經營論集, 第51卷 統合號 (2017年 12月)

보험 상품 가입 예측 모형*

송 인 성**

본 연구에서는 판별 분석, 로지스틱 회귀, 의사결정나무 모형, 그리고 신경망 모형 등의 데이터 마이닝 기법을 활용하여 자동차 보험 가입자들의 자차 보험 가입 여부를 예측하는 문제를 풀어 보았다. 사용된 데이터 마이닝 모형들이 예측력 측면에서 별로 우수하지 않은 것으로 나타났다. 입력 변수가 13개에 불과하지만 훈련 자료 내의 관측치가 7만 건 이상인 점을 고려하면, 이러한 낮은 예측력의 이면에는 활용된 자료 내의 입력 변수들이 출력 변수를 예측하는 데 있어 유용성이 높지 않을 가능성이 존재한다. 즉 보험 가입 내역에 나타난 고객 특성과 차량 특성들을 가지고 자차 보험 가입 여부를 추정하고 예측하는 데 한계가 존재한다는 것이 명확하다. 즉 고객들의 자차 보험 가입 여부는 관측되지 않은 다른 변수들에 의해 결정되는 부분이 매우 크기 때문에 자동차 보험 가입 내역에 관측된 변수들만 가지고 자차 보험 가입 여부를 예측하는 데에는 한계가 존재한다.

따라서 고객 행동에 대해 마케터가 더 잘 이해하고 예측할 수 있기 위해서는 자동차 보험 가입 내역에 들어 있는 고객 특성과 차량 특성 외에 추가적인 고객 데이터가 필요하다. 이러한 문제는 자동차 보험 회사뿐만 아니라 다양한 기업들이 현재 가지고 있는 문제라고 할 수 있다. 많은 기업들이 다년간 누적되어 온 고객 수준의 거래 자료를 활용하여 미래의 고객 행동을 이해하고 예측하려고 시도하지만 기업들이 가지고 있는 자료의 내용에서 한계가 존재하는 경우가 상당하다. 오늘날 많은 기업들이 가지고 있는 데이터의 문제는 관측치 개수의 문제가 아니라 관측 변수의 다양성과 이들의 품질에 관한 문제이다. 따라서 기존에 자료로서 고려하지 않았거나 계량적 코딩이 곤란해서 무시하였던 고객 관련 내용들을 포함하여 다양한 고객 행동 정보를 적극적으로 활용하여야 한다. 이러한 문제 극복을 위해서는 기존에 눈여겨보지 않았던, 소셜 미디어에 나타난 고객 의견이나 태도, 고객 접점 수준에서 고객 접촉 내용 데이터 등 비정형 데이터를 포함하여 이종의 고객 자료들을 융합(fusion)하는 데이터 융합 접근 방법이 필요하다.

주제어: 고객 선택 예측, 판별 분석, 로지스틱 회귀, 의사결정나무, 신경망

*본 연구는 서울대학교 경영연구소의 연구비 지원을 받아 수행되었다.

**서울대학교 경영대학 교수

I. 서론

기업이 마케팅 활동을 전개하는 데 있어 고객들의 상품 선택 행동에 대한 이해와 예측은 매우 필수적인 요소이다. 어떠한 특성을 가진 고객이 어떤 상품을 선택할 것인지를 이해하고 사전에 예측하는 능력은 고객별로 맞춤형 마케팅 프로그램을 제공하는 타깃 마케팅을 전개하는 데 있어 반드시 가져야 할 핵심적인 역량이기 때문이다. 전통적으로 마케팅 분야에서 고객 행동 예측에 대한 관심이 높았지만, 최근 정보 기술의 발달과 더불어 더욱 다양하고 방대한 고객 관련 자료들이 수집되고 누적되면서 고객 행동에 대한 보다 깊은 이해와 예측 가능성에 대한 관심이 더욱 더 높아지고 있다. 오늘날의 기업들은 이른바 빅데이터 시대라고 불리는 경영 환경을 마주하고 있는데, 기존에 수집할 수 없거나 활용할 수 없었던 다양하고 방대한 데이터들이 수집되고 있고 이를 유효 적절하게 경영 현장에서 활용하는 역량에 대한 경영자들의 관심이 고조되고 있다. 특별히 마케팅 분야는 이러한 방대한 데이터들이 생성되고 활용되는 최전선에 있다고 해도 과언이 아닐 정도로 다양한 고객 관련 데이터들이 생성되고 있다.

마케팅 분야에서는 1980년대 초반부터 마케팅 자료 분석에 대한 관심이 매우 높아지기 시작했는데, 소매 점포들이 스캐너 시스템들을 도입하면서 기존에는 없었던 스캐너 자료들이 등장하기 시작한 시점이다. 전통적으로 시장 조사에서 활용되어져 왔던 설문 조사를 통해 얻은 자료들은 응답자들의 구두 표현에 의존할 수밖에 없는 한계를 가지고 있기 때문에 응답자의 응답과 실제 고객과의 행동 양상과의 괴리라는 한계가 존재하는데, 스캐너 자료는 고객들의 실제 구매 자료이기 때문에 이러한 한계를 지니지 않으며 또한 설문 조사와는 다르게 매우 많은 고객들의 구매 행동을 오랜 기간에 걸쳐 파악할 수 있는 장점이 있다. 이러한 소매점 스캐너 자료를 활용하여 소비자 구매 행동에 대해 이해하고 이를 마케팅 프로그램에 반영하는 다양한 모형들이 마케팅 분야에서 개발되어 왔고, 이러한 의미에서 1980년대 이후 스캐너 자료 활용에 따른 계량 마케팅 분야의 발전을 마케팅 정보 혁명이라고 지칭하기도 한다(Blattberg, Glazer, and Little, 1994).

스캐너 자료는 슈퍼마켓에서 판매하는 상품들에 대한 소비자들의 구매 행동 자료에 국한되어 있는 반면, 2000년대 이후부터는 좀 더 다양한 산업 분야에서 고객 행동 자료들에 대한 관심이 높아졌다. 구매 대상이 되는 제품의 종류가 슈퍼마켓 소비재에서 다양한 카테고리의 제품으로 확대되었다. 예를 들면, 제약사에서 판매하는 전문의약품의 경우 환자들에 대한 광고 등의 직접적인 마케팅 활동이 규제 당국에 의해 제한되어 있기 때문에 제약사는 자사의

영업사원이 약품을 처방하는 의사에게 제품 관련 정보를 제공하는 디테일링(detailing)에 주로 의존하게 된다. 이러한 디테일링이 의사들의 처방 행동에 어떠한 영향을 미치는가를 이해하고자 하는 노력들의 일환으로 처방전 자료를 활용한 마케팅 연구들에서 이루어져 왔다. 고객 행동 자료에서 다양성의 확장은 제품 종류 확대에 그치지 않고 다양한 정보 채널의 활용에서도 나타나고 있다. 웹 브라우징 행동, 모바일 앱 사용 행동, 전자 상거래 구매 행동 등 인터넷과 모바일 부분에서 소비자들의 정보 탐색과 구매 행동에 대한 자료들이 활용 가능하게 되었으며, 게임이나 금융, 의료 등 서비스 산업 분야에서도 고객 자료들을 수집하고 이를 분석, 활용하고자 하는 필요성이 증가하고 있다.

본 연구에서는 금융 분야 특히 보험 산업에서 활용 가능한 고객 데이터를 이용하여 고객들의 보험 상품 선택 행동을 예측하는 마케팅 모형을 고찰하고자 한다. 개별 고객들과 기업의 다양한 접촉과 구매 및 거래 내역 자료를 활용하여 수익성 있는 고객을 파악, 발굴하여 그러한 고객들과 장기적 관계를 형성함으로써 고객 자산을 극대화하는 것이 데이터베이스 마케팅의 핵심이라고 할 수 있다(Blattberg, Kim, and Neslin, 2008). 금융 산업 분야는 이러한 데이터베이스 마케팅의 활용이 매우 중요한 산업 중의 하나이며 또한 데이터베이스 마케팅을 전개할 수 있는 고객 자료들이 일상적인 기업 운영 활동에서 자연스럽게 축적되고 있는 산업이다. 고객별로 맞춤형 마케팅 프로그램을 전개하는 데이터베이스 마케팅의 기본 개념을 활용하기 위해서는 고객들의 상품 선택 행동에 대한 정확한 예측이 선행되어야 하므로, 본 연구는 데이터베이스 마케팅 프로그램의 실질적 전개를 위한 필수 요소를 다루고 있다.

한편 본 연구에서는 최근에 각광받고 있는 다양한 데이터 마이닝 기법들을 활용하여 고객들의 보험 상품 선택 예측에 활용하고자 한다. 최근에 마케팅 자료의 종류가 다양해지고 자료의 크기가 점점 더 방대해지면서 전통적으로 사용된 분석적 기법들의 한계를 극복할 필요가 대두되고 있다. 자료의 종류가 다양해지면서 기존에 사용된 선형 모형들로서는 변수들 간의 다양한 상호 작용을 사전적으로 모형화하는 것이 여의치 않기 때문에 좀 더 유연한 모형들을 활용할 필요가 있다. 또한 방대한 자료를 이용하기 위해서는 이른바 규모성(scalability)에 대한 고려가 필요한데, 순수한 학술적 이해를 추구하는 모형에서는 자료 분석 및 모형 추정에 상당한 기간을 할애할 수 있겠으나 기업 의사 결정에 활용되기 위해서는 자료가 방대하다고 하더라도 상대적으로 짧은 시간에 자료 분석과 모형 추정을 할 수 있어야 한다. 이러한 측면을 고려하여 본 연구는 마케팅 분야에서 많이 활용되지 않았던 데이터 마이닝 기법들을 활용하여 이들 기법의 마케팅 분야 활용 가능성을 점검해 보고자 한다. 따라서 본 연구는 마케팅 자료 분석의 방법론적 경계를 확장하는 노력과도 관련이 있다.

