



빅데이터 처리 (화요일 (1:3교시))

실습: 영화 흥해 예측 보고서 2차

2022.04.12



Institutor ; JS LEE

개요

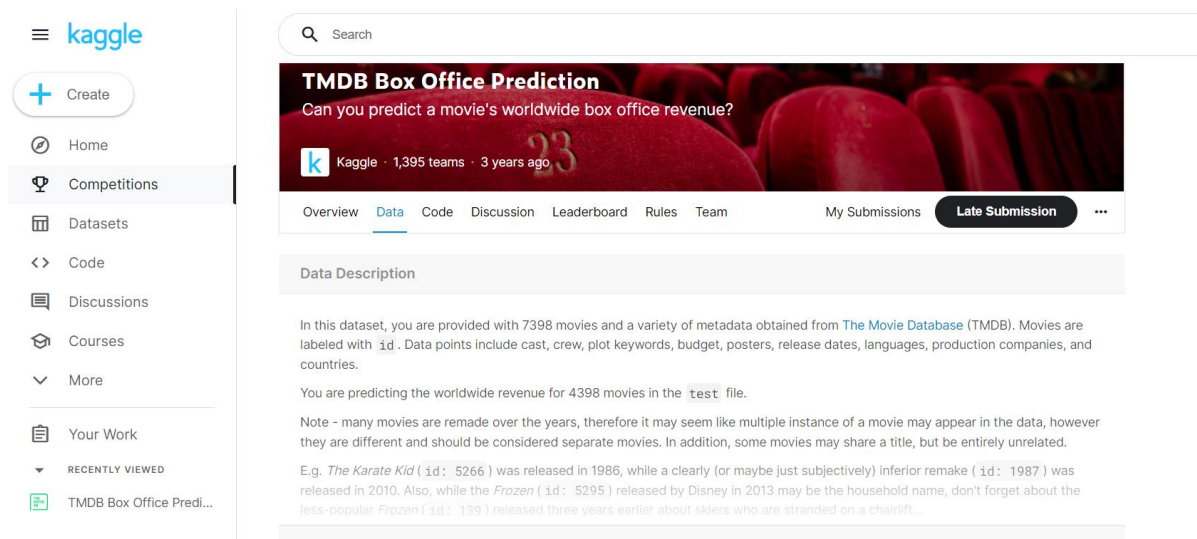
주제 : 분류 기법을 이용하여 영화 흥행 예측하기

- 영화 정보: 23개 변수와 revenue(label) . 2997건
- 흥행의 기준이 되는 매출액을 정하고 그에 따라 revenue 데이터를
1: 기준 초과/ 2: 기준 이하로 변경 (즉, 연속형 변수를 범주형 (factor) 변수화)
- 분류기법을 이용하여 영화의 흥행 예측 및 평가
 - 변수 선택, 파생 변수 생성, 모델 선정, 파라미터 설정 변경 등
 - accuracy, precision, recall, 등을 이용하여 모델 평가
- 모델링 과정과 모델링 결과, 평가 설명 등을 포함하여 보고서 제출
- 개인의 분석 Insight (분석 창의성)을 포함하여 보고서를 작성하여 주시기를 부탁드립니다.
- 7주차 진행할 R의 그래프 기능(ggplot등)을 추가하실 것을 권장합니다.
- 한글이나 워드 양식으로

기한 : ~ 5.10 까지

개요

- 분석 데이터 (train-box-office.csv)
 - 영화 데이터에 대한 자세한 내용은 아래 사이트를 참조한다.
 - <https://www.kaggle.com/c/tmdb-box-office-prediction/data>
 - <https://www.themoviedb.org/discover/movie>
 - Train (모델 구축용:3000건)와 적용 데이터 (4398건 : Revenue 변수 없음) 제공됨.
 - 관련 데이터 6주차 강의 폴더에 존재함.
 - 분석의 편의성의 위해 다소 수정작업 진행함



The screenshot shows the Kaggle interface for the 'TMDb Box Office Prediction' competition. The left sidebar contains navigation links: Home, Competitions, Datasets, Code, Discussions, Courses, and More. The main content area displays the competition title, a search bar, and tabs for Overview, Data, Code, Discussion, Leaderboard, Rules, Team, My Submissions, and Late Submission. The 'Data' tab is selected, showing a 'Data Description' section. The description states that the dataset includes 7398 movies with metadata from The Movie Database (TMDb), such as cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. It also mentions that the task is to predict the worldwide revenue for 4398 movies in the 'test' file. A note clarifies that many movies are remade over the years, so multiple instances of a movie may appear in the data. Examples are given: 'The Karate Kid' (id: 5266) released in 1986, its inferior remake (id: 1987) released in 2010, and 'Frozen' (id: 5295) released in 2013, compared to the less-popular 'Frozen' (id: 139) released in 2010.

개괄적 프로세스

1) 필요한 라이브러리 로딩

2) 데이터 준비

2-1)데이터 불러오기

2-2)데이터 일차적 탐색 및 변수 선택

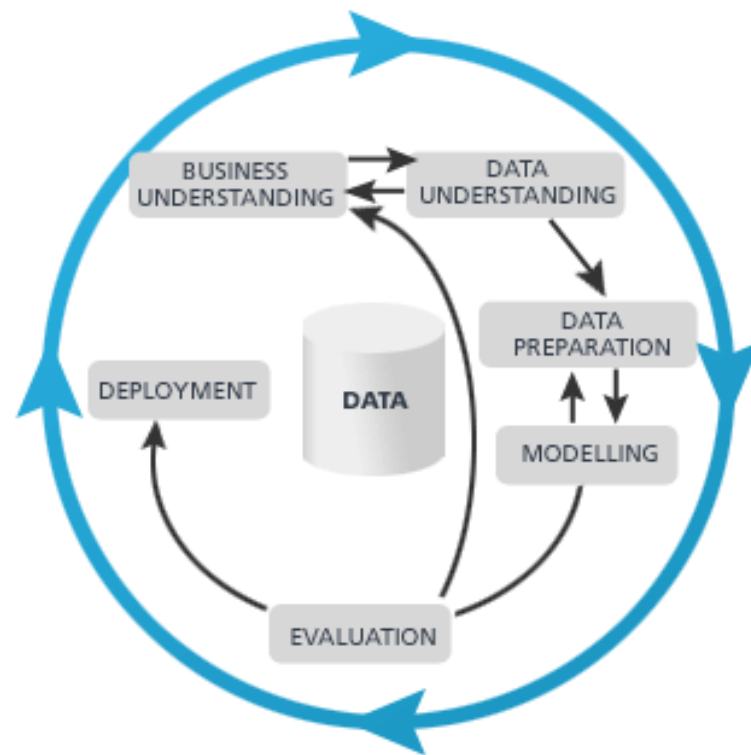
2-3) 파생 데이터 생성

예> 1. 타겟변수
2. 데이터 값이 텍스트 인 경우 다양한 방법으로 처리하여 구조화
등

3) 모델링

4) 모형평가

5) 최종 모델 선정 및 선정 사유



1) 필요한 라이브러리 로딩

7주차까지 접한 기본 패키지를 로딩>Loading) 하고, 그 외 각자가 더 필요한 패키지 설치 및 로딩 함

1.필요 라이브러리 로딩

#데이터 핸들링

```
library(readxl)  
library(dplyr)
```

기술통계량

```
library(psych)  
library(descr)
```

분류기법

```
library(rpart)  
library(rpart.plot)  
library(adabag)  
library(randomForest)  
library(caret)
```

시각화 (그래프로 표현)

```
library(ggplot2)
```



2) 데이터 준비

1. 구축용 모델 데이터 셋 불러오기

```
box_office <- read_excel("train_box_office.xlsx")
```

2. 데이터의 일차적 탐색 및 변수 선택하기

```
head(box_office)
```

```
#> # A tibble: 6 x 11
#>   id belongs_to_collection budget genres_code genres_desc homepage imdb_id original_language original_title overview popularity
#>   <dbl> <chr>                <dbl>    <dbl> <chr>      <chr>      <chr>      <chr>      <chr>      <chr>      <dbl>
#> 1     1 [{id: 313576, 'nam~ 1.4e7         35 Comedy    NA        tt2637~ en        Hot Tub Time ~ "when L~      6.58
#> 2     2 [{id: 107674, 'nam~ 4 e7          35 Comedy    NA        tt0368~ en        The Princess ~ "Mia Th~      8.25
#> 3     3 NA                    3.3e6         18 Drama      http://~ tt2582~ en        whiplash      "Under ~      64.3
#> 4     4 NA                    1.2e6         53 Thriller   http://~ tt1821~ hi        Kahaani       "Vidya ~      3.17
#> 5     5 NA                    0            28 Action     NA        tt1380~ ko        마린보이      "Marine~      1.15
#> 6     6 NA                    8 e6          16 Animation  NA        tt0093~ en        Pinocchio and~ "Pinocc~      0.743
#> # ... with 13 more variables: poster_path <chr>, production_companies <chr>, production_countries <chr>, release_date <chr>,
#> # runtime <dbl>, spoken_languages <chr>, status <chr>, tagline <chr>, title <chr>, Keywords <chr>, cast <chr>, crew <chr>,
#> # revenue <dbl>
```

```
tail(box_office)
```

```
#> # A tibble: 6 x 11
#>   id belongs_to_collection budget genres_code genres_desc homepage imdb_id original_language original_title overview popularity
#>   <dbl> <chr>                <dbl>    <dbl> <chr>      <chr>      <chr>      <chr>      <chr>      <dbl>
#> 2995 NA                    0            18 Drama      NA        tt0105~ en        School Ties    When Da~      7.44
#> 2996 NA                    0            35 Comedy    NA        tt0109~ en        Chasers        Militar~      9.85
#> 2997 NA                    0            18 Drama      NA        tt2364~ sv        Vi ar bast!    Three g~      3.73
#> 2998 NA                    6.5e7         80 Crime     NA        tt0116~ en        The Long Kiss~ Samanth~     14.5
#> 2999 NA                    4.2e7         35 Comedy    http://~ tt0343~ en        Along Came Po~ Reuben ~     15.7
#> 3000 NA                    3.5e7         53 Thriller   http://~ tt1600~ en        Abduction      A young~     10.5
#> .. with 13 more variables: poster_path <chr>, production_companies <chr>, production_countries <chr>, release_date <chr>,
#> runtime <dbl>, spoken_languages <chr>, status <chr>, tagline <chr>, title <chr>, Keywords <chr>, cast <chr>, crew <chr>,
#> revenue <dbl>
```

2) 데이터 준비

2) 데이터 일차적 탐색 및 변수 선택

View(box_office)

