대학생의 점심식사와 위치데이터와의 상관관계:

사례 연구*[1]

Team 8

2016100440 홍서영

2015105665 류연주

2017103723 박예린

2013100204 박완재

2013100273 윤수민

Contents

[1] 이 프로젝트에는 경희대학교 산업경영공학과 2019년도 데이터마이닝 수강생 40명의 2019년 3월 14일~4월 13일, 4월 14일~5월 13일의 평일 위치데이터를 활용하였다.

1. Research Background and Motivation(과제 수행 배경 및 동기)

본 프로젝트에서는 2019학년도 1학기 경희대학교 국제캠퍼스 산업경영공학과 데이터마이닝 과목을 수강하는 학생 40여명을 대상으로 'MapMyWalk' 스마트폰 어플리케이션을 이용하여 획득한 위치 데이터[2]를 활용하여 기업의 Business problem을 Data Science 관점에서 해결할 수 있는 사례를 찾던 중 '배달의 민족' 마케팅 팀과 접촉하여 점심시간의 '배달의 민족' 앱 이용률 증대를 위한 방안을 강구하기 위해 사례 연구를 진행하였다.

2. Nature and scope of the problem(문제 정의 및 도출)

'배달의 민족'에서 제공한 보고서의 선행 연구에 따르면 '배달의 민족'의 고객의 연령별 데이터에서 20대 비중이 남자 25.49%, 여자 16.23%로 많은 비중을 차지하고 있으며, 시간대별 데이터에서는 점심시간 대보다 저녁 시간대인 18시 가량에서 배달 주문이 많이 이루어지고 있다. 또한, 선행연구에서 실시한 설문조사에서 대학생 10명중 4명(47%)이 교외 시설에서 점심식사를 해결하는 이유로 다양한 메뉴를 고를 수 있다고 응답했다. 따라서 점심시간의 20대 고객을 확보하기 위한 마케팅 전략에 활용할 수 있도록 대학생의 점심시간과 위치 데이터 간의 유의미한 상관관계를 찾고자 한다.

〈주요 개념〉 대학생, 점심시간, 위치, 배달의 민족

3. Previous works and their problems(기존 연구 및 기술의 문제점)

대학생의 식사와 관련된 행동에 대해서는 주로 영양, 보건 측면에서 선행연구가 이루어졌다. 따라서 생활 패턴과 같은 요인에 따른 개개인의 식사 패턴보다는 대학생의 식생활 전반적인 측면을 다루는 경우가 많았다. 따라서 선행 연구에서는 점심식사 자체에서 이끌어낼 수 있는 대학생들에 대한 정보들이 부족하다. 또한 대학생의 교내 위치 데이터를 사용한 마케팅 전략 연구자료 역시 미비하여 참고할 수 있는 사안이 적다. 따라서 20대 학생의 교내 주간 시간대별 움직임과 위치를 나타낼 수 있는 자료들이 부족하기 때문에 이를 토대로 발견할 수 있는 유의미한 상관관계를 이끌어내기에 한계가 있다.

[2] 위도, 경도, 고도의 수치를 나타내는 데이터

4. Proposed Method(제안하는 방법의 개요)

어플리케이션을 이용해 수집한 해당 데이터 소유자가 해당 날에 밖에서 밥을 먹었는지, 안에서 먹었는지 확인한다. 그날의 생활 패턴을 파악한 후 선정한 attribute(①정문과의 거리, ②가용시간의 길이, ③기숙사 거주 여부, ④날씨, ⑤미세먼지)를 통해 Target attribute인 점심 식사 위치(outside/inside) 상관성을 분석한다. 교내에서 식사한 사람의 경우, 배달을 시켜먹을 가능성이 있다고 판단한다. 해당 target을 발견하기 위해 Classification Analysis에 기반한 model을 만들어 Target을 예측한다.

5. Principal results(논문 주요 결과의 개요)

가용시간, 정문과의 거리, 날씨 여부, 미세먼지 정도, 기숙사 거주 여부 총 5개의 Attribute를 가지고 R 프로그래밍의 R-part 패키지를 이용해 의사결정 나무 모델을 생성했다. Test set을 넣어 분석한 결과 도출해낸 모델은 약 75.17%의 정확도와 3.246e-08의 p-value값을 보여 유의미한 모델이라 판단하였다. 또한 해석을 하는 과정에 있어서 모순점이 발견되지 않아 다음과 같은 마케팅 전략을 수립하는데 유용한 모델이라 판단하였다. 첫째, 정문으로부터의 거리가 약 270m 이상인 앱 사용자들을 타겟으로 설정, 학식이 아닌 배달 음식 주문을 유도하는 push message를 보낼 수 있다. 둘째, 식사 위치에 대한 대략적인 가용시간 임계점인 2시간 10분을 참고, 기존 배달 시간보다 배달 시간을 줄인 대학생 대상 'on-time delivery' 도입을 고려해볼 수 있다.

Methods

경희대학교 산업경영공학과 데이터마이닝 과목에서 다루는 **CRISP-DM** 방법을 통한 문제 해결에 접근했다.

1. Business Understanding

1-1 Business Problem : '배달의 민족' 주문량이 저녁에 비해 점심에 적다. 고객 중 20대의 비율이 가장 높으므로 20대, 그 중에서도 대학생의 점심 주문을 늘리고자 한다.

1-2 Data science Problem : 대학생의 교내 위치 패턴의 데이터를 분석하여 영향을 끼치는 요인을 파악하고 결과치를 예측하고자 하기 때문에 Predictive Modeling 방법을 사용하고, Target Attribute가 존재하고 이산형의 형태를 지니고 있기 때문에 Classification Analysis 분석 방법을 사용한다.

2. Data Understanding

2-1 Raw material : 수집 기간 중 'Mapmywalk' 스마트폰 어플리케이션을 통해 획득한 데이터

2-2 Strengths & Limitations of data

2-2-1 Strengths : 시간, 위도, 경도, 고도 데이터가 하나의 데이터프레임에 저장된다. 대학생의 하루 전체에 대한 위치 데이터는 타 경쟁사에서 갖고 있지 않을 데이터이다.

2-2-2 Limitations : 데이터 조작, 측정 오차 발생 및 개인 프라이버시로 인한 소유자 불분명, 40명의 제한된 인원 수로 경희대학교 학생 전체를 대표하기에는 표본 수가 적다. 어플리케이션 조작 미숙으로 Missing data가 많을 수 있다.

2-3 Costs & benefits of data

2-3-1 Cost : 이용요금 월 ₩7,000 비용이 발생한다.

2-3-2 Benefits : 다른 프로젝트에도 활용할 수 있으며, 향후 마케팅에 활용했을 때 이를 통한 수익 증대를 기대해볼 수 있다.

3. Data pre-processing

▣ Data pre-processing은 **R**을 이용해 진행했다.

▣ 전제

1. 본 데이터마이닝 강의를 수강하는 모든 학생은 매일 한 끼의 20분[3] 이상의 점심식사를 한다.
2. 11:00 ~ 14:00 사이의 가용 시간은 점심 시간으로 활용한다.
3. 기숙사생을 제외한 학생이 11:00 ~ 14:00 사이에 학생회관과 제2기숙사에 머물렀다면, 이는 식사를 위해 방문한 것으로 간주한다.

3-1 GMT -> KMT 변환

수집된 시간 데이터가 GMT 표준시로 기록되어 있다. 한국의 대학생들을 대상으로 프로젝트가 진행되기 때문에 한국 표준시로 변환한다.

3-2 학교 외의 데이터 삭제

교내에 존재하는 학생 데이터를 사용하므로 학교 외의 위치해있는 위치데이터를 제외한다.

3-3 교내 위치해있는 장소와 시간을 추출

기록된 위도와 경도 값들을 토대로 학생이 머물렀던 장소(전자정보대학, 공과대학 등)의

---

[3] 이용숙, 이수현, "대학생의 점심식사 방식과 점심식사에 부여하는 의미: 문화기술적 연구", 서울대학교 비교문화연구소, "비교문화연구 제22집 2호, (2016): 340p 인용

위치를 파악하고 건물(Place)에 진입한 시점(Starting Point)와 나간 시점 (ending point)의 차이를 통해 머물러 있던 시간을 추출한다.

3-4 GPS 오류 전처리

어플리케이션 특성상 발생하는 이상치(같은 장소에 머무르고 있어도 이동이 기록되는 현상)을 보완하기 위해 주요 머무른 장소의 좌표 값을 평균을 토대로 구간 병합 처리하여 보완한다.

3-5 수집 기간 외 데이터 및 주말 데이터 삭제

평일 위치 데이터가 아닌 수집기간 외의 데이터, 주말 데이터를 삭제했다.

3-6 평균 가용 시간 구하기

- 점심식사를 할 수 있는 시간, 즉 수업 시간이 아닌 가용 시간을 파악한다.
- 데이터 수집 기간 중 같은 요일의 데이터를 분석, 해당 요일의 평균 가용 시간을 파악한다. 평균 가용 시간을 파악함으로써 해당 인물의 점심 시간대를 유추할 수 있다.

3-7 점심시간 설정

- 점심시간은 11:00 ~ 14:00로 설정한다. 학생회관의 학생 식당 점심 식사 제공이 13:30분에 마무리되는 점을 참고한다.

3-8 용어 정의와 Missing data 처리

①정문과의 거리, ②가용 시간의 길이, ③기숙사 거주 여부, ④날씨 ⑤미세먼지 ⑥식사 위치(outside/inside)

3-8-1 정문과의 거리

거점	X(경도)	Y(위도)
정문	127.078443	37.247441
멀티미디어관	127.07642	37.24459
공과대학	127.08067	37.24637
전자정보대학	127.08337	37.23930
예술디자인대학	127.08451	37.24172
국제대학	127.97126	37.24209
생명과학대학	126.08128	37.24286
체육대학	127.08035	37.24436
외국어대학	127.07764	37.24511

11~14시 시간 중 위치했던 거점을 파악하고 이들의 평균 좌표와 정문 좌표와의 거리를 미터(m)단위로 파악하여 점심식사를 외부에서 하기 위해 왕복해야하는 거리를 파악하여 수치로 정리한다. 이때, 정문과의 거리가 결측 처리되는 경우에 해당 데이터가 모델링에 사용되지 못하고 누락되어 활용 가능한 데이터의 수가 감소한다. 이를 보완하기 위하여 key값 별로 결측치를 제외한 정문과의 거리값의 평균을 구한 후, 결측치를 이 평균값으로 대체하였다.

3-8-2 가용 시간의 길이

데이터 수집 기간 중 해당 요일 별 주중 가용 시간의 평균을 구한다. 특정 요일에 대한 가용 시간이 없는 경우 해당 key의 전체 가용 시간의 평균으로 처리한다.

1. 수집한 dataset을 요일 별로 나누고, 해당 요일의 평균 가용 길이를 구하기 위해 해당 요일이 몇 번 존재하는지 확인하고 수업시간의 길이를 모두 더한다.
2. 해당 요일의 수업시간 길이의 합을 해당 요일의 카운트로 나누고 11시~14시 사이의 3시간인 10,800초에서 해당 길이를 빼서 가용 시간을 구한다.

3-8-3 기숙사 거주 여부

요일의 마지막 위치 레코드가 3번 이상 기숙사 좌표에 위치할 때 그 해당 데이터를 가진 사람을 기숙사에 거주하는 것으로 간주한다. 2차 프로젝트에서 전처리한 기숙사 데이터에서 남자 제 2기숙사의 좌표가 누락되었다는 문제가 발견되어 3차 프로젝트에서 문제점을 보완하였다. 이때 여자 제 2기숙사의 경우 1층에 학생 식당이 있기 때문에 교내 점심 식사 여부를 구분하기 위하여, 여자와 남자 기숙사를 각각 구분해 데이터 처리를 실시하였다.

3-8-4 날씨

날씨 웹 크롤링을 통해 해당 일의 날씨 데이터를 파악한 후 네이버의 날씨 구분을 그대로 적용한 후, 카테고리화 해서 점수를 부여한다. 맑음=5, 구름 조금 =4, 구름 많음=3, 흐림=2, 비=1, 눈=0점 순서로 부여하여 데이터 처리한다.

이전 연구에서는 텍스트를 기반으로 웹 크롤링을 실시하였으나, 처리 방식이 복잡하고 모호한 경우가 다수 발생한다는 한계점이 발견되어 이번 연구에서는 이미지 태그 기반으로 크롤링 방법을 변경하였다.

3-8-5 미세먼지

경기도 대기 환경 정보에서 11~14시의 수원시 영통구의 미세먼지 데이터를 파악, 평균값을 내고 카테고리화 하여 점수 부여하여 데이터 처리 점수 체계는 네이버 미세먼지에 근거한다.

0점부터 31점 미만은 3점(좋음), 31점부터 61점 미만은 2점(보통), 61점부터 151점 미만은 1점(나쁨), 151점 이상은 0점(매우 나쁨)을 부여한다.

3-8-6 식사 위치 판정

11시 ~ 14시 사이 교내에서 식사를 하는 학생을 이하 IS, 교외에서 식사를 하는 학생을 이하 OS로 정의한다.

