머신러닝, 딥러닝을 이용한 통신서비스 이용고객 분석 및 이탈 예측

김상휘*, 김기원**, 김유성**, 윤태영***, 전재완**
*동국대학교 산업시스템공학과
**인하대학교 정보통신공학과
***홍익대학교 컴퓨터공학과
hsk6913@naver.com, rockrool1@naver.com, helios1127@naver.com,
dhrmsry777@naver.com, beige481845@gmail.com

Analysis of customer churn prediction in telecom industry using Machine learning & Deep learning

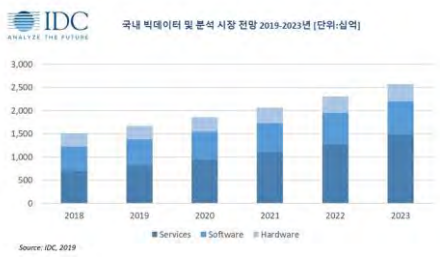
Sang-Hwi Kim*, Ki-Won Kim**, Yoo-Sung Kim**, Tae-Young Yoon***, Jae-Wan Jeon**
*Dept. of Industrial System Engineering, Dong-guk University
**Dept. of Information Engineering, In-Ha University
***Dept. of Computer Engineering, Hong-Ik University

요 약

최근 빅데이터 기술이 다양한 산업과 접목되고 있다. 그 중 고객 이탈 방지가 최우선인 통신사들 또한 예외가 아닐 수 없다. 이에 본 논문은 통신사 데이터에 머신러닝 알고리즘을 접목. 이탈 예측과 데이터 추이를 분석하고, 이를 시각화 하여 일목요연하게 표출하는 과정을 제공함으로서 통신사의 고객 유치 정책을 위한 토대를 마련할 것이다.

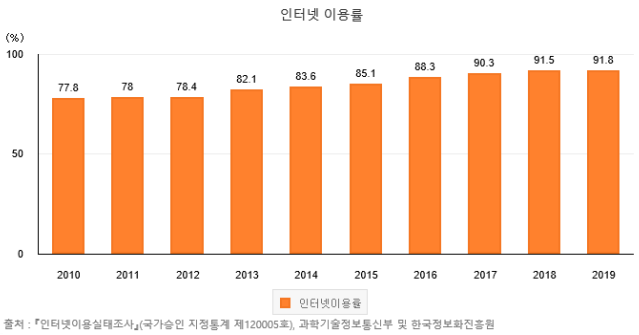
1. 서론

2020 년 1 월 22 일, 한국 IDC 는 최근 발간한 국내 빅데이터 및 분석 시장 전망 연구 보고서에서 국내 빅데이터 및 분석시장이 2023 년까지 연평균 성장률 11.2%를 기록하며 큰 규모로 성장할 것을 전망했다. 빅데이터 기반의 고급 분석 및 인공지능 시스템 구축으로 인한 시장의 규모는 점점 커지며, 기업은 이를 통해 복잡한 분석 및 비즈니스 의사 결정을 자동으로 해결함으로써 지속적인 수익 창출과 경쟁 우위를 확보할 수 있다고 강조했다. 이러한 흐름에 따라 정부의 정책적 지원과 더불어 기업 경영진 수준에서의 투자가 이러한 연구에 박차를 가하고 있다. 최근 통신에서 기계학습으로 인한 연구는 널리 활용되고 있다. 무선전송접속 기술, 이중망 기술, 네트워크 기술, 보안 기술, 5G PPP 프로젝트 등 통신분야에서도 널리 사용되고 있다.[1]



(그림 1) 국내 빅데이터 및 분석 시장 전망

현재 통신사업자 즉, 통신서비스제공자는 포화된 시장에서 경쟁 중이다. 2019 년도 기준 대한민국의 인터넷 이용률은 91.8%이고 가구 인터넷 보급률은 2019 년 기준 81.6%이다.[2] 이렇게 세계적으로도 인터넷 보급률에 대해 높은 수치를 가지고 있는 한국 통신사업자 시장에서 통신사업자는 고객의 충성도(Royalty) 및 이탈(Churn)에 대해 사업적으로 매우 중요하고 민감한 요소이다. 특히, 사업자의 매출 및 마케팅 비용절감 등의 효과로도 직결되는 요인이므로 통신사업자는 더더욱 정확한 빅데이터 및 분석을 원하고, 더 나아가 고객의 이탈을 잘 예측할 수 있길 원한다.



(그림 2) 국내 인터넷 이용률

우리는 이러한 빅데이터 및 분석 시장과 통신시장의 흐름에 따라 통신사업자의 빅데이터 및 분석과 기계학습을

통한 예측을 활용하여 자동화된 웹으로 표현하여 데이터분석에서 더 나아간 모습을 보여주고자 한다.

