Feature Selection Using Logistic Regression:

(Semiconductor Manufacturing Process Data)

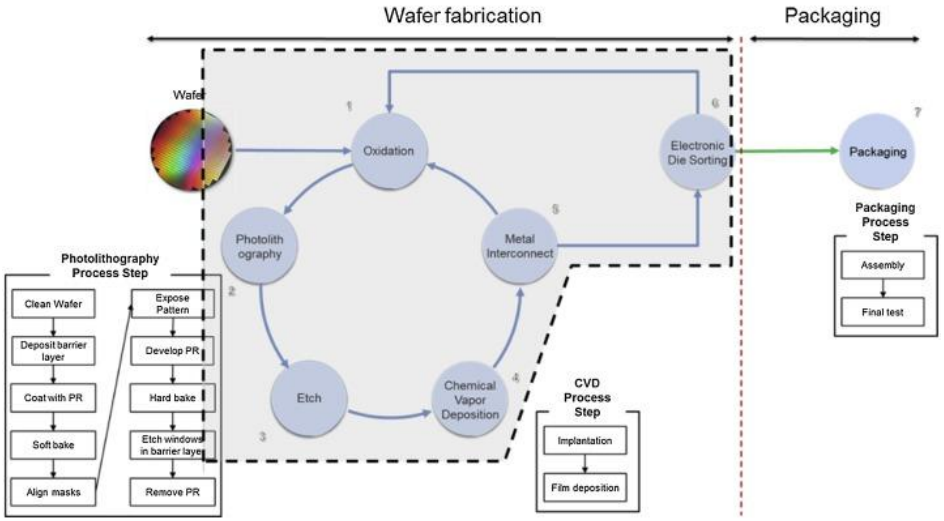
주건재

OVERVIEW

1. INTRODUCTION
1. DATA
1. MODEL
1. RESULT

1. INTRODUCTION

1.1 TOPIC



Research Question

Wafer 제조 공정 중 수집한 센서데이터로 Wafer결함을 예측하고
Logistic Regression 모델을 통해 중요한 Feature(센서)를 선택한다.

Goals

1st goal: 중요한 독립변수를 찾아낸다.

2nd goal: Imbalanced Data를 적절히 처리한다.

1.2 DATA OVERVIEW

Data Description

독립변수: Wafer 공정 중 수집한 590개 Sensor Data와 Time

종속변수: Wafer 불량 여부 (-1:양품, 1: 불량)

(각 Sensor가 무엇을 측정하는지는 모름)

Data Resource: uci ML Repository (https://archive.ics.uci.edu/ml/datasets/SECOM)

	Time	0	1	2	3	4	5	6	7	8	...	581	582	583	584	585	586	587	588	589	Pass/Fail
0	2008-07-19 11:55:00	3030.93	2564.00	2187.7333	1411.1265	1.3602	100.0	97.6133	0.1242	1.5005	...	NaN	0.5005	0.0118	0.0035	2.3630	NaN	NaN	NaN	NaN	-1
1	2008-07-19 12:32:00	3095.78	2465.14	2230.4222	1463.6606	0.8294	100.0	102.3433	0.1247	1.4966	...	208.2045	0.5019	0.0223	0.0055	4.4447	0.0096	0.0201	0.0060	208.2045	-1

Data Summary

```
Data shape: (1567, 592)
----------------------------
Data Types:
float64    590
object       1
int64        1
Name: Types, dtype: int64
----------------------------
```

	Types	Counts	Uniques	Nulls	Min	Max
0	float64	1561	1521	6	2743.24	3356.35
1	float64	1560	1505	7	2158.75	2846.44
10	float64	1565	393	2	-0.0349	0.053
100	float64	1561	36	6	-0.003	0.0023
101	float64	1561	30	6	-0.0024	0.0017
...	...	...	...	...	...	...
97	float64	1561	2	6	0	0
98	float64	1561	1421	6	-5.2717	2.5698
99	float64	1561	273	6	-0.5283	0.8854
Pass/Fail	int64	1567	2	0	-1	1
Time	object	1567	1534	0	2008-01-08 02:02:00	2008-12-10 18:47:00

1.3 Evaluation Method

Evaluation:

1) Imbalanced Data: 양품(1) 데이터가 대다수
2) Wafer 공정은 불량(0)을 걸러내는 것이 중요하다.

