

가을 학 기 A I T e r m P r o j e c t

불량품에 민감한 불량품 예측기

201502077 유 예 지
201601989 김 진 섭
201601998 박 경 신

CONTENTS

01

Introduction

- 프로젝트 주제
- 프로젝트 소개
- 프로젝트 기획

02

Problem Definition

- 문제 정의
- 해결 방안 제시

03

Data Explanation

- 데이터 소개
- 데이터 정제

04

Modeling Method

- 변수 선택
- Naïve Bayes
- 데이터 생성
- MLP

05

Model Assessment

- Confusion Matrix
- Accuracy 비교
- Recall 비교
- Precision 비교

06

Conclusion

- 팀 프로젝트 결과

1) 프로젝트 주제

- 작업자 데이터를 활용한 불량품에 민감한 불량품 예측기

2) 프로젝트 소개

- 공정 데이터의 양품은 항상 불량품에 비해 많음
- 클래스의 비율 차이가 크면, 제대로 된 모델을 생성 불가
(클래스 비율이 높은 데이터로 분류)
- 따라서, 불균형 데이터인 공정 데이터를 활용해 불량품을 민감하게 분류하는 모델을 생성하는 프로젝트 진행 계획

3) 프로젝트 기획

1. 변수 선택

- p-value와 RanfomForest를 활용해 변수 선택

2. 불량품 데이터 생성

- GAN 활용 : 학습에 많은 어려움이 있어서 해결을 못함.
- SMOTE 활용 : Over-Sampling을 활용하여 데이터 추가

3. 모델 생성

- Naïve Bayes : Prior를 활용해 불량품을 선택할 수 있도록 모델 생성
- MLP : 균형 데이터로 변경 후 MLP 모델 학습

> Problem Definition

문제 정의

- 추진 배경
 - 머신 러닝 기반 품질 문제 사전 예측 연구는 활발히 진행되고 있지만, 특정 분야(반도체)에서만 이루어지고 있으며 **작업자에 의존하는 조립 산업 적용을 위한 연구는 부족함.**
- 기술적 요인
 - 모델을 생성할 때 많은 변수 중 어느 변수를 사용할 지? (**과적합**)
 - **불균형 데이터**를 활용하여 학습을 할 경우, 비율이 높은 클래스로 선택할 확률이 높음 (**모델 생성에 제한**)

해결방안 제시(기술적 요인)

- 변수 선택
 - 로지스틱 회귀분석을 통해 얻은 **유의 확률(p값)**을 활용 ($\alpha = 0.05$ 일 경우, $\alpha > p\text{-val}$ 이면 연관성이 있음)
 - **RandomForest**분석으로 importance가 적은 변수는 삭제
- 불균형 데이터
 - **Synthetic Minority Over-sampling Technique**를 활용해 데이터의 균형을 맞춤.

- | baeCd | pGrpHn | qatDescr | sn | baeCase | stefm | abstExt | msDate | makeStk | veh | 기타정보 | 작성시간 | 작성데이터 | DB서버 | No | 법인명 | 이름 | No | 법인명 | 이름 | 자번호 | 부서 | 생년월일 | T08 | T020 |
|-------|--------|----------|----------|---------|-------|---------|---------------------|----------|------------|------|------|----------|----------|-------|-----|----|----|---------|---------------------|----------|------------------|------------|-----|------|
| C0112 | MAIN | 미국보훈 | VN12918A | 가운-양식 | 베트남 | 50 | 20181016 | 50 | HR | | | | | | | | No | 법인명 | 이름 | 자번호 | 부서 | 생년월일 | 남 | 여 |
| | | | | | | | | | | | | | | | | | No | Tên CTY | Họ và tên | Mã NV | BP | Ngày sinh | Nam | Nữ |
| | | | | | | | 2018-10-15 08:52:28 | autodata | manualdata | | | | 20180523 | | | | 1 | 베트남이력 | Nguyễn Thị Thoan | 11100011 | Phòng hành chính | 1982-08-14 | | |
| | | | | | | | 2018-10-15 18:47:54 | autodata | | | | 25180455 | | | | | 2 | 베트남이력 | Vũ Năng Hải | 11100012 | Phòng hành chính | 1977-04-15 | | |
| | | | | | | | 2018-10-12 17:38:25 | autodata | | | | 25080041 | 1003 | 베트남이력 | | | 3 | 베트남이력 | Nguyễn Thị Xuân | 12120010 | Phòng kế toán | 1983-02-03 | | |
| | | | | | | | | | | | | | | | | | 4 | 베트남이력 | La Thị Anh | 11100009 | Phòng kế toán | 1988-03-20 | | |
| | | | | | | | | | | | | | | | | | 5 | 베트남이력 | Nguyễn Mạnh Thắng | 13090011 | Phòng kế toán | 1985-11-08 | | |
| | | | | | | | | | | | | | | | | | 6 | 베트남이력 | Nguyễn Văn Quang | 15090007 | Phòng sản xuất | 1985-06-09 | | |
| | | | | | | | | | | | | | | | | | 7 | 베트남이력 | Nguyễn Hải Quỳnh | 15110015 | Phòng sản xuất | 1982-12-10 | | |
| | | | | | | | | | | | | | | | | | 8 | 베트남이력 | Nguyễn Thị Thu Hiền | 15080005 | Phòng sản xuất | 1983-02-26 | | |
| | | | | | | | | | | | | | | | | | 9 | 베트남이력 | Phạm Văn Hiệp | 15080003 | Phòng chất lượng | 1982-03-18 | | |
| | | | | | | | | | | | | | | | | | 10 | 베트남이력 | Nguyễn Văn Việt | 15120009 | Phòng chất lượng | 1986-01-17 | | |
| | | | | | | | | | | | | | | | | | 11 | 베트남이력 | Nguyễn Trọng Đạt | 14080006 | Phòng V&T | 1982-11-22 | | |
| | | | | | | | | | | | | | | | | | 12 | 베트남이력 | Quần Thị Hải | 14120005 | Phòng V&T | 1989-11-09 | | |
| | | | | | | | | | | | | | | | | | 13 | 베트남이력 | Nguyễn Thị Loan | 11200013 | Phòng chất lượng | 1988-10-10 | | |
| | | | | | | | | | | | | | | | | | 14 | 베트남이력 | Chu Thị Phương | 14080016 | Phòng V&T | 1986-11-16 | | |
| | | | | | | | | | | | | | | | | | 15 | 베트남이력 | Nguyễn Quang Khôi | 14150001 | Phòng V&T | 1981-03-12 | | |
| | | | | | | | | | | | | | | | | | 16 | 베트남이력 | Vũ Đức Duy | 15130023 | Phòng sản xuất | 1988-10-27 | | |
| | | | | | | | | | | | | | | | | | 17 | 베트남이력 | Phạm Thị Thảo | 15150001 | Phòng sản xuất | 1982-07-10 | | |
| | | | | | | | | | | | | | | | | | 18 | 베트남이력 | Nguyễn Thị Thuần | 14160001 | Phòng V&T | 1983-10-02 | | |
| | | | | | | | | | | | | | | | | | 19 | 베트남이력 | Đỗ Thị Nguyệt | 15080002 | Phòng sản xuất | 1983-02-28 | | |
| | | | | | | | | | | | | | | | | | 20 | 베트남이력 | Đỗ Quang Trung | 13160013 | Phòng kế toán | 1988-04-10 | | |
| | | | | | | | | | | | | | | | | | 21 | 베트남이력 | Nguyễn Ngọc Toàn | 13160014 | Phòng kế toán | 1991-09-11 | | |
| | | | | | | | | | | | | | | | | | 22 | 베트남이력 | Đỗ Thị Thy | 14120006 | Phòng V&T | 1990-09-26 | | |
| | | | | | | | | | | | | | | | | | 23 | 베트남이력 | Mạch Thị Trang | 16170001 | Phòng chất lượng | 1995-09-08 | | |
| | | | | | | | | | | | | | | | | | 24 | 베트남이력 | Nguyễn Thị Trang | 12170003 | Phòng kế toán | 1994-05-30 | | |

공정결과 데이터

작업자 데이터

데이터 정제

[illegible]

Bad	badCd	ptGroupN	qabDscr	Year	Female	Edu	Skill	Hand	N
	1 C0112	MAIN	피복밀림	33	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	35	1	1	1	2	1
	0 C0112	MAIN	피복밀림	33	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	29	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	29	1	1	1	2	1
	0 C0112	MAIN	피복밀림	23	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	36	1	1	1	2	1
	0 C0112	MAIN	피복밀림	24	1	1	1	2	1

[illegible]

1) NA 제거

2) 유도변수 생성 : 생년월일 -> 나이

3) 범주형 변수 처리 : 남/여 -> 0 or 1

1) 변수 선택

로지스틱 회귀분석 p-value

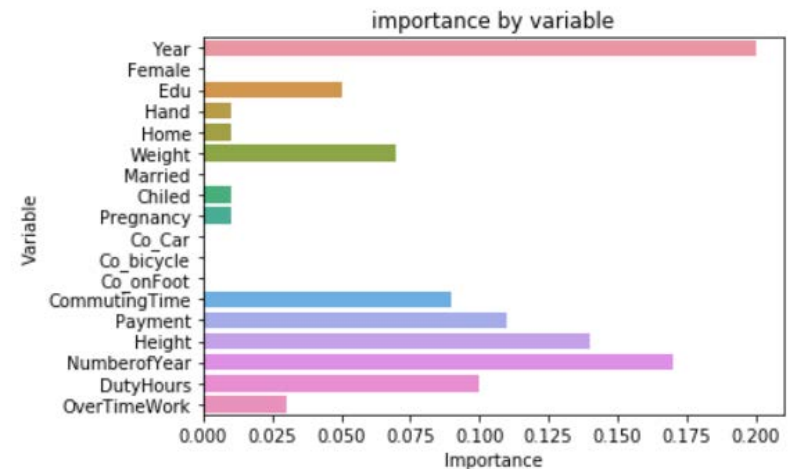
	coef	std err	t	P> t	[0.025	0.975]
const	0.9305	0.183	5.071	0.000	0.571	1.290
Year	0.0016	0.000	3.769	0.000	0.001	0.002
Female	-0.9547	0.176	-5.411	0.000	-1.300	-0.609
Edu	0.0058	0.004	1.411	0.158	-0.002	0.014
Skill	0.0122	0.008	1.607	0.108	-0.003	0.027
Hand	-0.0139	0.011	-1.276	0.202	-0.035	0.007
NumberofYear	-0.0070	0.004	-1.973	0.049	-0.014	-4.58e-05
Married	-0.0210	0.009	-2.224	0.026	-0.039	-0.002
Chiled	0.0289	0.007	4.361	0.000	0.016	0.042
Pregnancy	0.0394	0.009	4.164	0.000	0.021	0.058
CommutingTime	0.0816	0.009	8.716	0.000	0.063	0.100
Payment	-4.185e-09	4.75e-09	-0.882	0.378	-1.35e-08	5.12e-09
Height	-0.0002	0.000	-1.091	0.275	-0.001	0.000
Weight	0.0013	0.000	3.588	0.000	0.001	0.002
DutyHours	3.375e-05	4.56e-05	0.740	0.459	-5.57e-05	0.000
OverTimeWork	-0.0148	0.004	-3.664	0.000	-0.023	-0.007

P-value 값이 0.05보다 작은 변수 사용

2) Naïve Bayes

- Naïve Bayes를 사용한 이유
불량품 데이터가 양품 데이터에 비해 매우 적음
따라서, 생성된 모델은 대부분 양품 데이터로 분류
Prior를 불량품에 더 크게 주어, 불량품에 조금 민감한 모델을 생성
- Prior 비율선정
양품의 Prior : 0.4, 불량품의 Prior : 0.6

RandomForest분석 - importance



Randomforest 학습 후 importance가 0에 가까운 변수들은 학습에서 제외

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability

Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

04

3) 불량품 데이터 생성

Synthetic Minority Over-sampling Technique(SMOTE)

➤ Python imblearn.over_sampling 활용

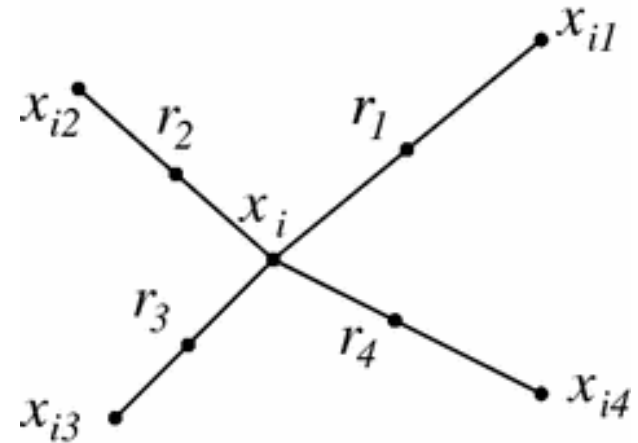
➤ SMOTE는 Oversampling 기법 중 합성데이터를 활용한 기법

After OverSampling, the shape of train_x: (46342, 15)

After OverSampling, the shape of train_y: (46342,)

After OverSampling, counts of label '1': 23171

After OverSampling, counts of label '0': 23171



4) MLP Classifier

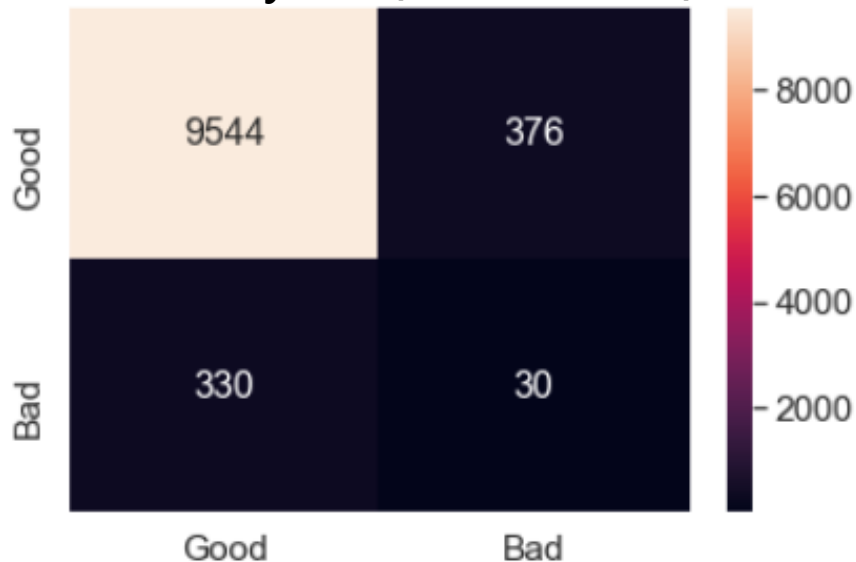
➤ Python sklearn 활용

➤ Hidden layer size : 100,200,300,100

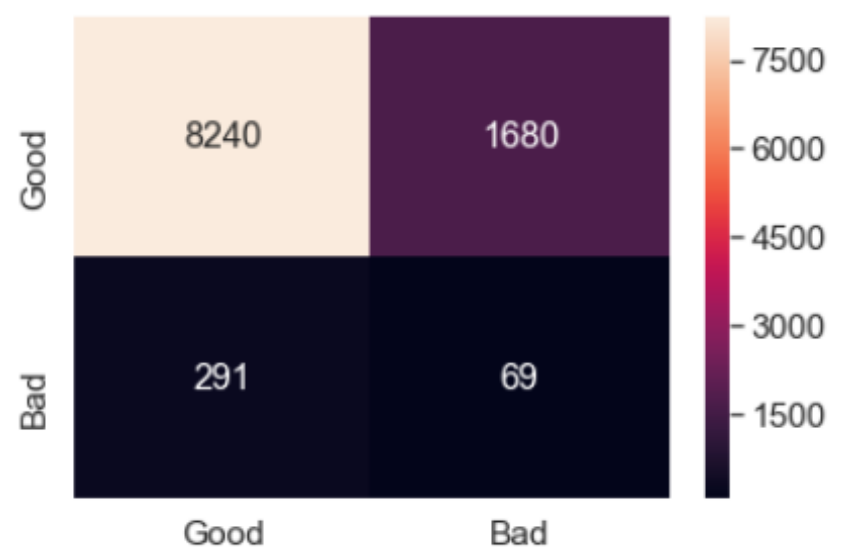
05

1) Confusion Matrix (Naïve Baysian)

Naïve Baysian (Prior 사용 안함)



Naïve Baysian Prior(0.6,0.4)

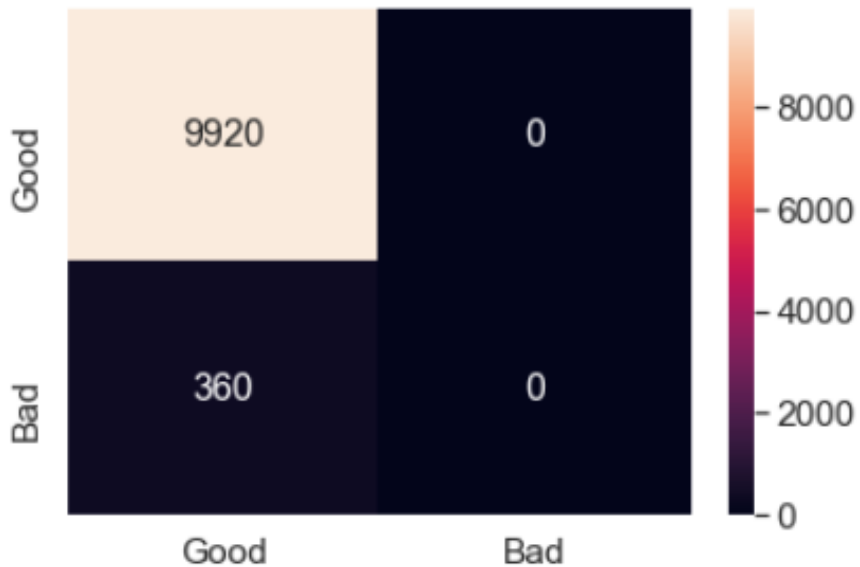


- Naïve Baysian의 Prior를 적용하지 않은 경우와 Prior를 적용한 경우
- 적용하지 않은 경우 Recall : 약 8.3%
- 적용한 경우 Recall : 약 19.2%

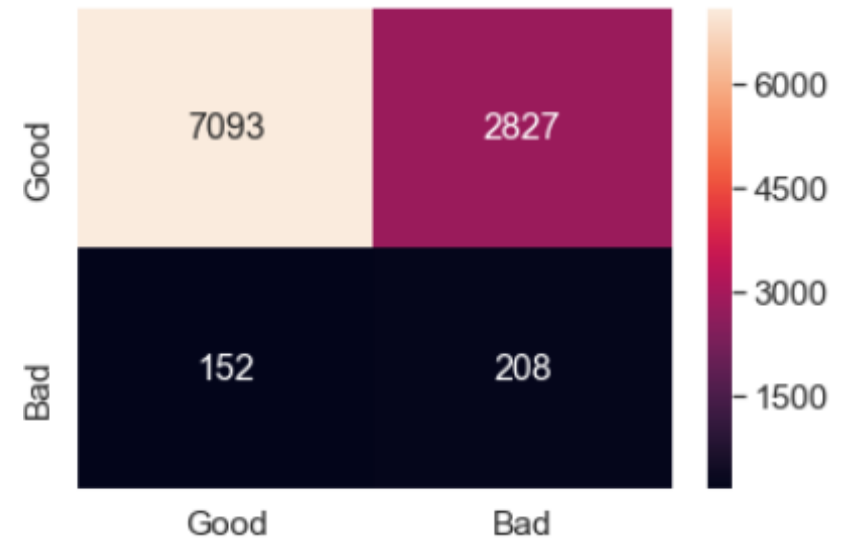
05

1) Confusion Matrix (MLP)

SMOTE 사용 안한 경우



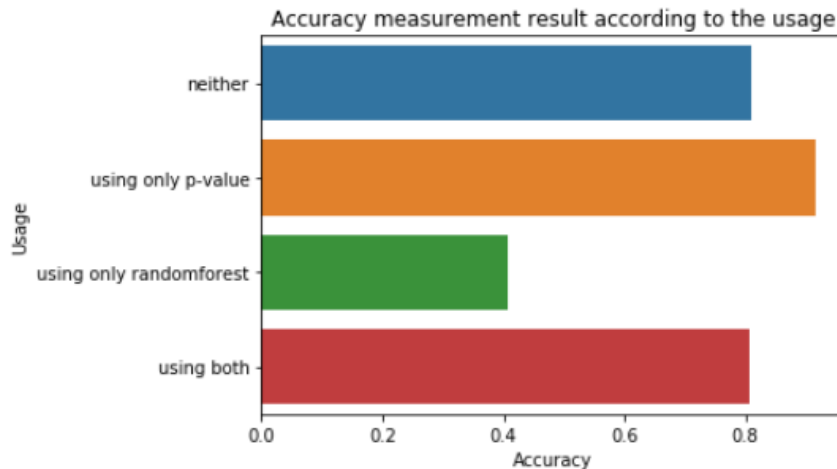
SMOTE를 사용한 경우



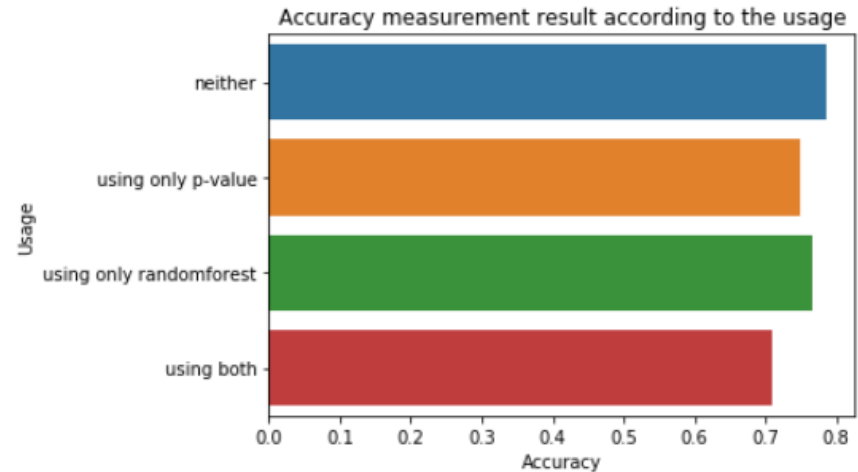
- Naïve Bayesian의 Prior를 적용하지 않은 경우와 Prior를 적용한 경우
- 적용하지 않은 경우 Recall : 0%
- 적용한 경우 Recall : 약 58%

2) Accuracy

Naïve Bayesian prior(0.6,0.4)



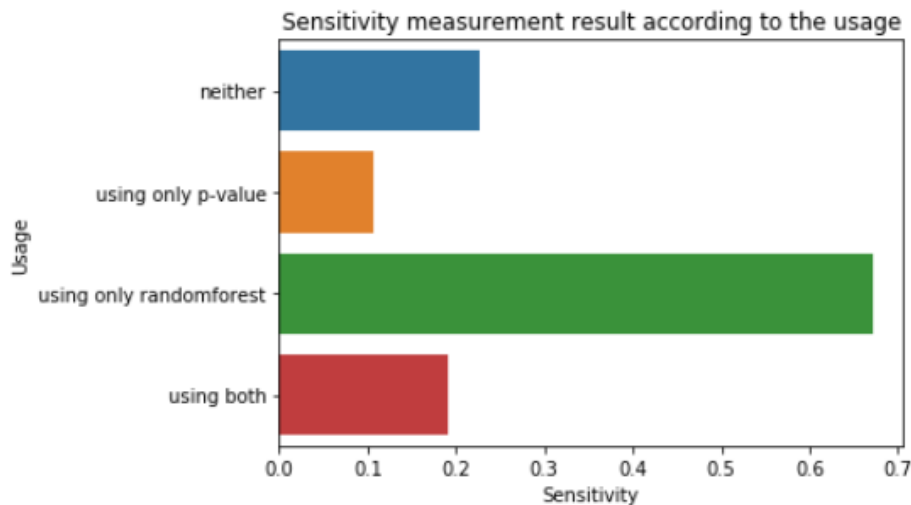
MLP를 활용한 경우



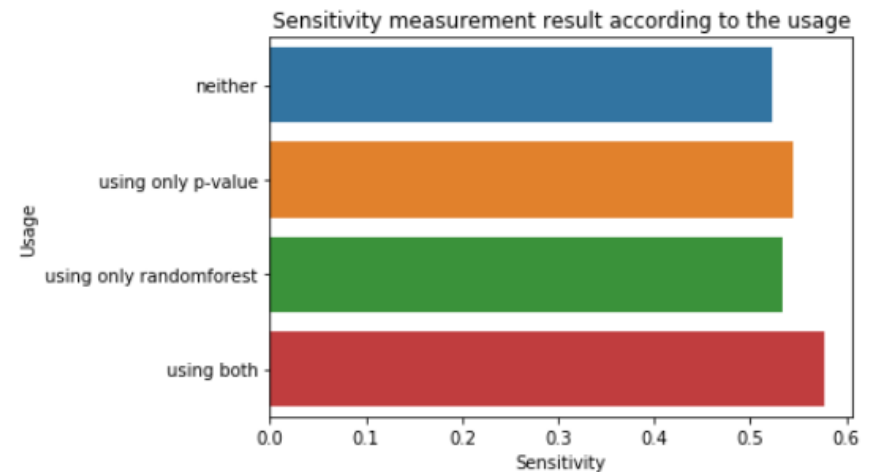
- 모델이 얼마나 잘 맞추었는가?
- 여러가지 변수선택법을 활용하여 분석한 Accuracy
- 위에서부터 변수 선택법을 모두 사용하지 않은 경우, P-value만 활용한 경우, RandomForest만 사용한 경우, 둘 다 모두 사용한 경우
- 정확도만으로는 모델을 분석하기가 어려움

3) Recall

Naïve Bayesian prior(0.6,0.4)



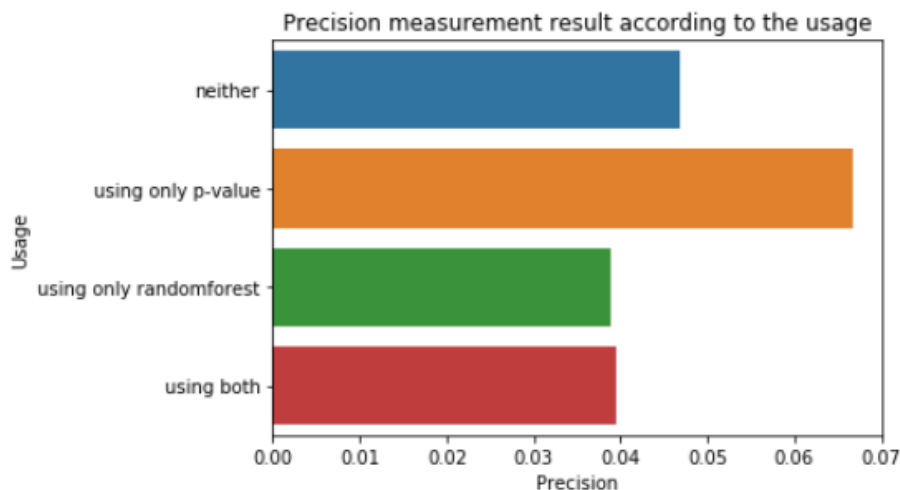
MLP를 활용한 경우



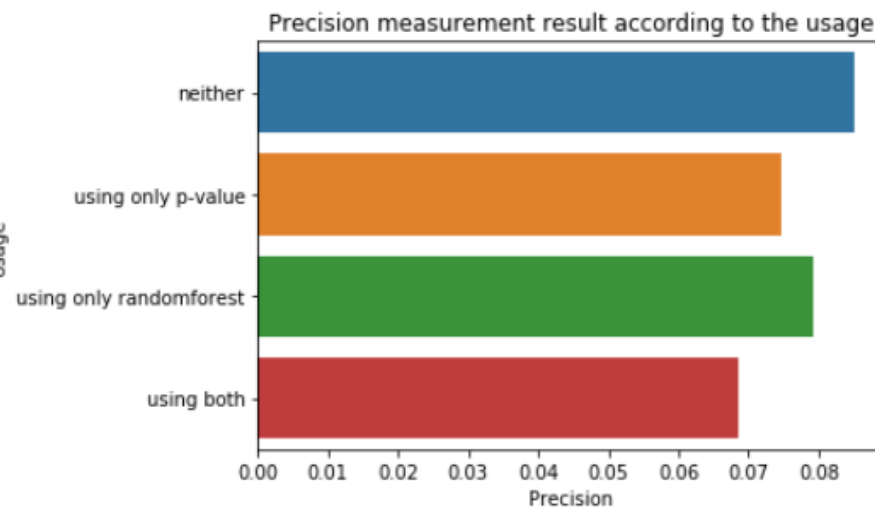
- 실제 불량품을 불량품으로 예측한 비율
- 클 수록 불량에 민감한 모델
- 이것만으로 모델을 평가하기는 어려움(모델이 모두 불량품이라고 하면, 높은 값을 가짐)

4) Precision

Naïve Bayesian prior(0.6,0.4)



MLP를 활용한 경우



- 불량품으로 검출한 것 중 실제 불량품인 비율
- Recall과 함께 분석을 해보면, p-value와 RandomForest를 활용해 변수를 줄여 학습을 수행하면, 좋은 결과를 기대할 수 있음.(모델의 크기는 줄이고, 그 만큼 성능은 향상 유지 가능)

06

[팀 프로젝트 결과]

1. 불균형 데이터 문제를 해결하는 방법 : SMOTE
2. 변수 선택법 : Randomforest importance, p-value
3. 변수 선택법을 사용하면 capacity ↓, 성능은 유지 및 향상

향후 계획

- GAN 활용 불량 데이터 생성
- 여러 모델 생성(변수선택법 검증)

불량품에 민감한 불량품예측기

THANK
YOU

인공지능 팀 프로젝트