데이터 분석 - 도입

시작 : 로또 예측은 가능할까?

옛날 이야기 몇 가지

사우스 시 컴퍼니(south sea company) 사건

사건 개요

- 남해회사 거품 사태라고도 부름
- 사우스 시 컴퍼니 - 영국 정부가 1711년 남아메리카 무역 독점권을 부여하며 설립된 회사
- 영국의 재정 상황을 개선하기 위해 설립된 회사로, 특히 노예 무역에 관한 특권을 보유
- 국가 부채를 주식으로 전환해서 발행
- 주식은 인기가 폭발해 1720년 8월 주식 가격이 700 파운드를 돌파
- 거품이 꺼지면서 1720년 12월 150파운드까지 추락

사건의 원인

- 가치를 평가할 수 있는 데이터 부재 → 기업 가치 평가와 투자 분석 방법론의 부재
    - “영국 정부가 보장하는 주식이다!” → 엄청난 인기를 끌어 많은 사람들이 할부로 주식을 매입
        - 1720년 8월 700파운드
        - 실제 가치나 수익성 분석 없이 주식 가격 상승이 투자자들을 재유인하는 요인
    - 정부 및 회사는 이익을 과장했고, 실제로는 무역이 거의 이루어지지 않음
        - 그 결과 1720년 12월 150파운드까지 추락

사건의 결과

- “거품경제”라는 용어가 이 사건에서 만들어짐
- 영국 정부는 주식회사 설립에 대한 규제를 도입하고 사우스 시 컴퍼니 경영진들의 재산을 몰수, 투자자들에게 벌금 부과

- 정부는 신뢰를 잃고 많은 사람들이 파산

교훈

- 과도한 기대와 과장된 정보에 기댄 불합리한 판단은 실패할 수 밖에 없음을 보여줌
- 객관적인 정보, 정확한 데이터가 필요

NASA Mars Climate Orbiter 탐사선 사건

사건 개요

- 1998년 나사는 화성 기후 연구를 위한 탐사선 Mars Climate Orbiter 탐사선 발사
- 1999년 9월 화성 대기권 진입 과정에서 궤도 비행 실패

사건의 원인

- 분석가들의 일관성 문제
  - 일부 팀은 국제 표준 단위계를 사용 (SI 단위계)
  - 일부 팀은 미국 관습 단위계를 사용 (파운드 피트 등)
- 이로 인해 탐사선 궤도 계산이 틀림

사건의 결과

- 탐사선은 예상보다 훨씬 가까운 거리에서 화성 대기에 진입
- 탐사선은 파괴되었음
- 손실 규모는 약 1억 2500만 달러 (약 1,625억원)

교훈

- 일관성 있는 분석 및 데이터 단위에 대한 중요성
- 복잡한 시스템의 분석에서 작은 실수가 얼마나 큰 영향을 끼치는지를 보여줌

2016년 미국 대선 예측 실패 사건

사건 개요

- 대다수의 여론조사 및 데이터 분석가들은 2016년 미국 대선은 힐러리 클린턴 후보가 도널드 트럼프 후보를 이길 것으로 예측

- 최종 결과는 도널드 트럼프 후보가 당선

사건의 원인

- 잘못된 데이터 품질
    - 많은 여론 조사가 특정 지역, 인구 그룹의 의견을 충분히 반영하지 못함 → 특히 백인 유권자들의 표본이 적게 반영되었고, 그들의 투표가 여론조사와 다르게 이루어짐
    - 데이터 정제 실수 : 수집된 데이터의 편향을 제거하는 과정에 중요한 변수를 고려하지 못해 제거되지 못한 편향성이 여론조사 결과에 그대로 반영

사건의 결과

- 여론조사에 대한 신뢰성 제고
    - 갤럽 스타일의 여론 조사를 극복하기 위한 다양한 여론조사 및 분석 방법이 제시
- 스타 데이터 분석가의 탄생 (네이트 실버)

교훈

- 방법론에 대한 지속적인 고민이 필요
- 데이터에 분석가의 편향을 더하면 안된다는 사례를 보여줌

2019년 영국 코로나 19 검사 결과 집계 오류

사건 개요

- 코로나19 판데믹 시즌, 영국에서 검사 결과 집계 과정에서 수만 건의 검사 결과가 누락

사건의 원인

- 엑셀 스프레드시트 파일의 크기 제한

사건의 결과

- 수만 건의 검사 결과 및 수천 명의 확진자가 통계에서 누락
- 판데믹 대응에 큰 차질이 발생

교훈

- 데이터 관리 도구에 대한 이해 및 적절한 도구 선택의 중요성

Moneyball

개요

- MLB 야구팀 오클랜드 애슬레틱스는 2000년대 초반 재정 위기 상황에서 팀 성적도 좋지 않은 약체 팀으로 분류
- 비스타 출신인 빌리 빈(Billy Beane)을 단장으로 임명했고, 빌리 빈은 새로운 데이터 분석 기법을 활용

원인

- 전통적인 선수 스카우팅 방법 대신 세이버매트릭스를 도입해 선수를 스카우팅
    - 타율, 홈런과 같은 전통적인 지표보다 출루율 등의 주목받지 않았던 지표가 팀 승리에 더 큰 영향을 미친다는 점을 분석
    - 출루율이 높으나 전통 지표로 좋은 평가를 받지 못해 연봉이 낮은 선수를 저렴한 비용으로 영입

결과

- 2002년 시즌, 아메리칸 리그 서부 지구에서 103승 59패 성적으로 1위를 차지
- 20연승의 대기록을 세우고, 포스트시즌 진출
- 오늘날 세이버매트릭스는 스카우팅 전략의 핵심 분석 기법으로 활용 중

교훈

- 합리적 데이터 분석의 중요성

구글의 HR 전략

개요

- 기존 채용 방식은 면접자의 직감이 포함된 평가 및 학력 중심의 채용
- 구글이 데이터 분석을 기반으로 한 인재 채용 프로세스를 도입

원인

- 구글의 기존 직원들의 인사 데이터 및 성과를 수집, 분석
- 면접 질문, 학력, 경력 등 입사 때 고려한 사항과 성과 간의 상관관계를 분석

- 분석 결과 학력 또는 특정 인터뷰 질문이 성과와 큰 연관이 없음을 밝혀냈고, 대신 사용할 수 있는 영향력 있는 변수를 도출

결과

- 핵심 인재 확보를 통한 회사의 실적 개선

교훈

- 맞다고 생각하는 것도 데이터 분석을 통한 검증이 필요

기타 국방 관련 데이터 분석 사례

펜타곤 패턴 분석 프로젝트

- 테러 활동 방지를 위한 패턴 분석
- 방대한 데이터를 수집하고, 이를 분석해 테러리스트들이 인터넷, 통신 등을 사용하는 방식의 패턴을 식별
- AI, 머신러닝, 데이터 분석 기법등을 총 망라해서 활용
- 테러리스트들의 이동, 자금 흐름, 커뮤니케이션 패턴을 찾아낼 수 있었고, 이를 재분석해 잠재적 테러 활동을 사전에 차단하는데 기여

AEO 프로젝트 (DARPA)

- 미국 국방고등연구계획국의 Adaptive Execution Office 프로젝트
- 데이터 분석 기벌을 사용, 실시간 전투 환경에서 적의 움직임과 전략을 분석하고 예측하는 시스템을 구축
- 작전 중 발생 가능한 위험을 사전에 예측하고 실시간으로 최적의 대응 전략을 제시

아프가니스탄 반군 분석 실패

- 미군은 아프가니스탄에서 반군 세력의 위치와 활동을 예측하는데 데이터 분석을 활용
- 반군의 실제 동선 및 공격 계획 예측에 실패
- 잘못된 데이터 품질 및 과도한 기술 의존으로 인한 실패
    - 잘못된 측정 지표 - 초기 적군 사상자 수, 개발 프로젝트의 완료 건수를 성과 지표로 삼았으나, 추후 분석을 통해 이 지표들은 결과와 큰 상관이 없는 데이터이며, 분석 단계에서 성과를 과대 평가하게 만드는 악영향을 끼침

    - 과도한 기술 의존 - 미국이 사용한 합리적 데이터 분석 방법은 오히려 현지 아프가니스탄 사회의 복잡한 부족 구조 및 전통적 권력 기반을 제대로 반영하지 못함. 이로 인해 현지 세력 간 역학을 이해하지 못했고, 반군간 전투에서 효과적인 전략을 세우는데 실패
- 아프가니스탄의 장기간 군사 작전은 효과를 거두지 못하고, 대표적인 미국의 군사 전략 실패 사례가 됨

1800년대 중반 런던 콜레라 유행 사건

개요

- 1854년 런던에 콜레라가 빠른 속도로 전파
- 당시 사람들은 공기를 통해 질병이 전염된다고 믿음
- 당시 영국에서 활동하던 마취 전문의 존 스노우(John Snow)는 런던의 지도와 런던의 수도 시스템 정보를 대조, 콜레라가 특정 수도 시스템을 중심으로 집중 발생된다는 사실을 깨달음
- 해당 수도를 차단한 후 콜레라 발생률이 급격히 줄어듬. 이를 통해 콜레라가 물을 통해 전파된다는 사실이 알려짐

세 가지 질문

- 존 스노우는 훌륭한 데이터 분석가였을까?
- 해당 수도를 차단한 후 콜레라 발생률이 급격히 줄어든 것을 기준으로 존 스노우의 주장인 “콜레라가 물을 통해 전파된다.”는 사실이 입증되었을까?
- 컴퓨터가 없던 시절 존 스노우의 데이터 분석과 오늘날의 데이터 분석의 차이는 무엇일까?