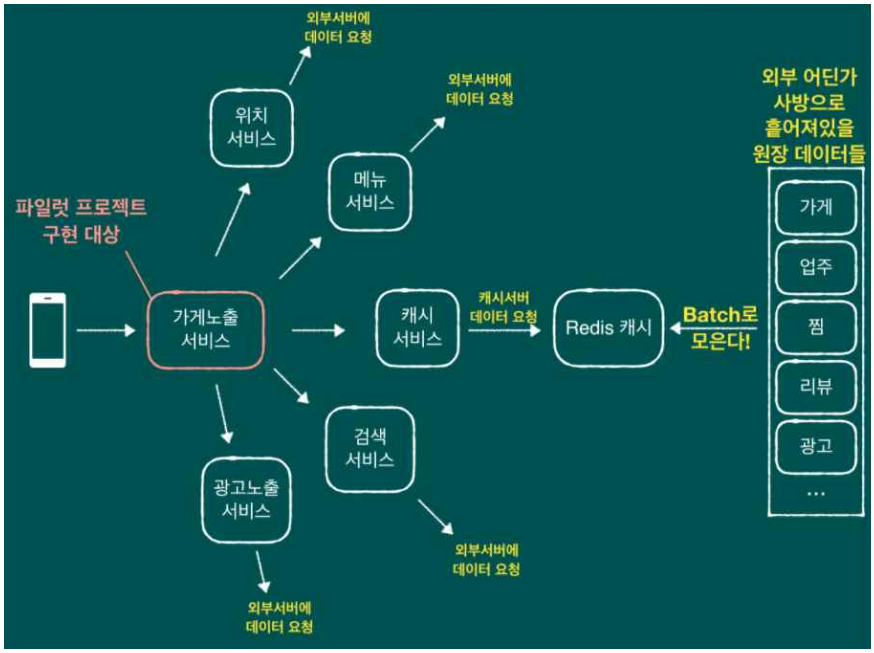
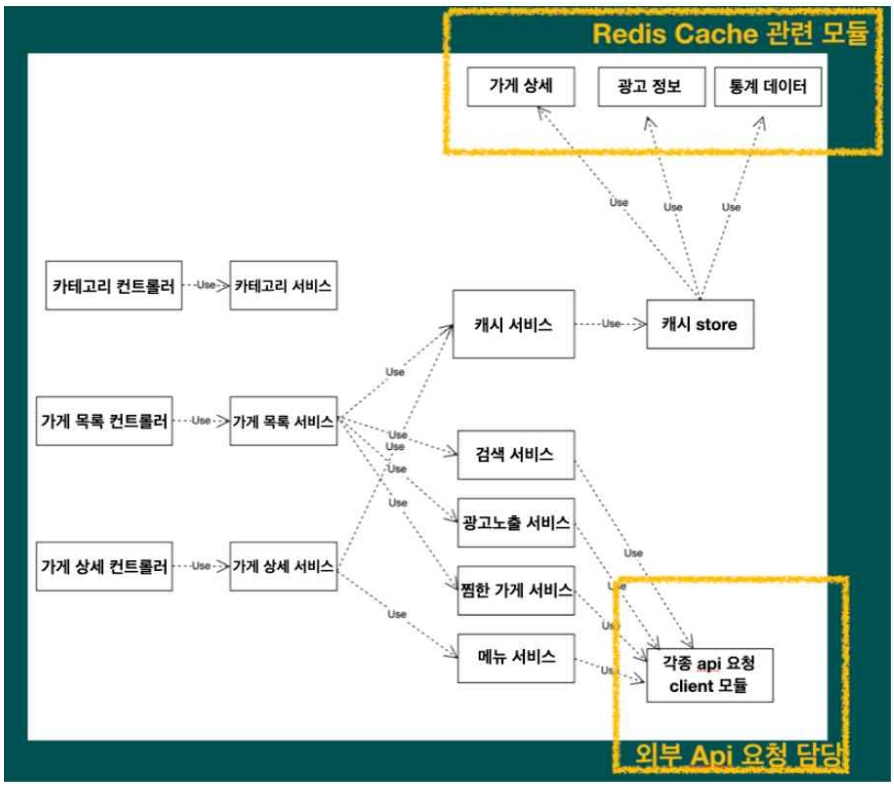
우아한 형제들 기술 블로그

