Data Science

데이터 분석

데이터 탐색

강사 양석환

데이터 전처리

데이터 전처리

- 데이터 전처리
  - 원시적인 형태의 데이터를 내가 원하는 형태로 변환하는 과정 및 작업으로 데이터 분석의 기초가 됨
  - 데이터 가공(Data Manipulation), 데이터 핸들링(Data Handling), 데이터 클리닝(Data Cleaning) 등의 이름으로도 불림

  - 데이터 전처리의 필요성
    - 수집한 데이터를 탐색해 보면 결측치, 틀린 값, 틀린 단위 등 손봐야 할 곳이 많음
    - 일반적으로 데이터는 그 데이터를 최초로 얻을 때의 목적에 맞게 구성되어 있기 때문에 이 데이터를 다른 목적으로 사용하려면 재가공이 필요함
    - 데이터가 처리하기 좋은 형식으로 되어 있는지, 새로 가공해야 하는지 등을 파악해야 함(상태 및 가치)

데이터 전처리

- **데이터 전처리의 일반적인 형태**
  - 필터링(Filtering): 필요한 데이터를 골라냄
  - 변환(Transformation): 데이터의 형식을 변경함
  - 통합(Integration): 여러 소스에서 온 데이터를 합침

- **데이터 전처리 예시**
  - 문자로 된 범주형 데이터의 경우, 처리를 편리하게 하기 위하여 숫자로 표현을 바꿈
  - 월요일은 1로, 화요일은 2로 인코딩 함
  - 수치 데이터의 분포를 정규화(Normalize) 함
  - 같은 대상에 대해서 10점 만점으로 처리한 것과 100점 만점으로 처리한 데이터를 같이 활용하기 위해서 동일한 분포로 변환

데이터 전처리

- 결측치 처리
  - 데이터셋에서 빠진 값(결측치, Missing Value)을 처리하는 작업
  - 데이터 전처리에서 가장 중요한 과정임
  - 결측치 처리 방법
    - 결측치가 포함되어 있는 항목을 모두 버림
    - 결측치를 적절한 값으로 대체함
      - 전체 평균값으로 대체
      - 인접한 값으로 대체 또는 인접한 값을 이용하여 추정치를 계산
    - 결측치를 NaN으로 표시하여 다음의 분석단계로 결측치 처리를 넘김

결측치를 함부로 대체하면
분석결과가 달라질 수 있으므로 주의가 필요함

예: 급여에 해당하는 데이터는 평균치나 0으로 대체하면 안됨!!!

결측치를 대체하는 경우, 이런 사실을 표시해주는 별도의 범주형 변수를 새로 정의하는 방법도 가끔 유용함

데이터 전처리

- 틀린(Invalid) 값 처리
  - 틀린 데이터란 잘못된 값이 들어있는 것을 의미함
  - 결측치는 값이 없는 것이므로 명백하게 발견할 수 있으나, 틀린 값은 바로 발견할 수 없음
    - 프로그램 등에 의해 일일이 어떤 기준이나 규칙에 따라 찾아내야 함
  - 틀린 값을 처리하는 방법
    - 틀린 값이 포함되어 있는 항목을 모두 버림
    - 틀린 값을 적절한 값으로 대체함
    - 분석단계로 틀린 값의 처리를 넘김

틀린 값의 처리도 결측치를 처리하는 방법과 같음

데이터 전처리

- 이상치(Outlier) 처리
  - 이상치란 값의 범위가 일반적인 범위를 벗어나 특별한 값을 가지는 것을 말함
  - 이상치를 찾아내는 것을 이상치 검출(Detection)이라고 부름

- 데이터 변환
  - 데이터를 분석하기 좋은 형태로 바꾸는 작업을 말함
  - 데이터의 분포를 고려하여 정규화 하거나 단위를 조정하거나 로그 스케일로 변환하는 등의 작업을 모두 데이터 변환이라고 함
  - 수치형→범주형으로 변환, 일반 정규화, Z-Score 정규화, 로그변환, 역수 변환 등

데이터 전처리

- 데이터 축소(Reduction)
  - 같은 정보량을 가지면서(유지하면서) 데이터의 크기를 줄이는 것
  - 대표적인 데이터 축소 방법: 주성분 분석(Principal Component Analysis, PCA):
    - 기존 데이터의 특징들을 대표하는 새로운 값을 추출하는 방법
- 샘플링(Sampling)
  - 구할 수 있는 전체 데이터 중에서 분석에 필요한 데이터를 취하는 것
  - 최소한의 샘플 데이터를 가지고 분석의 타당성을 조사하거나 분석 모델의 큰 방향을 결정할 때에도 필요함
  - 데이터 샘플링 시 주의할 점
    - 샘플링된 데이터가 전체 데이터의 특징을 계속 유지할 수 있도록 해야 함
    - 여러 소스에서 얻은 데이터를 다룰 때에는 각 소스에서 골고루 데이터를 취해야 하며, 여러 측면으로 균일해야 함

데이터 전처리

- 훈련 데이터와 테스트 데이터
  - 데이터 분석은 2단계의 절차가 필요함
    - 모델을 만드는 과정(Training, 훈련)
      - 각각의 모델을 구성하는 파라미터를 찾는 과정
    - 모델을 검증하는 과정(Test)
  - 훈련용 또는 테스트용 데이터를 준비할 때는 랜덤한 성질을 보장해야 함

데이터 정제

데이터 정제

- 데이터에 내재된 변수의 이해
  - 빅데이터 분석/전통적 통계분석을 통해 원하는 결과를 얻기 위해서는 모든 근간이 되는 자료의 이해가 필수
  - 데이터 관련 정의
    - 데이터(Data)
      - 이론을 세우는 기초가 되는 사실 또는 자료를 지칭함
      - 관심의 대상이 되는 사물이나 사건의 속성을 일정한 규칙에 의해 측정, 조사, 관찰하여 습득함
      - 컴퓨터와 연관되어 프로그램을 운용할 수 있는 형태로 기호화 · 수치화한 자료를 가리킴
    - 단위(Unit): 관찰되는 항목 또는 대상을 지칭함
    - 관측값(Observation): 각 조사 단위별 기록정보 또는 특성
    - 변수(Variable): 각 단위에서 측정된 특성 결과
    - 원자료(Raw Data): 표본에서 조사된 최초의 자료

