과목Ⅰ. 데이터 이해

제1장 데이터의 이해

제1절 데이터와 정보

1. 데이터의 정의

- 데이터의 의미는 과거 관념적이고 추상적인 개념에서 기술적이고 사실적인 의미로 변화
- 데이터란 추론과 추정의 근거를 이루는 사실(옥스퍼드 대사전) : 데이터를 단순한 객체로서 가치뿐만 아니라 다른 객체와의 상호관계 속에서 가치를 갖는 것으로 설명
- 데이터는 객관적 사실(fact, raw material)이라는 존재적 특성을 갖는 동시에 추론·예측·전망·추정을 위한 근거(basis)로 기능하는 당위적 특성을 가짐
- 논문, 경영전략, 정책수립 등 일련의 가치 창출과정에서 가장 기초를 이루는 것
- 데이터의 유형
  - 정성적 데이터(qualitative data) : 언어·문자 등 비정형 데이터, 상대적으로 많은 비용과 기술적 투자가 수반
  - 정량적 데이터(quantitative date) : 수치·도형·기호 등, 저장·검색·분석 활용에 용이
- 데이터는 지식경영의 핵심이슈인 암묵지와 형식지의 상호작용에 중요한 역할
  - 암묵지 : 학습과 체험을 통해 개인에게 습득된 무형의 지식(공유어려움), 과학적 발견
  - 형식지 : 형상화된 지식, 유형의 대상이 있어 지식의 전달과 공유가 매우 용이
- 암묵지와 형식지의 상호작용(→지식형성의 중요한 기초)

암묵지(tacit knowledge)	형식지(explicit knowledge)
공통화(Socialization)	표출화(Externalization)
내면화(Internalization)	연결화(Combination)

2. 데이터와 정보의 관계

- 데이터 : 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실
- 정보 : 데이터의 가공·처리와 데이터 간 연관관계 속에서 의미가 도출된 것
  정보가 내포하는 의미는 유용하지 않을 수 있음
- 지식 : 데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것
- 지혜 : 지식의 축적과 아이디어가 결합된 창의적 산물
- DIKW 피라미드 : 데이터, 정보, 지식을 통해 최종적으로 지혜를 얻는 과정을 계층구조로 설명

지혜(**W**isdom)	근본원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어
지식(**K**nowledge)	상호 연결된 정보패턴을 이해하여 이를 토대로 예측한 결과물
정보(**I**nformation)	데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 데이터
데이터(**D**ata)	존재형식을 불문하고 타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호를 의미

  - 데이터, 정보, 지식은 상호 관계 속에서 역할을 수행하며 인간이 사회활동에서 추구하는 가치창출을 위한 일련의 프로세스로 기능
- 데이터의 정확성은 향수 데이터 간의 관계 및 현상의 분석(정보)과 적용(지식), 더 나아가 미래를 예측하고 창의적 산물을 도출(지혜)하는데 지대한 영향을 미치며 가치창출에 핵심적인 역할 수행

제2절 데이터베이스 정의와 특징

1. 용어의 연역

- 1950s : 수집된 자료를 일컫는 데이터(data)의 기지(base)라는 뜻으로 데이터베이스(data base)가 탄생
- 1963년 : 미국 SDC가 개최한 심포지엄에서 공식적으로 사용됨(초기개념인 대량의 데이터를 축적하는 기지라는 의미에 머무름)
  GE의 바크만은 최초의 현대적 의미의 데이터베이스관리시스템인 IDS를 개발. 이를 통해 새롭게 저장장치에 직접 접근하기 위한 데이터 모델이 제안되었고 이후 다양한 데이터 모델을 기반으로 한 데이터베이스 관리시스템이 개발됨
- 65년 : 2차 심포지엄에서 시스템을 통한 체계적 관리와 저장 등의 의미를 담은 '데이터베이스 시스템'이라는 용어가 등장
- 70s초반 유럽 데이터베이스라는 단일어 일반화, 70s후반 미국 주요신문 등에 흔히 사용

2. 데이터베이스의 정의

- 단순한 데이터의 수집·저장을 위해 탄생한 데이터베이스는 이후 다양한 정보기술의 발달과 인터넷의 확산 등으로 인한 디지털 시대에 진입하면서 보다 현대적 개념이 적용된 용어로 정의됨
- 데이터베이스란?
  - 체계적이거나 조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물. EU『데이터베이스의 법적 보호에 관한 지침』
  - 소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것『국내 저작권법』·법률적으로 기술기반 저작물로 인정
  - 동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합『컴퓨터용어사전, 정보통신용어사전(TTA)』
  - 관련된 레코드의 집합, 소프트웨어로는 데이터베이스관리시스템(DBMS)을 의미『Wikipedia』
- 데이터베이스의 개념은 체계적으로 정렬된 데이터 집합을 의미하며 데이터양과 이용이 늘어나면서 대용량의 데이터를 저장·관리·검색·이용할 수 있는 컴퓨터 기반의 데이터베이스로 진화

- DBMS는 이용자가 쉽게 데이터베이스를 구축·유지할 수 있게 하는 소프트웨어로서 데이터베이스와 구분되며 일반적으로 데이터베이스와 DBMS를 함께 데이터베이스 시스템이라 함
- "문자, 기호, 음성, 화상, 영상 등 상호 관련된 다수의 콘텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집·축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체"로 정의 (콘텐츠란 다양한 의미전달 매체에 의해 표현된 데이터, 정보, 지식, 저작물 등의 인식 가능한 모든 자료를 의미)『ADP 가이드(KoDB)』

3. 데이터베이스의 특징

- 초기에는 일반적인 텍스트나 숫자, 그래프 형태의 데이터를 저장하였으나, 이후 정보기술이 발달하면서 그 범위는 이미지, 동영상 등을 포함한 멀티미디어로 확대. 단순한 데이터 저장에서 머무르지 않고 정보를 저장하는 지식베이스로 진화
- 데이터베이스가 단순한 저장소의 개념이 아니라 첨단 정보기술을 바탕으로 원하는 데이터를 저장·검색할 수 있는 복합체로 이해
- 데이터베이스의 일반적 특징
  - 통합된 데이터(integrated data) : 동일한 내용의 데이터가 중복되어 있지 않음
  - 저장된 데이터(stored data) : 컴퓨터가 접근할 수 있는 저장매체에 저장됨. 기본적으로 컴퓨터기술을 바탕으로 한 것
  - 공용 데이터(shared data) : 여러 사용자가 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용. 대용량화되고 구조가 복잡한 것이 보통
  - 변화되는 데이터 : 새로운 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터 유지
- 데이터베이스의 정보의 축적 및 전달 측면에서의 특성
  - 기계가독성 : 대량의 정보를 일정한 형식에 따라 컴퓨터 등의 정보처리기기가 읽고 쓸 수 있도록 함
  - 검색가능성 : 다양한 방법으로 필요한 정보 검색 가능
  - 원격조작성 : 정보통신망을 통하여 원거리에서도 즉시 온라인으로 이용 가능
- 정보이용측면 : 이용자의 정보요구에 따라 다양한 정보를 신속하게 획득할 수 있고 원하는 정보를 정확하고 경제적으로 찾아낼 수 있음
- 정보관리측면 : 정보를 일정한 질서와 구조에 따라 정리·저장하고 검색·관리할 수 있도록 하여 방대한 양의 정보를 체계적으로 축적하고 새로운 내용 추가나 갱신이 용이
- 정보기술발전의 측면 : 데이터베이스는 정보처리, 검색·관리 소프트웨어, 관련 하드웨어, 정보전송을 위한 네트워크 기술 등의 발전을 견인할 수 있음
- 경제·산업적 측면 : 데이터베이스는 다양한 정보를 필요에 따라 신속하게 제공·이용할 수 있는 인프라로서 특성을 가지고 있어 경제, 산업, 사회 활동의 효율성을 제고하고 국민의 편의를 증진하는 수단으로서 의미를 가짐

제3절 데이터베이스 활용

1. 기업내부 데이터베이스

- 정보통신망 구축이 가속화되면서 90년대에는 기업내부 데이터베이스(인하우스 DB)는 기업 경영 전반에 관한 모든 자료를 연계하여 일관된 체계로 구축, 운영하는 경영 활동의 기반이 되는 전사 시스템으로 확대됨
- OLTP(Online Transaction Processing)시스템 : 90년대 중반 이전, 정보의 수집과 이를 조직 내에서 공유하기 위한 경영정보시스템(MIS)과 생산자동화, 통합자동화 등 기업 활동의 영역별로 구축되던 시스템. 단순 자동화에 치우침
- OLAP(Online Analytical Processing)시스템 : 데이터 마이닝 등의 기술이 등장하면서 단순한 정보의 '수집'에서 탈피, '분석'이 중심이 되는 시스템 구축으로 변화하게 됨
- 2000년에 들어서며 기업 DB구축의 화두는 CRM(고객관계관리)와 SCM(공급망관리)로 바뀜

  유통·판매 및 고객데이터가 CRM과 연동되므로 CRM과 SCM은 상호 밀접한 관련을 가짐

인하우스 DB의 발전과정에서 나타난 산업 부문별 변화된 모습?	
제조부분	- 데이터베이스 기술의 가장 중요한 적용분야 00년 이전:부품테이블이나 재고관리 등 영역에서의 데이터베이스 활용이 중점 00년 이후:부품의 설계, 제조, 유통 전 공정을 포함하는 범위로 확대 - (초기)기업별 고유 시스템 형태로 구축 (이후)솔루션 유형으로 발전하게 됨 - 2000s 중반 이후: 중소기업에 대한 인하우스 DB 구축 투자 증가가 이루어짐 실시간 기업(RTE)이 대표적 화두 - 실시간 기업은 기업의 비즈니스 프로세스를 투명하고 민첩하게 유지하여 환경변화에 따른 적응 속도를 최대화하여 지연시간을 없애는 정보화 전략 → 대기업-중소기업 간 협업적 IT화의 비중 점차 확대 - (최근)제조부문의 ERP시스템 도입과 DW, CRM, BI등 진보된 정보기술을 적용한 기업내부 인하우스 DB 구축이 주류를 이룸
금융부분	- IMF 이후 금융사 간의 합병이나 지주회사 설립 등을 통해 총체적인 부실을 타파하기 위한 노력 지속 → 금융부문의 업무 프로세스 효율화나 e비즈니스 활성화, 금융권 통합 시스템 구축 등이 크게 확산 - 2000s 초반:EAI, ERP, e-CRM 등 데이터베이스 간 정보 공유 및 통합이나 고객 정보의 전략적 활용이 주된 테마 - 2000s 중반:DW를 적극적으로 도입해 관련 DB 마케팅 증대 위한 노력 가시화. 대용량 DW를 위한 최적의 BI기반 시스템 구축도 급속도로 퍼짐 - 향후 EDW(Enterprise Data Warehouse) 확장이 데이터베이스 시장 확대에 기여 예상
유통부분	- 00년 이후:전반적인 IT변화 환경에 맞물려 CRM과 SCM 구축이 이루어짐. 상거래를 위한 각종 인프라 및 KMS(Knowledge Management System) 위한 별도의 백업시스템 구축됨 - 2000s 중반:체계적인 고객정보 수집·분석과 상권분석 등으로 심화. 균형성과관리(BSC), 핵심성과지표(KPI), 웹 리포팅 등 다양한 고객 분석 툴을 통해 기존 데이터베이스와 연계 - 최근 전자태그(RFID)의 등장은 대량의 상품을 거래하는 유통부문에 적용되었을 때 파급효과가 매우 클 것으로 전망. 향후 이를 지원하는 대용량 데이터베이스를 지원하는 플랫폼이 요구되는 상황

2. 사회기반 구조로서의 데이터베이스

- 1990s 사회 각 부문의 정보화가 본격화되며 DB 구축이 활발하게 추진됨
- 무역, 통관, 물류, 조세, 국세, 조달 등 사회간접자본(SOC) 차원에서 EDI 활용이 본격화 되면서 부가가치통신망(VAN)을 통한 정보망 구축
- 지리, 교통부문의 데이터베이스는 고도화되고, 의료·교육·행정 등 사회 각 부문으로 공공 DB의 구축·이용이 확대

DB는 사회 전반의 기간재로 자리매김	
물류부분	'실시간 차량추적'을 위한 종합물류정보망 구축 CVO 서비스, EDI 서비스, 물류정보 DB 서비스, 부가서비스로 구성
지리부분	GIS 응용에 활용하는 4S 통합기술, LBS, SIM, 공간 DBMS 및 웹 GIS 지리정보유통망 가시화:지리정보통합관리소 운영, 지리정보 수요자에 정보 제공
교통부분	지능형교통정보시스템(ITS), 교통정보, 기초자료 및 통계 제공, 대국민 서비스 확대
의료부분	의료정보시스템 : 처방전달시스템, 임상병리, 전자의무기록, 영상처리시스템, 병원의 멀티미디어, 원격의료, 지식정보화 HL7 국내 표준화 작업에 따라 전국적인 진료정보 공유체계 구축계획 수립 u 헬스 실현에 기존 의료정보 데이터베이스 기반 활용
교육부분	첨단 정보통신기술(ICT)을 활용한 각종 교육정보의 개발 및 보급, 정보 활용 교육 대학정보화 및 교육행정정보화 위주로 사업 추진 교육행정정보시스템은 학사뿐만 아니라 기타 교육행정 전 업무를 처리하는 시스템

제2장 데이터의 가치와 미래

제1절 빅데이터의 이해

1. 정의

- 빅데이터(Big data) : 큰(big) 데이터
  - 단순히 용량만 방대한 것이 아니라 복잡성도 증가해 기존 데이터 처리 애플리케이션이나 관리 툴(tool)로는 다루기 어려운 데이터세트의 집합(collection of data sets)
- 빅데이터 현상은 다양한 영역에서 일어나고 있으며 정의 또한 다양
  - (일반적 정의)빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터 (Mckinsey, 2011) →활용하는 데이터 규모에 중점
  - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집·발굴·분석을 지원하도록 고안된 차세대 기술 및 아키텍쳐 (IDC, 2011) →분석비용 및 기술에 초점
  - 데이터와 데이터 처리, 저장 및 분석 기술 + 의미있는 정보 도출에 필요한 인재나 조직도 포함 (일본 노무라연구소) →정의자체가 포괄하는 범위 확대
- 3V : 빅데이터로 인한 새로운 도전과 기회를 요약(가트너그룹 더그래니)
  - 데이터의 양(Volume), 데이터 유형과 소스 측면의 다양성(Variety), 데이터 수집과 처리 측면에서 속도(Velocity) 3가지 측면의 급증으로 인한 현상
- 빅데이터를 보는 관점의 범위에 따른 정의
  - 3V로 요약되는 데이터 자체의 특성 변화에 초점을 맞춘 좁은 범위의 정의
  - 데이터 자체뿐 아니라 처리, 분석 기술적 변화까지 포함하는 중간 범위의 정의
  - 인재, 조직 변화까지 포함해 빅데이터를 넓은 관점으로 정의하는 방식

데이터 변화	기술 변화	인재, 조직 변화
· 규모(Volume) · 형태(Variety) · 속도(Velocity)	· 새로운 데이터 처리, 저장, 분석 기술 및 아키텍처 · 클라우드 컴퓨팅 활용	· Data Scientist 같은 새로운 인재 필요 · 데이터 중심 조직

• 기존 방식으로는 얻을 수 없었던 통찰 및 가치 창출
• 사업방식, 시장, 사회, 정부 등에서 변화와 혁신 주도

2. 출현 배경

- 빅데이터 현상은 없었던 것이 새로 등장한 것이 아니라 기존의 데이터, 처리방식, 다루는 사람과 조직 차원에서 일어나는 변화를 가르킴(패러다임 전환)
- 빅데이터 출현 배경
  • 산업계 : 고객데이터 축적(양질 전환 법칙)
  • 학계 : 거대 데이터 활용 과학 확산
  • 관련 기술 발전(디지털화, 저장기술, 인터넷보급, 모바일혁명, 클라우드 컴퓨팅)
- 개별 기업의 고객 데이터 축척 및 활용 증가, 인터넷 확산, 저장 기술의 발전과 가격 하락, 모바일 시대의 도래와 스마트 단말의 보급, 클라우드 컴퓨팅 기술 발전, SNS와 사물 네트워크 확산 등이 맞물려 데이터 생산이 폭발적으로 증가하면서 대세는 빅데이터 시대

3. 빅데이터 기능

- 산업혁명의 석탄, 철
- 21세기의 원유
- 렌즈 ex) 구글 'Ngram Viewer'
- 플랫폼

⇒ 차세대 산업 혁신에 꼭 필요한 요소
차세대 산업혁신에서 원재료 역할을 하면서 그 재료부터 가치를 추출하는 기법까지 아우르는 개념으로 폭넓게 쓰이고 일상생활 깊이 침투할 것

4. 빅데이터가 만들어 내는 본질적인 변화

- 사전처리 → 사후처리
  • 정해진 특정한 정보만 처리하는 것이 아닌, 가능한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아냄
- 표본조사 → 전수조사
  • 샘플링이 주지 못하는 패턴이나 정보를 얻을 수 있는 전수조사(complete enumeration)로 변화. 활용의 융통성 유지가능.
- 질 → 양
  • 대세에 영향을 주지 못하는 사례들일지라도 다른 변수에 대해서는 풍부한 정보를 갖고 있기 때문에 모든 데이터를 활용할 때, 훨씬 더 많은 가치를 추출할 수 있다는 관점
- 인과관계 → 상관관계
  • 데이터 기반의 상관관계 분석이 주는 인사이트가 인과관계에 의해 미래 예측을 점점 더 압도해 가는 시대 도래

제2절 빅데이터의 가치와 영향

1. 빅데이터의 가치

- 특정 데이터의 가치를 측정하는 것은 쉽지 않음
  • 데이터 활용 방식 : 재사용, 재조합(mashup), 다목적용 개발
    · 재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제·어디서·누가 활용할지 알 수 없음
  • 새로운 가치 창출 : 데이터가 기존에 없던 가치를 창출함에 따라 가치 측정이 어려움
  • 분석 기술 발전 : 클라우드 분산 컴퓨팅과 새로운 분석 기법의 등장으로 가치 없는 데이터도 거대한 가치를 만들어내는 재료가 될 가능성이 높아짐

2. 빅데이터의 영향

- 빅데이터가 가치를 만들어 내는 방식(빅데이터 보고서, 2011, 맥킨지)
  • 투명성 제고로 연구개발 및 관리 효율성 제고
  • 시뮬레이션을 통한 수요 포착 및 주요 변수 탐색으로 경쟁력 강화
  • 고객 세분화 및 맞춤 서비스 제공
  • 알고리즘을 활용한 의사결정 보조 혹은 대체
  • 비즈니스 모델과 제품, 서비스의 혁신 등
- 빅데이터가 시장에 미치는 영향
  • 기업 : 혁신과 경쟁력, 생산성 향상
  • 정부 : 환경 탐색, 상황분석, 미래대응
  • 개인 : 목적에 따라 활용
    ⇒ 효용 전이로 생활전반이 스마트화

제3절 비즈니스 모델

1. 빅데이터 활용사례

- 기업혁신 사례: 구글 검색 기능, 월마트 매출 향상, 질병 예후 진단 등 의료분야에 접목
- 정부 활용 사례: 실시간 교통정보수집, 기후정보, 각종 지질활동 등에 활용, 국가안전 확보 활동 및 의료와 교육 개선에 활용 방안 모색
- 개인 활용 사례: 정치인과 가수의 SNS 활용

2. 빅데이터 활용 기본 테크닉

- 연관규칙 학습(Association rule learning)
  • 어떤 변인들 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법
    ex) A를 구매한 사람이 B를 더 많이 사는가?
- 유형분석(Classification tree analysis)
  • 새로운 사건이 속하게 될 범주를 찾아내는 일
    ex) 이 사용자가 어떤 특성을 가진 집단에 속하는가?

- 유전 알고리즘 (Genetic algorithms)
  • 최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화시켜 나가는 방법
    ex) 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송
- 기계 학습 (Machine learning)
  • 훈련 데이터로부터 학습한 알려진 특성을 활용해 '예측'하는데 초점
    ex) 기존 시청기록을 바탕으로 시청자는 보유한 영화중 어떤 영화를 가장 보고 싶어 하는가?
- 회귀분석 (Regression analysis)
  • 독립변수를 조작하며, 종속변수가 어떻게 변하는지를 보며 두 변인의 관계를 파악
    ex) 구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?
- 감정분석 (Sentiment analysis)
  • 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석
    ex) 새로운 환불 정책에 대한 고객의 평가는 어떤가?
- 소셜 네트워크 분석 (Social network analysis)
  • 오피니언 리더, 즉 영향력있는 사람을 찾아낼 수 있으며, 고객들 간 소셜 관계를 파악
    ex) 특정인과 다른 사람이 몇 촌 정도의 관계인가?

제4절 위기요인과 통제방안

1. 위기요인

- 사생활 침해 : 데이터 수집이 신속 용이하고, 양이 증대됨에 따라 개인의 사생활 침해 위협뿐만 아니라 범위가 사회·경제적 위협으로 변형될 수 있음. 익명화 기술이 발전되고 있으나, 아직도 충분치 않음. 정보가 오용될 때 위협의 크기는 막대함
- 책임원칙 훼손 : 빅데이터 기반 분석과 예측 기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성도 높아짐
  • 빅데이터 시스템에 의해 부당하게 피해 보는 상황을 최소화할 장치마련이 반드시 필요
- 데이터 오용 : 데이터 과신, 잘못된 지표의 사용으로 인한 잘못된 인사이트를 얻어 비즈니스에 적용할 경우 직접 손실 발생

2. 통제방안

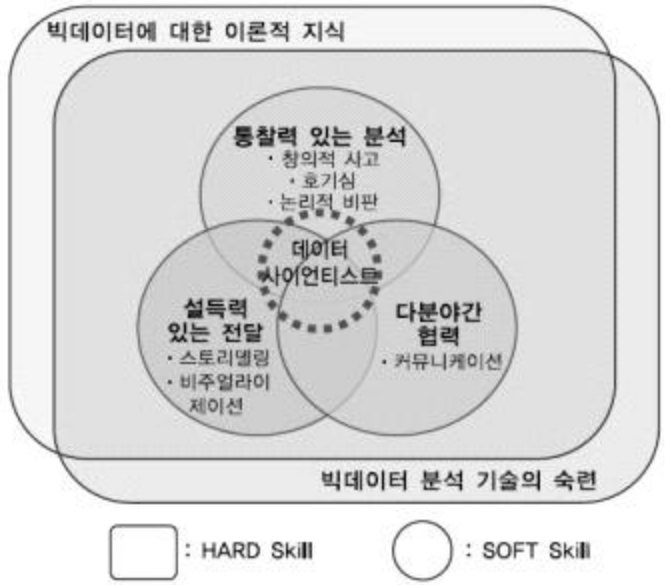
- 동의에서 책임으로
  • 개인정보 제공자의 동의'를 통해 해결하기보다 '개인정보 사용자의 책임'으로 해결
- 결과 기반 책임 원칙 고수
  • 특정인의 '성향'에 따라 처벌하는 것이 아닌 '행동 결과'를 보고 처벌
- 알고리즘 접근 허용
  • 알고리즘 접근권 보장 및 알고리즘에 의한 불이익을 당한 사람들을 대변해 피해자를 구제할 수 있는 능력을 가진 전문가로써, 컴퓨터와 수학, 통계학이나 비즈니스에 두루 깊은 지식을 갖춘 '알고리즈미스트' 대두

제5절 미래의 빅데이터

- 빅데이터 활용에 필요한 기본 3요소

구분	설명
데이터	모든 것의 데이터화 (Datafication) - 수많은 센서들이 인터넷에 연결되는 사물인터넷 시대
기술	진화하는 알고리즘, 인공지능(Artificial Intelligence) - 데이터가 알고리즘 성장의 영양분 역할: 알고리즘을 학습시킬 수 있는 데이터의 양의 증가로 알고리즘이 점점 스마트해지는 경향을 의미 - 인공지능 분야의 패턴인식, 자연어처리, 자동제어, 기계학습, 자동추론, 지능 엔진, 시멘틱 웹 등이 포함
인력	데이터 사이언티스트, 알고리즈미스트 (Algorithmist) 역할 증대

- 데이터 사이언티스트(scientist)
  • 빅데이터에 대한 이론적 지식과 숙련된 분석 기술을 바탕으로 통찰력·전달력·협업 능력을 두루 갖춘 전문인력을 의미
  • 빅데이터의 다각적 분석을 통해 인사이트를 도출하고 이를 조직의 전략 방향 제시에 활용할 줄 아는 기획자로서 전문가의 역할을 기대
- 데이터 사이언티스트의 역량과 조건



제3장 가치장조를 위한 데이터 사이언스와 전략 인사이트

제1절 빅데이터 분석과 전략 인사이트

1. 빅데이터 열풍과 회의론

- 시대의 분위기에 합류하기 위해 거액을 투자해 솔루션을 도입한 후 어떻게 활용하고 어떻게 가치를 뽑아내야 할지 첫번째 물음부터 다시 시작
- 현재 소개되는 많은 빅데이터 성공사례가 기존의 분석 프로젝트를 포장
- 과대포장은 빅데이터 분석 전체의 신뢰성에 의심을 갖게 만들거나 개념적 혼란을 불러일으켜 빅데이터 분석이 자리잡기도 전에 기반을 흔들 수 있음
- 빅데이터 분석도 데이터에서 가치, 즉 통찰을 끌어내 성과를 창출하는 것이 관건

2. 왜 싸이월드는 페이스북이 되지 못했나?

