수도권 아파트 전세가 예측 모델

Recsys-10조(물어봐조)
곽정무, 박준하, 박태지, 신경호, 이효준

1. 프로젝트 개요

1-1. 개요

전세는 주택 임차 계약 유형 중 하나로 세를 내지 않고 일정한 비용을 지불하고 계약 기간이 끝난 후 돌려받는 독특한 제도이다. 수도권에서는 전세 비율이 전체 전월세 계약에서 60% 이상을 차지하며 전세 시장 동향은 부동산에서 굉장히 중요한 지표로 간주된다. 이번 대회는 9월 30일부터 10월 24일까지 약 4주간 진행된 대회로 2019년 4월부터 2024년 4월까지의 부동산 전세계약 관련 데이터를 바탕으로 AI와 머신러닝을 활용해 전세가를 예측하는 알고리즘을 개발하는 대회이다.

1-2. 환경

협업 환경: GitHub, Optuna
의사 소통: Zoom, Slack, Jira, Confluence

2. 프로젝트 팀 구성 및 역할

공통	EDA 및 랩업 리포트 작성
곽정무	Confluence 템플릿 구축, 회의록 작성, Github 초기 세팅, 모델 결과분석
박준하	기본적인 프레임워크 구현, 외삽 모델 개발
박태지	AutoML, 클러스터링
신경호	hyperparameter 최적화 세팅, 이상치 탐지
이효준	Jira 세팅, 클러스터링 및 피처 엔지니어링, 랩업 리포트 관리

3. 프로젝트 수행 절차

3-1. Timeline

(1주차) 강의 듣기, Jira, Confluence, Github 세팅
(2주차) EDA 및 가설 설정
(3주차) 피처 엔지니어링 및 모델링
(4주차) 최종 모델 선정 후 하이퍼파라미터 튜닝

3-2. 협업 과정

[1] EDA, Feature Engineering 등 프로세스별로 관리자 배정
[2] 실험 과정을 Confluence에 기록하여 팀원들과 공유 및 관리자를 통한 모니터링
[3] 매 회의마다 회의록을 작성해 아이디어 기록
[4] 최종 모델 선정 후 역할 분배 통한 모델 발전 시도

4. 프로젝트 수행 결과

4-1. EDA

이번 대회는 180만개 이상의 대용량 데이터가 주어졌으며 시간적, 지리적 특성의 영향이 강한 데이터이기에 다양한 측면을 고려하여 EDA를 진행했다. 기본적인 계약 일자, 위치, 아파트 정보 등을 담은 전세 계약 데이터셋과 함께 지하철, 학교, 공원의 위치데이터, 월별 금리 데이터가 주어졌으나 위치데이터의 경우 위도, 경도만 주어져 피처엔지니어링 방향 설정에 어려움이 있었다.

EDA 진행 초기 주요 문제는 같은 위치, 같은 계약 일자, 같은 계약금으로 중복되는 데이터의 존재였다. 이에 대해 이상치라 판단해 데이터셋에서 제외하자는 의견도 있었으나 논의 끝에 신축 아파트의 사례가 있는 만큼 데이터 확인 후 판단하자는 결정을 내렸다.

또한, 건물의 나이가 0 미만, 층이 0 이하인 도메인 지식과 어긋나는 데이터가 존재했는데 조사를 통해 건물이 완공되기 전 계약하는 사례와 지하층 거래 내역이 있는 아파트의 존재를 확인하고 데이터를 보존한다는 결정을 내렸다.

EDA를 진행하며 위도, 경도를 활용한 아파트별 거래 계약에 집중했는데 같은 아파트, 월, 면적임에도 전세가가 1억이상 차이나는 등 모델에 혼선을 줄 수 있는 데이터가 다수 존재함을 확인했고 이에 따른 어려움이 예상됐다. 또한 테스트데이터셋이 학습데이터셋보다 시간상 후에 위치하기 때문에 신축인 경우 테스트데이터셋에는 존재하나 학습데이터셋에서는 존재하지 않는 아파트들이 존재했다. 따라서 추후 피처엔지니어링, 모델링, 모델 결과 분석에서는 각각의 아파트들과 아파트들이 위치한 지역을 고려하고자 했다.

4-2. 피처 엔지니어링

처음 데이터를 제공받았을 때, 서브 데이터의 경우 지역이 수도권으로 한정되지 않아 불필요한 장소를 먼저 제거했다.

지리데이터는 위도와 경도를 기반으로 아파트와 각 시설 간 최소 거리 및 특정 범위 내 시설 수를 변수로 추가했고, 거래 연월을 바탕으로 금리변수도 생성했다. 세부적으로는 초 ·중 ·고 분류와 공원 면적에 따른 거리 및 시설 수 변수를 함께 추가했다.

