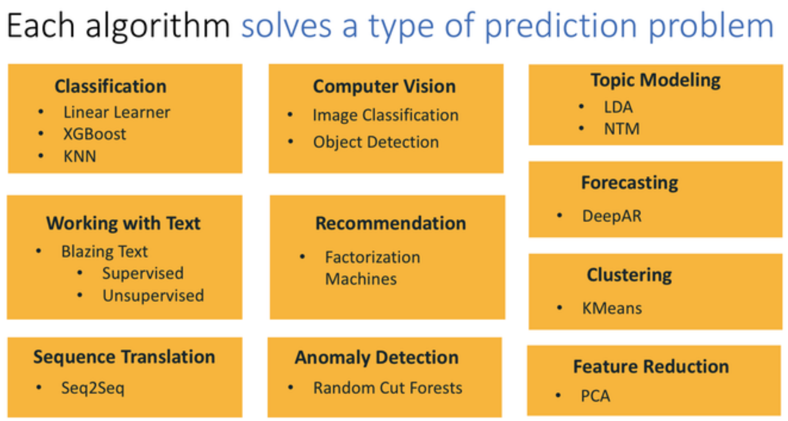
Bring Your Own Data Set: Tenets for a Successful Machine Learning Project Definition

Start with the business introduction.
Open the 2-page project description with a business introduction. What is the project about, and why is it valuable? Keep this to 5 sentences.

Move into the machine learning method.
Pick one type of machine learning solution class that your participants will focus on for the 3-day course. Keep this to methods that are generally known, so they don't have to re-invent the wheel and can focus on implementing a tried-and-tried method. You can use the chart below to help you determine which type of problem you would like to go after, and which type of algorithm you will most likely use to solve that problem.



Each algorithm solves a type of prediction problem

Classification
- Linear Learner
- XGBoost
- KNN

Working with Text
- Blazing Text
  - Supervised
  - Unsupervised

Sequence Translation
- Seq2Seq

Computer Vision
- Image Classification
- Object Detection

Recommendation
- Factorization Machines

Anomaly Detection
- Random Cut Forests

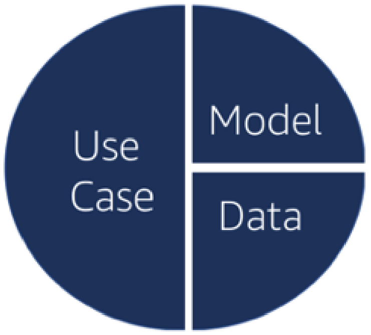
Topic Modeling
- LDA
- NTM

Forecasting
- DeepAR

Clustering
- KMeans

Feature Reduction
- PCA

This is a loose map that combines SageMaker algorithms with types of prediction problems and/or classes of machine learning methods. You can use this to begin to thing about problems that you can solve using your data set, and how they might map to business solutions you are interested in pursuing.



Describe your data set content.

Here you should go into some level of depth about your data set. What types of rows do you have, what types of columns? What are some example of the column headers? Are they integers, floats, strings? Are they specific to a time of day? When thinking about the type of data that you are going to use for your project, remember the following pie chart that explains the fundamental theory of machine learning. You need to use a data set that fully describes your use case, as thoroughly as you possibly can.

Connect your machine learning solution with your business problem
Clearly describe how your machine learning solution maps to solving your business problem. If you are using classification, how is your classification system assisting your business goals? Try to pick a data set that naturally maps to a set of machine learning solutions. When the data set is well aligned to the model, which is well aligned to your use case, the rest of the project will flow naturally.

Preparing your data for analysis
It should be somewhat easy to get the data into a Pandas data frame. That is, the data should be generally in a columnar format, with only a few additional steps necessary to being analyzing it. This step should still require some effort from your participants, otherwise they will not learn how to work with a real data set.

Select the right amount of data

The amount of data for your participants should be not too large, but also not too small. Anything over 10 GB is absolutely too much, and anything under 1 GB is absolutely too small. After engaging in this course you might discover that you really want to scale a solution, in which case you can use a larger data set. The 3-day course is perfect for discovering whether or not a given solution class is viable for your business problem, and which tools you would prefer to use for those.

Point to references where this problem has been attempted previously

Ideally, try to include links to either white papers or blog posts of practitioners attempting to solve this problem, along with what they learned along the way. This will help your participants develop more confidence in their approach.

Include reference code in Python

Lastly, if necessary, include reference code in Python that help your participants easily get started reading your files into the SageMaker notebook and playing with that data in Pandas.