- Data project portfolio -

통계 및 데이터 분석 포트폴리오

이다은

이다은
ekdms8306@gmail.com
GitHub here

학력

2018.02 반송고등학교 졸업
2018.03 ~ 2022.08 한신대학교 응용통계학과 학사 졸업
2022.09 ~ 2024.08 이화여자대학교 일반대학원 통계학과 석사 졸업 예정

skill

R python

어학 및 자격증

OPIc IM2
SQL 개발자 | 한국데이터산업진흥원
ADsP 데이터분석 준전문가 | 한국데이터산업진흥원
사회조사분석사 2급 | 한국산업인력공단

발표 및 교육

2024.07 한국통계학회 하계학술논문 발표회 - 포스터세션
2024.07 ~ ing 2024 Google ML Bootcamp

Applied statistics

- 수리통계학
- 회귀분석 및 실습
- 금융 포트폴리오 분석
- 시계열 분석
- 데이터마이닝
- 비모수 통계자료 분석

Statistics

- 이론통계학
- 베이지안통계
- 확률론
- 고급선형모형
- 통계계산특론
- [석사 학위 논문] 다양한 가설 검정 방법에서의 유의성 분석에 대한 효과크기 계산 및 app 구현

project

- 자동차 보험 데이터를 이용한 보험금 계산 및 app 구현
- 생명 보험 고객 데이터를 이용한 보험 해지 확률 계산
- 기대수명 데이터를 이용한 분류 학습 및 시각화
- [BIGCON] 예술의 전당 공연 예매 데이터를 이용한 공연 장르별 새로운 좌석 그룹화 및 시각화
- [데이콘] 신용카드 사기 거래 탐지 AI 경진대회
- Using MCMC to break classical ciphers - 논문 구현

[석사 학위 논문] 다양한 가설 검정 방법에서의 유의성 분석에 대한 효과크기 계산 및 app 구현

1

연구 배경 및 목적

본 연구의 목적은 귀무가설의 유의성 검정(NHST)에서 연구 결과에 대한 통계적 유의성을 정확히 해석하고 결과에 대한 자세한 내용을 제시하기 위한 방법으로 효과 크기를 설명하는 것이다.

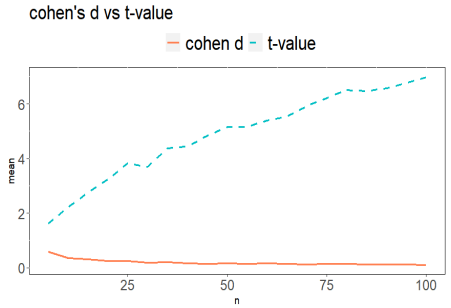
2

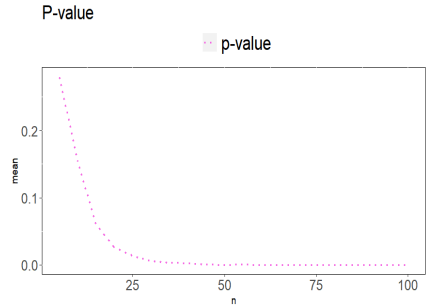
연구 방법

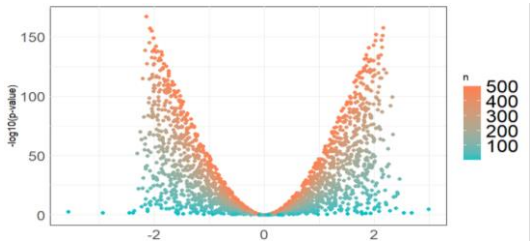
기본적인 통계 검정 방법인 t-통계검정, 회귀분석, ANOVA, 분할 표 검정에 따른 적절한 효과 크기 계산 방법을 세 가지 유형(d-family, r-family, c-family)으로 나누어 정의하고 그 의미를 설명한다. 간단한 데이터를 이용하여 각 효과 크기를 계산하고 결과를 비교해 본다.

3

연구 결과 (https://shinnylee.shinyapps.io/EffectSizeCalculator/)







4

결론

효과 크기는 연구 결과의 유의성의 정도를 나타내며, 유의수준과 함께 제시되어 결과의 실질적인 의미를 해석하는데 도움이 된다. 또한 표준화된 효과 크기는 다양한 연구 간의 결과를 비교할 수 있게 하며, 연구에 필요한 적절한 표본 크기를 결정하는 데에도 사용된다.

project

01 자동차 보험료 계산 및 app 구현

GLM

Application

02 생명보험 해지 확률 예측

GLM/GAM,

LDA/KNN/SVM/RF/XGB

03 기대수명 데이터를 이용한 분류 학습 및 시각화

PCA & k-means

LDA/RF, touring

| 01. 자동차 보험료 계산 및 app 구현

1

분석 목적

· 다양한 조건에서의 자동차 보험료를 계산하는 최적 모형을 추정

· 분석 내용을 쉽게 활용할 수 있도록 application 구현

2

데이터

· 주행거리, 운전 구역, 자동차 종류, 무사고 보너스, 총 사고 건수 등 총 7개의 변수로 구성된 1997년 스웨덴자동차 보험료 데이터

3

분석 방법

1. EDA

· **NA 및 이상치 제거** : 보험가입자수가 0 이하인 데이터 제거

· **변수 type 정리**

2. 분석 모델

1) **분포별 사고 빈도 GLM Model** – 포아송, 이항분포

2**) 분포별 사고 심도 GLM Model** – 포아송, 이항분포

GLM Model - 사고빈도

1) **사고빈도 확률변수가 포아송 / 이항 분포를 따를 때**의 각 회귀 모형 추정
2) Bonus(무사고 보너스)가 범주형/연속형인 경우, 교호작용 포함 여부 고려
3) AIC 및 BIC 기준으로 최적 모형 선택
4) Pearson 표준화잔차 및 Deviance 잔차를 이용하여 모형의 적합도 검토

Result

※ Bonus = 무사고 보너스

		AIC		BIC	
Bonus	구분	Value	선택한 변수 개수	Value	선택한 변수 개수
연속형	Model1	10653.42	25	10970.52	25
	Model2	**10243.59**	237	12648.04	189
범주형	Model3	11702.51	20	11956.19	20
	Model4	11235.50	142	12436.32	94

[포아송 분포]

		AIC		BIC	
Bonus	구분	Value	선택한 변수 개수	Value	선택한 변수 개수
연속형	Model1	10770.98	25	11088.08	25
	Model2	**10226.57**	237	12634.64	189
범주형	Model3	11972.20	20	12225.88	20
	Model4	11402.37	142	12606.98	94

Model1,3 = 1차항을 포함하는 모형(Simple Model)에서 변수선택을 적용한 후 모형
Model2,4 = 1차항과 교호작용항을 포함하는 모형(Full Model)에서 변수선택을 적용한 후 모형

[이항 분포]

- 모형 추정 결과, 두 경우 모두 AIC 기준으로는 Model2, BIC 기준으로는 Model1이 최적 모형으로 판단

- 그러나 교호작용항이 있는 모형(Modle2)은 모형의 복잡성을 증가시키고 변수별 영향력을 파악하기 어려우며, 회귀 모형 추정 결과 유의하지 않은 변수가 더 음

  → 1차항만 존재하는 Model1을 최종 모형으로 선택

GLM Model - 사고심도

1) **사고심도 확률변수가 감마 분포를 따를 때**의 회귀 모형 추정
2) Bonus(무사고 보너스)가 범주형/연속형인 경우, 교호작용 포함 여부 고려
3) AIC 및 BIC 기준으로 최적 모형 선택

Result

		AIC		BIC	
Bonus	구분	Value	선택한 변수 개수	Value	선택한 변수 개수
연속형	Model1	1879318	16	1879472	16
	Model2	1873661	94	1874567	94
범주형	Model3	1878541	21	1878743	21
	Model4	1863828	189	1865649	189

Model1,3 = 1차항을 포함하는 모형(Simple Model)에서 변수선택을 적용한 후 모형
Model2,4 = 1차항과 교호작용항을 포함하는 모형(Full Model)에서 변수선택을 적용한 후 모형

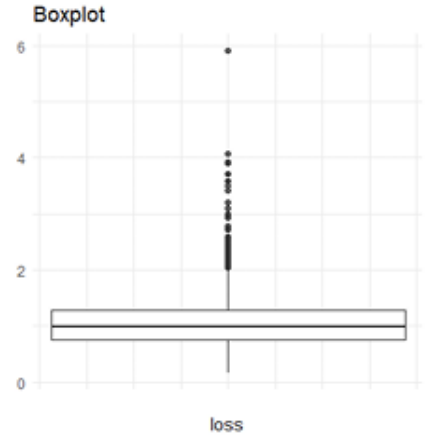
- 모형 추정 결과, AIC 기준으로는 Model4가 최적 모형으로 판단

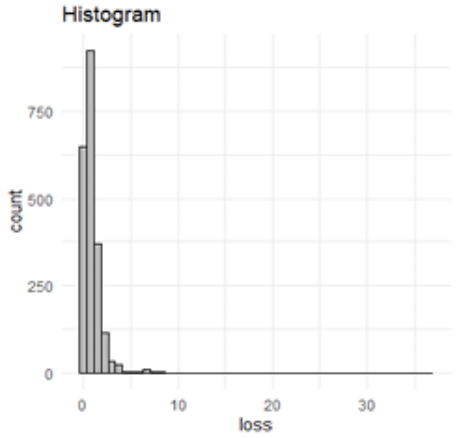
- 그러나 교호작용항이 있는 모형(Modle4)는 포함 변수가 너무 많아 모형의 복잡성을 증가시키고 변수별 영향력을 파악하기 어려우며, 회귀 모형 추정 결과 유의하지 않은 변수가 더 많음 → 교호작용을 포함하지 않는 Model3을 최종 모형으로 선택

