Data Science

데이터 모델링

강사 양석환

분석 모형 설계 (모형의 선정과 구축)

데이터 분석 모형의 선정

- **데이터 분석의 근본적인 목적**
  - **과거의 데이터를 토대로 원인에 대해 분석하고 그 결과로 미래를 예측하는 것**
  - **데이터는 후행성의 성격을 지니지만 선행성의 성격도 동시에 가지고 있음**
    - 예: 사람들이 포털에서 검색하는 것은 무엇인가를 알고 싶기 때문이며, 검색 키워드라는 후행성 데이터로 '왜 사람들이 그것을 알고 싶어할까?'라는 분석을 통해 미래에 일어날 일을 예측할 수 있음

특히 빅데이터의 분석은 통계에서 분석했던 방식과 함께 기존의 통계방식으로 분석할 수 없던 것도 분석이 가능하다.

데이터 분석 모형의 선정

- **데이터 분석 목적의 분류**
  - **의사 결정**
    - 여러 대안 중에서 하나의 행동을 고르는 일을 해 내는 정신적 지각 활동
    - 최종적으로 하나의 선택을 가지게 되고 이로 인한 결과가 도출됨
  - **불확실성 해소**
    - 의사 결정의 가장 큰 문제는 불확실성
    - 분석을 통해 불확실성을 제어한다면 시장 대응에 큰 도움이 될 수 있음
  - **요약**
    - 데이터 요약을 통해 현 상황을 쉽고 빠르게 파악하는 것이 가능하며, 다음 대응할 방안 등을 생각할 수 있음

데이터 분석 모형의 선정

- **데이터 분석 목적의 분류**
  - **인과관계 파악**
    - 단순히 요약 기능만 사용하기보다 데이터 간 연관관계 분석으로 원인과 결과를 파악할 수 있음
    - 인과관계 파악으로 세부적인 판단을 내릴 수 있음
  - **예측**
    - 원인과 결과로 어떤 패턴을 파악하게 된다면 다음에 생길 결과에 대한 예측 또한 가능함
    - 반드시 같은 패턴으로 이어지는 결과가 생기지는 않지만 향후 미래에 생길 결과에 대한 대비가 가능해짐

가트너 그룹에서 제시한 빅데이터 분석의 목적(2015)

- 고객 인사이트 (Customer Insight)
- 제품 및 절차 효율성 (Product & Process Efficiency)
- 디지털 제품 및 서비스 (Digital Products & Service)
- 운영의 탁월성 (Operational Excellence)
- 디지털 마케팅 (Digital Marketing)
- 위기 관리 시스템 (Risk Management and Compliance)

데이터 분석 모형의 선정

- **데이터 분석 모형이란?**
  - 분석 목표에 따라 데이터 특성을 도출하고, 가설 수립에 따라 전체적인 분석 방향을 정의하는 모델

- **데이터 분석 모형 정의 시 사전 고려사항**
  - 데이터 분석 모형을 정의하기 전에 분석이 실제 추진될 수 있을지 가능성을 타진하는 것이 중요함
  - 분석을 진행하기 이전 상황에 맞는 평가 기준표, 테이블을 작성하여 항목별로 점수를 부여하고 총점을 매긴 후 분석 모형 정의의 가능성을 판별할 수 있음
  - 추진 시급성과 구현 가능성만으로 데이터 분석 모형 정의를 위한 사전 판별 기준 활용이 가능함
  - 데이터 분석 모형 정의에 필요한 데이터가 충분히 확보되어 있는지를 판단하여 관련 과거 분석 사례 또는 솔루션을 최대한 활용할 수 있는지 검토한다면 보다 효율적인 데이터 분석 모형 설계를 진행할 수 있음

데이터 분석 모형의 선정

- **데이터 분석 모형 정의 시 사전 고려사항**
  - **데이터 분석 모형 정의와 판별을 위한 평가 기준**

평가 기준	판단 근거
필요성	개인이나 기관 관점에서 분석 과제가 필요한지 판단
파급효과	정성적, 정량적 기대효과의 정도 판단
추진 시급성	당장 해소되어야 할 사회 현안 여부 판단, 장기과제 성격 분리
구현 가능성	과제를 구현함에 있어서 어려움이 없는지 현실성 판단
데이터 수집 가능성	공공기관 협조나 데이터 확보, 데이터 구매 등 제약사항 판단
모델 확장성	과제가 시범과제로 끝나지 않고 전체 데이터 모델로 확장 가능한지 판단

