14주차:
구글 플레이 앱 스토어를 이용한 실전 프로젝트

ChulSoo Park

School of Computer Engineering & Information Technology

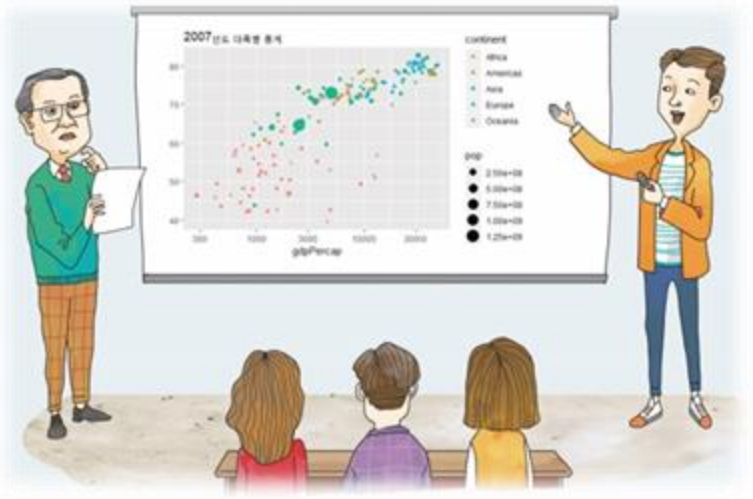
Korea National University of Transportation

학습목표 (14주차)

- 데이터를 이용한 실전 프로젝트 수행
- 데이터와 친해지기
- 전략적 통찰과 비즈니스에 집중한 분석
- 데이터 사이언스 메인 Process(pipeline) 정리

12 CHAPTER

실전 프로젝트



CONTENTS

12.1 프로젝트 소개
12.2 데이터 정제
12.3 탐색적 데이터 분석
12.4 모델링과 예측
요약

12.4 모델링과 예측

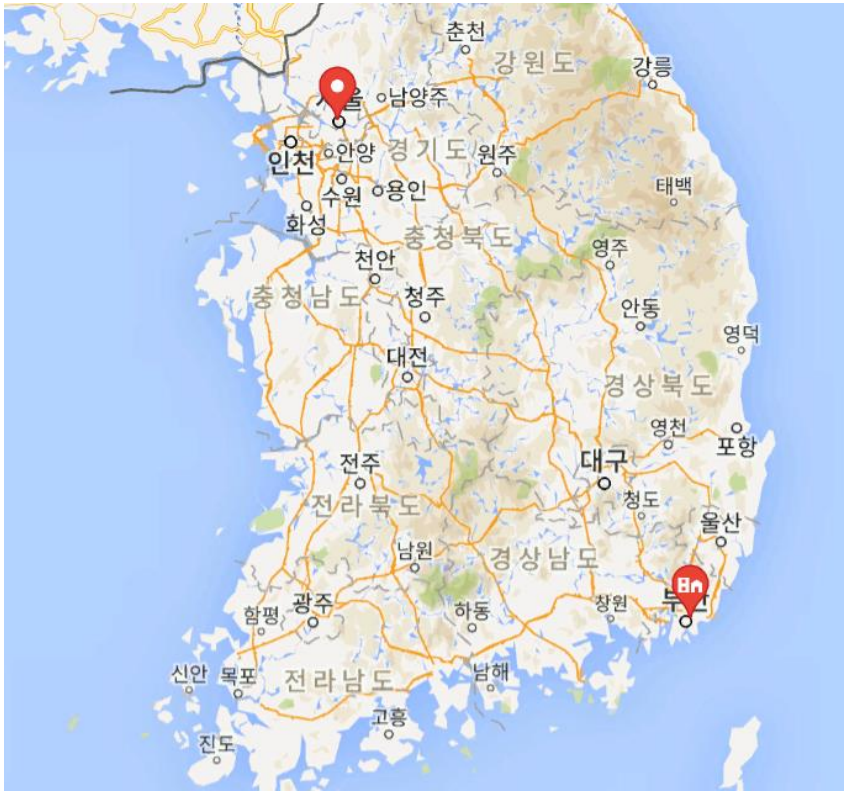
- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기

마케팅 부서의 중요 전략 수립 과정

- 제품의 목록
- 가격
- 크기
- 구매량
- 소비자 만족도
- 구매 후기

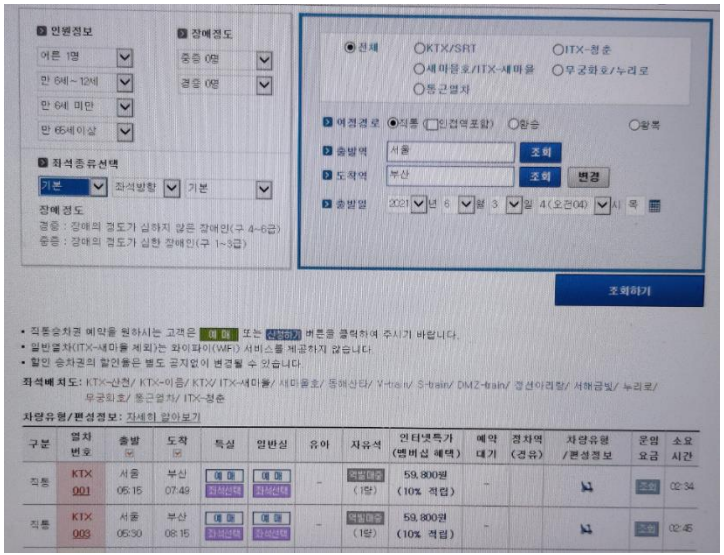
12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기



