Portfolio

JINHA

Data Scientist

현대카드
오진하

Contents

Ⅰ. 지원자 소개

지원자 프로필

지원자는 연세대학교를 졸업 후 현대카드에서 Data Scientist 로 근무하였습니다.
또한 Python 기반으로 분석을 수행하고 있으며 Spark, AWS 등의 Framework 를 사용 가능합니다.

기본 정보

- 생년월일 1995.01.09.
- 연락처 010.4149.5704
- 거주지 현재 서울시 관악구 신림동
- 기술스택 Python / Spark / AWS(Sagemaker)
- 전문분야 Machine Learning & Deep Learning
- 희망직무 ML&DL 개발
- 자격증 데이터분석준전문가(ADsP)

- 이메일 o41495704@gmail.com
- 블로그 lashid.github.io

약력

------- ○	2013.03 ~ 2020.02.	연세대학교 졸업
------- ○	2019.12 ~ 2020.03.	경영컨설팅펌 라이언앤코 인턴
------- ○	2020.10. ~ 현재	현대카드 Data Scientist

수행 프로젝트 목록

현대카드 재직 중 수행한 프로젝트 목록입니다.
주요 프로젝트는 M포인트몰 개인화 추천, 마케팅 오퍼 금액 최적화, 공통 고객 없는 고객사 마케팅 모델링 입니다.

기본 정보

프로젝트 요약	수행 기간	방법론	수행 역할
• **3층 개인화 추천**	2020.10. ~ 2021.02.	LGBM	추천 시스템 운영 및 일부 고객 분석
• **M포인트몰 개인화 추천**	2021.02. ~ 2022.02.	LGBM	연관 상품 추천 & AB 테스트 구조 설계 (Airflow)
• 마케팅 오퍼 금액 최적화	2022.02. ~ 2022.11.	Meta Learning	고객 결제 시퀀스 기반 임베딩 추출 & 모델링
• **프리미엄 카드 마케팅 (with Google Data)**	2022.11. ~ 2023.01.	Plate2Vec	결제 기록 기반 카드 종류별 임베딩 추출
• **공통 고객 활용한 고객사 마케팅 모델링**	2023.02. ~ 2023.03.	LGBM	서비스 운영 및 버전 업데이트
• 공통 고객이 없는 고객사 마케팅 모델링	2023.03. ~ 현재	LLM	전체 모델링 및 프로젝트 중간 관리자

Ⅱ. 수행 프로젝트

· 1. 2023 공통 고객이 없는 고객사 마케팅 모델링

1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – 개요 (1/2)

두 회사에 동시에 존재하는 고객이 없는 경우 (e.g 해외 고객사) 에도 당사의 데이터를 활용하여 고객사 고객에게 마케팅하는 프로젝트입니다.
관련 태그는 Unsupervised Domain Adaptation, Transfer Learning, Tabular Data 입니다.

Introduction

프로젝트 개요

수행 기간	2023.03. ~ 현재 (10개월)
팀 구성	총 2명
사용 모델	LLM(KoAlpaca)
목표값	고객사 마케팅 반응 여부 (Binary)
수집 데이터	현대카드 & 고객사 고객 데이터

데이터 세부 내용

성별	범주형 / 남자, 여자 (필수)
나이	수치형 (필수)
현대카드 Only	고객사에는 없고 현대카드에만 있는 데이터 (e.g 현대카드 결제 데이터)
고객사 Only	현대카드에는 없고 고객사에만 있는 데이터 (e.g 고객사 특화 데이터)

1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – 개요 (2/2)

두 도메인에 동시에 존재하는 고객이 없으며 Feature Set 이 완전히 다른 상황에서 정보를 Transfer 하기 위해서
언어라는 보편적인 도구를 채택했고 이를 위해 한국어로 Pre Trained 된 LLM(KoAlpaca) 을 사용하였습니다.

Problem Definition

✓ 프로젝트 목표

고객 및 데이터 구조가 다른 고객사에 우리 회사의 지식을 전달해주자.

