BIZ3199-01 AI비즈니스
김진우 교수님

YONSEI CHICKEN (연세_치킨)

2019.04.29 중간보고서 B1A5
김세정, 김소이, 송민수, 송자영, 유건욱, 정지혜

날씨가 안 좋아도 괜찮아. 우리에겐 치킨이 있잖아?

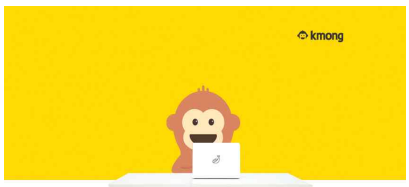
서비스 목적

: 본 서비스는 고객 정보 접근에 어려움을 겪어온 기존 서대문구의 치킨 업계 자영업자들에게 일차적으로 서대문구의 날씨 데이터(강수량, 기온, 미세먼지)에 따른 고객들의 행동 패턴 정보를 분석하여 제공하고자 한다. 이를 통해 서대문구의 치킨업계 자영업자들의 고객 정보를 활용한 마케팅 전략 수립을 돕고자 한다.

기존 유사 서비스의 한계 및 본 서비스의 차별성

1. 기존 유사 서비스

: '연세 치킨'과 유사한 기존의 서비스에는 '배달의 민족 사장님 사이트', '요기요 사장님사이트', '크몽(kmong)'이 있다. 해당 서비스들은 주로 자영업 사장님들을 대상으로 보편적 운영관리법, 마케팅 및 메뉴 전략 등에 대한 정보를 제공하고 있다.



2. 기존 유사 서비스의 한계점

: 기존 유사 서비스의 경우 주로 보편적이고 피상적인 운영관리법에 대한 정보를 제공하고 있다. 예를 들어 이들은 손익관리법, 광고 스팟 설정 방법, 광고 운영, 사이드메뉴 전략, 수수료 정보, 마케팅 전략, 전체 인기 순위와 같은 피상적 데이터분석 결과 등의 정보를 제공하고 있다. 이는 사실상 자영업자들에게 지나치게 포괄적인 정보들이라고 볼 수 있다. 따라서 기존 서비스에게는 실질적으로 필요한 개별 맞춤 정보들을 제공하지는 못하고 있다는 한계점이 존재한다. 뿐만 아니라 기존 서비스들이 세부적이고 전문적인 데이터를 제공하는 대신 수수료가 비교적 부담스러운 가격대에 형성되어 있다는 한계가 존재한다.

3. 기존 유사 서비스와의 차별성

: 기존 유사 서비스와 달리 본 '연세 치킨' 서비스는 자영업자들이 실질적으로 필요로 하는 상대적으로 저렴하고 간단한, 사업별 맞춤 데이터 정보를 제공하는 서비스를 제작하고자 한다. 이를 위해 판매 건당 수수료가 아닌 저렴한 고정 비용만을 설정하여 자영업자들로부터 수익을 창출하고자 한다. 뿐만 아니라 전체 사업 데이터가 아닌 의뢰에 따라 서대문구 내 개별 자영업 치킨 가게들의 데이터를 분석하고자 한다. 본 서비스가 다양한 변수들에 따른 각 데이터를 분석하고 이에 알맞은 간단한 제안 사항을 제공함으로써 자영업자들은 더 발전된 맞춤 전략을 수립할 수 있는 기회를 가지게 될 것이다.

BIZ3199-01 AI비즈니스
김진우 교수님

데이터 수집 방법

1. 독립변수 관련 – 서울 날씨 데이터 :

– 서울 미세먼지 데이터

출처 : 서울 열린 데이터 광장

https://data.seoul.go.kr/search/newSearch.jsp?query=%EB%AF%B8%EC%84%B8%EB%A8%BC%EC%A7%80

– 서울 최고/최저 기온, 강수량 데이터

출처 : 기상 공공데이터 포털

https://data.kma.go.kr/cmmn/main.do

날씨 데이터는 크게 (초)미세먼지, 최저/최고/평균 기온, 강수량을 기준으로 수집.

미세먼지 선정 이유 : 요즘 미세먼지에 대한 관심도가 높아지고, 그만큼 심각성 또한 부각되고 있는 상황이다. 그렇기 때문에 미세먼지의 농도가 높을 때에는 사람들이 외출을 하지 않고 실내에 있는 경향이 늘어나고 있고, 그럴수록 배달 음식에 대한 수요도가 높아질 것이라고 예상했기 때문에 미세먼지 데이터를 수집했다. (실제로 월별 평균 기온과 통화량의 상관관계의 절댓값은 0.8이고 미세먼지는 0.3정도 되었다.)

최저/최고/평균 기온 및 강수량 : 기온 데이터와 강수량 데이터는 계절과 월 단위의 평균적인 날씨들을 단적으로 보여줄 수 있는 가장 기본적인 날씨데이터라고 생각했기 때문에 최저/최고/평균 기온을 수집했다.

