앙상블 기법을 통한 여가활동만족도의 요인분석 및 예측: 20~30대를 중심으로*

김부현 (수원대학교 데이터과학부) • 김진흠 (수원대학교 데이터과학부, 지도교수)

1. Introduction

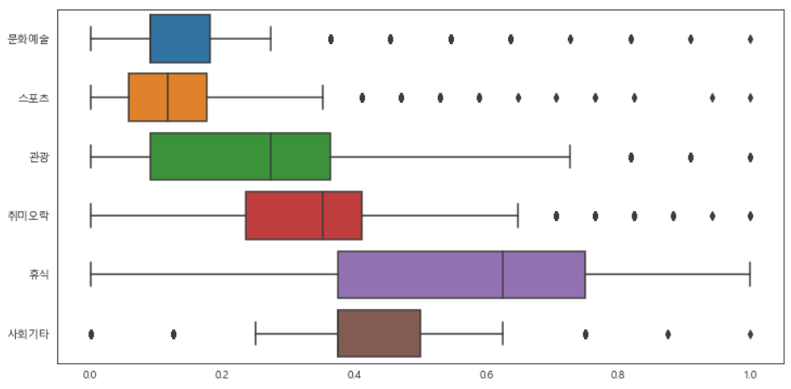
본 연구는 문화체육관광부에서 2020년도에 실시한 국민여가활동조사 자료를 이용하였으며 [1], 20~30대의 여가생활에 영향을 주는 요인을 찾고 여가생활만족도를 예측하고자 한다. 조사기간은 2019년 8월 1일 ~ 2020년 7월 31일, 조사대상자는 총 10,088명이며, 그 중 20~30대, 총 3,180명을 연구 대상자로 선정하였다.

연령	n(%)
15-19세	696(6.9)
20-29세	1,536(15.2)
30-39세	1,644(16.3)
40-49세	1,888(18.7)
50-59세	1,858(18.4)
60-69세	1,377(13.6)
70세 이상	1,166(11.5)
합계	10,088(100)

2. 자료변환

1. 여가활동유형

지난 1년간 한 번 이상 참여한 여가활동의 개수를 측정하였다. 각 영역별로 참여한 문화예술, 스포츠, 관광, 취미오락, 휴식, 사회기타활동으로 구분되며, Min-Max Scaler를 통해 정규화 처리하였다.



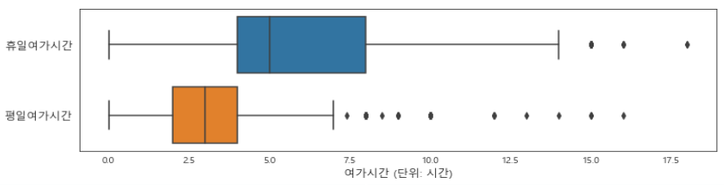
2. 여가비용

한 달 평균 얼마나 지출했는지 측정하였다. 분석에서는 다음과 같이 사분위수로 나누어 처리하였다.

비용	범주
6만원 이하	1
6-12만원	2
12-25만원	3
25만원 초과	4

3. 여가시간

하루 평균 여가시간과 7점 척도로 이루어진 충분도를 평일과 휴일로 구분하여 측정하였다. 분석에서는 국민여가활동조사 보고서를 참고하여 다음과 같이 나누어 처리하였다.



시간	범주
3시간 이하	1
3-5시간	2
5-7시간	3
7-9시간	4
9시간 초과	5

4. 공공여가시설충분도

생활권 내의 공공문화 여가시설 충분도에 대해 측정하였다. 공공문화 및 여가시설 이용 충분도와 공공문화 및 여가시설 프로그램 충분도 항목을 사용하였으며 해당 7점 척도의 평균값을 사용하였다.

5. 민간여가산업만족도

우리나라의 전반적인 여가산업 만족도에 대해 측정하였다. 여가 관련 공간산업, 용품산업, 서비스산업의 만족도 항목을 사용하였으며 해당 7점 척도의 평균값을 사용하였다.

6. 일과 여가의 균형

자신의 삶에서 일과 여가생활 간 균형이 잘 이루어지고 있는지 7점 척도로 측정하였다. 분석에서는 다음과 같이 3가지의 범주로 나누어 처리하였다.

7점 척도	내용	범주
1, 2, 3	일에 집중하는 집단	A
4	균형을 이루는 집단	B
5, 6, 7	여가에 집중하는 집단	C

7. 여가인식

여가인식은 7점 척도로 구성된 2가지의 설문으로 이루어져 있다. 여가활동이 삶의 필수적인 요건인지에 대해 측정한 문항과 여가활동이 삶에 긍정적인 영향을 끼치는지에 대해 측정한 문항이다.

8. 인구사회학적 요인

성별, 교육수준, 배우자유무를 포함하였다. 교육수준의 경우는 최종학력 고등학교 이하 또는 대학교 이상으로 구분하였다.

9. 주평균근무시간

주평균근무시간은 다음과 같이 나누어 처리하였다. 기준으로 삼은 주 40시간은 법정근무시간, 주 52시간은 연장근무시간이다.

시간	범주
0시간	0
0-40시간	1
40-52시간	2
52시간 이상	3

10. 변수정리

구분	변수	범주	범위	n(%) or M±SD
여가활동유형	문화예술		0-1	0.16±0.13
	스포츠			0.13±0.12
	관광			0.27±0.20
	취미오락			0.34±0.16
	휴식			0.56±0.21
	사회기타			0.40±0.14
여가비용		0-8만원		829 (26.1)
		8-15만원		766 (24.1)
		15-20만원		844 (26.5)
		20만원 이상		740 (23.2)
여가시간	평일여가시간	0-3시간		1926 (60.6)
		3-5시간		911 (28.6)
		5-7시간		225 (7.0)
		7-9시간		65 (2.0)
		9시간 이상		52 (1.6)
	휴일여가시간	0-3시간		712 (22.4)
		3-5시간		998 (31.4)
		5-7시간		575 (18.1)
		7-9시간		435 (13.7)
		9시간 이상		459 (14.4)
여가시간충분도	평일여가시간충분도		1-7	4.39±1.35
	휴일여가시간충분도		1-7	4.93±1.33
여가공간	공공여가시설충분도		1-7	4.18±1.19
	민간여가산업만족도		1-7	4.83±0.95
일과 여가의 균형			1-7	3.80±1.14
여가인식		삶의필수요건	1-7	5.70±0.92
		삶의 영향력	1-7	5.78±0.86
인구사회학적요인	성별	남		1663 (52.3)
		여		1516 (47.6)
	배우자유무	있다		1183 (37.2)
		없다		1996 (62.8)
	교육수준	고졸 이하		602 (18.9)
		대학 이상		2577 (81.0)
주평균근무시간		0시간		1212 (38.1)
		0-40시간		1060 (33.3)
		40-52시간		564 (17.7)
		52시간 이상		343 (10.8)

3. 자료분석 및 결과

1. 결과변수 - 여가생활만족도

자신의 여가생활에 대하여 전반적으로 만족하는지 7점 척도로 측정하였으며, 분석에서는 다음과 같이 이진형 범주로 나누어 처리하였다.

7점 척도	내용	범주
1, 2, 3, 4	불만족	0
5, 6, 7	만족	1

2. 초기 분석

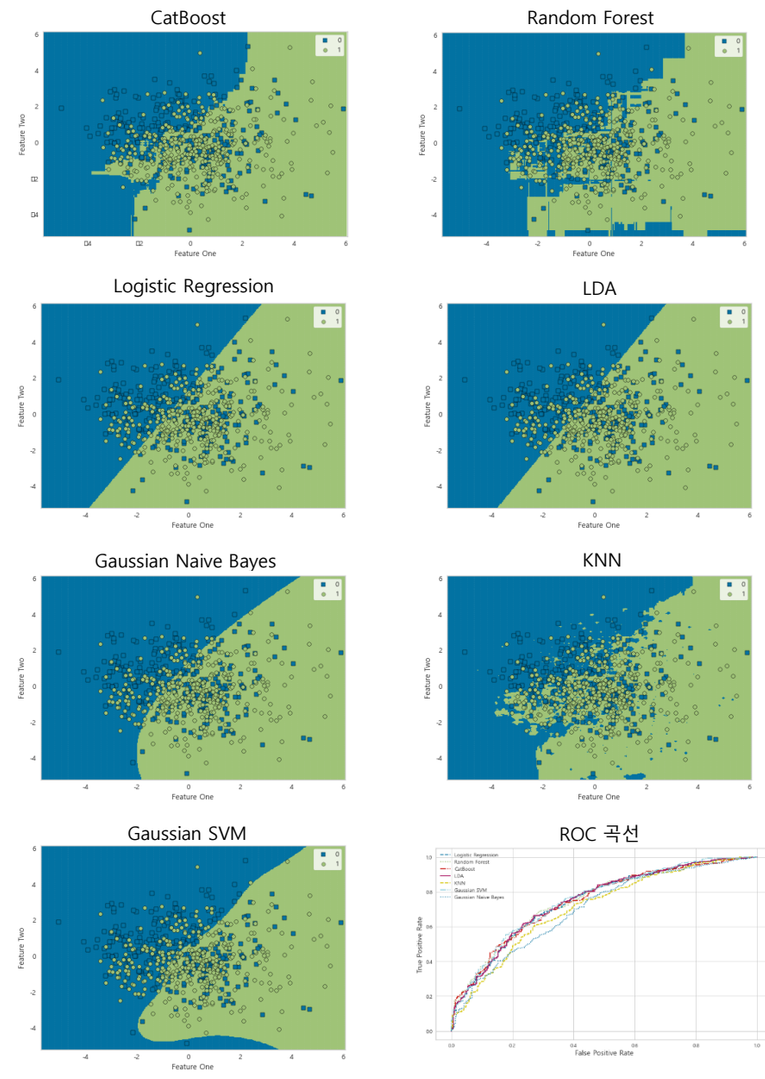
Method	Model	AUC	Accuracy	F1	MCC	Time
Boosting	CatBoost	0.789	0.727	0.769	0.440	16.9
	Gradient Boosting	0.783	0.718	0.761	0.421	1.9
	LightGBM	0.779	0.709	0.751	0.405	2.7
	XGBoost	0.760	0.702	0.744	0.388	5.6
	AdaBoost	0.771	0.719	0.762	0.423	3.1
Tree-based	Random Forest	0.780	0.721	0.763	0.427	3.8
	Extra Trees	0.780	0.719	0.762	0.423	3.7
	Decision Tree	0.629	0.633	0.672	0.257	0.6
Regression	Logistic Regression	0.768	0.711	0.757	0.406	1.1
Discriminant Analysis	LDA	0.768	0.713	0.759	0.409	0.5
	QDA	0.755	0.700	0.735	0.389	0.5
Naive Bayes	Gaussian Naive Bayes	0.750	0.689	0.722	0.368	0.4
KNN	K-Nearest Neighbors	0.710	0.662	0.712	0.307	1.0
SVM	Gaussian SVM	0.770	0.718	0.772	0.420	5.4
	Linear SVM	0.752	0.686	0.746	0.352	7.3
	Sigmoid SVM	0.459	0.390	0.462	-0.244	5.3

Training dataset 80%, test dataset 20%, stratified 10-fold 교차검증을 통해 초기분석을 진행하였다 [2,3]. 각 방법별로 가장 큰 AUC 값을 보여주는 모델을 채택하였다.

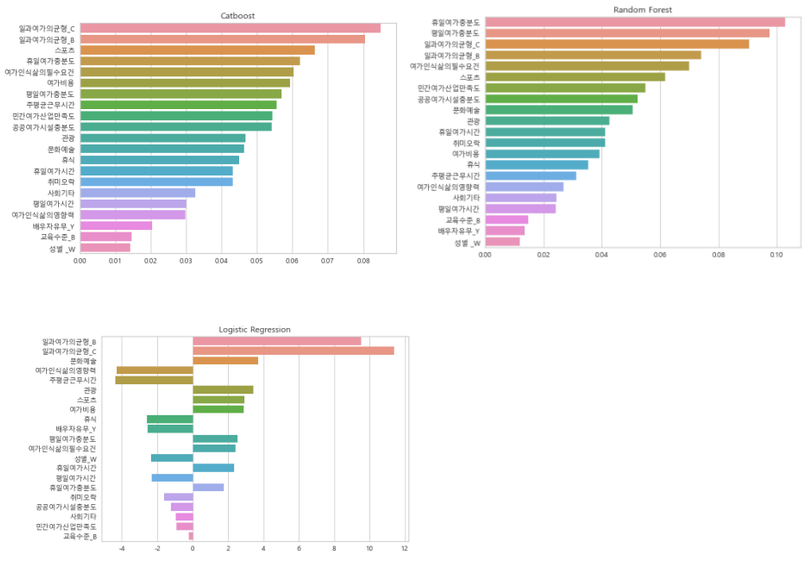
3. 하이퍼파라미터 튜닝 후 모형간 성능 비교

Model	Training				Test			
	AUC	Accuracy	F1	MCC	AUC	Accuracy	F1	MCC
CatBoost	0.786	0.723	0.764	0.432	0.747	0.671	0.713	0.335
Random Forest	0.787	0.720	0.752	0.431	0.749	0.678	0.704	0.350
Logistic Regression	0.769	0.711	0.756	0.407	0.747	0.687	0.730	0.367
LDA	0.769	0.712	0.758	0.408	0.749	0.689	0.731	0.370
Gaussian Naive Bayes	0.753	0.689	0.722	0.370	0.703	0.645	0.671	0.285
KNN	0.760	0.698	0.750	0.378	0.710	0.657	0.704	0.305
Gaussian SVM	0.778	0.707	0.742	0.405	0.755	0.692	0.717	0.379

4. 결정경계 및 ROC 곡선



5. 예측변수 중요도 Plot



4. Conclusion

- 하이퍼파라미터 튜닝 기준으로 training dataset은 Random Forest, CatBoost, Gaussian SVM 순으로, test dataset은 Gaussian SVM, Random Forest, LDA 순으로 AUC 값이 크게 나타났다.
- 여가활동에서 문화예술, 스포츠, 관광활동을 주로 하는 사람들은 취미오락, 휴식, 사회기타활동을 주로 하는 사람들에 비해 변수중요도 및 여가생활만족도가 높았다.
- 여가시간과 여가시간충분도가 증가할수록 여가생활만족도 역시 높아지며, 특히 충분도의 변수중요도가 높은 경향을 보였다. 주평균근무시간이 많은 사람들은 여가생활만족도가 낮았다.
- 일과 여가의 균형은 변수중요도가 전반적으로 매우 높으며, 여가활동을 많이 하는 집단이 일을 많이 하는 집단에 비해 여가생활만족도가 높았다.

References

[1] http://stat.mcst.go.kr/mcst/WebPortal/public/main/main.html.
[2] https://pycaret.org/.
[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1F1A1A01048397)