▣ 데이터마이닝(Data Mining)

1. 데이터마이닝의 개념

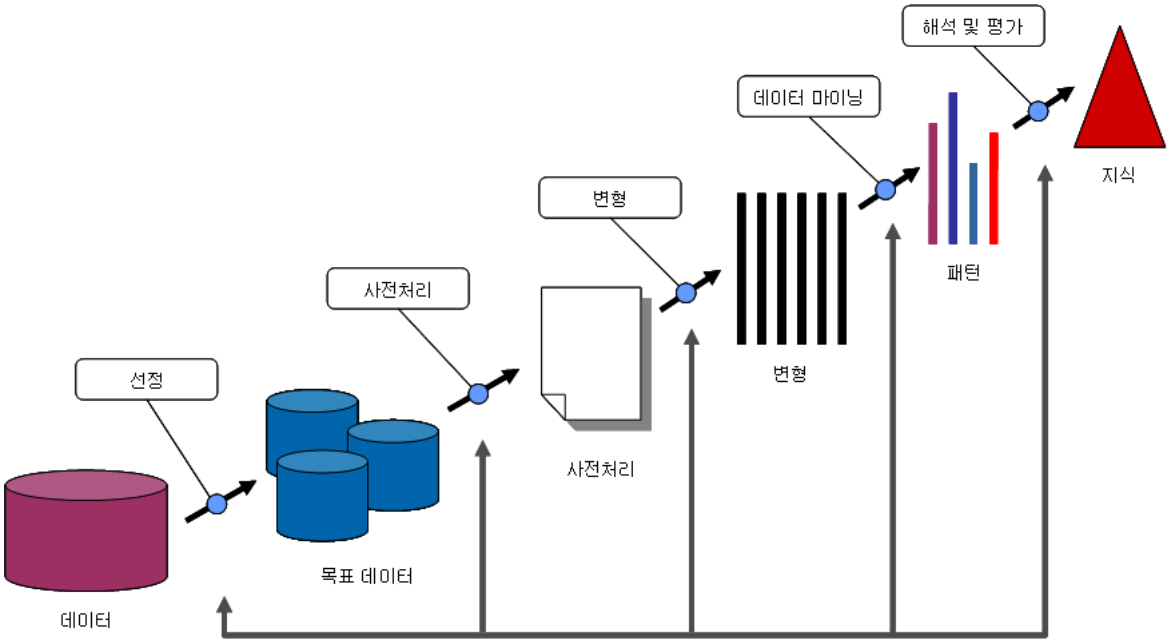
데이터마이닝이란 대용량의 데이터로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화 함으로써 유용한 지식을 추출하여 비즈니스 전략에 활용하기 위한 기술이라고 할 수 있다.

이러한 데이터마이닝이 최근 주목 받고 있는 이유는 크게 두 가지 요인 때문이다. 첫 번째 요인은 급격히 변화하고 있는 비즈니스 환경과 지속적으로 진보하고 있는 정보기술로 인한 것이다. 인터넷과 디지털로 대변되는 21세기 정보화 사회에서는 여러 가지 새로운 현상들이 나타난다. 소비자가 과거에 비해 엄청나게 많은 정보를 접하고, 또한 전자상거래를 통한 소비자와 생산자와의 직접적인 연결 가능성 등은 새로운 소비패턴의 등장을 초래하였다. 이미 많은 분야에서 시장이 포화상태에 접어든 것도 새로운 현상중의 하나이다. 시장의 포화상태에서는 시장 점유율을 늘이기가 매우 어려우며, 이를 극복하기 위하여 일대일 마케팅(one-to-one marketing)을 수행할 필요성이 크게 증가하였으며, 데이터마이닝을 통하여 여러 가지 새로운 일대일 마케팅 방법을 개발하는 방법이 크게 인기를 누리고 있다. 또한, 소비패턴의 다양화에 따른 새로운 시장의 형성이 있다. 새롭게 형성된 시장에서는 기존의 비즈니스모델이 비효율적이며 새로운 상황에 맞는 새로운 모델이 필요하다. 이러한 새로운 비즈니스 모델을 데이터마이닝을 통하여 찾고자 하는 요구가 증가하고 있다.

