4조_빅데이터엔지니어 25기

검토 준비 완료 | 공유 ••• 이 스페이스에서 별표 표시 취소

> “뛰어난 팀은 서로 감추지 않습니다. 치부를 드러내길 꺼리지 않습니다. 비난을 두려워하지 않고 자신의 실수, 약점, 걱정을 인정합니다.” — 미국 작가 Patrick Lencioni

프로젝트 주요 일정

2023.6/12 ~ 6/21

1. 6/12 ~ 6/14 : 매일 2시간
2. 6/15 ~ 6/21 프로젝트 몰입

- 컨플루언스에서 제공하는 로드맵 위젯이 일 단위 VIEW 지원이 안되니, 일자를 표기한 라인을 참조할 것!

extension

미션

해당 프로젝트의 궁극적인 목적이 무엇인지, 무얼 위해 하는지 항상 상기해야 방향을 잃지 않을 수 있습니다.

1. 이후 취업을 위한 면접 과정에서 본인이 맡은 프로젝트 내용을 어필할 수 있어야 합니다.

목표

1. **프로젝트의 성과 및 결과를 구체적으로 정의하고 Task를 정의합니다.**
   a. 정의한 문제와 해결하고자 하는 방향을 최대한 구체적으로 정리해야 합니다. (이 정도까지 정리한다고?) 할 정도로.. 그렇지 않으면 진행 과정에서 방향을 잃기 쉽습니다. 구체적인 성능기준이 필요하고 해당 기준이 곧 평가 시의 체크 항목으로 활용해서 객관적으로 프로젝트 결과를 평가해야 합니다.
   b. 반드시 성공할 필요는 없다고 생각합니다. 오히려 기한 내에 실패했다면 해당 원인에 대해 명확하게 분석하고 문서화가 필요합니다. 현업에서는 오히려 실패에 대한 회고를 자주 했던 기억이 있네요.
2. **정의한 Task 기준으로 일정에 맞추어 프로젝트를 진행합니다.**
   a. 계획하는 과정에서 이슈가 생기거나 생길 것 같다면 팀원에게 미리 공유합시다.
   b. 문제를 들고 있는 것 보다 신속하게 그리고 정확하게 공유하는 게 프로젝트 진행에 상당히 중요한 태도입니다.
3. **2023.06.21(수) 프로젝트 마감 기한 전날까지 모든 과정을 완료하고, 발표할 수 있어야 합니다.**
   a. 발표 자료는 곧 팀원들이 각 블로그(깃헙)에 정리할 아티클이 됩니다.

주요 인원

팀원	자기소개	MBTI
김지현	• 야구 경기(하이라이트 직관 노상관) 보는거 좋아합니당 • 나중에 한 줄 더 써 볼게요..!	ISFJ
이상훈	• 2년차 유부남입니다. • 경험이나 지식이 없지만 할 수 있는한 최대로 해보겠습니다!	ISFP or ENFP
정은아	• 도파민 중독 스포츠팬! 여자배구랑 야구 그리고 농구 • 추구미는 재미와 적당한 시련입니다.	INTP
이영직	• 취미는 농구 테니스 small talk • 빅데이터는 한글 처리 작업 해봤습니다 • 24일 빅분기 시험을 앞두고 있습니다	ENFP
김범기	• 유부남 경력 1년 차입니다. • 취미로 커피 하는 걸 좋아합니다:)	INFJ

액션 아이템

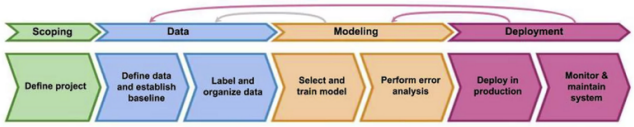
모든 회의 이후에는 회의록 및 액션 아이템이 정의되어야 합니다.

extension

b. 발표 목차를 정하고 내용을 정리해야 합니다.
c. 마지막 날에 발표내용을 정리하는 것 보다는 조금씩이라도 매일 진행 과정을 정리해야 합니다.
d. 발표자를 정합시다.
    i. 발표자는 꼭 1명이 아니어도 괜찮지 않을지?

머신러닝 프로젝트 프로세스

여기를 클릭하여 펼치기...



