The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.977
28(5), 977–990

A Study on the Number of Domestic Food Delivery Services

Jaeyoung Kwon[a] · Sinae Kim[a] · Eungee Park[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received July 29, 2015; Revised August 23, 2015; Accepted September 23, 2015)

Abstract

Food delivery services are well developed in the Republic of Korea, The increase of one person households and the success of app applications influence delivery services these days. We consider a prediction model for the food delivery service based on weather and dates to predict the number of food delivery services in 2014 using various data mining techniques. We use linear regression, random forest, gradient boosting, support vector machines, neural networks, and logistic regression to find the best prediction model. There are four categories of food delivery services and we consider two methods. For the first method, we estimate the total number of delivery services and the posterior probabilities of each delivery service. For the second method, we use different models for each category and combine them to estimate the total number of delivery services. The neural network and linear regression model perform best in the first method, this is followed by the neural network which is the best for the second method. The result shows that we can estimate the number of deliveries accurately based on dates and weather information.

Keywords: delivery services, linear regression, random forest, gradient boosting, support vector machines, neural network, logistic regression

1. 서론

우리나라는 세계적으로 배달문화가 발달한 나라로 다양한 종류의 음식에 대해서 배달이 가능하고 배달 가능 시간 또한 상당히 긴 편이다. 또한 최근에는 혼자 사는 일인가구가 늘어나고 맞벌이 부모들이 늘면서 간편하고 편리한 배달 음식에 대한 수요가 증가하고 있고, 따라서 배달 시장이 더욱 상승세로 접어들고 있다. 이와 더불어, 배달음식의 종류 또한 매우 다양해지고 있어서 가장 대표적으로는 치킨부터 심지어는 회를 배달해 주는 배달시스템까지 갖추고 있을 정도로 배달 가능한 음식의 범위가 점점 확대되고 있다. 이러한 상승세를 틈타 배달 앱 시장도 매우 활발하게 발전하고 있다. 스마트폰을 이용하여 다양한 업종별, 점포별 비교를 통해서 가장 합리적인 소비와 가장 신속한 배달을 할 수 있으며, 또한 음식점의 입장에서도 전화로 주문을 받는 시간과 비용의 불편함을 줄일 수 있고, 앱을 이용한 주문 수요가 증

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (No. NRF-2013R1 A1A2012817).
[1]Corresponding author: Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.
E-mail: josong@ewha.ac.kr

Table 1.1. Description of the regression models

선형 회귀 모형	설명변수와 반응변수간의 선형 관계를 가정하고 결과 해석이 용이. Stepwise regression, LASSO, Ridge regression 등과 같은 변수선택법 사용가능.
랜덤 포레스트	Bagging 방법론 중의 하나로 많은 수의 bootstrap sample을 이용하여 다수의 의사결정나무의 적합 결과를 이용하는 예측모형 (Brieman 등, 1984). 변수 선택 시 랜덤한 일부분의 변수만을 이용하여 de-correlated tree를 구축하여 추정치의 분산을 줄여주는 방법론.
그래디언트 부스팅	손실 함수의 경사도를 바탕으로 다수의 약한 예측 모형들을 단계적으로 생성하여 결합함으로써 강한 예측 모형을 생성 (Thomas, 2000). 일반적으로 아주 간단한 의사결정나무를 예측 모형으로 사용.
서포트 벡터 기계	데이터를 분리하는 초평면 중에서 마진이 큰 초평면을 선택하여 분리하는 방법으로서, 예측이 정확하고 여러 가지 형태의 자료에 대하여 적용이 용이 (Karatzoglou 등, 2006). 회귀모형에서는 가장 많은 자료 점들을 포함하는 참 회귀함수 둘레의 튜브를 이용하는 방법론.
신경망 모형	인간의 두뇌구조를 모방한 모형으로서, 독립변수와 반응변수간의 관계를 은닉층(hidden layer)이 있다고 가정함. 예측력이 좋은 반면 해석이 어렵다는 단점이 있음. 단층 은닉 신경망(single hidden layer)이 많이 사용됨.
로지스틱 회귀모형	반응변수가 범주형 자료의 경우에 사용되며 log odds ratio가 설명변수와 선형관계에 있다고 가정함. 선형회귀모형과 같이 결과해석이 용이하고 stepwise같은 변수 선택 방법론 적용 가능.

가하고 있기 때문에 경쟁력을 위해서 높은 수수료임에도 배달 앱에 가입하는 것이 불가피하다. 이들 앱에서는 이용가능 지역을 설정하면 그 지역에서 이용 가능한 음식점들을 보여주며, 이용시간 또한 제공해준다. 그리고 이 음식점을 이용해 본 사람들의 후기와 평점을 볼 수 있어서 이용에 참고할 수 있다. 마지막으로 배달시킬 음식을 선택한 후에는 결제를 앱을 통하여 할 수도 있고, 평소와 같이 배달음식을 받으면서 결제를 할 수도 있다. 하지만 아직도 많은 배달이 전화를 통해 이루어지고 있으며, 따라서 우리는 SKT에서 제공하는 배달 통화건수 자료를 이용하여 분석을 진행하였다.

우리는 이 논문을 통해서 배달음식의 이용건수를 시간과 날씨에 따라 예측해 봄으로써 판매자 측에서는 판매량을 예측하여 하루에 필요한 물량을 예측하고, 효율적인 시간 관리와 운영에 이익을 줄 것이다. 주문자 입장에서는 주문량이 많은 시간대를 피해서 주문을 하면 더 신속하게 배달음식을 이용할 수 있을 것이다. 또한 앱 개발자 입장에서는 평균 통화량과 평균 대기시간을 알려주어 이용자들이 더 편리한 선택을 할 수 있도록 도와주는 역할을 할 수 있다. 본 논문의 구성은 다음과 같다. 2장에서는 분석에 사용된 자료에 대한 설명으로 자료수집 과정과 총 배달건수에 대한 간단한 요약 및 설명을 한 후 변수 설명 및 데이터 전처리 과정에 대한 설명을 할 것이다. 3장에서는 통계프로그래밍 R (R Development Core Team, 2010)을 통하여 회귀 분석한 결과를 보여준다. 분석에 사용한 모형은 선형회귀모형, 랜덤 포레스트 (Breiman, 2001), 그래디언트 부스팅 (Friedman, 2002; Ridgeway, 2012), 서포트 벡터 기계 (Cortes와 Vapnik, 1995; Karatzoglou, 2006), 신경망 (Hastie 등, 2009; Park 등, 2011), 로지스틱 회귀모형 (James 등, 2013) 총 여섯 가지 모형으로, 많이 사용되는 회귀모형이므로 따로 설명하지 않고 Table 1.1에서 간략히 소개한다. 각각의 회귀모형은 다양한 tuning parameters가 있으나 이를 여기서 모두 소개하기에는 지면의 부족함이 있으므로 (Park 등, 2011)을 참조하기 바란다. 4장은 결론 부분으로 최종예측모형을 통하여 총 배달이용건수와 각 업종별 배달이용건수를 예측하여 실제 값과 비교한다. 마지막으로 이 논문이 가지는 시사점을 요약한다.