2. 활용 기술

2-1. XG Boost

XGBoost 는 앙상블 기법 중 부스팅을 사용하는 알고리즘인데, 앙상블 기법이란 동일한 학습 알고리즘을 사용해서 여러 모델을 학습하는 개념이다. 부스팅은 여러 트리를 합치는 방식(Additive)으로, 계속해서 발전해 나가는 방식을 쓴다. 어떻게 발전시키는지에 대해서는 바로 다음과 같은 목적함수을 최소화하는데, 시그마를 보면 한두개의 예측기가 아니라 여러 개의 예측기의 los 로부터 구함을 알 수 있다

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

<수식 1> 목적함수.

Regularized 된 objective 를 최소화하는 것이 목적. l 은 loss 이다.위의 식에서, 유클리디안 공간을 이용하는 optimization 은 쓸 수 없으니, 부스팅에서 사용하는 방법과 같이 조금씩 어떠한 식을 더해 나가면서 수정하는 모델을 만들어낼 수 있다.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

그리고 테일러 전개를 통해 2 차 다항식으로 근사한 후, 상수항을 빼 버리면 다음과 같다. 자세한 방법론은 테일러 정리를 찾으면 된다.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

그리고 고정된 트리 구조 q 에 대해서 정리하면 다음과 같다.

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

다만 생성가능한 모든 트리구조를 다 만들어 내는 것은 불가능하므로, 하나씩 더해가는 구조로 이를 greedy 방식으로 유추하자면 leaf 가 하나인 트리부터 점수를 매기는데 이를 분기할 때의 기준으로 삼을 수 있다. 따라서 최종적으로 분기된 나무와 그전의 나무의 차이를 구해주면 정보이득으로 계산한다. [3]

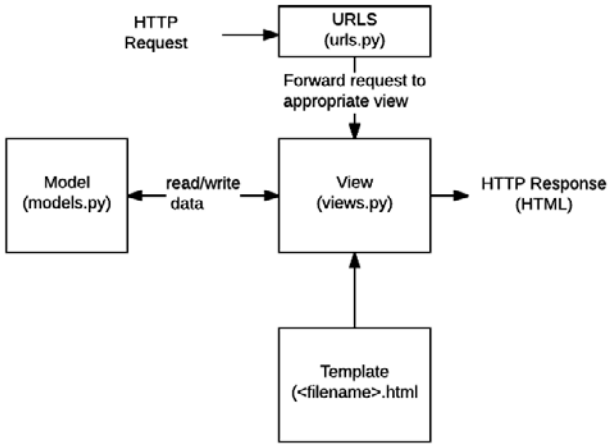
$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (7)$$

2-2. Tableau

최신 비즈니스 인텔리전스를 위해 시장에서 가장 많은 선택을 받는 Tableau 플랫폼은 빠르고 손쉽게 거의 모든 시스템에서 모든 종류의 데이터를 가져와 실행 가능한 인사이트로 전환하는 것으로 잘 알려져 있다. Tableau 의 다양하고 유용한 리소스, 교육 및 글로벌 데이터 커뮤니티는 고객과 고객의 분석 투자에 대한 지원을 제공한다. 데이터 활용을 위해 많은 곳에서 사용하는 데이터 시각화 BI 인 태블로로 데이터를 연결하고, 이를 시각화 할 수 있다.

2-3. Django

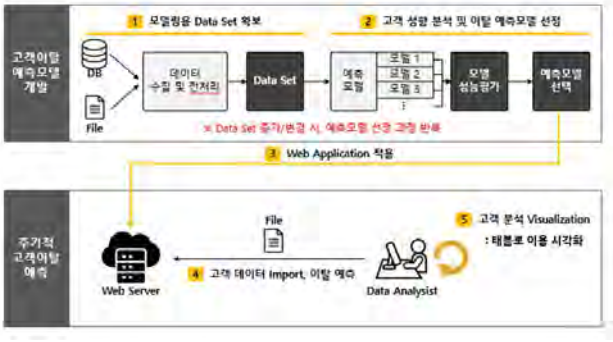
장고는 파이썬으로 작성된 무료 오픈 소스 웹 프레임워크로, 모델-뷰-컨트롤러 패턴을 따르고 있다. 현재는 장고 소프트웨어 재단에 의해 관리되고 있다. 고도의 데이터베이스 기반 웹사이트를 작성하는 데 있어서 수고를 더는 것이 장고의 주된 목표 이다.보안이 우수하고 유지보수가 편리한 웹사이트를 신속하게 개발하는 하도록 도움을 주는 Django 는 새롭게 웹 개발을 시작할 필요 없이 그저 프레임 워크를 활용하여 앱 개발에만 집중할 수 있다.전형적인 데이터 기반 웹 사이트에서 웹 어플리케이션은 웹 브라우저(또는 다른 클라이언트) 로부터 HTTP 요청(Request)을 기다린다. 요청을 받으면, 웹 어플리케이션은 URL 과 POST 데이터 또는 GET 데이터의 정보에 기반하여 요구사항을 알아낸다. 그 다음 무엇이 필요한 지에 따라, 데이터 베이스로부터 정보를 읽거나 쓰고, 또는 필요한 다른 작업들을 수행한다. 그 다음 웹 어플리케이션은 웹 브라우저에 응답(Response)을 반환하는데, 주로 동적인 HTML 페이지를 생성하면서 응답한다. Django 웹 어플리케이션은 전형적으로 아래와 같이 분류된 파일들에 대해 일련의 단계를 수행하는 코드로 구성되어 있다.



