감귤 착과량 예측 AI 경진대회

쥬혁이

Gamgyul

CONTENTS

001

데이터 전처리

1. 엽록소 관련 feature 제거

엽록소 측정에 오차 및 정확한 데이터라고 판단되지 않아 제거하는 방향으로 결정.

또한 제거하였을 때 성능이 올라가는 것을 확인하여 최종 데이터 전처리에 포함.

```
train.drop(train.filter(regex = '엽록소').columns, axis = 1, inplace=True)
test_x.drop(test_x.filter(regex = '엽록소').columns, axis = 1, inplace=True)
```

001

데이터 전처리

2. 새순 측정 feature 보완

데이터를 보았을 때 모두 소수점 한자리로 측정되는 것을 확인.

->이는 반올림이나 올림, 내림과 같은 어떠한 전처리를 거친 데이터일 수 있다는 가설을 세어 데이터 전처리를 진행.

또한 새순은 날짜가 지날수록 점점 감소하는 추세를 보이는 경향을 가짐.

->t+1 시점의 새순값이 t시점보다 새순값이 높다면 t시점의 새순값으로 변경

```
new_soon_tr = train.filter(regex = '새순')
new_soon_tr = new_soon_tr[sorted(new_soon_tr.columns)]
for i in range(len(new_soon_tr)):
  for j in range(len(new_soon_tr.columns) - 1):
    if(new_soon_tr.iloc[i,j] < new_soon_tr.iloc[i,j+1]):
      new_soon_tr.iloc[i,j+1] = new_soon_tr.iloc[i,j]
    if((new_soon_tr.iloc[i,j] == new_soon_tr.iloc[i,j+1]) | (new_soon_tr.iloc[i,j+1] == 0)):
      new_soon_tr.iloc[i,j+1] = new_soon_tr.iloc[i,j] - train.iloc[i,-1]
new_soon_tr[new_soon_tr < 0] = 0
new_soon_tr


new_soon_te = test_x.filter(regex = '새순')
for i in range(len(new_soon_te)):
  for j in range(len(new_soon_te.columns) - 1):
    if(new_soon_te.iloc[i,j] < new_soon_te.iloc[i,j+1]):
      new_soon_te.iloc[i,j+1] = new_soon_te.iloc[i,j]
    if((new_soon_te.iloc[i,j] == new_soon_te.iloc[i,j+1]) | (new_soon_te.iloc[i,j+1] == 0)):
      new_soon_te.iloc[i,j+1] = new_soon_te.iloc[i,j] - test_x.iloc[i,-1]
new_soon_te[new_soon_te < 0] = 0
new_soon_te
```

001

데이터 전처리

3. 2022-09-01 새순값에 대한 평균 기울기

새순은 날짜가 지날수록 점점 감소하는 추세를 보이는 경향을 가지고 있는 것을 확인하여

2022-09-01 새순값과 2022-11-28 새순값의 평균 기울기를 feature로 사용.

```
train['newsoon_ju'] = 0
train['newsoon_ju'][train['2022-09-01 새순'] < 3] = 0.0213
train['newsoon_ju'][(train['2022-09-01 새순'] >= 3) & (train['2022-09-01 새순'] < 3.5)] = 0.0275
train['newsoon_ju'][(train['2022-09-01 새순'] >= 3.5) & (train['2022-09-01 새순'] < 4)] = 0.0335
train['newsoon_ju'][(train['2022-09-01 새순'] >= 4) & (train['2022-09-01 새순'] < 4.5)] = 0.0395
train['newsoon_ju'][train['2022-09-01 새순'] >= 4.5] = 0.0461


test_x['newsoon_ju'] = 0
test_x['newsoon_ju'][test_x['2022-09-01 새순'] < 3] = 0.0213
test_x['newsoon_ju'][(test_x['2022-09-01 새순'] >= 3) & (test_x['2022-09-01 새순'] < 3.5)] = 0.0275
test_x['newsoon_ju'][(test_x['2022-09-01 새순'] >= 3.5) & (test_x['2022-09-01 새순'] < 4)] = 0.0335
test_x['newsoon_ju'][(test_x['2022-09-01 새순'] >= 4) & (test_x['2022-09-01 새순'] < 4.5)] = 0.0395
test_x['newsoon_ju'][test_x['2022-09-01 새순'] >= 4.5] = 0.0461
test_x['newsoon_ju']
```

002

모델

AutoGluon

다양한 알고리즘으로, customized parameter 범위내에서 최적의 알고리즘을 select 해주는 기법

```
train_data = TabularDataset(train)
test_data = TabularDataset(test_x)



predictor = TabularPredictor(label='착과량(int)', eval_metric='mean_absolute_error').fit(train_data, presets='high_quality', ag_args_fit={'num_gpus': 0})
y_pred = predictor.predict(test_data)
y_pred = pd.DataFrame(y_pred, columns=['착과량(int)'])
```

003

결과

V1

엽록소 관련 feature 제거
새순 측정 feature 보완
2022-09-01 새순값에 대한 평균 기울기

Autogluon

MAE : 30.0643

V2

엽록소 관련 feature 제거
Low Pass Filter

Autogluon

MAE : 29.3825

003

시도한 모델 및 feature

1. ML Model

Random Forest, Extra Tree, XGBoost, LightGBM, Gradient Boosting, CatBoost 등의 Machine Learning 모델들을 사용하였으나 Validation 성능이 좋지 않아 배제

2. Feature

2022-09-01 새순값에 대해 9월, 10월, 11월의 평균, 최대, 최소, 중앙값

수관폭1(min)과 수관폭2(max)에서 최댓값과 최솟값이 반대로 되어 있는 경우 처리

수관폭2(max)와 수관폭1(min)의 차이

2022-09-01 새순값에 대한 착과량을 비교하여 구간화 라벨링

2022-09-01 새순값에 따른 착과량(int)의 평균, 최소, 최대값

feature importance에 따른 drop

THANK YOU