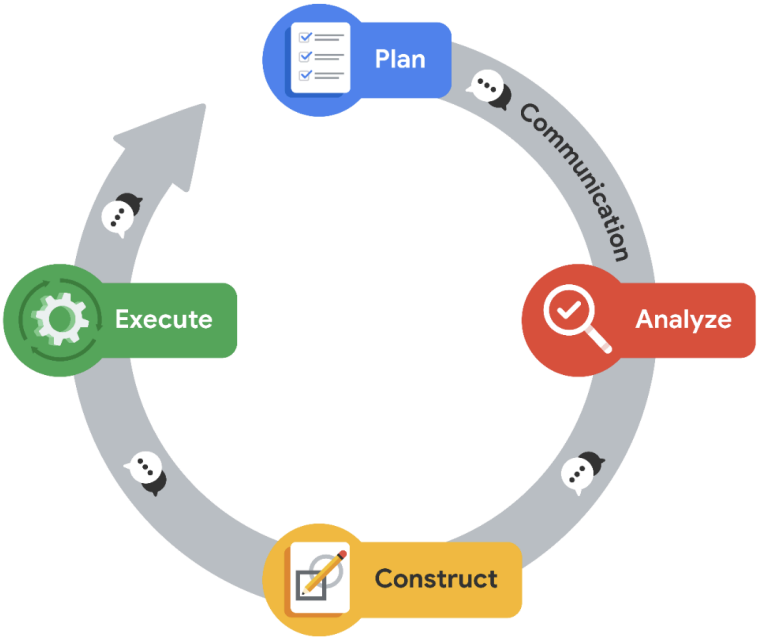
More about planning a machine learning project

The PACE workflow is something that can be used to keep the most experienced data professionals on track in their projects. In this reading, you will learn more about the Plan stage of PACE and the things that must be considered and determined to ensure a smooth and successful model development process.

The Plan Stage

The PACE workflow is something that you can use to keep you on track, no matter the project you're working on. Each step is important to get to your final product. However, like many things, the most important part is setting up the foundations of your project - The Plan Stage.

The Plan Stage is the part of the process where you first start thinking about what the problem actually is, and what needs to be done to find a solution. You start to consider what tools you have available to you, and how you'll need to manipulate the dataset. Sometimes this can be as straightforward as needing to create some visualizations for the data. Or, it can get as complex as needing to make a predictive model using the dataset.



The plan that you create during this stage will be carried through the whole process, so it is important to really make sure you've considered all the aspects and constraints of the project. However, that isn't to say that the plan you create must stay unchanging, you can absolutely reassess as you progress. It is there to get you started heading in the right direction.

What should your Plan Include?

This section of the course focuses on machine learning algorithms, so we will use those types of projects as the example here. However, you need to think about whether you need a model in the first place! Many analytical tasks do not require the creation of a model, and you could spend time creating something that is not necessary to what you're trying to achieve.

Knowing What You Need For a Problem

The first thing to do when forming your plan is to consider the end goal. What exactly are you trying to model, and what types of results from the model are needed? Something that can be determined immediately is what type of machine learning model you'll need. The two types that you've seen so far are Supervised and Unsupervised models.

Supervised models are used to make predictions about unseen events. These types of models use labeled data, and the model will use these labels and the predictor variables present to learn from the dataset. And when given new data points, they're able to make a prediction of the label. So, for example, if you're tasked with predicting rainfall amounts, you already know that you will need a supervised learning model.

Unsupervised models, on the other hand, don't really make predictions. They are used to discover the natural structure of the data, finding relationships within unlabeled data. So, for example, if you're tasked with discovering relationships between customer habits and segment users, you know you'll need an unsupervised model.

Now, let's go back to the rainfall example. Just from that problem statement alone, we know we need a supervised learning model. However, not all supervised learning models are the same. The two main types of supervised learning are Regression and Classification. There are different types of regression models that you have practiced, with different models able to perform regression or classification tasks.

Linear regression models are used when the result must be a continuous variable. As you have learned, continuous variables are numerical values that can have an unlimited number of values between the highest and lowest points of measurement. So if you need rainfall amounts in inches or centimeters, you know a linear regression model is needed.

However, what if we don't need exact rainfall amount predictions, but just whether or not it will rain that day? This is where a classification model, such as a logistic regression model, would be more appropriate. Classification models will deliver results as a categorical variable, where there is a finite set of values that the variable can be. In this example, the model would only ever predict two results: Will Rain or Won't Rain.

Figuring out the tools you need

After you've determined what type of model you're going to need, you must consider what you have at your disposal to complete the project. Most importantly, you need to figure out if you have the data you'll need to build the model.

If your dataset only has one or two predictor variables, it probably will not produce a model that will be useful. Or, if it has very few data points, the model's performance will similarly suffer. On the other hand, your dataset might be large and unwieldy, meaning that you'll either need to clean it up or cut it down to get it into a format that you can use to train the model. Having these issues means that you'll have to put in a little extra work to get it to usable form, or look elsewhere for data that will be helpful to create the model.

Key Takeaways

- The PACE workflow for machine learning is very useful for planning out and solving data driven problems.
- The Plan stage of PACE is one of the most important, setting you up for success throughout the rest of the process
- In the Plan stage, you first consider the problem at hand and what will be needed to solve it
- You also verify that you have the tools and resources you need to solve the problem
- The Plan is not set in stone. It just serves as a foundational starting point for the rest of the project