1. 가게 노출 시스템 아키텍처

- 배민의 서비스가 무엇이 있는지 알 수 있고 어떤 값들이 저장이 되어 있는지 확인 가능




※ 클래스 연관 관계도



2. B마트 물류센터 비용 효율화

다음으로, B마트 비용 효율화 업무에 대해 간단히 공유해드리고자 합니다. B마트의 물류 이동 프로세스를 간략히 살펴보면 아래와 같습니다.

1) 거래처에 발주 요청
2) 물류센터에 물품 입고
3) **입고된 물품을 발송 준비**
4) 주문한 고객에게 배송

위 과정 중 예측모델을 통해 효율화하고자 했던 부분은 3. 입고된 물품을 발송 준비하기 위해 투입되는 근무자의 시간이었습니다. 근미래에 발생할 것으로 예상되는 주문수 및 입고수를 기준으로 근무자의 스케쥴을 조정하여, 불필요한 시간의 낭비를 최소화하는 목적으로 예측모델을 개발하였습니다.

지금부터는 예측모델을 개발하고 이를 활용한 과정에 대해 기술하고자 합니다.

3.1 어떤 예측모델을 사용했는가?

3.1.1 학습 데이터

우리는 학습에 이용할 데이터를 만들기 위해 과거 6개월 정도의 데이터를 추출한 후 전처리를 진행했습니다. 이 과정에서 주로 고민했던 부분은 아래와 같습니다.

- 어떤 후보 변수를 모델에 적용하고 최종적으로 어떤 변수를 선택할 것인가?
- 예측력을 높이기 위한 변수 변환 작업을 어떻게 진행할 것인가?
- 이상치와 누락된 값을 어떤 기준으로 처리할 것인가?

주요 내용만 간단히 공유하면, 아래와 같은 전처리 과정을 거쳐 학습 데이터를 만들었습니다.

- 후보 변수 중 최종적으로 선택된 변수는 주로 기간이나 시점을 나타내는 변수이거나 최근의 변화량/트렌드와 관련있는 변수 => 미래 주문량을 예측하기 위해
- 예측력을 높이기 위해 분포 변환 및 표준화(변수의 값을 일정한 범위로 저정)하는 작업을 진행 => 시간 데이터이기 때문에 특정 패턴을 보이는 범위로 묶었을 것
- 이상치의 경우 IRQ 방식을 주로 이용, 누락된 값은 평균과 같은 통계치를 이용해 매꾸거나 때로는 샘플을 제거

> ※ 데이터 분포 4분위 중 1/4(Q1) ~ 3/4(Q3) 구간의 간격을 IRQ라고 하고,
> 상단 기준치 = Q3 + 1.5*IRQ
> 하단 기준치 = Q1 - 1.5*IRQ로 계산한다.
> 이상치는 상단 기준치보다 크거나, 하단 기준치보다 작은 값들이 된다.

3.1.2 모델 선정 및 검증

예측 대상인 주문수, 입고수의 경우 수치형 변수이므로 아래와 같은 회귀모델을 적용해 예측

력을 테스트해보았습니다.
- Linear Regression, Ridge, Lasso 등 선형 모델
- Random Forest, XGboost, LightGBM, KNN 등 비선형 모델

그 결과 LightGBM이 속도 측면이나 예측력, 일반화 가능성을 모두 고려했을 때 가장 좋은 모델로 판단되어 최종 모델로 선정되었습니다. 참고로, 위 과정에서 Pandas 및 Numpy, Scikit-learn을 주로 이용했고, 클래스 및 함수를 모듈로 저장해 엔지니어에게 전달하였습니다. 엔지니어는 테이블 자동화 작업을 진행하여, 운영 담당자가 직접 SQL을 통해 데이터를 추출할 수 있도록 지원하였습니다.

