Wrap-up Report

RecSys_3조 (ㄱ해줘)

강성택, 김다빈, 김윤경, 김희수, 노근서, 박영균

1. 프로젝트 개요

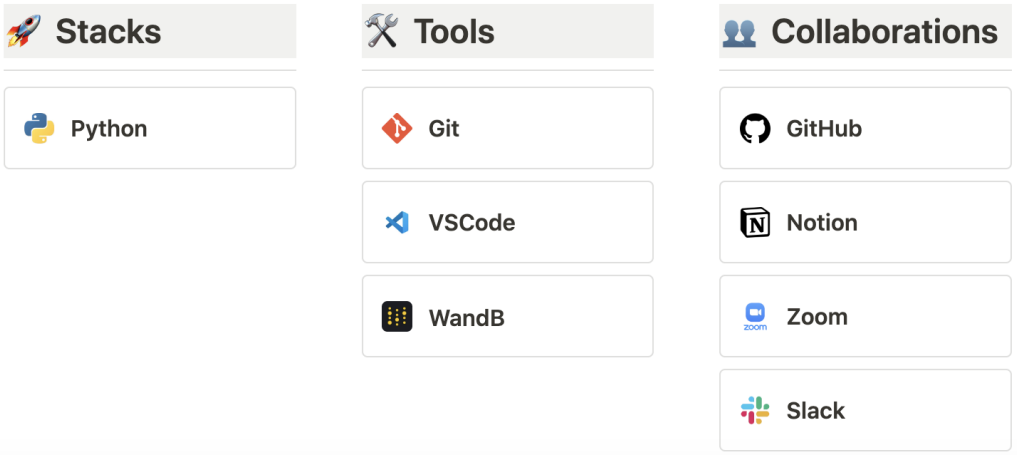
1.1 개요

전세 시장은 매매 시장과 밀접하게 연관되어 있어, 부동산 정책과 시장 예측 중요한 지표가 된다. 특히 전세 시장의 동향은 매매 시장과 밀접하게 연관되어 있어 부동산 정책 수립 시장 예측에 중요한 지표로 활용된다.

아파트의 주거 특성과 금융 지표 등 **다양한 데이터를 바탕으로 전세가를 예측하여 부동산 시장의 정보 비대칭성 해소에 기여**하는 것이 목표이다.

1.2 환경

- (팀 구성 및 컴퓨팅 환경) 6인 1팀, V100 서버를 VSCode와 SSH로 연결하여 사용
- (협업 환경) Notion, GitHub, WandB
- (의사 소통) Zoom, Slack



2. 프로젝트 팀 구성 및 역할

데이터팀 / 모델팀으로 3:3 나눈 후, 각 팀마다 한 명씩 짝지어서 페어 프로그래밍 방식으로 진행

공통

서버 구축, EDA

Data 팀

- 공통 - 베이스라인 구축, 모듈화
- **김희수**

  전처리(결측치 처리 · 급처매물 제거 함수 구현), 파생변수 생성(최단거리 장소 개수)
- **노근서**

  파생변수 생성(특정 반경 내 공공장소 수, 클러스터링 후 대장아파트까지의 최단거리, 특정 조건 만족하는 공원 인접 유무), 전처리(이상치 처리 함수, 위·경도 중복도가 낮은 데이터 제거 함수 구현), 모델(TabNet)

- **김윤경**

  파생변수 생성(최단 거리 및 위치), 클러스터링(K-means, DBSCAN, GMM) 함수 구현, 모델(GradientBoosting, TabNet)

Model 팀

- 공통 - 베이스라인 구축, 모듈화, Optuna, K-fold
- **김다빈**

  LightGBM, WandB, Random Sampling, DBSCAN, Classification 활용 Regression 예측

- **박영균**

  CatBoost, DenseNet, Ensemble(stacking, voting), 주변 인프라에 따른 Regression 예측

- **강성택**

  XGBoost, Randomforest, ArgParse, KMeans, GMM, Classification 활용 Regression 예측

---

3. 프로젝트 수행 절차 및 방법

3.1 팀 목표 설정

진행 기간 : 10/2 (수) 10:00 ~ 10/24 (목) 19:00

- (1주차) 강의 전부 듣기, 데이터셋 이해, 각자 EDA 완성, 각 EDA별 베이스라인 코드 작성
- (2주차) 과제 전부 완료, 데이터팀/모델팀 코드 모듈화
- (3주차) 각 팀마다 EDA를 통해 얻은 인사이트 + 파생변수를 활용해 실험 시작 및 앙상블 도입

3.2 프로젝트 계획

1. Project Rule, Github/Coding Convention 설정
2. 프로젝트 개발 환경 구축 - 서버 세팅, Github
3. 데이터셋 사전 공부
4. EDA 진행 및 인사이트 도출
5. 데이터팀/모델팀 베이스라인 코드 개발 & 모듈화
6. EDA별 실험 및 성능 비교(WandB 사용) & 모델 튜닝 : 교차 검증, 하이퍼파라미터 튜닝
7. 앙상블, 딥러닝 모델
8. 코드 리팩토링 & 프로젝트 코드 최종 모듈화

3.3 협업 방식

- 이슈 관리 - Github의 Issues + Notion 데이터베이스 보드로 진행
  - Github의 Issues : Issue Templete(개요, TODO 작성)
  - 총괄 보드 : 전체 진행 상황 관리(일정, 제출 관리 등)
  - Data 보드 : 데이터 전처리, 피처 엔지니어링 함수 및 모듈화 진행상황 관리
  - Model 보드 : 모델 설계, 모듈화, 앙상블 개발 등 진행상황 관리
- Github Convention에 따라 Github 관리
- Coding Convention에 따라 개발

- WandB를 활용한 실험 관리

---

4. 프로젝트 수행 결과

4.1 탐색적 데이터 분석(EDA)

- 오른쪽으로 꼬리가 긴 분포를 갖는 변수들은 로그 변환 또는 제곱근 변환 고려
- 건축 연도(`built_year`)와 건물 나이(`age`)간 상관계수가 -0.99이지만 `age`는 `built_year`로 만든 파생변수이므로 다중공선성을 고려해 하나만 사용
- 위·경도가 중복되는 수를 세어 중복이 낮은 것들을 제거했을 때 `MAE`가 낮아진다.
- 전세가(`deposit`), 위·경도(`latitude`, `longitude`)를 활용한 클러스터링을 적용해 지역별 범주화 필요

4.2 데이터 전처리

- EDA 실험 결과를 바탕으로 위·경도가 중복되는 수가 낮은(2~5) 데이터를 제거
- 건축 연도(`built_year`)가 2024인 데이터 또한 이상치로 판단하고 제거
- test data의 금리(`interest_rate`) 결측치를 평균값으로 대체

4.3 피처 엔지니어링

[파생변수 생성]

- 건물-공공장소 사이의 최단거리와 그 위치에 해당하는 공공장소의 위·경도
- 최단거리에 위치한 공공장소 수
- 건물 기준, 특정 반경 내 공공장소 수 (공원은 **100,000m^2 이상**인 공원만 활용 - 도시지역권 근린공원)
  - 특정 반경 내 도시지역권 근린공원 유무 변수
- 위 · 경도 기준 k-means 클러스터링 후,
  - 클러스터별 평균 전세가
  - 클러스터링 내 대장 아파트(최고 거래가 건물) 생성
  - 건물-대장 아파트 사이의 최단거리와 그 위치에 해당하는 대장 아파트의 위·경도

4.4 모델 개요

[아이디어]

1. 위·경도를 기준으로 클러스터링 후, 지역별 가격 편차를 반영해 평균 전세가(`region_mean`)를 계산하여 feature로 추가. 이는 지역별로 다른 전세가 트렌드를 반영할 수 있어 `deposit` 예측 정확도를 높이는 데 기여
2. 주거지의 접근성과 편의성을 고려해 주변 인프라(지하철역, 학교, 공원)까지의 거리와 특정 반경 내 인프라 수를 feature로 추가. 또한 각 건물별 대장 아파트와의 거리도 추가하여 고가 지역을 중심으로 한 전세가 패턴을 반영. 이러한 feature들은 입지의 가치를 반영하기 때문에 `deposit` 예측 정확도를 높일 것이라 예상
3. `deposit`을 로그 변환하여 N개의 `deposit_group`으로 나눈 후, 그룹 예측에는 분류 모델을 사용하고 그룹별로 개별 회귀 모델을 학습하여 최종 `deposit`값을 예측하는 방식 적용. 아파트마다 특징이 다르기 때문에 한 그룹으로 모델링을 하는 것 보다 여러 개의 그룹으로 나누어 모델링을 하는 것이 모델의 정확도를 높일 것이라 예상

4.5 모델 선정 및 분석

[모델 비교]

다양한 모델(XGBoost, LightGBM, Random Forest, CatBoost, TabNet, DenseNet, Gradient Boosting)을 활용하여 `deposit` 예측 성능을 비교. 모델 성능 비교에 사용한 평가지표는 `MAE`를 활용

[앙상블 기법 분석]

성능 개선을 위해 Stacking, Soft Voting 앙상블을 활용하여 성능을 높이고자 하였으나, 트리 모델 간 유사성이 높아 예상만큼의 성능 향상을 이끌어내지 못했다.

[최종 모델 선정]

XGBoost 모델은 데이터의 비선형성을 효과적으로 처리하며, 주요 하이퍼파라미터 최적화 후 높은 예측 정확도를 기록. 특히 `deposit_group` 분류와 그룹별 회귀 모델 모두에서 XGBoost의 성능이 두드러져 최종 모델로 선정

[모델 튜닝]

- **k-fold** 교차 검증 적용
- **Optuna**를 활용한 하이퍼파라미터 튜닝
- **Soft Voting**을 적용한 **앙상블**로 최종 결과 도출. 모델의 가중치는 **Optuna**로 최적화
- **Stacking**을 적용한 **앙상블**로 최종 결과 도출. 모델의 가중치는 **Optuna**로 최적화

4.6 모델 평가 및 개선 방법

[평가]

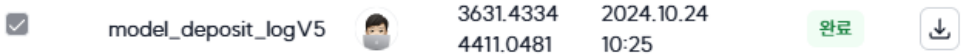
모델의 성능 평가는 두 가지 지표를 사용하여 수행했다.

