Logistic Regression을 이용한 이탈고객예측모형
Churn Prediction Model using Logistic Regression

정한나, 박혜진, 김남형, 전치혁, 이재욱
포항공과대학교 산업경영공학과
{aporia,iouhj,skagud,chjun,jaewool}@postech.ac.kr

Abstract

금융산업에서 고객의 이탈비율은 기대수익에 영향을 미친다는 점에서 예측이 필요한 부분이며 최근 들어 정확한 예측을 통한 비용관리가 이루어지면서 고객 이탈을 예측하는 것이 중요한 문제로 떠오르고 있다. 그러나 보험 고객 데이터가 대용량이고 불균형한 출력 값을 갖는 특성으로 인해 기존의 방법으로 예측 모델을 만드는 것이 적합하지 않다. 본 연구에서는 대용량 데이터를 처리하는 데 효과적으로 알려져 있는 Trust-region Newton method를 적용한 로지스틱 회귀분석을 통해 이탈고객을 예측하는 것을 주된 연구로 하며, 불균형한 데이터에서의 예측정확도를 높이기 위해 Oversampling, Clustering, Boosting 등을 이용하여 고객 데이터에 적합한 이탈 고객 예측 모형을 제시하고자 한다.

Keyword: Imbalanced Large-scale Data, Logistic Regression.

1. 서론

최근 금융 환경 위기 및 경기 불황으로 인하여 보험 계약이탈 및 신계약 감소현상이 나타나고 이에 대한 경쟁력 확보방안이 요구되고 있다. 시장의 움직임 또한 가격 및 상품 자유화, 진입 장벽 완화 등의 규제 완화 양상을 띄고 있으며 고객이 보험 회사의 수익과 직결되어 고객의 Bargaining power가 강화되고 있다. 따라서 전략적으로 고객과 시장을 세분화하여 Target 마케팅을 추진하는 것이 중요하다. 상품, 채널, 고객, 시장 분석을 통해 전략 수립 기능을 강화하고 고객 정보를 이용한 체계적인 영업 활동 지원을 통해 보유 고객을 유지하고 연계 가입을 유인하는 등 접촉 고객에 대한 로열티를 증대시킴으로서 고정 고객화해야 한다. 현재 보험사들의 CRM 활용 단계에 있어 데이터 마이닝을 통해 고객 분석을 하고 미리 이탈가능 고객을 예측함으로써 신뢰할 만한 결과를 얻어내고 있다.

고객의 이탈을 예측하려는 시도는 Churn prediction부분에서 자주 일어나고 있는 부분이며 이탈 고객에 대한 스코어링을 통해 기존 우량고객의 이탈을 미리 예측하고 기대수익 예측에도 도움이 될 것이라는 점에서 필요한 연구라 할 수 있겠다. 특히 이탈 고객에 대한 예측 모델을 구축함으로써 고객 이탈을 방지하고 이탈에 대한 재가입을 유도할 수 있어 그 기대효과가 크다.

그러나 예측 모델을 만들기 위한 학습데이터는 매우 불균형한 출력 값을 가질 뿐 아니라 너무 방대하기 때문에 기존의 방법으로는 효과적인 모델을 만들기 매우 어렵다. 본 연구의 분석에 사용한 데이터도 전체 데이터 중에서 약 4%정도만 이탈 고객에 해당하는 관측치로 매우 불균형적인 정보를 담고 있다. 또한 고객 정보 데이터의 Feature의 개수가 많아 Curse of dimensionality 문제를 야기 시킬 수 있으며 Numerical-valued 데이터와 Categorical-valued 데이터가 섞여 있는 형식으로 되어 있어 분석하기 매우 까다롭다.

이러한 문제점들을 극복하고 예측 모델을 만들기 위해 대용량 데이터를 처리하는 데 효과적으로 알려져 있는 Trust-region Newton method를 적용한 로지스틱 회귀분석 기법을 사용하였으며 불균형한 데이터에서의 예측정확도를 높이기 위해 Oversampling, Clustering, Boosting 등을 이용하였다.

2. 기존연구

2.1. Logistic Regression

Logistic regression model은 Two-class classification에 유용한 Tool의 하나로 가장

널리 사용되는 분류모형이다. 데이터에 대한 다변량 정규분포의 가정이 필요 없고, 설명변수로 연속형과 범주형자료를 다 포함시킬 수 있다는 장점을 가지고 있다. 본 연구에서 이탈고객예측에 사용한 데이터는 반응변수가 이항변수이며, 다양한 형태의 자료를 변수로 가지고 있으며, 각 변수의 분포가 고르지 못하므로 Logistic regression model이 적합할 것으로 판단되었다.

반응변수 y가 1과 0의 값을 가질 때 Logistic regression model은 다음과 같은 Probability model을 가지고 있다.

$$P(y=1)=\frac{1}{1+\exp(-(w^T x+b))}$$

여기에 Bias term인 b를 Weights w에 포함시키고 Negative log-likelihood를 구하고 Regularized term을 더해주면 다음과 같은 목적함수를 갖는 Regularized logistic regression form이 나오게 된다.

$$\min_w f(w) \equiv \frac{1}{2} w^T w + C\sum_{i=1}^{l} \log(1+e^{-w^T x_i})$$

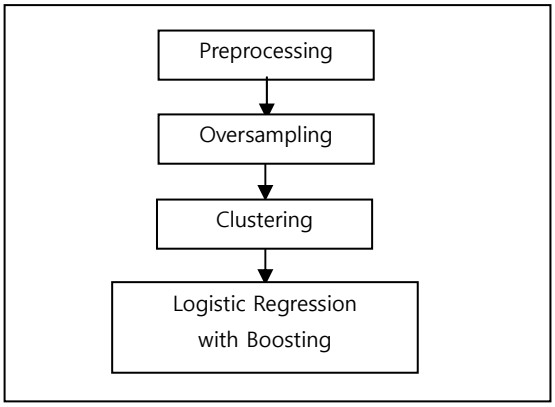
Logistic regression model을 풀기 위한 Optimization method들은 여러 가지가 있다. 그 중에서도 Truncated Newton method가 대용량 데이터의 분석에 가장 효과적인 것으로 알려졌으나 Logistic regression에서의 적용은 잘 이루어지지 않았다. 그러나 최근의 연구에 따르면 Logistic regression analysis에 Truncated Newton method를 적용한 Trust-region Newton method가 지금까지 가장 효과적이고 효율적이라고 알려진 LBFGS method보다 더 빠른 속도를 보이면서 비슷한 정확도를 보여 더 나은 방법이라는 연구 결과가 나왔다[1]. 이 Method는 Approximate direction을 사용하지만 Exact Hessian matrix를 사용하여 초기에 속도를 높이면서 정확한 결과가 나오게 되어 있다. 반면에 LBFGS method는 Approximate inverse Hessian matrix를 사용하여 Trust-region Newton method보다 느린 수렴 속도를 보인다. 본 연구에서는 대용량 데이터의 Logistic regression analysis에 Trust-region Newton method를 적용해 봄으로써 좀더 빠르고 정확한 결과를 내도록 하였다.

2.2. Imbalanced data

본 연구에서 사용한 고객정보 데이터는 이탈비율이 전체 고객 데이터의 약 4%를 차지하는 Imbalanced data로 정확한 예측을 하기가 까다롭다. 기존의 이탈고객 관리(Churn management) 에 관한 연구를 살펴보면 주로 이동통신에 관한 선행연구가 주를 이루었으며 Regression, Neural network, Decision tree등의 Data mining 기법들이 주로 사용되어왔다[2]. Imbalanced data를 처리하는 기법에는 Weighted accuracy또는 ROC curve를 구하거나 데이터를 Oversampling하는 기법들이 적용될 수 있다[3]. 본 연구에서는 이러한 기법들 중에서 보험고객 이탈 데이터 특성에 잘 맞는 기법들을 사용한 연구를 제안하고자 한다.

3. 제안 분석절차

고객정보 데이터의 Large-scale이며 Imbalanced class라는 특성상 기존의 Logistic regression의 방법으로는 분류가 쉽지 않아 추가적인 데이터마이닝 기법을 사용하여야 한다. 본 연구에서는 Preprocessing, Oversampling, Clustering, Logistic regression, Boosting을 사용하여 분석하였으며 개요는 [그림 1]과 같다.



