

2017

빅데이터 아카데미 우수 프로젝트 사례

**2017년,
우리가
빅데이터를
이끌다**

2017
빅데이터 아카데미 우수 프로젝트 사례



**2017년,
우리가
빅데이터를
이끌다**



과학기술정보통신부



한국데이터진흥원



BIGDATA ACADEMY
빅데이터아카데미

CONTENTS

- 3** 데이터 시대를 앞서 준비하는 빅데이터 아카데미

- 6** **빅데이터 분석 전문가 18기**
자동차 엔진오일 교환유형 분석을 통한 고객 이탈방지 및 서비스 개선

- 16** **빅데이터 분석 전문가 19기**
기계학습 방법을 활용한 신도시 아파트 가격변동 요인 분석

- 24** **빅데이터 분석 전문가 20기**
딥러닝 방법을 이용한 유방암 메디컬 이미지 분류·예측 모형화

- 34** **빅데이터 분석 전문가 21기**
어느 학교를 가야 할까요?

- 42** **빅데이터 기술 전문가 13기**
기상센서 데이터를 이용한 기후 정보 및 이상센서 알림 서비스

- 52** **빅데이터 기술 전문가 14기**
농구 국가대표 톱5 선발

- 60** **의료 빅데이터 융합 전문가 2기**
염증성 장질환^{IBD} 환자의 결핵 발병예측 분석

- 68** **금융 빅데이터 융합 전문가 1기**
양파 생산량 예측 기반 금융상품 제안

- 78** **금융 빅데이터 융합 전문가 2기**
스타트업 기업 지표와 투자유치 연관성 분석

- 86** **유통 빅데이터 융합 전문가 1기**
구매패턴 기반 구매감소 고객 예측

- 96** **제조 빅데이터 융합 전문가 2기**
빅데이터 분석을 통한 최적 사육 환경 조성

데이터 시대를 앞서 준비하는 빅데이터 아카데미

배움의 설렘과 성장의 기쁨이 있는 곳

빅데이터 아카데미는 과학기술정보통신부와 한국데이터진흥원의 시장 수요에 대응한 빅데이터 전문가 양성 프로그램입니다. 직무별·산업별 빅데이터 전문가로서 역량을 펼칠 수 있도록 세분화한 교육 프로그램을 운영해 지난 2013년 개소 이래 2017년까지 1,570명의 수료생을 배출했습니다.

빅데이터 아카데미가 걸어온 길

- 2013년** 빅데이터 아카데미 개소 및 출범
- 2014년** 2014년 빅데이터 아카데미 사례 발표회 개최
- 2015년** 빅데이터 기획 전문가 과정 신설
빅데이터 아카데미 성과 발표회 개최
- 2016년** 지역별(강원, 경기, 대전, 부산, 전북) 빅데이터 활용인력 양성과정 신설
빅데이터 융합(의료·제조·머신러닝) 전문가 과정 신설
빅데이터 아카데미 프로젝트 발표 및 시상식 개최
- 2017년** 빅데이터 융합(유통·금융·융합 기획) 전문가 과정 신설



BIG CHANCE BIG OPPORTUNITY

BIGDATA ACADEMY

빅데이터 아카데미는 지난 2013년 설립 이래 분석·기술·기획·융합·지역 전문가 등 세분화한 총 63회의 교육 과정을 개설·운영했습니다. 220건의 파일럿 프로젝트를 발굴하고 1,570명의 수료생을 배출하여 명실상부한 한국의 대표 빅데이터 전문가 양성 교육 프로그램으로 자리 잡았습니다.

빅데이터 아카데미 주요 성과

빅데이터 아카데미 교육과정 **63**회 운영

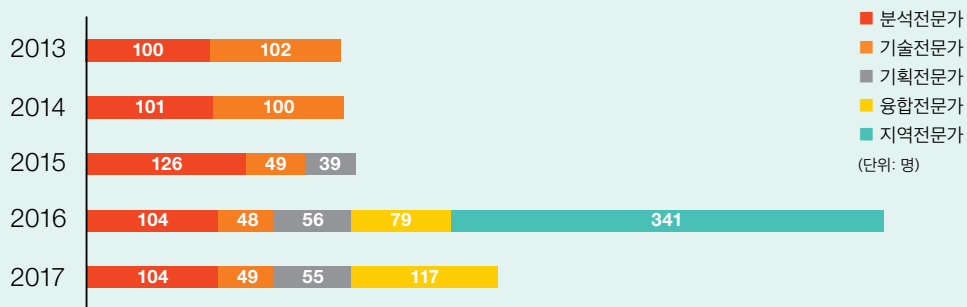


파일럿 프로젝트 **220**건 발굴

수료생 **1,570**명 배출



수료생 배출 현황



BIGDATA ACADEMY 2018

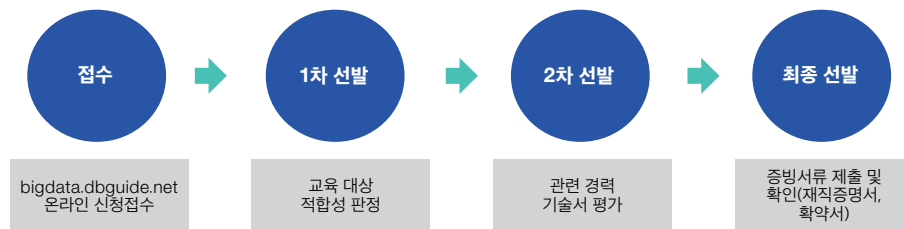
교육 과정

과정명	교육기간	과정 소개
빅데이터 기획 전문가	5주	빅데이터 프로젝트 기획에 대한 이해를 바탕으로 빅데이터 활용 기회 발굴, 사업 기획 및 관리가 가능한 빅데이터 기획 전문가 양성
빅데이터 기술 전문가	8주	빅데이터 수집·저장·처리 등 빅데이터 처리 기술에 대한 이해와 고급 기술 습득을 통한 실무 중심형 빅데이터 기술 전문가 양성
빅데이터 분석 전문가	8주	빅데이터 분석 도구, 빅데이터 분석 알고리즘 등에 대한 지식 함양을 기반으로 현업에서 빅데이터 분석을 통해 새로운 가치를 창출해낼 수 있는 빅데이터 분석 전문가 양성
빅데이터 융합 전문가	8주	산업 도메인별 빅데이터 특성을 이해하고 이를 산업현장에 융합하여 적용할 수 있는 빅데이터 융합 전문가 양성

교육 대상

구분	교육 대상
공통	빅데이터 프로젝트 수행 또는 예정인 산업계 재직자
빅데이터 기획 전문가	데이터 관련 사업 발굴, 관리 등 기획 업무 3년 이상 경력자
빅데이터 기술 전문가	개발자, DBA, SE(System Engineer) 등 IT 분야 3년 이상 경력자
빅데이터 분석 전문가	시장, 고객, 제품 등 데이터 분석 3년 이상 경력자
빅데이터 융합 전문가	산업 도메인별 데이터 분석 3년 이상 경력자

선발 프로세스



교육 단계

사전 교육(1주)	집체교육(1주~2주)	파일럿 프로젝트(2주~4주)	파일럿 프로젝트 공유(1주)
<ul style="list-style-type: none"> 선수 학습 지원 및 입과 역량 평준화 온라인 교육 온라인 교육 자료 제공 	<ul style="list-style-type: none"> 빅데이터 기획·기술·분석·융합 과정별 이론 및 실습 교육 시간: 09:30~17:30 장소: 한국데이터진흥원 	<ul style="list-style-type: none"> 현업 적용 고도화를 위한 팀별 파일럿 프로젝트 실시 시간: 매주 토요일 10:00~ 17:00 장소: 한국데이터진흥원 팀별 멘토강사 지원 	<ul style="list-style-type: none"> 팀별 프로젝트 발표·평가·결과 공유 프로젝트 결과 발표 및 공유 우수팀 선정, 시상 등

※ 상세 내용은 빅데이터 아카데미 홈페이지(<http://www.dbguide.net/bigacademy.db>)를 참조하세요

자동차 엔진오일 교환유형 분석을 통한 고객 이탈방지 및 서비스 개선



구분	빅데이터 분석
적용 도구	R studio, Rpart, randomForest, glm, nnet, h2o, PCA, K-means, GPLOT
수집 데이터	2006~2011년 사이에 출고된 차량 기준으로 2016년까지 총 11년간의 엔진오일 교환 이력 데이터
산출물	이탈 고객 예측, 타깃 마케팅 군집
지도	박원준
참여자	배창준 ^{조장} , 박정희, 안병진, 윤여갑, 이동주, 주인봉
프로젝트 소개	자동차 엔진오일의 교환 유형을 분석한 결과를 토대로 고객 이탈 예측과 방지 대책을 세울 수 있는 분석 모델을 구축했다. 타깃 마케팅을 위해 수집 데이터의 주성분을 분석해 속성을 파악하고자 했다. 군집분석으로 군집을 정의하고, 해당 군집의 특성을 시각화해 고객 특성에 맞는 마케팅 방안을 도출했다.

THE CHALLENGES

책이나 강연 등으로 접했던 빅데이터가 아니었다. 빅데이터 분석 전문가 교육 과정에 등록해 집체교육과 파일럿 프로젝트를 수행하면서 빅데이터를 몸으로 이해하게 됐다.

엔진오일 교환 패턴은 차주의 성향에 따라 매우 다양하다는 것을 알게 됐다. 그래서 이 분석의 끝을 보고자 파일럿 프로젝트의 주제를 ‘자동차 엔진오일 교환유형 분석을 통한 고객 이탈방지 및 서비스 개선’으로 정했다. 처음으로 100만 건 이상의 데이터를 R과 오라클로 다루면서 파생변수 도출을 위해 많은 시간을 보냈다.

‘할 수 있다. 지금 이 시간을 함께하고 있음에 의미를 갖자’는 문구를 새기며 분석에 들어갔다. 5가지의 분류 알고리즘의 5가지 기법(트리·랜덤포레스트·로지스틱·신경망·딥러닝)을 통해 오분류율을 구할 수 있었다. 주성분분석(PCA 분석)을 진행해 의미 있는 요인을 도출할 수 있었고, K평균 군집분석(k-means clustering)을 수행해 군집별 이탈결과를 확인할 수 있었다. 분석을 통해 얻은 결과를 스폿파이어(SpotPire)로 시각화해 결과를 더 쉽게 바라볼 수 있었다.

THE APPROACH

엔진오일 교환은 자동차 관리의 기본 사항이다. 엔진오일 교환 고객 유형 분석을 위해 △지정 네트워크에 주기적 방문 고객 △외부 정비소 방문 고객 △관리하지 않는 고객으로 구분해 특성 파악 및 이탈 고객 예측 모형을 개발하고자 했다.

데이터 수집

총 6년간의 출고 차량을 대상으로 지정 네트워크에 방문한 고객의 엔진오일 교환이력 데이터를 확보했다. 고객 정비소, 접근성, 차량정보 데이터만 수집 가능하다고 정의했다. 총 150만 건의 엔진오일 교환 이력을 수집했다. 5년 간의 데이터로 모델을 생성하고, 1년간의 데이터로 모델 검증기간을 정의했다. 개인

정보, 수리정보, 행사정보 등 추가정보를 확보할 수 없어서 제한된 데이터로 분석했다.

데이터 탐색

데이터 현황 파악 및 이상치 제거 분석집단 정의를 위해 엔진 교환 이력에 대한 시각화 분석을 했다.

엔진오일 교환 이탈률 분석

엔진오일 교환 시 직영 네트워크 이탈이 많은 집단을 개인속성, 차량속성 등으로 일차적으로 살펴보았다. 차량·차량 스타일·차량 크기별로 이탈률과 비중이 달랐으므로 사전 데이터 분석을 하여 분석 차량을 정의했다.

차령에 따른 오일 교환주기와 주행거리의 차이 분석

차령(車齡)이 증가할수록 평균 엔진오일 교환 주기가 길어지는 것으로 나타났다. 차량 이용 빈도가 낮아지면서 엔진오일 교환 주기도 길어지는 것이다. 하지만 평균 주행거리는 일정하게 유지되는 패턴을 보였다.

성별에 따른 교환 특성을 비교해 보았다. 여성 오너의 엔진오일 교환 패턴이 남성 오너에 비해 상대적으로 불규칙했다. 이는 차량에 대한 관심도가 남성에 비해 상대적으로 낮아서 발생하는 현상이 아닐까 하는 결론을 내렸다.

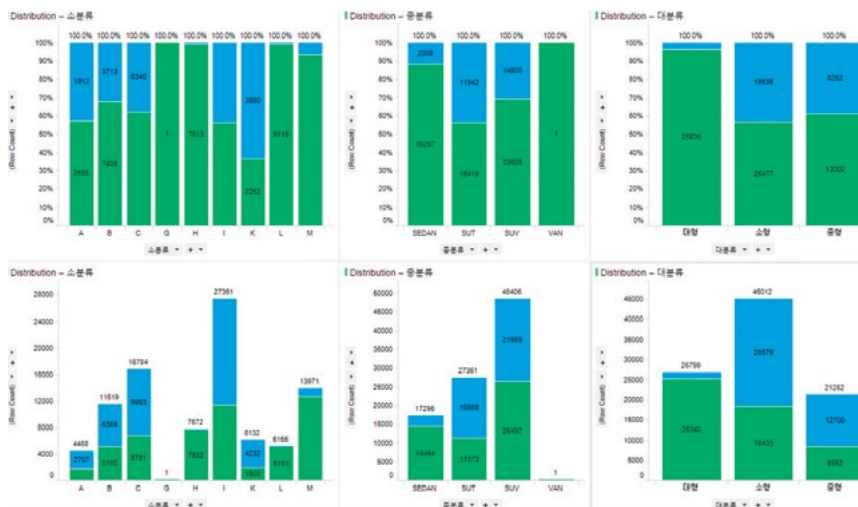


그림 1
엔진오일 교환
이탈률 분석

No.	구분	변수명	설명	예외처리
1	접근성	지정 네트워크 대비 비지정 네트워크 비중	지정 네트워크/비지정 네트워크	MISSING 처리
2	접근성	지정 네트워크 접근시간	고객 자택주소로부터 정비소까지의 시간	
3	차량정보	국토부 기준 교환유형	국토부 공시 교환주기 준수 (이상, 표준, 이하)	MISSING 처리
4	차량정보	평균 교환횟수 2년	이탈하지 않은 시점의 교환횟수 평균	연간 12회 이상 제외
5	차량정보	평균 주행거리	이탈하지 않은 시점의 주행거리 평균	50,000km 이상 제외
6	차량정보	차주 변경여부	성별 변경시 여부로 차주 변경이력을 유추	MISSING 처리

표 1
파생변수 리스트

고객 이탈에 영향을 줄 것으로 예상되는 파생변수 생성

수집된 데이터만 갖고는 이탈고객을 예측하는 데 한계가 있었다. 수집한 데이터 안에서 이탈과 관련이 높은 파생변수가 어떤 것인지를 고민했다. 외부정보 중 활용 가능한 정보를 추가로 수집해 모델의 예측력을 높이기로 했다. 교환이력 정보를 활용해 교환주기, 교환횟수 등의 변수를 생성했다. 외부정보로는 지역별(정비) 네트워크 영업소 수를 확인해 네트워크 접근성이라는 파생변수를 확보했다.

분석 목표 정의

분석 목표는 크게 이탈을 예측하는 모형을 생성하고, 세그먼트별로 고객의 특성을 정의해보는 것으로 잡았다. 이탈고객 예측을 위해 일반적으로 널리 사용되는 트리, 로지스틱, 신경망, 랜덤포레스트, 딥러닝 예측 알고리즘을 적용했다. 최근 많이 사용된 딥러닝 알고리즘을 특별히 예측 알고리즘으로 사용해 기존 머신러닝 성능과 비교했다.

타깃 마케팅 분석은 수집 데이터를 주성분 분석으로 데이터의 개념을 파악하는 것을 목표로 했다. 추가적인 비즈니스 도출을 위해 군집분석으로 특성을 분석하기로 했다. 도출된 결과는 시각화 툴을 이용하여 지역별·군집별 특성이 존재하는지 탐색하기로 했다.

예측 모델링

이탈 정의

2년 후 고객 이탈유무를 예측하는 모형을 생성하기 위해 직전 방문 이력이 18개월 이상이면서 직전 방문 시 교환거리 1만 5000킬로미터 이상이라는 두 조건을 만족하는 차량을 이탈로 정의했다. 우리나라는 도로가 좁고 짧으며, 신호가 많아서 급정거·급제동이 빈번하게 발생한다. 이에 따라 이탈을 정의하는 부분에서 어려움을 많이 겪었다. 또한 운전자의 성격에 따라 많게는 6개월에 2번 교환하는 소유주도 있었다. 반면 24개월이 지나도 한 번도 교환하지 않는 경우도 있었다. 이 부분을 어떻게 정의해야 할지 고민에 빠졌다. 하지만 데이터를 탐색하던 중, 18개월 이내에 교환 거리 1만 5000킬로미터 이내에 교환하는 경우가 약 70%를 차지하고 있음을 알게 됐다. 이에 따라 우리 조는 18개월 이상에 1만 5000킬로미터 이상인 차량들을 이탈로 정의하기로 결정했다.

데이터 기간 정의

이탈모형을 생성·테스트하기 위해 2006년에서 2010년까지 5년 간 출고된 차량들의 출고 시점으로부터 2년 도달 시 이탈 여부를 타깃으로 정의했다. 총 150만 건의 교환 이력을 차량별로 집계해서 약 20만 대의 차량정보를 얻을 수 있었다. 이탈모형 검증에 위해 테스트 기간을 2011년 1년으로 정의했다. 가장 우수한 모델로 테스트 데이터를 검증해 모델 적합도를 평가했다.

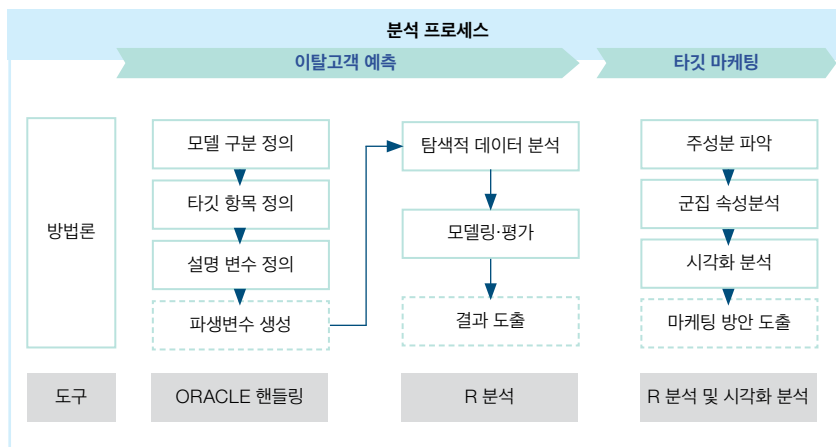


그림 2

자동차 엔진오일 교환
데이터 분석 프로세스

데이터 전처리

20대의 이탈률은 66%로 매우 높았다. 이탈모형 생성 시 적합력을 떨어트릴 수 있는 출고 후 5년 동안 4회 이내 방문 이력을 가진 차량을 제외했다. 성별·연령 등의 변수 중 오류 데이터를 제외해 총 10만 대의 차량을 분석 대상으로 정의해 이탈모형을 생성했다. 최종 분석 대상인 10만 대의 이탈률은 약 33%로 관측됐다.

모델 성능

모형 적합 결과 오분류율로 모델의 성능을 평가했으며, 리프트로 효율성을 검증했다. 그 결과 ▷딥러닝 0.18 ▷랜덤포레스트 0.20 ▷로지스틱 0.28 ▷트리 0.29 ▷신경망 0.30 순으로 나타났으며, 딥러닝을 적용했을 때 가장 뛰어난 성능을 보였다. 최근에 나온 랜덤포레스트와 딥러닝이 높은 적합도를 갖고 있음을 확인할 수 있었다. 가장 높은 성능을 보였던 딥러닝 알고리즘을 테스트 데이터에 적용해 보았다. 그 결과 오분류율은 0.25로 모형적합 결과 테스트를 했을 때보다 다소 높아졌다. 하지만 큰 차이를 보이지 않아 이탈모형을 적용하는 데 큰 문제가 없어 보였다. 리프트의 결과는 랜덤포레스트가 높은 성능을 보였다.

중요 변수

각 알고리즘을 통해 나온 중요변수를 정리했다. 교환패턴 속성과 차량속성 변수가 많이 도출됐으며, 각각 변수의 중요도는 트리분석을 통해 살펴보았다.

- 교환 패턴 속성: 교환거리, 교환횟수, 교환개월로 꾸준히 관리하는 군집의 이탈률이 낮음
- 차량 속성: 차종, 배기량 속성을 기준으로 대형 차종이 이탈률이 높음

트리모형 결과

가장 높은 영향력을 갖는 변수를 트리모형으로 살펴보았다. 트리모형의 결과는 높은 적합도를 보이지 못했다. 두 가지 변수만 도출되어, 가지치기 옵션을 조정해 추가적으로 유의한 변수를 조금 더 도출했다. 결과는 평균 교환거리, 교환횟수, 차종 등의 변수가 유의했다.

고객 분석과 시각화 분석

데이터의 속성을 파악하기 위해 PCA 요인분석을 했다. 의미 있는 요인으로서는 4~5년 교환주기, 이탈여부, 꾸준한 관리, 개인속성이라는 4개가 도출돼 데이터의 속성을 파악하는 데 도움을 얻을 수 있었다.

K평균 군집 분석의 결과로는 1)꾸준한 관리·교환거리는 증가 2)2년~5년 차에 반짝 방문 3)꾸준히 교환 안 함 4)3년부터 꾸준히 관리하는 교환패턴이 발견되었으며, 개인속성은 큰 영향을 주지 않는 것으로 분석됐다.

생성된 데이터는 최근 많은 분야에서 활용되고 있는 시각화툴인 스포트파이어를 활용했다. 30일 무료 버전을 이용해 군집별로 엔진오일 교환패턴이 지역에 따라 다른지를 살펴보았으며, 데이터의 드릴다운을 통해 추가적인 비즈니스를 도출했다.

분석결과 활용방안

연령대별 엔진오일 교환패턴을 분석해 활용하고, 차량 등록 대수가 적은 지역의 교환 패턴을 분석해 맞춤 서비스 등에 활용할 계획이다.

본 프로젝트를 통해 연령별도 마케팅 계획을 세워보았다. ~20대, ~50대, ~시니어층으로 구분하고 각각의 프로모션 활동을 정의했다. 호기심이 많은 20대에게는 마일리지별 교환쿠폰(1만 마일리지 운행기념)을, 차에 대해 자세히 모르지만 아끼는 ~50대에게는 마일리지별 자가 정비용품, 방문이 힘든 시니어

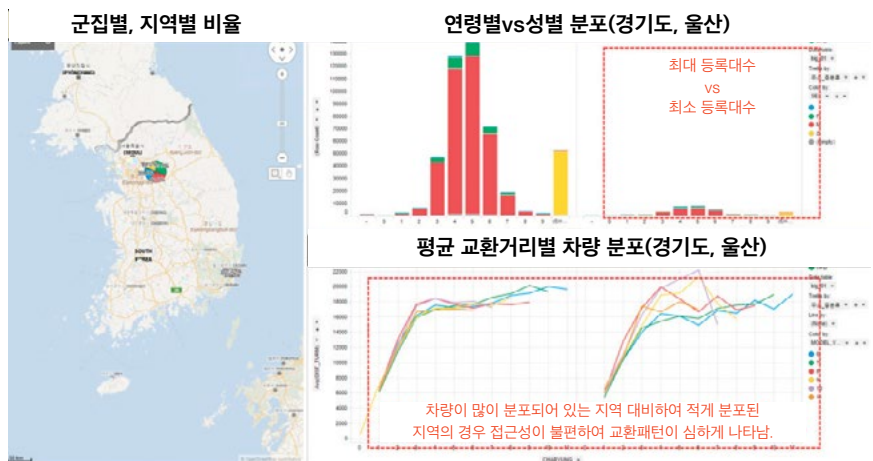


그림 3

군집별(K-Means)
차량 선호도 및
평균 교환거리

층에게는 찾아가는 서비스 또는 우선 예약제 등의 프로모션 안을 도출했다.

THE OUTCOME

2주 간의 집체교육을 받고 파일럿 프로젝트를 시작할 때 과연 우리가 잘해낼 수 있을까 하는 막연한 두려움과 설렘이 교차했다. 다행히 파일럿 프로젝트 시작 전에 많은 얘기들을 나누었고 조원들의 열정이 높아 분석 주제를 빠르게 정해서 진행할 수 있었다.

자동차 엔진오일 교환 패턴 데이터로 고객이탈을 예측하기 위해 중요변수를 도출하여 트리, 랜덤포레스트, 로지스틱, 신경망, 딥러닝의 알고리즘을 적용하고 마케팅 방안 도출을 위해 주성분분석과 K평균 군집분석 기법을 적용했다. 데이터를 분석한 결과, 고객 이탈 예측에서는 차주변경 여부와 차량 연식이 고객 이탈과 관련이 높았다. 차량 타입, 성별, 접근성에 주안점을 두고 마케팅 계획을 도출해야 할 것으로 나타났다. 특히 여성 자동차 소유주가 남성에 비해 오일교환 주기가 덜 규칙적이므로 여성 운전자에 특화된 생활 정보와 결합된 알림 서비스 등으로 타겟 마케팅 등을 기획하면 충성 고객 유치에 도움이 되지 않을까 하는 결론을 내렸다.

데이터 수집에 한계가 있어 다양한 분석을 하지 못한 점은 아쉬웠지만 향후 수집하지 못했던 데이터를 확보해 분석한다면 좀 더 다양한 결과를 도출할 수 있을 것이다.

분석 프로젝트를 경험해 본 사람이라면 공감할 것이다. 실제 데이터 수집과 분석에 들어가는 시간보다 고민하고 뭔가를 결정하는 데 더 많은 에너지와 시간을 투입할 수밖에 없다는 것을 말이다. 물론 쓸모없이 보낸 시간은 아니다. 그 시간은 우리가 물에 떠 있기 위해서 수면 아래에서 물갈퀴를 열심히 젓는 순간이 아닐까 한다. 우리가 그랬듯이 후배들도 고민의 시간이 필요할 것이다. 그 시간이 앞으로 나아가는 힘을 비축하게 할 것이라고 믿는다.

멋진 자세가 뿌듯한 결과를 가져오다

아이디어와 열정이 넘쳤다! 빅데이터 아카데미의 파일럿 프로젝트는 회사 업무와 병행하면서 해야 한다. 직장에서 대부분 허리 역할을 하고 있는 조원들은 자칫하다가 파일럿 프로젝트와 회사의 업무 사이에서 혼란(?)을 겪을 수 있겠다는 우려를 했다. 그래서 인지 팀원들 스스로 프로젝트 아이디어를 적극적으로 개선하고 서로 승부욕을 일으켜 세우기 위해 술선수범하는 모습을 보였다.

2017.3.31 적극적인 아이디어와 주제 개선

자신이 제안한 프로젝트 주제가 선정될 경우 데이터 확보와 진행에 더 노력을 해야 하지 않을까! 그럼에도 서로 자신의 아이디어를 주제로 하자는 '귀차니즘' 없는 조원들이 모였다. 아, 뭔가 되겠구나! 하는 생각이 들었던 순간이다. 3개의 아이디어가 한치 물러섬 없는 각축전을 벌이고 나서야 하나가 선정됐다.

2017.4.8 주제 확정과 자율적 역할 분배

프로젝트 경험이 많은 '프로 회사원'이자 '분석 전문가 입문자' 답게 데이터 확보의 편의성과 프로젝트의 원활한 진행 가능성, 유의미하고 재미있는 결과 도출 가능성을 기준으로 주제를 선택했다. 스스로 역할 분담을 할 수 있었던 것은 행운이었다.

2017.4.22 데이터와 첫 탐색전

확보된 데이터와의 첫 만남의 순간이 왔다. 사전 탐색 후 할 수 있는 것과 할 수 없는 것을 토의하고 개념 정의를 했다. 데이터 사전 탐색 중 누락된 범위와 잘못 가공된 데이터를 발견해 재가공해야 했다. 파생변수에 대한 브레인스토밍 기간이 매우 어려웠다. 그러나 이 과정이 결과적으로 프로젝트 만족도를 높여줬다.

2017.5.13 예상치와 결과 비교, 아쉬움

개인정보 확보가 어려워 고객별 유형을 파악해 볼 수 없었던 점이 아쉬웠다. 하지만 분석 전문가 교육생으로서 데이터 전처리와 분석 방법론을 적용해 보면서 흥미로운 결과와 값진 현장 경험을 했다.

“새로운 의미와 가치 창출의 시작, 빅데이터 아카데미”



배창준

포항제철소 설비기술부 데이터 및 시스템 표준화 담당 과장

1위를 차지할 수 있었던 힘은

무엇이라고 생각하나.

우리 조는 마치 프로젝트를 추진하기 위해 준비된 팀원을 갈았다. 빅데이터 분석가가 갖추어야 할 역량인 데이터 이해, 통계 및 분석방법 이해, 분석툴에 대한 이해, 비즈니스 커뮤니케이션 스킬을 적절히 갖춘 조원들이 모였기에 팀 프로젝트를 제대로 오케스트레이션할 수 있었다. 한 명 한 명의 역할도 빛났지만 한 뜻으로 모여서 하모니를 이뤘던 경험은 두고두고 기억에 남을 것이다. 약 두 달 간 지도해 주신 멘토의 지도도 많은 힘이 됐다.

빅데이터 아카데미 분석 전문가

과정에 들어온 목적은.

새로운 변화와 함께하는 것이었다. 전 세계적으로나 기업 내에서 화두가 되고 있는 빅데이터·스마트 팩토리·인공지능·인더스트리 4.0 등은 과거에는 논하지 않았던 새로운 기조로, 변화의 물결이 우리들 가운데 깊숙이 밀려왔다는 증거다. 이러한 변화에 동참함으로써 그동안 보지 못했던 새로운 개선 기회도 발굴할 수 있다고 생각했다.

프로젝트를 진행하면서 가장

기억에 남았던 점은.

우리 조의 특징 중 하나는 멀리서 참여한 조원이 많았다는 점이다. 대전, 일산... 이 가운데서도 교육장으로부터 멀기는 저 남쪽 포항에서 올라온 조장이 최고였다.

좋은 회의장소가 딱 한 군데 있었는데, 그 곳이 바로 중회의실이다. 이는 일찍 나오는 조가 차지한다. 조원들 사이에 ‘우리가 중회의실을 선점하자’는 이야기가 나왔다. 그 이후 한 명씩 돌아가면서 중회의실을 선점! 한 번은 다른조 멘토께서 도대체 몇 시에 왔느냐고 물었다. 다른 조원들께겐 죄송^^

교육을 받고 달라진 점은.

업무 연속 선에서 빅데이터 아카데미 교육과 파일럿 프로젝트를 진행했다. 회사에서 데이터 분석 기반의 스마트팩토리 업무를 하고 있는데 빅데이터 아카데미에서 얻은 경험이 큰 도움이 되고 있다. 혼자서 데이터 분석에 대해 공부를 했다면, 이해하기 쉽거나 관심이 가는 부분을 중심으로 접근할 수밖에 없었을 것이다. 여러 배경의 수강생들이 모인 데이터 분석 교육인 만큼, 데이터 분석이 무엇이고나 하는 전체 틀을 파악할 수 있었다. 데이터 분석의 전후 과정이 크게 머릿속에 그려진 것이 무엇보다 큰 소득이다. 앞서 소개했듯이 2주 간의 집체교육과 두 달 정도의 파일럿 프로젝트를 통해 데이터 분석을 잘하기는 쉽지 않다. 하지만 앞으로 어떻게 공부해야 할지를 나름대로 파악할 수는 있다. 공부방법을 배웠던 셈이다.

분석 프로젝트를 진행하는

후배들에게 조언을 한다면.

짧은 시간 안에 뭔가 결과를 얻을 수 있고, 자기 자신에게 도움이 될 수 있는 목표를 세워서 접근하기를 바란다. 집체교육

과정만으로 분석을 잘 할 수 있으면 좋겠지만, 그러기엔 너무 짧은 시간이다. 경험자로서 봤을 때, 빅데이터 아카데미의 교육 목표는 첫째 빅데이터 분석 전문가가 확보해야 할 최소한의 지식을 전달하는 것이고, 둘째는 파일럿 프로젝트를 하면서 동료들과의 커뮤니케이션 방법 등 분석가가 갖춰야 할 자세와 문제 해결방법을 터득하게 하는 것이 아닐까 한다.

현재 진행하는 스마트 팩토리

관련한 업무에 대해 간단히 소개하면.

중국의 급추격을 비롯해 국제 철강시장이 더 치열해지고 있다. 포스코도 줄기차게 개선해왔다. 스마트팩토리는 데이터에서 재도약의 발판을 찾는 일이다. 도약의 발판을 만드는 일에 참여한다는 자부심을 갖고 도전을 기쁘게 받아들이려고 한다. 회사에서 하는 일이 제조설비 관리 분석 중에 진동파 분석 업무다. 이 분야의 전문 업체들이 하는 전문 분석 교육도 있는데, 빅데이터 아카데미의 데이터 분석 교육을 먼저 받은 이유는 뭐든 기초를 든든하게 다지는 게 좋다고 봤기 때문이다. 2주간의 집체교육 중에 데이터 분석을 하려면 통계를 잘해야 한다는 것을 절절하게 느꼈다. 그래서 요즘은 아침 출근하기 전에 1시간씩 온라인 통계학 강의를 듣고 있다.

개인적으로 데이터와 관련해

향후 계획이나 바라는 바가 있다면.

데이터에서 해결의 실마리를 찾아 문제를 해결할 수 있는 진정한 해결사가 되고 싶다. 오래 전에 미국의 대기업 P사에 빅데이터 벤치마킹을 가서 그곳에서 느낀 점 두 가지가 있다. 그 회사는 모든 일의 시작과 끝을 디지털 데이터로 관리한다는 점이었다. 데이터를 근거로 모든 일이 이루어지고 있었다. 수작업 보고서도 과감히 디지털화를 했다. 또 한 가지는 분석 전문가가 모든 문제의 해결을 주도하고 있다는 점이었다. 우리나라도 데이터가 불러온 변화의 물결 속에서 많은 변화를 겪고 있다. 이제는 데이터 분석 전문가가 분명히 자리 매김할 때가 오고 있다.

기계학습 방법을 활용한 신도시 아파트 가격변동 요인 분석



구분	빅데이터 분석, 상관성 및 회귀분석, 분류 분석
적용 도구	상관분석, 선형회귀, 나이브 베이즈(Naive Bayesian), SVM, 디지전트리
수집 데이터	국토교통부 실거래가 공개시스템 월별 자료, 공동주택관리정보시스템, 신문기사, SNS 데이터
산출물	아파트단지 가격상승에 대한 유의미 요인 도출
지도	박진수
참여자	임영규 <small>조장</small> , 강광천, 송완영, 주기형, 허승표
프로젝트 소개	경기도 신도시를 중심으로 아파트 매매가격 변동에 영향을 미치는 주요 환경요인들이 무엇인지 기계학습을 이용해 분석했다. 새로운 인사이트 도출보다는 준비·집체교육 과정에서 습득한 분석 기법을 제대로 응용해 보는 데 중점을 두고 프로젝트를 진행했다.

THE CHALLENGES

데이터를 잘 다뤄보고 싶다는 목표를 갖고 모인 교육생들이 2주 동안 하루 8시간 넘게 수업을 받았다. 정보기술 업계에서 일하는 사람이 한 달의 절반을 일터가 아닌 교육장으로 나가는 경우는 흔한 일은 아니다. 이 귀중한 시간을 어떻게 활용해야 할까? 우선 재미가 있어야 뭔가 몰입할 수 있지 않을까 하는 생각에 이르렀다.

흥미로워야 한다는 기준에 맞춰 신도시 아파트 가격 변동요인 분석을 주제로 잡았다. 남의 얘기가 아닌 우리들의 얘기이자 관심거리이기에 프로젝트 과정 중에 배우는 것도 적지 않을 것이라는 기대도 했다. 그래서 서울 근교 신도시를 중심으로 아파트 매매가격 변동에 영향을 미치는 주요 환경요인들이 무엇인지 기계학습 기법을 이용하여 분석해 보았다. 분석 과정은 해당 아파트 단지에 대한 평판분석을 위해 신문기사, SNS 등에서 추출된 긍정 또는 부정적인 단어 횟수 추이와 실매매가 추이 간 연관성 분석 및 주변 환경요인과의 연관성을 중심으로 진행되었다.

주택가격 변동요인

주택가격에 관한 이론 또는 연구의 기본방향은 주택가격에 영향을 미치는 요인에 대한 분석(주택가격 형성)과 주택가격의 변동 및 상승 및 하락에 영향을 미치는 요인에 관한 연구(주택가격 변동) 등으로 나눌 수 있다.

주택가격 형성은 크게 사회적 요인, 경제적 요인, 행정적 또는 규제 요인, 그리고 토지 자체적 요인 등에 영향을 받는 것으로 알려져 있다.¹ 사회적 요인이란 인구, 가구 구성 등을 의미하며 경제적 요인은 저축, 소비성향 수준 및 물가, 임금수준, 고용환경 등이 주된 요인으로 꼽힌다.

반면 주택가격 변동에 영향을 미치는 요인은 수없이 많을 수 있다. 주택 수요자가 거주 형태 선택 시 선호하는 환경은 각각 다를 수밖에 없기 때문이다. 예를 들어 아파트 브랜드, 단지 규모, 전용 면적, 건축 년도, 층수 등 주택 자체에 대한 선호도와 학군 및 미래교육 여건 변화 가능성, 교통으로 대변되는 거

1 이래영, [부동산투자론], 삼영사, 2004, pp. 68-70.

주지 주변 접근 용이성과 개선 가능성(지하철 개통 예정 등), 주변 관공서 존재 여부, 생활편의 시설 등을 들 수 있다.

분석 목표

주택가격 변동 요인 분석에 초점을 맞췄다. 주택가격 변동은 실제 매매가격 변동률을 기준으로 했다. 변동 요인으로는 앞서 소개한 변동 요인 전체를 고려할 수도 있지만, 데이터 접근이 용이한 공개 데이터 위주로 선택했다. 분석 목표는 크게 두 가지로 1)특정 지역에 대한 부동산(아파트) 관련 기사와 커뮤니티에서 그 지역에 대한 긍정 또는 부정적 평가 정도(횟수)가 실제 매매가격 변동 패턴과 유의미한 상관관계가 있는지 여부, 2)매매가격 상승률을 상위/하위 등 일정 구간으로 구분한 뒤 군집분석을 통해 상승률 구분에 영향을 미치는 주요 환경 요인이 무엇인지 알아보는 것으로 목표를 수립했다.

THE APPROACH

기초 데이터의 분류

본 프로젝트 분석 목표에 필요한 기초 데이터는 크게 세 가지로 분류된다.

- ① 아파트 실거래가 데이터: 국토교통부 실거래가 공개 시스템 월별 자료
(<http://rtdown.molit.go.kr/download/downloadMainList.do>)
- ② 아파트 단지 주변 환경에 대한 정형 및 비정형 데이터: 공동주택관리정보시스템 (K-apt, <http://www.k-apt.go.kr>)
- ③ 특정 지역 아파트에 대한 기사, 의견 등 비정형 텍스트 데이터: 신문기사

기초 데이터 수집

기초 데이터 수집방법으로 채택한 크롤링(Crawling) 기법은 웹 페이지를 그대로 가져와 내재 데이터를 추출해 내는 방법이다. 먼저 국토교통부 실거래가 공개시스템(<http://rt.molit.go.kr>)에서 X 및 Y 아파트 실거래가에 대해 크롤링 작업을 실시했다. 실거래가 정보를 JSON 포맷으로 가져와 파싱 후 MariaDB에

저장하는 방식을 택했다.

한편 국토부 실거래가 공개시스템은 크롤링을 중간에 차단하므로 완벽하게 데이터를 수집할 수 없다. 아파트 단지에 관한 기사, 의견, 댓글 등에 대한 정보 수집은 아래와 같은 순서로 진행하였다.

- ① 신도시(X, Y)에 기사 데이터는 조선일보 사이트에서 크롤링으로 수집(10년 치 기사)
- ② 의견과 댓글 데이터는 네이버 '지식in'과 '82cook.com'에서 수집
- ③ 각각의 기사 및 의견·댓글에 대해서 월 단위로 데이터를 통합 분류해 감성분류기로 분석

기사 및 댓글 내용의 긍정적/부정적 요소를 도출하기 위한 감성 분석은 부동산 관련 평가 데이터를 구할 수 없었으므로 영화의 평가 데이터를 기준으로 실시했다. KoNLPy로 데이터 전처리 작업을 수행했고 NLTK로 데이터 탐색, 형태소로 tokenizing 과정을 거쳐 Naive Bayes Classifiers를 적용했다. 테스트 데이터에 대한 정확도는 0.80 정도였다.

특징 추출과 분석모델 설정

감성 분석

분석대상 아파트 단지는 경기지역 신도시 중 거래량이 상대적으로 많고 감성 분석 데이터가 많은 A지역 X단지, B지역 Y단지를 선택했다. 주요인 및 분석 방법론은 다음과 같다.

표 1

감성분석의 주요인 및
분석 방법론

구분	내용 및 방법론
Features	- 6개월 단위 85m ² 이하 실거래가 변동률 - 6개월 단위 85m ² 이하 거래량 변동률 - 아파트 단지에 관한 기사, 의견, 댓글 등으로부터 긍정 및 부정적 뉘앙스를 가진 단어 추출 - 출현 횟수 증가율은 실제 매매행위로 이어지는 시차가 존재할 것으로 예상되므로 연간 실매매가 기간보다 3개월, 6개월, 12개월 전까지의 누적 합산으로 계산하고 기간은 6개월로 선택
분석모델	- 뉘앙스 분석: 나이브베이즈(Naive Bayes) - 영화 별점 데이터 이용(train, test data) - 상관 분석: 디시전트리(Decision Tree)

각 단어출현 횟수 증가율과 실매매가 변동을 간 상관관계 분석을 위하여 파이썬에 내재된 통계모듈과 디시전트리 알고리즘을 이용했다. 일반적인 교차 상관관계 분석과 비교해 디시전트리 방법의 성능분석에 초점을 맞추고자 했다.

환경요인 분석

특정 단지 주변환경 요인이 실매매가 변동률에 미치는 영향을 분석하기 위해서는 먼저 변동률 자체를 구분할 필요가 있다. 어느 지역 전체 아파트 가격이 올랐을 경우, 특정 단지를 둘러싼 환경적 요인에 의해 그 단지의 오름세가 더 크게 나타날 수 있고 이때 그 요인이 무엇인지 알고자 하기 때문이다. 실매매가 변동을 및 분석에 이용한 feature 정의를 했다. 분

석기간은 2012년부터 2016년까지 총 6개 데이터세트를 마련했다.

각 요인들이 변동률 구간을 효과적으로 구분하는지 여부를 알기 위하여 본 프로젝트에서는 SVM(Support Vector Machine) 학습법을 적용시켜 보았다. 먼저 감성분석 결과를 살펴보자. 상기 데이터세트에 대한 디시전트리 학습 결과 Y지역의 경우 모델학습 정밀도(precision)는 99%를 보였으나 모델 유효성(validation) 결과는 76%로 감소했다. 이는 특정 지역에 대한 긍정 및 부정 출현 횟수 증가율 데이터가 부족하고, 과거 기간 또한 짧기 때문인 것으로 판단된다. X 지역 역시 비슷한 양상을 보였다.

아파트 거래량과 매매가격 사이의 관계를 알아보기 위해 Y지역 데이터를 이용한 SVM 분석을 해보았다. 통상적으로 뚜렷한 분류가 가능할 것으로 예상했으나 데이터 부족으로 만족할 만한 성과는 내지 못했다.

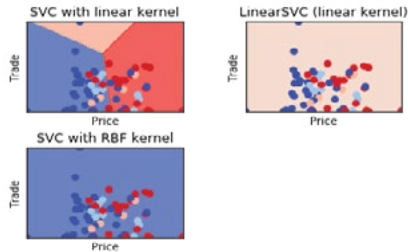
다음으로 주요 요인 간 상관분석을 실시했다. X지역 아파트 단지에 대해 아파트 가격과 브랜드 가치는 상관관계가 있으나(대립가설 성립), 설명력은 43%로 산출됐다. 가격 상승률과 브랜드 가치 역시 상관관계가 있으나(대립가설 성립), 설명력은 56% 정도였다.

그림 1
디시전트리 분석 결과



그림 2

SVC 분석결과



■ Y 지역 아파트 단지

- 브랜드 가치는 복지시설, 그리고 지하철 거리에서 강한 양의 상관관계가 있다.
- 거래량은 버스 정류장·역과의 거리에서 강한 상관관계가 있다.

■ X 지역 아파트 단지

- 브랜드 가치는 지하철 거리에서 강한 양의 상관관계가 있다.
- 그러나 학교와 복지시설과는 강한 음의 상관관계가 있다. 복지시설 분류의 제한에 의한 결과로 예상된다.

THE OUTCOME

파일럿 프로젝트에서 수도권, 특히 신도시 중심으로 아파트 매매가격 변동에 영향을 미치는 주요 환경요인들이 무엇인지 분석했다. 프로젝트의 첫 번째 목표는 집체교육 과정에서 습득한 기계학습 방법론을 실제로 적용해 보면서 분명하게 이해하는 것이었다.

진행 과정 및 결과 분석에서 도출된 시사점은 예상 외로 국내 주택관련 데이터의 정리와 체계가 매우 허술하다는 점이다. 주소 체계가 다른 점부터 시작해 같은 아파트 단지에 대한 환경적 요인이 데이터 제공 주체에 따라 다른 경우가 많아 어떤 기관의 데이터를 사용할지를 놓고 혼란스러웠다.

기계학습을 통한 감성분석과 환경요인을 분석해 주택가격 변동에 대한 민감도 분석을 할 수 있다. 이를 위해서는 요인 변동 이벤트, 예를 들어 지하철 개통예정 발표 이후 주택가격 변화 추이에 대한 데이터가 축적돼야 분석할 수 있다. 본 프로젝트 데이터세트에서는 방금 소개한 이벤트가 포함되지는 않았다.

요인분석의 또 다른 응용으로 관심지역의 환경요인 구조와 같은 구조를 가진 계산된 모델결과를 이용해 관심지역 주택가격 변동에 대한 확률적 추론을 생각해 볼 수 있다. 더 나아가 지역별 특성을 더 세분화해 분석함으로써 지역별 적정 주택가격 수준에 대한 지도(map)를 작성하고 실거래가 추이로부터 안정적인가, 과열인가, 투기 상태인가 등을 판정하는 보조 도구로 활용할 수 있을 것이다.

분석가 입장에서 원천 데이터의 중요성 실감

빅데이터 분석 전문가 19기 우수조는 함께하는 데 우선을 두고 기계학습 활용법에 대한 기본을 다지는 데 집중했다. 빅데이터 분석 프로세스의 기본 과정을 최대한 구현해 보며, 단계별 핵심 개념을 이해하는 데에 주안점을 두었다.

2017.5.27 주제 선정

첫 미팅 때 여러 가지 제안들이 나왔다. 그 중 데이터 입수가 상대적으로 쉽고 직관적 이해와 해석이 가능할 것이라는 판단에 따라 신도시 아파트 가격 분석을 프로젝트 주제로 잡았다.

2017.6.3 데이터 크롤링

분석 목표가 수립되었으니 관련 데이터를 수집하면 수월하게 마무리될 것으로 조원 모두 예상했다. 신문기사, SNS, 웹 페이지 내용 등을 크롤링 기법으로 가져오는 공개 API는 쉽게 구할 수 있었다.

2017.6.10 데이터 정형성 문제

크롤링 과정에서 해당 사이트의 접근 제한 조치 등 자동화의 걸림돌이 적지 않았다. 또 하나의 문제는 모든 데이터를 정형화하고 통일된 포맷으로 전환하는 일이었다. 부동산 소재지 주소 체계가 데이터 원천마다 다르고, 동일 원천에서조차 시간에 따라 내용이 변경된 데이터가 존재하는 등 기계학습에 필요한 데이터 프레임을 만드는 작업에 거의 대부분의 시간을 써야 했다.

2017.6.17 데이터 프레임 완성과 분석

마지막 주에 이르러서야 기계학습에 적합한 프레임이 완성됐다. 지역정보는 요인별 분류 작업에 썼고, 기사 및 평판 정보는 지역별 감성분석에 응용했다. 본 프로젝트의 목적을 새로운 결과 창출이라기보다는 습득한 기법 응용에 맞추었으므로 파일럿 프로젝트로서 매우 가치 있는 경험이었다고 생각한다.

“멋진 조원들과 쌓았던 실력과 자신감”

임영규
에이티맥스 부장



1위를 차지할 수 있었던 힘은

무엇이라고 생각하나.

조원 구성이 좋았다. 전제적인 흐름을 파악해 초기 설계를 담당했던 송영완 조원, 데이터 수집에 대한 방법 및 수집을 담당했던 허승표 조원, 감성분석과 발표 준비를 철저히 한 주기형 조원, 파이썬과 DB의 기술적 방향 제시에 힘써준 강관천 조원을 일일이 이름을 들어 칭찬하고 싶다. 늦은 시간까지 SNS와 전화로 파일럿 프로젝트의 완성을 위해 힘쓴 과정들이 좋은 결과로 이어졌다. 파일럿 프로젝트 1위를 차지할 수 있었던 것은 우리 조가 여러 조건이 잘 맞아 떨어진 결과다.

1위를 할 수 있을 것이라는

생각이 들었을 때는 언제인가.

파일럿 프로젝트 발표에서 다른 조도 높은 수준의 발표를 했다. 하지만 우리 조가 집체교육에서 배웠던 내용을 매우 충실하게 소화한 점이 눈에 띄지 않았나 싶다. 1위 조라는 발표가 나자 전혀 뜻밖이라 조원들 사이에 와! 하는 탄성이 터져 나왔다.

어느 과정이 가장 어려웠나.

분석에 필요한 아파트 실거래가 데이터는 쉽게 확보할 수 있었다. 하지만 아파트의 실거래가를 분석하고자 하는 김포·판교 신도시의 지역을 판단할 수 있는 주소 정보가 없었다. 결국 데이터를 다시 수집할 수밖에 없었다. 두 번째로는 감성분석을 위한 뉴스 및 SNS 데이터 크롤링 시 문제가 있었다. 일정량 이상의 데이터를 다운받을 수 없어 단순 반복 작업을 여러 번 해야 했다. 결국 많은 시간을 쓸 수밖에 없었다. 다른 조도 분석에 필요한 (데이터) 마트를 구축하는 데 고충이 컸을 거라고 본다.

프로젝트를 진행하면서

가장 기억에 남았던 점은.

처음 조별 미팅 자리에서 조장까지 정해야 했다. 조장은 집체교육을 마치고 프로젝트 진행 시 프로젝트 일정 및 조별 프로젝트의 결과물 발표까지 해야 한다. 나는 발표를 하지 않아 조장으로 책임자가 아니라고 했다. 그랬더니 조원 중 한 명이 자기가 발표하겠다고 하면서 조장과 발표자가 바로 결정됐다. 그 순간을 생각하면 입가에 미소가 그려진다.

분석 프로젝트를 진행할

후배들에게 조언한다면.

빅데이터 아카데미 교육 과정이 사전교육·집체교육·파일럿 프로젝트로 구성돼 있다. 결코 짧은 시간은 아니지만, 새로운 기술을 학습하기에는 부족하다. 굳이 조언하자면 사전교육과 집체교육에 충실하기를 당부 드린다.

향후 계획이 있다면.

현재 회사에서 빅데이터 프로젝트를 진행중이다. 빅데이터 아카데미에서 교육과 파일럿 프로젝트 진행 경험이 회사 프로젝트 진행에 큰 도움이 되고 있다. 앞으로 빅데이터 분석 프로젝트를 지속적으로 할 수 있었으면 한다. 멋진 조원들을 만나서 기대하지 않았던 1위를 차지했고, 좋은 인연을 맺은 것이 너무나 기쁘고 감사하다.

딥러닝 방법을 이용한 유방암 메디컬 이미지 분류·예측 모형화



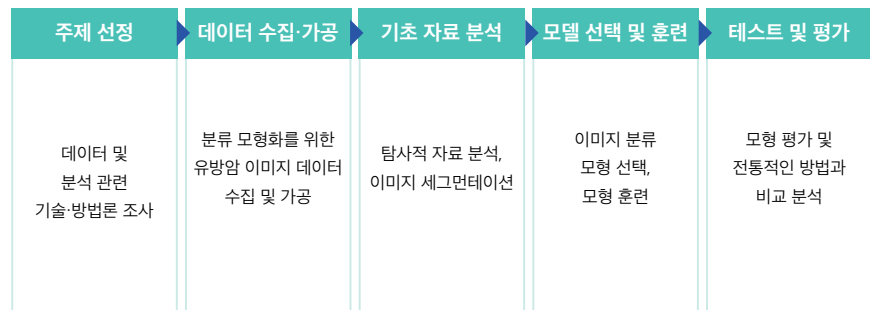
구분	빅데이터 이미지 분석, 판별분석
적용 도구	기계학습(CDNN, KNN, RF), R 패키지, Dicom 이미지 뷰어, 파이썬, 텐서플로
수집 데이터	TCIA(The Cancer Image Archive) 데이터베이스에 저장된 TCGA 139명 유방암 환자의 230,167개 이미지 데이터
산출물	유방암 MRI 이미지 판별 모델 및 향후 활용·적용 방안
지도	오원기
참여자	장인수 ^{초장} , 신단, 양승민, 임한경, 최순호
프로젝트 개요	유방암은 전 세계적으로 여성암 1위이다. 국내에서도 2012년 기준 1991년부터 20여년 간 발병률이 5배로 증가했다. 딥러닝을 적용하여 더 쉽고 빠르고 정확하게 유방암 여부를 예측할 수 있는 모형을 구축하는 프로젝트를 진행함으로써 조기 발견·치료에 도움을 주는 것을 목표로 했다.

THE CHALLENGES

파일럿 프로젝트를 진행함에 앞서 우리 조는 프로젝트 주제를 도전할 가치가 있고, 사회적으로 기여할 수 있는 것이면 좋겠다는 공통 의견을 모았다. 의료 분야라면 도전적 가치·사회 모두 만족할 것이라 생각해 의료 분야를 큰 주제로 선정했다. 그럼 다양한 의료 분야중 어떤 세부 주제를 선택할 것인가? 현대 인간 삶의 수준이 향상됨에 따라 건강한 삶의 욕구 및 관심도가 증가하고 있다. 하지만 건강에 대한 욕구와 다르게 환경적·유전적 요인 등 다양한 요인에 의해 암 발병률이 높아지고 있는 추세다. 그중 유방암은 전 세계적으로 여성암 1위를 차지할 정도로 발병률이 높다.

관련 자료를 조사하면서 유방암 의심 환자들 중 실제 암으로 판정받는 비율이 0.6%에 그친다는 신문 기사를 접할 수 있었다. 이렇듯 검사 비용 대비 판정률이 낮게 나오면서 의료보험 등 사회적 비용이 낭비되는 것은 아닌가, 실제 검강검진의 목적인 조기진단의 역할을 생각하게 되었다. 우리 조가 제시하는 모형이 오분류를 낮출 수 있다면, 검진으로 낭비되는 비용을 줄이고 더 빠르게 유방암을 판별해 조기 치료에 도움이 될 것으로 본다. 문제는 유방암 데이터 중 어떤 것을 사용할 수 있는지, 실제 데이터를 구할 수 있는지였다. 모든 의료 데이터는 인간 윤리 문제 때문에 심의를 받아야 사용할 수 있다. 하지만 미국 TCIA(The Cancer Image Archive)에서는 임상정보를 삭제한 유방암 MRI(magnetic resonance imaging) 및 일반 검사 이미지인 MG(MammoGraphy)를 서비스하고 있어 제약없이 분석에 사용할 수 있었다.

그림 1
프로젝트 분석
프로세스



우리 조의 프로젝트 목표는 실제 유방암 환자의 MRI 이미지를 이용해 판별 모형을 만들고 만들어진 모형의 오분류율을 최소화하여 임상 의사 또는 건강검진을 통해 얻은 유방 이미지를 소유한 일반인을 대상으로 유방 MRI 이미지를 이용해 빠르고 쉽게 암 유무를 판별할 수 있는 판별 모형을 구축하는 것이다.

THE APPROACH

데이터 수집

미국 TCIA 데이터베이스로부터 TCGA(The Cancer Genome Atlas) 유방암 이미지 데이터(BRCA)를 확보했다. 데이터는 139명 환자의 23만 167개 이미지(용량 88.58GB)로 MRI, MG(일반 유방암 검사) 검사 이미지다. 환자별로 수집에서 수백 장까지의 이미지를 갖도록 구성돼 있다. 유방암 MRI 이미지는 다양한 방향에서 촬영하는데 이번 분석에서는 분석 복잡도(complexity)를 줄이기 위해 [그림 2]의 박스로 구분한 이미지만 이용했다. 환자의 모든 이미지를 사용하지 않고 중앙 조직(tumor)이 직관적으로 확인되는 이미지만 선별했다. 최종 49명 환자에 대한 747개 이미지를 분석에 이용하기 위해 추출했다.

데이터 전처리

수집된 데이터는 49명의 유방암 환자 이미지이므로 실제 모형을 평가하기 위해 대조군인 정상인 유방 이미지가 필요했다. 온/오프라인에서 분석에 이용할 정상인 유방 MRI 이미지를 구하기 어려워 다음과 같이 한 가지 가정을 했다. ‘한 쪽 유방에만 암 조직이 있다

그림 2

TCGA-BRCA 유방암 이미지 타입

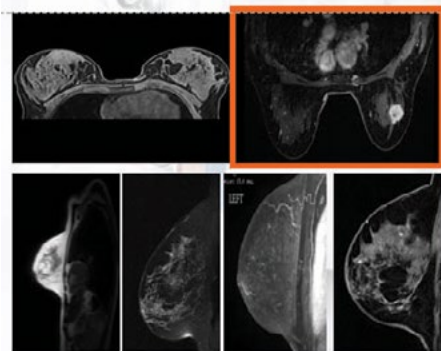
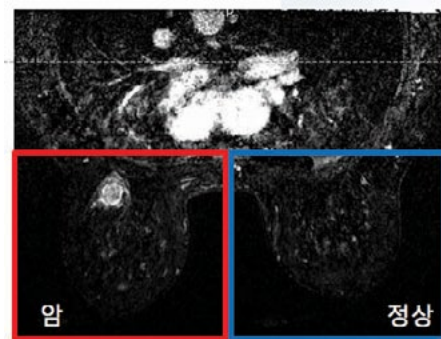


그림 3

암-정상 이미지 캡처



면 다른 쪽 유방은 정상이다.’ 따라서 [그림 3]과 같이 암 환자의 유방 이미지에서 한 쪽 유방에 종양 조직이 있다면, 종양 조직이 있는 유방을 암-유방, 그 반대쪽 유방을 정상-유방으로 분리하는 작업을 수행했다.

이미지 분리 작업은 R을 이용했다. TCIA에서 제공하는 이미지 데이터는 DICOM(Digital Imaging and COmmunications in Medicine) 포맷 파일로 ‘oro, dicom R’ 패키지를 이용해 이미지를 읽어 들이고 필요한 이미지를 부분 캡처해 JPEG 이미지로 변환·저장했다. 이렇게 모은 데이터는 [표 1]과 같다.

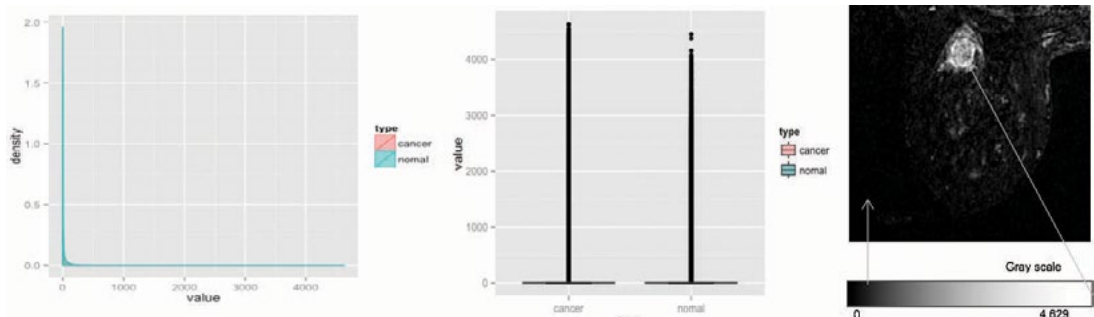
표 1
모델링에 이용할
데이터

구분	암	정상
샘플 수(명)	49	49
이미지 수(개)	747	747
이미지 포맷	JPEG	
이미지 크기(픽셀)	256×357(91,392)	
파일 크기(MB)	317	

데이터 탐색

일단 두 그룹, 즉 암 또는 정상 이미지의 모든 픽셀(pixel) 값 분포를 확인하기 위해 R 패키지를 이용해 상자그림(boxplot)과 밀도곡선(density curve)을 그려 보았다. 단순히 이미지를 살펴 보면 대부분의 픽셀이 검정색으로, 몇몇 부분에서 검정과 흰색 사이의 값으로 된 그림이다. 따라서 [그림 4]와 같이 밀도곡선이나 상자그림 모두 그룹 간 분포 차이를 가늠하기 힘든, 즉 검정(0)으로 많은

그림 4
이미지 분포 및
픽셀값 스케일



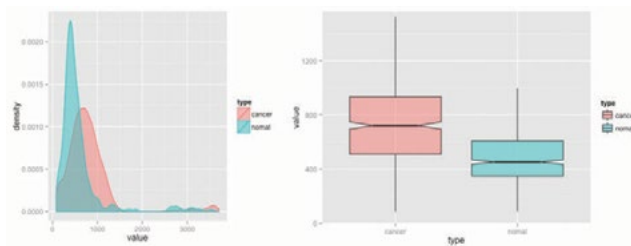
데이터가 치우쳐 있는 형태의 분포를 확인할 수 있다.

하지만 대부분의 종양은 이미지에서 밝기가 높은 픽셀값을 갖기 때문에 상위 몇 % 값의 차이를 보는 데에 의미가 있을 것으로 판단됐다. 이에 따라 모든 이미지에 대해 상위 20%의 픽셀값만 추출했다. 확보한

이미지 픽셀값 평균을 구해 분포를 살펴 보았다. ggplot2를 이용해 분포를 그려보니 [그림 5]와 같이 암과 정상 이미지의 상위 20% 픽셀 평균값의 분포에서 차이가 남을 확인할 수 있었다. T-검정을 이용해 두 그룹간 평균의 차이가 있는지 테스트한 결과 유의확률(p-value)이 $1.802e-12$ 로 아주 유의하게 나와 암/정상 간의 상위 20% 픽셀값의 평균은 차이가 있음을 알 수 있었다.

그림 5

상위 20% 픽셀의 평균값 분포



특징 세그멘테이션

그렇다면 실제 픽셀값의 조작으로 암/정상을 구분할 수 있는 특징(종양)을 찾을 수 있을까? 일단 하나의 이미지 샘플만 암/정상을 판별할 수 있는지 테스트했다. 위 데이터 탐색에서도 설명했듯이 종양의 픽셀값은 다른 조직보다 밝고 여러 개 픽셀이 모여 밀도가 높기 때문에 픽셀의 밝기 강도만을 이용해 종양 조직을 분할(segmentation)해 보았다. 일단 정상 이미지 픽셀의 최상위 값을 threshold값으로 하고, 암/정상 이미지 모두 threshold값보다 크거나 같은 픽셀값만 그대로 두고 나머지는 검정색으로 바꿔보았다. [그림 6]과 같이 종양 조직 부분만 정확하게 찾을 수 있었다.

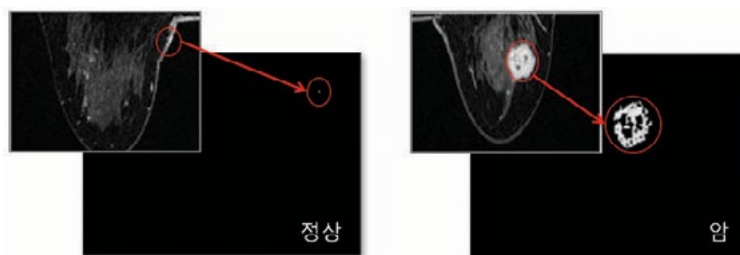


그림 6

특징 세그멘테이션

유방암 이미지 판별과 평가

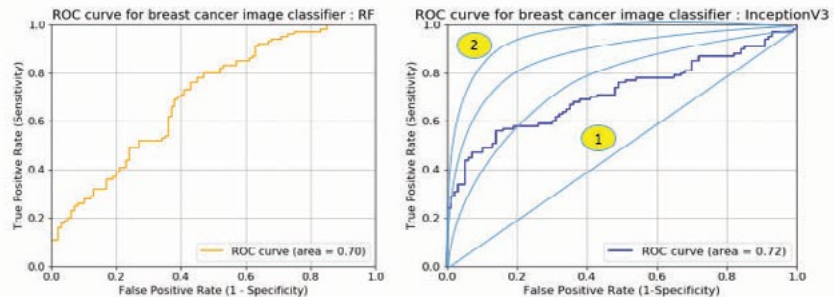
모형을 구축하기 위해 사전에 수집된 1,494 이미지 데이터는 모형을 훈련 시키기 위한 데이터 1,294개, 모형을 검증할 데이터 200개로 이미지를 구분했다. 판별모형 구현은 현재 딥러닝 및 AI 분야에서 많이 사용되고 있는 파이썬 기반의 텐서플로(TensorFlow)를 이용했다.

유방암/정상 이미지를 판별하기 위한 판별 모형을 3가지를 적용했다. 크게 딥러닝 이전에 전통적으로 사용했던 기계학습(machine learning) 방법인 kNN(k-Nearest Neighbors), RF(Random Forest)과 CDNN(Convolutional Deep Neural Network) 기반의 딥러닝 방법이다. 여기서 딥러닝 방법은 2012년 구글에서 자체 이미지 분류를 위해 구축해 놓은 모형인 Inception V3를 우리 조의 유방암 이미지를 이용해 다시 훈련시켜 사용했다.

