빅분기 9월 추가/수정 내용

2. 공공데이터와 같이 조직 외부의 데이터를 사용할 때의 장점으로 가장 적절한 것은?

1. 데이터 선택의 폭이 넓어진다
2. 데이터 제공을 받기 위해 외부 조직을 방문한다
3. 이용 가격이 저렴하다
4. 데이터 소유권을 가질 수 있다

조직 외부의 데이터 사용할 때의 장점
- 데이터 선택의 폭이 넓어진다
- 이용 가격은 무료/유료가 있으며, 이용 가격은 다양함 (23년 9월7일수정)
- 비용이 절감된다 (비용을 지불하더라도 직접 데이터를 수집, 생산해서 사용하는 것보다 비용이 절감됨)
- 신속한 의사결정에 도움이 된다
- 다양한 관점과 새로운 통찰력을 제공하고 새로운 아이디어와 혁신을 유발할 수 있다

설명을 수정했습니다.
23년 9월 7일

[S1-03] 3-1. 분석 마스터플랜 & ISP

9월 4일 내용 추가

분석 마스터 플랜이란?

- 일반적인 ISP 방법론을 활용하되, 데이터 분석 기획의 특성을 고려하여 수행함
- 기업에서 필요한 데이터 분석과제를 빠짐없이 도출한 후 과제의 우선순위를 결정하고 단기 및 중/장기로 나누어 계획을 수립하는 것
- 분석 마스터 플랜의 순서 : "중장기 마스터 플랜 수립 – 단기적인 세부 이행계획 수립 – 과제별 우선순위 설정"
- 분석 마스터 플랜의 모든 단계를 반복하기보다 데이터수집 및 확보와 분석 데이터를 준비하는 단계를 순차적으로 진행하고, 모델링 단계는 반복적으로 수행하는 혼합형을 많이 적용함

ISP(Information Strategy Planning, 정보 전략 계획)

- 기업의 경영목표 달성에 필요한 전략적 주요 정보를 포착하고, 주요 정보를 지원하기 위해 전사적 관점의 정보 구조를 도출하며, 이를 수행하기 위한 전략 및 실행 계획을 수립하는 전사적인 종합추진 계획
- 정보기술, 정보시스템을 전략적으로 활용하기 위해 조직 내/외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항 분석하여 시스템 구축 우선순위를 결정하는 등 중장기 마스터 플랜을 수립하는 절차
- 기업 및 공공기관에서는 시스템의 중장기 로드맵을 정의하기 위해 수행

[S3-03] 1-4-4 분산분석표 분석

- Cars 데이터에서 속도(speed)와 제동거리(dist)의 관계를 회귀모형으로 추정한 것이다.

```
> out <- lm(dist~speed, data=cars)
> anova(out)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value

MSE = 오차 분산의 불편추정량

Source	Df	Sum Sq	Mean Sq	F value
speed	k-1	SSR	MSR=SSR/(k-1)	MSR / MSE
Residuals	N-k	SSE	MSE=SSE/(N-k)	

- 회귀계수는 유의수준 5%에서 유의하다
- 관측치는 (48 + 1) + 1 = 50 이다 (Residuals Df + speed Df + 1개)
- 결정계수 = SSR/SST= $\frac{21186}{21186 + 11354} = 0.651$
- 오차 분산의 불편추정량은 'MSE' 오차제곱평균으로, Mean Sq 와 Residuals가 교차되는 지점에 236.5 이라고 써 있음

[S3-03] 1-4-5 분산분석표 분석

Wage 데이터셋을 사용한 anova 분석의 예

```
> A <-lm(wage~health + jobclass, data=Wage)
> anova(A)
Analysis of Variance Table

Response: wage
            Df  Sum Sq Mean Sq F value    Pr(>F)
health       1  121187  121187  74.153 < 2.2e-16 ***
jobclass     1  202930  202930 124.170 < 2.2e-16 ***
Residuals 2997 4897969    1634
```

p-value

MSE = 오차 분산의 불편추정량

- 모든 회귀계수는 유의수준 5%에서 유의하다
- 관측치는 (2997 + 1 + 1) + 1 = 3000 이다
- 결정계수 = SSR/SST
- $\frac{121187 + 202930}{121187 + 202930 + 4897969} = 0.062067$
- MSE => 1634

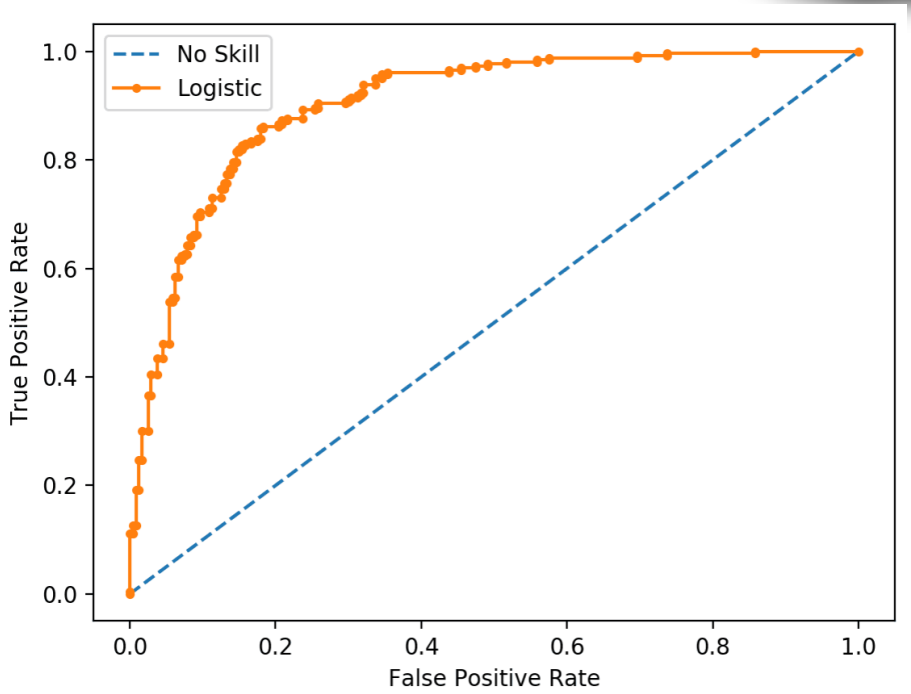
```
> A <-lm(wage~education + jobclass, data=Wage)
> anova(A)
Analysis of Variance Table

Response: wage
             Df  Sum Sq Mean Sq F value    Pr(>F)
education     4 1226364  306591 230.901 < 2.2e-16 ***
jobclass      1   20273   20273  15.268 9.535e-05 ***
Residuals  2994 3975448    1328
```

- 모든 회귀계수는 유의수준 5%에서 유의하다
- 관측치는 (2994 + 4 + 1) + 1 = 3000 이다
- 결정계수 = SSR/SST
- $\frac{1226364 + 20273}{1226364 + 20273 + 3975448} = 0.2387$
- MSE => 1328

14. ROC Curve에 대한 설명으로 올바르지 않은 것은?

1. 민감도가 0, 특이도가 1인 점을 지난다
2. 민감도가 1, 특이도가 0인 점을 지난다
3. 특이도가 증가하는 그래프이다
4. 가장 이상적인 그래프는 민감도가 1, 특이도가 1인 점을 지난다



ROC Curve
- X축이 FP rate, Y축이 TP rate인 그래프
- FP rate = 1 – specificity(특이도)
- 특이도 = 1 – FP rate
- FP Rate, 민감도(TP rate)가 증가하는 그래프

특이도 = 1 – FP rate 로 수정 (230829)

추가예정입니다!

[S4-02] 1-4 하이퍼 파라미터(Hyper-parameter) 최적화

- 하이퍼 파라미터의 최적값을 찾는데 도움이 되는 도구(자동이 아님에 주의!)

Manual Search	▪ 사용자가 수치를 여러 가지로 변경하여 적용해 보고 가장 좋은 성능을 갖는 경우를 찾아내는 방법 ▪ 단순한 방법이지만, 경험에 따라 최적의 조합을 찾아야 하므로 비효율적인 면이 있음
Grid Search	▪ 최적의 값이 있을 것으로 예상되는 구간과 간격 또는 목록을 함수로 전달해 차례로 적용해 보고 결과 목록에서 가장 좋은 성능을 보이는 경우를 찾는 방법 ▪ 구간, 목록에 있는 모든 경우에 대한 조합을 적용해 보기 때문에 가짓수가 많으면 시간이 오래 걸리며, 균일 간격, 목록 내(후보군)에 최적의 값이 존재하지 않을 수 있음
Random Search	▪ 최적의 값이 있을 것으로 예상되는 범위(Min, Max)을 정해두고 범위 내에서 무작위 값을 반복적으로 추출하여 최적의 조합을 찾는 방법 ▪ Grid Search는 정해진 선택지 중 최적의 값, Random Search는 예상 범위는 있지만 값은 Random으로 해서 최적의 값을 찾음
Bayesian Optimization	▪ 매 회 새로운 hyperparameter 값에 대한 조사를 수행할 시 '사전 지식'을 충분히 반영하면서, 동시에 전체적인 탐색 과정을 체계적으로 수행할 수 있는 방법론 ▪ 대체모델 (Surrogate Model) : 현재까지 조사된 입력값-함수결과값 점들을 바탕으로 목적 함수의 형태에 대한 확률적 추정을 수행하는 모델 ▪ 획득 함수(Acquisition Function) : 대체모델의 결과를 이용해 최적해를 찾는 데 유용한 다음 입력값 후보 추천