딥러닝 알고리즘에 기반한 퇴원 학생 예측모델 비교

고영상*, 임희석
*고려대학교 빅데이터융합학과
e-mail : koys007@korea.ac.kr

Comparison of Student Churning Prediction Models based on Deep Learning Algorithms

Young-Sang Ko*, Heui-Seok Lim
*Dept. of Big Data Convergence, Korea University

요 약

교육열이 강한 우리나라에서는 사교육은 언제나 뜨거운 감자이다. 교육대상 연령층의 인구수가 1990 년부터 빠르게 감소하기 시작했으며, 2005 년을 전후로 초등학생 수의 감소가 더욱 빨라지고 있다. 통계청 데이터에 따르면 2016 년 출생아 수는 40 만 6 천여명에서 2017 년은 35 만 7 천여명으로 향후에도 지속적으로 줄어들 추세이다. 이렇듯 매년 학생수가 감소함에도 불구하고 2018 년 사교육비 총액은 19 조 5 천억수준으로 2017 년 18 조 7 천억보다 8 천억원이 늘어 났다. 학생수는 전년보다 2.5% 줄었지만 사교육비는 반대로 4.4% 늘어났다. 이렇듯 사교육 시장이 심화 되게 되면 경쟁은 더욱 치열해 질 수 밖에 없으며 이 경쟁에서 살아 남기 위해서는 다양한 비즈니스 전략이 필요하며 특히 학생들의 이탈을 줄이는 것은 사업의 가장 중요한 포인트라고 볼 수 있을 것이다. 학원에서의 학생이 퇴원을 하는 이유에 대한 영향도를 분석하고 그 영향도 분석을 통해 학원 학생들의 퇴원 방지에 활용하고자 한다. 본 논문의 주요 연구 내용은 사교육을 대표하는 국내 사설 학원에서의 성적, 출결사항 및 학원 상담 내역 등의 다양한 학원 데이터들을 최적의 딥러닝 알고리즘 분석을 통한 퇴원 학생을 사전 예측하기 위한 논문임을 밝힌다.

1. 서론

출생률이 낮아 지고 한해 30 만명의 신생아도 태어나지 않는다는 기사가 쏟아지는 현 상황에서 학원 비즈니스 종사자들은 비즈니스 위기 의식을 느끼지 않을 수 없는 상황이다. 아이러니하게도 최근 교육부의 발표된 내용을 보면 2018 년 사교육비 총액이 20 조원을 육박하며 2 년 연속 상향세를 유지하고 있다. 또한 학생 1 인당 월평균 사교육비는 30 만원에 근접하며 역대 최고치를 기록했다. 이런 시장 상황에 변화를 초래하는 요인들이 증가 하는 시점에는 학생수에 민감하게 영향을 받은 경우 비즈니스의 안정성을 유지하는 방법을 찾는 것이야 말로 비즈니스의 승패를 좌우 할 수 있을 것이다. 이에 본 연구에서는 학원관리 시스템에 등록된 학생의 각 개인별 성적데이터, 출결데이터 및 학생 상담내역 데이터의 분석을 통한 이탈(퇴원) 학생 예측 모델을 만들어 보았다.

2. 예측 모형 구현 방법

2.1 본 연구의 방향

본 연구의 데이터 셋은 국내 국내에서 직영 및 프랜차이즈 학원을 운영하는 회사의 학생들을 관리하는 학원 관리시스템의 학생 데이터를 분석에 사용하였다. 학생 이탈의 예측을 위해 학생의 성적 데이터, 출결 데이터, 학생 학부모의 학생 상담 내역 데이터를 이용하여 예측 모델의 작성을 목표로 하였고 성적과 출결 데이터는 연속형(수치형) 데이터이고 상담내역 데이터는 텍스트기반의 이산형 데이터이므로 직접 데이터를 예측모델에 활용할 수 없어 수치형 데이터로 변환하기 위해서 텍스트 마이닝의 감성분석(Sentimental Analysis)모델을 적용하여 나온 수치형 데이터를 통해 예측모델을 생성하였다. 학생 상담내역을 학생 퇴원율 예측을 위해 이용한 이유는 이탈-항의 이론(Exit-Voice Theroy)의 그 사상적 배경을 이용하여 분석에 응용하였다. 학원의 경우 고객센터의 상담과는 다르게 학원의 관리데스크에서 상담한 내용을 기반으로 했으므로 기존 연구에서의 고객센터와는 성격이 다른 상담 내역들이 존재한다. 상담 내용에는 은유적인 표현들도 많이 있고 적극적인 의사 표현을 하지 않는 경우 들이 존재하기도 한다. 이에 감성분석(Sentimental Analysis)의 방법을 통해서 상담내역의 긍정적 상담과 부정적 상담의 내용으로 분류하고 지속적으로 부정적인 상담 이후에 학원 퇴원을 결심한다고 이탈-항의 이론을 가정하였다.