2. 분석 자료 설명

2.1. 자료수집과정

분석에 사용된 자료는 SKtelecom Bigdatahub 사이트(http://www.bigdatahub.co.kr)에서 제공하는 오픈소스 데이터인 '배달 업종 이용 현황분석 (2014년도)'를 기본 데이터로 하고, 기상청(http://www.kma.go.kr)에서 제공하는 일별, 시간대별 자료를 취합하여 만들었다. 배달 업종 이용 현황분석 데이터는 서울 지역 배달 업종에 대한 한 달간 요일/시간대별 이용현황 데이터로, 제공되는 변수는 기준일, 요일, 시간대, 업종, 통화량이다. 통화량은 T고객(발신) 기준 이용자의 배달 업종 통화건수로 통화량 5건 미만은 5건으로 표시된 데이터이다. 제공되는 데이터 중 2014년 1월 1일부터 12월 31일까지 총 1년 365일의 자료를 분석에 이용하였다. 기상청에서 제공되는 날씨 데이터 중에서 시간대별로 제공되는 시간별 기온, 풍향, 풍속, 습도, 미세먼지를 이용하였고, 일별로 제공되는 일별 평균기온, 최저기온, 최고기온, 운량, 강수량, 일조시간, 일출, 일몰과 같은 날씨타입 자료를 이용하였다. 운량에 따라서 0~2는 맑음, 3~5는 구름조금, 6~8은 구름 많음, 9~10이상은 흐림으로 날씨를 분류하였고, 이를 순서형 변수인 $1, 2, 3, 4$로 정의하였다. 기상청에서 제공되는 날씨타입 중에 눈, 비가 있는 경우를 1로 아닌 경우를 0으로 하여 눈, 비 여부를 표시하였다.

날짜에 따라서 파생되는 변수는 봄(3~5월), 여름(6~8월), 가을(9~11월), 겨울(12~2월) 계절 변수와, 공휴일, 기념일 변수이다. 2014년 공휴일은 네이버에 검색을 통해 확인한 결과 총 67일 이였고, 6월4일 지방선거를 포함하면 총 68일 이다. 공휴일 여부는 공휴일 전날을 2로, 공휴일을 1, 나머지를 0으로 하여 표시하였다. 기념일의 경우는 동계올림픽(2월 7일~2월 23일), 아시안 게임(9월 19일~10월 4일), FIFA월드컵(6월 13일~7월 14일), 블랙데이(4월 14일), 복날(7월 18일, 7월 28일, 8월 7일), 삼겹살데이(3월 3일), SKT 피자 멤버십 할인데이(1월~4월 마지막 주 수요일, 4월 14일~5월 31일 매주 금요일), 황금연휴(5월 1~6일, 6월 4~8일, 9월 6~10일) 등을 고려하여 1과 0의 범주형변수로 포함시켰다. 본 논문에서는 시간대별 자료를 이용하여 분석 할 것이다. 여기서 총 관측치의 개수는 33,281개이며, 2014년 1월 1일부터 2014년 12월 31일까지 서울의 배달음식(치킨, 피자, 족발/보쌈, 중국음식) 데이터를 사용하였다. 일 년 365일 동안 0~23시간대에 얻어지는 4가지 배달 업종에 대한 자료이므로 총 35,040개여야 하는데 데이터로 확인해본 결과 배달이용이 없는 시간대 이였기 때문에 결측치인 경우 1759개를 통화량 0건으로 처리하였다. 관측치를 주어진 시간대별로 묶어서 자료를 다시 정리하면 총 8,758개 행의 자료가 얻어지는데 이를 우리의 분석 자료로 이용하였다. 여기서는 1월 31일 6과 7시간대에 4업종 모두 결측값을 가지는 것으로 나타났기 때문에 두 건을 제외한 8,758개의 자료로 분석하였다.

2.2. 총 배달건수 요약

총 배달이용건수를 보면 시간대별 최대 이용건수 17,752건(10월 26일 오후 18시), 시간대별 최소 이용건수 5건 (9월 8일 오전 6시), 시간대별 평균 이용건수 2,226건으로 변동성이 매우 큰 변수이다. 하루 최대 이용건수 101,941건(10월 26일), 하루 최소 이용건수 9,383건(5월 8일), 하루 평균 이용건수 37,388.46건으로 일별 이용건수에도 변동성이 큰 것을 알 수 있다. 또한 총 배달이용건수의 월별 합을 구한 결과 월별 합이 가장 높은 달은 10월(총 1,897,206건)로 가장 작은 달인 2월(총 912,905건)의 2배에 달한다.

