**여가활동만족도 예측 및 분류 – 20, 30대를 중심으로**

2. 본론

2.1. 주제 및 연구대상

본 연구에서는 20~30대의 여가생활 조사대상자의 데이터를 활용하여 여가생활만족도를 예측하고자 한다.

문화체육관광부의 국민여가활동조사 자료를 이용하였으며, 본 연구는 2020년을 대상 연도로 하였다. 해당 자료는 홈페이지에서 다운받을 수 있다.

(홈페이지 링크: http://stat.mcst.go.kr/mcst/WebPortal/public/main/main.html)

조사 대상자는 총 10,088명이며, 본 연구는 그 중 20~30대, 총 3,180명을 연구 대상자로 선정하였다.

|  |  |
| --- | --- |
| 연령 | n(%) |
| 15~19세 | 696(6.9%) |
| 20세~29세 | 1536(15.2%) |
| 30세~39세 | 1644(16.3%) |
| 40세~49세 | 1888(18.7%) |
| 50세~59세 | 1858(18.4%) |
| 60세~69세 | 1377(13.6%) |
| 70세 이상 | 1166(11.5%) |
| 총합 | 10088 |

2.2. 도구설명

2.2.1. 인구사회학적요인

A. 인구사회학적 요인

성별, 교육수준, 배우자여부를 활용하였다. 성별은 남/여로, 배우자여부는 예/아니오, 교육수준의 경우는 최종학력 고등학교 이하/대학교 이상으로 구분된다.

배우자유무의 결측 값이 있는 대상 1명은 제외하였다. 총 3179명

B. 주평균근무시간

주평균근무시간은 다음과 같이 변수변환을 진행하였다.

|  |  |
| --- | --- |
| 시간 | 범주 |
| 0시간 (일을 하지 않음.) | 0 |
| 0시간 초과 40시간 이하 | 1 |
| 40시간 이상 52시간 미만 | 2 |
| 52시간 이상 | 3 |

2.2.2. 여가활동참여실태

B. 여가생활유형

여가생활은 각 항목별로 문화예술관람 및 참여활동, 스포츠관람 및 참여활동, 관광활동, 취미오락활동, 휴식활동, 사회 및 기타활동까지 총 6가지의 범주로 구분된다.

문화예술항목의 경우는 전시회 관람, 박물관 관람, 음악연주회관람, 전통예술공연관람, 연극공연관람, 무용공연관람, 연예공연관람 8가지 관람활동과 문학행사참여, 글짓기, 미술활동, 악기연주/노래교실, 전통예술 배우기, 사진촬영, 춤/무용의 7가지 참여활동까지 총 15가지의 세부 활동으로 구성된다. 스포츠 활동의 경우는 총 20가지의 활동이 존재한다.

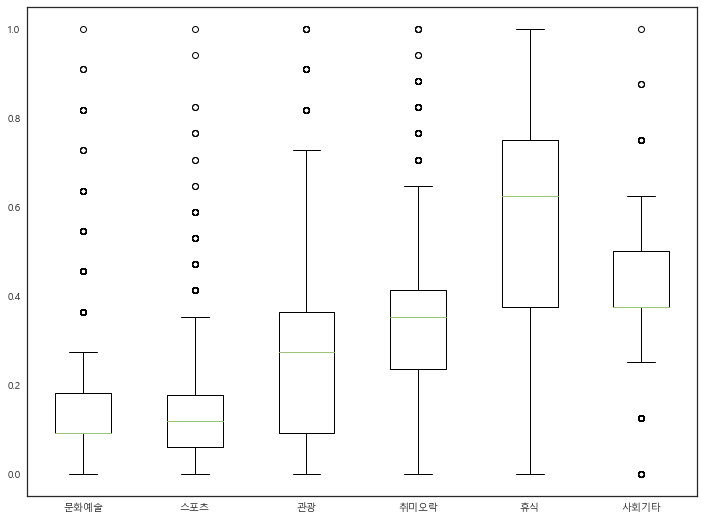
세부 구분의 경우 다음과 같다.

|  |  |
| --- | --- |
| 문화예술관람 | 전시회관람 |
| 박물관관람 |
| 음악연주회관람 |
| 전통예술공연관람 |
| 연극공연관람 |
| 무용공연관람 |
| 영화관람 |
| 연예공연관람 |
| 문화예술참여 | 문학행사참여 |
| 글짓기 |
| 미술활동 |
| 악기연주/노래교실 |
| 전통예술 배우기 |
| 사진촬영 |
| 춤/무용 |
| 스포츠관람 | 스포츠 경기 직접관람 |
| 스포츠 경기 간접관람 |
| 격투 스포츠 경기관람 |
| 온라인게임 경기관람 |
| 스포츠참여 | 농구/배구/야구/축구/족구 |
| 테니스/스쿼시 |
| 당구/포켓볼 |
| 볼링/탁구 |
| 골프 |
| 수영 |
| 윈드서핑/수상스키/스킨스쿠버다이빙/래프팅/요트 |
| 스노보드/스키 |
| 아이스스케이트/아이스하키 |
| 헬스/에어로빅 |
| 요가/필라테스/태보 |
| 배드민턴/줄넘기/맨손스트레칭체조/훌라후프 |
| 육상/조깅/속보 |
| 격투 스포츠 |
| 댄스스포츠 |
| 사이클링/산악자전거 |
| 인라인스케이트 |
| 승마/암벽등반/철인3종경기/서바이벌 |
| 관광활동 | 문화유적방문 |
| 자연명승 |
| 삼림욕 |
| 국내캠핑 |
| 해외여행 |
| 소풍/야유회 |
| 온천/해수욕 |
| 유람선 |
| 테마파크/놀이공원/동물원/식물원 |
| 지역축제 |
| 자동차드라이브 |
| 취미오락활동 | 수집활동 |
| 생활공예 |
| 요리 |
| 반려동물 |
| 노래방 |
| 인테리어 |
| 등산 |
| 낚시 |
| 홈페이지/블로그관리 |
| 인터넷검색/채팅/SNS |
| 게임 |
| 보드게임/퍼즐/큐브 |
| 바둑/장기/체스 |
| 겜블/복권 |
| 쇼핑/외식 |
| 음주 |
| 독서 |
| 미용 |
| 어학/기술/자격증 공부 |
| 이색/테마카페 체험 |
| 원예 |
| 휴식활동 | 산책/걷기 |
| 목욕/사우나/찜질방 |
| 낮잠 |
| TV시청 |
| 비디오/DVD/VOD 시청 |
| 라디오/팟캐스트 청취 |
| 음악감상 |
| 신문/잡지 |
| 아무것도 안 하기 |
| 사회기타활동 | 사회봉사활동 |
| 종교활동 |
| 클럽/나이트/디스코/캬바레 |
| 가족 및 친지방문 |
| 잡담/통화/문자 |
| 계모임/동창회/사교모임 |
| 데이트/미칭/소개팅 |
| 친구만남/동호회 |
| 기타 여가활동 |

지난 1년 간 한번 이상 참여한 여가 활동의 총합을 나타낸 값이다. 문화예술은 총 15가지의 활동으로 구성되었으니 모두 참여하였다면 값은 15가 될 것이다.

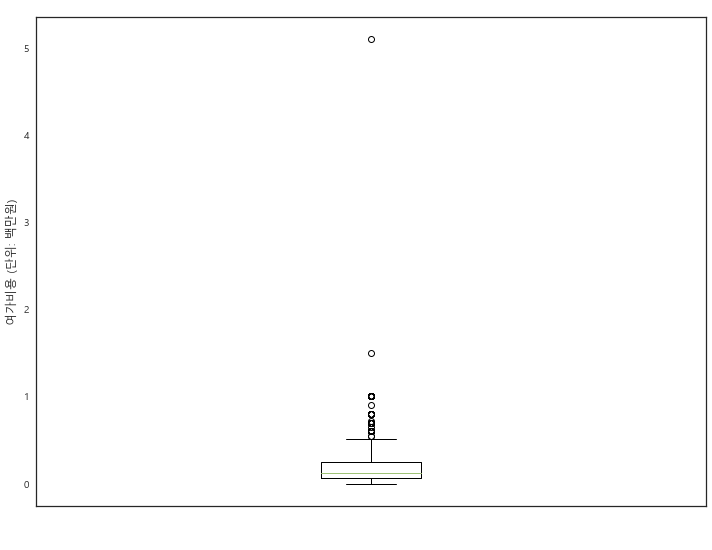
마지막으로 각 범주별로 MinMaxScaler를 사용하여 표준화 작업을 하였다. 각 활동별로 최대 개수가 차이가 나기 때문에 일치시키기 위함이다. 최댓값은 1, 최솟값은 0으로 구성된다,

다음 그림은 각 활동에 대한 상자 그림이다. 가로축은 여가생활유형, 세로축은 크기이다.



B. 여가비용

지난 1년동안 1달 평균 얼마나 지출했는지에 대한 값이다. 연구대상자의 여가비용을 상자 그림으로 나타내면 다음과 같다. 세로축 (또는 가로축) 이 실제 비용으로, 단위는 백만원이다.

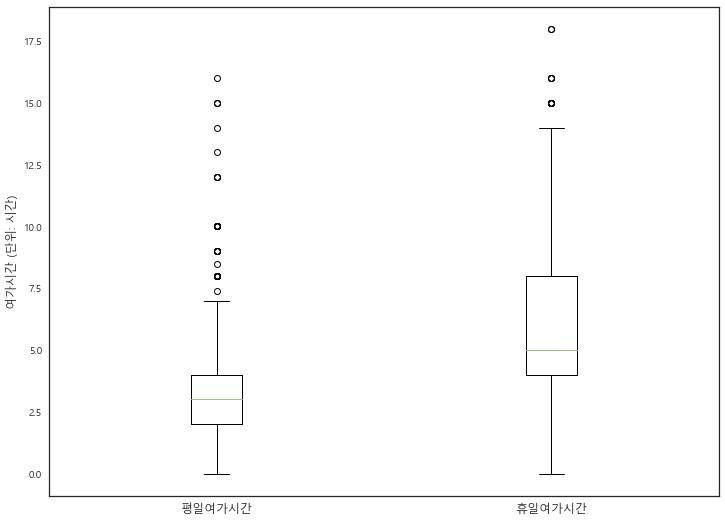


본 연구는 값을 그대로 사용하지 않고, 다음과 같이 사분위수로 나누어 처리하였다.

|  |  |
| --- | --- |
| 6만원 이하 | 1 |
| 6만원 초과, 12만원 이하 | 2 |
| 12만원 초과, 25만원 이하 | 3 |
| 25만원 초과 | 4 |

C. 여가시간 및 충분도

지난 1년간 하루 평균 여가시간에 대한 값이다. 평일과 휴일로 따로 구분되며, 본 연구는 하루 평균 여가 시간과 7점 척도로 이루어진 여가시간 충분도를 사용하였다. 대상자의 여가시간 분포도를 상자그림으로 나타내면 다음과 같다. 가로축이 평일과 휴일 구분, 세로축이 시간이다.



본 연구는 값을 그대로 사용하지 않고, 다음과 같이 시간 단위로 나누어 처리하였다.

|  |  |
| --- | --- |
| 3시간 이하 | 1 |
| 3시간 초과, 5시간 이하 | 2 |
| 5시간 초과, 7시간 이하 | 3 |
| 7시간 초과, 9시간 이하 | 4 |
| 9시간 초과 | 5 |

2.2.3. 여가공간

D. 공공여가시설충분도

공공여가시설 충분도는 생활권 내의 공공 문화여가시설 충분도에 관련된 값이다. 해당 설문은 7점 척도로 구성 되어있다. 본 연구는 공공문화 및 여가시설 이용 충분도와 공공문화 및 여가시설 프로그램 충분도 항목을 사용하였으며 해당 7점 척도의 평균값을 사용하였다.

해당 설문의 크론바하-알파값은 0.907로 내적 일관성이 충분히 높다고 할 수 있다.

E. 민간여가산업만족도

민간여가산업만족도는 우리나라 여가산업 만족도에 관련된 값이다. 해당 설문은 7점 척도로 구성 되어있다. 본 연구는 여가 관련 공간산업, 용품산업, 서비스산업의 만족도 항목을 사용하였으며 해당 7점 척도의 평균값을 사용하였다.

해당 설문의 크론바하-알파값은 0.865으로 내적 일관성이 충분히 높다고 할 수 있다.

2.2.4. 여가인식

E. 일과 여가의 균형

자신의 삶에서 일과 여가생활 간 균형이 잘 이루어지고 있는지에 대한 값이다. 해당 설문은 7점 척도로 구성 되어있다. 1번으로 갈수록 일에 집중하고 있다는 뜻이며, 7번으로 갈수록 여가에 집중하고 있다는 뜻이다. 본 연구는 다음과 같이 변수변환을 진행하였다.

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

F. 여가인식

여가인식은 2가지의 설문으로 이루어져 있다. 여가활동이 삶의 필수적인 요건인지에 대해 묻는 문항과 여가활동이 삶에 긍정적인 영향을 끼치는지 묻는 문항이다. 해당 설문은 7점 척도로 구성 되어있으며 별도의 변수변환은 하지 않았다.

모든 변수를 한번 더 정리해보면 다음과 같다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Main Category | Variable | Category | Range | n(%) or M±SD | 여가생활만족도 |
| 인구사회학적요인 | 성별 | 남 |  | 1663 (52.3%) | 4.63 |
| 여 |  | 1516 (47.6%) | 4.52 |
| 배우자여부 | 있다 |  | 1183 (37.2%) | 4.68 |
| 없다 |  | 1996 (62.8%) | 4.41 |
| 교육수준 | 고등학교 졸업 이하 |  | 602 (18.9%) | 4.34 |
| 대학교 이상 |  | 2577 (81.0%) | 4.63 |
|  | 주평균근무시간 | 0시간 (일을 하지 않음.) |  | 1212 (38.1%) | 4.62 |
| 0시간 초과 40시간 이하 |  | 1060 (33.3%) | 4.65 |
| 40시간 이상 52시간 미만 |  | 564 (17.7%) | 4.56 |
| 52시간 이상 |  | 343 (10.8%) | 4.23 |
| 여가생활유형 | 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 | 8만원 이하 |  | 829 (26.1%) | 4.34 |
| 8만원 초과, 15만원 미만 |  | 766 (24.1%) | 4.52 |
| 15만원 이상, 20만원 미만 |  | 844 (26.5%) | 4.66 |
| 20만원 이상 |  | 740 (23.2%) | 4.81 |
| 여가시간 | 평일여가시간 | 3시간 이하 |  | 1926 (60.6%) | 4.44 |
| 3시간 초과, 5시간 이하 |  | 911 (28.6%) | 4.71 |
| 5시간 초과, 7시간 이하 |  | 225 (7.0%) | 5.00 |
| 7시간 초과, 9시간 이하 |  | 65 (2.0%) | 5.06 |
| 9시간 초과 |  | 52 (1.6%) | 4.87 |
| 휴일여가시간 | 3시간 이하 |  | 712 (22.4%) | 4.29 |
| 3시간 초과, 5시간 이하 |  | 998 (31.4%) | 4.49 |
| 5시간 초과, 7시간 이하 |  | 575 (18.1%) | 4.61 |
| 7시간 초과, 9시간 이하 |  | 435 (13.7%) | 4.86 |
| 9시간 초과 |  | 459 (14.4%) | 4.89 |
| 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 |  |
|  | 일과여가의균형 | 일에 집중하는 집단 |  | 1058 (33.3%) | 3.98 |
|  | 균형을 이루는 집단 |  | 1378 (43.3%) | 4.81 |
|  | 여가에 집중하는 집단 |  | 743 (23.3%) | 4.99 |
|  | 여가인식 | 삶의필수요건 | 1-7 | 5.70±0.92 |  |
|  | 삶의영향력 | 1-7 | 5.78±0.86 |  |

|  |  |
| --- | --- |
| 일과여가의균형 | 여가생활만족도 |
| 1 | 3.07 |
| 2 | 3.80 |
| 3 | 4.20 |
| 4 | 4.81 |
| 5 | 4.94 |
| 6 | 5.10 |
| 7 | 5.62 |

2.3 상관계수, p-value

각 변수별로 상관계수와 p-value값을 나타내면 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **분류** | **문화예술** | **스포츠** | **관광** | **취미오락** | **휴식** | **사회기타**  **활동** | **여가비용** | **평일여가**  **시간** | **평일여가**  **충분도** | **휴일여가**  **시간** | **휴일여가**  **충분도** | **공공여가**  **시설충분도** | **민간여가**  **산업만족도** | **여가인식:삶의필수요건** | **여가인식:삶의영향력** | **주평균**  **근무시간** |
| **문화예술** |  | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | 0.204 | 0.112 | 0.203 | 0.131 | 0.006 | <0.001 | <0.001 | <0.001 | 0.001 |
| **스포츠** | 0.38 |  | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | 0.302 | 0.005 | 0.289 | <0.001 | 0.406 | <0.001 | <0.001 | <0.001 | 0.569 |
| **관광** | 0.36 | 0.31 |  | <0.001 | <0.001 | <0.001 | <0.001 | 0.115 | 0.01 | <0.001 | <0.001 | 0.507 | <0.001 | <0.001 | <0.001 | 0.524 |
| **취미오락** | 0.33 | 0.40 | 0.35 |  | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 | 0.52 | 0.872 | <0.001 | <0.001 | 0.018 |
| **휴식** | 0.16 | 0.20 | 0.33 | 0.50 |  | <0.001 | <0.001 | <0.001 | <0.001 | 0.8 | 0.947 | 0.172 | 0.548 | 0.002 | <0.001 | 0.704 |
| **사회기타활동** | 0.15 | 0.22 | 0.36 | 0.43 | 0.47 |  | <0.001 | <0.001 | <0.001 | 0.285 | 0.257 | <0.001 | 0.07 | <0.001 | <0.001 | 0.156 |
| **여가비용** | 0.14 | 0.18 | 0.16 | 0.12 | 0.11 | 0.08 |  | <0.001 | <0.001 | 0.154 | <0.001 | 0.789 | <0.001 | <0.001 | <0.001 | <0.001 |
| **평일여가시간** | -0.02 | 0.02 | 0.03 | 0.09 | 0.07 | 0.07 | 0.11 |  | <0.001 | <0.001 | <0.001 | 0.082 | <0.001 | <0.001 | <0.001 | <0.001 |
| **평일여가충분도** | -0.03 | 0.05 | 0.05 | 0.14 | 0.11 | 0.10 | 0.21 | 0.59 |  | <0.001 | <0.001 | 0.008 | <0.001 | <0.001 | <0.001 | 0.069 |
| **휴일여가시간** | 0.02 | 0.02 | -0.07 | 0.08 | 0.00 | -0.02 | 0.03 | 0.38 | 0.26 |  | <0.001 | <0.001 | <0.001 | 0.024 | 0.016 | <0.001 |
| **휴일여가충분도** | 0.03 | 0.07 | -0.07 | 0.07 | 0.00 | -0.02 | 0.08 | 0.33 | 0.47 | 0.69 |  | <0.001 | <0.001 | <0.001 | <0.001 | 0.019 |
| **공공여가시설충분도** | 0.05 | 0.01 | -0.01 | -0.01 | -0.02 | -0.06 | 0.00 | -0.03 | -0.05 | 0.19 | 0.12 |  | <0.001 | 0.052 | 0.02 | 0.178 |
| **민간여가산업만족도** | 0.07 | 0.08 | 0.07 | 0.00 | 0.01 | -0.03 | 0.10 | 0.06 | 0.09 | 0.13 | 0.16 | 0.37 |  | <0.001 | <0.001 | 0.405 |
| **여가인식:삶의필수요건** | 0.14 | 0.14 | 0.16 | 0.13 | 0.06 | 0.06 | 0.20 | 0.10 | 0.15 | 0.04 | 0.11 | -0.03 | 0.18 |  | <0.001 | 0.001 |
| **여가인식:삶의영향력** | 0.15 | 0.16 | 0.15 | 0.15 | 0.08 | 0.07 | 0.21 | 0.09 | 0.12 | 0.04 | 0.13 | -0.04 | 0.21 | 0.73 |  | 0.569 |
| **주평균근무시간** | -0.05 | 0.01 | 0.02 | -0.04 | -0.01 | -0.03 | 0.12 | -0.20 | -0.03 | -0.14 | -0.04 | -0.02 | -0.01 | -0.04 | -0.04 |  |

3. 자료분석

H. 여가생활만족도

본 연구의 타겟변수이다. 자신의 전반적인 여가생활에 만족하는지에 대한 7점 척도이며, 1점에 가까울수록 매우 불만족, 7점에 가까울수록 매우 만족이다. 여가생활만족도는 다음과 같이 변수변환을 진행하였다.

|  |  |  |
| --- | --- | --- |
| 7점 척도 | 내용 | 범주 |
| 1, 2, 3, 4 | 여가생활에 만족하지 않음. | A |
| 5, 6, 7 | 여가생활에 만족함. | B |

3.1. 초기 분석

Stratified 10-fold 교차검증을 사용하여 초기 분석을 진행하였다. 본 연구는 각 방법별로 가장 좋은 AUC 점수를 가진 모델을 사용할 것이다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Model** | **AUC** | **Accuracy** | **F1** | **MCC** | **Time** |
| Boosting | CatBoost Classifier | 0.789 | 0.727 | 0.769 | 0.44 | 17.1 |
| Gradient Boosting Classifier | 0.783 | 0.718 | 0.761 | 0.421 | 1.9 |
| Light Gradient Boosting Machine | 0.779 | 0.709 | 0.751 | 0.405 | 2.7 |
| AdaBoost Classifier | 0.771 | 0.719 | 0.762 | 0.423 | 3.1 |
| Extreme Gradient Boosting | 0.760 | 0.702 | 0.744 | 0.388 | 5.6 |
| Tree | Random Forest Classifier | 0.780 | 0.721 | 0.763 | 0.427 | 3.8 |
| Extra Trees Classifier | 0.780 | 0.719 | 0.762 | 0.423 | 3.7 |
| Decision Tree Classifier | 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 | Linear Discriminant Analysis | 0.768 | 0.713 | 0.759 | 0.409 | 0.5 |
| Quadratic Discriminant Analysis | 0.755 | 0.700 | 0.735 | 0.389 | 0.5 |
| Naïve Bayes | Gaussian Naïve Bayes | 0.750 | 0.689 | 0.722 | 0.368 | 0.4 |
| KNN | K Neighbors Classifier | 0.710 | 0.662 | 0.712 | 0.307 | 1.0 |
| Support Vector Machine | Gaussian SVC | 0.770 | 0.718 | 0.772 | 0.420 | 5.4 |
| Linear SVC | 0.752 | 0.686 | 0.746 | 0.352 | 7.3 |
| Sigmoid SVC | 0.459 | 0.390 | 0.462 | -0.244 | 5.5 |

Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz

3.2. 모델

A. Catboost (홈페이지: https://catboost.ai/)

러시아 회사 Yandex에서 2017년에 출시되었으며, Gradient Boosting Machines 알고리즘 기반의 모델이다. GBM의 단점인 학습 시간이 오래 걸린다는 점과, 과적합 문제를 개선하였다. 별도의 파라미터 조절 없이도 효율적인 성능을 보여주는 것으로 알려져 있다.

Catboost 알고리즘은 범주형 변수 전처리와 오버피팅 문제 해결에 초점을 둔 Ordered Boosting 기법이다.

Ordered Boosting은 모든 잔여 오차(Residual Error)를 차례로 학습하는 기존의 부스팅 모델과 달리 일부 데이터의 잔여 오차를 계산하여 모델을 만들며, 이 모델을 통해 남은 데이터의 잔여 오차를 계산하는 기법이다. 또한, Ordered Boosting에 Random Permutation을 통해 데이터 순서를 섞어줌으로써 오버피팅(Overfitting)을 방지한다[4].

범주형 변수 전처리를 위해서 Catboost 알고리즘은 Random Permutation을 거친 데이터셋에서 같은 범주를 가진 변수들의 평균 표본 값을 계산한다.

Catboost 알고리즘은 동일한 information gain을 가진 변수들을 하나로 묶는 변수 조합(feature combinations)을 통해 훈련 속도를 향상시킨다[4]. 또한, 최적의 하이퍼 파라미터를 찾기 위해 GridSearchcv나 RandomizedSearchcv를 사용하는 다른 앙상블 알고리즘과 달리 초기 하이퍼 파라미터값이 최적화 되어있어서 파라미터 튜닝 절차가 요구되지 않는다.

B. Random Forest

미국의 통계학자 LEO BREIMAN이 2001년 개발.

Decision Tree 알고리즘 기반이며, 여러 개의 나무가 각각 의사결정 후, 종합적으로 판단한다. 학습이 빠르며 효율이 뛰어난 것으로 알려져 있다.

C. LDA

데이터가 특정 범주로 나뉘어질 때, 선형을 이루는 결정경계를 사용해 분류하는 모델이다. 각 집단은 정규분포형태의 확률분포를 가지며, 설명 변수 간의 공분산 구조를 반영한다.

D. Naïve Bayes

나이브 베이즈는 베이즈 정리에 기반한 알고리즘이며 모든 특성에 대해 연관성 없이 독립을 가정한다.

E. KNN

새로운 데이터가 주어졌을 때, 데이터로부터 거리가 가까운 k개 이웃의 정보로 새로운 데이터를 예측하는 지도학습 알고리즘이다. 단순하며 훈련이 빠른 편이다.

F. SVM (구체적인 레퍼런스 필요)

서포트 벡터를 사용하여 결정경계를 정의하며 이 결정경계는 데이터의 속성에 따라 증감하며 결정경계가 고차원이 되면 초평면이 된다.

3.3. 혼동행렬 및 평가점수

A. 혼동행렬 (Confusion Matrix)

우선 혼동행렬을 정의하면 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 실제값 | |
| Positive | Negative |
| 예측값 | Positive | TP | FP |
| Negative | FN | TN |

TP는 양성으로 예측하고 실제로도 양성일 때,

TN은 음성으로 예측했고 실제로도 음성일 때,

FP는 양성으로 예측했지만 실제로는 음성일 때,

FN은 음성으로 예측했지만 실제로는 음성일 때를 나타낸다.

B. Accuracy

정확도를 나타내며, 전제 예측 중 얼마나 잘 맞추었는지 나타내는 지표이다.

(TP+TN)/All로 정의된다.

C. Recall

재현율을 나타내며, 실제 Positive를 얼마나 잘 예측했는지 나타내는 지표이다.

TP/(TP+FN)로 정의된다.

D. Precision

정밀도를 나타내며, Positive로 예측한 것을 얼마나 잘 맞추었는지 나타내는 지표이다.

TP/(TP+FP)로 정의된다.

E. Specificity

특이도를 나타내며, 실제 Negative를 얼마나 잘 예측했는지 나타내는 지표이다.

TN/(FP+TN)로 정의된다.

F. F1 Score

Recall과 Precision의 조화 평균이다.

G. AUC

우선 ROC는 x축은 1- Specificity, y축은 Recall로 나타내는 곡선이며, AUC는 그 ROC Curve의 밑면적을 계산한 값이다.

H. MCC (Matthews Correlation Coefficient)

정답 벡터와 예측 벡터가 얼마나 유사한지 나타내는 지표이다.

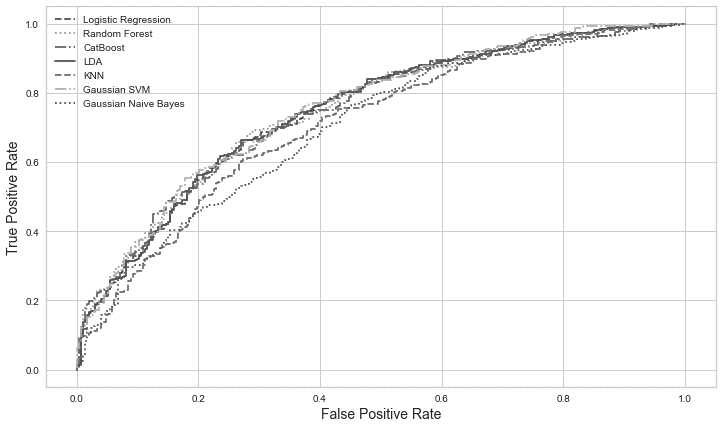
TP×TN−FP×FN√(TP+FP)(TP+FN)(TN+FP)(TN+FN)로 정의된다.

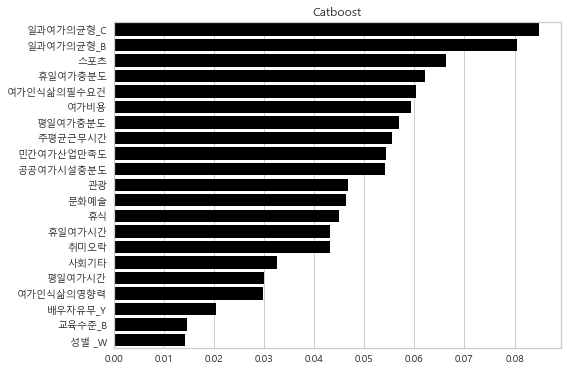
4. 하이퍼파라미터 튜닝

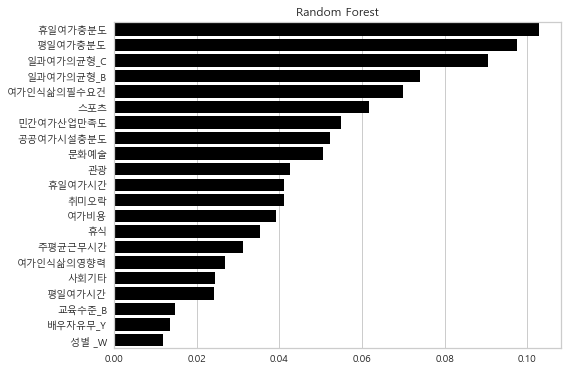
초기 분석에서 각 방법별로 가장 좋은 AUC 점수를 가진 모델을 선별한 후 하이퍼파라미터 튜닝을 진행하였다. 조정된 하이퍼파라미터는 다음과 같다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Hyper parameter** | **Default** | **Tuned** | **설명** |
| Catboost | learning\_rate | 0.015 | 0.1 | 학습률 |
| depth | 6 | 10 | 깊이 |
| max\_leaves | 64 | 1024 | 최대 잎 개수 |
| l2\_leaf\_reg | 3 | 8 | 손실함수에 가중치의 제곱을 더한다. 전체적으로 가중치를 작아지게 함으로 과적합을 방지하는 파라미터. |
| random\_strength | 1 | 0.8 | 트리 구조를 선택할 때 분할 평가 요소에 무작위 점수를 더함. 과적합을 방지하는 파라미터.  The amount of randomness to use for scoring splits when the tree structure is selected. |
| Random Forest | max\_depth | None | 10 | 최대 깊이 |
| bootstrap | True | False | Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree. |
| criterion | gini | entropy | The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. |
| Class\_weight | None | Balanced |  |
| max\_features | auto | log2 | The number of features to consider when looking for the best split |
| min\_samples\_leaf | 1 | 6 | 잎사귀 노드가 구성되기 위한 최소 표본의 수. |
| min\_samples\_split | 2 | 5 |  |
| n\_estimators | 100 | 270 | 결정 트리의 개수 |
| min\_impurity\_decrease | 0 | 0.0002 | 설정값 이상으로 불순도가 감소할 경우, 노드를 분할.  A node will be split if this split induces a decrease of the impurity greater than or equal to this value. |
| LDA | solver | svd | lsqr | ‘svd’: Singular value decomposition (default). Does not compute the covariance matrix, therefore this solver is recommended for data with a large number of features.  ‘lsqr’: Least squares solution. Can be combined with shrinkage or custom covariance estimator. |
| shrikage | None | auto | None: no shrinkage (default).  ‘auto’: automatic shrinkage using the Ledoit-Wolf lemma. |
| Logistic Regression | C | 1 | 9.463 | L1 또는 L2 규제 강도를 조절하는 alpha 값의 역수. C가 커질수록 규제 강도가 커짐.  Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization. |
| Class\_weight | None | {} | Weights associated with classes in the form {class\_label: weight}. If not given, all classes are supposed to have weight one. |
| KNN | n\_neighbors | 5 | 29 | 초기 K값, K의 값이 작다면 overfitting, 크다면 underfitting의 위험이 있다. |
| metric | minkowski | Manhattan | 거리측정 방식 |
| weights | Uniform | Distance | 예측 가중치, Uniform은 균등하게, Distance는 가까울수록 영향을 받는다. |
| n\_jobs | None | -1 | 이웃을 검색하기 위해 실행하는 병렬 작업 수 |
| Gaussian Naïve Bayes | var\_smoothing | 1e-09 | 0.005 | Portion of the largest variance of all features that is added to variances for calculation stability. |
| SVM | C | 1 | 4.71 | 규제 강도 조절 |
| class\_weight | None | Balanced |  |

5. 그래프 (ROC Curve, Feature Importance)







6. 최종 결과값

하이퍼파라미터 튜닝 후 결과값은 다음과 같다.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Model** | **Hyper Parameter** | **Train** | | | | **Test** | | | |
| AUC | Accuracy | F1 | MCC | AUC | Accuracy | F1 | MCC |
| Boosting | CatBoost | Default | 0.789 | 0.727 | 0.769 | 0.44 | 0.745 | 0.676 | 0.715 | 0.344 |
| Tuned | 0.786 | 0.723 | 0.764 | 0.432 | 0.747 | 0.671 | 0.713 | 0.335 |
| Tree | Random Forest | Default | 0.780 | 0.721 | 0.763 | 0.427 | 0.755 | 0.690 | 0.728 | 0.373 |
| Tuned | 0.787 | 0.720 | 0.752 | 0.431 | 0.749 | 0.678 | 0.704 | 0.350 |
| Regression | Logistic Regression | Default | 0.768 | 0.711 | 0.757 | 0.406 | 0.747 | 0.689 | 0.731 | 0.370 |
| Tuned | 0.769 | 0.711 | 0.756 | 0.407 | 0.747 | 0.687 | 0.730 | 0.367 |
| Discriminant Analysis | LDA | Default | 0.768 | 0.713 | 0.759 | 0.409 | 0.750 | 0.692 | 0.734 | 0.377 |
| Tuned | 0.769 | 0.712 | 0.758 | 0.408 | 0.749 | 0.689 | 0.731 | 0.370 |
| Naïve Bayes | Gaussian Naive Bayes | Default | 0.750 | 0.689 | 0.722 | 0.368 | 0.696 | 0.643 | 0.671 | 0.281 |
| Tuned | 0.753 | 0.689 | 0.722 | 0.370 | 0.703 | 0.645 | 0.671 | 0.285 |
| KNN | K Nearest Neighbors | Default | 0.710 | 0.662 | 0.712 | 0.307 | 0.666 | 0.626 | 0.666 | 0.243 |
| Tuned | 0.760 | 0.698 | 0.750 | 0.378 | 0.710 | 0.657 | 0.704 | 0.305 |
| SVM | Gaussian SVC | Default | 0.770 | 0.718 | 0.772 | 0.420 | 0.746 | 0.660 | 0.717 | 0.313 |
| Tuned | 0.778 | 0.707 | 0.742 | 0.405 | 0.755 | 0.692 | 0.717 | 0.379 |