DDWU_Bigdata 스터디 1번째 :

EDA 맛보기

Contents

kaggle

-2010년 설립된 빅데이터 솔루션 대회 플랫폼 회사

-2017년 3월, 구글에 인수

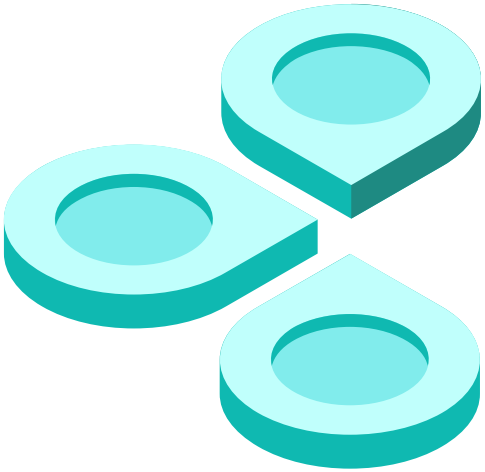
-Data science, ML, DL, AI등을 주제로 모인 커뮤니티 -〉 세계에서 가장 큰 규모의 커뮤니티

-현재 5만개 이상의 데이터 셋

왜 Kaggle을 할까?

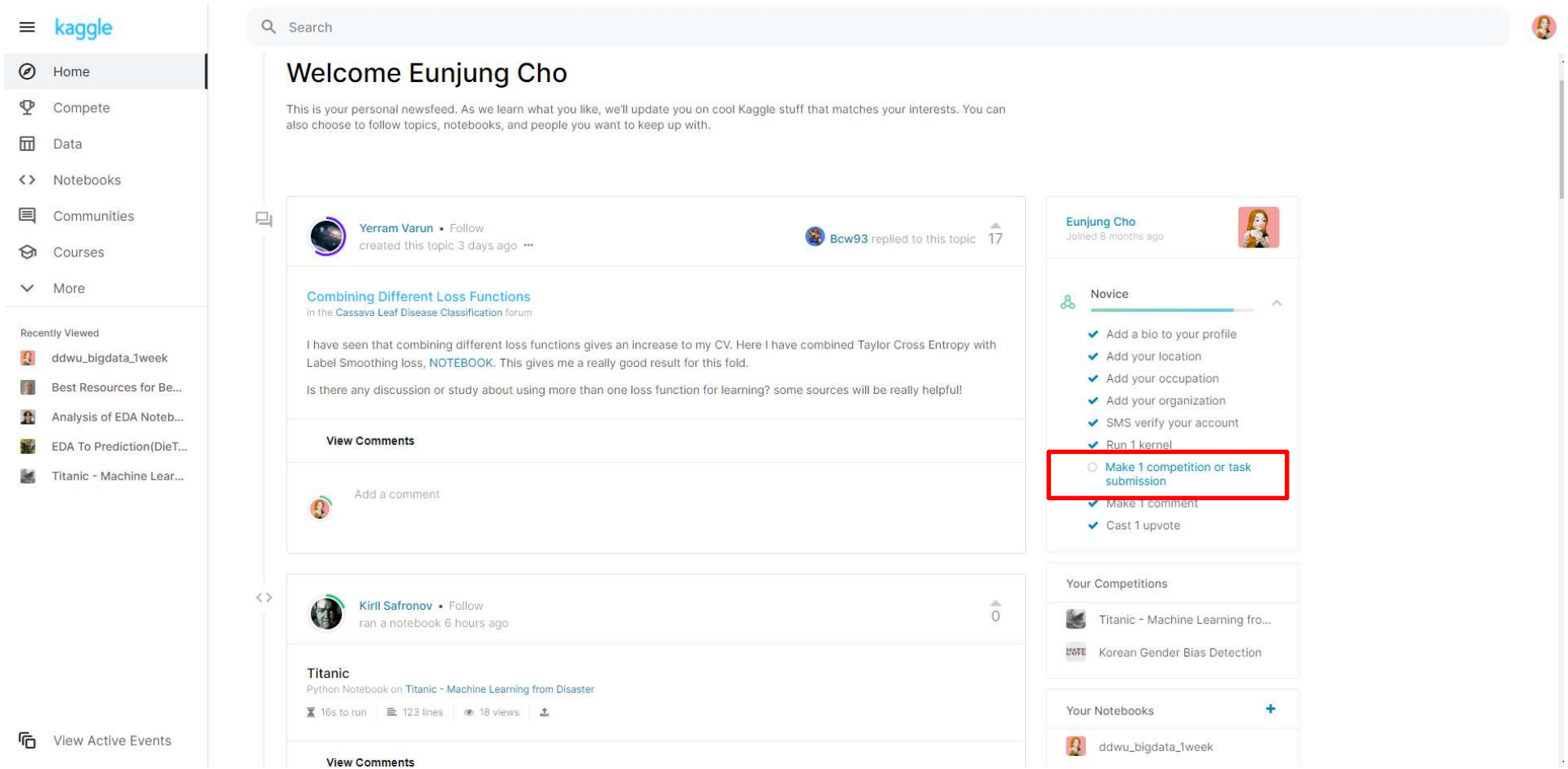
실 력 향 상 !

Machine Learning Certificate
머신러닝 경험해보았나요? -> 네!



Portfolio
어떤 분석을 해보았고 어떤 데이터를 다루었는지 증명!

Experience
여러 데이터셋을 다루면서 경험을 쌓을 수 있다.


kaggle
Search
Home
Compete
Data
Notebooks
Communities
Courses
More
Recently Viewed
ddwu_bigdata_1week
Best Resources for Be...
Analysis of EDA Noteb...
EDA To Prediction(DieT...
Titanic - Machine Lear...
View Active Events
Welcome Eunjung Cho
This is your personal newsfeed. As we learn what you like, we'll update you on cool Kaggle stuff that matches your interests. You can also choose to follow topics, notebooks, and people you want to keep up with.
Yerram Varun • Follow
created this topic 3 days ago
Bcw93 replied to this topic
17
Combining Different Loss Functions
in the Cassava Leaf Disease Classification forum
I have seen that combining different loss functions gives an increase to my CV. Here I have combined Taylor Cross Entropy with Label Smoothing loss, NOTEBOOK. This gives me a really good result for this fold.
Is there any discussion or study about using more than one loss function for learning? some sources will be really helpful!
View Comments
Add a comment
Kirill Safronov • Follow
ran a notebook 6 hours ago
0
Titanic
Python Notebook on Titanic - Machine Learning from Disaster
16s to run
123 lines
18 views
View Comments
Eunjung Cho
Joined 8 months ago
Novice
Add a bio to your profile
Add your location
Add your occupation
Add your organization
SMS verify your account
Run 1 kernel
Make 1 competition or task submission
Make 1 comment
Cast 1 upvote
Your Competitions
Titanic - Machine Learning fro...
Korean Gender Bias Detection
Your Notebooks
ddwu_bigdata_1week

데이터 분석...

어떤 것을 제일 먼저 해야 할까?

How make 경험 & 실력 for 정형데이터?

데이터 살펴보기!

1. Exploratory data analysis
   - Data visualization
     - Matplotlib, Seaborn, Plotly
     - Data mining
     - Pandas, numpy
3. Feature engineering
   - Time series features
   - Categorical features
   - Numerical features
   - Aggregation features
   - Ratio features
   - Product features

피처를 알아보자!

데이터준비하기
(불균형 어떻게 해결?)

2. Data preparation
   - Data augmentation (imbalance
     - Upsampling
     - Downsampling
     - SMOTE
4. Model development
   - Sklearn
     - Linear model
     - Non-linear model
     - Tree-model
   - Not sklearn
     - Xgboost
     - Lightgbm
     - Catboost
     - LibFFM

어떤 모델을 사용할까?
(이미 Kaggle 안에 잘 정리되어있다)

검증하기!

5. Model evaluation
   - Various metrics
     - Accuracy
     - Precision
     - Recall
     - F1-score
     - Etc.

- Other technique
  - Machine learning pipeline
    - My pipeline code
  - Feature management

EDA 란? – 탐색적 데이터 분석

수집한 데이터 :
다양한 각도에서 관찰하고 이해하는 과정
(with 그래프, 통계적인 방법)

과정

1) **분석 목적& 데이터의 변수 확인**

- 개별 변수의 이름이나 설명을 가지는지 확인

2) **데이터를 전체적으로 살펴보기**

- head, tail 먼저 보기
- 이상치, 결측치 등 확인

3) **데이터의 개별 속성값 관찰**

- 각 속성 값이 예측한 범위와 분포를 가지는지 확인

– if not ? -> 왜 그러는지 확인

4) **속성 간의 관계에 초점**

- 어떤 패턴이 있나? (상관관계, 시각화)

EDA 시작하기 가장 좋은 데이터!

Titanic Data Set

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S	
3	2	1	1	Cumings,	female	38	1	0	PC 17599	71.2833	C85	C	
4	3	1	3	Heikkinen	female	26	0	0	STON/O2.	7.925		S	
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S	
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S	
7	6	0	3	Moran, Mı	male		0	0	330877	8.4583		Q	
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S	
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S	
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S	
11	10	1	2	Nasser, Mı	female	14	1	0	237736	30.0708		C	
12	11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S	
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S	
14	13	0	3	Saundercc	male	20	0	0	A/5. 2151	8.05		S	
15	14	0	3	Andersson	male	39	1	5	347082	31.275		S	
16	15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S	

경험해보면 좋은 비기너용 데이터셋

여기서 사용한 알고리즘 및 모델 (tool) :
실제 현업에 적용!

Porto

고객이 내년에 자동차 보험금 청구를 할 것인가?

직방

아파트 거래가격 예측하기

Home Credit

고객이 앞으로 대출 상환을 할 것인가?

Elo

거래 내역 데이터를 가지고, 고객 충성도 예측하기

New York Taxi

Taxi 탑승시간 예측하기

Costa rican

고객의 소득 수준을 ML로 구분하기

Tips

여러가지 EDA 시작 참고:

https://www.kaggle.com/subinium/analysis-of-eda-notebooks?fbclid=IwAR22BnpaMYkdgDbcZFevrR308lr4RN3LJmzvmjT08uevxyHm-RYCqVMwiU4

TEAM EDA:

https://eda-ai-lab.tistory.com/13

Let’s start EDA!

지금부터 자유롭게 EDA를 해보세요!

- 필사적으로 필사하기!