포트폴리오

응용통계학과

신보현

Contents

0. 관심 분야

1. 직무 관련 역량 & 연관 포트폴리오 요약

2. 와이즐리 분석가 인턴십

3. SK C&C 분석가 인턴십

4. 데이터 파이프라인 프로젝트

5. 준실시간 시각화 프로젝트

6. 환경부 데이터 공모전

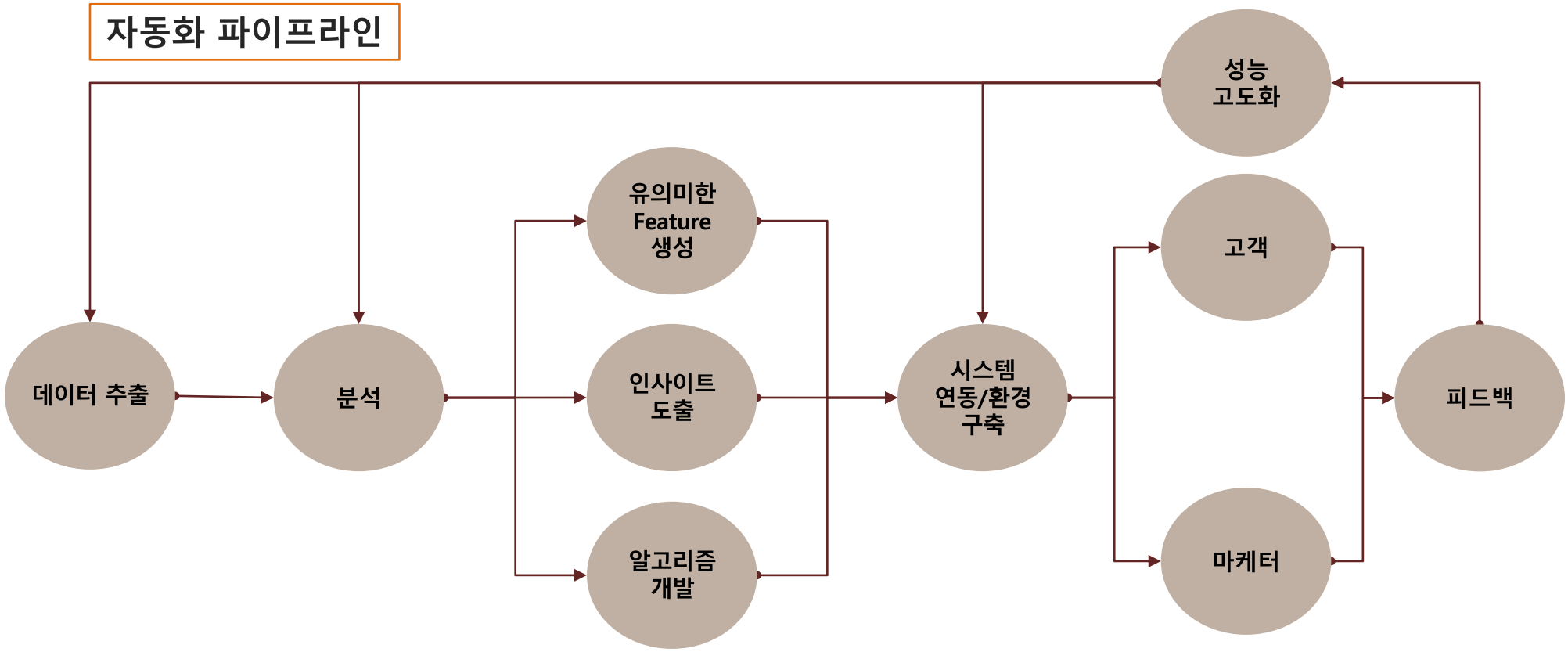
7. 빅데이터 콘테스트 공모전

8. Latent Dirichlet Allocation 연구

9. 측정 오차 모형 연구

※ 클릭하시면 해당 프로젝트 페이지로 이동합니다.

0. 관심 분야

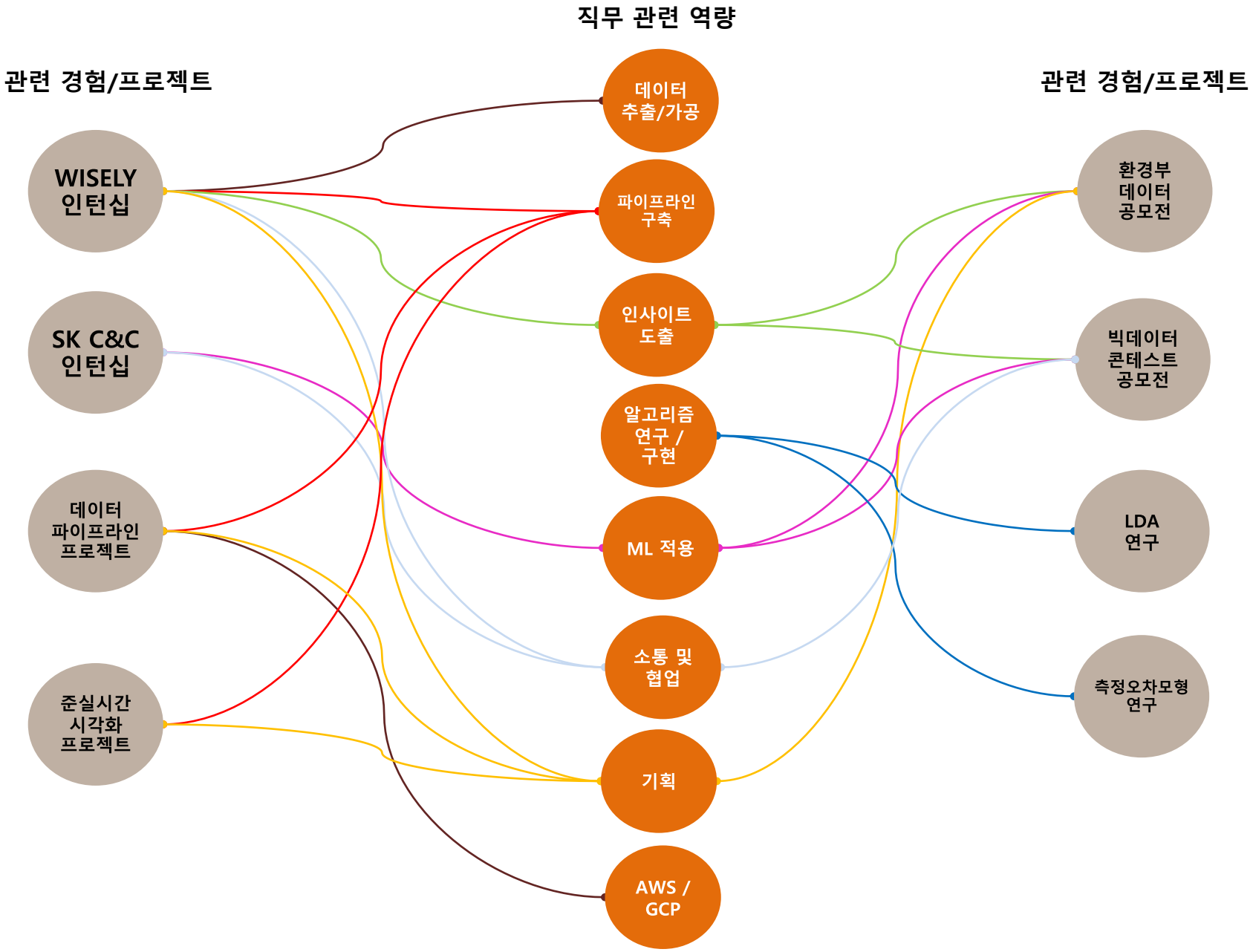


1. 분석에서 끝나는 것이 아닌, 분석 결과를 서비스화하고 제 3자에게 유용한 형태로 만드는 것
2. 계속된 유지 및 보수로 성능을 고도화하는 것
   - 최신 논문 리서치 및 새로운 알고리즘 구현
   - 사용자의 피드백을 반영하여 서비스를 개선
3. 이러한 전 과정을 자동화하여 사람의 수고가 최대한 덜 들어가게 개발하는 것

Chapter 1

직무 관련 역량 & 연관 포트폴리오 요약

1. 직무 관련 역량 & 연관 포트폴리오 요약



Chapter 2

WISELY 데이터 분석가 인턴십

상세 내용은 여기를 통해 확인 가능

2. WISELY 데이터 분석가 인턴십 (개요)

개발 역량

데이터 추출/가공
- RAW 데이터 추출
- 유의미한 지표로 가공

파이프라인 구축
- 데이터 자동 업데이트
- 고객 만족도 자동 발송

인사이트 도출
- 대시보드 제작
- 유의미한 고객군 발굴

소통 및 협업
- 타 부서와의 협업
- 적극적인 소통 자세

기획
- SQL 교육
- 인턴십 Final 프로젝트

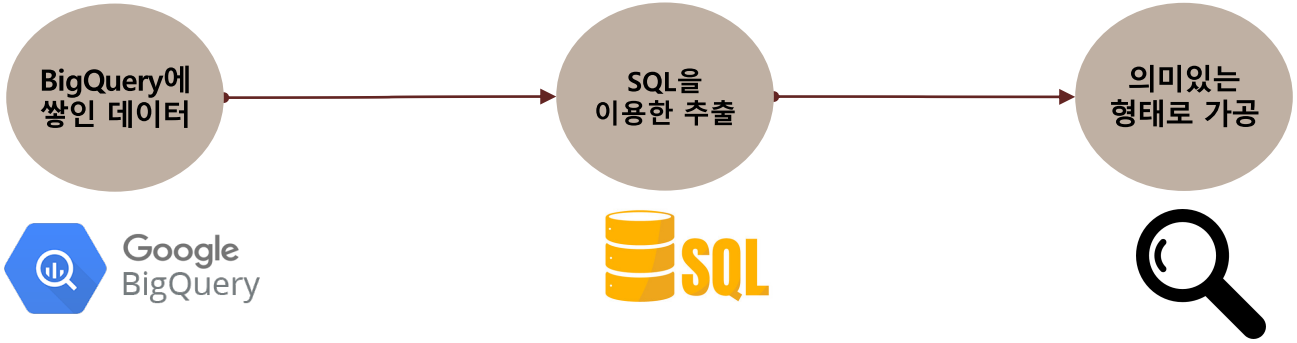
인턴십 개요

WISELY 회사 소개	• 비즈니스 모델이 면도기 구독 서비스인 스타트업 • 면도기 및 면도날의 독과점 시장에 맞서 품질 좋은 제품을 저렴한 가격에 유통하는 것이 주요 Mission • 데이터에 기반한 의사결정 문화가 정착되어 있음
기본 정보	• 포지션: 데이터 분석가 인턴 • 채용 기간: 21.3 ~ 21.6 (3.5개월) • 근무 부서: Insight & Data
주요 업무	• 팀원들의 데이터에 기반한 의사결정 지원 → RAW 데이터 추출 및 유의미한 지표로 가공, 분석 의뢰 수행 • Automatic 파이프라인 구축 → 데이터 자동 업데이트 및 DB와 연동, 고객 만족도 조사 발송 자동화 • 주요 프로젝트 → 전사 차원의 대시보드 제작, CX 팀의 대시보드 제작

2. WISELY 데이터 분석가 인턴십 (주요 업무)

1. 팀원들의 데이터에 기반한 의사결정 지원

1.1. RAW 데이터 추출 및 유의미한 지표로 가공



가공 데이터 예시	• 매출 • 재구매율 (By Cohort, sku) • 잔존율 또는 이탈율 (By Cohort, sku) • 장바구니 가격 (Basket Size) • 장바구니 sku 구성 • 고객 응대 시 요청율, 응답률 • CRM 가능한 고객의 수

고려 사항	• Cohort → 구매 회차, 구매 유형, 상품 구성 종류, sku • 대부분 담당자와 의논 후 지표를 계산 → 데이터 분석 의뢰 업무

2. WISELY 데이터 분석가 인턴십 (주요 업무)

1. 팀원들의 데이터에 기반한 의사결정 지원

1.2. 분석 의뢰업무 수행

팀원의 분석 의뢰 → **분석 방향 설정** → **분석 진행** → **결과 전달**

- 가설 A 증명 니즈

- 가설 A 내용 상호 확인
- 토의를 통해 의견 교환

- 필요하다면 설문조사 발송

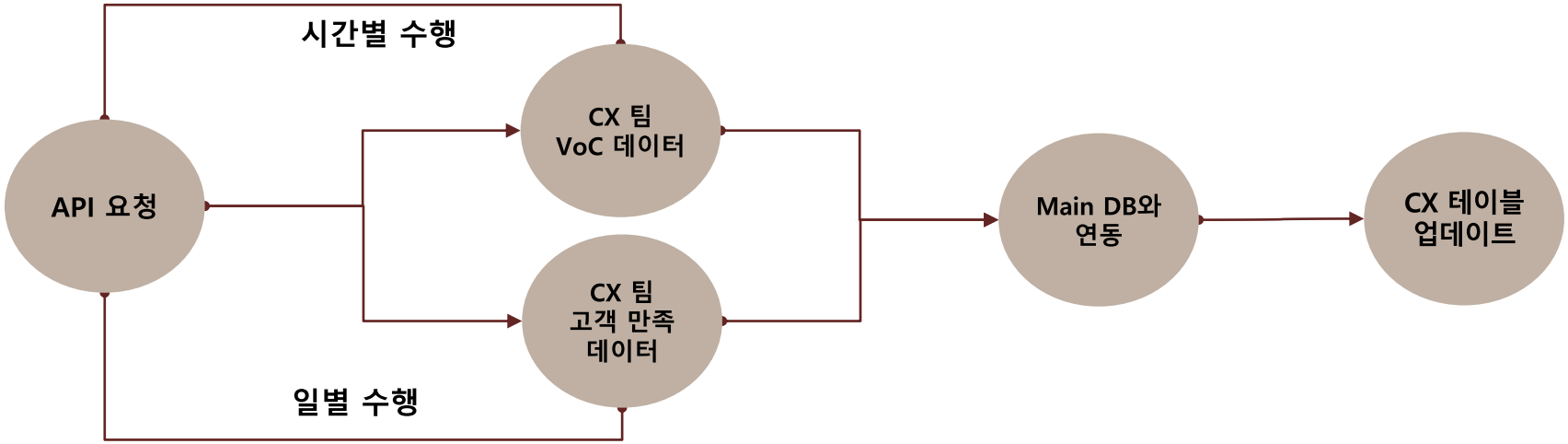
분석 의뢰	1. 첫 구매 형태에 따른 고객 만족도 변화 및 그 원인 2. 단일제품 구독자와 복수제품 구독자의 고객 경험 차이 3. A 브랜드의 잔존 및 이탈 고객 특성 파악을 통해 재구매로 유도할 수 있는지 가능성 검토 4. 브랜드의 lock-in window가 대략 언제 해당되고 이들의 재구매율 분석을 통해 재구매 가능성 고객들을 잃고 있는 것이 아닌지 검토

도출 Insight	1. 첫 구매가 1회성 구매인 고객의 제품 만족도가 높았지만 그 원인은 끝내 밝히지 못함 2. 복수제품 구독자의 구독 취소, 정지, 미루기 등의 횟수가 더 높게 나타났음 3. 다수의 고객이 무료체험 후 2주 이내에 이탈하는 것으로 나타났기 때문에 이를 golden time으로 정의함 4. 일자별로 확인한 결과, 2주 이내에 이탈 확률이 높아졌는데 이는 결제 3일전 알림톡이 주요 원인으로 밝혀짐. 또한 이들의 재구매율이 높았기 때문에 이 고객들을 잡는 것이 하나의 전략이 될 수 있음

2. WISELY 데이터 분석가 인턴십 (주요 업무)

2. Automatic 파이프라인 구축

2.1. 데이터 자동 업데이트 및 DB와 연동

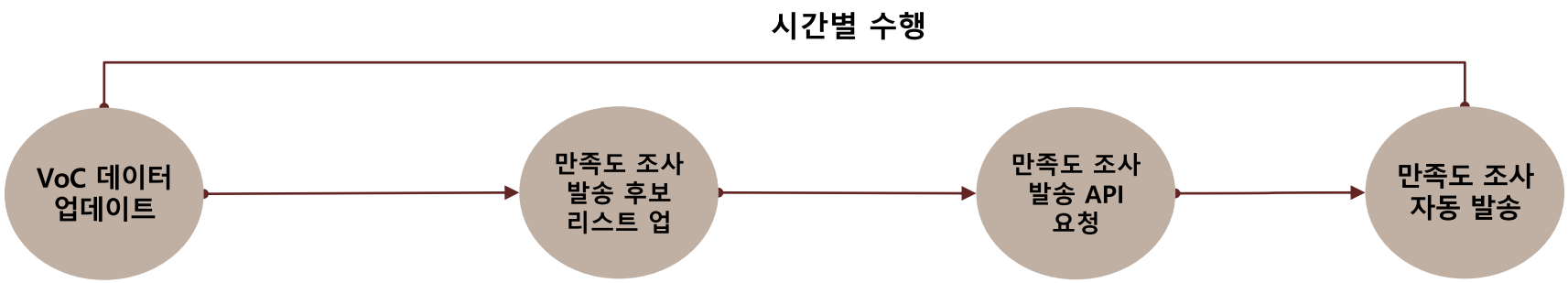


문제 정의	• CX 팀원들이 업무를 위한 데이터를 각자 수동으로 관리해 왔음 • 데이터를 매일 다운해야하는 불편함이 존재 • 데이터 팀에서 사용하는 DB와 연동을 할 수 없고 따라서 고객 경험 데이터와 구매, 구독 데이터를 연결하여 분석을 진행할 수 없었음
문제 해결 과정	• CX 팀원들과 인터뷰를 하며 그들의 업무 방식을 습득함 • 데이터 업데이트 과정에서 발생할 수 있는 이슈에 대해서 논의 → EX: 팀원이 하루 일과 종료 이전에 수정하는 VoC에 대한 처리 • CX 팀원들이 업무를 위한 데이터를 자동으로 다운받고 이를 DB와 연동함

2. WISELY 데이터 분석가 인턴십 (주요 업무)

2. Automatic 파이프라인 구축

2.2. 고객 만족도 조사 자동화

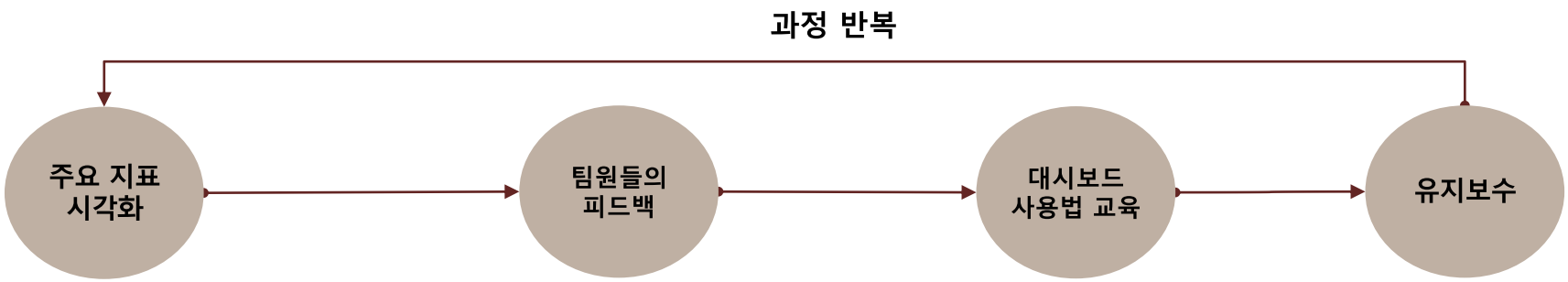


문제 정의	• CX 팀원들이 상담 이후 만족도 조사 알림톡을 직접 발송하였음 • 하루 상담 건이 많아질수록 이 작업은 업무 비효율을 가져옴 • 만족도 조사를 발송할 때, VoC KEY와 엮어서 발송하지 않았기 때문에 VoC와 통합된 분석 불가능
문제 해결 과정	• 업무 담당자와 인터뷰를 하며 만족도 조사 과정에 대해서 습득 • 만족도 조사 발송 과정에서 고려해야하는 점에 대해서 논의 → 발송 제외 대상 VoC 파악 • 시간별로 만족도 조사 알림톡을 자동으로 발송

2. WISELY 데이터 분석가 인턴십 (주요 업무)

3. 대시보드 제작

3.1. 전사 차원의 대시보드 제작

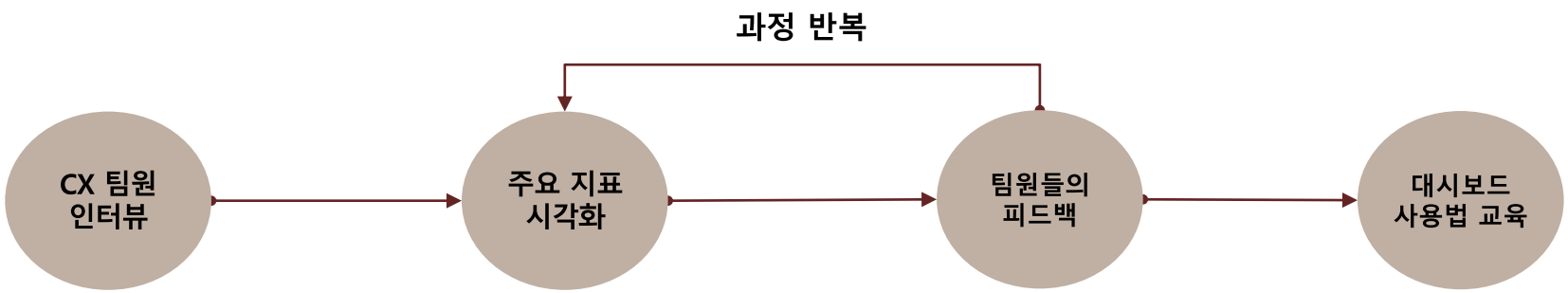


문제 정의	• 서로 다른 팀원들의 비슷한 요청을 처리하며 업무 비효율이 발생 • 겉으로 동일한 지표 A에 대해서 팀원들이 서로 다르게 정의하고 있었음 • 하나의 통일된 플랫폼으로 빠르고 일관된 의사결정을 지원하고자 함
문제 해결 과정	• 평소 분석 요청이 많았던 지표 중심으로 팀원들이 동의할만한 기준을 설정, 지표를 시각화 • 팀원들의 피드백 수용 및 추가 시각화 인풋을 받음 • 대시보드 사용법 교육을 통해 팀원들이 효과적으로 이용할 수 있도록 지원 • 이후, 대시보드 유지 및 보수
주요 Insight	• 고객들의 평균 정기 결제 Push을 새롭게 알 수 있었음 • sku 별로 잔존율이 다르게 나타나는 제품이 존재

2. WISELY 데이터 분석가 인턴십 (주요 업무)

3. 대시보드 제작

3.2. CX 팀 대시보드 제작



문제 정의	• 여태껏 CX 팀이 데이터를 별도로 관리해오며, 고객 경험 데이터에 대한 효과적인 분석 수행이 어려웠음 • 앞서, CX 팀의 데이터 파이프라인을 데이터 팀의 DB와 연동하였는데 이를 통해 데이터 팀에서 대시보드 제작이 가능해졌음
문제 해결 과정	• CX 팀원들과 인터뷰를 하며 그들의 업무 전 과정을 습득함 • 대시보드 또한 그들이 보는 것이기에 팀원들의 피드백 및 시각화 과정을 반복함 • 대시보드 사용법 교육을 통해 팀원들이 효과적으로 이용할 수 있도록 지원
주요 Insight	• 시간대별, 요일별 고객 응대 맞춤 전략 수립 → 요청율, 응답률 추이 분석 가능 • 주요 Event 발생 시, 고객 응대 Performance 피드백

2. WISELY 데이터 분석가 인턴십 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - SQL을 사용하여 RAW 데이터 추출 및 의미있는 형태로 가공하였음
  - 파이프라인 구축을 통해 데이터의 흐름을 파악하는 능력을 개발함
- **업무 수행적인 측면**
  - 프로젝트를 기획하고, 리드하고 끝까지 책임감있게 완수했기에 값진 경험을 할 수 있었음
  - 분석 과정에서 팀원들과 얼마나 원활하고 명쾌하게 소통하느냐에 따라서 결과가 달라짐을 깨달음
  - 소통 과정에서 상대방의 니즈를 파악하기 위해 노력하는 것이 중요함을 깨달음

한계점

- **기술적인 측면**
  - 고도화된 분석 기법을 적용하지 못한 점
- **업무 수행적인 측면**
  - 프로젝트 업무 배분을 적절히 하지 못한 점

    → 최종 프로젝트를 수행했을 때, 과제 단계별 예상 시간을 조금씩 초과하여 수행하였음

Chapter 3

SK C&C 분석가 인턴십

3. SK C&C 분석가 인턴십 (개요)

개발 역량

인사이트 도출
- 이상사건 관련 변수 탐색

소통 및 협업
- 팀원과의 소통을 통한 분석 방향 논의

ML 적용
- AutoEncoder 모형 개발

인턴십 개요

SK C&C 회사 소개	• SI 회사 • 고객사에게 Digital Solution을 제공하는 것이 주요 Mission
기본 정보	• 포지션: 분석가 인턴 • 채용 기간: 21.7 ~ 21.8 (1.5개월) • 근무 부서: Data분석Service팀
주요 업무	• 고객사의 공정데이터에서 이상사건을 감지하는 Rule 도출

3. SK C&C 분석가 인턴십 (내용)

인턴십 수행 과정

인턴십 프로젝트	• 투입된 팀의 프로젝트 결과물을 고도화하는 업무 • 이상사건 전후로 예측 모형의 정확도가 떨어지므로 이상사건을 탐지해서 해당 기간을 제외해야함 • 이상사건을 탐지하는 규칙을 생성하여 해당 기간에는 예측 모형이 구동되지 않도록 하는 것이 목적

인턴십 수행 내용

EDA를 통한 Rule 도출	• 관련있는 변수를 선정하고 EDA의 순서를 중요도에 따라서 결정 • 데이터를 Train / Test 기간으로 나누고, Train 데이터에서 이상사건 전후로 급격한 변곡을 보이는 변수를 확인 • 해당 변수들을 조합하여 Simple Rule 생성
DL 모형 개발	• Simple Rule을 통해서 감지되지 않은 이상사건을 탐지하기 위해 DL 모형 개발 • AutoEncoder 모형이 복원한 값이랑 실제 값이 차이가 크게 난다면 이상 사건이 발생했다고 판단
결과	• 한 건의 이상사건을 제외한 나머지 이상사건을 탐지할 수 있었음

3. SK C&C 분석가 인턴십 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - 현업 공정 데이터을 사용하여 EDA부터 DL 모형 개발까지 경험해보았음
- **업무 수행적인 측면**
  - 팀원과의 소통을 통해 기준을 정하고 결과물을 발전시켜나가는 경험을 해보았음

한계점

- **기술적인 측면**
  - 좀 더 적합한 AutoEncoder 모형을 탐색하지 못했다는 점
    - 정밀한 파라미터 튜닝 부족
    - 시계열 데이터에 적합한 AutoEncoder 모형 개발 부족

Chapter 4

데이터 파이프라인 프로젝트

상세 내용은 여기를 통해 확인 가능

4. 데이터 파이프라인 프로젝트 (개요)

개발 역량

파이프라인 구축

- 데이터 수집부터 간단한 시각화까지

기획

- 아키텍쳐 기획

AWS / GCP

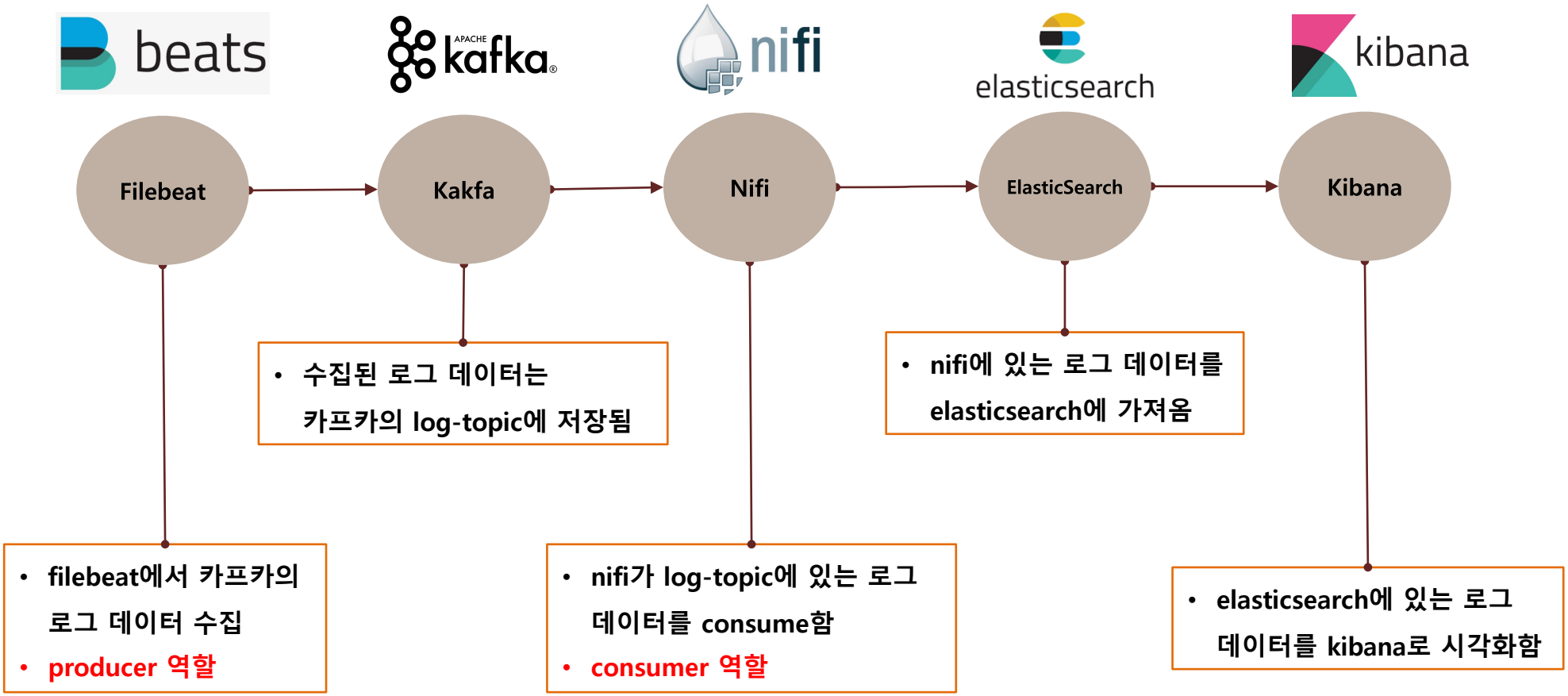
- 여러 대의 가상환경 구축 후 연동

프로젝트 개요

목표	**데이터를 수집하고 저장하며 시각화하는 일련의 파이프라인 개발**
기본 정보	• 개인 프로젝트 • 수행 기간: 21.2 (1개월) • 수집 데이터: Kafka 로그 데이터
동기	• Kafka 스터디가 끝나갈 쯤에 Kafka 클러스터가 중심이 되는 파이프라인을 개발하고자 하였음 → Why Kafka? Pub/Sub 모델의 Kafka가 빅데이터를 처리하는데 가지는 장점에 매력을 느낌 • 정적인 데이터에서 동적인 데이터로의 관심 변화 → 분석에서 끝나는 것이 아닌, 실시간 수집되는 데이터를 이용하여 분석 결과를 업데이트하고 서비스화하는 것에 흥미를 느끼기 시작
맡은 역할	• 파이프라인 아키텍쳐 기획 • 파이프라인 구현

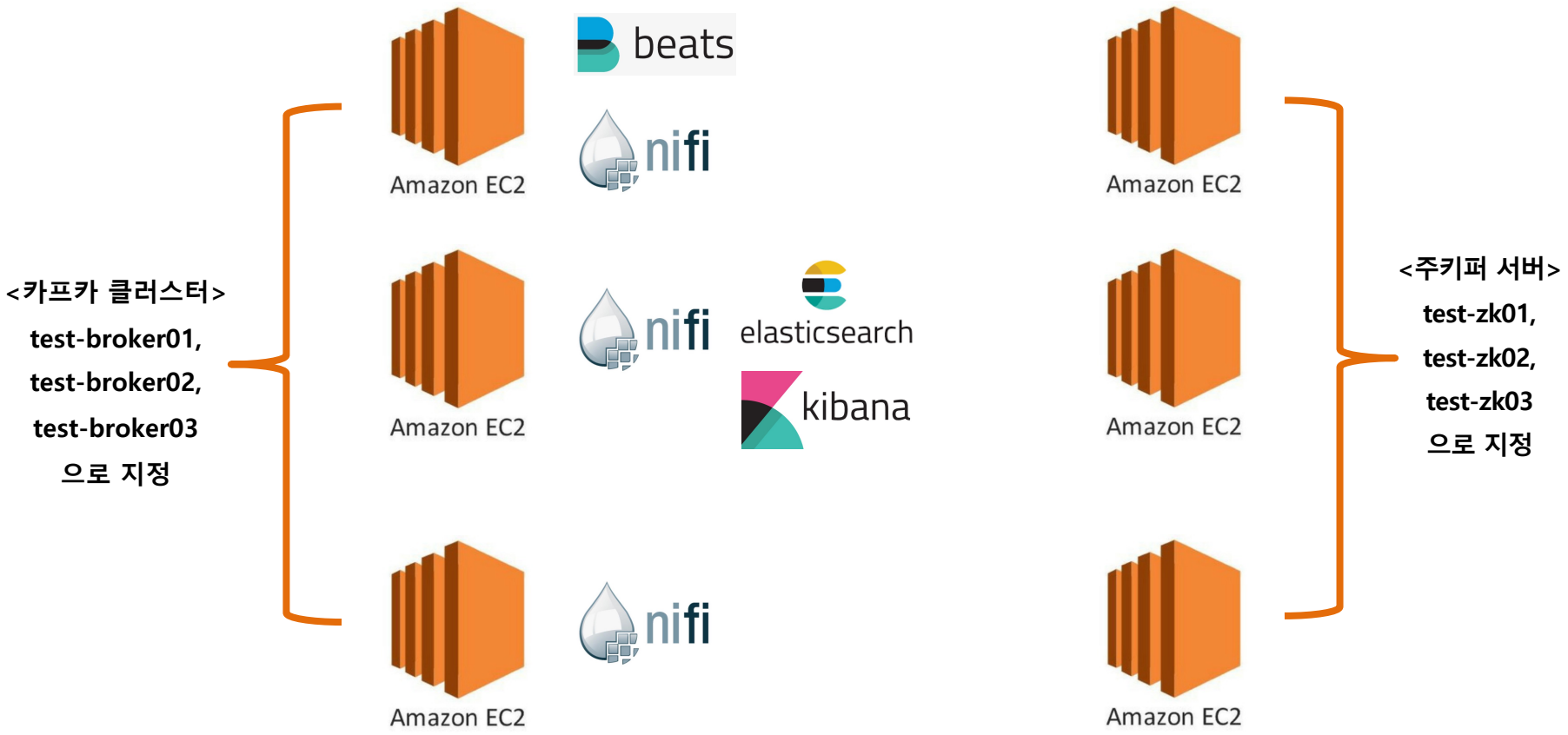
4. 데이터 파이프라인 프로젝트 (아키텍쳐)

파이프라인 아키텍쳐



4. 데이터 파이프라인 프로젝트 (개발 환경)

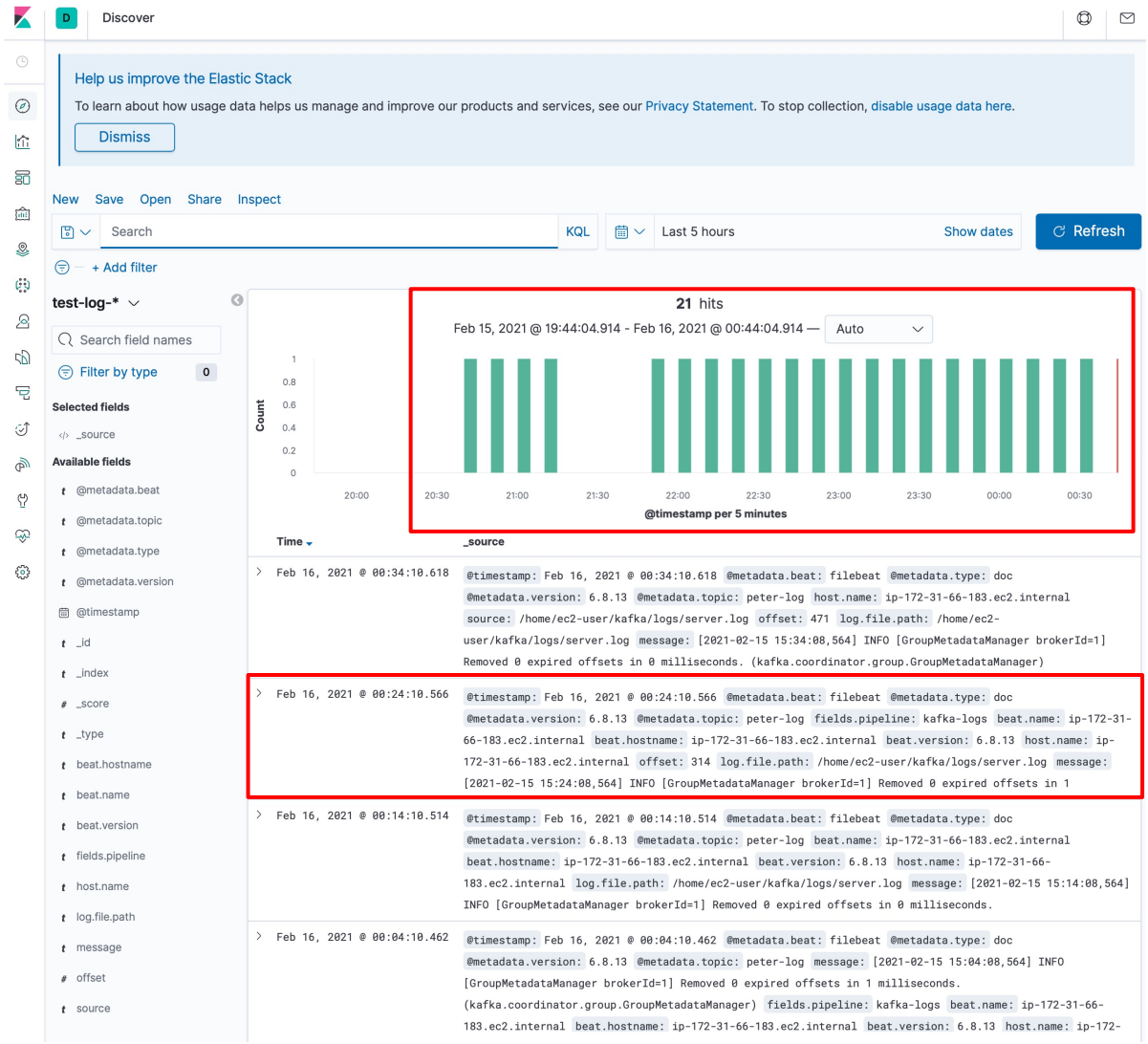
파이프라인 개발 환경



4. 데이터 파이프라인 프로젝트 (결과)

Kibana 시각화 결과

- 시간대별로 로그 개수를 확인할 수 있음
- 각 로그의 내용을 일목요연하게 확인할 수 있음



4. 데이터 파이프라인 프로젝트 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - AWS에서 Kafka 클러스터를 구축해보고 여러 시스템을 연결해봄
  - 트러블 슈팅 경험
    - ✓ 로그를 면밀히 분석하는 습관이 생김
    - ✓ 트러블 슈팅 경험을 기록하여 이후의 시간을 절약

한계점

- **기술적인 측면**
  - 좀 더 의미있는 데이터로 진행을 해보지 못한 점
  - ML Ops가 아닌, 단순 데이터 파이프라인이라는 점
  - 실제 구축해보고 운영해본 것이 아니기 때문에 유지 및 보수의 경험을 하지 못했다는 점

Chapter 5

준실시간 데이터 시각화 프로젝트

상세 내용은 여기를 통해 확인 가능

5. 준실시간 데이터 시각화 프로젝트 (개요)

개발 역량

파이프라인 구축

- 데이터 수집부터 간단한 시각화까지

기획

- 아키텍쳐 기획

프로젝트 개요

목표	**축구 데이터를 수집하고 저장하며 시각화하는 일련의 과정을 자동화하는 파이프라인 개발**
기본 정보	• 개인 프로젝트 • 수행 기간: 21.2 (1개월) • 수집 데이터: 해외 축구 데이터
동기	• 데이터 파이프라인 프로젝트를 종료하고 좀 더 의미있는 데이터로 다른 사람들에게 유용한 서비스를 제공할 수 있는 파이프라인에 대해서 고민 • 당시에 축구 뉴스 기사를 보며 외국 기사들을 기자들이 자의적으로 해석하는 것에 부정적인 시선이 있었음 • 이러한 불편함을 해소하기 위해 좋아하는 축구 팀의 뉴스 기사를 수집하고 시각화하는 파이프라인을 개발 → 일련의 과정은 모두 자동으로 진행
맡은 역할	• 파이프라인 아키텍쳐 기획 • 파이프라인 구현

5. 준실시간 데이터 시각화 프로젝트 (아키텍쳐)

파이프라인 아키텍쳐

- soccer-topic에 데이터 저장

- 실시간으로 카프카 토픽에 있는 메세지를 JDBC에 저장
- Consumer 역할

- Crontab을 통해 Python 크롤링 스크립트 스케쥴링
- Producer 역할

- Kafka Conncet을 이용해서 ROCKSET과 연결

- Tableau을 통한 시각화

5. 준실시간 데이터 시각화 프로젝트 (아키텍쳐)

Tableau 시각화 결과

- 탭에 따라서 서로 다른 시각화를 구성
- 5분마다 시각화 자동 업데이트
  - 준실시간 기능 구현

	Play	Points	Won	Drawn	Lost	For	Against	GD
	25	59	18	5	2	50	15	35
	25	49	14	7	4	53	32	21
	25	49	15	4	6	44	27	17
	25	45	13	6	6	39	29	10
	25	43	12	7	6	41	25	16
	25	40	11	7	7	45	34	11
	24	40	12	4	8	37	33	4
	23	36	11	3	9	37	26	11
	24	36	10	6	8	37	27	10
	25	35	11	2	12	43	43	0
	25	34	10	4	11	31	26	5
	25	33	9	6	10	26	32	-6
	25	32	9	5	11	29	43	-14
	25	30	8	6	11	31	43	-12
	25	28	7	7	11	18	30	-12
	25	26	5	11	9	26	32	-6
	25	25	7	4	14	26	43	-17
	25	22	4	10	11	21	32	-11
	25	14	2	8	15	19	55	-36
	25	11	3	2	20	15	41	-26

5. 준실시간 데이터 시각화 프로젝트 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - Kafka와 Tableau을 연결하는 파이프라인을 구축해봄
  - 실시간 데이터에 대한 고민
    - ✓ Crontab을 이용한 Job 스케쥴링 → 추후에는 AirFlow을 사용해볼 예정
    - ✓ 데이터를 실시간으로 가져오는 것과 동시에 분석 결과에 대한 실시간 업데이트의 고민

한계점

- **기술적인 측면**
  - ML Ops가 아닌, 단순 데이터 파이프라인이라는 점
  - 실제 구축해보고 운영해본 것이 아니기 때문에 유지 및 보수의 경험을 하지 못했다는 점

Chapter 6

환경부 데이터 공모전

상세 내용은 여기를 통해 확인 가능

6. 환경부 데이터 공모전 (개요)

개발 역량

인사이트 도출
- 실생활에 도움이 되는 결과 도출

ML 적용
- ML을 이용한 문제 해결

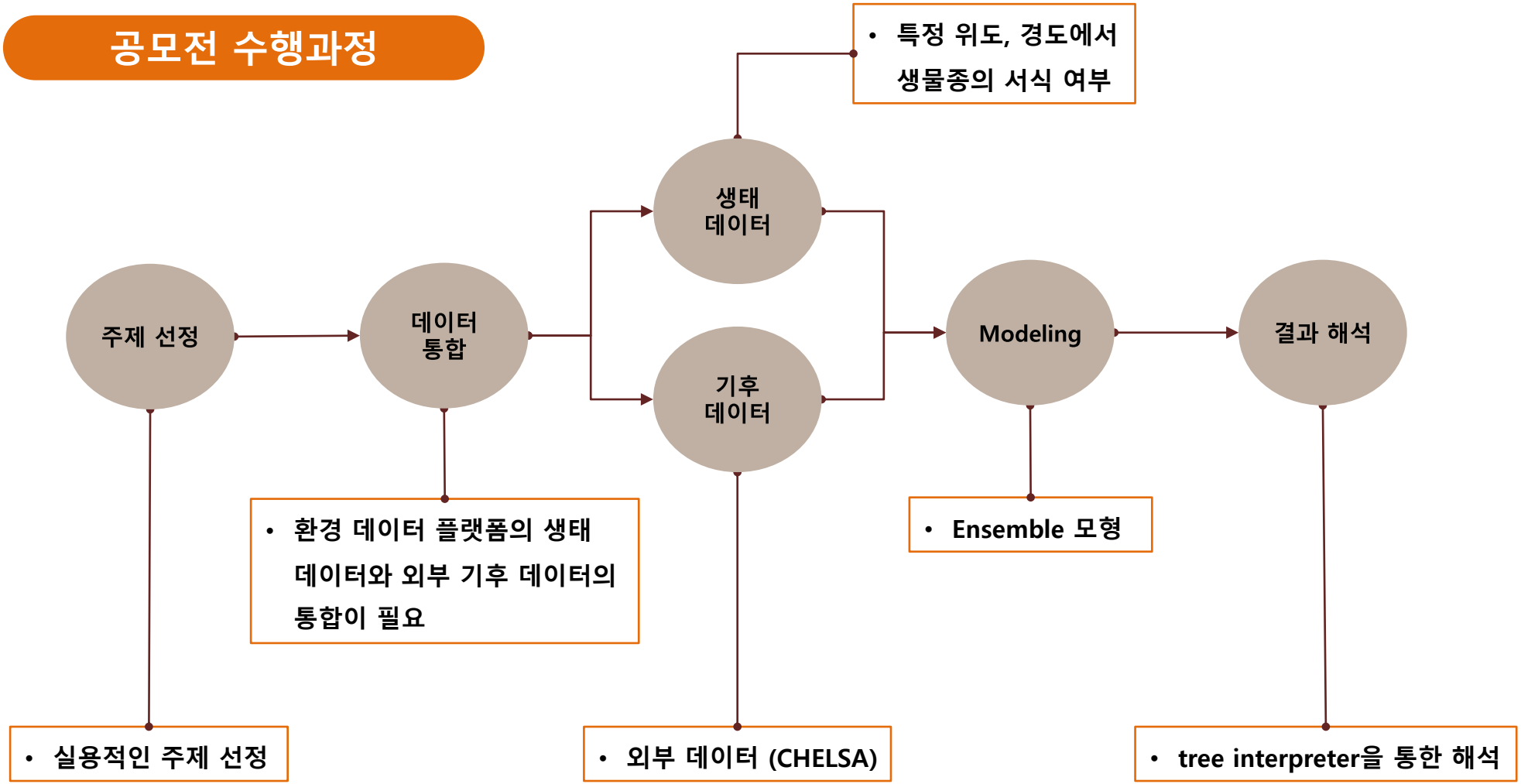
기획
- 주제 선정부터 결과 도출까지

공모전 개요

주제	**생태 데이터를 이용하여 생물종의 서식 여부 예측**
기대 효과	• 특정 지역의 생물 서식 여부를 미리 예측함으로써 생태학자의 효율적인 연구를 도울 수 있음 (달성) • 인천 수돗물 사태의 원인인 깔따구류와 같은 유해 생물의 서식 여부를 미리 예측할 수 있음 (달성) • 멸종위기 동식물의 서식 여부를 예측하여 이들을 보호하는데 기여할 수 있음 (미달성)
동기	• 주제 및 데이터를 직접 선정하여 문제를 정의하는 능력을 기르고자 함 • 학문적으로든, 실생활로든, 도움을 줄 수 있는 분석에 대해서 고민하고자 함
맡은 역할	• 주제 기획 • 데이터 통합 • Modeling • 결과 해석
수상	• 우수상 (3위)

6. 환경부 데이터 공모전 (수행 과정)

공모전 수행과정



CHELSA
Climatologies at high resolution for the earth's land surface areas

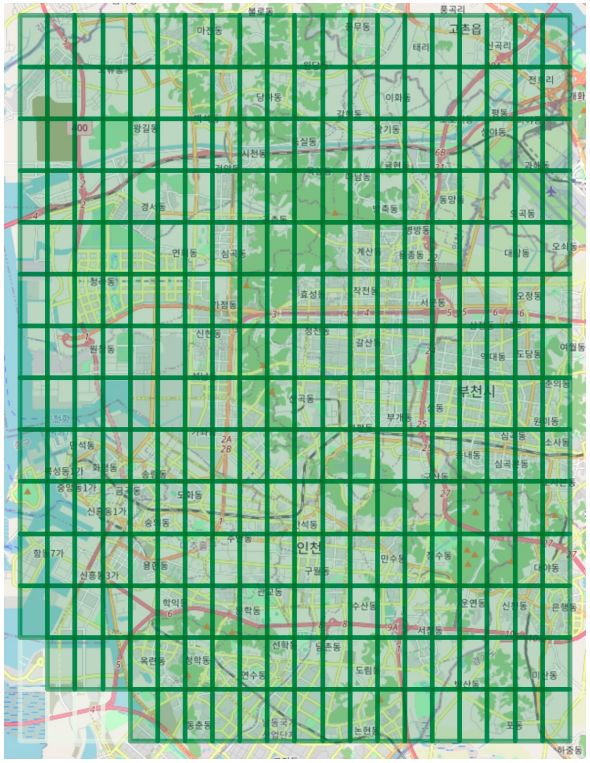
6. 환경부 데이터 공모전 (결과)

정량적인 결과

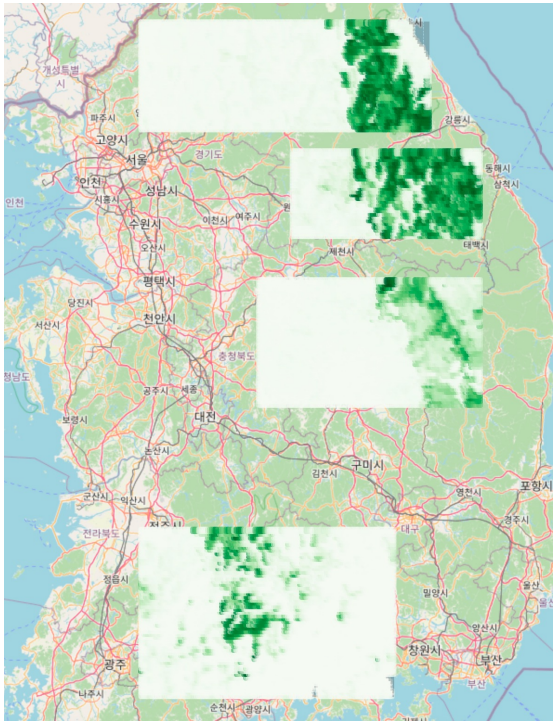
	조류	곤충	자서무척추동물	식물
고려한 종 개수 (수행한 모형 개수)	28	865	200	698
ROC AUC >= 0.7인 종 개수	10	80	71	61
ROC AUC 평균 값	0.86	0.77	0.77	0.78
ROC AUC 최대값	0.97 (섬휘파람새)	0.97 (제주왕딱정벌레)	0.99 (늪깔따구아과류)	0.97 (참식나무)

6. 환경부 데이터 공모전 (결과)

결과 활용 예시



- 위 그림은 인천 지역에 대해서 깔따구류가 서식할 확률을 시각화한 결과
- 색이 진할 수록 확률이 높은 것이고 인천 지역에 깔따구류가 서식할 확률이 높음을 알 수 있음
- 이와 같이 유해한 생물의 서식 여부를 미리 추론할 수 있다면 인천 수돗물 사태 등을 방지할 수 있을 것으로 기대됨



- 위 그림은 탐사되지 않은 지역에 섬휘파람새가 서식할 확률을 시각화한 것
- 색이 진할수록 섬휘파람새가 서식할 확률이 높음을 의미
- 이와 같이 생물 종이 특정 지역에 서식할 확률을 구할 수 있다면, 생태학자가 확률이 높은 지역을 중심으로 탐사하여 효율적인 연구를 진행할 수 있을 것으로 기대됨

6. 환경부 데이터 공모전 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - ML을 이용해서 문제를 해결
- **기획적인 측면**
  - 프로젝트 주제를 실용성에 근거하여 탐색함
  - 원하는 결과를 도출

한계점

- **기술적인 측면**
  - 멸종위기 생물종에 대한 모형은 극불균형 분류를 해야 했고 정확도 높은 모형을 확보하기 어려웠음
  - 분석 결과를 활용할 수 있는 Platform 개발을 하지 못했음

Chapter 7

빅데이터 콘테스트 공모전

상세 내용은 여기를 통해 확인 가능

7. 빅데이터 콘테스트 공모전 (개요)

개발 역량

인사이트 도출

- 각 변수 예측 시 주요하게 사용된 변수 탐색

ML 적용

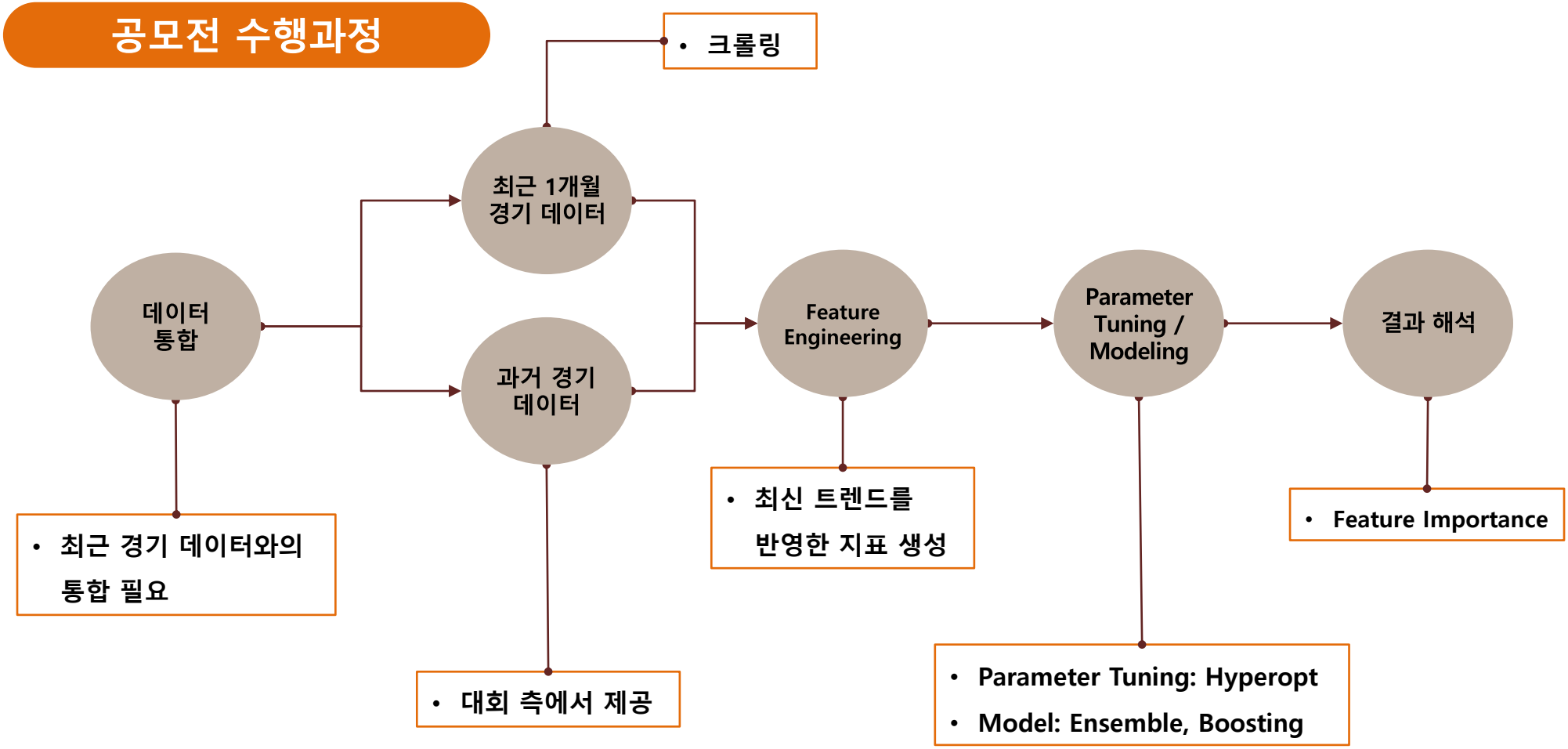
- ML을 이용한 문제 해결

공모전 개요

주제	**KBO 야구 데이터를 활용한 잔여경기의 타율, 방어율, 승률 예측**
기대 효과	• 잔여 경기에 대한 타율, 방어율, 승률을 정확하게 예측할 수 있다면 미래 경기에 대한 불확실성을 어느 정도 해소할 수 있음
동기	• 주제 및 데이터가 주어진 공모전이었으므로 정확도를 높이는데 초점을 맞춤
맡은 역할	• 추가 데이터 크롤링 • Feature Engineering • ML 모형 파라미터 튜닝 • Modeling • 결과 해석
수상	• 장려상 (7위)

7. 빅데이터 콘테스트 공모전 (수행 과정)

공모전 수행과정



7. 빅데이터 콘테스트 공모전 (결과)

정량적인 결과

	RMSE		
모델명	타율	방어율	승률
RF	0.016	0.918	0.098
Ada Boosting	0.0182	0.926	0.104
Gradient Boosting	0.0184	0.938	0.104
LGBoosting	0.0180	0.949	0.098
XGBoosting	0.0110	0.989	0.107

7. 빅데이터 콘테스트 공모전 (결과)

결과 활용 예시 (방어율)

모형별로 변수 중요도를 추출하고 그 순위를 변수에 따라 합한 결과

변수명	Random Forest	AdaBoost	GradientBoost	LightGBM	XGBoost	순위 합
홈런 개수	0	0	0	1	2	3
공인구 여부	2	1	2	8	0	13
피장타율	1	2	7	3	1	14
FIP	3	3	1	11	3	21
WPA	5	6	8	0	5	24
3루타 개수	9	5	3	5	8	30
삼진 개수	11	4	9	2	9	35
피OPS	4	11	15	6	4	40
볼넷 개수	12	7	13	4	11	47

해석 예시

- 투수의 퍼포먼스를 향상시키는 가장 큰 키워드는 '장타억제'로 볼 수 있다. 경기당 피홈런, 피장타율 등이 평균자책점에 큰 영향을 주는 것으로 분석되기 때문이다.
- 보통 야구에서는 투수에게 타자와의 승부에서 도망가지 말고 정면승부할 것을 최고의 덕목으로 여기지만, 분석 결과 볼넷이나 삼진 등이 큰 영향을 미치지 않는 것으로 보아 강타자에게는 볼넷을 내주더라도 유인구 위주의 까다로운 승부를 해야 한다는 것을 알 수 있다.

7. 빅데이터 콘테스트 공모전 (Lessons learned & 한계점)

Lessons learned

- **기술적인 측면**
  - ML을 이용한 예측 Modeling
  - 베이지안 최적화를 이용한 파라미터 튜닝
- **업무 수행적인 측면**
  - 팀원들에게 각자 역할을 적절히 배분함

    → 최신 야구 트렌드 조사, 변수 생성, 모델링, 결과 해석 등의 업무를 적절히 분담

한계점

- **기술적인 측면**
  - 다양한 모형 탐구가 부족하였음

    → 시계열 데이터에 맞는 모형 개발을 시도하지 못한 점
  - 분석 결과를 활용할 수 있는 Platform 개발을 하지 못했음

Chapter 8

Latent Dirichlet Allocation 연구

상세 내용은 여기를 통해 확인 가능

8. Latent Dirichlet Allocation 연구 (개요)

개발 역량

알고리즘 연구/구현

- ML을 이용한 문제 해결

연구 개요

연구명	**Confirmatory Latent Dirichlet Allocation**
주제	**LDA의 텍스트 요약 기능 극대화 (Short text 대상)**
문제 정의	• 기존 LDA의 단어-주제 분포가 uncoherent하게 추론되는 경우는 종종 발생 • 기존 LDA는 선행 지식 (Prior Knowledge)을 통합하기 어려움
문제 해결	• Short text에 맞는 Variational Inference 개발 • 주제별로 연관되는 단어를 Seed words의 형태로 모형에 통합
맡은 역할	• 선행 연구 Follow-up • 알고리즘 구현 / 시뮬레이션

8. Latent Dirichlet Allocation 연구 (내용)

CLDA 알고리즘

Algorithm 2: Variational Inference for CLDA

Input Document-term matrix $X_{i,j}$; Number of latent topics K; Seed words for each topic S_k

Output Document topic distribution ϕ_{dk}; Topic word distribution λ_{kw}

1. Setting prior

for $k \in \{1, \cdots, K\}$ **do**
  for $w \in \{1, \cdots, V\}$ **do**
    if $w \in S_k$ **then**
      $\eta_{kw} \propto A_{kw}$
    end
    else if $w \in S_{k' \in \{1,\cdots,K\} \setminus k}$ **then**
      $\eta_{kw} \propto \frac{1}{A_{kw}}$
    end
    else
      $\eta_{kw} = \alpha$
    end
  end
end

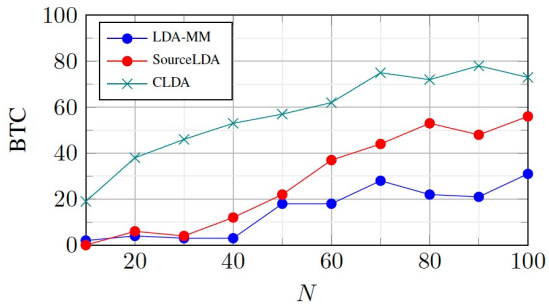
2. Inferring Variational Parameters

while *Relative change in* $\mathcal{L} > 0.0001$ **do**
  E-Step
  for $d \leftarrow 1$ **to** D **do**
    $\phi_{dk} \propto \exp\{E[\log \theta_k] + \sum_w n_{dw} E[\log \beta_{kw}]\}$
  end
  M-Step
  $\gamma_k = \alpha + \sum_d \phi_{dk}$
  $\lambda_{kw} = \eta_{kw} + \sum_d n_{dw} \phi_{dk}$
end

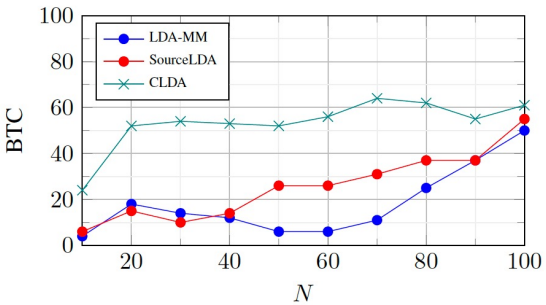
1단계 (Seed words Setting)	▪ 요약하고자 하는 주제 선정 ▪ 주제를 대표하는 단어들 (Seed words) 선정 → Word Embedding Vector 활용
2단계 (Prior Setting)	▪ 주제별 Seed words에 따라서 Prior을 다르게 세팅 ▪ 주제별로 연관된 단어가 다를 것이라는 선행 지식을 모형에 통합시키는 과정 ▪ Seed word는 하나의 주제를 대표하는 것으로 정했기 때문에 Prior을 세팅할 때도 이를 반영해야 함
3단계 (Variational Inference)	▪ 2단계에서 세팅한 Prior을 기반으로 추론 진행 ▪ Gibbs Sampling이 아닌, Variational Inference 수행 → 근사 추론이지만, 추론 시간이 월등히 빠름

8. Latent Dirichlet Allocation 연구 (내용)

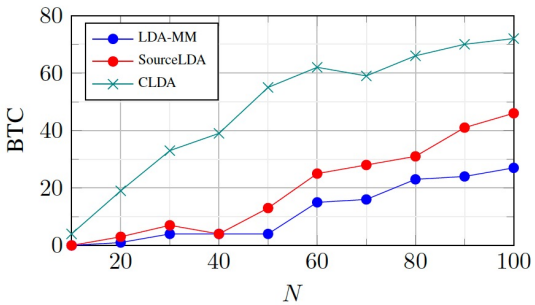
실험 결과



(a) Hotel Data



(b) Restaurant Data



(c) Car Data

Fig. 2: (a),(b),(c) - BTC results in each dataset

실험 과정	
	1. CLDA와 결과를 비교하기 위한 알고리즘으로 LDA, SourceLDA을 선택 2. 주제의 Coherence을 반영할 수 있는 Measure을 개발 → Between Topic Coherence 3. 해당 Measure을 통해 세 알고리즘을 비교해봤을 때, CLDA가 우세한 결과를 보임을 확인하였음 → 왼쪽 그림의 BTC가 높을 수록 주제가 coherent하다는 의미이고 CLDA가 두 알고리즘보다 높은 BTC을 가짐

8. Latent Dirichlet Allocation 연구 (Lessons learned & 한계점)

Lessons learned

- **연구적인 측면**
  - 기존 연구에 대한 Study, 구현, 비판적인 사고를 기를 수 있었음
- **기술적인 측면**
  - 연구를 통해 베이지안 계층 모형을 깊게 알게 되었음

    → 확장성이 높은 계층 모형을 이용하여 기존 연구를 발전시킬 계획

한계점

- **기술적인 측면**
  - Seed words와 Prior을 정하는 방식을 고도화할 필요가 있음
  - 좀 더 다양한 선행 지식을 통합할 필요가 있음

    → 단어들 간에 독립 가정을 Relax하는 word embedding vector 가정의 통합 고려

Chapter 9

측정 오차 모형 연구

상세 내용은 여기를 통해 확인 가능

9. 측정 오차 모형 연구 (개요)

개발 역량

알고리즘 연구/구현

- ML을 이용한 문제 해결

연구 개요

연구명	**Measurement Error Model with Two Proxies**
주제	**오차와 함께 관측되는 변수가 2개 존재할 때의 회귀분석에 대한 통계적 추론**
문제 정의	• 오차와 함께 관측되는 변수 (Proxy)를 통계적 조정 없이 독립변수로 사용하면 편향 (Bias)가 발생
문제 해결	• 두 개의 Proxies를 이용하여 회귀분석을 진행할 때, 베이지안 추론을 사용함
맡은 역할	• 선행 연구 Follow-up • 알고리즘 구현 / 시뮬레이션

9. 측정 오차 모형 연구 (내용)

측정 오차 모형 알고리즘

Model

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x + \epsilon \\
w &= x + v \\
z &= \gamma_0 + \gamma_1 x + \delta \\
\epsilon &\sim \mathrm{N}(0, \sigma_\epsilon^2) \\
x &\sim \mathrm{N}(\mu_x, \sigma_x^2) \\
v &\sim \mathrm{N}(0, \sigma_v^2) \\
\delta &\sim \mathrm{N}(0, \sigma_\delta^2)
\end{aligned}
$$

Prior

$$
\begin{aligned}
(\beta_0, \beta_1)^T &\sim \mathrm{N}_2(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \\
(\gamma_0, \gamma_1)^T &\sim \mathrm{N}_2(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}) \\
\sigma_\epsilon^2 &\sim \text{Inv-Gamma}(A_\epsilon, B_\epsilon) \\
\mu_x &\sim \mathrm{N}(0, \sigma_{\mu_x}^2) \\
\sigma_x^2 &\sim \text{Inv-Gamma}(A_x, B_x) \\
\sigma_v^2 &\sim \text{Inv-Gamma}(A_v, B_v) \\
\sigma_\delta^2 &\sim \text{Inv-Gamma}(A_\delta, B_\delta)
\end{aligned}
$$

Full Conditional Distribution

$$
\begin{aligned}
\boldsymbol{\beta} \mid rest &\sim N\left(\left(\mathbf{X}^T\mathbf{X}\sigma_\epsilon^{-2} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\mathbf{y}^T\mathbf{X}\sigma_\epsilon^2, \left(\mathbf{X}^T\mathbf{X}\sigma_\epsilon^{-2} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\right) \\
\boldsymbol{\gamma} \mid rest &\sim N\left(\left(\mathbf{X}^T\mathbf{X}\sigma_\delta^{-2} + \sigma_\gamma^{-2}\mathbf{I}\right)^{-1}\mathbf{z}^T\mathbf{X}\sigma_\delta^2, \left(\mathbf{X}^T\mathbf{X}\sigma_\delta^{-2} + \sigma_\gamma^{-2}\mathbf{I}\right)^{-1}\right) \\
\sigma_\epsilon^2 \mid rest &\sim IG\left(A_\epsilon + n/2, B_\epsilon + \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2\right) \\
\mu_x \mid rest &\sim N\left(\frac{\mathbf{1^T x}/\sigma_\mathbf{x}^\mathbf{2}}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}, \frac{1}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}\right) \\
\sigma_x^2 \mid rest &\sim IG\left(A_x + \frac{n}{2}, B_x + \frac{1}{2}||\mathbf{x} - \mu_x\mathbf{1}||^2\right) \\
\sigma_v^2 \mid rest &\sim IG\left(A_v + \frac{n}{2}, B_v + \frac{1}{2}||\mathbf{w} - \mathbf{x}||^2\right) \\
\sigma_\delta^2 \mid rest &\sim IG\left(A_\delta + \frac{n}{2}, B_\delta + \frac{1}{2}||\mathbf{z} - \mathbf{X}\boldsymbol{\gamma}||^2\right) \\
x_i \mid rest &\sim N\left(\frac{\beta_1(y_i - \beta_0)/\sigma_\epsilon^2 + w_i/\sigma_v^2 + \mu_x/\sigma_x^2}{\beta_1^2/\sigma_\epsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}, \frac{1}{\beta_1^2/\sigma_\epsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}\right)
\end{aligned}
$$

Gibbs Sampling 사용

9. 측정 오차 모형 연구 (내용)

실험 결과

Data Generation

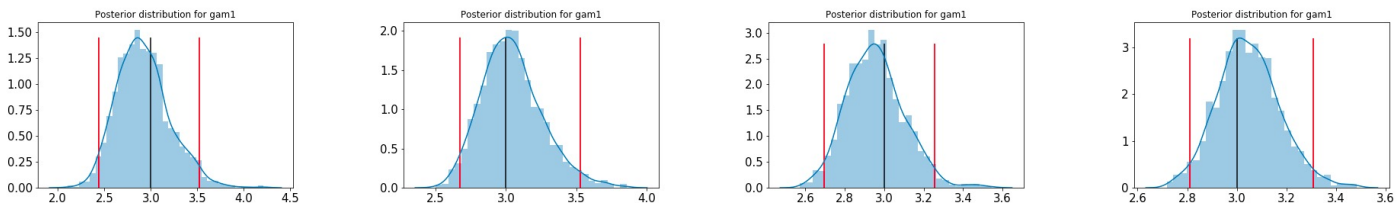
$$
\begin{aligned}
y_i &= -1 + 2x_i + \epsilon_i \\
w_i &= x_i + v_i \\
z_i &= -1 + 3x_i + d_i \\
x_i &\sim N(1/2, 1) \\
v_i &\sim N(0, 1) \\
d_i &\sim N(0, 1) \\
\epsilon_i &\sim N(0, 1)
\end{aligned}
$$

실험 과정

1. 모수를 임의로 정하고 이 모수를 적절히 추론할 수 있는지 Simulation을 수행
2. 각 모수에 대한 사후 분포를 구하고 95% credible interval이 true value를 포함하는지 확인
3. 1~2 과정을 다양한 sample size에 대해 수행

결과 예시

Results with γ_1



(a) (b) (c) (d)

Figure 4: (a) $n = 120$ (b) $n = 300$ (c) $n = 500$ (d) $n = 1000$

- 빨간색 수직선
  → 95% credible interval
- 검정색 수직선
  → True Parameter
- 모수에 대한 사후 분포의 95% credible interval이 True value을 적절히 포함하는지 확인하였다.

9. 측정 오차 모형 연구 (Lessons learned & 한계점)

Lessons learned

- **연구적인 측면**
  - 기존 연구에 대한 Study, 구현, 비판적인 사고를 기를 수 있었음
- **기술적인 측면**
  - 연구를 통해 데이터를 바라보는 새로운 시각을 습득

    → 데이터를 무조건적으로 받아들이는 것이 아니라 오차 존재 여부에 대한 비판적 사고

한계점

- **기술적인 측면**
  - 빠른 추론을 위해 Variational Inference 추론을 개발할 필요가 있음
  - 좀 더 다양한 Error model을 개발할 필요가 있음

    → 선형 모형이 아닌, Nonparametric 모형 등