**삼성전자 SQCI(Supplier Quality Control Innovation) 고도화 방안**

수석 최 원호

선임 임 재균

2022.2.16

*AItheNutrigene*

*요약*

건조기의 부품의 제조 불량 여부를 검출하는 과정은 비용의 절약을 목표로 할 뿐만 아니라, 완성된 제품을 보다 완벽히 생산하여 고객에게 제공하는 서비스의 편의를 극대화하는 데에 목적을 두고 있다. 제공된 표본을 대상으로 pilot testing을 실시하였으며, 부품의 약 1.22%의 불량률을 보였고, 제품의 정상 여부를 검출하는 과정에서 검출하는 방식에 따라 AD01과 AD02의 불량품을 검출하는데 차이가 있다는 통계적 결과를 발견했다. 두 검출방식에서도 정상품과 불량품의 차이도 통계적으로 유의했다. 약 5초간 66개 또는 67개의 제품을 생산할 경우 초기에 불량으로 인식될 확률이 높게 나타났으며 간략한 검출방안으로 AD02의 분포에서는 98% 이상의 정확도를 보였으나 AD01는 value의 분포가 여러 모달을 나타내는 관계로 정확도가 약 60%에서 89%의 범위를 보여, 단순한 통계적 규칙(Rule set)을 적용하기에는 무리가 있다는 인식과, 동시에 새로운 이상치 알고리듬이 필요하다는 결론에 도달했다. LSTM이나 AutoEncoder가 이상치 탐색으로 각광받고 있으며 이 둘의 결합 또한 새로운 알고리듬으로 활용되고 있다.

개요

가. 프로젝트 목표

: 협력업체에서 납품받는 부품에 대한 입고검사상에서 품질관리 고도화를 위한 신규 Rule Set 발굴과 알고리즘 개발

나. 기존에 수행했던 모형화 및 데이터 특성

- 자체 샘플링 모듈을 사용하여 입고 부품을 샘플링하여 품질 관리

- 제조사(협력업체) 데이터를 feature로 생산(공정)불량을 예측

- 부품별 검사항목 갯수가 다름, 즉 데이터의 컬럼수가 다를 수 있음

- 제조사(협력업체) 데이터는 검사항목 1개당 6개 컬럼이 있고 샘플당 최대

컬럼은 30개

- 분석 대상이 되는 부품은 삼성전자 내부에서 선별하여 진행예정

- 삼성SDS 측은 이상치 탐지로 접근하여 Tree계열(Random Forest /

Isolation Forest ) 적용 (공정데이터를 활용한 시장 불량 예측)

다. 표본의 통계량 및 정보

시장불량품(MARKET\_SN.txt)과 공정상에서 합격품이나 불량품과 일치하는 제품번호(PROD\_SN)가 없다. [표 1]에서는 features에 대한 간단한 요약 정보를 나타낸다. PASS\_YN는 검수의 양호 여부를 나타내고, INSP\_DTL\_SEQ는 0.1초에 검수하는 순번을 나타내는데 66번 또는 67번의 검수가 이루어지는 순환을, 제품 시리얼 번호(PROD\_SN)별로 반복해서 표기하고 있다. header, INSP\_NUM,INSP\_CODE ,INSP\_DTL\_CODE, IP\_ADDRESS는 하나의 값만 갖는 명목변수로 표본에서 어떤 정보를 내포하는지 추측하기 어렵다. value 특성은 최솟값 -145.1에서 1208의 범위를 나타내며, 평균값이 234.55이고 표준편차가 대략 254인 점을 감안하면 분포가 왼쪽(중앙값<평균)으로 치우쳐져 있으며 매우 넓게 퍼져 있다고 판단된다.

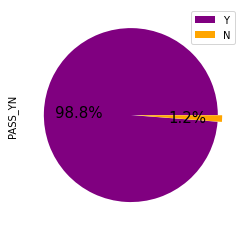
[표 1] 표본의 특징과 분포

|  |  |  |
| --- | --- | --- |
| **features** | **category** | **information** |
| header | CPC | 하나의 값 |
| INSP\_NUM | 2 | 하나의 값 |
| INSP\_CODE | RTO2 | 하나의 값 |
| INSP\_DTL\_CODE | ADO1, ADO2 | 두 개의 값 |
| VALUE | mean 234.553  std 254.749  min -145.100  25% 8.700  50% 19.000  75% 528.000  max 1208.000 |  |
| PROD\_SN | 5233 | 5233 SETS OF 66/67 (INSP\_DTL\_SEQ) |
| START\_TIME | 2021/10/22 11:53:25. | STARTING TIME OF SERIAL NUMBER |
| END\_TIME | 2021/12/07 11:14:36 | END TIME OF SERIAL NUMBER |
| PASS\_YN | Y/N | unbalanced distribution |
| INSP\_DTL\_SEQ | 66/67 | 0.1초 간격으로 66개 또는 67개의 제품 검수 |
| IP\_ADDRESS | 0.0.0.102 | 하나의 값 |

라 표본의 불량 정보

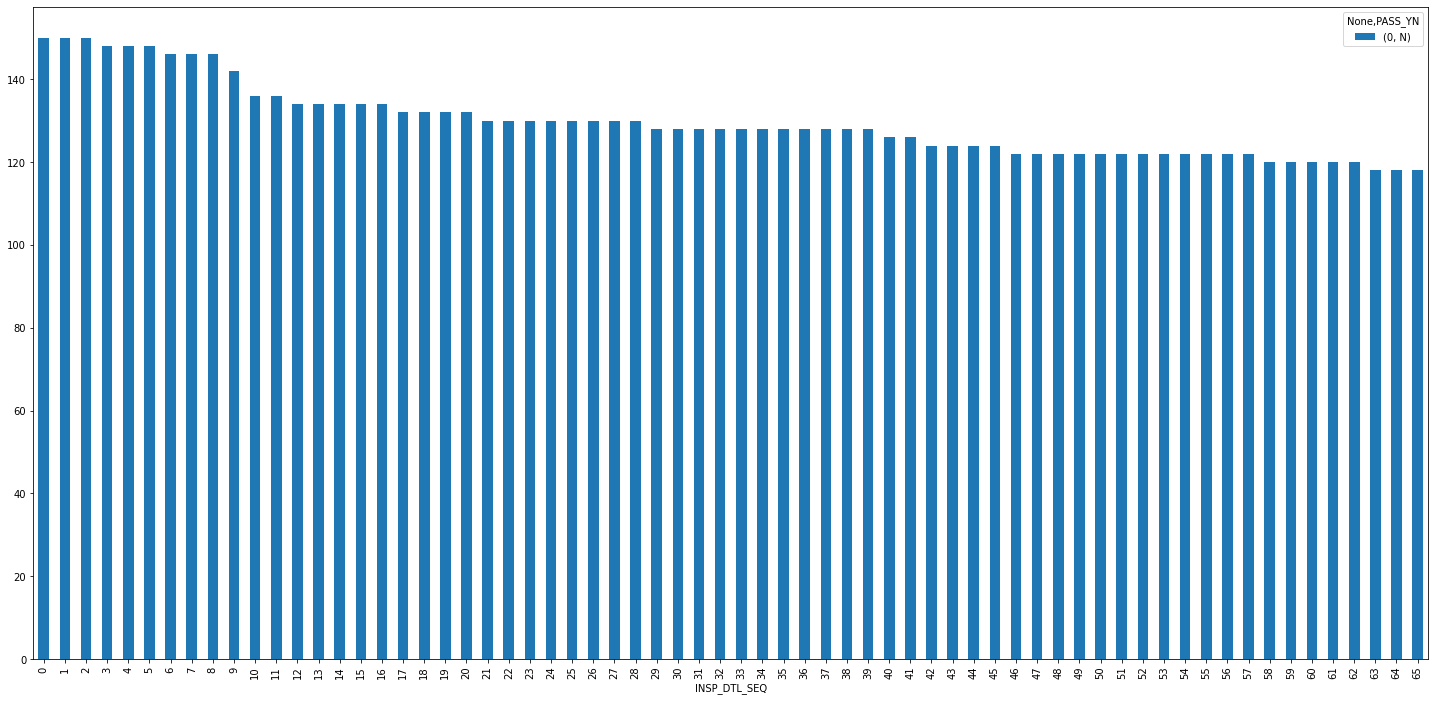
건조기 표본 관측 수는 700,385이고 이중 불량(N) 검수개수는 8558개로 나타나 불량률은 약 1.22% 정도이다[그림 1]. [그림 2]를 보면 초기에 제품의 불량 검수가 상대적으로 많이 발생하고 있다. 30번째 정도에 가까워져서야 평균과 같은 수치를 보인다. [그림 3]은 특정한 제품 번호 대에서 불량 검수가 많은지를 가늠하기 위해서 그려본 그래프이다. 하지만 뚜렷한 패턴이나 이상치를 찾기가 어려웠다.