[그림 1] 제안된 분석절차

3.1. Preprocessing

데이터의 특징을 파악하고 데이터 마이닝에 적합하도록 일부 변수에 정규화, 결측치 처리, 변수 재정의를 실행한다.

3.2. Oversampling

데이터의 Target class가 매우 작아 Logistic regression을 통해 분류할 때 제대로 된 결과를 낼 수 없으므로 Oversampling을 통해 30%까지 Target class의 데이터를 늘리는 것이 필요하다[4]. 본 연구에서는 Oversampling 방법 중 Data duplication을 이용하며 이는 Imbalance 의 정도가 매우 심할 경우 한 Class를 Outlier와 구별하지 못하는 단점을 해결할 수 있다.

3.3. Clustering

고객의 특성이 매우 다양하여 분류분석에 필요한 Key factor를 찾기 힘든 점을 방지하기 위하여 K-means clustering을 이용하였다. 이는 비슷한 특징을 가진 고객군 중에서 이탈과 지속을 나누는 Factor를 좀 더 자세하게 관찰할 수 있도록 하여 각 Cluster별 정확도를 높이며 궁극적으로 전체 정확도를 높일 수 있다[4]. K-means clustering은 속도가 빠르며 이해와 구현하기 쉽다는 장점이 있다[3].

본래 Classification에 Clustering방법을 이용하는 것은 Semi-supervised learning에서 제안된 것으로 Transduction라 한다[5]. 본 연구는 Semi-supervised learning은 아니나 각 Cluster별 Classification model을 만들었을 때 정확도에 유의할만한 향상을 가져와 제안하는 바이다.

3.4. Logistic Regression

Logistic regression은 Large-scale에 효과적이라고 알려진 Trust-region Newton method를 사용한다. 이때 각 Parameter는 TRON software에 사용되는 η_0=10^(-4), η_1=0.25, η_2=0.75, σ_1=0.25, σ_2=0.5, σ_3=4.0을 이용한다[6].

3.4. Boosting

Boosting은 보통의 방법으로 분류가 어려운 데이터를 제대로 분류하기 위해 데이터의 중요도를 조절하여 반복해서 분류하는 방법으로 분류분석의 정확도를 높이기 위하여 부가적으로 쓰인다[3].

Boosting의 대표적 알고리즘인 Adaboost는 초기에는 동일한 중요도를 부여하고 반복하면서 데이터의 중요도를 점차 늘려나가며 Training한다. 이 때 잘못 분류된 정도가 큰 데이터에 큰 중요도를 부여한다.

본 연구에서는 Adaboost를 사용하였으며 반복은 20번 시행되었다. 기본 알고리즘에서 중요도를 Reset하는 부분을 삭제하고 재계산되는 중요도의 비율을 80%로 다소 크게 두었는데 이는 정확도가 50% 이하로 떨어질 경우 재계산이 필요한 데이터의 수가 많기 때문이다.

4. 결과

4.1. 데이터분석

고객정보 데이터를 분석결과 전체 40001개의 데이터 중 1557개의 이탈고객 데이터, 즉 Target 데이터가 있었으며 이는 3.89%로 극도로 Imbalance한 데이터에 속한다. 이 중 Target의 비율을 30%정도로 늘리기 위해 Duplication를 실행하였으며 그 결과 34.50%의 비율의 이탈고객이 생성되었다. Sample에 따른 이탈고객과 지속고객의 비율은 [표 1]과 같다.

[표 1] 이탈고객과 지속고객의 비율

	Original Data	Oversampled Data
이탈 고객	1557 (3.892%)	20241 (34.48%)
지속 고객	38444 (96.10%)	38444 (65.50%)
전체 개수	40001	58692

데이터전처리 과정으로 데이터 결측치 처리, 변수재정의, 추가독립변수 정의를 수행하였으며 결과적으로 38개의 변수를 가지고 분석을 진행하였다.

K-means clustering으로 전체 고객을 8개의 Cluster로 나누었으며 Target의 비율은 각 Cluster별로 비슷하였다[표2]. 이것으로 고객정보로 알 수 있는 특징별로 Target값이 민감하게 다르지는 않다는 것과 각 특성의 고객별로 비율의 고객이 이탈한 것으로 한 사실로 보아 각 군집내에서 고객의 이탈여부에 따라 Classification modeling을 하였을 때 좀 더 자세한 Model이 만들어질 수 있음을 알 수 있다.

[표 2] Cluster별 개수와 target값

Cluster no.	전체 고객 수	이탈고객 수 (이탈비율)
Cluster1	15275	6396 (41.87%)
Cluster2	8544	2145 (25.10%)
Cluster3	6296	2496 (39.64%)
Cluster4	3918	1495 (38.15%)
Cluster5	4955	1053 (21.25%)
Cluster6	4031	1092 (27.09%)
Cluster7	7337	2275 (31.00%)
Cluster8	8336	3289 (39.45%)
Total	58692	20241 (34.48%)

4.2. 시간

본 알고리즘은 Matlab과 C++을 연동하여 수행되었으며 159.7초가 소요되었다. 이것은 기존의 Large-scale 데이터 분석에 가장 효과적으로 알려져 있던 LBFGS로 돌렸을 때에 비해 10배 이상 빠른 것으로 LBFGS의 경우 1793.8초를 기록하였다. 본 결과는 TRON logistic regression이 LBFGS에 비해 시간상으로 우수한 결과를 보인다는 연구를 뒷받침한다[1].

4.3. 정확도

정확도 계산시 2-fold Cross validation을 사용하여 Overfitting에 따른 정확도의 왜곡을 막도록 하였다. 제안된 각 단계를 지남에 따른 정오분류표를 [표 3]에서 [표 6]까지 나타내었다. Row는 실제 고객상태를 의미하며 1은 이탈고객, 즉 Target이며 0은 지속고객이다. Column은 분류 결과를 의미하며 1은 이탈 고객으로 예측한 것이다. 각 비율은 원래 고객 Class와 동일하게 예측한 고객의 비율을 뜻한다. Target정확도, 즉 True positive는 Bold체로 나타내었다.

[표 3]은 Original data를 Logistic regression으로 분류한 결과다. 이때 Target정확도는 0.128%로 전체 1557개 Target중에 2개밖에 Target으로 분류하지 못한 것을 볼 수 있다. 여기서 0을 0으로 분류한 것이 99.97%로 높은 정확도를 보여 전체 정확도는 96.09%에 달하게 되는데 이것은 모든 데이터를 같은 Class에 분류한 결과이기 때문에 신뢰할 수 없다. 따라서 Target 정확도에 초점을 맞추어 진행하였다.

[표 3] Original data의 정오분류표

	1	0	Sum
1	**2 (0.128%)**	1555 (99.87%)	1557
0	9 (0.023%)	38435 (99.97%)	38444
Sum	11	39990	40001

Oversampling한 후 Logistic regression으로 분류했을 때의 Target정확도는 19.27%로 Oversampling을 하기 전에 비해 150배의 증가를 보였다[표 4]. Oversampling을 한 후의 이탈고객 데이터는 21241개이나 Original data의 이탈고객 분류 결과와 직접적인 비교를 할 수 있도록 1557개로 보정하여 정오분류표에 나타내었고 이때 비율에는 변함이 없다.

[표 4] Oversampling 후 정오분류표

	1	0	Sum
1	**300 (19.27%)**	1256 (80.72%)	1557
0	3259 (8.477%)	35185 (91.52%)	38444
Sum	3559	36442	40001

Oversampling을 한 후 각 Cluster별로 다른 Logistic regression model을 세워 정확도를 구한 결과는 [표 5]과 같다. 각 Cluster별로 구해진 Target정확도를 비율에 따라 환산하여 구하였으며 이때 Target정확도는 29.68%로 증가하여 Transduction이 Classification 문제에서도 효과를 보인 것을 확인할 수 있다.

