

— TheoStat2 Project —

생명 보험 해지 확률 계산 및 app 구현

01

DATA

- Original data
- EDA

02

GLM / GAM

- Optimization Model
- Lift chart

03

Data mining - LDA/KNN/SVM/RF/XGB

- Optimization Model
- Lift chart

04

App

| 분석 목적

효과적인
고객 및 상품 관리

보험 해지 확률 계산

간단하게 활용 가능한
app

Original data

- 보험계약 해지여부를 포함한 12개의 변수로 구성된 데이터

변수명	이름	설명	타입
기존변수			
target	이탈여부	0: 이탈하지 않음, 1: 이탈	범주형
age	가입나이		연속형
cycle	납입방법	1: 월납, 3: 3개월납, 6: 6개월납, 12: 연납	범주형
pre_years	납입기간	단위: 년	연속형
t_method	수금방법	1: 방문, 2: 자동이체, 3: 지로, 5: 카드납부	범주형
premium	보험료	1회 납입 보험료	연속형
revival	부활여부	0: 없음, 1: 있음	범주형
cont_date	계약일자	계약한 날짜	날짜형
ins_exp_date	지급만기일자	무만기=9999XXXX, 수명을 77세로 간주해 만기일 설정	날짜형
real_pre_num	최종납입횟수		연속형
type_s	상품 중분류	0-5	범주형
type_m	상품 소분류	0-9	범주형

기존 변수 전처리

변수명	이름	설명	타입
cycle	납입방법	1: 월납, 3: 3개월납, 6: 6개월납, 12: 연납	범주형
t_method	수금방법	1: 방문, 2: 자동이체, 3: 지로, 5: 카드납부	범주형
ins_exp_date	지급만기일자	무만기=9999XXXX 수명을77세*로간주해만기일설정	날짜형
type_s	상품중분류	0-5	범주형
type_m	상품소분류	0-9	범주형

- 변수의 타입 정리
- NA 제거
- 계약상의 날짜인 0229가 윤년에만 존재하므로, 날짜로 바꿀 때 NA가 생성 → 2월 29일을 3월 1일로 정정

파생 변수 생성

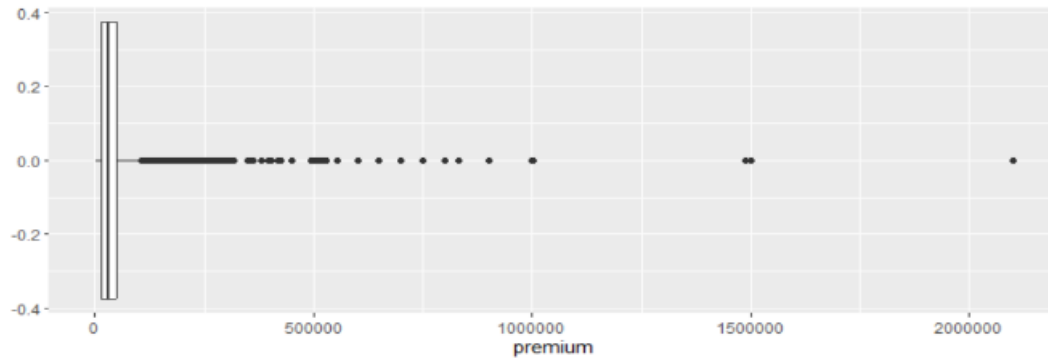
변수명	이름	설명	타입
cont_years	계약기간	계약 일자부터 지금 만기 일자까지 기간(년)	연속형
total_pre_months	최종 납입 개월수	최종 납입 횟수×납입 방법	연속형
pre_ratio	납입 비율	최종 납입 개월수/ 전체 납입 개월수	연속형
ins_exp_month	지급 만기 기간	2001년 6월부터 지급 만기일까지의 기간	연속형
delay	연체 횟수	계약상납입 횟수- 최종 납입 횟수	연속형
join_days	가입 일수	계약 일자로부터 2001.09.30까지의 일수	연속형
real_premium	총 납입 보험료	보험료×최종 납입 횟수	연속형

