이론통계학2

Project #7. 생명 보험 해지 - CRM

발표일 2021. 11. 10

1조

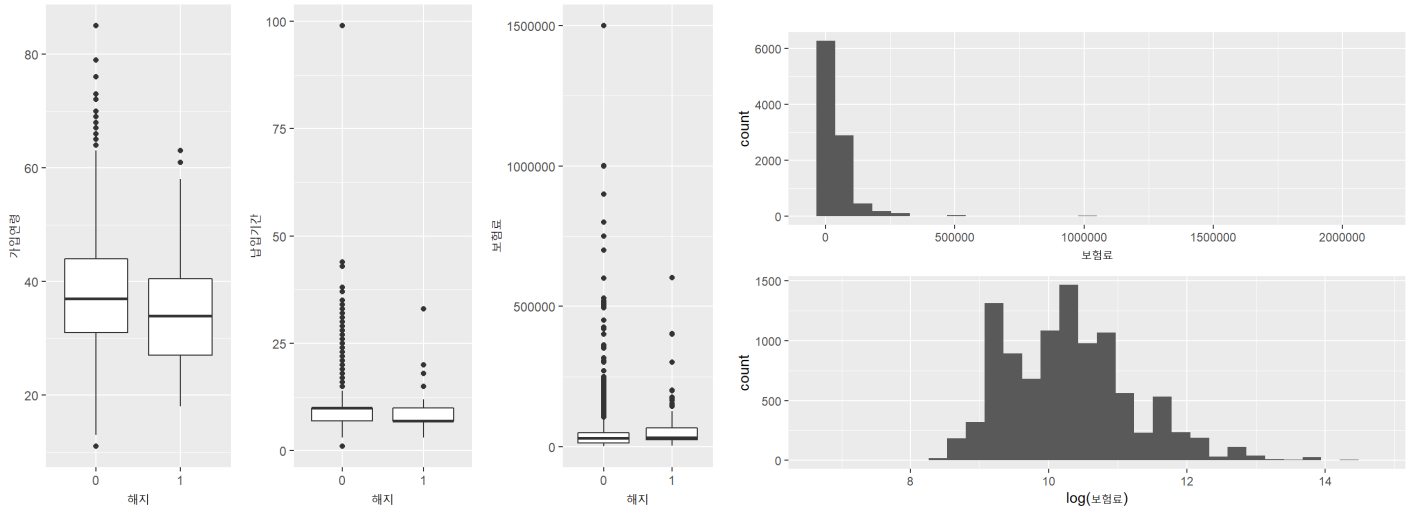
202STG26	박지윤
202STG27	이수현
212STG04	김이현
212STG12	박윤정

Part 1 : GLM/GAM을 이용한 고객의 보험 계약 해지 여부

1. GLM을 이용한 방법

변수 명		비고
기존 변수	가입연령	계약자의 가입 당시 연령
	납입방법	보험료 납입 방법
	납입기간/년	보험료 납입 기간, 단위 : 년
	수금방법	보험료 수금 방법
	보험료	1회 납입시의 보험료, 단위 : 원
	부활유무	실효일로부터 2년이내에 계약의 효력을 다시 발생시키는 것
	계약일자	계약 당시 일자
	지급만기일자	보험료 지급 만기 일자
	최종납입횟수	최종 보험료 납입 횟수
	상품중분류	가입한 보험 상품
	상품소분류	가입한 보험 상품
추가 변수	계약기간	계약일자로부터 지급만기일자까지 기간
	최종납입기간	최종납입횟수 * 납입 간격
	납입비율	최종납입기간/(납입기간*12)*100
	지급만기기간	2001년 6월로부터 지급만기일까지의 기간
	보험료	보험료/납입 간격 (기존 보험료 변수를 변경)
	연체	보험료 연체 횟수

기존의 예측 변수 외에 여러 파생 변수들을 추가해 분석을 진행하기로 한다. 추가 변수로는 계약기간, 최종납입기간, 납입비율, 지급만기기간, 보험료, 연체 횟수를 기존 변수로부터 생성하였다. 추가적인 변수 변환 여부를 검토하기 위해 기존 변수들의 box plot과 histogram을 그려보았다.



가입연령, 납입기간, 보험료의 boxplot과 보험료, log(보험료)의 히스토그램이다. 보험료의 로그변환의 필요성이 확인된다. 변수 추가 및 변수 변환 후, 적절한 link function을 결정하고 교호작용 포함 여부와 변수 선택 여부를 변화시켜 가며 최종 예측 모형을 결정한다.

