직장인을 위한 파이썬 데이터 분석

현업 데이터 사례 기반 문제해결 방법론

목차(Context)

· '데이터 분석가 이직 노하우 전자책'에 소개된 1차 실무 면접 질문 리스트를 기반으로 현업 문제해결 방법론을 데이터 분석 관점에서 정리한 책입니다.

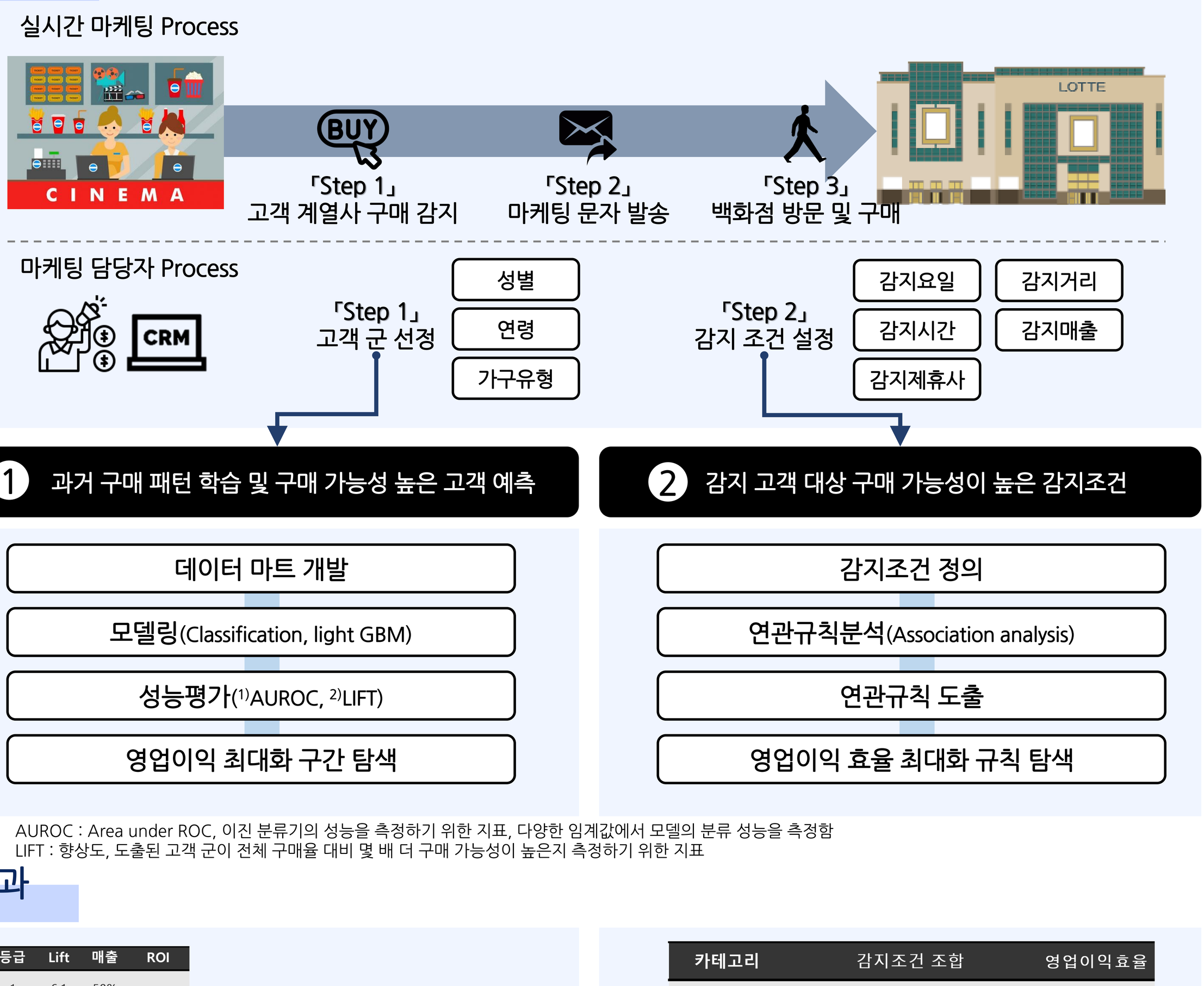
「유통」

실시간 마케팅 반응고객 예측 및 감지조건 최적화

프로젝트 개요

· 트랜드가 빠르게 변화하는 유통업(백화점)에서는 신규 고객을 확보하기 위해 계열사 이용 고객 대상으로 실시간 마케팅을 수행하고 있음. 실시간 마케팅을 효과적으로 활용하기 위해 1) 구매 가능성이 높은 고객 군과 2) 최적의 감지조건의 규칙을 찾고 영업이익을 최대화 하고자 함

개념 설계



① 과거 구매 패턴 학습 및 구매 가능성 높은 고객 예측

- 데이터 마트 개발
- 모델링(Classification, light GBM)
- 성능평가(1)AUROC, 2)LIFT)
- 영업이익 최대화 구간 탐색

② 감지 고객 대상 구매 가능성이 높은 감지조건

- 감지조건 정의
- 연관규칙분석(Association analysis)
- 연관규칙 도출
- 영업이익 효율 최대화 규칙 탐색

1) AUROC : Area under ROC, 이진 분류기의 성능을 측정하기 위한 지표, 다양한 임계값에서 모델의 분류 성능을 측정함
2) LIFT : 향상도, 도출된 고객 군이 전체 구매율 대비 몇 배 더 구매 가능성이 높은지 측정하기 위한 지표

분석결과

등급	Lift	매출	ROI
1	6.1	50%	-
2	4.0	70%	-
3	2.1	80%	-
4	1.3	90%	-
5	0.9		-
6	0.7		Max
7	0.5		
8	0.3		
9	0.1		
10	0		

· 모델 예측 결과를 예측확률(Probability) 기준으로 내림차순 정렬 후 고객 수를 10 분위수로 나눠 등급 부여

· 6등급 고객에게 까지 마케팅 수행했을 시 영업이익 최대의 효과를 기대할 수 있음

카테고리	감지조건 조합	영업이익효율
MALL(**종합몰**)	(MALL, 200M 이내, 12-18h, 5<=)	897 (5.9)
DSTR(백화점)	(DSTR, 200M 이내, 12-18h	862 (5.7)
토요일	(토요일, 12-18h, 200M 이내, 5<=)	770 (5.1)
금요일	(금요일, 200M 이내, 12-18h, 5<=)	756 (5.0)
기타	(200M 이내, 12-18h, 5<=)	742 (4.9)
Total	모든 규칙	151 (1.0)

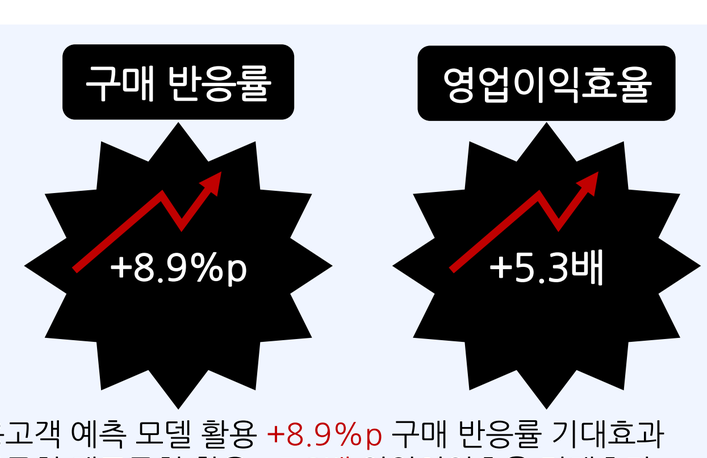
· 규칙을 설정하지 않은 것 대비 영업이익효율이 좋은 최종 대표 규칙 5종 도출