데이터 정제

- 데이터의 종류
  - 데이터의 종류는 변수들의 집합인 자료의 종류와 그 특성을 동일하게 가지기 때문에 데이터의 종류에 따라 적용 방법론이 다양하게 변화할 수 있음

  - 단변량 자료(Univariate Data): 자료의 특성을 대표하는 특성 변수가 하나인 자료
  - 다변량 자료(Multivariate Data): 자료의 특성을 대표하는 특성 변수가 두 가지 이상인 자료
  - 질적 자료(Qualitative Data)
    - 정성적 자료 또는 범주형 자료라고도 부름
    - 자료를 범주의 형태로 분류함
    - 분류의 편의상 부여된 수치의 크기 자체에는 의미를 부여하지 않음
    - 명목 자료, 서열 자료 등이 질적 자료로 분류됨

데이터 정제

- 데이터의 종류
  - 질적 자료(Qualitative Data)
    - 명목 자료(Nominal Data)
      - 측정대상이 범주나 종류에 대해 구분된 것을 수치 또는 기호로 표시한 자료
      - 명목 자료 처리 시 사용가능한 연산자는 [≠, =]
      - 예: 전화번호의 국번·지역번호
    - 서열 자료(Ordinal Data)
      - 명목 자료와 비슷하지만 수치 또는 기호가 서열을 나타내는 자료
      - 서열 자료 처리 시 사용가능한 연산자는 [≠, =, ≤, ≥]
      - 예: 기록 경기의 순위 등 일반적인 순위를 나타내는 대부분의 자료

데이터 정제

- 데이터의 종류
  - 수치 자료(Quantitative Data)
    - 정량적 자료 또는 연속형 자료라고도 부름
    - 숫자의 크기에 의미를 부여할 수 있는 자료
    - 구간 자료, 비율 자료가 수치자료에 속함

데이터 정제

- 데이터의 종류
  - 수치 자료(Quantitative Data)
    - 구간 자료(Interval Data)
      - 명목 자료, 서열 자료의 의미를 포함하면서 숫자로 표현된 변수에 대해서 변수 간의 관계가 산술적인 의미를 가지게 한 자료
      - 비율의 의미가 부여될 수 없는 자료
      - 구간 자료 처리 시 사용가능한 연산자는 [≠, =, ≤, ≥, +, -]
      - 예: 온도
    - 비율 자료(Ratio Data)
      - 명목 자료, 서열 자료, 구간 자료의 의미를 모두 가지는 자료로서 수치화된 변수에 비율의 개념을 도입할 수 있는 자료
      - 비율 자료 처리 시 사용가능한 연산자는 [≠, =, ≤, ≥, +, -, ×, ÷]

데이터 정제

- 데이터의 종류
  - 시계열 자료(Time Series Data)
    - 일정한 시간 간격 동안에 수집된 자료 (예: 일별 주가 데이터)
  - 횡적 자료(Cross Sectional Data)
    - 횡단면 자료라고도 부름
    - 특정 단일 시점에서 여러 대상으로부터 수집된 자료
    - 한 개의 시점에서 여러 대상으로부터 취합하는 자료를 지칭함
  - 종적 자료(Longitudinal Data)
    - 시계열 자료와 횡적 자료의 결합
    - 여러 개체를 여러 시점에서 수집한 자료

데이터 정제

- **데이터의 정제**
  - **수집된 데이터를 대상으로 분석에 필요한 데이터를 추출하고 통합하는 과정**
  - **데이터 정제의 필요성**
    - 데이터로부터 원하는 결과나 분석을 얻기 위해서는 수집된 데이터를 분석의 도구 또는 기법에 맞게 다듬는 과정이 필요함
  - **정제과정을 거치지 않은 데이터의 문제점**
    - 데이터 구성의 일관성이 없어지므로 분석의 처리에 어려움이 발생함
    - 도출된 결과의 신뢰성 저하가 발생함

데이터 정제

- **데이터의 정제**
  - **데이터 정제의 처리 과정(Processing)**
    - 다양한 매체로부터 데이터를 수집, 원하는 형태로 변환, 원하는 장소에 저장, 저장된 데이터의 활용 가능성을 타진하기 위한 품질 확인, 필요한 시기와 목적에 따라 사용이 원활하도록 관리하는 과정이 필요함
    - 시스템 내 · 외부에서 데이터를 수집하면 정형보다 비정형 데이터들이 많음. 비정형 데이터의 경우, 기본적으로 구조화된 정형 데이터로의 변환을 수행하고, 변환된 데이터에서 결측치나 오류의 수정과정을 거침
    - 기존 시스템 내의 데이터와 비교 분석이 필요한 경우, 레거시 데이터와 통합·변환의 과정이 발생할 수 있으므로 분석에 필요한 데이터를 추출하고 통합 · 변환하는 과정이 발생할 수 있음

데이터 정제

- **데이터의 정제**
  - **데이터 정제의 처리 과정(Processing)**

구분	수행 내용	Process
데이터의 수집	• 데이터의 입수 방법 및 정책 결정 • 입수경로의 구조화 • 집계 (Aggregation) • 저장소 결정	전처리 (Pre-Processing) 포함
데이터의 변환	• 데이터 유형의 변화 및 분석 가능한 형태로 가공 • ETL (Extract, Transform, Load) • 일반화 (Generalization) • 정규화 (Normalization)	
데이터의 교정	• 결측치 처리, 이상치 처리, 노이즈 처리 • 비정형 데이터 수집 시 필수 사항	
데이터의 통합	• 데이터 분석이 용이하도록 기존 또는 유새 데이터와의 연계·통합 • 레거시 데이터와 함께 분석이 필요할 경우 수행	

