Data Analyst
김홍범입니다

PORTFOLIO

CONTACT
billkim418@naver.com
010 7763 4231
https://github.com/billkim418

데이터 분석가 김홍범입니다

김홍범 / Hongbum Kim

1994.04.18 / 경기도 구리시

Tel. 010-7763-4231
Email. billkim418@naver.com
경기도 구리시 인창동

GRADUATION

2021 한국외국어대학교 산업경영공학과 졸업
2013 토평고등학교 졸업

SKILL

Python	95
R	95
SQL	80

INTERRESTS

Data Analysis
Explainable-AI
AI Safety

PROJECT

2020 에어비앤비 ML 비교
2020 Predict Raod Accidents
2021 코로나 시각화 경진대회
2021 날씨 빅데이터 콘테스트(본선 6위)
2021 삼성카드 데이터 분석
2021 한국어 음성인식 EMR 구축

ABOUT

제가 성장할수 있었던 2가지 프로젝트에 대한 포트폴리오입니다.
1. 한글로 진행한 날씨 빅데이터 콘테스트
2. 영어로 진행한 처인구 교통사고예측

PROJECT.1

날씨 빅데이터 콘테스트

01

2021년 4월 26일 ~ 7월 2일까지 진행되는 기상청 주최 날씨 빅데이터 콘테스트 공공협력 분야에 참가 하였습니다.
날씨, 임상도, 토양도 데이터를 기반으로 한 경상도 지역의 산사태 예측을 목표로 합니다.

ABOUT PROJECT

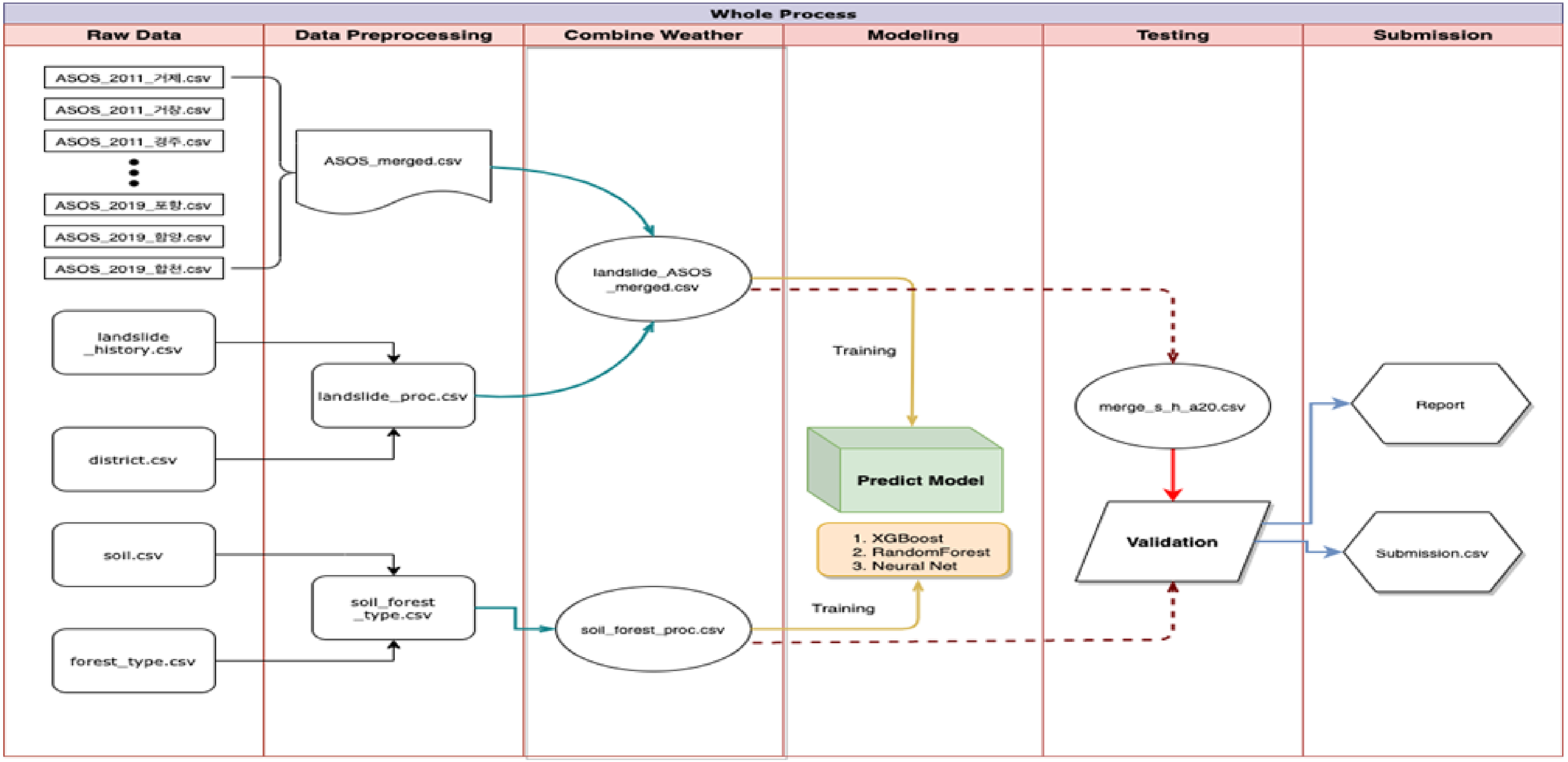
https://github.com/wonkwonlee/a2w-kma-big-data-contest