또한, 위경도를 binning하여 카테고리 변수를 생성하고, 좌표를 활용해 아파트 ID 변수를 만들었다. EDA 결과에서 도출한 인사이트를 반영해 위경도와 면적을 묶어 새로운 ID를 생성했으며, 환승역과의 거리 및 시설 수 변수, 제곱미터당 전세가 변수를 추가했다. 또한 시 ·군 ·구 ·동 같은 지역정보를 대체할 수 있는 클러스터링 변수도 고려했다. 클러

스터링 변수의 경우, 아파트ID를 기반으로 최적의 클러스터 수를 찾기 위해 엘보우 기법, 실루엣 스코어, Davies-Bouldin 스코어를 함께 사용했다.

분석 지표로는 대회 평가지표인 MAE를 RMSE, Adjusted R2과 함께 활용해 이상치에 대한 민감도와 과적합 문제를 함께 모니터링했다. 또한 평균 MAE가 상위 100개인 아파트를 따로 분석하여 MAE가 큰 아파트들에 초점을 맞춰 문제를 해결하고자 노력했다.

4-3. 모델별 분석

1) 지리 기반 피쳐 모델

1-1. 개요

baseline 모델에서는 주변 지리 관련 정보들이 모델에게 주어지지 않아 전세가를 오직 한정된 feature들로 예측하고 있다. 따라서 지하철, 학교, 공원 등의 지리 정보가 포함한 여러 feature들을 추가하여 모델의 예측 정확도를 높이려고 했다.

1-2. 구현

모델은 baseline과 마찬가지로 lightGBM을 사용하되, 지리 정보를 포함한 feature들을 추가하였다. 대표적으로 가장 가까운 학교, 공원, 지하철까지의 거리를 추가하였으며, 반경 10km 및 1km 이내에 존재하는 학교, 공원, 지하철의 개수 또한 feature로 두었다.

이때 가장 가까운 학교까지의 거리는 초중고 각각에 대해 계산하였으며, 공원의 경우에도 면적에 따라 세분화하였다.

1-3. 결과

validation set은 baseline과 마찬가지로 2023년 7월 1일부터 2023년 12월 31일까지의 데이터를 사용하였다. 또한 lightGBM의 하이퍼파라미터 중 num_leaves는 63으로 설정하였고 나머지는 기본값을 사용하였다. 그 결과, 추가적인 피처를 포함하지 않았을 때의 MAE는 4331이 나온 반면, 피처를 모두 사용하였을 때는 4400이 나왔다. 이로부터 추가된 피처들이 오히려 모델의 학습을 방해한다는 사실을 확인할 수 있었다.

따라서 피처들을 하나씩 추가하고 제거해보며 최적의 피처 조합을 찾는 과정을 거쳤다. 그 결과 특정 범위 내에 존재하는 시설 수만이 유효한 피처임을 확인할 수 있었으며, 기본 피처들 중 contract_day, contract_type, age 또한 불필요함을 알 수 있었다. 이러한 피처들을 제거하자 최종적으로 MAE를 4226까지 줄일 수 있었다.

2) 단순 시계열 모델

2-1. 개요

지리기반 피처를 추가해도 모델의 성능이 크게 향상되지 않았다. 직관적으로 생각해보아도, 모델에 지리 정보를 추가하는 것은 새로운 거래의 전세가를 예측하는 데에는 도움

되겠지만, 이전 거래의 전세가를 예측하는 데에는 큰 도움이 되지 않을 것이다. test data의 90%이상이 이전 거래가 존재하는 아파트이기에 문제가 더욱 두드러졌다. 따라서 시계열 모델을 사용해 이전 거래 전세가를 바탕으로 다음 거래 전세가를 예측하고자 했다.

2-2. 구현

복잡한 시계열 모델을 사용해볼 수 있겠지만, 일단 이전 거래들의 평균을 구하여 이를 다음 거래의 전세가로 예측하는 단순한 모델을 사용했다. 평균을 계산하는 다양한 방법들을 실험해본 결과, 이전 전세가들의 가중 평균을 사용하는 것이 가장 성능이 높았다. 이때 가중 평균은 파라미터 alpha를 0.5로 둔 pandas ewm 함수를 사용하여 계산하였다.

2-3. 결과

validation set은 2023년 7월 1일부터 2023년 12월 31일까지의 데이터 중 이전 거래가 존재해 가중 평균을 계산할 수 있는 데이터를 사용하였으며, 그 결과 MAE가 4572가 나왔다. 단순 rule-based 모델임에도 불구하고 성능이 꽤 괜찮게 나온 것을 확인할 수 있었으며, 시계열적인 접근방식이 유효할 수 있음을 방증해주는 결과였다.

