Data Science

# 분석 결과의 해석

## 분석 결과의 해석 및 활용

강사 양석환

# 분석 결과의 해석

# 분석 모델 별 결과 해석

- **회귀 모델의 해석**
  - **모델의 평가 지표와 해석을 위한 지표가 거의 같음**
    - 잔차 계열 지표: MAE, MSE, MAPE, RMSE, RMSLE 등
    - 결정계수 계열 지표: $R^2$, 수정된 $R^2$ 등

- **분류 모델의 해석**
  - **모델의 평가 지표와 해석을 위한 지표가 거의 같음**
    - 혼동행렬 기반 지표: 정확도, 정밀도, 재현율, F1-Score, ROC & AUC

# 분석 모델 별 결과 해석

- 군집분석 모델의 해석
  - 군집 그룹의 통계량을 요약하고 관측치의 공통점과 변동성을 확인함
  - 연속형 변수의 경우: 평균 또는 중앙값을 계산
  - 범주형 변수의 경우: 범주별로 각 군집의 분포를 사용

  - 평가 지표 및 해석 지표
    - 외부평가(External Evaluation)
    - 내부평가(Internal Evaluation)
    - 팔꿈치 기법(Elbow Method)
    - 실루엣 기법(Silhouette Method)

# 분석 모델 별 결과 해석

- **군집분석 모델의 해석**
  - **평가 지표 및 해석 지표**
    - 외부평가(External Evaluation)
      - 얼마나 유사하게 군집화가 되었는지 확인함
      - 자카드 지수(집합 간의 유사도 측정)를 활용함

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

    - 내부평가(Internal Evaluation)
      - 적절한 군집(클러스터) 개수 결정
      - Dunn Index(군집 간 거리가 멀수록, 군집 내부 분산 값이 작을수록 좋은 군집화 결과 반영)

$$\text{일반계산} = \sqrt{\frac{n}{2}} \quad \text{(n = 총 데이터 개수)}$$

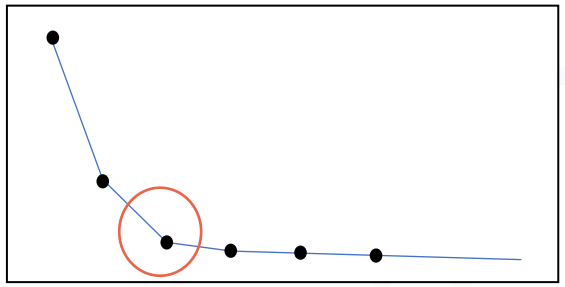
$$Dunn\ Index = \frac{min_{1 \le i \le j \le n} d'(i,j)}{max_{1 \le k \le n} d'(k)}$$

$$= \frac{\text{군집 간 거리 최소값}}{\text{군집내 요소간 거리 최대값}}$$

# 분석 모델 별 결과 해석

- **군집분석 모델의 해석**
  - **평가 지표 및 해석 지표**
    - 팔꿈치 기법(Elbow Method)
      - 팔꿈치(Elbow) 모습을 나타내는 곳의 값을 적절한 군집(클러스터) K값으로 지정함

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



    - 실루엣 기법(Silhouette Method)
      - 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐져 있다는 의미로 -1~1의 값을 가짐
      - 1에 가까울수록 최적화가 잘 되어 있는 것으로 해석함

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

a(i) = 데이터의 응집도를 나타낸 값

b(i) = 클러스터 간의 분리도

# 분석 모델 별 결과 해석

- 연관분석 모델의 해석
  - 두 개 또는 그 이상의 품목들 사이의 상호 관련성으로 해석
  - 지지도, 신뢰도 및 향상도가 높은 규칙들을 도출하고 규칙 별로 최소 기준점을 적용
  - 빈발 집합을 고려하여 연관규칙을 생성하는 Apriori 알고리즘 사용

  - 평가지표 및 해석지표
    - 지지도(Support)
    - 신뢰도(Confidence)
    - 향상도(Lift)

# 분석 모델 별 결과 해석

- **연관분석 모델의 해석**
  - **평가 지표 및 해석 지표**
    - 지지도(Support)
      - 전체 거래에서 품목 A와 B가 동시에 포함된 거래의 수(N=전체 거래 수)

$$Support = \frac{A \cap B}{N}$$

    - 신뢰도(Confidence)
      - 품목 A가 구매되었을 때 품목 B가 추가로 구매될 확률(조건부 확률)

$$Confidence = \frac{A \cap B}{A}$$

    - 향상도(Lift)
      - 품목 A를 구매할 때 B도 추가로 구매하는 지의 연관성을 파악하는 비율

$$Lift = \frac{A \cap B * N}{A * B} > 1 \quad \text{양의 상관관계}$$

$$= 1 \quad \text{독립적인 관계}$$

$$< 1 \quad \text{음의 상관관계}$$

# 비즈니스 기여도 평가

- 데이터 분석의 목적과 의의
  - 데이터 분석의 근본적인 목적은 과거의 데이터를 토대로 미래를 분석하는 것
  - 데이터 분석은 비즈니스에 도입, 활용함으로써 의사결정, 운영 프로세스의 효율화, 개선점 도출이 목적
  - 따라서 데이터 분석은 이러한 목적에 대한 기여도 평가가 필요함

- 분석 결과의 기여도 평가
  - 일반적으로 ROI(Return Of Investment, 투자 수익률) 또는 업무 효율성에 대한 비율로 측정
  - 그 외에 투자 이후 회수 기간과 전략적 기여도 기준 IT ROI 평가로 비용-효과 분석(Cost-Benefit Analysis), 정보경제학(Information Economics) 기반 방법론 등으로 기여도 평가 가능

# 비즈니스 기여도 평가

- 분석 결과의 기여도 평가
  - ROI(Return Of Investment, 투자수익률)
    - 투자한 자본에 대한 수익/손실의 비율
    - $ROI = \frac{\text{총체적인 금전적 이익} - \text{소요된 비용}}{\text{소요된 비용}} \times 100$
  - 업무 효율성 향상에 대한 비율
    - 분석 과제와 연관된 업무 효율성 향상 항목의 측정지표 기준 수립을 통해서 산정함

# 분석 결과의 활용

# 분석 모형 전개

- **빅데이터 분석 방법론**
  - **데이터 마이닝을 위한 방법론을 프로젝트 특성에 맞추어 적용함**
  - **대표적인 방법론: CRISP-DM, SEMMA, KDD 등**

  - **빅데이터 분석 방법론 참조모델 (데이터산업진흥원)**
    - 분석기획(Planning)
    - 데이터 준비(Preparing)
    - 데이터 분석(Analyzing)
    - 시스템 구현(Developing)
    - 평가 및 전개(Deploying)

# 분석 모형 전개

- 빅데이터 분석 방법론
  - CRISP-DM(Cross Industry Standard Process for Data Mining) 방법론
    - 비즈니스 이해(Business Understanding)
    - 데이터 이해(Data Understanding)
    - 데이터 준비(Data Preparation)
    - 모델링(Modeling)
    - 평가(Evaluation)
    - 전개(Deployment)

1996년 유럽연합의 ESPRIT 프로젝트에서 시작한 방법론
총 6단계로 구성되며 빅데이터 프로젝트에서 보편적으로 사용됨

# 분석 모형 전개

- 빅데이터 분석 방법론
  - SEMMA(Sampling Exploration Modification Modeling Assessment) 방법론
    - 샘플링(Sampling)
    - 탐색(Explore)
    - 전처리(Modify)
    - 모델링(Modeling)
    - 평가(Assess)

SAS사의 주도로 통계적 분석에 중심을 두고 있는 방법론
총 5단계로 구성됨

# 분석 모형 전개

- 빅데이터 분석 방법론
  - KDD(Knowledge Discovery in Database) 방법론
    - 데이터 추출(Select)
    - 전처리(Preprocessing)
    - 변환(Transformation)
    - 데이터 마이닝(Data Mining)
    - 해석/평가(Interpretation/Evaluation)

1996년 Fayyard가 정리한 데이터마이닝 프로세스
주로 데이터베이스 중심 시스템을 대상으로 적용됨
총 5단계로 구성됨

# 분석 모형 전개

- 전개(Deployment) 단계의 역할
  - 모델의 전개: 개발된 모델을 적용하여 결과를 확인하고 계속적인 관리를 위한 방법을 제시하는 단계
  - 방법론에 따라 명확하게 포함되지 않는 경우도 있음
  - 그러나 데이터 분석 프로젝트가 성공적으로 완료되기 위해 꼭 필요한 프로세스임

  - 개발된 모델의 성능은 실제 동작하는 운영 데이터의 특성과 품질에 따라 많은 영향을 받기 때문에 이를 주기적으로 모니터링하고 성능 개선을 위한 노력을 기울여야 함

# 분석 모형 전개

- 전개(Deployment) 단계의 역할
  - 모델의 전개 단계에서 이루어지는 작업
    - 분석 결과의 활용 계획 수립
      - 분석 결과를 어떻게 업무에 반영할 것인지에 대한 액션 플랜 수립
      - 업무성과를 지속적으로 모니터링하는 방안 수립
      - 분석 결과 활용 방안에 대한 시나리오 개발
      - 업무 적용 및 효과 검증
    - 분석 결과의 적용과 보고서 작성
      - 분석 결과 적용과 성과 평가
      - 프로젝트 진행 과정의 모든 산출물 및 프로세스 정리 → 자산화
      - 최종보고서 작성
        - 프로젝트 개요(목표, 범위, 일정, 비용), 수행 조직, 단계별 산출물 요약, 성과 평가 결과, 모니터링 및 개선 계획

# 분석 모형 전개

- **전개(Deployment) 단계의 역할**
  - **모델의 전개 단계에서 이루어지는 작업**
    - **분석 모형 모니터링**
      - **분석 모니터링의 주요 대상**
        - 서비스: 분석과제 발굴, 활용방안 마련, 성과관리 등
        - 분석 모델: 분석 알고리즘 주기, 변수, 소스(데이터 원천) 등
        - 데이터: 현 시점의 현행화 데이터 확인
      - **분석 서비스 유지관리의 주요 대상**
        - 정책/제도: 조직의 정책/제도 개발 및 적용
        - 업무: 신규 업무 반영, 기존 업무 업그레이드
        - 관련 시스템: 관련 시스템 변경 사항 반영
        - 인력: 업무 역량, 책임과 역할, 교육 훈련 등

# 분석 모형 전개

- **전개(Deployment) 단계의 역할**
  - **모델의 전개 단계에서 이루어지는 작업**
    - **분석 모형 리모델링**
      - **분석 모형 리모델링 과정**
        - 서비스 운영 과정에서 지속적인 데이터 유입, 정책/환경의 변화 등으로 분석 모델의 성능 하락 가능성이 높아짐
        - 데이터 수집, 전처리, 분석 방법론, 분석결과까지 과제 전반에 대하여 보완, 개선 방안 도출 → 리모델링 계획 수립
      - **분석 모형 리모델링 방법**
        - 분석 목적에 기반한 가설 및 추정 방법에 대한 재검토
        - 분석용 데이터의 범위 및 품질 검토
        - 과대적합과 과소적합 방지를 위한 알고리즘 개선
        - 분석 알고리즘과 매개변수 최적화
        - 분석 모형 융합과 재결합