대구 교통사고 데이터 분석

빅데이터개론 기말 프로젝트 결과 보고서
빅데이터 전공 20215123 김수연

ⓒSaebyeol Yu. Saebyeol's PowerPoint

목차

table of contents

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1

주제선정 배경/ 데이터 소개

ⓒSaebyeol Yu. Saebyeol's PowerPoint

주제 선정 배경

데이터 분석으로 유의미한 결과를 분석하고 예측하는 것은
결정에 도움을 주는 것에서 나아가,
사회 현상을 해결하는 데 도움이 됨.
해당 프로젝트에서는 교통사고 데이터를 분석함.
그 피해 정도를 예측해, 피해가 크고 작은 교통사고들에
영향을 끼치는 요인은 무엇인지 알아보고
피해를 줄이는 데 도움이 되고자 주제를 선택함.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 소개

```
train <- fread("/content/train.csv", header = T, encoding = "UTF-8") %>% as_tibble()
train %>% show()
```

```
# A tibble: 39,609 × 23
   ID              사고일시      요일  기상상태 시군구 도로형태 노면상태 사고유형
   <chr>           <chr>         <chr> <chr>    <chr>  <chr>    <chr>    <chr>
 1 ACCIDENT_00000 2019-01-01 00 화요… 맑음     대구… 단일로 … 건조     차대사람
 2 ACCIDENT_00001 2019-01-01 00 화요… 흐림     대구… 단일로 … 건조     차대사람
 3 ACCIDENT_00002 2019-01-01 01 화요… 맑음     대구… 단일로 … 건조     차대사람
 4 ACCIDENT_00003 2019-01-01 02 화요… 맑음     대구… 단일로 … 건조     차대차
 5 ACCIDENT_00004 2019-01-01 04 화요… 맑음     대구… 단일로 … 건조     차대차
```

- DACON에서 제공한 데이터.
- 요일, 기상상태, 시군구, 도로형태, 노면상태 등 총 23개의 독립변수, 39,610개의 행으로 구성되어 있다.

-데이터출처: https://dacon.io/competitions/official/236193/overview/description

ⓒSaebyeol Yu. Saebyeol's PowerPoint

2

데이터 취득과 정제 / 가공

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터취득과정제/가공

2-1) 컬럼 이름 공백 제거

```
# 컬럼 이름 공백 제거

colnames(train)[9]<-'사고유형세부분류'

colnames(train)[11]<-'가해운전자차종'
colnames(train)[12]<-'가해운전자성별'
colnames(train)[13]<-'가해운전자연령'
colnames(train)[14]<-'가해운전자상해정도'

colnames(train)[15]<-'피해운전자차종'
colnames(train)[16]<-'피해운전자성별'
colnames(train)[17]<-'피해운전자연령'

colnames(train)[18]<-'피해운전자상해정도'

head(train,2)
```

A tibble: 2 × 23

군구	도로형태	노면상태	사고유형	사고유형세부분류	법규위반	...	가해운전자상해정도	피해운전자차종	피해운전자성별	피해운전자연령	피해운전자상해정도	사망자수	중상자수	경상자수	부상자수	ECLO
r>	<chr>	<chr>	<chr>	<chr>	<chr>	...	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<int>	<int>	<int>	<int>
구역중대동	단일로-기타	건조	차대사람	길가장자리구역통행중	안전운전불이행	...	상해없음	보행자	여	70세	중상	0	1	0	0	5

- 컬럼 사용 시 공백이 있는 컬럼에서 오류가 나는 경우가 있어 컬럼 이름의 공백을 전부 제거했다.

2-2) 변수 factor 타입으로 변경

```
# 범주형 컬럼 factor 타입으로 변경

train$요일<- as.factor(train$요일)
train$기상상태<- as.factor(train$기상상태)
train$도로형태<- as.factor(train$도로형태)
train$노면상태<- as.factor(train$노면상태)
train$사고유형<- as.factor(train$사고유형)
train$사고유형세부분류<- as.factor(train$사고유형세부분류)
train$법규위반<- as.factor(train$법규위반)
train$가해운전자차종<- as.factor(train$가해운전자차종)
train$가해운전자성별<- as.factor(train$가해운전자성별)
train$가해운전자상해정도<- as.factor(train$가해운전자상해정도)
train$법규위반<- as.factor(train$법규위반)
train$가해운전자차종<- as.factor(train$가해운전자차종)
train$가해운전자성별<- as.factor(train$가해운전자성별)
train$가해운전자상해정도<- as.factor(train$가해운전자상해정도)
```

- 데이터의 원활한 처리를 위해 범주형 데이터를 가지는 컬럼을 모두 factor형으로 변경했다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터취득과정제/가공

2-3) 연령 관련 컬럼의 타입을 숫자형으로 변경

```
# 가해운전자연령, 피해운전자연령 int형으로 변경

train$가해운전자연령 <- substr(train$가해운전자연령, 1, 2)#"세"빼고 나이 부분만 자르기(앞 숫자 2개)
head(train$가해운전자연령)
train$피해운전자연령 <- substr(train$피해운전자연령, 1, 2) #"세"빼고 나이 부분만 자르기
head(train$피해운전자연령)
```

```
'51' '39' '70' '49' '30' '52'
'70' '61' '38' '36' '52' '35'
```

```
train$가해운전자연령 <- as.numeric(train$가해운전자연령)
train$피해운전자연령 <- as.numeric(train$피해운전자연령)
typeof(train$가해운전자연령)
typeof(train$피해운전자연령)

head(train,2)
```

피해운전자연령		피해운전자연령
<chr>		<dbl>
70세	→	70
61세		61

- 나이로 이루어진 '가해 운전자 연령', '피해 운전자 연령'열을 숫자로 처리하기 위해 "23세"와 같이 나이+"세"로 이루어진 데이터에서 substr()을 통해 "세"를 제외한 앞의 두 숫자를 분리했다. 그 후 as.numeric()을 이용해 숫자형으로 변경했다.
- 그 결과 열의 타입이 <chr>에서 <dbl>가 되었다.

2-4) tibble()형 데이터로 변환

```
train<- as.tibble(train)
head(train, 2)
```

- 데이터프레임 처리 시 편의성과 일관성을 위해 as.tibble()을 이용해 데이터를 tibble형으로 변환했다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터취득과정제/가공

2-5) 결측값 처리

```
[ ]  table(is.na(train)) # 결
```

```
 FALSE    TRUE
908491    2516
```

```
[ ]  train = na.omit(train)
     table(is.na(train))
```

```
 FALSE
853944
```

- 결측값 확인 결과 908,491개 데이터 중 2,516개의 결측값이 발견되었다. 2,516개의 데이터를 제거해도 충분한 양의 데이터가 확보되므로 na.omit()으로 제거했다.

2-6) summary, boxplot확인

-읽어온 데이터의 읽어온 데이터(train)의 summary와 train의 피해운전자연령, 가해운전자연령 열의 boxplot을 확인했다.

- summary 확인 결과 '기상상태'열의 대부분은 '맑음'으로 구성되어 있고, '노면상태'열의 대부분은 '건조'로 구성되어 있다. ECLO의 최솟갑이 1, 1st Qu.가 3, 평균이 4.8, 3st Qu.가 6, 최댓값이 74로 평균과 멀리 떨어진 최댓값이 존재하는 것을 알 수 있다.

- boxplot 확인 결과 피해운전자연령과 가해운전자연령 모두 대부분 30~60세에 분포한다.

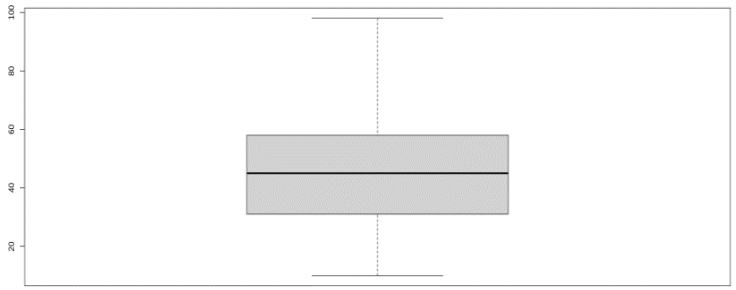
ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터취득과정제/가공

2-6) summary, boxplot확인

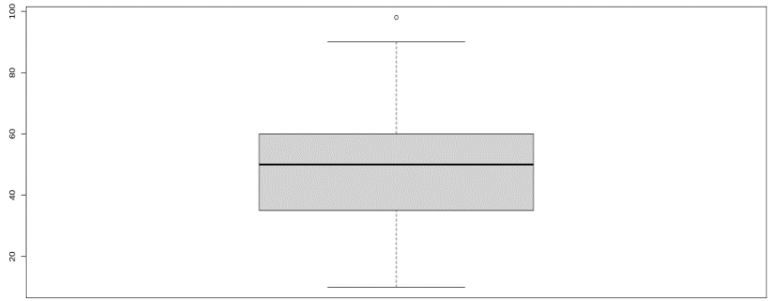
```
#16번
table(is.na(train$피해운전자연령))
boxplot(train$피해운전자연령) #B 이상값 여부 boxplot()으로 확인
```

```
FALSE
37128
```



```
table(is.na(train$가해운전자연령)) # NA의 개수 출력
boxplot(train$가해운전자연령) # 이상값 여부 boxplot()으로 확인
```

```
FALSE
37128
```



```
summary(train)
```

```
      ID              사고일시            요일          기상상태
 Length:37128       Length:37128       금요일:5803   기타:   53
 Class :character   Class :character   목요일:5401   눈  :    6
 Mode  :character   Mode  :character   수요일:5581   맑음:33954
                                       월요일:5544   비  : 2424
                                       일요일:3823   안개:    8
                                       토요일:5323   흐림:  683
                                       화요일:5653
     시군구                          도로형태          노면상태
 Length:37128       단일로 - 기타          :18012   건조     :34197
 Class :character   교차로 - 교차로안      : 9611   기타     :   42
 Mode  :character   교차로 - 교차로부근    : 5468   서리/결빙:   19
                    기타 - 기타            : 1771   적설     :    3
                    교차로 - 교차로횡단보도내: 1362   젖음/습기: 2865
                    단일로 - 지하차도(도로)내:  285   침수     :    2
                    (Other)                :  619
   사고유형                 사고유형세부분류                법규위반
 차대사람: 6058   측면충돌   :16861   안전운전불이행   :20021
 차대차  :31070   기타       : 9413   안전거리미확보   : 5206
 차량단독:    0   추돌       : 5796   신호위반         : 3694
                  횡단중     : 2186   교차로운행방법위반: 2721
                  정면충돌   :  828   기타             : 1147
                  후진중충돌:  610   보행자보호의무위반: 1020
                  (Other)    : 1434   (Other)          : 3319
 가해운전자차종   가해운전자성별   가해운전자연령   가해운전자상해정도
 승용   :25979   기타불명:    0   Min.   :10.00   경상     : 3854
 화물   : 3967   남      :27839   1st Qu.:35.00   기타불명: 1467
 이륜   : 3796   여      : 9289   Median :50.00   부상신고: 2226
 승합   : 1142                    Mean   :47.97   사망     :   54
 자전거 : 1101                    3rd Qu.:60.00   상해없음:28503
 건설기계:  422                   Max.   :98.00   중상     : 1024
 (Other) :  721
 피해운전자차종   피해운전자성별   피해운전자연령   피해운전자상해정도
 승용   :20049           :    0   Min.   :10.00           :    0
 보행자 : 6058   기타불명:    0   1st Qu.:31.00   경상     :24279
 이륜   : 5082   남      :26571   Median :45.00   기타불명:  257
 화물   : 2020   여      :10557   Mean   :45.16   부상신고: 1165
 자전거 : 1945                    3rd Qu.:58.00   사망     :  174
 승합   :  986                    Max.   :98.00   상해없음: 4182
 (Other):  988                                    중상     : 7071
   사망자수          중상자수          경상자수          부상자수
 Min.   :0.00000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.0000
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.: 0.0000
 Median :0.00000   Median :0.0000   Median : 1.000   Median : 0.0000
 Mean   :0.00703   Mean   :0.2645   Mean   : 1.102   Mean   : 0.1167
 3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 1.000   3rd Qu.: 0.0000
 Max.   :2.00000   Max.   :6.0000   Max.   :22.000   Max.   :10.0000

      ECLO
 Min.   : 1.000
 1st Qu.: 3.000
 Median : 3.000
 Mean   : 4.816
 3rd Qu.: 6.000
 Max.   :74.000
```

데이터취득과정제/가공

2-7) 학습에 사용할 열 추출

```
#사용할 피처 선택해서 train 데이터 다시 구성
train <- train %>% select(도로형태, 노면상태, 요일, 사고유형, 사고유형세부분류, 법규위반, 가해운전자차종, 가해운전자연령,
가해운전자상해정도, 피해운전자차종, 피해운전자연령, 피해운전자상해정도, 사망자수, 중상자수, 경상자수, 부상자수)
train %>% show()
```

```
# A tibble: 37,128 × 16
   도로형태      노면상태 요일  사고유형 사고유형세부분류 법규위반 가해운전자차종
   <fct>         <fct>    <fct> <fct>    <fct>          <fct>    <fct>
 1 단일로 - 기… 건조     화요… 차대사람 길가장자리구역… 안전운… 승용
 2 단일로 - 기… 건조     화요… 차대사람 보도통행중      기타     승용
 3 단일로 - 기… 건조     화요… 차대사람 차도통행중      안전운… 승용
```

- 학습에 사용할 열을 select()와 열 이름을 이용해 추출했다.

2-8) 열 추가

```
train$사망중상자여부 <- ifelse(train$사망중상자여부 > 0, 1, 0) # 사망중상자가 있으면 1, 없으면 0
summary(train$사망중상자여부)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2464  0.0000  1.0000
```

```
train$경상부상자여부 <- ifelse(train$경상부상자여부 > 0, 1, 0) # 사망중상자가 있으면 1, 없으면 0
summary(train$경상부상자여부)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  1.0000  1.0000  0.8178  1.0000  1.0000
```

```
train <- train %>% mutate(사망중상자여부 = 사망자수 + 중상자수)
train <- train %>% mutate(경상부상자여부 = 경상자수 + 부상자수)
# 사망자수, 중상자수, 경상자수, 부상자수 열 없애기
train <- train[, -(13:16)]
head(train)
```

- 여러 독립변수들과 사망자 및 중상자 여부, 경상자 및 부상자 여부의 관계를 알아보기 위해 mutate()를 이용해 두 열을 추가했다.
- 사망자 및 중상자 여부는 '사망자중상자여부'열에 사망자수+중상자수가 0 이상이면 1, 아니면 0이 되도록 만들었다. 경상자 및 부상자 여부는 경상자수+부상자수가 0 이상이면 1, 아니면 0이 되도록 만들었다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 취득과 정제/가공

2-9) 숫자형 변수 범주화

```
summary(train$가해운전자연령)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   35.00   50.00   47.97   60.00   98.00
```

```
train$가해운전자연령 <- ifelse(train$가해운전자연령 >= 20 &train$가해운전자연령 < 30, 20,
                        ifelse(train$가해운전자연령 >= 30 &train$가해운전자연령 < 40, 30,
                        ifelse(train$가해운전자연령 >= 40 &train$가해운전자연령 < 50, 40,
                        ifelse(train$가해운전자연령 >= 50 &train$가해운전자연령 < 60, 50,
                        ifelse(train$가해운전자연령 >= 60 &train$가해운전자연령 < 70, 60,
                        70)))))
summary(train$가해운전자연령)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   30.00   50.00   45.17   60.00   70.00
```

```
train$피해운전자연령 <- ifelse(train$피해운전자연령 >= 20 &train$피해운전자연령 < 30, 20,
                        ifelse(train$피해운전자연령 >= 30 &train$피해운전자연령 < 40, 30,
                        ifelse(train$피해운전자연령 >= 40 &train$피해운전자연령 < 50, 40,
                        ifelse(train$피해운전자연령 >= 50 &train$피해운전자연령 < 60, 50,
                        ifelse(train$피해운전자연령 >= 60 &train$피해운전자연령 < 70, 60,
                        70)))))
summary(train$피해운전자연령)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   30.00   40.00   43.17   60.00   70.00
```

- 범주화를 하지 않으면 히스토그램을 그리거나 연산 시에 어려움이 있어 숫자 데이터로 구성된 가해운전자 연령, 피해운전자 연령을 범주화했다.
- 20대, 30대, 40대, 50대, 60대와 그 이상의 나이는 모두 70대로 범주화했다.
- 70대 이상은 모두 70대로 범주화하여 80대 이상의 아주 적은 데이터가 통계적 의미를 가질 수 있도록 했다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터취득과정제/가공

2-10) 독립변수 간 통계 확인

```
train %>%
  group_by(피해운전자연령) %>%
    summarise(
      mean_사망중상자여부 = mean(사망중상자여부),
      sum_사망중상자여부 = sum(사망중상자여부),
      median_사망중상자여부 = median(사망중상자여부),
      n=n()) %>% show
#70대 이상 피해운전자의 평균 사망중상자수 평균이 다른 나이대에 비해 높음
```

```
# A tibble: 6 × 5
  피해운전자연령 mean_사망중상자여부 sum_사망중상자여부 median_사망중상자여부
           <dbl>               <dbl>               <dbl>                 <dbl>
1             20               0.184                1192                     0
2             30               0.180                1189                     0
3             40               0.216                1466                     0
4             50               0.263                1987                     0
5             60               0.307                1660                     0
6             70               0.383                1654                     0
# ℹ 1 more variable: n <int>
```

```
train %>%
  group_by(피해운전자연령) %>%
    summarise(
      mean_경상부상자여부 = mean(경상부상자여부),
      sum_경상부상자여부 = sum(경상부상자여부),
      median_경상부상자여부 = median(경상부상자여부),
      n=n()) %>% show
#70대 이상 피해운전자의 평균 경상부상자수 평균이 다른 나이대에 비해 높음
```

```
# A tibble: 6 × 5
  피해운전자연령 mean_경상부상자여부 sum_경상부상자여부 median_경상부상자여부
           <dbl>               <dbl>               <dbl>                 <dbl>
1             20               0.866                5602                     1
2             30               0.876                5781                     1
3             40               0.855                5801                     1
4             50               0.813                6135                     1
5             60               0.773                4180                     1
6             70               0.664                2865                     1
# ℹ 1 more variable: n <int>
```

- 관계를 확인하려는 변수끼리 group_by()후 summarise()해 통계 결과를 확인했다.
- 피해 운전자 연령과 사망 중상자 여부를 비교한 결과, 70대의 사망 중상자 여부 평균이 다른 나이대에 비해 높았다.
- 피해 운전자 연령과 경상 부상자 여부를 비교한 결과, 70대의 경상 부상자 여부 평균이 다른 나이대에 비해 높았다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

3

데이터 시각화

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

3-1) 데이터 시각화 과정에서 무엇을, 어떤 방법으로 할것인가?

- 피해 운전자 연령과 사망중상자여부 관계
- 피해 운전자 연령과 경상부상자여부 관계
- 어떤 피처와 사망중상자여부, 경상부상자여부와 관계가 높은가?

