공학전문석사 학위 연구보고서

고객이탈 예측 모델링 기반 기대수익 최적화 방안

- 고객이탈 방지를 위한 마케팅 비용 최적화 -

Profit based model selection and optimization

for customer retention

2020년 2 월

서울대학교 공학전문대학원

응용공학과 응용공학전공

김 재 엽

고객이탈 예측 모델링 기반 기대수익 최적화 방안

– 고객이탈 방지를 위한 마케팅 비용 최적화 –

지도 교수 구 윤 모

이 프로젝트 리포트를 공학전문석사 학위
연구보고서로 제출함
2020년 2 월

서울대학교 공학전문대학원
응용공학과 응용공학전공
김 재 엽

김재엽의 공학전문석사 학위 연구보고서를 인준함
2020 년 2 월

위 원 장 _____ 한 훈 _____ (인)

위 원 _____ 구 윤 모 _____ (인)

위 원 _____ 곽 우 영 _____ (인)

국문초록

국내 통신 3사의 마케팅 비용이 연 8조, 영업이익보다 2.3배 높은 수준(2018년 국정감사 자료 기준)으로 고객유치/이탈방지를 위해 총성 없는 전쟁을 벌이고 있다. 본 연구에서는 고객 이탈 방지를 위한 리텐션 마케팅 비용 최소화를 위한 방법으로 우선 다양한 머신러닝 기법을 적용하여 고객 이탈 확률을 예측하였고 이를 기반으로 통신사 입장에서 타겟 마케팅을 할 때 기대수익을 최대화 할 수 있는 최적화 실험을 진행하였다.

IBM Watson Analytics, Guide to Sample Data Sets의 고객지원(Customer Support) 파일을 데이터 세트(총 7,043명)로 활용하였고 머신러닝 결과 logistic 모델링이 accuracy 80.4%로 가장 높게 나타났다. 통신사 전환 비용을 약 10만원으로 가정하고 마케팅 비용 예산을 1억으로 한정했을 때, 고객이탈 확률을 줄이기 위해 고객 타겟 마케팅(보조금 지급) 방식을 4가지(총 예산/N, 고객 월 요금 비율, 이탈 확률 비율, 고객 월 요금*이탈 확률 비율)로 달리하여 기대 수익을 시뮬레이션하여 비교한 결과, 본 연구의 최적화 방안을 적용하면 기대수익이 22%~114%까지 향상되는 것을 확인 할 수 있었다. 또한 비용 예산을 10~100억으로 늘려 실험한 결과, 마찬가지로 최적화 기법 적용 시 기대수익이 가장 높았다. 특히 최적화 연산 수행 시간이 지수 형태로 증가하면 실제 적용이 어려울 수 있으나, 실험 결과 연산 복잡도가 샘플 수 또는 마케팅 비용에 대해 선형으로 달라지는 것을 확인하였기 때문에 실무 적용에 있어 계산에 문제가 없었다.

통신업계가 스마트폰 시장 포화로 신규 가입자 유입이 적은 상황에서 번호이동보다는 비용이 상대적으로 적게 드는 '기존 고객 유지 전략'이 마케팅 비용을 줄이면서 경쟁력을 이어갈 수 있다는 점을 고려할 때, 이동통신 마케팅 예산 전략/정책 입안자는 본 연구를 활용하여 한정된 마케팅 예산을 효율적으로 분배할 수 있는 근거를 마련하고 또한 예산을 사용하면서 고객의 이탈을 막을 수 있는 전략적

방향을 제시할 수 있다.

주요어 : 고객 이탈 예측, 머신러닝, 최적화, 기대수익, 리텐션 마케팅, 프로모션
학 번 : 2018-20096

목 차

표 목차

그림 목차

제 1 장 서 론

고객 리텐션은 기업 또는 단체가 기존에 확보한 고객들의 이탈률을 줄이기 위한 행위를 말한다. 고객 리텐션 프로그램의 목적은 기업들이 고객 이벤트 또는 브랜드 이벤트 등을 통해 최대한으로 기존 고객을 유지 할 수 있도록 하는 것이다[1]. 최근 이통사들이 IPTV와 인터넷을 포함한 결합상품으로 집토끼 지키기 전략에 더욱 힘을 싣는 부분도 기존 고객의 이탈을 막고 가족 단위로 고객을 유지함으로써 고객 로열티의 효과를 높이기 위해서이다.

일반적으로 기존 고객을 유지하면 새로운 고객 발굴과 비교해 수익성은 5~7배 향상된다. 베인앤드컴퍼니의 대표는 장기적인 고객의 가치를 마케팅 비용 절감, 네트워크 효과, 승수 효과를 신규고객 발굴 대비 수익성 향상의 근거로 제시하며 신규 고객을 늘리기보다 기존 고객 유지에 힘써야 한다고 말했다[2].

현재 다양한 조직에서 고객 이탈을 예측하고 방지해야 할 필요성이 높아지면서 많은 데이터 마이닝 기술과 머신 러닝 기술이 사용되고 있다. 고객 이탈을 예측하는 것은 고객 유지에 중요하며, 많은 산업에서 막대한 손실을 막기 위해 필수가 되었다. 고객 이탈을 예측할 수 있는 안정적인 모델 구축뿐만 아니라, 회사가 큰 손실을 피하기 위해선 고객 리텐션을 어떻게 효율적으로 할 것인가도 매우 중요하다[3].

본 연구에서는 IBM Watson Analytics의 샘플 데이터를 이용하여 이통사 환경에서 고객 이탈을 예측하고, 회사의 손실을 줄이기 위해 고객 리텐션 최적화 방안을 제안하고 한다. 또한 현실적인 고객 리텐션 운영이 반영된 시뮬레이션 환경을 가정하고 이를 최적화 방안과 비교 분석을 수행함으로써, 제안된 최적화 방안의 효과성을 검증하고 마케팅 비용 집행 시 고려되어야 할 전략적 방향을 제시하고자 한다.

제 1 절 연구의 목적과 의의

고객 이탈을 예측하는 것은 고객 유지에 중요하며 많은 산업에서 손실을 막기 위해 필수가 되었다. 이에 선행연구들에서는 고객 이탈 예측과 관련하여 여러 모형이 연구되어 왔지만, 구체적으로 고객 리텐션 과정에 관심을 기울이고 이에 초점을 맞춰 연구하는 경우가 많지 않았다.

따라서 실제 마케팅 비용 기획/시행 관련하여 경험하는 문제들을 해결하기 위해서 실질적이고 효율적인 고객 리텐션 방안에 바탕을 두고 이탈 문제를 분석하고, 분석한 결과를 토대로 구체적인 최적 안을 모색하는 접근이 필요하다고 할 수 있다. 단순히 고객 이탈률이 높은 고객 우선순위로 리텐션 프로모션을 할 것인지 등에 대해 실증 작업을 거치며, 그 과정에서 최적의 마케팅 비용 지출 방안을 파악함으로써 기업들이 겪고 있는 마케팅 비용 증가 문제를 해결할 방안을 효과적으로 모색할 수 있기 때문이다.

본 연구에서는 고객 리텐션 과정에서 고객들에게 마케팅 비용을 얼마나 어떻게 분배해야 효과적으로 이탈을 방지할 수 있는가 문제를 해결할 방안을 모색하는 데 초점을 맞추고자 한다. 고객 이탈과 관련하여 마케팅 비용 지급 최적화가 보다 중요한 이유는 이탈률을 미리 추정한 이후에는 정책적으로 효율적인 선제 대응 노력이 필요하기 때문이다. 또한 표 1-1과 같이 통신사의 마케팅 비용이 영업이익의 2배 이상에 이르는 상황[4]에서 기업 스스로가 마케팅 비용 지출 의사결정과정에서 합리적인 선택을 통해 기업의 수익 감소에 대처할 수 있는 자구책을 마련해야 하는 상황이기도 하다.

표 1-1. 통신3사의 마케팅 비용과 영업이익(2017년) (단위 :억원)

통신사	마케팅비용(a)	영업이익(b)	(a)/(b)
SKT	31,190	16,977	1.84
KT	26,841	9,521	2.82
LGU+	21,474	8,437	2.55
합계	79,505	34,935	2.28

제 2 절 연구 내용

본 연구에서는 고객 이탈을 예측하고 예측 결과를 바탕으로 고객 리텐션 최적화 방안을 찾기 위하여 다음과 같은 연구가 이루어졌다.

첫째, 데이터 세트(IBM Watson Analytics의 샘플 데이터)을 분석하고 마이크로소프트 Azure Analytics를 이용하여 고객 이탈 예측 모델링을 수행하였다. 이와 관련하여 Logistic Regression, SVM, Decision Forest, Neural Network, Decision Jungle 모델을 사용하였고 평가지표로서 Accuracy, Precision, Recall, F1 score, ROC(AUC) 등을 사용하여 비교하였다. 이를 바탕으로 가장 평가 점수가 높은 모델링 기법을 적용하여 데이터 세트 내 모든 고객의 이탈률을 추정하였다. 또한 고객 리텐션 최적화 즉 기대수익(회사 이익)과 관련된 기대수익 이론적 모델을 정립하였다.

둘째, 고객 이탈률과 이론적 모델을 기반으로 최적화 알고리즘을 구현하고 시뮬레이션을 수행 및 실증하였다. 특히 실질적으로 마케팅 예산이 한도가 있음을 고려하여 1억에서 최대 100억으로 세팅하여 금액에 갭을 두고 변화를 비교하였고, 고객 샘플 수도 변화를 주면서 시뮬레이션을 진행하였다. 위 과정은 모두 R 프로그래밍 언어를 활용하여 시뮬레이션 환경을 구축하였다.

셋째, 최적화 외 다른 마케팅 비용 지출 방안들을 가정하고 시뮬레이션 수행을 통하여 효과성을 검증하였다. 실제 기업에서 마케팅 비용을 지출하는 데 있어서 한정된 예산에 예산 배분과 집행 방식에 따라 기대수익이 천차만별로 달라질 것이다. 이러한 맥락에서 실제 지급 방식을 달리하여 시뮬레이션 환경을 구성하고 비교 분석하였다.

제 3 절 연구 보고서 구성

본 연구보고서는 총 5개 장으로 구성되어 있으며, 각 장은 다음과 같은 내용을 포함하고 있다.

1장에서는 본 연구를 수행하게 된 목적과 의의를 설명하고, 연구 내용에 대해 간략하게 설명한다.

2장에서는 본 연구를 이해하는데 필요한 배경 이론을 소개하고, 주요 관련 연구를 요약하여 본 연구의 차별성을 확인한다.

3장에서는 데이터에 대한 분석 및 예측 모델링을 수행하고, 최적화 방안에 활용될 기대 수익에 대한 이론적 모델을 정립한다.

4장에서는 이론적 모델을 바탕으로 최적화 알고리즘 설명 및 시뮬레이션 개발 내역을 소개한다. 그리고 최적화 외 비교할 수 있는 시뮬레이션 환경을 구성하여 결과를 분석하고, 최적화 방법과 비교 수행한다.

5장에서는 본 연구의 결과 및 성과에 대해 설명하고, 보완이 필요한 사항을 적시함으로써 실무 적용 시 참고 사항과 향후 연구의 방향성을 제시한다.

제 2 장 배경 이론 및 관련 연구

제 1 절 배경 이론

1. Measuring model performance

모델의 성능을 평가하기 위해서 본 연구에서는 이진 분류(class 0: 이탈, class 1: 잔존)의 성능 측정에 대해 집중하고자 한다. 고객 이탈 문제에 있어 분류의 목적은 Class 0과 1 즉, 이탈할 고객과 잔존할 고객을 정확하게 분류하는 것이다. 이를 위해서 logistic regression 등과 같은 binary classifier를 데이터 셋에 적용하여 0과 1 사이의 범위 안에서 각각의 고객들이 이탈할 확률 값을 추정할 수 있다. 이때 cutoff 값 $t \in [0;1]$를 정하여 t 보다 높은 값은 이탈 고객, t 보다 낮은 값은 잔존할 고객으로 분류한다. 표 2-1이 위와 같은 분류를 나타내는 confusion matrix이다[5].

표 2-1. Confusion Matrix

		Predicted Class	
		Class 0	Class 1
Actual Class	Class 0	$\pi_0 F_0(t)N$ [TP]	$\pi_0(1-F_0(t))N$ [FN]
	Class 1	$\pi_1 F_1(t)N$ [FP]	$\pi_1(1-F_1(t))N$ [TN]

[ ]사이에 True, False 그리고 Positive, Negative 약어로 표시

여기서 N은 모집단 크기를 나타내고, π_0과 π_1는 class0 과 1의 사전 확률이다. $F_0(t)$와 $F_1(t)$는 두 클래스의 누적 분포 함수(cumulative distribution function)를 나타낸다. $\pi_0 F_0(t)N$ 는 true positive로 classifier가 고객이 이탈할 것이라고 예측하고 실제로도 이탈한 고객 수를 나타낸다. $\pi_1 F_1(t)N$ 는 false positive로 classifier가 고객이 이탈할 것이라고 예측했지만 실제로는 잔존한 고객 수이다. 반대로 $\pi_0(1-F_0(t))N$ 는 false negative로 classifier가 고객이 잔존할

것이라고 예측했지만 실제로는 이탈한 고객 수, $\pi_1(1-F_1(t))N$ 는 true negative로 예측과 실제 모두 잔존한 고객 수를 나타낸다. 이를 간단히 평가지표로 나타내면 (2.1~4)로 표현할 수 있다.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.1}$$

$$\text{Recall} = \text{TP Rate} = \frac{TP}{TP+FN} \tag{2.2}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.3}$$

$$\text{F1 Score} = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP+\frac{FN+FP}{2}} \tag{2.4}$$

Precision은 classifier가 positive로 예측한 것 중 실제 정답인 비율을 나타내며 Recall은 실제 true 인 것 중 classifier를 정답을 맞춘 비율이며 둘 간은 trade off 관계를 갖고 있다. Accuracy는 전체 경우의 수중에 정답으로 분류한 비율로 사전적 의미상 Precision과 유사하지만 수식적으로 구분할 필요가 있다. F1 score는 precision과 recall의 밸런스를 고려한 조화 평균 값을 나타낸다.

cutoff 포인트 t값으로부터 독립적으로 classifier의 성능을 평가하기 위해 ROC(Receiver Operating Characteristic) curve가 자주 사용된다[7]. ROC 그래프를 만들 때 전형적으로 true positive rate(sensitivity = recall = $F_0(t)$)와 false positive rate(1−specificity= $F_1(t)$) 사용한다. 아래 그림 2−1은 ROC 그래프를 간략하게 표현하였다. ROC 공간에서는 왼쪽이나 위쪽에 있을수록 더 뛰어난 성능을 갖는다. 그리고 Curve가 대각선에 가까워질수록

무작위로 예측하는 경우와 다름 없기 때문에 classifier 성능이 좋지 않다. ROC curve의 하위 영역 AUC(Area Under the roc Curve)는 curve아래쪽 영역(면적)으로 0에서1사이의 값을 갖는다. ROC curve가 제공하는 정보는 많지만 성능을 단 하나의 수치로 요약한다거나 할 때에는 AUC가 유용하게 쓰인다.

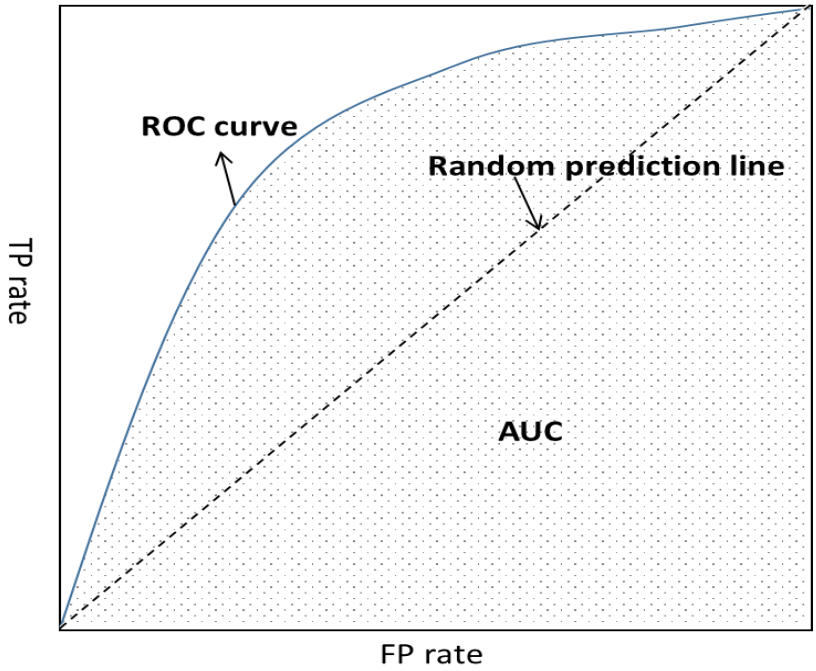


그림 2-1. ROC 그래프

2. 머신러닝 모델링 기법

머신러닝 모델링 기법 중 본 연구에서 적용할 Logistic Regression, SVM, Decision Forest, Neural Network 에 대해 설명하고자 한다.

첫째. Logistic Regression은 linear regression으로 fitting하기 어려운 문제를 해결할 수 있는 방법으로 어떤 class나 범주에 속할 확률을 추정하는 것이다. 구체적으로 말하면 선형 함수 f(x) 로 변수 x 가 어떤 class에 속할 사건을 Y라고 가정하면, Odds는 Y가 일어날 가능성 대 Y가 일어나지 않을 가능성에 대한 비율을 말한다. 하지만 Odds값은 0부터 ∞까지 밖에 표현 할 수 없기 때문에 Odds에 로그를 취해 −∞부터 ∞까지 표현할 수 있게 된다. 이것을 logit 변환이라고

부른다. 예를 들어 특징 벡터 x 로 표현한 고객이 계약기간 만료 전후로 이탈 할 것인지에 대한 logit 변환(natural log $of\ Odds$)을 통해 이탈 확률을 추정할 수 있다. 간단한 logistic 모델은 수학적으로 아래 (2.5)와 같이 표현 할 수 있다[8].

$$logit(Y) = \ln(\mathrm{Odds}) = \ln\left(\frac{p}{1-p}\right) = \mathrm{f(x)} = \alpha + \beta\,\mathrm{x} \qquad (2.5)$$

p 는 특징 벡터 x 로 표현한 데이터 객체가 어떤 사건이 발생할 확률을 말한다. 예를 들어 p 는 이탈 class에 속할 확률 추정치를 나타내고 $1-p$ 는 잔존 class에 속할 확률 추정치를 나타낸다. 위 (2.5)식으로부터 logistic function을 다음과 같은 식으로 표현 할 수 있다.

$$logistic\ \ funtion = Probability(Y|x) = p = \frac{e^{\alpha+\ \beta\mathrm{x}}}{1+e^{\alpha+\ \beta\mathrm{x}}} \qquad (2.6)$$

α는 f(x)의 intercept이고 β는 regression coefficient를 나타낸다. 간단한 logistic 모델에 다중 독립변수를 넣어 (2.7)과 같이 확장하여 표현 할 수 있다.

$$logit(Y) = \ln\left(\frac{p}{1-p}\right) = \mathrm{f(x)} = \alpha + \beta_1\mathrm{x}_1 + \beta_2\mathrm{x}_2$$

$$THEN\ \ Probability(Y|x_1, x_2) = p = \frac{e^{\alpha+\ \beta_1\mathrm{x}_1+\beta_2\mathrm{x}_2}}{1+e^{\alpha+\ \beta_1\mathrm{x}_1+\beta_2\mathrm{x}_2}} \qquad (2.7)$$

마찬가지로 α는 f(x)의 intercept이고 β_i는 regression coefficient, x_i 는 독립변수들을 나타낸다. 그리고 α 와 β_i 값은 전형적으로 ML(Maximum Likelihood) 방법을 사용하여 구하게 된다.

둘째, SVM(Support Vector Machine)은 데이터가 어느 class에 속할 것인지 판단하는 비 확률적 binary linear classification 모델을 만들게 된다. SVM은 두 class를 선으로 분할 하는 대신 두꺼운 막대로

분할한다. 이 개념은 그림 2-2에서 두개의 평행한 점선으로 표현할 수 있다.

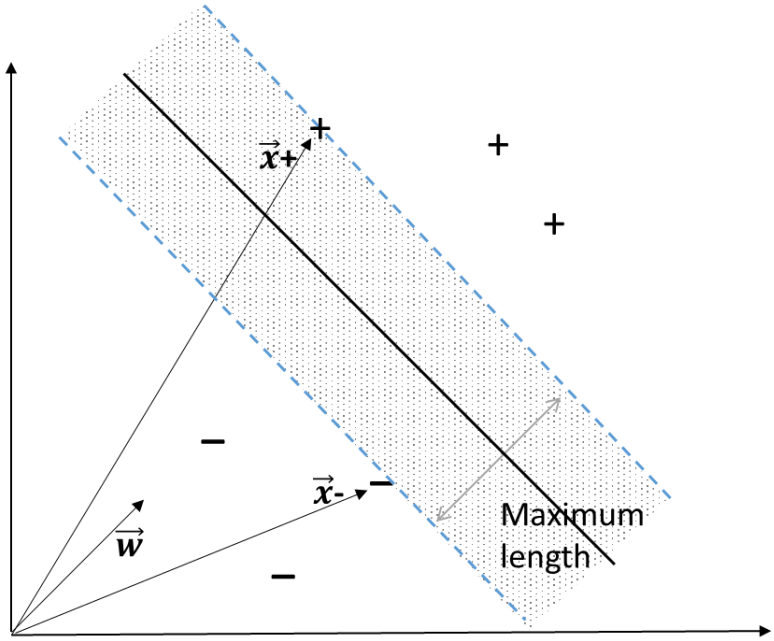


그림 2-2. SVM의 그래프

$\vec{w}$는 평행한 점선 폭의 중심선에 직교하는 벡터이고 $\vec{x}+$, $\vec{x}-$와 같이 경계선을 정하는 샘플들을 support vector라고 한다. SVM에서는 경계선의 두께가 두꺼울수록 더 좋기 때문에 이 폭을 최대로 만드는 것이 목적이다. 폭을 최대로 만드는게 좋은 이유는 다음과 같다. 예측 모델을 만들 때 훈련 데이터를 사용하여 위와 같은 support vector를 찾더라도 실제 데이터를 적용하게 되면 훈련 데이터와 다르게는 일부 데이터는 기존 양성, 음성으로 판별된 데이터보다 경계에 더 가까울 수도 있고 경계 안으로 들어갈 가능성도 있다. 심지어 폭이 좁다면 일부 데이터는 막대의 정반대쪽으로 들어가 classification의 오류가 발생하는 경우도 생길 수 있다. 따라서 경계의 폭이 넓을수록 경계에 가까이 있는 데이터들도 분류할 수 있는 여지가 훨씬 많아지게 된다. $\vec{w}$와 $\vec{x}_i$벡터 값은 아래 (2.8~2.9)수식으로 단순하게 표현할 수 있다[9].

$$\vec{\boldsymbol{w}} = \sum_{\boldsymbol{i}} \boldsymbol{\alpha_i y_i x_i}$$

$$for\ maximizing\ \alpha_i(Lagrange\ multiplier) \quad (2.8)$$

$$\boldsymbol{y_i(\overrightarrow{w} \cdot x_i + b) = 1}\ \ for\ \ x_i \in \textit{경계선}$$

$$with\ \ y_i = 1\ \ for\ \ +,\ \ y_i = -1\ \ for\ \ + \quad (2.9)$$

참고로 경계선 샘플들이 linearly separable한 경우라면 SVM이 매우 잘 동작하지만 non-linearly separable한 경우에는 적용 하기 어렵다. 이때 SVM이 잘 작동할 수 있도록 linearly separable한 공간으로 샘플들을 보내준 후 그 공간에서의 SVM을 적용하는 것을 커널 트릭이라고 한다.

셋째, Decision Forest(Random Forest)는 의사 결정 규칙과 그 결과들을 트리 구조로 도식화 하는데 트리 구조는 일련의 간단한 규칙을 적용하여 생성된 데이터의 분할을 나타내고 수많은 의사 결정 트리가 모여서 다수결 투표로 최종 의사결정을 내리는 방식이다. 우선 의사결정 트리에 대해 살펴보면 다른 모델링 기법에 비해 해석 가능한 규칙이나 논리 문장을 나타낼 수 있는 모델을 생산한다는 것이다. 게다가, 분류는 복잡한 계산 없이 수행될 수 있고 이 기법은 연속적 변수와 범주적 변수 모두에 사용될 수 있다[11]. 더욱이 의사결정 트리 모형 결과는 예측이나 분류에 중요한 요인의 중요성에 대한 명확한 정보를 제공한다. 그러나 일반적으로 비선형 데이터 및 노이즈가 많은 데이터에 취약하므로 이 기법은 범주형 결과를 예측하는 데 더 적합하며, 가시적 추세와 순차적 패턴을 이용할 수 없는 한 시계열 데이터에 적용하기엔 적합하지 않다[11].

또한 트리 하나로는 의사결정 과정에 hole 이 있을 수 있으니 그림 2-3과 같이 이런 트리를 여러 개 그려서 그 hole을 최대한 줄인 것이

Decision Forest이다. 배깅(bagging)[①]을 이용하여 T개의 결정트리들로 구성된 랜덤 포레스트를 학습하는 과정으로 S_0은 전체 학습 데이터 집합을 의미한다. S_0^t는 t 번째 결정 트리를 위해 배깅을 통해 임의로 선택된 학습 데이터들로, S_0의 부분집합이다.

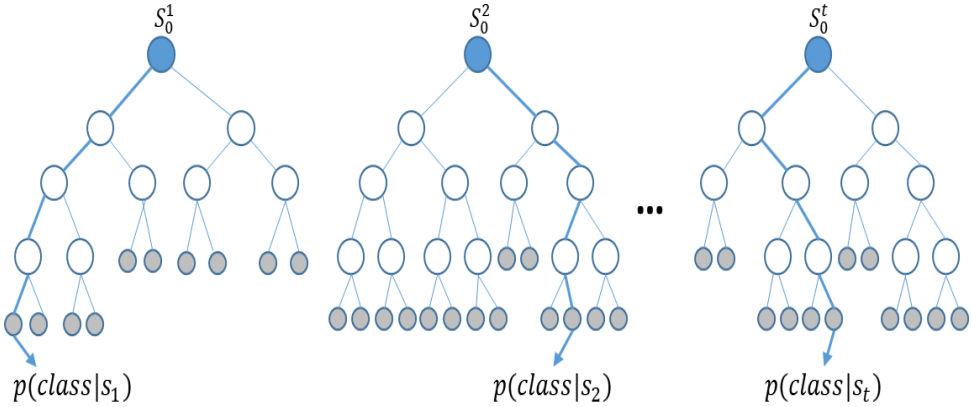


그림 2-3. Decision Forest 그래프

최종 예측 결과는 모든 트리의 예측 결과들의 평균으로 얻는다. 이를 수식으로 표현하면 (2.10)과 같다.

$$Decision\ Forest\ \ 예측\ 추정치 = \frac{1}{T}\sum_{t=1}^{T} p(class|s_t) \qquad (2.10)$$

Decision Forest의 가장 큰 특징은 랜덤성(randomness)에 의해 트리들이 서로 조금씩 다른 특성을 갖는다는 점이다. 이 특성은 각 트리들의 예측들이 비상관화(decorrelation) 되게 하며, 결과적으로 일반화 성능을 향상시키며 랜덤화는 노이즈가 포함된 데이터에 대해서도 강인하게 만들어 준다. 랜덤화는 각 트리들의 훈련 과정에서 진행되며,

① 배깅(bagging)은 bootstrap aggregating의 약자로, 부트스트랩(bootstrap)을 통해 조금씩 다른 훈련 데이터에 대해 훈련된 기초 분류기(base learner)들을 결합(aggregating)시키는 방법이다. 부트스트랩이란, 주어진 훈련 데이터에서 중복을 허용하여 원 데이터셋과 같은 크기의 데이터셋을 만드는 과정을 말한다

랜덤 학습 데이터 추출 방법을 이용한 앙상블 학습법인 배깅과 랜덤 노드 최적화(randomized node optimization)가 자주 사용된다. 이 두 가지 방법은 서로 동시에 사용되어 랜덤화 특성을 더욱 증진 시킬 수 있다[12].

하지만 Decision Forest는 기계 학습에 풍부한 역사를 가지고 있고 특히 컴퓨터 비전에 있어서 성공적이나, 데이터가 늘어난다면 노드 수는 깊이와 함께 기하급수적으로 증가할 것이다. 예를 들어 모바일 프로세서나 임베디드 프로세서의 특정 애플리케이션의 경우 메모리 자원이 제한됐기 때문에 트리의 기하급수적인 성장은 그 깊이와 정확도를 제한할 수 있다[13].

이에 반해 Decision Jungle은 DAG(Decision Directed Acyclic Graphs) ② 개념을 도입하여 모든 노드에 하나의 경로만 허용하는 기존의 의사결정 트리와 달리, 의사결정 정글의 DAG는 루트에서 각 노드에 이르는 여러 경로를 허용함으로써 Decision Forest대비 generalization을 개선하는 동시에 메모리를 상당히 적게 필요로 한다는 것을 보여준다[13].

넷째, Neural Network는 인지체계 및 뇌의 신경기능에서 학습하는 과정을 본뜬 분석 기법이며, 기존 데이터에서 이른바 학습 과정을 수행한 후 다른 관찰(동일한 변수 또는 다른 변수에 대한)에서 새로운 관찰(특정 변수에 대한)을 예측할 수 있다. 그림 2-4는 신경망의 흐름을 보여준다[11]. 신경망을 훈련시키는 것은 각 단위의 입력에 가장 적절한 가중치 (W_i) 를 설정하는 과정이며, feedforward 네트워크(Input에서 Output 방향)의 에러 gradient를 보정을 통해 가중치를 재계산하는 과정을 backpropagation(Output에서 Input 방향)이라 한다.

② DAG(Directed Acyclic Graph)는 비순환 그래프로, DAG알고리즘에서는 순환하는 싸이클이 존재하지 않고, 일 방향성만 가진다

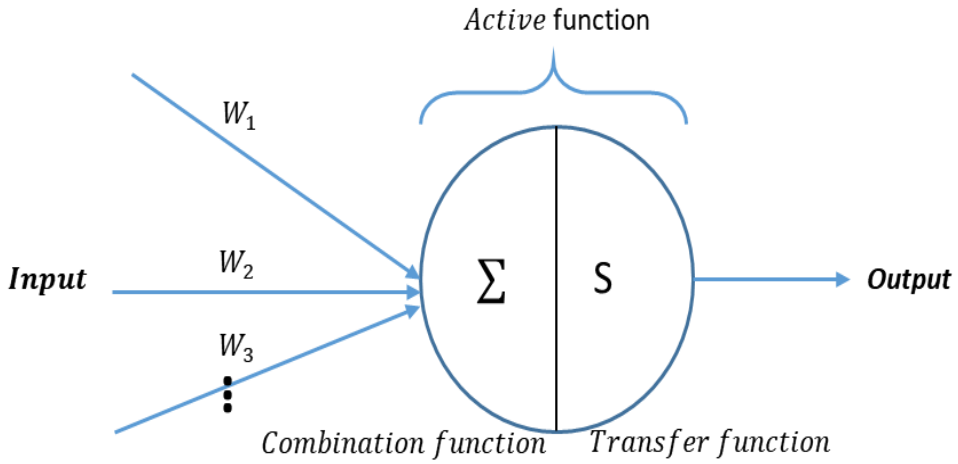


그림 2-4. Neural Network의 기본 unit

신경망은 기능적 형태가 비선형일 때 응용 분야에서 잘 수행된다. 그것들은 특히 입력과 출력 사이의 관계에 대한 수학 공식과 사전 지식이 알려지지 않은 문제를 예측하는 데 유용하다.

회귀 분석에 신경망을 사용할 때 단점은 계수 추정치의 유의성 검사에 p값을 제공하지 않는다는 것이다. 또한 학습 전 feature selection의 예비 단계가 필요하다. Hidden layer가 있는 인공신경망은 비선형 의사결정 Hyper-surfaces과 관련된 문제에 대한 classification는 우수하지만 해석하기는 훨씬 어렵다.

제 2 절 관련 연구

고객 이탈 예측 모델은 일반적으로 다양한 데이터 가운데 특징적인 변수들을 선택하여 AUC와 같은 통계적 기반 성능 측정을 사용하여 평가된다. 이는 그림 2-5와 같이 최적의 모델 선택과 가능한 많은 고객을 유지하기 위한 리텐션 마케팅으로 이어져야 한다. 물론 기업 측면에서 많은 고객을 유지하는 게 중요한 것이 아니라 고객의 가치 등을 고려하여 최대한 많은 수익을 낼 수 있도록 마케팅 비용을 지출하는 전략 수립이 가장 중요할 것이다.

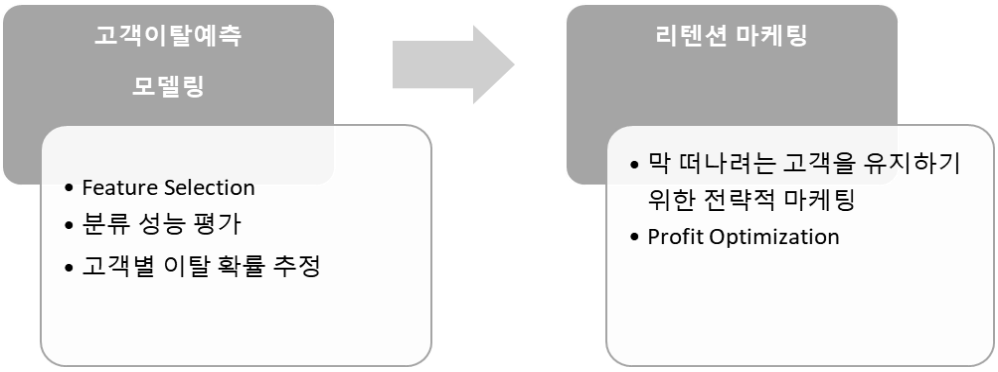


그림 2-5. 고객이탈 예측 및 리텐션 마케팅 Work Flow

따라서 고객의 가치를 생각해야 한다. CLV(Customer Lifetime Value)는 몇 년 동안 인기있는 연구 주제였다. 고객 한사람이 죽을 때까지 자사의 상품만을 구매한다고 했을 때의 매출액 혹은 이익을 의미, 즉 한 명의 고객이 고객으로 존재하는 기간 동안 만들어 내는 이익의 총합계로 단 한 번의 거래에 초점을 맞추는 것이 아니라 고객으로서 기업에 기여할 이익의 총계를 고려한 것이다. 이는 고객과 조직의 관계에 기인 한 모든 미래 현금 흐름의 현재 가치로 정의되며 가장 수익성 높은 고객을 식별하고 장기적으로 육성하기 위해 각 고객의 재무 가치를 평가할 수 있는 이점을 제공하며 수식적으로 (2.11)과 같이 표현 할 수 있다[15].

$$CLV = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+r)^t} - AC \qquad (2.11)$$

표 2-2. CLV 표기법

기호	내용
p_t	t시점의 고객 인당 매출
c_t	t시점의 고객 인당 비용
r_t	t시점의 할인율(이자율)
r	고객 유지 비율(retention rate), 즉 어떤 고객이 그 다음 해에도 여전히 고객으로 남아 있을 확률
AC	고객 획득 비용(Acquisition Cost). 고객이 첫 방문 또는 첫 구매를 하도록 하는데 드는 비용

그러나 다양한 유형의 고객 관계 및 거래 상황으로 CLV를 신중하게 모델링 해야 한다. Hyunseok Hwang 등은[16] 대부분이 과거 이익 기여에서 파생된 미래현금흐름에만 초점을 맞췄는데 경쟁이 치열하면 고객 이탈이 빈번해지고 수익성에 많은 영향을 미치기 때문에 CLV를 계산하기 위해 미래현금흐름만을 고려하지 않고 과거의 기여도, 잠재적 가치, 이탈 확률을 동시에 고려한 CLV 모델을 제안했다.

Verbeke 등은[17] 전 세계통신사업자 11개의 데이터 세트에 다양한 분류 알고리즘을 적용하여 수익 중심 측정과 통계적 성과 중심 측정 모두 평가하여 벤치마킹 실험을 한 결과 다음과 같은 결과를 도출하였다. 첫째 리텐션 고객 대상의 비율을 어떻게 최적으로 분류하느냐에 따라 창출되는 이익에 큰 영향을 미칠 수 있고 둘째 고객이탈 예측 모델의 성능에 대한 과잉 샘플링은 데이터 세트와 분류 기법에 따라 크게 달라질 수 있다. 셋째 분류 기법들 중 어느 기법이 크게 성능이 뛰어나거나 하지 않았지만 대체로 decision tree 기법이 전체적인 성능을 잘 발휘했다. 넷째 좋은 성능을 얻기 위해서 입력 변수의 선택이 중요하며 일반적으로 6~8개의 변수로도 높은 정확도를 예측하기 충분하다. Kirui 등은[14] 트래픽 수치와 고객 프로파일 데이터로부터 도출된 계약 관련, 통화 패턴 설명 및 통화 패턴 변경 설명 특징으로 분류되는 새로운 입력 변수를 제시했다. 이 특성은 두 가지 확률론적 데이터 마이닝 알고리즘인 Naive Bayes와 Bayesian

Network를 사용하여 평가되었으며, 많은 분류 및 예측 작업에서 널리 사용되는 알고리즘인 C4.5 decision tree를 사용하여 얻은 결과와 비교하였다. Bock 등은[18] 온라인 도박 산업에서 도박꾼들을 효율적으로 관리하기 위해 고객 이탈 예측을 시험하였고 앙상블의 성능이 단일 모델보다 더욱 우수한 것으로 평가했다. Coussement 등은[19] 고객 이탈 예측 성능에 데이터 준비에 따라 AUC 기준 14.5%까지 향상하는 것을 실험하였고 최적화된 logistic regression은 다른 알고리즘과 비교해도 손색이 없다고 판단하였다. Verbraken 등은[20] 다수의 베이지안 네트워크 알고리즘을 적용하여 고객 이탈 예측 성능을 조사하였다. 성능 평가 방법은 AUC와 최대 이익 기준으로 평가하였는데 여기서 최대 이익은 고객 수, 리텐션 마케팅을 통해 잔존할 비율(전환율)과 CLV, 리텐션 비용 그리고 전체 고객 중 이탈할 것으로 예측한 비율과 리텐션 마케팅 대상이 되는 비율 등을 토대로 결정된다. 그러나 모델 별로 최대 이익을 결정 짓는 파라미터는 리텐션 마케팅 대상이 되는 비율에 따라 결정될 뿐 실질적으로 각각의 고객에게 지급되는 리텐션 비용과 전환율 그리고 그에 따른 최대 이익이 어떻게 변하는가에 대한 상세한 방법은 확인할 수가 없기에, 해당 연구가 제안하는 최대 이익 기준을 활용하여 리텐션 마케팅을 수행하는 것은 현실적으로 어렵다고 할 수 있다. Ballings 등은[21] 고객 이탈 예측 연구의 대부분이 알고리즘 개선을 통한 모델 개선을 목표로 하고 있지만, 이 연구는 예측 성능에 대한 데이터 time window 최적화에 초점을 맞추고 있다. 고객 event log를 1년에서 16년으로 연장함으로써 모델 성능의 향상을 분석하였는데 그 결과 5번째 연도가 지나면, 예측 성능이 약간만 증가한다는 것을 보여주는데, 이는 본 연구의 회사가 예측 성능의 거의 감소 없이 데이터의 69%를 폐기할 수 있다는 것을 의미하며 데이터 관련 부담을 상당히 줄일 수 있다는 것이다. Eunjo Lee 등은[22] 게임 데이터 로그를 데이터 세트로 활용하여 고객이탈 예측을 하였다. 전체 고객과

오랫동안 이용한 충성고객을 나누어 모델에 따라 예측 성능을 비교한 결과 전체 고객을 대상으로 예측했을 때 성능이 더 나은 결과를 보였다. 하지만 리텐션 타겟팅을 했을 때 전환율에 따른 기대 수익 측면에서는 충성 고객의 기대수익이 훨씬 높은 결과를 나타냈다.

관련 연구 사례들을 요약하면, 우선 고객의 가치에 대해 정의하고 설계하는 다양한 연구들이 이루어졌다. 이러한 연구들은 기업이 최대 이익을 얻기 위해 필요한 고객의 가치에 대한 개념을 정립해주었다. 다음으로 고객 이탈 예측 모델링에 대한 다양한 연구 사례들이 있는데, 다수의 연구들은 알고리즘 개선을 통한 모델 개선 방법론에 초점을 맞추고 있다. 최근에는 이탈 예측과 고객의 가치를 반영하여 리텐션 마케팅을 했을 때 수익 관점에서 최적의 모델링을 다양한 데이터 셋에 적용하여 벤치마크하고 있다. 그러나 이러한 연구들은 수익과 비용 측면에 있어 얼마의 마케팅 비용을 어느 고객에게 얼마만큼 지불했을 때 얼마만큼의 수익을 기대할 수 있는지 실제로 고객 이탈을 예측하고 리텐션 마케팅을 적용하는데 있어서 한계가 있다.

본 연구는 여러 예측 모델 가운데 본 연구에 쓰이는 데이터 세트에 가장 적합한 모델링을 선정하고, 해당 모델링을 적용하여 각각의 고객들의 이탈률을 추정하고자 한다. 이를 바탕으로 최적의 기대수익을 얻을 수 있는 알고리즘을 제안하여 실제 리텐션 마케팅 정책을 입안, 시행하는데 있어 실질적인 방향성을 제시하고자 한다.

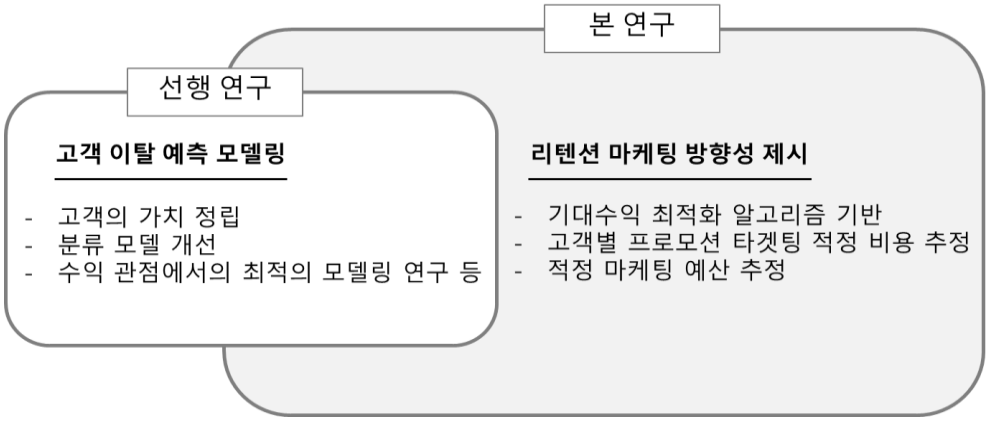


그림 2-6. 선행 연구와의 차별성

제 3 장 데이터 분석 및 기대수익

본 장에서는 탐색적 데이터 분석과 데이터에 대한 여러 모델링 기법을 적용하여 비교하고자 한다. 또한 모델링 결과로 나온 확률 추정치를 바탕으로 기대수익을 최적화 하기 위한 모형을 제안하고자 한다. 그리고 제안 모형을 구현하고 효과성을 검증하는데 필요한 이론적 모델을 정립하도록 하겠다.

제 1 절 데이터 분석 및 예측 모델링

IBM Watson Analytics, Guide to Sample Data Sets 의 고객지원 파일 내 7,043명의 이력 데이터를 데이터 세트로 다루며, 고객 데이터는 표 3-1에 나열된 변수로 기술된다.

표 3-1. 고객이탈 문제에 사용할 속성

변수	설명
customerID	고객ID
Gender	성별
SeniorCitizen	고령자 여부
Partner	결혼 여부 (Yes, No)
Dependents	부양가족 여부 (Yes, No)
tenure	계약유지기간
PhoneService	전화 가입여부(Yes, No)
MultipleLines	폰 멀티 라인
InternetService	인터넷 종류
OnlineSecurity	인터넷 온라인 보안
OnlineBackup	인터넷온라인백업
InternetService	인터넷 종류
DeviceProtection	디바이스보험 가입여부
TechSupport	기술지원 여부
StreamingTV	스트리밍TV가입여부
StreamingMovies	스트리밍무비가입여부
Contract	계약 형태
PaperlessBilling	청구 형태(Yes, No)
PaymentMethod	지불방법
MonthlyCharges	월 청구액
TotalCharges	전체 청구액
Churn	**이탈 여부 <- 타겟 변수**

위 변수들 중 Churn은 머신러닝 관점에서 보면 예측값과 비교되는 타겟 변수이고 그 외 나머지 변수들은 Churn을 예측하기 위한 입력 변수 역할을 하게 된다. 이 입력 변수들 가운데 수치형 변수는 standardization, 범주형 변수는 dummy 처리 등의 전처리 과정 거친 후 효율적인 모델링 분석을 위해 그림 3-1과 같이 Microsoft Azure Machine Learning Studio의 Logistic Regression, Decision Forest, Decision Jungle, SVM, Neural Network 알고리즘 모듈을 활용하여 학습을 진행하였는데 데이터 세트를 훈련용 80%, 나머지 20%는 테스트용으로 분할 진행하였다.

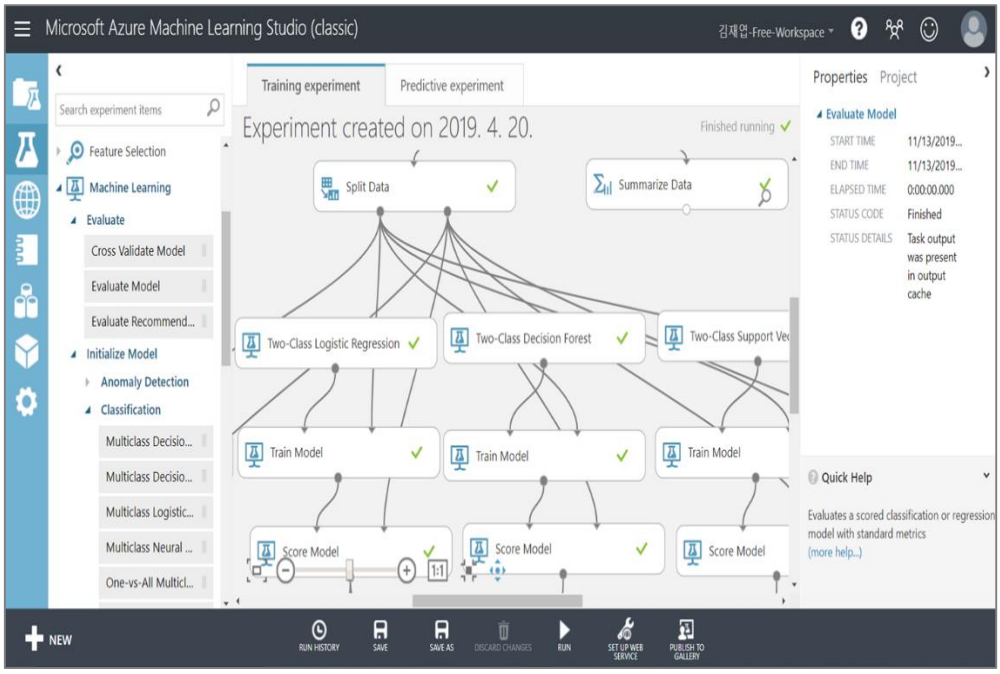


그림 3-1. Microsoft Azure Machine Learning Studio Work화면

상기 5가지 모형에 대한 학습 결과, 표 3-2와 같이 Logistic Regression이 accuracy 평가에선 80.2%, Decision Jungle이 precision 평가에서 81.3%로 가장 높은 결과가 나왔다.