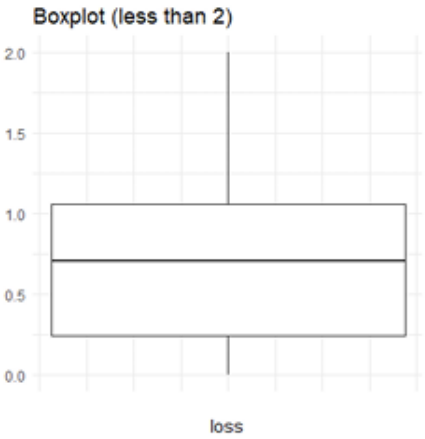
할증 보험료 자동계산 Application 개발

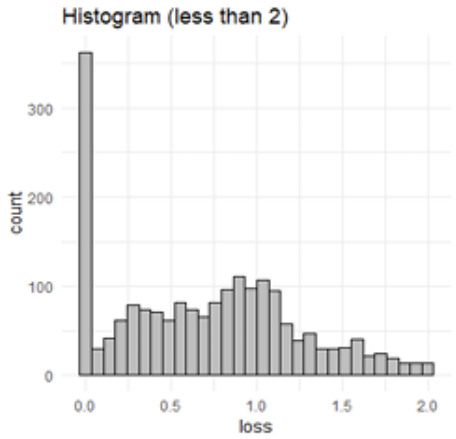
· 수준 조합별 적정 보험료 : 1인당 연간 평균사고빈도 X 1인당 평균사고심도

