데이터 분석을 직접 실행하고 검증할 수 있는 방안

현업에서 임직원들이 직접 데이터 분석을 실행해 보고 아이디어를 얻어서 업무에 활용할 수 있도록 적용 가능한 use case 99개와 실행해 볼 수 있는 시스템을 제공하고자 한다.

서론: 데이터 분석이 업무에 제대로 활용되지 않는 배경

■ 1) 데이터 분석을 위한 활용 방법(use case)을 모름

- 우리가 업무 중에 발생하는 수많은 데이터를 이용하여 업무에 활용할 수 있는 방법들이 많이 있음에도 제대로 활용하지 못하는데, 가장 큰 이유는 어떤 목적으로 어떤 데이터를 활용 어떻게 활용하는 지에 대한 아이디어가 없기 때문이다.

* use case란 데이터 분석으로 어떤 목적을 위해 활용할 방안으로 어떤 데이터를 어떤 방법으로 분석하여 어떤 목적에 활용할 수 있는 방법을 구체적으로 제시하는 것을 말함

- 분석을 위한 아이디어가 없으니 업무에 활용할 수 있는 중요한 데이터를 가지고 있다고 하더라도 가치있게 활용하지 못하고 있다.

■ 2) 데이터 분석 아이디어(use case)가 있다고 하더라도 이 분석을 위한 방법(개념, 통계, 딥러닝 등 분석 방법)을 모를 경우 활용하기가 쉽지 않음

- 이번에 소개할 "생산/품질의 양품 판정"에 로지스틱 분석을 활용할 수 있는 Use case에 대한 내용을 알게되었다 하더라도, 만일 동일한 생산 환경과 유사한 데이터가 아니라면 Use case에 소개한 개념만을 가지고 그대로 적용하거나 아니면 본인의 현장에 적용할 수 있는 적합한 아이디어를 도출하기가 쉽지 않다.

- 결국 Use case를 알게되더라도 정확한 이해를 위해선 해당 Use case에 적용한 분석 방법에 대한 최소한의 지식이 필요하다. 로지스틱의 분석의 경우 Odds, 로짓, 로지스틱 함수, ROC에 대한 개념을 이해할 필요가 있다. 만일 로지스틱 분석을 위해 딥러닝 기법을 사용했다면 해당 딥러닝에 대한 최소한의 개념을 알아야 정확히 Use case를 이해했다고 할 수 있을 것이다.

- 그리고 Use case를 이해하고 이제 이것을 자신의 업무나 현장에 적용할 수 있는 최소한 아이디어가 있다면 본인이 이것을 수행하지 못할 경우엔 외부 전문가를 활용 도움을 받을 수 있지만, 문제는 이것의 성공에 대한 확인이 아직은 부족하기 때문에 제일 좋은 것은 자신이 직접 실행해 보고 성공에 대한 확신이 있을 경우 외부 도움을 받는 것이 좋을 것이다.

■ 3) 분석을 위해 필요한 분석 Tool(통계, 딥러닝 등 분석 Tool)이 없거나 사용하기가 쉽지 않음

- 앞서 설명한 바와 같이 자신이 찾던 Use case를 찾았고 해당 개념에 대한 이해가 있다 하더라도 외부에 해당 내용을 맡기기 전에 자신이 먼저 가능성을 검증하고 싶을 경우가 있다.

- 문제는 직접 자신의 데이터를 가지고 어떻게 해당 개념을 스스로 어떤 방식으로 실행해 볼 것인 것 하는 문제에 부딪히게 된다.
- 보통 이러한 데이터 분석을 위해선 관련 Tool이 고가이고 사용을 위해선 사전 지식도 많이 필요한데, 특정 업무에 아주 특별한 Case에 한시적으로 사용을 할뿐이데 해당 Tool을 구매하고 관련 지식을 갖춘다는 것은 불가능한 가정이다.
- 따라서 이러한 목적에 부합되게 가능성만이라도 간단하게 확인할 수 있는 쉬운 시스템은 없는 것인가? 하는 것이다.

■ **이러한 문제를 극복하기 위해 영업, 구매 등 모든 업무에 적용 가능한 use case를 99개로 정리하고 스스로 실행해 볼 수 있는 아이디어를 체계적으로 시스템으로 정리하였으며 use case 중 하나 (로지스틱 분석)를 소개합니다.**

use case-1: 개념소개) 로지스틱 분석

로지스틱 분석이 여러 업무에 사용 가능한데, 여기선 영업과 생산에 적용할 경우에 가능한 use case 개념을 소개하고, 생산/품질의 양품 판정에 사용할 수 있는 Case에 대하여 보다 자세히 소개하고자 한다.

■ 영업부서에서 특정 마케팅을 위한 타겟 고객 선정

- 마케팅을 계획하고 있는 경우에 마케팅 프로그램의 비용이 클 경우 불특정 다수에게 해당 마케팅 프로그램을 진행하는 것은 상당한 비용을 수반하게 된다.
- 어떻게 하면 마케팅 효과를 극대화하기 위한 마케팅 대상 고객을 찾을 것인가하는 것은 아주 중요한 부분이다.
- 타겟 고객을 선정하기 위해 필요한 데이터 분석 내용은 다음과 같다.