데이터 정제

- **데이터의 정제**
  - **데이터 정제의 전처리와 후처리**
    - 전처리(Pre-Processing)
      - 데이터 저장 전의 처리과정
      - 대상 데이터와 입수방법 결정 및 저장방식, 장소를 선정
    - 후처리(Post-Processing)
      - 데이터 저장 후의 처리과정
      - 저장 데이터의 품질관리 등의 과정을 포함함

데이터 정제

- 데이터 결측값 처리
  - 데이터 분석에서 결측치(Missing Data)는 데이터가 없음을 의미함
    - 결측치를 임의로 제거 시: 분석 데이터의 직접 손실로 인해 분석에 필요한 유의수준 데이터 수집에 실패할 가능성이 발생함
    - 결측치를 임의로 대체 시: 데이터의 편향(Bias)이 발생하여 분석 결과의 신뢰성 저하 가능성이 있음
  - 결측치에 대한 처리는 임의 제거·대체의 방법을 사용할 경우, 상기 문제를 피할 수 있는 데이터에 기반한 방법으로 처리해야 함

데이터 정제

- 데이터 결측값 처리
  - 결측 데이터의 종류
    - 완전 무작위 결측(Missing Complete At Random, MCAR)
      - 어떤 변수 상에서 결측 데이터가 관측된 또는 관측되지 않은 다른 변수와 아무런 연관이 없는 경우
    - 무작위 결측(Missing At Random, MAR)
      - 변수 상의 결측 데이터가 관측된 다른 변수와 연관되어 있지만 그 자체가 비관측값들과는 연관되지 않은 경우
    - 비 무작위 결측(Not Missing At Random, NMAR)
      - 어떤 변수의 결측 데이터가 MCAR 또는 MAR이 아닌 결측 데이터로 정의되는 경우
      - 즉, 결측 변수값이 결측 여부(이유)와 관련이 있는 경우

데이터 정제

- **데이터 결측값 처리**
  - **결측 데이터의 종류에 따른 모델링의 예**

가정

나이대별(X), 성별(Y)체중(Z) 분석에 대한 모델링의 경우

가정의 내용

- X, Y, Z와 관계없이 Z가 없는 경우: 데이터의 누락(응답없음) → 완전 무작위 결측(MCAR)
- 여성(Y)은 체중 공개를 꺼려하는 경향: Z가 누락될 가능성에 Y에만 의존 → 무작위 결측(MAR)
- 젊은(X) 여성(Y)의 경우는 체중 공개를 꺼리는 경우가 더 높음 → 무작위 결측(MAR)
- 무거운(가벼운) 사람들은 체중 공개의 가능성이 적음: Z가 누락될 가능성이 Z값 자체에 관찰되지 않는 값에 달려 있음 → 비 무작위 결측(NMAR)

데이터 정제

- 데이터 결측값 처리
  - 결측 값 유형의 분석 및 대치
    - 결측치의 처리를 위해 실제 데이터셋에서 결측치가 어떤 유형으로 분류되는지 분석하고 결과에 따라서 결측치 처리 방법의 선택이 필요함
    - 일반적으로 결측·무응답을 가진 자료를 분석할 대는 완전 무작위 결측(MCAR) 조건 하에서 처리함. 즉 불완전한 자료는 무시하고 완전히 관측된 자료만을 표준적 분석 대상으로 지정하여 분석을 시행함
    - 결측치가 존재하는 데이터를 이용한 분석은 효율성(Efficiency), 자료처리의 복잡성, 편향문제의 세 가지 고려사항이 발생함
    - 결측치의 대치를 위한 방법으로는 단순 대치법과 다중 대치법이 있음

데이터 정제

- **데이터 결측값 처리**
  - **결측 값 유형의 분석 및 대치**
    - **단순 대치법(Simple Imputation)**
      - **기본적으로 결측치에 대하여 MCAR 또는 MAR로 판단하고 이에 대한 처리를 수행하는 방법**
      - **완전 분석(Completes Analysis)**
        - 불완전한 자료는 완전하게 무시하고 분석을 수행함
        - 분석의 용이성을 보장하지만 효율성 상실과 통계적 추론의 타당성에 문제가 발생할 가능성이 있음
      - **평균 대치법(Mean Imputation)**
        - 관측 도는 실험으로 얻어진 데이터의 평균으로 결측치를 대치해서 사용함
        - 평균에 의한 대치는 효율성의 향상 측면에는 장점이 있으나 통계량의 표준 오차가 과소 추정되는 단점이있음
        - 비조건부 평균 대치법이라고도 부름

데이터 정제

- **데이터 결측값 처리**
  - **결측 값 유형의 분석 및 대치**
    - **단순 대치법(Simple Imputation)**
      - **회귀 대치법(Regression Imputation)**
        - 회귀분석에 의해 결측치를 대치하는 방법
        - 조건부 평균 대치법이라고도 부름
      - **단순확률 대치법(Single Stochastic Imputation)**
        - 평균 대치법에서 추정량 표준 오차의 과소 추정을 보완하는 대치법
        - Hot-Deck 방법이라고도 부름
        - 확률 추출에 의해서 전체 데이터 중 무작위로 대치하는 방법

데이터 정제

- **데이터 결측값 처리**
  - **결측 값 유형의 분석 및 대치**
    - **단순 대치법(Simple Imputation)**
      - **최근접 대치법(Nearest-Neighbor imputation)**
        - 전체 표본을 몇 개의 대체군으로 분류하여 각 층에서의 응답자료를 순서대로 정리한 후, 결측값 바로 이전의 응답을 결측치로 대치하는 방법
        - 응답값이 여러 번 사용될 가능성이 있다는 단점을 가짐

