Multivariate Data Analysis Assignment #6

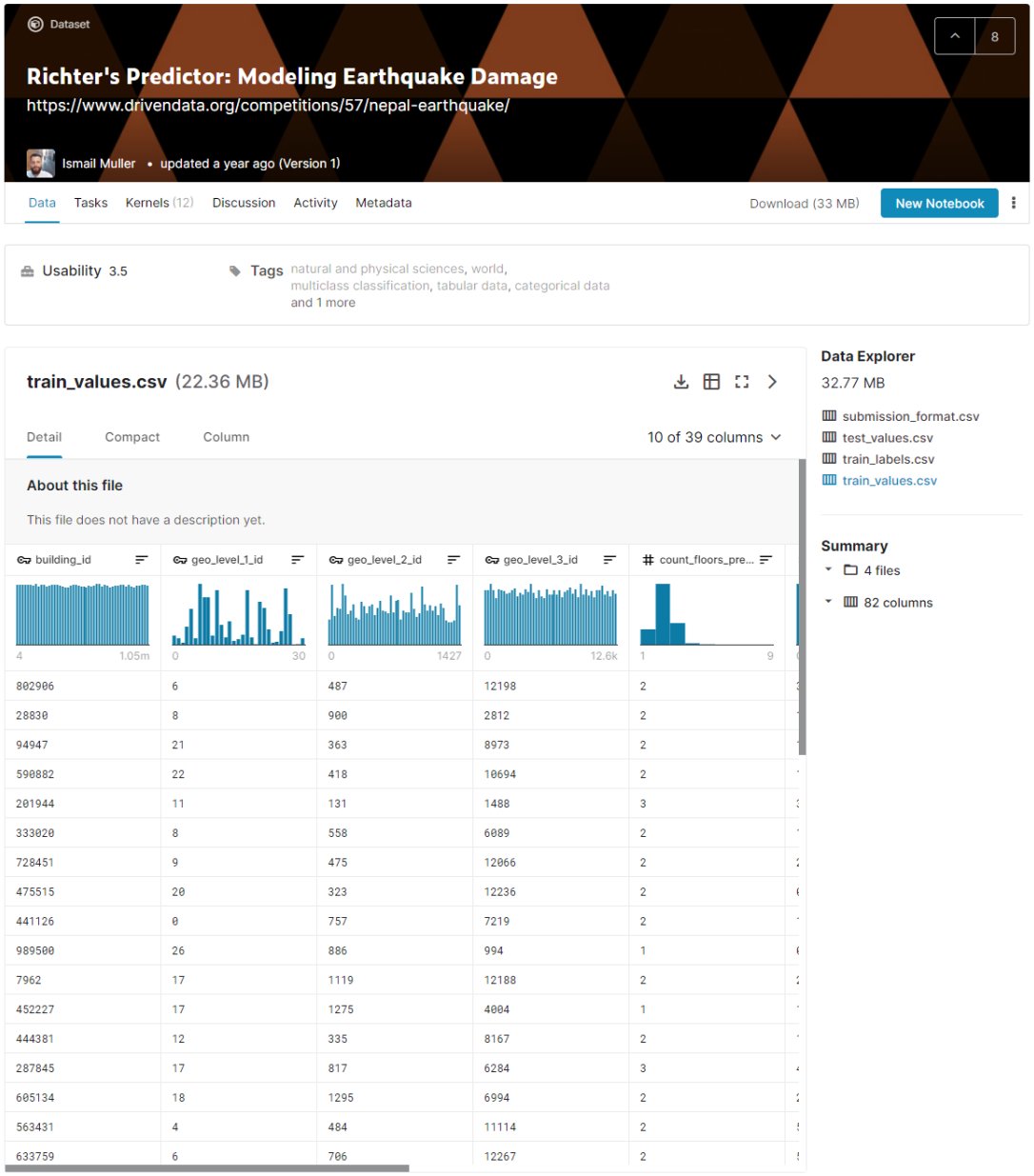
Ensemble Learning

Dataset: Richter's Predictor: Modeling Earthquake Damage

해당 데이터셋은 총 260,601개의 건물에 대한 지진 피해 정도를 기록한 데이터셋이다. 첫 번째 컬럼인 building_id는 각 건물의 일련번호이며, 마지막 컬럼인 damage_grade는 해당 건물의 지진 피해 정도로써 분류모형의 종속변수로 사용된다. 이 외 모든 컬럼은 입력변수이다.

(https://www.kaggle.com/mullerismail/richters-predictor-modeling-earthquake-damage?select=train_values.csv)

(주의) Kaggle에서 제공하는 데이터가 아닌 과제에서 제공된 Earthquake_Damage.csv 파일을 사용하시오.



전체 데이터 셋에 대해 다음 물음에 답하시오.

[사전 작업 사항]

[1] 입력 변수의 속성이 numeric 이 아닌 변수들에 대해 1-of-C coding (1-hot encoding) 방식을 통해 명목형(요인형) 변수를 범주의 개수만큼의 이진형(binary) 변수들로 구성되는 dummy variable 을 생성하시오.

[2] 전체 데이터셋을 임의로 150,000 개의 빌딩이 포함된 Training dataset 과 50,000 개의 Validation dataset, 그리고 60,601 개의 Test dataset 으로 구분한 뒤 다음 각 물음에 답하시오. 분류 성능을 평가/비교할 때는 3-class classification 의 Accuracy 와 Balanced Correction Rate (BCR)을 이용하시오.

[Q1] 다음과 같이 세 가지 단일 모형에 대하여 분류 모델을 구축하고 Accuracy 와 BCR 관점에서 분류 정확도를 비교해보시오. CART 와 ANN 의 경우 hyperparameter 후보 값들을 명시하고 Validation dataset 을 통해서 최적의 값을 찾아서 Test 에 사용하시오.

- Multinomial logistic regression
- Classification and Regression Tree (CART)
- Artificial Neural Network (ANN)

아래 질문들에 대해서는 Base Learner가 CART와 ANN인 경우 [Q1]에서 선택된 hyperparameter를 사용하여 실행하고 그 결과를 이용하여 답하시오.

[Q2] CART의 Bagging 모델을 Bootstrap의 수를 30부터 30단위로 300까지 증가시키면서 분류 정확도를 평가해보시오. 최적의 Bootstrap 수는 몇으로 확인되는가? 이 모델은 단일 모형과 비교했을 때 성능의 향상이 있는가?

[Q3] Random Forest 모델의 Tree의 수를 30부터 30단위로 300까지 증가시키면서 분류 정확도를 평가하고 다음 물음에 답하시오. 학습 과정에서는 변수의 중요도가 산출되도록 학습하시오.

[Q3-1] 최적의 Bootstrap 수는 몇으로 확인되는가?

[Q3-2] 최적의 Tree 수를 기준으로 이 데이터셋에 대해서는 CART Bagging과 Random Forest 중에서 더 높은 분류 정확도를 나타내는 모형은 무엇인가?

[Q3-3] 각 Tree의 수(Bootstrap의 수)마다 CART Bagging 모형과의 분류 정확도를 비교할 수 있는 그래프를 도시하시오. Tree의 수는 CART Bagging과 Random Forest는 성능 차이에 영향을 미친다고 볼 수 있는가?

[Q4] [Q1]에서 찾은 최적의 hyperparameter를 이용하여 ANN 단일모형을 30번 반복하여 테스트 정확도를 평가해보시오. Accuracy와 BCR의 평균 및 표준편차를 기록하시오.

[Q5] ANN Bagging 모델에 대해 다음 물음에 답하시오.

[Q5-1] Bootstrap의 수를 30부터 30단위로 300까지 증가시키면서 각 Boostrap 수마다 30회 반복 수행을 실시하여 Accuracy와 BCR의 평균 및 표준편차를 각각 기록하시오.

[Q5-2] 최적의 Bootstrap 수는 몇으로 확인되는가?

[Q5-3] 이 모델은 단일 모형의 30회 반복 결과와 비교했을 때 분류 정확도 및 성능의 편차 측면에서 어떤 차이가 있는가?

[Q6] Adaptive Boosting(AdaBoost)에 대해 다음 물음에 답하시오.

[Q6-1] Hyperparameter 후보 값들을 명시하고, Validation dataset을 통해 최적의 hyperparameter 값을 찾아보시오.

[Q6-1] 최적의 hyperparameter 값을 이용하여 AdaBoost 모델을 학습한 뒤, Test dataset에 적용하여 먼저 구축된 모델들과 분류 성능을 비교해보시오.

[Q7] Gradient Boosting Machine(GBM)에 대해 다음 물음에 답하시오.

[Q7-1] Hyperparameter 후보 값들을 명시하고, Validation dataset을 통해 최적의 hyperparameter 값을 찾아보시오.

[Q7-2] 최적의 hyperparameter 값을 이용하여 GBM 모델을 학습(변수의 중요도가 산출되도록 학습)한 뒤, Test dataset에 적용하여 먼저 구축된 모델들과 분류 성능을 비교해보시오.

[Q7-3] 산출된 변수의 중요도를 해석해보고, Random Forest 모델에서 산출된 주요 변수와 비교해보시오.

[Q8] 총 여덟 가지의 모델(Multinomial logistic regression, CART, ANN, CART Bagging, ANN Bagging, Random Forest, AdaBoost, GBM) 중 BCR 관점에서 가장 우수한 분류 정확도를 나타내는 모형은 무엇인가?

[Extra Question]

이 데이터셋은 아래 표와 같이 Class 2 > Class 3 > Class 1 순으로 높은 비중을 차지하고 있으며, 범주의 불균형이 상당한 수준이다. [Q8]에서 선정된 가장 우수한 모델(알고리즘 및 hyperparameter)에 대해서 데이터 전처리 관점에서 불균형을 해소하여 분류 성능을 향상시킬 수 있는 아이디어를 제시하고 실험을 통해 검증해보시오.

	Class 1	Class 2	Class 3
N. of instances	25,124	148,259	87,218
%	9.64%	56.89%	33.49%