3) 외삽 모델

3-1. 개요

미래 시점의 가격을 예측하는 것은 매우 어려운 문제이다. 사용할 수 있는 피처의 종류가 한정적이고, 데이터에 노이즈가 많은 현재 상황에서는 더욱 어렵다. 실제로 미래의 전세가를 예측하기 위해 복잡한 시계열 모델을 사용해보았지만 큰 효과를 보지 못했다.

따라서 거래가 드문 아파트의 전세가를 거래가 많은 아파트의 전세가를 바탕으로 외삽(extrapolation)하는 방법을 사용하고자 한다. 즉, '미래'가 아닌 비어 있는 '과거'의 전세가를 예측하는 것이다. 이는 최대한 멀리 예측해도 현재 시점의 전세가를 얻을 수 있는 것이기에 성능의 한계가 있겠지만, 직접적으로 미래를 예측하는 것보다는 쉬울 것이다.

3-2. 구현

모델의 전체적인 흐름은 아래와 같다.

1. 각 아파트와 각 시간대마다 임베딩을 둔다.
2. 대상의 특징을 잘 표현할 수 있도록 임베딩을 적절히 학습시킨다.
3. 아파트 임베딩과 시간 임베딩의 상호작용을 통해 전세가의 등락을 예측한다.

이해를 돕기 위해 모델이 어떤 방식으로 과거의 전세가 등락을 예측하는지 예를 들자면, 먼저 A아파트와 B아파트, 그리고 2019년에 대응되는 임베딩이 존재한다고 하자. 이때 정부가 2019년 3월에 고급 아파트에 대한 세금을 올리는 정책을 발표하였다. A아파트와 B아파트가 고급 아파트라면 2019년 한 해 동안 전세가가 크게 떨어질 것이다. 모델은

이 결과로부터 두 아파트의 임베딩을 유사하게 만들 것이고 2019년의 시간 임베딩은 고급 아파트에 해당하는 임베딩과 상호작용했을 때 '하락'을 예측하도록 학습될 것이다.

여기서 C아파트를 추가해보자. C는 고급 아파트이지만 이전 두 아파트와는 달리 거래가 드물어 2017년에 마지막 거래가 일어났다. 따라서 2019년 C아파트 전세가의 등락은 train data에 존재하지 않는다. 하지만 다른 데이터에서의 학습을 통해 C의 임베딩이 다른 고급 아파트들과 유사하게 만들어졌다면, 2019년의 시간 임베딩과 상호작용했을 때 '하락'을 예측할 것이다. 한마디로, 거래가 빈번한 아파트들로 시간 임베딩을 학습시킨 후 거래가 드문 아파트의 임베딩과의 상호작용을 통해 전세가의 등락을 예측하는 것이다.

이제 세부적인 모델 구현에 대해 설명하자면, 먼저 아파트 임베딩의 차원은 16으로 두었고, 시간 임베딩의 차원은 64로 둔다. 아파트 임베딩이 시간 임베딩에 비해 크기가 작은 이유는 아파트 임베딩이 학습되는 횟수가 시간에 비해 적기 때문이다. 시간 임베딩과 아파트 임베딩의 상호작용은 임베딩을 concatenate한 후 tower 형식의 fully connected layer를 3번 통과시키는 방법을 사용하였다. 구체적으로는 16+64=80차원의 임베딩을 입력을 받아 이의 절반인 40차원으로 줄이고 이를 다시 20차원으로 줄인 후 마지막으로 1차원으로 줄이는 방식을 사용하였다.

3-3. 결과

이전 실험들과 마찬가지로 validation set은 2023년 7월 1일부터 2023년 12월 31일까지의 데이터를 사용하였다. 다만 모델이 실제로 외삽을 올바르게 하는지 확인하기 위해 validation set 중에서 1년 정도의 거래 공백기가 있는 데이터만을 예측해보았다.

그 결과, MAE는 6000 전후로 굉장히 불안정하게 나왔다. 외삽을 하지 않고 그대로 예측한 결과가 5000 정도 나오기에 현재로서는 모델이 큰 효과를 보이지 못한 것이다. 다만 GCN을 통해 아파트 임베딩의 적은 학습을 보완하거나 임베딩에 아파트에 대한 추가적인 정보를 직접 넣는 등 다양한 방법을 통해 개선할 여지가 있다.