https://www.notion.so/b3f8cce82d074335bd563d99a1ffdf13?v=419b0115cd5a488e814552b9749039b4

날씨 빅데이터 콘테스트

전체 프로세스

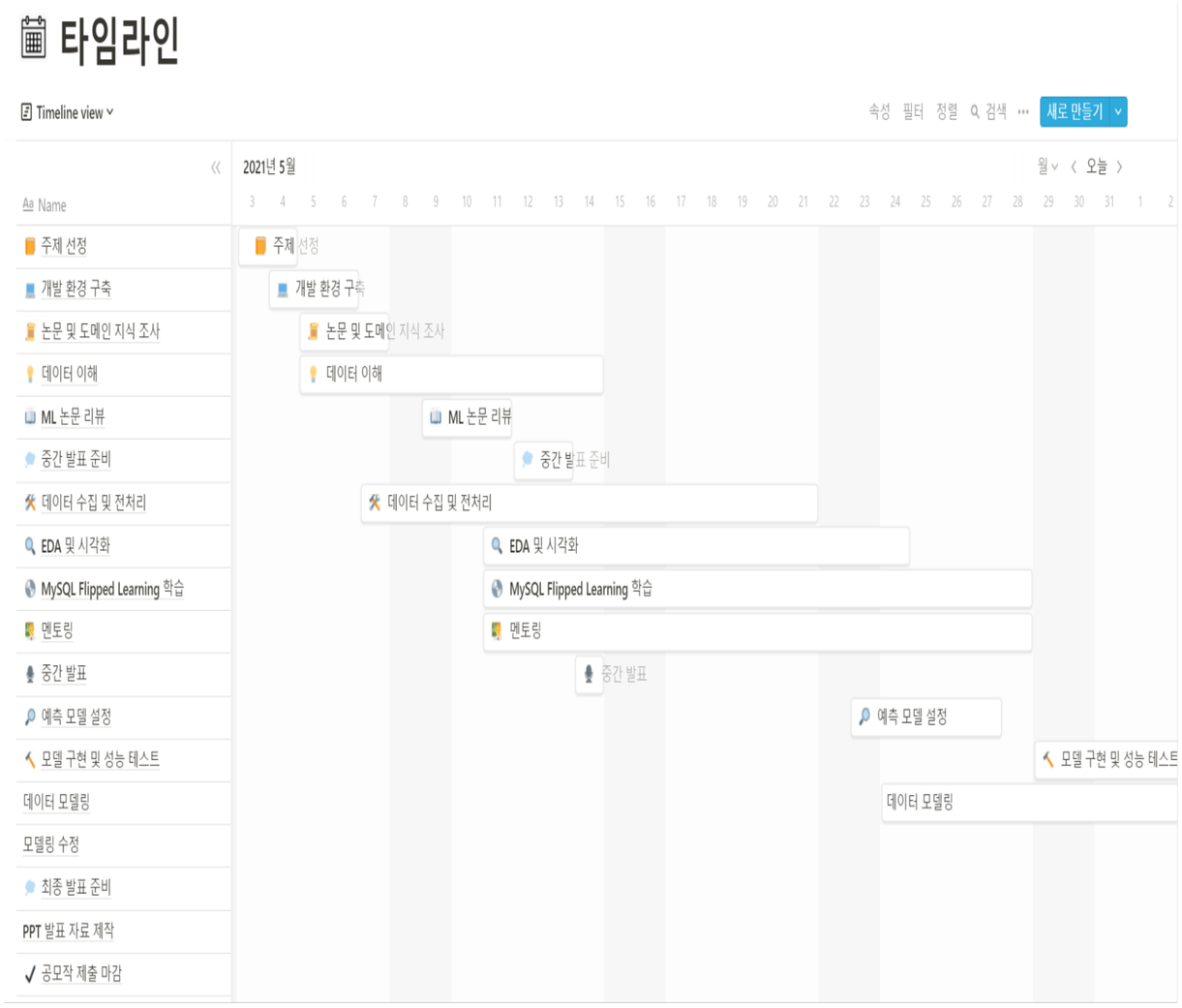
전체 프로세스



날씨 빅데이터 콘테스트

타임 라인 , 작업 일지- 노션

타임라인



작업 일지

팀 작업 일지

- 1주차 2021.05.03 ~ 2021.05.09
- 2주차 2021.05.10 ~ 2021.05.16
- 3주차 2021.05.17 ~ 2021.05.23
- 4주차 2021.05.24 ~ 2021.05.25

개인 별 작업 일지

곽희원

- 1주차 2021.05.03 ~ 2021.05.09
- 2주차 2021.05.10 ~ 2021.05.16
- 3주차 2021.05.17 ~ 2021.05.23
- 4주차 2021.05.24 ~ 2021.05.25

김홍범

- 1주차 2021.05.03 ~ 2021.05.09
- 2주차 2021.05.10 ~ 2021.05.16
- 3주차 2021.05.17 ~ 2021.05.23
- 4주차 2021.05.24 ~ 2021.05.25

이원권

- 1주차 2021.05.03 ~ 2021.05.09
- 2주차 2021.05.10 ~ 2021.05.16
- 3주차 2021.05.17 ~ 2021.05.23
- 4주차 2021.05.24 ~ 2021.05.25

