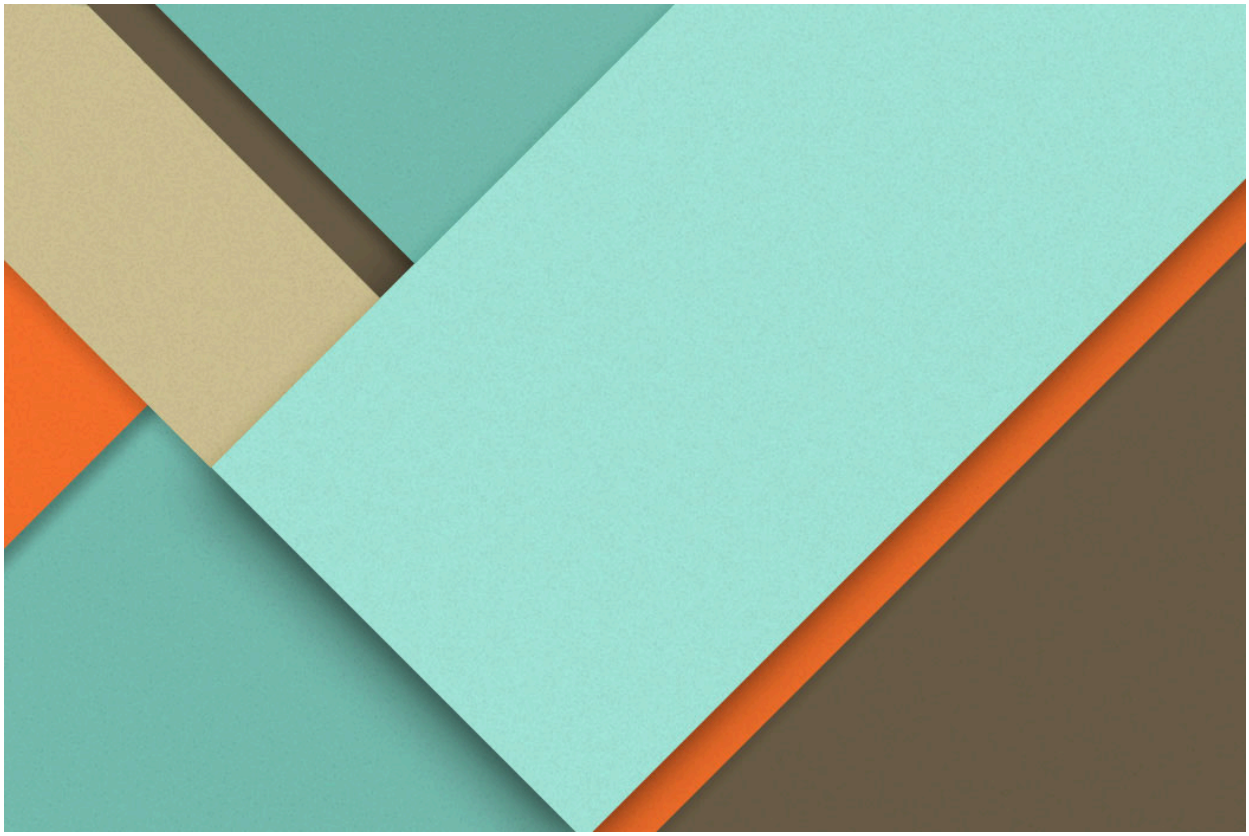


분석 프로젝트 1조

팀명 : TMI (We **t**ried **m**any t**i**mes, but failure is common)
2025년 1월 7일

작성자 이름

홍영일
장새영
김영재
이현승

목차

1. 분석 개요

1) 분석 목적

- **연령별 여가활동 분석**
  - 다양한 연령대가 어떤 종류의 여가활동에 참여하는지를 분석함으로써, 각 세대의 특성과 가치관을 이해
  - 이를 통해 보험사가 고객의 라이프스타일을 반영한 상품 개발에 필요한 기초 데이터를 제공

- **보험 가입 패턴 탐색**
  - 여가활동과 보험 상품 선택 간의 관계성 파악을 통해 특정 여가활동에 참여하는 고객이 선호하는 보험 상품의 특성을 규명

- **상관관계 분석**
  - 여가활동이 보험 상품 선택에 미치는 영향을 분석하고, 보험업계의 마케팅 전략 수립을 위한 인사이트를 제공

2) 기대 효과

- **고객 이해도 향상**
  - 고객의 선호도와 행동 패턴 파악을 통한 상품 개발로 고객 만족도를 높일 수 있을 것으로 추정

- **효율적인 마케팅 전략 수립**
  - 연령대와 여가활동 간 상관성 분석을 통해 맞춤형 마케팅 전략을 제시

- **상품 개발 및 개선**
  - 고객 니즈에 맞춘 신상품 개발 및 기존 상품 개선으로 보험 시장에서의 경쟁력 우위를 확보

- **세분화된 서비스 제공**
  - 고객의 여가활동 등을 고려한 맞춤형 보험 상품 제공을 통해 고객 충성도를 향상

- **데이터 기반 의사결정 강화**
  - 추후 보험사가 보험 상품 개발의 기초 자료로 활용

2. 데이터 개요

1) 데이터 명세서

- **보험 데이터 : '해빗팩토리' 데이터 활용**
  - 분석 항목 : 성별, 연령별 보험종류, 보장 항목, 계약 건수
  - 보험 계약자의 성별, 나이, 보험 종류, 보장상품 데이터 120만 건을 분석
  - 타겟인 20~30대와 이외 연령대의 보험 가입 비율 비교 분석
    - 2022년 기준 데이터 활용:
      보험 산업의 보수적 특성으로 최신 자료 수집에 한계
      → 2022, 2024년 유사 자료 통계 비교
    - 데이터 변화 추이 및 수치에 유의미한 차이 없음

- **여가활동 지속년수 데이터 : 'KOSIS' 데이터 활용**
  - 분석항목 : 여가활동 항목별 지속 기간 비율
  - 연령대별 여가활동 참여 지속 기간 분석
  - 2023년 기준 데이터 활용

- **여가활동 항목별 참여 비율 데이터 : 'KOSIS' 데이터 활용**
  - 분석항목 : 1년간 참여한 여가활동 종류
  - 연령대별 여가활동 빈도수 비율 분석
  - 2023년 기준 데이터 활용

- **여가활동 소비지출 금액 데이터 : 'KOSIS' 데이터 활용**
  - 분석항목 : 연령별 여가활동에 사용되는 소비지출 범위
  - 연령대별 여가활동에 소비하는 지출 범위 비율 분석
  - 2023년 기준 데이터 활용

2) 분석 결과 해석 시 고려사항

- **결과의 신뢰성**

  표본 수의 차이로 인한 분석 결과의 신뢰도를 판단해야 하며 결과가 통계적으로 유의미한지 확인하기 위해, p-value가 0.05 이하일 경우의 결과는 통계적으로 유의미하다고 판단할 수 있다.

- **데이터의 대표성**

  사용된 데이터가 전체 보험 가입자를 대표하는지 확인해야 하며 특정 연령대나 성별에 치우친 데이터는 결과의 일반화에 한계를 줄 수 있다.

- **상관관계와 인과관계**

  분석 결과가 상관관계를 나타내더라도, 인과관계가 존재하는지 여부는 추가적인 검토가 필요하다.

- **비즈니스 영향**

  결과가 비즈니스에 미치는 영향을 고려합니다. 예를 들어, 특정 취미 소분류와 관련된 보험 상품의 보장 항목이 소비자에게 긍정적인 반응을 얻었다면, 이를 바탕으로 마케팅 전략이나 제품 라인업을 조정할 필요가 있다.

- **대안적 설명**

  분석 결과에 대한 대안적 설명도 고려해야 한다. 특정 보장 항목의 선호가 취미 소분류 때문만이 아니라, 다른 요인 때문일 수 있다.

- **추가 분석 필요성**

  분석 결과가 명확하지 않거나 추가적인 통찰이 필요할 경우, 후속 분석을 제안해야 한다.

- **주요 구조 및 변수 설명**
  - 보험회사 가입 로우데이터를 성별로 먼저 나눈 후, 각 성별에 해당하는 연령별로 가입한 보험의 보장상품을 분류하였다.
  - 연령별 여가활동 비율 데이터를 통해 20~30대가 타 연령대보다 많이 참여하는 여가활동을 찾고 소득, 가구 형태에 따라 분류하였다.
  - 연령 데이터를 종속변수로 설정하여 그에 따른 각 열들의 상관관계를 분석하였다.

[표 1] 보험_가입자_정보.csv의 데이터 종류 (모든 변수는 범주형)

변수명	설명	예시 값
bth_yymm	생년월일	199704
age_val	나이	29
sex_cd	성별	M, F
gnt_itm_nm	보험 상품의 보장 항목	실손상해약제비
inco_typ_nm	보험 상품 종류	손해보험
inco_nm	보험 회사명	메리츠화재보험
category	여가생활 중분류	tour
sub_category	여가생활 소분류	camping
duration_year	여가생활 지속 기간	1~2, 5+
family_num	가족 수	1, 2, 3
money_spend	소비 금액	1-5, 5-10, 10-15, 15+

3) 데이터 전처리과정

- 데이터 수집 및 탐색
  - csv 파일을 통해 pandas data frame 형태로 약 120만 건의 보험 가입자 데이터를 수집
  - 정리한 데이터를 알고리즘으로 학습시키기 위해 중분류, 세부 분류, 지속 연도별로 분류
  - 각 열의 데이터 유형, 기본 통계량, 결측치 존재 여부 확인

- 결측치 및 이상치 처리
  - 데이터 탐색 단계에서 발견한 결측치는 Null 값으로 대체
  - 데이터 셋의 개수와 카테고리를 줄여 분석 정확도를 높이기 위해 특정 연령대를 포함하지 않는 구간(19이하, 60세이상)의 데이터 제거

- 데이터 변환
  - 데이터의 정확한 분석을 위해 일부 열의 데이터 유형을 범주형으로 변환

- 정규화/표준화
  - 범주형 데이터를 사용하여 정규화 과정은 제외

- 인코딩
  - 범주형 변수를 알고리즘으로 학습하기 위해 원-핫인코딩 사용

- 샘플링
  - 모델 학습의 효율성을 높히기 위해 오버샘플링과 언더샘플링을 이용하여 각 클래스의 샘플이 100개가 되도록 조정

- 훈련/테스트 데이터 셋
  - 훈련/테스트 데이터 셋을 8:2의 비율로 분할

4) 데이터 품질과 한계점

4-1) 데이터 품질

- 완전성

  데이터 셋은 전처리 과정을 통해 결측치가 없는 상태로 수집되어, 모든 변수에 대해 유효하다.

- 정확성

  데이터는 신뢰할 수 있는 출처에서 수집되었으며, 각 변수의 정의가 명확히 설정되어 분석의 일관성을 유지한다.

- 일관성

  데이터는 동일한 형식으로 수집되었으며, 데이터 유형이 적절하게 설정되어 있다.

- 대표성

  데이터가 전체 표본으로 대표성이 좋다.

4-2) 한계점

- 데이터 제한성

  보험산업의 보수성으로 인해 raw data 수집에 한계가 있어, 다양한 변수 간의 연관성을 가지는 구체적인 raw data의 제공에 제한이 있다.

- 범주형 변수의 세분화 부족

  'sub_category'와 같은 범주형 변수의 경우, 모든 여가생활 항목을 포함하지 않아 세분화가 부족할 수 있다.

- 새로운 변수 설정

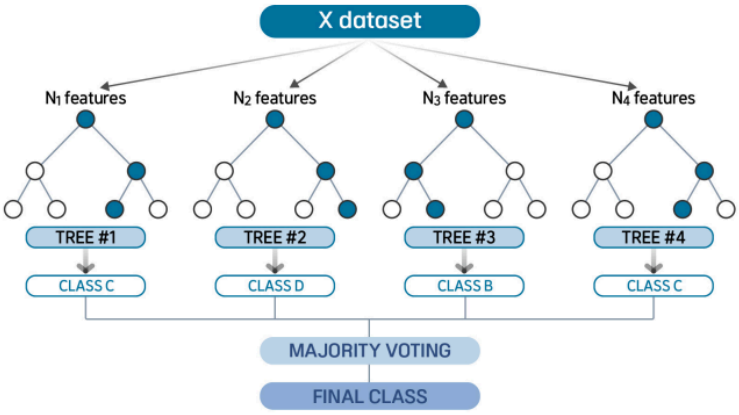
  얻고자 하는 정보에 대해 수치적 계산을 통한 새로운 변수를 설정하였으나 정확성에 한계가 있다.

3. 분석 방법론

1) 랜덤 포레스트(Random Forest)

- **설명**

  랜덤 포레스트는 여러 개의 결정 트리를 결합하여 예측을 수행하는 앙상블 학습 기법이다. 각 트리는 데이터의 무작위 샘플과 특성의 무작위 선택을 통해 독립적으로 학습되며, 이로 인해 과적합을 방지하고 예측 성능을 향상시킨다. 최종 예측은 각 트리의 예측 결과를 평균 내거나 다수결 투표를 통해 결정된다. 랜덤 포레스트는 분류와 회귀 문제 모두에 효과적으로 사용되며, 특히 고차원 데이터와 결측치에 강한 성능을 보인다.



자료 : https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

[그림 1] 랜덤포레스트 알고리즘 개념도

- **평가 지표 선정**

  본 과제에서는 랜덤 포레스트 모델의 성능을 객관적으로 평가하기 위해 다양한 지표를 활용했다. 먼저, 정확도, 정밀도, 재현율, F1-score 등을 계산했다.

$$\text{정밀도}(Precision) = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$재현율(Recall) = \frac{True\ positives}{True\ Positives + False\ Negatives}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

정확도는 전체 예측 중 올바르게 예측한 비율을 나타내는 지표로, 모델의 전반적인 예측 성능을 보여준다. 정밀도는 모델이 긍정 사례를 얼마나 정확하게 식별하는지를 측정하며, 재현율은 실제 긍정 사례 중 모델이 얼마나 많이 찾아내는지를 나타낸다. F1-score는 정밀도와 재현율의 조화평균으로, 모델의 균형 잡힌 성능을 평가할 수 있다.

이러한 다양한 평가 지표를 활용하여, 원본 보험 데이터(A)와 여가생활 데이터가 추가된 데이터(B)를 기반으로 한 랜덤 포레스트 모델의 성능을 비교 분석했다. 이를 통해 여가생활 정보 추가가 보험 보장 항목 예측 모델의 성능 향상에 기여하는지 여부를 객관적으로 평가할 수 있었다.

- **분석 기법 사용 과정**

1. KOSIS 통계자료를 통해 연령대별, 성별과 여가생활 종류 간 비율에서 눈에 띄는 특징(수치)을 발견할 수 있었다.
2. 여가생활 항목을 기준으로 다른 항목(ex, 성별, 연령별, 구성 가족원별, 여가문화 소비금액별) 비율과 연관성이 있을 것으로 예상하였다.
3. 이에 각 항목들을 독립변수로 설정하고 종속변수를 취미항목으로 설정하였다.
4. 초기에는 모든 카테고리를 독립변수로 넣어 수치를 확인하였다.
5. 유의미한 수치가 나올 수 있게 주성분분석을 통해 불필요한 변수를 제거하는 후진 제거법을 적용하였다.
6. 수치를 확인하여 여가 항목들을 가장 잘 분류할 수 있는 모델을 생성하였다.

2) A/B TEST(가설검정)

- **설명**

  A/B 테스트는 두 가지 변수를 비교하여 어떤 것이 더 효과적인지 확인하는 실험 방법이다. 가설 검정은 두 변수 간 차이가 통계적으로 유의미한지를 판단할 수 있다. A/B 테스트에서 나온 결과는 가설의 참/거짓을 파악한다.

- **평가 지표 선정**

  A/B 테스트의 결과를 객관적으로 평가하기 위해 다양한 지표를 활용했다. 통계적 유의성을 평가하기 위해 p-value(유의 확률)와 카이제곱 검정을 활용해 독립성을 검정하였다. p-value는 귀무가설 채택 여부를 판단하는 데 사용되며, 일반적으로 p-value 값이 유의수준 0.05 이하일 경우 귀무가설을 기각하며 대립가설을 채택하게 된다.

- **분석 기법 사용 과정**

  1. 랜덤포레스트 결과 분류가 가능하다는 것을 확인하여 성별, 연령별, 여가생활 카테고리별 연관성 및 패턴이 있음을 파악하였다.

  2. 만약 고객의 연령대별 여가생활과 보험의 보장항목 간 상관관계가 있을 경우, 여가생활 항목별 보험 보장항목 추천 등 중요한 인사이트를 도출할 수 있을 것이라고 판단하였다.

  - 가설을 검정을 위한 설정 :

    A : 나이(50~59), 성별(남자), 취미항목(등산), 가족 부양 수(1)

    B : 보험 보장 항목 중 '암' 항목

귀무가설 : 여가 생활 데이터(B)와 보험 보장 항목 데이터(A)는 연관이 없다.

대립가설 : 여가 생활 데이터(B)와 보험 보장 항목 데이터(A)는 연관이 있다.

3) 사용도구

- **pandas**

  데이터 프레임(DataFrame) 구조를 통해 대량의 데이터를 필터링 및 그룹화하였고 결측치 처리 등 데이터 전처리에 활용하였다.

- **scikit-learn**

  머신러닝을 위한 파이썬 라이브러리로 제공하는 모델 중 분류 모델을 활용하여 데이터 셋을 학습하고 평가 및 예측 등의 기능을 수행했다.

- **plotly**

  대화형 데이터 시각화를 위한 파이썬 라이브러리로 다양한 유형의 그래프와 차트를 쉽게 생성하여 시각화했다.

4. 결과 및 해석

1) 주요 결과

1-1) 랜덤포레스트(Random Forest)

- 여가생활 항목을 기준으로 다른 항목들을(ex, 성별, 연령별, 구성 가족원별, 여가문화 소비금액별) 독립변수로 설정하고 종속변수를 취미항목으로 설정하였다.
- 분석 결과, 분류 정확도가 0.73으로 나타났으며, 주요 특성으로는 age_group, sex_cd, duration_year가 확인되었다.

1-2) 모델 성능 평가

- 모델의 성능은 F1 Score를 통해 평가를 진행하였다. 본 모델의 F1 Score는 0.73으로, 모델이 전체적으로 좋은 예측 성능을 나타내고 있음을 의미한다.
- 이는 고객 특성별로 가장 참여 가능성이 높은 여가생활 항목을 예측하는 데 상대적으로 높은 정밀도와 재현율을 달성했음을 보여준다.

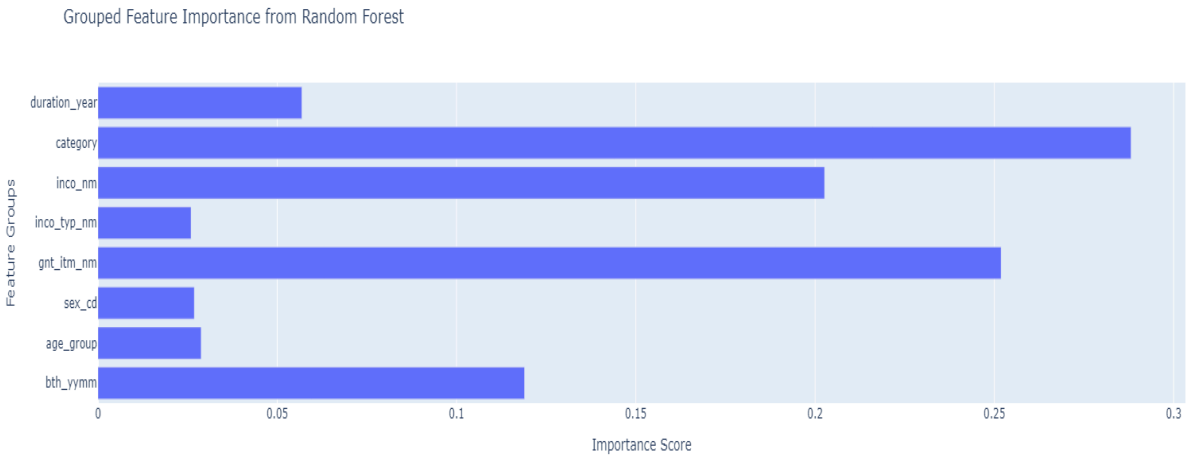
[표 2] Sub Category에 따른 F1 score

	Precision	Recall	F1 Score
camping	0.73	0.91	0.81
classic	0.77	0.77	0.77
concert	0.64	0.68	0.66
fishing	0.61	0.74	0.67
golf	0.75	0.72	0.73
hiking	0.82	0.66	0.73

musical	0.76	0.73	0.75
pet	0.66	0.73	0.69
picnic	0.84	0.59	0.69
work_out	0.72	0.68	0.70
yoga	0.74	0.81	0.78
accuracy(F1 average)			0.73

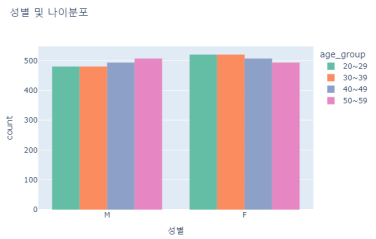
1-3) Feature Importance

- 랜덤 포레스트 모델을 통해 각 특성이 종속 변수에 미치는 영향을 평가하기 위해 Feature Importance를 분석하였다.

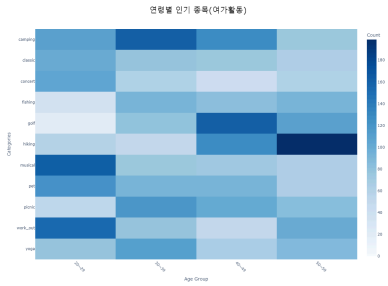


[그림 2] 랜덤포레스트 분석에서의 Feature Importance

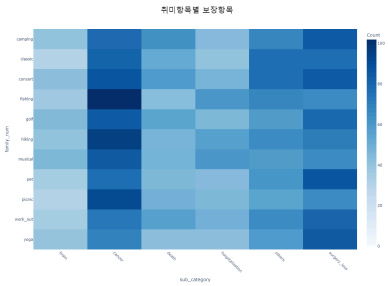
1-4) 패턴 분석



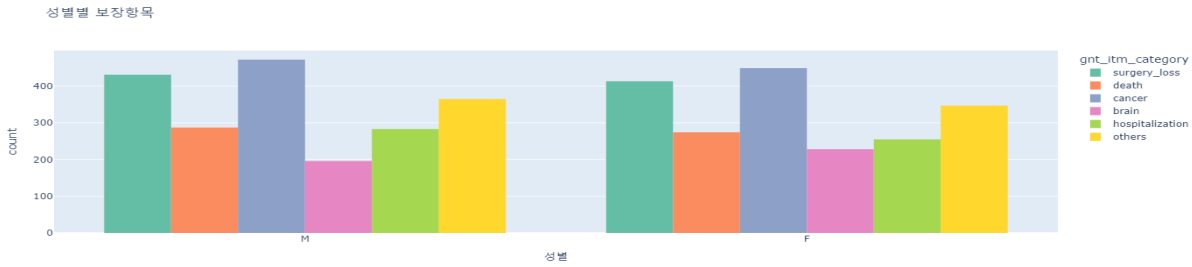
(a) 성별 및 나이에 따른 데이터 분포



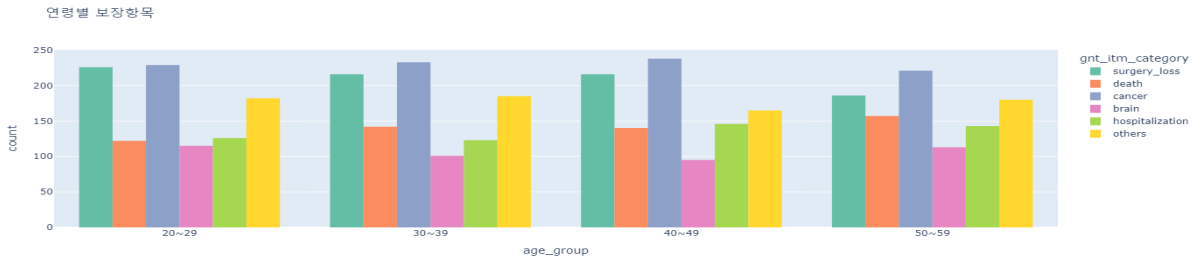
(b) 연령대별 인기 여가활동 항목



(c) 여가활동 항목별 보장항목



(d) 성별에 따른 보험 보장항목 데이터 분포



(e) 연령대별 보험 보장항목 데이터 분포

[그림 3] 보험가입자_정보.csv의 항목별 분포도

- [그림 3]을 통해 연령, 성별 중심으로 패턴을 파악할 수 있었다. 이를 통해 연령 '50~59', 성별 'M', 취미항목 'hiking', 가족 부양 수 '1' 이 가장 뚜렷한 특성을 나타낸다고 판단되어 이를 중심으로 가설을 세우기로 결정했다.

1-5) A/B test 선정이유

- 랜덤포레스트 결과 분류가 가능하다는 것을 확인하여 성별, 연령별, 여가생활 카테고리별 연관성 및 패턴이 있음을 파악하였다.

- 따라서 고객의 연령대별 여가생활과 보험의 보장항목 간 상관관계가 있을 것이라는 가설을 검정하기 위해 A/B 테스트를 진행하였다.

1-6) 가설 설정

- 본 과제에서 설정한 가설은 다음과 같다:
  - A : 나이(50~59), 성별(남자), 취미항목(등산), 가족 부양 수(1)
  - B : 보험 보장 항목 중 '암' 항목
- 가설 설정:
  - 귀무가설 : 여가 생활 데이터(B)와 보험 보장 항목 데이터(A)는 연관이 없다.
  - 대립가설 : 여가 생활 데이터(B)와 보험 보장 항목 데이터(A)는 연관이 있다.
- 테스트 실행:
  - 연령대 구간별로 구분한 데이터를 기준으로 50대 고객의 데이터를 기준으로 A/B 테스트를 실시하였다.

1-7) A/B test 결과

- 결과 수치 :

카이제곱 검정량	56.4101
p- value	0.2478

- p-value 값이 0.2478로 유의수준 0.05보다 커서 검정 결과가 유효하지 않은 것으로 나타남
- 이에 여가생활 데이터가 보험 보장 항목 예측 성능 향상에 유의미한 기여를 하지 못한 것으로 추정
- 또한 각 연령별, 성별, 취미항목, 가족 부양 수를 변경하여 수치를 재측정 하여도 마찬가지로 귀무가설을 기각하는 결과를 얻을 수 없었음
- 분석 결과를 바탕으로 귀무가설이 참으로 판명되었다. 즉, 고객의 여가 생활 패턴과 보험 상품 보장 항목 간에 유의미한 연관성이 없다는 것을
- 여가생활 데이터가 추가된 데이터 셋 모델을 돌린 결과, 보험상품 column의 포함 여부에 따라(A/B TEST) 정확도에 확연한 차이가 있었다.
- 각 항목별 상관분석 및 분류 분석을 적용한 결과 어떠한 유사성, 연관성, 패턴을 발견

2) 결과 해석

- 분석 결과, 연령별 여가활동과 보험 가입 간의 유의미한 상관관계가 발견되지 않음
- 특히, 각 연령대에서 선호하는 여가활동의 유형과 해당 연령대에서 가입하는 보험 상품의 종류 간 상관관계 분석에서 통계적 정확도가 낮게 나타남
- 이는 여가활동이 보험 상품 선택에 결정적인 영향을 미치지 않음을 의미

5. 결론 및 제언

1)결론

- 이번 프로젝트에서는 보험상품 선택과 여가생활 및 소비패턴에 대한 데이터셋 모델을 분석하여, 보험상품 항목의 포함 여부에 따른 정확도 차이를 확인하였다.

- A/B 테스트 결과, 보험상품 항목이 포함되지 않은 경우 모델의 정확도가 유의미하게 향상되었다. 반면, 보험상품 항목이 포함된 경우 상관분석 및 분류 분석을 통해서는 여가생활과 보장항목 간의 유사성이나 연관성을 발견하지 못했다.

- 특히, 나이대별로 여가생활과의 연관성이 존재하였지만, 이러한 패턴이 보험상품 선택으로 이어지지 않는 것을 확인하였다.

- 따라서, 초기 가설인 '여가생활과 보장항목 간의 연관성이 있다'는 주장은 사실이 아님이 증명되었다.

- 이는 보험 회사가 상품 개발 및 마케팅 전략 수립 시 고객의 라이프스타일 요인을 단독으로 고려하는 것은 한계가 있음을 시사한다.

- 추후 고객 중심의 세분화된 상품 개발을 위해서는 고객의 관심사, 구매 행동 등 다양한 측면을 반영하는 새로운 변수들을 고려할 필요가 있다.

2)제언

- **시장 세분화 재검토**
  여가활동과 보험 가입 간의 상관관계가 없다는 분석 결과를 통해 특정 연령대나 활동 유형에 따라 보험 상품에 대한 수요가 다를 수 있다는 결론을 도출할 수 있다. 따라서 보험사들은 고객의 여가활동이 아닌 다른 요소를 기반으로 시장을 분석할 필요가 있다.

6. 부록 및 참고자료

- KOSIS에서 사용한 통계자료

- 지난_1년_동안_가장_많이_참여한_여가활동__1순위__중분류
- 지난_1년_동안_한번이상_참여한_여가활동_유형_복수응답_관광활동
- 지난_1년_동안_한번이상_참여한_여가활동_유형_복수응답_문화예술관람활동
- 지난_1년_동안_한번이상_참여한_여가활동_유형_복수응답_스포츠참여활동
- 지난_1년_동안_한번이상_참여한_여가활동_유형_복수응답_취미오락활동
- 지속적_여가활동_기간

- 금융 빅데이터 플랫폼

- 성별, 연령별 보험종류, 보장항목, 계약 건수데이터(해빗팩토리)