LS 빅데이터 스쿨 3기
데이터 분석 실습

우리가 6주동안 함께 키워갈 능력

1. 데이터 수집 및 대시보드 시각화 능력
2. 데이터 기반 인사이트 획득 능력 ex) '특정 변수를 조정하면 불량률을 낮출 수 있다' 등
3. 비즈니스 솔루션 제안 기반 분석 및 결과 보고서 작성 능력

채용시장에서의 요구 조건 1 - 시각화 대시보드

주요업무

- 비즈니스 관점에서 데이터를 주도적으로 분석하고, 문제 해결 프레임워크를 적용하여 인사이트를 도출합니다.
- 주요 지표를 정의하고, 목표 달성을 위한 A/B 테스트를 설계하고 실행합니다.
- 합리적인 의사결정이 이루어질 수 있도록, 대시보드를 개발 및 모니터링합니다.
- 구성원이 데이터를 활용할 수 있도록 적극적으로 데이터셋을 구축하고 시각화합니다.

자격요건

- SQL을 사용해 직접 데이터 추출/정제/분석할 수 있는 분
- Excel을 활용해서 추출된 데이터를 자유롭게 가공할 수 있는분
- BI Tool(Looker Studio 등)을 활용한 데이터 시각화가 가능하신 분
- 대규모의 데이터셋을 핸들링해 본 경험이 있으신 분
- 통계학과 기본적인 머신러닝에 대한 지식이 있으신 분

주요업무

- 데이터 분석 결과 해석 및 보고서 작성 (데이터 대시보드 (Tableau/Redash 등) 운영 및 제작)
- 비모소프트가 서비스 중인 어플리케이션 유저들의 행태 DATA를 정제/가공/분석 및 인사이트 도출
- Product, Marketing, Development, HR, Finance 등 다양한 팀과의 협업을 통해 비즈니스 문제를 정의하고, 이를 해결하기 위한 가설 수립
- 가설 수립 및 검증을 통해 비즈니스 로직 개선 방안 제안
- 데이터를 수집, 분석 및 시각화하여 사용자 행동을 이해하고 서비스가 사용자의 요구를 충족하고 있는지 확인
- 주기적으로 데이터 분석 결과를 리포트하고 이를 협업 부서에 전달

자격요건

- 데이터 분석 관련 실무 경력 2년 이상이신 분
- 데이터 분석, 머신러닝 또는 데이터 기반 마케팅에 대한 전반적인 지식을 보유하신 분
- 데이터 추출, 전처리, 분석, 인사이트 도출까지 데이터 분석 전과정을 경험해보신 분
- SQL을 자유자재로 활용하여 Raw Data 처리 및 분석이 가능하신 분
- 분석 결과를 통한 결과 도출 및 논리적인 개선사항을 제시할 수 있는 분
- 데이터 시각화(Tableau, Amplitude 등)에 대한 경험을 보유하신 분
- 기획, 마케팅, 개발 등 다양한 직군과 원활한 커뮤니케이션 및 협업이 가능하신 분

주요업무

- 사업 및 프로덕트의 핵심 지표를 정의/측정하고 리포팅 및 시각화하여 관리해요
- 주요 이슈에 대한 가설을 세우고 실험 및 검증을 통하여 개선 실행안을 도출해요
- 사용자 데이터 분석을 통해 의사결정을 지원하고 인사이트를 도출해요
- 데이터를 이용해 서비스에 적합한 소프트웨어 로직을 만들어요

자격요건

- 관련 경력이 3년 이하이신 분
- 데이터를 원하는 형태로 쉽게 핸들링 할 수 있는 능력이 있으신 분
- 문제에 대한 가설을 세우고 검증 및 분석하는 역량이 있으신 분

주요업무

스포카의 데이터 분석가는 데이터를 이용하여 서비스의 성장을 만들어내는 중요한 역할입니다. 서비스를 직접적으로 만드는 PM, 디자이너, 개발자와 함께 소통하면서 데이터 중심의 의사결정에 주도적으로 역할을 진행합니다. 키친보드 서비스를 함께 성장시켜주실 분을 찾습니다.

- 데이터를 이용하여 서비스의 다양한 문제들을 해결하는 업무를 진행합니다.
- 서비스의 핵심 지표들을 정의하고, 추적하고, 시각화하여 팀원들에게 공유합니다.
- 서비스 이용 패턴 분석을 위해 데이터 로그를 설계하고 관리합니다.
- 사용자 분석을 통해 서비스에 직접적으로 영향을 줄 수 있는 인사이트들을 도출해서 공유합니다.
- 전사 및 부서별 KPI 정의하고 관리합니다.

자격요건

- 1년 이상 프로덕트 or 비즈니스 분석 경력을 보유하신 분
- SQL(BigQuery, PostgreSQL), Python 등을 활용한 데이터 분석 실무 경험이 있으신 분
- Redash를 통한 대시보드 구성 및 시각화에 익숙하신 분
- 모바일 서비스 로그 설계 혹은 분석 경험이 있으신 분
- 데이터와 인사이트를 분석하여 명확하고 실행 가능한 계획을 제시한 경험이 있으신 분

채용시장에서의 요구 조건 2 - 인사이트 도출 결과 보고서 작성 능력 요구

주요업무

< 이 업무를 같이 하실 분을 기다립니다 >
- 분석지원팀의 다양한 프로젝트 중 '게임 플레이 로그 데이터 분석'을 수행합니다.
- 분석을 통해 비즈니스 인사이트를 도출하고, 도출된 인사이트가 게임에 적용되어, 더 나은 방향으로 게임이 운영될 수 있도록 돕습니다.

