p13n-takehome-ml

Interview Take Home Project for the O'Reilly Personalization Team

For this project you will analyze part of a public "click through" dataset and build a model that predicts the probability that an ad will be clicked.

The data

The data for this take home can be found here:Click Data

The archive contains two files:

```
p13n-takehome-ml/
  sampled_training
  sampled_test
```

All data is in csv format.

Data Fields

```
 id -- ad identifier
click -- 0/1 for non-click/click
hour -- format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
C1 -- anonymized categorical variable
banner_pos
site_id
site_domain
site_category
app_id
app_domain
app_category
device_id
device_ip
device_model
device_type
device_conn_type
C14-C21 -- anonymized categorical variables
```

File Descriptions

`sampled_training` contains roughly 10 days worth of click data that has been downsampled from the original data file, all data is ordered chronologically. The first line in this file is the column headers.

`sampled_test` contains click data that you will make predictions against when you submit your project. All of the events it this dataset are ordered chronologically and occur after the events in `sampled_training`. This

file does not include column headers, but the data is in the same order as in `sampled_training` except for the fact that column 2 is excluded from the data set.

What you need to do

You need to analyze the dataset and develop a strategy for optimally predicting the probability a click will occur in the future. You should then train a model to predict the probability of click occuring based on the data in the dataset `sampled_test`. The output from this prediction should be a number between 0 and 1. This project should take you approximately 3 hours to complete. (Don't choose a compute intensive ml algorithm when training your model. We mostly care about your thought process when working on the problem. Don't be too concerned with squeezing every little bit out of your model.)

What you need to submit

You will submit (by executing a pull request on this project from your fork): * all code / notebooks etc. that you used in exploring the data and building your model. * a file containing the predictions from your model when executed on the `sampled_test` file.

All files should submitted as a **pull request** to this project. You *should not* submit the raw data as part of your project submission. Once you have submitted your project please notify your hiring contact that your take home project has been submitted. **If you do not submit your project as a pull request, your project will be ignored.**