1) 데이터 분석을 위해 독립 변수들은 무엇으로 할 것인가?

2) 개별 독립변수들 중에서 가장 중요한 변수와 최종 판정 변수에 어떤 정도로 영향을 주는가?

3) 독립변수들을 어떻게 정리할 것인가?

4) 로지스틱 분석을 통해 중요한 독립변수들을 찾고 모델의 성능을 판단 마케팅 전략을 수립할 수 있다.

- 분석이 필요한 시점은 신제품을 개발한 후에 어떤 고객을 대상으로 우선 타겟을 할 것인지 혹은 반드시 필요한 정보를 얻기위해 어떤 고객들에게 마케팅 프로그램을 실행할 지를 결정할 때 사용

■ 생산/품질의 양품과 불량을 판정하기 위한 양품 판단 숫치(모델) 결정

- 생산 라인에서 양품과 불량을 결정할 때에 기준이 되는 항목이 있고, 해당 항목의 숫치에 의거하여 양품을 결정하는 경우에 어떤 숫치로 결정하는 것이 가장 효과적인 지를 결정할 수 있다.
- 양품/불량을 결정할 때에 문제가 되는 것은 어떤 판정 항목에 어떤 기준을 적용할 때에 이것이 100% 양품과 불량을 결정할 수 있다면, 이러한 방법이 불 필요하다.
- 하지만, 보통은 기준에 의거 분류를 하지만 일부는 양품임에도 불량으로 판정하고, 일부는 불량임에도 양품으로 판정이 되는 경우가 있다. 이럴 경우 이 방법을 적용해 볼 수 있다.

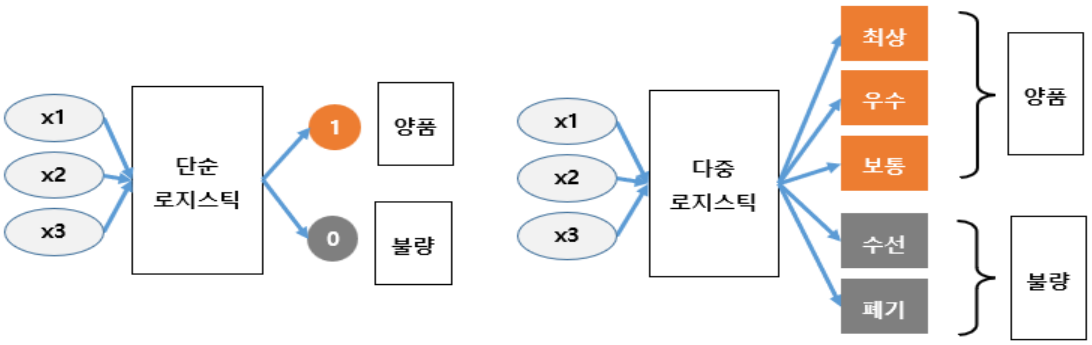
use case-1: 예시) 생산/품질의 양품 판정에 로지스틱 분석 활용

■ 1) use case 목적 및 절차 설명: 데이터 준비, 분석/해석 및 활용 (About)

- 로지스틱 분석은 이항형 로지스틱 회귀(binomial logistic regression)으로 종속 변수의 결과가 (성공, 실패) 와 같이 2개로 분류되며, 다항형 로지스틱 회귀는 종속형 변수가 3개 이상의 카테고리로 분류되는 것을 말하며 각각의 카테고리로 분류될 확률의 합은 1이 된다.

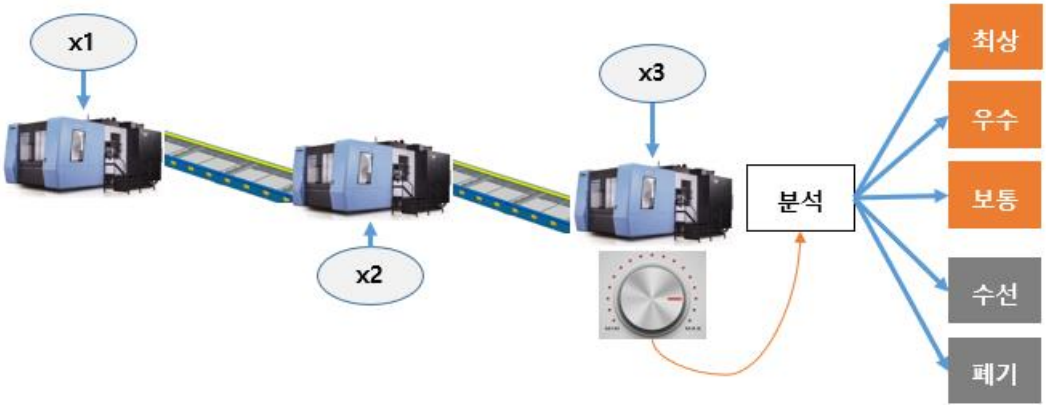
- 이번 use case는 생산 과정에서 여러 독립변수의 상태에 따라 최종 완성품의 품질을 판정하는 것에 로지스틱 분석을 사용하는 경우를 예시를 들고자 한다.

- 실제 현장에서는 다음 그림에서처럼 독립변수가 여러 개가 있으나 여기서 분석은 이해를 쉽게하기 위해 독립변수를 하나만 선정하여 분석할 수 있는 경우를 상정하였다.



(이미지 차이: 2개 범주로 분류하는 단순로지스틱 분석, 여러 범주로 분류할 수 있는 다중 로지스틱 분석)

- 실제 환경에서는 독립변수가 여러 개 있을 수 있고 종속변수의 범주가 여러 개 가 있을 수 있다. 이럴 경우에도 통계를 이용한 회귀분석 기법을 사용할 수 있거나 아니면 딥러닝 방법을 사용 할 수도 있다. 여기에선 통계적 기법을 이용한 단순 로지스틱 분서의 예시를 들고자 한다.



■ 2) 데이터 준비 및 분석 목적

- use case 목적에 따라 준비할 데이터들은 다양하다. 실제 현실에서는 독립변수들이 여러 개가 필요할 수 있기 때문에 아래 예시처럼 준비할 데이터에 대한 개별 속성에 대한 기준을 잘 정리해 두어야 한다.

독립변수	연속변수	Age
	비연속변수	Sex(1=남, 2=여) 비만지수(1= 00이하, 2=0-9.9, 3=10-19.9, 4=20-29.9, 5= 연령군(20=30미만, 30=30대, 40= 40대, 50= 50대, 60=6 혈당간격(1=1000이하, 2=100-120, 3=1200이상) 감마지티피(1=290이하, 2=30-59, 3=60-89, 4=900이상) 중성지방(1=1490이하, 2=150-199, 3=200-249, 4=2500이상
종속변수	이분된 비연속변수	지방간(1=정상, 2=지방간)

(Source: https://blog.naver.com/PostView.nhn?isHttpsRedirect=true&blogId=jowoon3&logNo=70005801982)

- 이해를 쉽게하기 위해 여기 Case에선 독립변수를 하나만 사용한 자료를 사용하였으며, 독립변수 하나(X)에 대하여 목표/결과 변수(Y)는 양품(Pass), 불량(Fail)으로 판정하는 자료이다.

- 아래 엑셀에 정리된 자료는 판정 기준(X)에 따라 분류된 양품의 개수와 불량품의 개수를 정리한 것이다. 현장의 양품과 불량을 판정하는 기기의 수치에 따라 아래와 같이 양품과 불량품의 개수가 분류되었다. 아래 데이터의 의미는 어떤 제품의 X의 값이 24인 제품이 53개가 발생했는데 그 중에서 양품(Pass)는 52이며 불량(Fail)이 1개 발생했다는 의미이다. 27의 경우는 양품(Pass)는 45이며 불량(Fail)은 3개 발생했다는 것이며 따라서 만일 판단 기준인 X의 값을 26으로 설정하면 25이하는 모두 양품으로 분류하고 그 이상은 모두 Fail로 분류를 한다는 의미이다.

X	Fail	Pass	X	Fail	Pass	X	Fail	Pass
24	1	52	28	6	44	32	55	11
25	1	50	29	7	35	33	74	4
26	1	46	30	33	24	34	77	3
27	3	45	31	45	20	35	25	2

*** 분석 목적**

- 이번 use case는 앞서 설명한 데이터를 기준으로 로지스틱 분석을 통하여 독립변수 X의 값을 어디로 결정하는 것이 양품과 불량을 구분하는 가장 최적인 지점(X 값)을 찾고자 하는 것이다.

  판단하기 위한 독립변수(X)를 어떤 지점으로 할 것인가에 따라 양품과 불량으로 분류되는 개수가 변화를 하기 때문에 최적의 지점을 찾는 것이 중요하다.

■ 3) 분석 절차 (상세한 내용은 생략)

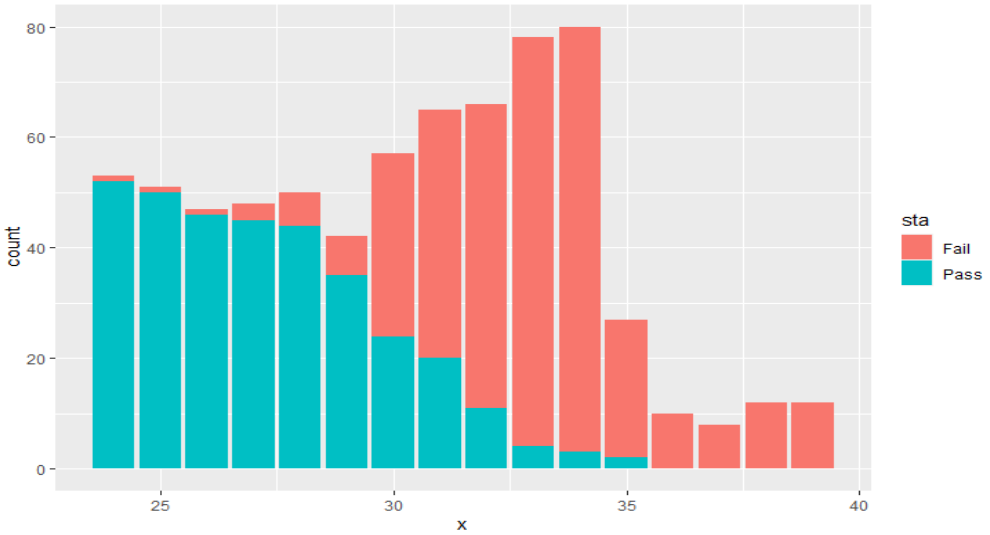
- 오픈 소스 통계 프로그램 R 이용 데이터에 대한 기술 통계를 통하여 개략적인 데이터의 구조와 일반적인 통계분석을 실시한다.
- 로지스틱 분석 함수를 이용 준비한 데이터를 분석하고, 결과 값의 해석을 통하여 해당 모델의 성능을 판단한다.
- 최적 Cut-off Value(Point)를 알기위한 ROC Curve 로 확인한다.
- 결과 값 확인 후 실제 업무에 적용하기 위한 방안을 수립한다.

■ 4) 데이터 구조 및 Summary (Structure & Summary)

- 분석을 위해 엑셀 자료를 upload 한 후에 전체 구조를 확인한 내용

```
'data.frame':   32 obs. of  3 variables:
 $ Var1: Factor w/ 16 levels "24","25","26",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Var2: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Freq: int  52 50 46 45 44 35 24 20 11 4 ...
```

- 판정 기준에 따른 양품(Pass)과 불량(Fail) 개수 변화



- 판단기준 X에 따른 양품(Pass)/불량(Fail)의 모습: 어떤 기준으로 하더라도 일부는 양품과 불량이 섞여있음을 볼 수 있다. 이럴 경우 X의 어느 수준에서 판단할 것인지가 중요하게 된다.
- X축의 판단 기준을 우측으로 옮길수록 양품으로 판단하는 개수는 늘어나고 더불어 불량품임에도 양품으로 잘 못 판단하는 경우도 높아진다. 극단적으로 가장 우측의 숫치로 판단할 경우 모든 제품을 양품으로 판단하게 된다.
- 결국, 판단 기준을 상기 그래프의 어느 지점으로 할 것인가 문제이다.

■ 5) 로지스틱 회귀분석 결과 값 (Analysis Output)

- 분석 결과: 통계분석 오픈소스 (R) 이용한 분석 결과

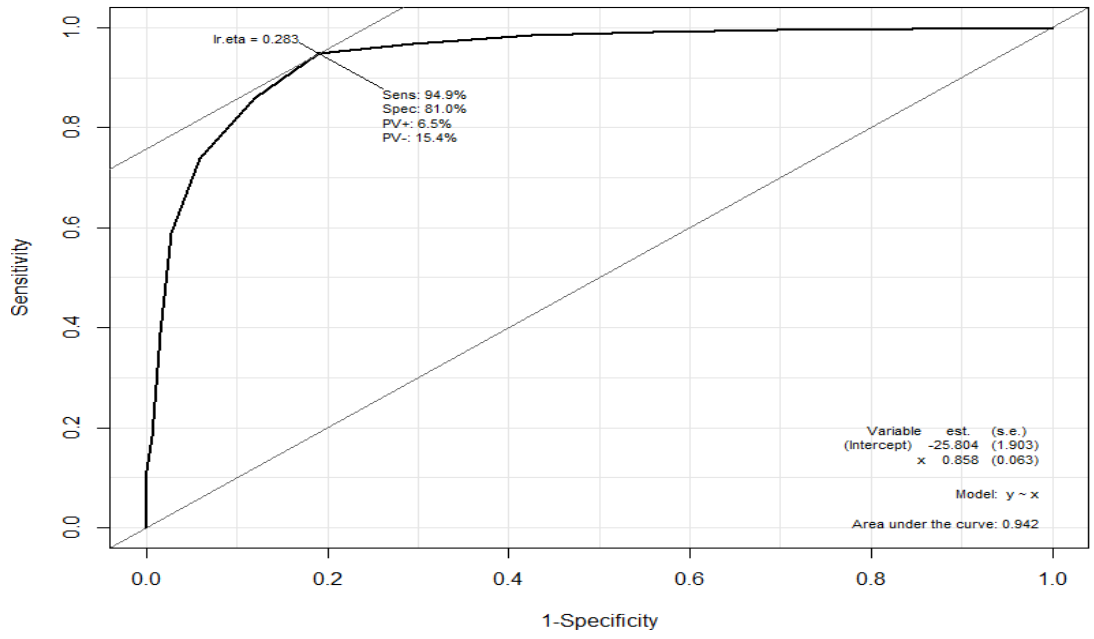
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9098  -0.3707   0.0474   0.3966   3.2318

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.80362    1.90259  -13.56   <2e-16 ***
x             0.85778    0.06261   13.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 977.09  on 705  degrees of freedom
Residual deviance: 426.42  on 704  degrees of freedom
AIC: 430.42