– 서대문구 동

https://www.bigdatahub.co.kr/product/view.do?pid=1002056

출처 : SKTelecom bigdata_hub

2. 반응변수 관련 – 치킨 통화량 데이터 :

– 치킨 통화량 데이터

https://www.bigdatahub.co.kr/product/view.do?pid=1002056

출처 : SKTelecom bigdata_hub

치킨 통화량 데이터는 동별 통화건수로 구분되어 나오기 때문에 각 동별로의 치킨 수요량을 확인할 수 있다. 치킨 배달업계는 대부분 통화로 배달을 받기 때문에 통화량은 치킨의 수요량과 같은 역할을 한다고 판단하였기 때문에 치킨 통화량 데이터를 수집했다.

날씨 데이터와 치킨 통화량 데이터 모두 2016년 6월부터 2018년의 데이터를 사용하여 알고리즘 모델 구현을 하는 데에 사용하였고, test set은 2019년 1,2월 데이터를 사용하였다.

사용 알고리즘 과정

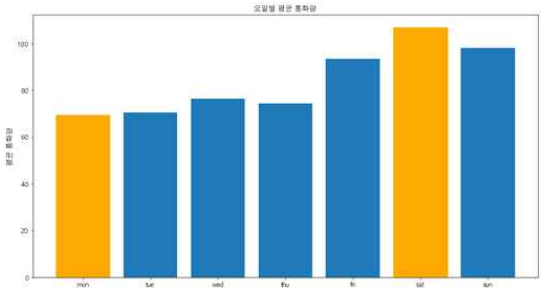
: Linear_Regression (선형회귀모델) / (추가 예상 모델 : Random Forest)

〈Feature engineering〉

1. 미세먼지 데이터와 기온 및 강수량 데이터를 하나의 데이터로 합쳤고,

2. 강수 데이터가 NA로 되어 있는 것은 0으로 대체하여 결측치 처리를 하였다.

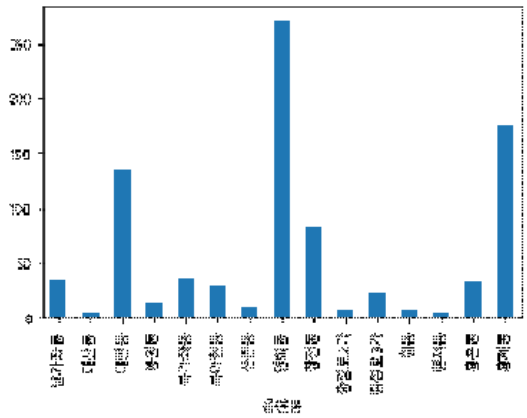
3. 금, 토, 일이 평균적으로 치킨 통화량이 많기 때문에 금, 토, 일을 1, 월~목을 0으로 dummy variable 형태로 변환하였다.



4. 각 동별로 날씨데이터가 동일하기에 생기는 문제를 완화시키기 위해 동을 치킨 통화량에 따라 4개의 그룹으로 묶었다.

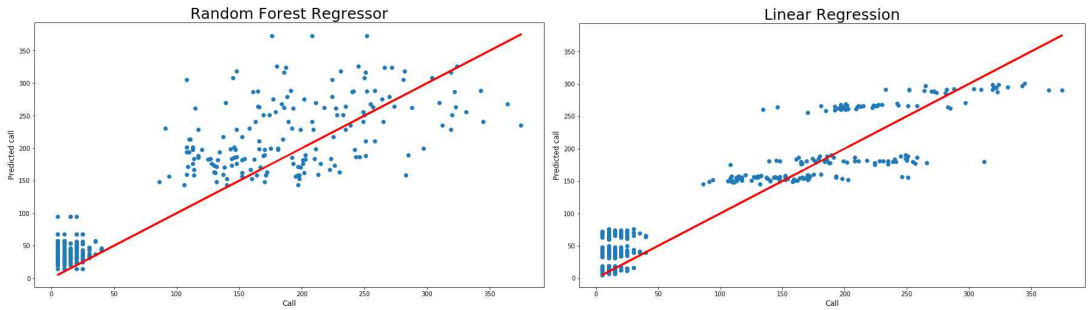
그룹은 다음과 같다.

연희동 / 대현동 홍제동 / 창천동 북가좌동 남가좌동 홍은동 / 나머지동



5. target값은 날짜, 읍면동그룹을 기준으로 하여 평균을 낸 치킨통화량으로 하였고 X값은 최종적으로 미세먼지, 평균기운, 일강수량, 읍면동그룹, 주말여부로 하였다.

〈중간 전 1차적 모델 구현〉



치킨 통화량 데이터는 양적 데이터이기 때문에 classification이 아니라 regression을 사용하였다.

우리가 구현한 모델은 random forest를 이용한 모델과 linear regression을 이용한 모델 두 가지였고, 두 가지 중 Linear regerssion model을 중간 발표 전 최종 모델로 선정하게 되었다. Linear Regression Model의 최종 R-squared는 0.65였다.

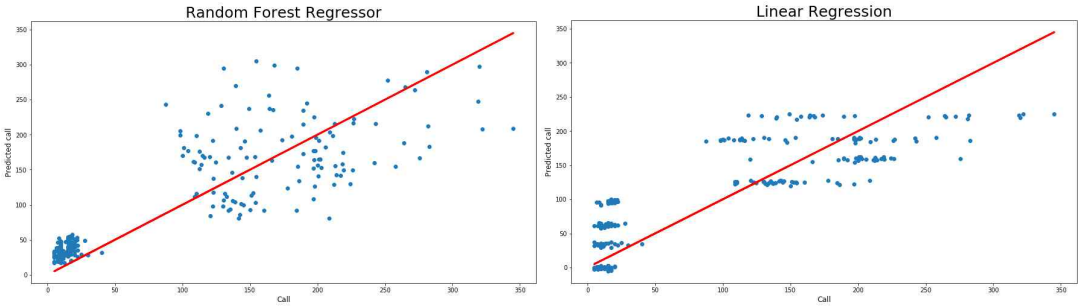
BIZ3199-01 AI비즈니스
김진우 교수님

본 알고리즘 사용 이유

: 치킨통화량(response)을 평균기온, 일 강수량, 미세먼지, 주말 여부 이용해 선형 관계로 설명하고 예측할 것이다.

- 설명변수와 반응변수 사이에 **선형관계**가 있다고 가정할 수 있다.
- 간단한 데이터이기 때문에 복잡한 알고리즘이 불필요하다.
- linear regression은 하이퍼 파라미터 튜닝의 영향이 상대적으로 적고 계산속도가 빠르다.
- 추후 input data에 대한 보충을 한 뒤에 black box model과 다르게 해석이 가능하다.(계수: 평균기온은 음수, 일강수량과 미세먼지는 양수)

〈중간 전 최종 모델 구현 결과〉



Model1 Test Mean squared error: 2370.91 (random forest)
Model2 Test Mean squared error: 2528.55 (linear regression)
Random Forest가 Linear Regression보다 test MSE가 약 200 정도 더 낮게 나왔으므로, Random Forest가 Linear Regression보다 예측률은 더 좋은 것으로 보인다.

그럼에도 불구하고 Linear Regression?

1. Ensemble model이나 Neural Net의 성능이 좋지만, 지금까지 수집한 데이터의 row와 column이 많지 않기 때문에 tree-based 나 Neural Net같은 복잡한 알고리즘이 불필요하다.

2. test MSE에서 엄청나게 큰 차이는 발생하지 않았기 때문에 복잡한 알고리즘을 사용하는 것보다 속도가 빠르고 더 간단한 linear regression을 사용하는 것이 더 효율적이라고 판단하였다.

따라서 속도가 빠른 선형 회귀 모델을 사용할 것이다.

하지만 계속해서 데이터를 다듬고 모델을 구현하다보면 random forest가 linear regression보다 훨씬 더 효율적인 예측을 할 수 있다는 가능성을 열어두고 있기 때문에 향후 random forest로 모델을 전환할 가능성이 높다.

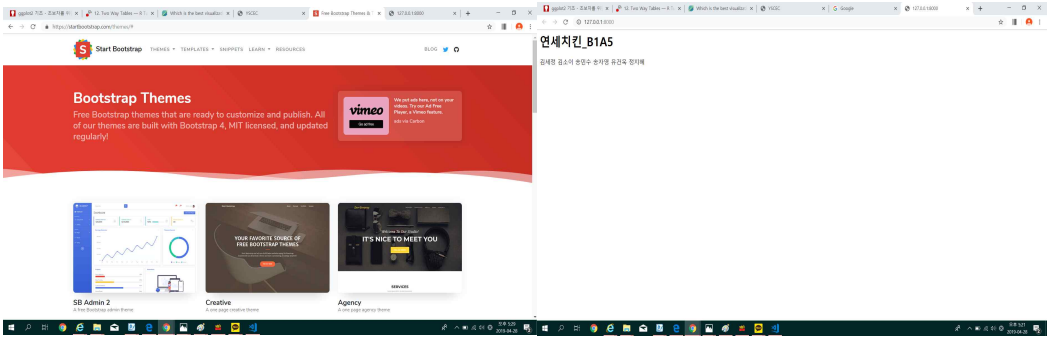
최종 서비스 구현 방안

: 서울의 날씨(미세먼지, 강수 등)와 서대문구 동을 독립변수로, 치킨 통화량 데이터를 반응 변수로 놓고 실행한 선형 회귀 모델 알고리즘을 통해 날씨, 지역과 치킨 판매량의 관계성을 알아본 뒤, 분석 결과를 웹으로 구현하는 방식으로 웹 방문 시 데이터 분석 결과를 확인할 수 있도록 한다.

1. 지역, 날씨예보, 요일을 작성하면 예상 통화량 제공

2. 지역, 날씨예보, 요일를 작성하면 성별, 나이대별 주문 비중 평균 제공(또는, 모델링을 통한 multi-classification)

BIZ3199-01 AI비즈니스
김진우 교수님



구현하려는 웹에 템플릿을 적용시켜 가독성을 높이고,(웹 디자인을 배워보는 노력을 할 것이고 만약 학기가 끝나기 전까지 웹 디자인에서의 완성이 되어있지 않다면 bootstrap 사이트를 이용하도록 한다.) aws(아마존 웹서비스)를 통하여 배포하는 것을 최종 구현 목적으로 한다.