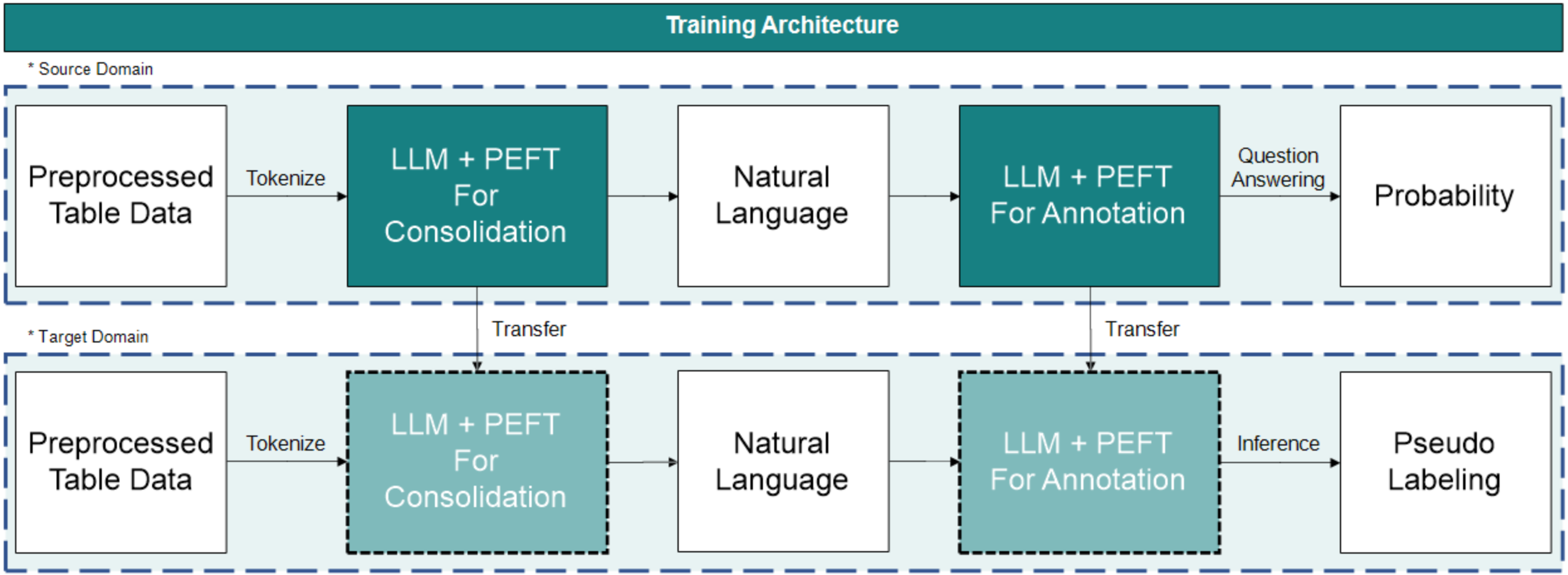
* 그리고 다음과 같은 조건을 가진다.

① 두 도메인에 동시에 존재하는 고객이 없어야 한다.
② 두 도메인의 데이터는 테이블 구조를 가진다.
③ 성별과 나이를 제외한 모든 Feature Set 이 다르다.
④ 데이터 이동은 없어야 한다.

Feature Name & Value 언어가 가진 보편적 의미를 반영해보자.

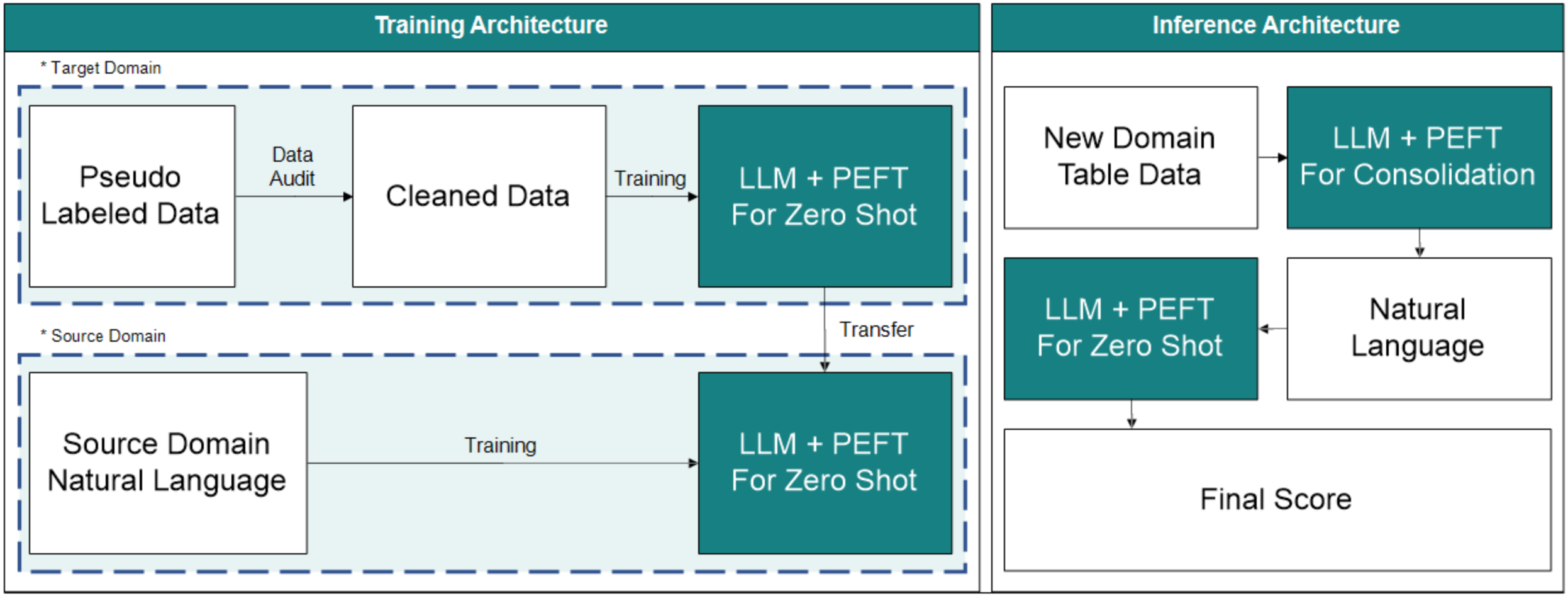
1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – Modeling (1/4)

