1과목

데이터 유형 : 정성적, 정량적
지식의 차원 : 암묵지, 형식지 (공통화 표출화 연결화 내면화) 공표연내
데이터 계층 구조 : DKIW
데이터베이스 특징 : 통합 저장 공용 변화
데이터베이스 설계 절차 : 요구분석 개념적 논리적 물리적
NOSQL ; 몽고DB HBASE REDIS
DW4대 특징 : 주제중심, 통합적, 시계열성, 비소멸성 주통시비
BI : 의사결정 정보 획득 및 경영활동 활용, 하나의 특정 비즈니스
BA :통계적이고 수학적 분석에 초첨, BI보다 큼

빅데이터 크키, 다양성, 속도, Value는 비즈니스 요소, 나머지는 ROI요소
빅데이터 활용기법 : 연관규칙, 유형분석, 유전알고리즘(최적화 필요한 문제 해결책)
기계학습(시청기록 바탕), 회귀분석(수치데이터 분ㄴ석), 감정분석(고객 원하는 것 찾아냄)

빅데이터 위기종류
사생활침해(동의제에서 책임제로), 책임원칙훼손(범죄예측프로그램), 데이터 오용(알고리즘 접근)
데이터사이언스 구성요소 : IT, 분석, 비즈니스컨설팅
데이터사이언티스트 역량 : 하드스킬(모델링등), 소프트스킬(커뮤니케이션, 스토리텔링)
데이터사이언티스트 분석모델 고려사항 : 모델 범위 바깥의 요인은 판단하지 않음
의사결정오류 : 로직 오류, 프로세스 오류,
가치패러다임의 변화 : 디지털제이션, 커넥션, 에이전시

3과목 데이터 분석 기획

분석주제유형
분석대상 방법 모두 알 때 최적화
분석대상 대상 알고 방법 모를 때 솔루션
분석대상 모르고 방법 알 때 통찰
분석대상 모르고 방법 모를 때 발견

목표시점별 분석기획 방안 - 과제단위 Speed와 TEST, 마스터플랜단위 정확성과 배포
분석기획시 고려사항 : 가용한 데이터(데이터의 유형분석이 선행적으로), 적절한 유스케이스(유사 분석 시나리오 및 솔루션 최대 활용), 장애요소들 사전계획 수립

데이터 유형, 저장방식
정형 : ERP, CRM, Demand Forcast
반정형 : Competior Pricing, Sensor

비정형 : email, S NS, IOT, NEWS

저장방식 : RDB, NOSQL, HDFS

분석방법론 구성요소 : 절차, 도구와 기법, 방법, 템플릿과 산출물
프레이밍효과 : 동일한 사건이나 상황임에도 사람들의 선택과 파단이 달라지는 현상, 시각에 따라 해석 달라진다는 이론

분석 방법론 3가지 : 폭포수, 나선형, 프로토타입
폭포수 : 하향식, 피드백, 순차
나선형 : 반복 점증, 복잡도 상승
프로토타입 : 상향식, 요구사항 규정 어려울 때 사용, 가설생성, 모형통해 가설확인, 개발검증과 양산검증 해야 시제품 됨

KDD : 비즈니스 도메인에 대한 이해와 프로젝트 폭표 설정
데이터선택> 전처리(잡음,이상치,결측치제거) > 변환(변수선택,차원축소) > 마이닝(패턴 예측), 평가

CRISP :
업무이해(목적파악>상황파악>마이닝목표설정>프로젝트계획수립) > 데이터이해(데이터수집,탐색품질확인) > 데이터준비(데이터셋편성, KDD의 변환과 같음) > 모델링(기법과 알고리즘선택, 파라미터최적화, 머신러닝) > 평가 > 전개

빅데이터 분석 방법론
분석기획 : 비즈스이해 및 범위설정, 프로젝트정의 계획수립(SOW), 위험계획 수립(회피 전이 완화 수용)
데이터준비 : 필요데이터 정의, 데이터 스토어 설계, 데이터 수집 및 적합성 점검
데이터분석 : 분석용데이터 준비, 텍스트분석, 탐색적분석, 모델링, 모델평가
시스템 구현 : 설계 및 구현, 시스템 테스트 운영
평가 및 전개 : 모델 발전, 평가보고

분석과제 도출 방법(하향식, 상향식)
하향식
절차 : 문제가 확실할 때(문제탐색 > 정의(데이터 및 기법정의) > 해결방안탐색 >타당성 검토)
★문제탐색종류
 비즈니스모델캔버스 : 업무, 제품, 고객, 지원인프라, 규제와 감시
 외부참조모델 기반 문제 탐색 : 벤치마킹, 빠르고 쉬움
 분석유즈케이스 : 문제 해결시 효과, 자금시재예측, 구매최적화
상향식 : 문제의 정의가 자체가 어려울 때
디자인싱킹 : 하향식 상향식 반복적 사용, 상향식의 발산, 하향식의 수렴단계 반복

해결방안 탐색 : 어떤 데이터 또는 분석 시스템 사용할 것인지 검토
분석역량 확보 – 기존 시스템 : 기존시스템 개선 활용
분석역량 확보 – 신규시스템 도입 : 시스템 고도화
분석역량 미확보 – 기존 시스템 : 교육 및 채용
분석역량 미확보 – 신규 도입 : 전문업체

상향식
문제의 정의자체가 어려움, 비지도학습, 발산, 인사이트 도출 후 시행착오 수정

지도학습 : 예측, 분류(이전까지 학습된 데이터 근거)
비지도학습 : 컴퓨터가 알아서 분류, 군집화

분석프로젝트 특징
영역별 관리
5가지 주요 특징 : Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision

- Accuracy : 분석의 활용적인 측면 (모델과 실제 값의 차이)
- Precision : 분석의 안정성 측면 (모델을 반복했을 때의 편차)

분석 마스터플랜 수립
– 분석과제 수행의 선/후행 관계 고려, 우선순위 조정
– 절차 : 분석과제 도출 >우선순위평가 > 우선순위 정련
– 우선순위 고려요소 : 중요도, ROI, 실행요이성

빅데이터 4V
투자비용 요소(Volume, variety, velocity), 비즈니스 효과 요소(value)

분석과제 우선순위 선정 기법
시급성 판단기준 : 전략적 중요도
난이도 : 분석비용과 적용범위 측면
시급성은 반시계 방향, 난이도는 시계방향

거버넌스 체계
거버넌스 : 기업, 기관등에서 규칙 규범 및 행동이 구조화 유지 규제되고 책임지는 방식 및 프로세스
분석거버넌스 : 데이터가 관리 유지 규제에 대한 내부 프로세스
데이터거버넌스 : 데이터가 적시에 필요한 사람에게 제공

분석거버넌스 체계 구성요소

Process ,Organization ,System ,Human Resource ,Data (포함되지 않는 것에 분석 비용 및 예산 나오면 답)

데이터 분석 준비도 : 분석 업무 파악, 인력 및 조직, 분석 기법, 분석 데이터, 분석 문화, IT 인프라
데이터 분석 성숙도 : CMMI 기반, 도입, 활용, 확산, 최적화

데이터 거버넌스 체계요소 : 데이터 표준화, 관리체계, 데이터 저장소관리, 표준화 활동
데이터 거버넌스 구성요소 : 원칙, 조직, 프로세스
데이터 분석 조직 구조 : 집중형(별도 독립, 회사 모든 분서 담당), 기능중심(해당 업무부서에 직접 할당), 분산(분산 인력들이 협업부서에 배치)

빅데이터 거버넌스 특징 : 분석 대상 및 목적 명확히 정의, 데이터 수명주기 관리 방안 수립, 변경사항 관리, 요소별 구분, 지속적 교육, 보안 확보

