

경력 포트폴리오

CONTENTS

01

번개장터

Data Team | Data Scientist Intern
(2021.09.27 ~ 2021.12.31)

02

카카오

Recommendation Team | Data Analyst Intern
(2022.01.03 ~ 2022.02.25)

03

쏘카

Value Discovery Team | Data Scientist
(2022.05.16 ~ 2022.09.08)

04

넥슨

Matching Team | Data Analyst
(2022.09.19 ~)

01. 번개장터

검색 광고 상품 배치 변경

목표

- 기존 검색 광고 상품의 배치를 변경하여 광고 효과 및 광고 수익을 증대하기 위함이다.

과정

- postgreSQL을 통해 두 달간의 검색 상품 위치별 클릭 수, 노출 수, 광고비 데이터를 추출하였다.
- 기존 검색 상품 위치의 클릭 수, 노출 수를 통해 번개장터 검색 광고/비광고 상품의 피드백 데이터 특징을 파악하고 검색 광고 상품이 다른 위치에 있을 때의 클릭 수, 노출 수를 추정하였다.
- 검색 상품은 상단에 위치할수록 클릭 수와 광고 수익이 높으나 광고 상품의 거부감으로 CTR은 낮아 추정된 클릭 수, 노출 수를 이용하여 최대한 광고 수익은 증가시키고 CTR은 감소시키지 않는 검색 광고 상품 배치안을 제시하였다.
- 총 3개의 배치안을 제시하였으며 각각의 배치안에 기존 유저 5%를 할당하여 AB Test를 진행하였다.

결과

- AB Test 결과, 최적의 배치안이 광고 상품의 CTR 하락 없이 광고 수익을 약 100% 증가시켰다.

회고

- 이번 프로젝트는 콘텐츠 기획팀과 지속적인 소통을 하며 진행되었으며 시니어분들의 회의 방식을 보며 소통하는 법을 배웠다.
인턴부터 시니어까지 모두가 평등한 입장에서 대화를 진행하였으며 남의 의견을 먼저 듣고 답하며 자연스럽게 자신의 의견으로 이동하는 대화 흐름이 이상적이었다.
의견의 충돌로 인해 언쟁이 될 수 있는 상황에서 이런 소통법은 건전하고 이상적인 회의를 가능케 하였으며 이후 항상 이 소통법을 신경 쓰며 회의에 임하는 계기가 되었다.

01. 번개장터

검색어 오타 생성 모델링

목표

- 검색어 오타 교정 모델 학습에 사용될 (검색어, 오타)쌍 데이터를 생성하기 위함이다.
- 검색어 오타 교정 모델이 기준 이하일 경우, (검색어, 오타)쌍 데이터를 통해 Rule based로 검색어를 교정하기 위함이다.

과정

- 초기에는 짧은 추론 시간을 위해 가벼운 RNN 모델인 LSTM과 자모 단위로 Tokenizer를 구성하여 학습하였으나 모두 같은 값을 예측하는 Local Minimum 문제가 발생하였다.
- 자모 단위 Tokenizer에서는 같은 입력으로 생성되는 다음 Sequence 패턴이 매우 다양한데, 가벼운 RNN 모델이 이를 학습하지 못하여 Local Minimum 문제가 발생했다고 판단했다.
따라서 Tokenizer를 음절 단위로 확대하고 LSTM보다 더 깊은 Transformer Encoder, Decoder를 학습하여 문제를 해결했다.
- 오타는 기존 검색어를 입력할 때, 키보드에서 가까운 자판을 누르거나 자판을 누르지 않음으로써 발생한다. 따라서 해당 특성을 반영한 키보드 자판 거리에 따른 Edit Distance를 정의하여 Edit Distance가 기준 이상인 생성 오타는 후처리로 제거하였다.
- Transformer 모델을 사용하여 커진 모델 사이즈와 느려진 추론 속도를 해결하기 위해 Quantization을 진행하여 모델 사이즈는 약 1/2배로 추론 속도는 1/20배로 축소하였다.

결과

- 기존에 존재하는 데이터로만 학습한 검색어 교정 모델은 영문과 숫자가 섞인 오타에서 낮은 성능을 보였는데, 오타 생성 모델로 생성된 오타를 학습하여 전체적인 성능 향상을 이루었다.

회고

- RNN이나 Transformer를 통한 모델링을 진행할 때는 Local Minimum 문제에 빠지는 상황이 많았다. Sequence 데이터는 같은 입력에도 다양한 결과 패턴이 있기 때문에 이를 잘 학습하려면 적합한 깊이의 모델과 Hyper parameter를 실험을 통해 찾아야 한다.
- 프로젝트에서 부족한 점을 모델링을 통해서 해결하는 것이 더 효율적일 때가 있다.
이번 프로젝트도 부족한 데이터를 모델을 통해 보충하였고 결국 성능 향상을 이루었다.
이러한 경험을 살려 넓은 시각으로 실험 계획을 할 것이다.