전처리된 Table Data는 Tokenize 후 Consolidation Model 을 통과하여 자연스러운 문장으로 변환됩니다.
이후 자연스러운 문장을 Input 으로 한 Question Answering 과정을 통해 Source 쪽 Target Label 지식을 학습합니다.
모델들은 Target Domain 에 Transfer 된 후 고객사 데이터를 통해 Pseudo Label 을 생성하게 됩니다.



1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – Modeling (2/4)

Target Domain 에서 생성된 Pseudo Label Data 는 Data Audit 과정을 거쳐 정제됩니다.
이렇게 정제된 Cleaned Data 와 Source Domain 의 원본 Data 를 학습시켜 Zero Shot Inference 가 가능한 최종 Model 이 탄생하게 됩니다.



1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – Modeling (3/4)

데이터와 모델 수행의 간단한 예시입니다.

Appendix. Example

	성별	나이
Data 1	남자	23살
Data 2	여자	34살

Table Data

Tokenize → **성별 남자, 나이 23살**
성별 여자, 나이 34살

실제로는 Token IDS

Consolidation → **23살 남자입니다.**
34살 여자입니다.

실제로는 Token IDS

Annotation → **23살 남자입니다.**
Response: 1
34살 여자입니다.
Response: 0

"Instruction: 다음 고객을 자연스러운 문장으로 표현해주세요"

"Instruction: 다음 고객이 ~ 마케팅에 반응할까요?!"

Data Audit

모델 성능에 영향이 적은 Data 삭제

23살 남자입니다.
Response: 1

Cleaned Data

Learning → **Trained Model**

Transfer →

47살 여자입니다.

New Data

Zero Shot Inference → **Trained Model**

47살 여자입니다.
Response: 0

"Instruction: 다음 고객이 ~ 마케팅에 반응할까요?!"

1. 2023 공통 고객이 없는 고객사 마케팅 모델링 – Modeling (4/4)

최종적으로 기존 Table 이 가진 한계를 문장화를 통해 극복하고 Pre Trained LLM 이 가진 보편성이라는 장점을 더해서
고객사에 우리가 가진 지식을 전달할 수 있었습니다.

Model Serving

✓ 모델 제공 프로세스

Table 을 문장으로 변환 → LLM 활용 Question Answering → 새로운 도메인에 적용

- 기존의 Table Data 의 한계점인 Column 별로 내재화된 의미 구조 대신 문장으로 표현.
- Pre Training Large Language Model 이 가지고 있는 Global Information 활용.

공통 고객이 없어도 현대카드의 데이터 및 모델 지식 전달 가능
(*Baseline 대비 평균 10% 이상 높은 AUROC)

* Baseline: Image 분야의 Unsupervised Domain Adaptation 기법 활용. 두 도메인 데이터의 분포를 일치시킨 후 동일 모델로 예측.

Ⅱ. PROJECT

· 2. 2022 마케팅 오퍼 금액 최적화

2. 2022 마케팅 오퍼 금액 최적화 – 개요 (1/2)

고객별 마케팅 최적 오퍼 금액을 예측하기 위한 프로젝트입니다.
관련 태그는 Prototype, Attention, Meta Learning, CounterFactual 입니다.