2.2 이론적 배경: 이탈-항의 이론

이탈 이론은 고객과 조직 간의 상호 관계성에 있어 기업 브랜드, 상품 및 서비스의 개선, 기업의 사회적 관계성, 고객 관리의 비용, 미래 잠재적 가치(손실) 등을 파악하는 기업 경영의 중요한 영역임을 인지하고, 다양한 고객 목소리를 관리하는 기업의 행태에 변화 연구하는 이론적, 실증적 체계라 할 수 있다. 이탈-항의 이론에 따르면 항의 (Voice)는 문제 해결을 한 극인 행동으로, 이탈(Exit)은 문제 해결을 한 노력 없이 그 상황을 회피하는 것으로 본다[1]. 이탈-항의 이론을 이론적으로 소비자 불만 행태 연구와 마케팅 채널 관계의 이론을 바탕으로 마케팅 측면을 고려해 보면 소비자들의 불만에 대한 이탈은 소비거부의 행태로 마켓을 이탈하는 형태의 행위가 나타나고, 항의는 기업과 상품에 대한 불만을 가지고 마켓에서 소비자의 목소리가 높아지는 것이라고 정의하고 있다. 본 연구에서는 이탈-항의 이론을 이론적 바탕으로 고객(학생 혹은 학부모)이 학원 서비스에 불만족했을 때의 피드백들, 즉 성적저하, 결석, 상담에 의한 불만 표출 등의 피드백으로 소극적 혹은 적극적인 항의를 할 것으로 가정하고 그 항의의 노력의 결과가 좋지 못할 때는 이탈이라는 극단적인 의사결정을 내릴 것이라는 가정을 하였다.

2.3 예측 모형 도출 방법

본 연구에서는 이탈-항의 이론을 바탕으로 고객(학생)이 학원서비스로부터 가질 수 있는 본질적인 불만사항이 되는 여러 가지 요인들, 즉 그 요인들로부터 파생되는 현상들의 데이터인 학생의 퇴원으로 이어지는 이탈에 영향을 미치는 각 학생들의 성적, 출결 등과 학생 학부모의 상담이력 중 부정적인 영향을 미치는 상담내역 데이터를 분석하고 학생 이탈에 대한 학원 관리 시스템의 결과를 최종 변인으로 하여 예측 모델을 작성 하였다. 학원관리시스템에 저장된 학생 성적, 출결, 상담 이력 데이터 들의 지속적으로 학원을 유지 하는 학생과 학원을 더 이상 다니지 않고 퇴원을 하는 학생의 데이터 조합으로 전체 대상이 되는 데이터(6,944 명)의 80%(5,555 명)는 훈련용 데이터로, 나머지 20%(1,389 명)는 테스트용으로 사용하였다. 예측 모형은 분류 예측 모델 분석에 주로 사용하는 Logistic Regression, SVM, DNN 등의 3 개의 모델을 사용하였고 Logistic Regression 모델은 epoch 를 10,000 수행, SVM 은 Gaussian Kernel 과 Sigmoid Kernel 의 두 가지 방식으로 분류 예측을 수행하였고 DNN 모델의 경우 batch size 를 100, global_step 을 10,000 번으로 모델 훈련을 수행 하였다.

3. 분석 데이터

본 연구에서는 국내에서 직영 및 프랜차이즈 학원을 운영하는 회사의 학생들을 관리하는 학원 관리시스템의 학부모 상담 내역 데이터를 분석에 사용하였다. 학원 사업의 경우 학생들의 세부 관리를 위해서 고객센터가 아닌 각각의 개별 학원(분원)에서 학원 선생님들과 학원 관리직원들이 학생의 성적, 출결은 물론 학부형들과 상세한 상담을 진행하며 그 상담 내역에 대해서 시간대 별로 기록을 해 놓는다. 상담 내역은 일상적인 내용부터 성적, 생활 등과 같은 내용은 물론 불만사항들까지 학부모들과의 상담 내역을 상세히 기록해 놓는다. 이 데이터 중 학생들의 불만과 연계될 수 있는 데이터를 이용하여 모델을 작성하였다.
데이터는 최종 3 개의 데이터를 활용하였고 최근 성적 데이터, 출결이력 데이터, 최근 상담이력 데이터를 가공하여 모델을 생성하였다. 정확도 향상을 위해 성적 데이터의 경우 최근 10 회 성적 데이터를 평균한 점수의 100 점 환산한 점수의 100 에 대한 평균점수의 보수 점수(100-평균점수)로 치환하여 점수가 높을수록 성적이 낮음을 표현한 점수를 이용하였고, 출결에 대한 데이터는 각 개별 학생별 최근 출결 데이터 10 회에 대해서 100 점 환산한 점수의 100 에 대한 평균출결 점수(100-평균출결점수)로 변환하여 점수가 높을수록 결석이 높은 점수로 환산 하였다. 상담내역 데이터는 최근 상담 내역 10 개의 텍스트 데이터를 추출하고 그 텍스트 데이터를 감성 분석(Sentimental Analysis)를 통해 부정 평가의 경우만 별도로 합산한 점수로 이 점수 또한 100 점 환산한 점수로 나타내었다. 대상이 되는 총 학생수는 6944 명이며 각 학생 별 사용된 데이터는 각 학생 별로 3 개의 항목(성적, 출석, 상담)의 시계열 데이터를 최근 데이터 기준 10 개씩 가공하여 사용 하였으므로 총 208,320 개의 데이터를 가공 처리하여 6944 명 학생의 수치형 데이터 셋으로 요약 정리 하였다. 학생 개인의 성적, 출결, 상담 내역들이 사용 되었지만 성적, 출결, 상담이력 데이터와 퇴원과의 관계에 대해서만 중점적으로 분석을 진행 하였으므로 개인을 식별할 수 있는 개인 정보는 별도로 사용되지 않았다.

4. 연구 결과

본 연구에서는 학원 퇴원율에 대한 예측을 기계학습 알고리즘 중 3 가지 Logistic Regression, SVM, DNN 등의 모델로 각각 예측모델을 작성하여 수행 하였다.

각 모델 별로 퇴원 예측 결과는 아래 표와 같다.

구분	예측 결과	비고
Logistic Regression	0.75	
SVM	0.77	
DNN	0.80	가장 높음

[표 4.1] 예측모델 비교 결과

위의 결과와 같이 모델 별 실행 결과가 조금씩 상의하게 나타났다. 성능적 측면에 있어서는 DNN 알고리즘이 80%로 가장 뛰어난 것으로 평가 되었다. 나머지 Logistic Regression 의 결과가 조금 낮았고 SVM 모델의 결과는 Sigmoid Kernel 의 경우 예측율이 50% 미만으로 예측모델로는 의미가 없었고 Gaussian Kernel 방식은 DNN 예측모델과 아주 큰 차이는 보이지 않는 수준으로 결과가 나왔다.

각각의 Confusion Matrix 는 아래 표들과 같았다.

구분	precision	recall	f1-score	support
0	0.81	0.82	0.81	924
1	0.63	0.61	0.62	465
Avg/total	0.75	0.75	0.75	1389

[표 4.2] Logistic Regression 결과

구분	precision	recall	f1-score	support
0	0.84	0.79	0.82	914
1	0.64	0.72	0.68	475
Avg/total	0.78	0.77	0.77	1389

[표 4.3] SVM 결과(Gaussian, Kernel)

구분	precision	recall	f1-score	support
0	0.84	0.86	0.85	883
1	0.74	0.71	0.72	506
Avg/total	0.80	0.80	0.80	1389

[표 4.4] DNN 결과

5. 결론

본 연구에서는 다수의 시계열 데이터를 단일 수치형 데이터로 변환하여 여러 모델에 의한 최적의 딥러닝 알고리즘을 기반으로 학원 학생들의 이탈 결과를 예측하였다는 점에서 의의가 있다. 모델의 예측 결과가 아주 높은 결과를 도출하지 못한 이유로는 감성분석 모델의 결과가 84%수준이었으며 그 영향도 어느 정도 미쳤을 것으로 보이며 또한 적극적인 의사 표명을 하는 고객들의 경우는 좀 더 고객의 불만 사항에 대한 대응이 쉬울 수 있으나 불만사항에 대한 적극적이지 않은 고객이 다수 존재 할 수 있어서 그 부분에 대한 상쇄 값이 어느 정도 반영 된 것으로 예측된다. 하지만 예측 모델로는 어느 정도 유 의미한 수준의 결과가 도출 된 것으로 생각 된다. 이 논문의 결과를 통해 다양한 학원 사업자들이 비즈니스 활성화에 도움이 되었으면 한다

참고문헌

[1] Hirschman, A., Exit, Voice and Loyalty: Responses to Decline in Firms, Organization and States, Cambridge, Mass: Harvard University Press, 1970.

[2] 장문경, 유병준, 이재환: 고객센터 상담내용 분석을 통한 이탈 요인에 관한 실증 연구, The Journal of Society for e-Business Studies, 2017.

[3] 송춘자: 고객 피드백에 대응하기 위한 LAD 기반 토픽분류 기법, 고려대 석사 논문, 2016.

[4] 이세희, 이지형: RNN 을 이용한 고객 이탈 예측 및 분석, 한국컴퓨터정보학회 학술대회, 2016.

[5] Yeon, J. H., Lee, D. J., Shim, J. H., Lee, S. G., “Product Review Data and Sentiment Analytical Processing Modeling,” The Journal of Society for e-Business Studies, Vol. 16, 2011