01. 번개장터

콜드 아이템 추천을 위한 GNN Content based 추천 모델링

목표

- 피드백 데이터가 없는 콜드 아이템을 추천하기 위해 기존 인기 상품과 비슷한 상품을 추천하는 Content based 추천 모델링 진행하였다.
- Content based 추천에 GNN 도입 가능성을 보기 위해 GNN과 ML 모델 성능 비교하였다.

과정

- postgresQL을 통해 **상품을 Node**로 하고 특정 유저가 **같이 클릭한 상품들을 Edge**로 이은 **Homogeneous graph** 데이터셋을 구성하였다.
- 상품(Node)의 Feature로는 상품 가격, 찜 수 등이 있으며 **상품 이름은 Fasttext를 통한 임베딩, 상품 카테고리는 Graph의 Random Walk를 통해 추출된 Sequence를 Word2Vec으로 학습하여 나타낸 임베딩**으로 하는 Feature Engineering을 진행하였다.
- Pytorch 기반의 GNN 라이브러리인 **DGL(Deep Graph Library)**을 사용하여 GCN, GraphSAGE 등 **9개의 GNN 모델을 실험**하였고 LGBM과 성능을 비교하였다.

결과

- 성능 비교 결과 LGBM의 성능이 우수했으나 GNN에서는 Homogeneous Graph보다 Node 타입이 여러 개인 Heterogeneous Graph에서 더 높은 성능을 보인다는 레퍼런스를 많기 때문에 개선 가능성은 충분했다.

회고

- 초기에 가벼운 실험부터 시작하기 위해 Homogeneous Graph를 데이터셋으로 하였지만, 제한된 기간으로 Heterogenous Graph를 통한 GNN을 실험하지 못한 것이 아쉬웠다.
- Homogeneous Graph는 노드 타입이 하나이기 때문에 그래프의 모든 정보를 포함하기엔 무리가 있어 ML 모델보다 성능이 좋지 않았다.
- GNN끼리의 성능 비교에서도 더 고도화된 모델이거나 논문을 읽고 해당 프로젝트에 더 적합하다고 생각된 모델이 무조건 좋은 성능을 보인 것이 아니기 때문에 **태스크에 맞는 모델링이 필요**했다.

02. 카카오

컨텐츠 품질의 정량적 지표 설정 및 고품질 컨텐츠 추출

목표

- 카카오 뷰의 컨텐츠인 **보드의 품질을 정량적 지표**로 설정하고 이를 통해 고품질 보드를 추출하여 **좋은 보드를 유저에게 많이 노출**하기 위함이다.
- 보드는 검수 과정을 거쳐 유저에게 노출되는데, **품질 지표 예측 모델을 통해 효율적인 검수에 도움**이 될 수 있다.

과정

- SQL과 파이썬을 통해 세 달간 데이터를 추출하고 분석에 맞게 처리하는 파이프라인을 구축했다.
- **보드 품질의 정량적 지표 설정**
검수 관련 컬럼의 요소별 **비중이 작을수록** 해당 요소의 특징을 가진 보드가 많아 **품질을 보장**한다는 특징을 발견했다. 따라서 검수 관련 컬럼의 요소별 비중에 따라 품질의 가중치를 달리 설정하였다. **비중이 작을수록 큰 가중치를, 비중이 클수록 작은 가중치를 부여**하기 위해 Poisson Distribution을 사용했다.
피드백의 경우에는 좋은 보드뿐만 아니라 선정적이고 자극적인 보드에도 높은 지표를 갖기 때문에 **검수 관련 컬럼이 긍정 지표인 보드의 한에서 피드백에 따라 품질 지표에 가산점을 부여**했다.
- **품질 지표 예측을 위한 Feature Engineering**
범주의 개수가 적은 변수는 One-Hot Encoding을 통해 Feature로 나타내었으며 다른 변수들은 지표와의 상관관계가 유의한 변수만을 추출하여 사용하였다.
텍스트 데이터의 경우에는 Fasttext, Bi-LSTM, KoBERT 순서로 임베딩 모델을 고도화하였으며 이 중 Loss가 가장 낮았던 **KoBERT를 사용하여 임베딩을 추출**했다.
- **Prediction**
추출한 Feature를 이용하여 LGBM이 품질 지표를 예측하도록 학습하였다. 검수 관련 컬럼이 품질뿐만 아니라 다른 기준으로 검수 되어 발생하는 “**검수 관련 컬럼과 실제 품질의 Bias 문제**”와 예측이 채널 Feature에 Robust하지 않아 **같은 보드라도 채널 정보에 따라 다른 예측을 한다는 문제가 존재**했다. (채널 : 보드 게시자)
이를 해결하기 위해 채널 **Feature를 제외**하고 학습한 LGBM을 사용하여 예측을 진행하였고 좋은 보드를 샘플링하는 **기준을 설정**하기 위해 기존 LGBM을 사용하였다.

결과

- 채널 정보에 Robust한 좋은 보드 샘플링 기준을 설정하였고 약 60,000개의 Validation 데이터셋 중 약 9,000개가 좋은 보드로 추출되었다.
해당 보드들의 실제 **검수 관련 컬럼은 대부분 긍정 지표**를 나타내었다.
- **검수 관련 컬럼과 실제 품질의 Bias 문제를 해당 모델로 보완**할 수 있었다.

회고

- 품질을 나타내는 지표가 없어 EDA를 통하여 품질을 나타낼 수 있는 여러 변수를 조합하여 품질 지표를 생성하였지만, 이는 실제 품질과의 Bias가 있어 아쉬웠다. 하지만 이를 모델링을 통해 보완할 수 있어 시간이 더 주어진다면 이를 해결하고 예측 모델링을 진행했을 것이다.
- 예측 모델 자체의 Loss가 낮지 않기 때문에 설득력이 부족하여 아쉬웠지만, 실제 예측한 품질 지표가 **매우 높은 것과 낮은 것은 모두 품질이 매우 좋지 않거나 매우 좋은 보드를 나타내어 실무 적용점은 다양**했다.
- 모델링이나 분석 같은 하드 스킬 뿐만 아니라 **논리적 근거를 들어 분석을 진행하는 소프트 스킬** 또한 많이 늘었다. 분석을 하다보면 논리에 따라 분석을 진행하는 것이 아닌 긍정적인 결과만 보며 분석에 논리를 맞추는 경우가 많은데, 이번 인턴에서 여러 피드백과 멘토들의 사고 방식을 통해 **논리적 근거를 잃지 않고 분석하는 법**을 배울 수 있었다.

03. 쏘카

차량 배터리 방전 정의 및 예측 모델링을 통한 방전 차량 선조치

목표

- 차량에서 발생하는 장애 중 배터리 방전은 빈도가 높다. 배터리 방전 장애는 차량 운행을 불가능하게 하는 치명적인 장애이며 **예약 직전에 발생한다면 큰 비용적 손해가 발생한다.**
- **배터리 방전 확률이 높은 차량을 예측하고 선조치하여 비용적 손해를 예방**하고 고객에게 쾌적한 경험을 제공하고자 한다.

과정

- 차량 배터리 방전에 따른 로그가 적재되지 않기 때문에 [신호 끊김], [낮은 전압], [전압 하락 폭], [시동 여부] 등의 정보를 조합하여 **방전 패턴을 찾아내 배터리 방전을 정의**하였다.
- [독립 변수 로그의 적재 주기], [선조치에 필요한 시간], [운행 중 배터리 방전 조치 여부]에 대해 차량 운영팀과 논의하였고 이를 반영하여 [독립 변수와 실제 방전의 시간 간격], [예측 배치 간격], [예측할 차량]과 같이 **선조치를 위한 룰을 설정**하였다.
- Class Imbalance 문제를 해결하기 위해 **Under Sampling**을 하였고 이로 인한 Negative Dataset(방전이 아닌 차량 데이터)의 정보 손실을 최소화하기 위해 다른 Negative Dataset으로 학습된 모델을 **Ensemble**하였다.
- 학습한 모델인 LightGBM의 Feature Importance와 예측 결과를 시각화하여 모델이 어떤 독립 변수를 중요하게 보고 어떤 전압 패턴을 방전으로 예측하는지에 대한 설명 자료를 작성하고 **기존 방법의 성능과 비교**하였다.
- 배치마다 Airflow를 통해 **배터리 방전 확률이 높은 차량 정보와 기존 배터리 방전을 판단하기 위해 확인하는 독립 변수를 DB에 적재**하여 차량 운영팀이 방전 차량을 확인하기 용이하게 하였다.

결과

- 배터리 전압이 낮아진 후에 배터리 방전을 조치하는 기존 방식이 아닌 전압이 낮아져 배터리가 방전될 것을 예측하여 선조치할 수 있는 방식을 제시하였다.
- 한 달 동안 배터리 방전 예측 모델링을 통해 차량을 선조치할 경우, 기존 방식 대비 약 20%의 비용을 절약할 수 있었다.
- 실제 배터리 방전이 일어나지 않았음에도 선조치한 차량이 있어 조치를 취한 차량의 수가 기존보다 더 많아졌다. 따라서 기간이 늘어남에 따라 잠재적 비용도 예방할 수 있다.

회고

- 배터리 방전에 대한 지표가 없었기 때문에 세 차례에 걸쳐 지표를 정의하고 성능을 실험했다. 이 과정을 통해 모델링 이전에 문제에 맞는 지표 정의의 중요성을 알게 되었다.
- 다른 팀(차량 운영팀)의 업무 처리 방식을 바꾸기 위해서는 비용 절약의 측면 뿐만 아니라 기존 방법에서 어떤 점이 개선되었는지 논리적으로 설득할 필요가 있었으며 이를 통해 자신의 생각을 데이터 시각화와 논리적 근거에 기반하여 글로 나타내는 능력을 키울 수 있었다.