< 업무 종류 >
- 분석 설계, 데이터 추출 및 가공, 적재
- 분석 리포트 작성 (논리적 스토리텔링)
- Tableau 기반 시각화
- 개발팀, 기획팀, 사업팀, 운영팀 등 유관 부서와의 커뮤니케이션 및 다양한 협업

< 분석 주제 예시 >
- 게임 플레이 허들 분석
- 게임 패치 및 프로모션 영향도 측정
- 지표 증감 영향 요인 분석 등

자격요건

< 이런 분을 찾고 있어요 >
- 넥슨 게임에 대한 깊은 관심과 이해가 있는 분 (게임 플레이 경험, 유튜브 시청 등의 간접 경험 포함)
- 데이터 분석 역량이 있는 분 (SQL, Python 등)
- 데이터를 통계적으로 이해하고 논리적으로 설명할 수 있는 분
- 문제 상황이 주어졌을 때, 문제 상황을 구체적으로 분석할 수 있는 형태로 정의하고, 논리적 스토리텔링이 가능한 구조로 고민하는 것을 즐기는 분

주요업무

- 데이터 분석 결과 해석 및 보고서 작성 (데이터 대시보드 (Tableau/Redash 등) 운영 및 제작)
- 비모소프트가 서비스 중인 어플리케이션 유저들의 행태 DATA를 정제/가공/분석 및 인사이트 도출
- Product, Marketing, Development, HR, Finance 등 다양한 팀과의 협업을 통해 비즈니스 문제를 정의하고, 이를 해결하기 위한 가설 수립
- 가설 수립 및 검증을 통해 비즈니스 로직 개선 방안 제안
- 데이터를 수집, 분석 및 시각화하여 사용자 행동을 이해하고 서비스가 사용자의 요구를 충족하고 있는지 확인
- 주기적으로 데이터 분석 결과를 리포트하고 이를 협업 부서에 전달

자격요건

- 데이터 분석 관련 실무 경력 2년 이상이신 분
- 데이터 분석, 머신러닝 또는 데이터 기반 마케팅에 대한 전반적인 지식을 보유하신 분
- 데이터 추출, 전처리, 분석, 인사이트 도출까지 데이터 분석 전과정을 경험해보신 분
- SQL을 자유자재로 활용하여 Raw Data 처리 및 분석이 가능하신 분
- 분석 결과를 통한 결과 도출 및 논리적인 개선사항을 제시할 수 있는 분
- 데이터 시각화(Tableau, Amplitude 등)에 대한 경험을 보유하신 분
- 기획, 마케팅, 개발 등 다양한 직군과 원활한 커뮤니케이션 및 협업이 가능하신 분

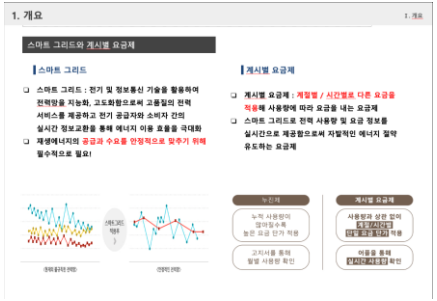
팀별 제출물 안내

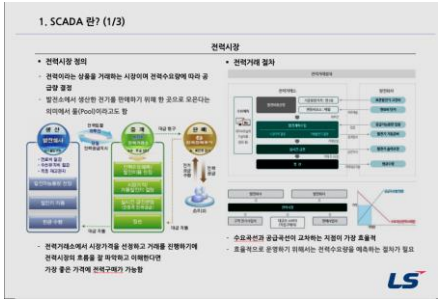
No.	제출물	제출 및 발표 여부
1	분석 주제 선정 보고서 (ppt)	제출
2	분석 코드 (.ipynb)	제출
3	최종결과보고서 (ppt)	제출, 발표

※) 보고서 양식 제공

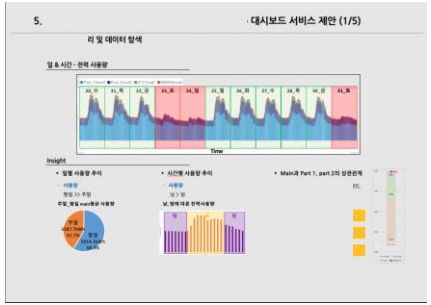
팀별 제출물 안내 – 1) 분석주제 선정 보고서 (ppt)

✓ 데이터를 탐색적 분석하여 얻은 **인사이트로 비즈니스 로직/솔루션(안)**을 제안하는 보고서





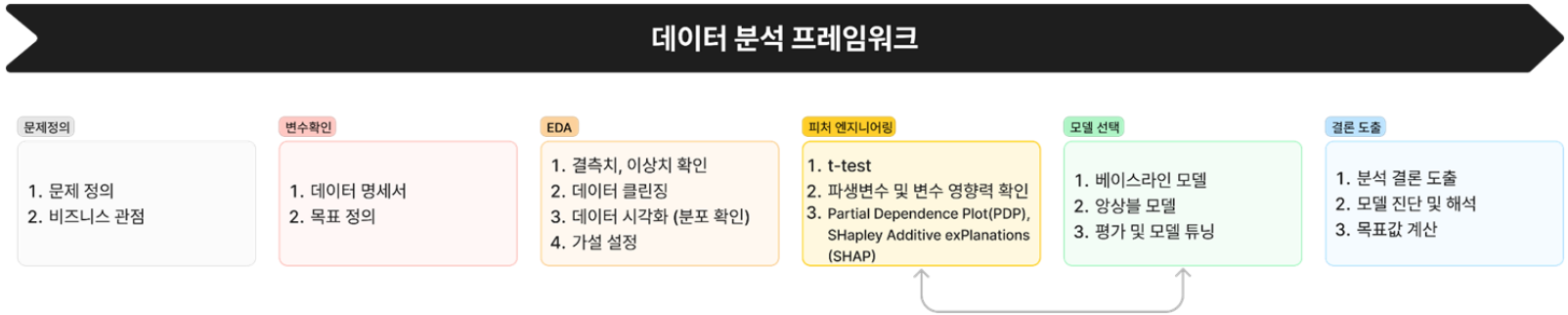
< 1기 분석주제 선정보고서 예시 >



< 결과보고서 예시 >

팀별 제출물 안내 – 2) 분석코드(.ipynb) (EDA & 모델 코드)

- ✓ EDA 코드 : 최종 독립변수를 정의하기까지의 모든 과정이 담긴 코드
- ✓ 모델 코드 : 코드를 실행하면, 보고서에 나오는 차트나 모델값이 모두 나오는 코드



팀별 제출물 안내 – 3) 최종 결과 보고서 (ppt)