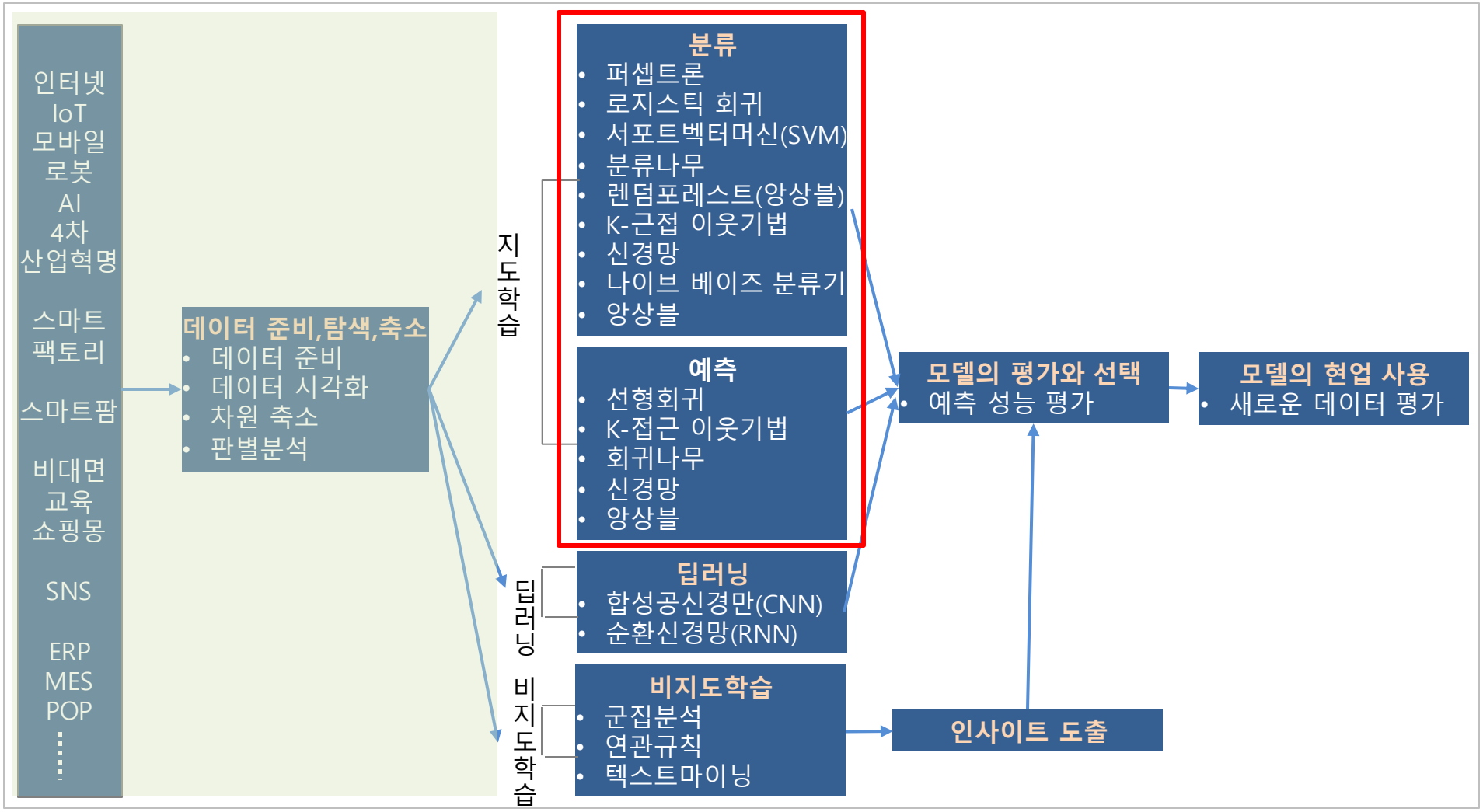
서울서 부산 가는 방법

1. 기차
2. 자동차
3. 버스
4. 비행기
5. 배
6. 자전거
7. 걸어서



12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기

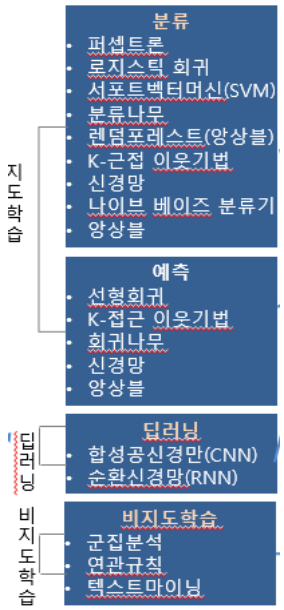


12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기

알고리즘 선택 과정

기본 모델 선택 → 하이퍼 매개변수 최적화 → 모델 평가



svm 커널 함수
rf 트리의 수 등

정확률
정밀도
재현율
ROC
AUC

12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기
  - ✓ 반응 변수와 설명 변수 선정

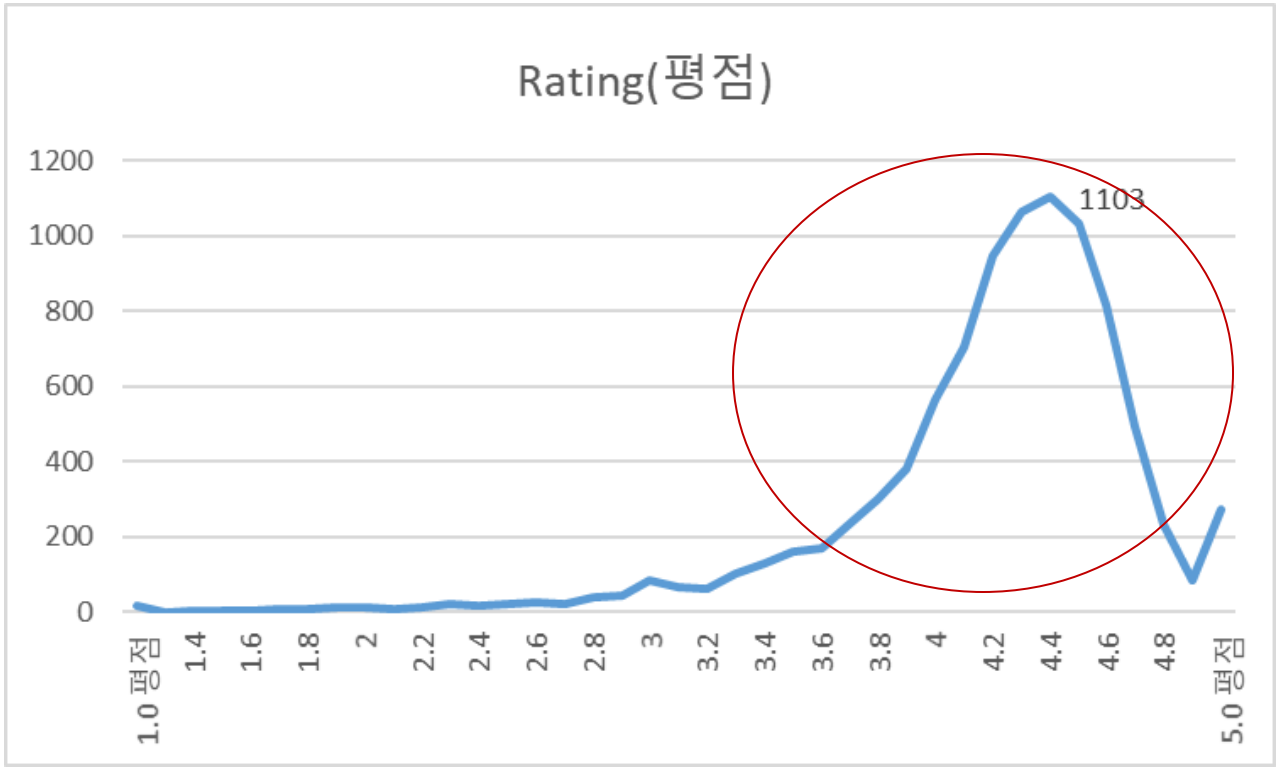
1. 반응 변수 :

2. 설명 변수 :

12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기
  - ✓ 반응 변수 살펴보기

평 균 : 4.19
최빈값 : 4.4



행 레이블	Rating(평점)
1.0 평점	18
1.2 평점	1
1.4	3
1.5	3
1.6	4
1.7	8
1.8	8
1.9	13
2	13
2.1	8
2.2	14
2.3	20
2.4	19
2.5	21
2.6	25
2.7	24
2.8	42
2.9	45
3	83
3.1	68
3.2	64
3.3	102
3.4	128
3.5	163
3.6	172
3.7	238
3.8	302
3.9	383
4	567
4.1	704
4.2	947
4.3	1066
4.4	1103
4.5	1035
4.6	818
4.7	493
4.8	233
4.9	87
5.0 평점	275

12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기
  - ✓ 국가직무능력표준(NCS)에서 모델링의 Data set 준비

학습 1	머신러닝 수행방법 계획하기(LM2001010507_15v1.1)
학습 2	**데이터 세트 분할하기 (LM2001010507_15v1.2)**
학습 3	지도학습 모델 적용하기(LM2001010507_15v1.3)