예측모델 테스트/활용 과정에서 제가 얻은 경험은 다음과 같습니다.
- 예측력을 높이기 위해 가장 효과적인 과정은 변수를 변환하거나 파생변수를 추가하는 것
- 하이퍼파라메터 튜닝이나 복잡하고 다양한 모델을 쓰는 것보다, 효과적인 변수를 추가하거나 변환하는 작업이 더욱 효과적
- 위 과정에서 도메인 지식이 중요한 역할 (유관팀을 귀찮게 하더라도 계속 물어보고 가설을 잘 설정하는 것이 중요)
- 한정된 업무 시간을 효율적으로 배분하기 위해 목적을 잊지 않고, 유관팀과 자주 커뮤니케이션하며 의견을 참고하는 것이 유용

4. 마무리
가용 가능 시간을 100으로 봤을 때, 아래와 같이 우선순위에 맞춰 시간 배분을 계획할 수 있습니다.
- 데이터 뭘리티 확보 및 전처리 (40)
- 변수의 생성 및 변환 (30)
- 모델 정교화 및 업데이트 (20)
- 문서화 및 커뮤니케이션 등 기타 작업 (10)

왓챠 기술 블로그
 => 추천 서비스를 위한 데이터는 어떤 것들을 고려해야 하는가
 => 비디오 데이터이므로 데이터는 배민 데이터로 생각해 볼 것.
추천 시스템 현장의 고민
1. 데이터셋은 어디에서 오는가
2020년의 주된 관심사는 시청기록을 활용하는 일이었습니다. 유저의 시청기록에 장,단기 패턴이 있다는 가정하에 다음에 볼만한 콘텐츠를 추천해주는 일인데요.
직관적으로 유저가 재생을 시작한 콘텐츠라고 생각해 볼 수 있습니다. 2시간 러닝타임의 영화를 5분 시청한 기록, 1시간 시청한 기록, 엔딩 크레딧까지 시청한 기록을 같은 위상의 데이터로 취급해도 괜찮을지가 첫 고민이었습니다. 1시간 시청한 유저는 왜 반 정도만 보고 이탈했을까요. 재미가 없어서, 졸려서, 약속 시간 때문에 등등 다양한 이유가 있습니다.
TV 드라마나 예능의 경우 더 복잡해집니다. <부부의 세계>는 16개 에피소드로 구성되어 있고 방송 직후 매주 새 에피소드가 업로드되었습니다. 유저가 5화까지 보고 시청을 안하고 있다면 유저는 <부부의 세계>를 얼마나 봤다고 말해야 할까요. <무한도전> 같은 예능은 1화부터 순서대로 보지 않고 100개가 넘는 에피소드 중 재밌는 에피소드만 골라 봅니다.
이외에도 며칠간의 기록을 사용할까? 실시간성은 얼마나 보장해야 하나? 시청기록이 적은 유저는 평가 데이터에서 빼야 하나? 등등의 명확하게 정할 점이 많았습니다. 뿐만 아니라 팀원들과 데이터에 대한 이해와 정의를 공유하고 합의해서 생각을 맞춰야 했습니다.

| |
|---|
| => 배민은 음식 주문 시스템으로 주문을 하면 끝이다.<br>=> 그럼 사용자들이 음식점에 들어간 클릭이 로그에 기록이 되는가?<br>=> 클릭 로그가 기록된다면 무엇을 보고 음식점을 선택하지 않았는가? -> 사용자가 무엇을 중요히 생각하고 고르는가를 알 수 있음 |

