Lab_assignment 4

2017010146 산업경영공학과 이시현

1. 과제3에서 유추할 수 있었던 사항들이 기계학습 결과와 어떻게 연계되는지 제시

1) 과제 3에서 유추했던 사항

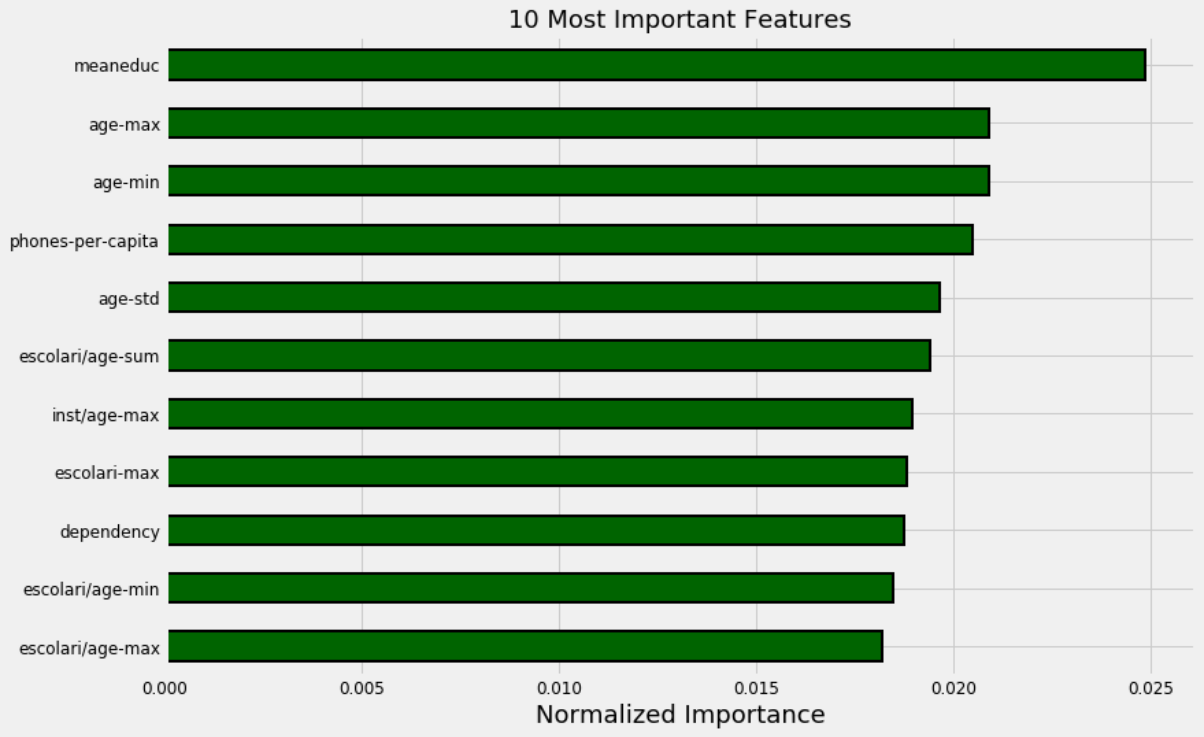
① meanedu(average years of education for adults (18+)) 변수는 빈곤도를 예측하는데 유용할 것이다.

② overcrowding(# persons per room) 변수는 빈곤도를 예측하는데 유용할 것이다.

③ 새로 생성한 walls+roof+floor 변수는 빈곤도를 예측하는데 유용할 것이다.

2) 기계학습 결과와의 연계

RandomForest를 이용하여 변수 중요도 상위 10개의 변수를 파악하였다.



① meanedu 변수는 가장 높은 중요도를 차지하였다.

위의 그림에서 살펴볼 수 있는 것과 같이, 18세 이상의 평균 교육 년 수를 의미하는 meanedu 변수가 가장 높은 중요도를 차지하였다. 또한, 학교를 다닌 년 수를 의미하는 escolari 변수를 age-max, age-min, age-sum으로 나눈 값과 escolari-max가 상위 10개의 중요 변수로 rank 되었다. 교육수준을 나타내는 inst 변수를 age-max로 나눈 값 또한 상위 10개의 중요 변수로 rank 되었다. 교육한 년 수가 빈곤도에 큰 영향을 끼친다는 것을 알 수 있다.