데이터 정제

- **데이터 결측값 처리**
  - **결측 값 유형의 분석 및 대치**
    - 다중 대치법(Multiple Imputation)
      - 단순 대치법을 복수로 시행함으로써 통계적 효율성 및 일치성 문제를 보완하기 위해 만들어진 방법
      - 복수(n) 개의 단순 대치를 통해 n개의 새로운 자료를 만들어 분석을 시행하고, 시행 결과로 얻어진 통계량에 대해 통계량 및 분산 결합을 적용하여 통합하는 방법
      - 1단계 – 대치 단계(Imputation Step): 결측값을 대치한 데이터 복수 개를 생성
      - 2단계 – 분석 단계(Analysis Step): 복수 개의 데이터셋에 대한 분석 시행
      - 3단계 – 결합 단계(Combination Step): 복수 개의 분석 결과에 대한 통계적 결합을 통해 결과 도출

데이터 정제

- 데이터 이상치 처리
  - 이상치(Outlier)란 데이터 전처리 과정에 발생 가능한 문제로 정상의 범주(데이터의 전체적 패턴)에서 벗어난 값을 의미함
  - 데이터의 수집 과정에서 오류가 발생할 수도 있기 때문에 이상치가 포함될 수 있음
  - 오류가 아니더라도 굉장히 극단적인 값의 발생으로 인한 이상치가 존재할 수 있음

  - 이상치의 종류
    - 단변수 이상치(Univariate Outlier): 하나의 데이터 분포에서 발생하는 이상치
    - 다변수 이상치(Multivariate Outlier): 복수의 연결된 데이터 분포 공간에서 발생하는 이상치

데이터 정제

- **데이터 이상치 처리**
  - **이상치의 발생 원인**
    - 비자연적 이상치 발생(Artificial/Non-Natural Outlier)
      - 입력 실수(Data Entry Error): 데이터의 수집 과정에서 발생하는 오류. 입력의 실수 등을 지칭함
      - 측정 오류(Measurement Error): 데이터의 측정 중에 발생하는 오류. 측정기 고장(이상 작동)으로 발생하는 문제
      - 실험 오류(Experimental Error): 실험과정 중 발생하는 오류. 실험환경에서 야기된 모든 문제점을 지칭함
      - 의도적 이상치(Intentional Outlier): 자기 보고 측정(Self-Reported Measure)에서 발생되는 이상치를 지칭함. 의도가 포함된 이상치로 예를 들어 남성의 키를 조사 시 의도적으로 크게 기입하는 경우 등이 있음
      - 자료처리 오류(Data Processing Error): 복수 개의 데이터셋에서 데이터를 추출·조합하여 분석 시, 분석 전의 전처리에서 발생하는 오류를 지칭함
      - 표본 오류(Sampling Error): 모집단에서 표본을 추출하는 과정에서 편향이 발생하는 경우를 지칭함

데이터 정제

- **데이터 이상치 처리**
  - **이상치의 발생 원인**
    - 자연적 이상치 발생(Natural Outlier)
      - 비자연적 이상치 이외의 이유로 발생하는 모든 이상치

데이터 정제

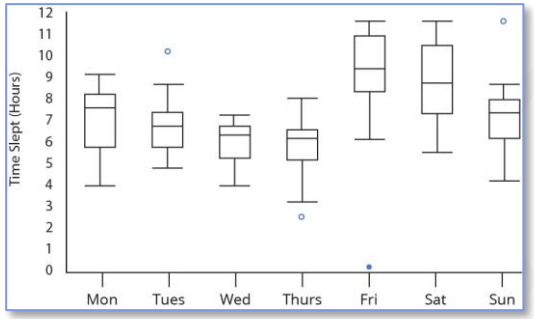
- **데이터 이상치 처리**
  - **이상치의 문제점**
    - 기초(통계적) 분석결과의 신뢰도 저하
      - 평균, 분산 등에 영향을 줌. 단 중앙값은 영향이 적음
    - 기초통계에 기반한 다른 고급 통계분석의 신뢰성 저하
      - 검정·추정 등의 분석, 회귀분석 등이 영향을 받음
    - 특히 이상치가 비무작위성(Non-Randomly)을 가지고 나타나게(분포하게) 되면 데이터의 정상성(Normality) 감소를 초래하며, 이는 데이터 자체의 신뢰성 저하로 연결될 수 있음

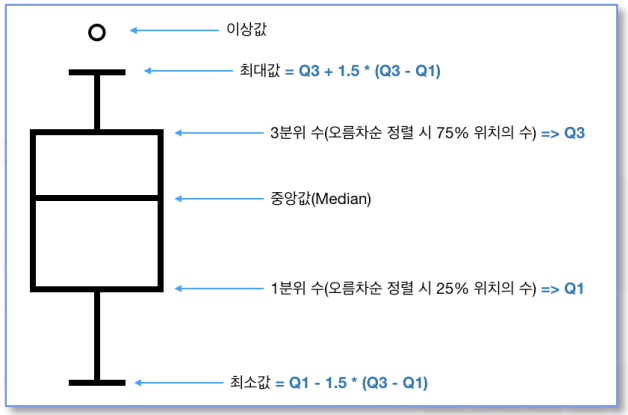
데이터 정제

- **데이터 이상치 처리**
  - **이상치의 탐지**
    - **종속변수가 단변량(Univariate)인지 다변량(Multivariate)인지 데이터의 분포를 확인하고, 모수적(Parametric) 또는 비모수적(non-Parametric)인지에 따라 다양한 방법으로 고려해야 함**
    - **시각화(Visualization)를 통한 방법(비모수적, 단변량(2변량)의 경우)**
      - **상자그림(Box Plot), 줄기-잎 그림(Stem and Leaf Diagram)**
        - 상자그림(Box Plot)은 상자수염그림(Box and Whisker Plot)이라고 부르기도 함
      - **산점도 그림(Scatter Plot): 비모수적 2변량인 경우 사용함**