| id | belongs_to_collection | budget | genres_code | genres_desc | homepage | imdb_id | original_language | original_title | overview | popularity | poster_pat |
|----|--|-----------|-------------|-------------|---------------------------------------|-----------|-------------------|--|---|------------|------------|
| 4 | NA | 1200000 | 53 | Thriller | http://kahaanithemovie.com/ | tt1821480 | hi | Kahaani | Vidya Bagchi (Vidya Balan) arrives in Kolkata from London t... | 3.174936 | /aTXRaPrV |
| 5 | NA | 0 | 28 | Action | NA | tt1380152 | ko | 마린보이 | Marine Boy is the story of a former national swimmer who f... | 1.148070 | /m2z7zvkl |
| 6 | NA | 8000000 | 16 | Animation | NA | tt0393743 | en | Pinocchio and the Emperor of the Night | Pinocchio and his friends, a glow worm and a marionette, se... | 0.743274 | /6IDQa1D; |
| 7 | NA | 14000000 | 27 | Horror | http://www.thepossessionmovie.com/ | tt0431021 | en | The Possession | A young girl buys an antique box at a yard sale, unaware th... | 7.286477 | /AQzFuaZ; |
| 8 | NA | 0 | 99 | Documentary | NA | tt0391024 | en | Control Room | A chronicle which provides a rare window into the internatio... | 1.949044 | /83BV8Gy |
| 9 | [[{"id": 256377, "name": "The Muppet Collection", "poster_path": "... | 0 | 28 | Action | NA | tt0117110 | en | Muppet Treasure Island | After telling the story of Flint's last journey to young Jim Ha... | 6.902423 | /5A8gKzO; |
| 10 | NA | 6000000 | 35 | Comedy | NA | tt0310281 | en | A Mighty Wind | In "A Mighty Wind", director Christopher Guest reunites the ... | 4.672036 | /xwO4ESpC |
| 11 | [[{"id": 1575, "name": "Rocky Collection", "poster_path": "/mCY5... | 1000000 | 18 | Drama | NA | tt0075148 | en | Rocky | When world heavyweight boxing champion, Apollo Creed w... | 14.774066 | /j5kiwd5sn |
| 12 | [[{"id": 48190, "name": "Revenge of the Nerds Collection", "post... | 0 | 35 | Comedy | NA | tt0093857 | en | Revenge of the Nerds II: Nerds in Paradise | The members of the Lambda Lambda Lambda fraternity trav... | 10.543750 | /1KT0wJQ |
| 13 | NA | 15000000 | 18 | Drama | http://www.dreamworks.com/ab/ | tt0168547 | en | American Beauty | Lester Burnham, a depressed suburban father in a mid-life c... | 20.726578 | /or1MP8B; |
| 14 | [[{"id": 91698, "name": "Chili Palmer Collection", "poster_path": "... | 53000000 | 35 | Comedy | NA | tt0377471 | en | Be Cool | Disenchanted with the movie industry, Chili Palmer tries the ... | 13.314233 | /ekKCH7Z; |
| 15 | NA | 102000000 | 28 | Action | NA | tt0181689 | en | Minority Report | John Anderton is a top 'Precrime' cop in the late-21st centur... | 20.666063 | /h3ip15n7 |
| 16 | NA | 500000 | 28 | Action | http://skinningmovie.com/synopsis.htm | tt1128437 | sr | Šćanje | Novica is a mathematics champion in a Belgrade high schoo... | 1.018477 | /88l8VBan |
| 17 | NA | 26000000 | 27 | Horror | NA | tt0421239 | en | Red Eye | After attending the funeral of her grandmother in Dallas, th... | 6.336927 | /7Td8Ald |
| 18 | NA | 0 | 18 | Drama | NA | tt1700845 | en | The Invisible Woman | In 1857, at the height of his fame and fortune, novelist and s... | 4.183558 | /clMdqrF; |
| 19 | NA | 8000000 | 35 | Comedy | NA | tt1487118 | en | Chalet Girl | While working a job at an exclusive ski resort to support her... | 6.496259 | /64AQz8E; |
| 20 | [[{"id": 9518, "name": "The Transporter Collection", "poster_path": "... | 32000000 | 28 | Action | NA | tt0388482 | en | Transporter 2 | Professional driver Frank Martin is living in Miami, where he ... | 11.359659 | /bZK0wFQ; |
| 21 | NA | 80000000 | 12 | Adventure | NA | tt0120738 | en | Lost in Space | The prospects for continuing life on Earth in the year 2058 a... | 12.000579 | /kk8DPKUi |
| 22 | [[{"id": 9735, "name": "Friday the 13th Collection", "poster_path": "... | 4000000 | 27 | Horror | NA | tt0083972 | en | Friday the 13th Part III | An idyllic summer turns into a nightmare of unspeakable ter... | 7.992290 | /5wg2NZy |
| 23 | [[{"id": 207621, "name": "V/H/S Collection", "poster_path": "/esf6... | 0 | 53 | Thriller | http://www.magnetreleasing.com/vhs/ | tt2105044 | en | V/H/S | When a group of misfits is hired by an unknown third party ... | 7.820787 | /y049x6Jhi |
| 24 | NA | 0 | 53 | Thriller | http://insightthemovie.com/ | tt1687277 | en | InSight | Kaitlyn, an ER nurse who is tending to a young stabbing vict... | 1.323333 | /7oKsA5B |
| 25 | NA | 0 | 27 | Horror | http://www.blacksheep-themovie.com/ | tt0779982 | en | Black Sheep | A genetic engineering experiment gone horribly awry turns ... | 7.434577 | /8y0L8W5I |
| 26 | NA | 0 | 35 | Comedy | NA | tt0104139 | en | Dr. Giggles | In 1937, Evan Rendell flees after his father is lynched for kill... | 1.252367 | /5JDLNXW |
| 27 | NA | 10000000 | 80 | Crime | NA | tt0120176 | en | The Spanish Prisoner | An employee of a corporation with a lucrative secret proces... | 4.305735 | /cwKWWf |
| 28 | NA | 11000000 | 18 | Drama | NA | tt1486634 | en | What If | Wallace, a medical school dropout, has been repeatedly bur... | 10.841891 | /g5vP8Cj; |
| 29 | NA | 6000000 | 16 | Animation | NA | tt0092106 | en | The Transformers: The Movie | The Autobots must stop a colossal planet-consuming robot ... | 6.759181 | /yGuQ2iz4 |
| 30 | NA | 45000000 | 28 | Action | NA | tt0264472 | en | Changing Lanes | A rush-hour fender-bender on New York City's crowded FD... | 7.818620 | /dCTVH1K; |

이 데이터 셋을 분석하기 위해 고려해야 할 점은?

결측치가 있는 변수들이 존재한다.

데이터 값이 텍스트로 이루어진 변수들이 존재한다. => 어떻게 구조화 할 것인가?

2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

| 변수명 | 사용여부 | 이유 | 변수명 | 사용여부 | 이유 |
|-----------------------|------|---------------------|----------------------|------|--------------------------|
| id | Yes | id 변수 | production_companies | No | 구조화가 힘들 |
| belongs_to_collection | No | 결측치 | production_countries | Yes | 구조화 가능 (동일위치에 나라명 존재) |
| budget | Yes | 예산 | release_date | No | 오류 데이터 |
| genres_code | Yes | 장르 코드 | runtime | Yes | 상영시간 |
| genres_desc | Yes | 장르 | spoken_languages | | 구조화가 힘들 |
| homepage | No | 결측치 | status | No | 데이터가 하나의 값의 비중이 너무 큼 |
| imdb_id | No | 의미 없음 | tagline | No | 결측치 |
| original_language | Yes | | title | No | 의미 도출이 어려움 |
| original_title | No | 구조화도 힘들며 의미 도출이 어려움 | Keywords | No | 결측치 |
| overview | No | 구조화도 힘들며 의미 도출이 어려움 | cast | No | 구조화도 힘들며 의미 도출이 어려움 |
| popularity | Yes | | crew | No | 구조화도 힘들며 의미 도출이 어려움 |
| poster_path | No | 구조화도 힘들며 의미 도출이 어려움 | revenue | Yes | 종속변수 (연속형 변수 이므로 범주화 필요) |

2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

- 변수별 결측치가 어느정도 인지 확인

건수 상이할 수 있음

```
colSums(is.na(box_office))
```

| | | | | | | |
|-------------------|-----------------------|------------------|-------------|-------------|----------------------|----------------------|
| id | belongs_to_collection | budget | genres_code | genres_desc | homepage | imdb_id |
| 0 | 2393 | 0 | 7 | 7 | 2053 | 0 |
| original_language | original_title | overview | popularity | poster_path | production_companies | production_countries |
| 0 | 0 | 8 | 0 | 1 | 155 | 55 |
| release_date | runtime | spoken_languages | status | tagline | title | keywords |
| 0 | 2 | 20 | 0 | 596 | 0 | 275 |
| cast | crew | revenue | | | | |
| 13 | 16 | 0 | | | | |

10% 정도가 na이면 제외 (각자 판단)

```
box_office1 <- box_office[, -c(2,6,19,21)]
```

*2,6,19,21,번째 컬럼 제외

- 모델링에 따라 결측치가 있는 경우 모델링이 안되는 경우 존재
- (예> RandomForest)
- 모델링에 따라 결측치가 있는 행은 제외하거나 대체해야 할 필요성 존재

2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

각자 가용성의 기준을 설정 후 변수 사용여부 결정

구조화된 Character 변수들은 freq()함수를 이용하여 빈도를 살펴본 후 사용여부 결정

구조화되지 않은 Character 변수들은 구조화 가능성을 판단하고 구조화가 되지 않으면 제외

```
box_office$status
  Frequency Percent
Released    2993   99.8665
Rumored       4    0.1335
Total       2997 100.0000
```

```
> freq(box_office$original_language)
```

```
box_office$original_language
  Frequency Percent
ar           1   0.03337
bn           1   0.03337
cn          20   0.66733
cs           1   0.03337
da           5   0.16683
de          18   0.60060
el           1   0.03337
en         2573  85.85252
es          42   1.40140
fa           5   0.16683
fi           2   0.06673
fr          78   2.60260
he           1   0.03337
hi          42   1.40140
hu           3   0.10010
id           1   0.03337
it          24   0.80080
ja          37   1.23457
ko          20   0.66733
ml           2   0.06673
mr           1   0.03337
```

2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

텍스트의 구조화

Production_Countries

```
[{'iso_3166_1': 'SE', 'name': 'Sweden'}]
```

```
[{'iso_3166_1': 'CN', 'name': 'China'}, {'iso_3166_1': 'HK', 'name': 'Hong Kong'}]
```

```
[{'iso_3166_1': 'FR', 'name': 'France'}, {'iso_3166_1': 'US', 'name': 'United States'}]
```

18,19위치에 국가명이 존재함 => 18,19 위치의 텍스트만 추출하여 새로운 변수로 생성하기

```
tbox_office2$production_countries1 <- substr(box_office2$production_countries,18,19)
```

2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

Revenue 변수 범주화 하기

```

      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
1.000e+00 2.400e+06 1.681e+07 6.678e+07 6.893e+07 1.520e+09
> describe(box_office2$revenue)
   vars  n   mean    sd  median trimmed   mad min      max   range skew kurtosis   se
X1      1 2997 66778537 137589660 16810383 35130206 24448056 1 1519557910 1519557909 4.54   27.69 2513289
> box_office2$revenue1 <- ifelse(box_office2$revenue >=16810383, "high","Low")
> box_office2$revenue1 <- as.factor(box_office2$revenue1)
> freq(box_office2$revenue1)
box_office2$revenue1

```

Revenue는 연속형 변수 = 범주형(이진형)으로 바꾸기 위한
기준을 잡기

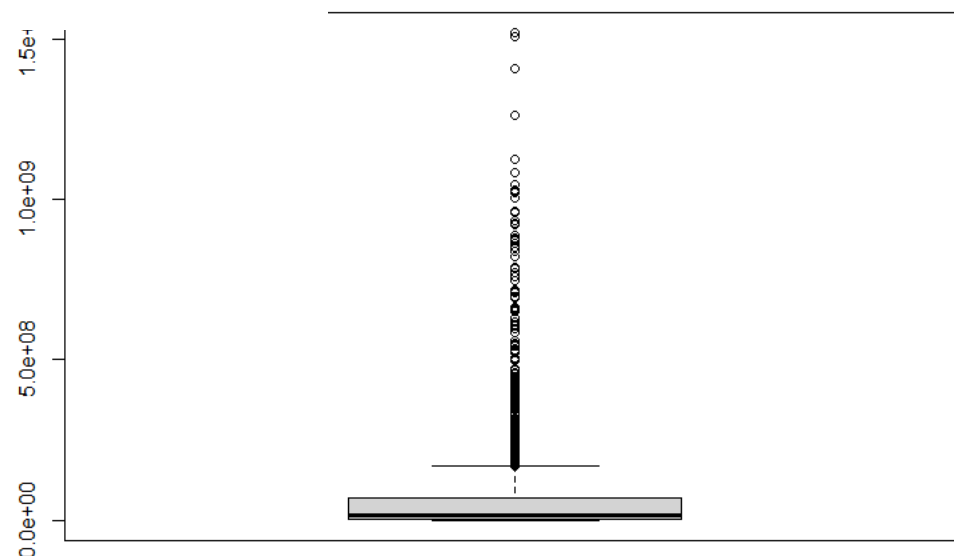
예1>상위 25%

예2> 평균값

예3> 중앙값

...

분석가의 판단에 의해 분류



2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

필요함수

ifelse (조건, True일 때 리턴할 값, False일 때 리턴할 값)

중앙값을 기준으로 분류

```
box_office2$revenue1 <- ifelse(box_office2$revenue >=16810383, "high", "Low")
box_office2$revenue1 <- as.factor(box_office2$revenue1)
```

새로운 종속변수가 분포 정도 살피기

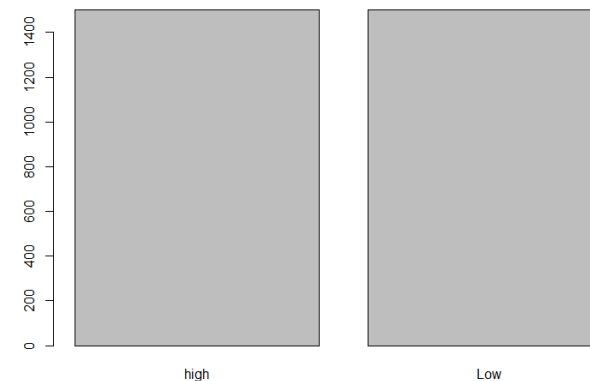
```
freq(box_office2$revenue1)
```

```
> freq(box_office2$revenue1)
```

```
box_office2$revenue1
```

| | Frequency | Percent |
|-------|-----------|---------|
| high | 1499 | 50.02 |
| Low | 1498 | 49.98 |
| Total | 2997 | 100.00 |

```
> |
```



2) 데이터 준비

2) . 데이터 셋 탐색하고 분석에 사용할 변수 선택하기

최종적으로 데이터 셋 정리 : box_office3완성

(분석에서 사용하지 않는 변수 제외 (예> id,evenue, production_coutries))

```

tibble [2,997 × 7] (S3: tbl_df/tbl/data.frame)
 $ budget      : num [1:2997] 0.00 0.00 1.02e+08 5.00e+05 3.20e+07 4.50e+07 1.40e+08 3.00e+07 5.00e+07 0.00 ...
 $ genres_desc : chr [1:2997] "Action" "Action" "Action" "Action" ...
 $ original_language : chr [1:2997] "ko" "en" "en" "sr" ...
 $ popularity    : num [1:2997] 1.15 6.9 20.67 1.02 11.36 ...
 $ runtime       : num [1:2997] 118 100 145 97 87 89 91 94 127 118 ...
 $ production_countries1: chr [1:2997] "KR" "US" "US" "RS" ...
 $ revenue1      : Factor w/ 2 levels "high","Low": 2 1 1 2 1 1 1 1 1 2 ...
> |

```

- id는 식별자 이므로 분석에는 사용하지 않음
- 모델링 특성에 따라, chr타입을 factor변경할 수도 있음
- ctree 함수는 character타입 사용 못 함

3)모델링 (모형구축)/ 평가/ 최종 선정

1

분석 데이터 셋을 모형 구축용과 모형 평가용으로 구분함

각자의 기준에 따라 구축용으로 사용할 비중과 평가용 비중을 정의함

2

모형 구축용 데이터 셋을 기준으로 분류 기법을 이용하여 모형 구축

(단독 모델링: raprt, ctree, 앙상블 모델링(RF개)

* 앙상블모델인 bagging 과 adaboodting은 시간이 오래 걸림

모델링의 변화 뿐 아니라, 변수의 조합을 상이하게 하여 모형을 구축하는 것을 권장함

3

모형평가용 데이터셋에 구축한 모형 3개 평가

: 평가 기준으로 3개의 모델별로 비교

4

최종 모델 선정한 기준 및 모형에 대한 설명 기술 (주요 분류 기준 , 모형 선정기준 등)

3)모델링 (모형구축)/ 평가/ 최종 선정

1

분석 데이터 셋을 모형 구축용과 모형 평가용으로 구분함

각자의 기준에 따라 구축용으로 사용할 비중과 평가용 비중을 정의함

1) 분석 데이터 셋을 모형 구축용과 모형 평가용으로 구분함

createDataPartition()또는 sample()이용

70% (구축용) , 30%(평가용)으로 하여, 데이터 셋을 구분함
70%, 30%에 준하게 추출됨.

```
# 데이터 셋 구분
# createPartion() 이용
box_row_idx <- createDataPartition(box_office3$revenue1, p=0.7, list=FALSE)
train <- box_office3[box_row_idx,]

table(train$revenue1)

# sample() 이용
ind <- sample(2, nrow(box_office3), replace=TRUE, prob=c(0.7, 0.3))

trainData <- box_office3[ind==1, ]
testData <- box_office3[ind==2, ]
```

```
> table(train$revenue1)
high Low
1050 1050
> table(trainData$revenue1)
high Low
1042 1066
.
```


3)모델링 (모형구축)/ 평가/ 최종 선정

2

모형 구축용 데이터 셋을 기준으로 분류 기법을 이용하여 모델 구축

(단독 모델링:2개, 앙상블 모델링(3개) 중 단독 모델링 1, 앙상블 모델링 2개 적용
모델링의 변화 뿐 아니라, 변수의 조합을 상이하게 하여 모형을 구축하는 것을 권장함

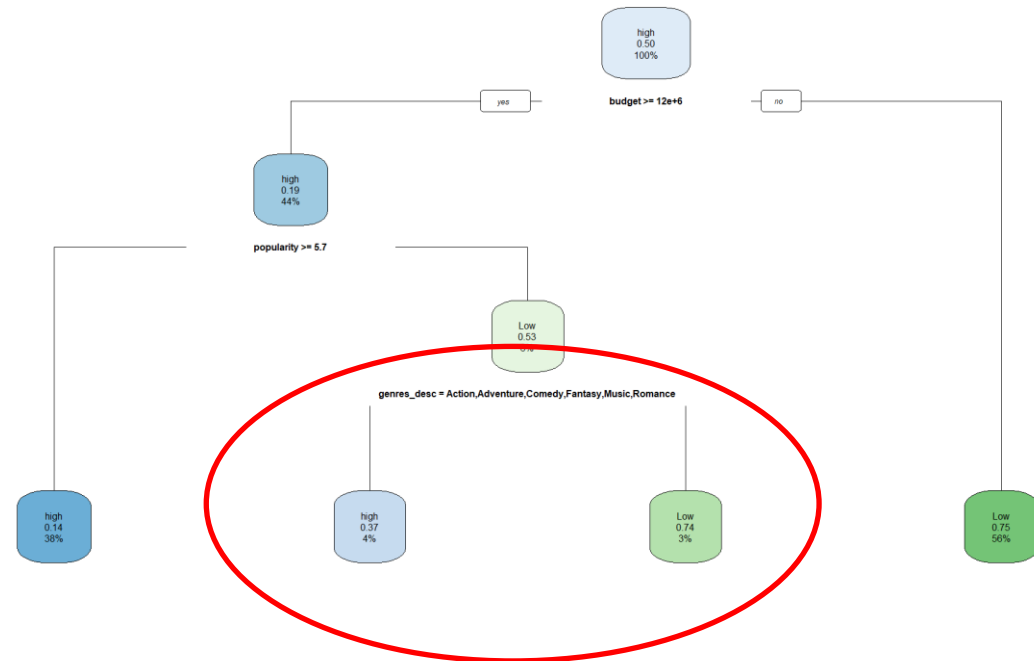
1) rpart()

```
result1 <- rpart(revenue1 ~., data=train, control=rpart.control(minsplit = 2))
result1
rpart.plot(result1)
```

n= 2100

```
node), split, n, loss, yval, (yprob)
* denotes terminal node
```

```
1) root 2100 1050 high (0.5000000 0.5000000)
 2) budget>=1.095e+07 962 194 high (0.7983368 0.2016632)
   4) popularity>=5.230208 851 128 high (0.8495887 0.1504113) *
   5) popularity< 5.230208 111 45 Low (0.4054054 0.5945946) *
 3) budget< 1.095e+07 1138 282 Low (0.2478032 0.7521968) *
```



3)모델링 (모형구축)/ 평가/ 최종 선정

2

모형 구축용 데이터 셋을 기준으로 분류 기법을 이용하여 모델 구축

(단독 모델링:2개, 앙상블 모델링(3개) 중 단독 모델링 1, 앙상블 모델링 2개 적용
모델링의 변화 뿐 아니라, 변수의 조합을 상이하게 하여 모형을 구축하는 것을 권장함

1) rpart()

result\$cptable

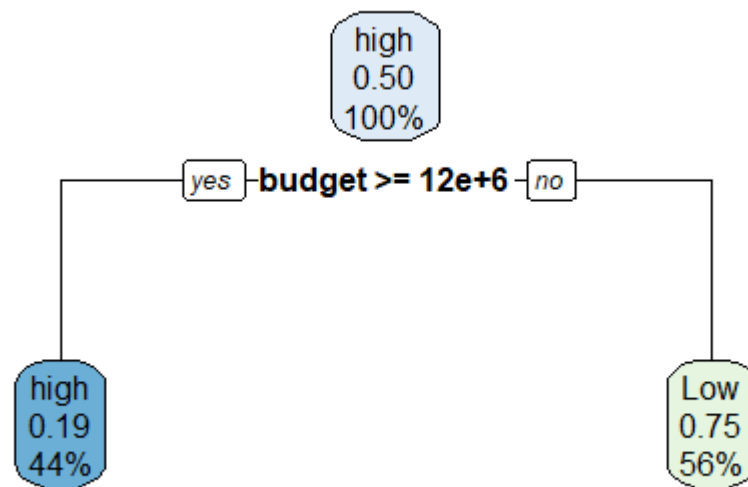
```

      CP  nsplit rel error   xerror   xstd
1 0.55714286    0 1.0000000 1.0361905 0.02180749
2 0.01380952    1 0.4428571 0.4580952 0.01833990
3 0.01000000    3 0.4152381 0.4428571 0.01812118
> |

```

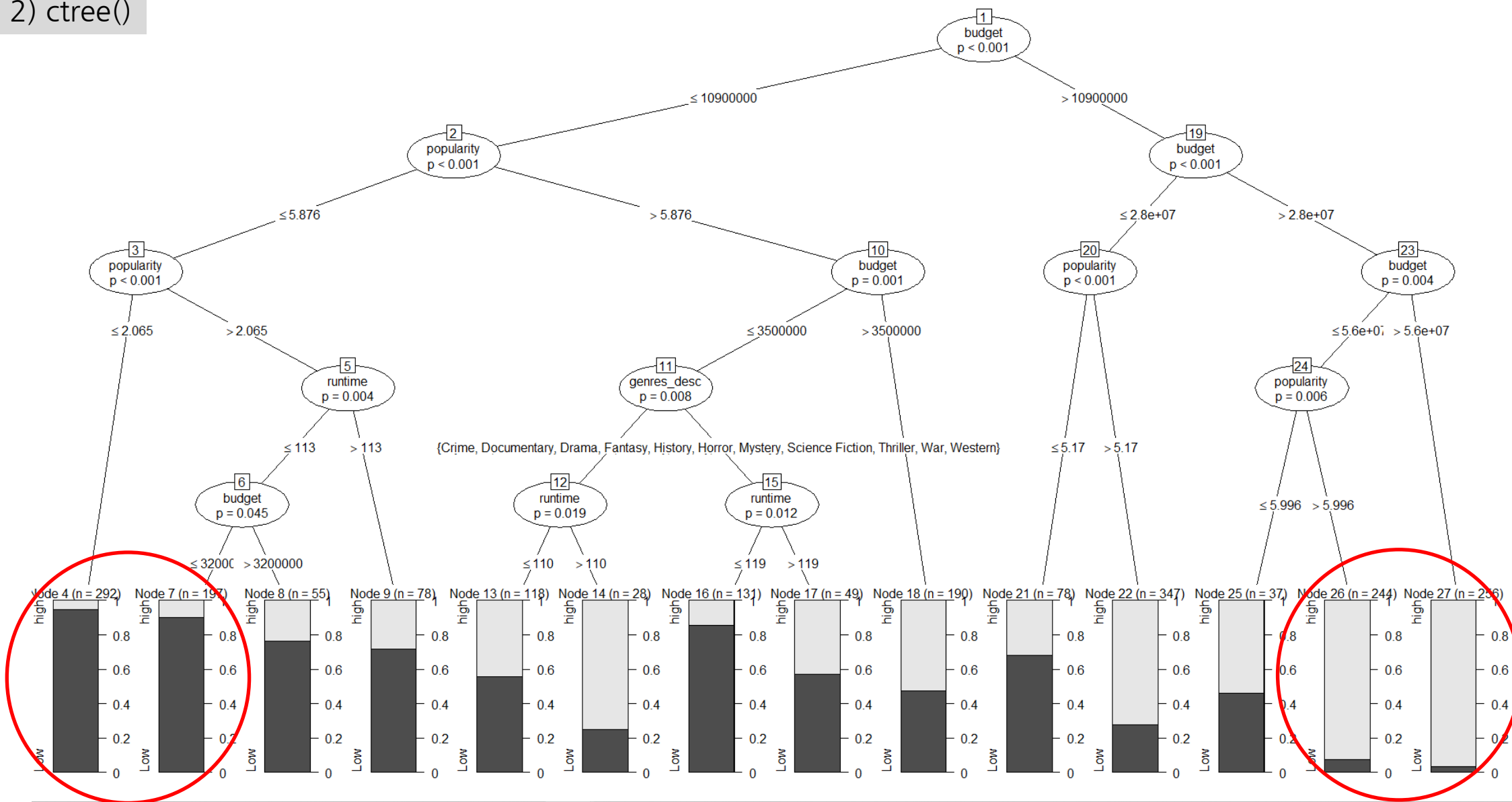
```
prune <- prune(result1,cp= 0.02)
```

```
rpart.plot(prune)
```



3)모델링 (모형구축)/ 평가/ 최종 선정

2) ctree()



3)모델링 (모형구축)/ 평가/ 최종 선정

2) RandomForest

결측 시 존재 시 모델링 에러 발생

```
> rf_result <- randomForest(revenue1 ~ ., data=train, ntree=100)
Error in na.fail.default(list(revenue1 = c(2L, 2L, 1L, 1L, 1L, 1L, 1L, :
missing values in object
```

| | budget | genres_desc | original_language | popularity | runtime |
|-----------------------|--------|-------------|-------------------|------------|---------|
| | 0 | 1 | 0 | 0 | 1 |
| production_countries1 | 43 | revenue1 | | | |
| | | 0 | | | |

* NA가 존재하는 행은 없앰 * 모형평가용 데이터 셋도 NA 제외해야 함

```
rf_train <- na.omit(train)
```

3)모델링 (모형구축)/ 평가/ 최종 선정

RandomForest

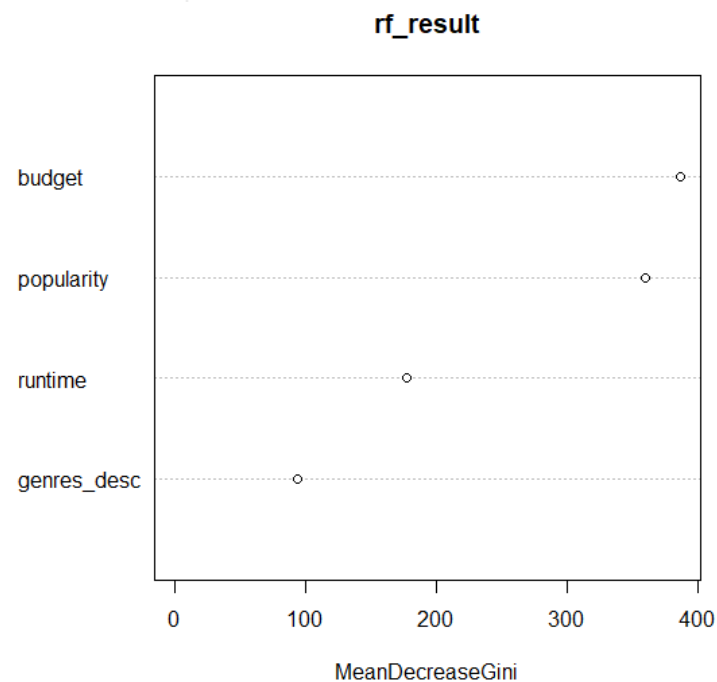
범주가 너무 많으면 수행 안됨 (53개 이하이어야 함)

* 범주가 많은 변수 제외하고 모델링

```
rf_train <- na.omit(train)
rf_result <- randomForest(revenue1 ~budget+genres_desc+popularity+ runtime, data=rf_train, ntree=100)
```

```
# 변수의 중요도
importance(rf_result)
varImpPlot(rf_result)
```

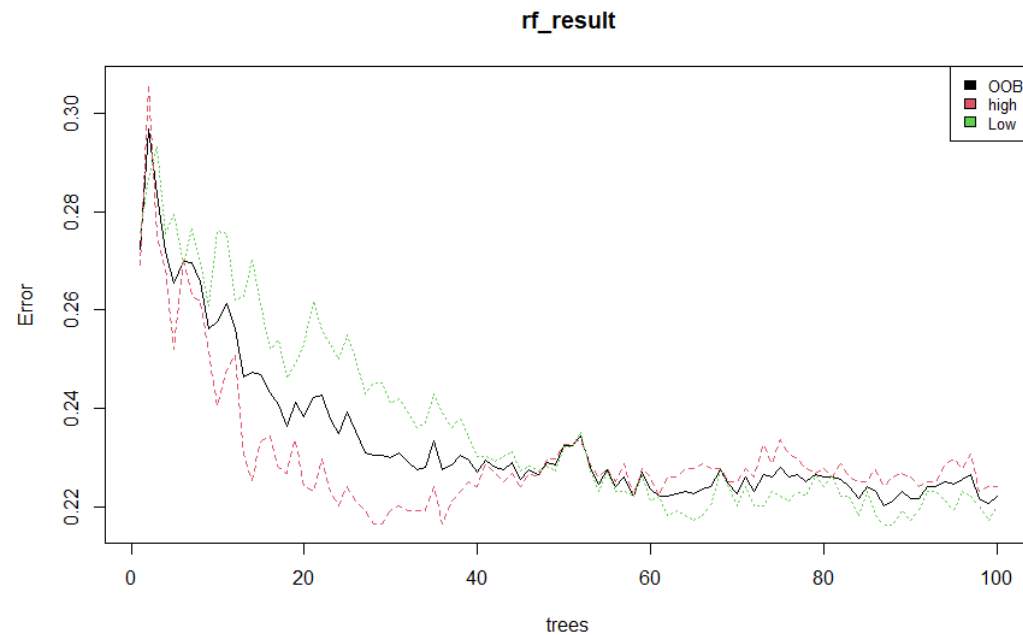
| | MeanDecreaseGini |
|-------------|------------------|
| budget | 386.74970 |
| genres_desc | 94.19774 |
| popularity | 360.22513 |
| runtime | 177.38448 |



3)모델링 (모형구축)/ 평가/ 최종 선정

RandomForest

```
plot(rf_result)  
legend("topright",colnames(rf_result$err.rate),cex=0.8,fill=1:3)
```



3)모델링 (모형구축)/ 평가/ 최종 선정

3

모형평가용 데이터셋에 구축한 모델 3개 평가

: 평가 기준으로 3개의 모델별로 비교

rpart

```
expect <- predict(prune, test, type="class")
actual <- test$revenue1

# 3. 혼동행렬을 이용하여 모형평가 하기

confusionMatrix(expect,actual,mode="everything")
```

```
Reference
Prediction high Low
high 340 103
Low 109 346

Accuracy : 0.7639
95% CI : (0.7347, 0.7
No Information Rate : 0.5
P-Value [Acc > NIR] : <2e-16

Kappa : 0.5278

McNemar's Test P-Value : 0.7313

Sensitivity : 0.7572
Specificity : 0.7706
Pos Pred Value : 0.7675
Neg Pred Value : 0.7604
Precision : 0.7675
Recall : 0.7572
F1 : 0.7623
Prevalence : 0.5000
Detection Rate : 0.3786
Detection Prevalence : 0.4933
Balanced Accuracy : 0.7639

'Positive' Class : high
```


3)모델링 (모형구축)/ 평가/ 최종 선정

3

모형평가용 데이터셋에 구축한 모델 3개 평가

: 평가 기준으로 3개의 모델별로 비교

ctee

```
expect1 <- predict(ctree_result, test1, type="response")
confusionMatrix(expect1,actual,mode="everything")
```

```

Reference
Prediction high Low
high 201 172
Low 248 277

Accuracy : 0.5323
95% CI : (0.499, 0.5653)
No Information Rate : 0.5
P-Value [Acc > NIR] : 0.0285479

Kappa : 0.0646

McNemar's Test P-Value : 0.0002526

Sensitivity : 0.4477
Specificity : 0.6169
Pos Pred Value : 0.5389
Neg Pred Value : 0.5276
Precision : 0.5389
Recall : 0.4477
F1 : 0.4891
Prevalence : 0.5000
Detection Rate : 0.2238
Detection Prevalence : 0.4154
Balanced Accuracy : 0.5323

'Positive' Class : high
```

3)모델링 (모형구축)/ 평가/ 최종 선정

3

모형평가용 데이터셋에 구축한 모델 3개 평가

: 평가 기준으로 3개의 모델별로 비교

RF

```
predict <- predict(rf_result, rf_test)
confusionMatrix(predict, rf_test$revenue1, mode = 'everything')
```

```

              Reference
Prediction high Low
high  422  28
Low   27  405

Accuracy : 0.9376
 95% CI : (0.9196, 0.9527)
No Information Rate : 0.5091
P-Value [Acc > NIR] : <2e-16

Kappa : 0.8752

McNemar's Test P-Value : 1

Sensitivity : 0.9399
Specificity : 0.9353
Pos Pred Value : 0.9378
Neg Pred Value : 0.9375
Precision : 0.9378
Recall : 0.9399
F1 : 0.9388
Prevalence : 0.5091
Detection Rate : 0.4785
Detection Prevalence : 0.5102
Balanced Accuracy : 0.9376

'Positive' Class : high
```

3)모델링 (모형구축)/ 평가/ 최종 선정

생성한 모델별 모형평가

| 알고리즘 | 사용변수 | 기준1 (accuracy) | 기준2 (F1) | 기준3.... |
|--------------|------|----------------|----------|---------|
| Rpart | | | | |
| Ctree | | | | |
| Randomforest | | | | |