해당 데이터가 11시 - 14시 사이에 OS/IS가 동시에 있다면 아래 기준에 따라 분류한다.

3-8-6.1 IS / OS 판단 근거

아래와 같은 기준을 통해 OS(0), IS(1)를 판단한다.

1. 나간 적 없는 경우 = 1

2. 나간 적 있는 경우

    1) 제2기숙사생 = 0

    2) 제2기숙사생이 아닌 경우

        (1) 학관/제2긱에 20분 이상 머무른 경우 = 1

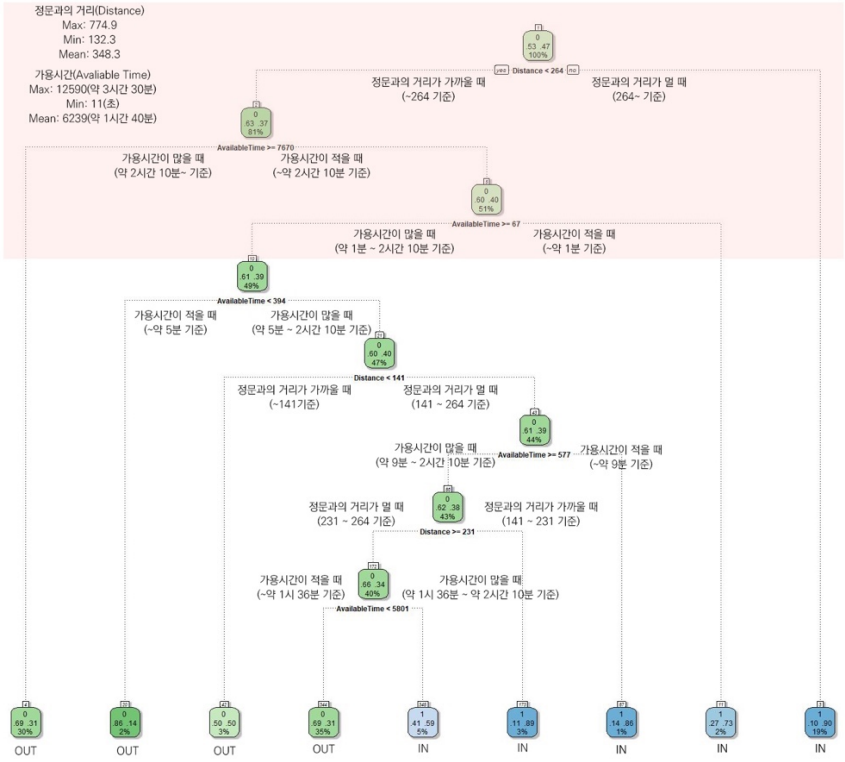
        (2) 학관/제2긱에 20분 이상 머무르지 않은 경우

            가. 11시~ 14시 사이에 나갔다 들어온 경우 = 0

            나. 11 시 ~ 14 시 사이에 나갔으나 14 시 내로 들어오지 않은 경우
                ㄱ. 14 시 이후에 학교 되돌아온 경우 = 0
                ㄴ. 14 시 이후에 학교 되돌아오지 않은 경우 = 1
            다. 11 시 이전부터 나간 경우
                ㄱ. 12 시 30 분 기준으로 이전에 들어온 경우 = 1
                ㄴ. 12 시 30 분 기준으로 이전에 돌아오지 않은 경우 = 0

4. Modeling

R에서 제공하는 rpart 패키지를 활용해 머신러닝으로 의사결정 나무(Decision tree)를 생성하여 Classification analysis 방법론을 적용하였다. 전체 966개의 데이터를 7:3으로 나누어 680개의 데이터와 286개의 데이터를 각각 training set과 test set으로 구성해 모델을 학습시켰다. 모든 set에서의 OS(0)과 IS(1)의 비율이 각각 0.55, 0.45로 동일하게 구성된다. 아래는 OS/IS 구분이 가장 잘 되는 attribute 값마다 유의미하다고 생성된 조건을 기준으로 우선순위를 정한 후, root node, internal node, leaf node를 생성해 도출된 모델들이다.



Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 152  52
         1  19  63

               Accuracy : 0.7517
                 95% CI : (0.6975, 0.8007)
    No Information Rate : 0.5979
    P-Value [Acc > NIR] : 3.246e-08

                  Kappa : 0.4582

 Mcnemar's Test P-Value : 0.000146

            Sensitivity : 0.8889
            Specificity : 0.5478
         Pos Pred Value : 0.7451
         Neg Pred Value : 0.7683
             Prevalence : 0.5979
         Detection Rate : 0.5315
   Detection Prevalence : 0.7133
      Balanced Accuracy : 0.7184

       'Positive' Class : 0
```

Distance 가 root node, 가용시간이 internal node 로 구성되어 두 attribute 가 모든 모델을 구성하였다. 그러나 외부 요인으로 작용할 것이라 예상했던 날씨와 미세먼지 정도, 기숙사 거주 여부가 배제되는 결과가 나왔다. R part 패키지를 활용하여 confusion matrix 를 만들고 test set 을 활용해 위 model 을 confusion matrix 에 따라 분석한 결과, 약 75.17%의 accuracy 와 3.246e-08 의 p-value 값을 갖는, 통계적으로 유의미한 모델임을 알 수 있다

Discussion

최종적으로 도출해낸 모델이 p-value 3.246e-08, 정확도 75.17%로 통계적으로 유의미한 모델로 보인다. 정문과의 거리도 최댓값 774.9(전자정보대학), 최솟값 132.3(공과대학)으로 수강생들의 특성을 반영한 결과를 보이고 있다. 구체적으로 해석하면, 정문과의 거리가 가까울 때, 가용시간이 많으면 밖에서 점심 식사를 해결하는 경우가 전체의 약 49% 임을 알 수 있다. 특히 공대생과 자대생이 대부분인 데이터 수집 대상 특성이 반영되어서, 모델의 root node 또한 우선적으로 공대와 자대로 분류되어 IS/OS 로 분류되었다. 이후. 정문에서 거리가 가까운 학생들(Distance 132.3~264)은 가용시간에 따라 IS/OS 가 분류되었다.

처음에 독립변수로 설정한 5 개의 Attribute(가용시간, 정문과의 거리, 기숙사 거주 여부, 날씨, 미세먼지 정도)에서 미세먼지, 날씨, 기숙사 거주 여부가 고려되지 않았다는 아쉬움이 있으나 Confusion Matrix 와 Statistics 결과에 의해 나머지 2 개 '가용시간'과 '정문과의 거리'가 Target Attribute 로 설정한 IOS 와 관련이 있다는 가설을 지지할 수 있다고 판단하였다. 기숙사 거주 여부의 경우, 이전 연구에서 누락된 남자 기숙사를 추가하였음에도 불구하고 전체 수강생 중 기숙사에 거주하는 학생이 5 명으로, 표본 크기를 고려했을 때 작은 숫자여서 모델링에서 고려되지 않았을 것이라 예상된다. 또한 외부요인으로 가정한 미세먼지와 날씨의 정도도 학교 내/외 식사 여부와 크게 상관이 없음을 알 수 있었다.

본 프로젝트의 결과를 바탕으로 다음과 같은 마케팅 활용방안을 제안한다. 11 시부터 2 시 사이에, 정문으로부터의 거리가 약 270m 이상인 앱 사용자들을 타겟으로 설정, 학식이 아닌 배달 음식 주문을 유도하는 push message 를 보낼 수 있다. 이 활용 방안은 캠퍼스가 넓어 교외로 나가서 식사하기가 번거로운 다른 대학에도 적용할 수 있을 것이다. 또한, 식사 위치에 대한 대략적인 가용시간 임계점인 2 시간 10 분을 참고하여, 기존 배달 시간보다 배달 시간을 줄인 대학생 대상 'on-time delivery' 도입을 고려해볼 수 있다. 가용시간이 적어도 2 시간은 있어야 교외에서 식사한다는 점을 고려, 2 시간 전에 예약 주문을 받고 수업이 끝나는 시간에 배달해 식사시간을 절약해주는 서비스를 생각해볼 수 있다. 이때 각 캠퍼스 근처의 식당과 연계하여 진행하면 상부상조로 더욱 효과적일 것이다.

1. 1 차 연구의 한계점

- 제 2 기숙사 남자동이 거점 목록에서 누락되어 남자 기숙사에 방문한 경우, 기록되지 않았다. 이에 따라 제 2 기숙사에 거주하는 남학생은 기숙사생으로 들어가지 않았다.
- p-value 가 0.3 이상으로 통상 기준치가 되는 0.05 보다 크게 높았다.
- 결측치를 따로 처리하지 않아, 모델링 과정에서 결측치를 가진 데이터가 누락되었다.
- 전체 데이터가 540 개로 적어, 랜덤으로 배정된 training set 에 어떤 데이터가 속했는지 따라 노드 기준이 완전히 바뀐다.

2. 2 차 연구에 보완된 점

- 약 한 달간의 데이터가 추가되어 총 966 개의 데이터로 분석할 수 있었다.
- 수집기간 외 데이터를 삭제하여 데이터의 통일성을 높였다.
- 가용시간과 정문과의 거리에 대한 결측치를 해당 key 의 평균값으로, 해당 key 에 해당 정보가 하나라도 존재하지 않는 경우는 NA 로 처리했다.
- 정문과의 거리 값에서 코드 오류로 생각되는 이상치를 여럿 발견해 제거, 정확도를 높였다.
- 남자 기숙사 좌표가 거점 목록에 추가되어, 남자 기숙사에 방문한 것을 기록할 수 있었다.
- 날씨 크롤링 방법을 기존 텍스트 크롤링에서 이미지 번호 크롤링으로 변경해 정확도를 높였다.
- OS 로 분류되어야 할 카테고리가 분류되지 않은 것을 발견, 해당 코드를 수정했다.
- p-value 가 0.32 에서 3.246e-08 로 매우 낮아졌다.

3. 여전히 남아있는 한계점

- 데이터마이닝 수업에 참가하는 모든 학생은 점심 식사를 거르지 않는다는 전제 하에 연구가 진행되었다. 학교생활 중의 점심식사 실태를 분석한 이희분, 유영상(1995)연구에 따르면 대학생들의 점심식사가 불규칙하게 이루어진다는 결과가 존재한다. 수업시간으로 인한 이유와 '배가 고플 때 점심 식사를 하기 때문'의 큰 이유가 있다. 따라서 점심식사를 실제로 하지 않는 학생들이 존재할 것으로 생각한다.
- 고도데이터의 전처리 불가로 인해 교내에서 식사한 학생(IS), 교외에서 점심식사 한 학생(OS) 구분에서 학생회관에서 20 분 이상 머무른 학생의 경우 모두 IS 처리를 했는데, 식사를 하지 않고 학생회관에 있는 동아리 방을 방문한 경우도 IS 처리를 하는 등의 한계가 존재한다.
- 제 2 남자 기숙사가 거점 목록에 추가됨에도 불구하고, 기숙사생의 표본이 여전히 예상보다 적어 일반화가 어려울 것으로 예상한다.

- R-part 패키지는 엔트로피, 지니계수를 기준으로 가지치기를 할 변수를 결정하기 때문에 상대적으로 연산 속도는 빠르지만 과적합화의 위험성이 존재한다는 한계점이 있다.
- 우리가 세운 기준에 따라 식사를 교외에서 했는지, 교내에서 했는지에 대한 결과를 얻었다. 하지만 이 결과는 추측으로, 해당 날에 실제로 어디서 식사를 했는지 수집이 불가능하므로 해당 attribute value 에 대한 정확도를 확인할 수 없다.
- 실제 머신 러닝에서는 최소한 10000 개의 sample data 를 사용해 학습시키는데, 그에 비해 사용가능한 데이터가 약 1000 개밖에 없어서 실제를 대변한다고 보기 어렵다.

Reference

논문 및 학술지

이용숙, 이수현, "대학생의 점심식사 방식과 점심식사에 부여하는 의미: 문화기술적 연구", 서울대학교 비교문화연구소, "비교문화연구 제22집 2호, (2016): pp. 329~390

대학내일, "대학생들은 학교에서 점심을 어떻게 먹을까?: 대한민국 대학생 점심 백서", 대학내일20대연구소, "연구리포트 2014-5"(2014) :12:

주성일, "위치정보기반 대학 캠퍼스 공간관리 시스템 구축에 관한 연구", 한양대학교 건축환경공학과, 2011:

김대룡, 김다영, 변수지, "날씨에 따른 배달음식 주문건수 예측", 한국기상학회,"한국기상학회 학술대회 논문집, 2016.10, pp 480~481.

왕망, 권혁준, 최재원, "위치기반 어플리케이션 서비스에서 제품의 희소성과 구매행동과의 영향", e-article 학술교육원, Information Systems Review, 30 June 2018, Vol.20(2), pp.209-226

기사

트렌드모니터, "불철주야 '배달음식'을 찾는 사람들, 전체 55.9%가 '배달앱(APP)' 사용경험 있어",

https://trendmonitor.co.kr/tmweb/trend/allTrend/detail.do?bIdx=1296&code=0301&trendType=CKOREA

Cheil magazine, "통계로 보는 배달 음식", http://blog.cheil.com/magazine/33722