(2) 머신러닝 프로젝트: 기획부터 배포까지

1. **프로젝트 정의(문제 정의)**
    a. 어떤 문제를 해결하고자 하는 것인가?
    b. 프로젝트를 통해 얻고자 하는 결과값은 무엇인가? (Y값)
    c. 현재 보유하고 있는 데이터가 무엇인가?(X 값)
    d. 성과지표를 설정한다.
        - Accuracy: 음성 인식 정확도가 어느정도여야 하는지?
        - Latency: 음성 인식을 하는데 시간이 얼마나 걸리는지?
        - Throughput: 초당 쿼리 처리수은 얼마나 되는지?
2. **데이터 정의**
    a. 보유 데이터(X)의 데이터 타입?
    b. 예측을 위한 데이터의 양은 충분한지?
    c. 데이터를 신뢰할 수 있는가?
    d. 데이터 레이블이 일정한가?
    e. 데이터 표준화에 대한 방법 고려
    f. 라벨링을 한다면 어떻게 할 것인가?
3. **베이스라인 설정**
    a. 기존방식은 성능이 어떤가?
    b. 기존보다 대략 얼마정도 성능을 개선할 것인가?
4. **라벨링 & 데이터 정제하는 단계**
    a. 수많은 데이터 중, 중요한 데이터는 무엇인가?
    b. 중요한 데이터가 있다면 그 이유는 무엇인가?
    c. 데이터 수집 가이드라인을 어떻게 세울것인가?
    d. 데이터에 라벨링 할 필요가 있다면, 어떻게 할 것인가?
5. **모델링**
    a. Train 모델 선정
        i. 어떤 모델을 선정할 것인가?
        ii. 선정한 이유는 무엇인가?
    b. 오류 원인 분석
        i. 성능이 베이스라인(Baseline) 혹은 목표치(Goal)를 달성할 만큼 우수한가?
        ii. 데이터 수가 충분하여 신뢰할만 한가? (부족하다고 판단될경우 전단계로 돌아간다)
6. **배포**
    a. 제품 배포
    b. 모니터링 & 시스템 유지

회의록

날짜	참석자	논의 사항	결정 사항	액션 아이템	비고
2023년 6월 12일	1. 김지현 2. 정은아 3. 이영직 4. 김범기 5. 이상훈	1. **조장 결정 → 이상훈** a. 발표는 별도 인원 지정 가능 2. **최종 결과물에 대한 주요 형식 정의** a. 참고 프로젝트 i. nlp_hate_speech/README.md at master · hayoon/nlp_hate_speech ii. 이루오_포트폴리오 3. **주제 논의 회의 6/13** a. 각자 하고 싶은 주제(약 2~3개) 선정 i. 정은아 : 교육 플랫폼 관련 주제 혹은 금융 쪽 공모전을 참고하는 방향 제시 ii. 이영직 : 금융 쪽 주제에 선호 iii. 김범기 : GCP나 AWS와 같은 클라우드 플랫폼 활용하여 데이터 파이프라인 구축 선호 b. 데이터 수집할지? i. 어디서 수집할 수 있는지? c. 사례 i. 실제 프로젝트 사례가 있는지? 4. **기타** a. 5명이서 분석 프로젝트를… 어떻게 분배할지? i. 시각화 ii. 담당자별 담당 모델링, 분석 결과 정리	**프로젝트 주제 후보** 1. **교육플랫폼 강의 추천 시스템(인프런,유데미 등)** a. 구글트렌드 등 이용 b. 깃(언어 점유율), 유튜브 강의 등/ 사람인 등 구직사이트 분석/현업자 기술블로그에서 키워드 크롤링 2. **환자정보와 기침소리를 이용한 covid-19 진단 모델** a. [토이프로젝트] 환자정보와 기침소리를 이용한 코로나 감염여부 판단 모델 (정형데이터 + 오디오 멀티모달 실습 코드, 머신러닝 딥러닝 프로젝트 주제)	☑ 6/13_프로젝트 주제 후보에 대한 예상 진행 과정 및 결과에 대한 조사 ☑ 6/13_ 프로젝트 주제 후보 중 주제 선정	
2023년 6월 13일	1. 김지현 2. 정은아 3. 이영직 4. 김범기 5. 이상훈	주제 브레인스토밍 1. 상훈 a. 검색을 토대로 한 자살확률에 대한 예측 모델 b. 장바구니 담겨있는 상품 중, 구매확률 예측 모델 2. 상품 묶음에 따른 판매확률 예측? a. 연관 분석	선정 주제 1. Coupon Purchase Prediction • Predict which coupons a customer will buy • k Coupon Purchase Prediction	☑ 테이블 스키마 및 컬럼 분석 ☑ 6/13_프로젝트 과정에 대한 역할 분담 ☑ 6/13_ 프로젝트 마일스톤 정리	k An EDA of coupon's transaction (in Vietnamese)
2023년 6월 14일	1. 김지현 2. 정은아 3. 이영직 4. 김범기 5. 이상훈	• EDA 목록 정하기 ◦ 이상치, 결측치 확인 ◦ 데이터 특징 및 패턴 파악 ◦ 데이터 시각화 ▪ 히트맵 ▪ 산점도 ▪ 히스토그램 • 전처리 ◦ 컬럼 내의 다중 값 추리기? ◦ NULL 값 변환? , 평균값? 중앙값? ◦ 전처리 할 목록 정하기- 일본어 전처리 - 이해할 수 있는 값으로 (영어 or 한국어) ◦ hash-id 필터링 ▪ hash 로 둘지? 이진값으로 변환할지? ◦ 테이블 별 변수 타입별 분류 ▪ 타입별 전처리 단계에서 방식을 구분하는 접근? ◦ 테이블 조인(어떤 테이블을 어떻게 조인할지) • 피처링 ◦ 모델에 넣을 피처, 라벨 정하기 ◦ 타겟(=라벨) 선정? ▪ 쿠폰 사용 카테고리 ▪ 쿠폰 카테고리 ▪ 쿠폰 할인 라격 ▪ 쿠폰 할인률 ▪ 쿠폰 사용 가능 일수 ▪ 홀리데이 사용 가능 여부 • 머신러닝 모델 선정 ◦ (딥러닝) 사용해서 모델 만들어라 ▪ 예)합성곱… ◦ • 모델 평가 • 모델 튜닝 및 개선 • 결과 해석	1. 피쳐 선정 진행 중 2. ML 모델 검토 진행 중 3. 빅쿼리 환경 세팅 완료 1. 전처리 수행 2. 훈련, 테스트 데이터 분리 3. 훈련 시작 …. .	☑ 6/13_프로젝트 작업 리스트업 @BK ☑ 6/14_VertexAI 환경 세팅 @BK ☑ 6/14_프로젝트 과정에 대한 역할 분담 ☑ 6/14_ 프로젝트 마일스톤 정리	
2023년 6월 15일		**TO_DO_List** 1. **분석 주제(가설) 구체적으로 설정** 분석의 목표/구체적으로 적용해볼 모델 종류/해당 모델에 넣은 피처(input)데이터 설정/모델을 통해 도출할 예측값(target)설정/모델 성능 평가 지표 설정(ex. AUC, confusion matrix etc…) • 캐글 컴페티션에서 요구 모델1 - 코사인 유사도 → 이론 및 코드 해석이 어려워 후순위로 • 캐글에서 요구 모델2 - **로지스틱**으로 구현 + 하이퍼 파라미터 튜닝 추가 - For 구매여부 예측		☐ coupon_list 테이블 usable_date ~ (na,0,1,2) 값 의미 파악_ @정은아 ☐ 각 맡은 테이블별 EDA 결과 6/16 공유 ☐ @지현 - Cupon_area, Cupon_list_ @정은아 ☑ @훈씨네마 - Coupon_detail ☑ @이영직 - Coupon_visit	

		• 캐글에서 요구 모델3 - 적용 가능한 다른 모델이 있으면 구현 → 쿠폰 추천 모델 - 분류**(RF, XGboost)** → 딥러닝**(후순위)** • 캐글에서 요구한 것 이외의 목표 설정 가능 ->**2. EDA 논의** • NA - 그래프 → NA값이 얼마나 있는지 확인하고 전처리 1의 과정이 얼추 정리 되었다면 해당 모델에 쓰일 유효한 컬럼 데이터를 중심으로한 시각화 진행(예를 들어 모델에 나이, 성별, 지역 데이터를 쓰기로 했다면 해당 데이터를 시각화) • 시각화 방법 - 사용할 그래프 종류		☑ @BK - User_list	
6월 16일				**coupon_visit -** @이영직 ☑ PURCHASE_FLG - 0,1 pie chart ☑ I_DATE - YEAR + MONTH별 쿠폰사용 ☑ I_DATE - DAY에 따른 쿠폰사용 ☑ I_DATE -시간에 따른 쿠폰사용 ☑ TOP10 쿠폰 **coupon_detail -** @혼씨네마 ☑ ITEM_COUNT - bar ☑ I_DATE - yy.mm - coupon_visit의 I_DATE와 함께 그리기 ☑ SMALL_AREA_NAME - 구매량 많은 상위 n개 ☑ USER_ID_hash 그룹바이 썸 ITEM_COUNT ☑ COUPON_ID_hash에서 발생 빈도 데이터 - 많이 사용된 상위 20개 쿠폰 **coupon_area -** @지현 ☑ SMALL_AREA_NAME - count ☑ PREF_NAME - 지역 **coupon_list - ALL** ☑ CAPSULE_TEXT ☑ GENRE_NAME ☑ PRICE_RATE ☑ DISPPERIOD - bar ☑ VALIDPERIOD - hist ☑ USABLE_DATE_요일 컬럼 모두 시각화 - 같은 fig ☑ USABLE_DATE_홀리데이 ☑ large_area_name ☑ ken_name ☑ small_area_name **user_list -** @BK ☑ EDA완료 ☐ 코사인 유사도 코드 분석 **오후** • EDA 얼추 마감 • 피처값 전처리(모델에 어떤 형태로 넣을 것인지 논의 및 전처리 수행)	
6월 17-18일 주말		• 모델링을 위한 분담 사항 - 피처 전처리 • 사용해볼 모델링 학습해보기 - (로직스틱, RF, XGboost) • EDA 그래프 보강		**1순위** - 전처리 맡은 것 처리(번역, 스케일링/정규화, 오버샘플링) ☑ PREF,SMALL,CAPSULE_TEXT/GENRE_NAME 번역_ @BK → 빅쿼리 테이블 업데이트 ☑ 새컬럼 추가) USERBLE SUM_ @BK → 빅쿼리 테이블 업데이트(SQL-저장) ☑ 성별 get_dummy 완료	

				☑ 새컬럼 추가) 유저의 뷰잉 횟수 @훈씨네마 → temp테이블에 추가/ 파이썬에서 코드로 구현 ☐ **범주형 get_dummy 처리** @정은아 • 거주 지역(PREF_NAME) • 쿠폰 종류 - CAPSULE_TEXT/GENRE_NAME • 쿠폰 발급 기간(월여부/이진) - DISPFROM/DISPEND - 범주 - date - get_dummy • 쿠폰 사용 가능 요일(count) - Usable_Date_sum • 쿠폰 사용 가능 지역 - small_area_name ☐ 나이 AGE -int(정규화/스케일링) - @이영직 ☐ 쿠폰 발급 기간 (일수) -DISPPERIOD - 연속형(필요시, 정규화/스케일링) @이영직 ☐ 구매여부 (PURCHASE_FLG)- 오버샘플링 @지현 **2순위** - 각자 EDA 보강 - 그래프 더 보기 편하게 수정(색상추가, 글자 깨짐 겹게 수정(색상추가, 글자 깨짐 겹침 수정) **3순위** - 적용할 모델링 학습해오기 • 로지스틱, 랜포, xgboost • (후순위/필수x) 코사인 유사도(추천시스템), 딥러닝(합성곱)	
6월 19일(월)					
6월 20일(화)					
6월 21일(수)					

2023년 6월 22일		금일 논의할 내용 요약
		발표자 방식 및 목차 정의
		• 발표시간 (총 20분)
		• ML 분석(?분)_'발표자?'
		  ◦ 로지스틱 회기
		  ◦ 랜덤 포레스트
		  ◦ XG Boost / LightGBM
		• 데이터 파이프라인(?분) _'발표자?'
		  ◦ 데이터 적재
		  ▪ bigquery
		  ▪ 구글 클라우드 스토리지
		  ◦ 파이프라인
		  ▪ vertex ai
		• Q&A (10분)
		  ◦ 모델 과정에 대한 질문
		  ◦ 랜덤 포레스트 - 이영직
		  ◦ XG Boost / LightGBM -김지현
		  ◦ 로지스틱 회기 - 이상훈
		  ◦ 딥러닝 - 정은아
		• 데이터 파이프라인 질문
		  ◦ 김범기
			06/22
			
			1. 14:00 ~ 16:50 : 발표 자료 구성
			2. 17:00 : 각 담당한 발표 목차 팀 내에 공유
			• 보완이 필요한 내용은 각자 집에서 추가 작업 후 06/23(금) 오전에 재공유
			06/23
			1. 09:00 ~ 10:00 : 보완된 내용 공유 및 파악
			2. 10:00 ~ 11:00 : 추가 수정
			3. 11:00 ~발표 및 예상 질문 정리
				1. 발표자료 완성
				미니프로젝트_4조_발표자료	

2023년 6월 23일		• 최종 발표			

참조

분류	요약	링크
공모전	신한AI, 보다 나은 금융 생활을 위한 AI 서비스 아이디어 경진	TopGun - 🚀🌏신한AI 글로벌 투자 원더랜드 🎩 - 마법같은 해외 뉴스 분석 주식 챗봇 🤖
프로젝트 레퍼런스	머신러닝 프로젝트 사이클	(2) 머신러닝 프로젝트: 기획부터 배포까지
프로젝트 레퍼런스	빅데이터 프로젝트 주제	내가 하려고 정리한 빅데이터 프로젝트 주제 (= 데이터분석 프로젝트 주제 = 머신러닝 프로젝트 주제)
프로젝트 레퍼런스	멀티모달 프로젝트	[토이프로젝트] 환자정보와 기침소리를 이용한 코로나 감염여부 판단 모델 (정형데이터 + 오디오 멀티모달 실습 코드, 머신러닝 딥러닝 프로젝트 주제)
프로젝트 레퍼런스	대기업 현직 데이터 분석가가 추천하는 데이터 분석 프로젝트 주제 + 활용 데이터셋	대기업 현직 데이터 분석가가 추천하는 데이터 분석 프로젝트 주제 | zero-base
데이터셋	음향데이터 COVID-19 검출	월간 데이콘 음향 데이터 COVID-19 검출 AI 경진대회
파이프라인	머신러닝 파이프라인 이해	[Machine Learning] 머신러닝 파이프라인의 이해 (MLOps) <인프런 - 머신러닝 엔지니어 실무>

2019년도 구축 빅데이터 플랫폼

- 통신 빅데이터 플랫폼: bdp.kt.co.kr
- 교통 빅데이터 플랫폼: 국가교통 데이터 오픈마켓
- 문화 빅데이터 플랫폼: init-page
- 환경 빅데이터 플랫폼: 환경 빅데이터 플랫폼
- 중소기업 빅데이터 플랫폼: datastore.wehago.com
- 지역경제 빅데이터 플랫폼: ggdata.kr
- 금융 빅데이터 플랫폼: fnbigdata.com
- 헬스케어 빅데이터 플랫폼: cancerportal.kr
- 유통소비 빅데이터 플랫폼: kdx.kr
- 산림 빅데이터 플랫폼: forestdata.kr–

2020년도 구축 빅데이터 플랫폼

- 소방안전 빅데이터 플랫폼: 소방안전 빅데이터 플랫폼
- 스마트치안 빅데이터 플랫폼: 스마트 치안 빅데이터 플랫폼
- 해양수산 빅데이터 플랫폼: 해양수산빅데이터 거래소
- 농식품 빅데이터 플랫폼: https://kadx.co.kr/
- 라이프로그 빅데이터 플랫폼: https://www.bigdata-lifelog.kr/portal
- 디지털 산업혁신 빅데이터 플랫폼: 디지털산업혁신 거래소

GCP 관련 링크

1. 딥러닝 프로젝트를 위한 클라우드 GPU 자원, Google Cloud Platform
2. 모듈 4: GCP를 활용한 머신러닝
3. GCP AI-Platform에서 딥러닝 학습하기
4. Google Cloud Big Data and Machine Learning Fundamentals 한국어
5. Google Cloud에서 머신러닝 구현을 위한 권장사항 | 클라우드 아키텍처 센터
    a. GCP