Number of Fisher Scoring iterations: 6
```

- Y 예측값을 위한 회귀식: -25.80362 + 0.85778*X

- 로지스틱 분석 결과: X값이 유의한 수준으로 확인이 되었으며 절편과 독립변수에 대한 계수가 구해졌다.

■ 6) ROC 분석 그래프 (Graph)

- 로지스틱 분석 결과을 이용 ROC 그래프를 확인하여 최적의 Cut-off value 선정: 민감도(Sensitivity)와 특이도(Specificity)의 합이 가장 큰 값을 구한다.



- ROC 그래프는 x축에 FPR (= 1-Specificity)을 기준으로, y축은 민감도를 보여주고 있다. 일반적으로 ROC 커브의 아래 면적이 사각형에 가까울수록 성능이 우수한 모델이며, 해당 커브가 대각선에 가까이 갈수록 해당 모델은 활용 가치가 없게된다.

- 따라서 ROC 커브의 가장 높은 지점이 얼마나 X,Y(0,1) 지점에 가까운지와 해당 커브의 아래 면적(AUC)를 보고 모델의 성능을 판단한다.

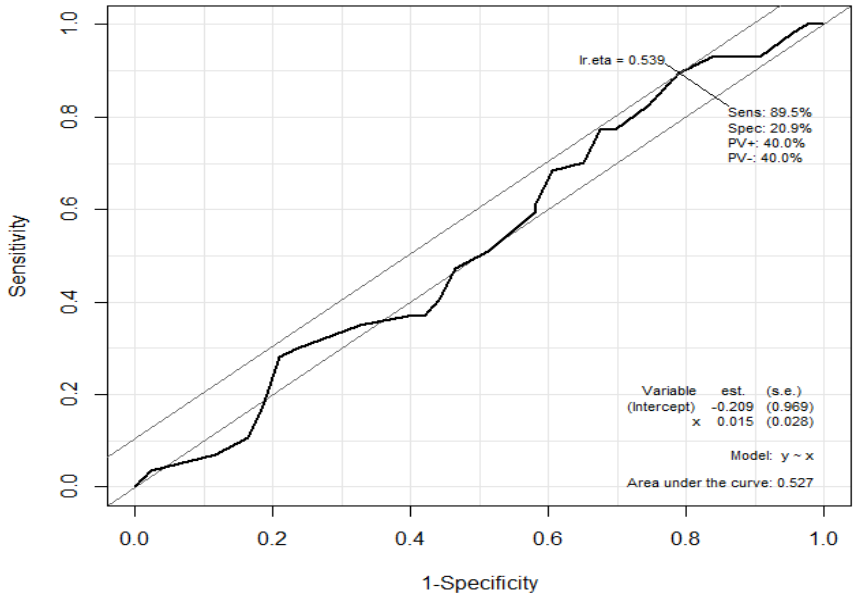
- 여기에 소개된 Use case의 경우 X축의 판단기준이 각각의 판단 모델이 되므로 해당 지점의 모형이 가장 최적의 판단 지점으로 사용할 수가 있다.

- 따라서, 최대한 FPR를 적게하면서 민감도가 높은 지점(Cut-off value)를 찾으면 해당 지점이 최적의 판단 지점인 것이다.

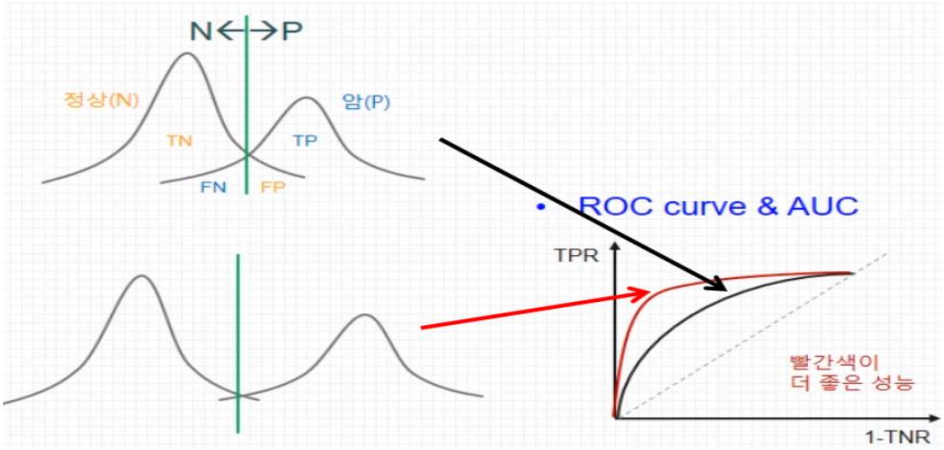
* 여기 예시에서는 독립변수(X)의 값이 30일 때 최적인 것을 확인할 수 있다.

참고) ROC 이해

ROC 이해를 돕기 위해 샘플로 X값을 25~45 사이의 무작위로 숫치를 정하고 해당 숫치별로 Y값(1,0)을 무작위로 배정한 샘플 데이터를 만들어 ROC 그래프를 확인한 결과: 거의 대각선에 가까운 커브 선을 보이고 있다. 이 의미는 어떤 모델을 적용하여 해당 데이터와 같은 판단을 하였다면 해당 모델은 사용 가치가 없음을 의미한다.



- ROC 커브와 AUC(Area Under the Curve)의 이해: 우리가 양품과 불량을 판단을 할 때 사용할 수 있는 여러 모델이 존재할 수 있는데, 여러 모델 중에서 어떤 모델을 선택할 지를 결정할 수 있다. 결국 AUC가 1에 가까운 모델을 선택하는 것이다. 아래의 경우 2개 모델의 ROC 커브가 있는데, 그 중에서 빨간 선의 모델의 AUC 가 다른 모델보다 우수하기 때문에 해당 모델을 선택하게 된다.



(Image Source: https://nittaku.tistory.com/297)

결론: use case를 직접 수행하기 위한 시스템/프로그램 소개

업무에 적용할 수 있는 Use case와 이것을 실행해 볼 수 있는 시스템이 있다고 하더라도 Use case별로 개략적인 이해를 위한 최소한의 설명은 필요하다. 이 부분은 적용할 기업과 웍샵을 통해 해소할 수 있으리라 생각한다.

■ 1. Use case 모음

영업, 구매, 생산 및 재무 등 업무별로 정리한 99개 use case들이 정리되어 있어 업무 담당자(임원/실무자)들은 use case 예시를 참조하여 본인이 직접 분석을 실행해 볼 수 있다.

* 업무 프로세스별로 use case가 준비되어 있음 (데이터 분석을 위한 참조 use case 제공)

- use case: 영업, 구매, 생산, 품질, 자재, 재무 등 업무 단위별로 99개 use case 정리

- use case list

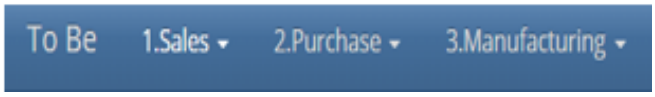
no	Use cases	목적 (현업 업무 중의 애로 사항)	Input data	방법론 및 솔루션	output Business use case 기분 결과 해석
1	일표본 t 검정	영업 : 마케팅 효과 검증 생산 : Lay out 및 공정 개선 전/후의 효과 검증	개선 전, 개선 후의 데이터	Statistics	일표본 t 검정 결과 분석 및 해석
2	EOQ	구매 혹은 자재 입장에서 품목별로 경제적 주문량을 확인하고자 함 * 현실에서는 공식에서 필요로 하는 입력값을 알 수가 없으며, 실제는 더 복잡한 상황	품목별로 기간(일/주)별 주문 수량 실적	Concept	1. 재고 유지비용, 주문 비용의 여러 경우에 따른 EOQ의 what if 분석 2. 재고 유지비용, 주문 비용에 해당 항목을 도출 3. 주문 변동에 미치는 요인 분석 (딥러닝)
4	PSM	기업에서 구매할 모든 품목에 대하여 절감을 위한 영역(대상, 범위) 분석	주요 제품의 BOM 정보와 품목별 구매원가 및 판매 단가	Program	1. Spend analysis로 구매 Spend 분석하여 구매 전략에 활용
5	안전 재고량		최근 1년 동안의 주문, 발주 및 재고관리 비용	Concept	제반 환경 (재고비용, 주문 미 대응 손실 등)을 고려한 적정 재고량 산출

■ 2. 자신이 직접 실행해 볼 수 있도록 시스템화 (분석 Tool 문제 해소)

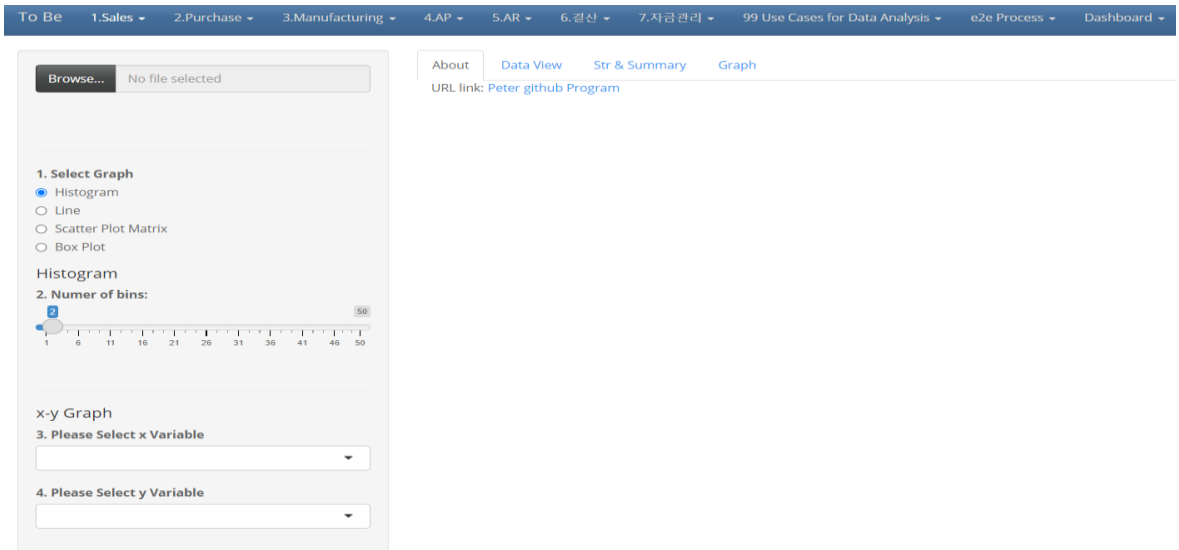
1) 회사 업무 영역 (영업, 구매, 생산, 품질, 자재, 재무) 별로 사용 가능한 use case들이 정리되어 있음

2) 분석을 위한 시스템은 오픈소스로 구성되어 있어 비용 부담이 없음

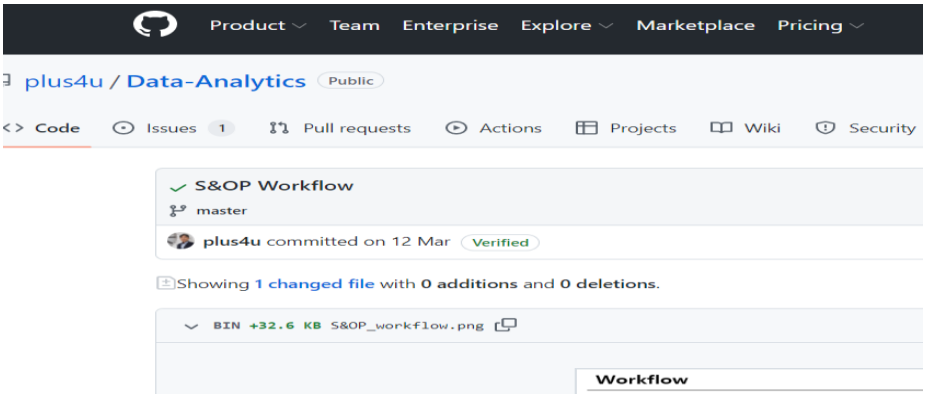
3) use case별로 About, Data View, Structure & Summary, 분석 결과, Graph 조회를 할 수 있음



- 시스템 전체 구성 모습 : 영업, 구매, 생산 및 재무 등 기업내 전체 프로세스별 Use case 정리



(1) About: use case에 대한 목적, 분석방법 및 절차에 대한 설명 (여기에 소개된 개념 설명 자료 (usecase-1 예시)를pdf 자료로 확인할 수가 있다.)

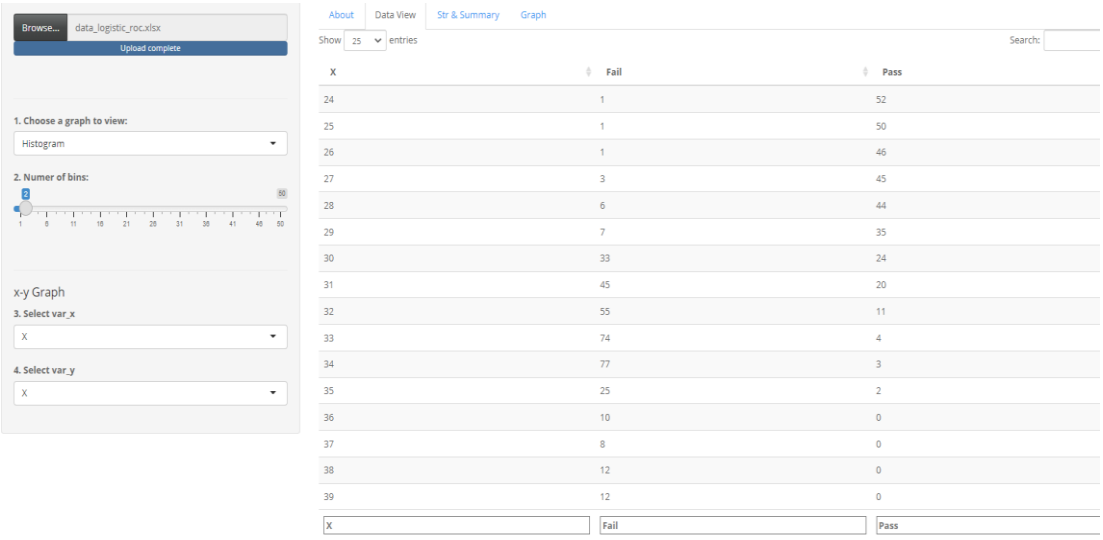


(2) Data upload

- Excel File upload : 데이터 분석을 위한 Source 데이터를 upload할 수 있다.

(3) Data View:

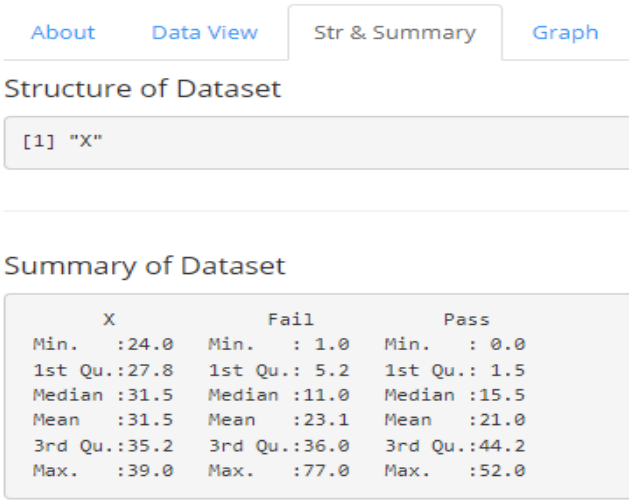
. 분석을 위한 Excel 데이터를 upload 후에 내용 조회 화면



(4) Structure & Summary

- 데이터를 읽어와서 데이터의 전체 구조와 기술통계(Data Summary) 내용 조회

- upload한 데이터의 기술통계



. 분석 결과: 분석결과는 사용한 분석 프로그램에 따라 해당 결과를 확인할 수 있음.

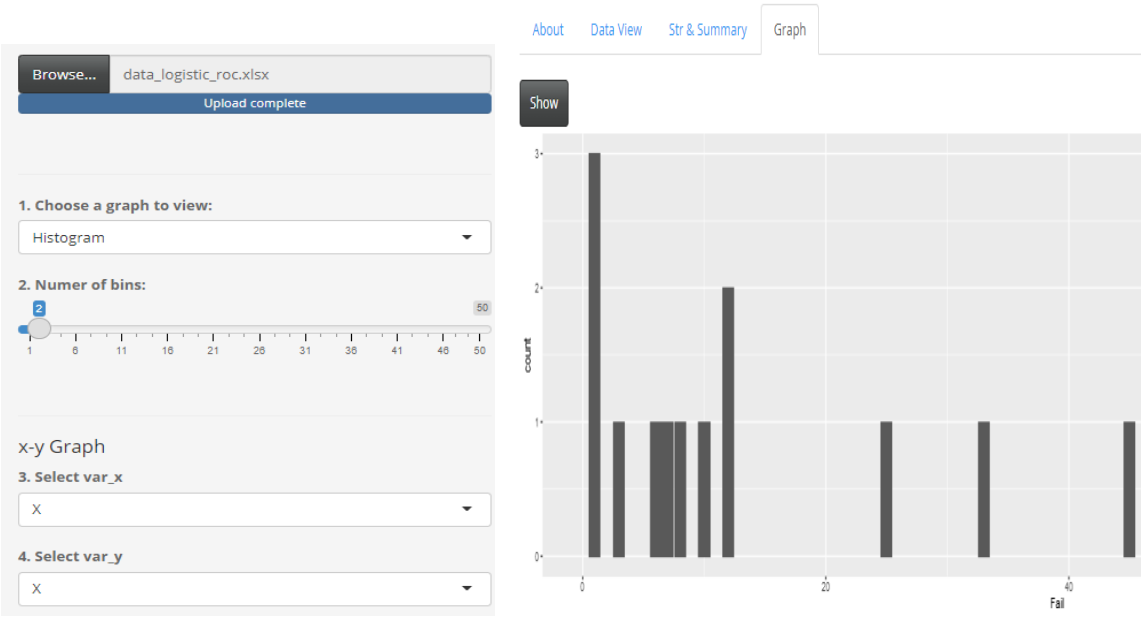
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9098  -0.3707   0.0474   0.3966   3.2318

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.80362    1.90259  -13.56   <2e-16 ***
x             0.85778    0.06261   13.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 977.09  on 705  degrees of freedom
Residual deviance: 426.42  on 704  degrees of freedom
AIC: 430.42

Number of Fisher Scoring iterations: 6
```