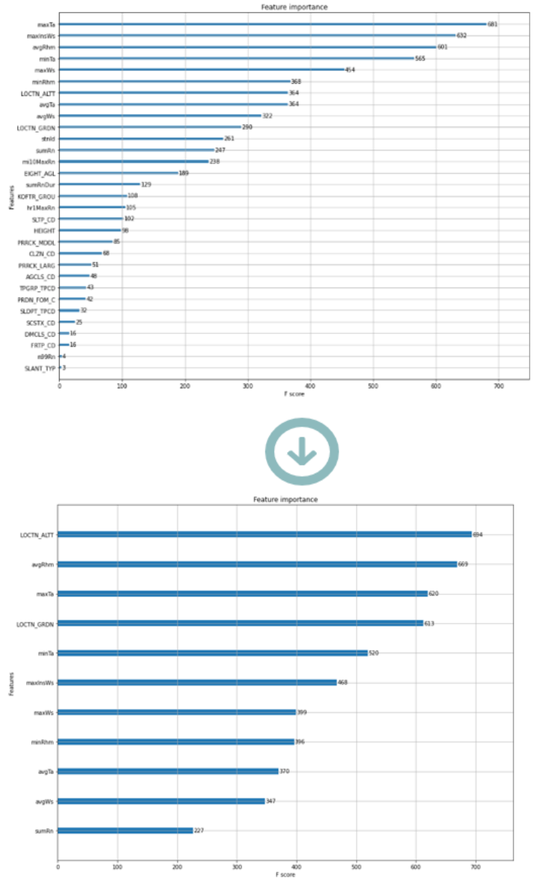
최디도

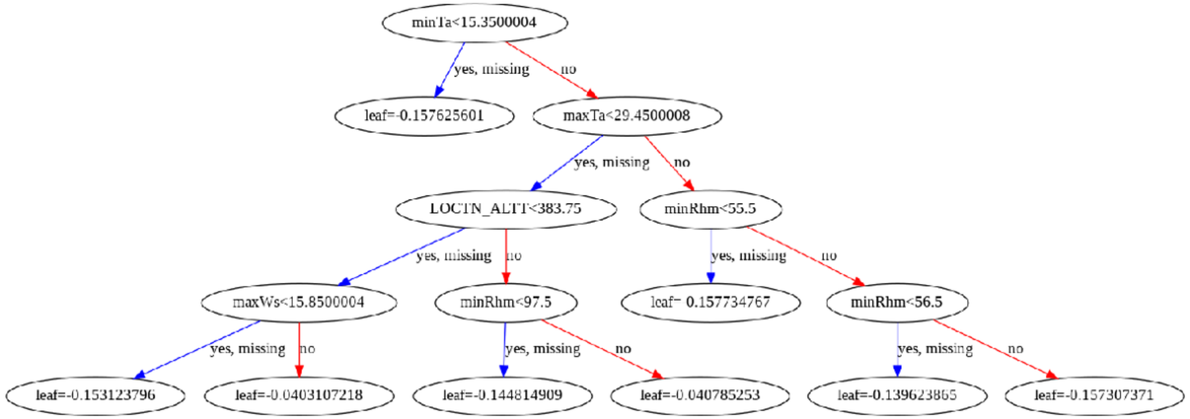
- 1주차 2021.05.03 ~ 2021.05.09
- 2주차 2021.05.10 ~ 2021.05.16
- 3주차 2021.05.17 ~ 2021.05.23
- 4주차 2021.05.24 ~ 2021.05.25

날씨 빅데이터 콘테스트
선정 모델

분석 기법 및 결과

03. XGBoost(eXtreme Gradient Boosting)





- Gradient Boosting 계열 알고리즘으로 학습 속도가 느리고 과적합 이슈를 보완
- Regression, Classification 문제를 모두 지원하며 성능과 자원 효율이 좋음
- GridSearchCV 를 이용하여 복수의 하이퍼 파라미터 최적화 (Train set : Test set = 75 : 25)
- XGBoost의 Feature Importance를 통해 전체 변수 중 중요 변수만 이용하여 새롭게 학습

날씨 빅데이터 콘테스트
결과

Under Sampling

분석 기법 및 결과
04. 결과-Under Sampling

모델	Under Sampling	Accuracy	CSI
DNN	1:30	70.26%	2.24%
RandomForest	1:10	95.1%	0%
RandomForest	1:20	94.78%	1.57%
XGBoost	1:5	87.68%	4.03%
XGBoost	1:10	93.27%	4.64%
XGBoost	1:15	90.81%	4.3%
XGBoost	1:20	93.07%	3.24%

Under Samping 모델의 평균 ACC는 93.27%, CSI는 4.64%

Over Sampling

분석 기법 및 결과
04. 결과-Over Sampling

모델	Over Sampling	Accuracy	CSI
Imb_XGBoost_weight	SMOTE	92.13%	4.95%
Imb_XGBoost_focal	SMOTE	92.42%	5.01%

최종 모델 : Imb_XGBoost(OverSampling)
최종 모델의 평균 ACC는 92.42%, CSI는 5.01%

날씨 빅데이터 콘테스트
피드백 및 최종데이터

피드백

제가 이번 대회를 하면서 가장 크게 느낀점은 전처리 및 EDA의 중요성입니다.

1. 데이터의 종류가 달라 병합시 어려움이 있었다.

-> Reference 및 정보 검색을 통해 병합 방법을 찾았 내었다. (Geopandas 이용)

2. 기존의 1일 후 예측이 아닌 2일후 예측이라 어려움이 있었다.

-> 날씨 데이터를 활용하여 해결하였지만 독립시행 혹은 종속시행 여부 검증이 안되었다.

3. 대회측에서 제공한 데이터인 행정동 변수와 그 외 변수가 담긴 데이터간의 지도정보(gis)가 다르게 주어졌다. 따라서 매칭이 어려웠다.

-> 실제 지도와 매칭해가며 인접지역으로 선정하였다.

4. 모델의 True Negative값이 너무 높게 나와 버렸다.

-> imbalance data를 보완하기 위해 실시한 Samling 기법들의 단점이었고, 이를 다른 전처리 혹은 모델로 보완가능할거 같다.

최종 데이터

활용 데이터

07. 최종데이터 Merge_final_real.csv

날짜	행정동	온도	···	강수량	임종	···	임상	토성	···	경사도	···	산사태
2011-01-01	울릉군 울릉읍	0.8	···	0	2	···	2	0	···	0	···	0
2016-10-05	울릉군 울릉읍	16.9	···	104.5	2	···	2	0	···	0	···	1
···	···	···	···	···	···	···	···	···	···	···	···	···
2019-10-03	영양군 일월면	16.9	···	21.9	2	···	2	4	···	20	···	1

- 행정동 변수 : 2개
- ASOS 날씨 데이터 : 14개
- 임상도 데이터 : 12개
- 토양도 데이터 : 7개

병합 후 약 10개의 ROW, 35개의 Feature

PROJECT.2

02

PREDICT ROAD ACCIDENTS

The purpose of the analysis is to create a model that can predict the casualty severity through the analysis and exa mine key indicators. The data can be used to determine t he ideal value of compensation according to the Injury Se verity of the customer, and to prepare for emergency trea tment. Also if we know the type of accident which has hi gh severity, we can prevent additional accidents.

ABOUT PROJECT

Written in english

https://github.com/wonkwonlee/a2w-km a-big-data-contest

PREDICT ROAD ACCIDENTS

Preprocessing

Preprocess for Machine Learning

PREDICT ROAD ACCIDENTS

Analysis

Apply Machine Learning model

Analysis

#Orange #R

01 Clustering

To identify similarities between the s everity of the accident and other fea tures

02 Association rule

To identify the rules of the condition s that occurred at the same time

03 Classification

To identify patterns with the severity of the accident and other features

04 Regression

To identify relations between the se verity of the accident and other feat ures

PREDICT ROAD ACCIDENTS

Limits

1. Because data is based on the target variable, classification is imbalanced. To supplement that, we can solve the data using python code. The most common technique to oversample a dataset is known as SMOTE. SMOTE-NC can handle a mix of categorical and continuous features, but it needs at least one continuous feature.

2. About clustering, there are a lot of clusters. so it is hard to understand
Since there are nine variables in the target variable, the number of clusters was also set at nine and clustering was carried out. As a result, interpretation of the results was not easy, and there will be ways to supplement them, such as using domain knowledge.

3. About association, even if the rules contain some high frequency condition in RHS. they can not associate with casu class. because it can be just coincidence. The interpretation was often meaningless. For example, Sunny accounts for most of the weather. In other words, the rules often show that there is a Sunny condition in the RHS, which is a meaningless association. To supplement that, The rhs of the rule excluded the condition car & car, Sunny, and passenger car, which account for the majority of each condition. Also when the casu class in lhs, the max sup is 18.4%, so after constrainting(ex, min conf) the supports became smaller. the sup means the percent of total, so if the sup is too small. it can also considered as coincidence

감사합니다!
잘 부탁드립니다!

~ 2021 PORTFOLIO