학습 4	자율학습 모델 적용하기(LM2001010507_15v1.4)
학습 5	모델성능 평가하기(LM2001010507_15v1.5)
학습 6	학습결과 적용하기(LM2001010507_15v1.6)

2-1. 데이터 세트 준비 및 분할

학습 목표	• 과적합과 일반화의 의미 및 파급효과를 이해하고, 이의 해결을 위한 데이터 세트 분할을 설계할 수 있다. • 분석하고자 하는 목적 및 데이터 세트 특성에 따라 머신러닝 기법 적용을 위한 훈련 데이터 세트와 테스트 데이터 세트 분할 기준을 판단할 수 있다. • 해결하고자 하는 이슈와 적용할 기법에 따라 교차검증 필요성을 판단하여, 훈련 데이터 세트와 검증 데이터 세트를 분할하고, 적합한 교차검증 K값을 결정할 수 있다 • 데이터의 특성을 고려한 예측 및 분류 목적 변수의 분포를 고려하여 데이터 세트를 분할하고 샘플링할 수 있다 • 데이터 세트 분할을 위한 다양한 샘플링 방법의 차이를 분석하여 샘플링 방법을 적용할 수 있다

출처 : https://www.ncs.go.kr

12.4 모델링과 예측

- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기
  - ✓ 국가직무능력표준(NCS)에서 모델링의 Data set 준비

필요 지식 /

1 데이터 세트 분할의 필요성

일반적으로, 머신러닝 기반 데이터 분석 진행시, 특히 지도학습 기
전체 데이터 세트를 사용하여 한꺼번에 분석하지 않고, 학습용 데
트) 데이터 세트로 분할하여 분석을 진행하게 된다.

1. 모델링의 과적합 및 일반화

머신러닝 기반 데이터 분석을 진행함에 있어서, 머신러닝 기법이
진 데이터 세트를 학습하여 최적의 모수(파라미터)를 도출하는
설명변수(혹은 특성(Feature))가 주어졌을 때 목적변수 (혹은 반
과정이라고 할 수 있다. 그런데 주어진 훈련 데이터 세트에 포함
면 '우연에 의해 얻어진 값' 이라고 볼 수 있으므로 새로운 목
의 값을 예측하기 위해 얻어지는 신규 데이터 세트는 원래의 훈

2 데이터 세트 분할 방법 및 절차

일반적으로 머신러닝 기반 분석 수행 시 훈련 데이터(혹은 학습 데이터)와 평가데이터를 나누어 모델링을 수행하는 과정은 다음과 같다.