4-4. 하이퍼파라미터 튜닝

하이퍼파라미터 튜닝을 위하여 optuna를 사용했다. 이전에 사용한 wandb는 파라미터별 학습 결과를 시각적으로 정리하기 좋다는 장점이 존재하지만 최적화된 파라미터를 확인하기 위해서는 web으로 이동해야 한다는 단점이 있었다. 이를 개선하기 위해 optuna 라이브러리를 사용하여 한 번의 실행으로 학습에 최적화된 파라미터를 반환할 수 있도록 코드를 구성했다. 하이퍼파라미터 튜닝 코드는 여러 모델 중 ML 경진대회에서 많이 사용되는 LightGBM, XGBoost, CatBoost를 타겟으로 설정했다. 또한, 하나의 모델을 선택할 수도 있으나, 3개의 모델을 다 실행한 후 그 중 최적의 모델과 파라미터를 선택할 수도 있도록 구현했다. 테스트 데이터와 검증 데이터를 분할하는 방법에는 교차검증을 위해 K-Fold 방식을 사용하였으며 평가 요소로는 이번 대회에서 사용한 MAE를 채택했다.

5. 자체 평가

5-1. Jira 효과적인 활용성 부족

이전 프로젝트의 아쉬운 점을 보완하기 위해 프로젝트 관리 도구인 Jira 를 활용하려고 했지만 그러지 못한 점은 이번 프로젝트의 아쉬움 중 하나이다. Jira 를 통해 업무의 배분과 진행 상황을 체계적으로 관리할 수 있었음에도 불구하고, 팀원 간의 의사소통과 업무 추적이 원활하지 않았다. 향후 프로젝트에서는 Jira 사용에 대한 이해도를 높이고, 프로젝트 일정을 관리하며 담당자를 지정하여 업무 프로세스를 개선할 필요가 있다.

5-2. Confluence 의 일관성 부족

Confluence 를 활용하여 프로젝트 관련 자료와 지식을 공유했지만, 문서의 구조와 형식에 일관성이 부족했다. 이는 정보 검색의 효율성을 떨어뜨리고, 이해도에 차이를 발생시켰다. 문서화는 협업의 핵심 요소이므로, 통일된 템플릿과 명확한 작성 가이드라인을 마련하여 모든 팀원이 일관성 있게 정보를 기록하도록 하는 것이 중요하다. 이를 통해 원활한 지식의 축적과 전달을 기대할 수 있을 것이다.

5-3. GitHub 코드 활용의 재사용성 부족

GitHub 를 사용하여 코드를 관리하였지만, 코드의 재사용성과 효율성에 대한 고려가 부족했다. 모듈화나 함수화를 하여 코드를 작성하였으나, 유사한 기능을 반복적으로 구현하는 비효율성이 나타났다. 특히 모델을 피클화(pickle)하여 저장하지 않아 동일한 조건의 모델을 반복적으로 학습해야 하는 상황이 발생했다. 이는 불필요한 시간과 자원의 낭비로 이어졌으며, 모델 버전 관리에도 어려움을 초래했다. 다음 프로젝트에서는 코드를 작성할 때 재사용성을 염두에 두고, 모듈화와 함수화를 적용할 예정이며 모델의 피클화를 통해 학습된 모델을 저장하고 재사용하여 효율성을 높일 계획이다.

5-4. 최종 모델 선정의 지연

프로젝트 일정 관리의 미흡으로 인해 최종 모델 선정이 지연되었다. 팀원들이 각자의 모델 검증에 몰두하느라 서로 간의 협업이 원활하지 않았고, 이로 인해 최종 모델 선정이 늦어졌다. 마지막 날에 최종 모델의 하이퍼파라미터 실험을 진행하는 등 계획이 원활하게 이루어지지 않았으며, 이는 모델의 성능 검증과 최적화에 충분한 시간을 투자하지 못하게 하였다.

개인회고 (곽정무_T7504)

이번 대회를 시작하기 전 각자 달성하고 싶은 부분 키워드를 적고 대회에 진입했는데 저의 경우, 상세하게 EDA 하기, 줌 미팅 세션 진행할 때마다 회의록 남기기, 실무 관련 구현해보기였습니다. 주로 저번 프로젝트에서 아쉬웠던 부분을 이번 대회에서 달성하고 싶은 키워드로 남겼는데 대회가 끝난 후 팀원들과 대회 회고를 하며 적어도 키워드 부분에서는 많은 부분 달성했음을 확인했습니다.

