

Big Data, Bigger Opportunities And The Biggest Value!

2014 우수 프로젝트 사례집

Big Data
2014
마하다!



Big Data, Bigger Opportunities And The Biggest Value!

2014 우수 프로젝트 사례집

Big Data
2014년
포토박스!



Big Data, Bigger Opportunities and The Biggest Value!

미래창조과학부와 한국데이터베이스진흥원은 시장 수요에 대응한 빅데이터 전문인력 양성 및 전문직업으로 육성 지원을 위해 2013년 6월 '빅데이터 아카데미'를 국내 최초로 개소하여 올해로 2년째를 맞이하고 있습니다.

올 한 해 동안 '빅데이터 아카데미'는 빅데이터 직무를 기획·처리·분석·시각화·운영관리로 세분화하여 직무 중심의 '빅데이터 처리 기술 전문가'와 '빅데이터 예측 분석 전문가'가 될 수 있도록 교육 과정을 보완하였으며, 금융·의료·제조 등 다양한 산업분야에 종사중인 201명의 인력을 빅데이터 전문가로 양성하였습니다.

또한 연수중 실시하는 빅데이터 프로젝트에 현업 빅데이터 전문기를 팀별 멘토로 지정하는 등 프로젝트의 품질 향상과 현업 적용도 제고를 위해 노력하였습니다.

그 결과, 빅데이터 아카데미에 참여한 연수생 중 85명이 77개 기업에서 실제 빅데이터 처리 및 분석 협업 프로젝트를 추진중인 것으로 조사되었습니다.

이에 연수중 실시한 우수 프로젝트 사례들의 기술적·분석적 노하우 공유와 신규 빅데이터 비즈니스가 발굴될 수 있도록 지원하기 위해, 「2014년 빅데이터 아카데미 우수 프로젝트 사례집」을 발간합니다.

본 사례집에 실린 8건의 사례는 2014년 한 해 '빅데이터 아카데미' 연수를 통해 개발된 파일럿 프로젝트 가운데 심사를 거쳐 선정된 것들로, 과제 발굴 단계부터 개발 과정과 프로젝트 수행 시 경험했던 문제점 등 프로젝트 전반을 소개하고 있습니다.

아무쪼록 본 책자가 전국 각 현장에서 빅데이터 프로젝트를 기획하고 준비하는 많은 분에게 좋은 참고 자료가 되기를 기원합니다.

감사합니다.

2014년 12월

한국데이터베이스진흥원장 서강수

빅데이터 2014년을 말하다
2014 빅데이터 아카데미 우수 프로젝트 사례집

CONTENTS

분석 전문가 과정

신간 서적의 판매량 예측 모형

베스트셀러는 빅데이터가 먼저 알아본다!

6 5기 우수 프로젝트

주가 예측 상관관계 분석

시스템 투자로 시장을 이기다!

16 6기 우수 프로젝트

일일 환율 예측 프로젝트

누구에게나 열려 있는 빅데이터 분석의 가능성!

22 7기 우수 프로젝트

자라섬 재즈 페스티벌 관람객 분석 및 예측

데이터 분석과 재즈 페스티벌이 만났을 때

30 8기 우수 프로젝트

빅데이터 아카데미 설립 배경 80

교육 대상과 참여 안내 81

한 눈에 보는 빅데이터 아카데미 83

기술 전문가 과정

증권사 고객 패턴 분석 시스템

데이터에서 증권사 고객 마케팅의 답을 찾다

5기 우수 프로젝트 42

윈도우 서버 감사 로그 분석 시스템

빅데이터 처리 기술로 윈도우 서버를 지켜라

6기 우수 프로젝트 52

방범시설과 범죄와의 상관관계 분석 시스템

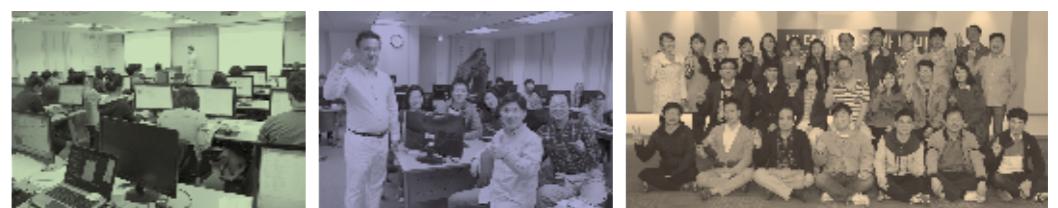
상관관계 속에서 범죄의 숨겨진 비밀을 캐다

7기 우수 프로젝트 62

자녀 교육특구 모델 찾기

워킹맘은 데이터에서 자녀 교육특구를 찾는다

8기 우수 프로젝트 72





신간 서적의 판매량 예측 모형

베스트셀러는 빅데이터가 먼저 알아본다!



08

글 KPDCH

한국의 인터넷 출판 유통 분야는 전체 시장의 50% 이상을 점유하며 어느 나라보다 활성화가 돼 있다. 하지만 인터넷 서점에서 시간 판매 예측은 담당 MD들의 경험과 직감에 의존하고 있다.

경험의 한계를 뛰어넘자! 빅데이터 분석 전문가 과정 5기 KPDCH팀이 '신간 서적의 판매량 예측 모형'을 개발하여 경험과 직감의 한계 뛰어넘기에 도전했다.

CHALLENGES

우리는 KPDCH이다

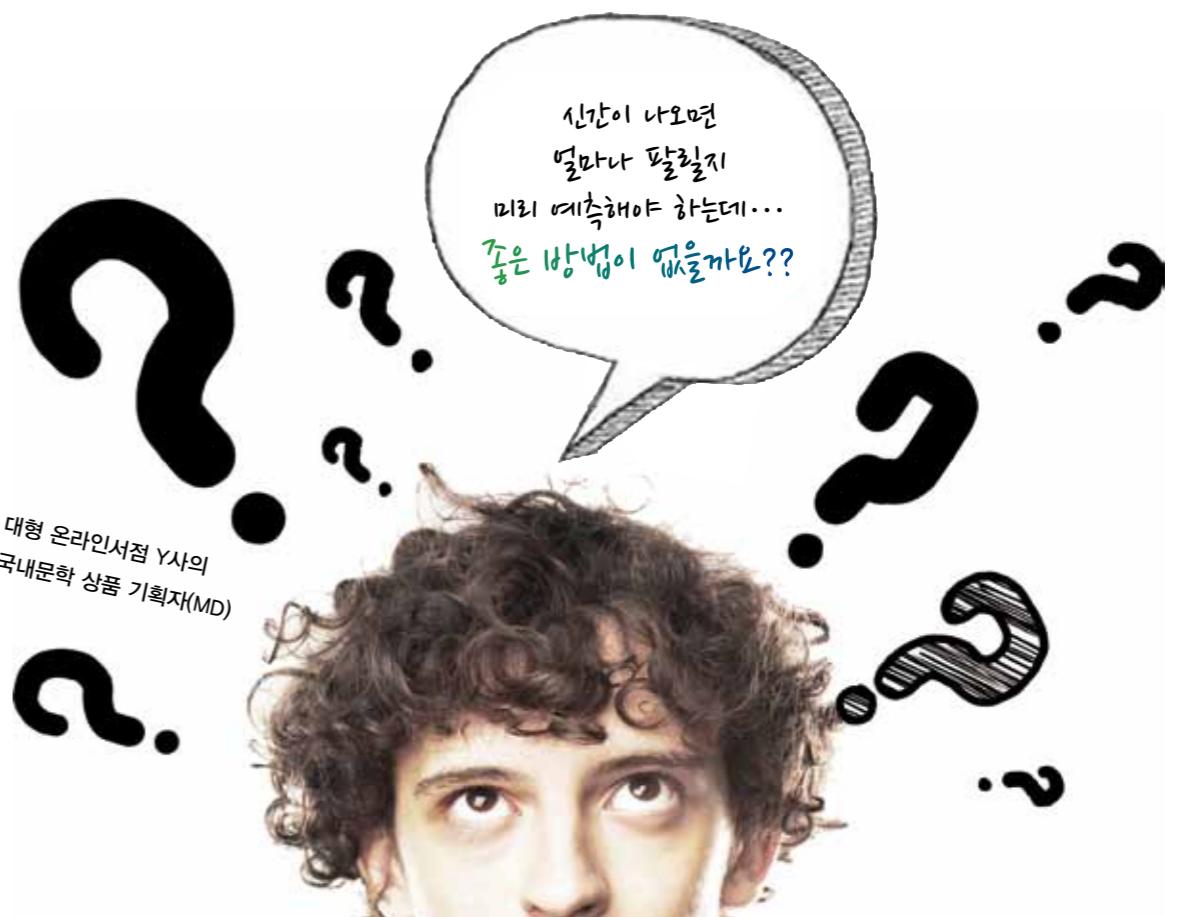
KPDCH. 이 발음하기 어려운 영어 단어는 팀 이름으로서 각 팀원들의 영어 이름 앞 글자를 따온 것이다. 팀원 모두는 향후 실력 있는 데이터 사이언스팀을 만들겠다는 비전을 담아 '(Dreaming) The Dream Team'이라는 캐치프레이즈를 내세웠다.

빅데이터 아카데미에서 함께 교육 받으며 3일쯤 지났을 때였다. 김경태 지도 교수께서 5명씩 한 팀을 꾸리고 프로젝트 주제를 미리 선정하라는 미션을 주었다. 대개의 경우가 그렇듯이 한 곳에 모여 앉아 있던 4명의 남자들은 수줍은 듯이 함께 팀을 구성하자는 이야기를 나누었고, 나머지 한 명은 누가 좋을지 망설이고 있었다. 때마침 'Peter(고민정)' 가 그 고민을 해결

해주겠다는 듯 저 멀리 대각선에 위치한 곳에서 남자들의 자리로 달려와 '함께 팀을 하시죠'라고 하여 5인 팀을 구성할 수 있었다.

팀이 구성된 이후 프로젝트 주제 선정 단계에서 소심한 갑론을박이 있었다. 각자의 관심이 다양했으나, 어느 누구도 강하게 밀어붙이지 못했다. 좋게 보면 상대를 배려해준 셈인데, 처음에만 그런 줄 알았지만 프로젝트가 끝날 때까지 그랬던 걸로 보아 다들 천성이 착한 사람들이었다.

KPDCH팀은 꽤 많은 시간을 들여 주제 선정을 위한 아이디어 회의를 가졌다. 금융회사 출신인 게빈은 공공 데이터를 활용해 대출을 받으려는 사람들의 신용을 평가하는 모형을 만들어 보길 원했고, 텍스트 마이닝 전문가인 피터는 6·4 지방 선거를 앞두고 소셜 네트워크의 감성분석을 통해 선거 결과를 예측해 보자는 의견을 냈다. 한편 온라인서점에서 개발자로 일했던 데이비드는 도서 추천 시스템이나 신간 서적의 판



09

매량을 예측해보는 것이 어떻겠냐고 했다. GIS 분야에 경험 있는 카멜은 상권분석 프로젝트 경험을 공유해 줬으며, 국 민건강관리공단에 다니는 하비는 건강검진 결과로 질병을 예측하는 것이 요즘 빅데이터의 주요 트렌드 중 하나라고 하자, KPDCH팀의 대화는 '당뇨병에 좋은 맛집 정보를 제공'하는 사업 아이템 이야기까지 빠져들어갔다.

결국 수많은 대안 중에서 데이터 수집이 가능한 프로젝트로 정하기로 의견이 모아졌다. 마침 1인 기업을 운영하고 있던 데 이비드가 과거 근무했던 온라인서점과 새로운 비즈니스를 시작하기 위해 몇 가지 아이템들을 제안했고, 온라인서점도 신 간 서적이 나올 때마다 얼마나 팔릴지 궁금했던 터라 서로의 입장이 일치했다. 이번 프로젝트가 잘 되면 데이비드의 사업에 도 작으나마 도움이 될 테고, 팀원들도 공통적으로 책에 대한 관심이 많았기 때문에 '2013년 신간 서적의 판매량 예측 모형 개발'로 주제를 결정했다.

SOLUTION

분류분석 vs 회귀분석?

프로젝트 시작 단계에서 분석 방법론으로써 분류분석과 회귀분석을 놓고 무엇으로 할지 고민했다. 사실 처음에는 분류분석에 무게를 두고 검토했다. 그 이유는 빅데이터 아카데미 교육 과정 중 주식 데이터를 갖고 간단하게 분류분석 실습을 해 본 경험도 있고, 회귀분석보다는 쉽게 할 수 있을 것이라고 판단했기 때문이다. 하지만 분류분석으로 정하기엔 몇 가지 걸리는 점이 있어 결국엔 회귀분석으로 가기로 했다.

먼저 분류분석은 종속변수가 Factor 형태여야 한다는 특징이 있다. 따라서 서적 판매량과 같은 Count data를 Factor 형태로 변환하면 몇 개의 구간으로 나눈 등급을 부여해야 한다. 이때 구간을 나누는 기준들이 다소 주관적으로 설정된다는 점이 문제였다. 또한 완성된 분류분석 모형으로 추정하면 판매량 등급(구간)을 얻을 수 있는 반면, 회귀분석은 구체적인 추정 값을 제공한다는 점에서 활용도가 더 높다는 결론을 내렸다.

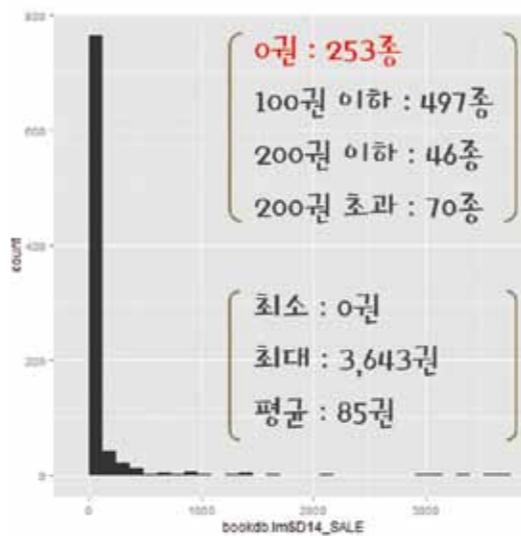
누군가 빅데이터 분석의 8할은 데이터의 수집과 정제에 있다 고 써놓은 것을 읽은 적이 있다. 이번에 한 번 해보니 절대 과한 이야기가 아니었다. KPDCH팀은 이번 프로젝트를 위해 42 개의 독립변수들을 만들었다. 결코 많다고 할 수는 없으나 KPDCH팀이 구글링과 브레인스토밍을 통해 당시 확보할 수 있었던 최대값이었고, 이 변수들을 생성하는 데에도 상당한 시간이 소요됐다.

80%의 노력은 데이터의 수집과 정제

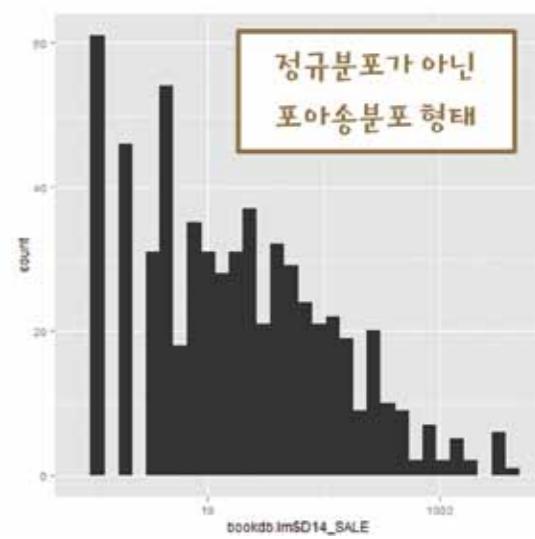
데이터 수집 과정을 설명하기에 앞서 먼저 종속변수를 정의해 보았다. 앞서 소개한 바와 같이 KPDCH팀이 알고 싶었던 것은 신간 서적의 판매량이다. 좀 더 구체적으로 정의하려면 '기간' 조건이 필요했다. KPDCH팀에게 이번 프로젝트를 의뢰한 온라인서점의 상품기획자(MD)는 '신간 서적이 판매되고 2 주 정도 지나면 이 책이 잘 팔릴 책인지 아닌지 여부를 경험으로 알 수 있다'고 하여, KPDCH팀은 '신간 서적의 출판일자로부터 14일간의 판매량'으로 종속변수를 정했다. 한편 온라인서점의 정형 데이터는 데이비드가 열심히 확보했다.

다음으로는 종속변수를 설명해줄 수 있는 독립변수들을 찾아 볼 차례다. KPDCH팀은 온라인서점으로부터 전체 서적의 저자와 출판사 정보 및 일별 판매량 자료를 제공 받아 여러 가지 파생변수들을 만들었다. 우선 저자와 출판사별로 '서적 종 수'와 '누적 판매량' 및 누적 판매량을 서적 종 수로 나눈 '평균 판매량' 변수들을 만들었다. 또 서적의 속성 중 장르와 가격도 독립변수에 포함했다. 처음에는 모든 장르의 서적을 아우르는 모형으로 만들고 싶었으나, 장르별로 판매되는 이유가 서로 달랐으므로 최종 모형에는 국내 문학 서적으로 한정해야 했다. 한편 저자와 출판사의 '누적 총 판매량'은 최신성을 반영하지 못하므로 신간 서적의 출판 월을 기준으로 최신 판매량 변수들을 새로 만들었다. 즉, 출판 전월과 3, 6, 12개월 전 판매량을 계산한 것이다.

다음은 심혈을 기울여 만든 비정형 변수들의 생성 과정을 소개하겠다. 사실 베스트셀러 작가의 책이 무명 작가의 책보다 많이 팔리는 건 당연하지만, 베스트셀러 정보가 신간 서적의 판매량에 어떻게 영향을 미치는지를 밝혀내고 싶었다. '네이버



●그림 1) 종속변수의 히스토그램. 우측 그림은 종속변수에 log10을 써운 형태



은 분기마다 통계청에서 발표하는 통계자료 중 가계 소득 및 지출(분기), 온/오프라인 서적 판매량(월), 소비자심리 지수(월) 및 실업률(월) 등을 독립변수로 추가했다. 이상과 같이 이번 프로젝트에 활용된 독립변수들은 총 42개였으나, 처음부터 데이터 분석에 활용된 것이 아니고, 분석 업무를 진행하는 도중에 필요에 따라 지속적으로 첨삭했다. 결국 프로젝트 9주 동안 데이터의 확보 및 정제 노력에 6주를 보낸 셈이다.

데이터 마트 구축

종속변수와 독립변수를 정의한 후 분석에 앞서 데이터 마트를 구축했다. 마침 팀의 데이터 모델러 카멜이 어려운 일을 금세 처리해주었다. 처음에는 오라클을 이용해 관계형 DB 형태로 만들겠다고 해 우리가 다루려는 데이터가 그 정도 대접을 받아도 되나 싶었다. 비록 프로젝트의 분석 대상인 2013년 국내 문학 신간 서적은 866권에 불과했지만, 온라인서점으로부터 받은 전체 데이터는 서적의 종류만 20만 권이 넘었고, 일별 판매량은 100만 건이 넘었으므로 데이터 모델러가 없었으면 작업이 불가능했다.

정형 데이터 (서점 DB)			비정형 데이터 (베스트셀러)			비정형 데이터 (통계청)		
독립변수	상관계수	유의확률	독립변수	상관계수	유의확률	독립변수	상관계수	유의확률
저자 출판 총 수	0.165	9.5e ⁻⁰⁷	판매지수(경영)	-0.027	0.436	가구 총소득(분기)	-0.037	0.270
저자 누적 판매량	0.433	2.2e ⁻¹⁶	판매지수(인문)	-0.018	0.587	가구 총소비(분기)	-0.034	0.452
저자 평균 판매량	0.408	2.2e ⁻¹⁶	판매지수(자/계)	0.038	0.268	도서 소비(분기)	-0.012	0.725
출판사 출판 총 수	0.080	0.019	판매지수(문학)	0.004	0.905	소비심리지수(월)	-0.014	0.680
출판사 누적 판매량	0.148	1.2e ⁻⁰⁵	저자지수(1W)	0.625	4.0e ⁻⁰⁹	경기예상지수(월)	0.004	0.898
출판사 평균 판매량	0.263	3.1e ⁻¹⁵	저자지수(2W)	0.794	2.3e ⁻¹⁴	도소매 판매액(월)	-0.012	0.732
가격(정가)	-0.021	0.536	출판사지수(3W)	0.277	2.2e ⁻¹⁶	온라인 판매액(월)	-0.032	0.342
최근 12M 판매량	0.280	2.2e ⁻¹⁶	출판사지수(4W)	0.240	7.6e ⁻¹³	실업률(월)	-0.015	0.664
:	:	:	:	:	:	:	:	:

●<표> 주요 독립변수 리스트. 종속변수와의 상관분석을 통해 얻은 유의확률을 선별 기준으로 사용

독립변수의 선별 기준, 상관분석

앞서 종속변수를 설명해줄 수 있을 것이라고 예상하고 모은 독립변수가 모두 42개라고 소개했지만, 회귀분석을 할 때 이 독립변수들을 모두 이용할 수는 없었다. 독립변수와 종속변수 간의 상관분석을 하여 선형관계가 있을 것으로 판단되는 변수들만 사용해야 했다. 유의수준을 0.05로 정했다. 종속변수인 '신간 서적의 출판일자로부터 14일간의 판매량'과 42개의 독립변수들을 차례로 상관분석을 해 유의확률이 0.05 미만인 변수들만 선별한 것이다. 실제로 이 작업을 거쳐 18개의 변수를 제외한 24개의 독립변수만 사용할 수 있었다.

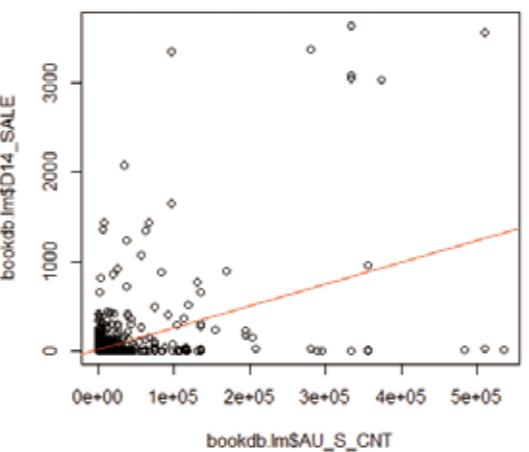
데이터 분석을 처음 해 보는 사람이라면 누구나 KPDCH팀과 비슷한 과정을 겪게 될 것이라 생각한다. 특히 R을 새로 배워 야 한다면 더더구나 그리할 것이다. KPDCH팀에게도 R은 생소했고, 회귀분석 방법론은 생각보다 방대했다. 프로젝트 해결을 위해 공부도 많이 하고 시행착오도 수없이 겪어야 했다. 최종 보고서와 함께 제출한 R 스크립트가 1000줄이 넘었다. 약간의 엄살을 더하면, 데이터 분석 과정에서 작성한 R 스크립트가 1만 줄이 넘었을 것이다.

무한 삽질의 반복, 데이터 분석

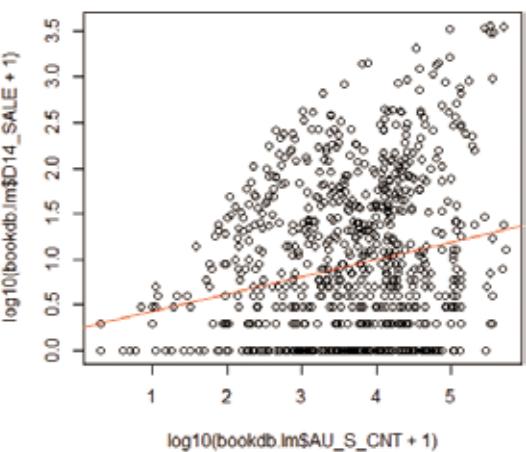
이렇게 부족한 KPDCH팀이 무사히 프로젝트를 마무리 할 수 있었던 것은 멘토와의 만남이 결정적이었다. 통계학 박사이면서 실력 있는 데이터 마이닝 전문가인 멘토께서 매번 올바른 방향을 제시해준 덕분에 시행착오 횟수를 크게 줄일 수 있었다. 첫 미팅 때, 분석 방법론으로 분류분석이 아닌 회귀분석이 더 좋겠다는 의견도 주었고, 다양한 회귀분석의 소개와 모형의 성능을 판별하는 여러 가지 기준에 대해서도 친절하게 소개해 주셨다.

특히 탐색적 데이터 분석 과정에서 큰 도움을 받았다. 처음에 종속변수와 상관관계가 큰 독립변수(저자 누적 판매량)와의 산점도를 그려보니 X-Y축에 가까운 'L' 자형 곡선으로 나타나자, 각 변수들 앞에 log10을 씌워서 다시 그려볼 것을 주문 했다. 실제로 log10을 씌워 다시 그려보니 두 변수 간에 선형 관계가 있음을 확인할 수 있었다. 이와 같은 멘토의 가이드가 없었다면 아마도 KPDCH팀의 프로젝트는 산으로 올라갔을지도 모른다. KPDCH팀은 매우 멘토와의 만남을 가졌고, 멘토가 가진 경험과 지식을 최대한 배우려 노력했다.

여기서 데이터 분석 과정을 상세하게 전달하려면 끝이 없겠지만, 전체 과정을 간단하게 요약해보면 다음과 같다. 먼저 회



●<그림 2> 종속변수와 독립변수(저자 누적 판매량)의 산점도(좌측)와 두 변수에 log10을 씌운 형태(우측)



귀 모형의 적합도를 나타내는 지표로 Adjusted R²를 사용하기로 하고, 목표수준을 0.8로 정했다. 그리고 회귀 모형 간 성능을 비교하는 추가 지표로 MAPE(Mean Absolute Percentage error)를 이용했다. 실제값과 추정값의 차이가 작을수록 MAPE 값도 작아지므로 이 값이 더 작은 회귀 모형을 선택하는 방식을 채용했다.

끝없는 도전

여기서 잠깐, KPDCH팀이 만든 첫 회귀 모형은 어느 수준이었을 것 같은가? 일단 온라인서점의 정형 데이터와 일부 통계청 데이터로 일반 선형 모형을 만들었을 때 Adjusted R²는 겨우 0.12가 나왔다.

생각보다 실망스러운 결과였지만, 좌절하지 않았다. 다음으로 각 변수 앞에 log10을 씌우니 Adjusted R²가 0.32로 3배 가까이 증가했지만, 여전히 팀의 목표와는 상당한 차이가 있었다. 당시 가지고 있는 독립변수들로는 더 이상 모형의 적합도를 상승시킬 수 없는 상황이라 판단했다.

KPDCH팀은 오랜 시간 공을 들여 만든 베스트셀러 정보를 추가하면, 회귀 모형의 적합도가 극적으로 상승할 것이라 기대 했다. 실제로 베스트셀러 정보를 독립변수로 추가해 보니 모형 적합도가 향상됐지만, 기대했던 만큼 극적으로 올라가지는

않았다. 정작 모형 적합도를 0.8 수준으로 끌어올릴 수 있었던 계기는 바로 저자 정보가 누락된 데이터들을 제외하고 회귀 모형을 만들었을 때였다. 온라인서점으로부터 받은 정형 데이터에는 저자 정보가 누락된 데이터가 전체의 85%에 달했다. 저자 정보가 있으나 저자의 속성을 설명해주는 변수들, 즉 누적 총 판매량, 베스트셀러 등록 횟수 등의 독립변수 값이 모두 0일 수밖에 없었다. 결국 저자 속성이 없는 데이터들이 전체 모형 적합도를 떨어뜨리고 있었던 것이다. 이렇게 N/A 데이터를 배제하고 회귀 모형을 만든 것은 불가피한 조치였다. 나중에 이 문제에 대해서는 KPDCH팀에게 프로젝트를 의뢰한 온라인서점에 알려 저자의 정보를 최대한 보완할 수 있도록 요청할 계획이다.

데이터 분석은 케빈이 담당했다. 회귀 모형의 적합도를 0.8로 끌어올리는 방법을 찾기 위해 여러 회귀 분석 모형에 대해 공부를 하면서 대략 10여 편의 영어논문을 읽었다고 한다. 한편 데이터 분석에 앞서 전체 데이터(866건)를 교사 데이터(Training Data, 80%)와 시험 데이터(Test Data, 20%)로 나누기 위해 난수를 생성해야 했다. 이때 케빈은 seed 번호로 212를 사용했다. 올해 2월 12일에 아들이 태어났기 때문에 212를 행운의 숫자로 여긴 셈이다.

최종 회귀 모형을 만들기까지 총 7주가 걸렸다. 일반 선형 모

형에서부터 GLM(Generalized Linear Model), Box-Cox Transformation 등 다양한 회귀분석 모형을 토대로 만들어 봤다. 결론적으로 모형 적합도가 가장 높고 MAPE가 가장 낮은 모형은 일반 선형 모형이었다. 이제 최종 회귀 모형과 해석 방법에 대해 간단하게 소개한다.

회귀 모형을 완성하다

다음은 최종 회귀 모형이다.

국내 문학 신간 서적의 출판 후 14일간 판매량

$$= 14.7 + (0.0004 \times \text{저자 누적 판매량}) \\ + (-0.02 \times \text{저자 최근 3개월 판매량}) \\ + (-47.2 \times \text{저자 베스트셀러 지수_1주}) \\ + (90.2 \times \text{저자 베스트셀러 지수_2주}) \\ + (215.7 \times \text{저자 베스트셀러 지수_3주}) \\ + (-30.3 \times \text{저자 베스트셀러 지수_4주}) \\ + (7.84 \times \text{출판사 베스트셀러 지수_4주})$$

위 수식에서 독립변수 앞에 굵은 표시된 것이 회귀계수다. 이 숫자들은 독립변수가 한 단위 증가할 때마다 종속변수에 영향을 미치는 크기로 이해할 수 있다. 총 7개의 독립변수 중 종속변수에 가장 크게 영향을 미치는 것은 바로 ‘저자 베스트셀러 지수_3주’다. 이 독립변수에 대한 해석은 이렇게 할 수 있다. 새로운 책을 출간한 저자의 기존 서적이 신간 서적 출판 3주 전에 6개 온라인서점에서 한 번이라도 베스트셀러에 등록되었으면, 신간 서적의 출판 후 14일간 판매량은 215.7권이 증가한다는 것이다. 이때 다른 모든 독립변수가 0이라고 한다면 이 회귀 모형으로 추정되는 종속변수의 값은 y절편 14.7에 215.7을 더한 230.4가 된다.

CONCLUSION

앞서 제시한 회귀 모형을 보면, 한 가지 의문이 들 것이다. ‘저자 베스트셀러 지수_1주’나 ‘저자 베스트셀러 지수_4주’ 같은 숫자가 커질수록 신간 서적의 판매량도 증가할 것 같은데, 즉 회귀 계수가 양수가 될 것 같은데 왜 음수가 되었는지… KPDCH팀은 이 현상을 ‘저자 베스트셀러 지수_3주’와 같은 독립변수의 회귀 계수가 상당히 큰 숫자로 산출된 것에 영향을 받은 것이라고 판단하고 있다. 그 결과 종속변수와 양의 상관관계에 있는 독립변수이지만 회귀 계수는 음수가 된 것이다.

KPDCH팀이 소개한 최종 회귀 모형의 결론이 별것 아님에 적잖이 실망했을지도 모르겠다. 신간 서적의 판매량을 결정하는 독립변수들이 누구나 쉽게 예상할 수 있는 저자의 속성들로 채워져 있기 때문이다. 누적 판매량이 많고, 베스트셀러에도 많이 등록된 저자는 당연히 유명할 테고, 그가 쓴 신간이 독자들의 선택을 받을 확률 또한 당연히 높다고 할 수 있다. 하지만 KPDCH팀이 이 회귀 모형에 나름의 의미를 부여하고 싶은 부분은 ‘저자 베스트셀러 지수의 주차별 회귀 계수’에 있다. 1주 전 또는 4주 전에는 음수였던 것이 2주 전과 3주 전에는 양수로 나타났다. 그것도 3주 전의 회귀 계수가 가장 크다.

이것은 신간 서적이 출판되기 직전(1주 전)에 베스트셀러 리스트에 등록된 저자보다 2~3주 전에 베스트셀러 리스트에 등록된 저자의 이름이, 마침 책을 고르는 독자의 머리 속에 보다 깊숙이 각인되어 있다고 정리할 수 있지 않을까?

이제 이 회귀 모형의 활용방안에 대해 언급할 차례다. 사실 우리가 이렇게 회귀분석을 하는 이유가 모형을 해석하고 실제 비즈니스에 활용하기 위함이기 때문에 이 부분이 가장 중요할 것이다. KPDCH팀은 다음과 같은 3가지 활용방안을 제시한다.

첫째, 매월 신간 서적들이 출판되기 전 예상 판매 순위를 예측할 수 있다. 교사 데이터(Training Data)로 회귀 모형을 만들고 시험 데이터(Test Data)로 검증했을 때, 실제 판매량이 가장 많은 1, 2위 서적을 정확하게 맞췄다.

둘째, 서점에서 프로모션 대상 서적을 선택하는 기준으로 활용할 수 있다. 회귀 모형으로 예상 판매량이 상위 10%인 서적들을 골랐을 때, 실제로 100권 이상 팔린 서적은 임의로 선택했을 때보다 무려 5.5배(Lift)나 많았다.

셋째, 출판사와 신간 서적의 이벤트 기간 등을 결정할 때 기초자료로 활용할 수 있다. 신간 서적의 출판을 앞두고 있는 저자의 기준 서적이 베스트셀러 리스트에 최근 등록되었다고 한다면, 신간 서적의 출판일자를 3주 뒤로 조정하는 것이 판매량 증대에 보다 효과적이라 판단할 수 있을 것이다.



소통이 팀워크를 단단하게 해준다

KPDCH팀은 이번 프로젝트를 제대로 해내기 위해 소통을 우선으로 했다. 소통이 제대로 기능할 때 팀 워크는 저절로 생긴다고 판단했기 때문이다. 전체 진행과정에 대해서 누구 하나 빠짐없이 잘 전달되어야 소외되는 팀원이 생기지 않고, 누가 무슨 얘기를 하더라도 전원이 바로 이해할 수 있기 때문에 속도도 빠르다. KPDCH팀은 2주간의 교육 과정이 종료되기 전에 KPDCH팀만의 소통 방법을 정했다. 페이스북에 새로운 그룹을 하나 만들어 각자 생성한 자료를 이 곳에 모두 업로드하였고, 모임 일정 등을 협의하는 메신저로는 네이버 밴드를 이용했다.

