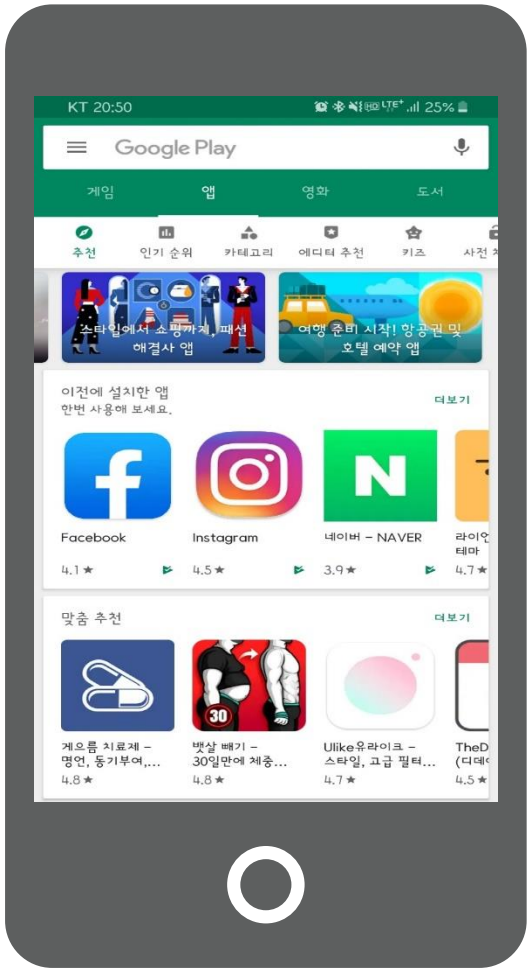
Google Playstore Application



응용통계학과
32150559 김 민 준
32173283 이 윤 정
32174871 한 재 리

00 목차

01 프로젝트 목적

02 데이터 탐색

03 데이터 전처리 및 축소

04 데이터 마이닝 문제 결정

05 데이터 마이닝 기법

06 알고리즘 과제 수행, 결과 해석

01 프로젝트 목적

관심분야

데이터 수집

kaggle

프로젝트 목적

Google

Google play

02 데이터 탐색

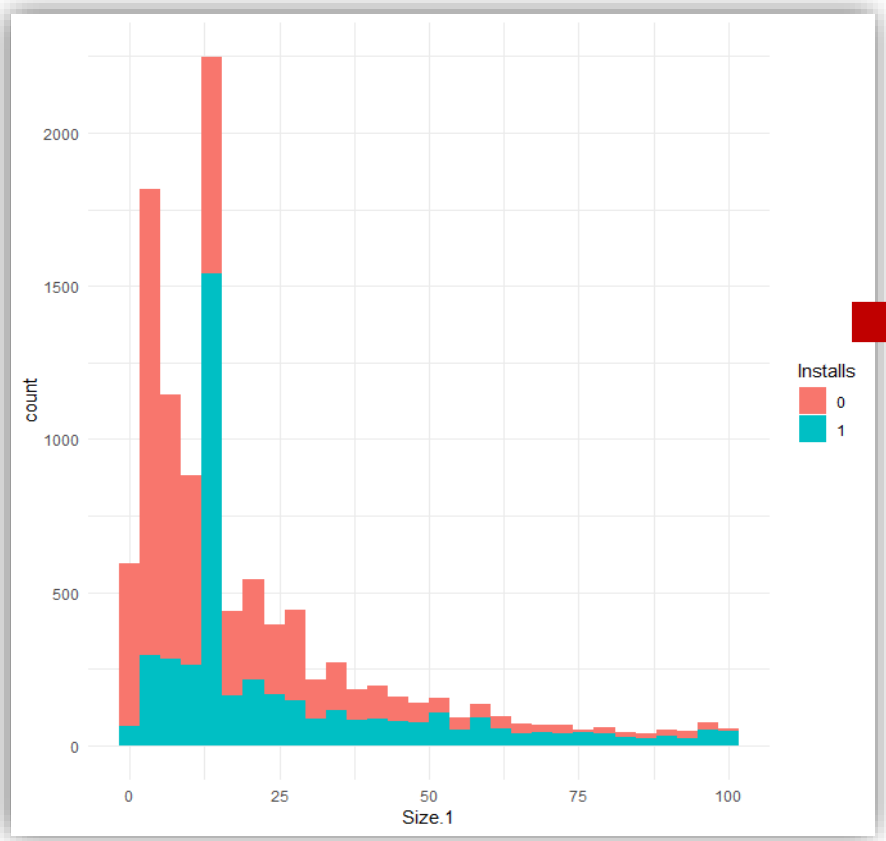
X : 12개, Y : 1개, 총 레코드 수: 10778개

변수 이름	변수 Type	변수 설명	Note
App	Factor	어플리케이션 이름	“Photo Editor & Candy Camera & Grid & ScrapBook”, “Coloring book moana”, ….
Category	Factor	어플리케이션들의 동일한 성질을 가진 부류나 범위	“ART_AND_DESIGN”, “AUTO_AND_VEHICLES”, “BEAUTY…”
Rating	Num	어플리케이션의 가치를 평하여 매긴 평균 점수	4.1, 3.9, 4.7, 4.5, 4.3, 4.4, 3.8, 4.1, 4.4, 4.7 ….
Reviews	Int	어플리케이션을 사용한 사람들이 작성한 후기 수	159, 967, 87510, 215644, 967, 167, 178, 36815…
Size	Factor	어플리케이션을 설치할 때 필요한 용량의 크기	19M, 14M, 8.7M, 25M, 2.8M, 5.6M, 29M, 33M….
Installs	Factor	얼마나 많은 사람들이 어플리케이션을 다운로드 받았는지 설치 수	0+, 1,000,000,000+, 1,000,000+, 10,000,000+….
Type	Factor	어플리케이션을 다운로드 받는 유형이 유료인지 무드인지를 나타내는 종류	“Free”, “Paid”
Price	Factor	어플리케이션을 다운로드 받을 때 가격이 얼마인지를 나타냄	$0.99, $1.00, 0, $3.90, $14.99, $2.99 …..