Result









Min	0.000
Q1	0.2775
Median	0.7883
Mean	1.0080
Q3	1.1737
Max	36.6802
NA	48

1) 각 수준조합별 손해율

· 대부분 0과 1 사이에 집중

→ 최적 회귀 모형들을 이용해 계산한 각 수준별 보험료율은 적절하다고 판단

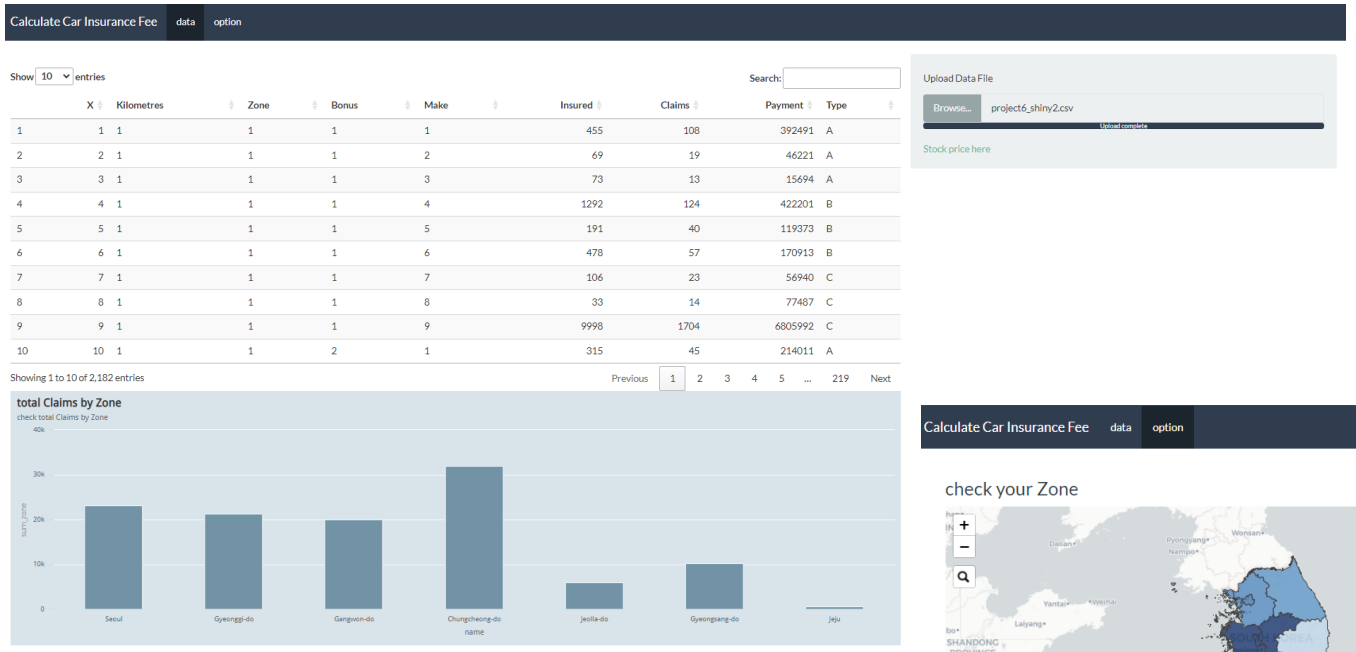
2) 전체손해율

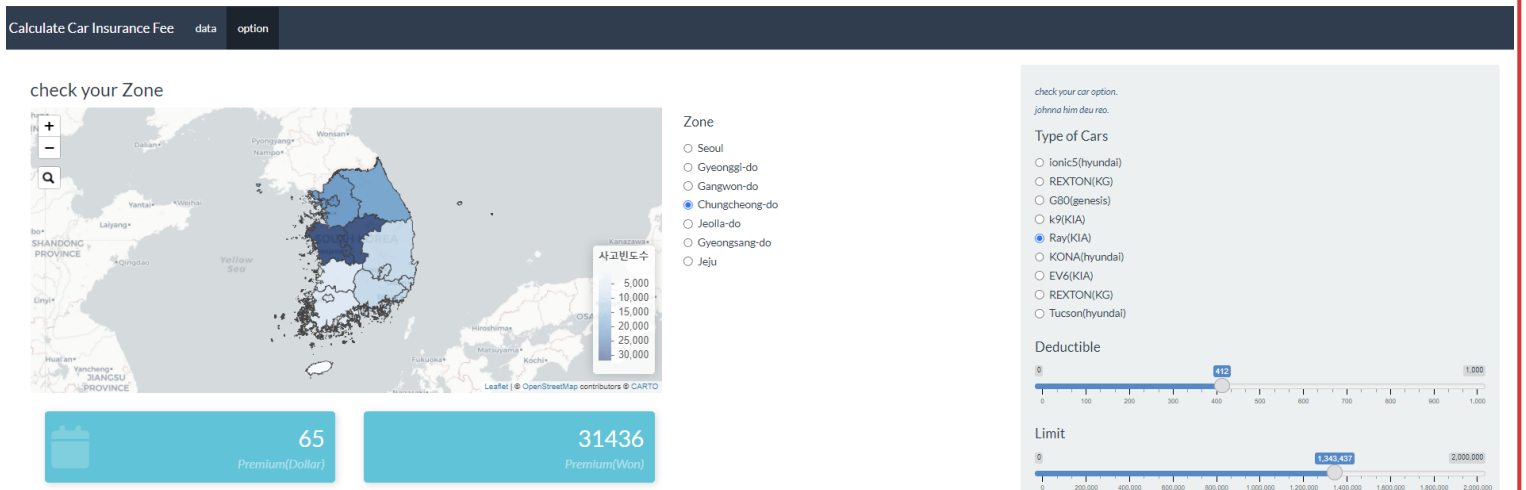
· 1.000009로 1에 매우 근사한 값

→ 이상적인 보험료가 산출되었고, 보험료율은 적정하다고 판단

| 01. 자동차 보험료 계산 및 app 구현

할증 보험료 자동계산 Application 개발





- 사용하고자 하는 데이터 파일을 업로드하여 간단하게 확인 가능
- 운전 지역, 자동차 종류, 보상한도, 자기공제액, 무사고 보너스를 입력하여 보험가입자의 적정보험료 계산 가능

| 02. 생명보험 해지 확률 예측

1

분석 목적

· 보험사 입장에서 효과적인 고객 관리를 위해 보험 해지 확률을 계산하는 최적 모형을 추정

2

데이터

· 나이, 보험료, 이탈 여부, 보험계약 해지여부를 포함한 12개의 변수로 구성된 데이터

3

분석 방법

1. EDA
  · **NA 및 이상치 제거** : 계약 날짜 0229 -> 0301로 변경
  · **총 7개의 파생변수 추가** 생성 : 계약기간, 납입 비율, 연체 횟수, 총 납입 보험료 등
  · 연속형 변수들에 대해 boxplot과 histogram을 비교한 후, 필요한 경우 **로그 변환 진행** : 보험료, 총 납입 보험료

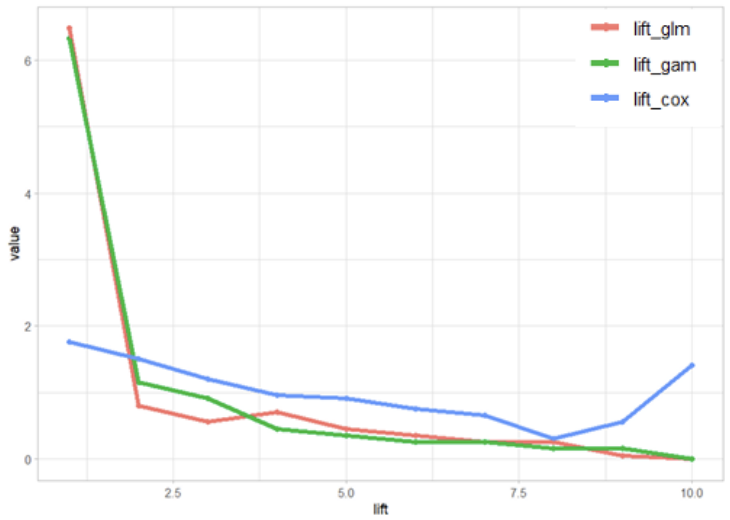
2. 분석 모델
  · **통계 모형** : GLM , GAM
  · **머신러닝 모형** : LDA, KNN, SVM, RF, NN, XGB