2. 모든 길은 AB 테스트로 통하지만
모의고사를 아무리 잘 받아도 결국엔 수능으로 승부가 나듯이 추천 모델도 결국엔 AB테스트를 통해 더 나은 모델이다/아니다를 정하게 됩니다. 실제로 테스트를 진행하려 하니 구체적으로 정해야 할 요소들이 많았습니다.
첫 번째는 “누구를 대상으로 몇 명이나 실험할까?”입니다. 모든 유저 중에 랜덤으로 뽑으면 되는 걸까요? 유저들이 서비스를 사용하는 패턴느 매우 다양합니다. 특히 서비스를 이용한 지 얼마나 되었는가에 따라 크게 달라집니다. 이제 막 가입한 신규 유저들은 대외적으로 많이 알려진 <왕좌의 게임>이나 <해리포터>, <킬링 이브> 같은 콘텐츠를 많이 봅니다. 반면 오랫동안 많이 이용한 헤비 유저들은 이미 유명한 시리즈는 다 보았을 가능성이 큽니다. 대신 신작으로 들어오는 콘테츠나 대중적으로 많이 보지 않는 콘텐츠를 적극적으로 탐색합니다.
유저를 특정 기준으로 나누어 실험해본 결과 유저군에 따라 추천 성능이 꽤 차이 났습니다. 모든 유저가 만족하는 단일한 모델ㅇ느 없다고 생각합니다. 여러 타겟 유저군을 정의하고 각 집단마다 다른 모델을 실험을 해봐야 합니다. 고민 끝에 ‘신규 유저’를 위한 추천 모델을 테스트하고 싶다고 가정하겠습니다. 몇 명을 대상으로 실험할까요? 샘플 수가 많아질수록 검정력이 강해지니 대략 10,000명으로 정했다고 해보겠습니다. (수치는 예시입니다.) 이런 경우 짧은 시간 내에 유저 10,000명이 가입해야 한다는 뜻이 됩니다. 서비스마다 이 수치가 쉬운 수치일 수도 있고 어려운 수치일 수도 있습니다. 서비스의 유저 흐름을 고려하고 필요한 유저의 숫자

를 조장해야 했습니다.
두 번째는 언제, 얼마나 오래입니다. 유저의 행동은 추천 시스템뿐만 아니라 다양한 요소에 의해 바뀝니다, 최근에는 <중경삼림 리마스터링>, <화양연화 리마스터링> 같은 왕가위 감독 특별전 영화가 마케팅을 통해 유저들에게 많이 노출되었습니다. 콘텐츠에 익숙해져서 추천에 반응을 더 쉬워졌을 수도 있고, 설 연휴처럼 쉬는 날이 많아서 서비스 이용이 늘 수도 있습니다. 따라서 유저에게 영향을 주는 요소들이 최대한 적은 날짜를 골라야 했습니다.
왓챠의 주 비즈니스 모델은 구독입니다. 나은 추천을 통해 유저의 만족도를 높여 구독 연장 비율을 늘리거나 전환 비율을 높일 수 있다면 최고의 성과로 볼 수 있습니다. “추천 모델의 성능이 구독 연장과 인과관계가 있다”라고 주장하려는 건 아닙니다. 그러나 이를 확인해보기 우해서라도 구독 지표를 보려면 최소 1달간 실험해야 합니다. 그렇다고 매번 실험할 때마다 1달씩 실험해야 한다면 실험 주기가 너무 길어지고 잘못하면 안 좋은 모델을 한 달간 실험해야 하는 상황도 생길지 모릅니다. 불상사를 방지하기 위해 구독 이외의 어떤 지표가 좋지 않을 때 조기 종료할지에 대한 계획도 세워야 했습니다.

| |
|---|
| 모든 추천 서비스는 A/B 테스트를 통해 좀 더 좋은 모델로 발전을 한다.<br>하지만 우린 데이터를 한번 받고 개인화 추천 서비스를 작성한다.<br>과거 1년치를 받고 분기별로 나눠서 3번의 A/B 테스트를 진행하는 방법은?<br>A/B 테스트를 진행할 때 고려해야 할 점이 나와 있다. |

3. 모델이 트렌드를 반영하려면 (모델이 만들어진 후 서비스화가 되면)

| |
|---|
| 트렌트를 반영하려면 트렌드 주기를 파악할 수 있어야한다.<br>그럼 꾸준한 데이터를 받아야 하지만 지금 단계에서는 진행하기 어려울 것으로 보임 |

Netflix 기술 블로그
=> 서비스 진행 후 Netflix의 홈페이지 구성에 대한 내용이 있음
=> 화면 구성은 개인 추천 랭킹 시스템도 있지만 국가별로 실시간 순위와 같은 대중적인 추천도 해줌