JKIICE

Short Paper

한국정보통신학회논문지 Vol. 27, No. 8: 1016~1019, Aug. 2023

# 기계학습에 기반을 둔 통신사 고객 이탈예측 방안

# Machine Learning based Churn Prediction in Telecommunications

**Jae-Hyuk Huh[1] · Woongsup Lee[2*]**

[1]Graduate Student, Graduate School of Information, Yonsei University, Seoul, 03722 Korea
[2*]Associate Professor, Graduate School of Information, Yonsei University, Seoul, 03722 Korea

## ABSTRACT

In this paper, we explore customer churn in a telecommunication company, measuring the probability of users discontinuing their service. Our investigation employs various machine learning models, including Decision Tree, Random Forest, XGBoost, LightGBM, SVM, Logistic Regression, and Deep Neural Network. To accomplish this, we leverage data collected from a California-based telecommunication company, initially containing 38 feature data. In order to reduce complexity, we apply Lasso regression to select the five most crucial features for determining customer churn: household situation, service satisfaction, loyalty, payment capability, and contract type. Through performance evaluation, we demonstrate that accurate predictions of customer churn can be achieved even with five features, emphasizing the significance of feature selection.

Keywords : Customer churn prediction, LASSO, deep learning, machine learning, feature extraction.

## Ⅰ. 서 론

최근 Mobile Virtual Network Operator (MVNO)를 비롯한 여러 통신사업자가 서비스를 제공함에 따라서 더 나은 서비스를 위해 고객들이 통신사를 옮기는 일이 많이 발생하고 있다 [1]. 이렇게 고객이 서비스 제공업체를 떠나 다른 업체로 이동하는 고객 이탈 (Customer churn)은 통신사의 수익성에 큰 영향을 미친다.

이러한 고객 이탈 예측관련 연구는 통신 산업뿐만이 아니라 E-commerce, 금융업, 보험 등 다양한 기존 산업 분야에서 진행되었다 [2]. 최근 빅데이터 분석의 발달을 기반으로 다양한 기계학습 알고리즘을 활용한 통신사 고객 이탈 예측 연구가 진행되고 있다. [3]의 연구에서는 기계학습의 한 방식인 랜덤 포레스트를 이용하여 88%의 확률로 통신사 고객 이탈을 예측할 수 있음을 보였다. 또한 [4]의 연구에서는 완전연결신경망을 활용하여 73.9%의 확률로 통신사 고객 이탈을 예측할 수 있었다.

통신사 고객 이탈을 예측하기 위한 기존 연구에서는 이탈 관련 많은 특징 데이터 (feature data)를 활용하였다. 예를 들어 [5]의 연구에서는 2,000여개의 특징 데이터를 기반으로 고객 이탈 예측을 수행하였다. 이렇게 많은 양의 특징 데이터를 이용하여 고객 이탈을 예측하면 예측 정확도는 높아질 수 있지만 과적합 (over-fitting) 문제가 발생할 수 있다. 더불어 실제 어떠한 요소가 고객 이탈에 영향을 주는지 인사이트 (insight)를 얻기 어려우므로 통신사의 고객 이탈 방지 전략을 세우는 데에도 활용되기 어려울 수 있다. 이러한 이유로 최근 머신러닝을 이용한 연구에서는 LASSO regression등을 이용하여 활용 변수의 개수를 줄이려는 시도가 이루어지고 있다 [6].

본 연구는 미국 캘리포니아에 위치한 통신사의 2022년 2분기 고객 데이터를 활용하여 고객 이탈 예측 모델을 개발하였다. 특히 로지스틱 회귀 (Logistic regression, LR), 서포트 벡터 머신 (Suppert Vector Machine, SVM), 결정트리 (Decision Tree, DT), 랜덤 포레스트 (Random Forest, RF), XGBoost (XGB), LightGBM (LGBM)과 같은 기계학습 알고리즘 및 심층 신경망 (Deep Neural Network, DNN) 기반 알고리즘을 고려하였다. 더불어 Lasso regression을 활용하여 실제 고객 이탈에 가장 큰 영향을 주는 5가지 요소를 도출하였고, 성능 분석을 통해서 이 5가지 요소만을 이용하여도 86%의 확률로 통신사 고객 이탈을 예측할 수 있음을 확인하였다.

