Meltingtank Dataset

전자제조데이터 분석 프로젝트

22510106 김신호

22532002 박가문날비

SEOULTECH

Contents

용해탱크(용해공정) 데이터셋(Meltingtank Dataset)

- 용해공정은 분말 원재료를 액상 원재료에 녹이는 공정으로 식품, 화학, 석유화학 등 다양한 분야에서 적용됨
- 본 프로젝트에서의 용해공정은 분말 유크림, 기능성 조제 분말등을 생산하는 식품제조업의 용해공정으로 SD/MSD 건조생산라인의 원료 전처리 작업의 첫 번째 단계에 속함

분석 목적

- 용해공정(용해탱크)에서는 분말 및 액상 원재료를 정제수 등에 용해/혼합 후, 후공정에서 다시 분말화하기 때문에 용해탱크에서 원재료가 균일하게 혼합되는 것이 매우 중요함
- 그러나 현장에서는 완제품의 주요 요인을 모두 고려하고, 설비운영 기준값에 따라 공정을 운영하여도, 용해품질에 영향을 미치는 다른 요인들이 존재하며, 경험과 노하우 등 암묵지에 의존하여 대처할 수 밖에 없으나 이마저도 인력 공백으로 인해 대처가 어려운 실정임

분석 목표

- 용해탱크 설비운영값(독립변수)과 주요 품질검사항목(종속변수) 의 결과값을 통해 생산품질을 예측할 수 있는 모델을 생성 후 검증을 진행
- 생산품질에 영향을 주는 여러 요인들을 분석

EDA

변수명	설명	데이터타입	데이터 개수	결측치 여부
STD_DT	날짜(YYYY-DDHH:MM:SS)	Object	835,200	Non-Null
MELT_TEMP	용해 온도	Int64 (연속형)		
MOTORSPEED	용해 교반속도	Int64 (연속형)		
MELT_WEIGHT	용해탱크 내용량(중량)	Int64 (연속형)		
INSP	생산품의 수분함유량(%)	Float64 (연속형)		
TAG	불량여부	Object (OK,NG)		

- 총 6개의 Attribute로 구성됨
- 4개의 독립변수와 1개의 종속변수로 구분됨
  - 독립변수 : MELT_TEMP, MOTORSPEED, MELT_WEIGHT, INSP
  - 종속변수 : TAG
- STD_DT : 시계열 모델 생성을 위한 데이터 셋의 인덱스로 사용함

Summary Statistics

	MELT_TEMP	MOTORSPEED	MELT_WEIGHT	INSP
count	835200.000000	835200.000000	835200.000000	835200.000000
mean	509.200623	459.782865	582.962125	3.194853
std	128.277519	639.436413	1217.604433	0.011822
min	308.000000	0.000000	0.000000	3.170000
25%	430.000000	119.000000	186.000000	3.190000
50%	469.000000	168.000000	383.000000	3.190000
75%	502.000000	218.000000	583.000000	3.200000
max	832.000000	1804.000000	55252.000000	3.230000

✓ MELT_TEMP, MOTORSPEED
- 소수점 1자리가 생략되어 있음

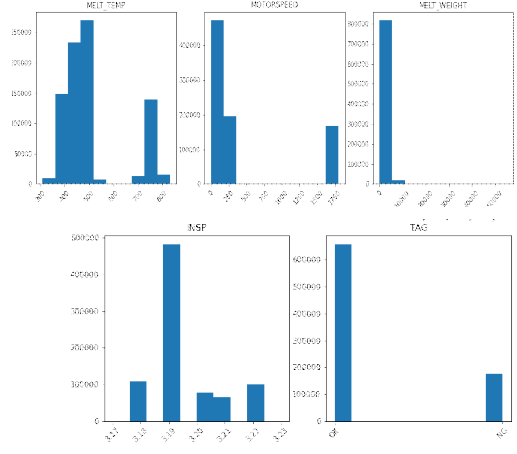
① MELT_TEMP
- Max : 평균에서 많이 벗어나는 이상치 존재

② MOTORSPEED
- min : 설비를 중지한 경우 0으로 표기
- Max : 180이 넘는 이상치 존재

③ MELT_WEIGHT
- Std : 1217 로 차이가 큰 편

① INSP
- Std : 0 으로 데이터들이 고르게 분포되어 있음

Histogram



✓ MELT_TEMP, MOTORSPEED, MELT_WEIGHT, INSP, TAG

① MELT_TEMP
- 비교적 균형적인 분포

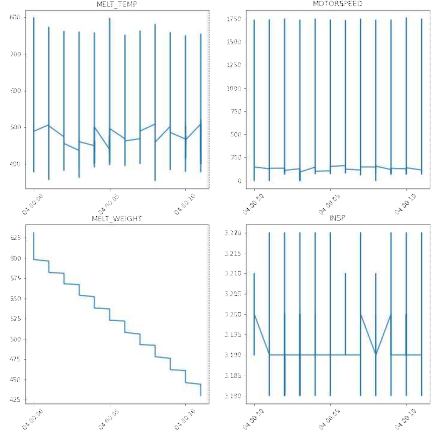
② MOTORSPEED
- 데이터가 극단적으로 분포되어 있음

③ MELT_WEIGHT
- 데이터가 한 쪽으로 치우쳐 있음.

④ INSP
- 비교적 균형적인 분포

⑤ TAG
- 종속변수인 TAG의 경우 클래스 불균형이 존재

Pattern Analysis



✓ MELT_TEMP, MOTORSPEED, MELT_WEIGHT
- 모두 특정 패턴을 보여주고 있음

① MELT_TEMP, MOTORSPEED
- 비교적 일정한 패턴

② MELT_WEIGHT
- 지속적으로 감소하는 패턴

③ INSP
- 비교적 불규칙한 패턴

Correlation Analysis(Pearson)

	NUM	MELT_TEMP	MOTORSPEED	MELT_WEIGHT	INSP	TAG
NUM	1.000000	0.000394	-0.000300	0.054676	-0.000551	0.070789
MELT_TEMP	0.000394	1.000000	0.219660	0.000080	0.620046	0.293098
MOTORSPEED	-0.000300	0.219660	1.000000	0.000144	0.287613	0.174996
MELT_WEIGHT	0.054676	0.000080	0.000144	1.000000	-0.000309	-0.036110
INSP	-0.000551	0.620046	0.287613	-0.000309	1.000000	0.244336
TAG	0.070789	0.293098	0.174996	-0.036110	0.244336	1.000000

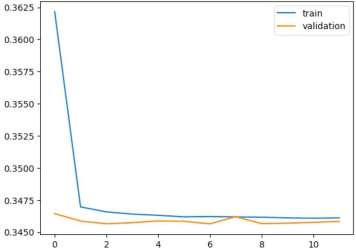
- 종속변수인 'TAG'와 독립변수들과 상관분석을 진행했을 때, 'MELT_WEIGHT' 를 제외하고는 모두 양의 상관관계
- 'TAG' 는 'MELT_WEIGHT' 를 제외한 나머지 변수들과 선형 관계를 갖는다고 할 수 있음.
- 상관관계는 낮게 나오지만 어떤 영향을 미치는지 알아보기 위해 'MELT_WEIGHT' 를 포함하여 모델분석 진행

Modeling

Data Preprosssing

Classification		
Processing	Machine Learning	Deep Learning
Columns	MELT_TEMP, MOTORSPEED, MELT_WEIGHT, INSP / TAG	
Window Function	-	Window size(timestep) : 10
Train / Test split	Train set : 70%, Test set : 30%	
Data Normalization	MinMaxScaler	
Imbalanced Data	SMOTE	
Modeling	Tree-based Methods(DT, RF, CB, XG, LGBM)	LSTM(Long Short Term Memory)

Model Verification – LSTM



train / validation loss plot

	실제값(N)	실제값(P)
실제값(N)	2155	785
실제값(P)	47,955	199,655

Confusion Matrix

f1-score : 0.8912
accuracy score : 0.8055
precision : 0.9961
recall : 0.8063

Modeling - Machine Learning(Tree-based Methods)

Model Verification : Confusion Matrix(f1-score, accuracy_score, precision, recall)

Xgboost - RandomForest - Decision Tree - LightGBM - Catboost

Decision Tree

22,600	30,385
39,840	157,735

f1-score : 0.8184
accuracy_score : 0.7204
precision : 0.8385
recall : 0.7994

RandomForest

22,565	30,420
39,629	157,946

f1-score : 0.8265
accuracy_score : 0.7344
precision : 0.8522
recall : 0.8023

Xgboost

22,798	30,187
29,137	168,438

f1-score : 0.8502
accuracy_score : 0.7632
precision : 0.8480
recall : 0.8525

Catboost

42,074	10,911
64,973	132,602

f1-score : 0.7775
accuracy_score : 0.6971
precision : 0.9239
recall : 0.6711

LightGBM

39,831	13,154
60,379	137,196

f1-score : 0.7886
accuracy_score : 0.7065
precision : 0.9125
recall : 0.6943

Modeling

VI(Variable Importance) – Permutation Importance

Decision Tree

0.0589 ± 0.0014	MELT_TEMP
0.0271 ± 0.0014	MOTORSPEED
0.0117 ± 0.0016	MELT_WEIGHT
0.0027 ± 0.0007	INSP

CatBoost

0.0699 ± 0.0013	MELT_TEMP
0.0103 ± 0.0005	MELT_WEIGHT
-0.0069 ± 0.0003	INSP
-0.0123 ± 0.0013	MOTORSPEED

RandomForest

0.0343 ± 0.0014	MELT_TEMP
0.0120 ± 0.0012	MELT_WEIGHT
-0.0032 ± 0.0013	INSP
-0.0111 ± 0.0008	MOTORSPEED

LightGBM

0.0641 ± 0.0012	MELT_TEMP
0.0127 ± 0.0005	MELT_WEIGHT
0.0008 ± 0.0005	INSP
-0.0153 ± 0.0009	MOTORSPEED

XGBoost

0.0411 ± 0.0009	MELT_TEMP
0.0137 ± 0.0013	MELT_WEIGHT
0.0000 ± 0.0010	INSP
-0.0017 ± 0.0004	MOTORSPEED

Feature Selection

1. MELT_TEMP
2. MELT_WEIGHT
3. ~~INSP~~
4. ~~MOTORSPEED~~

Conclusion

- 본 프로젝트에서는 용해탱크 데이터 셋을 통해 총 2가지의 분석 목표를 가짐

  ① 설비운영값과 주요 품질검사항목의 결과값을 통해 생산품질을 예측할 수 있는 모델을 생성 후 검증을 진행
  ② 생산품질에 영향을 주는 여러 요인들을 분석

- EDA

  Correlation : **MELT_WEIGHT(용해탱크 내용중량)** 독립변수가 종속변수와 가장 관련이 없는 것으로 보였음
  Pattern : **INSP(수분함유량)**이 가장 불규칙한 패턴을 보였음

- Modeling : Deep Learning > Machine Learning

  Deep Learning : LSTM
  Machine Learning(Tree-based) : Xgboost → RandomForest → Decision Tree → Catboost → LightGBM
  VI – Permutation Importance : MELT_TEMP, MELT_WEIGHT

- 설비운영값과 주요 품질검사항목의 결과값을 통해 분석을 진행한 결과 예측 모델에서는 Deep Learning(LSTM)의 성능이 가장 우수하였고, EDA에서는 MELT_WEIGHT가 종속변수와 관련 없는 변수로 도출되었으나 Machine Learning 모델의 VI를 통해 MELT_TEMP, MELT_WEIGHT의 독립변수가 성능 변화에 가장 큰 영향을 주는 요인으로 확인됨

THANK YOU!