기계학습 과제2

4조(이희구, 유제우)

○ 결정트리

1) 펌프식 용기의 치약 소비자 세분화

- 미국 Market Studies사는 대규모 소비용품업체가 의뢰한 2단계의 치약시장 조사를 통해 소비자 1000명을 대상으로 치약에 관한태도, 의견 및 행동을 파악하고, 기존의 로션 제품에서 사용하는 것과 유사한 “펌프식 용기”로 포장한 치약 제품에 대한 관심을 측정할 목적으로 면접조사를 실시하였다.

- 만약 펌프식 용기의 치약이 펌프식 용기의 액체 비누만큼 인기가 있는 것으로 판명된다면, 이것은 전체 치약 시장에서 하나의 중요한 세분 시장이 될 것으로 판단되었다. 펌프식 치약 제품에 대한 잠재시장을 구성하는 소비자들의 특성은 어떠한지를 파악하는 조사를 추가로 실시하였다.

- 응답자 구분방법은 펌프식 용기의 액체 비누를 알고 사용경험이 있으며, 펌프식 용기의 치약 아이디어를 “매우 좋아하거나”, “약간 좋아하며”, 이러한 제품을 “구입한다”거나 “구입가능성이 높은 ”응답자는 주 예상고객(prime prospect)으로 보았고, 펌프식 용기의 액체비누를 사용해 본 경험이 없으나 이러한 형태의 제품을 알고 있고, 펌프식 용기의 치약 아이디어를 “매우 좋아하거나”, “약간 좋아하며”, 이러한 제품을 “구입한다”거나 “구입가능성이 높은” 응답자는 좋은 예상고객(good prospect)으로 보았고, 펌프식 용기의 액체 비누를 알고 있고 사용해 본 경험이 있으며, 펌프식 용기의 치약 아이디어를“매우 좋아하거나” “약간 좋아 하며”, 이러한 제품을“구입가능성이 있거나”, “미정인”응답자는 평균적인 예상고객(fair prospect)로 판단하였고, 나머지 모든 응답자는 비 예상고객 (nonprospect)으로 보았다. 의사결정나무 분석에 사용한 고객구분은 주 예상고객과 좋은 예상고객을 주 예상고객으로, 평균적인 예상고객과 비 예상고객을 비 예상고객으로 구분하였다.

2) 맥주상표 인지도 조사

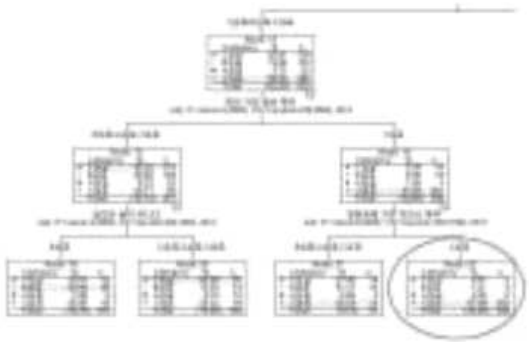
[19가지 광고표어]

- 가장 많이 판매되는 맥주
- 가장 오래된 전통맥주
- 매력적인 광고의 맥주
- 급조된 술이 아니다.
- 대중이 좋아하는 술
- 맛이 가장 좋은 맥주
- 가장 오래된 맥주회사
- 숙녀용 맥주
- 스포츠맨용 맥주
- 알코올 도수가 낮은 맥주
- 야외활동시 가장 맛있는 맥주
- 사교모임시 가장 맛있는 맥주
- 남자에 가장 잘 어울리는 맥주
- 가장 풍미가 도는 맥주
- 가장 숙성된 맛이 나는 맥주
- 젊은이용 맥주

〈그림-18 선호맥주에 따른 1차 마디 생성〉

- 미국 Brewer사는 아리조나, 뉴멕시코, 텍사스 주 지역에서 맥주류 제품을 판매하고 있다. 1000명의 남성 맥주 음주자를 대상으로 광고 카피 전략에 활용할 수 있는 상표인지도 조사를 실시하였다. 19가지 광고표어를 응답자에게 보여주고 각 광고 표어별로 먼저 생각나는 상표가 무엇인지를 묻고, 응답자의 맥주 음주량 수준, 가장 선호하는 맥주 상표를 물었다.

