21세기의 원유, 또는 기업 생존의 전략이라고 비유될 만큼 데이터의 가치가 중요시 되고 있지만, 아직도 빅데이터에 대한 올바른 이해와 실무에 대한 어려움이 있는 것도 사실입니다. 최근에는 인공지능이라는 키워드가 많이 회자되고 있는데, 이또한 빅데이터에 기반하고 있다는 사실을 잘 이해 못하는 경우도 많습니다.

데이터의 가치를 재인식하고 데이터 분석에 대한 이해를 높이는 목적

기업의 생존과 기업의 미래 가치에서 데이터가 갖는 의미

빅데이터와 4차 산업혁명의 이해 관계

빅데이터의 이해

아직도 빅데이터를 사이즈로만 구분하려고 하거나, 통계처리 한 과정, 단순한 데이터 분석 도구로 이해하는 부분. 빅데이터와 데이터 분석에 대한 정확한 이해가 필요한 시점.

의미있는 데이터 분석을 위해서는 데이터 전문가, 즉 데이터 사이언티스트가 그 중심에 있고 그들의 역량을 높이는게 무엇보다 중요하다. 데이터 분석가의 RNR

개인적으로 Citizen data scientist

Auto ML

데이터 분석 방법론

분석 방법론의 framework를 이해한다는 것은 대단히 중요하다

다양한 도메인에서 활용되고 있다

프로젝트에서도 관련된 모든 관계자들이 이 방법론이라는 걸 공유할 필요가 있다.

분석기획 및 데이터 수집 단계

기획은 모든 분야에서 중요. 머신러닝, 인공지능 이런 것들을 아트로 인식하는 경우가 있다. 빅데이터에서의 기획은 보다 구체적일 필요가 있다

또한 데이터 준비 단계도 중요하다

준비가 성공되지 않는다고 한다면 결코 빅데이터 프로젝트는 성공할 수 없다는 Garbage in garbage out이라는 절대 진리.

데이터가 준비되지 않은 빅데이터 프로젝트는 성공을 확신할 수 없다.

분석은 어떻게 하는지, 시각화가 얼마나 중요한지,

이제 데이터는 기업의 부산물이 아닌 가치 척도가 되고 생존의 방향타 역할을 할 수 있다.

Edyn 작은 스타트업

정원에 센서를 설치하고, 토양의 온도, 습도같은 정보를 측정해서 이 데이터를 스마트폰 앱이나 PC로 전달하는 아주 전형적인 IoT 디바이스를 판매하는 edyn의 비즈니스 모델은 너무 평범합니다. 사용된 기술도 뭐 wifi 통신, iot 센서, 모바일 엡, 태양열 배터리 정도? 이 기업의 기술적 장벽은 그리 높지 못하고, 비즈니스 모델 또한 그렇게 특별해 보이지는 않다. 그런데 만약 이 비즈니스 모델을 다음과 같이 바꾸면 이야기가 달라질 수 있다. 디바이스를 전략적인 가격으로, 한 마디로 조금 낮춰서 판매한다. IoT 디바이스에서 측정된 토양 정보를 edyn이 운영하는 혹은, 클라우드 시스템에 수년간 모은다. 그리고 이 데이터를 활용할 수 있는 서비스를 개발한다면 어떻게 될지 모른다. 상상하기 어렵다. 만약 단순히 앱과 센서를 공급하는 제품판매가 아닌, 상상하기 어려운 부가가치가 높은 데이터 서비스 회사가 될 수 있다는 가능성을 충분히 가지고 있는 기업의 비즈니스 모델.

Nest & Google

직원이 300명 정도 되는 네스트 라는 역시 작은 스타트업 기업을 3조 2천억 달러에 M&A한다(인수합병) 한국 국가 예산이 약 500조. 한 1퍼센트 가까운 엄청난 금액. 유튜브 인수금액 2배에 가까이 되는. 네스트의 온도 조절장치와 화재감지기 기술은 클라우드 시스템, 머신러닝 관련 알고리즘. 그렇게 특이해 보이지 않은 디바이스.

당시 국내외 언론에서는 향후 큰 시장이 형성될 것으로 예상되는 가정의 IoT 허브와 관련된 기술 확보로 해석.

그런데 실내 온도 조절 장치를 iot 허브로 생각하기에는 경쟁자가 많다. Ai speaker, 냉장고, 티비. 등등.

다른 가능성. 이 온도 조절 장치가 설치되어 있는 이 모든 가정에 실내 온도에 대한 변화 이런 걸 데이터로 확보. 이를 응용할 서비스를 기획하기 위한 M&A. 데이터의 가치를 중요하게 생각하고 있는 구글의 기업적 전략과 크게 다르지 않다.

2017말 2018초 신문에 크게 기재

네이버에서 리멤버라고 하는 작은 회사를 M&A

명함관리 기업. 100억 투자.

네이버의 투자 목적. 네이버는 리멤버가 가지고 있는 명함 데이터에 있다. 명함 자체의 속성 정보도 중요하지만 명함과 명함을 주고 받았던 네트워크, 커넥션 정보. 노드를 연결하는 방향성을 가지고 있는 엣지라는 정보.

Netatmo

측정된 정보를 보여주는 것이 IT에서 특별하지는 않다.

하지만 온도를 예측하고

10년전 글로벌 시가총액 상위 기업은

액손, GE, 공상은행 같이 금융과 제조업 위주

지금의 10대 기업을 살펴보면, 구글 애플, 페이스북, 텐센트, 아마존 등

모두 플랫폼과 컨텐츠 즉 데이터를 가지고 있는 기업들이 그 중심에 있다.

기업의 가치와 순위가 그 기업이 가지고 있는 데이터의 질과 양에 있다고 한다면

이제는 제조업도 제품을 생산하고 판매하는 것으로만 사업 영역을 한정한다면 결코 글로벌 기업이 될 수 없고 제품과 데이터를 같이 서비스할 수 있는 기업만이 생존할 수 있따는 걸 이해할 필요가 있다.

서버타이제이션, 프로덕타이제이션 컨셉 또한 프로덕트와 데이터가 하나의 서비스.

고품질 저가격에 제품을 만드는데 집중했고 부산물로 나오는 데이터의 가치를 중요시 여기지 않음.

하지만 이제 그 두개가 하나의 서비스라는 사실을 인지할 필요가 있다.

IT 시대 넘어가고 DT시대(data technology)

2. 빅데이터와 4차 산업혁명

1 ~ 4 notion

**5. 빅데이터 분석 방법론**

빅데이터 분석이 회자되고 업무에 적용되기 시작한지 10여년이 되었는데, 아직도 빅데이터 분석은 전문가의 영역으로 인식되고 있다. 심지어 데이터 사이언스를 이야기하면서 데이터 분석을 아트에 비유하기도 한다. 아트로 남아있어서는 진정한 사이언스가 될 수 없으며 엔지니어링은 더욱더 먼 길이다. 표준화 된 프로세스를 만들기 위한 빅데이터 분석 방법론의 개발과 활용이 필요하다.

빅데이터 분석의 기술적 난이도와 데이터 분석 전문가 부족이 원인이지만, 빅데이터를 만능의 해결도구로 또는 기존의 데이터 분석의 확장으로 인식.

방법론(Methodology) 개개인의 역량과 경험에 의존하지 않고 누가 수행하던 “일정수준의 질과 양”이 보장될 수 있는 체계(시스템)

우리는 어떤 업무를 새로 추진할 때, 절차와 방법을 찾게 된다. 개인의 역량과 경험에 의존하지 않고 조직에서 그 누가 업무를 진행해도 일정 수준의 품질을 보장받을 수 있는 체계를 방법론이라고 한다. 따라서 방법론을 통해 업무 수행 결과에 대한 최소한의 품질을 보장받을 수 있다. 기업에선 방법론 없이 운 좋게 또는 특정한 전문가에 의해 프로젝트가 성공할 수도 있다. 그러나 방법론이 부재된 상태에서 이런 성공은 일회성이 높고, 성취도가 낮은 단계라고 판단할 수 있다. 빅데이터 분석에서도 방법론을 활용하는 것은 빅데이터 분석 모델 품질에 많은 영향을 준다. 방법론은 절차와 방법, 도구와 기법, 산출물 이런 것들로 구성되어 있는 규정이다.

방법론을 어떤 과정을 거쳐 만들어지나?

A screenshot of a cell phone

Description automatically generated

일본의 노나카 교수는, 방법론은 4단계의 순환을 통해 발전한다고 얘기한다.

개인이 가지고 있는 암묵지를 조직에 공유함으로써 암묵적인 지식을 갖게 되고, 이를 다시 개인별 문서화를 통해 형식지가 되고, 이러한 형식지들이 결합해서 형식적인 지식, 방법론이 만들어지면 방법론은 다시 개인별로 내재화 되며 암묵화되는 선순환 과정에 들어간다는.

일반적인 방법론에 대한 이론들도 있지만 빅데이터 분석에 방법론도 노나카 교수의 사이클이 의미하는 것으로써 만들어지고 활용될 수 있다면 충분히 의미있게 방법론을 만들어갈 수 있다.

A screenshot of a cell phone

Description automatically generated

빅데이터 분석 방법론이 만들어진다면 위의 그림과 같이 단계, 태스크, 스텝으로 구성된다. 마지막 단계에서는(스텝) 거기에서는 input, process, output 그 다음 필요한 템플렛을 제공하는 형태를 가질 필요가 있다.

이러한 방법론의 모델은 여러가지가 있다. 가장 대표적인 모델이 폭포수 모델이다. 복잡도가 낮고 사례가 많은 장점을 가지고 있으나, 단계별로 완성을 해야 하는 수행의 문제점을 가지고 있다. 첫 번째 단계를 완성하지 않은 상태에서 다음 단계로 넘어가면 여러가지 이슈가 생긴다. 한 단계를 완료시키고 다은 단계로 넘어가야 하는 단계별 접근 방법인데, 한 단계를 완전히 완료시키고 다음 단계로 넘어간다는 것도 어렵다. 가장 심플한 모델임에도, 내재적인 문제를 가지고 있는 모형이 폭포수 모델이다.

다른 모델은 반복 구조적인 나선형과 프로토타입이다. 반복할 때마다 스파이크를 만들고 단계별 완성물이라고 할 수 있다.. 스파이크를 만들고 완성도를 높여가는 방법으로써 요구사항 충족에는 큰 장점이 있으나, 철저한 관리가 되지 않는다고 한다면 이 프로젝트의 목적도가 상당히 올라가고, 이터레이션 별 완성도가 정확히 측정되지 않는다고 한다면 완성도의 부담감이 뒤로 미루어지는 단점을 가지고 있다.

빅데이터 분석 방법론의 모델은 프로토타입이 좋다, 스파이럴이 좋다, 워터폴이 좋다 선택보다는 조직의 문화성, 내용, 경험 정도 이런 것들을 고류해서 가장 의미있는 모형을, 모델을 선택하는 것이 좋다. 처음에 시도한다 하면 폭포수 모델이 이해도가 높을 수 있기 때문에 처음 시도해볼 수 있는 모들이라 생각된다.

KDD(Knowledge Discovery in Database)

Selection: 분석을 위한 Data Set을 편성하거나 Sampling 또는 필요한 변수를 선택하는 과정

Pre processing: 일관성 있는 데이터 분석을 위하여 데이터를 정재하거나 선처리하는 과정

Transformation: 데이터의 차원 축소하거나 파생 데이터 생성하여 분석용 Data Set 생성

Data mining: 다양한 분석기법을 사용하여 데이터의 패턴을 찾고 모델링화

Interpretation/Evaluation: 분석된 데이터 패턴 및 모델을 해석하거나 평가

데이터 분석을 위한 방법론

KDD 방법론. 다섯 단계로 구성되어 있는 데이터 분석 방법론. 빅데이터 분석을 위해 만들어진 방법론은 아니다. 그동안 데이터 또는 데이터 베이스에서 탐색을 통해 패턴을 찾아내는 데이터 방법론으로서 많이 활용되어 왔다.

A screenshot of a cell phone

Description automatically generated

SAS에서 사용하는 SEMMA라는 방법론

A screenshot of a cell phone

Description automatically generated

CRISP-DM 방법론

데이터 분석에서 가장 많이 사용되고 있는 방법론

Business Understanding

* 비즈니스 관점에서 프로젝트의 목적과 요구사항을 이해하기 위한 단계
* 도메인 지식을 데이터 분석을 위한 문제정의로 변경
* 초기 프로젝트 계획을 수립

세부 테스크

* 업무 목적 파악
* 상황 파악
* 데이터 마이닝 목표 설정
* 프로젝트 계획 수립

분석 목적을 명확하게 인식하고 비즈니스 도메인 대한 충분한 이해 정도는 다음 단계를 수행하기 위한 가장 근본적인 절차이다. 데이터 언더스탠딩 단계를 수행하다가도 명확한 비즈니스 언더스탠딩이 필요하다고 판단된다면, 이 단계를 반복함으로써 도메인의 이해 완벽함을 추구할 필요가 있다.

Data Understanding

* 분석을 위한 데이터를 수집하고 데이터 속성을 이해하기 위한 과정
* 데이터 품질에 대한 문제점을 식별
* 데이터에 숨겨져 있는 Insight를 발견

세부 테스크

* 초기 데이터 수집
* 데이터 기술 분석
* 데이터 탐색
* 데이터 품질 확인

분석 모델링에 앞서서 데이터를 이해하는 차원에서 데이터 탐색을 수행하고 충분한 비즈니스 언더스탠딩이 만약 부족하다고 하면, 다시 전 단계로 돌아가는 과정으로 구성되어있다. 어떻게 보면 CRISP의 전체 과정 중에서 가장 핵심적인 요소. 만약 데이터 언더스탠딩에서 필요한 분석 데이터셋에 대한 정확한 이해, 혹은 데이터 확보가 되지 않는다고 하면 다음 단계로 넘어가는 것은 무의미할 수 있다. 다시 처음 단계로 돌아가는 것은 대단히 의미가 있다. 가능성을 가지고 다음 단계로 넘어간다, 데이터 확보될 것이다 전제로 넘어간다면 그만큼 위험한 프로젝트가 될 수 있다.

