

제목	[금융1기] 양파 생산량 예측 기반 금융상품 제안		
조회수	2384	작성일자	2018.05.16

금융 빅데이터 융합 전문가 1기

양파 생산량 예측 기반 금융상품 제안


THE CHALLENGES

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


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

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

추천 : 0회 [추천하기](#)



본 칼럼을 읽으신 여러분들의 의견을 댓글로 남겨주세요!
여러분들의 소중한 댓글이 칼럼자들에게 큰 격려가 됩니다.



양파 생산량 예측에 영향을 주는 요인들로 기상 데이터 및 SNS 등 연관 데이터들을 수집했다. 이 데이터에 대한 분석 범위를 정의하고 목적을 1)문제 정의, 2)솔루션 싱킹, 3)데이터 분석과 예측 모델링, 4)연계 금융 서비스 제안으로 잡았다.

THE APPROACH

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


문제 정의

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


양파는 다른 채소류와 달리 수분 함량이 높아 저장에 어렵다. 절단 건조를 통한 양념 원료 이용 외에는 활용 방법이 많지 않아 과잉 생산되면 폐기된다. 수요량을 정확하게 예측·생산하는 것이 효과적이나 기상 외의 변수들로 인해 인위적인 수요·생산 관리가 어렵다. 가뭄, 우박 등 자연재해에 따른 농작물 재해 보험과 같은 금융 서비스가 있기는 하다. 하지만 1년 만기 상품으로 자연 재해의 피해가 적은 해에는 가입율이 떨어진다. 무엇보다 상대적으로 높은 자기 부담비율로 농가에 부담을 주고 있어서 자체 개선 및 대체 가능한 금융 서비스가 필요한 상황이었다. 따라서 문제 정의를, 양파 생산량 예측 데이터의 정확도가 상대적으로 낮아 양파를 포함한 농작물 작황 부담을 경감시키는 금융 서비스의 필요성 대두로 정리했다.

솔루션 싱킹

추천 : 0회 [추천하기](#)



본 칼럼을 읽으신 여러분들의 의견을 댓글로 남겨주세요!
여러분들의 소중한 댓글이 칼럼자들에게 큰 격려가 됩니다.



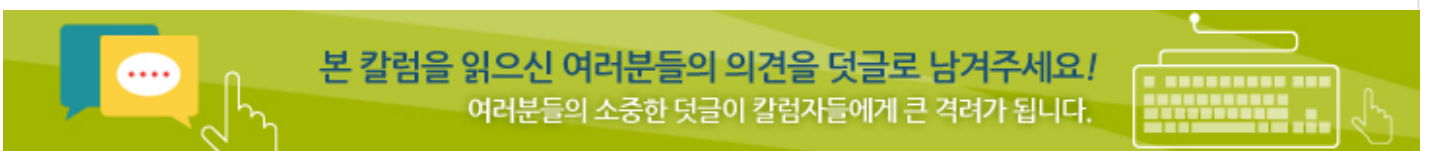
양파 생산량 예측의 정확도를 높이기 위한 예측 모델을 구축하기 위해 이에 영향을 주는 초기 데이터 수집 정의가 중요했다. 이상 현상을 예측하기 위해, 기상 데이터를 포함한 복합적인 요인들이 예측 모델에 영향을 주는 바로 가설을 세웠다. 이에 따라 기상청에서 구축한 자체 빅데이터 분석 인프라를 활용해 평균 온도, 최고 온도, 일교차, 강수량, 일조시간, 재배면적 데이터 항목을 도출했다. 양파 생산량 자체의 패턴 추이와 지역별 양파 재배 면적 대비 재배 비율을 추가했다. 이로써 양파 생산량 예측모델을 수립하는 데에 신뢰도를 높이고자 했다.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

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

도메인 룰: 성공적인 양파 재배를 위한 노하우

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

추천 : 0회 [추천하기](#)



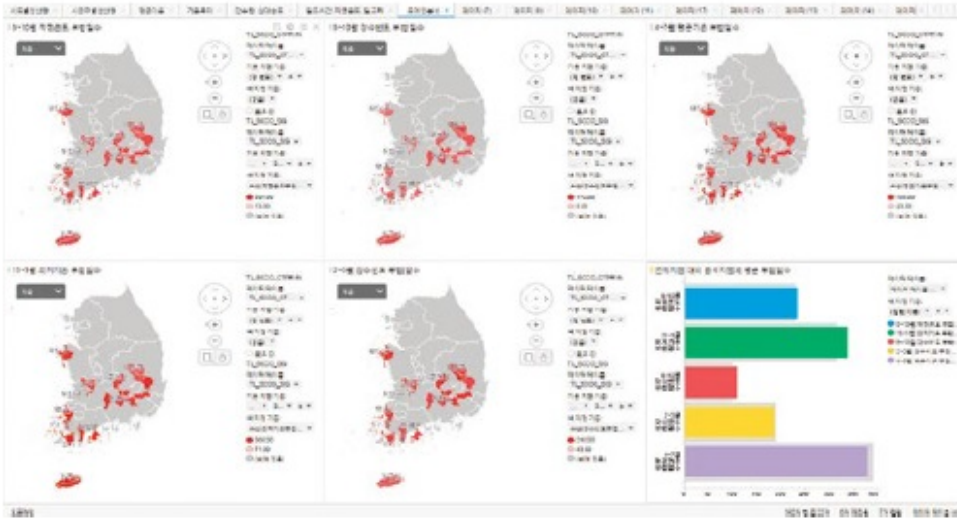


그림 1
기상 조건 일자
빈도수 분석

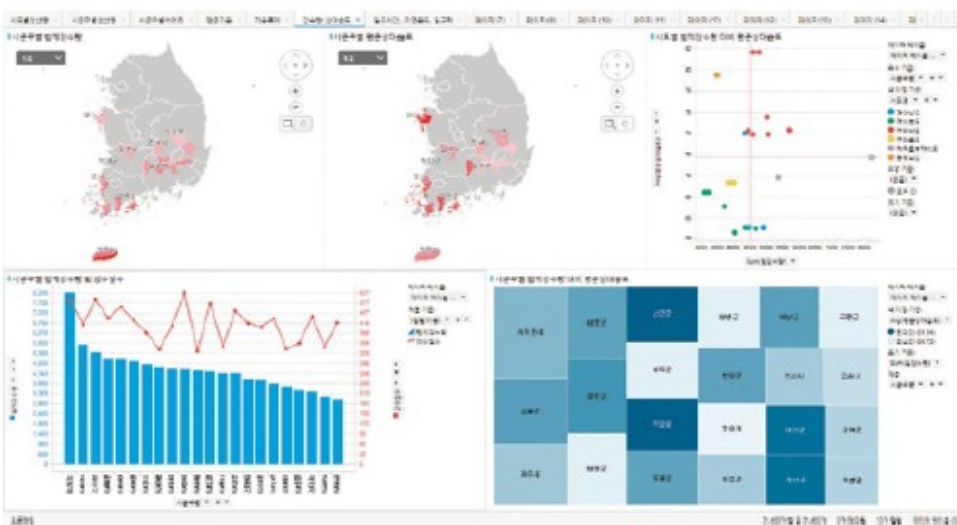


그림 2
시군구별 강수량,
상대습도

데이터 분석과 예측 모델링

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

기상 데이터와 양파재배 면적 데이터를 병합했다. 재배면적 비율만큼 가중치로 곱한 뒤 만든 데이터베이스, 지역별·시도·시군구 온도·일교차·강수량·일조시간·재배면적 합을 계산해 저장한 데이터베이스로 정리했다.

추천 : 0회 [추천하기](#)

본 칼럼을 읽으신 여러분들의 의견을 댓글로 남겨주세요!
여러분들의 소중한 댓글이 칼럼자들에게 큰 격려가 됩니다.

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

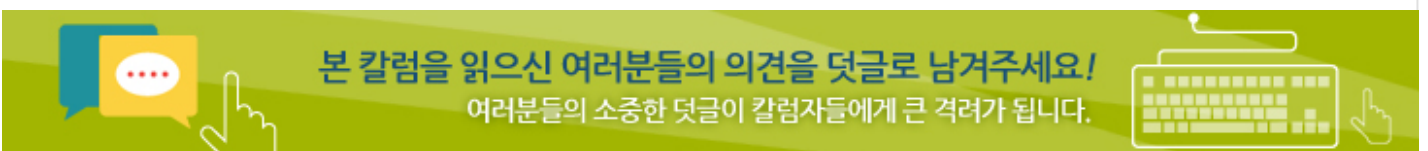
본격적인 데이터 분석과 예측모델을 구축하기 전에 다중공선성 여부를 확인하고자 함수를 적용해 다중공선성 연관관계를 분석했다. 다중공선성이란 입력변수들끼리 연관관계가 존재함으로써 모델의 정확도와 적합성을 떨어뜨리는 문제를 의미한다. 문제가 없음을 확인한 뒤 일반화 선형 분석(generalized linear model)으로 모델을 구축했다.

표 1
최종 도출된
데이터셋 테이블
설명과 구조

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

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

추천 : 0회 [추천하기](#)



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

그림 3

양파 생산량 예측 모델과 학습·검증을 위해 사용한 코드(부분)

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

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

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

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

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

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

일반화 선형 모델을 이용한 양파 생산량 예측 평가

일반화 선형 모델(Generalized Linear Model)을 이용한 양파 생산량의 예측 결과의 오차는 6% 이하로 정확도 검증에 완료했다. 도출된 데이터 예측 모델에 텍스트연관 분석을 추가해 SNS 키워드 동의어·불용어 처리를 했다. 2017년 7월 1일 기준 월간 검색어 기준으로 ‘양파’와 ‘가격’이란 키워드가 1만 7000회를 넘는 검색 횟수를 보여주고 있었다. 양파 가격과 연관성이 가장 높은 키워드 1위로 ‘당근’이 분류됐다. 이는 소비자 입장에서의 당근 가격이 인상될 때 양파 가격 또한 인상될 가능성이 높고, 양파 생산 유통 주기에 따라

추천 : 0회 [추천하기](#)

본 칼럼을 읽으신 여러분들의 의견을 댓글로 남겨주세요!

여러분들의 소중한 댓글이 칼럼자들에게 큰 격려가 됩니다.

생산량이 증가할 수 있음을 의미한다. 시간 여유가 있었다면, 당근 생산량 데이터 예측모 델과 양파의 생산량 데이터 예측모델을 수립해 비교 또는 병합이 가능하지 않 았을까 한다.

THE OUTCOME

2015년 전체 농가별 소득 집계 데이터를 분석한 결과 도시 근로자 대비 평균 소득은 64% 이중 양파 농가 소득은 농가 소득 평균 대비 58%수준에 그쳤다. 보험 가입율은 지속적으로 하락하는 반면, 양파 재배 농가 손해율은 타 작물 농가 손해율 대비 5배 이상으로 확인됐다. 그럼에도 양파 농가는 금융 서비스 의 사각지 대에 놓여 있다고 판단했기에 다음과 같이 세 가지 제안을 수립했다. 첫째, 보험 가입 제고를 통한 농가소득 리스크 헤지(Risk hedge)로써 보험 가입 농가는 재해로 인한 생산량 변동 불안이 내려간다. 보험사 또한 생 산량 예측 데이터를 활용해 이상 징후 발견 시 보험 가입을 위한 적극적인 홍보 활 동을 진행해 수익률을 제 고할 수 있을 것이다.

둘째, 보험상품 계리 심화를 통한 손해율(지급보험금 ÷ 거수보험료×100) 안 정화로써 기존 보험요율 산정 방식에 예측생산 모델링을 적용해 운영·검증함 으로서 손해율이 높은 포도, 복분자와 같은 타 작물 품목의 보험상품으로도 확 산할 수 있다.

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

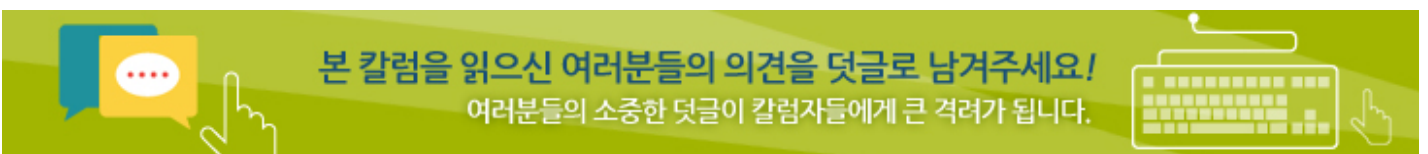
빅데이터 인프라 구축, 엔지니어링 부분도 중요하지만 앞으로 인프라·기 술의 표준화, 가이드라인 대중화로 서 데이터 자체의 가치를 도출해 내는 분석, 이를 지능적으로 해석하고 추론할 수 있는 인공지능이 맞물리 면 데이터 분석 은 지속적으로 발전할 것이다. 민간 영역에서도 데이터 공개와 활용이 활발해 지면, 실제 산 업과 사회에서 응용할 수 있는 데이터 분석 서비스들이 금융 산 업을 비롯해 여러 산업 분야에 큰 파급효과 를 불러올 것으로 기대된다

출처 : 한국데이터진흥원

제공 : 데이터 전문가 지식포털 DBguide.net

추천 : 0회

추천하기



댓글 남기기

댓글 쓰기

입력

한글 300자까지 입력 가능합니다.

* 욕설, 광고, 비방, 도배성 글 등은 자동삭제 대상입니다.
덧글은 한글 300자까지 입력 가능합니다.

이전글	[유통1기] 구매패턴 기반 구매감소 고객 예측
다음글	[분석19기] 기계학습 방법을 활용한 신도시 아파트 가격변동 요인 분석

스크랩

목록