② overcrowding 변수는 상위 10개의 변수에는 못 들었다.

하지만 상위 10개 중 가장 작은 중요도를 가진 escolari/age-max 변수와 0.005 정도 차이나는 0.01368의 변수 중요도를 가졌으므로 기계학습 결과에 어느정도는 중요하다고 볼 수 있다.

```
feature_importances[feature_importances['feature']=='overcrowding']
```

	feature	importance
71	overcrowding	0.01368

③ 새로 생성한 walls+roof+floor 변수는 상위 10개의 변수에는 못 들었다.

하지만 overcrowding 변수보다 높은 변수 중요도를 갖았으며, escolari/age-max 변수와 변수 중요도가 0.003 정도의 차이가 나므로 기계학습 결과에 어느정도 중요하다고 볼 수 있다.

```
feature_importances[feature_importances['feature']=='walls+roof+floor']
```

	feature	importance
95	walls+roof+floor	0.015055

2. 검증 및 평가 방법을 적용하여 알 수 있는 사실 제시

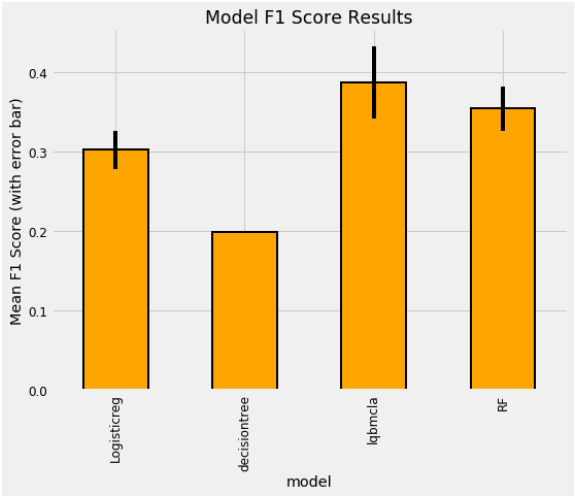
데이터 셋은 각 label 별 불균형한 데이터를 갖기 때문에 검증 및 평가 과정의 주요 성능 척도로 recall과 precision 모두 고려한 f1-score를 사용하였다.

1) 모델 검증 과정에서 알 수 있는 사실

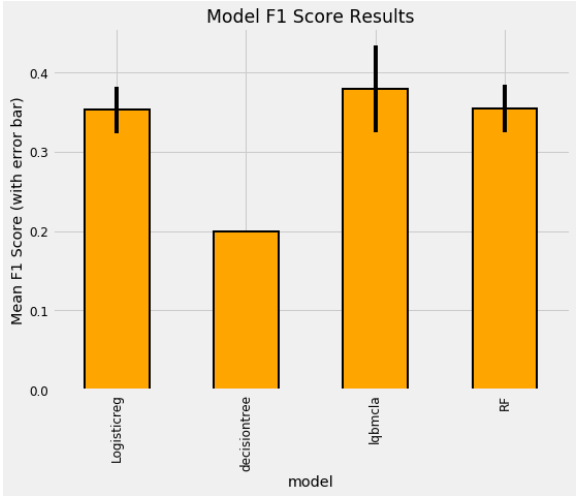
① 정규화 하기 전보다 정규화 한 후의 로지스틱 회귀 모델 성능이 향상되었다.

훈련 데이터를 정규화 하기 전과 정규화 한 후 성능을 10번 fold 한 교차 검증을 통해 비교해보았다.

<정규화 하기 전 성능>



<정규화 한 후 성능>



위의 그래프에서 볼 수 있듯이 다른 모델의 성능은 증가하는 모습이 보이지 않았지만, 로지스틱 회귀 모델에서 성능이 눈에 띄게 증가한 것을 볼 수 있다.

결정 트리나 랜덤 포레스트는 데이터 스케일에 영향을 받지 않기 때문에 정규화가 필요 없다. 하지만 로지스틱 회귀 같은 경우, 정규화 하지 않으면 변수 별 중요도가 스케일에 따라 달라질 수 있어, 중요한 변수라도 스케일이 작으면 모델에 잘 반영이 되지 않아 예측력을 떨어지게 한다.

이와 같은 이유로 정규화 한 후의 로지스틱 회귀의 성능이 향상된 것으로 보인다. 전반적으로 정규화 한 후에 성능이 올라갔거나 비슷하거나 하는 양상을 보였기 때문에 평가 모델에서는 정규화 한 모델을 사용했다.

② 정규화 하기 전, 한 후 모두 LGBM Classifier 모델의 성능이 가장 높다.

2) 모델 평가 과정에서 알 수 있는 사실

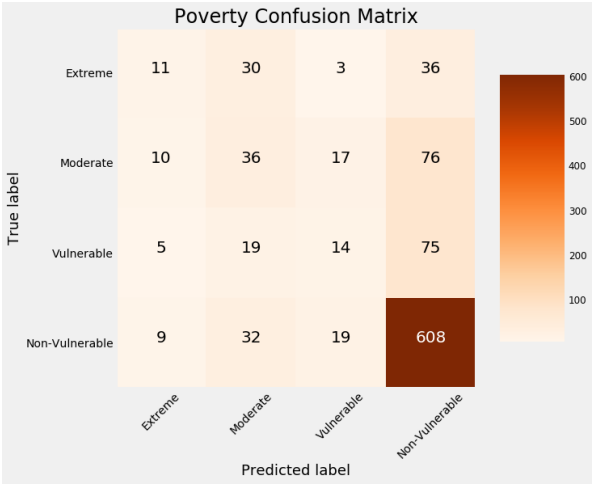
③ f1-score 기준으로 가장 높은 성능을 보인 것은 LGBM Classifier이다.

모델	모델 파라미터	F1-score
로지스틱 회귀	(random_state=13, solver='liblinear',C=10.0)	0.3681
결정 트리	(max_depth=2, random_state=13)	0.20024
랜덤 포레스트	(random_state=13, n_jobs=-1, n_estimators=100)	0.34688
LGBM	(n_estimators=1000, num_leaves=64, n_jobs=-1,boost_from_average=False)	0.38419

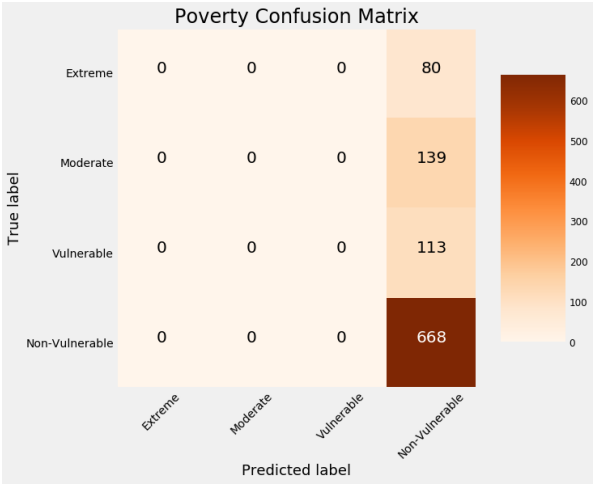
④ 결정 트리 예측 모델은 non-vulnerable 이외의 다른 label은 예측하지 못한다.

각 모델에 의해 예측된 값을 혼동행렬로 표현하면 아래와 같다.

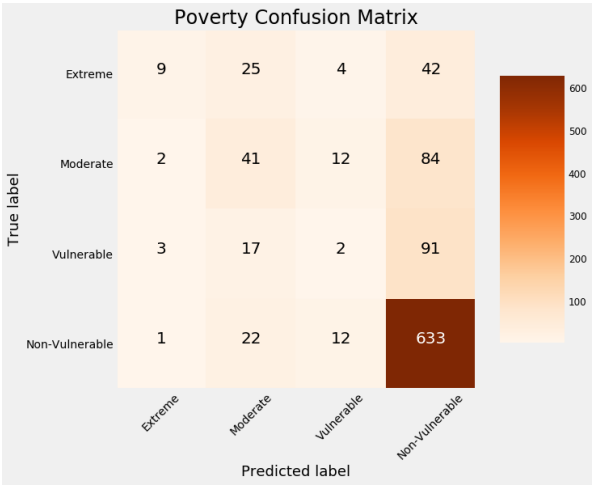
<로지스틱 회귀모델의 혼동행렬>



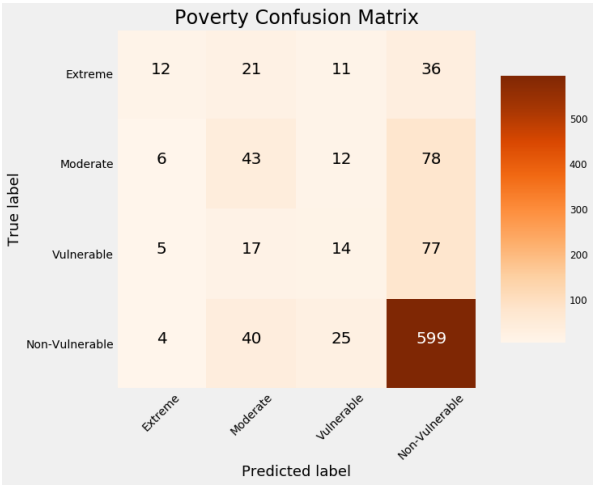
<결정 트리의 혼동행렬>



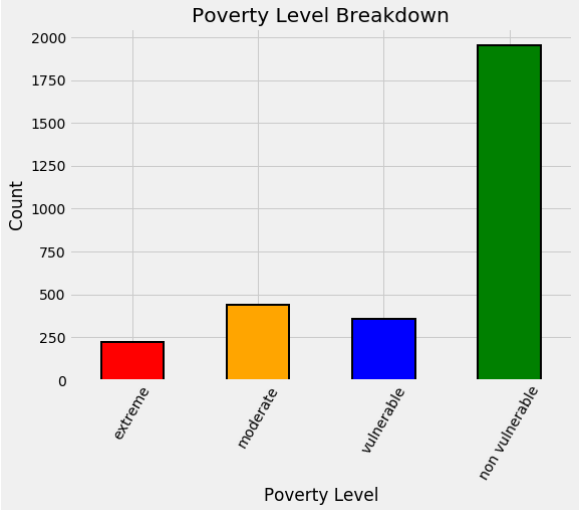
<랜덤 포레스트의 혼동행렬>



<LGBM Classifier의 혼동행렬>



<결정 트리의 혼동행렬>에서 볼 수 있듯이 모든 label을 non-vulnerable로 예측하였다. 클래스 불균형으로 인해 label이 non-vulnerable인 데이터가 전체의 65%를 차지하기 때문에 상대적으로 데이터 수가 적은 다른 label에 대한 학습이 안된 것으로 생각된다. (Label별 분포는 다음과 같다.)



이에 비해 결정 트리가 모여 만들어진 랜덤 포레스트는 모든 label에서 실제와 같은 label로 예측한 데이터들이 있으며, 결정 트리와는 0.146 정도의 성능 차이를 보인다. 랜덤 포레스트는 결정 트리가 훈련 데이터에 오버피팅 되는 단점을 해결한 모델이기 때문에 성능이 더 높게 나온 것 같다.

3. 다양한 결과 중 가장 나은 결과를 선택하고 그 이유를 제시할 것

네 가지 모델 중 가장 나은 결과는 LGBM Classifier로 만든 예측 모델이다.

모델 검증 과정에서도 LGBM Classifier는 모델의 성능이 가장 높았으며, 테스트 데이터로 모델 평가를 진행했을 때에도 f1-score값이 0.38419로 두번째로 좋은 성능을 보였던 로지스틱 회귀와 0.02 정도의 차이로 LGBM Classifier모델의 성능이 가장 높게 나왔다.

LGBM Classifier모델은 로지스틱 회귀보다 모든 label를 골고루 학습한 모델이다. label이 non-vulnerable인 데이터에 대한 예측력은 로지스틱 회귀보다 떨어졌지만, 다른 2개의 label에서는 실제와 같은 label로 예측한 데이터들이 로지스틱 회귀보다 8개 더 많았다.

모든 label 각각에 대한 설명력이 가장 높다고 판단하여 LGBM Classifier로 만든 예측 모델을 가장 나은 결과로 선택했다.