Cmd 1

Cmd 2

Baseline Solution

Objective: *Develop a baseline solution to a business problem.*

In this lab, you will complete a series of exercises to develop a baseline solution to determine whether health tracker users are from the United States or Canada.

Cmd 3

```
%run "../../Includes/Classroom-Setup"
```

▸ (5) Spark Jobs

Command took 3.70 minutes -- by tjamesbu@gmail.com at 4/9/2021, 4:20:39 PM on My Cluster

Cmd 4

Exercise 1

Summary: Randomly split health tracker users `dsfda.ht_users` into a training set (80 percent of users) and a test set (20 percent of users).

Cmd 5

```
%python
# ANSWER

# Load train_test_split
from sklearn.model_selection import train_test_split

# Split into training and test sets
ht_users_df = spark.sql("SELECT device_id, country FROM dsfda.ht_users").toPandas()
ht_users_train_df, ht_users_test_df = train_test_split(ht_users_df, test_size = 0.2, random_state = 42)

# Convert to Spark DataFrames
ht_users_train_sdf = spark.createDataFrame(ht_users_train_df)
ht_users_test_sdf = spark.createDataFrame(ht_users_test_df)

# Create tables for future SQL usage
ht_users_train_sdf.write.format("delta").mode("overwrite").save("/dsfda/ht_users_train")
spark.sql(
  "CREATE TABLE IF NOT EXISTS dsfda.ht_users_train USING DELTA LOCATION '/dsfda/ht_users_train'"
)
ht_users_test_sdf.write.format("delta").mode("overwrite").save("/dsfda/ht_users_test")
spark.sql(
  "CREATE TABLE IF NOT EXISTS dsfda.ht_users_test USING DELTA LOCATION '/dsfda/ht_users_test'"
)
```

▸ (9) Spark Jobs
▸ ht_users_train_sdf: pyspark.sql.dataframe.DataFrame = [device_id: string, country: string]
▸ ht_users_test_sdf: pyspark.sql.dataframe.DataFrame = [device_id: string, country: string]

```
Out[7]: DataFrame[]
```

Instrument ML code with MLflow: Use MLflow to track metrics, params, and models from your training code. Learn more | Don't show me this again

Command took 22.84 seconds -- by tjamesbu@gmail.com at 4/9/2021, 4:20:39 PM on My Cluster

Cmd 6

Exercise 2

Summary: Identify what proportion of health tracker users are from the United States and Canada, respectively.

Cmd 7

```
%sql
-- ANSWER
SELECT b.country, b.grouped_total / a.total AS proportion
FROM (SELECT count(*) as total FROM dsfda.ht_users_train) a,
     (SELECT country, count(*) as grouped_total FROM dsfda.ht_users_train GROUP BY country) b
```

▸ (4) Spark Jobs

	country	proportion
1	United States	0.9770833333333333
2	Canada	0.02291666666666665

Showing all 2 rows.

Command took 3.43 seconds -- by tjamesbu@gmail.com at 4/9/2021, 4:20:39 PM on My Cluster

Cmd 8

Exercise 3

Summary: Apply a most-common case baseline solution to the test set and save to a new table.

Cmd 9

```sql
%sql
-- ANSWER
SELECT device_id, "United States" as predicted_country
FROM dsfda.ht_users_test
```

(3) Spark Jobs

	device_id	predicted_country
1	f9a35106-e48a-11ea-8204-0242ac110002	United States
2	d621c744-e48a-11ea-8204-0242ac110002	United States
3	f7a1da76-e48a-11ea-8204-0242ac110002	United States
4	fce425f2-e48a-11ea-8204-0242ac110002	United States
5	d8e444ca-e48a-11ea-8204-0242ac110002	United States
6	0150756e-e48b-11ea-8204-0242ac110002	United States
7	dfca7fca-e48a-11ea-8204-0242ac110002	United States

Showing all 600 rows.

Command took 2.70 seconds -- by tjamesbu@gmail.com at 4/9/2021, 4:20:39 PM on My Cluster

Cmd 10

Exercise 4

Summary: Evaluate the baseline solution's accuracy on the test data.

Cmd 11

```sql
%sql
-- ANSWER
SELECT a.number_correct / b.number_total AS accuracy
FROM (SELECT count(*) AS number_correct
      FROM dsfda.ht_users_test
      WHERE country = "United States") a,
     (SELECT count(*) AS number_total FROM dsfda.ht_users_test) b
```

(4) Spark Jobs

	accuracy
1	0.97

Showing all 1 rows.

Command took 2.38 seconds -- by tjamesbu@gmail.com at 4/9/2021, 4:20:39 PM on My Cluster

Cmd 12

Great job completing the Baseline Solution lab! Continue on with the lesson to learn about measuring solutions in real-world settings.

Cmd 13

Shift+Enter to run