- 보다 정확한 예측을 위해 최종적으로 7개의 파생변수 생성
- 이때, 지급만기일자가2001.09.30 이전인 데이터는 이미 계약이 종료된 것으로 간주 → 제거

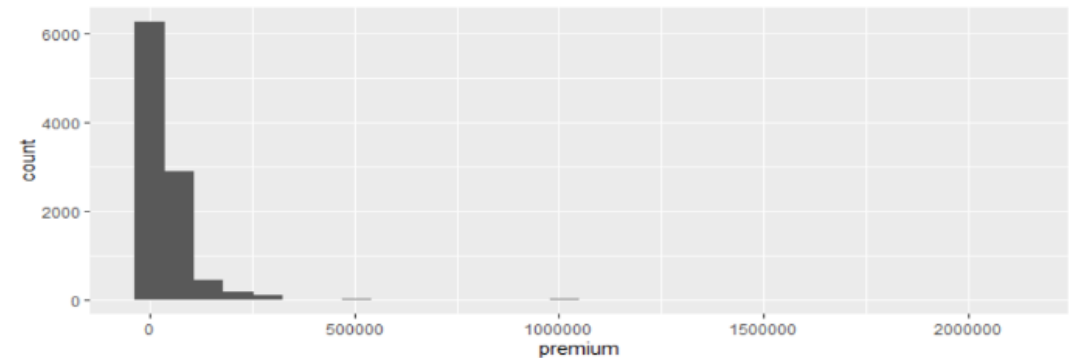
변수 변환

- 연속형 변수들에 대하여 원 데이터와 로그변환한 데이터의 boxplot과 histogram을 비교한 후, 변수 변환 진행
- premium 및 real_premium

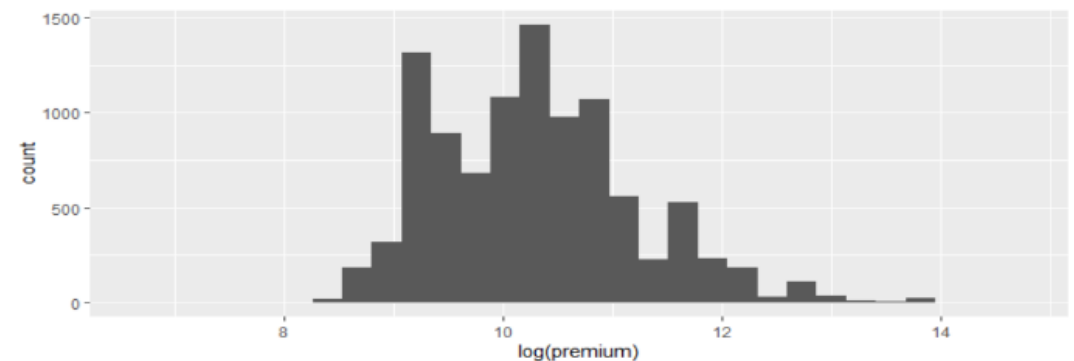
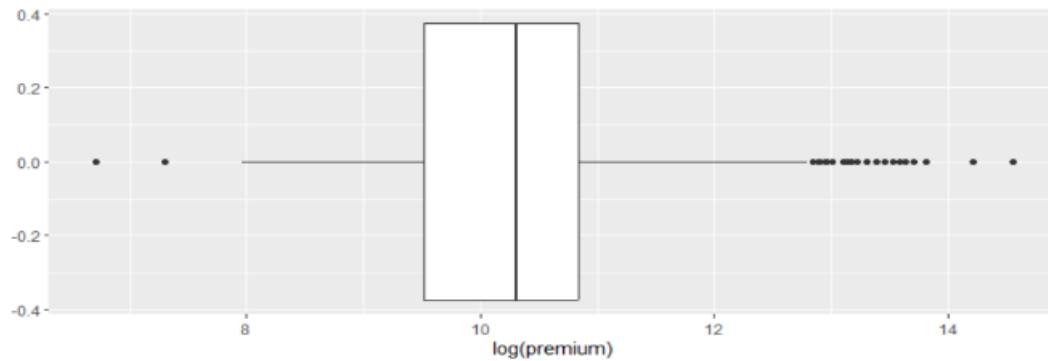
[premium]