   => False Positive 최소화

MCC: Imbalanced Data에 주로 사용됨

"It combines both the accuracy and the coverage of the prediction in a balanced way"[1)]

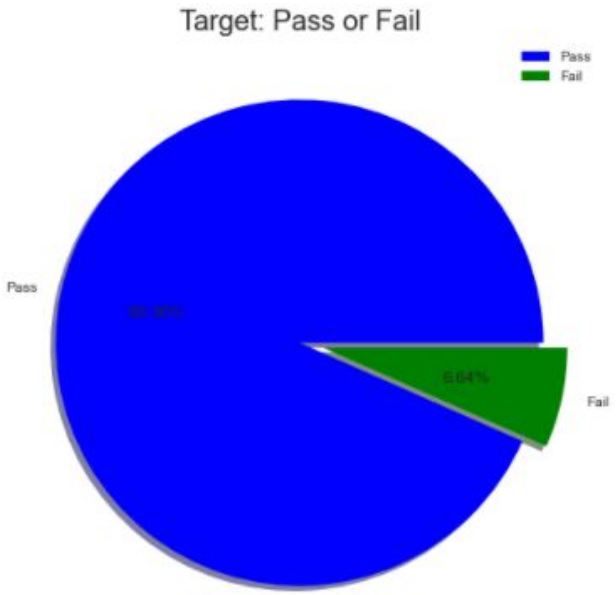
$$\text{Matthews correlation coefficient (MCC)} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

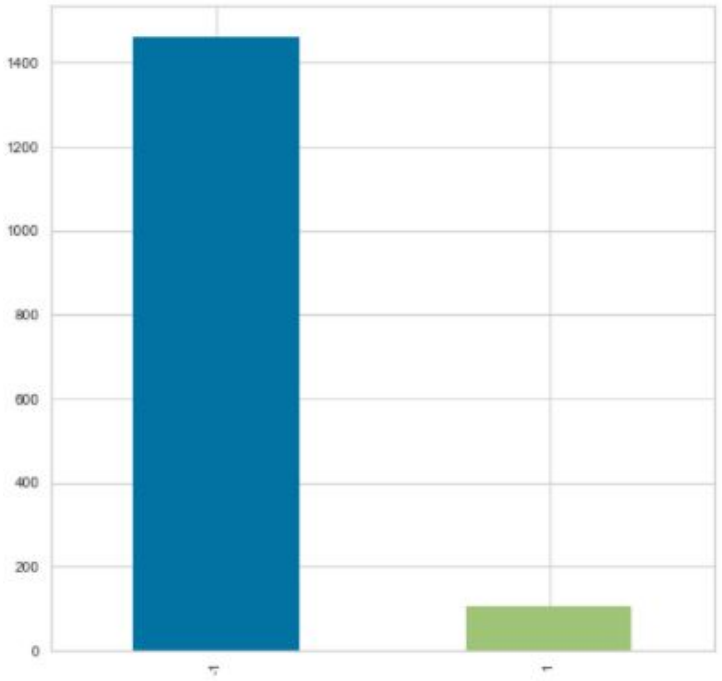
(worst value = −1; best value = +1).

1) Chang Cao 외2명, <The MCC-F1 curve: a performance evaluation technique for binary classification >,2020년 p4

2. DATA

2.1 Imbalanced Target





Imbalance

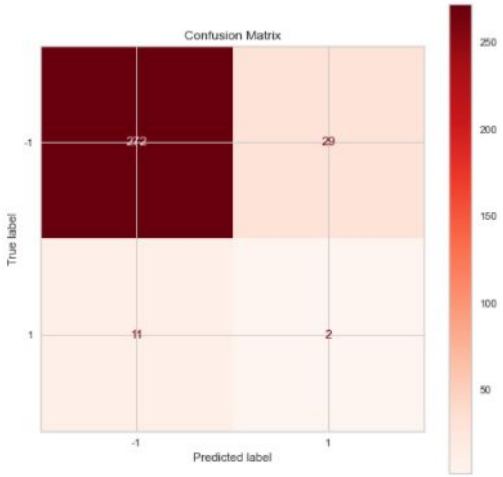
- SMOTE
- ADASYN
- Tomek
- Borderline SMOTE