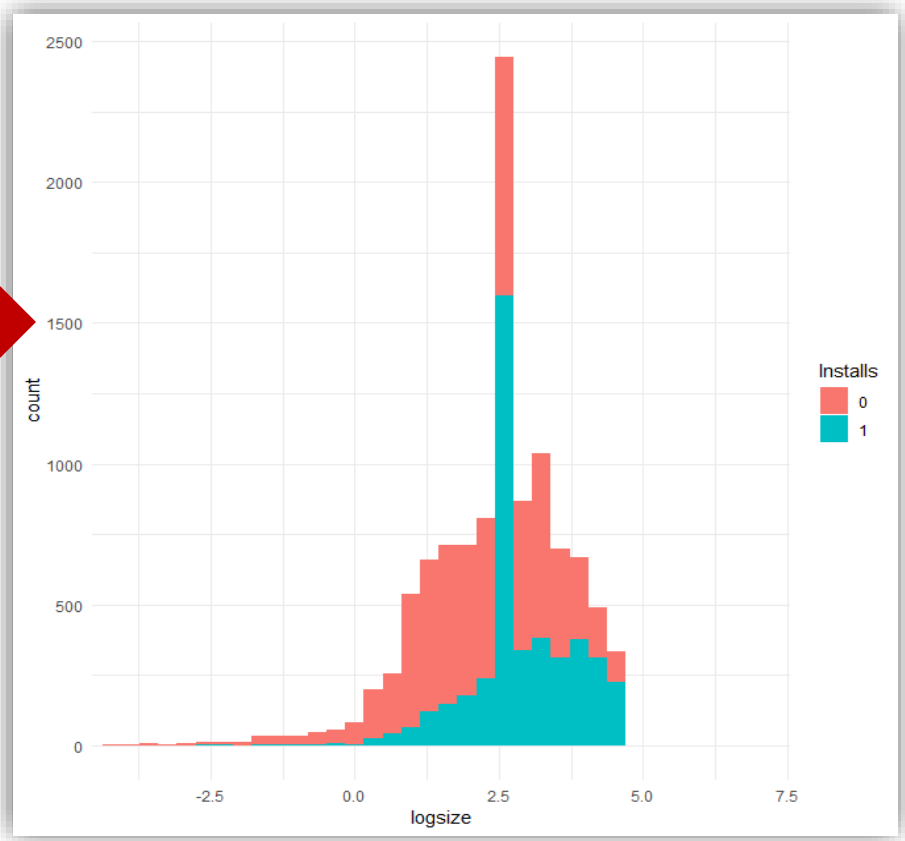
02 데이터 탐색

변수 이름	변수 Type	변수 설명	Note
Content.Rating	Factor	어플리케이션을 사용할 수 있는 연령대를 나타냄	“Everyone”, “Everyone 10+”, “Teen”, “Adults Only 18+”…..
Genres	Factor	어플리케이션들의 동일한 성질을 가진 부류나 범위 -Factor	“Action”, “Tools”, “ Role Playing”, “Business” …..
Last.Updated	Factor	어플리케이션이 가장 최근에 업데이트 된 날짜	“01-Apr-16”, “06-May-18”,”28-Feb-17”……
Current.Ver	Factor	어플리케이션 가장 최근 버전	“Varies with devices”, “1.0.5”, “3.3.0”,”1.0.12”…..
Android.Ver	Factor	어플리케이션이 실행될 수 있는 가장 낮은 안드로이드 버전	“Varies with devices”, “2.3 and up”, “4.4 and up”…