✓ 프로젝트 기간동안 분석한 결과물을 **가상의 고객사에게 보고하는 내용**으로 **논리적 스토리텔링**이 담겨있는 결과 보고서

< 최종 결과 보고서 예시 >

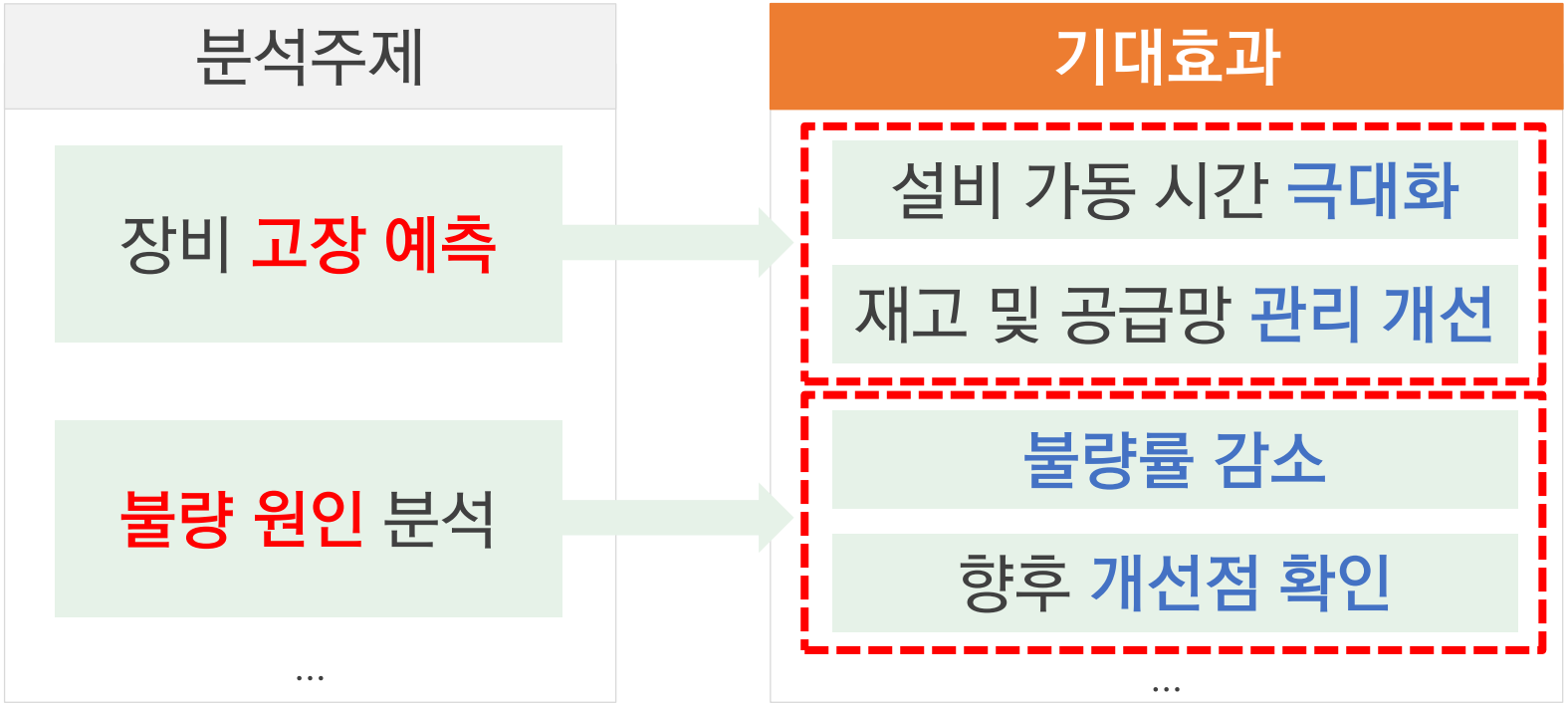
제조 데이터 특징

✓ 다양성 : 제조 과정에서 수집되는 데이터는 기계, 센서, 작업자 입력 등 다양한 출처에서 나오며 온도, 압력, 습도, 제품 이미지 등이 정형, 반정형, 비정형 데이터로 나눠짐

✓ 타겟데이터(결함)의 불균형 : 외관상 발생하는 결함 및 장비 고장 등 이상 데이터들이 굉장히 적은 수로 발생하며 불균형 데이터 처리 기법 등의 분석이 필요

