Exercise Overview

In this exercise we will play with Spark Datasets & Dataframes (https://spark.apache.org/docs/latest/sql-programming-guide.html#datasets-and-dataframes), some Spark SQL (https://spark.apache.org/docs/latest/sql-programming-guide.html#sql), and build a couple of binary classifiaction models using Spark ML (https://spark.apache.org/docs/latest/ml-guide.html) (with some MLlib (https://spark.apache.org/mllib/) too).

The set up and approach will not be too dissimilar to the standard type of approach you might do in Sklearn (http://scikit-learn.org/stable/index.html). Spark has matured to the stage now where for 90% of what you need to do (when analysing tabular data) should be possible with Spark dataframes, SQL, and ML libraries. This is where this exercise is mainly trying to focus.

Feel free to adapt this exercise to play with other datasets readily availabe in the Databricks enviornment (they are listed in a cell below). #####Getting Started To get started you will need to create and attach a databricks spark cluster to this notebook. This notebook was developed on a cluster created with:

- Databricks Runtime Version 4.0 (includes Apache Spark 2.3.0, Scala 2.11)
- Python Version 3

Links & References

Some useful links and references of sources used in creating this exercise:

Note: Right click and open as new tab!

1. Latest Spark Docs (https://spark.apache.org/docs/latest/index.html)
2. Databricks Homepage (https://databricks.com/)
3. Databricks Community Edition FAQ (https://databricks.com/product/faq/community-edition)
4. Databricks Self Paced Training (https://databricks.com/training-overview/training-self-paced)