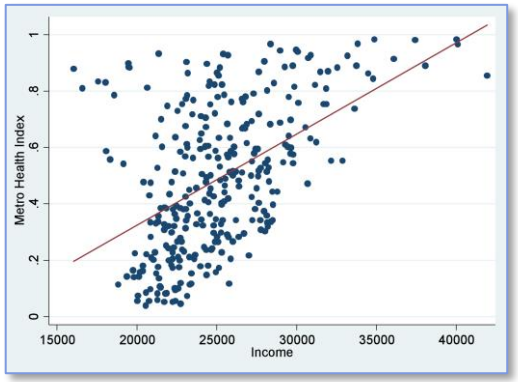
데이터 정제

- 데이터 이상치 처리
  - 이상치의 탐지
    - 시각화(Visualization)를 통한 방법(비모수적, 단변량(2변량)의 경우)
      - 상자그림(Box Plot), 줄기-잎 그림(Stem and Leaf Diagram)





      - 산점도 그림(Scatter Plot)



데이터 정제

- 데이터 이상치 처리
  - 이상치의 탐지
    - Z-Score를 통한 방법(모수적 단변량 또는 저변량의 경우)
      - 정규화를 통해 특정 Threshold(임계값)를 벗어난 경우를 이상치로 판별함
    - 밀도 기반 클러스터링 방법(Density Based Spatial Clustering of Application with Nise, DBSCAN)
      - 비모수적 다변량의 경우 군집 간의 밀도를 이용하여 특정 거리 내의 데이터 수가 지정 개수 이상이면 군집으로 정의하는 방법
      - 정의된 군집에서 먼 거리에 있는 데이터는 이상치로 간주함
    - 고립 의사나무 방법(Isolation Forest)
      - 비모수적 다변량의 경우 의사결정나무(Decision Tree) 기반으로 정상치의 단말 노드(Terminal Node)보다 이상치의 노드에 이르는 길이가 더 짧은 성질을 이용하는 방법

분석 변수의 처리

분석 변수 처리

- 변수의 선택
  - 통계적 분석 결과의 신뢰성을 위해서 기본적으로 데이터와 이를 특정 짓는 변수는 많을수록 좋음
  - 그러나 분석 모형의 구현과 사용에 지속적으로 필요 이상의 많은 데이터를 요구할 수 있다는 문제가 있음

- 회귀분석의 사례
  - 회귀모형에 의한 분석의 경우,
  - 최종 결과를 도출해 내기 위해서 사용된 독립 면수가 m개이고
  - 이를 통해서 얻어진 설명력이 R2=89%라고 했을 때,
  - m보다 작은 n개만을 사용할 때 동일한 설명력이 나온다면
  - 변수의 효율적 선택의 필요성이 증가함

분석 변수 처리

- 변수의 선택
  - 변수별 모형의 분류
    - 전체 모형(Full Model, FM): 모든 독립변수를 사용한 모형으로 정의
    - 축소 모형(Reduced Model, RM): 전체 모형에서 사용된 변수의 개수를 줄여서 얻은 모형
    - 영 모형(Null Model, NM): 독립변수가 하나도 없는 모형

분석 변수 처리

- **변수의 선택**
  - **변수의 선택 방법**
    - 전진 선택법(Forward Selection)
      - 영 모형에서 시작
      - 모든 독립변수 중 종속변수와 단순 상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 방법
      - 부분 F 검정(F Test)을 통해 유의성 검증을 시행, 유의한 경우는 가장 큰 F 통계량을 가지는 모형을 선택하고, 유의하지 않은 경우는 변수 선택 없이 과정을 중단함
      - 한 번 추가된 변수는 제거하지 않는 것이 원칙임

분석 변수 처리

- 변수의 선택
  - 변수의 선택 방법
    - 후진 소거법(Backward Elimination), 후진 선택법(Backward Selection)
      - 전체 모델에서 시작
      - 모든 독립변수 중 종속변수와 단순 상관계수의 절댓값이 가장 작은 변수를 분석 모형에서 제외시키는 방법
      - 부분 F 검정(F Test)을 통해 유의성 검증을 시행, 유의하지 않은 경우는 변수를 제거하고, 유의한 경우는 변수제거 없이 과정을 중단함
      - 한 번 제거된 변수는 추가하지 않는 것이 원칙임

분석 변수 처리

- **변수의 선택**
  - **변수의 선택 방법**
    - 단계적 선택법(Stepwise Selection)
      - 전진 선택법과 후진 소거법의 보완방법
      - 전진 선택법을 통해 가장 유의한 변수를 모형에 포함한 후, 나머지 변수들에 대해 후진 소거법을 적용하여 새롭게 유의하지 않은 변수들을 제거함
      - 제거된 변수는 다시 모형에 포함하지 않으며, 유의한 설명변수가 존재하지 않을 때까지 과정을 반복함

분석 변수 처리

- 차원 축소
  - 차원의 축소는 어떤 목적에 따라서 변수(데이터의 종류)의 양을 줄이는 것을 말함
  - 차원 축소의 필요성
    - 복잡도의 축소(Reduce Complexity)
    - 과적합(Overfit)의 방지
    - 해석력(Interpretability)의 확보
    - 차원의 저주(Curse of Dimensionality)

분석 변수 처리

- 차원 축소
  - 차원 축소의 필요성
    - 복잡도의 축소(Reduce Complexity)
      - 데이터를 분석하는데 있어서 분석 시간의 증가(시간복잡도: Time Complexity)와 저장변수 양의 증가(공간복잡도: Space Complexity)를 고려할 때, 동일한 품질을 나타낼 수 있다면 효율성 측면에서 데이터의 수를 줄여야 함
    - 과적합(Overfit)의 방지
      - 차원의 증가는 분석모델 파라미터의 증가 및 파라미터 간의 복잡한 관계의 증가로 분석 결과의 과적합 발생 가능성이 커짐. 이것은 분석 모형의 정확도(신뢰도) 저하를 발생시킬 수 있음
      - 작은 차원만으로 안정적인(Robust) 결과를 도출해 낼 수 있다면 많은 차원을 다루는 것보다 효율적임

분석 변수 처리

- 차원 축소
  - 차원 축소의 필요성
    - 해석력(Interpretability)의 확보
      - 차원이 작은 간단한 분석 모델일수록 내부구조의 이해가 용이하고 해석이 쉬워짐
      - 해석이 쉬워지면 명확한 결과도출에 많은 도움을 줄 수 있음
    - 차원의 저주(Curse of Dimensionality)
      - 데이터 분석 및 알고리즘을 통한 학습을 위해 차원이 증가하면서 학습 데이터의 수보다 차원의 수가 많아져 성능이 저하되는 현상
      - 해결을 위해서는 차원을 줄이거나 데이터의 수를 늘리는 방법을 이용해야 함

- **차원 축소**
  - **차원 축소의 방법**
    - **요인 분석(Factor Analysis)**
      - **요인 분석의 목적**
        - 변수 축소: 다수의 변수들의 정보 손실을 억제하면서 소수의 요인(Factor)으로 축약하는 것
        - 변수 제거: 요인에 대한 중요도 파악에 따라 중요하지 않은 변수를 제거함
        - 변수 특성 파악: 관련된 변수들이 묶임(군집화)으로써 요인 간의 상호 독립성 파악이 요구됨
        - 타당성 평가: 군집화하지 않은 변수의 독립성 여부를 판단함
        - 파생변수: 요인 점수를 이용한 새로운 변수를 생성함. 회귀분석, 판별분석 및 군집분석 등에 이용할 수 있음

분석 변수 처리

- **차원 축소**
  - **차원 축소의 방법**
    - 요인 분석(Factor Analysis)
      - **요인 분석의 특징**
        - 독립변수, 종속변수의 개념이 없음
        - 주로 기술 통계에 의한 방법을 이용함
      - **요인 분석의 종류**
        - 주성분 분석, 공통요인 분석, 특이값 분해(SVD), 행렬과 음수 미포함 행렬 분해(NMF) 등이 있음
        - 공통요인 분석: 분석대상 변수들의 기저를 이루는 구조를 정의하기 위한 요인 분석 방법으로 변수들이 가지고 있는 공통 분산만을 이용하여 공통요인만 추출하는 방법

- **차원 축소**
  - **차원 축소의 방법**
    - **주성분 분석(Principal Component Analysis, PCA)**
      - **주성분 분석 개요**
        - 서로 연관성이 있는 고차원 공간의 데이터를 선형연관성이 없는 저차원(주성분)으로 변환하는 과정을 거침(직교 변환을 사용함)
        - 기존의 기본변수들을 새로운 변수의 세트로 변환하여 차원을 줄이되 기존 변수들의 분포특성을 최대한 보존하여 이를 통한 분석결과의 신뢰성을 확보함
        - 2차원 좌표평면에 n개의 점 데이터들이 타원형으로 분포되어 있을 때, 이 데이터들의 분포특성을 2개의 벡터로 가장 잘 설명할 수 있는 방법은 두 개의 벡터를 좌표로 하여 그래프를 그려 데이터 분포를 설명하는 것임
        - PCA는 데이터 하나하나에 대한 성분을 분석하는 것이 아니라, 여러 데이터들이 모여 하나의 분포를 이룰 때, 이 분포의 주성분을 분석해 주는 방법임

분석 변수 처리

- **차원 축소**
  - **차원 축소의 방법**
    - **주성분 분석(Principal Component Analysis, PCA)**
      - **PCA의 특징**
        - 차원 축소에 폭넓게 사용됨
        - 어떤 사전적 분포 가정의 요구가 없음
        - 가장 큰 분산의 방향들을 주요 중심 관심으로 가정함
        - 본래의 변수들의 선형결합으로만 고려함
        - 차원의 축소는 본래의 변수들이 서로 상관 관계에 있을때만 가능함
        - 스케일에 대한 영향이 큼. 즉 PCA 수행을 위해서는 변수들 간의 스케일링이 필수임

- **차원 축소**
  - **차원 축소의 방법**
    - **특이값 분해(Singular Value Decomposition, SVD)**
      - **특이값 분해 개요**
        - 데이터 공간을 나타내는 m×n 크기의 행렬 M에 대해 다음과 같이 분해 가능함 ($M = U\Sigma V^t$)
        - 여기서 U 는 m×m 크기의 직교행렬이고 Σ 는 m×n 크기의 대각행렬, V^t 는 n×n 크기의 직교행렬임
      - **특이값 분해의 차원 축소 원리**
        - 수학적 원리: SVD 방법은 주어진 행렬 M(크기가 m×n인 행렬)을 여러 개의 (M과 동일한 크기를 갖는) 행렬로 분해할 수 있으며, 각 행렬의 원소 값의 크기는 대각행렬에서 대각 성분의 크기에 의해 결정됨
        - 데이터의 응용: 기존의 전차원의 정보를 포함하는 행렬 A를 SVD에 의해서 3개의 행렬로 분해하며, 적당한 k(특이값)만을 이용해 원래 행렬 A와 비슷한 정보력을 가지는 차원을 만들어 낼 수 있음
        - 즉, 큰 몇 개의 특이값을 가지고도 충분히 유용한 정보를 유지할 수 있는 차원을 생성해 낼 수 있음(차원 축소)