. Graph

. Use case에 따라 필요한 다양한 분석 그래프를 조회할 수 있다.

. Graph 선택: Histogram, Line, Scatter, Box Plot

- 분석에 필요한 x, y 축 변수 선택



요약:

■ Use case를 참고하여 현업에서 즉시 사용이 가능한 데이터 아이디어 도출:

- Use case의 제공된 개념설명과 분석 시스템을 통하여 직접 사용하면서 이해를 높이고 적용 가능성에 대한 판단을 스스로 할 수 있도록 한다.

■ 다만, Use case와 실제 적용 환경이 다르기 때문에 해당 Use case에 대한 최소한의 개념과 이해를 위한 협업은 필요할 수 있다.

- use case해석 결과에 대한 이해를 높일 수 있다.
- 이번 use case의 경우 각 분류 기준에 따른 cut off value(point)가 다음처럼 계산될 수 있다.

X	분류 기준	Fail	Pass	sensitivity	false positive	cut off point
24	1	1	52	1	1	1.0000
25	2	1	50	0.997	0.845	1.1521
26	3	1	46	0.995	0.696	1.2982
27	4	3	45	0.992	0.560	1.4324
28	5	6	44	0.984	0.426	1.5582
29	6	7	35	0.968	0.295	1.6729
30	7	33	24	0.949	0.190	1.7582
31	8	45	20	0.859	0.119	1.7404
32	9	55	11	0.738	0.060	1.6783
33	10	74	4	0.589	0.027	1.5624
34	11	77	3	0.389	0.015	1.3743
35	12	25	2	0.181	0.006	1.1751
36	13	10	0	0.114	0.000	1.1135
37	14	8	0	0.086	0.000	1.0865
38	15	12	0	0.065	0.000	1.0649
39	16	12	0	0.032	0.000	1.0324
	합계	370	336			

- 이 의미는 만일 분류기준을 24(1번)으로 선택하면 생산된 제품의 370개를 모두 Fail로 판정하게 되어 불량품은 100% 찾게되나, 양품으로 분류될 수 있는 336개도 모두 불량으로 판정한다는 의미이다.
- 만일 39를 기준으로 할 경우 16번 기준 이하는 모두 양품이고 오직 12개만 Fail로 판정하며, Fail(불량) 370 중에서 12개만 분류할 수 있다. 하지만 양품 336은 모두 양품으로 분류하게 된다.

■ Use case를 참조한 실제 자신의 현장에 적용하기 위한 구체적인 방안 협의