가처분 소득 추정 모델 개발에 관한 연구

임성준*, 임희석**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 컴퓨터정보대학 컴퓨터학과
e-mail : bigdata@korea.ac.kr

A Study on the Development of disposal income estimation model

SungJun Lim*, HeuiSeok Lim**
*Dept. of Big Data Convergence, Graduate School of Computer & Information Technology, Korea University
**Professor, Dept. of Computer Science & Engineering, College of Informatics, Korea University

요 약

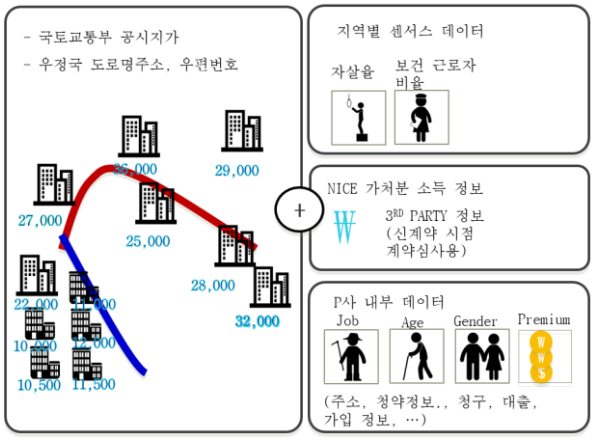
역사적으로 생명 보험은 상호부조의 형태로 갑작스럽게 어려운 상황이 발생하여도 경제적 어려움으로부터 가정을 지키는 역할을 해 왔다. 이는 평상 시에 만약의 경우를 대비하여 일정 비용을 지불함으로써 한 가정의 보장을 잘 준비하였기 때문이다. 하지만 한 가정의 경제적 상황은 지속적으로 변하기 때문에 시간이 지남에 따라 청약 당시의 보장 상태가 적절히 유지되고 있음을 확신할 수 없게 된다. 따라서 본 연구에서는 가처분 소득 추정 모델 개발을 통해 기존에 생명보험에 가입한 고객의 가처분 소득을 추정함으로써 고객에게 적절한 보장 강화의 기회를 제공하도록 한다.

1. 서론

불확실성이 큰 세상 속에서 한 가정이라도 더 지키기 위해 생명보험 회사는 보험의 본질인 보장을 조금이라도 더 전달하고자 전통적인 마케팅 방법부터 보험에 InsureTech 를 접목하는 등 다방면으로 노력한다. 일반적으로 보험회사로부터 보장을 받고자 하는 고객은 회사가 제공한 청약서 작성을 통해 이전에 발생한 병력이나 소득, 직업 등 다양한 정보를 제공하고 보험회사는 이를 기반으로 고객의 인수 여부를 결정한다. 이 중 청약서에 기재된 고객의 소득 정보는 신규 청약 시점 뿐만 아니라 향후 추가 청약의 기회가 되기 때문에 보험회사의 입장에서는 매우 중요하다고 할 수 있다. 하지만 보험 청약 당시 고객으로부터 작성된 소득 정보는 고객의 입장에서는 Private 한 민감 정보이기 때문에 정확하지 않는 경우가 있어 고지의 신뢰성에 대한 의문이 발생하는 경우가 많으며 청약 이후 시간이 갈수록 크게 의미 없는 정보가 되고 만다. 또한, 청약 이후 고객에게는 소득의 변화가 생기거나 가정에는 아이가 태어나는 등 지속적으로 상황이 변화하고 있지만 청약 시점의 보장 내역이 현재 시점을 반영하지 못하여 충분한 보장이 이루어지지 못하는 등의 문제가 있다. 만약 고객의 소득 데이터가 보장을 담당하는 설계사들에게 적절히 전달된다면 추가 보장의 제공 가능성에 대한 강력한 인사이트를 얻게 할 수 있다. 따라서 본 논문에서는 다양한 소득 정보 중 실제 소비가 가능한 정도를 의미하는 가처분 소득을 추정하는 모델을 만들어 제공하고자 한다.

