The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.3.391
32(3), 391–404

Consumer behavior prediction using Airbnb web log data

Hyoin An[a] · Yuri Choi[a] · Raeeun Oh[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received October 25, 2018; Revised January 15, 2019; Accepted January 19, 2019)

Abstract

Customers' fixed characteristics have often been used to predict customer behavior. It has recently become possible to track customer web logs as customer activities move from offline to online. It has become possible to collect large amounts of web log data; however, the researchers only focused on organizing the log data or describing the technical characteristics. In this study, we predict the decision-making time until each customer makes the first reservation, using Airbnb customer data provided by the Kaggle website. This data set includes basic customer information such as gender, age, and web logs. We use various methodologies to find the optimal model and compare prediction errors for cases with web log data and without it. We consider six models such as Lasso, SVM, Random Forest, and XGBoost to explore the effectiveness of the web log data. As a result, we choose Random Forest as our optimal model with a misclassification rate of about 20%. In addition, we confirm that using web log data in our study doubles the prediction accuracy in predicting customer behavior compared to not using it.

Keywords: web log, customer behavior prediction, machine learning, data mining

1. 서론

전통적으로 고객 행동에 대한 예측은 주로 고객이 가지는 고정적인 특성을 이용해왔으며, 예측의 대상은 고객의 특성에 따라 분류되는 '평균적인 고객'의 행동이었다. 그러나 빅데이터 시대가 도래하면서 방대한 양의 정보가 사용 가능해지고 각 고객 정보에 대한 기록이 용이해짐에 따라, 마케팅 분야에서 고객 행동에 대한 예측이 새로운 화두로 떠올랐다 (Pandagre와 Veenadhari, 2017). 특히 고객들의 활동이 오프라인에서 온라인으로 옮겨오면서 각 고객의 웹 로그 (Web log)를 추적 및 수집하는 것이 가능해 졌다. 이에 따라 웹 로그 데이터가 각 고객의 행동에 대한 숨겨진 패턴을 가지고 있는 핵심적인 정보로 여겨지면서 웹 로그 데이터를 고객의 공통적인 차원의 미래 행동을 예측하는 데 이용한 연구도 등장하였다 (Goel 등, 2010). 더 나아가 개별적인 차원의 고객 행동을 분석하는 데 활용하기 시작하였다. 미국의 가장 큰 온라인 쇼핑 플랫폼인 Amazon은 고객의 온라인 접속 및 구매 기록에 따라 알맞은 상품을 추천하는 알고리즘을 자체적으로 개발하여 큰 상업적인 성공을 거두었으며, 온라인 미디어 공급 사이트인 Netflix의 경우 2006년 100만 불 상당의 상금을 기반으로 한 경진대회를 통해 역시 개별 고객에게 적절한 영화 및 드라마 등의 컨텐츠를 추천해주는 알고리즘을 개발하였다. 하지만 구글 플루 트렌

[1]Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

드(Google Flu Trend; GFT)와 같이 사용자의 온라인 기록을 이용한 분석이 항상 성공적이지만은 않음을 보여주는 사례도 있었다 (Lazer 등, 2014). 이는 2008년 구글에 의하여 처음으로 시작된 프로젝트로 온라인 사용자들의 검색 기록을 이용하여 플루에 대한 예측을 하고자 하는 좋은 시도였지만 실제보다 두 배 이상 과대추정(overestimation)하는 등 목적에 걸맞은 정확한 예측을 하지 못하고 막을 내렸다. GFT의 실패의 원인으로는 플루(Flu)와 연결된 적절한 검색어를 사용하지 못한 것이 언급된다 (Harford, 2014).

웹 로그 데이터는 고객이 웹사이트에 접속하여 클릭함으로써 생성되는 모든 데이터에 대한 기록이다. 기록되는 형식에 따라 차이는 있지만, 일반적으로는 언제 접속했는지, 어디에서 접속했는지(IP주소), 어떤 방법으로 방문했는지, 어떤 브라우저(사파리, 익스플로러 등)를 통해 접속했는지, 어떤 페이지를 로딩했는지 등의 정보를 포함한다. 웹 로그 분석을 하기 시작한 초기에는 웹 로그 데이터를 정리하고 정리한 데이터를 바탕으로 기술적인 특성을 설명하거나, 웹 사이트 디자인에 반영하는 정도에 그쳤다. Igor (2000)는 고객들의 웹 사이트 방문 형태를 다양하게 분류하였으며, Kim (2002)의 연구에서는 웹 로그 파일을 이용한 방문자의 접속 관련 사항에 대한 시간별 기술통계를 이용하여 방문자의 접속실태를 파악하고 여행사의 인터넷 마케팅을 위한 사이트 운영전략에 관한 정책을 수립할 수 있다고 시사하였다. 그러나 점차 온라인 상에서의 고객행동이 복잡해져 감에 따라 최근의 웹 로그 분석은 고객의 행동 실태를 단순히 파악하는 것을 넘어 패턴을 분석하고 예측하는 데에 사용되고 있다. 예를 들어 Sujatha와 Punithavalli (2012)의 연구에서는 분류모형과 클러스터링을 기법을 이용하여 웹 로그 데이터를 분석하고 이를 토대로 개별 사용자가 해당 웹사이트에서 취할 행동을 예측하였다.

본 연구에서는 웹사이트 Kaggle에서 제공한 Airbnb 데이터 셋(https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data)을 이용하여 개별 고객 행동에 대한 예측 모형에서의 웹 로그 데이터의 효용성에 대하여 고찰하였다. 해당 데이터셋은 고객들의 성별, 연령 등의 기본 정보 및 웹 로그 기록을 포함하며, 우리는 첫 숙소 예약까지 걸리는 개인의 의사 결정 시간을 예측하고자 한다. 이 과정에서 데이터를 사전에 적절하게 가공하고 반응변수와 밀접하게 관련 있는 파생변수를 가능한 많이 생성하려는 시도를 통해 예측력을 높이려 하였다. 또한 Lasso (Tibshirani, 1996), SVM (Hearst 등, 1998), RandomForest (Breiman, 2001), GBM (Friedman, 2001), XGBoost (Chen과 Guestrin, 2016) 등 다양한 방법론을 활용하여 방법론에 따른 예측 오차의 차이를 비교하여 최적의 모형을 찾고자 한다.