[그림 1] 건조기 표본에서 불량률 비중

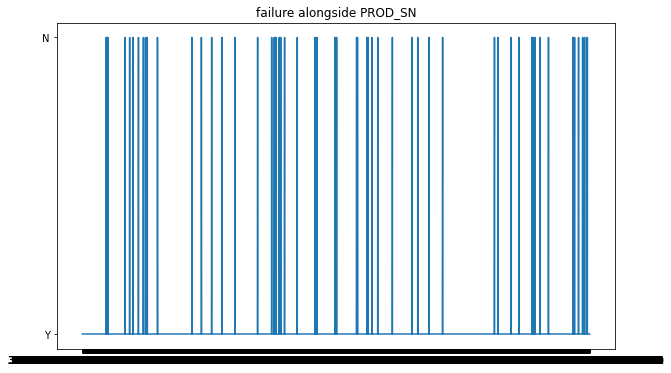


전체 불량률: 8558/700385=0.01221899

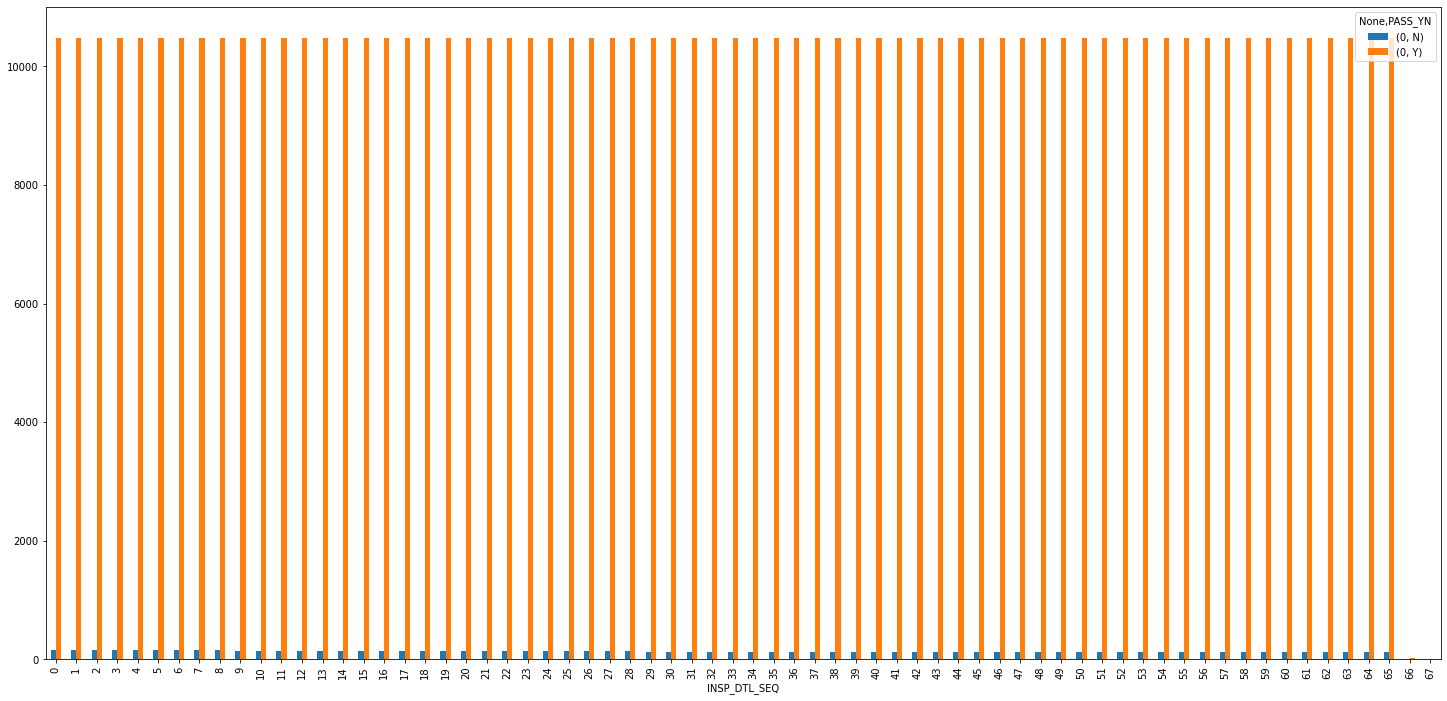
[그림 2] 제품 순번에 따른 불량품 발생 빈도



[그림 3] 시리얼 넘버에 따른 불량제품의 빈도

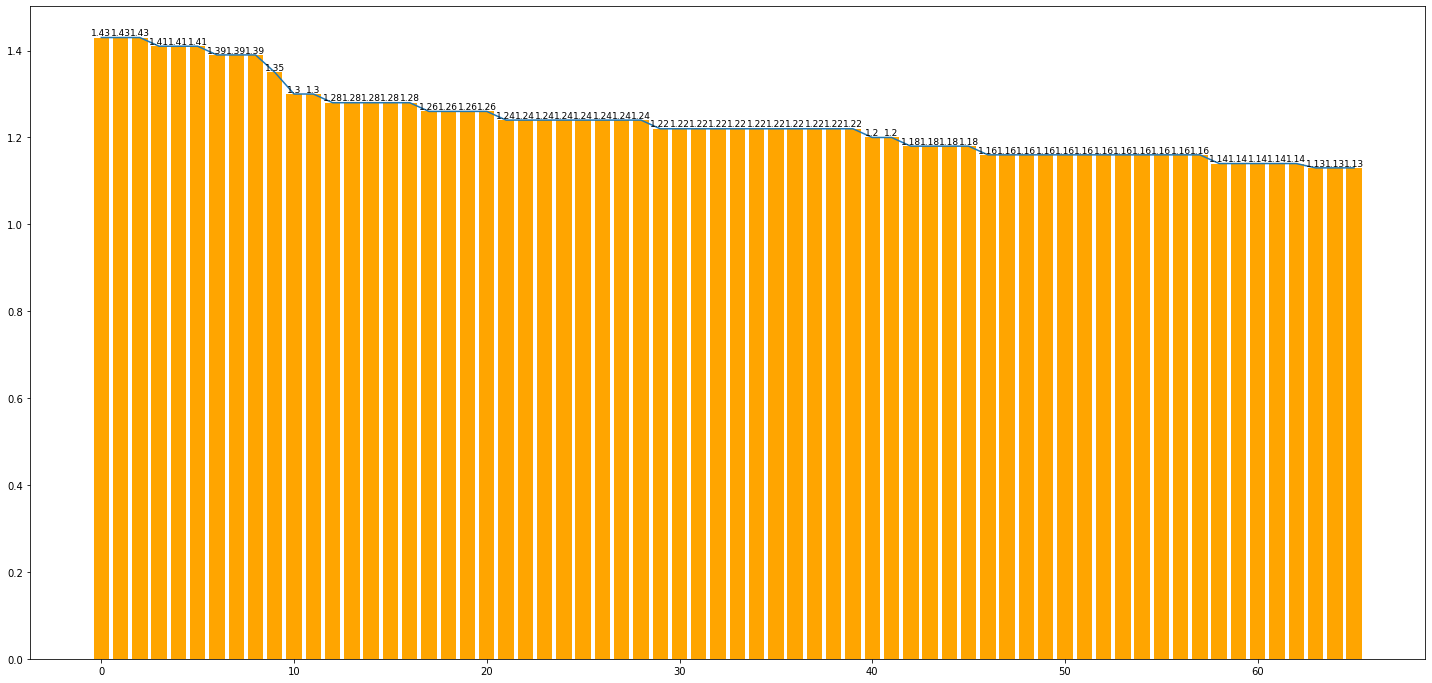


[그림 4] 제품 순번에서 불량 검수 수와 양호 검수 수



[그림 4]를 보면 매 순회마다 합격품 수는 10,482로 동일하나 불량 품수는 달라진다. 시리얼 넘버가 바뀌때마다 시작 시점에서 불량품이 많이 발생하는 경향을 드러내고 있다.

[그림 5] 불량 검수 대비 양호 검수 비율



[그림 6]은 INSP\_DTL\_CODE에 의해 구분된 두 코드별로 임의로 몇 개의 제품 시리얼 번호를 추출하여 실패한 제품과 통과한 제품의 value 분포가 어떻게 나타나는가를 조사를 하기 위해 그린 분포도이다. 표본에서 일부 추출된 AD01로 구분된 value의 분포는 비슷한 패턴을 보인다. 주황색 선으로 나타난 value의 분포가 눈에 띄는데 제품 검수 결과는 양호로 나타났다. 그러나 비슷한 분포를 띄더라도 제품 검수 결과가 불량으로 나타나는 경우도 있다(98\_n). 또한, AD02로 구분된 일부 표본을 추출하여 value의 분포를 그려보았다. 대부분 비슷한 패턴을 보이나 여기서도 불량과 양호 검수 결과가 동시에 발생해 불량여부의 특정패턴에 의해 발생한다는 가설을 지지하기가 어렵다. 패널 C에서 보면 불규칙한 value의 분포를 볼 수 있는데 이 표본은 불량 판정을 받았다. value 분포가 상대적으로 매우 불규칙하다면 불량 판정을 받을 가능성이 존재하는지 모형화를 통해 확인해 보아야 한다. 98번인 표본은 처음엔 불량 판정을 받았다가 다시 검수한 결과가 양호판정을 받은 경우이다. 두번 연속해서 불량 판정을 받으면 그 다음에는 검수를 하지 않는다. 물론 불량 판정이 한 번의 시퀀스에서 결정이 나는 경우도 있다. 불량이 나는 경우를 보면 어떤 특정 패턴을 발견하기가 쉽지 않다. 특히, AD01에서 불량 제품의 검수 시퀀스가 다양하여 Rule을 정형화하기가 쉽지 않다. 이러한 패턴이나 결과가 암시하는 바는 불량 객체를 검출하는 요인을 개선하고 시장에서의 불량 제품을 최소화하는 모형을 개발 또는 개선해야 한다는 점이다.

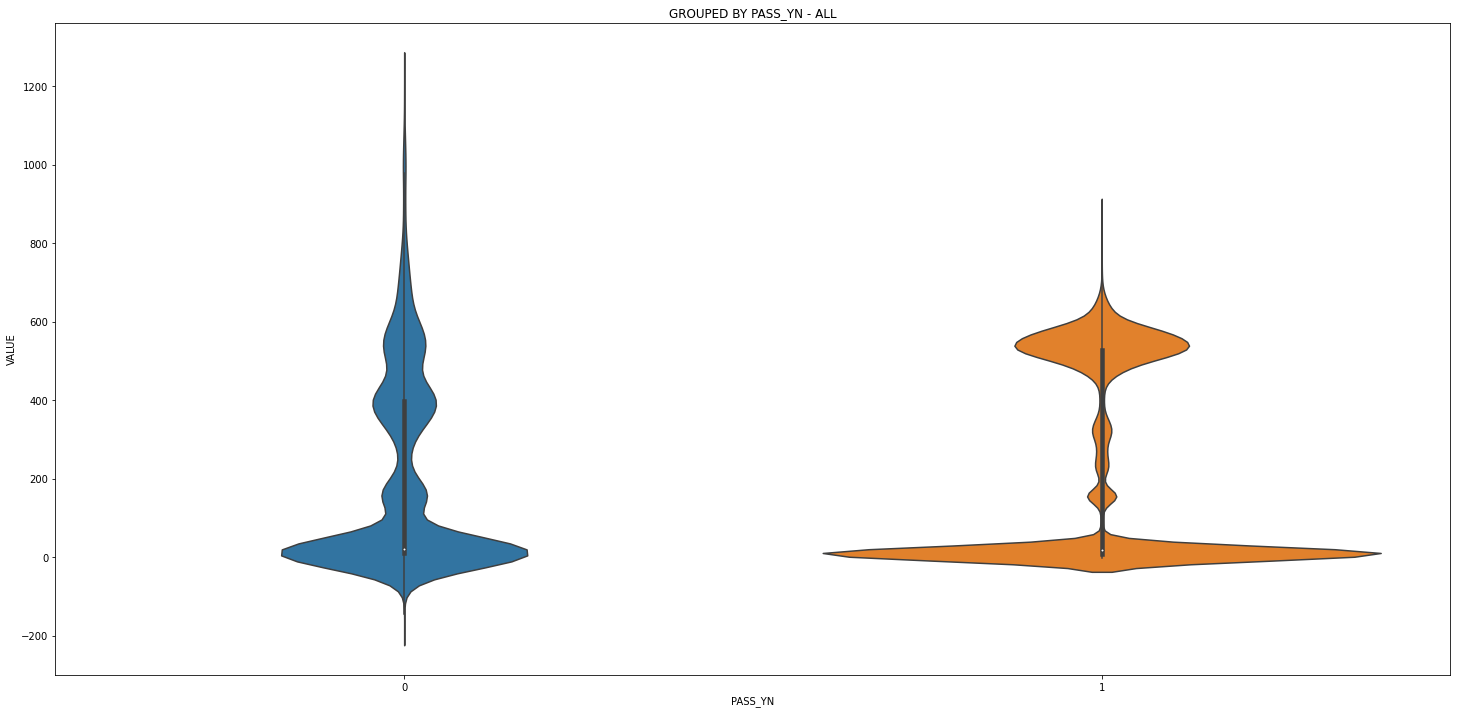
[그림 6] INSP\_DTL\_CODE(ADO1, ADO2 )의 구분에 따른 value 분포

|  |
| --- |
| A: value distribution by AD01 |
|  |
| B: value distribution by AD02 |
|  |
| B: value distribution by AD02\_2 |
|  |

Note: # 01, 02, 03, 04, 26 = Y만 있음, 20 : N만 있음, 13, 98 : N, Y 둘 다 있음

[그림 7]은 불량 검체와 양호 검체의 구분에 따른 value의 분포를 나타낸 그래프이다. value의 퍼짐이나 깊이가 서로 다르다는 걸 쉽게 파악할 수 있다. [그림 7-2]는 AD01과 AD02로 나누어 불량 검체와 양호 검체의 value 분포를 나타낸 그래프인데 양상이 매우 다르다. 서로 다른 퍼짐과 깊이가 뚜렷이 대비된다.

[그림 7] 불량여부에 따른 value의 분포



[그림 7-2] 불량여부에 따른 value의 분포

|  |  |
| --- | --- |
| AD01의 value 분포 | AD02의 value 분포 |
|  |  |

[표 2]는 [그림 6], [그림 7], [그림 7-2]를 바탕으로 두 AD01과 AD02의 검출 능력에 차이가 있는지를 분석한 결과이다. 건조기 부품의 양호가 두 검출방법에 차이가 두드러져 통계적으로 유의미했고 불량 검출에도 1% 유의수준에서 차이를 보였다.

[표 2] 두 코드 분류에 의한 성공과 실패의 차이

|  |  |  |
| --- | --- | --- |
|  | **성공** | **실패** |
| t value(AD01-AD02) | 1570.67 | 88.82 |
| p value | <0.0001 | <0.0001 |

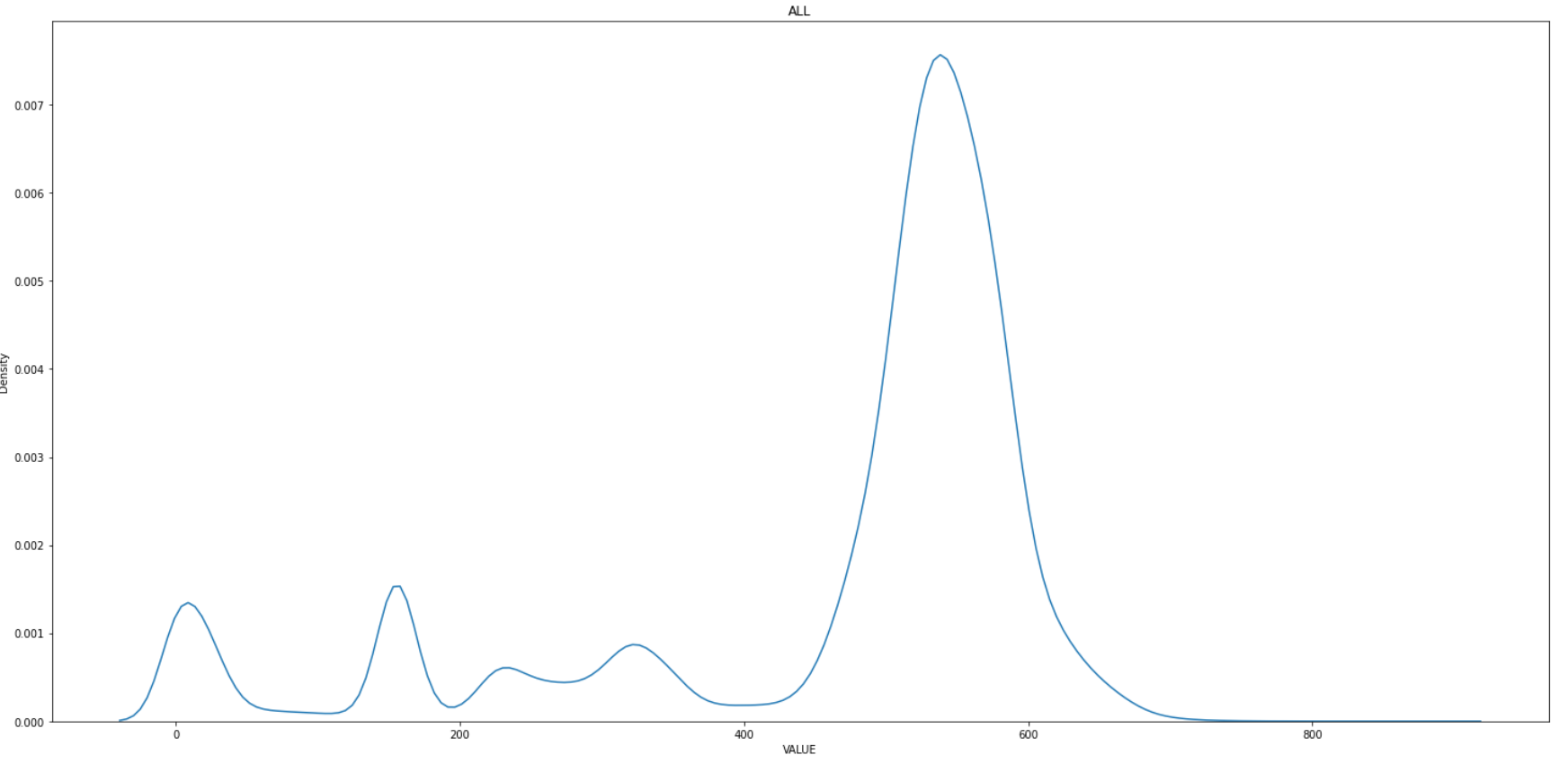
[그림 8]은 value와 검수 결과의 관계가 있는지를 순번 번호(INSP\_DTL\_SEQ)별로 상관계수를 시각적으로 나타낸다. [그림 5]에서 보면 초기에 불량제품 수가 많이 나온다는 증거를 바탕으로 둘 간에 어떤 관계성이 있는가를 파악하기 위해서이다. 불량제품 여부는 범주이므로 비모수통계인 spearman 상관계수와 Kendall-Tau 상관계수를 계산했다. 두 상관계수의 움직임이 비슷한 패턴을 보이나 계수 값 자체가 작아 영향관계가 적다고 판단된다. value가 계량적인 특성 변수이므로 pearson 상관계수도 실행해 보았으나 계수 자체가 작은 값을 가지고 있어 의미가 없어 보인다.

[그림 8 ] value와 PASS\_YN 상관계수

|  |
| --- |
| A: spearman correlation |
|  |
| B: kendall-tau correlation |
|  |
| c: pearson correlation |
|  |

[그림 9]는 value의 분포를 불량/양호 여부와 구분하지 않고 그린 그래프이다. 약 4개 정도의 모달이 있어 보인다. [글미 9-2]에서 왼쪽 value의 분포를 보면 전체의 value분포가 검출 방법인 AD01에 의해 기인하여 나타났다고 판단된다. AD01의 모달도 4개로 전체 value 분포와 비슷해 보이기 때문이다. 반면, AD02는 모달이 하나로 단순한 형상을 보이고 있다.

[그림 9] value 분포



[그림 9-2] 검출방법 AD01과 AD02에 의한 value 분포

|  |  |
| --- | --- |
| value distribution by AD01 | Value distribution by AD02 |
|  |  |

[그림 7]와 [그림 7-2], [그림 9]와 [그림 9-2]에서 검사방법인 AD01과 AD02의 value의 분포가 다르고 [표 2]에서 보다시피 양호와 불량을 검출하는 수량이 검사방법에 따른 차이를 보이고 있다는 점에서 착안하여 단순한 불량판단 여부를 확인할 수 있는 규칙을 세우고 적용해 보았다. [표 3]은 AD01의 모달을 보고 네 구간으로 구분하여 평균과 표준오차를 계산하였고 AD02는 모달이 하나이므로 전체를 대상으로 통계량을 계산하였다. 평균에서 1 표준편차 안에 value 값이 있으면 양호(Y)로 판단하고 범위 밖에 있으면 불량(N)으로 판단하는 규칙이다. AD02는 단순한 RuleSet만으로도 98.78%의 정확도를 보였다. 그러나 AD01은 value 분포의 퍼짐으로 인해 정확도가 60%에서 88%로 넘은 범위의 정확도를 나타낸다.

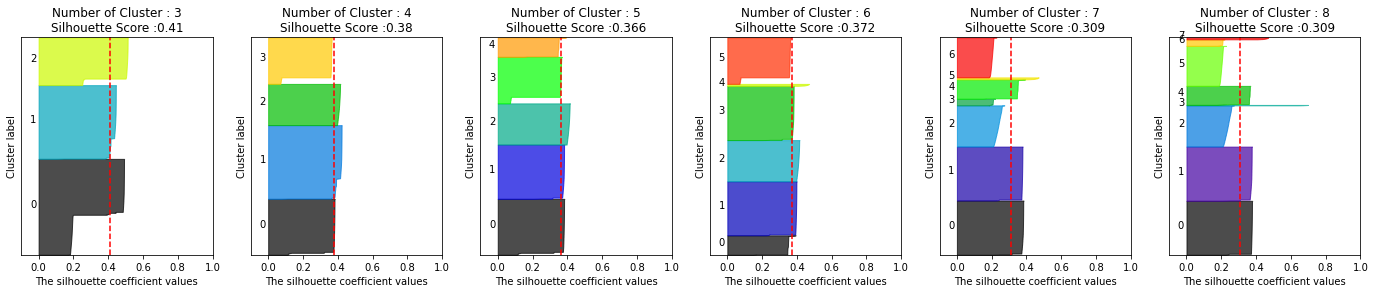
[표 3] INSP\_DTL\_CODE의 구분에 의한 Rule set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AD01** | | | | **AD02** |
| 구분(value) | value<100 | 100<=value<190 | 190<=v<400 | 400<=v<750 |  |
| 평균 | 16.97 | 154.30 | 294.85 | 543.27 | 10.01 |
| 표준편차 | 20.94 | 10.62 | 49.35 | 42.36 | 12.85 |
| 규칙 | 평균 ± 표준편차 | | | | |
| 정확도(%) | 88.47 | 89.33 | 60.55 | 72.94 | 98.78 |

마. 비지도 학습에 의한 데이터의 군집화

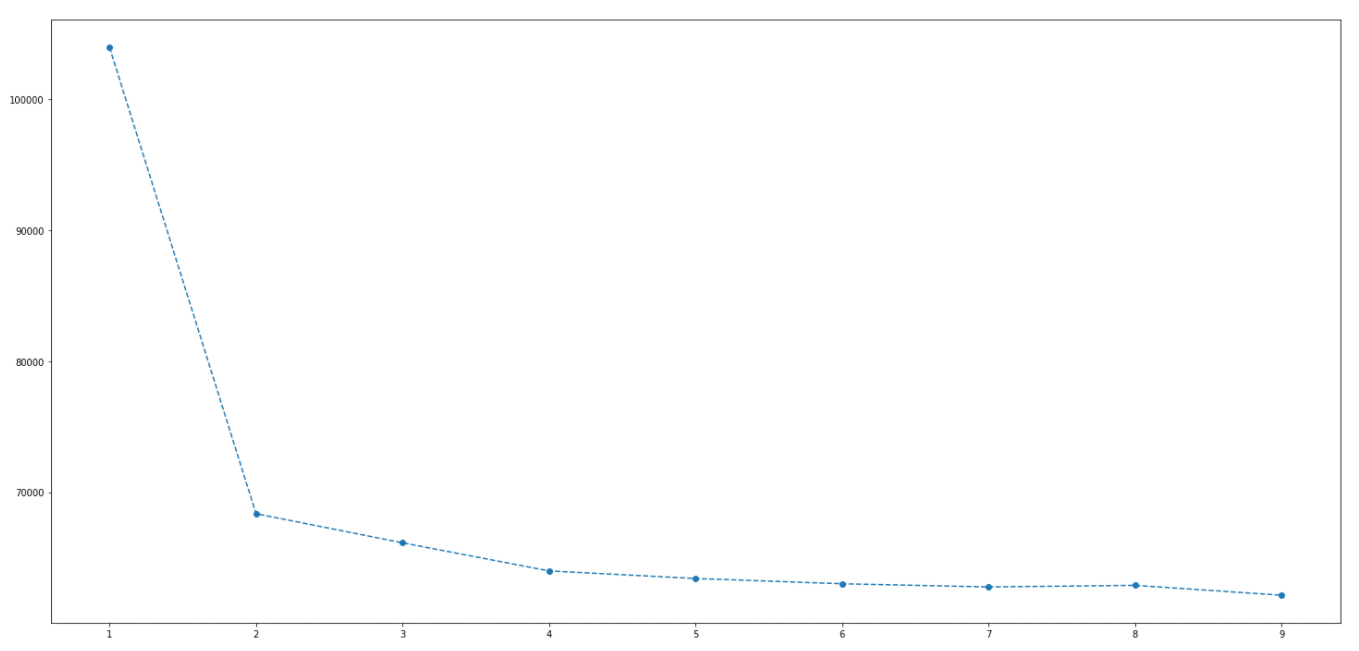
제품 번호와 INSP\_DTL\_CODE의 어떤 특성 조합이 건조기 검수의 실패와 통과를 결정하는지 비지도 학습을 통해 분석해 보았다. 우선, 표본을 500개 임의로 추출하여 one-encoding을 제품 번호(PROD\_SN)와 INSP\_DTL\_CODE, 두 특성 변수에 적용했다. 자료의 변동성을 95% 이상 설명하는 분산비를 기준으로 하면 차원의 수를 505개에서 460으로 줄일 수 있다. 클러스터링 개수를 정하기 위하여 실루엣 스코어와 데이터의 균질화를 기준으로 판단했다. [그림 10]은 클러스터링 갯수

[그림 10] PCA에 의한 클러스터링



[그림 11] K-means with PCA clustering

WCSS



Number of clusters

를 3개에서 8개로 변화했을 때, 실루엣 스코어와 데이터의 균질을 그래프로 나타낸다. 실루엣 스코어는 0.309에서 0.41로 대략 비슷하며 데이터의 균질화 측면을 목측으로 판단하면 4~6개 정도가 적당해 보인다. 추가적인 방법을 통해 클러스터링 개수를 결정해 본다. PCA를 통해 클러스터링은 WCSS(within-cluster sums of squares)에 기초하여 엘보우 메서드를 사용해서 결정했다[[1]](#footnote-1). [그림 11]에서 보면 4 이후에 부드럽게 곡선이 이어지므로 4개의 클러스터를 결정한다.

Approaches to Anomaly Detection

이상치 접근방법에는 크게 통계적인 방법, 기계학습/딥러닝, 하이브리드 방법이 있다. 하이브리드 방법은 시계열적인 특성을 갖는 데이터의 특징을 기계학습이나 딥러닝에 접목해 특성을 활용한다. 기계학습과 딥러닝에서 이상치 탐색방법으로 LSTM(Long Short Term Memory)나 AutoEncoder 방법을 최근 많이 활용하고 있다. LSTM은 알고리듬상 순차적인 데이터의 처리를 다루므로 시계열적인 특성을 녹여낼 수 있다. AutoEncoder 알고리듬은 특정한 잠재변수를 통해 입력된 데이터를 반영하여 출력값을 도출한다. 아래에서는 간략히 네 가지 알고리듬에 대한 개략을 기술했다.

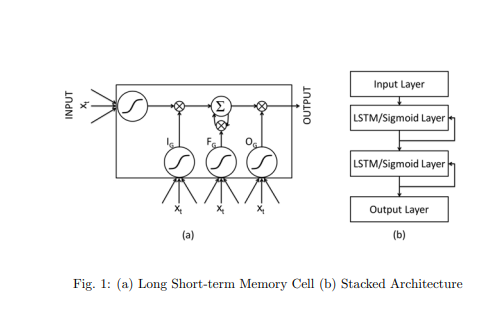
a.statistical modelling

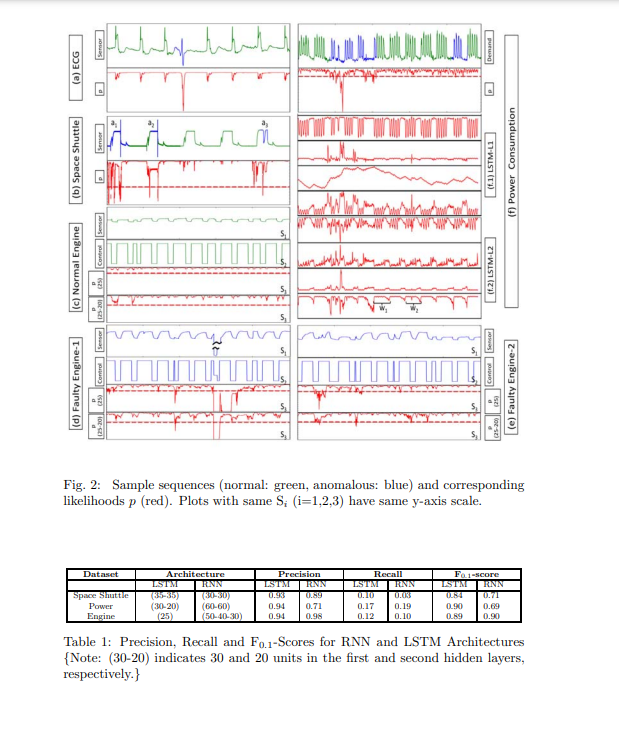
가.features characteristics

b.DL-based Moelling

가. features selection and extraction

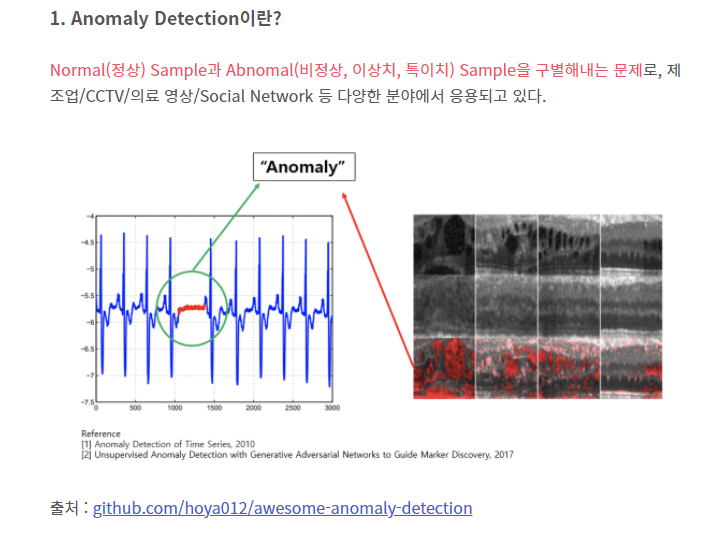
나. Anomaly Detection by LSTM



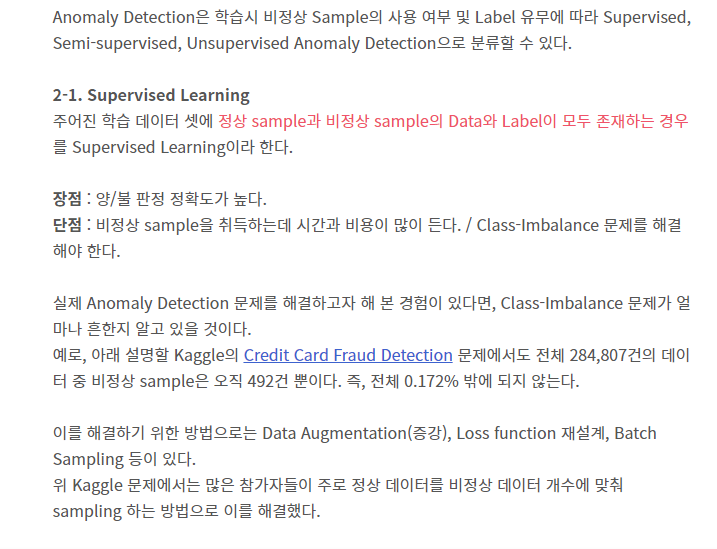


source : Malhostra, Vig, shroff, Agarwal, “Long Short Term Memory Networks for Anomaly Detection in Time Series”, ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligenceand Machine Learning. Bruges (Belgium), 22-24 April 2015

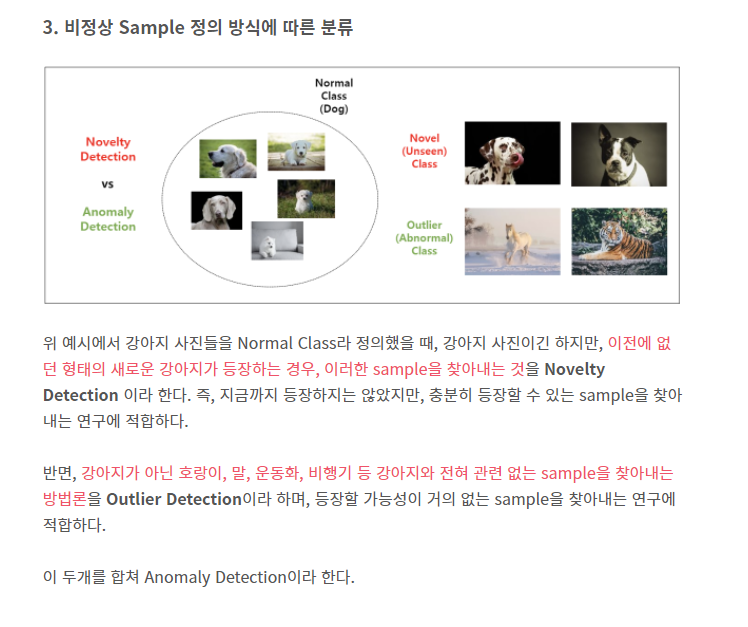
나. AutoEncoder



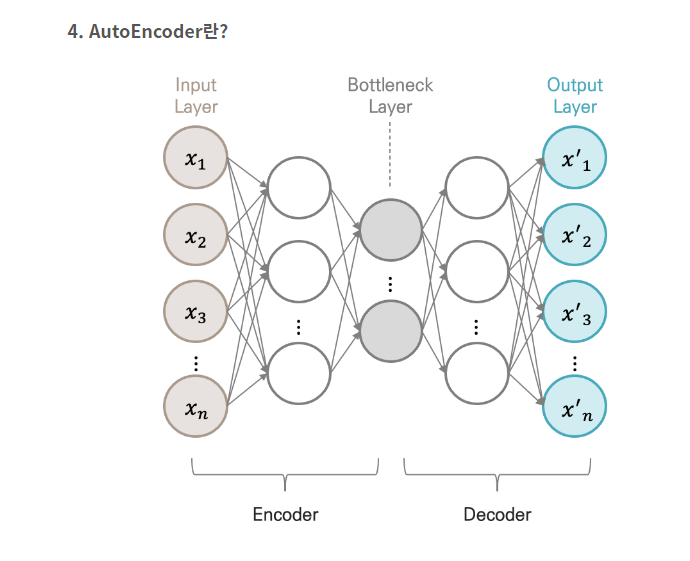
2. 표본 데이터의 특성: 불량과 양호의 제품표시가 있는 공정 데이터



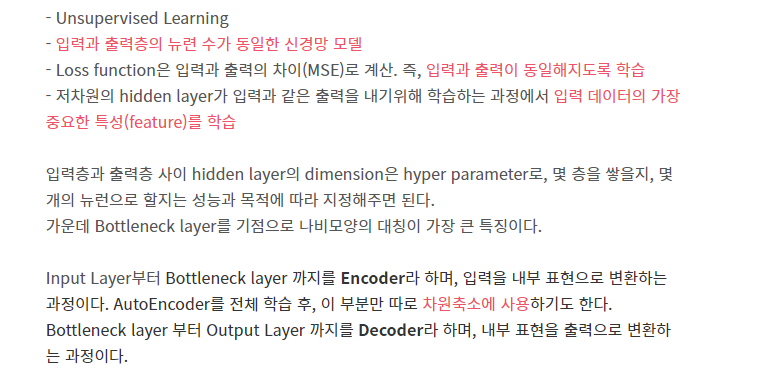
3. 이상치의 구별: 여기선 특이값 추출



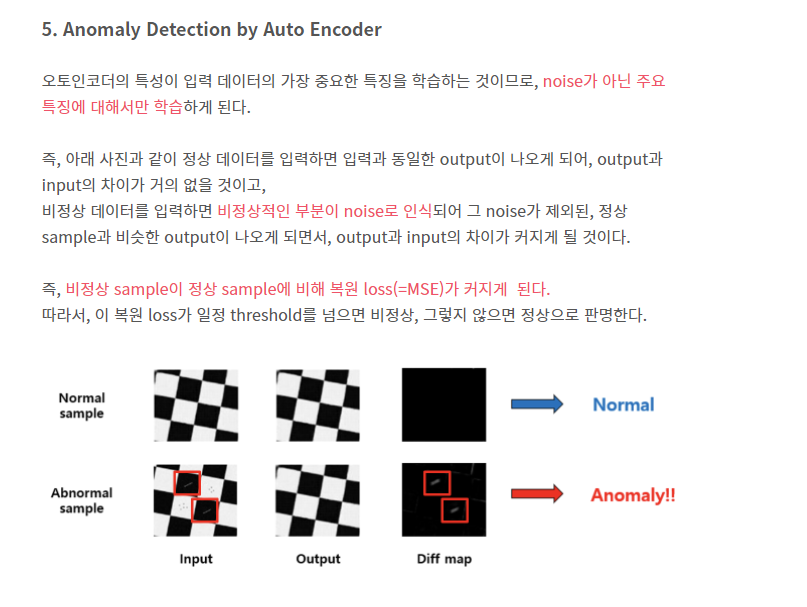
4. 검출 기술



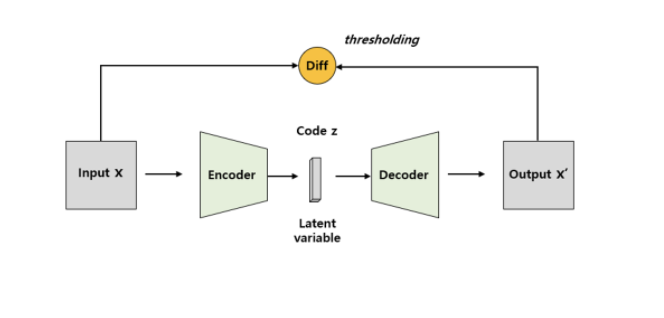
오토인코더의 특징:



5. Autoencoder 기술을 이용한 이상치 검출

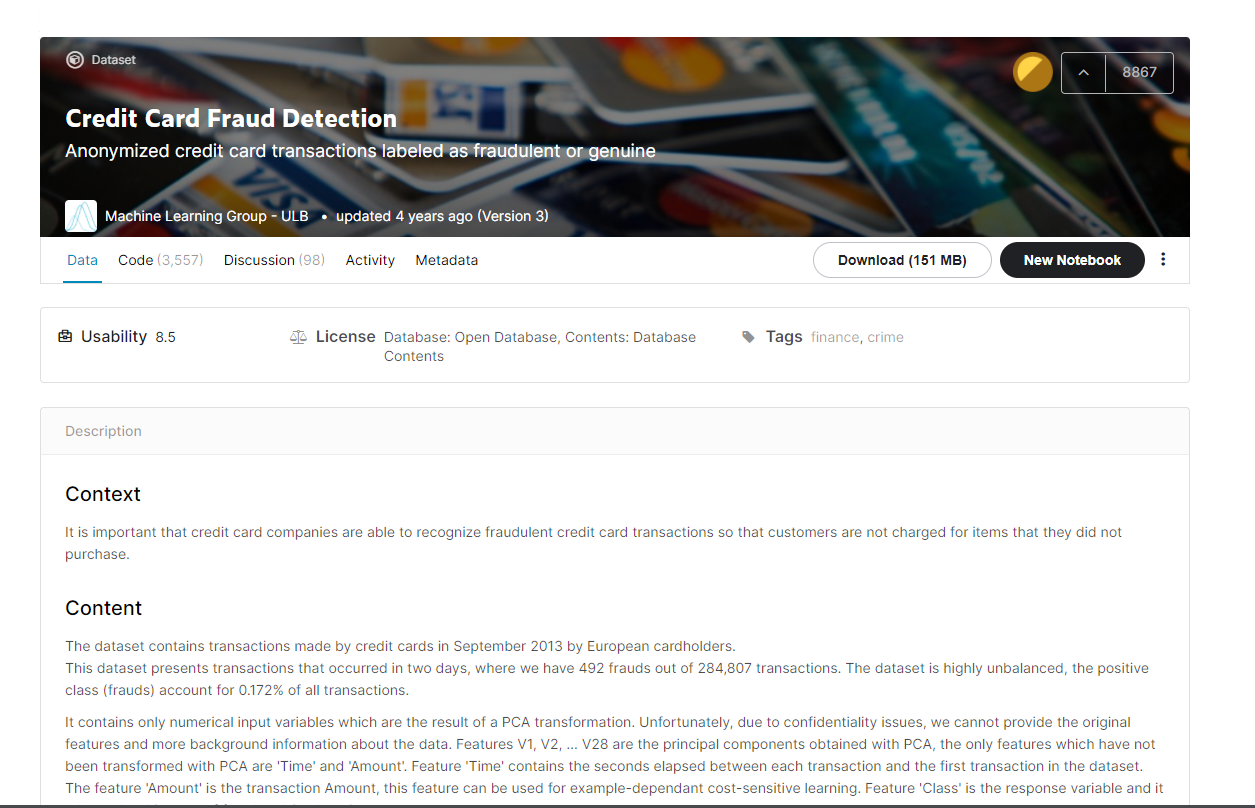


오토인코더의 다이어그램



자료출처 : <https://leedakyeong.tistory.com/entry/Anomaly-Detection-by-Auto-Encoder>

오토인코더 알고리듬을 사용한 사례 : 신용카드부정사용 검출



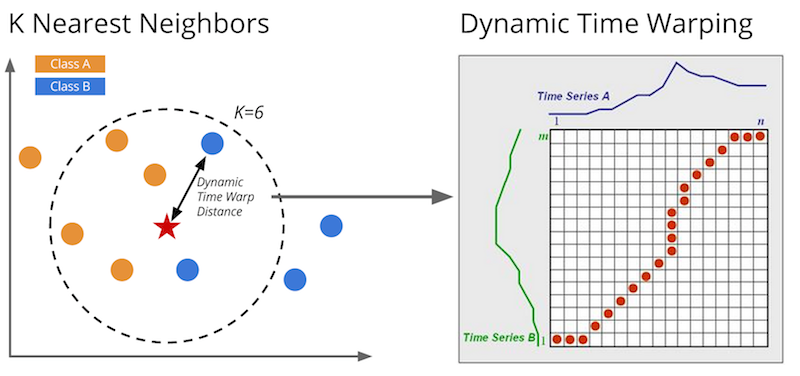
자료출처 : <https://www.kaggle.com/mlg-ulb/creditcardfraud>

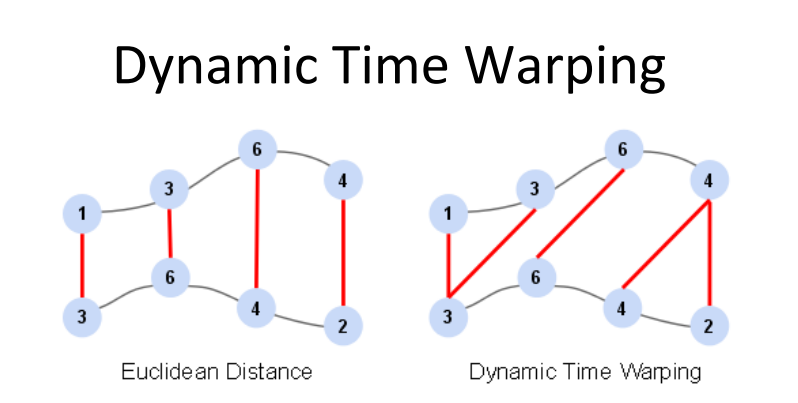
c. hybrid model of classification

가. K-Nearest Neighbors (with Dynamic Time Warping) for Time Series

KNN에서 유클리디안 거리 측정 대신 DTW를 사용한다. DTW는 두 시퀀스가 시간, 속도, 길이상에서 얼마나 유사한지를 측정한다.

- KNN with DTW is commonly used as a benchmark for evaluating time series classification algorithms because it is simple, robust, and does not require extensive hyperparameter tuning.





자료 출처: 구글 이미지

- While useful, KNN with DTW requires a lot of space and time to compute. During classification, the KNN-DTW compares each object with all the other objects in the training set. Further, KNN provides limited information about why a series was assigned to a certain class.

- KNN may also perform poorly with noisy series — the noise in a series may overpower subtle differences in shape that are useful for class discrimination

나. Time Series Forest Classifier

랜덤 포레스트를 시계열에 적용한 모형

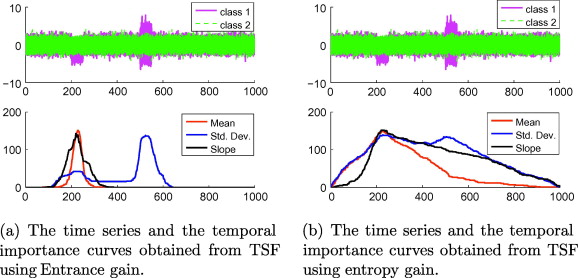
절차.

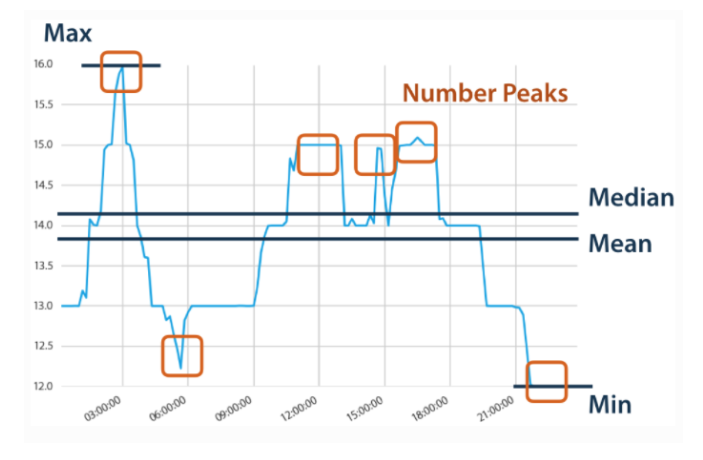
1.시계열 데이터를 출발시점과 길이를 임의적으로 정의해서 구간을 정한다.

2.각 구간에서 평균, 표준편차,기울기등 특성을 추출하여 하나의 벡터로 만든다.

3.추출한 특성에 의사결정나무를 훈련시킨다.

4.필요한 트리만큼 또는 시간이 마칠 때까지 1-3을 반복한다.





의사결정

새로운 시계열은 대표 투표(다수결 투표)에 의해 분류되고 예측은 랜덤 포레스트 모형에서 대부분의 트리가 예측한 값

- Experimental studies have demonstrated that time series forest can outperform baseline competitors, such as nearest neighbors with dynamic time warping

- Time series forest is also computationally efficient.

- Last, time series forest is an interpretable model. Time feature importance can be extracted from time series forest, as shown in the[sktime univariate time series classification demo](https://github.com/alan-turing-institute/sktime/blob/master/examples/02_classification_univariate.ipynb).

자료 출처: A Brief Survey of Time Series Classification Algorithms | by Alexandra Amidon | Towards Data Science

1. How to Combine PCA and K-means Clustering in Python? | 365 Data Science [↑](#footnote-ref-1)