**Received** 20 July 2023, **Revised** 29 July 2023, **Accepted** 2 August 2023

*** Corresponding Author** Woongsup Lee (E-mail:woongsup.lee@yonsei.ac.kr, Tel:+82-2-2123-4525)
Associate Professor, Graduate School of Information, Yonsei University, Seoul, 03722 Korea

Open Access http://doi.org/10.6109/jkiice.2023.27.8.1016 print ISSN: 2234-4772 online ISSN: 2288-4165

## Ⅱ. 활용 통신사 데이터 셋 및 전처리

본 연구에서는 미국의 캘리포니아에 위치한 통신사에서 수집된 고객 이탈 데이터를 활용하였다 [7]. 해당 데이터는 2022년 2분기동안 수집된 7,043개의 인스턴스(instance)로 이루어져 있으며, 454명의 가입 고객 및 1869명의 이탈고객 데이터를 포함하고 있다. 각 인스턴스는 고객의 행동, 선호도, 그리고 가치를 반영하는 38개의 특성을 포함하고 있다. 특성에는 고객의 전화 서비스 사용 여부, 다중 전화선 사용 여부, 프리미엄 기술 지원 서비스 이용 여부, TV 스트리밍 및 영화 구독 여부 등 고객의 통신사 서비스 이용 패턴과 선호도를 반영하는 정보가 포함되어 있다. 또한, 고객의 가치를 나타내는 중요한 지표인 이용 기간, 계약 유형, 결제 방법, 월별 요금 등의 계정 정보도 포함한다. 더불어 고객의 개인적 특성과 생활 상황을 반영하는 성별, 연령, 결혼 여부, 부양가족 수와 같은 인구 통계학적 정보도 포함하고 있다.

라벨 데이터로 활용되는 고객의 가입상태 정보(Customer status)는 Stay, Churn, 및 Joined의 3가지로 이루어져 있는데, 본 연구에서는 고객의 이탈 여부를 예측하는 방안을 고려하므로 Stay와 Churn 라벨을 지닌 데이터만 분석에 활용하였다. 본 분석에서는 고객 이탈을 나타내는 Churn을 1, 고객 유지를 나타내는 Stayed를 0으로 매핑하였다.

기계학습 분석을 위해서 성별 및 기혼 여부와 같은 범주형 데이터는 원-핫-인코딩 (one-hot-encoding)을 이용하여 전처리를 수행하였고 월별요금 및 이용 기간과 같은 수치형 변수들은 평균은 0, 표준편차를 1로 같은 값으로 정규화를 하여 수치형 변수들을 같은 척도로 비교할 수 있게 하였고 효율적 특징 추출이 가능케 하였다.

본 연구에서는 전체 데이터셋을 K개의 동일한 크기의 부분집합으로 나눈 후, 각 부분집합을 테스트 데이터셋으로 사용하고 나머지 부분집합을 훈련 데이터셋으로 활용하는 K-Fold 교차 검증 기법을 사용하여 성능을 검증하였다. 성능 분석에서는 K의 값을 5개로 설정하였다.

## Ⅲ. 기계학습 분석 및 특징선택

성능 검증을 위해서 이분 분류 문제 (Binary classification)을 위한 LR, SVM, DT, RF, XGB, LGBM, DNN 기계학습 모델을 활용하였다. LR은 로지스틱 함수를 사용하여 분류를 수행하고 SVM은 특징 데이터를 고차원 공간에 매핑한 후, 이 공간에서 최적의 결정 경계를 찾는 방법으로 분류를 수행한다. DT는 트리구조를 이용하여 분류를 수행하며 RF는 여러 개의 트리구조를 결합한 앙상블 기법을 이용하여 분류를 수행한다. XGB와 LGBM은 모두 Gradient Boosting 알고리즘을 활용한 머신러닝 모델로, 여러 개의 DT기반 예측 모델을 조합하여 분류를 수행한다. 마지막으로 DNN은 심층신경망을 이용한 분류 방안으로서 본 연구에서는 히든노드 수가 62개이고 5계층을 지닌 완전연결신경망 기반의 심층신경망을 고려하였다.

본 연구에서는 Lasso regression을 활용하여 전체 38개의 특징 중 가장 큰 영향을 미치는 요소들을 추출하였다. Lasso Regression은 회귀 계수의 정규화를 고려한 회귀방안의 한 종류로서, 회귀 계수의 절댓값에 대한 패널티를 부여한다. Lasso Regression의 목적함수는 다음의 수식과 같이 표현될 수 있다.

$$L(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \qquad (1)$$

여기서 Y는 라벨 데이터, X는 특징 데이터, β는 회귀계수를 의미하며 $\|\cdot\|_1$ 은 L1-norm을 나타낸다. λ는 Lasso Regression의 회귀 계수 규제 강도를 조절하는 파라미터로서 λ값을 증가시키면 더 많은 회귀 계수가 0으로 수렴한다. 본 연구에서는 λ값을 0.05로 설정하였으며 Lasso Regression을 통해서 얻은 회귀 계수 중 가장 큰 5개의 특징 데이터를 선택하였다. Lasso regression을 통해 얻어진 5개의 변수 들은 다음과 같다.

- Number of Dependents (의존인수): 고객과 함께 사는 의존인수로 자녀, 부모 등이 해당함. 고객의 가정 상황과 부양 의무에 대한 중요한 정보를 제공함.
- Number of Referrals (추천수): 고객이 지금까지 친구나 가족을 회사에 추천한 횟수를 나타냄. 이는 고객의 서비스 만족도와 그들의 추천 의향성을 보여줌.
- Tenure in Months (사용월수): 고객이 회사와 함께한 총 개월 수를 나타냄. 이 특징 데이터는 고객의 충성도와 장기적인 이용 여부를 보여줌.
- Monthly Charge (한달청구요금): 고객이 회사로부터 받는 모든 서비스에 대한 월별요금을 나타냄. 이는 고

객의 결제 능력과 서비스 가치 인식을 보여줌

- Month-to-Month (장단기계약): 월별 계약 여부를 나타내는 지표. 고객이 월별로 서비스 계약을 갱신하는지, 장기 계약을 체결하는지에 대한 정보를 제공.

## Ⅳ. 성능 분석 결과 및 논의

성능 분석에서는 전체 특징 데이터를 활용한 기계학습 방안과 Lasso Regression을 통해서 도출된 5개의 대표 특징들만을 활용한 방안의 성능을 비교하였다. 성능 분석척도로 정확도 (Accuracy), 정밀도 (Precision), 재현율 (Recall), F1 점수 (F1 Score) 및 수신자 조작 특성 곡선 하의 면적(AUC)을 활용하였다. 각 환경의 성능 분석 결과는 표 1과 2에 각각 나와 있다.

**Table. 1** Performance with full feature data.

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| LR | 0.86 | 0.75 | **0.74** | 0.74 | 0.91 |
| SVM | 0.86 | 0.78 | 0.71 | 0.74 | 0.91 |
| DT | 0.83 | 0.70 | 0.70 | 0.70 | 0.79 |
| RF | **0.87** | **0.83** | 0.67 | 0.74 | 0.92 |
| XGB | **0.87** | 0.8 | 0.71 | 0.75 | 0.92 |
| LGBM | **0.87** | 0.81 | 0.72 | **0.76** | **0.93** |
| DNN | 0.85 | 0.76 | **0.74** | 0.74 | 0.91 |

