

Appyhigh ML Assignment

May 03, 2020

Assignment Overview

The assignment is about building an AI model that could predict the likes of posts on our platform X. We power engaging content in various interests to our users across multiple countries. All of the content is being scrapped from the web from multiple sources. If we are able to predict the likes of any post, to some extent we can predict the level of the interest that the same post can create among the users and the relevancy of the content to them.

Data Set Introduction

To download the data, please find below the download link

<https://drive.google.com/file/d/1n3ctIDQFGOzAAcLJFsuM2dwZFfCGdHgL/view?usp=sharing>

There are two datasets shared

- **train.csv** - This is the training data set. Candidates can use this data to train & build the model. There are 369,921 posts in this data set
- **test_without_truth.csv** - After your model is ready and optimized, you can run your model on this test data and add the predicted likes against all the posts in this data set. The test data has 158,542 posts

Both the data sets have below features in them. Go through the below pointers to understand the data better

- **post_id** is the unique identifier for every post
- **user_id** is the unique identifier for the publisher who published the post
- **country** is the primary geolocation of the publisher/user
- **category** is the interest that best describes the post content

- **#views** is the number of times the post is viewed by the content consumers
- **#comments** is the number of times the content consumers commented on the post
- **#likes** is the number of times the content consumers liked the post. This is the variable that the model needs to predict

Evaluation Metric

Candidates are free to choose any machine learning technique for this whichever helps them in achieving better performance. The evaluation will be made on the basis of Root Mean Square Error (RMSE) for the predicted likes in the test data and also on the approach undertaken.

Submission -

Candidates are required to submit below two documents

- jupyter notebook that explains their code, thought-process and how the likes are predicted
- Updated test data CSV with likes predicted against all the posts

Please note that the candidates need to submit the assignment within 5 days of receiving this document. You can share your jupyter notebook and updated test data CSV at akshit@appyhigh.com and aneesh@appyhigh.com