**Got Data? Use Spark in Azure Databricks**

Challenge 4 – Introduction to Using Spark on Databricks for Machine Learning

What You’ll Need

To complete the labs, you will need the following:

• A web browser

• A Microsoft account

• A Microsoft Azure subscription

• A Windows, Linux, or Mac OS X computer

• Azure Storage Explorer

• The lab files for this course <only if you get stuck – proctors have access to this content>

**Note**: To set up the required environment for the lab, follow the instructions in the [**Setup**](https://microsoft-my.sharepoint.com/personal/laedell_microsoft_com/Documents/Hackathons/ML/Databricks/databricks_intro/databricks-introFinal/Setup%20Guide.docx)document for

this course. Specifically, you must have signed up for an Azure subscription.

Challenge Background:

Spark includes an API named Spark MLLib (often referred to as Spark ML), which you can use to create machine learning solutions. Machine learning is a technique in which you train a predictive *model* using a large volume of data so that when new data is submitted to the model it can predict unknown values.

The most common types of machine learning are supervised learning and unsupervised learning. In a

supervised learning scenario, you start with a large volume of data that includes both features

(categorical and numeric values that describe characteristics of the entity you’re trying to predict

something about) and labels (the value your model will predict. Training the model involves applying a

statistical algorithm that fits the features to the labels. Because your initial data includes known values

for the labels, you can train the model and test its accuracy with these known label values – giving you

confidence that the model will work accurately with new data for which the label values aren’t known.

Unsupervised learning is a technique in which there are no known label values, and the model is trained

to group (or cluster) similar entities together based on their features.

In this lab we’ll focus on supervised learning and specifically a type of machine learning called

classification in which you train a model to identify which category, or class an entity belongs to.

Business Challenge: can you accurately predict flight performance based on historical flight route patterns?

Questions:

1. What features are most important in your model relative to flights that are enroute to an airport?
2. What is the accuracy of your flight ontime predictions ? Can you think of ways to improve your accuracy?

To get you started, the 1st thing you will need to do is to Import your MachineLearning.py or .scala script into your Databricks Workspace. How? Ask your table mates.

\*\* remember: you can use scala or python; you can use Databricks notebooks or native scripts (.py, for example) to solve the Challenge \*\*

\*\*\*Lastly. If you get stuck, proctors can help you without providing the answer.

**Challenge 4:**

Steps TBD

END OF CHALLENGE 4