Thank you for accepting the rules.

$40,000 • 362 teams

Home Depot Product Search Relevance

Merger and 1st Submission Deadlin

Mon 18 Jan 2016 Mon 25 Apr 2016 (3 months to go)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Scripts
- New Script
- New Notebook

Leaderboard

My Team

My Submissions

Leaderboard

1. ==
2. SecondPlan
3. TSM
4. NxGTR
5. Dimitris Leventis
6. ponythewhite
7. metabyr
8. Arto
9. Andre Naef
10. Alvah

200 Scripts

sklearn_random_forest
20 Votes / 5 days ago / Python

data exploration
7 Votes / 3 days ago / Python

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
sample_submission.csv	.zip (226.76 kb)
train.csv	.zip (2.51 mb)
test.csv	.zip (4.74 mb)
product_descriptions.csv	.zip (34.77 mb)
attributes.csv	.zip (27.21 mb)
relevance_instructions	.docx (105.01 kb)

This data set contains a number of products and real customer search terms from Home Depot's website. The challenge is to predict a relevance score for the provided combinations of search terms and products. To create the ground truth labels, Home Depot has crowdsourced the search/product pairs to multiple human raters.

The relevance is a number between 1 (not relevant) to 3 (highly relevant). For example, a search for "AA battery" would be considered highly relevant to a pack of size AA batteries (relevance = 3), mildly relevant to a cordless drill battery (relevance = 2), and not relevant to a snow shovel (relevance = 1).

Each pair was evaluated by at least three human raters. The provided relevance scores are the average value of the ratings. There are three additional things to know about the ratings:

- The specific instructions given to the raters is provided in relevance_instructions.docx.
- Raters did not have access to the attributes.
- Raters had access to product images, while the competition does not include images.

Your task is to predict the relevance for each pair listed in the test set. Note that the test set contains both seen and unseen search terms.

Beginner Data Analysis
10 Votes / 6 days ago / RMarkdown

Search Word Cloud
4 Votes / 2 days ago / R

Benchmark Score Script
11 Votes / 8 days ago / R

Exploring the Home Depot Data
8 Votes / 7 days ago / R

Forum (13 topics)

A closer look at the data
1 hour ago

What kinds of models to start?
19 hours ago

Typos in the product descriptions.
yesterday

Can we use product attributes data as well?
3 days ago

How
3 days ago

Looking to Join a team
4 days ago

teams

players

entries

File descriptions

- **train.csv** - the training set, contains products, searches, and relevance scores
- **test.csv** - the test set, contains products and searches. You must predict the relevance for these pairs.
- **product_descriptions.csv** - contains a text description of each product. You may join this table to the training or test set via the product_uid.
- **attributes.csv** - provides extended information about a subset of the products (typically representing detailed technical specifications). Not every product will have attributes.
- **sample_submission.csv** - a file showing the correct submission format
- **relevance_instructions.docx** - the instructions provided to human raters

Data fields

- **id** - a unique Id field which represents a (search_term, product_uid) pair
- **product_uid** - an id for the products
- **product_title** - the product title
- **product_description** - the text description of the product (may contain HTML content)
- **search_term** - the search query
- **relevance** - the average of the relevance ratings for a given id
- **name** - an attribute name
- **value** - the attribute's value

© 2016 Kaggle Inc