✓ 속도 : 제조 공정에서 실시간 처리의 중요성은 높음. 공정 상태의 실시간 모니터링을 위한 빠른 데이터 처리 및 분석이 필요

제조 데이터분석 기대효과



CHAPTER

1주 실습 데이터 소개

1주 실습 데이터 소개(1) – 제조 불량 데이터

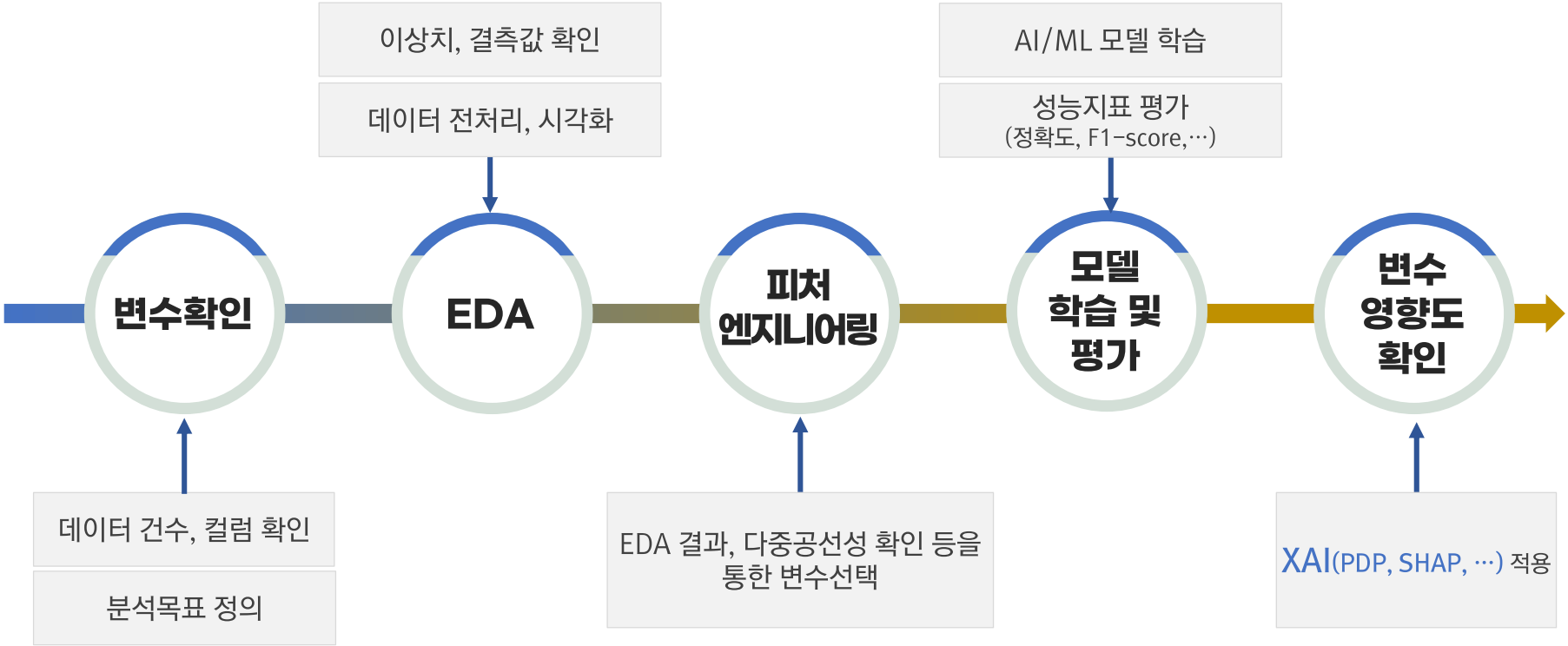
✓ 목표 : 불량을 야기하는 변수를 찾고, 불량품을 낮추기 위한 분석 진행

✓ **타겟변수**(종속변수) : 불량여부를 나타내는 이진데이터(0 또는 1)

✓ **설명변수**(독립변수) : 변수에 대한 설명이 없는 비식별화된 20개의 변수

설명변수					타겟변수
X1	X2	…	X19	X20	Y
0.3994	0.0013	…	0.683	0.033	1
…	…	…	…	…	…

1주 실습 데이터 소개(2) – 제조 불량 데이터



CHAPTER

2주 실습 데이터 소개

2주 실습 데이터 소개(1) – 전력사용량 데이터

✓ 목표 : 건물 정보와 기후 정보를 활용한 전력사용량 예측

✓ **타겟변수**(종속변수) : 건물의 날짜/시간 별 전력사용량(kWh)

✓ **설명변수**(독립변수) : 60개 건물에 대한 날짜/시간 별 건물 및 기상 데이터

설명변수

date_time	기온(°C)	…	비전기냉방설비운영	태양광보유
2020-06-01 00	17.6	…	0	0
…	…	…	…	…

타겟변수

kWh
8179.056
…

2주 실습 데이터 소개(2) – 전력사용량 데이터

✓ 목표 : 훈련 데이터의 feature들 + 시계열 데이터를 기반으로 전력사용량 예측

✓ 데이터 개요 : train.csv (122,400, 10) / test.csv (10,080, 9)

컬럼명	컬럼명 정의
전력사용량(kWh)	**전력사용량(kWh)**
Num	건물번호
date_time	시간
기온(°C)	기온(°C)
풍속(m/s)	풍속(m/s)

습도(%)	습도(%)
강수량(mm)	강수량(mm)
일조(hr)	일조(hr)
비전기냉방설비운영	운영여부 (0,1)
태양광보유	보유여부 (0,1)

※ 비고 : 기상 데이터는 train set에서는 1시간 단위로 제공 되나, test set에서는 3시간 단위로 제공 (강수량의 경우 6시간, 예보데이터)

CHAPTER

3주 실습 데이터 소개

3주 실습 데이터 소개(1) – 건설장비 데이터

✓ 목표 : 건설장비에서 작동오일 상태를 판단하기 위한 모델 개발

✓ **타겟변수**(종속변수) : 정상여부를 나타내는 이진데이터(0 또는 1)

✓ **설명변수**(독립변수) : 기계별 Features (Train 52개 / Test 18개)

설명변수

COMPONENT_ARBITRARY	ANONYMOUS_1	…	V40	ZN
COMPONENT3	1486	…	154	75
…	…	…	…	…

타겟변수

Kwh
0 (정상)
…

3주 실습 데이터 소개(2) – 건설장비 데이터

✓ 데이터 개요 : train.csv (14,095, 53) / test.csv (6041, 18)

✓ 특이사항 : 53개의 훈련세트 feature vs 18개의 테스트세트 feature (Highlighted)

컬럼명	컬럼명 정의
Y_LABEL	오일 정상 여부 (0 : 정상, 1 : 이상)
COMPONENT_ARBITRARY	샘플 오일 관련 부품 (Component 4종, 비식별화)
ANONYMOUS_1	무명 Feature 1. 수치형 데이터
YEAR	오일 샘플 및 진단 해
SAMPLE_TRANSFER_DAY	오일 샘플링 후 진단 기관으로 이동한 기간 (Days)
ANONYMOUS_2	무명 Feature 2. 수치형 데이터
AG	은 함유량
AL	알루미늄 함유량
B	붕소 함유량
BA	바륨 함유량
BE	베릴륨 함유량
CA	칼슘 함유량
CD	카드미늄 함유량
CO	코볼트함유량

CR	크로뮴 함유량
CU	구리 함유량
FH2O	물 수치(By FT-IR)
FNOX	질소산화물 수치(By FT-IR)
FOPTIMETHGLY	비식별화
FOXID	산화 수치(By FT-IR)
FSO4	황산 수치(By FT-IR)
FTBN	염기성 첨가제 수치(By FT-IR)
FE	철 함유량
FUEL	연료 함유량
H2O	물 함유량
K	포타슘 함유량
LI	리튬 함유량
MG	마그네슘 함유량
MN	망간 함유량
MO	몰리브데넘 함유량
NA	나트륨 함유량
NI	니켈 함유량
P	인 함유량

PB	납 함유량
PQINDEX	Particle Quantifier Index
S	황 함유량
SB	안티모니 함유량
SI	실리콘 함유량
SN	주석 함유량
SOOTPERCENTAGE	Soot(그을음) 함유량(%)
TI	티타늄 함유량
U100	100㎛ 이상 크기 입자 수
U75	75㎛ 이상 크기 입자 수
U50	50㎛ 이상 크기 입자 수
U25	25㎛ 이상 크기 입자 수
U20	20㎛ 이상 크기 입자 수
U14	14㎛ 이상 크기 입자 수
U6	6㎛ 이상 크기 입자 수
U4	4㎛ 이상 크기 입자 수
V	바나듐 함유량
V100	점도 (at 100 degrees)
V40	점도 (at 40 degrees)
ZN	아연 함유량

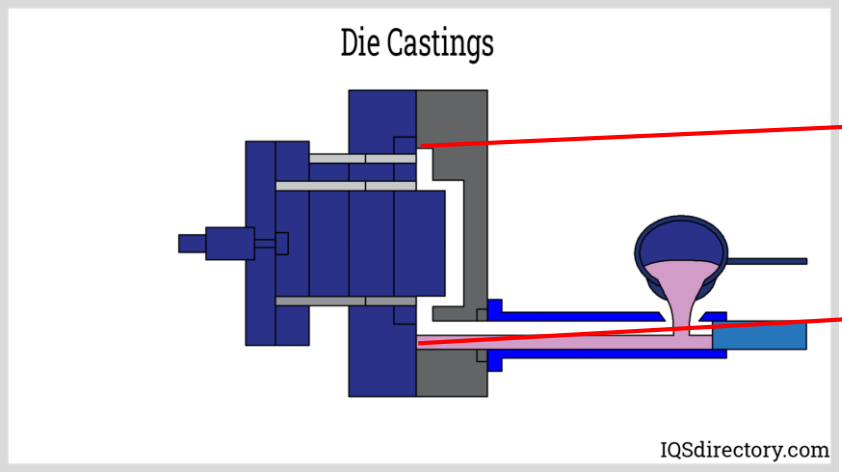
CHAPTER

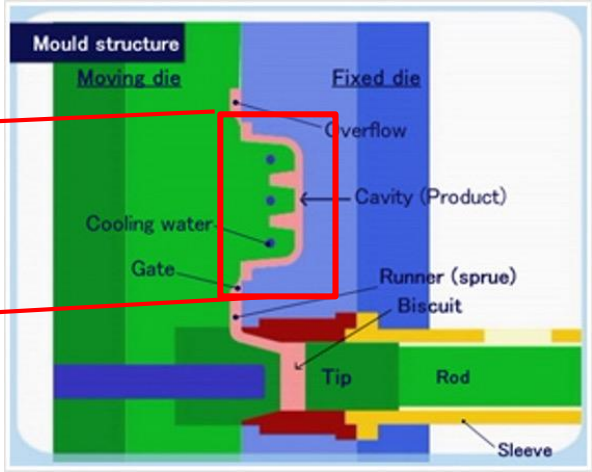
4, 5주 실습 데이터 소개

4주 실습 데이터 소개(1) – 다이캐스팅 데이터

✓ 용융된 금속을 금형에 밀어 넣는 금속 주조 공정

✓ 고품질 마감 처리로 치수가 정확한 정밀 금속 부품을 대량 생산하는데 주로 사용함





4주 실습 데이터 소개(2) – 다이캐스팅 데이터

✓ 목표 : 주조 공정에서 발생하는 불량품에 영향을 주는 변수를 찾고 최적화를 위한 주요 변수 구간 제시

✓ 데이터 레이아웃 : row 92,015개 X column 31개

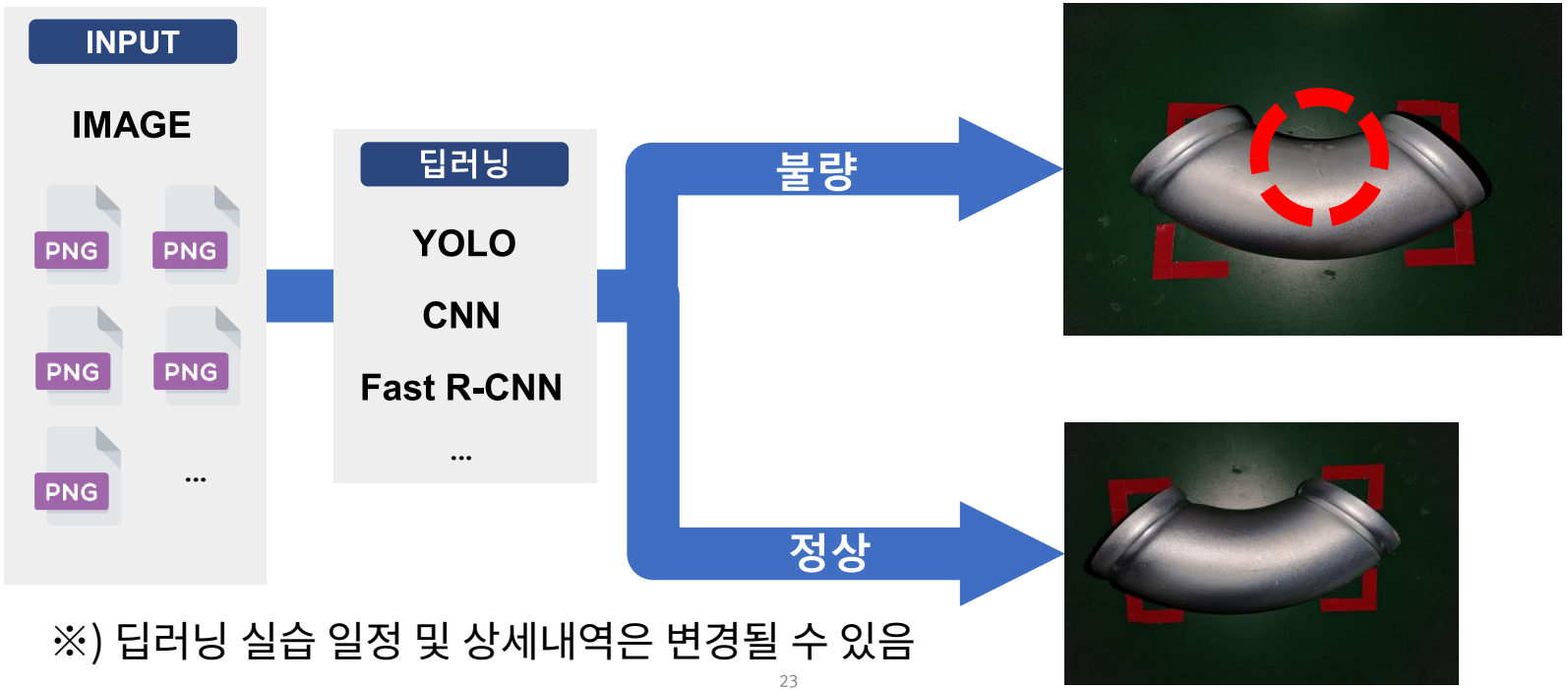
컬럼명	컬럼명 정의
passorfail	**양품불량판정**
line	작업라인
Name	제품명
Mold_name	금형명
Time	수집시간
Date	수집일시
Count	일자별 제품 생산 번호
Working	가동여부
Emergency_stop	비상정시
Molten_temp	용탕온도

facility_operation_CycleTime	설비 작동 사이클시간
production_CycleTime	제품생산 사이클시간
low_section_speed	저속구간속도
high_section_speed	고속구간속도
molten_volume	용탕량
cast_pressure	주조압력
biscuit_thickness	비스켓 두께
upper_mold_temp1	상금형온도1
upper_mold_temp2	상금형온도2
upper_mold_temp3	상금형온도3
lower_mold_temp1	하금형온도1

lower_mold_temp2	하금형온도2
lower_mold_temp3	하금형온도3
sleeve_temperature	슬리브온도
physical_strength	형체력
Coolant_temperature	냉각수 온도
EMS_operation_time	전자교반 가동시간
registration_time	등록일시
tryshot_signal	사탕신호
mold_code	금형코드
heating_furance	가열로

4주 추가 실습 소개(1) – 딥러닝 기반 이미지 데이터 처리

✓ 목표 : 크로메이트(금속표면 처리공정의 종류) 과정에서 발생하는 불량을 **딥러닝** 활용 탐지



※) 딥러닝 실습 일정 및 상세내역은 변경될 수 있음

CHAPTER

XAI의 이해

“ XAI의 이해 ”

Contents

XAI의 이해 – 1. AI의 성공과 보완점

Artificial Intelligence

인공지능

사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술

Machine Learning

머신러닝

컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술 방법

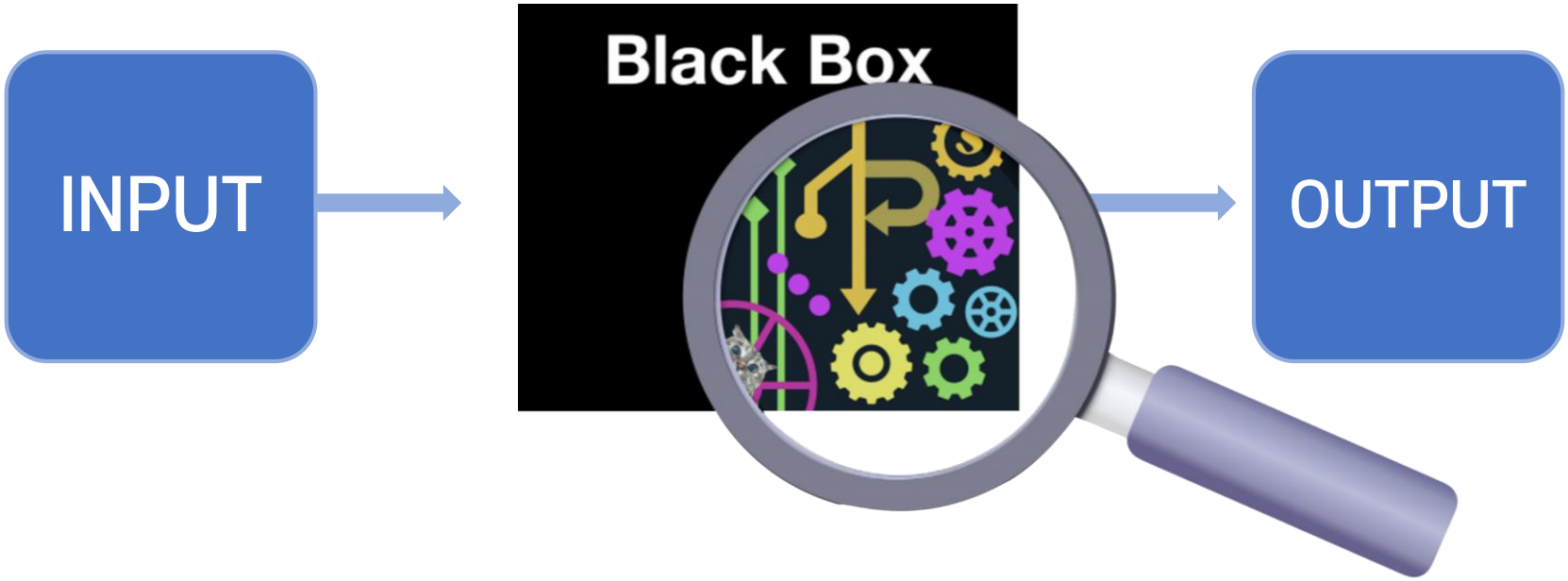
Deep Learning

딥러닝

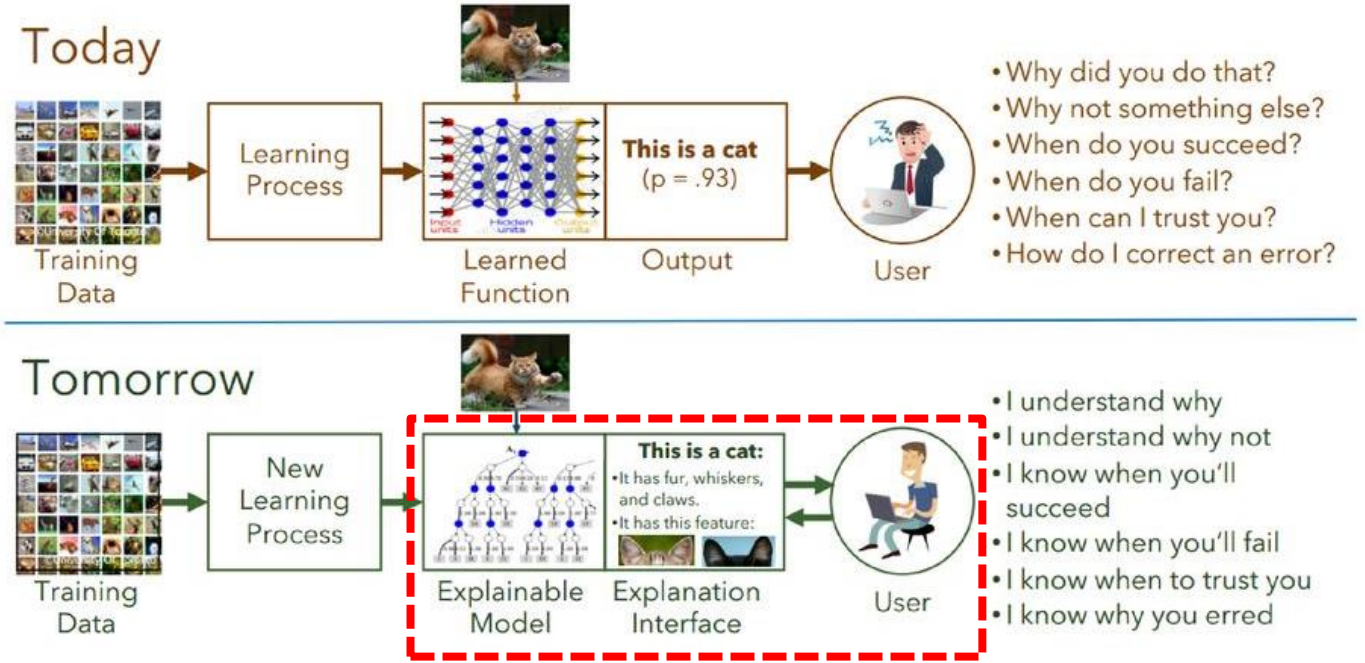
인간의 뉴런과 비슷한 인공신경망 방식으로 정보를 처리

XAI의 이해 – 2. AI의 활용 분야

XAI의 이해 – 3. XAI 기술이 왜 필요한가?



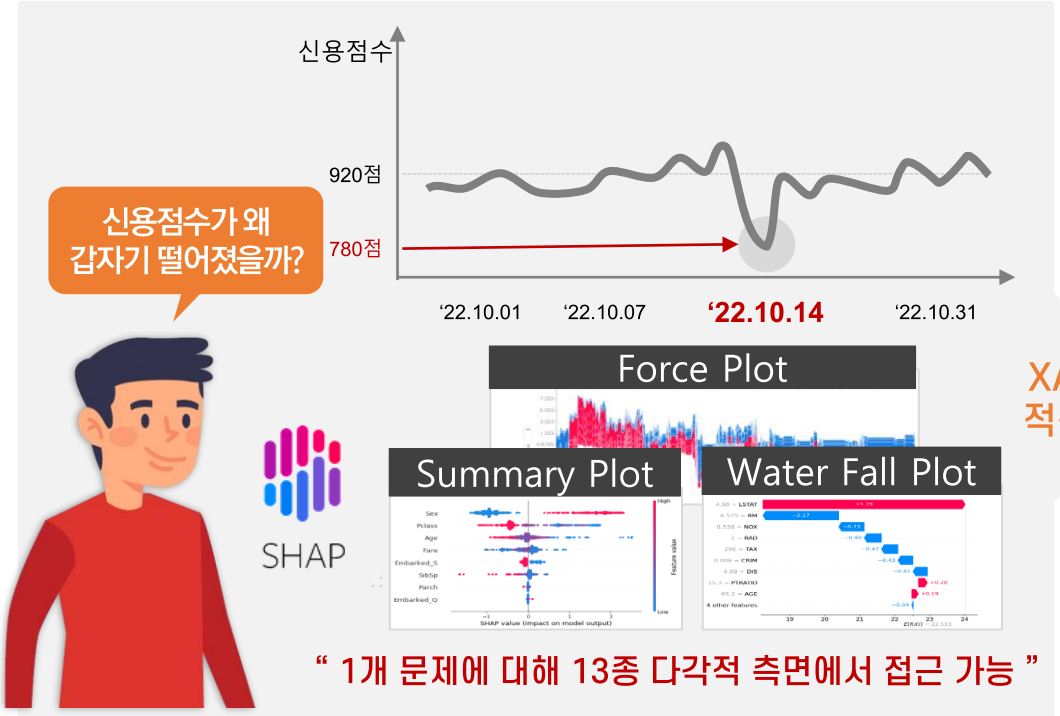
XAI의 이해 – 4. XAI(eXplanable AI)란 무엇인가?



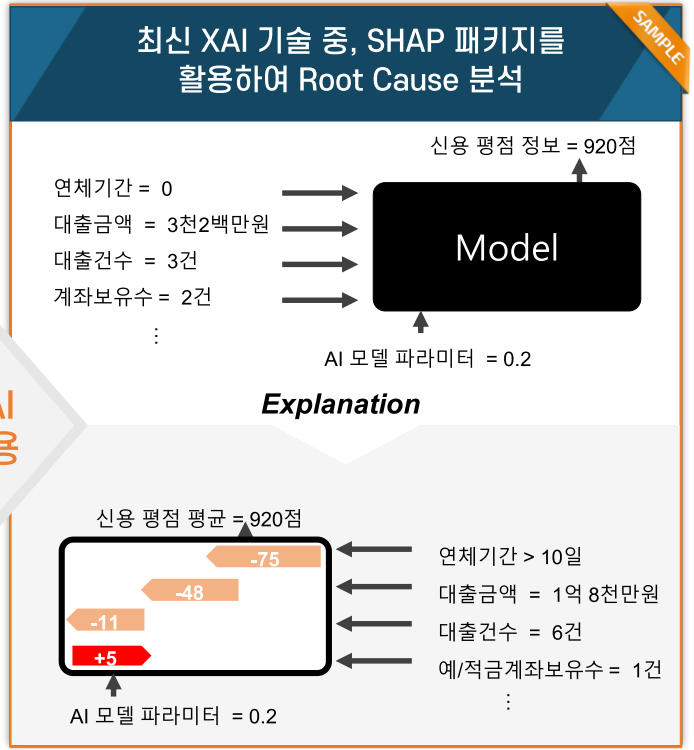
[그림1] 미국 방위고등연구계획국(DARPA)의 AI와 XAI의 개념 비교도

XAI의 이해 – 5. XAI의 비즈니스 활용

A회사의 XAI 기술의 비즈니스 활용 (Use SHAP-SHapley Additive Explanation)



XAI 적용



XAI의 이해 – 5. XAI의 비즈니스 활용(일하는 방식의 변화)

Near Real Time으로 문제의 근원 파악이 가능하며 업무의 효율성 증대 및 Deep-Dive 분석을 통해 새로운 Insight 도출 가능

AS-IS 기존에 일하는 방식

SAMPLE

가설/검증 Loop

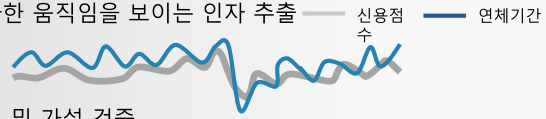
08:30 DB접속 및 이슈 발생한 기간 조회

09:30 담당자의 경험을 통해 신용도에 **영향을 줄 수 있는 여러 인자들을 선별**하여 차트로 **변동 추세 확인**

1) 신용점수, 연체횟수, 연체기간 등 변화 확인
→ 그 외 신용평점에 영향이 있는 사항 확인

10:30

가장 유사한 움직임을 보이는 인자 추출



11:30 원인 유추 및 가설 검증

" 가설 수립 → 검증 → 원인 규명 불가 → 가설 재 수립 → 재 검증 "

TO-BE XAI기반 일하는 방식

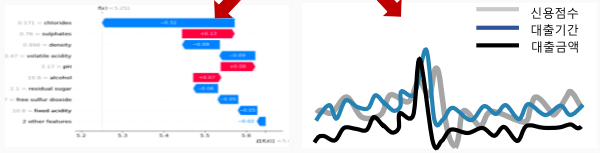
SAMPLE

08:30 이슈 발생 기간 포착 및 기간 설정, 분석 시작

❶ ForcePlot을 통한 신용도 하락에 미치는 전체 인자별 영향도 파악

Near Real-time Root Cause 도출

❷ Deep-Dive 기능을 통한 Root Cause 발굴



09:00

" 대출금액/대출기간/이자 상승 → 월납입 금액 증가(소득변동X)
→ 신용도 하락 "

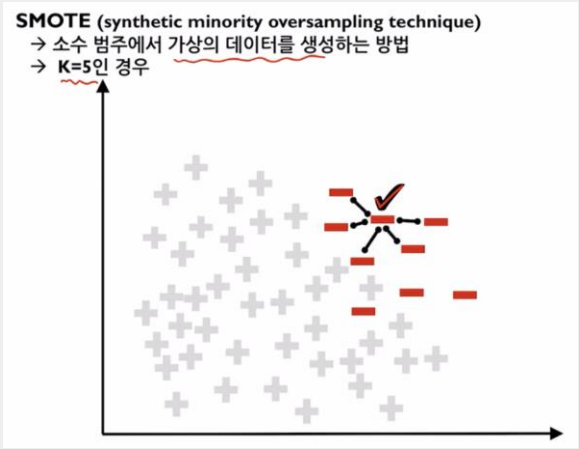
Q & A

CHAPTER

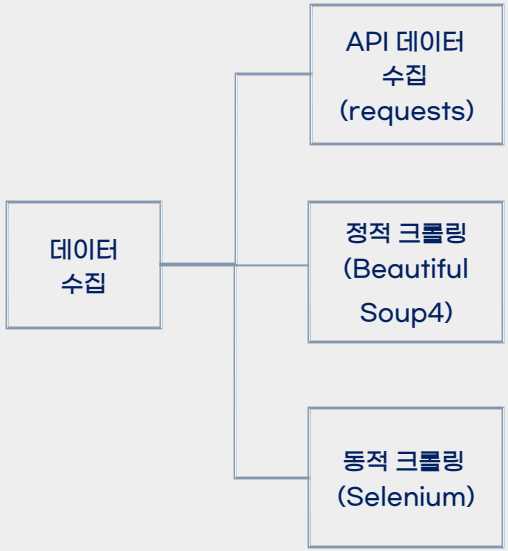
추가 실습 기법

추가 실습 기법

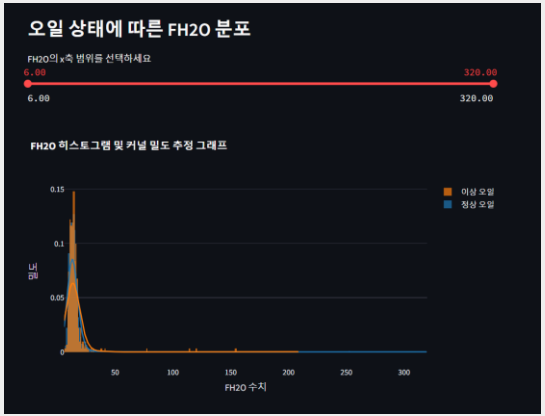
불균형 데이터 처리 SMOTE - 3주차 예정



크롤링 – 3, 4주차 예정



Streamlit – 5주차 예정



- Streamlit을 활용한 대시보드 생성 및 배포 실습