II. 문제 정의와 분석 자료

최근 계량 마케팅 분야의 핵심 개념은 데이터를 기반으로 한 마케팅 활동이라고 할 수 있는데, 이 개념하에서는 마케팅 데이터가 기업의 마케팅 활동을 이끄는 핵심 자원이 된다. 즉 데이터 기반 마케팅 개념하에서는 데이터를 기반으로 하여 마케팅 문제를 정의하고 이에 대응하는 마케팅 모형을 통하여 문제를 해결하는 과정을 거치게 된다. 본 연구에서도 이러한 맥락에서 데이터 중심으로 문제를 파악하는 것에서 논의를 시작한다. 보험 회사에서 보유하고 있는 고객의 보험 가입 자료는 고객의 인구 특성 변수 및 고객의 상품 구매 내역이 포함되어 있는데, 본 연구에서는 익명의 자동차 보험 회사로부터 얻은 기존 고객의 자동차 보험 가입 자료를 활용한다. 해당 보험 회사의 재구매 고객으로서 서울 지역 가입자 100,197건의 보험 가입 내역 자료를 분석 대상으로 하였다. 보험 회사가 보유하고 있는 내부 자료를 활용하는 경우에는 어떤 고객이 자동차 보험을 가입할 것인지에 대한 여부를 예측하는 문제를 다루는 것이 곤란하다. 왜냐하면 기본적으로 보험사 내부 자료는 이미 보험에 가입한 고객들로만 이루어져 있어서 보험에 가입하지 않는 잠재 고객들의 비가입 행동에 대한 정보가 없기 때문이다. 자동차 보험사의 내부 자료를 활용하는 본 연구에서도 이와 같은 자료의 특성을 고려하여 해당 보험사의 자동차 보험에 가입한 고객들의, 자동차 보험 가입 여부가 아닌 다른 의사 결정 변수에 대한 연구를 진행하여야 한다. 자동차 보험 가입 내역에는 여러 가지 다양한 하위 선택 행동에 대한 정보가 들어 있는데, 대물 가입 여부, 무보험 차량 손해 가입 여부, 자손 가입 여부, 자차 가입 여부 등 기본 보험 외에 추가적인 선택적 상품에 대한 가입 내역에 대한 정보가 들어 있다.

연구 문제로서 이러한 다양한 선택형 상품의 구매 여부를 분석하는 문제를 고려해 볼 수 있다. 만약 복수의 상품 각각을 독립적인 의사 결정으로 인식한다면 전통적으로 마케팅 분야에서 스캐너 자료를 이용하여 단일 카테고리 내에서 서로 완전 대체재인 경쟁 브랜드들 중 하나를 선택하는 소비자 선택 행동을 분석하는 전통적인 브랜드 선택 모형으로 귀결될 것이다. 그러나 만약 이종의 상품들을 선택하는 데 있어 상품 종류 간의 대체나 보완 관계가 존재함을 인식한다면 기존 브랜드 선택 모형과는 달리 여러 개의 상품 구매 의사 결정을 동시에 하는 소비자의 복수 카테고리 상품 선택 행동이 주제가 되는 분석이다. 이러한 복수 카테고리 상품 선택 행동에 대한 연구 역시 마케팅 분야에서 이루어져 왔는데, Kim, Allenby, and Rossi(2002)의 연구에서는 다양성 추구 행동 모형은 소비자가 여러 개의 대안을 동시에 구매하는 것이 최적이 되는 효용 함수를 제시하였다. Gentzkow(2007)의 보완재 선택 모

형에서는 소비자의 효용 함수를 테일러 전개를 통한 근사를 통해 선택 대안 간에 대체/보완 관계를 파악할 수 있도록 하였다. Song and Chintagunta(2006)의 연구는 Gentzkow(2007)의 접근 방법을 활용하여 여러 카테고리 간의 대체/보완관계를 분석하였다. Song and Chintagutna(2007)에서는 Hanemann(1984)이 제시한 구조적 모형을 복수의 카테고리에 일반화시켜서 여러 카테고리에 걸친 소비자의 구매 바스켓 의사 결정을 하나의 프레임웍으로 모형화하였다. Lee, Kim, and Allenby(2013)의 연구에서는 선택 대안 간 보완재 관계가 존재하는 경우에 소비자들의 구매 의사 결정을 직접 효용 함수를 이용하여 모형화하였다. 이들 문헌들을 정리 요약하면, 소비자들의 다수 대안 구매 의사 결정을 포함할 수 있는 미시 경제학적 기반의 구조적 효용 함수를 어떻게 구성할 것인가가 연구의 주제들이었다.

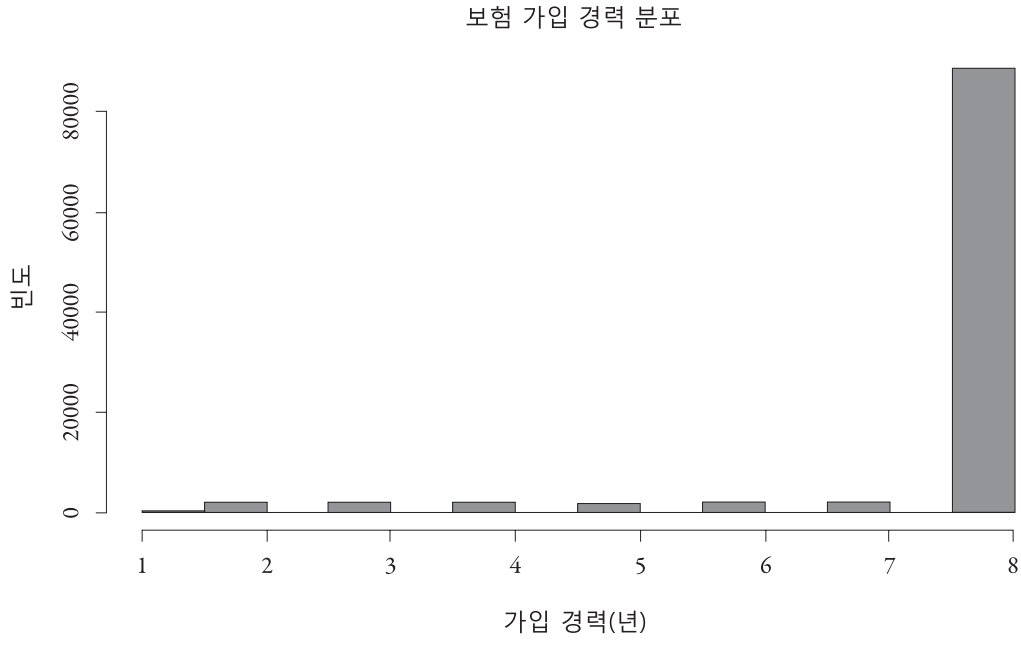
이러한 복수 카테고리 상품 선택 행동에 대한 분석은 학술적 중요도도 높고 동시에 교차판매(cross selling) 등의 실무적 응용 측면에서도 매우 의미 있는 주제이지만, 본 연구에서 다루는 특정한 자료에서는 적용이 무의미하다. 그 이유는 본 연구에서 사용되는 자료의 특성에서 기인한다. 자동차 보험 가입의 세부 내역별로 대물 가입 여부, 무보험 차량 손해 가입 여부, 자손 가입 여부, 자차 가입 여부 등에 대한 정보가 존재하지만, 자차 가입 정보에 따라 기타 상품들의 가입 여부가 사실상 정해져서 상품 간에 매우 강한 종속 관계가 존재하기 때문이다. 즉 본 연구의 자료에서는 자차 가입자의 99% 이상이 대물, 무보험 차량 손해, 자손 등의 하위 상품에도 가입한 것으로 나타나기 때문이다. 이러한 자료의 특성상 상기의 복수 카테고리 수준의 분석은 본 연구의 자료에서는 의미가 없다. 반면 자차 보험 가입 여부 자체는 가입자별로 상당한 다양성이 존재한다. 전체 100,917건의 고객 중 76%가 자차 보험에 가입을 하였고 나머지 34%는 자차 보험에 가입하지 않았다. 이러한 상당한 수준의 고객 간에 다양성이 존재하는 것은 고객 상품 선택 행동에 있어 체계적인 영향을 주는 요소들이 존재함을 의미하며, 이러한 요소들을 파악하여 고객들의 자차 보험 가입 여부를 예측하여 마케팅 프로그램에 활용할 여지를 남겨둔다고 할 수 있다. 따라서 본 연구에서는 해당 보험 회사의 자동차 보험 가입 고객 중 자차 보험 가입 여부를 예측하는 문제를 고찰하도록 하겠다. 보험 회사 내부 자료 중 자차 보험 가입과 관련하여 〈표 1〉에 제시된 다음과 같은 변수들을 활용한다.

