미국 총기 사고 예측

빅데이터 기술

201303009

이진웅



목차

[교수님 피드백 2](#_Toc517213108)

[발표 당시의 문제점 2](#_Toc517213109)

[수정한 점 2](#_Toc517213110)

[데이터 소개 3](#_Toc517213111)

[데이터 수정 3](#_Toc517213112)

[데이터 탐색 4](#_Toc517213113)

[데이터 예측 및 정확도 9](#_Toc517213114)

[결론 13](#_Toc517213115)

[느낀 점 13](#_Toc517213116)

# 교수님 피드백

## 발표 당시의 문제점

|  |
| --- |
| - 데이터 분석, 예측 시 제외하는 열 없이 분석, 예측 해보고  정확도 비교해보기  - 데이터 분할 시 비율이 너무 적다. |

## 수정한 점

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| * 문자열 열도 분석에 참여시키기 위해 데이터 형 변환  |  |  |  | | --- | --- | --- | | 열 이름 | 변환 전 | 변환 후 | | Intent(사망유형) | Suicide(자살)  Accidental(우연사)  Homicide(살인) | 0  1  2 | | Police(경찰개입여부) | 경찰 개입x  경찰 개입o | 0  1 | | Sex(성별) | 남자  여자 | 0  1 | | Race(인종) | White(백인)  Asian(아시아인)  Black(흑인)  Native American(미국인)  Hispanic(히스패닉) | 0  1  2  3  4 | | Place(사망 장소) | Home(집)  Street(거리)  Other specified(기타 장소)  Trade/service area(서비스지역)  School/instiution(학교)  Other unspecified(기타불특정) | 0  1  2  3  4  5 | | Education(최종학력) | 고졸미만  고졸  대재  대졸  확인불가 | 0  1  2  3  4 |  * 훈련데이터의 비율이 너무 적다.  |  | | --- | | 기존 훈련데이터(10%), 검증데이터(90%) 에서  변경 훈련데이터(70%), 검증데이터(30%) 으로 수정  ------------------------------------------------------------------------------  test\_idx <- createDataPartition(gun$intent, p=0.7)$Resample1  # 훈련데이터(70%), 검증데이터(30%) | |

# 데이터 소개

* 2012 ~ 2014년 미국의 총기 사망사고에 대한 정보

|  |
| --- |
|  |

# 데이터 수정

|  |
| --- |
| gun <- read.csv("C:/Users/kslbs/Desktop/guns수치화.csv", header = T, stringsAsFactors = FALSE)  #read.csv로 csv파일 불러오기 |
| gun <- gun[, !names(gun) %in% c("X","hispanic")]  #불필요 컬럼 제거 |
| gun <- na.omit(gun)  #gun 데이터에 결측치 제거 |
| gun$sex <- as.numeric(gun$sex)  gun$police <- as.numeric(gun$police)  gun$race <- as.numeric(gun$race)  gun$place <- as.numeric(gun$place)  gun$intent <- as.numeric(gun$intent)  gun$year <- as.numeric(gun$year)  gun$month <- as.numeric(gun$month)  gun$age <- as.numeric(gun$age)  gun$education <- as.numeric(gun$education)  #데이터 타입 지정 |

# 데이터 탐색

|  |
| --- |
| test\_idx <- createDataPartition(gun$intent, p=0.7)$Resample1  #Y 값을 고려한 데이터의 분할(훈련데이터70%, 검증데이터30%) |
| gun.test <- gun[test\_idx,]  gun.train <- gun[-test\_idx,]  nrow(gun.test)  nrow(gun.train)  #test데이터와 train데이터로 분리 |
| prop.table(table(gun.train$intent))  #gun데이터의 사망유형 비율 |
| createFolds(gun.train$intent, k=10)  #데이터 분리  create\_ten\_fold\_cv <- function() {  set.seed(137)  lapply(createFolds(gun.train$intent , k=10), function(idx) {  return(list(train=gun.train[-idx, ],  validation=gun.train[idx, ]))  })  }  #10겹 교차 검증 데이터를 만드는 함수 |
| summary(intent ~ year + month + police + sex + race + education, data = data, method = "reverse")  # 각 변수 값에 따른 사망유형 종류 |
| xtabs( ~ intent + race , data=data)  # 사망사유, 인종별 분할표(xtabs) |
| xtabs( ~ year + intent , data=data)  # 연도, 사망사유별 분할표(xtabs) |
| xtabs( ~ sex + year , data=data)  # 연도, 성별 별 사망자 수 분할표(xtabs) |
| xtabs( ~ year+ race , data=data)  # 연도, 인종별 사망자 수 분할표(xtabs) |
| xtabs(intent == "0" ~ year+ race , data=data)  # 연도, 인종별 사망자 수 분할표(사망 사유가 '자살') |
| xtabs(race == "White" ~ sex + year , data=data) / xtabs(race == "Black" ~ sex + year , data=data)  # 연도, 성별 별 사망자 수 분할표(백인 사망 수/흑인 사망 수 >>> 남자는 4~5배, 여자는 2~3배 차이) |

데이터 시각화

|  |
| --- |
| * featurePlot(   data.complete[,sapply(names(data.complete),function(n) { is.numeric(data.complete [, n]) })],  data.complete [, c("intent")], "ellipse")  #featureplot을 이용한 데이터 시각화   1. 데이터 수정 전   C:\Users\kslbs\Desktop\R기말프로젝트\피쳐플롯.PNG   1. 데이터 수정 후(참여 열 추가 + 형 변환)   C:\Users\kslbs\Desktop\수치화로 바꾸고 피쳐플롯.PNG |
| * mosaicplot(intent ~ race + sex, data = data, color=TRUE, main="guns")   #인종과 성별로 구분한 사망유형(mosaicplot), 사망유형과 성별로 비교한 모자이크플롯  C:\Users\kslbs\Desktop\R기말프로젝트\모자이크플롯.PNG |

.

# 데이터 예측 및 정확도

|  |
| --- |
| **Rpart 예측모델** |
| m <- rpart(intent ~ year + month + police + sex + age + race + place + education, data=gun.train)  p <- predict(m, newdata = gun.train, type = "class")  head(p)  # rpart 모델만듬 |
| folds <- create\_ten\_fold\_cv()  rpart\_result <- foreach(f=folds) %do% {  model\_rpart <- rpart(intent ~ year + month + police + sex + age + race + place + education,  data=f$train)  predicted <- predict(model\_rpart, newdata=f$validation, type="class")  return(list(actual=f$validation$intent, predicted=predicted))}  # folds 전체에 대한 결과를 리스트로 묶어서 변수에 저장 |
| evaluation <- function(lst) {  accuracy <- sapply(lst, function(one\_result) {  return(sum(one\_result$predicted == one\_result$actual)  / NROW(one\_result$actual))  })  print(sprintf("MEAN +/- SD: %.3f +/- %.3f", mean(accuracy), sd(accuracy)))  return(accuracy)  }  #평균과 표준편차를 계산한 뒤 Accuracy의 벡터를 결과로 반환 |
| evaluation(rpart\_result)  rpart\_accuracy <- evaluation(rpart\_result)  #rpart 모델의 성능 : 82.2%, 오차범위 : 0.010 |
| **ctree 예측모델** |
| **ctree\_result <- foreach(f=folds) %do% {**  **model\_ctree <- ctree(intent ~ year + month + police + sex + age + race + place + education,**  **data=f$train)**  **predicted <- predict(model\_ctree , newdata=f$validation, type="response")**  **return(list(actual=f$validation$intent , predicted=predicted))**  **}**  **#ctree :type에 response를 지정해야 class가 반환** |
| **ctree\_accuracy <- evaluation(ctree\_result)**  **#cpart 모델의 성능: 83.2%, 오차범위 : 0.008(rpart와 1%나 차이가난다.)** |
| **Rpart 와 ctree 정확도 비교** |
| 1. **데이터 수정 전**      1. **데이터 수정 후** (참여 열 추가 + 형 변환) |
| **피어슨 상관계수** |
| **corrg <-cor (gun[,c("intent","police","sex","age","race","education")])**  **#상관계수 비교할 열만 추출하여 변수에 저장** |
| **corrgram(corrg,type = "corr",upper.panel = panel.conf)**  **#corr상관계수 그래프 그리기**   1. 데이터 수정 전      1. 데이터 수정 후(참여 열 추가 + 형 변환) |
| **피어슨 상관계수를 통한 분할표** |
| **인종과 사망유형 별 분할표**     |  |  |  | | --- | --- | --- | | **백인과 흑인 자살율 1위**  **백인과 흑인 살해율 1위**  **백인과 흑인 돌발사 1위** | **Race(인종)**  **0 : 아시아인**  **1 : 백인**  **2 : 토종 미국인**  **3: 흑인**  **4: 라인계** | **intent(사망유형)**  **0 : 자살**  **1 : 타살**  **2 : 돌발사** | |
| **해를 거듭 할 수록 미국의 자살율이 늘고있다** |
| **#**  **예측한 값 중 정확히 예측한 값의 비율(정확도 = 0.6104497)** |

# 결론

|  |
| --- |
| 피드백을 받기 전까지 10만개가 넘는 데이터 중에 훈련데이터를 10%밖에 두지 않았다. 피드백을 받고 훈련데이터와 검증데이터의 비율을 7:3으로 바꾼 후 다시 분석 및 예측을 해보니 확실히 결과가 다르게 나왔다. Rpart와 ctree에서의 정확도도 살짝 다르게 나왔다. 또한 예측에 참여하는 열도 추가로 늘려서 분석을 해보니 결과가 바뀌는 것을 알 수 있었다. |

# 느낀 점

|  |
| --- |
| 한 학기 동안 배운 것을 기반으로 프로젝트를 진행하였는데 처음엔 예측 및 분석에 사용하기 적절한 데이터를 찾는 것에 시간이 오래 걸렸다. 검색 결과 한국 공공 데이터 말고도 해외 공공데이터를 구할 수 있는 사이트를 찾게 되었고, 그곳에서 미국의 총기 사고에 관한 데이터를 찾았다. 이 데이터로 프로젝트를 진행하기에 적합하다 생각되어 프로젝트를 진행하게 되었다. 데이터 탐색 부분까지는 순조롭게 진행을 했지만 데이터 예측(rpart & ctree) 쪽에서 많이 어려움을 겪었다. 교재와 인터넷으로 공부를 하며 잘못 된 부분을 찾을 수 있었고 예측 부분까지도 끝낼 수 있었다. 발표 후 데이터 열을 늘려보라는 것과 훈련 데이터 비율을 늘려보라는 교수님의 피드백을 받았고, 피드백을 기준으로 코드를 수정 해 나갔다. 코드 수정 후 결과를 보니 확실히 차이가 있었다. 10만줄을 빅데이터라고 부르기에는 터무니없지만 피드백을 통해 느낀 것은 데이터의 양이 많아 졌을 때 이런 사소한 설정으로도 차이가 크게 날 수 있다는 점이었다. |