업종별로 하루 평균 이용건수를 비교해 보면 치킨(16712.1건), 중국음식(13947.5건), 피자(4740.8건), 족발/보쌈 정식(1988.1건) 순으로 많았다. 업종별로 월별 합의 평균을 구한 결과 역시 치킨(508,326.8건), 중국음식(424,235.9건), 피자(144,199.7건), 족발/보쌈 정식(60,469.92건) 순이다. 4가지 업종 모두 10월에 총 이용건수 월별 합이 가장 많았으며, 치킨(2월)을 제외한 세 업종이 11월에 가장 적었지만

Table 2.1. Description of variables

X (설명변수)	변수 설명	타입
month	월 (m2: 9–10월 , m1: 나머지)	범주형변수
day	요일 (mon, tue, wed, thur, fri, sat, sun)	범주형변수
time	시간대 $(0, 1, \ldots, 24)$	범주형변수
season	계절 (봄, 여름, 가을, 겨울)	범주형변수
holiday	공휴일 (2: 공휴일 전날, 1: 공휴일, 0: 나머지)	범주형변수
holiday2	기념일 (1: 아시안게임, 복날, 블랙데이 등, 0: 나머지)	범주형변수
temp	기온	연속변수
wind_dir	풍향	연속변수
wind_spe	풍속	연속변수
hum	습도	연속변수
dust	미세먼지	연속변수
temp_aver	평균기온	연속변수
temp_min	최저기온	연속변수
temp_max	최고기온	연속변수
cloud	운량	연속변수
climate	날씨 (1: 맑음, 2: 구름조금, 3: 구름많음, 4: 흐림)	순서변수
rain_snow	눈비 (1: 눈 또는 비, 0 : 나머지)	범주형변수
rain_amount	강수량	연속변수
suntime	일조시간	연속변수
sunrise	일출	연속변수
sunset	일몰	연속변수
Y (반응변수)	변수 설명	타입
A	족발/보쌈정식에 대한 배달 이용건수	연속변수
B	중국음식에 대한 배달 이용건수	연속변수
C	치킨에 대한 배달 이용건수	연속변수
D	피자에 대한 배달 이용건수	연속변수
SUM_call	4가지 업종에 대한 총 배달 이용건수	연속변수

치킨의 이용건수가 많으므로 전체 데이터에서는 2월에 월별 합이 가장 작았다.

시간대별 총 이용건수가 만4천 건 이상인 이상점들이 있었다. 이 값은 총 이용건수의 평균값인 2,226건에 비해 매우 큰 값이며 따라서 총 이용건수가 가장 높은 8개의 이상점에 대해서 알아보도록 한다. 이 이상점들의 특징으로는 9월, 10월로 계절이 가을이며, 요일은 토, 일요일로 주말이며 시간대는 저녁시간인 18시가 대부분이다. 그 밖에 다른 특징은 보이지 않았다. 따라서 총 이용건수의 극대값에 관해서는 시간 변수에 큰 영향을 받고 있다고 생각할 수 있다.

2.3. 변수설명 및 데이터 전처리 과정

본 논문에 사용된 변수들의 설명은 Table 2.1에 나와있다. 특히 반응변수는 A(족발/보쌈 이용건수), B(중국음식 이용건수), C(치킨 이용건수), D(피자 이용건수)이고 전체 이용건수는 SUM_call로 정의하였다. 결측치가 존재하는 미세먼지(dust) 변수의 경우, KNN(K-nearest neighbor) 방법을 이용하여 값을 채워주었다. 결측치를 처리하는 많은 방법론 중에 KNN방법론을 사용한 이유는, 우선 결측치의 숫자가 많지 않은 경우에 가장 빠르면서 효과적으로 결측치를 해결할 수 있는 방법론이기 때문이다. 또한 month, day, time 변수는 범주형 변수로서, 범주의 수가 많아 용이한 해석을 위하여 상자그림과 GMM(Gaussian Mixture Model)을 이용하여 군집분석을 수행하고 그 결과를 이용하여 그룹화를 시도

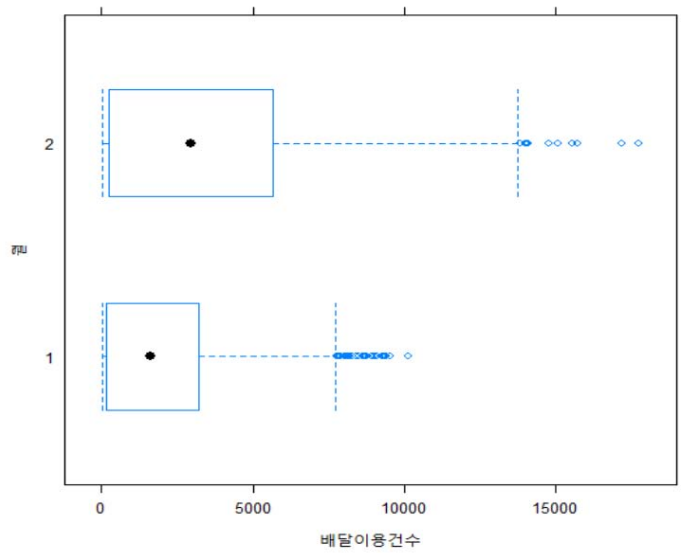


Figure 2.1. Boxplot of month variable after grouping.

해보았다. 그 결과, day, time 변수의 경우 그룹화 후의 예측력이 급격히 감소하여 그룹화 전의 원범주를 그대로 사용한다. 하지만 month 변수는 9월 10월을 m2로, 나머지 월을 m1로 한 범주형 변수로 그룹화한 결과, 원범주를 사용했을 때와 예측력이 크게 차이나지 않아 그룹화 후의 범주를 사용하였다. 그에 따른 상자그림은 Figure 2.1과 같다. 그림 상에서 각 변수 별로 차이가 있는 것을 확인할 수 있다.

분석을 위해 8,758개의 자료를 7 : 3 비율로 train과 test로 나누어 각각 train 6,130개, test 2,628개의 자료로 분석을 진행하였다. 설명변수 중에서 범주형 변수인 month, day, time, season, holiday, holiday2, rain_snow 총 7개에 대하여 factor 변환 하였다. 순서형 변수인 climate는 연속변수로 간주하여 분석에 사용하였다. 최종 모형에 사용된 독립변수 및 설명변수는 Table 2.1과 같다.

3. 분석 결과

이번 장에서는 두 가지 방법을 이용하여 각 배달업종별 이용건수를 예측해보도록 한다. 3.1절에서는 4가지 업종의 총 이용건수인 SUM_call을 예측하고, A, B, C, D 변수를 묶은 행렬을 반응변수로 하여 시간대별 각 배달업종의 이용비율을 예측하여 최종적으로 각 배달업종별 이용건수를 예측해본다(방법1). 3.2절에서는 각 업종별로 최적의 모형을 구축하여 각 배달업종별 이용건수를 예측하고 총 이용건수는 개별 업종별 이용건수의 합을 이용하여 예측해본다(방법2).

