Marketing

2020년 봄 학기

2020년 2월 29일

https://www.github.com/KU-BIG/KUBIG_2020_Spring

1. 데이터 사이언스(Data Science)

데이터 사이언스(Data Science)란, 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는 데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야이다. 그리고 데이터 사이언스는 데이터를 통해 실제 현상을 이해하고 분석하는 데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.

따라서 데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다.

2. 마케팅(Marketing).

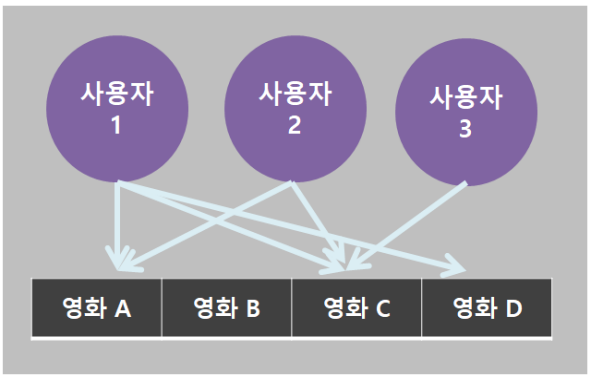
방대한 양의 데이터가 계속 생성이 되고, 이를 잘 활용하는 것이 점점 중요해지면서, 데이터 사이언스는 생물학, 의학, 공학, 사회학, 인문과학 등의 다양한 분야에 폭넓게 응용되고 있다. 특히 마케팅 분야에서 새로운 마케팅 방법으로 떠오르고 있는데 그 사례와 이론 방법에 대해 이야기 해보려 한다.

2.1넷플릭스(Netflix)의 추천 시스템

일정 금액을 지불하면 영상 컨텐츠를 맘껏 볼 수 있는 온라인 동영상 스트리밍 서비스인 넷플릭스는 데이터 과학을 활용하여 고객들에게 영상 추천을 해주는 시스템을 운영하고 있다. 넷플릭스 전체 시청의 75%가 추천을 통해 이루어질 정도로, 넷플릭스의 운영의 핵심은 추천 시스템에 있다고 해도 과언이 아니다. 추천 시스템에 활용되는 방식은 크게 '협업 필터링(collaborative filtering)', '내용 기반 필터링(content-based filtering)', 모델 기반 협력 필터링(Model-based collaborative filtering)으로 구분해볼 수 있다.

2.1.1협업 필터링(collaborative filtering):

협업 필터링은 기존 사용자의 행동 정보를 분석해 해당 사용자와 비슷한 성향의 사용자들이 기존에 좋아했던 항목을 추천하는 방식이다. 예를 들어 영화 A를 본 사람들이 영화 B를 시청한 경우가 많으면 영화 A를 보는 사람에게 영화 B를 추천해주는 방식이다. 이 알고리즘은 결과가 직관적이며 항목의 구체적인 내용을 분석할 필요가 없다는 장점을 가진다. 이러한 전략을 사용하는 경우, 비슷한 패턴을 가진 사용자나 항목을 추출하는 기술이 핵심적이기 때문에 행렬분해(Matrix Factorization), k-최근접 이웃 알고리즘(k-Nearest Neighbor algorithm; kNN)등의 방법이 많이 사용된다.



왼쪽의 그림은 이해를 돕기 위해 예시상황을 설정하고 이를 도식화한 것이다.

사용자1과 사용자2 모두 영화 A와 영화C를 둘 다 시청하였으므로, 영화 C를 본 사용자3에게 영화 A를 추천할 것이다.

그러나 협업 필터링은 기존의 자료를 필요로 하기 때문에, 기존에 없던 새로운 항목이 추가되는 경우, 추천이 곤란해진다. 이러한 문제점을 '콜드 스타트'라고 하는데, 이는 초기 정보 부족의 문제를 일컫는 말이다. 즉, 새로운 항목이 추가되는 경우, 이를 추천할 수 있는 정보가 쌓일 때까지 추천이 어려워진다.

또한, 협업 필터링은 계산량이 비교적 많은 알고리즘이기 때문에 사용자 수가 많은 경우에는 효율적으로 추천할 수 없다는 단점을 가진다.

마지막으로 롱테일 문제를 갖고 있다. 항목이 많고 다양하더라도, 사용자들은 소수의 인기 있는 항목에만 관심을 보이기 마련이다. 따라서 사용자의 관심이 적은 다수의 항목은 추천을 위한 충분한 정보를 제공하지 못할 수 있다.

2.1.2 내용 기반 필터링(content-based filtering):

내용 기반 필터링은 아이템의 특징을 기술하는 정보와, 사용자의 기호를 가지고 있는 프로파일을 비교하여 사용자에게 필요한 정보를 추천하는 방식이다. 쉽게 설명하자면, '슈퍼맨'이라는 영화에 대한 사전 분석을 통해 SF, 영웅 등의 특징을 기록하고 이 영화를 본 사람에게 '배트맨'이라는 비슷한 종류의 영화를 추천해주는 것이다. 이 기법은 항목 자체를 분석하는 것이 중요하므로 아이템 분석 알고리즘이 핵심적이며, 따라서 이를 위해 군집분석(Clustering analysis), 인공신경망(Artificial neural network), tf-idf(term frequency-inverse document frequency)등의 기술이 사용된다.

내용 기반 필터링은 내용 자체를 분석하는 것이기 때문에 협업 필터링에서 발생하는 콜드 스타트 문제를 자연스럽게 해결할 수가 있다. 그러나 다양한 형식의 항목을 추천하기 어렵다는

단점이 있다. 예를 들어 음악과 사진 비디오를 동시에 추천해야 하는 경우, 각각의 항목에서 얻을 수 있는 정보가 다르기 때문에 프로파일을 구성하기 매우 어려워진다.

2.1.3 모델 기반 협력 필터링(Model-based collaborative filtering):

각각의 필터링 방식이 갖는 문제점을 극복하기 위해, 넷플릭스는 새롭게 모델 기반 협력 필터링을 개발하여 활용하고 있다. 내용 필터링을 고도화한 이 방식은 현재 아주 높은 정확도와 만족도를 자랑하고 있다. 모델 협력 필터링은 기존 항목 간 유사성을 단순하게 비교하는 것에서 벗어나 자료 안에 내재한 패턴을 이용하는 방식이다. 다시 말해, 기존의 내용 필터링 방식이 사용자1에게 사용자1과 높은 유사도를 보이는 다른 사용자의 시청 영상을 그대로 추천했다면, 모델 필터링 방식은 이 정보를 단순하게 그대로 사용하는 것이 아니라, 주위의 정보를 이용해 세부적인 선호 이유를 유추하는 것이다. 사용자가 어떤 영화를 주연배우 때문에 좋아할 수도 있고, 노래 때문에 좋아할 수도 있고 그 장르를 선호해서 선택했을 수도 있다. 모델 필터링 방식은 많은 양의 정보를 분석함으로써 이러한 이유를 알아내고, 이를 추천에 이용하는 것이다. 즉, 자료에 내재되어 있는 패턴을 알아내는 것이 핵심적인 기술이기 때문에, LDA(Latent Dirichlet Allocation), 베이지안 네트워크(Bayesian Network) 등의 알고리즘이 사용된다.

2.2 통신사의 고객 이탈 방지

고객이 더 이상 서비스 이용이나 제품 구매를 멈추는 것을 우리는 '고객 이탈'이라고 말한다. 거의 모든 사업 영역에서 고객 이탈이 문제가 되지만, 특히 문제가 되는 분야가 바로 통신 산업분야이다. 현재의 통신 산업은 시장의 포화상태로 인해 성장이 둔화되고 있으며 최근에는 번호이동성 제도의 시행으로 인해 고객의 이탈이 매우 활발하게 일어나고 있다. 그러나 새로운 고객을 확보하는데 드는 비용이 기존 고객을 유지하는 비용보다 훨씬 크기 때문에, 이 '고객

이탈'을 잘 관리하는 것은 매우 중요하다. 기존 고객들 중 이탈할 가능성이 높은 고객들을 미리 알아내 막을 수 있다면 그 효과는 단순히 더 많은 신규 고객을 확보하는 것보다 클 수 있다. 따라서 통신사는 데이터 과학을 활용하여 이탈 고객을 예측하고, 그 고객이 이탈하지 않도록 하는 마케팅 전략(대체로 요금할인, 부가데이터 제공과 같은 인센티브 제공)을 수립한다.

2.2.1 로지스틱 회귀분석(Logistic regression):

로지스틱 회귀분석은 고객 이탈을 관리하는 데 활용되는 대표적인 방법 중 하나이다. 로지스틱 회귀분석이란 선형 회귀분석(Linear regression)과 달리 종속 변수가 범주형 데이터인 경우 사용하는 기법이다. 즉, 예측하고자 하는 것이 수치화 된 데이터가 아닌 변수일 때 쓰인다.

궁극적으로 로지스틱 회귀분석을 활용해 고객의 이탈 '확률'을 알고자 하는 것이며 이 확률을 통해 고객이 이탈할 지 유지할 지를 정할 수 있다. 일반적인 회귀분석으로 고객의 이탈을 예측할 경우 실제 확률이 0과 1 사이를 벗어날 수 있다. 그러나, 로지스틱 회귀모형에서는 Logit변환을 통해 확률을 0과 1사이로 제한을 할 수가 있다.

로지스틱 회귀분석을 통해 고객 이탈을 예측하는 과정을 간단히 설명하자면, 가장 먼저 데이터 탐색을 통해 이탈여부와 각 독립 변수간의 관계를 파악해야 한다. 그 다음 적절한 예측력을 가지고 있다고 생각되는 독립 변수를 선정해주는데, 일반적으로 일반적으로 IV(Information Value)가 0.1과 0.5 사이에 있다면 예측력이 적절하다고 판단한다. Train Data을 가지고 모델을 생성해주고, 독립 변수 간 다중공선성 문제도 해결해주면, 이탈 확률을 예측해주는 모델을 완성할 수가 있다.

2.3 CJ그룹의 SNS 소비 트렌드 분석

수치적인 데이터를 잘 분석하는 것도 중요하지만, 실제로 고객들의 목소리인 정성적 데이터들을 잘 수렴한다면 더욱 더 섬세하게 고객의 니즈를 만족시킬 수 있을 것이다. CJ그룹의 만두 마케팅은 이를 아주 잘 활용한 성공적인 사례에 속한다. CJ그룹은 비비고 만두를 출시한 뒤, 최근 3년 동안의 한국인의 만두 소비와 관련하여 각종 SNS 글 약 42억만 건을 조사하였다. 조사한 결과, '만두와 맥주 안주'를 키워드로 언급한 글이 3만 5천건 -> 4만 9천 건 -> 7만 3천건으로 점점 증가하는 것을 확인할 수가 있었다. SNS 분석을 통해 맥주 안주로 만두가 소비되고 있다는 트렌드를 읽은CJ그룹은 이에 맞추어'맥주 안주 마케팅'을 펼치기 시작했다. 그 결과 만두 매출이 늘었고, 비비고 브랜드가 만두 시장에서 압도적인 1등으로 자리매김할 수 있었다.

2.3.1 텍스트 마이닝(Text Mining):

CJ그룹이 고객의 언어를 모으고, 이를 통해서 인사이트를 도출할 수 있었던 비결은 텍스트 마이닝에 있었다. 텍스트 마이닝이란, 말 그대로 글을 캐낸다(마이닝은 'mining'으로 '(광산을) 채굴하다'라는 뜻) 의미이다. 즉, 정형화되지 않은 텍스트 데이터를 수집하여 구조화한 후 이로부터 의미 있는 정보를 추출하는 것이다.

텍스트 마이닝은 크게 자료 처리과정(data processing)과 자료 분석(data analysis)으로 나눌 수 있다. 자료처리과정은 비구조화 데이터를 분석에 용이하도록 가공 및 정제하는 단계이고, 자료 분석은 데이터마이닝, 머신러닝 등을 활용하여 텍스트로부터 유의미한 정보를 추출하는 단계이다.

텍스트 마이닝과 관련한 분석 기능으로는 전처리(Preprocess), 연관분석(Associate), 군집분석(Cluster), 요약(Summarize), 범주화(Categorize)가 있다. 전처리는 데이터를 준비하고

데이터를 해당 소프트웨어에 삽입, 가공 및 정제하는 일련의 과정이다. 연관분석은 단어들 간의 동시 발생하는 빈도수를 바탕으로 연관성을 찾아내는 것이며, 군집분석은 유사한 객체들까지 군집하여 동일한 그룹을 묶는 것이다. 요약은 텍스트에서 중요한 개념들을 요약한 것으로 일반적으로 빈도수가 높은 단어들을 반환하는 것이며, 범주화는 사전 정의된 범주로 텍스트들을 분류하는 것을 의미한다.

2.4신한카드의 고객 맞춤형 타깃 마케팅

신한카드는 데이터 분석을 통해 고객 맞춤형으로 타깃 마케팅을 진행하고 있다. 그 중 하나가 바로 Code9이다. 이는 고객의 특성에 따라 남녀고객을 아래와 같이 총 18가지로 분류하고, 각 분류 별로 특별한 명칭을 지어주는 것이다.

이렇게 분류한 기준에 따라 신한카드는 고객 특성 별로 특화된 카드 상품을 개발하였다. 고객마다 원하는 카드 혜택이 다르기 때문에, 고객 맞춤형으로 개발한 카드 상품은 누적 500만매를 돌파하며 성공적인 결과를 낳았다.

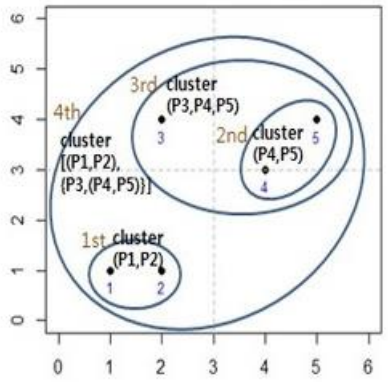
빅데이터를 활용한 신한카드의 또 다른 마케팅으로는 초개인화 서비스가 있다. 초개인화 서비스란, 비슷한 카테고리의 사람을 묶는 개인화 서비스를 넘어서, 고객 한 사람의 그때그때 상황에 따라 다른 서비스를 제공하는 것을 말한다. 그야말로 고객 단 한 명을 위한 서비스다.

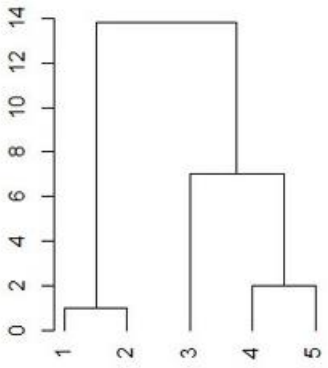
신한카드는 빅데이터 분석과 AI 알고리즘을 사용해 2만 5000개의 소비패턴을 정립하고, 고객의 취향과 상황에 따라 필요한 혜택을 제공하고 있다. 예를 들면, 똑같은 커피 할인 쿠폰이어도 고객의 소비패턴에 따라 어떤 고객에게는 출근시간에, 어떤 고객에게는 점심시간에 전달하는 것이다. 또한 이를 어떤 때에는 '오늘 커피 한잔 어떠세요'와 같은 부드러운 메시지로, 어떤 때는 '이 쿠폰 오늘 안 쓰면 손해입니다'와 같은 단도직입적인 메시지로 각각 다르게 전달한다.

2.4.1 군집분석(Clustering):

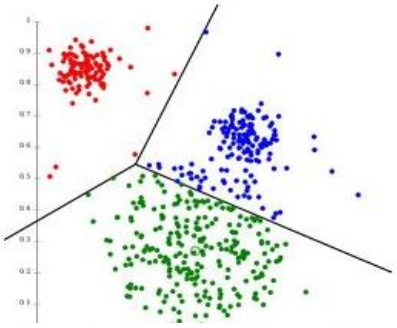
신한카드가 소비자의 성향에 따라 고객을 분류하고, 해당 군별로 마케팅 전략 수립을 다르게 할 수 있었던 것은, 고객정보와 카드 결제내역을 군집분석 하였기 때문이다. 군집분석이란 입력된 데이터들의 값에 따라 어떤 데이터들이 좀 더 비슷한 성질을 가지고 있는지 파악하여 비슷한 것들끼리 군집으로 묶어주는 분석방법이다. 이렇게 대상을 몇 가지 그룹으로 묶어서 일반화시키는 것은 복잡하고 다양하게 나타나는 대상들을 쉽게 이해하는 데 도움이 된다.

군집분석은 계층적 군집분석과 비계층적 군집분석으로 구분할 수가 있다. 먼저, 계층적 군집분석이란 한 군집이 다른 군집을 포함할 수 있는 구조로 군집을 만드는 방법이다. 따라서 아래와 같은 형태를 띤다.

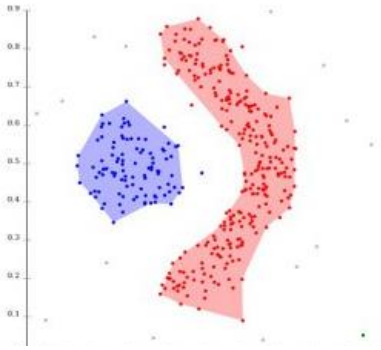




비계층적 군집분석이란 군집끼리 포함관계를 이루지 않고 서로 독립적인 한 군집으로 만드는 방법이다. 신한은행의 code9은 고객들의 특성을 서로 독립적인 군집들로 만든 것이기 때문에, 비계층적 군집분석을 활용한 것이라 할 수 있다. 비계층적 군집분석은 아래와 같은 형태를 띤다.



거리를 기반으로 군집화하는 방법(K-means)



밀도를 기반으로 군집화하는 방법(DB-SCAN)