1. 일정 비율로 학습용과 평가용 세트로 데이터 분할

데이터의 일부를 훈련 데이터, 나머지를 평가데이터로 분리한다. 특별한 경우가 아니라면 일반적으로 학습용과 평가용 데이터 각각의 분할은 전체 데이터에서 랜덤하게 특정 비율로 학습용 데이터를 추출하고, 학습용 데이터에 사용되지 않은 나머지 데이터를 평가용 데이터로 취하는 방법을 따른다. 이때 훈련 데이터와 평가데이터를 분할하는 비율은 정해진 원칙이 있는 것은 아니나, 모델을 훈련시키는 과정 자체에 더 많은 비중을 할당한다. 일반적으로 훈련 데이터를 60%~80%, 평가데이터를 40%~20% 정도로 할당한다. 그러나 절대적인 기준은 아니며, 실무 상황에서는 분석의 목적이나 연구수행자의 판단과 경험을 통해 분할 비율을 정하게 된다. 다만 데이터 세트 분할 시 중요한 점은 실제 훈련된 모델의 성능은 학습용 데이터 세트 크기가 작아질수록 나빠지게 되므로, 너무 많은 데이터를 평가용 데이터로 분할하는 것은 최종 성능에 오히려 나쁜 영향을 끼칠 수 있다는 점이다.

2. 학습(훈련) 데이터로부터 머신러닝 모델링 수행

앞의 1단계에 의해 추출된 훈련 데이터에 목적에 적합한 머신러닝 기법을 적용하여 머신러닝 모델링을 수행한다. 이때 여러 가지 기법을 적용하여 기법 간의 성능을 비교할 수도 있고, 동일 기법 내에서도 추정방법을 변경하거나 파라미터를 다양하게 변경하는 등의 과정을 거치게 된다. 여기서 모델링 성능에 대한 보다 정교한 검증을 위해 교차검증(Cross-Validation) 방법을 수행하는 경우도 있다. 교차검증은 훈련 데이터를 통한 모델링

12.4 모델링과 예측

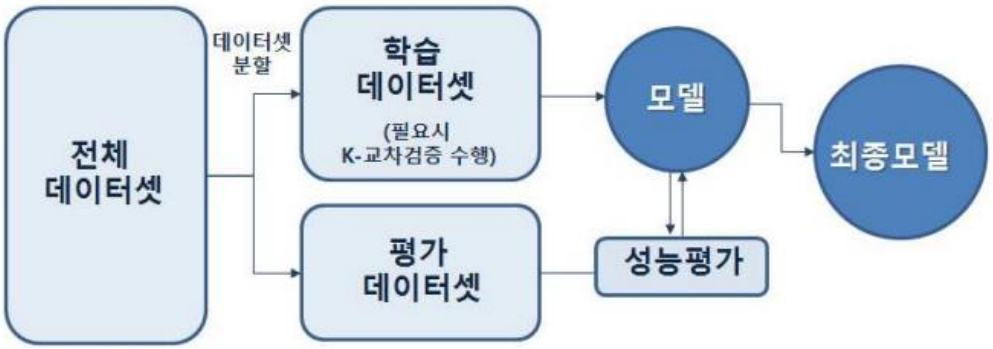
- 구글 앱의 데이터를 모델링하여 앱의 평점 예측하기
  - ✓ 국가직무능력표준(NCS)에서 모델링의 Data set 준비

3. 평가 데이터를 이용한 모델 성능 평가

2단계에서 만들어진 모델에 평가 데이터를 적용해 성능을 평가한다. 만일 성능이 만족스럽지 않다면, 앞의 2단계로 다시 돌아가게 된다. 여기서 평가 데이터는 원래의 전체 데이터 세트로부터 최초 분리해낸 뒤, 모델링 과정에서 이용되지 않다가 2단계 등을 통해 모델이 만들어지고 난 뒤, 해당 모델의 성능을 평가하기 위해 3단계에서 사용된다. 이런 점에서 모델링 과정 자체에서 검증의 목적으로 반복적으로 사용되는 검증 데이터와는 그 활용목적이 다르다고 할 수 있다.

4. 최종 모델 결과 제출

3단계에서 평가 데이터를 이용하여 수행한 성능 평가 및 예측결과가 기준치에 부합하거나, 목적에 적합하다고 판단될 경우 분석 모델링 과정을 종료하고, 최종 분석결과를 제출하게 된다.



[그림 2-2] 훈련 데이터와 평가 데이터 분할 통한 머신러닝 모델링 수행절차

12.4 모델링과 예측

> m = lm(Rating ~ Size + Content.Rating + Category, data = x)

```
Console  C:/RSources/
> # 선형 모델의 회구 분석
> m = lm(Rating ~ Size + Content.Rating + Category, data = x)
>
> # 선형 모델 계수
> m

Call:
lm(formula = Rating ~ Size + Content.Rating + Category, da

Coefficients:
                (Intercept)                        Size
                  4.727e+00                   1.373e-09
    Content.RatingEveryone    Content.RatingEveryone 10+
                 -3.850e-01                  -3.537e-01
  Content.RatingMature 17+          Content.RatingTeen
                 -4.120e-01                  -3.717e-01
     Content.RatingUnrated    CategoryAUTO_AND_VEHICLES
                 -2.913e-01                  -2.249e-01
           CategoryBEAUTY  CategoryBOOKS_AND_REFERENCE
                 -7.168e-02                  -4.325e-02
         CategoryBUSINESS               CategoryCOMICS
                 -2.453e-01                  -2.407e-01
     CategoryCOMMUNICATION                CategoryDATING
                 -2.571e-01                  -3.897e-01
          CategoryEDUCATION       CategoryENTERTAINMENT
                  1.458e-02                  -2.386e-01
            CategoryEVENTS                CategoryFAMILY
                  1.143e-01                  -1.964e-01
```

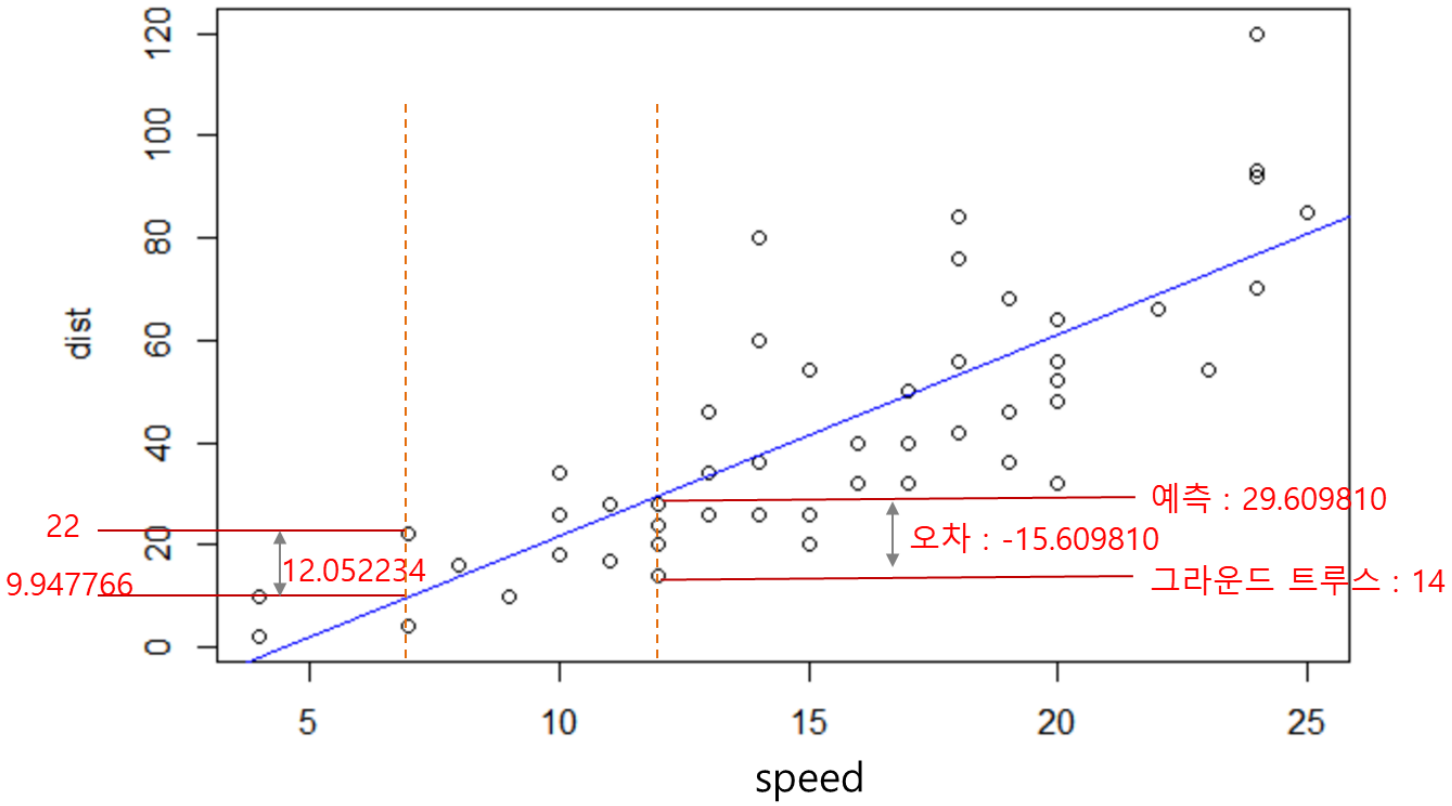
```
          CategoryFINANCE          CategoryFOOD_AND_DRINK
                -2.560e-01                      -2.792e-01
             CategoryGAME     CategoryHEALTH_AND_FITNESS
                -1.409e-01                      -1.524e-01
   CategoryHOUSE_AND_HOME     CategoryLIBRARIES_AND_DEMO
                -2.043e-01                      -1.625e-01
        CategoryLIFESTYLE    CategoryMAPS_AND_NAVIGATION
                -2.707e-01                      -3.525e-01
          CategoryMEDICAL     CategoryNEWS_AND_MAGAZINES
                -1.850e-01                      -2.209e-01
        CategoryPARENTING        CategoryPERSONALIZATION
                 5.360e-03                      -3.622e-02
     CategoryPHOTOGRAPHY           CategoryPRODUCTIVITY
                -2.219e-01                      -2.162e-01
       CategorySHOPPING                 CategorySOCIAL
                -1.394e-01                      -1.119e-01
         CategorySPORTS                  CategoryTOOLS
                -1.770e-01                      -3.448e-01
  CategoryTRAVEL_AND_LOCAL          CategoryVIDEO_PLAYERS
                -3.338e-01                      -3.400e-01
      CategoryWEATHER
                -1.206e-01
```

```
> # mse 산출
> deviance(m)/nrow(x)
[1] 0.2422341
```

12.4 모델링과 예측

```
> m = lm(Rating ~ Size + Content.Rating + Category, data = x)
```

mse : 평균 제곱 오차(Mean squared error)



12.4 모델링과 예측

Console C:/RSources/

121	4.4	BEAUTY	4200000	Everyone
122	4.6	BEAUTY	7100000	Everyone
123	4.5	BEAUTY	57000000	Everyone
124	NaN	BEAUTY	3700000	Everyone
125	3.9	BEAUTY	22000000	Everyone
126	4.4	BEAUTY	24000000	Everyone
127	NaN	BEAUTY	7400000	Teen
128	4.6	BEAUTY	21000000	Everyone
129	3.8	BEAUTY	3400000	Everyone
130	NaN	BEAUTY	2900000	Mature 17+
131	NaN	BEAUTY	3100000	Everyone 10+
132	4.0	BEAUTY	6400000	Everyone
133	4.3	BEAUTY	3200000	Everyone
134	4.5	BEAUTY	8200000	Mature 17+
135	NaN	BEAUTY	9900000	Mature 17+
136	4.1	BEAUTY	2900000	Everyone
137	3.7	BEAUTY	23000000	Everyone
138	4.7	BEAUTY	4600000	Everyone

NAN 제거

```
x = na.omit(x)
ps = select(x, Rating, Category, Size, Content.Rating)
```

12.4 모델링과 예측

3-2. 수치 예측 목적의 머신러닝 기법 적용

학습 목표

- 연속형 목적변수(혹은 반응변수)가 주어진 경우, 이의 문제 해결을 위해 다양한 수치 예측 모델을 비교해 보고 최적의 수치예측모델을 선정하여 적용할 수 있다.
- 주어진 상황과 데이터 특성에 따라 필요시, 여러 알고리즘이나 기법을 융합한 앙상블 모형을 적용하거나 새로운 방법론을 개발할 수 있다.

2 수치예측 목적을 위한 머신러닝

수치예측 목적의 머신러닝 기법분류기법은 설명하고자 하는 목적변수(혹은 반응변수)가 연속형 변수 형태를 가질 때 사용되는 기법으로서 데이터 분석에서 매우 자주 접하게 되는 문제라고 할 수 있다.

1. 수치예측 목적의 머신러닝 활용 영역

수치예측 목적의 머신러닝 알고리즘은 독립변수(혹은 설명변수, 예측변수)를 이용하여 관심 있는 목적변수(혹은 반응변수)의 수치값을 예측하는 형태의 문제라면 거의 대부분 적용할 수 있다. 대표적인 예시는 다음과 같다.

(1) 주식 가격 예측
(2) 경제 지표 예측
(3) 기업의 제품 판매량 및 가격 변화 예측
(4) 대출 채무 불이행에 대한 손실금액 예측
(5) 고객 LTV (Customer Lifetime Value) 예측
(6) 상품 구매 가능성 추천
(7) 인구통계 특성에 따른 의료비 증감 예측

위의 경우 외에도 관심 있는 목적변수(혹은 반응변수) 값의 변화를 설명하기 위해 적절한 독립변수(혹은 설명변수)를 활용하여 관계식(함수식)을 설정하여 예측할 수 있는 형태라면 무궁무진한 활용 예시를 만들어 낼 수 있을 것이다.

12.4 모델링과 예측

2. 수치예측 목적의 머신러닝 알고리즘 종류

수치예측 목적의 머신러닝 알고리즘은 통계학에서도 가장 많이 사용되는 회귀분석(Regression Analysis)이 가장 대표적인 기법이라고 할 수 있으며, 그 외에도 분류목적의 머신러닝에서도 활용되었던 의사결정트리나, 인공 신경망 기법, 서포트 벡터 머신, 랜덤 포레스트 등도 수치예측문제에 활용할 수 있다.

<표 3-8> 수치 예측 목적 주요 머신러닝 알고리즘 기법

종 류	개 념	비 고
회귀분석 (Regresssion Analysis)	관측된 사건들을 정량화해서 독립변수와 종속변수의 관계를 함수식으로 설명하는 방법. 해당 함수식이 모수에 대해 선형일 경우 선형회귀분석이라고 하며, 독립변수가 한 개인 경우 단순선형회귀, 여러 개일 경우 다중선형회귀분석을 적용함	추론통계기반 모형
의사결정트리 (Decision Tree)	목표변수와 가장 연관성이 높은 변수의 순서대로 나무 형태로 가지를 분할하면서 규칙을 만들어내지만, 분류목적의 의사결정트리가 지니계수나 엔트로피를 사용하는 것과는 달리 수치예측 목적일 때는 분산(혹은 표준편차)의 감소량(Variance Reduction)을 최대화하는 기준의 최적분리에 의해 회귀나무를 형성하게 됨	분할 정복기법 (Divide & Conquer)
인공 신경망 분석 (Artificial Neural Network)	인간의 뇌의 뉴런 작용 형태에서 모티브를 얻은 기법으로서, 입력 노드와 은닉 노드, 출력 노드를 구성하여 복잡한 수치예측 문제를 해결할 수 있도록 하는 분석 기법	블랙박스기법
서포트 벡터 머신 (Support Vector Machine, 혹은 Support Vector Regression)	분류문제에서는 서로 다른 분류에 속한 데이터 간의 간격(마진)을 최대화하는 초평면을 찾는 것이라면 수치예측문제에서의 서포트 벡터 머신은 데이터 점들의 분류가 아닌, 데이터 점들을 잘 적합할 수 있도록 가장 많은 데이터 점을 포함하는 튜브를 찾음	선형 및 비선형 (커널트릭)
랜덤 포레스트 (Random Forest)	주어진 데이터로부터 여러 개의 다양한 의사결정트리를 만들어 각 의사결정트리의 예측결과를 평균 내고 그 평균값을 최종결과로 결정하는 앙상블 형태의 기법(분류문제에서는 각 예측 분류결과를 투표하여 과반수 이상인 분류결과를 최종결과로 도출함)	앙상블 모형

12.4 모델링과 예측

수행 내용 / 수치 예측 목적의 머신러닝 수행하기

수행 순서

1 비즈니스 목적에 적합한 데이터 세트 추출 및 변환과 전처리가 완료된 데이터를 준비한다.
본 실습에 사용할 데이터는 R에서 다변량 분석 및 머신러닝을 위해 사용되는 MASS 패키지 내에 포함된 Boston 데이터 세트를 사용하기로 한다. Boston 데이터 세트는 미국 Boston 교외 506개 지역의 주택가격의 중앙값이(medv) 기록되어 있는 데이터 세트로 범죄율(crim), 주택당 평균 방의 개수(rm), 평균 주택 연령(age), 흑인 인구 비율(black), $10,000 당 세금 비율(tax) 등 13개 설명변수를 가지고 있다. 다음은 Boston 데이터의 일부분이다.