- 데이터 분석 기반 경영 문화의 부재 : 데이터 분석에 기초해 전략적 통찰력을 얻고, 효과적인 의사결정을 내리고, 구체적인 성과를 만들어 내는 체계가 없었기 때문
- 싸이월드 : 웹로그 분석과 같은 일차적인 분석이 이뤄지고 있었지만, 이는 경영진의 직관력을 보조하는 일부로서 활용
  - 사업 상황 확인을 위한 협소한 문제들에 집중하는 경향
- 성공적 인터넷 기업(구글, 링크드인, 페이스북)들은 대부분 데이터 분석과 함께 시작되고 분석이 내부 의사결정에 결정적 정보를 제공
- 분석 기반 경영이 도입되지 못하는 이유
  - 기존 관행 따를 뿐 중요한 시도 하지 않음
  - 경영진의 직관적인 결정을 귀한 재능으로 칭송
  - 적절한 방법조차 제대로 익히지 못한 사람에게 분석 업무 할당
  - 아이디어보다 아이디어를 낸 사람에게 관심을 두는 경향

3. 빅데이터 분석, 'Big'이 핵심 아니다

- 직관에 기초한 의사결정보다 데이터에 기초한 의사결정이 중요(데이터 자체의 중요성)
- 데이터의 양이 아니라 유형의 다양성과 관련
  - 중요 목표와 잠재적 보상은 다양한 데이터 소스와 신종 소스를 분석할 수 있는 능력이지 대용량 데이터 세트를 관리할 수 있는 능력이 아님
- 데이터는 크기의 이슈가 아니라, 거기에서 어떤 시각과 통찰을 얻을 수 있느냐의 문제
- 비즈니스의 핵심가치에 집중하고 이와 관련된 분석 평가지표를 개발하고 이를 통해 효과적으로 시장과 고객변화에 대응할 수 있을 때 빅데이터 분석은 가치가 있어짐
- 빅데이터와 관련된 걸림돌은 '비용이 아니라 분석적 방법과 성과에 대한 이해 부족'

4. 전략적 통찰이 없는 분석의 함정

- 데이터 크기를 떠나 전략적 분석이 주는 이점을 빠르고 구체적으로 이해해 받아들이는 것이 중요
- 분석이 사업성과에 미치는 효과(실증연구)
  - 기업이 양질의 데이터 기반을 구축하면 경영자들은 데이터 및 시스템을 활용해 더 나은 의사결정을 내리는 것에 관심의 초점을 옮김
- 분석활용과 사업성과의 상관관계
  - 성과가 우수한 기업들은 대부분 전략적으로 일상 업무에 분석을 활용
    - 성과가 낮은 기업들에 비해 무려 5배나 더 많이 전략적으로 분석을 활용했으며, 분석 지향성이 강할수록 재무성과도 우수
  - 성과가 높은 기업은 65%가 의사결정을 지원하는 역량이나 실시간 분석 역량 보유
    - 조직 전체적으로 분석을 활용하는 비율이 높음
  - 성과가 우수한 기업들도 가치 분석적 통찰력을 갖췄다고 대답한 비율이 매우 낮음
    - 기업의 핵심 가치와 관련해 전략적 통찰력을 가져다주는 데이터 분석 내재화는 쉬운 일이 아님을 나타냄
- 단순히 분석을 많이 하는 것이 경쟁우위를 가져다주지 않음
  - 소규모 부서단위에서 진행되는 분석 활동들과는 달리 좀 더 넓은 시야에서의 핵심적인 비즈니스 이슈에 답을 하는 분석은 기업의 경쟁전략을 이끌어나가는 중심

5. 일차적인 분석 vs 전략도출을 위한 필요 역량

- 빅데이터는 가치 창출이 가능해야 하고, 그 시점이 빠를수록 더 좋다
- 가치 창출을 위한 핵심 과제는 각 산업의 특성이나 경쟁의 정도, 분석의 목적, 분석을 활용하는 수준에 따라 다양
- 일차적 분석을 통해서도 해당 부서나 업무 영역에서는 상당한 효과를 얻을 수 있음
- 일차적인 분석을 통해 점점 분석 경험을 늘어가고 작은 성공을 거두면, 분석의 활용 범위를 더 넓고 전략적으로 변화시킴으로써 사업성과를 견인하는 요소들과 차별화를 꾀할 기회에 대해 전략적 인사이트를 주는 가치 기반 분석 단계로 나아가야 함
  - 이 단계에 도달하면 분석은 경쟁의 본질에 영향을 미치고 기업의 경쟁전략을 이끌어갈 수 있음
- 전략적 인사이트를 주는 가치 기반 분석을 위해 우선 사업과 이에 영향을 미치는 트렌드에 대해 큰 그림을 그려야 함
  - 인구통계학적 변화, 경제사회 트렌드, 고객 니즈의 변화 등을 고려하고, 또한 대변화가 어디서 나탈날지 예측을 통해 트렌드에 대한 큰 그림을 도출
- 전략적 수준에서의 분석은 사업성과를 견인하는 요소들, 차별화를 이룰 수 있는 기회에 대해 중요한 통찰을 줄 것

제2절 전략 인사이트 도출을 위한 필요 역량

1. 데이터 사이언스의 의미와 역할

- 데이터 사이언스란?
  - 데이터로부터 의미 있는 정보를 추출해내는 학문

• 통계학이 정형화된 실험 데이터를 분석 대상으로 하는 것에 비해, 데이터 사이언스는 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 대상으로 총체적 접근법을 사용
• 데이터 마이닝은 주로 분석에 초점되나, 데이터 사이언스는 분석뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 모두 포괄하는 개념
• 데이터공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문으로 정의

- 데이터 사이언스의 역할
• 전략적 통찰을 추구하고 비즈니스 핵심 이슈에 답을 하고, 사업의 성과를 견인
• 데이터 사이언티스트의 중요 역량 중 하나인 소통도 여기에 근거해 길러짐
- 훌륭한 데이터 사이언티스트는 비즈니스의 성과를 좌우하는 핵심요소를 정확하게 겨냥할 수 있고 이때 데이터 사이언스는 엄청난 위력을 발휘할 수 있음

2. 데이터 사이언스의 구성요소

- 데이터 사이언스의 핵심 구성요소
• 데이터 처리와 관련된 IT영역, 분석적 영역, 비즈니스 컨설팅 영역을 포괄
• Analytics : 수학, 확률모델, 머신러닝, 분석학, 패턴 인식과 학습, 불확실성 모델링 등
• IT(Data Management) : 시그널 프로세싱, 프로그래밍, 데이터 엔지니어링, 데이터 웨어하우징, 고성능 컴퓨팅 등
• 비즈니스 분석 : 커뮤니케이션, 프리젠테이션, 스토리텔링, 시각화 등
- 데이터 사이언티스트의 요구역량
• 하드 스킬(hard skill) : 데이터 처리나 분석 기술과 관련
• 소프트 스킬(soft skill) : 통찰력 있는 분석, 설득력 있는 전달, 협력 등

구분	요구 역량	내 용
하드 스킬	빅데이터에 대한 이론적 지식	관련 기법에 대한 이해와 방법론 습득
	분석 기술에 대한 숙련	최적의 분석 설계 및 노하우 축적
소프트 스킬	통찰력 있는 분석	창의적 사고, 호기심, 논리적 비판
	설득력 있는 전달	스토리텔링, Visualization
	다분야간 협력	커뮤니케이션

3. 데이터 사이언스: 과학과 인문의 교차로

- 데이터 사이언스 전문가들이 더 높은 가치를 창출해내고 진정한 차별화를 가져오는 것은 '사고방식(habits of mind), 비즈니스 이슈에 대한 감각, 고객들에 대한 공감능력' 등 전략적 통찰과 관련된 소프트 스킬
- 데이터 사이언스는 과학과 인문학의 교차로에 서있음
• 데이터 사이언티스트에게 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 글쓰기 능력, 대화 능력 등이 필요성 강조

4. 전략적 통찰력과 인문학의 부활

- 최근 사회경제적 환경의 변화
• 단순세계화에서 복잡한 세계화로의 변화(convergence→divergence)
• 비즈니스의 중심이 제품생산에서 서비스로 이동
• 경제와 산업의 논리가 생산에서 시장창조로 변화
- 인문학의 열풍
• 공급자 중심의 기술 경쟁 하에서는 '산출물'만을 중시하지만 소비자가 어디에서 재미와 편의를 느끼는지 이해하기 위해서는'창조과정'에 주목하는 인문학적 통찰력이 필요
• 기존의 사고의 틀을 벗어나 문제를 바라보고 해결하는 능력, 비즈니스 핵심가치를 이해하고 고객과 직원의 내면적 요구를 이해하는 능력 등 인문학의 역량이 점점 더 절실히 요구

5. 데이터 사이언티스트에 요구되는 인문학적 사고의 특성과 역할

- 정보 차원 : 단순히 정보를 활용한다고 할 수 있는 정도의 수준
- 통찰력 제시 : 사업 성과를 좌우하는 핵심적인 문제에 대해 대답
- 최고의 데이터 사이언티스트는 정량분석이라는 과학과 인문학적 통찰에 근거한 합리적 추론을 탁월하게 조합

구분	정보	통찰력
과거	무슨 일이 일어났는가? - 보고서 작성 등	어떻게, 왜 일어났는가? - 모델링, 실험설계
현재	무슨 일이 일어나고 있는가? - 경고	차선 행동은 무엇인가? - 권고
미래	무슨 일이 일어날 것인가? - 추출	최악 또는 최선의 상황은 무엇인가? - 예측, 최적화, 시뮬레이션

6. 데이터분석 모델링에서 인문학적 통찰력의 적용사례

- 인간을 바라보는 관점
• 타고난 성향의 관점 : 인간을 변하지 않는 존재로 상정
• 행동적 관점 : 한 사람의 행동을 지속적으로 관찰해 그 행동을 보고 사람을 판단하는 것이 더 정확하다는 관점
• 상황적 관점 : 특정한 행동을 지속하는 사람들도 주변 맥락이 바뀌면 갑작스레 행동 패턴이 변화(인간의 가변적 성향)
- 데이터 분석 모델링에서 인문학적 통찰력의 적용
• 모델의 예측력을 높이기 위해 '인간은 어떤 관점에서 바라봐야하나',
'이를 위해서는 어떤 데이터가 더 필요하며', '어떤 기술을 활용해야 할 것인가'라는 질문에 중요한 가이드 제공
- 인간에 대한 새로운 해석 관점의 제공 외에도 인문학은 '고정된 사고방식에서 벗어나 혁신을 생각하고 진부한 상상의 굴레에서 벗어난 창의성을 토대로 남보다 앞서 새로운 가치를 창출'하고자 하는 데이터 사이언티스트들에게 중요한 가치창출의 원천이 될 수 있음

제3절 빅데이터 그리고 데이터 사이언스의 미래

1. 빅데이터의 시대

- 2011년 기준 디지털 정보량 1.8 제타바이트
- 선거예측, 비용절감, 시간절약, 매출증대, 고객서비스 향상, 신규 비즈니스창출, 내부 의사결정 지원 등 상당한 가치 발휘

2. 빅데이터 회의론을 넘어: 가치+ 패러다임의 변화

- 내외부 환경의 급변할 때일수록 변화의 물결을 잘 읽어야 하며 예측하지 못했던 전환이나 위기에 빨리 적응할 수 있는 능력 필요
- 가치 패러다임 : 경제와 산업근저에는 다양한 가치 원천이 존재하며, 무작위로 작용하는 것이 아니라 특정기간 지배적으로 작용함. 이러한 가치원천은 일정기간 패러다임적인 존재로 강력한 힘을 행사하다가 효력이 다하면 다음의 가치 패러다임에게 지배적인 지위를 넘겨줌
- 가치 패러디임의 변화

구분	설명
디지털화 (Digitalization)	아날로그의 세상을 어떻게 효과적으로 디지털화하는가가 이 시대의 가치를 창출해 내는 원천 ex) 도스운영프로그램, 워드/파워포인트와 같은 오피스프로그램 등
연결 (Connection)	디지털화된 정보와 대상들이 서로 연결되어, 이 연결이 얼마나 효과적이고 효율적으로 제공해 주느냐가 이 시대의 성패를 가름 ex) 구글의 검색 알고리즘, 네이버의 콘텐츠
에이전시 (Agency)	사물인터넷(IoT)의 성숙과 함께 연결이 증가하고 복잡해짐 복잡한 연결을 얼마나 효과적이고 믿을 만하게 관리하는가가 이슈 데이터 사이언스의 역량에 따라 좌우

3. 데이터 사이언스의 한계와 인문학

- 데이터 사이언스의 한계
  - 정략적 분석이라도 모든 분석은 가정에 근거하며, 가정이 변하지 않는 동안에도 실제 외부 요인은 계속해서 변화함
  - 데이터 분석은 완벽하지 않으나, 정보가 뒷받침되지 않는 직관보다 낫다
- 데이터 사이언티스트의 역할
  - 훌륭한 데이터 사이언티스트는 인문학자들처럼 모델의 능력에 대해 항상 의구심을 가지고, 가정들과 현실의 불일치에 대해 끊임없이 고찰하고, 분석 모델이 예측할 수 없는 위험을 살피기 위해 현실 세계를 주시
  - 빅데이터와 데이터 사이언스가 빅데이터에 묻혀 있는 잠재력을 풀어내고, 새로운 기회를 찾고, 누구도 보지 못한 창조의 밑그림을 그리는 힘 발휘

과목Ⅲ. 데이터 분석 기획

제1장 분석과제 정의

제1절 개요

- 데이터의 핵심은 데이터 자체가 아닌 분석을 통한 의사결정 최적화
- 업무담당자가 의사결정을 내리기 위해 이벤트 발생부터 액션까지 지연시간(데이터지연, 분석지연, 의사결정의 지연)이 운영의 장애요인으로 발생

⇒ 데이터 기반의 분석업무 활용 체계 도입을 통해 지연시간을 줄일 수 있음

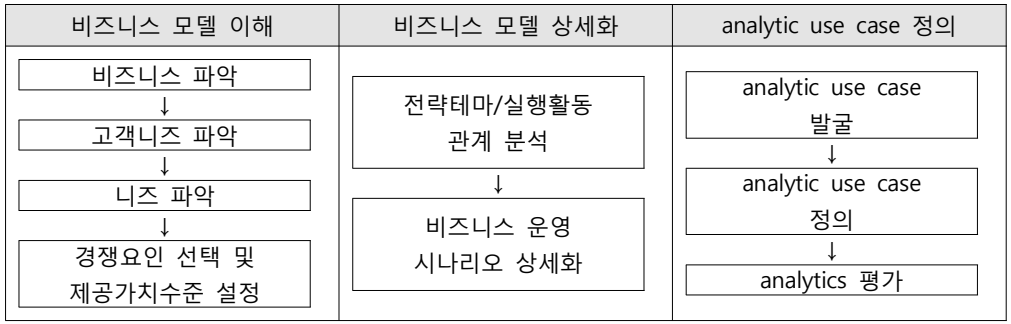
- 데이터 분석 도입의 성공요소
  - Question First 방식으로 접근 : 업무에 필요한 분석이 무언지를 찾기 위해 분석질문을 먼저 정의하고 분석하기 위해 필요한 데이터가 무엇인지 정의
  - 선택과 집중 : 핵심 분석 몇 가지만 잘해도 차별화된 복제할 수 없는 핵심 경쟁력 가짐
  - 자동화된 분석을 업무 프로세스에 내재화 : 분석은 업무 의사결정 프로세스의 일부

제2절 분석 기회 발굴

- 분석 기회 : 기업의 전사 또는 개별 업무별 주요 의사결정 포인트에 활용할 수 있는 분석의 후보들
- 데이터 분석의 업무 활용 체계 도입을 위한 접근방식
  - 톱다운 접근방식 : 전사 비즈니스 모델로부터 기업 경쟁력 향상을 도모할 수 있는 분석 기회를 발굴하고 전사 분석 체계 구현을 위한 거버넌스 체계 전반을 다룸
  - 보텀업 경로접근 : 주제별 분석 기회를 구현
- 분석 기회 발굴을 위한 3 가지 방법
  - 비즈니스 모델 분석을 통한 Top Down 방식 : 기업의 비즈니스 모델을 분석하여 경쟁력 강화를 위한 핵심 분석기회 식별
  - 대상 프로세스 선정.분석을 통한 Bottom Up Path-Finding 방식 : 특정 대상 프로세서를 선정한 후 주제별로 분석기회를 식별
  - 분석 유즈케이스 Benchmarking : 제공되는 산업별, 업무 서비스별 분석테마 후보 풀의 벤치마킹을 통한 분석 기회 식별

1. 비즈니스 모델분석을 통한 톱다운 접근방식

- 기업들이 당면한 새로운 경쟁방식을 정의하고 적용하기 위해 비즈니스 모델에 대한 검토 및 재설정이 필요
  - 핵심영역에 분석(Analytic) 도입으로 비즈니스 모델의 재설정을 가능하게 함
- 비즈니스 모델 분석을 통한 분석 기회 발굴 절차
  - 비즈니스 모델에 대한 이해→비즈니스 모델을 비즈니스 운영 시나리오를 통해 상세화→상세화한 운영 시나리오 기반으로 기업의 핵심 분석기회(Analytic Use Case) 도출
- 비즈니스 모델 분석을 통한 분석기회 발굴 절차



가. 비즈니스 모델 이해

- 기업의 비즈니스 모델을 이해하기 위해서는 먼저 당면한 비즈니스 컨텍스트를 산업요인, 시장요인, 거시경제요인, 주요 트렌드 관점으로 파악
- 가시화된 고객의 니즈 뿐 아니라 미충족 고객의 니즈라는 새로운 블루오션 영역 식별
- 이를 경쟁사와 비교하여 기업의 경쟁요인을 선택하고 제공 가치 수준을 설정

1) 비즈니스 컨텍스트 파악

비즈니스 컨텍스트	시장 요인 분석	• 마켓이슈: 고객.공급자 관점에서 시장을 이끌고 변화시키는 주요이슈 • 마켓세그먼트: 주요 마켓 세그먼트별 특성분석 및 신규 세그먼트 발굴 • 니즈와 수요: 시장 니즈 파악 및 충족수준 분석 • 전환비용: 고객이 경쟁자 쪽을 선택하게 되는 요소 분석 • 기대수익: 기대수익 및 가격 결정력 관련 요소 분석
	산업 요인 분석	• 기존경쟁사: 기존 경쟁자와 그들의 상대적인 경쟁요소 분석 • 신규진입자: 핵심 신규 진입자 확인 및 신규 진입자의 사업모델 분석 • 대체상품/서비스: 다른 시장을 포함한 잠재적인 대체 상품.서비스 분석 • 공급자/기타 value chain 내 주체들: 가치사슬상의 주요 주체 확인 • 이해당사자: 주요 이해 당사자 확인
	주요 트렌드 분석	• 기술트렌드: 사업을 위협하거나 발전.개선할 수 있는 기술 트렌드 분석 • 규제트렌드: 사업에 영향을 미치는 규정과 규제 트렌드 분석 • 사회/문화트렌드: 사업과 관련된 주요 사회.문화적 행동, 가치 변화 분석 • 경제트렌드: 사업과 관련된 주요 경제변화 패턴 분석
	거시 경제 요인 분석	• 글로벌시장 환경: 거시경제적 관점에서 현재의 전반적인 환경 분석 • 자본시장: 자사의 자본수요와 관련된 현재 자본시장 환경 분석 • 원자재 및 다른 자원: 자사에 필요한 자원의 현재가격과 가격추세 분석 • 경제 인프라: 시장의 경제 인프라에 대해 분석

2) 고객 니즈 파악

- 기업의 비즈니스 가치 창출의 시작점은 고객의 니즈가 무엇인가를 명확하게 파악하는 것부터 출발
- 고객 니즈의 4가지 가치유형
  - 기능적 가치(Functional Value): 물리적 속성, 성능, 특징 등 기능적 측면 (모바일 디바이스 이용기능, 이미지 데이터 업로드 기능 등)
  - 재무적 가치(Financial Value): 무료, 저가격, 고가격
  - 무형의 가치(Intangible Value): 공유성, 확장성, 이동성, 접근성
  - 감성 가치(Emotional Value): 즐거움, 자긍심, 자유

3) 미충족 니즈(Unmet Needs) 파악

- 분석된 고객 니즈를 경쟁 형태별로 매핑시켜 미충족된 고객 니즈 발굴
  - 경쟁사 선점 영역(Defeated Territory): 경쟁사는 충족, 자사는 미충족 영역
  - 공통 경쟁 영역(Fighting Territory): 자사와 경쟁사 공통으로 제공하는 고객니즈 영역(레드오션)
  - 자사 선점 영역(Sweet Territory): 자사는 충족, 경쟁사는 미충족 영역
  - 미충족 고객 니즈(Unmet Needs): 자사와 경쟁사 모두 아직 파악 못한 고객 니즈 영역. 기업의 새로운 비즈니스 전략 방향의 단초가 될 수 있는 영역

4) 경쟁요인 선택 및 제공가치 수준 설정

- 기업의 전략 목표 및 계획에 대해 조직의 모든 구성원이 쉽게 이해하고, 커뮤니케이션 하고, 창의적인 사고를 할 수 있도록 기업 전략을 수치가 아닌 시각적인 형태로 심플하게 표현해 주는 전략 캔버스를 활용하여 기업의 경쟁 방향 설정 가능
- 전략 캔버스를 통해 기업의 경쟁요소(Factors of Competition)별 제공되는 가치 수준을 자사, 경쟁사간 비주얼하게 비교함으로써 자사의 경쟁방향을 명확하게 이해 가능

나. 비즈니스 모델 상세화

- 비즈니스 모델 상세화를 통해 기업의 경쟁 요소를 구체적으로 정의 가능
  - 기업의 전략테마.실행활동 간 관계를 분석하고 비즈니스 운영 상세 시나리오 정의를 통해 구체화
    - 활동체계 지도(Activity System Map): 기업의 전략테마.실행활동 간 관계 분석 도구
    - 인과지도(Causal Loop Diagram): 비즈니스 운영 상세 시나리오 정의에 사용

1) 전략테마와 실행활동 간 관계 분석

- 활동체계 지도: 기업의 전략 방향을 전략테마와 실행활 동 간 관계를 통해 정의할 수 있도록 도와주는 도구
- 비즈니스 모델을 상세화
  - 비즈니스 모델단계에서 정의된 전략 캔버스의 경쟁요소에 따른 기업 가치 제공 수준을 바탕으로 전략테마 정의
  - 전략 테마를 실현하기 위한 실행활동 도출
  - 전략테마·실행활동 간 관계 정의
- 명확한 전략을 가진 조직은 몇 가지 높은 우선순위의 전략테마를 보유하고 있고 기업의 전략은 각 전략테마와 밀접하게 연결된 실행요소 군들을 통해 정의되고 실현 될 수 있음
  - 전략테마를 실현하기 위한 요소 관점에서 실행활동을 발굴하여 관계를 설정하고 관련성 있는 실행활동 및 전략 테마와의 관계를 추가 설정해 볼 수 있음

2) 비즈니스 운영 시나리오 상세화

- 전략테마와 실행활동 간 관계분석을 통해 도출된 전략테마와 실행활동을 바탕으로 선택(Choice)-이론(Theory)-결과(Consequence)의 형태로 비즈니스 운영 시나리오를 상세화해 정의할 수 있음
  - 선택(Choice) : 조직이 운영되어야 하는 방식에 대한 경영진의 의사결정 사항
    - 정책(Policies) : 기업의 모든 운영의 측면에서 일련의 실행활동
    - 자산(Assets) : 생산설비, 통신 시스템과 같은 유형의 자원
    - 거버넌스(Governance) : 자산을 사용하는 정책으로 계약(활용) 형태
  - 결과(Consequence) : 선택에 따른 결과를 발생시키는 것에 대한 이론
    - 민감한 결과(Flexible Consequence) : 선택의 변경에 민감하고 급속하게 변화되는 결과
    - 견고한 결과(Rigid Consequence) : 선택 변경이 발생해도 영향이 천천히 발생하는 결과
  - 이론(Theory) : 선택에 따른 결과를 발생시키는 것에 대한 이론
- 인과지도(Causal Loop Diagram) : 비즈니스 운용 시나리오를 상세화하기 위한 도구
  - 연관된 변수들이 서로 어떻게 영향을 미치는지를 시각적으로 표현한 다이어그램
  - 선택과 선택으로 파생된 결과를 화살표로 연결하여 인과관계를 표현한 것으로 비즈니스 운영 시나리오를 시각적으로 표현하고 이해하기에 용이
  - CLD 작성 절차
    - 선택(Choice) 목록 작성
    - 각 선택에 대한 직접적인 결과(Consequence) 찾기
    - 스텝 2 에서 발견한 결과가 자체적으로 중요한 결과 갖는지 판단
    - 결과가 없을 때까지 스텝 3 반복
    - 결과가 견고한(RIGID)한 것을 선별하고 주변에 박스 그리기
    - 선별된 결과가 몇몇 선택을 발생시키는지 체크(결과로부터 선택으로 화살표 그림)
    - 이 프로세스가 끝났을 때 선순환하는지 확인하고 순환이 얼마나 강한지 판단

다. 분석 유즈 케이스(Analytic Use Case) 정의 단계

- 분석 유즈 케이스를 발굴하고 상세정의를 통해 분석(Analytic)을 식별

1) 분석 유즈 케이스 발굴(CLD 분석)

- CLD 의 이론(Theory)을 분석하여 분석 유즈 케이스를 찾아냄
- 분석 유즈 케이스 : 분석을 적용했을 때 업무 흐름을 개념적으로 설명한 것
  - 분석 유즈 케이스는 비즈니스 모델을 구성하는 이론(Theory)을 설명
  - 분석 유즈 케이스는 하나 이상의 분석(Analytics)을 포함
  - 분석 유즈 케이스는 프로세스 혁신의 수단으로 사용 가능
- 분석 유즈 케이스는 비즈니스 용어로 명명
  - 재무 업무의 분석 유즈 케이스 : 자금 시재 예측, 서비스 수준 유지
  - 고객 업무의 분석 유즈 케이스 : 서비스 수준 유지, 고객만족 달성
  - 판매 업무의 분석 유즈 케이스 : 파이프라인 최적화, 영업성과 분석

2) 분석 유즈 케이스 정의(이벤트 반응 분석)