3.1. 전체모형(방법1)

위에서 설명한 바와 같이 이번 절에서는 우선 총 이용건수를 예측한 후에 업종 별 이용 건수를 예측하는 방법이다.

3.1.1. 총 이용건수 예측 5가지 회귀모형(선형회귀모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 신경망 모형)을 이용하여, 4가지 업종의 총 이용건수인 SUM_call을 예측해보았다.

Table 3.1. The result of linear regression (significant level = 0.05)

변수	회귀계수	p-value	변수	회귀계수	p-value
Intercept	10.697	<2.00E-16	time_15	0.843	<2.00E-16
month_m2	0.693	<2.00E-16	time_16	1.021	<2.00E-16
day_mon	−0.246	<2.00E-16	time_17	1.396	<2.00E-16
day_sat	0.260	<2.00E-16	time_18	1.822	<2.00E-16
day_sun	0.203	<2.00E-16	time_19	1.888	<2.00E-16
day_thur	−0.077	3.87E-12	time_20	1.668	<2.00E-16
day_tue	−0.154	<2.00E-16	time_21	1.473	<2.00E-16
day_wed	−0.089	4.90E-16	time_22	1.319	<2.00E-16
time_1	−0.845	<2.00E-16	time_23	0.810	<2.00E-16
time_2	−1.526	<2.00E-16	season_spring	0.138	2.20E-10
time_3	−1.968	<2.00E-16	season_summer	0.235	<2.00E-16
time_4	−2.299	<2.00E-16	season_winter	0.127	2.60E-12
time_5	−2.604	<2.00E-16	holiday_1	−0.109	6.50E-14
time_6	−2.758	<2.00E-16	holiday2_1	0.052	3.94E-10
time_7	−2.748	<2.00E-16	wind_spe	0.007	0.003077
time_8	−2.460	<2.00E-16	hum	0.001	1.95E-08
time_9	−1.481	<2.00E-16	temp_max	−0.006	9.52E-11
time_10	−0.296	<2.00E-16	cloud	−0.017	2.05E-07
time_11	0.878	<2.00E-16	climate	0.044	2.32E-05
time_12	1.290	<2.00E-16	rain_amount	0.002	0.000212
time_13	1.127	<2.00E-16	sunrise	−0.003	<2.00E-16
time_14	0.892	<2.00E-16	sunset	−0.003	<2.00E-16

먼저 단계적 선택법(stepwise selection)을 실시하여 최적의 선형회귀모형을 구해본 결과, adjusted-R^2가 0.982로 선형회귀모형이 데이터를 충분히 설명하고 있다고 판단된다. 유의 수준 0.05 하에서 유의한 변수들의 회귀 계수와 p-value는 Table 3.1과 같다.

최적 선형회귀모형의 중요변수에 대한 설명은 다음과 같다. 먼저 날짜 변수들을 살펴보면, 9월과 10월이 다른 달에 비해 배달이용건수가 많으며(month), 가을에 비해 봄, 여름, 겨울에 배달이용건수가 많다(season). 이 둘은 상충되는 결과이지만, month의 효과로 인하여 season의 회귀계수가 영향을 받은 것으로 보인다. 또한, 금요일에 비해 주말(토, 일)은 배달건수가 더 많고 평일(월, 화, 수, 목)에는 더 적으며(day), 새벽 12시를 기준으로 오전 시간대(1시–10시)에는 배달건수가 적고 오후 시간대(11시–23시)에는 더 많다(time). 휴일 관련 변수의 경우, 공휴일이 아닌 날에 비해 공휴일 전날에 배달건수가 증가하며(holiday), 기념일 당일에 배달건수가 증가한다(holiday2). 다음으로 날씨 변수들을 살펴보면, 습도가 높고(hum), 흐리며(climate), 강수량이 많을 때(rain_amount), 배달 건수가 증가한다.

Figure 3.1은 변수들의 상대적인 중요도를 알아보기 위하여 그려 본 랜덤 포레스트의 Variable Importance Plot이다. 이를 살펴보면, time, day, month, hum 변수 순으로 중요하며, 그 중 time변수가 가장 큰 영향을 미치는 변수임을 확인 할 수 있다.

다음으로, 위에서 언급한 5가지 회귀모형을 이용하여 모형을 적합 시키고 각 모형의 예측력을 비교해보고자 한다. 튜닝 모수가 필요한 모형에 대해서는, 10-fold cross validation을 이용하여 구한 최적의 튜닝 모수를 이용하였다. 최적의 튜닝 모수들을 살펴보면, 랜덤 포레스트의 경우 mtry = 17, 그래디언트 부스팅의 경우 shrinkage = 0.05, 서포트 벡터 기계의 경우 gamma = 2^{-6}, cost = 60, 신경망 모형의 경우 size = 8, decay = 0.1이다. 모형의 적합도 및 예측력 평가 척도는 평균 제곱근 오차(RMSE)를 이

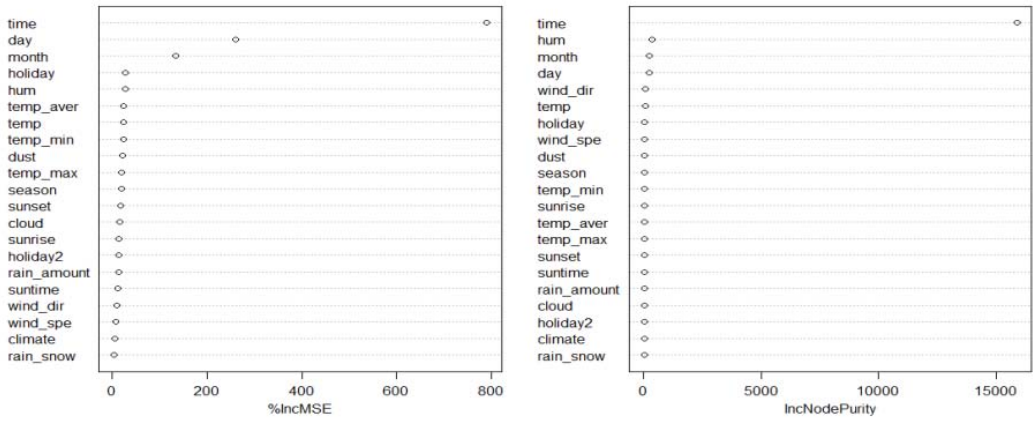


Figure 3.1. Relative importance of the independent variables using random forest.

Table 3.2. Train and test error of the models (Total number of delivery)

	Train error	Test error
선형 회귀 모형	644.2067	640.9477
랜덤 포레스트	437.6746	432.9345
그래디언트 부스팅	595.0497	611.5478
서포트 벡터 기계	360.2602	445.7141
신경망 모형	296.9019	297.0031

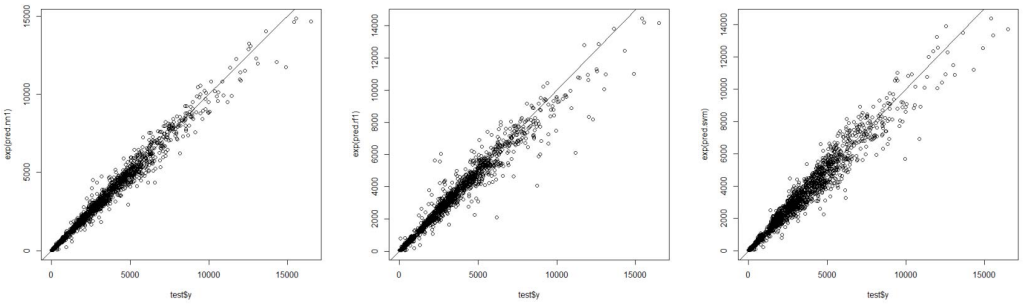
용하였다.

각 모형을 비교해 본 결과는 Table 3.2와 같다. 신경망 모형이 다른 모형에 비해 적합도와 예측력 모두에서 월등히 좋은 결과를 보이고 있으며, 랜덤 포레스트와 서포트 벡터 기계의 경우에도 나쁘지 않은 적합도와 예측력을 보임을 확인 할 수 있다. 이는 Figure 3.2에서도 확인 가능하다. 따라서 세 가지 모형(신경망 모형, 랜덤 포레스트, 서포트 벡터 기계)을 총 배달이용 건수 예측을 위한 최종 모형으로 선택하여 3.1.3장에서 각 배달음식별 이용건수를 예측해 보도록 한다.

3.1.2. 각 배달업종의 이용비율 예측 각 배달업종의 이용비율은 반응변수를 행렬 형태로 넣을 수 있는 다차원 선형회귀모형(Multivarite Linear Regression Model), 일반화 선형 모형(Generalized Linear Model)을 이용하여 예측했다.

1) 다차원 선형회귀모형(Multivariate linear regression model)

단계적 선택법을 실시하여 최적의 선형회귀모형을 구하였다. 선형회귀모형에서는 4종류 배달음식의 이용 비율(posterior probability)을 예측하기 위해서 반응변수를 이용건수의 weight 값으로 적합한다(weight 값 = 이용건수/네 업종 총 이용건수). 최종적으로 선택된 변수는 month, day, time, season, holiday, temp, wind_dir, wind_spe, hum, dust, temp_aver, temp_max, cloud, climate, suntime, sunrise, sunset이다. adjusted-R^2은 A(족발/보쌈 정식) 0.488, B(중국음식) 0.959, C(치킨) 0.973, D(피자) 0.729로 대체로 높았으나 중국음식, 치킨 업종 모형에서 적합도가 특히 높았다. 각 업종별 유의한 변수들의 회귀계수는 (http://home.ewha.ac.kr/∼josong/delivery/)에서 확인할 수 있다 (유의수준 0.05 기준).



(a) 신경망 모형 (b) 랜덤 포레스 (c) 서포트 벡터 기계

Figure 3.2. Predicted values VS True values (Total number of delivery).

Table 3.3. The influential variables for each delivery service

	증가		감소	
족발/보쌈 정식	기온별	최고기온	월별	9, 10월
	시간별	1시∼8시, 18시∼19시	요일별	토, 일요일
			시간별	10시∼16시 22시∼23시
중국음식	요일별	월∼목요일	요일별	토요일, 공휴일, 공휴일전날
	시간별	1시∼20시, 일출시, 일몰시	계절별	여름, 겨울
			기온별	최저기온
			시간별	21시∼23시
치킨	요일별	토, 일요일	요일별	월∼목요일
	기온별	최저기온	기온별	최고기온
	계절별	봄, 여름, 겨울	시간별	1시∼21시, 일출시, 일몰시
	월별	9, 10월		
	날씨별	미세먼지 양		
	시간별	23시		
피자	요일별	일요일, 공휴일	요일별	월∼목요일
	기온별	시간별 기온, 최저기온	기온별	최고기온
	날씨별	날씨 맑음	날씨별	운량
	시간별	2시∼23시, 일출시	계절별	여름
			월별	9, 10월

Table 3.3은 선형회귀모형의 회귀계수에 대하여 유의수준 0.05를 기준으로 배달업종별 회귀 계수가 양수인 변수들을 증가 항목으로, 음수인 변수를 감소 항목으로 정리한 결과표이다. 증가 변수이거나 증가 변수의 값이 커질수록 각 업종별 배달이용비율이 증가하며, 감소 변수이거나 감소 변수 값이 커질수록 배달이용비율이 감소한다. 결과를 보면 배달 업종별로 증가, 감소하는 변수가 각각 다른 것을 알 수 있다. 또한 모든 업종에 대해서 대체로 시간과 관련한 변수가 유의하다. 범주형 변수의 baseline 값은 각각 오전 0시(time), 금요일(day), 1월(month), 가을(season)이다.

2) 일반화 선형 모형(Generalized linear model)

Stepwise를 실시하여 최적의 일반화 선형 모형을 구하였다. 다항 로지스틱 회귀모형(Multinomial Logistic regression model)을 이용했으며 반응변수는 weight값이 아닌 각 업종별 이용건수의 행렬을 이용

Table 3.4. Test error of the models

	족발/보쌈 정식	중국음식	치킨	피자
선형회귀모형	0.02689662	0.05835995	0.04476176	0.02646191
일반화선형모형	0.05839920	0.11337370	0.07222395	0.07370148

Table 3.5. Test error using linear regression (Number of each delivery service)

	족발/보쌈정식	중국음식	치킨	피자	전체데이터
신경망 모형	**37.71563**	**151.7389**	**210.7867**	**61.11236**	**134.7339**
랜덤 포레스트	43.68084	157.6834	288.7749	69.48777	169.5520
서포트 벡터 기계	44.10318	170.7989	293.6540	76.80266	175.5340

Table 3.6. Test error using logistic regression (Number of each delivery service)

	족발/보쌈정식	중국음식	치킨	피자	전체데이터
신경망 모형	52.45214	175.5646	217.7885	65.15830	145.9894
랜덤 포레스트	58.25157	179.2984	290.5880	76.34362	177.3491
서포트 벡터 기계	56.90918	192.4844	299.1761	79.66012	184.4864

했다. baseline 반응변수로는 A(족발/보쌈정식)가 사용되었다. 최종적으로 선택된 변수는 time, holiday, wind_dir, wind_spe, hum, dust, temp_aver, temp_max, cloud, sunrise, sunset이다.

예측한 이용 비율의 test error를 구한 Table 3.4를 살펴보면, 일반화선형모형 보다는 선형회귀모형에서 대체로 작은 값이 나온 것을 알 수 있다. 따라서 선형회귀모형이 이용비율을 예측하는데 더 적합한 모형이라고 할 수 있다. 각 업종별로 보면 대체로 족발/보쌈 정식, 피자, 치킨, 중국음식 순으로 test error가 작았다.

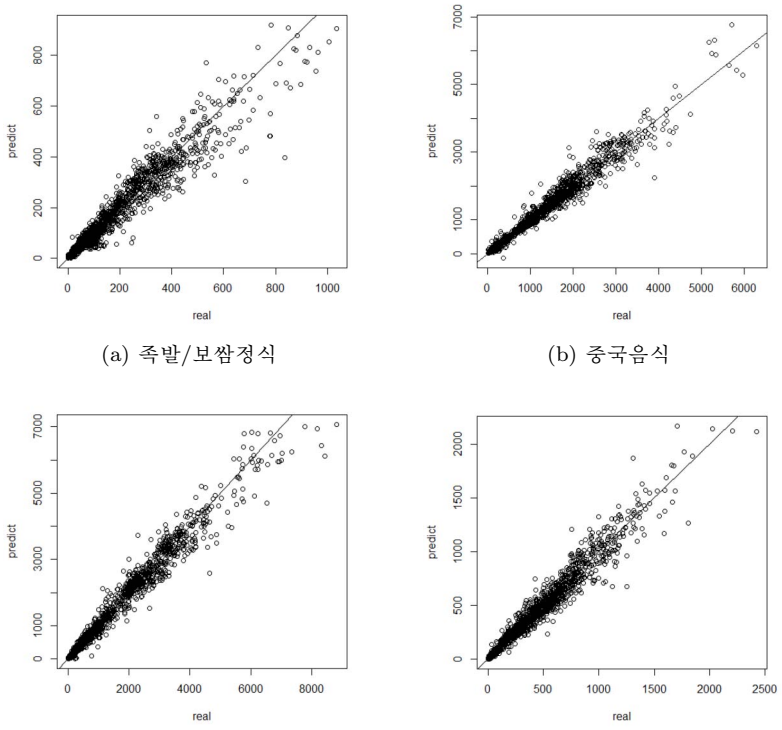
3.1.3. 각 배달업종의 이용건수 예측 3.1.1장의 최적의 모형 3가지(Neural Network Model, Random Forest, Support Vector Machine)로 예측한 총 이용건수와 3.1.2장의 모형 2가지(Linear model, Generalized linear model)로 예측한 각 배달업종의 이용비율을 곱하여 각 배달업종의 이용건수를 예측해본다. 예상했던 대로 선형회귀모형을 이용했을 때, 일반화 선형 모형에 비해서 모든 경우에서 test error가 작았다. 각 배달업종별 이용건수를 예측해 본 결과, 가장 좋은 결과를 보여주는 것은 3.1.1장에서 신경망 모형을, 3.1.2장에서 선형 회귀 모형을 선택했을 때이고 이때 test error가 가장 작았다 (Table 3.5 첫 번째 행). 따라서 위 데이터에 대한 최종 모형으로 신경망 모형 & 선형 회귀 모형을 선택한다.

각 업종별 이용건수의 실제 값과 예측 값을 비교한 Figure 3.3을 살펴보면, 모든 업종에 대하여 충분이 예측이 잘 되었음을 알 수 있다. 또한, 이용건수가 많은 중국음식, 치킨, 피자, 족발/보쌈정식 순으로 예측이 잘 되었다.

3.2. 개별 모형(방법 2)

방법 1에서 총 이용건수를 예측한 후에 개별업종의 배달비율을 예측하였다. 방법 2에서는 각 배달음식별로 모형을 구축해보고 이를 이용하여 개별업종과 총 이용건수를 예측해 보고자 한다. 분석에는 독립변수와 종속변수 간 관계 해석을 위한 선형 회귀 모형과 3.1장에서 가장 좋은 예측력을 보였던 신경망 모형만을 이용하였다.

3.2.1. 4가지 업종별 모형예측 각 배달음식에 대하여 단계적 선택법(stepwise selection)을 실시하여 최적의 선형회귀모형을 구해보았다. 그 결과, adjusted-R^2가 각각 0.9192, 0.9708, 0.9657, 0.9558로



(a) 족발/보쌈정식 (b) 중국음식

(c) 치킨 (d) 피자

Figure 3.3. Predicted values VS True values (Number of each delivery service).