1. `deposit_group` **분류 평가 :** `deposit`을 범주형 변수로 그룹화한 `deposit_group`을 예측할 때는 `mlogloss`를 사용. `accuracy`는 전체 데이터셋에서의 정확한 예측 비율을 보여주는 반면, `mlogloss`는 예측의 신뢰성을 평가하기 때문에 확률적 예측을 반영한다는 점에서 채택
2. **회귀 모델 평가 :** 최종적으로 예측된 `desposit` 값에 대해서는 `MAE`를 사용하여 모델 성능을 평가

[개선 방법]

- **feature selection :** EDA를 통해 예측과 관련성이 낮은 변수들을 제거하여 모델의 복잡성을 낮추고, 중요한 변수만을 사용함으로써 성능을 향상. 변수 선택을 통해 모델의 과적합을 방지하고, 보다 효율적인 feature set으로 예측 성능을 개선
- **로그 변환 :** EDA 결과, 오른쪽으로 꼬리가 긴 분포를 갖는 변수들은 로그 변환을 적용

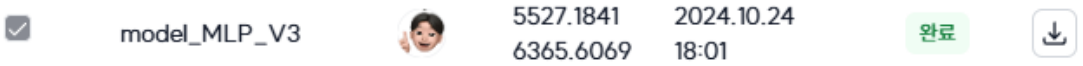
4.7 모델 성능 및 결과



제출 결과 1. deposit_group을 활용한 deposit 예측

deposit_group	valid MAE (mean)	public MAE	최종 MAE	group_score
n = 3	4410.33935	3631.4334	4411.0481	{0: 87.37584114074707, 1: 4407.927578889775, 2: 54825.825}

→ **valid MAE와 최종 MAE가 비슷하게 나왔다.**



제출 결과 2. 딥러닝 모델 테스트 결과

valid MAE	public MAE	최종 MAE
5895.7040	5527.1841	6365.6069

→ **최종 MAE를 보니 약간 뒤처지는 성능을 보였다.**

4.8 시행 착오

[문제]

- `deposit_group`을 15개의 세분화된 그룹으로 설정했을 때 모델의 valid 점수와 실제 public 점수간 격차가 크게 발생했다.

deposit_group	valid MAE (mean)	public MAE	최종 MAE
n = 15	2632.1164	3934.7548	4744.3762

[원인]

- 15개의 `deposit_group`으로 그룹화한 모델의 경우, 너무 세밀하게 그룹을 나눔으로써 과적합이 발생할 가능성 존재
- 잘못된 분류 예측으로 인한 `deposit` 예측에서의 loss 발생

[해결 방안]

- n = 3으로 설정하여 일반적인 경우/낮은 경우/높은 경우의 집값 예측 흐름을 반영하고자 했다.
- 포괄적인 그룹으로 범주화하여 분류 오류로 인한 `deposit` 예측의 손실을 줄이고자 했다.

[결과]

- valid 점수와 최종 점수가 거의 **일치**하는 모습을 확인

deposit_group	valid MAE (mean)	public MAE	최종 MAE
n = 3	4410.33935	3631.4334	4411.0481

[MLP 구현 과정]

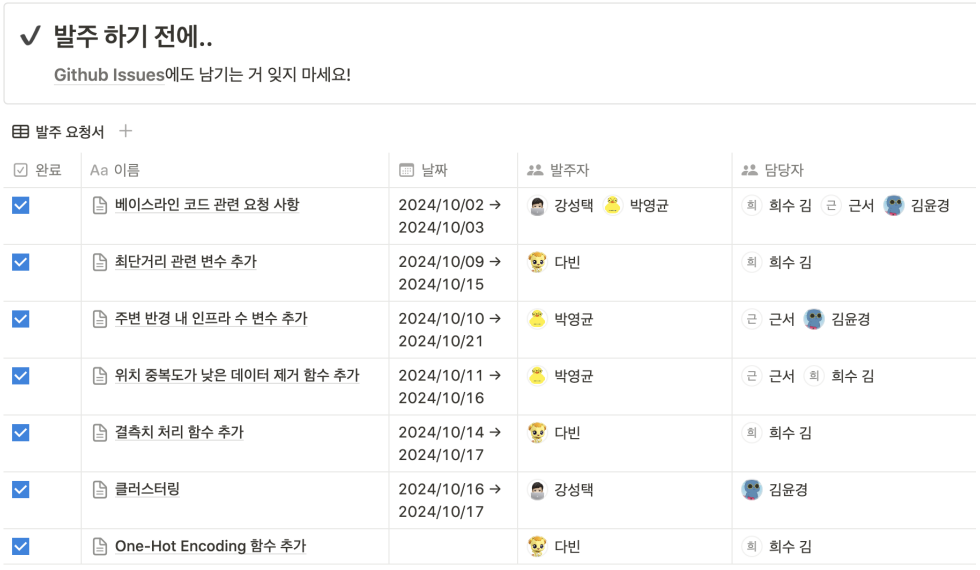
- 기본적인 MLP 모형을 이용해 예측을 수행하는 것이 복잡한 딥러닝 모델을 이용하는 것보다 효과적일 수도 있다고 생각하여 구현해보았다.
- 하지만 너무 간단한 모형이어서 그런 것인지 뉴런 개수 등을 조정하는 등의 조치로는 성능이 좋아지지 않았다.
- 이에 다방면으로 실험하면서, 피처/모델의 다양성을 고려하여 모델을 적용하면 보다 좋은 예측이 가능할 것이라는 가설을 세웠다.
- 최종적으로 정한 피처들의 2차항을 추가하여 총 90개의 피처로 만들어 사용하고, 모델의 마지막 활성화 함수를 ReLU 대신 SoftPlus로 대체하여 실험해보았다.
- 그 결과, 다섯 자리에서 네 자리까지 `MAE`를 낮추었고, 최종적으로 5000대까지 성능 향상을 이루었다.

---

5. 자체 평가 의견

잘한 점

- **EDA 기반 구현** : 각자의 EDA를 기반으로 데이터의 다양한 인사이트를 도출했다. 이를 바탕으로 가설을 설정한 뒤에 실험을 설계하여 모델 성능을 향상시킬 수 있었고 원인 • 결과 해석이 용이했다.
- **다양한 협업 툴 활용** : Github Issues, PR Templete을 적극적으로 사용해 서로 겹치는 작업 없이 협업을 진행할 수 있었다. 또한 실험 관리에 WandB를 사용해서 모델 학습 및 실험 결과를 이전 프로젝트보다 체계적으로 기록하고 시각화 할 수 있었다.
- **확실한 역할 분담** : 모델팀에서 필요한 파생변수가 생기면 회의를 거쳐 Issue와 발주서를 작성하고, 데이터팀에서는 담당자가 파생변수를 만들고 EDA를 진행 후 내용 검토를 받은 후 PR을 남겼다. 이후 모델팀에서 새 변수를 썼을 때 성능 변화를 WandB에 남겨 놓아 전체 작업 과정 흐름을 기록해 실험 추적이 가능해졌다.



- **사전 목표와 계획 설정** : 지난 프로젝트 회고에서 목표와 계획을 뚜렷하게 한 청사진을 그리고 시작해보자고 다짐했다. 프로젝트 기간 초반에 주차별 계획과 프로젝트 전반적인 목표를 먼저 설정해두고, 하루 단위로 일정 분배를 하고 매일 To-Do 리스트를 작성해 단기적인 목표를 따라가는 방식으로 진행했다. 그 결과, 지난 프로젝트는 4일간 스프린트를 한 반면 이번 프로젝트는 4주간 정해진 일정 안에서 움직여 제출 횟수 1위를 할 수 있었다.

아쉬운 점

- **GPU 사용** : 여러 시도를 했지만 결국 LightGBM 모델 학습시 GPU를 활용하지 못해 많은 학습을 시켜야 하는 실험 단계에서 주어진 서버 자원을 최대로 활용하지 못했다.
- **딥러닝 모델 :** 딥러닝 모델 특성상 긴 학습 시간이 필요하지만 늦게 시도해본 탓에 많은 실험을 해보지 못했다. 특히 모델의 복잡성과 한정된 시간으로 인해 최적의 하이퍼파라미터를 찾기 어려웠고, 결국 기대한 성능을 달성하지 못했다.
- **단일 평가지표 사용** : 회귀 모델 평가시 리더보드에 사용되는 `MAE`만 사용해 아쉬웠다. `MSE`, R^2 등 다양한 회귀 모델 평가지표가 있지만 이를 사용하지 않아, 큰 값을 갖는 오차나 모델의 설명력을 충분히 파악하지 못해 추가적인 성능 개선 기회를 놓쳤던 것 같다.

이번 프로젝트를 통한 배운 점

→ 3 GUNA!

1. 모듈화를 잘 하면 이렇게 좋구나!
2. WandB와 노션 데이터베이스를 활용했더니 이렇게 좋구나! (실험 및 제출 관리의 필요성)
3. 그리고 언제나 도메인 지식은 알면 알수록 중요하구나!

다음 프로젝트에 적용해볼 점

1. LightGBM을 제외한 나머지 모델에서는 GPU를 잘 활용하고 있다고 생각했는데 스페셜 피어세션 시간에 이야기해보니 GPU를 사용하지 않은 걸로 판단된다. 다음 프로젝트에 꼭 개선시켜야 할 부분이다.
2. 리더보드 평가지표 이외에 모델의 특징을 파악할 수 있는 다양한 평가지표 사용해보자. 그러면 다양한 평가지표를 활용해 보다 일반화 성능을 올릴 수 있을 거라 생각한다.
3. 사용하는 모델을 이해한 후 구현해보자. 그래야 성능이 좋든, 나쁘든 원인 분석과 개선 방안을 더 수월하게 할 수 있다.

---

6. 개인 회고

강성택_T7501

1. **나의 학습목표**
   이번 프로젝트의 학습 목표는 전세가 예측 모델을 구현하며, 다양한 피처 엔지니어링 기법과 하이퍼파라미터 튜닝을 통해 머신러닝 모델의 성능을 최적화하는 것이었다. 또한, 성능 향상을 위해 Optuna 최적화 및 앙상블 모델 적용과 같은 기술을 연습하는 것이 주요 목표였다. 이와 더불어, 모델링 전 과정에서 데이터 분석과 전처리의 중요성을 재확인하며, 이를 실제 상황에 적용하는 능력을 키우고자 했다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**
   저번 프로젝트에서는 데이터 팀으로서 프로젝트에 임했다면, 이번 프로젝트에서는 모델 팀의 역할을 맡아 모델 설계와 성능 향상에 더 깊이 관여했다. 또한 이번에는 WandB를 적극적으로 활용해 실험 추적과 하이퍼파라미터 튜닝을 관리하였으며, WandB의 실험관리 기능 덕분에 각 실험의 설정, 성능 지표, 손실 곡선등을 손쉽게 모니터링할 수 있었고, 최적 모델의 하이퍼파라미터를 빠르게 찾을 수 있었다.