〈표 1〉 자료 변수

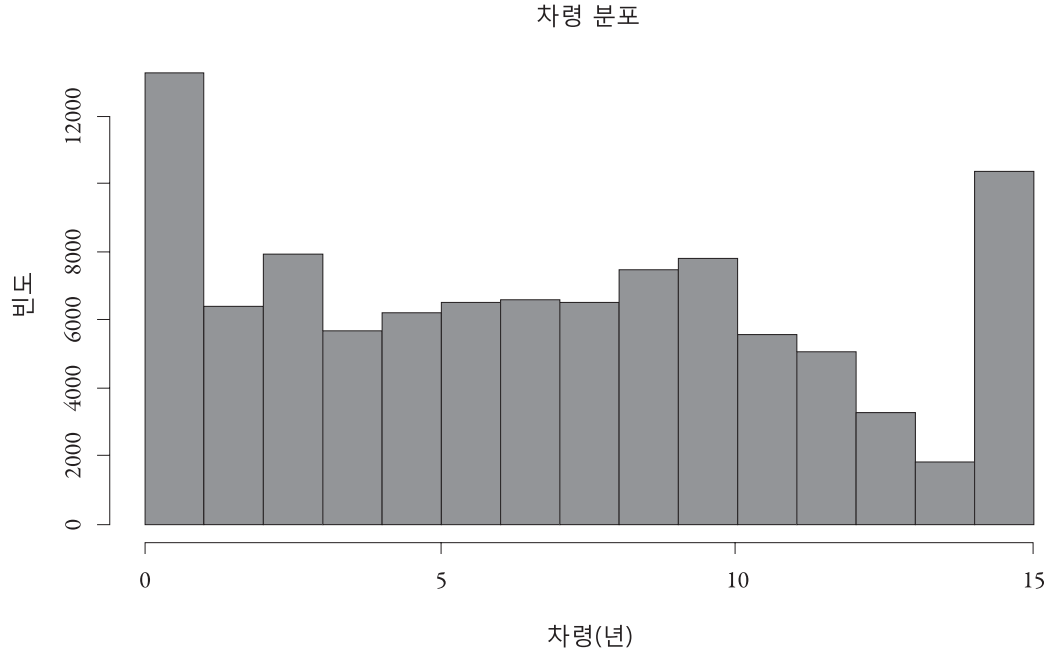
변수	변수 설명 및 단위
고객 연령	가입자 나이(년), 비율 척도
보험 가입 경력	자동차 보험 가입 경력, 서열 척도로 단계 구분
성별	남녀 구분
ABS 장착 여부	Y/N 구분
에어백 장착 여부	Y/N 구분
배기량	배기량(cc), 비율 척도
자동변속기 여부	Y/N 구분
차량 종류	자동차 종류(소형 A, 소형 B, 중형, 대형, 다목적), 명명 척도
차량 가액	금액(원), 비율 척도
차령	차량의 나이(년), 비율 척도
국산차 여부	Y/N 구분
도난 방지 장치 여부	Y/N 구분
물적 할증 기준 금액	단계별 서열 척도(50만, 100만, 150만, 200만 원 구간)
자차 보험 가입 여부	Y/N 구분

분석에 사용되는 자료에서 비율 척도로 측정된 변수들은 추후 모형 추정 단계에서의 안정성을 위해 평균을 빼고 표준 편차로 나누어 표준화시켰다. 그 외 명명 척도나 서열 척도로 측정된 변수들은 모형 추정에서 더미 변수들의 조합으로 재구성하게 된다. 한편 보험 가입 경력은 비율 척도로 측정될 수 있음에도 불구하고 서열 척도로 단계로 구분하였는데, 가입 경력 자료가 보험사 내부에서 8년 이상인 경우 특별히 구분하지 않고 '8년 이상'이라는 항목으로 처리하고 있고, 〈그림 1〉에서 나타난 바와 같이 이 항목에서 관측치가 집중되는 비대칭성이 매우 높기 때문에 서열 척도로 처리하였다. 한편 차량의 나이인 차령의 경우에도 〈그림 2〉와 같이 해당 보험사가 15년 이상인 경우에는 '15년 이상' 항목으로 단일 처리하고 있으나 보험 가입 경력의 경우와는 달리 그러한 처리가 해당 변수의 비대칭성을 높이지 않는 것으로 나타나 입력 변수의 숫자를 최소한으로 유지한다는 측면에서 비율 척도로 그대로 사용하였다.

한편 모형의 예측력을 평가하는 데 있어 모형 추정에 사용된 자료를 그대로 이용하게 되면 과적합(overfitting)하는 모형을 우수한 모형으로 인식하는 오류를 범하게 되므로 자료를 모형 추정에 사용되는 훈련 자료(training set)와 추정된 모형을 이용하여 예측력을 평가하는 평가 자료(test set)로 분리할 필요가 있다. Blattberg et. al(2008)에 따르면 일반적으로 관측치의 수치가 작을 때에는(예를 들면, $n < 100$) 전체 자료의 1/3이나 1/4 정도를 평가 자료로 보



〈그림 1〉 보험 가입 경력 분포



〈그림 2〉 차령 분포

유하고 나머지 자료를 훈련 자료로 사용하는 것이 추천되나, 자료의 크기가 커지면 훈련 자료와 평가 자료의 분리 비율에 크게 좌우되지 않는다고 한다(Blattberg et. al, 2008, 309쪽). 본 연구에서는 70%의 자료를 훈련 자료로, 나머지 30%의 자료를 평가 자료로 할당하였는데 각 관측치를 난수를 발생시켜 훈련 자료/평가 자료 중 하나에 할당하였다. 할당 결과 70,053개의 관측치가 훈련 자료로, 30,144개의 관측치가 평가 자료로 할당되었다. 각 자료군에서 자차 보험 가입률은 76.12%와 76.03%로서 두 자료군 간 차이가 거의 없는 것으로 나타났다.

III. 예측 모형 및 그 결과

자차 보험 가입 여부는 이진형(binary) 변수로서 이를 예측하는 문제는 연속형 변수를 다루는 회귀 모형과는 별도로 분류(classification) 문제로 다루어진다. 이러한 분류 모형으로서 본 연구에서는 판별 분석, 로지스틱 회귀, 의사결정나무 모형, 그리고 신경망 모형의 네 가지 접근 방법을 고려하였다. 판별 분석과 로지스틱 회귀는 선형 모형으로서 변수들 간의 상호 작용은 연구자가 사전에 모형에 명시하는 경우에만 고려된다. 즉 이러한 선형 모형에서는 입력 변수들 각각이 출력 변수인 자차 보험 가입 여부에 선형으로 영향을 준다는 인식을 하게 된다. 만약 변수들 간의 상호 작용 효과를 고려하려면 해당 내용을 연구자가 사전에 모형화하여야 하는데 사전에 알려진 이론이나 연구자의 판단에 의존하게 된다. 비율 척도가 아닌 명명 척도나 서열 척도로 측정된 변수들은 이진형 더미 변수들의 조합으로 재구성되기 때문에 본 연구에서 실질적으로 사용된 입력 변수들은 25개에 달하는데, 이 경우 어떤 변수들 간의 상호 작용을 고려하여야 하는지 사전에 정하는 것은 단순한 일이 아니다. 반면 의사결정나무 모형이나 신경망 모형은 비선형 효과 및 상호 작용 효과를 데이터를 이용하여 모형을 훈련하면서 스스로 탐지하도록 고안된 모형이라고 할 수 있다. 따라서 이들 모형에서는 연구자가 사전에 변수들 간의 상호 작용 내용을 모형화할 필요가 없는 장점이 있는 반면, 해석이나 모형의 복잡성 측면에서 단점이 있을 수 있다.

1. 선형 판별 분석

판별 분석은 출력 변수 y가 명명 척도로서 어떤 집단에 속하는지를 나타낼 때, 각 집단에 속한 관측치의 입력 변수가 다변량 정규 분포를 지니고 있다고 가정하여 각 관측치가 특정 집단

〈표 2〉 예측치와 실제 자료의 분류

구분		실제 평가 자료	
		가입하지 않음	가입
모형의 예측	가입하지 않음으로 예측	t_n	f_n
	가입으로 예측	f_p	t_p

에 속할 사후 확률을 추정하는 문제이다. 출력 변수가 자차 보험 가입 여부와 같은 이진형 변수로서 집단 1과 집단 2를 나타낼 때, 판별 함수에 근거하여 $x'\hat{\sum}^{-1}(\hat{\mu}_2-\hat{\mu}_1)>-\frac{1}{2}(\hat{\mu}_2+\hat{\mu}_1)'\hat{\sum}^{-1}(\hat{\mu}_2-\hat{\mu}_1)-\log(N_2/N_1)$의 조건이 성립하면 2번 집단으로 분류하고, 그렇지 않으면 1번 집단으로 분류하는 분석 방법이다. 여기에서 μ_k는 집단 k의 평균을 나타내고 Σ는 분산행렬이며 N_k는 훈련 자료에서 집단 k에 속한 관측치의 숫자이다. 판별 분석을 훈련 자료에 적용하여 판별 함수를 추정하고 추정된 판별 함수를 평가 자료에 적용하여 판별 분석의 예측력을 평가해 볼 수 있다. 예측력을 평가하는 지표로서, 자차 보험 가입 여부는 이진형 변수이므로 통상적으로 사용되는 평균오차제곱 등은 사용될 수 없으므로 적중률과 F1 정확도 측정치를 사용하였다.

〈표 2〉에서 제시된 바와 같이 예측 모형에서 가입으로 예측하였는데 실제 평가 자료에서도 가입으로 예측된 관측치의 경우는 True Positive(t_p)에 해당되며 예측 모형으로 가입하지 않을 것으로 예측하였는데 실제 가입하지 않는 경우는 True Negative(t_n)에 해당된다. 반면 예측이 틀린 경우로서, 가입으로 틀리게 예측한 경우는 False Positive(f_p), 가입하지 않을 것으로 틀리게 예측한 경우는 False Negative(f_n)에 해당된다. 적중률(Hit Ratio)은 전체 관측치 중에서 예측과 실제 관측치가 부합된 경우의 비율을 의미한다. 적중률이 높을수록 예측력이 높다고 평가할 수 있다.

$$Hit\ Ratio = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

한편 F1 정확도는 텍스트 마이닝 등의 머신 러닝 기법에서 주로 활용되는 예측 평가치로서 이른바 Precision과 Recall의 조화 평균(harmonic mean)으로 정의된다. Precision은 가입할 것으로 예측된 경우들 중에 실제로 가입으로 나타난 관측치의 비율이며, Recall은 실제 가입한 것으로 나타난 관측치들 중에서 가입으로 예측된 비율을 의미한다.