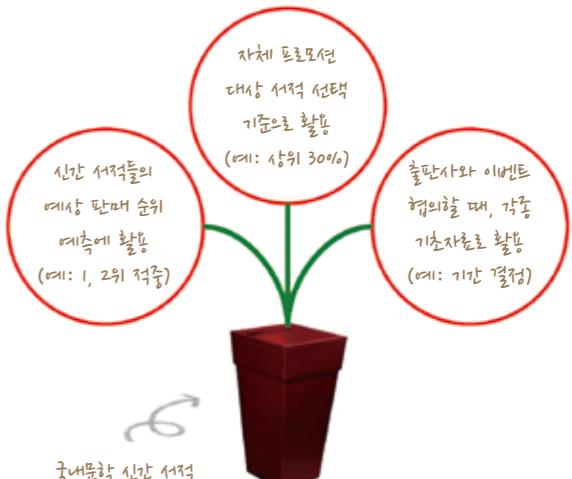
이제 막 첫 걸음을 떠다

맞다. 이제 겨우 첫 걸음을 떠었다. 하지만 빅데이터 분석 전문가로서 방향은 제대로 잡은 것 같다. 이제 속도를 내려면 실력을 키워야 할 차례다. KPDCH팀이 빅데이터 아카데미를 알기 전에는 회사에서 데이터 웨어하우스 형태로 보유하고 있는 데이터만 갖고 분석을 하려 했으나, 그나마도 제대로 활용하지 못했다. 그런데 이제 알게 되었다. 문제 해결을 위해서 얼마든지 다양한 소스로부터 데이터를 모을 수 있다는 것을…

요즘 ‘빅데이터’라는 키워드에 대해서 전문가들마다 의견이 엇갈리는 것 같다. 대개 2014년에 정점을 찍고 이제는 내려가는 추세라는 의견이 가장 많아 보였다. 그러나 ‘빅데이터’라는 트렌드가 내리막을 걷는다 해도 기업과 정부에서 ‘데이터 분석을 통해 인사이트를 발굴’ 하는 일이 없어지지는 않을 것이다. 오히려 새롭게 등장하는 분석 기법들과 데이터 확보 방법들로 더 에지 있는 모습으로 진화해 나갈 것이다. 자연스럽게 데이터 사이언티스트의 수요도 늘어나지 않을까?

그런데 데이터 사이언티스트에게는 상당한 역량이 요구된다. 전문가 수준의 수학과 통계학 실력에 해커의 손놀림, 인문학적 소양을 기초로 하는 창의력, 비즈니스에 대한 이해와 통찰력, 그리고 여러 분야와 협력 및 설득력 있는 전달력 등. 그런데 이런 역량을 모두 다 갖춘 전문가는 과연 얼마나 될까?

KPDCH팀은 ‘1명의 데이터 사이언티스트’ 대신 ‘다수로 구성된 데이터 사이언스팀’이 더 가능성 있고 현실적인 모델이라 생각한다. KPDCH팀은 또 한번의 도전으로 독자 여러분을 찾아뵙고 싶다. ●



●그림 3) 신간 예측 모델의 활용 분야

하나보다 둘이 강한 이유



나성호 팀장
하나금융경영연구소 수석연구원

'리더십을 갖췄다'는 평이 많았다.

일터나 모임에서도 나오는 것이 즐거워지면 저절로 일이 풀려 나간다는 평소 생각을 확인해 보고 싶었다. 먼저 챙기고 책임지는 모습을 바랐으므로 이번 분석 프로젝트를 진행하면서 나 스스로의 리더십을 테스트해 본 기회였다. 하나 더, 구성원 모두가 어떤 것인가 참여를 해야 하고, 또 공유해야 한다는 원칙을 세웠다. 어느 누가 일방적으로 분위기를 주도하거나 반대로 어느 누가 참여하지 않으면, 그 팀은 위험해질 수 있다고 본다.

1등을 차지한 원동력은?

팀워크가 가장 크게 작용했던 거 같다. 팀원 중에 '완벽한' 데이터 사이언티스트가 없는 상황이었으므로 참여자들의 장점을 최대한 끌어내야 했다. 그래서 나이와 경력을 떠나 수평적으로 지내는 게 좋겠다고 판단하여 이름과 직함 대신에 닉네임을 쓰기로 했다. 케번, 피터 등 닉네임을 부르니까 무척 편했다.

프로젝트 중에 어려움도 많았을 텐데.

팀원으로서 나는 데이터 모델링을 담당했다. 당초 기대했던 것과 너무나 달라서 중간에 주제를 바꿀 것인지를 놓고 심각하게 고민했다. 나중에 다른 팀과 얘기를 나눴는데 똑같은 얘기를 해서 웃었다(웃음). KPDCH팀이 확보한 데이터가

회귀분석으로 결과를 도출하기에는 쉽지 않았다. 확보했던 신간 서적 866종의 판매 데이터에는 14일 동안 한 권도 팔리지 않은 것이 전체의 30%(253종)나 되었고, 90%가 100권도 팔리지 않는 것으로 나와 있었다. 이렇게 한 쪽으로 치우친 상태의 데이터로 모델링할 수 있을까 싶었다. 하지만 이 과정에서 원가를 얻을 수 있겠다는 자신감이 있었다. 혼자가 아니었고 팀원과 지도 교수, 멘토가 있었으므로 그대로 밀고 나갔다. 불리한 데이터였지만, 거기서 인사이트를 뽑아내기 위해 다양한 회귀분석 이론들을 찾아서 적용하다 보니 R 스크립트를 1000줄 넘게 작성하게 됐고, 나 스스로도 정말 자신감을 많이 얻었다.

공부를 많이 했다고 들었다.

모델링을 하면서 통계분석 영어 논문을 10편 넘게 읽었다. 경영학 석사 과정에서 통계를 잠깐 접했을 뿐인데 다른 팀의 팀장이 찾아와 “어찌 (통계를) 그리 잘 아느냐?”고 물어 보더라 (웃음). 통계학을 공부하지 않았다면 누구나 비슷한 수준일 텐데, 본격적으로 데이터 분석을 하려면 어떤 형태로든 통계학은 넘어야 할 산임을 실감했다.

팀원들의 만족도도 높아 보였다.

빅데이터 아카데미가 아니라면 만날 수 없었던 사람들을 만났기 때문이지 않을까. 팀원, 지도교수 모두 훌륭했지만, KPDCH팀에서 가장 아쉬웠던 부분인 통계지식을 멘토였던 베가스 김도현 대표로부터 배워가면서 하나씩 풀어나갈 때 정말 즐겁고 벅찼다. KPDCH팀이 적극적으로 나오자 베가스에서 2명의 통계학 박사까지 추가 지원하는 등 열정적인 프로젝트가 이뤄졌다.

향후 계획은?

금융 CRM과 빅데이터 분석에 관심을 갖고 있다. 직장 생활을 시작한 지 13년차여서 새로운 지식을 채우기 위해 빅데이터 분석에 더 관심을 갖고 공부하고 있다. 앞으로는 데이터가 가진 스토리를 발견해 내는 일들을 가능한 많이 해보고 싶다. 필요하면 박사과정에 진학해 부족한 이론 공부도 더 해보고 싶다. ☺

신간 서적 출판 후
14일간 판매량 예측 분석

프로젝트 소개

866종의 국내 문학 신간 서적의 출간 후 14일까지의 판매 데이터를 회귀분석 기법으로 분석하여, 그 동안 서점 MD의 경험과 직감에 의존해 예측하던 서적 판매량을 더 정확하게 예측해서 인터넷 서점의 재고 부담 해소 등 효율화 도모

구분

분석 전문가 과정 : 다변량 회귀분석

프로젝트 기간

2014년 04~05월

멘토

김도현(베가스 대표)

적용도구

R, MS 엑셀

수집 데이터

- 정형 데이터 온라인서점 Y사의 국내 문학 신간 서적의 저자 및 출판사 정보 및 일별 판매량 데이터
- 비정형 데이터 '네이버 책' 사이트의 주간 베스트셀러 리스트를 크롤링하여 다양한 파생변수 생성

산출물

국내 문학 신간의 출판 후 14일 판매량 예측 모델

교육 참여형태

자발적 참여(5) / 회사 권유(0)

진행

- | | | | |
|----------|--------|-------------|--------|
| • 나성호 팀장 | 모델링 | 금융경영연구소 연구원 | 경력 13년 |
| • 고민경 팀원 | 텍스트マイ닝 | SI 업체 개발자 | 경력 10년 |
| • 박대건 | 데이터마트 | 벤처기업 대표 | 경력 12년 |
| • 최태웅 | 데이터마트 | SI 업체 개발자 | 경력 02년 |
| • 유성용 | 텍스트マイ닝 | 공기업 전산실 | 경력 02년 |

빅데이터
아카데미
수강 후 변화

- | | |
|------|---|
| 수강 전 | <ul style="list-style-type: none"> • 데이터 분석은 힘들다. • 데이터 마이닝은 새로운 발견을 도와주는 즐거운 놀이였다. • 정형 데이터 분석에 대해 경험해본 적이 없었는데 이번을 계기로 더 공부할 수 있게 됐다. |
| 수강 후 | <ul style="list-style-type: none"> • 데이터 분석은 여전히 힘들다. 하지만 팀워크를 통해 극복할 수 있다. • 학술적인 접근을 통해 데이터 마이닝에 대한 즐거움이 배가 되었고, 좀 더 진지해졌다. • 인생의 꿈이 선명해졌고, 집중적이고 체계적인 분석 공부를 할 수 있게 됐다. |





시스템 투자로 시장을 이기다

주가 예측 상관관계 분석



글 권오성 프리랜서

종목별로 측정 방법이 다른 호응도와 관심도 분석 대신 정량적 분석으로 주가 예측 분석을 했다. SVM(Supported Vector Machine)이라는 판별·회귀분석을 위한 학습모형으로 대용량 데이터에서 빠르고 쉽게 주가를 예측할 수 있는 시스템을 개발했다. TTR 패키지를 이용해 다양한 기술적 분석 파생변수를 생성해 현재부터 n일 간의 주가 변동을 예측하고, 투자 시뮬레이션 결과를 시각화할 수 있었다.

ALLEGES

메인 지식이 있는 주제를 선정

제 교육이 끝나갈 때 즈음, 팀원들이 모여 프로젝트 주제를 정하는 시간을 가졌다. 이때 주제를 데이터 확보와 가공 쉽고, 많은 참조 프로그램이 있으며, 그 결과가 비즈니스 이어질 수 있는 것을 기준으로 선정했다. 이 기준에서 보을 때 주가 예측이 가장 적절하다고 판단했다. 팀원 가운 몇몇은 이미 주식 투자 경험이 있어서 관련 지식을 활용할 있었다.

로젝트 초기 단계에서는 SNS와 뉴스 텍스트 마이닝을 통해 해당 기업, 브랜드, 상품에 대한 호응도와 관심도를 측정해 했다. 하지만 다음과 같은 몇 가지 이유 때문에 호응도와 관심도를 객관적으로 측정할 수 있는 프로그램을 2개월간 개발하기 어렵다고 판단했다.

내, 기업·상품·브랜드 각각에 대한 관심도를 측정하기 위해
연관어와 호응도를 판정하기 위한 궁부정어 정의가 종목마
달랐다. 둘째, 호응도와 관심도 측정 시스템을 구축하더라
소비재를 공급하는 기업의 종목이 아니라면 별 소용이 없었
즉 중공업 주 등 산업체에 대한 반응 측정은 쉽지 않았다.
내, 소비재 상품을 공급하는 기업에 대한 호응도는 이미 종
추세에 반영돼 주가수익비율(PER)로서 가치 측정이 된 경
우가 많다는 것이다 이를 도식화하면 <그림 1>과 같다

용도와 관심도는 가치나 성장성을 측정하는 기준이라 생각
수 있다. 하지만 이는 종목별로 측정 방법이 다르므로 일반
하기 어렵고 데이터를 통한 절량적 분석보다는 절성적 분석

기술적 분석 투자(패턴분석을 통한 투자)

그메터 투자(데마즈처럼 출률을 타는 투자)

기본적 분석 투자

第四章 项目管理与组织行为学

4.1.2.2.4. 허가증(면허증) 및 허가증(면허증)

[View Details](#) | [Edit](#) | [Delete](#)

그림 1) 투자자 유형별 구분

이 더 많은 시간을 들여야 하는 문제가 있었다.

이에 우리팀은 수료 프로젝트를 제한된 기간에 완료해야 했으므로 정성적 분석을 제외하고 정량적 분석을 통한 기술적 분석 투자방법을 주제로 선정했다.

SOLUTION

전략적 분석으로 좁혀서 접근

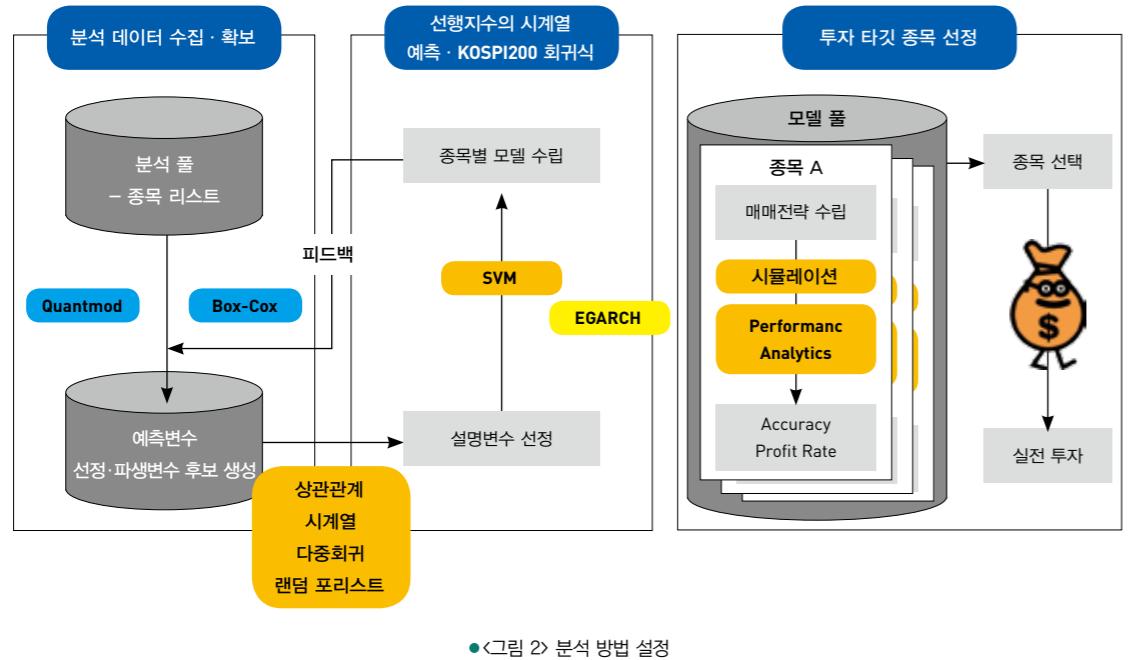
저 분석할 종목과 주가에 영향을 미치는 주가선행변수를 선언했다. 주가선행변수로서 GDP와 소비자물가지수 등 일시적으로 측정되는 지수 등은 제외했다. 대신 매일 그 변동치를 확인할 수 있는 데이터로 한정해 환율·유가·반도체지수 등을 주가선행변수로 선정하고, 종목은 KOSPI 대형주로 한정했다. 데이터 수집을 위해 처음에는 한국증권거래소(KRX)에서 엑셀로 데이터를 모았으나, 주가선행변수의 데이터를 얻을 수 없다는 문제를 발견했다. 이에 ‘fimport’라는 R 패키지로 애후 파이낸스에서 데이터를 직접 가져왔다.

기에 선정된 데이터는 OHLC(open-high-low-close) 모형에 따른 그날그날의 각 종목에 대한 시가·고가·저가·종가 데이터로 구성해 분석했다. 하지만 주가는 매수자와 매도 사이의 역학관계에서 결정되고, 이 힘을 좌우하는 주체는 개인보다는 기관 투자가와 외국인 투자가라는 점을 염두에 뒀다. 이에 따라 XML 패키지를 이용해 '다음 증권'의 데이터를 살피고, 기관투자가와 외국인 투자가의 수급량을 추가했다.

거래량 기준으로 모멘텀 변수 이용

수집한 원시 데이터로부터 40여 개의 파생변수를 생성했다.
각 파생변수는 OHLC 변수를 활용하는 TTR이라는 R 패키지
의 기술적 변수와, 기관투자가와 외국인 투자가의 거래량을
사용해 모멘텀 변수를 이용했다.

예측변수는 투자 여부를 결정할 수 있는 종속변수로서 미래 n 일 간의 주가 변동량의 합을 정의하고, 이 주가 변동량이 일정 (예 0.025 등) 이상일 때를 매입보유(long position)로 선정



일자	외국인		기관		증가	전일비	등락률
	보유주식수	지분율	순매수량	순매수량			
14.10.17	4,227,406	5.52%	+53,024	0	1,255	▲55	4.58%
14.10.16	4,174,382	5.45%	-2,694	0	1,200	▼60	-4.76%
14.10.15	4,177,076	5.45%	+26,831	0	1,260	▼40	-3.08%
14.10.14	4,150,245	5.42%	-87,097	0	1,300	▼20	-1.52%
14.10.13	4,237,342	5.53%	+49,909	0	1,320	▼10	-0.75%

●<표 1> '다음 증권' 데이터로 기관 투자가와 외국인 투자가의 수급량 추가

하는 매수 신호로 설정했다.

분석 절차

<그림 3>과 같은 방법으로 분석 절차를 수립하고, 모델 평가를 통한 피드백으로 모델을 구체화·정형화했다. 40여 개의 파생변수 중에서 실제로 모델에 적용할 설명변수를 선정하는 데 있어서는 종목별로 상이했다. 이때 랜덤 포리스트를 이용해 발굴한 중요변수로 종목별 설명변수를 선정했다. 기간설정 방법에 따라 중요도 값이 달라져, 변수에 대한 변

별력이 떨어졌다. 이는 결국 과거의 패턴에 기반해 가까운 미래를 예측할 때, 오래된 패턴보다 최근의 패턴이 더 중요할 것이라는 판단에 따라 최근의 트렌드를 따르는 데이터만 학습 데이터로 선정했다. 최근의 트렌드를 따르는 기간을 산정하는 방법으로는 차트를 보고 직감으로 추출했다. 대체로 1년 이내의 기간을 학습 데이터로 활용했다.

분석 모델의 구성

분석모델을 선정할 때 여러 논문과 증권사 보고서를 활용했

다. 이 가운데 SVM(Support Vector Machine)과 eGarch가 가장 우수한 것으로 판단했다. 그 중에서도 SVM 모델이 더 간단해서 이를 활용했다.

SVM은 대용량 데이터와 많은 설명변수에서도 빠르고 쉽게 적용할 수 있는 장점이 있다. 또한 SELL/BUY 시그널 두 개를 찾는 데 있어 판별할 수 있는 Hyperplane 튜닝도 비교적 어렵지 않게 할 수 있었다.

CONCLUSION

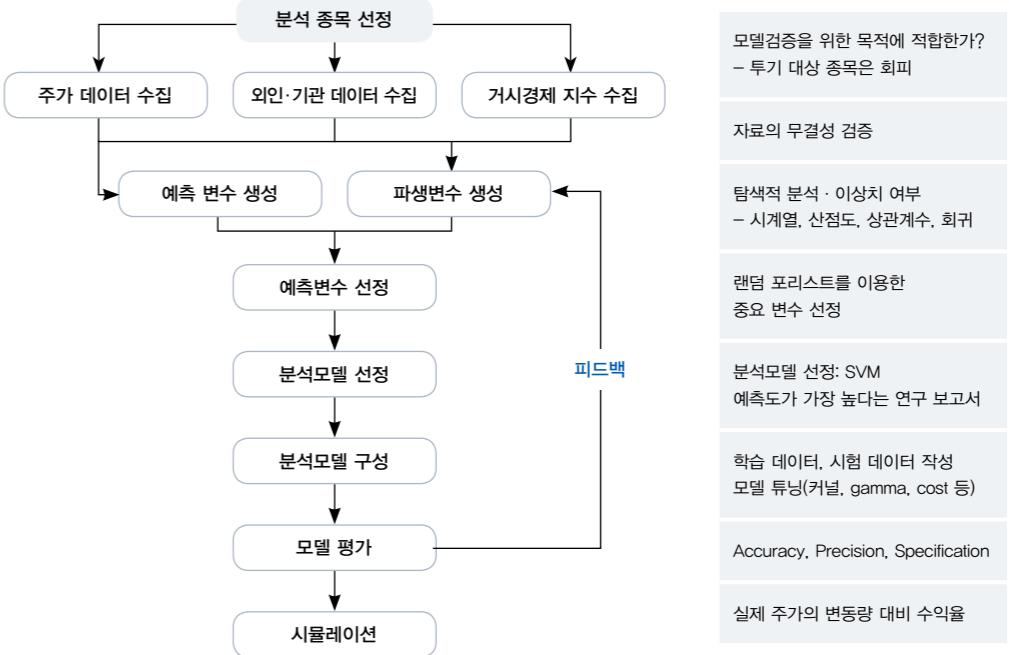
시뮬레이션을 통해 실제 종가의 매수에서 1주일의 등락을 예측해 거래했을 때, 삼성전자는 연초 대비 6월 24일 기준의 주가는 약 2.4%로 상승했으나 수익률은 11%로 나왔다. 주가가

1/3로 떨어진 현대증공업은 2.15% 손실을 기록했다. 하지만 아모레퍼시픽처럼 주가가 크게 오른 종목(54% 상승)에서 수익은 49%밖에 되지 않은 경우도 발생했다.

중요한 것은 손실 발생 종목에서도 그 손실이 크지 않았고, 전체적으로 상승 종목에서 상승폭 이상으로 수익을 낼 수 있다는 점이다. 특히 2014년 6월 24일 기준으로 KT 주가는 현재 연초 대비 -0.8%로 하락했으나, 모델을 통한 시뮬레이션 거래 결과 오히려 7.8%의 수익률을 거두었다.

따라서 본 주가예측 모델의 결과로 급등 종목에 대한 투자보다는 안정적인 종목에 투자했을 때 더 좋은 결과를 기대할 수 있다.

최대 수익을 목표로 주식을 투자하는 것보다 안정적인 종목에 투자하는 것이 유리하다는 판단과 더불어 우리팀의 주가 예측 모델은 안정적인 투자에 적합한 예측 시스템이라는 결론을 내렸다. ☺



- 모델검증을 위한 목적에 적합한가?
 - 투기 대상 종목은 회피
- 자료의 무결성 검증
- 탐색적 분석 · 이상치 여부
 - 시계열, 산점도, 상관계수, 회귀
- 랜덤 포리스트를 이용한 중요 변수 선정
- 분석모델 선정: SVM
예측도가 가장 높다는 연구 보고서
- 학습 데이터, 시험 데이터 작성
모델 튜닝(커널, gamma, cost 등)
- Accuracy, Precision, Specification
- 실제 주가의 변동량 대비 수익률

“분석, 그 깊은 맛을 내기 위한 조건은 숙성기간”



권오성 팀장
프리랜서

분석 전문가 과정에 등록한 배경은.

IT 전문가로서 평소 관심을 갖고 있던 주식투자를 요즘 유행하는 데이터 분석기법을 적용해 제대로 한번 해보고 싶었다. 팀원 중에는 실제 증권사 정보 시스템 구축 프로젝트에 참여했던 경험을 가진 사람도 있었다. 주식 투자를 실제로 해 보지 않았던 사람은 매우 낯설어 하고 심지어 사행성산업처럼 생각하는 경우도 있었다. 하지만 내가 잘 골라서 투자한 회사가 발전하면, 평균 이상의 수익을 올릴 수 있으므로 (투자 기업과) 동업 관계라고 볼 수 있다.

프로젝트를 우수하게 수료했는데 소감은.

집체교육 중에 배운 내용은 그 다음날이면 잘 기억나지가 않았다. ‘이렇게 하여 결과가 나올까?’ 싶었다. 하지만 수료 프로젝트를 하면서, 강의 시간에 배웠던 것이 헛되지 않았음을 알게 됐다. 직접 부딪치다 보니, 강의 때 들었던 내용이 ‘바로 이것이구나!’ 하고 떠오르면서 자신감도 불고 재미도 있었다. 분석을 하다 보면, 어느 순간 예측률이 놀랍도록 올라갈 때가 있다. 이때는 시간 가는 줄 모르게 된다. 반면 가설이 잘못돼 예측률이 올라갔음을 발견하는 순간 그 기쁨은 좌절로 뒤바뀐다. 이런 스릴 넘치는 과정의 연속이었다.

‘주가 예측 분석’은 자주 시도되는 주제인 만큼 차별화하지 못하면 위험했을 텐데.

흔한 주제라는 것은 그만큼 관심이 높은 주제라고 볼 수 있지 않겠는가? 그럼에도 주식의 ‘주’ 자도 모르는 팀원이 있었다. 그는 이번 프로젝트를 계기로 주식에 눈을 뜨고 직접 주식 투자를 하기 시작했다. 이미 주식에 관심이 있던 팀원들은 자신의 직관이나 투자 습관이 옳은지를 실제 데이터를 기반해 확인해 보는 재미있는 과정이었다.

기억에 남는 힘들었던 순간은.

모델을 설정할 때 어떤 것으로 해야 할지를 몰라서 막막했다. 시행 착오와 멘토의 지원을 받으면서 자연스럽게 알게 됐다. 주가 예측 분석 경험이 많은 멘토의 한마디의 조언으로 어떤 예측 모델을 우선 적용해 볼 것인지를 쉽게 결정할 수 있었다. 축박한 프로젝트 기간 중에 어떻게 해야 할지를 모르면, 난감하고 여러 생각이 교차한다. 금방 요리한 것은 양념 맛만 나지만, 숙성 기간을 거친 음식은 고유의 깊은 맛이 난다. 처음에는 양념만 치려고 했는데, 시행착오를 거치면서 하나씩 틀이 잡혔다. 아직 많이 부족하지만, 이런 과정을 통해 맛있는 데이터 분석이 되는 구나 하고 느꼈다.

분석 프로젝트 진행 후 달라진 점이 있다면.

주변 IT 엔지니어들 가운데 빅데이터를 버즈워드라고 생각하는 사람이 꽤 있다. 엔지니어들은 ‘하둡이나 NoSQL이 일반 데이터베이스, 즉 RDB 기술과 무슨 차이가 있느냐, 결국은 같은 것이 아니냐?’ 하고 말하기도 한다. 데이터 분석은 분석 플랫폼과는 분명히 다르다. 분석을 먼저 배우고 대용량 데이터 처리 플랫폼을 나중에 알아도 된다고 생각한다.

향후 계획은.

다양한 분석 모델을 만들어 실제 수익 창출로 연결해 보고 싶다. 수료 프로젝트를 완료하고 나서도 일주일에 한 번씩 전 팀원들이 모여서 개선 작업을 하고 있다. 수익 모델로 갈 수 있을 단계에 이르면, 외부에 서비스 모델로 공개해 볼 계획이다. ☺

주가 예측 상관관계 분석

프로젝트 소개

SVM(Supported Vector Machine)이라는 판별 · 회귀분석을 위한 학습모형과 범주를 판별하는 hyperplane으로 대용량 데이터를 분석해 주가를 예측하는 프로젝트. TTR 패키지를 이용해 다양한 기술적 분석 파생변수를 생성해 현재부터 n일간의 주가 변동을 예측하고 투자 시뮬레이션 결과를 시각화하는 시스템을 구현했다.

구분 분석 전문가 과정: 상관관계 분석

프로젝트 기간

2014년 05~06월

멘토 안정국(The ECG 상무)

작용도구 R, Rstudio, MS 엑셀

수집 데이터 KRX OHLC 모델의 주가 데이터, 한국은행의 환율, Yahoo Finance의 주가 데이터, 환율 · 유가 · 반도체 데이터, ‘다음 증권’의 외인 기관의 수급 데이터

신출물

- 개발 주식 종목의 n일 변동성 예측 모델
- 2014년 예측 결과

교육 참여형태 자발적 참여(3) / 회사 권유(1)

진행	• 권오성 팀장 • 박상현 팀원 • 곽재원 • 신은비	프리랜서 엔지니어링업체 연구원 SI 업체 과장 DB 컨설팅업체 과장	경력 15년 경력 15년 경력 7년 경력 5년
----	--	--	------------------------------------

빅데이터 아카데미 수강 후 변화

수강 전

- 데이터를 일상적인 업무의 일부로서 바라 봄
- 업무의 목적이 아닌 수단으로만 바라 봄

수강 후

- 데이터를 분석해 직접적인 수익 창출로 연결 가능
- 전문가의 자문을 통해 다양한 분야에서 데이터를 활용한 업무의 능력을 향상 가능
- 스스로 분석할 수 있다는 자신감 확보





누구에게나 열려 있는 빅데이터 분석의 가능성!



글 최선애 비스텔 부장

환율에 대해 잘 모르는 사람들이 모여 환율에 영향을 주는 원인들(환율에 선행되는 지수)을 함께 찾기 위해 맨 처음 한 일은 논문이나 자료를 찾아 보는 것이었다. 각자 찾은 자료를 공유하고 공부하는 과정을 거치면서 처음에는 잘 이해되지 않았던 내용들이 조금씩 눈에 들어오기 시작했다. 그 자료에 나온 내용을 기본으로 데이터를 모을 수 있었다. 짧은 기간이었지만, 이 프로젝트를 통해 특정 분야의 분석에 필요한 해당 도메인 지식을 쌓아 해당 분야의 데이터를 읽는 과정을 실제로 경험했다.

22

CHALLENGES

프로젝트 주제를 정할 때, 다양하고 흥미로운 제안이 많이 나왔다. 이 가운데 '이별 예측 커플', '연령·가족 구성원 등에 따른 최적의 이사장소' 등은 처음에는 꽤 가능성이 있어 보였다.

학위 논문을 써 본 사람이라면, 경험이 없는 상태에서 논문 주제 선정이 얼마나 위험한지를 알 것이다. 빅데이터 분석 프로젝트도 다를 게 없었다.

우선 '이별 예측 커플' 주제는 참신하고 재미있어 보였다. SNS에서 오가는 글을 텍스트 마이닝해 이별할 커플 사이에 어떤 말이 오갔는지를 찾아보고, 이별 전에 자주 쓰는 말을 주고 받는 커플은 이별 가능성이 높다고 가정했다. 하지만 이 주제가 얼마나 무모한 것이었는지는 얼마 가지 않아 드러났다. 데이터를 구할 곳이 없었던 것이다. 이별한 사람들은 SNS

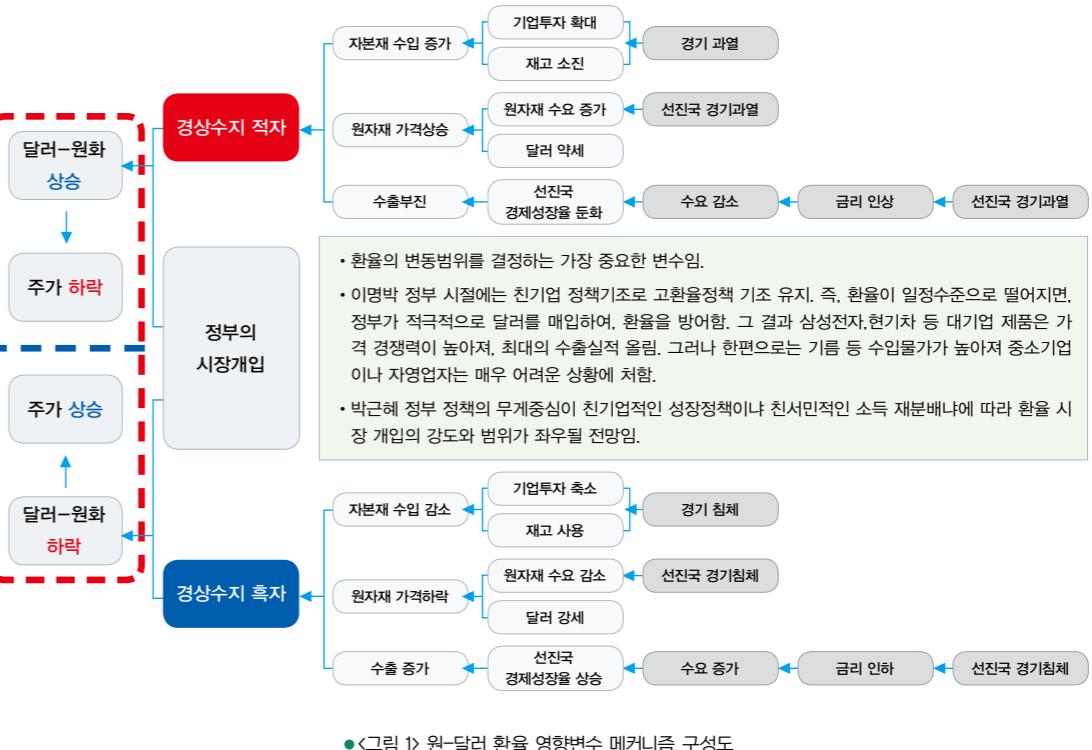
를 탈퇴하거나 잠수를 탔고, 무엇보다 연인과의 상태를 공개적으로 언급하는 사람이 별로 없었다.

데이터를 어디에서 어떻게?

역시 문제는 데이터였다. 새로운 주제를 떠올릴 때마다 '그럼 데이터를 어디에서, 어떻게 구하지?'라는 벽에 부딪혔다. 주제가 참신하고 재미있어도 결국 데이터를 확보할 수 있어야 뭐든 해볼 수 있음을 알게 됐다. 부동산 관련 주제나 여타 주제도 데이터를 구할 방법을 찾지 못해 포기하기에 이르렀다.

그럼 원-달러 환율 예측은 어떨까? '너무 어렵지 않을까?' 그 래도 성공하면 의미가 있고 가치도 높지 않을까?' 라며 다들 고민하던 찰나, 지도 강사의 '원-달러 환율 예측… 무난하게 갈 수 있을 듯 합니다'라는 한마디에 희망을 얻어 촉오톤은 환율에 대해 공부하기 시작했다.

이 분야에 관심을 갖고 있던 윤상택 팀원을 중심으로 환율 관련 논문과 자료를 찾아 나섰다. 주가 예측을 선정했던 조가



23

많았다. 환율 예측은 그동안 빅데이터 아카데미의 수료 프로젝트 주제로 다뤘던 적이 없는 새로운 프로젝트라 촉오팀에게는 ‘한번 해보자’는 의지를 다지게 했다.

SOLUTION

논문을 보면 <그림 1>처럼 많은 변수 중에서도 몇 가지 공통적으로 나오는 게 있다. 즉 금리 · S&P500지수 · 엔화 환율 · 기초 경제 변수 · 경상수지 등이 바로 그것이다. 그 데이터가 정확히 무슨 뜻인지는 몰랐지만, 일단 찾을 수 있는 데이터를 각자 찾아 보기로 했다. 엔-달러 · 원-달러 · 원-엔화 등 환율, 금리, 금값, KOSPI 지수 등을 찾기로 했다.

<그림 1>은 오른쪽 위편이 원인이 되어 원쪽의 원-달러 상승 또는 원-달러 하락을 불러온다고 볼 수 있다. 예를 들어 선진국에서 경기가 과열되면 금리가 올라간다. 이는 수요 감소로 이어져 선진국들의 경제 성장을 둔화로 연결된다. 이로

인해 수출이 영향을 받게 되어 경상수지가 악화되면, 원-달러의 환율이 올라가고 주가를 떨어뜨린다. 반대로 경기 침체는 기업 투자를 위축시켜 자본재의 수입 감소로 이어져 경상 수지 흑자로 연결된다. 결국 원-달러 환율이 하락하고 (원화 가치가 올라가므로) 주가는 상승한다. 이런 메커니즘 가운데 정부의 시장 개입의 강도와 범위에 따라서 조정되기도 한다.

데이터 구하기

일단 어떤 데이터를 확보해야 할지를 정했다. 분석할 데이터를 정하고 나니, 데이터를 가져오는 방법은 무척 다양했다. 촉오팀원들은 각자 나누어 데이터를 확보하기로 했다. 가장 먼저 확보할 데이터는 5년 이내로 지정했다. 이유는 2007년도에 서브프라임 모기지론 사태가 발생하면서 환율의 유동 폭이 너무 커졌기 때문이다. 이런 특이한 사태 발생으로 인해 정부가 정책적으로 환율을 관리하던 당시의 데이터를 사용하면, 예측률이 정확하지 않을 수 있었다. 데이터를 가져온 방법은 처음에는 복사/붙여쓰기로 관련 웹사이트의 자료를 가져왔다. 통

계청 사이트에서 일부 파일 데이터를 내려 받기도 했다. 프로젝트 사이사이 멘토의 도움으로 R 스크립트로 웹사이트의 자료를 읽어오고, QUANDL로 지정한 기간의 데이터만 골라 가져오는 방법도 적용했다.

하지만 이 중 어느 하나만을 사용할 수는 없었고 모든 방법을 동원해야만 했다. 우리가 원하는 데이터가 어느 한 곳에, 한 포맷으로 존재하지 않기 때문이다.

데이터 만들기 1, 합치기

열심히 모은 데이터를 한 곳에 모을 때, 촉오팀은 다시 벽에 부딪혔다. 데이터의 기준 날짜가 서로 달랐기 때문이다. 경상수지 · 소비자 물가지수는 월별 데이터이고, GPD 성장률은 분기별 데이터인 것처럼 날짜 기준이 제각각이었다. 이렇듯 일별 · 주별 · 월별 · 분기별로 다양한 데이터들을 어떻게 합쳐야 할지가 고민이었다.

결론부터 말하면, 일별 데이터는 통계에서 흔히 이용하는 평균값을 이용해 주별 데이터 또는 월별 데이터로 만들 수 있다. 하지만 주별 데이터를 일별로 이용하면, 데이터가 가진 성격을 흐트러뜨릴 수 있다. 이는 정확하지 않은 분석 결과로 이어진다는 사실을 멘토로부터 들었다. 그래서 촉오팀은 프로젝트 초기에 결정한 주간 예측에서 일간 예측으로 바꿨다. 일별 데이터가 주별 데이터보다 더 풍부한 것을 보여 준다는 생각에 과감히 일별 예측을 하기로 한 것이다.

변수 선택은 다음과 같은 과정으로 했다.

- ❶ 환율에 영향을 주는 변수를 선정해 인터넷에서 수집
- ❷ 항목별로 Daily/weekly/monthly/quarterly/half-yearly 별 다양한 데이터가 존재
- ❸ 최종적으로 주간 예측에서 일일 예측으로 전환. 일일이 아닌 지표들은 삭제

데이터 만들기 2, 유실 데이터 처리

일별 데이터들만 확인해 보니, 또 다른 문제가 기다리고 있었다. 해외 데이터의 날짜와 우리나라 데이터의 날짜가 서로 일치하지 않았다. 이유는 휴일이 다르기 때문이다. 맨 처음에는 모든 데이터에 존재하는 날짜를 기준으로 가져와 합쳤는데,

환율 데이터가 유실(missing)됨에 따른 새로운 문젯거리가 될 수 있었다. 예를 들어 원-달러 환율은 2014-02-02 데이터가 존재하는데, S&P500 지수는 해당 데이터가 없었다. 이 문제 해결을 위해 촉오팀이 첫 번째로 시도했던 방법은 공통 날짜의 데이터만 가져오는 것이었다. 이렇게 하면, 앞의 2014-02-02일의 원-달러 환율 데이터는 사라진다. 그래서 원-달러 환율의 날짜를 기준으로 다른 변수들의 데이터를 합쳤다. 그럼 2014-02-02의 원-달러 환율은 존재하는데, S&P500 지수는 NA로 존재하지 않게 된다. 이런 데이터들은 이전 날짜의 데이터로 채웠다.

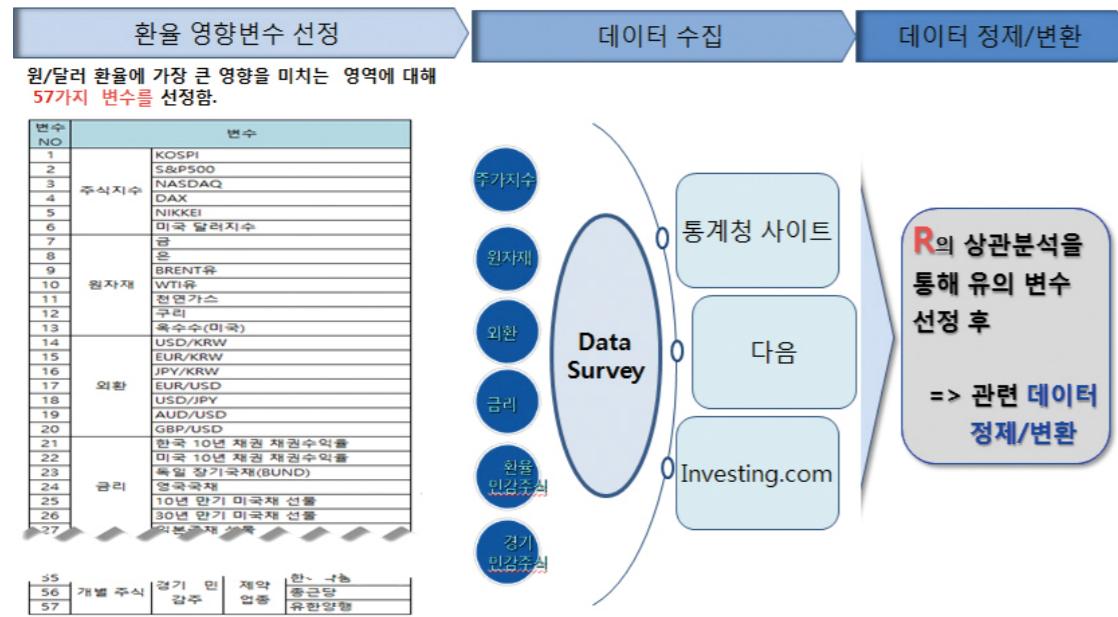
일단 25개 변수로 이루어진 데이터세트가 완성됐고, 이 데이터들을 어떤 모형으로 돌릴까 고민을 해 보았다. 하지만 환율은 시계열(time series) 데이터라고 하기엔 무리가 있었다. 나라별 경제 상황 · 주가 등 사람의 활동이나 경제 활동에 따라 변경되는 지표지수, 봄 · 여름 · 가을 · 겨울 등 어떤 계절적 요인(seasonality)에 영향을 받는 변수가 아니기 때문이다. 그래서 결정된 방법이 회귀분석(regression)이었다.

회귀분석을 하기에 앞서 우리팀은 상관분석을 해 보았다.

* USD_KRW 와 correlation 분석 결과

1. KOSPI_open -0.7899782
2. KOSPI_close -0.7855336
3. DWJ_open -0.7700833
4. DWJ_close -0.7677104
5. S&P500_open -0.7624288

●<그림 3> 상관관계



●<그림 2> 데이터 수집 과정

<그림 3>에서도 보듯이, KOSPI · DOW 지수 · S&P500 지수가 원-달러 환율에 영향을 많이 주는 것으로 보였다. 첫 번째 회귀분석 결과는 다음 <그림 4>와 같다. 가장 높은 상관 관계를 보인 것은 S&P500 High, Kospi open, Nasdaq open 지수로, 가장 큰 설명변수로 선정되었다. 그러나 R-squared 값은 0.747로 나타났다. 위 설명 변수

```

lm(formula = USD_KRW ~ ., data = subtdw)
Residuals:
    Min      1Q  Median      3Q     Max
-96.191 -17.237   2.621  18.186  71.227
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1582.93161 12.06050 131.249 < 2e-16 ***
snp_open    0.12704  0.33079  0.384  0.7010
snp_high   -0.10367  0.17881 -5.981  9.61e-09 ***
snp_close   0.56010  0.33244  1.685  0.0923 .
ksp_open   -0.24325  0.05434 -4.477  8.35e-06 ***
ksp_close   0.05670  0.05404  1.049  0.2942
dwj_open    0.01123  0.03148  0.357  0.7214
dwj_close   0.01091  0.03152  0.346  0.7292
nsq_open    0.15626  0.06496  2.406  0.0163 *
nsq_close   -0.11953  0.06516 -1.834  0.0669 .
---

```

●<그림 4> 첫 번째 회귀분석 결과

들로 원-달러 환율을 75% 정도의 정확도로 예측 가능하다는 말이다. 촉오팀의 목표는 90% 이상이었으므로 실망할 수밖에 없었다.

이때, 멘토의 조언이 따랐다. “대륙 대표 지수, 환율에 영향을 주는 추가 데이터를 추가해 보세요.”

그래서 촉오팀은 10년 채권 수익률, 5년 채권 수익률, 중국 상하이 지수, 유로/달러 환율, CAC40(프랑스 지수), 미국 국채 데이터, DAX(독일) CSI 300(중국) 등의 대륙 대표 지수, 다른 나라 화폐와의 환율, 주가 지수 등을 추가했다. 이로써 최종 사용된 변수는 52개였다.

```

> summary(lm.r0)
Call:
lm(formula = USD_KRW ~ ., data = exdf_bk.train[, -ind_org])
Residuals:
    Min      1Q  Median      3Q     Max
-40.670 -3.204  0.015  2.916  48.716
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.701e+02 1.044e+02 8.336 2.65e-16 ***
Call_rate_tomorrow_total -9.46e-02 7.217e-01 -1.175 0.24070
Call_rate_tomorrow_comissionmerchant_ -8.226e-01 7.217e-01 -1.118 0.269759
exchequerbond_3years -9.531e+00 6.197e+00 -1.538 0.124394
corporatebond_3years_AA_ -1.163e+01 4.644e+00 -2.505 0.012425 *
monetarystabilizationbond_91days 1.955e+00 5.029e+00 3.893 0.697560
monetarystabilizationbond_1year 1.859e+01 8.904e+00 2.088 0.097069 *
monetarystabilizationbond_2years 3.950e+00 7.757e+00 5.098 0.000257 **
exchequerbond_5years 3.467e+00 4.659e+00 0.538 0.460769
SSEC_low 3.655e-02 1.971e-02 0.955 0.623942
SSEC_high -1.398e-02 1.971e-02 0.447 0.455386
SSEC_Close -4.028e-02 1.693e-02 -2.373 0.017548 *
KDX_open 1.354e-02 5.015e-02 0.270 0.787187
KDX_close -1.171e-03 5.103e-02 -0.023 0.981692
KSP_open 5.089e-02 1.892e-02 2.690 0.007264 **
KSP_close -7.140e-02 1.847e-02 -3.865 0.000118 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
Residual standard error: 6.013 on 961 degrees of freedom
Multiple R-squared: 0.9771, Adjusted R-squared: 0.9757
F-statistic: 662.6 on 62 and 961 DF, p-value: < 2.2e-16

```

●<그림 5> 두 번째 회귀분석 결과

촉오팀은 놀랐다. R-Squared 값이 0.976 정도로 올라갔기 때문이다.

변수를 추가한 것밖에 없는데 결과는 놀라웠다. 표가 중간에 생략되어 잘 보이진 않지만, 가장 밀접한 중요변수로는 원-위

엔 · 위엔-달러 · 니케이(Nikkei) 225 변수 순으로 첫 번째 회귀분석 모델 안에서의 중요 변수들이 달라졌다.
이때 집체교육 중에 교수께서 했던 조언이 다시 떠올랐다. “미국 관련 지수들은 하루씩 늦게 적용해야 합니다.”

미국 관련 지수들은 하루씩 미흡



USDKRW 은 하루씩 달림

날짜	USDKRW	미국 지수	그 외
5/2			
5/3			
5/4			

5월 2일자 데이터를 이용해 5월 3일 USDKRW를 예측
한국 5월 2일자에 미국·유럽지수들은 5월 1일 데이터 이용

●<그림 6> 미국의 지수들을 하루씩 늦게 적용

<그림 6>을 보듯이, 5월 2일자 데이터를 이용해 5월 3일의 환율을 예측해야 했으므로 환율 데이터를 하루씩 뒤로 미루고, 한국이 5월 2일일 때 미국·유럽 지수들은 5월 2일 데이터이므로 하루씩 당겨 조정했다.

```

> summary(lm.r1)
Call:
lm(formula = NextDayFX ~ ., data = exdf_bk.train[, -ind_adj])
Residuals:
    Min      1Q  Median      3Q     Max
-46.855 -4.756  0.035  4.120  54.515
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.238e-02 1.447e-02 5.002 6.74e-07 ***
Call_rate_tomorrow_total -6.358e-02 1.000e-02 -0.064 0.945315
Call_rate_tomorrow_comissionmerchant_ 1.095e-01 1.002e-02 1.089 0.913045
exchequerbond_3years -1.852e-01 8.536e-02 -2.181 0.030307 .
corporatebond_3years_AA_ -1.852e-01 8.536e-02 -2.181 0.030307 .
monetarystabilizationbond_91days 1.513e+00 6.014e-02 2.516 0.022221 **
monetarystabilizationbond_1year 3.212e+00 1.240e-01 2.590 0.009744 **
monetarystabilizationbond_2years 1.123e+01 1.078e-01 1.047 0.295381
exchequerbond_5years 6.064e+00 6.531e-02 0.929 0.353364
SSEC_low 3.282e-02 2.604e-02 1.260 0.207870
SSEC_Close -4.554e-02 2.337e-02 -1.953 0.051145 .
KDX_open 1.847e-01 7.040e-02 2.622 0.008848 **
KDX_Close -1.395e-01 7.181e-02 -1.945 0.052303 .
KSP_Open 3.430e-02 2.605e-02 1.316 0.188356
KSP_Close -7.617e-02 2.543e-02 -2.995 0.002817 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
Residual standard error: 8.388 on 961 degrees of freedom
Multiple R-squared: 0.9955, Adjusted R-squared: 0.9927
F-statistic: 333.6 on 62 and 961 DF, p-value: < 2.2e-16

```

●<그림 7> 최종 회귀분석 결과

조정된 데이터를 가지고 마지막으로 돌려보았다. R-squared 값은 이전보다 조금 떨어졌지만, 그럼에도 많은 부분을 설명해 주었다. 만약 시간이 조금 더 있었더라면, 데이터를 정제하면서 실수는 없었는지, 다른 추가할 변수들을 찾아서 Adjusted R-squared 값이 0.99가 나을 때까지 도전해 볼 수 있었을 텐데 하는 아쉬움이 남았다.

CONCLUSION

회귀분석(regression analysis)은 관찰된 연속형 변수들에 대해 종속변수(우리가 알고자 하는 변수)와 독립변수(종속변수에 영향을 주는 변수) 사이의 상관관계를 나타내는 선형 관계식을 구하는 기법으로, 통계 예측에 많이 이용된다. 촉오팀이 사용한 회귀분석은 엄밀히 말해 하나의 종속변수와 여러 독립 변수 사이의 관계를 보는 다중 회귀분석이었다.

회귀분석을 이용하기 위해서는 여러 가정이 전제돼야 하지만, 여기서는 생략한다. 회귀 모형의 적합도는 R^2 값을 사용하는데, 이는 회귀모형의 독립 변수가 종속변수의 몇 %를 설명하고 있는지를 나타내는 지표다.

촉오팀의 프로젝트 결과는 다양하게 활용할 수 있다. 촉오팀이 아직 도전해 보지 못한 여러 예측 모델을 사용해서 예측을 해 보거나 또 다른 변수를 추가해서 R-squared 수치가 99% 가 될 때까지도 도전해 볼 수 있다. 그리고 데이터만 있다면, 다양한 방법으로 여러 산업 분야에서 활용할 수 있다.

첫째, 주·월 단위의 예측에 도전해 볼 수 있다

주간 · 월간 예측을 통해 기업에서 보유한 돈의 환율을 예측하여 외국환 거래 시 참고할 수 있을 것이다. 앞서 설명했듯이, 일별 데이터를 월간 데이터로 변형해 월 단위의 데이터까지 넣어 모델을 만들어 보고, 일별 · 월별 예측 데이터를 사용한다면 정확도는 올라갈 것으로 예상된다.

둘째, 추가 파생 변수들을 도출해 classification 방법으로 등락 예측할 수 있다

등락 폭으로 파생 변수를 생성하여 classification 방법으로 등락 자체 예측을 추가하면, 더 정교한 모델을 도출해 기업 투자에 활용할 수 있다.

예측에는 항상 한계와 위험성이 있게 마련이다. 환율의 변동의 차가 크지 않기 때문에 오차의 범위가 좀 더 클 수도 있다. 아무리 예측을 잘한다 해도 당시의 시장 상황이나 정부의 적극적인 시장 개입 등 고려해야 할 변수를 이번 모델에 적용하지 못한 점은 예측의 한계라고 볼 수 있다.

모두가 분석가가 될 수 있다!

우리 팀원들 중에 분석 전문가는 아무도 없었다. 그럼에도 8주라는 기간 동안 많은 것을 배웠고 각자 많은 것을 느꼈다. 요즘은 도구를 구입하지 않아도 되고, 자료를 얻기 위해 도서관을 들락거릴 필요도 없다. 누구에게나 데이터는 개방되어 있고, 누구나 사용할 수 있도록 제공되는 무료 분석 도구를 쉽게 접할 수 있다. 인터넷이 발달한 지금은 우리 모두에게 기회가 열려 있고, 우리가 만들고자 하는 것을 만들 수 있는 시대다. 앞서 알아 봤듯이 촉오팀에서 시도한 모델은 회귀분석 모델 가운데 하나다. 모델보다 선행되어야 할 중요한 것은 데이터를 모으고 정제하는 과정이다. 이 과정이 실제 분석의 80% 이상을 차지했다고 볼 수 있다. 무엇을 하기 위해 어떤 데이터가 필요하고, 어떻게 모을 것인가에 집중한다면, 그 이후에 모델을 찾고 분석하는 것은 틀이 해 줄 것이다.

예측 시스템이나 분석 프로그램은 아무리 아마추어라 하더라도 데이터를 모을 수 있는 끈기, 원하는 분야에 대해서 공부하려는 자세와 열정이 더해진다면, 촉오팀보다 더 좋은 결과를 이끌어 낼 수 있다고 생각한다. 두려워 말고 한번 도전해 보기를 바란다. 누구나 전문가의 길로 들어설 수 있는 시대가 왔다. 그리고 이 글로 인해 누군가에게 도움이 되길 바란다. ☺

“데이터 주도권의 흐름을 알고 싶었다”



백형충 팀장
케이씨에이 수석, 정보처리기술사

주제가 매우 전문 영역으로 느껴진다.

팀원들도 그렇게 생각했다. 이 분야에 대한 전문 지식을 가진 사람도 없었고, 한 팀원이 대기업의 전략기획실에서 일하며 국내외 각종 경제지표를 참고해 업무를 했던 경험을 갖고 있던 것이 그나마 다행이었다. 데이터 분석의 맛을 제대로 보기 위해서 익숙한 것보다는 조금 낯설더라도 도전적인 주제를 선정하고 싶었다. 팀원들에게 제대로 한번 해 보자고 제안했는데 흔쾌히 동의해 줬다.

환율 예측과 관련한 기존 연구와 데이터 분석으로 접근한 최오팀의 시도와는 어떤 차이가 있나.

기존 학술 연구에서는 통계적 접근을 하고 있다. 많은 데이터 중에서 일부를 샘플링해 전체를 예측하는 방법이었다. 반면 최오팀의 환율예측 프로젝트는 5년치의 실제 로우 데이터를 확보해, 수많은 변수들에 대해서 직접 탐색적 분석을 통해 상관관계가 있는 변수들을 도출해 보텀업 접근 분석 모델링을 만들었다는 점에서 차이가 있다. 당초 57개의 변수가 나중에는 60개 이상으로 늘어났다. 가장 영향력 있는 변수가 무엇인지 알아보기 위해 몇 편의 논문을 참고했는데, 대부분의 연구 논문에서 2~3개의 변수로 접근하고 있었다.

어떤 일을 하고 있나.

정보처리기술사로서 감리 업무를 하고 있다. 향후 데이터 아키텍트와 데이터 사이언티스트를 목표로 준비중이다. 이를 위한 과정으로 데이터 아키텍처 전문가(DAP, Data Architecture Professional)와 분석 전문가(ADP, Advanced Data Analytics Professional) 자격증 시험을 준비중이다.

교육 과정 수료자로서 바라는 바가 있다면.

2주간 실시되었던 집체교육의 밀도가 매우 높았다. 하지만 수료 프로젝트는 직장에서 실무를 하면서 진행해야 했는데 이것이 어려움으로 작용했다. 가능하다면 수료 프로젝트를 단기 1주일이라도 집체교육 방식을 통해 집중력과 효율성을 높일 수 있도록 시도해 보는 것도 좋을 것 같다. 혼자가 아닌 여러 명이 1주일 동안 집중하여 협업을 한다면, 떨어져서 했을 때보다 더 좋은 결과가 나오지 않을까.

특별히 기억에 남는 일은.

주제가 조금은 낯설었으므로 모델링하기까지가 어려웠다. 우리 팀의 멘토는 팀원들의 열띤 정반합의 과정을 통해 뭔가가 나타날 때까지는 전혀 도와주지 않았다. 하지만 어느 정도의 윤곽이 드러나려고 할 때, 살짝 던져주는 한마디가 막막한 어둠 속에서 항해를 하다가 등대 불빛을 만난 것처럼 길잡이 역할을 해 주었다.

데이터 분석을 놓고 망설이는 이에게 조언한다면.

(엔지니어링 관점에서) 자동차를 움직이게 하는 것이 엔진이듯, 정보 시스템의 비즈니스 가치를 만들어내는 핵심은 데이터다. 데이터가 얼마나 체계를 갖췄는지가 데이터 분석에 결정적으로 영향을 준다. 과거 데이터 웨어하우스나 비즈니스 인텔리전스(BI)가 데이터를 구축하는 데 역점을 둔 접근이었다면, 데이터 분석은 활용하는 데 중점을 둔 접근이라고 볼 수 있다. 데이터의 구축은 엔지니어의 영역이지만, 활용은 실무자의 영역이다. 시대가 바뀌면서 데이터의 권력이 어느 쪽으로 이동하는지를 지켜볼 필요가 있다.

원-달러 환율 데이터 분석을 통한 일일 환율 예측

프로젝트 소개

지난 5년 간 실제 원-달러 환율 데이터를 바탕으로 환율에 유의미한 영향변수를 연구·도출해 탐색적으로 분석하는 프로젝트. 최적의 상관변수를 선정·적용하여 일일 원-달러 환율 예측 모델을 구축해 환율 예측에 새로운 가능성을 제시했다.

구분

분석 전문가 과정 : 회기분석

프로젝트 기간

2014년 07~08월

멘토

안정국(The ECG 상무)

적용도구

R

수집 데이터

통계청·다음·Investing.com에서 수집한 최근 5년의 주가지수, 원자재, 외환, 금리, 환율 민감주식, 경기 민감주식 데이터

산출물

일일 KRW-USD 환율 예측 시스템

교육 참여형태

자발적 참여(5) / 회사 권유(0)

진행

• 백형충 팀장	IT컨설팅사 수석 연구원	경력 29년
• 윤상백 팀원	—	경력 27년
• 김성희	중공업업체 IT부서 대리	경력 6년
• 차순표	SI업체 차장	경력 10년
• 최선애	SI업체 부장	경력 14년

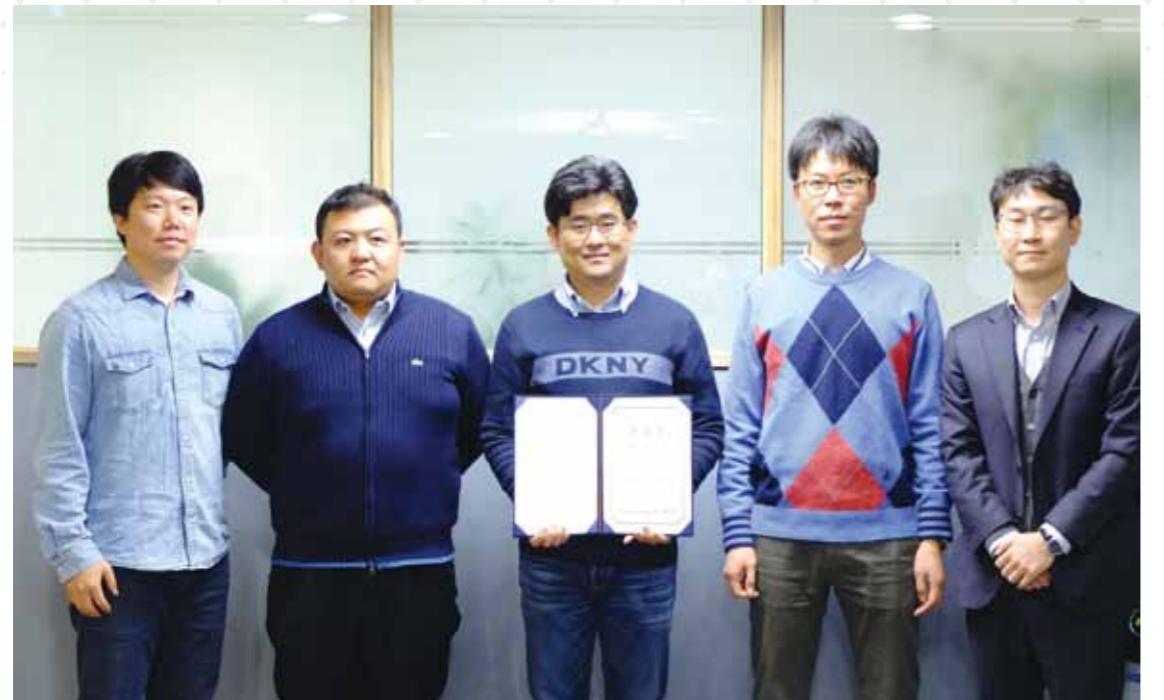
빅데이터 아카데미 수강 후 변화

- | | |
|------|---|
| 수강 전 | <ul style="list-style-type: none"> • 빅데이터 처리 분석은 굉장히 어려운 주제였다. • 빅데이터는 먼 곳에 있다. |
| 수강 후 | <ul style="list-style-type: none"> • 빅데이터 처리를 위한 다양한 기술 존재 • 도메인에 대한 이해만 있다면 얼마든지 도전 가능 • 빅데이터는 생각보다 가까이 있다. |





데이터 분석과 재즈 페스티벌이 만났을 때



글 안진훈 아이리치그린 대표

매년 28% 이상 관람객이 늘어나면서 누적 관객수 100만 명을 이미 돌파한 가평 '자라섬 재즈 페스티벌'. 대표적인 지역 축제로 자리매김한 이 행사를 분석해 발전에 필요한 사항들을 도출했다. 우리나라에서는 전국적으로 연간 550개가 넘는 크고 작은 축제가 개최되고 있다. 그 축제들에서 이 분석 모델을 참고할 수 있기를 기대하면서 수료 프로젝트를 진행했다.

CHALLENGES

빅데이터 분석 파이를 키우자

몇 년 전부터 기관·기업 할 것 없이 빅데이터를 얘기하고 있다. 널리 회자될 만큼 인기 있는 주제임에도 국내에서는 가시적인 성과가 드문 게 현실이다. 용어 자체는 일반화됐지만 성과는 오히려 정체된 느낌이다. 당연한 얘기지만, 빅데이터는 데이터 양뿐 아니라 데이터에 대한 새로운 시각까지 포함하고 있다.

팀원 6명 가운데 두 명을 제외한 4명이 이공계 출신인 우리 팀은 데이터 분석 활성화에 작으나마 보탬이 되어 이 분야의 산업 파이를 키워보자는 의미에서 팀 이름을 '빅파이'라고 지었다.

빅파이팀이 수료 프로젝트로 진행한 '자라섬 재즈 페스티벌 관람객 분석 및 예측'은 바로 도출했던 주제는 아니었다. 처음엔 한 빅데이터 경연대회 출전을 염두에 두고 접근해 보았다. 대회 주최측에서 데이터세트를 제공해 주는 행사였다. 하지만 빅데이터 아카데미 교육생으로서 멘토의 지도 아래 아마

추어 대회에 참가하는 것이 그 취지에 벗어난다고 판단해 주제를 바꿨다.

흥미롭고 유익한 주제 찾기

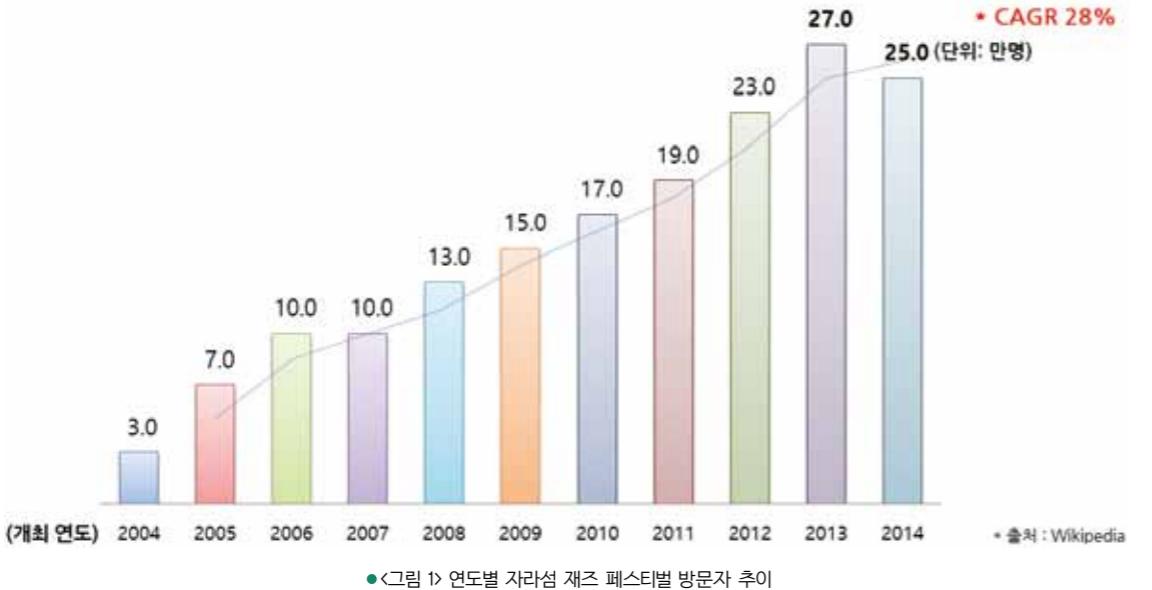
처음부터 주제 선정 과정을 너무 쉽게 생각해서 일까? 다른 주제를 찾는 데 많은 시간이 걸렸다. 무엇보다 주제를 선정할 때 어떤 대의명분을 가지고 프로젝트를 해야 하는지 몇 가지 원칙을 세울 필요가 있었다.

빅파이팀이 주제 선정 시 세운 원칙은 다음과 같다. 첫째, 8주 간의 프로젝트인 만큼 동기부여가 될 수 있도록 재미 있고, 데이터 분석 과정에서 배운 내용을 복습할 수 있어야 한다. 교육 시간에 처음 접해 본 분석도구 R을 프로젝트 실전에서 익혀 보려고 했다. 둘째, 다른 기수에서 시도하지 않았던 주제이거나 중에 자신뿐 아니라 소속된 조직에 도움을 줄 수 있도록 활용 범위가 넓어야 한다. 셋째, 데이터 분석 산업 활성화에 조금이나마 기여할 수 있고 향후 비즈니스 아이템으로 연결할 수 있어야 한다.

이런 주제로 고민하던 중 우연히 않게 프로젝트 기간 중에 '자라섬 재즈 페스티벌'이라는 지역 축제를 접하게 되었다.



• 이미지 출처: www.jarasumjazz.com



자라섬 재즈 페스티벌은 2004년에 시작되어 2013년 기준 누적 관객수 100만 명을 넘어선 행사다. 연평균 28% 이상의 놀라운 관람객 성장률을 기록하면서 한국을 대표하는 음악 축제로 자리잡았다. 2014년에는 음악 축제 최초로 최우수 축제로 지정되는 등 명성을 쌓아가고 있다.

프로젝트 주제 선정을 위해 국내에서 개최되는 여러 축제에 대해 조사해 보았다. 매년 전국적으로 550개 이상, 서울에서 만 연간 50개 이상의 지역 축제가 개최되고 있다. 하지만 이런 지역 축제에 대한 데이터 분석은 국내외를 막론하고 사례를 찾아보기 힘들었다. 최근에야 지역축제에 대한 몇몇 보고서가 나왔는데, 읽어보니 시간 · 연령 · 성별 유동인구와 매출 등 단지 현황 위주의 분석 결과로서 실제 지역 축제의 성공과 실패 요인 분석은 아니었다. 대략 8TB의 데이터를 수집 · 분석했다는데, 그 방대한 양의 데이터에 비해 분석 결과가 아쉬웠다. 여러 논문도 확인해 봤으나 상관관계나 예측 분석이 부족해 보였다. 이런 여러 상황을 반영해 빅파이팀은 '자라섬 재즈 페스티벌 관람객 분석 및 예측'을 최종 주제로 선정했다.

SOLUTION

시작은 해결할 문제의 이해

데이터 분석을 진행하는 단계는 해결해야 할 문제와 환경에 따라 세부적인 절차는 다를 수 있지만, 일반적으로 이해하기 쉽고 여러 산업 영역에 적용 가능한 CRISP-DM(Cross Industry Standard Process for Data Mining)을 프로젝트 방법론으로 적용했다. 이 방법론은 업무(문제)의 이해 → 데이터의 이해 → 데이터 준비 → 모델링 → 평가 → 적용을 따르도록 소개하고 있다.

빅파이팀의 멘토인 전용준 박사로부터 자문을 받아, 데이터 분석 프로세스의 첫 단계이자 가장 중요한 단계인 해결할 문제부터 알아 보았다. 기획과 가설 구상력(構想力)을 갖춘 다음, 분석 과제와 목적 및 우선순위를 결정했다. 업무 이해 단계에 수립된 가설에 필요한 데이터의 레이아웃을 정의했고, 이를 확보하기 위해 각 시스템의 데이터를 추출 · 가공해 그 결과를 검증하는 과정을 거쳤다.

가설에 기반해 처리 데이터로 모델을 구축했고, 집체교육 과정

에서 학습한 R과 분석 기법을 이용해 가설 검증을 반복했다. 이 과정에서 가설과 업무 현장의 경험과 직관을 융합해 모델을 보완하는 과정이 매우 중요함을 알게 됐다. 데이터 분석 방법은 멘토의 조언과 더불어 교육 과정에서 실습해 본 경험을 되살려보고자 군집 · 분류(의사결정나무, 랜덤 포리스트) · 회귀분석으로 접근해 보았다.

멀고도 험한 데이터의 정의와 수집

주제를 선정하고 나니, 수집해야 할 대상 데이터의 종류와 데이터세트 구성은 별로 어렵지 않을 거처럼 여겨졌다. 초반에 많은 시간을 쓰면서 주제 선정에 난항을 겪은 터라 데이터 수집 시간을 절약하고 좀 더 양질의 데이터를 확보할 수 있기를 바랐다. '자라섬 재즈 페스티벌 행사 사무국'으로부터 매출 데이터, 티켓 판매, 관람객 통계정보, 참여한 아티스트 정보, 시간별 공연 프로그램, 참여 스폰서 등의 원천 데이터 협조를 요청했다. 하지만 2주 후 데이터를 받을 수 없음을 알게 되었다. 너무 허탈했지만, 프로젝트 기간이 6주밖에 남지 않은 상

황이었기에 빠른 대안 마련이 시급했다. 필요한 데이터를 팀원들이 직접 수집했다. 최적화한 데이터가 아니었기에 데이터를 이해 · 수집하는 과정은 지루하다 못해 여러 생각(?)이 교차하게 했다.

팀원들과 머리를 맞대고 업무 이해 단계에서 수립된 가설에 필요한 데이터 레이아웃을 정의했다. 이를 확보하기 위해 각 시스템의 데이터를 추출 · 가공하고, 그 결과를 검증하는 일은 이번 프로젝트 업무에서 매우 큰 부분을 차지할 정도로 멀고 험난하기만 했다. 이 과정을 겪으며 팀원들은 데이터는 무궁무진하게 산재해 있지만, 필요한 데이터를 얻는 데에는 많은 시간과 노력이 필요함을 실감했다.

이 축제에 대한 일반정보는 자라섬 재즈 페스티벌 홈페이지에서, 뮤지션과 관련된 관심도 · 활동 경력 · 동영상 조회 수 등은 구글과 유튜브에서 구할 수 있었다. 티켓 예매량 데이터는 인터파크와 YES24에서 얻을 수 있었고, 연도별 와인 수입량과 국내 총생산 증가율 등을 통계청에서 가져올 수 있었다. 사회 · 문화 트랜드 및 날씨 · 교통 · 경제와 관련된 데이터는

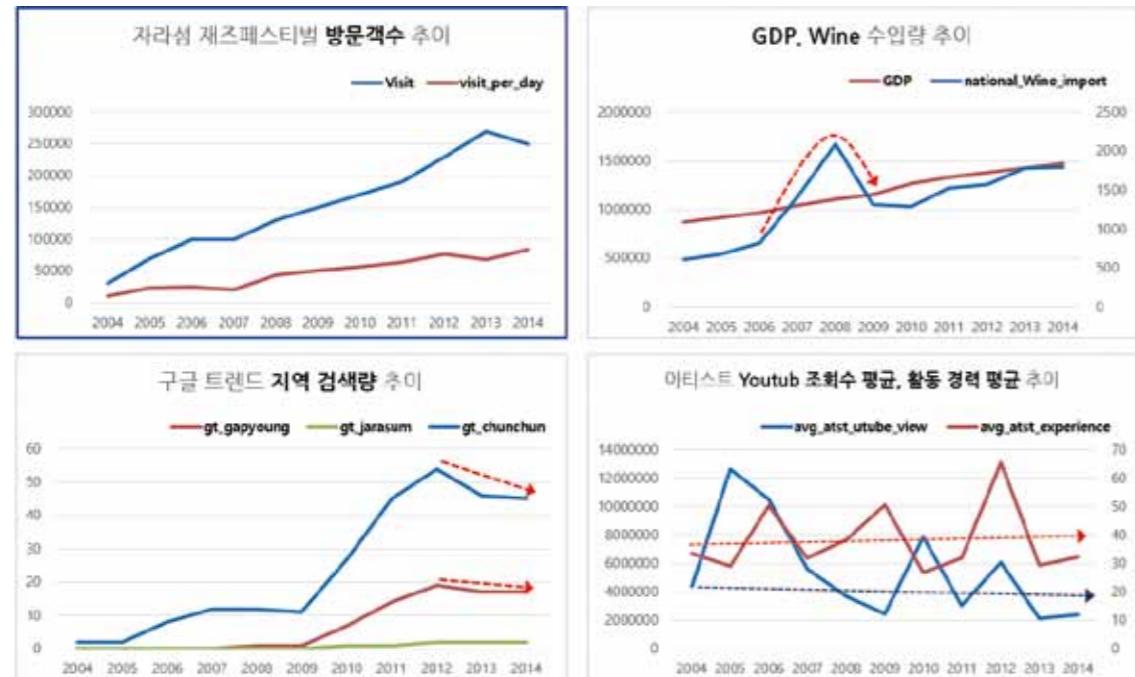
	30000	70000	100000	100000	130000	150000	170000	190000	230000	270000	250000
Visit_Wave	1	2	3	4	5	6	7	8	9	10	11
Year	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
duration_days	3	3	4	5	3	3	3	3	3	4	3
visit_per_day	10000	23333	25000	20000	43333	50000	56667	63333	76667	67500	83333
avg_GT_artistT3	3	6	4	4	2	1	3	3	9	3	3
avg_atst_utube_view	4398957	12661307	10497992	5612566	3654285	2446790	7856284	3030812	6085740	2161922	2413752
avg_atst_experience	33.666667	29.50.333333	32	38.3333333	50.6666666	26.6666666	32	65.3333333	29.3333333	32.3333333	3
max_youtube_index	5238265	33135430	21104314	8623592	5089836	3927620	18308970	6139135	12293783	3262653	3923829
atst_like_reg_na	0	1	0	1	0	0	0	0	1	1	1
atst_like_reg_eu	1	1	1	1	1	1	1	1	1	1	1
atst_like_req_ot	1	1	1	0	0	0	1	0	1	0	0
GT_Jazz	0	0	0	0	7	20	33	37	42	38	40
CumGT_Jazz	0	0	0	0	7	27	60	96	138	176	216
NTpc_Jazz	0	0	0	63	39	33	27	22	17	16	14
NTmb_Jazz	0	0	0	0	0	0	4	25	47	60	72
CumNTmb_Jazz	0	0	0	0	0	0	4	29	77	136	208
gt_gappyoung	0	0	0	0	1	1	7	14	19	17	17
gt_jarasum	0	0	0	0	0	0	1	1	2	2	2
gt_chunchun	2	2	8	12	12	11	27	45	54	46	45
av_temp	19	22.1	19.3	21	17.5	13.1	12.6	10.5	12.1	16.2	15.3
max_temp	26.8	27.4	28.2	27.8	24.9	20.8	21.2	19.6	22.2	27.6	21.4
min_temp	13.5	16.3	10.1	17.2	10.8	5.3	4.9	3.8	5.6	8.2	9.5
ls_rainy	1	1	0	0	0	0	0	0	0	0	0
rainfall	44	10.5	-	-	-	-	-	-	-	-	-
day_tempran_av	4.6	4.3	10.7	8.1	9.2	6.8	8	8.3	8.8	11.3	6.1
day_tempran_mx	13.3	7.7	16.4	10.6	13.9	14	16.3	15.8	15.4	16.9	11.9
train	0	0	0	0	0	0	1	1	1	1	1
highway	0	0	0	0	0	1	1	1	1	1	1
national_Wine_import	605	685	824	1411	2094	1313	1286	1523	1577	1785	1800
GDP	876,033	919,797	966,055	1,043,258	1,104,492	1,151,708	1,265,308	1,332,681	1,377,457	1,428,295	1,471,144
GDP growth	3	5	4	5	6	3	1	7	4	2	3

●그림 2) 구성된 데이터세트

정보 유형에 따라 구글·네이버 트랜드, 통계청, 기상청, 카인즈 언론기사 등에서 수집했다.

'스몰 데이터'에서 찾은 데이터 분석의 가치

전체 분석을 위해 회차별 관람객 수를 종속변수로, 관람객 증가에 영향을 미칠 수 있는 요인을 독립변수로 각각 정의했다. 자라섬 재즈 페스티벌 관련 기초 변수는 전체 11개의 표본에 29개의 독립변수로 정의했다. 많은 시간을 투자해 수집했지만 표본이 적었다. 아울러 상관관계 분석을 위해 재즈 페스티벌에 출연한 아티스트 관련 기초 변수를 뮤지션의 관심도, 동영상 조회수, 활동 경력, 선호 국(북미·유럽·기타) 등으로 정의했다. 여기서 63개의 표본에 18개의 변수를 도출했다. 이 또한 기초변수의 표본보다는 개수는 많지만, 분석을 위해서는 충분치 않은 양이었다. 이에 빅파이팀은 분석을 계속해야 할지를 놓고 고민할 수밖에 없었다. 왜냐하면 통계 분석을 위해서는 최소 표본 크기가 30개 이상이라는 조건이 있는데, 빅파이팀이 수집하고 정제한 데이터세트는 11개였기 때문이다.



●그림 3> 기초변수들의 연도별 추이

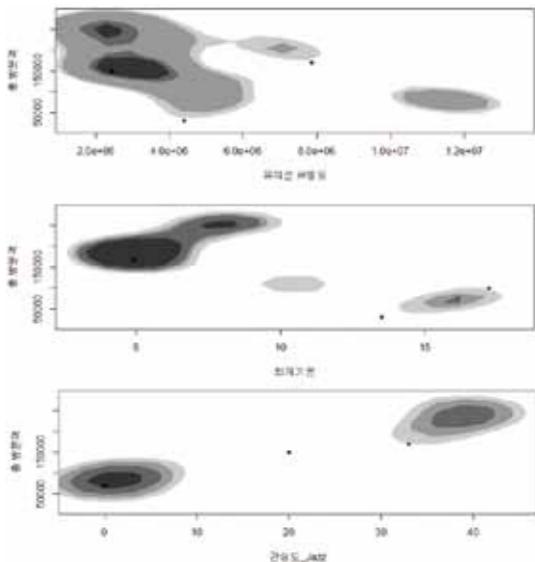
결국 고려대학교 통계학과 허명희 교수에게 자문을 구했다. '데이터 표본(n)이 10개 정도이면서 분석해야만 했을 때, 계량 분석이 통계적으로는 전혀 의미가 없다고 봐야 할까요?'라는 질문에 "표본(n)이 30개 이하는 별 의미가 없다. n 값이 10개 미만인 자료에 대해 랜덤 포리스트(Random Forest), 로지스틱스 등 통계적 모형(추론)이 만들어지면, 일반적인 문제는 예측의 변동성(variability)이 커진다는 점이다. 이에 따라 재현성은 감소한다. 예를 들어, n=10인 야구 결과에 대해 검증은 얼마든지 가능하다. 즉 두 팀의 저력이 동일하다는 가설에 대한 p-값 산출 정도는 문제가 없다. 다만 신뢰 구간이 너무 넓으므로 별 의미가 없어진다"는 답변을 받았다. 예측의 변동성이 커지면서 재현성은 감소하지만, 충분히 분석할 수 있다는 결론이다. 허 교수의 조언에 힘입어, 빅파이팀의 분석 방향은 표본(n)보다 변수(p)가 큰 유형의 분석 과제로 접근하기로 결정했다. 이를 통해 관람객 증가 요인을 파악하고자 했으며, 향후 주관 사무국이 어떻게 효율적으로 행사를 기획·운영해야 하는지 인사이트 제공을 목표로 잡았다.

분석 방향을 결정한 후, 먼저 기초 변수들의 연도별 추이를 살펴 봤다. 자라섬 재즈 페스티벌 방문객 수, 일 평균 방문객수는 꾸준히 증가 추세였다. 특히 GDP나 지역명 검색어 추이가 방문객 수와 좀 유사하게 변하고 있음을 알 수 있었다. 와인 수입량은 2008년 일시적 급등 이후 막걸리 유행 등에 따라 예전 수준으로 완만하게 상승하는 모습을 보였다. 지역명 검색에서는 '춘천' 대비 '가평'의 검색 건수의 감소폭이 낮았다. 더불어 '좀 더 유명한 뮤지션이 초청될수록 관람객이 증가할 것'이라는 가설을 분석해 보니 수집된 데이터에서는 이렇다 할 변화를 찾을 수 없었다.

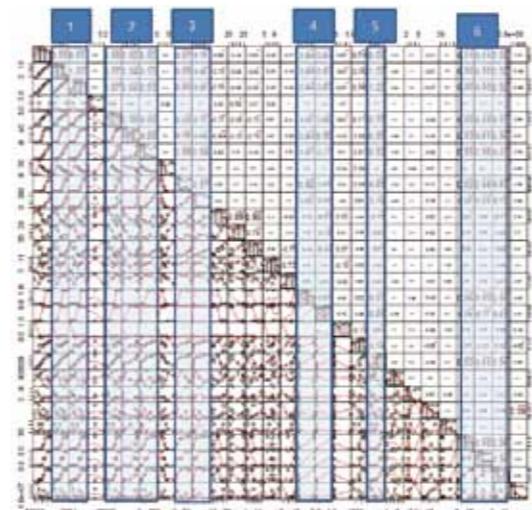
데이터 속에 숨어있는 패턴 찾기

1차 기초변수들의 추이를 파악한 다음, 빅파이팀이 구성해온 데이터세트 속에서 상식적으로 쉽게 예측하지 못했던 변수들 간의 상관관계를 분석했다. 이것은 어떤 두 사건의 발생 패턴을 분석하는 것으로 '패턴분석'이라고도 한다. 이런 상관관계 분석은 비교적 단순한 작업이지만, 분석할 대상이 많고 데이터 양이 많을 때는 점검도 어렵고 시간도 오래 걸린다. 이때 정확하고 빠른 상관관계 분석에서 힘을 발휘하는 것이 분석 도구 R이다. R에서 지지도와 신뢰도를 apriori() 함수로 간단히 구현할 수 있었다. 지지도는 0.001 이상, 신뢰도는 0.5 이상인

것들을 찾아냈으며, 결과는 <그림 4>와 같다. 총 방문객과 수집된 31개 변수들 사이의 상관관계를 분석한 결과, ①시계열적인 요인 ②재즈에 대한 관심도(구글 트랜드) ③재즈에 대한 네이버 모바일 관심 트랜드 ④교통 인프라(철도, 고속도로 개통) ⑤GDP ⑥가평·자라섬·춘천 등 페스티벌 인근 지역 관심도에 대한 구글 트랜드 등 총 6개 변수군에 대한 상관관계가 높은 것으로 나타났다. 이어서 방문객 수에 영향을 미치는 변수들을 확인하는 분석을 했다. 분석 결과, 유명 뮤지션 초청이 관람객 증가로 연결된다 는 가설과 야외 행사이므로 날씨가 추울수록 방문객이 줄어들 것이라는 가설 또한 기각됐다. 결국 '재즈에 대한 관심 증가가 방문객 증가와 관련이 높다'는 것을 <그림 5>의 등고선 그래프를 통해서 확인할 수 있었다.



●<그림 5> Plot 등고선 차트: 아티스트와 기온은 상관관계가 낮게 나타났다.



●<그림 4> 상관관계 분석: 총 방문객과 상관관계가 높은 주요 변수(군)

다음으로는 분류분석(Classification)에서 가장 많이 사용되는 의사결정나무(Decision Tree) 분석을 해보았다. 빅파이팀에서 작성한 데이터세트의 변수 종류가 그리 많지 않았고 복잡하지 않았기에 사용하기 적합하다고 판단해 R의 ctree() 함수로 의사결정트리를 도출했다.

잘 나눈 변수 그룹으로 의사결정 해법을 찾다

GDP 변수에 따라 관람객 수가 분리되지만, 95% 신뢰도에서 p-value가 0.05 이상으로 나왔다. 결국 의사결정나무는 유의하지 않은 것으로 분석됐다. 특히 GDP로만 잡힌다는 것은 시간이 지날수록 늘어날 가능성이 높으므로 별 인사이트가 되지 못한다고 판단했다. 좀 더 다양한 변수를 반영하고 정교하게 다양한 인사이트를 제공할 수 있는 모델링이 필요했다.

GDP 변수에 의해 방문자 수가 분리됐지만 95% 신뢰도에서 p-value가 0.05 이상으로 나와 의사결정나무 모형은 유의하지 않다고 판단했다.

결국 군집 및 의사결정나무의 대안이 필요했다. 멘토의 조언으로 랜덤 포리스트를 추천 받아 R 스크립트를 작성해 실행해 보았다. 100회 반복을 통해 얻은 모델은 20회 이후부터는 오차율에서 큰 변동이 없는 것으로 나타나 안정적이라고 판단했다.

랜덤 포리스트는 양상별 학습 기법을 사용한 모델이다. 양상별 학습은 주어진 데이터로부터 여러 모델을 학습하고 예측 결과들을 종합해 정확도를 높이는 기법이다.

이 기법은 두 가지 방법을 사용해 다양한 의사결정나무를 만든다. 첫 번째 방법은 데이터의 일부를 복원 축출로 꺼내 해당 데이터에 대해서만 의사결정나무를 만드는 것이다. 즉 각 의사결정나무는 데이터의 일부만 사용해 만들어진다. 두 번째 방법은 노드 안의 데이터로 자식 노드를 나누는 기준을 정하는 것이다. 즉 전체 변수가 아닌 일부 변수만 대상으로 해 가지를 나눌 기준을 찾는 방법이다. 새로운 데이터에 대한 예측을 수행할 때는 여러 개의 의사결정나무가 내놓은 예측 결과를 투

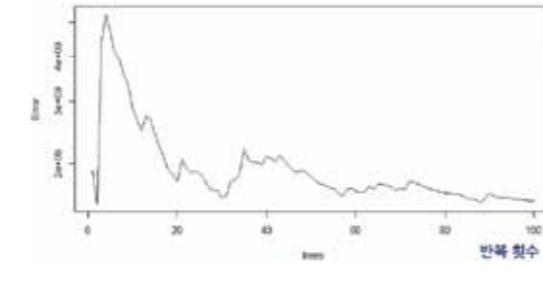
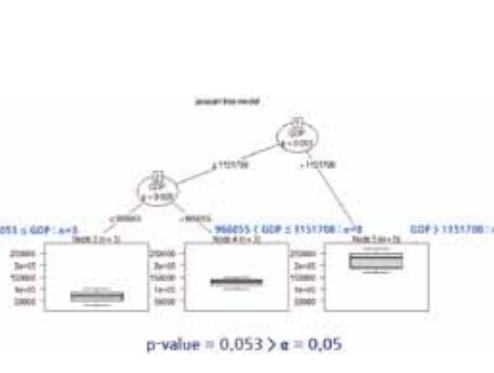
표(voting) 방식으로 합한다. 예를 들어, 총 5개의 의사결정나무 중 Y를 예측한 나무가 3개, N을 예측한 나무가 2개면 Y를 최종 결과로 결정한다. 랜덤 포리스트는 일반적으로 성능이 뛰어나고, 의사결정나무는 하나가 아니라 여러 개를 사용해 과적합(overfitting) 문제를 피한다.

결과가 나오기 시작하다

결정된 랜덤 포리스트 방식으로 우선 분석해 보았다. 역시 R을 접한 지 얼마 안된 터라 실전 코딩이 쉽지 않았다. 시행착오를 거쳐가며 데이터 분석을 위한 R 스크립트를 한 줄 한 줄 작성해가기 시작했다. 그나마 다행스럽게 다른 개발 도구를 사용해 본 개발자가 팀원 중에 있었고, 멘토께서 참고할 만한 소스코드와 분석 패키지를 제시해 주어서 속도를 낼 수 있었다.

랜덤 포리스트로 분석한 결과, MSE(mean square error, 평균 제곱 오차)의 퍼센트 증가(%IncMSE) 기준으로 ‘구글 트랜드 가평지수 > 구글 트랜드 누적 재즈 지수 > 구글 트랜드 재즈 지수’ 순으로 중요도를 보여줬다. 의사결정나무에서 도출한 GDP라는 기준보다 더 중요한 변수들이 있음을 랜덤 포리스트 분석으로 파악할 수 있었다. 기대한 대로 랜덤 포리스트 분석 모형이 다양한 변수의 영향을 상대적 중요도와 함께 제공해 줌을 알 수 있었다.

빅데이터팀은 좀 더 깊이 있고 정확성 높은 분석을 위해 추가 파생변수를 정의해 보았다. 기존 변수 중 상관분석을 통해 의미가 있을 거라 판단한 ‘최고 기온과 최저 기온의 차, 아티스트 경력에 대한 유튜브 조회 수, 춘천 구글 트랜드에 대한 가평



```

54 # Party 나무 모형 스크립트
55 jara_tree_der <- ctree(x.visit. ~ ., controls = ctree_control( mincriterion = 0.1, minsplitt
56 = 2, minbucket = 3), data=jarader)
57 plot(jara_tree_der, main="jarasum tree model(derived variable)")
58 # Random Forest 분석 스크립트
59 jarader_rf <- randomForest(x.visit.~., data=jarader, importance=TRUE, do.trace=5, ntree
=100)

```

●<그림 8> R로 구현한 의사결정나무 및 랜덤 포리스트 분석 스크립트

구글 트랜드 비율’ 같은 파생변수를 plyr 패키지의 ddply() 함수로 생성했다.

아티스트보다 행사 인지도가 더 중요

파생변수를 생성한 후 랜덤 포리스트로 2차 분석을 한 결과, ‘구글 트랜드 자라섬 지수 > 구글 트랜드 춘천 지수 > 구글 트랜드 누적 재즈 지수’ 순으로 중요도가 도출됐다.

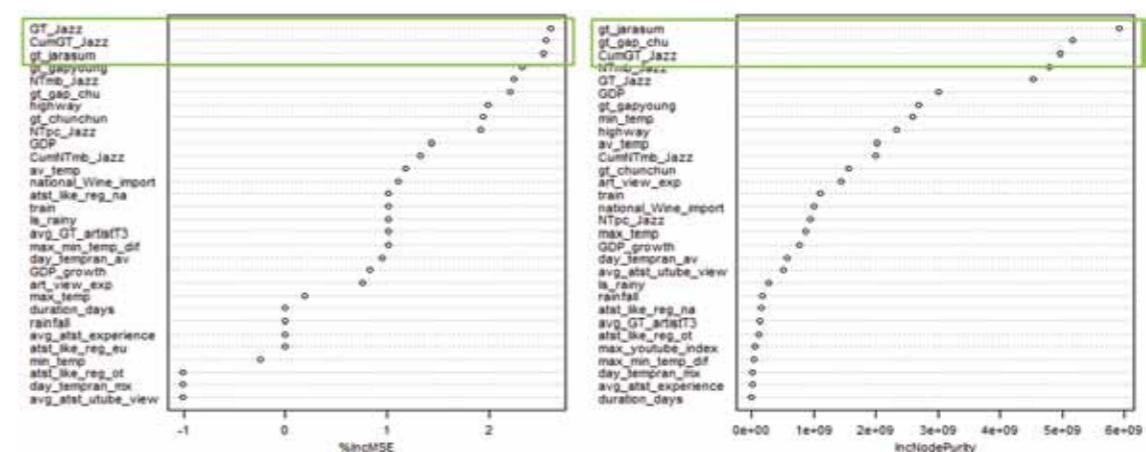
파생 변수	R Script
최고 기온과 최저 기온의 차	ddply(jarasum, .(Wave), transform, max_min_temp_dip=max_temp-min_temp)
아티스트 경력에 대한 유튜브 조회수 비율	ddply(jarasunder, .(Wave), transform, art_view_exp=avg_atst_utube_view/avg_atst_experience)
춘천 구글 트랜드에 대한 가평 구글 트랜드 비율	ddply(jarasunder, .(Wave), transform, gt_gap_chu=gt_gapyung/gt_chunchun)

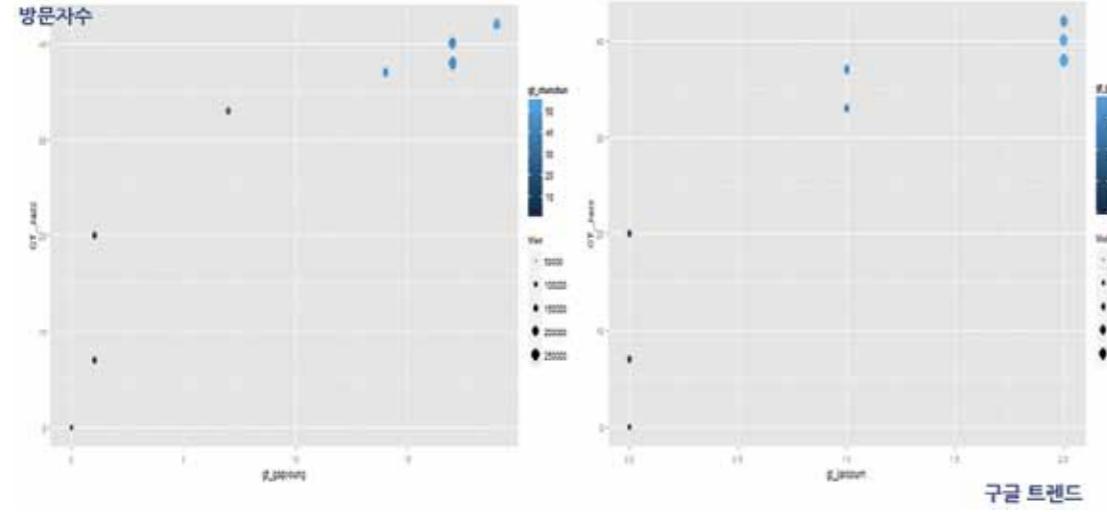
●<표 1> 파생변수 정의

이번 2차 분석에서는 추가된 파생변수가 중요한 요소로 등장했다. 이를 통해 파생변수 추가가 좀 더 깊이 있는 랜덤 포리스트 분석을 위해 매우 유용한 시도였음을 알게 됐다. 빅데이터팀이 첫 가설로 내세웠던 참가한 아티스트 변수가 매우 중요할 것 같지만 생각보다는 다양한 정보를 넣어 봤음에도 참여 아티스트보다 행사 인지도와 재즈에 대한 일반적 관심도 증가가 더 중요한 요인이었음을 알 수 있었다.

분석 결과 고찰: 어디에 집중해야 할까?

몇 주에 걸쳐 데이터 수집과 전처리 작업, 다양한 분석기법을 적용해 구현한 분석 모델을 검증해 ‘2015년 자라섬 재즈 페스티벌’의 방문자 수를 예측한 결과를 얻을 수 있었다. 좀 더 정확한 예측을 위해 다양한 파생변수를 생성해 더 깊이 상관분석을 해보고 싶었으나 프로젝트 완료 시간이 하락하지 않았다.





●<그림 10> 파생변수와 방문자 수의 상관관계

파생변수가 포함된 데이터세트에서 중요도가 높았던 가평·재즈·춘천 구글 트랜드 변수와 방문자 수 관계, 자라섬·재즈·가평·춘천 구글 트랜드가 높을수록 예측 방문자 수가 증가함을 아래 결과를 통해 알 수 있었다.

- 자라섬·가평·춘천에 대한 구글 트랜드 11회까지의 평균값을 넣고 재즈에 대한 구글·네이버 트랜드를 순차적으로 늘려본 결과, 예측 방문자 수가 증가했다.
- 재즈에 대한 구글·네이버 트랜드 값으로 11회 평균값을 넣고 자라섬·가평·춘천의 구글 트랜드값을 순차적으로 늘려본 결과, 예측 방문자 수가 증가했다.

- 재즈 트랜드와 자라섬·가평·춘천의 구글 트랜드값을 함께 늘려본 결과, 예측 방문자 수는 개별적으로 증가 시킬 때보다 더 높게 증가했다.

따라서 재즈와 자라섬, 가평, 춘천 구글·네이버 트랜드는 방문자 수 증감에 중요한 요인으로 작용하고, 개별적 트랜드보다 함께한 트랜드가 방문자 수 증감에 더욱 크게 영향을 미친다고 판단했다.

만일 빅파이팀이 '자라섬 재즈 페스티벌'의 행사 주최자라면 2015년을 위해 어떤 점을 중시해야 할지를 정리해 보았다. 이는 데이터를 분석해 도출한 내용이며, 절대적인 사항이 아님을

밝혀 둔다.

앞서 밝혔듯이 행사 주관 사무국으로부터 좀 더 정제된 매출 데이터, 티켓 판매, 관람객 통계정보(연령, 남/여 비율, 거주지 등), 참여한 아티스트 정보, 시간별 공연 프로그램, 참여 스폰서 등 원천 데이터를 제공 받았다면, 빅파이팀의 분석 결과가 어떠했을지 아쉬움이 남는다. 더불어 정부 3.0을 활용한 공간 데이터와 각종 통계 데이터를 결합해 봤으면 하는 생각이 든다. 그럼에도 자라섬 재즈 페스티벌은 다음과 같은 사항을 참고할 필요가 있다.

- 더 관객 동원력이 높은 뮤지션들을 재초청하는 방안을 적극 검토한다.
- 가평과 자라섬의 인지도가 급등했지만, 춘천만큼 높지 않으므로 '자라섬·가평·재즈' 키워드의 지속적인 노출(인터넷을 통한 미디어)이 필요하다.
- 고객관리 측면에서는 지난 행사의 추억을 상기시키는 콘텐츠를 제작해 페이스북, 카카오톡 등 SNS를 통해 타깃 고객을 관리한다.
- 카드사 및 이동통신사와 제휴해 추가 매시업 데이터를 확보·분석해 데이터 분석 모델 도출 및 가성비 최적화 시뮬레이션을 주기적으로 실행한다. 더불어 데이터 분석 측면에서는 그동안 행사를 진행해 오면서 터득한 경험과 직관, 경제적 효과를 종속변수로 한 추가 분석 모델을 개발할 필요가 있다.

CONCLUSION

빅파이팀은 8주 간의 수료 프로젝트를 진행하면서 빅데이터 분석의 맛을 보기 시작했다. 본업을 하면서 시간을 할애해 낯선 분석 프로젝트를 한다는 것이 말처럼 쉽지 않았다. 어려운 여건에도 우리팀은 서로 신뢰하며 적극 참여했다고 자랑하고 싶다. 각자가 보유한 비즈니스 도메인을 존중했고 그 신뢰를 바탕으로 신뢰를 다졌다.

빅파이팀의 멘토인 전용준 박사께서 한 주도 빠지지 않고 오프라인 모임에 참석해서 많은 조언과 가르침을 주셨기에 프로젝트를 완료할 수 있었다. 페이스북 그룹에 빅파이팀이 참고할 만한 자료를 올려 주셨고, 많은 질문에 대해 LTE급 속도로 답변을 주셨다.

조그마한 성과지만, '자라섬 재즈 페스티벌' 데이터를 분석해 성공/실패 요인도 도출해 보고 내년 관람객 수까지 예측해 봤던 것은 정말 즐거운 '여정'이었다. 이 여정을 통해 데이터 분석에 대한 새로운 경험을 했다. 이런 경험이 팀원 각자의 소속회사에서 새로운 기회를 발견하는 데 일조할 수 있기를 기대해 본다.

마지막으로, 이 프로젝트를 더 발전시켜 여러 지역 축제에 적용 가능한 예측 모델로 내놓자고 다짐했다. 실제로 빅파이팀의 분석 모델은 응용 범위가 넓다고 생각한다. 매년 550개 이상 열리는 지역 축제에서 이 모델을 활용할 수 있기를 기대해 본다. ☺



구분	1	2	3	4
재즈 트랜드	10	20	50	70
예측 방문자 수	177,730	179,990	209,500	209,653
자라섬·가평·춘천 트랜드	10	20	50	70
예측 방문자 수	189,542	195,244	199,471	199,471
재즈·자라섬·가평·춘천 트랜드	10	20	50	70
예측 방문자 수	182,323.3	190,471.3	224,291.7	224,445

●<표 2> 2015년 자라섬 재즈 페스티벌 관람객 예측 결과

“남이 가지 않는 길에서 분석의 소중한 경험을 견지다”



안진훈 팀장
아이리치그린 대표

프로젝트 진행 중에 어려웠던 점은.

분석할 데이터 확보가 쉽지 않았다. 데이터를 갖고 있는 행사 주관단체인 ‘자라섬청소년재즈센터’로부터 티켓 판매, 관람객 수, 시간별 공연, 총 매출액 등의 데이터를 협조 받지 못했다. 더불어 분석 전문가 팀이었지만 분석도구인 R을 아주 잘 쓰는 사람이 없었다. 결국 시행 착오를 겪는 수밖에 없었다. 멘토인 정용준 박사의 도움을 많이 받았다.

데이터 확보가 어려운 주제를 끝까지 유지해야 할 이유가 있었나.

실제 비즈니스 환경에서 시장 조사가 절실했는데도 데이터를 확보하기 어렵다고 분석을 하지 않는가? 수료 프로젝트는 완성도 높은 분석 결과물 도출보다는 짧은 기간에 여러 경험을 하는 데 초점을 두고 있으므로 끝까지 해보기로 했다. 포기하지 않았던 이유 한 가지가 더 있었다. 국내에서는 전국적으로 연간 550여 개의 지역 축제가 이뤄지고 있다. 서울에서만 연간 50개 이상의 지역 축제가 개최되고 있음에도 결과를 분석해 놓은 사례를 찾기 어려웠다. 가장 우수한 축제가 어떤 형태로 개최되고 있는지 분석해 보면, 유사한 지역 축제를 준비하는 곳에 도움이 될 것이다.

어떻게 데이터 수집을 했나.

확보할 데이터를 정의하여 팀원이 나눠서 대부분 수작업으로 수집했다. 네이버와 구글 등 포털에서 재즈 음악에 대한 관심도 동향 데이터를, 유튜브에서 참여 아티스트의 동영상 조회수, 활동 경력, 선호국가 데이터 등을 확보했다. 한국도로공사에서 일자별 교통량 정보를, 기상청에서 날씨 데이터도 확보했다. 티켓 예매량 데이터는 인터파크와 YES24에서 얻을 수 있었다. 연도별 와인 수입량, 국내 총생산 증가율 등을 통계청에서 가져왔다. 그러나 보니 프로젝트 진행 기간의 1/3을 데이터 수집에 썼다. 무엇보다 빅데이터 분석에서 매시업(Mash-up)이 얼마나 중요한지, 다양하고 관련성 있는 원천 데이터를 찾아내는가가 얼마나 중요한지를 체험했다.

분석할 데이터가 부족해 중간에 위기를 겪었다고 들었다.

표본이 10개 미만인 데이터를 가지고 분석을 하는 게 과연 의미가 있을지를 놓고 망설였다. 이때 고려대 통계학과 허명희 교수께 ‘표본이 10개 정도로 하는 계량분석이 의미가 있을지’를 여겼다. ‘표본이 10 미만일 때는 예측의 변동성이 커지는, 즉 재현성은 감소하지만 충분히 분석이 가능하므로 의미가 있다’는 답을 얻었다. 결국 표본보다 변수가 큰 유형의 분석 과제로 접근했다.

분석 프로젝트를 끝낸 소감은.

컴퓨터 프로그래밍은 책이나 동영상 강의를 보고 혼자서 공부할 수 있을지 모르지만, 데이터 분석은 누구의 지도와 도움이 꼭 필요한 영역이다. 경험이 정말 중요한 영역이기 때문이다. 데이터 확보가 쉽지 않은 주제더라도 분석이 가능하다는 것을 몸소 체험했다는 데 의미를 두고 있다. 분석 경험이 많은 멘토로부터 지도를 받을 수 있다는 점이 정말 행운이었다. 특정 기관으로부터 축제 분석 의뢰를 받으면, 결과 보고서까지 낼 수 있겠다는 자신감이 들었다. 아울러 이번 빅데이터 분석 과정을 통해, 책에서 다루지 않는 현실의 문제에 대한 해결 경험이 데이터 분석에서 중요하다는 점을 알게 됐다.

자라섬 재즈 페스티벌 관람객 분석 및 예측

프로젝트 소개

한국의 대표적인 음악 축제 가운데 하나인 ‘자라섬 재즈 페스티벌’을 분석해 관객 확대 등 성공적인 행사로 발전하기 위한 조건을 도출했다.

구분

분석 전문가 과정: 군집 · 분류 · 회귀분석

프로젝트 기간

2014년 10~11월

멘토

전용준(리비전컨설팅 대표)

작용도구

R, MySQL, MS 엑셀

수집 데이터

티켓 판매, 초청 뮤지션 정보(글로벌 관심도 Top3 뮤지션 기준), 자라섬 재즈 페스티벌 트랜드, 날씨 · 교통 · 연도별 와인 수입량, 국민총생산 데이터

산출물

2015년 관람객 예측 및 성공 요소

교육 참여형태

자발적 참여(6) / 회사 권유(0)

진행

- 안진훈 팀장 데이터 분석 IT 컨설팅사 대표 경력 20년
- 한경훈 팀원 모델링, 데이터마트 공기업 전산실 개발자 경력 8년
- 신정호 모델링, 데이터마트 여론조사업체 개발자 경력 10년
- 전영준 데이터마트 교육출판사 경력 10년
- 류경숙 데이터마트 IT 강리업체 경력 12년
- 송창열 데이터마트 온라인게임사 고객 서비스 경력 5년

빅데이터 아카데미 수강 후 변화

- | | |
|---|---|
| <ul style="list-style-type: none"> • 분석 과정과 R 프로그래밍이 쉽지 않다 • 데이터 분석에서 어떤 가치를 찾을 수 있을까 • 통계 지식이 없는데도 데이터 분석을 할 수 있을까? | <ul style="list-style-type: none"> • 빅데이터에 대한 새로운 접근과 기준 경험을 융합해 개인 역량을 높일 수 있다. • 멘토를 통해 통계적인 지식이 실전에 어떻게 적용되는가를 알다. |
|---|---|



데이터에서 증권사 고객 마케팅의 답을 찾다



글 박종욱 (주)두산 C&SI 사업부 대리

빅데이터라는 의미와 사고의 한계를 벗어던지고 더 넓은 곳으로 나아갈 수 있는 팀이 되고자 하는 기원을 담아 팀 이름을 'ThinkBig'이라고 정했다. 시작이 반이라고들 한다. 하지만 이번 프로젝트는 그 시작이 쉽지 않았다. 기술 프로젝트를 진행할 때 가장 큰 어려움은 주제를 선정하는 부분이다. IT·문화·통신·제조·유통·소비자·금융·인프라·정부 등에 대한 수많은 아이디어가 쏟아져 나왔다. 하지만 '고양이 목에 방울 달기'라고 '그 데이터를 어디서 구해야 하느냐?'는 벽에 부딪혔다. 여러 주제들 가운데 프로젝트 기간 내에 데이터를 구할 수 있었던 것은 하나도 없어 보였다.

CHALLENGES

'고양이 목에 방울 달기'

운 좋게도 한 팀원이 근무하는 증권사의 협조로 빅데이터라부를 수 있을 만한 방대한 데이터를 구할 수 있게 되었다. 하지만 개인정보보호에 대한 중요성이 강조되는 상황이고, 특히나 일련의 금융 보안사고들 때문에 모두들 보안에 매우 민감하게 반응할 수밖에 없었다. 개인을 식별할 수 있을 만한 모든 데이터는 제거하여 분석에 적용했으며, 분석 인프라는 증권사 안의 서버를 사용해 데이터가 밖으로 유출되지 않도록 하였다. 이는 많은 부분에서 제약으로 작용했다. ThinkBig팀에서 당초 기대를 걸었던 가입자의 인구통계학적 정보 분석에 한계가 따랐고, 빠른 진행이 사실상 불가능에 가까워졌다.

보안이 생명인 증권사 데이터를 다루면서...

게다가 2주 간의 집체교육 기간에 밀렸던 업무들이 산더미처럼 불어나 팀원들을 짊어 삼킬 듯 기다리고 있었다. 프로젝트를 진행할 시간은 사실상 주말밖에 없었다. 시간이 날 때마다 팀들이 진행하자는 생각은 접근성의 제한으로 사실상 불가능했다. 미팅에서 한 번의 오판은 일주일의 노력을 허사로 만들 수 있을 만큼 치명적이었다. 그 실수를 만회하기 위해 또 다른 일주일을 써야 했다. 그렇기에 방향성을 잡기 위한 매주 정기 미팅은 밤 12시가 넘어서도 끝날 것 같지 않았다. 매번 가설을 세우고 지난 가설을 검증하며, 검증된 가설을 확대하는 작업의 반복이었다.

하지만 이런 위기의식은 오히려 팀을 하나로 모으고 진행에 신중을 더하는 방향으로 이어졌다. 만약 팀원 서로가 공감했던 위기의식이 없었다면 더욱 많은 시행착오가 따랐을 것이고, 결국 기한 내에 의미 있는 결과를 얻기는 힘들었을 것이다. 어쩌면 ThinkBig팀에게 보안에 따른 접근성의 제한이라는 위기는 오히려 좋은 결과를 얻을 수 있었던 기회가 아니었나 생각해 본다.

SOLUTION

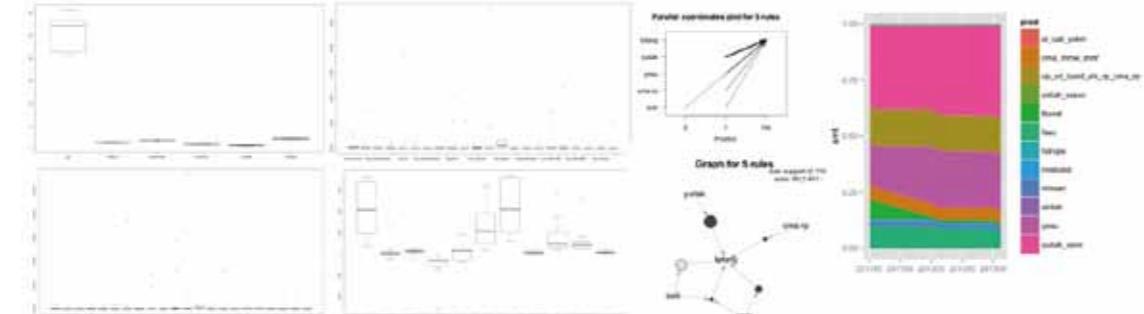
뼈대를 탄탄히 하다

집체교육이 끝나자 ThinkBig팀에게 주어진 임무는 평범한 과제가 아닌 하나의 프로젝트였다. 프로젝트 주제를 선정하고 WBS(Work Breakdown Structure)를 제출할 때까지만 하더라도 일반적인 프로젝트 관리 방법론만으로 충분할 것이라 생각했다. 하지만 막상 진행해 보니 당초 예상은 빗나갔다. 다른 프로젝트와는 달리 명확한 아웃풋 이미지가 없었으므로 일반적인 방법론을 바로 적용하기에는 일정관리가 쉽지 않았다. 따라서 기본적인 진행은 Agile/XP를 따라 1주 단위로 미팅을 통해 반복(Iteration)하는 것을 따르되, 세부적인 실행 내역에는 별도의 분석 방법론이 필요했다. 어렵게도 집체교육 때 배운 방법론은 분석 주제를 선정하는 비즈니스 모델링에 중점을 뒀기에, 현업 인터뷰가 불가능한 상황에서는 큰 의미를 가지지 못했다. 뿐만 아니라, 분석을 실행하고 시스템을 구축하는 실행 단계에 대한 명확한 가이드라인도 없었기에 방법론에 대한 폭넓은 조사가 필요했다. 현업 담당자들과의 인터뷰가 사실상 불가능했으므로 비즈니스 니즈에 따른 톱다운 방식의 방법론보다는 데이터 자체에 중점을 둔 바텀업 방식의 방법론이 필요했다. 따라서 데이터 분석에 특화된 SAS의 SEMMA 분석 방법론을 따르기로 결정하였다. SEMMA 방법론은 데이터로부터 어떻게 모델링을 이끌어내는 것에 집중되어 있다. 이 부분이 ThinkBig팀의 부족한 부분과 조금 더 맞는 느낌이었다.

프로젝트, 그 기나긴 여정**Sampling**

증권사로부터 받은 정보는 고객 마스터 데이터, 고객 포트폴리오(트랜잭션 정보), 기타 외부 정보, 이렇게 3가지로 구분할 수 있다.

고객 마스터 데이터에는 대표적으로 고객의 가치, 거래특성 정보 등을 담고 있으며, 고객 포트폴리오 정보는 고객별로 보유



● <그림 1> 데이터 탐색 과정에서 산출된 그래프들의 예

한 금융상품의 정보, 마지막으로 기타 정보에는 마케팅 캠페인 데이터, 고객별 채널 접촉 정보 및 종합 주가 지수 등의 정보가 담겨 있었다.

이 중에 가장 많은 비중을 차지하는 것은 고객별 월 평균 잔고였다. 증권사 DB는 각 트랜잭션 데이터를 모두 갖고 있었고, 그럼에도 그 정보를 그대로 전부 가져오지 않고 스쿱을 통해 ETL로 가져올 때, 고객·상품별 월 평균 잔고로 축약했다. 그럼에도 데이터 수가 1억 4000만 개가 넘었다.

Exploration

데이터를 가져와서 HDFS에 저장한 뒤에는 데이터의 속성을 파악하기 위해 모든 변수 각각에 대해 기초적인 분산분석을 포함해 상관성분석을 시행했다. 전체 과정 중에 어쩌면 가장 지루한 과정이 아니었나 싶다. 특히 데이터의 양이 많음에도 시스템이 3노드밖에 되지 않아 기초적인 분포를 보는 것만으로도 맵리듀스는 인내심을 시험하는 듯이 느렸다. 이 과정에서 나오는 그래프의 양도 상당했으므로 수많은 자료에서 무언가 인사이트를 잡아내기는 쉽지 않았다. 하지만 이 과정을 통해 각각의 데이터에 대한 기초적인 내용을 파악하고, 상식적으로 이해하기 힘든 이상치를 보정할 수 있었다.

이후 모든 변수에 대해서 서로간의 영향도를 파악하기 위해 상관성분석을 했다. 예상했던 것보다 서로간의 상관성이 높은 변수가 너무 많았다.

상관성이 높은 변수가 많다는 것은 분석적인 접근으로 봤을 때, 긍정적인 측면과 부정적인 측면 모두를 가지고 있다. 긍정

적인 측면으로는 변수 간에 어떠한 영향을 주고 받는지에 대한 개략적인 정보를 얻을 수 있고, 때로는 여기서 새로운 발견을 할 수 있다는 점이다. 하지만 부정적인 측면을 보면, 변수들끼리 서로 같이 움직인다는 의미다. 즉 분석의 입력 변수로서 모두를 같이 쓸 수 없다는 것으로 해석할 수도 있다. 상관성 분석을 통해 동일하게 움직이는 혹은 반대로 움직이는 변수들의 특성을 알 수 있었고, 고객의 성향을 분명히 갈라주는 요소를 확인할 수 있었다. 예를 들면, 주식과 채권에 투자하는 성향과 선물 옵션, 증권 위탁에 투자하는 성향은 서로 정반대의 특성을 가진다는 것이었다. 반면 주식과 채권끼리는 매우 유사한 고객의 투자 성향을 갖는 것으로 나타났다. 의외로 선물 옵션과 증권 위탁끼리 또한 유사한 투자 성향을 갖는 것을 알 수 있었다(상관성 90% 이상).

Modification

길고 험난한 데이터 탐색의 과정이 어느 정도 진행되자, 해당 변수만으로는 바로 분석에 활용할 수 없었다. 뿐만 아니라 변수들을 가공해 분석에 용이하도록 파생변수를 생성해야 할 필요가 생겼다.

여러 종류의 파생변수를 생성했다. 대표적으로는 총 거래일, 마케팅 휴면일, 이탈 소요기간, 최고자산 이탈 소요일, 최고자산 소요기간, 포트폴리오 비율 등과 같이 기존에 갖고 있는 데이터의 변수들을 비즈니스적으로 의미가 있으면서 결과 해석에 용이하도록 일정한 수식에 의해 생성해 분석에 적용했다.

이러한 파생변수들 또한 기초적인 분포분석 및 상관성 분석을

진행했다.

Modeling

이번 프로젝트의 결과는 크게 두 가지로 나눌 수 있다. 첫째는 고객 특성별 세그먼테이션에 따른 프로파일링이고, 둘째는 고객별 맞춤 상품 추천이다.

세그먼트를 정할 때는 특정한 Cut-off 요소를 찾는 방법을 적용했다. 비즈니스적으로 의미 있다고 파악되는 특정한 Cut-off 요소를 찾는 방법이다. 하지만 팀 입장에서는 분포분석을 통해 도출된 결과를 하나하나 비즈니스와 커뮤니케이션할 수 없었다. 그래서 수학적으로 합리적이라 생각되는 세그먼테이션 방법으로서, 클러스터링(K-Means) 알고리즘을 적용했다. K-Means 알고리즘을 사용하면, K의 값에 따라 나눠지는 군집의 개수와 특성에 영향을 많이 미치게 된다. 따라서 K 값을 점차 올리면서 Iteration 작업을 반복해 K에 따른 Sum of squares 값을 확인했다. 간략히 설명하면, Sum of squares 가 나타내는 것은 군집끼리의 Centroid(군집의 중간값)가 서로 얼마나 떨어져 있는지를 보여주는 값이라고 볼 수 있다. 군집의 개수를 늘렸을 때, Centroid가 일정 이상 가까워지거나, 혹은 갑자기 가까워지는 변곡점이 있으며 해당 군집의 개수를 활용하는 것이 가장 합리적이라 판단했다.

고객별 맞춤 상품을 추천하기 위해 Recommendation (Collaborative Filtering) 알고리즘을 사용했다.

상품을 추천할 때 전체 고객의 정보를 그대로 이용하는 방법도 있지만, 이전에 산출된 군집

별로 고객의 특성이 서로 다르다고 파악해 동일 군집 내에서의 추천 알고리즘을 사용했다. Collaborative Filtering을 계산하기 위해서는 고객의 상품 개입 매트릭스와 상품들 간의 상관성 매트릭스가 필요하다. 이 중에 문제가 되었던 부분은 상품들 간 상관성 매트릭스를 어떻게 구성할 것인가였다. 이에

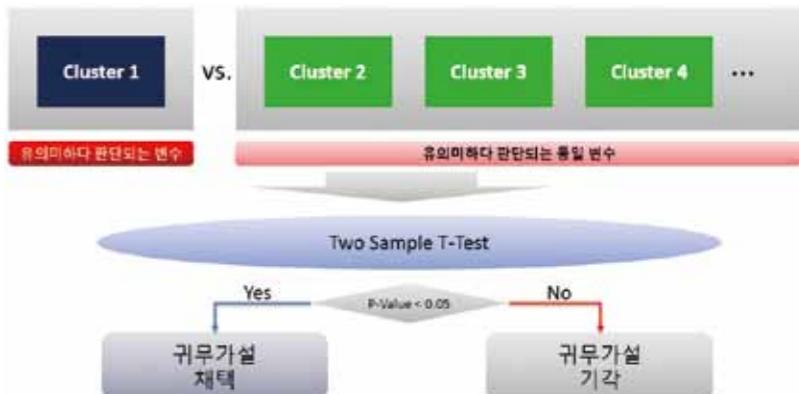
대한 많은 논의가 있었지만, 결국 머하웃(Mahout)에서 제공하는 모든 종류의 상관성 매트릭스를 다 구현했다. 따라서 Log Likelihood, Pearson, Tanimoto, Spearman, Euclidean Distance 모두를 다 사용해 5위까지의 결과를 도출했다.

Assessment

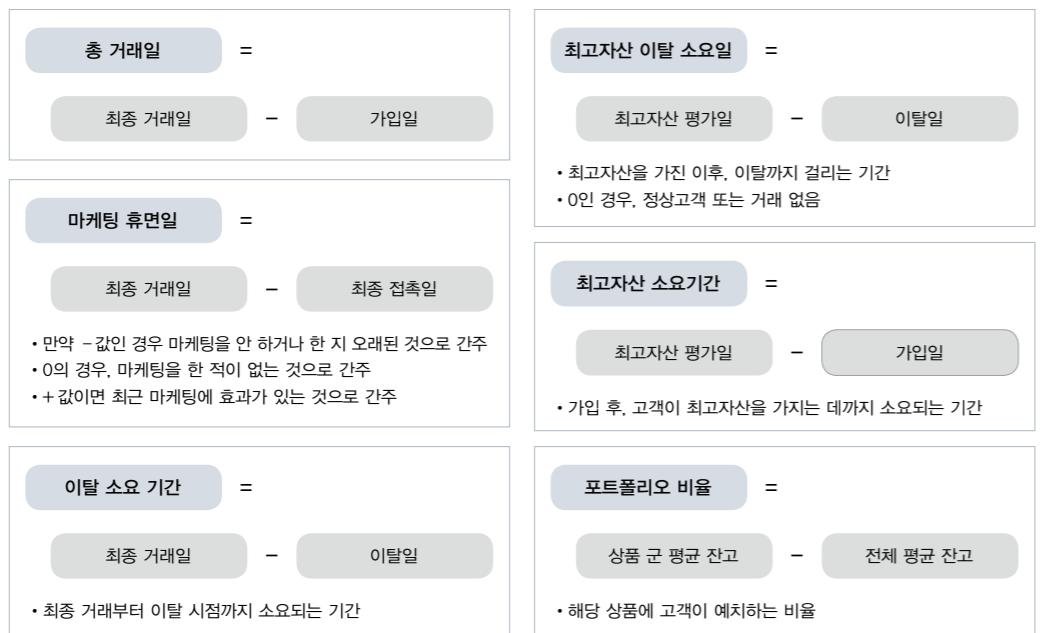
클러스터링을 통해 도출된 군집들은, 그 자체로는 별 의미가 없었다. 단순히 나눠놓은 것에 불과했으므로 이를 비즈니스적으로 재정의하고 각 군집의 특성을 프로파일링하는 것이 중요했다. 결국 ‘핵심 고객군’, ‘관리 사각지대 고객군’, ‘장려 고객군’이라는 3개 축으로 재정의했다. 해당 군집의 대표적 특성이라 보이는 점들을 해당 클러스터와 다른 클러스터 간 귀무가설을 토대로 T-검정을 시행했다. Two Sample T-Test 를 통해 P-Value가 0.05 미만으로 나온다면, 해당 귀무가설을 채택해 분석결과가 의미가 있음을 검증했다.

기술 프로젝트가 맞나요?

빅데이터 기술전문가 과정이라 해서 오로지 ‘기술’ 부분에만 집중했다면 결국 다른 사람들에게 어떤 의미로 다가설 수 있었을까? ThinkBig팀에게 프로젝트라는 것은 단순히 지금까지 배운 것을 확인하는 숙제로서의 의미뿐 아니라, 이를 뛰어넘어 ‘가치’와 ‘재미’를 담아 전달할 수 있는 결과물을 만드는 과정이었다. 어떻게 하면 프로젝트의 결과가 사람들에게 통찰력을 줄 수 있는가를 놓고 끊임 없이 고민했고, 메시지를 전달



● <그림 2> 클러스터별 특성 검증 과정



●<그림 3> 데이터의 특성을 파악하여 비즈니스적으로 의미 있는 파생변수의 도출

할 수 있는 스토리를 만들기 위해 노력했다.

필요에 따라 기술적인 요소를 적용하다 보니, ThinkBig팀이 적용한 기술이 어떤 것인지를 표현하기 어려웠다.

데이터 수집 계층(Data Gathering Layer, DGL)은 분석할 데이터를 가져와서 데이터 스토리지 계층(Data Storage Layer, DSL)으로 넘겨주는 논리 층이다. ThinkBig팀의 프로젝트에서는 증권사의 데이터를 가져오는 부분이다. 증권사 DB로부터 ETL 커넥션을 맺고 데이터를 지속적으로 끌어와서(Pulling) 스토리지에 넘겨 줄 수 있는 Sqoop을 사용했다. 이번 프로젝트를 위해서라고만 생각한다면, Sqoop을 활용해 주기적인 데이터 수집 절차를 거칠 필요는 없었을 것이다. 하지만 PoC(Proof of Concept) 개념으로, 이번 프로젝트 결과를 실제 비즈니스에 활용했을 때의 가치를 증명하고 싶었다. Sqoop은 비즈니스에서도 충분히 활용할 수 있다는 결론을 내렸다.

데이터 스토리지 계층은 데이터를 실제로 저장하는 논리 계층이다. ThinkBig팀은 HDFS(Hadoop Distribution File

System)로 구성했다. HDFS는 대용량 파일을 블록 단위로 쪼개어 저장하고, 각 블록을 여러 노드에 분산·중첩해 저장한다. 따라서 비즈니스 환경에서 폴트 툴러런스와 스케일아웃 기능을 제공할 수 있다.

데이터 프로세싱 계층(Data Processing Layer, DPS)은 DSL에 저장된 데이터를 필터링·요약·정렬·가공을 통한 파생 변수 생성 등의 과정을 거쳐 분석용 데이터 마트를 구성하는 논리 계층이다. ThinkBig팀은 맵리듀스와 하이브를 활용해 구현했다. 맵리듀스는 구글에서 발표한 논문을 기초로, 매플 작업과 리듀싱 작업을 활용해 각 노드 사이에서 분산작업을 통해 데이터를 분석하는 방법론이라 할 수 있다. HIVE는 HiveQL이라는 SQL과 비슷한 구문을 통해 맵리듀스를 직접 코딩하지 않고도 활용할 수 있도록 고안된 아키텍처다. 따라서 ThinkBig팀은 기존 맵리듀스만으로는 개발시간이 너무 길고 휴면에려가 높을 가능성이 있어, 복잡하고 다양한 1회성 검증작업 등에 HIVE를 많이 사용했다.

데이터 분석 계층(Data Analysis Layer, DAL)은 DPS에서

구성된 데이터 마트를 활용해 여러 분석 모델링을 실제 수행하여 결과를 얻어내는 논리 계층이다. ThinkBig팀은 머하웃과 R을 활용했다. 머하웃은 각종 머신 러닝에 해당하는 알고리즘의 집합체로서 분산처리에 적합한 환경을 제공한다. 특히 하둡 위에서 맵리듀스를 통해 분석할 수 있는 장점이 있다. ThinkBig팀은 머하웃을 활용해 고객 거래 특성을 토대로 고객 세그먼테이션을 위한 클러스터링(k-means)과, 고객에게 금융상품을 추천하기 위한 추천(Recommendation, Collaborative Filtering) 알고리즘을 사용했다. 또한 R을 활용해 소규모 마스터 데이터 탐색과 분포분석을 빠르게 진행할 수 있었다.

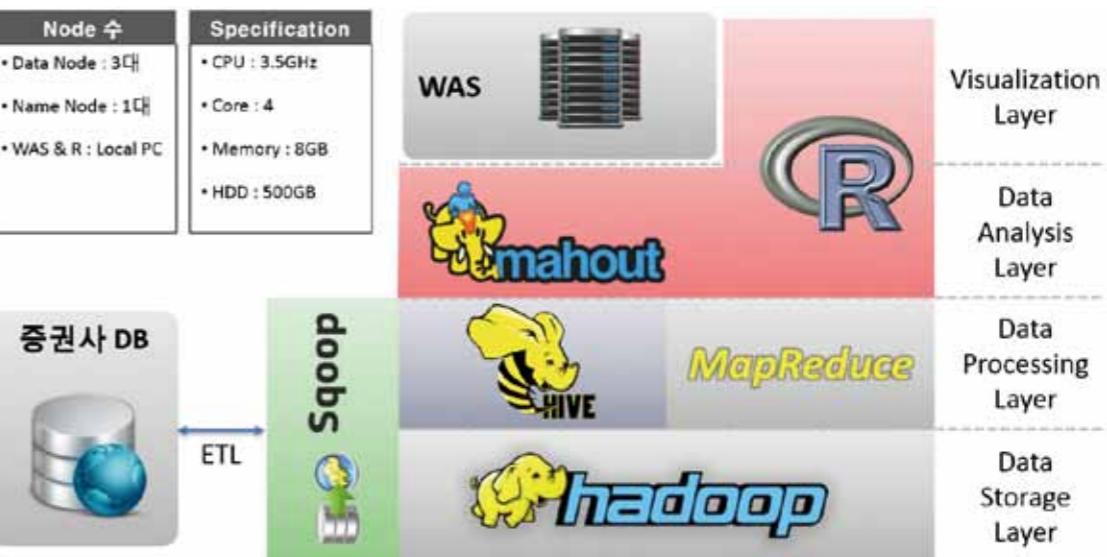
시각화 계층(Visualization Layer, VL)은 분석한 결과를 사용자가 직관적으로 이해할 수 있도록 시각화하는 논리 계층이다. 데이터 탐색과 분포분석 결과를 R로 확인했다. 사용자 클러스터와 추천 결과는 WAS(Web Application Server)를 통해 D3를 활용해 웹으로 시각화했다.

안전과 모험 사이에서

하둡 클러스터를 구축하고 그 안에 데이터를 밀어 넣는 것은

예상보다 빠르게 완료됐다. 시스템을 구축하고 데이터까지 확보되자, 분석 방법론대로 각 데이터의 분포부터 확인하는 길고 긴 데이터 탐색이 진행됐다. 이를 통해 이상치나 결측치를 제거하고 데이터에 대한 이해를 하는 작업까지는 다소 시간이 걸리더라도 순조롭게 진행되는 듯 하였다. 하지만 문제는 그 이후부터였다. 각 변수들의 상관성 분석까지 완료하고 나자, 적용할 만한 분석 모델을 바로 찾을 수가 없었다. 모두들 머리를 쥐어짜며 고민해 보았지만 시간은 하염없이 흘러가기만 했다. 결국 이러한 상황에서 팀의 분위기가 흐트러지기 시작했다.

결국 두 의견으로 엇갈리기 시작했다. 하나는 '일단 결과를 빨리 얻을 수 있는 추천분석에 집중하여 함께 완료한 뒤에 다른 주제거리를 찾자'는 의견이었다. 다른 하나는 '추천분석이라는 것 자체가 줄 수 있는 통찰(Insight)은 이미 예상되는 부분이고, 그러한 결과는 우리가 애초에 얻고자 했던 의미 있는 것이라고 볼 수 없으므로 다른 결과에 집중하는 것이 옳다'는 것이었다. 어쩌면 이는 무엇이든 안정된 결과를 하나 얻고자 하는 '안전'과 더욱 값진 결과를 얻기 위한 '모험' 사이의 갈등이었을지 모른다.



●<그림 4> ThinkBig 팀의 증권분석 시스템 아키텍처

하지만 무엇 하나만 추구하게 되면 그 자체만으로 '리스크'라고 생각했다. 밤새 갑론을박을 거쳐 역할분담을 하여 어느 정도는 정리되는 듯 보였다. 기술에 자신이 있었던 팀원들을 모아 머하웃을 통해 추천분석을 완성하기로 하고, 나머지 팀원들은 데이터를 계속 들여다 보면서 새로운 분석 주제를 찾기로 했다. 한 주가 지나자 모든 것이 원점이었다. 고객 대다수가 매우 적은 수의 상품만을 가입하였기에 추천분석의 예측치가 너무 낮아 바로 사용하기 힘들었다.

새로운 분석 주제를 찾기로 한 팀원들은 클러스터링을 거쳐 Iteration해 합리적이라 생각되는 12개나 되는 군집을 산출해 가져왔지만, 너무 많은 군집들 때문에 프로파일링을 통해 의미 있는 결과를 도출해내기가 쉬워 보이지 않았다. 하지만 계속하여 새로운 가설을 세우고 이를 검증하는 작업을 수 없이 반복한 결과 (소위 삽질이라 부르는) 시행착오 데이터가 쌓이게 되었고, 결국 분석된 결과들을 하나로 모아서 정리할 수 있음을 만한 스토리가 탄생했다.

만약, 안전과 모험 중에서 어느 하나만을 택하는 쉬운 길을 택한다면, 어쩌면 ThinkBig팀은 프로젝트 기간 추천분석이나 군집분석 하나의 결과만을 가지고 실제로 비즈니스에서 활용할 만한 애플리케이션을 도출하지 못했을지 모른다. 하지만 개인이 아닌 팀이라는 특성을 활용한 적절한 분업과 팀원들의 열정은, 두 가지 주제를 함께 해결할 수 있도록 하는 원동력이었다.

고객 세분화로 마케팅 최적화

데이터를 분석해 증권사 고객의 특성을 파악한 다음, 세분화한 고객별 프로파일을 바탕으로 최적의 마케팅 방향을 설정해 보기로 했다.

높은 연령대의 고객

가장 먼저 찾았던 내용은 증권사 고객의 연령 분포였다. 관계자로부터 들었던 고객들의 연령이 실제 데이터가 말해주는 것과는 달랐다. 경제활동이 활발한 30대 후반에서 40대가 많았음으로 예상했다. 분석결과 50대 중후반의 고객이 가장 많은 것을 확인할 수 있었다. 이를 바탕으로 실질적으로 경제활동

이 매우 활발한 30~50대 초반의 고객을 유치할 만한 대비책 마련이 매우 시급해 보였다.

고객군별 프로파일

거래특성별 클러스터링을 통해 도출된 7개의 군집을 비즈니스 적으로 이해할 수 있도록 다시 4개의 군집으로 묶었다.

첫 번째 군집은 전체 고객 중 0.25%를 차지하는 '핵심 고객군'이다. 증권사와 접촉이 매우 잦은 편이며, 거래가 매우 활발한 특성을 갖는다. 또한 포트폴리오 구성을 볼 때, 어느 한 쪽에 치우치지 않고 다양한 금융상품에 고르게 분산 투자하는 특성을 보인다. 안전 자산에 대한 선호도가 매우 높다. 전문 투자자문 서비스를 받고 있는 재력가이거나, 주로 자신의 증권을 위탁하는 기업 오너로 세부적으로 구분할 수 있다. 놀랄 만한 사실은 매우 적은 고객군임에도 고객의 잔고가 증권사 전체의 93%를 차지하고 있다는 점이다. 증권사의 수익 비중의 97% 정도를 차지하는 매우 중요한 고객군이기도 하다.

핵심 고객의 자위를 유지할 수 있도록 VIP 관리 서비스를 발굴 및 제공하는 것이 중요하다는 결론을 내렸다.

두 번째 군집은 전체 고객 중 5%를 차지하는 '장려 고객군'이다. 평균 거래 기간이 15년 정도로 길고, 최종 거래 경과일 수가 1달 이내로 대체로 거래가 활발한 고객군이다. 이 고객군은 대부분 금융사 등급이 높고, 우수고객으로 관리대상에 등록돼 있다. 안정 또는 위험 중립형의 거래 특성을 보이며, 전체 잔고 비중 6% 정도로 핵심 고객군 이외에 가장 견실한 고객군이라 볼 수 있다. 해당 고객군에게는 종합자산관리 마케팅 등을 통해 포트폴리오 다양화를 유도하고, 고객의 자산을 증권사의 상품으로 옮길 수 있는 방안 수립이 필요하다고 판단했다.

세 번째 군집은 전체 고객 중 8% 정도를 차지하는 '관리 사각지대 고객군'이다. 평균 거래기간이 17년 이상으로 충성도가 높음에도 평균 8개월 정도의 낮은 접촉 빈도를 가진다. 다른 군집은 대부분 1개월 내에서 지속적인 접촉이 이루어지는 반면, 거의 관리를 받지 못하는 고객으로 볼 수 있다. 포트폴리오 구성은 주로 주식관련 비중이 높아, 다소 공격적인 성향을 가진 유형으로 분석된다. 고객의 잔고는 0.14%를 차지하

며, 수익의 비중은 0.02% 정도로 미미하다. 이 고객군에게는 주식수익률대회 등 이벤트를 개최해 HTS나 MTS 등의 채널로 유도할 수 있는 마케팅 활동이 시급해 보였다. 또한 지속적인 접촉을 통한 활발한 거래를 유도하는 것도 중요하다는 판단을 내렸다.

마지막 군집은 87%로 거의 대부분의 고객을 차지하고 있는 일반/불량 고객군이다. 증권사와의 접촉이 매우 잦음에도 불구하고 거래가 거의 발생하지 않는다. 많은 분포도에 비해 뚜렷한 거래 성향을 보이지 않으며, 잔고는 1% 정도를 차지하고 증권사 수익에 0.2% 정도밖에 기여를 하지 않는다.

고객군별 프로파일링 결과, 해당 증권사는 일반 소규모 고객보다는 자산이 많은 우수 고객에 집중하여 영업하는 것을 확인할 수 있었다. 하지만 수익의 대부분이 다수의 고객이 아닌 일부 특수한 고객으로부터 나오는 점은, 소규모 고객의 이탈로 기업의 기반이 흔들릴 수 있으므로 다소 위험한 부분이다.

고객별 맞춤 상품 추천

다음 링크의 두 영상을 보자. 상품 추천은 고객의 번호를 입력하여 해당 고객에게 맞는 상품을 추천하는 고객별 상품추천과, 포트폴리오를 입력하면 (혹은 그렇게 가입을 원하는 고객이 있다고 한다면) 입력된 데이터를 바탕으로 상품을 추천하는 고객 포트폴리오별 상품 추천, 이렇게 두 가지로 구분돼 있다.

- <http://www.youtube.com/v/Z9jbzLUtdiE>
- <http://www.youtube.com/v/Fkf3TXuk1Gw>

CONCLUSION

프로젝트 중반쯤에 팀장이었던 필자에게 예상하지 못한 일이 생겼다. 필자의 아들이 세상이 너무나 궁금했는지 8주나 일찍 태어난 것이었다. 축복 받을 일이었음에도 인큐베이터에 안쓰럽게 들어가 있는 아이를 보니 제정신이 아니었다. 이 일 때문에 필자 주변의 모든 일이 멈춰버린 듯 했다. 당시 회사에서

중요한 프로젝트가 시작 단계에 있었고, 빅데이터 아카데미 수료 프로젝트에서도 팀장으로서 역할까지 해야 하는 상황이었는데, 모든 걸 놓아버리니 죄인이 된 기분이 들었다. 하지만 양해를 구할 때 모두들 너무나 당연한 듯 필자의 편의를 봐주는 모습이 정말 큰 위로가 되었다.

아무런 신경도 쓰지 못한 2주간의 공백 이후에 다시 모였을 때, 필자는 정말 깜짝 놀랐다. 생각보다 많은 부분이 진행되어서 오히려 따라잡기 힘들 정도였기 때문이다. 모두들 각자의 분야에서 전문가였기에, 확실한 책임배분(Role and Responsibility)을 토대로 묵묵히 자신의 역할에 최선을 다하고 있었던 것이다. 그 순간, ThinkBig팀의 가장 큰 강점을 발견할 수 있었다. 각자 시계 부품처럼 중요한 역할을 가지고 유기적으로 움직이지만, 공통된 목표를 바탕으로 서로의 빙자리마저 채울 수 있는 힘이 있었던 것이다.

다양함과 조화

필자가 생각하는 강한 조직이란 동일한 전문가로만 구성된 집단이 아니라, 다양한 사람이 모여서 상보적인 역할을 수행하는 조직이다. 이번 기술과정 5기에는 비슷한 사람들이 모여서 팀을 이룬 곳도 있었고, ThinkBig팀처럼 서로 다른 조직과 배경을 가진 사람이 모여서 팀을 이룬 곳도 있었다. 팀마다 각자의 강점을 갖고 있겠지만, ThinkBig팀이 강할 수 있었던 가장 큰 이유는 교육·기술·제조·금융·비즈니스 컨설팅 등을 아우르는 팀원들의 다양한 경험과 지식일 것이다. 그렇기에 한 사람의 시각으로 볼 수 없는 다양한 접근과 여러 분야에서의 검증, 서로간의 명확한 책임 배분이 이번 프로젝트를 성공적으로 이끌 수 있었다.

끝으로 고생했던 팀원들과 여러 조언을 아끼지 않은 멘토, 열정적으로 강의해주신 강사들, 빅데이터 기술 전문가 과정을 운영하신 KODB 관계자들께 감사의 말씀을 드린다. ☺



“분석 플랫폼의 힘을 실감하다”



박종욱 팀장
두산C&SI사업부 대리

교육 과정에서 가장 기억에 남은 것은.

열심히 해도 결과가 나오지 않을 때였다. H증권사로부터 받은 데이터로 프로젝트를 진행했는데 말 그대로 ‘빅데이터’이다 보니 해석조차 힘들었다. ‘제공한 데이터를 분석한 결과가 이런 의미 있는 결과가 나왔다’고 말할 수 있어야 하는데, 및 미한 결과만 나왔을 때, 그 심정… 솔직한 표현으로 ‘삽질’과 ‘막노동’을 엄청 했다(웃음). 어려움이 닥칠 때마다 팀원들은 ‘우리가 해낼 수 있을까?’ 하는 두려움에 휩싸였는데, 내가 배짱을 타고났는지 할 수 있겠다는 생각이 들었다. 그때 마다 ‘할 수 있다’고 말했다. 과제 제출 완료를 일주일 남겨 놨을 때다. 팀원들이 엄청나게 많이 도출해 놓은 결과물을 어떻게 정리해야 할지를 놓고 암담해 하고 있었다. 소위 ‘있어 보이지’ 않았다. 하지만 이것도 일주일 만에 해결됐다. 프로젝트 시작 전에 스토리라인을 짜뒀던 것을 활용했던 것이 큰 도움이 됐다. 물론 스토리라인은 중간 중간에 계속 수정됐다.

데이터 분석과 관련한 일을 하고 있나.

비슷한 일을 하고 있다. 컨설팅 업무를 하다 보니, 프로젝트 방법론은 낫설지 않았다. BI(Business Intelligence) 시절부터 데이터 분석 업무를 해와서 R이나 SAS 등 분석 도구 사용도 익숙한 편이다. 수료 프로젝트를 하다 보니 빅데이터를 BI라고 할 수 없지만, BI의 하나로서 볼 수 있겠다는 생각이 들었다.

실제 프로젝트를 진행해 보면서 느낀 점은.

수료 프로젝트는 비교적 쉽게(?) 확보한 데이터가 있었으므로 분석 인프라를 구성하여 결과만 도출하면 됐다. 프로젝트를 한번 경험하고 나니, 데이터 수집 과정부터 하나씩 직접 해 보면 비즈니스 아이템으로 확대할 수 있겠구나 하는 자신감이 들었다.

시각화도 완성도가 높다는 평가를 받았다.

멘토로부터 “모든 면에서 뛰어났다”고 들었다(웃음). 기술에 강한 팀원, 프로젝트 방법론과 프레임워크 구성 등 프로젝트 관리에 경험이 많은 팀원, 시각화에 강한 팀원 등 다양한 장점들이 어우러지면서 좋은 결과를 도출했다. 시각화는 D3 패키지를 썼다.

빅데이터 아카데미에 바라는 점은.

빅데이터 아카데미의 최대 장점은 ‘수료 프로젝트’라고 생각 한다. 수료 프로젝트의 의의를 완벽한 결과물 도출보다 분명 한 콘셉트 확보로 봤을 때, 8주라는 기간은 결코 짧지 않다. 이 기간 동안 같은 목적을 갖고 모인 사람들이 함께 프로젝트를 진행해 보는 건 정말 돈 주고 살 수 없는 경험이었다. 회사 팀원들이 외부 빅데이터 교육을 다녀와서 서로 정보를 주고 받는데, 교재나 지원·커리큘럼 측면에서 빅데이터 아카데미가 가장 체계적이라는 평이었다. 아쉬운 점은 교육장의 PC를 비롯해 VPN으로 접근해야 하는 가상서버(VM) 환경이 수료 프로젝트를 진행하기에는 성능이 약한 느낌이었다.

향후 계획은.

수료 프로젝트 팀이 힘을 합쳐 비즈니스 모델로 연결시키는 것을 놓고 의견을 모으고 있다. 데이터와 관련된 영역에서 발전의 기회를 찾고 싶다. 빅데이터의 실제를 경험하지 못했다면, 직접 한번 경험하고 나면 많은 영감이 떠오를 것이라고 확신한다. ☺

증권사 빅데이터를 활용한 고객 패턴분석

프로젝트 소개

1억 4000만 건이 넘는 증권사 고객 정보 데이터를 빅데이터 처리 인프라를 구축해 분석한 프로젝트. 고객군을 세분화해 고객군별 프로파일을 바탕으로 최적의 마케팅 방향을 설정하기 위한 데이터를 도출했다.

구분

기술 전문가 과정

프로젝트 기간

2014년 04~05월

멘토

이상훈(SK C&C 대리)

적용도구

- 시스템 계층: 데이터 노드 3대, 네임 노드 1대(3.5GHz CPU, 4-Core, 8GB RAM), WAS, R
- 스토리지 계층: 하둡
- 프로세싱 계층: 스쿱, 하이브, 맵리듀스
- 분석 계층: 머하웃, R

수집 데이터

1억 4000만 건의 증권사 고객 정보 데이터

산출물

- 고객별 맞춤 상품 추천 시스템
- 고객별 상품 추천
- 고객 포트폴리오별 상품 추천

교육 참여형태

자발적 참여(5) / 회사 권유(0)

진행

• 박종욱 팀장	SI 업체	PM, 분석 모델링, 프로파일링	경력 4년
• 김영민 팀원	증권사 IT부	하둡 클러스터	경력 9년
• 안병현	증권사 IT부	데이터 관리, SQL 작성, ETL	경력 15년
• 양승영	카드VAN사 IT부	WAS 개발, 머하웃 커넥터	경력 7년

빅데이터 아카데미

수강 후 변화

- | | |
|------|---|
| 수강 전 | <ul style="list-style-type: none"> • 빅데이터는 하둡으로만 처리할 수 있다. • 빅데이터 분석은 매우 어렵다. • 빅데이터는 먼 곳에 있다. • 빅데이터 처리를 위한 다양한 기술이 있다. • 도메인에 대한 이해만 있다면 얼마든지 도전할 수 있다. • 빅데이터는 생각보다 가까운 곳에 있다. |
| 수강 후 | |

원도우 서버 감사 로그 분석 시스템

빅데이터 처리 기술로 원도우 서버를 지켜라!



글 정하권 안랩 선임

서버 OS 시장에서 매우 높은 점유율을 자랑하는 마이크로소프트 원도우 운영체제. 널리 사용되는 운영체제인 만큼 원도우에는 수많은 보안 위협들이 존재한다. 현재 많은 기업이 이러한 보안 위협을 피하기 위해 적지 않은 비용을 투자한다. 그렇다면 빅데이터 처리 기술을 응용하여 더 현실적인 비용으로 원도우 운영체제의 허점을 이용한 외부 공격을 예방하고 침입을 탐지할 수 없을까? 빅데이터 기술 전문가 6기 원도우 감사 로그 분석팀은 이러한 문제를 해결하기 위해 ‘원도우 서버 감사 로그 분석’이라는 프로젝트에 도전했다.

CHALLENGES

어느 팀이나 프로젝트를 시작할 때는 ‘우리가 과연 할 수 있을까?’ 하는 두려움과 ‘최우수상을 타고 싶다’는 바람이 교차했을 것이다. 기술 전문가 집체교육이 중간 정도 진행되었을 쯤, 수료 프로젝트팀을 구성하라고 했다. 낯선 사람들이 모여서 어색한 인사를 나누던 때가 엊그제 같은데, 얼마 지나지 않아 서로를 이해하고 배려하는 단합된 팀으로 거듭났다. 감사 로그팀(이하 ‘우리팀’으로 통일)이 좋은 결과물을 낼 수 있었던 배경은 구성원 간 대화와 배려, 책임감, 열정이 잘 결합되었기에 가능한 일이었다.

원도우 보안로그 분석을 주제로 선정

학위 논문을 쓸 때, 주제 선정이 절반이라는 말이 있듯이, 빅데이터 아카데미의 수료 프로젝트의 주제 선정도 쉽지 않았다. 두 달 남짓한 기간에 그것도 직장 생활을 하면서 수행해야 했으므로 ‘과연 무엇을 얼마나 잘할 수 있을까’ 하는 고민을 계속 할 수밖에 없었다. 팀원 모두 한 개 이상의 주제를 갖고 토론의 시간을 가졌다. 데이터 처리 주제부터 분석 전문가 과정의 프로젝트에 더 어울릴 법한 다양하고 폭넓은 주제가 제안됐다. 다양한 의견이 쏟아 나오자 자칫 팀이 의도하지 않은 방향으로 흘러갈 수 있겠다는 생각이 들었다. 이에 팀원들은 프로젝트 주제의 선정 기준을 다음과 같이 더 분명히 하기로 했다.

- ❶ 집체교육에서 배운 빅데이터 처리 기술을 응용해 볼 수 있는 주제여야 한다.
- ❷ 제한된 일정과 리소스로 완성도 높은 결과물을 도출할 수 있어야 한다.

이 기준에 부합하는 주제 2개를 선정한 다음, 이 가운데 ‘운영체제(Linux, Windows) 보안로그 분석’을 최종 주제로 선정했다. 팀원 가운데 보안 전문업체에서 일하는 도메인 전문가가 있었으므로 제한된 일정에도 빠른 기획과 설계가 가능한 점과, 빅데이터 처리 기술을 다양하게 응용해 볼 수 있다는 점에서 최종 주제로 선정했다. 운영체제는 원도우와 리눅스를 놓고 논의를 거쳐 원도우를 최종적으로 선택했다. 원도우는 많은 보안 취약점이 존재하고, 다양한 로그 수집과 분석이 가능하기 때문이었다.

프로젝트 목표를 달성하기 위하여

빅데이터 기술 전문가 교육 마지막 날, 최종 결정된 팀 프로젝트 주제를 다음과 요약해 제출했다.

목표	원도우 운영체제의 감사 로그를 수집·통합·분석해 보안 위협 요소를 사전에 발견·대응할 수 있게 한다.
대상	원도우를 운영하는 모든 고객
가정	<ul style="list-style-type: none"> 많은 업체에서 주요 IT 서비스 플랫폼으로 원도우 운영체제를 사용한다. 그럼에도 원도우의 감사 이벤트 로그들이 관리되고 있지 않다.
조건	<ul style="list-style-type: none"> 예산 범위: 분석 시스템 4대 기간: 8주 수행 인원: 5명
시스템 구성	<ul style="list-style-type: none"> 로그 수집: 로그스타시, 플럼, Fluentd 중 선택 로그 검색과 저장: 엘라스틱서치, 하둡 중 택일 실시간 처리: 스톰
효과	<ul style="list-style-type: none"> IT 커뮤니케이션 차원에서 IT 시스템 로그 통합관리 계정 및 개체 접근 감사를 통한 내부 정보유출 위협 예방 로그온 및 시스템 이벤트를 통한 시스템 침입 위협 예방 정책 변경 감사 관리(권한, 방화벽 등)

● 표 1) 프로젝트 요약

SOLUTION

프로젝트에 들어가기에 앞서 프로젝트 시스템 기획과 설계 미팅을 가졌다. 집체교육이 끝나자 공용준 멘토를 중심으로 모든 팀원이 모여 첫 번째 미팅을 가졌다. 3시간 가까이 열띤 토론 끝에 프로젝트 시스템 구조와 팀원별 역할 및 진행 일정을 정리했다.

가장 적합한 도구를 찾아라!

무언가를 새로 시작하려고 할 때, 자신이 사용해 본 것을 먼

	로그스태시	이벤트와 로그를 관리하기 위한 도구. 검색을 비롯해 다양한 용도로 사용하기 위해 로그를 수집·저장·분석하는 데 사용된다. 아파치 2.0 라이선스를 사용하는 완전 무료 오픈소스이며, 현재는 엘라스틱서치 제품군의 일부가 되었다.
	엘라스틱서치	Shay Banon이 루신(Lucene)을 바탕으로 개발한 분산 검색엔진. 설치와 서버 확장이 매우 편리한 것이 장점이며, 탁월한 검색기능을 갖고 있다. 분산 시스템이므로 검색 대상 용량이 늘어났을 때, 유연하게 대응할 수 있는 것이 장점이다.
	스톰	분산 환경에서 사용할 수 있는 실시간 빅데이터 분석 플랫폼. 분산 환경에서 카assandra(Cassandra), 카프카(Kafka), 레디스 같은 다양한 데이터베이스와 연동할 수 있고, 다양한 프로그램 언어로 실시간 데이터 분석 알고리즘을 적용할 수 있다. 프로그램 개발자에게 다양한 빅데이터 분석의 장에 참여할 수 있도록 유도하고 있다.
	레디스	오픈소스로 BSD 라이선스를 사용하며, 키-값 저장소를 지원하는 기술. 문자열, 해시, 리스트, 세트, 정렬된 세트를 포함할 수 있으므로 데이터 구조 서버라고도 한다. 모든 데이터를 메모리에 올려 놓고 처리하고, 이벤트 기반의 네트워크 비동기 입출력 처리를 하여 초당 수만 건 이상의 요청을 처리할 수 있다. 또한 데이터의 가용성과 영속성을 위해 복제 및 RDB(Redis DB), AOF(Append Only File) 기능도 지원한다.

●<표 2> 원도우 감사로그 빅데이터를 분석하는 데 적용한 기술과 설명

저 찾게 되게 마련이다. 우리팀은 포함한 나머지 팀 또한 빅데이터 기술 전문가 집체교육 기간에 배운 기술을 먼저 검토했다. 플럼(Flume), 하둡(Hadoop), 스톰(Storm), R 등 다뤄 본 도구로 프로젝트를 수행할 수 있을지를 살펴 보았다. 그러나 보니 의도하지 않게 수료 프로젝트를 아는 기술 테두리에 맞춰서 접근하려는 ‘본말전도’ 상황이 벌어지기 시작했다. 당초 의도에서 벗어난 것이다. 이에 우리팀은 목표한 바를 이루기 위해서는 해당 주제를 선정했던 이유를 기억하며 더 적극적으로 파고 들어야겠다는 생각에 이르렀다. 결국 집체교육 중에 배우지 않았던 낯선 기술까지 적극 수용하기로 했다.

로그 수집도구: 로그스태시

첫 번째로 원도우 보안 이벤트 로그를 수집하기 위해 원도우 운영체제에서 이슈 없이 이벤트 로그를 수집·전달해 줄 수 있는 수집도구가 필요했다. 교육시간에 경험해 본 로그 수집 도구는 플럼이었다. 이 기술은 아파치 틱 프로젝트에 속할 만큼 강력한 도구지만, 우리팀의 주제를 수행하는 도구로서는 적합하지 않았다. 원도우 운영체제의 로그를 수집·전달할 수 있는 기능이 없었기 때문이다. 이에 원도우 로그 수집도구를 찾던 중 다양한 OS에서 이용할 수 있고 입출력 플러그인이 다양한 로그스태시라는 강력한 수집도구가 있음을 알게 됐다.

로그 저장·검색: 엘라스틱서치

이어서 로그를 저장·검색하기 위한 도구를 찾아 나섰다. 교육시간에 배운 대용량 데이터를 분산 저장하는 기술인 하둡을 놓고 적합성 검토를 했다. 하둡은 빠르게 늘어나는 대용량의 다양한 비정형 데이터를 저장하기 위해서는 의심할 여지 없는 선택이었다. 하지만 하둡 또한 원도우 이벤트 로그를 빠르게 검색하는 데는 적합하지 않았다. 하둡의 대안으로 루신 기반의 분산 검색엔진인 엘라스틱서치(Elasticsearch)를 선택했다. 엘라스틱서치는 안전한 분산 저장소와 강력한 검색기능을 지원하고 있었다. 더욱이 최근 로그스태시(Logstash)가 엘라스틱서치에 편입됨에 따라 로그 수집도구로 선택한 로그스태시와 조합하기에도 유리할 것이라고 생각했다.

실시간 처리 분석: Redis

세 번째로 스톰이 원도로 로그를 실시간 처리·분석하는 데 접합한지를 알아 보았다. 이 기술은 집체교육 시간에 배운 실시간 처리 기술로서 프로젝트 주제를 선정하기 이전부터 팀원들로부터 관심을 끌어왔다. 스톰은 프로그래밍이 가능하다면, 실시간 처리·분석에 얼마든지 응용할 수 있는 장점을 가지고 있으므로, 원도우의 보안 로그를 실시간 처리·분석하는 데 적합했다. 하지만 고민은 ‘스톰을 수집·저장·검색 도구와 어떻게 연동하느냐?’는 문제가 남아 있었다.

로그를 수집·전달하는 쪽에서 스톰과 연동할 수 있어야 했는데, 관련 사례가 없었고 로그 수집 툴인 로그스태시의 출력부분에서 스톰 연동 기능이 없었다. 결국 로그스태시에서 전달한 데이터를 저장할 수 있고, 스톰과 연동을 지원하는 매개체인 레디스(Redis)를 검토하게 되었다. 레디스는 인메모리 기반의 키값 저장소를 지원하고, 초당 수만 건 이상의 요청을 처리할 수 있다. 이에 따라 로그 수집도구에서 전달한 데이터를 중간에서 저장해 나를 수 있는, 데이터 버스 형태로 사용하기로 최종 결정했다.
우리팀이 여기까지 검토하고 적용한 기술을 간략하게 정리하면 <그림 1>과 같다.



●<그림 1> 실시간 저장·처리·분석하는 데 적용한 기술

원도우 보안 로그 분석

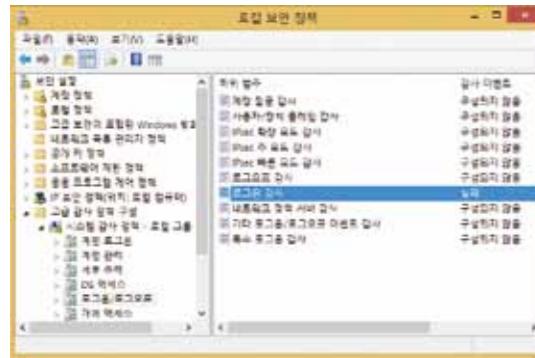
수료 프로젝트에 적합한 도구 선정이 끝나자, 팀원별로 배정받은 역할을 수행하기 위해 필요한 기술을 각자 검토해 취

합했다. 몇 차례의 모임에서 분석해야 할 다양한 원도우 보안 로그 중에서 어떤 것을 저장·분석할 것인지를 논의했다. 기본적으로 원도우 보안로그는 <표 3>에서 보는 바와 같이 다양한 감사로그 분류와 분류별 수백 개의 이벤트(EventID: 4000~6000)를 갖고 있다. 표본을 정의하지 않고 프로젝트를 시작하면, 제한된 일정 안에 목표한 분석 결과를 얻지 못할 것이라고 봤다.

우리팀은 보안 감사 로그 분류 중에서 사용자의 로그온/로그오프에 대한 Audit logon events를 토대로 로그를 수집·저장·검색·분석하기로 했다. 이 중에서 로그온 감사 실패 이벤트를 중심으로 보안 이슈를 발생시켜 시험·검증해 보기로 했다. 로깅을 구성하기 위해 먼저 대상 원도우 서버에 <그림 2>와 같이 로컬 보안정책의 고급 감사 정책을 적용해 로그온 감사 실패 이벤트를 발생시키도록 구성했다. 그리고 보안 위협을 발생시키기 위해, 특별한 도구를 사용하는 대신, 원도우 운영체제용 원격 데스크톱 연결 도구를 사용해 공격 대상 서버에 administrator(관리자 계정을 변경하지 않았다는 가정) 계정으로 암호를 무작위로 입력하는 브루트 포스 공격을 시도해 감사 이벤트를 발생시켜 보기로 했다. 더불어 감사 이벤트가 발생하는 즉시, 로그 수집도구인 로그스태시에서 파싱을 거쳐 데이터 버스 역할을 하는 레디스로 그것을 전달하도록 설정했다.

보안감사 로그 분류	세부 이벤트 설명
Audit account logon events	계정 로그온(자격 증명 유효성, 계정 로그온 이벤트)
Audit account management	계정 관리(사용자, 보안그룹, 컴퓨터, 메일그룹 관리·감사)
Audit logon events	로그온/로그오프(로그온, 로그오프, 특수 로그온, 계정잠금 감사)
Audit object access	개체 액세스(파일공유, 파일 시스템, 레지스트리, 이동식 저장소 감사)
Audit privilege use	권한 사용(중요한 권한 사용 감사)
Audit policy change	정책 변경(인증 정책, 감사 정책, 권한부여 정책 변경 감사)
Audit system events	시스템(시스템 무결성 감사, 보안상태 변경 감사)

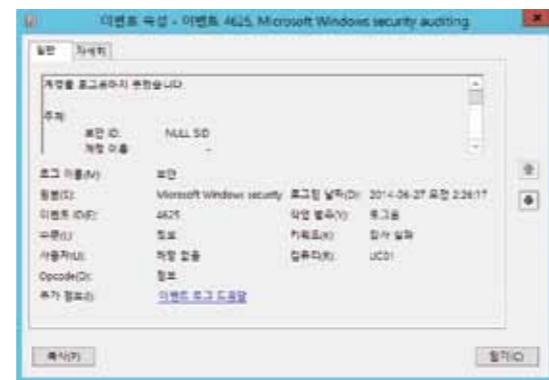
●<표 3> 원도우 보안감사 로그 분류



● <그림 2> 원도우 로그온 감사 실패 정책 설정

운영체제	원도우 서버 2008, 원도우 서버 2012
분류 - 하위 분류	로그온/로그오프 - Logon
유형	실패
원도우 2003 이전 버전에서 상응하는 이벤트	529, 530, 531, 532, 533, 534, 535, 536, 537, 539

● <표 4> 계정 감사 로그온 실패 이벤트 로그
(4625: An account failed to log on)



● <그림 4> 로그온 실패 감사 이벤트 로그(ID: 4625)



● <그림 3> 원격 데스크톱 연결을 사용한 로그온 감사 실패 테스트

보통 어떤 어떤 일을 맡았을 때, 기준에 갖고 있던 경험과 지식을 동원해 해결하려고 한다. 하지만 늘 새로운 기술이 나오는 IT 분야에서 만큼은 이 방법은 그리 잘 통하지 않는다. 늘 새로운 기술이 나오는 이유는 그만큼 예외 사항이라든가 요구 조건이 다양하다는 뜻이기도 하다. ‘감사 로그팀’도 수료 프로젝트를 진행하기에 앞서 조금이라도 알고 있거나 경험에 있는 기술과 도구를 적용하려 했다. 하지만 호미로 막을 일이 있고, 가래로 막을 일이 있음을 실감하고 처음부터 새로 시작

```

"host": "UC01",
"pubId": 1,
("Security",
  "Type": "Security-EventLog",
  "Timestamp": "2014-06-27T02:23:17.000+09:00",
  "Duration": "00:00:44",
  "Category": "로그온",
  "CategoryName": "로그온",
  "EventCode": "4625",
  "EventIdentifier": "4625",
  "EventTime": "2014-06-27T02:23:17.000+09:00",
  "EventTimeText": "2014-06-27T02:23:17.000+09:00",
  "EventTitle": "로그온 실패",
  "EventValue": "1",
  "EventValueText": "로그온 실패",
  "EventUser": "Administrator",
  "SourceName": "Microsoft-Windows-Security-Auditing",
  "SystemName": "UC01"
)
  
```

● <그림 5> 로그스태시에서 Redis로 전달하는 이벤트 로그

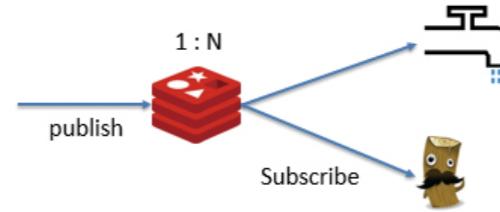
하여 하나씩 풀어나갔다.

실시간 처리 장벽에 부딪히다

레디스에 쌓인 로그를 다시 로그스태시를 사용해 엘라스틱서치에 저장하도록 설정하고, 다른 한쪽은 스톰에서 가져가도록 프로그래밍 하려 했다. 하지만 여기서 예상하지 못했던 난관에 부딪혔다. 스톰에서 레디스로부터 데이터를 가져오기 위해 레디스를 들여다 보면, 데이터가 없는 것이었다. 레디스에 로그 데이터를 보냈는데도 말이다. 문제가 해결되지 않아 계속 골치를 썩고 있을 때, 윤재문 팀원이 문제 해결의 실마리를 찾아냈다. 그것은 다름 아닌, 로그스태시에서 레디스로 전달하는 로그의 데이터가 ‘list’ 형태로 전달된다는 것이었다. 이 형식으로 로그를 전달하면, 중간에 한 명이 이 데이터를 가로채 가져가면서 메모리 기반의 레디스 저장소가 비게 되는 것이었다. 즉, 레디스에 전달된 데이터를 엘라스틱서치 저장소로 전송하기 위해 Indexer 역할을 하는 로그스태시에서 레디스의

쌓인 데이터를 가로채고 있었다.

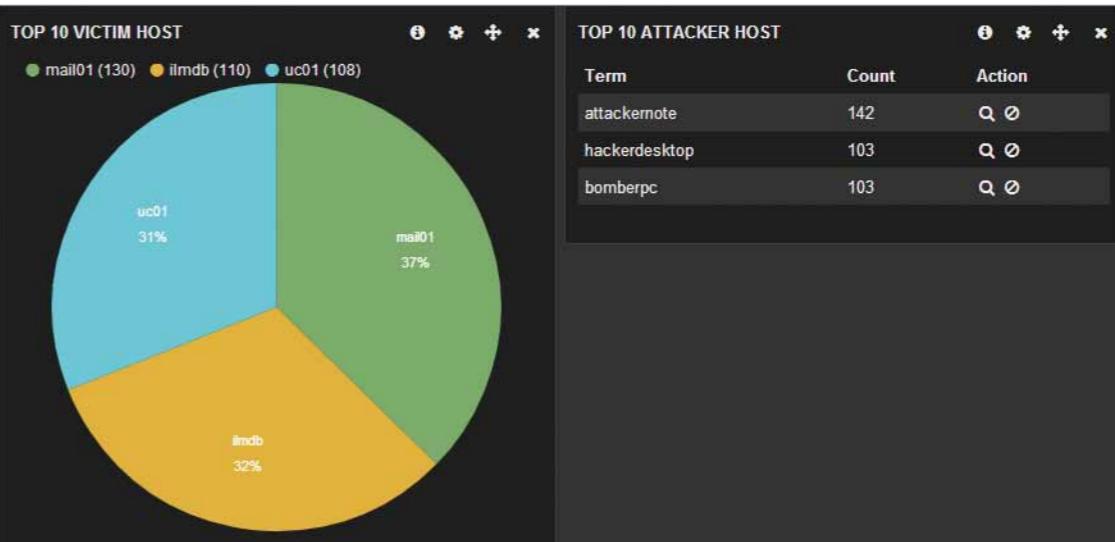
원인을 파악하고 난 뒤에 문제 해결은 그리 어렵지 않았다. 처음에 로그스태시에서 레디스로 로그를 전달할 때 ‘list’ 데이터 형태로 전달하는 대신에 ‘channel’로 전달되게 바꿨다. 이는 레디스에서 제공하는 1:N의 가입자/구독 형태 모델로서 channel에 가입한 Subscriber들 모두에게 특정 이벤트를 전달하게 하는 구조다. 따라서 스톰과 로그스태시를 통한 엘라스틱서치 저장소, 두 곳 모두에게 로그를 전달할 수 있게 됐다.



● <그림 6> Redis Pub/Sub 구조

실시간 보안 위협 대응 방안

로그스태시로 수집한 원도우 보안 이벤트 로그를 레디스에 저장한 다음, 원하는 최종 목적지로 데이터를 전달하는 과정까



● <그림 7> Kibana를 활용한 보안 대시보드 구현

지 기본 프레임워크가 완성되었다. 하지만 ‘보안 위협 요소를 사전에 발견하고 대응할 수 있어야 한다’는 최종 목표 지점에 도달하기 위해서는 위협 데이터를 가시화하고, 최종 사용자에게 알림 메시지를 줄 수 있어야 했다. 이에 우리팀은 두 가지 솔루션을 갖고 최종 마무리 작업을 진행했다.

첫 번째로 엘라스틱서치의 강력한 플러그인 중 하나인 Kibana를 활용한 보안 위협 대시보드의 구현이다. 앞서 언급한 바와 같이 엘라스틱서치는 다양한 플러그인을 갖고 있다. 그 중에서 Kibana는 데이터를 가시화해 대시보드 형태로 보여줄 수 있다. <그림 7>과 같이 엘라스틱서치에 저장된 로그 데이터를 가져와서 웹 UI에서 사용자가 손쉽게 검색하고 대시보드를 구현할 수 있는 것이다. 이에 따라 우리팀은 계정 로그온 감사 실패 이벤트를 이용해 보안 위협 대시보드를 구현하기로 했다. 우리팀은 무차별 브루트포스 공격을 체크하기 위해 계정 로그온 감사 실패 이벤트 로그 데이터에 존재하는 상태 값을 기준으로 대시보드를 구성해 보았다. 해당 이벤트 안의 상태 값은 다양한 종류가 있지만, 그 중 사용자 계정은 일치하나 패스워드가 틀린 항목에 대해 발생하는 상태 값인 0xC000006A 상태 값을 체크해(공격자가 기본 admin 계정에 무차별 공격을 시도했다고 가정) 해당 이벤트를 많이 발생

Status and Sub Status Codes	Description (not checked against "Failure Reason:")
0xC0000064	user name does not exist
0xC000006A	user name is correct but the password is wrong
0xC0000234	user is currently locked out
0xC0000072	account is currently disabled
0xC000006F	user tried to logon outside his day of week or time of day restrictions
0xC0000070	workstation restriction
0xC0000193	account expiration
0xC0000071	expired password
0xC0000133	clocks between DC and other computer too far out of sync
0xC0000224	user is required to change password at next logon
0xC0000225	evidently a bug in Windows and not a risk
0xC000015b	The user has not been granted the requested logon type (aka logon right) at this machine

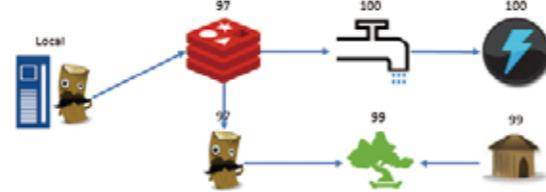
(출처: <http://www.ultimatewindowssecurity.com/securitylog/encyclopedia/event.aspx?eventid=4625>)

● <표 5> 계정 감사 로그온 실패 상태에 따른 설명
(Audit account Failed logon events (Ex : Status = 0xC000006A))

시킨 공격자의 호스트와 공격 대상이 된 호스트로 <그림 7>과 같이 간단한 대시보드를 구현했다.

두 번째로 실시간 처리 분석 기술인 스톰을 사용해 보안 위협을 사용자에게 실시간으로 알려줄 수 있는 기능을 구현했다. 엘라스틱서치를 실시간 검색하고 대시보드에서 확인해 보안위협 요소를 파악할 수 있지만, 이는 사용자가 해당 위협을 감지하기 위해 보안 위협에 대해 지속적으로 모니터링해야 하므로 현실적이지 못하다. 이 문제를 해결하기 위해 다음과 같은 고민을 하기 시작했다. ‘만약 짧은 시간 동안 특정 대상을 노리고 무차별 대입 공격으로 볼 수 있는 이벤트가 다양으로 들어오면, 보안 위협으로 판단해 해당 서버 운영자에게 실시간으로 알려준다’. 이러한 고민 끝에 현재 시점에서 ‘이전 3초 동안 시도한 대상 호스트의 접속 실패 카운트’를 실시간으로 스톰 서버 콘솔 화면에 출력하도록 프로그램을 설계했다. 즉, 계정 로그온 감사 실패 이벤트가 짧은 시간에 특정 호스트에서 얼마나 발생했는가를 화면에 출력해 보게 한 것이다. 사실 처음에는 어떤 식으로든 사용자에게 알릴 수 있도록 메일 또는 메시지를 연동해 보려고 했다. 하지만 프로젝트의 제약조건에 따라 최종적으로 원하는 바를 얻지 못한 점은 아쉬움으로 남아 있다. 그래도 처음에 기획했던 바를 일부 달성한 것만으로 팀원 서로를 격려했다. 실시간 처리 분석 기능을 구현하는 등 추가 발전방향에 대해 다시금 생각해 볼 수 있게 됐다.

우리팀은 위와 같은 기획과 설계, 테스트 과정을 거쳐 <그림 8>과 같은 프로젝트 시스템을 최종적으로 구축할 수 있었다.



● <그림 8> 프로젝트 시스템 최종 구축 설계도

CONCLUSION

프로젝트 팀이 꾸려지자마자 역시나 가장 걱정스러웠던 점은, 교육일정이 끝나고 현업으로 복귀한 뒤의 수료 프로젝트 지속 가능 여부였다. 현업에 복귀해서 각자 맡은 일을 하면서 수료 프로젝트의 발전을 위하여 멘토를 포함한 5명이 시간을 맞춰 한 자리에 모이기란 결코 쉬운 일이 아님을 팀원들 모두는 알고 있었다. 집체교육이 끝나갈 무렵, 프로젝트 주제 확정을 겸해 팀장을 선발하기 위한 모임을 가졌다.

팀원들간의 소통에 대해

‘윈도우 보안 로그 분석’이라는 주제를 기획했던 정하권 팀원이 팀장이 되면 어떻게 뛸지라는 의견이 나왔지만, 김태형 팀원의 결정적 의견으로 김용문 팀원이 팀장으로 선출됐다. 김태형 팀원이 프로젝트 주제를 제안한 사람이 팀장이 되는 것은 바람직하지 않다는 의견을 제시했다. 그 이유는 자칫 프로젝트가 개인 중심으로 흘러갈 수 있다는 것이었다. 따라서 팀장은 팀원이 다 같이 모일 수 있게 노력하고, 소통을 위해 노력하고 협신하는 사람이라면 한다는 의견을 내놓았다. 이 의견에 팀원 모두가 동의한 것이 수료 프로젝트 수행 최우수 팀으로 선정될 수 있었던 시발점이 아니었나 싶다. 프로젝트를 진행할 때, 말형인 김태형 팀원이 팀원들을 잘 조율해 준 덕분에 초기의 목적을 달성할 수 있었다.

우리팀은 집체교육이 끝나자마자 소통을 원활하게 하기 위해 ‘빅데이터 미니 프로젝트’라는 이름으로 ‘밴드’를 개설하는 한편, 각자가 테스트한 사항을 신속히 공유할 수 있도록 ‘N 드라이브’를 개설했다. 또한 정기적인 모임을 갖기 위해 모임 장소를 예약하고, 사전에 공지해 일정을 조율했다. 팀원들의 의견을 들어주고 프로젝트를 진행하는 데 어려움이 없도록 아낌없는 역할을 해준 김용문 팀장께 다시 한번 감사의 마음을 전한다.

프로젝트를 진행할 때 가장 어려운 점은 일정관리, 범위관리, 품질관리가 아닌, 바로 의사소통관리라고 생각한다. 프로젝트 팀이 TFT 형식으로 짧은 기간에 구성이 되든, 물리적으로 멀리 떨어진 가상팀(Virtual Team)으로 구성되든 팀원들 사이에 원활한 의사소통이 이루어지지 않으면, 해당 프로젝트는 결코 좋은 결과물을 도출해 낼 수 없다고 생각한다.

평일일 경우, 우리팀은 보통 오후 7시를 전후해 강남에서 모였다. 수료 프로젝트 팀이 구성되고 집체교육이 끝날 무렵, 김종희 팀원이 수원으로 발령받았음을 알게 되었다. 늦더라도 꼭 참석하려고 최선을 다하는 김 팀원의 모습에서 팀원들 모두가 힘을 얻었다. 윤재문 팀원은 현업에서 진행하는 별도의 프로젝트로 굉장히 바쁜 일정을 소화하고 있었음에도 스톰을 이용한 실시간 처리분석에서 만난 복병을 적극적으로 해쳐나가서 팀원들을 놀라게 했다.

마지막으로 공용준 멘토께 진심으로 감사를 드린다. 프로젝트의 전체 방향을 코칭하고, 직접 참석하기 어려울 때면 영상 통화를 통해서라도 팀원들의 문의에 대한 답변과 프로젝트 진행 방향을 점검해 주었다. 밴드를 통해서도 프로젝트에 도움이 될 만한 의견들을 지속적으로 제안해 주셨다. 공용준 멘토의 지원이 없었다면 최초 목표했던 방향에서 프로젝트를 얼마나 진행할 수 있었을지….

향후 발전 방향

윈도우 보안 로그 분석 프로젝트를 진행하면서 그 동안 몰랐던 빅데이터 처리 기술에 대해 많은 부분을 배우고 경험했다. 프로젝트의 제약조건(일정, 범위, 자원 등)으로 인하여 보다 많은 테스트를 진행해 보지 못한 부분이 다소 아쉽지만,

기회가 된다면 다음과 같은 형태로 프로젝트를 발전시켜볼 계획이다.

첫째, 보안 위협에 실시간 대응 및 처리 강화

- 윈도우 보안 로그의 상관관계를 분석해 보안 위협을 도출
- 해당 보안 위협들에 실시간으로 대응할 수 있도록 프로그래밍(스톰)
- 보안 위협을 사용자에게 신속하게 알려줄 수 있도록 연동(메일 또는 메시지)

둘째, 주기적 로그 분석을 통한 보안 위협 관리

- 엘라스틱서치의 실시간 검색뿐만 아니라 주기적으로 보안 로그 분석이 가능한 장치 마련
- 엘라스틱서치와 하둡을 연계(연계 시 MR, Hive, Pig 등을 사용)
- R을 이용한 다양한 분석

셋째, 간단한 설치 배포 및 사용 시나리오 작성

- 프로젝트에 사용된 시스템 구성을 간단하게 설치·배포할 수 있도록 구현
- 다양한 시나리오를 적용하여 많은 사용자들이 손쉽게 사용 할 수 있도록 제공

끝이 아닌 새로운 시작

당초 목표했던 바를 모두 이루지는 못했으므로 현재의 결과물에 만족하며 안주하지 않을 계획이다. 우리팀은 그동안 경험했던 지식을 꾸준히 다듬고 노력해 지속적으로 프로젝트 결과를 발전시켜 나가기로 했다. 빅데이터 기술 전문가 과정을 통해 많은 지식과 경험을 얻힐 수 있도록 기회를 주신 한국DB진흥원 관계자 분들께 감사하며, 두 달이라는 프로젝트 기간에 한 목표를 달성하기 위해 노력해주신 공용준 멘토와 김용문 팀장, 이하 모든 팀원께 다시 한 번 감사의 말씀을 드린다. ☺



“익숙한 것을 버리자 빅데이터가 가까이 다가왔다”



김용문 팀장
인터넷과사람들 대표

업무 환경에서 필요한 현실적인 주제라고 생각된다.

배치 처리와 실시간 처리를 동시에 할 수 있는 아이템을 찾다가, 실시간으로 수많은 로그 데이터가 생성되는 운영체제로 그 분석을 주제로 정했다. 대용량 로그 데이터 처리·분석에 특화된 상용 분석도구가 고가에 팔리고 있다고 들었다. 그만큼 니즈가 있다는 말이지 않겠나.

완료하기까지 고생이 많았다는 얘기를 들었다.

배우는 입장에서 접해 보지 않았던 빅데이터 기술을 총동원하다 보니 각오했던 어려움이었다(웃음). 원가 아는 테두리에서 맴돌다 보면, 안전하기는 하겠지만 그만큼 얻는 것도 없다고 생각한다. ‘수료 프로젝트를 이왕 할 거라면 힘이 들더라도 즐겁게 하고 다른 사람에게도 도움이 될 만한 것을 만들어 보자’는 의견에 팀원들이 공감하면서 목표 지점까지 함께 갈 수 있었다. 주제의 특성에 따른 어려움도 있었다. 분석 인프라를 구축하는 과정은 운영체제의 버전과 오픈소스 분석 도구의 버전 등에 매우 민감하다. 실수를 반복하는 과정의 연속이었으므로, 비슷한 프로젝트를 하는 사람들이 고생을 덜하도록 매뉴얼로 정리해 두면 좋겠다고 생각했다. 수료 프로젝트로 했던 결과물을 공개할 때, 매뉴얼까지 동시에 공개할 계획이다.

에피소드도 많았을 거 같다.

낮에는 회사에서 일하고 퇴근 후 시간과 주말을 이용해 과제를 수행해야 했기에 육체적으로도 많이 힘들었다. 그러다 보니 당연히 에피소드도 많았다(웃음). 수원에서 일하는 한 팀원이 버스를 타고 서울 모임장소로 오다가 잠이 들었다가 눈을 떠보니 수원으로 내려가는 차 안이었다고 했다.

어떤 것이 가장 기억에 남나.

어렵다고 느꼈던 것을 정면으로 마주쳐 하나씩 해결해 나갈 때의 뿌듯함이 기억에 남는다. 말 그대로 고뇌와 희열의 연속이었습니다. 예를 들어 원도우 서버 감사 데이터를 배치 처리와 실시간 처리를 할 때, 예전 방식으로 했다면 실시간 처리 서버와 배치 처리 서버를 따로따로 구현했을 수도 있었을 거다. 오픈소스 패키지를 이용함으로써 실시간 처리와 배치 처리를 한 곳에서 구현할 수 있었다.

빅데이터 아카데미 수료 후 달라진 점은.

팀원 중에 한 명이 보안 소프트웨어 업체에서 근무하는데, 빅데이터 아카데미에서 알게 된 대용량 데이터 처리 기법을 회사의 업무에 적용함으로써 바로 효과를 봤다고 한다. 로그 데이터를 적절히 처리하는 게 중요한데, 쏟아져 나오는 데이터를 어떻게 해야 할지 기존의 관점에서만 해결하려고 했는데 새로운 관점에서 접근할 수 있게 된 것이다. 자신의 생각에만 머물지 말고, 빅데이터를 현실적으로 바라보라고 조언하고 싶다. 친구들에게도 빅데이터에 대한 선입견을 버리고 현실적으로 받아들이라고 강조하는 사람이 되었다.

프로젝트 발전 계획이 있나.

당장 상용화하거나 비즈니스 아이템화할 계획은 없다. 하지만 원가 분명한 것을 하나 머릿속에 갖고 다니다 보면, 비즈니스 아이템으로 연결되곤 했던 기억이 있다. 그 사이에 오픈소스 형태로 미리 공개해 아이디어도 얻고 발전시켜 나가볼 생각이다. 현재 상태에서도 소스 데이터만 바꾸면 사용할 수 있을 정도로 만들어졌으니 누구나 사용해볼 수 있다. ☺

원도우 서버 감사 로그 분석 시스템

프로젝트 소개

원도우 보안 이벤트 로그를 실시간으로 수집·통합·분석해 보안 위협 요소를 사전에 발견·대응할 수 있는 분석 시스템을 구축한 프로젝트

구분

기술 전문가 과정

프로젝트 기간

2014년 05~06월

멘토

공용준(다음카카오 팀장)

작용도구

엘라스틱서치, 레디스, 로그스타시, 스톰

수집 데이터

원도우 OS에서 임의적으로 생성한 감사 로그 데이터

산출물

- 원도우 보안 로그 분석 샘플(로그스타시)
- 엘라스틱서치 Kibana 플러그인으로 구현한 대시보드
- 스톰을 활용한 실시간 분석 샘플 소스

교육 참여형태

자발적 참여(5) / 회사 권유(0)

진행

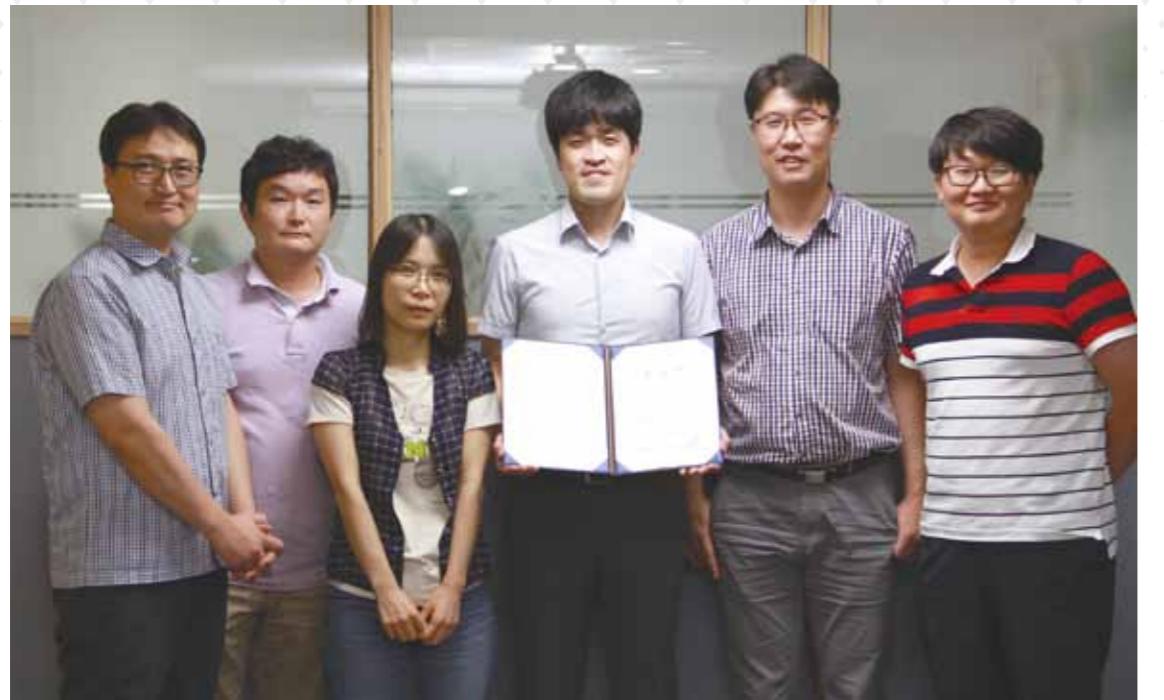
• 김용문 팀장	SI 업체 대표	경력 10년
• 김태형 팀원	—	—
• 윤재문	프린팅 서비스사 과장	경력 6년
• 김종희	SI 업체 과장	경력 6년
• 정하권	보안 SW사 선임 연구원	경력 7년

빅데이터 아카데미 수강 후 변화

- | | |
|-------------|---|
| 수강 전 | <ul style="list-style-type: none"> • 빅데이터 처리는 하둡으로 처리해야 가능하다. • 빅데이터 처리 분석은 굉장히 어려울 것이다. • 빅데이터는 먼 곳에 있다. |
| 수강 후 | <ul style="list-style-type: none"> • 빅데이터 처리를 위한 다양한 기술이 있다 • 도메인에 대한 이해만 있다면 얼마든지 도전할 수 있다. • 빅데이터는 생각보다 가까운 곳에 있다. |



상관관계 속에서 범죄의 숨겨진 비밀을 캐다



글 김대훈 프리랜서, BI 솔루션 엔지니어

범죄 예방을 위해 방범용 CCTV 설치 대수를 늘려가고 있지만 범죄율은 줄어들지 않고 있다. 빅데이터 기술을 활용해 범죄 발생에 대한 근본적인 원인 도출과 시민의 자발적 참여를 통한 공공기관 또는 지방자치단체에서 파악하지 못한 우범 지역을 파악해 미리 대처할 수 없을까 하는 생각에 시작한 프로젝트다.

CHALLENGES

하루가 멀다 하고 범죄 소식이 쏟아지고 있다. 프로젝트 팀원들 또한 그 불안으로부터 예외는 아니었다. 가족이 안전하길 바라는 마음은 누구나 마찬가지일 것이다. 이러한 불안감은 CCTV, IPTV, 차량용 블랙박스 등의 감시 카메라 시장으로 이어져 관련 산업이 매년 성장하고 있다.

하지만 아무리 자동차 블랙박스나 IPTV 성능이 좋아졌다 하더라도 공공의 목적이 약한 자동차 블랙박스와 IPTV로는 한계가 있다. 감시카메라 사각지대는 항상 존재하고 범죄에 악용되기까지도 한다. 개인의 안전을 목적으로 하는 감시카메라 설치도 중요하지만 공공의 목적으로 감시카메라를 설치하는 것 역시 중요하다. 범죄자들은 ‘거리의 눈’을 피해 범죄를 저지르기 때문이다.

방범용 CCTV를 설치한 지역은 범죄율이 다소 감소하지만, 그 주변의 범죄율이 증가하는 것을 범죄의 전이효과(crime displacement effect)라고 한다. 범죄의 사각지대는 항상 존재하기 마련이다. CCTV가 모든 범죄를 막을 수는 없지만 범죄 기회를 감소시킬 수 있다. 앞으로 CCTV와 같은 감시 장비와 더불어 도시환경 디자인을 개선해 나간다면 범죄율도 점차 줄어들 것이다.

이처럼 불안한 이 시기에 사회 안전에 관한 문제에 관심을 갖게 되었고, 본 프로젝트를 진행하게 되었다.

을 파악하고 이사하기로 결정했을 만큼 지역의 안전지표에도 민감하다.

그래서인지 처음 조원들에게 제시한 단어도 ‘범죄’였다. 그러자 ‘CCTV’ 의견이 나왔고 ‘어린이 보호구역’, ‘경찰서-지구대’ 등 다양한 의견이 줄어들어 나왔다.

우리는 제시된 단어들을 취합해 주제를 선정했고 흥미로운 프로젝트를 진행할 수 있는 초석이 되었다.

SOLUTION

프로세스와 계획 수립

HBase의 사용을 위해 Thrift Server와 zookeeper 서비스를 추가했지만, 일단 하이브에서 테스트하기로 했다.

수집된 비정형 정보를 정형화해 하이브의 테이블에 적재했고, 이를 RHadoop, RHIVE, RHdfs를 설정해 R에서 시각화하거나 GeoJSON과 JSON 파일을 추출해 D3.js로 시각화했다.

1,5개월이라는 기간을 알차게 활용하기 위해 일정을 수립해 지키려고 노력하면서 목표했던 결과물을 얻을 수 있었다.

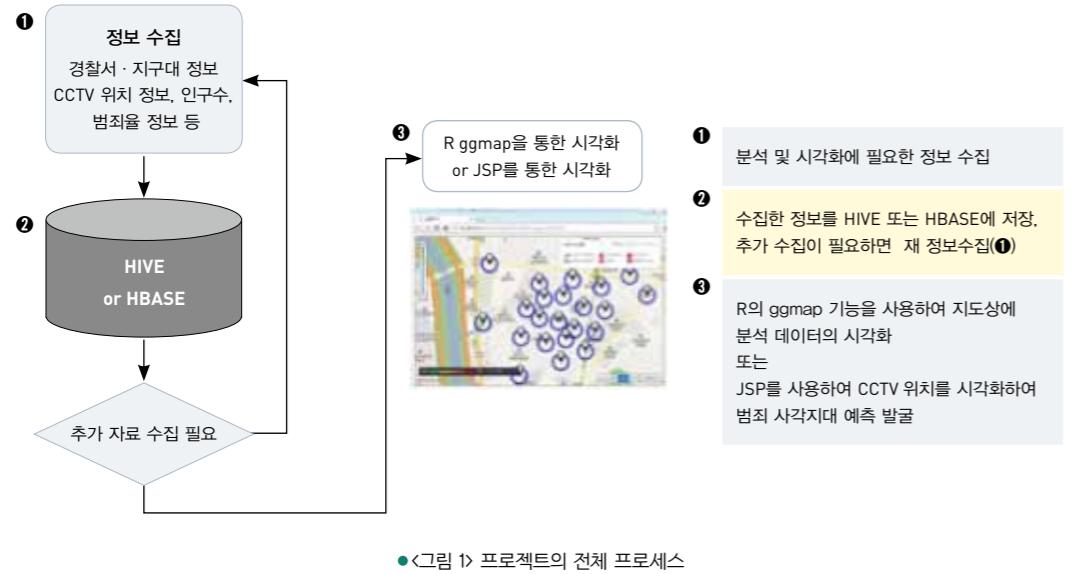
분석 도구 R에서 googleVis, ggplot2, ggmap 라이브러리로 CCTV 위치를 시각화해 범죄 사각지대를 도출했다. R 작업 후에는 D3.js와 Google Map API를 사용해 지역별 범죄율과 CCTV 설치 대수를 막대그래프와 지도에 색상을 입혀 시각화했다. D3.js에 대해 전혀 몰랐던 초기 단계에서는 JSP와 ajax로 개발을 시도했다. 하지만 집체교육 시간에 배웠던 R과, 다양한 형태로 시각화할 수 있는 D3.js로 진행하는 것이 본 프로젝트 취지에 맞다고 보고 R과 D3.js로 선호했다. JSP로 개발을 하면서 D3.js의 가치를 알게 되었다.

범죄 취약계층	2011년(건)	2012년(건)	증가율(%)	주요 범죄
노인	76,624	126,482	65.1	사기, 지능범죄
아동	7,508	14,416	92.0	폭력
여성	398,542	440,130	10.4	절도
합계	482,674	581,028	20.4	사기, 지능범죄

(출처: 2013년 10월 15일 경찰청이 국회 안전행정위원회 소속 강기윤 의원에게 제출한 국정감사 자료)

● <표 1> 취약 계층의 범죄발생 추이

팀장인 나는 평소 범죄와 사회안전기반시설에 관심이 많았다. 현 주거지 또한 어린이 보호구역이고 주변에 CCTV가 많은 것



에 따라 사용하는 서비스만 기동시켜며 작업했다.



테이블의 설계 및 데이터의 수집

플럼으로 수집한 트위터의 데이터는 선정적인 내용이 많았다. 그도 그럴 것이 범죄 관련 키워드로 수집했으므로 거친 내용이 많을 수밖에 없었다. 트위터 테이블을 제외한 CCTV, 상권, 경찰서, 학교 등의 위치정보 테이블은 각 지자체에서 받은 자료를 토대로 설계했다.

HUE 인터페이스 안에서 테이블 설계 및 수집 작업은 클라이언트 PC에서 사용하는 도구에 비해 불편했다. 하지만 셀 프롬프트에서 작업하는 것보다는 편리했고, 자동완성 기능과 그래프 등 생각보다 많은 기능을 제공했다.

플럼에서 범죄 관련 키워드를 설정해 〈그림 3〉과 같이 JSON 형식으로 데이터를 수집했다. 특정 디렉터리의 연도·월·일·시간별 데이터를 HDFS에 수집하도록 했다. 네임노드가 아닌, 3대의 서버에서 플럼 서비스를 구동했고 SNS 데이터를 트위터가 아닌 곳에서 확인할 수 있다는 것이 처음에는 낯설었지만 곧 익숙해졌다.



HDFS의 데이터를 하이브에서 확인하려면, 〈그림 4〉와 같이 Hive Editor의 설정 탭에서 Hive Serde를 설정하면 된다. Serde는 JSON 형식의 데이터를 읽고 쓸 때 사용하는 모듈이다.



〈그림 5〉와 같이 방범용 CCTV 설치 대수가 적은 서울의 기초 지자체를 확인할 수 있었다. 종로구와 송파구의 CCTV 대수는 의외였다. 개인적으로 상권이 형성되어 있거나 소위 '강남 4구'에 속하는 지역은 설치 대수가 많지 않을까? 하는 생각을 했기 때문이다. 현시점에도 방범용 CCTV 통합구축 사업을 진행하는 지자체가 있을 것이다. 이런 곳에서 빅데이터 분석을 토대로 CCTV를 설치하면 더 효율적이지 않을까.

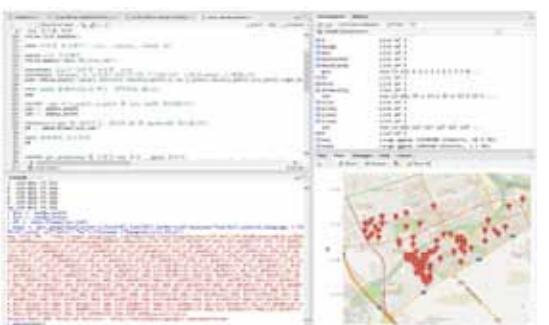
인포메티카(Informatica)의 '데이터 트랜스포메이션'을 사용해



DOC, PDF 형식의 CCTV 시설 위치 파일을 텍스트로 추출해 하이브에 저장했다. 데이터 트랜스포메이션 솔루션은 라이선스 문제가 되지 않는 수준에서 사용했다. 솔루션의 특징을 설명하자면 DOC, PPT, PDF, XLS 등 비정형 문서를 자동으로 정형화해 테이블에 적재할 수 있고, 수십만 개 이상의 파일도 자동으로 처리하는 편리한 솔루션이다.

〈그림 6〉은 각 구청에서 받은 비정형 파일을 인포메티카 데이터 트랜스포메이션 솔루션(〈그림 7〉 참조)의 파서로 정형화한 형식으로 변환하는 과정이다. 물론 이 과정을 자동화하려면 라이선스가 필요하기에 일부는 수작업으로 했다. 인터넷에 떠돌아다니는 밑을 수 없는 프로그램을 사용할 수는 없었으므로 직접 했다.

R의 웹 클라이언트 버전에서 RHIVE 라이브러리로 하이브에



●〈그림 8〉 R 웹 버전에서 CCTV, 경찰서, 지구대, 방범대의 위치 확인

접속해 〈그림 8〉과 같이 위도와 경도 정보를 조회했다. 이를 googleVis, ggplot2, ggmap의 R 라이브러리를 통해 지도에 시각화했다. CCTV 위치는 강남구에 한해 먼저 적색 마커로



●그림 9> CCTV, 경찰서, 지구대, 방범대의 위치 지도 확대



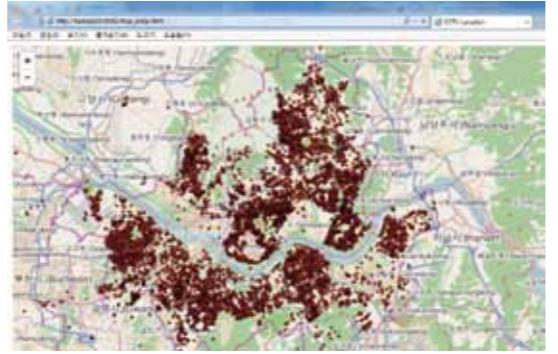
●그림 10> 초등학교, 유치원의 위치를 청색 마커로 표시



●그림 11> CCTV 사각지대 발굴

66

표시했다.
로컬에서 R의 PC 버전 클라이언트를 설치해, 하이브 ODBC를 설정할 필요 없이 편리하게 웹에서 작업할 수 있었다.
<그림 9>의 CCTV · 경찰서 · 방범대 · 지구대의 위치와 <그림 10>의 학교 · 유치원의 위치를 결합해 <그림 11>과 같이 CCTV 가 필요한 사각지대를 찾아낼 수 있었다. 그곳은 오래된 아파트 단지여서 주변에 방범용 CCTV가 없었지만, 근처에 학교가 있었다.
R로 분석하기 전에 기술검토 과정에서 JSP로 개발을 시도해



●그림 12> Google Map API와 D3.js를 사용한 CCTV 설치 위치 표시

보았다. JSP의 MVC 모델을 사용해 하이브의 데이터를 웹에 보여주거나 CCTV의 위치를 반경 50미터의 동그란 원으로 표시해 보았다. ROI 아닌, Google MAP API나 D3.js를 사용해도 시각화가 가능하겠다는 생각이 들었다.

분석 시스템 구성과 시각화

CDH5 기반에서 하이브와 플럼, R, 자바, JSP, D3.js를 사용했다. 처음엔 하둡 1.0 기반에서 플럼과 하이브를 사용했지만 유저 인터페이스와 관리가 편리한 하둡 에코시스템을 사용해 보고 싶었다. 그래서 멘토께서 제안해 주신대로 클라우데라의 CDH5를 기반으로 프로젝트를 진행했다.

처음 설치하고 구성하는 것이라 효율성은 다소 떨어졌지만, 편리한 클라우데라 서비스 관리 화면 UI를 통해 언제든지 원하는 대로 모듈構성을 변경할 수 있어서 좋았다.

<그림 13>은 클라우데라 관리 화면에서 확인할 수 있는 노드별 모듈 구성 화면이다. 이 화면을 통해 설치된 서비스를 한

눈에 확인할 수 있다. 설치하고 사용해 보니 불필요하게 구성된 서비스들이 눈에 들어오기 시작했다. 플럼이나 스파크 (Spark), 하이브의 구성을 줄이거나 네임노드의 서비스를 분산해야 했다. Hadoop01 노드에 있는 HIVEserver2와 Oozie, Sqoop, Hbase Master가 바로 그것이다.

SEQ	기술	내용
1	클라우데라 CDH5	에코 시스템
2	하이브	경찰서, 학교, CCTV 위치, 트위터 데이터 저장
3	플럼	범죄 관련 키워드로 트위터 글 수집
4	R(rhdfs, rHIVE, ggmap)	시각화와 분석
5	JSP, AJAX	하이브 데이터를 MVC 모델을 사용해 웹에 표현
6	D3.js + Google MAP API, GeoJSON, JSON	범죄율과 CCTV 설치 위치, 설치 대수를 막대 그래프와 지도로 시각화
7	Java + Google MAP API	주소(지번, 도로명) 좌표 변환 프로그램 개발
8	Informatica Data Transformation	PDF, HWP 문서에서 텍스트를 추출해 CSV로 변환

●표 2> 프로젝트에 사용된 기술(2014.08.08 기준)



●그림 14> PURE JAVA와 Google Map API를 사용한 좌표변환 프로그램

프로젝트 기간 내에 중점적으로 사용한 기술은 <표 2>와 같다. 짧은 기간에 좌표 변환을 수월하게 도와 주었던 지오서비스의 지오코더 프로그램은 <그림 14>와 같이 자바 프로그램을 직접 개발해 대체했다.



●그림 13> 노드별 구성 서비스

수행 과정에서 만났던 문제점과 해결 방법

기술

필자와 이상민 팀원은 클라우데라의 CDH5를 한국데이터베이스진흥원에서 제공받은 VM서버에 적용하기 전에 개인 노트북(i7 8core, 16Gb, 256 SSD)에서 4노드로 구성해 테스트해보았다. 하지만 리소스 문제로 에코시스템 전체 모듈의 완벽한 설치는 불가능했다. 전체 모듈이 아닌, 필요한 모듈만 설치 했다면 성공했겠지만, 전체를 설치해 다양한 기술 사항을 검토해 보고 싶었다.

제공받은 VM서버에는 이미 하둡 1.0과 플럼, 하이브가 동작하고 있었고, CDH5의 설치 테스트는 결과적으로 실패했다. 하지만 직접 확인해 보고 싶어서 VM서버에서 에코시스템을 설치해 보기로 했다.

VM서버에 설치돼 있던 하둡 1.0과 관련된 모듈을 모두 없애고 CDH5를 설치해 원하는 모듈 추가까지 아무 문제 없이 성공했다. 결국 전체 모듈을 설치하기에는 개인 노트북 리소스의 한계가 있었음을 알았다.

우리는 전체 모듈을 테스트 서버에 설치해 Hbase, Impala, Oozie, Sqoop 등을 시험해 볼 수 있었고, 단기간 내에 결과물을 낼 수 있는 모듈을 선정해 프로젝트를 진행했다.

시스템

플럼으로 트위터에서 키워드를 수집해 HDFS에 저장한 후, 하이브 SerDe(Serializer–Deserializer)를 통해 HUE에서 데이터를 조회할 수 있었다. 하지만 R과 RHIVE를 통한 데이터 조회와 동시에 진행하면서 HIVEserver2 서비스가 빈번히 다운되는 현상이 발생했고, 시스템의 메모리 스왑 문제도 함께 일어났다.

우리팀은 이에 따라 필요한 모듈만 기동하기로 하고 기술적 검토를 진행 중이던 Hbase, Impala, Oozie, Sqoop과 데이터를 수집중이던 플럼 서비스를 일시 중지시키고, 단기 목표인 CCTV 사각지대와 범죄를 분석에 초점을 맞춰 프로젝트를 진행했다.

국가정보공유포털(www.data.go.kr)에서 각 지방자치단체(이하 지자체)의 CCTV 데이터를 확보하려 했으나 대부분 미공개 상태였다. 각 지자체에 CCTV 데이터를 요청했으나 범죄목적으로 악용될 수 있고, 내부 자산이라서 공개가 어렵다는 답변을 받았다.

서울시 보안정책 담당관에게 유선상으로 문의했을 때도 범죄 목적으로 악용될 수 있으므로 불가능하다는 답변을 또 다시 받았다. 이에 이상민 팀원이 “CCTV의 개인정보 보호법에 대한 항목과 서울시 정보공개법 제9조 제1항 제2호”를 근거로 정보 공개를 요청해 CCTV 관련 데이터를 얻을 수 있었다. 이 자료가 없었다면, 프로젝트 주제를 바꿔야 하지 않았을까? 관련 법령을 찾아보고 자료를 재요청한 것이 주효했다.

좌표 오류의 문제

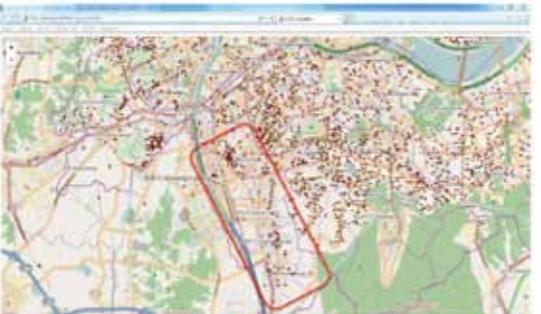
각 지자체에서 받은 좌표는 일정한 규격이 없었다. 파일 형식은 HWP, PDF, DOC, XLS에 이르기까지 다양했고, 내용 자체도 비정형이었다. 지자체마다 지번 주소, 도로명 주소 사용이 달랐고 심지어 혼합해 사용한 경우도 있었다. 이 가운데 시스템적으로 수작업 보정이 필요한 주소도 더러 있었다. 각 시설의 위치를 위도 경도로 표시하고, 모든 지자체에 적용 가능한 표준 문서형식이 필요함을 느꼈다. 향후 데이터 활용을 위해서는 데이터 표준안 마련도 중요한 부분이다. 수작업으로 데이터를 보정하고, 비정형 데이터를 정형화해 수집하는 데 많은 시간이 들었다.

시각화

D3.js를 처음 접했고 지도 좌표와 투영법에 대한 이해가 부족해서 처음에는 지도가 웹브라우저에서 한참 벗어나 보이지 않았다. 투영법 설정과 좌표를 수십 차례 재설정해 확인하는 과정을 거쳐 원하는 결과를 얻을 수 있었다.

D3.js와 하이브를 연결하는 작업에도 문제가 있었다. 당연히 연결될 것이라 생각해 몇 날 며칠을 고생했지만, 연결할 수 없었다. 구글에서 검색해 확인한 결과, 하이브와 연결했다는 정보는 없었고 JSON과 GeoJSON 파일을 사용해 시각화하는

방법을 찾을 수 있었다. D3.js에 대한 자신감을 높이기 위해 D3.js 관련 책을 구입해 공부했다.



- <그림 15> 금천구의 CCTV 현황



● <그림 16> 서울시 구별 범죄율

팀 프로젝트를 위한 커뮤니케이션

기술 문제를 해결하고 데이터 수집이 원활하게 이뤄지면 네이버 카페에 수시로 관련 사항을 올려서 팀원끼리 공유했다. 카카오톡으로 공지하고 의견을 나눴다. 부족한 부분에 대해서는 전화로 의견을 나누고, 그래도 부족했을 때는 직접 모여서 해결책을 도출했다. 팀원 모두가 서울에서 생활했으므로 모이는 것에는 크게 어려움이 없었다. 더러 회사 생활 때문에 바쁜 팀원이 있었지만, 네이버 카페에 기록된 내용들을 토대로 팀 일정을 좋아 올 수 있었다.

CONCLUSION

기술을 하나하나 적용해 나갈 때의 즐거움과 문제를 해결해 나갈 때의 희열로 시간 가는지도 모르게 프로젝트를 했다. 필자 개인적으로는 수료 프로젝트와 회사의 BMT 일정이 겹쳐 한때 바쁜 시간을 보냈지만, 좋은 경험이 되었다. 팀원들도 생업과 수료 프로젝트를 진행하느라 고생이 많았다. 주어진 시간에 모두들 힘을 모아 완료한 것에 의미를 두고 싶다. 프로젝트의 원활한 진행을 위해 협조해 주신 데이터베이스진흥원 담당자들께 감사 드린다.

더 밝고 건강한 사회를 위한 시도

미국의 범죄학자인 제임스 월슨과 조지 켈링이 1982년 3월에 생활 환경이 어수선한 동네에서 범죄가 많이 발생한다는 ‘깨진 유리창 이론’을 공동 발표했다. 이 이론을 따라 지자체별로 낙서, 깨진 유리창, 쓰레기가 방치된 장소 등의 위치를 주기적으로 기록하고 개선하면서 범죄율도 내려갔다고 한다. 우리팀은 R과 D3.js로 시각화해 범죄예방 설계에 관한 분석을 제시했다. 1.5개월이라는 짧은 기간에 여러 팀원이 프로젝트에 힘을 쏟았지만 여전히 아쉬움도 남아 있다. 좀 더 다듬어 멋진 결과물을 선보이고 싶다. R 분석을 이상민 팀원이 주도하고 정용주 팀원과 최기웅 팀원이 지원해 더 멋진 결과물을 얻었다. 데이터의 정제와 D3.js를 사용한 시각화에도 힘을 쏟아 짧은 기간에 기대 이상의 결과물을 얻어서 보람이 커다.

CCTV, 경찰서와 같은 전통적인 범죄예방 시설에서 더 나아가 환경을 개선해 더 안전한 생활 터전을 만들기 위한 범죄예방 설계를 셀테드(CPTED)라고 한다. 범죄율이 높은 지역에 놀이터, 공원, 체육시설을 설치한다든가 조명의 색 또는 시설물의 색상(노란색 등) 변경, 낙서 제거만으로도 범죄율을 줄일 수 있다. 이를 오픈소스 플랫폼에서 R과 D3.js로 분석·시각화해 사회 안전에 대한 프로젝트를 좀 더 깊게 진행할 수 있을 것이다.

더 나아가 이 프로젝트는 범죄 취약계층의 주거지 또는 어린이 보호구역의 감시 강화나 안전보행로 구현 등에 적용할 수 있

으로 확장성이 높다고 본다. 데이터 분석 관점에서 더 안전하고 쾌적한 삶의 터전을 마련하는 데 도움을 주기 위한 작은 시도였다. 생활 환경 개선과 더불어 체계적인 감시 시스템을 운영해 우발적이고 잠재적인 범죄 발생을 줄일 수 있다고 생각한다.

불안한 이 시기에 다 함께 힘을 합쳐 자발적으로 스마트폰 카메라를 사용한 ‘스마트 데이터’를 구축하는 것도 시도해 볼 만하다. 생활 주변의 정비되지 않은 시설물 DB를 구축해 단계별로 환경을 개선하고 발전시켜 나간다면 더 밝고 건강한 사회가 될 것이다. 정부나 지자체의 역할도 중요하지만, 시민이 자발적으로 참여해 개선 방안을 제시하면서 정보를 공유하는 것도 더불어 사는 사회의 모습이지 않을까.

두려움이 즐거움과 환희로 바뀐다

시작이 반이라 했다. 빅데이터 교육을 놓고 주위에도 망설이는 사람이 더러 있었다. 망설이는 가장 근본적인 이유는 금전적 부담과 생업 때문이었다. 필자도 그런 고민을 했지만 과감히 도전했고 정말 좋은 경험을 했다. 더욱이 빅데이터 아카데미의 교육과정은 무료이고, 교육 내용까지 알차다. 지도 강사의 학력과 실력은 궁금증과 모자란 부분을 채워주기에 충분하고도 남았다.

웠다. 프로젝트를 수료한 후, 지인이 근무하는 어느 한 기업으로부터 강의 요청을 받고 쉬지 않고 3시간 넘게 소개해 드렸던 즐거운 기억도 얻었다. 교육을 받기 전에 고민했던 당시의 막연함과 두려움이 즐거움과 환희로 다가왔다. 

“어려운 문제가 풀렸을 때의 기쁨으로 잠 못 이루다”



김대훈 팀장
프리랜서, BI 솔루션 엔지니어

‘빅데이터 플랫폼에 필요한 기본 기술을 총동원해 다시피 했다’는 심사평을 들었다.

빅데이터 인프라가 궁금했으므로 하나씩 직접 살펴보고 싶었다. 직장에서도 새로운 툴을 설치해 빨리 적용해야 하는 일을 하고 있기에 낯선 기술을 두려워하지 않는 편이다.

환경 구성 자료에 나온 기술들을 모두 적용했다.

아니다. 설치해 기능을 확인하기는 했지만 프로젝트 서버에서 그 많은 패키지를 모두 수용할 수 없었다. 다양한 기능을 일단 정지시켜놓고 플럼과 하이브, R 등 핵심 기술을 위주로 썼다. 향후 이 프로젝트를 발전시키면서 모두 적용해보고 싶다.

‘범죄 예방’이라는 주제가 좀 특이하다.

개인적으로 안전에 대해 민감한 편이다. 집을 구할 때도 (공) CCTV 위치와 외부 침입으로부터 안전한 구조인지를 면밀히 살펴볼 정도다(웃음). 프로젝트 주제로 ‘네트워크 로그 분석’을 하자는 제안도 나왔는데, 팀원들이 범죄 예방이라는 주제를 더 흥미로워해 최종 주제로 선정했다.

공공 프로젝트에 적합한 주제로 보인다.

그럴 수도 있다. 서울을 기준으로 놓고 볼 때, 구청별로

CCTV를 설치·운영해 이론적인 방법 취약지구는 이미 파악하고 있을 것이다. 하지만 CCTV 설치 대수가 범죄율과 직결되지 않은 걸로 알고 있다. 우리 팀에서 수행했던 프로젝트는 빅데이터 분석의 장점을 활용해 수치로 나타난 데이터의 허점을 뛰어넘고 싶었다. ‘상권, 유치원, 학교, CCTV, 범죄발생 위치’ 등의 키워드들 사이의 상관관계 속에서 방법 사각지대를 찾고 싶었다. 범죄 예방을 위해서는 CCTV 설치 대수 확대가 능사는 아니라고 생각한다.

프로젝트를 진행하면서 기억에 남는 것이 있다면.

기술적으로 어려움에 봉착했던 적이 있었다. 며칠 동안 잠이 부족한 상태에서 RHIVE와 R을 연결하려 아무리 해봐도 되지 않았다. 하루가 넘게 다양한 자료를 참고해서 시도해 봐도 안 됐다. 몸도 너무 피곤한 상태에서 생각했던 대로 일이 풀리지 않아 ‘그냥 넘어가야겠다’는 생각으로 잠자리에 들었다. 하지만 잠이 오지 않았다. 여기서 멈추면 그 동안 고생했던 것이 물거품이 될 거라는 생각에 다시 일어나 시도했다. 놀랍게도 연결되는 게 아닌가! 너무 기뻐서 R 개발을 담당하는 이상민 팀원에게 밤 10시쯤에 연락했는데, 그가 기다렸다는 듯이 개발에 들어간다고 바로 응답 메시지를 보내왔다. 어떻게 개발하고 있는지 너무 궁금해 새벽 4시에 일어나 보니 “R로 좌표의 위치를 표시했다”는 반가운 카카오톡 메시지가 도착해 있었다. 바로 바통을 이어받아 아침 7시 30분까지 작업을 했을 때, 정말로 쌓였던 피로가 사라지면서 날아갈 것 같았다. 팀원 간에 잠이 들고 깨는 타이밍까지 척척 맞아 떨어져 신기할 정도였다.

향후 발전 계획은.

지속적으로 업데이트해 하나의 과제로 완성해 나가고 싶다. 더불어 수료 프로젝트에서 분석 도구 R의 기능을 제대로 적용하지 못한 것이 아쉬웠다. 기술 전문가 과정이었으므로 분석에 포커스를 덜 뒀지만, R에는 그동안 상상할 수 없었던 신기한 기능이 정말 많았다. 긴 코드가 필수적인 부분을 코드 몇 줄로 처리하는 등 매력적인 기능이 많았는데 직접 하나씩 배워 볼 계획이다. ☺

방범시설과 범죄와의 상관관계 분석 시스템

프로젝트 소개

범죄로부터 안전 사각지대를 찾아내 사전에 범죄 예방조치를 취할 수 있도록 범죄율과 사회안전 기반 시설과의 상관관계를 분석한 프로젝트. 사회안전 기반시설(경찰서, 지구대, 방범대, CCTV)과 교육기관(학교, 유치원)의 위치 데이터를 토대로 분석했다.

구분

기술 전문가 과정: 상관관계 분석

프로젝트 기간

2014년 07~08월

멘토

신탁길(스냅데이터 대표)

작용도구

클라우데라 CDH5, 플럼, 하이브, RHIVE, R, JSP, Informatica DataTransformation

수집 데이터

트위터 데이터, 지자체에서 받은 CCTV · 상권 · 경찰서 · 학교 등의 위치 데이터

산출물

WBS, 설치 가이드, 중간 보고서, 종료 보고서

교육 참여형태

자발적 참여(5) / 회사 권유(0)

진행

• 김대훈 팀장	환경구성, 자료수집, 개발, R분석	프리랜서	경력 10년
• 이상민 팀원	환경구성, 자료수집, 정책 문제 해결, R 분석	SI업체 차장	경력 15년
• 정용주	자료수집, R 분석	SI업체 부장	경력 15년
• 최기웅	R분석	SI업체 차장	경력 15년
• 김마선(가명)	자바 개발	SI업체 차장	경력 10년

빅데이터 아카데미 수강 후 변화

수강 전	<ul style="list-style-type: none"> 하둡은 맵리듀스가 필수인지 알았다. UI 없이 콘솔 창에서만 개발하는지 알았다. 하둡, 하이브, 플럼 외의 모듈은 몰랐다. 맵리듀스를 직접 개발할 필요가 없음을 알았다. 에코시스템에서 편리한 UI를 제공해 주었다. 목적에 맞는 다양한 모듈이 있음을 알았다.
수강 후	





자녀 교육특구 모델 찾기

워킹맘은 데이터로 자녀 교육특구를 찾는다



글 김은희 삼성SDS 책임

여섯 살, 세 살배기 두 아이를 키우는 워킹맘으로 빅데이터 아카데미 기술 전문가 과정에 참여했다. 1~2년 전까지만 해도 필자의 관심사는 ‘아이들에게 어떤 좋은 음식을 먹일까, 어떻게 하면 감기에 걸리지 않게 할 수 있을까?’였다. 하지만 얼마 전부터 이런 관심사가 바뀌었다. 아이들 교육으로 눈길이 가기 시작했다. 자녀에게 좋은 교육 환경을 마련해 주고 싶은 마음은 모든 부모의 공통적인 바람이자 관심사가 아닐까? 이런 고민을 하고 있을 무렵, 한국DB진흥원의 빅데이터 아카데미를 알게 되었다.

CHALLENGES

호기심으로부터 시작

‘빅데이터라는 큰 물결이 몰려온다’는 소식을 접했을 때, 두려움보다는 호기심이 앞섰던 필자에게 빅데이터계(^^)에 첫 발을 내딛게 된 좋은 기회가 찾아온 것이다. 교육은 예상보다 재미 있었다. 강사들의 화려한 이력도 호기심의 대상이었고, 각양각색의 분야에 근무하면서 빅데이터 교육을 듣기 위해 무려 2주 동안의 실무를 뒤로하고 교육장에 앉아 있는 동기 교육생들도 흥미로웠다.

‘말이 씨가 되다’

집체교육 중에 팀원들과 함께 식사를 하던 중, 필자가 불쑥 꺼낸 한마디가 수료 프로젝트의 주제로 연결되기에 이르렀다. “요즘 제 걱정거리는 내년에는 어디로 이사가야 할지예요. 강남이나 목동이 교육환경이 좋다고 하지만, 그 중에서도 특히 어떤 아파트가 좋은지도 모르고, 그 동네 아파트 값이 너무 비싸서 좀 저렴하면서도 좋은 곳이 없는지 등 의외로 상세한

정보 찾기가 쉽지 않았어요. 원하는 교육 조건을 입력하면 순위별로 추천지역이 나왔으면 좋겠는데….” 뉴두리 투로 무심코 던진 필자의 한마디에 ‘우리, 교육 특구 모델을 만들어보면 어떨까요?’ 하고 한 팀원이 맞장구를 쳤다. 나머지 팀원들이 이에 적극 동조하면서 프로젝트 주제로 선정하기에 이르렀다.

SOLUTION

하둡이 능사는 아니다

빅데이터 아카데미 수료 프로젝트 성공사례마다 항상 등장하는 것이 데이터 확보 문제였다. 학교·아파트·주변시설 정보 등과 더불어 시설별 거리 계산을 위한 좌표 데이터가 필요했다.

데이터 찾기의 어려움을 익히 알았기에, 우리팀은 데이터 찾기는 최대한 단기간에 마무리하는 것을 목표로 했다. 다행히 팀원 중 한 명이 열심히 노력해 준 덕분에 원하는 데이터 찾기는 예상보다 쉽게 끝났다. 하지만 찾아 놓은 데이터를 분석해 보

유아 및 초등학생을 둔 부모의 교육환경 선택에 대한 잠재적 욕구 해소



이번에 이사를 가야하는데
우리 아이들 교육 환경이 좋은 곳으로
가고 싶은데 어디로 가야 할까?

니 흔히 알고 있는 위경도 좌표가 아닌 KATEC 좌표다. 이는 주변 분석을 위해서는 특정 지점을 기준으로 거리를 구해야 하는 좌표다.

팀원 중 한 명이 자바 프로그래머였다. 열심히 구글 검색을 하여 KATEC 좌표를 위경도로 전환하는 프로그램을 개발해, 무사히 모든 시설에 대한 위경도 값을 얻을 수 있었다. 이제 거리 계산이 남았다. 이미 데이터는 하둡에 올려 둔 상황이었으므로 거리 계산을 위해 맵리듀스(MapReduce) 프로그램을 짜기로 했다. 하지만 팀원 가운데 맵리듀스를 접해본 사람이 없었다. 몇 번의 실패를 거듭하고 시간이 흘러가자 대책을 세우기에 이르렀다. 발상의 전환이라고 해야 할까? 대용량 데이터를 분석하더라도 꼭 하둡에서 해야 한다는 법은 없다. 결과와 처리 시간이 중요하다는 생각으로 데이터를 다시 MySQL에 올리기로 했다. MySQL 함수를 이용하면 거리를 쉽게 계산할 수 있다는 것을 미리 알아들은 터였으므로 바로 적용해 보았다.

생각했던 것처럼 간단히 끝났다. 무사히 거리 계산까지 마무리하고, 원활한 분석을 위해 통합 테이블도 구성했다. 처음부터 아파트 단지를 기준으로 주변을 분석하려고 했다. 하지만 우리팀이 확보한 데이터는 특정 격자 지역의 아파트 개수였고, 아파트별 좌표 및 시세 현황 데이터는 유료였다. 수작업으로 하나씩 데이터를 모을 수는 없는 일이다. 동일한 분석 패턴일 터이므로 이미 확보한 학교 데이터를 기준으로 분석하기로 했다.

데이터세트 완료

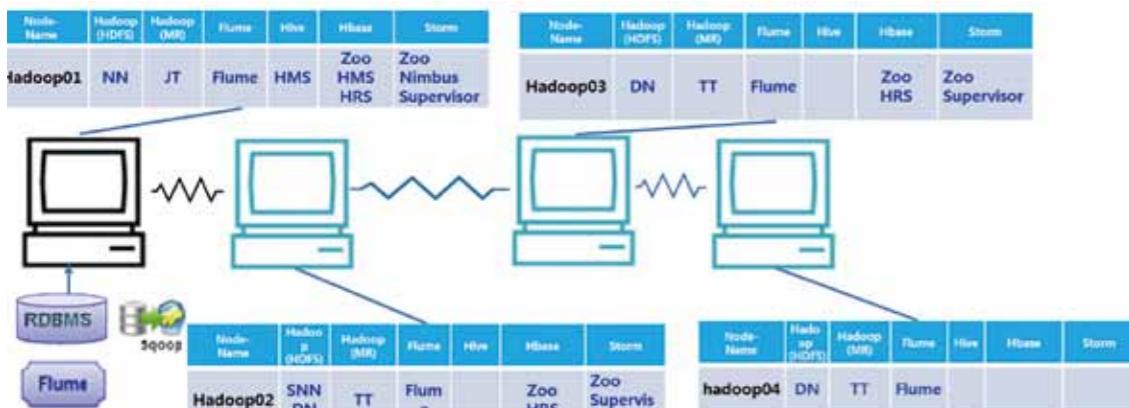
서울시 600여 개 학교를 기준으로 전국의 주변시설 정보를 루핑(Looping)해 두 지점 사이의 거리를 구했다. 이때부터 데이터 용량이 급격하게 늘어나기 시작했다. 이제부터 빅데이터 인가? 이 전까지는 빠른 분석을 위해 구해 놓은 두 지점 사이의 거리를 통합 테이블로 저장해 놓은 정도였다. 여기다 항후 대외적으로 서비스를 하면, 실시간 데이터를 추가해야 할 수도 있다. 이렇게 하려면 하둡 클러스터링과 실시간 처리 로직도 필요할 것이다.

이로써 데이터세트가 완료됐다. 이제 분석을 해 본 팀원이나 설 차례다. 우선 기초 데이터 분석을 해 보았다. 서울시 학교 현황, 학교 순위별 분포, 상하위 톱5 학교, 여기서 상위 학교 주변의 특징 등을 분석해 보았다.

데이터 분석의 미궁에 빠지다

하지만 1차 데이터 분석 결과는 참담했다. 인사이트를 확보할 수 있어야 하는데, 이렇다 할 것이 보이지 않았기 때문이다. 원본 데이터가 이상한가? 주변시설과 학교성적의 상관관계가 조금 약한 사립초등학교가 다수 포함되어 있기 때문이라고 결론을 짓기에 이르렀다. 대부분의 사립초등학교 학생들은 스쿨버스를 이용한다는 사실을 분석에 적용하기로 했다.

그렇게 해 보아도 데이터 분석의 미궁으로부터 벗어날 수 없었다. ‘꿈보다 해몽’이라고 이대로 마무리 하고 분석 결과 요약



●그림 2) 하둡 v1.2.1 에코시스템 구축



●그림 3) 데이터 정제와 통합

을 그럴듯하게 하고 넘어가야 하나? 수료 프로젝트가 이상할 정도로 순탄하게 진행되는가 싶었는데 분석에서 숨은 복병을 만나 앞으로 나아가지를 못했다. “군집분석을 해보면 어떨까요?” 정기 미팅에 참여해 주신 멘토의 이 한마디가 해매던 우리팀의 좌표가 되었다.

왜 그렇게 조언하는지 의미를 몰랐던 우리팀은 무조건 따라 해 보기로 했다. 다른 방법이 없었으므로 절대적으로 믿고 따를 수밖에.

처음부터 다시 시작해 학교순위를 크게 5~6가지 클래스로 나눠 등급을 부여하고, 등급별 주변 시설을 분석하기 시작했다. 와~ 처음과 달리 뭔가 보이기 시작했다. 예수님이 ‘당신의 고향에서는 환영을 받지 못했다’ 듯이, 늘 우리들 곁에서 듣기만 하던 멘토였으므로 그의 경험과 노하우를 은연중에 잊고 있던(?) 하룻강아지들이었던 셈이다^^ 역시 경험 많은 멘토님!

“군집분석을 해보면 어떨까요?”

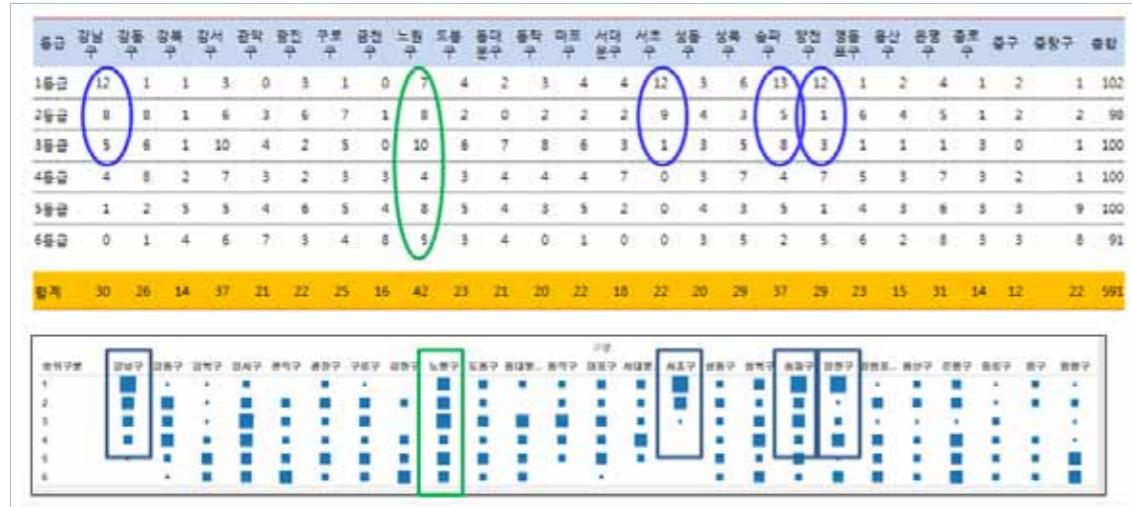
우리팀은 말 그대로 정말 기술자 집단이었다. 기술자 중심의 팀답게 분석 인프라는 며칠만에 똑딱 구축을 끝냈다.

수료 프로젝트에서 사용할 분석환경이든, 아니든 교육 때 배웠던 환경은 복습차원으로 모두 재설치하면서 빅데이터와 관련된 모든 기술은 한번씩은 직접 맛보기로 했다. 기술 전문가가 되겠다고 한 만큼 두려움 없이 이거저거 어떻게 돌아가는지는 확인해 보았다.

머리를 맞대면 해결책이 나온다

데이터 시각화 차원에서 스플렁크(Splunk)와 타블로(tableau)도 알아보았다. 다운로드 사이트를 찾아가서 설치하고, 기본 개념에 대한 동영상을 듣고, 여기저기 구글링하고, 지도에 좌표를 찍어보고, 그래프도 그려보았다. 하지만 기대한 만큼 예쁜 결과를 도출하지 못했다. 역시 짧은 시간에 완성도 있는 결과물은 쉽지 않다. 수료 프로젝트 후 제대로 한번 도전해 볼 계획이다. 어찌 보면, 빅데이터 아카데미는 데이터 분석 공부 방법을 파악하는 과정이었는지도 모르다. 이 차원이라면, 우리팀은 각자 소기의 목표를 달성한 셈이다. 팀원마다 관심 분야가 조금씩 다르지만, 각자 더 자세히 공부해 보고 싶은 분야 몇 개씩은 찾았으니 말이다.

모든 교육생이 동일하게 경험했던 어려움이었겠지만, 우리팀



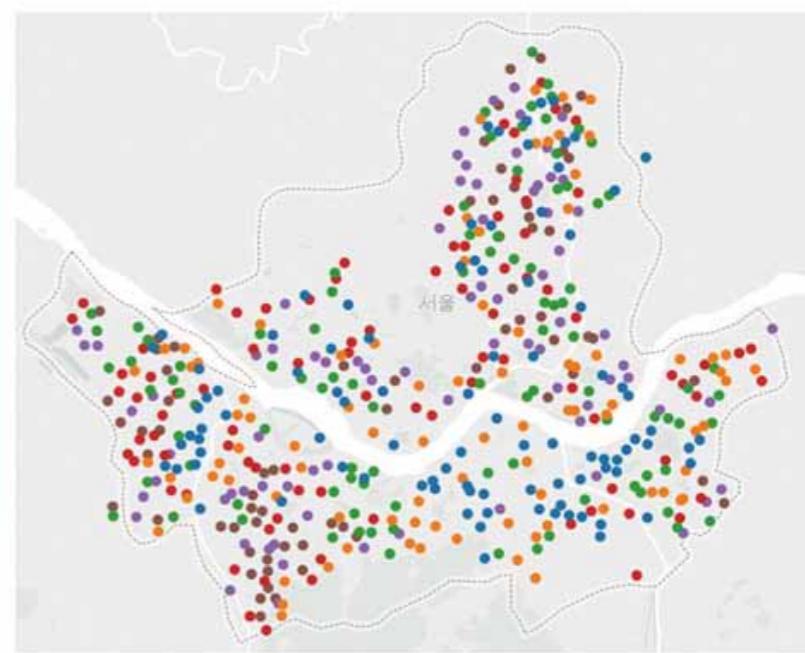
●<그림 4> 데이터 분석 : 강남구, 서초구, 송파구, 양천구에 상위 등급의 학교들이 많음 / 노원구는 다른 지역과 달리 등급이 골고루 분포되어 있음

에게도 제한된 시간에 끝내야 한다는 점이 제일 부담스러웠다. 팀원의 대부분이 각자의 직장에 돌아가 충족적 역할을 수행해야 하는 중견 직원들이었다. 업무시간 이후, 밤, 주말 짬짬이 시간을 내어 진행해야 했다. 혼자서 할 수 있는 일이 아니었기 때문에, 각자 맡은 바를 미리 준비해 와서 점검 · 협의하는 형태로 진행했다. 열정이 없으면 완료하기 어려운 숙제를 나눠서 한 셈이다. 매우 열정적으로 참여했던 팀원 가운데 한 명이 수료 프로젝트 시작과 동시에 지방 프로젝트에 투입돼서 나머지 팀원들을 긴장하게 했다. 이 팀원은 온라인으로 힘을 실어주었다. 또 다른 한 명의 팀원은 긴급 프로젝트에 투입돼서 자신은 물론, 주변 팀원을 안타깝게 했다.

'남에게 도움은 되지 못해도 피하는 주지는 않겠다'는 중견 IT 엔지니어로서 자존심의 발로였는지, 그 팀원은 앞이 보이지 않아 망연자실해 할 때, 슬그머니 해결 포인트를 제시하기도 했다. 수료 프로젝트의 가장 큰 산출물은 아마도 데이터 분석을 할 때 팀 플레이의 중요성을 알게 된 것이 아닐까 한다. 하나 더하기 하나는 둘이 아니라 100이 될 수 있음을 분석 과제를 하면서 몸소 체험했다.

'매우 빼먹지 않고 오프라인 모임을 갖는 팀'이라는 이유로 다른 팀과 멘토의 눈길을 끌지 않았나 싶다.

얼굴을 맞대고 찾아보면 뭔가 나오겠지 하는 적극적인 생각으로 숙제를 했든 못했든, 앞이 보이든 안 보이든 우선 만나서 함께했던 것이 좋은 결과를 낸 밑거름이었다고 생각한다. 정말



●<그림 5> 타블로로 데이터 시각화

바쁜 중에도 매주 화요일 DB진흥원 강의실에 모여 도시락을 먹으며, 함께 문제를 풀어내는 과정이 기억에 남아 있다. 힘들면 직장에서, 가정에서 있었던 일상대화를 나누는 과정에서 서로 힘을 얻기도 했다. 직장 생활을 많이 해 본 사람이라면 공통적으로 느끼겠지만, 적당한 수다는 팀 업무에 무척 큰 힘이 된다. 이것을 요즘은 '소통'이라고 한다. 물론 디지털 커뮤니케이션을 무시할 필요는 없다. 카톡, 온라인 카페, 메신저, N드라이브 등 온갖 온라인 매체를 필요에 따라 적극 활용했다.

CONCLUSION

이번 프로젝트는 여러 가지 상황상 실제로 구현하지 못한 부분도 있다. 알고 싶은 지역 혹은 Map 위치를 클릭하면 주변 학교 등급별 세부내역이 출력되는 WebUI 개발이 필요하고, 시각화 툴이나 스플렁크 등에 대해서도 더 공부해 보려고 한

다. 더불어 우리팀은 상용 모바일 버전을 선보일 계획으로, 팀원들이 계속 모임을 갖고 있다. 생각만 하는 것보다 일단 부딪쳐 보는 과정에서 발전이 함께함을 우리는 분명히 확인했다.

뛰어들어 본 자만이 안다

이제 우리팀원 모두는 빅데이터 바다에 풍덩 뛰어들었다. 데이터 분석이 뭔지 궁금하여 이 자료를 읽는 독자라면, 우리팀처럼 빅데이터 바다에 뛰어들 용기와 재능을 충분히 갖췄다고 믿어도 좋다고 생각한다.

프로젝트 기간 중에 아이들의 숙제도 잘 못 챙겨주고 놀아주지도 못한 나쁜 엄마가 되고 말았지만, '1등을 했다'는 말에 정말 좋아하던 아이들에게 그 날만큼은 세상 최고 멋쟁이 엄마가 되었다. 😊

“패션의 아닌 패러다임을 실감하다”



김은희 팀장
삼성SDS 책임

어떤 계기로 기술 전문가 과정에 지원했다.

생명보험 업체에서 DBA로 일하고 있다. 데이터 분석에 관심을 갖고 있던 참에 개인 역량강화 차원에서 회사에서 배려해 줘서 참여했다.

교육을 받기 전후 달라진 점이 있다면.

데이터 분석에 대한 기반 지식이 거의 없어서 따라가기에 벅찼다. 데이터 분석에 대한 공부 방법을 제대로 배웠다는 데 의미를 두고 있다. 수료 프로젝트를 진행하면서 집체교육 때 배웠던 기술을 직접 적용해 보려고 노력했다. IT 업무를 해 본 사람이라면 알겠지만, 익숙하지 않은 기술로 제한된 기간에 결과를 도출한다는 것은 말처럼 쉽지 않다. 그래서 배웠던 기술이 이런 것이구나, 이렇게 하면 되겠구나 하는 정도로 느낀 다음, 멘토에게 결과를 그때그때 보고했다. 실제 분석은 익숙한 SQL을 많이 활용했다.

팀장이 된 배경이 궁금하다.

회사에서 중간 직급의 사원으로서 일하고 있는데, 수료 프로젝트 팀에서는 막내였다. 그것도 여자는 한 명뿐이었다. 가장 고참 팀원께서 ‘팀장이 되면 팀원들이 적극 참여하도록 적극 돋겠다’고 하여 열렬결에 ‘하겠다’고 답하고 말았다. 8주 간

의 수료 프로젝트를 하면서 매주 KODB에서 만났는데 놀랍게도 참여율이 매우 높았다.

가장 어려웠던 점은.

분석 결과 데이터를 처음으로 봤을 때다. 원가 의미 있는 결과가 나올 것이라고 기대했는데, 그저 그런 결과가 나와서 팀원들의 사기가 급락했다. 어느 순간 포기하는 분위기로 흘러갈 것만 같았다. 이때 빠지지 않고 참여해 준 팀원들을 위해서라도 팀장인 내가 힘을 내야겠다고 생각했다. 데이터 시각화를 할 때였다. 5시에 퇴근하는 회사에 다니고 있어서 퇴근 후 시간을 활용해 시각화를 직접 처리했다.

분석에 더 역점을 둔 것이 아닌가.

기본적인 하둡 인프라 구축은 실제로 IT 분야에서 일하는 사람이라면 그리 어렵지 않게 할 수 있다. 요즘 들어서 분석 플랫폼 구축보다 분석에 관심이 집중되고 있다고 한다. 꼭 새로운 기술로 분석해야 할 이유도 없다. 필요할 때 적용해도 된다고 본다.

우수 프로젝트팀으로 선정된 결정적인 계기가 무엇이었다고 생각하나.

팀원들이 함께 데이터 분석과 관련한 다양한 기술을 적용하려 노력한 것에서 높은 점수를 얻지 않았나 싶다. 다른 팀에서 참신한 주제를 많이 제시했고, 발표도 참 잘해서 ‘우리팀은 열심히 했던 것에 의미를 두자’고 생각했는데 1등 팀이라고 하여 놀랄과 기쁨이 교차했다.

가장 낯설게 느껴졌던 것은.

다른 조에서는 분석 경험자가 있거나 R을 잘 다루는 팀원이 있었다. 하지만 우리 팀원 모두가 R을 낯설어했다. 나중에는 R을 써야 하는 이유와 왜 통계학이 필요한지를 알게 됐다. 기회가 된다면, 패턴 분석 등 데이터 분석을 제대로 한번 공부해 보고 싶다. 분석 플랫폼은 혼자서 공부할 수 있을 것 같은데, 분석은 누군가로부터 배우는 것이 훨씬 효율적이라고 생각한다.

자녀 교육특구 모델 찾기

프로젝트 소개

자녀 교육특구 예측모델을 찾는 프로젝트. 희망 지역의 아파트를 선택하면, 주변 환경을 분석해 정보를 보여 주는 소비자 맞춤형 서비스 제공이 가능하다.

구분

기술 전문가 과정

프로젝트 기간

2014년 09~10월

멘토

이상훈(SK C&C 대리)

작용도구

하둡, 하이브, MySQL 스플렁크, 타블로, R

수집 데이터

위치 데이터(유치원, 어린이집, 학교, 정류장, 편의점, 병원, 아파트, 은행), 지자체별 인구(직장인, 추정소득분위), 토지 데이터 등(DB 스토어 및 공공데이터, 구글링 등 포털 사이트에서 검색·취합)

산출물

- 교육 특구 예측모델
- 아파트·학교별 주변 시설정보

교육 참여형태

자발적 참여(6) / 회사 권유(0)

진행

• 김은희 팀장	데이터 분석	SI업체 책임	경력 13년
• 김성표 팀원	데이터 분석	리서치업체 실장	경력 15년
• 노병희 팀원	데이터 수집	통신장비업체 부장	경력 14년
• 노태상 팀원	—	—	—
• 전성종 팀원	데이터 변환	마케팅업체 본부장	경력 23년
• 최혁근 팀원	분석환경 구성	SI업체 본부장	경력 20년

빅데이터 아카데미

수강 전 • 낯선 기술을 적용하는 것이 데이터 분석이다

수강 후 변화

- 일단 부딪쳐 보라.
- 알고 있는 기술을 적용할 수 있다.
- 개인 역량 강화에 대한 확고한 방향성 설정



현재를 알고 미래를 읽는 즐거움!

Big Data, Bigger Opportunities And The Biggest Value!



빅데이터 시장은 2015년에만 세계적으로
18조 2000억 원, 한국은 약 300억 원 규모로 형성 전망



빅데이터 전문가 수요 급증과 경력자 부족 등으로
전문인력 수급 난항, 사회적 비용 증가 예상



미국은 '빅데이터 R&D 이니셔티브'를 발표하고 빅데이터 인력양성 등에
집중 투자, 한국의 빅데이터 인력양성 체계는 시작 수준



국내에서도 '신직업 발굴·육성 추진방안' 발표, 빅데이터 전문가를
신직업으로 정의, 신규 일자리와 고부가가치 창출 도모



미래창조과학부 등 관계부처 합동으로 데이터 전문인력 양성과
일자리 연계를 골자로 한 '빅데이터 산업 발전전략' 발표



빅데이터 전문가 양성 체계 수립 및 실무 전문가 양성을 통해
국내 빅데이터 기술 경쟁력을 강화하고 글로벌 시장 선점의 토대 마련 필요



오늘보다 나은 미래를 위한 선택

교육 대상

과정	내용
기술 전문가	<ul style="list-style-type: none"> 대상 : 개발자, DBA, SE 등 3년 이상 업무 경력자 선발 우대조건 <ul style="list-style-type: none"> - 빅데이터 프로젝트 수행인력 또는 예정 인력 - 하둡(MapReduce, HDFS), NoSQL 및 캐싱기술 유경험자 - 분산 파일시스템 또는 분산 데이터베이스 관리 유경험자
분석 전문가	<ul style="list-style-type: none"> 대상: CRM, 마케팅 및 기획 등 3년 이상 업무 경력자 선발 우대조건 <ul style="list-style-type: none"> - 빅데이터 프로젝트 수행인력 또는 예정 인력 - SAS, SPSS, R 등 통계분석 틀, 데이터 마이닝 유경험자 - SQL, OLAP, Query, Reporting 도구 유경험자

참여 방법(연수생 선발 절차)

구분	내용
수강신청	http://bigdata.dbguide.net 에서 원하는 교육과정을 신청
1차 선발	직무 및 업무경력 적합성 평가
2차 선발	프로젝트 경력 및 상세기술 온라인 질의 평가
최종 선발	2차 선발인원에 한하여 재직증명서 및 학약서 등 서류 제출

2014년 빅데이터 아카데미 연수생 현황

구 분	빅데이터 기술 전문가	빅데이터 분석 전문가	합 계
수료 인원	100명	101명	201명
교육생 연령	평균 39.8세	평균 40.1세	평균 39.9세
업무 경력	평균 11.3년	평균 10.7년	평균 11.0년
기업 분포	대기업	15%(15명)	16.8%(17명)
	중소기업	79%(79명)	68.3%(69명)
	공공기관	6%(6명)	14.9%(15명)
5대 직무별 분포	기획	12%(12명)	20.7%(21명)
	처리	59%(59명)	18.8%(19명)
	분석	14%(14명)	57.4%(58명)
	시각화	3%(3명)	0.9%(1명)
	운영관리	12%(12명)	2%(2명)
	의료건강	2%(2명)	2%(4명)
6대 분야별 분포	과학기술	36%(36명)	24.8%(25명)
	정보보안	29%(29명)	25.7%(26명)
	제조공정	10%(10명)	6.9%(7명)
	소비거래	3%(3명)	10.9%(11명)
	교통물류	2%(2명)	8.9%(9명)
	기타	18%(18명)	20.8%(21명)
			19.3%(39명)

2014 빅데이터 아카데미 우수 프로젝트 사례집

발행일 2014년 12월

발행처 미래창조과학부

427-140 경기도 과천시 관문로 47, 4동 / www.msip.go.kr

한국데이터베이스진흥원

110-799 서울시 종로구 종로 51 종로타워 19층 / www.kodb.or.kr

편 집 글봄크리에이티브 02-507-2340

디자인 page9 (studio.page9@gmail.com)



2014 우수 프로젝트 사례집

이 사례집은 미래창조과학부의 지원으로 제작되었습니다.



427-140 경기도 과천시 관문로 47, 4동
www.msip.go.kr



110-799 서울시 종로구 종로 51 종로타워 19층
TEL : 02-3708-5303 / www.kodb.or.kr