Introduction

프로젝트 개요

수행 기간	2022.02. ~ 2022.11. (10개월)
팀 구성	총 3명
사용 모델	Attention & DiceML
목표값	고객별 최적 오퍼 금액
수집 데이터	현대카드 고객 & 가맹점 데이터

데이터 세부 내용

고객 정보	성별, 나이 등
가맹점 임베딩	가맹점별 임베딩 값 (결제 기록 기반 추출)
오퍼 금액	고객이 마케팅 참여 시 얻는 혜택 금액
오퍼 허들	고객이 마케팅 참여하기 위해 사용해야 하는 최소 요건

2. 2022 마케팅 오퍼 금액 최적화 – 개요 (2/2)

고객별로 내재된 특성을 매핑하기 위해 결제 기록을 기반으로 Attention 기법을 활용하였습니다.
또한 고객별로 반응하는 금액의 경계선을 찾기 위해 DiceML 이라는 섭동 이론 기법을 활용하였습니다.

Problem Definition

✓ 프로젝트 목표

고객별로 마케팅에 반응하는 최소 금액을 찾아보자.

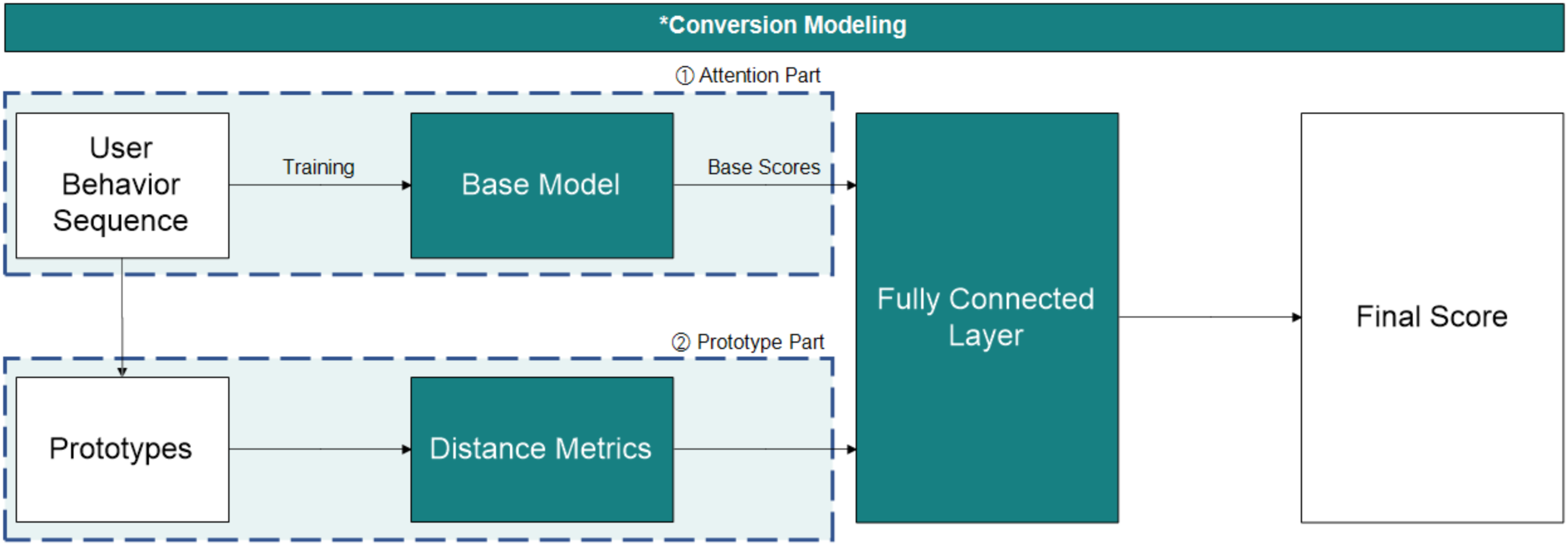
* 세부적으로는 아래 두 가지 경우가 생깁니다.

① 적은 금액에도 반응하는 고객에 대해서는 기준보다 낮은 금액을 제시하자.
② 높은 금액이어야 반응하는 고객에 대해서는 기준보다 높은 금액을 제시하자.

