석사학위논문

텔레마케팅 성과 개선을 위한 다목적 하이브리드 추천 시스템

2019년

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

박 의 범

HANSUNG
UNIVERSITY

석 사 학 위 논 문
지도교수 김형수

텔레마케팅 성과 개선을 위한 다목적 하이브리드 추천 시스템

Multi-purpose hybrid recommendation system
to improve telemarketing performance

2018년 12월 일

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

박 의 범

석 사 학 위 논 문
지도교수 김형수

텔레마케팅 성과 개선을 위한 다목적 하이브리드 추천 시스템

Multi-purpose hybrid recommendation system to improve telemarketing performance

위 논문을 산업경영공학 석사학위 논문으로 제출함

2018년 12월 일

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

박 의 범

박의범의 공학 석사학위논문을 인준함

2018년 12월 일

심사위원장 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

국 문 초 록

텔레마케팅 성과 개선을 위한 다목적 하이브리드 추천 시스템

한 성 대 학 교 일 반 대 학 원
산 업 경 영 공 학 과
산 업 경 영 공 학 전 공
박 의 범

협업 필터링 등 다양한 추천 알고리즘은 고객에게 상품을 추천하는 유용한 방법론이며, 텔레마케팅 산업에서도 추천 시스템을 활용하여 비즈니스 성과를 향상시킬 수 있다. 그러나 텔레마케팅의 특성상 고객과의 접촉이 실패하면 추천 시스템의 정확도가 아무리 높다고 하더라도 상품을 추천할 수 없다는 문제가 존재한다. 본 연구에서는 한정된 자원 내에서 최대한의 비즈니스 성과를 달성하기 위해 접촉 가능성이 높은 고객을 우선적으로 선별하고, 선별된 고객에게 상품을 추천하는 일련의 프로세스를 제시하였다. 또한 실험을 통해 본 연구에서 제안한 다목적 하이브리드 추천 알고리즘의 성과가 기존과 같이 랜덤하게 고객을 선별하여 상품을 추천하는 성과보다 더 우수함을 입증하였다.

Keywords: 추천 시스템, 머신러닝, 텔레마케팅, 대상고객 선별, 그래디언트 부스팅, 랜덤 포레스트

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1.1 연구의 배경

과거의 추천 시스템은 기존의 구매 트랜잭션이나 상품에 대한 선호도로 향후의 구매 가능성이 높은 항목을 예측하는 협업 필터링(Collaborative Filtering, CF)과 같은 방식이 주를 이루었다(Ekstrand, Riedl & Konstan, 2010). 이와 같은 알고리즘은 사용자에게 신뢰할만한 추천을 제시하기 위해 충분한 선호도 정보가 제공되어야 한다. 하지만 시스템에 새롭게 추가된 사용자나 항목은 충분한 정보를 갖지 못하므로 일반적으로 만족할만한 추천 항목을 얻을 수 없는 콜드 스타트나 희소성, 확장성 등의 문제가 발생한다(Hu & Pu, 2010, Reshma, Ambikesh & Thilagam, 2016). 최근 연구에서는 위와 같은 기존 추천 시스템의 문제점을 개선하기 위한 하이브리드 모형의 연구가 제안되어 왔으며, 영화(Christakou, Vrettos & Stafylopatis, 2007), e-러닝(Khribi, Jemni, & Nasraoui, 2011, Kardan, Abbaspour & Hendijanifard, 2009), 음악(Wang & Wang, 2014), 웹 서비스(Burke, 2007) 등 다양한 업종에 적용되고 있다.

텔레마케팅은 마케팅 캠페인을 위한 목적으로 가장 널리 사용되는 서비스 중 하나이다(Moro, Cortez & Rita, 2014). 따라서 기업 관점에서는 비즈니스 성과 향상을 위해 텔레마케팅을 효과적으로 운영할 필요가 있으며, 추천 시스템을 도입하는 것은 이를 위한 중요한 역량 중 하나가 될 수 있다.

또한 텔레마케팅은 고객 관점에서 기업이나 추천받는 상품에 만족하는 경우에는 유의미한 마케팅 수단이 될 수 있다. 그러나 접촉을 원하지 않은 고객에게 무분별하게 텔레마케팅을 시도하는 것은 고객의 피로도를 증가하여 기업에 대한 불만을 야기하여 역효과를 발생시키는 요인이 되기도

한다.

이처럼 텔레마케팅은 기업이 마케팅을 진행할 수 있는 효과적인 수단 중 하나이지만, 모든 고객이 원하지 않는다는 문제가 발생한다. 이러한 이유로 텔레마케팅에서는 추천 시스템을 활용하기에 앞서 고객이 텔레마케팅의 대상으로 적합한지를 먼저 고려하여야 할 필요가 있다. 여기서 텔레마케팅에 적합한 대상은 접촉에 긍정적인 반응을 보이는 고객을 말하며, 이러한 고객을 대상으로 텔레마케팅을 시도할 때 더 높은 반응률과 추천 성공률을 기대할 수 있다. 이를 위해 본 연구는 텔레마케팅의 성과 개선을 위해 접촉 가능성이 높은 고객을 선별하고 선별된 고객에게 상품을 추천하는 일련의 프로세스로, 상기 두 가지 목적을 동시에 달성하기 위한 다목적 하이브리드 추천 알고리즘을 제시하였다.

본 연구에서는 국내의 보험 상품을 판매하는 텔레마케팅 대행사의 실제 텔레마케팅 및 청약 데이터를 활용하였으며, 실험을 바탕으로 제안된 추천 시스템의 현실적인 적용 가능성을 검증하였다. 다목적 하이브리드 추천 시스템의 첫 번째 단계는 고객의 접촉 가능성을 예측하는 모형으로 여섯 가지의 머신러닝 알고리즘의 성능을 비교하여 최종 알고리즘을 결정하였다. 이때 접촉 가망 추출 모형은 텔레마케팅 대상으로 선별된 고객 중 긍정적인 반응을 보인 비율을 의미하는 정밀도로 모형의 성능을 평가하였다. 다음으로 보험 상품 추천 모형은 추천 성공률을 기준으로 모형의 성능을 평가할 수 있다. 그러나 본 연구와 같이 보험 등의 금융 업종에서 대면 상담을 진행할 때에는 첫 번째 추천 상품이 긍정적인 반응을 보이지 않는 등, 복수 개의 상품을 동시에 추천하여야 하는 경우가 존재한다(Lacerda, 2017). 일반적으로 한 고객에게 여러 상품을 추천하는 것은 마케팅 비용을 상승시킬 수 있는 문제이지만, 대면 상담의 성격상 하나의 상품만을 추천할 때와 여러 상품을 함께 추천할 때 소모되는 비용의 차이가 크지 않다는 점에서 후순위 상품을 연이어 추천할 수 있다. 그러나 지나치게 많은 수의 상품을 동시에 추천하는 것은 오히려 고객에게 반감을 일으킬 수 있어, 적정 수준으로 추천 상품의 수를 제한할 필요가 있다. 따라서 본 연구에서는 3순위의 상품까지만 추천할 수 있도록 허용하였고, 이러한 이유로

보험 상품 추천 모형의 성능을 평가할 때 1순위만 고려하는 것이 아니라 3순위 이내에서 청약 상품을 적중한 경우까지 포함하여 추천 정확도를 계산하였다.

Ⅱ. 문헌 연구

2.1 머신러닝 알고리즘

머신러닝(Machine Learning)은 데이터로부터 지식을 습득하는 과정을 컴퓨터를 통해 자동화하는 방법론으로(Langley & Simon, 1995), 데이터를 분석하기 위해 필수적으로 활용되는 도구이다(최영찬 & 이민수, 2010). 또한 디지털 혁명 또는 4차 산업혁명이 시작되면서 다양한 분야에서 데이터 분석이 요구되고 있고, 이러한 이유로 많은 기업이나 논문에서도 머신러닝 기반의 데이터 분석이 활발하게 연구되고 있다.

일반적으로 머신러닝은 알고리즘의 성격에 따라 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 분류된다. 지도 학습 모형은 종속변수, 즉 정답에 해당하는 데이터를 학습시키는 방법론으로, 대표적으로 분류(Classification)와 회귀(Regression)를 위한 알고리즘으로 구성된다. 여기서 분류는 이분형 변수 또는 다항의 범주형 종속변수를 예측하는 모형을 말하고, 회귀는 연속형 종속변수를 예측하는 모형을 의미한다. 반면 비지도 학습 모형은 종속변수가 존재하지 않은 방법론을 말하며, 대표적으로 군집화 알고리즘이 있다.

본 연구와 같이 고객의 과거 접촉 성공 여부를 활용하여 미래의 접촉 성공 여부를 예측하는 것은 지도 학습 모형 중 분류에 해당한다. 지도 학습의 분류를 위해 사용되는 알고리즘은 대표적으로 의사결정나무(Decision Tree, DT), 인공 신경망(Neural Network, NN), 서포트 벡터 머신(Support Vector Machine, SVM), 로지스틱 회귀(Logistic Regression, LR), 랜덤 포레스트(Random Forest, RF), 그래디언트 부스팅(Gradient Boosting, GB) 등이 존재한다.

2.1.1 의사결정나무

의사결정나무는 가장 널리 알려진 머신러닝 알고리즘의 하나로, 인공지능 분야 등 다양한 분야에서 분류 예측을 위해 활용된다(Elouedi, Mellouli & Smets, 2000). 이 알고리즘은 의사결정규칙이 나무와 유사한 구조로 표현되며, 마디(Node), 가지(Branch), 깊이(Depth)로 구성된다. 여기서 마디는 뿌리(Root Node)마디와 끝마디(Terminal Node) 그리고 부모마디(Parent Node)와 자식마디(Child Node)로 구분된다. 뿌리마디는 의사결정나무가 시작되는 마디이며, 끝마디는 각 가지의 끝에 위치한 마디를 말한다. 또한 자식마디는 하나의 마디로부터 분리된 2개 이상의 마디를 말하고, 이때 자식 마디로 분리되는 상위의 마디를 부모마디라고 부른다. 즉 의사결정나무는 하나의 뿌리마디에서 2개 이상의 자식마디로 분리되고, 분리된 마디는 부모마디가 되어 다시 자식마디로 분리되는 과정을 반복한다. 여기서 자식마디는 부모마디보다 불순도를 낮추는 방향으로 분리되며, 불순도는 지니 계수, 엔트로피 등을 통해 계산된다. 이러한 반복이 종료되어 더 이상 자식마디로 분리되지 않는 경우, 각 가지의 끝에 위치한 마디가 끝마디가 된다. 이와 같은 과정을 통해 의사결정나무가 학습되고, 학습된 의사결정규칙은 나무 형태로 표현되어 구조와 분리 기준을 이해하고 설명하기 쉬우며, 교호작용의 효과를 해석하기에도 용이하다. 이러한 장점으로 인해 의사결정나무는 분류 예측 목적 뿐 아니라 종속 변수를 예측함에 있어서 영향을 주는 요인을 파악하기 위한 용도로도 활용된다.

이처럼 의사결정나무를 학습하는 과정에서 모형의 성능을 향상시키기 위해 정지규칙과 가지치기를 선행하여야 한다. 정지규칙은 모형의 과적합을 방지하기 위해 일정 수준에서 더 이상 마디가 분리되지 않도록 깊이를 제한하거나, 끝마디의 최소 레코드 수를 결정하는 것을 말하고, 가지치기는 오분류가 발생할 위험이 큰 가지를 제거하는 것을 의미한다.

의사결정나무를 기반으로 이분형 종속변수를 분류하였던 연구의 예는 다음과 같다. 신경식과 안수산의 연구(2002)에서는 스팸 메일에 많이 포함될 것으로 판단되는 단어의 포함여부를 통하여 스팸 메일 여부를 예측하였다. 이때 의사결정나무와 인공신경망의 성능을 비교한 결과, 의사결정나무가 더 우수한

예측력을 보였다. 또한 채경희와 김상철(2010)은 의사결정나무를 통해 고객이 1년 후 우수고객이 될 것인가의 여부를 예측하였고, 규칙 기반 시스템과 비교하여 더 많은 타겟을 선정함과 동시에 높은 정분류율을 보이는 효과를 얻었다. 김석중 등(2000)은 일정금액 이상으로 우유를 구매하는 가구와 구매하지 않는 가구를 예측하였고, 의사결정규칙을 활용하여 우유 소비에 큰 영향을 미치는 요인을 확인하였다. 다음으로 최종후 등의 연구(2006)에서는 보다 정확한 선거예측조사 결과를 얻기 위해 자신의 지지후보를 밝히지 않는 의사 결정 유보층이 어떤 후보를 지지하는지 통계적으로 예측하는 방법론을 제안하였다. 한편 이혜주와 정의현(2013)은 중학생의 학업 성취에 영향을 미치는 요인들을 탐색하기 위한 방법으로 의사결정나무를 활용하였고, 임은정과 정순희(2015)는 중고령자의 은퇴를 결정하는 요인을 탐색하였다. 이를 위해 정의한 은퇴의사결정은 완전 은퇴, 점진적 은퇴, 은퇴하지 않음의 세 범주로 분류하였고, 모형을 개발하기 위한 독립변수로 인구 통계학적 변인, 제도 및 사회적 요인, 심리 및 기타 요인을 활용하였다.