4과목 데이터 분석

통계량 : 표본의 특성을 나타내는 수치
확률적 표본추출법 : 단순, 계통, 층화(중복 되지 않게), 군집(차이가 없는 여러 집단)
척도종류 : 명목(성별,혈액형), 서열(순위만제공,양적비교 불가), 등간(0존재하지 않음, 양적 비교 가능, 순위사이 간격 동일, 온도계),비율(0존재)

사건 종류
독립 : $P(B|A)=P(B)$, $P(A|B) = P(A)$, $P(A\cap B) = P(A)\cdot P(B)$ 성립
배반 : 교집합이 공집합, $P(A\cap B) = 0$
종속 : $P(A\cap B) = P(A|B) \cdot P(B)$

조건부확률 : $P(A|B) = P(A\cap B) / P(B)$, $P(B|A) = P(B)$, $P(A|B) = P(A)$, $P(A\cap B) = P(A)P(B)$

$P(A)=0.3$, $P(B)=0.4$ 이며 서로 독립일 때 $P(B|A)$? A, B가 독립사건일 때, $P(B|A) = P(B)$ 이다, 0.4!

이산형 확률 분포 : 이항(베루누이 시행 반복), 베르누이(모수 하나, 반복), 기하(베르누이 시행에서 처음성공까지 시도한 회수 분포), 포아송(단위시간에서 사건 몇 번 발생?), 초기
연속형 확률 분포 : 정규, 지수(다음 사건이 일어날 때까지 대기 시간), 연속균일, 카이제곱(분산의 특징을 확률분포), F분포(두지반 분산), t분포(표본을 많이 뽑지 못하는 경우에 대한 대응책)

통계적 추론 분류
모집단 가정여부에 따른 분류 : 모수적(모집단에 특정 분포 가정) 비모수적(모집단에 가정 x)

추론 목적에 따른 추론 분류 : 추정(점, 구간), 가설검정

좋은 추정량 판단기준 : 불편성, 효율성, 비편향성, 일치성

점추정 – 통계량 하나를 구하고 그것으로 모수 추정
점추정량은 적률법, 최대우도법, 최소제곱법으로 구함

구간추정 – 점추정의 정확성 보완 및 신뢰구간으로 추정
신뢰구간은 모수가 포함되는 기대 범위
신뢰수전 : 모수값이 정해져있을 때 신뢰구간 중 모수값 포함하는 신뢰구간이 존재할 확률

귀무가설 : 원래있던 가설, 부정하고자 하는 가설
대립가설 : 연구자가 연구를 통해 입증, 예상하는 주장
1종오류 : 귀무가설 참인데 기각
2종오류 : 귀무가설 거짓인데 채택
유의수준 : 1종오류 최대 허용한계
유의확률 : 1종오류 나올 확률, 귀무가설 지지정도, 판정이 잘못될확률, 유의확률이 유의수준보다 작으면 귀무가설 기각, 대립가설 채택

모수적 추론 : 모집단 특정푼포 가정, 모수에 대해 추론, 자료가 등간, 비율척도
비모수적 추론 : 모집단에 대해 특정 분포 가정안함, 표본수 적음, 명목, 서열척도
모수적 통계 조건 : 정규분포, 분산, 등간, 비율
모수적 검정방법: t-test, anova, 카이, f, t분포

t-test : 평균이 올바른지, 두집단의 평균차이가 있는지 검증
종류 : one sample t, paired t, two sample t

자유도 : 모집단에 정보를 주는 독립적 자료 수, df= n-1, 시험에 df가 100이면 자료수는 101개

데이터 정규성

QQ :시각적으로 정규성 확인. 대각선이면 정규성 만족
히스토그램 : 시각적으로 정규분포 확인
샤피로윌크 : 오차항의 정규분포 검정, 유의확률이 0.05보다 크면 정규성 가정
Kolmogorov smirnov test : KS TEST, 유의확률이 0.05보다 크면 정규성 가정

비모수적 검정

모집단 분포에 제약 하지 않고 검정 실시, 분포 검정, 모수적보다 단순
비모수적 검정종류
명목척도 : 카이스퀘어, mcnemar test, cochran test
서열척도 : Kolmogorov smirnov test, sign test, wilcoxon signed rank test, friedman

test, mann-whitney u test, kruskal-wallis h test (다시 들어가고 sign 들어가면 서열)

카이스퀘어 검정 : 한 개 범주형 변수와 특정 그룹 같은지 적합도 검정, 두 개 범주형 변수 독립적인지 독립성 검정
부호검정 : 두그룹 분포 차이에 대한 가설 검정(분포 나오면 부호)

회귀모형 가정
선형성, 독립성, 정규성, 등분산성, 비상관성

모형이 통계적으로 유의미한가 : F 통계량
회귀계수 유의미한가 : 회귀계수, 유의확률
모형의 설명력 : 결정계수 = SSR / SST
모형이 적합성 : 잔차 통계량

다중공선성 : 모형의 변수들끼리 상관관계 발생, 회귀계수 분산 증가하여 해석 어려움, vif > 10
해결은 설명변수제거, 단계적회귀분석

설명변수선택방법
- 모든 독립변수 조합에 대한 AIC, BIC , AIC BIC는 작은값일수록 좋음
- 후진제거법 : 독립변수 모두 포함 후 하나씩 제거
- 전진제거법 : 절편만 있는 모델에서 출발, 변수 차례로 추가
- 단계별 선택 : 모든 변수 포함 가장 영향력 없는 변수 삭제 또는 좋은 변수 추가

정규화 : 베타값에 패널티 부여
라쏘 회귀 : 변수 서택 가능, L1, 0이 되거나 0 가깝게 됨
릿지 회귀 : 변수 선택 불가능, L2, 0에 가까워질뿐 0은 안됨
엘라스틱넷 : 변수선택가능, 정규화, L1. L2

데이터 스케일링 : 데이터단위 불일치 무제 해결
정규화 : 0~1 변환(min-max)
표준화 : 정규분포 갖도록 변환, 평균은 0, 표준편차는 1

상관계수 : 두변수 관련성 정도, cor.test함수 사용, 피어슨은 선형적 크기만 측정, 스피어만은 비선형 관계도 포함, cor.test 함수 사용
스피어만은 서열척도 비선형 관계
피어슨은 등간척도 비율척도, 두변수간의 선형적 크기만 측정

차원축소 방법
주성분 분석 : 공분산, 상관계수 행렬 사용, 분산 극대화 변수로 축약
주성분 결정 기준 : 누적비율이 70~90%사이 되는 주성분 개수, 고유값이 1보다 큰 성분,

scree-plot 사용하여 정렬값 보여줌

시계열 분석

정상성 : 미래는 과거와 확률적으로 동일
정상성 조건 : 평균과 분산은 시점에 일정, 공분산은 시점에 의존하지 않고 시차에만 의존

정상시계열로 전환하는 방법
평균이 일정하지 않으면 차분사용 계절성 갖으면 계절차분, 분산이 일정하지 않으면 자연로그

AR모형은 백색잡음의 현재값만 사용
MA모형은 과거의 오차들에서 현재 상태 추론, 과거에 백색잡은 추가
ARIMA 모형은 비정상시계열 모형, ARIMA(1,2,3)이면 2번 차분하여 ARMA 모형으로 변경

분해시계열 : 시계열 영향끼치는 요인 분리
분해시계열요인
추세 - 오르거나 내리거나
계절 - 고정된 주기
순환 - 알려지지 않은 주기
불규칙 - 설명할수 없는 요인