2.2 Null Value

```
missing 10~20%: 20개
missing 20~30%: 0개
missing 30~40%: 0개
missing 40~50%: 4개
missing 50~60%: 4개
missing 60~70%: 16개
missing 70~80%: 0개
missing 80~90%: 4개
missing 90~100%: 4개
```

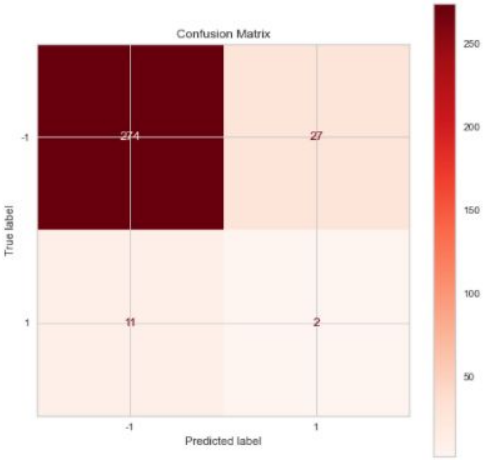
DROP

1. Fill Mean Value




```
Logistic Regression Evaluation
 F1 Score: 0.8726114649681529
 AUC: 0.55
```

2. KNN Imputer

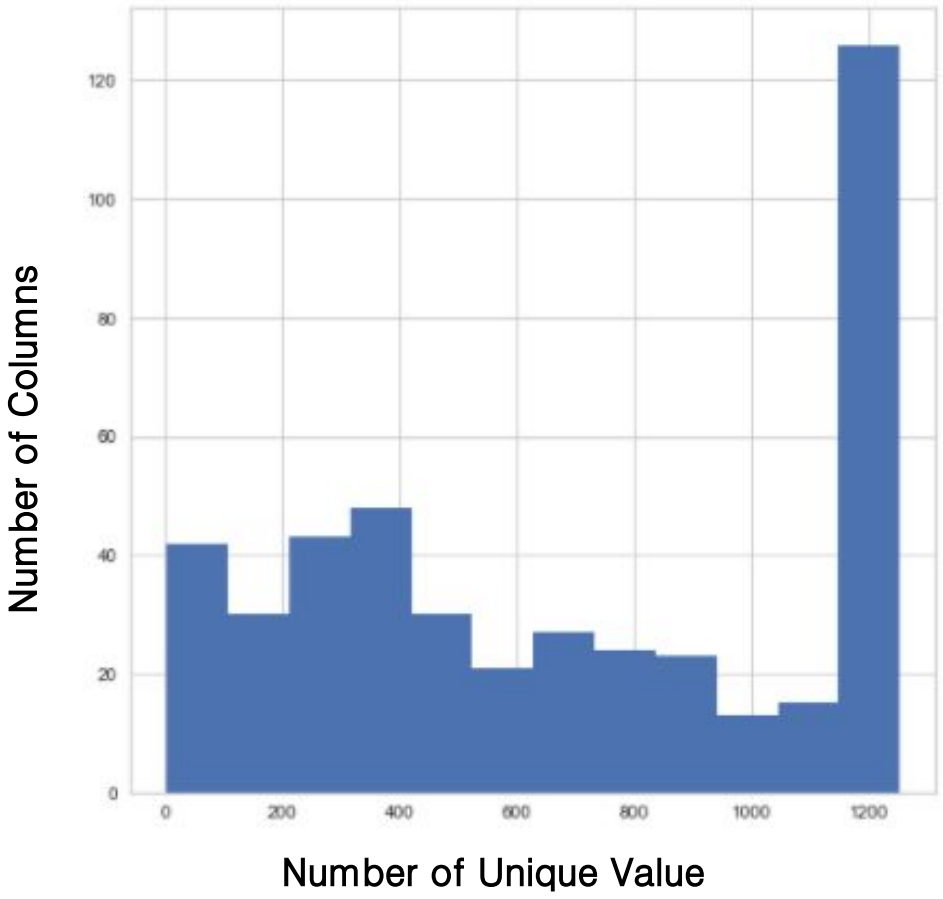



```
Logistic Regression Evaluation
 F1 Score: 0.8789808917197452
 AUC: 0.55
```

Result:

- 큰 차이 없어서 computation resource를 적게 사용하는 평균값 사용

2.3 Unique Value

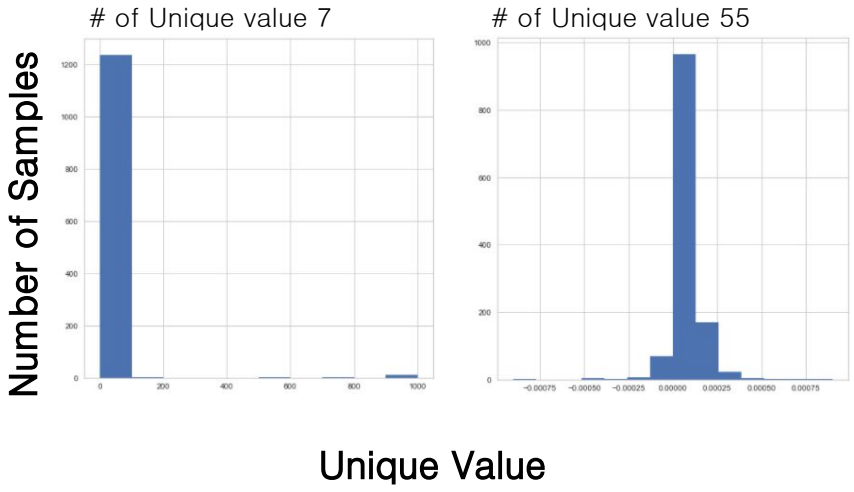


Graph Definition:

X_train shape (1253, 402)의 독립 변수 당 Number of Unique value의 Histogram

Result:

- unique value 100개 이하의 독립변수 40개를 drop

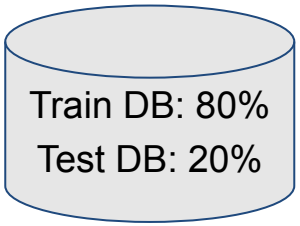


3. MODEL

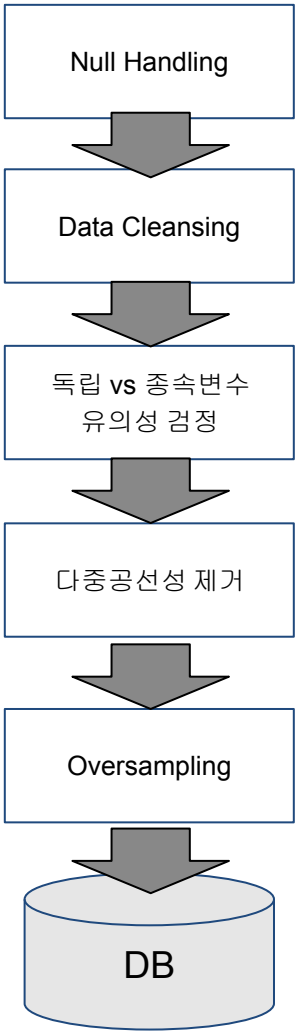
1.3 Model Overview

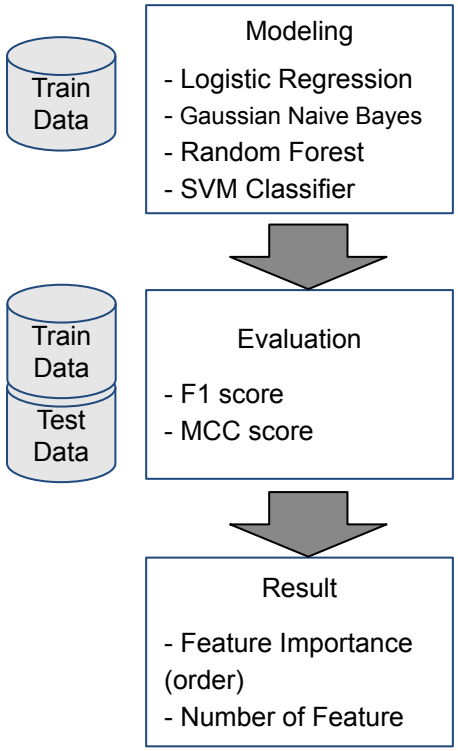
Data Understanding	Preprocessing	Modeling & Evaluation

Research Goal:
Find Important Sensor



특징:
- shape: (1567,590)
- masked data
- imbalanced
- garbage data 다수 존재





Cleansed Model

Preprocessing

Null Handling:
- Remove null value over 20% ot the col
- Remove null + 1 value data
- fill Mean Value

148개 변수 제거

Data Cleansing:

Remove under 100 unique values

근거:
- unique value 개수
- garbage data 다수 존재

40개 변수 제거

DB

특이사항:
- 독립변수 402개,
- Highly Imbalanced
- 종속변수와의 관계 미검정

Modeling & Evaluation

	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.944932	0.926752	0.485731	0.041957
Gaussian Naive Bayes	0.252993	0.162420	0.102539	-0.054267
RandomForest	0.996808	0.961783	0.976096	0.271982
SVM Classifier	0.996808	0.853503	0.976096	0.023212

Feature Selected Model

Preprocessing

Null Handling:
- Remove null value over 20% ot the col
- Remove null + 1 value data
- fill Mean Value

148개 변수 제거

Data Cleansing:

Remove under 100 unique values

근거:
- unique value 개수
- garbage data 다수 존재

40개 변수 제거

종속변수에 따른관계
WillCox two rank test

alpha = 0.05

근거:
- T-test: 정규성 조건 불만족

300개 변수 제거

다중공선성
VIF

10 이상 반복 제거

근거:
- 독립 변수간에 Highly correlated

30개 변수 제거

DB

특이사항:
- 독립변수 72개,
- Highly Imbalanced

Modeling & Evaluation

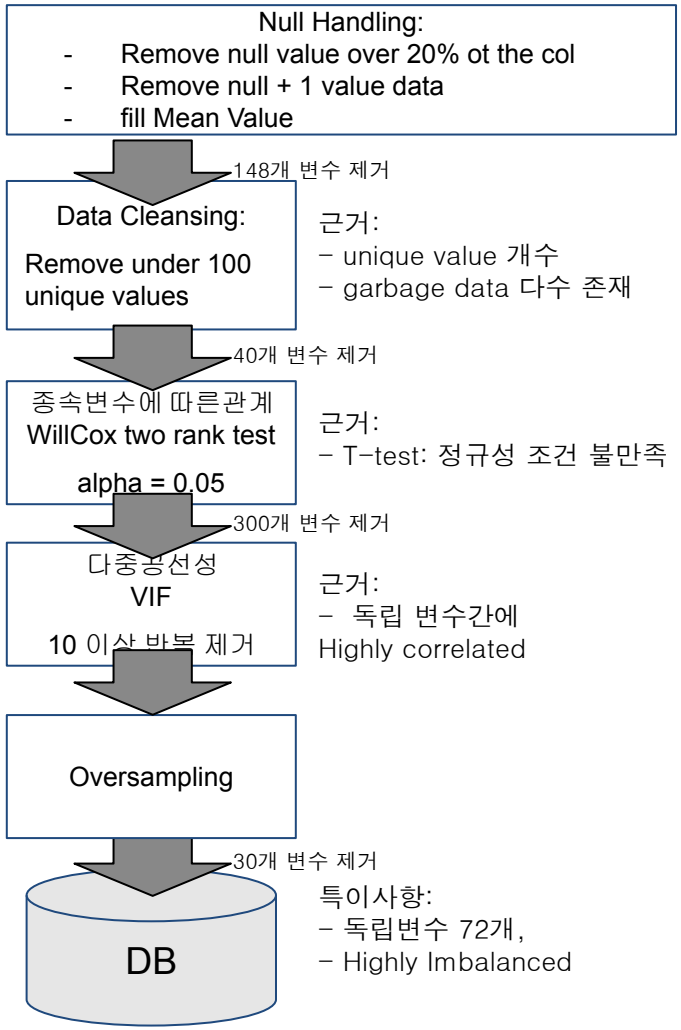
	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.926576	0.958599	0.079845	0.184308
Gaussian Naive Bayes	0.830806	0.818471	0.193743	0.117107
RandomForest	0.996808	0.942675	0.976096	0.221245
SVM Classifier	0.927374	0.958599	0.000000	0.000000

	Logistic Regression	RandomForest
0	511	59
1	59	348
2	577	279
3	102	477
4	348	102
5	460	40
6	135	200
7	510	577
8	200	295
9	78	510
10	428	434
11	279	428
12	213	468
13	477	282
14	282	78
15	79	425
16	467	213
17	468	79
18	425	460
19	295	511
20	434	467
21	40	135

mcc_test 값의 변화로 미루어 보아 다중공선성 제거는 Logistic Regression 모델을 크게 향상 시키고 Random Forest는 원래 다중공선성에 강하여 오히려 성능저하를 볼 수 있었다.

Oversampled Model

Preprocessing



Modeling & Evaluation

SMOTE

	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.727625	0.710191	0.455298	0.080389
Gaussian Naive Bayes	0.631670	0.321656	0.336677	0.031126
RandomForest	0.997418	0.885350	0.994850	0.099351
SVM Classifier	0.730637	0.694268	0.461274	0.103245

ADASYN

	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.743523	0.707006	0.487095	0.078667
Gaussian Naive Bayes	0.632988	0.315287	0.338210	0.028412
RandomForest	0.996978	0.869427	0.993973	0.167962
SVM Classifier	0.753022	0.726115	0.506043	0.089260

Tomek

	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.744699	0.703822	0.489374	0.076960
Gaussian Naive Bayes	0.624405	0.324841	0.314249	0.032473
RandomForest	0.997836	0.853503	0.995682	0.067148
SVM Classifier	0.748594	0.719745	0.497174	0.085658

Borderline SMOTE

	f1_train	f1_test	mcc_train	mcc_test
Logistic Regression	0.751291	0.742038	0.502851	0.098613
Gaussian Naive Bayes	0.608003	0.273885	0.295520	0.009973
RandomForest	0.998709	0.882166	0.997419	0.095601
SVM Classifier	0.755594	0.754777	0.511917	0.106502

4. RESULT

Research Result

- **1st goal: 중요한 독립변수를 찾아낸다.**

- 중요도가 낮은 변수를 제거하면 예측 성능이 크게 저하 한다.
- Model별 중요변수의 순위가 다소 다르다.

  분석: 모두 중요한 변수일것으로 예상되며 이를 정확히 설명하기 위해서는 데이터를 unmask하고 도메인 지식과 함께 보아야한다.

- **2nd goal: Imbalanced Data를 적절히 처리한다.**

- Oversampling 이론적 근거를 찾지 못했다.
- 여러 방법을 적용해 본 결과 예측 성능 저하로 이어졌다.

  분석: Oversampling에 적합하지 않은 데이터일 가능성이 농후하다.

Further Study

- Unique Value 100개 이하 독립변수 40개에 대한 각각의 전처리
- PCA 와 PCA loading을 활용한 Modeling 및 Feature Importance 방법
- Oversampling의 Convexity 등의 이론적 가능성을 먼저 검토후 적절한 처리

References

- Cao, C., Chicco, D., & Hoffman, M. M. (2020). The MCC-F1 curve: a performance evaluation technique for binary classification. *arXiv preprint arXiv:2006.11278*.
- Kim, J., Han, Y., & Lee, J. (2016). Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process. *Advanced Science and Technology Letters*, *133*, 79-84.
- Munirathinam, S., & Ramadoss, B. (2016). Predictive models for equipment fault detection in the semiconductor manufacturing process. *IACSIT International Journal of Engineering and Technology*, *8*(4), 273-285.
- Kerdprasop, K., & Kerdprasop, N. (2010, March). Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process. In *World Congress on Engineering 2012. July 4-6, 2012. London, UK.* (Vol. 2188, pp. 398-403). International Association of Engineers.

Q&A