  → 각 모형 추정 후, **AIC 및 Lift chart / Hit-ratio 비교하여** 최적 모형 결정

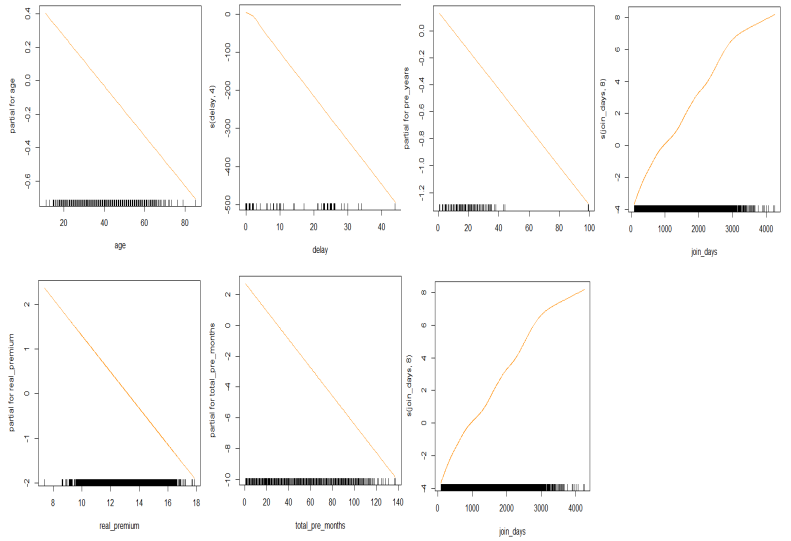
GLM , GAM

1) link function (logit, probit, gompit link)를 고려한 모형 생성
2) stepwise gam 함수를 사용하여 변수 선택 진행
3) AIC를 기준으로 모형 선택
4) 확률 및 점수 값의 크기순으로 전체 5000명의 고객을 10개의 구간으로 구분 → 각 구간 별 실제 해지 고객의 Lift 값 비교

Result



GLM	GAM
p=14 / interaction / link function="logit"	p=10 / interaction / link function=" probit"
AIC = 1144.066	AIC=1149.441
Hit-Ratio = 18.8	Hit-Ratio = 18.4



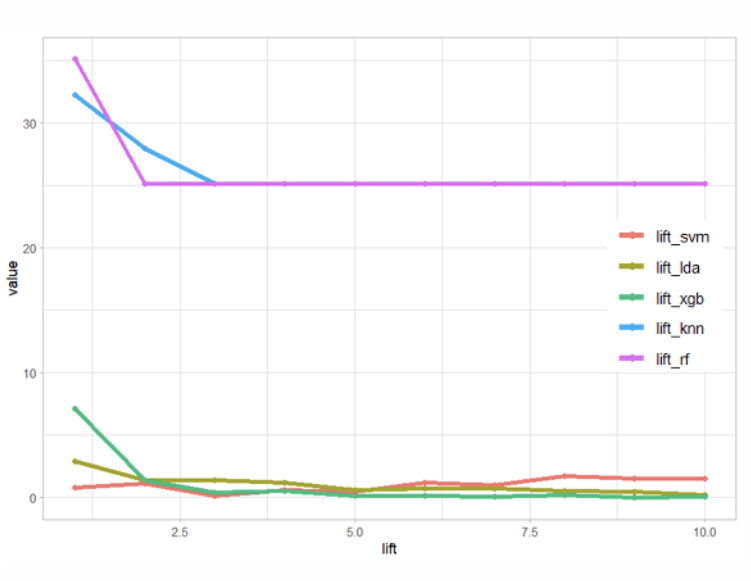
· GLM, GAM 모두 Lift 값이 감소하는 경향을 보이며, 특히 앞 구간에서 큰 폭으로 감소
· AIC와 test data를 이용한 Hit-ratio을 고려했을 때, GLM이 더 적절한 모형으로 판단

· GAM의 정량함수에 대한 최적 함수 결과 → 가입 연령, 연체 횟수, 납입 기간, 총 납입 보험료, 최종 납입 개월수가 증가하면 해지확률은 감소
→ 가입 일수, 보험료가 증가하면 해지확률 증가

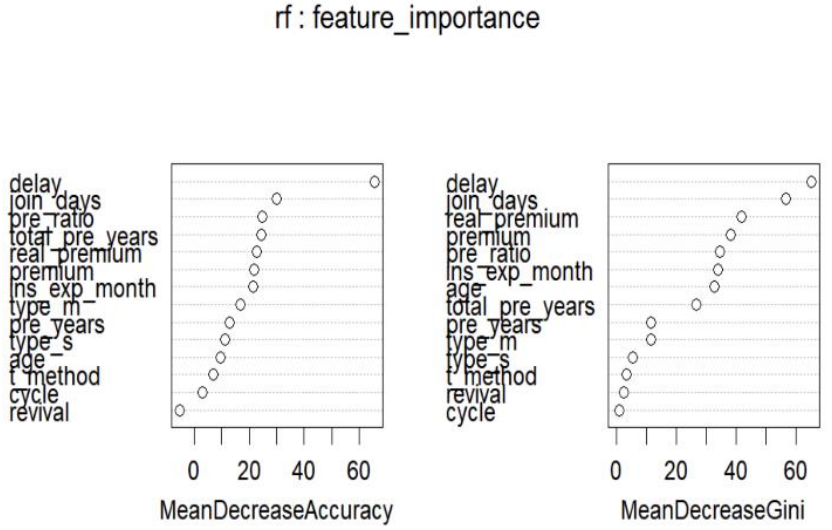
LDA, KNN, SVM, RF, NN, XGB

1) 각 머신러닝 방법을 이용하여 모형 추정
2) 보험 해지 확률 및 점수 값의 크기순으로 전체 5000명의 고객을 10개의 구간으로 구분 → 각 구간 별 실제 해지 고객의 Lift 값 비교

Result



	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 feature importance결과, 연체 횟수 > 가입 일수> 납입 비율 순으로 보험 해지 확률 모형에 중요한 영향을 미친다고 판단

Result

* GLM, GAM으로 추정한 각 구간 별 실제 해지 고객의 백분율 및 Lift 값

	GLM(logit)		GAM(probit)	
	백분율	Lift	백분율	Lift
Interval 1	0.2560	6.4232	0.252	6.3228
Interval 2	0.034	0.8548	0.046	1.1542
Interval 3	0.0301	0.7542	0.036	0.9033
Interval 4	0.0200	0.5028	0.018	0.4516
Interval 5	0.0180	0.4516	0.014	0.3513
Interval 6	0.0160	0.4023	0.010	0.2509
Interval 7	0.0100	0.2514	0.010	0.2509
Interval 8	0.0040	0.1006	0.006	0.1505
Interval 9	0.0080	0.2011	0.006	0.1505
Interval 10	0.0020	0.0502	0.000	0.000

| 03. 기대수명 데이터를 이용한 분류 학습 및 시각화

1 분석 목적

· 국가에 따라 유의미한 차이가 존재하는 건강 지표를 파악하고, 국가 별 특징을 탐색하기 위해 클러스터링 및 분류 분석 진행

2 데이터

· Life_Expectancy_Data.csv (kaggle)
· 여러 국가를 기반으로 한 면역 요인, 사망률 요인, 경제적 요인, 사회적 요인 및 기타 건강 관련 요인으로, 총 22개의 변수가 포함

3 분석 방법

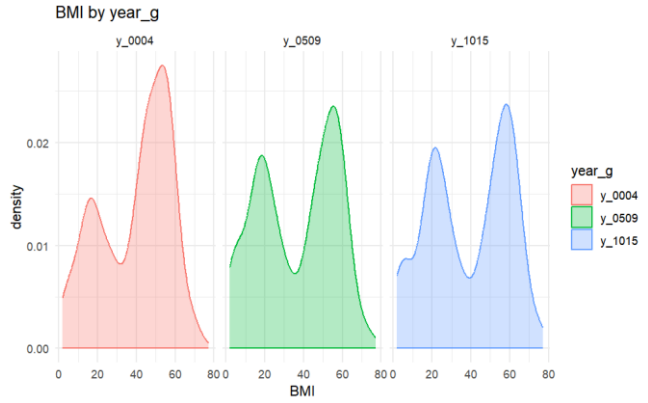
1. EDA
  · **범주화** : 국가 이름(country)를 6개의 대륙으로 재범주화
  · **변수 및 이상치 제거** : 변수 정의 및 plot을 이용하여 홍역 예방 접종률, 5세 이하 아동의 사망자 수, GDP 이상치 제거
  : BMI의 경우 이상치 비율이 높아 변수 제거
  · **다중 공산성 제거** : corrplot을 이용해 상관관계 확인 후, 0.95 이상의 높은 관계성을 가지는 변수 제거 (영아 사망수, 보건 예산 지출 비율)

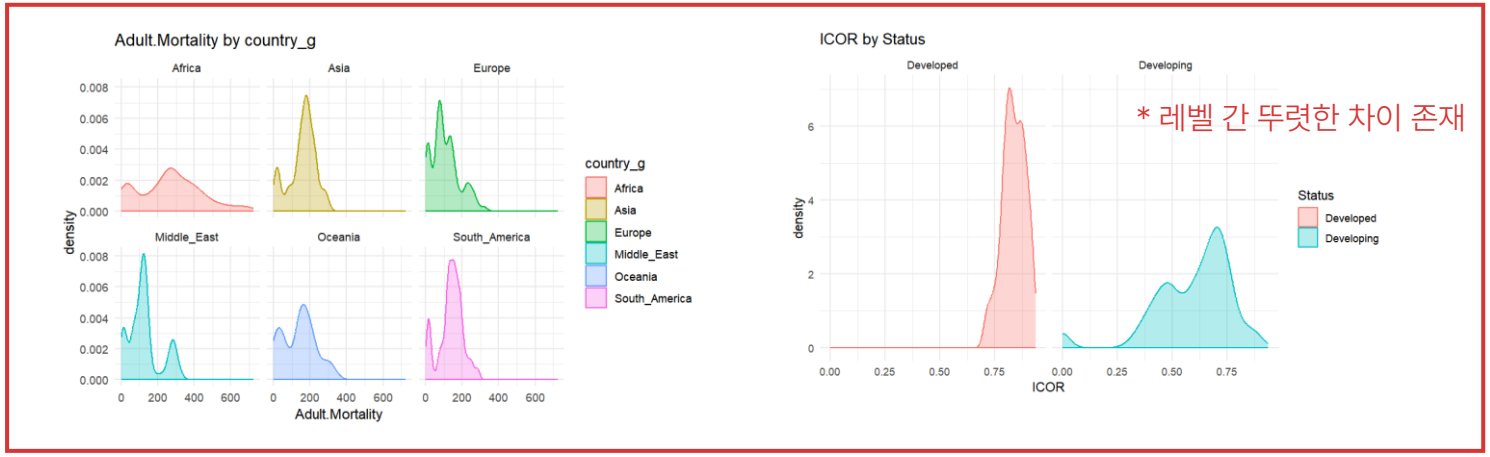
  · **데이터 특징 파악 : 대륙(country_g), 국가의 경제적 상태(status)**의 경우 레벨 간 뚜렷한 분포 차이가 존재하는 것을 확인
  → 두 개의 범주형 변수를 이용하여 분류 분석 진행

2. 분석 모델
  · 비지도 학습 : **PAC & k-means**
  · 지도 학습 : **LDA, RF**
→ 각 모델 추정 후, 성능을 비교하여 최적 모델 선택

데이터 특징

1) **대륙(country_g), 국가의 경제적 상태(status)**의 경우 레벨 간 뚜렷한 분포 차이가 존재하는 것을 확인 → 두 개의 범주형 변수를 이용하여 분류 분석 진행





2) 대륙별로 각 변수의 평균 값을 구하여 특성 확인
- **아프리카와 아시아에서는 건강(성인 사망률, 유아 사망률, HIV.AIDS)과 경제적인 측면(GDP, 기대 수명)에서 부정적**
- **유럽은 기대 수명이 가장 길고 경제적으로 안정**된 상태

2) 경제적 상태별로 각 변수의 평균 값을 구하여 특성 확인
- **Developing 사망 및 부정적인 건강 지표(성인 사망률, 유아 사망률, HIV.AIDS, 저체중율) 에서 부정적**
- 그에 비해 **Developed는 Life와 면역 요인을 나타내는 변수가 높게 나타남**

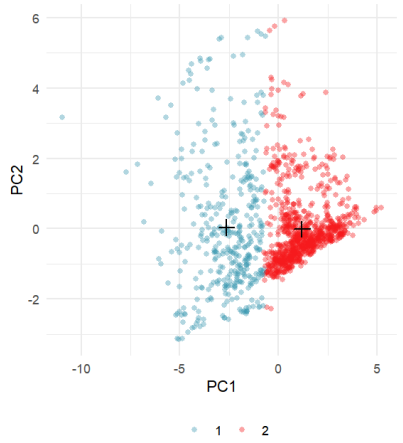
비지도 학습 - PCA & k-means

· PCA를 이용하여 연속형 **변수 15개를 8개의 주성분으로 차원 축소** 진행 (약 85% 누적 분산 기준)
→ 제 1 주성분은 종합적 반영, 제2 주성분은 면역 요인, 제3 주성분은 사망 정보와 관련

· 클러스터 개수를 각각 6개, 2개로 설정하여 k-means 진행 → touring으로 결과 시각화

Result

K = 2



· 각 군집끼리는 잘 뭉쳐져 있으나, 군집 간의 거리가 매우 가까워
경계부분에서는 데이터가 섞여 있는 모습

K = 6

· 클러스터를 6개로 설정했을 때보다 비교적 명확하게 군집 분리

→ k-means를 이용한 클러스터링은 군집 결과(색)이 각 대륙, 경제적 상황을 나타내는 것이 아니므로 실제로 원자료와 대조하여 비교하는 것은 어렵다.

지도학습 - LDA

· 기존 데이터를 스케일링 한 후, train data(7)와 test data(3)를 분할 생성하여 각각 LDA 모델 학습과 예측에 사용 → touring으로 결과 시각화

· LDA의 가정을 만족하지 못하는 상태 → 각 그룹의 분포를 살펴보면 대부분 클러스터의 모양이 일치하지 않는 모습

Result

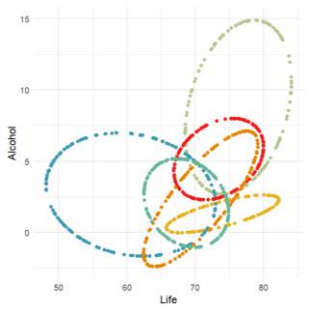
K = 2

· 판별 분석 결과, 면역, 사망, 경제 요인 등의 요인에 의해 영향을 받음
(Life, Alcohol, Measles, Under-Five Deaths)

· 예측 정확도는 67%

· 특히 Asia, South America에서 오분류 데이터 발생

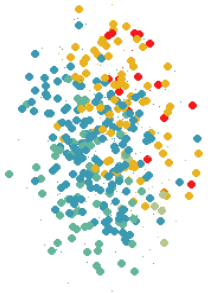
· touring 시각화 결과에서도 같은 군집끼리 뭉쳐져 있는 것이 보이기는 하지만 명확하지 않음



[LDA 가정]

Overall statistics	
Accuracy	0.06667
95% CI	(0.6169,0.7139)
No Information Rate	0.2572
P-value [Acc > NIR]	0.00000000000000022
Kappa	0.586
Mcnemar's Test P-Value	0.00000001025

[table 13. LDA statistics - country_g]



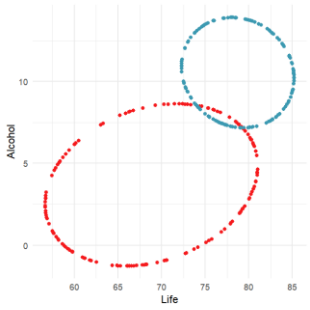
[touring]

K = 6

· 판별 분석 결과, 건강, 사회, 면역 요인에 의해 영향을 받음
(Alcohol, Life, GDP , Measles ,Population)

· 예측 정확도는 92%

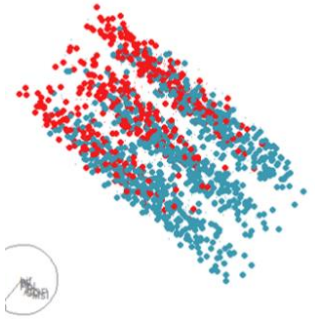
· LDA 가정이 만족되고 클러스터의 개수가 2개로 적은 편이기 때문이라고 판단



[LDA 가정]

Overall statistics	
Accuracy	0.9265
95% CI	(0.8955,0.9506)
No Information Rate	0.7402
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.7941

[table 14. LDA statistics - status]



[touring]

지도학습 – RF

- 기존 데이터를 스케일링 한 후 train data(7)와 test data(3)를 분할 생성하여, 각각 RF 모델 학습과 예측에 사용
- 모델의 성능을 향상시키기 위해 **RMSE 값을 가장 작게 만드는 방향**으로 파라미터 탐색 진행

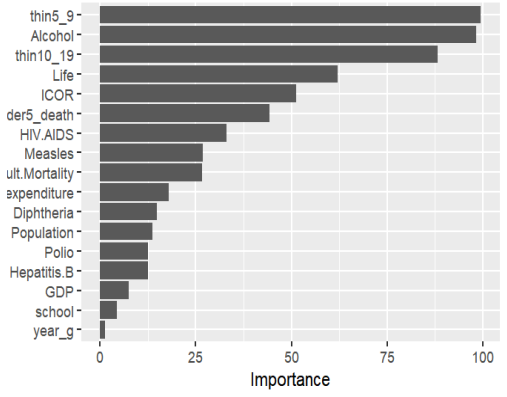
Result

K = 2

- ntree=690, mtry=5, nodesize-3, sample fraction =0.9(replace = FALSE)
- 예측 정확도는 0.9921로 향상
- Asia, South south America, Europe에서 오분류된 데이터 발생
- importance 결과, 저체중 비율(5-9세), 알코올 섭취량, 저체중 비율(10-19세), 기대수명 변수가 큰 중요도를 가짐

Overall statistics	
Accuracy	0.9921
95% CI	(0.9772,0.9984)
No Information Rate	0.294
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.99

[table 17. Rf statistics after – country_g]

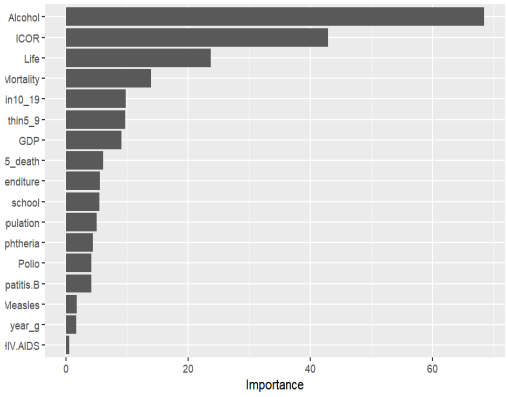


K = 6

- ntrees = 900, mtry = 5, nodesize = 5, sample fraction = 0.9(replace = FALSE)
- 예측 정확도는 0.9948로 향상
- importance 결과, 알코올 섭취량, 기대수명, 성인 사망수변수가 큰 중요도를 가짐

Overall statistics	
Accuracy	0.9921
95% CI	(0.9772,0.9984)
No Information Rate	0.294
P-value [Acc > NIR]	<0.00000000000000022
Kappa	0.99

[table 21. Rf statistics after– Status



결론

LDA		Random forest	
Country_g(6)	Status(2)	Country_g(6)	Status(2)
0.6667	0.9265	0.9921	0.9948
Life, Alcohol, Measles, Under-Five Deaths, Total_expenditure	Alcohol, Life, GDP , Measles ,Population	Thine 5-9, Alcohol, Thine10_19, Life	Alcohol, ICOR, Life , Adult.Mortality

[table 23. LDA vs RF]

- LDA에 비해 RF의 성능이 훨씬 좋음
- RF 결과, 국가를 **대륙별로 분류**할 때 가장 유의미한 변수 → **저체중율(5-9), 기대수명(Alcohol)**
  국가를 **경제적 상태로 분류**할 때 가장 유의미한 변수 → **기대수명(Alcohol), ICOR(소득 분배 및 자원 접근성 종합 지수)**
- 종합적으로, 대륙 간에는 5-9세의 저체중 비율과 기대 수명에 큰 차이가 존재하며, 경제적 상태에 따라 서는 기대수명과 ICOR에 큰 차이가 존재

- Data project portfolio -

감사합니다