[표 5] Clustering 후 정오분류표

	1	0	Sum
1	**462 (29.68%)**	1094 (70.31%)	1557
0	4997 (12.99%)	33447 (87.00%)	38444
Sum	3559	36442	40001

Oversampling과 Clustering을 거친 데이터를 Logistic regression을 이용하여 분류할 때에 Boosting을 사용하여 Iterative하게 정확도를 향상시킨 결과, Target정확도는 34.36%로 증가하였다[표 6]. 모든 정확도가 Test set의 정확도임을 감안할 때 Boosting의 단점이라 할 수 있는 Overfitting의 문제가 본 데이터에는 크게 영향을 주지 않았음을 알 수 있다.

[표 6] Boosting 후 정오분류표

	1	0	Sum
1	**535 (34.36%)**	1022 (65.63%)	1557
0	4447 (11.56%)	33997 (88.43%)	38444
Sum	3559	36442	40001

각 기법을 단계에 따라 적용하여 Target정확도와 그 증가율을 나타내었다[표 7]. 각 단계가 추가됨에 따라 Target정확도가 증가하였다. 특히 Oversampling을 하고 나서 150배 정도의 향상을 보였고 Clustering 후에는 이전 단계에 비해 53.9%, Boosting을 사용한 후에는 하기 전에 비해 15.8%의 증가를 보인 것을 확인하였다.

결과적으로 Target정확도는 34.36%까지 증가하였고 본 데이터가 실제의 고객정보를 담고 있는 것임을 고려할 때에 상당히 Acceptable한 결과임을 알 수 있다.

[표 7] 각 단계에 따른 정확도

	Target 정확도(%)	증가율 (%)
Original data	0.13	
Oversampling	19.28	14730
Clustering	29.68	53.9
Boosting	34.36	15.8

5. 결론

본 연구에서는 Logistic regression을 이용하여 이탈고객을 예측하는 방법을 제안하였다. 고객정보데이터의 특성이 Large-scale이며 Imbalance하다는 점에서 이에 적합한 분석방법을 제안하였다는 데 본 연구의 의의가 있다.

본 연구의 장점은 Large-scale matrix에 적합한 Logistic regression을 위해 TRON method를 이용하여 다른 방법보다 Computational time을 줄였다는 것과 고객정보 데이터와 같은 Imbalanced data를 분류하는 체계적인 방법을 제안하였다는 것에 있다. Oversampling, Clustering, Boosting등의 Technique이 사용되었고 결과적으로 Target 정확도를 높이는데 성공하였다.

Logistic regression은 해석하기 편하고 널리 사용되는 Classification방법이지만 동시에 Target정확도와 더불어 전체 정확도를 향상시키는 데는 한계가 있을 수 있는데 이것은 Linear method라는 특징 때문이다.

Logistic regression이 사용자에게 사용하기 쉽고 이해하기 편한 결과를 가져온다는 것을 고려할 때 Logistic regression을 이용한 고객 Classification 문제는 좀 더 논의될 여지가 있다. 좀 더 Accurate한 결과를 위해서는 목적식을 푸는 기법을 본 연구에서 사용한 Regularized logistic regression대신 Bayesian logistic regression 또는 Kernel logistic regression 등을 이용하여 좀 더 효과적이며 정확도가 높은 알고리즘을 만들 수 있다.

참고문헌

[1] Chih-Jen Lin, Ruby C. Weng, S. Sathiya Keerthi, “Trust region newton method for large-scale logistic regression”, *Journal of Machine Learning Research* 9(2008) 627-650.

[2] Shin-Yuan Hung, David C. Yen, Hsiu-Yu Wang, “Applying data mining to telecom churn management”, *Expert Systems with Applications* 31(2006) 515–524.

[3] P.Tan and M. Steinbach and V. Kumar, “Introduction to data mining”, *Addison Wesley*, 2006.

[4] Hyunjin Heo, Hyejin Park, Namhyoung Kim and Jaewook Lee, “Prediction of credit delinquents using locally transductive Multi-layer perceptron”, *ISNN*, paper no.136(2008).

[5] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zine, “Semi-Supervised Learning”, *MIT Press*, 2006

[6] Chih-Jen Lin and Jorge J.More, “Newton’s method for large-scale bound constrained problems”, *SIAM Journal on Optimization*, 9:1100-1127, 1999.