(그림 3) Django 동작도

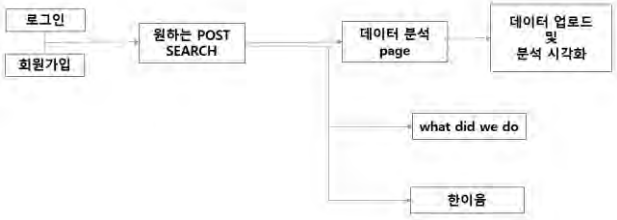
3. 설계 및 구현

3-1. 작품 구성도



(그림 4) 전체 작품 구성도

그림 4 는 전체 작품의 구성도를 나타낸 그림이다. 통신서비스 기업의 3 개년 918,607 개의 real data 로 전처리와 모델을 선정하여 개발하였다. 가장 성능이 좋은 모델을 선정하여 고객의 성향을 분석하였다. 이후 데이터 분석에 용이하게 이를 웹 페이지에 visualization 하였다.

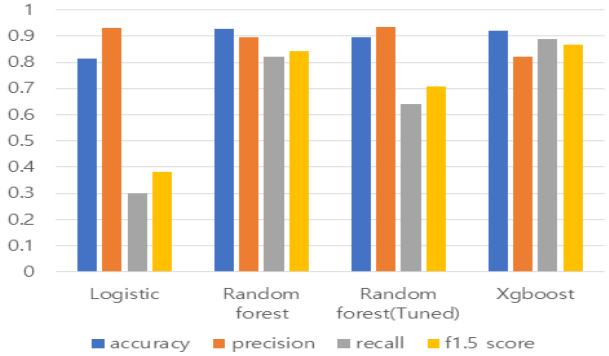


(그림 5) 웹페이지 구성도

그림 5 는 웹 페이지 구성을 구조적으로 나타낸 그림이다. 메인 화면에 로그인 창이 있고 로그인을 하게 되면 원하는 페이지를 배너를 통해 선택할 수 있게 된다. 이 때 선택할 수 있는 배너는 데이터 분석 페이지, 구현하기까지의 과정을 담은 페이지, 그리고 한이음 페이지까지 총 세개의 페이지를 선택할 수 있게 된다.

3-2. 기능 처리

Logistic Regression, Random Forest, XGboost, Catboost 총 4 개의 모델링 기법을 각각 파라미터 튜닝을 통해 data set 에 적용하였다. 이 때 '이탈할 사람을 이탈하지 않을 것이다' 라고 예측하는 것을 방지하는 것이 중요하다 판단하여 recall 에 비중을 많이 두고 precision 역시 고려하여 f1.5 score 가 가장 높은 XGBoost 를 선정하였다.

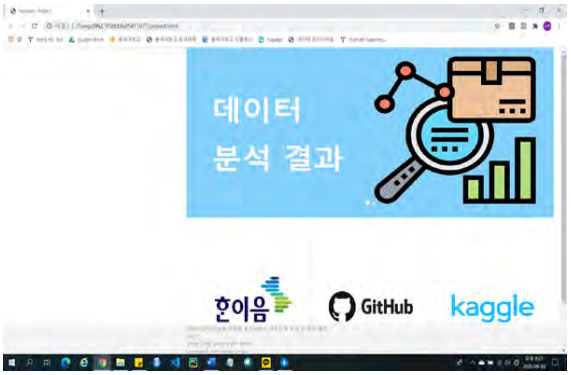


(그림 6) 모델 별 성능 비교

그림 6 은 모델별 성능을 비교한 것으로 XGBoost 의 f1.5 score 가 가장 높은 것을 볼 수 있다. 선정한 모델과 전처리 코드를 웹에 적용하여 타 데이터에도 적용할 수 있도록 하였다. 웹을 통해 사용자로부터 데이터를 수신하면 이를 분석하여 서버의 머신러닝 SW 를 통해 고객의 성향을 파악하고 이탈을 예측하게 된다. 이후 태블로를 통해 이탈 예측 데이터를 시각화하여 출력하게 된다.

