5/15 건축물 부분

1. 현재 박순만 교수님에게 메일로 질의 드린 상태. 아래는 내용.

선형회귀의 평가지표에서 R2의 기준을 지나치게 높게 잡은 것 같습니다.

애초에 데이터가 그렇게 정규성과 선형성을 가지고 있지 않을 거라 예상되는데… 일단 R2에 대한 기준 값은 무의미합니다. 데이터를 보고 결정해야 할 듯.

로지스틱에서 accuracy가 무엇을 의미하는지는 모르겠으나, 일반적으로 chi-sqaure 값이나 우도비(likelihood), pseudo-R2를 사용합니다.

그 외에 rmse나 AIC, BIC를 이용하기도 하는데, 다른 모델과의 score 비교를 한번 해보시죠.

이후 git link 다시 첨부해서 보내드렸는데 읽고 답장이 없으신 상태.

🡪 R2의 기준은 아래의 논문에서 보면 한 논문은 70%대고 한 논문은 90%대. 우리한테 더 잘 맞는 논문의 기준을 따라 가는 것이 좋을 듯.

🡪 선형성, 정규성 문제는 XGBoost가 Tree 기반 Algorithm이기 때문에 고려하지 않아도 준수한 성능을 보임.

🡪 XGBoost랑 이 모형의 발전 형태인 CatBoost도 같이 모델 비교하여 선택

🡪 PCA를 사용해서 변수간의 상관관계 고려한 예측 모델 구성할 예정

2. 우리 서버랑 어떤 방법으로 모델과 데이터를 공유할까에 대한 고찰.

* 서버에 직접 모델을 넣는 방법과 웹서버를 띄워 RESTful하게 서로 정보를 주고받는 방법이 존재.
* 전자로 간다면, aws 안에서 데이터들을 전처라하고 예측하는 모델을 만드는 pipeline을 구축해서 모델을 생성하고, 해당 모델을 저장해서 웹에서 사용하는 방식으로 사용.
* 후자로 간다면, GCP에 Flask로 웹서버를 띄우고 도메인 활용하여 GET, POST로 서로 상호소통하는 방식으로 사용.
* 뭐 방법은 백엔드 단이랑 이야기가 맞아야 하니 얘기하고 결정해야 할 듯.
* Flask Web server Reference: <https://towardsdatascience.com/how-to-build-a-machine-learning-api-using-flask-2fb345518801>
* 여러가지 방법 모음집: <https://www.baeldung.com/java-working-with-python>

3. 대망의 읽기 싫은 논문을 읽은 결과 우리의 검증 시나리오는 어떻게 할 것인가!

우선 첫번째 논문: 김상민(Sang Min Kim),박찬권(Chankwon Park),and 이지은(Ji-Eun Lee). "XGBoost 회귀를 활용한 편의점 계약전력 예측 모델의 최적화에 대한 연구." 한국IT서비스학회지 21.4 (2022): 91-103. (<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202230760671758&SITE=CLICK> )

내용을 살펴보면 해당 논문의 경우는 우리 처럼 부동산과 관련된 내용을 담지는 않았다. 하지만 우리가 모델의 선정 과정에 있어, 값을 활용하는데 정당성을 부여할 수 있는 자료다. 해당 논문의 내용을 발췌하면 “MSE, MAE, RMSE는 모델마다 결과 값이 다르고, 0에 근접한 값으로 표현되면 좋은 성능을 발휘하지만, 0에 얼마나 근접을 해야 우수한 성능으로 판단되는지에 대해 성능 판단에 어려움이 있을 수 있다” 라고 했으며, 분석의 결과를 확인할 때 값과 값을 활용했다. 해당 논문에서 가장 우수한 성능을 보인 값은 0.931 이였다.

텍스트, 영수증, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

두번째 논문: 김학현 ( Hakhyun Kim ),유환규 ( Hwankyu Yoo ),and 오하영 ( Hayoung Oh ). "딥러닝과 머신러닝을 이용한 아파트 실거래가 예측." 정보처리학회논문지. 소프트웨어 및 데이터 공학 12.2 (2023): 59-76.   
(<https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002935936> )

내용을 살펴보면 이 논문은 부동산의 실거래가를 예측하는 방법을 담은 논문으로 DNN과 Boosting 방법 3가지를 활용하여 예측하였다. 해당 논문의 프로세스는 전처리 후 다중공산성을 제거하고, 변수를 선정하는 과정을 거친 후 모델에 적용하는 방법이다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

해당 표는 변수 선택과정 후에 model을 적용시켰을 때 나타난 지표들이다. XGBoost를 사용한 경우 0.7965, CatBoost를 사용한 경우 0.8464의 값을 가지는 것으로 나타났다. 우리의 모델에 적용시키기 위해서는 feature selection 과정을 거쳐서 XGBoost, CatBoost를 사용해서 해당 지표를 참고할 수 있을 것 같다.

세번째 논문: 최재영,양희윤,and 오하영. "그래디언트 부스팅 모델을 활용한 상점 매출 예측." 한국정보통신학회논문지 25.2 (2021): 171-177.   
(<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202311540140522&SITE=CLICK> )

해당 논문은 점포의 특성에 따라 상권을 추천해주는 알고리즘에 대한 연구다. 해당 부분은 우리의 업종 추천 부분과 매우 유사하지만 우리의 과정의 역과정이다 (우리 과정은 건물의 특성을 기반으로 업종을 추천하기 때문). 그러나 순방향의 과정이 성립한다면 역방향의 과정도 비슷한 수준으로 예측 가능할 것이라고 가정을 한다면, 우리의 Accuracy 또한 해당 논문과 비슷한 수치를 구할 수 있을 것이다. 아래는 해당 논문의 Accuracy 표다.

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

해당 논문은 DT와 XGBoost를 사용했지만, 현재 구상중인 DNN을 사용해서 해당 수치를 목표로 Optimizing할 생각이다.

그래서 제 생각은

Classification threshold: Accuracy > 85%

Regression threshold: > 0.85 and smaller RMSE  
로 정하는 것이 가장 좋지 않겠나 하는 의견입니다.

4. 남은 3주간의 진도 계획

13주(~5/30): Regression Model 완성, 다양한 지표 검색

14주(~6/6): 모델 완성 및 threshold 만족하게 Optimization 및 data pipeline 구축

15주(~6/13): aws 모델 적재 및 테스트. Validation.

16주(~6/19): 왜 3주간 계획이지 남은건 2준데

혹시 읽으시면서 더 보완해야할 사항이나 궁금하신 내용 있으면 언제든 카톡, 전화 부탁드립니당ㅎㅎㅎ