두 번째 요인으로는, 급속히 발달한 컴퓨터와 더불어 성장한 데이터마이닝 도구의 발전을 들 수 있다. 모든 자료의 디지털화에 따른 거대자료의 출현은 새로운 종류의 거대한 자료의 출현을 이끌었다. 신용카드, 이동전화와 인터넷 등을 통하여 얻을 수 있는 소비자의 정보는 새로운 분석방법을 필요로 하게 되었다. 새로운 종류의 자료와 더불어 발전한 기계학습(Machine Learning)에 대한 이론, 데이터베이스(Database), 그리고 여러 가지 시각화 기술(Visualization technology)은 새로운 비즈니스 환경에서 요구되는 문제들을 거대한 자료를 바탕으로 효과적으로 풀 수 있는 방법을 제공할 수 있도록 하기 때문이다.

2. 데이터마이닝의 단계

데이터마이닝을 수행하기 위한 단계는 선정(Selection), 사전처리(Pre-processing), 변환(Transformation), 데이터마이닝(Data mining), 그리고 해석 및 평가(Interpretation and evaluation)의 5단계로 구성된다. 이러한 데이터마이닝의 각 단계에 대해 자세히 살펴보면 다음과 같다. ①선정(Selection) 단계에서는 데이터마이닝의 대상이 되는 데이터 집합을 어떠한 기준에 따라 선택하고 분할할 것인지를 결정한다. ②사전처리(Pre-processing) 단계에서는 데이터 집합 내에 존재하는 불필요한 데이터를 제거하여 데이터마이닝의 속도가 느려지지 않도록 미리 가공하는 데이터 클리닝(Data cleaning)의 과정을 거친다. ③변환(Transformation) 단계에서는 데이터마이닝 단계에서 이용되는 다양한 기법들에 동일하게 적용 가능하도록 하기위해 데이터 집합 내에서 동일한 의미를 지니는 변수들의 변환 및 중첩과 같은 작업을 수행한다. ④데이터마이닝(Data mining) 단계에서는 분류, 연관, 순차패턴, 군집화, 예측과 같은 다양한 기법들을 적용하여 데이터에서 의미 있는 정보를 추출하는 단계이다. ⑤해석 및 평가(Interpretation and evaluation) 단계에서는 데이터마이닝 단계에서 발견한 의미 있는 정보를 인간의 의사결정을 지원할 수 있도록 지식으로 해석하고, 이를 평가하는 과정을 수행한다. 아래 그림에서는 이러한 데이터마이닝의 단계를 보여주고 있다.



한편, 데이터마이닝 솔루션 업체인 SAS사에서는 SAS Enterprise Miner를 이용하여 데이터마이닝의 전체 과정을 효과적으로 수행할 수 있도록 하기위해, Sampling, Exploring, Modifying, Modeling, Assessing의 5단계로 구성되는 SEMMA 방법론을 제시하고 있다. Enterprise Miner는 SEMMA의 모든 단계에 해당하는 기능을 GUI 환경으로 제공하고 있고, 특히 회귀모형, 의사결정나무, 신경망 등 일반적으로 많이 이용하는 모델을 이미 제공하고 있기 때문에 분석 모델 구축에 소요되는 시간을 줄일 수 있다. 또한, 필요한 경우 분석자가 직접 분석 모형을 구축할 수 있도록 사용자 정의 모형 역시 제공하고 있다.

SEMMA 방법론의 각 단계에 대해 자세히 살펴보면 다음과 같다. Sampling 단계에서는 대량의 데이터로부터 분석에 이용할 데이터 집합을 정의한다. Exploring 단계에서는 정의된 데이터 집합을 통계적인 기법 혹은 그래프를 이용하여 탐색하여 데이터의 속성을 파악한다. Modifying 단계에서는 분석 알고리즘의 요구를 충족시키기 위해 데이터를 조정한다. 이 단계에서는 기존 변수 변환, 이상치 및 결측치 탐색 및 대체, 변수 선택 등과 같은 방법이 이용된다. Modeling 단계에서는 회귀모형, 의사결정나무, 신경망 등의 분석 모델을 이용하여 데이터로부터 지식을 추출한다. Assessing 단계에서는 분석 모델로부터 도출된 결과를 비교하고 평가한다.

3. 데이터마이닝의 기법

데이터마이닝의 근본적인 목표는 예측과 설명이다. 예측이란 관심 주제의 알려지지 않은 미래 값을 추정하기 위해 데이터베이스 내에 있는 기존의 변수를 이용하는 것이며, 설명이란 데이터와 이에 따른 결과로 발생하는 패턴의 발견을 목적으로 한다. 이러한 목표를 달성하기 위해 다양한 데이터마이닝 기법들이 이용되는데, 이를 정리하면 다음과 같다.

① 분류(Classification) : 분석의 대상이 되는 데이터 집합을 분류 기준에 따라 몇 개의 소집단으로 분류하거나 예측을 하기위해 이용되는 기법으로, 사전에 학습된 내용을 근거로 데이터를 분류하기위한 규칙을 유도하고, 이러한 규칙에 따라 입력되는 데이터가 어떠한 클래스로 분류되는지를 예측할 수 있다. 대표적인 분류기법으로는 의사결정나무와 인공신경망이 있다.

② 연관성(Associations) : 분석의 대상이 되는 데이터 집합에 있는 항목 혹은 개체 집합 내에서 빈번한 패턴, 연관성, 상관관계 등을 찾아내는 기법이다. 연관성을 평가하기 위해 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)라는 세 가지 척도를 사용하여 데이터 집합 사이의 유사성을 찾아낸다. 대표적인 기법으로는 장바구니 분석으로 알려진 연관성 분석이 있다.

③ 순차패턴(Sequential pattern) : 일정한 기간에 대한 동향을 분석하기 위해 데이터 집합을 분석하는 기법으로, 시간적인 개념이 추가되었다는 것을 제외하면 연관성과 동일하다. 대표적인 기법으로는 순차 연관성 분석이 있다.

④ 군집(Clustering) : 주어진 데이터 집합이 가지고 있는 이질적인 특성을 유사성을 바탕으로 동질적인 군집으로 분할하는 기법으로, 전체 데이터의 분포상태나 패턴 등을 찾아내는데 유용한 기법이다. 대표적인 기법으로는 클러스터링 분석이 있다.

⑤ 예측(Prediction) : 어떤 사건에 대한 가능성을 예측하기 보다는 어떤 변수에 대한 미래의 값을 예측하기 위한 기법이다. 대표적인 예측 기법으로는 회귀분석(Regression), 신경망(Neural Network) 등이 있다.

이러한 데이터마이닝 기법들은 대상이 되는 문제의 영역에 따라서 독립적으로 적용할 수 있으나 여러 기법들을 혼합해서 적용하기도 한다.

데이터마이닝을 이용한 콜센터 최적화

최근 IT기술이 급속도로 발전하면서 한꺼번에 많은 양의 정보가 보다 빠르게, 보다 정확하게, 고객이 원하는 정보만이 선별되어 전해지고 있다. 흔히 말하는 CRM의 한 방법이다. CRM의 제일선에 자리하는 콜센터는 고객들의 소리만 해결해주거나 단순히 전화 주문만을 받아서 처리하던 수준에서 이제는 적극적인 홍보의 가장 중요한 채널로 자리 잡아가고 있다.

CRM 마케팅 활동으로 보다 많은 고객들을 흡수하게 되면 부가적으로 고객의 소리 역시 그만큼 늘어나게 된다. 이에 따라 고객의 목소리를 들어주고, 고객에게 직접적으로 캠페인을 행하는 상담요원의 수요를 더욱 많이 요구하게 된다. 그러나 상담원수는 한정되어 있고 제한된 상담원으로 폭주하고 있는 콜량을 모두 수용하기란 거의 불가능에 가까운 일이다. 회사 입장에서도 늘어나는 콜에 대응하기 위해서 무한정의 투자를 지속하지는 못할 것이다. 제한된 콜센터의 규모와 동일한 수의 상담원 하에서 콜량만 늘어나면 우선적으로 피해를 보는 것이 고객이다. 고객이 상담전화를 했을 때 대기시간은 점점 늘어나고, 상담원들이 처리해야 할 업무량이 늘어나면서 고객응대의 질을 저하시킬 것이며, 이러한 것은 결국 고객 불만의 증대로 이어지고, 궁극적으로는 회사 이미지에 커다란 악영향으로 돌아오게 될 것이다. 더욱이 늘어나는 통화량과 과도한 업무증가로 인한 상담원의 높은 이직률 또한 콜센터가 당면한 문제점이다.

따라서 콜센터의 최적화를 통해서 저비용 고효율의 콜센터를 구축하고, 편리한 관리운영체제 확립 및 자원의 효율적 분배를 통해서 대고객서비스의 질적인 향상을 이루는 것만이 제반의 모든 문제점들을 해결할 수 있는 방안일 것이다. 결국 한정된 규모와 상담원을 이용해서 늘어나는 통화량을 처리할 수 있는 효율적인 방법이 요구된다.

이를 해결하기 위한 한 방법으로 콜센터의 방대한 통화 자료를 데이터마이닝을 이용해서 분석하고 이것의 결과를 이용해서 콜센터의 ROI를 올릴 수 있는 방법이 필요로 한다. 여기서는 아웃바운드 콜센터에서 데이터마이닝을 이용한 콜센터 최적화 사례를 소개하고자 한다.

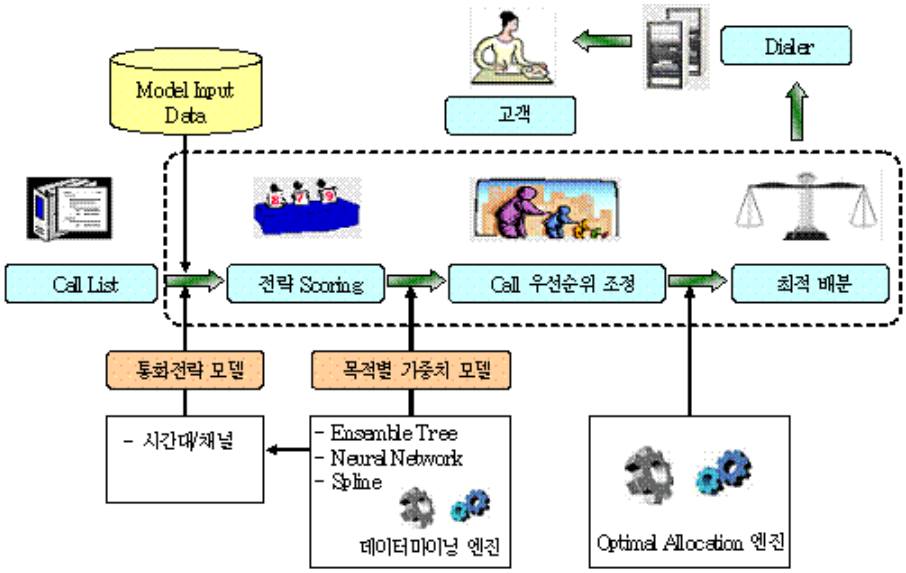
아웃바운드 콜센터의 경우 마케팅의 중요한 채널로서 새롭게 인식되고 있지만 현재까지 이를 제대로 활용하는 업체는 그리 많지 않은 듯하다. 2005년을 기준으로 볼 때 아웃바운드와 인바운드의 비율이 해외의 경우 4:6 정도이지만 국내의 경우

2:8 수준으로 아직까지는 아웃바운드를 이용한 캠페인보다는 걸려오는 고객에 대한 적절한 대응에 더 신경을 쓰고 있는 양상을 보인다.

또한, IT산업의 변화와 CRM의 발전은 현재 고객 데이터 분석 단계를 넘어 CRM과 고객과 직접 접촉하는 유력한 채널인 콜센터를 접목시키고 있다. 상담원을 통해 들어온 고객 정보는 바로 ODS(운영 데이터 저장고)에 저장되며 데이터마이닝과 같은 고객 정보 분석 솔루션을 통해 원투원 마케팅을 가능하게 해준다. 이러한 변화는 아웃바운드의 역할이 앞으로 점점 더 증폭되리라는 것을 예상할 수 있으며, 아웃바운드 콜센터의 운영에 있어서 효율화라는 문제가 더욱 심각하게 대두되고 있다.

아웃바운드의 경우 인바운드와 달리 현재의 고민거리는 대상 고객의 증가에도 불구하고 낮은 통화 연결율과 연결이 되더라도 본인 통화율이 떨어져 하나의 캠페인을 제대로 처리하는 비율이 떨어지고 있다는 것이다.

따라서, 이런 문제의 해결을 위해 Auto Dialer 도입을 통해서 캠페인의 처리율을 증대시킬 필요성이 제기되고 있다. 이를 위해서는 고객별 최적 통화가능 시간대와 채널을 예측하고, 예측된 결과를 이용해서 Auto Dialer와 연동하여 콜 목적에 따른 콜 우선순위를 배분하는 작업이 필요하다. 이에 대한 process를 그림으로 표현하면 다음과 같다.



콜 배분 프로세스에서는 우선 고객들의 프로파일 정보와 Dialer Log Data, 상담이력 데이터와 거래 데이터를 이용해서 데이터마이닝 기법을 이용한 통화 전략 모델을 만들게 된다. 다음으로 이제까지는 단순히 calling list의 순서대로 무작위로 뿌려지던 방식에서 통화전략 모델링에서 만들어진 score를 calling list와 접목시켜

콜 우선순위를 조정하고 캠페인의 처리율을 높이기 위한 최적배분을 하게 된다.
그 내용을 세부적으로 살펴보면 다음과 같다.

첫 번째, 통화전략모델은 어느 시간대에 어느 채널로 고객에게 전화를 했을 때 본인 통화율이 가장 높은가를 지정해주는 모델이다. 예를 들어, 회사원에게 평일 낮 시간에 자택으로 전화를 한다면 본인 통화율이 떨어질 것이다. 즉, 이제까지의 경험이나 동일한 그룹의 통화 형태를 이용해서 본인 통화율이 가장 높은 시간대와 채널을 찾는 모델이다.

두 번째로는 목적별 가중치 모델이다. 통화전략모델에서 아무리 본인 통화율이 높은 고객이라도 캠페인의 효과가 없는 고객에게 우선적으로 전화할 이유가 없다. 즉, 본인 통화율에 마케팅 효과가 높은 정도의 가중치를 부여함으로써 마케팅 효과가 높고 본인 통화율이 높은 고객에게 우선적으로 통화를 하자는 전략이다.

마지막으로 최적배분 효과에 대해서 간략히 살펴보자.

기본적인 아이디어는 각 시간대별 최대처리 가능 수와 통화전략 모델링에서 만들어진 score를 이용해서 전체적으로 가장 높은 통화연결 score를 가지는 전략을 만들어 내는 것이다.

예를 들어 아래 표와 같이 각 고객별 통화가능 score를 계산하였다고 하자. 그리고 오전/오후 각 실행 가능한 아웃바운드 콜 수가 2건이라고 가정하자.

고객 구분	오 전	오 후
고객 1	0.85	0.47
고객 2	0.26	0.64
고객 3	0.68	0.18
고객 4	0.70	0.68

각 고객 score는 이해하기 쉽게 하기 위해서 예측된 통화성공률이라고 하자. 가장 쉽게 순차적으로 배분을 하는 경우는 각 시간대별로 가장 높은 점수를 갖고 있는 고객들을 우선적으로 배분하는 것이다. 오전 시간대의 경우 가장 점수가 높은 고객1과 고객4가 배분될 것이고, 오후 시간대의 경우 이들을 제외한 고객 2와 고객 3이 배분될 것이다. 그리고 이 배분 방법을 평가하기 위해서 단순히 성공률의 합을 계산하면 2.37이 된다. 그러나 고객3의 경우 고객 4보다 성공률은 다소 낮지만 오후 시간대의 성공률이 현저한 차이를 보이고 있다. 만약 오전시간대에 고객1과 고객3을 그리고 오후시간대에 나머지 고객1과 고객 4를 배분한다면 이 배분방법에 의한 점수의 합은 2.85로서 순차적인 배분에 비해 약 20.3%의 성공률의 증대 효과

를 볼 수 있다. 최적 배분모델에서는 이러한 고객 통화율에 대한 최적의 배분을 추구한다.

이와 같은 콜센터 최적화 전략은 국내에서는 최초로 개발된 아이디어로서 이미 모 카드회사에서 전체적으로 10% 이상의 효과를 보았으며, 본 전략을 이용한 솔루션이 개발 완성 단계에 있다. 다만 이 기술을 시스템화하기 위해서는 calling list를 자동적으로 상담원과 연결해주는 기본적인 시설이 갖추어져 있어야 한다. 이 기술은 기본적으로 100석 이상의 상담원을 보유한 콜센터에서 보다 효율적으로 활용할 수 있으며, 최소 6개월 정도의 콜 데이터를 필요로 한다. 주요 수요처로는 개인 고객을 상대하는 금융기관의 서비스센터와 최근 새로운 수익 사업으로 떠오르고 있는 인터넷 쇼핑몰, 홈쇼핑 업체와 제조분야로 그 영역이 점차 확대되고 있는 추세다.