각기 다른 고객 특성을 새로운 공간에 매핑해서 비교해보자

2. 2022 마케팅 오퍼 금액 최적화 – Modeling (1/5)

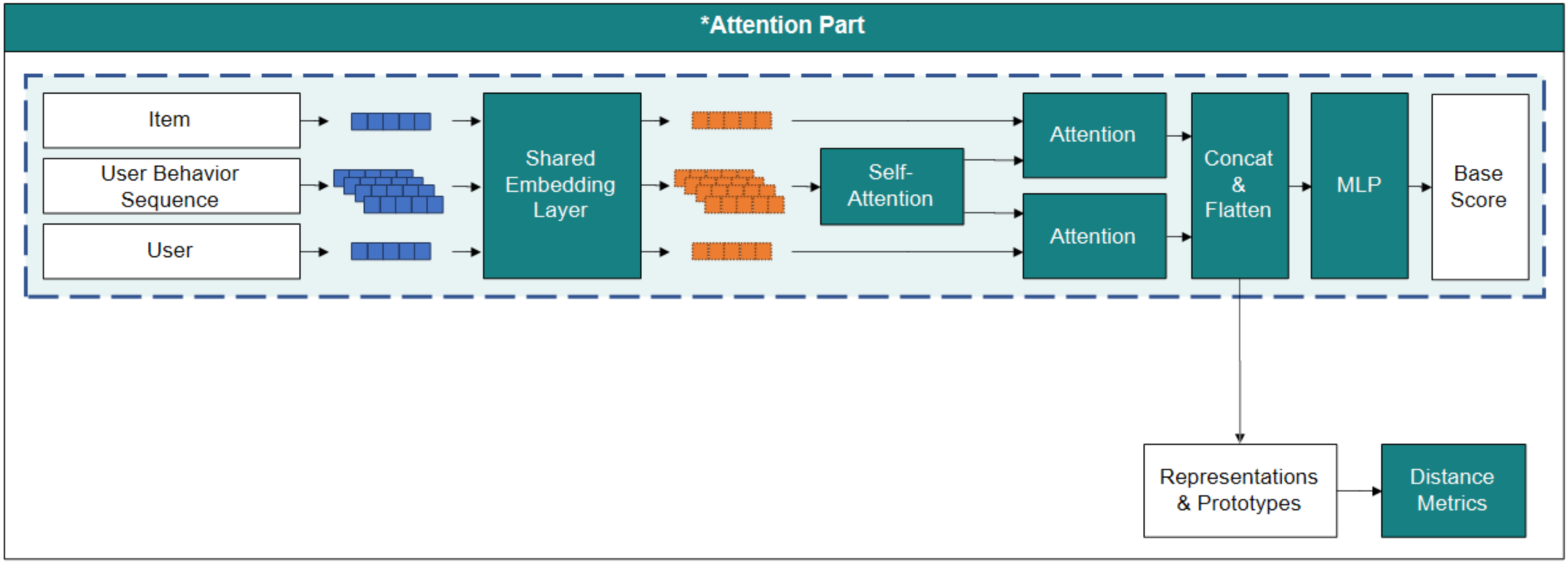
기본 예측 모델은 장기적인 유저의 행동 패턴을 Self-Attention 을 통해 Representation 을 뽑아내는 부분과
각 Scenario 별 Prototype 과의 Distance 를 반영하는 두 부분으로 이루어져 있습니다.



* Source: Xiaofeng Pan(2021). MetaCVR: Conversion Rate Prediction via Meta Learning in Small-Scale Recommendation Scenarios

2. 2022 마케팅 오퍼 금액 최적화 – Modeling (2/5)

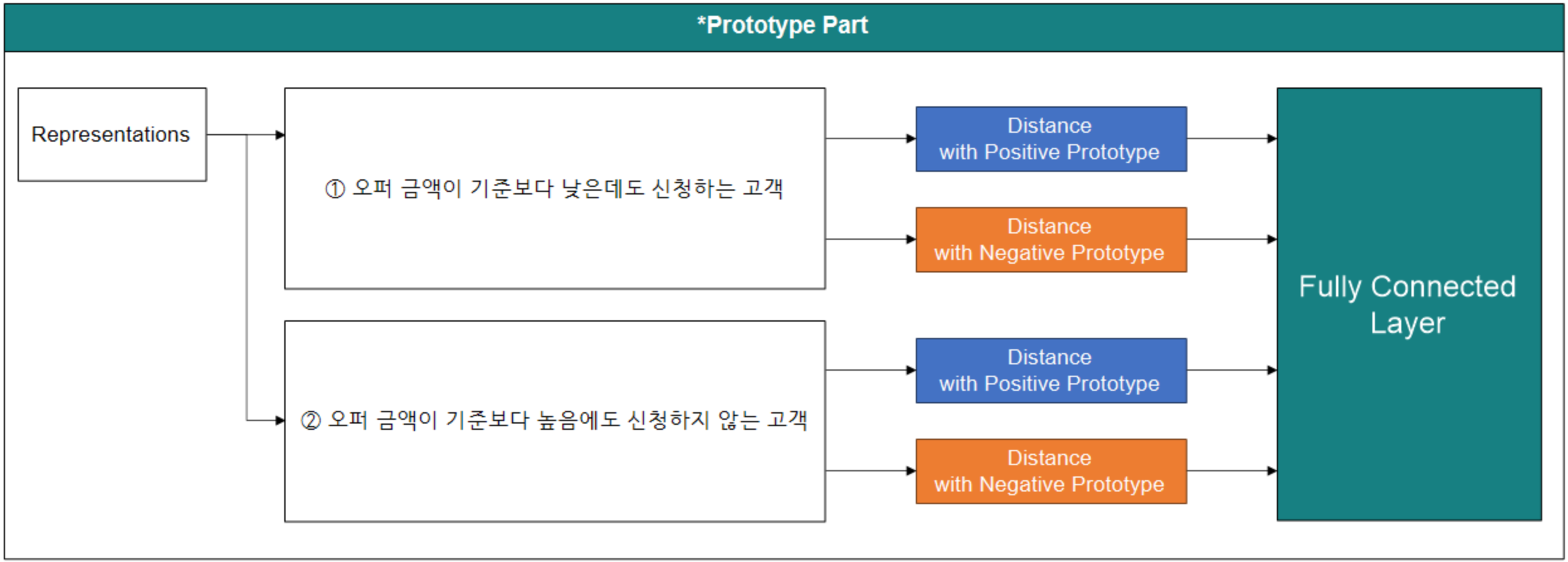
Attention Part 에서는 유저 행동 시퀀스에서 Self-Attention 을 통해 행동간의 관계성을 뽑아내고
이후 고객 정보와 가맹점 정보 Embedding 각각과 Attention 하여 고객 행동과의 관계성을 추출합니다.
중간 과정에서는 고객들의 Representation, Prototype 을 추출하고 / MLP 를 통해 Base Score 를 뽑습니다.



* Source: Xiaofeng Pan(2021). MetaCVR: Conversion Rate Prediction via Meta Learning in Small-Scale Recommendation Scenarios

2. 2022 마케팅 오퍼 금액 최적화 – Modeling (3/5)

Prototype Part 에서는 Attention Part 에서 추출한 Representations 와 각 Scenario 별 Prototype 과의 Distance 를 계산합니다.
그리고 이 결과는 Attention Part 의 결과와 함께 Fully Connected Layer 에 입력되어 최종 Score 를 도출하게 됩니다.



* Source: Xiaofeng Pan(2021). MetaCVR: Conversion Rate Prediction via Meta Learning in Small-Scale Recommendation Scenarios

2. 2022 마케팅 오퍼 금액 최적화 – Modeling (4/5)

Model 학습 후에는 Score 기반으로 고객별로 오퍼 금액 경계선 값을 찾게 됩니다.
반응은 반드시 Positive 하여야 하며, 기존 오퍼 금액에서 너무 차이가 나지 않으며, 반응할 확률이 가장 큰 금액을 선정하였습니다.

	오퍼 허들	오퍼 금액	모델 점수
A	10,000	2,000	0.2
B	10,000	3,000	0.3
C	5,000	1,000	0.5

Q: 그래서 어떤 오퍼 허들, 금액을 선택해야 하는가?

***DiceML**

① Hinge Loss Function
: 특정 Threshold 를 넘기면 같은 값을 부여, 넘기지 못하면 Panelty 부여

$$hinge\_yloss = max(0, 1 - z * logit(f(\boldsymbol{c})))$$

② Proximity
: 기준 오퍼 금액과 멀어질수록 Panelty 부여

$$Proximity := -\frac{1}{k}\sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}).$$

③ Model Score
: Model Score 값이 낮을수록 Panelty 부여

* Source: Ramaravind Kommiya Mothilal(2019). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

2. 2022 마케팅 오퍼 금액 최적화 – Modeling (5/5)

최종적으로 고객 행동 기반으로 추출한 Representation 를 활용하여 고객별 최적 오퍼 금액을 찾을 수 있었고
반응률 증가 및 단가 감소라는 결과를 가져올 수 있었습니다.

Model Serving

✓ 모델 제공 프로세스

고객 행동 기반 Representation 추출 → 각 Scenario 별 Prototype 과의 거리 비교
→ 두 정보를 모두 입력하여 최종 반응률 도출

- 단순 고객 및 가맹점 정보를 유저 행동 기반으로 추출하여 보다 깊은 정보 추출
- Case 수가 적은 Scenario 에 대해서 Distance 기반 Meta Learning 을 활용하여 빠르게 적합 가능

고객별 맞춤 오퍼 금액 최적화
(AB 테스트 시 *Baseline 대비 신청률 1.1배 건 별 단가 0.9 배 달성)

* Baseline: LGBM

Ⅱ. PROJECT

· 3. 2021 M포인트몰 개인화 추천

3. 2021 M포인트몰 개인화 추천 – 개요 (1/2)

M포인트몰 개인화 추천 영역 서비스 제공을 위한 프로젝트입니다.
관련 태그는 Recommendation System, E-commerce, Collaborative Filtering 입니다.

Introduction

프로젝트 개요

수행 기간	2021.02. ~ 2022.02. (13개월)
팀 구성	총 2명
사용 모델	LGBM
목표값	고객별 추천 상품 TOP 18 (6 Products * 3 Pages)
수집 데이터	현대카드 고객, M포인트몰 상품 데이터

데이터 세부 내용

Customer	성별, 나이 등
Item	가격, 카테고리 (대/중/소), 브랜드, 등
Interaction	고객별 / 상품별 클릭 수, 구매 수 등
Log	고객의 App 활동 기록

3. 2021 M포인트몰 개인화 추천 – 개요 (2/2)

연관 상품 추천에는 기본적인 Collaborative Filtering 에 더해 TF-IDF 에서 차용한 기법을 추가하였습니다.
개인화 상품 추천의 경우 고객 군집화를 기반으로 한 후보 선정 후 LGBM 으로 상품 예측하였습니다.

Problem Definition

✓ 프로젝트 목표

고객별로 클릭할 것 같은 상품을 예측하여 고객에게 추천하자

* 세부적으로는 다음 영역들이 있습니다.

① 연관 상품 추천 (함께 본, 함께 구매한 상품)
② 개인화 상품 추천

연관 상품 추천에는 Collaborative Filtering
개인화 상품 추천에는 LGBM 을 사용해서 성능을 높여보자

3. 2021 M포인트몰 개인화 추천 – Modeling (1/3)

연관 상품 추천 영역에서는 기본적으로 세션 내에서 동시에 클릭/구매한 상품을 추천했습니다.
다만 이 경우에 가벼운 생필품은 모든 상품에서 많이 발견되므로, 이에 대한 패널티 부여를 하여 연관 없는 상품의 추천을 예방했습니다.

Collaborative Filtering

클릭/구매 수	상품 A	상품 B	상품 C
상품 A	-	2	2
상품 B	2	-	1
상품 C	2	1	-

* 연관 상품

연관 순위	1등	2등	3등
상품 A	상품 C	상품 B	…
상품 B	상품 A	상품 C	…
상품 C	상품 A	상품 B	…

TF-IDF

* 상품 A가 생수, 라면 등의 누구에게나 보편적으로 클릭/구매되는 상품이라면
이는 추천의 다양성을 해치며 상품끼리 연관되어 있다고 보기 힘듦.

