Hello, Data Science

https://product.kyobobook.co.kr/detail/S000001057683

"데이터가 새로운 과학이다. 빅데이터는 모든 해답을 담고 있다" - 팻 잴싱어

"정량적 데이터 만으로 위대한 마케팅 의사결정이 내려진 경우는 단 한번도 없었다." - 존 스컬리(팹시&애플 CEO)

데이터를 통해서 당면 문제에 대한 해답 찾을 수 있음

- 자신에게 절실한 문제에서 출발한다.
- 데이터를 의사 결정과 실천으로 직결시킨다.
- 데이터 수집 및 활용을 일상화한다.

빅데이터는 비싸고, 느리고, 복잡하고, 어렵기에 스몰데이터로 시작하라

기계학습이나 딥러닝보다는 단순한 도구와 기술로 시작하라

" 데이터 과학자는 삶과 업무에서 끊임없이 데이터를 통해 가치를 창조하는 방법을 찾아내고, 이를 자신이 직접, 혹은 다른 사람을 통해 실천에 옮기는 사람이다."

데이터 과학자는 데이터가 주는 가능성 인지하고 탐혐하고 실제로 간단한 데이터 문제를 풀 능력을 갖추었으며 필요한 경우 전문가와의 협업을 통해 문제를 해결할 수 있는 사람

- 현상에서 데이터 발견 > 이를 통해 현상 제대로 이해 > 현상 개선하고자 함

사망률 개선한 나이팅게일

콜레라 이긴 존 스노우

플랭클린 다이어리

데이터 습관

- 수집 마인드
- 분석 마인드
- 실천 마인드
- 공유 마인드

린 스타트업

: 아이디어 - 만들기 - 제품 - 측정하기 - 데이터 - 학습하기

관련 지식과 기술 읽히기

- 해당 분야의 전문성
- 컴퓨터 응용 및 프로그램
- 통계 및 기계학습

"보조장치 없는 인간의 지적 능력은 미약하기 짝이 없다. 종이와 펜, 컴퓨터와 같은 인지보조 장치의 사용으로 인간의 기억과 사고력은 극적으로 향상된다." - 도널드 노먼

데이터 과학을 위한 도구

- 스프레드 시트(엑셀), 관계형 데이터베이스, R, 파이썬, 클라우드, 커스텀 코드

데이터 준비 - 하둡, 데이터베이스

탐색적 데이터 분석 - 엑셀, R

통계적 추론 / 예측 - 파이썬, R

해결책 구현 - 파이썬, 커스텀 코드

결과 소통 - 엑셀

엑셀의 장점

- 코드가 아닌 데이터 중심
- 올인원 솔루션
- 쉬운 결과물 공유

엑셀에서 R로 주사용이 바뀐 이유

- 메뉴에서 함수로
- 작업의 효율성
- 분석에서 예측

"애자일 운동을 방법론을 없애자는 이야기가 아니다. 사실 우리는 '방법론'이라는 말의 신뢰를 회복시키고 싶다." - 짐 하이스미스

"만약 나에게 문제 해결을 위해 한시간이 주어진다면, 나는 55분 동안 문제에 대해 생각하고 5분 동안 해결책에 대해 생각하겠다." - 알버트 아인슈타인

데이터 문제 해결 단계

- 주어진 문제 명확히 정의
- 문제 해결에 필요한 데이터 수집 및 추출
- 데이터 분석에 적합한 형태로 가공
- 가공된 데이터를 분석하여 해결책 유도
- 해결책을 여러 방식으로 구현
- 관계자에게 결과를 적절한 형태로 소통

데이터 분석 단계

- 탐색적 데이터 분석

- 통계적 추론
- 기계학습

데이터 문제 정의

- 문제의 목표
- 문제의 범위
- 문제 해결의 성공과 실패의 척도
- 문제 해결에 있어서 제약조건

http://www.netflixprize.com/assets/rules.pdf

"실험이 다 끝난 후에 통계학자를 부르는 것은, 의사에게 시체 부검을 부탁하는 것과 같다. 아마 통계학자는 왜 실험이 실패했는지를 알려줄 수 있을 테니까." - 로널드 피셔