[log(premium)]



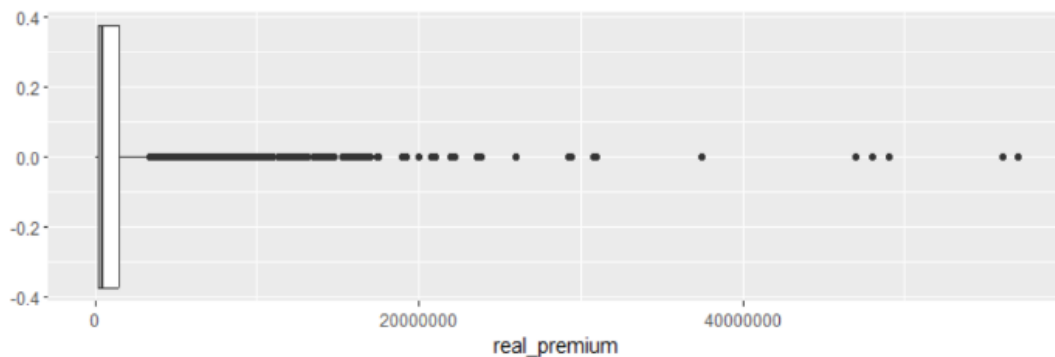
log(premium)



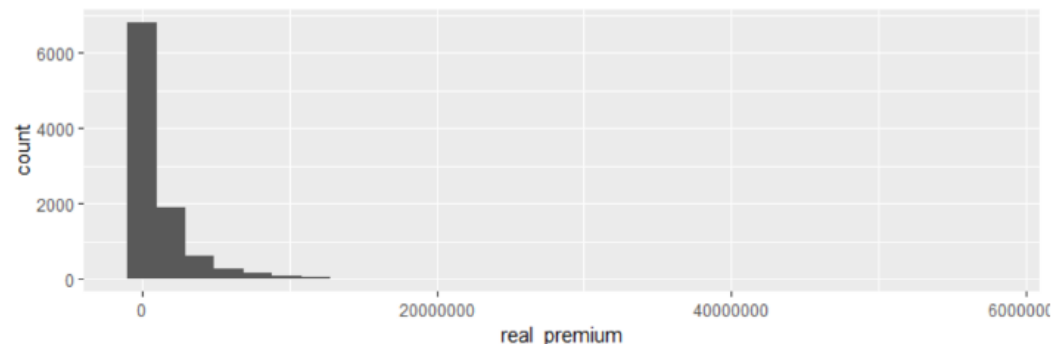
변수 변환

- 연속형 변수들에 대하여 원 데이터와 로그변환한 데이터의 boxplot과 histogram을 비교한 후, 변수 변환 진행
- premium 및 real_premium

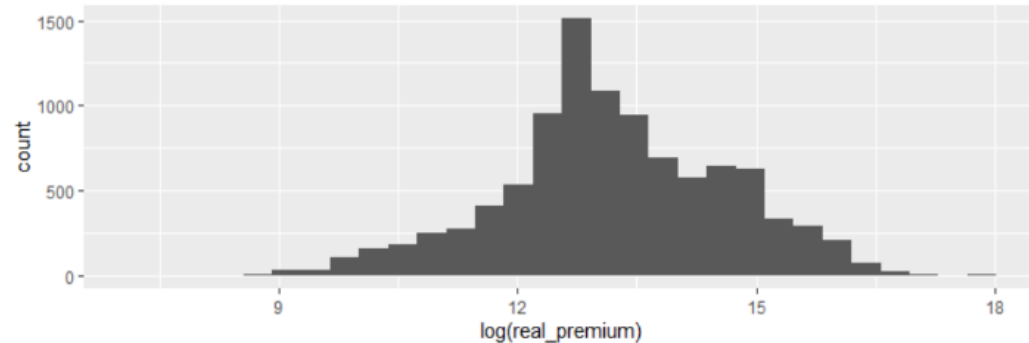
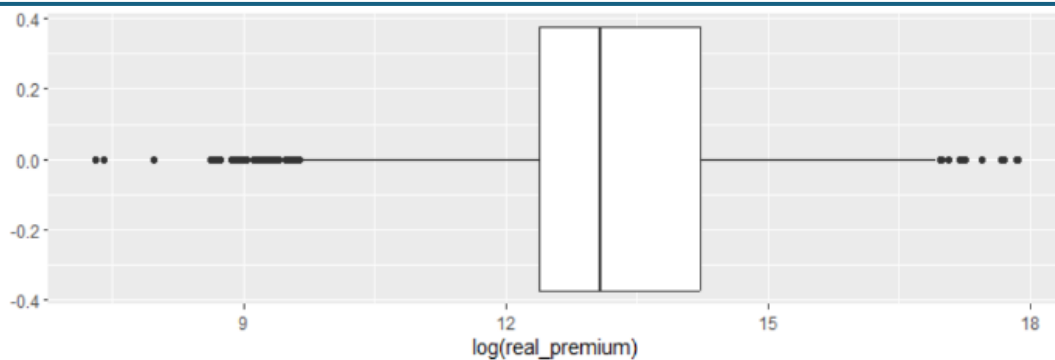
[real_premium]



[log(real_premium)]



log(real_premium)



모형 추정

- 1) link function (probit, logit, cloglog)과 교호작용 포함 여부를 고려한 총 12개의 모형 생성
- 2) step 함수를 사용하여 변수 선택 진행
- 3) AIC를 기준으로 모형 선택

link function	교호작용 포함 여부	(선택된) 변수 개수	AIC
"probit"	X	26	1622.224
		7	1600.812
	O	246	16207.033
		17	1145.874
"logit"	X	26	1622.335
		7	1602.097
	O	246	15630.334
		14	1144.066
"cloglog"	X	26	1622.164
		7	1602.682
	O	246	16062.858
		29	1553.250

최적 모형

· p=14 / 교호 작용 포함 / link function="logit"

GLM model: link function="logit", p=14, AIC=1144.066			
glm(formula = target ~ age + pre_years + premium + revival + total_pre_years + delay + join_days + real_premium + delay:join_days + total_pre_years:delay + delay:real_premium + revival:real_premium + pre_years:delay, family = binomial(link = "logit"), data = train)			
variable	coefficients	variable	coefficients
(intercept)	-6.780269	join_days	0.005847
age	-0.027659	real_premium	-0.187624
pre_years	-0.079667	delay : join_days	-0.013227
premium	0.608619	total_pre_years : delay	0.398201
revival1	-7.041467	delay : real_premium	-0.514926
total_pre_years	-0.173150	revival1 : real_premium	0.496323
delay	10.363008	pre_years : delay	0.073513

모형 추정 및 최적 모형

- 1) link function (logit, probit, gompit link)를 고려한 모형 생성
- 2) stepwise gam 함수를 사용하여 변수 선택 진행
- 3) AIC를 기준으로 모형 선택

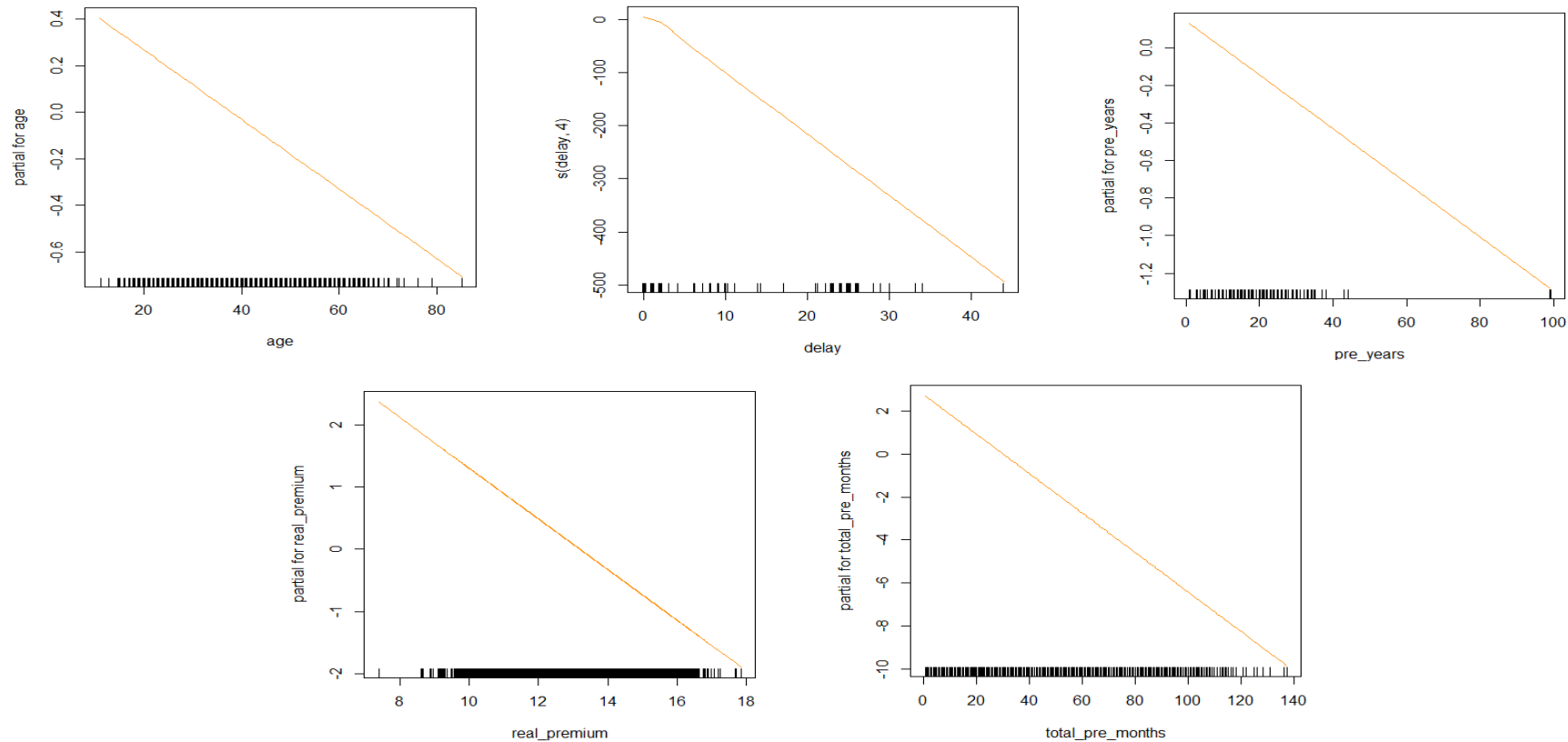
GAM model: link function="probit", p=10, AIC=1149.441

```
gam(formula = target ~ age + pre_years + s(premium, 4) + total_pre_months + s(delay, 4) +  
s(join_days, 8) + real_premium + cycle:delay, family = binomial(link = "probit"), data = train)
```

variable	coefficients	variable	coefficients
(intercept)	-2.510087399	s(join_days, 8)	0.003394933
age	-0.014980335	real_premium	-0.407216143
pre_years	-0.014339527	cycle1:delay	5.669578531
s(premium, 4)	0.565975323	cycle3:delay	9.114748756
total_pre_months	-0.091821306	cycle6:delay	0
s(delay, 4)	-4.387120863	cycle12:delay	0

| 02. GAM : Optimization Model

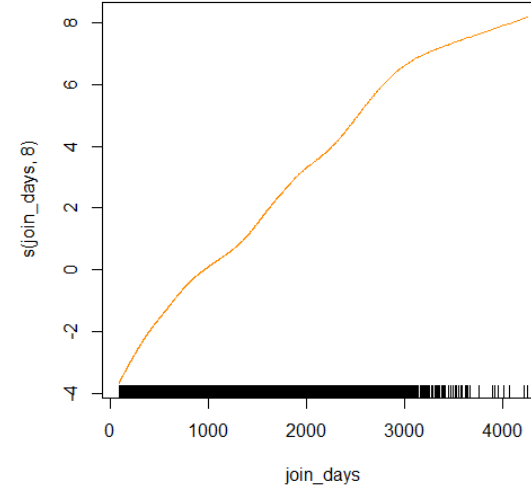
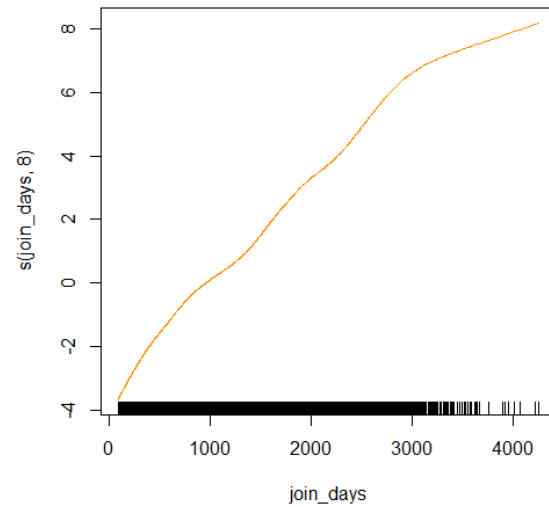
정량변수에 대한 최적 함수 그래프



가입연령, 연체횟수, 납입기간, 총납입보험료, 최종납입개월수 증가 ➡ 해지확률 감소

| 02. GAM : Optimization Model

정량변수에 대한 최적 함수 그래프

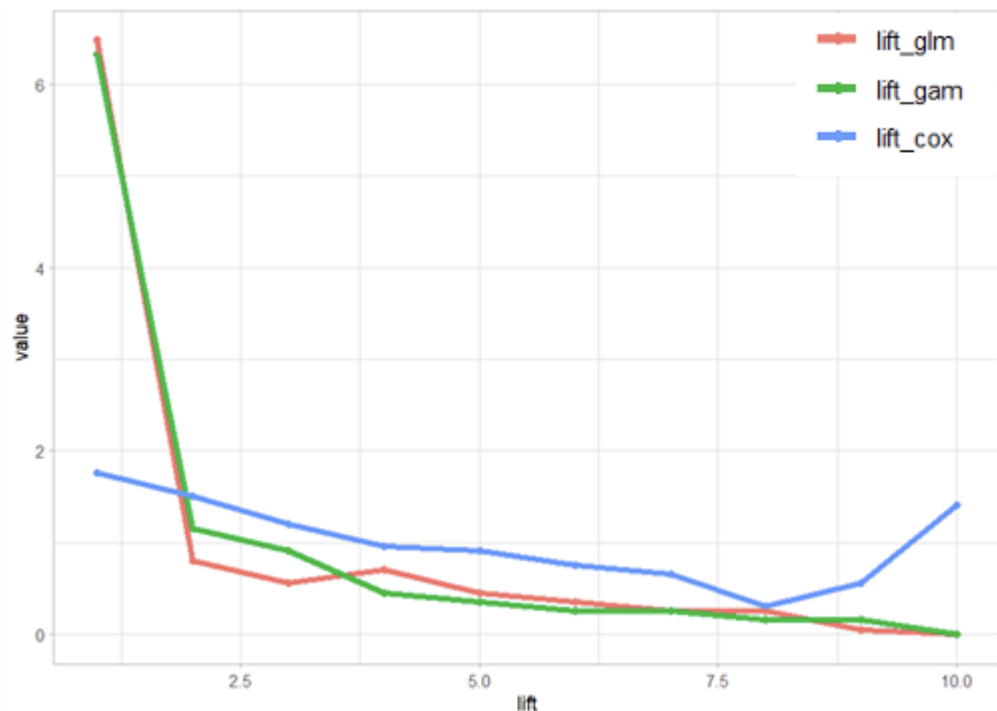


가입일수/ 보험료 증가 ➡ 해지확률 증가

| 02. GLM & GAM : Lift chart

GLM vs GAM

- 1) 확률 및 점수값의 크기순으로 전체 5000명의 고객을 10개의 구간으로 구분
- 2) 각 구간 별 실제 해지 고객의 Lift 값 비교



	AIC	Hit-Ratio
GLM(logit)	1144.066	18.8
GAM(probit)	1149.441	18.4

- GLM, GAM 모두 Lift 값이 감소하는 경향을 보이며, 특히 앞 구간에서 큰 폭으로 감소
- AIC와 test data를 이용한 Hit-ratio을 고려했을 때, **GLM이 더 적절한 모형**으로 판단

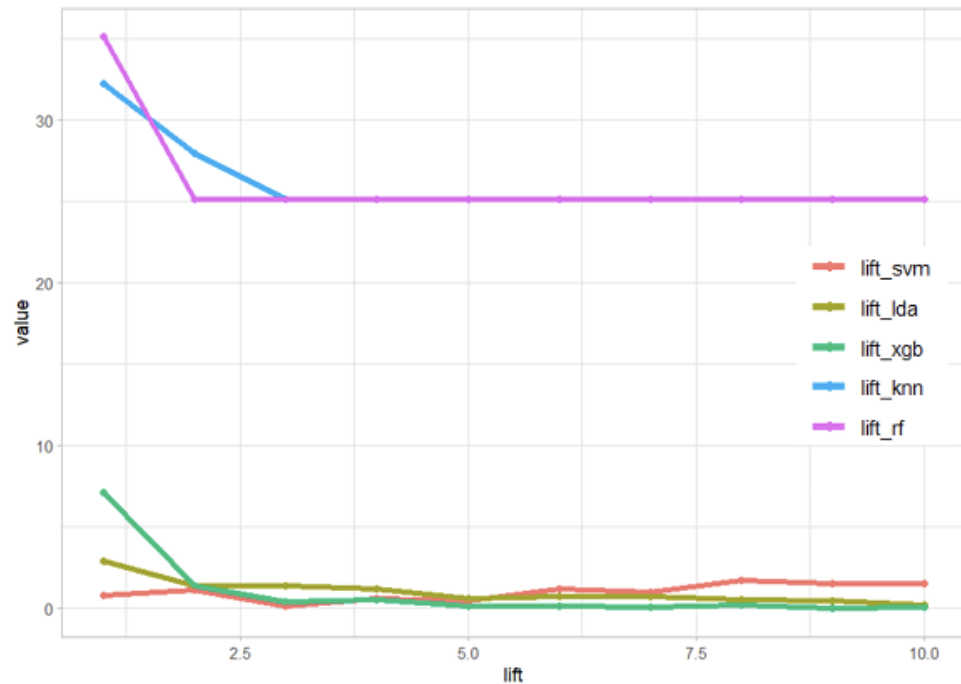
* 각 구간 별 실제 해지 고객의 백분율 및 Lift 값

	GLM(logit)		GAM(probit)		CoxPHM	
	백분율	Lift	백분율	Lift	백분율	Lift
Interval 1	0.2560	6.4232	0.252	6.3228	0.0800	2.0072
Interval 2	0.034	0.8548	0.046	1.1542	0.0580	1.4552
Interval 3	0.0301	0.7542	0.036	0.9033	0.0380	0.9534
Interval 4	0.0200	0.5028	0.018	0.4516	0.0580	1.4552
Interval 5	0.0180	0.4516	0.014	0.3513	0.0300	0.7527
Interval 6	0.0160	0.4023	0.010	0.2509	0.0320	0.8029
Interval 7	0.0100	0.2514	0.010	0.2509	0.0120	0.3011
Interval 8	0.0040	0.1006	0.006	0.1505	0.0260	0.6524
Interval 9	0.0080	0.2011	0.006	0.1505	0.0220	0.5520
Interval 10	0.0020	0.0502	0.000	0.000	0.0426	1.0688

