어떤 영화가 흥행할까?

2017-2

김순영 김어진 김응채 배현주

목차

1 연구 배경

2 데이터

수집과 전처리 과정

3 통계 분석

1) logistic 회귀 2) knn
3) lda, qda 4) decision tree

4 최종 결론

연구 배경

수많은 영화 중, 흥행하는 영화는 오직 소수

“어떤 영화가 흥행할까?”

데이터 설명

66 개 국가

5043 개 영화

27 개의 변수

IMDB 실제 이용자 중심 데이터

출처: kaggle.com

데이터 전처리

변수 선택

num_critic_for_reviews
director_facebook_likes
cast_total_facebook_likes
facenumber_in_poster
num_user_for_reviews
budget
imdb_score
movie_facebook_likes

actor_2_name
actor_1_name
actor_3_name
director_name
content_rating
genres
movie_title
plot_keywords
language
country
title_year
aspect_ratio
actor_3_facebook_likes
actor_1_facebook_likes
actor_2_facebook_likes
Color
movie_imdb_link

데이터 전처리

결측치 처리

gross, budget에 NA 존재

"mice" 패키지로 NA imputation

```
> summary(movie.n)
 num_critic_for_reviews    duration     director_facebook_likes     gross           cast_total_facebook_likes
 Min.   :  1.0          Min.   : 14.0   Min.   :    0.0         Min.   :      162   Min.   :     0
 1st Qu.: 66.0          1st Qu.: 93.0   1st Qu.:    7.0         1st Qu.:  1771133   1st Qu.:  1458
 Median :139.0          Median :103.0   Median :   48.0         Median : 18320696   Median :  3412
 Mean   :164.6          Mean   :106.5   Mean   :  569.4         Mean   : 42933315   Mean   : 10780
 3rd Qu.:230.0          3rd Qu.:116.0   3rd Qu.:  187.0         3rd Qu.: 53182670   3rd Qu.: 15046
 Max.   :813.0          Max.   :300.0   Max.   :23000.0         Max.   :760505847   Max.   :656730
 facenumber_in_poster num_user_for_reviews     budget            imdb_score    movie_facebook_likes
 Min.   : 0.000       Min.   :   1.0       Min.   :2.180e+02   Min.   :1.600   Min.   :     0
 1st Qu.: 0.000       1st Qu.:  65.0       1st Qu.:7.000e+06   1st Qu.:5.700   1st Qu.:     0
 Median : 1.000       Median : 162.0       Median :2.000e+07   Median :6.400   Median :   242
 Mean   : 1.431       Mean   : 287.3       Mean   :4.371e+07   Mean   :6.314   Mean   :  9518
 3rd Qu.: 2.000       3rd Qu.: 347.0       3rd Qu.:4.800e+07   3rd Qu.:7.100   3rd Qu.: 11000
 Max.   :43.000       Max.   :5060.0       Max.   :1.222e+10   Max.   :9.100   Max.   :349000
```

통계 분석

Logistic Regression

KNN

LDA & QDA

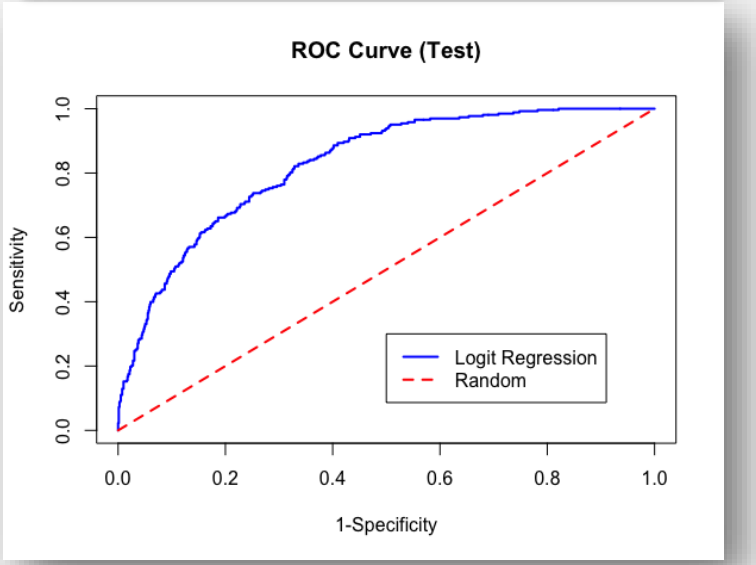
Decision Tree

로지스틱 회귀

Stepwise - 유의한 변수 고르기

```
> summary(fit2)

Call:
glm(formula = y ~ num_critic_for_reviews + cast_total_facebook_likes +
    num_user_for_reviews + imdb_score, family = binomial(link = "logit"),
    data = movie.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3411  -0.6484  -0.4622  -0.2831   2.3918

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.544e+00  3.188e-01  -4.843 1.28e-06 ***
num_critic_for_reviews    5.287e-03  5.612e-04   9.420  < 2e-16 ***
cast_total_facebook_likes 1.070e-05  2.678e-06   3.997 6.41e-05 ***
num_user_for_reviews      1.991e-03  2.115e-04   9.412  < 2e-16 ***
imdb_score               -2.052e-01  5.454e-02  -3.762 0.000168 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2789.8  on 2480  degrees of freedom
Residual deviance: 2206.9  on 2476  degrees of freedom
AIC: 2216.9

Number of Fisher Scoring iterations: 5
```

예측력 AUC: 0.82



Review 평론가 수 **Exp(0.005)**

Review 개수 **Exp(0.001)**

iMDB score **Exp(-0.2)**

영화 흥행 OR

KNN

01 Numeric 변수에 대해 scaling 진행

02 Loop문으로 k=1~10까지 KNN 진행
K에 따른 prediction accuracy 계산

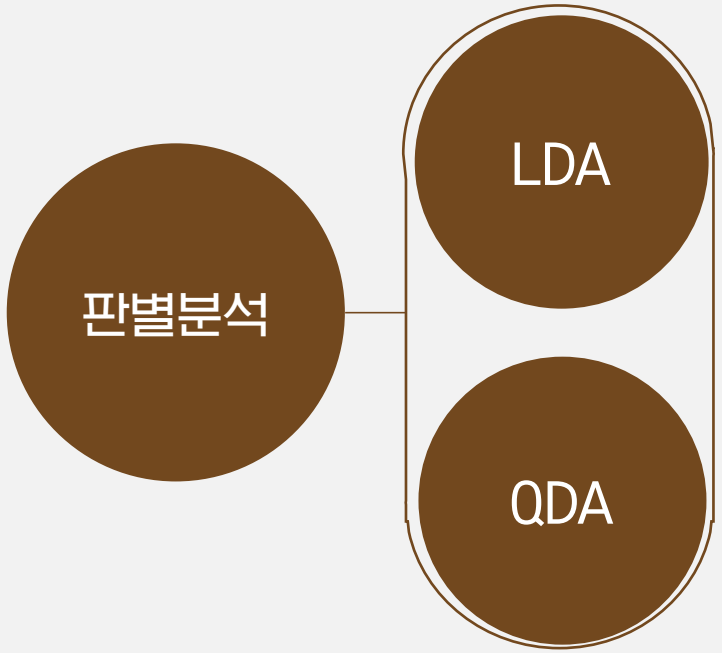
```
> for (i in 1:10){
+   fit.knn = knn(train=movie.train[,-9],test=movie.test[,-9],cl=movie.train[,9], k=i, prob=T)
+   yhat = fit.knn
+   ctable = table(movie.test[,9], yhat, dnn=c("Actual", "Predicted"))
+   miss.err = 1-sum(diag(ctable))/sum(ctable)
+   pred.acc = 1 - miss.err; round(pred.acc, 3)
+   print(pred.acc)
+ }
[1] 0.7190161
[1] 0.7114475
[1] 0.7341533
[1] 0.730369
[1] 0.7473983
[1] 0.7398297
[1] 0.7492904
[1] 0.7511826
[1] 0.7606433
[1] 0.7587512
```

K=9일 때 prediction accuracy가 가장 높음

LDA & QDA

LDA & QDA 란? 분석 결과 LDA vs QDA

집단에 대한 정보로부터
집단을 구별하는
판별함수를 만들고,
새로운 개체가 속할 집단을
판별하여 분류하는 기법



선형판별분석
- 정규분포, 등분산 가정을 기준으로함
- 적은 수의 모수를 예측하면 정확한 예측

이중판별분석
- 정규분포를 가정한다는 점에서 LDA와 유사
- LDA와 달리 등분산 가정이 필요하지 않음
- 많은 모수가 필요함

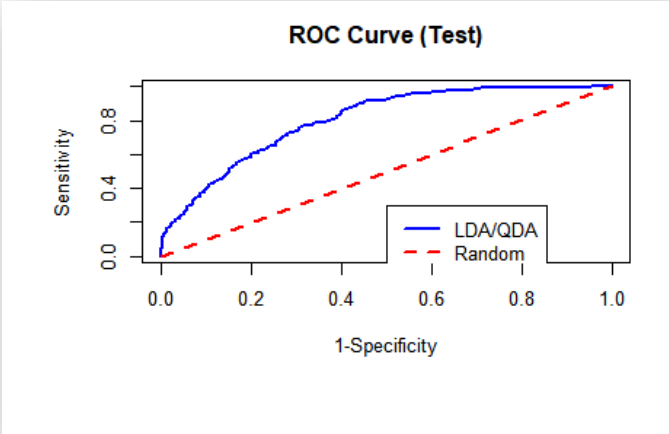
LDA

01 Test data 적용을 통한 분류: C-table

Actual \ Predicted	0	1
0	746	54
1	178	78

02 C-table의 정확도 0.78
Sensitivity 0.30
Specificity 0.93

03 AUC: 0.80



K=10

Train error 0.20598
Test error 0.20687

K=5

Train error 0.20619
Test error 0.20689

QDA

K=10

Test error 0.21645

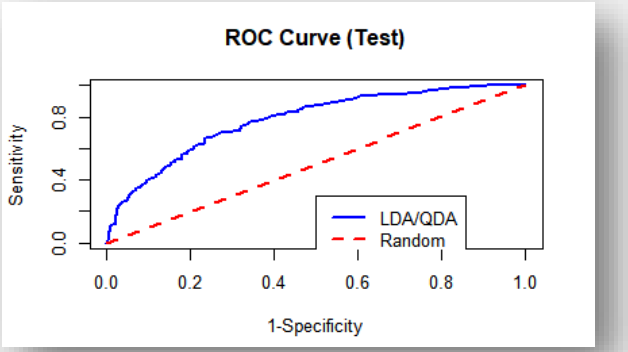
K=5

Test error 0.21883

01 Test data 적용을 통한 분류: C-table

Actual	Predicted 0	1
0	704	96
1	146	110

02 C-table의 정확도 0.77
Sensitivity 0.43
Specificity 0.88



03 AUC: 0.779

LDA vs QDA

LDA		QDA
0.78	ACCURACY	0.77
0.80	AUC	0.78
0.207	5-fold-cv test error	0.212
0.207	10-fold-cv test error	0.206

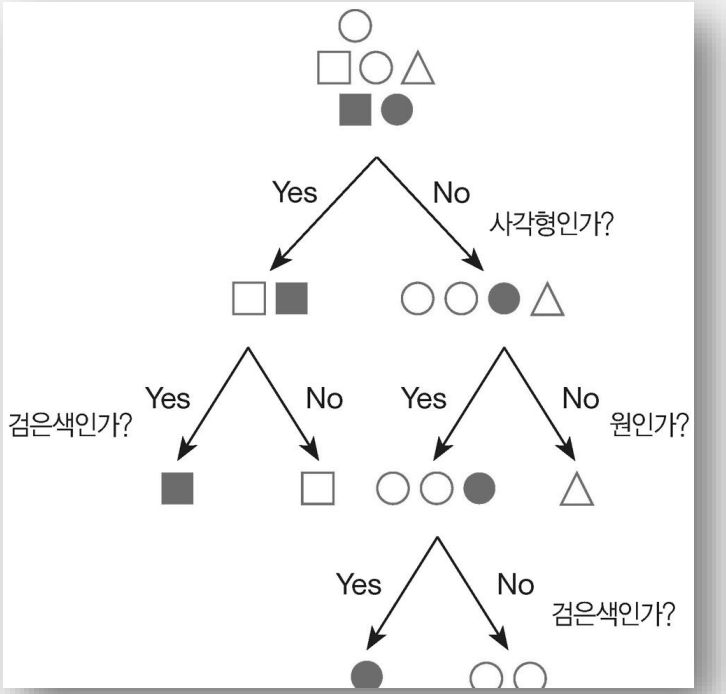
LDA가 전반적으로 더 잘 예측한다고 볼 수 있음
: Accuracy, AUC가 크고 test error가 낮음

그러나 LDA를 사용하기 위해서는 등분산성에 대한 검토필요

Decision Tree

의사 결정 나무 란? 분석 결과 정확도 검정

"데이터의 특징에 대한 질문을 하면서 응답에 따라 데이터를 분류해 나가는 알고리즘!"



장점

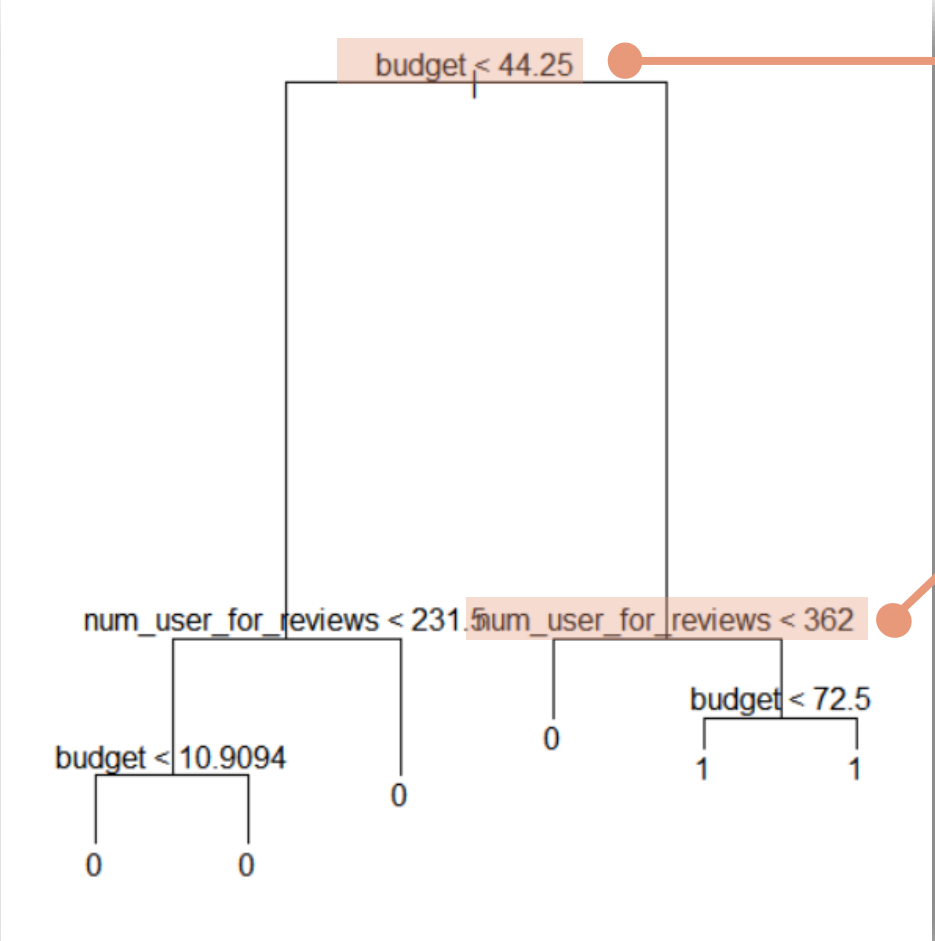
1) 분석 과정이 직관적, 이해하기 쉽다.
2) 수치형, 범주형 변수 모두 사용 가능
3) 계산 비용이 낮아 대규모의 데이터 셋에서도 빠르게 연산 가능

R에서 대표적인 패키지

CHAID | tree | rpart | party

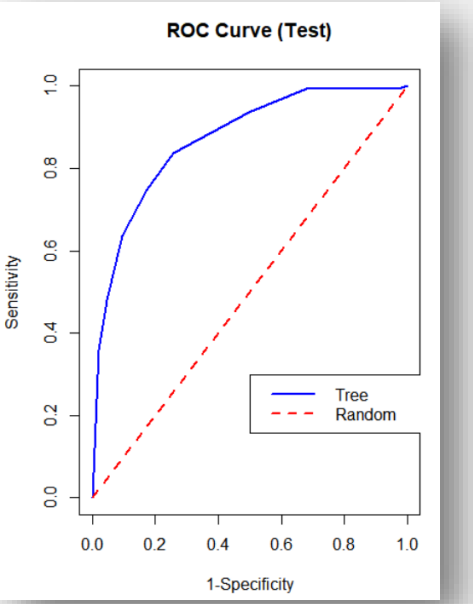
Tree

의사 결정 나무란? **분석 결과** 정확도 검정



가장 중요한 기준: 예산

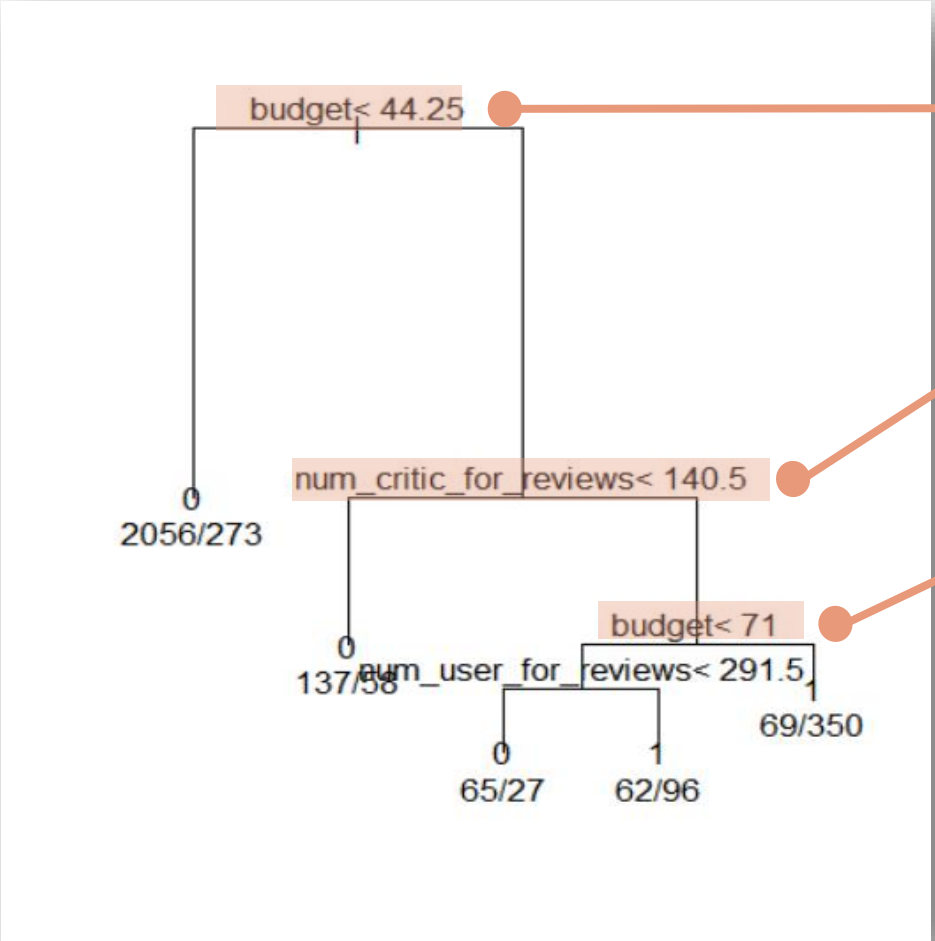
영화 제작 단계에서
투자를 많이 받고,
영화 제작 이후 리뷰 작성
평론가가 많아야 함!



ACCURACY	0.8111
AUC	0.869
5-fold-cv test error	0.1765
10-fold-cv test error	0.1777

rpart

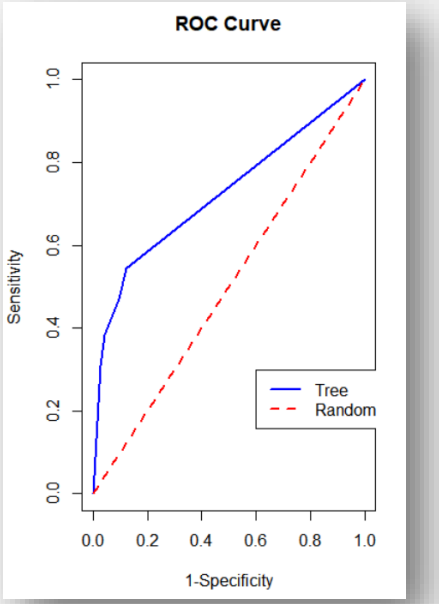
의사 결정 나무란? **분석 결과** 정확도 검정



가장 중요한 기준: 예산

예산이 44.25 초과,
리뷰 평론가 수 140.5 초과
할 때 흥행 가능성 높음

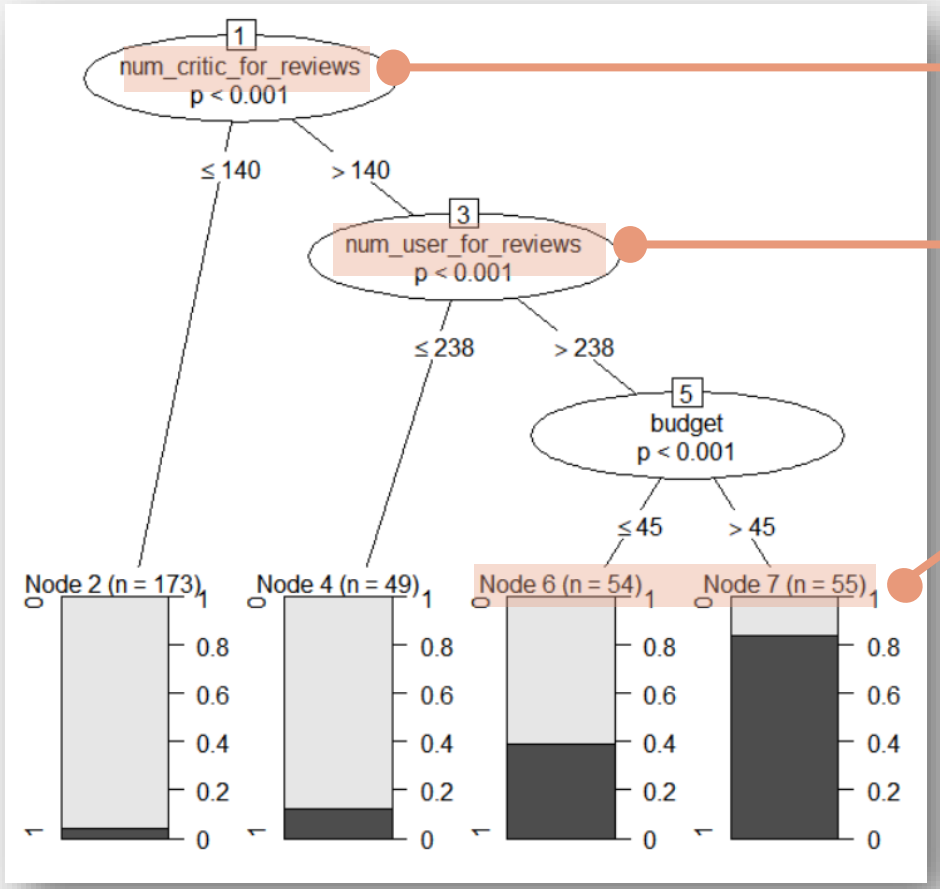
특히 예산이 71 넘으면
가장 많이 분류됨!



ACCURACY	0.840
AUC	0.723
5-fold-cv test error	0.174
10-fold-cv test error	0.158

party

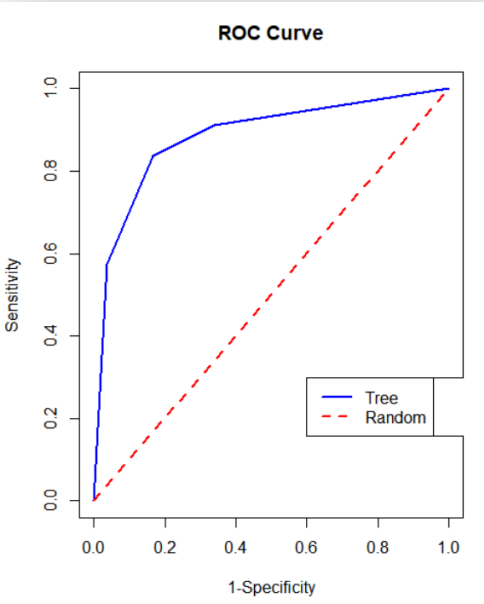
의사 결정 나무란? **분석 결과** 정확도 검정



가장 중요한 기준:
리뷰 평론가 수

그 다음 기준:
리뷰 일반인 수

평론가 리뷰 > 140
일반인 리뷰 > 238
→ 리뷰가 많을 때 흥행한다!



ACCURACY	0.849
AUC	0.885
5-fold-cv test error	0.182
10-fold-cv test error	0.181

Decision Tree

의사 결정 나무란? 분석 결과 **정확도 검정**

	Tree	Rpart	Party
ACCURACY	0.8111	0.840	0.849
AUC	0.869	0.723	0.885
5-fold-cv test error	0.1765	0.174	0.182
10-fold-cv test error	0.1777	0.158	0.181

모두 높은 accuracy 와 AUC값, 낮은 cv test error로
좋은 예측력을 가지고 있지만 party 패키지의 경우가 가장 좋음!

최종 결론

01 영화 흥행을 잘 예측하는 기법은?

	로지스틱	KNN	LDA	QDA	Tree	Rpart	Party
ACCURACY	0.8283	0.7686	0.78	0.77	0.8111	0.840	0.849
AUC			0.80	0.779	0.869	0.723	0.885
5-fold-cv test error			0.206	0.216	0.1765	0.174	0.182
10-fold-cv test error			0.206	0.218	0.1777	0.158	0.181

대체적으로 모든 기법의 예측력이 좋지만,
Accuracy와 AUC에서 decision tree의 party가 가장 좋았음!

최종 결론

02 영화 흥행에 영향을 미치는 요소는?

로지스틱
- Review개수
- iMDB score

의사결정 나무
- Budget
- Review 개수
- 평론가 review수

제언!

영화 제작 단계에서 투자를 탄탄히 받고,
개봉 이후 전문가 혹은 일반 관람객 리뷰를 활성화하는 방법을 모색하자!