02 데이터 탐색



Size 변수 분포

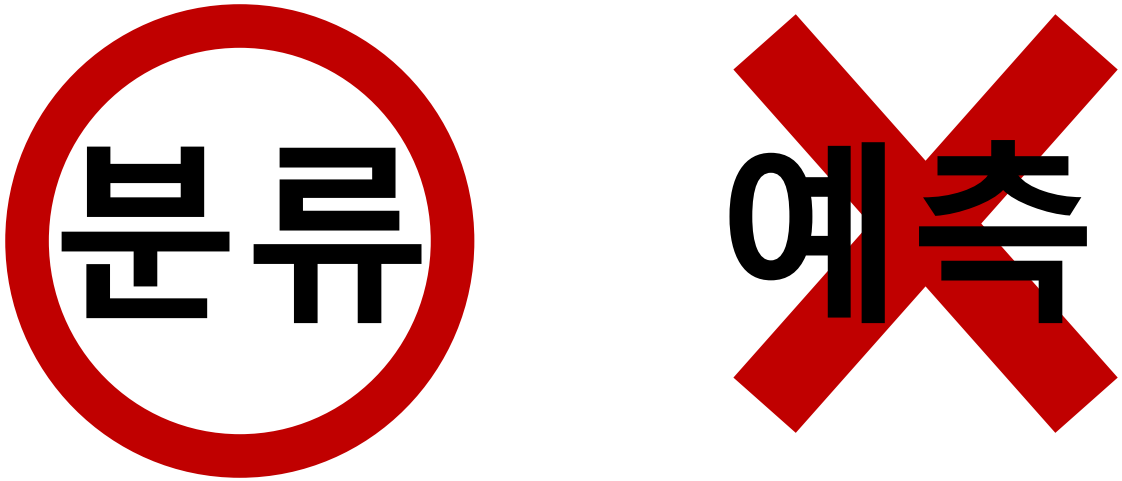


log(Size) 변수 분포

03 데이터 전처리 및 축소

제거한 변수	변수 Type	제거 이유	Note
App	Factor	앱의 이름은 정보를 제공할 뿐 인기의 척도에 영향을 주는 요소가 아니기 때문에 제거	“Photo Editor & Candy Camera & Grid & ScrapBook”, “Coloring book moana”, ….
Rating	Num	앱이 출시된 이후에 알 수 있기 때문에 예측 변수로 사용할 수 없다	4.1, 3.9, 4.7, 4.5, 4.3, 4.4, 3.8, 4.1, 4.4, 4.7 ….
Reviews	Int	앱이 출시된 이후에 알 수 있기 때문에 예측 변수로 사용할 수 없다	159, 967, 87510, 215644, 967, 167, 178, 36815…
Price	Factor	대부분이 “Free”로 9986개, “Paid”는 792개로 전체의 0.074%밖에 안되기 때문에 예측 변수로 사용할 수 없다	$0.99, $1.00, 0, $3.90, $14.99, $2.99 …..
Genres	Factor	Category변수와 겹치기 때문에 제거	“Action”, “Tools”, “ Role Playing”, “Business” …..
Last.Updated	Factor	앱이 출시된 이후에 알 수 있기 때문에 예측 변수로 사용할 수 없다	“01-Apr-16”, “06-May-18”,”28-Feb-17”……
Current.Ver	Factor	앱이 출시된 이후에 알 수 있기 때문에 예측 변수로 사용할 수 없다	“Varies with devices”, “1.0.5”, “3.3.0”,”1.0.12”…..

04 데이터 마이닝 문제 결정



05 데이터 마이닝 기법

1. 나이브 베이즈
2. 분류나무
3. 신경망

06 알고리즘 과제 수행, 결과 해석

1) 나이브 베이즈

Reference		
	0	1
0	3232	604
1	1262	1368

학습

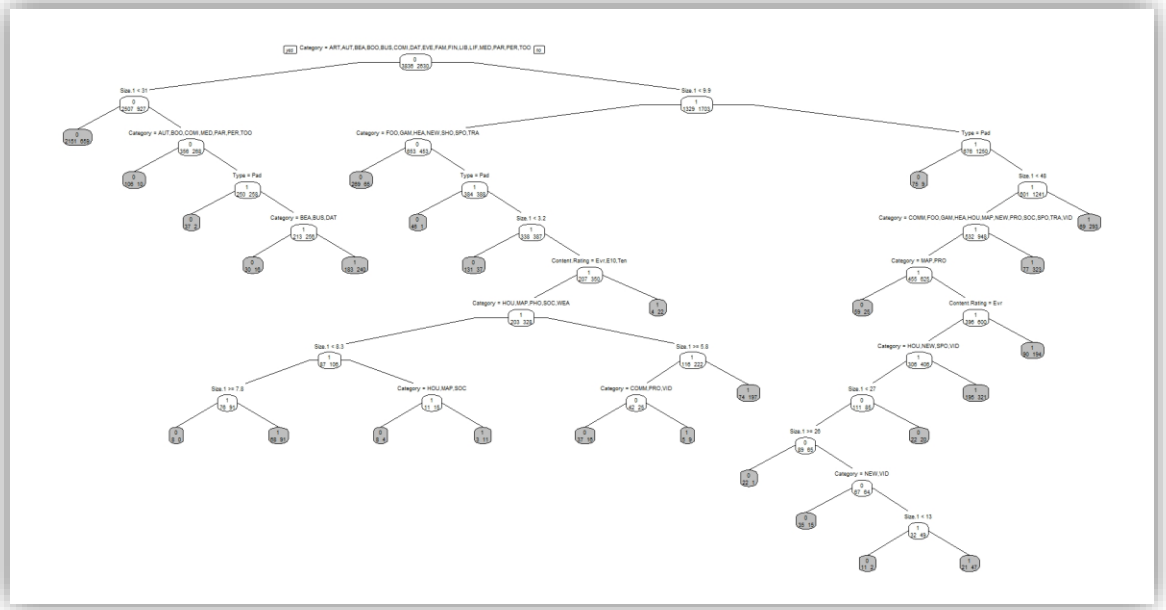
Reference		
	0	1
0	2167	391
1	884	870

검증

	Sensitivity	Specificity	Precision	F1	Accuracy	Misspecification
학습	0.8425	0.5202	0.7192	0.7760	0.7114	0.2886
검증	0.8471	0.4960	0.7130	0.7727	0.7043	0.2957

06 알고리즘 과제 수행, 결과 해석

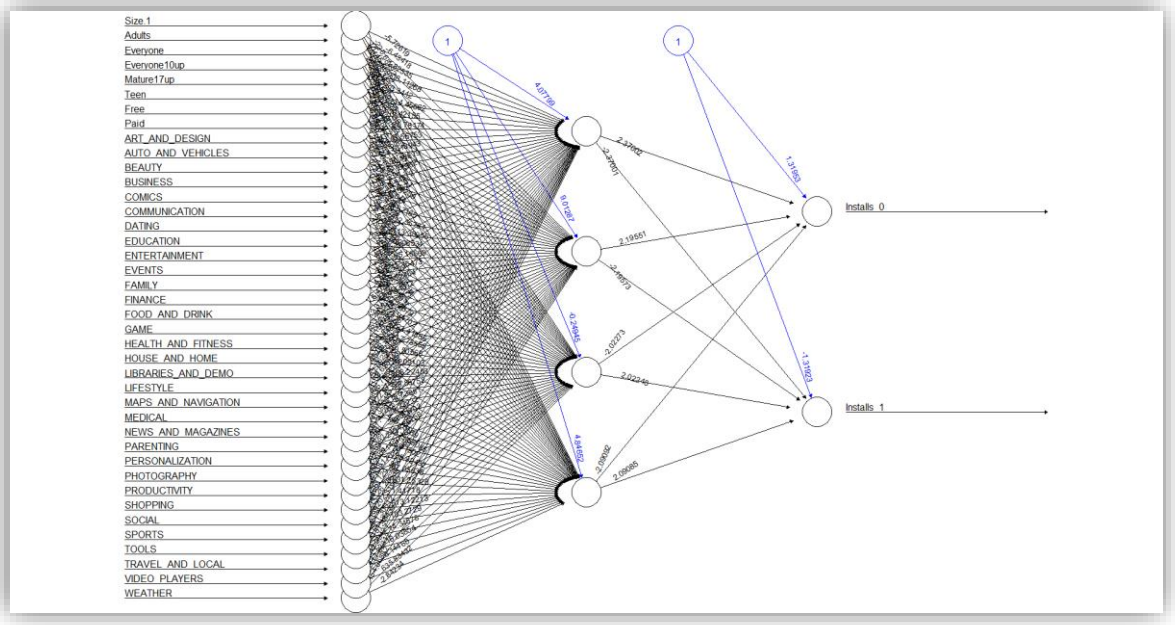
2) 분류 나무



	Sensitivity	Specificity	Precision	F1	Accuracy	Misspecification
학습	0.7755	0.6890	0.7943	0.7848	0.7416	0.2584
검증	0.7549	0.6671	0.7850	0.7696	0.7212	0.2788

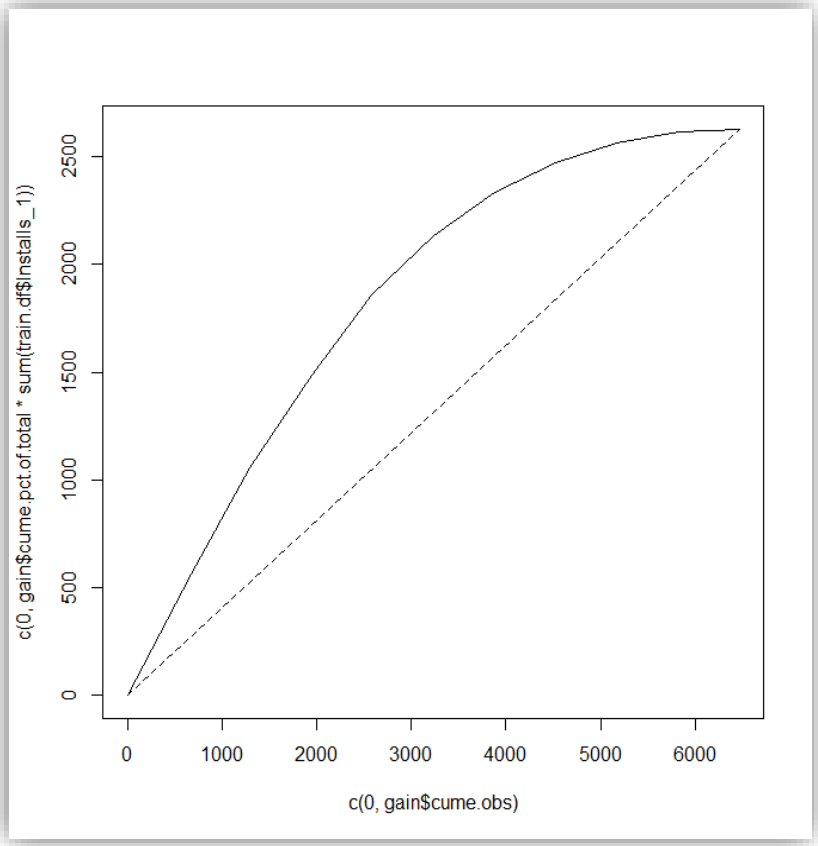
06 알고리즘 과제 수행, 결과 해석

3) 신경망

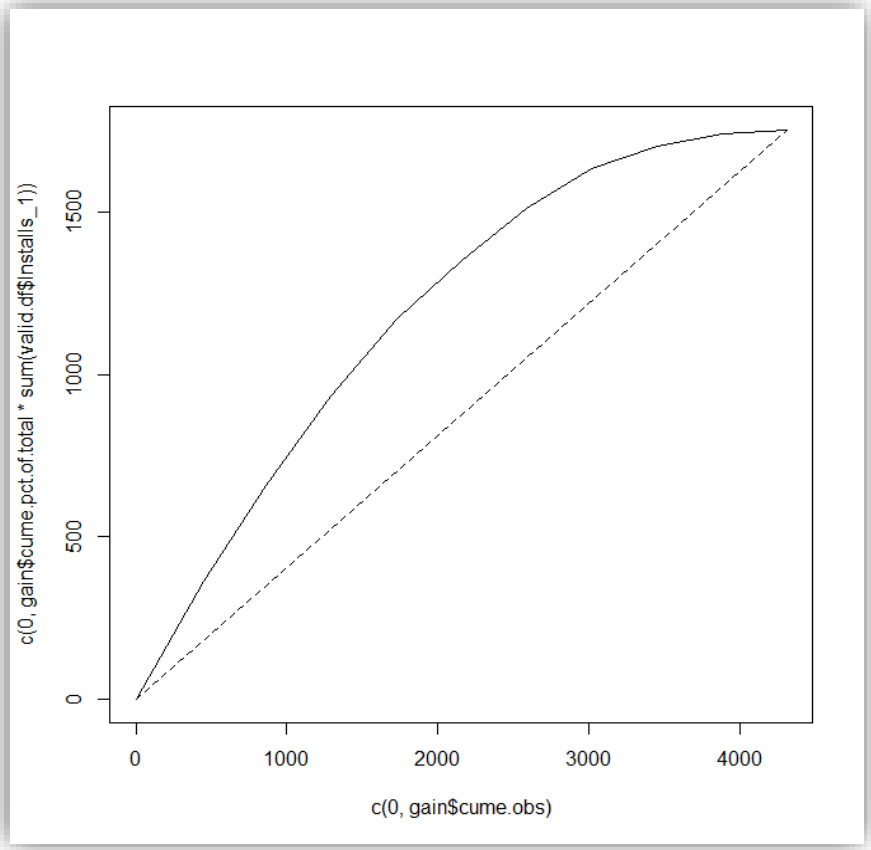


	Sensitivity	Specificity	Precision	F1	Accuracy	Misspecification
학습	0.7719	0.7555	0.8216	0.7960	0.7652	0.2348
검증	0.7498	0.7269	0.8002	0.7742	0.7405	0.2595

06 알고리즘 과제 수행, 결과 해석

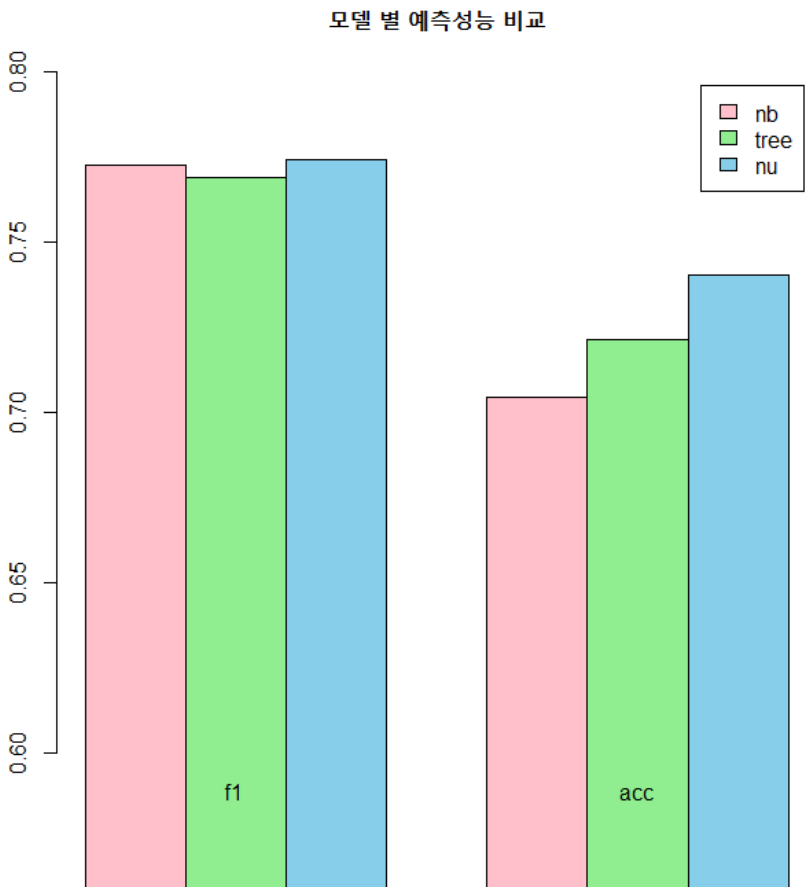


학습데이터
Lift Chart



검증데이터
Lift Chart

06 결과 해석



	F1-score	Accuracy
나이브 베이즈	0.7727	0.7043
분류나무	0.7696	0.7212
신경망	0.7742	0.7405

06 결과 해석

- 신경망이 가장 예측 성능이 좋다고 볼 수 있음
- 신경망은 변수 간 관계 파악이 불가하므로 모델에 대한 설명이 부족
- 분류나무에서 보면 Category, Size가 흥행 여부에 중요한 영향을 미치는 변수라고 볼 수 있다.
- If(Category= Education, Entertainment, Photography, Shopping, Weather) AND (Type=FREE) AND (size.1>=48) THEN 흥행 (클래스 = 1)
- 어플의 용량(Size)이 작다고 꼭 흥행하는 것은 아님을 알 수 있고, 유료(Type=Paid)보다는 무료(Type=Free)가 흥행에 영향을 미친다고 볼 수 있다. Category는 생활 관련 분야가 흥행에 영향을 미친다고 볼 수 있다.

THANKS