-> 연령대와 수치 데이터(연령)의 관계를 알아보기 위해 **barplot()**선택

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

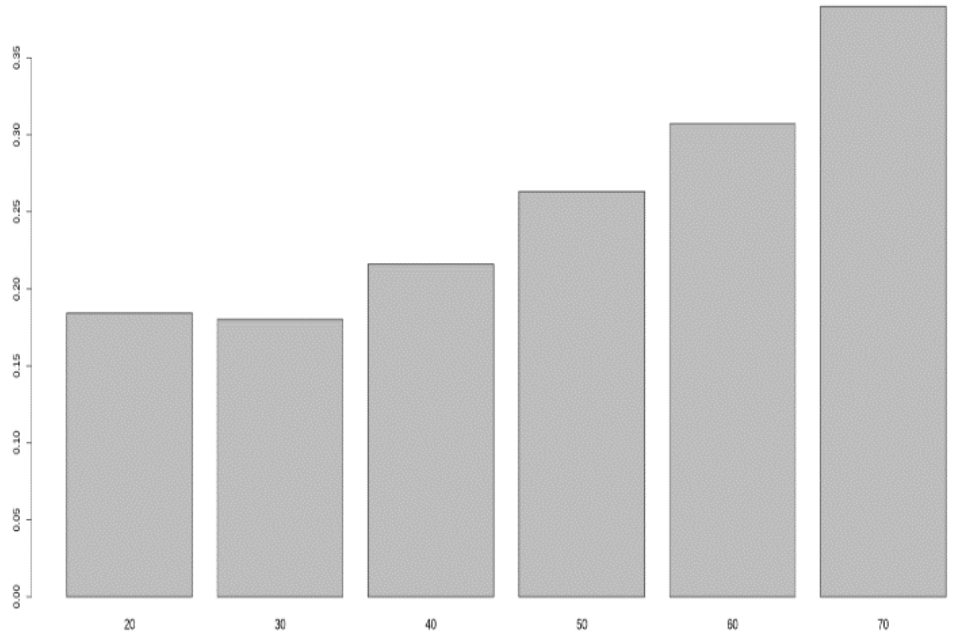
3-2) 피해 운전자 연령과 사망 중상자 여부 관계

```
#4. 피해 운전자 연령과 사망중상자수 관계
victim_age <- train %>%
  group_by(피해운전자연령) %>%
    summarise(mean_사망중상자여부 = mean(사망중상자여부))
victim_age
```

A tibble: 6 × 2

피해운전자연령	mean_사망중상자여부
<dbl>	<dbl>
20	0.1842634
30	0.1801242
40	0.2160012
50	0.2632137
60	0.3070094
70	0.3833140

```
barplot(victim_age$mean_사망중상자여부, names.arg = victim_age$피해운전자연령, cex.names=1.2)
# 피해운전자연령과 사망중상자수 관계 있어 보임
```



- 연령대가 높아질수록 사망 중상자 여부의 평균이 높아지는 것을 확인할 수 있다.
- 70대의 사망 중상자 여부 평균이 가장 높지만, 0.5를 넘지 않는 낮은 수치를 나타낸다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

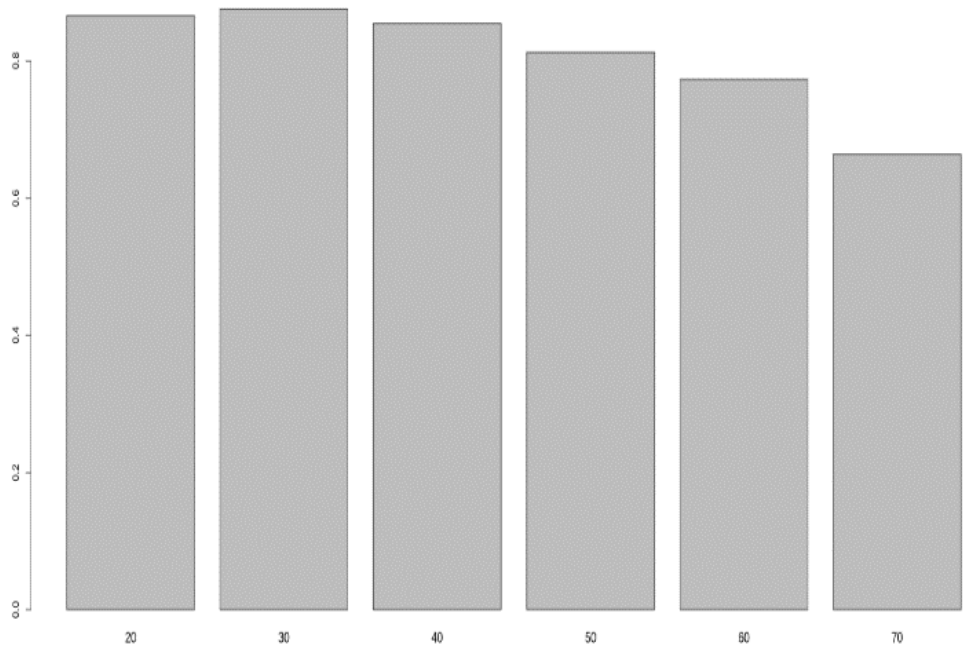
3-3) 피해 운전자 연령과 경상 부상자 여부 관계

```
# 피해 운전자 연령과 경상부상자수 관계
victim_age <- train %>%
  group_by(피해운전자연령) %>%
    summarise(mean_경상부상자여부 = mean(경상부상자여부))
victim_age
```

A tibble: 6 × 2

피해운전자연령	mean_경상부상자여부
<dbl>	<dbl>
20	0.8659762
30	0.8757764
40	0.8547223
50	0.8126904
60	0.7730719
70	0.6639629

```
barplot(victim_age$mean_경상부상자여부, names.arg = victim_age$피해운전자연령, cex.names=1.2)
```



- 연령대가 높아질수록 경상 부상자 여부의 평균이 낮아지는 것을 확인할 수 있다.
- 70대의 사망 중상자 여부 평균이 가장 낮지만, 0.6를 넘는 비교적 높은 수치를 나타낸다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

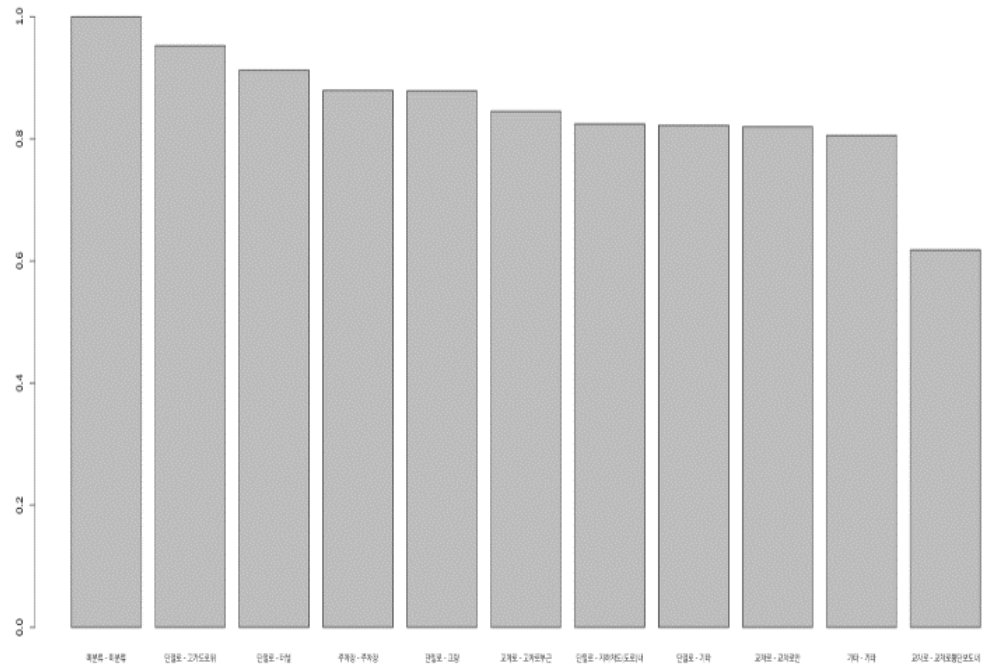
3-4) 도로 형태와 경상 부상자 여부 관계

```
# 도로 형태와 경상부상자여부 관계
road_type <- train %>%
  group_by(도로형태) %>%
    summarise(mean_경상부상자여부 = mean(경상부상자여부))
road_type <- road_type %>% arrange(desc(mean_경상부상자여부))
road_type
```

A tibble: 11 × 2

도로형태	mean_경상부상자여부
<fct>	<dbl>
미분류 - 미분류	1.0000000
단일로 - 고가도로위	0.9527559
단일로 - 터널	0.9122807
주차장 - 주차장	0.8790698
단일로 - 교량	0.8785047
교차로 - 교차로부근	0.8452816
단일로 - 지하차도(도로)내	0.8245614
단일로 - 기타	0.8220076
교차로 - 교차로안	0.8194777
기타 - 기타	0.8057595
교차로 - 교차로횡단보도내	0.6182085

```
barplot(road_type$mean_경상부상자여부, names.arg = road_type$도로형태, cex.names=0.7) # y축, x축, 폰트사
```



- 가장 높은 경상 부상자 여부 평균을 가지는 도로 형태를 알아보기 위해, arrange()를 이용해 높은 평균을 가지는 순서대로 내림차순 정렬해 시각화했다.
- 가장 높은 평균을 가지는 도로 형태는 미분류 다음으로 '단일로 – 고가도로위'이다.
- 가장 낮은 평균을 가지는 도로 형태는 '교차로 – 교차로횡단보도내'이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

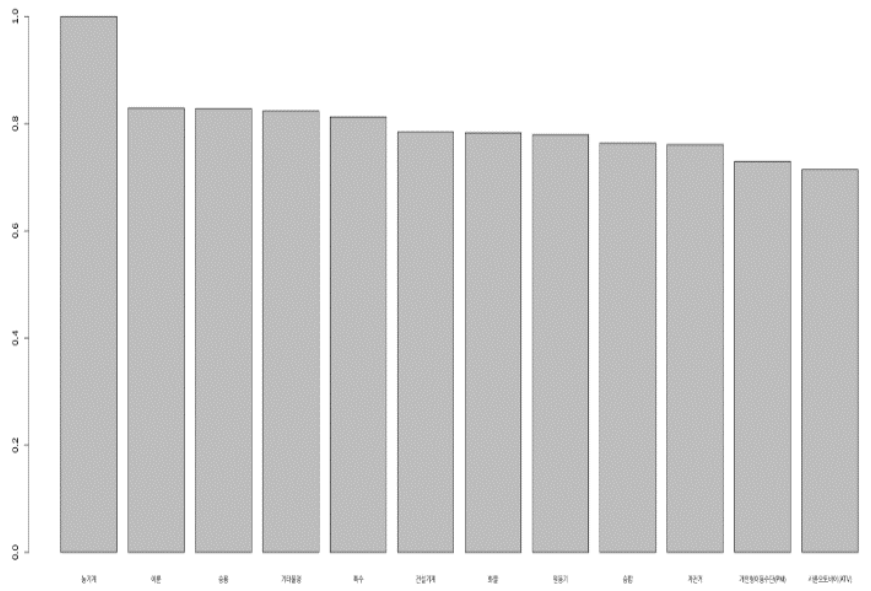
3-5) 가해 운전자 차종과 경상 부상자 여부 관계

```
# 가해운전자차종과 경상부상자여부 관
attacker_car <- train %>%
  group_by(가해운전자차종) %>%
    summarise(mean_경상부상자여부 = mean(경상부상자여부))
attacker_car <- attacker_car %>% arrange(desc(mean_경상부상자여부))
attacker_car
```

A tibble: 12 × 2

가해운전자차종	mean_경상부상자여부
<fct>	<dbl>
농기계	1.0000000
이륜	0.8292940
승용	0.8278995
기타불명	0.8235294
특수	0.8125000
건설기계	0.7843602
화물	0.7832115
원동기	0.7793765
승합	0.7635727
자전거	0.7611262
개인형이동수단(PM)	0.7297297
사륜오토바이(ATV)	0.7142857

```
barplot(attacker_car$mean_경상부상자여부, names.arg = attacker_car$가해운전자차종, cex.names=0.7) # y축
```



- 가장 높은 경상 부상자 여부 평균을 가지는 가해 운전자의 차종은 '농기계'이다.
- 가장 낮은 경상 부상자 여부 평균을 가지는 가해 운전자의 차종은 '사륜오토바이'이다.
- 가장 높은 경상 부상자 여부 평균을 가지는 가해 운전자의 차종과 두 번째로 높은 경상 부상자 여부 평균을 가지는 가해 운전자의 차종의 차이가 비교적 크고, 나머지 차종의 차이는 거의 없음을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

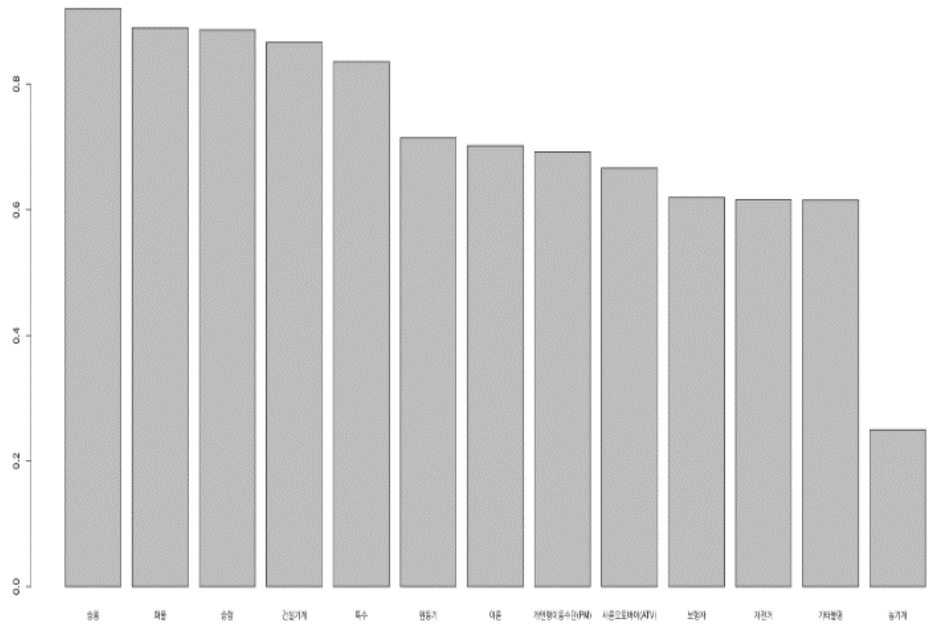
3-6) 피해 운전자 차종과 경상 부상자 여부 관계

```
#  피해운전자차종
victim_car <- train %>%
  group_by(피해운전자차종) %>%
    summarise(mean_경상부상자여부 = mean(경상부상자여부))
victim_car <- victim_car %>% arrange(desc(mean_경상부상자여부))
victim_car
```

A tibble: 13 × 2

피해운전자차종	mean_경상부상자여부
<fct>	<dbl>
승용	0.9203452
화물	0.8896040
승합	0.8864097
건설기계	0.8666667
특수	0.8356164
원동기	0.7145359
이륜	0.7020858
개인형이동수단(PM)	0.6923077
사륜오토바이(ATV)	0.6666667
보행자	0.6196765
자전거	0.6159383
기타불명	0.6153846
농기계	0.2500000

```
barplot(victim_car$mean_경상부상자여부, names.arg = victim_car$피해운전자차종, cex.names=0.8) # y축, x축
```



- 가장 높은 경상 부상자 여부 평균을 가지는 피해 운전자의 차종은 '승용'이다.
- 가장 낮은 경상 부상자 여부 평균을 가지는 피해 운전자의 차종은 '농기계'이다.
- 평균 경상 부상자 여부가 0.6을 넘는 나머지 차종과 달리, '농기계'의 평균 경상 부상자 여부는 0.25로 매우 낮은 수치를 가진다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

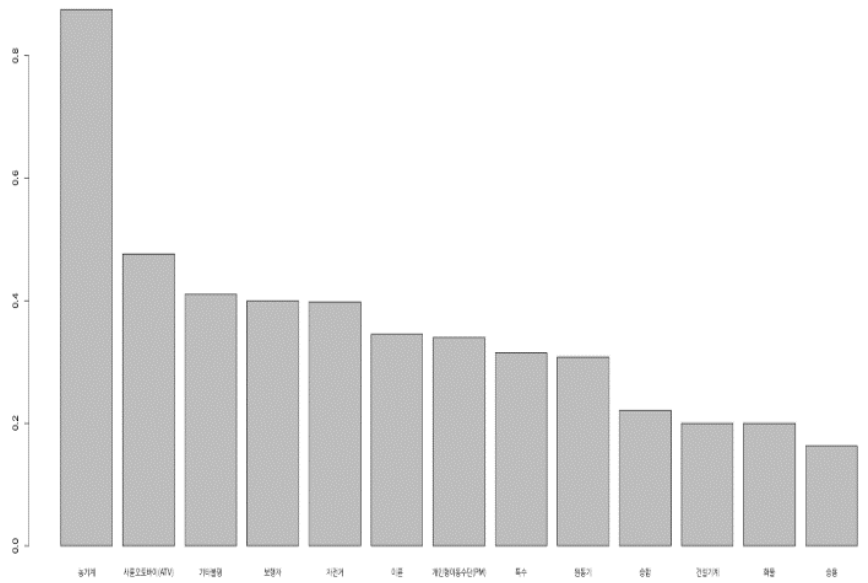
3-7) 피해 운전자 차종과 사망 중상자 여부 관계

```
# 피해운전자차종과 사망 중상자 여부 관계
victim_car <- train %>%
  group_by(피해운전자차종) %>%
    summarise(mean_사망중상자여부 = mean(사망중상자여부))
victim_car <- victim_car %>% arrange(desc(mean_사망중상자여부))
victim_car
```

A tibble: 13 × 2

피해운전자차종	mean_사망중상자여부
<fct>	<dbl>
농기계	0.8750000
사륜오토바이(ATV)	0.4761905
기타불명	0.4102564
보행자	0.3996368
자전거	0.3979434
이륜	0.3453365
개인형이동수단(PM)	0.3397436
특수	0.3150685
원동기	0.3082312
승합	0.2210953
건설기계	0.2000000
화물	0.2000000
승용	0.1629508

```
barplot(victim_car$mean_사망중상자여부, names.arg = victim_car$피해운전자차종, cex.names=0.8)
```