** TF-IDF 는 모든 문서에서 등장하는 단어의 중요도를 낮게 평가하는 기법
이를 연관 상품 추천에 적용시키자.

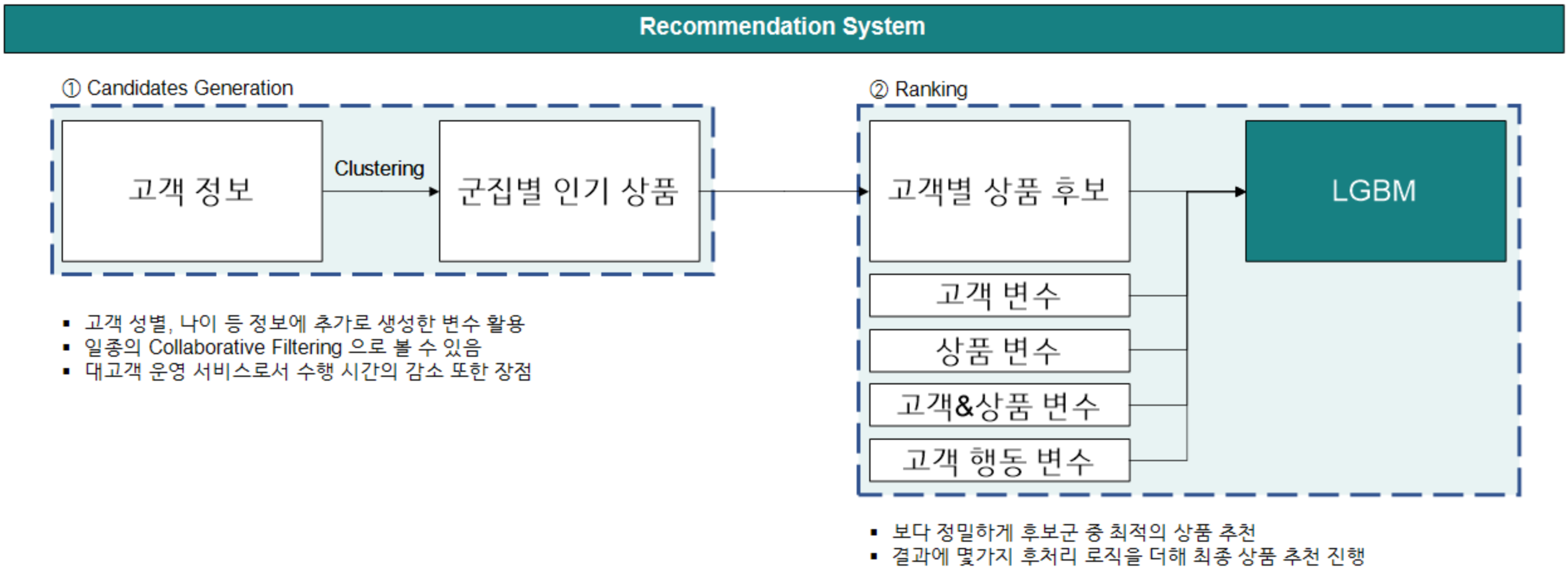
$$TF = \frac{\text{상품 X와 상품 Y가 동시에 클릭|구매된 수}}{\text{상품 X와 동시에 클릭|구매된 모든 상품 수}}$$

$$IDF = \log\left(\frac{\text{총 상품의 수}}{\text{상품 Y와 동시에 클릭|구매된 모든 상품 수}}\right)$$

*** 다른 여러 상품에서 많이 나타날수록 IDF 값은 작아지게 됨.
왼쪽 예시에서 상품 A만 상품 D, 상품 E에 동시에 클릭|구매 되었다면
상품 B의 TF-IDF 적용한 순위는 상품 C, 상품 A 순서가 됨.

3. 2021 M포인트몰 개인화 추천 – Modeling (2/3)

개인화 추천 영역은 고객별 후보 상품 선정 과정과 **Ranking** 과정을 거칩니다.



- 고객 성별, 나이 등 정보에 추가로 생성한 변수 활용
- 일종의 Collaborative Filtering 으로 볼 수 있음
- 대고객 운영 서비스로서 수행 시간의 감소 또한 장점

- 보다 정밀하게 후보군 중 최적의 상품 추천
- 결과에 몇가지 후처리 로직을 더해 최종 상품 추천 진행

3. 2021 M포인트몰 개인화 추천 – Modeling (3/3)

최종적으로 다른 MD 추천 영역에 비해 높은 CTR, CVR 을 보였으며
특히 연관 상품 추천의 경우 1.5배 이상 높은 성과를 나타냈습니다.

Model Serving

✓ 모델 제공 프로세스

연관 상품 추천 & 개인화 맞춤 추천 → Airflow 일 배치 → M포인트몰 앱 제공

- 고객 정보와 앱 로그 기반 추천으로 정확도 높임

***Baseline 대비 1.1배 ~ 1.5배의 CTR, CVR 성과 달성**

* Baseline: MD 상품 추천 영역 CTR, CVR

Ⅲ. SUMMARY

마무리 멘트

학문적 소양이 연구 자체로만 그치지 않고 실제로 활용될 수 있는 방법들을 고민해왔습니다.
누구보다 주도적이고 문제 해결할 줄 아는 지원자가 되겠습니다. 감사합니다.

Problem Solving

✓ 전달할 수 있는 공통 고객이나 변수가 없어.

→ 누구나 사용하는 언어 기반 모델로 해결하자.

✓ 누가 할인을 적게 해줘도 마케팅에 반응할 고객인지 모르겠어.

→ 결제 기록을 기반으로 고객을 새로운 공간에 매핑해서 생각하자.

✓ 모든 상품과 같이 클릭|구매되지만 실제로는 연관이 안 되는 상품들이 있어.

→ TF-IDF 에서 차용하여 모든 상품과 클릭|구매되는 상품에 대해 Panelty 를 부여하자.

감사합니다!

지원자 오진하 드림