첫 번째로 k=3인 kNN을 이용한 분류 결과는 [그림 7]과 같이 실제 암환자를 암으로 판단한 민감도(Sensitivity)가 100명 중 59명(59%), 정상인을 정상으로 판단한 특이도(Specificity)가 100명 중 61명(61%)으로 전체 정확도는 60% 정도를 보였다.

두 번째인 랜덤포레스트 모형(500번 반복, 100개 트리 생성, 500개 이미지 샘플

그림 7
모형 평가 결과



구분	kNN	Random Forest	Inception v3
민감도	59%	71%	87%
특이도	61%	42%	65%
정확도	60%	56.5%	76%
제 2종 오류 비율	41/200 (20.5%)	29/200 (14.5%)	13/200 (6.5%)
AUC	0.6	0.7	0.72

플링)을 적용한 결과에서는 암환자를 암으로 판단한 민감도가 100명 중 71명 (71%), 정상인을 정상으로 판단한 특이도가 100명 중 42명(42%)으로 전체 정확도는 56.5%를 보였다.

마지막으로 [그림 7]과 같이 딥러닝을 적용한 Inception V3 모형의 예측 결과는 암환자를 암으로 판단한 민감도가 100명 중 87명(87%), 정상인을 정상으로 판단한 특이도가 100명 중 65명(65%)으로 전체 정확도는 76%를 보였다.

의료 데이터에서의 모형 평가를 위해서는 얼마나 암 또는 정상을 정확하게 분류할 수 있는 정확도가 중요하지만, 제 2종 오류(Type II error), 여기서는 실제 암환자인데 정상으로 판정 받는 경우가 아주 위험(Critical)한 문제가 될 것이다. 따라서 제 2종 오류가 작은 모형이 좋은 모형이라고 볼 수 있다. [그림 7] 아래의 테이블에서 보는 것과 같이 실제 제 2종 오류도 딥러닝을 이용한 모형이 다른 2가지 모형보다 2~3배 이상 적게 검출됨을 확인할 수 있었다.

좀 더 시각화 해서 3가지의 모형을 평가해 보기 위해 ROC(Receiver Operating Characteristic) 곡선을 그려 보았다. [그림 7] 위쪽 그래프는 False Positive(제 1종 오류) 비율 대비 True Positive 비율을 그래프로 표현한 것이다. 그림에서 정상을 암으로 판단하는 비율(False Positive Rate)을 0.2로 고정시켜 보면, 랜덤포레스트는 약 40% 미만의 암환자를 암으로 분류하며, 딥러닝 방법은 60% 근처로 좀더 높은 비율로 암 이미지를 정확히 판단할 것으로 보인다. 또한 실제 그래프 아래쪽의 면적을 수치로 표현한 AUC(Area Under Curve) 값을 보더라도 딥러닝 방법이 가장 좋은 0.72로 나타나 3개의 모형 중 가장 뛰어난 것으로 나타났다.

분석의 마지막으로 그림 어떤 이미지들이 잘못 분류(암을 정상으로, 정상을 암으로)가 되었는지 오분류된 이미지를 살펴보았다. [그림 8]의 위쪽 그림과 같이 암환자가 정상으로 분류된 True Positive(제 2종 오류) 이미지를 살펴 보면 유방에 밝기강도(Intensity) 값이 높은 픽셀이 전반적으로 퍼져 있는 형상의 이미지였다. 반대로 아래쪽 그림에서 정상인이 암으로 분류된 False Positive(제 1종 오류) 이미지를 살펴 보

그림 8
오분류된 이미지



면, 전반적으로 검정색이지만 유독 몇몇 지역에 높은 픽셀값들이 모여 있는 것을 확인할 수 있다. 결과적으로 암/정상을 구분하는 요인이 밝기 강도가 높은 픽셀값들의 특징 지역 그룹화 여부로 구분하는 것으로 보인다.

THE OUTCOME

결론적으로 여러 가지 이미지 분류 모형을 이용해 본 결과, 세 가지 모형 중에서는 CDNN 의 파생형인 Google Inception V3 모형이 (속도 제외) 결과적으로 가장 좋은 성능을 보여주었다. 그럼 이 Inception v3 모형을 현장에 바로 적용할 수 있을까? 어디까지나 수치적으로 76%의 정확도와 87%의 민감도를 얻었을 뿐, 이 모형을 실제 의료 현장에 바로 적용할 수 있을 것이라고는 생각하지 않는다.

하지만 인간의 건강한 삶을 실현하기 위해서는 정확도는 높으면서 더 저렴하며 쉽게 조기에 질병을 진단할 수 있는 방법이 필요한 관점에서 가능성을 어느 정도 보여준 프로젝트라 생각한다. 이러한 시도와 도전 자체가 의미가 있다고 생각한다. 이번 프로젝트에 수행했던 유방암 이미지 판별 모형은 유방암뿐 아니라 다른 암과 질병으로 확대·적용해 분류할 수 있는 모형이라 생각한다.

물론 이미지를 이용한 질병 판별·예측 분야에도 넘어야 할 장애물이 많다. 예를 들어 의학 이미지에서 실제 비교 대상이 되는 부분(유방, 폐, 위 등) 이외의 다른 장기(Organ) 조직에 대한 처리, 노이즈, 여러 제조사의 의료 기계로부터 생산된 이질적인 이미지 처리 등 데이터의 품질부터 이를 처리할 수 있는 알고리즘 개발까지 다양한 장애물들이 존재한다. 이러한 부분에 대한 발전이 있다면 기계학습을 이용한 의료 진단 분야에서 획기적인 발전이 있을 것으로 기대된다.

마지막으로 우리가 구현한 유방암 이미지 판별 모형을 특정 이미지 형태만 이용하는 것이 아니라, 일반적인 유방암 이미지 전체를 이용해 재훈련시키고 모델을 최적화해 정확도를 높이고 오분류율을 낮추어 실제 의료 현장에서 이용될 수 있는 모형으로 발전시켜 나가고 싶다.

배움의 즐거움

다른 분야에서 일하는 조원 5명이 모여 공통의 관심사를 찾기 위해 머리를 맞댔다. 5명 모두 이미지 분석 경험 전무, 게다가 의료 분야 지식이 거의 없는 생소한 분야였던 터라 어떻게 접근을 해야 할지 막막했다. 하지만 조원 모두 배우고자 하는 열정과 해보자는 긍정적인 마인드가 있어 즐겁게 프로젝트를 수행할 수 있었다.

2017.10.14 사회적으로 도움이 될 만한 주제 선정

주제를 선정하면서 배경이 다른 조원들의 공통 관심사를 찾기 어려웠다. 하지만 조원 모두 공통적으로 사회에 도움이 될 수 있고, 도전할 만한 가치가 있는 프로젝트를 원했다. 결국 주제는 다양한 의견 중 암 관련 연구를 진행하고 있는 장인수 조장의 아이디어를 채택하여 유방암 판별 분석으로 정했다.

2017.10.21 데이터 이해와 확보

이미지를 이용한 유방암 판별 분석을 위해 자료 조사와 데이터를 수집했다. 문제는 대조군(정상) 데이터 확보가 어려웠다. 멘토 및 조원들과 상의 후 대조군 데이터는 유방암 환자 하나의 이미지를 암/정상 유방으로 각각 분리해 처리하기로 했다. 유방암 이미지 해석도 문제였다. 어느 부분이 암 조직인지 판단이 어려웠으나 간호사인 신단 조원 아내의 도움을 받을 수 있었다. 모든 이미지 데이터는 조원 모두 균등하게 할당해 암/정상 그룹 구분과 훈련, 테스트 데이터를 확보했다.

2017.11.14 경험이 전무한 이미지 분석

데이터 확보 후 위기가 찾아왔다. 이미지 분석을 어떻게 시작해야 할까? 조원 모두 이미지를 분석해본 경험이 전무했다. 또한 여러 장의 데이터를 핸들링하는 데 메모리 크기가 작거나, CPU의 성능이 좋지 않은 경우 멈추는 경우가 종종 발생해 좋은 컴퓨터 자원이 필요했다. 다행히 분석의 진행은 멘토의 도움으로 이미지를 같은 행렬 데이터로 인식해 처리함으로써 R과 파이썬을 이용해 큰 무리없이 분석할 수 있었다.

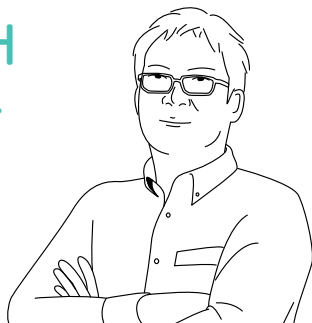
2017.11.11 분석 자료 취합 및 결과 도출

지금까지 분석했던 결과를 종합하는 시간이었다. 이미지 데이터의 분포 및 3가지 판별 모형의 결과를 종합하면서, 발표 자료를 하나씩 조원들과 준비해 나갔다. 추석과 데이터진흥원의 행사에 따라 프로젝트 일정이 늘어나면서 집중력을 잃지 않을까 걱정했다. 경험이 없는 분석을 진행하는 두려움이 있었지만, 그저 막막하게 보였던 프로젝트 결과물을 보니 감회가 남달랐다. 고생 많았던 5조 조원들 모두의 열정과 노력에 감사 드린다.

“협업으로 벽을 넘어 기계학습의 실체를 바라보다”

장인수

한국생명공학연구원 연구원



분석전문가 20기의 주제가

매우 신선했다는 평가를 받았다.

개인적으로는 암이 어떻게 발병하고 어떤 메커니즘으로 작동하는지에 관심이 많다. 주변의 지인들, 더 가깝게는 어머니께서 유방암을 앓고 계셔 개인적인 관심도 있다. 데이터를 이용해 암을 조기에 진단할 수 있는 통계 모형을 개발하고 싶었다. 이런 관심을 조원들과 주제 선정 과정에서 공유했다. 조원 모두가 사회에 기여할 수 있는 주제, 정말 도전할 만한 가치가 있는 주제를 선정하길 원했으므로 이 주제를 정할 수 있었다. 게다가 우리 프로젝트 주제가 신선했다는 평가는 대부분의 사람들이 관심을 갖는 주제였기 때문이지 않나 싶다.

프로젝트 과정에서 가장

어려웠던 순간과 그 해결 방법은.

가장 어려웠던 부분은 조원들이 서울, 대전, 안산, 부천에서 거주해 자주 모일 기회가 부족했던 점이다. 조원 모두 데이터의 특징을 잘 파악하고 이해해야만 좋은 분석 결과를 도출할 수 있을 것이다. 하지만 조원 모두 이미지 분석 경험과 의료 이미지 데이터 해석 경험이 없었다. 따라서 데이터의 특성을 파악하기 위해 스터디를 통한 지식 공유의 자리가 필요했다. 지식을 공유하고 이해를 도와야 조원들이 프로젝트에 더 관심을 갖고 진행할 수 있었을 텐데, 자주 모이기 힘들었던 점이 가장 아쉬웠다. 네이버 밴드(Band)를 이용해 지식을 공유하고 중간 결과 및 해석을 공유해 이해를 도왔다. 또한 모두 데이터

분석에 관심이 많았기 때문에 동일한 일을 하더라도 골고루 역할을 분배해 참여했다.

협업을 잘했던 팀이라고 들었는데

그 비결이 궁금하다.

다른 기수, 다른 조도 마찬가지로겠지만 공교롭게도 조원 5명 모두 다른 업무 배경을 갖고 있었다. 모두 관심 주제가 다들 텐데 관심 없는 주제가 선정됐더라도 배우고자 하는, 새로운 내용을 하나라도 더 해보자는 긍정과 열정을 갖고 적극적으로 참여했다. 주중에 출장이 있어 서울에 올 일이 있어서 ‘그동안의 결과를 공유하자’고 긴급 연락을 했을 때 이미 일정이 있었던 1명을 제외한 4명이 모두 참석했다. 관심과 열정이 없었다면 이렇게 협업이 잘 되진 않았을 것이다.

어떤 일을 하고 있고, 빅데이터 아카데미 분석 전문가 과정에 지원한 이유는.

생물정보학 분야에서 일하고 있다. 이는 다양한 생명현상을 컴퓨터로 해석하고 이해하는 학문 분야로 의료 분야와 밀접한 연관이 있다. 요즘 생명 현상을 연구하기 위해 수많은 자료가 전 세계적으로 쏟아져 나오고 있다. 이를 분석하고 해석할 사람이 필요한 시대가 되었다. 현재 이슈가 되고 있는 딥러닝 방법을 이용한 기계학습에 관심이 생겨 관련 교육을 찾아보게 되었다. 이 관점에서 빅데이터를 실제 이론 교육만이 아닌 현장 전문가로부터 조언을 받고 실제 데이터를 분석·운영해볼 수

있는 기회를 갖고자 빅데이터 아카데미에 지원하게 되었다. 수강 전 생각했던 것처럼 이론교육보다는 파일럿 프로젝트 과정 중에 익히고 배운 내용이 실제 현장 분석에 많이 도움이 될 것이라고 생각한다.

빅데이터 아카데미 수강 전과

후에 달라진 점이라면.

지금까지 통계 분석을 해왔지만, 내가 알고 있는 지식이 알고 공부해야 할 내용이 여전히 많음을 느낄 수 있었다. 데이터를 보는 시각이 달라졌다. 집체교육 중에 한 강사로부터 “데이터를 이해하기 위해 탐사적 자료 분석을 한 달 정도 수행한다”는 이야기를 듣고 좀 충격적으로 받아들였다. 현재, 내가 알고 있는 데이터라고 쉽게 넘어간 부분은 없나? 데이터를 이해하기 위해 좀 더 노력한다면 뭔가 더 많은 내용을 알아낼 수 있지 않을까 하는 생각을 했다. 또한 현재 인공지능·딥러닝 분야가 이슈가 되고 있지만 합성곱 신경망(convolutional neural network) 및 강화학습(reinforcement learning) 등 많은 분야가 학습의 의욕을 높이는 계기가 된 것 같다.

프로젝트를 주제를 심화하면 협업에 적용할 수 있어 보인다. 조원들과 향후 계획에 대해 얘기를 나눈 바가 있다면.

향후 계획을 얘기해본 적은 없다.

사실 학술적이든 상업적이든 인공지능, 특히 의료분야의 인공지능 분야의 연구가 많이 진행되고 있다. 지금 진행되고 있는 연구와 비교해 본다면 어느 정도의 수준인지는 가능하기 힘들다. 이번에 적용한 딥러닝 방법도 사실 2012년도에 발표된 Inception V3 모델이다. 현재는 V4 모델이 발표된 상태이므로 시대적으로도 상당히 뒤쳐진 것으로 생각된다. 물론 시작이 늦었지만 이번 교육을 통해 최신 수준의 기술 및 지식을 습득하는 데 도움이 된 것으로 본다. 또 이런 주제에 대한 개인적인 관심도도 높기 때문에 계속 발전시켜 현장에 적용 시키고 싶다.

어느 학교를 가야할까요?



구분	빅데이터 분석, 예측분석
적용 도구	패키지(Pandas, Scikit Learn, Numpy, TensorFlow) / 알고리즘(K-Means Clustering, Regression, PCA(주성분 분석), Random Forest) / 시각화(Scatter, 3D-Scatter, Pairplot 등)
수집 데이터	학교알리미 공공데이터, 지역별 상권 공공데이터
산출물	중학교 졸업생의 고등학교 진학에 미치는 영향력 분석(유형별 고등학교 진학률로 본 학교 내부 요인 영향력 분석)
지도	이강욱
참여자	최윤지 ^{조장} , 광동휘, 권도진, 김은희, 손성수, 황재호
프로젝트 개요	교육 과정 개편에 따라 데이터 분석 결과를 근거로 자녀가 진학할 고등학교 선택을 할 수 있도록 돕는 분석 프로젝트다. 학교별 정보가 공시된 학교알리미 공공데이터를 분석해 고교 선택에 참고할 만한 항목을 모아서 빅데이터 기법으로 분석했다.

THE CHALLENGES

최근 ‘어느 고등학교를 가야 할까요?’ 하는 3학년 중학생 또는 2학년 중학생 학부모들의 질문으로 각종 인터넷 커뮤니티 게시판이 뜨겁다. 2015 개정교육과정 도입, 수능 개편안 1년 유예에 자립형사립고·일반고 동시 선발까지 교육제도의 큰 변화에 맞춰 자녀에게 맞는 고등학교를 찾으려는 부모들의 움직임 때문이다.

고교 내신이 중요해졌고 수시 학생부종합전형 시대의 비교과 활동까지 해야 한다는 사실은 어느 정도 알고 있다. 그렇다면 자녀가 진학할 고등학교를 선택하는 데 입소문에 의지할 수 없다는 부모들이 질문을 올리고 있다. 분석 전문가 21기 우리 조는 학교별 정보가 공시된 ‘학교알리미’를 통해 고교 선택에 참고할 만한 항목을 모아 빅데이터 기법으로 분석해 보았다.

THE APPROACH

데이터 전처리

2017년 학교알리미 공공데이터 통합작업과 범주별 데이터 형태 변경작업을 했다. 모든 자료에 선형 변환을 적용해 전체 자료의 분포를 평균 0, 분산 1이 되도록 만드는 스케일링 작업을 거쳤다. 이는 언더플로우와 오버플로우 방지 및 독립변수의 공분산 행렬의 조건수(condition number)를 감소시켜 최적화 과정에서 안정성과 수렴 속도를 높이기 위한 일이다.

그림 1

분석 프로젝트
수행 절차



데이터 분석

데이터 분석에 활용된 알고리즘으로는 K평균(K-means), 상관관계, 주성분, 랜덤포레스트 등을 이용했다. 시각적인 분석을 위해 다양한 시각화 그래프를 이용했다.

종속변수 선정에 앞서, 우리 조는 졸업생의 진학률을 최대한 잘 설명할 수 있는 가설을 찾아야 했다. 이를 위해 주어진 데이터를 관련성 있는 특성으로 분류해 줄 수 있는 K-means 알고리즘을 활용했다. 여기서 최적의 K값을 도출하기 위해 오차제곱합(Sum of Squares Error, SSE)이 최소가 되도록 하는 elbow 그림 넣기, 클러스터링의 품질을 정량적으로 계산하는 실루엣 그림 넣기, 3D-스캐터 그래프 등 여러 가지 방법을 활용했다. 분석 과정에서 가지 수를 4개 위로 늘려도 분류의 의미가 무색해지는 모습이 나타났다. 이에 가장 최적의 K값을 '4'로 선정했다. 이렇게 4가지를 차례대로 '진학목적 상, 진학목적 중, 진학목적 하, 취업목적'이라는 라벨을 적용하였다.

종속변수가 설정된 후, 독립변수를 설정하는 단계에서 가장 먼저 각 변수들이 종속변수와 관련이 있는지 여부 파악을 위해 상관관계분석을 했다. 몇 가지 변수들이 종속변수와 관련성이 있음이 드러났다. 좀 더 자세히 검토하고자 '주성분 분석(PCA)'을 시도해 보았다. 주성분 분석은 방대한 데이터의 '차원수' 전체에서 벗어나지 않는 형태에서 차원수를 줄여나가면서 종속변수에 가장 크게 영향을 미치는 성분이 무엇인지 알 수 있게 해준다. 여기서의 약 세 가지 정도의 변수가 전체의 90%를 설명하는 것이 드러났다. 이를 토대로 독립변수의 구분을 '교육활동, 교육여건, 교사현황'이라는 3가지 속성으로 나누어 각 속성의 대표변수 선정에 할당했다. 산점도 분석을 통해 대표변수들이 전체를 대변할 수 있는지에 대해 판별했다. 그 결과 대표변수들이 전체변수들과 유사하게 나왔다.

이어서 우리 조의 분석 목적인 해당 변수들 중에서도 종속변수와 가장 영향력이 있는 변수를 찾아 나섰다. 여기서는 랜덤포레스트를 이용했다. 그 결과 1위가 '1인당

그림 2

최적의 K 값으로 4가지 라벨 적용

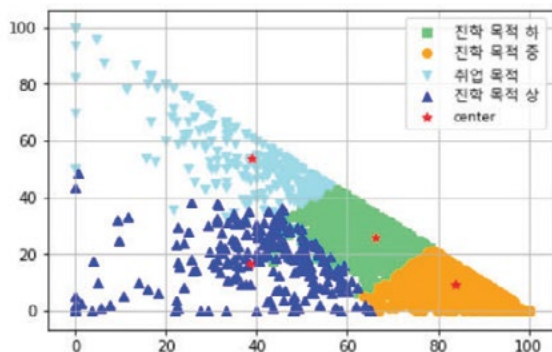
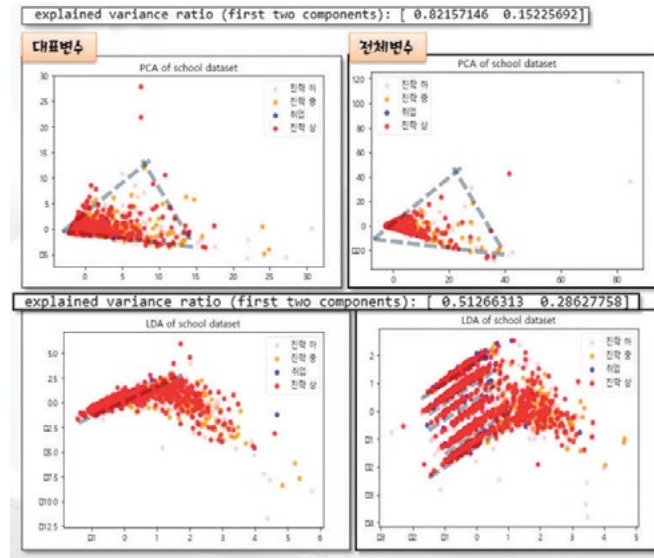


그림 3

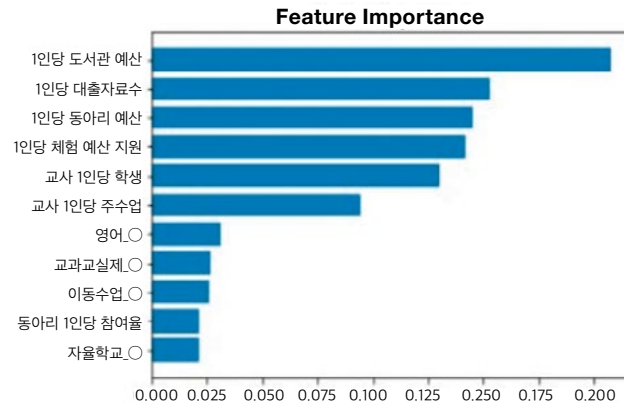
대표변수 판별을 위한
산점도 분석



도서관 예산', 2위가 '1인당 대출 자료 수'로 분석됐다. 이 중에서 '1인당 도서관 예산' 변수가 가장 중요도가 높았고, 동시에 평균편차보다도 차이가 많이 났다. 즉 학교 투자예산 수준이 고교 진학에 가장 영향력을 미친다는 것을 알 수 있었다. 이에 따라 각 학교의 투자예산 순위 분석 후 지역과 관계가 있는지를 추가적으로 분석했다. 그 결과 서울 강남에 위치한 중학교 혹은 수도권외의 중학교들이 상위권에 집중 분포된 것을 확인할 수 있었다.

그림 4

랜덤포레스트로
중속변수와 영향력
있는 변수 탐색



THE OUTCOME

우리 조는 최종적으로 중학교 졸업생 진학에 가장 영향을 미치는 요인을 다음과 같이 도출했다. 1)학교 내 예산투자, 2)학생 참여도, 3)교사가 핵심 요인이었다. 이 중에서 가장 영향력이 큰 것은 ‘학교 내 예산투자’이었다. 결국 교육 투자를 많이 하는 학교일 수록 해당 졸업생 진학의 질이 올라감을 알 수 있다. 다소 아쉬웠던 부분은 ‘학업성취도’ 부분이 비공개 데이터라 변수로 활용하지 못한 점이었다. 추후 해당 데이터를 추가한다면 더욱 정확한 분석이 가능할 것으로 보인다.

일선 학교를 위한 애플리케이션으로서 가능성

해당 데이터 분석 과정 및 결과가 활용될 수 있는 부분으로 학부모, 학생, 회사, 정부, 학교 등 크게 5가지로 볼 수 있다. 학부모들은 빠르게 변화하는 교육 환경 속에서 신뢰할 수 있는 데이터 분석 결과를 토대로 자녀가 최적의 고교를 선택할 수 있도록 지도할 수 있다.

어느 정도 목표가 있는 학생은 분석 결과를 토대로 향후 진로에 대한 고민을 줄일 수도 있다. 교육업체 등 대학 진학률을 높이려는 곳에서는 정확한 데이터에 근거해 학생들 지도에 유용한 자료로 활용할 수 있다. 더 나아가 신뢰할 만한 데이터를 토대로 마케팅 활동과 컨설팅도 가능할 것이다. 정부에서는 데이터를 기반으로 한 교육정책 수립도 가능하다. 일선 학교에서도 이러한 데이터 분석에서 나온 결과로 교육의 질 제고 방안을 도출할 수 있다.

분석 결과로만 끝나는 것이 아닌, 이를 자동화할 수 있다면 더 실용적인 것도 가능하다. 예를 들어 성적, 연령, 원하는 지역, 목표하는 학교, 장래 희망 등의 옵션을 지정해 검색하면, 자동으로 해당 학생에 맞춰 추천 학교가 순위별로 나오는 프로그램도 생각할 만하다. ‘최근 3년간 특목고를 가장 많이 보낸 학교의 특성’ 등과 같이 진학 관련 궁금증을 풀어줄 수 있는 애플리케이션을 만들어볼 수도 있다.

행복한 파일럿 프로젝트를 위하여

우리 조가 최우수 조가 될 수 있던 이유를 꼭 소개하고 싶다. 먼저 최적의 플랜

그림 5

진학지도 애플리케이션 안

◆나에게 맞는 학교 찾기◆

나이:

13

지역:

장래희망:

동아리활동:

도서관이용:

조회

★ 지역 및 학교별 특성 검색 ★

※궁금한 사항에 맞춰 분류하기

기간: 최근

4년

지역: 서울▼

전체▼

특성: 교사수업시수▼

조회

을 먼저 세워 역할을 분담하고, 해당 결과를 실제 사례에 어떻게 반영할 것인가를 도출한 것이 큰 힘이 됐다. 프로젝트를 진행하는 동안의 마음가짐도 중요하다. ‘이걸 어떻게 해?’라는 관점보다 ‘해보자’ 하고 생각을 바꾼다면 더욱 즐겁게 프로젝트를 진행할 수 있을 것이다.

사람은 무언가를 할 때 처음에는 어설피도 오랫동안 하다 보면 능숙해지게 마련이다. 뇌도 마찬가지다. 처음에는 조금만 생각해도 피곤하고 어려웠던 것이 점차 해당 생각을 떠올리는 데 익숙해지면 어려움과 피곤함이 덜하게 된다. 처음 프로젝트를 하는 것이니 당연히 내 마음대로, 내 생각대로 되지 않을 때가 있다. 그러나 이는 앞으로 내가 더 큰 세상으로 나아가기 위한 작은 발걸음 정도일 뿐이다. 부담보다는 즐겁게 임하는게 좋다. 다시 말해서 프로젝트의 완성도도 중요하지만 완벽보다는 내 스스로가 얼마나 알게 되었는가에 초점을 맞추는 것이 더욱 행복한 프로젝트 과정이 될 것이다.

최적 방법론과 전략·신속한 분석·정확한 결과!

약 한 달 간의 짧은 파일럿 프로젝트 기간이 주어졌음에도 우리 조는 단기간(속도)에 성과를 내면서 우수조(질)라는 두 마리 토끼를 잡았다. 조원들의 지혜로운 판단력과 이에 따른 선별적 전략으로 처음부터 잘 짜여진 베스트 플랜이 도출된 덕분이다. 조원 모두가 맡은 역할을 충실히 해주었기에 가능한 일이었다.

2017.11.10 달같이 먼저냐 닭이 먼저냐!

주제를 선정할 때 프로젝트 기간을 감안해 쉽게 확보할 수 있는 데이터세트를 기준으로 분석 주제를 정했다. 초반에는 야구, 부동산 등 여러 가지 아이템들이 나왔지만 양질의 데이터 확보가 어려웠다. 데이터를 보고서 주제를 선정해도 또 추가로 데이터 수집이 필요할 수밖에 없었다. 결국 일주일의 주제를 선정에 써버렸다. 얼마 후, 조원 중 한 명이 '학교알리미' 데이터세트로 해 보자는 제안을 했다. 이 데이터는 굳이 다른 데이터를 추가로 선정할 필요가 없을 정도로 다양했다. 모든 조원들의 찬성 아래 해당 데이터를 활용하기에 이르렀다.

2017.10.17 데이터를 어떻게 활용할 것인가

우리가 선택한 '학교알리미' 데이터세트는 종류가 매우 많다. 약 30여 개의 엑셀 파일은 연도별로 구분한다면 대략 100개의 엑셀 파일로 세분화할 수 있는 수준이었다. 그 중 어떤 것을 선택·분석해야 할지 막막했다. 이때 전체 데이터를 하나로 통합해 보자는 의견이 나왔다. 통합 작업은 변수 종류들이 모두 달라 힘들었다. 이를 황재호 조원의 주도로 성공적으로 해냈다.

2017.10.25 전처리 과정을 줄여 분석에 활용

데이터 수집과 전처리 과정이 비교적 수월하게 끝났다. 데이터 분석 과정은 생각보다 커다란 어려움이 없었다. 특히 분석 과정은 권도진 조원이 큰 힘을 발휘했다. 어떤 알고리즘을 사용하고 어떤 패키지를 사용할지 다들 머리를 맞대고 있는 동안 이미 통합 데이터세트를 갖고 이것저것 혼자서 분석을 하고 있었다. 여기에 다른 조원들의 의견이 더해지면서 거의 완성 단계까지 끌어올릴 수 있었다.

2017.11.05 다시 한번 의견 취합 후 자신감 있는 발표

발표 며칠을 앞두고 조원들과 결과에 대한 의견을 주고받고 정리했다. 실제로 만나서 확인해 봤더니 수정할 곳이 꽤 있었다. 1조였기에 가장 먼저 발표해야 했다. 마음이 촉박해졌다. 다행히 발표 전에 발표 자료를 잘 수정할 수 있었다. 프로젝트 전 과정을 총괄했던 조장 입장에서는 여유와 자신감을 갖고 발표할 수 있었다. 그 결과 '우수조'라는 행복한 타이틀을 거머쥔 수 있었다. 발표를 마치던 순간 너무나 뿌듯했다. 우수조라는 발표가 났을 때 조원들 모두가 기뻐서 서로를 축하해 줬던 기쁜 경험은 두고두고 힘이 될 것이다.

“당연한 것을 끄집어낼 수 있는 힘”

최윤지

SD경제연구소 금융사업본부 이사



분석 프로젝트 우수조로 선정된 소감은.

기쁘다. 프로젝트 조원들 중에 가장 나이가 어렸음에도 ‘조장’을 맡게 되어 처음에는 제대로 할 수 있을지 두려웠다. 프로젝트 조장으로서 가장 중시했던 부분은 조원별 의견과 진행중인 사항을 잘 종합해 전달하고 중재하는 것이었다. 중간에 정리가 제대로 이뤄지지 않았다면 아마 각자 생각하고 있는 목표가 다를 수 있었을 것이다. 결국 속도를 낼 수 없으면 모두가 불안해질 수밖에 없다. 설사 결과가 나왔다 하더라도 힘이 없는 결과가 되지 않았을까. 조원들 모두 성실히 임해준 덕분에 무사히 프로젝트를 마칠 수 있었고, 우수조로 선정되는 행운까지 얻을 수 있었다. 리더 역할에 대해 다시 생각해보 수 있는 기회였다. 앞으로 어떤 프로젝트의 리더가 된다 해도 두려움보다는 설렘으로 반길 수 있을 것 같다. 개인적으로 회사의 개발 업무를 총괄하는 위치에 있다 보니, 미래의 우리 세상이 어떻게 될까?를 놓고 끊임 없이 생각해야 한다. 성장 잠재력이 무한한 빅데이터와 인공지능 시장이 어떻게 발전해 나갈지에 대한 큰 그림은 그려볼 수 있게 되었다는 점이 무엇보다 도움이 됐다.

주제가 고교 선택을 위한 분석이었는데,

분석 전 예측과 분석 결과를 비교해 본다면.

정말 놀라울 정도로 달랐다. 우리 조가 처음에 이 주제를 도출했을 때만 해도 사실 ‘학생들의 도서관 이용현황’ 혹은 ‘학생들의 수업 양’ 등과 같이 ‘학생들의 교육 참여도’가 고교 진학에 크게 영향을

끼칠 것으로 예측했다. 그러나 분석 결과는 ‘학교가 학생들 교육의 질 향상을 위해 얼마나 투자했느냐’가 진학률에 큰 영향을 미치는 것으로 나왔다. 이 결과를 보았을 때 ‘신기하다’기보다는 ‘왜 이걸 몰랐지?’ 하는 생각이 들었다. 뭔가 당연한 결과라는 생각이 들었다. 하지만 이 당연함을 데이터 분석 프로젝트 과정을 거치기 전까지는 머리로 끄집어낼 수 없었다. 막상 생각을 했다고 해도 주관적임으로 설득력 있는 자료로 활용하기 힘들었을 것이다. 결국 데이터 분석의 필요성이 다시 한 번 드러난 셈이다.

