# 관광객 데이터를 이용한 한국 여행 추천 시스템


## Team Data Travel

### 오세인 손동희 양지연 김수진 김종혁


## 1. 서론

대한민국을 찾는 외래관광객은 매년 큰 폭으로 증가하고 있다(2018년 기준 전년 대비 28% 증가). 또한 방문한 외국인 관광객이 기존과는 다른 모습으로, 단순한 관람을 넘어 관심분야가 확대되고 있는 것이 사실이다. 정부와 지자체 차원에서는 계속해서 관광 지역별로 외국인 관광객들을 유인할 수 있는 그 지역만의 독특한 관광 상품을 개발하고 있으며, 최근 급격하게 여행 서비스 관련 소규모 창업부터 다양한 여행 플랫폼들이 증가하고 있는 추세이다.

이에 본 팀은, 관광산업은 양적 성장에서 질적 개선으로의 변화가 절실한 시점이라 판단을 하였다. 기존의 외국인 관광객들을 위한 여행 플랫폼은 단순히 대표적인 관광지를 이용자의 평점별로 추천하는 시스템이다. 그러나 이러한 플랫폼은 다양한 여행객들의 성향과 유형을 고려하지 않았기 때문에, 개인별 만족도의 편차가 매우 심할 것이라 생각하였다. 따라서 단순히 대표적인 여행지 추천에서 더 나아가 외국인 여행객들이 최고의 만족을 얻을 수 있도록 여러 요인을 모두 고려할 수 있는 방향으로 주제로 삼고 연구를 진행하였다.

본 팀의 분석 방향은 크게 세가지 측면으로 접근하였다. 첫째, 전체 데이터에 하나의 모형을 적합 시키는 방법 대신 군집분석을 통해 데이터를 네 개의 군집으로 나누고, 각 군집별로 모형을 적합 시켜 적합도와 예측력을 높이고자 하였다. 둘째, 예측을 위해 스터디에서 학습하였던 다양한 분석방법들을 유기적으로 연결하여 단계적으로 데이터에 적용하고자 하였다. 셋째, 적용한 통계분석을 통해 실제 외국인 관광객들이 활용할 수 있도록 어플리케이션 서비스로의 확장을 염두해두고 진행하였다. 이와 같은 방법으로 하나의 데이터를 다각적으로 분석할 수 있었다.

## 2. 연구 대상

### 2.1. 데이터 설명

데이터로는 2017년 외래관광객 실태조사를 사용했다. 해당 조사는 문화체육관광부와 한국관광공사가 주관하는 조사로 방한 후 출국하는 외래 관광객을 대상으로 하고 있다. 한국관광공사에 따르면 "본 조사는 우리나라를 방문한 외래관광객의 한국 여행실태, 한국내 소비실태 및 한국 여행 평가를 조사하여 외래관광객의 한국 여행성향

을 파악하고 연도별 변화추이를 비교, 분석"하고 있다고 나와있다.[1] 이는 8개의 공항에서 1년간 동일한 장소에서 수집된 자료이며 11 종의 언어로 되어있는 설문지를 활용하여 수집되었다. 또한 실태조사가 관광정책 수립의 기초가 되기 위해서인 만큼 본 연구에 적합하다고 생각되어서 사용하게 되었다.

관광객 실태조사는 여행사에서 자체적으로 실시하기도 하는데 표본조사인 만큼 조사과정에서의 오류로 인해 결과에 편향이 있을 수 있다는 점을 고려하여 해당 연구에서 얼마나 전문성을 확보하고 있는지 확인해보았다. 한국관광공사에 유선상으로 조사과정을 문의해보았으며, 설문조사 전담 팀이 별개로 존재하며 패키지 여행객만을 대상으로 한 것이 아닌 표본조사의 유의점을 고려한 실태조사가 이루어지고 있다는 확답을 얻은 후에 해당 연구를 진행하게 되었다.

해당 데이터에는 219개의 변수가 있었으나 본 연구에서는 전처리를 통해 분석에 용이한 방향으로 변형하여 사용했다. 변수 중 여행객에 대한 정보로는 국적, 성별, 동행자유형, 방문목적이나 동기, 등이 있다. 한국 여행실태에 대한 정보로는 체류기간, 방문지 목록, 활동 내역, 여행형태, 좋았던 방문지, 좋았던 활동 내역 등이 있다. 한국내 소비실태는 구체적으로 총 지출, 유형별 지출금액, 일일지출, 쇼핑품목, 등으로 이루어져 있다. 한국 여행 평가에 관련하여 만족도, 재방문의사, 한국여행 추천의사, 방문전/후 이미지가 있었으며 이 변수들을 추후에 반응변수로 활용하였다.

### 2.2. 분석 목표 및 흐름

추가적인 데이터의 보충 없이 실태조사 데이터에 있는 많은 수의 관찰 값과 변수들을 이용한 다방면의 분석으로 유기적으로 연결된 하나의 관광 추천 서비스를 완성시키고자 하는 것이 본 팀의 주요 분석 목표라고 말할 수 있다. 이를 위해 본 연구를 진행하는 과정에서 데이터에 주어진 변수를 최대한 버리지 않고 각각에 대한 의미를 부여하고 해석해내는데 큰 노력을 들였다.

분석의 흐름을 정하기 위해 본 팀원들은 관광객의 입장에서 어떠한 서비스가 제공된다면 좋을지에 대해 고려하는 동시에 개별적인 특성을 가지고 있는 관광객들을 어떻게 구별하여 서비스를 차별적으로 제공할 수 있을지에 대해 생각했다. 따라서 크게 관광객들의 유형별 고려와 관광지 및 예산 추천을 목표로 분석을 진행하기로 결정하였다. 먼저 관광객의 여행 유형에 맞는 맞춤형 만족 서비스를 제공하기 위해 관광객의 특성을 잘 나눌 수 있는 변수들을 선정하여 그것에 따라 클러스터링을 이용해 유형별 군집을 나누었다. 그리고 클러스터링에 의해 나뉘어 있는 군집에 따라 유형별로 관광객의 만족도에 어떤 변수가 어느 정도의 영향을 미치는지를 분석하기 위해 로지스틱 모델을 사용하였다. 그 후 유형별로 선호하는 관광지를 추천하기 위해 네트워크 분석을 이용해 시각적으로 이를 파악하였으며, 마지막으로는 전체 관광객을 상대로 여행 계획을 짜는데 얼마 정도의 예산을 책정해야 하는지 가이드라인을 제시해주기 위해 랜덤 포레스트, XG Boost 와 같은 머신 러닝 기법을 사용하여 개인별 적정 예산을 예측해주는 모델을 구축하였다.

[1] 한국 관광 공사 (https://kto.visitkorea.or.kr/kor/notice/data/statis/tstatus/forstatus/board/view.kto?id=429831)

## 3. 연구 과정

### 3.1 데이터 전처리 및 EDA

2017 년 외래 관광객 실태조사 데이터에는 값을 알 수 없어 NA 처리된 데이터는 존재하지 않았다. 다만 0 이거나 특정 값에 해당하지 않아 NA 처리가 된 값들이 있어 각각 0 으로 대체하고 더미 변수화 해주었다. 예를 들어, '방문횟수'의 경우 처음 한국을 방문한 관광객은 이전까지 방문횟수가 0 이었기 때문에 NA 처리 되어있었다. 이를 분석에 용이하게 0 으로 대체 해주었다. 또한, '숙박종류'와 '방문지역'의 경우 기존 데이터에는 해당되는 숙박과 방문지역에 각각 숙박 코드와 지역 코드로 기입되어 있어서 이를 해당하면 1 아니라면 0 으로 대체하는 더미 변수화를 진행하였다. 추가적으로 방문기간 내에 한국의 계절을 나타내는 더미변수(spring~winter), 방문한 관광지 수(Site_total), 방문한 지역 수(Area_total) 그리고 더미변수인 동행자 유형을 하나의 열로 보여주는 동행자 유형통합(Companion) 변수를 분석에 도움이 될 것이라 판단하여 생성해주었다.

간단한 데이터 전처리를 마치고 본격적으로 분석에 앞서 데이터의 특성을 파악하기 위해 시각화를 통해 데이터를 탐색해 보았다.

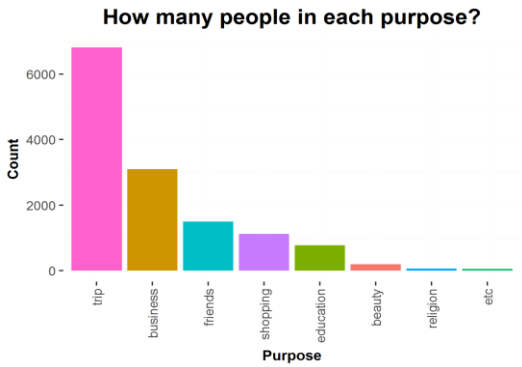



Figure 1. 방문 목적

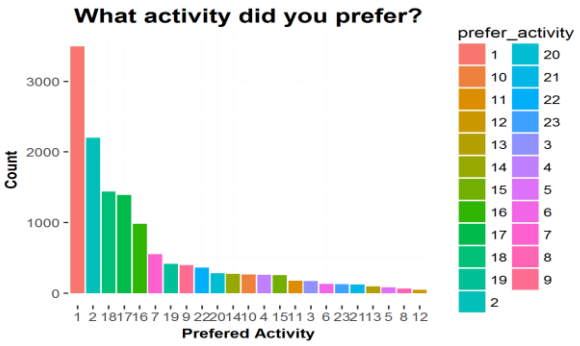



Figure 2. 선호 활동

먼저 Figure 1 은 외래 관광객들의 대한민국 방문목적 순위를 알려주는 히스토그램이다. 여행객들이 대한민국을 방문하는 주요 목적은 여가, 사업, 지인방문, 쇼핑 순으로 많았다. 또한 Figure 2 는 외래 관광객들이 대한민국에서 했던 활동 중 선호하는 활동의 순위를 나타낸다. 1 순위로 선호한 활동은 쇼핑이었으며, 식도락관광, 자경관 감상, 역사유적지 방문 등이 그 뒤를 따른 것을 알 수 있었다.

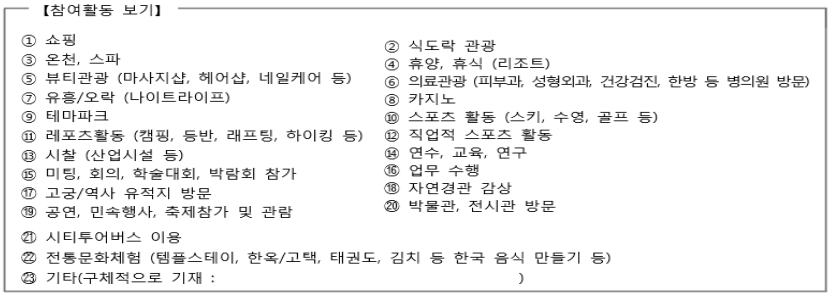



Figure 3. 선호 활동 범례

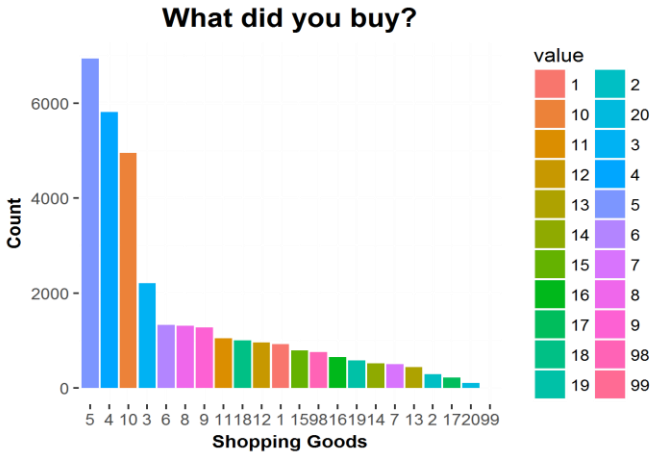
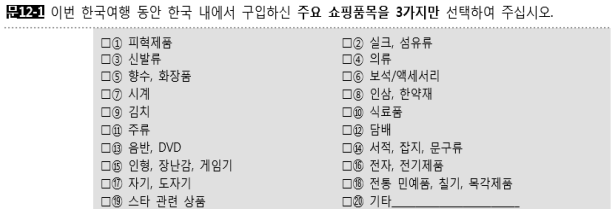



Figure 4. 쇼핑품목

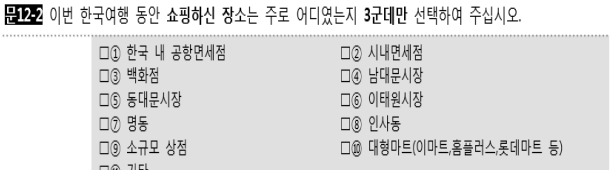

Figure 5. 쇼핑장소

다음은 쇼핑에 관련된 변수를 탐색해보았다. Figure 4 에서는 외래관광객이 구매한 쇼핑품목 순위를 나타낸다. 1 순위는 화장품이었고 그 뒤를 의류, 식료품, 신발류가 뒤따른 것을 볼 수 있다. 이런 품목들을 어느 장소에서 구매를 많이 할 지에 대해 탐색을 해 본 결과, Figure 5 에서 나타나듯이, 명동이 가장 많이 쇼핑을 한 장소임을 알 수 있고 공항 내 면세점, 대형 마트, 소규모 상점, 백화점 순으로 그 뒤를 따랐다. 마지막으로 동행자 수, 총 지출액 변수 탐색결과 이상치(outlier)라 판단되는 수치들이 있었기에 각각 43 명, 5000 달러 이상인 관측치를 삭제하였다.


문12-1 이번 한국여행 동안 한국 내에서 구입하신 **주요 쇼핑품목을 3가지만** 선택하여 주십시오.

| | |
|---|---|
| □① 피혁제품 | □② 실크, 섬유류 |
| □③ 신발류 | □④ 의류 |
| □⑤ 향수, 화장품 | □⑥ 보석/액세서리 |
| □⑦ 시계 | □⑧ 인삼, 한약재 |
| □⑨ 김치 | □⑩ 식료품 |
| □⑪ 주류 | □⑫ 담배 |
| □⑬ 음반, DVD | □⑭ 서적, 잡지, 문구류 |
| □⑮ 인형, 장난감, 게임기 | □⑯ 전자, 전기제품 |
| □⑰ 자기, 도자기 | □⑱ 전통 민예품, 칠기, 목각제품 |
| □⑲ 스타 관련 상품 | □⑳ 기타_____ |

Figure 6. 쇼핑품목 범례


문12-2 이번 한국여행 동안 **쇼핑하신 장소**는 주로 어디였는지 **3군데만** 선택하여 주십시오.

| | |
|---|---|
| □① 한국 내 공항면세점 | □② 시내면세점 |
| □③ 백화점 | □④ 남대문시장 |
| □⑤ 동대문시장 | □⑥ 이태원시장 |
| □⑦ 명동 | □⑧ 인사동 |
| □⑨ 소규모 상점 | □⑩ 대형마트(이마트,홈플러스,롯데마트 등) |

Figure 7. 쇼핑장소 범례

### 3.2. 클러스터링

관광객들은 그 나라를 방문하고자 하는 동기, 목적 등 다양한 이유에 따라 방문하는 관광지나 여행 지출에 차이를 보인다. 쇼핑을 목적으로 하는 관광객이 다른 목적을 가진 타 관광객에 비해 비교적 많은 지출을 하거나 역사 관광을 목적으로 하는 관광객들이 유적지에 좀 더 잦은 방문을 하는 등의 예시가 있다. 따라서 본 팀은 마찬가지로 한국을 방문하는 관광객들의 유형이 특정한 변수들에 의해 어떠한 특성을 가지고 구분되어질 수 있다는 가정 하에 군집을 나누기 위해 클러스터링 기법을 진행하였다.

본 팀은 유형별 군집을 나누기 위해 각 관찰지를 기준으로 거리(유사도)를 구하여 군집화하는 PAM(Partitioning Around Medoids) 기법을 사용하였다. 거리(유사도)를 구하는 방법은 범주형과 연속형 변수를 포함하고 있는 데이터 형태를 고려하여 고어 유사도(Gower's distance)를 사용하였으며 군집을 평가하는 기준으로는 1의 값에 가까울수록 군집화 품질을 높다고 평가하는 실루엣(Silhouette) 값을 이용하였다.

변수 선택 기준은 여행자의 유형을 파악이라는 분석 목표를 위해 여행자의 특성을 나타내는 변수들을 우선적으로 선정하였다. 결과적으로 나이, 국적, 목적, 동기, 머문 기간, 동행자 유형 등 이 외 5 개의 변수들이 선정되었다. 그 후 선정된 변수들 중 네다섯 개의 변수들을 랜덤으로 섞어가며 클러스터링을 진행하며 어떤

조합의 변수들이 군집화의 질을 크게 높일 수 있는지를 탐색하였다. 이 과정에서 나이, 국적, 직업과 같은 인구통계학적 변수들이 오히려 군집화의 질을 떨어뜨린 다는 사실을 발견할 수 있었다.

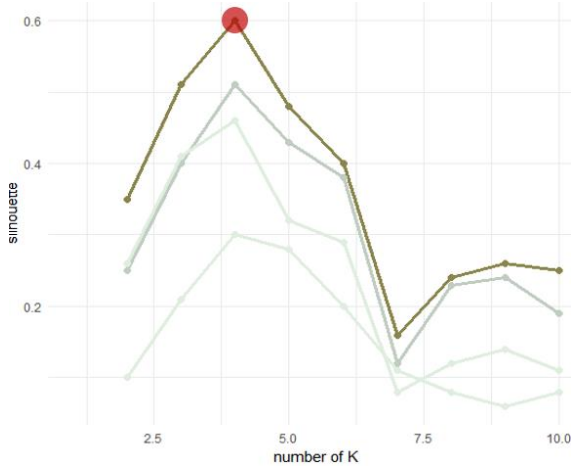



**Figure 1. 변수 조합별 평균 실루엣 값 비교**

평균 실루엣 값을 비교하여 결과적으로 가장 군집화의 질을 높게 한 변수의 조합은 동행자 유형(companion), 관광지 방문횟수(site_total), 총 지출(total_pay), 동기(motivation)로 그 중에서도 군집을 4 개로 설정하였을 때 실루엣 값이 0.6 으로 가장 높았다. 최적의 실루엣 값으로 나뉘어진 4 개의 군집은 동행자 유형(companion) 변수에 매우 큰 영향을 받는 것을 알 수 있었다. 각 군집의 특성에 대해 좀 더 자세한 해석을 해보았다.

```
> summary(group1)
 motivation1     site_total        total_pay          companion
 1      :724   Min.   : 1.000   Min.   :    0.101   1:   0
 2      :686   1st Qu.: 4.000   1st Qu.: 659.243   2:3714
 3      :350   Median : 6.000   Median :1062.000   3:   0
 4      :304   Mean   : 6.783   Mean   :1309.894   4:   0
 7      :208   3rd Qu.: 9.000   3rd Qu.:1680.636   5:   5
 (Other):820   Max.   :29.000   Max.   :5000.000
 NA's   :627
> summary(group3)
 motivation1     site_total        total_pay          companion
 2      : 333   Min.   : 1.000   Min.   :    0.454   1:4549
 1      : 281   1st Qu.: 2.000   1st Qu.: 500.000   2:   0
 3      : 205   Median : 4.000   Median : 900.000   3:   0
 13     : 147   Mean   : 5.297   Mean   :1242.493   4:   0
 4      : 115   3rd Qu.: 7.000   3rd Qu.:1660.000   5:   4
 (Other): 462   Max.   :40.000   Max.   :5000.000
 NA's   :3010
```

```
> summary(group2)
 motivation1     site_total        total_pay          companion
 2      :796   Min.   : 1.000   Min.   :    1.867   1:   0
 1      :607   1st Qu.: 4.000   1st Qu.: 619.245   2:   0
 4      :337   Median : 6.000   Median : 964.555   3:3237
 3      :210   Mean   : 6.939   Mean   :1218.917   4:   0
 7      :208   3rd Qu.: 9.000   3rd Qu.:1504.650   5:   4
 (Other):675   Max.   :33.000   Max.   :4999.500
 NA's   :408
> summary(group4)
 motivation1     site_total        total_pay          companion
 1      : 74   Min.   : 1.000   Min.   :    0.338   1:   0
 2      : 62   1st Qu.: 2.000   1st Qu.: 525.092   2:   0
 3      : 29   Median : 3.000   Median : 894.895   3:   0
 4      : 17   Mean   : 4.049   Mean   :1076.505   4:1229
 7      : 17   3rd Qu.: 6.000   3rd Qu.:1390.065   5:   5
 (Other): 69   Max.   :22.000   Max.   :5000.000
 NA's   :966
```

**Figure 2. 군집별 summary 결과**

첫번째 군집은 가족과 동행한 여행객 유형으로 총 지출은 평균이 1309.894 로 높은 편이었고 여행 동기는 주로 자연풍경과 쇼핑이라는 특징을 가지고 있었다. 두번째 군집은 연인 혹은 친구와 동행한 여행객 유형으로 관광지 방문 횟수는 평균 6.939 로 높은 편이었으며 여행 동기는 주로 쇼핑과 자연환경이라는 특징을 가지고 있었다. 세번째 군집은 혼자서 방문한 여행객 유형으로 관광지 방문 횟수나 총 지출에 뚜렷한 특징은 없었지만 여행 동기가 자연환경, 쇼핑, 역사, k-pop 등으로 비교적 고르고 다양하다는 특징을 가지고 있었다. 마지막 군집은 직장동료와 동행한 여행객 유형으로 관광지 방문 횟수와 총지출의 평균이 적은 편이라는 특징을 가지고 있었다. 직장동료와 동행한 여행객은 한국을 여행이 아닌 사업을 목적으로 방문한 것으로 예상할 수 있다.

마지막으로 우리는 연속형 변수인 관광지 방문횟수(site_total)와 총 지출(total_pay)이 sample 들로 판단한 평균의 값이 통계적으로 유의미한 차이를 보이는지를 확인하기 위해 ANOVA TEST 를 했다. 그 결과 각 p-value 값이 모두 0.05보다 작아 두 변수에 대해 모두 그룹 별 차이가 있음을 확인할 수 있었다.

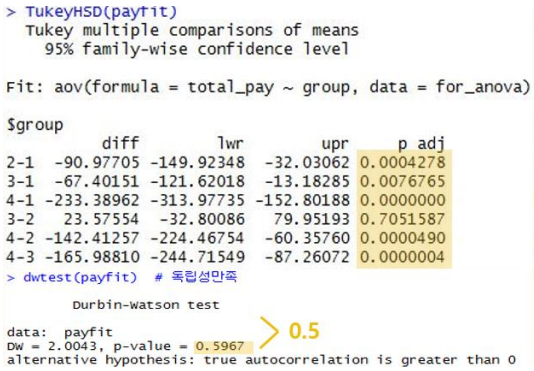



**Figure 6. 총지출 Tukey HSD 결과**

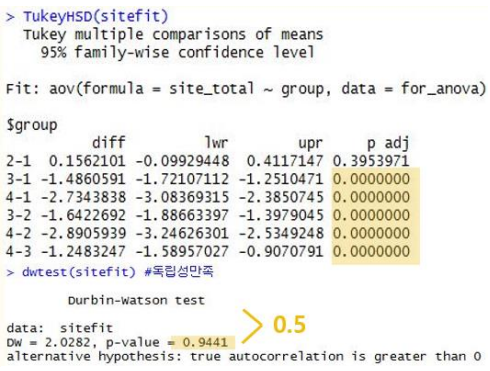



**Figure 7. 관광지 방문 횟수 Tukey HSD 결과**

더 나아가 Tukey HSD Test로 구체적인 그룹별 차이에 대해서 알아보았다. 총 지출(total_pay) 변수의 경우 친구, 연인과 동행한 두번째 그룹과 혼자서 방문한 세번째 그룹과의 차이를 제외한 모든 그룹의 평균 총 지출이 통계적으로 유의했다. 관광지 방문횟수(site_total) 변수의 경우 가족과 동행한 첫번째 그룹과 친구, 연인과 동행한 두 번째 그룹과의 차이를 제외한 모든 그룹의 평균 총 방문 관광지 수가 통계적으로 유의했다. 두 변수 모두 직장동료와 동행한 네 번째 그룹과 타 그룹과의 차이가 매우 큰 것을 확인할 수 있었다. Tukey HSD Test 결과를 통해 본 팀은 군집화로 나뉜 군집들의 차이가 통계적으로 유의미함을 확인할 수 있었다.

### 3.3. 다항 로지스틱 회귀분석

다항 로지스틱 회귀(Multinomial Logistic Regression))는 설명(독립)변수의 선형 결합과 다항 범주형 반(종속)변수를 가지는 모형으로 식은 다음과 같이 나타낼 수 있다.

$$g(y) = \beta 0 + \beta 1X1 + \cdots + \beta pXp$$

g(y)는 반응변수 Y의 함수로, 랜덤성분과 체계적 성분의 범위를 일치시키는 연결함수이다. 이 때, Y는 반응변수고, 랜덤 성분이라고 부른다. 우측 식, $\beta 0 + \beta 1X1 + \cdots + \beta pXp$ 은 설명변수 X와 계수 β 로 이루어진 부분으로 체계적 성분이라고 부른다. 다항 로지스틱 회귀의 장점은 선형성을 유지함으로써 해석이 간편하다는 특징을 가지기 때문에 대부분의 해석모형에서 사용된다.

본 팀의 반응변수 Y는 만족도 지표로서 세가지 범주로 표현하였고(1(불만족), 2(만족), 3(매우 만족)), 설명변수 X는 만족도의 요인이 될 수 있을 것이라 판단되는 변수들과 이후에 제언이 가능한 변수들로 택하였다. 이에 각 그룹별로 만족도에 미치는 요인이 다를 것이라 판단하여, 직관적인 해석이 용이한 다항 로지스틱 모형을 선택하였다. Y(만족도 지수)가 순서를 가지는 순서형 반응변수이기 때문에 Cut point를 설정하고 모델링을 진행했어야 하나 비례오즈 가정을 만족하지 못하였기 때문에 명목형 다항 로지스틱 모형에 적합 시켜 보았다.

| Y 변수 | X 변수 |
|---|---|
| **Satisfaction_index**<br>**(만족도 지수)**<br>- factor 변수 (1~3)<br>1.불만족 / 2.만족 / 3.매우 만족 | 1. **Num_companion(동행자 수)**<br>- numeric 변수 |
| | 2. **Type(여행 종류)**<br>- factor 변수 (1~3)<br>1.개별여행/ 2.단체여행/ 3.Air-tel Tour |
| | 3. **Period(체류 기간)**<br>- numeric 변수 |
| | 4. **Site_total(총 방문지 횟수)**<br>- numeric 변수 |

**Figure 1. 변수 선택**

해석하기에 앞서, 적합 시킨 회귀식이 과연 잘 적합 되었는가 검증하기 위해 LR test(Likelihood ratio test)를 시행해 보았고, 다중공선성을 확인하기 위해 VIF 검정을 진행하였다. LR test와 다중공선성을 모두 확인한 결과, 적합한 모델이 합리적이라고 판단을 내렸다.

```
Model 1: satisfaction_index ~ type + period + num_companion + site_total
Model 2: satisfaction_index ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  12 -3834.9
2   2 -3871.0 -10 72.219   1.65e-11 ***
```

**Figure 2. LR Test**

```
> vif(mulfit1)
                  GVIF Df GVIF^(1/(2*Df))
type          5.050539  2        1.499113
period        3.878400  1        1.969365
num_companion 5.833201  1        2.415202
site_total    8.497241  1        2.915003
```

**Figure 3. 다중공선성 확인**

```
> p1 < 0.05
  (Intercept) type2 type3 period num_companion site_total
2        TRUE FALSE FALSE  FALSE         FALSE      FALSE
3        TRUE  TRUE FALSE   TRUE         FALSE      FALSE
> exp(coef(mulfit1))
  (Intercept)     type2     type3   period num_companion site_total
2    2.788391 1.2245145 0.8315014 1.001585     0.9781940  0.9790109
3    2.096772 0.6726804 1.0991260 1.010803     0.9941215  1.0004055
```

**Figure 4. 다항 로지스틱 모형 - Group1 Summary**

상단 그림은 Group1의 만족도 지수에 대해 다항 로지스틱 모형을 적합 시킨 결과를 Summary 하여 발췌한 것으로 유의미한 변수는 여행 타입(type2 - 패키지)과 여행 기간으로 나타났다. 가족과 동행한 Group1의 경우 type2일 때 만족도3에 속할 오즈가 0.67배 높다고 얘기할 수 있으며, 패키지 여행일수록 높은 만족감을 느낄 확률이 더 낮다고 해석할 수 있다. 또한 여행기간이 한 단위 증가할 때 만족도2에 속할 오즈가 1.01배 높다고 얘기할 수 있기 때문에, 여행 기간이 길수록 높은 만족도를 느낄 확률이 높다고 해석하였다.

```
> p2 < 0.05
  (Intercept) type2 type3 period num_companion site_total
2        TRUE  TRUE FALSE   TRUE         TRUE      FALSE
3        TRUE FALSE FALSE  FALSE         TRUE      FALSE
> exp(coef(mulfit2))
  (Intercept)     type2     type3    period num_companion site_total
2    3.829558 2.207014 0.7572189 0.9933026     0.9517800  0.9851018
3    2.297141 1.468774 1.0655469 0.9974962     0.9447578  1.0195724
```

```
> p3 < 0.05
  (Intercept) type2 type3 period site_total
2        TRUE FALSE FALSE  FALSE       TRUE
3        TRUE  TRUE FALSE   TRUE      FALSE
> exp(coef(mulfit3))
  (Intercept)     type2     type3    period site_total
2    3.258815 0.7625320 0.7195200 0.9993829  0.9710277
3    3.321382 0.5056912 0.6143102 0.9964076  1.0116356
```

**Figure 5. 다항 로지스틱 모형 – Group2/3 Summary**

```
> p4 < 0.05
  (Intercept) type2 type3 period num_companion site_total
2        TRUE FALSE FALSE  FALSE         FALSE      FALSE
3        TRUE FALSE FALSE  FALSE         FALSE      FALSE
> exp(coef(mulfit4))
  (Intercept)    type2     type3   period num_companion site_total
2    2.331466 1.520032 0.8585039 1.006121     0.9822946  0.9839212
3    1.697750 0.575424 1.1118682 1.004347     0.9958184  1.0150354
```

**Figure 7. 다항 로지스틱 모형 – Group4 Summary**

Group1과 마찬가지로 각 그룹별로 만족도에 대해 다항 로지스틱 회귀 모형을 적합 시켜 보았고, 해석한결과는 다음과 같다.

- 연인, 친구와 여행 온 Group2의 경우 패키지 여행일수록 높은 만족감을 느낄 확률이 높고, 여행 기간은 길어질수록, 구성원이 많아질수록 불만족한 여행이 될 확률이 높다.
- Group3의 경우 혼자서 방문한 여행객 유형으로, 패키지 여행일수록 높은 만족감을 느끼기 어렵고, 여행 기간이 길어지고 관광지 수가 늘어날수록 불만족할 확률이 높을 것이다.
- 직장동료와 동행한 Group4의 경우, 유의한 변수들이 없었는데 이는 출장 차 방문했기 때문에 여행유형이나, 동행자의 수, 관광지 수 등의 변수가 만족도에 영향을 끼치지 않았을 것이다.

위 결과를 통해, 그룹별로 만족도에 영향을 끼치는 요인들이 모두 다르게 나타남을 알 수 있었다.

### 3.4. 네트워크 분석

#### 3.4.1 네트워크 분석

사회적 네트워크 분석이라고도 불리는 네트워크 분석은 네트워크를 구성하는 액터(actor)간의 관계와 구조를 파악하는 것이다. 액터들 간의 관계는 방향성이 있거나(directed) 방향성이 없는 것(undirected)으로 나누어 지며, 관계들에 가중치가 있는 경우와 없는 경우로도 나누어질 수 있다. 네트워크의 구조를 파악함으로써 액터들의 관계를 설명하거나 새로운 액터가 기존의 액터들과 어떤 관계를 맺을 지 예측하는 것이 네트워크 분석의 목적이 된다. 또한 이러한 분석을 통해 특정 액터가 네트워크에서 어떤 위치에 있는 지, 혹은 어떠한 중심도를 가지고 있는 지도 파악할 수 있다.

클러스터링 기법을 통해 구한 여행객 유형을 나누어 보았는데, 이러한 유형별로 선호 관광지를 조사하기 위해 네트워크 분석을 활용하였다. 이 분석 기법을 도입하게 된 계기는 특정 관광객이 선호했던 관광지들 간 관련성이 있을 것이라는 가설 때문이었다. 이에 따라 관광객들이 선호했던 관광지들을 액터로 설정하였고, 방향성이 없는 링크(edge)로 각 관광객이 선호하는 관광지들을 연결했다. 이때 특정 두 관광지를 선호하는 경우가 많은 경우, 관광객의 수 많은 가중치를 부여하였다. 이를 통해 네트워크를 시각적으로 구현하여 액터들 간의 구조를 보거나 중심성을 구하는 여러 방법 중 해당 데이터에 가장 적합했던 방법(Degree Centrality)을 통해 액터의 영향력을 파악해보았다.

네트워크 분석을 통해 특정 유형에 속하는 관광객이 하나의 관광지를 선호했을 때 다음 관광지를 추천해 줄 수 있게 되었다. 또한 선호되는 관광지들의 지리적인 분포를 파악하였으며, 어떠한 관광지들이 가장 빈번하게 선

호되는 지나 어떠한 두 관광지가 가장 높은 연관성을 가지고 있는 지도 시각적으로 파악할 수 있었다.

### 3.4.2 데이터 전처리

네트워크 분석을 진행하기 위해 선호 관광지1, 선호 관광지2, 선호 관광지3이라는 변수를 활용했다. 네트워크 분석에서의 링크(edge)는 두 액터 간의 관계를 지칭하는 것으로 세개의 선호 관광지를 선호 관광지1과 선호 관광지2, 선호 관광지2와 선호 관광지3, 선호 관광지1와 선호 관광지3의 세 조합으로 변형시켰다. 이 때 관광지의 조합이기 때문에 place1에 큰 숫자, place2에는 작은 숫자가 오도록 정돈하였다. 그 후 link_freq라는 변수를 만들어서 각 조합이 몇 차례 반복되었는지 저장하였다. 아래의 표에서 place1과 place2는 관광지 조합을 나타내며 link_freq는 횟수를 의미한다. 이때 관광지는 각 관광지를 나타내는 숫자로 보여지고 있다.

| place1 | place2 | link_freq |
|---|---|---|
| 80 | 79 | 24 |
| 132 | 131 | 15 |
| 128 | 127 | 26 |

**Figure 1. 관광지 조합과 조합별 개수**

### 3.4.3 Sociogram

네트워크 분석은 R 에서 'igraph' 패키지와 'visNetwork'패키지를 사용하였다. 우선 전체 데이터의 관계를 파악하기 위해 다양한 방식으로 시각화를 진행하였다. 유의미한 결과 해석을 위해 link_freq 가 15 이상인 경우만 표현하였으며, 노드를 연결하는 선의 두께는 link_freq 에 비례하도록 가중치를 부여하였다. 마지막으로 노드명을 관광지명으로 변경하였다. 이때 stringr 패키지를 통해 관광지명을 짧게 표기하였다. 관광지들의 중심도를 나타내는 그래프는 색의 진하기로 중심도(Degree Centrality)의 정도를 표현하였으며, 지리적 특성을 나타내는 그래프는 지역별로 다른 색을 부여하였다. 특히 레이아웃중에 "Force-directed"특성을 띄는 "Kamada-Kawai"레이아웃이나 "Fruchterman-Reingold"레이아웃을 사용하여 관계의 세기(force)에 따라 노드를 배치하지만 시각적으로 잘 나타내기 위해 노드들이 서로 포개어지는 것을 막는 배치방식을 택하였다.

### 3.4.4 그룹별 결과

그룹별 선호 관광지를 파악하기 위해 위의 두 그래프의 정보를 한번에 보여줄 수 있도록 중심도(Degree Centrality)는 노드의 크기, 지역은 노드의 색으로 표기하였다.

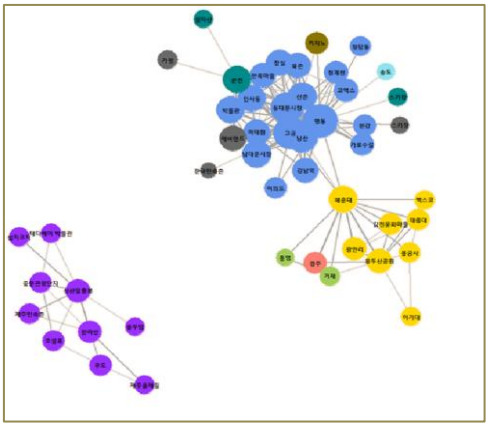


**Figure 2. 군집 1 선호 관광지**

첫번째 군집인 가족과 동행한 유형은 주로 아이들이 선호하는 관광지들의 중심도가 높았다. 예를 들어 다른 유형에서는 선호되지 않거나 선호도가 작았던 테디베어 박물관, 민속촌, 스키장이 높은 선호도를 보였다.

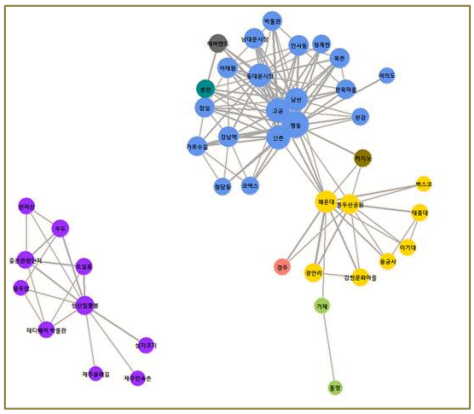


**Figure 3. 군집 2 선호 관광지**

두번째 군집인 연인이나 친구와 동행한 유형은 랜드마크를 방문하는 것을 선호했다. 이는 연인이나 친구가 동행할 경우 랜드마크를 주로 방문한다는 특성과 해당 유형의 경우 방문지보다는 동행자와의 추억을 더 기대하는 경우가 많다는 점 때문이라고 이해된다.

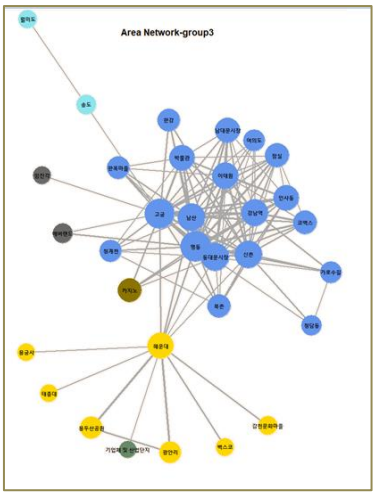



**Figure 4. 군집 3 선호 관광지**

세번째 군집인 혼자 여행 온 여행객 유형의 경우, 다른 유형들과는 다르게 제주도를 선호한 여행객이 많지 않았다. 이는 해당 유형의 여행 목적이 사업차, 교육, 친지방문의 경우가 많았기 때문에 이동시간이 긴 제주도까지 방문하지 않는 것으로 판단되었다. 또한 여행자들이 한국에 온 동기가 다양한 만큼 선호하는 방문지가 다양하며 노드들의 중심성이 낮아진 것을 파악할 수 있었다.

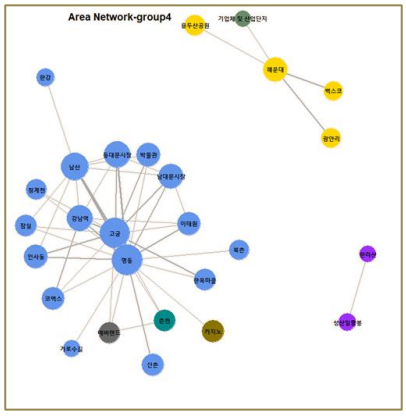



**Figure 5. 군집 4 선호 관광지**

네번째 군집인 직장동료와 동행한 유형은 지역간 이동이 적어서 노드들이 모두 연결되지 않고 지역별로 구분된 것을 볼 수 있었다. 또한 기업체나 산업단지를 선호하는 경우가 있었는데, 이는 방문지역이 출장 지역으로 한정되었기 때문이라고 생각된다.

### 3.5. 여행 경비 예측

여행을 계획하는데 있어 가장 중요한 요소는 여행 경비(예산)이다. 사전에 책정한 여행 경비에 따라 여행의 장소와 유형이 달라지기도 하며, 여행 경비에 의해 여행의 기간 또한 달라질 수 있다. 여행 경비는 쉽게 변화시킬 수 없는 부분이기도 하다. 본 연구는 최고의 만족을 위한 한국 여행 추천 시스템 구축을 목표로 하기 때문에 여행 경비를 예측하여 제공하고자 하였다.

앞서 진행한 분석들은 클러스터링을 통해 나누었던 군집별로 진행하였다. 그러나 여행 경비를 예측하는 데 있어서는 데이터를 하나로 통합하여 예측모델을 구축하였다. Y 변수로는 관광객들이 사용한 예산을 사용하였다. 이 예산을 사용하는 데 있어 쇼핑과 관련된 지출은 개인마다 특수성이 크고 예측이 어렵다고 판단하여 쇼핑을 제외한 지출을 Y 변수로 사용하였다. 또한, 노이즈를 감소시키고 예측력을 향상시키기 위해 Y 변수를 4 개의 범주로 재 분류하여 4 개의 범주를 예측하는 예측모델을 구축하였다.

| 최종 Y 변수(범주화 된 여행 경비) | 범위 | Obs 수 |
|---|---|---|
| 1 | 0$ ~ 350$ | 3517 |
| 2 | 350$ ~ 600$ | 3545 |
| 3 | 600$ ~ 950$ | 2812 |
| 4 | 950$ ~ | 2872 |

**Figure 1. 범주화 된 Y 변수**

각각 범주에 속하는 obs 가 균일하여 예측모델을 구축하는 데 있어 문제가 없을 것이라 판단하였다.

x 변수로는 개인이 고려하는 요인, 여행을 오기 전 예산을 책정할 때 별개로 영향을 미칠 수 있는 요인을 나타내는 변수들을 사용하였다. x 변수로 사용한 변수들은 다음과 같다.

- ✓ 여행객 유형 : nation, gender, num_companion, companion, decision_time
- ✓ 여행 목적/유형 : purpose, activity1, type,
- ✓ 숙박 유형 : accomodations_hotel, accomodations_guesthouse, accomodations_condo, accomodations_friend, accomodations_official, accomodations_temple, accomodations_homestay,
- ✓ 여행 시기 : period, spring, summer, fall, winter
- ✓ 여행 장소 : visit_place_1~48, area_total, site_total, visit_seoul

#### 3.5.1 Random Forest

여러 개의 decision tree 를 앙상블 하는 기법인 Random Forest 를 예산 예측에 사용하였다. Bagging 의 발전된 형태의 모델로 각 트리를 생성할 때 변수 또한 sampling 하여 여러 변수의 영향력을 고르게 반영할 수 있다는 장점이 있다.

모델을 학습시키고 평가하기 위하여 Train set 과 Hold out set 을 분할하였다. Hold out set 은 모델을 최종으로 결정하고 최종 정확도를 평가하는 데에만 사용하였다. 모델을 학습시키는데 있어서는 Train set 으로 5 - Fold cross validation 을 진행했다. 평가 metric 으로는 밸런스 데이터라 판단하여 Accuracy 를 사용하였다.

CV 를 통해 최적의 파라미터를 찾기 위해 'Caret' pakage 를 이용하여 Grid search 를 진행하였다. 조정 파라미터는 트리를 생성할 때 사용할 변수의 개수를 지정하는 mtry 하나로 10~40 까지 accuracy 를 비교하였다.

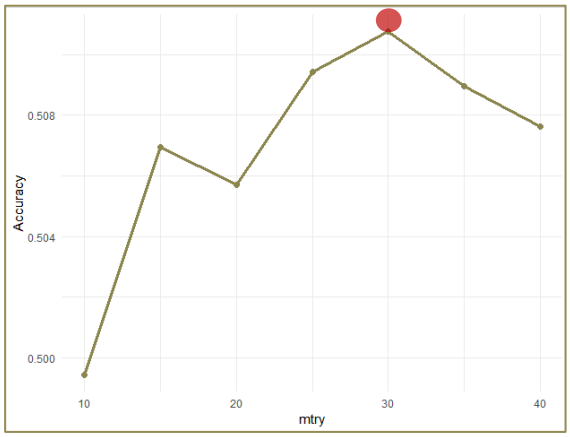
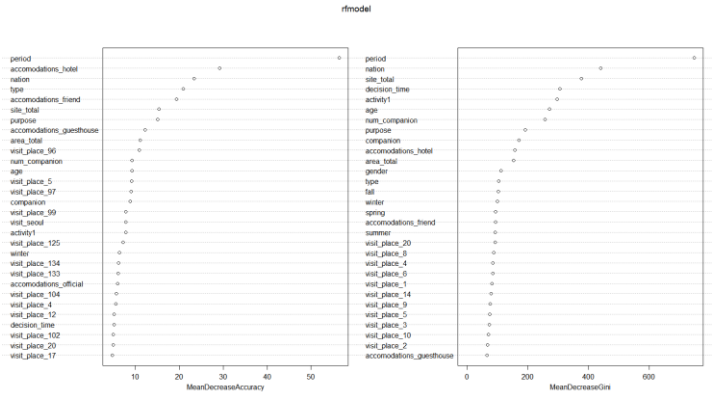



**Figure 2. mtry 별 Accuracy**

Grid search 결과 최적의 mtry는 30이라고 판단하여 mtry = 30, ntree = 100으로 모델을 최종 적합하였다. 최종 적합한 모델의 Importance plot은 다음과 같다




| Accuracy | Gini |
|---|---|
| Period | Period |
| Accomodation Hotel | Nation |
| Nation | Site_total |
| Type | Decision time |
| Accomodation Friend | Activity |

**Figure 3. importance plot & Top 5 variable**

### 3.5.2 XG boost

예측력을 더욱 높이기 위하여 Random Forest 와 유사한 tree based model 인 XG boost 를 추가로 사용하였다. XG Boost 는 gradient boosting 알고리즘을 기반으로 한다. gradient boosting 은 경사하강법을 사용한 Boosting 모델로서 트리 하나를 생성했을 때 발생하는 에러에 새로운 트리를 적합시키는 방식으로 에러를 줄여나간다.

XG Boost도 RF와 마찬가지로 hold out set을 제외라고 train set으로 5 - Fold cross validation을 진행하여 파라미터 튜닝을 실시하였다. 그러나 XG Boost는 RF와 달리 조정할 수 있는 파라미터가 많아 네 가지의 파라미터만(Max_depth,

Min_child_weight, Colsample_bytree, Subsample) 조정하여 모델을 적합시켰다.

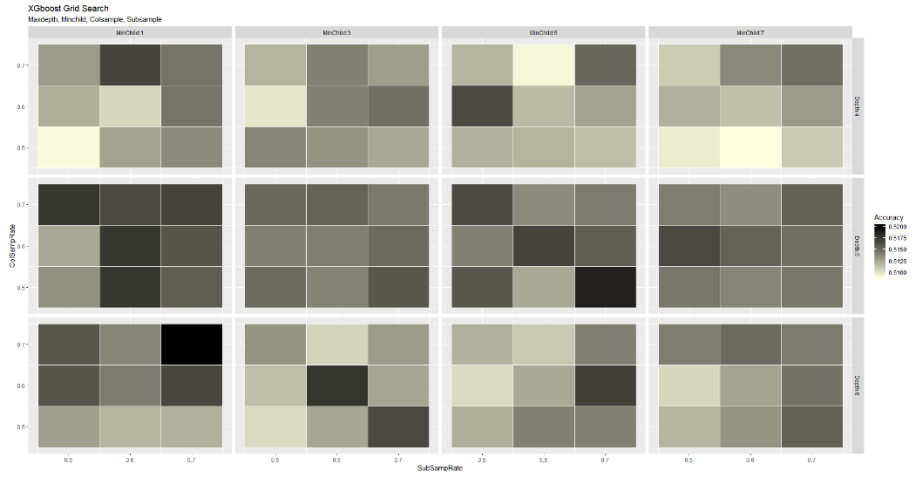



**Figure 4. Grid search 시각화**

Grid search 결과 Max_depth = 8, Min_child_weight = 1, Colsample_bytree = 0.7, Subsample = 0.7 일 때, Accuracy가 가장 높았으므로 이를 최적의 파라미터로 결정하였다. 최적의 파라미터로 early stopping을 진행하여 최적의 Nround를 121로 설정하였고 모든 결과를 종합하여 모델을 적합시켰다.

Random Forest와 XG Boost의 CV Accuracy를 비교한 결과 XG Boost가 더 좋은 결과를 보였으므로 XG Boost를 최종 모델로 선정하였다. 최종 모델을 이용하여 Hold out set의 Accuracy를 평가한 결과 53.45%의 정확도를 보였다. Y변수가 네 개의 범주였고 multiclassification임을 감안했을 때 그리 나쁘지 않은 결과라 판단하였다.

## 4. 결론

### 4.1. 활용방안

지금까지 서로 다른 특성을 가진 관광객 그룹을 구별하여 그에 따른 특징을 해석하였고 각 그룹에 맞게 관광지와 적정예산 범위를 제안하는 분석을 진행하였다. 이를 활용해 관광객의 입장과 각 그룹의 특성들을 모두 고려한 서비스를 제공하고자 하였고 그 방법으로 '트레블 메이커(Travel Maker)' 라는 어플리케이션을 고안하였다. 어플리케이션의 시스템은 다음과 같이 구축된다.

먼저 관광객이 원하는 동행 유형을 선택하면 관광객의 만족도를 위해 로지스틱 회귀 모형 결과를 활용해 여행에 관한 권장사항을 알려준다. 예를 들어 Figure 1 의 경우 친구 및 연인과 동행하고자 한다면, 패키지 여행보다는 개별여행을 추천하고 여행 기간은 길어질수록 만족도가 낮아질 수 있으므로 적절한 여행기간을 계획하는 것을 권장한다. 그리고 최대한 적은 수의 동행자와 함께 여행을 하는 것이 만족할 확률이 높으므로 적은 인원과의 여행을 추천하게 된다.

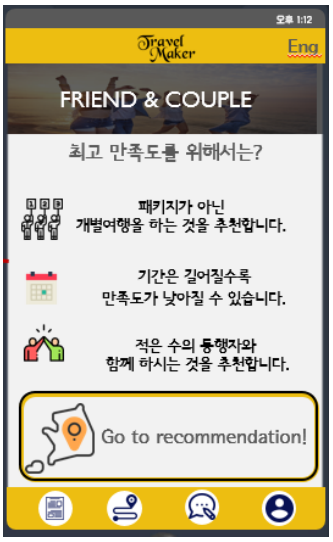



Figure 1. 만족도를 위한 조언(친구 및 연인)

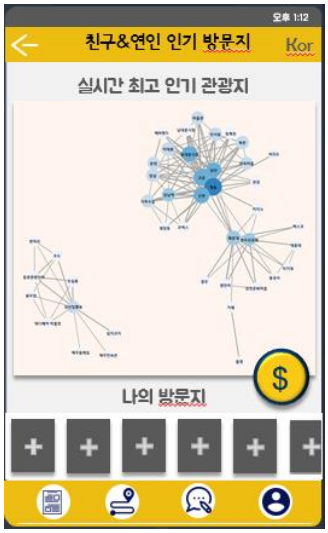



Figure 2. 인기관광지 추천

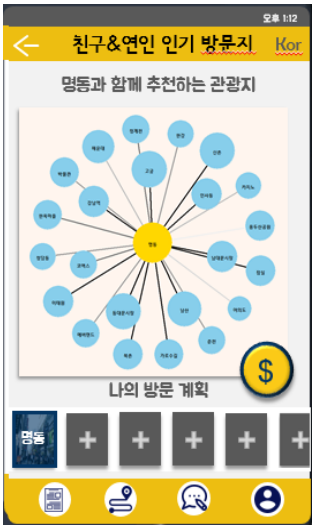



Figure 3. 특정 관광지와 연결된 관광지 추천

이후 관광객이 관광지 정보 버튼을 누르게 되면 Figure 2 처럼 네트워크 분석을 통해 도출된 동행 유형별 인기 관광지 정보가 네트워크 Plot 으로 보여진다. 이중에서 한 곳을 클릭해 나의 방문계획에 추가하면 Figure 3 과 같이 특정 관광지와 함께 추천할만한 관광지를 여러 개 제언해준다. 이를 통해 대한민국에 대해 무지한 관광객은 실시간으로 동행 유형에 따른 인기 관광지와 각 관광지와 엮인 잠재 관광지를 추천받게 되어 더욱 편리하게 여행을 계획할 수 있게 되고 이미 입국한 관광객들은 즉흥적으로 다른 관광지를 알아보고 방문할 수 있는 장점이 생기게 된다.


Figure 4. 추가 개인정보입력

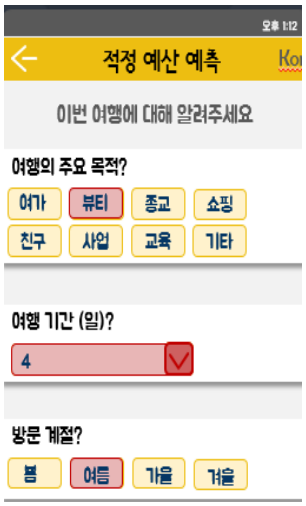



Figure 5. 추가 여행정보1


Figure 6. 추가 여행정보2

더불어 관광지 추천 인터페이스 상의 달러 버튼을 클릭하면 Figure 4, 5, 6 과 같이 추가적인 정보를 입력하는 화면이 나타난다. 이 화면에서 성별 / 국적 / 나이 의 약간의 개인정보와 여행목적 / 여행기간 / 숙소정보 등 추가적인 여행정보를 입력하게 되면 Figure 7 의 구현 예시와 같이 적정 예산 범위를 제시해주게 된다. 이러한 시스템을 통해 외래 관광객들은 대한민국 여행의 만족도를 높일 수 있는 조언을 받을 수 있게 되고 동행 유형별 인기•잠재 관광지 그리고 예산 범위까지 추천받을 수 있는 고객맞춤형 서비스를 제공받을 수 있게 될 것이다.

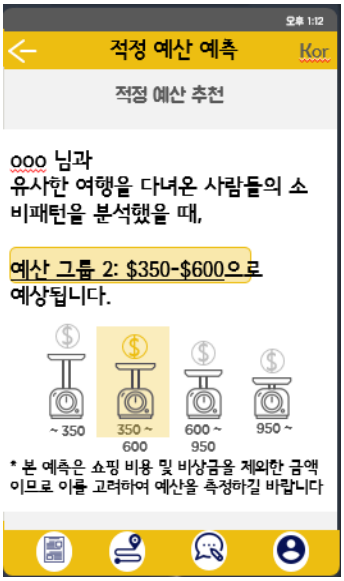



**Figure 7. 적정 예산 추천**

### 4.2. 한계 및 의의

본 연구는 외래 관광객 설문조사 데이터를 이용한 연구로 한국인들의 국내 여행 관련 데이터는 반영할 수 없었다. 만일, 한국 관광객의 데이터가 있다면 국내 여행 추천 시스템을 고안할 수 있을 것으로 기대한다. 또한, 본 연구는 2017년의 데이터만 사용한 연구로 이전의 데이터를 수집할 수 있었다면 더욱 세밀한 연구가 진행되었을 것이다. 또한 위도, 경도를 이용한 지리적 요인을 고려하지 못한 점이 본 연구의 한계점이라 할 수 있다.

그러나 본 연구의 추천 시스템을 이용하여 외국인 관광객을 타겟으로 한 정책 혹은 마케팅을 펼친다면 외국인 관광객 유치에 큰 보탬이 될 것으로 기대한다. 또한, 하나의 데이터를 다각적인 분석을 적용하여 많은 분석 모델을 경험하고 새로운 모델을 공부했다는 점에도 의의가 있다.

이 연구는 여행 추천 시스템에 아이디어를 제시하고 발전 방향을 제시했다는 점에도 의의가 있다. 현재 플랫폼에 존재하는 여행 추천 시스템들은 방문수가 가장 많거나 가장 인기가 많은 관광지 순으로 추천을 진행한다. 또한 만족도를 높이기 위한 방향 제시를 해주는 시스템은 거의 없다. 이 모든 과정을 하나로 통합하고 하나의 어플로 구현해냈다는 점에 큰 의의가 있다. 이 연구를 바탕으로 최고의 만족도를 끌어낼 수 있는 여행 추천 시스템들이 많이 연구되기를 소망한다.