과 중요한 입지 요인을 파악할 수 있었는데 이 점은 추후 아파트에 대한 가치 평가를 하는데 있어 활용될 수 있을 것으로 보임.