Technical Challenge Instructions

When developing a machine learning model from a cleaned dataset, you first need to split the data into three sets: The training set, the validation set, and the testing set. How you split the data is a very important issue.

For this project you will be using a stratification method to decide how to split the data. In the dataset, you will have several categorical variables (variables whose value takes on one of a discrete set of categories). Some of these variables are particularly important to our model and we would like to be sure that we preserve the proportions of these variables' categories from the original dataset within each of our new datasets.

For example, say you have an important categorical variable named *var1* which can take values in the set {A, B, C, D}. Further suppose that within the original dataset, 40% of rows have value A for *var1*, 25% have value B, 20% have value C, and 15% have value D. Your code should split the data such that for the training set, 40% of rows have value A for *var1*, 25% have value B, 20% have value C, and 15% have value D. The same should be true for the testing and validation sets.

We may want to preserve the proportions of an arbitrary number of variables, so your code should be scalable. The proportions do not have to be exactly the same in all the sets, as long as it is as close as possible. The training, validation, and testing sets should be disjoint and cover all rows in the original dataset. You should have the training data make up 70% of the original dataset, and the validation and testing sets should each make up 15% of the original dataset. Each dataset you create should be saved to its own .csv file.

I have provided a sample dataset for you to run your code on. I will be testing your code on this dataset and on other datasets of similar format that you will not have access to. All datasets will be in csv format. Remember that we may want to preserve the proportions of multiple variables in the dataset. In addition, the dataset may contain mixed categorical and numeric variables. We will only ever ask you to preserve the proportions of a categorical variable.

Along with the submission of your code, please also include a README file. In the readme you must give instructions on how to run your code including all inputs the code requires. You must also explain what output the code will provide. In addition, please briefly describe the algorithm you implemented. Also provide runtime bounds of your algorithm using big O notation.

Please write your code using python. If you think another language would be better suited to this task, use python anyway, but also explain in your README which other language you would have chosen and why you would have preferred it.

You may use any and all resources at your disposal, however you may not have another person write any of the code for you. If you have taken any code from a website, please include the link to this website in your README, and in a comment where the code is implemented.

Code is expected to be efficient, concise, and scalable. Please include your name at the top of all files you submit. Good luck, and feel free to email me (Jack Teitel at Jack_Teitel@urmc.rochester.edu) if you have any questions.