데이터 마이닝

5단계(목적정의, 데이터 준비, 데이터가공, 기법적용, 검증)
마이닝 기법
분류 - 기존의 분류
연관 - 교차판매, 공격적 판촉, 마케팅
예측 - 미래에 대한 예측, 장바구니, 의사결정, 신경망
군집 - 그룹화 및 이질성
기술 - 데이터가 가진 특징 및 의미 설명

분류분석

로지스틱 회귀(종속변수가 범주형, 이항변수, 분류하고자 할 때 사용), 절차(odds, logit, sigmoid)

의사결정나무(소집단으로 분류 및 예측 수행, 이해하기 쉬움, 직관적, 불순도 감소, 비모수적 모형, 범주형과 수치형 모두 사용, 오차가큼, 예측이 불안정)
정지규칙 - 불순도 감소량이 적을 때 정지
가지치기 - 노드가 많을수록 과적합하기에 가치치기 사용, 비용함수가 최소 되는 가지 찾도록 학습
지니지수 - 불순도 측정 지표 (1- ∑i=1 k P i2)
엔트로피 지수 - 불순도 측정지표 (- ∑i=1 k P i log 2 P I)

의사결정나무 알고리즘 : CART(지니지수) C5.0(엔트로피지수) CHAID(카이제곱통계량)

앙상블 모형 : 여러개의 분류 모형 종합, 과적합 감소
종류: 보팅, 배깅, 부스팅, 랜덤포레스트
보팅 – 서로 다른 여러개 알고리즘 분류기 사용(하드 보팅– 많은 것 선택, 소프트 보팅 – 확률 선택)
배깅 – 서로 다른 훈련 데이터로 훈련, 같은 알고리즘, 모델 병렬 및 집계
부스팅 – 모델이 순차적 학습, 가중치 부여, XGBoost, Light GBM
랜덤 포레스트 – 배깅에 랜덤, 여러개의 의사결정나무 사용, 과적합 방지

KNN – 근접 데이터 개수로 분류
SVM – 데이터 간격이 최대가 되는 선을 찾아 이를 기준을 분류
경사하강법 – 기울기르 낮은 쪽으로 이동, 기울기 최소값

모형평가

홀드아웃 – 원천 데이터 두 개로 분류 TRAIN, TEST
교차검증 – 데이터 충분하지 않을 때, 클래스 불균형 데이터에 미적합, 반복적, K-fold
부트스트랩 – 한번 이상 train 자료로 사용(복원추출), 63.2% 훈련데이터 사용하는 63.2 부트스트랩

ROC : X=0, Y=1일때 가장 우수, X는 1-특이도, Y는 민감도

군집분석

계층적 군집 : 응집형(최단,최장,평균,중심,WARD연결법), 분리형(다이아나)
– 유사도 판단은 거리기반, 유클리드, 맨하튼, 민코프스키, 마할노비스, 탐색적모형, 군집 후 개체 이동 불가
분할적(비계층) 군집 : 프로토타입(kmeans, k 중앙, k-medoid), 분포(혼합분포), 밀도(중심밀도, 격자밀도)

실루엣계수 : 군집성과 측정, 클러스터 안의 데이터들이 다른 군집과 비교해 비슷한 여부 평가
Dunn Index : 군집과군집사이 거리 최소 / 군집내 데이터 거리 최대값
덴드그램 : 높이에서 일직선 그어 선의 수가 군집수
SOM : 차원축소와 군집화 동시 수행, 비지도, 저차원 변환, 전방패스 이용하여 속도 빠름, 격자구성

연관분석 : 장바구니부석, 패턴규칙 발견 – 알포리알고리즘(발생 빈도기반)
FP Growth : 알포리 단점 보완, fp tree, node, link 구조, 조건 반응

지지도 : $P(A \cap B)$: A와 B가 동시에 포함된 거래 수 / 전체 거래 수
신뢰도 : $P(B|A) = P(A \cap B) / P(A)$: A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수
향상도 : $P(B|A)/P(B) = P(A \cap B) / (P(A)*P(B))$