데이터 분석 모형의 선정

- **데이터 분석 모형 정의를 위한 접근 방법**
  - **상향식(Bottom-Up) 접근**
    - 문제 정의가 어려울 경우, 많은 양의 데이터 분석을 통해 인사이트를 도출함
    - 특정 영역을 지정하여 의사결정 지점으로 진행하는 과정에서 분석 과제를 발굴할 때 많이 사용됨
  - **하향식(Top-Down) 접근**
    - 문제 정의가 가능할 경우, 문제 탐색과 연관되어 비즈니스 모델, 외부 참조 모델, 분석 유스케이스 기반 모델로 발굴하는 방식을 적용할 수 있음
      - 비즈니스 모델: 어떻게 수익을 창출할 것인가에 대한 검증으로 문제 해결을 위한 분석과제를 발굴
      - 외부 참조 모델: 벤치마킹으로 분석 테마 후보 Pool을 구축, 선택
      - 분석 유스케이스 기반 모델: 문제에 대한 상세 설명과 해결 시의 효과에 대해 명시함으로써 구체적인 분석 과제를 도출

데이터 분석 모형의 선정

- **분석 모형의 종류**
  - **예측 분석 모형**
    - 어떤 일들이 발생할 것인가?
    - 적조 예측, 날씨 예측, 주가 예측, 범죄/위험 예측, 쇼핑 아이템 추천 등 과거, 현재까지의 데이터와 상황에 따른 가설에 기반하여 미래에 대한 현상을 사전에 분류하고 예측하는 모형
  - **현황 진단 모형**
    - 과거에 어떤 상황이 왜, 어떻게 일어났는가? 그리고 현재는 어떠한 상태인가?
    - 과거 데이터를 통해 현재 상황을 객관적으로 진단하는 모형
    - 미래 예측이 아닌 현재를 이해하기 위해 활용함
  - **최적화 분석 모형**
    - 어떻게 하면 원하는 결과가 일어날 수 있을까?
    - 제한된 자원, 환경 내에서 최대의 효용성, 이익과 같은 결과를 생성하기 위해 분석 모델을 최적화하는데 중점을 둠

데이터 분석 모형의 선정

- **데이터 분석 모형의 선정**
  - 분석 기법 또는 분석 알고리즘을 적용하기 전에 분석 모형에 대한 선정이 필요함
  - 분석이 필요한 데이터 속성을 세부적으로 파악, 처리한 뒤에 분석 모형을 선정, 적합한 분석 기법을 선택함
  - 만약 데이터가 준비되어 있지 않다면 사전 분석 목적을 정확하게 파악해야만 문제 인식과 필요한 데이터의 준비에 따른 분석 모형 선정을 수월하게 진행할 수 있음

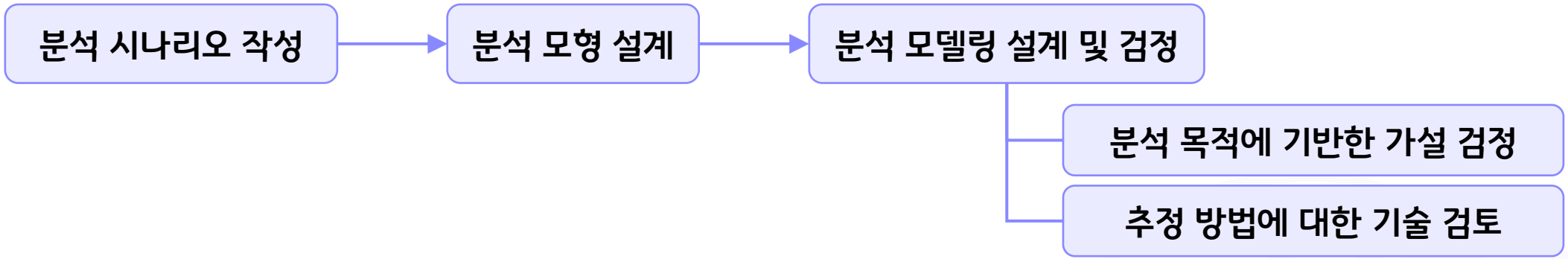
데이터 분석 모형의 선정

- **데이터 분석 모형의 선정 프로세스**

1. 문제요건 정의 또는 비즈니스 이해에 따른 대상 데이터 선정과 분석 목표/조건 정의
2. 데이터 수집, 정리 및 도식화
3. 데이터 전처리(데이터 정제, 종속/독립변수 선정, 데이터 변환, 데이터 통합, 데이터 축소 등)
4. 최적의 분석 모형 선정