프로젝트 과정에서 가장 어려웠던

순간과 그 해결 방법은.

가장 어려웠던 순간은 주제 선정이었다. 프로젝트 기간이 4주로서 짧았다. 조원들은 데이터 ‘수집’보다 ‘분석’에 더 강했기에 데이터를 어떻게 수집해야 할지도 막막했다. 머리를 맞댄 끝에 수집 과정을 거치지 않고 공공데이터와 같이 우리가 쉽게 구할 수 있는 데이터세트를 갖고 접근할 수 있는 주제로하기로 했다. 이렇게 생각하니 다음은 수월해졌다. 이용이 편리한 데이터세트를 찾기만 하면 됐다. 물론 중간에 어떤 데이터를 찾아야 할지, 즉 닙이 먼저나 달같이 먼저나 갈림길에서 망설이기도 했다. 하지만 결국 한 조원이 ‘학교알리미’라는 양질의 공공데이터를 추천해주면서 고민이 풀렸다. 이미 만들어 놓은 데이터세트를 중심으로 주제를 선정했으므로 데이터 수집 기간이 대폭 줄어들었다. 다른 조에서

데이터를 수집하고 있을 동안 데이터 전처리를 마칠 수 있었다. 프로젝트 기간이 길었다면 어렵더라도 데이터 수집 경험을 갖고 싶었다. 아쉽지만 이렇게 짧은 기간 프로젝트를 진행하다 보니 추후 어떤 급한 프로젝트를 맡는다 해도 단기간 프로젝트에 대한 자신감마저 들었다. 걱정보다는 할 수 있다는 자신감을 준 파일럿 프로젝트였다.

어떤 일을 하고 있고, 빅데이터 아카데미

분석 전문가 과정에 지원한 이유는.

증권과 파생상품에 대한 ‘로보 어드바이저’를 개발하는 SD경제연구소에서 일하고 있다. 로보어드바이저는 투자자들의 투자를 가이드해주는 프로그램이다. 과거에는 로봇보다는 사람들의 경험과 주관적인 판단 아래 매매했다. 하지만 이세들과 팔과 간 바둑 대결 이후 로봇의 정확성에 대해 더욱 신뢰도가 올라갔다. 이에 따라 SD경제연구소 역시 더 주목을 받게 되었다. 목표는 증권과 파생상품 투자에 있어 더욱 높은 승률을 얻고자 함에 있다. 일터에서는 데이터 분석 기술력을 접목해 승률을 더 높이려고 한다. 이를 위해 해당 기술에 대한 정확한 지식이 우선되어야 한다. 여러 책을 봤을 때 큰 흐름은 이해해도 뭔가 할 수 있겠다는 자신감이 들지 않았다. 실무에 접목해야 하므로 빅데이터의 실무를 알 필요성이 있었다. 마침 지인의 소개로 빅데이터 아카데미를 알고 선택했다. 너무나 도움이 되는 조원이었다.

빅데이터 아카데미 수강 후에

달라진 점이 있다면.

책에서 배웠던 빅데이터에 대한 내용과 실전이 달라 교육받는 내내 당혹스러웠다. 2주간의 빠듯한 집체교육 중에 들었던 내용은 오히려 내 머릿속을 뒤죽박죽 흔들어 놓았다. 그만큼 책과 분석 실전이 달랐다.

기상 센서 데이터를 이용한 기후 정보 및 이상센서 알림 서비스



구분	빅데이터 기술, 실시간 처리, 기계학습
적용 도구	Hadoop, Flume, Kafka, Storm, Hbase, Redis, Esper, Spark, Hive, ImpalaOozie, Zeppelin, AWS KINESIS, ElasticSearch, LogStash, Kibana
수집 데이터	국립재난안전연구원의 기상 데이터
산출물	기상 센서 데이터를 이용한 기후 데이터 수집 시스템
지도	이상훈
참여자	이동열 ^{조장} , 김경덕, 김재명, 안순희, 이인섭, 정인규
프로젝트 개요	국립재난안전연구원의 기상 데이터를 활용해 기후 정보 및 이상 센서 알림 서비스를 구축하는 프로젝트다.

THE CHALLENGES

빅데이터 기술 전문가 과정 강의를 들으며 학생 때 처음 배웠던 포트란이라는 프로그래밍 언어가 생각났다. 그 시절 낯설었던 느낌이 다시 떠올랐고, 어딘가 익숙한 모습에 반갑기도 했지만, 왜 우리는 각자의 도메인에서 빅데이터라는 영역에 대해 이 낯선 느낌을 극복하고 도입하려는지에 대한 의문이 생겼다.

빅데이터 관련 기술은 어느덧 신기술이 아닌 어찌 보면 우리들이 부르는 레거시 시스템의 단계로 넘어가고 있는 것은 아닐까. 신기술을 배우고자 한다면 요즘 핫키워드인 인공지능(AI), 머신러닝, 자율주행 자동차 같은 것을 접해야 했을 텐데 왜 빅데이터 기술을 선택했을까?

기술 전문가 과정 13기 참가자 모두가 동일하지는 않았겠지만, 빅데이터가 앞서 소개한 인공지능 등의 근본이 된다. 따라서 빅데이터의 기초를 제대로 다져야 인공지능 등으로 나아갈 채널이 열리지 않을까 하는 생각으로 빅데이터 아카데미 기술 전문가 과정을 노크했다.

파일럿 프로젝트를 준비하면서 분석에 초점이 맞추어져 있다는 생각이 많이 들었다. 상대적으로 데이터를 확보·정제·가공·시각화하는 부분은 가려져 있지 않나 싶었다. 이로부터 파일럿 프로젝트의 아이디어가 떠올랐다. 빅데이터 분석을 하기 위한 일련의 과정에서 분석 파트의 중요성을 간과할 수 없지만, 빅데이터 비즈니스가 분석 하나만으로 존재할 수 없다. 눈에 띄지 않지만 뒤에서 서포트하는 시스템을 알아보고 싶었다. 분석 플랫폼 구축과 데이터 수집·정제·적재·가공·분석·시각화 등 빅데이터 분석을 구성하는 일련의 과정에 대한 이해를 파일럿 프로젝트의 목표로 잡았다. 데이터 분석보다는 서비스 흐름 상의 관련된 기술을 다뤄보되, 조원 모두가 참여해야 한다고 조원들 간에 의견 일치를 보았다.

이 목표를 달성하기에 적합한 주제는 무엇일까? 하고 고민하던 중 기술 이해가 목적이라면 기상 데이터로도 문제가 없을 것이라는 의견도 나왔다. 이때 프로젝트 결과가 현실에 유용하게 활용될 수 있으면 좋겠다는 생각에 데이터를 찾아 나섰다. 안전재난연구원에서 일하는 조원이 앞장서 공공데이터포털에 공개된 기상청의 센서 데이터를 확보했다. 데이터가 이미 공개된 점이 아쉬웠다. 하지만 이것 또한 기회가 되지 않을까 싶었다. 선행 분석 결과와 우리 조의

분석 결과를 비교할 수 있겠다는 생각에 이르렀다.

파일럿 프로젝트 주제가 정해졌다. 남은 건 어떤 모습으로 만들지를 생각해야 한다. 이에 필수 주제 3개와 부가 주제 3개를 도출해 직접 해보기로 했다.

필수 주제	부가 주제
<ul style="list-style-type: none"> - 빅데이터 에코 시스템의 다양한 기술 사용 - 기상 데이터 수집·가공·적재, 실시간 데이터 처리, 이상 데이터 검출 - 기본 분석과 시각화 	<ul style="list-style-type: none"> - 머신러닝 기반 이상 센서 알림 - 실시간 모니터링 기술 - ELK를 이용한 시각화 구현

표 1

필수 주제와
부가 주제

THE APPROACH

프로젝트를 시작하기 전에 미팅을 갖고 주제와 방향을 정했다. 역할을 구분하고 조원별로 담당 업무를 배정하려 했다. 하지만 전반적인 이해만으로는 현업에 돌아가서 실무에 적용할 수 있겠느냐? 하는 의견이 나왔다. 이에 우리 조는 본 프로젝트가 진행되기 전에 K-ICT 빅데이터센터에 별도의 VM 서버를 팀원별로 신청·발급 받았다. 1주차 멘토링 시간에 팀원별로 구성된 VM 서버에 빅데이터 환경을 구축·테스트했다. 이 과정에서 지정 학습서를 참고해 2주 간의 집체교육 시간에 배웠던 내용들을 조금씩 내재화할 수 있었다.

멘토와 상의해 결가지들을 좀 더 정리했다. 하지만 멘토링을 받으면서 우리 조의 파일럿 프로젝트 범위가 만만찮다는 것을 알게 됐다. 그래도 할 수 있는 한 많은 것을 기술적으로 진행해 보자는 생각은 변함이 없었다. 이에 따라 멘토링이 진행될 때마다 다양한 기술적 접근에 대한 의견을 멘토에게 제시해 검증 받으며 세부적으로 구현해 나갔다.

필수 주제 영역

빅데이터 에코 시스템의 다양한 기술 사용

빅데이터 분석 플랫폼을 구성하는 다양한 기술 습득이 목표였으므로 하둡 에코 시스템의 구성 요소별 설치에 너무 많은 시간을 할당할 수 없었다. 그래서 분석 플랫폼 구축에 딱 적합한 클라우드라가 떠올랐다. 하둡 배포판인 클라우데라를 이용해 기본적인 빅데이터 환경을 빠르게 구축할 수 있었다. 하지만 시스템 설치가 간단하지 않았기에 관련 설정 정보들을 공유하며 설치를 완료할 수 있었다.

기상 데이터 수집·가공·적재

기상청의 데이터를 안전재난연구원을 통해 배치 또는 실시간으로 수집하려 했다. 보안상의 이슈로 해당 방법이 지원되지 않아 기상청의 기상 센서 데이터의 분단위 데이터를 수집하되, 임의의 시점에 배치용 데이터로 2016년에서 2017년까지 1년분의 데이터를 가져왔다. 실시간 데이터 처리를 위해 2017년 1~3월까지 3개월분의 데이터를 미리 내려 받아 프로젝트 파일서버에 배치용 데이터와 실시간용 데이터를 분리·저장했다.

배치 데이터 처리

저장된 파일서버에서 배치 데이터는 일일 배치 스크립트를 이용해 배치용 플럼(FLUME)이 바라고 보고 있는 경로에 적재했다. 이때 플럼(FLUME)은 해당 경로의 파일 생성 이벤트를 탐지해 파일을 읽어드린 후에 하둡파일시스템(HDFS)에 저장한다. 파일의 양이 적지 않았으므로 하둡 적재 시 파티션돼 저장될 수 있도록 플럼 인터셉터를 구현해 다음과 같은 구조로 파일이 저장될 수 있도록 했다.

폴더 구조의 예: /pilot-prj/log/year=2017/month=01/day=01/

실시간 데이터 처리

실시간 데이터 처리를 위해 TX, RX 실시간 전송 모듈을 개발했다. 실시간 데

이터 폴더에서 데이터를 읽어와 분 단위로 메시지를 플럼에 전달하도록 구현했다. FLUME SINK로 카프카(Kafka)를 지정해 실시간 데이터가 상당량 이상 발생했을 때 버퍼링할 수 있도록 했다. 이를 다시 STORM spout에 전달해 실시간 분석 과정을 수행하게 했다. Spout에서 Esper BOLT로 전달된 실시간 데이터는 Esper에 등록한 룰(5분간의 연속 데이터를 비교해 온도차가 1도 이상인 경우)에서 벗어나는 데이터가 들어오면 이상치(Outlier)로 판정한다. 이를 REDIS BOLT로 전달해 REDIS DB에 저장한다. 정상인 경우에는 HBASE에 저장하도록 했다.

분석 사전 작업

앞 과정으로 배치와 실시간 데이터는 HDFS, HBASE, REDIS에 저장됐다. 이 데이터들을 분석하기 위해 우리 조는 HUE라는 웹 인터페이스를 사용했다. HUE 웹 사용자 인터페이스는 하이버(HIVE), 임팔라(Impala) 등과 같이 빅데이터를 일반 RDBMS처럼 보여주거나, 데이터를 일부 시각화해 보여주는 기능을 제공한다. 하이버에서 데이터를 탐색하기 위한 메타 정의 작업을 하고, 해당 메타에 실제 HDFS 경로와 사상(MAPPING) 과정을 진행했다.

데이터 탐색과 분석

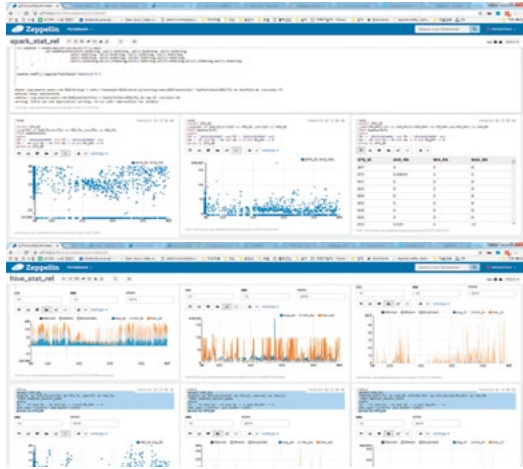
하이버에 정의된 메타 테이블로 데이터를 테이블 형태로 조회할 수 있다. 하지만 이때 실제 하둡 맵리듀스가 수행돼 하이버 결과가 도출되기까지는 요청 범위에 따라 수분 이상이 걸리는 경우가 대부분이다. 따라서 우리 조는 탐색 효율을 높이 위해 임팔라를 사용했다. 임팔라로 기상 데이터를 일별 평균 온도·강수량·풍속 정보와 각각의 최대값/최소값을 구했다. 마찬가지로 월별 데이터에 대한 정보도 확인했다.

데이터 시각화

하이버와 임팔라에서 탐색한 데이터를 시각화하는 과정이다. 이때 제플린(Zeppelin)을 사용했으며 일별·월별 평균 온도, 강수량, 풍속 데이터를 시각화 차트(라인차트, 바차트, 분포차트)로 표현했다. 추가로 스파크(Spark)로 HDFS에 적재된 데이터를 로딩해 분석한 정보를 메모리에 올려놓고, 필요 시 조회해 시

그림 1

제플린 기반의 데이터 시각화



각화하는 방식으로도 결과를 도출했다.

부가 주제 영역

필수 주제를 완료했지만, 눈과 손으로 확인하고 싶은 기술들을 그냥 둔 채 마무리 지을 수 없었다. 지금까지의 기술보다 난이도가 높아져서 완수하지 못하는 부분도 있을 거라는 두려움도 없지 않았다. 그러나 결과 도출이 우리 조의 목표가 아니었다. 따라서 완료하지 못해도 그 수행 기록을 나눔이 의미 있을 것이라는 생각으로 남은 부가 주제를 향해 나아갔다.

머신러닝 기반 이상 센서 알림

집체교육 과정에서 배운 머신러닝을 파일럿 프로젝트에 꼭 접목해보고 싶었다. 하지만 새로운 프로그래밍 언어와 수학 지식, 알고리즘 등 진입장벽이 만만치 않았다. 멘토의 많은 조언으로 케이스 스터디 수준으로는 진행해 볼 수 있겠다는 생각이 들었다.

1년치의 기상 데이터로 기본적인 조건은 갖추고 있었으므로 이상기후를 예측하는 서비스를 구축하기로 했다. 하지만 시계열 분석을 하기 위해서는 시계열 데이터도 있어야 했다. 1년치 데이터만으로는 데이터가 부족했다. 더불어 특정 기간의 이상기후 데이터를 라벨링할 수 있어야 했다. 하지만 현실적으로 관련 데이터를 입수해 기존 데이터에 사상한다는 것은 현실적으로 어려웠다.

1년치의 데이터와 라벨된 데이터 없이도 가능한 K평균값(K-Means) 클러스터링 알고리즘 기반으로 이상 센서를 분류해내기로 했다. K 값을 정하기 위해 2016년 12월 센서별 평균 데이터 분포도를 보면서 평균 온도 그룹이 정상 2개와 오류 1개로 나뉠 수 있다고 가정해 K값을 3으로 정했다. 정해진 K값으로 머신러닝을 하기 위한 데이터 준비 차원에서 벡터화와 실수화 작업을 했다. 이 과정을 거쳐 센서 데이터를 가공한 후 판정에 불필요한 데이터(센서 ID, 수집 시간 등)를 벡터 내에서 제거하고, 이를 K평균값 모델로 수행했다.

도출된 모델값과 실제 데이터를 비교해 클러스터 중심 거리와의 노드 거

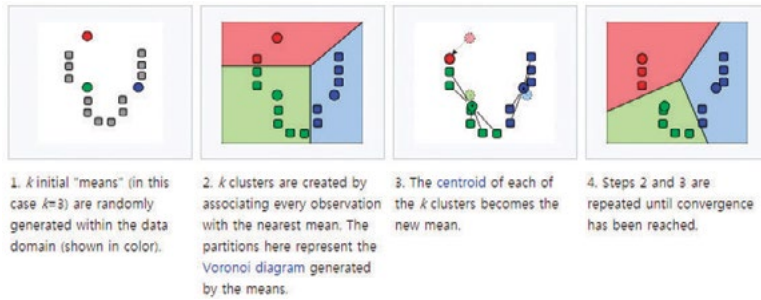


그림 2

K평균값 표준 알고리즘의 예
(출처: en.wikipedia.org/wiki/K-means_clustering)

리가 일정거리 이상(프로젝트에서는 거리 순위 200이 넘는 경우)이면, 오류 센서로 판정하게 했다. 해당 기술은 Spark ml과 스칼라(Scala)로 구현했다. 도출된 오류센서 데이터는 HDFS에 저장해 하이브에서 확인할 수 있도록 했다.

실시간 모니터링

초기 시작은 Spark Stream을 이용해 실시간 데이터가 들어오면 제플린에서 시각화하는 모습으로 가려고 했다. 하지만 Spark Stream과 제플린 연동 이슈가 있었으므로 동일 기능을 제공하는 AWS 키네시스(Kinesis)를 사용해보기로 했다. AWS 키네시스로 데이터를 전달하기 위한 데이터 전송(delivery) 모듈을 구현했다. 이 모듈로서 실시간 데이터 파일 서버에서 AWS 키네시스로 실시간 데이터를 전달했다. AWS 키네시스에서 버퍼링되는 데이터를 수신하도록 제플린에서 Spark interpreter를 사용해 데이터 수신부를 구현했다. 최종적으로 제플린에서 데이터를 시각화하는 형태로 구현했다.

ELK를 이용한 시각화

빅데이터 시각화의 꽃이라고 할 수 있는 파워풀한 시각정보를 제공하는 ‘키바나’. 이 존재가 우리 조에게도 무척 매력적이었다. 다만 일래스틱서치(Elastic search)라는 만만치 않은 산이 있음을 실감할 수 있었다. logstash를 이용해 시각화할 데이터를 수집한다. 이 데이터를 일래스틱 서치에 전달하면 데이터를 인덱싱해 최종적으로 키바나로 전달해 시각화하도록 구현했다. 하지만 인덱싱 처리 부분에 오류가 있어 넘겨진 데이터를 키바나가 처리하지 못해 시각화를 마무리하지 못했다. 처음에 도커 이미지를 사용해 구축하려 했으나, 도커 이미지를 사용하기 위해서는 centos 7이상을 사용해야 한다. 프로젝트에서

사용한 VM은 6.9라서 사용할 수 없다는 걸 알았을 때는 많은 시간이 흘러간 후였다. 실제로 ELK를 각각 설치하고 프로젝트를 진행한 시간이 넉넉하지 못했던 점이 아쉽다.

THE OUTCOME

필수 주제와 옵션 주제로 양분해 진행한 빅데이터 에코시스템에 대한 전반적인 기술 탐구의 종착역에 도착했다. 기반 인프라 구성으로부터 시작해 수집·실시간 분석·적재·탐색·분석·시각화까지 데이터의 흐름 상에 놓여 있는 기술들을 두루두루 다루어 볼 수 있었다. 그 과정에서 도출된 문제들을 멘토와 조원들이 함께 풀어간 경험이 좋은 기억으로 남아 있다. 옵션 주제를 수행하면서 상용 서비스 관점으로 접근해 머신러닝 기반으로 오류 센서를 검출하거나, 실시간 데이터 흐름을 시각화하는 것, ELK를 도입해본 것은 기본 영역에서만 머물지 않고 실제 서비스 영역으로 넘어가보고자 했던 즐거운 도전이었다.

기술 전문가 과정 13기 4조는 시작부터 일관되게 기술적인 접근을 했다. 그 영역을 점차 넓혀가는 방식으로 프로젝트를 진행했다. 요소별 욕심나는 부분이 없지 않았다. 하지만 하나에 몰입해 전체적인 흐름을 잃어 버리기보다는 파트별로 적절하게 리소스를 배분해 진행한 점, 전 조원이 본인 영역뿐 아니라 주변 동료 영역에 관심을 갖고 진행한 점, 결과보다는 과정에 충실했던 점들은 짧은 시간에 빅데이터 에코 시스템의 역할과 흐름의 이해뿐 아니라 집체교육 중 들었던 내용의 이해를 가능하게 했다.

기술 이해에 초점을 맞추고 프로젝트 진행

우리 조는 3T(1T: 분석보다는 테크닉/ 1T: 빅데이터 에코 시스템 테크닉/ 1T: 함께하는 테크닉)에 집중하며 파일럿 프로젝트를 진행했다. 빅데이터 에코 시스템을 기술 관점에서 접근하며 다음과 같은 터닝 포인트를 거처왔다.

2017.3.25 주제 선정

조별 첫 미팅 후, 조원 각자가 준비해온 주제를 갖고 다시 모였다. 여러 의견이 나왔으나 정인규 조원이 기상 센서 데이터 확보가 가능하다는 의견에 관심이 집중됐다. 기상 센서 데이터를 이용한 기후 정보 및 이상 센서 알림 서비스를 주제로 선정했다.

2017.4.1 아! 보안

주제가 선정되고 데이터까지 확보돼 순조롭게 진행되려던 참이었다. 보안 이슈가 프로젝트 진행을 가로막는 복병으로 등장했다. 안전재난연구원을 소스로 배치해 실시간 데이터를 전달 받으려 했었으나 보안상 이슈로 FTP, SCP, API 방식 모두가 어렵게 됐다. 차선으로 약 1년 치의 데이터를 별도로 내려 받아 프로젝트 파일서버에 구축하는 일과, 데이터 전달을 위한 별도의 스크립트와 소스 개발이 필요해졌다.

2017.4.8 옵션 주제들이 괴롭히다

머신러닝 기반 오류 센서 검출을 위해 수행한 프로그램이 1일이 넘어갔는데도 결과를 보여주지 않았다. Docker, 독하다 독해! 왜 버전 차별을 하는 거니(centos 6.9에서 도커 이미지 사용이 안 된다). Spark stream과 제플린도 특별한 이유 없이 대답이 없다.

2017.4.15 결과들이 나오다

멘토의 의견으로 스파크가 분산 모드에서 운영되는지 확인해 보았다. 입력 벡터를 줄여서 적용한 결과 1시간 30분만에 결과가 똑딱 나왔다. Docker만 있는 게 아니다. 실제 바이너리를 설치하면 될 뿐이었다.

“빅데이터 이해의 안개가 걷혔다”



이동열

티온미디어 연구소장

우수조가 될 수 있었던 힘은

무엇이라고 생각하나.

집체교육을 받으면서 기술 전문가 과정
13기 동기들 모두 열정이 대단하다고
느꼈다. 다만 우리 조가 조금 더 운이
좋았던 조원들에게 실례가 될지(^^).
실무에 적응을 염두에 두고 프로젝트를
수행했던 점과 하나하나 과정을 팀원 모두가
참여했다는 점이 좋은 결과로 이어졌다.

1위를 할 수 있을 것이라는

생각이 들었을 때는 언제인가.

1위를 했으면 좋겠다는 생각은 처음부터 할
수밖에 없지 않나^^ 어느 시점인가 1위에는
별 관심이 없어졌다. 파일럿 프로젝트 발표
자리에서 조원들 얼굴을 봤었는데, 밝게
웃고 있어서 기분이 좋았던 기억이다.

빅데이터 기술 전문가도 분석과

컨설팅까지 알아야 한다고 강조한 것이

인상적이었다. 분석 전문가 과정이 아닌

기술 전문가 과정을 택한 특별한 이유가 있나.

두 과정 모두 듣고 싶었다. 둘 중 하나를
선택해야만 했다. 순서상 기술을 먼저
이해하고 다음에 분석을 하면 좋을 것
같다는 생각으로 신청했다. 올해 기술
전문가 수료생을 위한 분석 특강이 준비돼
있다고 들었다. 꼭 들어볼 계획이다.

프로젝트를 플랫폼 구축 △데이터

수집·전처리 △분석·시각화로 구분했을

때, 각 영역에 어느 정도의 비중을 두고

했으며 어느 과정이 가장 어려웠다.

이상적으로는 모든 파트에 골고루 리소스를
배분하려고 했다. 실제로는 2:2:2:3 비율이
되지 않았나 한다. 마지막 3은 옵션 주제
부분에 대한 비율이다. 부가 주제(머신러닝,
실시간 데이터 모니터링, ELK)와 실시간
분석의 ESPER 구현이 어려웠다.

하둡 에코 시스템의 모든 것을

경험해보겠다고 했는데 경험해 본 소감은.

집체교육 강의를 들었을 때의 막막했던
느낌이 걷힌 느낌이다. 아직 못 다룬
기술들이 있고 새로운 기술이 더해지겠지만,
빅데이터 기반 프로젝트를 수행할 때
무엇을 시작하고 해야 하는지가 머릿속에
이미지화 됐다. 물론 실제 구현할 때는
또다른 문제를 만나게 될 것이다.

프로젝트를 진행하면서

가장 기억에 남았던 점은.

조원들의 커뮤니케이션 채널로 '잔디'라는
프로그램을 사용했다. 각자가 맡았던
작업들을 잔디 채널로 공유했다. 작업
내용뿐만 아니라 세미나 정보, 관련
기술 교육정보 등도 공유하는 모습을
보면서 자극을 받았던 게 기억난다.

기술 프로젝트를 진행할

후배들에게 조언을 한다면.

개인적으로 아쉬웠던 점은 빅데이터
아카데미 교육이 시작되기 전에 준비
학습을 좀 더 하고 왔었으면 하는 것이었다.
새로운 것들이 너무 많아 이해하고
적용하기에 시간이 많이 부족했다.

향후 계획이나 바라는 바가 있다면.

강의 시간과 파일럿 프로젝트에서 배운
내용을 기반으로 회사에 빅데이터 분석
시스템을 도입해 보는 것이다. ELK 기반
데이터 탐색 시스템을 구축해보고 싶다.

농구 국가대표 톱5 선발



구분	빅데이터 기술
적용 도구	HDP, 스프링부트, JSOUP, MySQL, Grafana
수집 데이터	한국프로농구연맹(KBL)의 2016~2017 시즌 선수 및 경기 데이터
산출물	HDP 설정 서버, 프로농구 우수 선수 분석결과 데이터 등
지도	이성윤
참여자	이동하 ^{주장} , 강민호, 김지훈, 김태형, 박상현, 이영호
프로젝트 개요	2016~2017 시즌 동안 10개의 남자 프로농구 팀에 소속되어 총 540 경기를 치렀던 173명의 선수 데이터를 분석해 우수 선수를 선정했다. 경기 시간, 필드골, 야투 성공률, 자유투 성공률, 수비 리바운드 등 여러 기록 데이터를 수집·분석해 공헌도 평가라는 기준을 세우고 '농구선수 톱5'를 선정했다.

THE CHALLENGES

집체 교육 이후 주어진 파일럿 프로젝트 일정은 9월말부터 11월 초까지 5주간의 주말 일정을 고스란히 바쳐야 하는 강행군이었다.

기술 습득 목적에 부합한 주제 선택

프로젝트 주제를 발굴하는 과정은 두 번의 시행착오 끝에 ‘농구 국가대표 톱5 선발(이하 농구선수 톱5)’라는 주제로 확정했다. 프로젝트를 진행하면서 각 조원의 기술 경험에 기초해 업무를 분장함으로써 결과물을 도출할 수 있었다.

주제 선택을 위해 토의하면서 나온 의견은 공공데이터를 활용하는 방안이었다. 공공데이터를 활용한 비즈니스가 언론에 자주 회자되고 있다는 점을 들어 우리 조도 공공데이터에서 뭔가를 도출해 보자는 의견이 나왔다. 그래서 처음 나온 주제는 서울시의 대기환경 데이터를 분석해 ‘서울에서 살기 좋은 동네’를 도출하는 분석시스템 구축이었다. 하지만 이 주제는 너무 많은 결정 사항을 요구했다. 공기의 질만으로 ‘살기 좋다’고 정의하는 것이 과연 의미가 있나? 빅데이터 기술 전문가 과정에서는 기술 이해 관점에서 접근해야 할 필요가 있었는데, 이 주제는 기술 전문가 프로젝트에는 부합하지 않다는 의견이 나왔다.

두 번째 나온 주제는 스포츠 분야에 관심이 많은 조원이 ‘축구 국가대표 베스트 11’를 선정하는 시스템을 구축해 보자고 제안했다. 이 주제는 웹 크롤링에 의한 데이터 수집이 더 적절해 보였다. 하지만 데이터 원천인 ‘대한축구협회(KFA) 사이트’로부터 데이터를 수집할 때 기술적인 난점이 발생했다. 그 문제를 해결하려면 많은 시간이 소요되고 해결하기도 쉽지 않아 보였다. 중요한 것은 데이터 수집에 많은 시간을 소요하게 되면 이후 진행이 어렵다는 것이었다. 결국 대한축구협회 사이트에서 프로젝트 기간 내에 원하는 데이터를 확보하기 어렵다고 판단했다. 데이터 수집이 용이한 주제를 선정해 데이터를 우선 수집한 후에 하둡의 다른 기술적 요소들을 적용하기로 결정했다. 그래서 이에 따라 상대적으로 쉽게 데이터를 쉽게 수집할 수 있는 ‘농구선수 톱5’를 최종 프로젝트 주제로 선정했다.

THE APPROACH

빅데이터 시스템 설계 시에 고민됐던 부분은 ‘아파치 하둡 에코시스템’ 선택이었다. 이미 안정화된 클라우데라(Cloudera) 또는 호튼웍스(Hortonworks) 배포판 중에 하나를 선택하고자 했다. ‘하둡 생태계와 하둡 사용자를 구축하고 오픈소스 코드를 발전시키겠다’는 호튼웍스의 기업 정신에 공감해 ‘호튼웍스(Hortonworks) 배포판’을 최종 선택했다. 물론 기술 전문가 과정이라서 하둡 배포판이 아닌, HBase 등 하둡 에코시스템 구성 요소를 하나씩 설치하면서 확인하는 방법도 있었다. 하지만 하나씩 설치하는 인프라 구축보다 활용 관점을 고려해 배포판을 설치하기로 했다. 각각의 구성 요소를 일일이 설치해 보는 것은 개별적으로 해 볼 수도 있다고 생각했다.

최근의 대용량 데이터 분석 기술 환경이 몇 년 전에 비해 보편화됐다. 설치에 많은 시간을 할애하는 대신에 분석을 하면서 하둡 에코시스템의 기술적 요소를 이해하는 것이 더 유리해 보였다. 설치에서 아낀 시간을 조원들이 모여서 함께하면 좋을 내용에 쓰기로 했다.

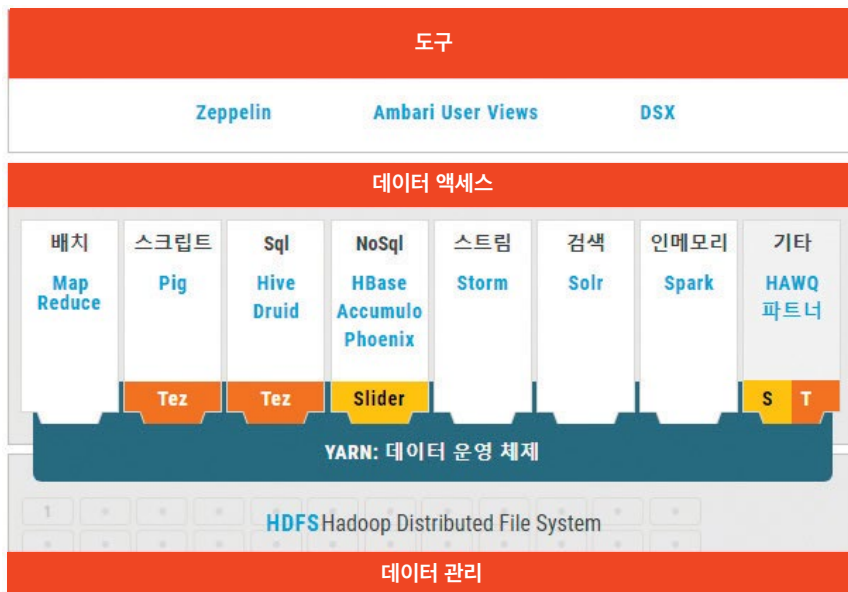


그림 1

하둡 에코시스템과
배포판

개발 서버 준비

파일럿 프로젝트 진행 시 사용하는 개발 서버는 여러 사람이 동시에 작업하기에는 어려운 환경이었다. 다양한 방법으로 네트워크 연결을 시도했지만 잘 되지 않았다. 이 과정이 길어지면서 분석을 위한 시스템 개발·배포·테스트에 이

그림 2

HDP ambari 대시보드



르는 환경 구성에 많은 시간을 쓰고 말았다. 결국 시스템 엔지니어로 일하는 조원이 주도해 개발 서버를 직접 구축·사용함으로써 앞에서 너무 많이 써버린 시간을 어느 정도 만회했다. 호튼웍스 배포판은 HDP 2.5.3 버전을 설치했으며, 설치에는 특별한 어려움이 없었다. 호튼웍스 배포판은 빅데이터 처리에 필요한 하둡에 코시스템의 대부분을 포함하고 있고 웹 UI 환경에서 관리할 수 있다.

개발한 웹 크롤러로 데이터를 수집해 HDFS에 저장

한국농구연맹 사이트(www.kbl.or.kr)에 등록된 10개 프로팀의 선수 및 경기 결과 데이터가 있었다. 경기 결과 데이터에는 2016~2017년 2년 동안의 총 540 경기에 대한 선수별 기록이 포함돼 있었다. 오픈소스를 사용해 다양한 소스로부터 관련 데이터를 수집할까 고민했다. 하지만 KBL 경기결과 데이터 수집에는 적합하지 않다는 판단에 따라 우리 프로젝트에 적합한 크롤링 프로그램을 직접 만들었다. 프로젝트 수행 기간이 길지 않아 빠르게 만들어 개발하기로 했다. Spring Boot, HttpClient를 이용해 웹 크롤링 배치 프로그램을 만들었다.

그림 3

하이브를 통한
분석결과 도출



Jsoup을 이용해 변환 과정을 거쳐 바로 하둡 파일 시스템(HDFS)에 저장했다. 크롤링 배치 프로그램을 직접 개발해 하나의 프로그램에서 데이터 추출·변환·적재라는 절차적 단계를 모두 처리하면서 원하는 데이터를 수집·적재할 수 있었다. 하이브(HIVE)를 이용해 HDFS에 수집·적재된 데이터를 처리한 후 시각화를 위해 MySQL DB에 저장했다.

공헌도 평가를 기준으로 톱5 선수 선정

2016~2017 시즌에 10개 팀 173명의 선수가 치른 총 540 경기 데이터에는 경기시간, 필드골, 야투 성공률, 자유투 성공율, 수비 리바운드 등의 기록 데이터가 포함돼 있었다. 데이터 분석은 해당 도메인에 대한 이해를 전제로 하는 것처럼, 데이터 분석 결과만으로 대표 선수를 선정하려면 프로농구에 대한 이해가 필수적이라는 생각이 들었다. 이에 우리 조는 ‘공헌도 평가’라는 기준을 세워 농구선수 톱5를 도출했다.

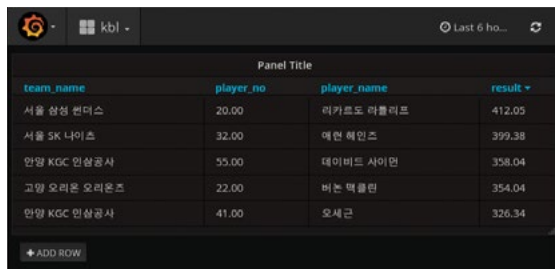
- 공헌도 평가 = 항목별 가산점 - 항목별 감점
- 항목별 가산점 = (득점 + 스틸 + 블록슛 + 수비 리바운드) × 1.0 + (공격 리바운드 + 어시스트 + 굿디펜스) × 1.5 + 출전시간(분, 초) ÷ 4
- 항목별 감점 = 턴오버(농구 경기에서 실책으로 인해 공격권 이전) × 1.5 + 2점슛 실패 × 1.0 + 3점슛 실패 × 0.9 + 자유투 실패 × 0.8

Grafana로 데이터 시각화

데이터 분석에서 시각화의 중요성은 늘 강조된다. 각종 시각화 패키지를 알아보았는데 Grafana 시각화 툴은 대용량 데이터 분석 시스템 이외에도 검색엔진 또는 다양한 데이터베이스와 연동이 가능했다. 우리조가 직접 개발한 웹크롤러에도 적용할 수 있었다. 최적화를 통해 데이터 수집 과정도 시각화해 관리할 수 있었다.

그림 4

HDFS에 저장된 수집 데이터



team_name	player_no	player_name	result
서울 삼성 썬더스	20.00	리카르도 라플리프	412.05
서울 SK 나이츠	32.00	마이클 헨리츠	399.38
안양 KGC 원상공사	55.00	데이비드 사이먼	358.04
고양 오리온 오리온즈	22.00	버논 맥클린	354.04
안양 KGC 원상공사	41.00	오세근	326.34

THE OUTCOME

빅데이터 기술 전문가 과정을 꼭 한번 수강하고 싶었다. 아니 그렇게 해야만 했다. 근래 수년간 빅데이터는 IT 환경 곳곳에서 필수 요소로 자리매김하고 있다. IT 분야의 개발자, DB 전문가, 시스템 관리자로서 십수 년을 근무해왔던 조원들이지만, 빅데이터 기술을 접하는 태도는 그야말로 초보자였다. 프로그래밍을 처음 접했을 때 두려움과 설렘 속에 어려움을 극복해 나갔던 순간들이 다시 떠올랐다. ‘농구선수 톱5’라는 주제를 선정해 데이터를 확보해 직접 분석하고 시각화까지 해보았다. 데이터 전처리는 프로그래머라면 크게 두려워할 필요가 없겠다는 결론에도 이르렀다. 자바나 MySQL 실력으로 원하는 데이터를 추출하고 간단한 분석 환경을 구축할 수 있기 때문이다.

우리 조는 기술 측면에서 프로젝트를 진행하면서 다음과 같은 결과를 도출했다. 빅데이터 시스템은 단위 기술보다는 수집·분석·가공·시각화의 요소별 모듈의 연계가 꼭 필요하다. 각 모듈들의 결합이 중요하다. 결합 시 여러 개의 라이브러리 중 필요한 것을 찾으려면 사전 학습과 벤치마킹도 필요하다. 어느 영역에서나 그러하듯이 협업 작업을 주로 해야 하는 IT 분야에서 의사소통은 프로젝트 성패를 가르는 요소가 된다. 오프라인 미팅이 쉽지 않은 상황에서는 다른 방법이 없었다. 네이버 밴드나 카카오톡, 행아웃 등의 커뮤니케이션 툴의 힘에 의존할 수밖에 없었다. 온라인 커뮤니케이션은 대면 소통에 비해 참가자들의 신뢰가 더 중요하다. 우리 조는 일단 정해진 기간 안에 파일럿 프로젝트를 완료함으로써 서로에 대한 신뢰가 있었음을 확인했으며, 앞으로의 삶의 여정에서도 좋은 경험을 했다.

유연한 대처는 데이터 분석에서도 필수

기술 전문가 과정 14기 1조는 시스템 엔지니어, 개발자, DBA 등 각 분야에서 10년 이상의 기술과 경험을 쌓아온 조원들로 구성됐다. 하지만 빅데이터 영역, 특히 분석 시스템 구축과 데이터 분석은 낯선 영역이었다.

2017.9.30 시행착오를 거쳐 주제 확정

공공데이터포털에는 여러 공공기관에서 공개한 다양한 데이터가 있다. 맨 처음 주제 선정에서 서울시 기상 데이터를 활용한 '서울에서 살기 좋은 곳'을 찾아보려고 했다. 그러나 너무 많은 선택 요건과 '살기 좋은'을 어떻게 정의해야 할지 등을 쉽게 도출하기 어려웠다. 방향전환 후 재선정한 '축구 국가대표 베스트 일레븐'을 분석하기 위한 데이터 수집에 웹 크롤링 방식을 적용했는데 마찬가지로 기술적인 문제점에 봉착했다. 결국 농구 국가대표 톱5 선발로 전환했다.

2017.10.14 개발 서버 준비로 많은 시간 허비

파일럿 프로젝트 진행 시 사용하는 개발 서버는 여러 사람이 동시에 작업하기에는 어려운 환경이었다. 많은 시간을 투자하면서 여러 시도를 했지만, 제안된 일정 안에 끝내야 할 분석 시스템 개발·배포·테스트용 시스템으로 적합하지 않다는 판단을 내렸다. 여기서 많은 시간을 써버리고 말았다. 이에 시스템 엔지니어 출신의 조원 주도로 개발 서버를 직접 구축해 시간을 단축할 수 있었다.

2017.10.16 자바로 데이터 수집 크롤러 개발

분석을 위한 데이터를 웹에서 얻으려면 웹 크롤링을 해야 한다. 교육과정에서 배운 파이썬 기반의 툴 외에도 개발자들에게 익숙한 자바 기반의 크롤러를 직접 개발했다.

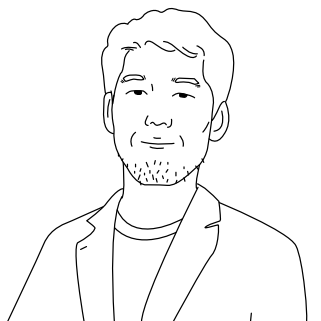
2017.10.30 쉽게 접근한 시각화

DBA로 활동하는 조원의 추천으로 교육과정에서 배우지 않았던 Grafana를 시각화 도구로 적용했다. 선행 학습 없이도 어렵지 않게 활용할 수 있었다. 새로운 기술 트렌드에 주의를 기울이고 있으면, 언제든지 새로운 시도를 할 수 있음을 깨달은 순간이었다.

“산에 오르려면 등산화라도 먼저 구입하라”

이동하

이씨플라자 기술연구소 부장



빅데이터 아카데미 수강 배경과

어떤 목표를 갖고 임했나.

요즘 IT 분야의 화두인 머신러닝과 인공지능 트렌드를 어떻게 받아들이야 할지 나름대로 막막했다. IT 분야는 늘 새로운 산을 올라가야 하는 분야이지 싶다. 데이터 분석과 인공지능이라는 새로운 산에 오르기 위한 등산화 한 켤레를 구입한다는 기분으로 도전했다. 등산화를 신어보고 싶어서라도 산에 오르지 않겠나! 커리큘럼을 비교하고 수료생들의 평판을 조사해 한국데이터진흥원의 '빅데이터 아카데미' 과정을 선택했다. 어느 기술을 습득하든지 그 과정에는 시간과 노력이 필요하다. 2주 온종일을 빅데이터 아카데미에서 교육을 받아야 하는 집체교육은 중간 관리자이자 현업 개발자인 나에게 약간의 부담스러웠다. 뭐든 제대로 하려면 과감한 결정과 약간의 희생은 필요하다. 파일럿 프로젝트는 어렵지만 그만큼 얻는 것도 많았다.

프로젝트 수행 우수조를 기대하고 있었다.

전혀 예상치 못한 결과였다. 프로젝트 결과물 제출일이 가까워 올수록 시간 안에 완료하지 못할 것 같다는 불안감이 밀려왔다. 이때가 되면 누구나 한번쯤 포기해버릴까! 하는 생각을 몰래 하지

않았을까! 우리 조는 특별한 욕심은 없었다. 다만 조원들끼리 정했던 결과물을 기간 내에 도출하기만 해도 성공이라는 분위기였다.

프로젝트 과정에서 기억에 남는 순간과

미처 생각하지 못했던 부분이 있었다면.

우리 조는 시스템 전문가, DB 전문가, 개발자 등 서로 다른 영역에서 일해온 조원들로 구성됐다. 각자의 영역에서 쌓아온 전문성과 자부심이 프로젝트 수행의 걸림돌이 될 수 있다고 생각했다. 하지만 조원 모두 목표가 무엇인지를 기준으로 겸손하게 각자의 역할을 충분히 해냈다. 멀리 나주에서 교육을 받으러 온 조원까지 있어서 주중 미팅은 사실상 힘들었다. 낮에는 회사일, 밤에는 프로젝트 준비를 위한 온라인 미팅을 하면서 치열한 시간을 보냈다. 지나고 보니 축복의 순간이었다.

조원들과 함께 프로젝트를 더

발전시킬 계획이 있다.

우선은 빅데이터 아카데미 교육을 받느라 밀린 회사일 처리에 집중하고 있다. 멀리 나주에서 일하는 조원까지 모이기로 했는데, 그때 얘기를 나눌 계획이다. 우리나라 스포츠는 대부분 남자 프로 경기에 집중했다는 것을 이번 프로젝트를 하면서

실감했다. 온 국민이 즐기는 스포츠가 되기 위해서는 저변 확대가 필요하다. 그 측면에서 프로·대학·실업·고등학교 여자 농구팀에서도 톱5 선수를 찾아내고 싶다.

빅데이터 아카데미에서 어떤 도움을

받았고 가능성을 발견했나.

빅데이터 아카데미는 전문가들의 실제 현장 이야기를 들을 수 있는 곳이다. 기술력과 이론을 겸비한 강사진이 포함된 커리큘럼은 이미 검증 받았다. 여러 유사 교육과정도 많이 생겼다. 여러 곳으로부터 교육 참가안내 매일도 자주 받는다. 교육은 내용뿐 아니라, 대외 공신력도 중요하다. 공인 기관에서 하는 빅데이터 교육이고, 수강생들 또한 저마다 자부심을 갖고 있었다. 짧지 않은 기간 열정적으로 참여한 강민호, 김지훈, 김태형, 박상현, 이영호 조원께 감사 드린다. 교육 일정을 챙기고 힘들 때마다 격려와 응원을 해 주신 빅데이터 아카데미 관계자들에게도 감사드린다.

염증성 장질환^{IBD} 환자의 결핵 발병예측 분석



구분	예측분석, 조작적 정의
적용 도구	조작적 정의(Operating Definition), Survival Analysis, Cox Model, Maria DB, R studio
수집 데이터	2010~2014년 건강보험심사평가원 전체환자 데이터, A대학병원 랜덤 추출 환자 데이터
산출물	IBD 환자 조작적 정의, IBD 환자들의 약제가 주는 결핵 위험도 파악, 약물 사용 패턴을 확인해 순간 위험률 확인
지도	안원철
참여자	김지연 ^{조장} , 강근원, 곽민섭, 이영주
프로젝트 소개	염증성 장질환은 20~30대에 발병하는 희귀성 난치 질환으로, 완치 없이 평생 약을 먹어야 한다. 최근 류마티스·장질환에 사용되는 항TNF 약물 부작용 중에 결핵 발병 위험성이 높아지는 문제가 있다. 건강보험심사평가원 명세서 데이터를 분석해 결핵 발병예측 모형 제작에 도전했다.

THE CHALLENGES

의료 빅데이터 전문가 집체교육을 마치고 파일럿 프로젝트가 시작됐을 때 어떤 주제로 분석을 진행해야 할지 고민이었다. 이때 건강보험심사평가원 전체 환자 데이터가 제시돼 여기서 무엇인가를 도출해 내는 프로젝트가 시작됐다. 데이터 명세서의 변수 명을 확인하면서 환자가 병원에 갔을 때 어떤 정보가 기록되고, 그 정보를 어떻게 이용할 수 있을지를 팀원들과 의논했다. 주제 선정은 현직 의사로 활동하는 광민섭 조원이 많은 도움을 주었다. 광민섭 조원은 이전부터 국가에서 제공하는 공공데이터에 관심이 많아서 제공 받은 데이터세트로 어떤 결과물을 도출할 수 있을지 정확하게 짚어 주었다. 논의 결과 염증성 장질환 환자 분석을 프로젝트 최종 주제로 선정했다.

데이터 분석으로 염증성 장질환 환자들의 약물 복용에 따른 결핵 발병에 측 모형을 제작하기 위해 우리 조는 다음과 같은 3가지 목표를 정했다. 첫째로 염증성 장질환 환자의 정확한 추출이 가능한 ‘염증성 장질환의 조작적 정의’를 완성한다. 둘째, 약제에 의한 ‘결핵 발병빈도를 파악하고 위험 인자를 확인한다. 마지막으로 염증성 장질환 환자에서 결핵 발병 예방을 위한 ‘새로운 약물 치료 지침’을 구축한다. 염증성 장질환 환자들의 약물 복용에 따른 결핵 발병 예측 모형을 제작하기 위해 우리 조는 다음과 같은 3가지 목표를 정했다. 첫째, 염증성 장질환 환자의 정확한 추출이 가능한 ‘염증성 장질환의 조작적 정의’를 완성한다. 둘째, 약제에 의한 ‘결핵 발병빈도를 파악하고 위험 인자를 확인한다. 마지막으로 염증성 장질환 환자에서 결핵 발병 예방을 위한 ‘새로운 약물 치료 지침’을 구축한다.

염증성 장질환이란?

염증성 장질환은 20~30대에 발병하는 희귀성 난치 질환으로, 완치 없이 평생 약을 먹게 된다. 최근 류마티스·장질환에 사용되는 항TNF 약물 부작용 중에 결핵 발병 위험성이 높아지는 문제가 이슈가 되고 있다.

만성적이고 반복적으로 장점막에 염증이 발생하며, 크론병(CD) 또는 궤양성대장염(UC)이라고 한다. 유전적인 요인과 이상 면역반응, 환경 요인, 장내 세균총 요인에 따라 발생하는 것으로 알려졌다. 치료는 Mesalazine, 면역억제제, 스테로이드, 항TNF 제제, 수술 등으로 한다. 하지만 완치가 거의 불가능하며 평생 약물 복용과 수술 치료가 필요하다.

THE APPROACH

사용 데이터

건강보험심사평가원으로부터 전체 환자의 2010~2014년 데이터세트(NPS)와 A대학병원으로부터 조작적 정의 유효성검사(Validation)로 무작위 추출(Random Sample)한 환자 데이터를 확보했다. 분석할 데이터가 코호트(cohort)가 아닌, 1년 단위로 끊어져 있었으므로 1년 안에 발생할 수 있는 질병인 결핵을 반응 변수로 선정했다.

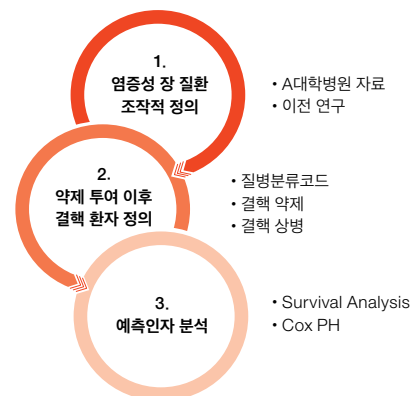
염증성 장질환·결핵 환자의 정의

데이터 분석 전에 가장 중요한 부분은 환자의 조작적 정의를 구축하는 것이다. 데이터에는 환자가 특정 질환을 앓고 있다는 기록은 나와 있지 않았다. 이에 따라 여러 변수를 이용해 해당 환자가 특정 질병을 가진 환자라고 정의를 해주어야 한다.

명세서에 나온 다양한 정보를 이용해 염증성 장질환 환자를 정의했다. 이때 정의한 조작적 정의가 얼마나 정확한지 확인하기 위해 A대학병원의 환자 정보를 이용해 민감도와 ‘병에 걸리지 않았는데 병에 걸렸다고 진단’하는 오진율(False Positive Rate, FPR)을 계산했다. 다양한 조작적 정의 중에 검증 데이터세트에서 민감도와 발병으로 오진하는 FPR을 가장 크게 하는 조작적 정의를 사용했다. 또한 결핵 환자의 정의도 필요했다. 약제 조건에서 최소한 결핵약을 두 가지 이상 처방 받거나 400 table의 상병 코드가 결핵인 경우 결핵 환자로 정의했다.

그림 1

분석 프로세스

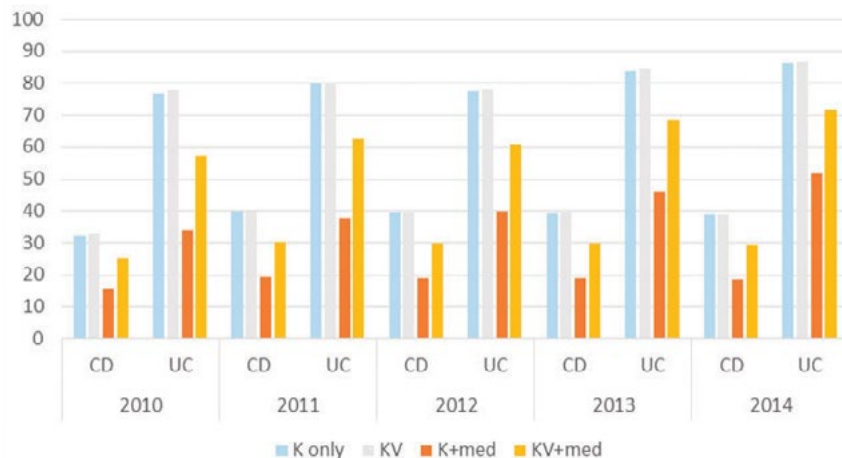


데이터 탐색

유병률 확인

유병률(Prevalence)은 정의한 염증성 장질환 환자에서 총 환자 수를 나누고 10만 명을 곱한 값으로, 10만 명당 환자 수를 의미한다. [그림 2]는 다양한 조

그림 2
IBD 유병률



작적 정의를 사용해 연도마다 IBD 환자의 유병률을 CD(Crohn Disease)와 UC(Ulcerative Colitis)를 나누어 정리한 것이다. 매해 환자 수가 증가 추세임을 확인할 수 있다. 기존 연구 결과에 나온 유병률 결과와도 비슷한 수치임을 확인했다.

파생변수 생성

약제 때문에 발생한 결핵을 확인하고 싶었으므로 최초 결핵 판정 시점과 최초 약 복용 시점 정보가 필요했다. 명세서에 나온 정보를 바탕으로 환자마다 시점에 대한 변수를 추출했다. 약제도 Mesalazine, 면역 억제제, 항TNF 제제라는 3개 카테코리로 구분해 얼마나 복용했는지 파생 변수를 생성했다.

약제 복용 패턴 확인

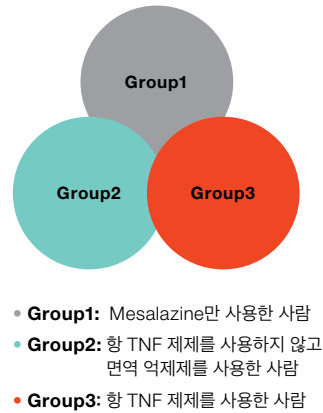
약제 복용 패턴을 보니 항TNF 제제만 복용한 사람의 빈도가 매우 낮았다. 이는 한국 의료 보험법상 다른 두 제제를 사용해야 그 이후에 항TNF 제제 처방이 나오기 때문으로 분석됐다. 다양한 약제 사용 패턴들을 비교하기 위해 새로운 그룹으로 재배분했다. 그룹1(G1)은 Mesalazine만 복용한 환자, 그룹2는 항TNF 제제를 사용하지 않고 면역 억제제를 사용한 환자, 그룹3은 항TNF 제제를 사용한 환자다. 약제 복용 패턴과 성별 나이를 고려해 결핵 발생 여부를 예측하려고 한다.

데이터 분석

약제 최초 복용 시점과 결핵 판정 시점의 차이를 이용해 생존 분석을 진행했다. 건강보험심사평가원 데이터는 연도별로 데이터가 끊겨 있어서 약제 최초 복용 시점과 결핵 판정 시점의 차이가 매해 0~400일 사이 값을 갖는다. 연도별로 분석하지 않고 각 연도를 합쳐서 데이터를 분석했다. 결핵 판정을 받지 않은 경우는 데이터가 중도 절단됐다고 처리했다. Cox-PH(Cox Proportional Hazard) 모형을 사용하여 방문 환자의 순간 위험율을 계산하였다. 결핵 판정 기준(Criteria)을 정할 수 있다면, 결핵 판정으로 예측된 환자에게 검사를 지도할 수 있다.

그림 3

새로운 그룹으로 재배분



THE OUTCOME

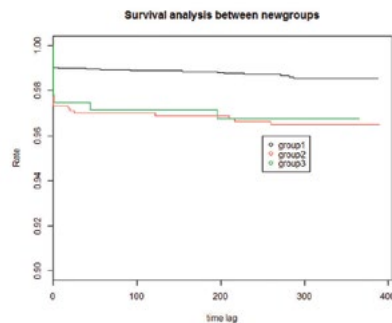
다양한 조작적 정의 중 K code와 약제 조건을 사용하여 IBD 환자를 정의한 경우, 밸리데이션 결과 민감도(sensitivity)가 0.99 이상, 양성예측도(Positive Predictive Value)가 0.93으로 가장 큰 값이 나왔다. 따라서 여러 조건 중에 K code와 약제 조건을 사용해 염증성 장질환(IBD) 환자를 정의했다.

약제 최초 복용 시점과 결핵 판정 시점의 차이를 시간으로 하여 카플란마이어 커브(Kaplan-Meier Curve)를 그려서 그룹 간 생존 커브(Survival Curve)가 다름을 확인할 수 있었다 로그순위시험(Log Rank Test) 결과로 그룹1과 그룹2, 그룹1과 그룹3 간에 유의한 차이를 확인할 수 있었다.

변수가 결핵 발생 여부에 어떻게 영향을 주는지 확인하기 위해 Cox-PH 모형을 적합했다(fitting). 하지만 Cox-PH 모형은 시간에 따라서 HR(Hazard Ratio)이 일정하다는 가정이 필요하다. 성별에 따라 HR이 일정하지 않아 성

그림 4

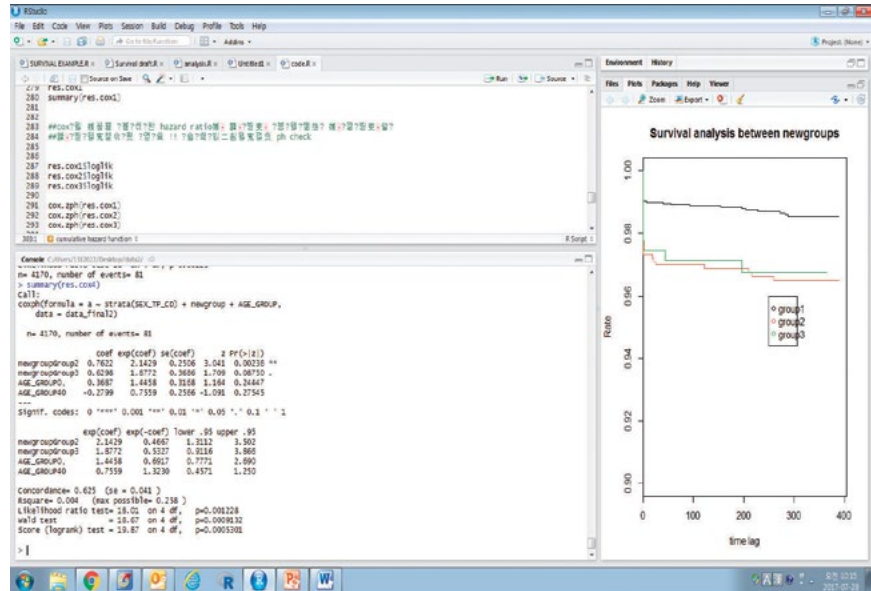
카플란마이어 추정량 (Kaplan-Meier estimator)



Log - Rank Test	1 vs 2	1 vs 3	2 vs 3
P-value	7.09e-05*	0.0135*	0.643

그림 5

층화 콕스 모형 결과 (Stratified Cox Regression)



별을 층화 변수로 보아 Stratified Cox Regression을 재적합(Re-fitting)했다. 그 결과 G1(그룹1, 이하 동일) 대비 G2 HR = 2.14, G1 대비 G3 HR = 1.88, 20~40세 대비 40세 이상 HR= 0.76, 20~40세 대비 0세~19세 HR =1.45를 확인할 수 있었다. 또한 구축된 모형을 통해 약물 사용패턴 성별 연령군 정보가 있으면, 순간 위험율을 계산할 수 있다. 나중에는 데이터셋트를 이용해 train과 test로 나눠 오분류율을 최소화하는 결핵 판정 여부 기준(criteria) 설정도 가능할 것으로 예상된다.

분석 대상 데이터가 1년치 데이터다 보니 추적할 수 있는 기간이 400일 안쪽으로 짧았다. Cohort 데이터로 실제 결핵이 발생하는 환자 수를 확보해 다시 분석해 보고 싶다. 현재 분석에는 결핵 환자 수가 상대적으로 적어 분석의 오류(bias)가 있다. 또한 결핵 예측을 확인할 수 있는 세팅 값을 정하고 싶다. 되도록 많은 데이터를 확보해 오분류율을 최소화할 수 있는지를 확인해 보고 싶다.

1+1=3!

우리 조는 서로의 의견을 존중하며 프로젝트의 완성을 최우선으로 생각했다. 다양한 배경에다 나이도 성별도 직업도 모두 달랐지만, 위화감 없이 자유롭게 의견 개진을 했고 열정적으로 프로젝트에 임했다.

2017.7.1 비교적 쉬웠던 주제 선정

주제 선정부터 조원들이 다양한 의견이 나왔다. 예상을 뒤엎고 모든 조원이 하고 싶은 주제에 대해 의견을 제시했다. 그 중에 이전부터 심평원 데이터를 이용해 분석을 하고 싶었던 조원의 의견이 가장 프로젝트로서 실현성이 높아 보였으므로 토론 결과 만장일치로 주제가 선정됐다.

2017.7.8 DB 프로그램 vs. R

우리 조에는 DB 전문가가 두 명이 있었다. 하지만 나중에 심사평가원에서 직접 데이터를 확인할 수 있게 R로도 DB 코드를 작성했다. R 분석은 통계 전공자가, DB 코드 작성은 DB 전공자가 각각 작성했다. 듀얼로 데이터베이스를 완성하고, 완성된 염증성 장질환 환자의 데이터를 이용해 본격적으로 분석에 들어갔다.

2017.7.15 파생변수 생성은 괴로워

염증성 장질환 환자의 약제 사용 패턴과 실제 약을 먹은 날짜의 확인이 필요했다. 분석을 위해 데이터 정리를 했다. 이때 어떻게 정보를 뽑아내야 할지 어려웠다. 하지만 멘토의 도움을 받아서 쉽게 데이터에서 변수를 추출해 낼 수 있었다. 이를 바탕으로 결핵 발생 시점과 약제 사용 시점을 비교할 수 있었고, 약제에 따른 결핵 발생 여부를 확인할 수 있는 데이터셋 생성도 가능했다.

2017.7.22 데이터셋의 실패

분석을 진행하다 보니 전혀 발생할 수 없는 약제 패턴들이 발견됐다. 구축한 데이터베이스에 문제가 생긴 거다. 발표 1주 전에 이 사실을 확인했기에 파일럿 프로젝트를 완료하지 못하는 상황이 발생할 수도 있었다. 하지만 천천히 코드를 다시 점검했고 수정해야 할 부분을 찾아서 데이터셋을 재가공했다. DB 전문 조원 2명이 듀얼로 데이터셋의 정확도를 다시 확인했다. 이 과정이 길고 매우 힘들었다. 이 기간을 무사히 잘 넘겨 프로젝트가 성공할 수 있었다.

2017.07.27 발표 전날 마무리를 위해 모이다

22일 이후로 수정된 데이터를 이용해 통계분석과 발표자료를 정리했다. 최종 발표 전에 자료를 점검하기 위해 모든 조원이 바쁜데도 강남에 모여 3시간 동안 자료 점검과 추가해야 할 부분을 챙겼다. 힘들었지만 뿌듯한 순간이었다.

“포기하지 않았기에 얻을 수 있었던 자신감”

김지연

셀트리온 CMC Statistics Team 대리



**의료전문가 과정은 아무나 접근하기
어려운 분야일 수 있다. 어떤 계기로 빅데이터
아카데미 의료전문가 과정을 선택했나.**

이전부터 회사에서 정부에서 공개하는
의료 데이터로 어떤 분석이 가능하고
어떤 정보를 추출할 수 있는지 관심이
많았다. 분석 실효성과 관련해 보고를 해야
했는데 의료 데이터를 직접 다뤄본 적이
없어서 전문적인 교육 과정을 찾고 있었다.
빅데이터 아카데미에 마침 의료 전문가 교육
과정이 개설된 것을 확인하고 지원했다.

**개인적으로 어떤 일을 하고 있으며,
빅데이터 아카데미 의료전문가
과정은 어떤 도움이 됐나.**

통계를 전공했고 제약회사 통계분석 팀에서
일하고 있다. 통계를 공부했다고 해서 모든
통계분석을 잘 아는 것은 아니다. 많이
사용하지 않는 분석은 공부도 더 필요했다.
빅데이터 아카데미에서 빅데이터 분석
교육을 받으며 직접 데이터를 다루고
분석할 수 있어서 좋았다. 실무에 하다
보면 학교에서 배운 내용만으로 해결되지
않는 애매한 부분들이 많다. 강사님들도
비슷한 고민을 했음을 알았다. 이와 관련해
많은 조언을 받았는데 도움이 됐다.