네 가지 배달음식 모두에서 높은 모형 적합도를 보이고 있다. 따라서 선형회귀 모형이 데이터를 잘 설명하고 있다고 판단된다. 배달음식별 선형회귀모형의 결과 유의 수준 0.05하에서 유의한 변수들의 회귀계수는 홈페이지(http://home.ewha.ac.kr/~josong/)에서 확인할 수 있다.

Figure 3.4는 선형회귀모형을 이용하여 각 배달업종별 2014년 1월 1일부터 12월 31일까지 365일의 이용건수를 예측하고, 하루당 평균 이용건수 나타내어본 그래프이다. 주말이 평일에 비해 이용건수가 높기 때문에 7일 단위로 주기성을 보인다. 또한 2.2장에서 언급했던 바와 같이 9월과 10월이 다른 달에 비해 배달음식 이용건수가 증가함(Figure 3.4 검은색 점선)을 확인 할 수 있다.

Figure 3.5는 선형회귀 모형의 회귀 계수들을 살펴본 결과, time변수에 따라 각 배달음식별 이용건수의 패턴이 다름을 확인하고, 각 배달음식 이용건수 예측치의 time별 평균을 나타낸 그래프이다. 전체적으로 모든 업종에서 점심시간대(11시–13시)와 저녁시간대(17시–20시)에 이용건수가 증가함을 확인할 수 있다. 또한 중국음식의 경우 점심시간대에 다른 업종에 비해 월등히 이용건수가 많으며, 치킨업종의 경우 점심시간부터 꾸준히 이용건수가 증가하여 저녁시간대에 가장 이용건수가 많고, 다른 업종과는 달리 밤시간대(21시–24시)까지도 꾸준히 많은 배달이 이루어지는 것이 특징이라 할 수 있다.

다음으로 신경망 모형을 이용하여 네 가지 배달음식 이용건수를 분석해보았다. 신경망 모형의 적합에 필요한 튜닝 모수(size, decay)의 경우, 각 배달음식별 10-fold cross validation 결과 얻은 최적의 튜닝 모수를 이용했을 때와 3.1장에서 이용한 튜닝모수(size = 8, decay = 0.1)을 이용했을 때의 적합도와 예측력 모두 크게 다르지 않아 3.1장에서 이용한 튜닝모수를 그대로 사용하도록 한다.

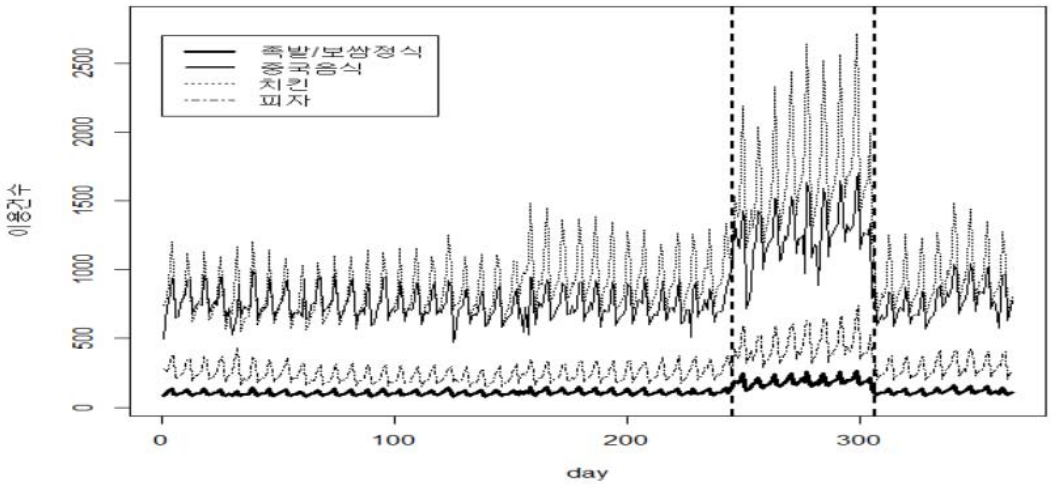


Figure 3.4. Predicted average number of each delivery service per day.

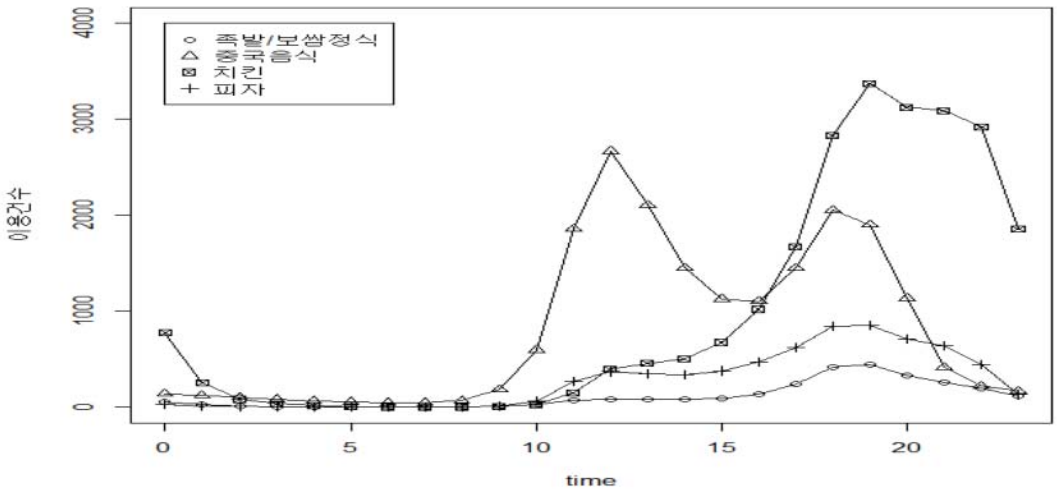


Figure 3.5. Predicted average number of each delivery service per time.

Table 3.7은 선형 회귀 모델과 신경망 모형을 이용해 구한 적합도 및 예측력을 나타낸 것이다. 이용건수가 많은 치킨(C), 중국음식(B), 피자(D), 족발/보쌈정식(A) 순으로 RMSE값이 높으며, 신경망 모형이 선형회귀 모형에 비해 좋은 예측력을 보이는 것을 확인할 수 있다. 따라서 3.2장의 최종모형은 신경망 모형으로 결정한다.

3.3. 모형 비교

Table 3.8은 앞에서 분석한 두 가지 방법론을 비교해 본 것이다. 방법 1과 방법 2에서 최종 선택된 모형의 test error를 비교하면 다음과 같다. 방법 1의 경우 총 이용건수를 신경망 모형으로, 업종별 이용비율을 선형회귀모형으로 선택하였으며, 방법 2의 경우 신경망 모형을 최종 모형으로 선택하였다. 두 방법

Table 3.7. Train and test error of the models (each delivery service)

	Train error		Test error	
	선형회귀 모형	신경망 모형	선형회귀 모형	신경망 모형
족발/보쌈정식(A)	45.80322	29.62474	48.25828	33.95777
중국음식(B)	260.02860	120.32880	240.21160	139.37720
치킨(C)	390.90930	248.83660	399.66450	264.47310
피자(D)	97.32933	64.64778	94.54840	68.81949
전체 데이터	240.82980	142.70220	239.11380	154.32210

Table 3.8. Test error of first and second methods

	방법 1	방법 2
족발/보쌈정식(A)	37.71563	33.95777
중국음식(B)	151.73890	139.37720
치킨(C)	210.78670	264.47310
피자(D)	61.11236	68.81949
전체 데이터	134.73390	154.32210

론의 최종 선택 모형을 비교하면, 족발/보쌈정식, 중국음식 업종의 경우 방법 2에서, 치킨, 피자 업종의 경우 방법 1에서 test error가 작았다. 전체 데이터의 test error를 비교해 보면 방법 1에서 더 작고, 개별 데이터에서는 방법 1이 A, B에서 성능이 조금 떨어지지만 방법 2의 값과 그 차이가 크지 않으므로, 방법 1을 최적모형으로 선택한다.

4. 결론

우리는 다양한 회귀모형을 사용해서 배달음식 건수를 예측해보았다. 우선 하나의 모형으로 전체 배달 건수를 예측한 후에 업종별 배달음식 건수를 예측하는 방법을 제시하였다 (방법 1). 그리고 업종별 배달음식을 개개의 다른 모형으로 예측한 후에 전체 배달 건수를 예측하는 모형을 제시하였다 (방법 2). 전체 배달 건수 예측에서는 방법 1이 더 우수하였고 업종별 배달음식 건수 예측에서는 족발/보쌈정식과 중국음식에서는 방법 2가, 치킨과 피자에서는 방법 1이 더 나은 성능을 보여주었다.

아주 흥미로운 것은 시간과 날씨에 관련된 설명변수만을 가지고 분석을 하였는데 상당히 정확한 예측이 가능하다는 사실이다. 물론 1년 치 데이터만을 이용해서 분석했으므로 연단위 시간이 지남에 따른 변화를 분석할 수는 없었지만 1년 시간 단위 안에서는 배달 음식 건수는 상당히 안정된 모형을 따른 다는 것을 알 수 있다. 최근의 논문 동향을 보면 랜덤 포레스트나 그래디언트 부스팅 같은 앙상불 방법론이 예측력에서 가장 우수한 성능을 보여주는 경우를 많이 볼 수 있는데 본 논문에서 가장 우수한 예측력을 보여준 모형이 신경망 모형이라는 사실도 특이한 점이라고 할 수 있다. 본 예측모형을 이용하면 배달 음식점들이 수요예측을 어느 정도 정확하게 할 수 있으므로 유용할 것이고, 배달 음식 이용추이와 현황을 파악함으로써 배달 업주들 뿐 만아니라 배달 이용 고객들에게도 유용한 정보를 알려줄 수 있을 것이라고 생각된다.

References

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman. J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Friedman, J. (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis*, **38**, 367–378.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York, USA.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer, New York, USA.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006). Support Vector Machines in R Journal of Statistical Software, 15(9).

Park, C., Kim, Y., Kim, J., Song, J. and Choi, H. (2011). *Datamining using R*, Kyowoo, Seoul.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ridgeway, G. (2012). Generalized Boosted Models: A guide to the gbm package.

Thomas, D. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, *Machine Learning*, **40**, 139–157.

국내 배달음식 이용건수 분석 및 예측

권재영[a] · 김시내[a] · 박은지[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2015년 7월 29일 접수, 2015년 8월 23일 수정, 2015년 9월 23일 채택)

요 약

우리나라는 세계적으로 배달음식 문화가 가장 많이 발달한 나라 중에 하나로 최근에는 일인가구의 증가와 배달앱 시장의 발달과 함께 그 성장 속도 또한 눈부시게 증가하고 있다. 따라서 배달음식 이용에 큰 영향을 미칠 것으로 예상되는 날씨와 날짜별 변수를 고려하여 시간대별 배달음식 이용건수를 예측함으로써 소비자와 생산자 모두에게 이익을 주는 예측모형을 찾고자 한다. 본 연구의 목적은 다양한 데이터마이닝 기법을 이용하여 2014년도 배달음식 통화건수를 예측하는데 있다. 예측에 사용되는 회귀 모형은 선형회귀모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 신경망, 로지스틱 회귀모형으로 총 6가지이다. 고려되는 배달음식 업종은 총 4가지(족발/보쌈정식, 중국음식, 치킨, 피자)로 크게 두 가지 방법을 이용하여 각 업종별 배달음식 이용건수를 예측하였다. 첫 번째 방법은 총 이용건수와 각 업종별 배달음식 이용비율을 곱하여 각 업종별 배달음식 이용건수를 예측하는 것이고, 두 번째 방법은 각 업종별 모형을 세워 각 업종별 배달음식 이용건수를 예측하는 방법이다. 최종적으로 선택된 모형은 방법 1에서는 신경망 모형과 선형회귀모형이며, 방법 2에서는 신경망 모형이었다. 방법 2보다는 방법 1로 구한 결과가 더 예측력이 좋은 것으로 나타났다.

주요용어: 배달음식 이용건수, 선형회귀모형, 랜덤 포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 신경망, 로지스틱 회귀모형

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2012817).

[1]교신저자: (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr