박사학위논문

피트니스 센터의 고객 세분화 및 재등록 예측에 관한 연구

- 피트니스 센터 소규모 운동 회원 정보를 이용하여 -

2024년

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

김 영 진

박 사 학 위 논 문
지도교수 홍정완

피트니스 센터의 고객 세분화 및 재등록 예측에 관한 연구

- 피트니스 센터 소규모 운동 회원 정보를 이용하여 -

A Study on Customer Segmentation and Re-registration Prediction of Fitness Center
- Using Fitness Center Group Exercise member information -

2023년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

김 영 진

박 사 학 위 논 문
지도교수 홍정완

피트니스 센터의 고객 세분화 및 재등록 예측에 관한 연구

- 피트니스 센터 소규모 운동 회원 정보를 이용하여 -

A Study on Customer Segmentation and Re-registration Prediction of Fitness Center
- Using Fitness Center Group Exercise member information -

위 논문을 컨설팅학 박사학위 논문으로 제출함

2023년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

김 영 진

김영진의 컨설팅학 박사학위 논문을 인준함

2023년 12월 일

심사위원장 이 상 복 (인)

심 사 위 원 박 인 채 (인)

심 사 위 원 신 채 상 (인)

심 사 위 원 유 연 우 (인)

심 사 위 원 홍 정 완 (인)

국 문 초 록

피트니스 센터의 고객 세분화 및 재등록 예측에 관한 연구

한 성 대 학 교 대 학 원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 컨 설 팅 전 공
김 영 진

본 연구는 피트니스 센터 소규모 운동 회원 정보를 이용하여 피트니스 센터의 고객 세분화 및 재등록 예측모델을 활용한 마케팅 방안을 제안하였다. 피트니스 산업은 다른 서비스 업종과 마찬가지로 고객과의 관계 형성이 중요하며, 고객의 특성을 반영한 고객 관계로 결국 재등록으로 이어져 수익성과 직결된다.

매번 설문 조사를 진행하여 재등록 여부를 예측하기에는 설문 응답의 한계가 있어, 평소 회원들의 센터 이용 데이터를 분석하는 체계적인 데이터 마케팅 전략이 현실적인 대안이 된다. 마케팅 분야에서 각광을 받는 마테크(Martech)를 활용하는 시도가 필요하다. 마테크(Martech)는 마케팅(Marketing)과 기술(Technology)의 합성어로 실시간으로 수집된 고객의 로그에서 특정 패턴 행동을 감지하고, 디지털 환경에서 잠재 고객의 니즈를 찾아

행동 패턴을 분석하고 타겟 마케팅 캠페인을 설계할 수 있다.

연구의 목적은 센터 내부의 고객 데이터를 수집 및 가공하여 센터 회원이 이탈되지 않고 재등록할 수 있는 프로모션, 이벤트 등 올바른 마케팅 의사결정을 할 수 있는 실무적인 측면에서 도움을 줄 수 있는 시사점을 제시하였다.

A 피트니스 프랜차이즈 N 지점을 조사 기간 (2020년 2월 ~ 2022년 1월/24개월)에 걸쳐 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 취합하였다.

피트니스 분야 선행 문헌 연구를 통해 총 7개 변수로 출석, 노쇼, 취소, 선호 강사, 선호 수업, 선호 시간, 선호 요일을 선정하였다. 범주형 변수인 선호 강사, 선호 수업, 선호 시간, 선호 요일에 대해 Frequency encoding을 통해 수치형 변수로 변환하였다. 왜도/첨도 및 KS(Kolmogorov-Smirnov) 검정을 통해 정규분포가 아닌 것으로 확인되어 비모수 통계를 이용해서 분석하였다. 비모수 검정을 진행하여 해당 변수가 재등록 여부에 의미가 있는지 확인하였다.

피처(Feature)는 데이터의 값을 잘 예측하기 위한 데이터의 특징으로, 정규분포를 하지 않아 최소-최대 스케일링을 진행하여 min_max값을 기반으로 후속 작업을 실시하였다.

군집 분석(Clustering)은 데이터 안에 존재하는 의미 있는 그룹을 찾아내는 작업으로, 팔꿈치 방법과 실루엣 계수를 비교하여 K=3일 때 실루엣 계수가 가장 크기에 3개의 군집으로 나누었다.

3개 이상의 집단에 대해 차이가 있는 지 없는 지를 알 수 있는 비모수적 분석방법으로 Kruskal-Wallis H 검정 결과 모두 유의확률이 0.000으로 귀무가설을 모두 기각하였다. 군집별 세부 분석을 할 때, 군집 1, 군집 3의 경우 취소 min_max는 유의확률이 0.05보다 작아 통계적으로 유의하여 취소 min_max 변수를 적용하며, 군집 2는 취소 min_max가 유의확률이 0.650으로 통계적으로 유의하지 않아 취소 min_max 변수를 제외하고 진행하였다.

군집 1은 총 7개 변수에 대해 노쇼 min_max는 다른 군집에 비해 높은 값, 취소 min_max는 군집2와 군집3의 중간값이 나타났다. 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 취약한 회원 군집이라 볼 수 있어

군집 1은 “취약 회원군”으로 이름 지었다. 군집 2는 총 7개 변수에 대해 속성값 H값(평균 순위)이 중간 위치로 총 3개 군집 중에서 중간 회원 군집이라 할 수 있다. 따라서 군집2는 “중간 회원군”으로 이름 지었다. 변수 군집 3은 노쇼 min_max를 제외하고 다른 군집에 비해 높은 값이 나타났다. 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 우수한 회원 군집이라 볼 수 있어 군집 3은 “우수 회원군”으로 이름 지었다.

의사결정나무(DT), 랜덤 포레스트(RF), 극한 경사도 부스팅 (XGB)에 대해 공통적으로 학습용 데이터(Training data) 70%, 평가용 데이터(Test data) 30%로 분리하였다. 일반적으로 사용하는 K값을 5로 지정한 교차검증(Cross Validation)을 통해 데이터를 여러 번 반복해서 나누어 여러 모델을 학습하여 예측 성능을 평가하였다.

재등록 예측 성능을 평가하고자 총 6가지 지표인 정확도, 정밀도, 재현도, 특이도, F1 스코어, AUC를 이용하였다. 단일 모형, 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군) 모두 DT 〈 RF 〈 XGB 순으로 성능이 우수한 것으로 확인되었다. XGB는 Boosting 앙상블 모델로 한 개의 예측 모델에 대해 error를 줄이는 방식이며, Bagging 앙상블 모델인 RF와 차이가 있다. 군집 1(취약 회원군) 〈 군집 2(중간 회원군) 〈 군집 3(우수 회원군) 순으로 예측 성능이 우수하였다.

재등록 여부를 알고 싶은 데이터의 경우 K-means 군집 분석을 통해 군집을 분류 후 해당 군집에서 최적 예측모델인 XGB 재등록 예측 확률을 확인하여 고객 관리에 활용할 수 있다.

군집별 타겟 마케팅 전략을 통해 이탈을 방지하고 재등록으로 연계하게 할 수 있다. 예약확인 안내 문자, 포인트 관리 등 각 군집의 변수 중요도를 고려하여 마케팅 방안을 실시할 수 있다.

학문적 시사점으로 첫째, 피트니스 분야 고객 세분화 재등록 예측에 있어 선행연구에 비해 고도화된 분석을 진행한 연구이다. 둘째, 비모수 통계기법을 이용하여 정규분포가 아닌 데이터에 대해 통계적으로 접근하였다. 셋째, 교차검증(Cross Validation)을 통해 특정 데이터셋의 과적합을 방지하고 일반화된 모델 생성이 가능하다. 넷째, 지도학습/비지도학습 간의 유기적인 연계를 손

꼽을 수 있다. 끝으로, 다양한 인공지능 알고리즘을 적용하여 성능을 비교한 결과 XGB가 가장 우수하였으며, XGB의 변수 중요도 (Feature Importance)를 통해 재등록에 영향에 주는 피처의 상대적인 중요도를 정량적으로 검토하여 향후 지표 관리에 활용할 수 있다.

실무적 시사점으로 첫째, 평소 고객 DB를 이용하여 자료 접근성을 개선하였다. 둘째, 고객의 특성을 고려한 고객 세분화를 통해 목표 고객을 집중적으로 관리할 수 있다. 셋째, 마케팅 관점에서 이용할 수 있는 참고자료로서 재등록 예측 확률 추정 등 실질적인 고객 관리에 초점을 맞추었다. 끝으로, 소상공인진흥공단 등 정부 기관에서 각 업종에 맞는 재등록(재구매) 예측 알고리즘을 개발하여 보급한다면 지속가능한 성장의 기반이 조성될 수 있다.

연구의 한계 및 향후 연구방향으로는 A 피트니스 센터 N지점의 소규모 운동에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건 데이터를 근거하였으나, 보다 나은 일반화를 위해 더 많은 데이터를 확보할 필요가 있다. 실증적 연구 결과의 일반화를 위해 다양한 피트니스 센터로 범위를 확대한다면 보다 가치 있는 연구가 될 것으로 기대된다.

【주요어】 마테크, 비모수 검정, 고객 세분화, 재등록 예측, 타겟 마케팅

목 차

I. 서 론 ····· 1

1.1 연구의 배경 및 목적 ····· 1
1.2 연구의 방법 및 구성 ····· 3

II. 이론적 배경 ····· 7

2.1 군집 분석을 통한 고객 세분화 ····· 7
2.1.1 군집 분석을 통한 고객 세분화 개요 ····· 7
2.1.2 군집 분석을 통한 고객 세분화 선행연구 ····· 9
2.2 재등록 예측 ····· 13
2.2.1 재등록 예측 개요 ····· 13
2.2.2 재등록 예측 인공지능 알고리즘 ····· 15
2.2.3 재등록 예측 선행연구 ····· 20
2.3 피트니스 센터 마케팅 ····· 31
2.3.1 피트니스 산업 ····· 31
2.3.2 피트니스 센터 타겟 마케팅 개요 ····· 32
2.3.3 피트니스 센터 타겟 마케팅 방안 ····· 33
2.3.4 피트니스 센터 마케팅 선행연구 ····· 37

III. 연구 설계 ····· 40

3.1 연구 모형 ····· 40
3.2 데이터 구성 ····· 41

IV. 연구 결과 ····· 50

4.1 기술 통계 ····· 50
4.2 데이터 전처리 ····· 52
4.3 재등록 영향 변수 선정 ····· 54

4.4 군집 분석을 통한 고객 세분화 ····· 62
4.4.1 K-평균 군집화 ····· 62
4.4.2 Kolmogorov-Smirov 정규성 검정 ····· 65
4.4.3 Kruskal-Wallis H 검정 ····· 67
4.5 재등록 예측 ····· 80
4.5.1 데이터 분할 ····· 80
4.5.2 k겹 교차검증 ····· 81
4.5.3 모델 성능 평가 척도 ····· 82
4.5.4 단일 재등록 예측 모형 ····· 85
4.5.5 군집별 재등록 예측 모형 ····· 87
4.5.6 재등록 예측 확률 ····· 96
4.5.7 개별 회원 수업 이력 시계열 예측 ····· 99
4.6 마케팅 전략 ····· 102
4.6.1 군집별 타겟 마케팅 개요 ····· 102
4.6.2 세부 마케팅 방안 ····· 104

V. 결 론 ····· 108

5.1 연구결과 요약 ····· 108
5.2 학문적 및 실무적 시사점 ····· 113
5.2.1 학문적 시사점 ····· 113
5.2.2 실무적 시사점 ····· 115
5.3 연구의 한계 및 향후 연구방향 ····· 116

참 고 문 헌 ····· 117

ABSTRACT ····· 141

표 목 차

[표 2-1] 군집 분석을 통한 고객 세분화 관련 선행연구 ····· 11
[표 2-2] 이탈예측모형 유형 ····· 21
[표 2-3] 재구매 의사에 영향을 미치는 요인 ····· 24
[표 2-4] 피트니스 분야 재등록 예측 관련 선행연구 ····· 27
[표 2-5] 피트니스 이외 분야 재등록 예측 관련 선행연구 ····· 29
[표 2-6] 피트니스 센터 마케팅 관련 선행연구 ····· 39
[표 3-1] 피트니스 재등록 관련 연구별 반영 변수 ····· 42
[표 3-2] 피트니스 재등록 관련 선행연구 ····· 43
[표 3-3] 범주형 변수 범주 ····· 45
[표 3-4] 주요 인코딩(encoding) 개념 및 특징 ····· 46
[표 3-5] Frequency encoding 예시 ····· 48
[표 3-6] 변수 설명 ····· 49
[표 4-1] 기술 통계 ····· 50
[표 4-2] 정규성 검정 ····· 51
[표 4-3] Mann-Whitney U 검정 의사결정 ····· 55
[표 4-4] Mann-Whitney U 검정 평균 순위 및 순위 합 ····· 56
[표 4-5] 취소의 군집에 따른 재등록 여부 범주 Kruskal-Wallis H 검정 의사결정 ····· 61
[표 4-6] K값에 따른 실루엣 계수 ····· 64
[표 4-7] 변수 군집별 정규성 검정 ····· 66
[표 4-8] 변수별 Kruskal-Wallis H 검정 의사결정 ····· 68
[표 4-9] 변수별 재등록 여부에 따른 군집 Kruskal-Wallis H 검정 의사결정 · 69
[표 4-10] 변수별 군집에 따른 재등록 여부 Kruskal-Wallis H 검정 의사결정 · 72
[표 4-11] 변수의 군집별 Kruskal-Wallis H 검정 ····· 75
[표 4-12] Kruskal-Wallis H 사후 검정 (군집 대응별 비교) ····· 77
[표 4-13] 변수 군집별 Kruskal-Wallis H 검정 속성값 ····· 78
[표 4-14] 혼동행렬 (Confusion Matrix) 구분 ····· 82
[표 4-15] 단일 재등록 예측 모형 성능 ····· 85

[표 4-16] 군집별 Test Data 재등록 예측 모형 성능 ····· 87
[표 4-17] Sobreiro 피트니스 재등록 예측 인공지능 분석 선행연구 ····· 88
[표 4-18] 군집별 Trainging Data와 Test Data 재등록 예측 모형 성능 ·· 89
[표 4-19] 단일 및 군집별 XGB 모형 예측 성능 비교 ····· 95
[표 4-20] XGB 모형 재등록 예측 확률 ····· 98
[표 4-21] 개별 회원 수업 이력 Raw-data 예시 ····· 99
[표 4-22] 개별 회원 수업 이력 시계열 작업 ····· 100
[표 4-23] 군집별 타겟 마케팅 방안 ····· 103
[표 4-24] 요일별 특화된 프로그램 ····· 107

그 림 목 차

[그림 2-1] 실루엣 지수 시각화 예시 ····· 9
[그림 2-2] 의사결정나무 가지치기 개념 ····· 16
[그림 2-3] 케어 디스크 종류 ····· 35
[그림 3-1] 연구모형 ····· 40
[그림 4-1] 최소-최대 스케일링 ····· 53
[그림 4-2] 표준화 스케일링 ····· 53
[그림 4-3] 출석 재등록 범주별 Mann-Whitney U 검정 ····· 57
[그림 4-4] 노쇼 재등록 범주별 Mann-Whitney U 검정 ····· 57
[그림 4-5] 취소 재등록 범주별 Mann-Whitney U 검정 ····· 58
[그림 4-6] 선호 강사 재등록 범주별 Mann-Whitney U 검정 ····· 58
[그림 4-7] 선호 시간 재등록 범주별 Mann-Whitney U 검정 ····· 59
[그림 4-8] 선호 수업 재등록 범주별 Mann-Whitney U 검정 ····· 59
[그림 4-9] 선호 요일 재등록 범주별 Mann-Whitney U 검정 ····· 60
[그림 4-10] 팔꿈치 방법(The Elbow Method) ····· 63
[그림 4-11] K-means 실루엣 (K=3) ····· 64
[그림 4-12] 군집별 Kruskal-Wallis H 속성 비교 ····· 78
[그림 4-13] k(5)-겹 교차검증 ····· 81
[그림 4-14] 단일 집단 XGB 모델 변수 중요도 ····· 86
[그림 4-15] 군집1(취약 회원군) XGB 모델 변수 중요도 ····· 91
[그림 4-16] 군집2(중간 회원군) XGB 모델 변수 중요도 ····· 92
[그림 4-17] 군집3(우수 회원군) XGB 모델 변수 중요도 ····· 93
[그림 4-18] XGB 모델 Flow ····· 96
[그림 4-19] XGB 모델 신규 데이터 추가 Flow ····· 96
[그림 4-20] XGB 모델 입력 데이터 변경 ····· 97
[그림 4-21] 개별 회원 시계열 재등록 예측 확률 ····· 101
[그림 4-22] 예약확인 안내 문자 예시 ····· 104

Ⅰ. 서론

1.1 연구의 배경 및 목적

피트니스 산업이 대표적인 서비스 산업으로 성장하고 있다. 피트니스 산업은 종사자의 전문지식이 필요하여 진입장벽이 높고, 주로 회원제로 운영하다 보니 평균 영업 기간이 10년이고, 2019년 기준 폐업률이 7.7%로 PC방 15.7%, 커피숍 14.4%, 제과업 11.0%에 비해 서비스 업종에서 안정적인 성장을 이어가고 있다. 피트니스 산업은 다른 서비스 업종과 마찬가지로 고객과의 관계 형성이 중요하며, 고객의 특성을 반영한 고객 관계로 결국 재등록으로 이어져 업장의 수익성과 직결된다. 코로나-19 등 환경 악화로 피트니스 산업 역시 어려움을 겪는 서비스 분야이다. 수동적인 고객관리로는 치열한 경쟁에서 생존이 어려우며, 고객지향의 경영 마인드를 기반으로 스포츠 마케팅 전략 차원의 접근이 필요하다. 재등록 시점에 임박하여 객관적이고 정량적인 정보 없이 회원에게 재등록 여부를 문의할 때 평소 고객관리가 부실하면 재등록으로 이어지지 않고 이탈로 끝나는 경우가 태반이다. 부적절하게 설계된 설문지는 고객이 만족 또는 불만족을 느끼고 있는 감정에 대해 어떻게 행동할 것인가를 파악하지 못하고 단순히 만족 또는 불만족을 아는 것은 유용성이 떨어진다.

매번 설문 조사를 진행하여 재등록 여부를 예측하기에는 설문 응답의 한계가 있어, 평소 회원들의 센터 이용 데이터를 분석하는 체계적인 데이터 마케팅 전략이 현실적인 대안이 된다. 마케팅 분야에서 각광을 받는 마테크(Martech)를 활용하는 시도가 필요하다. 마테크(Martech)는 마케팅(Marketing)과 기술(Technology)의 합성어로 실시간으로 수집된 고객의 로그에서 특정 패턴 행동을 감지하고, 디지털 환경에서 잠재고객의 니즈를 찾아 행동 패턴을 분석하고 타겟 마케팅 캠페인을 설계할 수 있다. 소비자 집단은 이질적인 요구를 가진 다양한 집합으로 구성되어 있지만, 특정 서비스에 대한 태도, 의견, 행동 등에서 비슷한 성향을 보이고 있다. 비슷한 성향을 가진 사

람들을 묶어 분석하면 세분화에 따른 맞춤형 서비스를 제공할 수 있다. 인구통계학적인 정보로 세분화하기도 하지만 직업, 연령, 소득 등 해당 정보 취합이 개인정보보호 차원에서 쉽지 않아 최근에는 고객의 행동 양식 및 이력 정보로 세분화하는 방법으로 세분화하는 것이 선호되고 있다.

일부 대형 프랜차이즈 피트니스가 있긴 하지만, 자영업자들이 운영하는 중소형 피트니스가 대다수를 차지하고 있다. 유통업, 금융업 등 고객 정보를 관리하는 산업의 경우 고객 세분화, 재등록 예측에 대해 인공지능 알고리즘 사용이 일반적이다. 피트니스 산업의 경우 고객 정보 필요성을 인식하고 있으나, 인공지능 도구에 대한 막연한 두려움으로 피트니스 분야에서 인공지능 분석이 사실상 전무하다. 이러한 영세한 사업장 특성은 자사가 보유한 데이터의 활용성이 낮아 지속 가능한 데이터 경영이 가능하도록 체계적인 컨설팅이 필요하다. 기존의 재등록(이탈) 예측 연구가 단순히 전체 고객집단을 하나의 그룹으로 다양한 모형 성능 비교에 맞추어, 고객의 군집 특성을 반영하지 못한 한계가 존재하였다.

본 연구에서는 피트니스 센터 소규모 운동(GX) 회원 정보를 기반으로 min-max 스케일링으로 데이터 전처리를 수행하고 K-means 군집 분석으로 고객 세분화를 진행하였다. 재등록 예측을 위해 인공지능 알고리즘을 통해 모형 간 비교와 재등록 예측 확률을 계산하였다. 이용 회차에 따라 재등록 예측 확률을 시계열 모니터링할 수 있다. 고객가치 증대 및 고객과의 관계 형성을 위한 고객 형태 예측의 중요성이 점차 커지고 있다. 회원의 행동 패턴을 실시간 분석하여 앞으로의 회원의 형태를 예측하는 것은 피트니스 센터의 현재 수익 증대에 기여할 뿐만 아니라 미래의 경쟁력 확보를 위한 필수 요건이다. 신규 고객을 찾기가 점점 어렵고, 고비용의 신규 고객 유치보단 재등록 유지가 훨씬 우선순위가 높다.

회원 전체를 대상을 매스 마케팅이 아닌 군집 특성에 맞는 마케팅 전략이 효과적이다. 회원 관리에 필요한 마케팅 예산이 제한적인 상황에서 예산을 효율적으로 집행할 수 있는 전략적 방향을 고려할 수 있다. 회원들의 다양한 성향과 서비스 욕구를 만족시킬 수 있는 서비스를 제공하고, 지속적으로 피트니스를 즐길 수 있는 참여 동기와 서비스 욕구 만족을 높이는 체계적인 마케팅

전략이 필요하다(임지수, 2020). 각 군집의 특성에 맞는 마케팅 방안으로 수립할 수 있다. 센터 내부의 고객 데이터를 수집 및 가공하여 센터 회원이 이탈되지 않고 재등록할 수 있는 고객관리에 적절한 마케팅 의사결정을 할 수 있는 실무적인 측면에서 도움을 줄 수 있는 시사점을 제시하시는 데 연구의 목적이 있다.

1.2 연구의 방법 및 구성

본 연구에서는 A 피트니스 프랜차이즈 N 지점을 조사기간 (2020년 2월~2022년 1월/24개월)에 걸쳐 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 취합하였다. 본 연구에서는 1,160명의 513,530건의 데이터를 교차검증(Cross Validation)을 통해 특정 데이터셋에 대한 과적합을 방지하고 더욱 일반화된 모델을 생성하였다. 또한, Random Forest, XGB 인공지능 알고리즘을 이용하여 여러 개 의사결정나무의 Bagging 앙상블(Random Forest), Boosting 앙상블(XGB)로 과적합을 해결하고 높은 예측 성능을 보였다. 교차검증, Random Forest, XGB 방식으로 일반화하여 소규모 그룹 수업(GX)을 하는 피트니스 센터 특성을 반영하고자 하였다. 조사 기간이 코로나 유행 기간으로 정부 방역조치에 따라 집합인원 제한의 영향을 고려해야 하는 데, 본 연구는 통상 6명 이하의 소규모 그룹 수업을 하는 피트니스 시설로 방역지침의 제한이 적고, 코로나로 인해 수업 취소 등은 자료에서 제외하여 코로나에 따른 데이터 영향을 최소화하였다.

재등록 여부를 예측하는 인공지능 알고리즘 성능 비교를 위해 재등록과 이탈 회원 수를 각 580명씩 같은 숫자로 분석하였다. 재등록/이탈고객의 회원 수가 다를 경우 즉 데이터가 불균일한 분포를 가지면 모델의 학습이 제대로 이루어지지 않을 확률이 높다. 이를 해결하기 위해 SMOTE(Synthetic Minority Over-sampling Technique)를 이용하여 낮은 비율 클래스 데이터의 최근접 이웃을 이용하여 새로운 데이터를 생성한다(안경민, 2021). 하지만 데이터의 불균형 상태에 따라 정밀도가 감소하고 재현율이 증가하는 등 문제가

생길 수 있어, 데이터의 불균형을 해소하고자 같은 수의 데이터를 적용하였다.

피트니스 분야 선행 문헌 연구를 통해 재등록에 영향을 주는 총 7개 변수로 출석, 노쇼, 취소, 선호 강사, 선호 수업, 선호 시간, 선호 요일을 선정하였다. 범주형 변수인 선호 강사, 선호 수업, 선호 시간, 선호 요일에 대해 Frequency encoding을 통해 수치형 변수로 변환하였다. 자료의 정규성 여부를 우선 검토해야 하는데, 크게 왜도/첨도, KS(Kolmogorov-Smirnov) 검정으로 정규성을 판단할 수 있다. 정규분포가 아닌 것으로 확인되어 비모수 통계를 이용해서 분석하였다. 표본의 크기가 충분히 크다면 표본평균의 분포는 근사적으로 정규분포를 따른다는 중심극한정리에 따라 모집단의 정규성을 가정하고 분석한다. 하지만, 정규분포를 가정하고 일반적인 통계 분석 방법을 사용하여 잘못된 결과를 얻는 등 분석의 신뢰도가 떨어질 수 있어, 정규성 여부를 확인하여 비모수 통계 방법을 세부 분석에 적용함이 바람직하다. 비모수 검정을 진행하여 해당 변수가 재등록 여부에 의미가 있는지 확인하였다.

피처 엔지니어링(Feature Engineering)은 기존에 존재하는 피처(변수)를 활용해서 새로운 정보를 추가로 생성하는 과정을 말한다. 피처(Feature)는 데이터 특성을 나타내는 값으로 변수(Variable)와 혼용되어 사용되는 데, 데이터 분석에서는 변수보단 피처라는 용어를 활용한다. 새로운 데이터를 추가로 수집하는 것이 아니라 사용 가능한 현재 데이터를 이용한다. 어떤 모델은 특정 타입의 피처에 더 잘 동작하며, 그 반대의 경우가 있다. 제대로 된 피처는 분석하고자 하는 과제에 적합하고 모형에 사용하기가 쉬워야 한다. 피처 엔지니어링은 주어진 데이터, 모델에 가장 적합한 피처를 만들어 내는 과정이다. 피처가 과도하게 많거나 대부분 피처가 관련성이 없으면 모델의 학습에 많은 비용과 어려움이 발생할 수 있다. 관련 선행연구를 통해 피처를 선택하여 분석에 적용하였다.

기계학습(Machine Learning)은 데이터의 수학적 모델을 사용하여 직접 명령없이 컴퓨터가 학습하도록 하는 컴퓨터 알고리즘이다. 과거에는 학문에 대한 깊이로 실생활과 거리가 있었다면, 최근에는 데이터 증가에 따른 활용방법과 관련 솔루션의 개발로 실용성이 높은 방법론으로 인식된다. 기계학습은 학

습 방법에 따라 크게 비지도학습과 지도학습 2가지로 나눈다. 비지도학습(Unsupervised Learning)은 출력 데이터에 대한 결과값이 없을 경우에 사용한다. 데이터에서 파악한 피처(Feature)를 바탕으로 유사한 데이터를 모아 같은 그룹을 묶는 군집 분석(Clustering Analysis) 등이 있다. 지도학습(Supervised Learning)은 입력 데이터와 출력 데이터를 통해 관계를 유추하여 올바른 해답을 제시한다(강성식, 2021). 정답을 알려주고 그에 따라 학습하는 방법으로 특정 label을 가진다. 데이터의 특징을 토대로 값을 예측하는 분류 분석(Classification)에 이용된다. 본 연구에서는 비지도학습인 군집 분석을 이용하였고, 지도학습으로 '재등록 여부'라는 해답을 특정 label로 하여 재등록 예측을 하는 분류 분석을 함께하는 OSMU(One Source Multi Use) 방식을 진행하였다.

군집 분석(Clustering Analysis)은 각 개체의 유사성을 측정하여 높은 대상 집단을 분류하고, 군집에 속한 개체들의 유사성과 서로 다른 군집에 속한 개체 간의 상이성을 규명하는 통계 분석 방법이다. 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 그룹을 발굴하는 데 사용한다. 군집 분석은 고객 집단별 형태적 특성을 파악하여 고객 세분화에 이용된다. K-평균(K means)은 군집 분석에서 가장 일반적으로 사용하는데, 알고리즘이 쉽고 간결한 특성이 있다. 몇 개의 군집을 선택할지 지정해야 하는 문제가 있어, 팔꿈치 방법과 실루엣 계수를 통해 군집 개수의 최적화를 진행하였다.

재등록 예측 (Re-registration prediction)은 고객에 관한 데이터를 활용하여 훈련 데이터와 학습 데이터로 7:3으로 나누어 모형을 생성하여 재등록에 대한 예측력을 검증하게 된다. 단 1번의 데이터 분리를 통해 모델 성능을 평가하는 것은 데이터를 어떻게 나누느냐에 따라 성능을 평가한 결과가 터무니없이 높거나 낮을 수 있어, 데이터를 나누는 방법이 중요하다. k-겹 교차검증(k-fold cross validation)은 데이터를 여러 번 반복해서 나누고 여러 모델을 학습하여 성능을 평가하는 방법으로, 여러 번 나눈 것의 평균적인 성능을 계산하면 한 번에 나누어서 학습하는 것에 비해 일반적인 성능을 얻을 수 있다. 본 연구에서는 k값을 5로 지정하여 5번 반복 진행하였다.

끝으로, 군집별 타겟 마케팅으로 고객 이탈을 방지하여 재등록을 유도할

수 있다.

본 논문의 구성은 총 5장으로 구성되어 있으며, 각 장의 내용은 다음과 같다.

제1장 서론에서는 연구의 배경 및 목적, 그리고 연구의 방법 및 구성을 기술하였다.

제2장 이론적 배경에서는 군집 분석을 통한 고객 세분화 개요, 군집 분석을 통한 고객 세분화 선행연구, 재등록 예측 개요, 재등록 예측 인공지능 알고리즘, 재등록 예측 선행연구 그리고, 피트니스 산업과 피트니스 센터 타겟 마케팅 개요, 피트니스 센터 타겟 마케팅 방안, 피트니스 센터 마케팅 선행연구를 정리하였다.

제3장 연구 설계에서는 선행연구를 토대로 연구 모형을 수립하고, 데이터 구성과 변수 선정 그리고 Frequency encoding 방법을 제시하였다.

제4장 연구 결과에서는 기술 통계, 데이터 전처리, 재등록 영향 변수 선정, 군집 분석을 통한 고객 세분화, 재등록 예측 그리고 마케팅 전략을 제시하였다.

제 5장 결론에서는 연구결과에 대한 요약, 학문적 및 실무적 시사점 제시, 연구의 한계 및 향후 연구 방향을 제언하였다.

Ⅱ. 이론적 배경

2.1 군집 분석을 통한 고객 세분화

2.1.1 군집 분석을 통한 고객 세분화 개요

군집 분석(Cluster Analysis)은 대표적인 비지도학습으로, 유사한 특성을 가진 대상들끼리 묶어 집단을 만드는 것을 말한다(임지수, 2020). 군집 분석은 주어진 개체 중에서 유사한 것을 몇 개의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적 데이터 분석방법이다. 고객 세분화(Customer Segmentation)는 다양한 변수들을 활용하여 비슷한 특징을 갖는 고객들을 그룹으로 분류하는 것을 말한다(김담희, 안가경, 2018).

대용량 데이터에 대해서는 개개의 관찰치를 요약하는 것보다는 전체를 유사한 관찰치들의 군집(cluster)으로 구분하여, 복잡한 전체보다는 그들을 잘 대표하는 군집들을 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 얻어낼 수 있다(강현철 외, 2006). 군집이 성공적으로 분류되면 군집들은 군집 동질성이 높아지고, 군집 간 이질성이 높아진다. 군집 분석은 해당 대상이 가지는 값을 거리(distance)로 환산해서 가까운 거리의 대상을 집단으로 분류하는 방법이다(임지수, 2020).

군집 분석은 우선 연구대상을 의미 있는 다수의 집단으로 구분할 수 있는 변수를 선정해야 된다. 변수를 고려해 대상 간 유사성을 의미하는 지표인 거리(distance)를 계산한다. 대상 간의 산출된 거리 기점으로 군집 분석(cluster analysis) 과정으로 계층적 방법(hierarchical method)과 비계층적 방법(nonhierarchical method)으로 구분된다. 계층적 군집 분석은 군집 형성의 위계로 인하여 모든 대상이 가까운 거리의 대상을 차례로 묶는다. 거리 측정방법에는 최단연결법, 최장연결법, 중심연결법, 평균연결법, 와드연결법이 있다.

모든 대상이 하나의 군집으로 모일 때까지 반복하며 흩어지지 않는다. 비계층적 방법은 군집의 수(K)를 미리 제공해야 하고, 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함하는 방식으로 군집을 형성한다(김담희, 안가경, 2018).

대표적인 군집 분석 방법은 K-means이며 단계별로 진행한다(임지수, 2020).

1단계 : 총 N개로 이루어진 데이터 군집을 위한 최초 군집 수 K를 확정한다.

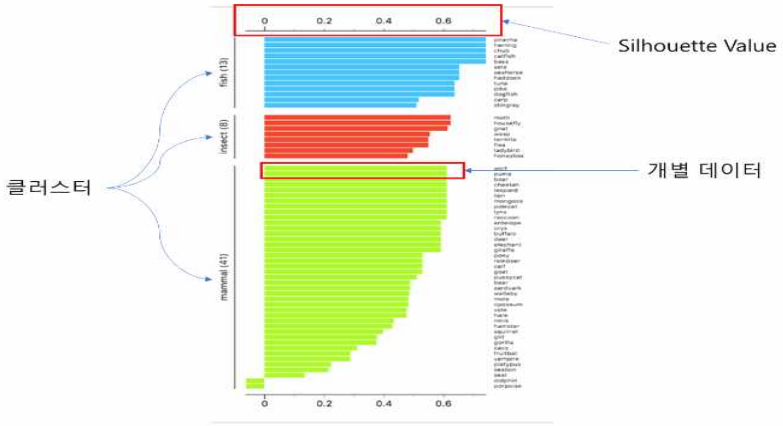
2단계 : 각 군집의 해당 변수의 중심점(Centroid)을 계산한다.

3단계 : 주어진 변수와 중심점과의 거리를 계산한다.

4단계 : 변수마다 3단계를 반복 시행한다.

5단계 : 어떤 변수도 다른 군집에 배정되지 않을 때까지 2~4단계를 반복한다. 군집 형성 후에도 일정 기준에 따라 이합집산을 반복한다.

K-mean의 K값을 미리 정해 주어야 하는 문제가 있어 이를 해결하고자 실루엣 지수를 고려할 수 있다. 실루엣 지수(Silhouette Index)는 군집 분석 결과를 평가하는 지표로써 각 데이터별로 그 데이터가 속한 군 내의 거리 기반 유사도와 인접한 군의 유사도를 비교하는 지표이다. 장점으로는 여러 군집 개수 후보에 대하여 실루엣 지수를 계산하고, 최적 군집 개수를 정할 수 있다. 군집별 색상을 정하고 각 데이터에 대한 실루엣 값을 x축으로 하여 막대 그래프 형태로 나타낼 수 있다. 이를 통하여 개별 데이터들이 1에 가까운 값을 갖는 지를 눈으로 확인할 수 있고, 그에 따라 군집분석 평가 결과를 시각적으로 이해할 수 있다(김담희, 안가경, 2018).



[그림 2-1] 실루엣 지수 시각화 예시

군집 분석은 인간의 직관과 짐작에 의존하는 전통적인 방법론에 비해 '인공지능 알고리즘'을 사용하여 고객 세그먼트(segment)를 생성하고, 고객 그룹 간의 차이를 구분하여 마케터가 알아차리지 못한 관계를 찾아 효율적인 마케팅 효과를 얻을 수 있다. 고객들에게 개선된 서비스 환경 및 품질을 제공을 위해서는 우선적으로 이용고객들에 대한 분석을 바탕으로 한 마케팅 전략이 필요하다(장하원, 2021).

2.1.2 군집 분석을 통한 고객 세분화 선행연구

고객 세분화는 인구통계학적 특성, 심리 분석적 특성, 행동적 특성 등의 기준으로 동질적인 다수의 하위 집단으로 분류하는 데, 인구통계학적 특성 기준은 소비자의 특성에 관한 것이고, 심리 분석적 특성 및 행동적 특성은 상품(서비스)에 대한 소비자 반응에 중심을 둔 것이다(김지승, 2020). 고객 세분화를 위해 군집 분석, 주성분 분석 등 다양한 분석이 이용되는데, 이 중 군집 분석이 가장 많이 사용되며, 특정 변수의 특성을 공유한 동질 집단을 서로 묶어 집단 내 통계적 동질성을 최대로 하고, 집단과의 동질성을 최소화하여 고객을 일정한 특성에 의해 구분할 수 있는 효과적인 방법으로 평가된다(김지승, 2020).

특성에 따라 분류된 고객들을 핵심 목표로 할 수 있고, 불특정 고객들 대상으로 마케팅하는 경우보다 높은 성과를 가져올 수 있다(김담희, 안가경, 2018). 고객은 기업이 제공하는 동일한 마케팅전략에 다르게 반응하기 때문에 고객의 특성에 맞는 적절한 고개관리 전략이 수립되어야 한다(홍태호, 김은미, 2010).

효과적인 고객 세분화를 위해 고려할 사항으로는 마케팅 활동이 적용 가능한 고객 대상으로 세분화 기준 변수로 사용할 수 있는 측정된 고객 특성을 고려하며, 마케팅 활동이 가능한 규모로 진행 후 마케팅 결과에 대한 차별화된 반응이 예상되어야 한다(김담희, 안가경, 2018). 마케팅전략을 제공할 목표 고객의 선정이 잘 이루어지면 기업은 수익을 증가시킬 수 있지만, 목표 고객을 잘 선정하지 못하면 기업은 마케팅 비용의 증가뿐 아니라 고객과의 관계도 악화될 수 있다(홍태호, 김은미, 2010). 어느 고객에게 어떤 마케팅 캠페인을 제시하는 것이 효과적인지를 판단하여 적절한 마케팅 캠페인을 제시해야 하며, 모든 고객에게 마케팅 전략을 제시하는 것보다 목표 고객을 결정하여 집중적으로 제시함으로써, 기업의 비용 절감은 물론 이윤을 극대화시킬 수 있다(Baesens et al., 2002). 고객 세분화는 이질적인 고객들을 보다 정확하게 이해할 수 있어 효율적으로 마케팅 전략을 개발하고 수행할 수 있게 한다(장남식, 2005).

군집 분석을 통해 관측 개체를 몇 개의 집단(군집)으로 나눔으로써, 대상 집단에 대한 이해 및 효율적인 활용이 가능하다(김윤진, 2005). 군집 분석에는 단계별로 다양한 방법이 존재하여, 분석방법의 선택에 따라 동일 자료에 대한 결과에 차이가 클 수 있어, 연구자는 각 단계별 방법과 적용 방법을 명확히 이해하고 분석해야 한다(안효경, 2009).

은행 내 다양한 빅 데이터를 융합하여 세분화를 통해 등급산출, 고객 이탈 예측 등 다양한 마케팅 모델을 개발하고자 군집 분석을 진행하였고, 특정 데이터를 활용하여 타겟 고객군을 만들어 낼 수 있다(장민석, 2017). 군집 분석을 통해 세분화된 집단에 대해 경제성과 효율성을 고려한 마케팅전략을 구사할 필요가 있다(조소려, 2020). 기존 고객 데이터에 대해 군집 분석을 진행하고, 고객의 성향에 적합한 차별화된 마케팅을 추진하여 충성고객을 확보해야

한다(김낙한, 2022). 특정 상품 또는 서비스에서 소비자들이 얻고자 하는 추구 편익이 비슷한 소비자끼리 세부 시장을 구분하여 마케팅 활동을 진행하는 것이 효과적이다(이정미, 2019). 고객들이 선호하는 참여 동기를 군집 분석을 통해 파악하여 고객 세분화 서비스 전략을 추진할 수 있다(임지수, 2020). 이용객의 선택속성 요인을 파악하기 위해 군집 분석을 실시하여 고객 만족을 극대화해야 한다(김지승, 2020).

[표 2-1] 군집 분석을 통한 고객 세분화 관련 선행연구

연구자 (연도)	논문명	연구내용	적용 분야
김수래 (2015)	고객 거래 데이터를 활용한 마케팅 자원 최적 재배치 방법론 개발	기업의 마케팅 활동과 예산을 최적화된 방향으로 재구성할 수 있는 마케팅 최적화 프로세스 방법론을 제시	유통
이수현 (2015)	빅 데이터의 군집분석을 위한 군집화 유효성 지수 개발과 응용	서포트 벡터 데이터 표현 개념을 반영한 군집화 유효성 지수 개발하여 활용사례 소개	제조업
장민석 (2017)	빅데이터를 활용한 은행권 고객 세분화 기법 연구	거래 패턴에 기반을 두고 적정 동료 그룹(peer group)을 찾아 금융등급 산정 및 상품추천 모형 제시	금융
김담희, 안가경 (2018)	구매이력 빅데이터를 활용한 타겟 마케팅 고객 세분화에 관한 연구 : K-means 클러스터링 방법을 중심으로	구매 이력 데이터를 기반으로 고객 세분화 모형을 활용하고, K-means 클러스터링을 통한 군집을 생성하여 실루엣 점수로 모형 비교	유통

이정미 (2019)	스파 방문객의 추구편익 및 시장세분화 분석	스파를 방문하는 고객의 추구편익을 기준으로 시장 세분화를 진행하며, 각 세분 시장에 효과적인 마케팅 활동을 추진	뷰티
조소려 (2020)	외식배달 서비스 고객 시장세분화 및 IPA 분석	외식배달 서비스 이용동기에 대해 군집 분석을 진행하여 3개 군집으로 시장세분화를 실시하며, 군집에 적합한 홍보 및 마케팅 전략 수립	외식
임지수 (2020)	실내테스니장 고객 특성에 따른 고객 세분화 전략	실내테니스장을 이용한 고객들이 선호하는 서비스 스케이프 요인에 따라 3개 군집으로 나누어 특성에 맞는 고객 세분화 마케팅 추진	피트니스
김지승 (2020)	패키지 여행상품 이용객의 선택속성에 따른 시장 세분화 연구	여행상품 이용객의 선택속성을 K-means 군집 분석을 진행하고 4개 군집에 대해 각 집단에 맞는 마케팅 전략 추진	관광
김낙한 (2022)	국내 MZ세대의 웰니스 라이프스타일과 숙박상품 선택속성에 따른 시장 세분화 연구	MZ세대의 웰니스 라이프 스타일 특성을 군집 분석을 통해 3개의 군집으로 시장 세분화를 추진하고 고객 경험에 대한 마케팅 차별화를 진행	관광

2.2 재등록 예측

2.2.1 재등록 예측 개요

재등록 예측(Re-registration prediction)은 이용고객의 구매 패턴을 분석하여 재등록(이탈)을 인공지능이나 수리적 모델을 활용하여 예측하는 것을 말한다(이식, 2023). 경쟁이 심화되고 있는 서비스 산업의 경우 기존 고객 유지를 위한 고객 만족이 핵심적인 경영전략이다.

행동 의도(Behavioral Intentions)는 소비자가 특성에 대한 태도를 형성한 뒤 특정한 미래행동에 나타내려고 하는 개인적인 의지와 신념으로 장기적인 관계성을 의미(이승준, 2020)하는 데, 노쇼 등 행동으로 재등록/이탈의 결과를 보여준다. 행동 의도는 소비자의 행동에 대한 예측 및 이해를 위한 태도-행동 간의 관계 연구는 사회과학의 주요 관심 분야이다(남수정, 2010). 행동 의도는 제공되는 서비스에 대한 지각에 의해 태도가 형성된 후 특정한 미래 행동에 대한 개인의 의지와 신념을 나타내는 행동을 예측하는 요소이다(이미혜, 2015). 행동 의도에 따라 고객이 기업의 상품이나 서비스를 이용할 것인가 아니면 다른 기업으로서의 전환할 것인가의 중요한 역할을 한다(김용운, 2022).

서비스 만족도는 고객들이 직·간접적으로 체험한 결과에 대한 반응으로서, 시간과 장소, 서비스 내용에 따라 다르며, 특히 개인의 경험적 수준에 따라 만족 수준이 다르다. 만족도는 소비자의 지각과 서비스 이후 소비자들이 받은 서비스 기대수준 차이의 정도로서 고객이 원하는 희망 수준이고, 기대와 지각된 서비스 성과의 일치-불일치 과정이다(Anderson, 2000).

재이용 의도는 소비자가 이용한 서비스에 대해 다시 이용하고 싶은 정도와 다른 사람에게 서비스를 이용하도록 권유하는 것을 말한다(McDougall & Levesque, 2000). 소비자에게 재구매 의도가 높다는 것은 상품(서비스)의 재구매뿐만 아니라 다른 소비자에게도 상품(서비스)을 추천할 가능성이 높다는 것을 의미한다(박현규, 2010). 재구매 의사는 고객이 제품이나 서비스에 대한 구매 경험의 평가를 근거로 재구매하려는 자신의 신념으로 소비자가 향후 미

래에도 제품이나 서비스를 반복하여 사용할 가능성을 말한다(이상우, 2015). 고객 만족은 소비자의 태도에 영향을 미치고 이러한 태도는 다시 재구매 의사에 영향을 미친다는 실증분석을 통해 입증하였다(Oliver, 1980). 기업 입장에서 서비스나 상품을 구매한 소비자에게 만족도를 높여줌으로써 향후 재등록 의도를 높이고자 노력하는 것이 현명한 판단이라 할 수 있다(이동건, 2017).

신규 고객 창출보다 기존 고객 관리가 더 효율적이고 양적 지표인 시장점유율보다 질적 성과지표인 고객 애호도가 기업 이익과 더 밀접한 관련이 있다(최상원, 2013). 제품이나 서비스를 재구매하는 경우 대다수 소비자들은 장기간 거래를 유지함으로써 기업의 수익에 기여하기에 고객의 충성도를 높여 기존 소비자를 유지하는 마케팅이 중요하다(송옥매, 2009). 고객이 경험하는 서비스와 품질, 신뢰는 고객 충성도에 밀접한 관계가 있어 두터운 충성도를 가지도록 관리해야 한다(오병임, 2016).

소비자와 지속적인 관계를 유지함으로 상호 간의 신뢰와 만족을 가질 수 있고, 재구매와 구전에 의한 효과를 얻을 수도 있다(오시현, 2023). 피트니스 센터 회원의 재구매 의도를 현재 피트니스 센터에 재등록하겠다는 의도로 볼 수 있는 데, 지도자의 서비스품질이 고객 만족과 재구매 의도에 긍정적인 영향을 준다(이유재, 공태식, 2004). 피트니스 분야에서 고객이 지도자와의 대면에서 제공하는 서비스품질에 긍정적 반응이 나타나도록 하는 것이 재등록 의도에 영향을 준다(오시현, 2023).

재구매 의도를 파악하는 것은 일회성 구매 고객이 아닌 고정 고객화할 수 있는 가능성이 높으며, 장기적인 수익 창출에 큰 도움이 된다(송종태, 2004). 피트니스 회원들이 인식하는 서비스품질의 만족 또는 불만족은 회원들의 재등록 의사에 직접적인 영향을 미치게 되며, 만족을 느낀 회원이 재등록을 한다면 투자없이 지속적인 산출을 얻는 것이라 피트니스 센터에서는 큰 이익을 확보하게 된다(백진우, 2005).

최근 기업들은 인공기법을 활용하여 고객의 활동유형 등 측정 가능한 변수로 계량화하는 방법을 적용하여 고객 관리에 이용하고 있다. 축적된 데이터베이스로부터 유용한 수요 패턴을 찾아내고 고객 수요 예측을 통한 데이터

마이닝 활용이 필요하다. 고객 재등록(이탈)을 사전에 예측하고 분석하여 경쟁이 심한 경영 환경 속에서 신속한 마케팅전략을 수립할 수 있는 기초적인 정보수단을 제공할 수 있다(김민창, 2023).

2.2.2 재등록 예측 인공지능 알고리즘

의사결정나무(Decision Tree) 분석은 데이터 분석 및 예측 모델링에 중요한 역할을 하는 알고리즘으로 계층적 구조를 가진 트리 형태의 그래프로 표현된다(왕반, 2023). 의사결정 규칙(Decision Rule)에 있어서 일반적인 나무(tree) 구조를 도표화하여 분류(Classification) 및 예측(Prediction)을 수행하는 방법이다(이희선, 2011). 분류(Classification) 및 예측(Prediction) 과정이 나무 구조와 유사한 방법으로 추론에 의한 규칙(Induction Rule)에 따라 표현되어 다른 방법론인 인공신경망, 판별분석 등에 비해 쉬운 이해와 설명력이 높은 장점이 있다(박지훈, 2020).

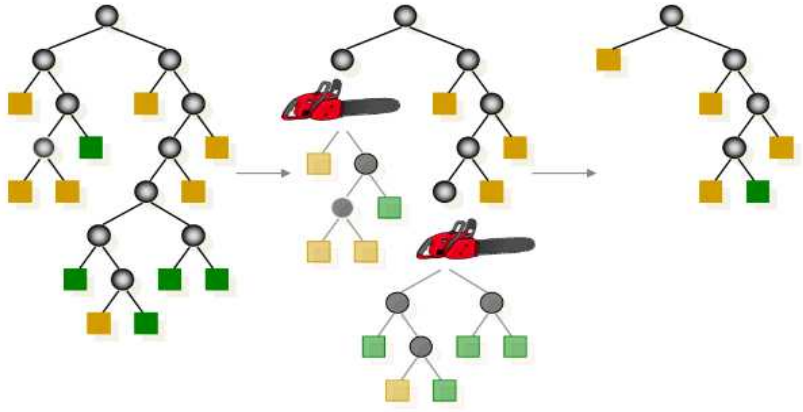
의사결정나무는 기존에 존재하는 데이터의 범주 관계를 파악하고, 새롭게 관측된 데이터의 범주를 스스로 판별하는 과정인 분류 학습과 레이블링된 값(출력값)이 연속형 데이터인 경우 학습한 모델은 입력값과 출력값 간의 일반적인 관계 특성을 도출하는 회귀 학습에 이용되고 있다(장현수, 2021).

의사결정나무 모델은 하나의 나무구조를 이루고 있으며, 마디(node)라고 불리는 구성요소들로 이루어져 있는 데, 나무구조가 시작되는 마디로서 분석 대상이 되는 전체 자료를 뿌리 마디(root node), 하나의 마디로부터 분리되어 나간 2개 이상의 마디를 자식 마디(child node), 자식 마디의 상위 마디를 부모 마디(parent node)라고 한다(이희선, 2011). 각 노드(node)마다 속성들을 테스트하는 의사결정 흐름도(decision flowchart) 또는 뒤집어진 나무(inverted tree) 형식을 취한다. 의사결정나무 경로의 맨 끝은 잎 노드(leaf node)라 하는 데, 의사결정 경로에 따라 구성된 조건들을 기반으로 타겟 변수를 예측할 수 있다.

노드는 데이터셋을 여러 개의 부분 집합(subset)으로 나눈다. 의사결정나무의 기본 아이디어는 데이터셋을 데이터의 순도/동질성(homogeneity)을 기준

으로 분할하는 것이다. 의사결정나무는 구분 뒤 각 영역의 순도(homogeneity) 증가, 불순도(impurity)/불확실성(uncertainty)이 최대한 감소하도록 하는 방향으로 학습을 진행한다.

의사결정나무가 고려해야 할 부분이 가지치기(pruning)인데, 의사결정나무 가지 수가 증가할 때 새로운 데이터에 대한 오분류율이 감소하나 일정 수준 이상이 되면 오분류율이 되레 증가하는 현상이 발생한다. 이러한 문제를 해결하기 위해 검증데이터에 대한 오분류율이 증가하는 시점에서 적절히 가지치기를 수행해야 한다. 마치 나뭇가지를 잘라내는 것과 의미에서 가지치기라는 명칭이 붙었는데 데이터를 버리는 것이 아니라 가지를 합하는 개념이다.



[그림 2-2] 의사결정나무 가지치기 개념

의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나, 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다(이희선, 2011). 의사결정나무는 직관적으로 이해하기 쉽기 때문에 설명이 필요한 경우에 많이 이용된다. 가령 병원에서 환자가 어떤 병에 걸렸을 때 이러이러한 조건에 부합하여 특정 질병이 의심된다고 설명할 때 손쉽게 이해할 수 있어 활용할 수 있다.

하지만, 의사결정나무의 가장 큰 단점은 과적합으로 정확도가 떨어지는 것이다. 과적합(overfitting)은 학습 데이터에 대한 정확한 예측을 제공하지만,

새로운 데이터에 대해서는 예측력이 떨어지는 바람직하지 못한 인공지능 학습 동작이다. 예측을 수행할 때 먼저 알려진 데이트 세트에 대해 모델을 훈련시키고, 이 정보를 기반으로 새 데이터셋에 대한 결과를 예측하는 데, 과적합 모형은 부정확한 예측을 제공할 수 있다. 과적합이 생기는 주된 이유는 다음과 같이 정리된다.

첫째, 훈련 데이터 크기가 너무 작고 가능한 모든 입력 데이터 값을 정확하게 나타나기에 충분한 데이터 샘플을 포함하지 않았다.

둘째, 훈련 데이터에 노이즈 데이터라고 하는 관련 없는 정보가 많이 포함되었다.

셋째, 모델이 단일 샘플 데이트셋에서 너무 오래 훈련하였다.

끝으로, 모델 복잡도가 높기 때문에 훈련 데이터 내 노이즈를 학습하였다. 이러한 과적합을 이유로 의사결정나무를 보완한 랜덤 포레스트 방법이 개발되었다.

랜덤 포레스트(Random Forest)는 다수의 의사결정나무를 이용해 얻어진 결과를 투표에 부쳐 결론을 내는 앙상블(Ensemble) 방법론으로 여러 개의 분류기를 만들어서 배깅(Bagging)으로 최종 결정하는 알고리즘이다(김민창, 2023). 랜덤 포레스트는 각기 다른 방향으로 과적합된 트리들을 묶음으로 평균을 내어 예측 정확성은 향상하면서 과적합을 줄일 수 있는데, 개별 결정 트리들을 많이 만들어야 하며 각각의 결정 트리는 다른 트리와 구별되는 특성을 가져야 한다(왕반, 2023).

앙상블 학습(Ensemble Learning)은 여러 개의 분류기(Classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법이다. 랜덤 포레스트는 동일한 데이터로부터 무작위 복원추출을 수행하여 각각의 하위 데이터셋에 의사결정나무를 적용하는 방식으로 동작하는 데, 일부의 나무가 특정 데이터셋에 과적합되더라도 나무들 사이의 투표를 통해 결과적으로 숲 차원에서 올바른 결정이 이루어진다(김민창, 2023).

어려운 문제의 결론을 내기 위해 여러 명의 전문가로 위원회를 구성해 다양한 의견을 수렴하고 결정하듯이 앙상블 학습은 다양한 분류기의 예측 결과를 결합함으로써 단일 분류기보다 신뢰성이 높은 예측값을 얻는다.

앙상블 학습의 유형은 크게 보팅(Voting), 배깅(Bagging), 부스팅(Boosting)의 3가지로 나눌 수 있다. 보팅과 배깅은 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식으로, 보팅은 서로 다른 알고리즘을 가진 분류기를 결합하는 것이고, 배깅은 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만, 데이터 샘플링을 서로 다르게 가져가는 학습을 하면서 보팅을 수행한다. 랜덤 포레스트가 대표적인 배깅 방식이다. 부스팅은 여러 개의 분류기가 순차적으로 학습을 수행하되, 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 다음 분류기에는 가중치(weight)를 부여하면서 학습과 예측을 진행한다.

배깅은 전체 데이터셋으로 부터 무작위 복원추출을 통해 T개의 학습용 하위 데이터셋을 생성하여 각각의 학습 데이터셋에 대응하는 T개의 의사결정나무를 훈련시킨다. 이후 학습이 완료된 의사결정나무를 결합하여 다수결 또는 평균 방식으로 투표를 진행한다(김민창, 2023). 랜덤 포레스트를 구성하는 각각의 의사결정나무는 무작위로 선택된 독립변수 집합에서 지니 불순도(Gini impurity)를 최소화하도록 독립변수를 선택해가며 구축하며, 각 의사결정나무는 Bootstrap sample로부터 구축되며, 분류를 수행하기 위해 각 트리는 특정 class에 투표하고 가장 많은 득표한 class로 분류한다(김경태, 이지형, 2018).

랜덤 포레스트의 기반 알고리즘은 결정트리(의사결정나무)로서, 의사결정나무의 쉽고 직관적인 장점을 그대로 가지고 있다. 랜덤 포레스트는 개별 트리가 학습하는 데이터셋은 전체 데이터에서 일부가 중첩되게 샘플링된 데이터셋이다. 여러 개의 데이터셋을 중첩되게 분리하는 것이 부트스트래핑(bootstrapping) 분할 방식이다.

배깅(bagging)은 bootstrap aggregating의 줄임말이다. 배깅 기법은 데이터에 무작위성을 부여하여 표본을 추출하고 의사결정나무 알고리즘과 같이 여러 차례 테스트를 통해 예측값을 도출하기 위해 연산을 반복하는 데, 이는 여러 개의 의사결정나무 분류기를 훈련시킨 후 그 예측값을 도출하고 학습 데이터 구성할 때 데이터 중복을 허용하는 특징을 가지며, 의사결정나무의 한계점인 과적합을 제어하여 의사결정나무보다 높은 성능을 보인다(장기만, 2020). 랜덤 포레스트의 특징은 Out-Of-Bag(OOB) 에러를 이용하는데,

OOB 에러를 통해 일반화 정도를 측정하며 독립변수의 중요도를 계산할 수 있다(김경태, 이지형, 2018).

랜덤 포레스트의 장점은 의사결정나무와 마찬가지로 아웃라이어에 거의 영향을 받지 않고 과적합이 잘 되지 않는다. 하지만, 데이터 수가 많아지면 의사결정나무에 비해 학습 속도가 상대적으로 느린 편이고, 개별 트리 분석이 어렵고 트리 분리가 복잡해지는 경향이 존재한다. 재등록을 예측하는 방법으로 인공지능 알고리즘을 주로 사용하는 데 Random Forest, XGB가 우수한 예측 성능을 보여준다.

극한 경사도 부스팅 XGB(Extreme Gradient Boosting)는 부스팅(Boosting) 기법을 이용하여 병렬 학습이 지원되도록 구현한 알고리즘이다(김민창, 2023). XGB는 GBM(Gradient Boosting Model)을 기반으로 여러 약한 학습기를 결합하여 강력한 예측모델을 구축하는데, 모델 구축 및 최적화 등에서 혁신과 개선이 이루어져 성능과 확장성이 향상되었다(왕반, 2023). 캐글 경연대회(Kaggle Contest)에서 상위를 차지한 데이터 과학자가 XGB를 이용하면서 널리 알려져 있다. XGB 이후에도 다양한 부스팅 모델이 개발되었지만 가장 인기 있는 모델로 다양한 참고자료가 나와 있다.

XGB는 다른 GBM과 비교했을 때 과적합을 방지하기 위해 더 많은 정규화 항을 사용하는 특징으로 더 좋은 성능을 보여준다(김경태, 이지형, 2018). 우선 하나의 트리에 의해 분류가 이루어지고 나면, 오분류된 사례에 더 큰 가중치를 부여하며, 다음에 만들어지는 트리는 큰 가중치가 적용된 오분류 사례에 집중하여 해당 사례를 더욱 잘 분류되게 한다(이재준, 2021).

부스팅은 약한 분류기를 순차적으로 연결하여 강한 분류기를 만드는 방법으로 A 분류기를 만든 후 결과를 예측하고, 그 정보를 바탕으로 B 분류기를 만든다. 다시 예측을 진행하고, 그 정보를 바탕으로 C 분류기를 만들어 이렇게 만들어진 분류기를 모두 결합하여 최종 모델이 생성된다.

주된 장점은 예측 속도가 상당히 빠르고 예측 성능이 좋으며, 변수 종류가 많고 데이터가 클수록 상대적으로 뛰어난 성능을 보인다. 지정된 반복 횟수가 아니라 교차검증을 통해 평가 데이터셋의 평가값이 최적화되면 반복을 중간에 멈출 수 있는 조기 중단(Early Stopping) 기능이 있다. 하지만, 복잡한 모

델인 만큼 해석이 어렵고, 더 나은 성능을 위한 하이퍼 파라미터 튜닝이 까다로운 단점이 존재한다.

2.3.3 재등록 예측에 관한 선행연구

재등록 예측은 재등록 또는 이탈에 대한 예측 모형으로 인과관계를 살피거나 이탈을 방지하고 재등록을 위한 전략을 제시할 수 있는 데, 연구방법론에서 가설검증을 통한 실증연구보다는 인공지능 기법 등을 이용한 예측 연구가 주가 된다. 인공지능 알고리즘을 결합한 모델로 높은 정확도의 예측 모형을 구축하며, 데이터의 불균형 문제를 개선하는 절차적인 차원의 개선을 이루고 있다(오세준, 2018). 인공지능 분석을 통해 고객의 구매 패턴 등을 파악하여 조기에 이탈하는 고객을 예측하며 효과적인 마케팅 전략을 세울 수 있다(왕반, 2023).

이탈예측모델은 [표 2-2]와 같이 목표로 하는 예측 대상에 따라 Sentiment Model(감정 모델), Event Model(사건 모델), Survival Model(생존 모델), Treatment Model(처치 모델)의 4가지 유형으로 나누며, 데이터셋의 특징과 비즈니스 목표에 따라 해당 알고리즘을 활용한다(김민창, 2023).

[표 2-2] 이탈예측모델 유형

구분	목표	적용 알고리즘
Sentiment Model (감정 모델)	이탈을 촉진시킨 주요 요인(이탈 사유)을 식별하고 이탈 신호를 감지	텍스트나 음성과 같은 비정형 데이터를 활용하여 자연어 처리(Natural Language Processing) 모델
Event Model (사건 모델)	단기적인 이탈 위험을 예측	Random Forest, Gradient Boosting 머신러닝 방법론 RNN, 트랜스포머
Survival Model (생존 모델)	장기적인 이탈 위험을 예측	Kaplan-Meier 추정법 Cox 비례위험모형
Treatment Model (처치 모델)	예측된 이탈 대상에게 가장 효과적인 이탈 방지	인과추론 (Causal Inference)

자료 : 김민창(2023)

이동통신의 통화 품질과 부가서비스, 고객 대응 서비스에서 오는 전반적인 서비스품질에 대한 만족도가 높을수록 전환장벽이 높아져 고객 유지에 긍정적인 영향을 미친다(남수태, 진찬용, 2013). 기존의 서비스를 이용하면서 서비스 제공자와의 지속적인 관계유지 가치에 대한 인식이 서비스 대체나 혹은 전환에 따른 금전적 이익보다 앞선다(오영삼, 2019). 통신영역에서는 시장 포화로 이동통신 시장이 둔화되고 최신 스마트폰의 경쟁적 출시로 고객의 약정 탈퇴가 발생하여 제조자들은 신뢰할 수 있는 고객 정보를 수집하고 보다 정확한 고객 이탈 예측모델을 구축하여 고객의 결정에 영향을 미치는 핵심 요인을 파악하는 것이 중요하다(Ullah et al., 2019).

금융권에서는 은행, 증권, 신용카드 업계를 중심으로 고객의 과거 금융 거래 내역을 분석하여 미래 형태를 예측하고 다양한 고객 관계 관리 활동에 활용하는 데, 고객 세분화 및 타겟 마케팅에 이용하고 있다(나광택 외, 2020).

게임 분야에서는 수익이 높은 유저들의 이탈 예측기법 및 이탈 예측 프레임워크를 제안하며, 이탈과 높은 상관관계가 있는 게임 활동의 세부 변수들을 도출하여 예측력을 높이는 연구가 진행되고 있다. 게임 로그 데이터를 이용하여 고객 이탈을 예측하는 딥러닝 기반의 예측모델과 고객 행동 변화와 통계적 특징을 활용하는 의사결정나무 기반의 예측모델이 이용되었다(김경태, 이지형, 2018).

태권도장에서 향상된 서비스 가치 인식을 높여줌으로써 태권도 수련생과의 우호적 관계를 형성하여 도장 이용을 증가하여 궁극적으로 재등록으로 유도할 수 있다(김용운, 2022). 피트니스 시설에 있어 정신적 또는 정서적 선호도가 높을수록 지속적인 이용과 재등록에 이어짐으로, 강사와의 관계를 통해 긍정적인 정서 구축이 재구매를 유도할 수 있다(고몽선, 2023).

재등록 의도를 높이기 위해 서비스품질의 모든 요인인 유형성, 신뢰성, 반응성, 확신성, 공감성 등의 중요성을 인지하여 지속적인 관리가 필요하다(안재혁, 2023). 피트니스 산업의 안정적인 경영 전략 및 장기적인 수익 증대를 위해 피트니스 회원의 서비스품질과 재등록 의도와의 관계를 파악하는 것이 중요하다(유서정, 2022).

재등록 의도는 주로 소비자 행동의 연구에서 고객 만족의 결과 변수로 논

의되고 있으며, 사회심리학에서의 사회적 교환이론에 따르면 관계를 지속시키기 위한 관계유지의 개념으로 이해되고 있다(서주환, 2014). 고객들의 재구매 의도는 1회 구매 고객이 아닌 고정 고객화할 수 있는 가능성이 높기 때문에 재구매 의도는 장기적인 수요 창출에 있어 기업 경영에 큰 도움이 된다(송종태, 2004). 새로운 고객을 유치하는 것이 기존 고객을 유지하는 것보다 비용이 5배나 더 들게 되어 소비자의 재등록 의도는 피트니스 업계에 중요한 부분으로 자리잡고 있다(이재현, 2007).

고객 이탈예측은 최적의 모델 선택이 필요하며 가능한 많은 고객을 유지하기 위해 리텐션 마케팅이 이어져야 하는 데, 고객가치 등을 고려하여 최대한 많은 수익을 낼 수 있도록 효과적인 마케팅 비용 지출 전략이 중요하다(김재엽, 2020).

만족의 정도에 따라 고객 충성도가 증가하여 관성적인 재구매가 발생하며, 소비자에게 만족도를 높여 재구매의도를 높여야 한다(김용운, 2022).

[표 2-3] 재구매 의사에 영향을 미치는 요인

구분	연구내용	연구자
서비스품질, 고객만족 →재구매	서비스품질이 고객만족에 긍정적인 영향을 미치고 고객만족은 구매의도에 긍정적인 영향을 줌을 증명	Cronin, Taylor (1992)
고객만족 →재구매	고객만족은 소비자의 태도에 긍정적인 영향을 주며, 고객만족의 정도가 높을수록 재구매의도 역시 증가	Oliver (1993)
	전반적인 서비스품질과 고객만족의 상호작용이 구매의도에 영향을 미침	John.Suh (2000)
	고객만족이 재구매의도에 긍정적인 영향을 미치는 것으로 확인	공지우(2008)
서비스품질, 회사신뢰 →재구매	서비스품질의 일부인 유지관리 요인, 보험판매원 요인, 우수고객 요인이 재구매의도에 영향을 미치는 것으로 나타났으며 이러한 신뢰는 재구매의도에 긍정적인 영향을 미치는 것으로 확인함	박균성(2016)
서비스 품질 →재구매	다른 고객에게 제품을 추천하고 구매하도록 긍정적으로 설득하여 긍정적인 추천은 기업이 장기적인 성공에 매우 중요한 역할을 수행함	김종택, 김석영 (2010)

자료 : 김용운(2022)

고객 이탈을 예측하는 것은 고객 유지에 중요하며 많은 산업에서 막대한 손실을 방지하는 데 필수적으로, 현재 다양한 영역에서 고객 이탈을 예측하고 예방하기 위해 데이터 마이닝 및 기계학습 기술이 사용되고 있다(Ahn et al., 2020). 첫째, 고객 유지를 가장 직접적인 중요한 이슈로 다루는 마케팅 분야의 방법론은 단순한 RFM(Recency, Frequency, Monetary) 모델로부터 고객 이탈을 예측하는 가장 좋은 방법을 찾는 데 중점을 둔 랜덤 포레스트와 같은 앙상블 머신러닝 방법에 이르기까지 많은 연구가 있다(Mirkovic et al., 2022). 고객 이탈예측을 위해 클러스터링 및 분류 알고리즘을 통합한 앙상블 학습 기법을 이용하는데, 분류 모델의 선택은 다른 분류 모델과 상관관계가 있거나 성능이 잘 나오지 않는 분류 모델을 제거하여 성능을 향상시킬 수 있다(Wang et al., 2022). 기계학습모델에서 피처 선택(feature selection)은 분류 정확도를 향상시키는 데 중요한 역할을 하며, 데이터셋의 차원을 줄이기 위해 GSA(Gravitational Search Algorithm)를 적용하기도 한다(서영정, 2023).

프로세스 마이닝은 실제 로그 데이터를 기반으로 기업의 실제 프로세스를 파악하고 경영 통찰력을 제시하는 데이터 분석이며, 고객의 전산 기록을 분석하여 고객이 언제 어떻게 어느 시점에서 이탈하는지를 '고객 여정'으로 시각화하여 고객 이탈을 파악할 수 있다(이식, 2023).

고객 이탈 방지에 대한 분석의 궁극적인 목적은 이탈 원인에 대한 분석과 이를 활용한 이탈 예상 고객군에 대한 활발한 관리 활동을 통하여 최종적으로 이탈고객을 최소화하는 데에 있다(배행수, 2005). 고객이 상품이나 서비스에 대해 긍정적인 생각을 가지고 지속적으로 이용할지 아니면 부정적인 생각으로 다른 상품이나 서비스로 전환할지에 대해 행동으로 표현하므로 사전에 적절히 파악하는 활동이 필요하다(김두일, 2018). 고객 정보 분석을 통하여 이탈 가능한 고객을 예측하고, 고객의 특성을 파악하여 고객관계관리(CRM) 수행에 체계적 활용할 수 있다(이훈영, 2002).

기계학습 방법론 중에서도 가장 널리 사용되는 알고리즘은 의사결정나무(DT: Decision Tree), 로지스틱 회귀(LR : Logistic Regression), 랜덤 포레스트(RF : Random Forest), 서포트벡터머신 (SVM : Support Vector

Machine) 등이다. 로지스틱 회귀와 의사결정나무, 의사결정나무를 여러 개 쌓은 랜덤 포레스트는 이진 분류 문제를 풀기 위해 베이스라인 모델로서 단순하면서 좋은 성능을 보여준다(Neslin & Charilotte, 2006). 고객 이탈 데이터의 특성상 대규모, 고차원, 비선형성, 비정규성, 시계열을 포함할 경우 모델의 구조적 위험을 최소화하면서 일반화 오류의 상한을 최대화하여 비선형성의 문제에도 강건한 서포트벡터머신(SVM)을 고객 이탈 모델로 도입했다(Xia & Jin, 2008). 로지스틱 회귀모형이 다중공선성이 있는 변수들로는 제대로 작동하지 않아 서로 상관성이 높은 변수를 제거하고 남은 변수들을 XGB 모델로 학습시키는 로지스틱 회귀-XGB 결합 모델로 결과를 개선하였다(Li & Li, 2019).

피트니스 분야 재등록 예측과 관련한 주요 선행연구를 정리하면 다음 [표 2-4]와 같다.

[표 2-4] 피트니스 분야 재등록 예측 관련 선행연구

연구자 (연도)	논문명	연구내용	적용 분야
남궁진 외 (2016)	확장된 목표지향적 행동모형을 적용한 피트니스센터 참여자들의 재등록 의도 분석	피트니스 센터 재등록에 영향을 주는 심리적 요인과 열망과의 관계를 목표지향적 행동 모형을 통해 파악	피트니스
정현아 (2018)	필라테스 퍼스널 트레이닝 스튜디오의 고객 만족, 재구매 및 추천의사에 관한 실증연구	고객층을 세분화하여 집단에 최적화된 다양한 특화 프로그램을 개발하여 고객의 기대수준에 충족	필라테스
최한나 (2019)	필라테스 회원이 지각하는 지도자의 전문성이 지도효율성 및 재등록의도에 미치는 영향	필라테스 지도자의 전문성은 지도효율성과 재등록의도에 영향을 주므로, 지도자 전문성 강화에 주력해야 됨	필라테스
이현국 (2020)	태권도장 주말 프로그램이 학부모만족 및 재등록의도에 미치는 영향	태권도장에 대한 학부모 만족을 높이기 위해 지도자 역량, 프로그램 구성 수준 등을 높일 필요가 있음	태권도
김용운 (2022)	초등학생의 태권도장 서비스 가치 인식이 재등록의사에 미치는 영향	태권도장에서 향상된 서비스 가치 인식을 높여줌으로써 수련생과의 우호적 관계로 재등록을 유도	태권도

강민지 (2023)	성인발레학원의 수업만족도 및 신체적 자기효능감이 재등록의사에 미치는 영향	수업만족도가 좋아지면 재등록 의사가 긍정적이며, 성인 발레의 지속 유지 전략을 추진	발레
고몽선 (2023)	선양시 태권도 지도자의 지도유형이 수련생의 훈련효과와 재등록의도에 미치는 영향	태권도 지도자의 지도유형이 재등록에 미치는 영향을 분석하며, 정신적 또는 정서적 선호도가 높을수록 재구매에 영향을 줌	태권도
소호진 (2023)	무도체육관 수련생의 선택속성이 수련만족과 재등록의도에 미치는 영향	무도체육관의 특화된 프로그램이나 교육시간 등 수련생이 만족할 수 있는 선택속성이 재등록 의도와의 관계를 파악	무도
지준철, 장세용 (2023)	골프아카데미 지도자의 전문성이 재등록 의도에 미치는 영향	골프지도자와 참여자의 관계를 고객 충성도를 매개하여 재등록 의도에 미치는 영향을 검증	골프

자료 : 선행연구를 토대로 연구자 재작성

피트니스 이외 분야의 재등록 예측과 관련한 주요 선행연구를 정리하면 다음 [표 2-5]와 같다.

[표 2-5] 피트니스 이외 분야의 재등록 예측 관련 선행연구

연구자 (연도)	논문명	연구내용	적용 분야
Mozer et al. (2000)	Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry	통신회사 고객 정보를 활용한 로지스틱 회귀, 의사결정나무, 인공신경망 및 여타 다른 부스팅 기법적용, 고객 불만족 정도 예측과 고객 유지 개선 방법을 제시	통신 회사
Hung et al. (2006)	Applying data mining to telecom churn management	통신회사 이용고객 세분화를 위한 신경망과 의사결정나무를 적용한 고객 이탈을 예측	통신 회사
Anil & Ravi (2008)	Predicting credit card customer churn in banks using data mining	신용카드 고객 이탈 예측을 위해 인공신경망, Random Forest, 의사결정나무, SVM 기반의 앙상블 예측	금융
Tang et al. (2014)	Assessing the impact of derived behavior information on customer attrition in the financial service industry	파생된 행동 정보를 사용하여 금융 서비스 산업의 고객이탈 모델 개선	금융
김경태, 이지형 (2018)	딥 러닝과 Boosted Decision Tree를 활용한 고객 이탈 예측 모델	게임 로그 데이터를 이용하여 고객 이탈을 예측하는 딥러닝 기반의 예측모델과 고객 행동 변화와 통계적 특징을 활용하는 의사결정나무 기반의 예측 모델 제안	게임

이정우 (2019)	데이터 마이닝과 AHP를 이용한 고객 이탈 분석에 관한 연구	카드사 이용고객 이탈분석을 위해 로지스틱 회귀모델, 의사결정나무 모델, 인공 신경망 모델의 성능을 비교하여 인공 신경망 모델의 적합성 확인	금융
윤성훈 (2019)	E-커머스의 차별화마케팅이 소비자만족과 재구매의도 및 경영성과에 미치는 영향	차별화 마케팅 전략이 소비자 만족과 재구매의도에 긍정적 영향을 미치며, 마케팅에 실제 적용된 효과를 검증	전자 상거래
장기만 (2020)	VOC(Voice of Customer) 빅데이터를 활용한 신용카드 사용자 이탈예측 모델	정형 데이터, 비정형 데이터를 분석하여 신용카드 사용자 이탈예측 프로세스를 생성하며, VOC 기반 경영으로 고객가치 차별화에 기여	금융
추희정 (2022)	온라인 쇼핑몰 고객의 재방문과 재구매 및 이탈 요인 연구	웹로그 데이터를 활용하여 생존분석 방법으로 재구매에 영향을 주는 변수를 선별	전자 상거래
이식 (2023)	고객이탈 예측을 위한 프로세스마이닝과 데이터마이닝 기반의 하이브리드 모델 개발	분석자의 도메인 지식으로 이탈고객을 정밀 분석할 수 있는 프로세스 마이닝을 적용하여 이탈예측을 구축하고 이탈 방지 전략 수립	교육
김민창 (2023)	행동 시퀀스 트랜스포머를 이용한 구독 비즈니스의 고객 이탈 예측	정기 납입자의 미납 가능성을 예측한 Churn Behavior Sequence Transformer 모델을 제안하며, 과적합을 해결하고 효과적인 예측 수행	전자 상거래

자료 : 선행연구를 토대로 연구자 재작성

2.3 피트니스 센터 마케팅

2.3.1 피트니스 산업

피트니스(Fitness)는 건강의 유지, 증진을 목적으로 한 운동을 총칭한다. 피트니스가 주로 헬스클럽에서 이루어지다 보니 '헬스클럽'이라는 인식이 강하지만, 헬스클럽에 국한하지 않고 관련 산업으로의 포괄적인 접근이 필요하다. 1960년대에 소개된 "에어로빅스(Aerobics)"가 대표적인 유산소 운동으로 일반인들에게 피트니스 운동이 본격적으로 보급되었다. 2000년 초반에는 근력운동과 심폐운동 및 고강도 운동이 유행하게 되었고, 몸짱 열풍으로 헬스클럽 이용이 증가되었다. 2010년에는 "Wellness" 개념이 대중화되면서 식습관 개선 등 라이프 스타일을 고려한 요가 등이 유행하였다. 2020년대에는 코로나 19 영향으로 오프라인 다중이용시설 이용 인원이 제한되고, 홈 트레이닝(Home Training) 열풍이 불기 시작하였다. 하지만, 운동 습관이 제대로 형성되지 못한 나홀로 홈 트레이닝의 경우 자칫 운동 동기가 상실되고, 잘못된 동작으로 부상 위험이 있어 기구 필라테스와 같은 소규모 그룹 운동(GX : Group eXercise)이 최근 트렌드가 되고 있다.

피트니스 센터의 고객 신뢰는 서비스 제공자에 대한 의존도를 높이고 고객 충성도를 유발시킬 수 있어 신뢰감 형성은 피트니스 시설 경쟁력을 높이기 위한 핵심 과제이다. 충성도의 대표적인 선행변수로서 고객 만족은 고객의 욕구나 기대를 충족시키는 데 중점을 두며, 현재 관계를 유지하기 위해 노력하는 의지라고 볼 수 있다.

피트니스 관여도는 고객이 특정 피트니스에 대해 느끼는 개인적 관심과 중요성으로 회원들의 판단, 의사결정을 변화시키는 핵심요소가 될 수 있다. 관여도가 높은 고객들은 서비스품질 요소 자체에 더 관심을 집중하게 되고, 관여도가 낮은 고객들보다 더 쉽게 만족할 가능성이 높다. 반면, 관여도가 낮은 고객들은 서비스 경험에 평가시 접근성, 시설, 홍보 활동 등 외부적 요소에 더 민감하게 반응할 수 있다(후앙이한, 2022).

2.3.2 피트니스 센터의 타겟 마케팅 개요

타겟 마케팅(Target Marketing)은 제품 또는 서비스를 판매하고자 하는 대상이 누구인지를 정하고 그들을 대상으로 마케팅 활동을 전개하는 것을 말한다(김담희, 안가경, 2018).

타겟 마케팅을 하기 전에 고객 세분화를 진행하는 데 고객 세분화의 장점은 다음과 같다.

첫째, 고객 세분화를 통해 목표 고객 그룹을 발견함으로써 적절한 마케팅을 배분하여 저비용과 고효율성 효과를 극대화할 수 있다.

둘째, 각 고객 그룹에 대한 니즈 파악과 분석을 할 수 있기 때문에 마케팅 활동에 대한 기본적인 자료가 되며, 마케팅 성과에 대한 전략과 피드백을 제공한다. 이러한 피드백을 통해 강점과 약점을 파악하고 고객 계층을 판단할 수 있다.

마지막으로 고객 세분화를 통해 각 고객 그룹에 적절한 전략과 성과에 대한 피드백으로 향후 고객들의 니즈에 대하여 먼저 적극적으로 어필하고 대응할 수 있게 된다(김담희, 안가경, 2018). 고객 세분화할 때 인구통계적, 구매행동, 심리학적 요인이 반영되는 데, 최근에는 구매 행동별 분석을 통해 고객 취향에 맞는 맞춤식 전략으로 이용된다(이찬도, 2020).

과거의 마케팅전략은 불특정 다수의 고객을 상대로 전단, 각종 매체 등을 통해 대대적인 광고 및 판촉 활동을 실시하여 매출 효과를 기대하는 매스 마케팅(Mass Marketing) 중심이었다. 하지만, 매스 마케팅 시행의 문제점이 부각되었다. 마케팅전략의 부재, 판촉 전략의 효율적 감소, 다양한 관점의 즉각적이고 직접적인 마케팅 분석 불가능, 내부직원들의 적극성 부족, 고객의 니즈에 대한 구체적인 대응이 부족하였다. 이러한 문제점을 해결하기 위해 고객 중심의 정보를 분석하여 선별된 고객의 가치를 극대화하여 긴밀한 관계유지로 고객 만족을 극대화하고자 하였다.

불특정 다수에서 특정 회원에 맞춘 타겟 마케팅이나 개성 있고 차별화된 고객 관계유지, 이탈 방지, 보상 마케팅이 진행될 수 있다. 전통적인 마케팅은 마케팅믹스(제품, 가격, 장소, 프로모션)을 통해 제품의 특징에 주안점을 두는

단기적인 시간의 틀(timeframe)을 가졌다면, 최근 마케팅은 '고객'에 관점을 두고 첫 번째로 고객을 파악(customer identification)하고 두 번째로 고객을 구분(customer differentiation)하고 세 번째로 고객과의 상호작용(customer interaction)을 통해 마지막으로 개별 고객들의 요구(needs)에 적합한 제품과 서비스를 제공하게 된다.

시장의 관점을 거시 시장(macro market)이 아닌 개별소비자(micro market)에 둠으로써 고객 개개인의 특성을 중요시한다. 고객과의 일대일 쌍방향 커뮤니케이션을 통해 고객만족도를 높이고, 지속적인 관계를 형성하여 고객가치를 극대화하게 되었다. 세부 그룹별 니즈 분석을 기반으로 타겟 고객군에 대한 차별화된 전략을 수립함으로써 효과적인 고객기반을 확대하고, 새로운 마케팅 전략을 시도할 수 있다(이희선, 2011).

2.3.3 피트니스 센터의 타겟 마케팅 방안

피트니스를 방문하는 회원들의 운동 목적이 각각 다를 수 있어 획일적인 프로그램으로는 지속적인 방문을 유도하기 어렵다. 본인들이 필요한 프로그램을 사전 안내하여 꾸준히 이용할 수 있도록 프로그램 구축이 요구된다. 카카오톡 프로필 사진 및 인스타그램 등 각종 SNS에 사진을 올리는 경우가 많아 바디 프로필 사진 촬영을 원하는 회원들에게 구체적인 목표를 수립하여 동기부여를 제공하여 프로그램 만족도를 제고할 수 있다.

닌텐도 위핏(wii fit)은 일본 닌텐도에서 개발한 헬스 보조용 소프트웨어 게임으로 요가, 근력 운동, 유산소 운동, 밸런스 게임 등으로 구성되어 코로나 상황에서 대표적인 홈 트레이닝(Home Training)으로 자리매김하였다. 동작인식 센서의 기술과 게임 요소를 접목하였다. 위핏이 가정용으로 보급되었다면, 피트니스 센터에는 위핏보다 3D 고도화된 동작인식 기술을 가진 장비로 동작 분석 및 소모 칼로리, 운동량 등을 분석하여 최적화된 개인 맞춤형 서비스를 제공할 수 있다. 체성분 분석기인 인바디(In-body)가 가정용으로 보급되어 있기에, 심층적인 분석이 필요한 장비로 센터 방문을 유인할 수 있다.

온라인 강의를 통한 홈 트레이닝을 활성화할 수 있다. 코로나-19의 영향

으로 일상생활에서 비대면 활동이 널리 펴지게 되었다. 피트니스 센터 이용이 제한되면서 집에서 하는 홈 트레이닝이 유행하였다. 하지만, 운동 습관이 제대로 형성되지 않은 상태에서 혼자 무리하게 동작을 하면서 부상 위험이 높아 홈 트레이닝이 작심삼일로 끝나는 경우가 비일비재하다. 온라인 강의 시스템을 통해 실시간으로 강사의 동작을 따라 하며 회원들과 소통하는 온라인 수업을 적극적으로 할 수 있다.

시간 마케팅으로 회원의 수업 참여를 확대할 수 있다. 시간 마케팅(Time Marketing)은 가격이나 품질뿐만 아니라 고객의 시간을 아껴 줌으로써 판매 촉진에 기여하는 전략이다. 유사한 사례로 패밀리 레스토랑에서 피크 파임을 피해 한가한 시간대에 생맥주 1+1 서비스를 제공하듯이 공짜라는 심리가 작용하여 매출 상승효과가 나타나기도 한다.

강사 대상 회원 응대 교육을 강화할 수 있다. 회원들과의 접점에 위치한 강사를 중심으로 회원 응대 교육을 강화한다. 피트니스 산업은 대표적인 서비스 산업이며, 회비가 적지 않은 경우가 많아 회원 문의 및 불만 접수가 많은 편이다. 매뉴얼 없이 개인의 감으로만 회원을 응대하기는 문제가 있을 수 있다. 고객과의 대화 스킬과 다양한 상황 응대 능력 향상이 필요하다.

회원 중심의 서비스 디자인 씽킹 접근을 고려할 수 있다. 서비스 디자인 씽킹(Service Design Thinking)은 '서비스 디자인'과 '디자인 씽킹'을 결합한 것으로 디자인 사고의 방법을 활용해 문제를 해결하고 고객의 경험을 최적화하는 과정을 말한다(양연경, 2023). 서비스는 무형이고 시간이 흐르면 존재하지 않지만 사람들은 유형의 증거(Physical Evidence)를 원하여, 서비스가 지닌 무형성을 디자인이 가진 물리적이고 유형적인 특성을 통해 고객에게 구체적인 경험을 제공하는 것이다.

쌍방향 커뮤니케이션 채널을 활용한다. 비대면 업무 방식의 확산으로 고객 경험 중심의 커뮤니케이션 방식을 진행한다. 일방적 메시지 전달이 아닌 고객의 의견을 청취하고 반영할 수 있는 쌍방향 커뮤니케이션 채널이 필요하다. 온라인 채널로 접근과 사용이 용이하며 고객에게 친숙한 디지털 통합 커뮤니케이션 채널을 제공한다. 고객의 개인화 정보를 바탕으로 고객 여정 전체의 매끄러운(seamless) 서비스를 제공한다. 별도 앱을 만들지 않고도 디지털 접점

채널을 소상공인들이 쉽게 접근할 수 있어 설문방식이나 대면 상담의 부담을 줄일 수 있다.

케어 디스크를 통해 맞춤식 관리하여 회원들이 불편한 부위에 대해 별도 이야기하지 않더라도 손쉽게 고지할 수 있다. 피트니스 센터를 방문하는 회원의 대부분이 특정 신체 부위가 불편한 경우 해당 근육 이완이 필요할 때가 많다. 크게 보면 목 어깨, 등, 허리, 고관절, 무릎, 발목 부위가 불편하다고 볼 수 있다.

[그림 2-3] 케어 디스크 종류

예약확인 안내 문자를 발송하여 노쇼를 줄일 수 있다. 노쇼(Noshow, 예약부도)는 예약을 해놓고 해당 일에 나타나지 않는 행동으로 서비스 이용과 관련하여 기차나 비행기, 레스토랑, 호텔, 공연 등에 좌석을 예약한 후 별도의 취소 통보 없이 나타나지 않는 행위이다(남유정, 2020). 예약 시스템은 다양한 분야에 널리 보급되어 손가락 클릭 몇 번으로 손쉽게 예약을 가능하여 소비자들의 예약에 대한 만족도는 점점 상승하지만, 그에 따른 문제점인 노쇼가 발생하여 손실이 발생한다(서미선, 2019). 노쇼는 예약제도를 도입한 곳이라면 언제든지 발생할 수 있으며, 서비스를 공급하는 측에는 추가적 업무에 대한 혼란을 가중시키며, 경제적인 손실을 발생시키고 다른 소비자에게도 예약 기회를 빼앗는 피해를 발생시킨다(Catrili & Veral, 2003).

항공업에서 노쇼는 최대의 관심사로 대두되고 있는 데, 재고를 통해 향후 부가적인 가치를 창출하는 다른 산업들과는 달리 항공업의 재고는 운송 서비스를 의미하며, 수익으로 직접적인 영향을 미치기 때문에 항공기의 빈 좌석을

최소화하는 것이 수익률을 높이는 데 큰 영향을 준다. 항공편 예정편 출발시간 이전에 취소를 하지 않거나, 항공편을 탑승하지 않을 경우 각 항공사별로 규정한 위약금을 징수하고 있다(안혜리, 2011).

피트니스 산업에서의 예약은 서비스 제공자와 미래의 특정 시점에 서비스를 이용하기 위한 회원 간의 공식적인 계약관계로 볼 수 있다. 계약관계를 가진 회원이 피트니스 사업자에게 별도의 취소 통보 등 아무런 조치없이 무단으로 예약을 이행하지 않아 다른 회원들의 이용 기회도 박탈한다는 점에서 경제적 손실로 이어진다.

노쇼의 심리에는 약속을 파기해도 자신에게는 아무런 피해가 없다고 생각하는 '이기심'과 약속에 대한 무책임감, 갑작스럽게 계획을 바꾸는 즉흥적 성향이 엿보인다고 한다. 이러한 노쇼가 증가하는 이유에는 '콜 포비아(Call phobia)' 현상과 관련이 있다. 콜 포비아는 전화 통보를 두려워하여 문자나 이메일로 소통하기 선호하는 현상인데, 어린 시절부터 SNS 등에 익숙한 젊은 층일수록 이런 문제가 생기기 쉽고, 심한 경우 진동 소리만 들어도 놀라면서 식은땀이 나고 두근거리는 신체 증상을 보이기도 한다. 내성적이고 경직된 전화통화 문화 그리고 코로나 19로 인한 비대면 활동 증가로 전화통화에 부담을 가진다. 콜 포비아가 있는 사람은 전화를 걸러 약속을 깨는 일을 불가능한 미션으로 인식한다. 전화통화 자체가 곤욕인데, 전화를 걸어 어려운 이야기까지 하는 것을 부담을 느껴 상황에 대처하는 대신에 피하고 숨어버리는 방법인 '노쇼'를 선택한다. 하지만 노쇼로 인해 피해가 생기는 상황이라 이를 방지하는 노력이 요구된다. 대표적으로 벌점, 이용제한 등의 페널티가 있으며, 장기적으로 볼 때 고객 불만으로 고객 이탈을 초래할 수 있어 회원 스스로 노쇼를 방지할 수 있는 방법이 필요하다.

병원 예약부도 감소를 위해 병원 이용객들이 스스로 예약에 대한 에티켓을 준수할 수 있도록 능동적인 동참의 자세가 중요하다(임지현, 2016). 노쇼 가해 경험이 있는 사람은 추후 이용 시에도 쉽게 반복할 수 있어, 고객 데이터를 DB화하여 지속적으로 관리하여 강력한 노쇼 집단을 사전에 방지해야 한다(천송이, 2013). 어플리케이션 알림을 이용하여 예약안내 전략이 필요하다(강용수, 2017).

포인트 관리를 통해 회원의 참여 동기를 부여할 수 있다. 노쇼시 페널티를 부여하는 제도를 도입하고 있지만, 이에 대해 SNS를 이용하여 악평을 하는 등 왜곡된 소비자의 권리를 주장하는 일이 빈번하고, 회원들에게 부정적 이미지로 고객 만족이 저하되고, 결국 이탈로 이어질 수 있어 널리 활용되지 못하고 있다(남혜정, 2022). 출석에 따라 포인트를 적립하며, 노쇼의 경우 포인트 차감 폭을 확대하여 손실에 대한 심리적 부담을 높이는 것이 효과적이다.

연계 마케팅의 할인 혜택이 도움을 줄 수 있다. 연계 마케팅(Joint Marketing)은 경쟁자가 아닌 같은 지역 내 또는 연관사업자와 공동으로 판촉을 하는 방식이다(이진아, 2007). 비용을 절감하는 동시에 시너지 효과를 발휘할 수 있어 많은 분야에서 즐겨 하는 프로모션이다. 백화점 고객, 카드회사, 호텔 회원고객 등과 연계한 마케팅을 주변에서 쉽게 볼 수 있다. 홍보 기간 동안 각 업체의 상품 및 서비스를 사인물에 노출시켜 고객에게 자연스럽게 홍보할 수 있는 효과를 거둘 수 있다.

재미와 사회적 교류를 통한 참여를 유발한다. 재미와 관련된 오락이 존재하여 피트니스가 산업으로서 자리매김하게 되었다. 프로야구 등 스포츠를 관람한 것 역시 멋진 플레이를 보면 즐기기 위해서 '재미'와 관련된 오락이 내재되어 있기 때문이다. 여가활동에서의 즐거움, 재미, 사교적인 요인이 고객 참여에 영향을 끼칠 수 있다.

2.3.4 피트니스 센터 마케팅 선행연구

피트니스 센터의 고객이 인식하는 관계 마케팅은 재구매에 관계를 주므로, 충성고객 확보를 위해 서비스 품질 관리가 중요하다(최영한, 2018). 피트니스 센터를 이용하는 회원 입장에서는 강사 요인이 센터 애착에 영향을 주어 운동 몰입과 충성도에 영향을 주므로 서비스 산업으로서 관계 형성에 주력해야 한다(박지윤, 2021). 강사와 고객과의 인적 유대관계를 강화하며, 개인별 중요도가 높은 마케팅 요인을 바탕으로 마케팅전략을 추진하여 충성고객으로 전환하여야 한다(하채원, 2023). 높은 수준의 서비스품질을 제공받은 태권도장 수련생은 만족을 통해 해당 태권도장에 충성 수련생으로 남을 수 있어 타 태

권도장보다 경쟁적 우위를 확보하여 태권도장 성장에 중요한 요인이 된다(심우극, 2015).

운동을 함께하는 팀원들의 상호 협력 능력이 신체적, 사회적, 정신적 성장을 도와 운동의 지속 의지에 영향을 미쳐 재등록 의도가 높아진다(유현민, 2018).

세분화된 고객군에 적절한 대응과 제안을 할 수 있게 하고 기업의 특정 마케팅 활동의 대상이 되는 핵심 고객군 선정을 위한 타겟팅이 가능하다. 고객 세분화를 통해 얻을 수 있는 장점을 정리하면 첫째, 핵심고객 집단을 발견함으로써 그에 맞는 마케팅으로 저비용, 고효율성 효과를 극대화할 수 있다. 둘째, 세분화된 고객 집단에 대한 요구를 분석할 수 있어 마케팅 전략을 통해 주된 고객 집단을 판단할 수 있으며, 끝으로 각 집단에 맞는 마케팅 전략과 향후 고객들의 요구에 적극적인 방안을 찾을 수 있다(임지수, 2020).

고객 전체에 대해 동일한 프로모션을 하기보다는 세분화된 집단별 고객 집단의 선정을 통해 고객의 특성에 맞는 맞춤식 전략을 제시할 수 있다고 보고, 충성도가 높아 굳이 프로모션을 제공하지 않더라도 구매가 이루어지는 고객보단 적극적인 프로모션 전략을 펼칠수록 성과가 나올 집단에 보다 적극적인 마케팅이 효과적이다(홍태호, 김은미, 2010).

[표 2-6] 피트니스 센터 마케팅 관련 선행연구

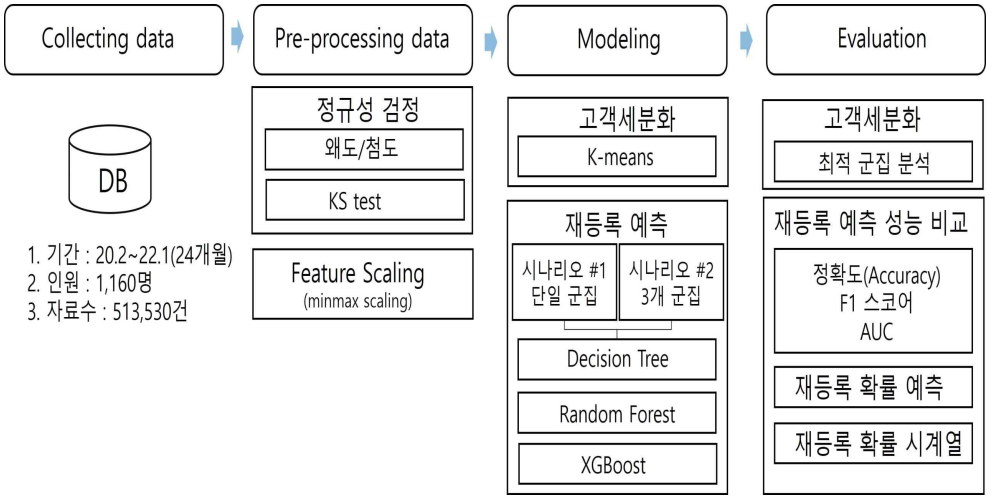
연구자 (연도)	논문명	연구내용
함명수 (2007)	태권도장수련 프로그램 지각이 수련지속 의사에 미치는 영향	수련 만족은 수련 활동을 하는 수련생이 느끼는 욕구의 충족이나 만족에 대한 개인의 인식으로, 태권도 수련프로그램이 수련생의 만족도에 긍정적인 영향을 줌
심우극 (2015)	교육서비스 품질과 서비스 가치, 고객만족, 브랜드 이미지, 자발적 행동 의도 간의 구조적 관계에 관한 연구	높은 수준의 서비스 품질을 제공받은 태권도장 수련생은 충성 수련생으로 태권도장 성장에 중요한 요인이 됨
유현민 (2018)	유소년 스포츠클럽 내 행사참여가 학부모의 만족도와 재등록 의도에 미치는 영향	운동을 함께 하는 동료들과의 교류를 통해 운동의 지속 의지에 영향을 주어 재등록 의도가 높아짐
윤현성 (2021)	피트니스센터의 온라인 마케팅 유형별 마케팅특성이 소비자 태도 및 구매의사 결정요인에 미치는 영향	마케팅특성이 피트니스센터 회원의 소비자 태도에 유의미한 영향을 주어 구매의사결정을 좌우함
정산 (2022)	COVID-19 시대의 피트니스 센터 유형별 이용객의 선택속성 중요도와 만족도 분석	지도자의 참여자 수준에 맞는 지도와 성의있는 지도가 피트니스 센터 선택속성에 중요한 요인으로 파악
하채원 (2023)	스포츠센터 마케팅믹스 요인이 만족도, 추천의도, 재구매의도에 미치는 영향	강사와 고객과의 유대관계를 강화 후 개인별 중요도가 높은 마케팅 요인을 바탕으로 마케팅 전략을 추진

자료 : 선행연구를 토대로 연구자 재작성

Ⅲ. 연구 설계

3.1 연구 모형

본 연구는 2020년 2월부터 2022년 1월에 걸쳐 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 취합 후 인공지능 분석 솔루션인 삼성 SDS Brightics AI v3.9를 이용하여 정규화 전처리를 과정을 진행하였다. K-means를 이용하여 고객 세분화를 진행하여 총 3개의 군집으로 나누었다. 총 3개의 군집에 대해 재등록 예측을 위해 의사결정나무, 랜덤 포레스트, XGB를 진행하였다. 고객 세분화 단계에서 최적 군집을 분석하고 각 군집 내 회원들의 특성을 분석하였다. 세분화하지 않은 전체 집단과 군집별 재등록 예측 성능을 비교하였다. 재등록 여부가 궁금한 회원에 대해 재등록 예측 확률을 계산하며, 시계열 데이터로 개별 회원의 재등록 예측 확률을 모니터링할 수 있었다.



[그림 3-1] 연구모형

3.2 데이터 구성

국내 A 피트니스 프랜차이즈 N 지점을 조사 기간 (2020년 2월 ~ 2022년 1월/24개월)에 걸쳐 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 취합하였다. 재등록 여부를 예측하는 인공지능 알고리즘 성능 비교를 위해 재등록과 이탈 회원 수를 각 580명씩 같은 숫자로 분석하였다. 인공지능 분석 솔루션 삼성SDS Brightics AI v3.9를 이용하였다.

본 연구에 사용하는 변수는 선행 문헌 연구를 통해 도출하였다. 피트니스 재등록 관련 선행연구에서 총 7개 변수 출석, 노쇼, 취소, 선호 강사, 선호 수업, 선호 시간, 선호 요일을 선정하였다. 피트니스 분야 재등록 인공지능 분석 사례가 제한적인 관계로 설문으로 진행한 선행연구에서 필요한 변수를 활용하였다. 선행연구에서 소개된 변수와 피트니스 센터 소규모 운동(GX) 예약 시스템에서 관리하는 항목이 중복된 변수를 선택하였다. 시설은 회원정보가 아니며, 예약 시스템에서는 시설과 관련해서 별도 관리되지 않아 “시설” 변수는 제외하였다.

[표 3-1] 피트니스 재등록 관련 연구별 반영 변수

구분	연구자 (연도)	출석	노쇼	취소	강사	수업	시간	요일	시설
설문	한아름 (2012)				○	○			
	하지선 외 (2014)				○				○
	황영주 (2015)				○	○			○
	김송은 (2016)				○	○	○	○	
	남궁진 외 (2016)				○				
	정현아 (2018)				○	○			○
	최진훈, 박주영 (2019)				○				
	김봉경 (2021)				○	○			○
	서일한 외 (2021)				○				
	하제현, 하재필 (2021)				○				○
	강민지 (2023)				○	○			○
	게세홍 (2023)				○	○			○
	지준철, 장세용 (2023)				○				
	하채원 (2023)	○							○
인공지능	Clavel et al. (2016)	○							
	Aldosary & Alrashdan (2021)	○	○	○					
	Sobreiro et al. (2021)	○							
본 연구		○	○	○	○	○	○	○	

국내 피트니스 분야 재등록 선행연구는 설문(다중회귀분석, 구조방정식)방식으로 진행되어 변수인 강사, 수업 등은 리커트 5점 척도로 집계하였다.

해외 피트니스 분야 재등록의 인공지능 분석에 대한 선행연구는 Clavel et al.(2016), Aldosary & Alrashdan(2021), Sobreio et al.(2021) 등에서 확인되었다. Clavel et al.(2016)은 재등록(이탈) 고객 회원 중에서 수치형 변수인 출석의 전체 횟수를 이용하였다. Alodary & Alrashdan (2021)는 재등록(이탈) 고객 회원 정보 중에서 수치형 변수인 출석, 취소, 노쇼의 전체 횟수를 집계하여 분석하였다. Sobreio et al.(2021)는 재등록(이탈) 고객 회원 중에서 수치형 변수인 출석의 전체 횟수를 집계하여 분석하였다.

[표 3-2] 피트니스 재등록 관련 선행연구

구분	연구자 (연도)	관련 변수	분석방법	결론
설문	한아름 (2012)	강사 출석 수업	구조방정식	강사, 출석, 수업 참가빈도는 재등록 의사에 통계적으로 유의
	하지선 외 (2014)	강사	다중회귀분석	강사는 재등록 의사에 통계적으로 유의
	황영주 (2015)	강사 수업	구조방정식	서비스 품질(강사, 프로그램)은 재등록 의사에 통계적으로 유의
	설민신 외 (2016)	강사	구조방정식	강사는 재등록 의사에 통계적으로 유의
	김송은 (2016)	강사 수업 시간 요일	구조방정식	서비스 품질(강사, 수업, 시간, 요일)은 재등록 의사에 통계적으로 유의
	남궁진 외 (2016)	강사	구조방정식	서비스 품질(강사)은 재등록하려는 회원 태도에 통계적으로 유의

설문	정현아 (2018)	강사 수업	다중회귀분석 / 구조방정식	강사, 프로그램 만족도가 높을수록 재등록 의사에 통계적으로 유의
	최진훈, 박주영 (2019)	강사	구조방정식	강사는 재등록 의사에 통계적으로 유의
	김봉경 (2021)	강사 수업	구조방정식	강사, 수업은 재등록 의도에 통계적으로 유의
	서일한 외 (2021)	강사	다중회귀분석	강사는 재등록 의사에 통계적으로 유의
	하제현, 하재필 (2021)	강사	다중회귀분석	강사는 재등록 의사에 통계적으로 유의
	강민지 (2023)	강사 수업	다중회귀분석	강사, 수업 만족도가 높을수록 재등록 의사가 높아짐
	게세홍 (2023)	강사 수업	다중회귀분석	강사, 수업 만족도가 높을수록 재등록 의사가 높아짐
	지준철, 장세용 (2023)	강사	구조방정식	강사에 대한 충성도가 높게 형성될수록 재등록 의도가 강함
	하채원 (2023)	출석	다중회귀분석	출석 횟수는 재등록 의사에 통계적으로 유의
인공 지능	Clavel et al. (2016)	출석	로지스틱 회귀	로지스틱 회귀분석 결과 정확도 65.6%, 민감도 80.0%로 집계
	Aldosary & Alrashdan (2021)	출석 노쇼 취소	Artificial Neural Network	출석, 노쇼, 취소 변수에 대해 ANN 분석 결과 이탈모형의 정확도 92.1%, 민감도 89.1%, 특이도 93.8%
	Sobreiro et al. (2021)	출석	의사결정나무 랜덤포레스트 XGB	의사결정나무 정확도 0.83 랜덤포레스트 정확도 0.92 XGB 정확도 0.95 XGB 정확도가 가장 우수

범주형 변수인 선호 강사는 범주 개수 8개, 선호 시간은 범주 개수 19개, 선호 수업은 범주 개수 5개, 선호 요일은 범주 개수 6개로 총 4개 범주형 변수에 대해 범주 개수는 38개로 집계된다.

[표 3-3] 범주형 변수 범주

변수	범주 개수	범주
선호 강사 (Teacher)	8	A강사, B강사, C강사, D강사, E강사, F강사, G강사, H강사
선호 시간 (Time)	19	07:00, 08:00, 09:00, 10:00, 11:00, 12:00, 13:00, 14:00, 15:00, 16:00, 17:00, 18:00, 18:30, 19:00, 19:30, 20:00, 20:30, 21:00, 21:30
선호 수업 (Class)	5	Beginner, Basic, Lv.1, Lv1.5, Lv.2
선호 요일 (Day)	6	월요일, 화요일, 수요일, 목요일, 금요일, 토요일

인코딩(Encoding)은 범주형 변수를 컴퓨터가 이해하기 쉬운 숫자 형태(정수형)으로 변환하는 것을 의미한다(이단비, 2020). 인코딩의 방법으로 One-hot encoding, Label encoding, Target encoding, Frequency encoding으로 구분할 수 있다. 인코딩 방법은 탐색적 데이터 분석(EDA)을 통해 데이터를 잘 분석하고, 인공지능 모델링 상황에 맞는 최적의 방법을 찾아야 한다.

주요 인코딩 방법인 One-hot encoding, Label encoding, Target encoding, Frequency encoding 개념 및 특징을 아래 [표 3-4]와 같이 정리하였다.

[표 3-4] 주요 인코딩(Encoding) 개념 및 특징

구분	개념	특징
One-hot encoding	피처 값의 유형에 따라 새로운 피처를 추가해 고유값에 해당하는 칼럼에만 1을 표시하고, 나머지 칼럼에는 0으로 표시	· 고유한 범주의 수가 많을 경우 데이터 세트 크기가 크게 증가 · 피처 수를 증가시켜 '차원의 저주' 문제 발생 가능 · 선형 모델 외에는 적합하지 않음
Label encoding	각 범주에 고유한 숫자값으로 변환	· 특정 순서를 가진 데이터에 잘 작동 · 각 범주에 임의로 정수를 할당하여 의도하지 않은 순서를 의미할 수 있음
Target encoding	범주형 자료값을 '타겟'에 해당하는 변수의 평균값로 변환	· '타겟 변수'를 사용하므로 데이터 누출의 위험이 존재 · 범주별 평균값을 계산하여 복잡함
Frequency encoding	각 범주를 데이트 세트에 표시되는 빈도수로 범주를 분류	· 범주의 빈도수를 처리하여 범주형 데이터를 단순화함 · 범주 수가 많더라도 차원을 그대로 유지한 채 빈도수로 계산하므로, 과적합을 방지할 수 있음

자료 : 선행연구를 토대로 연구자 재작성

One-hot encoding의 경우 각 범주에 해당하는 변수들이 모두 차원에 더해지게 되어 '차원의 저주(the curse of dimensionality)' 문제가 발생할 수 있기에, 일반적으로 범주의 수가 15개 이상인 경우에는 사용하지 않고, 선형 모델 외에는 적합하지 않은 문제가 있다(Pargent et al., 2022). One-hot encoding은 범주형 변수의 범주 수 만큼 차원이 증가하고, 0과 1이라는 극단적인 값은 비교열위를 갖게 하여 모델의 성능을 저하시켜, 범주 수가 많은 경우 One-hot encoding 이외의 다른 방법을 검토한다(정태일, 2019). 예를 들어 범주형 변수인 시간에 세부 범주 수가 많아 오전, 오후, 저녁 등 시간대로 별도 파생 변수를 생성하여 One-hot encoding을 진행할 때, 범주형 변수 시간 이외에도 전체적으로 파생 변수 수가 많으면, 차원 추가에 따라 XGB 등 인공지능모델 성능이 저하되는 문제점이 있다(이단비, 2020). Label encoding의 경우 각 범주에 고유한 숫자값을 순서로 인식할 수 있으며, 범주 수가 많을 경우 사용하지 않고, 숫자값을 가중치로 잘못 인식하여 예측 성능이 저하될 수 있다. Target encoding의 경우는 범주별 평균값을 계산하여 복잡하며, 과적합이 될 가능성이 높다(정태일, 2019).

One-hot encoding, Label encoding 등 추가되는 범주 수가 증가하면 랜덤 포레스트, XGB 등 인공지능모델 변수 중요도를 파악할 때 추가되는 범주 값에 의해 변수 중요도 해석이 왜곡될 수 있다(엄지성, 2023). 20개의 범주 수가 존재하는 SVM 인공지능 모델에 대해 One-hot encoding, Label encoding, Frequency encoding을 적용하여 성능을 비교한 결과 Frequency encoding이 우수한 성능을 보여, 범주 수가 많은 경우 일반적인 사용하는 One-hot encoding, Label encoding 대신 Frequency encoding 분석을 고려할 수 있다(Pargent et al., 2022).

본 연구에서는 크게 4개 범주형 변수의 총 범주 개수가 38개로 차원 증가에 따른 모델 성능 저하를 예방하고, 차원이 커져서 모델 학습에 어려움이 있어 One-hot encoding, Label encoding, Target encoding 대신 Frequency encoding 방법을 적용하였다.

Frequency encoding의 사용방법은 다음과 같다.
예를 들어 월요일, 화요일, 수요일, 목요일, 금요일, 토요일 6가지 범주를 가진 '요일'이라는 데이터 열에 대해 월요일이 25번, 화요일이 30번, 수요일이 20번, 목요일이 15번, 금요일이 16번, 토요일이 5번 나타나면, Frequency encoding에서는 '월요일'을 '25'로 '화요일'을 '30'으로 '수요일'을 '20'으로 '목요일'을 '15'로 '금요일'을 '16'으로 '토요일'을 '5'로 변환한다. 최종적으로 '요일' 변수의 frequency encoding 값은 빈도가 가장 높은 '30'으로 표현되었다.

[표 3-5] Frequency encoding 예시

요일	빈도
월요일	25
화요일	30
수요일	20
목요일	15
금요일	16
토요일	5

이런 식으로 범주형 데이터를 인공지능 모델이 이해하기 쉬운 수치 형태로 변화하여 범주형 데이터를 단순화하며, 본 연구와 같이 총 범주 개수가 38개가 존재하더라도 차원을 그대로 유지한 채 빈도로 분석할 수 있다.

목표지향적 행동모형에 따르면 행동 빈도가 의도, 열망에 영향을 주므로 반복 습관에 의한 강사, 수업 등을 선택하는 행동 빈도는 재등록 의도에서 고려할 중요한 요소이다(남궁진, 2017). 강사, 시간(요일), 수업에 대한 만족도

가 높을수록 선택하는 횟수가 증가하여 재구매에 이르는 것으로 분석되었다(정현아, 2018). 따라서, 범주형 변수인 선호 강사, 선호 시간, 선호 수업, 선호 요일은 Frequency encoding을 통해 빈도를 집계하여 분석을 진행하였다.

재등록을 종속변수로 재등록은 1, 이탈은 0으로 구분하며, [표3-6]에서 변수를 설명하였다.

[표 3-6] 변수 설명

변수	설명
재등록 (Re-registration)	(종속변수) 해당 회원권 기간 만료 후 재등록 여부
출석 (Attend)	해당 회원권 기간 동안 출석일 수
노쇼 (Noshow)	해당 회원권 기간 동안 노쇼일 수 (사전 통보없이 무단결석)
취소 (Cancel)	해당 회원권 기간 동안 수업 전 취소건 수
선호 강사 (Teacher)	해당 회원권 기간 동안 출석 수업에서 가장 선호하는 강사 선택일 수
선호 시간 (Time)	해당 회원권 기간 동안 출석 수업에서 가장 선호하는 시간 선택일 수
선호 수업 (Class)	해당 회원권 기간 동안 출석 수업에서 가장 선호하는 수업 선택일 수
선호 요일 (Day)	해당 회원권 기간 동안 출석 수업에서 가장 선호하는 요일 선택일 수

Ⅳ. 연구 결과

4.1 기술 통계

우선적으로 자료의 정규성 여부를 검토하였다.
왜도/첨도, 정규성 검정(KS test/SW test)으로 판단하였다.
첫째, Kline(2015)은 왜도, 첨도의 기준은 각각 ±3 이내, ±8 이내일 때 정규성이 있는 것으로 검토한다. 노쇼, 취소의 경우 첨도가 정규성 판단 기준에 부합하지 않아 정규분포(가우시안 분포)를 하지 않는 것으로 판단된다.
본 연구를 위해 조사한 자료의 기술 통계는 다음 [표 4-1]과 같다.

[표 4-1] 기술 통계

측정변수	개수 (N)	최대값 (Max)	최소값 (Min)	범위 (Range)	평균 (Average)	표준편차 (Stddev)	왜도 (Skewness)	첨도 (Kutosis)
출석 (Attend)	1,160	78	2	76	27.61	14.49	0.30	-0.95
노쇼 (Noshow)	1,160	15	0	15	1.42	2.02	2.67	10.01
취소 (Cancel)	1,160	257	0	257	33.28	26.20	2.09	8.66
선호 강사 (Teacher)	1,160	56	1	55	12.25	9.16	1.46	2.23
선호 시간 (Time)	1,160	54	1	53	12.37	9.10	1.43	2.05
선호 수업 (Class)	1,160	51	1	50	19.71	10.84	0.53	-0.47
선호 요일 (Day)	1,160	35	1	34	10.89	7.33	1.07	0.32

둘째, 정규성을 판단하기 위해 Kolmogorov-Smirnov(KS test)와 Shapiro-Wilk (SW test) 정규성 검정방법을 이용할 수 있다. Kolmogorov-Smirnov(KS test)는 표본 수가 50개 이상일 때 사용하며, Shapiro-Wilk(SW test)는 표본 수가 50개 미만일 때 주로 사용한다. 본 연구에서는 표본 수가 1,160개로 Kolmogorov-Smirnov(KS test)를 통해 정규성을 검정하였다.

H0 : 표본의 분포는 정규성을 만족한다.
H1 : 표본의 분포는 정규성을 만족하지 않는다.
5% 유의수준에서 유의확률이 0.000이므로 귀무가설이 기각되어 정규분포가 아닌 것으로 확인되어, 비모수 통계를 이용하여 분석했다.

[표 4-2] 정규성 검정

측정변수	Kolmogorov-Smirnov (KS test)		
	통계량	자유도	유의확률
출석 (Attend)	0.439	1,160	0.000
노쇼 (Noshow)	0.600	1,160	0.000
취소 (Cancel)	0.525	1,160	0.000
선호 강사 (Teacher)	0.491	1,160	0.000
선호 시간 (Time)	0.474	1,160	0.000
선호 수업 (Class)	0.389	1,160	0.000
선호 요일 (Day)	0.469	1,160	0.000

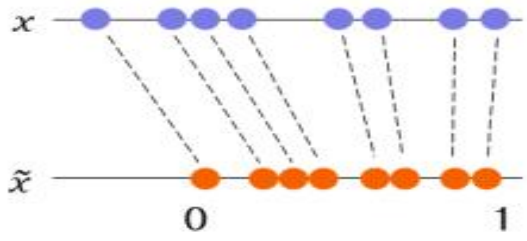
4.2 데이터 전처리

피처(Feature)는 데이터의 값을 잘 예측하기 위한 데이터의 특징으로 데이터 표에서 열(column)을 지칭한다. 열(column)은 피처(Feature), 속성(Attribute), 변수(Variable)로, 행(row)은 개체(instance), 관측치(observed value), 기록(record), 경우(case)로 사용한다. 피처는 항상 열(column) 위치하여 피처와 관측치를 열과 행 위치에서 바꾸면 안 된다. 피처 엔지니어링(Feature Engineering)은 가공되지 않은 데이터를 인공지능 모델에 사용할 수 있는 기능으로 전환하는 과정이라 할 수 있다. 피처 엔지니어링은 원시 데이터를 변환 및 통합하여 효과적인 피처를 만들어 인공지능 모델 훈련에 활용한다.

특정 데이터값에 따라 편향성을 갖고 전체 값과 최근 값 범위가 달라 그대로 해석하기에 무리가 있어 추가 작업이 필요하다. 피처 스케일링(Feature Scaling)은 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업이다. 변수 값의 범위 또는 단위가 달라서 발생 가능한 문제를 예방하고, 특정 데이터에 의해 편향성을 갖는 것을 방지할 수 있어 피처 스케일링을 진행한다. 종류는 크게 최소-최대 스케일링(정규화)과 표준화로 나눌 수 있다.

최소-최대 스케일링(Min-Max scaling)은 식 (1)을 적용하여 모든 피처의 값을 [그림 4-1]과 같이 [0, 1] 범위의 값으로 압축시키거나 확장하는 것이다.

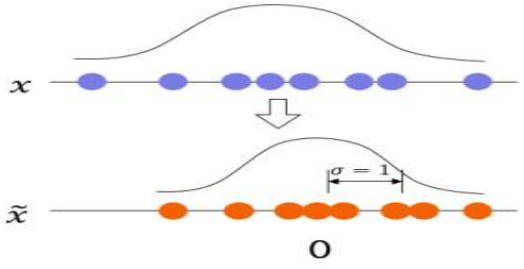
$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (1)$$



[그림 4-1] 최소-최대 스케일링

표준화(Standardization)는 데이터가 정규분포(가우시안 분포)일 일 때 평균이 0이고 분산이 1인 값으로 변환하여 모델의 예측 성능 향상에 도움이 된다.

$$\tilde{x} = \frac{x - mean(x)}{\sqrt[2]{var(x)}} \qquad (2)$$



[그림 4-2] 표준화 스케일링

본 연구는 기술 통계, Kolmogorov-Smirnov(KS test) 검토 결과 정규분포(가우시안 분포)를 하지 않는 것으로 확인되어 최소-최대 스케일링을 진행하였다. 향후 진행하는 군집 분석의 경우 관측값 사이에서 유사성을 찾아 군집을 형성하는 데, 넓은 범위를 가지는 변수는 군집에서 더 큰 영향을 줄 수 있어, min_max값을 기반으로 후속 작업을 실시하였다.

4.3 재등록 영향 변수 선정

종속변수 재등록 (Re-registration) 2개 집단(1: 재등록, 0: 이탈)에 대해 검정을 진행하였다. 일반적인 모수 검정에서는 정규성, 독립성, 등분산성의 3가지 기본 전제를 충족하는 x^2 (chi-square) 검정을 통해 기대빈도와 관찰빈도와의 부합 정도(차이)를 측정하며, 관찰빈도와 기대빈도 사이에 차이가 많이 날수록 값이 커지며 귀무가설을 기각한다. 본 연구에는 비모수 검정으로 x^2 (chi-square) 검정을 대신하여 Mann-Whitney U 검정을 사용하였다. Wilcoxon rank sum test를 이후에 Mann과 Whitney가 개정하여 U 통계량을 사용하였다.

각 변수를 min_max 변환 후 Mann-Whitney U 검정 값에 대해 유의확률을 확인 결과 취소 min_max를 제외하고 다른 변수 min_max 분포는 귀무가설을 기각하여 재등록 범주에 대해 통계적으로 유의한 것으로 파악되었다.

[표 4-3] min_max 변환 Mann-Whitney U 검정 의사결정

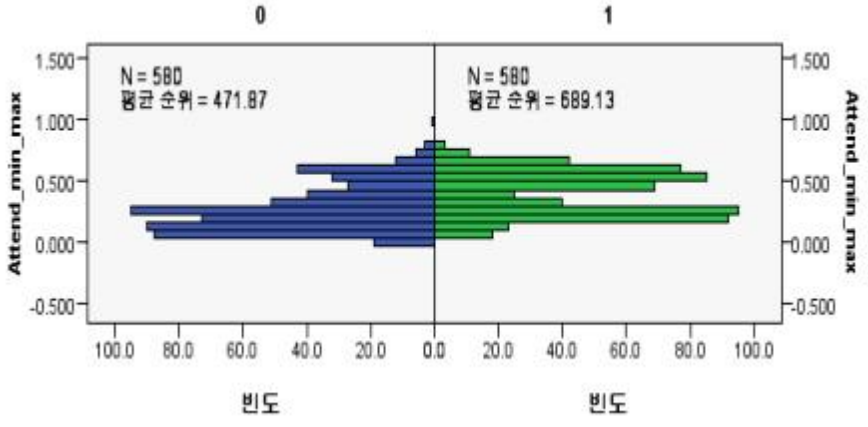
귀무가설	유의확률	의사결정
출석 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.
노쇼 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.
취소 min_max의 분포가 재등록의 범주에서 같다.	0.489	귀무가설을 채택한다.
선호 강사 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.
선호 시간 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.
선호 수업 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.
선호 요일 min_max의 분포가 재등록의 범주에서 같다.	0.000	귀무가설을 기각한다.

Mann-Whitney U 검정 값은 데이터를 순위별로 나열시켜 놓고 각 집단의 순위합계로부터 집단별로 U값을 계산하며, 출석 min_max 변수 경우 재등록 여부의 두 그룹 “1”(재등록)의 평균 순위가 “0”(이탈)의 평균 순위보다 높게 나타났다.

[표 4-4] Mann-Whitney U 검정 평균 순위 및 순위 합

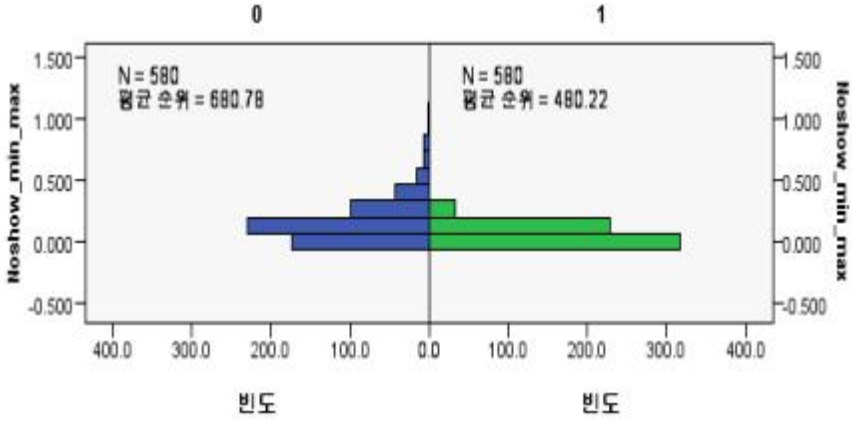
변수	ResultCode	개수	평균 순위	순위 합
출석 min_max	0	580	471.87	273,687.00
	1	580	689.13	399,693.00
노쇼 min_max	0	580	680.78	394,849.50
	1	580	480.23	278,530.50
취소 min_max	0	580	587.30	340,635.50
	1	580	573.30	332,744.50
강사 min_max	0	580	452.91	262,687.50
	1	580	708.09	410,692.50
시간 min_max	0	580	461.97	267,943.50
	1	580	699.03	405,436.50
수업 min_max	0	580	477.59	277,005.00
	1	580	683.41	396,375.00
요일 min_max	0	580	439.79	255,079.50
	1	580	721.21	418,300.50

[그림 4-3]은 출석 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 689.13으로 이탈의 평균 순위 471.87보다 높으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



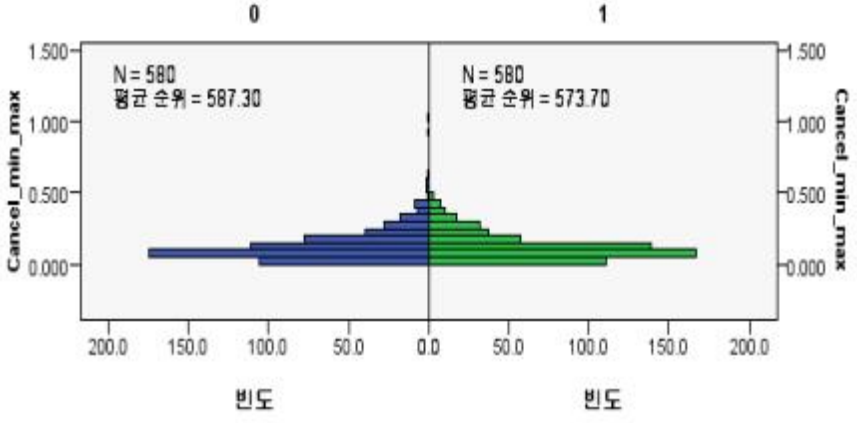
[그림 4-3] 출석 재등록 범주별 Mann-Whitney U 검정

[그림 4-4]는 노쇼 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 480.22로 이탈의 평균 순위 680.78보다 낮으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



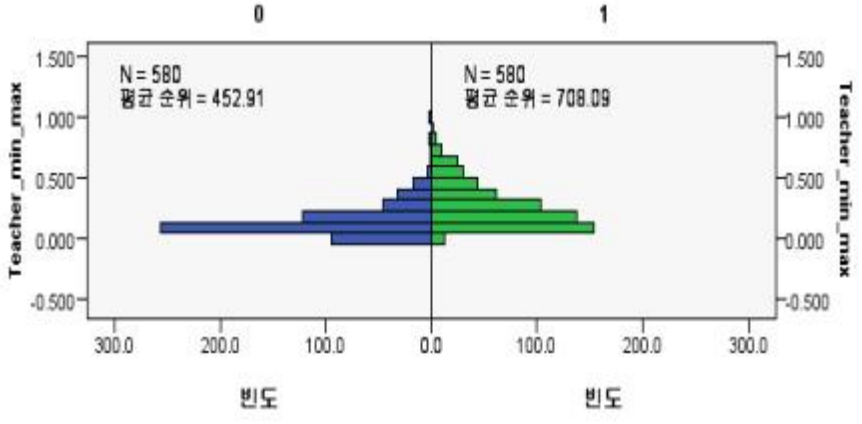
[그림 4-4] 노쇼 재등록 범주별 Mann-Whitney U 검정

[그림 4-5]는 취소 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 573.70으로 이탈의 평균 순위 587.30과 비슷하며, 재등록과 이탈의 빈도 분포가 유사하게 나타났다. 유의확률 0.489로 “취소 min_max의 분포가 재등록의 범주에서 같다.”는 귀무가설을 채택하여 통계적으로 유의하지 않음을 확인할 수 있었다.



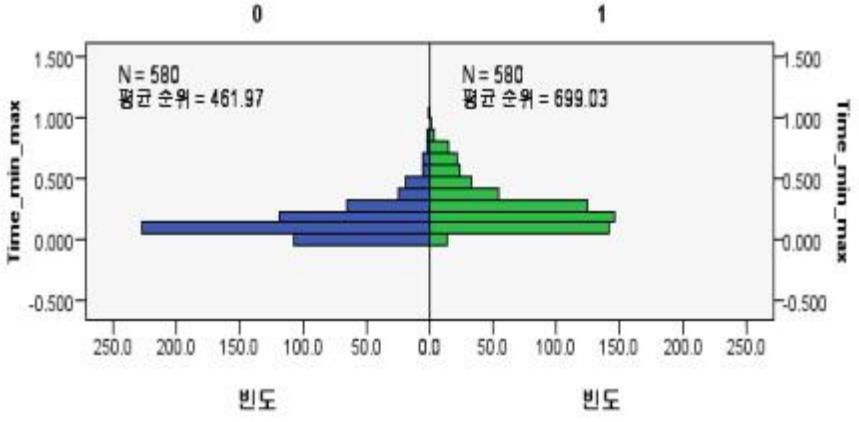
[그림 4-5] 취소 재등록 범주별 Mann-Whitney U 검정

[그림 4-6]은 선호 강사 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 708.09로 이탈의 평균 순위 452.91보다 높으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



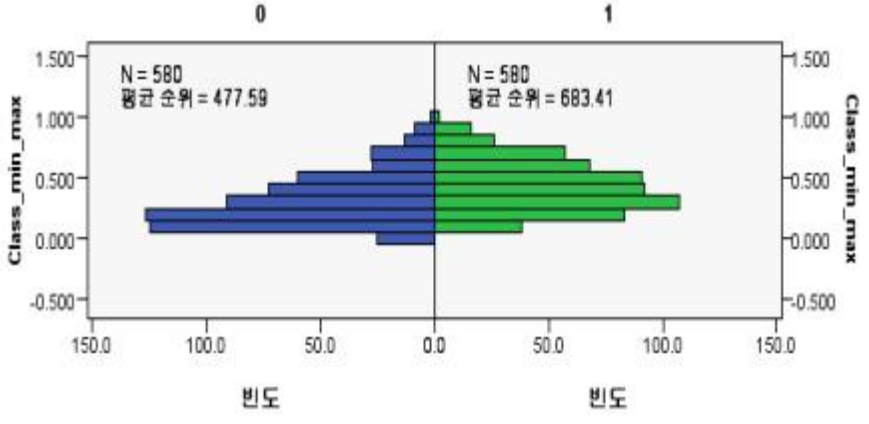
[그림 4-6] 선호 강사 재등록 범주별 Mann-Whitney U 검정

[그림 4-7]은 선호 시간 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 699.03으로 이탈의 평균 순위 461.97보다 높으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



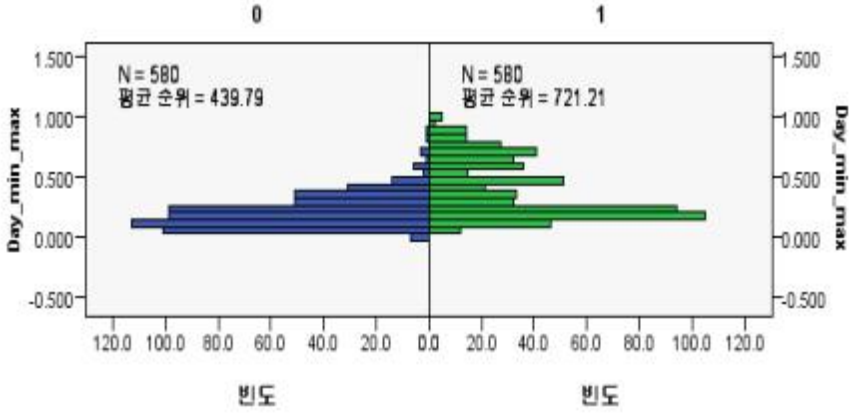
[그림 4-7] 선호 시간 재등록 범주별 Mann-Whitney U 검정

[그림 4-8]은 선호 수업 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 683.41로 이탈의 평균 순위 477.59보다 높으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



[그림 4-8] 선호 수업 재등록 범주별 Mann-Whitney U 검정

[그림 4-9]는 선호 요일 재등록 범주별 Mann-Whitney U 검정 결과 재등록의 평균 순위가 721.21로 이탈의 평균 순위 439.79보다 높으며 재등록과 이탈의 빈도 분포가 다르게 나타났다.



[그림 4-9] 선호 요일 재등록 범주 Mann-Whitney U 검정

취소 min_max의 분포가 재등록의 범주에서 같다는 귀무가설에 대해 유의확률은 0.489로 유의수준 0.05보다 큰 값으로 통계적으로 유의하지 않아 취소 min_max 변수를 향후 분석에서 제외할지를 검토해야 된다.
이후 진행하는 군집 분석에서의 Kruskal-Wallis H 검정에서 검토 결과 군집 2의 취소 min_max의 분포가 재등록 여부의 범주에서 같다는 귀무가설에 대해 유의확률이 0.650으로 유의수준 0.05보단 큰 값으로 귀무가설을 채택하여 통계적으로 유의하지 않는 것을 확인하였다. 따라서, 유의확률이 0.05보다 작은 군집 1, 군집 3의 경우 취소 min_max는 통계적으로 유의하여 취소 min_max 변수를 적용하며, 단일 집단, 군집 2는 취소 min_max가 통계적으로 유의하지 않아 제외하고 분석하였다.

[표 4-5] 취소의 군집에 따른 재등록 여부 범주 Kruskal-Wallis H 검정 의사결정

변수	귀무가설	검정 통계량	유의확률	의사결정
취소 min_max	군집 1의 취소 min_max의 분포가 재등록 여부의 범주에서 같다.	7.43	0.002	귀무가설을 기각한다.
	군집 2의 취소 min_max의 분포가 재등록 여부의 범주에서 같다.	0.21	0.650	귀무가설을 채택한다.
	군집 3의 취소 min_max의 분포가 재등록 여부의 범주에서 같다.	7.12	0.000	귀무가설을 기각한다.

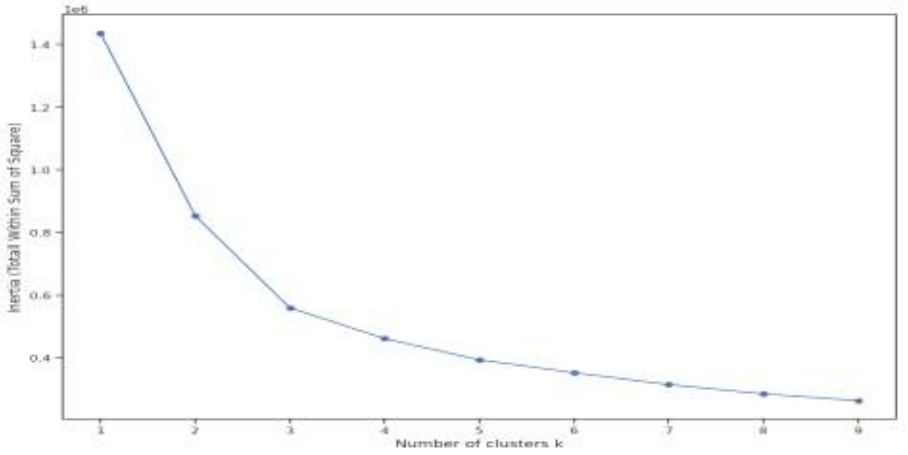
4.4 군집 분석을 통한 고객 세분화

군집 분석(Clustering)은 유사한 특성을 가진 대상들끼리 묶어 집단을 만드는 것을 말한다(임지수, 2020).

4.4.1 K-평균 군집화

K-평균 군집화(K-means Clustering)는 K값을 다르게 바꾸어가면서 군집화했을 때, 군집의 수 K별 “군집 내 군집과 개체 간 거리 제곱합의 총합(Inertia / Total within-cluster sum of squares)”의 그래프가 팔꿈치 모양으로 꺾이는 지점의 K를 최적 군집의 개수를 선택하는 방법을 팔꿈치 방법(The Elbow Method)이라고 한다. Inertia는 군집에 속한 샘플이 얼마나 가깝게 모여 있는지를 나타내는 값으로, 군집의 개수가 늘어나면 군집 개개의 크기는 줄어들기 때문에, Inertia가 줄어든다. 팔꿈치 방법은 군집의 개수를 늘려가면서 Inertia의 변화를 관찰하여 최적의 군집 개수를 찾아간다. 군집 개수를 증가시켜서 Inertia를 그래프로 그리면 감소하는 속도가 꺾이는 지점이 있다.

본 연구에는 군집의 개수가 3개일 때, 팔꿈치 모양으로 꺾이고, 군집 K가 4개 이상일 때부터는 Inertia의 변화가 매우 작으므로, 최적 군집의 개수를 3개로 하는 것이 바람직하다. 군집 수를 적게 하면 탐색은 빠르지만 좋은 모델을 안정적으로 찾기 힘들 수 있고, 군집 수가 많다면 탐색시간이 길어질 수 있다. 적절한 수의 군집 수를 선택하는 것이 핵심이다.



[그림 4-10] 팔꿈치 방법(The Elbow Method)

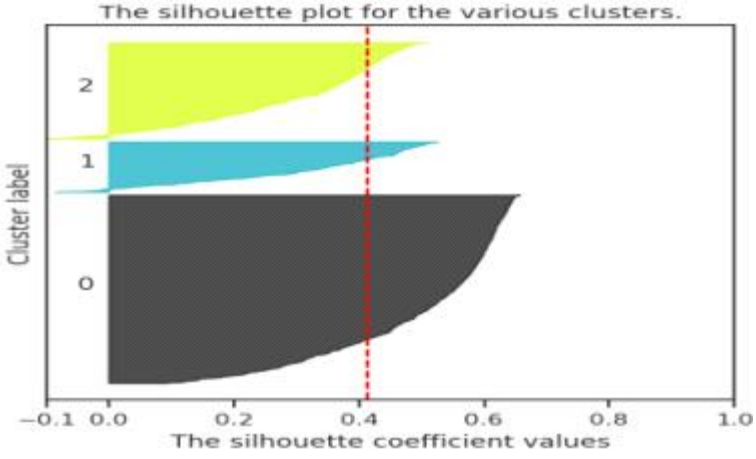
분할적 군집 분석(partitioning clustering)의 결과를 시각화하여 해석하고 평가할 수 있는 지표로 실루엣(silhouette)을 이용한다. 실루엣은 밀접성(tightness)와 분리성(separation)의 비교에 기반한 지표로, 실루엣은 어떤 객체가 군집 안에 잘 놓여 있는지, 그리고 어떤 객체는 단지 군집들 사이의 어딘가에 놓여 있는지를 보여준다. 전체 군집 분석의 실루엣들을 한 개의 그래프에 모아서 보여줌으로써, 데이터 구성과 형상의 개요(overview of the data configuration)와 군집들의 상대적인 질(the relative quality of the clusters)을 평가할 수 있다. 평균 실루엣 폭(the average silhouette width)은 군집화 결과에 대한 타당성(validity)을 평가하고, 또 적절한 군집 개수를 선택하는 데 활용할 수 있다.

X축은 실루엣 계수, Y축은 클러스터 ID를 의미하여, 빨간 선은 실루엣 계수의 전체 평균값을 의미한다. 실루엣 계수를 변경하면서 군집을 시각화하여 비교할 수 있도록 시뮬레이션하였다. 실루엣 계수가 -1에서 1 사이의 값을 가지며 1에 가까울수록 근처 군집과 멀리 떨어져 있음을 보여주는 데 값이 클수록 군집화가 잘 되었다고 판단한다. 대부분 객체가 높은 실루엣 값을 가진다면 군집의 구성이 적절하게 잘 되었다고 평가된다. K값에 따라 실루엣 계수를 비교하면 K값이 3일 때 실루엣 계수가 가장 크기에, 3개의 군집으로 나누었다.

[표 4-6] K값에 따른 실루엣 계수

K값	실루엣 계수
2	0.413
3	**0.415**
4	0.340
5	0.333
6	0.289
7	0.282
8	0.275
9	0.269

[그림 4-11]은 K=3일 때, 실루엣 계수가 0.415이며, 이때의 실루엣 시각화를 보여주었다.



[그림 4-11] K-means 실루엣 (K=3)

4.4.2 Kolmogorov-Smirnov 정규성 검정

3개 집단의 정규성을 판단하기 위해 Kolmogorov-Smirnov(KS test)와 Shapiro-Wilk (SW test) 정규성 검정방법을 이용할 수 있다. 본 연구에서는 앞서 분석과 같이 표본 수가 1,160개로 Kolmogorov-Smirnov(KS test)를 통해 정규성을 검정하였다.

H0 : 표본의 분포는 정규성을 만족한다.

H1 : 표본의 분포는 정규성을 만족하지 않는다.

5% 유의수준에서 유의확률이 0.000이므로 귀무가설이 기각되어 정규분포가 아닌 것으로 확인되어, 비모수 통계를 이용하여 분석하였다.

[표 4-7] 변수 군집별 정규성 검정

변수	군집	Kolmogorov-Smirnov (KS test)		
		통계량	자유도	유의확률
출석 min_max	1	0.433	652	0.000
	2	0.445	175	0.000
	3	0.484	333	0.000
노쇼 min_max	1	0.629	652	0.000
	2	0.595	175	0.000
	3	0.517	333	0.000
취소 min_max	1	0.530	652	0.000
	2	0.515	175	0.000
	3	0.480	333	0.000
선호 강사 min_max	1	0.434	652	0.000
	2	0.517	175	0.000
	3	0.538	333	0.000
선호 시간 min_max	1	0.415	652	0.000
	2	0.444	175	0.000
	3	0.498	333	0.000
선호 수업 min_max	1	0.351	652	0.000
	2	0.423	175	0.000
	3	0.448	333	0.000
선호 요일 min_max	1	0.392	652	0.000
	2	0.433	175	0.000
	3	0.448	333	0.000

4.4.3 Kruskal-Wallis H 검정

3개 이상의 집단에 대해 차이가 있는 지 없는 지를 알 수 있는 분석방법으로 크게 모수적 분석방법인 ANOVA(Analysis of Variance)와 비모수적 분석방법인 Kruskal-Wallis H 검정으로 나눌 수 있다. 본 연구는 비모수적 분석을 진행하므로, ANOVA 분석 대신 Kruskal-Wallis H 검정을 통해 3개 집단에 대해 비교하였다.

Kruskal-Wallis H 검정은 각 표본집단의 측정값을 집단이나 표본 크기에 관계없이 배열 후 크기 순으로 순위를 정하고, 각 변량에 대한 순위가 정해지면 각 집단별로 순위의 가중치를 구하고 서로 비교한다. Mann-Whitney U 검정의 확장으로 정규성 가정이 필요하지 않다.

비모수 분석인 Kruskal-Wallis H 검정 결과 모두 유의확률이 0.000으로 귀무가설을 모두 기각하였다. 즉, 통계적으로 유의한 차이가 있다.

[표 4-8] 변수별 Kruskal-Wallis H 검정 의사결정

귀무가설	검정 통계량	유의확률	의사결정
출석 min_max의 분포가 군집의 범주에서 같다.	793.524	0.000	귀무가설을 기각한다.
노쇼 min_max의 분포가 군집의 범주에서 같다.	21.760	0.000	귀무가설을 기각한다.
취소 min_max의 분포가 군집의 범주에서 같다.	511.483	0.000	귀무가설을 기각한다.
선호 강사 min_max의 분포가 군집의 범주에서 같다.	547.212	0.000	귀무가설을 기각한다.
선호 시간 min_max의 분포가 군집의 범주에서 같다.	558.772	0.000	귀무가설을 기각한다.
선호 수업 min_max의 분포가 군집의 범주에서 같다.	692.190	0.000	귀무가설을 기각한다.
선호 요일 min_max의 분포가 군집의 범주에서 같다.	730.733	0.000	귀무가설을 기각한다.

재등록 여부에 따라 변수별 군집 범주에 대해 Kruskal-Wallis 검정을 진행하였다. Kruskal-Wallis H 검정 결과 모두 유의확률이 0.000으로 귀무가설을 모두 기각하였다. 즉, 통계적으로 유의한 차이가 있다.

[표 4-9] 변수별 재등록 여부에 따른 군집 Kruskal-Wallis H 검정 의사결정

변수	귀무가설	검정 통계량	유의확률	의사결정
출석 min_max	재등록 여부가 이탈일 때 출석 min_max의 분포가 군집의 범주에서 같다.	341.80	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 출석 min_max의 분포가 군집의 범주에서 같다.	405.61	0.000	귀무가설을 기각한다.
노쇼 min_max	재등록 여부가 이탈일 때 노쇼 min_max의 분포가 군집의 범주에서 같다.	6.80	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 노쇼 min_max의 분포가 군집의 범주에서 같다.	3.25	0.000	귀무가설을 기각한다.
취소 min_max	재등록 여부가 이탈일 때 취소 min_max의 분포가 군집의 범주에서 같다.	257.51	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 취소 min_max의 분포가 군집의 범주에서 같다.	263.29	0.000	귀무가설을 기각한다.

선호 강사 min_max	재등록 여부가 이탈일 때 선호 강사 min_max의 분포가 군집의 범주에서 같다.	246.29	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 선호 강사 min_max의 분포가 군집의 범주에서 같다.	250.67	0.000	귀무가설을 기각한다.
선호 시간 min_max	재등록 여부가 이탈일 때 선호 시간 min_max의 분포가 군집의 범주에서 같다.	235.61	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 선호 시간 min_max의 분포가 군집의 범주에서 같다.	274.58	0.000	귀무가설을 기각한다.
선호 수업 min_max	재등록 여부가 이탈일 때 선호 수업 min_max의 분포가 군집의 범주에서 같다.	291.41	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 선호 수업 min_max의 분포가 군집의 범주에서 같다.	358.72	0.000	귀무가설을 기각한다.
선호 요일 min_max	재등록 여부가 이탈일 때 선호 요일 min_max의 분포가 군집의 범주에서 같다.	298.77	0.000	귀무가설을 기각한다.
	재등록 여부가 재등록일 때 선호 요일 min_max의 분포가 군집의 범주에서 같다.	386.60	0.000	귀무가설을 기각한다.

변수별 군집에 따른 재등록 여부 범주에 대해 Kruskal-Wallis 검정을 진행하였다. Kruskal-Wallis H 검정 결과 취소 min_max를 제외하고 다른 변수에 있어서는 모두 유의확률이 0.000으로 귀무가설을 기각하였다. 즉, 통계적으로 유의한 차이가 있다고 볼 수 있다. 군집 2의 취소 min_max의 분포가 재등록 여부의 범주에서 같다는 귀무가설에 대해 유의확률이 0.650으로 유의수준 0.05보단 큰 값으로 귀무가설을 채택하여 통계적으로 유의하지 않는 것을 확인할 수 있다. 따라서, 군집별 세부 분석을 할 때, 군집 1, 군집 3의 경우 취소 min_max는 유의확률이 0.05보다 작아 통계적으로 유의하여 취소 min_max 변수를 적용하며, 군집 2는 취소 min_max가 유의확률이 0.650으로 통계적으로 유의하지 않아 취소 min_max 변수를 제외하고 진행함이 필요하다.

[표 4-10] 변수별 군집에 따른 재등록 여부 Kruskal-Wallis H 검정 의사결정

변수	귀무가설	검정 통계량	유의확률	의사결정
출석 min_max	군집 1의 출석 min_max의 분포가 재등록 여부의 범주에서 같다.	3.96	0.000	귀무가설을 기각한다.
	군집 2의 출석 min_max의 분포가 재등록 여부의 범주에서 같다.	10.33	0.000	귀무가설을 기각한다.
	군집 3의 출석 min_max의 분포가 재등록 여부의 범주에서 같다.	58.86	0.000	귀무가설을 기각한다.
노쇼 min_max	군집 1의 노쇼 min_max의 분포가 재등록 여부의 범주에서 같다.	59.60	0.000	귀무가설을 기각한다.
	군집 2의 노쇼 min_max의 분포가 재등록 여부의 범주에서 같다.	25.76	0.000	귀무가설을 기각한다.
	군집 3의 노쇼 min_max의 분포가 재등록 여부의 범주에서 같다.	18.44	0.000	귀무가설을 기각한다.
취소 min_max	군집 1의 취소 min_max의 분포가 재등록 여부의 범주에서 같다.	7.43	0.002	귀무가설을 기각한다.
	군집 2의 취소 min_max의 분포가 재등록 여부의 범주에서 같다.	0.21	0.650	귀무가설을 채택한다.
	군집 3의 취소 min_max의	7.12	0.000	귀무가설을

	분포가 재등록 여부의 범주에서 같다.			기각한다.
선호 강사 min_max	군집 1의 선호 강사 min_max의 분포가 재등록 여부의 범주에서 같다.	9.76	0.000	귀무가설을 기각한다.
	군집 2의 선호 강사 min_max의 분포가 재등록 여부의 범주에서 같다.	16.15	0.000	귀무가설을 기각한다.
	군집 3의 선호 강사 min_max의 분포가 재등록 여부의 범주에서 같다.	81.28	0.000	귀무가설을 기각한다.
선호 시간 min_max	군집 1의 선호 시간 min_max의 분포가 재등록 여부의 범주에서 같다.	7.06	0.000	귀무가설을 기각한다.
	군집 2의 선호 시간 min_max의 분포가 재등록 여부의 범주에서 같다.	11.39	0.000	귀무가설을 기각한다.
	군집 3의 선호 시간 min_max의 분포가 재등록 여부의 범주에서 같다.	65.64	0.000	귀무가설을 기각한다.
선호 수업 min_max	군집 1의 선호 수업 min_max의 분포가 재등록 여부의 범주에서 같다.	1.12	0.000	귀무가설을 기각한다.
	군집 2의 선호 수업 min_max의 분포가 재등록 여부의 범주에서 같다.	2.81	0.000	귀무가설을 기각한다.
	군집 3의 선호 수업 min_max의 분포가 재등록 여부의 범주에서 같다.	51.66	0.000	귀무가설을 기각한다.

선호 요일 min_max	군집 1의 선호 요일 min_max의 분포가 재등록 여부의 범주에서 같다.	48.59	0.000	귀무가설을 기각한다.
	군집 2의 선호 요일 min_max의 분포가 재등록 여부의 범주에서 같다.	75.85	0.000	귀무가설을 기각한다.
	군집 3의 선호 요일 min_max의 분포가 재등록 여부의 범주에서 같다.	90.99	0.000	귀무가설을 기각한다.

Kruskal-Wallis H 검정을 통해 평균 순위 값을 속성값으로 다음 [표 4-11]과 같이 정리하였다.

[표 4-11] 변수의 군집별 Kruskal-Wallis H 검정

변수	군집	개수	검정 통계량 (H값) 평균 순위
출석 min_max	1	652	436.37
	2	175	522.67
	3	333	586.73
노쇼 min_max	1	652	593.87
	2	175	512.14
	3	333	377.28
취소 min_max	1	652	400.25
	2	175	390.11
	3	333	411.06
선호 강사 min_max	1	652	589.03
	2	175	671.25
	3	333	692.27
선호 시간 min_max	1	652	532.81
	2	175	651.47
	3	333	715.74
선호 수업 min_max	1	652	565.80
	2	175	661.43
	3	333	804.14
선호 요일 min_max	1	652	518.93
	2	175	542.16
	3	333	591.94

사후분석(Post-hoc Analysis)은 ANOVA 또는 Kruskal-Wallis H 검정을 시행하여 3개 집단 간 차이가 있다(적어도 하나는 차이가 있다)라는 결과가 나타났을 때 어떤 집단(간)이 다른 집단과 차이가 있는지 알아보고자 수행한다. 사후분석은 두 집단씩 짝지어서 여러 번 비교를 하는 것으로 t-test 또는 Mann-Whitney U 검정을 두 그룹씩 짝지어서 여러 번 수행한다. 제1종 오류 (귀무가설이 실제로는 참이어서 채택해야 함에도 불구하고 표본의 오차 때문에 이를 채택하는 않는 오류 / 유의수준(α)) 문제가 발생하므로 보정(adjusted)하는 방법이 필요하다. 보정하는 방법으로 Bonferroni correction method, Turkey, Duncan, Scheffe 방법 등이 있다. 모수적 분석방법인 ANOVA에서는 이러한 옵션을 지정하여 사후분석 결과를 출력하지만, 비모수 분석방법인 Kruskal-Wallis H 검정은 해당 옵션이 없어 직접 두 그룹씩 짝지은 비교 결과를 조정해야 한다.

각 변수에 대해 군집별 대응 비교 결과 유의확률이 0.000으로 유의수준 0.05미만으로 귀무가설을 기각하여 “세 군집 간에는 유의한 차이가 있다”라는 결론을 낼 수 있었다.

[표 4-12] Kruskal-Wallis H 사후 검정 (군집 대응별 비교)

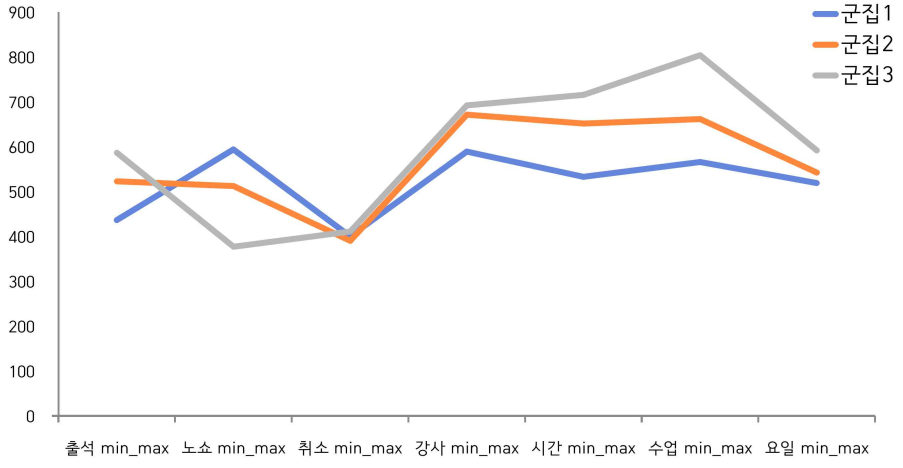
변수	군집 비교	검정 통계량 (H값)	표준 검정 통계량	유의확률
출석 min_max	군집1 - 군집2	-86.30	-3.54	0.000
	군집1 - 군집3	-150.36	-6.16	0.000
	군집2 - 군집3	-64.06	-2.63	0.000
노쇼 min_max	군집1 - 군집2	81.73	3.35	0.000
	군집1 - 군집3	216.59	8.88	0.000
	군집2 - 군집3	134.86	5.53	0.000
취소 mix_max	군집1 - 군집2	10.14	0.42	0.000
	군집1 - 군집3	-10.81	-0.44	0.000
	군집2 - 군집3	-20.95	-0.86	0.000
선호 강사 min_max	군집1 - 군집2	-82.22	-3.37	0.000
	군집1 - 군집3	-103.24	-4.23	0.000
	군집2 - 군집3	-21.02	-0.86	0.000
선호 시간 min_max	군집1 - 군집2	-118.66	-4.87	0.000
	군집1 - 군집3	-182.93	-7.50	0.000
	군집2 - 군집3	-64.27	-2.64	0.000
선호 수업 min_max	군집1 - 군집2	-95.63	-3.92	0.000
	군집1 - 군집3	-238.34	-9.77	0.000
	군집2 - 군집3	-142.71	-5.85	0.000
선호 요일 min_max	군집1 - 군집2	-23.23	-0.95	0.000
	군집1 - 군집3	-73.01	-2.99	0.000
	군집2 - 군집3	-49.78	-2.04	0.000

각 군집별로 Kruskal-Wallis H값을 [표 4-13]과 같이 정리하였다.

[표 4-13] 변수의 군집별 Kruskal-Wallis H 검정 속성값

변수		군집1	군집2	군집3
회원수		652	175	333
H값 (평균 순위)	출석 min_max	436.37	522.67	586.73
	노쇼 min_max	593.87	512.14	377.28
	취소 min_max	400.25	390.11	401.06
	선호 강사 min_max	589.03	671.25	692.27
	선호 시간 min_max	532.81	651.47	715.74
	선호 수업 min_max	565.80	661.43	804.14
	선호 요일 min_max	518.93	542.16	591.94

군집별 Kruskal-Wallis H 검정 속성값은 H값(평균 순위)를 군집별 도식화하면 다음 [그림 4-12]와 같이 나타났다.



[그림 4-12] 군집별 Kruskal-Wallis H 속성 비교

군집 1은 회원 수 652명으로 구성되어, 총 7개 변수에 대해 노쇼 min_max는 다른 군집에 비해 높은 값을 보여주고, 취소 min_max는 군집2와 군집3의 중간 값이 나타났다. 노쇼, 취소는 부정적인 변수로 작을수록 좋고, 나머지 변수는 긍정적인 변수로 클수록 좋은 특성이 있다. 긍정적인 변수와 부정적인 변수가 일관된 방향을 나타내고 있어 개별 변수로 판단하기 보다는 전체적인 경향으로 군집을 판단한다. 노쇼 min_max H값이 가장 높은데, 이는 다른 군집에 비해 노쇼 빈도가 높은 특성이며, 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 취약한 회원 군집이라 판단하여 군집 1은 “취약 회원군”으로 이름 지었다. 군집 1 내에서 상대적으로 노쇼 min_max H값이 593.87로 다른 변수 H값이 비해 높아 노쇼 속성이 군집 1에서 중요한 영향을 미칠 것으로 보인다

군집 2는 회원 수 175명으로 구성되어, 총 7개 변수에 대해 속성값 H값(평균 순위)이 중간 위치로 나타났다. 총 3개 군집 중에서 중간 회원 군집이라 판단하여 군집 2는 “중간 회원군”으로 이름 지었다. 군집 2 내에서 상대적으로 H값이 선호 강사 min_max 값이 671.25로 다른 변수 H값에 비해 높아 선호 강사 속성이 군집 2에서 중요한 영향을 미칠 것으로 보인다.

군집 3은 회원 수 333명으로 구성되어, 노쇼 min_max, 취소 min_max를 제외하고 다른 군집에 비해 높은 값을 보여주었다. 노쇼 min_max H값이 군집 내에서 가장 낮은 데, 이는 다른 군집에 비해 노쇼, 취소 빈도가 낮은 특성이며, 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 우수한 회원 군집이라 판단하여 군집 3은 “우수 회원군”으로 이름 지었다. 군집 3 내 상대적으로 선호 수업 min_max H값이 804.14로 다른 변수 H값에 비해 높아 선호 수업 속성이 군집 3에서 중요한 영향을 미칠 것으로 보인다.

4.5 재등록 예측

본 연구에서는 선행연구에서 진행한 기존 재등록(이탈) 예측 연구의 한계를 해소하고자 크게 2가지 시나리오로 총 3가지 인공지능 알고리즘을 통해 최적의 재등록을 예측하였다. 트리 기반의 인공지능 알고리즘 중 기본적인 의사결정나무(DT : Decision Tree)와 과적합을 방지하기 위해 개선된 랜덤 포레스트(RF:Random Forrest), XGB를 활용하였다. 시나리오 1은 전체 고객에 대해 하나의 집단으로 보고 시나리오 2는 고객 이용행태 기반의 3개의 군집으로 나누어 평가하였다. 재등록 예측 성능을 평가하고자 총 6가지 지표인 정확도(Accuracy), 정밀도(Precision), 재현도(Recall), 특이도(Specificity), F1 스코어(F1 Score), AUC(Area Under the ROC Curve)를 이용하였다.

4.5.1 데이터 분할

데이터 분할은 전체 데이터를 여러 부분으로 분할하는 것을 의미하며, 데이터 분석모델이 주어진 데이터에 대해서만 높은 성능을 보이는 과적합을 방지하기 위해 일부 데이터로 학습하고, 일부 데이터로 검증을 한다(강성식,2021). 데이터 분할시 고려해야 할 사항이 학습용 데이터(Training data)와 평가용 데이터(Test data)가 전체 데이터에 대한 대표성을 가져야 한다. 학습용 데이터(Training data)를 사용하여 모델을 학습시키고, 평가용 데이터(test data)로 모델 성능을 평가한다. 전체 데이터의 일부인 평가용 데이터(test data)를 사용해 모델 성능을 평가하는 것에 대한 문제는 평가용 데이터(test data)가 모델에 학습되지 않기 때문에 평가용 데이터(test data)의 선정에 따라 성능이 다르게 나타날 수 있어, 평가용 데이터(test data)에 대한 성능평가의 신뢰성이 떨어질 수 있다. 본 연구에서 전체 데이터의 70%는 인공지능 모형 학습용, 나머지 30%는 모형 평가용에 이용하였다. 이 데이터를 어떻게 나누느냐에 따라 성능을 평가한 결과가 터무니없이 높거나 낮을 수 있다.

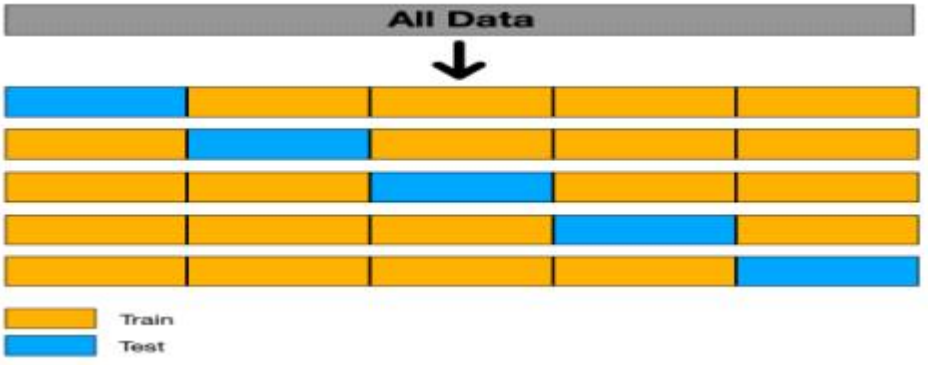
의사결정나무(DT), 랜덤 포레스트(RF), 극한 경사도 부스팅 (XGB)에 대해 공통적으로 학습용 데이터(Training data) 70% / 평가용 데이터(Test

data) 30%로 분리하였다.

4.5.2 k겹 교차검증

인공지능 성능평가에는 데이터를 어떻게 잘 나누는 지가 핵심이다. 이럴 때 우선 고려하는 것이 교차검증(Cross Validation)이라 할 수 있다. 데이터를 여러 번 반복해서 나누어 여러 모델을 학습하여 성능을 평가한다.

주로 k겹 교차검증이 안정적으로 뛰어난 평가기법으로 알려져 있다. 보통 k값은 5 또는 10을 사용하는 데, 본 연구에서는 k값을 일반적으로 사용하는 5로 지정하여 분할하였다. 데이터를 폴드(fold)라고 하는 거의 비슷한 크기의 부분 집합 다섯 개로 나눈다. 첫 번째 모델은 첫 번째 폴드를 평가용 데이터로 사용하고, 두 번째부터는 폴드(4개의 폴드)를 학습용 데이터로 사용한다. 그 다음 모델을 두 번째 폴드를 평가용, 1,3,4,5 폴드를 학습용 데이터로 사용한다. 이후에도 마찬가지고 수행하여 총 5개의 정확도를 얻을 수 있다. 5개의 결과물의 평균값이 모델의 성능이 된다.



[그림 4-13] k(5)-겹 교차검증

하지만, k겹 교차검증은 데이터 세트를 k번 나누어 학습하는 만큼, 일반적인 분할 방법에 비해 학습과 평가에 소용되는 연산비용이 증가하는 단점이 존재한다.

4.5.3 모델 성능평가 척도

혼동행렬(Confusion Matrix)은 예측값이 실제 값을 얼마나 정확히 예측했는지 보유주는 행렬로 모델의 성능을 평가할 때 주로 사용하는 지표이다.

[표 4-14] 혼동행렬 (Confusion Matrix) 구분

구분		실제값	
		Y	N
예측값	Y	TP (True Positive : 참긍정) 정분류	FP (False Positive : 거짓긍정) 오분류
	N	FN (False Negative : 거짓 부정) 오분류	TN (Ture Negative : 참부정) 정분류

혼동행렬을 읽는 방법은 좌상단부터 우하단으로 대각선 방향으로 진행한다. 이상적으로 분류된 결과에 대한 혼동행렬에서는 이런 주 대각 요소만이 0이 아닌 값을 가지고 있고, 비 대각 요소들은 0의 값을 가질 것이다. 즉 완벽한 분류기에서는 FP와 FN에 해당하는 분류결과는 0이다.

정확도(Accuracy)는 분류기가 선택해야 할 모든 케이스를 선택하고 거부해야 할 모든 케이스를 거부하는 능력이다. 어떤 방식으로 예측을 하였든 실제로 그러한 예측이 참이었는지 보유주는 지표로, 가장 직관적으로 상식적인 지표라고 할 수 있다. 100%의 정확도를 가지는 분류기에서는 FN=FP=0이다. 다만 결과만 맞는다면 과정에는 큰 관심이 없어 약점 또한 존재한다. 예를 들어 환자 10명 중 9명이 치명적인 질병에 걸렸다면, 인공지능 모델에서 환자들이 병에 걸렸다고 긍정 예측을 하여 90%의 정확도를 보이게 된다. 하지만 모델이 우수하지 않고 찍어서도 높은 정확도가 나타날 수 있어 모델의 변별력이 낮을 수 있다. 그래서 인공지능 모형 평가에서는 정확도만을 가지고 판

단하기에 문제가 있을 수 있어, 다른 지표들과 함께 고려해야 된다.

정밀도(Precision)는 결과 중에서 실제로 관련이 있는 케이스의 비율로 모델의 예측값이 얼마나 정확하게 예측됐는가를 나타내는 지표로 “예”라고 예측했을 때의 정답률이다. 실제로 어떤 방향(긍정/부정)의 예측에 관심을 두는 것이 아니라, 긍정적으로 예측하였다면 그러한 예측이 실제와 ‘일치하는지’에 관심을 두는 지표이다. 다만 부정으로 예측하였을 때 실제로 거짓일 확률을 계산하지 않는다는 한계가 존재한다.

재현율(Recall)은 분류기가 선택해야 할 모든 케이스를 선택하는 능력으로 민감도(Sensitivity)와 동일하며, 실제로 참일 때, 참으로 예측했을 확률을 계산한다. 정확도(Accuracy)보다 개선된 지표지만, 우연히 찍어서 높은 재현율을 보일 수 있는 한계가 존재한다.

특이도(Specificity)는 거부해야 할 모든 케이스를 거부하는 능력으로, 실제로 거짓일 때 거짓으로 예측했을 확률을 계산한다. 재현율과 대칭이 되는 지표로, 우연히 찍어서 특이도가 높게 나올 수 있는 한계가 존재한다.

F1 스코어(F1 score)는 정밀도와 재현율의 조화평균으로 구한 값이다. 조화평균의 특징이 분모의 값이 일정할 때, 분모의 두 값이 이질적일수록 전체적으로 더 낮은 값이 계산된다. 예를 들어, 정밀도와 재현율이 각각 0.9와 0.1일 떼 F1 스코어는 0.18이지만, 정밀도와 재현율이 각각 0.5와 0.5일 때, F1 스코어는 0.5값으로 분모값이 같더라도 두 값이 한쪽으로 편중되면 F1 스코어가 낮아지게 된다. 실제 참일 때 참으로 예측할 확률과 참으로 예측했을 때 실제 참일 확률은 인공지능 모형 예측력을 평가하는 데 동등하게 중요한 지표임을 고려해야 한다. 정밀도와 재현율은 둘 다 TP를 높이는 데 초점을 맞추지만, 정밀도는 FP를 낮추고, 재현율은 FN을 낮추는 데 초점을 맞춘다는 점이 서로 다른 특성이다. 정밀도와 재현율은 모형 평가에 대표적인 지표이지만, 한 지표값이 높아지면 다른 지표 값이 낮아질 가능성이 높아 관계를 보정하는 지표이다. 정밀도와 재현율이 모두 높을수록 1에 가까운 값(최대 1)을 가질 수 있다.

AUC (Area Under the ROC Curve)는 ROC Curve의 밑면적을 말한다. ROC(Receiver Operating Characteristic) Curve는 FPR(False Positive Rate)

과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프이다.

TPR(True Positive Rate)는 1인 케이스에 대해 1로 바르게 예측하는 비율 민감도(Sensitivity)이며, FPR(False Positive Rate)은 0인 케이스에 대해 1로 틀리게 예측하는 비율로 1-특이도((Specificity)이다. 민감도(Sensitivity)와 특이도(Specificity)는 반비례하며, 다시 말해 민감도(Sensitivity)와 1-특이도((Specificity)는 비례한다. ROC Curve는 모델의 판단 기준을 연속적으로 바꾸면서 측정했을 때 FPR와 TPR의 변화를 나타낸 것으로 (0,0)과 (1,1)을 잇는 곡선이다. ROC Curve가 좌상단으로 가장 많이 치우친 그래프를 갖는 모델이 가장 높은 성능을 보인다. AUC는 이러한 ROC Curve 밑면적으로 수치로 성능을 평가하는 데, 1에 가까울수록 그래프가 좌상단에 근접하여 좋은 모델로 판단된다.

변수 중요도(Feature Importance)는 어느 변수가 예측 성능에 중요한 역할을 하는 지를 보여주는 척도이다. 해당 변수가 상대적으로 얼마만큼 종속변수(label)에 영향을 주는 가를 보여주는 것으로, 어떻게 영향을 주는 지를 알려주지 않는다. 랜덤 포레스트나 XGB와 같은 트리 기반 모델에 대해 변수 중요도를 검토하였다.

4.5.4 단일 재등록 예측 모형

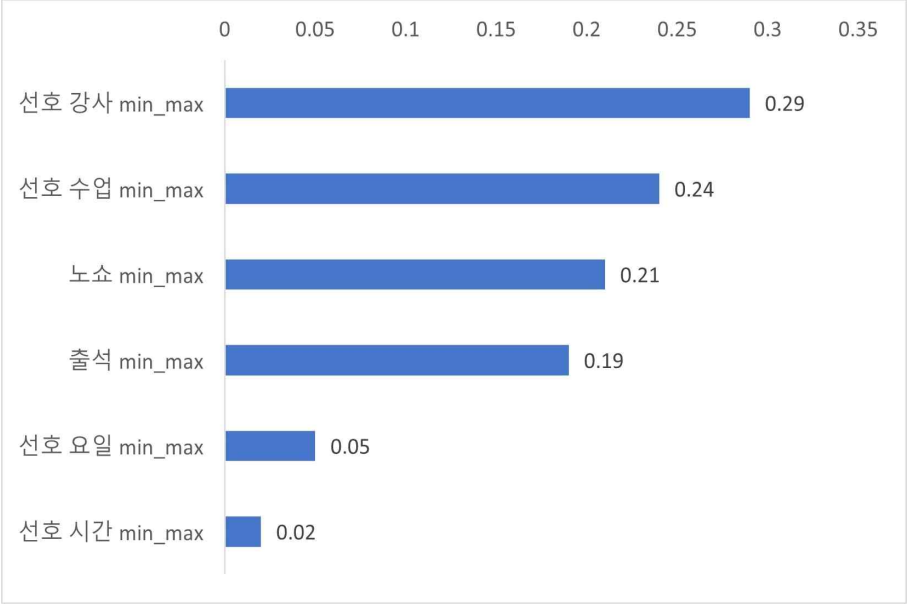
시나리오 1은 단일 재등록 예측 모형으로, 일반적인 재등록(이탈) 예측 연구에서 사용하고 있다. 단일 재등록 예측 모형은 전체 고객을 대상으로 하여 3가지 알고리즘을 적용하여 성능을 평가하였다.

[표 4-15]에서 단일 집단에 대한 재등록 예측 모형은 XGB가 가장 좋은 성능을 보였다. 성능평가 지표는 주로 정확도, 정밀도, 재현율, 특이도, F1 스코어를 비교하는 데 정확도를 비롯한 전반적인 지표에 대해 XGB가 다른 모형에 비해 우수한 값을 보였다. 예측 모형 성능을 평가하는 여러 지표 중에서 정확도, F1 스코어(정밀도, 재현율 활용), AUC(특이도 활용)를 중심으로 판단하는 데, Training Data에 비해 Test Data 성능이 개선됨을 확인할 수 있었다.

[표 4-15] 단일 재등록 예측 모형 성능

대상	데이터 개수	구분	모형		
			DT	RF	XGB
Training Data	812 (70%)	정확도 (Accuracy)	0.72	0.75	0.85
		F1 스코어 (F1 score)	0.72	0.75	0.85
		정밀도 (Precision)	0.73	0.77	0.83
		재현율 (Recall)	0.72	0.73	0.88
		특이도 (Specificity)	0.73	0.78	0.82
		AUC	0.73	0.76	0.85
Test Data	348 (30%)	정확도 (Accuracy)	0.73	0.77	0.87
		F1 스코어 (F1 score)	0.74	0.75	0.87
		정밀도 (Precision)	0.73	0.71	0.86
		재현율 (Recall)	0.75	0.79	0.88
		특이도 (Specificity)	0.73	0.76	0.85
		AUC	0.74	0.78	0.86

[그림 4-14]는 단일 집단의 최종 선택 모형인 XGB에서 재등록에 영향을 주는 변수 중요도(Feature Importance)를 보여 주는데, 선호 강사 min_max(0.29), 선호 수업 min_max (0.24), 노쇼 min_max(0.21) 순으로 나타났다. 취소 min_max는 통계적으로 유의하지 않아 변수 중요도에서 제외하여 분석하였다.



[그림 4-14] 단일 집단 XGB 모델 변수 중요도

4.5.5 군집별 재등록 예측 모형

시나리오 2는 군집별 재등록 예측 모형이며, K-means 실루엣 분석으로 총 3개의 적절한 군집 수로 고객을 세분화하여, 군집별 3가지 예측 알고리즘의 성능을 평가하였다. 군집1(취약 회원군), 군집2(중간 회원군), 군집3(우수 회원군)의 정확도, F1 스코어, AUC를 검토 결과 공통적으로 DT 〈 RF 〈 XGB 순으로 성능이 우수한 것으로 [표 4-16]에서 확인되었다.

[표 4-16] 군집별 Test Data 재등록 예측 모형 성능

군집	구분	모형		
		DT	RF	XGB
군집1 (취약 회원군)	정확도 (Accuracy)	0.78	0.84	0.90
	F1 스코어 (F1 score)	0.72	0.83	0.88
	정밀도 (Precision)	0.74	0.87	0.88
	재현율 (Recall)	0.70	0.78	0.87
	특이도 (Specificity)	0.78	0.79	0.87
	AUC	0.78	0.79	0.87
군집2 (중간 회원군)	정확도 (Accuracy)	0.79	0.85	0.91
	F1 스코어 (F1 score)	0.80	0.86	0.91
	정밀도 (Precision)	0.85	0.89	0.91
	재현율 (Recall)	0.76	0.83	0.90
	특이도 (Specificity)	0.78	0.81	0.90
	AUC	0.79	0.81	0.91
군집3 (우수 회원군)	정확도 (Accuracy)	0.84	0.88	0.92
	F1 스코어 (F1 score)	0.88	0.90	0.91
	정밀도 (Precision)	0.89	0.90	0.89
	재현율 (Recall)	0.88	0.91	0.92
	특이도 (Specificity)	0.80	0.82	0.91
	AUC	0.80	0.82	0.92

피트니스 재등록 인공지능 분석과 관련한 선행연구는 Clavel et al.(2016)의 로지스틱 회귀, Aldosary & Alrashdan (2021)의 ANN, Sobreiro et al.(2021)의 의사결정나무, 랜덤 포레스트, XGB가 있다.

본 연구에 사용한 알고리즘은 Tree 기반의 의사결정나무, 랜덤 포레스트, XGB로 Clavel et al.(2016)의 로지스틱 회귀, Aldosary & Alrashdan (2021)의 ANN 분석과 직접적인 예측 성능 비교는 무리가 있다. Clavel et al.(2016)의 로지스틱 회귀분석 결과 정확도가 65.6%, 재현율 80.0%로 본 연구의 DT, RF, XGB 정확도 성능이 더 우수하고, 군집 1의 DT, RF, 군집 2 DT를 제외하고 나머지 군집 및 알고리즘에서 재현율 성능이 더 우수한 것으로 확인되었다. Aldosary & Alrashdan (2021)의 ANN 분석 결과 정확도 92.1%, 재현율 89.1%, 특이도 93.8%로 본 연구의 정확도, 재현율, 특이도가 낮은 것으로 확인되었다.

[표 4-17]과 같이 Sobreiro et al.(2021)에서 사용한 의사결정나무, 랜덤 포레스트, XGB 모형을 비교하면 군집 3과 비슷한 예측 성능이 조사되었다. 전반적으로 의사결정나무 〈 랜덤 포레스트 〈 XGB 순으로 성능이 개선된 것으로 파악되었다. 이는 의사결정나무의 과적합을 방지하기 위한 랜덤 포레스트 배깅(bagging), XGB 부스팅(boosting) 앙상블 모델의 성능 개선 효과를 알 수 있었다.

[표 4-17] Sobreiro 피트니스 재등록 인공지능 분석 선행연구

구분	의사결정나무	랜덤포레스트	XGB
정확도	0.83	0.92	0.95
F1 스코어	0.86	0.90	0.95
정밀도	0.93	0.96	0.96
재현율	0.84	0.93	0.98
특이도	0.71	0.79	0.73
AUC	0.83	0.85	0.86

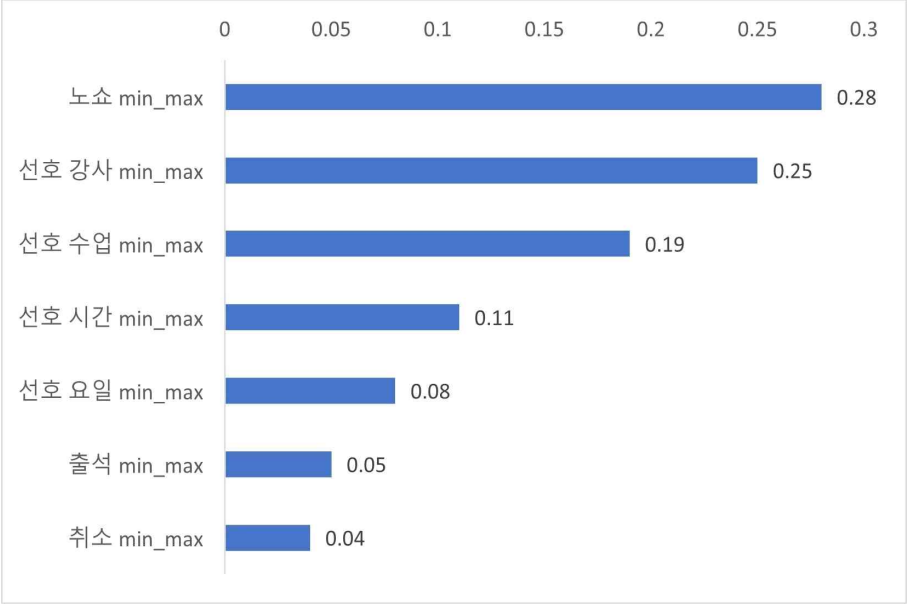
[표 4-18]은 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군) 재등록 예측 모형의 Training Data과 Test Data에 대한 성능을 정리하였으며, 공통적으로 Test Data의 예측 성능이 개선됨을 확인하였다.

[표 4-18] 군집별 Training Data와 Test Data 재등록 예측 모형 성능

군집	대상	데이터 개수	구분	모형		
				DT	RF	XGB
군집1 (취약 회원군)	Training Data	457 (70%)	정확도 (Accuracy)	0.75	0.83	0.88
			F1 스코어 (F1 score)	0.73	0.79	0.86
			정밀도 (Precision)	0.70	0.80	0.82
			재현율 (Recall)	0.75	0.78	0.90
			특이도 (Specificity)	0.75	0.78	0.85
			AUC	0.75	0.78	0.85
	Test Data	195 (30%)	정확도 (Accuracy)	0.78	0.84	0.90
			F1 스코어 (F1 score)	0.72	0.83	0.88
			정밀도 (Precision)	0.74	0.87	0.88
			재현율 (Recall)	0.70	0.78	0.87
			특이도 (Specificity)	0.78	0.79	0.87
			AUC	0.78	0.79	0.87

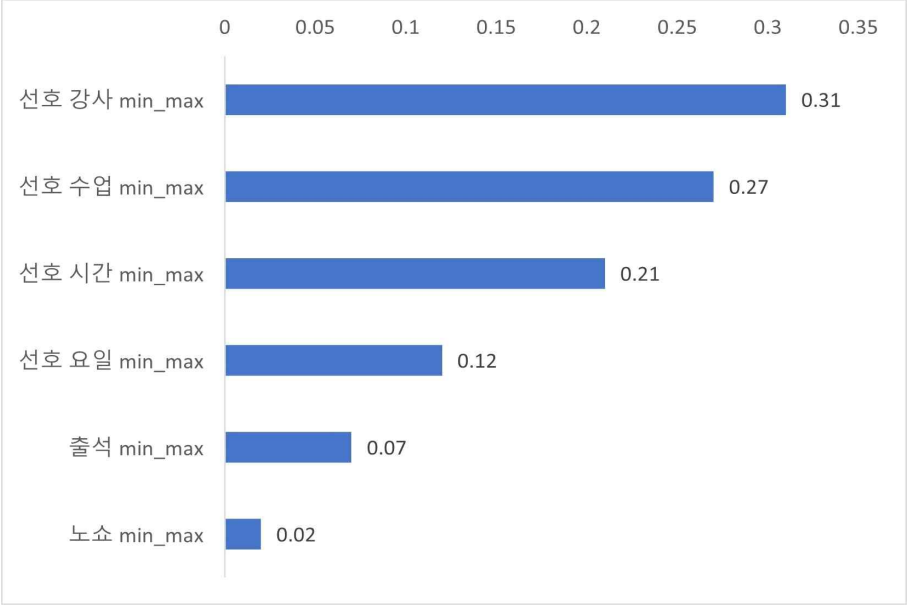
군집 2 (중간 회원군)	Training Data	122 (70%)	정확도 (Accuracy)	0.76	0.83	0.88
			F1 스코어 (F1 score)	0.73	0.84	0.86
			정밀도 (Precision)	0.75	0.86	0.88
			재현율 (Recall)	0.71	0.81	0.84
			특이도 (Specificity)	0.77	0.80	0.90
			AUC	0.77	0.80	0.90
	Test Data	53 (30%)	정확도 (Accuracy)	0.79	0.85	0.91
			F1 스코어 (F1 score)	0.80	0.86	0.91
			정밀도 (Precision)	0.85	0.89	0.91
			재현율 (Recall)	0.76	0.83	0.90
			특이도 (Specificity)	0.78	0.81	0.90
			AUC	0.79	0.81	0.91
군집 3 (우수 회원군)	Training Data	233 (70%)	정확도 (Accuracy)	0.81	0.86	0.90
			F1 스코어 (F1 score)	0.82	0.87	0.92
			정밀도 (Precision)	0.82	0.86	0.92
			재현율 (Recall)	0.83	0.87	0.93
			특이도 (Specificity)	0.78	0.79	0.88
			AUC	0.78	0.79	0.91
	Test Data	100 (30%)	정확도 (Accuracy)	0.84	0.88	0.92
			F1 스코어 (F1 score)	0.88	0.90	0.91
			정밀도 (Precision)	0.89	0.90	0.89
			재현율 (Recall)	0.88	0.91	0.92
			특이도 (Specificity)	0.80	0.82	0.91
			AUC	0.80	0.82	0.92

[그림 4-15]는 군집 1(취약 회원군)의 최종 선택 모형인 XGB에서 재등록에 영향을 주는 변수 중요도(Feature Importance)를 보여 주는데, 노쇼(0.28), 선호 강사 (0.25), 선호 수업(0.19) 순으로 나타났다. 군집 1(취약 회원군)에는 다른 변수에 비해 노쇼 min_max H값(평균 순위)이 높아 변수 중요도가 가장 높게 나타났다.



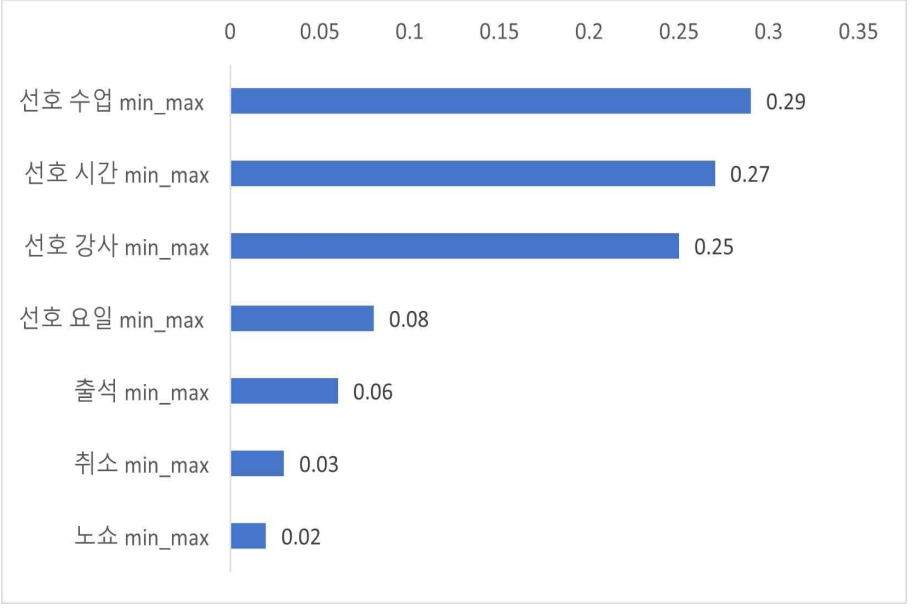
[그림 4-15] 군집 1(취약 회원군) XGB 모델 변수 중요도

[그림 4-16]은 군집 2(중간 회원군)의 최종 선택 모형인 XGB에서 재등록에 영향을 주는 변수 중요도(Feature Importance)를 보여 주는 데, 선호 강사(0.31), 선호 수업(0.27), 선호 시간(0.21) 순으로 나타났다. 군집 2의 경우 취소 min_max가 통계적으로 유의하지 않아 취소 min_max 변수를 제외하고 변수 중요도를 분석하였다. 군집 2에는 다른 변수에 비해 선호 강사 min_max H값(평균 순위)가 높아 변수 중요도가 가장 높게 나타났다.



[그림 4-16] 군집 2(중간 회원군) XGB 모델 변수 중요도

[그림 4-17]은 군집 3(우수 회원군)의 최종 선택 모형인 XGB에서 재등록에 영향을 주는 변수 중요도(Feature Importance)를 보여 주는 데, 선호 수업(0.29), 선호 시간(0.27), 선호 강사(0.25) 순으로 나타났다. 군집 3(우수 회원군)에는 다른 변수에 비해 선호 수업 min_max H값(평균 순위)이 높아 변수 중요도가 가장 높게 나타났다.



[그림 4-17] 군집 3(우수 회원군) XGB 모델 변수 중요도

단일 모형, 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군) 모두 정확도 등 XGB가 가장 뛰어나 전체적으로 예측 성능이 우수한 것으로 조사되었다. 의사결정나무가 대표적인 트리 기반 모델이지만 과적합 문제로 랜덤 포레스트가 선보였다. 랜덤 포레스트는 여러 개의 의사결정나무가 모인 bagging 앙상블 모델이다. XGB도 의사결정나무를 기반으로 하지만 boosting 앙상블 모델로 랜덤 포레스트와 차이점이 있다. boosting 방식은 한 개의 예측모델에 대한 error를 줄이는 방식이다. 예를 들어 첫 번째 의사결정나무가 있어 이를 통한 예측모델이 있을 때, 이것의 오차값을 다시 의사결정나무로 오차를 줄이고 그 오차를 똑같은 방식으로 줄이고 줄이는 방식이다. 근사분할 탐색 알고리즘, 희소성 고려 분할 탐색, 병렬 컴퓨팅, 자체적으로 규제를 추가하여 정확도를 높인다. XGB는 그래디언트 부스팅을 업그레이드한 모델로 2016년 캐글 대회에서 티엔치 첸이 우승한 뒤 XGB을 소개하면서 대표적인 인공지능 분석모델로 부각되었다.

군집별 재등록 여부 현황을 다음 [표 4-19]와 같이 정리하였다. K-means를 통해 최적 군집 3개로 나누어진 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군)이 단일에 비해 최종 모형로 선정된 XGB 모형의 정확도, F1 스코어, AUC가 높게 나타났다. 군집 분석이 각 데이터의 유사성을 측정하여 군집을 분류하므로 데이터를 유사성을 고려하지 않는 단일 군집에 비해 성능이 뛰어난 것으로 확인되었다.

전체 회원을 3개의 군집으로 나눈 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군)의 경우 군집 1(취약 회원군) 〈 군집 2(중간 회원군) 〈 군집 3(우수 회원군) 순으로 예측 성능이 우수한 것으로 보였다.

[표 4-19] 단일 및 군집별 XGB 모형 예측 성능 비교

구분 (회원수)	재등록 여부	회원 현황		XGB 모형		
		회원수 (명)	비율 (%)	정확도	F1 스코어	AUC
단일	재등록	580	50.0	0.87	0.87	0.86
	이탈	580	50.0			
	합계	1,160	100.0			
군집1 (취약 회원군)	재등록	261	40.0	0.90	0.88	0.87
	이탈	391	60.0			
	합계	652	100.0			
군집2 (중간 회원군)	재등록	86	49.1	0.91	0.91	0.91
	이탈	89	50.9			
	합계	175	100.0			
군집3 (우수 회원군)	재등록	233	70.0	0.92	0.91	0.92
	이탈	100	30.0			
	합계	333	100.0			

4.5.6 재등록 예측 확률

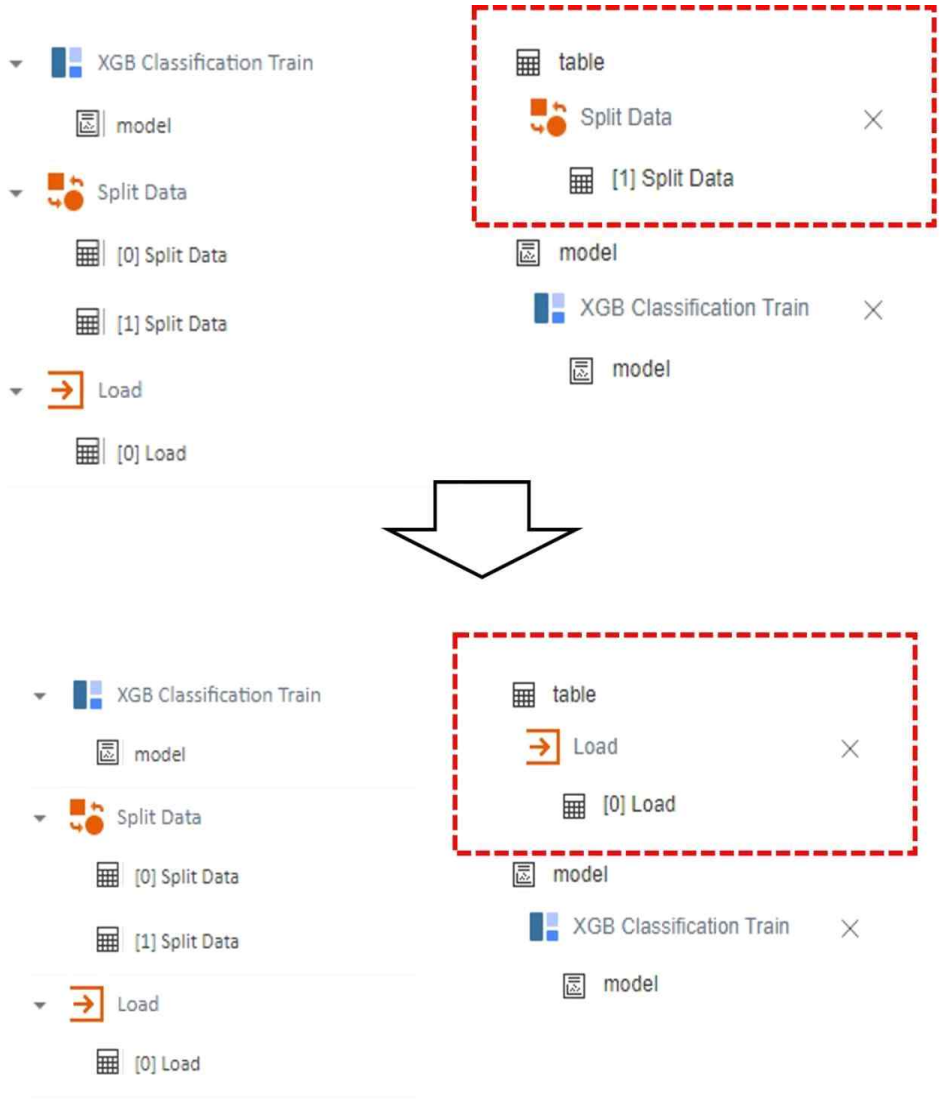
[그림 4-18]은 재등록 여부를 알고 싶은 데이터의 경우 K-means 군집 분석을 통해 군집을 분류 후 해당 군집에서 최적 예측모델인 XGB 모델을 적용하여 재등록 확률을 예측하였다.

[그림 4-18] XGB 모델 Flow

[그림 4-19]는 XGB 모델 신규 데이터 추가 Flow로 [그림 4-28] XGB 모델 Flow에 재등록 여부를 알 수 없는 신규 데이터를 별도 파일로 읽어서 “XGB Classification Predict”로 연결하였다.

[그림 4-19] XGB 모델 신규 데이터 추가 Flow

[그림 4-20]은 Test Data를 기존 split data 대신 재등록 여부를 알고자 하는 신규 데이터로 변경하는 작업을 진행하였다.



[그림 4-20] XGB 모델 입력 데이터 변경

상기 절차를 진행하여 재등록 여부를 알고 싶은 신규 데이터에 대해 재등록 확률을 [표 4-20]과 같이 확인하였다. 재등록(Re-registration)을 1로 구분하여 probability_1이 재등록 할 확률을 보여주는 데, E1 회원의 경우 probability_1값이 0.57로 재등록할 확률이 57%로 예측되었다.

[표 4-20] XGB 모델 재등록 예측 확률

User	Result	Attend	No show	Cancel	Teacher	Time	Class	Day	Prediction	Propability_0	Propability_1
E1		19	2	13	13	6	13	8	Re-registration	0.43	0.57
E2		18	2	12	11	6	12	8	Re-registration	0.32	0.68
E3		17	1	11	11	6	12	8	Re-registration	0.49	0.51
E4		50	4	40	15	22	32	25	Churn	0.75	0.25
E5		43	1	38	16	16	24	14	Re-registration	0.36	0.64
E6		39	1	35	13	15	20	13	Re-registration	0.18	0.82
E7		36	2	40	18	14	22	12	Re-registration	0.29	0.71
E8		35	3	32	12	14	18	12	Churn	0.84	0.16
E9		35	0	30	11	21	26	10	Re-registration	0.46	0.54
E10		22	5	3	15	11	14	12	Churn	0.81	0.19

4.5.7 개별 회원 수업 이력 시계열 예측

다음 [표 4-21]과 같이 개별 회원 수업 이력 Raw-data를 정리하였다.

[표 4-21] 개별 회원 수업 이력 Raw-data 예시

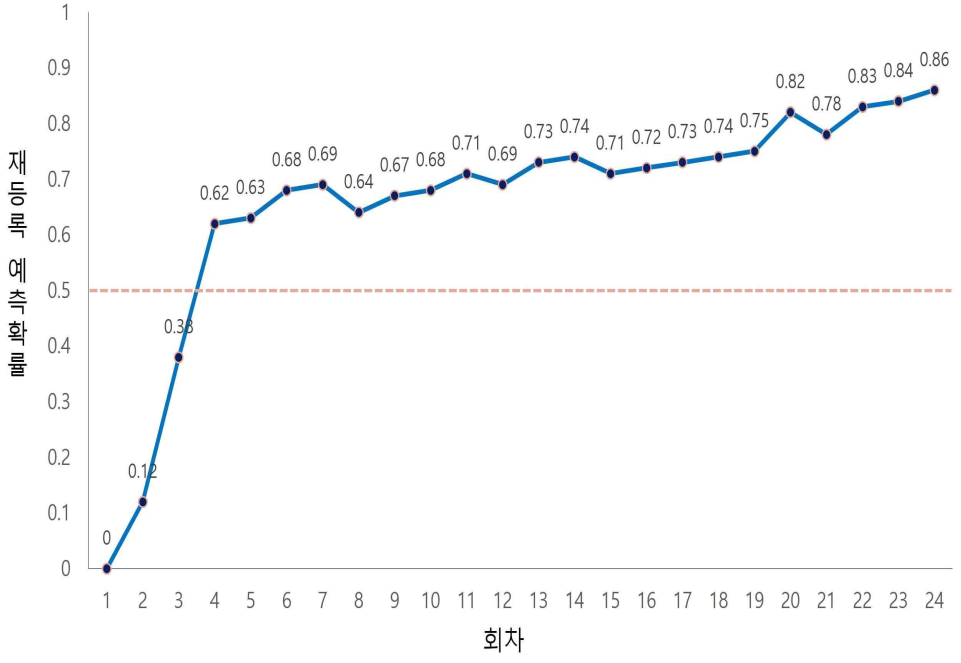
수업일	수업명	시간	강사	최종상태
2021-06-01	Basic	18:00	A	출석
2021-06-03	Basic	18:00	A	출석
2021-06-04	Basic	20:00	C	출석
2021-06-07	Basic	19:00	A	출석
2021-06-08	Basic	21:00	C	결석
2021-06-10	Basic	19:00	A	출석
2021-06-14	Basic	19:00	B	출석
2021-06-15	Lv.1	20:00	A	노쇼
2021-06-17	Lv.1	18:00	C	출석
2021-06-21	Lv.1	19:00	A	출석
2021-06-22	Basic	18:00	A	출석
2021-06-28	Basic	19:00	B	출석
2021-06-29	Lv.1	19:00	A	출석
2021-07-01	Basic	18:00	C	노쇼
2021-07-05	Basic	19:00	B	출석
2021-07-06	Basic	18:00	A	결석
2021-07-08	Lv.1	18:00	A	출석
2021-07-13	Basic	18:00	A	출석
2021-07-14	Basic	18:00	B	출석
2021-07-15	Basic	19:00	A	출석
2021-07-19	Lv.1	19:00	B	노쇼
2021-07-21	Lv.1	18:00	B	출석
2021-07-22	Basic	18:00	C	출석
2021-07-26	Lv.1	19:00	B	출석

시계열 분석을 위해 [표 4-22]와 같이 회차별 자료를 누적하여 재등록 확률 분석을 진행하였다.

[표 4-22] 개별 회원 수업 이력 시계열 작업

회차	출석	노쇼	선호 강사	선호 시간	선호 수업	선호 요일	재등록 확률
1	1	0	1	1	1	1	1.00
2	2	0	2	2	2	1	0.95
3	3	0	2	2	3	1	0.84
4	4	0	2	2	4	1	0.86
5	4	0	3	2	4	1	0.85
6	5	0	3	3	5	2	0.87
7	6	0	3	3	6	2	0.84
8	6	1	4	3	6	2	0.64
9	7	1	4	4	7	2	0.78
10	8	1	5	5	7	3	0.81
11	9	1	6	6	7	3	0.83
12	9	1	6	6	7	3	0.69
13	10	1	6	6	7	3	0.74
14	10	2	7	6	7	3	0.61
15	11	2	7	6	8	4	0.71
16	11	2	7	6	8	4	0.67
17	12	2	7	7	8	4	0.73
18	13	2	8	8	9	4	0.82
19	14	2	9	9	10	4	0.85
20	15	2	9	9	11	4	0.86
21	15	3	10	9	11	4	0.74
22	16	3	10	10	11	4	0.81
23	17	3	10	11	12	4	0.84
24	18	3	10	11	12	5	0.86

개별 회원 수업 이력 Raw-data를 예측 모형에 맞게 변수를 전처리 후 회차별 재등록 예측 확률을 시계열 도식화하면 다음 [그림 4-21]과 같이 확인하였다. 재등록 예측 확률 0.5를 기준으로 0.5 이상이면 재등록할 확률이 이탈할 확률보다 높을 것으로 수치로 예측할 수 있었다.



[그림 4-21] 개별 회원 시계열 재등록 예측 확률

특정 시점의 정적 데이터로는 재등록 예측 확률을 계산하기에 무리가 있어 과거부터의 누적 이력을 통해 재등록 예측 확률 경향을 파악할 수 있었다. 노쇼가 발생할 경우 재등록 예측 확률이 감소하므로, 노쇼가 일시적인 상황인지 여부를 종합적으로 판단할 필요가 있다. 재등록 예측 확률이 우하향으로 감소한다면 집중적인 관리가 필요하다. 노쇼 등 이상 신호에 대해 원인을 파악하고 적극적인 개선을 통해 시계열 데이터의 모니터링 효과를 배가할 수 있다.

4.6 마케팅 전략

4.6.1 군집별 타켓 마케팅 개요

군집 분석을 통해 총 3개의 군집으로 나누어 각 군집의 재등록, 이탈 회원 분포를 살펴보았다. 군집1(취약 회원군)은 652명에 대해 재등록 261명(40.0%), 이탈 391명(60.0%)이 분포하며, 군집 2(중간 회원군)는 175명에 대해 재등록 86명(49.1%), 이탈 89명(50.9%)이 분포하며, 군집 3(우수 회원군)은 333명에 대해 재등록 233명(70.0%), 이탈 100명(30.0%)이 분포하고 있다. DT, RF, XGB 모형 중 3개 군집 모두 XGB 모형의 예측 성능이 가장 우수하여 각 군집별 정확도, F1 스코어, AUC 지표의 성능을 정리하였다. 군집 1(취약 회원군) 〈 군집 2(중간 회원군) 〈 군집 3(우수 회원군) 순으로 예측 성능이 우수하다고 보였다. 군집 1(취약 회원군)의 주요 변수 중요도는 노쇼(0.28), 선호 강사(0.25), 선호 수업(0.19)이 군집 1(취약 회원군)에 영향을 주고, 군집 2(중간 회원군)의 주요 변수 중요도는 선호 강사(0.31), 선호 수업(0.27), 선호 시간(0.21)이 군집 2(중간 회원군)에 영향을 주고, 군집 3(우수 회원군)의 주요 변수 중요도는 선호 수업(0.29). 선호 시간(0.27), 선호 강사(0.25)가 군집 3(우수 회원군)에 영향을 주었다.

추가적인 고객의 데이터가 쌓이게 되면 기존 군집에 해당되는 지 판단하기 위해 몇 가지 방법을 사용할 수 있다. 첫째, 사전에 학습된 모델을 재사용하여 추가된 자료의 해당 군집을 알 수 있다. 둘째, 새로운 자료 포인트와 각 군집의 중심 사이의 거리를 측정하여 가장 가까운 군집을 찾을 수 있는 데, 거리가 가까운 군집에 해당하는 것으로 간주한다.

군집1(취약 회원군), 군집2(중간 회원군)의 고객을 군집3(우수 회원군)으로 이동할 수 있도록 하는 것이 가장 바람직한 마케팅 전략으로 개별 회원의 재등록 확률을 상시 시계열로 분석하고 peer group과의 비교를 통해 각 변수의 상대적인 위치를 파악할 수 있다. 취약한 변수에 대해 맞춤식 마케팅 전략을 집중적으로 시행하여 군집3(우수 회원군)으로 옮겨 오도록 관리한다.

각 군집별 맞춤식 마케팅 방안을 [표 4-23]과 같이 정리하였다.

[표 4-23] 군집별 타겟 마케팅 방안

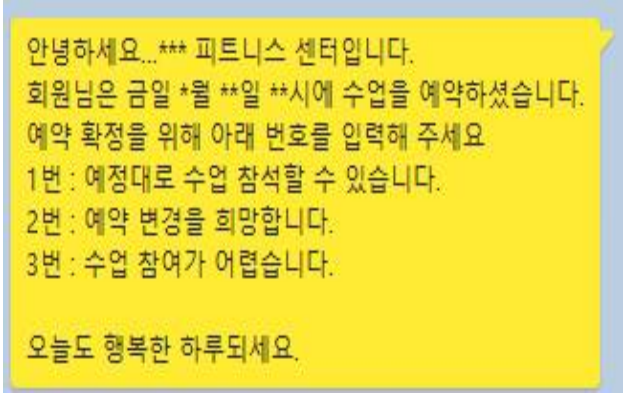
<table>
<tr><th>구분</th><th>군집1
(취약 회원군)</th><th>군집2
(중간 회원군)</th><th>군집3
(우수 회원군)</th></tr>
<tr><td>군집
분포</td><td>
<table>
<tr><th>구분</th><th>인원
(명)</th><th>비율
(%)</th></tr>
<tr><td>재등록</td><td>261</td><td>40.0</td></tr>
<tr><td>이탈</td><td>391</td><td>60.0</td></tr>
<tr><td>합계</td><td>652</td><td>100.0</td></tr>
</table>
</td><td>
<table>
<tr><th>구분</th><th>인원
(명)</th><th>비율
(%)</th></tr>
<tr><td>재등록</td><td>86</td><td>49.1</td></tr>
<tr><td>이탈</td><td>89</td><td>50.9</td></tr>
<tr><td>합계</td><td>175</td><td>100.0</td></tr>
</table>
</td><td>
<table>
<tr><th>구분</th><th>인원
(명)</th><th>비율
(%)</th></tr>
<tr><td>재등록</td><td>233</td><td>70.0</td></tr>
<tr><td>이탈</td><td>100</td><td>30.0</td></tr>
<tr><td>합계</td><td>333</td><td>100.0</td></tr>
</table>
</td></tr>
<tr><td>주요
변수
중요도</td><td>
<table>
<tr><th>노쇼</th><th>선호
강사</th><th>선호
수업</th></tr>
<tr><td>0.28</td><td>0.25</td><td>0.19</td></tr>
</table>
</td><td>
<table>
<tr><th>선호
강사</th><th>선호
수업</th><th>선호
시간</th></tr>
<tr><td>0.31</td><td>0.27</td><td>0.21</td></tr>
</table>
</td><td>
<table>
<tr><th>선호
수업</th><th>선호
시간</th><th>선호
강사</th></tr>
<tr><td>0.29</td><td>0.27</td><td>0.25</td></tr>
</table>
</td></tr>
<tr><td>마케팅
방안</td><td>· 예약확인 안내 문자
· 포인트 관리
· 연계 마케팅의 할인 혜택
· 재미와 사회적 교류를 통한 참여 유발</td><td>· 강사 대상 회원 응대 교육 강화
· 회원 중심의 서비스 디자인 씽킹 접근
· 쌍방향 커뮤니케이션 채널 활용
· 케어 디스크를 통한 맞춤식 관리</td><td>· 니즈에 맞는 다양한 프로그램
· 디지털 기기를 활용하여 수업 성과 관리
· 온라인 강의를 통한 홈트레이닝
· 시간 마케팅으로 회원의 수업 참여 확대</td></tr>
</table>

4.6.2 세부 마케팅 방안

군집1(취약 회원군)의 경우 노쇼 감소를 위해 크게 예약확인 안내 문자, 포인트 관리, 연계 마케팅의 할인 혜택, 재미와 사회적 교류를 통한 참여를 유발할 수 있다.

예약확인 안내 문자로 노쇼를 줄일 수 있다. 예약서비스에 대한 회원 인식 변화에 초점을 맞추어야 한다. 서비스를 이용해야 된다는 절박함보다는 예약을 통해 얻게 되는 심리적 혜택이 우선시되는 심리가 있다. 서비스의 절대적 가치에 대한 판단보다는 현시점에서의 상대적 가치가 높게 지각되는 요소에 의해 발생될 수 있다. 이용계획이 명확하지 않은 상황에서 무분별한 예약을 하게 되는 심리는 심리적 작용에서 비롯될 수 있다. 6명 이내의 소규모 그룹 수업을 하는 경우 많은 회원들의 참여를 유도하기 위해 회원별 주 3회로 이용횟수를 제한하는 것이 일반적이다. 주 3회 이용을 위해 무분별한 예약 후 노쇼가 아닌 실질적으로 수업 참여가 있는 예약을 하도록 회원들에게 인식시킬 필요가 있다. 회원들이 예약을 통해 받을 수 있는 다양한 혜택을 자신이 소유하고 있는 권리로 인지하게 하면 노쇼의 발생 확률이 낮아진다.

수업 당일 예약확인 안내를 보내는 방식이 필요하다. 콜 포비아 등 통화 자체에 부담을 가지는 경우가 많아 문자나 카카오톡 같은 SNS를 활용한 것이 선호된다.



[그림 4-22] 예약확인 안내 문자 예시

피트니스 센터의 경우 회원권을 선결제하며 별도 보증금을 두지 않아 금전적인 페널티보단 포인트 차감으로 직접적인 경제적인 손실은 없지만 심리적으로 손해라는 인식을 줄 필요가 있다. 항공사 마일리지 적립을 통해 공짜 항공권을 얻거나 카드 포인트를 현금 캐쉬백 받는 경우가 많아 포인트에 대해 사실상 현금으로 인식하고 있다. 노쇼로 인해 이러한 포인트를 차감된다면 현금 손실로 생각할 수 있어 노쇼를 줄이는 데 일조하게 된다.

연계 마케팅의 할인 혜택으로 참여를 적극 유도할 수 있다. 피트니스 센터를 방문하는 회원들은 체형관리를 위해 다이어트 식단에 관심이 많다. 기름진 음식보단 신선한 샐러드 중심으로 식단을 원하는 데, 어떤 메뉴의 식단을 얼마나 자주 먹어야 하며 경제적 부담을 줄이는 것에 고민을 하게 된다. 다이어트 식단 전문기업과 연계하여 체계적인 식단 관리를 저렴하게 제공할 수 있다. 피트니스 센터 인근 샐러드 전문점 방문시 할인 혜택으로 골목상권 상생의 기회를 가져올 수 있다. 출석 등 이용실적에 따라 적립한 포인트를 제휴매장에서 사용할 수 있게 하여, 노쇼를 하면 할인 혜택을 받지 못한다는 생각이 들게 할 수 있다.

재미와 사회적 교류를 통한 참여 유발할 수 있다. 피트니스는 오락성을 중심으로 하는 산업이라 할 수 있다. 직접 피트니스를 하는 목적이 건강을 증진하고 재미를 느끼고 싶은 욕구를 충족하기 위함이다. 가령 MZ세대들이 즐기는 할로윈(Halloween) 등 이벤트를 통해 회원들과 색다른 경험을 주는 재미를 겸비할 수 있다. 강사와 회원들이 분장을 하거나 드레스 코드를 같이하는 등 신나게 웃으면서 즐기는 추억을 만들 수 있다. 이러한 이벤트를 공지하여 평소 노쇼로 출석이 저조한 회원들에게도 센터 활동을 안내하고 센터 방문을 유도할 수 있다.

군집2(중간 회원군)의 경우 선호 강사, 선호 수업에 대해 관심을 높이기 위해 크게 강사 대상 회원 응대 교육 강화, 회원 중심의 서비스 디자인 씽킹 접근, 쌍방향 커뮤니케이션 채널 활용 그리고, 케어 디스크를 통한 맞춤식 관리가 필요하다.

강사 대상 회원 응대 교육을 강화해야 한다. 교육을 통해 기본적인 응대는 빠른 시간 내에 보여줄 수 있지만 기계처럼 매뉴얼만 따라 한다면 회원들은

거부감을 가질 수 있다. 교육을 통해 사람이 전달할 수 있는 '진심'에 대해 강조하며 진정성 있는 마음으로 고객을 대하는 태도를 가지도록 한다. 전화상담이나 대면 상담 이후에는 "Happy Call"이라 불리는 진심을 담은 전화 또는 문자 메시지를 통해 관계를 이어나갈 필요가 있다.

회원 중심의 서비스 디자인 씽킹 접근을 검토해야 한다. 다양한 분야의 방법과 도구를 사용해 서비스를 디자인할 때 수반되는 사용자 중심의 사고방식으로 '고객 형태를 이해하고 설명할 수 있는 분석 능력'과 '고객의 구매 경로 전반에 걸친 콘텐츠 기획 및 환경 조성'이 필요하다.

쌍방향 커뮤니케이션 채널을 활용할 수 있다. 최근 지능형 AI 챗봇 서비스가 가능하여 End-to-End 서비스를 제공하고 있다. 메신저의 특성상 실시간 응대가 중요한 데, 회원들의 문의에 원활히 대응하기 위해, 상담원의 실시간 응답이 어려울 때는 챗봇을 통해 과거 상담 이력을 바탕으로 고객 편의성을 확보할 수 있다.

케어 디스크를 통한 맞춤식 관리를 시행하면 효과적이다. 각자 불편한 부위가 상이한 상황에서 수업을 임할 때 강사와 회원 간의 사전 정보 공유가 필요한 데 매번 말을 하기엔 무리가 있다. 케어가 필요한 부위에 대해 미리 디스크를 해당 기구에 부착하여 커뮤니케이션을 강화할 수 있다. 해당 강사는 회원들의 몸 상태를 미리 알 수 있어 무리한 동작을 하지 않을 수 있어 회원들의 수업 부적응을 해소할 수 있다. 회원들의 몸 상태에 따라 맞춤식 지도가 가능하여 만족도를 높이는 데 기여한다.

군집3(우수 회원군)의 경우 선호 수업, 선호 시간에 대해 관심을 높이기 위해 크게 니즈에 맞은 다양한 프로그램, 디지털 기기를 활용하여 수업 성과 관리, 온라인 강의를 통한 홈 트레이닝, 시간 마케팅으로 회원의 수업 참여 확대를 손꼽을 수 있다.

다양한 통증과 원하는 운동 부위가 있어 매번 똑같은 수업을 진행하면 참여가 저조할 수 있다. 요일별 특화된 프로그램을 통해 회원들의 니즈에 적극 대응할 수 있다.

[표 4-24] 요일별 특화된 프로그램

요일	월요일	화요일	수요일	목요일	금요일	토요일
프로그램	복근	애플힙	하체강화	힐링	오프숄더	랜덤

3D 체형분석기 모티 피지오(Moti Physio)는 3장의 사진 촬영(앞,옆,뒤)으로 24개의 관절 포인트와 87가지 근육을 분석하여 내 몸의 불균형 부위, 약화된 근육들을 과학적이고 정확하게 분석할 수 있는 기계이다. 1회 촬영으로 끝나지 않고 지속적으로 촬영을 진행하여 전후 비교를 한 눈에 쉽게 볼 수 있어 회원들의 지속적인 센터 방문을 유도할 수 있다. 운동 전후 상태를 회원이 체감하며 기기에서 놓치기 쉬운 부분을 강사의 휴먼 터치를 추가한다면 피트니스 센터의 만족도를 높일 수 있다.

오프라인 수업의 보완재로서 온라인 강의를 통한 홈 트레이닝을 고려할 수 있다. 최근 코로나-19의 영향으로 일상생활에서 비대면 활동이 널리 퍼지게 되었다. 온라인 강의 시스템을 통해 실시간으로 강사의 동작을 따라 하며 회원들과 소통하는 온라인 수업을 적극적으로 할 수 있다. 센터를 직접 방문하지 않고 집에서 모니터를 보고 동작을 따라 할 수 있고, 잘못된 동작에 대해 실시간 피드백을 받을 수 있어 일방적인 동영상 수업에 비해 참여도를 높이게 된다. 또한 다른 회원들의 함께 온라인 수업을 들으면서 동기를 부여할 수 있다.

시간 마케팅으로 회원의 수업 참여 확대할 수 있다. 바쁜 일상 속에 피트니스 센터 이용 접근을 용이하게 하기 위해 점심시간 12:00시에 대해 수업 종료 후 샌드위치, 음료수를 무료로 제공하는 것으로 유인할 수 있다. 시간 마케팅은 정해진 시간에 정기적으로 진행하기 때문에 고객 충성도를 높이는데 효과적이다.

Ⅴ. 결론

5.1 연구결과 요약

본 연구는 피트니스 센터 회원들의 평소 이용실적을 분석하여 군집 분석, 재등록 예측을 활용하여 고객 관리 모델을 제안하였다. A 피트니스 프랜차이즈 N 지점을 조사 기간(2020년 2월 ~ 2022년 1월/24개월)에 걸쳐 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 취합하였다. 재등록 여부를 예측하는 인공지능 알고리즘 성능 비교를 위해 재등록과 이탈 회원 수를 각 580명씩 같은 숫자로 분석하였다. 재등록/이탈고객의 회원 수가 다를 경우 즉 데이터가 불균일한 분포를 가지면 모델의 학습이 제대로 이루어지지 않을 확률이 높다. 데이터의 불균형 상태에 따라 정밀도가 감소하고 재현율이 증가하는 등 문제가 생길 수 있어, 데이터의 불균형을 해소하고자 같은 수의 데이터를 적용하였다.

피트니스 분야 선행 문헌 연구를 통해 총 7개 변수로 출석, 노쇼, 취소, 선호 강사, 선호 수업, 선호 시간, 선호 요일을 선정하였다. 범주형 변수인 선호 강사, 선호 수업, 선호 시간, 선호 요일에 대해 Frequency encoding을 통해 수치형 변수로 변환하였다. 자료의 정규성 여부를 우선 검토해야 하는 데, 크게 왜도/첨도, KS test(Kolmogorov-Smirnov test)로 판단할 수 있었다. 정규성 판단 기준에 부합되지 않아 정규분포(가우시안 분포)를 하지 않는 것으로 확인되어 비모수 통계를 이용해서 분석했다. 표본의 크기가 충분히 크다면 표본평균의 분포는 근사적으로 정규분포를 따른다는 중심극한정리에 따라 모집단의 정규성을 가정하고 분석한다. 하지만, 정규분포를 가정하고 일반적인 통계 분석 방법을 사용하여 잘못된 결과를 얻는 등 분석의 신뢰도가 떨어질 수 있어, 비모수 검정을 진행하여 해당 변수가 재등록 여부에 의미가 있는지 확인하였다.

피처(Feature)는 데이터의 값을 잘 예측하기 위한 데이터의 특징으로, 피처 엔지니어링(Feature Engineering)은 가공되지 않은 데이터를 인공지능 모

델에 사용할 수 있는 기능으로 전환하는 과정이라 할 수 있다. 특정 데이터 값에 따라 편향성을 갖고 전체 값과 최근 값 범위가 달라 그대로 해석하기에 무리가 있어 추가 작업이 필요하다. 피처 스케일링((Feature Scaling)은 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업으로, 본 연구의 기술통계 검토 결과 정규분포를 하지 않아 최소-최대 스케일링을 진행하였다. 군집분석의 경우 관측값 사이에서 유사성을 찾아 군집을 형성하는 데, 넓은 범위를 가지는 변수는 군집에서 더 큰 영향을 줄 수 있어, min_max값을 기반으로 후속 작업을 실시하였다.

비모수 검정으로 χ^2 (chi-square) 검정을 대신하여 Mann-Whitney U 검정을 사용하였다. 각 변수를 min_max 변환 후 Mann-Whitney U 검정 값에 대해 유의확률을 확인 결과 취소 min_max를 제외하고 다른 변수 min_max 분포는 귀무가설을 기각하여 재등록 범주에 대해 통계적으로 유의한 것으로 파악되었다. 군집분석에서의 Kruskal-Wallis H 검정에서 검토 결과 군집 2의 취소 min_max의 분포가 재등록 여부의 범주에서 같다는 귀무가설에 대해 유의확률이 0.650으로 유의수준 0.05보단 큰 값으로 귀무가설을 채택하여 통계적으로 유의하지 않는 것을 확인할 수 있다. 따라서, 유의확률이 0.05보다 작은 군집 1, 군집 3의 경우 취소 min_max는 통계적으로 유의하여 취소 min_max 변수를 적용하며, 단일 집단, 군집 2는 취소 min_max가 통계적으로 유의하지 않아 제외하고 분석함이 바람직하다.

군집 분석(Clustering)은 데이터 안에 존재하는 의미있는 그룹을 찾아내는 작업이다. K-평균 군집화는 K값을 다르게 바꾸어가면서 군집화했을 때, 군집의 수 K별 “군집 내 군집과 개체 간 거리 제곱합의 총합(Inertia / Total within-cluster sum of squares)”의 그래프가 팔꿈치 모양으로 꺾이는 지점의 K를 최적 군집의 개수를 선택하는 팔꿈치 방법에서 군집의 개수가 3개일 때, 팔꿈치 모양으로 꺾이고, 군집 K가 4개 이상일 때부터는 Inertia의 변화가 매우 작으므로, 최적 군집의 개수를 3개로 하는 것이 바람직하다. 실루엣(silhouette) 계수를 비교하면 K=3일 때, 실루엣 계수가 가장 크기에 3개 군집으로 나누었다.

3개 이상의 집단에 대해 차이가 있는 지 없는 지를 알 수 있는 분석방법

으로 크게 모수적 분석방법인 ANOVA(Analysis of Variance)와 비모수적 분석방법인 Kruskal-Wallis H 검정으로 나눌 수 있다. 본 연구는 비모수적 분석을 진행하므로, ANOVA 분석 대신 Kruskal-Wallis H 검정을 통해 3개 집단에 대해 비교하였다. 비모수 분석인 Kruskal-Wallis H 검정 결과 모두 유의확률이 0.000으로 귀무가설을 모두 기각하였다. 즉, 통계적으로 유의한 차이가 있다고 볼 수 있다. 군집별 세부 분석을 할 때, 군집 1, 군집 3의 경우 취소 min_max는 유의확률이 0.05보다 작아 통계적으로 유의하여 취소 min_max 변수를 적용하며, 군집 2는 취소 min_max가 유의확률이 0.650으로 통계적으로 유의하지 않아 취소 min_max 변수를 제외하고 진행하였다.

군집 1은 총 7개 변수에 대해 노쇼 min_max는 다른 군집에 비해 높은 값, 취소 min_max는 군집2와 군집3의 중간 값을 나타났다. 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 취약한 회원 군집이라 볼 수 있어 군집 1은 “취약 회원군”으로 이름 지었다. 군집 2는 총 7개 변수에 대해 속성값 H값(평균 순위)이 중간 위치로 총 3개 군집 중에서 중간 회원 군집이라 할 수 있다. 따라서 군집2는 “중간 회원군”으로 이름 지었다. 변수 군집 3은 노쇼 min_max를 제외하고 다른 군집에 비해 높은 값이 나타났다. 종합적으로 변수 속성을 비교하면 총 3개 군집 중에서 가장 우수한 회원 군집이라 볼 수 있어 군집 3은 “우수 회원군”으로 이름 지었다.

본 연구에서는 선행연구에서 진행한 기존 재등록(이탈) 예측 연구의 한계를 해소하고자 크게 2가지 시나리오로 총 3가지 인공지능 알고리즘을 통해 최적의 재등록을 예측하였다. 트리 기반의 인공지능 알고리즘 중 기본적인 의사결정나무(DT : Decision Tree)와 과적합을 방지하기 위해 개선된 랜덤 포레스트(RF:Random Forrest), XGB(eXtreme Gradient Boosting)를 활용하였다.

의사결정나무(DT), 랜덤포레스트(RF), 극한 경사도 부스팅 (XGB)에 대해 공통적으로 학습용 데이터(Training data) 70% / 평가용 데이터(Test data) 30%로 분리하였다. 교차검증(cross validation)이라 할 수 있다. 데이터를 여러 번 반복해서 나누어 여러 모델을 학습하여 성능을 평가하는 데, 일반적으로 사용하는 k값을 5로 지정하여 분할하였다.

재등록 예측 성능을 평가하고자 총 6가지 지표인 정확도(Accuracy), 정밀

도(Precision), 재현도(Recall), 특이도(Specificity), F1 스코어(F1 Score), AUC(Area Under the ROC Curve)를 이용하였다. 시나리오 1인 단일 재등록 예측모형은 전체 고객을 대상으로 성능을 평가하였으며, XGB가 가장 좋은 성능을 보였다. 재등록에 영향을 주는 변수 중요도로 선호 강사 min_max(0.29), 선호 수업 min_max (0.24), 노쇼 min_max(0.21) 순으로 나타났다. 시나리오 2는 군집별 재등록 예측모형으로 3개 군집으로 세분화하여 성능을 평가하였다. 단일 모형, 군집 1(취약 회원군), 군집 2(중간 회원군), 군집 3(우수 회원군) 모두 DT 〈 RF 〈 XGB 순으로 성능이 우수한 것으로 확인되었다. XGB가 boosting 앙상블 모델로 한 개의 예측 모델에 대해 error를 줄이는 방식이며, bagging 앙상블 모델인 RF와 차이가 있다. 캐글 대회에서 XGB가 우승한 뒤 최근 대표적인 인공지능 분석모델로 부각되고 있다. 선행연구인 Sobreiro(2021)에서 사용한 의사결정나무, 랜덤 포레스트, XGB 모형을 비교하면 군집 3(우수 회원군)과 비슷한 예측 성능이 조사되었다. 전반적으로 의사결정나무 〈 랜덤 포레스트 〈 XGB 순으로 성능이 개선된 것으로 파악되었다. 이는 의사결정나무의 과적합을 방지하기 위한 랜덤 포레스트 배깅(bagging), XGB 부스팅(boosting) 앙상블 모델의 성능 개선 효과를 알 수 있었다.

군집1(취약 회원군), 군집2(중간 회원군), 군집3(우수 회원군)의 경우 군집 1(취약 회원군) 〈 군집 2(중간 회원군) 〈 군집 3(우수 회원군) 순으로 예측 성능이 우수한 것으로 보였다.

군집 1(취약 회원군)의 변수 중요도는 노쇼 (0.28), 선호 강사 (0.25), 선호 수업(0.19) 순으로 나타났다. 군집 2(중간 회원군)의 변수 중요도는 선호 강사(0.31), 선호 수업(0.27), 선호 시간(0.21) 순으로 나타났다. 군집 3(우수 회원군)의 변수 중요도는 선호 수업(0.29), 선호 시간(0.27), 선호 강사(0.25) 순으로 나타났다.

재등록 여부를 알 수 없는 회원들의 재등록 확률을 계산하여 고객 관리에 활용할 수 있다. 개별 회원 수업 이력을 회차별 누적하여 시계열 분석이 가능하다. 재등록 예측확률 0.5를 기준으로 0.5 이상이면 재등록할 확률이 이탈할 확률보다 높을 것으로 수치로 예측할 수 있다. 재등록 예측 확률이 우하향으

로 감소한다면 집중적인 관리가 필요하다. 노쇼 등 이상 신호에 대해 원인을 파악하고 적극적인 개선을 통해 시계열 데이터의 모니터링 효과를 배가할 수 있다.

군집별 타겟 마케팅전략을 통해 이탈을 방지하고 재등록으로 연계하게 할 수 있다. 예약확인 안내 문자, 포인트 관리, 연계 마케팅의 할인 혜택, 재미와 사회적 교류를 통한 참여 유발, 강사 대상 회원 응대 교육 강화, 회원 중심의 서비스 디자인 씽킹 접근, 쌍방향 커뮤니케이션 채널 활용, 케어 디스크를 통한 맞춤식 관리, 니즈에 맞는 다양한 프로그램, 디지털 기기를 활용하여 수업 성과 관리, 온라인 강의를 통한 홈 트레이닝, 시간 마케팅으로 회원의 수업 참여 확대 등 각 군집의 변수 중요도를 고려하여 마케팅 방안을 실시할 수 있다.

5.2 학문적 및 실무적 시사점

본 연구에서는 피트니스 센터 소규모 운동(GX) 회원 정보를 이용하여 고객 세분화 및 재등록 예측을 통해 군집 특성에 맞는 맞춤식 마케팅 방안을 제시하였다. 국내 피트니스 센터 회원을 대상으로 실증분석을 통해 학문적인 발전과 국내 피트니스 산업 혁신에 기여하고자 하였다. 본 연구의 학문적, 실무적 시사점을 다음과 같이 제시하였다.

5.2.1 학문적 시사점

본 연구는 국내 피트니스 센터를 이용하는 고객 데이터를 기반으로 고객 관리에 관한 연구로 기존의 선행연구들과 비교하면 다음과 같이 학문적 시사점을 얻을 수 있다.

첫째, , 피트니스 분야 고객 세분화 및 재등록 예측에 대해 복합적인 인공지능 모델의 선행연구에 비해 고도화된 분석을 진행한 연구이다. 유통 분야, 통신 분야, 금융 분야 등에서는 고객 세분화 및 재구매 예측을 인공지능 모델을 이용하여 진행하였지만 피트니스 분야는 주로 만족도 설문 방식을 해 왔다. 피트니스 분야에서도 고객 데이터를 활용하여 인공지능 분석 영역을 확대하였다.

둘째, 비모수 통계기법을 이용하여 정규분포가 아닌 데이터에 대해 통계적으로 접근하였다. 모집단이 정규분포를 보이지 않는 경우에 있어 Mann-Whitney U 검정, Kruskal-Wallis H 검정을 활용하여 통계적 유의성을 검토하였다. 비모수 검정을 진행하여 해당 변수가 재등록 여부에 의미가 있는지 확인하였다.

셋째, 교차검증(Cross Validation)을 통한 특정 데이터셋의 과적합을 방지하고 일반화된 모델 생성이 가능하다. 기존의 선행연구는 훈련 데이터와 학습 데이터를 7:3 또는 8:2로 분리만 진행하여 과적합의 가능성이 존재하였다. 본 연구는 5겹 교차검증으로 데이터 분리에 따른 성능 문제를 해소하였다. 훈련 데이터 불균형 분포가 있을 경우 결과가 왜곡될 수 있다. 오버 샘플링, 언

더 샘플링을 통해 조절하지만 생성된 데이터가 다른 클래스의 데이터와 중첩되어 노이즈가 생성할 수 있다. 본 연구에서는 재등록/이탈 데이터 수를 같이 하여 데이터 불균형으로 인한 문제를 해소하였다.

넷째, 지도학습/비지도학습 간의 유기적인 연계를 손꼽을 수 있다. 비지도학습인 군집 분석을 실시하여 적절한 고객 세분화를 진행하였다. 지도학습으로 의사결정나무, Random Forest, XGB를 활용한 재등록 예측 분류를 하여 유기적인 분석을 시도하였다. 기존 연구는 Random Forest, XGB 등 인공지능 모형 예측 성능 비교에만 끝나 학문적인 의미만 부여하는 경우가 많고, 모형의 단편적인 개별 연구에 그치고 있다. One Source Multi Use 전략으로 입체적 분석이 향후 후속 연구들에 도움이 될 것으로 기대한다.

끝으로, 기존 피트니스 분야 연구는 독립변수와 종속변수 간의 통계 유의성 여부를 살펴 어떤 요인이 종속변수인 재등록(이탈)에 얼마나 영향을 주는지를 수치적으로 확인이 어려웠다. 본 연구는 다양한 인공지능 알고리즘을 적용하여 성능을 비교한 결과 XGB가 가장 우수하였으며, XGB의 변수 중요도(Feature Importance)를 통해 재등록에 영향에 주는 피처의 상대적인 중요도를 정량적으로 검토하여 향후 지표 관리에 활용할 수 있다.

5.2.2 실무적 시사점

첫째, 평소 고객 DB를 이용하여 자료 접근성을 개선하였다. 피트니스 분야 재등록 연구는 만족도 설문방식을 진행하는 데, 매번 회원들의 이용 패턴과 재등록 의도를 설문으로 파악하기에는 어려움이 존재하였다. 그래서 유통분야, 통신 분야, 금융 분야 등 고객 정보를 체계적으로 관리하는 분야에서는 설문방식보단 고객 DB를 분석하고 있으며, 피트니스 분야에서도 데이터 경영의 기반을 조성하였다.

둘째, 고객의 특성을 고려한 고객 세분화를 통해 목표 고객을 집중 관리할 수 있다. 인구통계학적인 고객 세분화로는 실제 고객 이용 패턴을 파악하기 어렵다. 행동 특성을 고려한 고객 세분화를 진행하고 재등록 예측 분석이 효과적이다.

셋째, 실무 이용에 필요한 insight를 얻기 위해 첨단 기술 방법을 활용하였다. 선행 연구가 기술적인 성능 분석 결과에만 주력하였다면, 마케팅 관점에서 이용할 수 있는 참고자료로 이용하여 재등록예측 확률, 타겟 마케팅 등 실질적인 고객 관리에 초점을 맞추었다. 군집별 마케팅전략을 세분화하여 유한한 마케팅 역량을 적절히 활용해 성과를 낼 수 있는 기반을 제공하였다.

끝으로, 소상공인 육성 정책을 수립 및 운영하는 정부 기관에 효과적인 가이드가 될 수 있다. 정책자금 지원의 경우 예산의 한정으로 실제 혜택을 볼 수 있는 소상공인의 수가 제한적이다. 소상공인진흥공단 등 정부기관에서 각 업종에 맞는 재등록(재구매) 예측 알고리즘을 개발하여 보급한다면 지속가능한 성장의 기반이 조성될 수 있다.

5.3 연구의 한계 및 향후 연구방향

본 연구는 피트니스 센터 고객 관리에 대한 이론적 적응 및 실증을 통해 학문적, 실무적인 시사점을 제공하고 있다는 점에서 의의가 존재한다. 본 연구의 한계점과 향후 과제를 제시하면 다음과 같다.

첫째, 데이터 분석에 의한 연구의 공통적인 한계점은 데이터 수라고 할 수 있다. 피트니스 센터 소규모 운동(GX)에 참여한 재등록과 이탈이 확인된 회원 1,160명의 이력 513,530건에 대해 데이터를 근거하였으나, 보다 나은 일반화를 위해 더 많은 데이터를 확보할 필요가 있다. 더 많은 데이터가 확보되면 모집단이 정규분포가 될 수 있어 일반적인 모수적 통계기법으로 해석이 가능할 수 있다.

둘째, 특정 피트니스 센터 소규모 운동(GX)에 참여한 회원을 대상으로 분석하여 다른 피트니스 센터 소규모 운동(GX)에 그대로 적용하기는 어려움이 있다. 실증적 연구 결과의 일반화를 위해 다양한 피트니스 센터 소규모 운동(GX)에 참여한 회원으로 범위를 확대한다면 보다 가치 있는 연구가 될 것으로 기대된다.

셋째, 인공지능 분석을 위해 피처 엔지니어링을 진행하면서 모델을 보다 효과적으로 설명할 수 있는 피처를 활용해야 한다. 다양한 선행연구 분석으로 적절한 피처를 확인하며, 군집 특성을 고려할 수 있는 피처에 대한 심도 깊은 사례 연구가 필요하여 따라서 향후 연구에는 이를 반영하여 연구를 수행해야 하겠다.

참 고 문 헌

1. 국내문헌

강민지. (2023). "성인발레학원의 수업만족도 및 신체적 자기효능감이 재등록 의사에 미치는 영향". 한양대학교 대학원 석사학위논문.

강성식. (2021). "텍스트 분석과 기계학습 접근법을 활용한 건설업 재해분석". 부경대학교 대학원 박사학위논문.

강용수. (2017). “다차원척도법을 이용한 영화관 예약부도 개선방안 연구”. 중앙대학교 대학원 석사학위논문.

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경. (2006). 『고객관계관리(CRM)를 위한 데이터마이닝 방법론』. 서울: 자유아카데미.

게세홍. (2023). "태권도장의 서비스품질이 학부모 신뢰도 및 미래 소비행동에 미치는 영향". 우석대학교 대학원 석사학위논문.

고몽선. (2023). “선양시 태권도 지도자의 지도유형이 수련생의 훈련효과와 재등록의도에 미치는 영향”. 명지대학교 대학원 박사학위논문.

고보라, 허진. (2019). 피트니스센터의 고객지향성, 고객만족, 고객신뢰 및 충성도에 관한 연구. 『한국융합학회 논문지』, 10(2), 247-254.

권시현. (2022). 『데싸노트의 실전에서 통하는 머신러닝』. 서울: 골든래빗.

권오현, 박동현, 조신. (2021). 통합 로열티 프로그램 앱 로그 데이터를 이용한 이탈 고객 예측. 『한국정보기술학회논문지』, 19(8), 107-118.

권철민. (2022). 『파이썬 머신러닝 완벽 가이드』. 경기: 위키북스.

김경태, 이지형. (2018). 딥 러닝과 Boosted Decision Tree를 활용한 고객 이탈 예측 모델. 『한국지능시스템학회 논문지』, 28(1), 7-12.

김나현. (2019). "생존 분석을 활용한 모바일 온라인 게임 사용자 이탈 예측". 이화여자대학교 대학원 석사학위논문.

김낙한. (2022). “국내 MZ세대의 웰니스 라이프스타일과 숙박상품 선택속성에 따른 시장세분화 연구”. 경희대학교 대학원 석사학위논문.

김담희, 안가경. (2018). 머신러닝을 이용한 고객세분화에 관한 연구. 『융복합지식학회논문지』, 6(2), 115-120.

김도희. (2015). "스포츠센터의 고객관계관리(CRM) 요인과 고객만족, 고객신뢰 및 고객충성도의 관계 연구". 경희대학교 대학원 박사학위논문.

김동균. (2022). "대중제골프장 내장객의 충성도에 따른 서비스 경쟁력 강화 방안". 경희대학교 대학원 박사학위논문.

김두일. (2018). “태권도 수련생 학부모의 도장선택요인이 고객만족, 구전의도 및 재등록에 미치는 영향”. 조선대학교 대학원 석사학위논문.

김명중. (2018). "혁신 융합서비스의 시장 성숙도에 따른 고객이탈 행동에 관한 연구 -IPTV 서비스를 중심으로-". 이화여자대학교 대학원 박사학위논문.

김미숙. (2015). "스포츠센터의 인적서비스품질과 서비스회복이 고객충성도에 미치는 영향". 부산대학교 대학원 박사학위논문.

김민수, 정우석. (2019). 피트니스센터의 경쟁요인이 고객만족 및 재이용의사에 미치는 영향. 『한국콘텐츠학회 논문지』, 19(6), 632-643.

김민창. (2023). “행동 시퀀스 트랜스포머를 이용한 구독 비즈니스의 고객 이탈 예측”. 서울대학교 대학원 석사학위논문.

김봉경. (2021). 태권도장 선호요인이 만족을 통한 재등록의도에 미치는 영향. 『한국체육과학회지』, 30(2), 587-600.

김송은. (2016). "Zumba Fitness 참여자가 지각한 서비스품질이 전환장벽 및 고객충성도에 미치는 영향". 원광대학교 대학원 박사학위논문.

김수래. (2015). "고객 거래 데이터를 활용한 마케팅 자원 최적 재배치 방법론 개발". 한성대학교 대학원 석사학위논문.

김승수. (2018). "비정형 정보와 CNN 기법을 활용한 고객 행태 예측: 전자

상거래 사례를 중심으로". 이화여자대학교 대학원 박사학위논문.

김영교. (2018). "스포츠센터 입지 및 서비스품질이 이용만족도에 미치는 영향". 용인대학교 대학원 석사학위논문.

김용세. (2020). 『(디지털 트랜스포메이션 시대의) 고객경험 서비스 디자인 씽킹』. 서울: 생능출판.

김용운. (2022). “초등학생의 태권도장 서비스가치 인식이 재등록의사에 미치는 영향 : 운동열정과 수련만족의 매개효과”. 수원대학교 대학원 박사학위논문.

김윤진. (2005). "K-평균 군집분석 재현성 평가에 관한 연구". 중앙대학교 대학원 석사학위논문.

김은미. (2020). "국내 주택시장의 주택 보유기간 및 매도 의사결정에 대한 머신러닝 예측모델 비교". 이화여자대학교 대학원 박사학위논문.

김은지. (2019). "교사 딥러닝 기반 이상탐지 모델과 금융, 제조, 미디어 분야에의 응용". 서울대학교 대학원 박사학위논문.

김응용. (2016). "상업스포츠센터 경험마케팅이 고객태도, 고객만족 및 구매의사 결정에 미치는 영향". 건국대학교 대학원 박사학위논문.

김재엽. (2020). "고객이탈 예측 모델링 기반 기대수익 최적화 방안 : 고객이탈 방지를 위한 마케팅 비용 최적화". 서울대학교 대학원 석사학위논문.

김지승. (2020). “패키지 여행상품 이용객의 선택속성에 따른 시장세분화 연구”. 경희대학교 대학원 석사학위논문.

김지영, 박정주, 이지영, 김은혜. (2020). 다중근거이론(Mgt)을 활용한 댄스피트니스 소비자행동 분석. 『한국스포츠산업경영학회지』, 25(1), 107-123.

김진기. (2020). "태권도장 지도자와 수련생, 학부모의 관계마케팅, 관계품질, 고객충성도의 관계". 전북대학교 대학원 박사학위논문.

김현기. (2022). "정보보안 파라미터 및 부호화 알고리즘 식별을 위한 머신러닝 구성기법 연구". 국민대학교 대학원 박사학위논문.

김형수. (2020). 『Step by Step 비즈니스 머신러닝 in 파이썬』. 서울: PREDICS.

나광택, 이진영, 김은찬, 이효찬. (2020). 증권 금융 상품 거래 고객의 이탈 예측 및 원인 추론. 『한국빅데이터학회지』, 5(2), 215-229.

남궁진, 배정섭, 원도연. (2016). 피트니스센터 여성회원들의 서비스스케이프에 대한 중요도 및 만족도 분석. 『한국여성체육학회지』, 30(3), 209-231.

남수정. (2010). 계획적 행동이론을 적용한 메트로 섹슈얼 소비행동에 대한 연구. 『소비자학 연구』. 21(1), 161-180.

남수태, 진찬용. (2013). 기술수용모델 개념 간의 관계에 대한 메타분석. 『벤처창업연구』, 8(4), 67-79.

남유정. (2020). "뷰티산업에서 예약자의 노쇼(No-Show) 인식과 개선방안이 재방문 의도에 미치는 영향". 한성대학교 대학원 석사학위논문.

남혜정. (2022). “코로나-19 이후 뷰티샵 예약불이행(No-Show)이 고객감정과 행동의도에 미치는 영향”. 서경대학교 대학원 석사학위논문.

노수연. (2020). 필라테스 센터의 서비스품질과 관계의 질 및 구매후 행동의 관계분석. 『한국체육과학회지』, 29(4), 485－496.

마효예. (2020). "필라테스 트레이닝 스튜디오의 서비스스케이프, 서비스가치, 스포츠관여도 및 장기지향성의 관계". 경희대학교 대학원 석사학위논문.

문찬석. (2012). "상업스포츠센터 고객 마케팅 요인이 고객 만족 및 구매행동에 미치는 영향". 서남대학교 대학원 박사학위논문.

박경원. (2017). "모바일커머스의 품질이 고객만족과 재구매의도에 미치는 영향 : 모바일친숙도의 조절효과를 중심으로". 영남대학교 대학원 박사학위논문.

박기연. (2018). "웹 로그 분석을 통한 사용자 행동 예측: 온라인 쇼핑 중 구매 결정 지원 기능을 이용한 트래픽을 중심으로". 이화여자대학교 대학원 석사학위논문.

박대호. (2016). "피트니스클럽 관계마케팅과 고객만족, 고객신뢰 및 행동의도의 관계". 단국대학교 대학원 석사학위논문.

박상용. (2020). "스크린골프장 서비스품질에 대한 인식이 고객만족 및 재방문에 미치는 영향". 이화여자대학교 대학원 석사학위논문.

박상윤. (2020). "온라인 여행상품 고객이탈 분석 연구 -데이터마이닝 기법 활용-". 이화여자대학교 대학원 석사학위논문.

박수연. (2023). “고객 이용 로그와 순환신경망을 활용한 이커머스 고객 이탈 예측”. 고려대학교 대학원 석사학위논문.

박수정. (2018). "스크린야구장 이용자의 선택속성, 서비스 품질, 소비감정, 고객만족, 재방문 의도간의 인과관계". 전남대학교 대학원 박사학위논문.

박영표. (2002). "스포츠센터 고객의 라이프스타일 특성에 따른 서비스만족 및 센터 재구매". 전남대학교 대학원 박사학위논문.

박일봉. (2017). "운동센터 회원의 신뢰도와 만족도 및 충성도의 구조적 관계". 부산대학교 대학원 박사학위논문.

박지윤. (2021). "필라테스 센터의 서비스스케이프가 고객 만족, 애착, 운동몰입 및 충성도에 미치는 영향". 경희대학교 대학원 박사학위논문.

박지훈. (2020). "화물자동차 운송시장의 화물정보 통합망 활성화 연구". 한국해양대학교 대학원 석사학위논문.

박현규. (2010). “브랜드 이미지가 관계품질과 재방문의도에 미치는 영향에 관한 연구”. 건국대학교 대학원 석사학위논문.

배재희. (2019). "기계학습을 이용한 스마트폰 재구매 패턴 분석 및 예측". 이화여자대학교 대학원 석사학위논문.

배행수. (2005). "고객 이탈 예측 데이터마이닝 기법 비교 연구". 연세대학교

대학원 석사학위논문.

백진우. (2005). “공공스포츠센터의 서비스 품질이 재구매의도 및 구전의도에 미치는 영향”. 『한국체육학회』, 44(1), 455-464.

서미선. (2019). "미용실 예약시스템 인식과 노쇼가 노쇼 페널티에 미치는 영향". 서경대학교 대학원 석사학위논문.

서영정. (2023). “인공지능 기반 고객 이탈 예측 기술 동향 및 발전방향”, 『디지털콘텐츠학회논문지』, 24(7), 1617-1627.

서웅식. (2018). "피트니스 센터 참여자의 지각된 케어링 분위기가 운동 만족과 운동 지속에 미치는 영향: 자율성의 매개 역할 검증". 순천향대학교 대학원 석사학위논문.

서일한, 이승만, 김선희. (2021). 골프연습장의 서비스스케이프가 고객만족과 재등록의도에 미치는 영향. 『골프 연구』, 15(2), 33-44.

서주환. (2014). "코레일 예약발매 서비스 편의성이 서비스가치와 고객만족, 재이용의도에 미치는 영향". 경기대학교 대학원 박사학위논문.

서현지. (2017). "빅 데이터를 이용한 고객 행태 분석에 대한 연구 : 유통업 사례를 중심으로". 가천대학교 대학원 석사학위논문.

서호진. (2023). "품질향상을 위한 데이터 전처리 프로세스에 관한 연구". 경상대학교 대학원 석사학위논문.

설민신, 임정원, 이성균. (2016). 스포츠센터 제공 서비스품질이 이용만족과 재등록에 미치는 영향. 『한국체육과학회지』, 25(6), 875-886.

소호진. (2023). “무도체육관 수련생의 선택속성이 수련만족과 재등록의도에 미치는 영향”. 계명대학교 대학원 박사학위논문.

송영일. (2012). "골프연습장의 서비스품질과 관계혜택이 전환장벽, 고객만족 및 고객충성도에 미치는 영향". 군산대학교 대학원 박사학위논문.

송옥매. (2009). “물리적 환경이 고객만족과 재구매의도에 미치는 영향에 관한 연구”. 숭실대학교 대학원 석사학위논문.

송유정. (2021). "신경망 기반 주가예측 성능 향상 기법". 숙명여자대학교 대학원 박사학위논문.

송종태. (2004). “서비스품질이 고객만족과 재구매 의도 및 구전 커뮤니케이션에 미치는 영향”. 원광대학교 대학원 박사학위논문.

송현수. (2020). "필라테스 선택속성과 서비스품질이 필라테스회원의 고객만족도 및 재등록의도에 미치는 영향". 경기대학교 대학원 석사학위논문.

심우극. (2015). “교육서비스품질과 서비스 가치, 수련만족, 브랜드 이미지, 자발적 행동의도간의 구조적 관계에 관한 연구 : 가치일치의 조절효과를 중심으로”. 공주대학교 대학원 박사학위논문.

안경민. (2021). "통계적 매칭과 머신러닝 앙상블 기법을 활용한 기업혁신 및 경영성과 예측 모형 개발". 동국대학교 대학원 박사학위논문.

안광성. (2017). "스포츠센터의 인적서비스품질이 이용객의 재이용의도에 미치는 영향". 한국교통대학교 대학원 석사학위논문.

안재혁. (2023). “태권도장의 교육서비스 품질이 도장이미지, 학부모만족, 재등록의도에 미치는 영향”. 부산대학교 대학원 석사학위논문.

안혜리. (2011). “예약부도를 고려한 항공사의 최적의 초과예약 방법에 관한 연구-저가항공사의 사례를 중심으로”. 부산대학교 대학원 석사학위논문.

안효경. (2009). "K-평균 군집분석과 이단계 군집분석을 이용한 고객세분화 비교 연구". 연세대학교 대학원 석사학위논문.

앨리스 젱, 아만다 카사리. (2019). 『피처 엔지니어링, 제대로 시작하기』. 서울: 에이콘출판.

양연경. (2023). 지속가능한 미술관 전시 큐레이팅 서비스 디자인을 위한 디자인 씽킹 연구. 『한국디자인문화학회지』, 29(1), 183-197.

엄지성. (2023). “머신러닝 기법을 활용한 수입 수산물 통관검사 부적합 영향요인 분석”. 충북대학교 대학원 석사학위논문.

오경택. (2009). "상업스포츠센터의 고객만족과 전환장벽이 재구매의도에 미치

는 영향". 세종대학교 대학원 박사학위논문.
오병임. (2016). "피트니스 센터의 고객경험관리와 즐거움, 고객태도, 관계몰입 및 고객충성도에 관한 연구". 단국대학교 대학원 석사학위논문.
오세준. (2018). MMORPG 사용자 유형 분류를 통한 이탈 예측 모델 생성 및 평가. 『정보과학회지』, 24(5), 220-226.
오시현. (2023). “필라테스센터의 서비스스케이프와 인적서비스품질이 고객만족 및 재등록의도에 미치는 영향”. 영남대학교 대학원 박사학위논문.
오영삼. (2019). 고객가치와 전환장벽이 이동통신 5G 서비스 수용태도에 미치는 영향 연구. 『유통경영학회』, 22(4), 101-108.
왕반. (2023). "머신러닝 기법을 활용한 이동통신사업자의 5G 요금제 잠재고객 이탈요인에 관한 연구". 한양대학교 대학원 석사학위논문.
유기성. (2018). 피트니스 센터 이용자가 인식한 시설이용 만족 및 불만족 요인 탐색. 『한국스포츠학회지』, 16(1), 53-63.
유박. (2018). "헬스클럽의 서비스품질이 고객충성도에 미치는 영향". 원광대학교 대학원 석사학위논문.
유서정. (2022). “필라테스 센터의 서비스 품질이 재구매 의도에 미치는 영향”. 충남대학교 대학원 석사학위논문.
유현민. (2018). “유소년 스포츠클럽 내 행사참여가 학부모의 만족도와 재등록 의도에 미치는 영향”. 한국체육대학교 대학원 석사학위논문.
유현탁. (2022). "데이터 기반 진단을 위한 기계학습 분석 프로세스 개발". 경상대학교 대학원 박사학위논문.
유호길. (2020). 피트니스센터의 서비스지향성, 직무만족, 조직몰입 및 조직성과의 관계. 『한국체육과학회지』, 29(6), 547-559.
윤선우. (2013). "피트니스 센터의 서비스스케이프 및 인적서비스가 고객만족, 신뢰, 몰입 및 충성도에 미치는 영향". 국민대학교 대학원 박사학위논문.
윤성훈. (2019). "e-커머스의 차별화마케팅이 소비자만족과 재구매의도 및 경

영성과에 미치는 영향". 이화여자대학교 대학원 박사학위논문.

윤영구. (2021). 휘트니스센터 소비자의 성향과 소비만족 및 구매의도 간의 인과관계. 『한국체육과학회지』, 30(1), 415－430.

윤현성. (2021). "피트니스센터의 온라인 마케팅유형별 마케팅특성이 소비자 태도 및 구매의사 결정요인에 미치는 영향". 경기대학교 대학원 박사학위논문.

이단비. (2020). “XGBoost 알고리즘에 기반한 직원 근무성향 분석에 대한 연구”. 남서울대학교 대학원 석사학위논문.

이동건. (2017). "유소년 종합 스포츠클럽의 서비스스케이프, 고객 만족, 스포츠관여도 및 재등록의도의 구조적 관계". 경희대학교 대학원 박사학위논문.

이동수. (2017). "서비스직원의 관계마케팅 실행요인이 관계품질을 매개로 고객충성도에 미치는 영향 : 스포츠 센터를 중심으로". 전북대학교 대학원 박사학위논문.

이미혜. (2015). 여행서비스 과정에서 여행객의 자기결정성에 따른 계획된 행동이 행동의도에 미치는 영향연구. 『관광연구저널』, 29(12), 33-47.

이병관, 이병수. (2014). 태권도장 수련생 학부모의 태권도장 선택속성과 고객만족, 재구매의도와 구조적 관계. 『한국체육과학회』, 23(1), 823-838.

이상우. (2015). "체육대학 입시학원의 서비스품질, 고객만족, 재등록의도간의 관계: 전환장벽의 조절효과". 이화여자대학교 대학원 석사학위논문.

이수현. (2015). "빅 데이터의 군집분석을 위한 군집화 유효성 지수 개발과 응용". 전남대학교 대학원 박사학위논문.

이승은, 총루이, 최인식. (2017). 스포츠센터의 관계마케팅 요인이 고객신뢰 및 재등록의사에 미치는 영향. 『디지털융복합연구』, 15(8), 435-444.

이승준. (2020). “계획행동이론(TPB)과 감정반응(PA)을 적용한 스쿼시 참여동기 및 행동의도 연구”. 국민대학교 대학원 박사학위논문.

이식. (2023). “고객이탈 예측을 위한 프로세스마이닝과 데이터마이닝 기반의 하이브리드 모델 개발”. 가톨릭관동대학교 대학원 박사학위논문.

이예준, 김상유, 김수현. (2020). 스포츠클라이밍센터 선택속성과 고객만족, 재등록의사의 관계 연구. 『한국융합학회논문지』, 11(9), 163-168.

이예진, 이용수. (2021). 퍼스널 트레이너의 전문성이 재등록의사에 미치는 영향 : 충성도를 매개효과로. 『한국체육과학회지』, 30(2), 513-527.

이유재, 공태식. (2004). 서비스 공정성이 고객 불량행동에 미치는 영향에 관한 연구 : 부정적 감정과 고객몰입의 매개효과. 『마케팅연구』, 19(4), 185-210.

이은정. (2011). "스포츠센터 퍼스널 트레이너 관계마케팅 실행요인 및 실무인식을 통한 고객관계관리(CRM)방안". 경희대학교 대학원 박사학위논문.

이은주. (2012). "스포츠센터 CRM의 하위요인과 고객만족, 고객신뢰, 고객충성도간의 인과관계". 경성대학교 대학원 박사학위논문.

이재준. (2021). “XGBoost와 SHAP 기법을 활용한 근로자 이직 예측에 관한 연구”. 국민대학교 대학원 석사학위논문.

이재현. (2007). “여성전용 피트니스센터의 서비스품질의 소비자 만족도 및 재구매의도에 미치는 영향”. 한국체육대학교 대학원 석사학위논문.

이정미. (2019). “스파 방문객의 추구편익 및 시장세분화 분석”. 경기대학교 대학원 석사학위논문.

이정우. (2019). "데이터 마이닝과 AHP를 이용한 고객 이탈 분석에 관한 연구 - 카드 고객 사례를 중심으로". 이화여자대학교 대학원 박사학위논문.

이준호. (2016). "스마트 스포츠센터의 서비스품질이 고객만족 및 재방문에

미치는 영향". 이화여자대학교 대학원 석사학위논문.

이지열. (2017). "국내 스키리조트 이용객이 인지하는 서비스 품질과 고객충성의 관계에서 고객만족의 매개효과 검증". 연세대학교 대학원 석사학위논문.

이진아. (2007). “패밀리 레스토랑의 마케팅 전략 및 개선방안 연구”. 단국대학교 대학원 석사학위논문.

이찬도. (2020). 문화관광형시장의 고객 소비형태별 군집분석과 시장 세분화 전략. 『문화산업연구』, 20(4), 87-96.

이현국. (2020). "태권도장 주말 프로그램이 학부모만족 및 재등록의도에 미치는 영향". 이화여자대학교 대학원 석사학위논문.

이훈영. (2002). e-마케팅 플러스. 『마케팅』, 36(10), 70-76.

이희선. (2011). "타겟 마케팅을 위한 캠페인 반응 예측모형 개발에 관한 연구". 호서대학교 대학원 석사학위논문.

임원현. (2015). "스포츠센터 퍼스널 트레이너의 인적서비스요인이 소비자의 감정 및 충성도에 미치는 영향". 경희대학교 대학원 박사학위논문.

임지수. (2020). "실내테니스장 고객 특성에 따른 고객 세분화 전략". 연세대학교 대학원 석사학위논문.

임지현. (2016). “병원예약부도 감소를 위한 연구”. 연세대학교 대학원 석사학위논문.

임진훈. (2017). "크로스핏센터 관계마케팅요인과 고객만족, 고객충성도 간의 관계". 전남대학교 대학원 석사학위논문.

장기만. (2020). "VOC(Voice of Customer) 빅데이터를 활용한 신용카드 사용자 이탈예측 모델". 이화여자대학교 대학원 석사학위논문.

장남식. (2005). 사전 세분화를 통한 고객 분류모형의 효과성 제고에 관한 연구. 『Information Systems Review』, 7(2), 23-40.

장민석. (2017). "빅데이터를 활용한 은행권 고객 세분화 기법 연구". 고려대

학교 대학원 박사학위논문.

장원용. (2018). 피트니스 센터 서비스스케이프가 고객만족, 센터 이미지 및 재 등록의도에 미치는 영향. 『한국체육과학회지』, 27(5), 849-862.

장하원. (2021). 레스토랑의 서빙로봇 속성 평가에 따른 고객 세분화. 『호텔경영학연구』, 30(1), 49-63.

장현수. (2021). "인공지능을 이용한 이동통신 가입자 이탈 예측 모델 정확도 비교분석 연구". 이화여자대학교 대학원 석사학위논문.

장혜원. (2019). "필라테스 교습의 문제점과 개선방안 : 지도자를 중심으로". 경북대학교 석사학위논문.

장호중, 김영수. (2020). 태권도 지도자의 지도유형과 수련만족 및 재등록의사의 관계. 『한국스포츠학회지』, 18(3), 783-792.

전상민. (2021). "데이터기반 이미지를 활용한 고객신용등급 분류 예측". 이화여자대학교 대학원 석사학위논문.

전윤경. (2021). "스포츠센터의 서비스 스케이프가 감성경험, 공간몰입, 만족, 충성도에 미치는 영향_지식동기의 조절효과를 중심으로". 부경대학교 대학원 석사학위논문.

전한나. (2022). "재구매 고객 만족도 요인 다차원 분석". 이화여자대학교 대학원 석사학위논문.

정난희. (2021). 필라테스 이용자가 지각한 서비스 품질이 이용자의 재이용의도에 미치는 영향: 기능적 가치와 감정적 가치의 매개효과 연구. 『한국콘텐츠학회논문지』, 21(5), 681－694.

정민규. (2020). "다양한 앙상블 모델을 이용한 개인신용평가 모델 개발". 이화여자대학교 대학원 석사학위논문.

정산. (2022). "COVID-19 시대의 피트니스센터 유형별 이용객의 선택속성 중요도와 만족도 분석". 동아대학교 대학원 석사학위논문.

정성훈. (2020). "머신 러닝 방법을 이용한 오피스 임대료 산정 -랜덤포레스

트,인공신경망,서포트벡터머신 활용을 중심으로-". 이화여자대학교 대학원 석사학위논문.

정은교, 황인선, 이상일. (2020). 필라테스 서비스품질이 가치, 만족 및 재이용에 미치는 관계 분석. 『한국체육과학회지』, 29(6), 561-575.

정태일. (2019). "다수의 범주형 설명변수를 갖는 데이터에 대한 서포트 벡터 머신 연구". 성균관대학교 대학원 석사학위논문.

정현아. (2018). "필라테스 퍼스널 트레이닝 스튜디오의 고객 만족, 재구매 및 추천의사에 관한 실증연구". 경희대학교 대학원 박사학위논문.

조성흠. (2020). "시니어 운동프로그램 이용특성이 브랜드 자산 및 만족도에 미치는 영향_코오롱스포렉스 '골드라이프'사례를 중심으로". 한국체육대학교 대학원 석사학위논문.

조소려. (2020). “외식배달 서비스 고객 시장세분화 및 IPA 분석”. 배재대학교 대학원 박사학위논문.

지준철, 장세용. (2023). 골프아카데미 지도자의 전문성이 재등록 의도에 미치는 영향: 고객 충성도의 매개효과 검증. 『스포츠사이언스』, 41(1), 93-100.

천송이. (2013). “리조트 예약부도 방지를 위한 연구”. 경희대학교 대학원 석사학위논문.

최경근. (2007). "유-소년 스포츠센터의 Service Encounter와 고객만족도, 고객충성도 및 재이용의사와의 관계". 경희대학교 대학원 박사학위논문.

최명희. (2006). "휘트니스클럽의 고객전환장벽 분석을 통한 마케팅전략". 동아대학교 대학원 박사학위논문.

최민규, 변경원. (2019). 피트니스 센터 진정성과 지속의사에 관한 연구: 브랜드 태도와 브랜드 애착의 다중매개 모형 검증. 『한국스포츠학회』, 17(1), 227-245.

최상원. (2013). “스포츠센터 고객접점서비스와 고객만족, 전환장벽 및 재구매

의 관계". 용인대학교 대학원 박사학위논문.

최영한. (2018). "피트니스센터 고객의 관계마케팅 경험과 서비스품질인식이 소비행동에 미치는 영향: 계획행동이론의 적용". 상명대학교 대학원 박사학위논문.

최은진. (2020). "머신러닝 기반의 중소기업 대출 약정해지 예측모델에 관한 연구". 이화여자대학교 대학원 박사학위논문.

최재호. (2011). "공공체육시설의 마케팅믹스요인이 고객만족도 및 재구매에 미치는 영향". 이화여자대학교 대학원 석사학위논문.

최진훈, 박주영. (2019). 골프연습장의 서비스 품질이 서비스 가치, 고객만족 및 재이용 의도에 미치는 영향. 『한국체육과학회지』, 28(3), 319-333.

최한나. (2019). "필라테스 회원이 지각하는 지도자의 전문성이 지도효율성 및 재등록의도에 미치는 영향". 경희대학교 대학원 석사학위논문.

추희정. (2022). “온라인 쇼핑몰 고객의 재방문과 재구매 및 이탈 요인 연구”. 서울과기대학교 대학원 석사학위논문.

코리 웨이드. (2022). 『XGBoost와 사이킷런을 활용한 그레디언트 부스팅』. 서울: 한빛미디어.

코투, 데쉬판데. (2016). 『데이터 과학』. 서울: 한티미디어.

트렁 도안 총빈. (2020). "골프장 소비자의 서비스 질 및 가격이 고객 만족도 및 재방문 의도에 미치는 영향". 이화여자대학교 대학원 석사학위논문.

피채희. (2022). "고객관계관리(CRM) 활동의 관계편익이 브랜드 이미지와 추천의도에 미치는 영향 : 차량판매 딜러를 중심으로". 숭실대학교 대학원 박사학위논문.

하제현, 하재필. (2021). 유아스포츠센터의 선택속성이 고객만족과 재등록의도에 미치는 영향. 『한국사회체육학회지』, 83, 147-157.

하지선, 최영희, 김주영. (2014). 테니스연습장의 마케팅전략이 고객만족과 재

등록의사에 미치는 영향. 『한국체육과학회지』, 23(1), 291-302.

하채원. (2023). “스포츠센터 마케팅믹스 요인이 만족도, 추천의도, 재구매 의도에 미치는 영향”. 세종대학교 대학원 석사학위논문.

한아름. (2012). "피트니스 프로그램에 참가하는 여가소비자의 정서표현성, 여가만족도, 감정적충성도 및 재구매의도의 관계: GroupExercise(G·X) 프로그램을 중심으로". 동덕여대학교 대학원 박사학위논문.

한철, 김대훈. (2021). 유소년 축구클럽 서비스품질과 고객만족 및 재등록의도와의 관계. 『한국여가레크리에이션학회지』, 45(2), 137-146.

함명수. (2007). “태권도장수련 프로그램 지각이 수련지속 의사에 미치는 영향”. 용인대학교 대학원 석사학위논문.

허용성. (2022). "중국 상업스포츠센터의 관계마케팅, 서비스품질, 소비성향 및 고객충성도 간의 구조적 관계". 우석대학교 대학원 박사학위논문.

홍승우. (2021). "이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론". 이화여자대학교 대학원 석사학위논문.

홍태호, 김은미. (2010). “데이터마이닝을 이용한 세분화된 고객집단의 프로모션 고객반응 예측”. 『Information Systems Review』,12(2), 75-88.

황영주. (2015). "스포츠센터 서비스 품질과 센터이미지, 고객만족도 및 고객충성도와의 관계". 동신대학교 대학원 박사학위논문.

황우현. (2018). "고객세분화를 위한 기대가치 산출모형에 관한 연구". 부경대학교 대학원 박사학위논문.

후앙이한. (2022). "스포츠센터의 서비스품질이 고객만족, 신뢰, 고객몰입, 충성도에 미치는 영향: 스포츠 관여도의 조절효과를 중심으로". 부경대학교 대학원 박사학위논문.

2. 국외문헌

Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data Platform. *Journal of Big Data*, 6(28), 6-28.

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. *IEEE Access*, 8, 220816-220839.

Albari, A., & Kartikasari, A. (2019). The influence of product quality, service quality and price on customer satisfaction and loyalty. *Asian Journal of Entrepreneurship & Family Business*, 3(1), 49-64.

Aldosary, M., & Alrashdan, A. (2021). Churn prediction for gym members using artificial neural networks assisted with the psychological concept of habit formation in the fitness industry. *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, 3996-4005.

Almohaimmeed, B. (2019). Pillars of customer retention: An empirical study on the influence of customer satisfaction, customer loyalty, customer profitability on customer retention. *Serbian Journal of Management*, 14(2), 421-435.

Alshurideh, M.T. (2016). Is customer retention beneficial for customers: a conceptual background. *Journal of Research in Marketing*, 5(3), 382-389.

Anderson, T. W. (2000). Antecedents to Satisfaction with Service Recovery. *European Journal of Marketing*, 23(12), 156-175.

Anil Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1), 4-28.

Arora, L., Singh, P., Bhatt, V., & Sharma, B. (2021). Understanding and managing customer engagement through social customer relationship management. *Journal of Decision Systems*, 30(2), 215-234.

Baesens, B., Viaene, S., Van, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, 138(1), 191-211.

Bahri, N., & Bilgihan, A. (2019). Customer retention to mobile telecommunication service providers: the roles of perceived justice and customer loyalty program. *International Journal of Mobile Communication*, 17(1), 82-107.

Bartha, É., & Perényi, S. (2015). Communication of trainers as one of the means for Customer retention in fitness Clubs. *Applied Studies in Agribusiness and Commerce*, 9(1), 35-40.

Catrili. T, & Veral. E. (2003). Outpatient scheduling in health care: A review of the literature. *Production and Operations Management society,* 12(4), 519-549.

Chavent, M., Genuer, R., & Saracco, J. (2021). Combining clustering of variables and feature selection using random forests. *Communications in Statistics-Simulation and Computation*, 50(2), 426-445.

Chayomchai, A., & Charnarpas, M. (2021). The service quality

management of the fitness center: the relationship among 5 aspects of service quality, *International Journal of Current Science Research and Review*, 4(6), 514-521.

Chouiekh, A. (2020). Deep convolutional neural networks for customer churn prediction analysis. *International Journal of Cognitive Informatics and Natural Intelligence*, 14(1), 1-16.

Clavel San Emeterio, I., Iglesias-Soler, E., Gallardo, L., Rodriguez-Cañamero, S., & García-Unanue, J. (2016). A prediction model of retention in a Spanish fitness centre. *Managing sport and leisure*, 21(5), 300-318.

Dias, C., Ferreira, A., Pereira, A. R., & Fonseca, A. M. (2019). Examining the relationship between perceived service quality, satisfaction and renewal intention in Portuguese fitness centers. *Journal of Sports Psychology*, 28(2), 49-58.

Elyan, E., & Gaber, M. M. (2017). A genetic algorithm approach to optimising random forests applied to class engineered data. *Information sciences*, 384, 220-234.

Ferrand, A., Robinson, L., & Valette, P. (2010). The intention-to-repurchase paradox: A case of the health and fitness industry. *Journal of Sport Management*, 24(1), 83-105.

Gattermann, T., & Thonemann, U. W. (2021). How training on multiple time slices improves performance in churn prediction. *European Journal of Operational Research*, 295(2), 664-674.

Ghadiri, M.S., Fahiminejad, A., Bagheri, H., & Tayebi, S. M. (2021). Market-retention in fitness centers: Identifying and ranking the factors affecting customer retention. *Journal of New Studies in*

Sports Management, 2(1), 95-110.

Goncalves, C., & Diniz, A. (2015). Analysis of member retention in fitness through satisfaction, attributes perception, expectations and wellbeing. *Revista Portuguesa de Marketing*, 18(34), 65-76.

Gu, Z., & He, C. (2021). Application of fuzzy decision tree algorithm based on mobile computing in sports fitness member management. *Wireless Communications and Mobile Computing*, 2021, 1-10.

Hassan, M., & Mizra, T. (2018). Churn Prediction in Banking Sector using Bayesian Neural Networks. *International Journal of Research in Advent Technology*, 6(12), 3343-3346.

Howat, G., & Assaker, G. (2016). Outcome quality in participant sport and recreation service quality models: Empirical results from public aquatic centres in Australia. *Sport Management Review*, 19(5), 520-535.

Hu, C., & Steingrimsson, J. A. (2018). Personalized risk prediction in clinical oncology research: applications and practical issues using survival trees and random forests. *Journal of biopharmaceutical statistics*, 28(2), 333-349.

Hung, S. Y., Yen, D. C., & Wang, H. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515-524.

Hurley, T. (2004). Managing customer retention in the health and fitness industry: A case of neglect. *Irish Marketing Review*, 17(1), 23-29.

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing*, 80(6), 36-68.

Larson, J., Jaworski, B. J., & Larson, J. (2021). Managing your most loyal customer relationships. *Business Horizons*, 64(1), 141-147.

Li, X., & Li, Z. (2019). A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des Systèmes d'Information*, 24(5), 525-530.

MacIntosh, E., & Law, B. (2015). Should I stay or should I go? Exploring the decision to join, maintain, or cancel a fitness membership. *Managing Sport and Leisure*, 20(3), 191 – 210.

McDougall, G. H. G., & Levesque, T. (2000). Customer satisfaction with services: Putting perceived value into equation. *Journal of Service Marketing*, 14(4), 392-410.

Milan, G.S., Eberle, L., & Bebber, S. (2015). Perceived value, reputation, trust and switching costs as determinants of customer retention. *Journal of Relationship Marketing*, 14(2), 109-123.

Mirkovic, M., Lolic, T., Stefanovic, D., Anderla, A., & Gracanin, D. (2022). Customer churn prediction in b2b non-contractual business settings using invoice data. *Applied Sciences*, 12(10), 500-502.

Moreira, A. C., & Silva, P. M. (2015). The trust-commitment challenge in service quality-loyalty relationships. *International Journal of Health Care Quality Assurance*, 28(3), 253-266.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting Subsriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Transactions on Neural Networks*, 11(3), 690-696.

Nalatissifa, H., & Pardede, H. F. (2021). Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data. *Jurnal Elektronika dan Telekomunikasi*, 21(2), 122-127.

Neslin, S. A., & Charilotte, H. M. (2006). Defection Detection : Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journals of marketing research*, 43(2), 204-211.

Neto, M. P., & Paulovich, F. V. (2020). Explainable matrix-visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1427-1437.

Oblander, E. S., Gupta, S., Mela, C. F., Winer, R.S., & Lehmann, D. R. (2020). The past, present, and future of customer management. *Marketing Letters*, 31(2), 125-136.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4), 460-469.

Ozdemir, O., & Yildiz, S. M. (2020). Evaluation of fitness center service quality through important-performance analysis. *European Journal of Physical Education and Sport Science*, 6(6), 46-58.

Pargent, F., Pfisterer, F., Thomas, J., & Bischi, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, 37(5), 2671-2692.

Peitzika, E., Chatzi, S., & Kissa, D. (2020). Service quality expectations in the fitness center context: A validation of the expectations

component of the SERVQUAL scale in Greece. *Services Marketing Quarterly*, 41(2), 89-104.

Rodrigues, R., Pinheiro, P., Gouveia, A., Brás, R., O'Hara, K., Duarte, P. & Esteves, M. (2017). Segmentation of Portuguese customers' expectations from fitness programs. *Journal of International Studies*, 10(3), 234-249.

Simões, D., & Nogueira, J. (2022). Learning about the customer for improving customer retention proposal of an analytical framework. *Journal of Marketing Analytics*, 10(1), 50-63.

Sobreiro, P., Guedes-Carvalho, P., Santos, A., Pinheiro, P., & Gonçalves, C. (2021). Predicting Fitness Centre Dropout. *International journal of environmental research and public health*, 18(19), 46-56.

Srivastava, P. R., & Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach. *Journal of Global Information Management*, 29(6), 1-29.

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operational Research*, 236(2), 624-633.

Tran, V. D. (2020). Assessing the effects of service quality, experience value, relationship quality on behavioral intentions. *The Journal of Asian Finance, Economics & Business*, 7(3), 167-175.

Tsitskari, E., Quick, S., & Tsakiraki, A. (2014). Measuring exercise involvement among fitness centers' members: is it related to their

satisfaction?. *Services Marketing Quarterly*, 35(4), 372-389.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, 7, 60134-60149.

Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2009). A Frequency Based Encoding Technique for Transformation of Categorical Variables in Mixed IVF Dataset. *In 2009 annual international conference of the IEEE engineering in medicine and biology society*, 6214-6217.

Wang, R., Cheng, M. N., Loh, Y. M., Wang, C., & Cheung, C. F. (2022). Ensemble learning with a genetic algorithm for surface roughness prediction in multijet polishing. *Expert Systems with Applications*, 207, 260-272.

Wei, P., Hung, H., Yang, H., & Jui, Y. (2010). Examination of the influence of service quiality on membership renewal in fitness centers in San Francisco Bay Area. *Journal of Service Science*, 3(2), 13-24.

Wouter, V., David, M., Christophe, M., & Bart, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3), 2354-2364.

Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, 9, 62118-62136.

Xia, G. E., & Jin, W. D. (2008). Model of customer churn prediction on

support vector machine. *Systems Engineering-Theory & Practice*, 28(1), 71-77.

Yusof, A., Popa, A., & Geok, S. K. (2018). Perception of service quality and behavior intentions of campus fitness center users in Malaysia and Thailand. *International Journal of Academic Research in Business and Social Sciences*, 8(12), 187-197.

Zdravevski, E., Lameski, P., Apanowicz, C., & Ślęzak, D. (2020). From Big Data to business analytics: The case study of churn prediction. *Applied Soft Computing*, 90, 106-164.

ABSTRACT

A Study on Customer Segmentationysis and and and Re-registration Prediction of the Fitness center - Using Fitness Center Group Exercise member information -

Kim, Young-Jin

Major in Smart Convergence Consulting

Dept. of Smart Convergence Consulting

The Graduate School

Hansung University

This study proposed a marketing plan using the fitness center's customer segmentation and re-registration prediction model using the fitness center small-scale group exercise member information. In the fitness industry, like other service industries, it is important to establish relationships with customers, and customer relationships reflecting customer characteristics eventually lead to re-registration, which is directly connected to profitability.

Since there is a limit to the survey response to predict whether or not to re-register by conducting a survey each time, a systematic data marketing strategy that analyzes the data of members' use of the center is a realistic alternative. It is necessary to try to utilize Martech, which is in the spotlight in the marketing field. A compound word of marketing and

technology, Martech can detect specific pattern behavior in real-time collected customer logs, analyze behavior patterns, and design target marketing campaigns by finding the needs of potential customers in a digital environment.

The purpose of the study was to collect and process customer data inside the center, suggesting practical implications for making proper marketing decisions such as promotions and events that allow center members to re-register without churn. Data were collected on the history of 513,530 cases of 1,160 members who were re-registered and exited from small-scale group exercise (GX) over the period of investigation of the N branch of Fitness Franchise A during 2020.2~2022.4(24 months).

Attendance, no-show, cancellation, preferred instructor, preferred class, preferred time, and preferred day of the week were selected with a total of 7 variables through prior literature research in the fitness field. The categorical variables such as preferred instructor, preferred class, preferred time, and preferred day of the week were converted into numerical variables through frequency encoding. It was confirmed that it was not a normal distribution through skewness/ kurtosis and KS (Kolmogorov-Smirnov) tests, so it was analyzed using nonparametric statistics. Nonparametric tests were performed to confirm whether the variables were meaningful to re-register.

Feature is a characteristic of data for predicting the value of data well, and since it was not normally distributed, minimum-maximum scaling was performed to perform follow-up work based on the min_max value.

Clustering is the work of finding meaningful groups in the data, comparing the elbow method and the silhouette coefficient, and dividing it into three clusters with the largest silhouette coefficient when K=3.

With a nonparametric analysis method that can determine whether

there is a difference between three or more groups, all of the Kruskal-Wallis H test results had a significance probability of 0.000, and all the null hypotheses were rejected. In the detailed analysis by cluster, in the case of cluster 1 and cluster 3, the cancellation min_max variable was applied statistically significantly because the significance probability was smaller than 0.05, and cluster 2 excluded the cancellation min_max variable because the cancellation min_max had a significance probability of 0.650, which was not statistically significant.

Cluster 1 showed a higher value for a total of 7 variables, no-show min_max was higher than that of other clusters, and cancellation min_max was the middle value of cluster 2 and cluster 3. Overall, when comparing variable attributes, it can be seen as the most vulnerable member cluster among the total three clusters, so cluster 1 was named "vulnerable member group". Cluster 2 has an attribute value H (average ranking) of 7 variables in the middle position, and can be said to be an intermediate member cluster among the total three clusters. Therefore, cluster 2 was named "intermediate member group". Variable cluster 3 showed higher values than other clusters except for no-show min_max. Overall, when comparing variable attributes, it can be seen as the best member cluster among the total three clusters, so cluster 3 was named "Excellent member group".

Decision Tree (DT), Random Forests (RF), and eXtreme Gradient Boosting (XGB) were commonly divided into 70% of training data and 30% of evaluation data. The predictive performance was evaluated by repeatedly dividing the data several times through cross validation in which the commonly used K value was designated as 5.

A total of six indicators, accuracy, precision, recall, specificity, F1 score, and AUC, were used to evaluate the re-registration prediction performance. The single model, cluster 1 (vulnerable member group),

cluster 2 (intermediate member group), and cluster 3 (excellent member group) were all confirmed to have excellent performance in the order of DT < RF < XGB. XGB is a boosting ensemble model that reduces errors for one predictive model, and is different from RF, a bagging ensemble model. Cluster 1 (vulnerable member group) < cluster 2 (intermediate member group) < cluster 3 (excellent member group) showed excellent prediction performance in the order of cluster 1.

In the case of data that want to know whether to re-register, the cluster can be classified through K-means cluster analysis, and then the XGB re-registration prediction probability, which is the optimal prediction model, can be checked in the cluster and used for customer management.

Through target marketing strategies for each cluster, departure can be prevented and linked to re-registration. Marketing plans can be implemented in consideration of the each cluster feature importance, such as reservation confirmation guidance text messages and point management.

As an academic implication, first, it is a study that conducted an advanced analysis compared to previous studies in predicting re-registration in the fitness field. Second, non-parametric statistical techniques were used to statistically approach non-normal data. Third, it is possible to prevent overfitting of a specific data set through cross validation and generate a generalized model. Fourth, the organic linkage between supervised learning and unsupervised learning can be cited. Finally, as a result of comparing the performance by applying various artificial intelligence algorithms, XGB was the best, and the relative feature importance affecting re-registration through feature importance of XGB can be quantitatively reviewed and used for future index management.

As practical implications, first, data accessibility was improved by using the usual customer DB. Second, it is possible to intensively manage target customers through customer segmentation considering the characteristics of the customer. Third, as a reference material that can be used from a marketing perspective, it focused on practical customer management such as estimating the probability of re-registration prediction. Finally, if government agencies such as the Small Business Corporation develop and distribute re-registration (re-purchase) prediction algorithms suitable for each industry, the foundation for sustainable growth can be established.

The limitations of the study and future research directions were based on 513,530 data on the history of 1,160 members who were re-registered and departed from small-scale exercise at N branch of A fitness center, but more data needs to be secured for better generalization. It is expected that the study will be more valuable if the scope is expanded to various fitness centers to generalize the empirical research results.

【Key words】Martech, Non-parametric test, Customer Segmentation, Re-registraion prediction, Target marketing