| 03. Data mining : Lift chart

GLM vs GAM

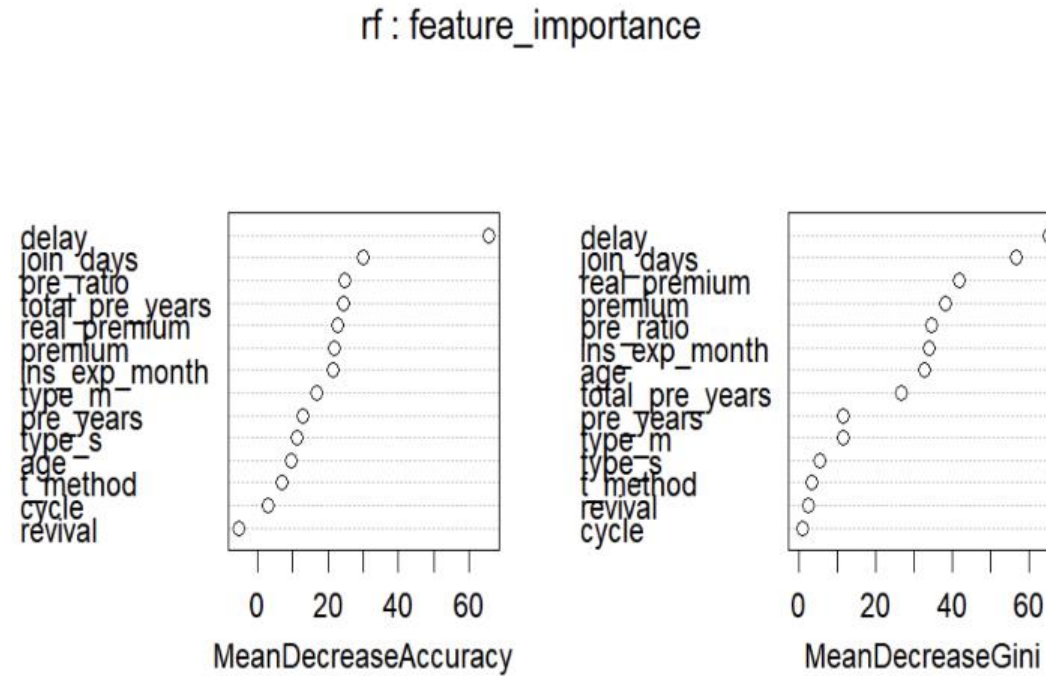
- LDA, KNN, SVM, Random Forest, NN, XGBoost
- 각 구간 별 실제 해지 고객의 Lift 값 비교



	Hit-Ratio
SVM	15.6
LDA	8.6
KNN	4.8
Random Forest	18.8
XGBoost	17.6

- 전반적으로 Lift 값이 감소하는 경향
- test data를 이용한 Hit-ratio을 고려했을 때, Random Forest이 가장 적절한 모형으로 판단

Random Forest



delay > join_days > pre_ratio > total_pre_years > real_premium

- 임의의 보험가입자에 대해 3개월 이내에 보험을 해지할 확률을 자동으로 계산하는 Application

CRM

Upload Data File

Browse... No file selected

고객님의 이름을 입력하세요

삼식이

분석방법

☐ GLM

☐ GAM

☐ CoxPHM

☐ LDA

☐ KNN

☒ RandomForest

☒ SVM

☐ XGBoost

Info1 Info2 Info3

연령

0 40 100

납입 기간 (연)

0 20 100

최종 납입 횟수

0 50 200

보험료

10000

부활유무

☒ Yes


☐ No

Table

names	values
이름	삼식이
연령	40
납입 주기	월납
지급 기간	20년납
수급 방법	방문
보험료	10000
부활유무	Yes
계약일자	2001-01-01
지급 만기 일자	2020-11-01
최종 납입 횟수	50
상품종분류	암
상품소분류	갱신형

0.3%

삼식이님의 보험 해지확률입니다



chlekgp.shinyapps.io/Insurance_Cancellation/

— TheoStat2 Project —
감사합니다