2. **개인 학습 측면**
   개인 학습 측면에서는 주어진 문제를 다양한 각도에서 분석하며, 문제 해결에 있어 데이터를 다루는 기법과 모델 최적화 방법을 체계적으로 연습할 수 있었다. 특히, Optuna를 통한 최적화는 이전에 비해 모델 성능을 극대화할 수 있는 방향성을 제시해 주었고, 이를 통해 모델링에 있어 하이퍼파라미터의 영향력을 실감할 수 있었다. 또한, 데이터 시각화와 EDA 과정을 통해 데이터의 패턴을 더욱 깊이 있게 이해하게 되었으며, 예측 모델의 신뢰성을 높이는 데에 큰 도움이 되었다.

3. **공동 학습 측면**
   팀원들과 함께 실험 및 결과 분석을 공유하는 과정을 통해 다양한 피드백을 얻을 수 있었다. 각기 다른 접근 방식을 바탕으로 토론하며 효율적인 피처 엔지니어링 방법과 하이퍼파라미터 조정의 중요성을 다시 한번 배울 수 있었다. 또한, 공동 학습을 통해 최적의 조합을 찾기 위한 다양한 접근법과 아이디어를 공유하며 보다 나은 결과를 도출할 수 있었다.

4. **기억에 남는 시도**
   가장 기억에 남는 시도는
   `deposit_group`을 세분화하여 전세가 수준에 따른 그룹별 맞춤형 회귀 모델을 사용했던 것이다. 이 과정에서 최적의 그룹 개수를 찾는 시도를 여러 번 거듭하며, 예측의 성능과 실제 적용 가능성을 모두 고려한 결정을 내려야 했다. 이를 통해 데이터의 특성에 따라 분류와 회귀를 결합하는 방법의 장점을 체감할 수 있었고, 다양한 실험을 통해 최적의 결과를 얻을 수 있었다.

5. **한계와 아쉬움**
   프로젝트를 진행하면서 피처 엔지니어링과 그룹화의 중요성을 깨달았지만, 복잡한 데이터를 다루면서 최적의 피처를 찾는 데 있어 한계가 있음을 느꼈다. 특히, 지역 특성을 반영하는 피처 생성에서 추가적인 데이터가 부족해 성능을 더 향상시키지 못한 점이 아쉬움으로 남았다. 또한, 비슷한 트리 기반 모델을 사용한 앙상블 방식에서 성능 개선의 한계를 느꼈으며, 다양한 모델을 조합하여 최적의 앙상블을 시도해 보지 못한 점이 아쉬웠다.

6. **앞으로의 도전**
   앞으로는 다양한 데이터를 추가하여 더 세분화된 피처를 만들고, 더 나은 예측력을 갖춘 모델을 개발하는 데 도전하고자 한다. 특히, 이번에 사용하지 못한 다양한 앙상블 기법이나 딥러닝 기반의 예측 모델을 사용하여 모델링의 효율성을 높이고자 한다. 또한, 여러 산업 분야에서 적용할 수 있는 고도화된 데이터 전처리 기법을 배우고, 최적화의 효율성을 높이는 연구를 지속적으로 해 나갈 계획이다.

김다빈_T7558

1. **나의 학습목표**

   이번 프로젝트에서는 팀원 모두가 일정 수준의 도메인 지식을 갖추고 있어, 이를 효과적으로 활용하여 의미 있는 결과를 도출하는 것이 목표였다. 팀원들과 함께 EDA 단계에서부터 논의하고 가설을 제시하는 과정이 매우 흥미로웠고, 이를 통해 더욱 깊이 있는 통찰을 얻을 수 있었다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

   이전 프로젝트에서는 데이터 전처리 작업을 주로 담당했지만, 이번에는 모델링을 맡게 되어 큰 변화가 있었다. 다양한 모델을 실험하며 각 모델의 특성과 장단점을 분석하는 데 중점을 두었다. 이전 프로젝트에서 사용한 Tree 모델에만 의존하지 않고, 여러 모델을 시도해보려 했으나, 최종적으로는 Tree 모델에 국한된 점은 아쉬움으로 남는다. 다양한 접근 방식을 통해 더 나은 결과를 도출할 수 있었다면 좋았을 것 같다.

2. **개인 학습 측면**

   담당한 부분은 충분히 이해하고 다른 팀원들에게 설명할 수 있을 정도로 학습하자를 목표로 했다. 모델팀으로서 실험에 어떤 모델을 사용하면 왜 그 모델을 사용했는 지, 어떤 효과를 기대하는 지를 설명하려고 노력했다.

3. **공동 학습 측면**

   이전 프로젝트와 마찬가지로 역할 분담을 분명히 해서 공동 학습을 효율적으로 진행했다. 특히 이번 프로젝트에서는 WandB를 도입해서 실험 추적을 자동화하는 것과 동시에 다른 팀원이 어떤 실험을 진행하고 있는 지도 실시간으로 확인할 수 있어 좋았다. 추가로 Notion를 메인으로 GitHub Issues와 Pull Request를 활용해 프로젝트 진행상황과 각 함수에 대한 설명을 기록해 좀 더 체계적인 틀 안에서 협업했다.

