**DS Automation Assignment**

**Using our prepared churn data from week 2:**

- use pycaret to find an ML algorithm that performs best on the data

- Choose a metric you think is best to use for finding the best model; by default, it is accuracy but it could be AUC, precision, recall, etc.

- save the model to disk

- create a Python script/file/module with a function that takes a pandas dataframe as an input and returns the probability of churn for each row in the dataframe

- your Python file/function should print out the predictions for new data (new\_churn\_data.csv)

- the true values for the new data are [1, 0, 0, 1, 0] if you're interested

- test your Python module and function with the new data, new\_churn\_data.csv

- write a short summary of the process and results at the end of this notebook

- upload this Jupyter Notebook and Python file to a Github repository, and turn in a link to the repository in the week 5 assignment dropbox

\*Optional\* challenges:

- return the probability of churn for each new prediction, and the percentile where that prediction is in the distribution of probability predictions from the training dataset (e.g. a high probability of churn like 0.78 might be at the 90th percentile)

- use other autoML packages, such as TPOT, H2O, MLBox, etc, and compare performance and features with pycaret

- create a class in your Python module to hold the functions that you created

- accept user input to specify a file using a tool such as Python's `input()` function, the `click` package for command-line arguments, or a GUI

- Use the unmodified churn data (new\_unmodified\_churn\_data.csv) in your Python script. This will require adding the same preprocessing steps from week 2 since this data is like the original unmodified dataset from week 1.