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

2 수치 예측 목적의 머신러닝 기반 분석을 위해 적합한 분석 기법을 선정한다.
수치 예측 목적의 다양한 머신러닝 기법이 있으나, 기법들이 대부분 앞서의 분류목적의 머신러닝 기법들과 유사하거나 중복되는 측면이 있으므로, 여기서는 수치예측 목적의 다중회귀분석, 의사결정트리 기법, 인공 신경망 기법, 랜덤 포레스트 기법을 적용해 본다.

3 각 기법을 사용하여 수치예측을 실시한다.
주택가격 중앙값(medv)을 목표변수로 하고 나머지 13개 변수를 설명변수(독립변수)로 하여 다양한 수치예측 머신러닝 기법을 적용해 보고자 한다.

12.4 모델링과 예측

Console C:/RSources/

```
> # 10-fold cross-validation for linear regression
> ps = select(x, Rating, Category, Size, Content.Rating)
> library(caret)
> data = ps[sample(nrow(ps)), ]
> k = 10
> q = nrow(data)/k
> l=1:nrow(data)
>
> te_total = c()
> pe_total = c()
> for(i in 1:k) {
+   test_list = ((i-1) * q+1) : (i * q)
+   testData = data[test_list,]
+   train_list = setdiff(l, test_list)
+   trainData = data[train_list, ]
+
+   m = lm(Rating~., data = trainData)
+   # print(residuals(m))
+   te = deviance(m)/nrow(trainData)        # mean squared error
+   te_total = c(te_total, te)
+
+   prd = predict(m, newdata = testData)
+   pe = mean((prd-testData$Rating)^2)       # mean squared error
+   pe_total = c(pe_total, pe)
+ }
```

```
> (te_total)
 [1] 0.2914398 0.2942018 0.2942018 0.2942018 0.2942018 0.2960985 0.2942018
 [8] 0.2942018 0.2942018 0.2942018
> (pe_total)
 [1] 0.3209624 0.3211810 0.3077886 0.2974638 0.3218347 0.2789736 0.3161342
 [8] 0.2419674 0.2650408 0.2778200
```

12.4 모델링과 예측

Console C:/RSources/

```
> # 10-fold cross-validation for linear regression
> #ps = select(x, Rating, Category, Size, Content.Rating)
> ps = select(x, Rating, Category)
> library(caret)
> data = ps[sample(nrow(ps)), ]
> k = 10
> q = nrow(data)/k
> l=1:nrow(data)
>
> te_total = c()
> pe_total = c()
> for(i in 1:k) {
+   test_list = ((i-1) * q+1) : (i * q)
+   testData = data[test_list,]
+   train_list = setdiff(l, test_list)
+   trainData = data[train_list, ]
+
+   m = lm(Rating~., data = trainData)
+   # print(residuals(m))
+   te = deviance(m)/nrow(trainData)        # mean squared error
+   te_total = c(te_total, te)
+
+   prd = predict(m, newdata = testData)
+   pe = mean((prd-testData$Rating)^2)       # mean squared error
+   pe_total = c(pe_total, pe)
+ }
```

```
> (te_total)
 [1] 0.2960331 0.2959144 0.2959144 0.2959144 0.2959144 0.2958140 0.2959144
 [8] 0.2959144 0.2959144 0.2959144
> (pe_total)
 [1] 0.2977961 0.3294332 0.3554719 0.2534552 0.3096741 0.2995680 0.3554783
 [8] 0.2446201 0.2784609 0.2446043
```

12.4 모델링과 예측

Console C:/RSources/

```
> # 결정 트리의 모델
> library(rpart)                          # rpart 라이브러리
>
> r = rpart(Rating~., data = trainData)                  # 모델 학습
> prd1 = predict(r, newdata = testData)                  # 모델에 으한 예측
>
> # 랜덤 포스트 모델(랜덤 포스트 라이브러리
> library(randomForest) # 랜덤 포스트 라이브 러리
>
> f = randomForest(Rating~., data = trainData)  # 모델 학습
> prd2 = predict(f, newdata = testData) # 모델에 으한 예측
>
>
> # 랜덤 포스트 라이브러리(SVM 라이브러리)
> library(e1071)                  # SVM 라이브러리
>
> f = svm(Rating~., data = trainData)           # 모델 학습
> prd3 = predict(f, newdata = testData)         # 모델에 으한 예측
```