데이터 분석 모형의 구축

- **분석 모형 구축 절차**



  - **분석 시나리오 작성**
    - 분석 시나리오 작성을 통해 분석 과정과 결과가 어떻게 활용되는지 명확하게 이해할 수 있음
    - 데이터 분석 대상 및 범위를 요구사항에 맞게 정의
    - 분석을 통해 해결할 수 있는 문제와 목표, 그리고 분석 목표 별 구현 모델과 예상 결과를 작성
    - 분석 과정에 필요한 데이터, 절차, 분석 기법 등의 세부 사항들을 정의
    - 데이터의 경우 사전 확보 및 유형 분석이 필요함
    - 기존에 잘 구현되어 활용되고 있는 유사 분석 시나리오 및 솔루션을 고려할 수 있음

데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모형 설계** (분석 대상 및 범위를 정하여 분석 목적을 구현하기 위한 분석 방법론을 설계하는 단계)
    - **분석 모형 설계 시 사전 확인 사항**
      - 필요한 데이터 항목이 정해졌는가?
      - 데이터 단위를 고려, 항목에 다른 표준화 방법을 정하였는가?
      - 데이터를 수집한 항목에 따라 단계별로 모델이 설계되었는가?
      - 분석 검증 통계 기법을 선정하였는가?
    - **분석 모델링 설계와 검정**
      - 분석 목적에 기반한 가설 검정 방법 수립
      - 추정 방법에 대한 기술 검토
      - 분석 모델링 설계와 검정 방법 수립

데이터 분석 모형의 구축

- 분석 모형 구축 절차
  - 분석 모형 설계
    - 분석 모델링에 적합한 알고리즘 설계
      - 비지도학습: 군집 분석, 연관성 분석, 오토 인코더 등
      - 지도학습: 의사결정트리, 랜덤 포레스트, 서포트 벡터 머신, 회귀분석 등
      - 준지도학습: 셀프 트레이닝, 적대적 생성 모델 등
      - 강화학습: Q-Learning, 정책경사(Policy Gradient, PG) 등
    - 분석 모형 개발 및 테스트
      - 모듈 기능 정의
      - 모듈 설계
      - 모듈 개발 결과물과 모델 설계의 일치 여부 확인
      - 모듈의 정상 동작 여부 검증

데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **분석 목적에 기반한 가설 검정 방법 (총 5단계의 절차)**

1. 유의수준 결정, 귀무가설과 대립가설 설정

- 유의수준: 통계적 가설검정에서 제1종 오류와 제2종 오류가 있을 때, 제1종 오류를 범할 확률이 최대 허용값
- 귀무가설: 직접 검정 대상이 되는 가설. 대립 가설에 상반되는 가설로서 기각될 것이라고 예상되는 가설을 말함
  ('표본의 관찰을 통해 모집단은 **할 것이다'라고 내린 가설)
- 가설 검정을 시행할때는 귀무가설이 옳다는 가정 하에 시작함.
  이것을 반대로 생각하면 진실일 가능성이 적어 처음부터 기각될 것이 예상되는 가설임
- 대립가설은 귀무가설이 기각될 때 받아들여지는 가설로 정의함

데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **분석 목적에 기반한 가설 검정 방법**

2. 검정통계량의 설정

- 검정통계량은 가설을 검정하기 위한 기준으로 사용하는 값을 의미함
- 검정통계량이 확률분포 상에 어디에 위치하는지에 따라 귀무가설을 기각하거나 또는 기각하지 않음

3. 기각역의 설정

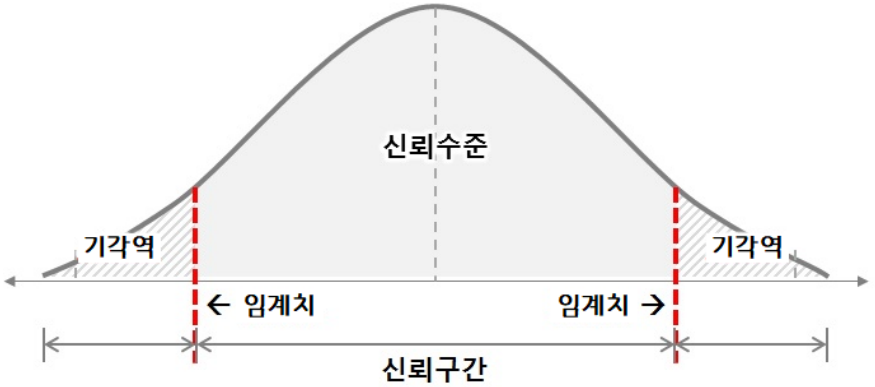
- 기각역은 확률분포에서 귀무가설을 기각하는 영역을 말함
- 기각역에 검정통계량이 위치하면 귀무가설을 기각함

데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **분석 목적에 기반한 가설 검정 방법**

4. 검정 통계량 계산

- 검정통계량의 계산식 : $\frac{(\text{표본평균} - \text{모평균})}{(\text{표본 표준편차})}$



- 신뢰수준: 가설을 검정할 때 어느 정도로 검정할 것인지에 대한 수준
- 유의수준: 가설을 검정할 때 일정 수준을 벗어나면 귀무가설이 오류라고 판단하는 수준

  (유의수준의 수학적 의미는 기각역들의 합이며, 1에서 신뢰수준을 뺀 값이기도 함)

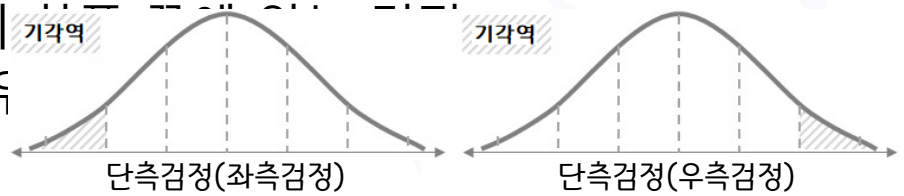
데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **분석 목적에 기반한 가설 검정 방법**

5. 통계적인 의사결정(가설 검정)

- 가설 검정에서의 검정 방법은 양측 검정과 단측 검정의 두 가지가 있음
- 양측검정
  - 귀무가설을 기각하는 영역이 양쪽에 있는 검정
  - 만약 대립가설이 ~가 아니다(크거나 작다)라면 이

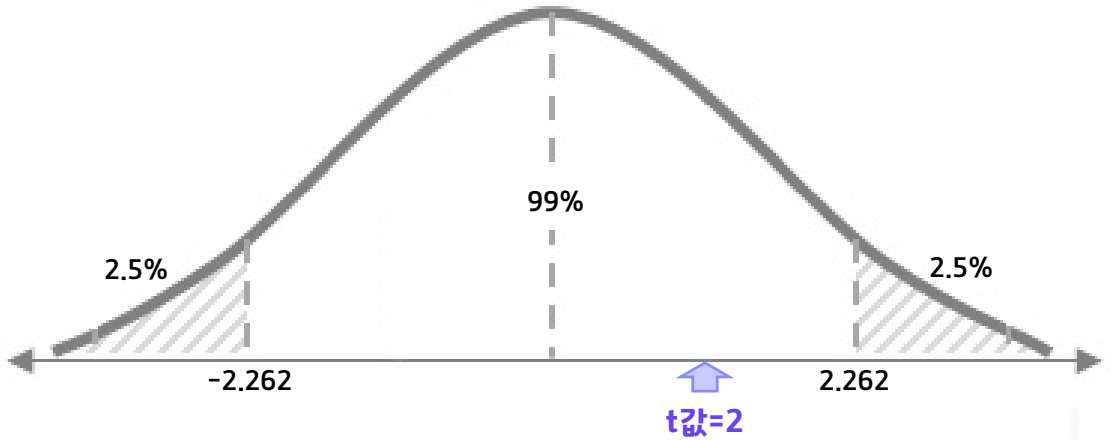
- 단측검정
  - 양측검정과 달리 귀무가설을 기각하는 영역이
  - 만약 대립가설이 ~보다 작다 또는 크다인 경우



- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **분석 목적에 기반한 가설 검정 방법**

5. 통계적인 의사결정(가설 검정)

- 통계적인 의사결정 단계에서는 계산한 검정통계량을 t값 분포도와 비교하여 기각역에 속하는지 아닌지를 판단함



검정통계량(t값)에 따른 귀무가설 기각 결정

데이터 분석 모형의 구축

- **분석 모형 구축 절차**
  - **분석 모델링 설계와 검정**
    - **추정 방법에 대한 기술 검토**
      - **전체적으로 데이터에 대한 전처리 과정을 마치게 되면 모형에 활용될 후보 변수와 후보 분석 모형에 사용할 알고리즘을 파악하게 됨**
      - **기초 통계, 데이터 검증, 데이터 정제 등의 데이터 변환 과정을 거치면 후보 변수는 전처리 과정에서 선정됨**
      - **분석 모형은 크게 분류예측 추천 등의 예측 분석, 과거 데이터를 기반하여 현재를 진단하는 현황 진단, 시뮬레이션과 제한된 환경 최적화를 모색하는 예측 최적화로 나누어짐**
        - 분석 모형을 선정하는 문제는 비즈니스 환경 여건이나 종속 변수의 유무에 따라 달라지기 때문에 종속 변수가 있는지 없는지를 살펴보아야 함
        - 종속변수가 없으면 사용할 수 있는 알고리즘이 군집과 원인분석, 이상치, 연관 법칙 등으로 제한됨
        - 변수의 속성에 따라서도 알고리즘의 선택이 달라짐