

8일차. 캐글 데이터 분석 실습

Bike Sharing Demand 예측

심선영 교수

강의 목표

- ❖ 캐글 -Bike Sharing Demand Competition 실습을 통하여 데이터 분석 역량을 높인다.
- ❖ 데이터 전처리, 분석 모델 변경에 따라 예측 성능이 향상되는 과정을 실습해 본다.
- ❖ 캐글 Competition에 직접 submission하여 본인의 순위를 확인해 본다.

강의 목차

❖ 분석 실습

- New Notebook 생성
- 디렉토리 구조 확인
- 실습

❖ Submission

강의 스케줄

시간	목차	활동
1h	강의 Overview	PPT학습
3h	Bike Demand Sharing 분석 실습	파이썬 실습
3h	Bike Demand Sharing 분석 실습	파이썬 실습
1h	Wrap-Up	학습 정리

분석 실습

분석 실습

New Notebook 생성

❖ Code > New Notebook

The screenshot shows the Kaggle interface for the 'Bike Sharing Demand' competition. The header includes the competition title, a subtitle 'Forecast use of a city bikeshare system', and the Kaggle logo with statistics: '3,242 teams · 7 years ago'. Below the header is a navigation bar with tabs: 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', 'Rules', and 'Team'. The 'Code' tab is selected and highlighted with a red dashed box. To the right of the tabs is a 'New Notebook' button, also highlighted with a red dashed box. Below the navigation bar is a search bar labeled 'Search notebooks' and a 'Filters' button. Under the search bar are tabs for 'All', 'Your Work', 'Shared With You', and 'Bookmarks'. The 'All' tab is selected. Below these tabs is a list of notebooks. Each notebook entry includes a profile picture, the notebook title, the time it was updated, the number of comments, and the competition name. The first notebook is 'Bike Sharing Demand only data analsis and EDA', updated 19h ago, with 3 comments. The second is '2022 SMARCLE Kaggle Study_Bike Demand_KIY', updated 14h ago, with 1 comment. The third is 'CWS_bikesharing', updated 12h ago, with 0 comments. The fourth is 'First_bike_notebook', updated 3d ago, with 0 comments. Each entry also shows a medal icon (Bronze for the first) and a 'Hotness' dropdown menu.

Bike Sharing Demand
Forecast use of a city bikeshare system

Kaggle · 3,242 teams · 7 years ago

Overview Data **Code** Discussion Leaderboard Rules Team

New Notebook ...

Search notebooks Filters

All Your Work Shared With You Bookmarks Hotness ▾

- Bike Sharing Demand only data analsis and EDA**
Updated 19h ago
3 comments · Bike Sharing Demand
Bronze ...
- 2022 SMARCLE Kaggle Study_Bike Demand_KIY**
Updated 14h ago
1 comment · Bike Sharing Demand
...
- CWS_bikesharing**
Updated 12h ago
0 comments · Bike Sharing Demand
...
- First_bike_notebook**
Updated 3d ago
0 comments · Bike Sharing Demand
...

분석 실습

New Notebook 생성

❖ 파일명을 “BikeSharing1”으로 수정

❖ 첫 코드 → 데이터 파일 위치 확인

- /kaggle/input/bike-sharing-demand/sampleSubmission.csv
- /kaggle/input/bike-sharing-demand/train.csv
- /kaggle/input/bike-sharing-demand/test.csv

❖ 두 번째 코드 → 노트북 작업 위치 확인

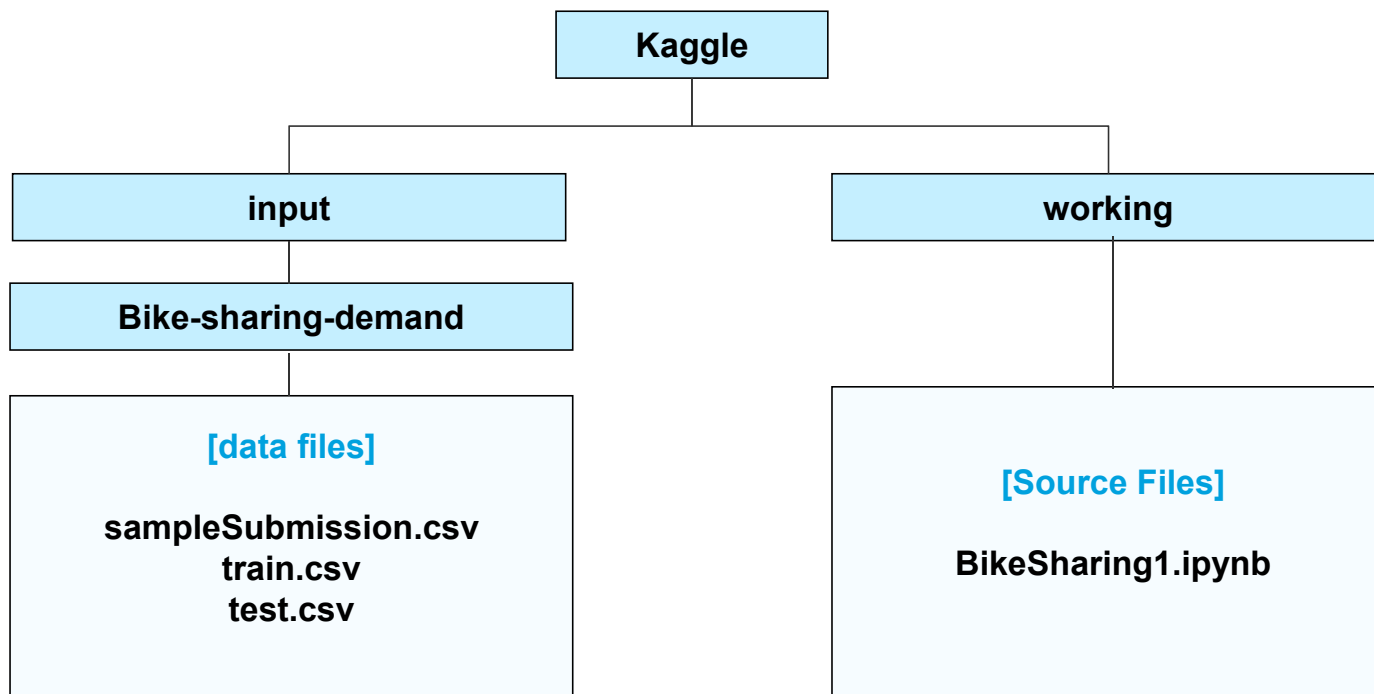
- /kaggle/working

The screenshot shows a Kaggle Notebook interface. The notebook title is "BikeSharing1" (highlighted with a red dashed box). The first code cell contains comments and Python code to list files in the /kaggle/input directory. The output of this cell is a list of three files: /kaggle/input/bike-sharing-demand/sampleSubmission.csv, /kaggle/input/bike-sharing-demand/train.csv, and /kaggle/input/bike-sharing-demand/test.csv (highlighted with a red dashed box). The second code cell contains the command !pwd. The output of this cell is /kaggle/working (highlighted with a red dashed box).

```
[1]:  
# This Python 3 environment comes with many helpful analytics libraries installed  
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python  
# For example, here's several helpful packages to load  
  
import numpy as np # linear algebra  
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)  
  
# Input data files are available in the read-only "../input/" directory  
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory  
  
import os  
for dirname, _, filenames in os.walk('/kaggle/input'):  
    for filename in filenames:  
        print(os.path.join(dirname, filename))  
  
# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a new notebook  
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session  
  
/kaggle/input/bike-sharing-demand/sampleSubmission.csv  
/kaggle/input/bike-sharing-demand/train.csv  
/kaggle/input/bike-sharing-demand/test.csv  
  
!pwd  
  
/kaggle/working
```

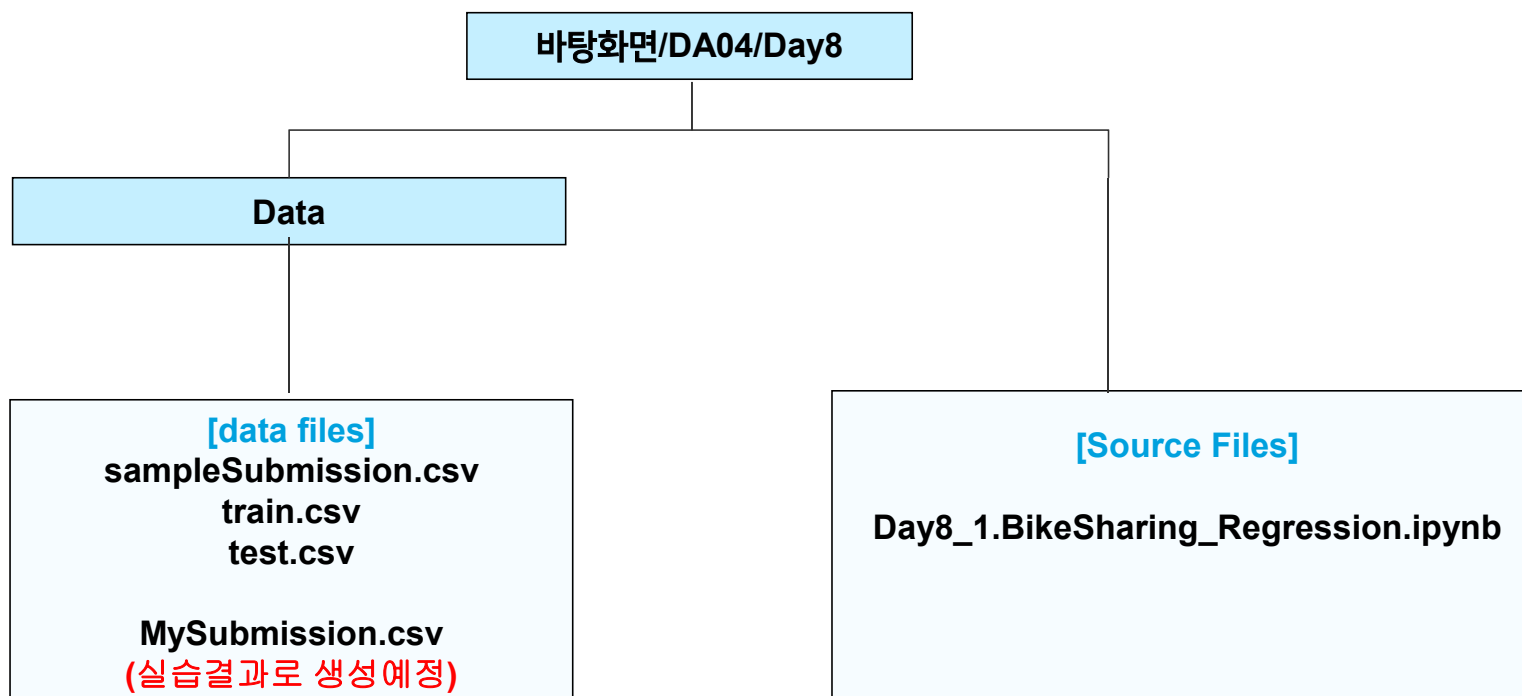
분석 실습

Kaggle 노트북 작업 시 디렉토리 구조



분석 실습

쥬피터 노트북 작업 시 디렉토리 구조



분석 실습

실습 시작

- ❖ 다음의 실습 파일들을 이용하여 실습을 진행함
- ❖ Day7_3.BikeSharing_EDA.ipynb – 처음~ EDA까지
- ❖ Day8_1.BikeSharing_Regression.ipynb – 분석모델 다양화를 통한 성능 고도화

분석 실습

예측성능 향상과정

step	작업	학습/예측 모델	RMSLE	RMSE	MAE
1	대부분의 feature로 일단 시작 (아웃라이어 제거, 인코딩, log변환 등 아무것도 안 한 상태)	Linear Regression	1.159	141.136	106.056
2	아웃라이어 제거		1.162	130.812	98.611
3	Target 값 로그 변환		1.015	153.000	103.929
4	Category형 feature 원-핫 인코딩		0.582	93.652	60.183
5	앙상블 적용 (회귀 트리의 랜덤 포레스트)	RandomForestRegressor	0.334	46.020	28.217

Submission

분석 실습

Submission

❖ 대회가 종료되었기에 'Late Submission' 클릭

The screenshot shows the Kaggle Playground Prediction Competition page for 'Bike Sharing Demand'. The page title is 'Bike Sharing Demand' with the subtitle 'Forecast use of a city bikeshare system'. It is a Kaggle competition with 3,242 teams and was created 7 years ago. The navigation bar includes 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and 'Late Submission'. The 'Late Submission' button is highlighted with a red dashed border. Below the navigation bar, the 'YOUR RECENT SUBMISSION' section shows a submission named 'bike_submission.csv' with a score of 0.42092, submitted by 'seonyoungs' just now. A button 'Jump to your leaderboard position' is also visible.

Playground Prediction Competition


Bike Sharing Demand

Forecast use of a city bikeshare system

Kaggle · 3,242 teams · 7 years ago

Overview Data Code Discussion Leaderboard Rules Team My Submissions **Late Submission** ...

YOUR RECENT SUBMISSION

 **bike_submission.csv** Score: 0.42092
Submitted by seonyoungs · Submitted just now

↓ Jump to your leaderboard position

분석 실습

Submission











❖ Submission 후 'My Submission'에서 Score 확인

The screenshot shows the Kaggle Playground Prediction Competition page for 'Bike Sharing Demand'. The page header includes the competition title and a description: 'Forecast use of a city bikeshare system'. Below this, it indicates the competition is on Kaggle, with 3,242 teams and was created 7 years ago. A navigation bar at the top contains links for Overview, Data, Code, Discussion, Leaderboard, Rules, Team, My Submissions (highlighted with a red dashed box), Late Submission, and a menu icon. Below the navigation bar, the 'YOUR RECENT SUBMISSION' section displays a green checkmark icon, the filename 'bike_submission.csv', the submission details 'Submitted by seonyoungs · Submitted just now', and the score 'Score: 0.42092'. At the bottom of this section is a button that says '↓ Jump to your leaderboard position'.

분석 실습

Submission

❖ 'Leaderboard'에서 순위 확인

Overview	Data	Code	Discussion	Leaderboard	Rules	Team	My Submissions	Late Submission	...
This competition has completed. This leaderboard shows the final standings.									
#	Team	Members	Score	Entries	Last	Code			
1	Bolaka Mukherjee		0.33756	26	7Y				
2	Logical Guess		0.34821	25	7Y				
3	Louis Martin		0.34834	18	7Y				
4	张 李		0.34928	14	7Y				
5	rediculous		0.34928	5	7Y				
6	just_did		0.34930	17	7Y				
7	Greg		0.35570	47	7Y				
8	allmi		0.35595	9	7Y				
9	Gopal Joshi		0.35704	5	7Y				
10	Steven Lee		0.35783	23	7Y				

분석 실습

Wrap-Up

❖ 8일차 설문

■ <https://forms.gle/v5TUEpmx3Tdo8vs19>