기대효과

구분	대상	반응률
기존(Base)	전체 고객	23.6%
반응고객 예측 모델	모델링을 통해 Scoring한 1~6등급 고객	**32.5%**

구분	대상	1)영업이익 효율
기존(Base)	전체 고객	151
연관규칙분석모델	대표 규칙 행동패턴을 수행한 고객	**2)805**

1) 영업이익 효율 : 영업이익/마케팅 투자 비용(판매관리비)
2) 805 : 최종 도출된 대표규칙 5개의 평균 영업이익 효율



· 반응고객 예측 모델 활용 +8.9%p 구매 반응률 기대효과
· 연관규칙 대표규칙 활용 +5.3배 영업이익효율 기대효과

「유통」

옴니채널 활용 360° 고객 이해 및 타겟 마케팅

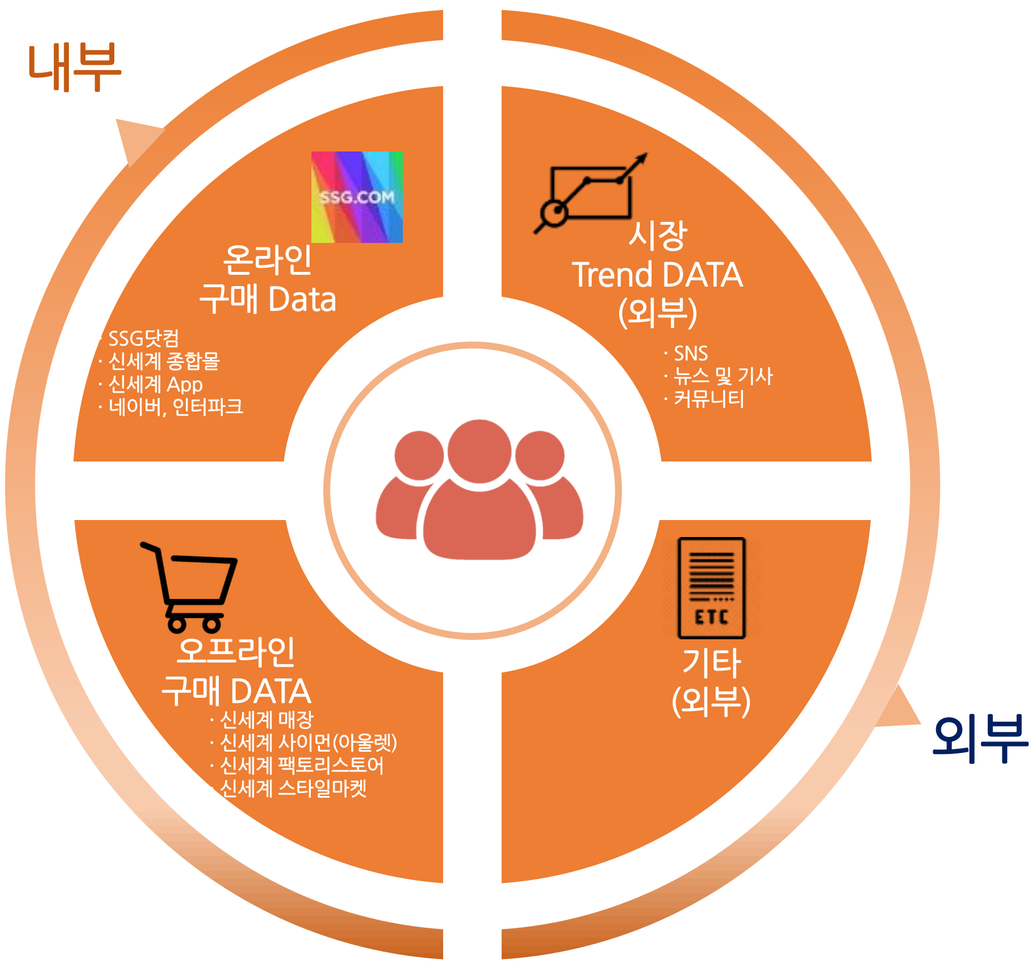
프로젝트 개요

・빠르게 변화하는 고객 트랜드에 대응하기 위해서 1) 내/외부 추가 데이터 수집 및 활용 확대를 통해 다양한 관점에서 고객을 이해하고 2) 미래 채널변화에 따라 채널 별 관리 고객 및 서비스 제공 방식 변화가 요구됨

개념 설계

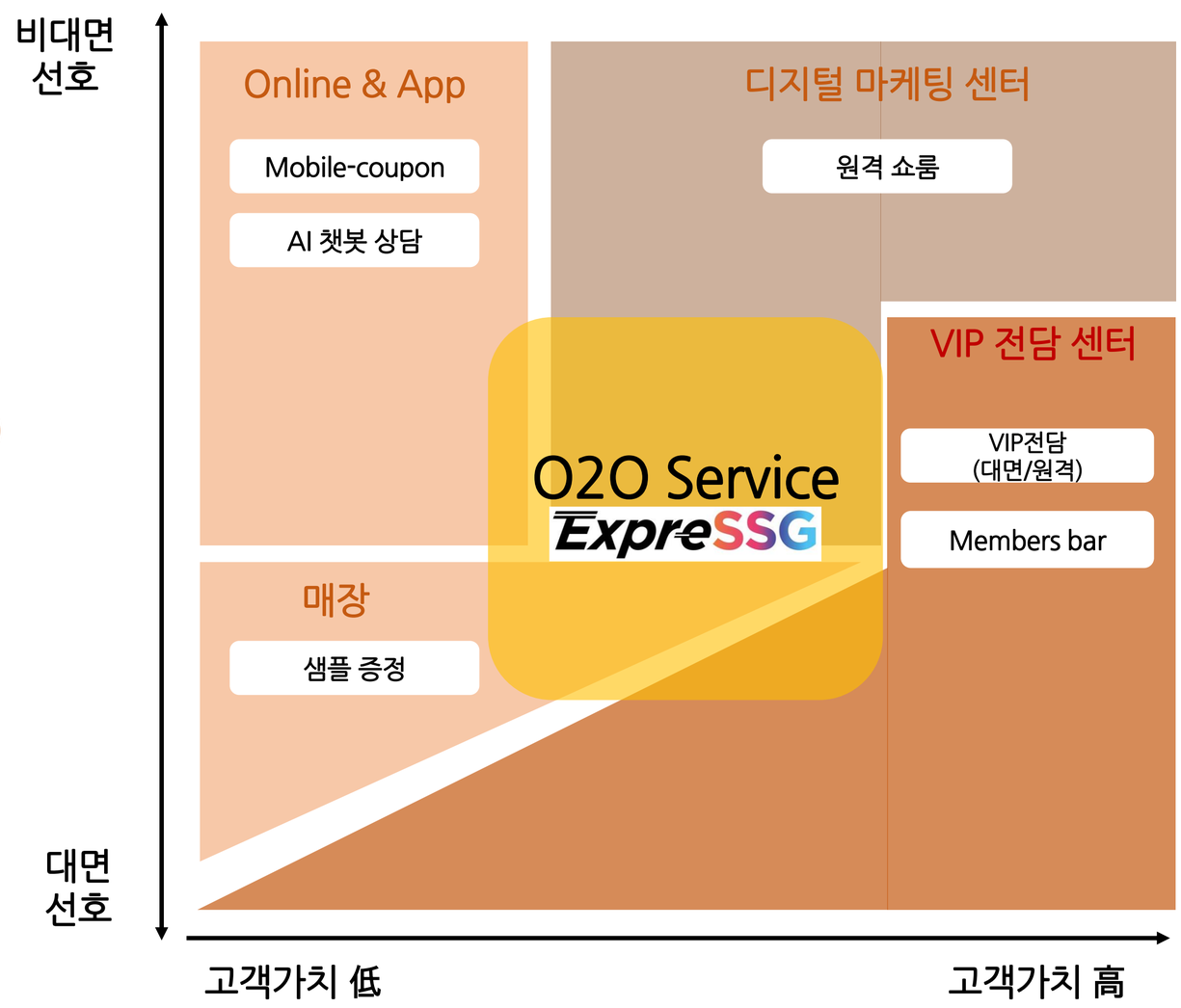
① 데이터 기반의 360° 고객 이해

・옴니채널 활용 다양한 내/외부 데이터 수집 및 고객 분석



② 채널별 관리 고객 및 차별화된 서비스 제공

・선호 채널에 따른 차별화된 서비스 제공 및 O2O 서비스 공략



기대효과

③ 마케팅 반응률 증대

・360 ° 고객 이해를 바탕으로 예측 모델 활용, 상위 Score 고객 대상 타겟 마케팅 진행 시 기존 대비 **반응률 증대** 가능

④ 마케팅 비용 효율성 증대

・마케팅 추진 시 반응률 저조 고객군 제거 및 고객군별 선호 채널 마케팅으로 **저비용 고효율 성과** 달성

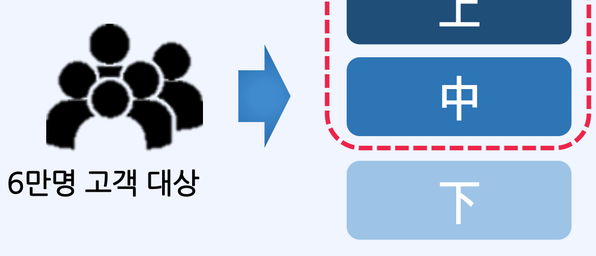
*Case1) 기존 전수 고객 대상 *LMS 마케팅 진행 시*

6만명 고객 대상

LMS(장문 문자 메세지)
☞ 25원/건

마케팅 반응률	2% 가정
마케팅 비용	150만원
반응 고객수	1,200명

Case2) 상위 Score 타겟 마케팅 진행 시(1)예측 모델 활용)



6만명 고객 대상

18,000명 상위 Score 고객 추출
☞ 마케팅 반응률이 높은 고객군을 선별하여 타겟 마케팅 진행

마케팅 비용 절감	마케팅 반응률 증대	반응 고객수
70% ↓	4.0%p ↑	90%
150만원 → 45만원	2% → 6%	1,080명

1) 머신러닝 예측 모델을 활용하여 기존 반응률 대비 약 3배 높은 반응률을 보이는 고객군을 추출

2) 기존 반응 고객수 1,200명 대비 90% 수준(1,080명)

「이커머스」
전염병 시기 특수 품목 수요 예측

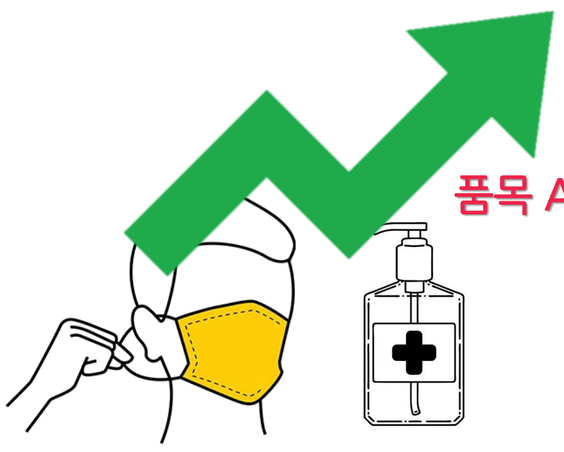
프로젝트 개요

· 전염병 발병 시 특정 품목 판매량 증가에 따른 품절 상태 발생으로 인해 손해 발생을 방지하기 위해 과거 전염병 기간 동안 발생한 거래 데이터를 활용하여 판매량 예측 모델을 생성하고자 함

개념 설계

① 품목 선정

· 전염병 기간 동안 판매 민감도 높은 품목 선정



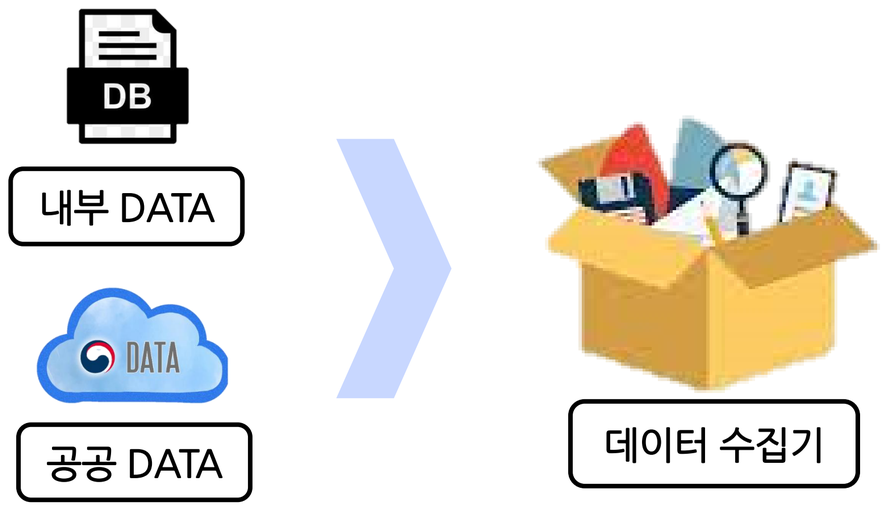
품목 A 거래량 > (품목 A 평균 거래량) x 2

평균 판매량 대비 급격하게 거래가 증가한 품목 탐색

· 과거 전염병 기간 동안 급격하게 판매량 증가 품목 탐색
· 해당 품목 중 품절 횟수가 가장 높은 품목 예측 대상 선정
· 품목 구분 어려울 시 상위 카테고리를 예측 대상으로 선정

② 데이터 수집

· 공공 데이터 및 거래 데이터 수집



· 발병일, 치명률 등 전염병 특성 관련 공공 데이터 수집
· 예측 품목 대상 발병 전/후 N개월 거래 데이터 수집

③ 데이터 마트 개발

· 가설 수립 및 데이터 마트 개발

가설

"전염병 발병 시 품절 품목은 첫 국내 환자 확진 후 판매량의 변화가 있었을 것이다."

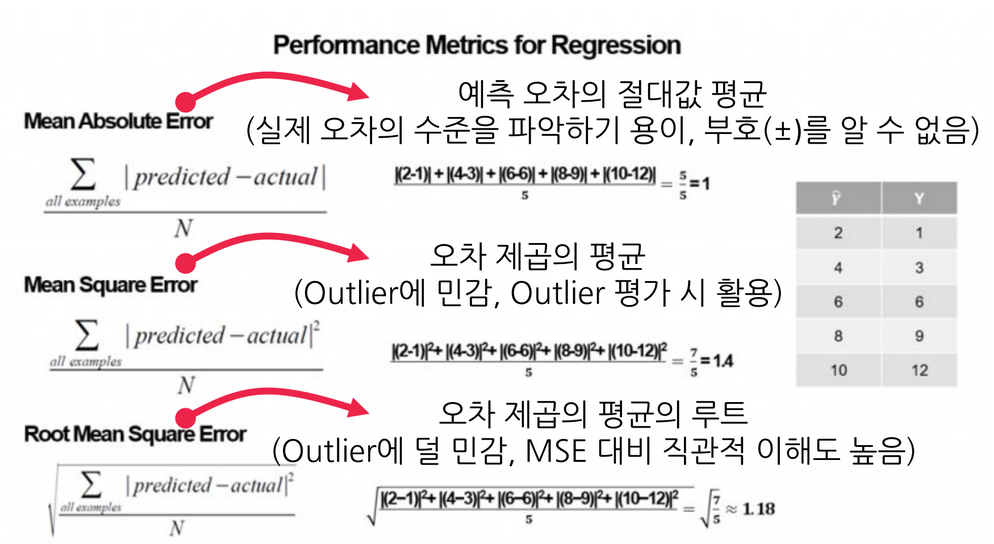
DATA

'국내 환자 확진일 기준 판매량 증감률 "

· 모델 학습을 위한 데이터 마트 개발
· 다양한 가설을 수립하고 관련된 변수를 개발
· 가설 기반 변수의 집합 → 데이터 마트

④ 모델 학습 및 평가

· 모델 학습 및 평가, Hyper parameter 최적화



· 판매량 예측 문제이기 때문에 Regression 알고리즘 탐색
· MAE, MSE, RMSE, R2 Score 등 평가지표 활용 모델 평가
· Hyper parameter 최적화를 통한 모델 성능 향상

기대효과

특정 이벤트 발생시 수요 대응

코로나 상황 적용 및 기대 매출 분서

효율적 재고관리

재고 대응 매출 상승

「이커머스」
기업 매출 감소 원인 분석

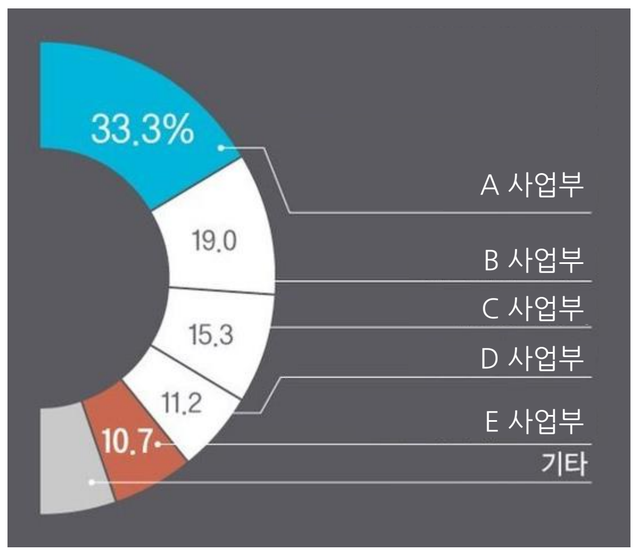
프로젝트 개요

· 기업 매출 감소에 따라 데이터 분석을 통해 매출 감소 원인을 파악하고 대책안(案) 수립을 하려고 함

개념 설계

❶ 매출 구성요소 파악

· 기업 매출을 발생시키는 주요 사업 영역 구분



A 사업부(33.3%)
B 사업부(19.3%)
C 사업부(15.3%)
D 사업부(11.2%)
E 사업부(10.7%)

· 기업의 매출이 발생하는 구조를 파악
· 사업 영역 혹은 사업부처럼 구조화 할 수 있는 단위로 분석

❷ 사업 영역 선정

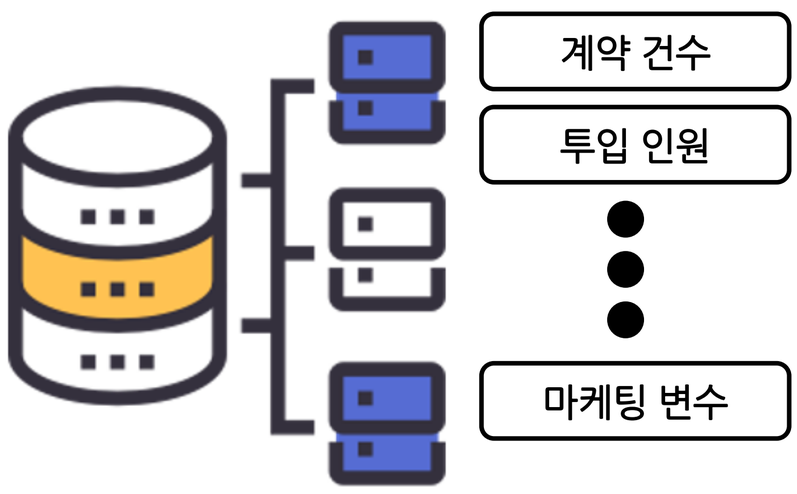
· 매출 감소가 큰 사업 영역 분석 대상으로 선정

B 사업부
분석 대상 선정

· 단위 부서 별 매출 감소가 가장 큰 사업 영역을 분석 대상 선정
· 전체 매출 대비 감소 비중이 큰 영역을 우선 선정

❸ 데이터 마트 개발

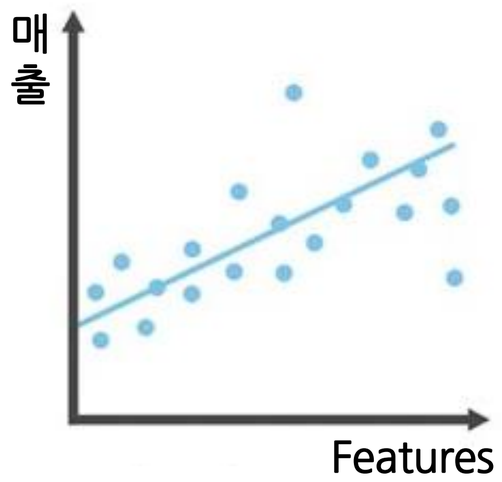
· 사업 영역 매출과 관계 있는 파생 변수 생성



· 선정된 사업 영역의 매출 발생과 관련된 파생 변수를 생성
· 계약 건수, 투입 인원, 마케팅 비용 등 매출에 영향을 끼칠 수 있는 다양한 변수를 탐색하고 데이터를 생성

❹ 매출 감소 원인 파악

· 예측 모델링 활용 중요변수 도출 및 감소 원인 파악



데이터 부족 → 상관관계 분석

리포트 작성

· 충분한 데이터가 있다면 예측 모델링을 수행
· 데이터 부족 시 상관관계 분석을 통해 매출 감소 원인 파악
· 분석된 결과 기반 인사이트 리포트 작성

기대효과

매출 예측 모델링 수행
분석 리포트 기반 대책 수립

→ 매출 감소 원인 파악 / 차년도 계획 수립

「핀테크」
결제 데이터 활용 다양한 파생변수 생성

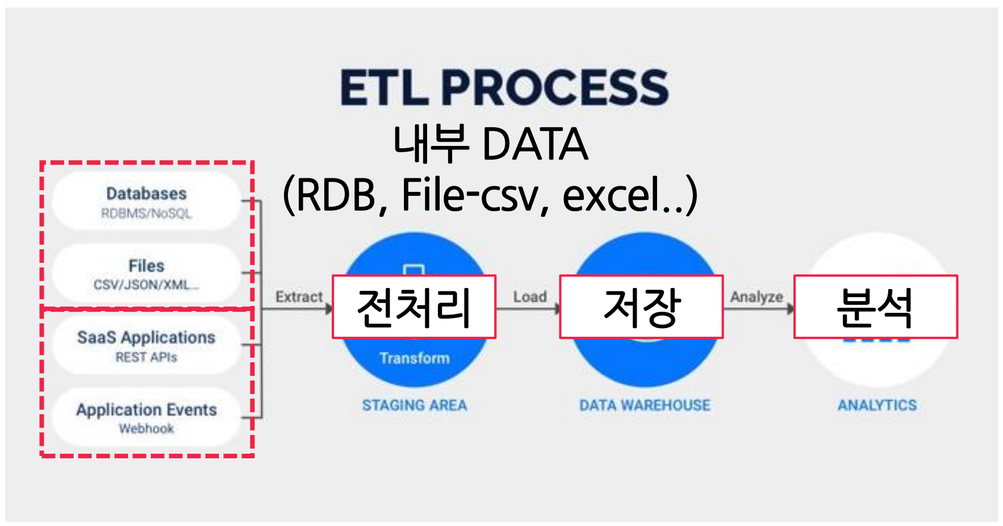
프로젝트 개요

· 데이터 분석을 위한 학습 데이터 변수 부족 문제를 해결하기 위해 기존 데이터를 활용하여 다양한 파생변수를 생성하고 데이터 분석 프로젝트에 활용하려고 함

개념 설계

① 데이터 현황 파악

· 수집 활용 가능한 변수 리스트 확인



· 현재 수집되어 활용 가능한 변수 리스트와 앞으로 수집 가능한 변수 리스트 모두 확인
· 활용할 수 있는 변수가 적다면 공공데이터 활용 추가 확보

② 가설 수립

· 파생 변수 생성을 위한 가설 수립

가설
"과거 구매 형태에 따라 고객의 구매 패턴이 다를 것이다."

DATA
최근 3개월 구매금액, 최근 6개월 구매금액

· 파생 변수 생성을 위한 가설 수립
· 기간(최근 N개월)을 활용하여 다양한 파생 변수 생성
· 과거 대비 현재 구매금액 증감률과 같이 증감 개념을 활용하여 추가 파생 변수 개발

③ 파생 변수 생성 및 평가

· 파생 변수 생성 및 의미 있는 변수 선택

· 결제 데이터는 결제일이라는 시간에 종속된 데이터이므로 기준년월을 기준으로 데이터를 생성해야 함
· 생성된 파생변수 활용 전 사전 EDA과정을 통해 의미 있는 변수 선택

④ 모델링 및 성능 평가

· 최종 변수 활용 모델 학습 및 성능 평가

· 파생 변수 평가 과정을 통해 의미 있는 파생 변수들을 활용하여 해결하려는 예측 모델링에 사용
· 기존 변수와 파생 변수를 활용하여 모델링 학습 후 성능 평가

기대효과

다양한 파생 변수 생성
모델링 학습 및 평가

→ 풍부한 Feature set 확보
→ 모델 성능 향상

「핀테크」
데이터 분석 리포트 활용 이익(Profit) 창출

프로젝트 개요

· 점점 증가하는 기업 내 데이터 분석팀 운영 비용을 충당하기 위해 자사 데이터를 분석하여 인사이트 리포트를 개발하고 판매를 통해 운영 비용에 일부를 충당하고 새로운 비즈니스 기회를 창출

개념 설계

① 목적 설정

· 리포트 목적 설정

Case 1
부가 서비스

Case 2
인사이트 리포트

· 인사이트 리포트를 작성하여 유상계약에 부가 서비스로 제공하거나 리포트 자체를 판매하여 수익을 얻을 수 있음
· 기업에서 판매하고 있는 유형의 제품이나 서비스의 판매를 촉진 하거나 리포트 판매로 추가 수익 달성

② 컨텐츠 기획

· 리포트 컨텐츠 설계

· 제품 및 서비스의 판매 촉진에 목적이라면 상품을 강화시키는 컨텐츠 기획
· 인사이트 리포트 판매가 목적이라면 판매 대상을 선정하고 선정 대상의 Needs에 맞춘 컨텐츠 기획

③ 리포트 개발

· BI 툴 활용 리포트 개발

태블로

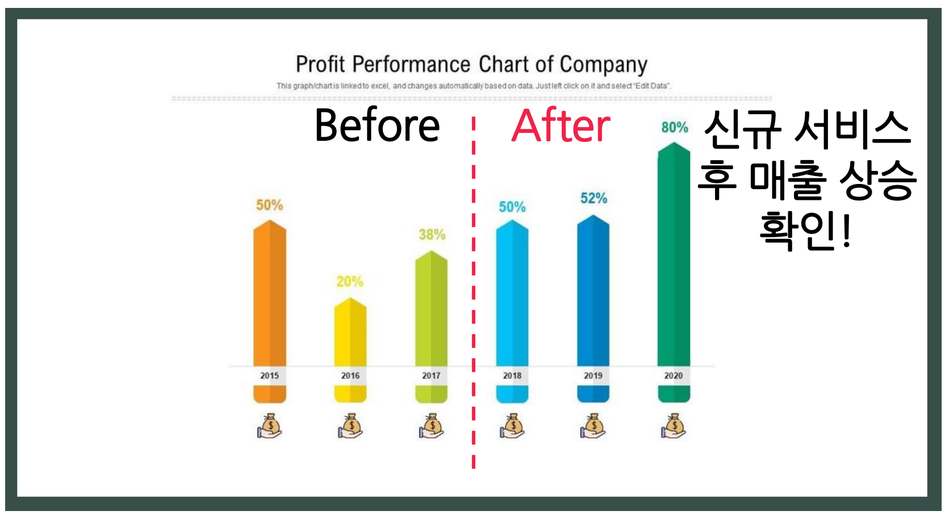
Splunk

Power BI

· 리포트는 계속 활용 될 수 있도록 BI 툴을 이용해서 개발
· BI 툴 사용 시 데이터를 변경하면 동일한 Frame에 리포트 반복 생산 가능의 장점

④ 영업 활동 및 성과 측정

· 영업 및 매출 모니터링



· 영업 활동과 함께 리포트를 고객에게 전달하고 매출 모니터링을 통해 효과 측정
· 인사이트 리포트 판매의 경우 판매 건수 및 매출 측정

기대효과

「금융」

예측 모델(Black box model) 설명력 확보하기

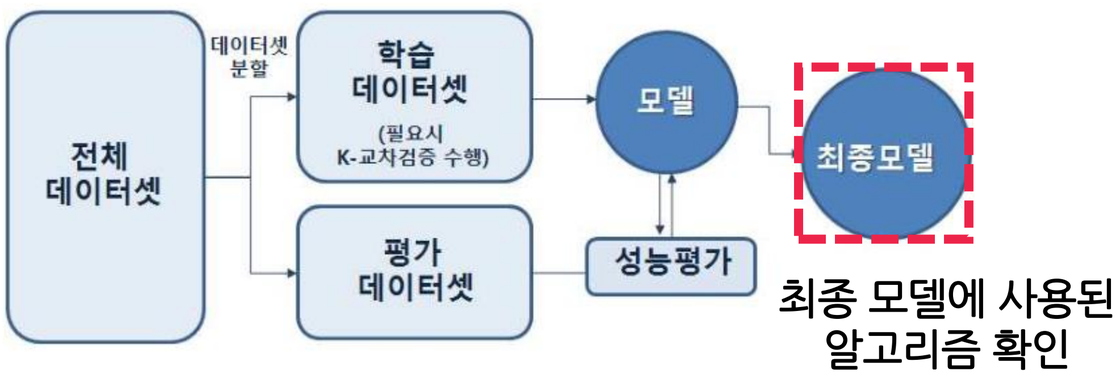
프로젝트 개요

· 선형 회귀, 로지스틱 회귀를 제외한 대부분의 머신러닝, 딥러닝 알고리즘은 Black box 모델인 경우가 많다.
현업에서 모델을 활용하기 위해서는 설명이 필요하며 Black box 모델의 설명력을 확보하기 위한 분석은 꼭 필요하다.

개념 설계

❶ 모델링 결과 확인

· 최종 모델 확인

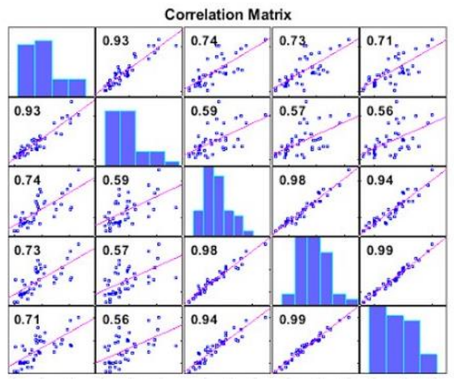


최종 모델에 사용된
알고리즘 확인

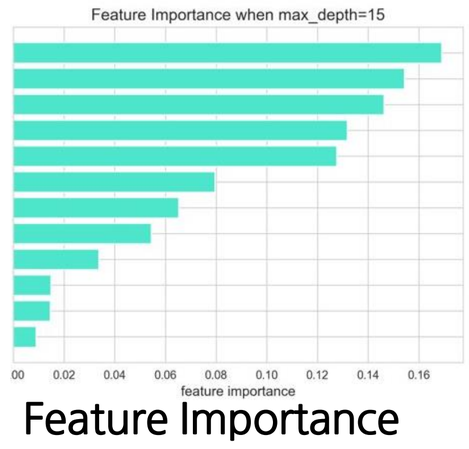
· Hyper parameter 튜닝을 통해 완성된 최종 모델 확인
· 최종 모델에 사용된 알고리즘을 파악
→ 사용된 알고리즘에 따라 해석방법이 상이하기 때문

❷ Feature IMP 분석

· 중요 변수의 선형 관계 파악



Scatter plot



Feature Importance

· 회귀 계열 알고리즘 사용 시 회귀계수를 통한 설명력 확보
· Tree 계열 알고리즘 사용 시 Feature IMP로 도출된 중요 변수에 대해 산점도(Scatter) plot 및 선형관계 파악

❸ 추가 방법 탐색

· XAI 알고리즘 활용 설명력 확보

LIME

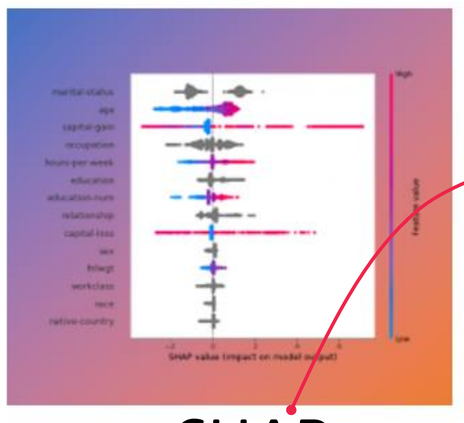
SHAP

PDP PLOT

· XAI ? 해석할 수 없는 Black box 모델을 해석하기 위한 방법론
· 선형관계가 없다면, XAI알고리즘을 활용해 추가 설명력 확보
· XAI 대표 알고리즘 (LIME, SHAP, PDP …)

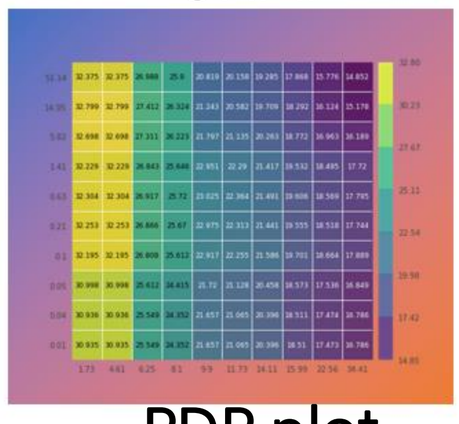
대표 XAI 알고리즘

예측에 대한 각 변수에 기여도를 계산하여 왜 그렇게 나왔는지 해석할 수 있는 방법론



SHAP

LIME



PDP plot

관심 대상인 변수와 Target간에 어떠한 관계가 있는지 plot으로 확인하는 방법

기대효과

모델 설명력 확보

현업 도메인 전문가 미팅 및 피드백

→ 모델 활용 고객 이해도 상승

→ Black box 모델의 약점 보안

「금융」

RFM 모델 활용 우량 고객 정의

프로젝트 개요

· 우량 고객(VIP) 고객 관리를 위해 RFM 기법을 활용하여 우량 고객의 기준을 수립하고 우량 고객 관리를 통해 매출을 향상 시키고자 함

개념 설계

① RFM 변수 개발

· RFM 활용 우량 고객 데이터 Set 구축

· 최근성(Recency), 구매빈도(Frequency), 금액(Monetary) 3가지 지표들을 통해서 고객 점수부여 및 등급화

② 우량 고객 정의

· 우량 고객 선정 기준 수립

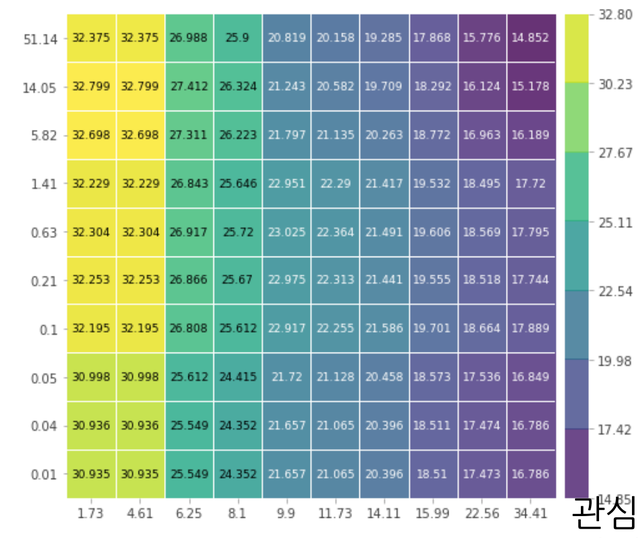
90점 이상일 시 Very Strong(우량고객)으로 정의

고객번호	R	F	M	RFM	Grade
100034	-	-	-	98	Very Strong
100012	-	-	-	70	Strong
100001	-	-	-	60	Normal

· R, F, M 변수 값을 모두 더해서 RFM 변수를 생성
· 고객별 RFM 점수 계산 및 점수 구간에 따른 서비스 등급 부여

③ 우량 고객 & 매출 분석

· 우량 고객과 매출 간 상관관계 분석 및 영향도 파악



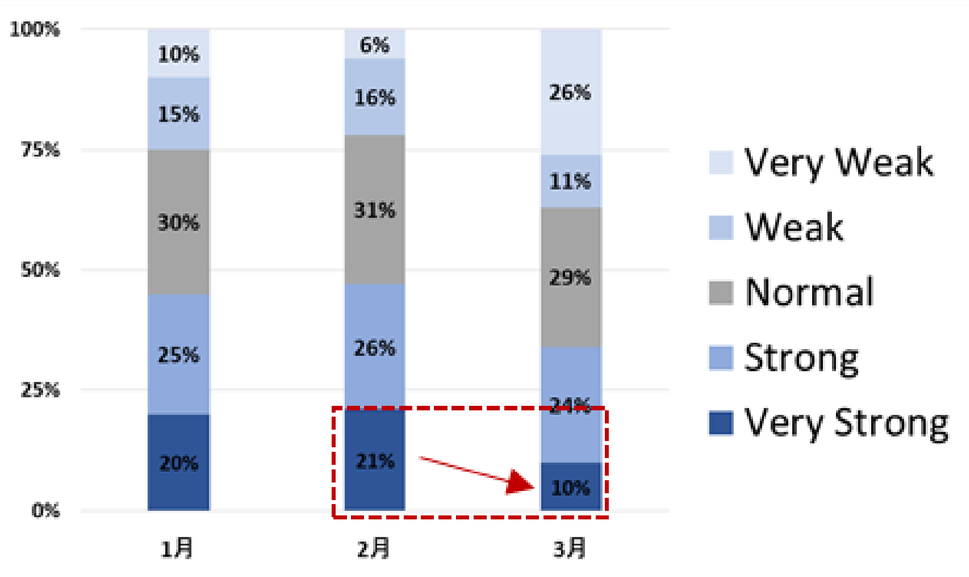
RFM 점수와 매출간에 PDP plot을 통해 상관관계 파악

※ PDP plot
관심 대상인 변수와 Target간에 어떠한 관계가 있는지 plot으로 확인하는 방법

· 우량고객과 매출간에 상관관계 분석을 하고 분석결과 상관관계가 높다면 우량고객을 관리해야 한다는 논리를 확보

④ 관리 체계 구축

· 우량 고객 관리 방안 수립



· 각 등급별 고객 비중 월 별 모니터링을 통해 이슈 파악
· Very Strong(우량고객) 비중 감소 원인 파악

기대효과

정량화된 고객 지표 선정 및 활용 가능

등급별 맞춤 관리 방안을 통해 충성 고객 증대

등급 모니터링을 통해 Issue 사전감지 및 대응

「제조」
품질 중요인자 도출

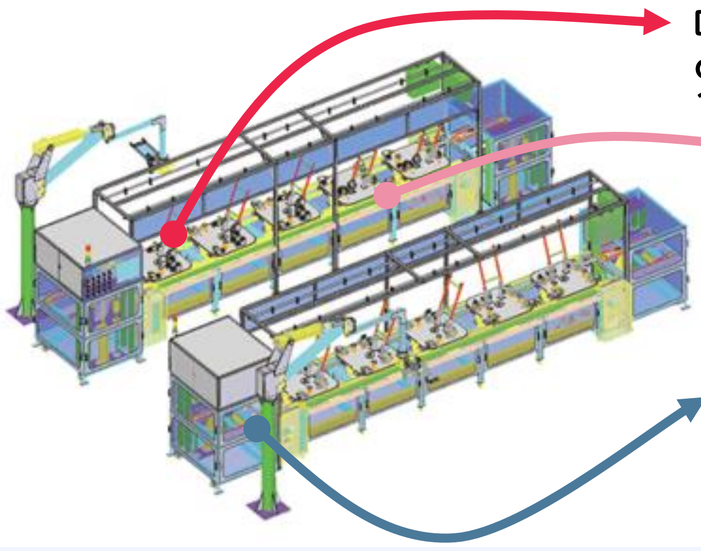
프로젝트 개요

· 품질 이슈 발생으로 생산량 감소, 폐기 비용 발생의 문제를 해결하기 위해 모델링을 통해 품질 중요인자를 도출하고 중요 인자 모니터링 시스템을 구축하려고 함

개념 설계

① 데이터 수집

· 데이터 수집 방법(I/F) 결정 및 수집



· 품질 데이터 적재 DB 확인 및 실시간 수집 시스템 구축
· 실시간 수집 체계 구축 전 Dump 형태로 데이터 전달 받은 후 분석 진행

② 모델링

· 머신러닝 알고리즘 활용 예측 모델링

learn Random Forest

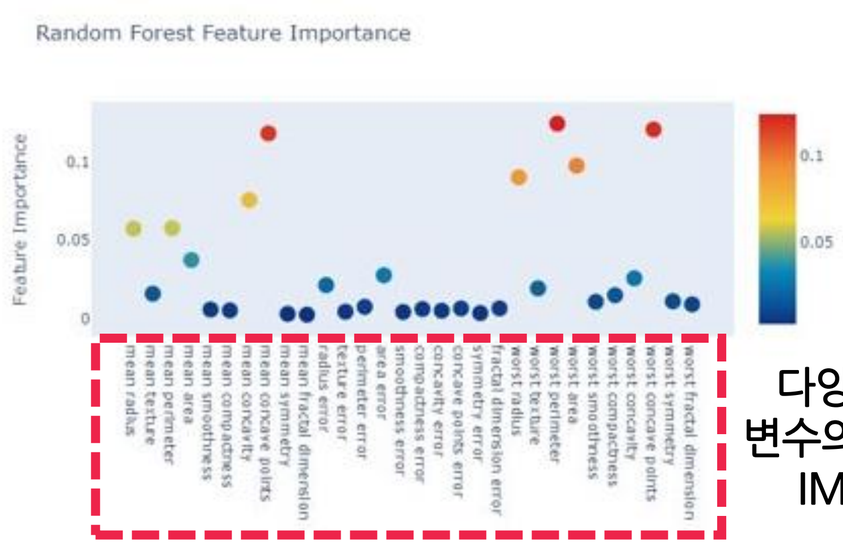
LightGBM

dmlc XGBoost + scikit learn

· 생산품의 무게나 성능 같은 연속형 Target을 예측할 때는 회귀 계열 알고리즘을 사용
· 양품과 불량품을 예측할 때는 분류(Classification) 계열 알고리즘 사용

③ 중요인자 도출

· Feature IMP 분석 및 중요인자 도출

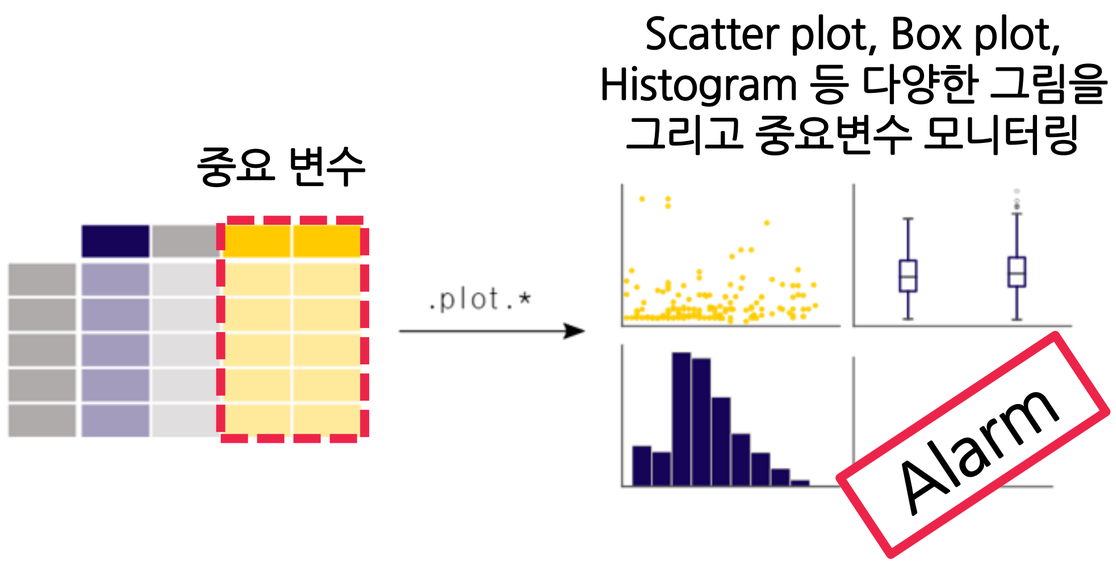


다양한 공정 변수의 Feature IMP 비교

· 모델링 완료 후 회귀 계수, Feature IMP 활용 중요인자 도출
· 중요인자 Scatter plot을 통한 모델 설명력 확보

④ 모니터링 시스템

· 중요인자 관리를 위한 모니터링 시스템 구축



· 품질 중요인자 관리를 위한 공정 변수 모니터링 시스템 개발
· 중요변수가 특정 수치를 넘어갈 시 경고를 전송하여 점검할 수 있도록 운영 체계 구축