첫 대회에 대한 설렘과 추석이라는 기간이 겹쳐 혼자 llm 을 활용해 빠르게 eda 를 한 후 모델링에 바로 진입했던 저번 대회와 달리 이번 대회는 협업을 우선시하고자 처음 다뤄보는 협업 툴인 Jira 와 Confluence 에 대한 습득부터 진행이 됐습니다. 대회가 시작되고 일주일 채 안 되어 휴가 기간이 시작될 예정되었고 협업 툴을 제가 주도적으로 관리하는게 팀원 간 합의가 되어 프로젝트 기간 할당된 수업도 제쳐두고 협업 툴이 우선시 되었고, confluence 와 관련되어 팀 협업 및 아이디어 보존에는 긍정적인 영향을 끼쳤으나 스스로 대회 자체에 들어가야 되는 집중이 분산됨을 적잖게 느꼈습니다.

대회와 관련해서는 제가 주로 시간과 집중을 할애했던 부분은 지역에 관한 부분이였습니다. 위치에 관한 데이터가 위도/경도로만 주어졌고 도메인 지식 상 부동산 가격은 위치가 제일 중요한 요소 중 하나이므로 어떠한 방법으로 지역을 나눌까에 대한 고민이 지속되었고 단순하게 k-means clustering 이나 dbscan 으로 지역을 나누는 것에 대한 의문이 많이 따랐습니다.

처음에는 아파트에서 가장 가까운 초등학교를 이용해 이를 아파트가 속해 있는 지역으로 자연스럽게 클러스터링을 하려고 했으나 초등학교의 수가 많아서인지 큰 효과를 발휘하지 못했고 지하철 역 또한 똑같은 결론이 나왔습니다. 이후 들었던 의문이 전체 데이터셋을 활용하는데 대한 의문이였습니다. 주어진 데이터셋은 수도권 전체 전세 거래인데 보통 한 아파트에 대한 전세 거래가를 알아보려 부동산에 방문한다면 부동산에서 얻는 정보는 당연하게도 그 아파트가 속한 동네 전세가 동향입니다.

따라서, 테스트 데이터셋에 있는 전세가를 예측할 때 필요한 정보는 예측하고 싶은 아파트 주변 정보라 판단하고 반경 1km 아파트 거래로 제한해 18000 개의 unique 한 아파트 거래가 있다면 18000 개의 다른 데이터셋으로 lightgbm 이나 부동산 정보는 변하지 않는 정보이므로 시계열 특성을 활용하기 위해 LSTM 활용해 보았으나 lightgbm 같은 경우 많은 시간이 소요되지 않았으나 LSTM 은 너무 오랜 시간이 걸려 끝까지 시도해보지 못한 아쉬움이 따랐습니다. 새로 만든 데이터셋들 중 샘플이 1000 개 정도만 있는 경

우도 있기에 MAE 가 높은 아파트들에 대한 앙상블을 고려해봄직 했으나 시간 부족으로 시도해 보지 못한 점은 굉장히 아쉽습니다.

개인적으로 이번 프로젝트는 저번 프로젝트와 비교해 협업 부분에서 많은 발전이 따랐으나 여전히 Jira 와 Github 부분에서 발전할 부분이 보여 다음 프로젝트에서 성과를 보여주고 싶습니다. 개인적인 측면에서는 스스로 많은 아쉬움과 실망감을 많이 느꼈습니다. 이번에 휴식 기간이 길었고 저번 프로젝트에 많은 몰입과 시간을 할애했던 때문인지 대회 초기 집중하는데 스스로 어려움을 겪었고 대회 내내 충분한 몰입이 안되었다 느껴 팀원들에게 미안함을 느꼈습니다.

다행히도 대회 기간이 지나가면서 다시 열정이 살아났고 이번 대회에서 아이디어 공유는 활발했으나 팀원들끼리 아이디어를 구현한 코드를 공유해 실험을 해 볼만큼 재사용성에 대한 충분한 고찰이 뒤따르지 않았다 느껴 다음 프로젝트에서는 실험 기록과 이를 다른 팀원이 재현하고 재사용할 수 있게 처음부터 고려하며 프로젝트를 진행하는게 주요 목표입니다.

개인회고 (박준하_T7523)

이번 프로젝트에서 가장 중점을 둔 부분은 외삽(extrapolation) 모델 개발이었습니다. 거래량이 많은 아파트의 가격 변화를 활용해 거래량이 적은 아파트의 가격 변화를 예측하는 방식이 직접적인 시계열 예측보다 간단하고 효과적일 것이라 판단했기 때문입니다. 하지만 기대와 달리 예측 성능이 저조해 여러 방안을 추가로 시도했습니다. 예를 들어, 아파트 임베딩에 관련된 feature 를 추가하거나, GCN 을 통해 인접한 아파트 간 임베딩 학습을 촉진하는 방법 등을 적용했습니다. 비록 이 방법들로도 성능 개선에 성공하지는 못했지만, 신경망 학습과 GNN 에 대한 이해를 높이는 데 큰 도움이 되었습니다.

이번 프로젝트에서 발전시킨 중요한 요소는 코드의 구조화와 모듈화였습니다. 이전에는 모델 학습과 추론을 주피터 노트북에서 진행해 코드 구조가 복잡하고 재사용이 어려웠습니다. 이번에는 모델을 관리하는 매니저 클래스를 만들어 학습과 추론을 체계적으로 관리할 수 있게 했습니다. 전처리 과정도 Preprocessor 클래스를 통해 일괄적으로 관리하여 각 모델에 필요한 전처리만 수행함으로써 코드 가독성과 재사용성을 높일 수 있었습니다.

저번과 달리 Confluence 를 통해 아이디어와 실험 결과를 공유하고 피드백을 주고받는 것도 이번 프로젝트에서 큰 도움이 되었습니다. 팀원들과의 소통을 통해 서로의 아이디어를 공유하고 발전시킬 수 있었으며, 다양한 시도를 통해 성능을 개선하는 데 도움이 되었습니다. 아예 기록 자체를 하지 않았던 저번 프로젝트에서 크게 발전한 것 같다고 생각합니다.

프로젝트를 진행하며 느낀 아쉬움은, 실패할 때마다 바로 포기하고 새로운 아이디어로 넘어가려 했던 점입니다. 외삽 모델을 개발하면서 여러 시도를 했지만, 성능이 나오지 않으면 쉽게 다른 접근으로 바꾸려는 경향이 있었습니다. 그러나 돌아보니 아이디어 자체는 유효했으며, 더 깊이 파고들고 발전시켰다면 더 좋은 성과를 낼 수 있었을 거라는 아쉬움이 듭니다. 이번 경험을 통해 앞으로는 실패하더라도 쉽게 포기하지 않고 아이디어를 충분히 탐구하고 개선해 나가는 노력이 중요하다는 것을 깨달았습니다.

다음 프로젝트에서는 코드의 구조화와 모듈화를 한층 더 발전시키고, 한 가지 아이디어를 깊이 있게 탐구하며 프로젝트에 임하겠다고 다짐합니다.

개인회고 (박태지_T7524)

이번 프로젝트에서 수도권 아파트 전세가격 예측을 수행하면서 여러 가지 부분을 얻을 수 있었다. 특히 데이터 선정과 모델링 과정에서의 의사결정, 그리고 협업과 기록의 중요성에 대해 깊이 있게 성찰하게 되었다.

1. 금리에 대한 지나친 집중과 다른 데이터 등한시

프로젝트 초기부터 금리가 부동산 시장에 미치는 영향에 주목하여 금리 데이터에 집중하였다. 금리는 전세가격에 큰 영향을 미치는 요인 중 하나이지만, 이에 지나치게 몰두한 나머지 다른 중요한 변수들을 상대적으로 등한시하였다. 예를 들어, 학교 데이터, 공원 데이터, 아파트 거래량 등 다양한 요인을 충분히 고려하지 못했다. 이는 모델의 예측 성능에 부정적인 영향을 미쳤으며, 복합적인 요인들이 상호 작용하는 부동산 시장의 특성을 충분히 반영하지 못하는 결과를 초래했다. 특정 변수에 과도하게 집중하기보다 다양한 변수들을 종합적으로 고려하여 모델의 전반적인 성능을 향상시킬 필요가 있다.

2. 얕은 도메인 지식에 의존한 제한적인 데이터 활용

개인적인 도메인 지식에 기반하여 2024 년을 예측하는 데에는 2023 년 데이터가 큰 영향을 끼칠 것이라고 판단하여, 2023 년 데이터만을 활용하여 모델을 구축하였다. 이는 과거의 추세와 패턴을 충분히 반영하지 못하게 하였으며, 데이터의 시간적 다양성을 확보하지 못하는 한계를 가져왔다. 부동산 시장은 장기적인 트렌드와 주기성이 존재하므로, 더 긴 기간의 데이터를 활용하여 모델의 안정성과 예측력을 높일 필요가 있었다. 추후 프로젝트에서는 개인적인 지식에만 의존하지 않고, 데이터 분석의 기본 원칙에 따라 충분한 양의 데이터를 수집하고 활용해야 할 것이다.

3. 클러스터링에 대한 과신과 성능 개선의 한계

아파트 단지별로 특성이 다를 수 있다는 판단 아래 클러스터링 기법을 도입하여 모델의 성능을 개선하고자 했다. 그러나 클러스터의 수를 아파트 단지 개수인 14,000 개까지 늘리는 등 성능 향상이 나타나지 않음에도 불구하고 클러스터링이 효과적일 것이라는 주장을 고집하였다. 이는 데이터 분석에서의 유연한 사고와 실험 결과에 대한 객관적인 평가의 중요성을 간과하였다. 모델의 성능이 개선되지 않는다면 접근 방식을 재고하고 다른 방법을 모색해야 함에도 불구하고, 특정 방법에 집착함으로써 시간과 자원을

비효율적으로 사용하였다. 실험 결과를 기반으로 한 객관적인 판단과 다양한 방법론의 적용을 통해 문제 해결에 접근해야 할 것이다.

4. AutoML 에 대한 이해 부족과 과도한 신뢰

AutoML 을 활용하여 모델링 과정을 자동화하고자 하였으나, 충분한 이해도 없이 지나치게 신뢰하였다. AutoML 은 모델 선정과 하이퍼파라미터 튜닝 등을 자동으로 수행해주지만, 데이터의 특성이나 문제의 도메인 지식이 반영되지 않을 수 있다. 이에 따라 모델의 성능이 기대에 미치지 못하였고, 문제의 원인을 파악하고 개선하는 데 어려움을 겪었다. AutoML 은 도구일 뿐이며, 최종적인 의사결정과 모델의 해석은 개인의 몫임을 인식하고, 결과를 비판적으로 검토하고 보완할 필요가 있다.

5. 모델 실행 계획과 결과 기록의 미흡

모델의 실행 계획과 결과를 상세하게 기록하지 않은 점은 큰 아쉬움으로 남는다. 어떤 실험을 수행하였는지, 그 결과는 어떠하였는지에 대한 명확한 기록이 남아 있지 않아 실험의 재현성과 결과의 해석에 어려움이 있었다. 이는 팀원 간의 지식 공유를 저해하고, 동일한 실험을 반복하게 만드는 비효율성을 초래하였다. 데이터 분석에서는 실험 과정과 결과를 철저하게 문서화하여 향후 참고할 수 있도록 하는 것이 중요하다. 추후에는 실험 계획 단계에서부터 상세한 기록을 남기고, 결과를 체계적으로 정리하여 공유함으로써 협업의 효율성을 높이도록 할 것이다.

6. 결론

이번 프로젝트를 통해 데이터 분석에서의 객관성, 유연한 사고, 철저한 기록의 중요성을 다시 한번 깨닫게 되었다. 특정 변수나 방법론에 지나치게 의존하기보다 다양한 관점을 수용하고, 실험 결과를 기반으로 한 객관적인 판단을 내릴 수 있어야 한다. 또한 도구의 활용에 있어서도 그 한계를 인지하고, 도구에 대한 깊은 이해를 바탕으로 활용해야 한다. 마지막으로, 체계적인 기록과 지식 공유를 통해 협업의 시너지를 극대화할 수 있도록 노력할 것이다. 이번 교훈을 통해 다음 프로젝트에서는 부족한 점을 보완하며 더욱 발전된 모습으로 임할 수 있을 것이다.

개인회고 (신경호_T7530)

프로젝트를 진행하면서 데이터에 관한 고심이 더 깊어졌습니다. 비트코인 가격 예측에서 주어진 데이터는 raw 해서 수많은 피처 엔지니어링이 가능하였지만, 큰 영향을 주는 피처를 찾기가 어려웠습니다. 이번에 주어진 데이터는 정제되어서 초반에 생각하였던 피처 엔지니어링 이후 새로운 피처를 생성하기가 어려웠습니다.

목표를 달성하기 위해서 하이퍼파라미터 튜닝과 이상치 탐지 위주의 실험을 진행하였습니다. 이전에는 wandb 를 사용하여서 실험 결과를 시각화 하는 것에는 좋았지만, 한번의 실행으로 파라미터를 도출하여 모델에 학습하는 파이프라인을 구축하기 어려웠습니다. 그래서 이번 프로젝트는 optuna 를 사용하여 하이퍼파라미터 서치에서 학습까지의 파이프라인을 구축하였습니다.

EDA 를 진행하면서 같은 면적, 층, 위치의 데이터가 다른 가격으로 거래된 데이터를 발견하였습니다. 이에 의문을 품고 데이터를 분석하니 다른 동의 거래도 같은 위치로 저장된 것을 발견하였습니다. 이를 보고 이전에 동일한 데이터가 거래된 이력도 이상치가 아닌 정상적인 거래라는 것을 파악할 수 있었습니다.

