

SLICED S00 E04

Data - Meta Kaggle

<https://www.kaggle.com/kaggle/meta-kaggle>

Specifically, this is the `Datasets.csv` and some related dataset CSV. The dataset is a running account of all the uploaded Kaggle datasets and all related info, including upvotes, downloads, kernels created, and dataset sizes.

Target - TotalViews

The object is to predict `TotalViews`, which is how many times a particular dataset gets viewed.

You are to maximize `R2`. Specifically, you can use Pearson R or Spearman R to calculate `R2`; either will be accepted.

s00e04-sliced_data.csv

Title: str title of the dataset
Subtitle str subtitle of the dataset (if available)
Name: str tag name (think like primary category name)
Id: int dataset unique ID
CreationDate: date created
VersionUpdate: date when dataset was last updated
DaysSinceCreation: int days since `CreationDate`
DaysSinceLastUpdate: int days since `VersionUpdate`
TotalViews: int total views for the dataset
TotalDownloads: int total downloads for the dataset
TotalVotes: int total votes for the dataset
TotalKernels: int total kernels created using the dataset
DatasetTagCount: int the estimated number of datasets using the given mix of tags for the dataset
CompetitionTagCount: int the estimated number of competitions using the given mix of tags for the dataset
KernelTagCount: int the estimated number of kernels using the given mix of tags for the dataset
TotalCompressedBytes: float size of the dataset, compressed (i.e. zip file compression)
TotalUncompressedBytes: float size of the dataset, uncompressed (i.e. raw files)

s00e04-holdout-data.csv

All columns are the same except the following column, which is NOT IN this file:

`TotalViews`

THE ABOVE COLUMN IS NOT IN THE HOLDOUT DATASET