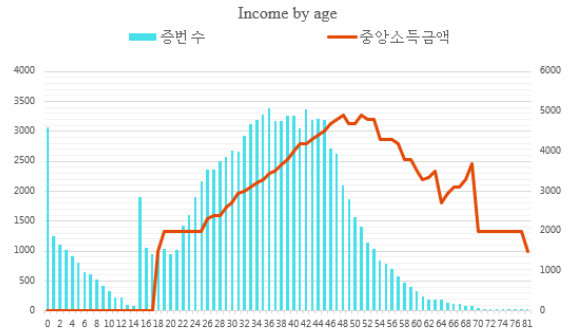
2. 소득 추정 방법

소득에 대한 추정은 고객 주소에 대한 국토해양부 공시지가 정보를 통한 소득 추정이 가능하다. 즉, 모든 부동산 실거래가에 대해 주소 및 거래일자별 데이터를 제공하고 있고, 우정사업 본부에서는 도로명 주소 정보를 배포하고 있는데 이 정보와 부동산 실거래가를 연결시키면 어느 도로 쪽이 평균 부동산 지가가 높은지를 알 수 있다. 또한 신용 정보사로부터 구매했던 신계약 시점에서의 가처분 소득과 생보사의 내부 데이터, 예를 들면 고객의 주소, 직업, 연령, 성별, 보험료, 약관대출, 보험금, 가입 상품 등의 정보를 매핑하면 소득 추정이 가능하다. 이를 위해 본 논문에서는 저자가 속한 P 보험사의 정보를 활용하였고 NICE 데이터 기준이 2016 년이기 때문에 2016 년도 신계약 데이터 1 년치의 정보를 활용하였으며 이를 통해 약 150 여가지의 변수를 채택해 모델을 개발하였다.



(그림 1) 소득 추정 방법

3. 모델 구성



(그림 2) 2016 년도 연령별 증번 수와 소득 중위 수

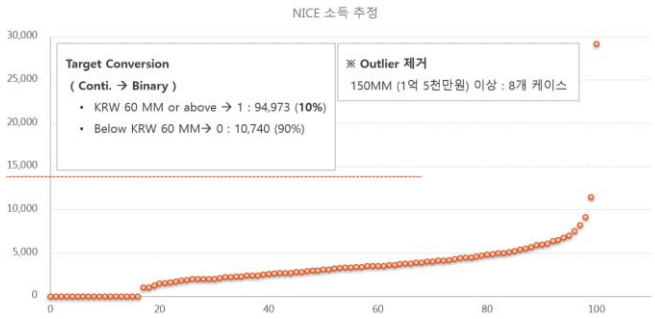
구분	50 세 이하	50 세 초과
대상	96,775 명 (92%)	8,958 명 (8%)

<표 1> 50 세 기준 대상 분포

(그림 2)에서 보는 것처럼 P 사의 2016 년도 신계약 대상자 약 10 만명의 소득 중위 수 값으로 보았을 때 소득은 50 세를 기준으로 급격히 감소하는 것을 확인할 수 있었다. 따라서 본 모델은 하나의 모델로 구성하지 아니하고 명예 퇴직, 실직 등 소득의 패턴이 바뀌는 50 세를 기준으로 2 개의 구간으로 나누었다. 데이터는 학습/테스트/검증를 위해 4:3:3 으로 나누었다.
아래 그래프는 DNN 기법의 설명력을 보완하기 위해 dual 로 개발한 일반화 가법모델인 [1]GAM 모델 (Generalized Additive Model)이다.

50 세 이하 모델은 GAM 에 의해 설명력이 강한 10 개의 변수를 채택하였다. 연령이 가장 중요한 변수였고, 집 값의 중위 수, 직업 카테고리, 피보험자의 성별, JOB 의 생존 등급, 일반 사망 금액, 교육 중심 지역, 납입 보험료 수준, 노령화 지역이 주요 변수였다. 또한 50 세 초과 GAM 모델은 50 세 이하 모델과 유사하나 연령과 교육 중심 지역의 설명력이 높았던 것에 비하면 떨어졌다.

또한 신용 정보사인 NICE 의 가처분 소득 정보는 아래와 같이 100 구간으로 나누어 데이터를 조정하였다.

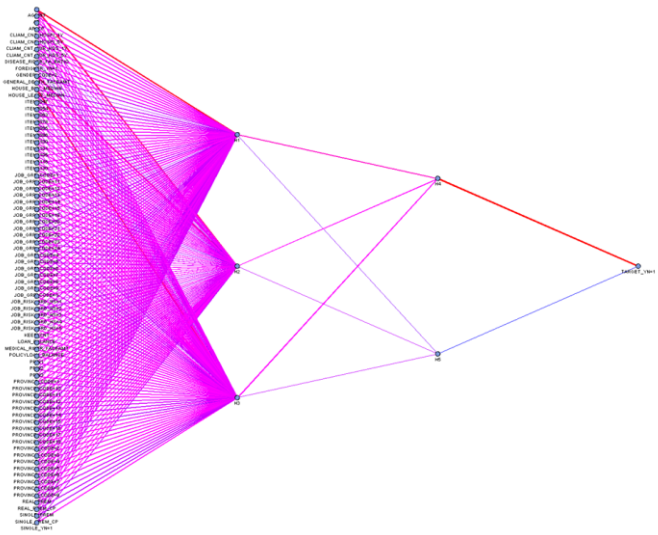


(그림 3) 100 구간으로 나눈 NICE 의 소득 추정

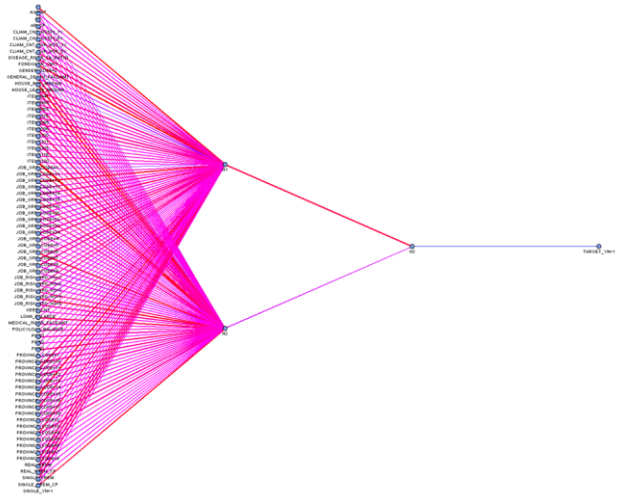
신용 정보사로부터 구매한 신계약 시점에서의 가처분 소득과 P 사의 내부 데이터, 예를 들면 연령, 성별, 보험료, 약관 대출, 보험금, 가입 상품 등의 정보를 매핑하여 약 150 개 정보를 채택하였다.

4. 모델 결과

2016 년 1 월~12 월 신계약 대상자 약 10 만명을 대상으로 증권 번호와 피보험자 이름과 증권의 발행 일자를 Key 값으로 하였고 Target 은 Nice 의 소득 예측으로 하였다.
32 개는 P 사의 내부 데이터인 고객정보(9), 보험료(11), 보험금 청구(9), 대출(3)로부터, 그리고 나머지 121 개의 변수는 국가통계포털인 KOSIS 로부터 117 개, 국토교통부와 우정국으로부터 4 개의 input factor 를 받아 학습에 사용하였다.
학습은 다층 퍼셉트론 딥러닝 방법론과 참조 모델로 GAM(Generalized Additive Model) logistic 을 같이 사용하였다.
모델링 Tool 로는 SAS Enterprise Miner 를 이용하였으며 사용된 알고리즘은 Decision Tree, Random Forest, SVM, Gradient Boosting, Deep Neural Network 이었다. 이 중 극단치 추정에 있어 다른 모델보다 우수하고 개발 및 유지보수가 상대적으로 쉬운 Neural Network 이 선택되었다. 레이어와 뉴런 수는 수작업으로 옵션을 변경해 가며 수행한 결과 50 세 이하는 아래 (그림 4)와 같이 2 layer, 5 Neurons 가 사용되었고 50 세 초과는 (그림 5)와 같이 2 layer, 3 Neurons 가 사용되었다.

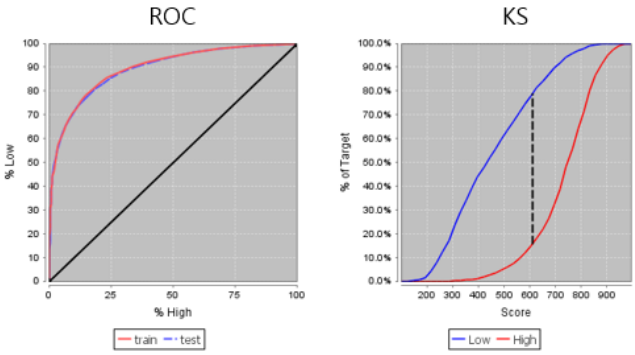


(그림 4) 50 세 이하 Neural Network



(그림 5) 50 세 초과 Neural Network

GAM 을 이용하여 50 대 이하와 초과의 소득 추정은 아래와 같다. 모델을 수행한 결과 우선 50 대 이하의 소득 추정의 신뢰도는 아래와 같았다.



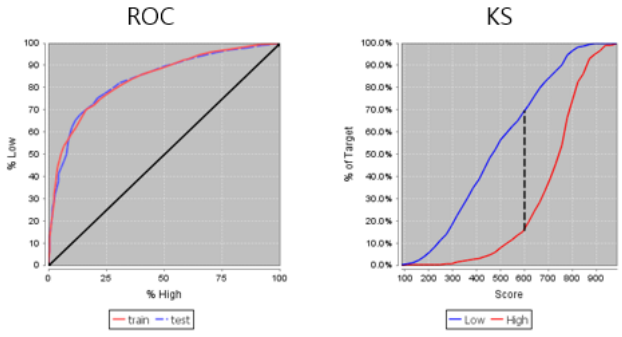
(그림 6) 50 대 이하 대상 ROC, KS

50 대 이하

Gubun	*Divergence*	*ROC*	*Gini*	*KS(%)*	*RMSE*	*NSSE*
Train	3.511	0.896	0.791	73.3%	513.782	0.815
Test	3.433	0.893	0.786	73.2%	513.791	0.815

<표 2> 50 대 이하 모델링 결과

50 대 초과는 아래와 같다. (9 개의 선택된 변수)



(그림 7) 소득 추정 방법

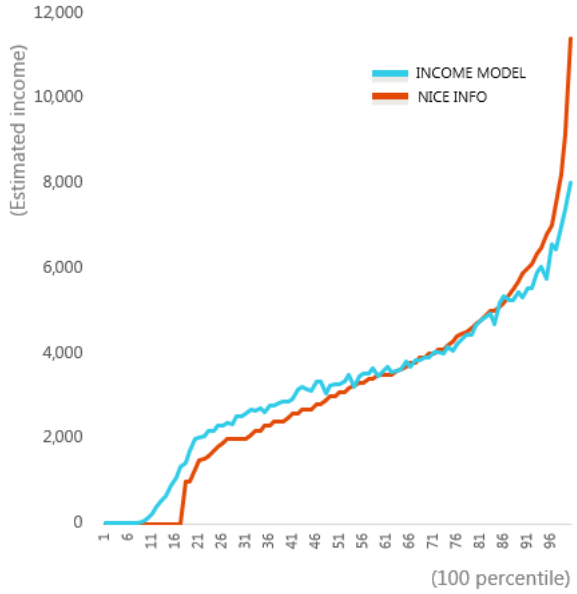
Gubun	*Divergence*	*ROC*	*Gini*	*KS(%)*	*RMSE*	*NSSE*
Train	2.050	0.837	0.675	57.3%	571.74	0.764
Test	2.039	0.836	0.672	55.2%	566.345	0.762

<표 3> 50 대 초과 모델링 결과

50 대 이하에서는 89%, 50 대 초과에서는 84%정도의 신뢰도를 보였다.

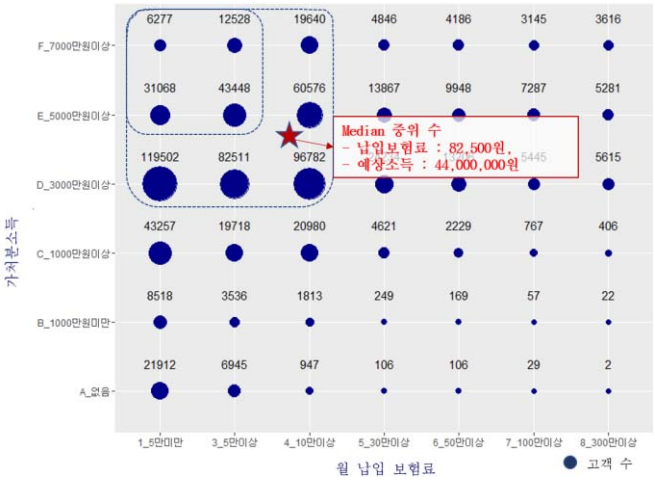
5. 결론

검증(Verification) 데이터로 NICE 신용 평가사와 본 모델에서 개발된 가처분 소득의 가격 곡선은 아래 (그림 8)에서 보는 것처럼 전 구간에 걸쳐 유사한 것을 확인할 수 있었다.



(그림 8) Verification Data 를 통해 확인된 NICE 정보와 소득 추정 모델의 비교 추이

또한 납입 보험료와 가처분 소득과의 관계를 아래 (그림 9)와 같이 그려본 결과 소득 추정 모델에 따라 확인된 P 생보사 약 71 만명 고객의 가처분 소득 중위수는 4400 만원이고 납입보험료 중위수는 82,500 원으로 나타났다. 납입 보험료 수준이 높아질수록 대체로 가처분 소득도 높은 것을 볼 수 있다. 또한 중위 소득 및 보험료 기준으로 소득 수준은 높으나 납입 보험료가 낮은 고객들은 점선 박스 안으로 나타나는데 P 사의 고객 71 만명 중 작은 박스의 고객은 약 9 만 3 천여명이고, 좀 더 범위를 넓혀본 큰 박스는 약 47 만명에 해당되며 이 고객들은 본 모델에서 확인하고자 한 결과이자 보장을 추가적으로 전달하기 위한 Target 이 될 수 있다고 볼 수 있다.



(그림 9) 납입 보험료 별 가처분 소득

6. 향후 과제 및 계획

고객의 소득정보 추정 데이터는 보험 회사가 upselling 등을 통해 고객에게 더 나은 보장을 전달할 수 있으며 설계사 이탈로 인해 발생한 한번도 만나본 적 없는 [2]고아 계약(Orphan policy) 고객에게 Contact 하는 주요 정보로 활용할 수도 있을 것이다. 또한 더불어 마케팅을 위한 고객 segmentation 및 profiling 시 중요한 정보로 활용할 수 있을 것이다.

하지만 소득 정보를 추정하기 위해서는 회사가 [3] NICE 와 같은 3rd Party 데이터를 지속적으로 구매해야 한다는 점, 고객이 신용 정보 활용을 동의해야 한다는 점은 해당 모델의 follow up 과제로 남는다고 할 수 있다.

참고문헌

[1] GAM (Generalized Additive Model)
(https://datascienceplus.com/generalized-additive-models/)
[2] Orphan policy : IIFL
(https://www.indiainfoline.com/article/news-top-story/understanding-‘orphaned’-policies-113110106327_1.html)
[3] NICE Credit Information Service :
(https://www.niceamc.co.kr/kr/index.do)