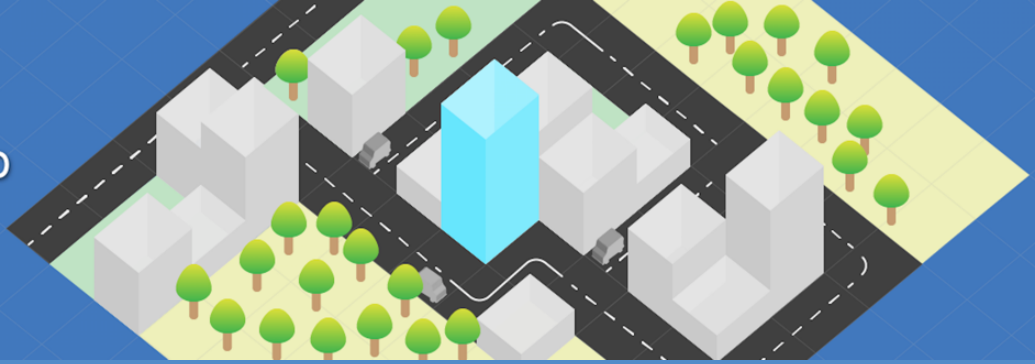
IBM

Machine Learning **Hub**

Learn. Innovate. Apply.



Machine Learning 101

Óscar Lara-Yejas, Ph.D.
Senior Data Scientist, Machine Learning Hub
odlaraye@us.ibm.com

Hands-on exercise # 0: Load

. Create an R notebook

. Use **`download.file()`** function

. Data location: https://github.com/olarayej/spark/raw/master/airline.csv

. Use R's function **`read.csv()`**. Assign the result to a variable named **`air`**

. Use functions **`str()`** and **`head()`** on **`air`**

Hint: Always use function head() to visualize the data contents. Otherwise, the entire dataset may be shipped to the browser.

Hands-on exercise # 1: Curate

- **Remove columns with more than 70% missing values**
  - In R, missing values are marked as `NA`. Use function `is.na()` (works for vectors and data.frames)
  - Use function `colSums()` to get the number of NA values per column.
  - Use function `nrow()` to get the number of rows of a data.frame.
  - Use operator `[,]` to remove columns. Example: `data[ , c(-1,-2)]` removes columns 1 and 2
- **Remove all rows with missing values**
  - Use function `na.exclude()`
- **Remove all canceled and diverted flights**
  - Use columns Cancelled and Diverted
  - Use operator `[,]` for filtering with predicates
  - Example: to filter rows in data.frame:

```
data[data$name == "value" & data$age > value2, ]
```

- Store the result in a variable called `airCurated`

Not a syntax error

Hands-on exercise # 2: Explore

- Plot a histogram of ArrDelay
  - Project a column using operator $. Example: `data$col`
  - Use function `hist()`
- Plot flights per year
  - Get the distinct counts for a column using function `table()`
  - Use function `barplot()` to plot the result of `table()`
- Find the busiest airlines
  - Use function `sort()` with parameter decreasing=T
- Find the busiest times to fly
- Find the busiest airports
- Find which columns are correlated with ArrDelay
  - Use `cor()` function

Hands-on exercise # 3

. Remove non-meaningful attributes from the airline dataset.

. Create a new attribute **`class`** from **`ArrDelay`** with three possible values: *delayed* (>15 min), *early* (<0), and *on-time* (otherwise).

. To add a column to a dataset in R: **`data$new_col <- value`**

. Use function **`ifelse()`**.

. Plot the distribution of the three classes.

Hands-on exercise #4: Modeling

- Build a SVM model for the airline dataset
  - Work on a smaller sample of the data (first 10,000 rows)
  - Load SVM package using `library(e1071)`
  - Train a model:

```
model <- svm(y ~ ., dataset)
```

Label column

Formula notation

Nope. No syntax error. The dot is to indicate all features will be used as predictors

- Make predictions on the dataset
  - Use function `predict()`. Parameters are model and data.
- Evaluate the model
  - Compare predictions with labels. Compute overall accuracy.
  - Use `table()` function with two parameters: labels and predictions. The result is called Confusion Matrix.

Hands-on exercise #5

. Split the dataset randomly into 70% for training and 30% for testing.

. Build a SVM model on the training set.

. Make predictions on the testing set.

. Evaluate the model.

Hands-on exercise #6

. Remove attributes that are unknown in practice.

. Split the dataset randomly into 70% for training and 30% for testing.

. Build a SVM model on the training set.

. Make predictions on the testing set.

. Evaluate the model.