〈그림-19 선호맥주에 따른 2차 마디 생성〉



- “순수한 만족”이란 광고 카피로 인지하는 맥주 상표의 종류에 따라 응답자들의 맥주상표 선호도 차이가 나는 것으로 나타났다. <그림-19 선호맥주에 따른 2차 마디 생성> Brewer사의 X상표의 맥주를 가장 선호하는 사람들은 X상표가“순수한 만족”, “맛이 가장 좋은 맥주”, “일한 후에 가장 맛있는 맥주”로 인지하고 있다.

- 맥주 음주량의 수준별로 상표인지도를 살펴보았을 때 대량소비 응답자집단은 C상표를“가장 풍미가 도는 맥주”로 인식하는 집단이다. 가장 대량 소비를 하는 응답자 집단은 C상표를 “가장 풍미가 도는 맥주”로 인식하고 경쟁상표(B상표,A상표,C상표)를“야외 활동 시 가장 맛있는 맥주”로 인식하고, B상표, X상표를“맛이 가장 좋은 맥주”로 인식하는 집단이다.

- 가장 대량 소비를 하는 응답자 집단은 C상표를 “가장 풍미가 도는 맥주”로 인식하고 경쟁상표(B상표,A상표,C상표)를“야외 활동 시 가장 맛있는 맥주”로 인식하고, B상표, X상표를 “맛이 가장 좋은 맥주”로 인식하는 집단이다.

- 1인당 한 달 평균 맥주 구입비용이 소량 소비 집단 1만원, 보통 소비 집단 2만원, 대량 소비 집단 5만원으로 가정 하였을 때“가장 풍미가 도는 맥주”, “일한 후 가장 맛있는 맥주”로 X상표를 인지하는 집단의 1인당 추정 맥주 구입비용은 3.5만원인 것으로 나타났고 이는 전체 집단 평균의 1.26배에 해당하는 것으로 나타났다. 1인당 추정 맥주구입 비용이 4.18만원으로 가장 큰 집단은 C상표를“가장 풍미가 도는 맥주”로 인식하고 X상표를 “야외 활동 시 가장 맛있는 맥주”로 인식하는 집단이다.

3) 가정용품 구매 정보원

- 미국 가정용품회사의 마케팅관리자는 구매자들이 구매결정을 할 때 몇 개의 다른 정보원을 활용하는지 조사하였다. 활용하는 정보원의 수에 차이가 있는 구매자 집단 간에 어떤 차별 특성이 있는지 분석해 보면 보다 효과적인 광고 및 판촉 캠페인을 설계하는데 도움이 될 것으로 판단되었다.

- 표본조사를 통해 최근에 가정용품을 구입한 1000가구를 대상으로 1)친구와 이웃사람들, 2)책과 잡지, 3)매체광고, 4)광고 팜플렛과 소책자, 5)소매점의 점원, 6)인터넷 등의 정보원 활용 여부, 구입 고려한 가정용품의 수, 가정용품 구입비, 고려했던 상표의 수, 과거와 동일상표 구입여부, 가장의 연령과 학력, 가족소득, 자녀의 수 등을 포함하는 25개 변수에 관한 정보를 수집하였다.

- 응답자의 특성을 나타내는 변수의 수가 25개나 되므로 교차 집계로 중요 변수를 파악하기는 힘들다. 의사결정 나무분석의 CHAID 알고리즘에 의해 모든 2원 분할 중에 가장의 학력이 가장 설명력이 높은 변수로 선정되었다. 정보원의 대량 사용자는 가장의 학력이 고졸 이상이라는 증거를 갖는 셈이 된다. CHAID 알고리즘은 전체표본을 하위표본의 종속변수 평균 간에 통계적으로 유의미한 차이가 되도록 2개의 하위표본으로 분할하는 변수만을 선정한다. CHAID 알고리즘은 2개 이상의 독립변수 조합이 종속변수에 미치는 효과를 잘 이해할 수 있게 해준다. 가장 정보원을 많이 활용하는 집단은 고졸 이상의 가장, 2개 이상의 상표를 고려하고, 가정용품 구입 예상 수량이 2개 이상인 가구 집단이다.

○ 앙상블 학습과 랜덤포레스트

1) 키넥트에서의 신체 트랙킹

- 엑스박스 360에서 사용되는 모션 캡처 주변기기인 키넥트에서는 랜덤 포레스트를 이용하여 주어진 입력에서 신체의 각 부분을 분류한다. 훈련 단계에서는 미리 신체부분들이 라벨화(pre-labeled) 되어있는 깊이 지도(depth map)에서 2,000개의 픽셀(pixel)을 임의로 추출해 300,000장의 깊이 지도당 트리 하나씩 총 세개의 깊이(depth) 값이 20인 트리를 구성한다. 트리 구성 시 1,000개의 코어 클러스터(core cluster)를 사용한 GPU연산으로 하루정도의 시간이 걸린다. 테스트 단계에서는 앞서 구성한 트리들을 이용하여 임의의 입력 깊이 사진의 배경을 제외한 모든 픽셀들을 분류하는데 엑스박스 GPU에서 5밀리초(200프레임/초)의 시간 성능을 보인다.

2) 컴퓨터 단층 촬영 영상 내 해부학 구조 검출 및 위치파악

- 안토니오 크리미니시(Antonio Criminisi) 등은 랜덤 포레스트를 이용하여 3차원 컴퓨터 단층 촬영 영상(Computed Tomography, CT) 내에서 주어진 복셀에 대해 해당 되는 해부학 구조가 어디인지 검출하고 해당 위치를 파악하는 방법을 제안하였다. 훈련 단계에서는 해당 복셀이 어떤 해부학 구조인지 그리고 해당 구조의 바운딩 박스(bounding box) 정보를 가지는 55개의 복셀 포인트들을 가지고 6개의 트리를 통해 동시에 학습한다. 테스트 단계에서는 각 트리로부터 얻어진 주어진 복셀이 각 해부학 구조에 포함될 확률들의 평균을 통해 최종 확률을 계산하여 이 확률이 해당 해부학 구조에 속할 가능성($\in[0, 1]$)이 0.5 이상인지 아닌지를 확인하여 결정한다.

3) 다채널 자기공명영상 내 고악성도 신경교종 검출

- 마이크로소프트 연구소, 브라운 대학교, 그리고 캠브리지 대학교 연구팀은 다채널 자기공명영상(Multi-channel Magnetic resonance image)으로 촬영된 뇌 영상에서 고악성도 신경교종(High-grade gliomas)를 검출하기 위해 랜덤 포레스트를 적용하였다. 기존의 많은 종양 검출 연구들이 전체적인 종양덩어리를 검출하는데 초점을 맞췄다면 이 연구팀에서는 랜덤 포레스트가 내재적으로 다중 클래스 특성을 지니는 점을 이용하여 서로 다른 종류의 신경 조직을 동시에 검출하는 연구를 진행하였다. 특히 기존의 신경 조직 검출 알고리즘에 비하여 전처리과정 등이 많이 요구되지 않으며, 비교적 간단한 복잡도 모델로 높은 검출 정확도를 보이는 결과를 얻어 랜덤 포레스트의 유용성을 증명하였다.

4) 기업 부도, 주가 예측

- 금융 분야에서 랜덤포레스트의 응용은 매우 다양하다. 랜덤포레스트는 부도 예측과 주가 예측과 같은 금융 응용 분야에서 중요한 역할을 한다.

- 예를 들어, 금융 회사 XYZ사는 랜덤포레스트를 활용하여 부도 예측 모델을 구축했다. 이 모델은 기업의 재무 정보, 시장 동향, 경영 효율성 등 다양한 데이터를 종합적으로 분석하여 기업이 부도나 신용 불량 상태에 빠질 가능성을 예측한다. 이를 통해 XYZ사는 부도 위험을 줄이고 금융 거래의 안전성을 확보합니다.

- 뿐만 아니라, 주가 예측에서도 랜덤포레스트가 활용된다. 금융 기관과 투자 회사는 랜덤포레스트를 사용하여 주가 예측 모델을 개발하며, 이 모델은 주가의 움직임을 예측하여 투자 전략을 수립하는 데 도움을 준다.

- 랜덤 포레스트를 활용한 이러한 금융 모델은 데이터 기반의 의사결정을 향상시키고, 금융 기관과 투자자들이 더 나은 의사결정을 내릴 수 있도록 도움을 주며, 금융 분야에서 중요한 도구로 자리 잡고 있다.

4) 질병진단 및 환자 예후 예측

- 랜덤포레스트는 의료 분야에서 매우 중요한 역할을 하며, 이를 통해 질병 진단과 환자 예후 예측이 개선된다. 의료 영상 데이터 분석 분야에서 랜덤포레스트는 종양, 질병 또는 이상을 탐지하고 예측하는 데 사용된다.

- 예를 들어, 의료 기관 ABC병원은 랜덤 포레스트를 도입하여 의료 영상 데이터를 자동으로 분석하고 종양을 자동으로 감지하는 의료 영상 분석 시스템을 개발했다. 이 시스템은 의료 전문가에게 정확한 정보를 제공하여 조기 진단과 치료 기회를 높이며, 환자의 생존율을 향상시킨다. 랜덤포레스트를 활용한 의료 분야의 이러한 응용은 의료 현장에서 중요한 도구로 사용되고 있다.

4) 판매 및 마케팅

- 랜덤포레스트는 판매와 마케팅 분야에서 고객 세그멘테이션 및 제품 추천과 같은 중요한 과제에 활용된다.

- 예를 들어, 소매업체 DEF사는 랜덤 포레스트를 사용하여 고객 데이터를 체계적으로 분석하고, 특정 고객 그룹에 대한 맞춤형 제품 추천을 구현하여 판매량을 상당히 증가시켰다. 이러한 데이터 기반 의사결정은 고객 만족도를 향상시키고 매출을 증가시킵니다.

4) 환경 모니터링

- 랜덤 포레스트는 환경 모니터링 분야에서도 큰 역할을 한다. 주로 대기질 예측 및 기후 모델링에 사용되며, 환경 모니터링 기관 EFG기관은 랜덤 포레스트를 활용하여 대기질 데이터를 체계적으로 분석하고, 대기 오염 이벤트를 정확하게 예측하여 대중에게 경고를 제공한다.

- 이를 통해 환경 보호가 강화되고 시민 건강이 보호되며, 랜덤 포레스트는 환경 모니터링 분야에서 중요한 도구로 활용되고 있다.

5) 제조업 품질 향상

- 제조업체 GHI사의 경우, 랜덤 포레스트를 활용하여 제품의 불량률을 예측하고 품질 관리를 개선하였다. 이 과정에서 제조과정 중 다수의 변수와 센서 데이터를 모니터링하며, 랜덤포레스트를 이용하여 어떤 인자가 제품 불량과 관련이 있는지를 식별했다.

- 이 결과로 불량률이 감소하고 제품의 품질이 상당히 향상되었다. GHI사는 고객 만족도를 향상시키고 불량으로 인한 손실을 줄이는데 크게 기여하였다.

5) 교통 혼잡 예측

- 교통 기관 JKL기관은 랜덤 포레스트를 활용하여 도로에서 발생 가능한 교통 혼잡을 예측하고 효과적인 교통 흐름 관리를 수행하고 있다. 이 모델은 교통사고 발생 가능성과 교통 흐름에 영향을 미치는 다양한 변수를 고려하여 도로 안전성을 향상시키고 교통 혼잡을 줄이는데 기여하고 있다. 이로써 교통 유동성이 향상되고 교통사고 발생 가능성이 감소하고 있습니다.

6) 신용평가 및 대출 승인

- 신용평가 기관 MNO는 랜덤 포레스트를 활용하여 대출 신청자의 신용등급을 예측하고 대출 승인 결정에 활용한다. 이 모델은 대출 신청자의 신용 이력, 소득, 고용 상태, 부채 정보, 거주지 정보 등 다양한 요인을 종합적으로 고려하여 대출 승인 여부와 대출 조건을 결정한다.

- 이를 통해 MNO는 신용 등급이 높은 대출 신청자에게 더 낮은 금리와 더 나은 조건을 제공하고, 신용 위험이 높은 대출 신청자에게는 조건을 조정하여 위험을 최소화한다. 이는 대출 거래의 안정성을 확보하고 동시에 고객 만족도를 향상시키는 데 기여한다.