**조원들 간 역할 분담, 커뮤니케이션,
미팅 등을 어떻게 진행했는가.**

분석이나 발표자료는 구글 드라이브와
트렐로를 이용해 공유했다. 주로 카카오톡을
통해서 의견을 교환했다. 다들 본업이
바쁘다 보니 만나서 해야 할 일과 혼자서도

할 수 있는 일을 구분하는 것이 중요했다.
파일럿 프로젝트 기간에는 일주일에 한 번
교육장에서 만날 때, 구성원 각자가 해야
할 일을 정했다. 더불어 일주일 동안 할
일을 진행하면서 궁금했던 점이나 정리가
필요한 부분을 의논했다. 하지만 프로젝트를
진행하다 보니가 카카오톡에서 의논하는
것보단 만나서 말하는 것이 커뮤니케이션
에러를 줄일 수 있어서 더 효율적이라는
생각이 들었다. 개인적으로는 일주일에 한
번 만나는 것보다는 매일 만나서 총기간을
단축해 끝내는 것도 좋을 것 같다. 역할분담은
조원들이 무엇을 할 수 있는지 없는지
파악해서 해야 할 일의 목록이 나오면 사람에
맞추어 일을 배분하는 것이 중요하다. 우리
조에는 다행히 다양한 배경과 지식을 갖춘
사람들이 있어서 주제 선정부터 데이터
분석 준비까지 원활하게 진행할 수 있었다.

어떤 어려움이 가장 컸나.

프로젝트 마무리를 해야 하는 4주차에
가장 힘들었다. 팀에서 정한 소주제 3개
정도가 있었는데 발표 1주 전에 분석용
데이터셋을 만드는 코드가 완성됐다.
팀원들이 원하는 것들 중에 지금 할 수 있는
것과 할 수 없는 것을 정해 일을 끊고, 업무를
배분하는 것이 가장 힘들었다. 원하는 3개
주제 중에 하나를 포기하고, 할 수 있는
나머지 2주제에 대해 집중하고 몰입한
결과 프로젝트를 잘 마무리 할 수 있었다.

기억에 남을 만한 에피소드도 있었을 거 같다.

4주차에 알고 있는 분석 방향과 프로젝트

진행 방향이 다를 수 파악했다. 이제까지
작성한 R 코드를 처음부터 다 수정해야 하는
상황이었다. 눈 앞이 캄캄했다. 그날 스터디가
4시 반쯤 끝났다. 문제가 발생했는데 무엇이
문제인지를 파악하지 못한 채 헤어졌다. 당장
다음주에 발표를 해야 하는데 프로젝트가
얼어지기 직전이었다. 포기할까 하고 생각도
해 보았다. 이때 코드에서 문제 지점만
찾아낸다면 프로젝트를 완성할 수 있겠다는
생각이 들었다. 다시 마음을 추스르고 코드를
찬찬히 확인해 가면서 수정했다. 이때 남은
시간에 할 수 있는 것과 할 수 없는 것을
구분해 할 수 있는 것 위주로 정리했다.
오류 점검에 착수한 당일에 오류를 발견할 수
없었다. 그 다음날 차분히 보니 오류 지점이
보였다. 그 다음부터는 물 흐르듯 문제가 다
해결돼 프로젝트를 마무리할 수 있었다.

**빅데이터 아카데미 교육 수강을 검토중인
분들에게 해주고 싶은 말이 있다면.**

개인적으로 공부를 하다 보면 여러 제약
때문에 중간에 그만 두기 쉽다. 하지만 단체
교육은 웬만하면 정해진 커리큘럼대로
따라갈 수 있으므로 포기하지 않고 들으면
얻어 가는 것이 많다. 집체교육에서
전문적인 지식들을 배울 수 있고,
파일럿 프로젝트에서 앞서 배운 내용을
적용·활용할 수 있다. 가능하면 빅데이터
아카데미 교육을 수강하라고 권한다.

기타 하고 싶은 말은.

프로젝트를 진행하면서 귀찮은 기색 하나
없이 열심히 노력하고 따라와 주셨던
조원들에게 감사한다. 열정적인 조원들
덕분에 프로젝트를 완성할 수 있었다.
수평적인 위치에서 자유롭게 의견을
교환하고 요청하는 것을 최대한 들어주려
했던 조원들 덕분에 즐겁게 했다. 그리고
교육 기회를 준 회사와 교육 때문에 업무가
많아진 팀원들에게도 감사한다. 회사의
팀원들이 편의를 많이 봐주어서 프로젝트에
집중할 수 있었다. 팀 프로젝트에서 함께
고생하신 안원철 멘토께 감사한다. 정말
이거 저거 많이 요구했는데 그때마다 귀찮은
기색 없이 다 확인하고 조언해 주셔서
프로젝트를 완성하는 데 큰 힘이 됐다.

양파 생산량 예측 기반 금융상품 제안



구분	빅데이터 분석, 예측모델, 기계학습
적용 도구	R, H2O 기계학습 플랫폼
수집 데이터	기상 자료 개방 포털, 기상·기후 데이터 플랫폼, 통계청 KOSIS(온도평균, 최고 온도, 최저 온도, 일교차, 강수량, 일조 시간, 재배 면적, 양파 생산량, 지역별 양파 재배면적 대비 재배 비율)
산출물	기상 데이터 및 연관 데이터 분석 결합 양파 생산량 예측 기반 금융상품 제안
지도	안상선
참여자	이혜영 조장 , 강하늘, 권혜윤, 김성학, 김한용, 이주광, 이태연
프로젝트 소개	기상청과 통계청의 기상 데이터, 양파 재배면적, 생산량과 SNS 연관 데이터 복합 분석을 통해 양파 생산량 예측모델 수립과 검증을 했다. 이를 활용해 농가를 위한 금융 서비스를 제안하기 위한 분석이다.

THE CHALLENGES

비금융권 영역에서 데이터 분석·활용을 통한 금융상품의 가치를 찾자!

금융 빅데이터 융합 전문가 1기 과정을 시작하면서 함께 만난 조원들 간 어떻게 프로젝트 주제를 만들고 진행할지 의견을 나누다 보니 금융 상품, 크라우드펀딩, 데이터 분석 및 서비스 기획 등의 영역에서 활동하는 각자의 역량을 결합하면 좋겠다는 공감대가 형성됐다. 비금융권 분야에서 빅데이터 기술 적용 전후 모습의 차이가 큰 주제로서 어떤 게 있을까를 놓고 뉴스, 논문 등을 수집·조사했다. 농업 분야에서는 농민들이 실제로 체험하는 기초 데이터를 수집·처리·분석하는 인프라가 상대적으로 부족함을 알게 됐다.

이에 따라 농업 분야로 대주제를 잡고 세부 방향을 정하기 위해 다시 조사했다. 빅데이터 분석에서 중요한 첫째 단계가 문제 정의다. 하지만 해결 과제를 데이터 분석으로 풀어볼 수 있는 문제로 표현하는 것이 쉽지 않았다. 지역 날씨 모니터링, 농업 데이터 분석에 기반해 농업인들이 휴대전화나 태블릿 PC에서 원격으로 날씨와 토지 상태를 점검할 수 있고, 기후에 따른 토양 분석으로 최적의 파종 시기와 종자의 종류를 예측할 수 있는 미국의 클라이미트 스타트업과 같이 손쉽게 방향을 잡아갈 수 있으리라 생각했다. 하지만 농업에서 데이터가 활용될 수 있는 잠재 영역들이 너무나 다양하고 방대했다. 몇 차례 반복적인 브레인스토밍에서 의견 도출과 검증을 거쳐 멘토의 지원까지 받아 데이터 분석과 기계학습을 통한 예측이 효과적으로 적용될 수 있는 ‘농산물 생산량 예측’으로 좁힐 수 있었다. 이어서 농산물마다 재배 조건, 주기, 지역 등 편차가 있으므로 재배 기간이 길면서 농업인들에게 생산량 예측 실패에 따른 영향력이 큰 작물을 탐색했다.

때마침 조사 시점에 여러 뉴스에서 양파의 상품성 저하에 따른 문제점이 보도됐다. 2016년 대비 30%까지 떨어진 강수량으로 인하여 수확량의 40% 이상이 상품성이 떨어진다는 내용이었다. 이는 농가뿐 아니라 소비자 부담 증가로 연결되므로 대책 마련을 촉구하고 있었다. 양파는 중국집 등에서 흔히 보던 채소류이지만, 소매가 추이가 30% 이상 가파르게 치솟고 있었다. 생산량과 소비자 물가의 상관성이 높은 작물로 양파를 프로젝트 핵심 주제로 선정했다.

양파 생산량 예측에 영향을 주는 요인들로 기상 데이터 및 SNS 등 연관 데

이터들을 수집했다. 이 데이터에 대한 분석 범위를 정의하고 목차를 1)문제 정의, 2)솔루션 싱킹, 3)데이터 분석과 예측 모델링, 4) 연계 금융 서비스 제안으로 잡았다.

THE APPROACH

프로젝트 범위에 따른 추진 계획을 세우고 완수에 필요한 업무를 나눠 조원들 간 의견을 공유했다. 조원별로 선호하는 역할을 할당해 자율적인 환경에서 프로젝트를 진행했다. 분석 인프라 구축보다는 분석·예측에 따른 서비스 기획에 대한 비중이 높았다. 이에 따라 분석 대상의 데이터 소스 수집과 전처리, 분석 모델 설계 구현과 이를 응용한 서비스 기획에 초점을 맞추었다. 추진 계획을 세워 일정별 추진 사항을 기록해 공유했다. 완벽하지 않더라도 조원들 각자가 맡은 부분을 수행하면서 필요 시 다른 조원들과 협의하면서 프로젝트를 진행했다.

문제 정의

현실의 문제를 데이터 분석 문제로 정의하는 것을 먼저 해야 했다. 양파는 가격 하락, 재배 감소, 가격 상승, 재배 증가의 순환 과정을 2년 주기로 나타나는 작물로 알려졌다. 그럼에도 순서대로 진행하지 않고 생산량 예측 데이터가 과거 경험을 쉽게 대변하지 못하는 현상을 보여주고 있었다. 2014년은 농업관측센터에서는 5월 상순에도 기상 여건이 좋아 양파 단수가 증가할 것이라 전망했다. 하지만 7월 관측 결과 5월 중순 이후 고온과 가뭄에 따라 단수가 감소했다. 이에따라 농업인들로부터 ‘근거 없는 예측 발표로 생산량 감소와 가격폭락을 불러왔다’는 비판을 듣기도 했다.

양파는 다른 채소류와 달리 수분 함량이 높아 저장에 어렵다. 절단 건조를 통한 양념 원료 이용 외에는 활용 방법이 많지 않아 과잉 생산되면 폐기된다. 수요량을 정확하게 예측·생산하는 것이 효과적이나 기상 외의 변수들로 인해 인위적인 수요·생산 관리가 어렵다. 가뭄, 우박 등 자연재해에 따른 농작물 재

해 보험과 같은 금융 서비스가 있기는 하다. 하지만 1년 만기 상품으로 자연 재해의 피해가 적은 해에는 가입율이 떨어진다. 무엇보다 상대적으로 높은 자기 부담비율로 농가에 부담을 주고 있어서 자체 개선 및 대체 가능한 금융 서비스가 필요한 상황이었다. 따라서 문제 정의를, 양파 생산량 예측 데이터의 정확도가 상대적으로 낮아 양파를 포함한 농작물 작황 부담을 경감시키는 금융 서비스의 필요성 대두로 정리했다.

솔루션 싱킹

양파 생산량 예측의 정확도를 높이기 위한 예측 모델을 구축하기 위해 이에 영향을 주는 초기 데이터 수집 정의가 중요했다. 이상 현상을 예측하기 위해, 기상을 포함한 복합적인 요인들이 예측 모델에 영향을 주는 바로 가설을 세웠다. 이에 따라 기상청에서 구축한 자체 빅데이터 분석 인프라를 활용해 평균 온도, 최고 온도, 일교차, 강수량, 일조시간, 재배면적 데이터 항목을 도출했다. 양파 생산량 자체의 패턴 추이와 지역별 양파 재배 면적 대비 재배 비율을 추가했다. 이로써 양파 생산량 예측모델을 수립하는 데에 신뢰도를 높이하고자 했다. 추가로 양파 생산량과 연관성이 높은 SNS 키워드들을 추려내 관계 규칙을 분석함으로써 예측 모델의 정확도를 개선하고자 했다. 가설 수립과 더불어 현황 데이터, 즉 도메인 분석으로 주요 데이터 항목에 대한 특성치를 추출해 시각화해 사전 데이터 분석 수집·처리를 했다.

연관 금융 서비스 제안으로는 농업 수입보장보험과 같은 상품들이 생산량 예측 치 반영 부족과 보험료 할증 부담 등의 개선이 필요하다고 봤다. 이에 따라 생산량 예측 데이터 제공과 연계된 금융 서비스로서 대체 보험 상품 제안 및 크라우드펀딩 플랫폼을 활용하기로 했다.

도메인 톨: 성공적인 양파 재배를 위한 노하우

1. 유묘기: 8월 중순에서 10월 하순, 발아 적정 온도는 15~25도, 강수 빈도가 낮으면 품질 저하
2. 활착기·월동기: 11월 상순 ~ 1월 하순, 최저 기온이 영하 9도씨이면 동해 발생
3. 경엽 신장기: 2월 상순 ~ 3월 상순, 강수 빈도가 높으면 습해 발생
4. 구 비대기: 4월 하순 ~ 7월 상순, 평균 기온이 25도 이상이면 고온 장애 발생

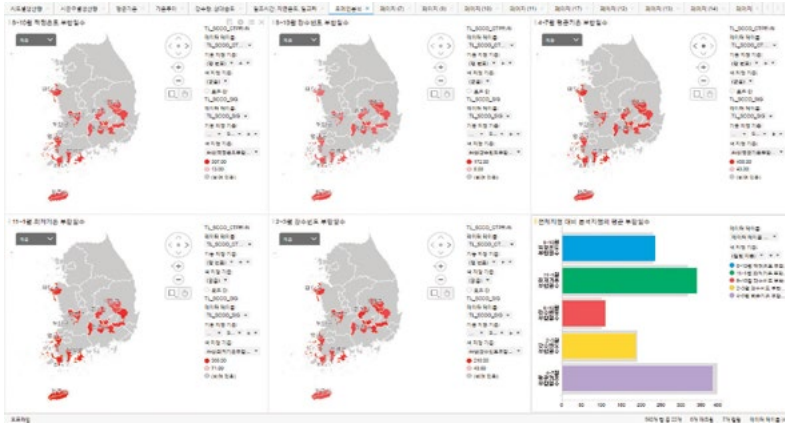


그림 1
기상 조건 일차
빈도수 분석

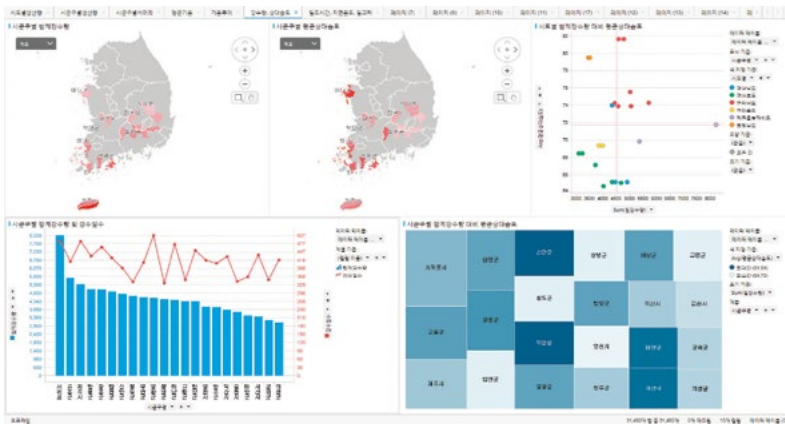


그림 2
시군구별 강수량,
상대습도

데이터 분석과 예측 모델링

데이터 분석 기상자료 개방포털, 기상기후 데이터 플랫폼, 통계청 국가통계포털(KOSIS) 플랫폼을 활용해 정의한 데이터 수집 항목에 매칭했다. 시간은 월, 공간은 도(국내)를 기준으로 했다. 2013년 8월에서 2016년 6월까지 양파의 작황 주기에 따른 데이터를 수집했다. 1차 데이터로 기상 데이터, 양파 재배면적 대비 재배 비율 및 생산량 데이터를 병합, 결측치 처리를 했다. 또한 관측 장비에서 관측될 수 없는 이상치 데이터를 Null로 구분·변환했다. 파생변수 생성을 위해 관측 장비별로 연별·월별로 평균 온도, 온도의 최대/최소값, 강수량 합, 일조량 평균, 일교차 평균 등을 계산·처리해 데이터베이스를 만들었다.

기상 데이터와 양파재배 면적 데이터를 병합했다. 재배면적 비율만큼 가

중치로 곱한 뒤 만든 데이터베이스, 지역별·시도·시군구별·연월별로 평균 온도의 합, 최고 온도와 최저 온도·일교차·강수량·일조시간·재배면적 합을 계산해 저장한 데이터베이스로 정리했다.

양파는 보통 6월에 수확하므로 전년 9~12월 날씨는 다음해의 양파 생산량에 영향을 미친다. 날씨 정보를 내년 생산량과 재배면적과 연동했다. 2013년도 데이터는 전년도 데이터가 필요해 삭제했다. 이를 기상 데이터와 재배면적 데이터를 결합한 데이터베이스와 양파 생산량 데이터에 각각 병합해 데이터셋으로 저장함으로써 최종 변수 생성을 완료했다. 최종 변수인 데이터셋은 지역(시도, 구군)별 양파 재배면적, 해당 연도 생산량, 1월에서 12월까지의 평균 기온, 최고/최저 기온, 일교차, 강수량 합, 일조량 합으로 구성됐다.

본격적인 데이터 분석과 예측모델을 구축하기 전에 다중공선성 여부를 확인하고자 함수를 적용해 다중공선성 연관관계를 분석했다. 다중공선성이란 입력변수들끼리 연관관계가 존재함으로써 모델의 정확도와 적합성을 떨어뜨리는 문제를 의미한다. 문제가 없음을 확인한 뒤 일반화 선형 분석(generalized linear model)으로 모형을 구축했다.

표 1

최종 도출된
데이터셋 테이블
설명과 구조

DataSet	region_1	지역(시, 도)
	region_2	지역(구, 군)
	area	양파 재배 면적
	year	연도
	y	해당년도 생산량
	TAD_01~TAD_12	해당지역 해당년도의 01월~12월 평균기온으로 각각의 칸에 저장되어 있음
	TAmx_01~TAmx_12	해당지역 해당년도의 01월~12월 최고기온으로 각각의 칸에 저장되어 있음
	TAmin_01~TAmin_12	해당지역 해당년도의 01월~12월 최저기온으로 각각의 칸에 저장되어 있음
	DTD_01~DTD_12	해당지역 해당년도의 01월~12월 일교차로 각각의 칸에 저장되어 있음
	RAINSUM_01~RAINSUM_12	해당지역 해당년도의 01월~12월 강수량 합으로 각각의 칸에 저장되어 있음
	SUM_SS_HR_01~SUM_SS_HR_12	해당지역 해당년도의 01월~12월 일조량 합으로 각각의 칸에 저장되어 있음

기계학습 라이브러리 플랫폼인 H2O를 연동해 분석 예측모형에 중요도가 높은 순위별로 변수들을 정렬·처리했다. 이어서 양과 생산량을 예측하기 위해 2개 변수간의 상관관계(피어스 상관계수 0.45 이상)가 높은 변수들은 영향력이 중복됨에 따라 1개만 활용했다. 따라서 최종적으로 예측모델의 구성 변수들로는 2월, 9월의 강수량과 11월~1월 하순 최저 기온이 9 미만인 날의 개수 변수들로 확정됐다.

확정된 모델 변수들의 학습을 통한 예측모델 검증에 위해 H2O 플랫폼과 연동해 회귀분석모델을 생성했다. 검증 방법은 2014년 날씨 정보를 이용해 회귀 방정식을 만들고, 2015년 날씨로 입력했을 때 예측한 양과 생산량과 실제 2015년의 양과 생산량 사이의 오차를 이용해 회귀식이 얼마나 정확한지를 살펴봤다.

일반화 선형 모델을 이용한 양과 생산량 예측 평가

일반화 선형 모델(Generalized Linear Model)을 이용한 양과 생산량의 예측 결과의 오차는 6% 이하로 정확도 검증을 완료했다. 도출된 데이터 예측 모델에 텍스트연관 분석을 추가해 SNS 키워드 동의어·불용어 처리를 했다. 2017년 7월 1일 기준 월간 검색어 기준으로 ‘양과’와 ‘가격’이란 키워드가 1만 7000회를 넘는 검색 횟수를 보여주고 있었다. 양과 가격과 연관성이 가장 높은 키워드 1위로 ‘당근’이 분류됐다. 이는 소비자 입장에서의 당근 가격이 인상될 때 양과 가격 또한 인상될 가능성이 높고, 양과 생산 유통 주기에 따라 생산량이 증가할 수 있음을 의미한다. 시간 여유가 있었다면, 당근 생산량 데이터 예측모델과 양과의 생산량 데이터 예측모델을 수립해 비교 또는 병합이 가능하지 않았을까 한다.

그림 3

양과 생산량 예측 모델과 학습·검증을 위해 사용한 코드(부분)

```
Coefficients: glm coefficients
names coefficients standardized_coefficients
1 Intercept 6443.210124 6496.454545
2 TA_WINTER 29.498434 64.272964
3 RAINSUM_12_2 -32.125775 -66.094480
4 RAINSUM_09_2 3.115240 67.110711
5 RAINSUM_02_0 9.136983 44.093241
```

```
for(Signifi_Year in min(Dataset$Year):max(Dataset$Year)){
  #현재 년도에서 최소년도의 자+1 이 줄의 인덱스 j 입니다
  j <- Signifi_Year-min(Dataset$Year)+1
  #Signifi_Year(테스트데이터)를 제외한 학습데이터를 생성합니다.
  Train <- subset(Dataset,
    year!=Signifi_Year,
    select=c("Y", Select_imp_RF$variable))
  #학습데이터를 출력합니다
  str(Train)

  #학습데이터를 h2o 형식으로 변환합니다.
  Train <- as.h2o(Train)
  #Signifi_Year의 예측하고자 하는 양과 생산량(y)를 Test_Y 에 저장합니다
  Test_Y <- subset(Dataset, year==Signifi_Year, select=c("Y"))

  #테스트 데이터에서 독립변수들을 저장합니다. 이러한 독립변수들을 이용
  #예측이 얼마나 되는지를 테스트 할 것입니다.
  Test_X <- subset(Dataset, year==Signifi_Year,
    select=c(Select_imp_RF$variable))

  #Test_X 를 h2o 형식으로 변환합니다.
  Test_X <- as.h2o(Test_X)

  # 학습데이터를 이용한 glm 모델을 생성합니다
  GLM_Validation <- h2o.glm(
    y=1, x=2:ncol(Train),
    training_frame = Train,
    family = "gaussian",
    link = "family_default",
    fold_assignment = "AUTO", lambda = 0.04849)
```

THE OUTCOME

2015년 전체 농가별 소득 집계 데이터를 분석한 결과 도시 근로자 대비 평균 소득은 64%, 이중 양파 농가 소득은 농가 소득 평균 대비 58% 수준에 그쳤다. 보험 가입율은 지속적으로 하락하는 반면, 양파 재배 농가 손해율은 타 작물 농가 손해율 대비 5배 이상으로 확인됐다. 그럼에도 양파 농가는 금융 서비스의 사각지대에 놓여 있다고 판단했기에 다음과 같이 세 가지 제안을 수립했다.

첫째, 보험 가입 제고를 통한 농가소득 리스크 헤지(Risk hedge)로써 보험 가입 농가는 재해로 인한 생산량 변동 불안이 내려간다. 보험사 또한 생산량 예측 데이터를 활용해 이상 징후 발견 시 보험 가입을 위한 적극적인 홍보 활동을 진행해 수익률을 제고할 수 있을 것이다.

둘째, 보험상품 계리 심화를 통한 손해율($\text{지급보험금} \div \text{거수보험료} \times 100$) 안정화로써 기존 보험요율 산정 방식에 예측생산 모델링을 적용해 운영·검증함으로써 손해율이 높은 포도, 복분자와 같은 타 작물 품목의 보험상품으로도 확산할 수 있다.

셋째, 클라우드펀딩으로 데이터 예측 상품(양파)을 등록·운영함으로써 양파 수급에 대한 공공 자산과 같은 인식으로 투자해 상호 이익을 공유할 수 있도록 설계할 수 있다. 이는 데이터 유통 거래 플랫폼의 활성화가 멀지 않은 미래에 다각화된 포트폴리오 구성으로 편입할 수 있을 것이다.

빅데이터 인프라 구축, 엔지니어링 부분도 중요하지만 앞으로 인프라·기술의 표준화, 가이드라인 대중화로서 데이터 자체의 가치를 도출해 내는 분석, 이를 지능적으로 해석하고 추론할 수 있는 인공지능이 맞물리면 데이터 분석은 지속적으로 발전할 것이다. 민간 영역에서도 데이터 공개와 활용이 활발해지면, 실제 산업과 사회에서 응용할 수 있는 데이터 분석 서비스들이 금융 산업을 비롯해 여러 산업 분야에 큰 파급효과를 불러올 것으로 기대된다.

고생 끝에 낙이 온다

비금융권 영역에서 데이터 분석·활용을 통한 금융상품의 가치를 찾자는 미션으로 프로젝트를 시작했다. 조원들·멘토 간의 커뮤니케이션을 통해 분석 대상 데이터 범주와 프로젝트 과제를 정의하기 위해 다사다난한 초기단계를 경험했다. 물론 이후는 순탄하게 흘러갔다. 고생 끝에 낙이 온다는 말이 틀리지 않음을 새삼 느낀다.

2017.6.11 분석 대상 선정

농업 분야에서의 데이터 분석 가치가 높음을 조 내부에서 공감하고 있었다. 하지만 기후 데이터를 통한 농작물 수확 시기 예측, 구제역 이동 경로 예측, Home-IoT 확산 예측에 따른 금융 해킹보험 제안, 금융상품 추천 어드바이저, 벼 재해보험 리모델링 등 여러 아이디어가 나왔다. 하지만 데이터 분석·예측 모델과 금융 서비스 연계라는 목표에 매칭되는 분석 대상을 선정하기가 쉽지 않았다. 멘토의 지원으로 농작물 생산량 예측에 따른 수급 불균형으로 방향을 조정했다. 마침 가뭄 등으로 인해 양파 생산량 예측이 어긋나 농가의 피해가 보도되고 있었다. 이에 조원 과반수 이상 동의로 '양파' 분석을 주제로 선정했다.

2017.6.14 기상청 데이터 분석 플랫폼

김한용 조원 추천으로 알게 된 기상청 데이터 분석 플랫폼(bd.kma.go.kr)을 각 조원들이 데이터 분석·활용 인프라로 사용하고자 했다. 하지만 계정 사용권한 부여가 바로 진행되지 않았다. 특정 데이터를 받고자 해도 보안상 이슈로 락이 걸려 있어 다운로드되지 않았다. 통계청 플랫폼을 포함해 최종 데이터까지 통합할 수 있었지만, 해당 플랫폼은 R 코드 내용 참조에 그쳐 여러모로 아쉬움이 남았다.

2017.6.25 각자 영역에 올인!

데이터 소스 수집, 데이터 수집 처리, 분석 모델 수립, 분석 및 시각화, 분석 모델 검증, 금융상품 기획까지 일정표를 만들었다. 하지만 조원들 간 진행사항 공유 및 Q&A로 카톡 메시지는 여전히 24시간 가동 중!

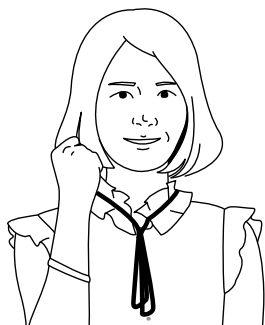
2017.6.28 R 데이터 분석 완료, 금융 서비스 기획

H2O 플랫폼을 적용하여 회귀분석 모델링으로 예측모델 검증 완료!

“생각과 실제가 다름을 확인했던 프로젝트”

이혜영

알스퍼릿 대표



이 주제를 선정하게 된 배경과

어떤 목표를 갖고 했다.

공개된 데이터 중에서 신뢰도가 높고
확보하기 쉬운 기상 데이터로 선정 한 후,
이를 어떻게 활용할지를 고민하는 형태로
접근했다. 기상 데이터가 누구에게 어떤
도움이 될까? 하고 고민하다가 기상 조건에
영향을 많이 받는 농업으로 영역을 좁혔다.
배추와 양파 중에서 중국집에서 필수
채소인 양파를 분석 대상 작물로 확정했다.
마침 2017년은 가을까지 심해 양파 수급
난이 예상돼 더 관심을 갖게 됐다.

금융융합전문가 과정 1기의 우수조로

뽑혀 소감이 남다를 거 같다.

처음 시작은 파일럿 프로젝트를 끝내기만
해도 다행일거라고 생각했다. 막상
프로젝트를 진행하면서 하나씩 길을
찾아나갈 수 있었다. 다른 조와 차별화한
주제 선정과 안상선 멘토의 적극적인
지원으로 프로젝트를 진행할수록 1위도
가능하겠구나 하는 자신감이 들었다.
결과가 좋게 나와 어려운 순간에 힘이
되는 좋은 경험 하나를 얻게 돼 기쁘다.

프로젝트 진행 중 어려움이 많았다고 들었다.

데이터 분석 중에서도 예측 분석을 농업

영역에 적용하는 예가 드물어 수십
번에 걸쳐 가설을 수립하고 수정하기를
반복했다. 멘토의 지도를 받으면서 길을
찾아나갔다. 여러 생각이 교차하기 마련인
어려운 순간에도 노력은 헛되지 않다는
믿음으로 희망을 놓지 않았다. 우수조로
호명되자 조원 모두가 너무나 할 것 없이
일어나 환호성을 질렀다. 그동안 몸 고생,
마음 고생했던 것에 대한 보상이지 않았나
싶다. 금융이라는 특정 도메인으로 특화된
과정이었던 만큼, 긴장했을 조원들이 서로
존중하면서 맡은 역할을 불평없이 한 것이
우수조로 선정된 요인이 된 것 같다. 온라인
소통의 힘도 확인할 수 있었던 기회였다.

기억에 남는 순간은.

프로젝트 발표 당일 이동 과정에서 이슈가
발생했다. 발표자인 내가 조금 늦게
발표장에 도착하는 바람에 조원 가운데
한 명이 대신 발표하고 있었다. 충분히
공유하고 있었으므로 다른 조원이 대신
발표할 수 있다고 봤다. Q&A 시간에
조장이 발표자로 참여함으로써 자연스럽게
마무리할 수 있었다. 주말마다 교육장으로
나와서 어떻게 하면 가치 있는 파일럿
프로젝트가 될 수 있을지를 놓고 고민하고
토론했던 순간이 기억에 남는다.

**주제를 발전시키면 자연재해로 인한 농민
피해를 줄이는 데 도움이 될 것 같다.**

기회가 된다면, 제한된 시간 때문에 시도하지
못했던 나머지 분석 과정을 조원들과 함께
해보고 싶다. 지난해에 폭등했던 작물을
이듬해에도 계속 재배해 가격이 폭락하는
패턴에서 벗어나는 토대로 우리 조의
파일럿 프로젝트가 작은 본보기가 되었으면
한다. 기상 빅데이터를 분석해 향후 작물
생산량을 예측하고, 가격 폭등이나 폭락에
대비한 농산물보험을 연계하면 농업계의
안정적인 소득이 이뤄질 것이라 본다.

빅데이터 아카데미 교육을 검토중인

분들에게 해주고 싶은 말이 있다면.

빅데이터 아카데미처럼 2주간의 집체교육과
2개월에 가까운 파일럿 프로젝트로 구성된
교육과정을 접할 기회는 직장인에게
그리 많지 않다. 배움의 열기가 가득한
빅데이터 아카데미에 들어오는
것만으로 변화에 동참하는 것이다.

스타트업 기업 지표와 투자유치 연관성 분석



구분	금융 빅데이터 분석, 투자 연관 요소 해석 및 예측 가능성 검증
적용 도구	R, 파이썬
수집 데이터	더브이씨 기업 데이터(2014), 은행 DW 기업 데이터, 중소기업현황정보시스템, NICE 신용정보, 서울기업정보망, 취업포털 커리어, 취업포털 잡플래닛
산출물	스타트업 기업 지표와 투자유치 연관성 분석
지도	이성윤
참여자	정호열 ^{초장} , 신동범, 이지태
프로젝트 소개	스타트업마다 투자 받는 금액의 차이에 대한 이유를 분석해 내고 이를 바탕으로 투자할 만한 스타트업을 미리 알아낼 수 있는지에 대해 예측하는 분석이다.

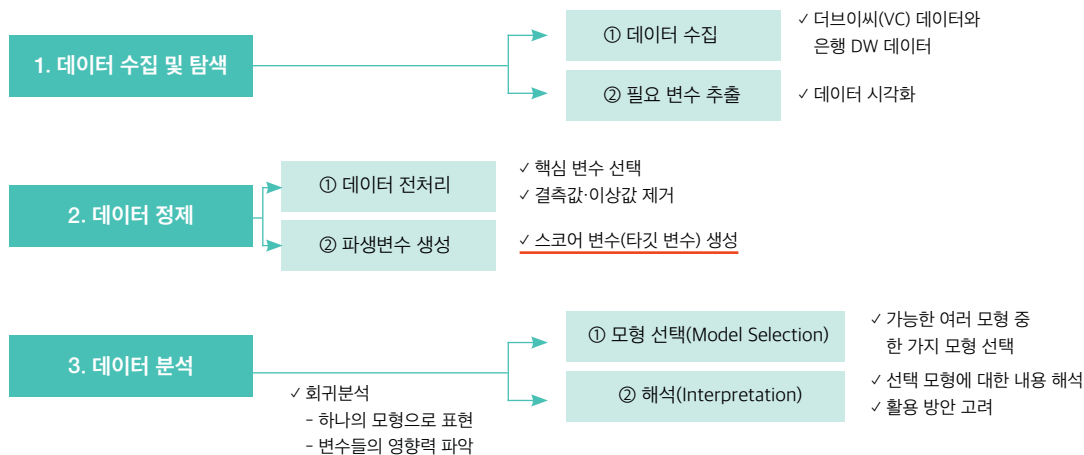
THE CHALLENGES

현재 금융권들은 기존의 예대마진(대출이자에서 예금이자를 뺀 나머지 부분) 중심의 매출로는 더 이상 미래에 생존할 수 없다는 사실을 인지하고 4차산업 관련 신기술과 융합한 신사업 발굴에 매진하고 있다. 이를 위해 신기술 경험 보유자 영입, 내부 인재 중심의 디지털 교육확대, 기업가치가 높은 스타트업과 협업해 협업 모델 찾기에 고심하고 있다. 현재 스타트업 투자를 위한 가치평가는 표준화한 기준 없이 동일 산업군 내 이미 성사된 과거의 딜을 근거로 한다. 데이터 분석 기반보다 엄밀한 기준 체계가 필요한 상황이다. 본 분석 프로젝트의 결과물을 스타트업 선발 및 투자(인수)에 활용할 수 있다면, 신기술 내재화를 강화하는 데 기여할 수 있을 것이다. 또 기업·고객에게 투명한 기업 가치평가, 정당한 조건의 투자유치를 할 수 있을 것이라고 봤다. 그래서 이 프로젝트에서 스타트업 현황 지표 중 투자에 영향을 주는 핵심 변수가 무엇인지 탐색했다. 현재 확보된 기업들 이외에 새로운 기업을 평가할 상황이 온다면, 과연 투자받을 만한 조건을 갖춘 기업인지 예측 가능성을 검증해 본 분석이었다.

THE APPROACH

제일 먼저 ‘더브이씨’라는 벤처캐피탈사로부터 무료로 제공되는 2014년 기업 데이터를 확보했다. 이 데이터를 중심으로 첫 번째로 시도한 기업가치 측정 요소는 상장여부·합병·투자라운드-시리즈 B 이상이었다. 하지만 결과가 만족스럽게 나오지 않았다. 결국 기업가치를 설명할 수 있는 다른 변수들을 찾아내야 했다. 정호열 조장이 일하는 회사의 연구소 인공지능팀과 사내 대학원팀에게 협조를 요청했다. 어렵게 은행 재무 데이터를 확보해 조원들과 함께 협업할 수 있는 발판을 마련했다.

우리 조는 먼저 데이터를 확보해 기업 현황 지표들의 영향력과 연관성을 파악·분석·시각화하고, 가장 타당한 모델을 수립해 결과를 분석했다. 이 결과를 활용해 실무에 적용할 수 있는 스타트업 투자 기준·체계를 마련하기 위해서였다.



하지만 현실은 그리 녹록하지 않았다. 문제를 정의하고 가설을 세우고 상관관계를 분석해 나가면서, 이에 대한 해석력 부족으로 가설을 새로 세우고 검증하는 작업을 되풀이해야 했다. 수집된 데이터가 우리에게 던진 질문은, 좋은 기업 가치를 가진 회사란 어떤 조건을 갖춘 곳일까? 우리가 갖고 있는 데이터만으로 기업가치를 정의할 수 있을까? 높은 기업 가치를 갖춘 회사는 어떤 핵심 변수를 갖고 있나?에 대해 여러 논의를 거듭한 결과, 다음과 같은 가설을 세웠다.

그림 1
3단계로 구분해
가설 검증



그림 2
하나의 데이터세트
만들기

- 투자금액은 현재의 기업가치를 일정하게 반영한다.
- 좋은 기업 가치를 가진 기업은 투자금액이 높다.

우리의 가설을 검증하기 위해 [그림 1]과 같은 3단계 작업을 수행해 나갔다. 이어서 투자 데이터로 타깃변수를 생성한 후 재무 데이터와 조인해 하나의 데이터세트를 만들어 냈다.

THE OUTCOME

투자 데이터와 재무 데이터 총 389건을 수집해 이상치를 제거해 154건으로 선별했다. 재무 데이터는 은행 DW(Data Warehouse)와 나이스기업정보 데이터를 보완해 110개 데이터로 통합했다. 데이터를 정제해 최종 92개의 데이터를 만들었다. 이후 기업가치 산정을 위해 투자 데이터를 전처리했다. 총 31개 변수 중 24개의 변수(예: 투자자명)를 제거하고 7개의 변수를 추출했다.

데이터 탐색 전략은 다음과 같다. 1)투자 데이터의 ‘투자금액’을 활용해 기업가치를 설정한다. 2)기업가치에 영향을 주는 핵심 변수 요인은 재무 데이터에서 추출한다. 3)투자금액은 투자라운드·섹터별로 기업가치가 반영돼 있다.

그림 3
7개의 변수 추출

1) 독립변수: 6개 변수 중 ‘라운드, 업력, 기술영역’ 선택

투자_추정라운드	투자 라운드(단계)를 나타내는 범주형 변수
투자대상_업력	투자 대상의 당시 업력을 나타내는 수치형 변수
투자대상_이름	투자 대상(회사명)을 나타냄
투자대상_기술영역	투자 대상의 섹터를 나타내는 범주형 변수
투자대상_분야1	투자 대상의 섹터를 나타냄(중분류)
투자대상_분야2	투자 대상의 섹터를 나타냄(소분류)

2) 종속변수: ‘투자금액’ 선택

투자금액	투자 대상 기업의 투자유치 금액을 나타냄
------	------------------------

첫 번째로 섹터별 투자유치 금액의 분포를 파악했다. 그 결과 섹터와 관련없이 투자유치 금액이 고르게 분포돼 있음을 확인했다. 두 번째로 업력별로 투자유치 금액의 분포를 파악했다. 그 결과 업력은 투자금액과 관련이 있으나 데이터 설명력이 부족해 검토해볼 의미가 없다고 판단해 제거했다. 세 번째로 라운드별로 투자유치 금액의 분포를 살펴봤다. 그 결과 투자 라운드별 투자금액·투자금액이 스케일링된 파생변수가 필요했다. 즉, 관하게 기업가치를 설명할 수 있는 변수이어야 하 가치를 스코어링했다. 이후 최종 데이터세트를 생 스코어에 영향을 주는지를 분석했다. 스케일된 스 ‘자산’ 변수가 유의미하게 영향을 미침을 확인할 소거법, 단계적방법을 사용해 자산, 부실예측구간 후 스코어 팩터에 대한 로지스틱 회귀 분석결과와 자 원수를 핵심변수로 채택했다. 최종적으로 Scaled 공통으로 중요한 자산, 부실예측구간을 핵심변수로

그림 4
라운드별 투자 유치금액 분포

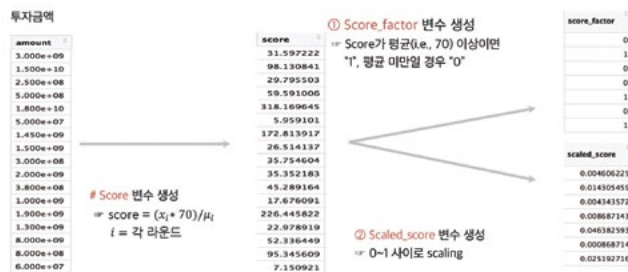


그림 5 타깃변수 생성

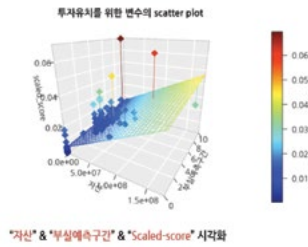


그림 6

스코어에 영향을 주는
독립변수 탐색

그림 7

자산, 부실예측구간, Scaled score의 시각화



자산, 부실예측구간, Scaled score를 시각화해 [그림 7]과 같은 결과를 얻었다. 이후 전체 데이터에 대한 변수 간 상관관계를 분석한 결과 각 변수가 서로 영향을 미치고 있었다. 이에 투자라운드별로 변수 간 상관관계 분석을 했다. 그 결과 pre-A에 대해 scaled_score 회귀분석 결과 자본과 부실예측 구간이 유의미한 변수로 확인됐다. Series A에 대해 회귀분석 결과 자산-매출액-자본-부실예측구간 순으로

유의미한 변수가 확인됐다.

결과를 요약하면 다음과 같다. 초기단계(pre-A)에서 자본, 부실예측구간이 투자금액에 영향을 미치나 전체 기업을 대표할 변수는 아니다. 자본이 있거나 기술력 등 외부 요인이 기업가치에 영향을 미칠 수 있었다. 중기라운드(Series A)에서는 자산과 매출액이 초기단계 대비 더욱 중요한 변수가 되고 있었다. 초기 단계보다 회사 내재가치를 더욱 중요하게 생각하는 단계이기도 하다. 순이익의 영향은 없었다.

분석 한계와 발전 방향

분석의 한계는 기업의 가치를 투자금액 기준으로 설정했다는 점이다. 투자금액은 트렌디한 섹터와 연관성이 높다고 볼 수 있다. 더불어 기초 회귀분석을 통한 검증이었으므로 한계가 따른다. 재방문율과 클릭 수 같은 기업 외부요인이 포함되지 않는 데이터이므로 초기 라운드의 기업가치 분석에 한계가 있다.

우리 조는 이 분석을 다음과 같이 발전시킬 계획이다. 기업 가치에 대한 투자금액 기반 기준을 보완할 수 있는 타깃 변수를 재산정하고, 다른 회귀분석 방법으로 결과를 분석할 계획이다. 또한 핵심변수 중 부실예측구간은 범주형 변수이므로, 그룹별로 분석하고 스타트업 투자 절차에 맞춰 결과에 대해 해석해볼 계획이다. 핵심 변수들을 중심으로 기업들을 클러스터링해 예측모델 확장 가능성도 검토해 보고 싶다.

현재의 데이터로 접근할 수 있는 것에 집중하다

우리 조는 서로의 장단점을 사전에 파악해 각각의 역할을 초기에 나눴다. 일해온 분야와 나이대도 달랐지만 서로를 배려해 줌으로 써 자유롭게 의견을 주고 받으며 열정적으로 프로젝트에 임할 수 있었다.

2017.11.10 주제 선정

주제를 선정하기 위해 다양한 의견이 나왔다. 주제 선택의 기준은 결과가 만족스럽게 도출되지는 않더라도 우리가 하고자 하는 방향성이 좋다면, 나중에 우리의 수고로움에 대한 보상이 될 것이므로, 결코 후회하지 않을 것을 하기로 했다. 이로 인해 만장일치로 주제를 선택할 수 있었다.

2017.11.17 데이터세트의 실패

수집된 데이터만으로 우리가 목표로 했던 분석을 수행하기에는 턱없이 부족했다. 그래서 비용을 지불하더라도 필요한 데이터를 확보하고자 노력했다. 하지만 너무 비싼 가격에 좌절을 해야 했다. 처음부터 완벽한 데이터세트를 확보하고 시작하는 분석이란 없다는 조언을 듣고 당장 확보한 데이터를 갖고 분석 가능한 부분을 먼저 정의한 후 가능성이 보이면 그때 추가로 데이터를 확보해 확장해 나가도 된다는 생각으로 분석 범위와 가설을 수정해 나갈 수 있었다.

2017.11.25 데이터 분석 연합군 모집

데이터 분석 작업을 거둬해 나가면서, 금융권 내부 데이터 확보의 필요성을 깨닫게 됐다. 또한 분석을 위한 데이터 전처리와 후처리에 생각보다 많은 시간이 필요함을 느끼게 됐다. 이로 인해 주변의 도움을 받을 수 있는 채널을 확보해야 본 프로젝트를 성공적으로 마무리할 수 있다는 확신에 따라 연합군 모집에 온 힘을 쏟았다.

2017.12.05 데이터 분석 중간 리뷰

연합군들이 각각의 역할을 해 나가면서 우리 조는 중간 리뷰를 했다. 놀라운 결과를 만났다. 프로젝트를 시작했을 때만 해도 뭘, 어떻게 해야 할지도 모르고, 분석 방법론이 뭔지도 몰랐다. 하지만 서로 의견을 조율해 가면서 놀라운 변화를 경험했다. 생각하지 못했던 의미미한 결과들이 나오고 있었다. 서로간의 역할과 의무대로 한 사람이라도 만약 시간 내에 완수하지 못했다면, 본 프로젝트를 성공적으로 완수할 수 없었을 것이다.

2017.12.10 발표 전날 마무리를 위해 모이다.

마지막 발표를 남겨두고, 최종 점검을 위해 바쁜 일정에도 불구하고 신촌의 한 스터디 모임 장소에 모여 자료를 점검하고 보완할 사항들을 챙겼다. 힘들었지만 뿌듯한 순간이었다.

“아이디어와 데이터 확보 위해 뛰어다닌 보람 느껴”

정호열

신한데이터시스템 디지털전략연구소 수석



금융 빅데이터 융합

2기 우수조로 선정된 소감은.

생각지도 못했는데 우수조로 선정됐다는 발표를 듣고 놀랐다. 함께 고생한 조원들 모습들을 떠올려보니 우리가 그동안 했던 노력이 헛되지 않았구나 하는 생각에 가슴 한 칸이 뜨거워졌다. 서로 다른 환경에서 만난 사람들이지만, 하나의 목표를 갖고 열성적으로 할 수만 있다면 어떤 힘들고 어려운 상황이 닥쳐와도 극복해 낼 수 있다는 확신을 갖게 됐다.

금융 빅데이터 융합 전문가

과정을 수강하게 된 배경은.

금융권에 종사하지만 금융 빅데이터 분석 프로젝트를 데이터 수집부터 분석까지 풀 프로세스를 경험해 보기는 사실 쉽지 않다. 그래서 분석 경험을 쌓기 위해 빅데이터 아카데미를 선택했다. 집체교육 때는 예상 외로 깊이 있는 내용을 이해하기가 힘들기도 했다. 이를 잘 극복했기에 좀 더 유의미한 결과를 얻을 수 있지 않았나 싶다.

파일럿 프로젝트를 진행하면서 가장

힘들었던 부분과 어떻게 극복했나.

아이디어 선정과 데이터 확보가 어려웠다. 아이디어와 데이터를 확보하기 위해 데이터 분석 업무를 수행하고 있는 여러 사람들과

만나서 얘기하기 시작했다. 열띤 토론을 거쳐 아이디어를 도출해 낼 수 있었다. 우리 조가 목표로 했던 성과 달성을 위해 열성적으로 달려들었던 노력은 예상 외의 많은 로우 데이터 확보로 연결됐고, 그 결과 기대 이상의 분석 결과를 도출했다고 본다.

빅데이터 아카데미 수료 전과

후에 달라진 점이 있다면.

빅데이터 분석 프로젝트를 바라보는 시각이 달라진 것을 들 수 있다. 이론으로 아무리 많은 지식을 습득하더라도 직접 프로젝트를 수행한 경험이 없다면, 주제 선정과 데이터 확보가 얼마나 어렵고 고된 작업인지, 얼마나 중요한 작업인지를 몰랐을 것이다. 이러한 과정에서 많은 사람들의 협조가 필요하고, 소통이 원활히 되지 않는다면 아무리 멋진 알고리즘을 활용한다 한들 의미 있는 결과를 도출해 내기 힘들다는 사실을 알게 됐다.

파일럿 프로젝트를 발전 시킬 계획이 있다면.

주제 선정을 할 때부터 일회성 주제가 아닌 지속적으로 발전 가능한 주제를 찾으려 했다. 본문에 소개했듯이 아직 풀어나가야 할 숙제가 많다. 사내 학습 동호회(Community of Practice, COP), 공모전 아이디어 사업화 등 다양한 경로를 통해, 의지가 있고 뜻이 있는 사람들을 모아 발전시켜 나갈 계획이다.

빅데이터 아카데미 교육을 검토중인

분들에게 해주고 싶은 말이 있다면.

단순히 분석 알고리즘이나 분석 기술을 배우고 싶어서 왔다면, 많은 부분을 놓치게 될 것이라라는 말을 해 주고 싶다. DBMS에서 처리할 수 없는 빅데이터를 다루는 기술을 배우고 싶어서 지원했다면, 본 과정의 취지와 맞지 않았을 것이다. 본 과정은 현업의 입장에서, 내가 데이터 분석을 통해 무엇을 얻고자 하는지를 명확한 목적을 갖고 해야 뭔가를 얻을 수 있지 않을까 하는 생각을 했다. 인공지능 알고리즘들을 왜 배워야 하고, 어떻게 적용하는 것이 좋을지에 대해 끝없이 도전하고 싶다.

어떤 일을 하고 있으며, 데이터 분석과

관련해 어떤 목표를 갖고 있다.

금융권에서 빅데이터 관련 신사업 발굴 업무를 맡고 있다. 스타트업과 협업해 금융업과 상생할 수 있는 모델을 만들어 나가는 일을 하고 있다. 금융관련 빅데이터 사업 발굴을 위해 본 과정에서 경험했던 노하우를 적극 반영해 업무에 임해 나갈 계획이다. 좀 더 유망한 스타트업을 알아보고 발굴해 내고 싶다.

구매패턴 기반 구매감소 고객 예측



구분	빅데이터 분석, 예측 분석
적용 도구	기계학습(rf, ligit, dt), R 패키지, SQL 등
수집 데이터	2014~2015년 통합 멤버십 4개 제휴사 고객 정보 및 구매 이력 데이터, 국내외 유통사업 현안 및 전략 관련 자료, 소셜 데이터(네이버, 구글 트렌드)
산출물	구매패턴 기반 구매감소 고객 예측모형 및 대응방안
지도	이이백
참여자	박기범 ^{초장} , 변성원, 최정원, 최종현, 황덕열
프로젝트 소개	유통업계의 주요 이슈인 지속적인 구매감소 고객의 구매패턴을 분석해 구매감소 고객을 미리 예측하는 모형을 개발하는 프로젝트다. 구매 패턴에 기반한 적절한 대응책을 강구해 미리 대응함으로써 구매감소 고객을 최소화하기 위한 기반으로 활용할 수 있다.

THE CHALLENGES

프로젝트 시작 시점에 조원들이 브레인스토밍 방식으로 토론해 ‘유통 패러다임의 급격한 변화 및 유통사들의 대응 방안’을 관심 주제로 선정했다. 주제 관련 데이터를 입수·탐색하면서 프로젝트 과제를 구체화했다. 고객 만족의 중요성은 더욱 강조되는 반면, 비정형적·복합적인 고객 행동에 대한 유통업체 대응은 점점 어려워지고 있다. 이로 인한 구매감소 고객의 지속적인 발생이 유통업체의 현안이라는 사실에 주목했다. 실제로 국내 및 글로벌 대표 유통 브랜드의 온라인, 온오프 통합 분야에 대한 고객 반응도가 각각 상이하게 나오는 것을 네이버와 구글 트렌드 분석으로 확인했다. 고객은 채널과 무관하게 만족할 수 있는 쇼핑경험을 제공하는 기업에 반응하고, 그 반대의 경우는 외면하게 된다. 고객이 만족할 수 있는 경험가치를 어떻게 제공할 수 있을까? 그 해답은 다양한 고객 니즈와 구매 패턴에 기반한 고객 경험을 제공해야 한다는 사실에서 찾아볼 수 있다.

몇 차례 논의와 수정을 거듭한 후에 ‘구매패턴 기반 구매감소 고객 예측 및 대응방안’을 프로젝트 과제로 선정하는 데 전원 의견 일치를 보았다. 데이터 수집, 전처리, 변수 선정, 예측분석, 대응방안 수립 등 본격적인 프로젝트에 착수했다. 예상하지 못한 문제에 직면해 상이한 해법으로 갈등을 겪기도 했다. 하지만 과제 선정 과정에서 형성된 조원간 공감과 신뢰를 바탕으로 문제를 하나씩 풀어나갈 수 있었다.

그림 1

다양한 고객 니즈와
구매 패턴의 상존



THE APPROACH

데이터 수집

고객 구매패턴 변동을 파악할 수 있는 데이터로 유통업체의 통합 멤버십 고객 정보, 구매 이력 데이터, 비즈니스 도메인 정보로 온/오프라인 유통사업 현안 자료를 수집했다. 네이버·구글 트렌드 분석 데이터도 활용했다. 학술연구 목적 이고 공개된 데이터를 수집하는 경우였지만 불필요한 오해를 방지하기 위해 특정 업체명은 모두 익명 처리했다.

데이터 탐색 및 변수화

데이터 수집 후 탐색 및 변수화 과정에서 직면한 첫 번째 이슈는 수집된 데이터의 전처리 문제였다. 통합 멤버십 고객 2만 명이 4개 제휴사와 거래한 2800만 건의 방대한 구매이력 데이터를 탐색하면서 고객별로 데이터 정리가 쉽지 않았다. 엑셀이나 R에서는 데이터 처리 자체가 불가능했다. SQL로 고객별 월별 데이터로 1차 가공을 했다. 가공된 데이터를 R로 가져와서 어렵게 탐색을 수행할 수 있었다.

두 번째 이슈는 업체별로 상품 분류 기준이 상이한 문제였다. 상품별로 고객 구매 패턴을 분석하지 못할 수도 있겠다는 불안감이 밀려왔다. 의미 있는 데이터 해석의 전제 조건인 표준화한 상품 분류 기준을 작성했다. 관련 데이터를 새로운 기준에 맞춰 다시 분류할 필요가 있었다. 변성원 조원의 헌신적인 노력으로 상품 유형별·수준별로 표준화한 상품 분류기준을 완성하면서 문제 해결의 실마리를 찾을 수 있었다. 결국 상이한 4개 제휴사 상품 분류기준을 14개 카테고리, 3개 레벨로 표준화할 수 있었다.

세 번째 이슈는 구매패턴의 변동요인을 반영할 수 있는 변수를 도출하는 문제였다. 고객별 구매변동 패턴을 반영하는 구매변동 및 증감지수를 산출하면서 분석의 기초가 되는 45개의 변수를 얻을 수 있었다. 데이터 탐색 및 변수화 작업의 마지막 이슈는 변수의 적합성 검증이었다. rf의 importance 기능을 활용해 변수의 중요성을 확인하고, 45개 변수 중 정확도와 불순도 개선 기여도 측면에서 중요한 32개의 변수를 추려냈다. 변수의 분산 값이 0에 가까운 의미 없는 변수 3개를 추가로 제외한 후에 29개의 변수를 최종 선정했다. 최종

고객속성 변수		구매증감 변수		구매변동 변수	
변수명	설명	변수명	설명	변수명	설명
Cust_No	고객번호	Net_H_A	반기 증감(총구매금액)	Fluc_H_A	반기변동(총구매금액)
Gender	성별	Net_H_Q	구매건수 증감	Fluc_H_Q	구매건수 변동
Mobile	모바일 이용횟수	Net_lux_idx	사치품(분기) 증감	Fluc_lux_idx	사치품(분기) 변동
lux_Ratio	사치품 비중	Net_cho_idx	선매품(분기) 증감	Fluc_cho_idx	선매품(분기) 변동
Cho_Ratio	선매품 비중	Net_life_idx	일용품(분기) 증감	Fluc_life_idx	일용품(분기) 변동
Food_Ratio	식품 비중	Net_A_Channel	증감지수_A사	Fluc_A_Channel	변동지수_A사
		Net_B_Channel	증감지수_B사	Fluc_B_Channel	변동지수_B사
		Net_C_Channel	증감지수_C사	Fluc_C_Channel	변동지수_C사
		Net_food_high	식품(고가) 구매금액 증감	Fluc_food_high	식품(고가) 구매금액 변동
		Net_food_mid	식품(중가) 구매금액 증감	Fluc_food_mid	식품(중가) 구매금액 변동
		Net_food_low	식품(저가) 구매금액 증감	Fluc_food_low	식품(저가) 구매금액 변동

Y (Target Var.) 구매감소 고객(감소:1)

표 1

29개의 변수를
최종 선정

선정된 변수 29개는 구매감소 고객 여부를 대변하는 반응변수 1개와 설명변수 28개다. 설명변수는 고객속성 변수 6개, 구매증감 변수 11개, 구매변동 변수 11개로 구성됐다.

예측모형 구축

최종 선정된 29개 변수를 적용·분석한 예측모형 수행의 첫 단계는 ‘구매감소 고객을 어떻게 정의할 것인가?’였다. 2년간 고객 데이터를 반기별로 4기로 구분해 1~3기는 데이터 예측 모형, 4기는 평가용으로만 각각 활용하기로 했다. 1기와 4기를 비교해 평균 구매 증감을 대비 하락한 고객을 구매감소 고객으로 정의했다.

두 번째 단계는 예측모형 수행 방법이었다. 기본적으로 1기와 4기를 비교해 구매감소한 고객의 1~3기 기간 중 구매패턴을 기계학습으로 훈련시켜 4기

에 대한 예측모형을 생성하는 방식이었다. 예측모형 평가를 위해 전체 데이터를 7:3으로 랜덤 샘플링해 트레인 데이터와 테스트 데이터로 분리했다. 트레인 데이터를 적용해 생성된 예측모형을 테스트 데이터로 검증했다.

세 번째 단계로 분석할 예측모형을 선정했다. 분류분석의 대표적인 모형인 랜덤포레스트(Random Forest), 로지스틱 회귀분석(Logistic Regression), 디시전 트리(Decision Tree)를 선택했다. 랜덤포레스트로 예측모형을 수행한 결과, 예측 정확도는 상대적으로 양호한 78.05%로 나왔다. rf 일차 수행 시 중요 파라미터인 ntree(나무의 수)와 mtry(노드 분류기준으로 고려할 변수의 개수)는 각각 디폴트 값으로 500과 5가 적용됐다.

```

cust_rf <- randomForest(y~.,data=train01,importance=TRUE)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 5

OOB estimate of error rate: 21.76%
Confusion matrix:
  0   1 class.error
0 4707 1673  0.2622257
1 1266 5860  0.1776593

```

ntree와 mtry가 각각 디폴트값으로 500과 5로 수행된 것에 대한 적합성을 검증했다. ntree는 500까지는 미세하게 에러가 축소되고 있었다. mtry는 5일 때 에러가 최소화되는 것을 확인함으로써 예측모형 수행 결과를 신뢰할 수 있었다. 로지스틱 회귀분석(Logistic Regression) 모형 수행 과정에서, 일차적으로 28개 설명변수 모두를 트레인 데이터에 적용한 logit 모델을 step function에 적용했다. AIC(Akaike Information Criterion) 값이 최저 수준인 12,758일 때의 16개 변수를 얻을 수 있었다. Summary를 통해 베타 계수 값

그림 2

500까지 미세하게 에러가 축소된 ntree

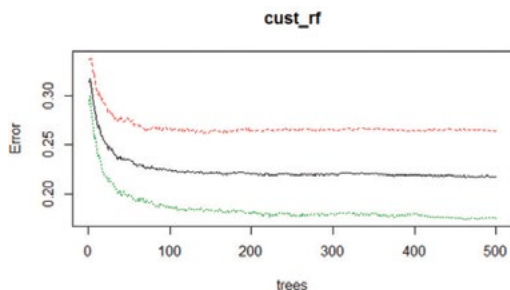


그림 3

5일 때 에러가 최소화된 mtry

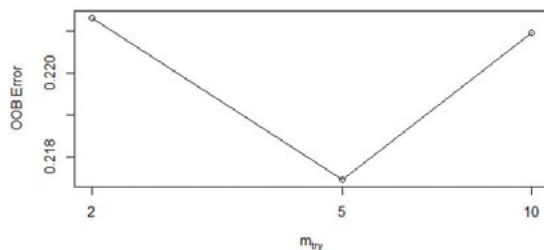


그림 4

AUC가 0.86으로 양호해 정확한 예측모형으로 판단

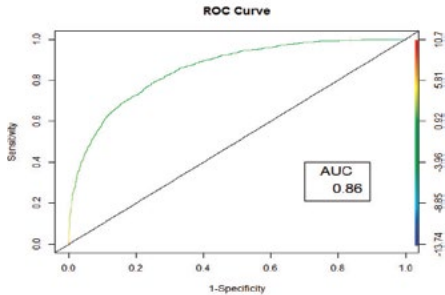


표 2

각 모형의 인사이트를 구매감소 고객에 대한 대응방안 수립에 활용

구분	RF	Logit	DT(ctree/rpart)
예측모형 평가	Accuracy : 0.78	Accuracy : 0.77	Accuracy : 0.74
	Sensitivity : 0.73	Sensitivity : 0.72	Sensitivity : 0.71
	Specificity : 0.83	Specificity : 0.81	Specificity : 0.76
활용변수	29개	16개(최적 AIC 기준)	29개
인사이트	구매증감지수 중요도 1순위 - 전체>A사>B사의 순으로 중요 - 일용품 및 저가식품 증감지수 유의 필요	구매증감지수 기여도 1순위 - 전체>A사>B사>C사의 순으로 중요	구매증감지수 중요도 1순위 - 전체>A사>B사의 순으로 중요
	제품 레벨 비중의 영향도 상위권 위치 - 사치품 비중, 중요도 6위 - 선매품 비중 및 증감, 변동지수 유의미	구매건수 증감지수 기여도 상위권 위치 제품 레벨 비중의 기여도 유의미 - 사치품 비중, 기여도 상위권	구매감소 예상고객 사전 인지 기준 제시 구매증감지수 ≤ -1 구매증감지수 > 1, A사증감지수 ≤ -1 A사증감지수 > 1, 구매건수증감지수 ≤ 1 구매건수증감지수 > 1, B사증감지수 ≤ -1
	제품 카테고리별 저가식품 증감, 변동지수	제품 카테고리별 증가식품 증감지수	
	모바일의 영향력이 온라인보다 중요	모바일 이용고객, 구매감소 가능성 낮춤	

을 확인하는 과정에서 16개 변수의 예측 기여도를 파악하고 인사이트를 발굴할 수 있었다.

생성된 logit 모형에 테스트 데이터를 적용해 예측점수를 산출했다. ROC Curve를 사용하기 위해 예측점수와 테스트 데이터의 실제 y값으로 prediction 객체를 만들었다. 이를 performance() 함수에 넘겨 tpr(구매 감소하지 않는 고객을 구매 감소하지 않았다고 예측) 및 fpr(구매 감소하는 고객을 감소하지 않았다고 예측)을 구했다. tpr과 fpr을 각각 X, Y축으로 하는 ROC Curve

를 이용해 성능을 최대화하는 Cutoff value 0.04를 선택했다. ROC Curve 아래 면적을 의미하는 AUC가 0.86으로 양호해 예측모형은 정확하다고 판단할 수 있었다.

Cutoff value 0.04를 기준으로, 예측 값을 0과 1로 분류해 주고 실제 y값과 같이 confusionMatrix를 수행한 결과 예측모형의 정확도 76.99%를 구할

구분	구매 패턴			비중	대응 방안
	1차	2차	3차		
구매 감소	구매증감지수 ≤ 0	A사 증감지수 ≤ 0	구매증감지수 ≤ -1	19%	<ul style="list-style-type: none"> - 85% 이상 구매 감소 가능고객으로 우선적으로 대책 추진 - 일용품, 선매품, 저가식품 증감지수 및 A사 사치품 비중 추가 분석 - 전체 유통 구매 감소 Pain points 대응 프로그램 실행
	구매증감지수 ≤ 0	A사 증감지수 ≤ 0	구매증감지수 > -1 A사 증감지수 ≤ -1	9%	<ul style="list-style-type: none"> - 85% 이상 구매 감소 가능고객으로 A사 구매 감소가 주요인 - A사 사치품 및 선매품 비중, 선매품 증감 변동 요인 반영한 A사 구매 감소 Pain points 대응 프로그램 실행
	구매증감지수 ≤ 0	A사 증감지수 ≤ 0	구매증감지수 > -1 A사 증감지수 > -1 B사 증감지수 ≤ 0 구매건수증감지수 ≤ 1 B사 증감지수 ≤ -1	6%	<ul style="list-style-type: none"> - 70% 이상 구매 감소 가능고객, B사 구매 감소가 주요인 - 일용품, 선매품, 저가식품 증감지수 추가 분석 - 제품 카테고리별 감소, 특히 식품 구매증감 요인 파악후 B사 구매 감소 Pain points 대응 프로그램 실행
	구매증감지수 ≤ 0	A사 증감지수 ≤ 0	구매증감지수 > -1 A사 증감지수 > -1 B사 증감지수 ≤ 0 구매건수증감지수 ≤ 1 B사 증감지수 > -1	20%	<ul style="list-style-type: none"> - 60% 이상 구매 감소 가능고객이며 고객 비중이 높은 유형으로 효과적인 대응 위한 상세 추가 분석 후 대응 프로그램 실행 - B사, C사 구매 증감지수 - 일용품, 선매품, 저가식품 증감지수 - 구매건수의 감소 요인
	구매증감지수 > 0			22%	<ul style="list-style-type: none"> - 구매감소 가능성 20% 미만
구매 증가	구매증감지수 ≤ 0	A사 증감지수 > 0		16%	<ul style="list-style-type: none"> - 구매감소 가능성 30% 미만이나 A사 외 B사, C사 구매 감소 발생 가능성에 대한 대응 방안 필요
	구매증감지수 ≤ 0	A사 증감지수 ≤ 0	구매증감지수 > -1 A사 증감지수 > -1 BTK 증감지수 > 0	7%	<ul style="list-style-type: none"> - 구매감소 가능성 40% 미만으로 구매 감소 전환 리스크 대비 - A사 구매 감소 요인 점검 및 대응 필요

수 있었다. 디시전 트리 모형은 예측 정확도가 74%로 조금 떨어졌다. 하지만 구매감소 고객 예측에 적용되는 핵심 변수의 기준 값을 구체적으로 제시해 준다는 측면에서 유용하게 활용했다.

표 3

구매감소 고객의 구매 패턴 유형별 대응방안

예측모형 평가 및 인사이트 도출

3가지 예측모형의 분석결과를 종합적으로 평가해 정확도가 상대적으로 양호한 rf 모형을 기본 예측모형으로 사용하기로 했다. rf는 물론 logit이나 dt 모형에서 파악한 인사이트를 종합해 구매감소 고객에 대한 대응방안 수립에 활용했다.

구매패턴 유형별 구매감소 고객 대응방안

예측모형에서 파악한 인사이트를 기반으로 구매감소 고객의 구매패턴을 유형화하고 대응방안을 마련했다. dt의 노드 분류 기준 값이 더 구체적인 유형화의 기준을 제공해 주었다. 고객의 구매패턴을 감소고객 4개, 증가고객 3개를 포함한 7개 유형으로 구분했다. 유형별 구매패턴에 대한 대응 방안을 수립했다. 각 유형의 구매패턴을 보이는 고객의 미래 구매감소 여부를 예측하고, 대응 조치를 사전적으로 실행할 수 있는 가이드가 마련됐다.

THE OUTCOME

요즘 들어 고객 경험의 가치가 더욱 더 중요해지고 있다. ‘구매패턴에 기반한 구매감소 고객 예측모형’은 시장환경에서 유통업체들이 고객의 구매행동을 더 구체적으로 이해하고 고객의 어려움(Pain Points)에 대해 사전에 대응함으로써 구매감소를 최소화할 수 있다는 점에서 의미를 찾을 수 있다. 이번에 만든 모형은 다양한 고객의 구매패턴 및 업종별 비즈니스 도메인의 특성을 반영할 수 있는 추가 변수를 개발해 좀 더 정교하고 완성도를 높여갈 계획이다. 특히 급변하는 유통 환경에서 고객 경험 가치 제고를 더 이상 방치할 수 없음에도 구체적인 실행 방안의 부재로 어려움을 겪고 있는 중소기업 및 소상공인에게 특화된 ‘구매패턴 기반 구매감소 고객 예측모형’으로 발전시켜 나갈 계획이다.

반전의 반전을 거듭한 프로젝트

우리 조는 다양한 배경을 가진 조원들이 모여서 과제를 함께 풀어 나가면서 시각 차이는 물론 가능성과 포기를 넘나드는 반전에 반전을 거듭하여 목표 지점에 도달했다. 조원들 사이에 서로 신뢰를 바탕으로 각자 맡은 일을 끝까지 수행한 결과 프로젝트를 완성할 수 있었고, 소중한 배움의 기회도 얻었다.

2017.09.23 프로젝트 과제에 대한 자유로운 의견 교환

관심 주제에 대해 제한을 두지 않고 자유롭게 토의하면서 흥미를 끌 만한 주제들이 거론됐다. 이때까지는 조원 모두 프로젝트 수행에 대한 의지가 넘쳤다.

2017.09.30 데이터 탐색 과정에서 드러나기 시작한 우려

프로젝트 과제 확정 후 데이터를 탐색하면서 데이터 전처리, 변수 작업 과정에서 여러 가지 이슈에 직면했다. 이때부터 '그건 어렵다!'는 목소리가 흘러 나오기 시작했다. 특히 상이한 상품분류 기준으로 구매패턴 변동에 상품을 포함시켜야 할지 여부가 이슈였다. '나중에 수정하더라도 할 일은 우선 다해 보자'는 입장을 견지하면서 상품 변동 내역도 변수화하기로 조원들과 어렵게 의견 일치를 보았다. 그래도 분위기는 우려감으로 상당히 무거웠다.

2017.10.14 상품 분류기준의 표준화로 본 궤도에 오른 변수화 작업

상품 분류기준을 표준화하면서 침체되었던 분위기가 반전됐다. 구매패턴의 중요한 항목인 상품 구매변동 내역을 변수화할 수 있었다. 본격적인 변수 선정을 위한 전처리 작업이 진행됐다.

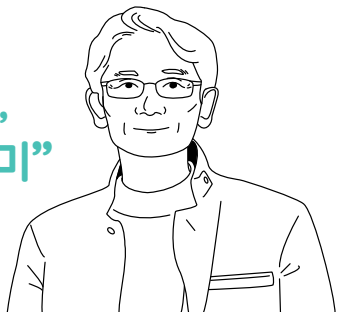
2017.10.21 분석모형의 선정과 부족한 프로젝트 수행 시간

1차 완료된 45개의 변수를 활용해 logit과 디지전트리라는 2개의 예측 모형을 수행했다. 하지만 결과가 다소 미흡해 구매패턴을 유형화하기에는 역부족이었다. 분위기는 다시 가라앉았다. 랜덤포레스트 모형의 importance 기능으로 변수의 중요도에 대한 인사이트를 추가할 필요가 있다는 멘토의 제안이 나왔다. 남은 시간은 일주일뿐이었다. 고민 끝에 결국 rf를 수행했고 멘토의 제안이 적중했음을 확인했다.

2017.10.26 최선을 다한 후 덕담으로 마무리 한 그룹콜

그룹콜로 최종 마무리할 발표자료를 조원들이 함께 점검하면서 '그간 수고했다'고 서로를 격려하면서 프로젝트가 잘 마무리됐음을 실감했다.

“새로운 의미와 가치 창출의 시작, 빅데이터 아카데미”



박기범

비엔बी컨설팅 대표

빅데이터 아카데미 수강 배경과

어떤 목표를 갖고 임했나.

개인적으로 '데이터 기반 리테일 매장 운영' 프로젝트를 진행하고 있다. 여러 가지 부족함을 느끼고 있던 차에 빅데이터 아카데미에 유통전문가 과정이 개설된다는 것을 알고 기대를 갖고 지원했다. 빅데이터 아카데미 교육 및 프로젝트에서 유통 현안을 데이터 과제화하고 데이터 수집·분석·해결책 도출까지의 전 과정을 수행할 수 있는 역량 확보의 기회로 삼겠다는 자세로 몰입했다.

심사위원 전원 최고 점수로

우수조로 선정됐다고 들었다.

첫 각오는 굳게 했으나 프로젝트 진행 과정, 특히 전처리 과정에서 여러 가지 난관에 부딪치면서 변수 선정문제에 있어서 일부 타협도 생각했다. 분석에 필요한 주요 변수를 절대 포기하지 않고 시간이 걸리더라도 마침내 도출해 예측모형 수행에 적용한 것이 양호한 분석 결과와 의미 있는 대응방안으로 연결됐다고 본다.

프로젝트 과정에서 기억에 남는 순간과 미처 생각하지 못했던 부분이 있었다면.

데이터 전처리 과정에서 분석 대상인 4개 제휴사의 상품 분류기준이 상이한 점을 해결하는 과정에서 서로 의견이 엇갈렸다.

그럼에도 끝까지 문제를 풀어나가려는 의지를 지켜 나가면서 조원들간 공감대를 조성하려고 노력했다. 마침내 기대 이상의 표준화한 상품 분류기준이 만들어졌다. 이로써 고객 구매감소 패턴에 있어서 더 의미 있는 상품 측면의 인사이트를 얻을 수 있게 된 점이 가장 기억에 남는다.

프로젝트 조원들 구성 시에

에피소드도 있었지 싶다.

운 좋게도 우리 조는 프로젝트를 하는 데 최적의 조합이었다. 리테일 매장 운영과 컨설팅을 하는 조장과 빅데이터 서비스를 하는 최종현 조원이 비즈니스 도메인을 담당했다. IT 솔루션 회사 임원인 황덕열 조원, DB 정제와 통계 및 분석 전문가인 최정원 조원, 홈쇼핑 PI 매니저인 변성원 조원이 프로젝트의 전반적 운영으로 자연스럽게 역할을 분담했다. 혼자서는 도저히 해결하기 어려운 문제를 협업 과정에서 해결의 실마리를 찾아냈다. 협업의 시너지 효과를 경험했다.

주제를 발전시켜 실무에도 적용할 수 있다고 했는데 조원들과 함께 발전 계획이 있나.

충분히 가능하다고 본다. 우선 이번 프로젝트가 유통 대기업을 대상으로 수행됐지만 정작 고객 기반 데이터 마케팅의

사각지대인 중소 유통업체나 매장 운영을 하는 소상공인들에게도 확대할 수 있다. 개인적으로 진행하고 있는 프로젝트도 '리테일 매장 운영을 데이터 기반으로 수행'인데 이번 과정에서 습득한 역량을 적극 활용해 데이터 수집, 변수개발, 예측모형 활용 및 고객 구매패턴에 기반한 마케팅 고도화 등 여러 분야의 개선에 활용할 계획이다. 더 나아가 중소기업, 소상공인에 특화된 '구매패턴 기반 구매감소 고객 예측모형'은 물론, 업종별 모형도 개발할 수 있다. 감에 의존해 매장 운영과 고객 대응을 해야 하는 소상공인 및 중소기업체들을 지원하는 일을 할 수 있지 않나 생각한다.

빅데이터 아카데미 수강을 검토중인 분들에게 해주고 싶은 말이 있다면.

내 주변에는 화려한 비즈니스 경험을 제대로 활용하지 못하는 이들이 적지 않다. 경험을 통해 구축한 비즈니스 통찰력을 바탕으로 핵심 변수를 선정하고 데이터 과제화해서 의미 있는 결과를 도출할 수 있을 텐데도 시도조차 못하는 것을 보곤 한다. 그런 이들에게 빅데이터 아카데미 교육은 새로운 의미와 가치를 만들어 내는 기회를 제공할 것이다.

빅데이터 분석을 통한 최적 사육 환경 조성



구분	구조적 정의
적용 도구	Oracle RDBMS, Spotfire, R studio
수집 데이터	2016년 1~12월 사이 36개 표본 농가의 농장 구성 및 개체별 종량 데이터
산출물	개체 균일도의 침도(분포의 뽀족한 정도를 나타내는 통계량)를 높이는 영향인자 확인
지도	안진훈
참여자	박정훈 ^{초장} , 김영진, 김자중, 박성주, 이만영, 최병환
프로젝트 소개	닭의 사육 생산성을 확인하는 요소로서 체중별 산포균일도, 체중, 일별증체증량 등이 있다. 이중 체중별 산포균일도를 높일 수 있는 방안들의 연구는 적었다. 이에 표본 농가들의 데이터를 분석해 침도가 높은 농가를 확인했다. 그 결과에 영향을 준 요소가 무엇인지 영향 인자를 찾아 최적의 사육 환경을 조성하기 위한 프로젝트였다.

THE CHALLENGES

처음 2주간 진행되는 집체교육 스케줄과 교육 후 5주간 진행되는 파일럿 프로젝트 설명을 들었을 때 걱정부터 앞섰다. 2주간 집체교육 후 5주간 진행되는 파일럿 프로젝트에 대한 심리적 부담이 첫날부터 머릿속을 가득 채웠다. 이곳 교육장에서 처음 만난 사람들과 짧은 기간에 분석 결과물을 도출하는 게 가능할까 싶었다. 하지만 빅데이터 교육 참석 전에 막연하게나마 직접 해보고 싶다는 주제 하나씩은 가지고 있었기에 정면으로 부딪히면서 결과를 뽑아내야 할 수밖에 없었다. 일종의 벼랑 끝 전술이라고 해야 할까. 제조 전문가 교육이라는 기준에 맞춰 제조와 관련된 주제로 한정해야 하는 부분이 너무 큰 부담으로 다가왔다.

닭 생육·가공·유통업에 몸담고 있던 조장이 조심스럽게 제안했던 주제 ‘빅데이터 분석을 통한 최적 사육 환경 조성’이 선정됐다. 우리 조장의 얼굴은 금방 환해졌다. 엉뚱한 주제로 받아들여지지 않을까 하고 염려했던 조장의 예상과 달리, 조원들 모두 재미있을 것 같다고 환영했기 때문이다. 뭔가 일을 낼 수 있겠구나 하는 희망의 빛이 보이기 시작했던 순간이었다. 데이터와 해당 분야에 대한 지식을 갖고 시작한다는 점은 매우 큰 장점이다. 직면할 위험을 크게 낮출 수 있기 때문이다.

THE APPROACH

사육 환경에서 생산성을 판단하는 기준은 여러 가지가 있다. 사료 요구율, 출하 시점의 중량, 일별 증체량, 육성율 등 여러 판단 기준이 존재한다. 닭의 개체별 균일도 관점은 그동안 접근하기 어려운 부분이었다. 성적이 좋다, 나쁘다는 기준을 세우기가 어렵다. 이에 조장은 근무하는 회사에서 1월에서 12월까지 월별로 이틀 간 도계(屠鷄)한 농가들의 데이터를 확보했다. 각 농가의 표준편차를 조사해 침도가 높은 농가순으로 비교했다. 농장 환경, 온도, 습도, 모계의 주령(나이), 시설을 구성하는 요소를 대입해 어떠한 부분이 균일도(침도)를 높일 수

있는지 분석해 보았다.

제조실행시스템(Manufacturing Execution System, MES)에서 뽑아온 2016년 도계한 농가 대상의 개체별 중량 데이터였다. 주제가 선정됐고 데이터까지 쉽게 확보했지만 문제는 이 순간부터 찾아왔다. 진도를 빼지 못하고 갈팡질팡하기 시작했다. 주제와 데이터를 순조롭게 확보하면서 겪지 못한 어려움을 데이터 전처리 과정에서 만난 것이다.

데이터 정제 작업에서 시행착오가 반복됐다. 조원들도 처음에 보였던 열의 대신 불안해 하기 시작했다. 전환점 모색이 필요한 순간이었다. 이때 조원들이 모여 해당 도메인 지식을 다시 한 번 공유하고, 조원들 각자의 목표가 무엇인가?에 대해 생각해 보기로 했다. 조장이 닭 사육·가공·판매 업무에 대해 조원들에게 다시 한 번 상세하게 소개했다. 분석 포인트가 잡히지 않아서 발생한 문제일 수 있다고 봤기 때문이다. 최대한 밝은 분위기 속에서 질문과 답을 주고 받으며 도메인 지식을 공유했다. 이 과정에서 생각하지 못했던 뭔가가 도출되기를 바랐다. 조장도 분석 목표를 다시 한 번 명확히 생각해 보기로 했다. 당연하다고 받아들였던 것에서 발생 가능한 실수를 찾아보기 위해서였다. 이 과정을 거친 후 분명한 변화가 나타났다. 마음을 가다듬고 먼저 농가별 성적 분석을 실행해 성적에 대한 결과 인자가 무엇인지 알아보았다. 그에 맞춰 데이터 정제 작업을 몇 번 반복해 원하는 데이터 집합을 만들어 낼 수 있었다

데이터 탐색과 변수 도출

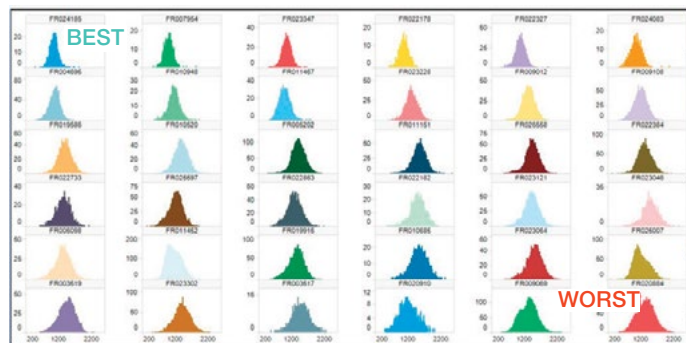
해당 농가에 대한 농장의 환경, 즉 환경적·시설적인 요소들을 확보해 유효인자를 최대한 도출하기로 했다. [표 1]의 변수들을 기준으로 농장별 닭의 체중 분포에 영향을 주는 인자를 파악했다. 침도가 높게 분포된 농장은 시장 수요에 맞춰 원활하게 닭을 공급할 수 있다. 반면 침도가 낮게 분포된 농장은 표준 대비 균일도에 대한 차이가 크다. 따라서 시장에서 원하는 중량의 닭을 원활하게 공급하기 어렵다. 이는 곧 손실로 연결된다.

농장별로 닭의 체중 분포를 확인한 결과는 [그림 1]과 같다. 좌측 상단이 가장 좋은 분포를 보이는 농장들이다. 우측 하단이 가장 나쁜 분포를 보이는 농장이다. 과연 어떠한 인자들이 이러한 분포도를 도출하게 했는지 주요 인자를 분류하고자 했다.

표 1
변수와 설명

구분	인자	정의
지역	지역	시·군 단위 지역
농장	농장 역할 번호	농장 번호, 농장을 구분하는 용도
사육 규모	계사 평수 합계	농장의 계사 합계 평수
	계사 수 합계	농장을 구성하는 계사의 개수
	사육 규모 수	사육 규모 입주 가능 수수로서 규모에 따른 성적 차이
농장시설 형태	종합 계사 측벽 형태 코드	계사의 측벽 형태(시설 형태에 따라 사육 성적 연관성)
	종합 계사 지붕 형태 코드	계사 지붕 형태(시설 형태에 따라 사육 성적 연관성)
	종합 계사 환기 형태 코드	계사 환기 형태(시설 형태에 따라 사육 성적 연관성)
	계사 단열 계수 합계	단열재에 따라 온도 관리 주요 요소가 높을수록 성적에 비례하는지 검토
	계사 형태 구분 코드	계사의 형태, 무창/유창에 따른 성적 분석, 무창(無窓) 계사에 대해 신축 유도
	시설 등급 코드	계사의 측벽, 지붕, 환기 등급 등을 종합한 전체 시설 등급
	종합 등급 코드	시설 등급, 원가 등급 등을 종합한 전체 등급
	용지 형태 코드	계사 위치와 용지 형태(용지 형태에 따른 성적 분석)
평가 방법 코드	사육 정산 평가 방법 코드	평가 방식에 따른 성적 분석, 절대평가 성적순, 상대평가 성적상 대비율 적용
온 습도	온 습도	해당 지역의 최대/최소 온도 및 습도
모계영향	종계 주령	종계 모계의 주령(나이)에 따른 연관성

그림 1
농장별
닭의 체중 분포



분석 목적이 예측이 아닌, 주요 변수를 도출하고 그 변수의 관리선을 정하
고자 했기에 머신러닝·딥러닝 알고리즘보다는 신뢰도가 높고 전통적인 방법

인 회귀분석과 의사결정나무를 이용했다. 연속형 변수인 표준편차를 3개의 범주형 변수인 그룹으로 나눈 다음, 다양한 방면으로 예측했다. 연속형 변수는 회귀분석으로, 범주형 변수는 의사결정나무로 각각 분석했다.

랜덤포레스트 분석으로 주요 변수 탐색

균일도 그룹을 분류하는 가장 중요한 인자들을 찾기 위해 랜덤포레스트 분석을 했다. 여기서 찾은 인자들을 토대로 데이터 선별(feature selection) 분석을 했다. 분석 결과는 프로젝트 초반에 예상했던 계절적 요인이나 농장 종합등급 코드와 다르게 나왔다. 즉 사육 규모나 계사 평수 합계가 주요한 변수로 예측됐다. 산출된 주요 변수를 가지고 다음 과정의 분석에 착수했다.

의사결정나무로 변수 간 원인 탐색

분류되는 변수와 수치를 통해 농장을 관리할 수 있을 것으로 예상하고 의사결정나무(Decision Tree) 분석을 했다. 의사결정나무는 어떤 변수들이 어떤 수치로 분류되는지 균일도 등급을 탐색하기 위해 하는 분석이다. 분석 결과 계사 평수 합계가 가장 주요한 변수로 떠올랐다. 분리의 좌우측의 변수들을 살펴보니 사육규모 수가 한쪽에만 치우쳐 있었다. 당연한 결과이겠지만 교호작용(reciprocal action, 둘 이상의 사물·현상이 서로 작용해 원인이 됨)이 있을 것으로 예상됐다.

다중공선성 분석으로 독립변수들 간 관계 확인

회귀분석 이전에 다중공선성 분석을 했다. 다중공선성 분석은 변수 간 다중공선성(multicollinearity, 독립변수가 다른 독립변수들과 완벽한 선형 독립이 아닌 경우)을 확인하여 이를 제거함으로써 회귀분석의 결정계수(Adjust R square) 값을 높이는 과정이다. 이를 통해 변수들 사이의 선형 관계를 파악할 수 있다. 분석 결과 계사 평수 합계와 사육 규모 사이에 높은 다중공선성이 있음을 확인했다. 변수를 후진 소거하며 선형회귀 분석을 했다.

선형회귀 분석으로 계사의 크기와 개체 균일도 관계 파악

계사의 평수 합계가 개체 균일도와 어떤 관계를 갖는지 파악할 필요가 있었다.

이를 회귀식으로 표현하기 위해 선형회귀 분석을 했다. 선형회귀 분석 결과 계사 평수 합계와 표준편차 사이 음의 관계가 있는 것으로 나타났다. 또한 유의수준(Significant Level)을 통해 변수의 중요도를 재확인했다.

THE OUTCOME

분석을 종합해 볼 때 사육 규모와 계사 평수 합계가 작을수록 침도가 높은 이상적인 표준편차를 갖는 농장으로 나타났다. 즉 작은 평수의 양계장이 더 우수한 농장이라는 의미다. 대규모 농장에서는 온도, 습도, 이산화탄소 등의 관리가 소규모 농장에 비해 어렵다. 그래서 소규모 농장에서 기른 닭의 체중 분포가 고르게 나왔을 것이다. 당초 예상했던 주요 변수와 다른 변수가 도출됨으로써 데이터 분석의 힘을 느꼈던 순간이었다.

예상과 다르게 나온 분석 결과

농장 내 지점마다 농장의 상세 환경을 알 수 있는 변수가 있다면 추가 분석을 할 수 있었을 텐데 그 변수가 없었다. 또한 농장 내 온·습도의 데이터가 구축되지 않아 더 정확한 분석도 어려웠다. 프로젝트 중에 데이터 정제 시간이 많이 부족했다. 이 것이 못내 아쉽다. 이미 관리되고 있는 데이터라고 해도 정제는 반드시 필요함을 알게 됐다. 분석 전에 유효인자 데이터를 미리 수집해야 하고, 데이터 전처리 과정에 힘을 쏟아야 함을 몸소 체험했다.

최종 보고서를 완료하고 유의미한 인사이트를 얻지 못한 거 같아서 아쉬움도 있었다. 이때 한 조원이 ‘이렇게 짧은 기간에 유의미한 정보를 얻는 것 자체가 위험하지 않나?’ 하는 의견을 내놓았다. 일리가 있는 말이었다. 학습 차원에서 진행했던 파일럿 프로젝트였지만, 그래도 아쉬운 건 어쩔 수 없다. 향후 보완해야 할 부분과 분석 방법론을 학습했다는 것이 가장 큰 소득이었다.

협업의 힘을 실감했던 순간들

제조 빅데이터 융합 전문가 과정 2기 2조는 빨리 친해졌다. 서로 의견을 존중하고 배려하며 유의미한 결과를 얻기 위해 열정적으로 파일럿 프로젝트에 임했다.

2017.10.10 주제를 확정해 놓고도 교차했던 불안감

처음 조별 모임에서 제조에 특화된 주제를 놓고 의견을 주고 받았다. 제조와 관련된 주제로 좁혀지면서 '주제 도출이 너무 어렵다'는 의견이 나왔다. 이때 박정훈 조장이 머릿속에 그려왔던 주제를 조원들에게 조심스럽게 털어놓았다. 너무 낯설다고 할 법한 주제에 대해 조원들이 예상 외로 흔쾌히 동의했다. 한편으로 너무 특정 회사에 치우친 주제가 아닐까? 하는 걱정과 함께 과연 결과를 잘 도출해낼 수 있을까? 하는 불안감이 교차했다.

2017.10.17 표준 데이터 확보를 놓고

주제를 제안한 박 조장이 데이터를 확보해야 했다. 교육 시작 전, 조장은 소속된 회사가 매우 많은 데이터를 갖고 있을 거라고 생각했다고 한다. 하지만 실재는 그렇지 않았다. MES를 통해 축적되는 닙의 개체별 종량 데이터는 수천만 건이 존재했다. 하지만 그 데이터를 분석하기 위한 유효인자 성격의 변수들은 막상 펼쳐 놓고 보니 너무 부족했다. 주제 선정을 잘못했나? 하는 불안감이 밀려왔다. 어떤 분석도 완벽한 상황에서 진행되지 않을 것이라고 생각하고 다시 힘을 내기로 했다. 현재 수준에서 최선을 다해 준비하기로 했다.

2017.11.25 데이터 분석

주제와 목표를 명확히 했더니 가고자 하는 방향 또한 명확해 졌다. 데이터 집합을 잘못 수행해 흘려 보낸 시간을 만회하고 싶었다. 잘못된 데이터로 분석했을 때의 경험을 염두에 두고 하다 보니 생각보다 빨리 분석작업이 진행됐다. 재능을 가진 조원들이 많아 다양한 방법으로 분석과 시각화를 했다.

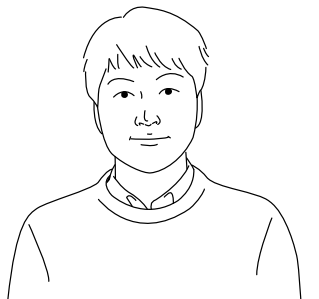
2017.12.02 힘을 합치면 문제 해결이 쉬워진다

주제를 선정하고 주차별 프로젝트를 진행할수록 유효 인자가 무엇일까? 이 거다 하고 생각되는 부분이 잡히지 않았다. 분석을 진행하면서 변수들 간의 관계를 파악할 수 있었고, 변수들 간의 상관관계 분석을 하면서 희미하게나마 유의미한 결과에 접근해 갈 수 있었다. 발표자료 준비에 들어갈 때만 해도 어떻게 마무리할 수 있을지 고민했다. 하지만 조원들간 분석 결과를 놓고 의견을 나누다 보니 놀랄 만큼의 내용이 나오기 시작했다. 갑자기 분위기가 되살아났다. 혼자서 했다면 과연 어디까지 나아갈 수 있었을까, 할 정도로 힘을 발휘하기 시작했다. 협업의 힘이 바로 이런 것이구나 하는 경험을 제대로 했다.

“경험지식이 농친 부분을 찾아낸다”

박정훈

하림 기획조정실 정보전략팀 차장



**주제를 미리 정하고 교육에 참석할
것으로 알고 있다. 그 배경이 궁금하다.**

일하고 있는 회사가 시장에서 원하는 수요와 공급 간의 갭을 줄이기 위한 일환으로 검토하던 주제 중 하나였다. 하지만 회사에서조차 도계 닭의 균일도 관리에 대한 뾰족한 접근방법이 없는 상태였다. 어떤 방법론을 적용해 이 문제를 풀어야 할지 모두 난감해하던 상황이었다. 개체 균일도를 정확하게 판단하기 위한 데이터가 모두 준비된 상황도 아니었다. 하지만 MES를 통해 모든 닭의 개별 종량정보를 관리하고 있었다. 또한 ERP(Enterprise Resource Planning)를 통해 관리되는 농장의 위치, 시설 등의 데이터를 중심으로 접근했다. 5주간 진행할 수 있는 파일럿 프로젝트의 범위를 염두에 둔 상태에서 직원들이 적극 참여하면서 나름대로의 결과를 도출할 수 있었다.

**주제와 데이터까지 확보한
상태였다는 점이 다른 조에 비해
유리한 요소로 작용하지 않았나.**

처음 주제를 정하는 과정은 분명 다른 팀에 비해 빠른 편이었다. 하지만 주제에 대해 막연하게 생각했을 뿐이었다. 접근 과정까지 생각한 것은 아니었다. 막상 진행하면서 우리가 해결하려는 문제가 당초 생각했던 것보다 많은 데이터와 유효인자들을 원한다는 것을 알게 됐다. 몇 번에 걸쳐 주제를 바꿔야 하는 거 아닌가 하고 망설였던 것도 사실이다. 결과적으로 주제를 사전에 선정한 부분은 파일럿 프로젝트를 진행하는 시간을 많이 줄여준 것은 사실이다.

**프로젝트 진행중에 어려웠던
점과 그 해결 방법은.**

누구나 비슷하겠지만 회사에 다니면서 파일럿 프로젝트를 진행해야 하는 부분이 적지 않은 부담으로 작용했다. 회사에서도 프로젝트를 진행중이었다. 파일럿 프로젝트에 많은 시간을 할애할 수 없었다. 그중 이미 알고 있는 부분이라고 생각했던 것이었는데, 현실과는 달라서 데이터 수집 시 오류로 연결되기도 했다. 결국 직원들이 함께하는 프로젝트에서 시간을 낭비하는 요인이 됐다. 비슷한 상황을 반복하지 않기 위해 직원들 간 협의해가면서 데이터 준비부터 정제까지 최종의 데이터 집합을 만들어 낼 수 있었다. 센서 데이터 분석이 아닌 이상, 양계에 대한 분석이었으므로 도메인 지식, 즉 닭의 사육·생산·가공 지식이 풍부해야 한다.

**이 프로젝트는 회사 차원에서 관심을
갖고 있을 거 같다. 발전 계획이 있다면.**
현재 회사에서 관리하는 데이터만으로는 유의미한 결과 도출에 어려움이 있다. 닭의 생육 환경에 가장 큰 요소인 계사 내부의 온도, 습도 데이터에 대한 관리가 중요하고 이 데이터를 향후 IoT 기술을 접목해 관리할 필요성을 느꼈다. 데이터 정제를 위해 현업에서 관리할 부분 역시 필요함을 느꼈다. 부족한 부분을 확인했으므로 현업 부서와 함께 이 프로젝트를 심화해볼 계획이다.

**빅데이터 아카데미 수강 전과
후에 달라진 점이라면.**

IT 업무 종사자이므로 빅데이터라는 말은 익숙하지만, 실제로 접할 기회는 별로 없었다. 책을 봤을 때는 ‘이게 뭐구나’ 할 정도로 머리에서 이해한 수준에 그쳤다면, 2주 동안의 이론교육과 5주 동안의 파일럿 프로젝트로 구성된 빅데이터 아카데미는 온몸으로 받아들였던 프로젝트였다. 직장인은 물론 누구에게나 교육은 발전의 토대가 된다. 본인에게 필요한 교육이라고 느낀다면 망설일 필요가 없지 않을까. 처음에는 할 수 있을까 하는 막연한 두려움이 들겠지만, 빅데이터 교육 과정에 과감히 도전해 성취감을 맞볼 수 있기를 바란다.

**지방에서 근무하면서 파일럿 프로젝트를
진행하기가 쉽지 않았을 거 같다.**

익산에서 근무한다. 교통 여건이 좋아지면서 그리 멀지 않게 느껴졌다. 1시간 30분 정도 소요되는 거리였지만 문제될 건 없었다. 집체교육 2주, 파일럿 프로젝트를 5주 동안 진행하면서 거리 제약으로 인한 불편함은 크게 없었다.

2017

빅데이터 아카데미 우수 프로젝트 사례

발행일 2017년 12월 22일

발행처 과학기술정보통신부
13809 경기도 과천시 관문로47, 5동
국번 없이 1335
www.msit.go.kr

한국데이터진흥원
04513 서울시 중구 세종대로 9길 42 부영빌딩 7층
02-3708-5371
www.kdata.or.kr

편집 글봄크리에이티브
07025 서울시 동작구 동작대로7길 80 3층
02-507-2340 www.mustree.com

디자인 황지원 jiwon.works@gmail.com

2017

빅데이터 아카데미 우수 프로젝트 사례



과학기술정보통신부



data 한국데이터진흥원