김성연 개인회고

- **나는 내 학습목표를 달성하기 위해 무엇을 어떻게 했는가?**

우리 팀원들과 대회 형 프로젝트를 처음 하는 만큼 지금까지 배운 이론을 실전 모델에 적용하고 5명의 팀으로써 대회를 치뤄보자가 큰 목표였던 것 같아요.

머신러닝을 활용한 대회는 경험이 많지만 딥러닝은 저에게 큰 도전이였습니다. 딥러닝과 친해지고 엔지니어라는 말을 쓸 수 있을정도로 상세하게 쓰여진 딥러닝 모델을 여러 관점에서 많이 만져보았습니다.

이전에 참가했던 대회들도 팀으로 참가하긴 했지만 깃허브 등을 사용해서 협업했다는 느낌보다는 주기적으로 아이디어를 공유하고 각자 발전시킨다는 느낌이 강했습니다. 조금 어색하더라도 CLI 환경, 깃허브를 통한 협업, .py 파일 디렉토리 형식을 최대한 적응하려고 노력했습니다.

완벽하게 이뤄내진 못했지만 대회기간 전과 비교했을 때 발걸음은 뗐다고 생각합니다. 앞으로도 험난하겠지만 지금 경험이 큰 도움이 될 것 같아요.

- **나는 어떤 방식으로 모델을 개선했는가?**

타겟 데이터 분포를 먼저 관찰했던 것 같습니다. 조금 편향되어있구나 정보를 얻었고 트레인 데이터와 테스트 데이터의 분포를 비교했습니다. 이 대회에서 예측해야하는 테스트 데이터는 단순히 랜덤분할한거구나 정보를 얻었습니다. 다음으로 추천모델을 선정하는데 가장 중요한 점 중 하나인 얼마나 희소행렬인지 여부를 살펴봤습니다. 생각보다 유저수와 아이템 수 대비 데이터가 적었고, MF 모델이나 파라미터가 많은 딥러닝을 적용하면 오버피팅이 날 수 있다는 직관을 얻었습니다.

이전 대회에서 얻은 경험으로 이런 경우 부스팅 모델이 괜찮을 것 같다는 직관을 얻었고 카테고리 변수가 많았기 때문에 이에 유리한 CatBoost 모델을 선정하게 되었고 성능도 괜찮게 나왔던 것 같습니다. CatBoost 모델을 빠르게 찾아 시간을 벌어서 다양한 시도를 할 여유가 생겼는데요. CatBoost가 해결하지 못하는 유저와 아이템 관계에 집중한 딥러닝 모델과 앙상블 한다면 큰 효과를 볼 수 있다는 아이디어를 얻고 딥러닝 모델에 도전했습니다. 실제 이 부분이 리더보드에 큰 효과가 있었습니다.

하이퍼파라미터 최적화 등 모델 고도화에 시간을 넉넉히 잡고 여유롭게 임했던 것도 좋은 결과로 이끌었던 것 같습니다. K-Fold와 WandB Sweep, Optuna 툴을 이용해 모델 고도화를 성공적으로 진행했습니다.

- **내가 한 행동의 결과로 어떤 지점을 달성하고, 어떠한 깨달음을 얻었는가?**

대회 초기 성능이 좋지 않아 낙담했을 때 실패를 두려워 하지 않고 다양한 관점에서 시도했던 것이 좋았던 것 같습니다. 이전에도 얼핏 알고 있었지만 10번 시도하면 1번 성공하긴 하지만 10번 시도를 해야지만 성공을 할 수 있다는 것 입니다. FM, FFM, CNN_FM, DeepCoNN 등 다양한 딥러닝 모델 사용과 여러 관점에서 생각해보기, 부스팅 기반 머신러닝 모델 사용해보기를 통해 많은 실패를 겪었으나 결국 성공했습니다. 그리고 이러한 경험이 10번 시도 시 성공할 횟수를 늘려줄 것 입니다.

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

적극적인 깃허브 사용과 VsCode 터미널 환경 적응 그리고 코드를 깔끔히 정리해 팀원과 소통하며 집단 지성 얻기 입니다.

Push와 Pull 만 사용했었고 얼마전에 이고잉님 특강에서 이론으로만 배웠던 깃허브를 고급 기능까진 모르겠지만 필수로 다뤄야 하는 기능은 사용해보면서 많이 배웠던 것 같습니다.

ssh라는 용어도 어색하고 CLI 환경 이름도 처음 들어봤던 통계학과 출신 저에게 VsCode를 활용한 서버 사용은 큰 산이였습니다. 하지만 앞으로 계속 써야하는 거 지금 제대로 적응해보자는 생각으로 도전해봤고 많은 오류와 싸워왔지만 부스트캠프 내 동료들 덕분에 잘 이겨냈던 것 같습니다. 잘한다곤 하지 못하겠지만 한번 경험했으니 무지에서 오는 두려움은 사라진 것 같습니다.

매일 진행되는 피어세션 시간을 이용해 팀원들과 제 생각을 적극적으로 공유하자는 목표를 달성하기 위해 열심히 노력했던 것 같습니다. 제 아이디어를 기반으로 팀원들이 저 혼자 생각했다면 하지 못할 여러가지 관점을 생각해 실제 모델 성능 향상에 도움이 많이 된 것 같습니다.

- **마주한 한계는 무엇이며, 아쉬웠던 점은 무엇인가?**

Valid Score가 높게 나온 모델이 있는데 왜 Test Score는 다르게 나오는지 시간을 많이 들여서 여러가지 각도로 원인 분석을 했는데 이해를 못한 부분이 아쉽습니다.

팀원들과 깃허브를 사용했는데요. 나쁜 점은 없었지만 뚜렷히 어떤 효과를 보았다고 할 만큼 잘 사용하진 못했습니다. Master 브랜치를 땡겨온 뒤 개인 브랜치에 merge 하는 작업을 잘 하지 않아 서로의 작업이 따로 놀게 되었고 이를 인지한 순간 너무 충돌이 많이 나 시도조차 못했던 것 같습니다.

스스로가 딥러닝 모델을 제대로 고도화하지 못했던 점이 많이 아쉽습니다. 제가 생각하는 직관이 맞는지 확인을 해야하는데 시간 관계상 진행하지 못했던 것 같습니다.

- **한계/교훈을 바탕으로 다음 프로젝트에서 스스로 새롭게 시도해볼 것은 무엇일까?**

다음 프로젝트에선 하루에 마무리를 팀원들과 함께 깃허브 브랜치 정리로 마무리 하려합니다. 능숙해진다면 이렇게 하지 않겠지만 팀원들과 함께 깃허브 버전 관리를 습관화를 하려고 노력할 것 같습니다. 추가로 코드 리뷰를 더 적극적으로 진행할 것 같습니다. 내가 가진 생각과 팀이 가진 생각 간 버전 관리를 잘할때 팀으로써에 효율이 극대화 될 것 같습니다.

딥러닝 모델을 Low 레벨부터 도전해보려합니다. 만약 이번 대회와 같이 베이스라인이 잘 주어진다면 베이스라인 코드 리뷰부터 꼼꼼하게 진행할 것 같습니다. 이러한 경험이 쌓인다면 딥러닝 모델에 대한 직관도 높아질 것 같고 고도화에도 뛰어난 기초체력이 될 것 같습니다.