4. **기억에 남는 시도**

   모델 베이스라인을 모듈화하고 이를 단계적으로 수정하면서 실험을 진행하려고 했으나, 실제로는 베이스라인과 크게 다른 방식으로 실험하게 되어 ipynb를 사용한 점이 아쉬움이 남는다. 특히 `deposit`을 직접 예측하는 것이 아니라 `deposit_group`으로 세분화한 후 예측하는 방식으로 접근한 점이 기억에 남는다. 이러한 새로운 시도를 통해 다음에는 모듈화 설계 시 다양한 실험 가능성을 고려해야겠다고 다짐했다.

5. **한계와 아쉬움**

   WandB를 활용했지만, 그 기능을 최대한으로 활용하지 못한 점이 아쉽다. 모델 생성 시 사용한 하이퍼파라미터를 기록하고 활용하려 했지만, 데이터를 리스트 형태로 넘기면서 복사하는 과정에서 어려움을 겪었다. 기록으로서의 역할은 충분했지만, 향후에는 JSON 형태로 데이터를 넘기는 등 이전 실험 기록을 다른 실험에 효율적으로 활용하는 방법을 모색해야겠다고 생각했다. 또한, 다양한 EDA를 시도했지만 개인적으로 집중했던 부분에 대한 실험을 하지 못한 점도 아쉽다.

6. **앞으로의 도전**

   실험을 진행하면서 valid MAE를 기반으로 public MAE를 예측하여 성능 향상이 기대되는 실험을 선정하려 했으나, 각 실험마다 valid MAE와 public MAE 간의 차이가 커서 활용에 어려움을 겪었다. 따라서 다음 프로젝트에서는 다양한 평가지표를 활용하여 일반화 성능을 체크할 수 있도록 해서, 더욱 신뢰성 있는 결과를 도출할 수 있도록 노력해야겠다.

김윤경_T7511

1. **나의 학습목표는 무엇이었나?**

   이번 프로젝트 목표는 데이터 관련 프로세스를 전반적으로 이해하고 구현해보는 것이였다. 특히 EDA를 통해 얻은 인사이트를 바탕으로 feature engineering을 진행하고, 도메인 지식을 적용해 모델이 필요한 입력 데이터를 준비하는 데 중점을 두었다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

   전 프로젝트에서는 주로 모델링 역할을 담당해 알고리즘 선택과 하이퍼 파라미터 튜닝에 집중했다면 이번 프로젝트에서는 데이터 관련 프로세스를 담당했다는 점에서 큰 변화가 있었다. 이를 통해 데이터 품질과 특성, 도메인 지식의 중요성을 깨달았고, 모델링 경험을 기반으로 적합한 변수를 생성함으로써 데이터의 의미를 강화할 수 있었다.

3. **개인 학습 측면**

   개인적으로 도메인 지식에 맞는 파생 변수를 생성하면서 공간적 데이터 처리에 대해 자세히 배울 수 있었다. 특히 건물과 공공장소 간 최단거리 변수를 생성하기 위해 Balltree를 이용해 각 지점 간 거리 계산을 효율적으로 할 수 있었고, 실제 지리적 정보(위도,경도)로 구면 거리를 계산하기 위해 Haversine 공식을 적용하였다. 이를 통해 데이터의 공간적 관계를 잘 표현하는 변수를 생성할 수 있었다.

4. **공동 학습 측면**

   프로젝트의 협업 효율성을 향성시키기 위해 집중했다. 먼저, 모듈화를 진행하고 WandB를 도입하였다. 이를 통해 코드의 재사용성과 유지보수성이 높아졌고, 더 체계적인 실험 관리를 할 수 있었다. 또한, Github의 issue, PR 템플릿을 활용해 팀원 간 피드백 과정을 명확히 할 수 있었다. 결과적으로 전 프로젝트보다 더 체계적인 협업을 할 수 있었다.

5. **기억에 남는 시도**

   각각의 클러스터링 기법(K-means, DBSCAN, GMM)에 대해 최적의 클러스터 수를 찾는 함수로 모듈화 한 후 적용하면서 각 알고리즘의 특성을 비교할 수 있었다. 결과적으로 대용량 데이터를 이용하다보니 계산 속도가 상대적으로 빠른 K-means를 선택하게 되었는데, 다른 팀 사례에서 DBSCAN이 효과적으로 군집화를 형성했다라는 점에서 놀랐다. 이를 통해 너무 속도만을 고려한 선택이였나 라는 생각을 했고 앞으로의 프로젝트에서는 데이터의 특성을 좀더 고려해야겠다는 다짐을 하게 되었다.

6. **한계와 아쉬움**

Tabnet 모델이 정형 데이터를 잘 다룰 수 있다는 점에서 사용해 보았지만 학습 시간이 너무 길어 파라미터 조정에 많은 시도를 하지 못한 점이 아쉬웠다. 특히 Tabnet은 데이터 특성과 중요성을 학습해 feature selection을 따로 진행하지 않아도 된다는 장점이 있어 선택하게 되었는데, 학습 속도가 느리다는 점을 고려하지 못했던 것 같다. 이로 인해 여러 파라미터 조합을 실험할 기회를 갖지 못해 충분한 성과를 내지 못해 아쉬움이 남랐고 향후 프로젝트에서는 Tabnet의 잠재력을 최대한 활용할 수 있도록 노력해야겠다.