- 도출된 분석 유즈 케이스를 상세히 정의
- 필요한 분석을 찾아내기 위해 프로세스 흐름을 시작부터 종료까지 표현하는 이벤트 반응 다이어그램(Event Response Diagram) 활용
  - 이벤트 반응 다이어그램 : 프로세스 흐름을 시작부터 종료까지 표현하는 다이어그램
    - 분석 유즈 케이스를 상세하게 정의하고 필요한 분석을 찾아내기 위해 사용
    - 액터 : 이벤트 주체. 사람이나 기업 등의 실체 또는 정보시스템.
      이벤트를 촉발시키는 역할 또는 반응 결과를 접수하는 역할
    - 이벤트 : 반응을 촉발시키는 것(화살표 도형)
      - 외부이벤트 : 시스템 외부에서 발생하는 것. 항상 액터 존재, V+N 형태로 명명
      - 시간이벤트 : 특정 조건 또는 시기로서 액터가 존재하지 않으며 ~시기 형태로 명명
    - 반응 : 수행하는 활동(모서리 둥근 직사각형), N+V 형태로 표현.
      반응이 실행되면 새로운 이벤트 또는 다른 반응이 시작됨
    - 분석 : 가치 있는 결론을 도출하는 과정(원), N+V+분석형태로 표현.
      분석이 실행되면 반응에서 분석 수행
    - 흐름 : 액터, 이벤트, 반응 등의 선후관계와 연결관계(선). 명칭 부여 안함.
      흐름을 통해 데이터(정보)가 전달되면 데이터(정보) 이름을 적음
      하나의 반응에서 두 가지 이상의 결과가 발생하는 경우 분석 조건 표시

3) 분석 평가

- 분석 유즈 케이스 정의를 통해 발굴된 분석을 평가하여 핵심 분석(Critical Analytics)을 발견해야 함
- 핵심분석 : 비즈니스 모델의 경쟁요인과 관련되는 분석으로 가장 많은 경쟁요인과 관련될수록 핵심분석
- 분석평가는 중요도, 영향도, 난이도 별로 기업 특성에 따라 가중치를 부여하고 가중 평균값으로 평가하여 분석 우선순위 평가

2. 대상 프로세스 선정·분석을 통한 보험업 경로 접근 방식

- 특정 업무 영역을 대상으로 주제를 정하여 분석기회를 발굴하는 접근방식
- 프로세스 분석을 통한 분석 기회 발굴 절차
  - 프로세스 분류 :전사업무를 가치사슬→메가 프로세스→메이저 프로세스→프로세스 단계로 구조화해 업무 프로세스 정의
  - 프로세스 흐름 분석 :프로세스 맵을 통해 업무 흐름 상세화
  - 분석 요건 식별 :프로세스 맵 상의 주요 의사결정 포인트 식별
  - 분석 요건 정의 :의사결정 시점에 의사결정을 할 수 있게 하는 분석의 요건을 정의하고 분석의 요건을 분석 기회화 함

3. 분석 유즈 케이스 벤치마킹을 통한 발굴

- 제공되는 산업별, 업무 서비스별 분석 테마 후보 그룹을 통해 Quick and Easy 방식으로 필요한 분석기회가 무엇인지에 대한 아이디어를 얻고, 기업에 적용할 분석테마 후보 목록을 브레인 스토밍을 통해 빠르게 도출하는 방법

제3절 분석 기회 구조화

1. 유저스토리 정의

- 식별된 핵심 분석 기회(주제)별로 유저 스토리 작성 방법을 통해 분석자의 역할, 의사결정 사항, 분석을 통해 추구하는 목표 가치를 기술해 봄으로써 분석하고자 하는 바를 명확히 함
  - 업무담당자 입장에서 무엇을 의사결정 해야 하는지 정의
  - 이 업무를 잘 수행하기 위해 업무 담당자는 무엇을 알아야 하는지 정리
    - 분석기회 :알아야하는 무엇을 찾는 방법(의사결정요소 산출을 위한 데이터 분석 포인트)

→ 유저스토리를 통해 분석기회는 명확히 정의될 수 있음

2. 목표가치 구체화

- 유저 스토리를 통해 명확히 정의된 분석 기회의 목표 가치를 지표화함으로써 분석을 통해 달성하려는 사업성과를 구체화
  - 리소스 투입대비 업무적 성과 평가는 반드시 필요
- 성과는 측정 가능한 형태인 지표로 정의될 때 가장 잘 관리될 수 있으므로 분석기회를 통해 달성해야 할 목표 가치를 측정가능한 형태인 지표로 정의하여 관리

3. 분석질문 구체화

가. 분석 질문 도출

- 유저스토리를 통해 정의된 의사결정 사항의 달성을 위해 답해야 하는 질문이 무엇인지 식별

- 분석질문 :의사결정에 필요한 정보를 얻기 위해 필요한 분석을 찾기 위한 질문
  • 분석질문을 잘 만들어야 필요한 분석을 찾아낼 수 있으므로 분석 기회를 정의하는 데 있어 매우 중요한 과정
- 질문을 구체화하는 방식

1) 연속질문 방식

- 업무 담당자 입장에서 도출한 분석 기회의 답을 얻기 위해 많은 다른 분석들이 필요하며 이를 위해 질문에서 질문으로 연관된 질문들을 계속 만들어 가는 것이 중요
  • 분석기회의 답을 얻기 위해 필요한 분석 질문들을 계속함
- 분석 질문 도출 시 분석 컨텍스트 및 필요 분석까지 같이 도출 가능

2) 에이전트, 객체·장소, 이벤트 관점의 질문

- 분석질문을 만들어 나가는 방식에 대한 보완방법으로 질문을 만들어 나갈 때 에이전트(Agent), 객체·장소(Object/Place), 이벤트(Event) 관점으로 질문과 관련된 요소들이 무엇인지 식별하는 것을 통해 질문이 추가적으로 더 도출될 수 있음
- 관련 있는 분석 질문을 빠짐없이 도출하고자 하는 목적의 도구이므로 이를 사용해 분석 질문을 정의하는 것은 필요에 따라 활용여부 판단 가능
  • 에이전트 : 분석결과를 활용하거나 분석결과에 영향을 받는 이해 관계자
  • 객체·장소 : 궁극적인 분석의 답에 대한 실마리(시그널)를 확보할 수 있는 대상 및 장소(정적)
  • 이벤트 : 정책 수혜 대상자의 일련의 활동(동적)

나. 분석질문 정련

- 질문의 답이 우리가 필요한 의사결정에 필요한 답인지 재검토
  • 궁극적으로 알고자 하는 분석 기회 질문에 대한 답을 도출할 수 있는 질문인지 확인

제4절 분석방안 구체화

1. 의사결정 요소 모형화

- 분석 컨텍스트 간 상관관계를 모형화하여 의사결정을 위한 일련의 제 요소와 요소 간 관계 구체화 가능
  • 분석의 핵심 이슈와 의사결정을 위한 필요요소를 한 장의 그림으로 분명하게 설명 가능
  • 최적의 의사결정을 위해 필요한 분석(의사결정 요소) 도출

2. 분석체계 도출

- 정의된 의사 결정 모형의 분석 컨텍스트별로 수행할 분석을 정리하여 의사결정을 위한 전체 분석세트와 관계를 도출
  • 의사결정의 각 분석체계는 한번에 확정되지 않고 지속적으로 보완되는 과정을 거쳐 정렬됨

3. 분석필요 데이터 정의

- 분석체계에 따라 분석에 필요한 데이터 및 데이터의 유형을 식별하여 현재의 기업에서 보유한 데이터와 외부에서 확보해야 할 데이터 정의
  • 데이터 확보 가능성과 비용요인을 고려해 향후 분석 우선순위와 범위 조정에 활용 가능

4. 분석 ROI 평가

- 분석에 대한 경제성 평가

제5절 분석 활용 시나리오 정의

- 분석 컨텍스트를 기반으로 도출된 분석체계를 종합적으로 고려하여 업무적인 분석 활용 시나리오를 정의
- 주요 업무 의사 결정에 분석 결과가 어떻게 활용되어 업무가 효과적으로 수행할 수 있는지 명확히 이해할 수 있도록 도와줌
- 분석을 업무 운영 프로세서 반영할 때 기존 프로세서의 변경 및 신규 프로세서가 생성되는 등의 업무 프로세서의 변화가 발생하기도 함
  • 분석 업무 프로세서를 내재화하면 운영업무의 후행 액션이 분석에 의해 자동으로 실행되는 형태로 프로세서가 지능화 됨
- 분석의 업무 활용 시나리오 정의 시, 분석으로 인한 업무 프로세스 변화를 명확히 식별하고 재설계 방안을 정의해야함

제6절 분석 정의서 작성

- 분석별로 필요한 소스 데이터, 분석방법, 데이터 입수 및 분석의 난이도, 분석수행 주기, 분석결과에 대한 검증 오너십, 상세 분석과정을 정의

제7절 전사관점 분석적용 시 고려요소

- 분석업무 적용에 따른 다양한 변화 요구에 신속하게 대응하고 효율적인 분석업무를 수행하기 위해서는 아키텍처를 새로운 관점에서 정의해야 함
  • 기업의 전사 목표를 최적화하는 관점에서 조정과 정련이 필요
- 다양한 분석 간의 선순환 관계 정의, 분석 내재화를 통한 업무 운영 프로세스 정의, 전사 공통적으로 적용되는 분석 패턴을 서비스화하여 정의할 수 있는지 고려해야 함

1. 분석 선순환 구조 맵

- 분석결과에 따른 의사결정이 각 조직별로 어떻게 상충되는지의 여부를 맵을 통해 연관성을 파악해보고 각 조직의 목적에 부합될 수 있도록 상충되는 부분은 조율을 통해 전사관

점에서 최적화 될 수 있도록 해야 함
- 모델링 시 선택과 결과들이 선순환 구조가 되는지 확인하고 각 분석에 따른 결과 요소들 간에 상충되는 요소가 있는지 확인하여 조정 및 관리할 수 있도록 정/부관계 표현
• 전사 목표를 최적화하기 위해 전사관점에서 균형성 있는 분석들을 배치하고 관리할 수 있음

2. 분석 내재화 프로세스 정의
- 분석운영 프로세스의 내재화를 통해 프로세스 지능화 관점에서 실시간 의사결정을 위한 분석정보들을 제공할 수 있음
- 차별화된 경쟁력을 확보하기 위해서는 합리적 의사결정을 위한 분석을 운영업무 프로세스에 내재화함으로써 실시간 의사결정에 따른 실행과 연속적인 피드백이 이루어질 수 있도록 해야 함
• 데이터 기반의 합리적인 의사결정이 가능하도록 프로세스를 지능화하고 최적화하는 형태로 전환

3. 분석 패턴 서비스 아키텍처 정의
- 중복 분석과 일관성 문제를 해소하기 위해 전사차원에서 발생하고 있는 세부 분석 요소들을 정련한 후 공통적인 분석요소를 식별하여 패턴화하고 이를 전사관점에서 공통서비스로제공하는 방안 고려
- 유저스토리별로 도출된 분석 세트들의 분석 패턴 체계를 정리하고 서비스화하여 관리할 수 있는 분석 패턴 서비스 체계를 정의
• 유저 스토리별로 정의된 분석 세트를 전사 분석맵 상의 분석 테마와 연결하여 기업의 전략목표 달성을 위한 전체 분석 서비스 흐름을 파악하고 관리 할 수 있도록 함

제2장 분석 마스터 플랜

제1절 마스터 플랜 수집

1. 마스터 플랜 수립 개요

- 데이터 기반 구축을 위해서는 분석과제를 대상으로 전략적 중요도, 비즈니스 성과 및 ROI, 분석 과제의 실행 용이성 등 다양한 기준을 고려해 적용 우선순위를 설정할 필요
- 우선순위 뿐 아니라 분석의 적용 범위 및 방식에 대해서도 종합적으로 고려하여 데이터 분석을 구현하기 위한 로드맵 수립

2. 우선순위 평가

- 우선순위 평가 : 정의된 데이터 과제에 대한 실행 순서를 정하는 것
  - 업무 영역별로 도출된 분석과제를 우선순위 평가 기준에 따라 평가하고 과제 수행의 선후행 관계를 고려하여 적용 순위를 조정해 최정 확정
  - 일반적 IT 프로젝트는 과제의 우선순위 평가를 위해 전략적 중요도, 실행용이성 등 기업에서 고려하는 중요 가치기준에 따라 다양한 관점에서의 우선순위 기준을 수립하여 평가
- 빅데이터의 4V(Volume, Variety, Velocity, Value)를 고려한 ROI 관점
  - 투자비용(Investment)측면의 요소
    - 크기 : 데이터 규모 및 양
    - 다양성 : 데이터의 다양한 종류와 형태
    - 속도 : 데이터 생성 속도 또는 데이터 처리 속도
  - 비즈니스 효과(Return) 측면 요소
    - 가치 : 분석 결과 활용 및 실행을 통한 비즈니스 가치
      데이터 분석을 통해 추구하거나 달성하고자 하는 목표 가치
- 데이터 분석 과제를 추진할 때 우선 고려해야하는 요소
  - 시급성 : 전략적 중요도가 핵심. 분석과제의 목표가치(KPI)를 함께 고려하여 판단
  - 난이도 : 데이터를 생성, 저장, 가공, 분석하는 비용과 현재 기업의 분석수준을 고려
    데이터 분석의 적합성 여부. 해당기업의 상황에 따라 조율 가능
- 포트폴리오 사분면(Quadrant) 분석을 통한 과제 우선순위 선정 기법

난이도		
	Ⅰ • 전략적 중요도가 높아 경영에 미치는 영향이 크므로 현재 시급하게 추진 필요 • 난이도가 높아 현재 수준에서 과제를 바로 적용하기 어려움	Ⅱ • 현재 시점에서는 전략적 중요도가 높지 않지만 중장기적 관점에서는 반드시 추진돼야 함 • 분석과제를 바로 적용하기엔 난이도가 높음
	Ⅲ • 전략적 중요도가 높아 현재 시점에 전략적 가치를 두고 있음 • 과제 추진의 난이도가 어렵지 않아 우선적으로 바로 적용 가능할 필요성이 있음	Ⅳ • 전략적 중요도가 높지 않아 중장기적 관점에서 과제추진이 바람직함 • 과제를 바로 적용하는 것은 어렵지 않음

현재 ---------- 시급성 ---------- 미래

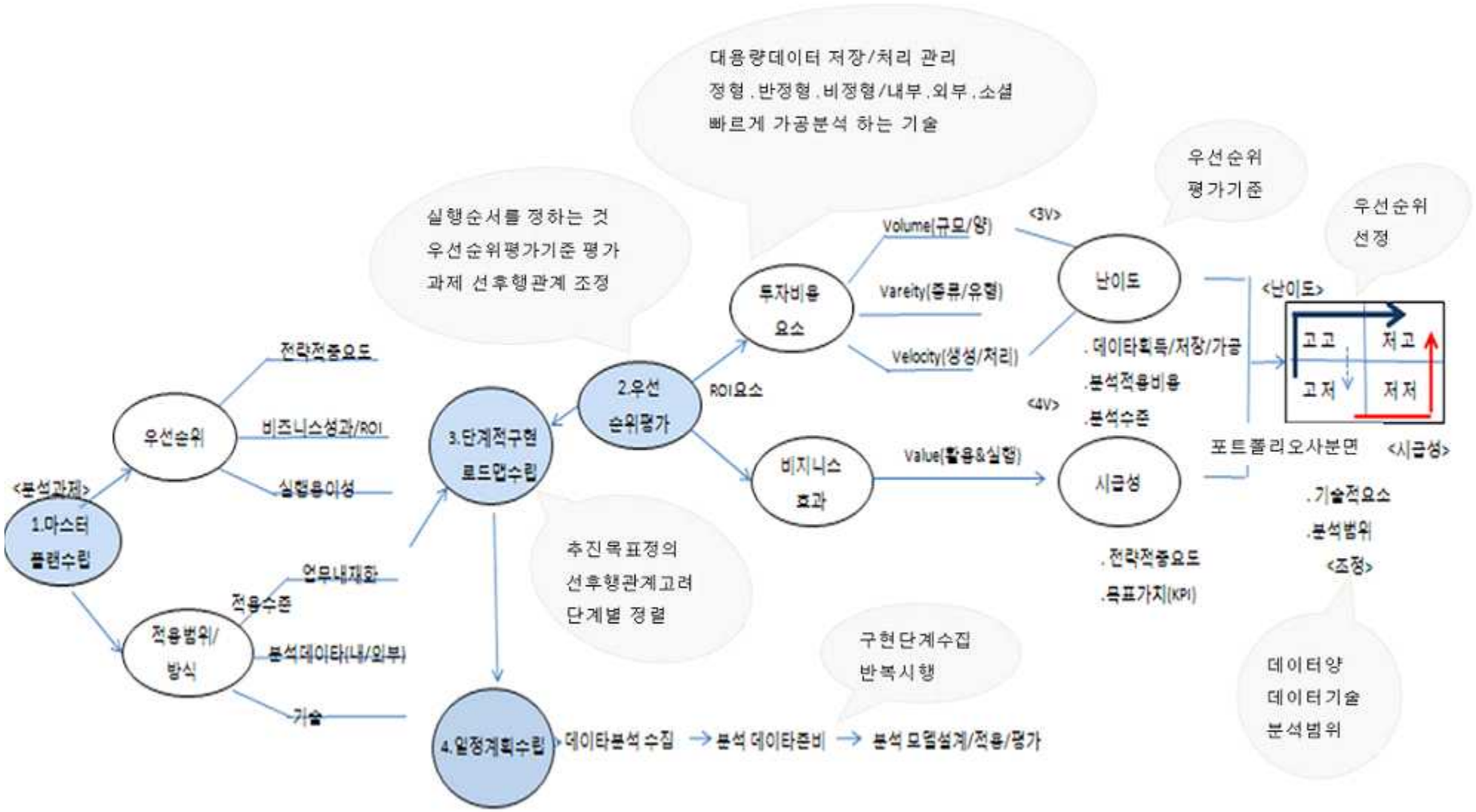
  - 시급성 기준 : Ⅲ→Ⅳ→Ⅱ 순 의사결정
  - 난이도 기준 : Ⅲ→Ⅰ→Ⅱ
  - 시급성 및 난이도가 높은 1 사분면은 경영진 또는 실무 담당자의 의사결정에 따라 적용 우선순위 조정 가능
  - 분석을 위한 기술적 요소, 분석 범위에 따라서도 분석과제 적용 우선순위 조정 가능

3. 단계적 구현 로드맵 수립

- 분석과제에 대한 포트폴리오 사분면 분석으로 1차적 우선순위를 결정하고 분석 과제별 적용범위 및 방식을 고려해 최종 실행 우선순위를 결정한 후 단계적 구현 로드맵 수립
  - 단계별 추진 목표를 명확히 정의하고 추진 과제별 선후행 관계를 고려해 단계별 추진 내용 정렬

4. 일정 계획 수립

- 구현단계 : 분석을 위한 데이터를 수집·확보하고 이를 분석을 위한 형태로 준비한 후 분석모델을 상세하게 설계
  - 준비된 데이터를 통해 모델에 적용해 보고, 적용 결과를 평가하는 과정을 반복 시행
  - 분석의 구현일정은 반복 정련과정을 고려해 수립하고 최종적으로 세부 일정계획을 수립


대용량데이터 저장/처리 관리
정형,반정형,비정형/내부,외부,소셜
빠르게 가공분석 하는 기술
실행순서를 정하는 것
우선순위평가기준 평가
과제 선후행관계 조정
우선순위
평가기준
우선순위
선정
<분석과제>
1.마스터
플랜수립
우선순위
전략적중요도
비즈니스성과/ROI
실행용이성
적용범위/
방식
업무내재화
적용수준
분석데이타(내/외부)
기술
3.단계적구현
로드맵수립
2.우선
순위평가
ROI요소
투자비용
요소
비지니스
효과
Volume(규모/양)
Variety(종류/유형)
Velocity(생성/처리)
Value(활용&실행)
<3V>
<4V>
난이도
시급성
. 데이타획득/저장/가공
.분석적용비용
.분석수준
. 전략적중요도
.목표가치(KPI)
<난이도>
고고
저고
고저
저저
포트폴리오사분면
<시급성>
. 기술적요소
.분석범위
<조정>
데이터양
데이터기술
분석범위
추진목표정의
선후행관계고려
단계별 정렬
구현단계수집
반복시행
4.일정계획수립
데이타분석 수집
분석 데이타준비
분석 모델설계/적용/평가

제2절 분석 거버넌스 체계

1. 거버넌스 체계 개요

- 데이터 분석과 활용에 대한 체계적인 관리의 중요성으로 분석 관리체계 수립 필요
  - 지속적 분석 고도화, 분석과제 추가발굴 등 기업 문화로 정착, 안정적으로 분석운영에 필요
- 분석의 지속적인 개발, 확산 및 서비스 관리를 위한 분석 거버넌스 체계는 기업의 현 분석수준을 정확히 진단, 분석 조직 및 분석 전문인력 배치, 분석 관련 프로세스 및 분석 교육 등 의 관점에서 정의 가능
- 분석의 지속적 개선/개발, 확산 및 서비스관리를 위한 거버넌스 체계
  - COA(Center Of Analysis) : 분석조직, 분석 수준진단, 분석교육, 분석 개발확산/평가 프로세스, 분석전문 인력

2. 데이터 분석 수준집단

- 데이터 분석.활용여부가 기업의 경쟁력을 좌우하는 궁극적 요소
  - 데이터 분석 수준진단을 통해 데이터 분석기반 구현을 위한 준비, 보완 등 분석의 유형 및 분석의 방향성을 결정
- 분석을 위한 준비도 및 성숙도 진단 궁극적 목표 : 각 기업이 수행하는 현재 분석 수준을 명확히 이해하고 수준진단 결과를 토대로 미래 목표수준 정의
- 수준진단을 통해 데이터 분석을 위한 기반 또는 환경이 어느 수준이고 데이터를 활용한 분석의 경쟁력 화보를 위해 어떤 영역에 선택과 집중을 해야는지, 어떤 관점을 보완해야 는지 등 개선상안 도출 가능
- 분석 수준진단 프레임워크
  - 6 개 분석준비도 + 3 개의 성숙도등 9 개 영역의 70 여개 항목으로 수행

가. 분석준비도(readiness) : 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단방법

분석업무파악	분석 인력 및 조직	분석기법
•발생한 사실 분석 업무 •예측 분석 업무 •시뮬레이션 분석 업무 •최적화 분석업무 •분석 업무 정기적 개선	•분석전문가 직무존재 •분석전문가 교육훈련프로그램 •관리자 기본분석능력 •전사총괄조직 •경영진분석업무이해	•업무별 적합한 분석기법 사용 •분석 업무 도입 방법론 •분석기법 라이브러리 •분석기법 효과성 평가 •분석기법 정기적 개선
분석데이터	**분석문화**	**IT 인프라**
•분석업무를 위한 데이터 충분성/신뢰성/적시성 •비구조적 데이터 관리 •외부데이터 활용 체계 •기준데이터 관리(MDM)	•사실에 근거한 의사결정 •관리자의 데이터중시 •회의 등에서 데이터 활용 •경영진 직관보다 데이터활용 •데이터공유 및 협업 문화	•운영시스템 데이터 통합 •EAI,ETL 등 데이터유통체계 •분석 전용 서버 및 스토리지 •빅데이터/통계/비쥬얼 분석환경

나. 분석성숙도 모델

- 데이터 시대는 분석능력 및 분석결과 활용에 대한 조직의 성숙도 수준 평가해 현재 상태 점검
  - 분석수준은 성숙 단계에 따라 점차 진화하며 산업 및 기업의 특성에 따라 각 성숙 단계의 내용은 약간 상이할 수 있음

단계	도입	활용	확산	최적화
설명	분석 시작, 환경과 시스템구축	분석결과를 업무에 적용	전사차원에서 분석 관리, 공유	분석을 진화시켜 혁신 및 성과향상에 기여
비즈니스 부문	•실적분석 및 통계 •정기보고 수행 •운영 데이터 기반	•미래결과예측 •시뮬레이션 •운영데이터 기반	•전사성과 실시간분석 •프로세스혁신3.0 •분석규칙관리 •이벤트관리	•외부환경분석 활용 •최적화업무 적용 •실시간 분석 •비즈니스모델진화
조직·역량 부문	•일부부서에서 수행 •담당자역량에 의존	•전문담당부서수행 •분석기법 도입 •관리자가 분석수행	•전사 모든 부서 수행 •분석 COE 운영 •데이터 사이언티스트 확보	•데이터 사이언스 그룹 •경영진 분석 활용 •전략 연계
IT 부문	•데이터 웨어하우스 •데이터 마트 •ETL/EAI •OLAP	•실시간대시보드 •통계분석 환경	•빅데이타 관리 환경 •시뮬레이션·최적화 •비쥬얼 분석 •분석 전용 서버	•분석 협업환경 •분석 SandBox •프로세스 내재화 •빅데이타 분석

다. 분석 수준 집단 결과

- 분석준비도와 성숙도 진단 결과를 토대로 기업의 현재 분석 수준을 객관적으로 파악 가능
  - 유관 업종, 경쟁사와 비교하여 분석 경쟁력 확보 및 강화를 위한 목표수준 설정 가능
- 분석관점에서 4가지 유형으로 분석 수준진단 결과를 구분하여 데이터 분석 수준에 대한 목표방향을 정의하고 유형별 특성에 따라 개선방안 수립
  - 준비형 : 낮은 준비도, 낮은 성숙도
    · 분석을 위한 데이터, 조직 및 인력, 분석업무, 분석기법이 적용되지 않음으로 사전준비가 필요
  - 정착형 : 준비도는 낮은편. 조직, 인력, 분석업무, 분석기법을 제한적으로 사용
    · 우선적으로 분석의 정착이 필요한 기업
  - 도입형 : 분석업무 및 분석기법 부족, 조직 및 인력 등 준비도가 높음
    · 데이터 분석을 바로 도입할 수 있는 기업
  - 확산형 : 6 가지 분석 구성요소 모두 갖춤
    · **지속적 확산이 가능한** 기업

3. 데이터 분석 조직 및 인력

- 기업의 차별화된 경쟁력을 확보하는 수단으로 데이터를 효과적으로 분석·활용하기 위한 기획. 운영 및 관리 전담 전문 분석조직 필요 제기
- 분석조직의 개요
  - (목표) 기업의 경쟁력 확보를 위해 비즈니스 질문과 이에 부합하는 가치를 찾고 비즈니스를 최적화하는 것
  - (역할) 전사 및 부서의 분석업무를 발굴하고 전문적 기법과 분석도구를 활용하여 기업 내 존재하는 빅데이터 속에서 insight 를 찾아 전파하고 이를 Action 화 하는 것
  - (구성) 기초통계학 및 분석방법에 대한 지식과 분석경험 가진 인력으로 전사 또는 부서 내 조직으로 구성하여 운영
- 분석 조직구조 3 가지 유형

집중구조	기능구조	분산구조
·전사 분석업무를 별도의 분석 조직에서 담당 ·전략적 중요도에 따라 분석조직이 우선순위를 정해 진행 ·현업 업무부서와 이원화/이중화 가능성 높음	·일반적 분석수행구조 ·별도 분석조직 없고 해당 업무부서에서 분석 수행 ·전사적 핵심 분석 어려움 ·과거실적에 국한된 분석 수행 가능성 높음	·분석조직인력을 현업부서로 직접 배치해 분석업무 수행 ·전사차원의 우선순위 수행 ·분석결과에 따른 신속한 Action가능 ·베스트프랙티스 공유 가능 ·부서 분석업무와 역할분담 명확화 필요

- 분석조직의 인력구성 : 전문 역량을 갖춘 각 분야의 인재들을 모아 조직을 구성하는 것이 바람직
  - 비즈니스, IT 기술, 분석전문, 변화관리, 교육담당 인력 등 다양하게 구성해 분석 조직의 경쟁력 극대화 가능

4. 분석교육 및 변화관리

- 기업 내 모든 직원이 분석, 업무활용을 위한 분석문화 정착을 위한 변화
- 기업에 맞는 적합한 분석업무 도출
- 분석조직 및 인력에 대한 지속적 교육과 훈련 실시
- 경영층이 사실 기반(fact-based) 의사결정 할 수 있는 문화 정착
- 분석교육의 목표 : 분석역량을 확보하고 강화하는 것
  - 분석기획자 : 분석 큐레이션 교육
  - 분석 실무자 : 데이터 분석기법 및 툴
  - 업무 수행자 : 분석기회 발굴, 구체화, 시나리오 작성법
- 분석적 사고를 업무에 적용할 수 있도록 다양한 교육을 통해 조직 구성원 모두에게 분석 기반의 업무를 정착시킬 수 있어야 함
  - 데이터를 바라보는 관점, 데이터 분석과 활용 등이 기업 문화로 자연스럽게 스며들게 확대되어야 함

※ 프로세스 혁신 3.0 이란?

통합된 데이터를 통한 분석결과에 따른 의사결정을 프로세스에 내재화시켜 혁신하는 것
통합데이터→분석→의사결정→프로세스 적용

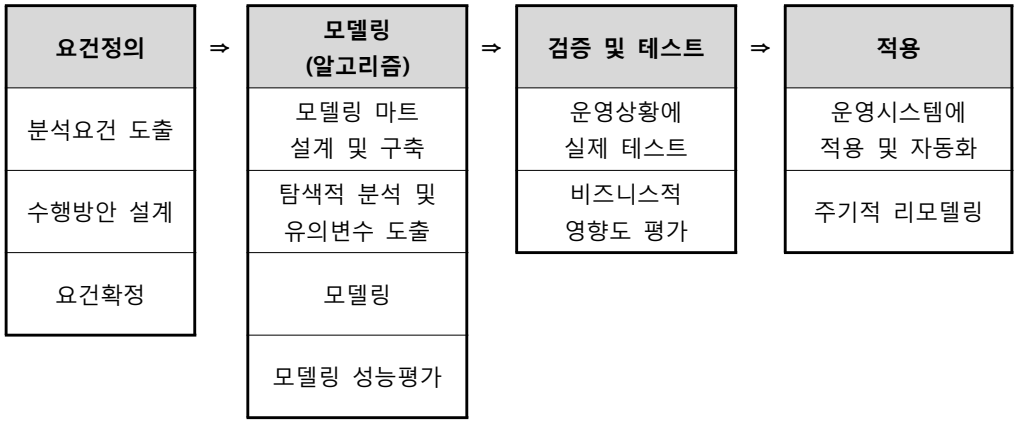
과목Ⅳ. 데이터 분석

제1장 데이터 분석 개요

제1절 데이터 분석 프로세스

1. 요건정의

- 분석요건을 구체적으로 도출·선별·결정하고, 분석과정을 설계하고, 구체적인 내용을 실무담당자와 협의하는 업무
- 광범위하고 다양한 정보를 다루고 문서화 작업의 비중이 높음
- 전체 프로세스 중에서 가장 중요한 부분으로, 빅데이터 분석업무의 성패 좌우



가. 분석요건 도출

- 요건은 비즈니스 이슈로부터 도출
  - 이슈 : 업무를 수행하는 데 있어 수익 증가, 비용증가, 상황변화, 처리속도 지연 등을 발생시키는 항목→전사적 측면에서 개선돼야할 사항
  - 단순 불편 사항이나 불만사항을 요건으로 정의하면 비즈니스적 의미가 낮아지고 분석결과 보고나 실행의 타당성 잃을 가능성 높음
- 다양한 이슈에서 진정한 요건이 될 수 있는 항목 선정하는 것 매우 중요

이슈 리스트 작성 → 핵심 이슈 정의 → 이슈 그룹핑 → 해결방안 정의

- 분석요건 도출단계는 기획단계와 유사 but 상세하게 접근하고 실무 측면으로 진행
- 분석요건의 조건은 문제를 해결했을 때 투자수익(ROI)으로 증명 가능해야 함
- 요건정의 단계
  - 상세한 분석보다 문헌조사 및 이해와 간단한 기초분석 수행
    · 요건으로 제시된 내용에 대한 사실 확인 및 통찰 도출로 방향성 설정에 필요한 수준
    · 요건정의에 많은 시간 할당하면 전체 업무진행에 차질
    · 전문가의 방향성 제시와 이해관계자들 간의 합의가 중요
- 기획단계 요건정의
  - 캠페인 반응율을 개선해야한다. 수준이나 재구매 유도 캠페인을 개선·강화해야한다. 정도
- 데이터마이닝 단계 요건정의
  - 캠페인 반응율 개선을 통한 CRM 업무 효율성 증대, 캠페인 채널 비용 절감, 캠페인 대상 20% 증대 방안 정도
- 분석단계 요건정의
  - 재구매 유도 캠페인의 대상 고객 20% 확대 방안에 대해 전체 고객구성 상황, 현재 재구매 캠페인 대상, 캠페인별 대상이 어떻게 정의되고 변경되는지, 성과는 어떤지, 미흡한 점, 어디서 재구매 대상 늘리고 이에 대한 비용은 어디서 보충할 것인지 등 정의
- 요건
  - 데이터마이닝의 요건 : 캠페인 반응율 개선을 통한 업무 효율성 또는 비용절감
  - 시뮬레이션 요건 : 의약품 분리장비 추가도입에 따른 업무시간 및 재무효과의 변화 검토
  - 최적화 : 병원의 간호사 배치에 대한 진료과별 최적 할당, 근무 시간표 최적할당 등
- 요건에 대한 현재의 이슈와 실상은 무엇인지, 어떻게 개선할지, 어느 정도 개선가능한지 등을 보완자료로 추가

수행준거	고려사항
- 데이터 분석 업무의 배경, 주요이슈, 기대효과, 제약사항 파악 가능 - 이해관계자들과 의사소통을 통해 데이터 분석요건 식별 가능 - 식별된 데이터 분석요건을 현업의 문서를 수집해 일부 수행함으로써 기획단계에서 간과할 수 있는 사항을 상세화·구체화 가능 - 상세화·구체화한 데이터 분석요건 명세화 가능 - 종합적으로 분석요건의 적합성 평가 가능해야 함	- IEEE에서 정의한 요구공학 프로세스를 고려 ・요구사항 추출, 분석, 명세, 검증, 유지보수 등 - IIBA에서 식별한 일반적으로 인정된 요구사항 도출 테크닉 고려 ・브레인스토밍, 기존문서검토, 외부 인터페이스 분석, 집중 집단 인터뷰, 관찰 또는 직무체험, 요구사항 도출 워크숍, 인터뷰, 설문 등 - 개별 분석요건에 대한 지나친 상세화보다 기존 분석 자료와 정보를 기반으로 분석요건 항목들을 누락 없이 식별하는 것에 집중 - 분석요건 분석과정에서 더 구체화되고 수정되는 것이 타당 - 데이터 분석 업무 이해 당사자들과의 긴밀한 커뮤니케이션이 필수적 - 데이터 분석 기대 효과에 대한 명확한 사전 정의와 협의가 필수적 - 개인정보 보호, 접근 통제 등 정보 보안 정책과 충돌할 수도 있기 때문에 이에 대한 사전 확인·협의가 필수적

나. 수행방안 설계

- 정의한 분석요건에 따라 구체적인 수행방안 설계
- 분석을 구체적으로 수행하기 위해서 간단한 탐색적 분석을 수행하며 미리 가설을 수립해

어떤 분석을 수행할지 틀을 잡아야 함

• 분석요건이 정해졌다고 수행방안이 확정되는 것은 아님

• 절차와 방안을 수립해야 하는 이유

· 탐색적 분석을 하며 분석자체가 의미 없다는 것을 미리 파악할 수 있는 기회를 얻을 수 있음 ⇒ 자원과 비용, 시간낭비 방지 가능

· 미리 가설을 수립해 수행방안을 설계하지 않고 진행하면 분석 필수항목과 선택항목, 일정, 필요한 자원의 양 등 계획 수립이 어려워짐 ⇒ 품직이나 납기 준수 어려워짐

• 반드시 선험적 지식을 통해 수행방안을 구체적으로 설계해야 함

- 반드시 분석기법을 정의하고 진행해야 하며 결정시 해당 분석기법에 대한 전문지식을 갖춘 인력이 참가해 검토해야 함

• 다양한 분석기법을 이해한 전문가가 적합한 분석기법을 다양한 측면에서 검토해 가장 적합한 방법 제시할 수 있어야 함

- 빅데이터 기획단계에서는 전체 로드맵과 선행 및 후행 과제만 정의됨

• 수행방안의 최종 산출물: 분석계획서와 WBS(Work Breakdown Structure)

· 일(day) 단위, 상위 기획단계에서 미처 고려못한 구체적 업무와 자원, 선행관계 등의 충돌로 일정이 부족할 수 있음

· 분석계획서 : 핵심적 분석항목과 구체적 분석범위를 지정해 분석범위를 명확히 하고 관련 업무와의 선·후행 관계를 검토하기 위해 이에 대한 WBS 를 일단위로 작성

<분석수행 관리를 위한 WBS 예시>

Phase	Task	Method	Resources	W1				
				D1	D2	D3	D4	D5

- WBS 작성시 우선 Forward 방식으로 전개를 해보고 납기를 만족시킬 수 있는지 확인

• 납기 초과할 경우 납기기준으로 Backward 전개해 언제 특정업무를 시작해야는지 파악

· 자원추가, 일정조절 등 요건 조절로 일정상 충돌 해결

- 인력이 인프라 기술과 분석기술 양 업무를 동시에 수행하는 것은 피해야 하며 기술인력은 특정기간에 제한적으로 필요한 경우가 많으므로 해당 시점에만 투입할 수도 있음

- 동일업무에 대해 기술 담당자와 분석 담당자 누가 해야 할지에 대해서는 처리속도 및 IT 자원의 효율적 활용이라는 기준에서 결정해야 함

• 수백 TB 데이터는 Hive 등에서 처리 요약해 1TB 이하로 만든 다음 R 에서 처리가 적합

수행준거	고려사항
- 권한 및 계정을 확보해 DB 접근 환경 구축 가능 - 분석 대상 데이터의 존재 여부와 품질 확인 가능 - 간단한 기초분석을 통해 분석수행 타당성 확인 - 분석기법, 수행단계 및 절차, 인도 산출물, 주요일정, 수행 인력을 식별하고 구성해 분석 방법론 구축 가능 - 구축된 분석 방법론을 기반으로 분석 프로젝트 수행계획 수립 가능	- 분석 수행 방법론 구축 시 프로젝트 관련 지식 체계를 참조 및 활용 · 프로젝트 통합관리, 범위관리, 시간관리, 비용관리, 질관리, 인력관리, 의사소통 관리, 위험관리, 조달관리, 이해당사자 관리 - 분석 프로젝트에는 일정계획, 수행 조직 및 역할·책임, 표준인도 산출물, 품질 관리 계획, 위험관리계획, 의사소통 계획 등이 포함될 수 있음 - 필수 분석 항목과 선택 분석 항목을 사전에 구분해 우선순위를 부여하고 우선순위가 높은 필수 분석항목들이 작업 대상에서 누락되지 않도록 함 - 예상 결과가 나오지 않을 경우 대안적 접근 방안으로 분석 항목들 추가·식별 가능 - 데이터 오류 또는 분석 수행 오류 등으로 인한 재작업 시간을 분석일정에 반영 - 데이터 오류 및 시스템 성능 부족 현상 발생 등 관련 위험들 사전 식별하고 대응 방안 수립

다. 요건 확정

- 요건도출과 분석계획을 수립하면 어떻게 요건에 접근하고 어떤 정량·정성적 효과 나올지 기획안이 나옴

• 이를 통해 분석 요청 부서와 IT 부서, 기타 연관 부서와 공유해 최종 요건 확정

- 대론(對論) 기획단계에서 나온 분석과제가 기각될 수도 있음

• 자세한 현황과 내용을 정의하는 과정에서 대론 기획 단계의 오류 발견 가능

• 사전에 충분히 소통하지 않을 경우 요건 확정이 어려우므로 사전에 지속적으로 대화·조율하며 요건 확정

· 분석은 복잡성과 전문성이 필요해 짧은 시간 안에 상대방으로부터 이해 구하기 어려움

• 한번 확정된 요건을 종료(closing)해 이후 변경하는 일이 없도록 해야 함

· 확정된 요건이 바뀌기 시작하면 다시 반복 작업으로 시간 보낼 수 있으므로 요건을 명확히 처리·결정

• 실무에서 모델링 과정 중 요건이 변경되는 일은 빈번히 발생. 프로젝트 완료일을 준수할 수 있는 범위에서 조율

수행준거	고려사항
- 상세화·구체화·명세화한 데이터 분석요건 항목을 기준으로 추진 의미가 있는지 최종 결정 - 이해 관계자들에게 설명할 수 있음 - 공식 변경 관리를 통해 데이터 분석요건 항목들 변경 가능 - 분석요건에 대한 적합성·타당성·일정 계획에서의 제약을 종합해 업무범위 조정 가능 - 확정 데이터 분석요건 항목들을 변경 이력 및 추적성을 확보해 현행화 가능 - 데이터 분석요건을 문서화해 이해관계자들 간 공식적으로 확정 가능	- 데이터 분석요건 변경은 반드시 공식 변경관리 절차에 따라 이뤄져야 함 - 데이터 분석요건은 특정 이해관계자의 의견 위주로 확정하기보다 참여자들의 다양한 시각과 의견이 폭 넓게 수집·수렴·고려해 확장 - 이해관계자들 간의 의견 불일치를 최소화하고 만약 의견 대립시 이를 적극 조율 - 요건 확정 이후에 데이터 분석요건 변경은 전체 프로젝트에 큰 영향(대부분 부정적)을 미치므로 모든 이해관계자들의 공감대 아래 진행

2. 모델링

- 요건정의에 따라 상세 분석기법을 적용해 모델을 개발하는 과정
  - 모델링을 거치면 필요한 입력 데이터에 대한 처리가 매우 용이해짐
  - 시뮬레이션이나 최적화에서 필요한 자료가 빅데이터 분석 시스템에 이미 존재할 가능성 높음
  - 최적화에서도 제약조건에 해당하는 값이 실제 어떠했는지 시스템에 존재
- 가정이나 인터뷰해 값을 구할 일이 없어져 모델링 시 데이터 획득 및 검증에 소요되는 시간 크게 감소
- 모델링은 해당기법에 대한 전문 지식이 필요

가. 모델링 마트 설계와 구축

- 어떤 모델링 기법을 사용하든 모델링을 위한 데이터를 준비해 시스템에 체계적으로 준비해 놓으면 모델링 용이해짐
  - 모델링 도구에 따라 DBMS 에서 직접 값을 가져와 반영할 수 있는 기능도 제공
- 모델링 진행 전에 필요한 데이터의 마트를 설계해 비정규화(De-normalized) 상태로 처리하면 사용이 편리
  - 데이터 마이닝에서 지도학습(Supervised Learning)은 모델링 마트를 직접 이용해 모델 개발 가능

수행준거	고려사항
- 다양한 원천 데이터로부터 분석대상 데이터 획득 가능 - 분석 대상 데이터를 탐색·정제·요약 등 전처리해 변수들 식별 가능 - 분석 대상 데이터를 구조화하는 모델 마트 설계 가능 - 전처리한 분석 대상 데이터를 적재해 모델 마트 구축 가능	- 데이터 원천은 관계형 DB, 데이터 웨어하우스, 시스템 로그, 비정형 데이터 등 다양한 형태로 존재 가능 - 분석 대상 데이터(변수)는 연속형과 범주형으로 구분 가능 ・연속형 : 주어진 범위 내 연속되는 실수로 구성 ・범주형 : 수치형과 텍스트형으로 구분, 명목형과 순위형 변수로 구분 - 재활용성이 높은 모델 마트 설계·구축을 위해 원천 데이터에 대한 명확한 이해가 선행되어야 함 - 기존 정보시스템 내의 데이터를 최대한 활용·확장하는 접근을 하며 신중히 채택된 가설 기반으로 마트를 설계해 작업 효율성 최대화 - 데이터 획득·수정·확정이 지연될 우려가 크므로 계획된 기간 내에 데이터 획득과 확정을 강제해 현실적인 작업 수행 유도 - 데이터 정제 시 1단계(데이터요약), 2단계(파생변수도출), 3단계(변수확대)의 단계별 접근 기법 권고

나. 탐색적 분석과 유의변수 도출

- 데이터 마이닝에 해당하는 업무로 해당 비즈니스 이해와 분석요건에 대한 구체적 팩트를 발견해 통찰을 얻기 위해 수행하는 업무, EDA(탐구 데이터분석, Exploratory data analysis)
  - EDA 는 시간이 많이 필요한 일로 최근에는 EDA 를 자동으로 신속하게 수행해 유의미한 값만 파악해 데이터 마트로 만든 후 모델링 업무로 진행하는 게 일반적
- 유의미한 변수를 파악하는 방안
  - 목표값(target value)별로 해당 변수가 분포된 값을 보고 해당 변수의 구간에서 차이가 큰지 파악 → 구간 존재하면 유의미한 변수임을 시각적으로 알아볼 수 있음
  ・이 단계와 최종 분석결과를 산출해 결과를 공유하는 단계는 시각화가 매우 중요한 역할 → 전문적 지식이 없는 사람들의 이해 도울 수 있음(효율적 정보 제시, 전문적 시각화×)
- 시각화로 정보 제시 시 유의할 점
  - 모양보다 팩트와 통찰을 전달할 수 있는 것에 중점 → 단순 그래프 출력 지양
  - 시각화와 제시하고자 하는 정보의 차이 존재
  ・추세변화, 비교에 적합한 그래프 형식 선택은 필수적, 불필요한 스케일 조절은 지양

수행준거	고려사항
- 분석 목적과 요건, 데이터 특성을 기반으로 적합한 데이터 분석기법 선정 가능 - 선정된 데이터 분석기법을 기준으로 분석 모형 설계 가능 - 설계한 분석 모형을 기준으로 유의성을 분석해 높은 유의성을 보유한 변수들 식별	- 분석 모형 설계·구축 시에는 해당 모형의 학습·평가·검증을 통해 최적 모형을 선정 및 적용하기 위해 하나 이상의 모형을 준비하는 것이 타당 - 탐색적 분석을 통해 준비된 데이터의 가설 적합성과 충분성을 사전 검증해야 함

가능 - 높은 유의성을 보유한 최소한의 변수들로 분석 모형 구축 가능	- 변수의 유의성 검증 후 유의성이 높은 최소한의 변수들로 분석모형을 검증할 것을 권장 - 시뮬레이션을 통해 기 수립된 분석 모형의 타당성과 적합성을 판단해 반복적으로 보정 - 최소한 시간에 탐색적 분석을 완료하는 것이 성공적 분석의 관건으로 단위 분석에 대한 예상 소요 시간을 추정해 필요 시 샘플링 권고 - 탐색적 분석과 유의변수 도출 과정에서 정보의 부족함 식별 시 신속하게 추가변수를 개발해 데이터마트에 반영

다. 모델링

- 개념적인 모델링도 있지만 결국 이를 구현해 적용 가능해야 함
  - 전체 내용을 제대로 제시하려면 특정 도구를 사용해야 함
  - SQL 은 차이가 거의 없고 표준이라 할 수 있는 ANSI SQL 이 있으나 주요 DBMS 공급사들은 자사 특성에 따라 다양한 기능을 추가·제시해 ANSI SQL 로 활용 및 적용에 대한 정보를 제시할 수 있는 것은 매우 제한적
    → SQL 의 경우도 특정 공급사의 SQL 을 이용해 제시함으로써 이해 및 실습과 적용에 도움줄 수 있음
  - 가장 광범위하게 사용되고 학습을 위해 획득이 용이한 DBMS 를 선택해야 함
- 데이터마이닝, 시뮬레이션, 최적화별로 산업에서 시장 점유율이 높은 분석도구들이 다양하게 있고 일부는 데이터마이닝 도구에서 시뮬레이션이나 최적화를 지원하기도 함 ex) R
  - R : 오픈소스, 데이터 입수 및 변환, 분석용 마트생성, 기초통계 및 다양한 분야의 시각화, 시뮬레이션, 최적화 지원
  - 시뮬레이션은 매우 전문적인 불연속(discrete) 시뮬레이션 모델이 가장 많이 사용됨

수행준거	고려사항
- 다양한 모델링 기법을 능숙하게 다뤄 업무 특성에 적합한 기법을 선택하거나 모델링 기법을 결합해 적용할 수 있어야 함 - 선택된 모델링 기법을 이용해 모델링 - 미래값을 예측하는 데 프로세스적인 측면이 없으면 데이터마이닝 모델링을 수행 - 프로세스 및 자원에 대한 제약이 있고 입력값이 확률분포를 갖는 경우 시뮬레이션 기법 선택 - 프로세스 및 자원에 대한 제약이 있고 상수값을 가질 때는 최적화 기법 사용 - 경우에 따라 시뮬레이션과 최적화를 결합해 접근 가능	- 데이터마이닝 모델링은 통계적 모델링이 아니므로 지나친 통계적 가설이나 유의성에 집착하지 말아야 함 - 충분한 시간이 있으면 다양한 옵션을 줘서 시도하며 일정 성과가 나오면 해석과 활용 단계로 진행할 수 있도록 의사결정해야 함 - 분석 데이터를 훈련 및 테스트 데이터로 6:4, 7:3, 8:2 비율로 상황에 맞게 실시 - 훈련 및 테스트 성능에 큰 편차가 없고 예상 성능을 만족하면 중단 - 과도한 성능에 대한 집착으로 분석 모델링의 주목적이 실무 적용에 있음을 간과하고 시간을 낭비하면 후속 검증 및 적용에 지연 발생 가능

라. 모델링 성능평가

- 모델링 성능을 평가하는 기준은 분석 기법별로 다양
- 데이터마이닝은 정확도, 정밀도, 디텍트 레이트(detect rate), 리프트(lift) 등 값으로 판단
- 시뮬레이션에서는 Throughput, Average Waiting Time, Average Queue Length, Time in System 등의 지표 활용
- 최적화에서는 최적화 이전 Object Function Value와 최적화 이후 값의 차이를 구해 평가

수행준거	고려사항
- 분석 모형 적합성 판단 기준 수립 가능 - 분석 모형별 학습용 데이터 집합 구축 가능 - 구축된 학습용 데이터로 분석 모형 조정 가능 - 학습용 데이터를 활용해 조정한 분석모형에 검증용 데이터를 적용해 학습용 데이터 기반 결과와 검증용 데이터 기반 결과를 비교·분석 가능 - 검증 결과에 따라 필요 시 분석 모형과 데이터(항목, 건수)를 조정해 최적화 가능 - 선정된 기법(방법)으로 분석모형을 실제 운영환경에 적용할 수 있으며 오픈소스 R을 이용할 때는 샤이니(Shiny)를 이용해 배포 가능	- 업무 특성에 따라 다양한 모델링 기법을 선택하거나 결합해 적용가능해야 함 - 미래 값을 예측하는 데 프로세스적 측면이 없으면 데이터마이닝 모델링 수행 - 프로세스 및 자원에 대한 제약이 있고 입력 값이 확률분포를 가지면 시뮬레이션 기법 선택 - 프로세스 및 자원에 대한 제약이 있고 상수값을 갖는 경우는 최적화 기법 사용 - 경우에 따라 시뮬레이션과 최적화를 결합해 접근 가능 - 데이터마이닝 모델링은 통계적 모델링이 아니므로 지나친 통계적 가설이나 유의성에 집착하지 말아야 함 - 다양한 옵션에 대한 시도는 충분한 시간이 있으면 실시하며 일정 성과가 나오면 해석 및 활용적 측면 단계로 옮겨가야 함 - 훈련 및 테스트 데이터의 비중은 6:4, 7:3, 8:2 비율로 프로젝트 수행 경험에 비춰 최적의 조합으로 구성해 수행할 것 권고 - 훈련 및 테스트 성능에 큰 편차가 없고 예상 성능을 만족하는 시점에 작업 완료 가능 - 성능에 대한 과도한 집착으로 인해 분석 모델링의 실무 적용이라는 핵심 목적이 간과되고 후속 검증 및 적용에 지연이 발생 가능함을 염두

3. 검증 및 테스트

가. 운영상황에서 실제 테스트

- 업무 프로세스에 가상으로 적용해 검증, 분석과 운영 간 연계 검증 및 전체적인 흐름을

통합적으로 시험 하는 과정

수행준거	고려사항
- 구축 및 조정된 분석 모형을 테스트하기 위한 유사 운영환경 구축 - 분석 모형을 테스트하기 위한 절차 설계 - 설계된 절차에 따라 테스트하고, 결과 분석 - 테스트 결과를 기반으로 분석 모형을 조정해 반복 테스트 - 최종 테스트 결과를 기분으로 실제 운영환경 적용 여부를 판단 가능	- 모형의 유형에 따라 과적합화(overfitting)가 발생할 수 있음 - 실제 운영환경 성능 테스트는 사전 시나리오를 따라 1주일 정도 실시 - 일 단위 측정이 가능한 경우, 1주간의 성능이 일관됨을 확인 할 것 - 결과는 일 단위로 공유해 실무적용의 객관성 유지 - 조직변화관리와 병행 - 성능테스트는 최소 3회 이상, 테스트 기간은 최소 1주 이상 - 외부 이해관계의 개입을 최소화 또는 차단해, 결과 왜곡 방지

나. 비즈니스 영향도 평가

- 분석 결과의 정확성을 높여 만족도 개선·추가 수익 창출 등 비즈니스 영향도와 효과를 산출 할 수 있어야 함. 테스트를 통해 나온 최종 결과를 기반으로 정량적 효과 도출 가능

수행준거	고려사항
- 모델링 성과에서의 검출률(Detection rate)이 증가하거나 Lift가 개선 돼 발생하는 정량적 효과 제시 - 타 모델링과의 중복에 따란 효과를 통제·제시 할 수 있어야 함 - 기대효과는 수익과 투자대비효과(ROI, Return On Investment)로 제시	- 투자대비 효과 정량화 기법: 총 소유비용(TCO, Total Cost of Ownership), 투자대비효과(ROI), 순현재가치(NPV, Net Present Value), 내부수익률(IRR, Internal Rate of Return), 투자회수기간(PP, Payback Period) - 데이터마이닝모델링에서는 Detection rate 이 증가하거나 Lift 개선돼 발생되는 정량적 효과 제시 - 시뮬레이션에서는 처리량, 대기시간, 대기행렬의 감소를 통한 정량적 효과 제시 - 최적화에서는 목적함수가 증가한 만큼의 정량적 효과 제시

4. 적용

- 분석결과를 업무 프로세스에 완전히 통합해 실제 일·주·월 단위로 운영하는 것
- 분석 시스템과 연계돼 사용될 수 있고 별도 코드로 분리돼 기존 시스템(legacy system)에 별도 개발해 운영 가능

가. 운영시스템에 적용과 자동화

- 운영시스템에 적용해 운영하면 실시간 또는 배치 스케쥴러(Batch Scheduler) 실행하고 주기별로 분석모델의 성과가 예상했던 수준으로 나오고 있는지 모니터링 할 수 있도록 DBMS에 성과자료 누적하고 이상현상이 발생하면 자동으로 경고(Alert) 하도록 함
- 분석모델은 개발된 내용이 많아질수록 상시 파악이 자동으로 이뤄지고 이상 시에만 확인하도록 프로세스를 수립해놔야 분석업무를 다양한 분야에 적용하고 정교화를 계속해 지속적인 성과를 거둘 수 있음
- R을 이용해 이 단계를 단순화 할 수 있으며 R studio에서 제공하는 샤이니(Shiny)를 이용해 모델링 결과를 사용자 작업파일과 서버상의 파일을 이용해 간단히 배포할 수 있음

수행준거	고려사항
- 분석 모형 적용에 따른 기존 업무 프로세스 영향도와 개선 기회 분석 가능 - 식별된 기존 업무(비즈니스) 프로세스 영향도와 개선 기회를 바탕으로 목표 업무(비즈니스) 프로세스 설계와 문서화 가능 - 분석 모형의 운영환경 적용을 위한 다양한 방법들의 특징·장단점 비교 분석 가능 - 비교·분석 결과를 기준으로 분석모형 적용 기법(방법) 선정 가능 - 선정된 기법(방법)으로 분석모형을 실제 운영환경에 적용 가능	- 최종 모델링 결과를 실제 운영 정보 시스템에 적용하는 단계로 상용 또는 오픈소스 도구의 활용 또는 자체 개발 고려 가능 - 모델 적용 자동화 및 모델 갱신 자동화를 고려할 수 있으나 전용(상용 또는 오픈소스) 도구에서 해당기능 제공시에만 적용하는 것이 타당 ・적용하는 것으로 결정할 경우 적용 대상 데이터의 볼륨과 처리 속도를 고려해야 함 - 시뮬레이션은 모델 적용을 위한 프로세스와 업무 규칙이 문서화되고 이해관계자 간 공유돼야 함 - 최적화는 최적화 솔루션의 결과를 시스템과 인터페이스 할 수 있도록 데이터베이스 연동 프로그램을 개발해야 함

나. 주기적 리모델링

- 비즈니스 상황 변화나 분석결과 적용에 따른 주변 요인들, 분석결과 적용 시 고객의 행동패턴 변화 등은 자연스러운 성과(부정적×)로 이런 변화에 시스템이 대응 가능해야함
- 성과 모니터링이 지속적이어야 하고 일정수준 이상의 편차가 지속적으로 하락하는 경우 리모델링을 주기적으로 수행해야 함
- 일반적으로 주기적 리모델링은 분기, 반기, 연 단위로 수행
  - 데이터 마이닝 : 평균 분기별로 수행하는 것이 적합
  - 시뮬레이션 : 주요 변경이 이뤄지는 시점과 반기 정도가 적합
  - 최적화 : 1 년에 1 번 정도가 적합
- 리모델링시 수행하는 업무
  - 데이터 마이닝 : 동일 데이터를 이용해 다시 학습하는 방법, 변수 추가로 학습하는 방법
  - 시뮬레이션 : 이벤트 발생 패턴 변화, 시간지연(delay) 변화, 이벤트 처리하는 리소스 증가, Queuing Priority, Resource Allocation Rule 변화 등 처리
  - 최적화 : Object Fuction 의 계수 변경, Constraint 에 사용하는 제약값 변화와 추가

수행준거	고려사항
- 분기·반기·연 단위로 정기적인 분석 모형 재평가 실시, 성능 편차 발생을 분석·식별 할 수 있어야 함 - 업무 IT 환경에 주요 변화 발생 시, 분석 모형 재평가 실시하고 성능 편차 발생을 분석·식별 할 수 있어야 함 - 정기·비정기 분석 모형 재평가 결과에 기반해 모형 조정 및 개선 작업 수행, 분석모형 전면 재구축 위한 독립 프로젝트 계획 수립해 추진 가능	- 데이터마이닝, 최적화 모델링 결과를 정기적으로(분기,반기,연) 재평가해 결과에 따라 필요시 분석 모형 재조정 - 데이터 마이닝은 최신데이터 적용이나 변수 추가 방식으로 분석모형 재조정 가능 - 시뮬레이션은 업무 프로세스 KPI의 변경, 주요 시스템 원칙 변경, 발생 이벤트 건수 증가에 따라 성능 평가 및 필요시 재조정 - 최적화는 조건 변화나 가중치 변화시 계수 값 조정 또는 제약 조건 추가로 재조정 가능 - 업무특성에 따라 차이가 있으나 일반적으로 초기에는 모형 재조정을 자주 수행, 점진적으로 그 주기 길게 설정 가능 - 관리 대상 모델이 월 20개 이상이거나 기타 업무와 병행해서 수행해야하는 경우 도구를 통한 업무 자동화 권고

*주요 역량 소개(286~287p)

제2절 데이터 분석 기법의 이해

1. 개요

- 데이터 분석에 대한 정의는 매우 다양하고 수준과 복잡성, 목적도 다름
- 분석은 일반적으로 조회와 고급분석으로 양분되며 고급분석은 20개 이상의 변수와 수천 건 이상의 데이터를 이용해 인사이트를 얻거나 의사결정을 하는데 직접 사용됨

2. 기초 지식과 소양

- 평균과 분산에 대한 이해를 토대로 집단 간 평균과 분산의 차이, 상관관계, 독립/종속 변수를 이용한 회귀분석 이해, R square와 p값에 대한 이해, 클러스터링(clustering)
- 진정 필요한 추가 지식은 다양한 산업에 대한 이해
  - 상식수준에서 벗어난 해당 업계 신입사원 수준의 산업 분야 이해가 필요
- 평상시 관심을 갖고 업무와 관련지어 조금씩 늘 학습할 것 추천

3. 데이터 처리

- 분석을 위해 분석방법에 맞게 데이터를 수집·변형하는 과정이 필요하고 때론 잘 정리된 데이터 마트(data mart)가 필요
- 신규 시스템이나 DW(data warehouse)에 포함되지 못한 자료가 있으면 기존 운영 시스템(Legacy)에서 직접 가져오거나 ODS(Operational Data Sore)에서 운영 시스템과 거의 유사한 정제된 데이터를 가져와 DW에서 가져온 내용과 결합(데이터 수집과정)
- 문서를 받고 데이터를 처리하는 과정에 분석자가 충분히 관여해야 함
- 데이터 입수과정이 완료되면 최종 데이터 구조로 가공하는 과정을 거치며 이는 분석기법에 의존
  - 원하는 데이터 형태로 가공하는 과정은 분석결과의 품질과 성능에 크게 영향을 미쳐 분석가가 많은 노력을 해야하는 단계
- 비정형 데이터는 적합한 DBMS에 저장됐다가 텍스트 마이닝을 거쳐 데이터 마트와 통합
- 관계형 데이터는 DBMS에 저장돼 사회 신경망 분석을 거쳐 분석결과 통계값이 마트와 통합돼 다른 분석기법과 연계·활용됨
- 데이터 처리과정에 가장 좋은 방법은 원시모형(prototype)을 만드는 것이며 분석의 질이 중요해도 데이터 처리를 제대로 거치지 않으면 분석 자체의 의미가 사라짐
  - 데이터 처리와 분석은 트레이드오프 관계
- 데이터를 분석도구보다 DBMS에서 처리하면 풍부한 기능을 효율적으로 활용 가능
  - 일반적으로 DBMS에 메모리와 CPU를 더 많이 할당한 경우가 많으므로 DBMS에서 1차처리해 분석도구로 가져오는 것이 현실적
- 테이터 처리에서 성능 튜닝은 주기적이어야지 매번 시도하면 끝이 없음(분석이 우선)

4. 시각화

- 가장 낮은 수준의 분석이지만 잘 사용하면 복잡한 분석보다도 더 효율적
  - 대용량 데이터를 다루는 빅데이터 분석에서는 시각화의 활용률이 높음
  - 탐색적 분석을 할 때 시각화는 거의 필수적

5. 공간분석

- 낮은 수준의 분석
- 공간적 차원과 관련된 속성들을 시각화해 추가한 것
  - 지도 위에 관련 속성 생성, 크기, 모양, 선 굵기 등으로 구분하면 노출도 향상되고 이를 통해 인사이트 얻을 수 있음

6. 탐색적 분석

- 하나하나 탐색하면서 분석하는 방식
  - 다양한 차원과 값을 조합해가며 특이한 점이나 의미있는 사실을 도출하고 분석의 최종 목적을 달성해가는 과정
  - 매우 많은 시간과 자원이 필요하고 해결하려는 분야에 대한 지식, 사실을 확보하는 단계
- 일정 가설과 시나리오를 갖고 제한적인 범위에서 주어진 목적을 달성 가능하도록 조절
- 효율적인 탐색적 분석을 위해 의미 있을 것 같은 변수집단과 아닌 집단을 1차구분하고 그래도 변수가 많으면 우선순위 2단계로 의미 있을 1차 집단 우선수행, 의미 없을 것 같은 변수5개정도 선별해 확인

- 선별된 차원과 값들에 대해 자동으로 탐색적 분석 결과를 테이블과 그래프로 산출하는 스크립트를 실행해 의미 있는 내용을 걸러줄 수 있는 자동화 방법 필요
- 탐색적 분석 마친 후 분석 시나리오 만들어야 함
  • 무슨 기법으로 어떤 것을 분석해 목적을 달성할 것인지 결정하는 단계
  • 상세하고 현실적인 WBS(Work Breakdown Structure) 나옴

7. 통계분석

- 샘플이 충분히 크기 때문에 빅데이터 분석하는데 모집단과 샘플은 고려대상 아님
  • 모집단의 속성과 샘플링 결과가 일치하는지 주요변수에 대해 반드시 확인

8. 데이터 마이닝

- 대표적 고급분석. 데이터에 있는 패턴을 파악해 예측하는 분석
  • 상황분류, 집단간 차이를 갖고 클러스터링해 구분, 이전 값들의 패턴으로 미래 값 예측, 입력변수와 종속변수 관계이용해 미래 값 예측, 동시 발생 이벤트와 시차 갖고 발생하는 이벤트 이용해 어떤 이벤트 발생할지 파악 등
- 데이터가 크고 정보가 다양할수록 보다 활용하기 유리한 최신 기법

9. 시뮬레이션

- 복잡한 실제상황을 단순화해 컴퓨터상의 모델로 만들어 재현하거나 변경함으로써 현상을 보다 잘 이해하고 미래의 변화에 따른 결과를 예측하는데 사용하는 고급분석 기법
- 시뮬레이션 기법과 최적화 기법이 결합되면서 규칙이나 조건을 정교화해 효과 높임

10. 최적화

- 오랜 역사 가진 고급 분석기법으로 목적함수 값의 최대화/최소화를 목표로 함
  • 제약조건 하에서 목표값을 개선하는 방식으로 목적함수와 제약조건을 정의해 문제 해결

11. 배포 및 운영

- 사용자가 개발된 데이터와 모델을 이용해 활용하는 환경도 구축해야 함
  • 분석 및 마이닝 모델에 직접 접근해 데이터를 조회하고 마이닝 결과를 적용해 결과를 조회할 수 있는 인터랙티브한 환경개발(RStudio Shiny)

제3절 분석환경 이해와 기본 사용법

- 데이터 분석은 SQL 수준의 교육과 달리 분석도구가 다양하고 표준이 없음

	SAS	SPSS	R
프로그램 비용	유료, 고가	유료, 고가	오픈소스
설치 용량	대용량	대용량	모듈화로 간단
다양한 모듈지원 및 비용	별도 구매	별도 구매	오픈소스
최근 알고리즘 및 기술반영	느림	다소 느림	매우 빠름
학습자료 입수의 편의성	유료 도서 위주	유료 도서 위주	공대 논문 및 자료 많음
질의를 위한 공개 커뮤니티	NA	NA	매우 활발

1. 분석환경의 이해

가. 통계패키지 R

- 오픈소스 프로그램으로 통계, 데이터 마이닝과 그래프를 위한 언어
- 다양한 최신 통계분석과 마이닝 기능을 제공
- 전 세계적으로 사용자들이 다양한 예제를 공유
- R의 특징
  • 다양한 최신 통계 분석 및 마이닝 기능을 R 플랫폼에서 제공
  • 다양한 최신 알고리즘을 제공해 다양한 시도 가능
  • 기능들의 자동화가 비교적 쉬움
  • 사용자들이 여러 예시를 공유

나. R 스튜디오

- 오픈소스이고 다양한 운영체제를 지원
- R스튜디오는 메모리에 변수가 어떻게 되어 있는지 타입과 무엇인지를 볼 수 있고, 스크립트 관리와 도큐먼테이션이 편하다.
- 메모리에 변수가 어떻게 돼있는지 타입이 무엇인지 볼 수 있고 스크립트 관리와 문서화가 편해 R studio 사용
- 래틀과 R의 장단점

	래틀(Rattle)	**R**
장점	처음 접근하고 데이터를 다루는 것이 쉬움	유연하고 자주 업그레이드됨
단점	패키지의 정해진 기능 사용, 업그레이드가 제대로 안되면 통합성에서 문제 발생	코딩을 해야함

다. 데이터 소스 및 분석 IT 아키텍처

- 작업환경은 업무 규모와 본인에게 익숙한 환경이 무엇인지를 기준으로 선택
- 기업환경에서는 64bit 환경의 듀얼코어, 32GB RAM, 2TB 디스크, 리눅스 운영체제를 추천

• 가용 물리적 메모리 크기는 x86 64비트 시스템에서는 128TB가 한계
• 64비트 윈도우 운영체제에서는 8TB의 메모리까지 지원

2. 기본사용법

가. R 언어와 문법

1) R 스튜디오 설명

- 스크립트 : R 명령어 입력하는 창. 명령어를 실행할 때는 실행하려는 문장에 커서를 두고 Ctrl+Enter↵
- 콘솔 : 스크립트 창에서 실행한 명령문이 실행되는 것을 볼 수 있는 곳. 오류 있으면 에러 메시지 뜸. 명령어 직접 입력하면 저장 안돼 재실행 불가
- 워크스페이스 : 할당된 변수와 데이터 나타남
- Search Results : 설치된 패키지와 help 등 볼 수 있음

2) 변수와 벡터 생성

- 변수명 <- or = 임의값
- 생성확인 : 변수명 입력 또는 프린트 명령어 이용

① 프린트

• print(변수명) : 변수의 값을 출력

② c() : 벡터 생성

• c(값1, 값2, …,) : 하나의 변수에 여러 값을 할당
  · 문자형 값은 ""써줘야 문자형 값으로 인식
  · 함수 내 연산 가능
  · 논리형 값(TRUE,FALSE)은 ""필요 없음
  · 변수 결합 가능

3) 수열

• n:m :n, n±1, n±2,…,m
• seq(from=시작점, to=끝점, by=간격) : 일정한 간격으로 숫자를 나열
• seq(from=시작점, to=끝점, length=값의 길이) : 값의 길이만큼 숫자 나열
• rep(반복할 내용, 반복수) : 같은 값의 단순 반복
  · rep(1,times=5) : 벡터 값 5번 반복(1 1 1 1 1)
  · rep(1:2,each=2) : 각각 2번씩 반복(1 1 2 2)

4) 데이터 유형과 객체

① Numeric : 숫자형. integer(정수), double(소수점포함)
② Character : 문자형. ""표시. ex) "a", "abc"
③ Paste("붙일 내용","붙일 내용",sep="") : 데이터 결합
• 기본적으로 공간 삽입됨. 자동으로 변수명 만들 때 유용(paste(A,10,sep="")
  결과> 붙일내용 sep 붙일내용
④ Substr(문자열, 시작, 끝) : 시작과 끝에 해당하는 하위 문자열 추출
⑤ 논리값 : True(T), False(F)
⑥ Matrix : 벡터에 차원정해주면 행렬로 변환 가능
• matrix(이름, 행 수, 열 수) ex) matrix(theData,4,5), matrix(1:20,4,5)
  · row 먼저 채우고 column에 값 들어감
• dim(행렬) : 행렬의 행과 열 수를 반환
• diag(행렬) : 행렬의 대각선에 있는 값을 반환
• t(행렬) : 전치행렬
  · 행렬 곱 : 행렬%*%행렬
• colnames(행렬) : 열 네임을 조회
  · 열 네임붙이기 : colnames(mat)<-c("IBM","MSFT","GOOG")
• rownames(행렬) : 행 네임을 조회
  · 행 네임붙이기 : rownames(mat)<-c("IBM","MSFT")
• 행렬조회 : 행렬이름 입력
  · 행 조회 : 행렬이름[#,]
  · 열 조회 : 행렬이름[,#]

⑦ list(, , , ...)

- list 만들 땐 list 원소들에 태그 부여해야 함. 서로 다른 데이터 오브젝트 결합 가능
  · 각 요소 확인 : 리스트이름$요소(or 리스트이름[[#]])
  unlist() : 리스트형식의 데이터를 벡터로 변환

* 벡터와 리스트의 차이
: 벡터에서 모든 원소는 같은 모드, 리스트는 원소들이 다른 모드여도 OK

⑧ 데이터 프레임 : 관찰된 결과(observation)로 된 테이블. 행렬×

- 가장 자주 사용되고 편리한 데이터 처리방식. 모든 측면에서 매우 직관적
• data.frame(, , ,...) : 여러 열로 정리된 데이터를 데이터프레임으로 조립
• rbind(dfrm1,dfrm2) : 두 데이터프레임의 행 쌓기
• cbind(dfrm1,dfrm2) : 두 데이터프레임의 열 이어붙이기
• subset(dataframe, select=열 이름) : 데이터세트에서 조건에 맞는 내용 조회
  (리스트와 벡터에서도 선택 가능)
• with(dataframe, 열 이름)
• merge(df1, df2, by="df1과 df2의 공통된 열 이름")
  : 행 정렬 안 되거나 동일한 순서로 안 나타나도 상관없음
• grep(조회할 문자패턴, data)

⑨ 벡터에 있는 원소 선택

- 벡터 내 값 조회에 유용. 대괄호와 간단한 인덱스 사용
- 벡터 값에서 특정 값 가져오기 : [], c() 함수 이용

- []안에 조건문 넣고 조건 만족하는 값 가져올 수 있음. &,+
- 이름으로 원소 선택 가능

⑩ 자료형 데이터 구조 변환 : 데이터 구조를 다른 구조로 바꾸고 싶을 때
변환 적용 안 되면 NA값 나타남

- as.data.frame(x) : 데이터프레임 형식으로 변환
- as.list(x) : 리스트 형식으로 변환
- as.matrix(x) : 행렬 형식으로 변환
- as.vector(x) : 벡터 형식으로 변환
- as.factor(x) : 팩터(factor) 형식으로 변환
- as.numeric(논리값) : FALSE=0, TRUE=1
- as.character(숫자) : 숫자를 문자로

⑪ 문자열을 날짜로 변환

- 2013-08-13처럼 문자열 표현으로 된 날짜를 Date 객체로 변환
- Sys.Date() : 현재 날짜를 반환
- as.Date() : 날짜 객체로 변환(yyyy-mm-dd)
- format : 날짜 스타일 변환("%m/%d/%Y"→08/13/2013")

⑫ 날짜를 문자열로 변환

- format(날짜, 포맷)
- as.character(Sys.Date())
- format(Sys.Date(),'%a') : 요일 조회 → 월
- format(Sys.Date(),'%b') : 축약 월 이름 조회 → 3
- format(Sys.Date(),'%B') : 전체 월 이름 조회 → 3월
- format(Sys.Date(),'%d') : 두자리 숫자로 된 일 조회 → 17
- format(Sys.Date(),'%m') : 두자리 숫자로 된 월 조회 → 03
- format(Sys.Date(),'%y') : 두자리 숫자로 된 연도 조회 → 14
- format(Sys.Date(),'%Y') : 네자리 숫자로 된 연도 조회 → 2014

⑬ Missing : missing 데이터(/0, infinite)

⑭ 벡터의 기본 연산

- mean() : 평균
- sum() : 합
- average() : 평균
- median() : 중앙값
- log() : 로그
- sd() : 표준편차
- var() : 분산
- cov() : 공분산
- cor() : 상관계수(–1≦rxy≦1)
- length() : 변수의 길이 값 반환
- sapply(c,log) : 벡터 c에 log 적용

5) 알아두면 유용한 기타 함수들

- write.csv(변수 이름,"지정할 파일이름.csv") : 변수를 csv 파일로 저장
- read.csv("저장된 파일이름.csv") : csv 파일을 R로 읽음
- save(변수이름, file="지정할 데이터 파일이름.Rdata") : R데이터 파일로 저장
- load("저장된 파일이름.Rdata") : R 데이터를 읽어 들이는 방법
- rm() : 데이터 삭제, 선택 변수만 삭제
- rm(list=ls(all=TRUE)) : 모든 변수를 삭제
- data()
- summary() : 한 번에 간단한 통계량들을 데이터세트 열마다 요약
- head() : 데이터세트의 6번째 행까지 조회
- install.packages("패키지 이름") : R 패키지 설치
- library(패키지 이름) : R에 패키지를 불러오는 함수
- vignette("알고 싶은 package 이름") : 정보 간단 요약본
- q() : 작업종료. 종료시 작업환경 변수 저장여부 물음
- setwd("~/") : R 데이터와 파일등을 로드하거나 저장할 때 워킹 디렉터리 지정
- ? 명령어 : 도움말 로드
- ?? 명령어 : 명령어 검색

3. 래틀

- R을 GUI환경에서 편리하게 사용할 수 있게 도와주는 패키지

가. 래틀설치

나. 래틀 기본사용법

다. 래틀에서 데이터 불러오기

1) 라이브러리의 데이터 불러오기

2) csv 파일 불러오기

① data Explore

② 테스트

③ 데이터 변환

④ 클러스터

⑤ 모드

제2장 통계분석

제1절 통계분석의 이해

1. 통계

- 특정집단을 대상으로 수행한 조사나 실험을 통해 나온 결과에 대한 요약된 형태의 표현

- 표본조사 : 대상 집단의 일부를 추출해 어떤 현상을 관측/조사해 자료 수집하는 방법
  • 표본추출방법
  ・단순랜덤추출법 : n개의 번호를 임의로 선택해 해당 원소를 표본으로 추출
  ・계통추출법 : N개 원소로 구성된 모집단에서 k개씩 n개구간 나누고 첫 구간에서 하나 임의 선택 후 k개씩 띄어 표본 추출
  ・집락추출법 : 모집단이 집락(cluster)의 결합으로 구성돼있는 경우 일부 집락을 랜덤으로 선택하고 선택된 각 집락에서 표본 임의 선택
  ・층화추출법 : 각 계층 고루 대표할 수 있게 표본 추출. 이질적 모집단 원소를 유사한 것끼리 몇 개의 층(stratum)으로 나눈 후 각 층에서 랜덤하게 표본 추출
- 실험 : 특정 목적 하에서 실험 대상에게 처리한 후 그 결과 관측해 자료 수집
- 측정(measurement) : 표본조사나 실험을 실시하는 과정에서 추출된 원소/실험 단위로부터 주어진 목적에 적합하도록 관측해 자료를 얻는 것
  • 명목척도(nominal scale) : 측정대상이 어느 집단에 속하는지 분류할 때 사용
  • 순서척도(ordinal scale) : 측정대상 특성의 서열관계를 관측하는 척도
  • 구간척도(interval scale) : 측정대상이 갖은 속성의 양 측정. 관측값 사이 비율 의미×
  • 비율척도(ratio scale) : 절대적 기준값(0) 존재, 사칙연산 가능. 가장 많은 정보 갖는 척도
  ・질적자료/이산형자료 : 명목척도, 순서척도
  ・양적자료/연속형자료 : 구간척도, 비율척도

2. 통계분석(statistical analysis)

- 특정한 집단이나 불확실한 현상을 대상으로 자료를 수집해 대상 집단에 대한 정보를 구하고 적절한 통계분석방법을 이용해 의사결정을 하는 과정(통계적 추론)
  • 대상 집단에 대한 정보 : 자료를 요약·정리한 결과, 숫자/그림으로 정리된 각종 통계
  • 통계적 추론 : 수집된 자료를 이용해 대상 집단(모집단)에 대해 의사결정을 하는 것
  ・추정(estimation), 가설검정(hypothesis test), 예측(forecasting)
  • 기술통계(descriptive statistic) : 수집된 자료를 정리·요약하기 위해 사용되는 기초통계
  ・자체로도 여러 용도에 쓰이나 대게 자세한 통계적 분석을 위한 전단계 역할

3. 확률 및 확률분포

- 확률 : 특정사건이 일어날 가능성의 척도
  • 표본공간(sample space,Ω) : 나타날 수 있는 모든 결과들의 집합
  • 원소(element) : 나타날 수 있는 개개의 결과
  • 사건(event) : 표본공간의 부분집합
- 확률변수(random variable) : 특정값이 나타날 가능성이 확률적으로 주어지는 변수
  (정의역이 표본공간, 치역이 실수값인 함수)
  • 이산형 확률변수(discrete r.v.) : 0이 아닌 확률 값을 갖는 셀 수 있는 실수값
  • 연속형 확률변수(continuous r.v.) : 특정 실수구간에서 0이 아닌 확률을 갖는 확률변수
  • 결합확률분포(joint probability distribution) : 두 확률변수의 결합확률분포
- 통계분석에서 수집된 자료에서 어떤 정보를 얻고자 할 때는 항상 수집된 자료가 특정 확률분포를 따른다고 가정
  • 이산형 : 베르누이, 이항분포, 기하분포, 다항분포, 포아송분포 등
  • 연속형 : 균일분포, 정규분포, 지수분포, t분포, 분포, F분포 등

4. 추정과 가설검정

- 각 확률분포는 평균, 분산 등의 모수(parameter)를 갖음
- 확률표본(random sample) : 특정 확률분포로부터 독립적으로 반복해 표본을 추출하는 것
  각 관찰값들은 서로 독립적이며 동일한 분포
  • 모수 : 모집단의 특성을 나타내는 값(일반적으로 알려져 있지 않음)
  표본추출에 의해 모수 추정
  ・점추정(point estimation) : 모수가 특정한 값. 얼마나 추정이 정확한지 판단 불가
  ex) 표본평균, 표본분산
  ・구간추정(interval estimation) : 확률로 표현된 믿음의 정도 하에서 모수가 특정 구간에 있을 것. 분포에 대한 전제 필요. 구해진 구간(신뢰구간) 안에 모수가 있을 가능성의 크기(신뢰수준) 주어져야함.
- 가설검정 : 모집단에 대한 어떤 가설을 설정한 후 표본관찰을 통해 가설의 채택여부 결정
  • 검정하고자 하는 모집단의 모수에 대한 가설 설정이 가장 기본적
  ・귀무가설(H0) : 모수에 대한 가설 중 간단하고 구체적인 표현 설정
  ・대립가설(H1)
  • 검정통계량(test statistic) : 검정에 사용되는 통계량
  • 유의수준(significance level) : H0이 옳은데 이를 기각하는 확률의 크기
  • 기각역(critical region) : H0이 옳다는 전제에서 구한 검정통계량의 분포에서 확률이 유의수준인 부분
- 오류(error)
  • Type Ⅰ : H0가 맞는데 기각하는 오류
  • Type Ⅱ : H0가 틀린데 채택하는 오류
  ・상충관계, 일반적으로 1종오류(α) 크기 고정시키고 2종오류(β) 최소화되게 기각역 설정

5. 비모수 검정

- 모수적 검정방법 : 검정하고자 하는 모집단의 분포에 대한 가정 하에서 검정통계량과 그 분포를 유도해 검정 실시
- 비모수적 검정 : 자료가 추출된 모집단의 분포에 아무 제약 않고 검정 실시
- 모수 & 비모수 차이점

	모수	비모수
가설설정	가정된 분포의 모수	분포의 형태

검정	관측된 절대적 크기 자료 이용	관측값의 순위나 차이의 부호 등 이용

제2절 기초통계분석

1. 기술통계(Descriptive Statistics)

- 자료를 요약하는 기초적 통계
- 데이터 분석에 앞서 데이터의 대략적인 통계적 수치를 계산해봄으로써 데이터에 대한 대략적 이해와 분석에 대한 통찰력을 얻기에 유리
- 데이터 마이닝에 앞서 데이터의 기술통계를 확인해보는 것이 좋음
  - head : 데이터를 기본 6줄 보여줘 데이터가 제대로 import됐는지 살펴볼 수 있는 함수
  - summary : 데이터의 컬럼에 대한 전반적인 기초 통계량 보여줌
  - 데이터의 특정 컬럼 선택 : 데이터네임$column명

2. 인과관계의 이해

- 용어
  - 종속변수(반응변수,y) : 다른 변수의 영향을 받는 변수
  - 독립변수(설명변수,x) : 영향을 주는 변수
  - 산점도(scatter plot) : 좌표평면 위에 점들로 표현
    · 두변수 사이의 선형관계, 함수관계, 이상값 존재, 몇 개의 집단으로 구분 되는가 확인
- 공분산(covariance) : 두 확률변수 X, Y의 방향의 조합(선형성)
  - $v(X,Y)=E[(X-\mu\ )(Y-\mu_Y)]$
  - X, Y가 독립이면 Cov(X,Y)=0

$$Cov(X,Y)=\sigma_{XY}=E[XY]-E[X]E[Y]$$

3. 상관분석(Correlation Analysis)

- 데이터 안의 두 변수 간의 관계를 알아보기 위함
- 두 변수의 상관관계를 알기위해 상관계수(correlation coefficient) 이용
  - 피어슨 상관계수 : 등간척도 이상으로 측정되는 두 변수의 상관관계 측정
  - 스피어만 상관계수 : 서열척도인 두 변수의 상관관계 측정
  - 1(-1)에 가까울수록 강한 양(음)의상관관계를 나타내고 상관관계 없으면 r=0

가. 피어슨의 표본상관계수

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{=1}^{n}(x_i-x)^2\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}},\ -1\le r\le 1$$

나. 스피어만 상관계수

$$\theta=\frac{\sum(r_i-\bar{r})(s_i-\bar{s})}{\sum(r_i-r)^2\sqrt{\sum(s_i-\bar{s})^2}},\ -1\le\theta\le 1$$

4. 회귀분석

가. 단순회귀분석과 중회귀분석(다중회귀분석)의 개념

- 회귀분석 : 하나나 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정하는 통계기법
- 단순선형회귀 : $y_i=\beta_0+\beta_1x_i+\epsilon_i$
- 다중회귀(중회귀)분석 : $y_i=\beta_0+\beta_1u_i+\beta_2v_i+\beta_3\omega_i+\epsilon_i$
- 찾은 선이 적절한지 확인
  - 모형이 통계적으로 유의미한가? F통계량(p값) 확인
  - 회귀계수들이 유의미한가? 계수의 t값, p값 또는 신뢰구간 확인
  - 모형이 얼마나 설명력을 갖나? 결정계수(R-square) 확인
  - 모형이 데이터를 잘 적합하고 있나? 잔차 그래프 그리고 회귀진단
  - 데이터가 전제하는 가정을 만족시키나?
    · 가정
    선형성, 독립성(잔차와 독립변인 값 독립), 등분산성(오차 분산 일정),
    비상관성(잔차끼리 상관×), 정상성(잔차가 정규분포)

나. 회귀분석의 종류

종류	모형	
단순회귀	$\beta\ +\beta_1X+\epsilon$	설명변수가 1개이며, 반응변수와의 관계가 직선
다중회귀	$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\epsilon$	설명변수 k개이며, 반응변수와의 관계가 선형
다항회귀	$k=2$이고 2차 함수인 경우, $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_{11}X_1^2+\beta_{22}X_2^2+\beta_{12}X_1X_2+\epsilon$	설명변수 k개이며, 반응변수와의 관계가 1차 함수 이상 (단, $k=1$이면 2차 함수 이상)
곡선회귀	2차 곡선인 경우, $Y=\beta_0+\beta_1X+\beta_2X^2+\epsilon$ 3차 곡선인 경우, $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\epsilon$	설명변수가 1개이며 반응변수와의 관계가 곡선
비선형회귀	$Y=\alpha e^{-\beta}\ +\epsilon$	회귀식의 모양이 미지의 모수 들의 선형관계로 이루어져 있지 않은 모형

다. 최적회귀방정식의 선택: 설명변수의 선택

- 회귀모형 설정 변수 선택 원칙
  - y에 영향 미칠 수 있는 모든 설명변수 x들을 y값 예측에 참여시킨다.
  - 설명변수 x가 많아지면 관리하는데 노력이 많이 요구되므로 가능한 범위 내에서 적은 수의 설명변수를 포함시켜야 한다.

1) 선택방법

① 모든 가능한 조합의 회귀분석

- 가능한 모든 독립변수 조합에 대한 회귀모형 분석해 가장 적합한 회귀모형 선택

② 단계적 변수선택

- 전진선택법 : 상수모형부터 시작해 중요하다고 생각되는 설명변수부터 차례로 추가
- 후진제거법 : 독립변수 후보 모두 포함한 모형에서 시작해 가장 적은 영향주는 변수부터 제거하면서 더 이상 제거할 변수 없을 때 모형 선택
- 단계별방법 : 전진선택법에 의해 변수 추가하면서 새롭게 추가된 변수에 기인해 기존 변수가 그 중요도가 약화되면 그 변수 제거하는 등 단계별로 추가/제거되는 변수 여부 검토해 더 이상 없을 때 중단

* step(lm(종속변수~설명변수, 데이터세트), scope=list(lower=~1,upper=~설명변수),direction="변수선택방법") 함수로 변수 쉽게 선택 가능
* R에서 구체적 디렉터리 설정해 외부 데이터세트 읽을 때 ₩를 2번 해줘야함(C:₩₩~)

제3절 시계열 분석

1. 정상성(stationarity)

- 시계열 자료 : 시간의 흐름에 따라 관찰된 값들
  - 비정상성 시계열 : 시계열 분석하는데 다루기 어려운 시계열 자료. 정상시계열로 만들어 분석
  - 정상성 시계열 :
    - 약한의미의 정상성 : 모든 시점에 평균일정, 시점과 분산 독립, 공분산은 시차에만 의존
- 비정상→정상 : 변환(transformation), 차분(difference)
  - 변환 : 분산이 일정하지 않은 비정상 시계열
  - 차분(t1-t0) : 평균이 일정하지 않은 비정상 시계열

2. 시계열 모형

가. 자기회귀모형(AR모형)

- p시점 전의 자료가 현재 자료에 영향을 주는 자귀회귀모형을 AR(p) 모형이라 함

$$\phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t$$

a_t : white noise process(백색잡음과정)

- 자귀회귀모형 판단 조건
  - 자기상관함수(ACF) 빠르게 감소하고 부분자기상관함수(PACF)는 어느 시점에 절단점 갖음

나. 이동평균모형(MA모형)

- 유한한 개수의 백색잡음의 결합. 항상 정상성 만족

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_p a_{t-p}$$

- ACF에서 절단점 갖고 PACF가 빠르게 감소

다. 자기회귀누적이동평균모형(ARIMA(p,d,q)모형)

- 가장 일반적인 모형으로 비정상시계열 모형으로 차분이나 변환을 통해 AR/MA/ARMA로 정상화 가능
- p는 AR, q는 MA와 관련있는 차수로 ARIMA에서 ARMA로 정상화할 때 차분한 횟수 의미

라. 분해 시계열

- 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법
  - 회귀분석적 방법 주로 사용
- 시계열 구성 요소

1) 추세요인(trend factor) : 자료가 어떤 특정한 형태를 취할 때
2) 계절요인(seasonal factor) : 고정된 주기에 따라 자료가 변화
3) 순환요인(cyclical factor) : 알려지지 않은 주기를 갖고 자료가 변화
4) 불규칙요인(irregular factor) : 회귀분석에서 오차에 해당하는 요인

- 분해시계열분석법에서는 각 구성요인을 정확히 분리하는 것이 중요
  - 요인 정확히 분리하기 쉽지 않으며 이론적 약점 존재 but 많이 사용됨

$$Z_t = f(T_t, S_t, C_t, I_t)$$

마. 시계열 실습

1) 시계열 자료읽기

- 시계열 분석 패키지 : TTR, forecast
- ts : 데이터를 시계열 형식으로 변환

2) 그래프

- plot.ts
- log변환하면 안정적인 패턴의 그래프 얻을 수 있음

3) 분해시계열

- TTR 패키지 SMA 함수 이용해 시계열 트렌드 보여주는 분해시계열의 MA그래프 그리기

① decompose non-seasonal data

- 비계절성 띄는 시계열 자료 : 트렌드 요소, 불규칙 요소로 구성

② decompose seasonal data

- 계절성 띄는 시계열 자료 : 경향성 요소, 계절성 요소, 불규칙 요소로 구성

③ seasonally adjusting
• 계절성 띄는 자료는 계절성 요소를 추정해 그 값을 raw에서 빼면 적절히 adjust 가능

4) ARIMA 모델

① 차분
- ARIMA는 정상시계열에 한해 사용
• 비정상시계열자료는 차분해 정상성 만족시키는 시계열로 변환

5) 적합한 ARIMA 모델 결정
- ACF와 PACF를 통해 적합한 ARIMA 모델 결정 가능
- 모델링의 기본은 모수(parameter)들이 적을수록 단순하고 이해쉬움
• forecast package에 내장된 auto.arima()함수 사용하면 적절한 모형 찾을 수 있음
- 미세한 차이로 약간의 오차발생해도 모수 적게 사용한 모델링 보정(fitting)이 적절하다면 간단하고 이해하기 쉬운 모델 선택이 관례

6) ARIMA 모델을 이용한 예측
- 주어진 시계열에 ARIMA 모델 채택했다면 ARIMA 모델의 모수로 미래 값 예측가능
- fitting(보정) 후 forecast package의 forecast.Arima()함수 이용해 미래값 예측 가능

제4절 다차원척도법(Multidimensional Scaling, MDS)
- 여러 대상 간의 관계에 대한 수치적 지료를 이용해 유사성에 대한 측정치를 상대적 거리로 시각화 하는 방법
• cmdscale()

제5절 주성분분석(Principal Component Analysis, PCA)
- 상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수로 선형결합 해 변수 축약하는데 사용
• 데이터 내부 구조를 파악할 수 있는 방법으로 예측모델 만들 때 주로 사용
• 보통 3개 이내의 변수로 축약하고 이로 인한 정보손실은 20% 정도로 함

제3장 데이터 마트

제1절 데이터 변경 및 요약

1. R reshape를 활용한 데이터 마트 개발
- 고객 데이터 마트 생성하는 일은 CRM(Customer Relationship Management) 관련 업무 중 핵심
- 마트 만드는 일 접근법
• 작게 시작해 크게 만들어 나가는 노력 필요
・요약변수→파생변수→모델링
• 빠르고 간편한(quick and dirty) 방법
・미리 검증해보고 변수를 더 만들어 나가는 것이 효율적
- 마트는 담당자의 역량에 따라 수준차이가 큼

가. 요약변수(summary variables)
- 가장 기본적인 변수로 고객·상품·채널을 종합(aggregation)한 변수
• 단순한 구조이므로 자동화하기 쉬워 조금만 고민하면 상황에 맞게 또는 일반적인 자동화 프로그램 만들 수 있음
• 요약변수만으로도 세분훈화하거나 행동 예측을 하는데 큰 도움받을 수 있으나 기준값(threshold value)의 의미해석이 애매할 수 있음
• 연속형 변수를 자동으로 타깃에 맞춰 그루핑해주면 좋음
- 많은 모델에서 공통적으로 많이 사용될 수 있으며 재활용성도 높음
- 다양한 모델을 개발해야 하는 경우 효율적으로 사용 가능
- 마트를 만드는데 시간과 공간의 제약이 덜한 상황이라면 다양한 조합의 요약변수를 자동으로 만드는 것이 적합
• 너무 오랜 기간을 포함한 요약변수는 별 의미 없음
- SQL 튜닝이 잘못됐다면 전문가에게 요청하는 것이 나음
• 수천만 건을 30분내에 처리 못하면 DBA나 시스템 관리자의 잘못 확률 높음
• SAS로 처리하는게 훨씬 빠르나 DB에서 데이터 가져올 때 고생
• 결측값과 이상값은 있는 그대로 놔두는 게 더 효율적이고 효과적 그룹핑에서 자동처리 되도록 함

나. 파생변수(derived variables)
- 특정한 의미를 갖는 작위적 정의에 의한 변수
• 사용자가 특정조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수
• 매우 주관적인 변수일수 있으므로 논리적 타당성을 갖춰야 함
- 파생변수 자체로도 분석가능하나 이를 이용하면 데이터마이닝에 기여하는 바가 큼
- 상황에 따라 특정 상황에만 유의미하지 않게 대표성을 나타나게(robust)할 필요
- 보다 많은 변수를 잘 활용하는 것이 요즘 대세

다. reshape
- 기존 거래 데이터(TR)구조를 column-wise하게 전환하는데 크게 melt와 cast단계로 구분
• melt : 기준되는 변수를 제외한 여러 Factor변수를 하나의 Dimension 변수와 하나의 Factor 변수로 변환하는 함수
• cast : 엑셀 피벗팅 하듯 자료 변형

2. sqldf를 이용한 데이터 분석
- 표준 SQL에서 사용되는 문장이 모두 가능하고 데이터 이름에 "."같은 특수문자가 들어간 경우 ''로 묶어주면 테이블처럼 간단히 처리 가능

3. plyr

- 데이터를 분리하고 처리한 다음 다시 결합하는 등 가장 필수적인 데이터 처리기능 제공
  - apply 함수와 multi-core 사용 함수 이용하면 for loop사용하지 않고 간단하고 빠르게 처리 가능
- apply함수에 기반해 데이터와 출력변수를 동시에 배열로 치환하여 처리하는 패키지
  - set.seed : R에서 난수생성시 일정하게 고정시켜주는 역할(생성시 같은 값 난수 생성)
  - #runif(생성할 난수의 개수, 최소값, 최대값)
  - ddply : plyr을 올리고 dataframe에서 dataframe으로 입출력 하는 함수

4. 데이터 테이블

- 데이터프레임과 유사하지만 보다 빠른 그루핑과 ordering, 짧은 문장 지원 측면에서 데이터프레임보다 매력적
  - 무조건 빠른 것이 아니므로 특성에 맞게 사용(64bit RAM 충분할 때는 효율적)
- 데이터 테이블을 데이터프레임처럼 사용하면 성능은 비슷해짐. 무조건 빨라지는 것 아님

제2절 데이터 가공

1. Data Exploration

- summary : 데이터가 어떻게 분포돼 있는지 보여줌
  - 디멘젼변수 : 각 멤버의 갯수, 결측치 개수(NA's)
  - 메져변수 : 최소값(Min), 1st Q(1사분위값), 중앙값(Median), 평균값(Mean), 3rd Q(3사분위값), 최대값(Max), 결측치 개수(NA's)
- plot : 차트

2. 변수중요도

- 개발 중인 모델에 준비된 데이터를 기준으로 한 번에 여러 개의 변수를 평가
  - 변수 중요도 평가 : 패키지로 평가, 모델링 실행해 평가
  - 모델링 실행 결과(Decision Tree)로 의미 있는 것들에 대해 변수 중요도파악이 일반적
  - 휴면고객 분류 모델개발: 특성 유사한 것끼리 그룹 만들어 실행 후 모아 최종모델 개발
  - 개발 모형 개선 위해 파생변수 추가 : 기존 최종변수에 파생변수 1개 추가해 돌려보고 의미 없으면 버리고 다른 변수 개발해 돌려보고 의미 있으면 선택해 어느 정도 개선되는지 보고 결정하는 식으로 반복 작업
- klaR 패키지
  - greedy.wilks : 모델링 목적에 따른 변수 선택 방법. 모델링을 정의하고 이에 따라 변수를 stepwise하게 투입해 의미있는 변수 순서대로 보여줌
    →효율적으로 정확도를 최소한 희생하면서 초기 모델링 빨리 실행 가능
- 일반적으로 구간화 개수가 증가하면 정확도는 높아지나 속도가 느려지고 추정오차(overestimation) 발생 가능
  - 기본적으로 40개정도를 구간화하고 이를 대상(target)과 비교해 유사한 성능 보이는 인접 구간을 병합하는 것이 적합

제3절 기초분석 및 데이터 관리

1. 데이터 EDA

- 데이터 분석 전에 대략적 특성을 파악하고 데이터에 대한 통찰을 얻기 위해 다각도로 접근

가. 데이터 기초통계

- head(iris) : 데이터 앞 6줄 보여줌
  - head(iris,10) : 숫자 넣어주면 원하는 개수만큼 볼수 잇음
- str(iris) : 데이터 구조 파악
- summary : 데이터 기초통계량
- cor(x,y) : 상관 계수
- cov(x,y) : 공분산

2. 결측값 처리

- 결측값 처리에 시간 많이 쓰는 것은 비효율적
  - 가능하면 결측값은 제외하고 처리하는 것이 적합(결측값 자체가 의미 있는 경우 있음)
- 결측값 어떻게 처리하냐는 전체 작업 속도에 많은 영향
  - 이 부분을 자동화하면 업무 효율성 매우 향상됨
- R 결측값 처리 관련 패키지 : Amelia II, Mice, mistools 등
  - 결측값 : NA(not available), 불가능한 값 : NaN(not a number)
  - NA로 결측값 입력, is.na로 결측값 여부 확인
- 평균 산출 등 데이터 처리에서 결측값으로 인한 문제 해결 : 해당값 제외
  - complete.cases() : 결측값 포함 레코드 삭제
  - imputation : 많은 자료 삭제 방지차원에서 해당 변수의 대푯값으로 대체

■ 이상값(Outlier) 검색

- 분석에서 전처리를 어떻게 할지 결정할 때와 부정사용방지 시스템(Fraud Detection)에서 규칙을 발견하는데 사용
  - 의도치 않게 잘못 입력한 경우 (bad data)
  - 의도치않게 입력됐으나 분석 목적에 부합되지 않아 제거해야 하는 경우 (bad data)
  - 의도되지 않은 현상이나 분석에 포함해야 하는 경우 (이상값)
  - 의도된 이상값 : fraud (이상값)
- 관련 알고리즘 : ESD(extreme studentized deviation), MADM 등
- 이상값 찾는데 너무 많은 시간쓰는 것 비추

• 변수들에 대해 summary 정도로 mean과 median값 파악해 Q1, Q3보고 1차 판단(분포)
• 좀 더 시간되면 주요 dimension별로 플롯해보며 특성파악 가능
  · 부정사용방지 프로젝트(Fraud Detection project)는 여기 많은 시간 할당
- 일정 기간을 할애해 분석 기준 수립해 해당 기준에 의해 안드러나는 것은 무시하고 진행
• 그렇지 않으면 분석 데이터와 결과 자체가 엉망이 돼 관리 불가

제4장 정형데이터 마이닝

제1절 데이터 마이닝의 개요

- 데이터 마이닝 : 대용량 데이터에서 의미있는 데이터 패턴을 파악하거나 예측을 위해 데이터를 자동으로 분석해 의사결정에 활용하는 방법
- 통계분석과 비교해 데이터마이닝의 큰 차이
• 가설이나 가정에 따른 분석이나 검증, 통계학 전문가가 사용하는 도구도 아님
• 다양한 수리 알고리즘을 이용해 DB의 데이터로부터 의미있는 정보를 찾아내는 방법 통칭
- 정보 찾는 방법론에 따라
• 인공지능, 의사결정나무, K-평균군집화, 연관분석, 회귀분석, 로짓분석, 최근접이웃 등
- 분석 대상이나 활용목적, 표현 방법에 따라
• 시각화분석, 분류(classification), 군집화(clustering), 포케스팅(forecasting)
- 사용하는 분야 매우 다양
- 데이터마이닝 도구가 매우 다양하고 체계화돼 도입환경에 적합한 제품을 선택·활용가능
• 데이터 마이닝을 통한 분석 결과의 품질은 분석가의 경험과 역량에 따라 차이
• 분석대상의 복잡성이나 중요도가 높으면 풍부한 경험을 가진 전문가에게 의뢰할 필요
- 통계학 전문가와 대기업 위주시장, 쓰기 힘들고 단순 반복 작업이 많아 실무에서 적극 이용되기 어려움, 데이터 준비위한 추출·가공 부담, 경영진과 어려운 소통, 데이터 핸들링에만 사용, 신뢰 부족

1. 데이터 마이닝 추진 단계

- 데이터 마이닝은 일반적으로 목적 정의, 데이터 준비, 데이터 가공, 데이터 마이닝 기법 적용, 검증 단계로 추진

■ 1단계 : 목적 설정
- 도입 목적을 분명히. 데이터마이닝을 통해 무엇을 왜 하는지 명확한 목적 설정
• 목적 정의 단계부터 시작. 목적은 이해 관계자 모두가 동의하고 이해 가능
• 가능하면 1단계부터 전문가가 참여해 목적에 따라 사용할 데이터 마이닝 모델과 필요 데이터를 정의하는 것이 바람직

■ 2단계 : 데이터 준비
- 데이터 정제를 통해 데이터의 품질을 보장하고 필요하다면 보강해 데이터의 양을 충분히 확보해 데이터 마이닝 기법을 적용하는데 문제없도록 해야 함
• 고객정보, 거래정보, 상품 마스터 정보 등 필요. 웹로그 데이터, SNS데이터도 활용 가능
• 대부분 용량이 크므로 IT부서와 사전 협의해 데이터 접근 부하가 시한 일을 해도 문제없도록 일정 조율하고 도움 요청
• 필요하면 데이터를 다른 서버에 저장 운영

■ 3단계 : 가공
- 모델링 목적에 따라 목적변수를 정의하고 필요한 데이터를 데이터마이닝 sw에 적용할 수 있도록 적합한 형식으로 가공
• 모델 개발단계에서 데이터 읽기, 데이터 마이닝에 부하 걸림 → 모델링 일정계획을 팀원 간 잘 조정

■ 4단계 : 기법 적용
- 앞 단계를 거쳐 준비한 데이터와 데이터 마이닝 sw를 활용해 목적하는 정보 추출
• 적용할 데이터 마이닝 기법은 1단계에서 이미 결정됐어야 바람직
- 데이터 마이닝 모델을 목적에 맞게 선택하고 sw사용하는데 필요한 값 지정
• 어떤 기법을 활용하고 어떤 값을 입력하느냐 등은 데이터 분석가의 전문성에 따라 다름
• 데이터 마이닝 적용 목적, 보유 데이터, 산출되는 정보 등에 따라 적절한 sw와 기법 선정

■ 5단계 : 검증
- 마이닝으로 추출한 정보를 검증하는 단계
• 테스트 마케팅이나 과거 데이터 활용 가능
- 검증됐으면 자동화 방안을 IT부서와 협의해 상시 데이터 마이닝 결과를 업무에 적용할 수 있게 해야 하며 보고서를 작성해 경영진에게 기대효과를 알릴 수 있어야 함

2. 데이터 마이닝을 위한 데이터 분할

- 결과 신빙성 검증을 위해 일반적으로 데이터를 구축용(training), 검정용(validation), 시험용(test)으로 분리
• 구축용 : 초기의 데이터 마이닝 모델 만드는데 사용. 추정용, 훈련용(50%)
• 검정용 : 구축된 모델의 과잉 또는 과소맞춤 등에 미세조정 절차 위해 사용(30%)
• 시험용 : 데이터 마이닝 추진 5단계에서 검증용으로 사용.(20%)
- 데이터 양이 충분치 않거나 사용sw입력 변수에 대한 설명이 충분할 경우 구축용과 시험용으로만 분해 사용하기도 함
- 필요에 따라 구축용과 시험용을 번갈아가며 사용(교차확인;cross-validation)을 통해 모형평가
• 최근에는 구축용과 시험용으로만 분리해 사용하는 추세

3. 데이터 마이닝 모형 평가

- 데이터 마이닝 프로젝트의 목적과 내용에 따라 적합 모형 다름
• 몇가지 모형 대안 놓고 어느 것이 적합한지 판단하는 가장 보편적 기준 : 손익비교
- 모델링은 변경 주기가 있으며 근본적으로 정확도의 편차가 급증하는 시점에 실행
• classification : 최소1년2번, 연관성규칙: 비즈니스특성에 따라 1주/1개월,
  forecasting : 일·주·월 단위 등 모델링 기준에 따라 다름

- 성공적 데이터마이닝 핵심 : 전반적인 비즈니스 프로세스에 대한 이해
  • 각 프로세스에서 어떤 형태로 데이터가 발생돼 변형·축적되는지 이해하고 필요한 데이터 선별가능 해야 함
  • 데이터에 대한 전반적 파악, 팩트와 특이사항 파악해 브레인 스토밍, 마트 잘만들기(자동화), 모델링(처음부터 전체 데이터 접근×,샘플링 최대한 활용)

제2절 분류분석(Classification Analysis)

1. 분류모델링

- 분류분석 : 데이터의 실체가 어떤 그룹에 속하는지 예측하는 데 사용하는 데이터마이닝 기법
  • 특정 등급으로 나누는 점에서 군집분석과 유사하나 각 계급이 어떻게 정의되는지 미리 알아야 함
- 분류(classification) : 객체를 정해놓은 범주로 분류하는데 목적
  • CRM에서는 고객행동예측, 속성파악에 응용. 다양한 분야에서 활용 가능
- 많은 경우 분류모델 개발할 때는 train data/test data 구분지어 모델링
  • 전체 데이터를 7:3, 8:2 등으로 나눠 train 해서 최적모델 확정짓고 test로 검증
  • train과 test간 편차 없어야 하며 성능은 test가 다소 낮게 나오는 경향
- 분류를 위해 사용되는 데이터마이닝 기법
  • 최근접이웃(nearest neighborhood), 의사결정나무(decision tree), 베이지안 정리를 이용한 분류, 인공신경망(artificial neural network), 지지도벡터기계(support vector machine), caret(classification and regression tree) 등
  · 상황판단, 속하는 분류 집단 특성, 예측 등에도 사용

가. 의사결정나무

- 분류함수를 의사결정 규칙으로 이뤄진 나무 모양으로 그리는 방법
  • 나무의 구조에 기반한 예측모델을 갖는 데이터를 분류하기 위한 질문, 잎은 분류 결과에 따라 분리된 데이터 의미
  • 연속적으로 발생하는 의사결정문제를 시각화해 의사결정이 이뤄지는 시점과 성과를 한눈에 볼 수 있게 하며 계산결과가 의사결정나무에 직접 나타나 분석 간편

나. 의사결정나무의 활용

- 세분화(segmentation) : 데이터를 비슷한 특성 갖는 몇 개 그룹으로 분할해 그룹별 특성 발견, 각 고객이 어떤 집단에 속하는지 파악
- 분류(classification, stratification) : 관측개체를 여러 예측변수들에 근거해 목표변수의 범주를 몇 개 등급으로 분류하고자 하는 경우
- 예측 : 자료에서 규칙 찾고 이를 이용해 미래 사건 예측
- 차원축소 및 변수선택 : 매우 많은 예측변수 중 목표변수에 큰 영향 미치는 변수 골라냄
- 교호작용효과의 파악(interaction effect identification) : 여러 예측변수들을 결합해 목표변수에 작용하는 규칙 파악
- 범주의 병합 또는 연속형 변수의 이산화 : 범주형 목표변수의 범주를 소수의 몇 개로 병합하거나 연속형 목표변수를 몇 개 등급으로 이산화

다. 의사결정나무의 특성

- 의사결정나무 모형 결과는 누구에게나 설명이 용이
- 의사결정나무 알고리즘 모형 정확도는 다른 분류모형에 뒤지지 않음
- 만드는 방법은 계산적으로 복잡하지 않아 대용량데이터에서도 빠르게 만들 수 있고 한번 모델링하면 소속집단을 모르는 데이터 분류 작업도 빠르게 할 수 있음
- 의사결정나무 알고리즘은 비정상적인 잡음 데이터에 대해서도 민감함 없이 분류 가능
- 한변수와 매우 상관성 높은 다른 불필요한 변수가 있어도 의사결정나무는 크게 영향 안받음 but 불필요변수 많아지면 나무크기 커질 수 있으니 분류 전 불필요 변수 제거작업 필요

- R에서 지원되는 분류 방법
  • rpart, rpartOrdinal, randomForest, party, Tree, marginTree, MapTree 등 다양

2. 성과분석과 스코어링

가. party 패키지를 이용한 의사결정나무

- party 패키지 핵심 : 의사결정나무(사용편한 다양한 부류 패키지 중 하나)
  • (문제) 분실값(missin value) 잘 처리 못함, tree에 투입된 데이터 표시 안되거나 predict 실패, 명목변수의 테스트 데이터 train과 다르게 처리 등

나. rpart를 이용한 의사결정나무

- rpart는 Recursive Partitioning and Regressin Tree로 CART와 유사한 트리
  • 예측오차 최소화 가능

다. 랜덤 포리스트

- random input에 따른 forest of tree를 이용한 분류방법, 랜덤한 forest에는 많은 tree생성
- 새로운 오브젝트를 분류하기 위해 forest에 있는 트리에 각각 투입해 각각의 트리들이 voting함으로써 분류하는 방식
  • 대용량 데이터에서 효율적으로 실행, 수천개의 변수를 통해 변수제거없이 실행돼 정확도 측면에서 좋은 성과, 특히 unbalanced된 클래스의 모집단 잘 지원
  • (제약) 각 category variable의 value 종류가 32개 넘을 수 없음
  • (대안) party 패키지의 randomforest 사용

라. ROCR 패키지로 성과분석

- 성과분석(performance analysis) : ROC analysis, Lift analysis 등

• ROCR 패키지는 binary classification 만 지원

마. caret(classification and regression tree)
- 분류 관련 알고리즘 수 십 가지가 각각 형식이 달라 혼란스러움
- 전체적으로 동일한 형식 사용할 수 있게 한 caret패키지가 나음

제3절 예측분석(Prediction Analysis)
- 분류 : 불연속적 값 / 예측 : 연속적 값
- 예측 : 시계열분석으로 시간에 따른 값 두 개만을 이용해 앞으로의 매출/온도 등을 예측
  • 두 접근방법은 모델링하는 입력 데이터가 어떤것인지에 따라 특성이 다름
  • 예측은 여러개의 다양한 설명변수가 아닌 하나의 설명변수로 생각

1. 활용분야

가. 행동예측 유형
- 휴면·이탈, 등급변동, 특정상품 구매, 특정금액 이상 구매, 특정시점의 특정조건에 해당되는 행동 예측 등 다양한 경우 가능
  • 특정행동 예측능력이 뛰어날수록 더욱 정교한 고객관계 활동 전개 가능
- 행동예측이 행동의 결과를 모두 상식적으로 설명가능해야 한다는 것은 잘못된 생각
  • 맞으면 됨. 이해할 수 있는 논리 제공하면 좋지만 당연한 것은 아님

나. 휴면·이탈 예측
- 이미 고객을 돌이킬 수 없는 상태까지 가기 전에 보유(retention)하기 위한 방안, 고객의 거래주기를 단축시키지 위한 방안
  • 단기휴면은 예측이 다소 어려우나 매우 도움이 되며 장기휴면 예측은 의미없다고 생각할 수 있으나 효율성 측면에서 의미가 있고 일부 고객(1년간 거래×)은 매우 유용한 정보
  • 구매주기 길고 상품구매 다양성 부족, 구매금액 감소 성향 거래 매장 한정 : 적합한 개인화한 상품과 서비스 추천, 정보제공 중요

다. 등급변동 예측
- 등급하락 고객을 예측해 재구매나 연쇄판매로 등급을 유지하고 등급상승 가망 고객에 보다 더 상승할 수있도록 동기부여 필요

라. 신규고객 우수가망 예측
- 우수고객이 될 고객의 거래 특성은 이미 1년전의 1~2개월간 거래 패턴으로 파악 가능

마. 상품구매 예측
- 개인화한 상품추천을 위한 중요 요소
  • 단순 교차판매의 사례가 아니라 고객이 구매할 만한 상품을 전체적으로 봐야함

바. 캠페인 반응예측
- 반응할 가망성이 높은 고객에게 캠페인을 해야 더 좋은 반응을 얻게되고 의사결정할 임계치를 넘어섬
  • 캠페인에 반응할 사람과 자발적 구매가능성이 낮은 고객을 대상으로 진행하는 것이 적합

2. party 패키지를 이용한 airquality 데이터 선형모델링(lm)

제4절 군집분석(Clustering Analysis)

1. 군집분석 개요
- 특성에 따라 고객을 여러 개의 배타적인 집단으로 나누는 것
  • 결과는 구체적인 군집분석 방법에 따라 다름
  • 군집 개수, 구조에 대한 가정없이 데이터로부터 거리 기준에 의해 자발적인 군집화 유도
- 군집분석의 목적
  • 적절한 군집으로 나누는 것
  • 각 군집의 특성, 군집간의 차이 등에 대해 분석
- 나누는 방법에 따른 군집화 구분
  • 임의적 방법 : 논란여지 많으나 많이 사용되옴
  • 통계적 기법 활용 : 1, 2세대 알고리즘 이용해 사용돼왔으나 실무적용성에 대한 논란

2. 전통적 군집분석

가. 기존 세분화 방법의 유형
- 임의로 나누는 방법 : 고객등급/고객구분(신규/기존), 4분면, 9개 집단 등 다양
- 통계적 기법 : clustering, k-means 등

나. 전통적 세분화 방법의 문제점
- 변수를 선정하고 구간대로 나눈 다음 이를 기준으로 격자형으로 단순히 나누고 집단이 적으면 병합(merge)하는 방식과 단순 clustering, k-means
- 단순 격자형 : 작업에 오랜시간 소요, 후처리로 병합할 때 원칙이 명확하지 않음
  • 분리된 격자 셀의 프로파일을 보고 유사한 근처 집단으로 나눠야는데 차이 안나는 경우 존재
  • 세분화 변수와 프로파일링 변수는 달라야 함
- 격자, clustering, k-means : 변수의 특성으로 인한 변동에 따라 의미없이 고객집단 이동
  • 세분화를 안정적으로 관리하면서 전략 수립과 액션을 할 수 없고 자연스런 변화에 마치 의미있는 것처럼 끌려다님. 세분집단 수가 많은 경우 더 심함
- (해결) k-means에 SRM 결합한 방식 : 세분집단의 변화가 많지 않음. 집단은 안정적으로 유지되며 해당 집단에 속한 고객이 변화 → 세분화를 통한 고객관리 가능

다. 목표기반(target-based) 세분화 방법

- 고객가치 또는 특정상품을 구매하는 고객을 타깃으로 세분화하는 방법
  - 해당 집단이 많이 존재하는 집단/그렇지 않은 집단으로 구분, 이 집단들도 다른 변수에 의해 집단의 특성이 구분됨

라. 프로파일링 방법

- 집단 간에 동일한 변수로 할 수도 있고 서로 다른 변수로 집단의 특성을 규명할 때 사용할 수도 있음
  - 동일변수를 기준으로 집단을 비교할 때 집단 간에 차이가 명확할 수 있으며 그런 변수들이 프로파일링 변수로 선정되며 집단별로 유의미한 변수가 다를 수 있음
  - 격자방식에서는 집단 간에 차이가 나지 않거나 프로파일링 변수가 나오지 않을 수 있으며 때에 따라서는 무의미한 변수로 집단 간에 차이가 있다고 몰아가기도 함
    →누가 사용해도 동일한 결과가 나오는 프로파일링 기법 필요
- 자동화한 방식으로 세분화 되고 프로파일링 돼야 동일한 데이터에 대해 일관된 품질의 결과가 나오고 세분화한 집단의 프로파일링은 집단을 변별하는 가장 유의미한 변수 순서로 표시돼야하며 이것도 자동으로 이뤄져야함
  - 고정된 변수로 다양한 세분집단을 비교 가능해야 함

마. 세분화 수행기간

- 세분화는 데이터 입수가 된 순간에서 마트 생성에 반나절~1일, 군집분석을 통한 세분화 및 보고서 작성에 1일이면 됨
- 군집분석 전 프로파일링하는 것은 무모
  - 프로파일링 먼저 하려면 구조적 특성이나 도메인에 대한 이해 등 시간이 걸림
- (대안) 군집분석 : 일단 군집을 3~10개정도사이로 나눠 군집내 분산크기 통계에 따라 몇 개로 나누는 경우가 적합한지 판단가능, 군집 개수 내에서의 프로파일 보면 한 번에 데이터 전체에 대한 특성 파악 가능
  - 군집분석 실행시간 30분이내, 보고서 작성까지 최대 1일정도
  - 프로파일링은 군집분석 보고서에 자동으로 군집별 measure들의 평균값 나와 평균의 차이 보여줌

바. 세분집단 개수

- 전략적으로 집단을 MECE(Mutually Exclusive and Collectively Exhaustive)하게 나누는데 어느 정도 규모 갖춰야 의미 있으므로 보통 3~10개정도로 나뉨
  - 효율성을 판단하는 정형화한 방법 : 2~15개정도로 군집개수늘렸을 때 집단 내 분산의 크기가 즐어듦을 파악

사. 거리

- 군집분석에선 관측 데이터 간 유사성이나 근접성을 측정해 어느 군집으로 묶을 수 있는지 판단. 측도로 데이터 간의 거리(distance) 이용
- 관측값들이 얼마나 유사한지 측정하는 방법
  - 유클리드 거리
  - 표준화 거리
  - 마할라노비스 거리
  - 체비셰프 거리
  - 맨하탄 거리
  - 캔버라 거리
  - 민코우스키 거리

3. 계층적 군집방법

- n개의 군집으로 시작해 점차 군집의 개수를 줄여나가는 방법
  - 관측벡터간의 거리뿐만 아니라 군집간 거리에 대한 정의 필요

가. 최단연결법

- n×n 거리행렬에서 거리가 가장 가까운 데이터가 U와 V라면 먼저 두 데이터를 묶어 군집형성 후 군집과 나머지(n-2)개의 다른 데이터 또는 군집과 거리 계산
  - 수정된 거리행렬에서 거리가 가장 가까운 데이터 또는 군집을 새로운 군집으로 함

나. 최장연결법

- n×n 거리행렬에서 거리가 가장 먼 데이터가 U와 V라면 우선 두 데이터를 묶어 군집형성 후 군집과 나머지 다른 데이터 또는 군집과의 거리 개산
  - 수정된 거리행렬에서 거리가 가장 가까운 데이터 또는 군집을 새로운 군집으로 함

다. 평균연결법

- n×n 거리행렬에서 거리가 가장 가까운 데이터가 U와 V라면 두데이터 묶어 군집형성 후 나머지 다른 데이터와 군집과 거리계산해 평균거리를 구한후 가장 가까운 데이터와 다시 군집 형성
  - 과정 반복하면 모든 데이터를 포함하는 하나의 군집 형성

라. 와드연결법

- 군집내 편차들의 제곱합을 고려. 군집간 정보 손실을 최소화하기 위해 군집화 진행

4. 비계층적 군집방법

- n개의 개체를 g개의 군집으로 나눌 수 있는 모든 가능한 방법을 점검해 최적화한 군집을

형성하는 것

• K-평균법(K-means method)가 대표적

・원하는 군집개수, 초기값 정해 seed 중심으로 군집 형성→각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류→각 군집의 seed값 다시 계산→모든 개체가 군집으로 할당될 때까지 반복

・K-평균법은 한 개체가 속해있던 군집에서 다른 군집으로 이동해 재배치가 가능. 초기값에 의존. 군집의 초기값 선택이 최종 군집 선택에 영향 미침. 몇 가지 초기값 선택 후 결과 비교하는게 유용

가. 비계층적 군집화의 장점

- 주어진 데이터의 내부구조에 대한 사전정보없이 의미 있는 자료구조 찾을 수 잇음
- 다양한 형태의 데이터에 적용 가능
- 분석방법의 적용이 용이

나. 비계층적 군집화의 단점

- 가중치와 거리정의가 어려움
- 초기 군집수를 결정하기 어려움
- 사전에 주어진 목적이 없으므로 결과 해석이 어려움

5. 최신 군집분석 기법들의 적용

가. K-means

- 일반적으로 K-means사용할 경우 최적 k값의 정확도에 많은 영향. 데이터 학습할 때 일정한 변화줘서 정확도가 어떻게 변하는지 보면 됨

나. PAM(Partitioning Around Medoids)

- 좀더 탄탄한(robust) k-means. 결측값 허용. 프로파일링 시 실제 관측값으로 표현
- 대용량 데이터 처리 시간 급상승하는 단점

다. overall cluster.R

라. Hierarchical Clustering

마. Density-based Clustering

바. Fuzzy Clustering

- 숫자변수만 가능하고 NA 허용됨. k개의 cluster가 생성되는데 개수는 관측값/2개까지 가능

제5절 연관분석(association Analysis)

1. 연관성 규칙

가. 연관성 규칙의 개념

- 장바구니분석(MKT basket analysis), 서열분석(Seq. Analysis)이라 불림
- 포괄적 개념. 흔히 기업의 DB에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용
  • (마케팅) 손님 장바구니에 들어있는 품목 간 관계를 알아본다는 의미에서 장바구니분석
  • 장바구니에 뭐가 같이 들어있나(장바구니분석), A사고 B산다(연관성분석)
- 어느 고객이 어떤 제품을 같이 구매할까? → 연관성 분석 실시
  • 분석을 통해 제품 간 연관성 파악하면 세트메뉴 구성/쿠폰발행 등 교차판매(cross selling)할 때 효과적
- 연관성 규칙의 일반적인 형태 : 조건과 반응(if-A then B) : 연관규칙
  • 모든 규칙에 유용하진 않음
- 유용한 규칙이 되기 위한 조건
  • 두 품목 A와 B를 동시에 구매한 경우의 수가 일정 수준 이상
  • 품목 A를 포함하는 거래 중 품목 B를 구입하는 경우의 수도 일정 수준 이상
- 연관성분석을 통해 쿠폰발행, 가까운 곳 배치 등 의사결정도 가능

나. 연관성분석의 측도

- 연관성규칙을 이용할 수 있는 데이터는 판매시점에 기록된 거래와 품목에 관한 정보를 담고 있어야 함
  • 인구통계학적 자료를 비롯한 기타 정보를 필요로 하진 않음
  • 측정의 기본은 얼마나 자주 구매했는가 하는 빈도(count)
- 연관성규칙 자체를 이해하는 것은 어렵지 않으나 모든 규칙이 유의미한 것이 아니므로 산업의 특성에 따라 지지도(suppert), 신뢰도(confidence), 향상도(lift)값을 잘 보고 규칙을 선택
  • 연관성분석 Average Duration을 고려해 적용
- 지지도 : 전체 거래 중 항목 A와 B를 동시에 포함하는 거래의 비율
  • 전체 거래 중 A와 B를 동시에 포함하는 거래가 어느 정도인지 나타내며 전체 구매 경향 파악 가능
  • 같이 많이 판매되고 있다는 뜻으로 Association Rule이 나왔을 때 적용성이 있는지 판단 가능하고 불필요한 분석을 대폭 줄일 수 있음

$$\text{지지도}=P(A\cap B)=\frac{\text{와}B\text{가 동시에 포함된 거래수}}{\text{체거래수}}$$

  • 얼마나 빈번하게 나타나는 경우인지 설명하는 상대적인 값
- 신뢰도 : 항목 A를 포함한 거래중 항목 A와 B가 같이 포함될 확률. 연관성 정도 파악가능

$$\text{신뢰도}=\frac{P(A\cap B)}{P\ A)}=\frac{A\text{와 }B\text{를 동시에 포함하는 거래수}}{A\text{를 포함하는 거래수}}$$

  • A를 산 고객이 B를 산 비율
  • 조건부확률로 A한사람이 B하더라. 이 값이 높아야함
- 향상도 : A가 주어지지 않았을 때의 품목 B의 확률에 비해 A가 주어졌을 때의 품목 B의

확률의 증가 비율

• 관련없는 경우 향상도=1, 향상도>1 우연보다 우수, 향상도<1 우연적기회보다 도움 안됨

・향상도 1이면 서로 독립적 관계, 1보다 작으면 음의 상관관계, 크면 양의 상관관계

$$\text{향상도} = \frac{P(B|A)}{B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A\text{와 } B\text{를 포함하는 거래수}}{A \quad \text{포함하는 거래수} \times B\text{를 포함하는 거래수}}$$

- 규칙'A→B'가 의미있다면 전체 거래에서 품목 B를 포함한 거래의 비율보다는 품목 A가 구매된 거래내에서 품목B를 호함한 거래의 비율이 더 클 것
- 연관성분석을 수행할 때 모든 경우의 수를 분석하는 것은 매우 불필요. 최소 지지도(min. support)를 정해 규칙(rule)을 도출

• 처음엔 5%정도로 임의 설정해 산출해보고 현실적인지, 규칙은 충분히 도출됐는지에 따라 지지도 조절해 다양한 시도해봐야함. 처음에 너무 낮은 지지도 선정은 매우 불필요

• 실제로 높은 값에서 낮은 값으로 설정해 처리속도와 규칙 개수 파악한 다음 낮추는 방법 필요

• 컨설팅 프로젝트에서는 4시간이내에 답이 나와야하고 운영모드는 24시간 걸려도 적합

• 선행사건(antecedent)→후건(consequent), 때에따라 선행/후건 기준으로 결과봐야함

다. 연관규칙분석 절차

- Apriori : 최조 지지도를 갖는 연관규칙을 찾는 대표적인 방법

• 최소 지지도보다 큰 집합만을 대상으로 높은 지지도를 갖는 품목 집합 찾음

• 분석절차

①최소 지지지도 정함

②개별 품목 중 최소 지지도 넘는 모든 품목 찾음

③2에서 찾은 개별 품목만 이용해 최소 지지도 넘는 2가지 품목 집합 찾음

④위의 두 절차에서 찾은 품목집합을 결합해 최소 지지도 넘는 3가지 품목집합 찾음

⑤반복적으로 수행해 최소 지지도가 넘는 빈발품목 집합 찾음

라. 연관규칙의 장점

- 탐색적인 기법 : 조건반응(if-then)으로 표현되는 연관성 분석의 결과 이해 쉬움
- 강력한 비목적성 분석기법 : 분석방향이나 목적이 특별히 없는 경우 목적변수가 없어 유용
- 사용이 편리한 분석 데이터 형태 : 거래내용에 대한 데이터를 변환없이 그 자체로 이용가능한 간단한 자료구조 갖는 분석방법
- 계산의 용이성 : 분석을 위한 계산이 상당히 간단

마. 연관규칙의 단점

- 상당수의 계산과정 : 품목수가 증가하면 분석에 필요한 계산은 기하급수적으로 증가
- 적절한 품목의 결정 : 너무 세분화한 품목으로 연관성 규칙을 찾으면 의미없는 분석가능성 있음
- 품목의 비율차이 : 거래량이 적은 품목은 포함된 거래수가 적고 규칙발견시 제외하기 쉬움

바. 순차패턴

- 동시에 구매될 가능성이 큰 상품군을 찾아내는 연관성측정에 시간개념을 포함시켜 순차적 구매 가능성이 큰 상품군을 찾아내는 것

• A가 구매되면 일정 시간이 경과한 다음 B가 구매된다

- 구매의 순서가 고려되 상품간 연관성이 측정되고 유용한 연관 규칙을 찾는 기법

• 연관성측정 데이터에 각 고객의 구매시점 정보 포함

2. 기존 연관성분석의 이슈

- SAS E-miner와 Clementine의 문제 : 대용량 데이터에대한 연관성 분석 불가능

• 오래된 알고리즘인 apriori 사용해 SKU레벨의 연관성분석시 시스템 먹통

3. 최근 연관성분석 동향

- KXEN : 처음부터 3세대 FPV를 이용해 메모리를 효율적으로 사용함으로써 SKU레벨의 연관성분석을 성공적으로 적용

4. 연관성분석 활용방안

- 장바구니분석은 실시간 상품추천을 통한 교차판매에 응용가능

• 최근 실시간추천이 가능해짐에 따라 활용도가 올라갈 것으로 예상

- 시차분석은 A를 구매했는데 B를 구매안한 경우 B를 추천하는 교차판매 캠페인에 사용가능
- 기업이 갖고 있는 데이터가 바로 연관성분석에 사용될 수 없음

• 정보를 가공해 연관성 규칙을 사용할 수 있는 데이터로 전환

- 결과 검증 : 테스트 마케팅

• 기존방식에 대한 반응율이 얼마인지 정확한 기준으로 평가하고 연관성규칙 적용한 테스트 마케팅 기획

• 연관성규칙 도출과 타겟팅해 컴페인기획하는 것은 별개

• 무조건 고객에게 규칙을 적용해 추천하지 않아야 함

• 채널에 대한 민감도 고려

・50대에 이멜 효과 없음, SMS는 과거이력 기반으로 타겟팅, 활동고객 기준

- 연관성 규칙의 효과와 타깃팅 효과를 결합해 테스트 하는 것이 필요

• 테스트 마켓팅 사이즈를 충분히 크게 해야 비율의 증가가 검증 가능

제5장 비정형 데이터 마이닝

제1절 텍스트 마이닝

- 텍스트로부터 고품질의 정보를 도출하는 과정. 입력된 텍스트를 구조화해 그 데이터에서 패턴을 도출한 후 결과를 평가·해석하는 일련의 과정 통칭

• 다양한 포맷의 문서로부터 데이터를 획득해 이를 문서별 단어의 매트릭스로 만들어 추가 분석이나 데이터 마이닝 기법을 적용해 통찰을 얻거나 의사결정을 지원하는 방법
- 다양한 포맷의 문서로부터 텍스트를 추출해 이를 하나의 레코드로 만들어 단어 구성에 따라 마트를 구성하고 이들 간의 관계를 이용해 감성분석(sentiment analysis)이나 워드 클라우드(word cloud)를 수행하고 이 정보를 클러스터링이나 분류와 사회연결망 분석에 활용 가능
• 예) 주고 받은 문장을 API로 읽어 분석해 평판관리와 마케팅 활동을 실시간 관리 가능, 경쟁사 브랜드에 대한 반응 모니터링으로 경쟁 전략 수립 가능, 효율적 검색을 위해 주제어 분리해 체계적 관리방안에 활용, 특정 분야의 전문가 알아내는 데 활용 가능 등

1. 텍스트 마이닝 기능 요약

- 문서요약(summarization)
- 문서분류(classification)
- 문서 군집(clustering)
- 특성 추출(feature extraction)

- 해당 언어에 대한 깊이 있는 이해와 문화와 관습에 대한 이해 필요
• 국가별로 다른 접근 방식의 분석을 수행해야함(어려운 점)

2. 정보검색(information retrieval)의 적절성
- 정확도와 재현율(Precision & Recall) : 자연어 처리 분야에서 분석 결과를 평가하기 위해 사용하는 대표적 방법
• 정확도 : 분석모델이 결과 중에서 정답과 일치하는 비율
• 재현율 : 실제 정답 중에서 분석모델에서 정답이라고 내놓은 결과의 비율
• 일반적으로 정확도와 재현율은 반비례 관계

가. Corpus
- 데이터 마이닝의 절차 중 데이터의 정제, 통합, 선택, 변환의 과정을 거친 구조화된 단계로서 더 이상 추가적 절차 없이 데이터 마이닝 알고리즘 실험에서 활용가능한 상태
• 텍스트 마이닝 패키지(TM)에서 문서를 관리하는 기본구조, 텍스트 문서들의 집합
- VCorpus로 메모리에서만 유지되는 Corpus와 R외부의 DB나 파일로 관리되는 PCorpus로 나뉨
- Corpus를 다른 object에서 가져온 경우 default working directory에 개별파일로 저장됨
- stop word : (한글)조사, 띄어쓰기, (영어)띄어쓰기, 시제 등 내용 제거 & 표준화

나. Create Term-Document Matrix
- 읽어들인 문서를 plain text전환, space제거, lowercase로 전환, punctuation제거, stopword 처리 stemming 등 처리한 후 문서번호와 단어간 사용 여부, 빈도수를 이용해 matrix 만드는 작업

다. Dictionary
- 복수의 문자들의 집합. 텍스트 마이닝에서 분석에 사용하고자 하는 단어들의 집합
• 단어 추가 가능 : 분석하고자 하는 단어들을 별도 사전으로 정의해서 해당 단어들에 대해서만 결과를 산출해 볼 때 사용

라. 감성분석(Sentiment Analysis)
- 흔히 Opinion Mining 등으로 언급
- 문장에서 사용된 단어의 긍정과 부정여부에 따라 얼마나 긍정적인 단어가 많은지 소스를 부여해 긍정 문장인지 평가
• 브랜드 평판 분석 가능. 주체에 따라 다르게 해석 가능'
- 복잡한 문장을 분석할 때 개별 문장이나 문서에 대해서는 오류발생 가능
• 개별문장 분석 오류 나도 수많은 문서나 데이터 가공하면 그 추이 파악에는 무리없어 감성분석에 대해 부정적일 필요는 없음
- 트위터에서 자료가져오는 방식
• 웹페이지에서 HTML을 데이터로 가져와 파싱(parsing)하는 방식
• API를 이용해 자료를 가져오는 방식(인증필요, 데이터양 제한)
• callback URL 사용 불가

마. 한글처리
- KoNLP가 대표적. JRE(java runtime environment) 반드시 설치

바. 워드 클라우드
- 문서에 포함된 단어의 사용 빈도를 효과적으로 보여주기 위함
• available.package : R에서 사용가능한 패키지 리스트 출력

제2절 사회연결망 분석(Social network analysis, SNA)
1. 사회연결망 분석 정의
가. SNA 정의
- 개인과 집단들 간의 관계를 노드와 링크로서 모델링해 그것의 위상구조와 학산 및 진화 과정을 계량적으로 분석하는 방법론
• 사회연결망 : 개인의 인간관계가 인터넷으로 확대된 사람 사이의 네트워크. 다양한 분야에서 응용
• 기존 사회연결망에 대한 관심이 주로 그룹 간 또는 그룹 안의 개인에 집중한 반면 처음 사회연결망 용어 사용한 Barnes는 독립 네트워크 사이의 관계에 대해 집중
- 분석방법 : 집합론적 방법, 그래프 이론에 의한 방법, 행렬을 이용한 방법 등

1) 집합론적 방법
- 객체들의 집합에서 각 객체들 간 관계를 관계 쌍(pairs of elements)으로 표현

2) 그래프 이론을 이용한 방법

- 객체는 점(꼭지점, 노드)으로 표현되며 두 객체 간 연결망은 선으로 표현

3) 행렬을 이용한 방법

- 각 객체를 행렬의 행과 열에 대칭적으로 배치하고 I번째 객체와 j번째 객체가 연결망으로 연결돼 있으면 행렬의 (i,j)에 1넣고, 없으면 0 넣음
- 분석하고자 하는 데이터는 행렬로 표현
  - 행과 열이 만나는 셀에 특정 값을 입력해 행과 열 사이의 관계 나타냄
  - 행과 열에 같은 개체 배열 : 1원(1 mode)자료, 다른 개체배열 : 2원(2 mode)자료
  - 관계를 표현하는 기본적 방법 : 관계 존재하면 1 존재하지 않으면 0 입력
- 준연결망(quasi network) : 직접적인 상호작용 없어도 관계를 인위적으로 설정해 관계를 나타낸 네트워크
- 네트워크 구조를 파악하기 위한 기법 : 중심성(Centrality), 밀도(density), 구조적 틈새(structural hole), 집중도(centralization) 등
  - 연결정도 중심성(degree centrality) : 한 점에 직접적으로 연결된 점들의 합
  - 근접 중심성(closeness centrality) : 각 노드 간 거리를 근거로 중심성 측정(직간접 거리 합산)
  - 매개 중심성(betweenness centrality) : 네트워크 내 한 점이 담당하는 매개자 혹은 중개자 역할의 정도로 중심성 측정
  - 위세 중심성(eigenvector centrality) : 연결된 노드의 중요성에 가중치를 둬 중심성을 측정하는 방법

나. SNA 적용

- 소셜 네트워크는 노드 또는 점(vertex), 링크 또는 에지로 구성된 그래프
  - 링크 방향성 여부에 따라 방향(direct) 그래프와 무방향(undirected) 그래프로 구분 가능
- 분석용 솔루션으로 KXEN, SAS, XTRACT, Indiro, Onalytica, Unicet 등이 있으나 데이터 로딩 속도, 시각화 기능 등 제약이 있어 더 발전이 필요
  - 소규모 네트워크 대상 분석 수행은 쉽지만 실제 기업 수준 애플리케이션에서 분석 수행하는 데에는 많은 어려움 존재
- R은 모든 데이터를 메모리에 로드해야 분석가능하므로 대규모 데이터를 분석하려면 분산처리 프레임워크가 반드시 필요
  - 분산처리기술인 하둡 MapReduce 활용/하둡기반 그래프 프로세싱 프레임 워크 Giraph로 R에서 처리 가능한 수준까지 정제한 후 분석 및 가시화 수행 가능

다. 단계

- 사회연결망 분석 단계
  - 그래프 생성
  - 그래프를 목적에 따라 가공해 분석
  - 커뮤니티를 탐지하고 각 객체 또는 노드의 롤을 정의해 어떤 롤로 영향력을 보다 효율적으로 줄 수있는지 정의
  - 데이터화해 다른 데이터 마이닝 기법과 연계
  - ·소셜네트워크 분석결과로 얻어진 커뮤니티 프로파일을 고객 프로파일 평균값으로 산출해 각 그룹에 속한 개별 고객 속성에 그룹넘버와 롤을 곁바해 속성을 추가하는 업무

2. R을 이용한 SNA 기본 사용법

가. 데이터 로딩

- read.table()
- URL에서 데이터 직접 로딩

1) 그래프 로딩

2) 그래프에 꼭지점 속성 추가

3) 네트워크 시각화

4) 그래프를 다양한 타입으로 저장

나. 데이터 로드

다. 노드 레벨 통계값

- subcomponent : 입력으로 주어진 그래프에 대해 특정 vertex와 연결된 vertex 집합(connected component)을 반환해 연결정보 알수 있게 해줌

라. 네트워크 레벨 통계량

- 대부분 네트워크 분석은 distances와 reach로 시작해 점차 글로벌한 summary로 나아감

1) Degree

2) Shortest paths

3) Reachability

4) density

5) Recipreocity

6) Transitivity

7) Triad census

마. Clustering

바. community detection

1) COMMUNITY DETECTION: WALKTRAP

- walktrap 알고리즘은 일련의 random walk 과정을 통해 커뮤니티 발견

2) COMMUNITY DETECTION: EDGE BETWEENNESS METHOD

3. 활용방안

- 네트워크를 구성해 몇 개의 집단으로 구성되는지, 집단간 특성, 해당집단에서 영향력있는 고객,

시간의 흐름과 고객상태의 변화에 따라 누가 다음 영향을 받을지 기반으로 churn/acquisition prediction, fraud, product recommendation 등에 활용 가능

가. 단어 간 연관성을 이용한 소셜 네트워크 분석
나. 트위터 검색을 통한 사용자 간 소셜 네트워크

제6장 시뮬레이션 및 최적화
제1절 빅데이터와 시뮬레이션
- 데이터 마이닝 : 대용량 DB에서 숨어있는 예측 가능한 정보를 자동으로 추출하는 데이터 분석방법
  • 데이터에서 쉽게 발견하기 어려운 정보를 발견하거나 특정상황을 예측하는 것
- 데이터 마이닝은 모든 분야 Business Intelligence는 기업에서 데이터 분석을 통해 효율적인 의사결정을 하도록 지원하는 시스템과 기술
- 데이터 마이닝을 정의하는 핵심용어 : 자동화(automated), 숨겨진(hidden), 예측가능(predictive)
  • 이전에 발견되지 않았던 데이터들 간의 상호관계를 분석하는 것
- 더 많은 데이터는 시뮬레이션 예측의 정확도를 높임
- 시뮬레이션 : 실제 테스트해보기 어려운 초대형 프로젝트나 위험한 테스트 등을 대신해 행하는 모의실험
  • 실제상황을 컴퓨터 모델로 축약해 표현하는 방법을 통해 물리적 변화없이 가상으로 실제상황을 재현함으로써 문제점을 발견하고 해결하거나 예측하는 기법
- 시뮬레이션 활용 분야 : 다양한 분야에서 문제해결을 위해 이용
- 시뮬레이션으로 상황을 결정하는 각 요인이 결과에 어떤 영향을 미치는지 알아볼 수 있음
- 통계학 관점에서 상황예상하는 dynamic 시뮬레이션, 미분방정식 등 이용하는 확정적 방법 등
  • 모델에 따라 연속형/이산형으로 나누기도 함
- 시뮬레이션 프로그램 구현시 부분적으로 최적화 방법 활용하기도 함
  • 최적화 : 어떤 제약조건이 있을 수도 있는 상황에서 목적함수의 최대값과 최소값을 찾는 것
  • 목적함수에 포함된 변수들은 제약조건에 해당 변수들에 대한 제약이 표현돼있고 이를 만족해야 함
- 최적해를 찾기위해 최적화할 때 보통 혼합정수프로그래밍(MIP)이나 선형 프로그래밍(LP) 사용
  • 최적화는 주로 네트워크 최적화, 정유사 터미널 배치 등 할당관리, 물류등 경로 최적화, 소매 네트워크의 체인 재고관리 등에 사용
- 시뮬레이션은 전체 공급 체인의 각 변수변화가 전체에 어떤 영향을 끼치는지 알수 있게함
  • 근본적인 질문에 대한 답을 주나 최적화가 최적해를 주는 것은 아니며 다만 최적화를 통해 도출된 최적해가 어떻게 동작하는지 시각적으로 보일수 있음
- 시뮬레이션과 최적화는 근본적으로 빅데이터가 필수는 아님
  • 입수 가능한 데이터가 풍부해짐에 따라 시뮬레이션이 더욱 각광받고 있으며 최적화 또한 함수에 포함된 계수 등을 정확히 추정하는데 큰 역할
- 일반적으로 충분히 많은 데이터가 있다면 그 데이터는 정규분포를 따름
  • 이 분포를 이용해 평균 대기시간을 추정해 대기시간 최소화 가능
- 최적화는 상시 운영하는 경우가 드물고 주기적으로 수행
- 데이터 분석과 시뮬레이션의 관계
  • 데이터 분석의 흐름에 따라 여러 프로세스의 기기에서 나오는 머신 데이터를 저장 관리할 수 있게 되면서 전책적인 흐름이 명확하고 속성의 변화나 확률적 분포가 empircal distribution으로 대체 가능해짐
  • 정보의 지속 공급이 모델에 반영되며 미래를 예측하고 폐회로(closed loop) 액션을 하는게 용이해져 분석적 업무를 운영업무 수준에 가깝게 관리 가능해짐
- 최적화와 데이터 분석의 연관점 : 입력데이터 측면과 활용 측면
  • 입력데이터 측면 : 입력데이터가 최적화 모델링의 목적함수나 제약조건의 계수값에 사용되는데 데이터 분석으로 값의 산출이 쉬워짐. 다양한 머신 데이터에서 정보 획득 가능해 최적화의 제약조건에 반영 가능해짐
  • 활용측면 : 점점 많은 정보 발생하고 있고 다양한 문제를 해결해야 하는 시대가 돼 모든 문제에 집중할 수 없으므로 취사선택해야 하는 일이 늘어남. 어디에 집중해야는지 얼만큼 자원할당해야는지 최적화 접근이 더욱 필요해짐

제2절 시뮬레이션
1. 시뮬레이션이란?
- 실제상황을 수학적으로 모델화하고 그 모델을 컴퓨터에 프로그램으로 저장한 후 일어날 수 있는 가능한 모든 상황을 입력함으로써 각각의 경우에 어떤 결과가 도출되는지 예측
  • 시뮬레이터 : 시뮬레이션 모델에 대한 프로그램을 사용자들이 편리하게 사용하고 그 결과를 시각적으로 볼 수 있도록 만든 컴퓨터 기능
- 실제 상황을 모델링하고 프로그램하기 위해 고급 인력을 써야므로 초기 비용이 많이 듦
  • 일단 프로그램화하면 사용자가 여러 경우를 맘대로 가정해 결과확인이 가능하기 때문에 결과적으로 비용절약이 가능하고 특히 짧은 시간에 미래예측에 효과적

가. 시뮬레이션의 정의
- 활용분야에 따라 여러 의미로 정의가능. 일반적으로 주어진 조건 하에서 실제 상황 속에서 모의실험을 통해 정보를 얻는 수리적 실험기법

1**) 시뮬레이션 모델 구분**(일반적)
- 정적 시뮬레이션 모델과 동적 시뮬레이션 모델
  • 정적 : 어떤 정해진 시간 안에서 시스템이나 시간이 필요 없는 시스템(ex. 몬테칼로)
  • 동적 : 시간에 따른 현상 파악하기 위한 모델
- 결정론적 모델

• 랜덤변수와 같은 확률변수를 포함하고 있지 않은 모델
- 연속형 모델과 이산형 모델
• 연속형 : 상태변수가 연속형
• 이산형 : 상태변수가 시간으로 분리된 점에서 순간적으로 할 때 표현
• 연속형과 이산형 모델이 완전히 분리될 필요 없음

나. 시뮬레이션의 장점
- 복잡한 현실문제는 추리적 방법으로 해결책 못 구할수 있는데 이때 유일한 해결책
- 여러 대안 쉽게 비교가능
- 현실문제와 근접하게 만들 수 있어 이해와 사용이 편리하며 문제 해결방법에 대해 의사결정자와 대화가 용이
- 많은 시간이 지난 후 결과를 알 수 있는 문제를 시뮬레이션으로 단시간에 결과 예측 가능(시간단축효과), 반대로 시간 확장시켜 시뮬레이션 가능(시간확장효과)

다. 시뮬레이션의 단점
- 모형 개발에 많은 경험과 노력이 필요
- 확률적 시스템을 시뮬레이션할 때 관찰한 입력 자료를 사용해 얻은 결과는 하나의 표본값에 해당하므로 여러 개의 표본값을 구해 통계처리를 해야며 시간이 많이 소요됨

- 시뮬레이션 프로그램은 대부분 복잡하고 매우 방대해 결과가 예상과 다르게 나오거나 중간에 문제가 생겼을 때 이를 해결하는 것도 쉽지 않음
• 중간에 문제발생시 과정이 명확하다면 단계별 검정으로 쉽게 문제점 찾을 수 있음

라. 시뮬레이션의 과정
- 1단계 : 문제 정의와 모델의 필요조건 규명
• 문제를 명확하게 구체화해 정의 : 문제에 대한 목적이 명확·구체화 됐을 때 적합한 방법론을 적용해 유용한 결과를 얻은 가능성이 높음
• 문제가 명확하고 구체화될수록 시뮬레이션의 활용도도 높음
- 2단계 : 기대와 손실에 관한 평가
• 복잡하고 규모가 큰 시뮬레이션은 비용과 노력이 많이 소요되므로 초기 단계에서 시뮬레이션 수행에 따른 손익에 관한 타당성 평가를 하는 것이 바람직
- 3단계 : 시뮬레이션 모델 개발의 방법 결정
• 모델 개발에 플로차트(flow chart)와 같이 모델 과정의 각 단계를 적절한 기호로 표시하는 방식이 많이 쓰임
• 플로차트 방식 : 시뮬레이션의 논리적 과정을 도표로 단순화하고 시각적으로 표현가능
• 데이터가 모델개발뿐 아니라 모델 적합성 평가에도 활용되기 때문에 모델 개발에 필요한 데이터 수집이 특히 중요
• 시뮬레이션 모델개발 : 기능별 상향식(bottom up) 개발방식, 총괄적 하향식(top down) 개발방식
・기능별 상향식 : 기능별로 분석·정립된 소규모 단위의 모델들을 전체 논리적 흐름에 맞춰 연결
· 복잡한 총체적 모델을 일괄적으로 분석하는 것 보다 훨씬 용이
· 하위모델의 연결도 총체적인 모델의 수정이 필요할 때 쉽게 할 수 있게 신축성있게 수행
- 4단계 : 모델의 프로그램화
• 개발한 시뮬레이션 모델을 컴퓨터 프로그램으로 전환
· 전문 프로그램 언어는 비용과 시간절약 뿐 아니라 프로그램의 유연성과 표준화한 시뮬레이션 전문용어 등에 따른 시스템의 호환성을 높여주므로 전문 프로그램언어를 사용하는 것이 바람직
- 5단계 : 모델의 적합성 평가
• 일반적으로 시뮬레이션 모델의 적합성 평가는 실험적 실행결과를 비슷한 조건에서 얻은 실제 자료와 비교해 파악 가능. 테스트 결과 검정은 실제 문제의 본질을 잘 파악하고 있는 관리자가 주로 수행
- 6단계 : 시뮬레이션 모델 실행
• 시뮬레이션 모델 프로그램 작성과 적합성 조사가 끝난 후에 수행하며 선정된 계획과 목적에 입각해야 함
• 여러 실험계획에 따른 모델 실행이 이뤄지며 결과 분석에 입각해 모델 수정 가능
• 실질적 데이터를 얻기 위해 시스템의 실제 사용자가 참여하는 것이 바람직하며 실행시간은 모델 정립 목적에 따라 결정
• 확정적 모델이면 한 차례 실행으로 가능하나 확률적 모형일 경우 다양한 매개변수가 파생되므로 필요한 결과가 도출될때까지 여러번 실행
- 7단계 : 시뮬레이션 실행결과 분석
• 타당성 여부는 그 결과의 분석이 얼마나 현실적으로 활용될 수있냐에 따라 결정됨
• 분석결과 현실적 활용성이 높을수록 조정의 필요성이 감소할뿐아니라 상대적으로 위험률도 낮아지며 반대의 경우 현실적 타당성과 활용성에 문제가 있다는 의미로 문제의 심각성 정도에 따라 부분적/전면 재검토해 수정하거나 다른 기법 적용 고려

- 시뮬레이션 과정은 문제의 본질, 모델의 형태, 필요한 자료 및 변수들의 수집 가능성과 상호연관성 등에 따라 실행절차가 보다 단순화되거나 각 단계의 실행순서가 바뀔 수 있음
- 시뮬레이션은 현실적 타당성과 활용성이 높은 결과 도출을 위해 상황에 적합하고 유연성있게 수행되는 것이 바람직
• 불규칙 특성을 반영하기 위한 데이터 수집방법 : 인터뷰나 실제 관측
→ 확률분포를 가정해 모델링에 반영
・발생되는 랜덤 number는 자기상관이 없어야하며 장기적으로 반복되는 패턴이 나오면 안됨(run tset로 검정 가능)
- 대부분 모델링은 실제와 완전 동일하게 만들지 않음 : 실제와 동일하면 모델의 복잡성이 증가해 실험 수행 시간이 증가할뿐 아니라 정확성 측면에서 오류가능

• 모델을 얼마나 정교하게 할지 전체 모델에 걸쳐 일정 수준으로 유지해야함
- 시뮬레이션은 현재 상황을 모델링하고 예측해 개선안을 도출함
• 개선안 적용하기 전에 개선 후 상황까지 시뮬레이션 해 현재와 비교
• 개선결과 분석할 때는 집단 간 평균차이를 보는 t검정 등 통계분석 적용
- 시뮬레이션 모델링의 핵심 : 어렵고 복잡한 일을 쉽고 단순하게 분석하는 것
• 확률적 내용은 이산형으로 처리, 인과관계 표현되게 규칙 정의, 언제든 간단히 수정가능한 모델로 만들어야 함
- 하나하나의 관찰에 집착하지 말고 패턴을 읽을 수 있도록 접근해야 함
• 분석할 데이터가 너무 크다면 효율적으로 샘플링 해야 하는지 숙고

제3절 최적화

- 최적화 기법으로 체계적으로 접근해 결정하기는 쉬운일이 아니며 결정의 질 또한 평가하기 어려움
- 최적화방법으로 선형계획법(수리계획법 분야의 한 종류)을 가장 많이 사용
- 최적화 모델은 목적식을 최대화/최소화 하기도 하며 등식/부등식이라는 제약식을 가짐
• 최적화 적용함으로 최적 경영기법을 구할 수 있는 것은 아님
- 모델의 기본적인 두 가정
• 계수의 확실성 : 불확실성이 존재할 경우 최적화 방법 적용 불가
・민감도 분석 : 자료 미비로 계수의 정확도를 알기 어려울 경우 계수 값을 여러 가지로 추정하면서 해의 결과가 어떻게 변하는가를 보고 의사결정
• 명확한 함수 형태
• 최적화는 문제의 성격과 목적에 따라 최소점/최대점을 찾는 방법(근×)
・제약조건 만족시키는 범위에서 목적함수의 값을 극대화/극소화 하는 방법
• 몬테칼로 시뮬레이션과 최적화를 결합해 안정적인 모델구축 방법 적용
• 민감도 분석 : 최적모델링의 입력 값이 어느 범위에서 최적이 유지되는지에 대한 분석
・shadow value : 독립변수가 한 단위 증가했을 때 변화되는 목적 값의 변화량
- 복잡한 문제를 풀어가는 과정은 단계별로 명확히 밝아나가는 것이 차후 문제 발견이나 다른 문제에 재적용하기에 좋음
- 최적화 성공수행 단계별 과정
• 1단계 : 문제이해. 분석하려는 문제점이 무엇인지 충분히 이해
• 2단계 : 의사결정을 위한 변수 정하기. 어떤값을 구하는지 명확히 정함
• 3단계 : 해의 우열을 결정하는 기준 선택. 최대/최소 찾기 위해 반드시 필요
• 4단계 : 3단계에서 정한 기준을 의사결정변수들의 함수식으로 표현하고 목적함수가 분명히 나타나도록 함
• 5단계 : 모든 조건이 의사결정변수 식으로 나타나도록 제약식 만들기
• 6단계 : 입력 자료를 수집하거나 추정. 앞서 수립한 모형에 필요한 자료를 모두 수집가능한지 수집에 얼마나 걸리는지 확인(사내자료, 필요자료 추출, 외부자료)
• 7단계 : 모형 개발 후 최적해 구함

- 최적화할 대상을 적합하게 표현하는 모델이 핵심
- 제약조건에 따라 해가 없는(infeasible) 경우가 나올 수 있음
• 모델링 때 현상을 잘못 파악해 발생 가능
• 강제로 모델 변경하면 안됨
- 실무와 관련된 복잡한 모델은 간단한 모델에서 복잡한 모델로 점진적으로 복잡성을 올려가며 적합한 선에서 해결