Reproducible Research is key to the advancement in Sciences and especially Data Science.

Goal/Problem Statement: If you are doing research with an objective in mind, make sure to state it. You might also want to state ahead of time how you will measure success. It has been seen that researchers, intentionally or not, change their standards or ways of measuring to get a conclusion that is convenient. You can remind yourself not to do this but setting your standard ahead of time.

There are times when you are just exploring and are not led by a clear objective, you can still capture your intent in these cases.

Data Sources: Your research is probably based on data. Make sure to keep a reference to the data. If possible, save a copy of it. If your data comes from external sources, it is possible that these sources can disappear or not retain data that your research was based upon. Prefer publicly available data without personal information if possible.

Cleaning/Feature Engineering: Typically, you will end up cleaning and transforming your data. You will end up doing feature engineering and possible dimension reduction. Be sure to save well commented code that does these transformations. Resist the temptation to do any manual cleaning or transformation. For this and other reasons below, you might want to document your work in a notebook and use Markdown.

Libraries and environments: You will likely use libraries, compilers, models, model implementations, Operating Systems, virtual machines, databases…. that you did not develop but chose to take a dependency on. Make sure to attribute these tools that enabled you. Also, capture detailed information about them, the version numbers, package names, sizes and other details so your code can be run under the right environment. If possible, you can save docker images or setup instructions/scripts so someone else can get the same environment ready.

Sampling/training data/test data: You will often divide your data into training data and test data, be clear on the strategy you use, so that you do not bias the model and yourself. Document the strategy you used.

Labeling: You may need to label data at times. For example: if you are developing an image detection model to distinguish cats and dogs, you need some data that has been labelled already. If you end up using humans to label the data, you should capture details of the labelling process and the results. Ensure that you get more than one labeler for the same data so you can be sure that the labels are accurate.

Model comparisons: You will have a choice of models that you can leverage. Often, you will try to make a prediction or classification with several similar models. You will compare the effectiveness of the model using measures sch as precision, recall, accuracy, gain, ROC curves, F1 scores… Even though you may take the winning model for the rest of your research – save your results from models your tried.

Champions/Challengers: Once you have a winning model, do not settle there. Try to improve it. Try to get more data. Your current champion may be displaced by a challenger. A good model is one that has typically won a few rounds of competition.

Interpret your results: You should make sound conclusion from your data science work. You have to remain factual (mathematically and statistically clean). Try not to sugar coat or use weasel words to describe your conclusions.

Next steps: You should clearly document what future researcher should pursue. Maybe they can solve the same problem but for a different domain. Maybe they need to focus on getting more data. Maybe they need to focus on ethical application of the work or improving the model's accuracy or performance. Give others some thing that they can build upon. Remember, if your work is used in someone else's research, it is credit for you as well.

Lessons Learnt: While the project is still fresh in your mind, write down things that went well and things that can be improved. How could someone else learn from you without having the experience of doing this work themselves.

Kaushik Pushpavanam – Principal Project Manager, Big Data, Microsoft

December 17, 2020