분석 변수 처리

- **차원 축소**
  - **차원 축소의 방법**
    - **음수 미포함 행렬 분해(Non-Negative Matrix Factorization, NMF)**
      - **음수를 포함하지 않은 행렬 V를 음수를 포함하지 않은 두 행렬의 곱으로 분해하는 알고리즘**
      - **일반적으로 행렬 분해는 정확한 해가 없기 때문에 대략적인 해를 구하게 됨**
      - **NMF의 이해**
        - 일반적으로 W의 열 개수, H의 행 개수가 WH=V가 되도록 결정함
        - 기존 행렬 V와 분해한 음수 미포함 행렬 W와 H의 곱과의 차이를 오차 U로 지정함
        - V=WH+U 이며 U의 원소들은 양수나 음수가 될 수 있음
        - W와 H의 크기가 V보다 작기 때문에 저장하거나 다루기에 용이함
        - 또한 V를 원래 정보보다 상대적으로 적은 정보로 표현하여 분해한 행렬 하나가 전체 정보의 대략적인 정보를 제시할 수 있음

분석 변수 처리

- **차원 축소**
  - **차원 축소의 방법**
    - **음수 미포함 행렬 분해(Non-Negative Matrix Factorization, NMF)**
      - **NMF의 차원 축소**
        - 행렬 곱셈에서 곱해지는 행렬은 결과행렬보다 훨씬 적은 차원을 가지기 때문에 NMF가 차원을 축소할 수 있음

분석 변수 처리

- 파생변수의 생성
  - 데이터 분석 시 주어진 원 데이터를 그대로 활용하기 보다는 분석의 목표에 적합하도록 계속해서 데이터 형태를 수정 및 보완할 필요가 있음
  - 요약변수와 파생변수는 분석 모델을 구축하는데 있어서 핵심인 환경과 문제를 잘 해석할 수 있는 변수를 찾는데 의의가 있음

분석 변수 처리

- **파생변수의 생성**
  - **파생변수**
    - 사용자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여하는 변수
    - 매우 주관적일 수 있으므로 논리적 타당성을 갖출 필요가 있음
    - 세분화 고객행동예측, 캠페인 반응예측 등에 활용할 수 있음
    - 특정 상황에만 유의미하지 않게 대표성을 나타나게 할 필요가 있음
  - **요약변수**
    - 수집된 정보를 분석에 맞게 종합한 변수
    - 데이터 마트에서 가장 기본적인 변수
    - 많은 분석 모델에서 공통으로 사용될 수 있어 재활용성이 높음

분석 변수 처리

- **파생변수의 생성**
  - **요약변수 VS 파생변수**
    - 요약변수 처리 시의 유의점
      - 처리(단어의 빈도, 초기 행동변수, 트렌드 변수 등) 방법에 따라 결측치의 처리 및 이상값 처리에 유의해야 함
      - 연속형 변수의 구간화 적용과 고정된 구간화를 통한 의미 파악 시 의미 있는 구간을 찾도록 해야 함
    - 파생변수 생성 및 처리의 유의점
      - 특정 상황에만 의미성을 부여하는 것이 아닌 보편적이고 전 데이터 구간에 대표성을 가지는 파생변수의 생성을 위해서 노력해야 함

분석 변수 처리

- **파생변수의 생성**
  - **파생변수의 생성 방법**
    - 한 값으로부터 특징을 추출함
    - 한 레코드 내의 값들을 결합함
    - 다른 테이블의 부가적 정보를 결합함
    - 다수의 필드 내의 시간 종속적인 데이터를 선택(Pivoting)함
    - 레코드 또는 중요 필드를 요약함
    - 상기 이외에도 다양한 형태의 파생 변수를 목적에 맞게 생성하는 것이 가능함

분석 변수 처리

- **변수 변환**
  - **변수 변환의 개념**
    - 데이터를 분석하기 좋은 형태로 바꾸는 작업
    - 수학적 의미에서의 변환(Transformation)은 기존의 변수 공간에서는 해결하거나 관찰할 수 없는 사실을 영역을 달리 하는 것으로(변환) 해석이 용이해 지거나 취급이 단순해지는 장점이 있음
    - 데이터의 전처리 과정 중 하나로 간주됨

분석 변수 처리

- 변수 변환
  - 변수 변환의 방법
    - 범주형 변환
      - 연속형 변수 중에서 변수 자체로의 분석보다는 분석 결과의 명료성 및 정확성을 배가시키기 위해 범주형으로 변환시키는 방법
    - 정규화
      - 데이터가 가진 스케일이 심하게 차이나는 경우 그 차이를 그대로 반영하기보다는 상대적 특성이 반영된 데이터로 변환하는 것이 필요함

분석 변수 처리

- **변수 변환**
  - **변수 변환의 방법**
    - **정규화**
      - **일반 정규화**
        - 수치로 된 값들을 여러 개 사용할 때 각 수치의 범위가 다르면 이를 같은 범위로 변환해서 사용하는 방법
      - **최소-최대 정규화(Min-Max Normalization)**
        - 데이터를 정규화하는 가장 일반적인 방법
        - 모든 Feature에 대해 최소값 0, 최대값 1로, 그리고 다른 값은 0과 1사이의 값으로 변환하는 것
        - 이상치의 영향을 많이 받는다는 단점이 있음

분석 변수 처리

- **변수 변환**
  - **변수 변환의 방법**
    - **정규화**
      - **Z-점수(Z-Score) 정규화**
        - 이상치 문제를 피하는 데이터 정규화 전략 ($Z = \frac{X-\mu}{\sigma}$)
        - 만약 데이터의 값이 평균과 일치하면 0, 평균보다 작으면 음수, 평균보다 크면 양수로 계산됨
        - 이때 계산되는 음수와 양수의 크기는 그 데이터의 표준편차에 의해 결정됨
        - 만약 데이터의 표준편차가 크면(값이 넓게 퍼져 있으면) 정규화 되는 값이 0에 가까워짐

분석 변수 처리

- 변수 변환
  - 변수 변환의 방법
    - 로그 변환(Log Transformation)
      - 데이터 분석에서 로그를 취하면 그 분포가 정규 분포에 가깝게 분포하게 되는 경우기 있음. 이런 분포를 로그정규분포(Log-Normal Distribution)를 가진다고 말함
      - 로그 변환 분포를 사용하는 전형적인 데이터
      - 국가별 수출액, 사람의 통증 정도 수치화, 개별 주식의 가격 이용 변동성 분석 등
      - 데이터 분포의 형태가 우측으로 치우친 경우, 정규분포화를 위해 로그변환을 사용함 ($X \sim \ln(X)$)

분석 변수 처리

- **변수 변환**
  - **변수 변환의 방법**
    - 역수 변환(Inverse Transformation)
      - 어떤 변수를 데이터 분석에 그대로 사용하지 않고 역수를 사용하면 오히려 선형적인 특성을 가지게 되어 의미를 해석하기 쉬워지는 경우를 말함
      - 데이터의 분포 형태가 극단적인 우측으로 치우친 경우, 정규분포화를 위해 역수 변환을 사용함 ($X \sim \frac{1}{X}$)
    - 지수 변환(Power Transformation)
      - 어떤 변수를 데이터 분석에 그대로 사용하지 않고 지수를 사용하면 오히려 선형적인 특성을 가지게 되어 의미를 해석하기 쉬워지는 경우를 말함
      - 데이터의 분포 형태가 극단적인 좌측으로 치우친 경우, 정규분포화를 위해 지수 변환을 사용함 ($X \sim X^n$)

분석 변수 처리

- **변수 변환**
  - **변수 변환의 방법**
    - 제곱근 변환(Square Root Transformation)
      - 어떤 변수를 데이터 분석에 그대로 사용하지 않고 제곱근을 사용하면 오히려 선형적인 특성을 가지게 되어 의미를 해석하기 쉬워지는 경우를 말함
      - 데이터의 분포 형태가 우측으로 약간 치우친 경우, 정규분포화를 위해 제곱근 변환을 사용함 ($X\sqrt{X}$)

분석 변수 처리

- 변수 변환
  - 변수 변환의 방법
    - 분포형태별 정규분포 변환
      - 모집단의 분포형태별로 사용 가능한 변수변환이 달라짐
      - 최종적으로는 정규분포 형태를 지향함
      - 기본적으로 단일 집단의 정규성 검정은 데이터 분포의 형태를 눈으로 확인할 수도 있지만 샤피로 테스트(Shapiro Test) 또는 큐큐 플롯(Q-Q Plot)을 이용해 확인 가능하며, 결과에 따라 적당한 변수 변환식을 사용하여 정규분포 형태로 변환이 가능할 수 있음
    - 기타 방법
      - 데이터 축소 등

변수 변환 전 분포	사용하는 변수 변환식	변수 변환 후 분포
좌로 치우침	X^3	정규분포화
좌로 약간 치우침	X^2	
우로 약간 치우침	$\sqrt{X}$	
우로 치우침	$\ln(X)$	
극단적 우로 치우침	$\frac{1}{X}$	

분석 변수 처리

- **불균형 데이터 처리**
  - **어떤 데이터에서 각 클래스(주로 범주형 반응 변수)가 갖고 있는 데이터의 양에 차이가 큰 경우, 클래스 불균형이 있다고 말함**
  - **불균형 데이터의 문제점**
    - 데이터 클래스 비율이 너무 차이가 나면(Highly-Imbalanced Data) 단순히 우세한 클래스를 택하는 모형의 정확도가 높아지므로 모형의 성능 판별이 어려워짐
    - 즉 정확도(Accuracy)가 높아도 데이터 개수가 적은 클래스의 재현율(Recall-Rate)이 급격히 작아지는 현상이 발생할 수 있음

분석 변수 처리

- **불균형 데이터 처리**
  - **불균형 데이터의 문제점**
    - **혼동행렬 (Confusion Matrix)**

		사실	
		참 (Positive)	거짓 (Negative)
실험 결과	참 (Positive)	TP (True Positive)	FP (False Positive)
	거짓 (Negative)	FN (False Negative)	TN (True Negative)

- 정확도 (Accuracy) = $\frac{TP+TN}{TP+TN+FP+FN}$
- 재현율 (Recall) = $\frac{TP}{TP+FN}$

분석 변수 처리

- **불균형 데이터 처리**
  - **불균형 데이터의 처리 방법**
    - **가중치 균형 방법(Weighted Balancing)**
      - **데이터에서 손실(Loss)을 계산할 때 특정 클래스의 데이터에 더 큰 손실값을 갖도록 하는 방법**
      - **데이터 클래스의 균형이 필요한 경우이며, 각 클래스별 특정 비율로 가중치를 주어서 분석하거나 결과를 도출하는 것**
      - **고정 비율 이용**
        - 클래스의 비율에 따라 가중치를 두는 방법
        - 예: 클래스의 비율이 1:5라면 가중치를 5:1로 줌으로서 적은 샘플 수를 가진 클래스를 전체 손실에 동일하게 기여하도록 할 수 있음
      - **최적 비율 이용**
        - 분야와 최종 성능을 고려하여 가중치 비율의 최적 값을 찾으면서 가중치를 찾아가는 방법

분석 변수 처리

- 불균형 데이터 처리
  - 언더샘플링(Undersampling)과 오버샘플링(Oversampling)
    - 비대칭 데이터는 언더샘플링이나 오버샘플링을 사용하여 데이터 비율을 맞추면 정밀도(Precision)가 향상됨
    - 언더샘플링
      - 대표 클래스(Majority Class)의 일부만을 선택하고 소수 클래스(Minority Class)는 최대한 많은 데이터를 사용하는 방법
      - 이때 언더샘플링된 대표 클래스 데이터가 원본 데이터와 비교해 대표성이 있어야 함
    - 오버샘플링
      - 소수 클래스의 복사본을 만들어 대표 클래스의 수만큼 데이터를 만들어 주는 방법
      - 똑같은 데이터를 그대로 복사하는 것이기 때문에 새로운 데이터는 기존 데이터와 같은 성질을 가지게 됨