BerkeleyX: CS190.1x Scalable Machine Learning

LAB4: CLICK-THROUGH RATE PREDICTION PIPELINE (100/100 points)

Once you have successfully run the test notebook, you can submit to the course autograder by first **exporting lab4 as a Python (.py) file**, and then using the file chooser to select your file and clicking *"Check"* to submit your code.

Before submitting your assignment, please ensure that your submission follows these guidelines:

- Only use the following libraries: standard python libraries, numpy, pyspark, and test_helper (the autograder library).
- Don't include any extraneous code as the autograder will timeout if a submission takes too long.
- Only change sections of code where you see FILL IN. Changing other parts of the code, including directory paths, may cause the code to fail the autograder's tests.
- The autograder grades submissions using the same tests that are included in your notebook. Hence, please run the local tests before submitting to the autograder.

For further background on the autograder please visit this page with the information from Week 0 and the autograder FAQ. To check the status of your submission, please visit the autograder submission website.

This is the autograder only for your submission of "Lab4 - CTR Prediction Pipeline". Please **DO NOT submit other labs to this autograder.**

Choose Files No file chosen

```
One-hot-encoding (1a)
---------------------
All tests passed
Sparse Vectors (1b)
-------------------
All tests passed
OHE Features as sparse vectors (1c)
-----------------------------------
All tests passed
Define an OHE Function (1d)
---------------------------
All tests passed
Apply OHE to a dataset (1e)
---------------------------
All tests passed
Pair RDD of (featureID, category) (2a)
--------------------------------------
All tests passed
OHE Dictionary from distinct features (2b)
------------------------------------------
All tests passed
```

```
Automated creation of an OHE dictionary (2c)
--------------------------------------------
All tests passed
Loading and splitting the data (3a)
-----------------------------------
All tests passed
Extract features (3b)
---------------------
All tests passed
Create an OHE dictionary from the dataset (3c)
----------------------------------------------
All tests passed
Apply OHE to the dataset (3d)
-----------------------------
All tests passed
Handling unseen features (3e)
-----------------------------
All tests passed
Logistic regression (4a)
------------------------
All tests passed
Log loss (4b)
-------------
All tests passed
Baseline log loss (4c)
----------------------
All tests passed
Predicted probability (4d)
--------------------------
All tests passed
Evaluate the model (4e)
-----------------------
All tests passed
Validation log loss (4f)
------------------------
All tests passed
Hash function (5a)
------------------
All tests passed
Creating hashed features (5b)
-----------------------------
All tests passed
Sparsity (5c)
-------------
All tests passed
Logistic model with hashed features (5d)
----------------------------------------
All tests passed
Evaluate on the test set (5e)
-----------------------------
All tests passed
-- 24 cases passed (100.0%) --


Your submission token ID is 2058574-34effac15fa589270976853767b5ee9c:ip-172-31-16-68
Please include this submission token ID when you need support for your code submission.
Your anonymous student ID is d9e65bc8d252ec579ef766790c87772e. Do not post this ID on
Piazza.
```

CHECK **SAVE** *You have used 1 of 10 submissions*