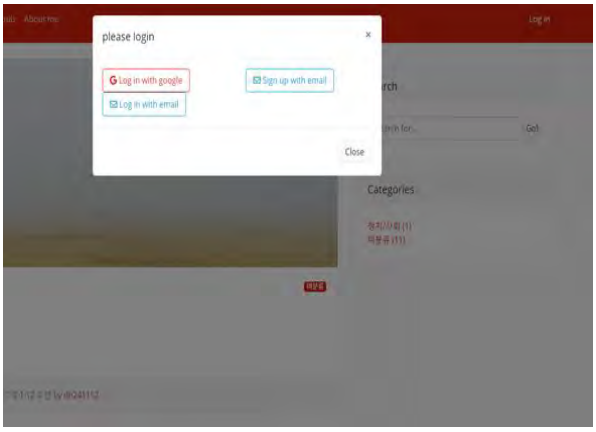
3-3. 구현

그림 7은 메인화면 페이지의 구현이다. 메인화면 페이지에서는 배너를 활용하여 분석 결과 페이지, 깃허브 등 여러 사이트를 쉽게 방문할 수 있도록 디자인했다.



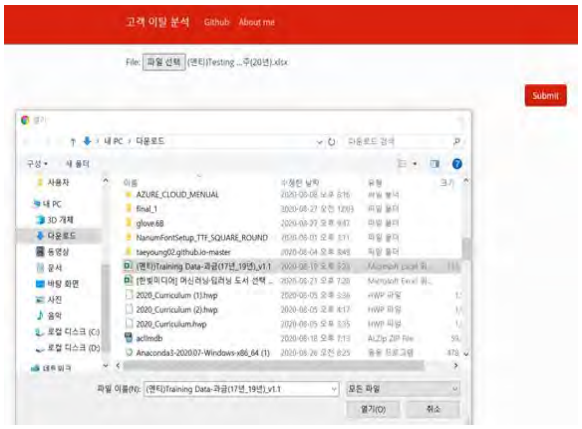
(그림7) 메인 화면

'데이터 분석 결과' 배너를 클릭하면 데이터 업로드 및 분석 결과를 확인할 수 있는 사이트로 이동한다.이때 구글 이메일을 통해 로그인을 해야 한다. 그림 8은 사이트로 이동한 후, 로그인하기 위한 창이다.



(그림 8) 로그인 화면

그림 9는 분석할 데이터를 업로드하는 화면이다. 분석을 시작하기 위해서는 데이터가 로컬 환경에서 존재해야 하고, [upload] 버튼을 클릭하면 데이터가 업로드 된다.



(그림 9) 데이터 업로드

데이터를 업로드하면 미리 작성해둔 코드를 통해 자동으로 전처리와 모델링이 된다. 분석된 결과는 태블로를 통해 시각화 된다. 대쉬보드는 월평균 요금(그림 10), 이탈확률(그림 11), 고객 종류(그림 12), 지역(그림 13)을 기준으로 분석결과를 시각화 한다.

(그림 10) 월평균 요금

(그림 11) 이탈 확률

(그림 12) 고객 종류

(그림 13) 지역

4. 결론

본 논문에서는 통신서비스 이용고객의 행동 패턴을 분석하고 이탈을 예측할 수 있는 이탈 방지 시스템을 시각화하고 웹에 구현하였다. 본 논문에서 설계하고 구현한 프로그램은 이탈 여부를 예측할 뿐만 아니라 이탈확률을 제공함으로써 높은 확률로 예측된 고객을 우선적으로 관리할 수 있다. 또한, 4가지 기준으로 시각화된 대쉬보드를 통해 고객을 세분화 및 그룹화할 수 있고 이를 통해 고객 관리의 효율이 높아질 것이라 기대한다. 본 논문에서 설계하고 구현한 프로그램은 '무작위적 고객 관리'를 '선택적 고객 관리'로 발전시키면서 고객의 만족도와 신뢰도를 높일 것이라 예상한다.

[본 논문은 과학기술정보통신부 정보통신창의인재양성 사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다]

참고문헌

[1] 김근영 외 5 인, 기계학습을 활용한 5G 통신 동향, 전자통신동향분석, 31 권, 5 호, 1-10 쪽, 2016

[2] 『인터넷이용실태조사(국가승인지정통계 제 120005 호), 과학기술정보통신부 및 한국정보화진흥원

[3] Tianqi Chen & Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016, 8/13-17, 13 Page.