실험 결과, 정확도 측면에서 RF, XGB, 및 LGBM 모델이 모두 0.87로 가장 큰 값을 보이는 것을 확인할 수 있었다. 또한, DT 모델이 0.83으로 가장 낮은 정확도를 보였지만 RF, XGB 및 LGBM과 비교하여 크게 떨어지지 않는 정확도를 보이는 것을 볼 수 있다. 이를 통해서 기계학습을 이용하여 높은 정확도로 통신사 고객 이탈을 예측하는 것이 가능한 것을 확인할 수 있다. 또한, 전반적으로 보았을 때 LGBM방식이 가장 높은 성능을 보이는 것을 볼 수 있다. 더불어 정밀도보다 재현율이 낮은 것을 확인할 수 있는데 이는 데이터의 비대칭성 (이탈 고객 데이터가 더 많음)에 의해서 발생하는 현상이다.

**Table. 2** Performance with 5 selected feature data.

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| LR | 0.84 | 0.71 | 0.72 | 0.71 | 0.90 |
| SVM | 0.85 | 0.74 | 0.75 | 0.74 | 0.90 |
| DT | 0.83 | 0.70 | 0.70 | 0.70 | 0.79 |
| RF | 0.85 | 0.76 | 0.69 | 0.72 | 0.91 |
| XGB | **0.86** | 0.78 | **0.72** | **0.75** | 0.92 |
| LGBM | **0.86** | **0.79** | 0.71 | 0.74 | **0.93** |
| DNN | 0.83 | **0.79** | 0.67 | 0.68 | 0.90 |

5개의 특징만을 활용한 표 2의 결과를 통해서 전체 특징 데이터를 활용했을 때에 비해 전반적인 성능이 조금 감소한 것을 확인할 수 있으나 그 정도가 0.01 정도로 매우 미미한 것을 확인할 수 있다. 이를 통해서 Lasso regression에 의해 선택된 5개의 변수가 통신사의 고객 이탈을 예측하는 데 있어서 중요한 요소인 것을 확인할 수 있고, 이러한 대표 특징 추출을 통해 기계학습 기반

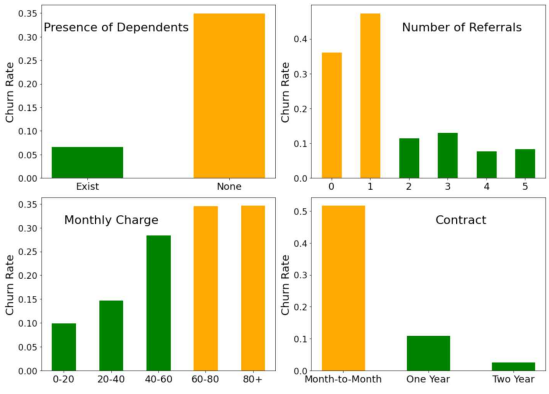


**Fig. 1** Illustration of relation between 4 features and churn rates

예측의 효율성을 개선하는 동시에 적은 특징 사용으로 인한 성능 저하를 최소화할 수 있음을 확인할 수 있다.

마지막으로 그림 1에는 DT를 활용한 예측모델에서 의존인수, 추천수, 한 달 청구요금, 장단기계약에 따른 고객 이탈 예측 확률 결과를 보여준다. 사용월수 특징의 경우 장단기계약 특징과의 상관계수가 0.65로 높은 관련성을 보여주므로 고려하지 않았다.