표 3-2. 모델별 성능 평가 결과표

모델	Accuracy	Precision	Recall	F1 score
Logistic	0.802	0.704	0.485	0.574
SVM	0.784	0.651	0.466	0.544
Decision Forest	0.774	0.733	0.284	0.409
Neural Network	0.789	0.678	0.451	0.542
Decision Jungle	0.738	0.813	0.067	0.124

하지만 해당 평가의 수치는 cutoff 값에 따라 달라지기 때문에 어느 모델이 우수한지 해당 수치만으로는 판단하기 어렵다. 이런 경우 cutoff값에 독립적인 ROC 그래프 및 AUC 수치를 통해 불확실성을 수용하여 적합한 모델을 선정할 수 있다. 그림 3-2와 같이 Logistic 모델과 그 외 다른 모델들의 ROC 그래프를 비교하였다

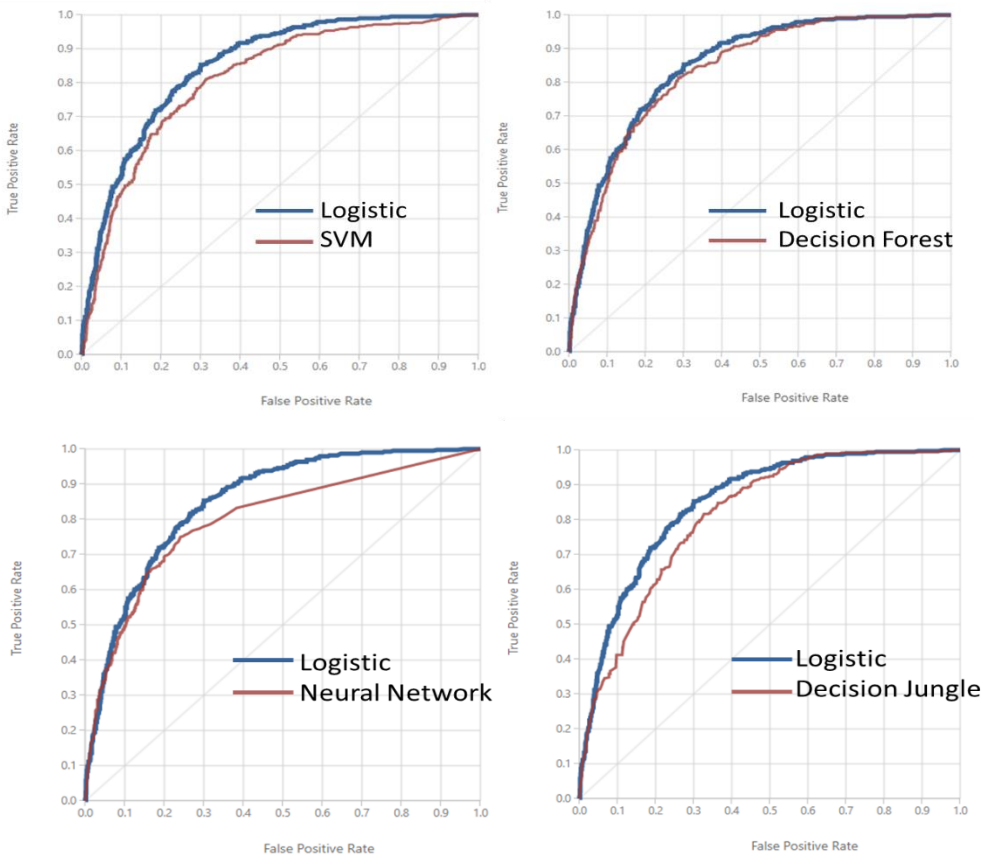


그림 3-2. 모델별 ROC 그래프 비교

ROC 그래프를 보면 Logistic 모델의 곡선이 전반적으로 다른 4개의 모델 곡선보다 위에 있는 것을 확인 할 수 있다. 이는 Logistic 모델의 True Positive Rate 즉 실제 이탈한 고객 중 모델이 제대로 예측한 적중률이 다른 모델보다 높다는 뜻이다. 또한 그림 3-3과 같이 ROC 곡선의 하위 영역인 AUC 값을 비교해 보면 Logistic 모델이 1에 가장 가까운 0.851로 classification 성능이 가장 우수한 것으로 판단되었다.

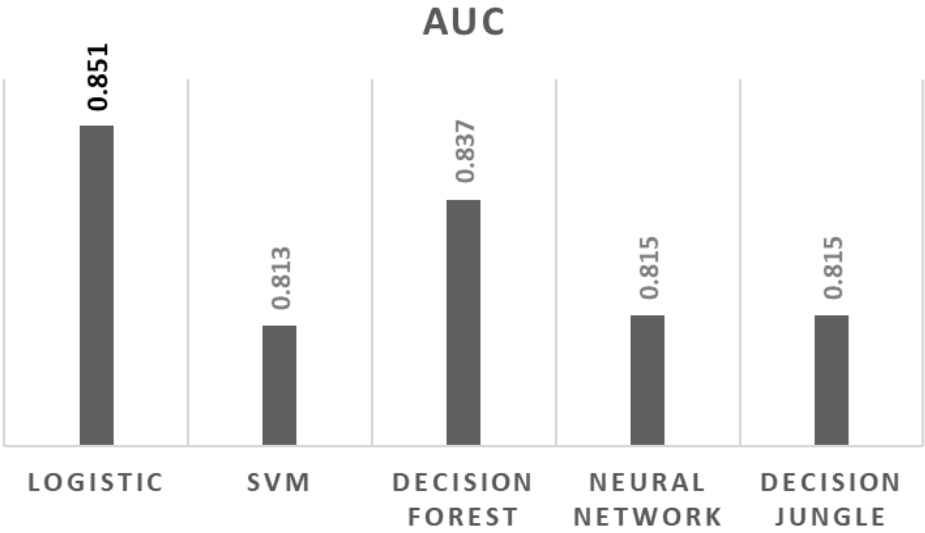


그림 3-3. 모델별 AUC값 비교 그래프

따라서 본 연구에서는 Logistic Regression을 최종 모형으로 선정하였고 예측 성능 향상과 과적합 방지를 위해서 통계 분석 도구(R)의 MASS 패키지 내 stepAIC()함수를 사용해서 최종 모형에 들어갈 독립변수를 선택하였다. 그 결과 그림 3-4와 같이 선택된 변수(dummy변수 포함)들 모두 p-value가 0.05 이하로 유의미한 결과를 나타냈다. 하지만 해당 모형으로 성능을 측정했을 때 accuracy는 80.4%로 이전 값인 80.2% 과 비교하여 의미 있는 차이는 보이지 않았다.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8763  -0.6806  -0.2764   0.7371   3.3770

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.11146    0.20745 -10.178  < 2e-16 ***
tenure                        -1.42780    0.17978  -7.942 1.99e-15 ***
MonthlyCharges                -0.68967    0.21010  -3.283 0.001029 **
TotalCharges                   0.70358    0.18966   3.710 0.000207 ***
Dependents_Yes                -0.21816    0.09662  -2.258 0.023947 *
MultipleLines_Yes              0.29510    0.10503   2.810 0.004962 **
InternetService_Fiber.optic    1.26113    0.23100   5.459 4.78e-08 ***
InternetService_No            -1.35652    0.21297  -6.369 1.90e-10 ***
OnlineSecurity_Yes            -0.31815    0.10809  -2.943 0.003246 **
TechSupport_Yes               -0.32660    0.11071  -2.950 0.003179 **
StreamingTV_Yes                0.45932    0.11790   3.896 9.78e-05 ***
StreamingMovies_Yes            0.45767    0.11560   3.959 7.53e-05 ***
Contract_One.year             -0.79791    0.13269  -6.013 1.82e-09 ***
Contract_Two.year             -1.35909    0.20917  -6.497 8.17e-11 ***
PaperlessBilling_Yes           0.40681    0.08943   4.549 5.40e-06 ***
PaymentMethod_Electronic.check 0.38493    0.08280   4.649 3.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

그림 3-4. Logistic Regression stepAIC 결과 summary

중요한 것은 본 연구가 제안하는 기대 수익 최적화 알고리즘을 적용하기 위해서는 각각의 고객들이 이탈 확률 추정치가 필요하다는 점이다. 즉 고객들의 이탈 확률 값이 어떻게 분포 되어 있는지 탐색할 필요가 있다. 이를 위해 변수 선택을 한 Logistic 모형을 데이터 세트의 전체 7,032명 고객에 적용하여 이탈 확률을 추정하였다.

그 결과 그림 3-5의 왼쪽 그래프 Normal Q-Q Plot[③]을 보면 이탈 확률 값은 정규분포의 성격을 띠고 있음을 알 수 있고 오른쪽 boxplot를 통해 이탈 확률의 중간 값은 19.8%이고 최고 약 83%까지 이탈 확률이 높은 고객이 있음을 확인 할 수 있다.

③ R을 이용하여 정규분포의 성격 여부를 판단하는 그래프로 확률 추정 값들이 빨간 대각 기준선을 따라 분포되어 있으면 데이터가 정규 분포의 성격을 띠고 있다고 판단한다

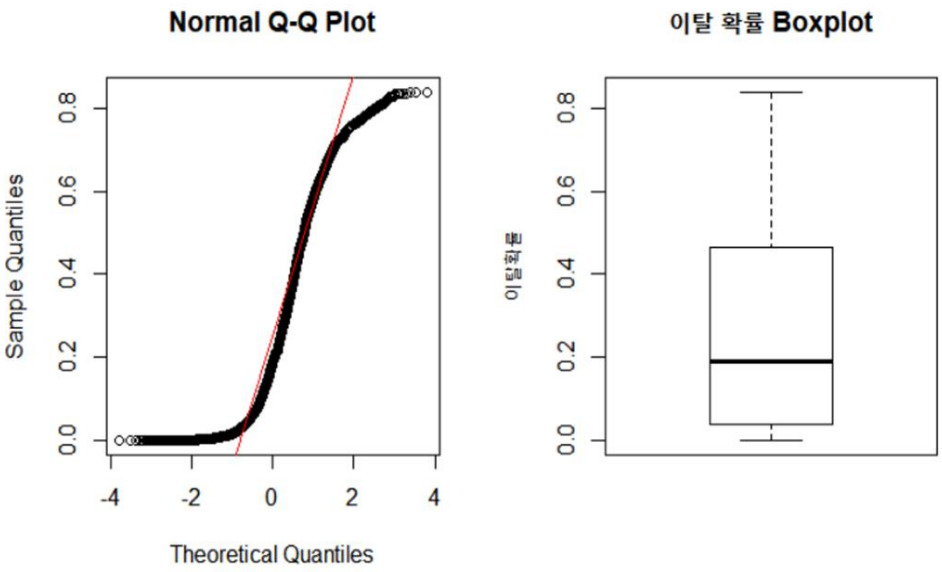


그림 3-5. 전체 고객 이탈 확률 추정치의 정규 분포 확인 및 boxplot

또한 그림 3-6의 히스토그램과 밀도 그래프를 보면 고객들의 이탈 확률이 약 0~20%에 가장 많이 밀집해 있는 것을 확인 할 수 있다.

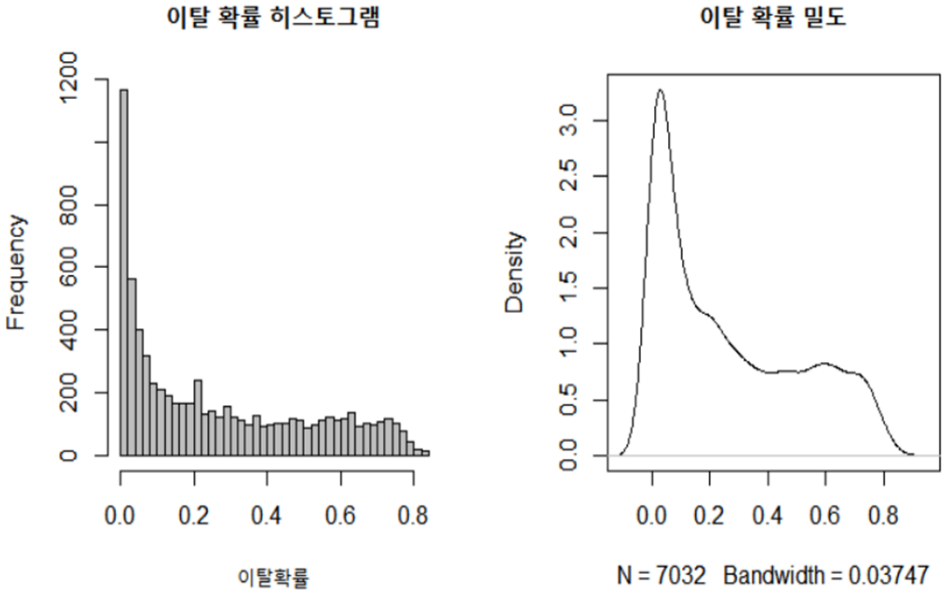


그림 3-6. 전체 고객 이탈 확률 추정치의 히스토그램

제 2 절 기대 수익에 대한 이론적 모델 정립

앞에서 우리는 Logistic regression 모형을 통해 고객들이 통신사를 이탈할 확률을 예측(추정)하였다. 일반적으로 계층을 예측한 후에는 예측 결과에 따라 행동으로 이어지는 경우가 많다. 그렇다고 해서 통신사를 바꿀 가능성이 큰 고객을 타겟팅 해서 프로모션을 제안해야만 하는 것은 아니다. 실제 비즈니스 관점에서 보면 많은 고객을 유지하는 것보다 수익이 줄어들지 않도록 하는 것이 더 바람직하기 때문이다. 따라서 통신사 입장에서는 수익이 줄어들지 않도록 하는 일에 중점을 둬야 한다. 프로모션을 통해 고객 x에게 리텐션 타겟팅 할 때 통신사에 남을 확률을 p(retain|x), 고객 x가 통신사에 주는 수익을 v(x)라 가정하자. 사례를 단순화하기 위해 고객이 이탈하는 경우의 가치는 0으로 가정하면 기대수익은 (3.1) 같이 표현 할 수 있다.

$$\textit{프로모션 타겟팅 기대수익} = p(retain|x) \cdot v(x) \qquad (3.1)$$

위 식을 보면 오히려 유지할 가능성이 높은 고객에게 프로모션 타겟팅을 해야 하는 이상한 결론에 이른다. 사실 타겟팅 하지 않았는데도 그 고객이 남아 있다면 0보다 큰 가치를 얻게 된다. 따라서 제대로 계산하려면 타겟팅 하지 않을 때의 수익을 추정해 타겟팅 할 때와의 수익과 비교해 이 고객을 타겟팅 할지 말지를 결정해야 한다. 타겟팅 여부에 따라 고객이 이탈할 확률도 달라질 것(즉 고객에게 주는 혜택이 효과가 있다)이라고 가정하고, 회사가 고객이 유지하든 떠나든 지출하는 비용을 c라고 하자. 타겟팅 할 때의 기대수익 EP_{target} 와 타겟팅 하지 않았을 때의 기대수익 $EP_{not\ target}$ 를 (3.2)와(3.3)로 표현 할 수 있다.

$$EP_{target} = \ p(retain|x, target) \cdot v(x) - \mathrm{c} \qquad (3.2)$$

$$EP_{not\ target} = \ p(retain|x, not\ \ target) \cdot v(x) - c \qquad (3.3)$$

타겟팅해 수익이 가장 많이 발생할 고객은 $EP_{target} - EP_{not\ target}$ 값이 가장 큰 고객이다. 이러한 내용을 바탕으로 통신사 기대수익을 최대화하는 이론적 모델을 표 3-3과 같이 요약할 수 있다.

표 3-3. 기대수익 최대화 모델

Decision variables : $p(\mathrm{r}|\mathrm{x}_i,\mathrm{t})$, $p(\mathrm{r}|\mathrm{x}_i,\mathrm{nt})$, $v(x_i)$, $c(x_i)$ $(\mathrm{i} = 1,2,3\ldots,\mathrm{N})$

Objective :

$$\mathrm{Max}\left[\ \sum_{i=1}^{N}\Big(\big(p(\mathrm{r}|\mathrm{x}_i,\mathrm{t}) - p(\mathrm{r}|\mathrm{x}_i,\mathrm{nt})\big)\cdot v(x_i)\Big)\ \right] - \sum_{i=1}^{N} c(x_i)$$

$$= \mathrm{Max}\left[\ \sum_{i=1}^{N}(\Delta \mathrm{p}_i \cdot v(x_i)\ \right] - \sum_{i=1}^{N} c(x_i)$$

Subject to : $0 < p(\mathrm{r}|\mathrm{x}_i,\mathrm{t}) < 1$

$0 < p(\mathrm{r}|\mathrm{x}_i,\mathrm{nt}) < 1$

$c(x_i) \ \geq 0$

$\sum_{i=1}^{N} c(x_i) = C$

표 3-4. 기대수익 최대화 모델 표기법

기호	내용
$p(\mathrm{r}\,\vert\,\mathrm{x}_i,\mathrm{t})$	리텐션 타겟팅 한 이후 고객이 잔존할 확률
$p(\mathrm{r}\,\vert\,\mathrm{x}_i,\mathrm{nt})$	리텐션 타겟팅 하기 전 고객이 잔존할 확률
$v(x_i)$	고객가치
$c(x_i)$	회사가 고객에게 지출하는 비용
C	회사의 가용 지출 비용

그렇다면 $p(\mathrm{r}\,\vert\,\mathrm{x}_i,\mathrm{nt})$ 확률 값은 어떻게 얻을 수 있을까? 비즈니스 환경에 문제가 될만한 변화가 없다고 가정하면 앞서 예측 모델링을 통해 추정한 고객들의 이탈 확률을 $p(\mathrm{r}\,\vert\,\mathrm{x}_i,\mathrm{nt})$ 로 가정할 수 있다. 반면 $p(\mathrm{r}\,\vert\,\mathrm{x}_i,\mathrm{t})$ 확률을 추정하기 위한 데이터는 아직 존재하지 않는다. 해당 프로모션은 새롭기 때문에 어떤 고객도 이 프로모션에 접한 적이 없기

때문이다. 이에 본 연구에서는 $p(\mathrm{r}|\mathrm{x}_i,\mathrm{t})$의 확률 변화를 추정하기 위해 2017년 통신사의 적정 보조금을 추정하고자 일반 국민 558명을 대상으로 설문조사를 실시한 통신사 전환 비용 추정 연구 결과를 활용하였다. 전환 비용이란 고객에게 얼마의 금액이 보조금으로 추가 제공된다면 통신사를 이동할 것인지를 나타내는 금액이다[23].

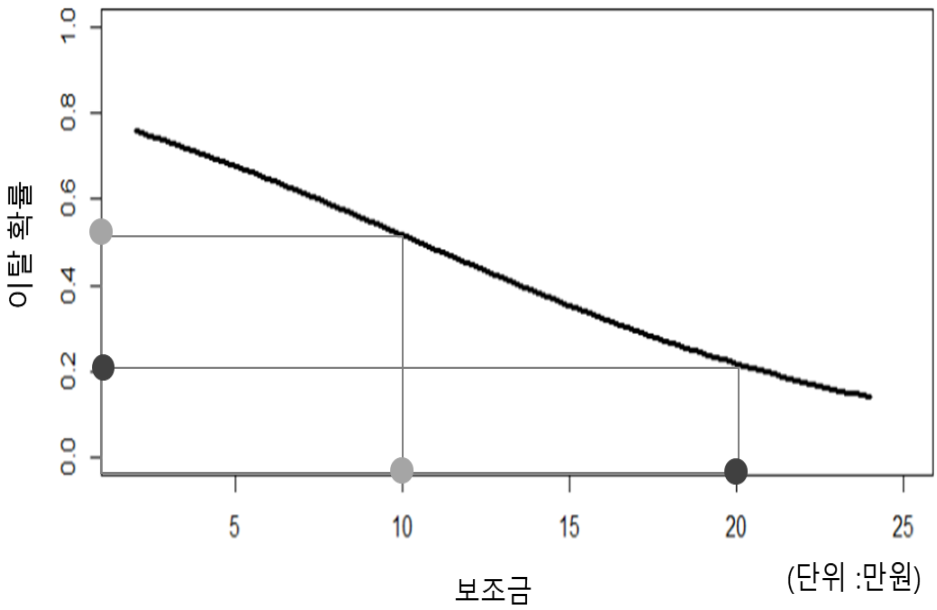


그림 3-7. 보조금 지급에 따른 고객의 이탈 확률 추정치

위 그래프에서 보는 바와 같이 전환 비용은 로지스틱 분포를 따르며 가입자당 약 10만 원 정도의 추가 보조금이 지급된다면, 고객은 통신사 전환을 할 의사가 있음을 알 수 있다. 제시된 금액에서 이탈할 확률은 보조금 액수가 올라갈수록 떨어진다. 예를 들어, 고객에게 20만 원의 보조금 혜택이 지급된다고 하면 이탈할 확률이 약 20%로 낮아진다고 말할 수 있다. 이 정보를 바탕으로 프로모션 타겟팅을 통해 고객에게 혜택을 주었을 때 $p(\mathrm{r}|\mathrm{x}_i,\mathrm{t})$ 확률 추정치를 얻을 수 있다.

그림 3-8은 잔존 확률(이탈하지 않을 확률)을 CDF(Cumulative Distribution Function)로 구현한 그림이다. 그림 3-7의 통신사 전환 비용 추정치와 R을 이용하여 평균 10만 원, 표준편차 11만 9천 원을

갖는 정규분포에 대해 $p(\mathrm{r}|\mathrm{x}_i, \mathrm{nt})$ 의 확률 값에서 Inverse CDF 값을 계산한 결과이다. 로지스틱 분포와 정규분포는 거의 비슷한 모양을 갖고 있기 때문에 정규분포 CDF로 변환해도 무방하다. 단 성별, 연령, 통신사 등의 환경변수를 고려하지 않는 결과값이므로 실제 타겟팅 환경에 따라 평균과 표준편차는 달라질 수 있으며, 이는 누적 확률 분포상 x축의 중심과 분포 스케일이 변할 수 있음을 가정한다. 정규분포와 로지스틱 분포의 CDF 식은 (3.4)와(3.5)로 표현된다.

$$\Phi(\mathrm{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{3.4}$$

$$\mathrm{F}(\mathrm{x}; \mu, \mathrm{s}) = \frac{1}{1+e^{-\frac{x-\mu}{s}}} \tag{3.5}$$

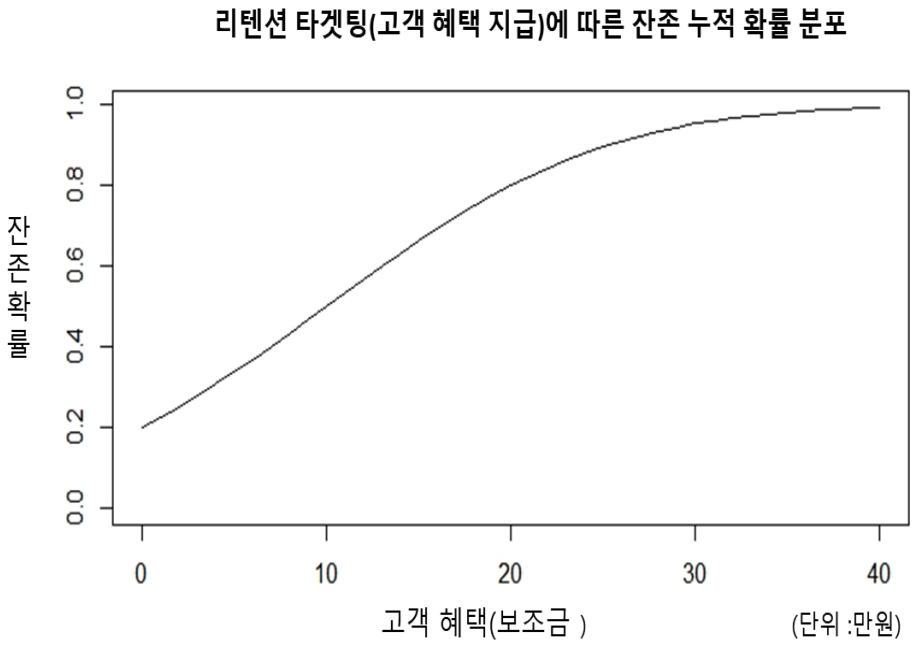


그림 3－8. 보조금 지급에 따른 고객의 잔존 누적확률분포

고객 혜택이 커질수록 고객이 이탈하지 않을 확률 값은 점점 증가하다가 고객 혜택이 약 40만원이 넘어가면서 확률 값은 1에 가까운 포화 상태에 이르게 된다.

이제 위 그래프에 표 3-3의 기대수익 최대화 모델 표기법에 따라 표기를 할 수 있다. 예를 들어 $p(\mathrm{r}|\mathrm{x}_i, \mathrm{nt})$ 확률 값이 60%인 고객에게 보조금을 30만원까지 지급했을 때의 $p(\mathrm{r}|\mathrm{x}_i, \mathrm{t})$ 확률 값과 $c(x_i)$를 그림 3-9과 같이 표현할 수 있다.

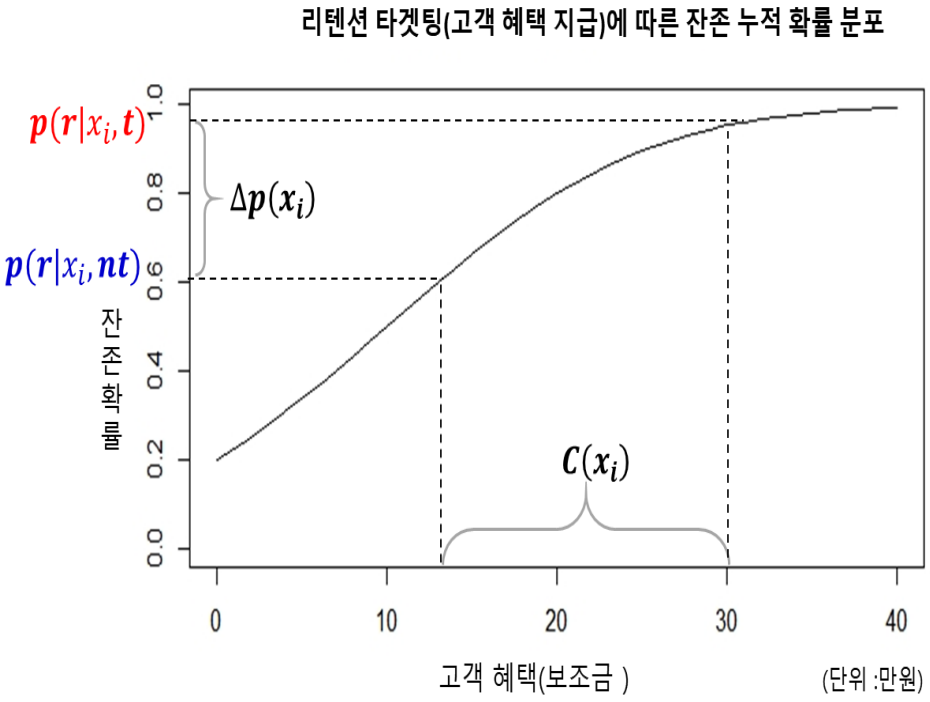


그림 3-9. 보조금 지급에 따른 고객의 잔존 누적확률 분포 모델 표기

추가로 17년 9월 이후 단통법(단말기 유통구조 개선에 관한 법률) 환경에서는 공시된 단말 가격 이외 보조금 지급을 허용하지 않기 때문에 실질적으로 국내에서는 이동통신사가 고객 이탈을 막기 위해 보조금을 지급하는 것은 불가능할 것이다. 하지만 본 연구에서는 무료 통화/데이터, 쿠폰, 멤버십 등 보조금에 상당하는 고객 혜택을 제공함으로써 보조금과 동일한 효과를 줄 수 있다고 판단하였다. 따라서 보조금은 고객 혜택과 동일하게 여기고 본 연구에서는 혼용 표기하였음을 참고 바란다.

제 4 장 최적화 시뮬레이션

기대수익 최대화 모델은 회사의 가용 지출 비용($\sum_{i=1}^{N} c(x_i)$)의 한계선이 없이 설계된 모델링이므로, 실질적으로 통신사 적용 가능 여부의 효과성 검증을 위해서는 한정된 예산 범위 안에서 최대의 기대수익을 낼 수 있도록 최적화하여야 한다. 이와 관련하여, 본 장에서는 최적화 시뮬레이션 프레임워크의 설계 및 구현 내역에 대해 상세하게 살펴보고자 한다.

제 1 절 시뮬레이션 개요

본 연구의 최적화 시뮬레이션 목표는 한정된 예산 내에서 기대수익을 최대화(Max[$\sum_{i=1}^{N}(\Delta p_i \cdot v(x_i)$]) 할 수 있는 방법을 찾는 것이다. 우선 두명의 고객 x1, x2에게 같은 보조금을 지급했다고 가정하자. 그림 4-1과 같이 이때 고객 x1은 λ_1 → λ'_1로 고객 x2는 λ_2 → λ'_2 로 이동하게 됨에 따라 고객 x1, x2의 잔존 확률은 $\Delta p(x_1)$, $\Delta p(x_2)$ 만큼 차이가 발생하게 된다.

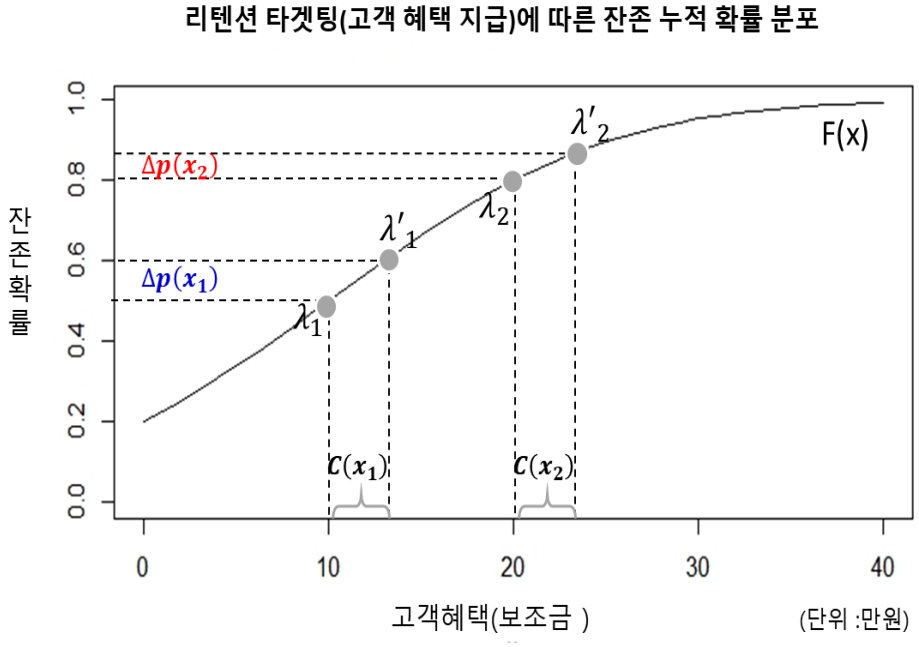


그림 4-1. 같은 고객혜택에도 Δp_i가 달라지는 예시

하지만 그림 4-1의 잔존 누적 확률분포 곡선은 linear하지 않기 때문에 각각의 고객에게 똑같은 $c(x_i)$를 지급하더라도 $p(\mathrm{r}|\mathrm{x}_i, \mathrm{nt})$가 같지 않은 한, 즉 타겟팅 하지 않았을 때 잔존할 확률이 다르다면 $\Delta \mathrm{p}_i$ 값은 고객마다 서로 다르다. 실제 예를 들어 $\lambda_1 = 10$만원, $\lambda_2 = 20$만원이라 가정하고 $c(x_1) = c(x_2) = 3$만원을 대입했을 때 $\Delta \mathrm{p}(\mathrm{x}_1) = 0.099$, $\Delta \mathrm{p}(\mathrm{x}_2) = 0.063$ 로 두 값이 틀림을 확인하였다. 따라서 한정된 예산에서 기대수익 최적화를 하려면 어느 고객에게 우선적으로 보조금을 지급해야 하는지가 관건이다.

이를 구현하기 위한 이론적 모델 정립을 위해 우선 전체 예산을 M 등분한 값을 c로 가정하자. 그리고 c만큼의 보조금을 지급했을 때 $\Delta p(x_i)$ 값이 누가 가장 큰지 비교하는 과정이 필요하다. 그래서 $\Delta p(x_i)$ 값이 가장 큰 고객에게 우선적으로 보조금을 지급해야 한다. 위 과정을 예산 한도에 이르기까지 M번 반복해야 한다. 여기서 M회 동안 고객 x_i가 받는 보조금 횟수를 j라 하고 이를 x_i^j로 가정하면, 매 회 보조금을 받는 고객의 $x_i^j = 1$이 되고 그 외 고객들은 x_i^j가 0 될 것이다. 이는 그림 4-1에서 x축의 변화를 미분하여 N명의 고객 중 M회 마다 $F(x)$ 값의 변화가 최대가 되는 고객을 찾아서 보조금을 우선적으로 지급해야 함을 의미한다. $F(x)$ 는 $[x_i^{j-1}, x_i^j]$ 구간에서 연속이고 구간에서 미분 가능한 함수이므로 (4.1)과 같이 표현 할 수 있다.

$$argmax\left(F\left(x_i^j\right) - F\left(x_i^{j-1}\right)\right) = argmax\left(\Delta p(x_i^j)\right) \text{ for M iteration,}$$
$$x_i^j \in \{0,1\} \ (\forall \mathrm{i} = 1,2,3,\ldots,\mathrm{N}),\ (\mathrm{j} = 1,2,3\ldots,\mathrm{M}) \qquad (4.1)$$

그리고 전체 고객 N명중에서 고객 x_i에게 j번째 지출하는 비용은 $\mathrm{c} \cdot x_i^j$이고 (4.2)와 같이 표현 할 수 있다.

$$\sum_{j=1}^{M}\sum_{i=1}^{N} \mathrm{c} \cdot x_i^j = Budget \qquad (4.2)$$

이러한 내용을 바탕으로 통신사 기대수익을 최적화하는 이론적 모델을 표 4-1과 같이 요약할 수 있다.

표 4-1. 기대수익 최적화 모델

Decision variables : x_i^j $(\forall \mathrm{i} = 1,2,3,\ldots,\mathrm{N}),\ (\mathrm{j} = 1,2,3\ldots,\mathrm{M})$

Objective :

$$\mathrm{Max}[\ \sum_{j=1}^{M}\sum_{i=1}^{N}(p(\mathrm{r}|\mathrm{x}_i^j,\mathrm{t}) - p(\mathrm{r}\,|\,\mathrm{x}_i^j,\mathrm{nt}))\cdot v(x_i)\ ] - \sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{c}\cdot x_i^j$$

$$= \mathrm{Max}[\ \sum_{j=1}^{M}\sum_{i=1}^{N}\Delta p(x_i^j)\cdot v(x_i)\ ] - \sum_{j=1}^{M}\sum_{i=1}^{N}\mathrm{c}\cdot x_i^j$$

Subject to : $0 < p(\mathrm{r}|\mathrm{x}_i^j,\mathrm{t}) < 1$

$0 < p(\mathrm{r}|x_i^j,\mathrm{nt}) < 1$

$\mathrm{c} = \frac{Budget}{M}$

$x_i^j \in \{0,1\}$

표 4-2. 기대수익 최적화 모델 표기법

기호	내용
$p(\mathrm{r}\|\mathrm{x}_i^j,\mathrm{t})$	고객 x_i에게 고객 혜택을 j번째 지급 했을 때 고객이 잔존할 확률
$p(\mathrm{r}\,\|\,\mathrm{x}_i^j,\mathrm{nt})$	고객 x_i에게 고객 혜택을 j-1번째 지급 했을 때 고객이 잔존할 확률 (j=1일 경우 타겟팅 하지않았을 때 잔존할 확률)
$v(x_i)$	고객가치(회사에 주는 수익)
$\mathrm{c}\cdot x_i^j$	고객 x_i에게 j번째 지출하는 비용
$Budget$	회사의 가용 지출 비용
N	전체 고객 수
M	한도 예산까지 고객에게 비용 지급 가능한 횟수

제 2 절 시뮬레이션 구현

여기에서는 기대수익 최적화 시뮬레이션 환경 구현의 핵심요소라고 할 수 있는 최적화 구현 알고리즘에 대해 소개하고자 한다. 표 4-3에는 최적화 알고리즘이 R 코드로 표현되어 있다.

표 4-3. 기대수익 최적화 알고리즘

ALGORITHM: 기대수익 최적화 FRAMEWORK

Input:

c: cost

Inverse CDF: x

$p(\mathrm{r}|\mathrm{x}_i^j,\mathrm{t})$: prob_t

$p(\mathrm{r}\,|\,\mathrm{x}_i^j,\mathrm{nt})$: prob_nott

$p(\mathrm{r}|\mathrm{x}_i^j,\mathrm{t})-p(\mathrm{r}\,|\,\mathrm{x}_i^j,\mathrm{nt})$: delta_p

$v(x_i)$: mcharge

$\mathrm{argmax}(revenue(x_i))$: n

기대수익: revenue_sum

C : cost_sum

```
repeat {
    cost = 1000
    x<-qnorm(prob_nott, mean=100000, sd=119000)
    prob_t<-pnorm(x+cost, 100000, 119000)
    delta_p = prob_t-prob_nott
    revenue = delta_p*mcharge*1100*26
    n = which.max(revenue)
    mylist[j] = n
    j=j+1
    x<-qnorm(prob_nott[n], mean=100000, sd=119000)
    prob_t[n]<-pnorm(x+cost, 100000, 119000)
    delta_p[n] = prob_t[n]-prob_nott[n]
    revenue[n] = delta_p[n]*mcharge[n]*1100*26
    prob_nott[n] = prob_t[n]
    cost_sum = cost_sum+cost
    revenue_sum = revenue_sum+revenue[n]
    if(cost_sum >= 1,000,000,000) break  }
```

기대수익 최적화 알고리즘은 타겟팅 프로모션을 통하여 고객 이탈을 방지하고자 할 때 통신사 입장에서 한정된 예산 범위 내 최적의 의사 결정을 수행할 수 있도록 구현하였다. Input 변수를 최적화 모델 표기와 매핑하여 살펴보면, 우선 cost는 N명의 고객에게 전체 예산을 M회 마다 나눠 지급하는 보조금 혜택으로 모든 고객에게 동일한 금액으로 주어져야 한다. 이와 더불어 타겟팅 이전의 확률 값(prob_nott)에 해당하는 보조금액(x), 해당 보조 금액에 cost를 타겟팅 프로모션으로 지급했을 때의 잔존 확률(prob_t), 타겟팅 전후의 잔존 확률 차이(delta_p), 고객이 회사에 주는 고객가치(mcharge), 어느 고객의 revenue가 가장 큰지 알 수 있는 인덱스(n), 예산 한도 범위까지 cost를 지급하고 난 후 revenue의 총합(revenue_sum) 등이 입력 변수로 활용된다. 또한 필요한 경우 고객별로 cost를 지급한 횟수(mylist)를 입력할 수 있다.

알고리즘의 흐름을 의사 코드의 행 번호를 기준으로 살펴보면, 우선 cost 값을 정하고 모든 고객에게 cost를 지급 했을 때 잔존 확률을 구하는 단계(1-3행), cost 지급 전후 잔존 확률의 차이 delta_p를 구하여 고객별 기대수익을 구하는 단계(4-5행), 어느 고객이 가장 높은 기대수익을 주는지(delta_p 값이 큰지) 인덱스 처리하는 단계(6-8행), 기대수익이 가장 높은 고객에게 실제 cost를 지급하는 단계(9-12행), cost를 지급받은 고객의 prob_t를 prob_nott로 업데이트하는 단계(13행), cost와 revenue의 누적 합을 구하고 cost의 누적 합이 예산 범위 내 있는지에 따라 위 과정을 다시 반복하는 단계(14-16행)로 진행된다. 이 중 7-8행에 언급된 mylist 설정은 어느 고객에게 몇번의 cost를 지급하였는지 확인할 수 있는 count 변수로 필수 input 변수는 아니다. 참고로 cost는 지급 가능한 최소 단위로서 1000원으로 정했고 revenue는 고객의 가치를 월 청구액으로 평가하였고, 이를 국내 가치로 환산하기 위해 환율 1100원에 약정 기간 포함하여 26개월치를 곱하였다. 위 과정을 그림 4-2와 같이 표현 할 수 있다.

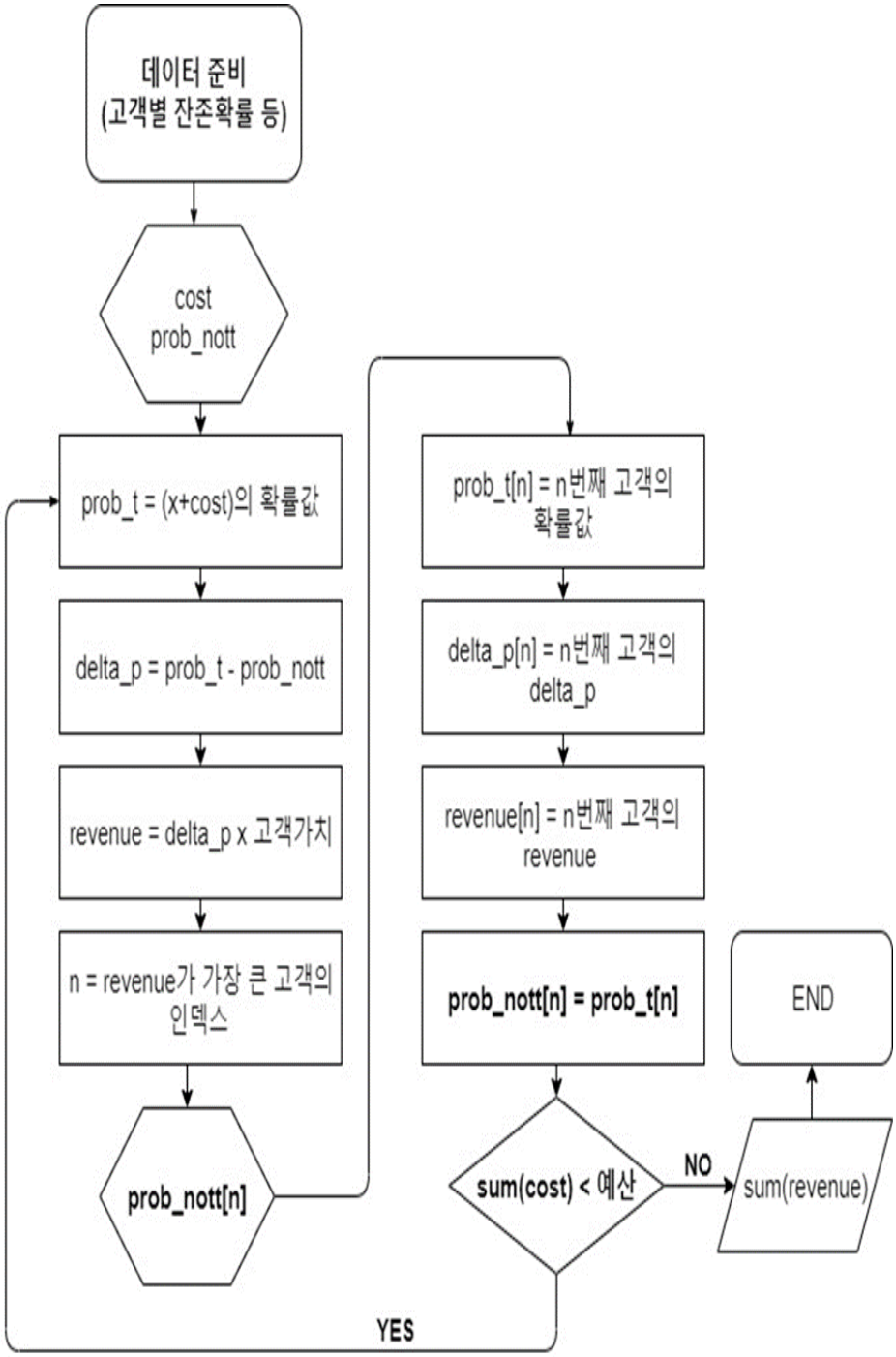


그림 4-2. 기대수익 최적화 알고리즘 순서도

제 3 절 시뮬레이션 수행 조건

본 연구는 표 4-4와 같이 프로모션 고객 혜택(보조금) 지급 방식을 최적화 방식 외 4가지 항목을 기준으로 시뮬레이션을 수행하였다. 최적화 알고리즘을 통해 한정된 예산에 대한 최대의 기대수익을 도출하고(1번 항목), 보조금 지급 방식을 달리하여 각 항목별 기대수익 및 연산 처리 시간을 비교 분석하였다(2~5번 항목). 이를 기반으로 최적화 알고리즘의 효율성을 확인 할 수 있다.

표 4-4. 최적화 외 시뮬레이션 항목

<table>
<tr><th>번호</th><th>시뮬레이션 항목
(보조금 지급 방식)</th><th>실험 변수</th><th>측정지표</th></tr>
<tr><td>1</td><td>최적화 알고리즘</td><td>예산
고객 샘플 수</td><td rowspan="5">기대수익
및
연산 처리 시간</td></tr>
<tr><td>2</td><td>총 예산/고객 수</td><td rowspan="4">$cost(x_i)$</td></tr>
<tr><td>3</td><td>고객 월 요금 비율</td></tr>
<tr><td>4</td><td>고객 이탈 확률 비율</td></tr>
<tr><td>5</td><td>고객 월 요금*이탈 확률 비율</td></tr>
</table>

1번 항목에서는 최적화 알고리즘의 경향성을 분석하고자 한다. 여기에서는 총 예산과 고객 수를 실험 변수로 설정하고, 기대수익 및 연산 처리 시간을 측정하고자 한다.

2번~5번 항목에서는 보조금 지급 방식을 달리하여 기대 수익과 처리 시간을 측정하고자 하기 때문에 고객별 지급받는 보조금을 $cost(x_i)$, 총 예산을 budget, 고객 수를 N 이라고 가정하고 나머지는 표 3-4의 표기법을 준용한다고 했을 때 2번 항목의 $cost(x_i)$ 변수는 (4.4)과 같이 설정 할 수 있다.

$$cost(x_i) = \frac{budget}{N} \tag{4.3}$$

3번 항목에서는 고객의 가치 비율(여기서는 월 요금 비율)로 지급했을 $cost(x_i)$는 (4.5)과 같이 설정 할 수 있다.

$$cost(x_i) = \text{budget} \cdot \frac{V(x_i)}{\sum_{i=1}^{N} V(x_i)} \tag{4.4}$$

4번 항목에서는 이탈 확률 비율을 구하기 위해 (1−잔존확률)의 비율로 식을 대체하였다. $cost(x_i)$는 (4.6)과 같이 표현 할 수 있다.

$$cost(x_i) = \text{budget} \cdot \frac{1 - p(\mathrm{r}|\mathrm{x}_i, \mathrm{nt})}{\sum_{i=1}^{N}(1 - p(\mathrm{r}\,|\,\mathrm{x}_i, \mathrm{nt}))} \tag{4.5}$$

5번 항목에서는 고객 가치 비율과 이탈 확률 비율의 곱을 통해 cost는 (4.7)과 같이 표현 할 수 있다.

$$cost(x_i) = \text{budget} \cdot \frac{V(x_i)}{\sum_{i=1}^{N} V(x_i)} \cdot \frac{1 - p(\mathrm{r}|\mathrm{x}_i, \mathrm{nt})}{\sum_{i=1}^{N}(1 - p(\mathrm{r}\,|\,\mathrm{x}_i, \mathrm{nt}))} \tag{4.6}$$

제 4 절 시뮬레이션 수행 결과

1번 항목에서 최적화 알고리즘에 따라 예산 1억, 5억, 10억으로 고객 샘플 수를 1,406명, 7,032명으로 실험 변수로 설정하고 시뮬레이션 한 결과표 4-5와 같다. 고객 샘플 수가 1,406명 일 때 예산을 1억에서 10억으로 늘리면 기대 수익은 약 35% 증가하고, 특히 음영 부분과 같이 5억에서 10억으로 예산 변경하면 기대 수익의 변화가 거의 없는 것을 확인할 수 있다. 반면 고객 샘플 수 7,032명으로 고정하고 예산을 1억에서 10억으로 늘리면 기대수익은 약 346% 증가하고 5억에서 10억으로 예산이 늘어나도 기대수익 증가율의 변화가 큰 것을 확인하였다.

표 4-5. 최적화 시뮬레이션 결과

구분	1,406 명	7,032 명
1 억	579,497,348 원	826,685,783 원
5 억	779,436,043 원	2,904,327,822 원
10 억	779,557,111 원	3,687,540,288 원
1 억→10 억 기대수익 증가율	**35%**	**346%**

이렇게 고객 샘플 수, 예산의 정도에 따라 증가율 차이가 각기 다른 이유는 보조금 지급이 증가하더라도 $\Delta p(x_i^j)$는 그림 4-3과 같이 거의 0으로 수렴되기 때문에 계속 지급하다 40만 원 가까이 되면 $\Delta p(x_i^j)$는 매우 작은 값으로 계속 낮아져 보조금 지급 대비 기대 수익의 효율은 떨어지는 것으로 추정할 수 있다. 또한 샘플 고객 수 1,406명, 예산이 10억일 때 기대수익이 약 7.8억으로 투자 원금에 못 미치는 현상이 나타났다. 이는 마찬가지로 모든 고객이 $\Delta p(x_i^j)$ 가 0으로 수렴되기 때문에 보조금을 계속 지급하더라도 기대수익은 saturation에 이르게 되는 것이다.

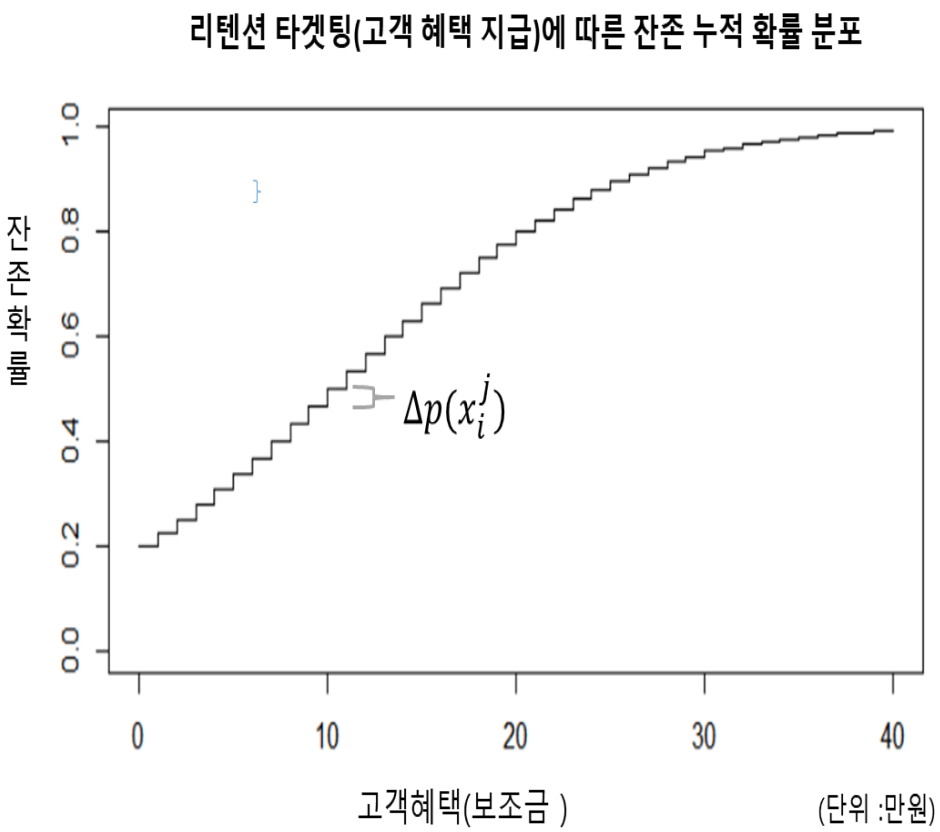


그림 4-3. 보조금 증가에 따른 $\Delta p(x_i^j)$ stair steps

따라서 실제 통신 사업자 입장에서 마케팅 비용 지출을 고려할 경우 무작정 예산을 늘리는 것보다 언제 $\Delta p(x_i^j)$ 값이 포화 상태에 이르는지를 고려해야 한다. 결국 실익(기대수익 - 예산) 관점에서 보면 실익이 증가하다가 감소하게 되는 변곡점이 생기게 마련인데 이 변곡점을 찾으면 $\Delta p(x_i^j)$ 값이 포화 상태에 이르기 전 필요 예산이 얼마인지를 확인 할 수 있을 것이다.

변곡점에 해당하는 보조금을 a, 실익을 나타내는 곡선은 $f(a)$라고 가정하면 변곡점에 해당하는 $f(a)'=0$의 a값을 찾아야 최대 실익을 구할 수 있다. 이러한 내용을 그림 4-4로 표현할 수 있다.

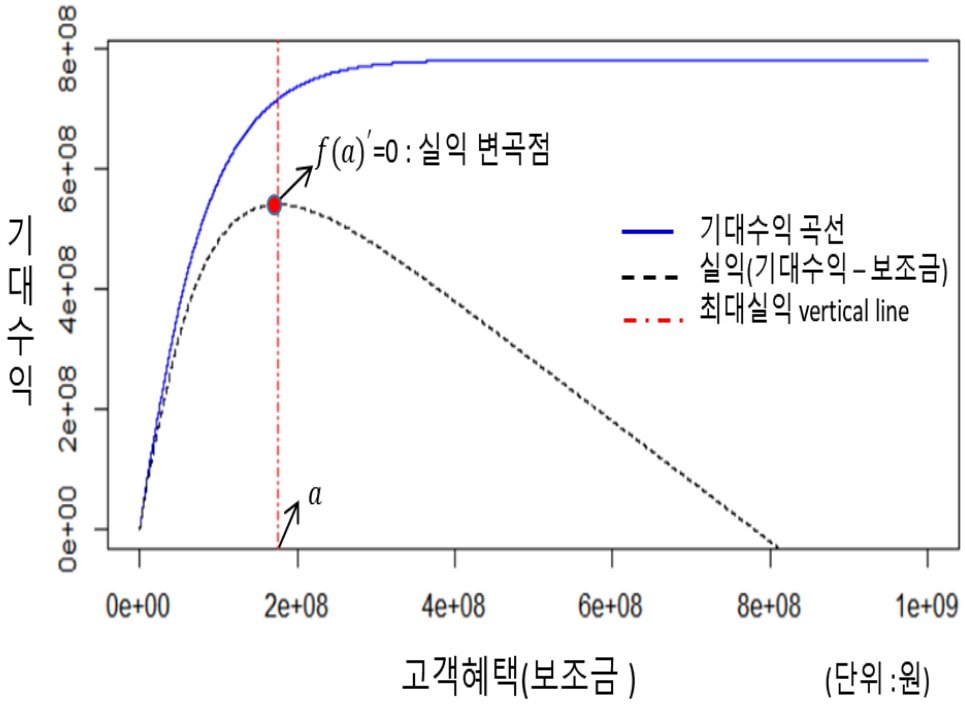


그림 4-4. 실익 곡선과 변곡점

본 연구에서는 a값을 구하기 위해 $f(a)$ 식을 구하지 않고 간단히 R 코드로 구현할 수 있었다. 우선 기대 수익과 보조금의 차이가 가장 큰 포인트를 찾도록 했고 해당 포인트의 index를 통해 a값을 찾았다. 또한 그 index에 해당하는 기대수익값을 찾아서 둘간의 차이를 구하면 최대 실익을 얻을 수 있다. 고객 샘플 수 1,406명, 예산 10억, cost 1만원 기준으로 시뮬레이션한 결과 그림 4-5과 같다. 보조금 지급액이 약 1.75억을 넘어서는 순간 최대 실익은 약 5.39억이고 그 이후 감소하는 모습을 보이고 있다.

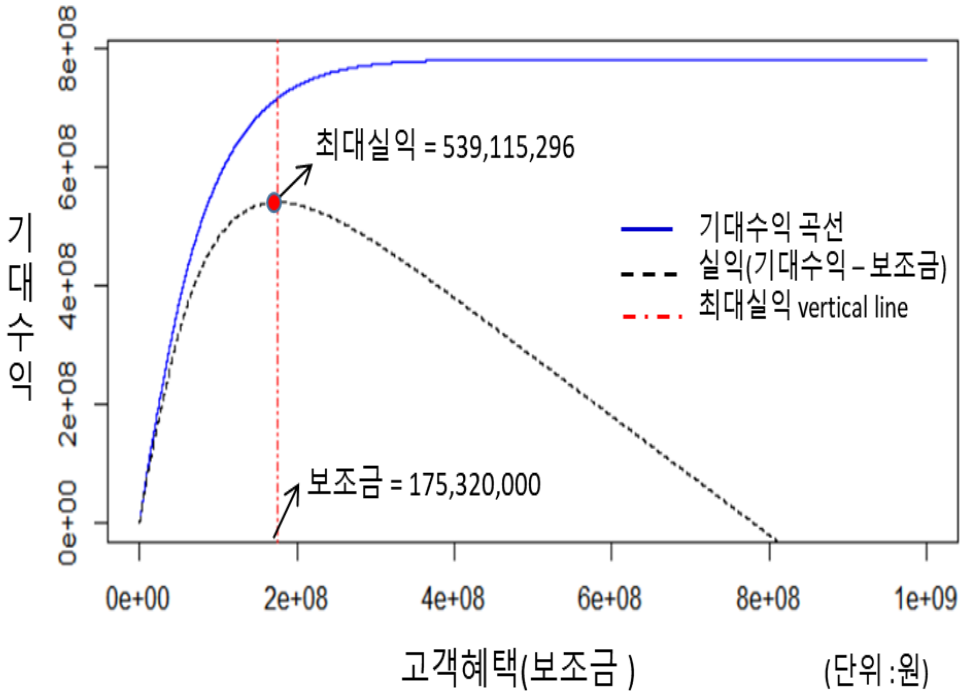


그림 4-5. 보조금, 기대수익, 실익 곡선
(조건: 고객 샘플 수 1,406명, 예산 10억, cost 1만원)

고객 샘플 수를 7,032명으로 늘려서 같은 환경으로 시뮬레이션 진행한 결과 그림 4-6와 같다. 고객 샘플 수가 1,406명 일 때 보다 최대 실익 vertical line이 오른쪽으로 shift 되었다. 이는 보조금을 지급해야 할 대상 수가 늘어났기 때문에 모든 고객의 $\Delta p(x_i^j)$가 포화 상태에 도달하려면 보조금을 전보다 더 많이 지급할 수 밖에 없는 예상된 결과이다. 그림 4-7는 두 경우의 기대수익 곡선을 한 눈에 비교하기 쉽도록 나타낸 그림이다.

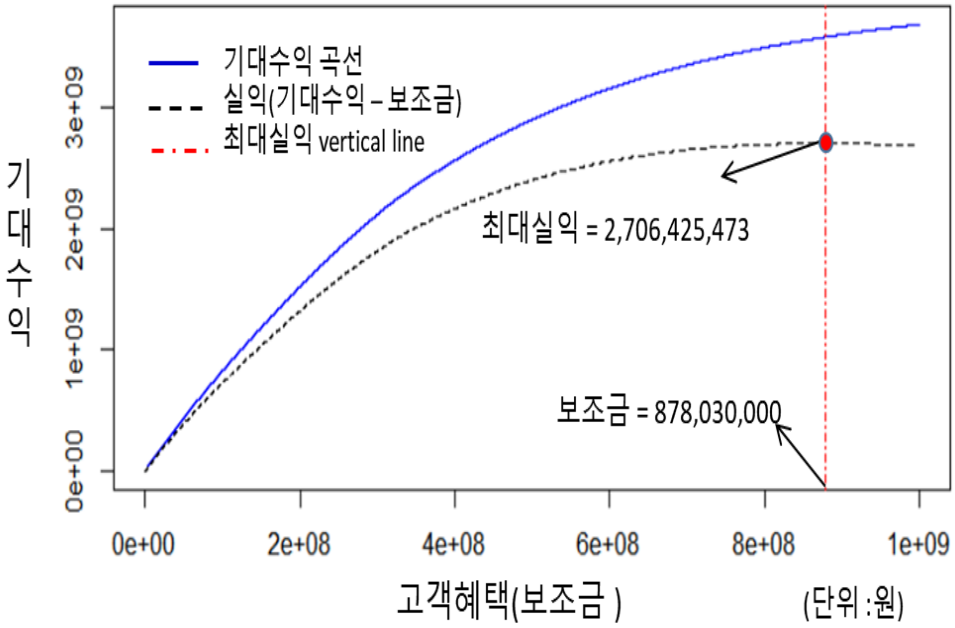


그림 4-6. 보조금, 기대수익, 실익 곡선
(조건: 고객 샘플 수 7,032명, 예산 10억, cost 1만원)

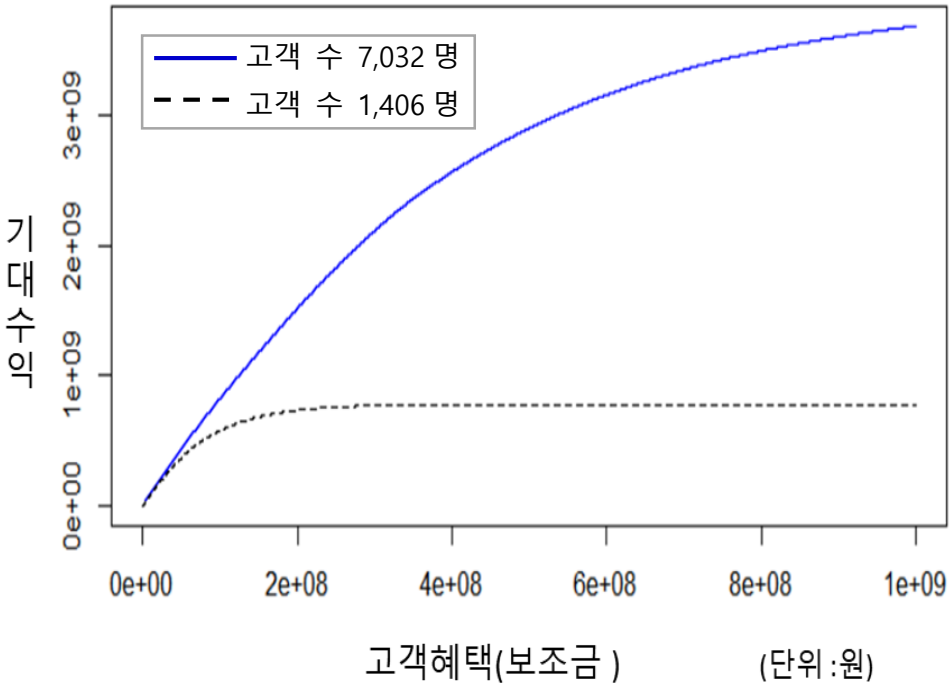
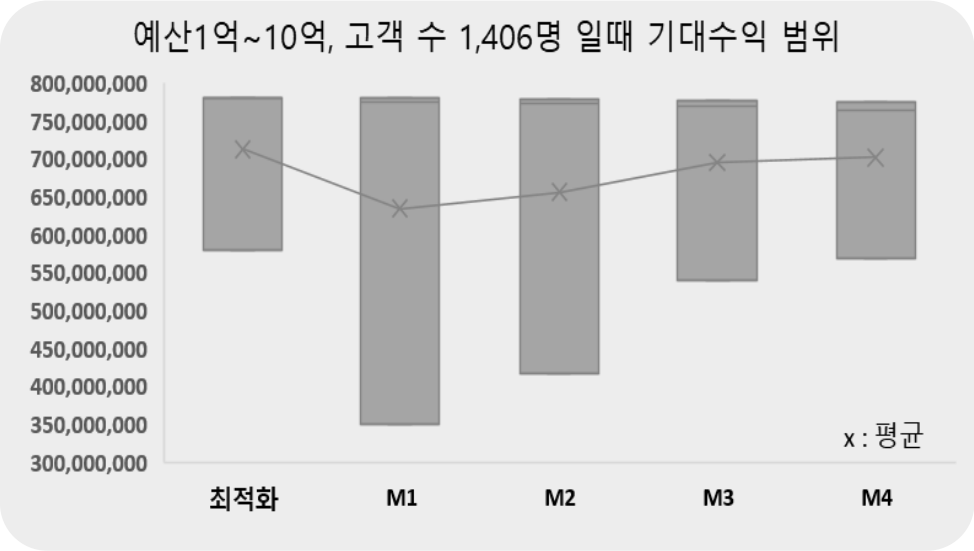


그림 4-7. 고객 샘플 수에 따른 기대수익 곡선 비교

최적화 시뮬레이션 외 나머지 2~5번 항목을 동일하게 고객 수를 1,406/7,032명으로 두고 예산을 1/5/10억으로 변경하며 수행한 결과를 그림 4-8과 같이 시각화 하여 비교하였다. 고객 수, 예산을 변경하여도 최적화 방식의 기대수익이 항상 높은 것을 확인 할 수 있다. 상세 내역은 표 4-6에서 살펴볼 수 있다.



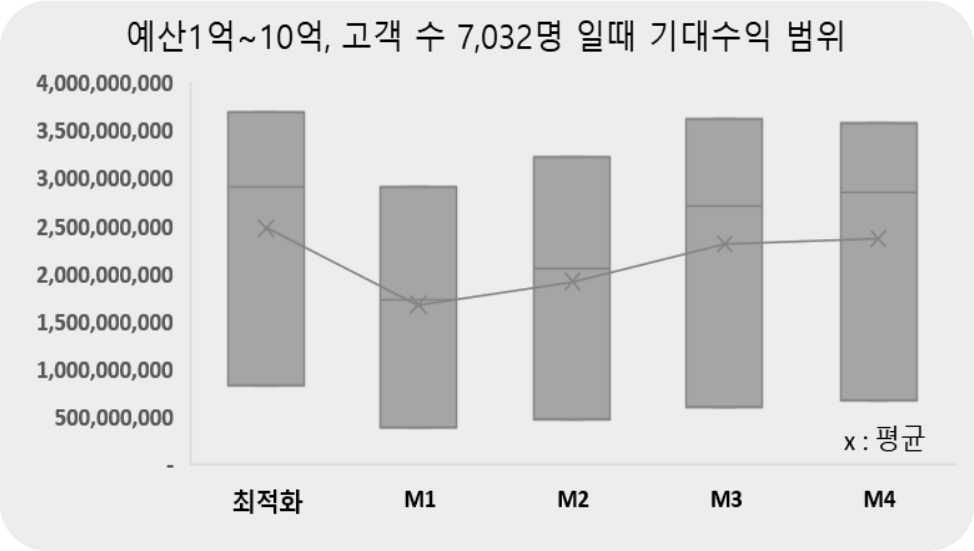


그림 4-8. 최적화 외 시뮬레이션 항목 기대수익 범위 비교

[M1: 총 예산/고객 수 , M2: 고객 월 요금 비율,
*M3: 고객 이탈 확률 비율, M4: 고객 월 요금*이탈 확률 비율]*

표 4-6. 최적화 외 시뮬레이션 항목 비교 표

No	실험 변수 [$c(x_i^j)$= 1 천원]	보조금 지급 방식	기대수익(원)	M1~M4 대비 최적화 효율비	연산 수행 시간(sec)
1	예산 1 억, 고객 수 1,406 명	최적화	579,497,348		37.89
		M1	349,025,514	66.0%	0.59
		M2	416,707,132	39.1%	0.67
		M3	539,998,387	7.3%	0.63
		M4	567,924,371	2.0%	0.72
2	예산 5 억, 고객 수 1,406 명	최적화	779,436,043		225.28
		M1	774,864,510	0.6%	0.57
		M2	773,182,979	0.8%	0.67
		M3	769,463,866	1.3%	0.66
		M4	763,376,820	2.1%	0.57
3	예산 10 억, 고객 수 1,406 명	최적화	779,557,111		452.49
		M1	779,557,081	0.0%	0.67
		M2	779,221,030	0.0%	0.59
		M3	776,701,793	0.4%	0.61
		M4	774,148,939	0.7%	0.65
4	예산 1 억, 고객 수 7,032 명	최적화	826,685,783		212.84
		M1	385,749,073	114.3%	12.9
		M2	478,793,028	72.7%	11.91
		M3	606,447,025	36.3%	12.07
		M4	676,695,967	22.2%	11.83
5	예산 5 억, 고객 수 7,032 명	최적화	2,904,327,822		907.44
		M1	1,732,160,377	67.7%	12.17
		M2	2,059,493,083	41.0%	12.74
		M3	2,710,281,572	7.2%	13.93
		M4	2,845,389,669	2.1%	14.04
6	예산 10 억, 고객 수 7,032 명	최적화	3,687,540,288		2081.07
		M1	2,900,536,038	27.1%	13.35
		M2	3,212,400,805	14.8%	13.69
		M3	3,610,722,924	2.1%	13.71
		M4	3,575,178,330	3.1%	14.03

실험 케이스를 예산 1/5/10억, 고객 샘플 수를 20%, 100%로 나누어서 진행하였고 기대수익을 비교한 결과, 최적화 알고리즘 적용 시 가장 높은 기대수익을 나타냈으며 실험 케이스 4번을 보면 최적화 방식이 다른 보조금 지급 방식 대비 최대 114.3%의 효율성을 나타내었다. 참고로 보조금 지급 방식 중 M1, 2 방식은 고객 이탈 확률 추정 없이 기본 데이터 세트를 통해 산출 가능한 방식이고 M3, 4는 머신 러닝을 통해 고객 이탈 확률 추정치가 필요한 방식이다.

최적화 방식을 제외한 기대수익이 가장 높은 시뮬레이션 항목은 음영부분으로 처리한 것과 같이 실험 케이스마다 각각 다르다(1번 실험에서는 M4, 2번 실험에서는 M1, 3번 실험에서는 M1, 4번 실험에서는 M4, 5번 실험에서는 M4, 6번 실험에서는 M3가 기대수익이 가장 높다). 실제로 마케팅 예산 분배 및 정책 실무 적용할 때 통신 기업마다 예산 및 고객 수는 다를 것이기 때문에 이 결과가 시사하는 바는 굉장한 의미를 부여한다. 가령 실험 1번 케이스의 환경에 처한 기업이 마케팅 정책으로 보조금 지급 방식을 M4의 방법으로 지급하여 최대 기대수익을 올렸다고 가정하자. 해당 기업이 고객 수도 늘고 마케팅 예산도 늘어나 실험 6번 케이스로 되었을 때, 정책 담당자는 최대 기대수익을 얻으려면 M3의 방식으로 보조금 지급 정책을 펼쳐야 하는 상황이지만 과거 경험에 따라 M4 방식으로 보조금 지급할 가능성이 크다. 즉 기업 환경에 따라 최대 기대수익을 보장할 수 없고, 추정할 수 있는 방식이 매번 달라지는 셈이다. 하지만 본 연구에서 제시하는 최적화 알고리즘을 적용하면 어떠한 환경변수에서도 최대 기대수익을 보장할 수가 있다.

연산 수행 시간은 최적화 알고리즘 방식이 다른 보조금 지급 방식보다 오래 걸리지만, 실험 결과 복잡도가 샘플 수나 마케팅 비용에 대해 선형으로 달라지는 것으로 가정하면 고객 샘플 수를 1,406명에서 7,032명으로 늘렸을 때와 예산을 1억에서 10억으로 늘렸을 때 각각 500%, 1,000% 연산 시간이 늘어나겠지만, 표 4-7과 같이 실제 증가율은 이보다도 더 적은 것을 확인하였다. 따라서 최적화 알고리즘의

실무 적용에 있어 computation상의 문제는 없을 것으로 예상된다.

표 4-7. 최적화 연산 수행 시간 비교

구분	1억	10억
1,406명	37.89	452.49
7,032명	212.84	2081.07

(단위:sec)

구분	1,406명 →7,032명	1억 →10억
예상 증가율	500%	1000%
실제 증가율	**460%**	**978%**

실제로 약 10배가량 scale up(기존 데이터의 평균과 표준편차를 이용하여 정규분포를 가정하고 랜덤 생성)하여 진행한 결과 표 4-8로 나타내었다. 연산 수행 시간은 약 9시간으로 실무 작업 환경상 무리가 없을 것으로 판단된다.

표 4-8. 10 배 Scale up 하였을 때 최적화 외 시뮬레이션 항목 비교 표

No	실험 변수 [$c(x_i^j)$= 1 만원]	보조금 지급 방식	기대수익(원)	M1~M4 대비 최적화 효율비	연산 수행 시간(h)
7	예산 100 억, 고객 수 84,658 명	최적화	**43,195,986,793**		9.34
		M1	36,302,307,738	19.0%	0.56
		M2	38,372,053,858	12.6%	0.58
		M3	41,661,138,694	3.7%	0.53
		M4	42,392,110,319	1.9%	0.88

만약 실무에서 예산과 고객 수를 위 케이스들보다 훨씬 더 scale up 하여 computation에 문제가 발생한다고 가정하면, 최적화 알고리즘을 배제하고 다른 지급 방식을 선택해야 할 것이다. 실험 2, 3 케이스와 같이 예산이 많고 고객 수가 적은 경우는 M1 방식처럼 예산을 균등 배분하여 보조금을 지급하는 것이 최선의 방법이겠지만 현실적으로 이런 케이스는 거의 없을 것으로 생각된다. 반면 실제 상황과 유사한 실험 1, 4, 6, 7 케이스와 같이 제한된 예산 범위 내에서 고객 수가 많은 경우는 M4 방식 즉, 고객의 가치(월 요금)와 고객 이탈률에 비례하여 보조금을 지급하는 방식이 대안이 될 것이다.

제 5 장 결 론

제 1 절 연구 성과

본 연구는 고객 이탈률을 머신러닝을 통해 예측하고, 이를 분석하여 고객 이탈 방지를 위한 프로모션 비용 지급 방식을 기대수익을 최대화 할 수 있는 최적화 알고리즘을 제안하였다. 또한, 이와 관련하여 실제 예산 집행 환경 입각한 시뮬레이션을 진행하였으며, 이를 활용하여 기대수익 관점에서의 실질적인 효율성 검증 및 분석을 수행하였다.

본 연구를 통하여, 최적화 알고리즘 적용하면 최적의 예산 배분으로 기대수익을 높일 수 있음을 확인하였다. 특히 적은 예산으로 많은 고객에게 보조금을 지급해야 하는 상황일수록 더 우수한 효율성을 나타내었다. 시뮬레이션 수행 결과 케이스 4번을 보면 단순히 예산을 1/N으로 나눠 보조금을 지급하는 경우 대비 최적화 알고리즘의 효율성이 114% 가량 우수함을 확인하였고, 고객 월 요금이 높은 순으로 지급하는 방법 대비 72.7%, 고객 이탈률이 높은 순으로 지급하는 경우 대비 36.3% 그리고 최적화 알고리즘의 대안 방법으로 제시한 고객 월요금과 이탈률의 곱이 높은 순서대로 지급할 경우보다 22.2%가량 더 많은 기대수익을 얻을 수 있었다.

기존에 이루어진 대다수의 고객 이탈 예측 관련 연구들은 이탈 예측 정확도 및 방법론적 최적화에 초점을 맞추었으며, 실제 이탈 예측한 결과를 바탕으로 통신사의 예산 환경을 고려하여 이탈 방지를 위한 프로모션 비용 지급 방식을 유형별로 비교 분석을 수행한 연구 사례를 찾을 수는 없었다. 이러한 관점에서 본 연구 결과는 향후 이루어질 마케팅 비용 절감 관련 연구들에 많은 시사점을 제공할 수 있을 것으로 기대된다.

특히 본 연구는, 이동통신 기술의 5G 전환과 동시에 통신사별 신규 가입자 유치 및 기존 가입자 전환 유지가 적극적으로 검토되고 있는 현시점에서, 고객의 이탈률을 사전 분석하고 리텐션 타켓팅하는 방안을 제시하였다는 점에 큰 의미를 부여할 수 있으며, 이동통신 사업자의

비용 기획 및 절감 방안에 있어서 고려되어야 할 가입자 유지(유치)와 비용 간의 trade-off 문제를 해결하는데 도움이 될 것으로 기대된다.

위와 같은 내용을 표 5-1에서 간략하게 정리하였고, 본 연구 내용을 실무에 적용했을 때 예상되는 전후 모습을 기술하였다.

표 5-1. 연구 성과와 실무 적용 시 전후 비교

성과 구분	AS-IS(예시)	TO-BE
과학적 예산 기획 및 배분	작년 비용 대비 % 증감 지역별 가입자수 고려하여 예산 1/N	가입자의 이탈율 사전 분석 ⬇ 기대수익 최적화 알고리즘 적용 ⬇ 소요 비용 기획
정밀한 타겟 프로모션	약정 만료(또는 예정) 고객 무작위 프로모션	고객별 최적의 전환 비용 가이드라인 제공 ⬇ 고객 리텐션을 위한 정밀한 프로모션 상담 가능
저비용 고수익	영업이익 대비 높은 마케팅 비용 지출	한정된 예산 환경 내 최적의 기대수익 창출

최적화 알고리즘 UPDATE

프로모션 결과 데이터 자산화

제 2 절 향후 연구 계획

본 연구에서는 고객 이탈률 예측, 그리고 고객 이탈 방지를 위한 마케팅 비용 투자 효율성 향상이라는 두 가지 문제를 해결하는데 있어서 각기 다른 방법론을 적용하였다. 고객 이탈률 예측 문제에 대해서는 머신러닝을 통하여 다양한 모델링을 비교 분석하였고, 투자 효율성 향상의 문제는 최적화 기법을 통하여 효과성을 확인하였다. 서로 다른 특성의 문제를 해결하는데 있어서 공통적으로 필요한 것은 고객의 이탈 여부, 고객가치 등이 포함된 다양한 데이터이다. 실제 이동 통신 환경에서 벌어지는 다양한 고객 데이터들에 대한 분석이 이루어져야 한다. 이에, 후속 연구에서는 실제 이동 통신사의 데이터들에 대한 체계적인 분석을 수행하고, 이탈률에 영향을 미치는 다양한 독립변수들을 활용하고 비용을 연계하여 보다 의미 있는 연구 성과를 도출하고자 한다.

아울러, 본 연구에서는 고객의 가치를 고객의 월 요금으로 대체하였지만 신뢰성을 부여하기 위해서는 CLV 관점에서의 보다 구체적인 설계가 이루어져야 한다. 또한 타겟팅 하는데 프로모션에 응할 가능성이 낮은 고객을 잘못 타겟팅하면 비용만 발생하고 수익을 낼 수 없기 때문에 고객의 반응을 구별해낼 방안을 도출하여야 한다. 이에, 후속 연구에서는 고객 가치에 대한 면밀한 검토를 바탕으로 기대수익 모형을 구체화 시키고, 타겟팅에 소요되는 비용 요소를 포함하여 기대수익 최적화를 효과적으로 적용하는 방법론에 대해 연구하고자 한다.

참 고 문 헌

[1] Molly Galetto, https://www.ngdata.com/what-is-customer-retention/, 2015년 6월 25일

[2] 이성용, 베인앤드컴퍼니 대표, “신규 고객 늘리기보다 기존 고객 유지 힘써라”, https://www.bain.com/ko/insights/customer-lifecycle/

[3] Santharam, A., & Krishnan, S. B. (2018). Survey on Customer Churn Prediction Techniques. International Research Journal of Engineering and Technology, 5(11), 3.

[4] 통신3사 마케팅 비용과 영업이익, 2018년 10월8일, http://www.ddaily.co.kr/news/article/?no=173376

[5] Óskarsdóttir, M., Baesens, B., & Vanthienen, J. (2018). Profit-based model selection for customer retention using individual customer lifetime values. Big data, 6(1), 53-65.

[6] T. Verbraken, W. Verbeke, and B. Baesens, “A novel profit maximizing metric for measuring classification performance of customer churn prediction models,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 961–973, 2013.

[7] T. Fawcett, “An introduction to roc analysis,” Pattern recognition letters, vol. 27, no. 8, pp. 861–874, 2006.

[8] Peng, Cyj. "An Introduction to Logistic Regression Analysis and Reporting." The Journal of Educational Research. 96.1 (2002): 3. Web.

[9] MIT OpenCourseWare(Producer).(2014). *MIT 6.034 Artificial Intelligence, Fall 2010* [Youtube]. Available from https://www.youtube.com/watch?v=_PwhiWxHK8o&t=6s

[10] K. -. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf, "An introduction to kernel-based learning algorithms," in IEEE Transactions on Neural Networks, vol. 12, no. 2, pp. 181-201, March 2001.doi: 10.1109/72.914517

[11] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[12] 위키피디아, "랜덤포레스트", https://ko.wikipedia.org/wiki/랜덤_포레스트

[13] Shotton, J., Sharp, T., Kohli, P., Nowozin, S., Winn, J., & Criminisi, A. (2013). Decision jungles: Compact and rich models for classification. In Advances in Neural Information Processing Systems (pp. 234-242).

[14] Kirui, C., Hong, L., Cheruiyot, W., & Kirui, H. (2013). Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI), 10(2 Part 1), 165.

[15] Óskarsdóttir, M., Baesens, B., & Vanthienen, J. (2018). Profit-based model selection for customer retention using individual customer lifetime values. Big data, 6(1), 53-65.

[16] Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert systems with applications, 26(2), 181-188.

[17] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

[18] Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629-1636.

[19] Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

[20] Verbraken, T., Verbeke, W., & Baesens, B. (2014). Profit optimizing customer churn prediction with Bayesian network classifiers. Intelligent Data Analysis, 18(1), 3-24.

[21] Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough?. Expert Systems with Applications, 39(18), 13517-13522.

[22] Lee, E., Kim, B., Kang, S., Kang, B., Jang, Y., & Kim, H. K. (2018). Profit Optimizing Churn Prediction for Long-term Loyal Customer in Online games. IEEE Transactions on Games.

[23] 이현송. (2018). 기업전용 LTE 선호도 분석 및 통신사 전환비용 추정에 관한 연구 (Doctoral dissertation, 서울대학교 대학원).

Abstract

Profit based model selection and optimization for customer retention

Jaeyeop Kim

Department of Engineering Practice

Graduate School of Engineering Practice

Seoul National University

The marketing costs of the three Korean telecommunication companies are 8 trillion won per year and 2.3 times higher than their operating profits (based on the audit data of the state administration in 2018), which is waging a war without guns to attract customers/to prevent them from going out of the company. In this study, various machine learning techniques were applied to minimize marketing costs to prevent customer churn and based on these methods, the company conducted optimization experiments to maximize expected revenue from the perspective of telecommunication companies.

IBM Watson Analytics, Guide to Sample Data Sets' Customer Support files were utilized as a dataset (population: 7,043) and machine learning showed the highest logistic modeling at 80.4%. Assuming a carrier switching cost of approximately 105,000 won and limiting the marketing cost budget to 100 million won, we found that the expected return would be improved by 22% to 114% as a result of benchmarking expected returns by using four different customer retention marketing (i.e. subsidy payment, coupon, free data) approaches (total budget/N, customer monthly rate ratio, and

customer monthly rate*departure rate) to reduce the probability of customer departure. In addition, experimenting with increasing the cost budget to 1 billion won showed the highest expected return on the application of optimization techniques. In particular, the actual application may be difficult if optimized computational performance time increases in the form of an index, but there was no problem with the calculation in practice because the experimental results showed that the complexity varies linearly with respect to the number of samples or marketing costs.

Considering that an 'existing customer retention strategy' that costs relatively less than number portability can continue to be competitive while reducing marketing costs in the face of the low influx of new subscribers due to smartphone market saturation, the mobile communications marketing strategy/policymaker can use this study to provide a strategic direction to prevent customers from leaving the market while efficiently using marketing costs.

Keywords : machine learning, optimization, retention, marketing cost
Student Number : 2018-20096