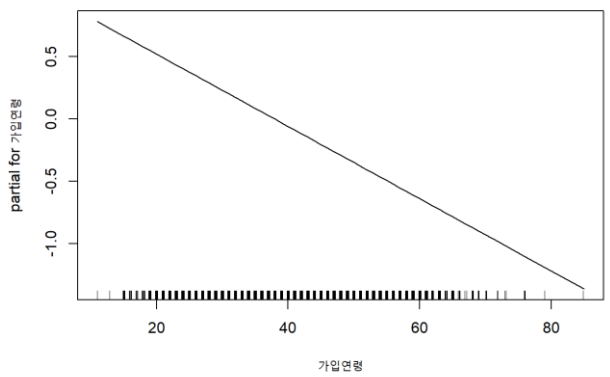
Link Function	교호작용	변수 선택	선택된 변수 개수	AIC
Probit	미포함	x	24	1636.339
		o	5	1614.200
	포함	x	199	1722.267
		o	32	1575.658
Logit	미포함	x	24	1636.235
		o	5	1615.099
	포함	x	199	17482.692
		o	27	1576.026
Gompit	미포함	x	24	1636.201
		o	5	1615.354
	포함	x	199	25412.295
		o	27	1574.976

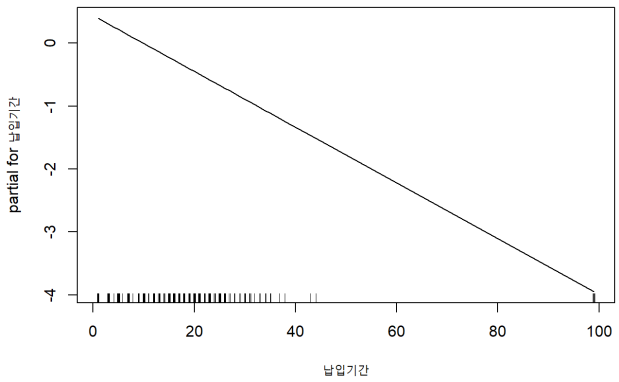
AIC를 기준으로 Link function은 Gompit을 사용, 교호작용은 포함하고 변수선택을 한 모형을 최종 예측 모형으로 선택한다. 이때의 AIC 값은 1574.976이다.

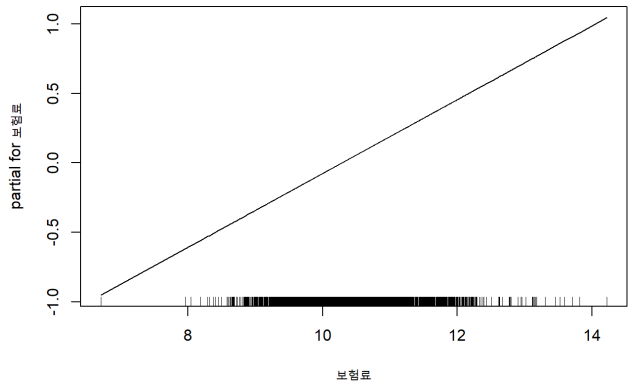
2. GAM을 이용한 방법.

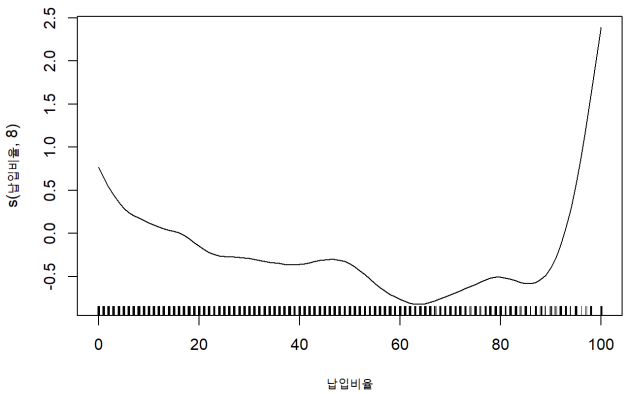
선택된 변수	AIC
가입연령, 납입기간, 보험료, 납입비율, 연체	1180.387

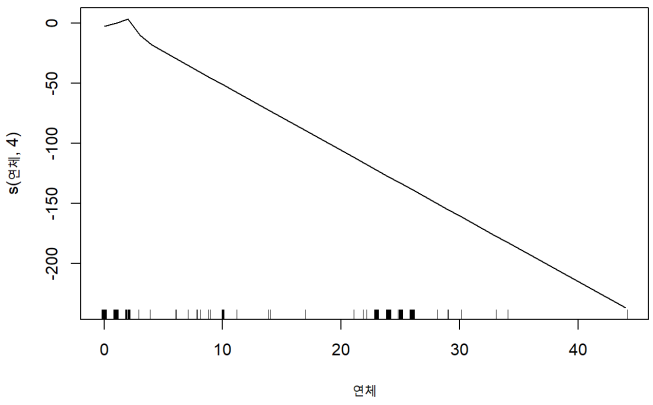
GAM 모형을 적합 후 변수 선택을 한 결과 가입연령, 납입기간, 보험료, 납입비율, 연체 변수가 선택되었고, AIC값은 1180.387이었다.











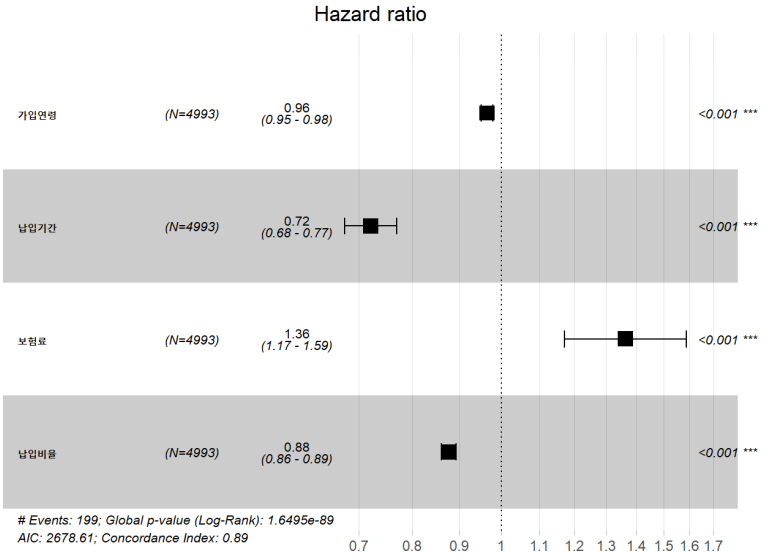
모형 내에서 각 변수의 효과를 나타낸 그래프이다. 가입연령, 납입기간, 연체가 증가할수록 보험해지 확률이 낮아지고, 보험료는 증가할수록 보험해지 확률이 높아진다. 연체 횟수가 증가할수록 보험해지 확률이 낮아지는 것은 일반적인 상식에 어긋나는 결과이나, 이는 3개월 미만으로 연체를 하되 보험해지는 하지 않는 고객의 경우가 많음에서 비롯된 것이라고 판단하였다. 납입비율이 작을수록 해지확률이 작아지다가 60부근을 지나면 납입비율이 높아질수록 해지확률이 높아진다.

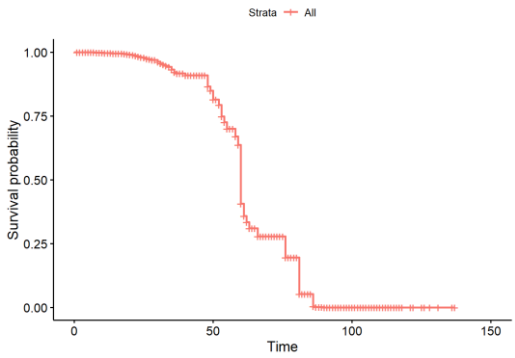
Part 2 : 생존분석 기법을 이용한 해지 시점 예측 방법

1. **순간 해지 확률에 대한 최적 Cox PHM을 찾아보시오.**

선택된 변수	AIC
가입연령, 납입기간, 보험료, 납입비율	2678.61

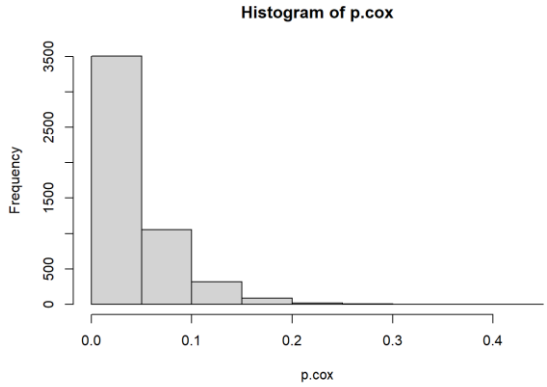
Cox PHM 모형을 적합 후 변수 선택을 진행한 결과 가입연령, 납입기간, 보험료, 납입비율 변수가 선택되었고, AIC값은 2678.61이었다.



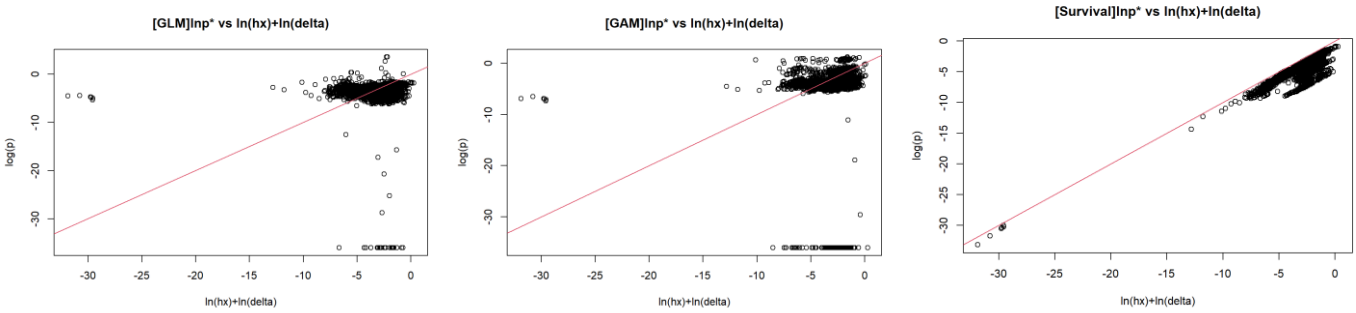


최종 모형의 Hazard ratio와 Survival probability의 그래프이다. 가입연령, 납입기간, 납입비율은 위험도가 1보다 작기 때문에 작을수록 보험해지율이 높아지고, 보험료는 위험도가 1보다 크기 때문에 커질수록 보험해지율이 높아짐을 해석할 수 있다. 생존확률 그래프를 보면 시간이 지날수록 생존확률이 감소함을 알 수 있다.

2. **보험 가입 후 t 시점에 보험 계약을 유지하고 있는 고객이 향후 3개월 내에 보험을 해지할 확률을 구해 연체 확률 사이의 관계식을 확인하고 의미를 분석하시오.**



Cox PHM 모형을 사용했을 때, 고객이 향후 3개월 내에 보험을 해지할 확률의 histogram이다. 0.1 이하의 확률이 가장 많음을 알 수 있다.



[GLM] [GAM] [Cox PHM]

위의 세 모형에서 구한 연체확률을 이용하여 다음의 관계식이 성립함을 검토하였다.

$$\ln\left(-\ln(1-p^*)\right) \cong \ln p^* \cong \ln h\left(t|x_1,\dots,x_p\right) + ln\Delta t = lnh(t) + \sum ß_j x_j + ln\Delta t$$

세 경우 비교적 점들이 y=x 근처에 존재했다. 특히 Cox PHM의 경우 위의 관계식이 강하게 성립함을 확인할 수 있었다. 이는 고객이 3개월 기간 이내에 보험을 해지할 확률은 고객의 순간 해지확률의 3배에 근사함을 뜻한다.

Part 3 : 해지 확률 계산 및 Lift Chart를 이용한 다양한 방법의 예측력 비교

1. **위에서 구한 최적 예측모형을 이용하여 추정용 자료에서 예측변수들을 이용하여 구한 해지확률 및 Cox PHM score 가 상위 10% 에 해당하는 500 명 고객들의 지시변수(indicator variables) 를 구하시오.**

사용 모형	GLM	GAM	Cox PHM
고객(상위 5명)	3680, 3883, 3253, 4539, 4509, 4564	4685, 4155, 4712, 4175, 4653, 4429	3346, 2734, 4616, 4854, 4982, 4992

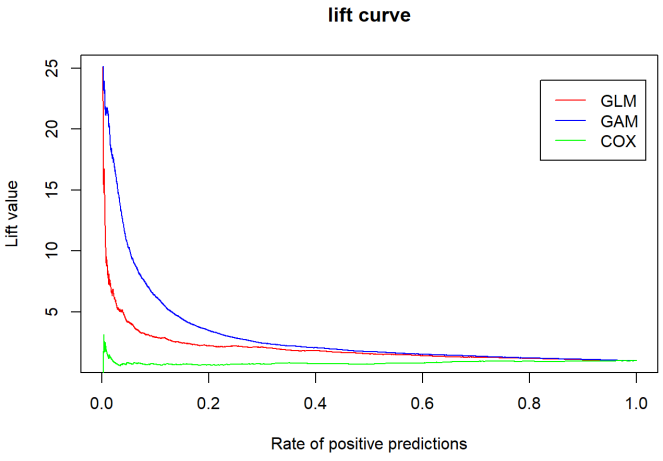
위에서 구한 세 모형을 이용했을 때 보험해지 확률과 Cox PHM score가 상위 10%에 해당하는 고객들을 추출해보았다. 위의 표는 그 중 상위 6명의 고객들만 추출한 표이다.

2. **위에서 구한 GLM/ GAM/ CoxPHM 확률 및 점수값의 크기순으로 전체 5000명의 고객을 10개의 구간으로 나눈 후 분석용 자료를 이용하여 각 구간별 실제 해지고객의 백분율 및 Lift 값을 구하여 세 방법을 비교하는 표를 만드시오.**

구간(상위)	GLM		GAM		CoxPHM	
	실제 해지고객 백분율	Lift	실제 해지고객 백분율	Lift	실제 해지고객 백분율	Lift
10%	0.118	2.961	0.250	6.273	0.082	2.057
10% - 20%	0.060	1.505	0.026	0.652	0.056	1.405
20% - 30%	0.074	1.857	0.014	0.351	0.036	0.903
30% - 40%	0.040	1.004	0.040	1.004	0.060	1.505
40% - 50%	0.016	0.401	0.018	0.452	0.030	0.753
50% - 60%	0.030	0.753	0.018	0.452	0.028	0.703
60% - 70%	0.018	0.452	0.016	0.401	0.018	0.452
70% - 80%	0.020	0.502	0.008	0.201	0.020	0.502
80% - 90%	0.008	0.201	0.008	0.201	0.024	0.602
90% - 100%	0.014	0.356	0.000	0.000	0.045	1.120

전체 5000명의 고객 데이터를 보험해지 확률과 socre값 크기순으로 10개의 구간으로 나눈 후, 각 구간별로 실제 해지 고객의 백분율과 Lift값을 구한 표이다. 세 모형 모두 상위권으로 갈수록 Lift값이 높으므로 분석이 적절히 시행되었음을 판단할 수 있다.

3. **위에서 구한 표를 그래프로 그린 Lift Chart를 서로 겹쳐서 그려보고 각 방법의 장단점을 서로 비교 검토하시오.**



세 모형의 Lift Chart를 겹쳐 그린 표이다. GLM과 GAM에 비해 Cox PHM의 성능이 떨어짐을 파악할 수 있었다. GLM의 경우는 가장 적합이 쉬운 모델이라는 장점이 있다. GAM 모델의 경우 각 변수에 비선형 함수를 적합하여 예측력을 높일 수 있다는 장점이 있다. Cox PHM의 경우는 non parametric 모델이기 때문에 예측력이 좋은 모델이라는 장점이 있으나 해당 데이터의 경우 변수 선택 과정에서 변수가 많이 빠짐으로 인해 성능이 떨어지는 것이라 판단하였다.

4. **위에서 최종 선택한 GLM/ GAM/ Cox PHM 예측모형의 AIC값을 각각 제출하시오.**

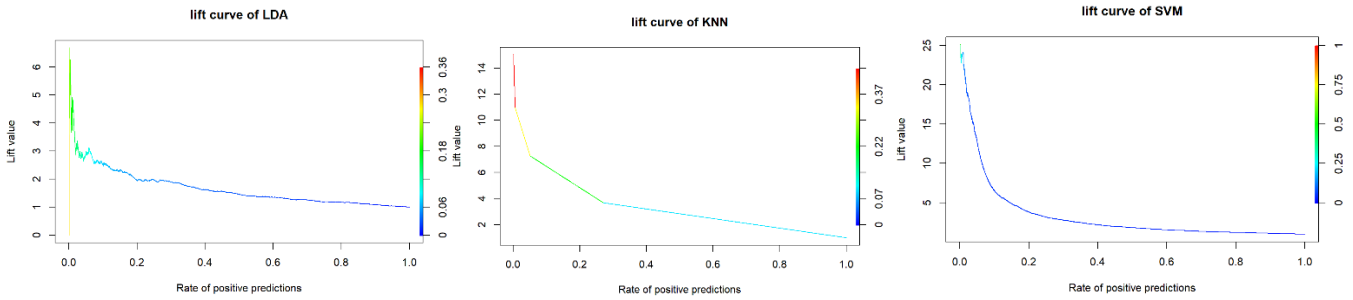
모형	GLM	GAM	Cox PHM
AIC	1574.976	1180.387	2678.61

각 모형의 AIC 값이다. GAM의 AIC값이 가장 작으므로 가장 적절한 모형이라 판단하였다.

5. **위에서 사용한 3가지 방법 외에 아래의 다양한 Data Mining 기법을 이용한 분석을 추가하여 Lift Chart를 서로 겹쳐서 그려보고 각 방법들의 장단점을 서로 비교 검토해 보시오.**

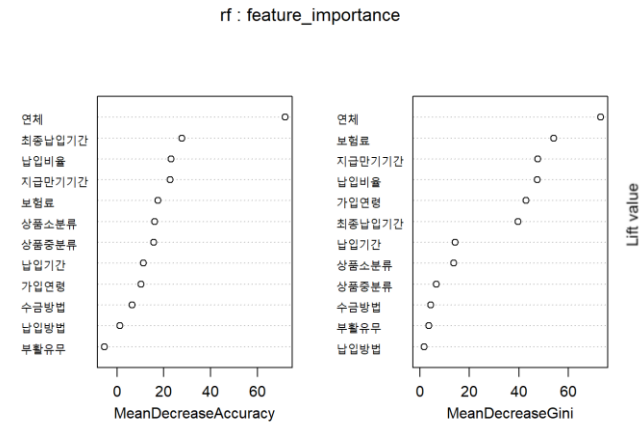
위에서 사용한 모형 외에 LDA, KNN, SVM, Random Forest, Neural Network, XGBoost 분석을 추가로 시행하여 보험 해지 확률을 예측해보기로 한다.

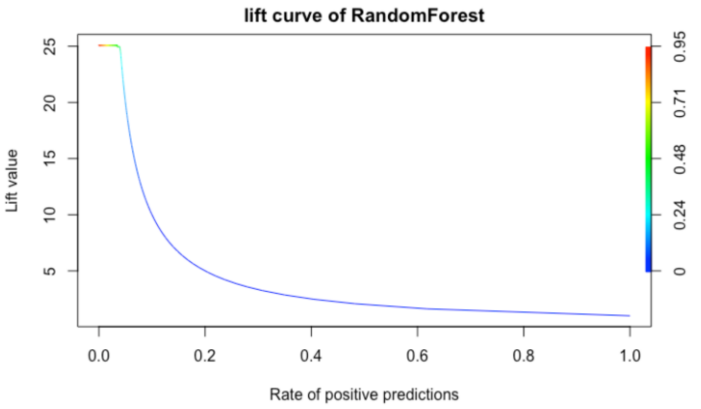
- LDA / KNN/ SVM

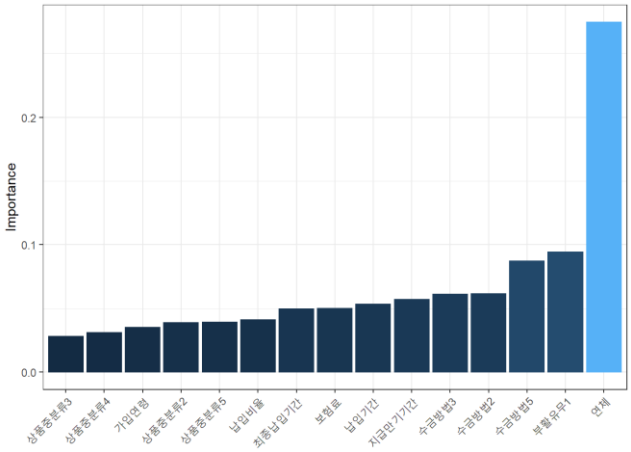


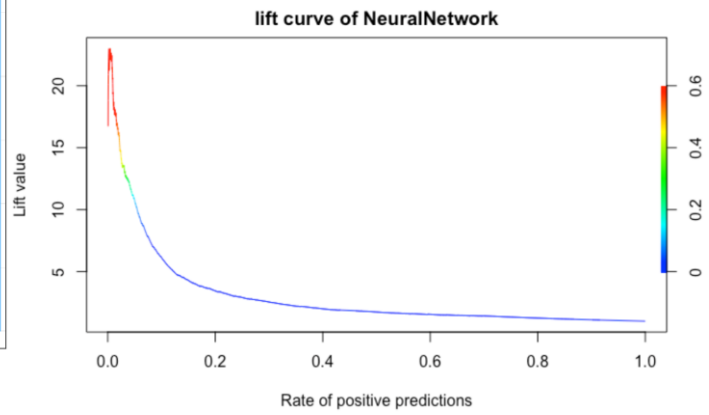
LDA, KNN, SVM의 Lift Chart이다.

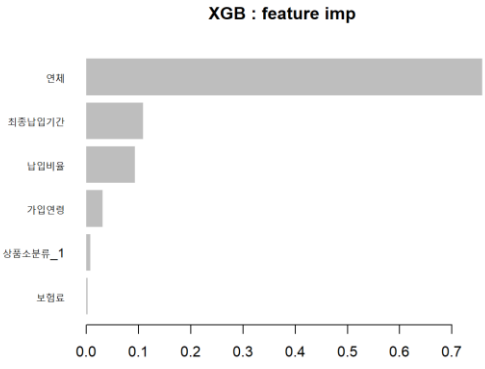
- Random Forest / Neural Network / XGBoost

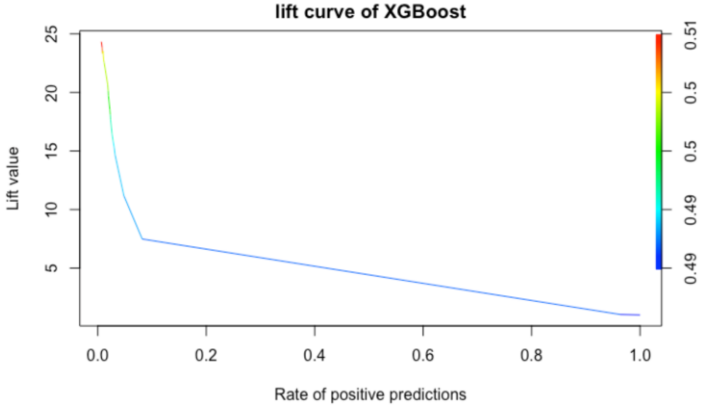




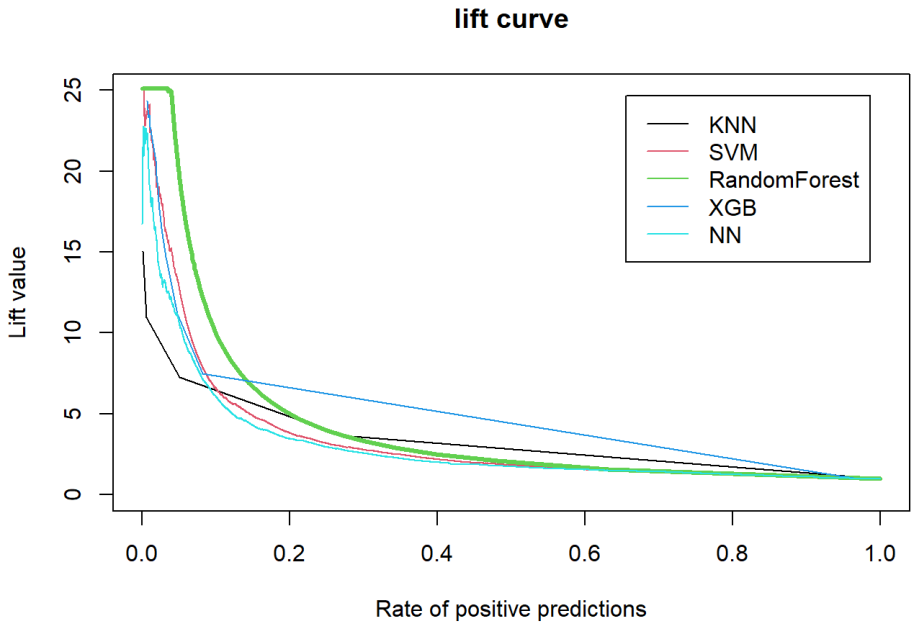








Random Forest, Neural Network, XGBoost의 변수 중요도를 그린 그래프와 Lift Chart이다. 세 모델에서 모두 연체와 보험료가 보험해지율 예측에 중요한 변수임을 확인할 수 있다.

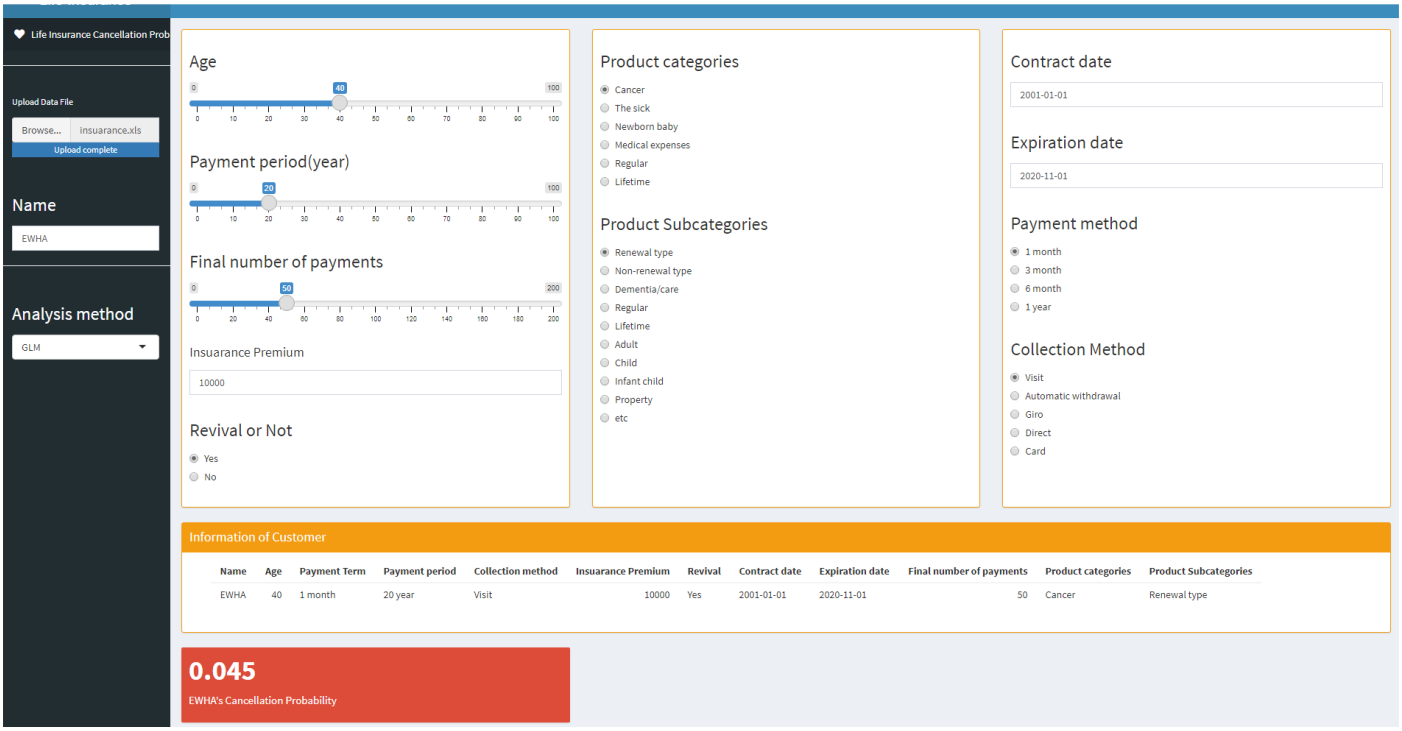


분석을 시행한 후 각 모델 별 Lift Chart를 겹쳐 그려보았다. LDA의 경우 다른 모델들과 scale의 차이가 크고 성능이 좋지 않음을 고려해 최종 그래프에서 제외하였다. 모든 모델이 비교적 좋은 성능을 내고 있음을 파악할 수 있었다.

모형	장점	단점
LDA	• 차원축소를 통해 class 분류에 용이함	• 두 집단 사이 비슷한 형태의 공분산 가정이 필요
KNN	• 기존 데이터에 기반하여 분석하기 때문에 데이터에 대한 가정이 불필요함 • 각 관측치에서 가까운 상위 k개의 데이터만 활용하기 때문에 오류 데이터에 영향을 받지 않음	• 기존의 모든 데이터를 비교하기 때문에 데이터가 클 경우 처리 시간이 길
SVM	• 기존 데이터에 기반하여 분석하기 때문에 데이터에 대한 가정 불필요함 • 분류와 예측 동시 가능함	• 기존의 모든 데이터를 비교하기 때문에 데이터가 클 경우 처리 시간이 길
Random Forest	• 변수 제거 없이 실행되므로 정확도가 높음 • 분류 모델에서 중요 변수 선정 및 랭킹이 가능함	• 텍스트 데이터 등 고차원의 모델에는 부적합함
Neural Network	• 비선형 모델에 용이함 • 입력변수와 출력변수의 이산/범주 형태에 상관없이 모두 출력 가능함	• 결과에 대한 해석 어려움 • 최적 모형 도출이 어려움 • 데이터 정규화 하지 않을 경우 Local minimum에 빠질 위험 있음
XGBoost	• 과적합 규제가 가능함 • Tree puning으로 효율적인 분할을 시행함 • 자체 내장된 교차검증을 통해 조기 중단이 가능함	• 처리 시간이 길

해당 분석들의 장단점을 정리한 표이다. 이를 고려하여 데이터의 특성과 분석의 용도에 맞게 적절한 모델을 선택한다면 예측력이 높은 모델을 적합할 수 있을 것이라 기대된다.

Part 4 : R- Shiny



고객의 정보를 입력하면 해당 고객이 3개월 이내에 보험을 해지할 확률을 각 모형별로 계산해주는 어플리케이션이다.