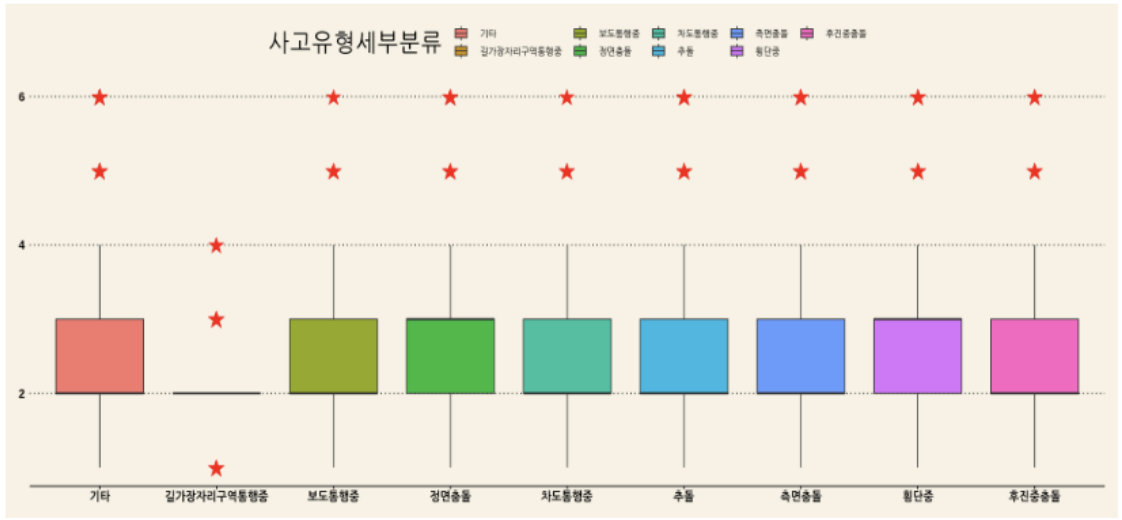
- 가장 높은 사망 중상자 여부 평균을 가지는 피해 운전자의 차종은 '농기계'이다.
- '농기계'는 다른 차종에 비해 매우 높은 가장 높은 사망 중상자 여부 평균 수치를 가진다.
- 가장 낮은 사망 중상자 여부 평균을 가지는 피해 운전자의 차종은 '승용'이다.
- 이전 '5-5) 피해 운전자 차종과 경상 부상자 여부 관계'에서 가장 높은 수치를 가지는 차종이 '승용', 가장 낮은 수치를 가지는 차종이 '농기계'였던 것과 반대되는 결과임을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

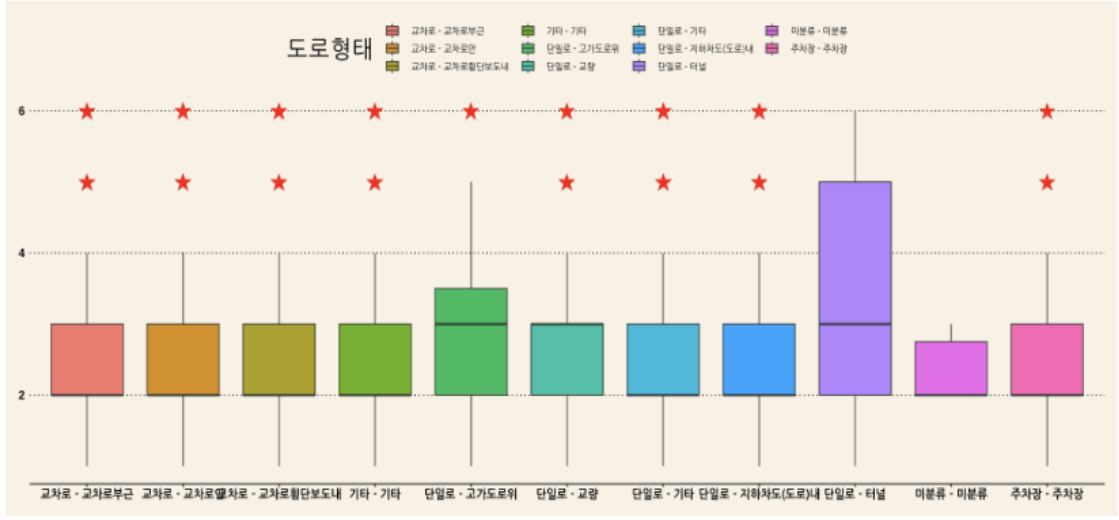
데이터 시각화

3-8) geom_boxplot()

```
ggplot(data=train, mapping=aes(x=사고유형세부분류, y=ECLO)) +
geom_boxplot(mapping=aes(fill=사고유형세부분류), outlier.color="red", outlier.shape="★", outlier.size=7) +
#geom_text(mapping=aes(label=is_out), na.rm=T, vjust=-1) +
theme(axis.title.x=element_text(family = fonts()[1])) + # 한글 설정.
theme_wsj()
```




```
ggplot(data=train, mapping=aes(x=도로형태, y=ECLO)) +
geom_boxplot(mapping=aes(fill=도로형태), outlier.color="red", outlier.shape="★", outlier.size=7) +
#geom_text(mapping=aes(label=is_out), na.rm=T, vjust=-1) +
theme(axis.title.x=element_text(family = fonts()[1])) + # 한글 설정.
theme_wsj()
```



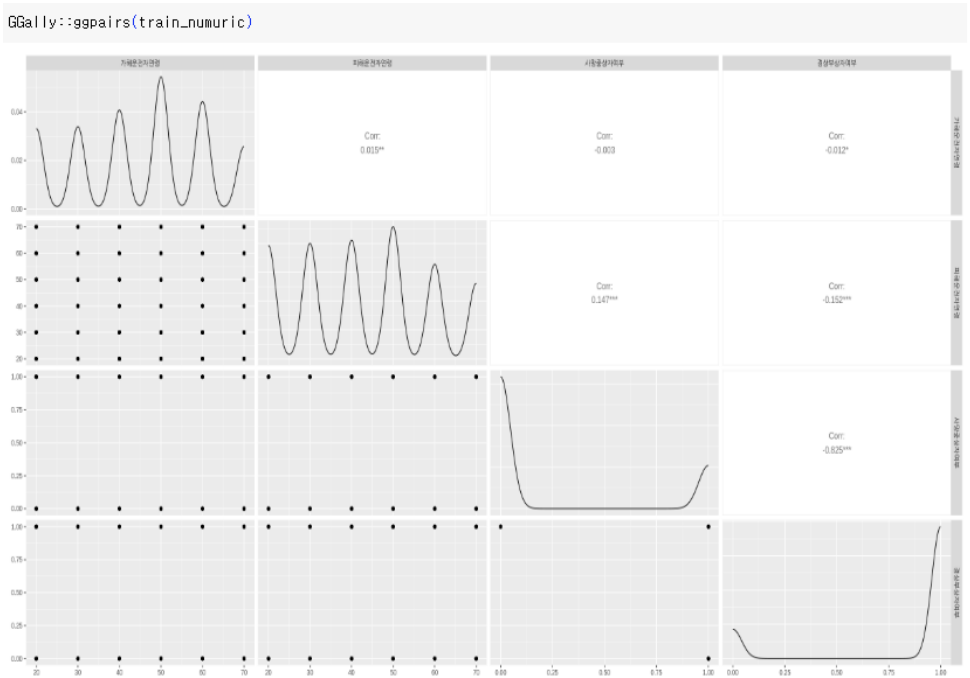
- geom_boxplot()을 이용해 '사고 유형 세부 분류' 열과 '도로 형태' 열이 사고 피해 정도를 의미하는 ECLO 수치와 가지는 관계를 확인했다.

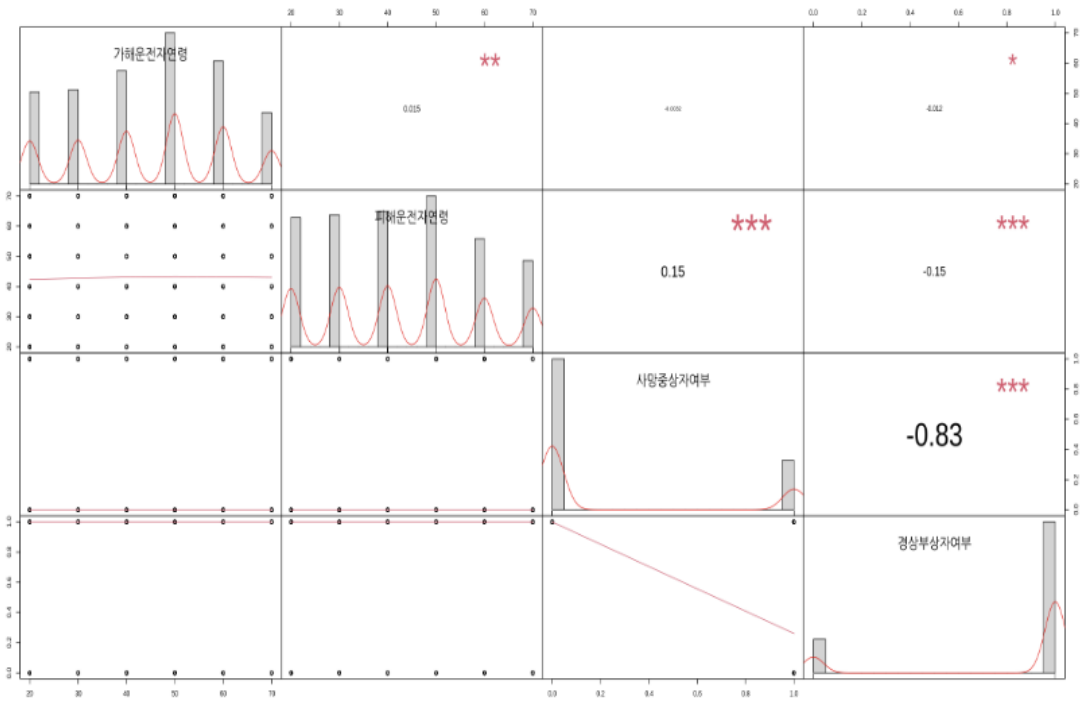
ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

3-9) 상관분석

```
GGally::ggpairs(train_numuric)
```





- 숫자 데이터를 가지는 가해 운전자 연령, 피해 운전자 연령, 사망 중상자 여부, 경상 부상자 여부에 대해 cor(), corrplot(), GGally의 ggpairs(), PerformanceAnalytics의 chart.Correlation()를 사용해 상관관계를 확인했다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

데이터 시각화

3-9) 상관분석

```
train_numuric <- train %>% select(가해운전자연령, 피해운전자연령,사망중상자여부, 경상부상자여부)
head(train_numuric)
```

A tibble: 6 × 4

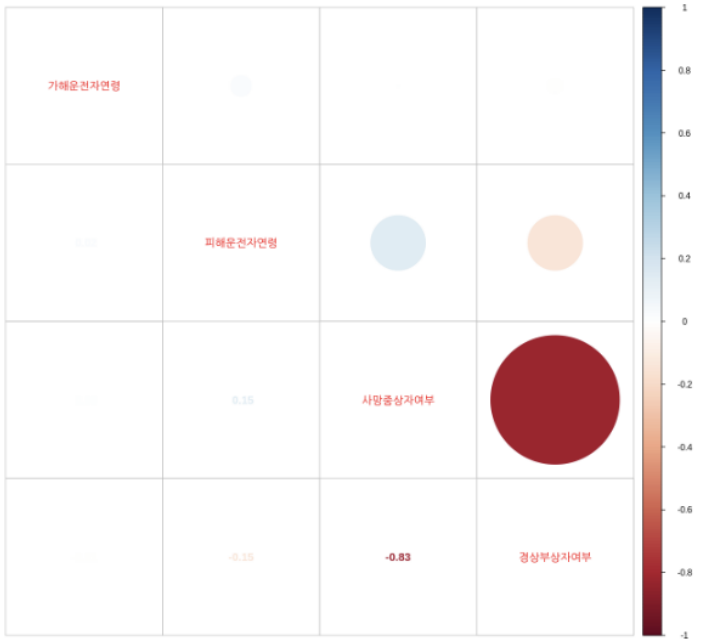
가해운전자연령	피해운전자연령	사망중상자여부	경상부상자여부
<dbl>	<dbl>	<dbl>	<dbl>
50	70	1	0
30	60	0	1
70	30	0	1
40	30	1	0
30	50	0	1
50	30	0	1

```
train_numuric_cor <- cor(train_numuric) %>% round(2) #선형회귀 관련
train_numuric_cor
```

A matrix: 4 × 4 of type dbl

	가해운전자연령	피해운전자연령	사망중상자여부	경상부상자여부
가해운전자연령	1.00	0.02	0.00	-0.01
피해운전자연령	0.02	1.00	0.15	-0.15
사망중상자여부	0.00	0.15	1.00	-0.83
경상부상자여부	-0.01	-0.15	-0.83	1.00

```
corrplot.mixed(train_numuric_cor)
```



- 다음은 가해 운전자 연령, 피해 운전자 연령, 사망 중상자 여부, 경상 부상자 여부에 대해 그래프를 바탕으로 분석한 결과이다.
- 피해 운전자 연령과 사망 중상자 여부는 양의 상관 관계가 있지만 높지 않다.
- 피해 운전자 연령과 경상 부상자 여부는 양의 상관 관계가 있지만 높지 않다.
- 사망 중상자 여부와 경상 부상자 여부는 매우 높은 음의 상관관계를 가진다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4

가설 설정 및 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 1. 귀무가설 : 사망 중상자 여부와 피해 운전자 연령은 연관이 있다.
가설 1. 대립가설 : 사망 중상자 여부와 피해 운전자 연령은 연관이 없다.

가설 2. 귀무가설 : 경상 부상자 여부와 피해 운전자 연령은 연관이 있다.
가설 2. 대립가설 : 경상 부상자 여부와 피해 운전자 연령은 연관이 없다.

유의수준 : 0.05 / 검정 방법 : 카이제곱 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 3. 귀무가설 : 사망 중상자 여부와 가해 운전자 연령은 연관이 있다.
가설 3. 대립가설 : 사망 중상자 여부와 가해 운전자 연령은 연관이 없다.

가설 4. 귀무가설 : 경상 부상자 여부와 가해 운전자 연령은 연관이 있다.
가설 4. 대립가설 : 경상 부상자 여부와 가해 운전자 연령은 연관이 없다.

유의수준 : 0.05 / 검정 방법 : 카이제곱 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 5. 귀무가설 : 피해 운전자 차종과 사망 중상자 여부는 연관이 있다.
가설 5. 대립가설 : 피해 운전자 차종과 사망 중상자 여부는 연관이 없다.

가설 6. 귀무가설 : 피해 운전자 차종과 경상 부상자 여부는 연관이 있다.
가설 6. 대립가설 : 피해 운전자 차종과 경상 부상자 여부는 연관이 없다.

유의수준 : 0.05 / 검정 방법 : 카이제곱 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 7. 귀무가설 : 가해 운전자 차종과 경상 부상자 여부는 연관이 있다.
가설 7. 대립가설 : 가해 운전자 차종과 경상 부상자 여부는 연관이 없다.

가설 8. 귀무가설 : 가해 운전자 차종과 사망 부상자 여부는 연관이 있다.
가설 8. 대립가설 : 가해 운전자 차종과 사망 부상자 여부는 연관이 없다.

유의수준 : 0.05 / 검정 방법 : 카이제곱 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 9. 귀무가설 : 도로 형태와 경상 부상자 여부는 연관이 있다.
가설 9. 대립가설 : 도로 형태와 경상 부상자 여부는 연관이 없다.

가설 10. 귀무가설 : 도로 형태와 사망 중상자 여부는 연관이 있다.
가설 10. 대립가설 : 도로 형태와 사망 중상자 여부는 연관이 없다.

유의수준 : 0.05 / 검정 방법 : 카이제곱 검정

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 1. 귀무가설 : 사망 중상자 여부와 피해 운전자 연령은 연관이 있다.
가설 1. 대립가설 : 사망 중상자 여부와 피해 운전자 연령은 연관이 없다.

```
print("피해 운전자 차종과 사망 중상자여부 연관 관계")
gmodels::CrossTable(train$사망중상자수, train$피해운전자차종, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "피해 운전자 차종과 사망 중상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$피해운전자차종
train$사망중상자수 | 개인형이동수단(PM) |          건설기계 |          기타불명 |              농기계 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
           0 |         103 |          96 |          23 |           1 |        3637 |          11 |       16782 |
             |     117.563 |      90.433 |      29.391 |       6.029 |    4565.364 |      15.826 |    15109.110 |
             |       1.804 |       0.343 |       1.390 |       4.195 |     188.782 |       1.472 |     185.224 |
             |       0.003 |       0.003 |       0.001 |       0.000 |       0.098 |       0.000 |       0.452 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
           1 |          53 |          24 |          16 |           7 |        2421 |          10 |        3267 |
             |      38.437 |      29.567 |       9.609 |       1.971 |    1492.636 |       5.174 |     4939.890 |
             |       5.518 |       1.048 |       4.250 |      12.830 |     577.408 |       4.501 |     566.523 |
             |       0.001 |       0.001 |       0.000 |       0.000 |       0.065 |       0.000 |       0.088 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
Column Total |         156 |         120 |          39 |           8 |        6058 |          21 |       20049 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|---


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2104.252     d.f. =  12     p =  0
```

자유도 d.f. : 12
P-value : 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 사망 중상자 여부와 피해 운전자 연령은 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 2. 귀무가설 : 경상 부상자 여부와 피해 운전자 연령은 연관이 있다.
가설 2. 대립가설 : 경상 부상자 여부와 피해 운전자 연령은 연관이 없다.

```
print("경상부상자여부와 피해운전자연령의 연관 관계")
gmodels::CrossTable(train$경상부상자수, train$피해운전자연령, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "경상부상자여부와 피해운전자연령의 연관 관계"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$피해운전자연령
train$경상부상자수 |        20 |        30 |        40 |        50 |        60 |        70 | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
           0 |       867 |       820 |       986 |      1414 |      1227 |      1450 |      6764 |
             |  1178.526 |  1202.574 |  1236.459 |  1375.281 |   985.050 |   786.109 |           |
             |    82.347 |   121.708 |    50.734 |     1.090 |    59.428 |   560.674 |           |
             |     0.023 |     0.022 |     0.027 |     0.038 |     0.033 |     0.039 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
           1 |      5602 |      5781 |      5801 |      6135 |      4180 |      2865 |     30364 |
             |  5290.474 |  5398.426 |  5550.541 |  6173.719 |  4421.950 |  3528.891 |           |
             |    18.344 |    27.112 |    11.302 |     0.243 |    13.238 |   124.898 |           |
             |     0.151 |     0.156 |     0.156 |     0.165 |     0.113 |     0.077 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |      6469 |      6601 |      6787 |      7549 |      5407 |      4315 |     37128 |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  1071.118     d.f. =  5     p =  2.401422e-229
```

자유도 d.f. : 5
P-value ≈ 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 경상 부상자 여부와 피해 운전자 연령은 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 3. 귀무가설 : 사망 중상자 여부와 가해 운전자 연령은 연관이 있다.
가설 3. 대립가설 : 사망 중상자 여부와 가해 운전자 연령은 연관이 없다.

```
print("사망중상자여부와 가해운전자연령 연관 관계")
gmodels::CrossTable(train$사망중상자수, train$가해운전자연령, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "사망중상자여부와 가해운전자연령 연관 관계"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$가해운전자연령
train$사망중상자수 |        20 |        30 |        40 |        50 |        60 |        70 | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
           0 |      4001 |      4082 |      4889 |      6518 |      5360 |      3130 |     27980 |
             |  3989.607 |  4085.315 |  4912.778 |  6556.399 |  5335.553 |  3100.348 |           |
             |     0.033 |     0.003 |     0.115 |     0.225 |     0.112 |     0.284 |           |
             |     0.108 |     0.110 |     0.132 |     0.176 |     0.144 |     0.084 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
           1 |      1293 |      1339 |      1630 |      2182 |      1720 |       984 |      9148 |
             |  1304.393 |  1335.685 |  1606.222 |  2143.601 |  1744.447 |  1013.652 |           |
             |     0.100 |     0.008 |     0.352 |     0.688 |     0.343 |     0.867 |           |
             |     0.035 |     0.036 |     0.044 |     0.059 |     0.046 |     0.027 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |      5294 |      5421 |      6519 |      8700 |      7080 |      4114 |     37128 |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3.128449     d.f. =  5     p =  0.6801904
```

자유도 d.f. : 5
P-value ≈ 0.68

P-value 가 유의수준 0.05보다 크므로 대립가설 채택.