7. **앞으로의 도전**

   모델링 역할과 데이터 전처리 역할을 모두 담당해보았으니 앞으로의 프로젝트에서는 지금까지의 경험과 피드백을 활용하여 좀더 효율적이고 더나은 결과를 도출하기 위해 노력해야겠다.

김희수_T7513

1. **나는 내 학습목표를 달성하기 위해 무엇을 어떻게 했는가?**

   저는 이번 프로젝트에서 데이터팀에 들어가 모델팀에서 요구사항이 적힌 발주서를 처리하면 데이터 전처리에 필요한 함수를 만들거나 feature engineering을 하였다. 또한, 제가 생각한 모델을 모델팀에서 사용해주길 원해서 제가 생각한 모델의 구현안을 알려주었다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

   저번 프로젝트에서는 EDA에서 추측하여 feature를 생성하였지만, 이번 프로젝트에서는 제가 생각한 EDA를 통계적 방법을 통하여 가설 검증을 하고 새로운 feature를 생성하려 하였다. 하지만 데이터 불균형으로 인해 모델에 feature를 넣어도 좋은 성능을 나타내지 못했다.

3. **마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?**

   모델링을 하는 중 팀원이 랜덤포레스트를 사용하였는데 성능이 좋지 않았다. 여기서 모델의 성능을 올리기 위해 팀원들과 상의를 해보았는데 부트스트랩 방법을 사용하면 성능이 올라갈거라고 결론이 나왔다. 하지만 나는 부트스트랩을 사용해도 랜덤포레스트에선 성능이 좋아지지 않을거라 생각하였지만 제 생각과는 반대로 성능이 올라갔다. 이런 결과로 인해 모델이 어떤 원리를 통해 작동하는지를 알아야 하는것을 깨달았다.

4. **한계/교훈을 바탕으로 다음 프로젝트에서 시도해보고 싶은 점은 무엇인가?**

   이번 프로젝트에서의 평가지표는 MAE라 모델 성능을 확인할 때 MAE만 사용하였다. 이렇게 MAE만 사용하면 모델에 오차가 큰 값이 있을 경우 파악하기 어렵다. 따라서 다음 프로젝트에서는 여러가지 지표를 사용해 볼 것이며, 모델의 작동 방식에 대해서도 자세하게 알 필요가 있다고 생각한다.

노근서_T7514

1. **나의 학습 목표는 무엇이었나? 달성하기 위해 무엇을 어떻게 했는가?**

   팀 구성시 데이터팀 역할을 맡고 페어 프로그래밍을 하기로 한 시점부터 모델팀을 서포트 해준다는 생각으로 프로젝트에 임하려고 했다. 마침 EDA가 끝난 후 모델팀으로부터 흩어져 있는 csv 파일을 하나의 데이터프레임으로 통합해달라는 요청을 받고, 모델팀 → 데이터팀으로 Issues와 노션을 활용해 발주서를 남기는 아이디어를 생각해냈다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

   지난 프로젝트와 달리 모델팀이 아닌 데이터팀으로 참여했다. 위 목표를 실천하기 위해 모델팀에서 발주서를 남기면 담당 데이터팀 파트너가 필요한 전처리 함수 또는 파생변수를 생성해 데이터를 불러오기만 하면 바로 사용할 수 있게끔 작업을 했다. 위 과정을 Issues, PR, 노션에 함께 기록했더니 어떤 작업을 하는지, 만든 이유는 무엇인지 추적하기 쉽고 팀끼리 분업화가 잘 이뤄지도록 할 수 있었다.

2. **마주한 한계는 무엇이며 아쉬웠던 점은 무엇인가?**

   프로젝트 마지막 주 월요일이 돼서야 TabNet 모델을 사용했고 MAE 5000 수준까지는 낮췄지만, 결국 앙상블에 사용할 수 있을 만큼 원하는 결과를 내지는 못했다. 딥러닝 모델 특성상 긴 학습 시간이 발목을 잡았던 점, 모델에 대한 지식이 부족해 성능을 보다 이끌어 내지 못한 점이 못내 아쉬웠다.

3. **한계/교훈을 바탕으로 다음 프로젝트에서 시도해보고 싶은 점은 무엇인가?**

   모델을 보다 잘 활용하려면 무작정 사용하는 것보다 이해를 하고 적용하는 게 중요하다는 걸 지난 프로젝트에 이어 다시금 느낄 수 있었다. 한정된 시간 내에 딥러닝 모델을 효과적으로 사용하기 위해선 프로젝트 초반부터 도입해야 실험에 필요한

긴 시간을 커버할 수 있다는 교훈도 얻을 수 있었다. 다음 프로젝트에서 딥러닝 모델을 사용할 경우, 두 교훈을 꼭 되새기고 적용할 수 있도록 노력해야겠다.

4. **나는 어떤 방식으로 모델을 개선했는가?**

   건물과 공공장소 사이의 거리를 계산할 때, Scikit-learn에서 제공하는 Ball Tree와 K-D Tree를 활용해 계산 속도를 줄이는 데 성공했다. 처음 강의 자료와 같이 Brute-Force 방식으로 코드를 짰을 때는 몇시간을 돌려도 작업이 끝나지 않았지만, 위 알고리즘을 적용한 결과 2분 내외로 연산을 끝낼 수 있었다.

5. **내가 해본 시도 중 어떤 실패를 경험했는가? 실패 과정에서 어떤 교훈을 얻었는가?**

   기존에는 Ball Tree를 활용해 거리를 계산했는데 그 이유는 haversine 거리를 metric으로 지정할 수 있었기 때문이다. 반면 K-D Tree는 Ball Tree보다 초기 트리 구성 비용이 낮고, 저차원 데이터에서 빠른 연산 속도를 갖지만 haversine 거리를 metric으로 지원하지 않았다. 하지만 데이터가 2차원(위・경도)이고 유클리드 거리를 사용할 때 연산 속도는 K-D Tree가 빨랐던 점을 이용하고 싶었다. 그래서 haversine 거리로의 변환만 가능하면 보다 빠른 연산이 가능할 거라 생각하고 최적화를 시도해봤지만, 결국 위・경도의 중복을 제거하면 기존에 사용한 Ball Tree의 연산이 K-D Tree보다 빨라 새로 시도한 방법을 적용하지는 못했다. 그렇지만 처음 알게 된 두 알고리즘의 동작 원리도 공부하면서 새로운 개념을 습득할 수 있어서 좋았고, 2분 남짓한 연산 속도를 1초만에 끝내도록 성능 개선도 성공한 점이 가장 만족스럽다.

박영균_T7520

1. **나의 학습목표는 무엇이었나?**
   경제학도로서 매우 흥미로운 주제를 받았다고 생각했다. 마침 학사 졸업논문으로 생각 중이었던 주제와 일치해서 더욱 마음이 갔다. 이번 프로젝트 뿐만아니라 이후에도 쓸 수 있는 인사이트를 얻는 것이 나의 목표였다.

2. **전과 비교해서, 내가 새롭게 시도한 변화는 무엇이고, 어떤 효과가 있었는가?**

   도메인 지식의 측면에서 접근해보려 많이 노력했다. 금리와 부동산은 밀접한 연관을 가지고 있다는 것을 알고 있었기에, 이 관계를 어떻게 주어진 데이터에 적용할 수 있을지 많은 고민을 했다. 결국 주어진 데이터 정보가 부족해 해당 프로젝트에서 유의미한 결과를 얻지는 못했지만, 차후 외부 데이터 등을 활용해 더 질 좋은 예측을 위한 방법에 대해 생각해볼 수 있었다.

3. **개인적으로 어떤 것을 배웠는가?**

   부동산에 대한 도메인 지식을 얻는데 도움 많이 되었다. 내가 알고 있던 경제학적 지식만으로는 부동산 가격의 상승 및 하락 요인을 설명할 수 없다는 점에서 새로웠다. 금리와의 상관관계 및 정치와의 연관성 등이 유기적으로 연결되어 있다는 사실을 알았고, 앞으로의 공부에서도 적용해볼 만한 좋은 지식들을 가질 수 있었다.

4. **협업 측면에서 얻은 것이 있는가?**

   이전 프로젝트에서는 팀원 간의 의사소통이 잘 되지 않았다고 생각한다. 서로 자기의 일만 하다보니 서로의 진행상황을 공유받지 못했고, 이를 하나로 통합하는 과정에서도 소통의 부재가 문제를 일으켰다. 이번 프로젝트에서는 내가 맡은 일이 아니라도 여유가 있을 때 들어보려 노력했다. 그러고 나니 나도 새로운 아이디어를 많이 얻었고, 나와 소통하는 다른 팀원들에게도 나의 다를 시각을 이야기해주며 더 나은 작업을 할 수 있었다.

5. **프로젝트 중 기억에 남는 것이 있는가?**

   다른 팀의 프로젝트 방향성에서 생각지 못했던 방법을 들었다. 머신러닝 기법을 쓰지 않고, 일반적인 알고리즘 기법을 활용해 예측한 결과가 예상 외로 성능이 높았다는 것이 흥미로웠다. 나도 제법 기발한 방법이라고 하며 우리 프로젝트를 수행했지만, 이렇게도 생각할 수 있구나 하는 생각이 들었다. 다음부터는 생각을 좀 더 열어봐야겠다는 다짐을 했다.

6. **프로젝트에서 아쉬웠던 점은?**

   딥러닝과 관련해 여러 시도를 해보지 못한 것이 못내 아쉽다. 시간 관리 측면에서 저번 프로젝트와 비교해서는 여유롭게 진행했지만, 그 여유로움에 오히려 풀어져 많은 작업을 스킵한 면도 있었다. 앞으로의 프로젝트에서는 좀 더 굳은 마음가짐을 가져야 할 듯하다.

7. **다음 프로젝트에서 시도해볼 것은?**

   무엇이 되었든 내가 해보지 않았던 것을 섞어볼 것이다. 아무래도 모델일 확률이 높아 보인다. 성능이 잘 나온다는 이유로 여지껏 트리 모델을 주로 써 왔지만, 이 틀에 갇혀 오히려 다른 많은 모델을 써볼 기회를 놓치고 있다는 생각이 들었다. 아무튼, 모든 답변에서 그렇지만 다양한 시각에서 문제를 살펴보는 자세가 필요하다고 본다.