기대효과

주요 공정 변수 모니터링 전/후 생산량 비교

주요 공정 변수 모니터링 전/후 불량률 비교

주요 공정 변수 모니터링 전/후 폐기비용 비교

→ 생산량 증가 / 불량률 감소 / 폐기비용 감소

「제조」

장비 사전 이상진단을 통한 고장 방지

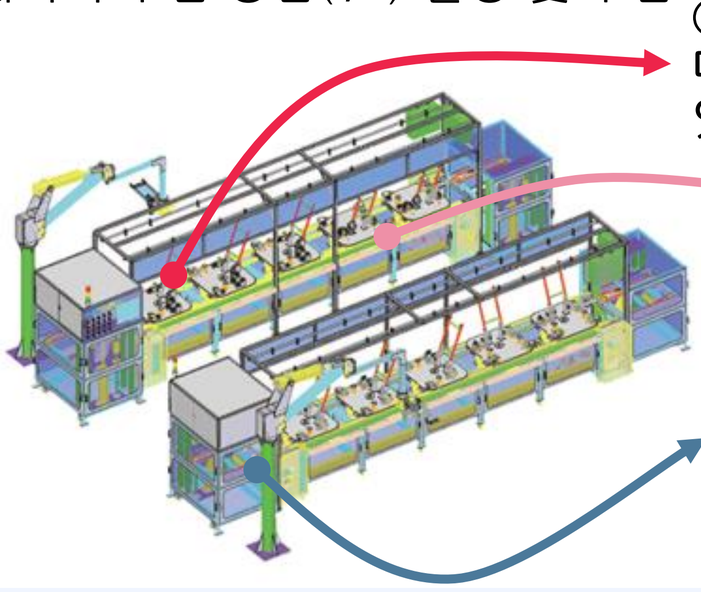
프로젝트 개요

· A공장은 잦은 고장으로 인하여 산발적으로 Line stop이 발생하고 있다. Line stop으로 인해 생산량 감소, 폐기비용이 증가하고 있으며 이를 해결하기 위해 장비 센서데이터 활용 이상증상을 예측하고 사전 점검을 통해 고장을 방지하려 함

개념 설계

① 센서 데이터 수집

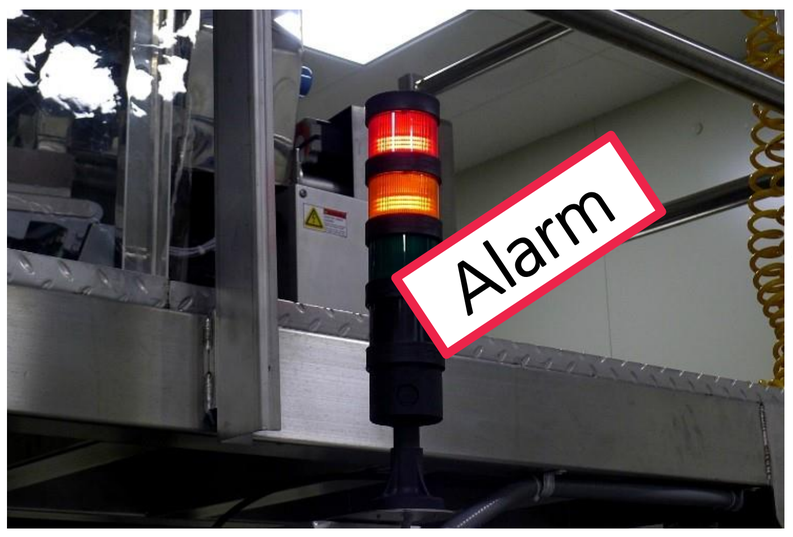
· 데이터 수집 방법(I/F) 결정 및 수집



· 공정별 실시간 센서 데이터 수집 체계 구축
· 실시간 수집 체계 구축 전 Dump 형태로 데이터 전달 받은 후 분석 진행

② 이상 증상 정의

· 고장 나기 전 이상 증상에 대한 정의

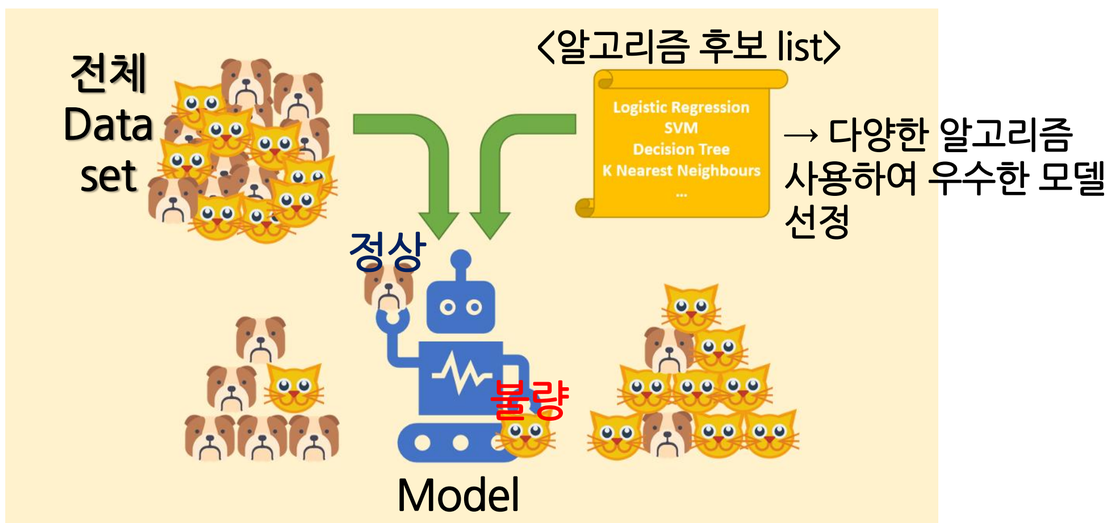


고장 발생 전 수집되는 알람코드를 활용하는 것도 Tip

· 현장의 엔지니어와 미팅을 통해 사전 이상증상을 정의
· 고장 발생하기 전 센서데이터의 Trend를 분석하여 데이터 기반으로도 이상을 정의할 수 있음

③ 예측 모델링

· 이상증상을 예측하기 위한 Classification(분류) 모델링



· Binary Classification(이진분류) Model을 생성
· 다양한 Tree 계열의 알고리즘 사용 추천

④ 성능 평가 및 현장 적용

· 중요인자 관리를 위한 모니터링 시스템 구축

· 고장이 발생하는 것을 막기 위한 활동이므로 Recall을 중점적으로 평가
· 현장에 파일럿 테스트 운영 및 세세한 운영사항 정비

기대효과

주요 공정 변수 모니터링 전/후 생산량 비교

주요 공정 변수 모니터링 전/후 Line stop 횟수

→ 생산량 증가, Line stop 감소