수집 방법 결정

- 자동 수집, 수동 수집

추상적인 대상을 구체적인 측정의 대상으로 바꾸기

- 만약 중요한 일이라면 어떤 식으로든 관찰할 수 있다
- 관찰 가능한 일은 수치 혹은 범위로 표현할 수 있다
- 수치 혹은 범위로 표현될 수 있는 일은 측정할 수 있다

수집 환경 결정하기

: 표본의 대표성

관찰형 연구

- 무작위 디자인 (A/B 테스트)
- 블록 디자인

데이터 품질 점검

- 완전성, 정확성, 일관성

메타데이터

"대단한 성취의 이면에는 대부분 사소해 보이는 준비 과정이 필효다아." - 로버트 슐러

데이터 선택, 추가, 집계

- 항목 추출, 속성 선택

"진정한 발견은 새로운 장소를 찾는 것이 아니라, 새로운 관점을 갖는 것이다." - 마르셀 프루스트

탐색적 분석

- 데이터 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해하고 데이터에 잠재된 문제 발견 가능

탐색적 데이터 분석의 주된 수단

- 원본 데이터 관찰
- 요약 통계값 사용
- 적절한 시각화

데이터 개관하기 > 속성 분석하기 > 속성 간 관계 분석하기

데이터형의 조합에 따라 주로 사용되는 요약 통계 및 시각화 방법

- 카테고리와 카테고리 : 요약 통계 = 교차 테이블, 시각화 = 모자이크 플롯
- 수치와 수치 : 요약 통계 = 상관 계수, 시각화 = 스케터 플롯
- 카테고리와 수치 : 요약 통계 = 카테고리 별 통계값, 시각화 = 박스 플롯

관점을 갖되 편견은 금물

"통계로 거짓말을 하기는 쉽지만, 진실을 말하기는 어렵다." - 안드레아 덩켈스

모수적 방법 : 중심극한정리

비모수적 방법 : 표본 재추출법

통계적 추론의 유형

- 신뢰구간 구하기
- 가설 검정하기

"어떤 가설의 유효성에 대한 유일한 검정 방법은 예측값과 실제 결과를 비교하는 것이다." - 밀턴 프리드먼

기계학습

- 지도 학습 / 자율 학습

"데이터는 어떤 일이 일어나는지를 알려준다. 스토리는 왜 그것이 중요한지를 말해준다." - 저크 매킨리

새부사항 및 가술적인 내용은 최대한 자제

내용과 형식을 독자 중심에서 선별하고 재구성

메시지를 중립적인 관점에서 전달

"공개 데이터는 과학 및 사회 전체를 발전시키는 데 매우 중요하다. 연구 데이터는 다른 용도로 사용되면 그 가치가 늘어난다. 연구의 투명성은 결과에 대한 대중의 신뢰를 위해서도 중요하다. 연구 데이터 관리는 연구자, 발주자, 연구기관, 도서관 그리고 공공의 책임이다."

캐클 참가자의 성공 비결

- 문제와 데이터에서 단서 찾기
- 빨리 시작해서 반복 개선
- 다른 사람들을 통해 학습

검색엔진을 만들기 위한 단계와 필요한 데이터 과학의 기술

- 데이터 수집 (빅데이터 처리) . 데이터 색인 (빅데이터 저장) > 검색결과 생성 (예측 모델링) > 검색결과 평가 (통계적 추론)

검색 엔진 데이터 과학자로서 업무 프로세스

- 고객의 요구사항 청취 > 평가 데이터 수집 > 평가 지표 디자인 > 평가 지표 검증 > 고객에게 데이터 및 지표 전달

도메인 전문가 V.S. 데이터 전문가