2.1.2 인공신경망

인공신경망은 인간의 신경세포 뉴런(Neuron)과 유사한 구조를 가진 학습 알고리즘으로, 1980년대에 컴퓨터의 처리 속도가 인공신경망의 요구를 충족시키기 시작하면서(Cochocki & Unbehauen, 1993) 활용되기 시작하였다. 이 알고리즘은 대표적으로 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)으로 구성된 다층 퍼셉트론(Multi-Layer Perceptron)의 구조를 가진다. 여기서 입력층은 입력변수가 1:1로 대응되는 노드를 의미한다. 따라서 입력층의 개수는 입력변수의 수와 동일하며, 출력층은 종속변수의 데이터 타입에 따라 노드의 개수가 결정된다. 마지막으로 은닉층은 입력층과 출력층 사이에 존재하는 모든 노드를 말한다. 이와 같은 요소로 구성된 인공신경망은 잡음이 많은 데이터에서도 안정적인 결과를 산출할 수 있고, 입력변수와 결과변수의 관계가 복잡할 때에도 우수한 예측성과를 보일 수 있는 장점이 있다. 실제 비즈니스에서는 환경의 변화로 인한 데이터에 잡음이 발생하고, 고려하여야 할 변수들이 너무 많으며, 그 중 정성적인 변수들이 포함된다

는 어려움이 존재한다(조홍규, 2003). 그러나 인공신경망은 이러한 상황에서도 적용이 가능하다는 점에서 다양한 분야의 예측을 위한 알고리즘으로 활용되고 있다.

또한 인공신경망은 종속변수의 데이터 형태에 따라 연속형 변수의 예측과 범주형 변수의 분류가 모두 가능하며, 본 연구와 마찬가지로 데이터를 분류하기 위한 목적으로 인공신경망을 활용한 연구는 다음과 같다. 유은정 등의 연구(2013)에서는 인공신경망을 활용하여 신체적 행동을 기반으로 고객의 감정을 분류하는 모형을 제안하였고, 감정에 따라 표출되는 행동의 크기와 지속시간의 차이를 고려하여 각 감정을 예측할 수 있는 최적의 프레임 수를 결정하였다. 또한 정윤과 황석해(1999)는 재무적 지표와 비재무적 지표를 활용하여 부실기업 여부를 예측하였고, 로지스틱 회귀분석과 인공신경망의 성능을 비교하여 인공신경망 모형의 성능이 우수함을 확인하였다. 다음으로 서상완 등의 연구(2013)에서는 보청기를 사용하는 환경에 따라 음성 신호를 다르게 처리하기 위한 예측 모형을 제안하였다. 보청기를 사용하는 환경은 음성만 있는 상황, 음성과 잡음이 함께 있는 상황, 잡음만 있는 상황, 음악을 듣는 상황으로 제시되었으며, 입력되는 소리의 특징을 추출하여 제시된 네 가지의 음성 인식 환경을 예측하였다. 마지막으로 손호선 등(2018)은 위암의 조기 진단을 위하여 위암을 유발하는 유전자를 찾고, 이를 활용하여 위암의 발병 여부와 생존여부를 예측하는 모형을 제안하였다.

2.1.3 서포트 벡터 머신

서포트 벡터 머신은 머신러닝 알고리즘 중 강력한 분류 성능을 보이는 방법론 중 하나로(Cortes & Vapnik, 1995), 인공신경망에서 발생할 수 있는 부분 최적화 문제를 해결할 수 있어, 세계적인 최적화 알고리즘으로 자리를 잡게 되었다(Richhariya & Tanveer, 2018). 이 알고리즘은 주로 패턴 인식의 목적으로 활용되는(Burges, 1998) 지도 학습 모형으로, 주어진 데이터의 범주 간 여백(Margin)을 최대화하면서 오분류율을 낮추는 초평면(Hyperplane)을 찾는 과정을 통해 모형이 학습되며, 이때 최종적으로 결정된 초평면을 기준으로 범주를 분류한다. 여기서 여백은 초평면과 서포트 벡터 사이의 간격을 말

하며, 서포트 벡터는 초평면에서 가장 근접한 곳에 위치한 데이터를 의미한다. 또한 서포트 벡터 머신이 발전함에 따라 비선형 분류 문제를 해결하기 위해 커널함수를 활용하는 방법론이 등장하였으며, 대표적인 커널함수로는 다항식, 방사 기저 함수, 쌍곡 탄젠트 등을 활용한다.

본 연구와 같이 분류 예측을 위한 목적으로 서포트 벡터 머신을 활용한 연구의 예는 다음과 같다. 강남이 등의 연구(2012)에서는 자연자원이나 환경관리를 위해 인공위성의 영상을 활용하여 논, 하천, 임목지, 주거지 등 8가지의 토지피복을 분류하는 모형을 제안하였고, 분석결과 최대우도 분류법과 인공신경망에 비해 더 우수한 결과를 도출하였다. 또한 박선미 등(2008)은 영상기반의 감시 시스템을 위한 의상 특징 분석에 대한 연구를 진행하였다. 이 연구는 의상 영역에 대한 특징 벡터를 구성한 뒤 서포트 벡터 머신을 활용하여 양복과 비양복을 성공적으로 분류하였다. 다음으로 김상균과 장준혁(2008)은 기존에 존재하던 SMV코덱의 음성/음악 분류 성능을 향상시키기 위해 서포트 벡터 머신을 활용하였다.

이처럼 서포트 벡터 머신은 주로 영상인식, 음성인식 등 다양한 패턴인식 분야에서 가장 활발히 연구되고 발전되어 왔다. 그러나 이는 패턴인식 외의 분야에서 서포트 벡터 머신의 예측 성능이 낮기 때문이 아니며, 경영 분야 등 다양한 분야에서도 함께 연구되고 있다. 그 예로 최하나와 임동훈(2013)은 서포트 벡터 머신을 기반으로 재무적인 자료를 활용하여 기업의 부도를 예측하였다. 이 연구에서는 예측 성능을 특정 커널함수에 의존하는 것을 피하기 위해 다양한 커널함수를 통해 생성된 모형들을 과반수 투표 방식으로 결과를 선정하는 앙상블 모형을 제안하였다.

2.1.4 로지스틱 회귀

로지스틱 회귀 모형은 회귀분석 중 하나로, 1970년대 이후 통계학 및 데이터마이닝을 중심으로 많은 연구가 진행되어 왔으며, 비즈니스 및 금융, 범죄학, 공학, 의료 등 다양한 분야에서 널리 사용되는 알고리즘이다(Menard, 2002). 회귀분석 중 가장 많이 알려진 선형 회귀분석은 연속형 변수를 예측하는 반면 로지스틱 회귀분석은 주로 이분형 변수를 분류하는 목적으로 사용

된다. 일반적으로 선형 회귀모형은 다음과 같이 정의된다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

이러한 회귀식에서 종속변수는 독립변수의 값에 따라 [−∞,∞] 사이의 값을 가질 수 있다. 하지만 로지스틱 회귀모형은 사용 목적상 독립변수의 값에 관계없이 항상 [0,1] 사이의 값으로 예측되어야 하는 문제가 발생한다. 이러한 목적을 달성하기 위해 다음과 같은 로지스틱 함수를 사용하는 방법론이 제안되었다.

$$Logistic\,function = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}}$$

앞서 설명한 것처럼 일반적인 로지스틱 회귀분석은 이분형 변수를 예측하는 용도로 사용된다. 예를 들어 김수진과 김보영(2013)은 개인의 인구 통계적 특성과 서베이를 통한 개인에 대한 우울 여부를 예측하고 그 요인을 찾는 연구를 진행하였고, 로지스틱 회귀 모형과 의사결정나무의 의사결정규칙을 함께 활용하여 사용된 독립변수 중 우울여부에 유의한 영향을 주는 변수를 찾아내었다. 다음으로 김세형 등(2008)은 한국프로농구에서 공개한 경기별 공격과 수비에 관련된 14개의 경기기록을 활용하여 승패를 예측하였다. 이 연구에서는 전진선택법을 통해 설명력이 높은 변수를 선택하였고, Wald값을 기준으로 각 변수가 경기 승패에 미치는 영향력의 우선 순위도 함께 제시하였다.

반면 로지스틱 회귀를 활용하여 범주가 셋 이상인 다항형 종속변수도 예측할 수 있다. 이를 다항 로지스틱 회귀분석이라고 부르며, 다음과 같은 연구들을 예로 들 수 있다. 이영섭과 김희경의 연구(2007)에서는 설문을 통해 응답받은 땀, 수면, 소화 등의 소증 항목을 활용하여 태양인, 태음인 등 사상체질을 예측하는 모형을 제안하였다. 이때 모형의 정확성을 향상시키기 위해 로지스틱 회귀모형을 활용한 각 소증 항목별 변수 선택 과정을 거쳤으며, 그

결과 로지스틱 회귀모형이 의사결정나무나 인공신경망에 비해 더 높은 정확도를 보였다. 마지막으로 손소영과 신형원(1998)은 사망, 중상, 경상, 부상신고, 물적피해라는 다섯 가지의 사고 심각도를 예측하였다. 그러나 일부 범주에 속하는 사고의 수가 극히 낮아 치명적 상해, 경미한 상해, 물적 피해로 묶은 것과 신체 상해, 물적 피해로 묶은 두 가지 경우에 대해 각각의 예측 모형을 생성하였다. 이 연구에서는 의사결정나무를 통해 독립변수를 선정하였으며, 모형 성능 비교 결과 세 가지의 항목으로 구분하였을 때에는 의사결정나무, 두 가지의 항목으로 구분하였을 때에는 로지스틱 회귀모형의 정확도가 우수한 것으로 검증되었다.

2.1.5 랜덤 포레스트

앞서 설명한 의사결정나무는 해석과 사용이 간단하다는 장점이 존재하였다. 그러나 상대적으로 인공신경망이나 서포트 벡터 머신 등에 비하여 성능이 떨어지며, 학습 데이터의 작은 변경으로도 의사결정규칙이 크게 달라질 수 있는 불안정성도 함께 존재한다(Li & Belford, 2002). 랜덤 포레스트는 의사결정나무의 이러한 한계를 극복하기 위하여 제안된 알고리즘으로, 여러 개의 의사결정나무를 통해 도출된 각각의 결과에서 투표(Voting) 방식으로 최종 결과를 결정하는 방법론이다. 이때 생성되는 각각의 의사결정나무는 임의로 샘플링된 데이터와 변수들을 사용하며, 그 결과 각각의 의사결정나무는 서로 다른 예측 값을 도출한다. 따라서 이를 앙상블한 랜덤 포레스트는 의사결정나무에 비해 과적합이나 데이터 변동으로 인한 모형의 변화 등의 측면에서 안정성을 가지며, 일반화되는 효과를 얻을 수 있다. 또한 매개변수를 결정하는 것이 다른 머신러닝 알고리즘에 비해 간단하다는 장점이 존재한다. 랜덤 포레스트를 활용하여 모형을 생성할 때 고려하여야 하는 대표적인 매개변수로 숲의 크기와 최대 허용 깊이가 있다. 여기서 숲의 크기는 모형을 결합하는데 사용되는 의사결정나무의 수를 말하며, 최대 허용 깊이는 각 의사결정나무의 최대 깊이를 말한다. 이와 같은 장점으로 인해 랜덤 포레스트는 분류 및 예측을 위한 용도로 활발히 활용되고 있다. 또한 Breiman(2001)에 의하면 랜덤 포레스트를 활용하여 예측 모형에 투입되는 변수들의 중요도를 추출하는 용도로도 사

용할 수 있다.

랜덤 포레스트를 분류 예측 목적으로 사용한 예는 다음과 같다. 김태진 등의 연구(2018)에서는 모기업의 거래처가 가치사슬에 속하는가를 분류하기 위한 모형을 제안하였다. 이를 위해 전자세금계산서를 통해 발생하는 값들을 독립변수로 활용하였고, 전문가를 통해 거래가 발생한 기업의 가치사슬 포함 여부를 판단하였다. 또한 MDG(Mean Decrease Gini)와 MDA(Mean Decrease Accuracy)를 활용하여 모형에 투입된 독립변수별 중요도를 함께 파악하였다. 다음으로 악성코드를 통해 PC의 기능을 제한하거나 일부 중요 파일을 암호화하여 이용할 수 없게 만드는 랜섬웨어를 탐지하기 위한 연구에서도 랜덤 포레스트 알고리즘이 활용되었다(이승환 & 황진수, 2018). 이 연구에서는 패킷 시그니쳐 등 랜섬웨어 프로그램에 내장된 정보를 이용하는 기존의 정적인 분석이 아닌 프로그램이 실제 실행되는 과정에서 발생되는 동적 정보를 분석하여 작동을 저지하는 동적 분석을 활용하였다. 마지막으로 김성진과 안현철(2014)은 기업채권등급을 평가하는 모형을 제안하였다. 이 연구에서 활용된 기업의 신용등급은 5가지이지만, 데이터 수의 불균형으로 인해 총 4가지로 분류된 신용등급을 예측하였으며, 데이터의 희소성에 의한 과적합을 막기 위해 5-fold cross validation을 수행하였다. 그 결과 함께 비교한 다변량 판별분석, 인공신경망, 서포트 벡터 머신에 비하여 랜덤 포레스트가 더 높은 예측결과를 산출하는 것을 확인하였다.

2.1.6 그래디언트 부스팅

그래디언트 부스팅은 여러 개의 약한 모델(Weak Learner)을 결합하여 정확도가 향상된 하나의 강력한 모델(Strong Learner)을 생성하는 알고리즘이다. 이는 랜덤 포레스트와 마찬가지로 여러 개의 의사결정나무를 묶어 하나의 모형으로 구성하는 앙상블 방법이다. 그러나 랜덤 포레스트는 임의성에 의해 생성된 각각의 의사결정나무 결과에서 투표를 통해 최종 결과를 도출하는 반면, 그래디언트 부스팅은 앞 단계에서 생성된 모형의 손실 함수(평균 제곱 오차)를 최소화할 수 있는 새로운 약한 모델을 반복적으로 추가하여 점점 더 모형 정확도를 개선해가는 방식으로 각각의 모형이 결합된다(Touzani,

Granderson & Fernandes, 2018).

그래디언트 부스팅을 분류 예측 목적으로 활용한 예는 다음과 같다. 김희종과 김형도의 연구(2014)에서는 대출자 또는 대출희망자가 채무를 이행하지 않을 가능성을 예측하는 채무 불이행 예측 모형을 제안하였고, 다양한 머신러닝 알고리즘의 성능을 비교한 결과 그래디언트 부스팅이 의사결정나무, 선형회귀, 선형 판별분석, 서포트 벡터 머신에 비해 더 우수한 성능을 보였으며, 랜덤 포레스트보다도 약간 나은 결과를 보였지만, 통계적으로는 예측 성능의 차이가 없다는 결론을 내렸다. 다음으로 한은정과 김동건(2015)은 노인의 장기요양보험의 상담이 필요한 대상을 선정하기 위해 그래디언트 부스팅 알고리즘을 활용하였다. 이 연구는 상담 대상이 가족인 경우와 부양자가 존재하지 않은 독거노인을 위한 경우로 나누어 모형을 개발하였으며, 두 모형에서 AUC를 기준으로 평가하였을 때 그래디언트 부스팅이 가장 성능이 우수한 모형인 것으로 확인되었다.

2.2 추천 시스템

인터넷과 관련 기술의 발전으로 전자상거래의 확장을 위한 기반이 되었으며, 모든 전자상거래 웹사이트는 암시적 또는 명시적으로 사용자의 관심 항목을 발견하고 추천하는 자체 추천 시스템을 보유하고 있다(Tewari & Barman, 2018). 이처럼 추천 시스템은 온라인 환경에서 비즈니스 성과를 위해 필수적인 역량으로 자리잡았다. 이를 위한 대표적인 알고리즘으로 협업 필터링과 장바구니 분석을 들 수 있으나, 최근에는 상품을 추천하기 위한 방법론이 발전함에 따라 여러 기법들을 응용하거나 추가적인 컨텍스트 정보를 활용하는 하이브리드 기반 접근법이 추천 시스템을 위한 알고리즘으로 가장 활발히 연구되고 있다.

2.2.1 협업 필터링 기반 접근법

협업 필터링은 가장 알려져 있는 추천 시스템 기술의 하나이며, 다른 사용자와의 유사도를 기반으로 사용자가 관심을 가질만한 항목을 추천하

는 알고리즘이다(Hu & Pu, 2010, Bobadilla, Ortega, Hernando et al., 2013). 일반적으로 협업 필터링은 다음과 같이 분류된다(Breese, Heckerman & Kadie, 1998).

1. 메모리 기반 협업 필터링: 메모리 기반 협업 필터링은 사용자-아이템의 등급 매트릭스를 기반으로 하며(Ghazarian & Nematbakhsh, 2015), 대표적으로 사용자 기반 협업 필터링(User-based CF)과 아이템 기반 협업 필터링(Item-based CF)로 구분된다. 사용자 기반 협업 필터링은 각 사용자가 아이템별 선호도를 선택한 뒤, 평점 매트릭스를 기준으로 각 사용자 사이의 유사도를 측정한다. 이때 유사도가 높은 사용자들이 가장 선호하는 아이템을 추천한다. 반면 아이템 기반 협업 필터링은 각 아이템별 선호도를 조사한 뒤, 아이템별 유사도를 측정한다. 이때 가장 유사한 아이템을 추천 항목으로 도출한다. 이처럼 사용자 기반 협업 필터링과 아이템 기반 협업 필터링은 모두 유사도를 기반으로 추천 항목이 계산되며, 유사도를 계산하기 위해 대표적으로 코사인 유사도, 피어슨 상관관계, 유클리드 거리 등을 활용한다(Sarwar, Karypis, Konstan et al., 2001, Arsan, Koksal & Bozkus, 2016).
2. 모델 기반 협업 필터링: 모델 기반 협업 필터링은 모델을 기반으로 사용자의 선호도를 분석하는 접근법이다(Langseth & Nielsen, 2015). 이를 위한 대표적인 모델로 베이지안 네트워크(Bayesian Network), 클러스터링 모델(Clustering Model), 잠재 변수 모델(Latent Variable Model) 등이 활용된다(Breese, Heckerman & Kadie, 1998, Langseth & Nielsen, 2015, Mobasher, Burke & Sandvig, 2006).

서론에서 언급한 것처럼 추천 시스템에서의 협업 필터링은 콜드 스타트나 희소성 등의 문제가 발생한다. 이 때문에 오래전부터 협업 필터링을 개선하기 위한 연구가 진행되어왔다. 예를 들어 Sarwar (2001)는 콘텐츠 기반 등급과 SVD(Singular Value Decomposition) 기반 알고리즘과 아이템 기반 협업 필터링으로 희소성을 해결하였고, SVD 기반 차원 축소와 클러스터 기법을 활용하여 확장성을 향상시켰다. 또한 Acilar와 Arslan (2009)은 공간의 분포와

클러스터의 상호 관계를 포함하여 개선된 협업 필터링 모형을 제안하였다. 다음으로 이재식과 박석두(2007)는 MovieLens 데이터를 활용하여 장르별로 영화를 추천하는 시스템을 연구하였다. 이 연구는 추천 시스템이 처음부터 모든 영화를 대상으로 적용되는 것이 아니라, 최근접이웃을 기반으로 고객별 상위 3개의 추천 장르를 선정하고, 선정된 장르에서 상위 10개의 영화를 추천하는 프로세스로 진행된다. 이처럼 추천 장르를 먼저 선택하는 과정을 통해 고객과 아이템 매트릭스의 차원을 축소하였고, 이를 통해 매트릭스의 희소성과 확장성을 대처하는 방안을 제시하였다.

이와 같이 협업 필터링 모형은 다양한 접근법이 연구되어 실제 서비스에 활용되고 있다. 대표적으로 Amazon.com은 사용자가 구매하거나 평가한 상품과 유사도가 높은 상품을 추천하는 아이템 기반 협업 필터링을 활용하여 성공적으로 추천 시스템을 도입한 기업으로 알려져 있다(Linden, Smith & York, 2003).

2.2 규칙 기반 접근법

연관성 분석 또는 장바구니 분석은 트랜잭션 데이터에서 연관이 있거나 동시 발생한 항목을 추출하고 고객의 구매패턴을 발견하여 연관규칙을 생성하는 알고리즘이다(Chen, Tang, Shen et al., 2005). 연관 규칙은 항목의 출현 패턴과 두 패턴 사이의 조건부 확률에 기초하여 매우 간단하게 해석할 수 있으며, 더 쉽게 이해할 수 있는 IF-THEN 형식으로 표현된다(Borah & Nath, 2018). 예를 들어 “상품 A를 구매하고 상품 B도 함께 구매한다.”라는 연관규칙은 A→B(if A, then B)로 나타낸다. 연관성 분석은 일반적으로 지지도, 신뢰도, 향상도를 통해 규칙을 평가하며, 추출된 연관규칙집합에서 세 지표를 통해 중요 규칙을 선별한다(Ordonez, 2006, Mossong, Hens, Jit et al., 2008). 연관규칙은 A→B일 때, 지지도는 전체 거래 중 상품 A와 B를 함께 구매한 거래의 비율을 의미하며, 빈발 상품 집합을 판별하기 위해 사용할 수 있다. 이때 지지도는 다음과 같은 식으로 표현한다.

$$s(A \rightarrow B) = P(A \cap B)$$

다음으로 신뢰도는 상품 A를 구매한 거래 중에서 B도 함께 구매한 비율으로 계산된다. 이는 상품 집합 간 연관성의 강도를 판단하기 위해 사용할 수 있으며, 신뢰도는 다음과 같은 식으로 표현한다.

$$c(A \rightarrow B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

마지막으로 향상도는 상품 A를 구매했을 때 그 거래가 상품 B를 함께 구매한 경우와 임의로 구매된 경우의 비율로 계산된다. 향상도는 상품 A와 B이 서로 독립적인지, 상관관계가 있는지를 판단할 수 있으며, 그 값이 1보다 클수록 양의 상관관계를 의미하고, 1일 때에는 독립의 관계, 1보다 작으면서 0에 가까울수록 음의 상관관계를 갖는다. 향상도는 다음과 같은 식으로 표현된다.

$$Lift(A \rightarrow B) = \frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

연관성 분석은 상품의 수가 많아질수록 연관규칙의 양이 기하급수적으로 증가하여 추천 시스템의 성능을 저하시킬 수 있다. 따라서 연관성 분석은 최소 지지도와 최소 신뢰도를 기준으로 연관규칙을 필터링하여야 하고, 최소 기준은 추천 시스템을 활용하는 사용자가 결정할 수 있다(Wang, Chuang, Hsu et al., 2004).

연관성 분석을 활용한 규칙 기반 접근법에서도 콜드 스타트 등의 문제를 해결하기 위해 다양한 방법론이 연구되고 있다(Shaw, Xu & Geva, 2010, Min & Zhu, 2013). 예를 들어 Voditel & Deshpande (2013)은 주식 시장의 포트폴리오 추천 시스템을 위해 지지도와 신뢰도를 바탕으로 주식 시장의 기하급수적인 항목 집합을 축소하는 연관규칙 필터링 방법을 개발하였다. 또한 Lin, Alvarez & Ruiz (2002)는 항목 간의 연관성과 함께 사용자 간의 연

관성을 고려하였고, 최소 지지도를 미리 설정하여야 하는 기존의 연관성 분석과 달리 연관 규칙 마이닝 프로세스 중 최소 지지도를 조정하여 지정된 범위 내의 규칙 수를 갖는 방법론을 제안하였다. 다음으로 이동원의 연구(2017)에서는 모바일 상거래 상에서 한정된 수의 추천 상품을 노출시키기 위해, 연관성 분석의 결과와 회귀분석을 결합하여 상품별 추천 순위를 계산하는 방법론을 제안하였다. 여기서 연관성 분석의 결과는 지지도, 신뢰도, 향상도를 말하며, 이를 독립변수로 활용한 선형 회귀분석 모형을 생성하였다. 이처럼 보편적으로 활용하는 지표들로 모형을 구성함으로써, 모형에 대해 직관적으로 이해하기 쉬우며, 실무에서도 활용하기 간편한 방안을 제시하였다. 마지막으로 류기동 등의 연구(2016)에서는 ARS(Automatic Response System) 환경에서 고객 만족도를 향상시키기 위해 노출되는 메뉴, 특히 실제 고객의 서비스가 이루어지는 최종 단계에서의 메뉴 순서를 추천하기 위한 방법론을 제안하였다. 이를 위해 연관성 분석이 활용되었으며, 전체 71개 메뉴 중 최소 지지도, 최소 신뢰도, 최소 향상도를 만족하는 14개의 메뉴가 최종적으로 도출되었다. 이때 연관성 분석에 의해 추천 메뉴가 생성되지 못한 나머지 메뉴들은 피어슨 상관 분석을 통해 위 14개 메뉴와의 유사도를 측정하였고, 가장 높은 유사도를 보인 메뉴의 추천 메뉴를 도출하도록 결정하였다.

2.3 하이브리드 기반 접근법

하이브리드 기반 추천 시스템은 개별적으로 작용하던 다양한 알고리즘을 결합하여 하나의 프로세스를 구성하거나, 선호 등급과 같은 기본적인 데이터 이외에 추천 시스템의 성능을 향상시킬 수 있는 보조 데이터를 활용하는 방법론을 말하며, 개별 알고리즘에서 존재하던 단점을 개선하여 결과적으로 추천 시스템의 성능을 향상시킬 수 있다(Burke, 2002, Wang, Jhou & Tsai, 2018). 이러한 장점으로 인해 하이브리드 기반 접근법은 추천 시스템에서 가장 활발히 연구되고 있는 영역이다.

하이브리드 기반 추천 시스템에서 여러 추천 알고리즘을 결합한 예로 Chen, Niu, Zhao et al. (2014)의 연구가 있다. 이 연구는 e-러닝 환경에서

사용자에게 학습 항목을 추천할 때, 아이템 기반 협업 필터링을 사용하여 관련 아이템 셋을 발견하고, 이를 순차 패턴 마이닝에 적용하여 아이템을 필터링하는 두 단계의 프로세스를 제안하였다. 또한 Kim, Park, Oh et al. (2017)은 CNN (Convolutional Neural Network)과 PMF(Probabilistic Matrix Factorization)를 활용하여 문서의 문맥을 인식하여 추천에 반영하는 하이브리드 알고리즘을 연구하였다. 다음으로 손창환 등의 연구(2006)에서는 사용자 기반 협업 필터링과 아이템 기반 협업 필터링을 함께 사용하는 방법론을 제안하였다. 이 연구는 먼저 사용자 기반 협업 필터링을 통해 사용자 사이의 유사도를 계산하고, 대상 고객과 일정 기준 이상의 유사도를 보인 사용자들의 리스트를 선정하였다. 그 다음 선정된 사용자와 대상 고객의 아이템별 선호도를 활용하여 아이템 기반 협업 필터링을 적용하였다. 그 결과 단순히 사용자 기반 협업 필터링 또는 아이템 기반 협업 필터링을 활용하였을 때보다 제안된 하이브리드 협업 필터링에서 가장 높은 추천 정확률을 보였다.

다음으로 보조적인 데이터를 활용한 하이브리드 추천 시스템의 예로, Paradarami, Bastian & Wightman (2017)은 식당의 선호 등급을 예측하기 위해 외식 산업의 데이터를 활용하였다. 이 연구에서는 외식 산업을 대상으로 생성된 알고리즘을 건강/의료, 홈서비스, 쇼핑 등 10가지 비즈니스 영역에 적용하여 추천 시스템의 실현 가능성을 검증하였다. 또한 Yang, Cheng & Dia (2008)은 모바일 쇼핑에서 추천을 위한 위치 정보를, Hu, Guo & Zhang (2009)은 음악 추천을 위해 사용자의 선호도와 음악의 컨텐츠 정보를 함께 활용하였다. 다음으로 김현희 등(2013)은 사용자의 청취 습관과 태그 정보를 사용하여 음악을 추천하는 방법론을 제시하였다. 여기서 사용자의 청취 습관이란 사용자가 각 음악을 청취한 횟수를 선호도로 측정한 것을 말하며, 태그 정보는 각 음악에 대해 사용자가 텍스트 형태로 입력한 정보이다. 이러한 태그 정보 중 일부는 긍정적인 감정, 또는 부정적인 감정을 나타내는 감정 태그이며, 이를 통해 음악에 대한 선호도로 사용하였다. 또한 이를 혼합하여 하이브리드 추천을 위한 프로파일을 생성하였다. 이는 청취 습관과 태그 정보가 동시에 존재하는 아이템이 포함되며, 두 정보의 선호도에 가중치를 곱하여 새로운 선호도를 계산하였다. 이와 같이 세 가지 방식으로 생성된 추천 알고리

증의 결과를 분석하여 제안된 하이브리드 추천 알고리즘의 성능 향상을 확인하였다.

Ⅲ. 실험

3.1 데이터

본 연구는 제안된 추천 시스템을 성능을 검증하기 위해 국내 보험 업종의 텔레마케팅 대행사로부터 제공받은 고객 마스터 정보와 텔레마케팅 이력, 포인트 이력, 그리고 보험 청약 이력을 활용하였다. 그 중 2016년 10월부터 2017년 9월 사이에 발생한 텔레마케팅 이력에서 보험 청약까지 성공한 39,839건과 성공하지 못한 39,839건의 트랜잭션을 선정하였다. 이때 선정된 트랜잭션에 해당하는 69,338명의 고객의 정보들도 함께 수집하였다.

데이터 수집이 완료되면 분석을 위한 데이터 셋을 구성한다. 본 연구의 초기 데이터 셋에는 140개의 변수가 존재하였으나 추천 시스템의 속도와 예측 정확도를 향상시키기 위하여 특징 선택(Feature Selection)을 통해 불필요한 변수를 제거하였다. 그 결과 표 3-1과 같이 51개의 변수가 선정되었다.

[표 3-1] 데이터 스키마

데이터 셋	변수명	변수 타입
고객 마스터	나이	연속형
	세그먼트	범주형
	포인트	연속형
	동의일자	범주형
	가입기간	연속형
포인트 이력	주요포인트사용처	범주형
	포인트사용처개수	연속형
	지불수단개수	연속형
	평균누적포인트	연속형

[표 3-1] 계속

데이터 셋	변수명	변수 타입
포인트 이력	주유매출빈도	연속형
	주유지불수단개수	연속형
	평균주유사용포인트	연속형
	총주유매출금액	연속형
	평균주유매출금액	연속형
	총주유누적포인트	연속형
	평균주유누적포인트	연속형
	주요주유상품코드	범주형
텔레마케팅 이력	접촉성공여부	범주형
	주요통화분류코드	범주형
	통화결과코드개수	연속형
	평균통화시간	연속형
	평균통화성공시간	연속형
	통화성공율	연속형
	긍정적통화성공율	연속형
보험 청약 이력	주요청약사	범주형
	청약횟수	연속형
	주택보험청약수	연속형
	상해보험청약수	연속형
	암보험청약수	연속형
	실버보험청약수	연속형
	어린이보험청약수	연속형
	운전자보험청약수	연속형
	치아보험청약수	연속형
	종신보험청약수	연속형
	주택보험청약이력	범주형

[표 3-1] 계속

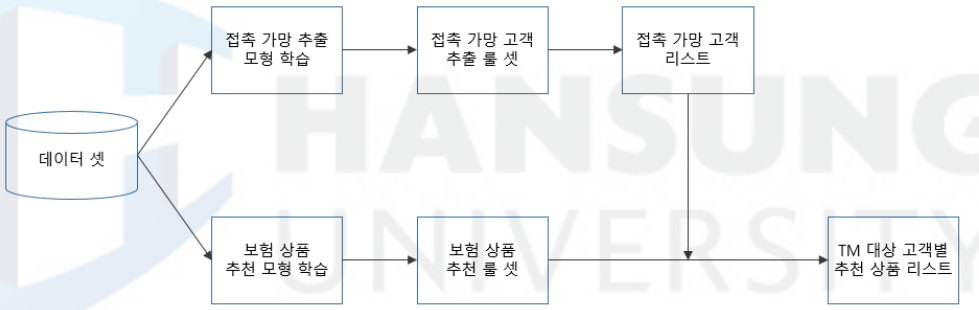
데이터 셋	변수명	변수 타입
보험 청약 이력	상해보험청약이력	범주형
	암보험청약이력	범주형
	실버보험청약이력	범주형
	어린이보험청약이력	범주형
	운전자보험청약이력	범주형
	치아보험청약이력	범주형
	종신보험청약이력	범주형
	청약보험종류개수	연속형
	주택보험청약여부	범주형
	상해보험청약여부	범주형
	암보험청약여부	범주형
	실버보험청약여부	범주형
	어린이보험청약여부	범주형
	운전자보험청약여부	범주형
	치아보험청약여부	범주형
	종신보험청약여부	범주형

고객 마스터 테이블은 나이, 성별, 거주지역 등의 인구통계적인 정보와 함께 세그먼트, 동의일자, 가입기간 등 고객별 특성을 관리하고 있는 테이블을 말한다. 또한 포인트 이력 테이블은 해당 기업에서 운영하고 있는 제휴 멤버십 포인트에 대한 적립, 사용 이력이 누적된 테이블을 말한다. 포인트 이력은 크게 일반 포인트 이력과 주유 포인트 이력으로 구분된다. 다음으로 텔레마케팅 이력은 고객에게 텔레마케팅을 시도하였을 때 발생한 변수들을 포함한다. 여기서 접촉성공여부는 각 고객이 텔레마케팅의 대상으로 선정된 1개월간의 접촉 시도에서 한 번 이상 접촉이 이루어졌을 경우를 성공으로 정의하였다. 마지막으로 보험 청약 이력은 텔레마케팅에서 접촉이 성공하였을 경우 이어지는 보험 상품의 청약 권유에서 발생한 변수들을 포함한다. 여기서 보험 상

품은 주택, 건강 및 상해, 암, 실버, 어린이, 운전자, 치아, 종신보험의 총 여덟 개 카테고리로 구분된다.

3.2 다목적 하이브리드 추천 모형 개발

본 연구의 목적은 접촉 가능성이 높은 고객을 우선적으로 선별하고, 선별된 고객에게 청약 가능성이 높은 개인화된 보험 상품을 추천하는 것이다. 이를 위해 본 연구에서는 그림 3-1과 같이 접촉 가능성이 높은 고객을 예측하는 모형과 개인화된 상품을 추천하는 모형을 결합한 두 단계의 다목적 하이브리드 추천 알고리즘을 개발하였다.



<그림 3-1> 다목적 하이브리드 추천 시스템 개념도

3.2.1 접촉 가망 추출 모형

접촉 가망 고객을 선별하기 위한 모형은 데이터 셋 생성, 모형 학습, 모형 적용 및 접촉 가망 고객 선별의 프로세스로 진행된다. 이때 모형을 학습하는 알고리즘으로 의사결정나무, 인공 신경망, 서포트 벡터 머신, 로지스틱 회귀, 랜덤 포레스트, 그래디언트 부스팅이 고려되었다.

3.2.1.1 데이터 셋 생성

먼저 접촉 가망 추출 모형을 학습하기 위한 데이터 셋을 구성한다. 데이터

셋은 고객 마스터 정보와 포인트 이력을 독립변수로, 텔레마케팅 이력의 접촉 성공 여부를 종속변수로 활용하였다.

또한 예측 모형의 일관성 있는 성능 비교를 위해 데이터 셋을 트레이닝 셋과 테스트 셋으로 분할한다. 여기서 트레이닝 셋은 모형을 학습하기 위한 용도로, 테스트 셋은 학습된 모형을 적용하여 성능을 평가하기 위한 용도로 사용하며, 트레이닝 셋과 테스트 셋의 비율은 7:3으로 분할하였다.

3.2.1.2 모형 학습

이 단계는 위에서 선정된 여섯 가지 후보 머신러닝 알고리즘을 기반으로 접촉 가망 추출 모형을 학습하는 단계이다. 본 연구에서는 반복 실험을 통해 각 후보 알고리즘의 파라미터를 결정하고, 이를 통해 생성된 모형의 성능을 평가한다. 그 결과 가장 우수한 성능을 보이는 알고리즘을 접촉 가망 추출 모형의 알고리즘으로 선정한다.

3.2.1.3 모형 적용 및 접촉 가망 고객 선별

모형의 학습과 최종 알고리즘의 선정이 완료되면 테스트 셋에 모형을 적용하여 고객별 접촉 성공 가능성을 예측한다. 예측 결과는 0과 1사이의 스코어로 생성되며, 예측 스코어가 높은 순서대로 접촉 가망 고객으로 선별할 수 있다.

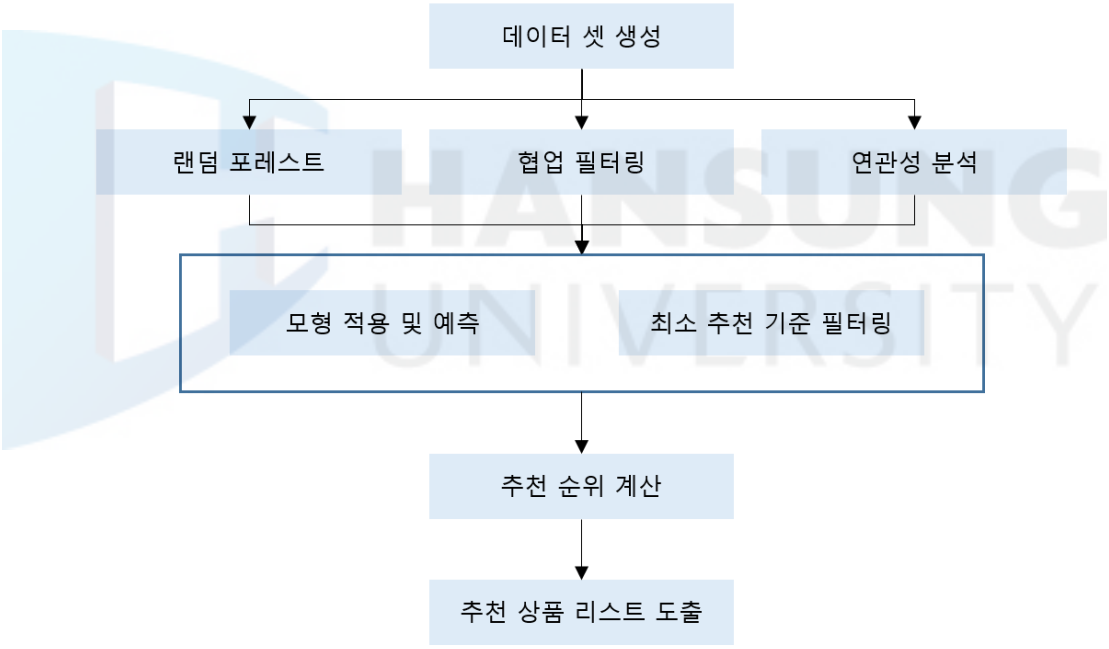
3.2.2 보험 상품 추천 모형

본 연구의 보험 상품 추천 알고리즘은 그림 3-2와 같은 프로세스를 통해 개발된다. 알고리즘 개발 프로세스는 총 여섯 단계로 구성되어 있으며, 랜덤 포레스트와 협업 필터링, 연관성 분석을 기반으로 보험 상품을 추천하는 모형을 생성한다. 그 결과 가장 우수한 성능을 보이는 모형을 보험 상품 추천 모형의 최종 알고리즘으로 선정한다.

3.2.2.1 데이터 셋 생성

추천 알고리즘을 개발하고 성능을 비교하기 위한 프로세스는 학습 모형에

활용할 데이터 셋을 생성하는 것으로부터 시작된다. 이 단계에서는 고객 마스터 정보, 텔레마케팅 이력, 청약 이력을 활용하여 모형별 분석용 데이터 셋을 구성한다. 그러나 랜덤 포레스트 모형과 협업 필터링, 연관성 분석은 분석을 위해 요구되는 데이터 셋의 형태가 다르다. 랜덤 포레스트 모형은 통합된 데이터 셋의 모든 변수를 포함하는 데이터프레임, 협업 필터링은 청약 이력만을 포함하는 등급 매트릭스, 연관성 분석은 청약 이력의 트랜잭션 형태로 변환하여야 한다. 또한 추천 성능을 평가하기 위한 모형에서 청약을 실패한 데이터는 필요하지 않으므로 청약에 성공한 트랜잭션만 추출하여 모형 학습에 활용하였다.



<그림 3-2> 보험 상품 추천 알고리즘 개발 프로세스

3.2.2.2 추천 모형 생성

본 연구에서 성능을 비교하고자 하는 추천 알고리즘은 랜덤 포레스트 모형, 협업 필터링 모형, 연관성 분석 모형이며, 협업 필터링과 연관성 분석의 경우 모형을 생성하기 위한 방법론을 결정하여야 한다. 협업 필터링은 유사도

를 계산하는 메트릭을 선정하여야하며 대표적으로 피어슨 상관관계, 코사인 유사도, 유클리드 거리 등을 사용한다. 또한 연관성 분석 모형은 신뢰도, 지지도, 향상도를 기준으로 연관규칙의 우선순위를 선택할 수 있다. 이러한 방법론은 반복 실험을 통해 성능을 비교/평가할 수 있고, 그 중 가장 우수한 결과를 보이는 방법론을 활용하여 모형의 최종 평가를 위한 알고리즘을 생성한다.

3.2.2.3 모형 적용 및 최소 추천 기준 필터링

추천 시스템을 위한 추천 모형이 학습되면, 각 모형을 테스트 셋에 적용하여 트랜잭션별 추천 상품을 예측한다. 그 결과로 랜덤 포레스트 모형은 추천 상품별로 0과 1사이의 추천 스코어가 도출되고, 협업 필터링은 추천 등급, 연관성 분석은 연관규칙이 생성된다.

보험 상품 추천 알고리즘의 목표는 제안된 추천 시스템을 활용하여 텔레마케팅의 성과 개선을 위해 청약 가능성이 높은 보험 상품을 추천하는 것이다. 그러나 우선순위만을 기준으로 하는 추천은 상품간의 상대적인 순위만을 고려하며, 각 상품의 절대적인 추천 정도는 알 수 없다. 따라서 순위 기반 추천 시스템은 추천 스코어가 낮더라도 순위가 높은 상품을 추천하게 되는 문제가 발생하고, 이는 추천 성공률의 하락과 추천 시스템 자체의 신뢰성 저하로 이어진다. 이러한 이유로 절대적인 추천 스코어 값이 일정 기준 이하로 내려갈 경우 추천 대상에서 제거할 필요가 있다. 본 연구의 경우 랜덤 포레스트 모형은 추천 스코어를 기준으로, 협업 필터링 모형은 추천 등급을 기반으로, 연관성 분석 모형은 지지도나 신뢰도를 기준으로 최소 추천 기준을 결정할 수 있다. 이때 추천을 필터링하기 위한 지표나 값의 결정 기준은 존재하지 않지만, 산업, 상품, 상황, 사용자의 경험 등에 의해 적절한 기준을 설정할 수 있다.

이 단계에서 랜덤 포레스트 모형과 협업 필터링 모형은 테스트 셋으로 예측 결과를 생성한 뒤 최소 기준에 의해 추천 상품을 필터링하는 순서로 진행된다. 하지만 연관성 분석 모형은 알고리즘 특성상 연관규칙의 필터링을 거친 후 테스트 셋에 적용하여야 하는 차이가 존재한다.

3.2.2.4 추천 순위 계산

테스트 셋에 추천 모형 적용을 통한 추천 스코어 예측과 최소 기준에 의한 상품 필터링이 완료되면 생성된 추천 결과를 활용하여 추천 상품 순위를 계산할 수 있다. 이때 랜덤 포레스트 모형은 추천 스코어가 높은 상품부터 추천 순위를 부여할 수 있으며, 협업 필터링 모형은 상품별 추천 등급을 기준으로, 연관성 분석 모형은 연관 규칙을 기준으로 개인화된 보험 상품의 추천 순위를 계산한다. 동일한 순위로 예측되는 상품은 미리 결정된 상품의 중요도에 따라 순위를 결정할 수 있다. 본 연구에서는 보험 상품별 청약 성공률과 객단가 등을 기준으로 현업 종사자와의 논의를 통해 보험 상품별 추천 우선순위를 결정하였다.

3.2.2.5 추천 상품 리스트 도출

앞의 모든 단계가 완료되면 테스트 셋의 각 트랜잭션마다 추천 상품 리스트가 도출된다. 이를 활용하여 추천 모형의 성능을 비교하고 평가하여 가장 우수한 성능을 보이는 알고리즘을 보험 상품 추천 모형에 활용한다.

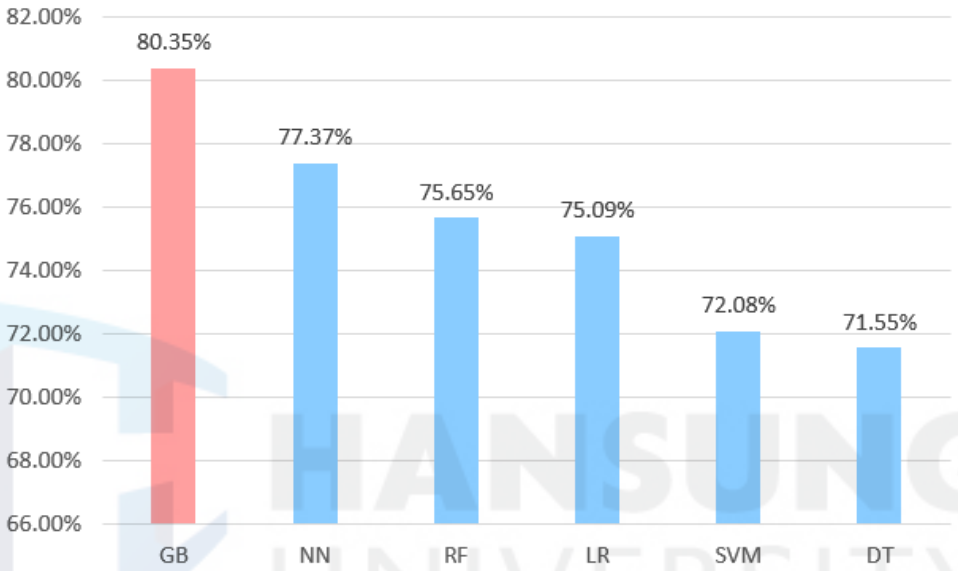
3.3 모형 성능 평가 및 최종 알고리즘 선정

다목적 하이브리드 추천 시스템의 알고리즘이 생성되면 모형의 성능을 평가하여 최종 알고리즘을 선정하여야 한다. 일반적으로 예측 모형의 성능을 평가하는 지표는 재현율, 특이도, 정밀도 등이 사용되며, 목적에 따라 적절한 평가 지표를 결정하여 모형의 성능을 비교하여야 한다.

3.3.1 접촉 가망 추출 모형 성능 평가

본 연구를 위해 데이터를 제공한 텔레마케팅 대행사에서 한 달 동안 접촉을 시도할 수 있는 고객의 수는 전체 보유 고객의 20% 정도에 해당한다. 이와 같이 일정 기간 동안 접촉을 시도할 수 있는 역량 이상으로 보유 고객이 존재할 때, 접촉 가망 추출 모형을 통해 텔레마케팅 대상 고객을 효과적으로 선별할 수 있어야 하며, 얼마나 정확하게 유효한 고객을 선별하였는지에 따라

모형의 성능이 결정된다. 이러한 관점에서 접촉 가망 추출 모형의 성능은 선별된 고객 중 유의미한 고객의 비율인 정밀도로 판단할 수 있다. 그림 3-3과 같이 여섯 가지 후보 알고리즘을 정밀도를 기준으로 비교한 결과 그래디언트 부스팅(GB)이 가장 우수한 성능을 보였다.



<그림 3-3> 접촉 가망 추출 모형 알고리즘별 성능 비교

3.3.2 보험 상품 추천 모형 성능 평가

보험 상품 추천 모형의 경우 추천 성공률로 성능을 평가할 수 있다. 본 연구에서는 3순위 이내의 순위에서 상품 추천이 적중한 경우 추천 성공으로 판단하였으며, 연관성 분석과 협업 필터링의 모형에 대한 성능을 평가한 결과는 다음과 같다.

보험 상품 추천 모형의 첫 번째 알고리즘인 연관성 분석 모형은 생성된 연관규칙을 신뢰도, 지지도, 향상도를 통해 우선순위를 결정할 수 있다. 본 연구에서는 실험을 통해 세 가지 지표의 추천 성능을 표 3-2와 같이 비교하였다. 그 결과 신뢰도를 기준으로 연관규칙의 우선순위를 결정하였을 때 접촉

가망 추출 모형에서 46.85%, 랜덤 추출에서 40.61%로 가장 높았으며, 지지도는 각각 45.39%, 39.72%, 향상도는 21.52%, 18.33%로 신뢰도를 기준으로 생성된 모형보다 추천 정확도가 낮게 나타났다. 따라서 본 연구에서는 신뢰도를 연관규칙의 우선순위 생성 기준으로 결정하였다.

[표 3-2] 연관성 분석 모형에 대한 추천 정확도

연관규칙 평가 지표	접촉 가망 추출 모형	랜덤 추출
신뢰도	46.85%	40.61%
지지도	45.39%	39.72%
향상도	21.52%	18.33%

다음으로 협업 필터링은 본 연구의 데이터 셋을 활용하기 가장 적합한 사용자 기반 협업 필터링을 추천 시스템을 위한 모형으로 결정하였다. 또한 모형에서 사용자간의 유사성을 측정하기 위한 방법으로 대표적인 피어슨 상관계수, 코사인 유사도, 유클리드 거리를 선택하였다.

[표 3-3] 협업 필터링 모형에 대한 추천 정확도

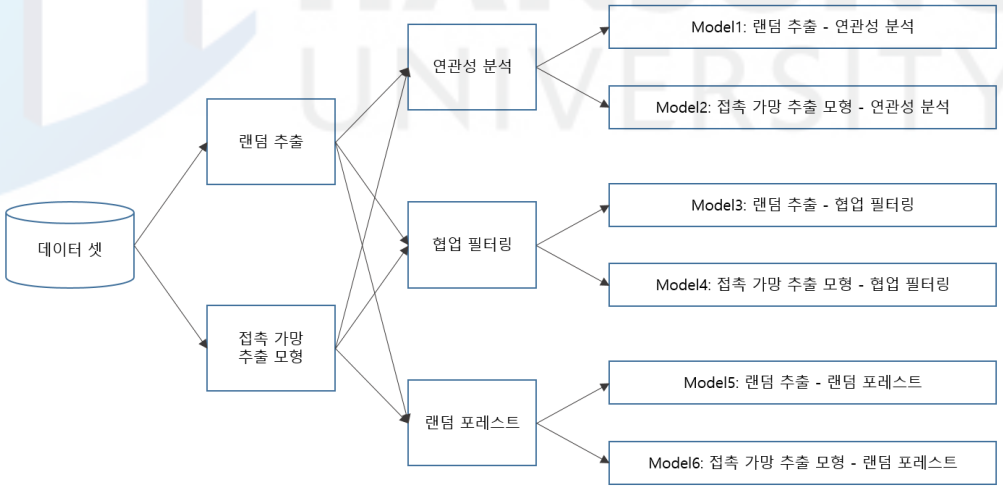
유사도 메트릭	접촉 가망 추출 모형	랜덤 추출
피어슨 상관계수	52.80%	45.57%
코사인 유사도	51.49%	45.27%
유클리드 거리	49.35%	42.60%

본 연구에서는 실험을 통해 각 유사도 메트릭으로 생성한 추천 모형의 성능을 표 3-3과 같이 비교하였다. 그 결과 피어슨 상관계수를 활용하였을 때 접촉 가망 추출 모형에서 52.80%, 랜덤 추출에서 45.57%의 추천에 성공하여 가장 우수한 성능을 보였고, 코사인 유사도는 각각 51.49%, 45.27%, 유클리드 거리는 49.35%, 42.60%의 예측에 성공하여 피어슨 상관계수보다는 낮은 성능을 보였다. 따라서 본 연구에서는 유사도 메트릭 중 가장 추천 정확도가

높은 피어슨 상관 계수를 사용자 기반 협업 필터링을 학습하는 메트릭으로 결정하였다.

Ⅳ. 다목적 하이브리드 추천 모형 성능 평가

접촉 가망 추출 모형과 보험 상품 추천 모형의 알고리즘이 모두 결정되면 두 모형을 결합한 다목적 하이브리드 추천 모형의 성능을 평가한다. 텔레마케팅 대상 고객을 선정하는 단계는 접촉 가능성이 높은 고객을 추출하는 것과 무작위로 고객을 추출하는 두 가지 방법의 성능을 비교하고, 보험 상품을 추천하는 단계는 연관성 분석, 사용자 기반 협업 필터링, 랜덤 포레스트의 세 가지 알고리즘의 성능을 비교한다. 따라서 두 단계의 결합으로 그림 4-1과 같이 총 여섯 개의 모형이 생성되며 그 중 본 연구에서 제안하고자 하는 모형은 접촉 가망 추출 모형과 랜덤 포레스트를 결합한 여섯 번째 모형이다.



<그림 4-1> 다목적 하이브리드 추천 시스템 연구 모형

모든 모형의 학습이 완료되면 각 모형의 성능을 평가하기 위해 별도의 테스트 셋을 생성하여야 한다. 우선 전체 테스트 셋에 접촉 가망 추출 모형을

적용한 뒤, 접촉 가능성이 높은 상위 50%에 해당하는 고객을 선별하여 접촉 가망 추출 모형을 평가하기 위한 테스트 셋으로 활용한다. 또한 테스트 셋에서 무작위 추출을 통해 50%의 고객을 선별하여 랜덤 추출을 평가하기 위한 테스트 셋을 구성한다. 그 결과 접촉 가망 추출 모형의 테스트 셋은 11,910건 중 9,570건의 접촉 성공 트랜잭션이 포함되어 80.35%의 접촉 성공률을 보였고, 랜덤 추출 기반 테스트 셋에서는 8,406건이 포함되어 70.58%의 접촉 성공률을 도출하였다.

4.1 전체 상품의 모형 성능 비교

서론에서 설명한 것처럼 대면 상담에서는 하나의 추천에서 끝나는 것이 아니라 복수 개의 상품을 연속적으로 추천할 수 있다. 따라서 본 연구에서는 모형의 정확한 성능 비교를 위해 3순위 상품까지의 추천 성공률을 모두 평가하였다.

먼저 첫 번째 모형인 랜덤 추출-연관성 분석 결합 모형을 테스트 셋에 적용하여 추천 상품을 예측한 결과, 1순위 보험 상품만 고려하였을 때 2,667건의 추천 성공(22.39%)으로 나타났다. 다음으로 1순위 추천에서 성공한 이력을 제외한 뒤 나머지 트랜잭션에 대해 2순위 추천 상품으로 추천 성공 여부를 판단한다. 이때의 성공 횟수는 1,388건으로 1순위와 합산하여 총 4,055건의 추천 성공(34.05%)을, 같은 방법으로 3순위의 추천 상품까지 고려하였을 때 총 4,837건(40.61%)의 추천 성공을 보인다.

두 번째 모형인 접촉 가망 추출 모형-연관성 분석 결합 모형도 위와 같은 방법으로 성능을 평가한다. 그 결과 1순위 추천에서 3,631건(30.49%), 2순위에서 누적 5,067건(42.54%), 3순위에서 누적 5,580건(46.85%)의 추천에 성공하였다. 이때 두 번째 모형의 결과는 모든 구간에서 첫 번째 모형의 결과보다 우수한 추천 성능을 보인다.

다음으로 사용자 기반 협업 필터링을 활용한 세 번째, 네 번째 결합 모형의 성능을 평가한다. 세 번째 모형의 결과는 1순위 추천에서 2,812건(23.61%), 2순위 추천에서 누적 4,446건(37.33%), 3순위 추천에서 누적

5,427건(45.57%)의 추천에 성공하였다. 또한 네 번째 모형의 결과는 1순위 추천에서 3,701건(31.07%), 2순위 추천에서 누적 5,325건(44.71%), 3순위 추천에서 누적 6,288건(52.80%)로 추천에 성공하였다. 이처럼 사용자 기반 협업 필터링을 활용한 결합 모형은 연관성 분석의 결과와 마찬가지로 랜덤 추출보다 접촉 가망 추천 모형이 우수한 추천 성능을 보였고, 모든 케이스에서 사용자 기반 협업 필터링이 연관성 분석 모형에 비해 우수한 성능을 도출하였다.

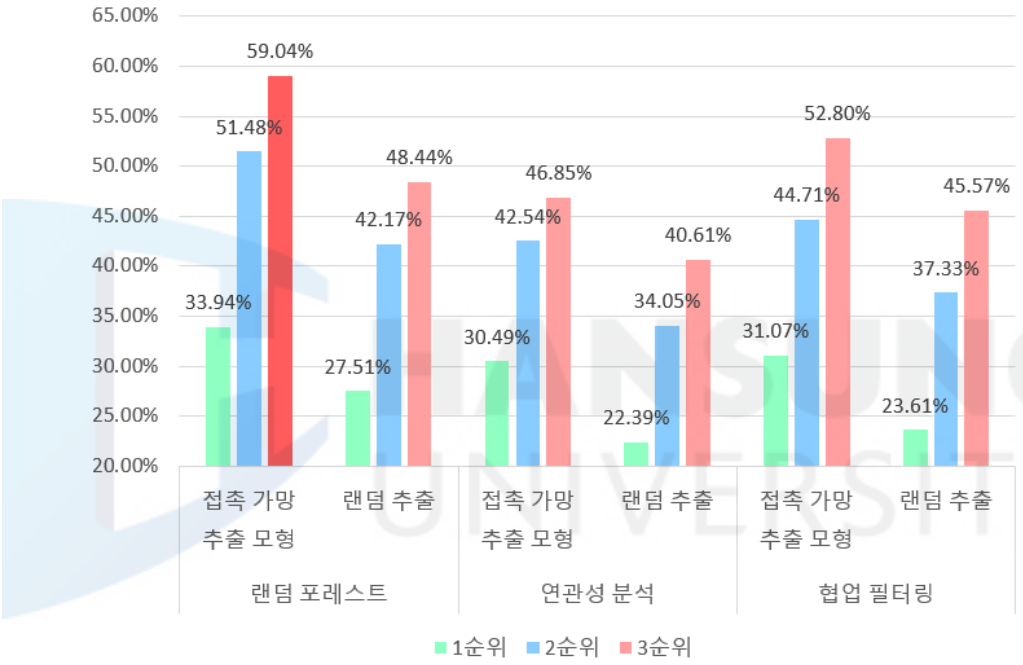
마지막으로 랜덤 포레스트를 활용한 모형의 결과를 평가한다. 랜덤 추출과 랜덤 포레스트를 결합한 다섯 번째 모형의 결과는 1순위 추천에서 3,276건(27.51%)의 추천 예측을 성공하였다. 이는 연관성 분석 모형에 비해 22.83%, 사용자 기반 협업 필터링 모형에 비해 16.50%의 추천 성공률의 향상을 보인다. 다음으로 2순위 추천에서 누적 5,022건(42.17%)의 추천에 성공하여 경쟁 모형에 비해 각각 23.85%, 12.96%의 추천 성공률의 향상을 보였으며, 3순위 추천에서 누적 5,769건(48.44%)의 추천 성공으로 경쟁모형과 비교하였을 때 각각 19.27%, 6.30%의 추천 성능의 향상을 보였다.

또한 본 연구에서 제안하고자하는 접촉 가망 추출 모형-랜덤 포레스트 모형을 결합한 여섯 번째 모형의 결과는 1순위 추천에서 4,042건(33.94%)의 추천에 성공하였다. 이는 연관성 분석 모형에 비해 11.32%, 사용자 기반 협업 필터링 모형에 비해 9.21%의 추천 성공률이 향상되었고, 랜덤 추출-랜덤 포레스트 결합 모형에 비해 23.38%가 향상된 결과이다. 또한 2순위 추천에서 누적 6,161건(51.48%)의 추천에 성공하여 경쟁 모형에 비해 21.00%, 15.14%의 성능이 향상되었으며, 랜덤 추출보다 22.08%의 추천 성능이 향상되었다. 다음으로 3순위 추천까지 모두 누적한 결과 11,910건의 트랜잭션 중 7,032건(59.04%)의 추천에 성공하여 아주 우수한 성능을 보였다. 이는 경쟁 모형에 비해 26.02%, 11.83% 향상, 랜덤 추출에 비해 21.89%가 향상된 결과이며, 여섯 가지 모형의 성능을 비교한 결과는 표 4-1과 같다.

[표 4-1] 전체 상품의 추천 성능 비교

1단계	2단계	순위	대상 수	성공 수	실패 수	누적 성공 수	추천 성공률
랜덤 추출	연관성 분석	1순위	11910	2667	9243	2667	22.39%
		2순위	9243	1388	7855	4055	34.05%
		3순위	7855	782	7073	4837	40.61%
	협업 필터링	1순위	11910	2812	9098	2812	23.61%
		2순위	9098	1634	7464	4446	37.33%
		3순위	7464	981	6483	5427	45.57%
	랜덤 포레스트	1순위	11910	3276	8634	3276	27.51%
		2순위	8634	1746	6888	5022	42.17%
		3순위	6888	747	6141	5769	48.44%
접촉 가망 추출 모형	연관성 분석	1순위	11910	3631	8279	3631	30.49%
		2순위	8279	1436	6843	5067	42.54%
		3순위	6843	513	6330	5580	46.85%
	협업 필터링	1순위	11910	3701	8209	3701	31.07%
		2순위	8209	1624	6585	5325	44.71%
		3순위	6585	963	5622	6288	52.80%
	랜덤 포레스트	1순위	11910	4042	7868	4042	33.94%
		2순위	7868	2089	5779	6131	51.48%
		3순위	5779	901	4878	7032	59.04%

본 연구에서 제안한 다목적 하이브리드 추천 모형의 순위별 추천 성공률을 비교한 결과는 그림 4-2와 같으며, 모형별 성능을 비교한 결과 랜덤 추출보다 접촉 가망 추출 모형에 기반하여 고객을 선별할 때 우수한 성능을 보였다. 또한 랜덤 포레스트 알고리즘이 연관성 분석이나 협업 필터링과 비교하였을 때 모든 케이스에서 우수한 추천 성공률을 도출하였다.



<그림 4-2> 전체 상품의 추천 성능 비교

4.2 저인기 상품의 모형 성능 비교

물론 추천 시스템에서 전체 트랜잭션에 대한 추천 성공률은 매우 중요한 요소이다. 그러나 롱테일 법칙(Brynjolfsson, Hu & Simester, 2011)에서 알 수 있듯이 저인기 상품에 대한 추천 성공률 또한 무시할 수 없다. 특히 머신러닝과 같이 과거의 데이터를 학습하여 새로운 데이터를 예측하고자 할 때에

는 종속변수의 비중이 높은 항목은 예측이 용이하지만, 비중이 낮은 항목은 상대적으로 예측이 어렵다는 문제점이 발생한다.

본 연구의 경우 여덟 가지의 보험 상품 중 청약 비중이 높은 상해, 암, 치아, 실버 보험이 전체 청약 트랜잭션의 대부분을 차지하며, 주택, 어린이, 운전자, 종신 보험 등 나머지 네 가지 보험 상품은 전체 11,910건 중 약 400~600건의 비교적 적은 청약 건수를 보인다. 따라서 본 연구에서는 청약 비중이 낮은 네 가지 보험 상품의 추천 성공률도 확인할 필요가 있다. 이때 랜덤 추출 방식은 전체 11,910건의 트랜잭션 중 저인기 상품 청약으로 이어진 횟수는 408건으로 나타났다. 반면 접촉 가망 추출 모형에 의해 선별된 11,910건의 트랜잭션에서는 총 526건의 저인기 상품 청약 트랜잭션이 포함되었다. 이와 같이 저인기 상품 측면에서도 랜덤하게 대상 고객을 선별하는 것보다 접촉 가망 추출 모형에 의해 선별된 결과가 더 우수한 결과를 나타내는 것으로 확인되었다.

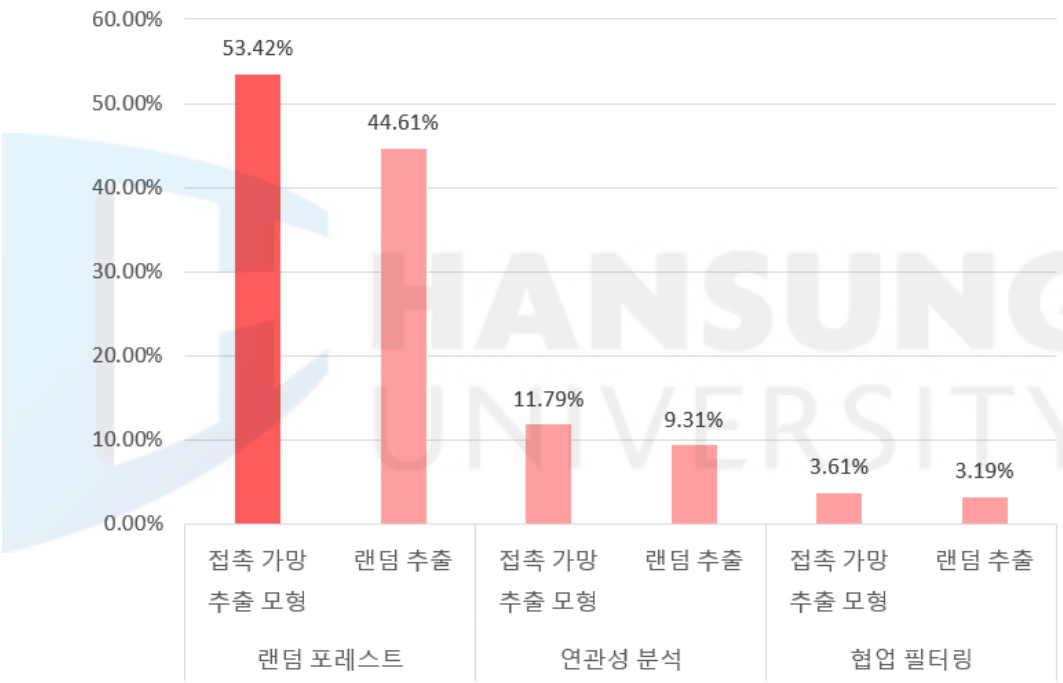
[표 4-2] 저인기 상품의 추천 성능 비교

1단계	2단계	대상 수	성공 수	실패 수	추천 성공률
랜덤 추출	연관성 분석	408	38	370	9.31%
	협업 필터링	408	13	395	3.19%
	랜덤 포레스트	408	182	226	44.61%
접촉 가망 추출 모형	연관성 분석	526	62	464	11.79%
	협업 필터링	526	19	507	3.61%
	랜덤 포레스트	526	281	245	53.42%

저인기 상품의 추천 성능을 비교한 결과는 표 4-2와 같으며, 저인기 상품은 1순위나 2순위로 추천되는 경우가 많지 않아 3순위의 추천 상품까지 한 번에 고려하였다. 첫 번째 모형은 전체 408건의 트랜잭션 중 38건(9.1%)만이 추천에 성공하였다. 또한 접촉 가망 추출 모형을 기반으로 하는 두 번째 모형은 전체 526건의 트랜잭션 중 62건(11.79%)의 추천에 성공하였다. 다음으로

사용자 기반 협업 필터링 모형을 활용한 세 번째, 네 번째 결합 모형의 결과는 각각 13건(3.19%), 19건(3.61%)의 추천에 성공하여 오히려 연관성 분석을 활용한 모형보다도 낮은 결과를 보였다.

반면 다섯 번째 결합 모형은 전체 408건의 트랜잭션 중 182건(44.61%), 여섯 번째 결합 모형은 전체 526건의 트랜잭션 중 281건(53.42%)의 추천에 성공하였다. 이를 비교한 결과는 그림 4-3과 같으며, 랜덤 포레스트 모형이 다른 모형에 비해 월등한 추천 결과를 산출하였다.



<그림 4-3> 저인기 상품의 추천 성능 비교

이처럼 접촉 가망 추출 모형에 기반한 고객의 선별이 무작위로 선별하는 방법에 비해 전체 상품과 저인기 상품의 두 가지 측면에서 모두 좋은 결과를 보였다. 따라서 개인화된 상품을 추천하기 전에 앞서 접촉 가능성이 높은 고객을 선별하는 알고리즘은 텔레마케팅에서 추천 시스템을 효과적으로 활용할 수 있는 접근으로 판단된다. 또한 다양한 컨텐츠 정보를 활용한 랜덤 포레스

트 모형이 모든 케이스에서 연관성 분석이나 협업 필터링에 비해 우수한 결과를 도출하여 본 연구에서 제안한 다목적 하이브리드 추천 시스템의 최종 알고리즘으로 선정되었다.

Ⅴ. 결론 및 향후 연구

5.1 결론

본 연구에서는 텔레마케팅 영역에서 비즈니스 성과를 향상시키기 위해 추천 시스템을 활용할 때 무작위로 텔레마케팅 대상 고객을 선별하는 것이 아니라 접촉 가능성에 기반하여 대상 고객을 선별한 뒤 선별된 고객에게 보험 상품을 추천하는 두 단계로 구성된 프로세스의 활용 가능성을 제시하였다.

첫 번째 단계인 접촉 가망 추출 모형은 머신러닝 알고리즘을 기반으로 고객별 접촉 가능성을 예측하였으며, 반복적인 실험을 통해 여섯 가지 후보 알고리즘 중 그래디언트 부스팅이 가장 높은 예측 성능을 보이는 것을 확인하였다. 그래디언트 부스팅 모형을 통해 추출된 접촉 가망 고객은 무작위로 텔레마케팅 대상 고객을 추출할 때보다 더 높은 비율로 보험 상품을 청약한 고객을 포함하여 효과적으로 텔레마케팅에 유효한 고객을 선별할 수 있는 방법론으로 확인되었다.

다목적 하이브리드 추천 시스템의 두 번째 단계인 보험 상품 추천 모형에서는 고객 마스터 정보, 텔레마케팅 이력 등의 컨텍스트 정보를 활용하여 기존 추천 알고리즘의 콜드 스타트와 희소성 문제를 해결하였다. 또한 실험 결과, 추천 알고리즘으로 활용된 랜덤 포레스트 모형은 기존의 추천 알고리즘인 연관성 분석이나 협업 필터링 기반 추천 시스템보다 우수한 예측 결과를 산출하여 본 연구의 추천 시스템이 텔레마케팅의 성과를 향상시킬 수 있음을 함께 보여주었다.

위와 같이 본 연구에서 제안된 다목적 하이브리드 추천 시스템은 전체 보험 상품을 대상으로 측정한 결과에서 비교 모형보다 더 우수한 성과를 보였다. 그 뿐 아니라 저인기 상품을 대상으로 유의미한 고객을 선별하는 변별력

과 추천 성공률 관점에서도 우수한 성능을 보여 기존 방식의 추천 시스템에 비교하여 높은 활용성을 기대할 수 있다.

5.2 한계 및 향후 연구

본 연구는 실제 텔레마케팅 대행사로부터 제공받은 데이터를 통해 모형이 개발되었다. 따라서 제안된 다목적 하이브리드 추천 시스템을 실제 텔레마케팅 과정에 적용하여 모형의 효용성을 확인할 수 있다. 또한 제안된 추천 알고리즘을 응용하여 기업이 보유하고 있는 동일한 고객군을 대상으로 리스 상품의 텔레마케팅에서 추천을 위한 알고리즘으로 확장할 수 있을 것으로 기대된다.

추천 시스템을 개발하기 위해 제공받은 데이터에서 보험 상품은 수백 가지로 구성되어 있었으며, 각 보험 상품의 목적상 유사성을 기반으로 여덟 가지의 카테고리로 분류하였다. 그러나 동일한 카테고리로 분류되더라도 보험사에 따라 취급하는 상품의 성격이 다를 수 있음을 배제하였다. 따라서 단순한 유사성에 의해 보험 상품을 카테고리로 분류하는 것이 아니라, 군집분석 등의 분석 알고리즘 통해 통계적인 관점의 유사 카테고리로 분류하는 과정을 통해 모형을 정교화할 수 있다. 그리고 제공받은 데이터 중 일부 보험사에서 데이터를 수집한지 오래되지 않아 충분한 데이터가 발생되지 못하였던 문제가 존재하였다. 따라서 보험사별로 충분한 데이터가 존재한다면 각 보험사를 위한 별도의 모형을 개발하여 추천 시스템의 품질 향상을 기대할 수 있다.

또한 본 연구에서는 텔레마케팅의 영역 중 보험 상품을 위한 추천 시스템의 성능만을 확인하였으며, 이를 모든 텔레마케팅 영역에 적용이 가능하다고 일반화하기에는 어려움이 존재한다. 따라서 추천 시스템의 일반화를 확인하기 위하여 상조, 카드, 통신 등의 계약 서비스 업종의 텔레마케팅 데이터를 활용하여 제안된 추천 시스템의 적용 가능성을 확인해볼 필요가 있다.

마지막으로 본 연구의 보험 상품 추천 모형의 알고리즘으로 랜덤 포레스트가 결정되었다. 이를 다항 분류가 가능한 SVM 등의 다른 머신러닝 알고리즘을 활용하여 새로운 하이브리드 추천 시스템을 개발하여 성능 개선을 시도

할 수 있다. 또한 본 연구에서 활용한 고객 마스터, 포인트 이력 등의 컨텍스트 정보와 함께 또 다른 부가 데이터를 활용하여 추천 알고리즘을 개선하고자 한다.

【참고문헌】

1. 국내 문헌

강남이, 박정기, 조기성, 유연. (2012). IKONOS 영상을 이용한 토지피복분류 기법 분석.『한국지형공간 정보학회지』, 20(3), 65-71.

김상균, 장준혁. (2008). SMV코덱의 음성/음악 분류 성능 향상을 위한 Support Vector Machine의 적용. 『電子工學會論文誌-SP』, 45(6), 142-147.

김석중, 신인자, 이병오. (2000). 우유의 소비행태 변화에 관한 연구: 의사결정나무분석기법을 이용하여.『농업경영정책연구』, 27(1), 148-161.

김성진, 안현철. (2014). 랜덤 포레스트를 활용한 기업채권등급평가 모형.『한국지능정보시스템학회 학술대회논문집』, 2014(5), 371-376.

김세형, 강상조, 박재현, 김혜진. (2008). 한국프로농구 경기기록 분석에 의한 승패결정요인.『한국체육측정평가학회지』, 10(1), 1-12.

김수진, 김보영. (2013). 로지스틱 회귀분석과 의사결정나무 분석을 이용한 일 대도시 주민의 우울 예측요인 비교 연구.『한국콘텐츠학회논문지』, 13(12), 829-839.

김태진, 홍정식, 전윤수, 박종률, 안태욱. (2018). 랜덤포레스트를 이용한 모기업의 하향 거래처 기업의 분류.『한국전자거래학회지』, 23(1), 1-22.

김현희, 김동건, 조진남. (2013). 사용자 청취 습관과 태그 정보를 이용한 하이브리드 음악 추천 시스템. 『한국컴퓨터정보학회 논문지』, 18(2), 107-116.

김희종, 김형도. (2014). 그라디언트 부스팅과 균형 분류를 이용한 채무 불이행 예측.『한국정보기술학회논문지』, 12(1), 155-164.

류기동, 김종명, 금영정, 강필성, 김우제. (2016). 연관 규칙 분석을 활용한

ARS 추천 메뉴 시스템 연구. 『한국정보기술학회논문지』, 14(3), 127-136.

박선미, 김구진, 최유주. (2008). Support Vector Machine을 이용한 의상 분류 기법.『한국정보과학회 학술발표논문집』, 35(2), 335-340.

서상완, 육순현, 남경원, 한종희, 권세윤, 홍성화, 김동욱, 이상민, 장동표, 김인영. (2013). 인공 신경망을 이용한 보청기용 실시간 환경분류 알고리즘.『대한의용생체공학회』, 34(1), 8-13.

손소영, 신형원. (1998). 데이터 마이닝을 이용한 교통사고 심각도 분류분석.『대한교통학회 학술대회지』, 1998(3), 373-381.

손창환, 김기수, 문석환. (2006). 추천 시스템의 예측 정확도 향상을 위한 하이브리드 협업 필터링.『한국산업경영학회 발표논문집』, 2006, 561-587.

손호선, 박재성, 김경옥, 차은종, 김경아. (2018). 인공신경망을 사용한 위암 유전자 데이터의 분류분석.『대한전기학회 학술대회 논문집』, 2018, 1235-1236.

신경식, 안수산. (2002). 데이터 마이닝 기법을 활용한 스팸 메일 분류 및 예측모형 구축에 관한 연구.『經營論叢』, 20(2), 89-105.

유은정, 안현철, 김재경. (2013). 고객 맞춤형 서비스를 위한 관객 행동 기반 감정예측모형.『지능정보연구』, 19(2), 73-85.

이동원. (2017). 연관상품 추천을 위한 회귀분석모형 기반 연관 규칙 척도 결합기법.『지능정보연구』, 23(1), 127-141.

이승환, 황진수. (2018). 랜섬웨어 탐지를 위한 동적 분석 자료에서의 변수 선택 및 분류에 관한 연구.『응용통계연구』, 31(4), 497-505.

이영섭, 김희경. (2007). 데이터마이닝 기법들을 이용한 한의학에서의 체질분류모형 개발: 소증항목을 중심으로.『Journal of the Korean Data Analysis Society』, 9(2), 597-609.

이재식, 박석두. (2007). 장르별 협업필터링을 이용한 영화추천시스템의 성능향상.『지능정보연구』, 13(4), 65-78.

이혜주, 정의현. (2013). 데이터마이닝을 이용한 학업성취 결정요인 탐색.『아동교육』, 22(2), 5-18.

임은정, 정순희. (2015). 의사결정나무 분석을 통한 한국 중고령자의 점진적 은퇴의사결정에 관한 연구. 『Financial Planning Review』, 8(4), 167-195.

정윤, 황석해. (1999). 인공신경망을 이용한 부실기업예측모형 개발에 관한 연구. 『한국데이타베이스학회』, 1(1), 415-421.

조홍규. (2003). 인공지능 방법을 이용한 신용평가 모형에 대한 개관. 『나이스채권평가 금융공학연구소』.

채경희, 김상철. (2010). 의사결정나무 기법을 활용한 백화점의 고객세분화 사례연구. 『유통과학연구』, 8(1), 13-19.

최영찬, 이민수. (2010). 모돈도태 의사결정지원을 위한 머신러닝모델 응용. 『농업경영·정책연구』, 37(3), 387-410.

최종후, 강현철, 한상태. (2006). 선거예측조사 의사결정유보층 분류 및 예측을 위한 의사결정나무모형의 비교와 평가. 『Journal of the Korean Data Analysis Society』, 8(1), 167-178.

최하나, 임동훈. (2013). 앙상블 SVM 모형을 이용한 기업 부도 예측. 『한국데이터정보과학회지』, 24(6), 1113-1125.

한은정, 김동건. (2015). 노인장기요양보험 이용지원 상담 대상자 선정모형 개발. 『응용통계연구』, 28(6), 1063-1073.

2. 국외 문헌

Acilar, A. M., & Arslan, A. (2009). A collaborative filtering method based on artificial immune network. *Expert Systems with Applications*, 36(4), 8324-8332.

Arsan, T., Koksal, E., & Bozkus, Z. (2016). Comparison of collaborative filtering algorithms with various similarity measures for movie recommendation. *International Journal of Computer Science, Engineering and Applications*, 6(3), 1-20.

Bobadilla, J., Ortega, F., Hernando, A., & Gutierrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46, 109-132.

Borah, A., & Nath, B. (2017). Rare association rule mining: a systematic review. *International Journal of Knowledge Engineering and Data Mining*, 4(3/4), 204-258.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence,* 43-52.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32.

Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science*, 57(8), 1373-1386.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2), 121-167.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331-370.

Burke, R. (2007). Hybrid Web Recommender Systems. *The Adaptive*

Web, 377-408.

Chen, W., Niu, Z., Zhao, X., & Li, Y. (2014). A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, 17(2), 271-284.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2), 339-354.

Christakou, C., Vrettos, S., & Stafylopatis, A. (2007). A hybrid movie recommender system based on neural networks. *International Journal on Artificial Intelligence Tools*, 16(5), 771-792.

Cochocki, A., & Unbehauen, R. (1993). Neural networks for optimization and signal processing, (1st), John Wiley & Sons, Inc, New York, NY, USA.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2010). Collaborative Filtering Recommender Systems. *Foundations and Trends in Human-Computer Interaction*, 4(2), 81-173.

Elouedi, Z., Mellouli, K., & Smets, P. (2000). Classification with belief decision trees. Aritificial Intelligence: Methodology, Systems, and Applications, *Springer*, 80-90.

Ghazarian, S., & Nematbakhsh, M. A. (2015). Enhancing memory-based collaborative filtering for group recommender systems. *Expert Systems with Applications*, 42(7), 3801-3812.

Hsu, M. H. (2008). A personalized english learning recommender system for ESL students. *Expert Systems with Applications*, 34(1), 683-688.

Hu, R., & Pu, P. (2010). Using personality information in collaborative filtering for new users. *Recommender Systems and the Social Web*, 17-24.

Hu, B., Guo, M., & Zhang, H. (2009). A hybrid music recommendation system by M-LSA. *The International Conference on Computational Intelligence and Natural Computing*, 1, 129-132.

Kardan, A. A., Abbaspour, S., & Hendijanifard, F. (2009). A Hybrid Recommender System for E-learning Environments Based on Concept Maps and Collaborative Tagging. *The 4th international conference on virtual learning*, 300-307

Khribi, M. K., Jemni, M., & Nasraoui, O. (2007). Toward a Hybrid Recommender System for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 6136-6145.

Kim, D., Park, C., Oh, J., Lee, S., & Yu, H. (2016). Convolutional matrix factorization for document context-aware recommendation. *Proceedings of the 10th ACM Conference on Recommender Systems*, 233-240.

Lacerda, A. (2017). Multi-Objective Ranked Bandits for Recommender Systems. *Neurocomputing*, 246, 12-24.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11), 54-64.

Langseth, H., & Nielsen, T. D. (2015). Scalable learning of probabilistic latent models for collaborative filtering. *Decision Support Systems*, 74, 1-11.

Li, R. H., & Belford, G. G. (2002). Instability of decision tree classification algorithms. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM*, 570-575.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1),

76-80.

Menard, S. (2002). Applied logistic regression analysis(v.106). Sage.

Min, F., & Zhu, W. (2013). Cold-start recommendation through granular association rules. https://arxiv.org/abs/1305.1372. Accessed 18 December 2018.

Mobasher, B., Burke, R., & Sandvig, J. (2006). Model-based collaborative filtering as a defense against profile injection attacks. *Proceedings of the 21st National Conference on Artificial Intelligence*, 2, 1388-1393.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., & Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS MEDICINE*, 5(3), 381-391.

Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, 10(2), 334-343.

Paradarami, T. K., Bastian, N. D., & Wightman J. L. (2017). A hybrid recommender system using artificial neural networks. *Expert Systems With Applications*, 83, 300-313.

Paranjape-Voditel, P. & Deshpande, U. (2013). A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, 13(2), 1055-1063.

Reshma, R., Ambikesh, G., & Thilagam, P. S. (2016). Alleviating data sparsity and cold start in recommender systems using social behaviour. *2016 International Conference on Recent Trends in*

Information Technology, 1-8.

Richhariya, B., & Tanveer, M. (2018). EEG signal classification using universum support vector machine. *Expert Systems with Applications*, 106, 169-182.

Sarwar, B. (2001). Sparsity, scalability, and distribution in recommender systems. PhD thesis, University of Minnesota.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285-295.

Tewari, A. S., & Barman, A. G. (2018). Sequencing of items in personalized recommendations using multiple recommendation techniques. *Expert Systems with Applications*, 97, 70-82.

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buidings*, 158, 1533-1543.

Wang, H. C., Jhou, H. T., & Tsai, Y. S. (2018). Adapting topic map and social influence to the personalized hybrid recommender system. *Information Sciences*, 1-17.

Wang, X., & Wang, Y. (2014). Improving Content-based and Hybrid Music Recommendation using Deep Learning. *Proceedings of the 22nd ACM international conference on Multimedia*, 627-636.

Yang, W. S., Cheng, H. C., & Dia, J. B. (2008). A location-aware recommender system for mobile shopping environments. *Expert Systems with Applications*, 34(1), 437-445.

ABSTRACT

Multi-purpose hybrid recommendation system to improve telemarketing performance

Park, Eui-Beom

Major in Industrial & Management Engineering

Dept. of Industrial & Management Engineering

The Graduate School

Hansung University

A variety of recommendation algorithms such as collaborative filtering are useful methodologies for recommended products to customers. The telemarketing industry can also use the recommendation system to improve business performance. However, due to the nature of telemarketing, there is a problem that if the contact with the customer fails, the product can not be recommended even if the accuracy of the recommendation system is high. In this study, we propose a series of processes to select customers with high possibility of contact and recommend products to selected customers in order to achieve maximum business performance within limited resources. In addition, we have demonstrated through experiments that the performance of the multi-objective hybrid recommendation algorithm proposed in this study is superior to those of

randomly selecting customers and recommended products.

Keywords: recommender system, machine learning, telemarketing, target customer selection, gradient boosting, random forest