우선 예측 결과를 통해 의존인수가 없는 경우에 이탈률이 35%로 의존인수가 있는 경우의 평균인 6.5%에 비

해 5배가 넘는 유의미하게 높은 수치를 보이는 것을 확인할 수 있다. 따라서 혼자 사는 고객들에 대해서 통신사 이탈을 막기 위한 전략이 필요하다는 것을 알 수 있다. 또한 추천수가 적은 고객의 이탈률이 약 40%를 나타내고 추천수가 2 이상이 경우에는 고객 이탈률이 급격하게 감소되는 것을 확인할 수 있다. 이는 고객 만족도와 고객 이탈률 사이의 연관성을 시사한다. 더불어 한 달 청구 요금과 이탈률이 비례하는 것을 확인할 수 있으며, 특히 60 달러 이상의 요금이 청구된 경우는 35%의 높은 확률로 고객이 이탈하는 것을 확인할 수 있다. 즉, 높은 서비스 요금을 부담하는 고객을 유지하기 위한 전략이 필요할 수 있음을 시사한다. 마지막으로 월 단위 계약자의 경우, 51.7%의 이탈률을 보였지만 1년 단위의 경우 10.9%, 2년 단위 계약자의 경우 2.6%로 낮아지는 것을 확인할 수 있다. 장기 계약자일수록 이탈률이 급격히 낮아지는 결과를 보이므로, 이탈률을 줄이기 위해서 장기계약을 유도하는 것이 바람직함을 확인할 수 있다.

## Ⅴ. 결론

본 연구에서는 기계학습을 활용한 통신사 고객 이탈 예측 방안을 제안하였다. 특히 Lasso Regression을 통한 대표 특징 선택을 통해서 전체 38개의 특징 중 5개의 핵심 특징들을 도출하였다. 미국 캘리포니아에 있는 통신사업자 데이터를 활용한 성능검증을 통해서 기계학습 기반 예측모델을 활용하여 통신사업자의 고객 이탈을 정확하게 예측할 수 있음을 보였고 5개의 핵심 특징들만을 이용해서도 고객 이탈을 정확하게 예측할 수 있음을 확인하였다. 본 연구의 결과는 통신사업자들이 고객 이탈을 미리 인지하고 대응 전략을 세우는데 이바지할 수 있다. 후속 연구에서는 제안 방안의 다양한 산업 분야로의 확장과 개선된 특징 선택 방법의 개발을 수행할 계획이다.


## ACKNOWLEDGMENTS

This work was carried out with the support of "Cooperative Research Program for Agriculture Science & Technology Development (Project title: Advancement of Hanwoo beef tracing and management technique based on data, Project No. PJ0170202022)" Rural Development Administration, Republic of Korea.


## REFERENCES


[1] J. Ahn, S. Han, and Y. Lee. “Customer Churn Analysis: Churn Determinants and Mediation Effects of Partial Defection in the Korean Mobile Telecommunications Service Industry,” *Telecommunications policy*, vol. 30, no. 10, pp. 552-568, Nov. 2006. DOI: 10.1016/j.telpol.2006.09.006.

[2] J. Ahn, J. Hwang, D. Kim, H. Choi and S. Kang, “A Survey on Churn Analysis in Various Business Domains,” *IEEE Access,* vol. 8, pp.220816-220839, Jan. 2020. DOI: 10.1109/ACCESS.2020.3042657.

[3] I. Ullah, , B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. Kim, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,” *IEEE Access*, vol. 7, pp. 60134-60149, May 2019. DOI: 10.1109/ACCESS.2019.2914999.

[4] S. W. Fujo, S. Subramanian, and M. A. Khder, “Customer Churn Prediction in Telecommunication Industry Using Deep Learning,” *Information Sciences Letters,* vol. 11, no. 1, pp. 185-198, Jan. 2022. DOI: 10.18576/isl/110120.

[5] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform,” *Journal of Big Data,* vol. 6, no. 28, pp. 1-24, Mar. 2019. DOI: 10.1186/s40537-019-0191-6.

[6] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques,” *IEEE Access*, vol. 9, pp. 19304-19326, Jan. 2021. DOI: 10.1109/ACCESS.2021.3053759.

[7] S. L. Zhuang, Telecom Customer Churn Prediction, 2023. [Internet]. Avaiable:https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics