**제조공정에서 제품품질에 영향을 미치는 변수의 탐색 및**

**제품품질 회귀모델**

**Studying the variables that affect product quality and Creating Regression model on the Manufacturing process**

고가연,국명준,김지호,김하영,성현우

중앙대학교 CUAI DA 2팀

**0. 초록**

DACON의 **<스마트 공장 제품 품질 상태분 류>**에서 제공되는 비식별화된 공정변수 데이 터를 기반으로 제품 품질을 예측하는 ML 모델 을 만들기 위한 탐색적 데이터분석(EDA)를 진 행한다. 해당 과정에서는 각 제품군별로 제조 되는 Line을 분석하여 공정 process를 나타 내는 비식별화된 변수인 X\_featur에 집중하여 결측치를 최대한 정확하게 채우는 EDA를 진 행했다. 또한 VIF, PLS-VIP및 피어슨 상관계수 등을 활용하여 공정변수들의 다중공선성을 개선함과 동시에 2800여개에 달하는 설비변수 들중 품질에 결정적인 영향을 주는 혐의설비를 찾아내고자 하였고 최종적으로 분류 (Classific ation) 모델 중 XGBoost 및 GradientBoosting 모델을 적용하여 완성한 머신러닝 모델을 DA CON에 제출하였다.

**1. 서론**

최근 코로나 펜데믹과 AI 기술의 발전으로 제조운영과 공급망에서의 디지털 혁신이 중요 한 화두로 떠오르고 있다. 대표적으로 스마트 공장을 통한 공정 데이터에서의 인사이트 발견 및 해석의 중요성이 강조되고 있으며 특히 스 마트폰 등 복잡한 정밀기기의 경우에는 제조공 정에서 사소한 불량설비 및 요인을 찾아 내는 것 만으로 전반적인 수율개선 및 수익 증가에 큰 영향을 줄 수 있다. 연구의 2장에서는 LG 사에서 제공한 각종 공정변수 및 품질수준에 관한 데이터를 EDA로 분석함으로써 데이터의 특성을 파악하고 knn방식을 통해 결측치를 보 간하며 3장에서는 머신러닝 모델을 통하여 해 당 데이터에서 공정변수에 따른 품질을 예측 한다. 4장에서는 해당 데이터세트의 다중 공선 성을 제거하고 변수를 대폭 축소한 모델을 추 가 예측에 활용하며 5장에서는 2~4장에서 소 개된 모델의 예측 결과를 DACON의 PRIVATE SCORE 기준으로 비교한다. 마지막 6장에서는 유의설비의 후보 변수를 파악하여 공정개선에 도움이 되고자 한다.

**2. 탐색적 데이터 분석(EDA)**

**2.1 이상 피처 제거**

주어진 Dataset의 핵심 feature는 ‘LINE’, ‘PRODUCT\_CODE’, ‘X\_1~X\_2870’이다.

매우 많은 피처가 있는 만큼, 최대한 유효한 데이터를 우선적으로 선별하는 것을 우선으로 수행하였다. 각 피처에서 유효 데이터가 40개 미만인 피처를 제거했으며, 해당 피처는 총 데 이터의 10%를 차지했다. 또한 분산이 1인 피 처도 Classification 모델에 도움이 되지 않으니 372개의 피처를 제거했다.

이렇게 보정한 데이터의 피처는 trainset에서 2337개, testset에서는 2335개였다.

.

**2.2 결측치관련 통계량 파악**

trainset에서는 2개의 PRODUCT\_CODE인 A\_31과 T\_31이 있었으며, A\_31 제품은 T010305, T010306, T050304, T050307 4개의 LINE에서만, T\_31 제품은T100304, T100306 2개의 LINE 에서만 제조된다는 것을 확인 했다. 결측치보간 측면에서 비식별화된 변수인 X\_feature는 공정 Process를 나타내며 이들의 배열이 공정 순서를 나타내는 것은 아니지만, 연관이 없는 값들끼리 이웃한 관계는 아니라는 점은 중요한 정보이다. 따라서 제품별로 특정 LINE을 거쳐 공정이 이루어지는데, LINE 별로 모든 X\_feature를 거치는 것이 아니며 특정 스크림에서만 공정이 이루어진다고 가정을 했 다. 즉 각 피처별로 분포하는 결측치가 단순한 Missing 값이 아닌 공정이 이루어지지 않는다 는 의미일 수 도 있다는 의미이다.

이를 확인하기 위해 LINE 별로 진행되는 공정 X\_feature들을 결측치의 분포에 따라 추정하였 다. (표1)은 ‘LINE’ 피처별로 그룹화 하였을때 X\_feature 중 그 값이 아예 존재하지 않는 것

의 개수를 세어 나타낸 도표이다.

|  |  |
| --- | --- |
| **LINE** | **Missing Value** |
| T010305 | 1577 |
| T010306 | 1579 |
| T050304 | 641 |
| T050307 | 643 |
| T100304 | 1783 |
| T100306 | 1784 |

**(표1) 값이 아예 존재하지 않는 X\_피처의 개수**

|  |  |  |
| --- | --- | --- |
| **LINE1** | **LINE2** | **결측치 분포 차이** |
| T010305 | T010306 | **14** |
| T050304 | T050307 | **3** |
| T100304 | T100306 | **369** |

**(표2) 각 LINE 별 결측치 분포의 차이가 나는 피처의 개수**

표에서 확인할 수 있는 것처럼 A\_31에 관여 하는 LINE인 "T010305" &"T010306" 과 "T050304" & "T050307", T\_31에 관여하는 LIN E인 "T100304" & "T100306"는 각각X\_featur e가 결측되어 있는 개수가 매우 비슷하다. 또한 이를 통해 각LINE별로 그룹화하여 결측 치분포 차이를 확인한 결과 (표2)는 결측치 분 포가 매우 유사한 양상을 보인다는 것이다. 이를 통해 앞서 진행한 가정인 “제품별로 특정 LINE을 거쳐 공정이 이루어지는데, LINE 별로 모든 X\_feature를 거치는 것이 아니며 특정 스크림에서만 공정이 이루어 진다”를 증명 할 수 있다. 다만, Missing Value의 개수가 비슷 하지만 일부 차이가 존재하며 각 그룹별, 총 3개의 그룹에 대해서 다시 한 번 그룹화하여 최종적으로 결측치를 채우는 과정을 진행한다.

**2.3 결측치 채우기**

결측치를 채우는 핵심은 2가지이다. 첫번 째는 LINE별로 line\_groups라는 3개의 그룹 을 형성해 각각의 LINE 별로 존재하는 피처끼 리 결측치를 채운다. 이때 KNN Imputer 알고 리즘을 사용한다. 이는 NA 값 가장 가까운 주 변 k개의 평균을 NA 값으로 대체하는 알고 리즘이다. X\_feature의 순서에 의미가 없는 것이 아니기 때문에, 상관관계가 높은 피처 들 사이에서 KNN Imputer를 적용한다면 결측 치를 효과적으로 채울 수 있다. LINE 별로 그 룹화된 피처들 중 서로간의 상관계수가 0.6이 상인 피처들을 필터링하여 해당 피처들의 결 측치를 KNN imputer(n\_neighbors=5)로 채운 다. 그리고, 앞서 말한 것과 같이 나머지 피처 들의 결측치들은 해당 스크림에서 공정이 진 행되지 않았음으로 판단하고, 모두 0으로 채 우면서 최종적으로 EDA를 마무리한다.

**3. ML 모델**

DACON에 제출하기 위한 머신러닝 모델은 Classification 중 XGBoost를 적용한다. 해당 모델은 트리 기반 학습이므로 일반적으로 별 도의 정규화, 피처 스케일링이 필요없기에 수 치형 데이터간 차이가 심한 본 데이터 세트에 서는 효율적일 것으로 예상된다. 본 EDA방식 은 다중 피처들에 대해서 모두 고려한 학습 방식이며 2,000개가 넘는 피처를 학습하는 것은 과적합(Overfitting) 가능성이 높아진다. 따라서 이후에는 다중공선성을 이용한 데이

터 선별 과정 후 동일 ML 모델을 적용하며 학습을 수행하는 모델과 대조하는 연구를 진행한다.

**4. 다중공선성 개선**

**4.1 Background**

4.1.1 VIF

다중 공선성은 하나의 피쳐가 다른 피쳐의 조합으로 표현될 수 있는 경우를 의미한다. 머신러닝 모델을 사용할 때 사용자는 피처 간에는 서로 독립인 상태라고 예상한다. 반면 상관관계가 존재하는 피처로 회귀모델을 만들면 그 피처에 과대적합된 결과를 예측하 고, 이는 예측 결과의 부정확성을 높인다. 따라서, 많은 연구자들은 다중 공선성을 줄이 기 위해 주로 VIF 알고리즘을 사용하여 피처 들간의 상관관계를 높이는 피처를 찾아 제거하는 방법을 사용한다.

VIF=  **(1)**

식 (1)에서 R\_i 값은 독립변수 X\_i를 다른 변수로 선형 회귀하는 성능을 의미한다. 즉, X\_i와 다른 변수 간의 상관관계가 높다면 R\_i값은 높아지고, 결과적으로 VIF의 값도 커지게 된다. 단순하게 목표 값과 상관관계가 높은 변수만을 선택한다면 그 변수끼리

VIF값이 클 가능성이 높으므로 이 값이 일정 수준 이상 크다면 그 변수는 제외하는 것이 바람직하다.

4.1.2 PLS-VIP

Partial least squares (PLS) 회귀는 변수 추출 법의 하나로, 다중 선형 회귀 모델을 통해 독립변수인 X와 종속변수인 Y의 상관관계를 반영하여 회귀 모형을 추정하는 방법이다. 독립변수와 종속을 변수를 잘 설명할 수 있는 새로운 잠재변수를 도출하여 의미 있는 분석을 가능하게 한다. 전통적 다중 회귀와는 다르게 다중공선성, 노이즈를 갖는 대량의 독립변수 데이터도 분석이 가능하다.Variable importance in projection (VIP) 값은 독립변수 X가 PLS 모델에 미치는 영향을 정리한 것으로, 잠재변 수가 종속변수인 Y의 분산에 미치는 영향을 고려한 PLS 가중치의 제곱의 가중합으로 계산 할 수 있다. VIP score는 종속변수 Y의 분산에 영향력을 미치는 독립변수의 중요도를 평가하 는 지표로 사용할 수 있다.

**4.2 다중공선성 개선 및 피처 줄이기**

2장과 3장에서 학습한 데이터세트로 품질 예측이 가능하지만 2300여개의 X피처를 사용 한다는 점이 실제 상황에서 설비의 점검을 부 담스럽게 만든다고 생각하여 해당 피처 중 품 질을 나타내는 Y\_CLASS 피처와 피어슨 상관 계수가 0.6 이상인 176개 X피처 를 선별하여 다중공선성 개선을 진행하였다.

Condition number = **(2)**

한편 식 (2)는 최대Eigen Value 와 최소 Eigen

 Value의 비를 나타내는 Condition number로 해당 값이 클수록 높은 다중공선성을 지닌 모 형일 가능성이 높고 회귀에서 과적합을 유발 할 수 있다. 따라서 VIF를 적용하기 전에 X 피처들을 정규화하여 condition number를 줄여 다중공선성을 1차로 개선시킨다. 이후 해당 데이터세트를 Kmeans clustering을 통하여 176개의 X피처를 4개의 집합으로 구분하였 고 그 중 대부분인 162개의 피처가 하나의 그룹으로 분류되었다. 따라서 나머지 그룹의 피처들은 제거하지 않고 162개의 피처를 포함 한그룹에서 VIF값이 높은 피처들 중 내림차 순으로 나열했을 때 홀수번째 위치한 피처들 을 제거해 99개의 피처를 선별하였다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

**(그림 1) 공정변수들의VIF**

마지막으로, 선별된 99개의 X피처 중 테스트 세트에서 대부분이 결측치로 존재하는 피처를 제거하고 LINE 피처와 PRODUCT\_CODE피처를 추가하여 총16개 피처만으로 Xgboost회귀를

진행하였으며 해당 피처는 아래와 같다.

**[선별된 X피처 14개]**

**X\_318 X\_367 X\_368 X\_372 X\_373**

**X\_374 X\_240 X\_1833 X\_2466 X\_2779**

**X\_2787 X\_2841 X\_130 X\_131**

**5. 예측 결과**

각 실험은 DACON의 <스마트 공정 제품 품질 분류 데이터>에서 train set (598 rows, 2881 columns), test set (310 rows, 2879 col umns; Y\_quality, Y\_Class feature 제외)를 기준

데이터로 사용하였다. 모든 실험은 2~4장에서 각각 언급한 결측치 처리를 적용하여 진행하였

고 후술할 조건 외에는 모두 같은 조건 내에서 진행하였다.

2장에서는 전처리한 데이터를 XGBC와 GBC모델을 통한 앙상블로 예측하는Raw test 모델을 만들었다. Train set에서 NULL값이 유 의미하게 존재하는 feature를 선별하여 약 2 0 %의 feature를 제외한 2347개의 feature를 가지고 예측 모델을 만든것으로 DACON의m1 score(Private score)를 측정하였다.

(그림 2. Raw test)

다음으로, 전처리한 데이터를 VIF를 통해 다중 공선성을 제거하고 XGBC와 GBC 모델을 이용해 예측 모델을 만들었다. 상기 과정을 통해 남은 2347개의 feature에서 VIF 값을 측정하여 가장 상관관계가 높은 feature를 제거했다. 이때, VIF가 높은 모든 feature를 제거하면 예측 모델에 사용할 feature가 너무 적어질 수 있으니 전체 feature 중에서 절반만 선택하여 VIF가 높은 feature를 제거하여 m1 score를 측정했다. (그림 2. VIF)

마지막으로 4장에서 다루었던 모델로

앞선 VIF 모델과 달리 Y\_quality (목표 값)와의 피어슨 상관계수가 높은 피처에 한하여 VIF 까지 고려한 모델을 활용하였다. 이는 최소 한의 feature 16개만을 이용한 예측 모델로, 역시 m1 score를 측정했다.

(그림 2. VIF\_least\_feature)

측정 결과는 그림(3)에서 확인할 수 있다

대조군으로 사용한 Raw test 실험에 비해서 VIF 실험은 m1 score의 개선이 크게 이루어진 것을 통해 다중 공선성을 제거하면 예측 결과 향상에 도움이 된다는 것을 확인할 수 있다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Raw test** | **VIF** | **VIF\_least\_feature** |
| **Number**  **of used features** | 2347 | 1910 | 16 |

**(그림 2) 실험에서 사용한 차원 크기**

**텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명 (feature 개수) (上)**

**(그림 3) 실험에서 test set를 이용해 얻은**

**m1 score (下)**

또한, VIF\_least\_feature 실험의 결과를 보면 Raw test 실험과 비슷한 m1 score를 보이는 것을 확인할 수 있다. 그런데, 사용한feature

의 수가 146배이므로 VIF\_least\_feature의 결 과는 생성된 예측 모델이 가벼우면서 예측 성 능은 유지한다는 의미이다. 이 결과는 추후 실 험을 통해 Y\_quality와 상관관계가 높은 독립 변수의 수를 늘린다면 현재 얻은 m1 score 보 다도 더 많이 향상될 가능성이 있다는 것을 시 사한다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명스크린샷, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

**(그림 4) 실험에서 사용한 차원 크기**

추가로 데이터 세트 中 각 feature의 VIP 방식을 통해서도 측정해본 결과 <그림 4>와 같이 나타났다. VIP 값이 0.83 이상인 피처는 719개로, 해당 피처 를 통해 학습시켜 얻은 m1 score는 0.4944이다. 앞서 VIF를 통해 다 중공선성을 제거한 실험들과 비교했을 때 m1 score가 낮은 수치임을 확인할 수 있다. 하지 만 앞서서 VIF 알고리즘을 통한 다중공선성 제 거 후의 score가 Raw test의 score보다 더욱 높았던 것으로 보아, 공정 데이터의 피쳐는 다 중 공선성 문제를 가지고 있음을 알 수 있고

해당 719개의 피쳐 역시 다중공선성을 추가로 줄이는 방식을 통해 예측 성능을 향상시킬 수 있음을 기대한다.

**6. 유의설비**

5장에서 다룬 VIF\_least\_feature 모델에서

다중공선성과 상관계수를 고려하여 14개의

핵심적인 X feature를 선별하였는데 해당 X feature와 높은 유의성을 가지는 feature를 선별한 후 ‘유의설비집합’ 이라고 이름붙혔다. 높은 유의성의 기준은 피어슨 상관계수로, 2800여개의 모든 X feature中 해당 14개 설비와 피어슨상관계수 0.8 이상인 총 50개의 feature들이 선별되었다. 유의 설비집합은 불량률에 큰 영향을 미치는 대표적인 피처를 선별한 것이므로 엔지니어의 입장에서 불량률이 높다면 유의설비 집합에 속하는 X피처의 점검을 우선시하는 것을 권한다.

**(그림 5) 유의 설비 집합 선정**

**7. 결론**

본 연구는 DACON이 제공하는 <스마트 공 장 제품 품질 상태분류> data set를 활용한 연 구로 공정의 결과인 ‘품질변수’ 및 이와 관련된 2800여개의 설비 변수를 다양한 방식을 통해 전처리 및 회귀분석하고 그 결과를 비교하는 내용을 담고있다.

전처리 과정에서 EDA방식을 통하여 데이터 의 독특한 분포 특성을 발견할 수 있었고 이를 결측치 보간에 활용할 수 있었다. 한 편으로는 다중공선성을 개선함과 동시에 품질에 영향을 크게 미치는 설비변수들을 선정하여 품질의 회 귀 예측성능을 향상시켰다. 특히 다중공선성의 개선은 회귀예측성능에 중요한 영향을 미치는 것으로 확인 되었고 이는 독립변수간 상관관계 가 높은 다른 데이터 세트의 회귀 분석에도 활 용될 수 있을 것 으로 보인다.

또한, 이 과정에서 2800여개의 설비 변수 중 품질에 핵심적 인 영향을 미치는 몇 가지 설비변수들을 찾아 낼 수 있었는데 이를 활용 하여, 방대한 종류의 설비변수 데이터를 해석 및 점검 해야 하는 엔지니어들이 부담을 덜 수 있기를 기대한다.

**참고문헌**

최두원 외2인(2013),Discovery of Process Equipments Causing Product faults in a Nano-Scale Manufacturing Line, 한국경영과학회

Gi-Sung Lee1 , Jong-Chan Lee(2022),Data Analysis and AI Model for Defect Prediction in the Injection Process, Journal of the Korea Academia-Industrial cooperation Society

Il-Gyo Chong, Chi-Hyuck Jun (2005), Performance of some variable selection methods when multicollinearity is present, Chemometrics and Intelligent Laboratory Systems, Volume 78, Issues 1–2

Svante Wold, Michael Sjöström, Lennart Eriksson(2001), PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, Volume 58, Issue 2