$$Precision = \frac{t_p}{t_p + f_p}$$

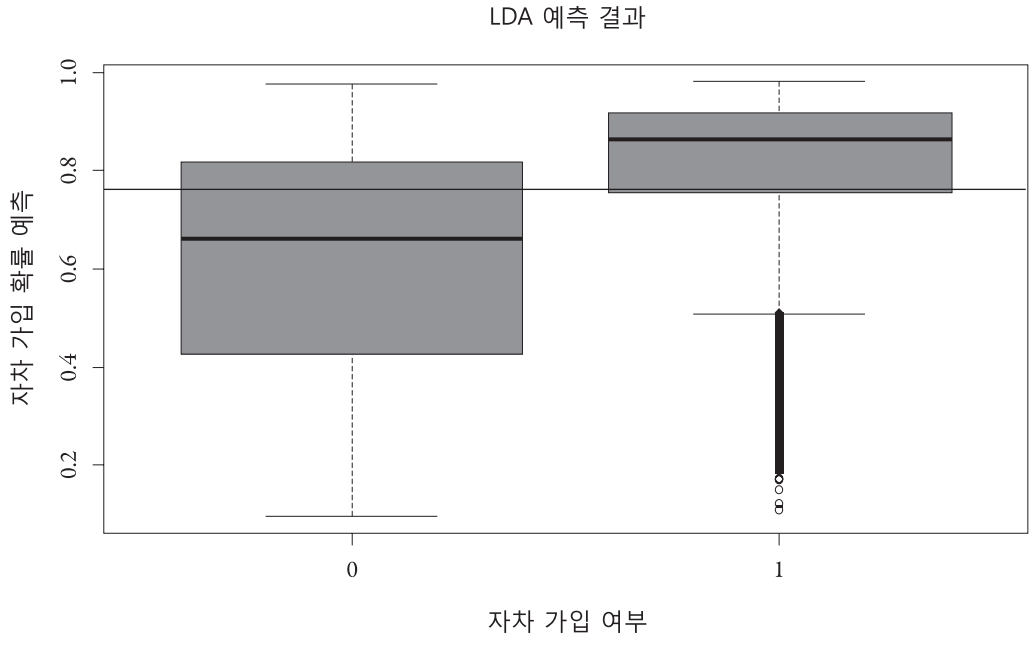
$$Recall = \frac{t_p}{t_p + f_n}$$

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

예측이 매우 정확한 경우에는 F1 정확도 값이 1에 가까우며, 예측이 매우 부정확한 경우에는 F1 정확도 값이 0에 가깝게 된다.

예측력을 평가하는 데 있어 자료의 특성에 맞추어 기준을 설정할 필요가 있다. 즉 본 연구에서 사용된 자료에서 평가 자료의 경우 사후적으로 76%의 고객이 자차 보험에 가입해 있음을 알고 있기 때문에, 어떤 모형에도 의존하지 않고 모든 고객이 자차 보험에 가입할 것이라고 예측한다면 76%의 경우에 예측이 적중하고 34%의 경우에는 예측이 빗나갈 것이다. 이 경우에 예측 정확도는 76%에 이른다. 반면 아무도 자차 보험에 가입하지 않을 것이라고 무조건 예측하는 경우에는 적중률이 34%에 이르게 된다. 만약 동전 던지기 등의 무작위 실험을 통해 50%의 확률로 자차 보험 가입 여부를 예측한다면 그러한 예측의 정확도는 평균적으로 50%에 이를 것이다. 훈련 자료와 평가 자료가 동질적인 경우에, 복잡한 데이터 마이닝 기법에 의존하지 않고 훈련 자료의 평균을 기준으로 하면 자차 보험 가입 확률이 평균적으로 76%임을 알기 때문에, 모든 고객이 자차 보험에 가입할 것이라고 예측하는 극단의 경우에도 예측력이 상당히 높다. 이 경우 적중률이 76%이고 F1 정확도값은 86.36%에 이르게 된다. 물론 이러한 예측의 경우에는 34%의 경우에 예측이 확실하게 빗나가는 것을 사전에 인지할 수 있으나, 데이터 마이닝 모형을 활용하여 예측을 하는 경우 이러한 맹목 예측(blind prediction)보다 예측력이 우수하여야 함을 인지하여야 한다.

판별 분석을 본 연구의 자료에 적용한 결과 〈그림 3〉과 같이 예측되었다. 판별 분석을 통하여 평가 자료의 각 관측치의 입력 변수값에 근거하여 각 관측치의 출력 변수값이 1번이 되는 경우, 즉 자차 보험을 가입할 확률을 예측하였고, 실제 자차 보험에 가입하지 않은 경우를 0, 실제로 가입한 경우를 1로 하여 실제 관측치들을 두 집단으로 구분하여 각 집단별 관측치에 해당하는 판별 분석의 예측 가입 확률을 박스플롯으로 나타내었다. 자차 보험에 실제로 가입한 집단의 경우 확률값이 대부분 높은 것으로 분포되었으며, 자차 보험에 가입하지 않는 집단에서는 가입 확률값이 상대적으로 낮으며 또한 그 분산이 큰 것으로 나타났다. 박스플롯에서는 선형 판별 분석 모형이 실제 가입 여부를 상당히 잘 예측하는 것처럼 보여지지



〈그림 3〉 선형 판별 분석의 예측 결과

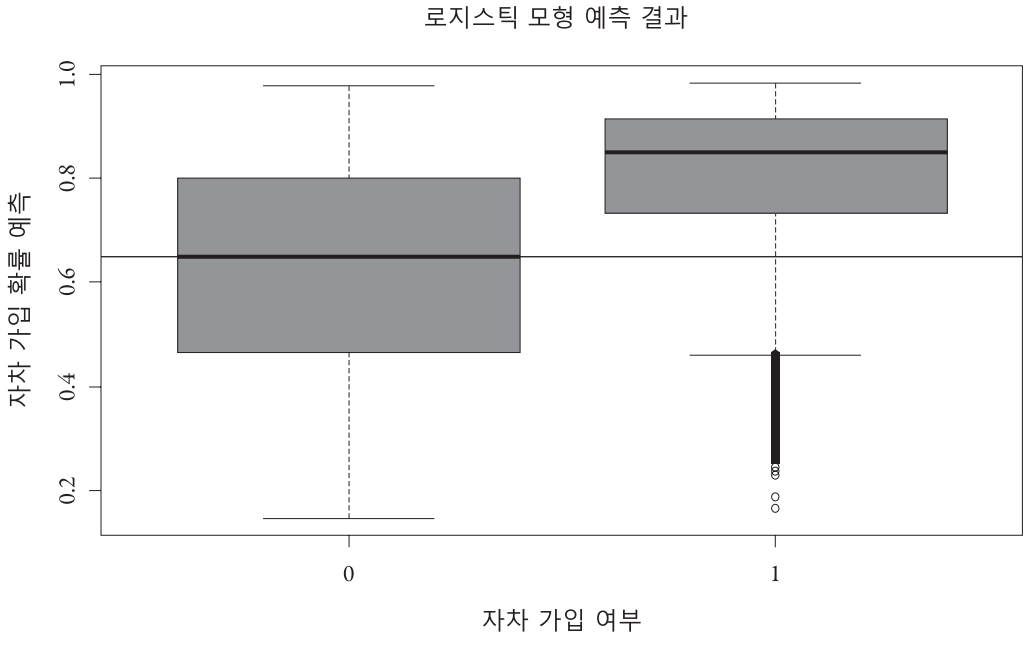
만, 판별 분석을 통한 예측의 적중률과 F1 정확도 값은 각각 77.81%와 86.37%로 나타났다. 이 수치는 앞서 언급한 맹목형 예측의 예측력을 간신히 상회하는 정도라고 할 수 있을 것이다.

2. 로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 0과 1 값만을 가지는 이진형 출력 변수에 대하여, 출력 변수가 1의 값을 가질 확률을 다음과 같은 로지스틱 함수로 모형화한다.

$$Pr(Y=1|X)=\frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)}$$

로지스틱 회귀 모형은 기본적으로 출력 변수가 1의 값을 가질 확률을 계산한다. 따라서 예측 역시 0과 1 사이의 확률값으로 나타나게 되는데, 이러한 확률값을 기초로 가입 여부에 대한 이진형 예측을 하여야 한다. 확률값이 어떤 특정한 기준값(cutoff)보다 크면 가입할 것으로 예측하고 그렇지 않으면 가입하지 않을 것으로 예측하게 되는데, 이 기준값을 평가 자료를 가지고 사후적으로 정하는 것은 예측의 의미에 맞지 않기 때문에 훈련 자료를 이용하여



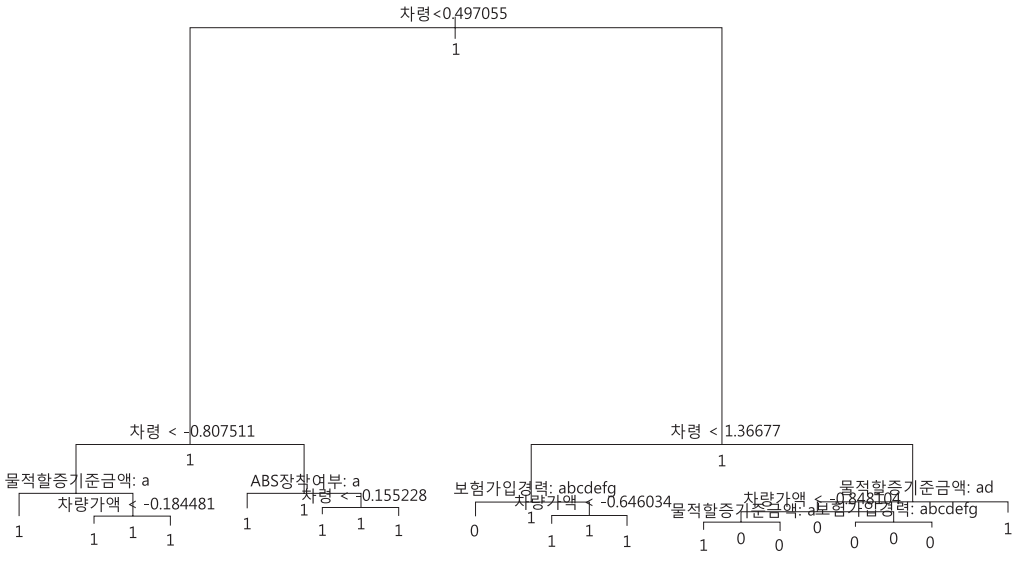
〈그림 4〉 로지스틱 회귀 모형의 예측 결과

다음과 같이 기준값을 결정한다. 훈련 자료에서 76.12%의 경우에 자차 보험을 가입하였으므로 평가 자료에서 예측된 가입 확률값들의 76.12% 분위수를 기준값으로 하여 그에 해당하는 확률보다 예측된 확률값이 크면 가입으로 예측하고 그렇지 않으면 가입하지 않을 것으로 예측한다. 즉 확률값이 76.12%보다 크거나 작은지 여부를 따지는 것이 아니라, 모든 확률값들을 내림차순으로 정리하였을 때 76.12% 분위에 해당하는 확률값이 기준값이 된다는 것이다. 〈그림 4〉와 같이 예측 결과가 나타났는데 76.12%에 해당하는 기준 확률이 76.12%보다는 낮은 것으로 나타났으며 그 양상은 선형 판별 분석 결과와 유사하다. 이 경우에도 적중률과 F1 정확도 값은 각각 76.56%와 84.62%로 나타났는데 마찬가지로 맹목 예측을 간신히 상회하는 정도에 그치고 있다.

3. 의사결정나무(Decision Tree) 모형

의사결정나무를 통한 분류 예측 기법은 전체 자료를 출력 변수의 값이 비슷한 하위 집단으로 분화하는 과정을 거치게 되는데, 그 성격상 자연스럽게 입력 변수 간의 상호 작용 효과를 모형화하게 된다. 의사결정나무에서는 전체 관측치들을 출력 변수의 값이 비슷한, 즉 순

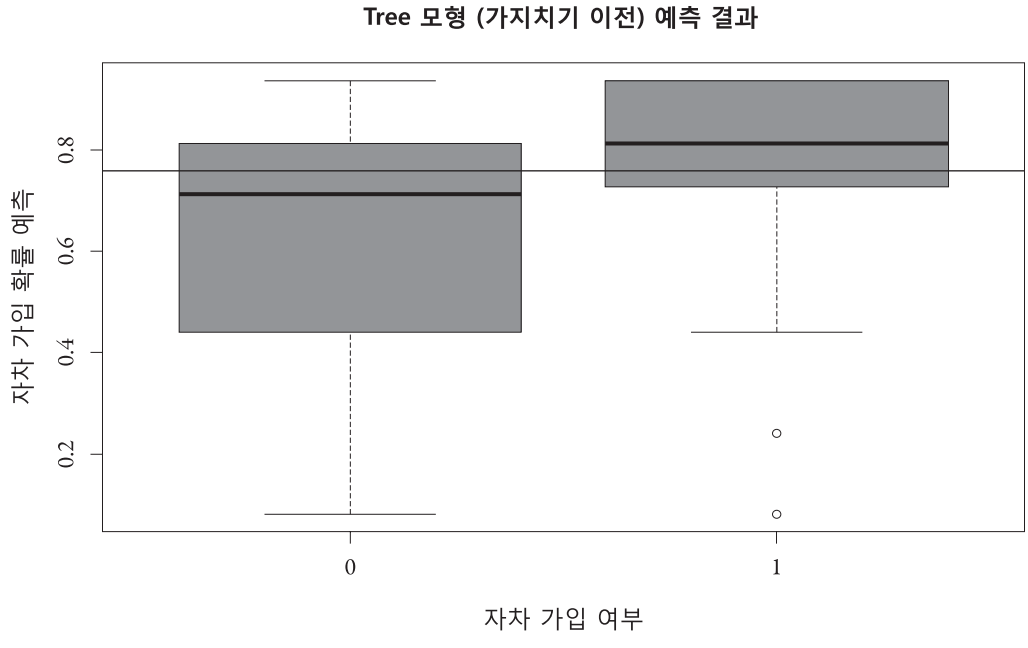
Tree 모형 결과: 가지치기 이전



〈그림 5〉 CART 모형(가지치기 이전)

수도(purity)가 높은 하위 집단으로 지속적으로 분화시키면서 커지게 되는데, 순수도를 측정하는 방법에 따라 그리고 분화시키는 알고리즘에 따라 다양한 의사결정나무 기법이 발달하였다. 의사결정나무는 출력 변수가 연속형인 경우와 이산형인 경우 모두를 다룰 수 있는, 분류 문제를 해결하는 본 연구에서는 CART(Classification and Regression Tree)와 C5.0 기법을 활용하였다. CART 기법에서는 일부러 크기가 큰 나무(oversized tree)를 생성한 후, 과적합 문제를 해결하기 위해 가지치기(pruning)를 하면서 오분류 오차를 적게 내는 간단한 나무를 생성하는 기법이다. C5.0 기법은 각 노드에서 다지 분리(multiple split)가 가능하게 하는 모형이다.

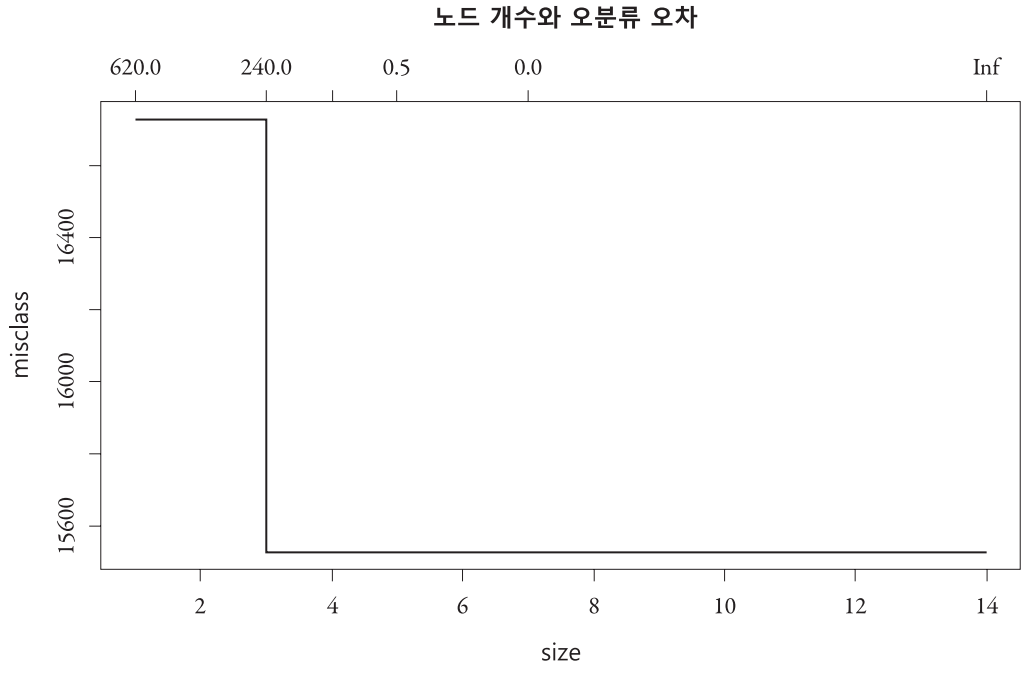
먼저 CART 방법을 적용하여 훈련 자료에 적용해 보면 〈그림 5〉와 같이 자차 보험 가입 여부를 분류하고 있는 것으로 나타난다. 이 경우에 14개의 최종 노드가 있는 이 나무 모형을 통하여 자차 보험 가입 여부를 예측한 결과의 박스플롯이 〈그림 6〉에 제시되어 있다. 박스플롯상에서는 실제로 자차 보험에 가입한 경우에 가입 확률 예측치가 높게 나타나고 있고, 실제로 가입하지 않은 경우에는 가입 확률 예측치가 낮으면서 분산도 크게 나타나고 있다. 이 모형의 적중률은 78.09%이며 F1 정확도는 86.58%이다. 마찬가지로 이 모형도 본 연구에서 사용된 자료의 자차 보험 가입 여부를 예측하는 데 있어 맹목 예측을 간신히 상회하는 정도



〈그림 6〉 CART 모형에 의한 예측 결과(가지치기 이전)

의 예측력을 보이고 있다.

한편 CART 방식에서는 이러한 분류 나무의 과적합 문제가 존재할 가능성을 인지하고 이를 가지치기를 통하여 더 간단한 모양의 나무로 축약하는데, 이때 오분류 오차를 크게 증가시키지 않으면서 더 간단한 형태의 나무 구조를 찾는 것이 가지치기의 기본 개념이다. 가지치기를 위해서는 14개의 최종 노드를 가진 복잡한 나무 구조에서 노드의 수를 줄여 후보 나무 구조들을 생성한 다음, 훈련 자료를 전체 자료인 것처럼 가정하여 K 폴드 교차 타당성 평가를 통해 오분류 오차를 각 후보 나무 구조에 대해 계산한다. 그리고 후보들 중에서 오분류 오차가 작으면서도 간단한 나무 구조를 최종 모형으로 선정한다. 그 결과가 〈그림 7〉에 제시되어 있는데, 가로축은 노드의 개수를 세로축은 오분류 오차를 나타낸다. 이 결과에 의하면, 노드의 수가 1개나 2개인 모형에 비해 노드 수가 더 많은 모형이 오분류 오차가 확연히 줄어들지만, 일단 노드의 수가 3개 이상이 되면 노드의 수가 증가되더라도 오분류 오차가 더 줄어들지 않는 것으로 나타났다. 즉 노드의 수가 3개인 간단한 모형에 비해 그보다 더 노드가 많은 복잡한 모형이 추가로 제공하는 설명력은 거의 없는 것으로 보이므로 따라서 최종 노드가 3개인 모형으로 가지치기를 하는 것으로 결정한다.

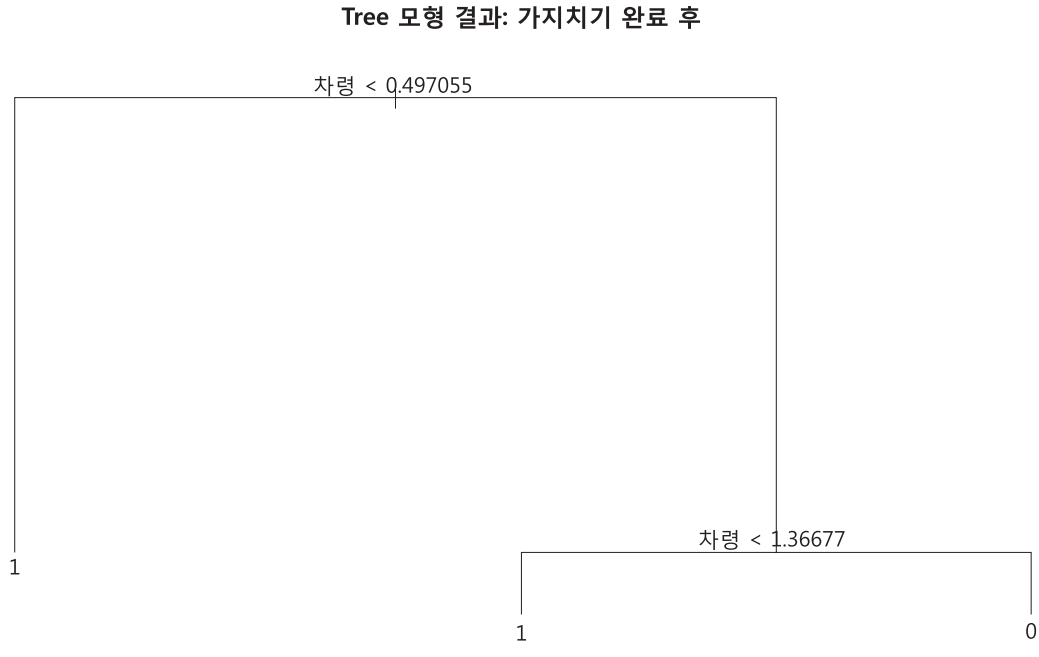


〈그림 7〉 CART 모형의 노드 개수와 오분류 오차

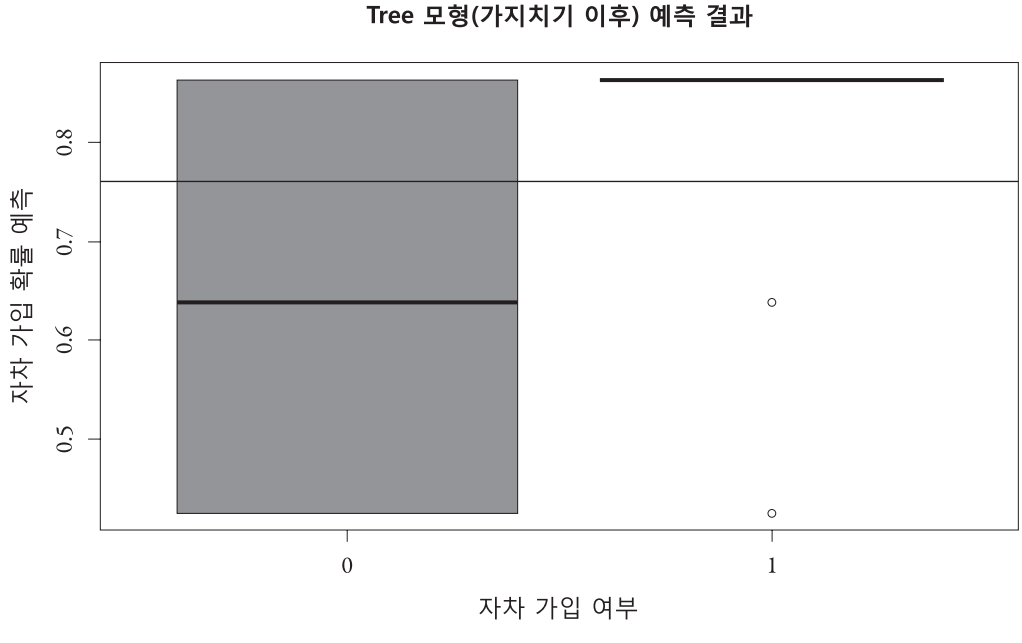
〈그림 8〉에 제시된 것처럼, 노드의 수가 3개인 CART 모형은 차령에 의해 간단하게 자차 보험 가입 여부를 예측한다. 이 모형에 의하면, 표준화된 차령값이 0.497 이하이면 자차 보험에 가입할 확률이 0.8620으로 예측되며, 표준화된 차령값이 이보다 큰 경우에는 다시 차령값에 근거하여 1.367을 기준으로 나누는데 표준화된 차령값이 0.497에서 1.367 사이에 해당하면 자차 보험에 가입할 확률이 0.6399으로 예측되며, 차령값이 1.367보다 크면 자차 보험에 가입할 확률이 0.4268로 예측된다. 제시된 그림에는 이진형 분류 예측치가 표시되어 있는데 예측된 확률이 0.5보다 크면 1, 그렇지 않으면 0으로 이진 분류하고 있다.

이러한 예측치를 실제 가입 자료와 비교한 것이 〈그림 9〉에 제시되었는데 실제로 자차를 가입한 관측치에 대해서 자차 가입 확률이 매우 높으며 그 분산은 매우 작게 나타나고 있고, 자차를 가입하지 않은 경우에는 분산이 매우 크게 나타나고 있다. 한편 이 간단한 모형의 적중률과 F1 정확도는 각각 77.43%와 86.22%로서 가지치기를 하기 이전보다 오히려 예측력이 하락하였다. 나아가 F1 정확도는 맹목 예측의 경우보다 소폭 낮은 것으로 나타났다.

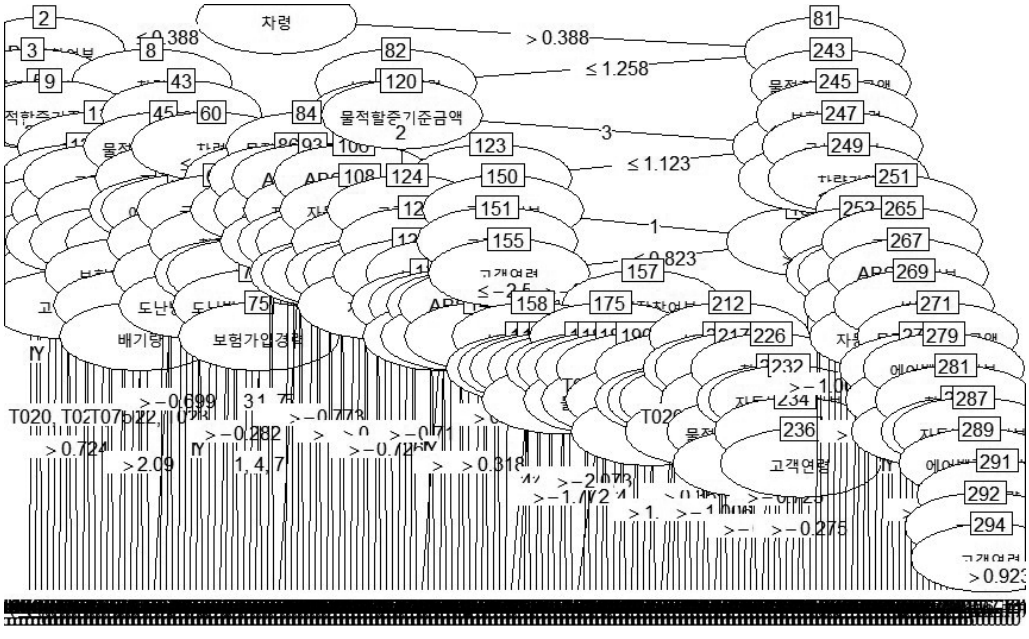
C5.0 모형을 이용하여 자차 보험 가입 여부를 예측하는 경우에, 먼저 훈련 자료에서 분류나무를 추정하였는데 이 모형에서는 노드의 수가 164개에 이르는 매우 복잡한 모습의 나무



〈그림 8〉 CART 모형(가지치기 완료 후)



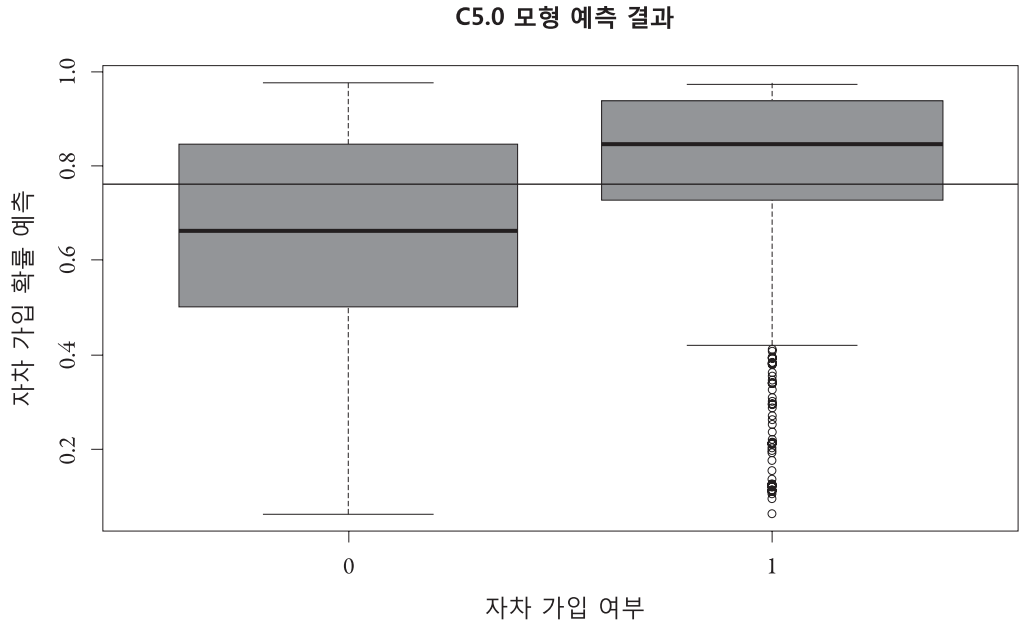
〈그림 9〉 CART 모형에 의한 예측 결과(가지치기 완료 후)



〈그림 10〉 C5.0 모형에 의한 분류 나무 구조

구조가 추정되었다. 〈그림 10〉이 C5.0 모형에 의해 추정된 분류 나무 모형의 구조인데, 그림에서 전체 노드가 완벽하게 표현되지 않을 정도로 매우 복잡한 나무 구조임을 알 수 있다

C5.0 모형에서 추정된 분류 나무 구조를 이용하여 평가 자료의 자차 보험 가입 여부를 추정한 결과는 〈그림 11〉에 제시되어 있다. 앞의 모형들과 마찬가지로 실제로 자차 보험에 가입한 관측치의 경우에는 가입 확률이 높게 예측되었고 그렇지 않은 경우에는 가입 확률이 낮게 예측되었다. 한편 예측력 측면에서 C5.0 모형이 다른 모형보다 상대적으로 좋은 것으로 나타났는데, 적중률과 F1 정확도가 각각 78.35%, 87.00%로 나타났다. C5.0 모형에서 분류 나무 구조가 너무 복잡하여 사전적으로 과적합의 문제가 존재할 가능성이 있었으나 실제 예측 결과에 의하면 이 자료에 관한 한 모형의 복잡성이 예측력을 떨어뜨리지는 않는 것으로 나타났다.



〈그림 11〉 C5.0 모형에 의한 예측 결과

4. 신경망(Neural Networks) 모형

신경망 모형은 인체 내 신경 조직의 기본 단위인 뉴런의 작동 원리를 응용한 지도 학습 기법으로서 컴퓨터 비전 등의 머신 러닝 분야에서 자주 활용되고 있다. 신경망 모형은 미지의 비선형 함수를 훈련 자료를 이용하여 근사하는 데 유용하며, 소비자 선택 모형에서 자주 활용되는 로짓 모형을 일반화한 형태로 이해되기도 한다. 본 연구에서는 신경망 중에서도 입력 변수와 출력 변수 사이에 은닉층(hidden layers)이 존재하는 다층 퍼셉트론(multilayer perceptron) 모형을 사용하는데, 특히 한 개의 은닉층을 가진 모형을 사용하였다. 다층 퍼셉트론에서 은닉층의 수는 사전적으로 정해지지 않으나 은닉층의 수가 많아질수록 과적합 문제가 발생할 가능성이 높다는 점과, 하나의 은닉층을 가지는 신경망이 임의의 연속 함수를 원하는 정확도로 근사할 수 있다는 범용 근사 정리(universal approximation theorem)에 의하면 하나의 은닉층도 충분할 수 있다는 점을 고려하여 하나의 은닉층을 가진 신경망 모형을 사용하기로 하였다. 다만 은닉층 내 은닉 노드의 수는 사전에 정하지 않고 평가 자료에서의 예측력을 기준으로 정하였다. K개의 입력 변수를 활용한 이진형 분류 문제의 경우, M개의 은닉 노드가 있는 신경망 모형은 다음과 같은 형태를 지닌다.

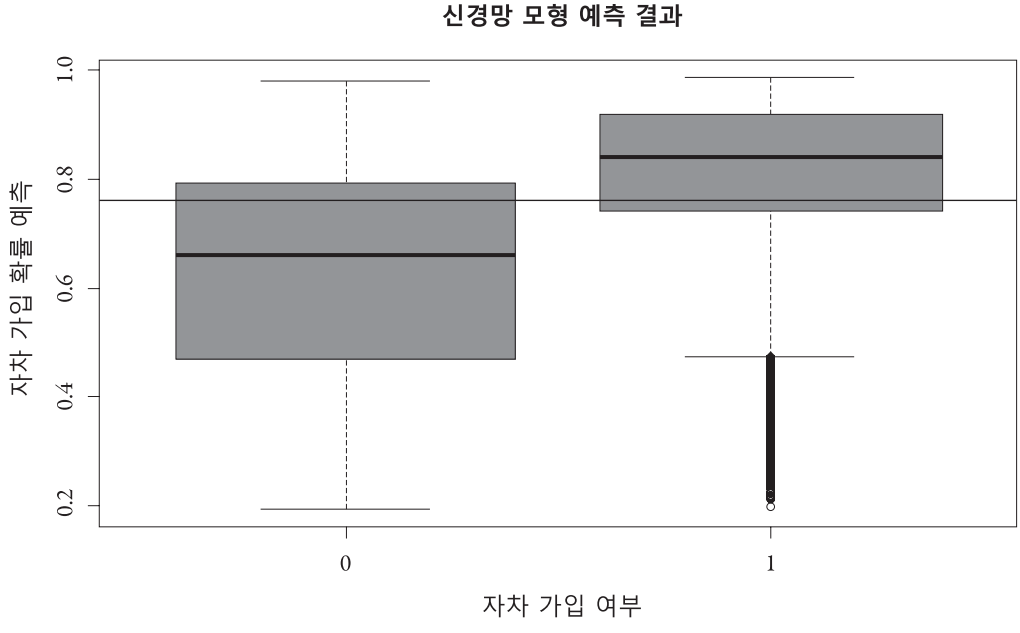
- 요약 함수: $v_m = \alpha_{0m} + \sum_k \alpha_{km}\chi_k, m = 1,\cdots,M$
- 활성화 함수: $z_m = \varphi_m(v_m)$, $m = 1,\cdots, M$. 신경망에서 활용되는 활성화 함수의 종류는 다양한데 그중 많이 활용되는 시그모이드 함수를 본 연구에서도 사용하였다. 시그모이드 활성화 함수는 $\varphi_m(v_m) = \sigma(v_m) = (1+\exp(-v_m))^{-1}$의 형태로 주어진다.
- 출력 변수 예측: $Pr(y = 1) = \sigma(\beta_0 + \beta_1 z_1 + \cdots + \beta_M z_M)$

위에서 은닉 노드의 수 M을 얼마로 하는가에 따라 추정해야 하는 모수인 신경망 가중치 $\{\alpha, \beta\}$들의 개수가 달라진다.

본 연구에서 사용된 입력 변수는 13개이지만, 앞에서 서술한 바와 같이 비율 척도가 아닌 명명 척도나 서열 척도로 측정된 변수들은 이진형 더미 변수들의 조합으로 재구성되기 때문에 본 연구에서 실질적으로 사용된 입력 변수들은 25개에 이른다. 은닉층에 있는 은닉 노드가 하나만 있으면(M=1), 바이어스라고도 불리는 절편을 포함하여 α는 26개이며 β는 2개에 달하여 총 28개의 신경망 가중치를 훈련 자료를 이용하여 추정하여야 한다. 따라서 은닉 노드의 수가 M인 경우에는 26*M개의 α들과 M+1개의 β들을 추정해야 하므로 총 27*M+1개의 신경망 가중치를 훈련 자료를 이용하여 추정하여야 한다. 신경망의 특수한 구조를 이용한 역전파(backward propagation) 알고리즘이 활용된다. 은닉 노드 수인 M을 결정하기 위해서, 먼저 M=1인 신경망 모형을 훈련 자료를 이용하여 적합시키고 이 적합된 결과를 평가 자료에 활용하여 해당 모형의 예측력을 평가한다. 그리고 M의 수를 하나씩 늘려가는데 예측력이 더 증가하지 않을 때까지 M의 수를 늘려가는 전략을 취하였다. 〈표 3〉에 나타난 바와 같이 은닉 노드의 수를 증가시킬수록 훈련 자료상 손실 함수의 값은 감소하지만, 은닉 노드의 수가 증가할수록 과적합 문제가 발생할 가능성이 있다. 평가 자료에서의 적중률과 F1 정

〈표 3〉 신경망 모형 적합 결과

은닉 노드의 수 (M)	손실 함수 (훈련 자료)	적중률 (평가 자료)	F1 정확도 (평가 자료)
1	32885.26	78.26%	86.77%
2	32611.45	78.36%	86.82%
3	32548.76	78.60%	86.98%
4	32410.58	78.43%	86.84%
5	32367.67	78.35%	86.80%
6	32302.21	78.19%	86.68%



〈그림 12〉 은닉 노드가 3인 신경망 모형에 의한 예측 결과

확도 측면에서 보면 은닉 노드의 수가 3일 때에 적중률과 F1 정확도 값이 최고조에 이르게 되며, 은닉 노드의 수가 3보다 크면 적중률과 F1 정확도가 낮아지는 것을 볼 수 있다. 따라서 본 연구에서는 은닉 노드의 수가 3인 모형을 신경망 모형의 최종 모형으로 확정한다.

〈그림 12〉에 은닉 노드가 3인 신경망 모형에 의한 자차 보험 가입 확률 예측값과 실제 자차 보험 가입 자료를 박스플롯한 결과를 제시하였다. 박스플롯에서는 모형의 예측력이 상당히 좋아 보이나, 마찬가지로 계량인 정확도 측정치 차원에서는 적중률이 78.60%이고 F1 정확도는 86.98%로서 맹목 예측을 소폭 상회하는 수준으로 나타났다.

5. 예측 결과 종합

전체 100,197개의 관측치를 70,053개의 훈련 자료로, 30,144개의 평가 자료로 나누어 훈련 자료를 이용하여 각 모형을 추정하고 추정된 모형을 평가 자료에 적용하여 각 모형의 예측력을 평가한 결과가 〈표 4〉에 나타나 있다.

자동차 보험 가입자 중에 자차 보험에 가입한 사람이 그렇지 않은 사람보다 많으므로 무조건 자차 보험에 가입할 것이라는 맹목 예측을 하면 적중률은 76.03%이고 F1 정확도는

〈표 4〉 예측 기법들의 예측력

예측 기법	적중률	F1 정확도
1. 맹목 예측	76.03%	86.38%
2. 선형 판별 분석	77.81%	86.37%
3. 로지스틱 회귀	76.56%	84.62%
4. CART(가지치기 이전)	78.09%	86.58%
5. CART(가지치기 완료 후)	77.43%	86.22%
6. C5.0	78.35%	87.00%
7. 신경망(은닉 노드 수 = 3)	78.60%	86.98%

86.38%에 달한다. 물론 이러한 맹목 예측은 마케팅 측면에 실질적인 도움이 전혀 되지 않는다. 고객 특성이나 고객이 소유한 차량의 특성들을 전혀 고려하지 않는 이러한 맹목 예측은 마케팅 의사 결정 측면에서 어떠한 지침도 주지 않기 때문이다. 다만 이러한 맹목 예측의 예측력를 비교 기준으로 이용하여 다른 모형들의 예측력을 평가해 볼 수는 있다.

〈표 4〉로부터 두 개의 요점을 추출하게 되는데, 먼저 가장 주목해야 할 점으로는, 본 연구에서 사용된 네 가지 종류의 모형들, 즉 판별 분석, 로지스틱 회귀, 의사결정나무 모형, 신경망 모형은 모두 예측력 측면에서 맹목 예측 대비 의미 있을 만큼 우수한 성능을 보이지 못하고 있다. 적중률은 모든 모형이 맹목 예측을 소폭 상회하였으나 F1 정확도의 경우에서는 판별 분석이 맹목 예측에도 미치지 못하였다. 둘째, 사용된 모형들 간에 상대적인 성능을 비교해 보면, 의사결정나무 모형과 신경망 모형이 판별 분석과 로지스틱 회귀보다 상대적으로 나은 것으로 나타나고 있다. 이러한 두 가지 결론은 서로 연관되어 있다고 판단된다.

먼저 연구에 사용된 자료에서 모형들의 예측력이 맹목 예측에 비해 우수하지 못하다는 점은 예측에 사용된 모형들의 일반적인 특성이라고 볼 수 없다. 저조한 예측 성능은 본 연구에 사용된 자료에 국한된 문제라고 볼 수 있다. 즉 본 연구에서는 자동차 보험 가입자의 자동차 보험 내역에서 추출된 고객 특성과 차량 특성들로 구성된 13개의 입력 변수들을 이용하여 자차 보험 가입 여부를 예측하였는데, 13개의 입력 변수가 자차 보험 가입 여부라는 출력 변수를 예측하는 데 유의미한 도움을 주지 못하고 있다는 것이다. 7만 건이 넘는 훈련 자료와 3만 건이 넘는 평가 자료를 활용하였고 입력 변수의 숫자 대비 자료의 크기가 충분히 크기 때문에 과적합 문제에서 예측력 저하가 발생하였다고 볼 수 없다. 이 경우에는 과대 적합(overfitting)보다는 과소 적합(underfitting)이 오히려 문제인 경우라 할 수 있다.

한편 모형들 간의 상대적 예측 성능을 비교해 보면, 의사결정나무 모형과 신경망 모형이

판별 분석과 로지스틱 분석보다 본 자료에서 더 우수한 것으로 나타났다. 마찬가지로 입력 변수의 숫자 대비 자료의 크기가 매우 큰 본 연구의 특성을 고려하면 판별 분석이나 로지스틱 등의 선형 모형들은 자료를 분석하는 데 유연성이 부족한 것으로 판단된다. 반면 의사결정나무나 신경망 모형처럼 자연스럽게 비선형 관계와 입력 변수 간 상호 작용을 포함하는 모형이 상대적으로 우수하다는 것은 본 연구에서 활용된 자료에 있어서는 유연성이 높은 모형이 더 필요하다는 것의 반증이다. 즉 자료 내 관측치의 수가 상당하지만 입력 변수의 수가 상대적으로 매우 작은 본 자료와 같은 경우에는 과대 적합보다는 과소 적합이 예측력 저하의 원인이 될 가능성이 높으므로, 비선형성과 상호 작용을 자연스럽게 포함하는 유연성이 높은 모형이 훨씬 더 유용하다고 할 수 있다.

IV. 결론

본 연구에서는 판별 분석, 로지스틱 회귀, 의사결정나무 모형, 그리고 신경망 모형 등의 데이터 마이닝 기법을 활용하여 자동차 보험 가입자들의 자차 보험 가입 여부를 예측하는 문제를 풀어보았다. 10만 건이 넘는 보험 가입 자료를 활용하여 모형을 추정하고 예측하였는데, 사용된 모형의 예측력이 맹목적인 예측보다 크게 우수하지 않는 것으로 나타났다. 이러한 낮은 예측력의 이면에는 활용된 자료 내의 입력 변수들이 출력 변수를 예측하는 데 있어 유용성이 높지 않을 가능성이 존재한다. 입력 변수가 13개이고 훈련 자료 내의 관측치가 7만 건 이상인 점을 고려하면 이러한 가능성에 무게가 실린다. 즉 보험 가입 내역에 나타난 고객 특성과 차량 특성들을 가지고 자차 보험 가입 여부를 추정하고 예측하는 데 한계가 존재한다는 것이 명확하다. 즉 고객들의 자차 보험 가입 여부는 관측되지 않은 다른 변수들에 의해 결정되는 부분이 매우 크기 때문에 자동차 보험 가입 내역에 관측된 변수들만 가지고 자차 보험 가입 여부를 예측하는 데에는 한계가 존재한다.

따라서 고객 행동에 대해 마케터가 더 잘 이해하고 예측할 수 있기 위해서는 자동차 보험 가입 내역 외에 추가적인 고객 데이터가 필요하다. 고객의 경제 및 인구통계적 계량 자료 외에도 소셜 미디어에 나타난 고객 의견이나 태도, 고객 접점 수준에서 고객 접촉 내용 데이터 등 비정형 데이터 등을 활용하여 고객 행동 예측의 품질을 높여야 한다. 이러한 문제는 자동차 보험 회사뿐만 아니라 다양한 기업들이 현재 가지고 있는 문제라고 할 수 있다. 많은 기업들이 다년간 누적되어 온 고객 수준의 거래 자료를 활용하여 미래의 고객 행동을 이해하

고 예측하려고 시도하지만 기업들이 가지고 있는 자료의 내용에서 한계가 존재하는 경우가 상당하다. 오늘날 많은 기업들이 가지고 있는 데이터의 문제는 관측치의 수의 문제가 아니라 오히려 관측 변수의 다양성과 이들의 품질 문제이다. 따라서 기존에 자료로서 고려하지 않았거나 계량적 코딩이 곤란해서 무시하였던 고객 관련 내용들을 포함하여 다양한 고객 행동 정보를 적극적으로 활용하여야 한다. 이러한 문제를 극복하기 위해서는 기존에 눈여겨 보지 않았던 비정형 자료 등을 포함하여 이종의 고객 자료들을 융합(fusion)하는 데이터 융합 접근 방법이 필요하다.

참고문헌

Blattberg, R. C., R. Glazer, and J. D. C. Little (1994), *The Marketing Information Revolution*, Harvard Business School Press.

Blattberg, R. C., B. Kim, and S. A. Neslin (2008), *Database Marketing*, Springer.

Gentzkow, M. (2007), "Valuing New Goods in a Model with Complementarity: Online Newspapers," *American Economic Review*, 97(3), 713–744.

Hanemann, M. (1984), "Discrete/Continuous Models of Consumer Demand," *Econometrica*, 52(3), 541–561.

Kim, J., G. Allenby, and P. Rossi (2002), "Modeling Consumer Demand for Variety," *Marketing Science*, 21(3), 229–250.

Lee, S., J. Kim, and G. M. Allenby (2013), "A Direct Utility Model for Asymmetric Complements," *Marketing Science*, 32(3), 454–470.

Song, I. and P. K. Chintagunta (2006), "Measuring Cross–Category Price Effects with Aggregate Store Data," *Management Science*, 52(10), 1594–1609.

Song, I. and P. K. Chintagunta (2007), "A Discrete: Continuous Model for Multicategory Purchase Behavior of Households," *Journal of Marketing Research*, 44(4), 595–612.

Predictive Modeling of Customers' Insurance Purchase Behaviors

Inseong Song*

This study investigates the approaches for predictive modeling of customers' purchase for optional collision and comprehensive (CNC hereafter) coverage when buying automobile insurances. The training data set includes 13 input variables with more than 70,000 observations along with the output variable of whether a customer purchases CNC. Linear discriminant, logistic regression, decision tree, and neural networks are utilized but their prediction performance turns out low. Since the size of data is sufficiently large, the low predictive power seems to stem from underfitting rather than overfitting. That is, the input variables are not useful to predict the output variables. The input variables are consumer and automobile characteristics retrieved from car insurance transaction records. The prediction results indicate such type of information has little power in predicting customer purchase choice for CNC. So it is obvious that marketers should collect more information from other sources and include those additional variables into the model to improve the predictive accuracy.

Many firms have accumulated customer transaction records and now want to utilize such data in optimizing their marketing programs. But it occasionally occurs that such transaction records alone are not sufficient to predict customer behaviors of interest. In order to overcome such problems, firms need to collect additional data from various sources including nonstructured sources and fuse them through an integrated data fusion framework.

*Professor, College of Business Administration, Seoul National University

Keywords: predicting customer purchase, discriminant analysis, logistic regression, decision tree, neural networks