12.4 모델링과 예측

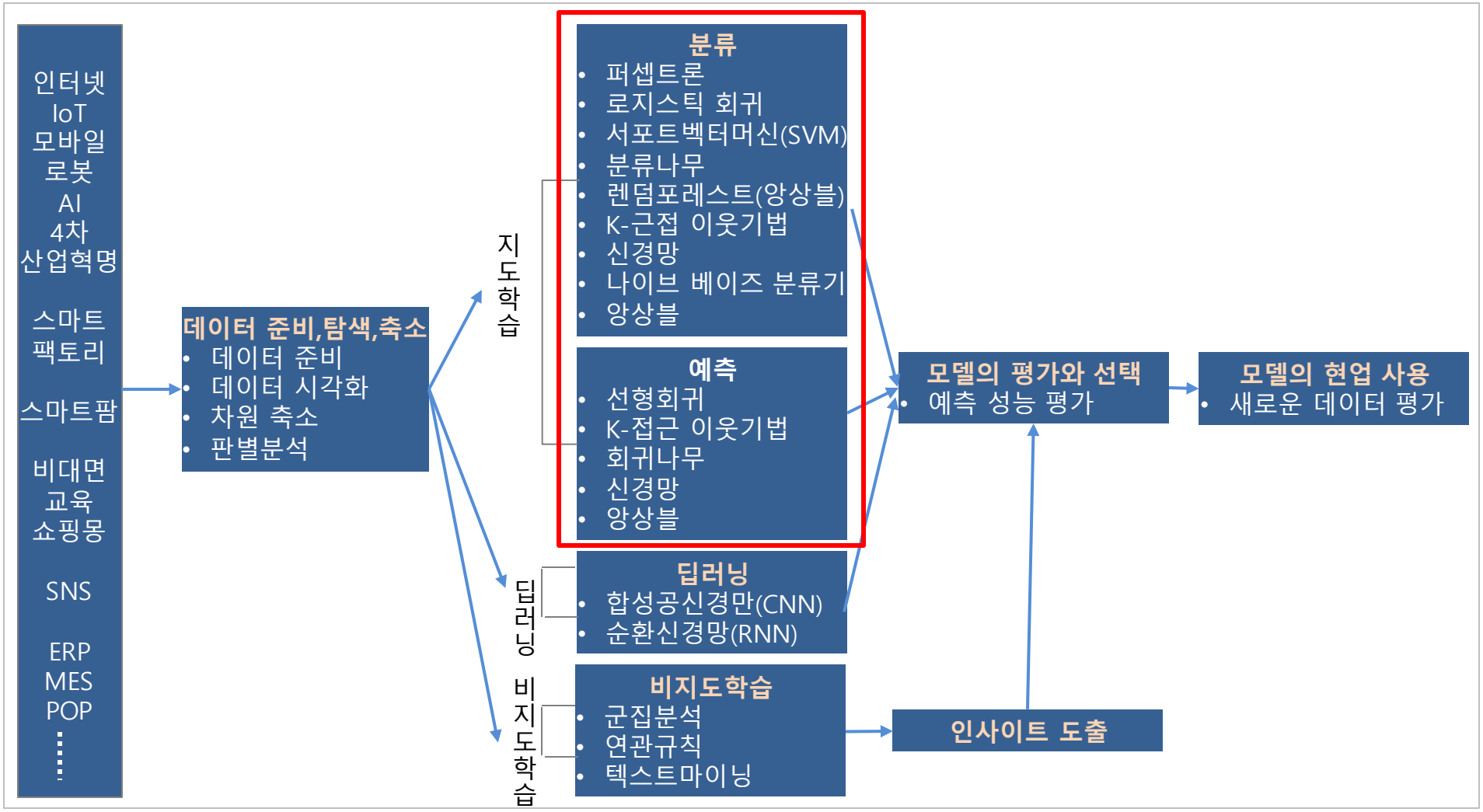
```
> # ---------- svm 모델 -------------
> library(e1071)                # SVM 라이브러리
> ps = select(x, Rating, Category, Size, Content.Rating)
에러: Can't subset columns that don't exist.
x Column `Rating` doesn't exist.
Run `rlang::last_error()` to see where the error occurred.
> data = ps[sample(nrow(ps)), ]
> k = 10
> q = nrow(data)/k
> l=1:nrow(data)
>
> te_total = c()
> pe_total = c()
> te_total1 = 0
> pe_total1 = 0
> for(i in 1:k) {
+   test_list = ((i-1) * q+1) : (i * q)
+   testData = data[test_list,]
+   train_list = setdiff(l, test_list)
+   trainData = data[train_list, ]
+
+   f = svm(Rating~., data = trainData)
+   te = deviance(f)/nrow(trainData)        # mean squared error
+   te_total = c(te_total, te)
+   te_total1 = te_total1 + te
+
+   prd = predict(f, newdata = testData)
+   pe = mean((prd-testData$Rating)^2)      # mean squared error
+   pe_total = c(pe_total, pe)
+   pe_total1 =pe_total1 + pe
+ }
>
> pe_total1=pe_total1/k
> pe_total1
[1] 0.311252
> te_total
numeric(0)
> pe_total
 [1] 0.2951518 0.4152235 0.2712578 0.2973073 0.3151249 0.3017356 0.3319172
 [8] 0.3129575 0.3080545 0.2637895
```

12.4 모델링과 예측

알고리즘(모델 기법)에 따른 성능

설명 변수 조합	모델 성능 (예측 값의 mse)			
Category				
Size				
Content.Rating				
Category + Size				
Category + Content.Rating				
Size + Content.Rating				
Category + Size + Content.Rating	0.29683	0.29926	0.29677	0.31125

요약


인터넷
IoT
모바일
로봇
AI
4차
산업혁명
스마트
팩토리
스마트팜
비대면
교육
쇼핑몽
SNS
ERP
MES
POP
데이터 준비,탐색,축소
데이터 준비
데이터 시각화
차원 축소
판별분석
지도학습
분류
퍼셉트론
로지스틱 회귀
서포트벡터머신(SVM)
분류나무
렌덤포레스트(앙상블)
K-근접 이웃기법
신경망
나이브 베이즈 분류기
앙상블
예측
선형회귀
K-접근 이웃기법
회귀나무
신경망
앙상블
딥러닝
딥러닝
합성공신경만(CNN)
순환신경망(RNN)
비지도학습
비지도학습
군집분석
연관규칙
텍스트마이닝
모델의 평가와 선택
예측 성능 평가
모델의 현업 사용
새로운 데이터 평가
인사이트 도출

Thank you


continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
90
70
60
50
40
300
1000
3000
10000
30000
gdpPercap