Data Preparation

* 분석을 위하여 수집된 데이터에서 분석툴에 적합한 데이터셋을 편성
* 데이터 준비 과정에 많은 시간이 소요
* 데이터 준비는 모델링과 연계되어 반복 수행

세부 테스크

* 분석용 데이터셋 선택
* 데이터 정제
* 분석용 데이터셋 편성
* 데이터 통합
* 데이터 Format

수집된 데이터에서 분석툴에 적합한 데이터셋을 편성하는 과정으로써 많은 시간이 소요될 수 있다. 따라서 적정한 도구의 사용과, 데이터 편성에 대한 기술적인 사전 지식을 갖출 필요가 있다. 또한 이 단계는 모델링 단계와 계속 반복해서 수행되어야 한다.

데이터 분석 과정에서 비정형 분석이 포함된다고 하면, 더욱더 이 데이터 준비 과정에 대단히 많은 시간이 소요된다. 빅데이터 프로젝트 수행 계획서, 일정표에 반영을 해야 뒷 단계에서 받는 부담감을 줄일 수 있다.

Modeling

* 다양한 모델링 기법과 알고리즘을 선택하고 모델을 최적화
* 모델링 과정에 데이터셋이 추가로 필요한 경우 Data Preparation 단계를 반복
* 모델은 테스트 데이터셋으로 평가하여 Overfitting 문제 발견

세부 테스크

* 모델링 기법 선택
* 모델 테스트 계획 설계
* 모델 작성
* 모델 평가

핵심 단계. 준비된 훈련용 데이터셋으로 모델 생성하고 과적합 문제를 객관적으로 평가하기 위해 테스트용 셋으로 평가를 최종적으로 하게된다. 데이터 셋이 분석에 충분하지 못하다고 판단해서 준비된 테스트용 데이터셋을 훈련에 넣는 이런 우를 범해서는 절대 안 된다. 이런 문제가 사전에 발견될 때에는 cross validation이라던지 다른 방법을 시도해야 한다.

Evaluation

* 모델링을 통한 최적의 모델을 찾고 모델이 프로젝트의 목적에 부합되는 평가
* 데이터 마이닝의 결과를 수용할 것인지 최종 판단

세부 테스크

* 분석결과 평가
* 모델링 과정 평가
* 모델 적용성 평가

이 단계는 모델링 과정을 수행하여 최적화된 모델을 만든 후에 그 모델이 프로젝트의 목적에 부합되는지 평가하는 단계이다. 평가를 통해 합목적이지 못하다고 한다면 즉시 CRISP의 첫 단계인 비즈니스 언더스탠딩으로 돌아가는 것이 이 단계의 해심이라고 할 수 있다. 빅데이터 분석은 사용자의 요구사항을 찾아 시스템을 구현하는 일반적인 SI성 프로젝트와는 다르기 때문이다. 사용자의 요구사항은 있지만, 주어진 데이터가 충분하지 못하거나, 데이터 품질등의 문제로 적용한 알고리즘이 올바른 답을 주지 못하는 등의 문제는 충분히 발생한다. 이러한 문제를 해결하고 무리한 모델링을 수행하는 과적합은 더 큰 문제를 만든다. 따라서 평가를 통해 우리가 목적하는 바가 발생되지 않는다면, 다시 첫 단계로 돌아가서 비즈니스 언더스탠딩, 데이터 준비, 모델링 과정을 거쳐 다시 평가가 필요하다. 이러한 과정 반복에 프로젝트 비용 부분을 간과할 순 없지만, 그런 것 까지 전제한 프로젝트 계획을 처음부터 기획하는 것은 대단히 중요하다.

Deployment

* 모델링과 평가를 통해 완성된 모델을 실 운영환경에 적용
* 모델의 모니터링 및 유지보수 방안 마련
* 프로젝트 종료 프로세스 진행

세부 테스크

* 전개 계획 수립
* 모니터링과 유지보수 계획 수립
* 프로젝트 종료보고서 작성
* 프로젝트 리뷰

CRISP-DM이 빅데이터 분석 방법론을 위해 만들어진 것은 아니지만, 의미있는 레퍼런스 일 수 있다.

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

Planning: 분석 과제 발굴과 과제 목적, 목표를 설정해서 프로젝트 계획을 수립하는 단계. 분석 계획이 수립되면 데이터 준비 단계를 수행하게 된다.

데이터 준비 단계에서는 가장 많은 리소스가 투입 될 수 있다.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated