Classification FAQs

**Question 1:**

How do I know which model to choose for my problem?

**Answer 1:**

The first step is to figure out whether your problem is linear or non-linear, i.e., do the variables have a linear relationship or not?

If they do, logistic regression would be the first port of call, then SVM. If the relationship between variables is non-linear, then implement K-NN, Naïve Bayes, Decision Tree or Random Forest.

**Question 2:**

How can I improve each of these models?

**Answer 2:**

To select a model, apply a model selection process called k-Fold Cross Validation. Then from a business point of view you would rather use:

Logistic regression or naïve bayes when you want to **rank your predictions** **by their probability**. For example, if you want to rank your customers from the highest probability that they buy a certain product, to the lowest probability. Then you can target your marketing campaigns at the customers most likely to purchase.

SVM when you want to predict which **segment** your customers belong to.

Decision tree when you want to have **a clear interpretation** of your model results.

Random forest when you are just look for **high performance** with less need for interpretation.

**Question 3:**

How can I improve each of these models?

**Answer 3:**

To improve a model, you can carry out a process called Parameter Tuning, that will allow you to improve the performance of your models, by tuning them.

A model is composed of two types of parameters:

* The parameters that are learnt, i.e., coefficients.
* The hyperparameters.

The hyperparameters are the not learnt and require fix values inside model equations. For example, the regularization parameter lambda or the penalty parameter C.