같은 매물의 거래 금액이 1 ~ 24000 까지의 차이가 났기에 결과에 많은 영향을 준다고 생각하여 평균값 변경, 최대값 변경, IQR 등 이상치를 변경하였지만 결과에 부정적인 영향을 주었습니다. 다음에는 Z-score, 평균분산 등 더 다양한 방법을 적용하겠다고 생각하였습니다.

실험 과정과 결과를 공유하기 위하여 confluence 를 만들었는데, 성능을 개선하지 못한 실험을 많이 기록하지 못하였습니다. 제가 한 실험이 다른 팀원에게 새로운 영감을 줄 수 있는데 이를 기록하지 않았던 부분이 아쉬웠으며, github 컨벤션도 익숙하지 않아서 작은 실수가 있었습니다.

하지만, 팀원이 구축한 모듈화 프로세스를 해석하며 필요한 부분을 제작하여 새로운 모듈을 제작한 경험은 차후에도 많은 도움이 될 것이라 생각합니다. 다음번에 기회가 되면 제가 모델 프로세스를 구축하고 싶습니다.

개인 회고 (이효준_T7545)

지난 프로젝트에서는 진행과정 중 소실된 아이디어들에 대한 아쉬움이 컸기에, Jira와 Confluence를 적극 활용해 모든 아이디어와 실험 과정을 기록하고자 했다. 이를 위해 프로젝트 시작 전, 팀원들과 협업 도구를 설정하고 준비하는 회의를 먼저 진행했다. 또한, 다양한 실험을 통해 프로젝트 경험을 쌓는 것을 목표로 설정하여 여러 가설을 교차 검증하며 모든 변수와 모델 조합 중 최적의 결과를 도출하는 데에 중점을 두고 프로젝트를 진행했다.

EDA 결과, 데이터에 이상치가 많이 존재함을 확인했으며 이를 관리하기 위해 RMSE와 Adjusted R^2을 평가지표로 함께 사용했다. 또한, 주요 변수인 지역 정보를 대체하기 위해 클러스터링 기법을 적용했고, 최적의 클러스터 수를 찾기 위해 일주일간 24시간 내내 클러스터 수를 조정하며 최적 값을 찾고자 노력했다.

클러스터링 변수 생성 과정에서는 큰 성능 향상을 이루지는 못했지만, 클러스터링에 활용되는 지표들에 대해 학습할 수 있었다. 특히, 실루엣 스코어와 Davies-Bouldin 스코어를 함께 사용하여 포인트와 클러스터 단위를 모두 고려해 최적의 클러스터 수를 선정하고 클러스터링 품질을 높이기 위해 노력했다. 이 과정을 통해 여러 지표를 상호보완적으로 해석하여 평가의 질과 일관성을 높일 수 있었다.

이전 프로젝트에서는 타겟 변수를 다양한 관점에서 해석하지 못한 아쉬움이 있었기에, 이를 보완하고자 면적당 전세가가 기존 전세가에 비해 변동 폭이 작다는 인사이트를 바탕으로 면적당 전세가 예측을 시도했다. 교차 검증 결과, 면적당 전세가 예측이 기존 전세가 예측에 비해 MAE 기준으로 약 100 정도의 성능 향상을 보여주었고, 특히 평균 MAE가 높은 아파트의 변동성을 안정화하는데 도움이 되었다.

클러스터링 변수 생성 과정에서 계산 비용 문제를 해결하기 위해 GPU 기반의 cuML을 사용하고 클러스터링에 활용하는 변수조합을 변경하는 등 여러 방안을 활용해 최적화를 진행했다. 이러한 과정에서 하나의 기법에만 집중하기보다, 초기 실험 단계에서 다양한 기법을 로깅하여 비교 후 최적의 기법을 선정한 다음, 정밀 실험을 진행하는 방식이 효과적이라는 것을 배웠다.

협업 과정에서는 회의록과 가설을 정리하는 데 있어 이전보다 발전이 있었지만, Jira 사용에 있어 다소 한계를 느꼈고, 프로젝트 초기 단계 각자의 가설에 대한 실험을 진행하는 과정에서 Github을 효율적으로 활용하는 데 어려움이 있었다. 개인적으로는 여러 가설을 코드로 재현하는데 한계를 느꼈으며, 다양한 모델을 구현해보지 못한 데에 대한 아쉬움이 있다.

다음 프로젝트부터는 본격적인 추천시스템 프로젝트가 진행되는데, 이전에 논문에서 리뷰한 추천시스템 모델들을 최대한 많이 구현해 보는 것이 주요 목표이다. 또한, 추후에도 사용할 수 있도록 팀의 Github 구조에 맞춰 해당 모델들을 체계적으로 구성하고 관리하는 것이 또 다른 목표이다.