즉, 사망 중상자 여부와 가해 운전자 연령은 연관이 없다

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 4. 귀무가설 : 경상 부상자 여부와 가해 운전자 연령은 연관이 있다.
가설 4. 대립가설 : 경상 부상자 여부와 가해 운전자 연령은 연관이 없다.

```
print("경상부상자여부와 가해운전자연령 연관 관계")
gmodels::CrossTable(train$경상부상자수, train$가해운전자연령, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "경상부상자여부와 가해운전자연령 연관 관계"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$가해운전자연령
train$경상부상자수 |       20 |       30 |       40 |       50 |       60 |       70 | Row Total |
-------------|----------|----------|----------|----------|----------|----------|----------|
           0 |      875 |      955 |     1244 |     1648 |     1307 |      735 |     6764 |
             |  964.464 |  987.601 | 1187.635 | 1584.971 | 1289.838 |  749.491 |          |
             |    8.299 |    1.076 |    2.675 |    2.506 |    0.228 |    0.280 |          |
             |    0.024 |    0.026 |    0.034 |    0.044 |    0.035 |    0.020 |          |
-------------|----------|----------|----------|----------|----------|----------|----------|
           1 |     4419 |     4466 |     5275 |     7052 |     5773 |     3379 |    30364 |
             | 4329.536 | 4433.399 | 5331.365 | 7115.029 | 5790.162 | 3364.509 |          |
             |    1.849 |    0.240 |    0.596 |    0.558 |    0.051 |    0.062 |          |
             |    0.119 |    0.120 |    0.142 |    0.190 |    0.155 |    0.091 |          |
-------------|----------|----------|----------|----------|----------|----------|----------|
Column Total |     5294 |     5421 |     6519 |     8700 |     7080 |     4114 |    37128 |
-------------|----------|----------|----------|----------|----------|----------|----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  18.4208     d.f. =  5     p =  0.002462736
```

자유도 d.f. : 5
P-value ≈ 0.002

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 경상 부상자 여부와 가해 운전자 연령은 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 5. 귀무가설 : 피해 운전자 차종과 사망 중상자 여부는 연관이 있다.
가설 5. 대립가설 : 피해 운전자 차종과 사망 중상자 여부는 연관이 없다.

```
print("피해 운전자 차종과 사망 중상자여부 연관 관계")
gmodels::CrossTable(train$사망중상자수, train$피해운전자차종, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "피해 운전자 차종과 사망 중상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$피해운전자차종
train$사망중상자수 | 개인형이동수단(PM) |          건설기계 |          기타불명 |             농기계 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
           0 |         103 |          96 |          23 |           1 |        3637 |          11 |       16782 |
             |     117.563 |      90.433 |      29.391 |       6.029 |    4565.364 |      15.826 |    15109.110 |
             |       1.804 |       0.343 |       1.390 |       4.195 |     188.782 |       1.472 |      185.224 |
             |       0.003 |       0.003 |       0.001 |       0.000 |       0.098 |       0.000 |        0.452 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
           1 |          53 |          24 |          16 |           7 |        2421 |          10 |        3267 |
             |      38.437 |      29.567 |       9.609 |       1.971 |    1492.636 |       5.174 |     4939.890 |
             |       5.518 |       1.048 |       4.250 |      12.830 |     577.408 |       4.501 |      566.523 |
             |       0.001 |       0.001 |       0.000 |       0.000 |       0.065 |       0.000 |        0.088 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|---
Column Total |         156 |         120 |          39 |           8 |        6058 |          21 |       20049 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|---


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2104.252     d.f. =  12     p =  0
```

자유도 d.f. : 12
P-value : 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 피해 운전자 차종과 사망 중상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 6. 귀무가설 : 피해 운전자 차종과 경상 부상자 여부는 연관이 있다.
가설 6. 대립가설 : 피해 운전자 차종과 경상 부상자 여부는 연관이 없다.

```
print("피해 운전자 차종과 경상 부상자여부 연관 관계")
gmodels::CrossTable(train$경상부상자수, train$피해운전자차종, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "피해 운전자 차종과 경상 부상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$피해운전자차종
train$경상부상자수 | 개인형이동수단(PM) |          건설기계 |          기타불명 |            농기계 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-
           0 |          48 |          16 |          15 |           6 |        2304 |           7 |        1597 |
             |      28.420 |      21.862 |       7.105 |       1.457 |    1103.650 |       3.826 |    3652.538 |
             |      13.489 |       1.572 |       8.773 |      14.158 |    1305.523 |       2.634 |    1156.795 |
             |       0.001 |       0.000 |       0.000 |       0.000 |       0.062 |       0.000 |       0.043 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-
           1 |         108 |         104 |          24 |           2 |        3754 |          14 |       18452 |
             |     127.580 |      98.138 |      31.895 |       6.543 |    4954.350 |      17.174 |   16396.462 |
             |       3.005 |       0.350 |       1.954 |       3.154 |     290.823 |       0.587 |     257.692 |
             |       0.003 |       0.003 |       0.001 |       0.000 |       0.101 |       0.000 |       0.497 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-
Column Total |         156 |         120 |          39 |           8 |        6058 |          21 |       20049 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  4191.466     d.f. =  12     p =  0
```

자유도 d.f. : 12
P-value : 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 피해 운전자 차종과 경상 부상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 7. 귀무가설 : 가해 운전자 차종과 경상 부상자 여부는 연관이 있다.
가설 7. 대립가설 : 가해 운전자 차종과 경상 부상자 여부는 연관이 없다.

```
print("가해 운전자 차종과 경상 부상자여부 연관 관계")
gmodels::CrossTable(train$경상부상자수, train$가해운전자차종, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "가해 운전자 차종과 경상 부상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$가해운전자차종
train$경상부상자수 | 개인형이동수단(PM) |          건설기계 |          기타불명 |              농기계 |  사륜오토바이 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----
           0 |          40 |          91 |           6 |           0 |           2 |        4471 |         270 |
             |      26.963 |      76.880 |       6.194 |       0.547 |       1.275 |    4732.869 |      208.050 |
             |       6.304 |       2.593 |       0.006 |       0.547 |       0.412 |      14.489 |       18.446 |
             |       0.001 |       0.002 |       0.000 |       0.000 |       0.000 |       0.120 |        0.007 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----
           1 |         108 |         331 |          28 |           3 |           5 |       21508 |         872 |
             |     121.037 |     345.120 |      27.806 |       2.453 |       5.725 |    21246.131 |      933.950 |
             |       1.404 |       0.578 |       0.001 |       0.122 |       0.092 |       3.228 |        4.109 |
             |       0.003 |       0.009 |       0.001 |       0.000 |       0.000 |       0.579 |        0.023 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----
Column Total |         148 |         422 |          34 |           3 |           7 |       25979 |        1142 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  115.4844     d.f. =  11     p =  1.463245e-19
```

자유도 d.f. : 11
P-value $\approx$ 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 가해 운전자 차종과 경상 부상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 8. 귀무가설 : 가해 운전자 차종과 사망 부상자 여부는 연관이 있다.
가설 8. 대립가설 : 가해 운전자 차종과 사망 부상자 여부는 연관이 없다.

```
print("가해 운전자 차종과 사망중상자여부 연관 관계")
gmodels::CrossTable(train$사망중상자수, train$가해운전자차종, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "가해 운전자 차종과 사망중상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$가해운전자차종
train$사망중상자수 | 개인형이동수단(PM) |          건설기계 |          기타불명 |             농기계 |  사륜오토바
-------------|-------------|-------------|-------------|-------------|-------------|-------------|----
           0 |         105 |         300 |          27 |           3 |           4 |       19906 |         809 |
             |     111.534 |     318.023 |      25.623 |       2.261 |       5.275 |   19578.012 |     860.622 |
             |       0.383 |       1.021 |       0.074 |       0.242 |       0.308 |       5.495 |       3.096 |
             |       0.003 |       0.008 |       0.001 |       0.000 |       0.000 |       0.536 |       0.022 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|----
           1 |          43 |         122 |           7 |           0 |           3 |        6073 |         333 |
             |      36.466 |     103.977 |       8.377 |       0.739 |       1.725 |    6400.988 |     281.378 |
             |       1.171 |       3.124 |       0.226 |       0.739 |       0.943 |      16.806 |       9.470 |
             |       0.001 |       0.003 |       0.000 |       0.000 |       0.000 |       0.164 |       0.009 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|----
Column Total |         148 |         422 |          34 |           3 |           7 |       25979 |        1142 |
-------------|-------------|-------------|-------------|-------------|-------------|-------------|----


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  84.03169     d.f. =  11     p =  2.438602e-13
```

자유도 d.f. : 11
P-value $\approx$ 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 가해 운전자 차종과 사망 부상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 9. 귀무가설 : 도로 형태와 경상 부상자 여부는 연관이 있다.
가설 9. 대립가설 : 도로 형태와 경상 부상자 여부는 연관이 없다.

```
print("도로형태와 경상 부상자여부 연관 관계")
gmodels::CrossTable(train$경상부상자수, train$도로형태, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "도로형태와 경상 부상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$도로형태
train$경상부상자수 |       교차로 - 교차로부근 |        교차로 - 교차로안 | 교차로 - 교차로횡단보도내 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
           0 |             846 |            1735 |             520 |             344 |               6 |
             |         996.163 |        1750.937 |         248.130 |         322.642 |          23.137 |
             |          22.636 |           0.145 |         297.882 |           1.414 |          12.693 |
             |           0.023 |           0.047 |           0.014 |           0.009 |           0.000 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
           1 |            4622 |            7876 |             842 |            1427 |             121 |
             |        4471.837 |        7860.063 |        1113.870 |        1448.358 |         103.863 |
             |           5.042 |           0.032 |          66.357 |           0.315 |           2.828 |
             |           0.124 |           0.212 |           0.023 |           0.038 |           0.003 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
Column Total |            5468 |            9611 |            1362 |            1771 |             127 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  427.0048     d.f. =  10     p =  1.669817e-85
```

자유도 d.f. : 10
P-value ≈ 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 도로 형태와 경상 부상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

가설 설정 및 검정

가설 10. 귀무가설 : 도로 형태와 사망 중상자 여부는 연관이 있다.
가설 10. 대립가설 : 도로 형태와 사망 중상자 여부는 연관이 없다.

```
print("도로형태와 사망 중상자여부 연관 관계")
gmodels::CrossTable(train$사망중상자수, train$도로형태, chisq = T, expected = T, prop.r = F, prop.c = F)
```

```
[1] "도로형태와 사망 중상자여부 연관 관계"
Warning message in chisq.test(t, correct = FALSE, ...):
 "Chi-squared approximation may be incorrect"


   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  37128


             | train$도로형태
train$사망중상자수 |       교차로 - 교차로부근 |        교차로 - 교차로안 | 교차로 - 교차로횡단보도내 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
           0 |            4266 |            6995 |             797 |            1378 |             107 |
             |         4120.735 |         7242.937 |         1026.416 |         1334.642 |          95.708 |
             |           5.121 |           8.487 |          51.277 |           1.409 |           1.332 |
             |           0.115 |           0.188 |           0.021 |           0.037 |           0.003 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
           1 |            1202 |            2616 |             565 |             393 |              20 |
             |         1347.265 |         2368.063 |         335.584 |         436.358 |          31.292 |
             |          15.663 |          25.959 |         156.835 |           4.308 |           4.075 |
             |           0.032 |           0.070 |           0.015 |           0.011 |           0.001 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----
Column Total |            5468 |            9611 |            1362 |            1771 |             127 |
-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  315.3227     d.f. =  10     p =  8.91898e-62
```

자유도 d.f. : 10
P-value ≈ 0

P-value 가 유의수준 0.05보다 작으므로 귀무가설 채택.

즉, 도로 형태와 사망 중상자 여부는 연관이 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

5

모델링 및 분석

–로지스틱 회귀, 결정트리,Random Forest, SVM

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '사망중상자여부'인 경우

```
train1 <- subset(train, select=-경상부상자여부)
head(train1)
```

A tibble: 6 × 11

도로형태	노면상태	요일	사고유형	사고유형세부분류	법규위반	가해운전자차종	가해운전자연령	피해운전자차종	피해운전자연령	사망중상자여부
<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>
단일로 - 기타	건조	화요일	차대사람	길가장자리구역통행중	안전운전불이행	승용	50	보행자	70	1
단일로 - 기타	건조	화요일	차대사람	보도통행중	기타	승용	30	보행자	60	0
단일로 - 기타	건조	화요일	차대사람	차도통행중	안전운전불이행	승용	70	보행자	30	0
단일로 - 기타	건조	화요일	차대차	추돌	안전운전불이행	승용	40	승용	30	1
단일로 - 기타	건조	화요일	차대차	추돌	안전운전불이행	승용	30	승용	50	0

```
train_set <- sample_frac(train1, 0.7)
test_set <- setdiff(train1, train_set) #setdiff: 차집합

head(train_set)
```

A tibble: 6 × 11

도로형태	노면상태	요일	사고유형	사고유형세부분류	법규위반	가해운전자차종	가해운전자연령	피해운전자차종	피해운전자연령	사망중상자여부
<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>
교차로 - 교차로안	건조	목요일	차대차	측면충돌	안전운전불이행	화물	60	이륜	40	0
단일로 - 교량	건조	목요일	차대차	기타	안전거리미확보	승용	20	승용	50	0
교차로 - 교차로안	건조	화요일	차대차	측면충돌	안전운전불이행	화물	50	자전거	70	0
단일로 - 기타	건조	목요일	차대차	정면충돌	안전운전불이행	이륜	50	승용	40	0

0) 로지스틱 회귀 모델 선택 이유

- 종속변수가 여러 독립변수에 대해 선형적으로 변화하는 형태가 아닌, 두 종류로 분류되는 형태이기 때문

1) 데이터 분리

- 종속변수가 '경상 부상자 여부 ' 인 경우와 '사망 중상자 여부 ' 인 경우로 나누어 분석하기 위해 전체 데이터셋에서 '사망 중상자 여부 ' 열을 제외한 데이터셋 생성

- train 데이터와 test 데이터를 7:3 비율로 분리

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '사망중상자여부'인 경우

2) 로지스틱 회귀 모델 m 생성

```
m <- lm(formula = 사망중상자여부 ~ ., data = train_set)
summary(m)
```

```
Call:
lm(formula = 사망중상자여부 ~ ., data = train_set)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8620 -0.2570 -0.1534  0.0364  1.0333

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.941e-01  6.256e-02   9.496  < 2e-16 ***
도로형태교차로 - 교차로안        2.604e-02  9.083e-03   2.867 0.004150 **
도로형태교차로 - 교차로횡단보도내 5.715e-03  1.646e-02   0.347 0.728369
도로형태기타 - 기타              -4.655e-02  1.363e-02  -3.415 0.000639 ***
도로형태단일로 - 고가도로위      -1.017e-02  4.424e-02  -0.230 0.818243
도로형태단일로 - 교량            -1.209e-03  3.629e-02  -0.033 0.973419
도로형태단일로 - 기타            -3.568e-03  7.683e-03  -0.464 0.642391
도로형태단일로 - 지하차도(도로)내 -7.080e-03  2.879e-02  -0.246 0.805752
도로형태단일로 - 터널            -3.430e-03  6.172e-02  -0.056 0.955677
도로형태미분류 - 미분류          -9.725e-02  2.907e-01  -0.334 0.738012
도로형태주차장 - 주차장          -5.744e-02  3.499e-02  -1.642 0.100628
노면상태기타                      9.789e-02  7.646e-02   1.280 0.200444
노면상태서리/결빙                 2.658e-02  1.188e-01   0.224 0.822877
노면상태적설                     -7.477e-02  4.109e-01  -0.182 0.855617
노면상태젖음/습기                 2.166e-02  9.601e-03   2.256 0.024083 *
노면상태침수                     -2.833e-01  2.913e-01  -0.973 0.330772
요일목요일                       -3.678e-04  9.293e-03  -0.040 0.968431
요일수요일                        1.387e-04  9.200e-03   0.015 0.987970
요일월요일                       -5.301e-03  9.212e-03  -0.576 0.564949
요일일요일                        1.450e-02  1.021e-02   1.419 0.155776
요일토요일                        8.483e-03  9.278e-03   0.914 0.360585
요일화요일                        5.472e-03  9.138e-03   0.599 0.549312
사고유형차대차                    2.010e-02  4.007e-02   0.502 0.615910
사고유형세부분류길가장자리구역통행중 -4.142e-02  2.617e-02  -1.583 0.113455
사고유형세부분류보도통행중        4.100e-02  2.564e-02   1.599 0.109804
사고유형세부분류정면충돌          7.603e-02  1.873e-02   4.060 4.92e-05 ***
사고유형세부분류차도통행중        1.288e-01  2.322e-02   5.549 2.91e-08 ***
사고유형세부분류추돌              3.543e-02  9.104e-03   3.892 9.96e-05 ***
사고유형세부분류측면충돌          3.753e-03  7.101e-03   0.529 0.597112
사고유형세부분류횡단중            1.700e-01  1.578e-02  10.776  < 2e-16 ***
사고유형세부분류후진중충돌        -8.310e-02  2.114e-02  -3.931 8.49e-05 ***
법규위반교차로운행방법위반        -3.948e-01  4.486e-02  -8.801  < 2e-16 ***
법규위반기타                      -3.626e-01  4.628e-02  -7.834 4.90e-15 ***
법규위반보행자보호의무위반        -3.620e-01  4.721e-02  -7.668 1.82e-14 ***
법규위반불법유턴                  -3.606e-01  4.990e-02  -7.228 5.05e-13 ***
법규위반신호위반                  -2.634e-01  4.449e-02  -5.921 3.25e-09 ***
법규위반안전거리미확보            -3.979e-01  4.434e-02  -8.973  < 2e-16 ***
법규위반안전운전불이행            -3.906e-01  4.386e-02  -8.906  < 2e-16 ***
법규위반중앙선침범                -2.639e-01  4.696e-02  -5.619 1.94e-08 ***
법규위반직진우회전진행방해        -4.019e-01  4.667e-02  -8.612  < 2e-16 ***
법규위반차로위반                  -3.817e-01  4.671e-02  -8.171 3.20e-16 ***
가해운전자차종건설기계            6.247e-02  4.751e-02   1.315 0.188600
가해운전자차종기타불명           -7.289e-02  8.924e-02  -0.817 0.414105
가해운전자차종농기계              -1.899e-01  4.130e-01  -0.460 0.645666
가해운전자차종사륜오토바이(ATV)   5.201e-03  2.095e-01   0.025 0.980198
가해운전자차종승용                -4.554e-02  4.139e-02  -1.100 0.271250
가해운전자차종승합                -1.457e-02  4.379e-02  -0.333 0.739374
가해운전자차종원동기              5.476e-03  4.792e-02   0.114 0.909027
가해운전자차종이륜                -2.784e-02  4.202e-02  -0.663 0.507545
가해운전자차종자전거              -6.031e-03  4.368e-02  -0.138 0.890190
가해운전자차종특수                1.487e-02  6.365e-02   0.234 0.815262
가해운전자차종화물                7.255e-03  4.201e-02   0.173 0.862883
가해운전자연령                    -8.493e-05  1.661e-04  -0.511 0.609214
피해운전자차종건설기계            -1.680e-01  5.790e-02  -2.901 0.003725 **
피해운전자차종기타불명            4.762e-02  9.976e-02   0.477 0.633134
피해운전자차종농기계              3.870e-01  1.878e-01   2.060 0.039363 *
피해운전자차종보행자                    NA         NA      NA       NA
피해운전자차종사륜오토바이(ATV)   3.474e-02  1.204e-01   0.289 0.772894
피해운전자차종승용                -1.769e-01  3.876e-02  -4.563 5.07e-06 ***
피해운전자차종승합                -1.161e-01  4.171e-02  -2.783 0.005382 **
피해운전자차종원동기              -3.264e-02  4.365e-02  -0.748 0.454629
피해운전자차종이륜                1.983e-02  3.918e-02   0.506 0.612775
피해운전자차종자전거              4.226e-02  4.021e-02   1.051 0.293295
피해운전자차종특수                -3.831e-02  6.978e-02  -0.549 0.582950
피해운전자차종화물                -1.491e-01  4.013e-02  -3.716 0.000203 ***
피해운전자연령                    2.898e-03  1.658e-04  17.478  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4107 on 25925 degrees of freedom
Multiple R-squared:  0.09977,   Adjusted R-squared:  0.09755
F-statistic: 44.89 on 64 and 25925 DF,  p-value: < 2.2e-16
```

- *은 해당 변수가 유의미한 정도를 나타내므로, 주로 도로 형태, 사고 유형, 법규 관련 독립변수와 종속변수의 관련성이 높은 것을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '사망중상자여부'인 경우

3) 로지스틱 회귀 모델 mback 생성

```
print("backward model mback")
mback <- step(m, direction = "backward")
```

```
[1] "backward model mback"
Start:  AIC=-46195.38
사망중상자여부 ~ 도로형태 + 노면상태 + 요일 +
    사고유형 + 사고유형세부분류 + 법규위반 +
    가해운전자차종 + 가해운전자연령 + 피해운전자차종 +
    피해운전자연령


Step:  AIC=-46195.38
사망중상자여부 ~ 도로형태 + 노면상태 + 요일 +
    사고유형세부분류 + 법규위반 + 가해운전자차종 +
    가해운전자연령 + 피해운전자차종 + 피해운전자연령

                Df Sum of Sq    RSS    AIC
- 요일           6     0.860 4373.1 -46202
- 노면상태       5     1.301 4373.6 -46198
- 가해운전자연령  1     0.044 4372.3 -46197
<none>                       4372.3 -46195
- 도로형태      10     5.836 4378.1 -46181
- 가해운전자차종 11    11.531 4383.8 -46149
- 사고유형세부분류  8   33.824 4406.1 -46011
- 법규위반      10    53.842 4426.1 -45897
- 피해운전자연령  1    51.518 4423.8 -45893
- 피해운전자차종 12   152.697 4525.0 -45327

Step:  AIC=-46202.27
사망중상자여부 ~ 도로형태 + 노면상태 + 사고유형세부분류 +
    법규위반 + 가해운전자차종 + 가해운전자연령 +
    피해운전자차종 + 피해운전자연령
```

- 모델 m에서 "backward" 방식을 이용하여 필요 없는 설명변수를 제거하여 mback모델 생성

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '사망중상자여부'인 경우

3) 예측

```
predict_value <- predict(mback, test_set, type = "response") %>% tibble(predict_value= .)
predict_value %>% show()
```

```
# A tibble: 7,299 × 1
   predict_value
           <dbl>
 1         0.119
 2         0.0927
 3         0.359
 4         0.165
 5         0.348
 6         0.226
 7         0.224
 8         0.258
 9         0.163
10         0.365
# ℹ 7,289 more rows
```

```
predict_df <- test_set %>% select(사망중상자여부) %>% dplyr::bind_cols(., predict_value)
predict_df %>% show()
```

```
# A tibble: 7,299 × 2
   사망중상자여부 predict_value
            <dbl>         <dbl>
 1              0         0.119
 2              1         0.0927
 3              0         0.359
 4              1         0.165
 5              1         0.348
 6              0         0.226
 7              0         0.224
 8              0         0.258
 9              0         0.163
10              1         0.365
# ℹ 7,289 more rows
```

```
predict_df <- predict_df %>%
  mutate(predict_사망중상자여부 = as.factor(ifelse(predict_value > 0.5, 1, 0)))

predict_df %>% show()
```

```
# A tibble: 7,299 × 3
   사망중상자여부 predict_value predict_사망중상자여부
            <dbl>         <dbl> <fct>
 1              0         0.119 0
 2              1         0.0927 0
 3              0         0.359 0
 4              1         0.165 0
 5              1         0.348 0
 6              0         0.226 0
 7              0         0.224 0
 8              0         0.258 0
 9              0         0.163 0
10              1         0.365 0
# ℹ 7,289 more rows
```

- 정답, 모델이 예측한 확률, 모델이 예측한 클래스 순서로 predict_df에 저장한다.

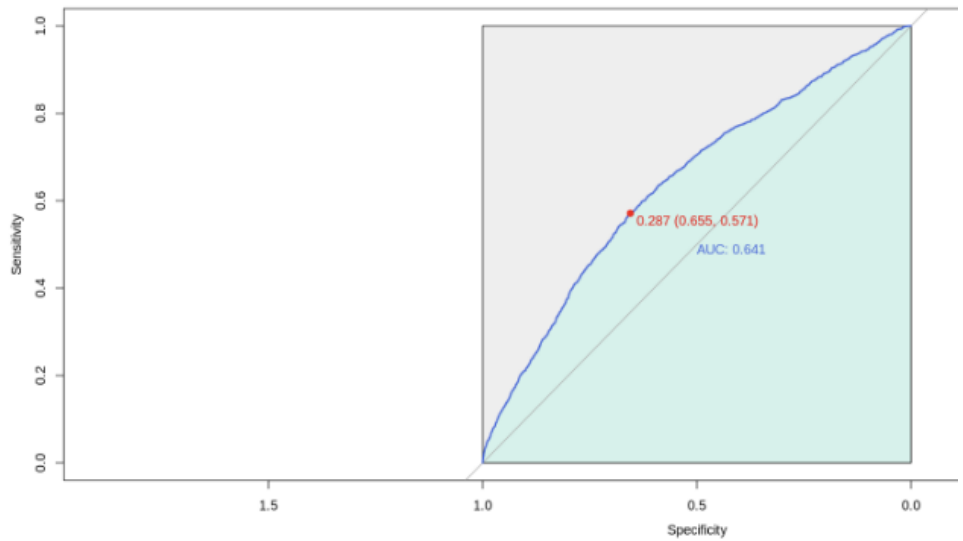
ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '사망중상자여부'인 경우

4) ROC curve

```
roc_c <- pROC::roc(predict_df$사망중상자여부 , predict_df$predict_value )
pROC::plot.roc(roc_c,
               col = " royalblue",
               print.auc=TRUE,
               max.auc.polygon=TRUE,
               print.thres=TRUE, print.thres.pch=19, print.thres.col="red",
               auc.polygon=TRUE, auc.polygon.col="#D1F2EB")
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

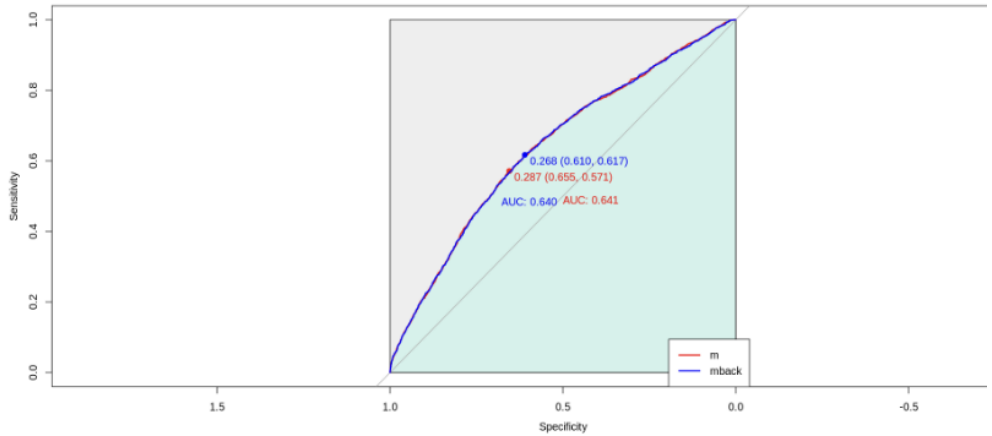



```
pROC::coords(roc_c, "best", ret="threshold", transpose=F)
```

A data.frame:
1 × 1

threshold
<dbl>
0.2873478

```
pROC::plot.roc(roc_c,
col = "red",
print.auc=T,
max.auc.polygon=T,
print.thres=T, print.thres.pch=19, print.thres.col = "red",
auc.polygon=T, auc.polygon.col="#D1F2EB")

pROC::plot.roc(roc_c_mback, add=T, # 기본 그래프에 추가할 수 있도록 설정
 col="blue", print.auc=T, print.auc.adj=c(1.11,1.2), print.thres=T, print.thres.pch=19, print.thres.col = "blue")

legend("bottomright", legend=c("m", "mback"), col=c("red", "blue"), lwd = 2)
```



- ROC 곡선이 가운데 대각선에 위치할수록 낮은 성능을 나타내므로, 현재 모델은 높은 성능을 가지지 않음을 알 수 있다.
- 모델 m과 mback은 거의 비슷한 형태의 ROC curve와 성능을 가짐을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱 회귀 – 종속변수가 '사망중상자여부'인 경우

5) Confusion Matrix

```
caret::confusionMatrix(predict_cutoff_roc$사망중상자여부, predict_cutoff_roc$predict_사망중상자여부)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2961 2063
         1  839 1461

               Accuracy : 0.6038
                 95% CI : (0.5925, 0.615)
    No Information Rate : 0.5188
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1963

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7792
            Specificity : 0.4146
         Pos Pred Value : 0.5894
         Neg Pred Value : 0.6352
             Prevalence : 0.5188
         Detection Rate : 0.4043
   Detection Prevalence : 0.6860
      Balanced Accuracy : 0.5969

       'Positive' Class : 0
```

- 약 60%의 정확도를 가진다.
- P-value가 0과 거의 같으므로 통계적 유의성을 가진다.
- Kappa가 0.1로 매우 낮다. 때문에 모델의 예측이 우수하다고 할 수 없다.
- 약 78%의 민감도를 가진다.
- 약 41%의 특이도를 가진다.
- 민감도가 높고, 특이도가 낮은 것으로 보아 양성 데이터는 잘 분류하지만 음성 데이터는 잘 분류하지 못하는 것을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

1) 데이터 분리

- train 데이터와 test 데이터를 7:3 비율로 분리

```
train2 <- subset(train, select=-사망중상자여부)
head(train2, 2)
```

A tibble: 2 × 11

도로형태	노면상태	요일	사고유형	사고유형세부분류	법규위반	가해운전자차종	가해운전자연령	피해운전자차종	피해운전자연령	경상부상자여부
<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>
단일로 - 기타	건조	화요일	차대사람	길가장자리구역통행중	안전운전불이행	승용	50	보행자	70	0
단일로 - 기타	건조	화요일	차대사람	보도통행중	기타	승용	30	보행자	60	1

```
train_set2 <- sample_frac(train2, 0.7)
test_set2 <- setdiff(train2, train_set2) #setdiff: 차집합

head(train_set2, 2)
```

A tibble: 2 × 11

도로형태	노면상태	요일	사고유형	사고유형세부분류	법규위반	가해운전자차종	가해운전자연령	피해운전자차종	피해운전자연령	경상부상자여부
<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>
단일로 - 기타	건조	수요일	차대차	정면충돌	안전거리미확보	승용	40	화물	40	1
교차로 - 교차로안	건조	금요일	차대차	측면충돌	신호위반	승용	20	승용	60	1

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

2) 로지스틱 회귀 모델 m2 생성

```
m2 <- lm(formula = 경상부상자여부 ~ ., data = train_set2)
summary(m2)
```

```
Call:
lm(formula = 경상부상자여부 ~ ., data = train_set2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04082  0.02287  0.08491  0.20099  0.72105

Coefficients: (1 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.982e-01  5.388e-02   9.246  < 2e-16 ***
도로형태교차로 - 교차로안       -1.683e-02  7.937e-03  -2.121 0.033951 *
도로형태교차로 - 교차로횡단보도내  2.668e-03  1.411e-02   0.189 0.849991
도로형태기타 - 기타             1.905e-02  1.198e-02   1.590 0.111757
도로형태단일로 - 고가도로위      1.076e-01  4.068e-02   2.646 0.008157 **
도로형태단일로 - 교량            1.151e-02  3.019e-02   0.381 0.702975
도로형태단일로 - 기타            2.097e-03  6.704e-03   0.313 0.754421
도로형태단일로 - 지하차도(도로)내  4.201e-02  2.552e-02   1.646 0.099802 .
도로형태단일로 - 터널            8.712e-02  5.853e-02   1.488 0.136647
도로형태미분류 - 미분류          1.421e-01  2.076e-01   0.685 0.493605
도로형태주차장 - 주차장          2.204e-02  2.904e-02   0.759 0.447988
노면상태기타                    -9.243e-02  6.456e-02  -1.432 0.152233
노면상태서리/결빙                1.052e-01  9.958e-02   1.056 0.290847
노면상태젖음/습기               -1.703e-02  8.357e-03  -2.038 0.041599 *
노면상태침수                     1.555e-01  2.544e-01   0.611 0.541094
요일목요일                      -2.001e-03  8.110e-03  -0.247 0.805103
요일수요일                      -7.297e-03  8.064e-03  -0.905 0.365580
요일월요일                      -3.120e-03  8.062e-03  -0.387 0.698784
요일일요일                      -3.710e-03  8.984e-03  -0.413 0.679661
요일토요일                      -6.646e-03  8.143e-03  -0.816 0.414386
요일화요일                      -9.630e-04  8.040e-03  -0.120 0.904659
사고유형차대차                  -1.517e-02  3.623e-02  -0.419 0.675450
사고유형세부분류길가장자리구역통행중  6.562e-02  2.274e-02   2.886 0.003905 **
사고유형세부분류보도통행중      -3.448e-02  2.291e-02  -1.505 0.132306
사고유형세부분류정면충돌        -1.445e-02  1.605e-02  -0.900 0.367908
사고유형세부분류차도통행중      -1.416e-01  2.014e-02  -7.030 2.11e-12 ***
사고유형세부분류추돌             7.807e-03  7.982e-03   0.978 0.327989
사고유형세부분류측면충돌         6.886e-03  6.178e-03   1.115 0.265057
사고유형세부분류횡단중          -1.684e-01  1.382e-02 -12.187  < 2e-16 ***
사고유형세부분류후진중충돌       6.559e-02  1.871e-02   3.505 0.000457 ***
법규위반교차로운행방법위반       2.494e-01  3.895e-02   6.403 1.55e-10 ***
법규위반기타                    2.265e-01  4.032e-02   5.618 1.95e-08 ***
법규위반보행자보호의무위반       2.304e-01  4.092e-02   5.629 1.83e-08 ***
법규위반불법유턴                1.824e-01  4.333e-02   4.210 2.56e-05 ***
법규위반신호위반                1.812e-01  3.863e-02   4.690 2.75e-06 ***
법규위반안전거리미확보          2.503e-01  3.846e-02   6.508 7.77e-11 ***
법규위반안전운전불이행          2.379e-01  3.805e-02   6.253 4.10e-10 ***
법규위반중앙선침범              1.781e-01  4.087e-02   4.357 1.32e-05 ***
법규위반직진우회전진행방해      2.674e-01  4.043e-02   6.615 3.78e-11 ***
법규위반차로위반                2.170e-01  4.050e-02   5.358 8.50e-08 ***
가해운전자차종건설기계          -4.556e-02  4.098e-02  -1.112 0.266265
가해운전자차종기타불명           4.480e-02  8.770e-02   0.511 0.609453
가해운전자차종사륜오토바이(ATV)  -1.260e-01  1.844e-01  -0.683 0.494487
가해운전자차종승용              7.663e-02  3.534e-02   2.168 0.030149 *
가해운전자차종승합              5.200e-02  3.751e-02   1.387 0.165602
가해운전자차종원동기            1.584e-02  4.118e-02   0.385 0.700423
가해운전자차종이륜              7.875e-02  3.590e-02   2.194 0.028269 *
가해운전자차종자전거            1.466e-02  3.736e-02   0.392 0.694872
가해운전자차종특수              1.693e-02  5.202e-02   0.325 0.744911
가해운전자차종화물              3.555e-02  3.589e-02   0.991 0.321935
가해운전자연령                  6.676e-05  1.447e-04   0.461 0.644459
피해운전자차종건설기계          1.488e-01  5.310e-02   2.802 0.005085 **
피해운전자차종기타불명          -1.089e-01  7.244e-02  -1.503 0.132793
피해운전자차종농기계            -6.439e-01  1.642e-01  -3.921 8.84e-05 ***
피해운전자차종보행자                NA         NA         NA       NA
피해운전자차종사륜오토바이(ATV)  -2.852e-02  1.028e-01  -0.277 0.781553
피해운전자차종승용              2.232e-01  3.510e-02   6.358 2.08e-10 ***
피해운전자차종승합              2.097e-01  3.748e-02   5.596 2.21e-08 ***
피해운전자차종원동기            3.199e-02  3.922e-02   0.816 0.414792
피해운전자차종이륜              -9.110e-05  3.545e-02  -0.003 0.997949
피해운전자차종자전거            -4.869e-02  3.630e-02  -1.342 0.179765
피해운전자차종특수              1.365e-01  6.123e-02   2.230 0.025777 *
피해운전자차종화물              1.994e-01  3.623e-02   5.505 3.72e-08 ***
피해운전자연령                  -2.152e-03  1.446e-04 -14.877  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3585 on 25927 degrees of freedom
Multiple R-squared:  0.1407,    Adjusted R-squared:  0.1387
F-statistic: 68.49 on 62 and 25927 DF,  p-value: < 2.2e-16
```

- *은 해당 변수가 유의미한 정도를 나타내므로, 주로 도로 형태, 사고 유형, 법규, 피해 운전자의 차종 관련 독립변수와 종속변수의 관련성이 높은 것을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

3) 로지스틱 회귀 모델 mback2 생성

```
print("backward model mback2")
mback2 <- step(m2, direction = "backward")
```

```
[1] "backward model mback2"
Start:  AIC=-53258.53
경상부상자여부 ~ 도로형태 + 노면상태 + 요일 +
    사고유형 + 사고유형세부분류 + 법규위반 +
    가해운전자차종 + 가해운전자연령 + 피해운전자차종 +
    피해운전자연령


Step:  AIC=-53258.53
경상부상자여부 ~ 도로형태 + 노면상태 + 요일 +
    사고유형세부분류 + 법규위반 + 가해운전자차종 +
    가해운전자연령 + 피해운전자차종 + 피해운전자연령

                  Df Sum of Sq    RSS    AIC
- 요일             6     0.176 3332.5 -53269
- 가해운전자연령    1     0.027 3332.4 -53260
- 노면상태          4     0.988 3333.3 -53259
<none>                         3332.3 -53259
- 도로형태         10     3.297 3335.6 -53253
- 가해운전자차종   10    12.030 3344.4 -53185
- 법규위반         10    16.487 3348.8 -53150
- 사고유형세부분류  8    28.289 3360.6 -53055
- 피해운전자연령    1    28.448 3360.8 -53040
- 피해운전자차종   12   224.325 3556.7 -51589

Step:  AIC=-53269.16
경상부상자여부 ~ 도로형태 + 노면상태 + 사고유형세부분류 +
    법규위반 + 가해운전자차종 + 가해운전자연령 +
    피해운전자차종 + 피해운전자연령

                  Df Sum of Sq    RSS    AIC
- 가해운전자연령    1     0.027 3332.5 -53271
- 노면상태          4     0.988 3333.5 -53269
<none>                         3332.5 -53269
- 도로형태         10     3.298 3335.8 -53263
- 가해운전자차종   10    12.005 3344.5 -53196
- 법규위반         10    16.498 3349.0 -53161
- 사고유형세부분류  8    28.279 3360.8 -53066
```

- 모델 m2에서 "backward" 방식을 이용하여 필요 없는 설명변수를 제거하여 mback2모델 생성

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

3) 예측

```
# 예측
predict_value2 <- predict(mback2, test_set2, type = "response") %>% tibble(predict_value2= .)
predict_value2 %>% show()
```

```
# A tibble: 7,135 × 1
   predict_value2
            <dbl>
 1          0.729
 2          0.985
 3          0.923
 4          0.496
 5          0.866
 6          0.877
 7          0.902
 8          0.957
 9          0.685
10          0.563
# ℹ 7,125 more rows
```

```
predict_df2 <- test_set2 %>% select(경상부상자여부) %>% dplyr::bind_cols(., predict_value2)
predict_df2 %>% show()
```

```
# A tibble: 7,135 × 2
   경상부상자여부 predict_value2
            <dbl>          <dbl>
 1              0          0.729
 2              0          0.985
 3              0          0.923
 4              1          0.496
 5              1          0.866
 6              1          0.877
 7              1          0.902
 8              1          0.957
 9              1          0.685
10              0          0.563
# ℹ 7,125 more rows
```

```
predict_df2 <- predict_df2 %>%
  mutate(predict_경상부상자여부 = as.factor(ifelse(predict_value2 > 0.5, 1, 0)))

predict_df2 %>% show()
```

```
# A tibble: 7,135 × 3
   경상부상자여부 predict_value2 predict_경상부상자여부
            <dbl>          <dbl> <fct>
 1              0          0.729 1
 2              0          0.985 1
 3              0          0.923 1
 4              1          0.496 0
 5              1          0.866 1
 6              1          0.877 1
 7              1          0.902 1
 8              1          0.957 1
 9              1          0.685 1
10              0          0.563 1
# ℹ 7,125 more rows
```

- 정답, 모델이 예측한 확률, 모델이 예측한 클래스 순서로 predict_df2에 저장한다.

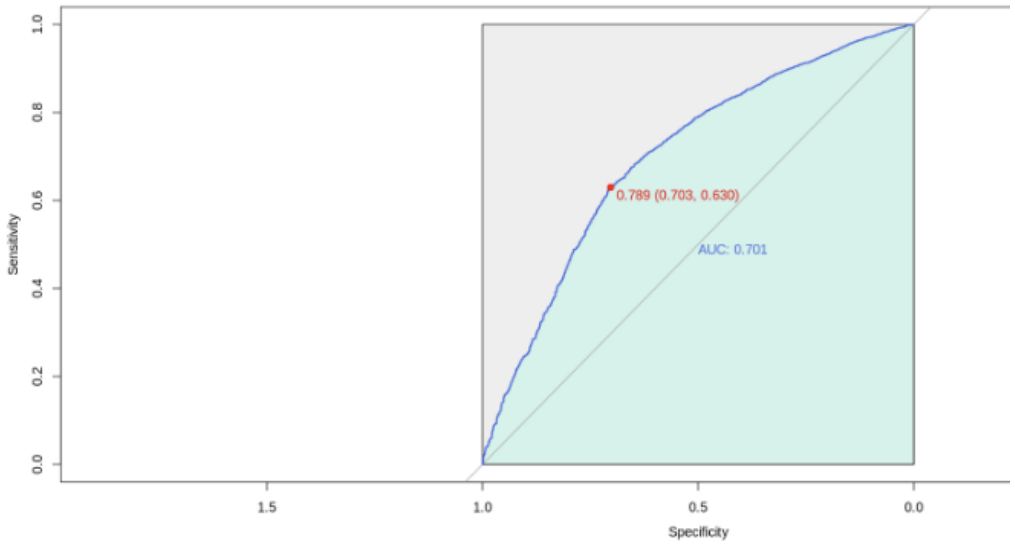
ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

4) ROC curve

```
roc_c2 <- pROC::roc(predict_df2$경상부상자여부 , predict_df2$predict_value2 )
pROC::plot.roc(roc_c2,
               col = " royalblue",
               print.auc=TRUE,
               max.auc.polygon=TRUE,
               print.thres=TRUE, print.thres.pch=19, print.thres.col="red",
               auc.polygon=TRUE, auc.polygon.col="#D1F2EB")
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

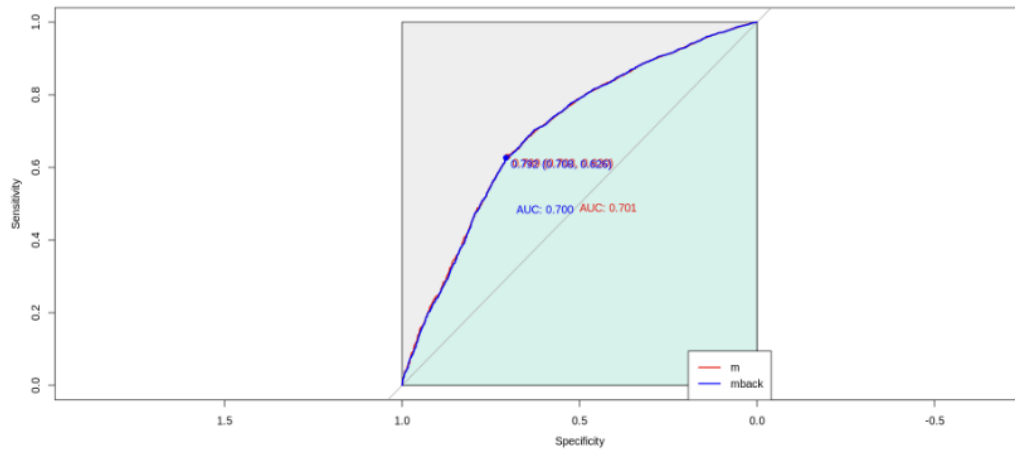



```
pROC::coords(roc_c2, "best", ret="threshold", transpose=F)
```

A data.frame:
1 × 1

threshold
<dbl>
0.7887513

```
pROC::plot.roc(roc_c2,
col = "red",
print.auc=T,
max.auc.polygon=T,
print.thres=T, print.thres.pch=19, print.thres.col = "red",
auc.polygon=T, auc.polygon.col="#D1F2EB")

pROC::plot.roc(roc_c_mback, add=T, # 기본 그래프에 추가할 수 있도록 설정
 col="blue", print.auc=T, print.auc.adj=c(1.11,1.2), print.thres=T, print.thres.pch=19, print.thres.col = "blue")

legend("bottomright", legend=c("m", "mback"), col=c("red", "blue"), lwd = 2)
```



- ROC 곡선이 가운데 대각선에 위치할수록 낮은 성능을 나타내므로, 현재 모델은 높은 성능을 가지지 않음을 알 수 있다. 하지만 종속변수가 '경상부상자 여부 ' 인 경우보다는 높은 성능을 가진다.
- 모델 m과 mback은 거의 비슷한 형태의 ROC curve와 성능을 가짐을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

1.로지스틱회귀 – 종속변수가 '경상부상자여부'인 경우

5) Confusion Matrix

```
caret::confusionMatrix(predict_cutoff_roc2$경상부상자여부, predict_cutoff_roc2$predict_경상부상자여부)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0    4 1750
         1    2 5379

               Accuracy : 0.7544
                 95% CI : (0.7443, 0.7644)
    No Information Rate : 0.9992
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0029

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6666667
            Specificity : 0.7545238
         Pos Pred Value : 0.0022805
         Neg Pred Value : 0.9996283
             Prevalence : 0.0008409
         Detection Rate : 0.0005606
   Detection Prevalence : 0.2458304
      Balanced Accuracy : 0.7105952

       'Positive' Class : 0
```

- 약 75%의 정확도를 가진다.
- Kappa가 0.0029로 매우 낮다. 때문에 모델의 예측이 우수하다고 할 수 없다.
- 약 67%의 민감도를 가진다.
- 약 75%의 특이도를 가진다.
- 민감도가 낮고, 특이도가 높은 것으로 보아 양성 데이터에 비해 음성 데이터를 더 잘 분류하는 것을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

2. 결정트리 – 종속변수가 '사망중상자여부'인 경우

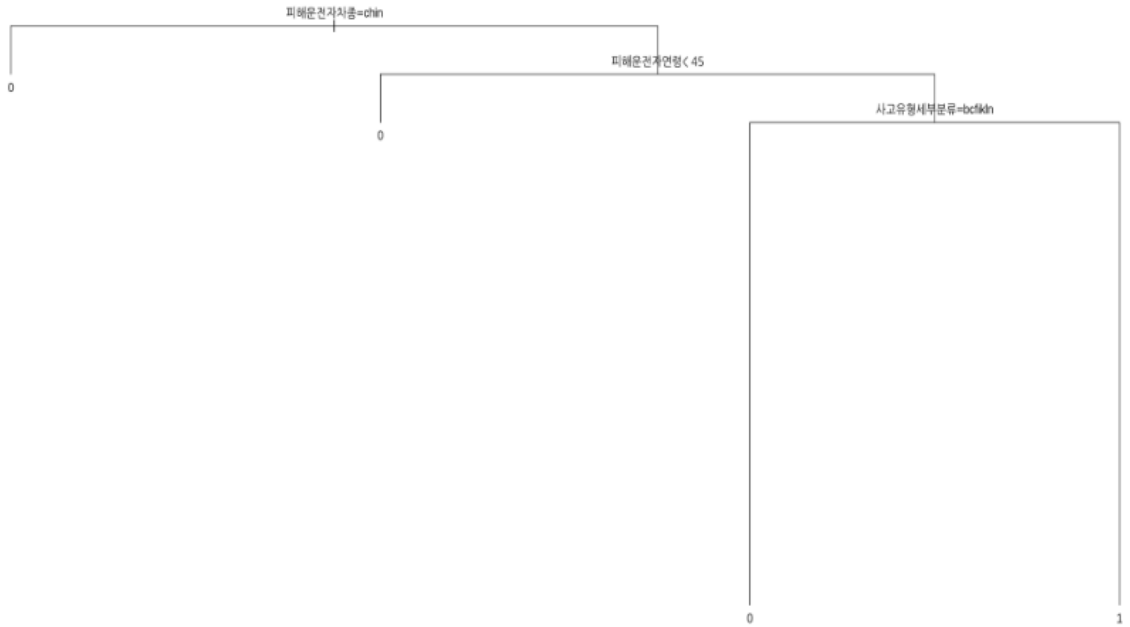
1) 모델 생성

```
model_dt <- rpart(formula = 사망중상자여부 ~ ., data = train_set, method = "class")
summary(model_dt)
```

```
Call:
rpart(formula = 사망중상자여부 ~ ., data = train_set,
    method = "class")
  n= 25990

          CP nsplit rel error    xerror       xstd
1 0.01154937      0 1.0000000 1.0000000 0.01077974
2 0.01000000      3 0.9653519 0.9712297 0.01067402

Variable importance
  피해운전자차종         사고유형   피해운전자연령 사고유형세부분류
              42               18               16               15
        법규위반         도로형태
               6                2

Node number 1: 25990 observations,    complexity param=0.01154937
  predicted class=0  expected loss=0.2487495  P(node) =1
    class counts: 19525  6465
   probabilities: 0.751 0.249
  left son=2 (16301 obs) right son=3 (9689 obs)
  Primary splits:
      피해운전자차종   splits as  -RLRRRRLLRRRRL, improve=535.7388, (0 missing)
      사고유형세부분류 splits as  -LL--R--LRLLRL, improve=289.8201, (0 missing)
      사고유형         splits as  RL-, improve=243.1619, (0 missing)
      법규위반         splits as  RLRRLRLLRLL, improve=203.4320, (0 missing)
      피해운전자연령   < 45 to the left,  improve=185.8117, (0 missing)
  Surrogate splits:
      사고유형         splits as  RL-, agree=0.791, adj=0.438, (0 split)
      사고유형세부분류 splits as  -RR--R--LRLLRL, agree=0.726, adj=0.266, (0 split)
      피해운전자연령   < 65 to the left,  agree=0.689, adj=0.167, (0 split)
      법규위반         splits as  LLRRLLLLLLL, agree=0.668, adj=0.111, (0 split)
      도로형태         splits as  LLRLLLLLLLL, agree=0.650, adj=0.061, (0 split)
```

2) plot()을 이용하여 DecisionTree 그리기

```
options(repr.plot.width=20, repr.plot.height=10)
plot(model_dt)
text(model_dt)
```

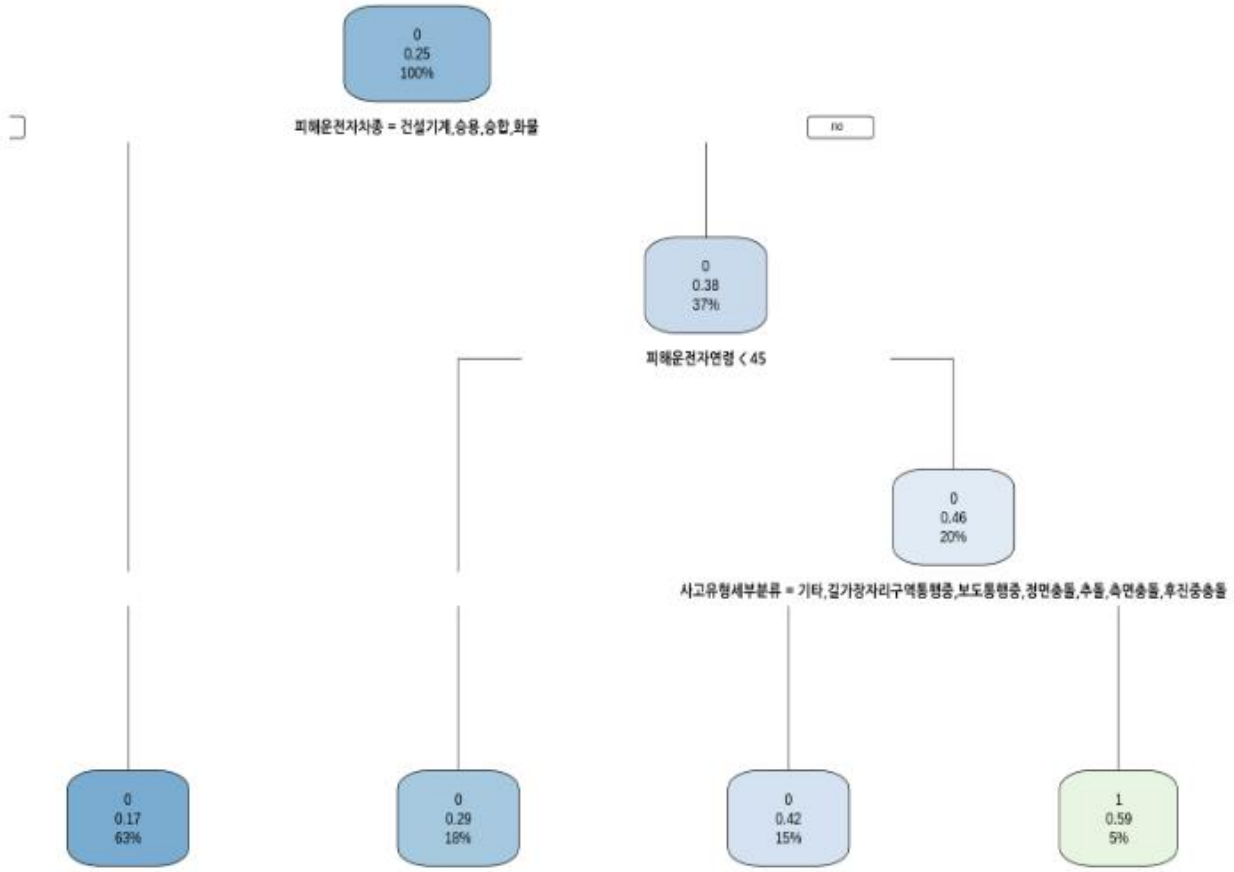


ⓒSaebyeol Yu. Saebyeol's PowerPoint

2. 결정트리 – 종속변수가 '사망중상자여부'인 경우

3) rpart.plot() 을 이용하여 DecisionTree 그리기

```
options(repr.plot.width=20, repr.plot.height=10)
rpart.plot(model_dt)
```



- 피해 운전자 차종, 피해운전자 연령, 사고유형 세부 분류 변수를 기준으로 노드가 분리되었다.

- 가장 왼쪽부터 세개의 노드는 0으로, 가장 오른쪽 노드는 1로 분류되었다.

- 네 개의 리프노드 중 맨 왼 쪽 리프 노드는 다른 클래스 데이터가 해당 노드에 존재하는 비율이 17%로, 리프노드 중 가장 잘 분류된 것을 알 수 있다.

- 맨 왼쪽 리프노드에는 가장 많은 63%의 데이터가, 두번째 리프노드에는 18%, 세번째 리프노드에는 15%, 마지막 리프노드는 5%가 존재한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

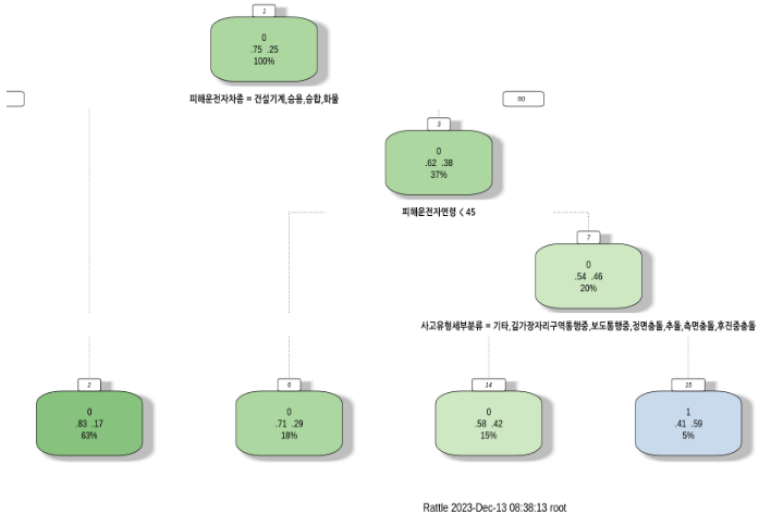
2. 결정트리 – 종속변수가 '사망중상자여부'인 경우

4) 가지치기 후 DecisionTree 그리기

```
min_xerror_cp <- model_dt$cptable %>% as_tibble() %>% filter(xerror == min(xerror))  %>% pull(CP)
print("min_xerror_cp = ")
min_xerror_cp
```

```
[1] "min_xerror_cp = "
0.01
```

```
model_pr <- rpart::prune(model_dt, cp = min_xerror_cp)
fancyRpartPlot(model_dt)
fancyRpartPlot(model_pr)
```



- rpart의 prune()을 이용해 가지치기 후 DecisionTree를 그렸다.
- 가지치기를 하기 전의 DecisionTree와 동일하게 분류되었다.

5) Confusion matrix

```
cm <- caret::confusionMatrix(predict_check$predict_value, predict_check$사망중상자여부)
cm
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4900 1990
         1  199  210

               Accuracy : 0.7001
                 95% CI : (0.6894, 0.7106)
    No Information Rate : 0.6986
    P-Value [Acc > NIR] : 0.395

                  Kappa : 0.0734

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.96097
            Specificity : 0.09545
         Pos Pred Value : 0.71118
         Neg Pred Value : 0.51345
             Prevalence : 0.69859
         Detection Rate : 0.67132
   Detection Prevalence : 0.94396
      Balanced Accuracy : 0.52821

       'Positive' Class : 0
```

- 70%의 정확도를 가진다.
-96%의 매우 높은 Sensitivity와
0.095%의 매우 낮은 Specificity를 가진다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

2. 결정트리 – 종속변수가 '경상부상자여부'인 경우

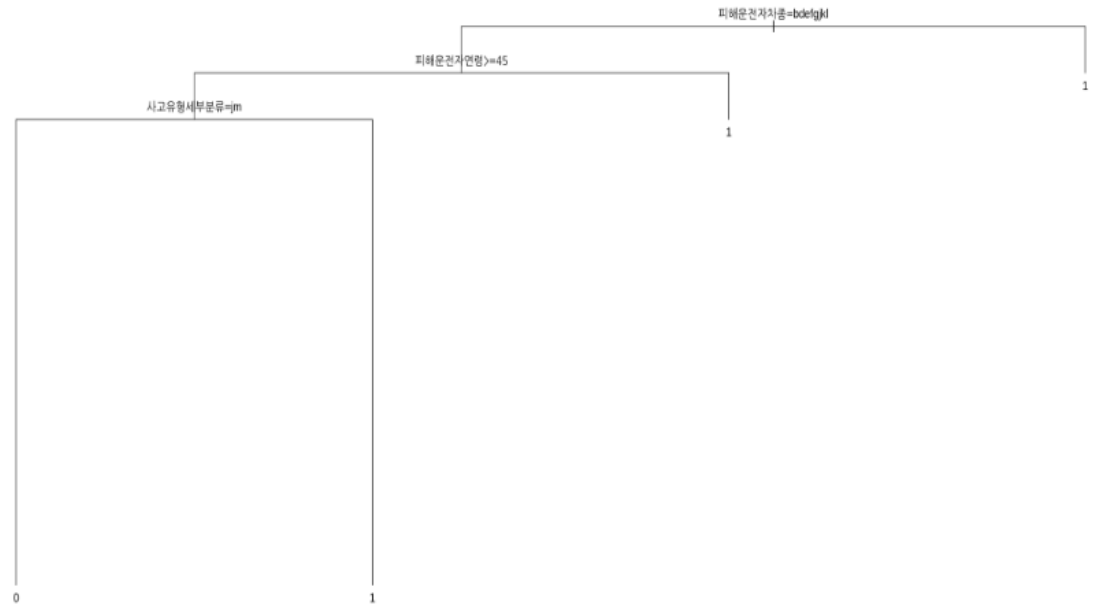
1) 모델 생성

```
model_dt2 <- rpart(formula = 경상부상자여부 ~ ., data = train_set2, method = "class")
summary(model_dt2)
```

```
Call:
rpart(formula = 경상부상자여부 ~ ., data = train_set2,
    method = "class")
  n= 25990

          CP nsplit rel error    xerror       xstd
1 0.01074585      0 1.0000000 1.0000000 0.01317622
2 0.01000000      3 0.9677625 0.9913043 0.01313144

Variable importance
  피해운전자차종         사고유형 사고유형세부분류   피해운전자연령
              44               19               15               13
        법규위반         도로형태
               6                3

Node number 1: 25990 observations,    complexity param=0.01074585
  predicted class=1  expected loss=0.1814159  P(node) =1
    class counts:  4715 21275
   probabilities: 0.181 0.819
  left son=2 (9712 obs) right son=3 (16278 obs)
  Primary splits:
      피해운전자차종   splits as  -LRLLLLRRLLLRR, improve=835.5095, (0 missing)
      사고유형         splits as  LR-, improve=416.2752, (0 missing)
      사고유형세부분류 splits as  -RR--L--RLRRLR, improve=389.6266, (0 missing)
      피해운전자연령   < 65 to the right, improve=172.9184, (0 missing)
      법규위반         splits as  LRRLRRRRRRR, improve=115.0149, (0 missing)
  Surrogate splits:
      사고유형         splits as  LR-, agree=0.790, adj=0.438, (0 split)
      사고유형세부분류 splits as  -LL--L--RLRRLR, agree=0.724, adj=0.261, (0 split)
      피해운전자연령   < 65 to the right, agree=0.688, adj=0.166, (0 split)
      법규위반         splits as  RRLLRRRRRRR, agree=0.666, adj=0.107, (0 split)
      도로형태         splits as  RRLLRRRRRRR, agree=0.651, adj=0.066, (0 split)
```

2) plot()을 이용하여 DecisionTree 그리기

```
options(repr.plot.width=20, repr.plot.height=10)
plot(model_dt2)
text(model_dt2)
```

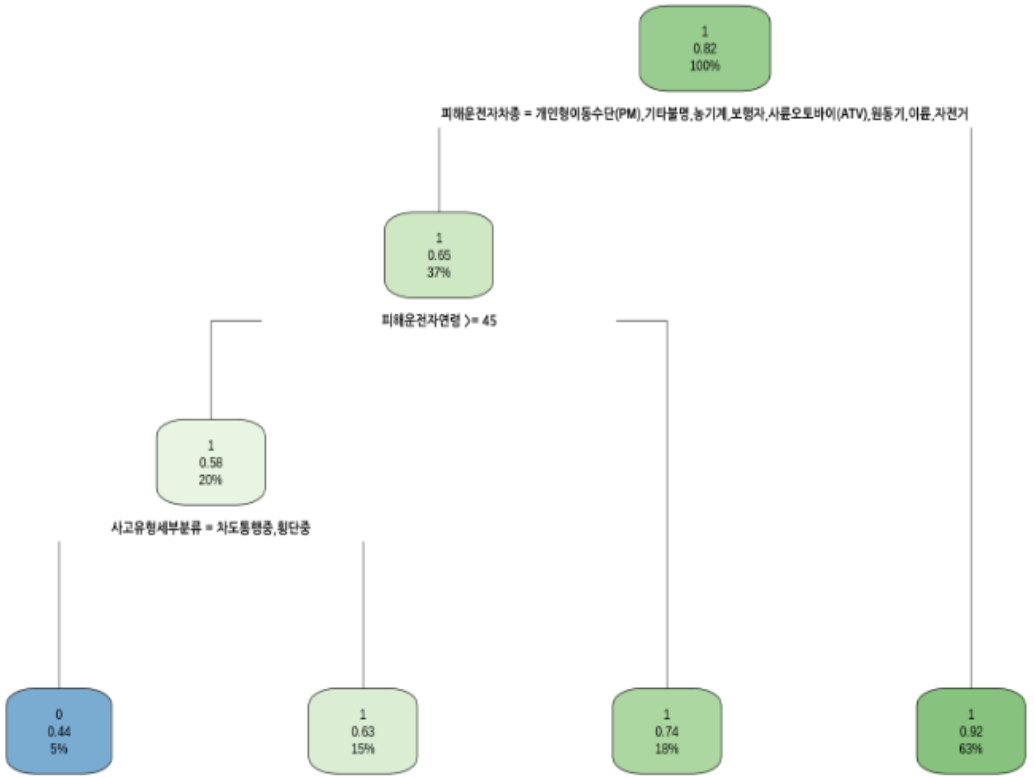


ⓒSaebyeol Yu. Saebyeol's PowerPoint

2. 결정트리 – 종속변수가 '경상부상자여부'인 경우

3) rpart.plot() 을 이용하여 DecisionTree 그리기

```
] options(repr.plot.width=20, repr.plot.height=10)
  rpart.plot(model_dt2)
```



- 피해 운전자 차종, 연령, 사고유형 세부분류 변수를 기준으로 노드가 분리되었다.

- 가장 오른쪽부터 세개의 노드는 1로, 가장 오른쪽 노드는 0으로 분류되었다.

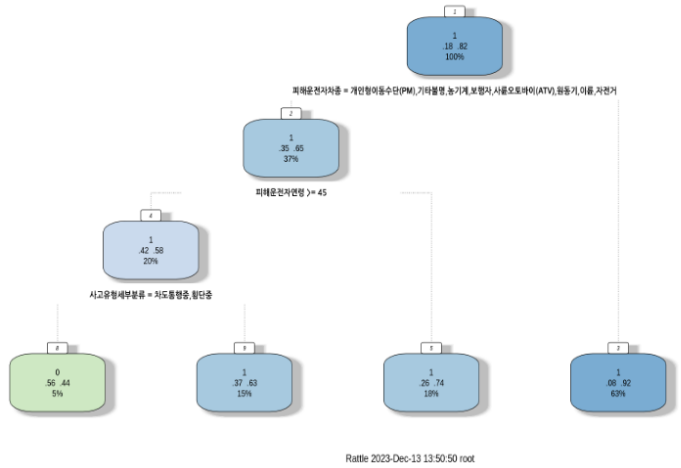
- 맨 오른쪽 리프노드에는 가장 많은 63%의 데이터가, 두번째 리프노드에는 18%, 세번째 리프노드에는 15%, 마지막 리프노드는 5%가 존재한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

2. 결정트리 – 종속변수가 '경상부상자여부'인 경우

4) 가지치기 후 DecisionTree 그리기

```
min_xerror_cp <- model_dt2$cptable %>% as_tibble() %>% filter(xerror == min(xerror))  %>% pull(CP)
print("min_xerror_cp = ")
min_xerror_cp

[1] "min_xerror_cp = "
0.01

model_pr2 <- rpart::prune(model_dt2, cp = min_xerror_cp)
fancyRpartPlot(model_dt2)
fancyRpartPlot(model_pr2)
```



- rpart의 prune()을 이용해 가지치기 후 DecisionTree를 그렸다.

- 가지치기를 하기 전의 DecisionTree와 동일하게 분류되었다.

5) Confusion matrix

```
cm2 <- caret::confusionMatrix(predict_check2$predict_value2, predict_check2$경상부상자여부)
cm2

Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  225  212
         1 1550 5170

               Accuracy : 0.7538
                 95% CI : (0.7437, 0.7638)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : 0.3669

                  Kappa : 0.1169

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.12676
            Specificity : 0.96061
         Pos Pred Value : 0.51487
         Neg Pred Value : 0.76935
             Prevalence : 0.24801
         Detection Rate : 0.03144
   Detection Prevalence : 0.06106
      Balanced Accuracy : 0.54369

       'Positive' Class : 0
```

- 약 75%의 정확도를 가진다.
-1%의 매우 낮은 Sensitivity와
96%의 매우 높은 Specificity를 가진다.

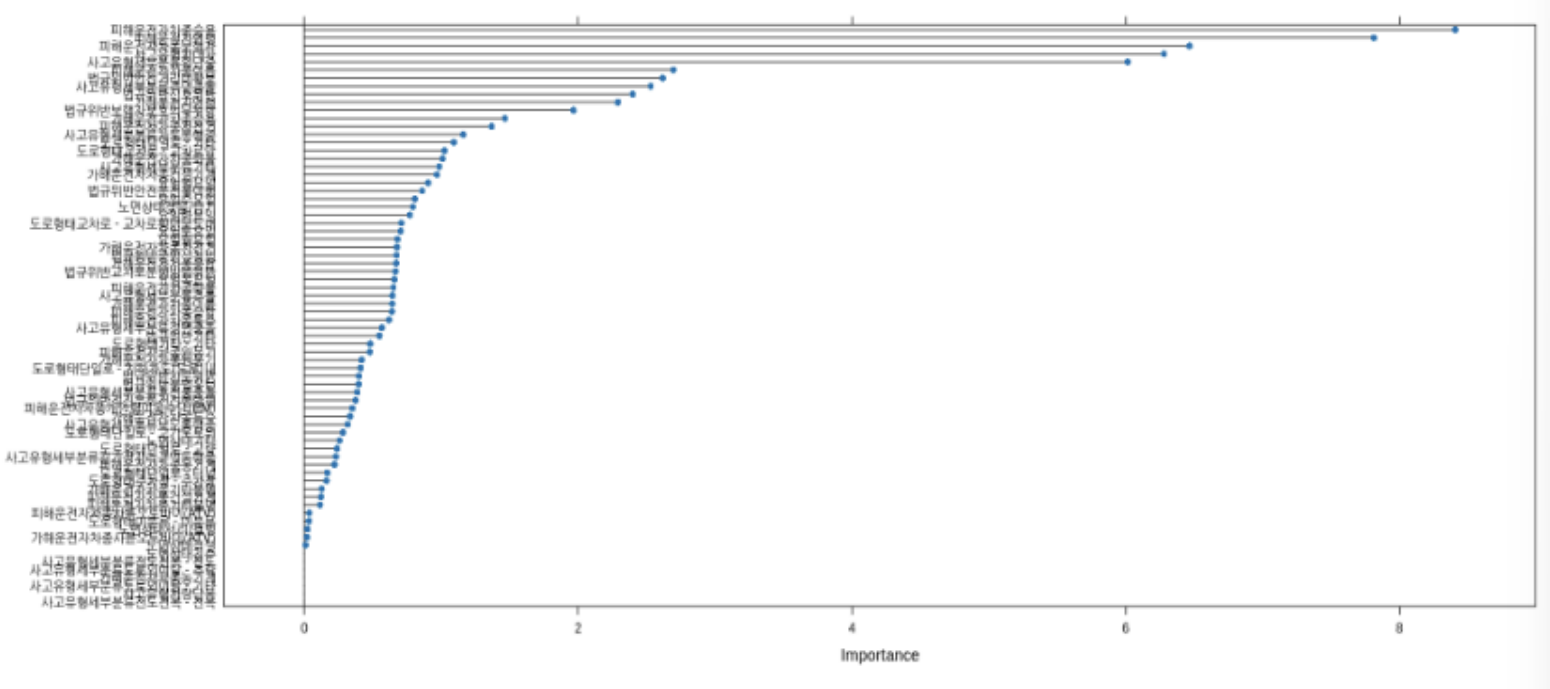
ⓒSaebyeol Yu. Saebyeol's PowerPoint

3. Random Forest

- 변수의 중요도 측정

```
#모형훈련
model_rf = train(사망중상자여부~., data=train_set, method='rf')

#변수의 중요도 측정
importance <- varImp(model_rf, scale=FALSE)
print(importance)
plot(importance)
```

```
rf variable importance

  only 20 most important variables shown (out of 72)

                          Overall
피해운전자차종승용          8.4038
피해운전자연령              7.8119
피해운전자차종보행자        6.4643
사고유형차대차              6.2776
사고유형세부분류횡단중      6.0119
피해운전자차종이륜          2.6960
법규위반안전거리미확보      2.6170
사고유형세부분류측면충돌    2.5304
법규위반신호위반            2.3985
가해운전자연령              2.2907
법규위반보행자보호의무위반  1.9662
가해운전자차종승용          1.4656
피해운전자차종자전거        1.3683
사고유형세부분류차도통행중  1.1617
도로형태단일로 - 기타       1.0925
도로형태교차로 - 교차로안   1.0240
가해운전자차종화물          1.0102
사고유형세부분류기타        0.9859
가해운전자차종건설기계      0.9683
요일일요일                  0.9034
```



- 피해 운전자의 차종이 가장 높은 중요도를 가지는 변수임을 알 수 있다.
- 요일이 가장 낮은 중요도를 가지는 변수임을 알 수 있다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

3. Random Forest – 종속변수가 '사망중상자여부'인 경우

1) 배깅, 예측

```
set.seed(66)
model_bagging <- ipred::bagging(사망중상자여부~., data=train_set, nbagg = 100)
```

```
summary(model_bagging)
```

```
        Length Class      Mode
y       25990  -none-     numeric
X          10  data.frame list
mtrees    100  -none-     list
OOB         1  -none-     logical
comb        1  -none-     logical
call        4  -none-     call
```

```
predict_value_bagging <- predict(model_bagging, test_set, type = "class") %>% tibble(predict_value_bagging = .)

predict_value_bagging <- ifelse( predict_value_bagging>0.5, 1, 0)
predict_value_bagging
```

```
0
0
0
0
0
0
```

```
predict_check_bagging <- test_set %>% select(사망중상자여부) %>% dplyr::bind_cols(., predict_value_bagging)
head(predict_check_bagging)
```

A tibble: 6 × 2

사망중상자여부	predict_value_bagging
<dbl>	<dbl>
1	0
0	0
1	0
0	0
0	0
0	0

- bagging 모델을 생성하고 예측 결과를 저장.

2) Confusion Matrix

```
cm <- caret::confusionMatrix(predict_check_bagging$predict_value_bagging, predict_check_bagging$사망중상자여부)
cm
```

```
Warning message in confusionMatrix.default(predict_check_bagging$predict_value_bagging, :
 "Levels are not in the same order for reference and data. Refactoring data to match."
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5442 2511
         1    0    0

               Accuracy : 0.6843
                 95% CI : (0.6739, 0.6945)
    No Information Rate : 0.6843
    P-Value [Acc > NIR] : 0.5054

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.6843
         Neg Pred Value :    NaN
             Prevalence : 0.6843
         Detection Rate : 0.6843
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

- 모든 데이터를 0으로 예측하는 모델이 생성되었다.
- Sensitivity가 100%이고 Specificity가 0%로 성능이 매우 낮은 모델이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

3. Random Forest – 종속변수가 '경상부상자여부'인 경우

1) 배깅, 예측

```
set.seed(66)
model_bagging2 <- ipred::bagging(경상부상자여부~., data=train_set2, nbagg = 100)
```

```
summary(model_bagging2)
```

```
       Length Class      Mode
y      25990  -none-     numeric
X         10  data.frame list
mtrees   100  -none-     list
OOB        1  -none-     logical
comb       1  -none-     logical
call       4  -none-     call
```

```
predict_value_bagging2 <- predict(model_bagging2, test_set2, type = "class") %>% tibble(predict_value_bagging2 = .)

predict_value_bagging2 <- ifelse( predict_value_bagging2 > 0.5, 1, 0)
predict_value_bagging2
```

A matrix: 7135 × 1 of type dbl

predict_value_bagging2
1
1
1

```
predict_check_bagging2 <- test_set2 %>% select(경상부상자여부) %>% dplyr::bind_cols(., predict_value_bagging2)
head(predict_check_bagging2)
```

A tibble: 6 × 2

경상부상자여부 <dbl>	predict_value_bagging2 <dbl>
0	1
0	1
0	1
1	1
1	1
1	1

```
train$경상부상자여부 <- ifelse(train$경상부상자여부 > 0, 1, # 사망중상자가 있으면 1, 없으면 0
                        0)
summary(train$경상부상자여부)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  1.0000  1.0000  0.8178  1.0000  1.0000
```

```
predict_check_bagging2$predict_value_bagging2 <- as.factor(predict_check_bagging2$predict_value_bagging2)
predict_check_bagging2$경상부상자여부 <- as.factor(predict_check_bagging2$경상부상자여부)
```

- bagging 모델을 생성하고 예측 결과를 저장.

2) Confusion Matrix

```
cm2 <- caret::confusionMatrix(predict_check_bagging2$predict_value_bagging2, predict_check_bagging2$경상부상자여부)
cm2
```

```
Warning message in confusionMatrix.default(predict_check_bagging2$predict_value_bagging2, :
 "Levels are not in the same order for reference and data. Refactoring data to match."
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0    0    0
         1 1754 5381

               Accuracy : 0.7542
                 95% CI : (0.744, 0.7641)
    No Information Rate : 0.7542
    P-Value [Acc > NIR] : 0.5064

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7542
             Prevalence : 0.2458
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

- 모든 데이터를 1으로 예측하는 모델이 생성되었다.
- Sensitivity가 0%이고 Specificity가 100%로 성능이 매우 낮은 모델이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM

- **Data set 크기 줄이기**

  - SVM의 튜닝은 매우 높은 시간복잡도를 가지므로 적은 수의 데이터로 나누어 학습했다.

```
# 데이터 작게 나누기
train1_2 <- sample_frac(train1, 0.1) #train1_2를 train, test 전체 데이터셋으로 사용
train_set <- sample_frac(train1_2, 0.7)
test_set <- setdiff(train1_2, train_set) #setdiff: 차집합

train_set
```

A tibble: 2599 × 11

도로형태	노면상태	요일	사고유형	사고유형 세부분류	법규위반	가해운전자차종	가해운전자연령	피해운전자차종	피해운전자연령	사망중상자여부
<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<dbl>	<fct>	<dbl>	<dbl>
단일로 - 기타	건조	목요일	차대차	측면충돌	안전거리미확보	승용	70	승용	60	0
교차로 - 교차로부근	건조	금요일	차대차	기타	안전운전불이행	승용	40	승합	50	1
단일로 - 기타	젖음/습기	월요일	차대차	기타	기타	이륜	40	승용	50	0
교차로 - 교차로안	건조	토요일	차대차	기타	교차로운행방법위반	화물	40	승용	20	0
단일로 - 기타	건조	수요일	차대차	기타	안전거리미	승용	70	승용	30	0

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '사망중상자여부'인 경우

1) 모델 생성

```
svm_basic <- e1071::svm(formula = 사망중상자여부~., data = train_set, type = "C-classification", kernel="radial")

summary(svm_basic)

print("svm_basic:train 데이터 분류 결과")
table(predict(svm_basic,train_set),train_set$사망중상자여부)

print("svm_basic : train 데이터 confusionMatrix 결과")
cm <- caret::confusionMatrix(predict(svm_basic, train_set), train_set$사망중상자여부)
cm
```

```
Call:
svm(formula = 사망중상자여부 ~ ., data = train_set, type = "C-classification",
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1491
```

- SVM()을 이용해 모델을 생성한다.

2) Confusion Matrix

```
cm <- caret::confusionMatrix(predict(svm_basic, test_set), test_set$사망중상자여부)
cm
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 706 322
         1   3   4

               Accuracy : 0.686
                 95% CI : (0.6567, 0.7142)
    No Information Rate : 0.685
    P-Value [Acc > NIR] : 0.4883

                  Kappa : 0.0109

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99577
            Specificity : 0.01227
         Pos Pred Value : 0.68677
         Neg Pred Value : 0.57143
             Prevalence : 0.68502
         Detection Rate : 0.68213
   Detection Prevalence : 0.99324
      Balanced Accuracy : 0.50402

       'Positive' Class : 0
```

- 생성한 SVM 모델로 test data에 대해 예측을 수행하고, confusion matrix를 계산한다.
- Sensitivity가 99%이고 Specificity가 1%로 성능이 매우 낮은 모델이다

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '사망중상자여부'인 경우

3) 튜닝

```
tuned <- e1071::tune.svm(사망중상자여부~., data=train_set, gamma = 10^(-8:1), cost= 1:30)
```

```
summary(tuned)
```

```
Parameter tuning of  'svm' :

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
  0.01   13

- best performance: 0.2408554

- Detailed performance results:
    gamma cost     error dispersion
1   1e-08    1 0.2570181 0.03029200
2   1e-07    1 0.2570181 0.03029200
3   1e-06    1 0.2570181 0.03029200
4   1e-05    1 0.2570181 0.03029200
5   1e-04    1 0.2570181 0.03029200
6   1e-03    1 0.2570181 0.03029200
7   1e-02    1 0.2570181 0.03029200
8   1e-01    1 0.2481631 0.02263225
9   1e+00    1 0.2654871 0.02336233
10  1e+01    1 0.2662489 0.03331077
11  1e-08    2 0.2570181 0.03029200
12  1e-07    2 0.2570181 0.03029200
13  1e-06    2 0.2570181 0.03029200
14  1e-05    2 0.2570181 0.03029200
15  1e-04    2 0.2570181 0.03029200
16  1e-03    2 0.2570181 0.03029200
17  1e-02    2 0.2531675 0.02795310
18  1e-01    2 0.2497015 0.02182708
19  1e+00    2 0.2785670 0.02299487
20  1e+01    2 0.2662489 0.03331077
```

- e1071의 tune을 이용해 최적의 파라미터를 찾는다.

- 튜닝 결과 최적의 gamma는 0.01, cost는 13으로 도출되었다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '사망중상자여부'인 경우

4) 최적 파라미터 가지는 모델 생성

```
best_param <- summary(tuned)$best.parameters
best_param
```

A data.frame: 1 × 2

	gamma	cost
	<dbl>	<int>
127	0.01	13

```
# 9번
svm_best <- e1071::svm(사망중상자여부~., data = train_set, type = "C-classification", kernel="radial",
                       gamma=best_param[1,1], cost = best_param[1,2])

summary(svm_best)

print("svm_best:train 데이터 분류 결과")
table(predict(svm_best,train_set),train_set$사망중상자여부)

print("svm_best : train 데이터 confusionMatrix 결과")
cm <- caret::confusionMatrix(predict(svm_best, train_set), train_set$사망중상자여부)
cm

#  6,9번 Accuracy 비교 / 7, 10 Accuracy 비교
```

```
Call:
svm(formula = 사망중상자여부 ~ ., data = train_set, type = "C-classification",
    kernel = "radial", gamma = best_param[1, 1], cost = best_param[1,
        2])


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  13
```

- 최적 파라미터 gamma 0.01, cost 13을 이용해 svm 모델을 생성한다.

5) Confusion Matrix

```
cm_svm_train <- caret::confusionMatrix(predict(svm_best, test_set), test_set$사망중상자여부)
cm_svm_train
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 685 280
         1  24  46

               Accuracy : 0.7063
                 95% CI : (0.6775, 0.7339)
    No Information Rate : 0.685
    P-Value [Acc > NIR] : 0.07446

                  Kappa : 0.1361

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9661
            Specificity : 0.1411
         Pos Pred Value : 0.7098
         Neg Pred Value : 0.6571
             Prevalence : 0.6850
         Detection Rate : 0.6618
   Detection Prevalence : 0.9324
      Balanced Accuracy : 0.5536

       'Positive' Class : 0
```

- 생성한 SVM 모델로 test data에 대해 예측을 수행하고, confusion matrix를 계산한다.
- Sensitivity가 96%이고 Specificity가 14%로 튜닝 전에 비해 성능이 향상되었지만, 여전히 성능이 낮은 모델이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '경상부상자여부'인 경우

1) 모델 생성

```
svm_basic2 <- e1071::svm(formula = 경상부상자여부~., data = train_set2, type = "C-classification", kernel="radial")

summary(svm_basic2)

print("svm_basic:train 데이터 분류 결과")
table(predict(svm_basic2,train_set2),train_set2$경상부상자여부)

print("svm_basic : train 데이터 confusionMatrix 결과")
cm2 <- caret::confusionMatrix(predict(svm_basic2, train_set2), train_set2$경상부상자여부)
cm2
```

```
Call:
svm(formula = 경상부상자여부 ~ ., data = train_set2, type = "C-classification",
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1114

 ( 661 453 )


Number of Classes:  2
```

- SVM()을 이용해 모델을 생성한다.

2) Confusion Matrix

```
cm2 <- caret::confusionMatrix(predict(svm_basic2, test_set2), test_set2$경상부상자여부)
cm2
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0   0   0
         1 204 819

               Accuracy : 0.8006
                 95% CI : (0.7748, 0.8247)
    No Information Rate : 0.8006
    P-Value [Acc > NIR] : 0.5187

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.8006
             Prevalence : 0.1994
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

- 생성한 SVM 모델로 test data에 대해 예측을 수행하고, confusion matrix를 계산한다.
- Sensitivity가 0%이고 Specificity가 100%로 성능이 매우 낮은 모델이다

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '경상부상자여부'인 경우

3) 튜닝

```
tuned <- e1071::tune.svm(경상부상자여부~., data=train_set2, gamma = 10^(-8:1), cost= 1:30)
```

```
summary(tuned)
```

```
Parameter tuning of  'svm' :

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
   0.1    3

- best performance: 0.1738996

- Detailed performance results:
    gamma cost     error dispersion
1   1e-08    1 0.1742842 0.01820025
2   1e-07    1 0.1742842 0.01820025
3   1e-06    1 0.1742842 0.01820025
4   1e-05    1 0.1742842 0.01820025
5   1e-04    1 0.1742842 0.01820025
6   1e-03    1 0.1742842 0.01820025
7   1e-02    1 0.1742842 0.01820025
8   1e-01    1 0.1754366 0.02049485
9   1e+00    1 0.1792857 0.01874057
10  1e+01    1 0.1808271 0.01718193
11  1e-08    2 0.1742842 0.01820025
12  1e-07    2 0.1742842 0.01820025
13  1e-06    2 0.1742842 0.01820025
14  1e-05    2 0.1742842 0.01820025
15  1e-04    2 0.1742842 0.01820025
```

- e1071의 tune을 이용해 최적의 파라미터를 찾는다.

- 튜닝 결과 최적의 gamma는 0.1, cost는 3으로 도출되었다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

4. SVM– 종속변수가 '경상부상자여부'인 경우

4) 최적 파라미터 가지는 모델 생성

```
best_param2 <- summary(tuned)$best.parameters
best_param2
```

A data.frame: 1 × 2

	gamma	cost
	<dbl>	<int>
28	0.1	3

```
svm_best2 <- e1071::svm(경상부상자여부~., data = train_set2, type = "C-classification", kernel="radial", gamma=best_p

summary(svm_best2)

print("svm_best:train 데이터 분류 결과")
table(predict(svm_best2,train_set2),train_set2$경상부상자여부)

print("svm_best : train 데이터 confusionMatrix 결과")
cm2 <- caret::confusionMatrix(predict(svm_best2, train_set2), train_set2$경상부상자여부)
cm2
```

```
Call:
svm(formula = 경상부상자여부 ~ ., data = train_set2, type = "C-classification",
    kernel = "radial", gamma = best_param2[1, 1], cost = best_param2[1,
        2])


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  3

Number of Support Vectors:  1234

 ( 783 451 )
```

- 최적 파라미터 gamma 0.1, cost 3을 이용해 svm 모델을 생성한다.

5) Confusion Matrix

```
cm_svm_train2 <- caret::confusionMatrix(predict(svm_best2, test_set2), test_set2$경상부상자여부)
cm_svm_train2
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  23  19
         1 181 800

               Accuracy : 0.8045
                 95% CI : (0.7788, 0.8284)
    No Information Rate : 0.8006
    P-Value [Acc > NIR] : 0.3949

                  Kappa : 0.1276

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.11275
            Specificity : 0.97680
         Pos Pred Value : 0.54762
         Neg Pred Value : 0.81549
             Prevalence : 0.19941
         Detection Rate : 0.02248
   Detection Prevalence : 0.04106
      Balanced Accuracy : 0.54477

       'Positive' Class : 0
```

- 생성한 SVM 모델로 test data에 대해 예측을 수행하고, confusion matrix를 계산한다.
- Sensitivity가 11%이고 Specificity가 97%로 튜닝 전에 비해 성능이 향상되었지만, 여전히 성능이 낮은 모델이다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

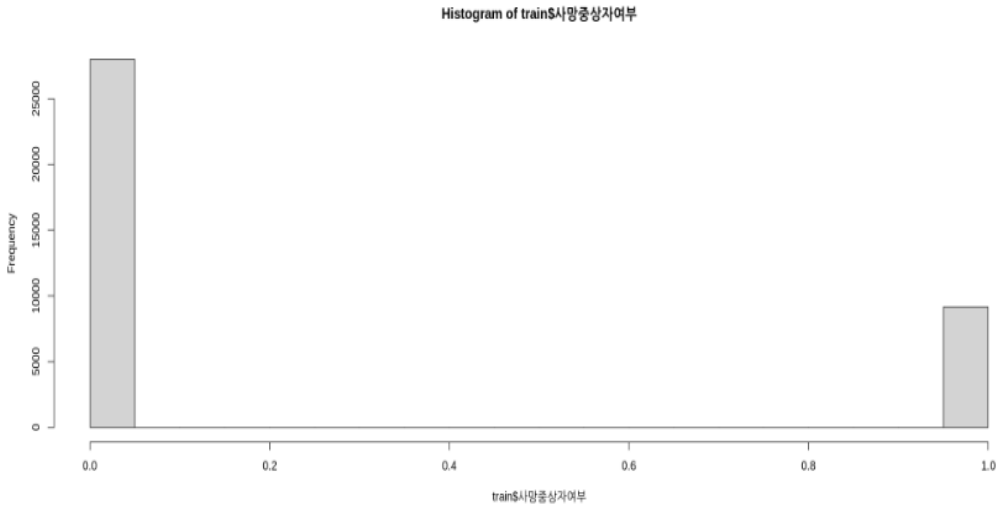
6

분석 결과 해석

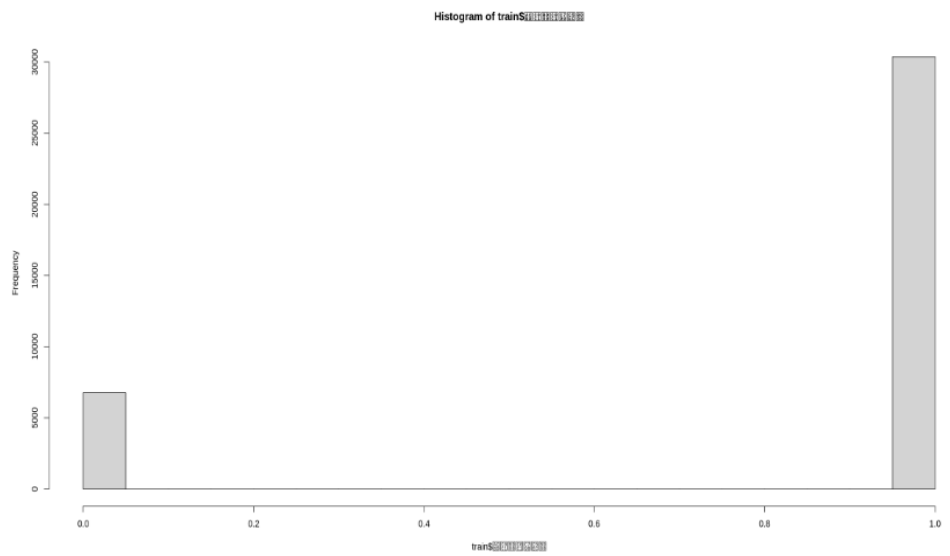
ⓒSaebyeol Yu. Saebyeol's PowerPoint

분석 결과 해석

```
hist(train$사망중상자여부)
```




```
hist(train$경상부상자여부)
```



- 로지스틱 회귀, 결정트리, Random forest, SVM 모두 종속변수가 '사망 중상자 여부' 인 경우에는 높은 민감도, 낮은 특이도를 보였다. 반면 종속변수가 '경상 부상자 여부 ' 인 경우에는 낮은 민감도, 높은 특이도를 보였다.

- 민감도가 높고 특이도가 낮다는 것은 양성 데이터는 잘 분류하지만 음성 데이터는 잘 분류하지 못한다는 것을 의미한다. 네 모델에서 이러한 결과가 나온 이유는, '사망 중상자 여부 ' 열과 '경상 부상자 여부' 열 데이터의 불균형때문이다. '사망 중상자 여부'의 경우 음성 데이터가 양성 데이터에 비해 훨씬 많다. 반면 '경상 부상자 여부'의 경우 양성 데이터가 음성 데이터에 비해 훨씬 많다.

- 때문에 이러한 데이터 양의 불균형 문제를 해결하려면 upSampling이나 down Sampling을 통해 데이터의 수를 맞추는 등의 과정으로 해결할 수 있을 것으로 예상한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

분석 결과 해석

- 가설을 설정할 때 했던 예상과 달리, 사망 중상자 여부와 가해 운전자 연령은 상관관계가 없었다. 하지만 사망 중상자 여부와 피해 운전자 연령은 연관이 있으므로, 고령의 운전자가 교통사고 발생 시 더욱 위험성이 높음을 알 수 있다.

- 카이제곱 검정을 이용한 가설 검정과 로지스틱 회귀의 모델에 대한 summary(), 결정 트리의 노드를 나누는 기준, Random Forest의 변수의 중요도 측정 과정을 통해 종속변수에 유의미한 영향력을 가지는 변수를 알 수 있다. 이를 통해 알게 된 중요도가 높은 독립 변수들은 도로 형태, 사고 유형,피해 운전자 차종, 법규가 있다.

- 이러한 결과를 데이터의 분석을 통해 교통사고의 피해를 줄이고자 했던 목적에 비추어 분석해 볼 수 있다. 종합적으로 사고 피해에 가장 큰 영향을 미치는 요인은 피해 운전자의 차종이었다. 때문에 개인적인 차원에서 교통사고의 피해를 줄이려면, 차량의 안정성을 차량 구매 시의 중요한 요인으로 정하는 것이 중요하다.

- 위반한 법규의 종류에 따라 피해의 정도에 영향이 크게 나타나는 것을 알 수 있다. 때문에 사회적 측면에서 교통사고 피해를 줄이려면, 어겼을 때 피해가 특히 큰 법규의 단속과 처벌을 강화하는 것으로 교통사고 피해를 감소시킬 수 있을 것이라고 예상한다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

감사합니다.

빅데이터 전공 20215123 김수연

ⓒSaebyeol Yu. Saebyeol's PowerPoint