따라서 본 연구에서는 각 고객의 개인적 특성뿐만 아니라 웹 로그 데이터를 추가적인 정보로 이용하여 각 고객의 특정 행동까지 걸리는 시간을 예측하고, 더 나아가 그를 이용해 고객 분류 모형을 개발하고자 한다. 2장에서는 분석에 사용된 데이터에 대해 자세히 소개하고, 웹 로그 정보를 예측 모형에 반영하기 위해 생성한 파생 변수에 대해 기술하였다. 3장에서는 이를 이용해 자료를 분석한 결과를 회귀 모형과 분류 모형으로 나누어 제시하였다. 마지막으로 4장에서는 본 연구의 결론 및 의의에 대해 서술하였다.

2. 자료 탐색 및 가공

2.1. 자료 탐색

2.1.1. 고객 기본 정보 본 연구에서는 Kaggle에서 제공하는 Airbnb에 관한 데이터셋을 이용하였으며, 이 자료는 성별, 연령 등 각 고객의 고정적인 특성 및 각 고객의 웹 로그 기록을 모두 포함하는 자료이다 (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data). 이 자료는 총 두 개의 데이터 테이블을 포함하고 있으며, 각 데이터 테이블에 속한 변수를 Table 2.1 및 Table 2.2에 나열하였다. 예측할 대상이 되는 고객의 행동은 '개인의 첫 숙소 예약까지 걸리는 일 수(duration)'으로 하였으며, 이 반응변수는 0부터 365까지의 정수 값을 가진다.

Table 2.1. Descriptions on customer information variables

Variable	Class	Description
id	Character	각 고객의 고유 아이디
date-account-created	Date	각 고객이 Airbnb 계정을 만든 일시
timestamp-first-active	Double	각 고객이 Airbnb에서 첫 활동을 시작한 일시
date-first-booking	Character	각 고객이 처음으로 Airbnb에서 예약한 일시
gender	Character	각 고객의 성별
age	Double	각 고객의 연령
signup-method	Character	각 고객의 회원가입 방법 (예: 기본, 페이스북)
signup-flow	Character	각 고객의 회원 가입 경로
language	Character	각 고객의 사용 언어
affiliate-channel	Character	Airbnb 광고 형태
affiliate-provider	Character	Airbnb 마케팅 제휴 회사
first-affiliate-tracked	Character	각 고객이 처음 Airbnb에 접속한 마케팅 경로
signup-app	Character	각 고객의 회원 가입 어플리케이션
first-device-type	Character	각 고객의 첫 기기 종류 (예: 데스크탑, 아이폰, 아이패드 등)
first-browser	Character	각 고객의 첫 브라우저 (예. 인터넷 익스플로러, 사파리 등)
country-destination	Character	각 고객의 첫 방문 국가

Table 2.2. Descriptions on web log information variables

Variable	Class	Description
user-id	Character	각 고객의 고유 아이디
action	Character	각 고객의 'Action' 웹 로그
action-type	Character	'Action'과 연결되는 'Action-type' 웹 로그
action-detail	Character	'Action'과 연결되는 'Action-detail' 웹 로그
device-type	Character	웹 로그 생성 당시의 접속 기기
secs-elapsed	Double	해당 세션 소요 시간

총 28,409명의 개별 고객 중, 기본 정보 및 웹 로그 세션 정보를 포함하며 이상치를 제외한 20,559명의 개별 고객의 데이터를 분석에 사용하였다. 기본 정보는 Table 2.1에 정리하였다. 변수 date-account-created는 각 고객이 에어비앤비 계정을 만든 일시이고, timestamp-first-active는 각 고객이 에어비앤비에서 첫 활동을 시작한 일시이다. 두 변수 모두 2014년 1월 1일부터 2014년 6월 30일까지의 값을 가지며, 20,559명의 개별 고객 모두에 대해 두 변수의 값이 동일하다. 따라서 반응변수(duration)는 각 고객이 처음으로 에어비앤비에서 예약한 일시인 date-first-booking에서 date-account-created 또는 timestamp-first-active를 뺀 값으로 생각할 수 있다. 반응 변수 생성에 대해서는 다음 절에서 더욱 자세히 논의할 것이다. gender는 각 고객의 성별을 나타내는 변수인데, 37%의 고객은 Male, 42%의 고객은 Female, 21%의 고객은 unknown으로 분류되어 있다. age 평균은 35.7세이며, 잘못 기입된 경우는 분석 대상에서 제외하였다. 10세 미만과 80세 이상, 그리고 연도를 기입한 경우에는 잘못 기입되었다고 판단하여 분석에서 제외하였다. signup-method는 각 고객이 어떤 방법으로 회원가입을 했는지를 나타내는데, 대다수는 새로운 계정을 생성하거나 페이스북 아이디를 연동하여 가입하였으며 구글 계정을 통해 회원가입한 고객도 소수 존재한다. affiliate-provider 및 affiliate-channel는 에어비앤비와 마케팅 제휴를 맺은 회사 및 광고 형태를 나타내는 변수이다. first-device-type은 각 고객이 회원가입 시 어떤 기기를 이용하였는지(예: 데스크탑, 아이폰, 아이패드)를 나타내고, first-browser는 어떤 브라우저를 통해 들어왔는지(예: 인터넷 익스플로러, 사파리)를 나타낸다. 이 외에도 사용하는 언어를 나타내는 language, 각 고객이 처음으로 방문했던 국가인 country-destination 변수가 있다.

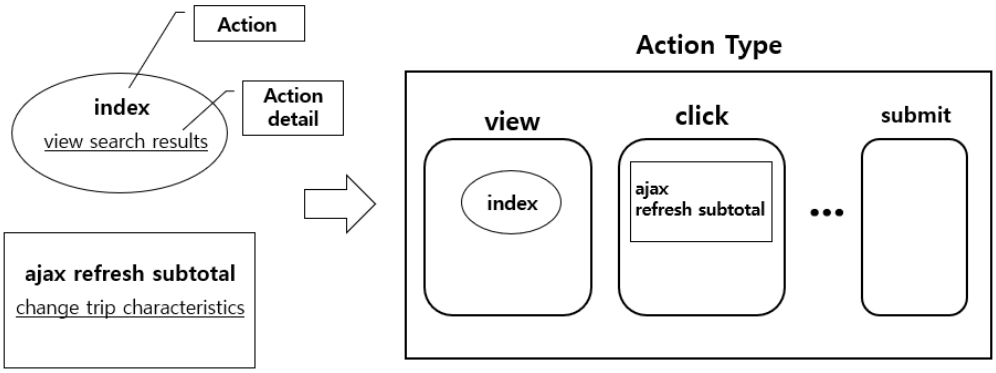


Figure 2.1. Visualization of web log data structure.

2.1.2. 고객 웹 로그 정보 웹 로그 정보가 기록되는 형태는 웹사이트마다 다르다. 우리가 분석에 사용한 에어비앤비 사이트에 기록된 웹 로그 데이터의 구조를 Table 형식으로 정리한 것이 Table 2.2이다. user-id는 각 고객에게 할당되는 고유 아이디로 Table 2.1의 id와 동일하다. 이 id를 기반으로 고객이 온라인 상에서 특정 행동을 취하면 그 웹 로그 정보가 기록되는데, 이 웹 로그 정보는 303개의 활동(action)과 11개의 활동속성(action-type), 그리고 126개의 상세활동(action-detail)으로 구성되어 있다.

예를 들어 Figure 2.1을 보면, 어떤 고객이 'index'의 활동을 했을 때 해당 활동이 'view'라는 활동속성을 가지며 'view search results'라는 상세활동 내역이 남는다. 따라서 이 고객이 웹에서 검색한 결과를 보고 있음을 알 수 있다. 만일 어떤 고객이 'ajax refresh subtotal'의 활동을 했다면 이는 'click'이라는 활동속성을 가지고, 'change trip characteristics'의 세부활동을 가지므로 고객이 여행의 세부적인 사항들을 변경하려 한다는 사실을 알 수 있다. 따라서 고객마다 취한 활동, 활동속성 및 상세활동은 그 내용만 다른 것이 아니라 횟수도 모두 다르다. 예를 들어 분석에 사용한 데이터에서는 단 하나의 웹 로그 기록만을 가진 고객부터 최대 2643개의 웹 로그 기록을 가진 고객까지 존재한다. 평균적으로는 한 고객당 86개의 웹 로그 기록을 가지고 있다. 웹 로그 기록이 5개 이하인 고객은 정보가 불충분하다고 생각하여 분석 대상에서 제외하였다. 각 웹 로그 활동 소요시간이 24시간 초과인 경우도 일반적이지 않다고 생각하여, 그러한 웹 로그 기록의 경우 소요시간을 24시간으로 치환하였다.

2.2. 분석목적 및 반응변수 설정

본 연구에서는 각 고객의 정보가 포함되어 있는 데이터 테이블 및 각 고객이 온라인에서 활동한 웹 로그 자료를 이용하여 '첫 숙소 예약'까지 걸리는 개인의 의사 결정 시간을 예측하기로 한다. 또한 각 시간을 반응변수로 하는 회귀 모형과 각 고객 그룹을 반응변수로 하는 분류 모형을 상정해 다각도로 고객 행동에 대해 분석을 시도하였다. 이에 따라 먼저 회귀 및 분류 모형에 대한 반응 변수를 아래와 같이 각각 생성하였다.

1. Duration: Airbnb 데이터셋에 포함된 date-first-booking(처음 예약한 날짜) 변수와 timestamp-first-active(첫 온라인 활동 날짜 및 시간) 변수의 차이를 이용하여 Duration(온라인 활동 시작 후 처음으로 예약하는 데까지 걸리는 시간) 변수를 생성하였다. date-first-booking이 결측치인 경우는 데이터에서 제외하였다. Duration 변수의 범위는 0(1일 이내)부터 365(1년)이며, 평균은 50.68, 중

Table 2.3. Percentage of customers by Duration's category

Fast (2일 이내)	Slow (2일 이상)
39.06%	60.94%

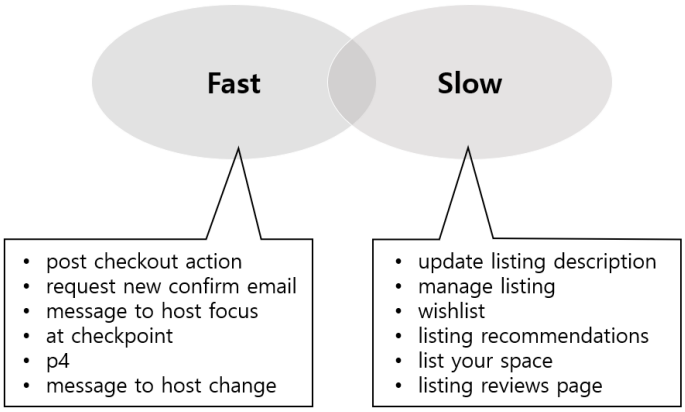


Figure 2.2. Generation of derived variables using 'Anti join' method.

앙값은 4이다.

2. Duration group: 고객 분류를 하기 위하여 고객 그룹을 Duration을 기준으로 2개로 나누었다. Fast (2일 이내), Slow (2일 이상)으로 분류하였다. 각 범주별 고객의 비율은 Table 2.3과 같다.

2.3. 변수 선택 및 파생 변수 생성

고객들의 기본 정보 외에 웹 로그 데이터를 통계 모형에 반영하기 위하여 상세활동에 대한 파생 변수를 생성하기로 하였다. 고객의 duration별로 다른 값을 가지는 파생변수들을 생성하기 위해서 많은 시도를 하였고, 그 중에서 총 5가지 방법을 선별하여 모형의 예측력을 높이는 데에 기여하는 파생변수들을 생성하였다. 각 파생 변수들의 생성 과정은 아래와 같다.

2.3.1. 기초 작업: 상세활동에 대한 가공 고객 웹 로그 정보 테이블의 상세활동 열은 총 126가지의 상세활동을 가진다. 개인 별 가지는 상세활동의 종류 및 횟수가 다르기 때문에, 이 정보를 다른 형태의 테이블로 정리할 필요가 있었다. 따라서 본 프로젝트에서는 126가지의 상세활동을 각각의 열로, 20,559개의 ID를 행으로 하는 20559×127 크기의 테이블을 생성하였고 (ID 열 포함, ID 열 제외시 20559×126), 그 값으로는 각 개인이 특정 상세활동을 한 횟수를 넣었다. 이 테이블을 편의상 **D**라고 정의한다. 만약, 어떤 개인이 특정 상세활동을 한 번도 하지 않았다면, **D**에 들어가는 값은 0으로 기록된다.

2.3.2. 차집합을 이용한 'Anti join' 방법 고객 그룹 간 유의한 차이를 보이는 상세활동들을 선별하기 위하여 차집합을 이용하였다. 먼저 각 그룹별로 빈도가 높은 상위 50개의 활동들을 선택한 후, 의사결정이 빠른 그룹의 활동들을 기준으로 상대적으로 의사결정이 느린 그룹이 하지 않은 활동들을 선별하고 반대의 경우도 같은 방식으로 진행하였다. 이 때 선별된 12가지 활동들은 각 그룹별로 다른 특성들을 반영하는 중요 변수로 고려하였다. Figure 2.2를 보면, 'Fast' 그룹 기준 'Slow' 그룹이 하지 않

은 상세활동으로 'post checkout action', 'request new confirm email', 'message to host focus', 'at checkpoint', 'p4', 그리고 'message to host change'의 여섯 가지 상세활동들이 선별된 것을 확인할 수 있다. 선별된 세 가지 상세활동들은 고객들의 의사결정을 빠르게 하는 중요한 변수가 될 수 있다. 대체적으로 의사결정이 빠른 그룹은 직접적으로 호스트와 연락하는 등 주로 적극적인 활동을 많이 하고, 의사결정이 느린 그룹은 다른 고객들의 후기를 보거나 wishlist에 추가하는 등 주로 탐색하는 활동을 많이 한다고 볼 수 있다.

2.3.3. 고객의 Duration에 따른 가중치를 부여한 'Scores' 방법 기록된 웹 로그 파일에는 하나의 세션에서 얼마나 많은 시간을 보냈는 지에 대한 정보인 'time-elapsed'라는 변수가 존재한다. 이를 모형에 반영하기 위하여 개인이 하나의 웹 로그 상세활동을 몇 번 했는지에 대한 값에 시간을 가중치(weight)로 줄 수 있는 방법을 고안하였다. 이를 이용하여 파생 변수를 생성하는 방법은 아래와 같다.

- Step 1. 아래와 같이 Score 계산에 필요한 새로운 변수들을 생성한다.
  - customer group: $g \in \{f(Fast), s(Slow)\}$
  - $medtime$: median of the time elapsed for each action-detail of the customers in the group g
  - cnt: the number of each action-detail of the customers in the group g
  - $\mathbf{T}_g$: $timepercnt = medtime/cnt$ in the group g if $cnt > 500$

ID	Action detail	Secs elapsed
1	signup_login_page	--
	change_password	--
2	account_transaction_history	--
	change_password	--
	user_wishlists	**513**
	message_post	--
	user_wishlists	**1047**
3	airbnb_picks_wishlists	--
	user_wishlists	**286**
	profile_verifications	--
4	**user_wishlists**	**202**
	message_post	--
⋮	⋮	⋮

(예) Fast 그룹

① 'user wishlist'의 소요시간(초)의 중앙값($medtime$) 계산

② 'user wishlist'의 총 횟수(cnt) 계산

$$\mathbf{T}_f \equiv timepercnt = \frac{medtime}{cnt}$$

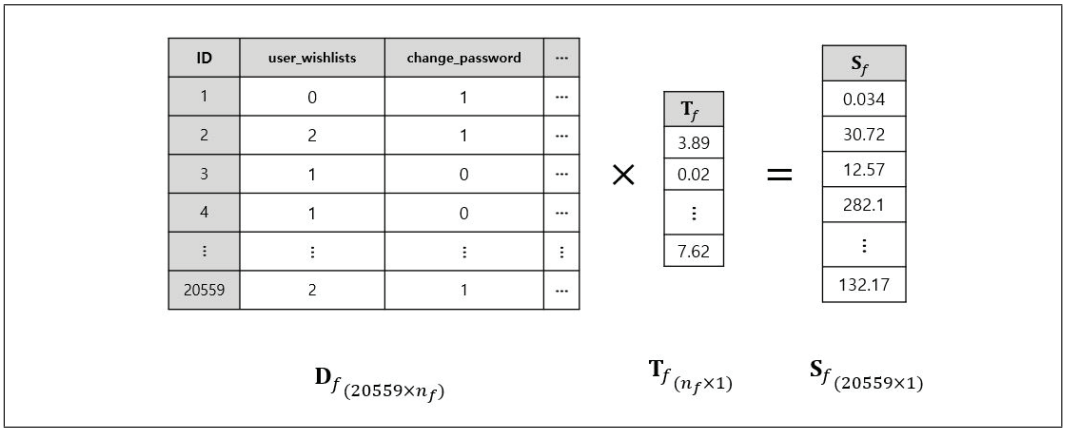
'user wishlist' 대표 소요시간(초)

- Step 2. n_g를 그룹 g에서 $cnt > 500$인 상세활동의 개수라고 정의하고, 기초 작업 단계에서 생성한 행렬 $\mathbf{D}$의 열 중에서 그룹 g 내에서 존재하는 상세활동만을 열로 선택한 행렬을 $\mathbf{D}_g$라고 정의한다.
- Step 3. $\mathbf{D}_g$ 행렬과 $\mathbf{T}_g$를 아래와 같이 곱하여 모든 고객 ID에 부여하는 각 그룹에 대한 Score인 $\mathbf{S}_g$를 계산한다. 수식 아래 첨자에 행렬 차원을 함께 나타냈다. Fast 그룹에 대한 예시는 아래 그림과 같다.

$$\mathbf{S}_{g(n\times 1)} \equiv \mathbf{D}_{g(n\times n_g)}\mathbf{T}_{g(n_g\times 1)} \tag{2.1}$$

Table 2.4. Average score by customer groups

Duration group	$\mathbf{S}_{\text{Fast}}$	$\mathbf{S}_{\text{Slow}}$
Fast	76.1	86.4
Slow	41.8	42.9



생성한 파생변수인 두 개의 Score를 고객 군별로 살펴본 결과를 아래 Table 2.4에 나타냈다. 모든 파생변수에 대해서 Fast 고객군의 평균 값이 Slow 고객군의 평균 값보다 더 높다. 다시 말해 고객의 예약까지 걸리는 시간이 빠를 수록 Score값이 커지는 경향이 있었다. 따라서 이 파생변수가 각 고객군의 차이를 반영한다고 판단하여 예측 모형에 사용하기로 하였다.

2.3.4. 각 그룹 별 유의한 차이의 횟수를 가진 행동들에 대한 선별 우리는 각 그룹 별로 특정 상세활동을 실행한 횟수가 그룹 간 차이를 반영할 것으로 기대하였다. 예를 들어 상세활동 A에 대하여 Fast 그룹에 속한 고객들은 평균적으로 5회 해당 상세활동을 하였고, Slow 그룹에 속한 고객들은 평균적으로 20회 해당 상세활동을 하였다고 가정해 보자. 그렇다면 어떤 고객을 두 개의 그룹 중 하나로 분류하는데 있어서 상세활동 A를 한 횟수가 적을수록 상대적으로 Fast에 속할 가능성이 커질 것이고, 반대로 상세활동 A를 한 횟수가 많을수록 Slow 그룹에 속할 가능성이 높아질 것이다. 따라서 아래와 같은 알고리즘을 통해 126개 상세활동 중 그룹 별로 유의한 횟수 차이를 가질 것으로 기대되는 상세활동만을 선택하였고, 이를 최종 모형에 반영하였다.

또한, 2.3.2장의 'Anti join' 방법에서는 그룹별로 빈도가 높은 상위 50개 상세활동들만을 먼저 선택한 후 이들을 비교하였지만, 2.3.4장의 방법에서는 전체 상세활동에 대해 평균 횟수를 먼저 계산한 후 126개의 모든 상세활동을 그룹별로 비교하였기 때문에, 'Anti join' 방법에서 빠뜨린 중요한 상세활동을 추가적으로 선별할 수 있다.

- Step 1. Duration-group이 Fast, Slow인 경우 각각에 대해 상세활동에 대한 평균 횟수를 구한다.
- Step 2. 평균 횟수가 0.5 이하인 상세활동에 대해서는 충분한 정보력이 없다고 판단하여 제외하고, 0.5 초과인 상세활동만을 2개 그룹별로 추출한다.

ex) Fast 그룹

ID	read_policy_click	change_password	user_wishlists	message_post	···
1	0	1	1	1	···
4	0	0	1	1	···
⋮	⋮	⋮	⋮	⋮	⋮
19527	2	1	5	0	···

평균횟수	0.34	0.22	0.75	1.54	···

평균 횟수가 **0.5 이상**인 Action detail 1차적인 선별

- Step 3. 추출된 상세활동의 평균 횟수를 그룹별로 비교하여 선택하는데, 다음 조건 중 하나라도 만족하면 해당 상세활동을 선택한다.
  - 모든 그룹에서 해당 상세활동이 선택되었고, 2개 그룹 간 차이가 1 이상인 경우
  - 최소 한 그룹에서 해당 상세활동이 선택되지 않은 경우

1차적으로 선별된 Action detail

	Fast	Slow	
pending	0.6273	.	→ 최종선택
confirm_email_link	0.8081	0.7041	→ X
⋮	⋮	⋮	
wishlist_content_update	5.7937	7.6045	→ 최종선택
⋮	⋮	⋮	
change_contact_host_update	.	0.8104	→ 최종선택
⋮	⋮	⋮	

- Step 4. 그 결과 선택된 변수는 다음과 같다:

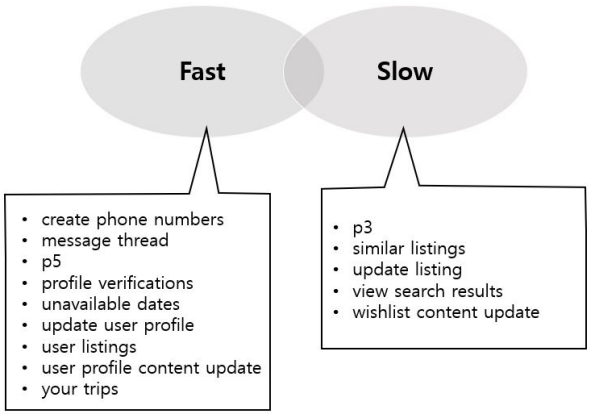


Table 3.1. Comparison of 10-fold CV error and Test error by regression models

Regression	CV error		Test error	
	웹 로그 정보 있음	웹 로그 정보 없음	웹 로그 정보 있음	웹 로그 정보 없음
Multiple linear	84.8964	88.9452	**84.9355**	90.7857
Lasso	84.9812	88.9653	86.8818	90.7406
SVM	92.5478	96.8626	95.3234	99.4000
Random Forest	85.3149	90.9242	85.5741	92.6140
GBM	**83.9581**	90.8012	85.7169	92.0252
XGBoost	86.6157	91.2880	86.6157	92.4707

2.3.5. 총 웹 로그 활동 수 및 총 웹 로그 활동 소요 시간 각 고객별로 수행한 모든 웹 로그 활동의 횟수(total-activity) 및 총 소요 시간(total-time)이 고객의 의사 결정 시간과도 연관이 되어 있을 것이라고 판단하여 이 두 변수 역시 모형의 변수로 사용하였다.

2.3.6. 주성분 분석 고객 웹 로그 정보 테이블에서 303개의 활동에 대한 정보를 추가적으로 반영하기 위하여 고객별로 각 활동에 대한 횟수를 세어 303개 활동의 이름을 열로 가지는 테이블을 만들었다. 그리고 주성분 분석(principal component analysis; PCA)를 이용하여 분산비가 각각 0.4955, 0.4273로 전체 분산의 약 92%를 설명하는 2개의 주성분(PC1과 PC2)을 모형의 변수로 사용하였다.

3. 결과

3.1. Duration 회귀 모형

먼저 개별 고객이 최종 숙소 예약을 할 때까지 걸리는 시간을 예측하기 위해 다중 선형 회귀 모형, 라소(Lasso) 모형, 서포트 벡터 기계(support vector machine; SVM), 랜덤 포레스트(random forest) 및 그래디언트 부스팅(gradient boosting) 모형을 고려하였다. 추가적으로 그래디언트 부스팅의 응용 모형인 XGBoost 까지 이용하여 도합 6가지의 모형을 적합하였다. 웹 로그 정보의 효용성을 알아보기 위해, 각 모형에 대해 고객의 기본 정보만 이용하였을 때와 기본 정보 및 웹 로그 정보 모두를 이용하였을 때의 두 경우로 나누어 결과를 정리하였다. 분석 순서는 다음과 같다. Train 및 Test Dataset을 7 : 3의 비율로 나누어, Train Set에서 10-fold Cross-Validation Error를 구하였다. 이 때 Error는 root mean squared error (RMSE)이다. 그리고 Train Set을 이용하여 적합한 모형으로 Test Set에서의 RMSE를 계산하였다.

Table 3.1은 웹 로그 정보를 이용하였을 때와 그렇지 않았을 때 Train Set과 Test Set에서 각 모형의 RMSE를 정리한 결과이다. Table 3.1을 살펴보면 웹 로그 정보를 사용한 모형이 사용하지 않은 모형보다 RMSE가 낮은 것을 알 수 있다. 따라서 모형의 종류와 상관없이 웹 로그 정보는 예측 정확도를 향상시킨다고 할 수 있다. 모형에 따른 RMSE를 비교한 결과 Train Set에서 GBM의 10-Fold CV RMSE가 83.9581로 가장 낮고, Test Set에서는 Multiple Linear의 RMSE가 84.9355로 가장 낮다. Stepwise Regression을 이용하여 다중 선형 회귀분석을 한 결과, at-checkpoint와 p5의 상세활동을 많이 할수록 예약을 빠른 시일 내에 하는 경향이 있었다. 반대로 message-to-host-focus, manage-listing, list-your-space, listing-reviews-page, create-phone-numbers의 상세활동을 많이 할수록 예약을 결정할 때까지 시간이 더 많이 걸리는 경향이 있는 것을 알 수 있었다.

3.2. Duration group 분류 모형

다음으로는 각 고객이 어느 Duration Group에 속하는 지를 예측하기 위한 분류 모형을 상정하였다. 분

Table 3.2. Comparison of 10-fold CV error and Test error by classification models

Classification	CV error		Test error	
	웹 로그 정보 있음	웹 로그 정보 없음	웹 로그 정보 있음	웹 로그 정보 없음
Multinomial	0.2171	0.3971	0.2215	0.4063
Lasso	0.2165	0.4041	0.2261	0.4014
SVM	0.2199	0.4039	0.2289	0.4014
Random Forest	**0.2080**	0.4144	**0.2085**	0.4295
GBM	0.2099	0.4010	0.2132	0.4027
XGBoost	0.2127	0.4082	0.2130	0.4126

Table 3.3. Confusion matrix of the Random Forest model

		True	
		Fast	Slow
Predicted	Fast	1600	410
	Slow	876	3282

류를 위한 모형으로는 다항(multinomial) 분류모형, 라소 모형, 서포트 벡터 기계, 랜덤 포레스트 및 그래디언트 부스팅 모형, 그리고 XGBoost 모형까지 총 7가지의 모형을 고려하였다. 회귀 모형과 마찬가지로 각 모형에 대해 고객의 기본 정보만 이용하였을 때와 기본 정보 및 웹 로그 정보 모두를 이용하였을 때를 구분하여 결과를 정리하였으며, 분석 순서는 다음과 같다. Train 및 Test Dataset을 7 : 3의 비율로 나누어, Train Set에서 10-fold Cross-Validation Error를 구하였다. 이때의 'Error'는 오분류율(misclassification rate)이다. 그리고 Train Set을 이용하여 적합한 모형으로 Test Set에서의 오분류율을 계산하였다.

Table 3.2는 웹 로그 정보를 이용하였을 때와 그렇지 않았을 때 Train Set과 Test Set에서 각 모형의 오분류율을 정리한 결과이다. 여기서 주목할 만한 점은 고객 기본 정보만 모형을 적합한 경우는 Test Set에서 모형에 상관없이 오분류율이 40% 이상인 것에 비해 고객 기본 정보에 추가적으로 웹 로그 데이터까지 추가적으로 이용하여 모형을 적합한 경우, 모형의 종류에 상관없이 오분류율이 23% 미만으로 이전보다 최소 17% 이상, 최대 두 배 정도 감소한다는 점이다. 따라서 분류 모형에서도 모형의 종류와 상관없이 웹 로그 정보는 예측 정확도를 향상시킨다고 할 수 있다. 모형 별 오분류율을 비교해보면, Train Set에서는 2개 범주에 대한 랜덤 포레스트의 10-Fold CV Error가 약 20.8%로 가장 낮고, Test Set에서도 역시 오분류율이 약 20.85%로 가장 낮았다. 따라서 Train Set과 Test Set에서 모두 성능이 가장 좋은 랜덤 포레스트를 최종 모형으로 선택한다. 표3.3에 랜덤 포레스트 모형의 분류 결과 표를 나타냈다. 이를 살펴보면 실제 2476명의 'Fast' 그룹 고객 중 64.62%를 'Fast' 그룹으로 제대로 분류하였고, 35.38%를 'Slow' 그룹으로 잘못 분류한 것을 알 수 있다. 또한 실제 3692명의 'Slow' 그룹 고객 중 76.27%를 'Slow' 그룹으로 제대로 분류하였고, 23.73%를 'Fast' 그룹으로 잘못 분류하였다.

Figure 3.1의 변수 중요도 그림에서 PC1, PC2가 높은 중요도를 가진 것으로 보아 모형에서 중요한 변수로 고려된 것을 관찰할 수 있다. 이를 통해 웹 로그의 활동 정보를 사용한 것이 모형 예측력을 높이는 역할을 하는 것을 알 수 있다. 마찬가지로 'Scores' 방법으로 생성된 파생변수인 fscore 및 sscore의 중요도 역시 높은 것으로 보아, 우리가 생성한 파생변수가 각 고객을 'Fast' 그룹과 'Slow' 그룹으로 분류하는 데에 도움이 된 것으로 보인다.

변수 중요 그림에서 중요하다고 여겨진 변수들이 대부분 해석이 어려운 파생 변수이므로, 종속성 그림으로 그려서 설명하기 어려운 경우가 많았다. 따라서 설명이 가능한 상세활동 중 대표적인 두 개의 변수에

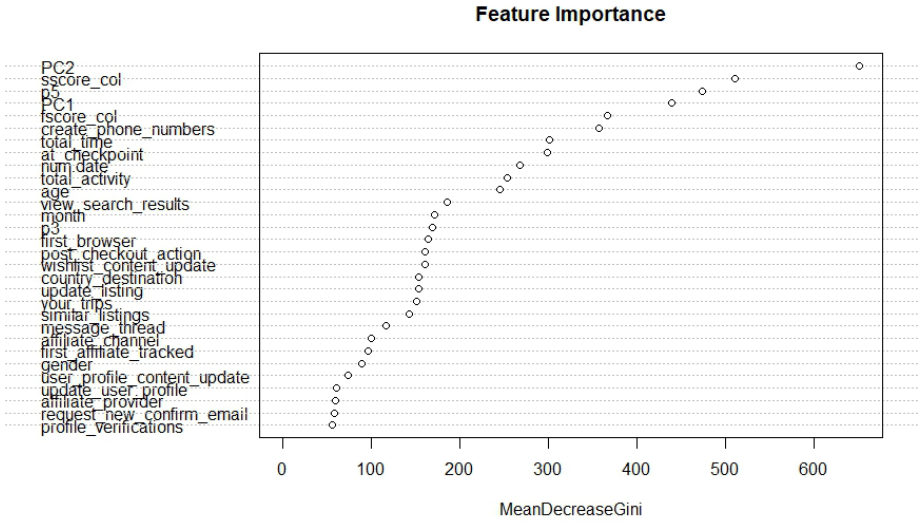


Figure 3.1. Variable importance plot of the Random Forest model.

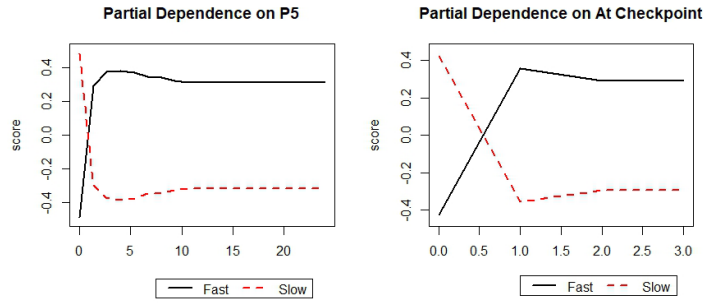


Figure 3.2. Partial Dependence plot of variables with high importance.

대하여 종속성 그림을 그렸고, 이를 Figure 3.2에 나타내었다. 이를 통해 각 상세활동이 고객 그룹을 어떻게 나누는 지 알 수 있다. 먼저 'p5'는 실제 자료에서 'requested(요청됨)'이라는 활동과 연결되어 있는데, 이 상세활동을 많이 할수록 예약을 빠른 시일 내에 하는 그룹으로 분류되는 경향이 있다. 이는 예약 과정에서 무언가를 '요청'하는 적극적인 행동을 보이는 고객이 더 신속하게 의사결정을 하기 때문인 것으로 생각된다. 다음으로 'at-checkpoint(저장 시점)'는 'booking-request(예약 요청)'이라는 활동과 연결되어 있는데, 이 상세활동 역시 예약으로 이어지는 적극적인 행동이므로 빠른 시일 내에 예약을 하는 그룹이 이 상세활동을 많이 하는 경향이 있다.

참고로 303개의 웹 로그 활동에 대한 PC 파생 변수들이 높은 중요도를 보였기 때문에, 126개의 웹 로그 상세활동에도 PCA를 취하여서 파생변수로 추가한 후 분석해 보았다. 그 결과 Test Set에서의 오분류율이 0.2049로 1% 미만의 미미한 향상력을 보였다. 또한, 상세활동에 대한 다른 파생변수를 사용하지 않고, 웹 로그 상세활동 및 활동속성에 대한 PC 파생 변수들만 사용한 경우의 오분류율은 0.2258로 모형의 예측력이 다소 하락한 것을 확인하였다.

4. 결론

회귀 모형의 경우 웹 로그 데이터를 사용한 경우에 모형에 상관없이 RMSE가 사용하지 않은 경우보다 더 낮았다. 그러나 모형 간에 큰 차이가 없었으며, 이를 통해 이 데이터를 이용하여 회귀 모형을 적합하여 예측하는 것에는 한계가 있다고 판단하였다. 특히 부스팅 모형과 같이 복잡한 모형이 Test Data에서는 다중 선형 회귀 모형 보다도 성능이 더 안 좋은 것을 통해 모형의 복잡성이 이 데이터에서 예측의 정확도에 기여하지 않는다는 것을 알 수 있었다. 서포트 벡터 기계의 경우는 Train Data와 Test Data 모두에서 다중 선형 회귀 모형의 RMSE보다 높은 RMSE값을 보였다. 반면 분류모형의 경우 회귀 모형과 마찬가지로 모형 간의 예측력이 큰 차이는 없었으나, 2개 범주에 대한 분류 모형 문제에서 예측치에 대해 무작위로 하나의 범주를 부여할 시 오분류율의 기댓값이 50% 인 것을 고려했을 때 예측 모형의 성과가 있다고 말할 수 있다. 우리는 최종모형으로 랜덤 포레스트 분류모형을 선택하였으며, 이 때 오분류율은 20.85%였다. 그리고 변수 중요도 그림을 통해 활동 정보를 활용한 것이 모형 예측력 향상에 도움이 되고, 생성한 Scores 파생변수 역시 고객 그룹을 분류하는 데에 주요 변수로 작용한 것을 확인할 수 있었다. 또한 적극적인 행동과 관련된 상세활동들이 의사결정을 빨리 하는 그룹의 특징을 잘 보여주는 변수임을 확인하였다.

본 연구가 갖는 의의는 웹 로그 데이터가 잠재적으로 갖는 효용성을 보여주는 것에 있다. 주목할 만한 점은 그림 3.1의 변수 중요도 그림에서 상위 5개의 변수가 모두 웹 로그 활동 또는 이에 대한 파생변수라는 점이다. 그리고 웹 로그 데이터를 이용하여 고객 개개인의 행동을 예측한 결과 웹 로그를 사용하지 않은 경우와 비교해 예측의 정확도가 최대 두 배 더 높아졌다. 이는 웹 로그가 개개인 행동 양상을 파악하여 의사결정을 예측하는 데에 중요하게 활용될 수 있음을 시사한다. 본 분석에서는 반응변수로 '예약까지 걸리는 시간'을 설정하였지만 기업의 필요에 따라 원하는 반응변수를 매번 새로 설정하는 것이 가능하므로 개인의 행동을 예측하는 측면에 있어서는 더 많은 분야에서 활용이 가능하다. 이를 통해 기업은 특정 고객을 상대(예. 의사 결정에 소요되는 시간이 애매한 고객들)로 더 활발한 프로모션을 진행하는 등 더욱 생산적인 마케팅 전략을 세울 수 있을 것이다. 또한 각 웹 로그 기록 시각, 예약의 목적, 예약 기간 등 다양하고 많은 정보를 가지고 있는 고질의 데이터가 아닌 한정적인 양의 정보를 지닌 데이터만을 이용하여 예측력을 개선하였다는 점에서도 의의를 갖는다.

본 연구에서는 주어진 웹 로그 데이터에 대한 횟수 및 소요 시간에 대한 정보는 이용하였으나, 일련의 웹 로그 활동들의 발생을 독립적으로 상정하여 각 로그 활동이 갖는 순서적 흐름(sequence)의 정보는 고려하지 않았다. 따라서 앞으로의 연구에서는 일련의 웹 로그 활동의 순서적 흐름에 대한 정보를 반영한 모형을 고려할 수 있을 것이며, 이것이 모형의 예측력을 더욱 높일 것으로 생각된다.

References

Breiman, L. (2001). Random forests, *Machine Learning*, **13**, 5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 .

Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine, *The Annals of Statistics*, **29**, 1189–1232 .

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. In *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 17486–17490.

Harford, T. (2014). Big data: are we making a big mistake?, *Significance*, 14–19.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines, *IEEE Intelligent Systems and their Applications*, **13**, 18–28.

Igor, V. C., Scott, G., and Smyth, P. (2000). A general probabilistic framework for clustering individuals and objects. In *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining 2000*, 140–149.

Kim, J. K. (2002). A study of web log file analysis for internet marketing of travel agency, *Journal of Tourism and Leisure Research*, **13**, 147–160 .

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis, *Science*, **343**, 1203–1205.

Pandagre, K. N. and Veenadhari, S. (2017). Data mining techniques with web log, *International Journal of Advanced Research in Computer Science Transactions on Pattern Analysis and Machine Intelligence*, **8**, 384–386.

Sujatha, V. and Punithavalli (2012). Improved user navigation pattern prediction technique from web log data, *Procedia Engineering*, **30**, 92–99.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

에어비앤비(Airbnb) 웹 로그 데이터를 이용한 고객 행동 예측

안효인[a] · 최유리[a] · 오래은[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2018년 10월 25일 접수, 2019년 1월 15일 수정, 2019년 1월 19일 채택)

요 약

그동안의 고객 행동에 대한 예측은 주로 고객이 가지는 고정적인 특성을 이용해왔다. 최근에는 점차 고객들의 활동이 오프라인에서 온라인으로 이동하면서 각 고객의 웹 로그를 추적하는 일이 가능해졌다. 그러나 방대한 양의 웹 로그 데이터를 수집할 수 있게 된 반면, 이에 대한 연구는 로그 데이터를 정리하거나 기술적인 특성만을 설명하는 것에 그쳤다. 본 연구에서는 웹사이트 Kaggle에서 제공하는 Airbnb 고객들의 성별, 연령 등의 기본 정보 및 웹 로그가 포함된 데이터셋을 이용하여 첫 숙소 예약까지 걸리는 개인의 의사 결정 시간을 예측하였다. Lasso, SVM, Random Forest, XGBoost 등 다양한 방법론을 활용하여 최적의 모형을 찾고, 웹 로그 데이터의 유무에 따른 예측 오차를 비교하여 웹 로그의 효용성을 확인하였다. 결과적으로 오분류율이 약 20%로 낮은 랜덤 포레스트 분류모형을 최적모형으로 선택하였다. 또한, 웹 로그 데이터를 이용하여 고객 개개인의 행동을 예측한 결과 사용하지 않은 경우와 비교해 예측의 정확도가 최대 두 배 더 높아진 것을 확인할 수 있었다.

주요용어: 웹 로그, 고객 행동 예측, 기계학습, 데이터 마이닝

[1]교신저자: (03760) 서울시 서대문구 대현동 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr